Lessons learned from gene identification studies in Mendelian epilepsy disorders

Lessons learned from gene identification studies in Mendelian epilepsy disorders~~: a guide to future next generation sequencing projects~~

Hardies Katia[1,2], Weckhuysen Sarah[1,2], De Jonghe Peter[1,2,3] and Suls Arvid[1,2,*]

1 Neurogenetics Group, Department of Molecular Genetics, VIB, Antwerp, Belgium

2 Laboratory of Neurogenetics, Institute Born-Bunge, University of Antwerp, Antwerp, Belgium

3 Division of Neurology, Antwerp University Hospital, Antwerp, Belgium

* Corresponding author: Dr. Arvid Suls, Neurogenetics group, VIB-Department of Molecular Genetics, University of Antwerp – CDE (Building V), Antwerp, Belgium, email: arvid.suls@uantwerpen..be, Tel: +32 (0)3 265 10 22, Fax: +32 (0)3 265 11 12

Running title	Gene identification studies in Mendelian disorders
Key words	Next Generation Sequencing; Mendelian disorders; Epilepsy; Gene identification
Number of text pages	13
Number of words (summary)	241
Number of words (main text)	3141
Number of tables	1
Number of figures	2
References	68
Supplementary information/references	1

ABSTRACT

Next Generation Sequencing (NGS) technologies are now routinely used for gene identification in Mendelian disorders. Setting up cost-efficient NGS projects and managing the large amount of variants remains, however, a challenging job. Here, we provide insights in the decision-making processes before and after the use of NGS in gene identification studies.

Genetic factors are thought to play a role in about 70% of all epilepsies, and a variety of inheritance patterns have been described for seizure-associated gene defects. We therefore chose epilepsy as disease model and selected 35 NGS studies that focused on patients with a Mendelian epilepsy disorder. The strategies used for gene identification and their respective outcomes were reviewed.

High-throughput NGS strategies have led to the identification of several new epilepsy-causing genes, enlarging our knowledge on both known and novel pathomechanisms. NGS findings have furthermore extended the awareness of phenotypical and genetic heterogeneity. By discussing recent studies we illustrate: (I) the power of NGS for gene identification in Mendelian disorders, (II) the accelerating pace in which this field evolves, and (III) the considerations that have to be made when performing NGS studies.

Nonetheless the enormous rise in gene discovery over the last decade, many patients and families included in gene identification studies still remain without a molecular diagnosis; hence further genetic research is warranted. Based on successful NGS studies in epilepsy, we discuss general approaches to guide human geneticists and clinicians in setting up cost-efficient gene identification NGS studies.

INTRODUCTION

Unravelling the genetic cause of a patient's suspected inherited disorder is one of the major practices in human genetic research. Not only will the genetic background and cellular pathways involved in a specific disorder contribute to our understanding of human physiology, it will also influence the accuracy of prognosis, guidance in treatment decisions, and genetic counseling.

Because Mendelian or monogenic inherited disorders are caused by a single gene defect, they form a unique model for assessing direct cause-effect relationships. Traditional genetic approaches such as linkage analysis and candidate gene screenings have implicated causal genes in the etiology of many Mendelian inherited disorders (http://omim.org/). Since 2005, the availability of next generation sequencing (NGS) technologies has revolutionized the field of human genetics. They have enabled high-throughput gene identification studies and thereby introduced novel strategies to unravel the genetic etiology of Mendelian inherited disorders[1,2,3,4].

Since targeted resequencing of selected genes, whole exome and genome sequencing (WES/WGS) have become affordable, the number of individuals that are being studied has rapidly enlarged. Henceforth more and more researchers have been generating, analyzing, and interpreting NGS data. Setting up NGS projects and managing the large amount of variants remains, however, a challenging job. By referring to recent studies that used NGS technologies to identify causal gene defects in individuals with a presumed Mendelian

epilepsy disorder, we provide insights in how the field is evolving. By discussing several of the pitfalls, we also aim to illustrate how more successful and efficient NGS projects can be conducted.

METHOD

In 70% of epilepsy patients genetic factors are thought to underlie the seizure phenotype[5]. Mendelian epilepsy syndromes are rare, but causal gene defects following all monogenic inheritance patterns have been described in a subset of patients. Moreover, complicating factors such as a reduced penetrance, phenocopies and genetic heterogeneity are common (e.g. [6] and [7]). Epilepsy is thus an ideal disease model to discuss the implications of NGS on gene identification studies in Mendelian disorders. With a Medline search using the key words *epilepsy AND* respectively *whole genome*, *exome*, or *panel sequencing* about 100 reports were obtained. Thirty-five were selected based on the NGS methodology used to identify gene defects associated with the patient(s) epilepsy phenotype (i.e. study design). The overall study design for gene identification in Mendelian disorders is commonly determined by the characteristics of the investigated phenotype (such as severity, onset age, and incidence) in combination with the structure of the pedigree (including consanguinity, gender and number of affected individuals per generation). Together this information is suggestive for the most likely mode of inheritance, which in turn points towards a starting NGS strategy. The most frequently used strategies are further referred to as the autosomal recessive (AR) homozygous, AR compound heterozygous, autosomal dominant (AD) heterozygous, AD *de novo*, and candidate approach strategy[1]. Despite the importance of a good starting strategy, NGS permits to investigate different hypotheses in a single experiment.

