Data driven approaches towards computational genome interpretation for identification of disease causing mutations

---

Data-gedreven strategien voor computationele genoominterpretatie om ziekte veroorzakende mutaties te identificeren

Thesis submitted in order to obtain the degree of Doctor in Medical Sciences from the University of Antwerp

Ajay Anand Kumar

Promoters:
Prof. Dr. Bart Loeys
Dr. Geert Vandeweyer
Dr. Maaike Alaerts

UNIVERSITEIT ANTWERPEN
Faculty of Medicine and Health Sciences

Members of the evaluation committee

Internal jury

Prof.Dr. Bart Loeys (UZA/ University of Antwerp)
Dr. Geert Vandeweyer (CMG,University of Antwerp)
Dr. Maaike Alaerts (CMG, University of Antwerp)
Prof.Dr. Kris Laukens (ADReM, University of Antwerp)
Prof.Dr. Wim Wuyts (CMG, University of Antwerp)

External jury

Prof. Dr. Christian Gilissen (Radboud UMC, Nijmegen)
Dr. Alejandro Sifrim (K.U Leuven/Sanger Institute,Cambridge)

कर्मण्येवाधिकारस्ते मा फलेषु कदाचन ।
मा कर्मफलहेतुर्भूर्मा ते संगोऽस्त्वकर्मणि ॥ ४७॥
द्वितीयोऽध्यायः, भगवद्गीता

You have rights for the action and never to its’ fruits. Don’t be impelled by the fruits of action, at the same time don’t be tempted to withdraw from the action.

Ch.2, verse 47, Bhagavad Gita

Summary

Deciphering the ground truth for any underlying physical or biological phenomenon is the core goal for any scientific endeavor. Formulating hypotheses, collecting evidence and subsequent reasoning formulate the core principles through which this truth is established. Generally, scientific reasoning could be classified as deductive or inductive. Deductive reasoning is the process in which initially a hypothesis is formulated together with expected consequences, and based on that supporting evidence is collected. On the contrary, for an inductive reasoning is a bottom-up approach which utilizes the data to infer the underlying truth by generating several plausible hypotheses. Over the years, deductive reasoning has been readily successful in practice and subsequently led to many scientific discoveries, but they suffer with limitations as the layers of complexity of the underlying system increase. For example, deciphering the underlying cause of human diseases is challenging, as the mechanisms by which the normal cell functions and interacts is full of complexities. The advent of next generation sequencing (NGS) technologies has paved the way to understand disease mechanisms at a much faster scale than anticipated earlier, thereby leading towards discovery of a great number of disease causing mutations. The success of NGS technologies can be seen as an example of an inductive reasoning process, where the underlying genomic data drives the formulation of suitable hypotheses by which the mechanism of disease can be explained. In the current thesis, I introduce data driven approaches for the identification of disease causing mutations, and address the challenges associated with their development.

In **chapter 1**, I provide a brief overview about the complexities underlying the human genome and the functionality of genes. The fundamental concepts behind classification of different types of genomic sequence variations such as single nucleotide variations (SNVs) and structural variations (SVs) in the genome that can cause disease, were introduced. Detecting these variations is a challenging task and it is equivalent to the problem of finding needles in a haystack. Hence I introduce a conceptual framework behind the implementation of linkage based analysis and NGS technologies for their

detection. These technologies are fast and generate high throughput genomic data. Hence it requires the development of automated procedural routines that formulate the roadmap from discovery of these variants to their functional interpretation. Gene prioritization and burden analysis constitute two of the most important strategies in this roadmap of causal SNV discovery. I present the ideas and challenges behind gene prioritization and burden analysis and how together these strategies can be used to identify causal SNVs. Additionally, I present conceptual ideas and strategies that need to be taken into account for the detection of copy number variations (CNVs) from NGS data.

After having introduced these concepts, I formulate the objective of my thesis in **chapter 2**. There are two main contributions to the field of medical genetics presented in this thesis. First, the development of novel computational tools for the interpretation of SNVs and genes associated with bicuspid aortic valve (BAV) with thoracic aortic aneurysm (TAA) disease from NGS data. Second, the development of a novel statistical method for identification of CNVs from targeted resequencing NGS data.

Chapters 3 & 4 presents the practical implementation of how conceptual ideas behind gene prioritization and mutation burden analysis on NGS data together helped in pinpointing the *SMAD6* gene to be associated with BAV/TAA disease. I present a novel gene prioritization tool (in **chapter 3**) named **pBRIT** which integrates 10 different annotation sources to prioritize candidate genes through a Bayesian regression model. I explored the utility of our method on several retrospective and prospective benchmark datasets and compared its performance with several existing methods. The dynamic implementation of **pBRIT** enables users to perform large scale exome prioritization and enables them to intuitively explore the results.

Mutation burden tests constitute an association based approach to identify disease associated rare variants. Combination of **pBRIT's** gene prioritization and burden analysis helped in elucidating the role of *SMAD6* gene as an important contributor towards the pathogenesis of BAV/TAA disease (see **chapter 4**). Additionally, it also exemplifies the applicability of data driven approaches in disease gene identification.

In chapter 5, we demonstrate the implementation of another novel statistical method named **varAmpliCNV** for the detection of CNVs from amplicon-based targeted resequencing NGS data. We describe how various technology-specific biases arising from enrichment design protocols can obfuscate the underlying CNV-related genomic signal. The novelty in our method is utilizing the design pattern of the enrichment protocol together with a PCA/MDS based method to control the variance present in the data. Comparison with three existing tools demonstrates the superior performance of our method with respect to speed, predictability and interpretation of CNVs.

Finally, we conclude that in the era of high throughput NGS the huge amount of data being generated stresses the necessity for the development of robust data driven approaches that can easily be scaled and generalized to a wide range of problems in the domain of genetics research. Formulating new hypotheses induced directly from the data alone is indeed innovative, but caution should be taken with respect to contextual interpretation and plausibility of the obtained results.

Samenvatting

Het onderliggende mechanisme van een fysisch of biologisch fenomeen volledig ontrafelen is het hoofddoel van elk (biomedisch) wetenschappelijk onderzoek. Een hypothese formuleren, gegevens verzamelen en analyseren en daaropvolgende kritische interpretatie van de resultaten, zijn de basisprincipes om dit te bereiken. In het algemeen kan elk wetenschappelijk proces gecategoriseerd worden als ’deductief’ of ’inductief’. Deductief is het proces waarin eerst een hypothese en verwachte consequenties worden geformuleerd en vervolgens data worden verzameld om deze hypothese te bewijzen of weerleggen. Daartegenover is een inductief wetenschappelijk proces een ’bottom-up’ benadering waarbij men kijkt naar wat de verzamelde data ’vertelt’ om mogelijke hypotheses op te stellen en zo de waarheid te achterhalen. Het deductieve redeneringsproces is in de praktijk zeer succesvol gebleken en heeft tot vele wetenschappelijke ontdekkingen geleid, maar het heeft zijn nadelen wanneer de complexiteit die aan de basis van een fenomeen ligt, toeneemt. Bij voorbeeld, de onderliggende oorzaken van ziekten bij de mens achterhalen en het mechanisme van deze ziekten ontrafelen is erg uitdagend, omdat cel fysiologie en cel-cel interacties een zeer complex interagerend systeem vormen. De ontdekking en toepassing van nieuwe generatie sequeneringstechnieken (’next-generation sequencing’, NGS) heeft de mogelijkheid gecreerd om ziekte veroorzakende moleculaire mechanismen op een veel snellere manier te onderzoeken dan verwacht, en heeft dan ook geleid tot de ontdekking van een groot aantal nieuwe ziekte veroorzakende mutaties. Het succes van de toepassing van NGS technieken kan gezien worden als een voorbeeld van een inductief proces waarbij de onderliggende genomische data tot een interessante hypothese leiden die het ziekte mechanisme kan verklaren. In deze thesis worden data-gedreven strategien gentroduceerd en de uitdagingen besproken die bij de ontwikkeling van deze strategien voor de identificatie van ziekte veroorzakende mutaties komen kijken.

In het eerste **hoofdstuk** geef ik een kort overzicht van de complexiteit van het humane genoom en de functie van genen. De fundamentele concepten die aan de basis liggen van de classificatie van de verschillende types varianten in de genoomsequentie die ziekte kunnen veroorzaken, zoals basepaarvarianten ('single nucleotide variant', SNV) en structurele varianten (SV), worden gentroduceerd. Het identificeren van deze ziekte veroorzakende varianten is een zeer uitdagende taak, gelijkaardig aan het zoeken naar een naald in een hooiberg. Ik introduceer het achterliggende conceptuele kader voor het gebruik van koppelingsanalyse en NGS voor deze variant identificatie. NGS technieken zijn zeer snel en genereren een enorme hoeveelheid genomische data. Hierdoor zijn er automatische methodes nodig die de volledige weg van het ontdekken van genetische varianten tot hun functionele interpretatie mogelijk maken. Gen prioritizatie en 'burden' analyse horen bij de belangrijkste strategien in dit kader. Ik stel de conceptuele ideen en de uitdagingen voor die samengaan met de ontwikkeling van gen prioritizatie en burden analyse technieken en hoe deze samen kunnen gebruikt worden om causale SNVs te detecteren. Daarnaast introduceer ik ook de conceptuele ideen en strategien die belangrijk zijn voor de detectie van 'copy number' variatie (CNV) in NGS data.

Nadat ik deze concepten heb voorgesteld, formuleer ik de doelstellingen van mijn thesis in **hoofdstuk 2**. De eerste doelstelling is de ontwikkeling van nieuwe computationele 'tools' voor de detectie en interpretatie van SNVs geassocieerd met bicuspiede aortaklep ('bicuspid aortic valve', BAV) in combinatie met thoracaal aorta aneurysma (TAA) uit NGS data. De tweede doelstelling is de ontwikkeling van een nieuwe statistische methode voor de identificatie van CNVs in NGS data van specifieke 'resequencing' genenpanels.

In **hoofdstukken 3 & 4** wordt de praktische implementatie beschreven van de conceptuele ideen en hoe gen prioritizatie en mutatie burden analyse uitgevoerd op NGS data samen bijdroegen tot de ontdekking van het *SMAD6* gen als mogelijke oorzaak van BAV/TAA. Ik stel een nieuwe gen prioritizatie tool voor, genaamd pBRIT (**hoofdstuk 3**), dat 10 verschillende annotatie databanken integreert om kandidaatgenen te prioritizeren door middel van een Bayesiaans regressiemodel. Ik onderzocht de toepassing van onze methode op verschillende retrospectieve en prospectieve 'benchmark' datasets en vergeleek de performantie met verschillende bestaande methodes. De dynamische implementatie van **pBRIT** stelt gebruikers in staat om prioritizatie uit te voeren op grote exoom datasets en om de bekomen resultaten intutief te exploreren.

Mutatie burden analyse is een test gebaseerd op statistische associatie die gebruikt wordt om zeldzame genetische varianten te selecteren die met de ziekte geassocieerd kunnen zijn. De toepassing van **pBRIT** tesamen met burden analyse leidde tot de ontdekking van *SMAD6* als een belangrijke speler in de pathogenese van BAV/TAA (**hoofdstuk 4**). Daarenboven toont het de toepasbaarheid van data-gedreven strategien aan om ziekte veroorzakende genen te identificeren.

In **hoofdstuk 5** bespreek ik de ontwikkeling en toepassing van een andere nieuwe statistische methode, genaamd **varAmpliCNV**, voor de detectie van CNVs in specifieke genenpanels die gebaseerd zijn op NGS van amplicons verkregen via een specifiek aanrijkingsprotocol. Ik beschrijf hoe verschillende verstorende factoren ('biases') die eigen zijn aan de technologie die wordt gebruikt voor de aanrijking, het onderliggende CNV-gerelateerde genomische signaal kunnen verdoezelen. Het vernieuwende aan onze methode is het gebruik van het specifieke design van het aanrijkingsprotocol samen met een PCA/MDS gebaseerde methode die de variantie in de data controleert. Een vergelijking met drie bestaande tools toont de superieure performantie van onze methode aan wat betreft snelheid, voorspelbaarheid en interpretatie van de CNVs.

Tot slot concluderen we dat in het tijdperk van hoge doorvoer NGS de enorme hoeveelheid data die hiermee gecreerd wordt robuuste data-gedreven strategien noodzakelijk maakt in genetisch onderzoek. Optimaal zouden deze strategien ook flexibel moeten zijn en toepasbaar op een brede waaier van vraagstellingen. Het formuleren van nieuwe hypotheses rechtstreeks uit data is inderdaad zeer innovatief, maar voorzichtigheid is steeds geboden in verband met contextuele interpretatie en de aannemelijkheid van de bekomen resultaten.

Contents

Summary vii
Samenvatting xi
Table of Contents xiii
List of Figures xix
List of Tables xxiii
List of abbreviations xxv

General Introduction 3
1 Introduction 5
- 1.1 The human genome 6
- 1.2 Genetic Variation 9
  - 1.2.1 Nature of variants and their classification 10
- 1.3 Genetic Disease 13
  - 1.3.1 Detecting disease-causing variants 13
  - 1.3.2 Genetic linkage analysis 14
  - 1.3.3 NGS Technologies 18
- 1.4 Methods for NGS Data analysis 20
  - 1.4.1 WES Analysis pipeline 20
  - 1.4.2 Gene Prioritization 21
  - 1.4.3 Mutation burden analysis 27
  - 1.4.4 Computational detection of CNVs 28

Aims and objectives 33
2 Aims and objectives 35

Gene prioritization and Mutation burden analysis 37
3 pBRIT: Gene Prioritization 39
- 3.1 Abstract 41
- 3.2 Introduction 41
- 3.3 Materials and Methods 43
  - 3.3.1 Internal representation of annotation sources 44
  - 3.3.2 Information-Theoretic model for feature mining 45

3.3.3 Modelling feature interdependencies and sparsity 45
3.3.4 Data Fusion 46
3.3.5 Prioritization using Bayesian ridge regression model 46
3.3.6 Cross-validation strategy 48
3.3.7 Validation datasets 50
3.3.8 Implementation of pBRIT 51
3.4 Results 51
3.4.1 BRR and SVD allows accurate and stable prioritization 51
3.4.2 Intermediate fusion provides uniform prioritization 53
3.4.3 Effect of annotation changes on prioritization 53
3.4.4 Effect of training set size and annotation bias 54
3.4.5 Real-World Performance Evaluation 54
3.4.6 Results exploration and visualization 55
3.5 Discussion 56
3.6 Conclusion 59
3.7 Acknowledgements 59
3.8 Supplementary Data 60
3.8.1 Annotation Sources used in pBRIT 60
3.8.2 Sparseness and dimensionality of annotation sources 62
3.8.3 Proportion of Variance Explained after SVD 62
3.8.4 Processing Annotation sources 64
3.8.5 Validation Strategy 64
3.8.6 Stability of AUC score for Goh et. al benchmark dataset 70
3.8.7 Benchmark results. 71
3.8.8 pBRIT benchmark performance metrics 74
3.8.9 Summary of performances 77
3.8.10 Negative Control Experiment 77
3.8.11 Effect of changing annotations on prioritization 79
3.8.12 Effect of number of training genes on prioritization 94
3.8.13 Visualization of gene prioritization through pBRIT 97
3.8.14 Toy example explaining how pBRIT works 99
4 SMAD6: BAV/TAA 105
4.1 Abstract 107
4.2 Introduction 107
4.3 Materials and Methods 108
4.3.1 Study Cohort 108
4.3.2 Targeted Enrichment 109
4.3.3 Data analysis and filtering 110
4.3.4 Validation by Sanger sequencing 110
4.3.5 Segregation analysis 111
4.3.6 Statistical Analysis 111

4.4 Results 111
4.5 Discussion 116
4.6 Ethics Statement 119
4.7 Author Contributions 119
4.8 Funding 119
4.9 Conflict of Interest Statement 120
4.10 Supplementary Material 120
    4.10.1 Selection of candidate genes from MIBAVA exomes 120

Statistical approach towards detection of CNVs **131**
5 varAmpliCNV: CNV detection **133**
5.1 Abstract 135
5.2 Introduction 136
5.3 Materials & Methods 139
    5.3.1 Validation sets 139
    5.3.2 VarAmpliCNV workflow 139
    5.3.3 Input files 140
    5.3.4 Read counting 141
    5.3.5 Quality control 141
    5.3.6 Read count normalization 141
    5.3.7 Controlling variance 141
    5.3.8 Computation of Ratio score (Log_2R) 142
    5.3.9 CNV calling using segmentation and filtering 142
    5.3.10 Annotation and validation of CNV segments 143
    5.3.11 Comparison with ONCOCNV, CoNVADING, DECoN (TAAD panel only) 143
    5.3.12 Computing optimal thresholds (TAAD panel only) 143
5.4 Results 144
    5.4.1 TAAD panel data analysis 144
    5.4.2 Application of DS-AOF on deafness panel data 147
5.5 Performance comparison with ONCOCNV, CoNVADING, DECoN (TAAD panel only) 148
5.6 Discussion 150
5.7 Conclusion 153
5.8 Supplementary data 153
    5.8.1 Information flow for the amplicon design patterns 153
    5.8.2 GC content correction 155
5.9 Controlling the variance using PCA/MDS 156
    5.9.1 Eigen Value decomposition of Covariance matrix is basic PCA analysis 158
    5.9.2 Using MDS gives identical result as PCA 159

5.10 Computing AOF 160
5.10.1 Application of DS-AOF 161

General discussion and conclusions 163

6 General discussion and conclusions 165
6.1 Complementary strategies for disease gene identification 166
6.1.1 Gene Prioritization 167
6.1.2 Mutation burden analysis 170
6.2 Copy number variation analysis 172
6.3 Interpretation of variants 174
6.4 Conclusion and future perspective 176

Bibliography 179

Bibliography 181

Acknowledgments 209

Curriculum Vitae 217

List of Figures

1.1.1 Representative structure of DNA. 7
1.1.2 Schematic representation of central dogma of molecular biology 8
1.3.1 NGS workflow and data analysis steps. 19
1.4.1 Prioritizing candidate genes. 22
1.4.2 Computational gene prioritization principles. 23
1.4.3 Categorization of CNV calling algorithms 30

3.2.1 Schematic workflow of pBRIT 44
3.3.1 Bayesian ridge regression 49
3.4.1 ROC plot of pBRIT benchmark performance 52
3.4.2 Impact of training set size 55
3.4.3 Exploring prioritization results using heatmap plot 56
S1.2 Proportion of variance explained by k components 63
S1.3 Schematic workflow for processing annotation sources 65
S3 Stability of AUC scores for the Goh et. al benchmark dataset 70
S4.1 LO-OCV result on Goh *et al.* and HyDRA benchmark datasets 71
S4.2 LO-OCV result on DisGeNET benchmark dataset 72
S4.3 Performance on prospective HPO benchmark dataset 73
S4.3 Summary of AUC scores for all the methods 77
S4.3 Negative Control Experiment 78
S5.1 Workflow for studying the effect of changes in GO and HPO based annotation sources on prioritization results 79
S5.2.1 Stratification of changes in the number of GO Terms and its effect in prioritization (TF-IDF_Test.N.Na) 82
S5.2.2 Stratification of changes in the number of GO Terms and its effect in prioritization (TF-IDF_Test.ALL.Na) 83
S5.2.3 Stratification of changes in the number of GO Terms and its effect in prioritization (TF-IDF→SVD_Test.N.Na) 84
S5.2.4 Stratification of changes in the number of GO Terms and its effect in prioritization (TF-IDF→SVD_Test.ALL.Na) 85
S5.2.5 Stratification of changes in the number of HPO Terms and its effect in prioritization (TF-IDF_Test.N.Na) 86
S5.2.6 Stratification of changes in the number of HPO Terms and its effect in prioritization (TF-IDF_Test.ALL.Na) 87

S5.2.7 Stratification of changes in the number of HPO Terms and its effect in prioritization (TF-IDF→SVD_Test.N.Na) 88
S5.2.8 Stratification of changes in the number of HPO Terms and its effect in prioritization (TF-IDF→SVD_Test.ALL.Na) 89
S5.2.9 Stratification of changes in the combined number of HPO and GO Terms and its effect in prioritization (TF-IDF_Test.N.Na) 90
S5.2.10 Stratification of changes in the combined number of HPO and GO Terms and its effect in prioritization (TF-IDF_Test.ALL.Na) 91
S5.2.11 Stratification of changes in the combined number of HPO and GO Terms and its effect in prioritization (TF-IDF→SVD_Test.N.Na) 92
S5.2.12 Stratification of changes in the combined number of HPO and GO Terms and its effect in prioritization (TF-IDF→SVD_Test.ALL.Na) 93
S6.1 Correlation plot between the number of training genes and AUC scores under the TF-IDF_Test.N.Na scenario during DisGenet based benchmarking 94
S6.2 Correlation plot between the number of training genes and AUC scores under the TF-IDF_Test.ALL.Na scenario 95
S6.3 Correlation plot between the number of training genes and AUC scores under the TF-IDF→SVD_Test.N.Na scenario 96
S6.4 Correlation plot between the number of training genes and AUC scores under the TF-IDF→SVD_Test.ALL.Na scenario 97
S7.1 Exploring prioritization results for *KCNA2* gene 99
S8.1 Computation of TF-IDF weights of an example binary valued annotation matrix 100
S8.2 Computation of gene-by-gene proximity profiles 101
S8.3 Computation of gene-by-gene proximity profiles using SVD 102
S8.4 Toy example demonstrating prioritization using Bayesian ridge regression. 103

4.4.1 Proportion of variant alleles per gene in the patient group, control group and ExAC cohort. 113
4.4.2 Graphical representation of the identified SMAD6 variants 115
4.10.1 Workflow regarding candidate gene selection through the WES pipeline 121
4.10.2 Resulting list of candidate genes selected through prioritization and burden analysis. 123
4.10.3 Family pedigree for segregation analysis 130

5.3.1 Workflow of varAmpliCNV 140
5.4.1 CNV evaluation using direct segmentation (DS) approach 146
5.4.2 Evaluation of CNVs using amplicon overlap filtering approach (AOF) . 147
5.4.3 Evaluation of CNVs using amplicon overlap filtering approach (AOF) and lenient pre-filtering 148

5.4.4 Evaluation of CNVs using amplicon overlap filtering approach (AOF) . 149
5.8.1 Flow of information in amplicon sequencing data. 154
5.8.2 Correlation plot of GC content fraction with target read counts 155
5.8.3 Correlation plot of GC content fraction with target read counts. 156
5.10.1 Flow of information in amplicon sequencing data. 161
5.10.2 Visualization of an example CNV segment using DS-AOF approach . . 162

List of Tables

1.1 Summary of available linkage analysis tools. 18
1.2 List of freely available gene prioritization tools 25

S1 Annotation sources used in pBRIT 60
S1.1 Dimensionality and degree of sparsity 62
S4.1 Performance of pBRIT on the Goh *et. al* benchmark set 74
S4.2 Performance of pBRIT compared to Endeavour, ToppGene and HyDRA 74
S4.3 Performance of pBRIT using DisGeNET disease database 75
S4.4 Performance of pBRIT on novel disease genes 75
S4.5 Comparative performance on subset of 693 HPO benchmark data set . . 76

4.1 Genes included in the targeted gene panel and the criteria on which their selection was based. 109
4.2 Variant burden comparisons per gene between patients and either study controls or ExAC controls. 114
4.3 List of training genes for prioritization of MIBAVA exomes 121
4.4 Patient cohort overview. 122
4.5 Overview of the identified variants in the genes from the targeted gene panel and their phenotypic data 129

5.1 Performance comparison of varAmpliCNV with three competing methods.150

List of abbreviations

AOF	Amplicon overlap filtering
array-CGH	array-based comparative genomic hybridization
AS	De-novo assembly
AUC	Area under the curve
BAM	Binary aligned map
BAV/TAA	Bicuspid aortic valve associated Thoracic aortic aneurysm
BDA	Big data analytics
BLAST	Basic Local Alignment Search Tool
bp	basepairs
BT	Boundary threshold
CAFA	Critical assessment for functional annotation
CAST	Cohort allelic sum test
CBS	Circular binary segmentation
CCA	Canonical correlation analysis
CMAT	Cumulative minor-allele test
CMC	Combine multivariate and collapsing
CNVs	Copy number variations
DAG	Directed acyclic graph
DGV	Database of genomic variants
DIR	Direct integration of ranks
DL	Deep learning
DNA	Deoxy Ribonucleic Acid
DO	Disease ontology
DOC	Depth of coverage
DS	Direct segmentation
DT	Detection threshold
EL	Elston-Stewart algorithm

EVD	Eigen value decomposition
ExAC	Exome aggregate consortium
FISH	Fluorescence in situ hybridization
FNs	False negative
FPs	False positives
GAD	Genetic Association Database
GATK	Genome analysis tool kit
GBA	Guilt by association
gnomAD	Genome aggregation database
GO	Gene ontology
GWAS	Genome wide association analysis
HC	HaplotypeCaller
HPO	Human phenotype ontology
HuGe	Human Genome Epidemiology
ICA	Independent component analysis
Kb	Kilobases
LG	Lander-Green algorithm
LOD	Logarithm of Odds
LSM	Latent semantic model
MAF	Minor allele frequency
MAQ	Multiplex amplicon quantification
MCMC	Markov-Chain Monte Carlo
MDS	Metric dimensional scaling
MIBAVA	Mechanistic interrogation of Bicuspid aortic valve associated aortapathy
MKL	Multiple kernel learning
MLPA	Multiplex Ligation-dependent probe amplification
MPO	Mammalian phenotype ontology
MRF	Markov random field
mRNA	messenger Ribonucleic Acid
MRR	Mean rank ratio
NGS	Next generation sequencing
NMD	Nonsense mediated decay
NPL	Non-parametric linkage analysis
pBRIT	Prioritization using Bayesian ridge regression and information theoretic model
PCA	Principal component analysis

PCR	Polymerase chain reaction
PCs	Principal components
PEM	Paired end mapping
PPI	Protein-protein interactions
pre-mRNA	precursor messenger RNA
RC	Read count
RD	Read depth
ROC	Receiver operation characteristics
ROI	Region of interest
RVAS	Rare variant association analysis
RWR	Random walk with restart
RWR-M	Random walk with restart on multiple heterogeneous networks
SKAT	Sequence kernel association test
SNPs	Single nucleotide polymorphisms
SNVs	Single nucleotide variants
SO	Sequence ontologies
SR	Split-read
ST	Segmentation threshold
SVD	Singular value decomposition
SVs	Structural variants
TAA	Thoracic aortic aneurysm
TAAD	Thoracic aortic aneurysm and dissection
Test.ALL.Na	Test gene all NA (phenotype score)
Test.N.Na	Test gene not NA (phenotype score)
TF-IDF	Term frequency Inverse document frequency
TF-IDF$\rightarrow$SVD	SVD transformed TF-IDF matrices
TFs	Transcription factors
TGF-β	Transcription growth factor beta
TPs	True positives
TR	Targeted re-sequencing
UG	UnifiedGenotyper
VCF	Variant calling file
VEP	Variant effect predictor
VT	Variable minor allele frequency threshold
WES	Whole exome sequencing
WGS	Whole genome sequencing

WS	Weighted Sum

General Introduction

Chapter 1

Introduction

> *What I cannot create, I do not understand*
>
> Richard Feynman

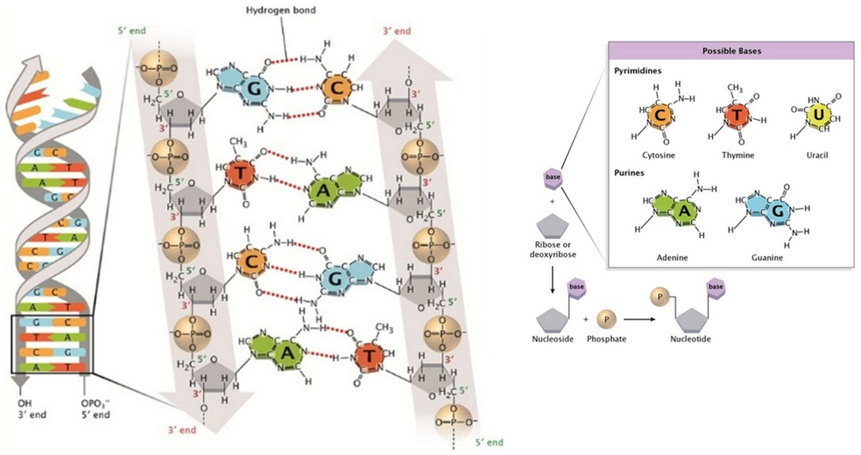
The scientific outlook in the second decade of the 21st century got revamped by the landmark discovery of the Higgs-Boson particle in 2012 and gravitational waves in 2016. At the level of sub-atomic space or quantum realm the Large Hadron Collider (LHC) experiment investigated why some fundamental particles in nature have mass. Whereas at the macroscopic level detection of gravitational waves provided the experimental justification of Einsteins theory of general relativity explaining that the macroscopic mass bends the space-time curvature. An interesting observation that can be deduced from both these discoveries is that the theoretical models for these phenomena were proposed many decades before the technological advances could experimentally substantiate them. The success of the development of theoretically sound models for such sub-atomic and macroscopic object entities was only possible because these entities have very unique, orderly and structured behavior under the constraints of the natural physical law. Because of this orderliness behavior the formulations of these models have universal appeal. However, can this similar approach of theoretical derivation of models be applied in sub-macroscopic space?

A sub-macroscopic space is the length scale that lies between the sub-atomic and macroscopic space which comprises of the molecules, their environment and interactions, together constituting an ever evolving, dynamic biological entity. Understanding the functioning of these biological entities is a challenging task due to a wide range of concerns, from prospective to predictive and causal questions addressing the complexities involved in deciphering their inherent mechanisms . For example, finding the mechanism by which diseases affect humans, so far the theoretical models that have been developed have shortcomings in effective and precise explanation of the underlying cause as no single hypothesis can be universally applied to explain the disease mechanism. The advent of 'omics' technologies has led to an alternative approach called data-driven hypothesis generation which is principally based on let the data speak for itself without any a priori assumptions. It provides a multi-view perspective to understand human diseases at the sub-macroscopic level. The high-throughput generation of biological data through different 'omics' technologies has increased the demand for development for large-scale statistical frameworks that can explain these intricate relationships, correlations and causations between the data and ultimately explain the underlying biological phenomenon. The success of data- driven approaches can be best demonstrated by the LHC collision experiment where mining 15 million gigabytes of data led to the discovery of the Higgs-Boson particle, thereby validating the theoretical predictions.

This thesis is an attempt to address the development and implementation of data-driven approaches towards computational interpretation of human genome data to identify causal genes for human diseases. Section 1.1 presents the basic architecture of the human genome thereby explaining its complexities and the flow of information from genes to their functional form. Section 1.2 explains basic concepts and definitions regarding different types of sequence variations in the genome and how they can be involved in genetic diseases. Details about the methods to detect disease-causing genetic variants are presented in section 1.3. Furthermore, section 1.4.2 - 1.4.4 presents novel statistical frameworks that implement data-driven principle to computationally interpret the human genome, which in fact substantiates the ultimate aim of this thesis.

1.1 The human genome

From the simplest to the most complex organism, the life on planet earth has evolved through the fundamental principle of inheritance and adaptation to the environment. The unit of heredity is called a gene and was first described by Gregor Mendel in 1865 and can exhibit different phenotypic forms called alleles. In mid-40s the basic element of the gene was discovered where Avery et. al.[11] demonstrated that the phenotypic expression of bacterial colonies could be transmitted from cell to cell through deoxy ribonucleic acid (DNA) alone. Subsequently, the experiments from Watson and Crick in 1952 [221] elucidated the structure of DNA that eventually opened the door to understand the mechanism how this double stranded helical structure can act as the agent of inheritance.



A Double-helical structure of DNA **B** Composition of nucleotide bases

Figure 1.1.1: **Representative structure of DNA.** (A) The three dimensional structure of DNA where nucleotide bases (A, G, C, and T), based on their complementarity are held together as a pair by hydrogen bonds get stacked to give a double helical structure. (B) The nucleotide bases are made up of three components: a nitrogen-containing base, a five-carbon sugar (a ribose sugar in RNA or a deoxyribose in DNA) and a phosphate group. The nitrogenous base are further classified as purine or pyrimidine base. (Figure adapted from article by Leslie A Pray[173]).

DNA consists of a pair of strands of a sugar-phosphate (deoxyribose) backbone attached to a set of pyrimidine and purine bases. Each DNA strand consists of alternating deoxyribose molecules connected by phosphodiester bonds from the 5′ position of one deoxyribose to the 3′ position of the next. The two DNA strands are bound together by hydrogen bonds between adenine (A) and thymine (T) bases and between guanine (G) and cytosine (C) bases leading to a double stranded helical structure (Figure 1.1.1). The complementary matching of the bases (A−T and G−C) through hydrogen binding enables the DNA replication process using one strand as a template to generate or replicate a new strand. The compact stacking of the nucleotide bases in the double stranded DNA serves the purpose of coding of genetic information. The biological function is further enabled by another class of biomolecules called proteins which are the end products of the translation mechanism. Proteins consist of a chain of amino acids, each of which is encoded by a triplet of nucleotides, also called codons (see Figure 1.1.2), corresponding to one of the 20 different amino acids, a startcodon or one of the stopcodons. This genetic code is degenerate or redundant because more than one codon can code for the same amino acid. The sequence of codons are first transcribed as ribonucleic acid or messenger RNA (mRNA) by a protein called RNA polymerase. The mRNA is then translated into proteins by the ribosomal machinery. Overall, the process of flow of information through transcription of the DNA code to translation to a polypeptide chain of amino acids is called the central dogma of molecular biology as shown in Figure 1.1.2

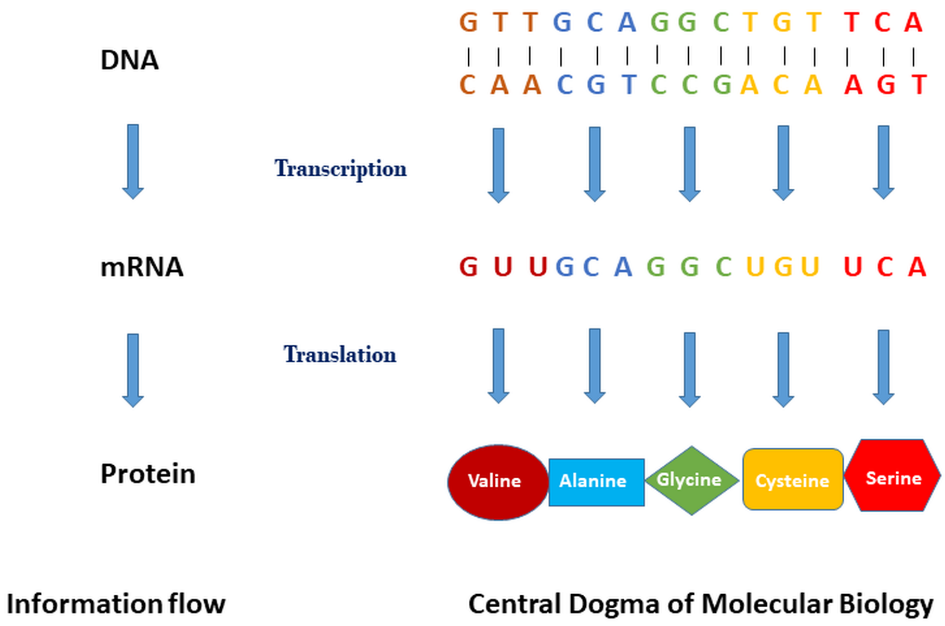

G T T G C A G G C T G T T C A
DNA
C A A C G T C C G A C A A G T
Transcription
mRNA
G U U G C A G G C U G U U C A
Translation
Protein
Valine
Alanine
Glycine
Cysteine
Serine
Information flow
Central Dogma of Molecular Biology

Figure 1.1.2: **Schematic representation of central dogma of molecular biology**. It involves a two step process of transcription and translation, by which the information in genes flows into proteins mediated by intermediary mRNA.

With the exception of germline cells, red blood cells and platelets, every human cell contains two copies of the genome, one maternal and one paternal, therefore also called a diploid genome. The genome is organized into 46 chromosomes, arranged in 22 pairs of autosomal chromosomes and a set of sex chromosomes called the X and Y chromosomes (female: XX, male: XY). At any given locus of a chromosome, the homologous chromosome of that pair, has either identical or slightly different forms of the same gene (with differences in the sequence of bases), called alleles. The set of alleles that collectively constitute the genetic makeup either at all loci or at single locus gives the genotype of the individual. Similarly, the phenotype is an observable expression of genotype as morphological, clinical, cellular or biochemical trait. If an individual has identical alleles at a given locus then this is called homozygous and if he/she carries different alleles this is called heterozygous with respect to that locus. A typical haploid human genome consists of approximately 3 billion nucleotides. So far it has been estimated that roughly 1% of the genomic sequence is constituted by approximately 22000 to 25000 protein coding genes. The remaining 99% of the genome was earlier considered to be junk-DNA, but recent large scale projects such as Encyclopedia of DNA Elements[63] (ENCODE) have deciphered important functions of this non-coding DNA, such as regulating gene expression.

1.2 Genetic Variation

No two individual genomes are identical. The composition of base sequence in the genome varies from one individual to another, defining genetic variability. The creation of genetic variation is a natural process occurring continuously in both germline and somatic cells. The study of genetic variation can be done from an evolutionary perspective e.g. to investigate migration of ancient human populations and for medical applications e.g. to determine the molecular basis of genetic diseases. In this thesis we mainly focus on the genetic variants potentially causing human disease. Many human diseases have a genetic component. The contribution of this genetic component can range from small, modest to very large in some diseases. For example, diseases such as Huntington disease[5, 155, 164], Tay-sachs disease[160] and cystic fibrosis[179] exemplify the consequence of strong genetic variation (so called mutations). Whereas, the contribution of the genetic component for diseases such as cancer [143], diabetes[82], cardiovascular diseases[13] is modest as other factors which include environment and lifestyle also contribute towards their etiology. Lastly, for most of the infectious diseases there is no or an extremely small contribution of a genetic component with a potential increase in susceptibility for infection in some individuals[39].

Independent assortment, DNA mutation, and recombination (via sexual reproduction) events are some of the causes of genetic variation. Natural selection is the key driving force that results in interaction between genetic variants and environment contributing towards the variations in the genome. The causes behind occurrence of mutations can be broadly categorized as: (a) spontaneous mutations arising due to molecular decay, (b) DNA replication error, (c) translation error, (d) error prone DNA repair mechanisms and finally (e) mutagens such as exposure to radiation and chemicals that can induce breaks in DNA strands. Genetic recombination contributes towards genetic diversity during meiosis, where pairs of homologous chromosomes exchange genetic material. This passes on from parent to offspring, thereby generating a new set of genetic combinations. Recombination events can occur at any location in the genome and the probability is directly linked to the distance between two loci. Complex recombination processes can lead to large scale alteration of chromosome structure including insertions, deletions, inversions and translocation events. In subsequent sections these different types of variants are discussed along with their consequences.

1.2.1 Nature of variants and their classification

Variants can be classified based on their size/nature and their location in the genome which subsequently defines their functional effect. Substitutions, insertions or deletions of one base pair are generally called point mutations or Single Nucleotide Variants (SNVs). Nucleotide substitutions involve alteration in the sequence but not the number of nucleotides. If pyrimidine and purine bases get substituted to the same chemical category it is called a transition. In contrast, when purine and pyrimidine bases are interchangeably substituted this is called transversion. Deletions and insertions of a few nucleotides are called indels. Large scale deletions, insertions or translocations of genomic segments leading to alteration in the overall structure of the genome are called Structural Variants (SVs). Some SVs are classified as Copy Number Variations (CNVs) because the number (copies) of these segments present in the genome differs from the normal two copies.

Variation in the coding region: Nucleotide level variation can have functional consequences. As per the central dogma of molecular biology, genes are transcribed and translated to yield proteins which in fact determine the functionality of the gene. Variation in a single nucleotide can alter the amino acid chain and affect the final gene product. When a point mutation in the codon leads to the identical amino acid in the polypeptide chain as the wild type it is defined as a synonymous or silent variant. Whereas if such point mutation leads to substitution to a different amino acid it is called a non-synonymous variant or missense mutation because it alters the ″sense″ of the coding strand. Synonymous variants are generally considered to be benign or have neutral effect. However, there is also plenty of evidence that synonymous SNVs can affect the functionality of the gene by altering splicing and gene expression[40, 42, 183]. They may alter the splicing patterns, or change the structure of pre-mRNA by altering the folding energy, which subsequently affects the translation dynamics. E.g. synonymous mutations in the *CFTR* gene associated with cystic fibrosis were found to alter the translation efficiency[14, 188].

The functional impact of non-synonymous variations depends upon which region of the protein is affected and the properties of the amino acids. When an amino acid important for the functionality of the protein is substituted with another one with similar chemical characteristics this is known as a conservative substitution. In this case the hydrophobicity and molecular bulk of the amino acid side chain remains unaltered. For example, replacement of an aspartate (Asp) with glutamate (Glu) is a conservative substitution as they are both negatively charged amino acids. On the other hand, a semi-conservative substitution involves replacement of an amino acid with another one that results in a similar steric conformation of the protein, but its overall biochemical property is changed. For example, substitution of cysteine (Cys) for alanine or leucine (Leu). Finally, a non-conservative substitution involves a change in amino acids with radically different chemical properties. Substitution of valine (Val) to arginine (Arg) demonstrates a non-conservative substitution.

Another class of non-synonymous variants are known as nonsense variants in which a point mutation generates a stop codon when actually there was none (stop gain). Alternatively, exchanging an existing stop codon for an amino acid coding codon is called stop loss. The stop gains may lead to degradation of the messenger mRNA (process called nonsense mediated decay, NMD). The stop losses can produce an appended polypeptide chain. Point mutations can as well alter the specific sequences important for mRNA processing steps, including 5′ capping, polyadenylation and splicing. Mutations located near (approximately 50 bp) the exon-intron (5′ donor site) or intron-exon boundaries (3′ acceptor site) can disturb splicing and are known as splice-site mutations. Small indels where the number of bases involved is not a multiple of three leading to an alteration of the reading frame are called frameshift mutations. Indels that remove or add a number of bases divisible by three can cause increase or decrease of expression of genes.

Variation in non-coding region: Approximately 80% of the non-coding region of the genome potentially plays a role as regulatory elements[63]. The ENCODE project provided useful insights about the nature of transcription, chromatin structure and histone modification in the human genome. The noncoding genomic DNA includes functional RNA genes, introns, pseudogenes, repeat sequences, transposons and telomeric repeated elements. Noncoding RNA ranging from small microRNAs (miRNA) to long non-coding RNAs (lncRNA) regulate the translation activity of protein coding genes. For example, genome wide association studies (GWAS) identified high risk schizophrenia associated SNVs/SNPs in miRNA[59] and lncRNA[21] and deciphered their putative role in reducing the expression of coding genes.

Enhancers are generally short (50-1500bp) regions that can be bound by proteins (activators) or transcription factors (TFs) thereby increasing the transcription of particular genes. Generally these enhancers are cis-acting which means they regulate nearby genes, but they can also be located very distant from the gene they regulate. These enhancers are scattered across 98% of the genome[216] and their location relative to the target gene(s) is also highly variable as they can be found in upstream, downstream and also within introns of the gene. In comparison to sequence of protein coding genes the sequence code of enhancers is poorly understood, which makes their computational identification from the DNA sequence very challenging. The trans-regulatory elements are the DNA sequences that encode the TFs. They regulate the expression of gene(s) distant from the genes from which they were transcribed. There can be often more than one or more trans-acting factors that can bind to the cis-regulatory elements.

During the transcription process a precursor messenger RNA (pre-mRNA) is produced that has exons and introns. The introns are spliced from the pre-mRNA to yield mature mRNA by a process called RNA splicing. The precise excision of introns require ribonucleoprotein machinery called the spliceosome. Additionally, the intron-exon boundaries are delimited by short consensus sequences at 5′ (donor) and 3′ (acceptor) splice sites that modulate the recognition of the spliceosome. Point mutation in or near the intron exon boundaries or the consensus sequence recognized by the spliceosome can therefore alter the splicing pattern of pre-mRNA and are also recognized as a causal mechanism for hereditary disorders[41, 167]. Also mutations in the binding sites for splice regulatory proteins contribute towards aberrant splicing which can lead to a disease phenotype[37].

SVs in the form of duplications, deletions and inversions in the intergenic regions can also result in disease phenotype by altering the regulatory mechanism. For example presence of SVs in locus spanning *WNT6/IHH/EPHA4/PAX3* gene resulted into variable phenotypes of limb malformations[137]. It was found that the pathogenic CNVs (deletions) in 250kb upstream of *FOXF1* gene resulted in lethal lung development disorder[199]. The deletions in this cis-regulatory region which also harbors lncRNA genes affect the regulation of the *FOXF1* gene. Although evidence is accumulating regarding the various roles these non-coding sequence play in disease gene associations, given the amount of complexities involved in deciphering their regulatory mechanism the interpretation is generally limited. Hence in most of the genetic research studies non-coding regions are generally ignored, but with application of whole genome sequencing (WGS), information about these regions should definitely be accounted for understanding of the underlying disease etiology.

Mutations in coding and non-coding regions produce variation in the human genome. Natural selection is the driving force that determines which of these mutations will be favored and passed on to the next generation and which are unfavorable and get eliminated. Each new mutation creates a new allele that can be characterized by the selection coefficient, which measures the expected change in an allele's frequency over time. Therefore the knowledge of mutational rates is key for understanding the progress and evolution of diseases especially in case of cancer and inherited disorders[206]. The mutation rate is generally expressed as the number of new mutations occurring per locus per generation. For a given gene it varies from 10^{-4} to 10^{-7} mutations per locus per generation. The reason for this variation is attributed to: (a) gene size, fraction of mutant alleles that gives particular observable phenotype, (b) the age and gender of the parents, (c) the mutational mechanism and (d) presence or absence of mutational hotspots such as methylated CG nucleotide repeats in the gene. Evolutionary process has ensured that there is steady influx of new nucleotide variants adding to a high degree of genetic diversity. For protein coding regions the mutation rate is much lower and variants are under rigid selective pressure during the course of evolution. When a variant is very common with a frequency of $>1\%$ of chromosomes in the general population then it is categorized as genetic polymorphism. Whereas alleles having a frequency less than 1% are conventionally classified as rare variants. The allele frequency seldom does not allow classification of any variant to be deleterious as many of the rare variants appear to have no deleterious effect and some of the common polymorphic variants are found to increase susceptibility to disease.

In the next section we discuss about how these rare variants are involved in genetic diseases and which techniques can be used to identify these variants.

1.3 Genetic Disease

The expression of any human phenotypic characteristic depends upon the genotype at the locus or loci that are associated with it. There can be one or multiple genes playing a role along with other factors such as environment, age of onset, and demography for complete manifestation of the disease. In the previous section, we described different types of variants which can lead to manifestation of the disease phenotype. Based on their etiology and involvement of mutations in the number of genes these disorders can be broadly classified as monogenic or Mendelian disorder (involvement of single gene) and oligogenic or polygenic disorders (involvement of multiple genes) or multifactorial disorders (involvement of multiple genes and environment).

1.3.1 Detecting disease-causing variants

We discussed in the previous sections the complexity of the human genome and different types of genetic variants that can lead to dysfunction of genes. At the molecular level, the functional products of genes, e.g. proteins interact in networks by exhibiting several interdependent functionalities and at the macro level their function is governed by several factors such as environment. Given the complexities of human genome architecture and mechanisms by which the genomic elements interact, the precise determination of causal variants regarding their role in research on and diagnosis of genetic diseases is a challenging task and can be equated to finding the needle in the haystack. The identification of the causal mutation and/or gene helps in elucidating the disease mechanism, improving molecular diagnosis, and ultimately designing better drug therapies. Traditionally hypothesis driven approaches such as functional cloning based approaches, based on sequence or structural similarity to known genes, were used as probe to identify new genes that might share similar functions. Alternatively, the positional cloning based approaches which include linkage analysis, were developed and used the exact chromosomal location instead of function to guide gene identification. With advent of DNA sequencing technology the automated generation of high throughput genomic data has provided an opportunity to perform large scale genome wide association studies to detect causal variants. In the following sections, we will focus on conceptual ideas behind genetic linkage analysis and the possibilities that next generation sequencing has opened for gene identification.

1.3.2 Genetic linkage analysis

Linkage analysis is based on the principle that during meiosis DNA sequences that lie close together on a chromosome get inherited together. During first stage of meiosis, the homologous chromosomes pair up along the mitotic spindle. The paternal and maternal homologues exchange homologous segments by crossing over and create new chromosomes consisting of alternating portions of ancestral chromosomes. This process is known as chromosomal recombination. Linkage analysis theoretically deviates from classical Mendelian law of independent assortment which states that alleles for separate traits are passed independently of each other from parents to siblings. Alleles at loci on different chromosomes assort independently and alleles at loci on the same chromosome assort independently if at least one crossover occurs between them in every two meiosis. Thus, the frequency of recombination (θ) provides a measure on how far apart are two given loci. If two loci are closely linked with no recombination happening between them then $\theta = 0$ and $\theta = 0.5$ if they are far apart with 50% chance of recombination happening, as if they are located on different chromosomes. The lower the frequency, the closer the two loci are located and vice versa. Such loci are said to be tightly linked if they are physically located very close to each other and unlinked when they are far apart and thus can assort independently. Measurement of the θ requires statistical methods to determine it accurately and reliably from the given familial data. Statistical determination of linkage relies on two steps. First, it is ascertained whether the recombination fraction θ between two loci deviates significantly from 0.5 and second, if θ is less than 0.5 it is calculated what will be the best estimate of θ that may explain how close or far apart the loci are.

Linkage analysis is computationally performed on a pedigree. The members of each pedigree are divided into founder and non-founder members. A set of known markers for which the set of possible alleles and their frequencies in the population are known is chosen. Additionally, the recombination fraction between these markers is also determined. Next, individuals of the pedigree who are accessible are genotyped for these markers. This results in a set of sequence of unordered allele pairs one for each marker. For a given pedigree it is often difficult to determine which of these alleles comes from which of these parents because not all the individuals could be genotyped. Therefore, precise knowledge of pattern of inheritance in the pedigree helps in determining the assignment of these unordered alleles to parents through a probability distribution. From a given family pedigree data, the number of children that show or do not show any recombination events between a set of loci are counted and the likelihood of observing the data at various possible values of θ between 0 and 0.5 is computed. Next, under the null hypothesis that two loci are unlinked ($\theta = 0.5$) the second likelihood is computed. The ratio between these two likelihoods is defined as the logarithm of odds or precisely as LOD score. This is given by:

$$\textbf{LOD score (Z)} = \textbf{Log}\left(\frac{\text{Likelihood of the data if loci are linked at particular } \theta}{\text{Likelihood of the data if loci are unlinked}}\right) \tag{1.1}$$

The strength of linkage can be assessed from the values of LOD score. Any score above 3.0 (equivalent to 1000:1 odds in favor of linkage) is considered as definitive evidence of strong linkage. Score between 1.81 and less than 3.0 it is considered as suggestive or non-conclusive evidence of linkage and any value which is less than -2.0 is considered as no evidence for linkage. For Mendelian diseases where the pattern of inheritance can be well defined, model-based or parametric linkage analysis is generally performed. The pedigree tree, the affection status, frequency of disease alleles and penetrance are necessary parameters required for parametric linkage analysis. The estimated LOD scores are highly sensitive to any of these parameters. In case of complex disease traits where the pattern of inheritance is unknown, a model-free analysis is performed and called non-parametric linkage analysis. For qualitative or quantitative traits the computation of non-parametric LOD (NPL) score allows mapping of disease genes in which variants contribute towards susceptibility for diseases or towards physiological measurements. NPL scores evaluates for allele sharing among affected individuals. These individuals are related to each other such as relatives or pair of siblings in a family who show greater similarity for any quantitative trait. If the allele sharing at a certain polymorphic marker is significantly higher than expected, then it is indicative that the disease locus is closely located near the marker. The evidence of this increased allele sharing is quantified through the NPL score of 3.6 and strength of this evidence is given by score greater than 5.4.

Combining all of these information through a marker map where each of these markers are investigated for linkage. It basically forms the underlying core principle of most of the available computation tools for linkage analysis. If only one marker is investigated independently of other loci it is called single-point analysis. When the information about the neighboring markers is taken into account it is called multi-point analysis. Most of the contemporary methods for doing multi point linkage analysis are based on Elston-Stewart[61] (EL) or Lander-Green algorithm[55] (LG). For a marker map with 6-8 markers EL is suitable whereas with LG thousands of markers on any chromosome can be analyzed. However, with increased size of pedigree or higher density of markers, the computational complexity of these algorithms also rises. Hence, as alternative to these algorithms, Markov-Chain Monte Carlo (MCMC) based methods have been developed which approximate the linkage likelihoods and are non-deterministic in nature. A short summary of various available methods for linkage analysis is presented in Table 1.1. The table provides details on the method of analysis and the ease of usability. For most of the practical applications, it is indicated to use more than one software program in order to formulate a consensus report on the detected region. For example, findings from one of deterministic algorithm such as Merlin should be replicated several times using approximation based methods such as Simwalk2 in order to report the true consensus peak of the LOD score as demonstrated by Monteferrario *et. al.*[150].

Linkage Analysis tools	**Method description**	**Availability**	**Accessibility**
easyLINKAGE[133]	Software suite that integrates several programs such as Allegro, Merlin, SimWalk, GeneHunter, SuperLink, FastLink, SPLink to perform linkage analysis. Also generates input files for drawing pedigree, colored representation of markers, position and haplotypes through external programs such as HaploPainter[204]	`https://sourceforge.net/projects/easylinkage/`	Executable binaries available for Windows OS platform with an interactive user-interface. Cannot be implemented for parallelization.
GeneHunter[110] **Allegro**[75]	It implements Lander-Green algorithm. Suitable for both parametric and non-parametric linkage analysis. It can handle medium size pedigrees and large number of markers. Provides extensive set of association tests. Allegro is faster version of GeneHunter having 30 fold increase in execution time.	`http://www.broad.mit.edu/ftp/distribution/software/genehunter/` **Allegro:** Currently deprecated	Binaries available for Linux OS platform. Command line execution.
LINKAGE[122] and **Fastlink**[123]	It implements Elston-Stewart algorithm. Suitable for parametric linkage analysis. First tool to perform multi-point linkage analysis. It can handle large pedigree structure but has limited performance with large set of markers. Additionally, it can be used to detect Mendelian-inconsistent genotyping errors.	`http://linkage.rockefeller.edu` (LINKAGE) and `https://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html`	Binaries available for Linux and Windows NT, 2000, Vista, XP OS platforms. Can be parallelized in the cluster computing nodes.
Mendel[119]	It implements Elston-Stewart and Lander-Green algorithm. Provides both parametric and non-parametric linkage analysis for both pedigree and population based data. It also provides a large range of statistical test for association analysis.	`https://www.genetics.ucla.edu/software/mendel`	Binaries available for Linux./Mac/Windows OS platforms. Suitable for parallelization. Graphical user interface (GUI) available.
Merlin[2]	It implements Lander-Green algorithm. Provides parametric and non-parametric linkage analysis. Suitable for medium-sized pedigrees. Execution time increases with large number of markers.	`http://csg.sph.umich.edu/abecasis/merlin/index.html`	Binaries available for Linux/Windows OS platforms. Command line execution
Plink[177]	Software tool suite for whole genome association analysis for genotype/phenotype data. Can be used for data management, population stratification testing, Case/Control and family based association tests, haplotypic tests, CNV analysis.	`http://pngu.mgh.harvard.edu/purcell/plink/`	Binaries available for Linux and Windows platforms. Command line execution.
SEQLinkage[219]	It implements Elston-Stewart algorithm. Provides parametric linkage analysis for the WGS data using collapsed haplotype pattern method. It takes sequence data in VCF format and perform two-point linkage analysis.	`http://bioinformatics.org/seqlink/start`	Binaries available for Linux and Mac OS platform. Command line execution

SimWalk2[197]	It is based on MCMC algorithm. Provides both parametric and non-parametric linkage analysis. It can handle large pedigrees and an intermediate number of markers.	`http://www.genetics.ucla.edu/software/`	Binaries available for Linux and Windows OS platform. Command line execution. Suitable for parallelization in cluster computing framework.
TLINKAGE	It implements Elston-Stewart algorithm. Extension of LINKAGE program thereby suitable for only parametric linkage analysis. It can handle large pedigrees.	`http://www.jurgott.org/linkage/tlinkage.htm`	Binaries available for both Linux and Windows OS platforms. Command line execution. Suitable for parallelization.

Table 1.1: **Summary of available linkage analysis tools.** The tabulated presentation of 9 most frequently used software programs for linkage analysis. For each of these tools a short description about the underlying algorithm that has been implemented along with various analysis these programs offer is presented. Additionally, the availability and accessibility regarding ease of usage is also indicated.

1.3.3 NGS Technologies

Although historically the positional cloning based approaches which include linkage analysis have been instrumental for disease gene identification, recent technologies based on DNA sequencing have revolutionized the field. DNA sequencing is a process of precisely determining the order of nucleotides within a DNA molecule. The first generation sequencing technology or Sanger sequencing is based on incorporation of dideoxynucleotide chain terminators followed by electrophoretic separation of chain-termination products of individual sequencing reactions. Current Next generation sequencing (NGS) technologies across different platforms is characterized by parallelizing millions of sequencing reactions thereby generating high throughput sequencing data. NGS is fast and scalable with an impressive increase in throughput, accuracy and coverage of the targets.

The different NGS platforms such as Illumina/Solexa, Roche 454, Life Technologies SOLiD, Life Technologies Ion Proton Helicos Bioscience, PacBio provide commercial kits for massive parallel sequencing and differ in their exact sequencing protocols. The internal functioning of these technologies involves three major steps: first, generation of sequencing libraries, second by parallelizing the amplification of template DNA and finally sequencing the DNA by synthesis yields high throughput sequencing data. Subsequently using commercial software frameworks or customized bioinformatics pipelines the data is processed for identification of candidate genes. An example workflow describing usage of different NGS strategies and various stages of sequencing data analysis to detect causal variants is presented in Figure 1.3.1.

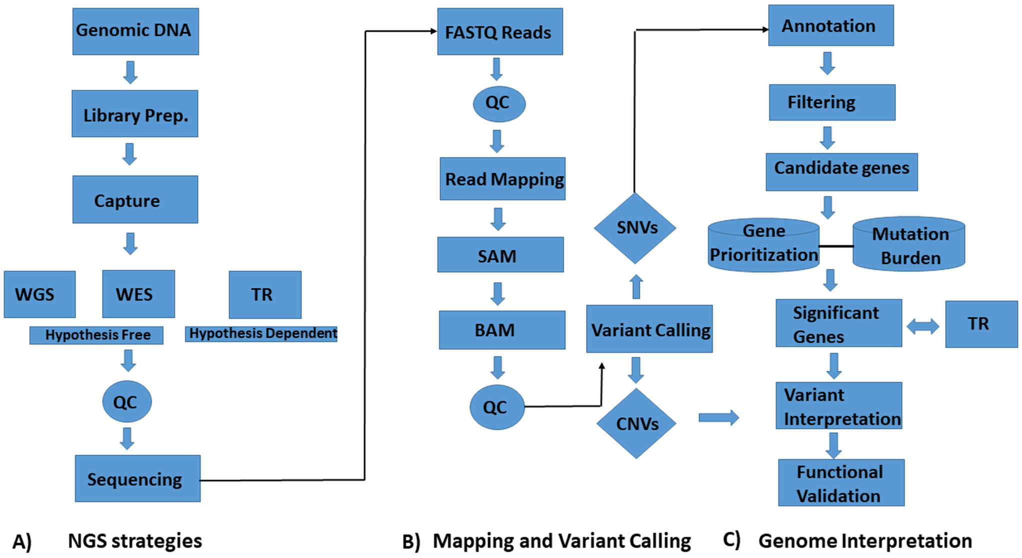


Figure 1.3.1: **NGS workflow and data analysis steps.** The three stages of the NGS workflow include (A) NGS strategies such as incorporation of WGS, WES or TR depending upon the research question being addressed. (B) Variant calling using standard bioinformatics tools and (C) Interpreting the genome through annotation and prioritization of genes/variants.

Especially when the underlying research question is hypothesis free, whole genome sequencing (WGS) or whole exome sequencing (WES) are ideal methods. Entire information on the genome that includes protein-coding and non-coding regions can be obtained using whole genome sequencing (WGS) approaches. Whereas through WES, only information about protein coding regions (with inclusion of 10-20 bp of intronic region) can be obtained. Due to cost effectiveness and scalability to sequence larger cohorts WES is often preferred over WGS. Subsequently targeted resequencing (TR) based strategy can be used where panels of known genes are sequenced making it a completely hypothesis driven approach.

In order to get from the high throughput sequence data to variant discovery, automated bioinformatics pipelines are required. The analysis steps include tasks such as variant calling, annotation, filtering strategies and gene prioritizations, which helps in narrowing down the search for candidate genes. The identified variants are then interpreted for their pathogenicity and subjected towards in vivo functional validation. In the next section we introduce and describe the analysis steps incorporated in a classical WES pipeline. In sections 1.4.2 and 1.4.3 we present the conceptual explanation behind the data analysis steps for identification of SNVs such as gene prioritization and mutation burden analysis. In addition, we introduce various analysis strategies to identify copy number variation (CNVs) specifically from TR data in section 1.4.4.

1.4 Methods for NGS Data analysis

1.4.1 WES Analysis pipeline

The enormous high-throughput data generated by NGS platform has led to a surge in the development of bioinformatics tools and analysis pipelines. The standard WES pipeline involves three important phases namely, preprocessing the raw sequencing data, genome alignment and variant discovery, annotation and prioritization.

(a) Quality control and preprocessing: The raw fastq reads obtained from the sequencing machine like Illumina is subjected to quality check for quality of the sequences. Low quality reads, PCR primers, adaptors or duplicates can affect the downstream analysis. Hence, in order to estimate the effect of these biases a fastQC report can aid as diagnostic tool that helps in understanding the underlying properties of the raw data such as base quality scores, GC content, sequence length distribution, sequence duplication levels. Based on these reports subsequent steps such as adapter trimming is applied using widely used existing methods such as Trimmomatic[29], cutadapt[141] etc. After these preprocessing steps the raw data is ready for the alignment. It is important to note that based on the choice of enrichment protocols and the NGS platform the quality check and preprocessing can vary.

(b) Genome alignment and variant calling: The processed raw fastq reads are next aligned with wild-type reference genome depending upon the build version hg18 from 2006, hg19/GRCh37 from 2009, hg 38/GRCh38 from 2013. Using burrow wheeler aligner[128] (BWA) or bowtie2[120] which utilizes Burrows-Wheeler transform to index the reference genome and has the advantage of being memory efficient and fast. The short reads are then aligned to the indexed reference genome. Internally, for bowtie2 the scoring scheme is similar to that of Needleman-Wunsch[157] and Smith-Waterman[196] for global and local sequence alignment respectively. Alternatively, other tools like NUCmer[54], BLAT[101] or BLAST[7] can also be used to align the short reads to the reference genome but are relatively slow as compared to bowtie2 or BWA. The alignment results are output in a sequence alignment map (SAM) format which is then sorted and merged to get a compressed binary aligned map (BAM) format file. Next, the existing WES pipeline available from best practices of the genome analysis tool kit (GATK)[6] is incorporated for variant calling. For germline variants it incorporates HaplotypeCaller (HC)[172] based method and performs joint genotyping on the cohort. The joint genotyping step is an important feature of HC where the genotyping is done for all the samples together which enables better statistical confidence score of the genotypes. When the cohort is larger, the confidence for calling the variants is better but it is also adds to the computational complexity. For detecting somatic variants it incorporates MuTect2[48]. The key feature of HaplotypeCaller is that it calls SNP and indel variants simultaneously. It has better sensitivity in detecting indels[87] in comparison to the previous version of GATK variant caller called UnifiedGenotyper (UG)[56] which is currently made obsolete. The list of variants emitted from this pipeline is stored in a variant calling format (VCF4.0) file. This is a tab separated file which holds information of all the variants with respect to their coordinates, genotype information, and variant quality metrics. Next this file is annotated for subsequent variant filtering.

(c) Annotation and Filtering: An average WES pipeline results in 50,000$\sim$60,000 variants per exome. Analyzing such a large list is time consuming, cost ineffective and a computational burden for any downstream analysis. Hence, it is customary to incorporate some filtering criteria to narrow down the search of causal variants that can be functionally validated. Hence, the filtering process requires annotation of the variants either using ANNOVAR[220] or variant effect predictor (VEP)[146] annotation provided by the Ensembl database. ANNOVAR provides gene-based annotation (RefSeq symbols) for identifying protein coding variants. It also provides region-based annotation that helps in identifying variants in the given genomic regions. Finally, filter-based annotation can help in distinguishing rare and frequent variants from the list by comparing the allele frequency in existing databases such as dbSNP[192], Exome Aggregation Consortium[62] (ExAC), or 1000 Genomes Project[1]. Additionally, the functional effect of variants (eg. being benign or pathogenic or unknown) can also be annotated using scores obtained from the prediction tools like SIFT[158], PolyPhen[3], LRT[47], MutationTaster[187] etc. Together with these functionalities the variants are annotated, and then subsequent filters can be applied based on the underlying research question. Based on sequence quality annotation the variants are filtered with respect to allelic depth, allelic balance (preferably above 25% for heterozygous variant and above 80% for homozygous variants).

Even after a traditional WES pipeline, the list of variants still contains dozens of variants. To further narrow down the list of candidate genes, complementary strategies are being developed. In this thesis, we implemented gene prioritization tool and mutation burden analysis. We introduce the conceptual ideas behind the implementation of these strategies in sections 1.4.2 and 1.4.3 respectively. The actual results obtained from the application of gene prioritization and burden analysis is presented in **chapter 4**.