NGS findings in Mendelian epilepsy disorders

All reported studies have been successful in associating a single gene defect with a Mendelian epilepsy disorder. Table 1 gives an overview of the selected 35 studies, their methodology, starting strategy and findings. A study was considered "successful in gene identification" in the following situations: (1) a variant was found in a gene previously associated with the same phenotype as that under investigation (i.e. known gene – known phenotype); (2) a variant was found in a gene previously associated with another neuronal phenotype as that under investigation (i.e. known gene – novel phenotype); (3) a variant was found in a gene not previously associated with any phenotype, but conclusive genetic evidence and/or substantial functional evidence was provided to support a causal relation with the phenotype (i.e. novel gene – novel phenotype). Variants found in a single individual without any further evidence of causality are considered 'candidate genes'.

Family based studies in monogenic epilepsy disorders searching for either AR or AD variants seem to lead to a molecular diagnosis for nearly 100% of the included patients. Naturally this is just apparent, as negative reports are rarely published for family studies. Of note, when NGS identifies a known gene – known phenotype variant missed by traditional genetic approaches, the family will also not be reported. This publication bias makes it difficult to estimate the true success rate in family based studies; so what can we expect? Based on our experience, we are able to identify a (likely) pathogenic variant in about 25% of AD families and up to 40% of AR families (unpublished data). Large scale studies on isolated patients or those who only include the index case in the initial analysis often do

report positive and negative results simultaneously. This leads to a more accurate estimation of the success rate, which ranges from 10% to 48%[8,9,10,11]. Most of the early reports are, however, enriched for patient samples prescreened for established epilepsy genes, lowering the calculated success rate. Similar to our own experiences, the more recent NGS studies – including patients without preceding molecular investigations - provide a molecular diagnosis for about 30% of the investigated epilepsy patients[12,13].

Setting up NGS gene identification studies

The power of NGS in gene identification studies is clearly illustrated by the progress in epilepsy genetics. NGS can partially overcome the large clinical heterogeneity, genetic variability, and other complicating factors often seen in Mendelian disorders. The current results should nevertheless caution researchers and clinicians. They imply that up to 70% of the patients remain without a molecular diagnosis after inclusion in a NGS based gene identification study. High-throughput is not an equivalent for infallibility, therefore we will discuss some of the considerations that have to be made when conducting a gene identification study using NGS.

1) Sampling and phenotyping

When setting up any genetic study, sample collection is the first and most important step: DNA material of key family members will be a necessity for interpretation of candidate variants. Although this might seem obvious, it is often not evident. Cooperation of both affected and unaffected family members is warranted, along with high quality and quantity of the DNA samples. A huge number of families are estranged from each other, either by distance or disputes. Additionally, needle phobia is quite common but can be overcome by extracting DNA from saliva; a technique that also allows sampling of very young children. Manual extraction and purification of a sufficient amount of saliva can render the required DNA material necessary for NGS.

Collaborating with motivated clinicians is therefore key for starting medical genetic research. Also, efficient gene identification goes hand in hand with deep phenotyping. A specific gene defect can be associated with different types of neurological phenotypes. Vice versa, different gene defects can be associated with the same phenotype. These obstacles can partially be overcome by NGS and might support the idea "to sequencing everyone". Specific clinical hallmarks can nevertheless, be indicative for a defined genetic syndrome and subsequent genetic defect. Selecting phenotypically matched cases can enlarge the chance of finding independent variants in the same gene (e.g.[14]).

Overall, the rise of NGS technologies has broadened the focus of study populations. Large multi-generation families were favored in the past, but a combined approach of linkage analysis with NGS of positional candidate genes remains the major approach in AD families. Problems such as reduced penetrance and variable expression within these families remain, however, problematic. Identifying highly penetrant variants in patients with a severe phenotype, either presenting as sporadic patients (i.e. *de novo* analyses) or as siblings (i.e. *de novo* analyses with germline mosaicism or recessive inheritance), is more straightforward. Hereby making isolated cases and sib pairs most popular in NGS studies.

2) NGS strategy

After collecting and thoroughly assessing the study cohort, the most likely inheritance pattern can be determined for each family. Figure 1 is provided as a guide to help deciding which strategy can best be applied to analyze specific sample sets. The choice of strategy will define which individuals should be sequenced: a trio-approach is required for a

straightforward *de novo* analysis, whereas studies on families with multiple affected individuals are less indicative towards the subjects to be sequenced. The earliest NGS studies showed that including sequencing data of an increasing number of family members gives considerably more power to any genomic analysis[15]. Depending on the pedigree structure some individuals will, however, be more informative than others. Several tools exist for automated selection, such as the statistical framework GIGI-Pick, although they mainly focus on large and complex pedigrees[16]. Studies including AD families mostly opt for sequencing two or three distantly related relatives and look to identify variants according to the kinship. Alternatively, some studies select unrelated but phenotypically or genetically (through linkage analysis) matched index cases, and search for a shared gene defect. Because gene identification in AR families relies on the presence of two hits in the same gene, sequencing one or two individuals is often sufficient in order to reach a molecular diagnosis.

Conclusive genetic proof of pathogenicity can only be established when the same gene defect is identified in independent patients. The yield of pathogenic variants in established epilepsy genes and for specific patient subgroups can, however, vary from <1% to up to 80%. Phenotypic and genetic heterogeneity further complicate an accurate estimation of the required amount of patients to sequence, in order to find variants in the same gene. Data sharing between genetic centers is indispensable in the future to achieve a higher yield of molecular diagnoses (e.g. [11] and [17]). Several initiatives have recently been set up to facilitate multi-center collaborations between genetic diagnostic and research center (e.g. [18], https://genematcher.org/ and http://www.matchmakerexchange.org/).

3) *Sequencing*

The 35 selected studies have collectively sequenced 1895 individuals belonging to 1244 different epilepsy families on a NGS platform. WES is by far the most used sequencing technique in research: 73% compared to 11% WGS and 16% targeted resequencing. Because the exome represents an enriched part of the genome for disease associated variants, this preference is justifiable in gene identification studies. Loss of information will nevertheless have to be taken into account, because not the complete exome will be optimally covered. Due to dropping sequencing costs and optimization of bioinformatics tools, WGS is most likely to take over in the future[19]. In the mean while, targeted resequencing can also provide a higher coverage of selected regions. For large families where linkage data is available, the combination with targeted resequencing can be very powerful. Remarkably, WES is also preferred when linkage data is available (Table 1: 5/8 studies). The time needed to optimize custom gene panels, in combination with having direct access to data on the entire exome if no (likely) pathogenic variant can be found in the captured region(s) probably underlies this choice. The increasing speed and decreasing cost have also led to the implementation of NGS in clinical genetic services. Targeted gene panels long had an ethical and cost-efficient advantage over WES/WGS by screening only the relevant disease genes[20,21,22]. Due to their limited power in variant yield for heterogeneous disorders (such as epilepsy), diagnostic labs are now moving towards the use of WES/WGS[23,24]. Validation of candidate variants and segregation analysis in additional family members is generally still done by direct Sanger sequencing.

Nowadays, NGS data is often generated in service facilities, therefore the actual sequencing procedure will not be further discussed here. Differences in variant callings due to capturing, sequencing technology, and mapping algorithms have been extensively reported[25,26,27]. To handle and analyze the generated data most labs have also developed custom pipelines over the years. If such pipeline is not available, we recommend the user-friendly and web-based tool Galaxy and Broad's Genome Analysis Toolkit (GATK) variant caller to compile, annotate, and analyze data.

4) NGS variant management

After collecting the right samples, determining whom to sequence and generating the data, most of the work still has to start. Interpretation of NGS data and selecting the causal gene defect has become more time-consuming than the data generation itself. A set of highly recommended guidelines for interpretation of sequence variants was recently published by the ACMG. They include comprehensive lists of (I) population, disease, and gene specific databases, (II) *in silico* prediction algorithms, and (III) criteria for classifying candidate variants[28]. Here, we further discuss some general principles of a standard filtering cascade used during variant prioritization (Figure 2).

A variant-based selection on the quality of the variant calling (e.g. coverage), frequency in population databases and predicted impact on the encoded protein is usually the first step. Setting quality cut offs pursues an optimal balance between sensitivity and specificity. Trying different settings and analyzing multiple samples side-by-side will prove useful to reduce false positive and false negative results. In addition, joint calling of different

samples versus single sample calling will generally provide better quality calls. The most widely used variant caller, GATK, has an active community that is constantly guiding users in how to optimize quality thresholds. We encourage new users to participate actively (http://gatkforums.broadinstitute.org/). Cut offs regarding the frequency and predicted impact of a variant will mainly depend on the assumed inheritance pattern and incidence of the investigated phenotype. For example, when looking for *de novo* variants in an individual with a devastating phenotype, candidate variants should be absent in control databases. On the contrary, when looking for AR or AD variants with a reduced penetrance a certain percentage of heterozygous carriers can be expected. Publicly available databases including the most recently released ExAC browser (http://exac.broadinstitute.org, Cambridge, MA), are constantly expanding. Yet, setting up an in-house variant database is equally important as it has the additional advantage to enable exclusion of NGS platform dependent variants. This can furthermore provide information on variant frequencies in distinct ethnical groups.