1.4.2 Gene Prioritization

Based on a stringent choice of WES filters, the list of variants can get significantly reduced to approximately 200$\sim$300 variants per sample. However, this new list is still very large in order to carry out any functional validation. Hence, to narrow down the search for the potential gene which causes the phenotype it is important to know functional information about the gene. These functional data are spread across several manually curated online databases containing high quality information. The functional information about the candidate genes can be used to manually rank (involving multiple experts) them according to their priority for subsequent analysis. The ranking of candidate genes based on functional information criterion is called gene prioritization as shown in Figure 1.4.1. This information is retrieved from various different annotation sources that can be broadly categorized under literature, expression databases, gene ontology, pathway databases, protein-protein interaction databases, phenotype or disease ontologies, and protein sequence similarity information.

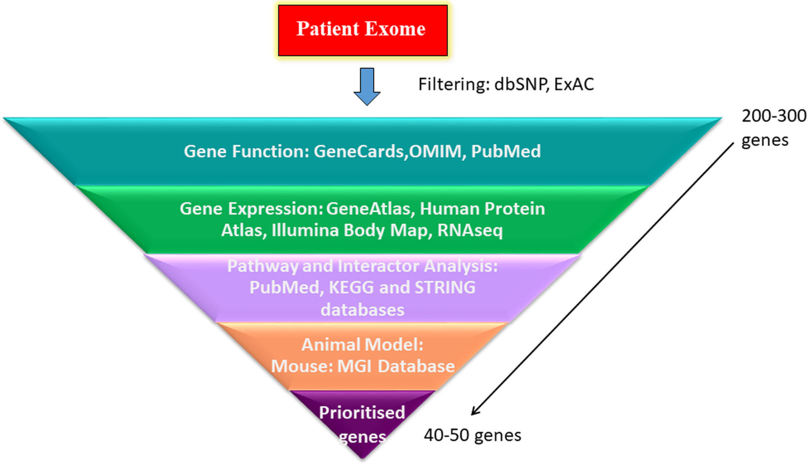

Patient Exome
Filtering: dbSNP, ExAC
200-300 genes
Gene Function: GeneCards,OMIM, PubMed
Gene Expression: GeneAtlas, Human Protein Atlas, Illumina Body Map, RNAseq
Pathway and Interactor Analysis: PubMed, KEGG and STRING databases
Animal Model: Mouse: MGI Database
Prioritised genes
40-50 genes

Figure 1.4.1: **Prioritizing candidate genes.** The patients' exome processed using WES pipeline results in a large list of variants. For detecting rare causal variants, the common variants are filtered against dbSNP and/or ExAC with certain cut-off of minor allele frequency (MAF). Post filtering the list of remaining variants are filtered through knowledge based analysis. Figure adapted from Nikhita *et.al*[28]

Multi-expert level ranking of this list of genes (∼200 to 300) per patient for a large cohort or large family is very time-consuming and induces individual biases that can subsequently affect the downstream analysis result. Hence, computational algorithms based on machine learning principles can help in automating the overall process. State-of-the-art methods such as Endeavour[4, 208] and ToppGene[45] suite are few of the primary tools developed for addressing the gene prioritization problem. These methods were first to demonstrate that integration of annotation sources under machine learning principles can effectively solve the gene prioritization problem. In subsequent years the advancement in the machine learning research has led to a surge of many sophisticated algorithms having superior performance on several existing benchmark dataset. A detailed review on gene prioritization algorithms based on their internal design and different types of data they integrate was previously published[151, 209]. The majority of the available tools incorporate the guilt-by-association (GBA) principle for prioritization, which states that the genes that share similar functional patterns (such as similar interaction partners, enrichment in same pathways) are likely to be involved in similar diseases. The overall principle of these tools, as shown in Figure 1.4.2, is to first integrate different annotation sources by enabling learning of functional aspects associated with the gene. In a second stage, using a set of training or so called seed genes (derived from GBA principle), a discriminatory model can be trained on the integrated annotation sources. The trained model could be a regressor or a classifier which, when applied on a set of candidate genes, eventually leads to their prioritization. Although the prioritization principle is apt, straightforward and similar to any existing classification or recommender algorithm applied in a wide variety of domains such has object detection, IMDB movie recommender system, internet search engines etc., it is very sensitive towards the underlying data upon which it has been designed (in this case the function of genes). Since for some genes little information about their function is available, it is customary to integrate information from different sources effectively for prioritization.

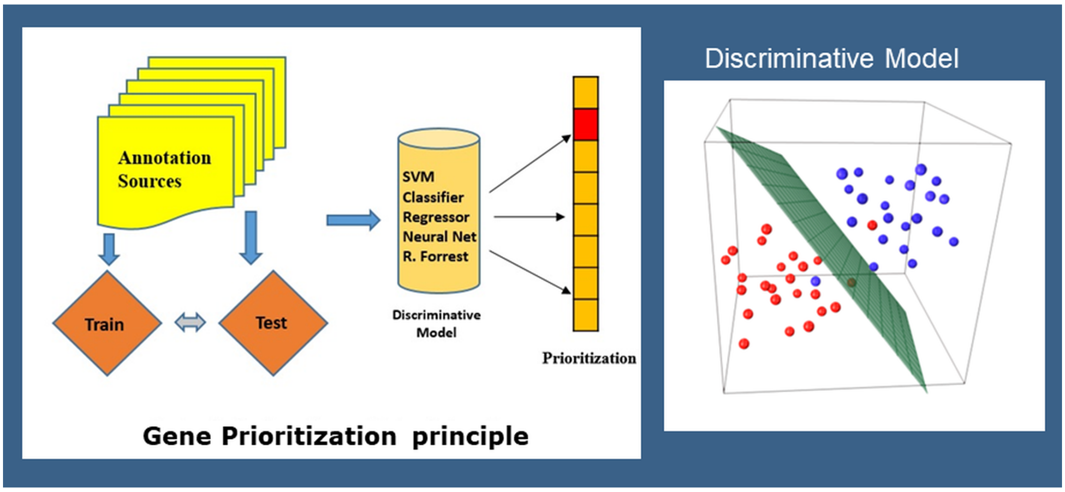


Figure 1.4.2: **Computational gene prioritization principles.** Computational gene prioritization approach requires integration of annotation sources. Based on the choice of training genes the information is extracted from the integrated annotation sources and a regressor or a classifier (discriminative model as shown in right) is trained. The test genes are then prioritized based on similarity to the training genes by applying the learned classifier.

Eventually, this led us to develop a novel prioritization tool named pBRIT, which is an acronym for prioritization using Bayesian Ridge Regression and Information-Theoretic approach. It incorporates an intermediate integration strategy based approach to fuse 10 annotation sources categorized under phenotype annotation and functional annotation. It implements the existing biological hypothesis that genes that are functionally related could also be phenotypically correlated. Utilizing the functional and phenotypic correlation, a linear regression model in bayesian framework enables to learn the underlying linear mapping with respect to a given set of training genes. Internal design of pBRIT handles feature dependencies and sparsity present in the annotation sources. Moreover, it is fast and scalable, allowing the prioritization of a large set of candidate genes. The details of this method and corresponding results are presented in **chapter 3** of this thesis.

Table 1.2: **List of freely available gene prioritization tools**

Prioritization Tool	**Method Description**	**Annotation sources**	**Integration strategy**	**Availability & Accessibility**
Endeavour (2006)[4]	Rank aggregation from individual annotation sources using order statistics to yield global prioritization ranks.	20 different annotation sources categorized under Text, PPI, Pathways, Expression, Sequence specific scores from Prospectr & Ouzounis, Interpro, Swissprot, EnsemblEST, Sequence similarities, Motif	Late Integration	Deprecated
Endeavour (2016)[208]	Kernel methods utilizing one class support vector machine (SVM) to perform prioritization	42 different annotation sources which includes above 20 annotation sources. Newly added annotation sources include phenotype, expression, pathways, chemical information.	Late integration	`https://endeavour.esat.kuleuven.be/Endeavour.aspx`
ToppGene-suite[45]	Portal for: (1) Gene list functional enrichment. (2) Candidate gene prioritization using functional annotation or network analysis. (3)I dentification and prioritization of novel disease candidate genes in the interactome. Utilizes fuzzy-similarity for semantic annotations, extended versions of PageRank and HITS algorithm and K-step Markov method to prioritize candidate genes	14 annotation categories which includes Gene Ontology, Human Phenotype, Mouse Phenotype, Protein domains, Pathway, Pubmed, PPI, Cytoband, Transcription Factor Binding Site, Gene Family, Co-Expression, Co-Expression Atlas, Computational prediction scores from MSigDb, MicroRNA, Drug, Disease.	Late integration	`https://toppgene.cchmc.org/prioritization.jsp`
pBRIT[112]	Implements Information-Theoretic and Bayesian Ridge regression approach to prioritize candidate genes	Integrates 10 different annotation sources that are categorized under phenotypic and functional annotation sources.	Intermediate Integration	`http://biomina.be/pBRIT`
Collage[231]	It utilizes collective matrix factorization to compress data and chaining to relate dfferent data object types in the annotation sources	14 annotation sources under 10 categories which includes 3 whole genome RNA-seq experiments, PPI from STRING, gene mention from research articles, MeSH headings, Pathway-KEGG, Reactome, HPO, GO	Intermediate integration	`ps://github.com/marinkaz/collage-dicty`
HyDRA[103]	Ensemble of rank-aggregation methods which includes Lovsz-Bregman, Hybrid Borda, Hybrid Kendall. Hybrid Kendall is the preferred method suggested by the authors	Acts on ranking on Endeavour and ToppGene which are in fact generated by all of their internal sources.	Late Integration	NA

…continued

Prioritization Tool	Method Description	Annotation sources	Integration strategy	Availability & Accessibility
F_3PC[43]	Logistic regression that that predicts the phenotypic label of candidate genes from the integrated multiple network.	Incorporate OMIM, PPI-HPRD, BioGrid, IntAct, Pathway-KEGG, Reactome, PharmaGKB and Expression-BioGPS to create gene networks.	Early integration.	Available as standalone MATLAB implementation. Requires customization for large scale exome prioritization.
MRF[44]	Applies Markov Random Field model on integrated network	Incorporate OMIM, PPI-HPRD, BioGrid, IntAct, Pathway-KEGG, Reactome, PharmaGKB and Expression-BioGPS to create gene networks	Early integration	Available as standalone MATLAB implementation. Requires customization for large scale exome prioritization.
DIR[46]	Provides a unified graphical representation to integrate heterogeneous networks using diffusion kernel measure. The integration strategy utilizes most informative evidence among the set of data sources for prioritization	Incorporate OMIM, PPI-HPRD, BioGrid, IntAct, Pathway-KEGG, Reactome, PharmaGKB and Expression-BioGPS to create gene networks	Early integration	Available as standalone MATLAB implementation. Requires customization for large scale exome prioritization.
RWR[107]	Implements an iterative randowm walker on graphs constructed upon only PPI networks using a diffusion kernel. The network nodes are ranked according to the proximity to known disease-assoicated genes or seed genes.	Incorporate individual networks of PPI component of STRING and text mining component of STRING database.	Early integration	Available as standalone MATLAB implementation. Requires customization for large scale exome prioritization.
RWR-M[211]	Implements an iterative randowm walker on graphs constructed upon multiplex networks using a diffusion kernel. The network nodes are ranked according to the proximity to known disease-assoicated genes or seed genes.	Incorporates PPI, Disease-Disease similarity network obtained from HPO, Pathways, Co-expression.	Early integration	Available as standalone R implementation. Requires customization for large scale exome prioritization.

1.4.3 Mutation burden analysis

Association of a genetic variant with a particular disease can be determined by computing the increase or decrease in frequency of its specific alleles in affected individuals compared to control individuals. Association studies of common variants (having MAF $\geq 5\%$) are often referred to as genome wide association studies (GWAS) and studies involving a set of rare variants (MAF $\leq 1\%$) in coding regions through WES technologies are referred to as rare variant association studies (RVAS). There are many more rare compared to common variants in the genome and their contribution to disease should not be ignored. NGS allows enumeration of both these types of variants directly for each sample for the association studies. But, the single marker based test used for common variants association analysis are unsuitable for RVAS due to reduced power directly related to the low allele frequency.

Therefore the development of a conceptual framework of statistical methods for RVAS relies on taking into account of following: (a) **type of variants:** assessing the effect of variants: disruptive/damaging alleles (stop, frameshift, splice-site mutations), or protective alleles in the form of missense mutations or benign alleles with no risk, (b) **frequency threshold:** choosing the optimal threshold of the frequency to maximize the power of association, (c) **sample size:** determining total number of cases to detect significant associations, (d) **whole genome analysis:** extending the current exome based RVAS analysis to incorporate non coding variants as well and finally (e) **other strategies:** studying isolated populations, specific gene sets or de novo mutations.

For establishing the significance of the associations many statistical methods have been developed and can be broadly classified into two types namely, BURDEN tests[127, 153, 175] and non-burden tests[124] such as C-alpha[156] or the sequence kernel association test (SKAT)[222]. The BURDEN test, also known as mutation burden test, involves the strategy of collapsing a group of rare variants in a gene or a pathway and statistically comparing the difference between cohort of affected (cases) and non-affected individuals (matched controls) and test for significance. The standard statistical test could either be a general Fishers Exact test for low sample size or Chi-square test when number of cases and controls are relatively much higher (average cell size have a value ≥ 5) also known as the cohort allelic sums test (CAST)[152]. The combined multivariate and collapsing (CMC)[127] method extends the basic principle of CAST by collapsing the rare variants within a gene and the information of both collapsed rare and common variants is used in the association test. Similarly, incorporating the information of allele frequencies, both the rare and common variants can be weighted (based on their inverse allele frequencies) for the association test. This is known as the weighted sum (WS) method[139]. Other methods such as the variable minor allele frequency threshold (VT) method[175] and the cumulative minor-allele test (CMAT)[226] utilize the MAF threshold and aggregation of minor alleles across all the sites for cases and controls respectively for the association test.

The basic assumption of the collapsing based methods is the fact that rare variants are either deleterious or protective to some diseases, meaning that effect of all rare alleles act in one direction. However, when both risk and protective variants are present, then above mentioned methods are underpowered because the opposite effects will counteract each other[190]. Additionally, the effect of neutral alleles is generally ignored by these methods but their presence can dilute the estimating power by adding noise in the underlying statistical models.

Alternative to collapsing based methods, non-burden test such as SKAT give superior performance as they include the directional effects on association arising due to presence of both risk (disruptive and protective) and neutral variants in their underlying model. SKAT incorporates a multiple regression framework where the phenotype is directly regressed upon the genetic variants in a region and on covariates, and thus allows the different variants to have different direction and magnitude of effects. If above mentioned assumptions of the collapsing burden test are met, then SKAT based tests are less powerful.

Each of the above gene-based collapsing methods reports a p-value of the association of multiple variants to the disease but detection of a small group of true causal variants still remains a challenging task. In the current thesis as described in chapter 4 collapsing variants based methods have been applied to determine the causal deleterious variants in BAV/TAA disease. Further we demonstrate how the two step complementary strategy using gene prioritization and mutation burden analysis could aid in identifying these causal variants.

1.4.4 Computational detection of CNVs

Copy number variations (CNVs) are genetic variants in which a section of the genome is altered either via duplication or deletion leading to increase or decrease in number of inherited copies. The size of these CNVs typically ranges from 50 bp to several kilo-base pairs. They can be spanning across one or several genes, resulting in several functional consequences like change in gene expression due to alteration of gene dosage and fusion and disruption of genes.

Array-based comparative genomic hybridization (arrayCGH), fluorescence in situ hybridization (FISH) or SNP arrays have traditionally been used to detect somatic[218] and germline CNVs[191]. However, they are suitable mostly for detecting large chromosomal aberration events (FISH: $>$100 kb; arrayCGH: $>$10Kb; SNP arrays: $\sim$10Kb) and suffer from poor sensitivity in detecting small CNVs (single exon deletion/amplification events). The size and breakpoint resolution of CNVs are correlated to probe density of SNP arrays at given loci[182].

The advent of NGS approaches, especially WGS, has promised to detect CNVs with far greater resolution in comparison to the contemporary methodologies. Although, WES and WGS are excellent methods providing comprehensive analysis to detect CNVs, in a clinical setting TR is often used. In TR, only known candidate genes for the disease, are sequenced to overcome constraints of cost per patient, to time bound results and to reach high depth of coverage (DOC). CNV analysis from TR data currently still poses an important challenge.

There are many new methods being developed and applied on a varied range of NGS datasets to identify CNVs[229]. These methods can be categorized into five different strategies: (a) paired-end mapping (PEM) (b) split-read (SR) based approach (c) read depth-based (RD) approach (d) de novo assembly and (e) combinations of any of the above approaches.

Paired-end mapping (PEM): In the paired-end sequencing the DNA fragments are sequenced from both the ends of the molecule thereby generating a pair of reads. These pairs of reads are jointly mapped to the reference genome. The CNVs could be identified by evaluating if the distance or orientation between the paired reads are significantly different from the predetermined insert size. The expected insert size is an intrinsic property of the sequencing library. If the distance between aligned paired reads is significantly larger than the expected insert size this indicates potential deletions. Conversely, if the aligned pairs are significantly closer then this is indicative of putative duplications. An advantage of using PEM based approach is that it can identify not only insertions or deletions but also inversions and tandem duplications. An example signature of paired-end approach is shown in Figure 1.4.3A.

Split read-based (SR) approach: Split read based methods are capable of CNV break-point resolution with high precision. One of the reads from the read pairs is uniquely mapped to the reference genome and another one either fails to map or gets partially mapped. The incompletely mapped reads are split into multiple fragments. Next, among these fragmented reads the first and the last fragments are aligned to the reference genome independently. This remapping thus provides the precise start and end point of insertion or deletion. The signature of SR based method for identifying deletion or duplication is shown in Figure 1.4.3B.

Read depth (RD) approach: Read depth-based methods essentially compute the number of reads aligned to a particular genomic position. Theoretically, the RD is proportional to the underlying copy number, which in fact is directly correlated to depth

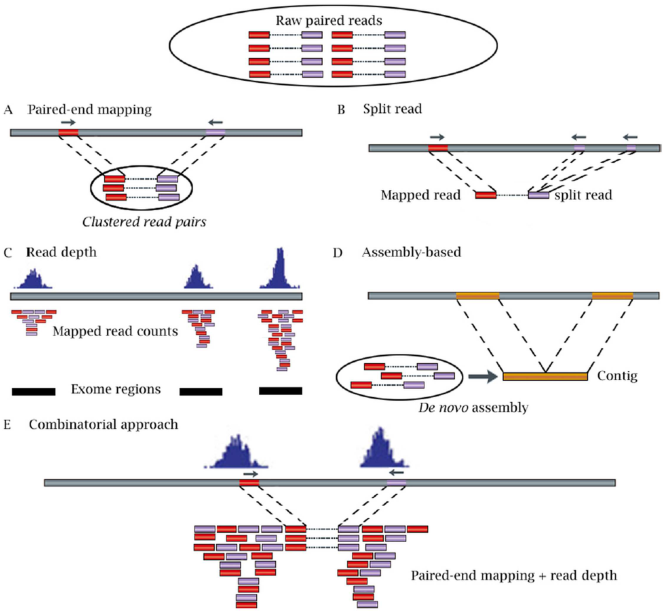


Figure 1.4.3: **Categorization of CNV calling algorithms.** (A) Denotes the signature of PEM strategy for CNV calling. (B) Denotes the signature of SR based strategy where incompletely mapped reads are used to identify the break points of CNV segment. (C) Denotes the signature of RD based strategy where counting the paired reads mapped to the genomic regions are used to detect CNVs. (D) Denotes the signature of de novo assembly based approach where contigs are mapped to the reference genome to detect CNVs. (E) Signature of a combinatorial approach that combines RD and PEM information to detect CNVs. The above figure is adapted from the article by Zhao *et. al.*[229]

of coverage (DOC). The DOC is calculated based on the Lande-Waterman equation[117]:

$$\textbf{DOC} = \frac{\text{Number of mapped reads} \times \text{average read length}}{\text{Length of reference sequence}} \tag{1.2}$$

In comparison to PEM/SR based methods the RD based approaches can detect exact copy number events. Additionally, detection of large insertions or deletions in complex genome regions is better in comparison to PEM/SR based approaches. For reliable detection of CNVs this approach requires substantially high DOC (normally 100x). The signature of RD-based approaches for detecting CNVs is shown in Figure 1.4.3C.

De Novo assembly (AS): Apart from PEM, SR and RD based approach which rely on alignment of reads to the reference genome to detect CNVs, another class of methods is based on de novo assembly of these reads without using the reference genome. These methods first reconstruct assembly of contigs from overlapping reads. The assembled contigs are then compared to the reference genome and thus the genomic regions with discordant copy numbers are identified. Figure 1.4.3D presents an example signature of AS based approach.

Combination of above methods: Each of the above four methods have distinctive advantages and merits but their predictive performance varies from case to case. Hence, for predicting full spectrum of CNVs it is good practice to combine either of these methods for better predictive performance. For example, the PEM based approach is good at predicting insertions, deletions, translocations, inversions, interspersed and tandem duplication, but has poor resolution in detecting average number of copies of CNV segments. Incorporation of the RD based approach, which can accurately predict the number of copies can complement the PEM in detecting CNVs. An example of combinatorial based approach is presented in Figure 1.4.3E.

Among these strategies only RD based approaches can be successfully applied to WES or TR in comparison to WGS data[202]. Computational detection of CNVs requires adequate addressing and handling of inherent biases arising from sequencing platforms. The normalization procedure includes accounting for biases associated to the enrichment protocol being used, non-uniform depth of coverage variability, sequence properties of regions of interest (ROIs) such as GC content and presence of repetitive elements. Overall the goal is to normalize the input data such that variability of RD is minimized, followed by detection of CNVs by comparing the RD of the patient samples to matched control samples processed in a similar way. This is accomplished by computing the logarithmic-ratio score, which represents deviation of the patient's copy number state from normal. Positive log-ratios indicate a region of DNA copy number gain and a negative log-ratio indicates DNA copy number loss.

In this thesis a novel tool named **varAmpliCNV** is introduced, which predicts potential CNVs from TR data. It has been designed specifically for detecting CNVs from amplicon-based TR data (Haloplex^{TM} based enrichment protocol) and for the special case where matched controls are not available. The details of the method and analysis results are discussed in **chapter 5**.

Aims and objectives

Chapter 2

Aims and objectives

Identification and interpretation of genetic variants (single nucleotide variation and structural variation) associated to disease is challenging task. Application of NGS technologies has provided the unique opportunity to simultaneously screen and analyze large number of these variants to uncover underlying genetic and molecular mechanisms pertaining to pathogenesis of disease. The high throughput data produced by NGS workflows require development of computational tools that can automate the detection and interpretation of the functionality of the variants. Thus, the main objectives of current thesis work are:

1. To develop computational framework for the identification and prioritization of single nucleotide variants (SNVs) from whole exome sequencing (WES) data. The results are presented in **chapter 3**.

2. To perform mutational burden analysis and to apply the computational framework from aim (1) to WES and resequencing data obtained in a large cohort of BAV/TAA patients. The analysis is explained in **chapter 4**.

3. To develop a statistical model to predict copy number variations (CNVs) from targeted resequencing NGS data thereby addressing the challenges arising due to inherent biases associated with specific enrichment technologies. The details are provided in **chapter 5**.

Gene prioritization and Mutation burden analysis

Chapter 3

pBRIT: Gene prioritization by correlating phenotypic and functional annotations

To build a truly intelligent machine, teach them cause and effect

Judea Pearl (2018)

Ajay Anand Kumar[1,2], Lut Van Laer[1], Maaike Alaerts[1], Amin Ardeshirdavani[3,4], Yves Moreau[3,4], Kris Laukens[3], Bart Loeys[1] and Geert Vandeweyer[1,2]

Bioinformatics 1 (2018): 9

[1]Center of Medical Genetics, University of Antwerp & Antwerp University Hospital, Antwerp, Belgium. [2]Biomedical Informatics research network Antwerp (biomina), University of Antwerp, Antwerp, Belgium. [3]Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium. [4]imec, Leuven, Belgium. [5] ADReM Data Lab, University of Antwerp, Antwerp, Belgium.

- This chapter was adapted from the article published online at **Bioinformatics** 2018 Feb 14; 1:9, `https://doi.org/10.1093/bioinformatics/bty079`
- Supplementary File mentioned mentioned in this chapter can be accessed online: `https://tinyurl.com/y7s3ljct`

3.1 Abstract

Motivation: Multiple computational gene prioritizers fusing various genomic annotation sources exist to aid in disease gene identification. Here, we propose pBRIT (prioritization using Bayesian Ridge regression and Information Theoretic model), a novel adaptive and scalable prioritization tool, integrating Pubmed abstracts, Gene Ontology, Sequence similarities, Mammalian and Human Phenotype Ontologies, Pathway, Interactions, Disease Ontology, Gene Association database and Human Genome Epidemiology database, into the prediction model. We explore and address effects of sparsity and inter-feature dependencies within annotation sources, and the impact of annotations changing over time on rank stability.
Results: pBRIT models annotation feature dependencies and sparsity by an Information-Theoretic (data driven) approach, allowing effective feature mining and intermediate integration based data fusion. Following the hypothesis that genes underlying similar diseases will share similar functional and phenotype characteristics, it incorporates Bayesian Ridge regression to learn a linear mapping between functional and phenotype annotations. Genes are prioritized on phenotypic concordance to the training genes. We evaluated pBRIT against 7 existing methods, and on over 2,000 HPO-gene associations retrieved after construction of pBRIT data sources. We achieved maximum AUC scores ranging from 0.92 to 0.96 against benchmark datasets and of 0.80 against the time-stamped HPO entries, indicating good performance with high sensitivity and specificity. Our model shows stable performance with regard to changes in the underlying annotation data, is fast is scalable for implementation in routine pipelines
Availability: pBRIT is freely available at `http://biomina.be/apps/pbrit/`

3.2 Introduction

Whole-exome sequencing (WES) is the current standard approach to identify causal variants in genes underlying human genetic disorders, but returns a large number of variants. Databases of known variants such as ExAC [62] provide a powerful first filter. However, finding the true causal variant often remains a time consuming and challenging task. For small sample sizes, it often involves manual evaluation of functional and phenotypical aspects of genes using literature and curated biological databases, but for large datasets computational tools are necessary.

The core principle of computational gene prioritization is to rank candidate genes based on annotation patterns using some discriminatory statistical model. Additionally, these methods can help in generating interesting non-trivial hypotheses for novel gene functions. The predictive ability of these tools heavily depends on both the choice of annotation sources and the technique used to mine the patterns.

Tranchvent *et.al.*[209] presented an overview of existing gene prioritizers classified with respect to integrated annotation sources. Based on the presence or absence of a training set [151], these tools are broadly classified as supervised (e.g Endeavour [4, 208]; ToppGene [45]) or unsupervised models (e.g Biograph [131]).

Next to the learning approach, prioritization results depend on two other aspects: annotation sources can be integrated using early, intermediate and late integration [166], and a wide range of statistical methods can be used as the underlying model to rank the genes. Network-based prioritization tools [107, 115, 130, 223, 228], incorporating both protein-protein interaction and phenome networks, are examples of early integration based approaches. Among these, Random Walk with Restart (RWR) gives robust performance with higher predictive accuracy, but it is typically only applicable to single networks and often incorporates only direct neighbourhood information. For multiple networks, Direct Integration of Ranks (DIR)[46] and Markov Random Field (MRF)[44] were proposed which automatically assign weights to different networks for integration. Recently, a new version the RWR algorithm was proposed that also incorporates multiple heterogeneous networks (RWR-M)[211]. Chen *et.al.*[43] proposed a logistic regression based model that utilizes direct and higher-order neighbourhood information in the network for prioritization, together with pathway and expression profiles.

Early integration based approaches can represent topological relationship of entities, but often require complex feature construction during data fusion. In contrast, late integration approaches compute ranks on individual annotation sources and then integrate them to obtain an overall ranking. Rank fusion can become computationally challenging when the number of annotation sources and genes to be prioritized is large. Recently, Zitnik *et. al*[231] proposed a midway approach, termed intermediate data integration. The main idea is to fuse annotation sources while retaining the overall data structure, thereby capturing internal structures and latent dependencies. Despite the broad range of available methods, most current implementations ignore these internal structural representations (like hierarchical ontologies) and latent inter-feature dependencies during fusion.

It should be noted that updates to annotation sources can eventually alter biological meanings associated with the functionality of any gene. Furthermore, Schnoes *et. al*[185] pointed out that the advent of next generation sequencing created a large gap between computationally predicted annotations and their experimental validation. For example, three studies [72, 74, 111] discussed how changes in the internal directed acyclic graph structure of Gene Ontology (GO) terms over an interval of ten years can impact subsequent functional analyses. The dynamic nature of biological annotation sources will thus inevitably lead to annotation errors, with a significant potential impact on downstream analyis [217]. Although gene-by-gene proximity profiles are at the core of all available prioritization tools, the uncertainty on the proximity scores related to these changes is typically not taken into account, which might impact the prioritization results and lead to less stable ranking.

Another important aspect that should be addressed is the issue of annotation sparsity. Annotation features describing gene functionalites are typically sparsely distributed when considering genome wide data, making feature mining computationally intensive. Moreover, current regression based methods [223, 228] assume there is no multi-collinear effect of the independent variables (training genes) in the analysis. When multi-collinearity is present however, this might lead to inflated values for the regression coefficient estimates, which might in turn lead to over-fitting.

In order to address the above issues, we propose a new computational gene prioritization tool named pBRIT, which applies an Information-Theoretic approach for effective feature mining and Bayesian Ridge Regression (BRR), leading to an intermediate data integration based prioritization model. In this work we explore the efficiency of text mining methods such as TF-IDF (Term Frequency-Inverse Document Frequency) and latent semantic models (LSM) in gene prioritization. We apply TF-IDF for feature extraction and LSM to address sparsity and feature dependencies. Different aspects of pBRIT were evaluated on two separate tasks. First, we compared pBRIT performance with 7 existing methods on their original benchmark datasets. Second, we approximated a prospective evaluation using time-stamped benchmark data derived from HPO and compared performance with 2 additional recent state-of-the-art methods (Endeavour-v3.71 and RWR-M). Finally, we demonstrate the applicability of pBRIT in result visualization and exploration. pBRIT is implemented on a high-performance computing platform, freely available at `http://biomina.be/apps/pbrit`.

3.3 Materials and Methods

pBRIT offers a three staged gene prioritization, as represented in Figure 3.2.1 Unsupervised feature mining, assigning statistical weights to features in the individual annotation sources, is followed by intermediate data fusion. A Bayesian ridge regression model is then built to prioritize candidate genes under a supervised approach. This

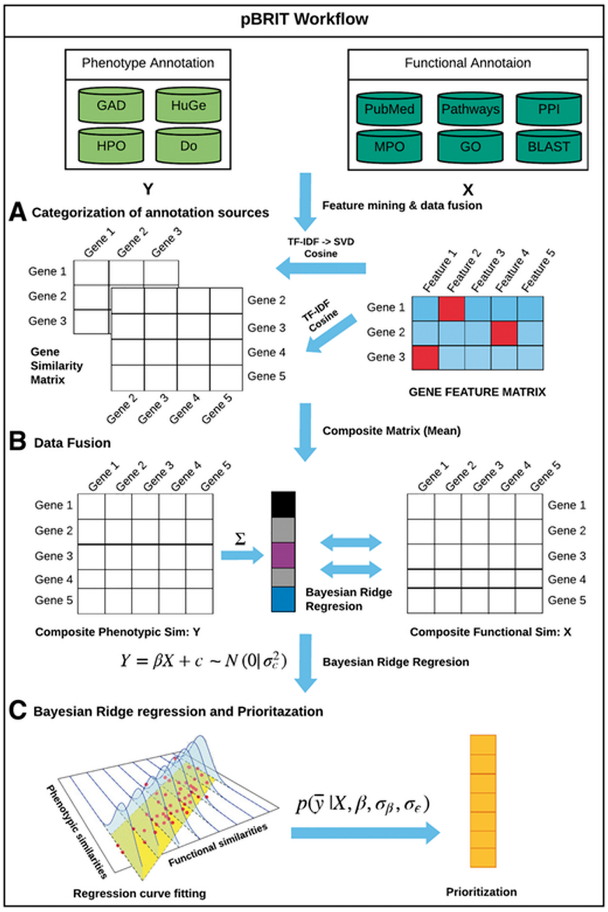


Figure 3.2.1: **Schematic workflow of pBRIT.** (A) Categorization of annotation sources as functional or phenotypic, (B) Gene-by-gene proximity profile computation using TF-IDF and TF-IDF→SVD, followed by intermediate data fusion, (C) Bayesian ridge regression based candidate gene prioritization

framework aids in modelling parameter uncertainties arising due to implicit annotation changes or errors.

3.3.1 Internal representation of annotation sources

We integrated 10 annotation sources, categorized as phenotypic or functional (Figure 3.2.1A Phenotypic annotations include human phenotype ontology (HPO), HuGe disease navigator (HuGe), the gene association database (GAD) and the disease ontology (DO). For functional annotations, we incorporated Pubmed abstracts, pathway databases, protein-protein interactions (PPI), protein sequence similarities (BLAST), mammalian phenotype ontology (MPO) and gene ontology (GO). All annotation sources were downloaded between January 6th, 2014 and January 26th, 2015. (See S1 and table S1.1)

Annotation sources were pre-processed using a generalized version of GOParGenPy

[111] to obtain sparse binary matrices with rows representing gene names (mapped to Ensembl ids) and columns representing specific annotation features (figure S1.3). Entries of 0 and 1 represent feature absence and presence respectively. For PubMed abstracts, the entries were generalized to the number of feature occurrences per abstract. One exception to the sparse representation was BLAST, for which normalized bit scores from pairwise sequence alignment of all human proteins were used as similarity scores. The matrix is treated as a full matrix (table S1.1).

3.3.2 Information-Theoretic model for feature mining

We computed TF-IDF based statistical weights for features in the sparse annotation matrices (equation 3.1). TF-IDF is based on the relevance and frequency of feature occurrences in the corpus. Features that are less frequent indirectly imply an annotation specific to a gene.

$$\begin{aligned}\mathbf{TF}(f,g) &= 1 + \mathbf{Log}\left(\mathbf{tf}_{feature,gene}\right)\\ \mathbf{IDF}(f,G) &= \mathbf{Log}\left(\frac{|G|}{1+|\{g\epsilon G : f\epsilon g\}|}\right)\\ \mathbf{W}(f,G) &= \mathbf{TF}\times\mathbf{IDF} \end{aligned} \tag{3.1}$$

For all sources except PubMed, the term frequency (tf) is equal to one due to the binary data format. IDF(f,G), or inverse document frequency, denotes the inverse frequency of a particular feature (f) across all genes (G). Hence, it describes the specificity of a feature. W(f, g) gives the statistical weight of feature (f) for a given gene (g). Using TF-IDF, specific features get higher weights, contributing more to the final similarity score used in ranking.

3.3.3 Modelling feature interdependencies and sparsity

Singular Value Decomposition (SVD) is a matrix factorization technique that reduces the sparsity and can model co-occurrences of the feature concepts [84]. Through SVD, high dimensional matrices are transformed to a lower dimension, where each original row and column can be represented as a linear combination of latent concepts in the new singular vector space. This linear combination of latent concepts indirectly models any co-occurring or semantically related features. The final number of vectors (k) defines both the complexity of the model and the accuracy of representing the original feature space.

Using SVD, each annotation matrix was decomposed in k singular values and then projected in those directions. The optimal choice of k corresponds to a maximal preservation of variance in the data. Mathematically, this can be expressed as:

$$A_{m\times n} \approx U_{m\times k}D_{k\times k}V_{k\times n}; \tilde{A}_{m\times k} \approx A_{m\times n}V^{T}_{k\times n} \tag{3.2}$$

Where, U is an $m \times k$ unitary matrix with k columns as left singular vectors. V is a $k \times n$ unitary matrix with k rows as right singular vectors. D is a $k \times k$ diagonal matrix holding k singular values.

Table S1.1 presents the average number of non-zero features per gene in each annotation source used in pBRIT, which ranges from 236 (Pubmed) to 10 (GAD). From Figure S1.2, it can be seen that a uniform proportion of variance is explained for all sources with k set to 200. Hence, we generalized the choice of k equal to 200 for all TF-IDF weighted matrices. Gene-by-gene proximity profiles were obtained using cosine similarity on both TF-IDF and SVD transformed TF-IDF matrices, represented throughout the text as TF-IDF and TF-IDF→SVD, respectively.

3.3.4 Data Fusion

In order to perform Bayesian ridge regression, we compute the composite matrices for the independent and dependent variables in the regression model by averaging the gene-by-gene proximity profiles :

$$X_{composite} = \frac{\sum_{f}^{F} X_f}{F}; Y_{composite} = \frac{\sum_{p}^{P} Y_p}{P} \tag{3.3}$$

where, F and P denote total number of functional and phenotypic annotation sources respectively. X_f and Y_p represent gene-by-gene proximity profiles for all f functional and p phenotypic annotations sources, following equations 3.1 and 3.2.

3.3.5 Prioritization using Bayesian ridge regression model

pBRIT implements the underlying hypothesis that the biological function of a gene is correlated to phenotypic characteristics presented by deregulation of that gene. Mathematically, this can be formulated by a regression between functional and phenotypic annotations. However, the parameters of such a regression are intrinsically affected by uncertainties in the model arising due to incomplete annotations and changes in the annotation corpus. Regression under a Bayesian framework can model these uncertainties while learning the linear mapping between functional and phenotypic annotation sources. Specifically, we want to model the mean of conditional $E(Y|X)$, i.e. the expected distribution of phenotype similarities given the functional annotation information. This is represented by $E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$. For any given n training genes and m test genes which needed to be prioritized, we extract the respective composite matrices for both functional and phenotypic annotations using equation 3.3.

The response, or dependent variable vector of the regression model is obtained by $Y_{(n+m)\times 1} = \sum_{j=1}^{n} y_{ij}$. The independent, or predictor variables are the gene-by-gene proximity profiles with respect to n training genes, forming the design matrix $X_{(n+m)\times n}$. The overall regression model is thus given by:

$$\mathbf{Y}_{(n+m)\times 1} = \beta \mathbf{X}_{(n+m)\times n} + \boldsymbol{\varepsilon}; \text{where, error term } \varepsilon \sim N(0, \sigma_\varepsilon^2) \tag{3.4}$$

The unknowns, the regression coefficient β, its corresponding variance σ_β^2 and the residual variance σ_ε^2 can be estimated uniquely from the above regression settings. The regression model of pBRIT uses proximity profiles of both training and test genes in the design matrix. The relatedness of the selected training genes gives a high likelihood of dependencies among the predictor variables. Sometimes, this leads to over-fitting and multi-collinearity of the regression model. Ultimately, multi-collinearity of the predictor variables can lead to inaccurate estimation of regression coefficients, inflated standard error estimates and degradation of model predictability. In order to overcome these problems, we propose a Bayesian ridge regression model. We regularize the estimates by adding a parameter $\tilde{\lambda}$ which is given by the ratio of $\frac{\sigma_\varepsilon^2}{\sigma_\beta^2}$. As the σ_β^2 increases to larger values the solution to find optimal $\hat{\beta}$ approximates ordinary least squares estimates. Requirements for the optimal choice of $\hat{\beta}$ are given by:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n+m} (y_i - x_i{}^T \boldsymbol{\beta})^2 + \tilde{\lambda} \sum_{j=1}^{n} \beta_j{}^2 \right\} \tag{3.5}$$

$$E(\boldsymbol{\beta} \mid y) = \hat{\boldsymbol{\beta}} = \left[\mathbf{X}^T \mathbf{X} + \tilde{\lambda} \mathbf{I} \right]^{-1} \mathbf{X}^T \mathbf{y} \tag{3.6}$$

In bayesian setting the likelihood of the model is given by:

$$\text{Likelihood :} \qquad p(\mathbf{y} \mid \beta, \sigma_\varepsilon{}^2) = \prod_{i=1}^{n+m} N \left[y_i \mid \sum_{j=1}^{n} x_{ij} \beta_j, \sigma_\varepsilon{}^2 \right] \tag{3.7}$$

$$\text{Prior :} \qquad p(\beta \mid \sigma_\beta{}^2) = \prod_{i=1}^{n} N(\beta_i \mid 0, \sigma_\beta{}^2) \tag{3.8}$$

$$p(\sigma_\beta^2) = \chi^{-2}(\sigma_\beta^2 | df_\beta, S_\beta) \tag{3.9}$$

$$p(\sigma_\varepsilon^2) = \chi^{-2}(\sigma_\varepsilon^2 | df_\varepsilon, S_\varepsilon) \tag{3.10}$$

We assume NIG (Normal Inverse-Gamma) density priors on unknown regression parameters. The joint posterior distribution of the vector of unknowns, represented by $\theta \epsilon (\beta, \sigma_\beta^2, \sigma_\varepsilon^2)$ in the model, is proportional to the product of the likelihood and the prior distribution, given by:

$$p(\theta|y) \propto \underbrace{\prod_{i=1}^{n+m} N(y_i|\sum_{j=1}^{n} x_{ij}\beta_j)}_{\text{Likelihood}} \times \underbrace{\prod_{i=1}^{n} N(\beta_i|0,\sigma_\beta^2)\chi^{-2}(\sigma_\beta^2|df_\beta,S_\beta)}_{\text{Prior on }\beta} \times \underbrace{\chi^{-2}(\sigma_\varepsilon^2|df_\varepsilon,S_\varepsilon)}_{\text{Prior on }\varepsilon} \tag{3.11}$$

Since the posterior distribution does not have a closed form, a Gibbs sampler was used. Regression analysis was performed using an adapted version of the BLR package [53] in R. Once the parameters are estimated, the corresponding phenotype concordance score y_{pred} can be predicted by:

$$E(X\boldsymbol{\beta} \mid y,\sigma_\varepsilon^2,\sigma_\beta^2) = \mathbf{X}E(\boldsymbol{\beta} \mid y,\sigma_\varepsilon^2,\sigma_\beta^2) \tag{3.12}$$

$$y_{pred} = E(X\boldsymbol{\beta} \mid y,\sigma_\varepsilon^2,\sigma_\beta^2) = \mathbf{X}\left[\mathbf{X}^T\mathbf{X} + \tilde{\lambda}\mathbf{I}\right]^{-1}\mathbf{X}^T\mathbf{y} \tag{3.13}$$

Prior to regression, the dependent variable Y and independent variable X were transformed by taking the square root of their values, in order to reduce any non-linearity effects. We follow the BLR guidelines for initializing the priors [53]. The prior on residual variance is indicated by two parameters: Scale, S_ε and degree of freedom, df_ε. The prior variance of the residuals is given by V_ε which is assigned as the variance of the phenotypic concordance score of the training genes. Together, they can be expressed as: $S_\varepsilon = V_{\varepsilon(\text{Train})}(\text{df}_\varepsilon + 2)$. Similarly, the prior on the regression coefficient can be expressed as: $S_\beta = \frac{\text{Var}(\text{Y}_{\text{Train}})\times(df_\beta+2)}{\sum_j^n \text{Var}(\text{X}_{\text{Train}j})}$. In this work, we chose $df_\varepsilon = df_\beta = 3$. For the Gibbs sampling we chose a total number of iterations of 100,000, a burn-in period of 30,000 and a thinning parameter of 10. The algorithmic details can be found in section S2.

3.3.6 Cross-validation strategy

The overall performance of pBRIT was evaluated by performing leave one-out cross-validation (LO-OCV) on several benchmark sets. For a given disease, with n known associated genes, we trained our model with n-1 genes and placed the query gene (known gene whose ranking is to be determined) in a list of 99 Test genes randomly selected across the genome. We removed direct contribution of known phenotypic associations of the query gene to the remaining training genes during validation experiments by setting all proximity scores to 'Na' (indicating phenotype information 'Non available'). Hence, the model purely predicts the phenotype concordance score of the query gene, without bias to prior knowledge (See section S2 for details).

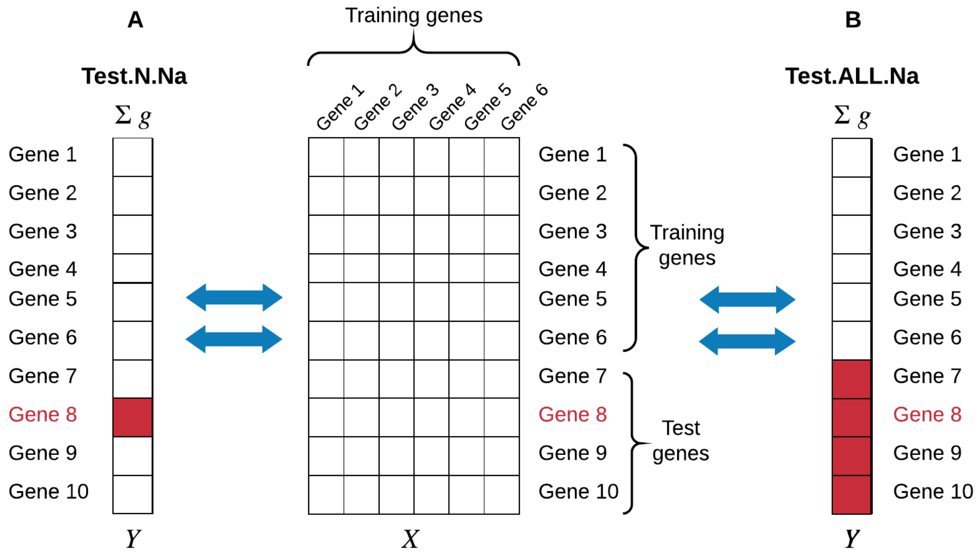


Figure 3.3.1: **Bayesian ridge regression.** The design matrix (X) contains similarity scores of both training and test genes to training genes. The phenotypic concordance score vector is indicated by Y. For LO-OCV, the summed phenotypic score of the n^{th} query gene (A. Test.N.Na) or all test genes (B. Test.ALL.NA), corresponding to prior phenotypic knowledge, is removed (red box) during regression parameter estimation.

We explored the effect of the regression model design on the prediction efficiency in two cases. In the Test.N.Na case (Figure 3.3.1A), the known phenotypic associations of all 99 test genes were taken into account in the regression model, discarding only the known associations of the n^{th} query gene. In contrast, in the Test.ALL.Na case (Figure 3.3.1B), the phenotypic association of all the test genes, along with the query gene, is discarded. Both Test.N.Na and Test.ALL.Na were then combined with either TF-IDF or TF-IDF→SVD based proximity profiles to evaluate the effect of the feature extraction methodology, leading to four analysis scenarios in total. The TF-IDF→SVD_Test.N.Na scenario, reflecting all pBRIT functionality, is referenced as the full pBRIT model hereafter. (See section S2 for algorithmic details).

LO-OCV analysis yields ranks of all the training genes per studied disease. Query gene ranks were normalized to rank-ratios by dividing them with the total number of test genes (typically n=100) and evaluated by two criteria. First, the mean rank ratio (MRR) of all training genes for a given disease was calculated. The MRR is computed by taking average of rank ratios per disease class and is a metric of efficiency, estimating how many candidates a user must review before the true positive candidate is encountered. Second, the Area Under the Curve (AUC), which measures the prediction accuracy of the model, was obtained from plotting the Receiver Operation Characteristic (ROC) curves. ROC curve analysis measures the trade-off between True positive rate (TPR, sensitivity) and False positive rate (1-specificity). The sensitivity is measured as the percentage of query genes that were ranked above a given threshold. The specificity is defined as the percentage of randomly selected test genes ranked below the threshold [4]. Performance differences were evaluated by a two-sided paired Wilcoxon signed-rank test. (For details see section S8). Additionally, we performed a control experiment on the DisGeNET data, replacing the query gene by a random gene not associated with any given *UMLS* class during LO-OCV (File S6, sheet 7; section S4.4).

3.3.7 Validation datasets

As a first benchmark dataset, we obtained 1,154 genes associated to 12 disease classes [73] used to validate previous prioritization tools [43] (File S1: sheet 6-7). The dataset is referenced throughout the text as the Goh *et. al* dataset. Included disease classes are Cardiovascular, Connective tissue, Dermatological, Development, Endocrine, Hematological, Immunological, Metabolic, Muscular, Ophthamalogical, Renal, and Skeletal. On average, 100 training genes were available per disease class.

A second benchmark dataset was obtained from the authors of HyDRA [103]. It consists of 8 diseases: Autism, Breast cancer, Colorectal cancer, Endometriosis, Ischaemic stroke, Leukemia, Lymphoma and Osteoarthritis (File S2 sheet 6) and was previously used to evaluate performance of HyDRA against Endeavour and ToppGene. ToppGene and Endeavour are supervised prioritization methods fusing 18 and 20 annotation sources respectively. In this work, we considered only scores obtained by the respective full annotation models.

Third, we extracted 9,414 curated genes, associated with 779 *UMLS* coded diseases from DisGeNET [169] (File S6 sheet 5). Within DisGeNET, we considered only diseases with 4 to 51 associated genes, resulting in a minimum of 3 and a maximum of 50 training genes during LO-OCV.

Finally, we simulated a prospective benchmark dataset, derived from HPO. For this, we extracted 2,025 HPO terms with 2,484 novel unique gene-phenotype associations added between January 2015 and February 2017 (File S7 sheet 1-2). For each selected HPO term, we extracted associated genes from the January 2015 release as training genes and performed genome wide prioritization of the novel gene. Similar to the 4 LO-OCV scenarios, we performed prioritization with and without inclusion of phenotype data from the test genes (labeled Test.Pheno.Include and Test.Pheno.Discard respectively). Additionally, we extracted a subset of 693 HPO terms having 1,111 unique gene associations to evaluate performance of pBRIT in Test.Pheno.Include mode to Endeavour-v3.71 (with usage of 24 annotation sources) and RWR-M (built with four annotation sources).

3.3.8 Implementation of pBRIT

Generation of sparse annotation matrices was done in python using a customized version of GOParGenPy [111]. TF-IDF and TF-IDF→SVD computation was done in R using the 'snow' [180] package to parallellize processing and 'irlba' [12] for TF-IDF→SVD computation. The web interface was developed using PHP as front-end and MySQL as back-end, connected to a torque/pbs job manager for prioritization job execution on a high-performance computing cluster.

3.4 Results

pBRIT was benchmarked against a set of published datasets. The individual datasets were chosen to range from very broad disease categories (Goh *et al.*, HyDRA), often with well known causative genes, to very specific diseases with a minimal number of known involved genes (DisGeNET, HPO). As such, the benchmark data represent an increasingly challenging validation trajectory. Similarly, competing methods were selected to either allow objective comparison on the respective benchmark data (Goh *et al.*, HyDRA), or to represent alternative state of the art methodologies in real life scenario's (Endeavour-v3.71; RWR-M). pBRIT is available as a web-interface and using a command line interface (batch mode). Prioritization of 100 test genes using 30 training genes takes on average 47.8 seconds using the web-interface. However, using the command line interface, prioritizing 10 similar sets of 100 test genes took approximately 83 seconds in total. Afterwards, results can be visualized using the web-interface.

3.4.1 BRR and SVD allows accurate and stable prioritization

LO-OCV on the Goh *et. al* data showed that most of the query genes were ranked among the top 15% highest scoring test genes, with a minimum AUC score of 0.86, under all 4 analysis scenarios (Table S4.1 and figure S3-A). Despite the broad disease classes and large amount of training genes per disease class, these results already highlight the relevance of different aspects of the pBRIT methodology. First, considering phenotype association scores of random test genes during regression improves AUC scores. This can be seen by comparing Test.N.Na and Test.ALL.Na scenarios, showing effects up to 7%, accompanied by an improvement in MRR from 0.148 to 0.075 (p-value = 3.3E-61, File S1: sheet 5). Second, singular value decomposition on the gene-by-feature profiles yields a better resolution of the similarity profiles, reflected in the slight improvement of AUC and MRR values over all disease classes when changing from TF-IDF to TF-IDF→SVD based feature extraction. Although the impact of SVD on the final prioritization results

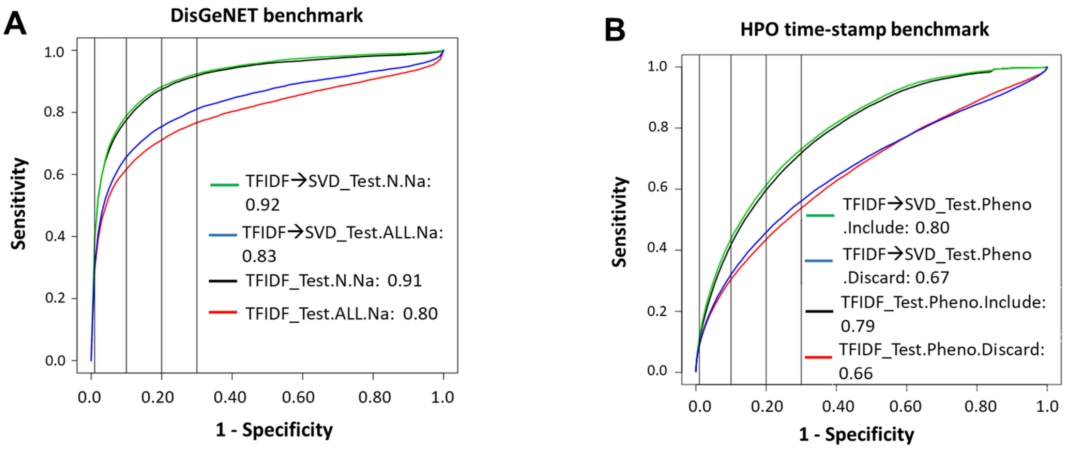


Figure 3.4.1: **ROC plot of pBRIT benchmark performance.** (A) 779 *UMLS*-coded disease classes obtained from DisGeNET and (B) 2,025 time-stamped HPO terms. The four vertical lines indicate the top1%, top10%, top20% and top30% of query genes which were prioritized.

is rather limited, the difference is significant (p-value = 4.86E-10, File S1: sheet 5). Furthermore, the higher gene-by-feature resolution will also help in the interpretation of the results (see section 3.6).

The dataset was already applied to benchmark 4 other methods, all applying early or intermediate data integration [43]. These methods were a) logistic regression based fast F_3PC algorithm b) Markov random field (MRF) c) Random walk with Restart (RWR) based network integration and d) Direct integration ranking (DIR) algorithm. The previously reported maximum AUC score on this dataset was 0.83, achieved by F_3PC. For MRF, RWR and DIR, the AUC scores were 0.731, 0.711 and 0.716 respectively. In our analysis, pBRIT performs better under all scenarios, with a maximum AUC score of 0.94 using the full model (TF-IDF→SVD_Test.N.Na).

Additionally to higher overall AUC scores, they show a lower variance over the individual disease classes compared to the competing methods (Figure S3-A,B). The global AUC score standard deviation of 0.015 under the full model indicates that pBRIT is not biased towards specific medical domains. In contrast, the F_3PC algorithm, being the best performing overall method, showed a maximum AUC score of 0.92 under the immunological disease class and a minimum AUC score of 0.68 under the developmental disease class, whereas pBRIT reaches AUC scores of 0.95 and 0.94 for these classes respectively.

3.4.2 Intermediate fusion provides uniform prioritization

Subsequently, we wanted to evaluate pBRIT's intermediate data integration against three methods representing late integration. For this, we used another benchmark dataset, previously used to evaluate Endeavour, ToppGene and HyDRA performance. ToppGene and Endeavour integrate ranks computed on individual annotation sources, while HyDRa is an ensemble of rank aggregation methods applied directly on the ranks computed from Endeavour and ToppGene.

The reported AUC score for Endeavour and ToppGene using full annotation models were 0.908 and 0.951 respectively. The best AUC values for HyDRA, using Weighted Kendall, were 0.91 and 0.947 respectively, based on Endeavour and ToppGene ranks. pBRIT has at least similar performance to these late integration methods, with an overall minimal AUC score of 0.93 and a maximum of 0.96 using the full model (see figure S3-B, table S4.2 and file S2). No significant improvement was observed using SVD for either N.Na or ALL.Na mode (p-value=0.91 and 0.12 respectively). However, there is a significant difference ($p\text{-value} < 0.0002$) between N.Na and ALL.Na mode for both feature mining methodologies (See File S2: sheet 5). A more in depth comparison, based on the MRR is available in table S4.2, showing improved MRR values compared to Endeavour for 7/8 included diseases. For 4/8 diseases, the full pBRIT model outperforms both Endeavour, ToppGene and HyDRA based rank aggregation methods. These results indicate that our regression approach after intermediate integration provides a uniform prioritization strategy independent of ensemble methods, with at least similar performance.

3.4.3 Effect of annotation changes on prioritization

Due to regular updates to the ever expanding biological knowledge base, annotation sources used in gene prioritization are highly dynamic. This is reflected in the monthly archives of ontology based annotation sources such as GO and HPO. Consequently, computing similarity profiles based on these ontologies will also be subjected to changes. As Bayesian Ridge Regression should help in modeling uncertainties related to changing annotations, we explored the potential impact of changing annotations on the prioritization results (section S5). Based on computational feasibility and data availability, we selected GO as part of the functional annotations and HPO as part of phenotypic annotations to construct yearly versions of the annotation framework, ranging from 2009 to 2014, keeping the remaining 8 annotation sources stable.

Ranking results of 250 genes from 8 disease classes of the HyDRA based dataset are summarized in File S3, S4 and S5, showing a variance of < 0.0002 on the overall

AUC scores over the included timeframe. Additionally, no significant correlation was observed between annotation changes and the overall change in gene ranking (Figures S5.2.1-5.2.12).

3.4.4 Effect of training set size and annotation bias

Although an ongoing debate in the machine learning domain is whether robust prediction requires more training data or better algorithms [230], the amount of training data is important for any supervised learning method. In the above benchmark sets, the number of training genes per disease class was large, especially for the Goh *et al.* data, and often involved well studied disease genes. Here, we evaluated pBRIT performance using limited training sets, targeting individual disease-gene associations extracted from the DisGeNET database [169] According to Figure 3.4.1A, table S4.3 and file S6-sheet 6, a small but significant ($p < 0.0005$) improvement in performance is seen between TF-IDF and TF-IDF→SVD feature extraction, using either regression strategy. On the other hand, the results again illustrate the importance of including phenotype association scores of both training and test genes during prioritization, with an overall improvement of over 10% in AUC ($p \simeq 0$). Analysis of MRR values (Figure 3.4.2) shows that this effect flattens out past 25 training genes. This is also reflected in a positive correlation between AUC scores and number of training genes for 'All.NA' setups (Pearson's product-moment correlation, $p < 0.01$, Figure S6.2, S6.4), which was absent for both 'N.Na' setups (Figure S6.2, S6.4).

These results demonstrate that the potential effect of annotation bias is minimal in pBRIT.

3.4.5 Real-World Performance Evaluation

LO-OCV has long been a standard approach to evaluate gene prioritization tools. Since well characterized genes tend to dominate prioritization results, LO-OCV estimates might be over-optimistic. Therefore, a real test for any prioritization tool should be its capacity to prioritize newly discovered genes with minimal disease association information. To achieve this, we evaluated pBRIT performance on HPO to gene associations assigned after creation of pBRIT's annotation database (January 2015, table S1). pBRIT was used to prioritize genes in the context of individual HPO phenotypic terms, instead of multi-phenotype diseases. A maximum AUC score of 0.80 and minimal MRR of 0.205 was obtained with the full pBRIT model (Figure 3.4.1B table S4.4, file S7). SVD had a small but significant positive effect on prioritization (p-value = 2.00E-56). Inclusion of phenotype data during regression again resulted in significantly better results for either

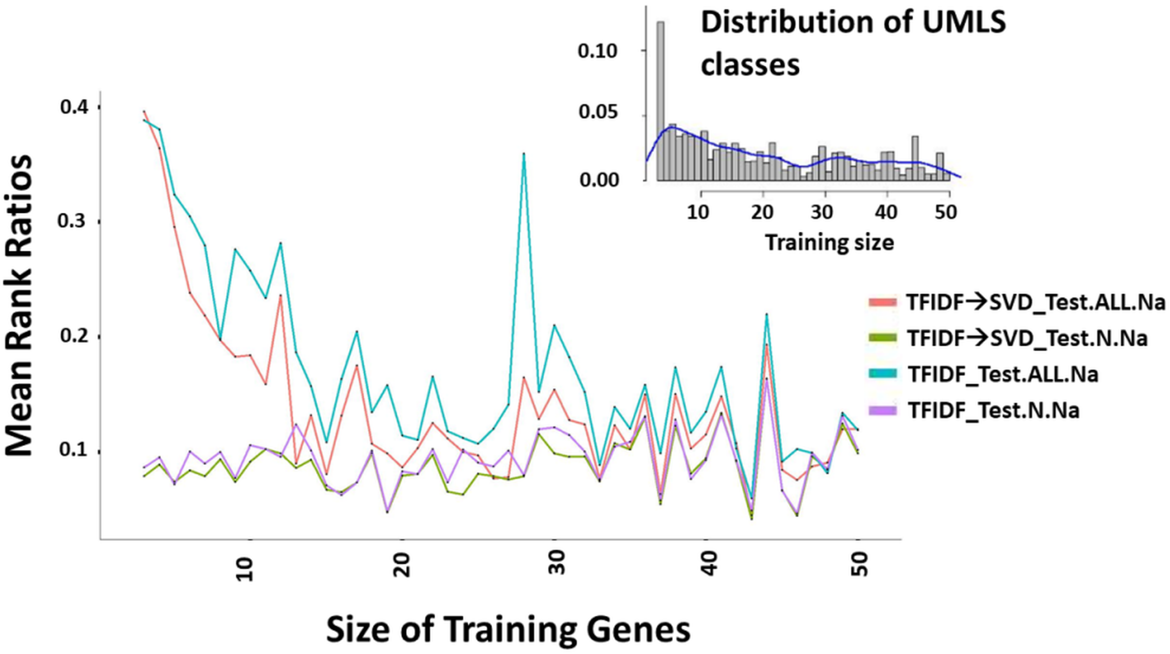


Figure 3.4.2: **Impact of training set size.** Main: Mean rank ratio (MRR) versus number of training genes. Incorporation of test gene phenotypic information (N.NA) in the regression model results in a low and stable MRR, irrespective of feature extraction methodology. Without phenotypic information (All.Na), MRR decreases with increasing number of training genes. Insert : Distribution of training sizes per disease class.

feature mining methodology, similar to the retrospective validations. Lastly, pBRIT (with an annotation release updated in December 2016) was directly compared with two recent tools, Endeavour-v3.71 and Random Walk with Restart on multiple networks (RWR-M), which both have internal annotation sources built in or before December 2016. We achieved a maximum AUC score of 0.87 in comparison to 0.85 for Endeavour ($p < 0.0004$) and 0.68 for RWR-M methods ($p < 7.666348e-196$). (See section S4.2.1 and Figure S4.3B for further details).

3.4.6 Results exploration and visualization

Researchers designing experiments based on prioritization results need insight into which annotation sources and training genes contribute more towards the ranking of specific genes. Although early and intermediate data fusion can obfuscate interpretation, we provide an interface to intuitively explore and explain these individual contributions.

As an example, prioritization results for *KCNA2* in the context of epileptic encephalopathy are shown in Figure 3.4.3 (For details, see section S7). The heatmap explains the gene-by-gene similarities. Darker shades indicate a larger contribution to the prioritization. *KCNA2* (marked in green) is top ranked mainly because of a higher

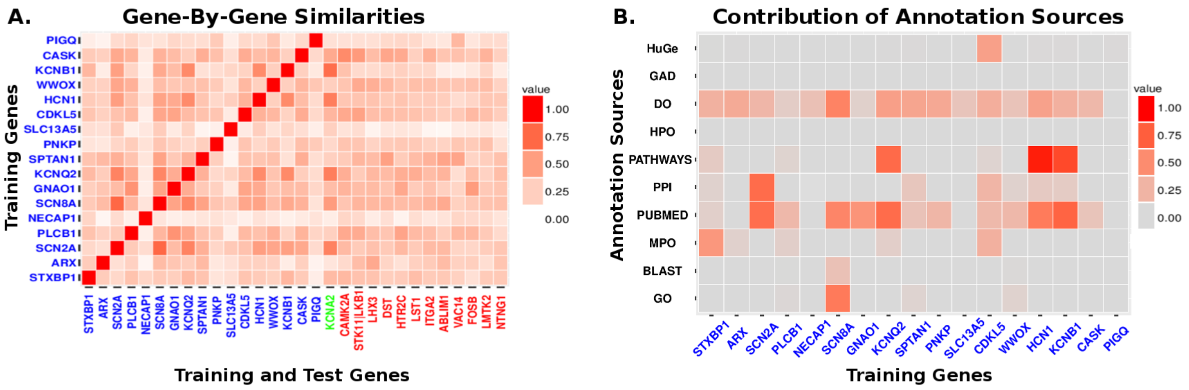


Figure 3.4.3: **Exploring prioritization results using heatmap plots.** A. The functional annotation matrix X, illustrating contribution of individual training genes (red) during regression, using the full TF-IDF→SVD_Test.Pheno.Include model. Darker shades indicate higher contributions. The example gene to be prioritized (*KCNA2*) is marked in green. B. Contribution of individual annotation sources for each training gene to the ranking of *KCNA2*.

similarity to *KCNB1*,*HCN1*, *KCNQ2* and *SCN2A*. Despite direct evidence in the literature of disease association for *NECAP1*, functional similarities to *KCNA2* are negligible. Comparison of Figures 5 and S7.1 shows that SVD transformation of the gene-by-feature matrices results in visibly more pronounced similarity scores. Second, pBRIT provides heatmaps of similarity scores per individual annotation source (Figure 3.4.3B These gene-specific plots highlight the training genes and annotation sources contributing most to the ranking of that particular gene. Again, it can be seen from Figures 5 and S7.1 that SVD provides more pronounced similarity profiles.

Finally, the pBRIT web-interface provides actual overlapping features between training and test genes, with the corresponding TF-IDF scores.

3.5 Discussion

We present a novel gene prioritization tool, based on Bayesian Ridge regression and utilizing an information-theoretic approach towards feature extraction followed by intermediate data integration. We compared pBRIT performance to 9 current state-of-the-art methods under a variety of conditions, reflecting both different aspects of our methodology and varying degrees of prior evidence.