A second variant-based prioritization, here referred to as genetic-selection, is based on the assumptions made regarding inheritance patterns. Although assumptions regarding the mode of inheritance is the most empirical method to identify disease-causing variants, it can be misleading. Analysis of the data under different inheritance models is recommended, as has been nicely illustrate in several recent NGS studies (e.g.[3,13]).

Next, data mining on all genes harboring remaining candidate variants will enable researchers to prioritize genes of interest. Several programs can help to rapidly pinpoint the

genes most likely associated with the investigated phenotype (e.g. ToppGene and Biograph) or evaluate specific expression patterns (e.g. EvoTol, GTex and Human Protein Atlas). Of note, data-mining is quite subjective and often correlates with the experience of the researcher. Automated probability interpretations have the advantage of speed and throughput over manual labor, but share a large bias towards existing knowledge on gene function and disease pathomechanisms. Therefore it is recommended to level different programs and manually check the ranked candidate genes.

The final interpretation about the potential impact of a specific gene defect is partly variant- and partly gene-based. Ideally such interpretation is supported by *in vitro* or *in v*ivo experiments, but due to practical considerations *in silico* predictions can provide a first and fast indication. The recently released integrated framework Combined Annotation–Dependent Depletion (CADD) unites the most widely used interpretation tools. Again a bias towards current knowledge and the limitation of most *in silico* tools towards variants with a potentially large effect (e.g. protein altering variants) has to be taken into account. Several scoring systems also exists for ranking genes according to their tolerance for genetic variations or likelihood to be involved in disease pathways. These work quite well for *de novo* variants, but are less indicative when examining recessive gene defects. The existence of multiple transcripts and variable expression patterns can furthermore cause contradictions between predictions algorithms.

It is clear that NGS variant management is not a rigid science for which a standard protocol can be followed. General methodologies for prioritizing an interpreting variants are

nevertheless proposed. Depending on the nature of the gene identification study, different cut offs will proof most valuable. Getting acquainted with frequently used annotation sets and prediction tools is necessary when learning how to interpret NGS data. Basic genetic ideas remain valid in NGS studies and researchers should always be critical about early assumptions: there are always several options and NGS gives you the power to investigate them all trough a single experiment. Finally, an interpretation will have to be made for determining the nature of a variant based on combined genetic, clinical, and preferably functional data[29].

5) Technical pitfalls

Even when multiple individuals of the same family are sequenced, the data is analyzed under different inheritance models, and filtering strategies have been worn out, it remains possible that no variant is interpreted as (likely) pathogenic. When working with NGS technologies it is important to realize that no single technique will cover everything. Identification of somatic and germline mosaic events or copy number variants (CNVs) remain complicated despite the progress made in the field[30,31]. CNVs have an important contribution to many Mendelian disorders[32,33,34] and are often overlooked or disregarded in NGS studies. The pathogenic variant can also be unsequenced due to lack of targeting, capturing, or bad mapping quality. Variant callers and annotation tools might furthermore be wrong or misleading. A very useful tool to check the nomenclature of a variant, together with its impact, is the Mutalyzer 2.0.9[35]. Ideally, each individual should be sequenced and analyzed on multiple platforms to reduce the chance of missing causal variants due to technical errors[27]. Unfortunately this is not a realistic situation; most researchers have a

budget that is preferably used to investigate as many families as possible. Analyzing NGS data with different variant callers and annotation tools is for the moment the best way to partially compensate for limited resources.

CONCLUSION

High-throughput screenings have led to the identification of many new disease causing genes for Mendelian diseases. The large phenotypical and genetic heterogeneity of human disorders will nevertheless make it necessary to collaborate on a large scale and pool data in a joined effort to support causality of specific gene defects (e.g.[11]). In the end, accumulating genetic information will expand our knowledge on normal human body functions and eventually lead to a better understanding of disease pathways.

NGS is a continuous evolving field that is becoming increasingly available, faster and cheaper. Henceforth, more and more researchers will be setting up NGS projects to give a molecular diagnosis to their patient(s). By summarizing recent findings in patients with Mendelian epilepsy disorders and discussing the used NGS research designs, we highlight several considerations that have to be taken into account before and during a gene identification study. It is clear that NGS has provided many insights and opportunities in genetic research, but with still more than half of the patients without a molecular diagnosis, further genetic research is a priority.

ETHICAL GUIDELINES

We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

To gain experience in NGS data analysis and gene identification in Mendelian epilepsies the involved researchers had the chance to be part of a fruitful, large scale consortium. Therefore we thank the Eurocores program EuroEPINOMICS of the European Science Foundation (G.A.136.11.N and FWO/ESF-ECRP), all partners and in particular Rudi Balling and Patrick May of the Family Genomics group at the Institute for Systems Biology (Luxembourg). We also acknowledge the constant support of the Fund for Scientific Research Flanders (FWO), the University of Antwerp and the centralized services of the VIB-Department Molecular genetics with special mentioning of Peter De Rijk for developing our in-house NGS pipeline (GenomComb). K.H. is a PhD fellow of the Institute for Science and Technology (IWT)-Flanders and A.S. is a postdoctoral fellow of the FWO.

Reference List

1. Gilissen C, Hoischen A, Brunner HG *et al*: Disease gene identification strategies for exome sequencing. *Eur.J.Hum.Genet.* 2012; **20**: 490-497.

2. Rabbani B, Mahdieh N, Hosomichi K *et al*: Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *J.Hum.Genet.* 2012; **57**: 621-632.

3. Yang Y, Muzny DM, Xia F *et al*: Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* 2014; **312**: 1870-1879.

4. Koboldt DC, Steinberg KM, Larson DE *et al*: The next-generation sequencing revolution and its impact on genomics. *Cell* 2013; **155**: 27-38.

5. Hildebrand MS, Dahl HH, Damiano JA *et al*: Recent advances in the molecular genetics of epilepsy. *J.Med.Genet.* 2013; **50**: 271-279.

6. Escayg A, MacDonald BT, Meisler MH *et al*: Mutations of SCN1A, encoding a neuronal sodium channel, in two families with GEFS+2. *Nat.Genet.* 2000; **24**: 343-345.

7. Claes S, Devriendt K, Lagae L *et al*: The X-linked infantile spasms syndrome (MIM 308350) maps to Xp11.4-Xpter in two pedigrees. *Ann.Neurol.* 1997; **42**: 360-364.

8. Allen AS, Berkovic SF, Cossette P *et al*: De novo mutations in epileptic encephalopathies. *Nature* 2013; **501**: 217-221.

9. Michaud JL, Lachance M, Hamdan FF *et al*: The genetic landscape of infantile spasms. *Hum.Mol.Genet.* 2014; **23**: 4846-4858.

10. Lemke JR, Riesch E, Scheurenbrand T *et al*: Targeted next generation sequencing as a diagnostic tool in epileptic disorders. *Epilepsia* 2012; **53**: 1387-1398.

11. EuroEPINOMICS-RES Consortium, Epilepsy Phenome/Genome Project, Epi4K Consortium.: De Novo Mutations in Synaptic Transmission Genes Including DNM1 Cause Epileptic Encephalopathies. *Am.J.Hum.Genet.* 2014; **95**, 360-370.

12. Mercimek-Mahmutoglu S, Patel J, Cordeiro D *et al*: Diagnostic yield of genetic testing in epileptic encephalopathy in childhood. *Epilepsia* 2015; **56**: 707-716.

13. Muona M, Berkovic SF, Dibbens LM *et al*: A recurrent de novo mutation in KCNC1 causes progressive myoclonus epilepsy. *Nat.Genet.* 2015; **47**: 39-46.

14. Suls A, Jaehn JA, Kecskes A *et al*: De novo loss-of-function mutations in CHD2 cause a fever-sensitive myoclonic epileptic encephalopathy sharing features with Dravet syndrome. *Am.J.Hum.Genet.* 2013; **93**: 967-975.

15. Roach JC, Glusman G, Smit AF *et al*: Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 2010; **328**: 636-639.

16. Cheung CY, Marchani BE, Wijsman EM: A statistical framework to guide sequencing choices in pedigrees. *Am.J.Hum.Genet.* 2014; **94**: 257-267.

17. Zhang X, Ling J, Barcia G *et al*: Mutations in QARS, encoding glutaminyl-tRNA synthetase, cause progressive microcephaly, cerebral-cerebellar atrophy, and intractable seizures. *Am.J.Hum.Genet.* 2014; **94**: 547-558.

18. Fokkema IF, Taschner PE, Schaafsma GC *et al*: LOVD v.2.0: the next generation in gene variant databases. *Hum.Mutat.* 2011; **32**: 557-563.

19. Gilissen C, Hehir-Kwa JY, Thung DT *et al*: Genome sequencing identifies major causes of severe intellectual disability. *Nature* 2014; **511**: 344-347.

20. Sun Y, Ruivenkamp CA, Hoffer MJ *et al*: Next-generation diagnostics: gene panel, exome, or whole genome? *Hum.Mutat.* 2015; **36**: 648-655.

21. Vrijenhoek T, Kraaijeveld K, Elferink M *et al*: Next-generation sequencing-based genome diagnostics across clinical genetics centers: implementation choices and their effects. *Eur.J.Hum.Genet.* 2015; **23**: 1142-1150.

22. Lohmann K, Klein C: Next generation sequencing and the future of genetic diagnosis. *Neurotherapeutics.* 2014; **11**: 699-707.

23. Della ME, Ciccone R, Brustia F *et al*: Improving molecular diagnosis in epilepsy by a dedicated high-throughput sequencing platform. *Eur.J.Hum.Genet.* 2014; **23**: 354-362.