Although the Goh *et al.* [73] benchmark set does not represent a typical gene prioritization use case due to extensive and curated gene lists associated to high level disease classes, important conclusions could be drawn from it. First, it provides initial evidence that the implemented TF-IDF approach is feasible, as pBRIT globally outperforms four existing methods using alternative approaches, which were originally benchmarked on this dataset. It thus demonstrates the validity of applying TF-IDF in discriminatory mining of genomic features other than textual information, for which it was originally presented. In our case, these features are structured concepts holding specific details about gene functionality or phenotype associations. Furthermore, leveraging of phenotypic information and performing SVD transformation of the feature-by-gene matrices, being two of the key characteristics of pBRIT, improves AUC scores by approximately 9%. Finally, stable AUC scores across individual disease classes indicates that prioritization is not biased towards particular disease classes.

The singular value decomposition of TF-IDF weighted gene-by-feature matrices, prior to data integration is a novel characteristic of pBRIT. We applied SVD transformation because TF-IDF based weights show two limitations. First, weights are generally computed for every individual feature assuming feature independence. However, features do co-occur in biological data and as such contain additional functional information about the gene. Second, since most features are rather specific, the binary gene-by-feature matrix holds many zeros. This might impact feature learning and generally leads to less cohesive gene clusters, in turn affecting the overall gene-by-gene similarity profiles used to prioritize candidate genes. SVD allows a reduction in the feature space dimensionality, thereby reducing sparseness and implicitly modelling co-occurrences and latent relationships. Despite a limited 1-4% gain in performance over TF-IDF alone, the benefits are twofold. First, latent relations can be helpful to identify candidate genes in rare diseases having related disorders with overlapping phenotypes. Here, the choice of SVD based feature mining can improve prioritization based on training genes implicated in those related disorders. Second, as shown in Figure 3.4.3A,B, SVD increases the resolution of the gene-by-gene similarity profiles, facilitating result interpretation using the provided visualization tools.

In reality only a limited set of training genes can be defined for most genetic disorders. These genes are often less studied and reflect a subjective measurement of how well they describe the underlying disease etiology. We simulated this by selecting DisGeNET disease classes with maximally 51 associated genes. The obtained results illustrate the power of the second key characteristic of pBRIT. Indeed, we observe a significant improvement of approximately 11-14% (table S4.3) when phenotypic information from the 99 random test genes is taken into account (Test.N.Na) during regression. As this effect is more pronounced for smaller training sets, it makes pBRIT a valuable asset in the study of rare and less studied disorders.

Given the potential impact of annotation changes over time on functional interpretation [217], it should be noted that pBRIT results are not subject to such changes. This is likely attributable to BRR, which implicitly captures the uncertainties in the model arising due to changes, and stabilizes the ranking. Although all annotation sources get updated, our simulation was limited to HPO and GO due to the availability of archived and quantifiable data over a fixed time interval. Therefore, we can not fully exclude that the impact is nullified by the contribution of the 8 annotation sources that were kept stable during the experiment.

Overall, pBRIT performs equally well, and often better than competing methods on cross-validation studies. Although we removed prior knowledge on phenotypic association between test and training genes during regression, this knowledge still contributed indirectly to similarity scores through the IDF component of the TF-IDF calculations. To exclude any indirect contribution of prior knowledge, the real test is therefore to prioritize genes that have been published after construction of the internal annotation database. For these genes, the tools by definition lack any prior knowledge of the gene-disease association. Hence, we performed a final validation of pBRIT on 2,025 HPO terms having 2,484 novel gene-phenotype associations. Interestingly, the obtained maximum AUC score of 0.80 using the full pBRIT model, is lower in comparison to the performed LO-OCV based analyses, confirming the tendency of LO-OCV to overestimate performance. A subset of 693 HPO terms was also analyzed using Endeavour-v3.71 and RWR-M, with neither method outperforming the full pBRIT model. The inferior performance of RWR-M, using only 4 annotation sources, mainly demonstrates that integration and fusion of more and relevant annotation sources has a distinctive advantage. Since Endeavour-v3.71 and pBRIT both use approximately similar annotation sources, our results indicate that intermediate integration combined with BRR offers a valuable alternative to late integration and rank fusion while offering superior computation speed.

Additionally, we want to highlight the importance of community driven data competitions in this context (like CAFA [94] for developing true prospective benchmark datasets, such that future function prediction and prioritization tools can be properly evaluated on high quality and unbiased datasets.

Despite the promising performance of pBRIT, we believe that further improvements are possible. First, alternative to empirical selection of k in SVD, one might apply a probabilistic generative model, either using classical Latent Dirichlet Allocation [24] or aspect Bernoulli models [23]. Second, we focused on giving equal weights to all annotation sources during the construction of composite similarity matrices. However, data driven approaches for optimal weight selection might further improve performance. Finally, it might be interesting to investigate the influence of incorporating informative priors such as Zellner's g-prior [78] in the Bayesian ridge regression.

3.6 Conclusion

In conclusion, our results present pBRIT as robust and performant. Its performance was competitive, or better, compared to current state-of-the-art methods when applied to their benchmark datasets. We demonstrated performance of pBRIT both at the level of the information-theoretic model, by evaluating TF-IDF and SVD as feature extraction approaches, and by contrasting intermediate data fusion to other data fusion methodologies, and at the level of the regression model, by evaluating the effect of incorporating phenotypic information from test genes into the model. Additionally, we explored the predictive power of pBRIT to detect novel disease causing genes without prior information in the internal database. We demonstrated that regression under the Bayesian framework has an advantage in handling uncertainties and errors in the annotation sources, while incorporation of a ridge regression model helps in alleviating the problem of over-fitting and multi-collinearity in the model. Ultimately, these aspects lead to a more robust prediction. We can therefore conclude that each aspect of the pBRIT methodology provides distinct and additive benefits, making the TF-IDF→SVD.Pheno.Include approach, referenced as the full model, the method of choice in real-world application. Finally, we extended the prioritization task to providing insight in the resulting gene ranks through visualization. Using heatmap plots showing both pre- and post-integration similarity scores, together with actual feature matches between training and test genes, interpretation of gene ranks becomes intuitive.

3.7 Acknowledgements

This research was supported by funding from the University of Antwerp (Lanceringsproject), the Fund for Scientific Research, Flanders (FWO, Belgium, G.0221.12, 1513715N and 12D1717N), The Dutch Heart Foundation (2013T093) and the Fondation Leducq (MIBAVA-Leducq 12CVD03). Work by KULeuven researchers was supported by KU Leuven CELSA/17/032, the Flemish Government (IWT 150865, FWO 06260) and VIB: ELIXIR. BL is senior clinical investigator of the Fund for Scientific Research, Flanders and holds a starting grant from the European Research Council (ERC-StG-2012-30972-BRAVE).

3.8 Supplementary Data

3.8.1 Annotation Sources used in pBRIT

Annotation Sources	Download date	Download links
Pubmed abstract	18.05.2014	Customized Script
Pathway	07.01.2014	`http://cpdb.molgen.mpg.de`
Protein-Protein Interactions	08.01.2014	`http://cpdb.molgen.mpg.de`
Gene Ontology (GO)	04.04.2014	`www.geneontology.org`
Mammalian phenotype ontology (MPO)	04.09.2014	`ftp://ftp.informatics.jax.org/pub/reports/index.html#pheno`
Disease Ontology (OBO)	26.01.2015	`https://github.com/DiseaseOntology/HumanDiseaseOntology/blob/master/src/ontology/HumanDO.obo?raw=true`
Human Phenotype Ontology (HPO) OBO file	06.01.2014	`http://purl.obolibrary.org/obo/hp.obo`
HuGe Navigator	10.09.2014	`https://phgkb.cdc.gov/HuGENavigator/downloadCenter.do`
Genetic Association Database (GAD)	10.09.2014	`https://geneticassociationdb.nih.gov/data.zip`
Uniprot (BLAST)	06.02.2014	`www.ensembl.org/biomart`

Table S1: Annotation sources used in pBRIT.

pBRIT integrates 10 annotation sources categorized as being phenotypic or functional in nature. The phenotypic annotation sources are: Genetic Association Database (GAD) [16], Human Genome Epidemiology (HuGe) navigator [225], Disease Ontology (DO) [186], Human Phenotype Ontologies (HPO) [108]. The functional annotation sources include: Pubmed abstracts [184], Pathway databases [98] and Protein-Protein Interactions (PPI) [98], Mammalian Phenotype Ontologies (MPO) [195], Gene Ontology (GO)[9] and Uniprot [203] protein sequences similarities. Protein sequence similarities were generated using offline BLAST [7] tool. PPI and Pathway databases were download from the ConsensusPathDB[98] online database. The pathway database from ConsensusPathDB includes information from Biocarta, EHMN, HumanCyc, INOH, KEGG, NetPath, PharmGKB, PID, Reactome, Signalink, SMPDB, WikiPathways. For PPI, ConsensusPathDB includes information from PhosphoPOINT, PDZBase, NetPath, PINdb, BIND, CORUM, Biogrid, InnateDB, MIPS-MPPI, Spike, Manual upload, MatrixDB, DIP, IntAct, MINT, PDB, HPRD.

Table S1 provides a list of all annotation sources, their download date and a link for download. For pubmed abstracts an automated script in python was used to download abstracts based on the pubmed-id linked to the genes (See section 1.3 for extraction and processing details).

3.8.2 Sparseness and dimensionality of annotation sources

This section describes the dimensionality of each of the annotation sources and the corresponding degree of sparseness present in them. Computation of gene-by-gene proximity profiles yields less cohesive clusters due to presence of sparseness.

Annotation Sources	Dimensionality	Summary statistics of columns with non-zero entries			Non-zero entries (%)
		Mean	Median	Max	
Pubmed Abstract	12254 × 61814	236	172	3015	0.38
Pathway	10529 × 3479	11.21	5	381	0.32
Protein-Protein Interaction (PPI)	15021 × 16298	17.61	6	8523	0.10
Gene Ontology (GO)	22100 × 17790	69.09	50	940	0.39
Mammalian Phenotype Ontology (MPO)	7543 × 5926	21.80	13	383	0.37
Disease Ontology (DO)	15095 × 4686	53.30	31	1011	1.13
Human Phenotype Ontology (HPO)	2872 × 6615	88.80	60.50	716	1.34
Human Genome Epidemiology (HuGe)	11692 × 2675	36.3	18.0	1089	1.35
Genetic Association Database (GAD)	11899 × 3124	10.38	9	730	0.33
Protein Sequence Similarities (BLAST)	21994 ×21994	12120	16120	18300	55.10

Table S1.1: **Dimensionality and degree of sparsity**. The dimensionality and degree of sparsity present across different annotation sources used in pBRIT. Summary statistics was used to compute the degree of sparsity.

3.8.3 Proportion of Variance Explained after SVD

We applied Singular Value Decomposition (SVD) to address the presence of sparsity and to capture dependencies between the features in a given annotation source. Given the high number of rows and columns in the annotation matrices, we applied the irlba package in R for faster computation of truncated singular value decomposition of the sparse matrices. As described in the manuscript, let $A_{m \times n}$ be the matrix with TF-IDF scores where m and n are the total number of genes and corresponding associated features respectively. Using equation 2 of the manuscript, with optimal choice of k the matrix is decomposed into the left singular matrix (represented by $U_{m \times k}$), a diagonal matrix (represented by $D_{k \times k}$) and the right singular matrix (represented by $V_{n \times k}$). We empirically chose $k = 200$ for truncated SVD, as there has no single best strategy or algorithm been defined to determine the optimal value of k with respect to prediction or classification. Our choice of k is based on our goal of addressing annotation sparsity and modelling co-occurrences. The summary statistics of columns (features associated with genes in the original TF-IDF matrices) for each of the annotation sources are shown in Table S1.1. From there it can be seen that the maximum (200 non zero columns) is reached for Pubmed abstracts and the minimum (10.38) for the GAD database. Hence, by choosing K=200 we can model sparsity and co-occurrences. Figure S1.2 demonstrates the proportion of variance explained by the first 200 singular values after SVD based decomposition for 9 annotation matrices, illustrating the validity of our empirical selection of k.

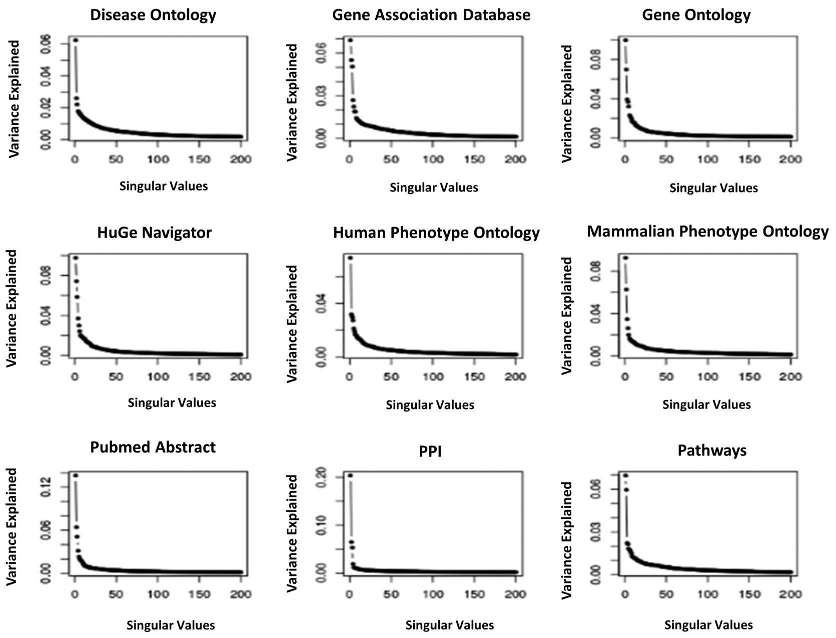


Figure S1.2: **Proportion of variance explained by k components**. k was empirically set to 200 components in singular value decomposition (SVD) computation. Using this setting, SVD was uniformly applied on all annotation sources except BLAST.

3.8.4 Processing Annotation sources

Figure S1.3 presents a schematic workflow for processing annotation sources. All annotation sources except the BLAST annotation source were pre-processed using GOParGenPy to obtain a sparse binary representation. GOParGenPy was initially designed for processing GO data. We incorporated its OBO processing functionality and extended it to process OBO structures of other ontologies. Additionally, we also utilized its sparse matrix generation functionality for non-ontological annotation sources.

For PubMed, we only extracted abstracts that have a corresponding GO-id linked to it. The mapping between pubmed-id and GO terms was retrieved from Uniprot. The text-mining of Pubmed abstract involves two stages. First, all high frequent stop words were removed. Second, a stemming algorithm (from Snowball incorporating English vocabulary) was applied to process the lexical forms of words to their base or root form. Post stemming, a gene-by-word-feature sparse matrix is created. For ontology based annotations (such as HPO, GO, DO, MPO), the DAG structure was parsed to retrieve parental terms. Other annotation sources such as GAD, HuGE, Pathways, PPI were directly reformatted to gene-by-feature sparse binary matrices. An information theoretic approach for feature mining using TF-IDF and TF-IDF→SVD methodology was then applied uniformly to these gene-by-feature matrices, on a per annotation source level. Using matrix dot product, gene-by-gene similarity matrices were then computed. For protein sequence data, all human protein sequences, obtained from UniProt, were BLASTed against themselves and their normalized bit-score (by taking unit-norm) was used to compute the final gene-by-gene similarity scores. Default parameter settings[1] were used for protein BLAST (*blastp*). For a given gene pair, the bit-score of the most similar protein/gene sequences (over all isoforms) were divided by the square root of the summed squares. This is basically the same as computing unit-norm such that the normalized score lies between 0 and 1.

The gene-by-gene proximity profiles for each of these annotation sources were then merged for either the functional or phenotypic annotation category to create composite matrices that will be utilized in regression. This procedure was performed both for TF-IDF and for TF-IDF→SVD based feature extraction.

3.8.5 Validation Strategy

The overall performance of pBRIT was evaluated by performing a leave one-out cross-validation (LO-OCV) and compared with 7 different tools using their respective benchmark datasets. Additionally, we simulated a real life scenario to evaluate how well

---

[1] `https://www.ncbi.nlm.nih.gov/books/NBK279684/`

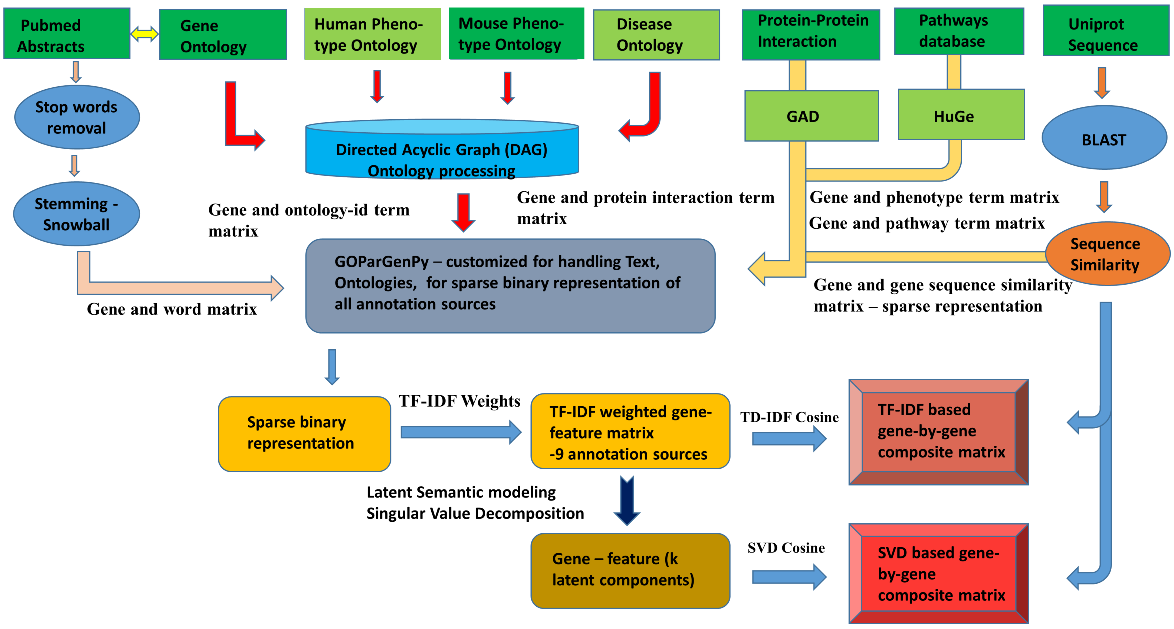

Pubmed Abstracts
Gene Ontology
Human Pheno-type Ontology
Mouse Pheno-type Ontology
Disease Ontology
Protein-Protein Interaction
Pathways database
Uniprot Sequence
Stop words removal
Directed Acyclic Graph (DAG) Ontology processing
GAD
HuGe
BLAST
Stemming - Snowball
Gene and ontology-id term matrix
Gene and protein interaction term matrix
Gene and phenotype term matrix
Gene and pathway term matrix
Sequence Similarity
GOParGenPy – customized for handling Text, Ontologies, for sparse binary representation of all annotation sources
Gene and gene sequence similarity matrix – sparse representation
Gene and word matrix
Sparse binary representation
TF-IDF Weights
TF-IDF weighted gene-feature matrix -9 annotation sources
TD-IDF Cosine
TF-IDF based gene-by-gene composite matrix
Latent Semantic modeling
Singular Value Decomposition
Gene – feature (k latent components)
SVD Cosine
SVD based gene-by-gene composite matrix

Figure S1.3: **Schematic workflow for processing annotation sources.** The steps involve pre-processing of all the annotation sources to obtain sparse binary representations, with exception of BLAST similarity scores. Subsequently, TF-IDF and TF-IDF→SVD methodology is applied to compute gene-by-gene proximity matrices. Finally, a composite matrix is created for functional and phenotypic categories.

pBRIT can prioritize newly discovered disease genes without having any prior information on the association with the disease or phenotype. Algorithm 1 describes the complete procedure of prioritization steps using pBRIT. It requires a) a set of input training genes, b) test genes, c) choice of feature mining methodology which could be either TF-IDF or TF-IDF→SVD and d) the selected analysis scenario. These scenarios can be i) N.Na and ALL.Na for LO-OCV and ii) Pheno.Include and Pheno.Discard for real-world usage. The output is a prioritized list of test genes sorted according to phenotypic concordance score after regression.

step 1: Based on the chosen feature mining methodology (which are computed as given by equation 1 and 2 in the manuscript) the proximity scores for the given set of training and test genes are extracted from the respective composite matrices (given by equation 3).As seen from the regression design given in equation 4, the prioritization result of pBRIT is based on the phenotypic concordance score of the test genes. Hence, it is important to highlight how much of this information can influence the prioritization. Therefore, in **step 2** the regression design is altered for prioritization based on the chosen analysis mode. To evaluate the performance during LO-OCV analysis, existing information of disease/phenotype is removed for the query gene only, resulting in N.Na mode, or for all test genes (including query gene), resulting in the ALL.Na mode. In the real-world analysis scenario, we want to evaluate how well pBRIT can prioritize newly discovered gene associations without any prior knowledge of the respective association. Hence, we translated the respective modes into the similar setups of Pheno.Include (use phenotypic information) and Pheno.Discard (discard all phenotypic information of the test genes).

It is noteworthy to mention how these analysis modes get implemented into the regression model. **Step 3** yields the regression model (as shown in equation 4) for the given training and test genes with the functional annotation matrix represented as $X_{(n+m)\times n}$ and the phenotypic concordance vector as $Y_{(n+m)\times 1}$:

$$\mathbf{Y}_{(n+m)\times 1} = \beta \mathbf{X}_{(n+m)\times n} + \boldsymbol{\varepsilon}; \text{where, error term } \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

In this model, the following parameters are the respective unknowns: regression coefficient β, its variance σ_β and variance of error term σ_ε. The coefficient of regression can be uniquely estimated and is provided by equation 5 and 6 in the manuscript which indicates the incorporation of $Y_{(n+m)\times 1}$ and $X_{(n+m)\times n}$ in the estimation. Hence, under the respective analysis modes N.Na and ALL.Na, the corresponding "NA" values of the query gene or all test genes are substituted by the weighted.mean of all *non-NA* values of the $Y_{(n+m)\times 1}$ vector. This is internally done by the BLR package. The customized version of ridge regression functionality used in pBRIT in the BLR package can be found here [2].

Under the bayesian framework the regression is performed where the likelihood of the model is given by equation 7. The prior on β is given by equation 8, the prior on variance of regression coefficient is given by equation 9 and the prior on variance of the error term is given by equation 10. The posterior distribution of parameters conditional to the given phenotype information is formulated using equation 11. Utilizing the property of conjugacy in bayesian frameworks, we assume NIG priors for these unknown parameters. Additionally, as the number of the test genes grows, as seen for genome wide prioritization, the closed form of equation 11 becomes intractable. Hence, incorporation of a Gibbs sampler eases the estimation of posterior parameter distribution. We incorporate the overall functionality of the ridge regression BLR package to perform regression analysis. As per the guidelines of the package, the initialization of parameters is done in **step 5**. For the respective priors, the initial values were derived from the training genes only. Actual regression is performed in **step 6** of the algorithm.

---

[2] `https://bitbucket.org/medgenua/pbrit/raw/d1b403582263548f5469b04cc33b0313956c61f5/Analysis_Files/BLR_RIDGE.R`

Finally, using the learned parameters, phenotypic concordance scores are predicted using equation 12 and 13. The predicted scores are sorted in non decreasing order to give the final prioritized list of test genes as shown in **step 7**.

```
Algorithm 1: Algorithm for Cross Validation analysis
  Input: List of parameters and input dataset
  $S_{Train} \leftarrow \{G_1, G_2, G_3, G_4, ....G_n\}$ Set of $n$ training genes
  $S_{Test} \leftarrow \{G_{1'}, G_{2'}, G_{3'}, G_{4'}, ..G_q...G_m\}$, Set of $m$ test genes that need to be prioritized
  $analysis\_mode \leftarrow$ {N.Na or ALL.Na; Pheno.Include or Pheno.Discard}
  $method \leftarrow$ {TF-IDF or TF-IDF→SVD}
  $query\_gene \leftarrow G_q; num\_of\_iterations, burnIn, thin$
  Result: Set of Prioritized Test list
  $S_{RankTest} \leftarrow \{G_{p1'}, G_{p2'}, G_{p3'}, G_{p4'}, ..G_{pq}...G_{pm}\}$ Final prioritized list
  Step 1: Extract proximity profiles for Training and Test genes under the
    regression design setting as mentioned in equation 4 in the manuscript.
  if $method = TF\text{-}IDF$ then
      $X_{Train(n \times n)} \leftarrow X_{composite-TFIDF}[S_{Train}, S_{Train}]; Y_{Train(n \times n)} \leftarrow Y_{composite-TFIDF}[S_{Train}, S_{Train}]$
      $X_{Test(m \times n)} \leftarrow X_{composite-TFIDF}[S_{Test}, S_{Train}]; Y_{Test(m \times n)} \leftarrow Y_{composite-TFIDF}[S_{Test}, S_{Train}]$
  else
      $method =$ TF-IDF→SVD
      $X_{Train(n \times n)} \leftarrow X_{composite-SVD}[S_{Train}, S_{Train}]; Y_{Train(n \times n)} \leftarrow Y_{composite-SVD}[S_{Train}, S_{Train}]$
      $X_{Test(m \times n)} \leftarrow X_{composite-SVD}[S_{Test}, S_{Train}]; Y_{Test(m \times n)} \leftarrow Y_{composite-SVD}[S_{Test}, S_{Train}]$
  end
  Step 2: Regression design, based on the chosen analysis mode scenario.
  if $analysis\_mode = N.Na$ then
      $Y_{Test}[G_q] \leftarrow$ NA, Used for all LO-OCV scenarios
  else if $method = ALL.Na$ then
      $Y_{Test} \leftarrow$ NA, Used for all LO-OCV scenarios
  else if $method = Pheno.Discard$ then
      $Y_{Test} \leftarrow$ NA, Used for time-stamped HPO benchmark
  else
      $Y_{Test} \leftarrow Y_{Test}$, Used for time-stamped HPO benchmark
  end
  .. continued
```

```
Step 3: Combine and arrange matrices for multiple regression according to
  equation 4 and the analysis scenario as mentioned in Step 2.
$Y_{(n+m)\times n} \leftarrow \{Y_{Train(n\times n)}, Y_{Test(m\times n)}\}$
$Y_{(n+m)\times 1} = \sum_{j=1}^{n} y_{ij}$
$X_{(n+m)\times n} \leftarrow \{X_{Train(n\times n)}, X_{Test(m\times n)}\}$
Step 4: Reduction of non-linearity effects
$Y_{(n+m)\times 1} \leftarrow \sqrt{Y_{(n+m)\times 1}};$
$X_{(n+m)\times n} \leftarrow \sqrt{X_{(n+m)\times n}};$
Step 5: Initialization of regression parameters:
$df_\varepsilon = df_\beta = 3;$
$V_\varepsilon \leftarrow Var(Y_{Train(n\times 1)});$
$S_\varepsilon \leftarrow V_\varepsilon(df_\varepsilon + 2);$
$S_\beta = \frac{\text{Var(Y}_{\text{Train}}\text{)}\times(df_\beta+2)}{\sum_j^n \text{Var(X}_{\text{Train}j}\text{)}};$
Step 6: Regression using BLR package. It predicts the posterior expected value
  of phenotypic concordance score according to equation 12 and 13.
$Y_{Pred(n+m)\times 1} \leftarrow$
  $BLR(Y_{(n+m)\times 1}, X_{(n+m)\times n}, (df_\varepsilon, S_\varepsilon), (df_\beta, S_\beta), numOfIterations, burnIn, thin);$
;
Step 7: Prioritization of Test genes. Separate the $Y_{Pred}$ scores into respective
  Training and Test labels.
$Y_{PredTrain(n\times 1)} \leftarrow Y_{Pred(n+m)\times 1};$
$Y_{PredTest(m\times 1)} \leftarrow Y_{Pred(n+m)\times 1};$
$Y_{RankTest(m\times 1)} \leftarrow ReverseSort(Y_{PredTest(m\times 1)});$
```

3.8.6 Stability of AUC score for Goh et. al benchmark dataset

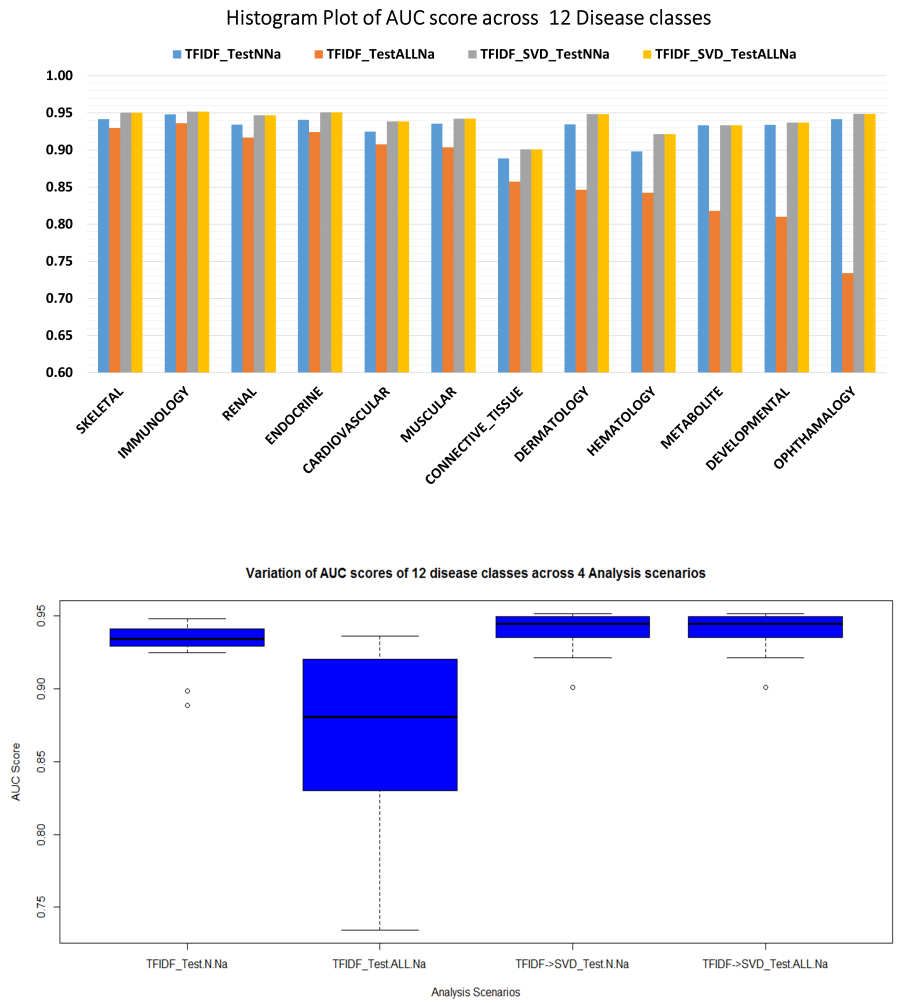


Figure S3: **Stability of pBRIT AUC scores for the Goh et. al benchmark dataset** (a) Histogram plot of AUC scores across 12 disease classes under the four available analysis scenarios in pBRIT. (b) Boxplot of AUC scores for the four available analysis scenarios indicating the AUC scores are more stable in N.Na mode than ALL.Na modes

3.8.7 Benchmark results.

3.8.7.1 ROC curve plots for performed benchmark analyses

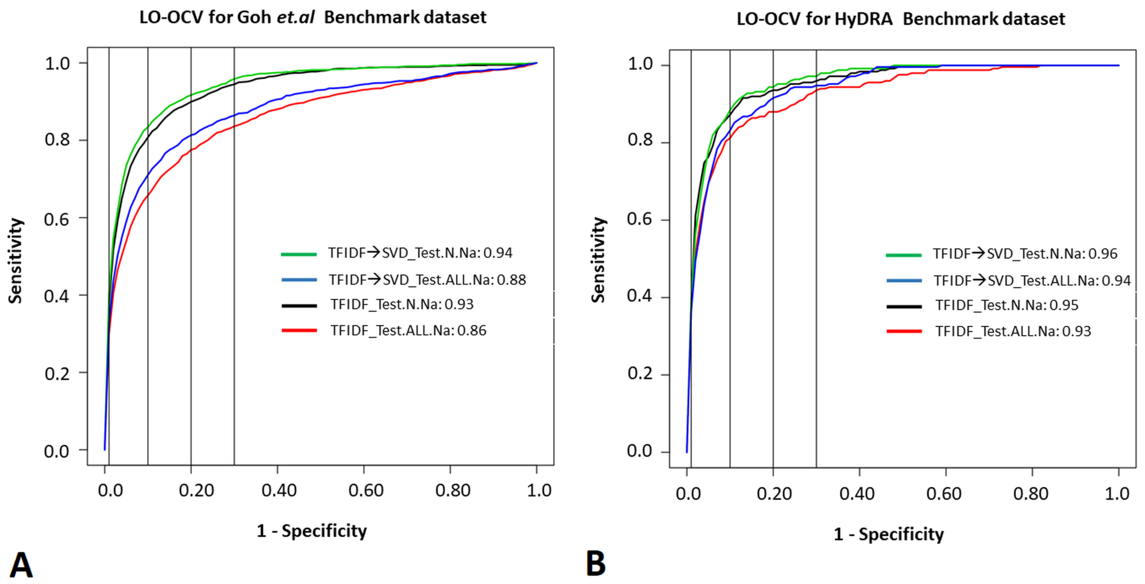


Figure S4.1: **LO-OCV result on Goh *et al.* and HyDRA benchmark datasets:**A) Goh et. al benchmark of 1,154 disease genes associated to 12 disease classes. B) HyDRA prioritization tool dataset having 250 genes associated to 8 disease classes. The four vertical lines indicate top 1%, top 10%, top 20% and top 30% of query genes which were prioritized

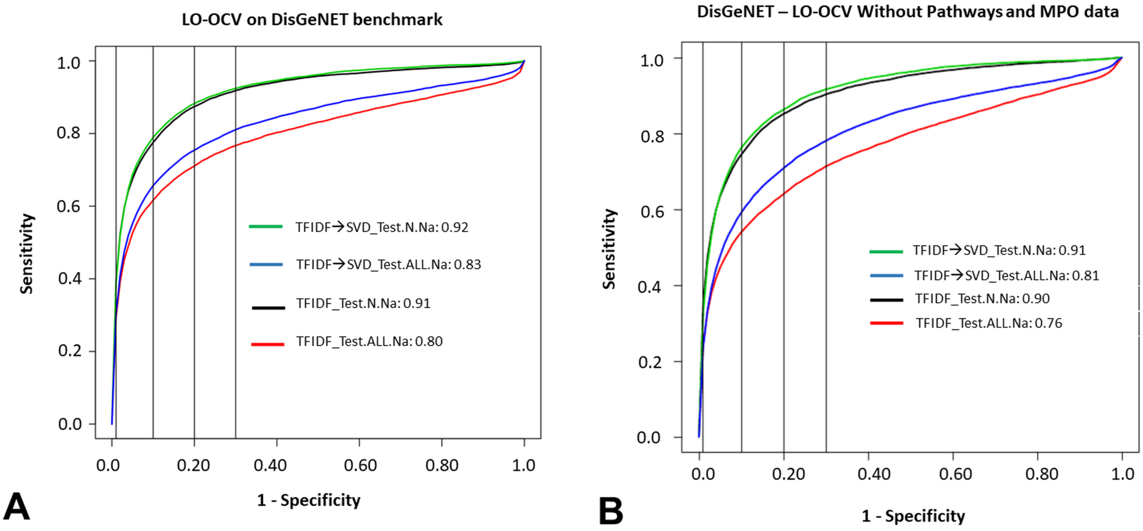


Figure S4.2: **LO-OCV result on DisGeNET benchmark dataset**, **A**) 9,414 genes associated with 779 *UMLS* classes were prioritized uing all 10 annotation sources. **B**) 9,121 genes associated to 767 *UMLS* classes were prioritized without using pathways and mammalian phenotype ontology (MPO) annotation sources. The four vertical lines indicate top 1%, top 10%, top 20% and top 30% of query genes which were prioritized.

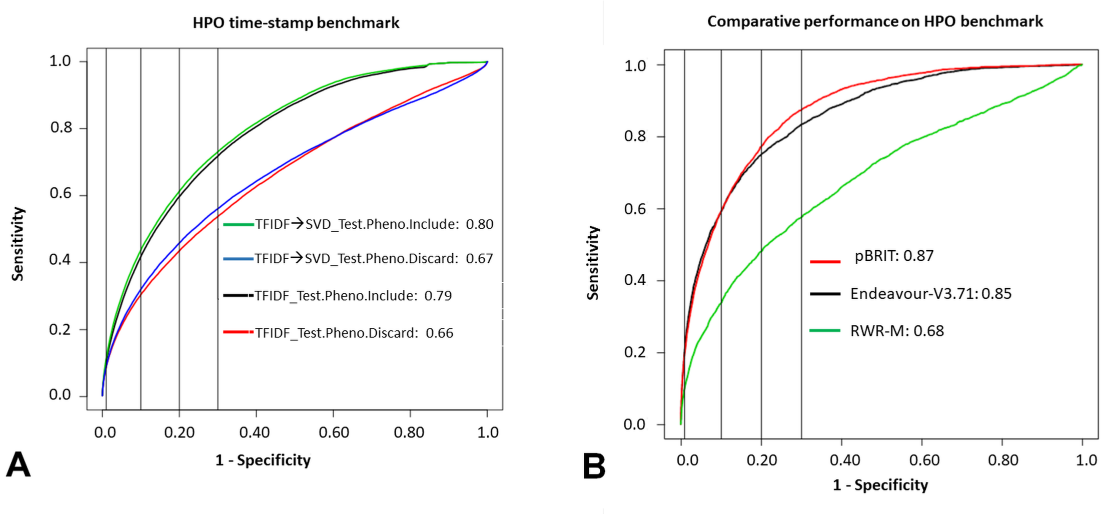


Figure S4.3: **Performance on prospective HPO benchmark dataset.** **A**) Real world usage performance of pBRIT on a benchmark dataset of 2,025 prospective HPO classes, using the January 2015 annotation release of pBRIT for genome wide prioritization. **B**) Performance of pBRIT (TFIDF→SVD.Pheno.Include method) in comparison to Endeavour-v3.71 and Random-walk with restart (RWR-M) on a random subset (693 HPO classes) of the previous HPO benchmark dataset. Here, the December 2016 annotation release was used for genome wide prioritization. The prospective HPO data was downloaded on March 2017 (see Supplementary File S7 for details). The four vertical lines indicate top 1%, top 10%, top 20% and top 30% of query genes which were prioritized.

3.8.8 pBRIT benchmark performance metrics

Analysis Scenario	top1%	top5%	top10%	top30%	MRR	AUC Score
TF-IDF_Test.N.Na	37.61	69.49	80.76	94.45	0.075	0.93
TF-IDF_Test.ALL.Na	30.00	53.94	65.74	83.60	0.148	0.86
TF-IDF→SVD_Test.N.Na	40.43	73.50	83.33	95.84	0.066	0.94
TF-IDF→SVD_Test.ALL.Na	32.69	58.99	71.02	86.41	0.128	0.88

Table S4.1: **Performance of pBRIT on the Goh *et. al* benchmark set.** Results of TPR for LOOCV with respect all 4 analysis scenarios (see methods). The reported values reflect the average percentage of training genes ranked in the top X% of test genes over all disease classes. MRR: Mean-Rank-Ratio. Rank ratio is defined as the obtained rank of the test gene, divided by the number of test genes; AUC: Area under the curve. Ranks of all prioritized genes, and the corresponding classes are available in S1 File.

Disease Class	# Training Genes	ToppGene	Endeavour	Lovasz-Bregman HyDRA	Hybrid Borda HyDRA	Hybrid Kendall HyDRA	pBRIT TFIDF_TestNNA	pBRIT TFIDF_Test.ALL.Na	pBRIT TFIDF→SVD_Test.N.Na	pBRIT TFIDF→SVD_Test.ALL.Na
Autism	40	7.275	17.96	11.2	9.75	6.85	9.5	13.47	8.5	8.72
Breast Cancer	10	4.6	14.4	7.1	12	2.5	7.5	7.0	5.3	7.0
Colorectal Cancer	20	7.3	8.55	5.2	7.85	8.7	5.7	7.1	5.05	9.65
Endometriosis	43	6.46	5.3	8.63	10.63	7.74	4.76	5.76	4.09	7.09
Ischaemic stroke	44	5.61	6.18	7.25	9.25	6.05	3.04	6.21	2.84	4.08
Leukemia	10	5.5	13.7	12	6.6	10.2	14.6	13.4	15.1	13.9
Lymphoma	42	3.74	9.57	6.45	9.26	2.93	3.65	5.14	3.90	5.81
Osteoarthritis	41	6.44	5.56	6.32	7.46	5.41	3.09	8.92	2.99	4.02

Table S4.2: **Performance of pBRIT compared to Endeavour, ToppGene and HyDRA** Mean Rank Ratios (MRR) for 8 diseases using ToppGene, Endeavour, 3 different implementations of HyDRA and four different analysis strategies of pBRIT. For Endeavour and ToppGene, analysis was performed using all the annotation sources present. For pBRIT, ranks of all prioritized genes, and the corresponding diseases are available in S2 File.

Analysis Scenarios - A	top1%	top5%	top10%	top30%	MRR	AUC Score
TF-IDF_Test.N.Na	39.17	67.52	77.54	91.82	0.094	0.91
TF-IDF_Test.ALL.Na	29.41	52.59	61.62	76.67	0.208	0.80
TF-IDF→SVD_Test.N.Na	39.04	68.72	78.93	92.38	0.088	0.92
TF-IDF→SVD_Test.ALL.Na	30.34	55.19	65.65	81.04	0.174	0.83
Analysis Scenarios - B	**top1%**	**top5%**	**top10%**	**top30%**	**MRR**	**AUC Score**
TF-IDF_Test.N.Na	35.25	63.45	74.40	90.30	0.103	0.90
TF-IDF_Test.ALL.Na	24.11	44.61	53.96	71.33	0.243	0.76
TF-IDF→SVD_Test.N.Na	32.22	63.77	76.18	91.62	0.096	0.91
TF-IDF→SVD_Test.ALL.Na	23.57	47.61	59.10	78.02	0.195	0.81

Table S4.3: **Performance of pBRIT using DisGeNET disease database** TPR, MRR and AUC results for LO-OCV of prioritization on the manually curated dataset from DisGeNet. **A)** When all 10 annotation sources were used across all four analysis scenarios and **B)** when Pathway and Mammalian Phenotype Ontology (MPO) annotation sources were removed for all four analysis scenarios. The reported values reflect the average percentage of training genes ranked in the top X% of test genes over all diseases. Ranks of all prioritized genes, and the corresponding classes are available in S6 File

Analysis Scenario	top1%	top5%	top10%	top30%	MRR	AUC Score
TF-IDF_Test.Pheno.Include	10.29	27.86	41.81	71.78	0.215	0.79
TF-IDF_Test.Pheno.Discard	08.84	21.18	30.49	53.74	0.341	0.66
TF-IDF→SVD_Test.Pheno.Include	11.30	29.82	43.74	73.10	0.205	0.80
TF-IDF→SVD_Test.Pheno.Discard	08.92	22.03	32.05	56.08	0.335	0.67

Table S4.4: **Performance of pBRIT on novel disease genes** TPR, MRR and AUC results of prioritization on novel disease genes, published after construction of pBRIT annotation databases. Ranks of all prioritized genes, and the corresponding HPO classes are available in S7 File.

3.8.8.1 Comparative evaluation of pBRIT

For comparative performance evaluation of pBRIT with Endeavour-v3.71(published in 2016) and Random Walk with Restart on multiplex networks (RWR-M) method, we extracted a random set of 693 HPO terms with 3,037 novel associations (1,111 unique genes) from the set of 2,025 prospective HPO terms of the previous benchmark dataset (see Table S4.4 and supplementary sheet 8 for details). For these three tools their internal database was constructed in or before December 2016. Full genome wide prioritization was done using default settings for all three methods. Due to computational complexity of using all 44 annotation sources of Endeavour-v3.71, only 24 annotation sources were used. The selected ones are at the level of Protein function: Gene Ontology, Uniprot, Text-mining, Interpro, Reactome; Pathways: Reactome, Wikipathways, RGD pathways, BioCarta, ConsensusPathway database, hiPathDB; Phenotype information: GAD, OMIM, RGD MPO, RGD-RDO; Protein-Protein Interaction: String, BioGrid, I2D, IntAct, iRefIndex, Mint, HPRD, MIPS, GeneRIG; Sequence based features: BLAST. The annotation sources were selected to overlap with the internal annotation sources used by pBRIT. For RWR-M, its internal annotation source consist of Pathways, PPI, Co-expression and Disease similarity (from HPO) annotation sources. The detailed list of prioritization results obtained from these three tools can be found in supplementary File S7. Additionally, the ROC curve comparison can be found in Figure S4.2B.

Comparative Tools	top1%	top5%	top10%	top30%	MRR	AUC Score
RWR-M	10.14	24.03	33.61	57.49	0.319	0.68
Endeavour-v3.71	21.00	45.83	58.97	83.27	0.142	0.85
pBRIT	19.88	43.63	59.10	87.32	0.128	0.87

Table S4.5: **Comparative performance on subset of 693 HPO benchmark data set.** TPR, MRR and AUC results of prioritization on novel disease genes associated with 693 HPO classes, published after construction of pBRIT, Endeavour and RWR-M internal annotation databases (see Supplementary File S7, sheet 8). Ranks of all prioritized genes, and the corresponding HPO classes are available in S7 File.

3.8.9 Summary of performances

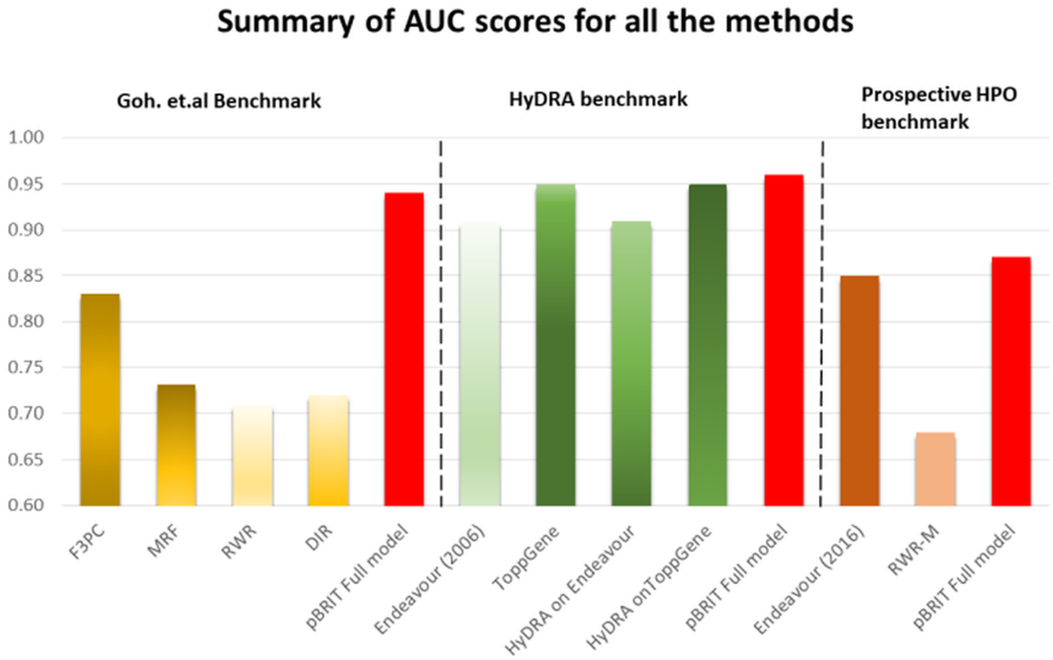


Figure S4.3: **Summary of AUC scores for all the methods** The performance of pBRIT (red-colored bar) in comparison to 9 different competing methods on their own benchmark dataset and prospective HPO benchmark data set. For pBRIT, the full TF-IDF→SVD_N.Na method is used for all LO-OCV analyses against Goh.*et.al.* and HyDRA datasets. For the prospective HPO dataset, we applied the TF-IDF→SVD.Pheno.Include method for genome-wide prioritization.

3.8.10 Negative Control Experiment

As an additional test for the relevancy of the obtained AUC scores, we performed a control experiment on the DisGeNET data, replacing the query gene by a random gene not associated with any given *UMLS* class during LO-OCV. As expected, we achieved an average AUC score of approximately 0.5, corresponding to random guessing, for all analysis scenarios (Supplementary File S6, sheet 7).

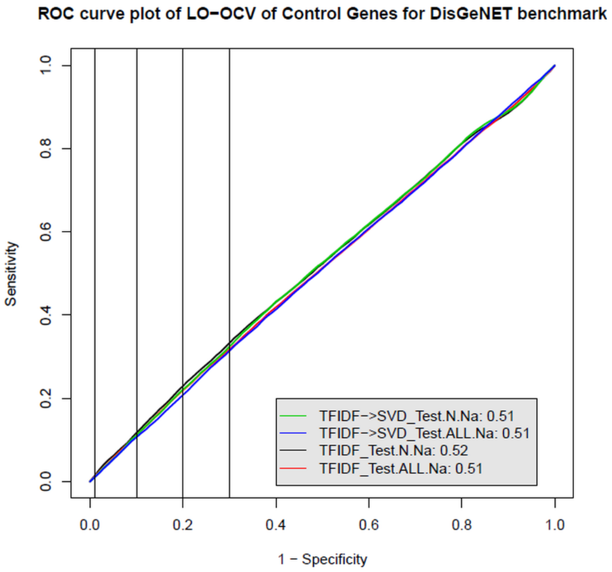


Figure S4.3: **Negative Control Experiment** For the DisGeNET data, the query gene was replaced by a random gene not associated with any given *UMLS* class during LO-OCV.

3.8.11 Effect of changing annotations on prioritization

Studying the effect of changes in annotation sources is a complex task and is limited by two factors, as mentioned in the manuscript. First, the choice of annotation sources must be made such that the effect of changes in annotation features can be easily quantified. Second, the monthly or yearly releases must be available for the annotation source. Therefore, we opted to use GO and HPO in this analysis. Figure S5.1 presents the schematic workflow followed.

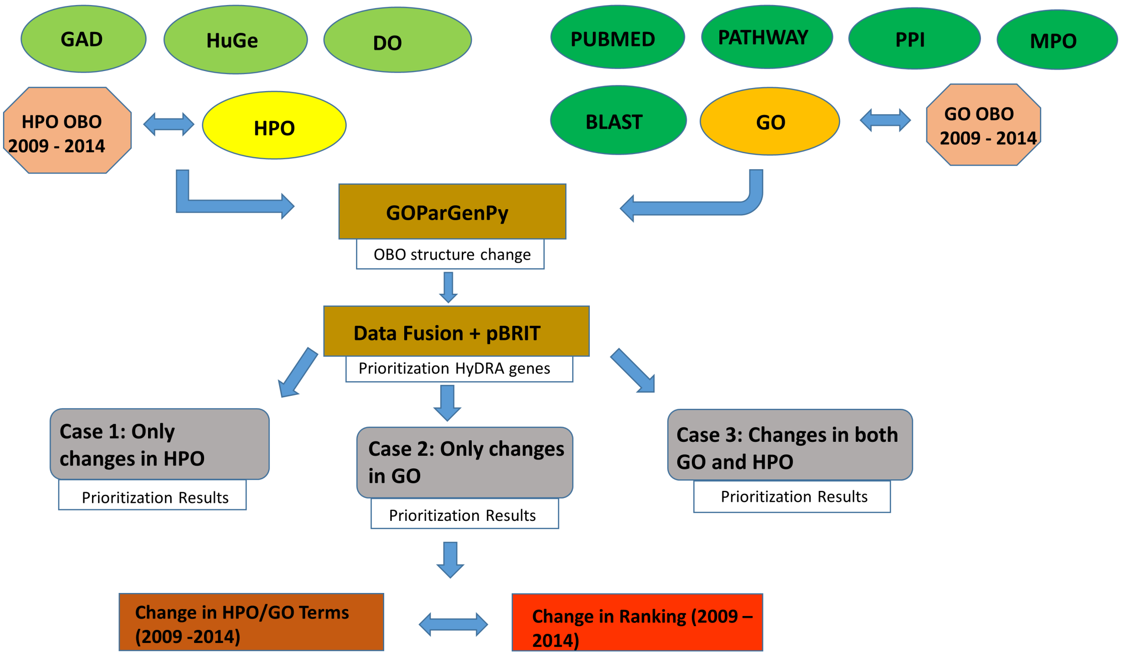


Figure S5.1: **Workflow for studying the effect of changes in GO and HPO based annotation sources on prioritization results:** All the annotation sources are incorporated in the model. Retrospectively, corresponding yearly OBO representations of GO and HPO from 2009 to 2014 are used to construct the annotation model and used for prioritization. The HyDRA based benchmark dataset was used in this analysis. Three analysis case scenarios were devised to evaluate the effect of changes in GO and HPO terms with respect to changes in ranking.

We first downloaded the corresponding OBO structure for HPO and GO ontologies from 2009 2014. For HPO, the OBO structures were retrieved from 01.01.2009 to 01.01.2014) and for GO it was retrieved from 01.04.2009 to 01.04.2014. OBO files hold all the dependencies for each of the ontology classes. Hence, we incorporated the functionality of GOParGenPy to capture this hierarchy and dependencies. From the base annotation file, all parent terms were retrieved for any given gene, for both GO and HPO. This was done for each year from 2009 to 2014. From here, the yearly annotation sources were fused with the remaining annotation sources to create composite matrices (phenotype and functional annotation category) and the full pBRIT model was applied for prioritization.

To measure the effect of changes of annotation sources we formulated three case scenarios:

1. **Case 1**: When we only replace HPO terms using 2009 to 2014 releases and keep the rest of the annotation sources to the current version of 2014
2. **Case 2**: When we only replace GO terms using 2009 to 2014 releases and keep the rest of the annotation sources to the current version of 2014.
3. **Case 3**: When we replace both HPO and GO terms using 2009 to 2014 releases and keep the rest of the annotation sources to the current version of 2014.

For each of these three cases, we used the HyDRA based benchmark data consisting of 8 disease classes and 250 gene. The prioritization results can be found in Supplementary File (S4 for Case 1, S5 for Case 2 and S6 for Case 3).

Analysis steps include:

1. For each of the given genes, GOParGenPy produces tab separated output as:
    (a) gene_id base ontology terms parent ontology terms
2. Between subsequent years we computed the difference in associated ontology terms (GO or HPO) for the given gene.
    (a) **Ontology set**: Let $O_g(i)$ be the set of all ontology terms (direct and parents) associated to a gene g for a given year $i\epsilon\{2009..2014\}$.
    (b) **Computing difference in ontology set**: Let $N_{O_g}(i, i+1) = n(O_g(i) - O_g(i+1))$ for $i\epsilon\{2009..2013\}$. Hence, $N_{O_g}(i, i+1) = n(O_g(i) \cup O_g(i+1) - O_g(i) \cap O_g(i+1))$ gives the unique difference in number of ontology terms between subsequent years for a given gene. Therefore, $N_{Sum_{O_g}} = \sum_j N_{O_g}(j)$ where $j\epsilon\{09-10, 10-11, 11-12, 12-13, 13-14\}$ gives the total count of unique changes of ontology terms between subsequent years.
    Similarly, $N_{Avg_{O_g}} = \frac{\sum_j N_{O_g}(j)}{6}$ for $j\epsilon\{2009, ..2014\}$ gives the average number of associated ontology terms for a given gene. Finally, we compute $N_{Norm_{O_g}} =$

$\frac{N_{sum_{O_g}}}{N_{Avg_{O_g}}}$ which gives the normalized count of unique changes according to the ontology for a given gene.

3. **Rank difference**: Compute the difference in ranking across 2009 to 2014 after prioritization. Let $R_g(i)$ be the rank of query gene in the prioritization results for given year $i\epsilon\{2009..2014\}$. Then $N_{Rank_g}(i, i+1) = R_g(i) - R_g(i+1)$ gives the difference in the rank for a given gene for subsequent years. Therefore, $N_{RankSum_g} = \sum_j N_{Rank_g}(j)$ for $j\epsilon$ {09-10,10-11,11-12, 12-13, 13-14} gives the summed rank of subsequent years for a given gene.

4. **Regression**: Finally, under each of the three cases, a linear regression model was used to explain the relationship between variables $N_{Norm_{O_g}}$ and $N_{RankSum_g}$ across all analysis scenarios of pBRIT.

Ranking results of 250 genes from 8 disease classes of the HyDRA based benchmark dataset are summarized in Supplementary file S3, S4 and S5, showing a variance of < 0.0002 on the overall AUC scores over the included time-frame. Figures S5.2.1-S5.2.12 present the results explaining the relationship between changes in the number of associated ontology terms and changes in gene ranks computed for 2009 to 2014. Clearly, it can be seen that there is almost no correlation between these two variables in any of the pBRIT analysis scenarios.

Looking at individual genes however, we can notice some differences in behavior. Most genes showed substantial annotation changes over the included time-span, while the respective prioritization ranks remained stable. *TP53* for example, associated with Lymphoma and Colorectal cancer, had an average of 444 GO terms and 87 HPO terms associated to it between 2009-2014, while 334 GO terms (75%) and 99 HPO terms (>113%) (file S3, S4) were altered. This could have potentially affected *TP53* ranking. However, the ranks provided by the full pBRIT model remained stable with minimal variance over the included disease classes. As mentioned above, this observation can be generalized, as no significant correlation between annotation changes and ranking was identified.

However, a few genes presented highly variable ranking between ontology releases, while the number of actually changed terms was limited. For example, *NAT2*, associated with Endometriosis, has the highest variance in ranking among all tested genes using the full model, mainly attributed to a 0.39 difference in rank ratio (see TFIDF→SVD_N.Na results in S3, S4 file) between 2013 and 2014. Nevertheless, only one GO term (GO:0044763; single-organism cellular process) changed between 2013(present)-2014(absent) and none of the HPO terms changed. Since this particular GO term is near to the root term (GO:0008150;biological process) in the hierarchy, it corresponds to a very generic term. Consequently, it will be associated to many genes, resulting in a low IDF score. Hence, this improved ranking can either be due to loss of this generic GO term, thereby improving the proximity scores, or due to changes in the ontology-terms related to the remaining training genes.

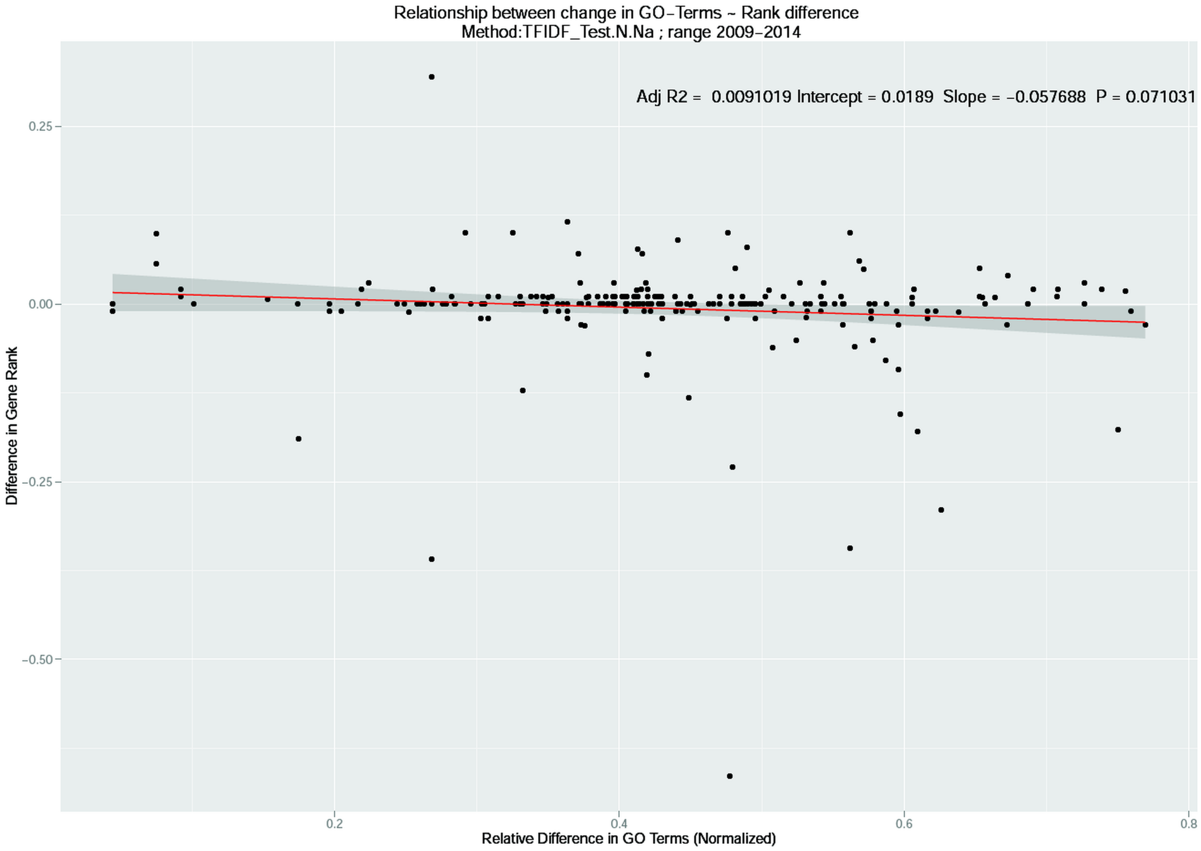


Figure S5.2.1 : **Stratification of changes in the number of GO Terms and its effect in prioritization:** Correlation plot of absolute change in the number of GO Terms across 2009-2014 releases of GO for genes and their respective ranks using the HyDRA based benchmark dataset for TF-IDF_Test.N.Na analysis scenario.

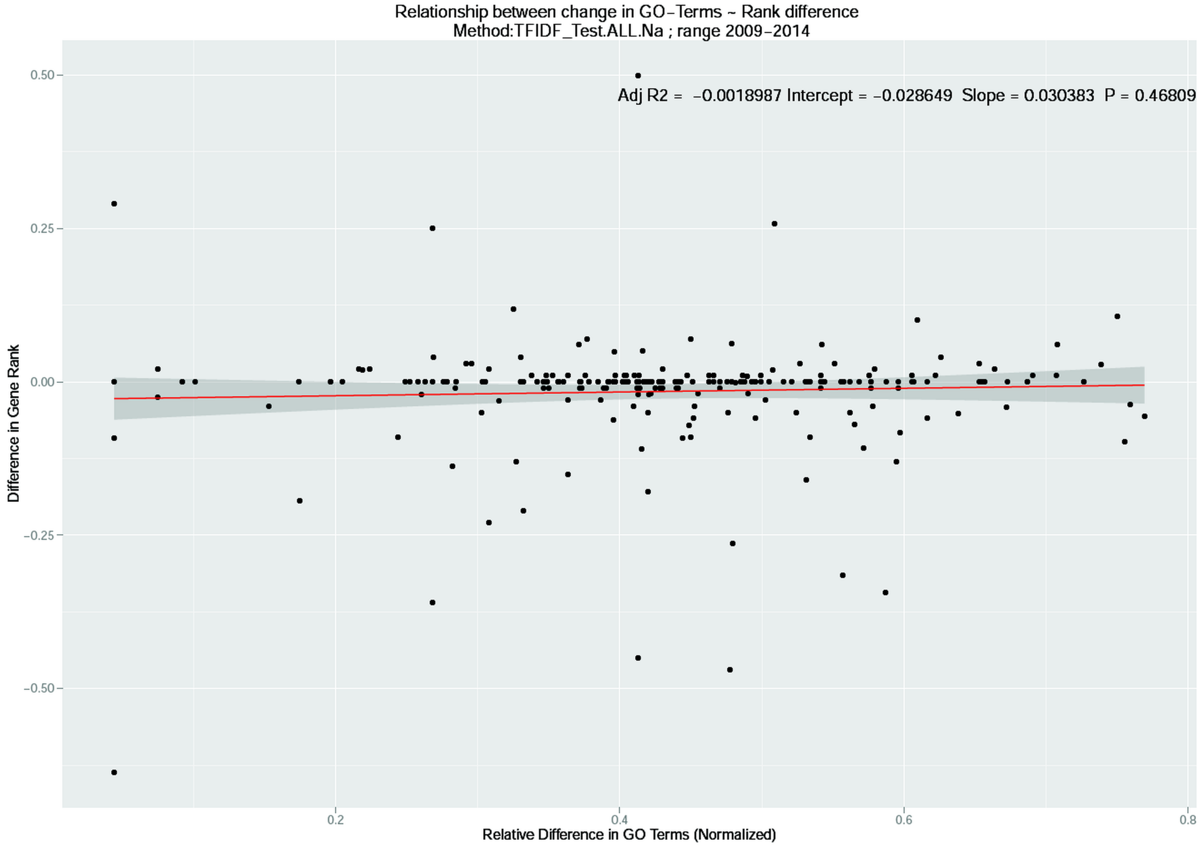


Figure S5.2.2: **Stratification of changes in the number of GO Terms and its effect in prioritization:** Correlation plot of absolute change in the number of GO Terms across 2009-2014 releases of GO for genes and their respective ranks using the HyDRA based benchmark dataset for TF-IDF_Test.ALL.Na analysis scenario.