24. Wang J, Gotway G, Pascual JM *et al*: Diagnostic yield of clinical next-generation sequencing panels for epilepsy. *JAMA Neurol.* 2014; **71**: 650-651.

25. Metzker ML: Sequencing technologies - the next generation. *Nat.Rev.Genet.* 2010; **11**: 31-46.

26. Clark MJ, Chen R, Lam HY *et al*: Performance comparison of exome DNA sequencing technologies. *Nat.Biotechnol.* 2011; **29**: 908-914.

27. O'Rawe J, Jiang T, Sun G *et al*: Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med.* 2013; **5**: 28.

28. Richards S, Aziz N, Bale S *et al*: Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet.Med.* 2015; **17**: 405-423.

29. Sunyaev SR: Inferring causality and functional significance of human coding DNA variants. *Hum.Mol.Genet.* 2012; **21**: R10-R17.

30. Rios JJ, Delgado MR: Using whole-exome sequencing to identify variants inherited from mosaic parents. *Eur.J.Hum.Genet.* 2014; **23**: 547-550.

31. Tan R, Wang Y, Kleinstein SE *et al*: An evaluation of copy number variation detection tools from whole-exome sequencing data. *Hum.Mutat.* 2014; **35**: 899-907.

32. Mefford HC, Yendle SC, Hsu C *et al*: Rare copy number variants are an important cause of epileptic encephalopathies. *Ann.Neurol.* 2011; **70**: 974-985.

33. Steffens M, Leu C, Ruppert AK *et al*: Genome-wide association analysis of genetic generalized epilepsies implicates susceptibility loci at 1q43, 2p16.1, 2q22.3 and 17q21.32. *Hum.Mol.Genet.* 2012; **21**: 5359-5372.

34. de Kovel CG, Trucks H, Helbig I *et al*: Recurrent microdeletions at 15q11.2 and 16p13.11 predispose to idiopathic generalized epilepsies. *Brain* 2010; **133**: 23-32.

35. Wildeman M, van OE, den Dunnen JT *et al*: Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum.Mutat.* 2008; **29**: 6-13.

36. Vaher U, Noukas M, Nikopensius T *et al*: De Novo SCN8A Mutation Identified by Whole-Exome Sequencing in a Boy With Neonatal Epileptic Encephalopathy, Multiple Congenital Anomalies, and Movement Disorders. *J.Child Neurol.* 2013; **29**: NP202-NP206.

37. Veeramah KR, Johnstone L, Karafet TM *et al*: Exome sequencing reveals new causal mutations in children with epileptic encephalopathies. *Epilepsia* 2013; **54**: 1270-1281.

38. Hackenberg A, Baumer A, Sticht H *et al*: Infantile Epileptic Encephalopathy, Transient Choreoathetotic Movements, and Hypersomnia due to a De Novo Missense Mutation in the SCN2A Gene. *Neuropediatrics* 2014; **45**: 261-264.

39. Nakajima J, Okamoto N, Tohyama J *et al*: De novo EEF1A2 mutations in patients with characteristic facial features, intellectual disability, autistic behaviors and epilepsy. *Clin.Genet.* 2014; **87**: 356-361.

40. Vanderver A, Simons C, Schmidt JL *et al*: Identification of a Novel de Novo p.Phe932Ile KCNT1 Mutation in a Patient With Leukoencephalopathy and Severe Epilepsy. *Pediatr.Neurol.* 2014; **50**: 112-114.

41. Baasch AL, Huning I, Gilissen C *et al*: Exome sequencing identifies a de novo SCN2A mutation in a patient with intractable seizures, severe intellectual disability, optic atrophy, muscular hypotonia, and brain abnormalities. *Epilepsia* 2014; **55**: e25-e29.

42. Lee H, Lin MC, Kornblum HI *et al*: Exome sequencing identifies de novo gain of function missense mutation in KCND2 in identical twins with autism and seizures that slows potassium channel inactivation. *Hum.Mol.Genet.* 2014; **23**: 3481-3489.

43. Saitsu H, Kato M, Osaka H *et al*: CASK aberrations in male patients with Ohtahara syndrome and cerebellar hypoplasia. *Epilepsia* 2012; **53**: 1441-1449.

44. Nakamura K, Kodera H, Akita T *et al*: De Novo mutations in GNAO1, encoding a Galphao subunit of heterotrimeric G proteins, cause epileptic encephalopathy. *Am.J.Hum.Genet.* 2013; **93**: 496-505.

45. Barcia G, Fleming MR, Deligniere A *et al*: De novo gain-of-function KCNT1 channel mutations cause malignant migrating partial seizures of infancy. *Nat.Genet.* 2012; **44**: 1255-1259.

46. Milh M, Falace A, Villeneuve N *et al*: Novel compound heterozygous mutations in TBC1D24 cause familial malignant migrating partial seizures of infancy. *Hum.Mutat.* 2013; **34**: 869-872.