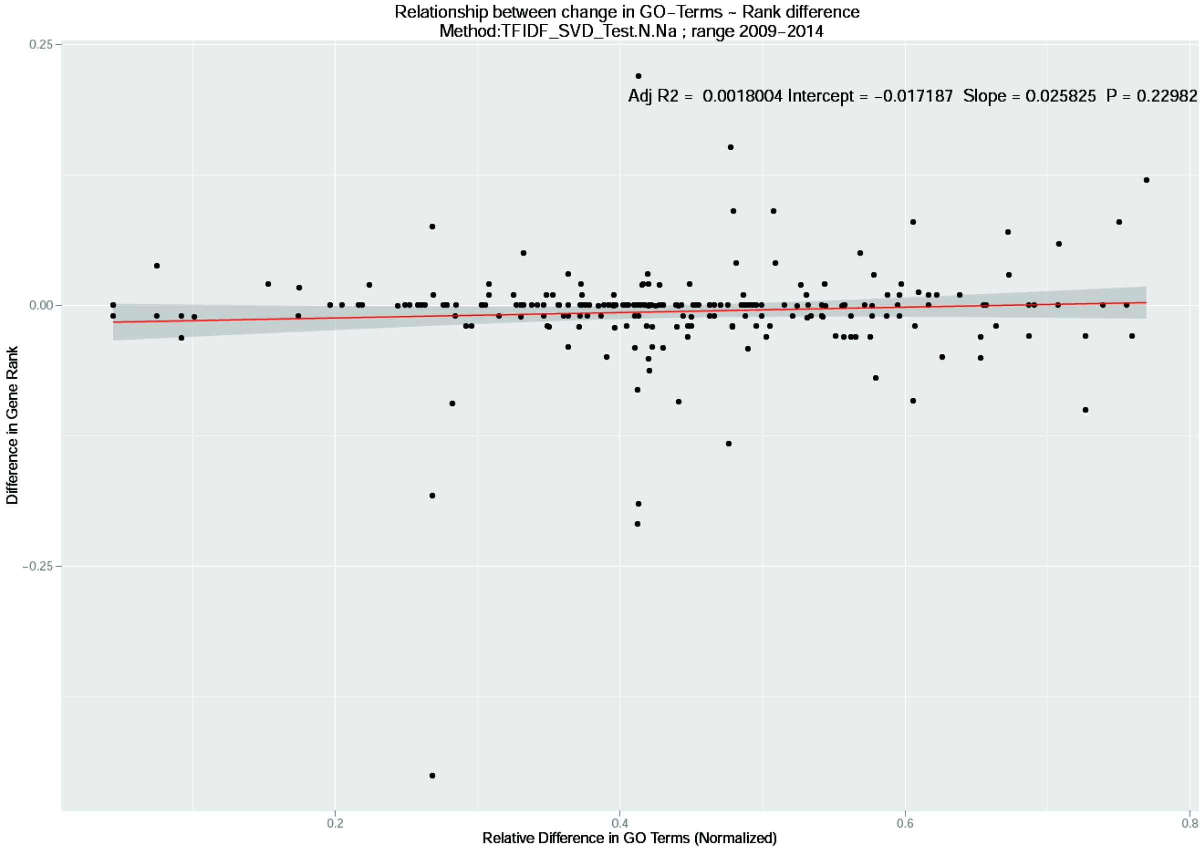


Figure S5.2.3: **Stratification of changes in the number of GO Terms and its effect in prioritization:** Correlation plot of absolute change in the number of GO Terms across 2009-2014 releases of GO for genes and their respective ranks using the HyDRA based benchmark dataset for TF-IDF→SVD_Test.N.Na analysis scenario.

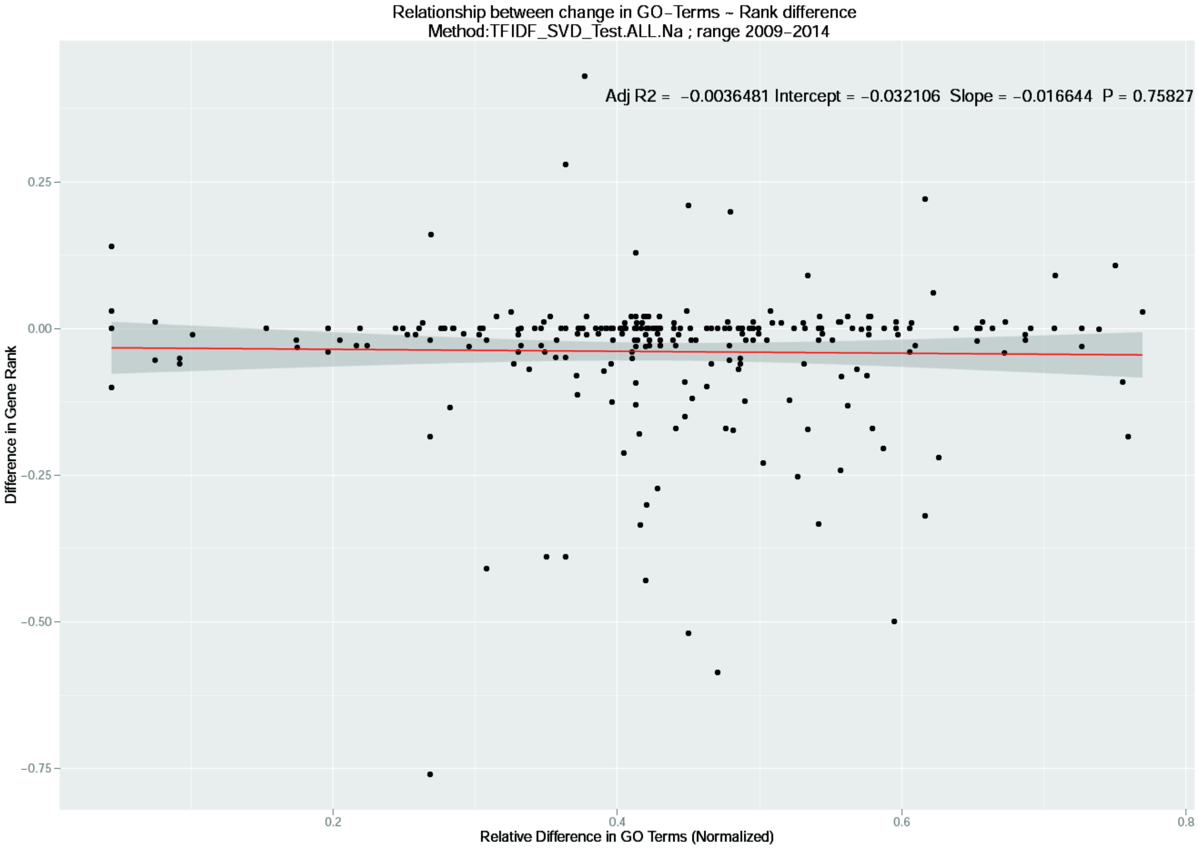


Figure S5.2.4: **Stratification of changes in the number of GO Terms and its effect in prioritization:** Correlation plot of absolute change in the number of GO Terms across 2009-2014 releases of GO for genes and their respective ranks using the HyDRA based benchmark dataset for TF-IDF→SVD_Test.ALL.Na analysis scenario.

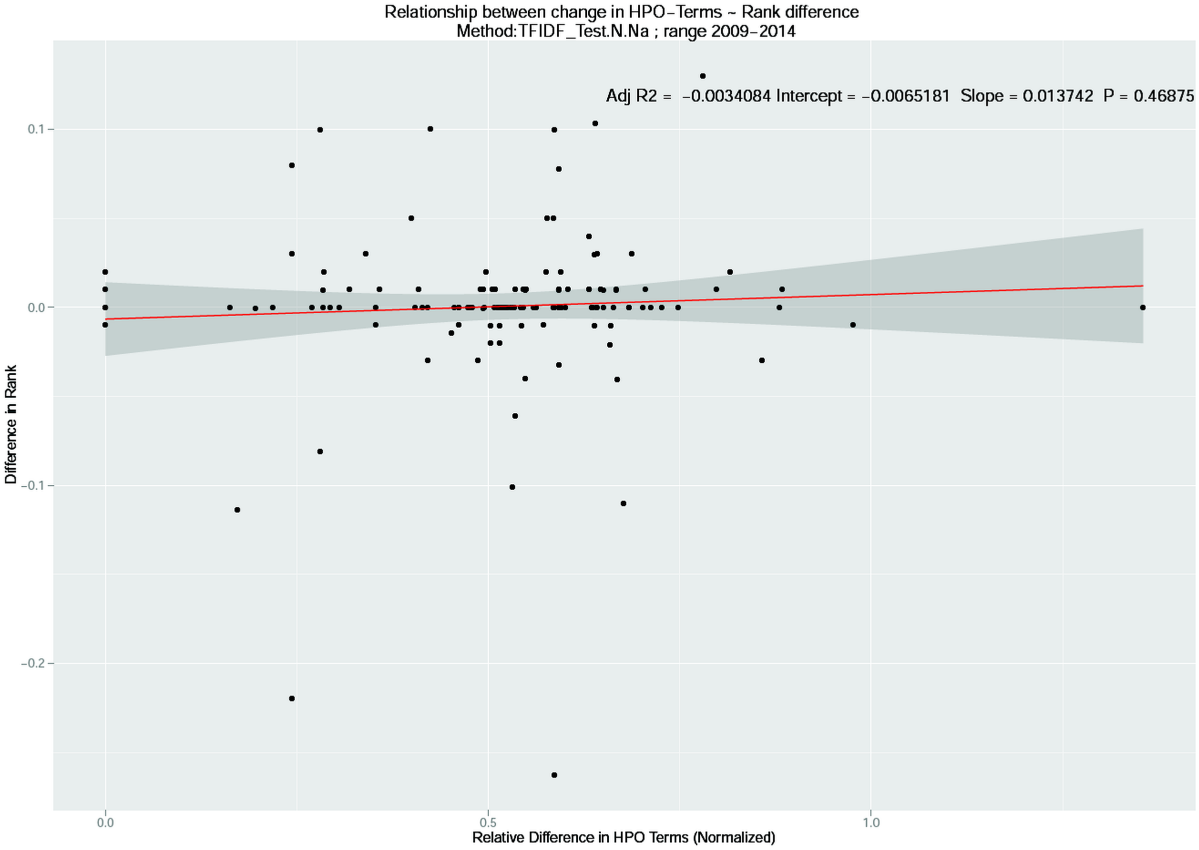


Figure S5.2.5: **Stratification of changes in the number of HPO Terms and its effect in prioritization:** Correlation plot of absolute change in the number of HPO Terms across 2009-2014 releases of HPO for genes and their respective ranks using the HyDRA based benchmark dataset for TF-IDF_Test.N.Na analysis scenario.

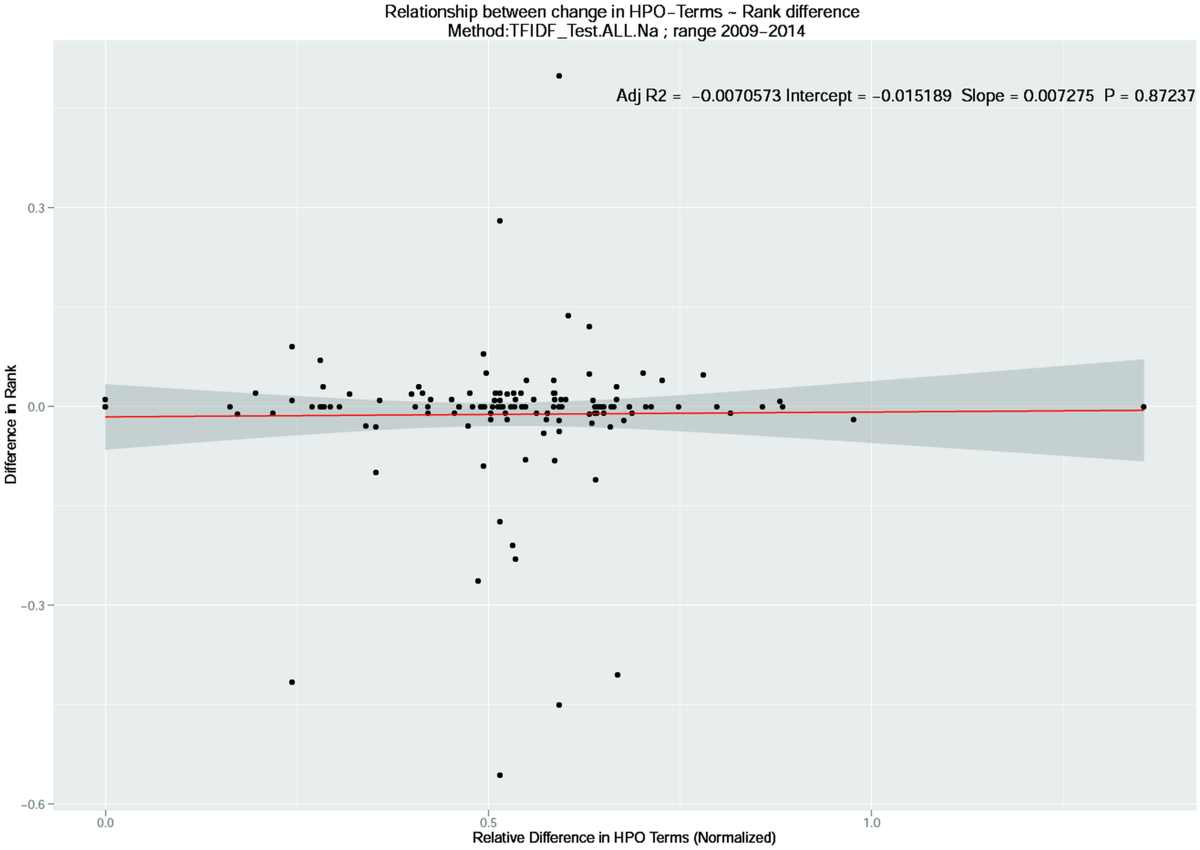


Figure S5.2.6: **Stratification of changes in the number of HPO Terms and its effect in prioritization:** Correlation plot of absolute change in the number of HPO Terms across 2009-2014 releases of HPO for genes and their respective ranks using the HyDRA based benchmark dataset for TF-IDF_Test.ALL.Na analysis scenario.

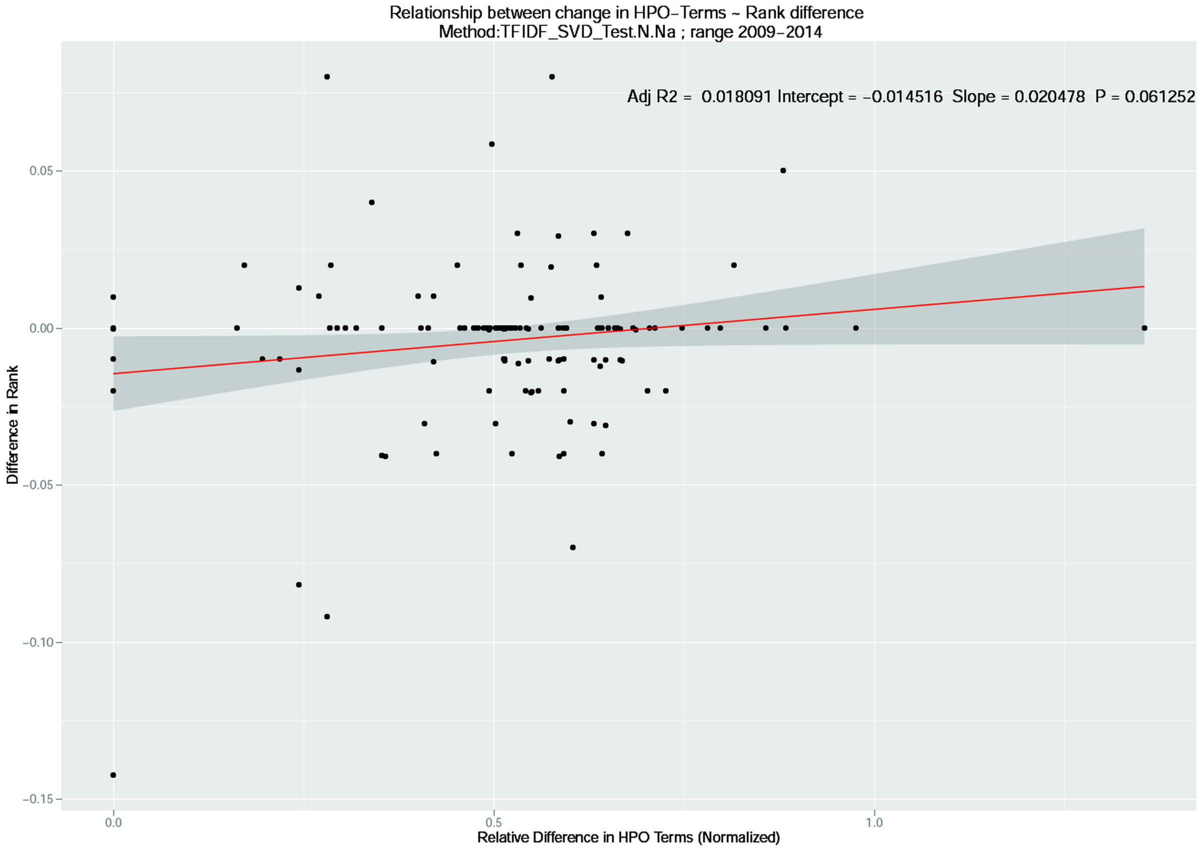


Figure S5.2.7: **Stratification of changes in the number of HPO Terms and its effect in prioritization:** Correlation plot of absolute change in the number of HPO Terms across 2009-2014 releases of HPO for genes and their respective ranks using the HyDRA based benchmark dataset for TF-IDF→SVD_Test.N.Na analysis scenario.

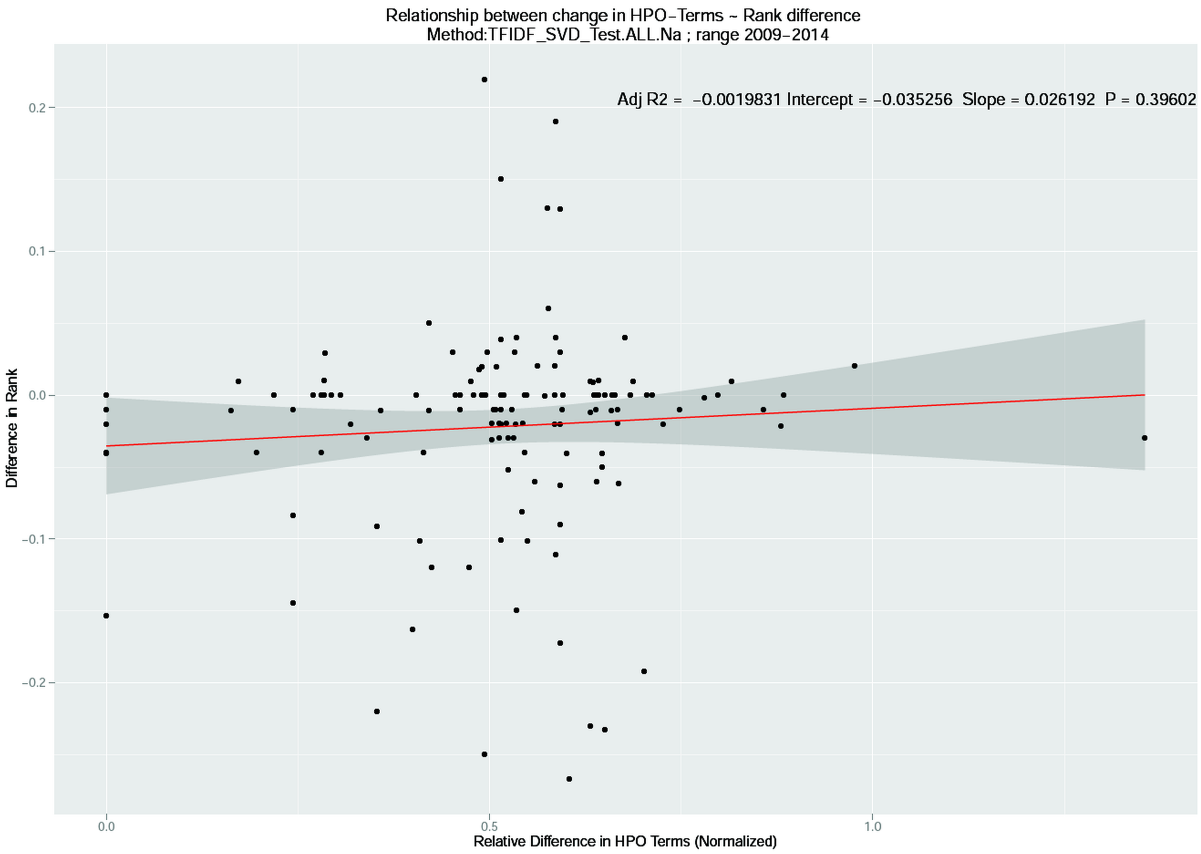


Figure S5.2.8: **Stratification of changes in the number of HPO Terms and its effect in prioritization:** Correlation plot of absolute change in the number of HPO Terms across 2009-2014 releases of HPO for genes and their respective ranks using the HyDRA based benchmark dataset for TF-IDF→SVD_Test.ALL.Na analysis scenario.

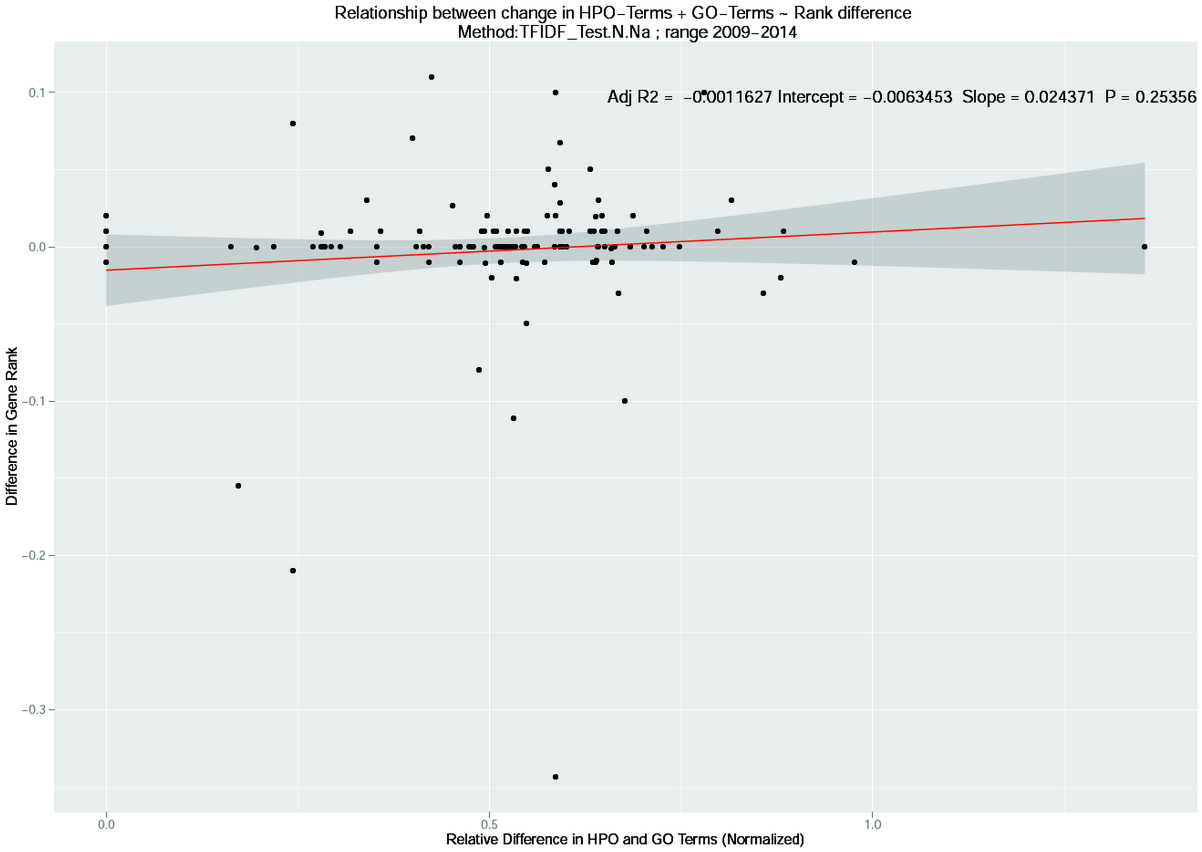


Figure S5.2.9: **Stratification of changes in the combined number of HPO and GO Terms and its effect in prioritization:** Correlation plot of absolute change in the number of combined HPO and GO Terms across 2009-2014 releases for genes and their respective ranks using the HyDRA based benchmark dataset for TF-IDF_Test.N.Na analysis scenario.

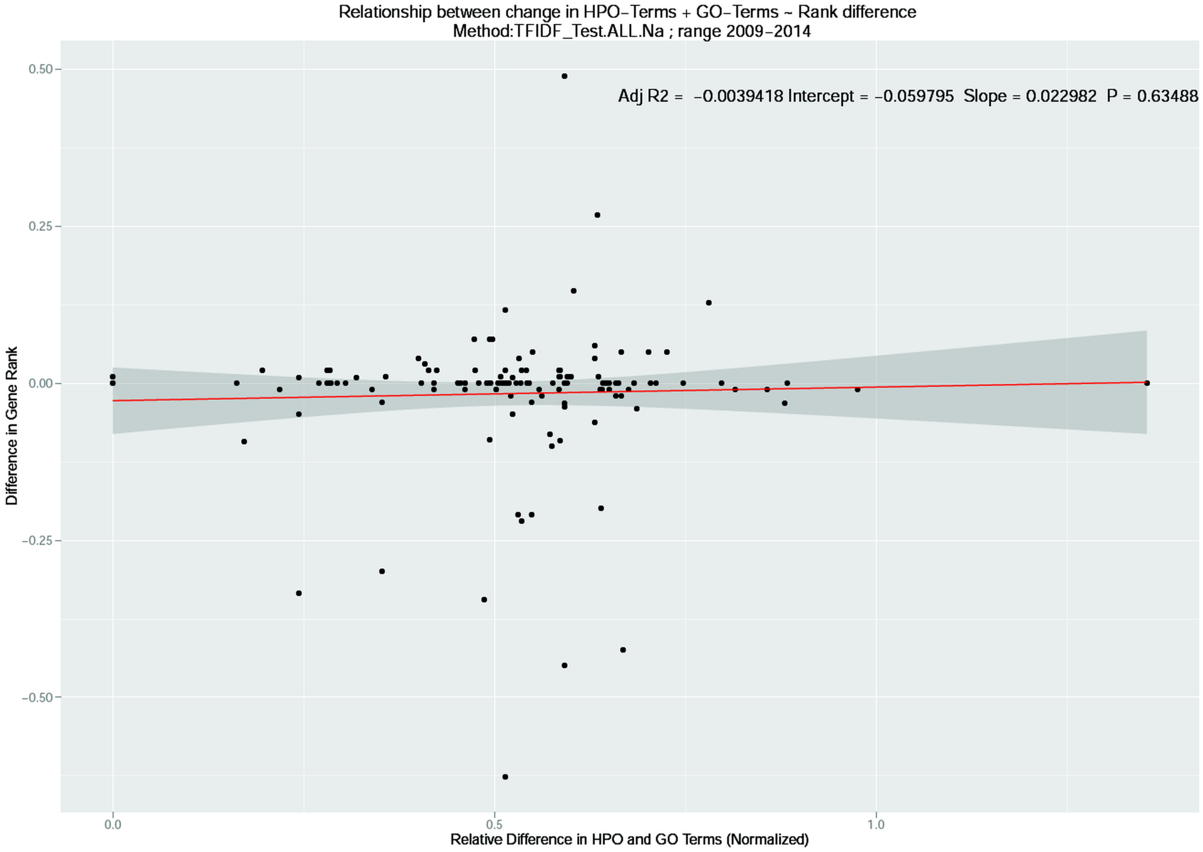


Figure S5.2.10 : **Stratification of changes in the combined number of HPO and GO Terms and its effect in prioritization:** Correlation plot of absolute change in the number of combined HPO and GO Terms across 2009-2014 releases for genes and their respective ranks using the HyDRA based benchmark dataset for TF-IDF_Test.ALL.Na analysis scenario.

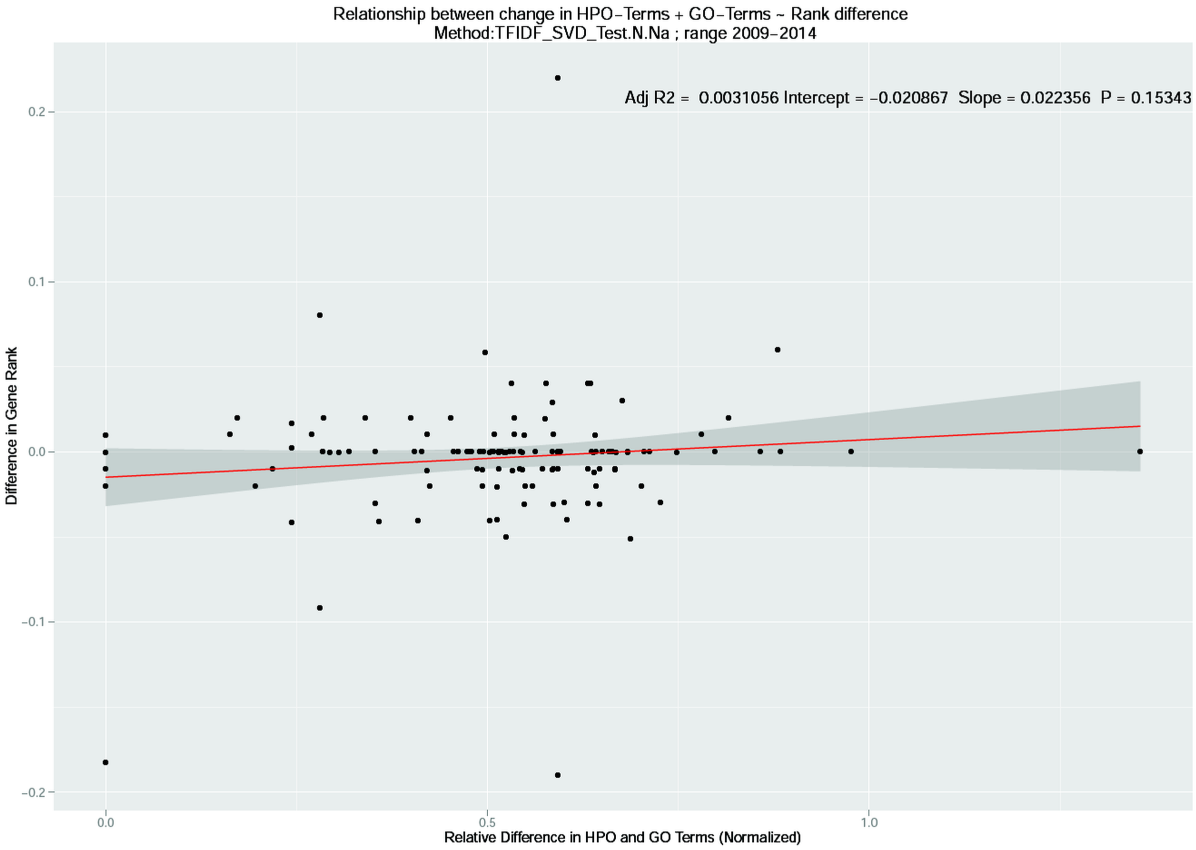


Figure S5.2.11: **Stratification of changes in the combined number of HPO and GO Terms and its effect in prioritization:** Correlation plot of absolute change in the number of combined HPO and GO Terms across 2009-2014 releases for genes and their respective ranks using the HyDRA based benchmark dataset for TF-IDF→SVD_Test.N.Na analysis scenario.

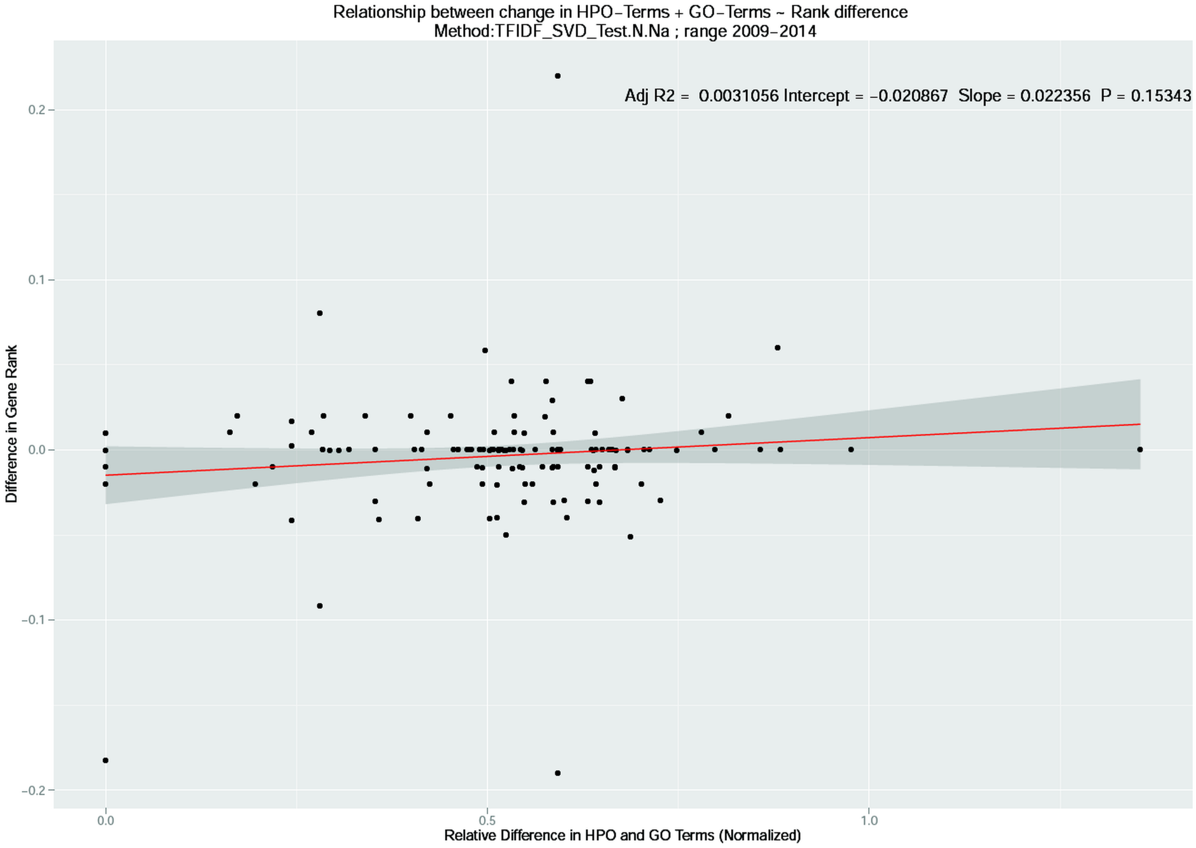


Figure S5.2.12 : **Stratification of changes in the combined number of HPO and GO Terms and its effect in prioritization:** Correlation plot of absolute change in the number of combined HPO and GO Terms across 2009-2014 releases for genes and their respective ranks using the HyDRA based benchmark dataset for TF-IDF→SVD_Test.ALL.Na analysis scenario.

3.8.12 Effect of number of training genes on prioritization

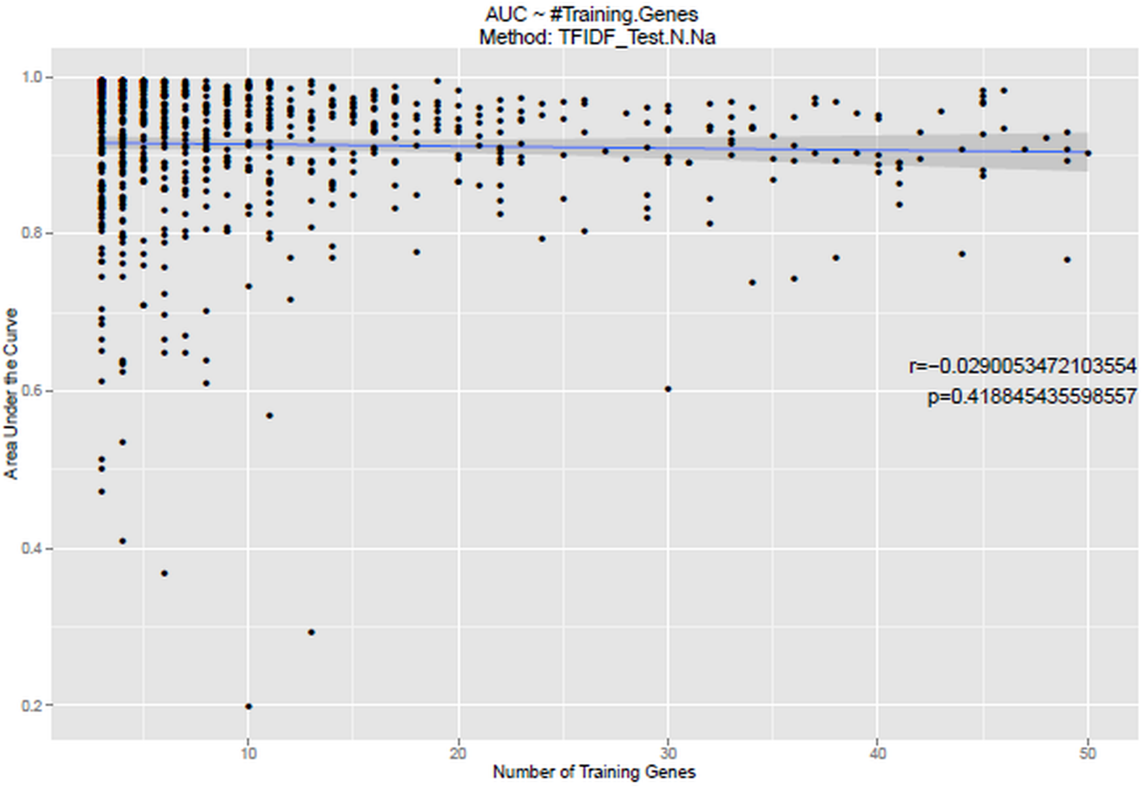


Figure S6.1: Correlation plot between the number of training genes and AUC scores under the TF-IDF_Test.N.Na scenario during DisGenet based benchmarking.

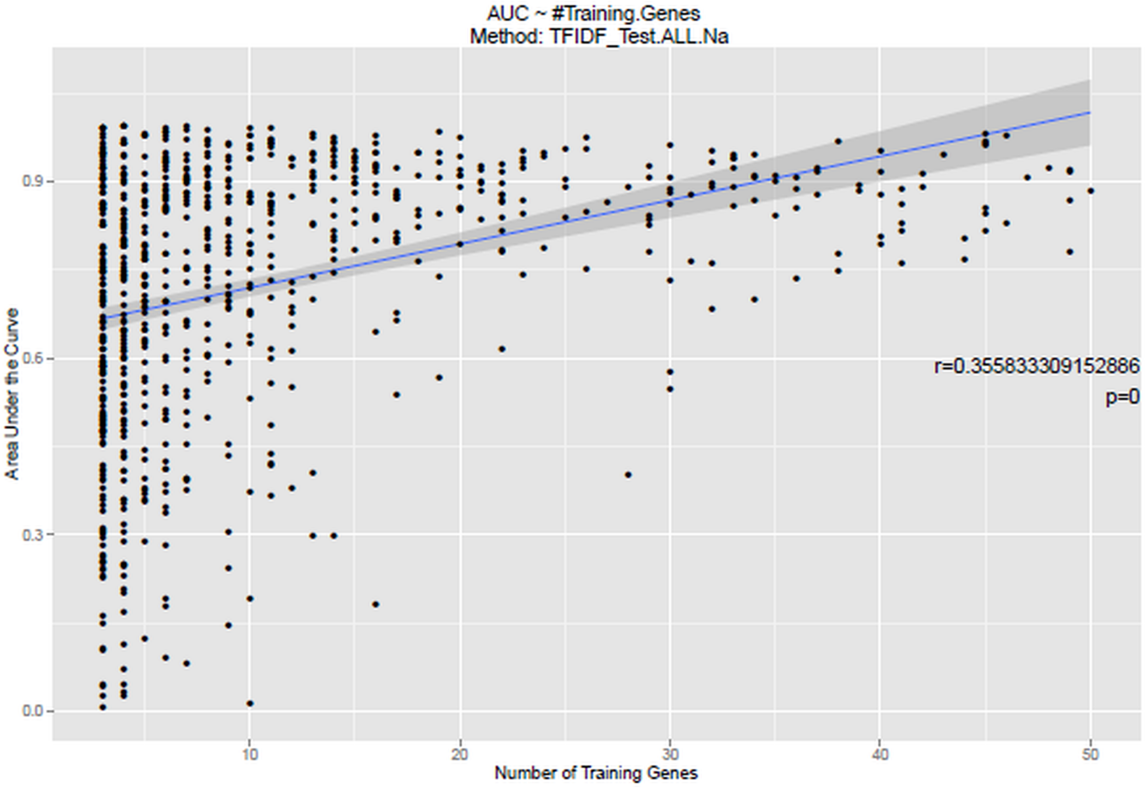


Figure S6.2: Correlation plot between the number of training genes and AUC scores under the TF-IDF_Test.ALL.Na scenario during DisGenet based benchmarking.

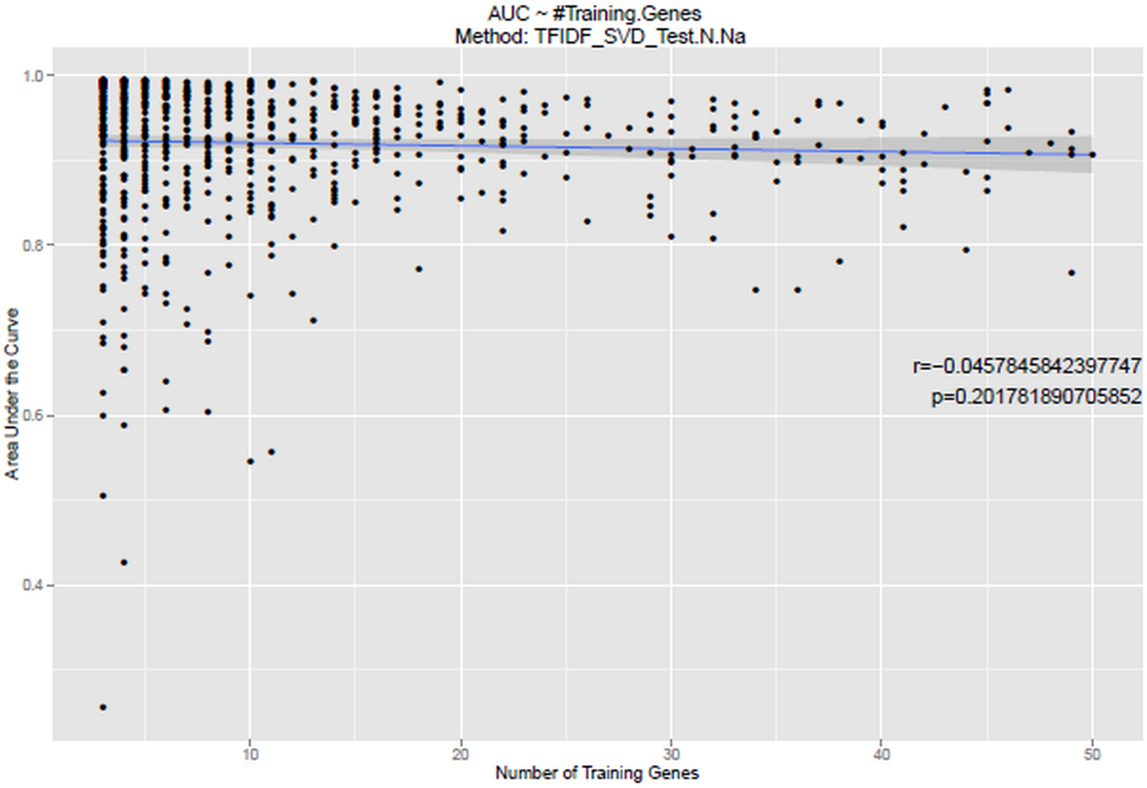


Figure S6.3: Correlation plot between the number of training genes and AUC scores under the TF-IDF→SVD_Test.N.Na scenario during DisGenet based benchmarking.

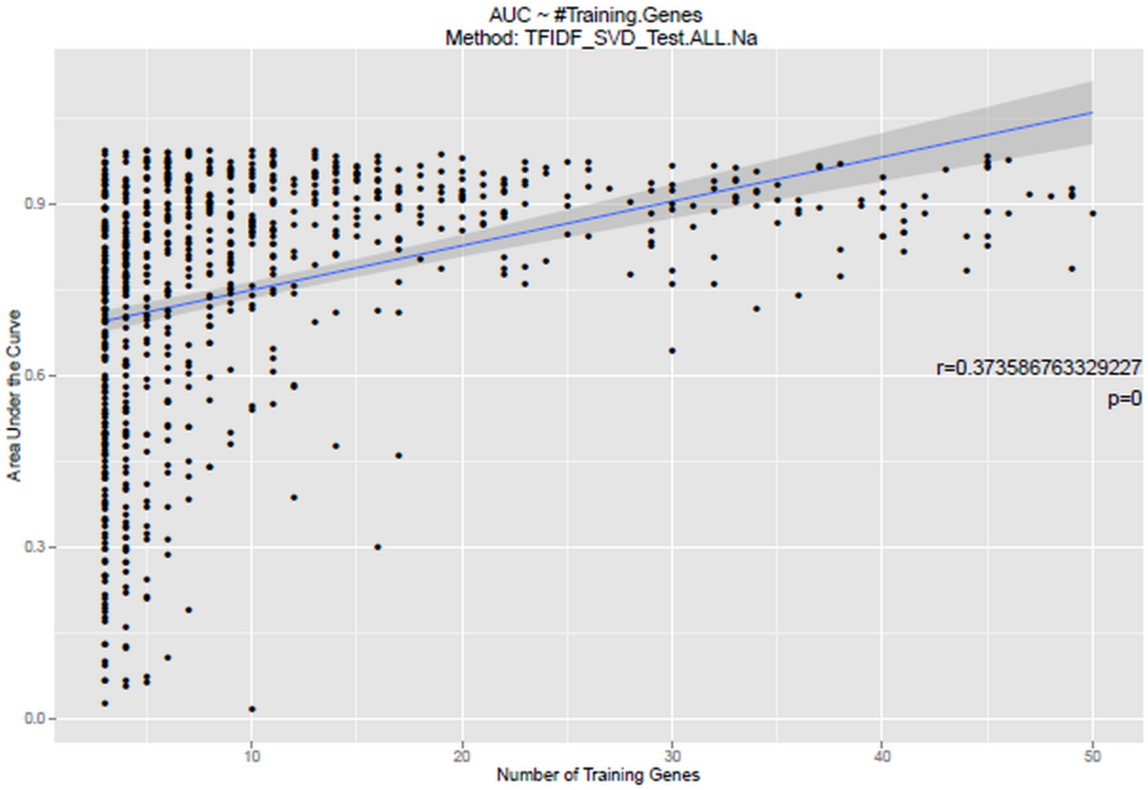


Figure S6.4: Correlation plot between the number of training genes and AUC scores under the TF-IDF→SVD_Test.ALL.Na scenario during DisGenet based benchmarking.

3.8.13 Visualization of gene prioritization through pBRIT

Analysis details: We present an example data set to explain how pBRIT offers exploration of the prioritization results. pBRIT is based on Bayesian regression, and as such the training genes (from the functional annotation) are the independent variables to predict the phenotypic concordance score of the input test genes.

Using heatmaps (see figure S7.1) we can examine which of the training genes are responsible for predicting the phenotpyic concordance score of the test genes.For example: In February 2015, it was published that *de novo* loss or gain of function mutations in *KCNA2* gene cause epileptic encephalopathy [198]. No information of disease association for this gene is present in our internal pBRIT model, since all the annotation source have been constructed before January 2015.

Hence, the goal for pBRIT would be to prioritize this gene for possible disease association. pBRIT requires a set of training genes and a choice of prioritization methodology (TFIDF→SVD and TFIDF, with the option of Pheno.Include and

Pheno.Discard).

The training genes were retrieved manually using keywords present in related abstracts and existing knowledge of epileptic encephalopathy. Keywords present in the abstracts included: Ohtahara syndrome, Epeleptic encephalopathies, intellectual disabilities, neurodevelopmental features.

Training genes:
STXBP1,ARX,SCN2A,PLCB1,NECAP1,SCN8A,GNAO1,KCNQ2,SPTAN1,PNKP, SLC13A5,CDKL5WWOX,HCN1,KCNB1,CASK, PIGQ
Test gene: 99 genes were selected randomly across the genome such that none of the training genes and query gene are present in this list.

The top heatmap explains gene-by-gene similarities between both the training (red) and test (black) genes to the training genes. Darker shades indicate a larger contribution to the prioritization. *KCNA2* (marked in green) is top ranked mainly because of a higher similarity to *KCNB1*,*HCN1*, *KCNQ2* and *SCN2A*. Despite direct evidence in the literature of disease association for *NECAP1*, functional similarities to *KCNA2* are negligible. It can be seen that SVD transformation of the gene-by-feature matrices results in visibly more pronounced similarity scores. Second, pBRIT provides test-to-training similarities according to individual annotation sources through heatmap plots (bottom plot). These gene-specific plots highlight the most contributing training genes and annotation sources to the ranking of that particular gene. Again, it can be seen that SVD provides more pronounced similarity profiles. For the *KCNA2* example, additional contributions show up for Pubmed and PPI, together with more pronounced scores for other annotations.

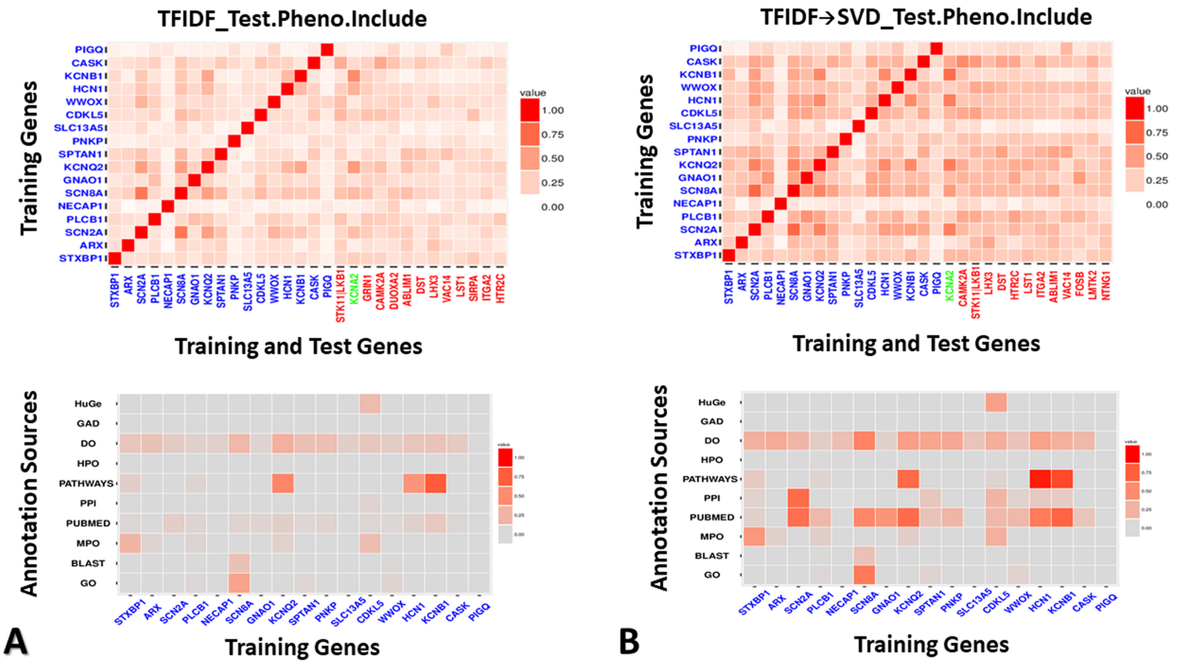
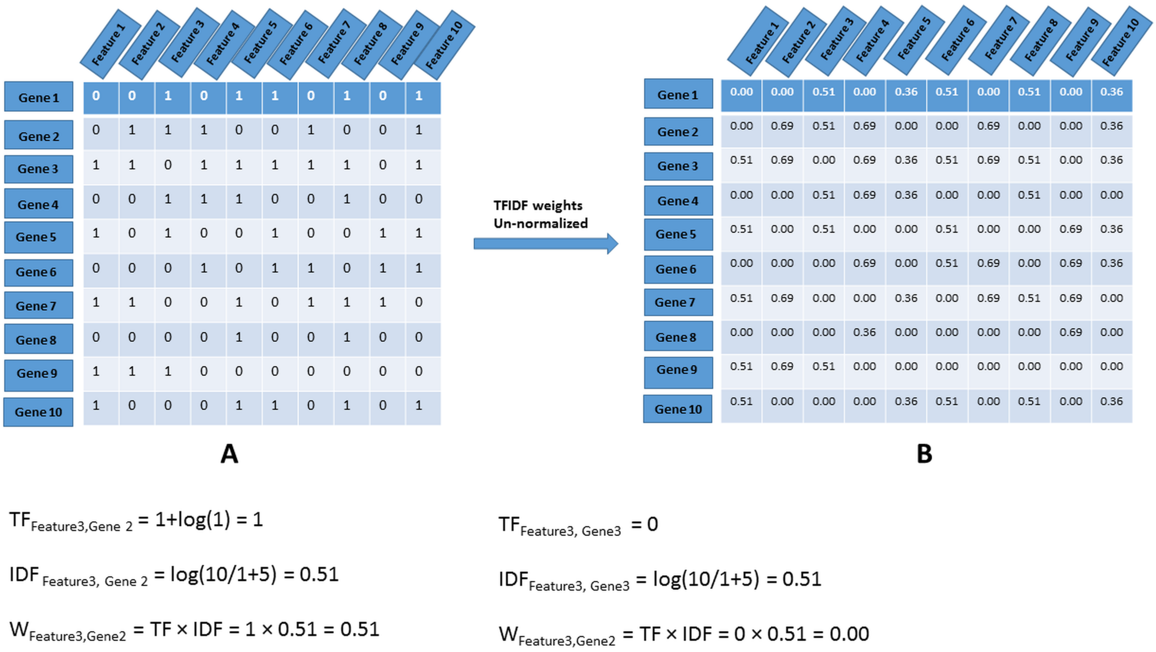


Figure S7.1: **Exploring prioritization results for *KCNA2* gene**. Feature mining methodology (a) TF-IDF_Test.Pheno.Include. (b) TF-IDF→SVD_Test.Pheno.Include. The top image describes the functional annotation matrix X, illustrating contribution of individual training genes (red) during regression. The example gene to be prioritized (*KCNA2*) is marked in green. The bottom image indicates contribution to the ranking of *KCNA2* of distinct annotation sources for each training gene.

3.8.14 Toy example explaining how pBRIT works

Computing TF-IDF weights for binary value represented annotation matrices. Matrix A (Figure S8.1) is an example representation of binary valued annotation source whose rows are the 10 genes annotated with 10 features represented by columns. 0 and 1 indicate presence and absence of annotation features. TF-IDF weights are computed based on equation 1 of the manuscript. For example for Gene2 annotated with Feature 3 has term frequency (TF) of 1. The inverse document frequency (IDF) is computed as logarithmic ratio of how many times the Feature 3 is frequently annotated across all 10 genes. Hence, the corresponding IDF value of Feature 3 for the given Gene 2 is computed as 0.51. Together, the TF-IDF weight is equal to 0.51. Similarly, for the all the annotation features their corresponding TF-IDF weights are computed and are shown by matrix B.



A

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10
Gene 1	0	0	1	0	1	1	0	1	0	1
Gene 2	0	1	1	1	0	0	1	0	0	1
Gene 3	1	1	0	1	1	1	1	1	0	1
Gene 4	0	0	1	1	1	0	0	1	0	0
Gene 5	1	0	1	0	0	1	0	0	1	1
Gene 6	0	0	0	1	0	1	1	0	1	1
Gene 7	1	1	0	0	1	0	1	1	1	0
Gene 8	0	0	0	0	1	0	0	1	0	0
Gene 9	1	1	1	0	0	0	0	0	0	0
Gene 10	1	0	0	0	1	1	0	1	0	1

TFIDF weights
Un-normalized

B

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10
Gene 1	0.00	0.00	0.51	0.00	0.36	0.51	0.00	0.51	0.00	0.36
Gene 2	0.00	0.69	0.51	0.69	0.00	0.00	0.69	0.00	0.00	0.36
Gene 3	0.51	0.69	0.00	0.69	0.36	0.51	0.69	0.51	0.00	0.36
Gene 4	0.00	0.00	0.51	0.69	0.36	0.00	0.00	0.51	0.00	0.00
Gene 5	0.51	0.00	0.51	0.00	0.00	0.51	0.00	0.00	0.69	0.36
Gene 6	0.00	0.00	0.00	0.69	0.00	0.51	0.69	0.00	0.69	0.36
Gene 7	0.51	0.69	0.00	0.00	0.36	0.00	0.69	0.51	0.69	0.00
Gene 8	0.00	0.00	0.00	0.36	0.00	0.00	0.00	0.00	0.69	0.00
Gene 9	0.51	0.69	0.51	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Gene 10	0.51	0.00	0.00	0.00	0.36	0.51	0.00	0.51	0.00	0.36

$TF_{Feature3,Gene\ 2} = 1+\log(1) = 1$

$IDF_{Feature3,\ Gene\ 2} = \log(10/1+5) = 0.51$

$W_{Feature3,Gene2} = TF \times IDF = 1 \times 0.51 = 0.51$

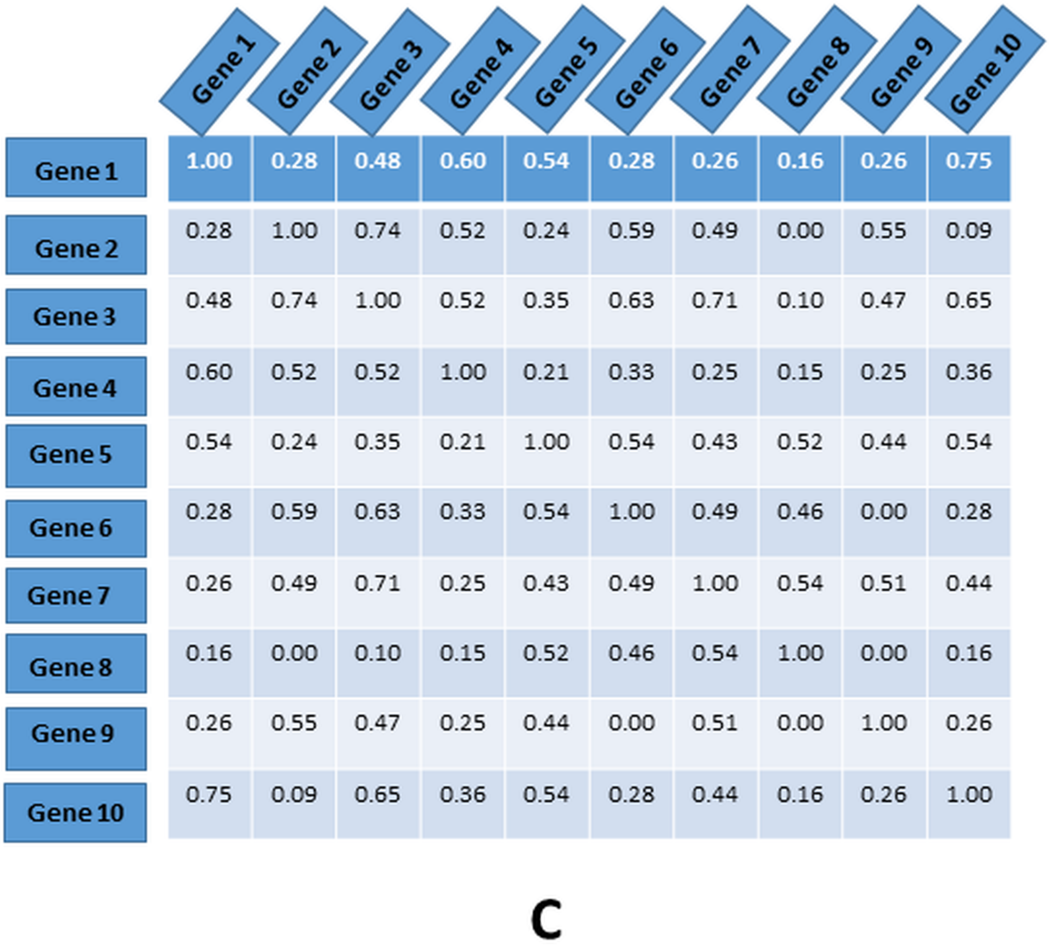
$TF_{Feature3,\ Gene3} = 0$

$IDF_{Feature3,\ Gene3} = \log(10/1+5) = 0.51$

$W_{Feature3,Gene2} = TF \times IDF = 0 \times 0.51 = 0.00$

Figure S8.1: Computation of TF-IDF weights of an example binary valued annotation matrix

Computing gene-by-gene proximity profile matrix. After computing TF-IDF matrix, a unit vector normalization was applied and then *cosine* or dot product was taken to compute gene-by-gene proximity profiles. Figure S8.2 shows the corresponding gene-by-gene similarity matrix (matrix C) for the TF-IDF weighted matrix B. The similarity score ranges between 0 and 1 representing dissimilarity and similarity between genes respectively.



	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7	Gene 8	Gene 9	Gene 10
Gene 1	1.00	0.28	0.48	0.60	0.54	0.28	0.26	0.16	0.26	0.75
Gene 2	0.28	1.00	0.74	0.52	0.24	0.59	0.49	0.00	0.55	0.09
Gene 3	0.48	0.74	1.00	0.52	0.35	0.63	0.71	0.10	0.47	0.65
Gene 4	0.60	0.52	0.52	1.00	0.21	0.33	0.25	0.15	0.25	0.36
Gene 5	0.54	0.24	0.35	0.21	1.00	0.54	0.43	0.52	0.44	0.54
Gene 6	0.28	0.59	0.63	0.33	0.54	1.00	0.49	0.46	0.00	0.28
Gene 7	0.26	0.49	0.71	0.25	0.43	0.49	1.00	0.54	0.51	0.44
Gene 8	0.16	0.00	0.10	0.15	0.52	0.46	0.54	1.00	0.00	0.16
Gene 9	0.26	0.55	0.47	0.25	0.44	0.00	0.51	0.00	1.00	0.26
Gene 10	0.75	0.09	0.65	0.36	0.54	0.28	0.44	0.16	0.26	1.00

Figure S8.2: Computation of gene-by-gene proximity profiles.

Computing Latent Semantic similarity using SVD. Latent semantic modelling is basically computing singular value decomposition to find the optimal k components of the data. Using equation 2 in the manuscript, the SVD is directly applied to TF-IDF weighted annotation matrices. Here, we present a notional representation of the approach (Figure S8.3) instead of the toy example because originally the SVD was done on very large annotation matrices as presented in Table S1.

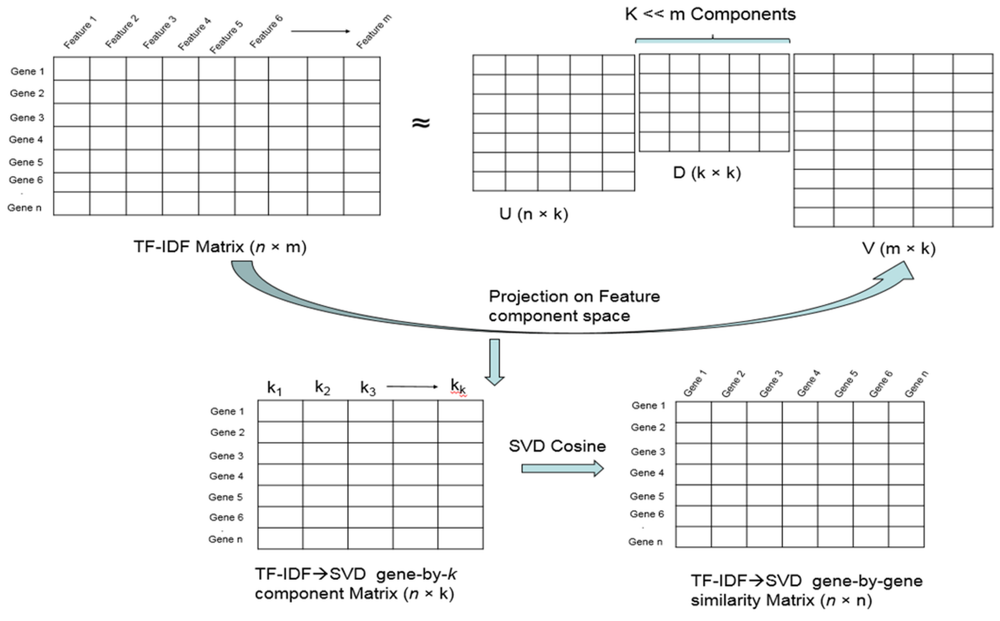


Figure S8.3: Computation of gene-by-gene proximity profiles.

Computing composite matrices Gene-by-gene proximity profiles are obtained for each of the 9 annotation matrices under TFIDF and TFIDF→SVD based feature extraction methodology and categorized as either phenotype or functional annotation sources. The gene-by-gene proximity profile matrix corresponding to the sequence similarity based annotation source (BLAST) is categorized as a functional annotation. For each of these categories, the gene-by-gene similarity matrices are fused together using equation 3 of the manuscript thereby yielding composite matrices.

Prioritization using Bayesian Ridge regression

To demonstrate how the Bayesian ridge regression works in the overall work flow a toy example is demonstrated in Figure S8.4. Let a set of genes: Gene1, Gene2. Gene3, Gene4 constitute the training genes and let genes: Gene 5,Gene 6...Gene N be a set of test genes whose ranking needs to be determined.

For the given training genes (Gene 1, Gene 2, Gene 3 and Gene 4) and test genes (Gene 5, Gene 6 ... Gene N), the proximity scores with regard to used training genes are obtained from the composite matrices corresponding to the phenotypic and functional category. However, to suit the multiple regression design setting of equation 4 in the manuscript, the phenotypic vector ($Y_{N\times 1}$) is obtained by summing the scores across the columns for the given phenotypic similarity matrix. Additionally, the corresponding regression design matrix is given by $X_{N\times 4}$. The ridge regression step is computed using BLR package with the given strategy to initialize the priors.

After regression, the phenotypic vector is sorted in non-increasing order according to their predicted scores thereby giving the final ranking of test genes.