47. Nakamura K, Osaka H, Murakami Y *et al*: PIGO mutations in intractable epilepsy and severe developmental delay with mild elevation of alkaline phosphatase levels. *Epilepsia* 2014; **55**: e13-e17.

48. Paciorkowski AR, Weisenberg J, Kelley JB *et al*: Autosomal recessive mutations in nuclear transport factor KPNA7 are associated with infantile spasms and cerebellar malformation. *Eur.J.Hum.Genet.* 2013; **22**: 587-593.

49. Hitomi Y, Heinzen EL, Donatello S *et al*: Mutations in TNK2 in severe autosomal recessive infantile onset epilepsy. *Ann.Neurol.* 2013; **74**: 496-501.

50. Campeau PM, Kasperaviciute D, Lu JT *et al*: The genetic basis of DOORS syndrome: an exome-sequencing study. *Lancet Neurol.* 2014; **13**: 44-58.

51. Basel-Vanagaite L, Hershkovitz T, Heyman E *et al*: Biallelic SZT2 mutations cause infantile encephalopathy with epilepsy and dysmorphic corpus callosum. *Am.J.Hum.Genet.* 2013; **93**: 524-529.

52. Poduri A, Heinzen EL, Chitsazzadeh V *et al*: SLC25A22 is a novel gene for migrating partial seizures in infancy. *Ann.Neurol.* 2013; **74**: 873-882.

53. Pippucci T, Parmeggiani A, Palombo F *et al*: A Novel Null Homozygous Mutation ConfirmsCACNA2D2 as a Gene Mutated in Epileptic Encephalopathy. *PLoS.One.* 2013; **8**: e82154-

54. Heron SE, Smith KR, Bahlo M *et al*: Missense mutations in the sodium-gated potassium channel gene KCNT1 cause severe autosomal dominant nocturnal frontal lobe epilepsy. *Nat.Genet.* 2012; **44**: 1188-1190.

55. Ishida S, Picard F, Rudolf G *et al*: Mutations of DEPDC5 cause autosomal dominant focal epilepsies. *Nat.Genet.* 2013; **45**: 552-555.

56. Dibbens LM, de VB, Donatello S *et al*: Mutations in DEPDC5 cause familial focal epilepsy with variable foci. *Nat.Genet.* 2013; **45**: 546-551.

57. Edvardson S, Oz S, Abulhijaa FA *et al*: Early infantile epileptic encephalopathy associated with a high voltage gated calcium channelopathy. *J.Med.Genet.* 2013; **50**: 118-123.

58. Alazami AM, Hijazi H, Kentab AY *et al*: NECAP1 loss of function leads to a severe infantile epileptic encephalopathy. *J.Med.Genet.* 2014; **51**: 224-228.

59. Corbett MA, Bahlo M, Jolly L *et al*: A focal epilepsy and intellectual disability syndrome is due to a mutation in TBC1D24. *Am.J.Hum.Genet.* 2010; **87**: 371-375.

60. De FM, Vago R, Striano P *et al*: The alpha2B-adrenergic receptor is mutant in cortical myoclonus and epilepsy. *Ann.Neurol.* 2014; **75**: 77-87.

61. Heron SE, Grinton BE, Kivity S *et al*: PRRT2 mutations cause benign familial infantile epilepsy and infantile convulsions with choreoathetosis syndrome. *Am.J.Hum.Genet.* 2012; **90**: 152-160.

62. Carvill GL, Heavin SB, Yendle SC *et al*: Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1. *Nat.Genet.* 2013; **45**: 825-830.

63. Touma M, Joshi M, Connolly MC *et al*: Whole genome sequencing identifies SCN2A mutation in monozygotic twins with Ohtahara syndrome and unique neuropathologic findings. *Epilepsia* 2013; **54**: e81-e85.

64. Lemke JR, Hendrickx R, Geider K *et al*: GRIN2B mutations in west syndrome and intellectual disability with focal epilepsy. *Ann.Neurol.* 2013; **75**: 147-154.

65. Lemke JR, Lal D, Reinthaler EM *et al*: Mutations in GRIN2A cause idiopathic focal epilepsy with rolandic spikes. *Nat.Genet.* 2013; **45**: 1067-1072.

66. Veeramah KR, O'Brien JE, Meisler MH *et al*: De novo pathogenic SCN8A mutation identified by whole-genome sequencing of a family quartet affected by infantile epileptic encephalopathy and SUDEP. *Am.J.Hum.Genet.* 2012; **90**: 502-510.

67. Lee HY, Huang Y, Bruneau N *et al*: Mutations in the novel protein PRRT2 cause paroxysmal kinesigenic dyskinesia with infantile convulsions. *Cell Rep.* 2012; **1**: 2-12.

68. Martin HC, Kim GE, Pagnamenta AT *et al*: Clinical whole-genome sequencing in severe early-onset epilepsy reveals new genes and improves molecular diagnosis. *Hum.Mol.Genet.* 2014; **23**: 3200-3211.