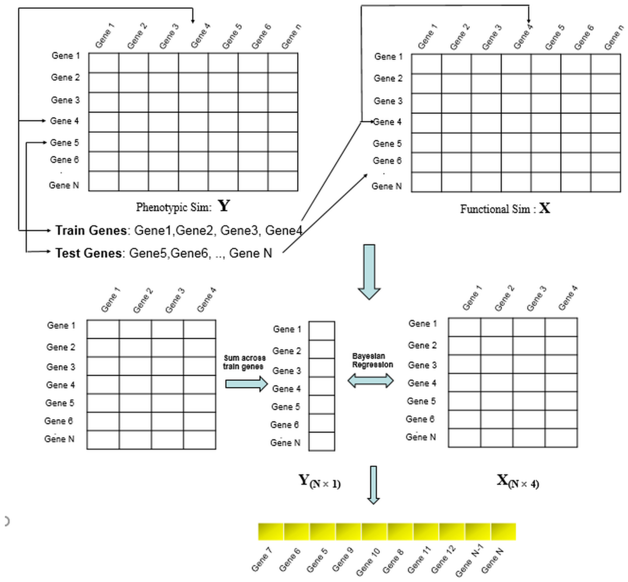


Figure S8.4: Toy example demonstrating prioritization using Bayesian ridge regression.

Chapter 4

Candidate gene resequencing in a large Bicuspid Aortic Valve-associated Thoracic Aortic aneurysm cohort: *SMAD6* as an important contributor

> *Mathematicians have been hiding and writing messages in the genetic code for a long time, but it's clear they were mathematicians and not biologists because, if you write long messages with the code that the mathematicians developed, it would more than likely lead to new proteins being synthesized with unknown functions.*
>
> Craig Venter

Gillis E*[1], **Kumar AA***[1], Luyckx I[1], Preuss C[2], Cannaerts E[1], van de Beek G[1], Wieschendorf B[1,3], Alaerts M[1], Bolar N[1], Vandeweyer G[1], Meester J[1], Wnnemann F[2], Gould RA[4], Zhurayev R[5], Zerbino D[5], Mohamed SA[3], Mital S[6], Mertens L[6], Bjrck HM[7], Franco-Cereceda A[8], McCallion AS[4], Van Laer L[1], Verhagen JMA[9], van de Laar IMBH[9], Wessels MW[9], Messas E[10], Goudot G[10], Nemcikova M[11], Krebsova A[12], Kempers M[13], Salemink S[13], Duijnhouwer T[13], Jeunemaitre X[10], Albuisson J[10], Eriksson P[7], Andelfinger G[2], Dietz HC[4,14], Verstraeten A[1], Loeys BL[1,13]; Mibava Leducq Consortium.
Frontiers in physiology 8 (2017): 400

[1]Faculty of Medicine and Health Sciences, Center of Medical Genetics, University of Antwerp and Antwerp University Hospital Antwerp, Belgium. [2]Cardiovascular Genetics, Department of Pediatrics, CHU Sainte-Justine, Universit de MontrealMontreal, QC, Canada. [3]Department of Cardiac and Thoracic Vascular Surgery, University Hospital Schleswig-HolsteinLbeck, Germany. [4]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of MedicineBaltimore, MD, United States. [5]Department of Clinical Pathology, Lviv National Medical University after Danylo Halytsky Lviv, Ukraine. [6]Cardiovascular Research, SickKids University HospitalToronto, ON, Canada. [7]Cardiovascular Medicine Unit, Department of Medicine, Karolinska InstituteStockholm, Sweden. [8]Cardiothoracic Surgery Unit, Department of Molecular Medicine and Surgery, Karolinska InstituteStockholm, Sweden. [9]Department of Clinical Genetics, Erasmus University Medical CenterRotterdam, Netherlands. [10]Assistance Publique-Hpitaux de Paris, Hpital Europen Georges Pompidou; Universit Paris Descartes, Paris Sorbonne Cit; Institut National de la Sant et de la Recherche Mdicale, UMRSParis, France. [11]Department of Biology and Medical Genetics, 2nd Faculty of Medicine-Charles University and Motol University HospitalPrague, Czechia. [12]Institute of Clinical and Experimental MedicinePrague, Czechia. [13]Department of Human Genetics, Radboud University Medical CentreNijmegen, Netherlands. [14]Howard Hughes Medical InstituteBaltimore, MD, United States.

This chapter was adapted from the article published online at **Frontiers in physiology**. 2017 Sep 25;8:730., `https://tinyurl.com/yaqar3sf`

4.1 Abstract

Bicuspid aortic valve (BAV) is the most common congenital heart defect. Although many BAV patients remain asymptomatic, at least 20% develop thoracic aortic aneurysm (TAA). Historically, BAV-related TAA was considered as a hemodynamic consequence of the valve defect. Multiple lines of evidence currently suggest that genetic determinants contribute to the pathogenesis of both BAV and TAA in affected individuals. Despite high heritability, only very few genes have been linked to BAV or BAV/TAA, such as *NOTCH1*, *SMAD6*, and *MAT2A*. Moreover, they only explain a minority of patients. Other candidate genes have been suggested based on the presence of BAV in knockout mouse models (e.g., *GATA5*, *NOS3*) or in syndromic (e.g., *TGFBR1/2*, *TGFB2/3*) or non-syndromic (e.g., *ACTA2*) TAA forms. We hypothesized that rare genetic variants in these genes may be enriched in patients presenting with both BAV and TAA. We performed targeted resequencing of 22 candidate genes using Haloplex target enrichment in a strictly defined BAV/TAA cohort (n = 441; BAV in addition to an aortic root or ascendens diameter 4.0 cm in adults, or a Z-score 3 in children) and in a collection of healthy controls with normal echocardiographic evaluation (n = 183). After additional burden analysis against the Exome Aggregation Consortium database, the strongest candidate susceptibility gene was *SMAD6* (p = 0.002), with 2.5% (n = 11) of BAV/TAA patients harboring causal variants, including two nonsense, one in-frame deletion and two frameshift mutations. All six missense mutations were located in the functionally important MH1 and MH2 domains. In conclusion, we report a significant contribution of SMAD6 mutations to the etiology of the BAV/TAA phenotype.

4.2 Introduction

With a prevalence of 12% in the general population, bicuspid aortic valve (BAV) is the most common congenital heart defect. It has a 3:1 male preponderance and is characterized by an aortic valve with two cusps instead of the normal three. BAV often coincides with aortic manifestations such as coarctation of the aorta and thoracic aortic aneurysm (TAA)[215].The latter can lead to lethal dissections if left untreated. Although first described over 400 years ago and high heritability (89%) [52], the genetic etiology of BAV, with or without TAA, remains largely elusive. It was initially suggested that TAA results from altered blood flow dynamics imposed by the abnormal bicuspid valve. Changes in shear stress were presumed to weaken the aortic wall, resulting in dilatation and rupture. At present, common genetic risk factors for BAV and TAA are proposed [83], based on the following observations: (i) the aortic valve and the aorta share common embryologic origins [i.e., the cardiac neural crest (CNC) and the second heart field] [142], (ii) family members of BAV/TAA probands show TAA without valve abnormalities and/or BAV without aneurysmal disease [136], and (iii) TAA formation in BAV probands that previously underwent valve replacement has been reported [34].

Transmission of BAV/TAA mostly complies with an autosomal dominant inheritance pattern, displaying reduced penetrance and variable expressivity [50, 86]. Few genes have been robustly linked to the BAV phenotype to date. *NOTCH1* is often considered the sole established BAV gene, either as an isolated finding or in association with early onset valve calcification, TAA, or other left-sided heart defects [30, 65, 68, 70, 100, 102, 145, 148]. *SMAD6* [201] and *MAT2A* [76] have also been implicated in BAV, but only in a very limited number of patients. A dozen candidate genes emanated from knockout mouse models with increased BAV occurrence [22, 113, 114, 125, 149, 178, 205]. The prevalence of BAV in these knockout models is often low (range: 2-42% in single knockouts) (Table 4.1), probably due to reduced penetrance and/or activation of compensatory mechanisms. Mutations in some syndromic [10, 36, 134, 159, 168, 213] or non-syndromic [77] TAA genes also associate with increased BAV occurrence. (Table 4.1).

To date, no major BAV/TAA gene has emerged. The described genes have been associated with BAV, but their contribution to the etiology of BAV/TAA has never been examined systematically. Here, we evaluate this contribution in 22 BAV-associated genes (Table 4.1) using a targeted gene panel and variant burden approach.

4.3 Materials and Methods

4.3.1 Study Cohort

Genomic DNA (gDNA) of 441 BAV/TAA patients was collected through a collaborative effort involving 8 different centers (Supplementary Table 4.4. Patients were selected based on the presence of BAV and either an aortic diameter at the sinus of Valsalva or the ascending aorta of at least 4.0 cm in adults, or a Z-score exceeding 3 in children. Aortic diameter dimensions were determined using echocardiography, computed tomography or magnetic resonance imaging. A positive family history was defined as having at least one first- or second-degree relative with BAV and/or TAA. Control gDNA was obtained from 183 cancer patients who presented at the SickKids Hospital, Toronto, Canada. None of the controls showed structural heart disease upon examination with echocardiography. All study participants or their legal guardians gave informed consent

Context	Gene	Incidence	References	Selection Criteria
BAV in humans	*NOTCH1*	Mutations found in 27 BAV patients	[30, 65, 68, 70, 100, 102, 145, 148]	Literature
	SMAD6	Mutations found in 2 BAV patients	[201]	Prioritization
	MAT2A	Mutations found in 1 BAV patient	[76]	Literature
BAV in mice	*ACVR1*	BAV in 78-83% of $Alk2^{FXKO}/Gata5^{-Cre+}$ mice	[205]	MIBAVA Consortium
	GATA4	BAV in 43% of $Gata4^{+/-};Gata5^{+/-}$ mice	[114]	Literature
	GATA5	BAV in 25% of $Gata5^{-/-}$ mice	[113]	Literature
	GATA6	BAV in 25% of $Gata5^{+/-};Gata6^{+/-}$ mice	[114]	Literature
	MATR3	BAV in 12% of $Matr3^{+/-}$ mice	[178]	Literature
	NKX2-5	BAV in 2-20% of $Nkx2-5^{+/-}$ mice	[22]	Literature
	NOS3	BAV in 42% of $Nos3^{-/-}$ mice	[125]	Literature
	ROBO1	BAV in 100% of $Robo1^{-/-};Robo2^{-/-};$ mice	[149]	Literature
	ROBO2	BAV in 100% of $Robo1^{-/-};Robo2^{-/-};$ mice	[149]	Literature
BAV in (non)syndromic TAA cases	*FBN1*	Occasional BAV in Marfan syndrome	[10, 159, 168]	Literature
	ACTA2	7% BAV in non-syndromic TAA	[77]	Literature
	ELN	Occasional BAV in cutis laxa	[36]	Prioritization
	FLNA	Occasional BAV X-linked valve disease	[92]	CNV analysis
	MYH11	Occasional BAV non-syndromic TAA	Personal observation	Prioritization
	SMAD3	3-11% BAV in Loeys-Dietz syndrome	[213]	Literature
	TGFB2	8-13% BAV in Loeys-Dietz syndrome	[134]	Literature
	TGFB3	4% BAV in Loeys-Dietz syndrome	Personal observation	Literature
	TGFBR1	8-12% BAV in Loeys-Dietz syndrome	Personal observation	Literature
	TGFBR2	8-12% BAV in Loeys-Dietz syndrome	Personal observation	Prioritization

[1] BAV: Bicuspid aortic valve

[2] TAA: Thoracic aortic valve

Table 4.1: Genes included in the targeted gene panel and the criteria on which their selection was based.

at the respective sample-contributing centers

4.3.2 Targeted Enrichment

Genes (n = 22) were selected for targeted resequencing based on the following criteria: (i) mutations occur in human BAV cases (n = 3), (ii) knockout mouse models present with incomplete penetrance of BAV (n = 9), and (iii) occasional or increased BAV manifestation occurs in patients with mutations in known TAA genes (n = 10) (Table 4.1. Enrichment of all exons of these candidate genes, including ±10 nucleotides of adjacent intronic sequence, was performed with a custom Haloplex target enrichment kit per instructions of the manufacturer (Agilent Technologies, USA). Probe design covered a theoretical 99.7% of the complete target region (560 kb). Pooled samples were sequenced either on a HiSeq 2500 (Illumina, USA) with 2 × 150 bp reads or on a HiSeq 1500 (Illumina, USA) with 2 × 100 bp reads.

4.3.3 Data analysis and filtering

The raw data were processed using an in-house-developed Galaxy-based pipeline, followed by variant calling with the Genome Analysis Toolkit Unified Genotyper [56]. Variants were subsequently annotated and filtered with the in-house developed database VariantDB [214], which uses ANNOVAR. Heterozygous coding or splice site (±2 bp from exon-intron boundaries for nucleotide substitution, and ±5 bp for multi-bp deletions or insertions) variants with an allelic balance between 0.25 and 0.85 (FLNA in males: 0.751) and a minimum coverage of 10 reads were selected. Finally, we included variants that fitted within at least one of the following three categories; unique variants [absent in the Exome Aggregation Consortium (ExAC) database [126], variants with an ExAC Minor Allele Frequency (MAF) lower than 0.01% or variants with an ExAC MAF between 0.01% and 0.1% that had a Combined Annotation Dependent Depletion (CADD) [104] score above 20. All splice region variants underwent splice site effect prediction using ALAMUT (Interactive Biosoftware, France). Synonymous variants outside of splicing regions were not taken into account.

The ExAC database was used as an independent control dataset. The raw data of variants (∼all ExAC datasets) fulfilling ExAC's quality control parameters ("PASS") were extracted from the offline version of ExAC v0.3.1. Since the ExAC variants were annotated using VEP, whereas our patient variant annotation was ANNOVAR-based, we re-annotated the ExAC variants with ANNOVAR. The same variant filtering strategy as described for the patient cohort was subsequently applied. For each selected ExAC variant, the allele frequency was determined by computing the ratio of the Mutant Allele Count (mAC) and Total Allele Count (tAC). Next, we re-scaled each variant's mAC by multiplying its computed allele frequency by its respective tAC_Adj, i.e., the tAC average of all variants in that specific gene. Finally, the variant counts for each panel gene were obtained by summing up the re-scaled mACs.

4.3.4 Validation by Sanger sequencing

Variants discussed in the results section were confirmed with Sanger sequencing. Primers were designed using Primer3 software [210] v4.0.0 and polymerase chain reaction (PCR) products were purified with Calf Intestinal Alkaline Phosphatase (Sigma-Aldrich, USA). Sequencing reactions were performed using the BigDye Terminator Cycle Sequencing kit (Applied Biosystems, Life Technologies, USA), followed by capillary electrophoresis on an ABI3130XL (Applied Biosystems, Life Technologies, USA). The obtained sequences were analyzed with CLC DNA Workbench v5.0.2 (CLC bio, Denmark).

4.3.5 Segregation analysis

When family members were available, Sanger sequencing of the SMAD6 variants identified in the proband was performed in additional relatives to check if the phenotype segregated with the variant.

4.3.6 Statistical Analysis

We performed burden analyses comparing frequencies of the variants fulfilling the three criteria that were mentioned in Section Data Analysis and Filtering between patients and controls. Whereas the Fisher's Exact Test was used to statistically compare variant frequencies in the patient cohort to those in the study control cohort, the Chi-Square Test with Yates' correction was used for the patient-ExAC comparison. No p-values were calculated if the number of variants in patients and/or controls was zero. Fisher's Exact statistics were also used to determine if significant variant type enrichment and/or domain clustering of variants occurs in patients. Statistical significance was considered when $p < 0.05$.

4.4 Results

The patient cohort consisted of 441 BAV/TAA patients (75% males and 25% females) with an average age at inclusion of 63.5 ± 14.4 years. For these patients, the most common associated feature was coarctation of the aorta (2.9%, n = 13). About 3% (n = 14) had other additional findings such as mitral valve prolaps, aortic stenosis, dilated cardiomyopathy, aortic insufficiency, patent ductus arteriosus or intracranial aneurysm. 46.7% (n = 206) had a left-right leaflet BAV orientation, 15.9% (n = 70) had a right-non-coronary leaflet BAV orientation and for 37.4% (n = 165) of the patients the subtype of valve leaflet morphology was not specified. A positive family history was known for 9.3% of the patients, whereas for the remainder the family history was negative or unknown. The study control cohort (n = 183) consisted of 58% males and 42% females. The average age at inclusion of this control cohort was 13.1 ± 5.1 years.

Targeted gene panel sequencing reached an overall coverage at 10x of 99.13% of the targeted regions. In total, 169 variants passed our selection criteria in our patient and control group (Supplementary Table 4.5). Of these, 112 variants were identified in 441 patients. They included 101 missense, 2 nonsense, 2 splice-site, 5 in-frame indel, and 2 frameshift variants. The 183 study controls contained 57 variants including 53 missense, 1 nonsense, 2 splice-site, and 1 frameshift variant. After applying the identical filtering criteria to the ExAC control cohort, 15,660 variants were retained in on average 54,940 individuals: i.e., 14,931 missense, 190 splice-site, 72 nonsense, 10 no-stop, 204 frameshift, and 253 in-frame indel variants.

To validate our control cohort, we compared its variant frequencies for the 22 selected candidate genes to those of the ExAC cohort. No significant differences were observed (Figure 4.4.1). We then performed a variant burden analysis equating the numbers of patient variants per gene to the numbers found in the control cohort (Table 4.2. Results are graphically presented in Figure 4.4.1, showing the proportion of variants per gene in the three different cohorts. Although a few genes (e.g., *FLNA*) showed trends toward significance when comparing our study patient and control cohort, we decided to focus on the patient-ExAC comparison because of the larger number of controls in the ExAC cohort and hence, higher power. Only *SMAD6* reached significance (p = 0.002) in the patient-ExAC comparison. Remarkably, a protective effect for *NOS3* and *NOTCH1* variants was suggested (p = 0.06 and p = 0.05, respectively).

We identified 11 *SMAD6* variants in 441 patients (2.5%). These included two frameshift deletions, two nonsense mutations, one in-frame deletion, and six missense variants (Figure 4.4.2. Only a single individual (0.55%) in the study control cohort harbored a *SMAD6* missense variant. The ExAC database harbored 450 *SMAD6* variants in 47,389 individuals (0.9%). Whereas 36.4% (n = 4/11) of the *SMAD6* mutations in the patient cohort were loss of function (LOF; frameshift, nonsense or splice site) mutations, truncating *SMAD6* mutations were found in only 4.0% (n = 18/450) of the ExAC individuals, demonstrating a clear enrichment in BAV/TAA patients compared to controls (p = 0.001).

The SMAD6 c.726del variant leads to a frameshift (p.Lys242Asnfs*300) and a predicted protein with a C-terminal extension due to loss of the intended stop codon. The c.454_461del frameshift variant (p.Gly166Valfs*23) causes the introduction of a premature stop codon, most likely resulting in haploinsufficiency due to nonsense-mediated mRNA decay (NMD). Also the two nonsense variants (p.Tyr279* and p.Tyr288*) are predicted to lead to NMD. All of the missense variants cluster in the functionally important MH1 and MH2 domains [140] (amino acids 148275 and 331496, respectively), which is not the case for the sole missense variant (p.Ser130Leu) found in a control individual (Figure 4.4.2. All but one (p.Arg443His) of the identified variants were absent in the ExAC control cohort (v0.3.1; Supplementary Table 4.5. Moreover, the missense variants in the patient cohort (7/7) are enriched in the MH1 and MH2 domains when compared

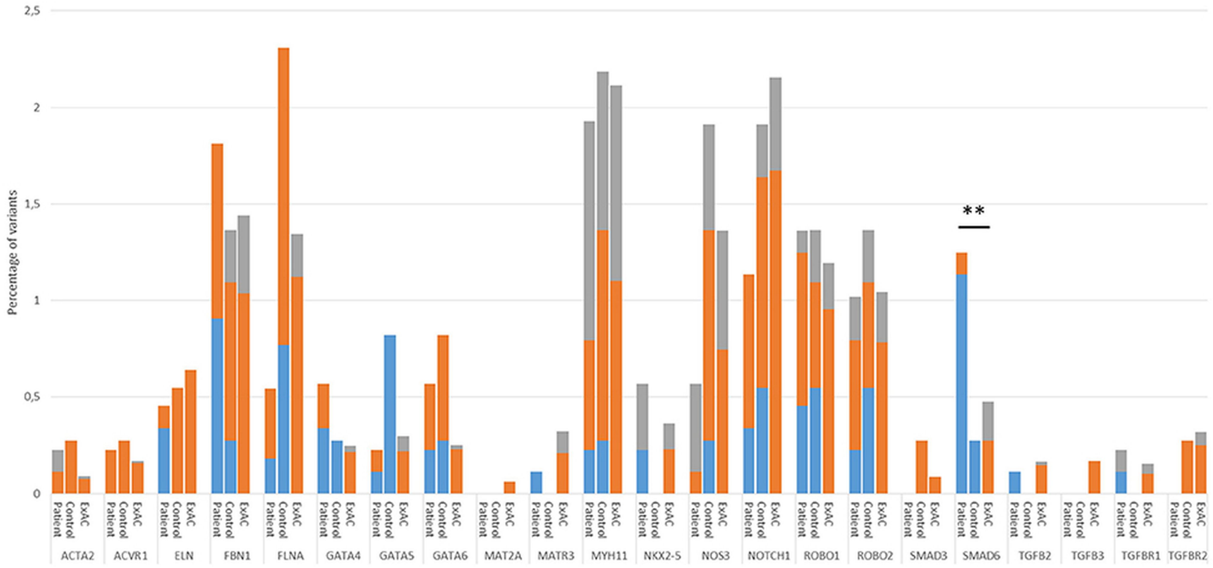


Figure 4.4.1: **Proportion of variant alleles per gene in the patient group, control group and ExAC cohort.** Variants were selected as follows: First, we selected heterozygous coding or splice site variants with an allelic balance between 0.25 and 0.85 (FLNA in males: 0.751) and a minimum coverage of 10x. Next, we made three variant groups based on their frequency in the ExAC database; that is, variants that are absent from the ExAC control dataset (blue), variants with an ExAC MAF lower than 0.01% (orange) and variants with an ExAC MAF between 0.01% and 0.1% that had a CADD score above 20 (gray). Only statistics of the patient-ExAC comparison are shown (**$p \leq 0.01$). No statistically significant differences in allele frequencies were observed between our control cohort and the ExAC controls. Abbreviations: ExAC, Exome Aggregation Consortium; MAF, Minor Allele frequency; CADD, Combined Annotation Dependent Depletion.

to ExAC controls (n = 228/430; p = 0.02).

For two *SMAD6* mutation carriers (P89, p.Gly271Glu; P99, p.Gly166Valfs*23), gDNA of family members was available for segregation analysis (Supplementary Figure 4.10.3). Although neither of these probands had a documented family history of BAV/TAA, a brother of P89 has been diagnosed with a sinus of Valsalva aneurysm (45 mm) and carried the *SMAD6* mutation. The mutation was also observed in an unaffected daughter (age 28) of the proband (Supplementary Figure 4.10.3). Three unaffected siblings at ages 54, 58, and 64 did not carry the mutation. No gDNA was available from a sister of P99 with unspecified aortic valve problems. The p.Gly166Valfs*23 mutation was found in an unaffected daughter (age 39) of P99 but was absent in his 39 year-old unaffected son (Supplementary Figure 4.10.3).

Intriguingly, two genes (*NOTCH1* and *NOS3*) that previously had been associated with increased BAV risk in humans [65, 70, 145, 148] and/or mice [31, 125] revealed borderline significance for protection from BAV/TAA (p = 0.05 and p = 0.06, respectively). Analysis of *NOTCH1* identified 10 variants in patients (2.3%), including two splice-site

Gene	Number of variants in 882 patient alleles	Number of variants in 366 control alleles	Number of variants in ExAC alleles	p-value patient-controls	p-value patients-ExAC
ACTA2	2	1	109 in 120,631	1.00	0.44
ACVR1	2	1	202 in 120,994	1.00	0.98
ELN	4	2	728 in 113,954	1.00	0.63
FBN1	16	5	1,740 in 120,988	0.81	0.43
FLNA	3	6	1,133 in 84,359	**0.03**	0.15
GATA4	5	1	260 in 105,980	0.68	0.11
GATA5	2	3	259 in 86,819	0.15	0.94
MAT2A	0	0	74 in 116,667	/	/
MATR3	1	0	382 in 119,089	/	/
MYH11	17	8	2,513 in 119,001	0.82	0.79
NKX2-5	5	0	360 in 98,978	/	0.47
NOS3	5	7	1,390 in 102,070	0.05	0.06
NOTCH1	10	7	2,181 in 101,245	0.29	0.05
ROBO1	12	5	1,354 in 113,390	1.00	0.77
ROBO2	9	5	1,245 in 119,282	0.57	0.95
SMAD3	0	1	95 in 111,500	/	/
SMAD6	11	1	450 in 94,779	0.20	**0.002**
TGFB2	1	0	192 in 117,070	/	0.71
TGFB3	0	0	205 in 121,315	/	/
TGFBR1	2	0	181 in 118,320	/	0.90
TGFBR2	0	1	366 in 115,147	/	/

Variant burden analyses were performed by comparing frequencies of the variants fulfilling the three criteria that were mentioned in "Section Data Analysis and Filtering" between patients and controls. Whereas, the Fisher's Exact Test was used to statistically compare variant frequencies in the patient cohort to those in the study controls cohort, the Chi-Square Test with Yates' correction was used for the patient ExAC comparison. No p-values were calculated if the number of the variants in the patients and/or controls was zero. Statistical significance was considered when $p < 0.05$. The asterisk denote that in these cases the number of alleles is consistent with the number of X-chromosomes, i.e., 553 patient alleles and 260 control alleles were checked for variants. Statistically significant p-values are represented in bold.

Table 4.2: Variant burden comparisons per gene between patients and either study controls or ExAC controls.

variants, vs. seven variants (all missense) in controls (3.8%) and 2,181 (4.3%) variants in ExAC. One variant in the patient cohort (c.5167+3_5167+6del) leads to complete loss of the 5′ donor splice site of intron 27, predicted to result in skipping of exon 27 (149 bp) and hence a frameshift. For the second variant (p.S784S), the predicted effect on splicing is more ambiguous. If loss of the 5′ donor splice site of intron 14 would occur, skipping of exon 14 (146 bp) would again lead to a frameshift event. Unfortunately, cDNA to

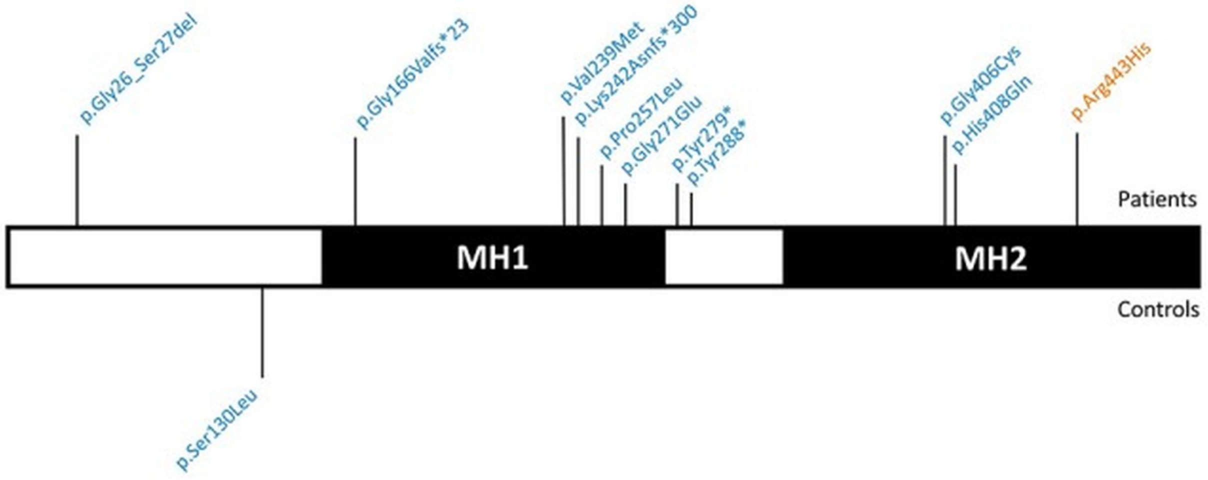


Figure 4.4.2: **Graphical representation of the identified *SMAD6* variants**. *SMAD6* has two major protein domains, a DNA-binding MH1 domain and a MH2 domain that interacts with components of the TGF $-\beta$ and BMP signaling pathways. Variants above the protein have been found in patients, while those below the protein occurred in control individuals. Variants in blue are absent from the ExAC database, variants in orange have an ExAC MAF below 0.01%. Abbreviations: TGF $-\beta$, Transforming growth factor $-\beta$; BMP, Bone morphogenetic protein; ExAC, Exome Aggregation Consortium; MAF, Minor Allele frequency.

reliably determine the precise effect of these mutations on splicing is not available. None of the *NOTCH1* variants that we identified in BAV/TAA patients has previously been reported in the literature. We did not observe any variant-domain clustering or significant differences in CADD scores when comparing the patient and control NOTCH1 variants. Similarly, for *NOS3* a total of five missense variants (1.1%) was found in patients, whereas the control cohort harbored seven variants (3.8%), including one out-of-frame mutation (p.Leu927Hisfs*32). In the ExAC control cohort, 1,390 NOS3 variants (2.7%) were found in 51,035 individuals.

Based on statistical analyses of BAV/TAA heritability and the fact that BAV/TAA shows prominent gender bias, oligogenic inheritance of BAV/TAA is an emerging concept [8, 215]. To test for such oligogenic patterns, we determined the number of patients and controls in our study cohort with variants in at least two out of the 22 analyzed genes. In the patient cohort, 10 patients presented with two variants (2.3%), while the control group harbored 7 individuals that carried two variants (3.8%). Based on these data, there is no evidence for a digenic or multigenic model in the analyzed genes (p = 0.29).

4.5 Discussion

So far, no gene with a contribution of more than 1% to BAV or BAV/TAA has been identified in humans. Gene identification has been hampered by low penetrance, variable clinical expressivity, the likelihood of BAV-phenocopies within individual families and, most likely, substantial locus heterogeneity. [215]. *NOTCH1* has been suggested as a BAV(/TAA) gene, but does not contribute greatly to disease etiology. About 20 other genes have been associated with BAV in humans and mice (Table 4.1), but few of them also showed association with TAA. This suggests that whereas some disease genes might be linked to both BAV and TAA, others increase risk for only one of the component phenotypes. In this study, we used a targeted gene panel approach to study the prevalence of mutations in genes that previously have been associated with BAV and/or TAA in people or mice in a cohort of BAV/TAA patients. In total, 22 genes were sequenced in 441 BAV/TAA patients and 183 controls. *SMAD6* was identified as the most important known gene in the etiology of BAV with associated TAA. With 11 mutation-carrying probands, *SMAD6* offers a molecular explanation for 2.5% of our study population. For two of the variants segregation analysis in relatives could be performed, revealing the presence of one of the respective *SMAD6* mutations in a TAA patient and two rather young individuals (age 28 & 39) that might still develop TAA later in life. Four unaffected individuals (age 37, 54, 58, 64) did not carry a *SMAD6* mutation. As two nonsense and two frameshift *SMAD6* variants in our cohort are predicted to lead to haploinsufficiency, LOF is the most likely mechanism. All the patient-specific missense variants (n = 7) are in the functionally important MH1 and MH2 domains of *SMAD6* [140]. LOF missense mutations in *SMAD2* and *SMAD3* causing Loeys-Dietz syndrome, another syndromic TAA form, are also located in the MH1 and MH2 domains [147, 212]. The MH1 domain of SMAD6 binds DNA (Bai and Cao, 2002), while the MH2 domain interacts with key components of the transforming growth factor (TGF)- and bone morphogenetic protein (BMP) signaling cascades [79, 96, 132]. In 2012, two missense variants in the MH2 domain of SMAD6 were identified in two patients with BAV in association with mild to moderate aortic stenosis [201]. Interestingly, in our cohort, one *SMAD6* patient (p.Tyr288*) presented with coarctation in addition to BAV and TAA. Moreover, mice lacking expression of the murine orthologue of *SMAD6*, i.e., Madh6/ mice, also present with cardiovascular pathologies, including abnormal vascular smooth muscle cell relaxation, thickening of the cardiac valves and misplaced septation and ossification of the outflow tract (OFT) [69]. As such, our findings confirm a role for *SMAD6* mutations in the etiology of BAV and expand the spectrum of SMAD6-related cardiovascular manifestations with

BAV-related TAA.

SMAD6 is highly expressed in the cardiac valves and OFT of the embryonic heart, in the late-embryonic, and adult vascular endothelium as well as in the vascular smooth muscle cells of the adult aortic root [57, 69]. Upregulation in response to laminar shear stress has been reported [207]. *SMAD6* encodes an inhibitory SMAD protein which negatively regulates BMP signaling by binding to BMP type I receptors or by establishing competitive interactions for *SMAD4* [81, 90]. In doing so, SMAD1/5/8 phosphorylation and/or nuclear translocation are prevented. Additionally, *SMAD6* cooperates with *SMURF* E3 ubiquitin ligases to prime ubiquitin-mediated proteasomal degradation of BMP receptors and SMAD effector proteins [154], including *SMAD1* and 5. BMP signaling has previously been independently implicated in BAV- and TAA-related processes [35, 71]. In addition to mediating CNC cell migration into the cardiac cushions and differentiation to smooth muscle cells, BMP signaling promotes endothelial-to-mesenchymal transition and instigates mesenchymal cell invasion [71, 97]. While *SMAD6* and *SMAD7* are thought to have a predominant negative regulatory effect on BMP and $\mathrm{TGF}-\beta$ signaling, respectively, there is strong evidence that this specificity is not absolute and that SMAD6 can directly suppress the $\mathrm{TGF}-\beta$ signaling cascade. Important crosstalk between BMP, $\mathrm{TGF}-\beta$ and NOTCH signaling has been reported [71]. Many syndromic forms of TAA are caused by mutations in genes encoding effectors or regulators of the $\mathrm{TGF}-\beta$ signaling pathway (including *TGFB2/3*, *TGFBR1/2*, *SMAD2/3*, *SKI*) [20, 27, 38, 58, 134, 135, 147, 212], with increased activity observed in aortic specimens from people and mice with these conditions. An increased prevalence of BAV has been observed in patients carrying mutations in these genes (Table 4.1). Overall, these results imply that mutations in *SMAD6* likely cause BAV/TAA through impaired negative regulation of BMP and/or $\mathrm{TGF}-\beta$ signaling.

Multiple studies have previously reported a link between *NOTCH1* mutations and BAV [65, 70, 145, 148]. In 2005, a nonsense and a frameshift *NOTCH1* mutation were found to segregate with BAV associated with early onset valve calcification in the respective families [70]. Since the initial report, multiple *NOTCH1*, mostly missense, variants have been associated with BAV, BAV/TAA, aortic valve stenosis, coarctation, and hypoplastic left heart [65, 68, 89, 91, 145, 148, 174]. In addition to these mutations in association with left-sided heart defects, frameshift and nonsense mutations were also identified in patients with right-sided heart defects affecting the pulmonary valve and conotruncal disease including pulmonary atresia with intact ventricular septum, tetralogy of Fallot, and truncus arteriosus, and other congenital heart diseases, such as anomalous pulmonary venous return, atrial septal defect, and ventricular septal defect [102]. Mouse models have confirmed a role for Notch1 in the development of the aortic valve and the cardiac OFT [106]. Unexpectedly, in our dataset *NOTCH1* did not stand out as a prominent BAV/TAA gene, with the suggestion that *NOTCH1* variants might even be protective. Sample selection bias might contribute to this observation as *NOTCH1* variants appear to associate with early and severe valve calcification and seem to be enriched in families with highly penetrant BAV but far lower penetrance of TAA[100]. Given that our study did not select for valve calcification and prioritized the BAV/TAA phenotype, it is understandable that NOTCH1 variants would be underrepresented. It also seems notable that only missense variants were seen in controls, while multiple variants in the patient cohort are predicted to have a more overt impact on protein expression and function.

Similarly, our variant burden test suggested that NOS3 variants might be protective for BAV/TAA development. *NOS3*, the endothelial specific nitric oxide (NO) synthase, is important in balancing NO production and in the reduction of oxidative stress [66]. Its role in cardiac development is demonstrated by the formation of BAV in Nos3-targeted mice (Table 4.1). Furthermore, it has already been shown that specific *NOS3* polymorphisms can affect NO production [161], and increased NO levels have been found in a MFS mouse model and in Adamts1-deficient mice that develop TAA [162]. Pharmacological inhibition of *NOS2* in mice led to a protective effect in aortic aneurysm development [162]. This supports the importance of NO levels and nitric oxide synthases in aneurysm pathology. The variants in *NOS3* identified in the current study may lead to less active *NOS3* and as such may protect against development of aortic aneurysm.

Our study has several methodological limitations: (i) The small number of genes included in our study, as well as the patient cohort size, precludes the ability to detect oligogenic inheritance or gene-gene interactions involved in BAV/TAA. An extended experiment in a larger BAV/TAA cohort, including BAV-related pathways instead of selected genes, could give us more insight regarding how genes work together in BAV and/or TAA development; (ii) The size of the patient and study/ExAC control cohort only allows us to detect BAV/TAA genes with a fairly large contribution (variant burden in patients: 3%&2%, respectively); (iii) The control cohort consists of younger, adolescent patients that did not show cardiac complications at the time of investigation but may still develop complications such as TAA later-on in life. Therefore, the ExAC database was used as an additional dataset for allele frequencies in a cohort without gross developmental defects.

Our study specifically assesses the presence of pathogenic variants in BAV-associated genes in a large BAV/TAA cohort. We conclude that *SMAD6* is currently the most important contributor to the genetic architecture of BAV/TAA. More research and larger cohorts will be needed to fully elucidate the genetic architecture of this common but complex cardiovascular pathology.

4.6 Ethics Statement

This study was carried out in accordance with written informed consent from all subjects. All subjects gave written informed consent in accordance with the Declaration of Helsinki. The protocol was approved by the Ethics Committee of the Antwerp University Hospital and all participating centers.

4.7 Author Contributions

All authors revised the work critically. All authors provided final approval of the version for publication. All authors agreed to be accountable for all aspects of the work and ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. More specific contributions to the work are: EG, AAK, IL, CP, EC, NB, FW, RG, LV, SAM, SM, LM, HB, AF, AM, PE, GA, HD, AV, and BL contributed to conception and design of the work. EG, AAK, IL, EC, MA, NB, GvdB, BW, GV, JM, RZ, DZ, SAM, SM, LM, JV, IV, MW, EM, GG, MN, AK, MK, SS, TD, XJ, JA, PE, AV, and BL contributed to acquisition of the data. EG, AAK, CP, FW, GA, LV, HD, AV, and BL contributed to analysis of the data. EG, IL, CP, EC, MA, FW, RG, RZ, DZ, SAM, SM, LM, HB, AF, AM, LV, JV, IV, MW, EM, GG, MN, AK, MK, SS, TD, XJ, JA, PE, HD, AV, and BL contributed to interpretation of the data.

4.8 Funding

This research was supported by funding from the University of Antwerp (Lanceringsproject), the Fund for Scientific Research, Flanders (FWO, Belgium, G.0221.12), The Dutch Heart Foundation (2013T093), the Foundation Leducq (MIBAVALeducq 12CVD03). BL is senior clinical investigator of the Fund for Scientific Research, Flanders (FWO). AV and GV are FWO postdoctoral researchers and JM (FWO) and IL (FWO-SB) hold a PhD mandate. BL holds a starting grant from the European Research Council (ERC- StG-2012-30972-BRAVE).

4.9 Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

4.10 Supplementary Material

4.10.1 Selection of candidate genes from MIBAVA exomes

As part of Mechanistic Insight for Bicuspid Aortic Valve Aortopathy (MIBAVA) consortium whole exome sequencing (WES) was performed on 196 patients (with BAV/TAA phenotype) and 193 controls (with non-cardiovascular disease phenotype) in order to select candidate genes. The selected candidate genes were next subjected to targeted resequencing in a relatively larger replicative cohort for identification of rare causal variants associated with BAV/TAA. The workflow for selecting candidate genes from the WES pipeline is presented in Figure 4.10.1. In the first stage the WES data from the cases and controls are processed using standard GATK software tools (with similar settings) for variant calling. Subsequently all the variants were annotated with RefSeq through VariantDB framework. For rare variant association analysis the variants were filtered for rareness and genotype quality. Based on this we devised three frequency based filters i.e variants that are: (a) unique or absent in any public databases (b) have a MAF $\leq 0.1\%$ and (c) MAF $\leq 1\%$. Combining unique and either of two MAF filters gives two filter settings based on which the respective variants for the 196 BAV/TAA cases and 193 controls were further processed. After filtering we obtained on an average 2000-3000 variants in 150-200 genes per sample. This list of variants is trimmed by incorporating gene priortization (through pBRIT) and burden test (using CAST approach) based computational methods which eventually yield a list of the most promising candidate genes. Finally, the obtained list of candidate genes are further analyzed for the variants regarding their deleteriousness and pathogenicity.

Prioritization using pBRIT: The gene prioritization was performed using pBRIT by selecting a list of 30 training genes. Since so far only NOTCH1 gene has been known to be associated with BAV disease hence the training set was expanded by adding genes by searching information from literature about BAV/TAA, genes that are involved in TGFB-pathway, exhibit function in Extra cellular matrix (ECM), transcription factors, smooth muscle cells (SMC) and other functionalities. The overall list of genes summarized according to these criteria is presented in Table 4.3.

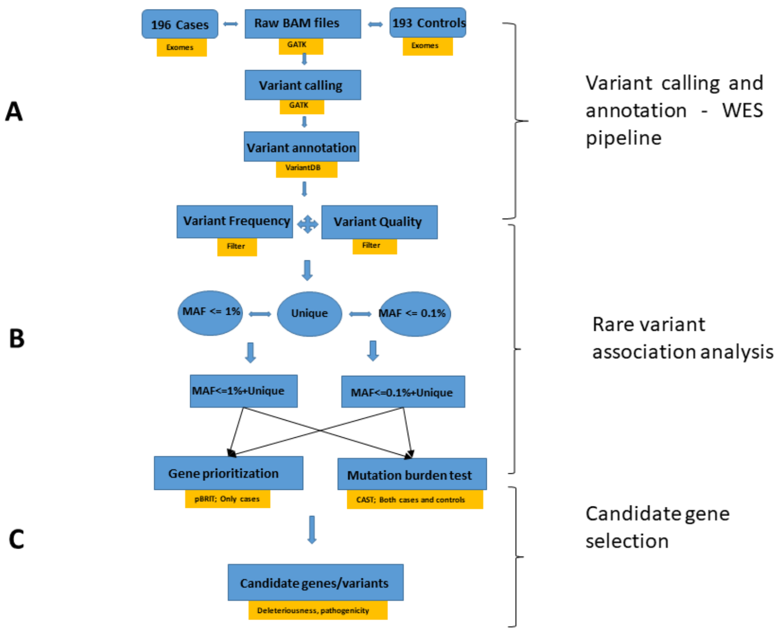


Figure 4.10.1: **Workflow regarding candidate gene selection through a WES pipeline** (A) **Variant calling and annotation:** WES data from 196 BAV/TAA patients and 193 controls were processed for variant calling using GATK based software tools. The obtained variants were then annotated using VariantDB tool. (B) **Rare variant association analysis (RVAS):** The variants obtained per gene per patient sample were filtered based on it's frequency and genotype quality. The filtered variants collapsed per gene were then subjected to prioritization and frequency count based mutation burden test to yield significant candidate genes/variants (C) **Candidate gene selection:** The significant genes/variants were further subjected for deleteriousness and pathogenicity analysis in order to select best candidate genes for targeted resequencing.

Literature	**TGFβ Pathway**	**Extra Cellular Matrix (ECM)**	**Transcription Factors (TF)**	**Smooth Muscle Cell (SMC)**	**Others**
NOTCH1	SMAD3	COL3A1	HOXA1	ACTA2	JAG1
NOS3	TGFB2	FBN1	NKX2-5	MYH11	FGF8
UFD1L	TGFBR1	FLNA	GATA5	MYLK	
FN1	TGFBR2	EFEMP2	GATA6		
AXIN1	SKI	ELN			
PDIA2	SLC2A10				
ENG					
IGFBP2					
IGF1					
HSP27					

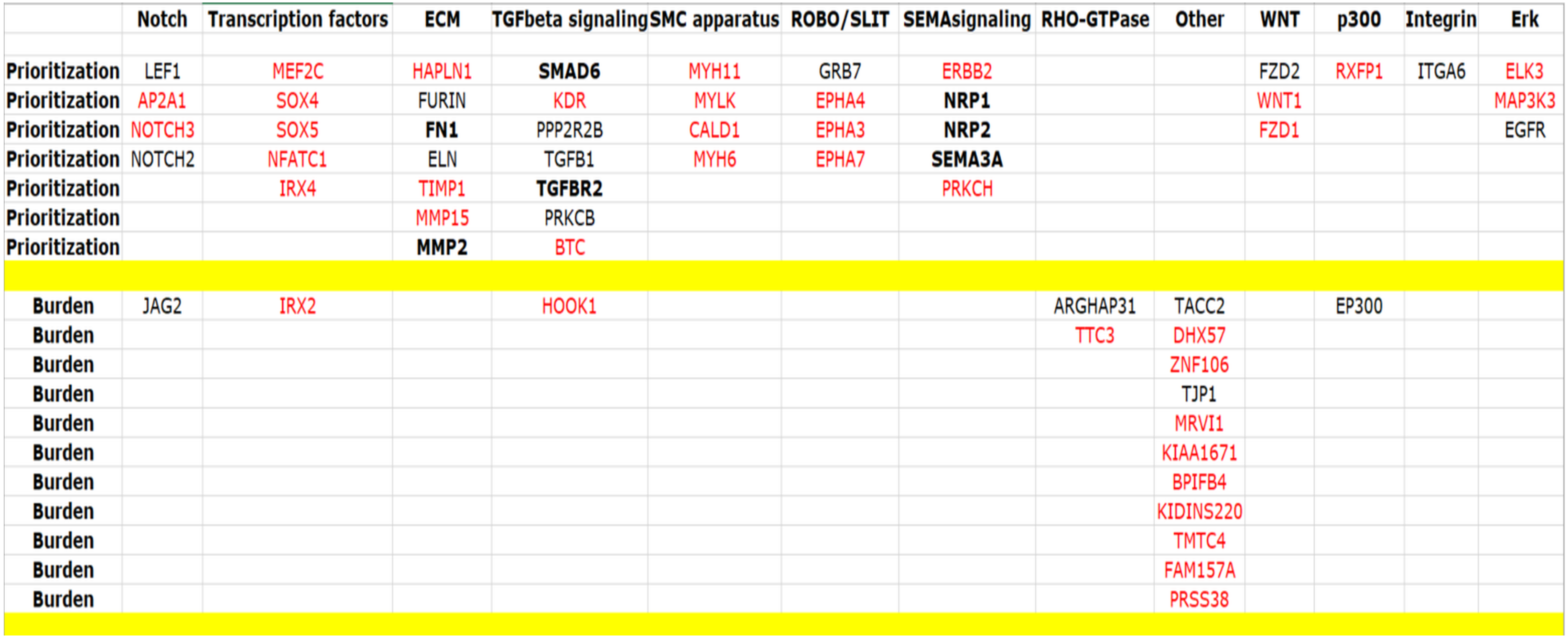
Table 4.3: List of training genes for prioritization of MIBAVA exomes

Mutation burden analysis: For the rare variant association analysis we incorporated CAST based approach (section 1.4.3 of **chapter 1**) for performing mutation burden test. The rare variants from the filtered list from 196 unrelated patients and 193 controls were collapsed per gene and frequency count of the variants per gene was compared across cases and controls for significant difference. We only considered those genes whose count of variants were significantly higher in cases in comparison to the controls.

Resulting candidate genes: Using the complementary strategy of gene prioritization and mutation burden test, 61 candidate genes (44 genes from prioritization and 17 genes from burden analysis) were selected for further downstream analysis through TR approach. The resulting list of genes are categorized under the various signaling pathways that are relevant to BAV and presented in Figure 4.10.2. The selected genes are color coded in bold-black: higher priority, red:intermediate priority and black:normal priority. Next other strategies such as extensive literature search, CNV analysis etc were incorporated to select genes that can be appended to the existing list of targeted panel. Overall in total 147 genes to constitute the targeted resequencing panel out of which 61 genes were contributed through the prioritization and burden analysis. In this thesis we only present a subset of 22 genes out of 147 that explain their contribution towards BAV/TAA disease in the resequencing analysis.

Center	**City, Country**	**#**
Radboud University Medical Centre	Nijmegen, the Netherlands	27
APHP-Hopital Europeen Georges Pompidou	Paris, France	59
Erasmus University Medical Center	Rotterdam, the Netherlands	30
University of Luebeck	Luebeck, Germany	87
Institute for Clinical and Ex- perimental Medicine	Prague, Czech Republic	16
Sickkids Hospital	Toronto, Canada	62
Karolinska University Hospital, Karolinska Institutet	Stockholm, Sweden	156
Lviv National Medical University after Danylo Halytsky	Lviv, Ukraine	4
Total		441

Table 4.4: Patient cohort overview.



	Notch	Transcription factors	ECM	TGFbeta signaling	SMC apparatus	ROBO/SLIT	SEMAsignaling	RHO-GTPase	Other	WNT	p300	Integrin	Erk
Prioritization	LEF1	MEF2C	HAPLN1	**SMAD6**	MYH11	GRB7	ERBB2			FZD2	RXFP1	ITGA6	ELK3
Prioritization	AP2A1	SOX4	FURIN	KDR	MYLK	EPHA4	**NRP1**			WNT1			MAP3K3
Prioritization	NOTCH3	SOX5	**FN1**	PPP2R2B	CALD1	EPHA3	**NRP2**			FZD1			EGFR
Prioritization	NOTCH2	NFATC1	ELN	TGFB1	MYH6	EPHA7	**SEMA3A**						
Prioritization		IRX4	TIMP1	**TGFBR2**			PRKCH						
Prioritization			MMP15	PRKCB									
Prioritization			**MMP2**	BTC									
Burden	JAG2	IRX2		HOOK1				ARGHAP31	TACC2		EP300		
Burden								TTC3	DHX57				
Burden									ZNF106				
Burden									TJP1				
Burden									MRVI1				
Burden									KIAA1671				
Burden									BPIFB4				
Burden									KIDINS220				
Burden									TMTC4				
Burden									FAM157A				
Burden									PRSS38				

Figure 4.10.2: **Resulting list of candidate genes selected through prioritization and burden analysis.** The genes are categorized based on important biological pathways they are known to be associated with it. The genes are color coded based on the priority by which they were chosen for the targeted resequencing panel. Bold-black colour: High priority; Red-colour: Intermediate priority; Black colour: Normal priority

Gene	ID	Sex	BAV type	Nucleotide change	Protein change	Classification	MAF ExAC	CADD score
ACTA2	P11	M	NA	c.678A>C	p.Glu226Asp	Missense	1/121150	15.8
	P281	M	L-R	c.977C>A	p.Thr326Asn	Missense	16/121284	21.2
	C102	F	TAV	c.959C>T	p.Thr320Met	Missense	7/121328	19.1
ACVR1	P25	F	NA	c.636G>C	p.Glu212Asp	Missense	5/121074	18.7
	P286	M	R-N	c.718C>T	p.Arg240Cys	Missense	1/121392	19.8
	C172	M	TAV	c.1394C>T	p.Pro465Leu	Missense	7/121410	23.9
ELN	P311	F	L-R	c.1114_1125del	p.Ala372_Lys375del	In-frame deletion	Absent	/
	P349	M	NA	c.1114_1125del	p.Ala372_Lys375del	In-frame deletion	Absent	/
	P179	F	R-N	c.1421_1422insTCCTGGTGTCGGCGTGGC	p.Pro475_Ala480dup	In-frame duplication	Absent	/
	P381	F	L-R	c.1909G>A	p.Ala637Thr	Missense	1/17222	12.2
	C165	M	TAV	c.1021G>A	p.Val341Ile	Missense	4/121176	0
	C95	F	TAV	c.1883G>C	p.Gly628Ala	Missense	6/111550	11.8
FBN1	P117	F	NA	c.1158C>G	p.Asn386Lys	Missense	8/119050	13.8
	P126	M	NA	c.1472T>C	p.Val491Ala	Missense	Absent	16
	P227	F	L-R	c.1651G>A	p.Gly551Ser	Missense	1/121162	36
	P431	M	L-R	c.2115G>A	p.Ala705Ala	Splice site	Absent	11.5
	P287	M	R-N	c.2315A>G	p.Asn772Ser	Missense	Absent	23.7
	P174	M	R-N	c.3142A>G	p.Ile1048Val	Missense	Absent	18.8
	P28	M	NA	c.3382G>A	p.Val1128Ile	Missense	Absent	8.4
	P132	M	NA	c.3797A>T	p.Tyr1266Phe	Missense	12/121316	12.5
	P137	F	NA	c.4340T>C	p.Ile1447Thr	Missense	Absent	19
	P196	M	NA	c.4609G>T	p.Asp1537Tyr	Missense	Absent	22.1
	P426	M	NA	c.4727T>C	p.Met1576Thr	Missense	10/121398	6.5
	P117	F	NA	c.5123G>A	p.Gly1708Glu	Missense	1/121388	23.7
	P85	F	NA	c.6595G>A	p.Gly2199Ser	Missense	1/121280	36
	P332	M	L-R	c.6783A>C	p.Lys2261Asn	Missense	4/121108	16
	P150	F	R-N	c.7846A>G	p.Ile2616Val	Missense	8/121320	3

Table 4.5 continued from previous page

Gene	ID	Sex	BAV type	Nucleotide change	Protein change	Classifi-cation	MAF ExAC	CADD score
	P41	M	NA	c.8232G>C	p.Gln2744His	Missense	Absent	8.4
	C32	M	TAV	c.185G>A	p.Arg62His	Missense	7/119890	17
	C28	M	TAV	c.716T>C	p.Ile239Thr	Missense	2/121008	13.8
	C55	F	TAV	c.1118C>T	p.Ala373Val	Missense	1/121226	9.2
	C29	M	TAV	c.6163+2dupT	/	Splice site	Absent	/
	C167	M	TAV	c.6832C>G	p.Pro2278Ala	Missense	19/121266	25.7
FLNA	P209	F	L-R	c.2906T>C	p.Leu969Pro	Missense	4/86745	20.3
	P434	M	R-N	c.5908G>A	p.Asp1970Asn	Missense	Absent	23.5
	P51	M	NA	c.7172G>A	p.Arg2391His	Missense	7/86932	18.2
	C160	F	TAV	c.C901C>T	p.Arg301Trp	Missense	6/85694	11.6
	C132	F	TAV	c.1270A>G	p.Met424Val	Missense	2/85774	0.1
	C136	F	TAV	c.2738G>C	p.Gly913Ala	Missense	2/86788	11.8
	C55	F	TAV	c.3346G>C	p.Asp1116His	Missense	Absent	19.4
	C81	M	TAV	c.4520A>G	p.Gln1507Arg	Missense	Absent	18.2
	C167	M	TAV	c.4711G>A	p.Asp1571Asn	Missense	2/82272	35
GATA4	P339	M	L-R	c.142G>T	p.Val48Leu	Missense	Absent	0.2
	P370	M	NA	c.173G>T	p.Gly58Val	Missense	Absent	10.2
	P177	M	L-R	c.611A>G	p.Asn204Ser	Missense	Absent	5
	P29	M	NA	c.939G>T	p.Glu313Asp	Missense	8/120012	16.5
	P30	M	NA	c.939G>T	p.Glu313Asp	Missense	8/120012	16.5
	C155	M	TAV	c.175G>T	p.Ala59Ser	Missense	Absent	0
GATA5	P438	M	L-R	c.472C>T	p.Pro158Ser	Missense	Absent	15.4
	P98	F	NA	c.616G>A	p.Gly206Ser	Missense	0.00003	36
	C161	F	TAV	c.287C>G	p.Ala96Gly	Missense	Absent	3.9
	C104	M	TAV	c.395G>A	p.Arg132Gln	Missense	Absent	16.1
	C152	F	TAV	c.1153G>T	p.Ala385Ser	Missense	Absent	0
GATA6	P138	M	NA	c.148G>A	p.Gly50Arg	Missense	1/95966	18.8
	P395	F	NA	c.271C>T	p.Pro91Ser	Missense	6/73362	12
	P133	M	NA	c.706G>T	p.Gly236Cys	Missense	Absent	11.5

Table 4.5 continued from previous page

Gene	ID	Sex	BAV type	Nucleotide change	Protein change	Classifi-cation	MAF ExAC	CADD score
	P289	M	L-R	c.968_976delACCACCACC	p.His324_His326del	In-frame deletion	Absent	0
	P98	F	NA	c.1555A>G	p.Thr519Ala	Missense	3/121370	12.4
	C166	M	TAV	c.89G>A	p.Arg30Gln	Missense	1/106184	16.4
	C94	M	TAV	c.352C>T	p.Leu118Phe	Missense	6/103504	14.7
	C130	F	TAV	c.727G>T	p.Gly243Cys	Missense	Absent	9.1
MAT2A								
MATR3	P255	M	L-R	c.35G>A	p.Arg12Lys	Missense	Absent	16.4
MYH11	P377	F	L-R	c.2026C>T	p.Arg676Cys	Missense	70/121144	21.5
	P167	F	R-N	c.2026C>T	p.Arg676Cys	Missense	70/121144	21.5
	P181	M	R-N	c.2026C>T	p.Arg676Cys	Missense	70/121144	21.5
	P252	M	L-R	c.2026C>T	p.Arg676Cys	Missense	70/121144	21.5
	P372	M	L-R	c.2981T>A	p.Ile994Asn	Missense	Absent	12.4
	P153	M	L-R	c.3784A>G	p.Lys1262Glu	Missense	1/121400	29.4
	P252	M	L-R	c.3826A>G	p.Ser1276Gly	Missense	6/121368	15.4
	P160	M	L-R	c.3848C>T	p.Ala1283Val	Missense	6/121224	7.2
	P7	F	NA	c.3917C>A	p.Ala1306Asp	Missense	Absent	19.3
	P314	F	L-R	c.4531C>T	p.Arg1511Trp	Missense	7/121404	19.6
	P79	F	L-R	c.4624C>T	p.Arg1542Trp	Missense	15/121346	21.1
	P57	F	NA	c.4694C>T	p.Thr1565Met	Missense	90/121408	23.2
	P223	M	R-N	c.4694C>T	p.Thr1565Met	Missense	90/121408	23.2
	P88	M	NA	c.4681G>A	p.Ala1568Thr	Missense	15/121408	34
	P64	M	L-R	c.5247G>C	p.Glu1749Asp	Missense	66/90680	21.3
	P341	M	L-R	c.5294G>A	p.Arg1765Gln	Missense	25/116672	33
	P391	M	L-R	c.5687C>T	p.Ala1896Val	Missense	6/121208	25.9
	C110	F	TAV	c.33G>T	p.Glu11Asp	Missense	7/121350	11.6
	C23	F	TAV	c.1223T>C	p.Ile408Thr	Missense	1/121412	21
	C40	F	TAV	c.1934C>T	p.Ser645Leu	Missense	17/106978	24.8
	C81	M	TAV	c.1934C>T	p.Ser645Leu	Missense	17/106978	24.8
	C173	M	TAV	c.3430G>T	p.Ala1144Ser	Missense	1/121412	18.6
	C95	F	TAV	c.3583C>T	p.Arg1195Trp	Missense	20/121406	24.1

Table 4.5 continued from previous page

Gene	ID	Sex	BAV type	Nucleotide change	Protein change	Classifi-cation	MAF ExAC	CADD score
	C34	F	TAV	c.4599C>T	p.Asn1533Asn	Splice site	Absent	8.5
	C143	M	TAV	c.5687C>T	p.Ala1896Val	Missense	6/121208	25.9
NKX2-5	P30	M	NA	c.61G>C	p.Glu21Gln	Missense	92/114290	22.9
	P352	M	L-R	c.61G>C	p.Glu21Gln	Missense	92/114290	22.9
	P273	M	L-R	c.89C>A	p.Ala30Asp	Missense	Absent	16.1
	P157	F	R-N	c.358G>T	p.Val120Leu	Missense	Absent	4.8
	P257	F	NA	c.650G>A	p.Arg217Lys	Missense	21/92066	24.6
NOS3	P343	M	L-R	c.466G>A	p.Glu156Lys	Missense	26/94176	36
	P336	M	R-N	c.668A>G	p.Asn223Ser	Missense	2/112186	22.6
	P65	M	L-R	c.1267G>A	p.Ala423Thr	Missense	73/120868	25.1
	P175	F	R-N	c.2457C>G	p.Asp819Glu	Missense	5/11892	23.7
	P382	M	NA	c.2642C>T	p.Ala881Val	Missense	14/119328	22.4
	C19	M	TAV	c.466G>A	p.Glu156Lys	Missense	26/94176	36
	C6	M	TAV	c.638A>G	p.Asn213Ser	Missense	Absent	13.7
	C68	M	TAV	c.1267G>A	p.Ala423Thr	Missense	73/120868	25.1
	C85	M	TAV	c.2471C>T	p.Thr824Ile	Missense	1/11184	12.7
	C152	F	TAV	c.2546G>A	p.Arg849Gln	Missense	1/115870	36
	C151	M	TAV	c.2776_2776delinsCCA	p.Leu927Hisfs*32	Frameshift	1/84330	/
	C182	F	TAV	c.3589G>A	p.Gly1197Ser	Missense	3/110874	7.3
NOTCH1	P113	M	NA	c.983C>G	p.Thr328Ser	Missense	1/119274	18.3
	P373	F	L-R	c.1951G>A	p.Asp651Asn	Missense	Absent	14.7
	P344	F	L-R	c.2352C>T	p.Ser784Ser	Splice site	7/118434	9.1
	P423	M	L-R	c.4013C>T	p.Ala1338Val	Missense	Absent	19.1
	P155	M	R-N	c.4021G>A	p.Glu1341Lys	Missense	3/66948	7.7
	P128	M	NA	c.5047C>T	p.Arg1683Trp	Missense	1/119254	29.4
	P202	F	L-R	c.5167+3_5167+6del		Splice site	Absent	/
	P134	M	NA	c.5248G>A	p.Val1750Met	Missense	3/100758	17.3
	P420	M	L-R	c.5414T>C	p.Leu1805Pro	Missense	3/119106	22.1
	P106	M	NA	c.6413C>T	p.Pro2138Leu	Missense	1/114110	3.6
	C103	M	TAV	c.121A>G	p.Asn41Asp	Missense	Absent	12.8
	C45	M	TAV	c.800A>G	p.Lys267Arg	Missense	1/119480	11.1

Table 4.5 continued from previous page

Gene	ID	Sex	BAV type	Nucleotide change	Protein change	Classifi-cation	MAF ExAC	CADD score
	C162	M	TAV	c.2003C>T	p.Pro668Leu	Missense	4/118872	17.7
	C158	M	TAV	c.2003C>T	p.Pro668Leu	Missense	4/118872	17.7
	C54	M	TAV	c.5273G>A	p.Arg1758His	Missense	13/107714	27.1
	C101	M	TAV	c.7361A>G	p.His2454Arg	Missense	Absent	11.9
	C49	F	TAV	c.7372C>A	p.Pro2458Thr	Missense	2/78680	12.9
ROBO1	P344	F	L-R	c.153C>A	p.Asp51Glu	Missense	Absent	22
	P252	M	L-R	c.497C>G	p.Ala166Gly	Missense	Absent	21.9
	P5	M	NA	c.703G>A	p.Ala235Thr	Missense	5/118284	34
	P432	M	L-R	c.818T>C	p.Val273Ala	Missense	15/120622	23.1
	P374	M	R-N	c.979T>C	p.Ser327Pro	Missense	23/120206	21.2
	P328	M	NA	c.979T>C	p.Ser327Pro	Missense	23/120206	21.2
	P69	F	R-N	c.1432G>C	p.Ala478Pro	Missense	Absent	33
	P5	M	NA	c.1616A>G	p.Tyr539Cys	Missense	5/120482	15.9
	P191	M	NA	c.3007G>A	p.Asp1003Asn	Missense	3/120414	19
	P384	M	L-R	c.3259A>C	p.Met1087Leu	Missense	8/120756	7.1
	P348	M	NA	c.3472A>G	p.Ser1158Gly	Missense	Absent	16.1
	P34	M	NA	c.4821G>A	p.Met1607Ile	Missense	2/120590	19.4
	C99	M	TAV	c.394A>G	p.Ile132Val	Missense	Absent	18.9
	C105	M	TAV	c.508G>A	p.Asp170Asn	Missense	Absent	24.3
	C145	M	TAV	c.818T>C	p.Val273Ala	Missense	15/120622	23.1
	C60	M	TAV	c.2987C>T	p.Thr996Met	Missense	2/120118	17.8
	C158	M	TAV	c.3259A>C	p.Met1087Leu	Missense	8/120756	7.1
ROBO2	P48	F	NA	c.1238C>T	p.Thr413Ile	Missense	10/120348	14.2
	P423	M	L-R	c.1859C>T	p.Pro620Leu	Missense	Absent	21.4
	P364	F	NA	c.2018G>T	p.Arg673Leu	Missense	Absent	31
	P432	M	L-R	c.2897C>T	p.Thr966Met	Missense	5/120752	21.3
	P164	M	L-R	c.3229C>G	p.Pro1077Ala	Missense	6/120516	13.1
	P29	M	NA	c.3230C>A	p.Pro1077His	Missense	2/120524	17.2
	P180	F	L-R	c.3857G>T	p.Arg1286Leu	Missense	67/120712	27
	P250	F	L-R	c.3857G>T	p.Arg1286Leu	Missense	67/120712	27
	P168	M	R-N	c.4063C>T	p.Arg1355Cys	Missense	3/120728	15
	C90	F	TAV	c.406C>T	p.Arg136*	Nonsense	Absent	37

Table 4.5 continued from previous page

Gene	ID	Sex	BAV type	Nucleotide change	Protein change	Classifi-cation	MAF ExAC	CADD score
	C20	M	TAV	c.2018G>A	p.Arg673His	Missense	16/120314	32
	C125	F	TAV	c.2390G>A	p.Arg797Gln	Missense	6/120660	21
	C74	M	TAV	c.2902C>G	p.Leu968Val	Missense	Absent	17.7
	C5	F	TAV	c.3230C>A	p.Pro1077His	Missense	2/120524	17.2
SMAD3	C72	M	TAV	c.448T>C	p.Phe150Leu	Missense	2/121372	11.9
SMAD6	P128	M	NA	c.73_79del	p.Gly26_Ser27del	In-frame deletion	Absent	/
	P99	M	NA	c.454_461del	p.Gly166Valfs*23	Frameshift deletion	Absent	/
	P94	M	NA	c.715G>A	p.Val239Met	Missense	Absent	24.4
	P231	M	L-R	c.726del	p.Lys242Asnfs*300	Frameshift deletion	Absent	/
	P12	F	NA	c.770C>T	p.Pro257Leu	Missense	Absent	16
	P89	M	NA	c.812G>A	p.Gly271Glu	Missense	Absent	23.6
	P308	F	L-R	c.837C>A	p.Tyr279*	Nonsense	Absent	38
	P180	F	L-R	c.864C>G	p.Tyr288*	Nonsense	Absent	38
	P367	M	R-N	c.1216G>T	p.Gly406Cys	Missense	Absent	19.6
	P67	F	R-N	c.1224C>G	p.His408Gln	Missense	Absent	18.3
	P201	M	NA	c.1328G>A	p.Arg443His	Missense	1/106314	23.7
	C148	M	TAV	c.389C>T	p.Ser130Leu	Missense	Absent	12.8
TGFB2	P133	M	NA	c.1048C>T	p.Leu350Phe	Missense	Absent	18.8
TGFB3								
TGFBR1	P334	M	L-R	c.119T>A	p.Leu40His	Missense	Absent	9.5
	P105	M	NA	c.926C>T	p.Thr309Met	Missense	2/121364	24.9
TGFBR2	C72	M	TAV	c.1090C>T	p.Arg364Trp	Missense	6/120734	14.7

Table 4.5: Overview of the identified variants in the genes from the targeted gene panel and their phenotypic data

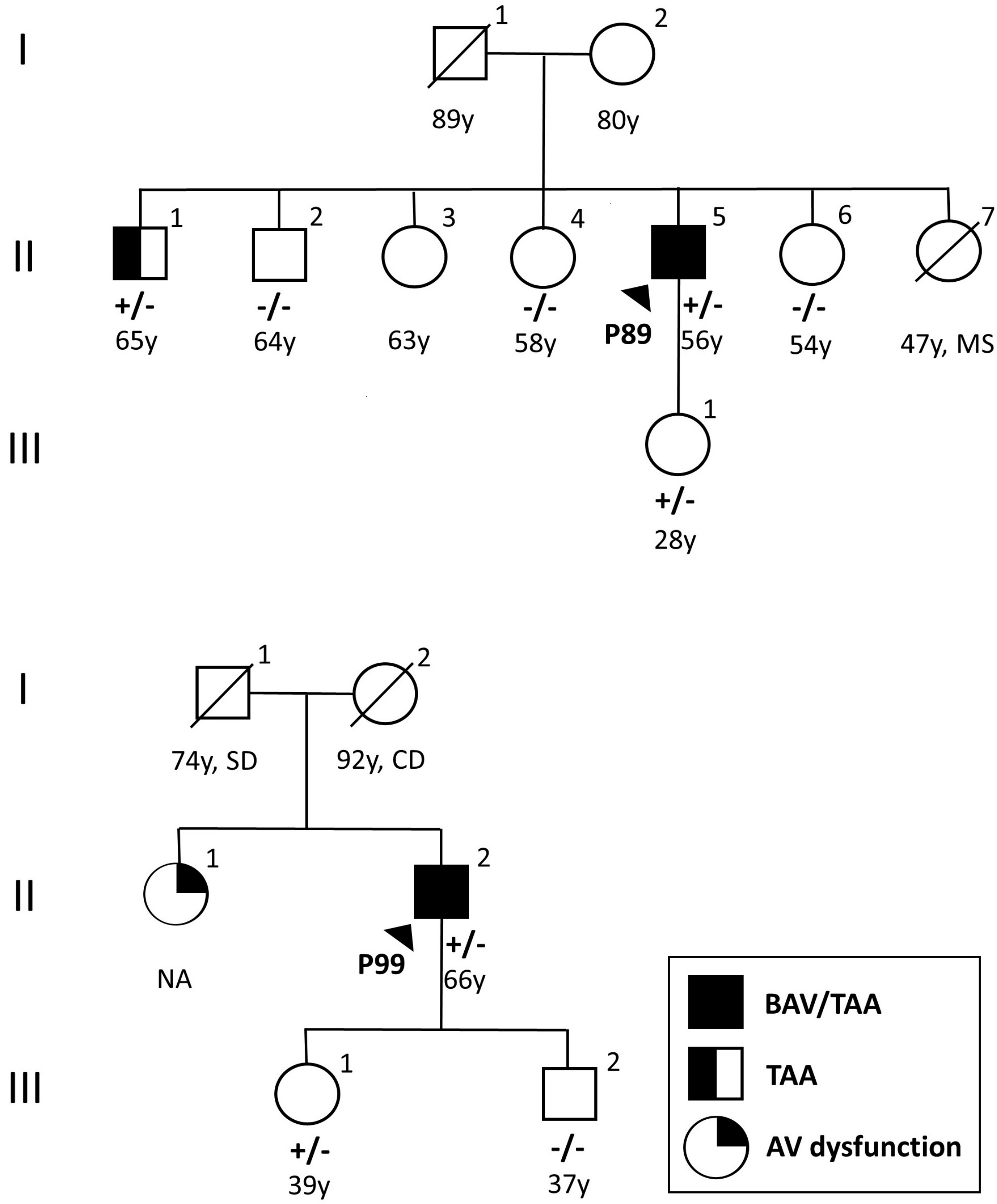


Figure 4.10.3: **Family pedigree for segregation analysis**

Statistical approach towards detection of CNVs

Chapter 5

varAmpliCNV: Analyzing Variance of Amplicons to detect CNVs in targeted NGS data.

> *When we try to pick out anything by itself, we find that it is bound fast by a thousand invisible cords that cannot be broken, to everything in the universe.*
>
> John Muir, 1869

Ajay Anand Kumar[1,2], Bart Loeys[1,2], Gerarda Van De Beek[1,2], Nils Peeters[1,2],
Wim Wuyts[1,2], Lut Van Laer[1,2], Maaike Alaerts[1,2] and Geert Vandeweyer[1,2]

[1]Department of Medical Genetics, University of Antwerp & Antwerp University Hospital, Antwerp, Belgium. [2]Biomedical Informatics research network Antwerp (biomina), University of Antwerp, Antwerp, Belgium.

The current work is in the stage of submission of the manuscript

The supplementary data present in this chapter can be accessed at this link:

- Supplementary file 1: `https://tinyurl.com/ybbb3smd`
- Supplementary file 2: `https://tinyurl.com/yddgyz6w`

5.1 Abstract

Introduction: Computational identification of copy number variants (CNVs) in resequencing data is a challenging task. Existing methods developed for detection of CNVs in NGS data (targeted, exome and whole genome) take into account various sources of variation and perform different normalization strategies to detect CNVs. However, their applicability and predictions are limited to the usage of specific enrichment protocols. Here, we introduce a novel tool named varAmpliCNV, which has been designed specifically for detecting CNVs from amplicon-based targeted resequencing data (Haloplex^{TM} enrichment protocol) in the absence of matched control samples.

Material & Methods: VarAmpliCNV incorporates three analysis steps. (1) **Read counts:** The depth of coverage signal in the target region is decomposed to read count per amplicon, by uniquely assigning reads to amplicons. (2) **Bias correction/Normalization:** Read counts are normalized by average depth and corrected for GC content per amplicon. Principal component analysis (PCA) and Metric Dimensional Scaling (MDS)-based variance control is applied to model the variability of amplicon read counts. (3) **CNV detection:** Log$_2$R ratio per sample is computed using a leave-one-out approach on a cohort scale, and finally, circular binary segmentation is used to detect CNVs. We used varAmpliCNV to analyze 167 samples screened with a Haloplex^{TM}-based panel including 30 genes. Nine samples containing MLPA-validated CNVs were included as positive controls. The same data were analyzed using 3 competing methods (ConVADING, ONCOCNV, DeCoN). Additionally we validated the performance on a large deafness panel of 145 genes run on 138 samples, containing 4 positive controls.

Results: VarAmpliCNV achieved higher sensitivity and specificity in comparison to three existing methods. These methods do not account for the specific Haloplex^{TM} amplicon-design information, leading to inflated read counts and reduced signal to noise ratios. Our PCA/MDS-based approach is useful for controlling the variance of amplicon-level normalized read counts at different cut-off levels, thereby giving robust predictions. Visualization of amplicons in the predicted CNV regions is included through plots as a downstream strategy to filter out false positive results.

5.2 Introduction

Copy number variants (CNVs) are a class of structural variants involving deletion or duplication of specific DNA-segments, leading to alterations in the number of copies of these segments present in the genome. The size of CNVs typically ranges from 50 to several thousand basepairs (bp), potentially including several genes [138]. CNVs are known to be associated with various diseases including congenital heart disease (CHD)[51], Parkinson [165], diabetes mellitus[93], autism [170, 191] and cancer[99].

Array-based comparative genomic hybridization (arrayCGH), fluorescence in situ hybridization (FISH) and SNP arrays have traditionally been used to detect somatic [218] and germline CNVs [191]. However, they are mainly suitable for the detection of large chromosomal aberration events (FISH: $>$100 kb; arrayCGH: $>$10Kb; SNP arrays: $\sim$10Kb) and suffer from poor sensitivity in detecting shorter CNVs (single exon deletion/amplification events). Furthermore, the size and breakpoint resolution for SNP arrays are correlated to the unequal probe density across the genome[191]. The advent of next generation sequencing (NGS) approaches has promised to detect CNVs with far greater resolution in comparison to the traditional methodologies. Whole-genome sequencing (WGS) can examine sequence and structural variation present in both coding and non-coding regions of the genome. Due to the high cost of WGS, in practice more cost effective methods, such as whole exome sequencing (WES) and targeted resequencing (TR) on customized gene panels, are preferred to screen large sample cohorts for mutations. WES and WGS data can comprehensively be analyzed to detect CNVs, while TR data pose extra complications. But in a clinical diagnostic setting, when candidate genes for the disease are known, TR is preferred because it overcomes constraints of sequencing cost per patient, need for time bound results and high depth of coverage (DOC).

When using TR, the choice of capture protocol for the customized gene panel is important for effective detection of CNVs. There are two main categories of enrichment protocols to capture a given region of interest (ROI): (a) amplicon-based and (b) hybridization capture-based. The amplicon-based technologies (e.g. Haloplex$^{\text{TM}}$, AmpliSeq$^{\text{TM}}$) use oligonucleotide as PCR primers to capture the target ROIs. Specifically in the Haloplex$^{\text{TM}}$ method the genomic DNA is fragmented with restriction enzymes and subsequently oligonucleotides complimentary to the 5′- and 3′-ends of each fragment are used as PCR primers for amplifications. The hybridization capture based technologies (e.g. SeqCap$^{\text{TM}}$, SureSelect$^{\text{TM}}$) use sonication-based fragmentation to shear the genomic DNA and generate random size DNA fragments. Next, specific oligonucleotide probes are hybridized and used to capture the target ROIs. WES data obtained with these four categorizations enrichment protocols were assessed regarding their performance in CNV calling in exome sequencing using VarScan2 [105] compared against SNP arrays. It was found that all these technologies were highly concordant and could be used equally well for detection of copy number gains and losses. However, the comparison also showed that capture-based assays have advantage over amplicon-based assays with respect to coverage uniformity and enrichment library complexity, though amplicon-based assays are preferred in the laboratory for their simplicity in sample preparation. The assessment of CNV detection was done on the cancer related WES data but applying the same on TR panels with smaller ROIs could help in differentiating the performance of these enrichment assays.

Computational detection of CNVs from NGS data is a challenging task. There are many new methods being developed and applied on a varied range of datasets to identify CNVs [17, 26, 67, 95, 129, 200, 202, 229]. These methods can be categorized into five different strategies: (a) read depth (RD), (b) paired-end (PEM), (c) split read (SR), (d) de novo assembly and e) combinations of any of the above approaches[229]. Among these strategies only RD-based approaches can be successfully applied to WES or TR data while all are applicable to WGS data[202]. The main reason being that the ROI in WES and TR is only a percentage of that in WGS. Hence, capturing the re-arrangements using PEM/SR is more effective in WGS than WES/TR data. Additionally, RD-based approaches incorporate the counting of number of reads aligned to a given ROI. This has been empirically determined to be proportional to the number of copies of genomic segments, which helps in directly quantifying the CNVs with respect to read depth.

Computational detection of CNVs through RD requires normalization of the input data such that variability of RD is minimized, followed by detection of CNVs by comparing to control samples processed in a similar way. The normalization procedure includes accounting for biases associated to the enrichment protocol, non-uniform depth of coverage across the ROI and sequence properties such as local GC content and presence of repetitive elements. Most of the existing methods take these biases into account in their analysis pipelines. For example, ONCOCNV [26] incorporates multi factor normalization to detect CNVs on amplicon sequencing based tumor data. In their approach they normalize the tumor samples for potential enrichment biases and noises, and then perform principal component analysis (PCA) on a set of normal controls to extract the baseline coverage. Subsequently, segmenting the logarithmic ratio between the tumor and normal samples gives the putative CNVs. ONCOCNV was developed and tested on samples subjected to amplicon-based enrichment (AmpliSeq$^{\mathrm{TM}}$). The read counting is done at the amplicon level in the sense that each read is assigned to the amplicon it overlaps most with. In case of overlapping amplicons, they are merged if overlapping more than 75%. Similarly, CoNVADING [95] was developed specifically for hybridization-based (SureSelect$^{\mathrm{TM}}$) targeted resequencing data to detect single exon copy number events. The normalization procedure involves two stages where in a first step the data is normalized using all targets within the sample and a second step uses all targets within the gene. Additionally, it incorporates stringent quality control (QC) metrics to select the most informative control samples such that distribution patterns of read depth are highly similar between query and control samples used to compute the coverage ratio. Another tool called DECoN [67] adapts the ExomeDepth [171] package, which internally fits a beta-binomial model to describe the read distribution of the samples to detect exon copy number variations. Originally it was designed and validated on hybridization-based enrichment protocols.

Many of these tools give robust performance on their own benchmark dataset, sequenced using a specific enrichment protocol, with high sensitivity and specificity. However, their specificity declines dramatically when they are applied on gene panels enriched using different technologies, leading to a large number of false positive CNV calls (FPs). The reason for this drop in the performance can be attributed to the protocol-specific internal design pattern of enrichment probes [202]. For example, CoNVADING was developed and tested for the SureSelect$^{\mathrm{TM}}$ enrichment protocol, which is a hybridization-based capture technique, resulting in randomly fragmented DNA and bell-shaped RD profiles around the targets. For ONCOCNV, the samples were subjected to sequencing after enrichment using the AmpliSeq$^{\mathrm{TM}}$ protocol. This protocol involves PCR-based amplification with uniform amplicon length and limited overlap. To our knowledge however, currently no method exists which can be applied directly on HaloPlex$^{\mathrm{TM}}$-based amplicon sequencing data, which follows a hybrid enrichment strategy. First, genomic DNA undergoes restriction-based (non-random) fragmentation, followed by hybridization-based capture and PCR amplification of the fragments. This approach results in a design pattern of multiple overlapping amplicons of highly variable length. The resulting RD profiles show a wide range in coverage, and complex overlapping patterns makes the methodology of ONCOCNV unsuitable for HaloPlex$^{\mathrm{TM}}$ data.

Hence, we introduce a novel tool called varAmpliCNV, specifically designed to detect CNVs in HaloPlex$^{\mathrm{TM}}$ enriched panel data by analyzing the variance of depth of coverage of individual amplicons. The internal design principle of varAmpliCNV harnesses the amplicon design information and accounts for the overall RD variability using PCA and metric multi-dimensional scaling (MDS). VarAmpliCNV is fast and scalable in comparison to existing methods and can predict CNVs in extensive gene panels. Finally, varAmpliCNV provides visualization of each predicted copy number segment, by plotting RD profiles of individual amplicons in the context of the amplicon design pattern. This visualization serves as a post-hoc filter for pruning out false positive CNV calls. The tool is accessible at : `https://bitbucket.org/aakumar/varamplicnv`

5.3 Materials & Methods

5.3.1 Validation sets

TAAD panel: Targeted resequencing of a panel of 30 thoracic aortic aneurysm and dissection (TAAD) genes using HaloPlex^{TM} enrichment was performed on 167 samples. The samples were divided into five batches according to different sequencing experiments. Nine samples contained a CNV validated in our laboratory using MLPA and MAQ assays. These include two full gene duplications and a single exon duplication in the *MYH11* gene, four multiple exon deletions, amplifications and a single full gene amplification of the *FBN1* gene and one two-exon duplication in the *TGFBR2* gene and a single exon duplication in the *MYH11* gene (see supplementary file 1: sheet 8). The first analysis was done blindfolded with regard to information about known CNVs and corresponding samples. All samples were analyzed using varAmpliCNV and three other competing methods. Each batch was analyzed independently to avoid any batch specific biases in the analysis.

Deafness Panel: Targeted resequencing of a panel constituting of 145 genes involved in deafness (see supplementary file 2: sheet 5) using HaloPlex^{TM} enrichment was performed on 138 samples divided into four batches. Four samples across these batches contained an arrayCGH validated CNV. These include full gene deletions of *OTOA*, *POU3F4*, *EYA4* and *EYA1*. The main objective of incorporating this validation data set was to validate the thresholds for deletion and amplification that were derived from the TAAD panel.

5.3.2 VarAmpliCNV workflow

VarAmpliCNV analysis consists of five stages, shown in Figure 5.3.1: (a) Processing the input BAM files using the amplicon design to obtain read counts per amplicon. Additionally sample specific QC metrics are applied to filter out samples with low average coverage (b) normalizing the read counts for enrichment biases (c) controlling variance using PCA or MDS and subsequent Log$_2$R computation (d) segmentation of the Log$_2$R profile and filtering CNV segments on QC metrics and (e) annotation of predicted CNV segments and visualization.

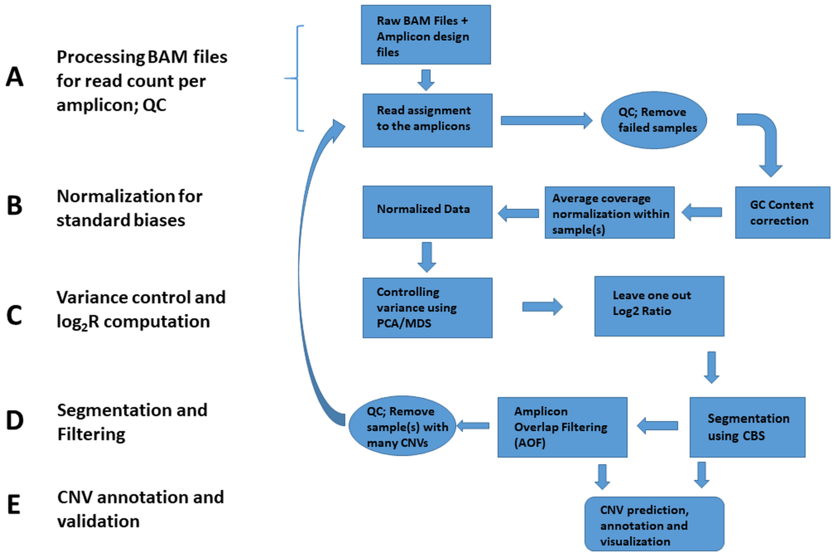

A
Processing BAM files for read count per amplicon; QC
B
Normalization for standard biases
C
Variance control and log2R computation
D
Segmentation and Filtering
E
CNV annotation and validation
Raw BAM Files + Amplicon design files
Read assignment to the amplicons
QC; Remove failed samples
GC Content correction
Average coverage normalization within sample(s)
Normalized Data
Controlling variance using PCA/MDS
Leave one out Log2 Ratio
Segmentation using CBS
Amplicon Overlap Filtering (AOF)
QC; Remove sample(s) with many CNVs
CNV prediction, annotation and visualization

Figure 5.3.1: **Workflow of varAmpliCNV:** (A) BAM files are processed to obtain read counts per amplicon. Subsequently, QC metrics are applied to remove bad quality samples from the analysis. (B) Read counts are normalized for average coverage and GC corrected. Normalization is performed separately for autosomal and sex chromosome targets. (C) Remaining variance is reduced using PCA/MDS, followed by a leave-one-out approach to compute Log_2R ratios per amplicon. (D) Detection of CNVs using CBS, optionally followed by Amplicon Overlap Filtering. (E) Annotation and visualization of CNV segments.

5.3.3 Input files

VarAmpliCNV uses BAM files and a BED file. The BAM files were obtained from raw fastq reads using an in-house pipeline customized for Haloplex$^{\text{TM}}$ enrichment data (Supplementary Figure S1 of [176]). The BED file contains the amplicon design, in the form of individual amplicon coordinates spanning a given region of interest (ROI). In our case the ROI was the coding region of 30 genes related to thoracic aortic aneurysm and dissection (TAAD) or 145 genes related to deafness, including 51 base pairs flanking either side of the exons. A total of 4701 amplicons were present in the TAAD panel BED file and 37,383 amplicons in the deafness panel BED file to cover the full ROI. The design file was sorted per chromosome and duplicate coordinates were removed. A representative visualization of the structure of a design file is present in Supplementary Figure 5.8.1, illustrating that amplicons have an overlapping structure and non-uniform distribution of amplicon lengths.

5.3.4 Read counting

Read counting is done at the amplicon level, using the PySAM package in python `https://github.com/pysam-developers/pysam`. The processing step involves unique assignment of reads to amplicons by exactly matching the respective start and end coordinates. Our approach removes the impact of mutual dependency due to overlap between the amplicons (see the information flow in the dependency model in Supplementary Figure 5.8.1 B). Consequently, unstable amplicons will not impact the signal of overlapping amplicons. Furthermore, it reduces the noise arising from unassigned reads, which are typically artefacts from aspecific amplifications. The read count is stored as a read count matrix and by convention we represent the amplicons as rows and samples as columns. Here, each amplicon is considered as an independent data point, defined by the start coordinate.

5.3.5 Quality control

Since for both panels the targeted resequencing was performed at 4000x, we expect for each sample to have an average coverage of at least 100 reads per amplicon . We removed all samples from the analysis that did not meet this criterion. Similarly, amplicons not having any assigned reads across all samples were pruned out from the read count matrix. Next, an additional sample specific QC metric was formulated related to predicted CNV segments, discarding samples containing a significantly higher number of CNVs than generally expected. Significance was evaluated using a student t-test, comparing the CNV count of each sample against the remainder of the batch.

5.3.6 Read count normalization

The read count data is corrected for inherent biases associated to local GC-content of the amplicons using Loess based linear regression [18, 26]. The GC corrected matrix was separated for autosomal and sex chromosomes, and normalized for average coverage of the sample. This is accomplished by dividing the read counts of each amplicon by the average coverage within a given sample.

5.3.7 Controlling variance

The most fundamental problem inherent to the existing methods is to control the variance of coverage in the targeted regions. We deal with this issue by formulating an objective function which can transform the normalized read count matrix in such a way that variance is minimized. We accomplish this primarily by principal component analysis (PCA). Using PCA we estimate the orthogonal principal components (PCs) and arrange them according to the proportion of variance they explain. We chose to remove those PCs that account for approximately 80% of the variance (see results for evaluation of this choice). This reverse denoising step is computed using equation 5.15 of supplementary data section 5.9. Since PCA can become computationally intensive as the number of amplicons increases, we implemented an alternative approach based on metric multi-dimension scaling (MDS), providing identical results with faster execution time. To obtain identical results with MDS and PCA, it is required that the Euclidian distance measure between the data points (amplicon with read counts) is used. Using the Euclidean distance is apt for our data because it is non-spatial and does not represent any three dimensional coordinate system where usage of different distance measures between the data points can affect the results. The mathematical details regarding the implementation can be found in section 5.9 of the supplementary data.

5.3.8 Computation of Ratio score (Log_2R)

All existing methods require some control dataset for comparing the normalized RD of a sample using ratio scores. The logarithm of this ratio score or Log_2R gives a distinctive negative value for a deletion and a positive value for an amplification. If a matched control sample, as evident in a tumor/normal pair, is not available, it can be simulated by pooling normal samples. In our case, we follow a leave-one-out approach to select the reference samples. This means that for each amplicon the denominator of the ratio of the normalized read counts is computed by taking the average normalized read count of the remaining n-1 samples (where n is total number of samples passing QC present in a given experiment). If any samples are discarded after CNV calling (due to large number of CNVs per sample), Log_2R calculation and segmentation is repeated to exclude impact of low quality samples on the reference.

5.3.9 CNV calling using segmentation and filtering

Segmentation is applied directly on Log_2R values to detect change point events. For varAmpliCNV we incorporate circular binary segmentation (CBS) [163] available from the DNACopy package [189] in R. We provide the user a choice to select: (a) direct segmentation (DS) using CBS or (b) applying a post processing filtering on CBS-segmented CNVs, called amplicon overlap filtering (DS-AOF). For the DS method the CBS algorithm with default setting is used which eventually results in a list of putative CNV segments. Segments with Log_2R values below -0.5 are reported as deletions and those with values above 0.5 are reported as duplications. Additionally, we discard segments covered by less than 10 amplicons and having a standard deviation greater than 1. For the DS-AOF approach information related to the overlapping structure of amplicons is used (see section 5.8.1 of supplementary data). AOF utilizes this information to recalculate average logarithmic ratios of CNVs predicted by DS. The recalculated segment Log_2R is a weighted average, where the Log_2R of each amplicon overlapping the segment contributes relative to the amount of overlapping positions. The conceptual details of this approach can be found in supplementary data, section 4.

5.3.10 Annotation and validation of CNV segments

The output format is a tab-separated list of CNV segments with sample names, coordinates and summary statistics. Segments are annotated with gene names, exon numbers, and number of involved amplicons. Additionally, each CNV segment is visualized through plots, presenting RD and gene and amplicon structure.

In the TAAD panel analysis, all CNV segments predicted by varAmpliCNV and passing filtering on SD (≤ 1) and amplicon (≥ 10) were experimentally validated using MLPA (MRC Holland) and MAQ assay (Agilent) protocols according to manufacturers' instructions. Similarly, congruent CNVs predicted by at least two competing methods and missing from varAmpliCNV were validated using MLPA/MAQ (see supplementary file 1: sheet 6). In total, 23 putative CNVs were experimentally evaluated, but none were confirmed as true CNVs. Based on the validation results, we determined the threshold for an optimal sensitivity/specificity balance.

5.3.11 Comparison with ONCOCNV, CoNVADING, DECoN (TAAD panel only)

We compared varAmpliCNV to three existing methods: ONCOCNV, CoNVADING and DECoN. We used these tools with their recommended default settings. Since our exclusion of bad quality samples is inherent to the design principle of varAmpliCNV and depends on the read counts per amplicon, we cannot apply the same QC metric to filter out bad samples from the competing methods. We applied the same leave-one-out principle to define a reference set for ONCOCNV. For CoNVADING and DECoN we selected sample pooling to create a reference set.

5.3.12 Computing optimal thresholds (TAAD panel only)

The initial analysis with all the tools was made blindfolded and once the CNV predictions were made, the TPs were revealed. For sensitivity and specificity analysis default parameter settings of the competing methods were used. For varAmpliCNV, which is based on a two step prediction strategy (PCA/MDS and DS/DS-AOF) we define three thresholds respectively that can maximize its performance. These are the detection threshold (DT), the segmentation threshold (ST) and the boundary threshold (BT). The DT is defined as the optimal cut-off of the percentage of variance that needs to be removed by principle component analysis, such that the maximum number of TPs can be detected. DT thus reflects the experiment specific number of principal components removed, using a stable metric. We define ST as a primary cut-off for the average normalized Log_2R ratio of a CBS-generated segment above (for amplification) or below (for deletion) which it is predicted to be a CNV. STs +0.5 and -0.5 and +0.1 and -0.1 are used in our analyses. Finally, the BT is defined as the decision boundary values for the average Log_2R values (DS) or recalculated Log_2R values (DS-AOF) at which the number of TPs and FPs are maximized and minimized respectively.

Since initially the analysis was blindfolded the analysis was repeated by removing 0 to 4 PCs (see supplementary file 1: sheet 7) thereby removing a proportion of variance from 56% to 96% across five batches. Analyzing this result helped in establishing the optimal value for DT such that all TPs are detected (sensitivity analysis). For segmentation using CBS the performance was analyzed by initializing the values for ST with -0.5,+0.5 and -0.10 +0.10 for deletions and amplifications respectively. Finally for all the batches it required to formulate optimal decision boundary (BT value) such that detection of true amplification and deletion related events is maximized.

5.4 Results

5.4.1 TAAD panel data analysis

5.4.1.1 Sample selection

167 samples, corresponding to 5 experimental batches, were processed to obtain amplicon-based read counts. Samples having average read counts less than 100 were not included in the analysis. Six, one, zero, one and two samples were excluded from the different batches respectively, making the total number of samples 157. Each batch was analyzed independently and the very first analysis was done blind-folded with regard to sample id and prior knowledge on validated CNVs.

5.4.1.2 Effect of GC content

GC correction is an important step in the overall normalization procedure and has been widely adopted in existing methods. In our analysis pipeline, we did not find any significant correlation between GC content and read depth. From supplementary Figure 5.8.2 and 5.8.3 it can be seen that for all batches the average correlation is almost zero for autosomal target regions and approximately 0.20 for sex chromosomal targets. However, we still correct for this effect using Loess based measure, as correlation might be higher in other target regions.

5.4.1.3 PCA/MDS based normalization and usage of Amplicon Overlap Filtering

The core of the varAmpliCNV pipeline is controlling the variance present in the depth of coverage of TR panel data. We account for this by performing PCA/MDS, with removal of an optimal number of principal components. We strategized to remove the number of PCs that most closely corresponds to a removal of 80% of the variance. The number of removed PCs is batch specific, since each of them was analyzed independently. We tried other fractions of variance as well and our choice of using approximately 80% variance removal as optimal DT can be deduced from supplementary file 1: sheet 7 for respective batches. It can be seen that all the true positives were correctly retained in the data when we removed approximately 80% of variance. Additionally, in the analysis pipeline we accounted for removal of samples containing an excessive number of CNVs. We found that for the first and second batch there were one (s25) and two samples (s14 and s25) respectively, behaving aberrantly and containing an unexpectedly high number of CNVs. After removal of these samples the analysis was repeated from the beginning.

The proportion of variance explained by the removal of number of PCs for each batch was 82.22%, 85.19%, 77.21%, 0% (no PCs were removed in this case) and 78.17% for autosomal targets. Next, it is important to deduce a threshold (BT) for the (DS or AOF-corrected) segment Log_2R, resulting in optimal sensitivity and specificity. Details of all CNV segments corresponding to varying numbers of removed PCs can be found in supplementary file 1: sheet 1-5 for all the five batches. CNV segments enclosed with green rectangles in this supplementary file are presented in Figure 5.4.1 and Figure 5.4.2. Figure 5.4.1 describes the results obtained with DS by the average segment Log_2R of the predicted CNV segments, per batch. It can be seen that by choosing the BT value of +0.51 for duplications and -0.61 for deletions we retain all the nine TPs (solid blue colored data points), while five FPs (unfilled data points) passed through.

Application of AOF yields comparatively better results, by removing more FPs without compromising sensitivity. Figure 5.4.2 describes the analysis result. It can be seen that the predicted CNV segment scores get altered and hence the cut-off threshold also gets changed. The new BT values for duplication and deletion events, can be set to a value greater than +0.38 and less than -0.50 respectively. With these settings all the

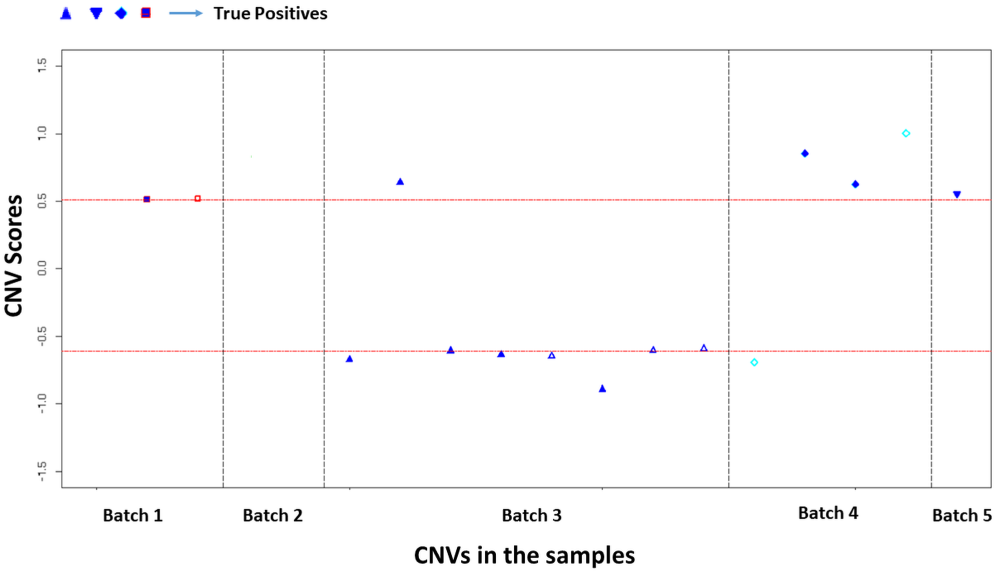


Figure 5.4.1: **CNV evaluation using direct segmentation (DS) approach:** The X-axis represents the CNVs present in the samples with Log_2R values above 0.5 or below -0.5 (segmentation threshold or ST) for each of the 5 batches (separated by vertical lines). The Y-axis corresponds to the average Log_2R value of predicted CNV segments, given by the CBS algorithm. The boundary threshold (BT) values (horizontal red dotted lines) for duplications and deletions are +0.51 and -0.61 respectively in order to maximize the detection of all 9 true positive CNVs and minimize the number of false positives. The proportion of variance removed for Batch 1, Batch 2, Batch 3, Batch 4 and Batch 5 samples are 82.22%, 86.74%, 77.21%, 0% and 78.17% for autosomal genes.

TPs were retained and only one FP passed through. More specifically, AOF filters out FPs present in Batch 1, Batch 2 and Batch 3, increasing overall specificity.

In the above analysis, the ST value of -0.50 and 0.50 was used to generate putative CNV segments that can be validated using experimental protocols. To differentiate the performance of DS and AOF approach it is necessary to try other values of ST, as more false positive AOF results might have uncorrected Log_2R values below the original ST values. We therefore repeated the analysis with a less stringent ST value of -0.10 and +0.10 and calculated the AOF values (Figure 5.4.3, supplementary file 1:sheet 3) As expected, more CNV segments obtained from DS approach pass this new threshold (see Supplementary file 1:sheet 3). We re-determined the optimal BT decision boundary based on the 374 AOF corrected segment scores, as -0.53 for deletions and +0.38 for duplications. With these new BT values, we observed a slightly reduced performance in comparison to the previous analysis by retaining two additional FP segments in Batch 4

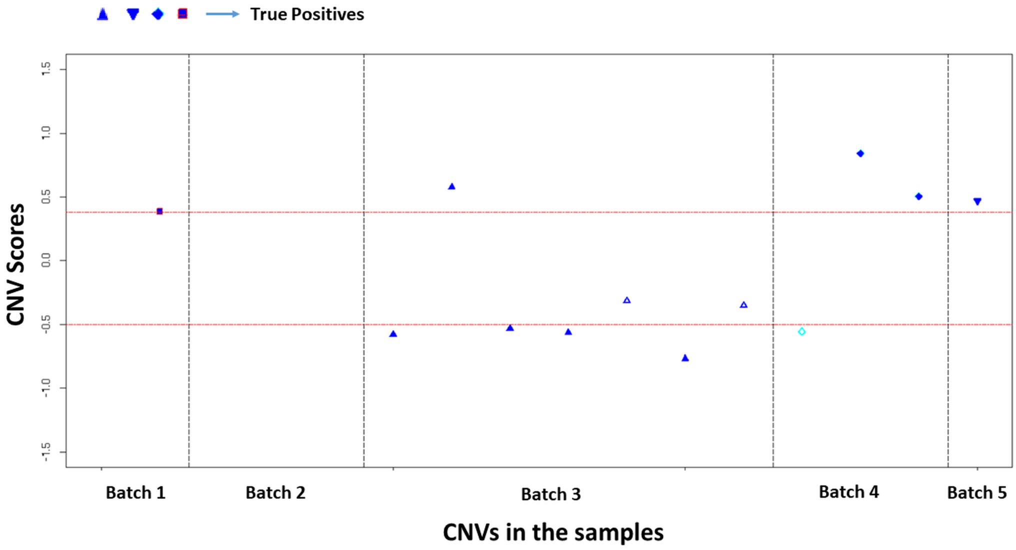


Figure 5.4.2: **Evaluation of CNVs using amplicon overlap filtering approach (AOF):** The X-axis represents the CNVs present in the samples with Log_2R values above 0.5 or below -0.5 (segmentation threshold or ST) for each of the 5 batches (separated by vertical lines). The Y-axis corresponds to the AOF-corrected average Log_2R value of predicted CNV segments. The BT values (horizontal red dotted lines) for duplications and deletions are now set at +0.38 and -0.50 respectively in order to maximize the detection of all 9 true positive CNVs and minimize the number of false positives. The proportion of variance removed is identical to Figure 5.4.1

5.4.2 Application of DS-AOF on deafness panel data

From the TAAD panel analysis we derived the optimal values for DT (approximately 80 % variance removal from both autosomal and sex chromosomal analysis), ST (≤ -0.5 for deletion and $\geq +0.5$ for amplification) and BT (obtained after applying AOF; ≤ -0.50 and ≥ 0.38 for amplification) as shown in Figure 5.4.2. Based on these threshold values we predicted CNV segments were predicted for a second gene panel including 145 genes associated to deafness, run on 138 samples grouped into four batches.

By applying DS-AOF we predicted 19 CNV segments as described in supplementary file 2: sheet 5. Three out of four arrayCGH validated CNVs were among this list of 19 CNVs and thus successfully predicted. One deletion present in Batch 4 was not detected with the current threshold settings. The remaining 15/19 CNV segments were inspected using the commercial program SeqPilot (JSI medical systems) and showed no indication of deletions or amplifications. We thus speculate they are more likely to be FPs. The overall results are summarized in Figure 5.4.4 along with the details presented in supplementary file 2.

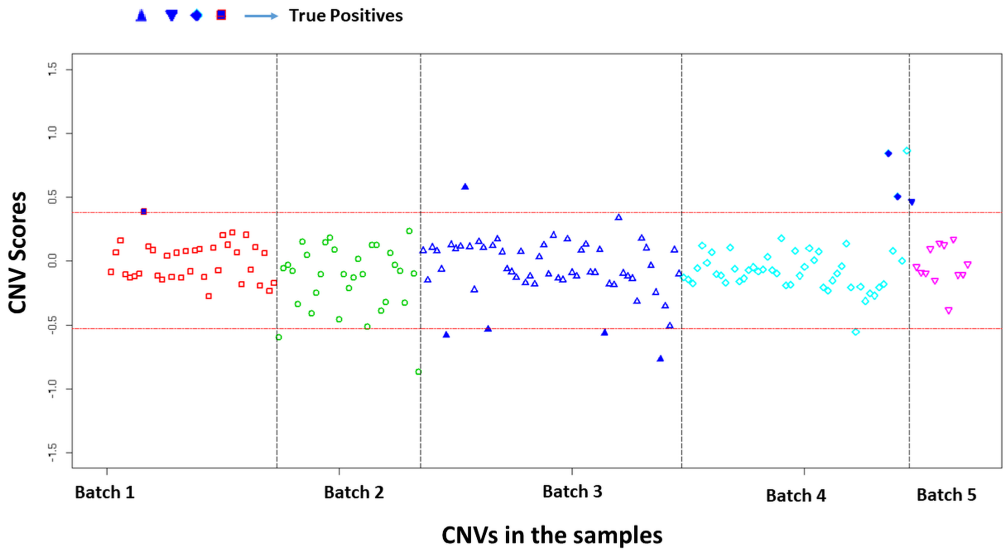


Figure 5.4.3: **Evaluation of CNVs using amplicon overlap filtering approach (AOF) and lenient pre-filtering:** The X-axis represent the CNVs present in the samples with Log_2R values above 0.1 or below -0.1 (segmentation threshold or ST) for each of the 5 batches (separated by vertical lines). The Y-axis corresponds to the AOF-corrected average Log_2R value of predicted CNV segments. The BT value boundaries (horizontal red dotted lines) for duplications and deletions are now set at +0.38 and -0.53 respectively in order to maximize the detection of all 9 true positive CNVs and minimize the number of false positives. The proportion of variance removed is identical to Figure 5.4.1

5.5 Performance comparison with ONCOCNV, CoNVADING, DECoN (TAAD panel only)

The predictive performance of varAmpliCNV, ONCOCNV, CoNVADING and DECoN was evaluated on 9 MLPA and MAQ validated CNVs (see supplementary file 1: sheet 9). The CNVs that were predicted by the four methods were validated using MLPA and MAQ assays. Since apart from varAmpliCNV, all methods predicted many CNVs, individual validation was practically infeasible. Hence, for the competing methods, only a subset of CNVs congruent between at least two methods were validated.

For varAmpliCNV, we validated all predicted CNVs under ST value that corresponds to a Log_2R ratio below -0.5 or above +0.5. We determined the sensitivity and specificity of these tools by calculating the number of true positives (TPs), false positives (FPs), false negatives (FNs) and true negatives (TNs). When the known validated CNVs were correctly predicted in the blindfolded analysis, these are classified as true positives (TPs). Predicted CNVs that could not be confirmed by either MLPA or MAQ assays, are classified as false positives (FPs). Subsequently, known CNVs that were not predicted

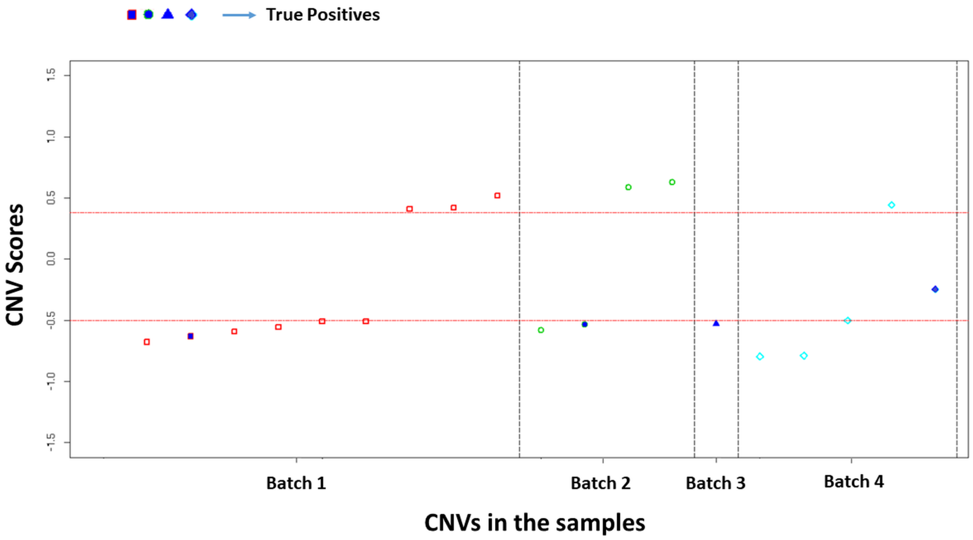


Figure 5.4.4: **Evaluation of CNVs using amplicon overlap filtering approach (AOF):** The X-axis represents the CNVs present in the samples with Log_2R values above 0.5 or below -0.5 (segmentation threshold or ST) for each of the 5 batches (separated by vertical lines). The Y-axis corresponds to the AOF-corrected average Log_2R value of predicted CNV segments. The BT values (horizontal red dotted lines) for duplications and deletions are now set at +0.38 and -0.50 respectively in order to maximize the detection of all 9 true positive CNVs and minimize the number of false positives. The proportion of variance removed is identical to Figure 5.4.1 . The proportion of variance removed for Batch 1, Batch 2, Batch 3 and Batch 4 samples are Auto=0%; Sex=78%, Auto=82.43%; Sex=81.38%, Auto=79.14%, Sex= 80.48% and Auto=84.71%; Sex=80.69% for autosomal and sex chromosomes respectively

are classified as false negatives (FNs). Finally, all given autosomal and sex chromosomal targets (each exon is defined as a target) for which no CNV calls were made by any of the methods were marked as true negatives (TNs). Sensitivity was calculated as #TP/(#TP + # FN) and specificity as #TN/(#TN + #FP). The analysis results of varAmpliCNV and competing methods are presented in Table 5.1. For varAmpliCNV, one CNV detected in batch 4 sample turned out to be false positive (also shown in Figure 5.4.2 but all the known TPs were predicted correctly. ONCOCNV could predict 6/9 known TPs and 6 additional CNVs were not confirmed by the validation assay. For CoNVaDING only 3 of the known TPs got predicted correctly and additionally a long list of potential FPs was also generated. Finally, DECoN could predict all the 9 confirmed TPs but also gives a long list of potential FPs. The list of all the CNVs analyzed in this performance analysis is presented in Supplementary File: sheet 6.

	varAmpliCNV (with AOF)	**ONCOCNV**	**CoNVaDING**	**DECoN**
True positives (TP)	9	6	3	9
False positives (FP)	1	6	6	9
False negatives (FN)	0	3	6	0
True Negatives (TN)	458	255	175	403
Total CNV calls	10	41	445	291
Sensitivity	100%	66.67%	33.33%	100%
Specificity	99.78%	97.77%	96.68%	97.81%

Table 5.1: Performance comparison of varAmpliCNV with three competing methods. The counts for TPs and FPs for varAmpliCNV were obtained using the boundary threshold (BT) value of +0.38 (for amplification) and -0.50 (for deletion) as shown in Figure 5.4.2. For ONCOCNV, CoNVaDING and DECoN the analysis was done with their respective default settings. For the competing methods, sensitivity and specificity were calculated for validated CNVs only.

5.6 Discussion

CNV detection in TR data is a challenging task and it is limited due to variability in the high read depth coverage, biases associated with specific enrichment protocols and lack of matched controls especially in clinical diagnostic settings. We have developed varAmpliCNV, a novel approach towards predicting CNV segments from TR sequencing data, designed specifically to handle amplicon-based sequencing NGS data enriched by Haloplex^{TM} technologies. The corresponding enrichment protocol results in a unique overlapping structure of the amplicons and contributes towards the variability in read depth associated with it. VarAmpliCNV incorporates a unique two step strategy to model the flow of information from read depth count to the genomic position that is mediated by the overlapping amplicons. In the first step reads are uniquely assigned to the amplicons. Then it implements PCA/MDS based methodology as normalization to control the variance of read depth associated with the amplicons which upon segmentation eventually leads to putative CNV segments. In a second step, it utilizes the dependencies between the amplicons to filter out the potential false positive segments thereby detecting CNVs with higher sensitivity and specificity in comparison to competing methods.

For our validation dataset, which consists of high-coverage NGS data from a targeted gene panel, the initial blind-folded prediction analysis using varAmpliCNV retrieved all the known true positives with high sensitivity (100%) and specificity (99.78%) in comparison to the three competing methods. ONCOCNV which was also designed for handling amplicon sequencing data on targeted cancer gene panel, achieved relatively much lower sensitivity (66.67%) and specificity (97.77%) scores. However, performance of DECoN in detecting all TPs was on par with that of varAmpliCNV thereby achieving 100% sensitivity but comparatively it detected slightly more false positives leading to a low specificity (97.81%). Finally, CoNVaDING which was primarily designed to detect single exon CNV events in amplicon-based TR data achieved the lowest sensitivity (33.33%) and specificity (96.68%) among all the competing methods. This sensitivity and specificity analysis clearly indicates the poor applicability of the existing methods for reliably predicting CNVs from Haloplex^{TM} based amplicon sequencing data.

Current state-of-the-art methods are robust in handling standard sequence specific biases, but show limitations in handling variability associated to high depth of coverage. For example, ONCOCNV incorporates PCA only to capture the baseline coverage using a set of PCs (by default first 3 PCs) from the matched control set to compute ratio scores. In contrast, varAmpliCNV has been designed to predict CNVs when a matched control set is not available. Hence, the principle of PCA is applied directly on all the samples in a single batch to minimize or control the variability of read counts . Second, enrichment designs are typically handled specifically for the targeted protocol. In ONCOCNV, offering the closest match to the varAmpliCNV methodology the read counts per amplicon are generated by assigning reads to the amplicon to which it maximally overlaps. In case of more than 75% overlap between amplicons (tiled design pattern amplicons in AmpliSeq^{TM}) they are merged into a single amplicon. In varAmpliCNV, the read counts are generated by uniquely assigning the reads directly to each amplicon by exact matching of the start and end coordinates. This results in the removal of low quality, unmapped or partially overlapping reads, which increases the signal to noise ratio. Moreover, the MDS based approach has been incorporated in the workflow, giving identical results as with PCA, but eventually increasing the prediction speed several folds so varAmpliCNV is suitable to handle large sets of amplicons. Together with this unique design principle varAmpliCNV scores above ONCOCNV with respect to sensitivity and specificity analysis.

Similarly, varAmpliCNV outperformed DECoN, which achieved the same sensitivity in detecting known TPs, by detecting a smaller number of FPs achieving higher specificity. The essential feature of DECoN is that it incorporates functionalities of the ExomeDepth package in R to predict CNVs. Internally ExomeDepth models the read depth count ratio using a beta binomial distribution that accounts for its over dispersion. Comparatively, varAmpliCNV uses a non-parametric approach by using PCA/MDS method without fitting any predefined standard distribution to control the variances in the read depth count across the samples. Although both of these methods have equal sensitivity (100%) in predicting CNVs, varAmpliCNV performs betterin specificity (99.68%) in comparison to DECoN (97.81%). This is primarily because varAmpliCNV utilizes the overlapping structure of amplicons to prune out the majority of FPs.

Finally, varAmpliCNV outperforms CoNVADING both in terms of sensitivity and specificity. The internal design principle of CoNVADING is based on a comparison of the distribution score of target read count with a set of matched controls. It also ignores the variability associated with the underlying enrichment design protocol. The read counts are generated at the exon level but usage of an amplicon based design file (as suggested in the online tutorial [1]) leads to prediction of CNVs per amplicon, resulting in a long list of CNVs that could be potential FPs.

Although, the sensitivities of recently developed methods are increasing, controlling the specificity (the number of FPs) attributes as the limiting factor regarding the applicability of these tools for reliable prediction. The competing methods do not incorporate any post-processing step for pruning the FPs. VarAmpliCNV provides a unique novel strategy called AOF to deal with filtering of FPs by utilizing the overlapping design pattern of amplicons that can be currently applied for Haloplex$^{\text{TM}}$ technologies. Principally this can also be extended towards other amplicon-based sequencing technologies such as AmpliSeq$^{\text{TM}}$. The utility of AOF in pruning the number of FPs (4 FPs were pruned) can clearly be seen from Figure 5.4.1 (applying DS approach) and Figure 5.4.2 (applying DS-AOF approach). Together, availability of this functionality as graphical user interface enables non computational users to explore the predicted CNV segments and then optimally select them for wet lab validation and subsequent clinical reporting.

In spite of achieving high sensitivity and specificity there are some potential limitations to varAmpliCNV. Foremost, the cut-off boundaries for duplication and deletion have been optimized for the current TR panel of 30 genes. Hence, generalizing the applicability of the same cut-off values on other targeted panel data could lead to potential false negatives. As a proof of concept, we applied the derived cut-offs on an independent deafness panel. According to Figure 5.4.4, it can be seen that 3/4 known TPs (arrayCGH validated) could be identified correctly, but a single CNV (complete deletion of EYA1 gene) was not detected. A possible reason for this missed deletion might be the lower overall average coverage (computed in the normalization step B of the workflow) of the sample containing this CNV in comparison to other samples. This can lead to skewing of the estimation of proportion of variance in PCA/MDS, leading to a loss of true signal during noise removal. In such lower quality scenarios we recommend to inspect both DS and DS-AOF results. Additionally, users can remove less than the recommended 80% variance, such that the true signal is retained at the cost of some additional FPs. In addition to the high sensitivity, we obtained only 15 CNV segments that were considered to be false positive, from this large panel data. This number is considerably lower in comparison to the long lists of CNVs reported by the evaluated competing methods. This is illustrated by the minimal number of reported

---

[1] https://github.com/molgenis/CoNVaDING/blob/master/docs/README.md

CNVs, namely 41, achieved by ONCOCNV based on a gene panel containing just a fourth of the genes. This higher specificity makes varAmpliCNV feasible for wet-lab validation and clinical diagnostic reporting. Incorporation of additional information, such as the allele frequencies of informative heterozygous SNVs present in the predicted CNV segments can further aid in discriminating TP and FP CNVs.

Incorporation of additional information, such as the allele frequencies of informative heterozygous SNVs present in any predicted CNV segment regions can further aid in discriminating the TP and FP CNVs.

5.7 Conclusion

The varAmpliCNV workflow provides a modular approach to handle variability and complexities in amplicon sequencing data, in the context of CNV prediction. The information granularity, flowing from sequencing reads to exon-level targets is indirectly mediated by the enriched amplicons. Leveraging the design information enables us to capture underlying dependencies in the data. The presented PCA/MDS based method captures variability at the individual amplicon level, while the positional dependency between the amplicons is used to filter out FPs. varAmpliCNV is easy to use via command line and Galaxy. The analysis result can be visualized, making interpretation straightforward for both bioinformaticians and lab technicians. Together varAmpliCNV presents a novel approach in detecting CNVs with high sensitivity and specificity on amplicon sequencing data applicable in both research and clinical diagnostic settings.

5.8 Supplementary data

5.8.1 Information flow for the amplicon design patterns

Haloplex based enrichment provides a unique design pattern of the amplicons in the form of overlapping mesh like structure targeting the region of interest (ROI). This overlapping structure allows for characterization of CNVs with good precision and accuracy. An example representation of this structure is shown in Figure 5.8.1A for one of the exons of FBN1 gene. The reads are generated with respect to each of the amplicons. The amplicons have discrete start and the end coordinates and with their overlapping structure encompassing the whole ROI. Hence, the flow of information can be traced from read count describing about the targeted ROI via the amplicons. This flow can be encoded using the graphical structure as shown in Figure 5.8.1B. The given graph has a directed acyclic structure with nodes and directed edges. In the bottom there is single node representing the Read Count (RC). The amplicons A1, A2, A3 An represent the set of amplicons for given ROI thereby constituting the intermediate layer. Finally, the targeted positions P1, P2, P3,..Pm or ROIs form the top layer of the graph. The dependencies in this graph can be described in two ways:

1. The amplicons are independent of each other with respect of read counts. This means any two or more amplicons does not depend upon which reads they are assigned.
2. The amplicons are dependent with respect to each other with respect to position as they form an overlapping structure as shown in Figure 5.8.1A. For example, for position P1 is encompassed by amplicon A1, A2 and A3. Similarly, position P2 is encompassed by amplicon A2, A3 and A4.

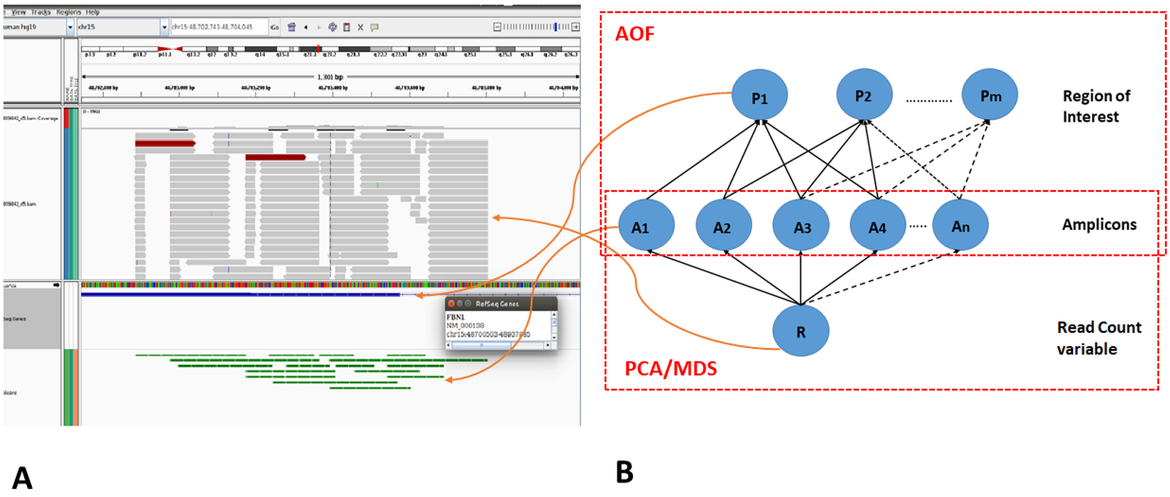


Figure 5.8.1: **Flow of information in amplicon sequencing data.** (A) Example depth of coverage representation for targeted exonic region of FBN1 gene plotted using IGVTools. The green horizontal bars in the bottom are overlapping structure of the amplicons. The middle represents the region of interest targeted by these amplicons. The top of panel shows the reads that are aligned to the position encoded by the amplicons. (B) The flow information in panel A is encoded by the graphical representation (directed acyclic graph) encapturing the various dependencies. The nodes of this graphical structure are Read counts, Amplicons and Positions or the targeted region.

With this graphical structure it is much easier to implement various steps of varAmpliCNV pipeline. The independency with respect to read counts enables assignment of reads to each amplicons matching their start and end coordinates respectively thereby treating each amplicon as unique data point. This helps in applying PCA/MDS and other subsequent steps for CNV detection in the pipeline (see lower part of graphical structure of Figure 5.8.1B). The dependency of amplicons with respect to position is incorporated as filtering step (AOF) to prune out false positives (see upper part of the graphical structure Figure 5.8.1B). Overall, encoding of this dependency structure of the amplicons is core of the varAmpliCNV pipeline.

5.8.2 GC content correction

GC content correction is done by loess based linear regression method to account for biases introduced by GC rich region in the target region. In order to measure the effect of GC content on the read count data it is important to know how much of these fraction correlate with the read counts for all the target regions. Figure 2 and 2b shows the correlation between these two entities computed for each samples across all 5 batch sets divided into autosomal and sex chromosomal targets.

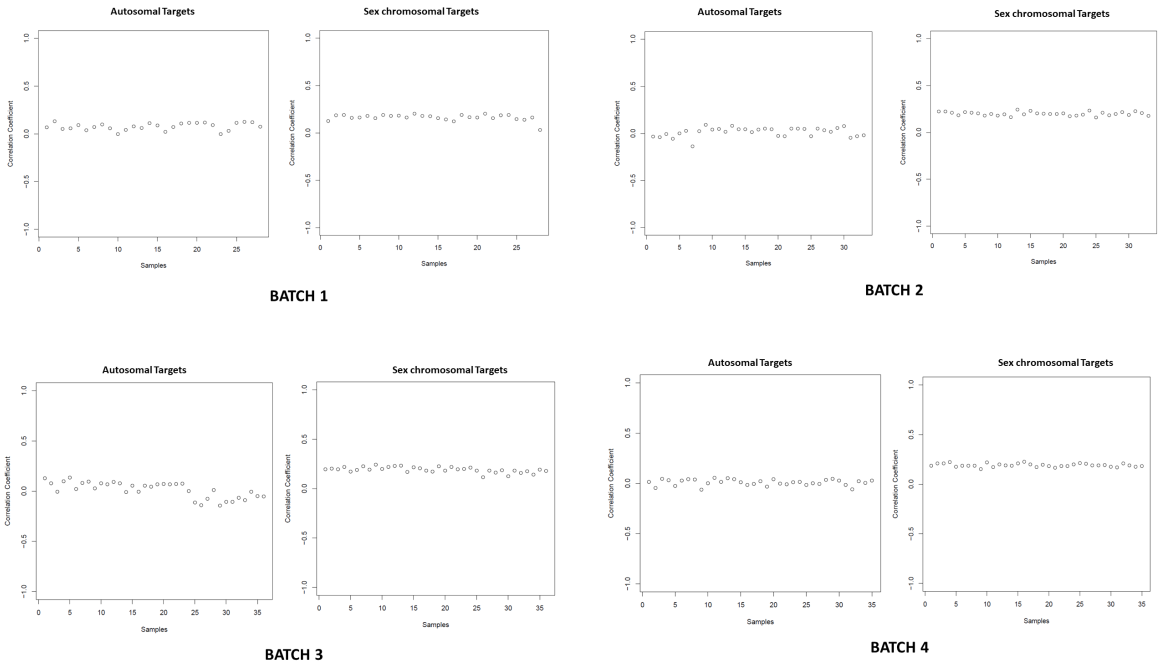


Figure 5.8.2: **Correlation plot of GC content fraction with target read counts.** For samples of each of the batches 1-4 the GC fraction computed for each of the target amplicons were correlated with their respective read counts. The correlation plot was obtained separately for autosomal and sex chromosomal targets. **.**

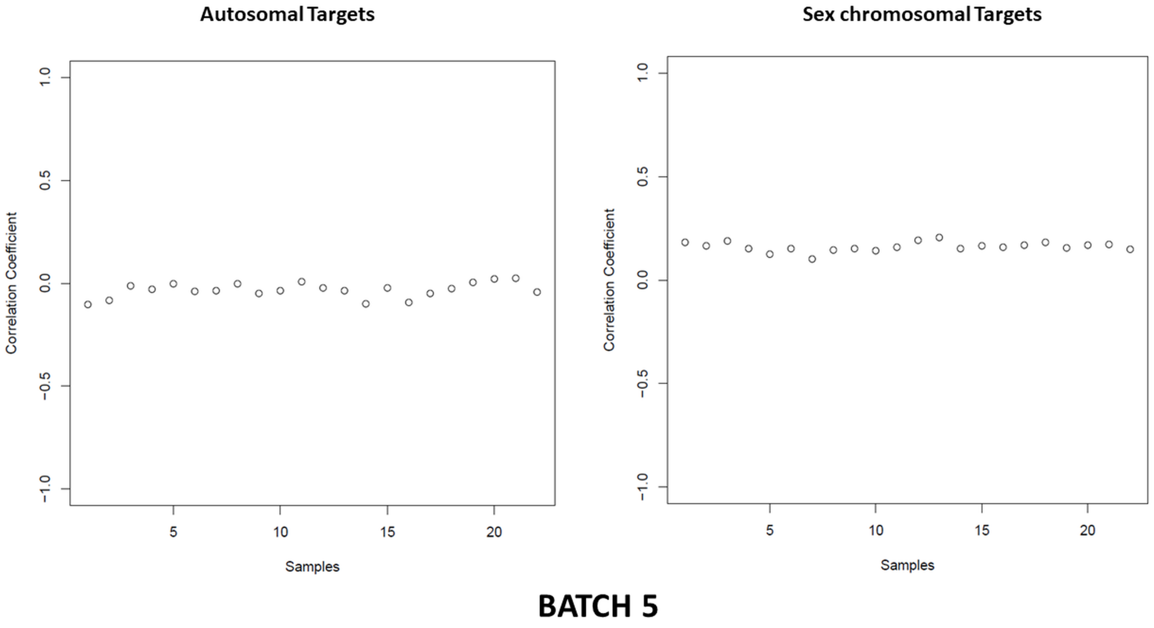


Figure 5.8.3: **Correlation plot of GC content fraction with target read counts.** For samples of the batch 5, the GC fraction computed for each of the target amplicons were correlated with their respective read counts. The correlation plot was obtained separately for autosomal and sex chromosomal targets. **.**

5.9 Controlling the variance using PCA/MDS

In this case Principal Component Analysis (PCA) has been implemented to capture the true underlying variance structure present in the data. From theoretical perspective, the goal is to handle the uneven variation present in the data. Hence, the ideal way to deal with such situation is to explore the variation present in the amplicons within the sample and across the samples. This can be done using Principal Component Analysis. Let's discuss the principles behind it.

For any data cloud, it can be ideally characterized using mean and variance. Mean of the data tells where exactly the data is centered and the variance tells about the spread (elongation and deflation) of data points (x_i of random vector $\mathbf{x}$) from it's mean. Zero mean centering i.e subtracting mean from the random vector, so that each x_i has zero mean. Thus, we can better concentrate on the structure which is present in the data in addition to the mean.

Principal components are the linear combination $s = \sum_i w_i x_i$ that explain as much of the variance of the input data as possible. The amount of variance explained is directly related to the variance of the component. Each of the w_i are the respective principal component weights. Constraint of unit norm is imposed on these weights so that:

$$||w|| = \sqrt{\sum_i w_i^2} = 1 \tag{5.1}$$

This definition gives only one principal component. Other principal components are obtained by deflation approach which can be found by orthogonal transformation to the linear combination of maximum variance. This yields second principal component. This procedure can be repeated to obtain as many components as there are dimension in the data space. The k principal components are given by weight vectors $w_1, w_2, w_3, ...w_k$. The $k+1$-th principal components weight vector is defined by:

$$\max_{\mathbf{w}} var(\sum_i w_i x_i) \tag{5.2}$$

under the constraints:

$$||w|| = \sqrt{\sum_i w_i^2} = 1 \tag{5.3}$$

$$\sum_i w_{ji} w_i = 0 \text{ for all j = 1,...k} \tag{5.4}$$

Variance of any random variable x_1 is defined as:

$$var(x_1) = E\{x_1^2\} - (E\{x_1\}^2) \tag{5.5}$$

This can also be written as $var(x_1) = E\{(x_1^2 - E\{x_1\}^2)\}$, which clearly shows that variance measures the average deviation from the mean value. For more than one random variable it is useful to analyze the covariances given by:

$$cov(x_1 x_2) = E\{x_1 x_2\} - E\{x_1\}E\{x_2\} \tag{5.6}$$

For our Amplicon-by-Sample Read depth count data matrix, the major goal is to understand the variability of the read depth of amplicons within the sample and across the sample. Specifically, we want to know the structure of the variance of the amplicons in the data. This can be better understood by computing the sample covariance matrix of amplicons. If we assume $x_1, x_2.., x_n$ as set of our amplicons then

$$\mathbf{C(x)} = \begin{bmatrix} cov(x_1, x_1) & cov(x_1 x_2) & cov(x_1 x_3) & \dots & cov(x_1 x_n) \\ cov(x_2, x_1) & cov(x_2 x_2) & cov(x_2 x_3) & \dots & cov(x_2 x_n) \\ \dots\dots & \dots\dots & \dots\dots & \dots & \dots\dots \\ cov(x_n, x_1) & cov(x_n x_2) & cov(x_n x_3) & \dots & cov(x_n x_n) \end{bmatrix} \tag{5.7}$$

By combining equation 8 and 7 and extending it as matrix notation we get:

$$\mathbf{C(x)} = E\{\mathbf{x}\mathbf{x}^T\} - E\{\mathbf{x}\}E\{\mathbf{x}\}^T \tag{5.8}$$

Eventually, if the variables are uncorrelated, the covariance matrix is diagonal which means they are correlated to themselves. If they are all further standardized to unit variance then covariance matrix is identity matrix.

5.9.1 Eigen Value decomposition of Covariance matrix is basic PCA analysis

Maximization and minimization of variance of any linear combination of random variable can be computed by optimization of the covariance matrix of the data. Consider any linear combination $w^T x = \sum_i w_i x_i$ we can compute its varaince simply by:

$$E\{(\mathbf{w}^T\mathbf{x})^2\} = E\{(\mathbf{w}^T\mathbf{x})(\mathbf{x}^T\mathbf{w})\} = E\{\mathbf{w}^T(\mathbf{x}\mathbf{x}^T)\mathbf{w}\} = \mathbf{w}^T E\{\mathbf{x}\mathbf{x}^T\}\mathbf{w} = \mathbf{w}^T\mathbf{C}\mathbf{w} \tag{5.9}$$

Assuming mean is zero hence $E\{x\} = 0$. From equation 9 we can see that basic problem of PCA can be seen as optimizing Covaraince matrix for some optimal weight vectors **w**. From linear algebra we can see that the covariance matrix C can be decomposed as:

$$\mathbf{C} = \mathbf{U}\mathbf{D}\mathbf{U}^T \tag{5.10}$$

where U is an orthogonal matrix, and $\mathbf{D} = diag(\lambda_1, ..., \lambda_m)$ is diagonal. The columns of U are the *eigenvectors* of **C**, and the λ_i are the corresponding *eigenvalues*.

From equation 11 and 12 we can solve the PCA problem easily:

$$\mathbf{w}^T\mathbf{C}\mathbf{w} = \mathbf{w}^T\mathbf{U}\mathbf{D}\mathbf{U}^T w = \mathbf{v}^T\mathbf{D}\mathbf{v} = \sum_i v_i^2 \lambda_i \tag{5.11}$$

Substituting, $m_i = v_i^2$ we get

$$\max_{m_i \geq 0, \sum m_i = 1} \sum_i m_i \lambda_i \tag{5.12}$$

All the principal components can be found by ordering eigenvectors $\mathbf{u}_i$, $i = 1, ...m$ in **U** so that the corresponding eigenvalues are in decreasing order. If **U** is ordered then the i-th principal component $\mathbf{s}_i$ is equal to:

$$\mathbf{s}_i = \mathbf{u}_i^T \mathbf{x} \tag{5.13}$$

Now, the columns of the matrix **U** are arranged according to corresponding eigen values hence these are the directions that explain maximum variance.

Our aim is to reduce the variance hence we remove first k columns of matrix $\mathbf{U}$ by projecting the input data matrix to the reduced column matrix of U. The reduced column matrix of eigen vectors is given by:

$$\mathbf{A}_{n\times k} = \mathbf{U}_{n\times k} \tag{5.14}$$

Finally, in order to get the denoised original amplicon-by-sample matrix we do this by:

$$\hat{\mathbf{X}} = AA^TX \tag{5.15}$$

Here, the matrix A is obtained from equation 5.14. In the current analysis we aim to remove approximately 80% of the variance present in the data. Hence, choosing first k columns of the eigen vector we double re-project the unit-norm and zero mean centered data on to it. This denoised matrix is finally used for computing Log_2R ratios and subsequent segmentation using CBS algorithm.

5.9.2 Using MDS gives identical result as PCA

One limiting factor for performing PCA is computing the covariance matrix XX^T and subsequent EVD step. Conventionally, we represent this matrix as read count matrix whose rows are amplicons and columns as samples. Analyzing large gene panels involve large set of amplicons and thus computing of covariance matrix and EVD becomes computational intensive. Hence, we implemented metric multi-dimensional scaling (MDS) approach to address this issue by using euclidean distance measure between the data points. MDS and PCA are connected as they both address towards solving of x^Tx or xx^T matrix. Euclidean distance measure of the amplicon read count matrix is given by:

$$d_{ij} = \text{distance between data points } x_i \text{ and } x_j. \tag{5.16}$$

If we denote each data point as $\mathbf{x}_i$ then euclidean distance between them is given by:

$$d_{ij} = ||x_i - x_j||^2 = ||x_i||^2 + ||x_j||^2 - 2x_i^Tx_j \tag{5.17}$$

If we now normalize these distance such that each row and column sum is zero then we obtain matrix X^TX multiplied with some constant. EVD of this matrix is same as that of covariance matrix XX^T. The only difference is that in this case the EVD is done on the column side of the matrix (X^TX). If we recall the amplicon-by-sample read count matrix the rows are amplicons and columns are samples, then this representation using MDS approach reduces the computational complexity. Hence, we obtain identical result as that of PCA. Finally, once the eigen vectors have been computed then we perform same step as in equation 5.15 for denoisation of the read count matrix.

5.10 Computing AOF

We described two approaches for predicting CNV segments : (1) using direct segmentation (DS) and (2) using amplicon overlap filtering (DS-AOF) approach. According to the Figure 5.10.1 the flow of information in the amplicon sequencing data as represented using graphical model shows the dependency structure of amplicons with respect to the genomic position (upper half of the graph). The aim of DS-AOF approach is to utilize this dependency structure to average out the Log_2R ratios of each of these amplicons. The DS-AOF approach is applied to filter out the potential false positive segments. It works in three stages as shown in Figure 5.10.2 for an example CNV segment predicted by DS approach in Batch 3 (see supplementary file 1: sheet 3). The CNV segment is of FBN1 gene having coordinates as chr15:48780441-48782301 encompassed by 14 amplicons.

- **Overlapping amplicons retrieval**: For the given CNV segment all the 14 amplicons are retrieved as shown in blue horizontal lines starting from amplicon AMP_1 (start of the segment) and AMP_14 (end of the segment). There are eight other amplicons marked in green which either start or end segments partially overlap with the predicted CNV segment and are called as "flanking" amplicons. In the DS approach these flanking amplicons were not included for predicting the CNV segments. However according to dependency model of flow of information as shown in Figure 5.10.1 these flanking amplicons are related to the included amplicons because they overlap. Hence for computing the final segmental average Log_2R ratios should include the contribution from the flanking amplicons. Thus for the given predicted CNV segment all the amplicons that overlap with this region are retrieved.

- **Partitioning the segments**: After retrieving all the segments (including flanking and included amplicons) that encompass with the predicted CNV segment we partition it. The segment is partitioned according to start and and end of the amplicons such that each partition has fixed width or equal length of amplicons. An example demonstration of this partitioning step is demonstrated in Figure 5.10.2 where the predicted CNV segment is partitioned into 19 intervals denoted as $\lambda_1 ... \lambda_{19}$.

- **Averaging out the Log_2R ratios**: For each of these partitioned interval we average the Log_2R associated with each of the amplicons for a given interval. The averaged

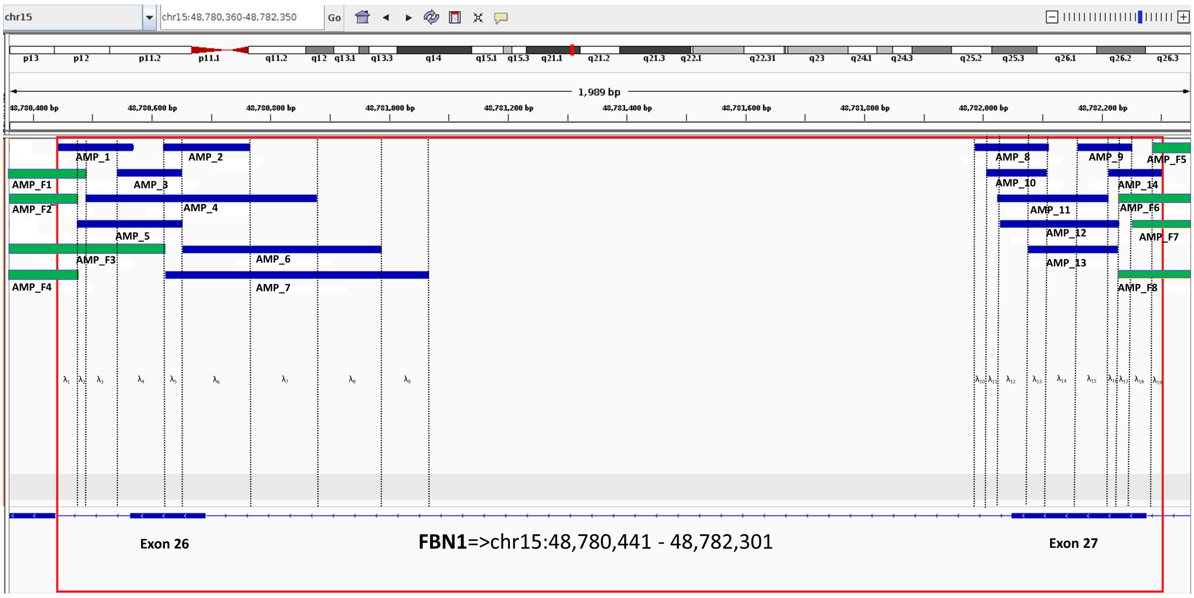


Figure 5.10.1: **Flow of information in amplicon sequencing data. A**) Example depth of coverage representation for targeted exonic region of FBN1 gene plotted using IGVTools. The green horizontal bars in the bottom are overlapping structure of the amplicons. The middle represents the region of interest targeted by these amplicons. The top of panel shows the reads that are aligned to the position encoded by the amplicons. **B**) The flow information in panel A is encoded by the graphical representation (directed acyclic graph) encapturing the various dependencies. The nodes of this graphical structure are Read counts, Amplicons and Positions or the targeted region.

Log_2R ratio for the overall segment is given by :

$$\text{Avg. Log}_2\text{R} = \sum_{j}^{N} \frac{\sum_i^{M_j} \text{Log}_2\text{R}_{\text{AMP}_{ij}} \times \lambda_{ij}}{M_j} \tag{5.18}$$

From equation 5.18 we obtain the new averaged Log_2R ratio score for the predicted segments. Thus application of amplicon overlap filtering (AOF) helps in providing realistic estimate of the segment means as compared to those obtained by using direct segmentation (DS) approach by using CBS algorithm. Together we define this strategy for filtering out the segments as DS-AOF.

5.10.1 Application of DS-AOF

We demonstrate the application of DS-AOF approach in filtering out the CNV segment (FBN1 gene; chr15:48780441-48782301) using DS approach as shown in Figure 5.10.2. Here the amplicons are plotted according to their Log_2R scores having similar intuitive representation as shown in Figure 5.10.1. We can see that the 14 amplicons that led DS-based approach to predict as potential CNV segment are marked with blue horizontal lines. These amplicons result in segmental average Log_2R score of -0.61 (blue dotted line) which is less than the boundary threshold (BT) for deletion (≤ -0.50) thereby making a mis-classification.

With incorporation of Log_2R score values from the "flanking" amplicons and combining their overlap dependencies with the 14 amplicons results in the segmental average Log_2R score of -0.3163 (brown dashed line). This value is higher than the BT value of (≤ -0.50) and thus helps in classifying this CNV segment as a false positive. Thus this makes DS-AOF strategy a useful approach in filtering out the false positives.

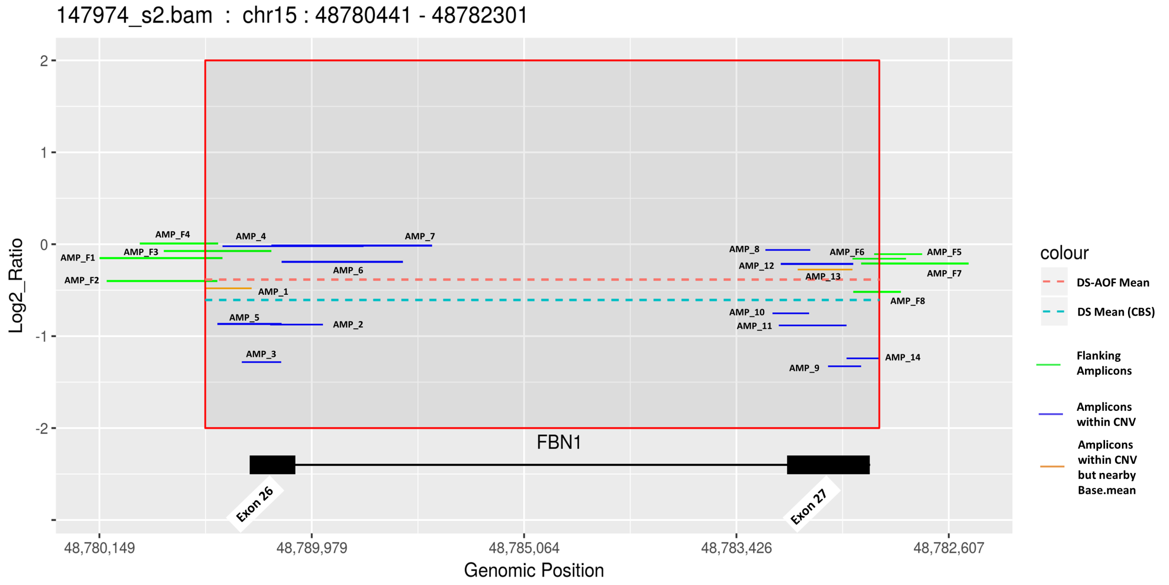


Figure 5.10.2: **Visualization of an example CNV segment using DS-AOF approach** The CNV segment corresponds to *FBN1* gene having coordinates as chr15:48780441-4878230 enclosed in red box. The X-axis represent the positions of the respective exons and the Y-axis represent the actual Log_2R ratio values of amplicons obtained after PCA/MDS based normalization. The amplicons are plotted as horizontal lines (blue/green/orange) according to their genomic coordinates. Amplicons that are marked as blue are the ones that led to the generation of this particular CNV segment using the CBS algorithm. And the green amplicons correspond to those amplicons that did not participate in CNV segment generation but overlaps with this CNV segment. The dotted lines marked in as light-blue and brown colour represent the average segmental Log_2 R scores obtained using DS-approach or CBS algorithm and DS-AOF approach respectively. Finally the brown amplicons are the ones that are plotted nearby to the average segmental Log_2R scores (brown dotted line) obtained using DS-AOF approach.

General discussion and conclusions

Chapter 6

General discussion and conclusions

All models are wrong but some are useful

BOX G.E.P (1976)

Genetic disorders can be complex and depending upon the kind of inheritance model they follow, the underlying pedigree patterns, the age of onset, the locus and phenotype heterogeneity underlines the complexities involved in detection of associated causal variants. In order to decode the mechanism by which these variants cause disease, it is important to delineate their characteristics, their prevalence in patients and control populations and their interactions with the environment.

In **chapter 1** we introduced complexities associated with the human genome. We explained the fundamental concepts about different types of genetic variants (for brevity variants), including single nucleotide variants (SNVs) and structural variants (SVs), and their potential functional consequences. We also highlighted the importance and challenges associated with identifying disease-causing variants using various strategies. These strategies ranged from traditional functional cloning to candidate gene approaches utilizing current state-of-the-art NGS technologies such as whole exome (WES) and whole genome sequencing (WGS). Recently WES based approaches have been reported[33] to successfully identify the cause of Mendelian disorders under different inheritance models. Due to these successes, NGS technologies have become the standard practice in disease gene discovery over the last 15 years. Moreover, as the sequencing cost reduced dramatically, the way was paved to upscale the analysis and interpretation of the human genome to increasingly large patient cohorts. In-depth analysis of the large amounts of high-throughput data these technologies generate, requires advanced computational frameworks and methodologies. These frameworks offer an automated process to computationally interpret genome-wide SNVs and SVs with increased accuracy. However, despite the advancements so far, fewer than 50% of Mendelian disorders have been resolved, indicating that there is significant room for improvement over the available methods.

In this thesis we addressed the issue of computational genome interpretation by implementing a complementary approach which includes the development of a novel computational candidate gene prioritization tool (**chapter 3**) and application of statistical mutation burden analysis to pinpoint causal variants involved in bicuspid aortic valve (BAV) associated with thoracic aortic aneurysm (TAA) or BAV/TAA disease (**chapter 4**). Additionally, in **chapter 5** we introduced a novel statistical model to detect SVs in the form of copy number variations (CNVs) from targeted NGS data. In the following sections we discuss some key aspects of our approaches, their applications and future perspectives.

6.1 Complementary strategies for disease gene identification

Accurate interpretation of the genome is key to identify the causal variant in NGS data. In **chapter 1** we presented the basic NGS workflow which describes the different steps between sequencing and variant discovery. Given the underlying research question, analysis can be hypothesis free, such as for whole genome or whole exome sequencing, or hypothesis driven, using customized sequencing on set of targeted panel. For either case, the amount of data to be analyzed is enormous, making it unrealistic to manually grasp its full implications. For example, through WES analysis we only obtain information about exonic regions, which corresponds to 1-2% of genome, but finding candidate genes in the thousands of returned variants is a challenging task. Although the list can be narrowed down through optimal design of filtering strategies, pinpointing the actual causal variants that can be functionally validated and predicting their effects still remains computationally challenging.

To pinpoint disease associated genes in BAV/TAA, we applied two complementary approaches, namely computational gene prioritization and mutation burden testing. pBRIT, a novel gene prioritization method, constituted the first part of the complementary approach and was used to prioritize the candidate genes obtained from a WES pipeline. The results were combined with mutation burden analysis on the same data, resulting in a set of candidate genes. These candidate genes were subsequently analyzed for rare variant association in a replicative cohort using a customized targeted gene panel. The successful application of this strategy on WES data helped to identify *SMAD6* as a significant and important contributor towards BAV/TAA disease. The results support our hypothesis that in the NGS era, data driven approaches are key towards disease gene identification. However, certain considerations and challenges need to be addressed adequately to correctly interpret the findings of these methodologies in the context of genetic research. Some of these considerations are discussed below.

6.1.1 Gene Prioritization

Computational candidate disease gene prioritization is a broad field with many available methods and approaches. Despite this range in methods, some challenges are common. First, many implement the guilt-by-association principle, which means that genes that share similar functional aspects also share the association with similar phenotypic aspects. Utilizing this principle to cluster the network of genes can help in prioritizing candidates. Second, storing and analyzing multiple annotation sources for the full set of human genes requires a rigorous computational framework to efficiently handle the presence of sparsity and modeling co-occurrence and dependencies among the features within the annotation sources. Finally, a very important aspect that can dramatically impact the performance of gene prioritization methods is the effect of changing annotation sources.

Annotation sources are dynamic and keep on changing on regular basis. Hence, the application of regression in Bayesian framework aids in modeling uncertainties associated to estimation of the regression parameters. We specifically aimed to address these challenges in **chapter 3**, where we presented our novel gene prioritization tool named pBRIT. It utilizes an information theoretic approach to integrate 10 different annotation sources and a linear regression model in Bayesian framework to prioritize candidate genes.

As mentioned above, the guilt-by-association principle has been the core of the majority of gene prioritization algorithms developed so far. This means that genes that are known to be directly associated with the disease are often chosen as training or seed genes and based on similarity to these genes, the candidate genes are prioritized. Often it has been alleged that methods based on this principle are biased towards well studied genes (annotation bias) and the potential to discover new genes becomes limited. Therefore, it is necessary to circumvent this bias by facilitating an unsupervised approach towards the retrieval of these associations. Although the functionality of pBRIT is also partially based on the guilt-by-association principle, we circumvent the issue of annotation bias by incorporation of an information-theoretic (data driven) approach through TFIDF and TFIDF→SVD based methodology for mining features, done solely in an unsupervised way to yield genome-wide gene-by-gene proximity profiles. Eventually these similarity profiles result in small clusters or networks of genes, which can be visualized to offer the user a way to interpret the pattern of relatedness.

Next to the unsupervised feature selection, two questions still arise related to possible biases to the mined annotation sources: is the integration of large annotation sources necessary and viable, and what is the effect on prioritization when a smaller number of annotation sources are used? The answer to these questions can be found in the fact that the precise determination of the functionality of a gene is a complex question. Functional experiments, complemented with transcriptomics, metabolomics and pathway analysis, have provided a plethora of information about genes regarding molecular function, involvement in certain biological processes and precise localization in the cell. This information is widely curated within high quality annotation sources (e.g. GO, KEGG, IntAct etc.) and distributed across different biological databases. It is thus required to have a multi-view approach to capture these different aspects or views on the functionality of genes, to obtain robust gene prioritization tools

Hence, we categorized the gene level proximity profiles obtained using unsupervised methods for 10 curated annotation sources, into functional and phenotype annotations. This combination of data driven based mining of features through TF-IDF methodology, followed by a multi-view style integration of the proximity profiles, together leads to an intermediate integration based fusion of the included annotation sources. This intermediate fusion approach further circumvents potential annotation bias and contributes to improvement of the predictive performance of pBRIT.

With regard to the number of annotation sources needed, we demonstrated that the integration of a more annotation sources certainly plays a role in improving the predictive performance. For example, the performance of pBRIT is 22% higher than an RWR-M based approach that integrates only 4 annotation sources when performing prospective benchmarking. Here, the information from additional views will provide more comprehensive knowledge to correctly prioritize the novel genes. On the other hand, pBRIT performs on par with Endeavour, integrating 44 annotation sources, including broader and less curated sources such as experimental gene expression datasets. This indicates the necessity and importance of high quality data, in addition to merely a higher number of data sources, for an effective predictive model.

Beside the data integration it is important to take the dependency or relatedness between the annotation sources into consideration. We used the information-theoretic model of pBRIT to capture dependencies between features within a single annotation source. However, relatedness between annotation sources can provide an advantage to the prioritization model when modelled at the data fusion level. Earlier research has demonstrated that different annotation sources are non- orthogonal and hence tend to correlate with each other with respect to gene(s). For example, functionality of genes represented by GO terms (molecular function, biological process and cellular localization) and sequence similarities are correlated and this has been used to correctly transfer the annotation in homology-based sequence similarity searches[109]. Capturing these dependencies between the annotation sources can be done by several machine learning techniques. Co-training[25], multiple kernel learning[116] (MKL) and canonical correlation analysis[85] (CCA) based approaches are examples whose main principle is to find a certain set of features which can maximize the mutual agreement between them thereby giving a holistic multi-view approach towards learning the pattern in the data. Incorporating these methods would be potential areas for future development of pBRIT.

Furthermore it is also worth to mention that the information related to the functionality of genes distributed across various databases is not static. The widening gap between the speed with which genomic data is generated and subsequent availability of curated annotations, defines the presence of uncertainty in the underlying functionality of genes. A recent study[217] analyzing the impact of outdated gene annotations on pathway enrichment tools claimed that the majority of tools incorporate outdated annotation sources, thereby making the predictions highly unreliable. In such a scenario it is customary for the developers of prioritization tools to either update their internal annotation sources periodically, or at least include some customized improvement in the internal algorithm design that can incorporate the uncertainties arising due to implicit changes in the annotation. pBRIT addresses uncertainties arising due to implicit changes in annotation sources by performing linear regression under a Bayesian framework, which aids in modeling the uncertainties associated with regression parameters. In our work we explored and measured the effect of 5 year changes in the annotation of human phenotype ontology (HPO) and gene ontology (GO) terms on prioritization ranks. We achieved almost no correlation between the obtained rank and changes in annotation terms. However, our study is limited to only GO and HPO, which are both relatively stable, whereas other annotation sources holding information about sequence similarity or associated phenotypes keep on changing. Incorporating changes from these annotation sources could be computationally intensive, but better representations and statistical modelling of the changes could aid in better understanding of their effects on prioritization performance.

By addressing these key aspects, coupled to the intermediate data fusion, pBRIT provides a modular tool that can prioritize large amounts of high throughput data. Its performance was compared with 9 different competing methods, using different data integration principles and/or different prioritization algorithms. The obtained AUC score ranging from 0.92 to 0.96 on cross-validation benchmarks and of 0.80 and 0.87 on time-stamped and prospective HPO terms predictions clearly demonstrates the stable performance of pBRIT on various disease data sets, and superior predictive ability on unseen or newly discovered disease gene associations (real world usage scenario).

6.1.2 Mutation burden analysis

In addition to gene prioritization, there are different statistical methods for estimating the association of genetic variants with diseases based on the comparison of allele frequencies in a cohort of affected individuals and healthy control individuals. These methods include the basic approach widely applied in rare variant disease association studies called burden test or more precisely cohort allelic sum test[152] (CAST). Practical application of this method is demonstrated in **chapter 4**. After exome sequencing of a BAV/TAA cohort of 196 unrelated individuals and 193 control individuals, the application of filtering criteria for the rare variant association analysis (MAF $\leq$ 1%) yielded a list of on an average 2000-3000 variants in 150-200 genes per sample. These filtered genes were further trimmed through the complementary strategy incorporating gene prioritization (using pBRIT) and mutation burden analysis using the CAST approach (see section 3.10.1 of **chapter 4**). Together both of these strategies resulted in a list of 61 genes (44 from prioritization and 17 from burden test) for targeted resequencing.

Next based on extensive literature review and CNV analysis experts of the MIBAVA consortium selected an additional 86 genes to incorporate into final targeted panel containing 147 genes. This panel was applied to a larger replicative cohort of 441 patients with BAV/TAA disease and 183 controls (non-cardiovascular phenotype). CAST based burden testing was performed between the patient cohort and control frequencies obtained from the ExAC[62] database. In the current thesis we summarize our findings for a selection of 22 genes (see **chapter 4**) out of 147 genes that can be associated to BAV/TAA disease. Among these 22 genes we found that a significantly higher number of variants in *SMAD6* gene in BAV/TAA patients than controls. The selection of *SMAD6* gene in the panel was mainly contributed from the gene prioritization. Three other genes that resulted through prioritization (see Table 4.1 of **chapter 4**) failed to survive the burden test. The other candidate genes obtained from burden test of variants from WES data did not survive the burden test in the targeted resequencing analysis. Hence they were excluded from the list of selected 22 genes to explain the hypothesis behind etiology of BAV/TAA disease. Nevertheless, from the perspective of methodology the burden analysis is powerful tool in the rare variant association studies. Combining this with gene prioritization approaches can together help in pinpointing the disease causing genes from a large set of candidate genes. Identification of *SMAD6* gene as an important contributor for BAV/TAA disease exemplifies the importance and utility these complementary strategies in disease gene identification from NGS data.

Given the genetic etiology of BAV with or without TAA being elusive, enrichment of *SMAD6* variants in our replicative cohort is an important finding. Earlier *SMAD6* has been implicated in BAV without TAA, but mutations were observed in a limited number of patients. Additionally, so far no known gene contributing to more than 1% of BAV or BAV/TAA disease has been identified in humans. In the current analysis *SMAD6* with 11 variants provides a molecular explanation for 2.5% of our study population. This supports *SMAD6* as an important contributor towards development of BAV-related TAA disease. Moreover, two other genes from the targeted panel, namely, *NOTCH1* and *NOS3*, had a border line statistically significant result (p-value   0.05). *NOTCH1*, which has been previously determined as the only BAV gene, unexpectedly turned out to have a protective effect, with the number of variants significantly higher in the control dataset compared to the cases. Similarly, variants in *NOS3*, a gene that has also been previously demonstrated to play role in formation of BAV in *Nos3* targeted mice, turned out to be protective.

From a methodological point of view there are certain limitations to this study, as presented in **chapter 4**. First, the number included genes in the targeted gene panel and the patient cohort size is rather small to significantly detect oligogenic inheritance or gene-gene interactions involved. This limitation can be ameliorated primarily by selecting larger cohorts during the study design and secondarily by selecting genes directly from BAV related pathways which can also be used as training genes at the gene prioritization step. Second, although the CAST method is very elegant and powerful, it has some shortcomings. Its underlying assumption is that all variants in a gene influence the phenotype in the same direction. However there are certain cases where the same gene can carry alleles having opposite directional effect[19]. The modular structure of collapsing the variants at gene level ignores these directional effects. In order to account for this, several advanced statistical models have been devised recently which intrinsically model the distribution of these direction effects. In chapter 1 we introduced basic concepts behind these methods and their applicability towards rare variant association analysis. Among these, the sequence kernel association test (SKAT) [222] test is most widely used, as it incorporates the directional effects of the variants along with the covariates and allows population stratification. We envisage that in our future work incorporating SKAT test on MIBAVA exome and targeted panel could result in further identification and elucidation of disease causing variants.

6.2 Copy number variation analysis

In the previous section we discussed data driven approaches contributing to the computational interpretation of the genome in order to identify causal mutations in BAV with TAA disease. Next, we discuss another computational approach named varAmpliCNV that was developed to identify copy number variants (CNVs) from targeted resequencing NGS data using HaloPlex enrichment. In **chapter 1**, we highlighted the importance and benefits of using NGS techniques in detecting CNVs with high resolution in comparison to traditional methods such as karyotyping, FISH and SNP arrays. NGS techniques facilitate a genome wide scan for putative CNV regions through distinct mapping of high throughput sequencing reads to genomic regions. We described distinct sequencing features that can be used as signatures of structural variants such as deletion, amplification, inversion, translocation and tandem repeats. These features were aberrant insert size clusters in paired-end mapping (PEM), split-reads (SR), and read depth (RD). Many tools were developed over the years for handling WES or WGS data, incorporating either of these sequencing features, de novo assembly, or a combination of any of them to distinctively predict the SV segments with good accuracy.

However, for targeted resequencing data only read depth (RD) can be applied as reliable predictor, allowing only detection of unbalanced events. Additionally, we highlighted various forms of systemic biases, such as presence of GC rich regions, enrichment protocol design and variability in read depth, that need to be addressed adequately in the analysis pipeline. Among these concerns our focus in the development of varAmpliCNV, presented in **chapter 5**, was mainly on biases arising due to the design pattern of the specific enrichment protocol provided by the HaloPlex^{TM} technology. We do so by incorporation of a unique strategy to directly assign sequencing reads to individual amplicons to obtain amplicon-level read counts. This way, we can decompose the signal of multiple overlapping amplicons into individual data points. Next, it incorporates relevant quality control metrics at several stages of the analysis to prune out bad quality samples, and addresses the above mentioned standard systemic level biases. Using a novel statistical model based on PCA/MDS, we further de-noise the read count data by controlling the variance. Potential CNV segments are obtained by statistically comparing change point events of RD ratios between the sample of interest and the remainder of samples in the experiment. Finally, an amplicon overlap filtering (DS-AOF) approach that harnesses the dependency or overlapping structure of amplicons representing a given genomic region is available to filter out putative false positive CNV regions.

Although all of these features are important in our pipeline, it is worth to go into detail about the steps that represent the underlying flow of information, mediated via the amplicons and going from the sequencing reads to the genomic position. We encoded this flow of information using a three layered directed acyclic (DAG) graphical model where the lower nodes represents the read count data, intermediary nodes are the amplicons and upper nodes are the respective genomic positions covered by each of these amplicons (see Supplementary figure 5.8.1 B of **chapter 5**). This modular representation as a graphical model helps in dividing the CNV prediction task in two global stages, thereby capturing inherent dependencies and independencies present in the data. Coupled to the PCA/MDS based normalization approach, this exemplifies the data driven approach for CNV detection, making varAmpliCNV different from competing methods.

varAmpliCNV was evaluated against three competing methods, namely ONCONV, CoNVADING and DECoN. Each of these methods can accurately predict CNVs from specific TR data but all were designed for different enrichment protocols. ONCOCNV was constructed to analyse amplicon-based data generated using AmpliSeq, while CoNVADING and DECoN work optimally on hybridization based data obtaind using SureSelect^{TM}. To our knowledge, varAmpliCNV is the first tools that has been designed specifically to predict CNVs from targeted NGS data enriched with Haloplex^{TM}.

varAmpliCNV outperforms both ONCOCNV and CoNVADING in sensitivity and specificity. Although DECoN is equivalent with respect to detection of true positives, varAmpliCNV scores significantly better in terms of specificity, as the AOF approach effectively prunes out the number of false positives. Moreover, we only considered variants not called by varAmpliCNV for wet-lab validation if they were predicted by at least two competing methods. If we would assume that all single-method calls were false positives, the specificity of ONCOCNV, CoNVADING and DECoN would drop to approximately 89.8%, 28.4% and 59.6%, as compared to 99.78% for varAmpliCNV.

Computational efficiency is of less concern while analyzing TR data, because the ROI of a targeted panel is much smaller than that of WES or definitely WGS. However, for amplicon based sequencing, an increase in the number of amplicons could also substantially increase the computational complexity of the prediction model. Incorporation of MDS methodology in varAmpliCNV anticipates this by significantly reducing the computational time without affecting sensitivity and specificity.

Despite having key novelties and exhibiting robust performance on Haloplex^{TM}

enriched TR data, there are some limitations to varAmpliCNV. First, the cut-offs determined to filter amplification and deletion events might be over optimistic due to overfitting on the benchmark dataset of 30 genes related to TAAD. Applying the same criteria to other targeted panel data might lead to a decrease in performance. This was demonstrated when we applied this derived cut-offs on large deafness panel data consisting of 145 genes. In this case we correctly predicted 3/4 arrayCGH validated CNVs or TPs and failed to predict single TP. Overall for such large panel data we could only obtain 15/19 additional CNVs which is relatively much lesser in comparison to existing methods thereby making it more feasible for the wet lab validation for clinical reporting under diagnostic settings. Nevertheless, the derived cut-off values is just an option in the program, it is not compulsory and can be adjusted for different TR panel data. This makes the varAmpliCNV tool a completely data driven and finds what is essentially present in the underlying data without any predefined hypothesis.

Second, removal of systemic noises that are correlated to or dependent on the underlying genomic signal could pose a potential challenge in the read count normalization step. Such noise could arise at various experimental levels, such as during the sample preparation stage. Currently in our analysis tool we ignore such noises and we only deal with noises (GC content biases, read depth variability etc) that are not significantly correlated to the genomic signal. Hence usage of linear methods such as PCA helps in controlling these biases. Accurate noise analysis to identify relevant patterns is non-trivial and requires more sophisticated statistical approaches such as independent components analysis (ICA)[88] to separate statistically independent sources of noise. We envisage that the future development of varAmpliCNV should take into account this aspect of noises.

6.3 Interpretation of variants

In previous sections we discussed different computational approaches and challenges involved in interpreting the causality of identified variants, including SNVs and CNVs from WES and TR data. Interactive, web-based platforms and command line tools are often clubbed together in decision support frameworks to systematically evaluate the sheer amount of variants generated. Delineating the direct connections from the variant to the disease phenotype for clinical reporting, incidental findings and further research precisely defines the importance of variant interpretation.

In order to understand the effect of the variants it is important to understand how these variants gets annotated. There are currently three major tools that provide the annotations namely, Annovar[220], SnpEff[49] and Variant Effect Predictor

(VEP). ExAC [62](used as control dataset) utilizes VEP annotation and our in-house tool VariantDB[214] (through which all the samples were processed) uses Annovar. Moreover, the set of gene transcripts used by a tool can differ, using for example Ensembl or RefSeq transcript information. Recently it has been shown that there is significant impact on the classification of variants between the usage of either transcript set, or the usage of either VEP and Annovar based annotation[144]. For example, the Ensembl transcript set is larger, resulting in functional effects for transcripts not present in RefSeq. Second, internal precedence rules might result in different annotations. For example, Annovar inspects the transcript sequence generated by a frameshift event for the generation of stop codons at the mutation site. If present, a stop-gain mutation is reported, while VEP simply reports a frameshift mutation. Although different, both annotations are principally correct. In order to ameliorate the discrepancy arising due to these differences, incorporation of sequence ontologies(SO) [60] could be used as an alternative, as it provides uniform comparison of the annotations across different tools. The SO provides the hierarchical representation of information flows about different types of variants and describes the sequence alteration is seen in the sequence (such as in transcript or exons). Currently SO terms are generally used by most of the genetic variant databases such as ClinVar[118], dbVar[121], dbSNP[192] and Ensembl Variation[227].

Furthermore, prediction programs are often used to prioritize the variants based on the estimated deleteriousness and pathogenicity. The usage of such functional effect predictions such as CADD, SIFT, Polyphen and Mutation Taster can help in narrowing down the list of variants in the candidate genes but caution must be exercised as causality cannot solely be based on prediction programs. Because these scores are only related to alteration of the gene function, they are generally not indicative of the ability of the variant to cause disease. Hence it is required to have automated and dedicated bioinformatics tools[193, 194, 224] to evaluate the causality of variants. Because these tools can differ in reporting of their interpretation, uniform guidelines have been formulated, such as those from American College of Medical Genetics (ACMG)[15]. By standardizing the interpretation workflows, accuracy will be increased. The ACMG provided a scheme of ranking evidence used to make disease association assertions. The evidence could be categorized into four classes namely supporting, moderate, strong and very strong. Subsequently the outcome can be further categorized into five classes being either benign, likely benign, variant of uncertain significance (VUS), likely pathogenic and pathogenic. In summary, the prioritization of variants should take into account the genotype, the disease prevalence, family history and phenotype. In addition, prioritized variants should be further subjected for manual expert interpretation through literature review and validation through in vivo and in vitro experiments.

Interpretation of SVs is also complicated by the fact that they are almost not represented in population-scale variant databases such as ExAC. Alternatively databases such as dbVar, database of genomic variants[138] (DGV) and DECIPHER[64] can be used to compute population frequencies for SVs. However while using these databases several considerations should be taken into account for the analysis. First, the size of CNVs reported in the database could differ due to technical limitations of the different array platforms used[80]. Second, information related to gender is not always provided and this is particularly important when analyzing cases of X-linked CNVs in males, as many of these CNVs can be seen in healthy females. Finally, it should be noted that many of the reported CNVs from large population studies have not been validated. In the near future, with the upcoming large scale WGS projects such as the 100,000 Genome project (UK 100K) and the NHLBI Trans-Omics for Precision Medicine (TOPMed), useful resources of accurate NGS-based SV frequencies in the general population might become available.

6.4 Conclusion and future perspective

The current thesis work demonstrates the successful implementation of data driven approaches towards interpretation of variants detected in the genome. We introduced a novel tool named **pBRIT** which was used to prioritize a list of candidate genes resulting from a WES pipeline and enabled us to detect *SMAD6* as an important contributor towards BAV/TAA disease. We discussed the fundamental design principle of the statistical models and highlighted their key limitations that need to be addressed adequately for effective prediction. We demonstrated that multi-view data driven approaches and unified computational pipeline can aid in processing exome data to pinpoint causal genes for a specific genetic diseases. Additionally the implementation of a non-parametric statistical model in our tool called **varAmpliCNV**, to identify CNVs from Haloplex enriched TR data exemplified key considerations to be taken into account for successful identification of variants from noisy data. The advent of NGS technologies has indeed revolutionized the fundamental understanding of the way we look at our genome and functionalities of genes. The massive deluge of genomic data from NGS technologies on the one hand has provided an opportunity to decipher the mechanisms underlying genetic diseases with greater precision, but on other hand it also poses challenges towards development of an effective framework to store the huge amounts of data and robust computational methods to interpret them.

Many challenges related to the interpretation of the sheer numbers of variants generated by these sequencing technologies remain unsolved. We discussed various approaches and guidelines that are currently being followed to make a uniform assessment across different clinical research and diagnostic labs. The confluence of availability of the genomic data, the sophisticated analytical methods and growing need for quality health care has led to emerging of an era of big data analytics (BDA) in the field of medical sciences. The BDA can be briefly summarized with three Vs[181] which are volume, variety and velocity of the data. The volume defines size and rate at which the genomic data is arising due to large scale sequencing using different NGS technologies or -omics experiments. The availability of different types of data such as through multi-omics experiments which are structured or unstructured and can be combined for the analysis underscores the definition of variety in data. Finally the speed (velocity) of analyzing this large volume and different types of data require faster sophisticated analytical frameworks for yielding timely information.

With the outlook of BDA being promising it has brought back the debate whether future medical science would still be a hypothesis driven (deductive reasoning) or a data driven (inductive reasoning) approach. We envisage that with growing advances in analytics, especially in the domain of statistical machine learning - such as the re-invention of neural networks in the form of Deep Learning (DL) - could lead to development of better decision frameworks for addressing the challenge towards interpretation of genetic variants. However, no matter how robust the underlying statistical model, the quote from G.E Box that All models are wrong but some are useful[32] is still applicable. Although the data driven approaches are useful in inducing a hypothesis, they are truly subjective in the context of underlying data. If the data is noisy and incomplete then often these models result in false hypotheses. Nevertheless, these models have the possibility to create multiple hypothesis and choose the most parsimonious model that explains the data in the best possible way.

Bibliography

Bibliography

[1] 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

[2] Abecasis, G. R., Cherny, S. S., Cookson, W. O., and Cardon, L. R. (2002). Merlinrapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics*, 30(1):97–101.

[3] Adzhubei, I., Jordan, D. M., and Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics*, (SUPPL.76).

[4] Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L. C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., and Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology*, 24(5):537–544.

[5] Albin, R. L. and Tagle, D. A. (1995). Genetics and molecular biology of Huntington's disease. *Trends in Neurosciences*, 18(1):11–14.

[6] Alkan, C., Coe, B. P., and Eichler, E. E. (2011). GATK toolkit. *Nature reviews. Genetics*, 12(5):363–76.

[7] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–10.

[8] Andelfinger, G., Loeys, B., and Dietz, H. (2016). A Decade of Discovery in the Genetic Understanding of Thoracic Aortic Disease.

[9] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L.,

Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: Tool for the unification of biology.

[10] Attias, D., Stheneur, C., Roy, C., Collod-Béroud, G., Detaint, D., Faivre, L., Delrue, M. A., Cohen, L., Francannet, C., Béroud, C., Claustres, M., Iserin, F., Khau Van Kien, P., Lacombe, D., Le Merrer, M., Lyonnet, S., Odent, S., Plauchu, H., Rio, M., Rossi, A., Sidi, D., Steg, P. G., Ravaud, P., Boileau, C., and Jondeau, G. (2009). Comparison of clinical presentations and outcomes between patients with TGFBR2 and FBN1 mutations in marfan syndrome and related disorders. *Circulation*, 120(25):2541–2549.

[11] Avery, O. T. (1944). STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES: INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *Journal of Experimental Medicine*, 79(2):137–158.

[12] Baglama, J. and Reichel, L. (2005). Augmented Implicitly Restarted Lanczos Bidiagonalization Methods.

[13] Barreirinho, S., Ferro, A., Santos, M., Costa, E., Pinto-Basto, J., Sousa, A., Sequeiros, J., Maciel, P., Barbot, C., and Barbot, J. (2003). Inherited and acquired risk factors and their combined effects in pediatric stroke. *Pediatric Neurology*, 28(2):134–138.

[14] Bartoszewski, R. a., Jablonsky, M., Bartoszewska, S., Stevenson, L., Dai, Q., Kappes, J., Collawn, J. F., and Bebok, Z. (2010). A synonymous single nucleotide polymorphism in DeltaF508 CFTR alters the secondary structure of the mRNA and the expression of the mutant protein. *Journal of Biological Chemistry*, 285(37):28741–8.

[15] Bean, L. and Bayrak-Toydemir, P. (2014). American College of Medical Genetics and Genomics Standards and Guidelines for Clinical Genetics Laboratories, 2014 edition: technical standards and guidelines for Huntington disease. *Genetics in medicine : official journal of the American College of Medical Genetics*, 16(12):e2.

[16] Becker, K. G., Barnes, K. C., Bright, T. J., and Wang, S. A. (2004). The Genetic Association Database. *Nature Genetics*, 36(5):431–432.

[17] Bellos, E., Kumar, V., Lin, C., Maggi, J., Phua, Z. Y., Cheng, C. Y., Cheung, C. M. G., Hibberd, M. L., Wong, T. Y., Coin, L. J., and Davila, S. (2014). cnvCapSeq: detecting copy number variation in long-range targeted resequencing data. *Nucleic acids research*, 42(20):e158.

[18] Benjamini, Y. and Speed, T. P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Research*, 40(10).

[19] Benjannet, S., Hamelin, J., Chrétien, M., and Seidah, N. G. (2012). Loss- and gain-of-function PCSK9 variants: Cleavage specificity, dominant negative effects, and low density lipoprotein receptor (LDLR) degradation. *Journal of Biological Chemistry*, 287(40):33745–33755.

[20] Bertoli-Avella, A. M., Gillis, E., Morisaki, H., Verhagen, J. M., de Graaf, B. M., van de Beek, G., Gallo, E., Kruithof, B. P., Venselaar, H., Myers, L. A., Laga, S., Doyle, A. J., Oswald, G., van Cappellen, G. W., Yamanaka, I., van der Helm, R. M., Beverloo, B., de Klein, A., Pardo, L., Lammens, M., Evers, C., Devriendt, K., Dumoulein, M., Timmermans, J., Bruggenwirth, H. T., Verheijen, F., Rodrigus, I., Baynam, G., Kempers, M., Saenen, J., Van Craenenbroeck, E. M., Minatoya, K., Matsukawa, R., Tsukube, T., Kubo, N., Hofstra, R., Goumans, M. J., Bekkers, J. A., Roos-Hesselink, J. W., van de Laar, I. M., Dietz, H. C., Van Laer, L., Morisaki, T., Wessels, M. W., and Loeys, B. L. (2015). Mutations in a TGF-β ligand, TGFB3, cause syndromic aortic aneurysms and dissections. *Journal of the American College of Cardiology*, 65(13):1324–1336.

[21] Bhartiya, D., Jalali, S., Ghosh, S., and Scaria, V. (2014). Distinct patterns of genetic variations in potential functional elements in long noncoding RNAs. *Human Mutation*, 35(2):192–201.

[22] Biben, C., Weber, R., Kesteven, S., Stanley, E., McDonald, L., Elliott, D. A., Barnett, L., Köentgen, F., Robb, L., Feneley, M., and Harvey, R. P. (2000). Cardiac septal and valvular dysmorphogenesis in mice heterozygous for mutations in the homeobox gene Nkx2-5. *Circulation Research*, 87(10):888–895.

[23] Bingham, E., Kabán, A., and Fortelius, M. (2009). The aspect Bernoulli model: Multiple causes of presences and absences. *Pattern Analysis and Applications*, 12(1):55–78.

[24] Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I., and Edu, J. B. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

[25] Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory - COLT' 98*, pages 92–100.

[26] Boeva, V., Popova, T., Lienard, M., Toffoli, S., Kamal, M., Le Tourneau, C., Gentien, D., Servant, N., Gestraud, P., Frio, T. R., Hupé, P., Barillot, E., and Laes, J. F. (2014).

Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data. *Bioinformatics*, 30(24):3443–3450.

[27] Boileau, C., Guo, D. C., Hanna, N., Regalado, E. S., Detaint, D., Gong, L., Varret, M., Prakash, S. K., Li, A. H., D'Indy, H., Braverman, A. C., Grandchamp, B., Kwartler, C. S., Gouya, L., Santos-Cortez, R. L. P., Abifadel, M., Leal, S. M., Muti, C., Shendure, J., Gross, M. S., Rieder, M. J., Vahanian, A., Nickerson, D. A., Michel, J. B., Jondeau, G., and Milewicz, D. M. (2012). TGFB2 mutations cause familial thoracic aortic aneurysms and dissections associated with mild systemic features of Marfan syndrome. *Nature Genetics*, 44(8):916–921.

[28] Bolar, N. A. (2017). *Molecular Genetic Dissection of Disease in the Era of Next-generation Sequencing: Proefschrift*. PhD thesis.

[29] Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114–2120.

[30] Bonachea, E. M., Zender, G., White, P., Corsmeier, D., Newsom, D., Fitzgerald-Butt, S., Garg, V., and McBride, K. L. (2014). Use of a targeted, combinatorial next-generation sequencing approach for the study of bicuspid aortic valve. *BMC Medical Genomics*, 7(1).

[31] Bosse, K., Hans, C. P., Zhao, N., Koenig, S. N., Huang, N., Guggilam, A., LaHaye, S., Tao, G., Lucchesi, P. A., Lincoln, J., Lilly, B., and Garg, V. (2013). Endothelial nitric oxide signaling regulates Notch1 in aortic valve disease. *Journal of Molecular and Cellular Cardiology*, 60(1):27–35.

[32] Box, G. E. P. (1976). Science and Statistics. *Science and Statistics Author Journal of the American Statistical Association*, 71(356):791–799.

[33] Boycott, K. M., Vanstone, M. R., Bulman, D. E., and MacKenzie, A. E. (2013). Rare-disease genetics in the era of next-generation sequencing: Discovery to translation.

[34] Braverman, A. C., Güven, H., Beardslee, M. A., Makan, M., Kates, A. M., and Moon, M. R. (2005). The Bicuspid Aortic Valve. *Current Problems in Cardiology*, 30(9):470–522.

[35] Cai, J., Pardali, E., Sánchez-Duffhues, G., and Ten Dijke, P. (2012). BMP signaling in vascular diseases.

[36] Callewaert, B., Renard, M., Hucthagowder, V., Albrecht, B., Hausser, I., Blair, E., Dias, C., Albino, A., Wachi, H., Sato, F., Mecham, R. P., Loeys, B., Coucke, P. J., De

Paepe, A., and Urban, Z. (2011). New insights into the pathogenesis of autosomal-dominant cutis laxa with report of five ELN mutations. *Human Mutation*, 32(4):445–455.

[37] Caminsky, N. G., Mucaki, E. J., and Rogan, P. K. (2014). Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. *F1000Research*.

[38] Carmignac, V., Thevenon, J., Adès, L., Callewaert, B., Julia, S., Thauvin-Robinet, C., Gueneau, L., Courcet, J. B., Lopez, E., Holman, K., Renard, M., Plauchu, H., Plessis, G., De Backer, J., Child, A., Arno, G., Duplomb, L., Callier, P., Aral, B., Vabres, P., Gigot, N., Arbustini, E., Grasso, M., Robinson, P. N., Goizet, C., Baumann, C., Di Rocco, M., Sanchez Del Pozo, J., Huet, F., Jondeau, G., Collod-Beroud, G., Beroud, C., Amiel, J., Cormier-Daire, V., Rivière, J. B., Boileau, C., De Paepe, A., and Faivre, L. (2012). In-frame mutations in exon 1 of SKI cause dominant shprintzen-goldberg syndrome. *American Journal of Human Genetics*, 91(5):950–957.

[39] Carrington, M. and O'Brien, S. J. (2003). The influence of HLA genotype on AIDS. *Annual review of medicine*, 54:535–551.

[40] Cartegni, L., Chew, S. L., and Krainer, A. R. (2002). Listening to silence and understanding nonsense: Exonic mutations that affect splicing.

[41] Chabot, B. and Shkreta, L. (2016). Defective control of pre-messenger RNA splicing in human disease.

[42] Chamary, J. V., Parmley, J. L., and Hurst, L. D. (2006). Hearing silence: Non-neutral evolution at synonymous sites in mammals.

[43] Chen, B., Li, M., Wang, J., Shang, X., and Wu, F. X. (2015). A fast and high performance multiple data integration algorithm for identifying human disease genes. *BMC Medical Genomics*, 8(3).

[44] Chen, B., Wang, J., Li, M., and Wu, F.-X. (2014). Identifying disease genes by integrating multiple data sources. *BMC Medical Genomics*, 7(Suppl 2):S2.

[45] Chen, J., Bardes, E. E., Aronow, B. J., and Jegga, A. G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, 37(SUPPL. 2).

[46] Chen, Y., Wang, W., Zhou, Y., Shields, R., Chanda, S. K., Elston, R. C., and Li, J. (2011). In silico gene prioritization by integrating multiple data sources. *PLoS ONE*, 6(6).

[47] Chun, S. and Fay, J. C. (2009). Identification of deleterious mutations within three human genomes. *Genome Research*, 19(9):1553–1561.

[48] Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3):213–219.

[49] Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92.

[50] Clementi, M., Notari, L., Borghi, A., and Tenconi, R. (1996). Familial congenital bicuspid aortic valve: A disorder of uncertain inheritance. *American Journal of Medical Genetics*, 62(4):336–338.

[51] Costain, G., Silversides, C. K., and Bassett, A. S. (2016). The importance of copy number variation in congenital heart disease. *npj Genomic Medicine*, 1(1):16031.

[52] Cripe, L., Andelfinger, G., Martin, L. J., Shooner, K., and Benson, D. W. (2004). Bicuspid aortic valve is heritable. *Journal of the American College of Cardiology*, 44(1):138–143.

[53] De Los Campos, G., Pérez, P., Vazquez, A. I., and Crossa, J. (2013). Genome-enabled prediction using the BLR (Bayesian Linear Regression) R-package. *Methods in Molecular Biology*, 1019:299–320.

[54] Delcher, A. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research*, 30(11):2478–2483.

[55] Deng, Y., Wang, H., and Guo, B. (2015). BDD algorithms based on modularization for fault tree analysis. *Progress in Nuclear Energy*, 85:192–199.

[56] Depristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., Del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–501.

[57] Dickel, D. E., Barozzi, I., Zhu, Y., Fukuda-Yuzawa, Y., Osterwalder, M., Mannion, B. J., May, D., Spurrell, C. H., Plajzer-Frick, I., Pickle, C. S., Lee, E., Garvin, T. H., Kato,

M., Akiyama, J. A., Afzal, V., Lee, A. Y., Gorkin, D. U., Ren, B., Rubin, E. M., Visel, A., and Pennacchio, L. A. (2016). Genome-wide compendium and functional assessment of in vivo heart enhancers. *Nature Communications*, 7.

[58] Doyle, A. J., Doyle, J. J., Bessling, S. L., Maragh, S., Lindsay, M. E., Schepers, D., Gillis, E., Mortier, G., Homfray, T., Sauls, K., Norris, R. A., Huso, N. D., Leahy, D., Mohr, D. W., Caulfield, M. J., Scott, A. F., Destrée, A., Hennekam, R. C., Arn, P. H., Curry, C. J., Van Laer, L., McCallion, A. S., Loeys, B. L., and Dietz, H. C. (2012). Mutations in the TGF-β repressor SKI cause Shprintzen-Goldberg syndrome with aortic aneurysm. *Nature Genetics*, 44(11):1249–1254.

[59] Duan, J., Shi, J., Fiorentino, A., Leites, C., Chen, X., Moy, W., Chen, J., Alexandrov, B. S., Usheva, A., He, D., Freda, J., O'Brien, N. L., McQuillin, A., Sanders, A. R., Gershon, E. S., Delisi, L. E., Bishop, A. R., Gurling, H. M., Pato, M. T., Levinson, D. F., Kendler, K. S., Pato, C. N., and Gejman, P. V. (2014). A rare functional noncoding variant at the GWAS-Implicated MIR137/MIR2682 locus might confer risk to schizophrenia and bipolar disorder. *American Journal of Human Genetics*, 95(6):744–753.

[60] Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., and Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. *Genome biology*, 6(5):R44.

[61] Elston, R. C. and Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human heredity*, 21(6):523–542.

[62] Exome Aggregate Consortium (2016). ExAC Browser.

[63] Feingold, E. A., Good, P. J., Guyer, M. S., Kamholz, S., Liefer, L., Wetterstrand, K., Collins, F. S., Gingeras, T. R., Kampa, D., Sekinger, E. A., Cheng, J., Hirsch, H., Ghosh, S., Zhu, Z., Patel, S., Piccolboni, A., Yang, A., Tammana, H., Bekiranov, S., Kapranov, P., Harrison, R., Church, G., Struhl, K., Ren, B., Kim, T. H., Barrera, L. O., Qu, C., van Calcar, S., Luna, R., Glass, C. K., Rosenfeld, M. G., Guigo, R., Antonarakis, S. E., Birney, E., Brent, M., Pachter, L., Reymond, A., Dermitzakis, E. T., Dewey, C., Keefe, D., Denoeud, F., Lagarde, J., Ashurst, J., Hubbard, T., Wesselink, J. J., Castelo, R., Eyras, E., Myers, R. M., Sidow, A., Batzoglou, S., Trinklein, N. D., Hartman, S. J., Aldred, S. F., Anton, E., Schroeder, D. I., Marticke, S. S., Nguyen, L., Schmutz, J., Grimwood, J., Dickson, M., Cooper, G. M., Stone, E. A., Asimenos, G., Brudno, M., Dutta, A., Karnani, N., Taylor, C. M., Kim, H. K., Robins, G., Stamatoyannopoulos, G., Stamatoyannopoulos, J. A., Dorschner, M., Sabo, P., Hawrylycz, M., Humbert, R., Wallace, J., Yu, M., Navas, P. A., McArthur, M., Noble, W. S., Dunham, I., Koch, C. M.,

Andrews, R. M., Clelland, G. K., Wilcox, S., Fowler, J. C., James, K. D., Groth, P., Dovey, O. M., Ellis, P. D., Wraight, V. L., Mungall, A. J., Dhami, P., Fiegler, H., Langford, C. F., Carter, N. P., Vetrie, D., Snyder, M., Euskirchen, G., Urban, A. E., Nagalakshmi, U., Rinn, J., Popescu, G., Bertone, P., Hartman, S., Rozowsky, J., Emanuelsson, O., Royce, T., Chung, S., Gerstein, M., Lian, Z., Lian, J., Nakayama, Y., Weissman, S., Stolc, V., Tongprasit, W., Sethi, H., Jones, S., Marra, M., Shin, H., Schein, J., Clamp, M., Lindblad-Toh, K., Chang, J., Jaffe, D. B., Kamal, M., Lander, E. S., Mikkelsen, T. S., Vinson, J., Zody, M. C., de Jong, P. J., Osoegawa, K., Nefedov, M., Zhu, B., Baxevanis, A. D., Wolfsberg, T. G., Crawford, G. E., Whittle, J., Holt, I. E., Vasicek, T. J., Zhou, D., Luo, S., Green, E. D., Bouffard, G. G., Margulies, E. H., Portnoy, M. E., Hansen, N. F., Thomas, P. J., McDowell, J. C., Maskeri, B., Young, A. C., Idol, J. R., Blakesley, R. W., Schuler, G., Miller, W., Hardison, R., Elnitski, L., Shah, P., Salzberg, S. L., Pertea, M., Majoros, W. H., Haussler, D., Thomas, D., Rosenbloom, K. R., Clawson, H., Siepel, A., Kent, W. J., Weng, Z., Jin, S., Halees, A., Burden, H., Karaoz, U., Fu, Y., Yu, Y., Ding, C., Cantor, C. R., Kingston, R. E., Dennis, J., Green, R. D., Singer, M. A., Richmond, T. A., Norton, J. E., Farnham, P. J., Oberley, M. J., Inman, D. R., McCormick, M. R., Kim, H., Middle, C. L., Pirrung, M. C., Fu, X. D., Kwon, Y. S., Ye, Z., Dekker, J., Tabuchi, T. M., Gheldof, N., Dostie, J., and Harvey, S. C. (2004). The ENCODE (ENCyclopedia of DNA Elements) Project.

[64] Firth, H. V., Richards, S. M., Bevan, A. P., Clayton, S., Corpas, M., Rajan, D., Vooren, S. V., Moreau, Y., Pettett, R. M., and Carter, N. P. (2009). DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *American Journal of Human Genetics*, 84(4):524–533.

[65] Foffa, I., Ait Alì, L., Panesi, P., Mariani, M., Festa, P., Botto, N., Vecoli, C., and Andreassi, M. G. (2013). Sequencing of NOTCH1, GATA5, TGFBR1 and TGFBR2 genes in familial cases of bicuspid aortic valve. *BMC Medical Genetics*, 14(1).

[66] Förstermann, U. and Münzel, T. (2006). Endothelial nitric oxide synthase in vascular disease: From marvel to menace.

[67] Fowler, A., Mahamdallie, S., Ruark, E., Seal, S., Ramsay, E., Clarke, M., Uddin, I., Wylie, H., Strydom, A., Lunter, G., and Rahman, N. (2016). Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. *Wellcome Open Research*, 1:20.

[68] Freylikhman, O., Tatarinova, T., Smolina, N., Zhuk, S., Klyushina, A., Kiselev, A., Moiseeva, O., Sjoberg, G., Malashicheva, A., and Kostareva, A. (2014). Variants

in the NOTCH1 gene in patients with aortic coarctation. *Congenital heart disease*, 9(5):391–396.

[69] Galvin, K. M., Donovan, M. J., Lynch, C. A., Meyer, R. I., Paul, R. J., Lorenz, J. N., Fairchild-Huntress, V., Dixon, K. L., Dunmore, J. H., Gimbrone, M. A., Falb, D., and Huszar, D. (2000). A role for Smad6 in development and homeostasis of the cardiovascular system. *Nature Genetics*, 24(2):171–174.

[70] Garg, V., Muth, A. N., Ransom, J. F., Schluterman, M. K., Barnes, R., King, I. N., Grossfeld, P. D., and Srivastava, D. (2005). Mutations in NOTCH1 cause aortic valve disease. *Nature*, 437(7056):270–274.

[71] Garside, V. C., Chang, A. C., Karsan, A., and Hoodless, P. A. (2013). Co-ordinating Notch, BMP, and TGF-β signaling during heart valve development.

[72] Gillis, J. and Pavlidis, P. (2013). Assessing identity, redundancy and confounds in Gene Ontology annotations over time. *Bioinformatics*, 29(4):476–482.

[73] Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabási, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21):8685–8690.

[74] Groß, A., Hartung, M., Prüfer, K., Kelso, J., and Rahm, E. (2012). Impact of ontology evolution on functional analyses. *Bioinformatics (Oxford, England)*, 28(20):2671–7.

[75] Gudbjartsson, D. F., Jonasson, K., Frigge, M. L., and Kong, A. (2000). Allegro, a new computer program for multipoint linkage analysis.

[76] Guo, D. C., Gong, L., Regalado, E. S., Santos-Cortez, R. L., Zhao, R., Cai, B., Veeraraghavan, S., Prakash, S. K., Johnson, R. J., Muilenburg, A., Willing, M., Jondeau, G., Boileau, C., Pannu, H., Moran, R., Debacker, J., Bamshad, M. J., Shendure, J., Nickerson, D. A., Leal, S. M., Raman, C. S., Swindell, E. C., and Milewicz, D. M. (2015). MAT2A mutations predispose individuals to thoracic aortic aneurysms. *American Journal of Human Genetics*, 96(1):170–177.

[77] Guo, D.-C., Pannu, H., Tran-Fadulu, V., Papke, C. L., Yu, R. K., Avidan, N., Bourgeois, S., Estrera, A. L., Safi, H. J., Sparks, E., Amor, D., Ades, L., McConnell, V., Willoughby, C. E., Abuelo, D., Willing, M., Lewis, R. a., Kim, D. H., Scherer, S., Tung, P. P., Ahn, C., Buja, L. M., Raman, C. S., Shete, S. S., and Milewicz, D. M. (2007). Mutations in smooth muscle alpha-actin (ACTA2) lead to thoracic aortic aneurysms and dissections. *Nature genetics*, 39(12):1488–1493.

[78] Hanson, T. E., Branscum, A. J., and Johnson, W. O. (2014). Informative \$g\$ -Priors for Logistic Regression. *Bayesian Analysis*, 9(3):597–612.

[79] Hanyu, A., Ishidou, Y., Ebisawa, T., Shimanuki, T., Imamura, T., and Miyazono, K. (2001). The N domain of Smad7 is essential for specific inhibition of transforming growth factor-β signaling. *Journal of Cell Biology*, 155(6):1017–1027.

[80] Haraksingh, R. R., Abyzov, A., Gerstein, M., Urban, A. E., and Snyder, M. (2011). Genome-wide mapping of copy number variation in humans: Comparative analysis of high resolution array platforms. *PLoS ONE*, 6(11).

[81] Hata, A., Lagna, G., Massagué, J., and Hemmati-Brivanlou, A. (1998). Smad6 inhibits BMP/Smad1 signaling by specifically competing with the Smad4 tumor suppressor. *Genes and Development*, 12(2):186–197.

[82] Herder, C. and Roden, M. (2011). Genetics of type 2 diabetes: Pathophysiologic and clinical relevance.

[83] Hinton, R. B. (2012). Bicuspid aortic valve and thoracic aortic aneurysm: Three patient populations, two disease phenotypes, and one shared genotype.

[84] Hofmann, T. (2004). Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems*, 22(1):89–115.

[85] Hotelling, H. (1936). Relations Between Two Sets of Variates. *Biometrika*, 28(3/4):321.

[86] Huntington, K., Hunter, a. G., and Chan, K. L. (1997). A prospective study to assess the frequency of familial clustering of congenital bicuspid aortic valve. *Journal of the American College of Cardiology*, 30(7):1809–12.

[87] Hwang, S., Kim, E., Lee, I., and Marcotte, E. M. (2015). Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific Reports*, 5.

[88] Hyvärinen, A. and Oja, E. (2000). Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13(45):411–430.

[89] Iascone, M., Ciccone, R., Galletti, L., Marchetti, D., Seddio, F., Lincesso, A. R., Pezzoli, L., Vetro, A., Barachetti, D., Boni, L., Federici, D., Soto, A. M., Comas, J. V., Ferrazzi, P., and Zuffardi, O. (2012). Identification of de novo mutations and rare variants in hypoplastic left heart syndrome. *Clinical Genetics*, 81(6):542–554.

[90] Imamura, T., Takase, M., Nishihara, A., Oeda, E., Hanai, J. I., Kawabata, M., and Miyazono, K. (1997). Smad6 inhibits signalling by the TGF-β superfamily. *Nature*, 389(6651):622–626.

[91] Irtyuga, O., Malashicheva, A., Zhiduleva, E., Freylikhman, O., Rotar, O., Bäck, M., Tarnovskaya, S., Kostareva, A., and Moiseeva, O. (2017). NOTCH1 Mutations in Aortic Stenosis: Association with Osteoprotegerin/RANK/RANKL. *BioMed Research International*, 2017.

[92] Jefferies, J. L., Taylor, M. D., Rossano, J., Belmont, J. W., and Craigen, W. J. (2010). Novel cardiac findings in periventricular nodular heterotopia. *American Journal of Medical Genetics, Part A*, 152(1):165–168.

[93] Jeon, J.-P., Shim, S.-M., Nam, H.-Y., Ryu, G.-M., Hong, E.-J., Kim, H.-L., and Han, B.-G. (2010). Copy number variation at leptin receptor gene locus associated with metabolic traits and the risk of type 2 diabetes mellitus. *BMC genomics*, 11:426.

[94] Jiang, Y., Oron, T. R., Clark, W. T., Bankapur, A. R., D'Andrea, D., Lepore, R., Funk, C. S., Kahanda, I., Verspoor, K. M., Ben-Hur, A., Koo, D. C. E., Penfold-Brown, D., Shasha, D., Youngs, N., Bonneau, R., Lin, A., Sahraeian, S. M., Martelli, P. L., Profiti, G., Casadio, R., Cao, R., Zhong, Z., Cheng, J., Altenhoff, A., Skunca, N., Dessimoz, C., Dogan, T., Hakala, K., Kaewphan, S., Mehryary, F., Salakoski, T., Ginter, F., Fang, H., Smithers, B., Oates, M., Gough, J., Törönen, P., Koskinen, P., Holm, L., Chen, C. T., Hsu, W. L., Bryson, K., Cozzetto, D., Minneci, F., Jones, D. T., Chapman, S., Bkc, D., Khan, I. K., Kihara, D., Ofer, D., Rappoport, N., Stern, A., Cibrian-Uhalte, E., Denny, P., Foulger, R. E., Hieta, R., Legge, D., Lovering, R. C., Magrane, M., Melidoni, A. N., Mutowo-Meullenet, P., Pichler, K., Shypitsyna, A., Li, B., Zakeri, P., ElShal, S., Tranchevent, L. C., Das, S., Dawson, N. L., Lee, D., Lees, J. G., Sillitoe, I., Bhat, P., Nepusz, T., Romero, A. E., Sasidharan, R., Yang, H., Paccanaro, A., Gillis, J., Sedeño-Cortés, A. E., Pavlidis, P., Feng, S., Cejuela, J. M., Goldberg, T., Hamp, T., Richter, L., Salamov, A., Gabaldon, T., Marcet-Houben, M., Supek, F., Gong, Q., Ning, W., Zhou, Y., Tian, W., Falda, M., Fontana, P., Lavezzo, E., Toppo, S., Ferrari, C., Giollo, M., Piovesan, D., Tosatto, S. C., del Pozo, A., Fernández, J. M., Maietta, P., Valencia, A., Tress, M. L., Benso, A., Di Carlo, S., Politano, G., Savino, A., Rehman, H. U., Re, M., Mesiti, M., Valentini, G., Bargsten, J. W., van Dijk, A. D., Gemovic, B., Glisic, S., Perovic, V., Veljkovic, V., Veljkovic, N., Almeida-e Silva, D. C., Vencio, R. Z., Sharan, M., Vogel, J., Kansakar, L., Zhang, S., Vucetic, S., Wang, Z., Sternberg, M. J., Wass, M. N., Huntley, R. P., Martin, M. J., O'Donovan, C., Robinson, P. N., Moreau, Y., Tramontano, A., Babbitt, P. C., Brenner, S. E., Linial, M., Orengo, C. A.,

Rost, B., Greene, C. S., Mooney, S. D., Friedberg, I., and Radivojac, P. (2016). An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology*, 17(1).

[95] Johansson, L. F., van Dijk, F., de Boer, E. N., van Dijk-Bos, K. K., Jongbloed, J. D., van der Hout, A. H., Westers, H., Sinke, R. J., Swertz, M. A., Sijmons, R. H., and Sikkema-Raddatz, B. (2016). CoNVaDING: Single Exon Variation Detection in Targeted NGS Data. *Human Mutation*, 37(5):457–464.

[96] Jung, S. M., Lee, J. H., Park, J., Oh, Y. S., Lee, S. K., Park, J. S., Lee, Y. S., Kim, J. H., Lee, J. Y., Bae, Y. S., Koo, S. H., Kim, S. J., and Park, S. H. (2013). Smad6 inhibits non-canonical TGF-β1 signalling by recruiting the deubiquitinase A20 to TRAF6. *Nature Communications*, 4.

[97] Kaartinen, V., Dudas, M., Nagy, A., Sridurongrit, S., Lu, M. M., and Epstein, J. A. (2004). Cardiac outflow tract defects in mice lacking ALK2 in neural crestcells. *Development*, 131(14):3481–3490.

[98] Kamburov, A., Stelzl, U., Lehrach, H., and Herwig, R. (2013). The Consensus-PathDB interaction database: 2013 Update. *Nucleic Acids Research*, 41(D1).

[99] Kandoth, C., McLellan, M. D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J. F., Wyczalkowski, M. A., Leiserson, M. D., Miller, C. A., Welch, J. S., Walter, M. J., Wendl, M. C., Ley, T. J., Wilson, R. K., Raphael, B. J., and Ding, L. (2013). Mutational landscape and significance across 12 major cancer types. *Nature*, 502(7471):333–339.

[100] Kent, K. C., Crenshaw, M. L., Goh, D. L. M., and Dietz, H. C. (2013). Genotype-phenotype correlation in patients with bicuspid aortic valve and aneurysm. *Journal of Thoracic and Cardiovascular Surgery*, 146(1).

[101] Kent, W. J. (2002). BLAT - The BLAST-like alignment tool. *Genome Research*, 12(4):656–664.

[102] Kerstjens-Frederikse, W. S., Van De Laar, I. M., Vos, Y. J., Verhagen, J. M., Berger, R. M., Lichtenbelt, K. D., Klein Wassink-Ruiter, J. S., Van Der Zwaag, P. A., Du Marchie Sarvaas, G. J., Bergman, K. A., Bilardo, C. M., Roos-Hesselink, J. W., Janssen, J. H., Frohn-Mulder, I. M., Van Spaendonck-Zwarts, K. Y., Van Melle, J. P., Hofstra, R. M., and Wessels, M. W. (2016). Cardiovascular malformations caused by NOTCH1 mutations do not keep left: Data on 428 probands with left-sided CHD and their families. *Genetics in Medicine*, 18(9):914–923.

[103] Kim, M., Farnoud, F., and Milenkovic, O. (2015). HyDRA: Gene prioritization via hybrid distance-score rank aggregation. *Bioinformatics*, 31(7):1034–1043.

[104] Kircher, M., Witten, D. M., Jain, P., O'roak, B. J., Cooper, G. M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315.

[105] Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576.

[106] Koenig, S. N., Bosse, K., Majumdar, U., Bonachea, E. M., Radtke, F., and Garg, V. (2016). Endothelial notch1 is required for proper development of the semilunar valves and cardiac outflow tract. *Journal of the American Heart Association*, 5(4).

[107] Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the Interactome for Prioritization of Candidate Disease Genes. *American Journal of Human Genetics*, 82(4):949–958.

[108] Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., Black, G. C., Brown, D. L., Brudno, M., Campbell, J., Fitzpatrick, D. R., Eppig, J. T., Jackson, A. P., Freson, K., Girdea, M., Helbig, I., Hurst, J. A., Jähn, J., Jackson, L. G., Kelly, A. M., Ledbetter, D. H., Mansour, S., Martin, C. L., Moss, C., Mumford, A., Ouwehand, W. H., Park, S. M., Riggs, E. R., Scott, R. H., Sisodiya, S., Vooren, S. V., Wapner, R. J., Wilkie, A. O., Wright, C. F., Vulto-Van Silfhout, A. T., Leeuw, N. D., De Vries, B. B., Washingthon, N. L., Smith, C. L., Westerfield, M., Schofield, P., Ruef, B. J., Gkoutos, G. V., Haendel, M., Smedley, D., Lewis, S. E., and Robinson, P. N. (2014). The Human Phenotype Ontology project: Linking molecular biology and disease through phenotype data. *Nucleic Acids Research*, 42(D1).

[109] Koskinen, P., Törönen, P., Nokso-Koivisto, J., and Holm, L. (2015). PANNZER: High-throughput functional annotation of uncharacterized proteins in an error-prone environment. *Bioinformatics*, 31(10):1544–1552.

[110] Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., and Lander, E. S. (1996). Parametric and nonparametric linkage analysis: a unified multipoint approach. *American journal of human genetics*, 58(6):1347–63.

[111] Kumar, A. A., Holm, L., and Toronen, P. (2013). GOParGenPy: a high throughput method to generate gene ontology data matrices. *BMC bioinformatics*, 14(1):242.

[112] Kumar, A. A., Van Laer, L., Alaerts, M., Ardeshirdavani, A., Moreau, Y., Laukens, K., Loeys, B., and Vandeweyer, G. (2018). pBRIT: Gene Prioritization by Correlating Functional and Phenotypic Annotations Through Integrative Data Fusion. *Bioinformatics*.

[113] Laforest, B., Andelfinger, G., and Nemer, M. (2011). Loss of Gata5 in mice leads to bicuspid aortic valve. *Journal of Clinical Investigation*, 121(7):2876–2887.

[114] Laforest, B. and Nemer, M. (2011). GATA5 interacts with GATA4 and GATA6 in outflow tract development. *Developmental Biology*, 358(2):368–378.

[115] Lage, K., Karlberg, E. O., Størling, Z. M., Ólason, P. Í., Pedersen, A. G., Rigina, O., Hinsby, A. M., Tümer, Z., Pociot, F., Tommerup, N., Moreau, Y., and Brunak, S. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, 25(3):309–316.

[116] Lanckriet, G. R., Cristianini, N., Bartlett, P., El Ghaoui, L., Jordan, M. I., Jordan Lanckriet, M. I., and Ghaoui, E. (2004). Learning the Kernel Matrix with Semidefinite Programming. *Journal of Machine Learning Research*, 5:27–72.

[117] Lander, E. S., Waterman, M. S., Gu, H., Gnirke, A., Meissner, A., Lowe, C., Wenger, A., Bejerano, G., Rongione, M., Webster, M., Ji, H., Potash, J., Sabunciyan, S., and Feinberg, A. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, 2(3):231–239.

[118] Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., and Maglott, D. R. (2014). ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(D1).

[119] Lange, K., Papp, J. C., Sinsheimer, J. S., Sripracha, R., Zhou, H., and Sobel, E. M. (2013). Mendel: The Swiss army knife of genetic analysis programs. *Bioinformatics*, 29(12):1568–1570.

[120] Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25.

[121] Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J. D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G., Paschall, J., Ananiev, V., Flicek, P., and Church, D. M. (2013). DbVar and DGVa: Public archives for genomic structural variation. *Nucleic Acids Research*, 41(D1).

[122] Lathrop, G. M., Lalouel, J. M., Julier, C., and Ott, J. (1984). Strategies for multilocus linkage analysis in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 81(11):3443–3446.

[123] Lathrop, G. M., Lalouel, J. M., Julier, C., and Ott, J. (1985). Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *American journal of human genetics*, 37(3):482–98.

[124] Lee, S., Wu, M. C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775.

[125] Lee, T. C., Zhao, Y. D., Courtman, D. W., and Stewart, D. J. (2000). Abnormal aortic valve development in mice lacking endothelial nitric oxide synthase. *Circulation*, 101(20):2345–2348.

[126] Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., O'Donnell-Luria, A. H., Ware, J. S., Hill, A. J., Cummings, B. B., Tukiainen, T., Birnbaum, D. P., Kosmicki, J. A., Duncan, L. E., Estrada, K., Zhao, F., Zou, J., Pierce-Hoffman, E., Berghout, J., Cooper, D. N., Deflaux, N., DePristo, M., Do, R., Flannick, J., Fromer, M., Gauthier, L., Goldstein, J., Gupta, N., Howrigan, D., Kiezun, A., Kurki, M. I., Moonshine, A. L., Natarajan, P., Orozco, L., Peloso, G. M., Poplin, R., Rivas, M. A., Ruano-Rubio, V., Rose, S. A., Ruderfer, D. M., Shakir, K., Stenson, P. D., Stevens, C., Thomas, B. P., Tiao, G., Tusie-Luna, M. T., Weisburd, B., Won, H. H., Yu, D., Altshuler, D. M., Ardissino, D., Boehnke, M., Danesh, J., Donnelly, S., Elosua, R., Florez, J. C., Gabriel, S. B., Getz, G., Glatt, S. J., Hultman, C. M., Kathiresan, S., Laakso, M., McCarroll, S., McCarthy, M. I., McGovern, D., McPherson, R., Neale, B. M., Palotie, A., Purcell, S. M., Saleheen, D., Scharf, J. M., Sklar, P., Sullivan, P. F., Tuomilehto, J., Tsuang, M. T., Watkins, H. C., Wilson, J. G., Daly, M. J., and MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291.

[127] Li, B. and Leal, S. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83:311–321.

[128] Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760.

[129] Li, J., Lupat, R., Amarasinghe, K. C., Thompson, E. R., Doyle, M. A., Ryland, G. L., Tothill, R. W., Halgamuge, S. K., Campbell, I. G., and Gorringe, K. L. (2012). CONTRA: Copy number analysis for targeted resequencing. *Bioinformatics*, 28(10):1307–1313.

[130] Li, Y. and Patra, J. C. (2010). Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, 26(9):1219–1224.

[131] Liekens, A. M. L., De Knijf, J., Daelemans, W., Goethals, B., De Rijk, P., and Del-Favero, J. (2011). BioGraph: Unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biology*, 12(6).

[132] Lin, X., Liang, Y.-Y., Sun, B., Liang, M., Shi, Y., Brunicardi, F. C., Shi, Y., and Feng, X.-H. (2003). Smad6 recruits transcription corepressor CtBP to repress bone morphogenetic protein-induced transcription. *Molecular and cellular biology*, 23(24):9081–93.

[133] Lindner, T. H. and Hoffmann, K. (2005). easyLINKAGE: A PERL script for easy and automated two-/multi-point linkage analyses. *Bioinformatics*, 21(3):405–407.

[134] Lindsay, M. E., Schepers, D., Bolar, N. A., Doyle, J. J., Gallo, E., Fert-Bober, J., Kempers, M. J., Fishman, E. K., Chen, Y., Myers, L., Bjeda, D., Oswald, G., Elias, A. F., Levy, H. P., Anderlid, B. M., Yang, M. H., Bongers, E. M., Timmermans, J., Braverman, A. C., Canham, N., Mortier, G. R., Brunner, H. G., Byers, P. H., Van Eyk, J., Van Laer, L., Dietz, H. C., and Loeys, B. L. (2012). Loss-of-function mutations in TGFB2 cause a syndromic presentation of thoracic aortic aneurysm. *Nature Genetics*, 44(8):922–927.

[135] Loeys, B. L., Chen, J., Neptune, E. R., Judge, D. P., Podowski, M., Holm, T., Meyers, J., Leitch, C. C., Katsanis, N., Sharifi, N., Xu, F. L., Myers, L. A., Spevak, P. J., Cameron, D. E., De Backer, J., Hellemans, J., Chen, Y., Davis, E. C., Webb, C. L., Kress, W., Coucke, P., Rifkin, D. B., De Paepe, A. M., and Dietz, H. C. (2005). A syndrome of altered cardiovascular, craniofacial, neurocognitive and skeletal development caused by mutations in TGFBR1 or TGFBR2. *Nature Genetics*, 37(3):275–281.

[136] Loscalzo, M. L., Goh, D. L. M., Loeys, B., Kent, K. C., Spevak, P. J., and Dietz, H. C. (2007). Familial thoracic aortic dilation and bicommissural aortic valve: A prospective analysis of natural history and inheritance. *American Journal of Medical Genetics, Part A*, 143(17):1960–1967.

[137] Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., and Mundlos, S. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025.

[138] MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L., and Scherer, S. W. (2014). The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Research*, 42(D1).

[139] Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genetics*, 5(2).

[140] Makkar, P., Metpally, R. P. R., Sangadala, S., and Reddy, B. V. B. (2009). Modeling and analysis of MH1 domain of Smads and their interaction with promoter DNA sequence motif. *Journal of Molecular Graphics and Modelling*, 27(7):803–812.

[141] Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10.

[142] Martin, P. S., Kloesel, B., Norris, R. A., Lindsay, M., Milan, D., Body, S. C., and Maslen, C. L. (2015). Embryonic Development of the Bicuspid Aortic Valve. *J. Cardiovasc. Dev. Dis*, 2:248–272.

[143] Mavaddat, N., Antoniou, A. C., Easton, D. F., and Garcia-Closas, M. (2010). Genetic susceptibility to breast cancer.

[144] McCarthy, D. J., Humburg, P., Kanapin, A., Rivas, M. A., Gaulton, K., Cazier, J. B., and Donnelly, P. (2014). Choice of transcripts and software has a large effect on variant annotation. *Genome Medicine*, 6(3).

[145] McKellar, S. H., Tester, D. J., Yagubyan, M., Majumdar, R., Ackerman, M. J., and Sundt, T. M. (2007). Novel NOTCH1 mutations in patients with bicuspid aortic valve disease and thoracic aortic aneurysms. *The Journal of Thoracic and Cardiovascular Surgery*, 134(2):290–296.

[146] McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, 17(1).

[147] Micha, D., Guo, D.-c., Hilhorst-Hofstee, Y., van Kooten, F., Atmaja, D., Overwater, E., Cayami, F. K., Regalado, E. S., van Uffelen, R., Venselaar, H., Faradz, S. M., Vriend, G., Weiss, M. M., Sistermans, E. A., Maugeri, A., Milewicz, D. M., Pals, G., and van Dijk, F. S. (2015). SMAD2 Mutations Are Associated with Arterial Aneurysms and Dissections. *Human Mutation*, 36(12):1145–1149.

[148] Mohamed, S. A., Aherrahrou, Z., Liptau, H., Erasmi, A. W., Hagemann, C., Wrobel, S., Borzym, K., Schunkert, H., Sievers, H. H., and Erdmann, J. (2006). Novel missense

mutations (p.T596M and p.P1797H) in NOTCH1 in patients with bicuspid aortic valve. *Biochemical and Biophysical Research Communications*, 345(4):1460–1465.

[149] Mommersteeg, M. T., Yeh, M. L., Parnavelas, J. G., and Andrews, W. D. (2015). Disrupted Slit-Robo signalling results in membranous ventricular septum defects and bicuspid aortic valves. *Cardiovascular Research*, 106(1):55–66.

[150] Monteferrario, D., Bolar, N. A., Marneth, A. E., Hebeda, K. M., Bergevoet, S. M., Veenstra, H., Laros-van Gorkom, B. A., MacKenzie, M. A., Khandanpour, C., Botezatu, L., Fransen, E., Van Camp, G., Duijnhouwer, A. L., Salemink, S., Willemsen, B., Huls, G., Preijers, F., Van Heerde, W., Jansen, J. H., Kempers, M. J., Loeys, B. L., Van Laer, L., and Van der Reijden, B. A. (2014). A Dominant-Negative <i>GFI1B</i> Mutation in the Gray Platelet Syndrome. *New England Journal of Medicine*, 370(3):245–253.

[151] Moreau, Y. and Tranchevent, L. C. (2012). Computational tools for prioritizing candidate genes: Boosting disease gene discovery.

[152] Morgenthaler, S. and Thilly, W. G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1-2):28–56.

[153] Morris, A. P. and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, 34(2):188–193.

[154] Murakami, G. (2003). Cooperative Inhibition of Bone Morphogenetic Protein Signaling by Smurf1 and Inhibitory Smads. *Molecular Biology of the Cell*, 14(7):2809–2817.

[155] Myers, R. H. (2004). Huntington's disease genetics. *NeuroRx : the journal of the American Society for Experimental NeuroTherapeutics*, 1(2):255–62.

[156] Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., and Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genetics*, 7(3).

[157] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

[158] Ng, P. C. and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Research*, 31(13):3812–3814.

[159] Nistri, S., Porciani, M. C., Attanasio, M., Abbate, R., Gensini, G. F., and Pepe, G. (2012). Association of Marfan syndrome and bicuspid aortic valve: Frequency and outcome.

[160] O'Brien, J. and Okada, S. (1970). Tay-Sachs Disease: Detection of Heterozygotes and Homozygotes by Serum Hexosaminidase Assay. *New England Journal of Medicine*, 283(1):15–20.

[161] Oliveira-Paula, G. H., Lacchini, R., and Tanus-Santos, J. E. (2016). Endothelial nitric oxide synthase: From biochemistry and gene structure to clinical implications of NOS3 polymorphisms.

[162] Oller, J., Méndez-Barbero, N., Ruiz, E. J., Villahoz, S., Renard, M., Canelas, L. I., Briones, A. M., Alberca, R., Lozano-Vidal, N., Hurlé, M. A., Milewicz, D., Evangelista, A., Salaices, M., Nistal, J. F., Jiménez-Borreguero, L. J., De Backer, J., Campanero, M. R., and Redondo, J. M. (2017). Nitric oxide mediates aortic disease in mice deficient in the metalloprotease Adamts1 and in a mouse model of Marfan syndrome. *Nature Medicine*, 23(2):200–212.

[163] Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572.

[164] Palfi, S. and Jarraya, B. (2008). Huntington's disease: Genetics lends a hand.

[165] Pankratz, N., Dumitriu, A., Hetrick, K. N., Sun, M., Latourelle, J. C., Wilk, J. B., Halter, C., Doheny, K. F., Gusella, J. F., Nichols, W. C., Myers, R. H., Foroud, T., and DeStefano, A. L. (2011). Copy number variation in familial parkinson disease. *PLoS ONE*, 6(8).

[166] Pavlidis, P., Weston, J., Cai, J., and Noble, W. S. (2002). Learning Gene Functional Classifications from Multiple Data Types. *Journal of Computational Biology*, 9(2):401–411.

[167] Pedrotti, S. and Cooper, T. A. (2014). In Brief: (Mis)splicing in disease. *Journal of Pathology*, 233(1):1–3.

[168] Pepe, G., Nistri, S., Giusti, B., Sticchi, E., Attanasio, M., Porciani, C., Abbate, R., Bonow, R. O., Yacoub, M., and Gensini, G. F. (2014). Identification of fibrillin 1 gene mutations in patients with bicuspid aortic valve (BAV) without Marfan syndrome. *BMC Medical Genetics*, 15(1).

[169] Piñero, J., Queralt-Rosinach, N., Bravo, À., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F., and Furlong, L. I. (2015). DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015.

[170] Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T. R., Correia, C., Abrahams, B. S., Almeida, J., Bacchelli, E., Bader, G. D., Bailey, A. J., Baird, G., Battaglia, A., Berney, T., Bolshakova, N., Bölte, S., Bolton, P. F., Bourgeron, T., Brennan, S., Brian, J., Bryson, S. E., Carson, A. R., Casallo, G., Casey, J., Chung, B. H., Cochrane, L., Corsello, C., Crawford, E. L., Crossett, A., Cytrynbaum, C., Dawson, G., De Jonge, M., Delorme, R., Drmic, I., Duketis, E., Duque, F., Estes, A., Farrar, P., Fernandez, B. A., Folstein, S. E., Fombonne, E., Freitag, C. M., Gilbert, J., Gillberg, C., Glessner, J. T., Goldberg, J., Green, A., Green, J., Guter, S. J., Hakonarson, H., Heron, E. A., Hill, M., Holt, R., Howe, J. L., Hughes, G., Hus, V., Igliozzi, R., Kim, C., Klauck, S. M., Kolevzon, A., Korvatska, O., Kustanovich, V., Lajonchere, C. M., Lamb, J. A., Laskawiec, M., Leboyer, M., Le Couteur, A., Leventhal, B. L., Lionel, A. C., Liu, X. Q., Lord, C., Lotspeich, L., Lund, S. C., Maestrini, E., Mahoney, W., Mantoulan, C., Marshall, C. R., McConachie, H., McDougle, C. J., McGrath, J., McMahon, W. M., Merikangas, A., Migita, O., Minshew, N. J., Mirza, G. K., Munson, J., Nelson, S. F., Noakes, C., Noor, A., Nygren, G., Oliveira, G., Papanikolaou, K., Parr, J. R., Parrini, B., Paton, T., Pickles, A., Pilorge, M., Piven, J., Ponting, C. P., Posey, D. J., Poustka, A., Poustka, F., Prasad, A., Ragoussis, J., Renshaw, K., Rickaby, J., Roberts, W., Roeder, K., Roge, B., Rutter, M. L., Bierut, L. J., Rice, J. P., Salt, J., Sansom, K., Sato, D., Segurado, R., Sequeira, A. F., Senman, L., Shah, N., Sheffield, V. C., Soorya, L., Sousa, I., Stein, O., Sykes, N., Stoppioni, V., Strawbridge, C., Tancredi, R., Tansey, K., Thiruvahindrapduram, B., Thompson, A. P., Thomson, S., Tryfon, A., Tsiantis, J., Van Engeland, H., Vincent, J. B., Volkmar, F., Wallace, S., Wang, K., Wang, Z., Wassink, T. H., Webber, C., Weksberg, R., Wing, K., Wittemeyer, K., Wood, S., Wu, J., Yaspan, B. L., Zurawiecki, D., Zwaigenbaum, L., Buxbaum, J. D., Cantor, R. M., Cook, E. H., Coon, H., Cuccaro, M. L., Devlin, B., Ennis, S., Gallagher, L., Geschwind, D. H., Gill, M., Haines, J. L., Hallmayer, J., Miller, J., Monaco, A. P., Nurnberger, J. I., Paterson, A. D., Pericak-Vance, M. A., Schellenberg, G. D., Szatmari, P., Vicente, A. M., Vieland, V. J., Wijsman, E. M., Scherer, S. W., Sutcliffe, J. S., and Betancur, C. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*, 466(7304):368–372.

[171] Plagnol, V., Curtis, J., Epstein, M., Mok, K. Y., Stebbings, E., Grigoriadou, S., Wood, N. W., Hambleton, S., Burns, S. O., Thrasher, A. J., Kumararatne, D., Doffinger, R., and Nejentsev, S. (2012). A robust model for read count data in exome sequenc-

ing experiments and implications for copy number variant calling. *Bioinformatics*, 28(21):2747–2754.

[172] Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., and Banks, E. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *doi.org*, page 201178.

[173] Pray, L. A. (2008). Discovery of DNA Structure and Function: Watson and Crick. *Nature Education*, 1(1):6.

[174] Preuss, C., Capredon, M., Wünnemann, F., Chetaille, P., Prince, A., Godard, B., Leclerc, S., Sobreira, N., Ling, H., Awadalla, P., Thibeault, M., Khairy, P., Loeys, B., Dietz, H., Franco-Cereceda, A., Eriksson, P., Mohamed, S. A., McCallion, A. S., Mertens, L., Van Laer, L., Mital, S., Samuels, M. E., and Andelfinger, G. (2016). Family Based Whole Exome Sequencing Reveals the Multifaceted Role of Notch Signaling in Congenital Heart Disease. *PLoS Genetics*, 12(10).

[175] Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L. J., and Sunyaev, S. R. (2010). Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *American Journal of Human Genetics*, 86(6):832–838.

[176] Proost, D., Vandeweyer, G., Meester, J. A. N., Salemink, S., Kempers, M., Ingram, C., Peeters, N., Saenen, J., Vrints, C., Lacro, R. V., Roden, D., Wuyts, W., Dietz, H. C., Mortier, G., Loeys, B. L., and Van Laer, L. (2015). Performant Mutation Identification Using Targeted Next-Generation Sequencing of 14 Thoracic Aortic Aneurysm Genes. *Human Mutation*, 36(8):808–814.

[177] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3):559–575.

[178] Quintero-Rivera, F., Xi, Q. J., Keppler-Noreuil, K. M., Lee, J. H., Higgins, A. W., Anchan, R. M., Roberts, A. E., Seong, I. S., Fan, X., Lage, K., Lu, L. Y., Tao, J., Hu, X., Berezney, R., Gelb, B. D., Kamp, A., Moskowitz, I. P., Lacro, R. V., Lu, W., Morton, C. C., Gusella, J. F., and Maas, R. L. (2015). MATR3 disruption in human and mouse associated with bicuspid aortic valve, aortic coarctation and patent ductus arteriosus. *Human Molecular Genetics*, 24(8):2375–2389.

[179] Rommens, J., Iannuzzi, M., Kerem, B., Drumm, M., Melmer, G., Dean, M., Rozmahel, R., Cole, J., Kennedy, D., Hidaka, N., and al. Et (1989). Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science*, 245(4922):1059–1065.

[180] Rossini, A. J., Tierney, L., and Li, N. (2007). Simple parallel statistical computing in R. *Journal of Computational and Graphical Statistics*, 16(2):399–420.

[181] Rumsfeld, J. S., Joynt, K. E., and Maddox, T. M. (2016). Big data analytics to improve cardiovascular care: Promise and challenges.

[182] Samorodnitsky, E., Datta, J., Jewell, B. M., Hagopian, R., Miya, J., Wing, M. R., Damodaran, S., Lippus, J. M., Reeser, J. W., Bhatt, D., Timmers, C. D., and Roychowdhury, S. (2015). Comparison of custom capture for targeted next-generation DNA sequencing. *Journal of Molecular Diagnostics*, 17(1):64–75.

[183] Sauna, Z. E. and Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease.

[184] Sayers, E. (2012). E-utilities Quick Start Entrez Programming Utilities Help Entrez Programming Utilities Help. *The Journal of Systems and Software*, 85:1930–1952.

[185] Schnoes, A. M., Brown, S. D., Dodevski, I., and Babbitt, P. C. (2009). Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology*, 5(12).

[186] Schriml, L. M., Arze, C., Nadendla, S., Chang, Y. W. W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W. A. (2012). Disease ontology: A backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1).

[187] Schwarz, J. M., R??delsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations.

[188] Scott, A., Petrykowska, H. M., Hefferon, T., Gotea, V., and Elnitski, L. (2012). Functional analysis of synonymous substitutions predicted to affect splicing of the CFTR gene. *Journal of Cystic Fibrosis*, 11(6):511–517.

[189] Seshan, V. E. and Olshen, A. B. (2014). DNAcopy : A Package for Analyzing DNA Copy Data. *Bioconductor Vignette*, pages 1–7.

[190] Sha, Q. and Zhang, S. (2014). A Novel Test for Testing the Optimally Weighted Combination of Rare and Common Variants Based on Data of Parents and Affected Children. *Genetic Epidemiology*, 38(2):135–143.

[191] Shaikh, T. H., Gai, X., Perin, J. C., Glessner, J. T., Xie, H., Murphy, K., O'Hara, R., Casalunovo, T., Conlin, L. K., D'Arcy, M., Frackelton, E. C., Geiger, E. A., Haldeman-Englert, C., Imielinski, M., Kim, C. E., Medne, L., Annaiah, K., Bradfield, J. P., Dabaghyan, E., Eckert, A., Onyiah, C. C., Ostapenko, S., Otieno, F. G., Santa, E., Shaner, J. L., Skraban, R., Smith, R. M., Elia, J., Goldmuntz, E., Spinner, N. B., Zackai, E. H., Chiavacci, R. M., Grundmeier, R., Rappaport, E. F., Grant, S. F., White, P. S., and Hakonarson, H. (2009). High-resolution mapping and analysis of copy number variations in the human genome: A data resource for clinical and research applications. *Genome Research*, 19(9):1682–1690.

[192] Sherry, S. T. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research*, 29(1):308–311.

[193] Sifrim, A., Popovic, D., Tranchevent, L. C., Ardeshirdavani, A., Sakai, R., Konings, P., Vermeesch, J. R., Aerts, J., De Moor, B., and Moreau, Y. (2013). EXtasy: Variant prioritization by genomic data fusion. *Nature Methods*, 10(11):1083–1086.

[194] Singleton, M. V., Guthery, S. L., Voelkerding, K. V., Chen, K., Kennedy, B., Margraf, R. L., Durtschi, J., Eilbeck, K., Reese, M. G., Jorde, L. B., Huff, C. D., and Yandell, M. (2014). Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *American Journal of Human Genetics*, 94(4):599–610.

[195] Smith, C. L. and Eppig, J. T. (2009). The mammalian phenotype ontology: Enabling robust annotation and comparative analysis. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 1(3):390–399.

[196] Smith, T. F., Waterman, M. (1981). Smith-Waterman Algorithm. *Mol. Biol. ()*, (147):195–197.

[197] Sobel, E. and Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American journal of human genetics*, 58(6):1323–1337.

[198] Syrbe, S., Hedrich, U. B. S., Riesch, E., Djémié, T., Müller, S., Møller, R. S., Maher, B., Hernandez-Hernandez, L., Synofzik, M., Caglayan, H. S., Arslan, M., Serratosa, J. M., Nothnagel, M., May, P., Krause, R., Löffler, H., Detert, K., Dorn, T., Vogt, H., Krämer, G., Schöls, L., Mullis, P. E., Linnankivi, T., Lehesjoki, A.-E., Sterbova, K., Craiu, D. C., Hoffman-Zacharska, D., Korff, C. M., Weber, Y. G., Steinlin, M., Gallati, S., Bertsche, A., Bernhard, M. K., Merkenschlager, A., Kiess, W., Balling, R., Barisic, N., Baulac, S., Caglayan, H. S., Craiu, D. C., De Jonghe, P., Depienne, C., Gormley, P.,

Guerrini, R., Helbig, I., Hjalgrim, H., Hoffman-Zacharska, D., Jähn, J., Klein, K. M., Koeleman, B. P. C., Komarek, V., Krause, R., LeGuern, E., Lehesjoki, A.-E., Lemke, J. R., Lerche, H., Marini, C., May, P., Møller, R. S., Muhle, H., Palotie, A., Pal, D., Rosenow, F., Selmer, K., Serratosa, J. M., Sisodiya, S. M., Stephani, U., Sterbova, K., Striano, P., Suls, A., Talvik, T., von Spiczak, S., G Weber, Y., Weckhuysen, S., Zara, F., Gonzalez, M., Züchner, S., Palotie, A., Suls, A., De Jonghe, P., Helbig, I., Biskup, S., Wolff, M., Maljevic, S., Schüle, R., Sisodiya, S. M., Weckhuysen, S., Lerche, H., and Lemke, J. R. (2015). De novo loss- or gain-of-function mutations in KCNA2 cause epileptic encephalopathy. *Nature Genetics*, 47(4):393–399.

[199] Szafranski, P., Dharmadhikari, A. V., Brosens, E., Gurha, P., Kolodziejska, K. E., Zhishuo, O., Dittwald, P., Majewski, T., Mohan, K. N., Chen, B., Person, R. E., Tibboel, D., de Klein, A., Pinner, J., Chopra, M., Malcolm, G., Peters, G., Arbuckle, S., Guiang, S. F., Hustead, V. A., Jessurun, J., Hirsch, R., Witte, D. P., Maystadt, I., Sebire, N., Fisher, R., Langston, C., Sen, P., and Stankiewicz, P. (2013). Small noncoding differentially methylated copy-number variants, including lncRNA genes, cause a lethal lung developmental disorder. *Genome research*, 23(1):23–33.

[200] Talevich, E., Shain, A. H., Botton, T., and Bastian, B. C. (2016). CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Computational Biology*, 12(4).

[201] Tan, H. L., Glen, E., Töpf, A., Hall, D., O'Sullivan, J. J., Sneddon, L., Wren, C., Avery, P., Lewis, R. J., ten Dijke, P., Arthur, H. M., Goodship, J. A., and Keavney, B. D. (2012). Nonsynonymous variants in the SMAD6 gene predispose to congenital cardiovascular malformation. *Human Mutation*, 33(4):720–727.

[202] Tan, R., Wang, Y., Kleinstein, S. E., Liu, Y., Zhu, X., Guo, H., Jiang, Q., Allen, A. S., and Zhu, M. (2014). An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. *Human Mutation*, 35(7):899–907.

[203] The UniProt Consortium (2015). UniProt: a hub for protein information. *Nucleic acids research*, 43(Database issue):D204–12.

[204] Thiele, H. and Nürnberg, P. (2005). HaploPainter: A tool for drawing pedigrees with complex haplotypes. *Bioinformatics*, 21(8):1730–1732.

[205] Thomas, P. S., Sridurongrit, S., Ruiz-Lozano, P., and Kaartinen, V. (2012). Deficient signaling via Alk2 (Acvr1) leads to Bicuspid aortic valve development. *PLoS ONE*, 7(4).

[206] Tomlinson, I. P., Novelli, M. R., and Bodmer, W. F. (1996). The mutation rate and cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 93(25):14800–14803.

[207] Topper, J. N., Cai, J., Qiu, Y., Anderson, K. R., Xu, Y. Y., Deeds, J. D., Feeley, R., Gimeno, C. J., Woolf, E. A., Tayber, O., Mays, G. G., Sampson, B. A., Schoen, F. J., Gimbrone Jr., M. A., and Falb, D. (1997). Vascular MADs: two novel MAD-related genes selectively inducible by flow in human vascular endothelium. *Proc Natl Acad Sci U S A*, 94(17):9314–9319.

[208] Tranchevent, L. C., Ardeshirdavani, A., ElShal, S., Alcaide, D., Aerts, J., Auboeuf, D., and Moreau, Y. (2016). Candidate gene prioritization with Endeavour. *Nucleic acids research*, 44(W1):W117–W121.

[209] Tranchevent, L. C., Capdevila, F. B., Nitsch, D., de Moor, B., de Causmaecker, P., and Moreau, Y. (2011). A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics*, 12(1):22–32.

[210] Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., and Rozen, S. G. (2012). Primer3-new capabilities and interfaces. *Nucleic Acids Research*, 40(15).

[211] Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., Cau, P., Remy, E., and Baudot, A. (2017). Random Walk With Restart On Multiplex And Heterogeneous Biological Networks. *bioRxiv*, pages 1–31.

[212] Van De Laar, I. M., Oldenburg, R. A., Pals, G., Roos-Hesselink, J. W., De Graaf, B. M., Verhagen, J. M., Hoedemaekers, Y. M., Willemsen, R., Severijnen, L. A., Venselaar, H., Vriend, G., Pattynama, P. M., Collée, M., Majoor-Krakauer, D., Poldermans, D., Frohn-Mulder, I. M., Micha, D., Timmermans, J., Hilhorst-Hofstee, Y., Bierma-Zeinstra, S. M., Willems, P. J., Kros, J. M., Oei, E. H., Oostra, B. A., Wessels, M. W., and Bertoli-Avella, A. M. (2011). Mutations in SMAD3 cause a syndromic form of aortic aneurysms and dissections with early-onset osteoarthritis.

[213] van de Laar, I. M., van der Linde, D., Oei, E. H., Bos, P. K., Bessems, J. H., Bierma-Zeinstra, S. M., van Meer, B. L., Pals, G., Oldenburg, R. A., Bekkers, J. A., Moelker, A., de Graaf, B. M., Matyas, G., Frohn-Mulder, I. M., Timmermans, J., Hilhorst-Hofstee, Y., Cobben, J. M., Bruggenwirth, H. T., van Laer, L., Loeys, B., De Backer, J., Coucke, P. J., Dietz, H. C., Willems, P. J., Oostra, B. A., De Paepe, A., Roos-Hesselink, J. W., Bertoli-Avella, A. M., and Wessels, M. W. (2012). Phenotypic spectrum of the SMAD3-related aneurysms-osteoarthritis syndrome. *Journal of Medical Genetics*, 49(1):47–57.

[214] Vandeweyer, G., Van Laer, L., Loeys, B., Van den Bulcke, T., and Kooy, R. F. (2014). VariantDB: A flexible annotation and filtering portal for next generation sequencing data. *Genome Medicine*, 6(10).

[215] Verstraeten, A., Roos-Hesselink, J., and Loeys, B. (2016). Bicuspid aortic valve. In *Clinical Cardiogenetics*, pages 295–308. Springer.

[216] Visel, A., Rubin, E. M., and Pennacchio, L. A. (2009). Genomic views of distant-acting enhancers.

[217] Wadi, L., Meyer, M., Weiser, J., Stein, L. D., and Reimand, J. (2016). Impact of outdated gene annotations on pathway enrichment analysis.

[218] Walter, M. J., Payton, J. E., Ries, R. E., Shannon, W. D., Deshmukh, H., Zhao, Y., Baty, J., Heath, S., Westervelt, P., Watson, M. a., Tomasson, M. H., Nagarajan, R., O'Gara, B. P., Bloomfield, C. D., Mrózek, K., Selzer, R. R., Richmond, T. a., Kitzman, J., Geoghegan, J., Eis, P. S., Maupin, R., Fulton, R. S., McLellan, M., Wilson, R. K., Mardis, E. R., Link, D. C., Graubert, T. a., DiPersio, J. F., and Ley, T. J. (2009). Acquired copy number alterations in adult acute myeloid leukemia genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 106(31):12950–5.

[219] Wang, G. T., Zhang, D., Li, B., Dai, H., and Leal, S. M. (2015). Collapsed haplotype pattern method for linkage analysis of next-generation sequence data. *European Journal of Human Genetics*, 23(12):1739–1743.

[220] Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 38(16).

[221] Watson, J. D. and Crick, F. H. C. (1953). Molecular structure of nucleic acids.

[222] Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, 89(1):82–93.

[223] Wu, X., Jiang, R., Zhang, M. Q., and Li, S. (2008). Network-based global inference of human disease genes. *Molecular Systems Biology*, 4.

[224] Yang, H., Robinson, P. N., and Wang, K. (2015). Phenolyzer: Phenotype-based prioritization of candidate genes for human diseases. *Nature Methods*, 12(9):841–843.

[225] Yu, W., Gwinn, M., Clyne, M., Yesupriya, A., and Khoury, M. J. (2008). A navigator for human genome epidemiology [2].

[226] Zawistowski, M., Gopalakrishnan, S., Ding, J., Li, Y., Grimm, S., and Zöllner, S. (2010). Extending rare-variant testing strategies: Analysis of noncoding sequence and imputed genotypes. *American Journal of Human Genetics*, 87(5):604–617.

[227] Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., Gil, L., Gordon, L., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., To, J. K., Laird, M. R., Lavidas, I., Liu, Z., Loveland, J. E., Maurel, T., McLaren, W., Moore, B., Mudge, J., Murphy, D. N., Newman, V., Nuhn, M., Ogeh, D., Ong, C. K., Parker, A., Patricio, M., Riat, H. S., Schuilenburg, H., Sheppard, D., Sparrow, H., Taylor, K., Thormann, A., Vullo, A., Walts, B., Zadissa, A., Frankish, A., Hunt, S. E., Kostadima, M., Langridge, N., Martin, F. J., Muffato, M., Perry, E., Ruffier, M., Staines, D. M., Trevanion, S. J., Aken, B. L., Cunningham, F., Yates, A., and Flicek, P. (2017). Ensembl 2018. *Nucleic Acids Research*.

[228] Zhang, W., Chen, Y., Sun, F., and Jiang, R. (2011). DomainRBF: A Bayesian regression approach to the prioritization of candidate domains for complex diseases. *BMC Systems Biology*, 5.

[229] Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives - Springer. *BMC bioinformatics*, 14 Suppl 1(Suppl 11):S1.

[230] Zhu, X., Vondrick, C., Ramanan, D., and Fowlkes, C. (2012). Do We Need More Training Data or Better Models for Object Detection? *Procedings of the British Machine Vision Conference 2012 (BMVC12)*, pages 80.1–80.11.

[231] Žitnik, M., Nam, E. A., Dinh, C., Kuspa, A., Shaulsky, G., and Zupan, B. (2015). Gene Prioritization by Compressive Data Fusion and Chaining. *PLoS Computational Biology*, 11(10).

Acknowledgments

Acknowledgments

This thesis work is an important milestone in my life and it would not have been possible without contribution of several people.

First and foremost I would like to thank and pay tribute to my promoter Prof. Dr. Bart Loyes for giving me an excellent opportunity to pursue doctoral research in his cardiogenetics group (formally aorta research group). Without his guidance, critical reviews and support I could never imagine finishing up my thesis. I am very grateful to you for your generous act by twice extending my contract and helping me overcome the unwanted 'visa' related stress during the course of thesis writing. Beside academic support, I admire his generous and caring side which he exhibited not only upon me but for all other group members as well. Be it offering little snacks while working late in the lab or inviting for several football matches at Gent pushed me to work harder and give my full potential towards finishing up of the projects. I will carry with me these experiences and learning which will definitely help me shaping up my career in science and pursue more fundamental research.

Working hard is the norm in any scientific discipline but how to work smart can be best seen in my co-promotors Geert and Maaike. I really feel blessed to have them by my side during nerve cracking marathon thesis writing process. The precise planning, fruitful discussion and timely encouragements deeply motivated me. Be it resolving issues related to servers or simplifying complex concepts or improving my writing skills, each time I was enriched with a new learning experience. I hope to carry these virtues throughout my academic career. Additionally, I would like to thank Aline for providing time to time useful tips and tricks in addressing issues in the group.

I am grateful to Dr. Lut Van Laer for being co-promoter during my first two years of doctoral studies. Your guidance and support throughout this tenure immensely benefited me in understanding the bridge between the clinical and biological aspects of research here at CMG. Thank you for the small token help on my first day arrival here at Antwerp in 2013 and will cherish these memories throughout my life.

I would like to thank the members of my internal jury members Prof. Kris Laukens and Dr. Wim Wuyts for critically reviewing my progress reports, providing useful suggestions for improving my manuscripts and current thesis as well. I would also like to thank external jury members: Prof. Christian Gilissen and Dr. Alejandro Sifrim for patiently reading up my thesis, acknowledging it and providing very useful comments related to underlying research work.

I have always believed and strive for team work and this research could not be successful without the support of my research group members. I convey my special hug and thanks to all my aorta research group G.230 room members: Nikhita, Dorien.S, Dorien. P, Liesbeth, Jeannette, Ilse Luyckx, Elyssa, Gerarda, Charlotte and Eva for chit-chatting, cracking jokes, morale boosting which I greatly enjoyed and found much needed support in carrying out research. Thanks to Nikhita for passing the mantle of being the last person to routine tasks for closing CMG at the end of day. Your never give up attitude, persistence and being articulate deeply helped in shaping up my research and learnt a lot from you. I still remember the way you and Liesbeth contributed towards evaluation of gene prioritization tool sitting next to me which immensely helped in achieving one of the best results. Liesbeth, working with you was an exhilarating experience and your suggestions, tips and tricks during all MIBAVA, BAV meetings indeed helped a lot to me.

With ever smiling Dorien.S it was always a pleasure to talk and felt next door friend to knock for any help related to research, discussing off-science topics, organizing after work parties, cracking jokes with sarcasm made enjoyable experience in the lab. Dorien. P, thank you for support and help in understanding the concepts behind CNV related data analysis during initial days. Jeannette, special thank you for all the help and support you provided during my doctoral studies (especially in final stages of writing thesis). Moreover I really enjoyed your company and the fun we all had at the #ESHG-2016 meeting and will definitely cherish those memories for long time. Thanks to Gerarda for being my next neighbor in this room and being digital RJ made me to listen all dutch/english radio songs which sometimes provided soothing relief while writing the codes. The #Efteling trip which you organized was superb and will cherish those memories for long.

Special thanks to Ilse.L for being my fellow companion as MIBAVA presenter/traveler/Volleyballer/shuttler/mid-way biker and soon a thesis defender. Wish you a good luck for it!! Your helping hands in organizing things (along with Maaike, Aline) meant a lot to me specially during these crucial time of my defense. Had really great time during morning run at Barcelona-ESHG conference. Thanks for all the memories and will cherish those lifelong. Elyssa also in the league of being next thesis defender, thank you for all the talks and discussions which indeed helped me a lot. Had really a great time with you too at all MIBAVA, ESHG meetings. Wish you all the best with your thesis and future thereafter. Thanks to Charlotte for all the help and conversations in my initial days of research at CMG. I still cherish the memories of small sapling plant which you gave but unfortunately could not sustain it further.

Thanks to Ola, Celine, Ellen, Joke and Ann-Katrin for providing company at the fag end corner room with all lunch-talks, discussions on various related/un-related topics which indeed helped in shaping up of the momentum. Thanks to current new members Ewa, Melanie, Eline, Jarl for giving the company and wish you all the best with your current research work. I would like to thank all the members of MIBAVA-Leducq consortium members for providing a true international collaborative research ambience that helped me gaining insight of clinical and fundamental research.

I am grateful to Prof. Geert Mortier, Prof. Frank, Prof. Guy and Prof. Wim for providing excellent courses in genetics and useful feedbacks during genetic seminars that helped me in understanding and learning fundamental concepts.

I convey my thanks to current members of Bioinformatics group Matthias-B, Mathias-H, Joe, Jules and Philip for wonderful memories and discussions over the past one year. Matthias-B you are a true friend, an excellent researcher and great person whose company I really cherished and felt honored sharing desk space. Thanks for all the precious bioinformatics, biology oriented ideas, discussions, cracking sarcastic jokes were really awesome. This indeed helped me shaping several ideas especially during development of varAmpliCNV tool. Wish you loads of success and all the best for future with research and biking. Thanks to Joe and Mathias-H for lovely discussions, weekend outing and helping me with several other random stuffs. Special thanks to Philip and Jules for helping with discussions and several server related issues. You guys were very helpful and I will always cherish those lighter moments spent with you guys which made a conducive environment for doing research.

I convey my thanks to Erik for helping a lot during the courses in Statistics, Linkage analysis and other random topics discussions. These were indeed very useful for me when applying them in my research work. Arvid, thank you for being an excellent research manager and helping and guiding all of us with several tasks. I learned a lot from you during organization of NGS course during past two years. Thanks to all the other including current and former colleagues: Nele, Haane, Ellen, Manou, Ken, Timmon, Eveline, Marieke, Lieslot, Gretl, Igor, Raphael, Anne, Ilse VW, Ankke, Elisa, Esther, Amber, Silke, Sara, Gitta, Elke, Jolien for productive discussions either related to my research work, thesis related queries or some other random stuffs which were indeed very helpful.

I would like to thank all the technical support staff members of CMG for providing all help related to my access at CMG and resolving daily routine tasks.

I would like to thank all my teachers and research labs at University of Helsinki which provided useful bioinformatics courses/experiences that immensely helped me during my doctoral studies. Special thanks to Prof. Jarkko Salojarvi at plant stress group as a mentor, friend and teacher who gave brief window of opportunity in September 2012 eventually opening the gateway for me to learn statistical machine learning, probabilistic modeling in the domain of bioinformatics. I am grateful to you for incorporating me into GnViz project which in fact provided me the platform to work upon the gene prioritization tool. I immensely benefited from your PhD thesis for learning fundamentals of Bayesian statistics and it's applications and it was worthy treasure. Thanks to my other colleagues and friends at Helsinki: Shaman, Chakri, Poojitha, Naveen and Kirti Prakash for wonderful philosophical discussions.

I would like convey my thanks to Shinjini and Abhishek for being my first friends in Belgium and had really great time discussing, analyzing other worldly affairs besides research. Shinjini thanks for delicious food and motivating me to apply for current doctoral position in Belgium!!

I express my thanks to all my Indian friends from Antwerp: Amit, Gaurav Zinta, Tanmay, Kumud, Shweta, Sandeep, Khusboo, Chandra, Ajay Palagani, Xavier Britto who never let me feel alone and always helped and guided me during my initial days. Special thanks to current Indian friends: Gaurav, Garima, Prerna, Parul, Jyotsana, Rohit, Shyam, Ravi Soni, Nitin and Madhuri. Hanging out with you guys felt like home coming. Gaurav & Garima thank you for all the useful scientific discussions and providing helping hands in various activities. Special thanks to Prerna for providing the cherished friendship of lifetime. Your help and encouragement was indeed motivating especially during thesis cover design phase. Courteous thanks to Parul for being a great little friend, making delicious food, gathering and most importantly helping with writing equations using LaTeX for my research paper. Soon you will also be graduating and wish you all the best for it. Thanks to Shyam, Ravi, Rohit, Sreejith, Nadeesh for having wonderful time spent here at Antwerp. Love you guys.

Coming from a very small town full of coalmines (#KedlaNagar) and having an ambition to get doctorate was an almost an impossible dream. Hence this thesis is also a tribute to all my school teachers, childhood friends at Kedla. Especially I would like to thank my teacher Mr. Navin Sinha who taught us Physics so brilliantly and generated the spark to pursue research. I would also express my gratitude to my mathematics teacher Mr. Pratyush Kumar Singh at Ranchi for making us fall in love with abstract algebra, probabilities etc. This immensely helped me in grabbing several of these concepts in advanced mathematics/ statistics during my research.

Most importantly I tribute my doctoral thesis to my wonderful parents, elder brother and sister who supported me emotionally, financially throughout my stay outside India. Its’ a dream come true for them as well. Ma and Daddy you have been an inspiration for me since childhood the way you taught the values of patience, honesty, perseverance and hard work always motivated me to push harder and pursue my dreams. I can understand being away from home for more than 15 years might have been tough time with you but I am so thankful for understanding and believing on my ambitions and dreams. Vijay bhaiya and Sulekha didi you guys are amazing siblings and I thank you for being part of my journey. This has been possible only because you guys set an example in my early childhood and shown me the path to follow upon. Hope my research work will prove a stepping stone for inspiring your kids. I would also like to thank all my paternal and maternal relatives, cousins for all the encouragements throughout my studies.

Curriculum Vitae

Ajay Anand Kumar

University of Antwerp
Centrum Mediche Genetica (UZA/UA)
Prins Boudwijnlaan 43/6, Edegem-2650
Antwerpen, Belgium
Phone: +32465182459,+919006482012
DOB: 17.07.1983

aakumar1707@gmail.com
Linkedin: www.linkedin.com/in/aakumar17/
Skype: infogistt17
https://bitbucket.org/aakumar17/
Twitter: www.twitter.com/ajay_1707
Nationality: INDIA

Education

2013-2018 **PhD Bioinformatics**
Faculty of Medicine, **University of Antwerp**, Belgium
Thesis: *Data driven approaches towards computational genome interpretation for identification of disease causing mutations*
Key competencies: NGS data analysis, Machine Learning, Statistical genetics, Human Genetics
Promoters: Prof.Dr. Bart Loeys, Dr. Geert Vandeweyer, Dr. Maaike Alaerts

2009-2012 **MSc. Bioinformatics**
Faculty of Mathematics & Natural Sciences, **University of Helsinki**, Finland
Thesis: *Correlation analysis for evaluation of functional annotation methods of protein sequences*
Principal subjects: Computational Systems Biology, Machine Learning, Computational Statistics, Biophysics
Supervisors: Dr. Petri Toronen, Prof. Liisa Holm
Thesis grade: Magna Cum Laude Approbatur (MCLA)

2003-2007 **B.Tech Bioinformatics**
SASTRA University
Thesis: *Molecular dynamics study of fluid permeation through Carbon Nanotubes*
Principal Subjects: Computational structural biology, Protein Modeling, Drug designing, Molecular Biology, Genetics.
GPA: 8.4/10

1999-2001 **Higher Secondary School (XII$^{\text{th}}$)**
TATA D.A.V Public School, Jharkhand, India
Percentage score: 81% (3rd Rank)

-1999 **Secondary School (Xth)**
Holy Cross School, Jharkhand, India
Percentage score: 82%

Research/Industry Experience

2013-2018 **Thesis:** *Data driven approaches towards computational genome interpretation for identification of disease causing mutations*
Role: PhD researcher
Supervisors: Prof.Dr. Bart Loeys, Dr. Geert Vandeweyer, Dr. Maaike Alaerts, Dr. Lut Van Laer
Research group: Cardiogenetics group, **University of Antwerp**, Belgium

2011-2013 **Project 1:** *AraGnViz: Bayesian approach towards predicting functionality of Arabidopsis Thaliana genes.*
Role: Research assistant
Supervisor: Dr. Jarkko Salojarvi
Project 2: *Characterization and annotation of genes of plant pathogen Taphrina Deformans.*
Role: Resarch assistant
Supervisors: Dr. Kirk Overmeyer, Dr. Jarkko Salojarvi
Research group: Plant stress group, **University of Helsinki**, Finland

2010-2011 **MSc Thesis**: *Correlation analysis for evaluation of functional annotation methods in proteins.*
Role: Research Assistant
Supervisor: Dr. Petri Toronen, Prof. Liisa Holm
Research group: Bioinformatics group, **University of Helsinki**, Finland

2007-2009 **Project:** *Automation of requisition, spend analysis, e-Forms in Supply Chain management.*
Role: Project Engineer/ Ariba Consultant
Employer: WIPRO Technologies, Bangalore, India

2006-2007 **Project:** *siRNA design for effective RNAi the structural way.*
Role: Research Assistant (Indian Academy of Sciences, summer research fellow)
Supervisor: Prof. Dr. Manju Bansal
Research group: Bioinformatics group, **Indian Institute of Sciences**, Bangalore, India

Computational skills

Programming Languages	Python/Perl, Java/J2EE, C/C++, R/MATLAB, Latex
Databases	Oracle 9i, MySQL
Linkage analysis	Merlin, Simwalk2
Cluster computing	Torque-PBS
NGS	BWA, GATK, IGVTools, Galaxy
Sequence Assembly tools	Trinity, Samtools, SoapDenovo
Molecular Dynamics tools	NAMD/VMD, INSIGHT-2, CHARMM22
Protein modeling tools	MODELLER, PyMOL
Operating Systems	UNIX/Linux, Windows PC

Internships/Fellowships

2010	Intern at Prof. Liisa Holms Bioinformatics Group, University of Helsinki
2006	Indian Academy of Sciences summer research fellowship pursued at Indian Institute of Sciences, Bangalore, India
2006	Selected for summer research fellowship at National Brain Research Center, Delhi, India

Summer school/Courses/Symposiums

2017	”Advanced course in Python for Data Processing”, VIB, Leuven, Belgium
2017	ESHG training course in Cardiogenetics, University of Antwerp, Antwerp, Belgium
2017	Probabilistic graphical model (PGM) - Inference, Stanford University, Coursera
2016	”Introduction to HPC, Tips and Tricks”, University of Antwerp, Belgium
2015	DENIS summer school on Deep Learning, Aalto University, Espoo, Finland
2015	”Generalized Linear Models” - STATUA, University of Antwerp, Belgium
2015	”From Big-Data to Bedside” - Symposium on translational bioinformatics in cancer research, Ghent, Belgium
2014	Basic Principles of Statistics, STATUA, University of Antwerpe, Belgium
2014	Data science summer school 2014, Saarland University, Saarbrucken, Germany
2013	NGS Course Medical Genomics, University Hospital, KU-Leuven, Belgium
2013	”Next Generation Sequencing” Workshop, Antwerp University Hospital, Belgium

Co-curricular activities

2017	External reviewer of Master thesis from Department of Mathematics and Computer science, University of Antwerp, Belgium.
2016-2017	Course instructor/organizer on 'Introductory course to Next-Generation Sequencing data analysis', University of Antwerp, Belgium.
2007-2009	Successfully handled the offshore team for all the deliverables at WIPRO Technologies, India
2007	Cluster coordinator for event ChromoXone at National level techfest DAKSH 07 (www.daksh.sastra.edu) at SASTRA University, India
2007	Designed gaming event EDIFICE for DAKSH07 where participants has to find the correct structure of proteins based on its structural features such as loop, beta-sheets, alpha-helices etc.

Languages

English (Proficient), Hindi (Mother tongue), Dutch-Level1

Publications

During PhD

1. **Ajay Anand Kumar**, Bart Loeys, Gerarda Van De Beek, Nils Peeters, Wim Wuyts, Lut Van Laer, Maaike Alaerts, Geert Vandeweyer. *varAmpliCNV: Analyzing Variance of Amplicons to detect CNVs in targeted NGS data*. [Manuscript in submission]

2. Russell Gould, Hamza Aziz, Courtney Woods, Elizabeth Sparks, Dr. Christoph-Preuss, Mr. Florian Wnnemann, Miss Djahita Bjeda, Cassandra Moats, Sarah McClymont, Rebecca Rose, Dr. Nara Sobreira, Dr. Hua Ling, Mrs. Gretchen Mac-Carrick, **Ajay Anand Kumar**, Ilse Luyckx, Elyssa Cannaerts, Aline Verstraeten, Hanna Bjrk, Ann-Cathrin Lehsau, Vinod Jaskula-Ranga, Henrik Lauridsen, Asad Shah, Dr. Christopher Bennett, Patrick Ellinor, Honghuang Lin, Eric Isselbacher, Christian Lino Cardenas, Jonathan Butche , G. Chad Hughes, Mark Lindsay, Luc Mertens, Anders Franco-Cereceda, Judith Verhagen, Marja Wessels, Salah Mohamed, Per Eriksson, Seema Mital, Lut van Laer, Bart Loeys, Gregor Andelfinger, Andrew Mc-Callion, Harry C Dietz. *ROBO4 Mutations Predispose Individuals to Bicuspid Aortic Valve and Thoracic Aortic Aneurysm.* **Nature Genetics**, 2018 [Under review]

3. **Ajay Anand Kumar**, Lut Van Laer, Maaike Alaerts, Amin Ardeshirdavani, Yves Moreau, Kris Laukens, Bart Loeys, Geert Vandeweyer. *pBRIT: gene prioritization by correlating functional and phenotypic annotations through integrative data fusion.* **Bioinformatics** 1 (2018): 9.
4. Gillis, Elisabeth*, **Ajay A. Kumar***, Ilse Luyckx, Christoph Preuss, Elyssa Cannaerts, Gerarda van de Beek, Bjrn Wieschendorf1, Maaike Alaerts, Nikhita Bolar, Geert Vandeweyer, Josephina Meester, Florian Wnnemann, Russell A. Gould, Rustam Zhurayev, Dmytro Zerbino, Salah A. Mohamed, Seema Mital, Luc Mertens, Hanna M. Bjrck, Anders Franco-Cereceda, Andrew S. McCallion, Lut Van Laer, Judith M. A. Verhagen, Ingrid M. B. H. van de Laar, Marja W. Wessels, Emmanuel Messas, Guillaume Goudot, Michaela Nemcikova11, Alice Krebsova, Marlies Kempers, Simone Salemink, Toon Duijnhouwer, Xavier Jeunemaitre, Juliette Albuisson, Per Eriksson, Gregor Andelfinger, Harry C. Dietz, Aline Verstraeten, Bart L. Loeys and Mibava Leducq Consortium. *"Candidate gene resequencing in a large bicuspid aortic valve-associated thoracic aortic aneurysm cohort: SMAD6 as an important contributor."* **Frontiers in physiology** 8 (2017): 400.
5. van der Werf, Ilse M., Anke Van Dijck, Edwin Reyniers, Cline Helsmoortel, **Ajay Anand Kumar**, Vera M. Kalscheuer, Arjan PM de Brouwer, Tjitske Kleefstra, Hans van Bokhoven, Geert Mortier, Sandra Janssens, Geert Vandeweyer, R. Frank Kooy. *"Mutations in two large pedigrees highlight the role of ZNF711 in X-linked intellectual disability."* **Gene** 605 (2017): 92-98.

During Master's

6. **Kumar, Ajay Anand**, Liisa Holm, and Petri Toronen. *"GOParGenPy: a high throughput method to generate Gene Ontology data matrices."* **BMC bioinformatics** 14, no. 1 (2013): 242
7. Ciss, Ousmane H., Joo MGCF Almeida, lvaro Fonseca, **Ajay Anand Kumar**, Jarkko Salojrvi, Kirk Overmyer, Philippe M. Hauser, and Marco Pagni. *"Genome sequencing of the plant pathogen Taphrina deformans, the causal agent of peach leaf curl."* **MBio** 4, no. 3 (2013): e00055-13.

Bioinformatics acknowledgement:

8. Mukherjee, Shinjini, Timo Sipil, Pertti Pulkkinen, Kim Yrjl. *"Secondary successional trajectories of structural and catabolic bacterial communities in oilpolluted soil planted with hybrid poplar."* **Molecular ecology** 24.3 (2015): 628-642.

*- Equally contributed to the work.

Presentations

Oral presentation

1. Computational tools for gene prioritization and mutation burden analysis update 8th MIBAVA-Leducq conference, Baltimore,USA, (2016)
2. pBRIT: gene prioritization by correlating functional and phenotypic annotation through an an integrative data fusion. Invited talk at Biomina lunch talk series, University of Antwerp (2016)
3. Computational tools for gene prioritization and mutation burden analysis update 7th MIBAVA-Leducq conference, Lubeck, Germany (2015)
4. Information-Theoretic model for gene prioritization. Biomina research day, University of Antwerp, Belgium (2015)
5. Information-Theoretic model for gene prioritization. 10Th BeNeLux Bioinformatics conference, Student symposium, Antwerp, Belgium (2015)
6. Computational tools for gene prioritization and mutation burden analysis update. 5th MIBAVA-Leducq conference, Toronto, Canada (2015)
7. Computational tools for gene prioritization update. 4th MIBAVA-Leducq conference, Stockholm, Sweden (2014)
8. Designing potential inhibitor/ligand for PTP-B enzyme: Mycobacterium tuberculosis at National level technical symposium SHRISTI (2005)

Poster presentations

1. varAmpliCNV: Analyzing Variance of Amplicons to detect CNVs in targeted NGS data Applied Bioinformatics in Life Sciences, Leuven, Belgium (2018)
2. varAmpliCNV: Analyzing Variance of Amplicons to detect CNVs in targeted NGS data Benelux bioinformatics conference, Leuven, Belgium (2017)
3. pBRIT: gene prioritization by correlating functional and phenotypic annotation through an integrative data fusion approach. 17Th Annual Belgian Society of Human Genetics meeting, Louvain-la-Neuve, Aula Magna, Belgium (2017)
4. pBRIT: an advanced gene prioritization tool using Bayesian Regression and Information-Theoretic model approach. European Society of Human Genetics, Barcelona, Spain (2016)

5. Information-Theoretic model for gene prioritization. 10Th BeNeLux Bioinformatics conference, Antwerpen, Belgium (2015)
6. Bayesian regression approach towards gene prioritization. 23Rd annual Belgian-Dutch Conference on Machine Learning (BENELEARN), Brussels, Belgium (2014)
7. Sequencing of Taphrina Genome at 11th European Conference on Fungal Genetics, Germany 2012
8. Albumin as a cross linker in Hydrogels for sustained drug release at XVI National SBAOI Conference at I.I.T Delhi, April 2006.