TITLES AND LEGENDS TO FIGURES

Figure 1: Visual aid to the guide NGS projects

Orange boxes are indicators for the project strategy. Blue boxes indicate to choose a sequencing technique; depending on this choice different results will be obtained. Green boxes contain the different project strategies. Red lines indicate the minimum sample requirement, sequencing more individuals of the same family will render more power, but is not necessary to find a molecular diagnosis in the majority of patients.

Figure 2: General filtering cascade

NGS data sets generate a large amount of variants. A standard prioritization scheme consists of four different layers, which are further separated into variant-based and gene-based.

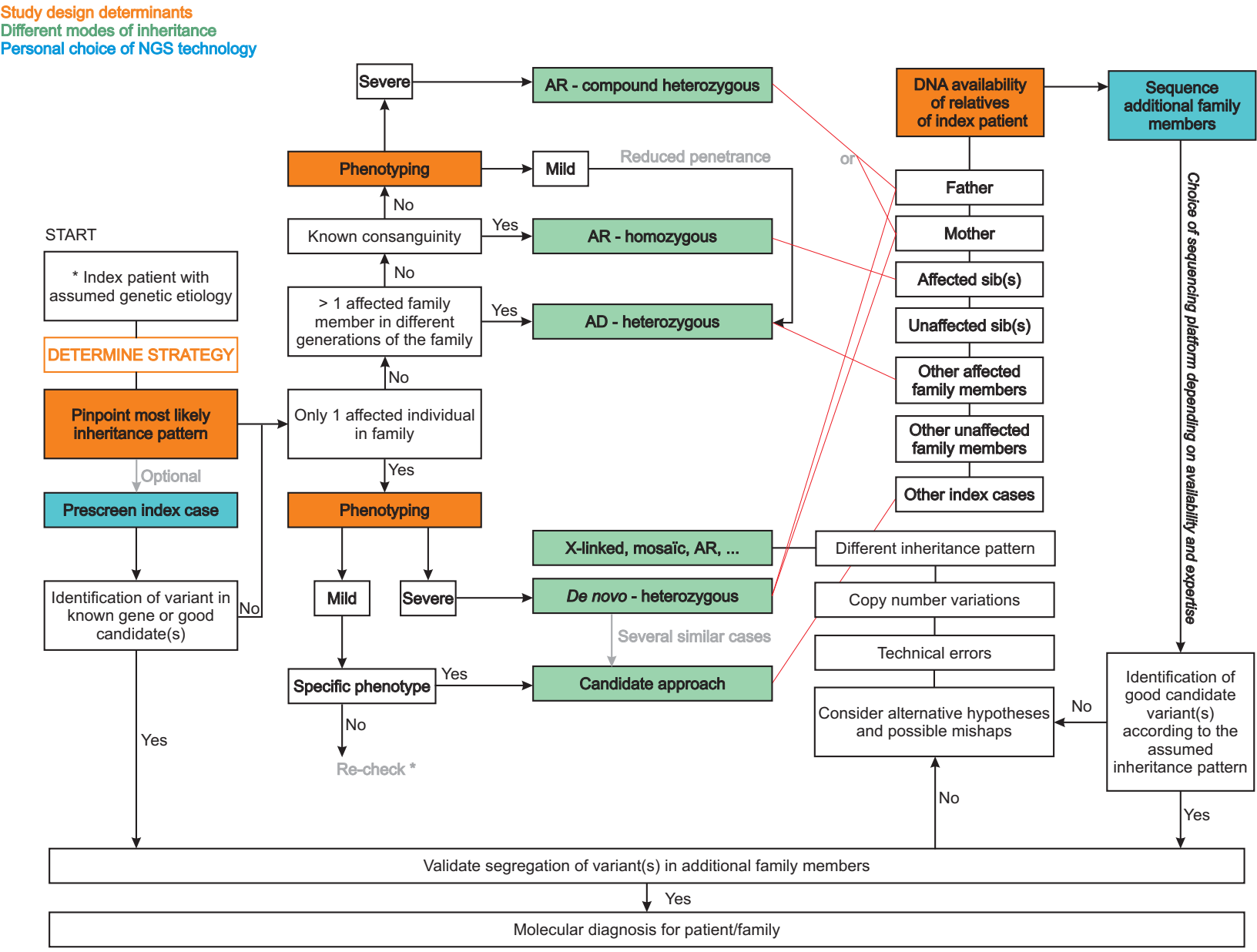
1) A variant selection is based on quality, frequency in a control population and predicted impact on the protein.
2) Based on the assumed mode of inheritance the genetics selection can reduce the number of candidate variants.
3) Extensive literature study on the genes harboring the remaining variants can lead to a reduction of candidate genes.
4) Different tools can be used to help interpret the possible impact of the gene defects and assess the most likely associations with the phenotype of interest

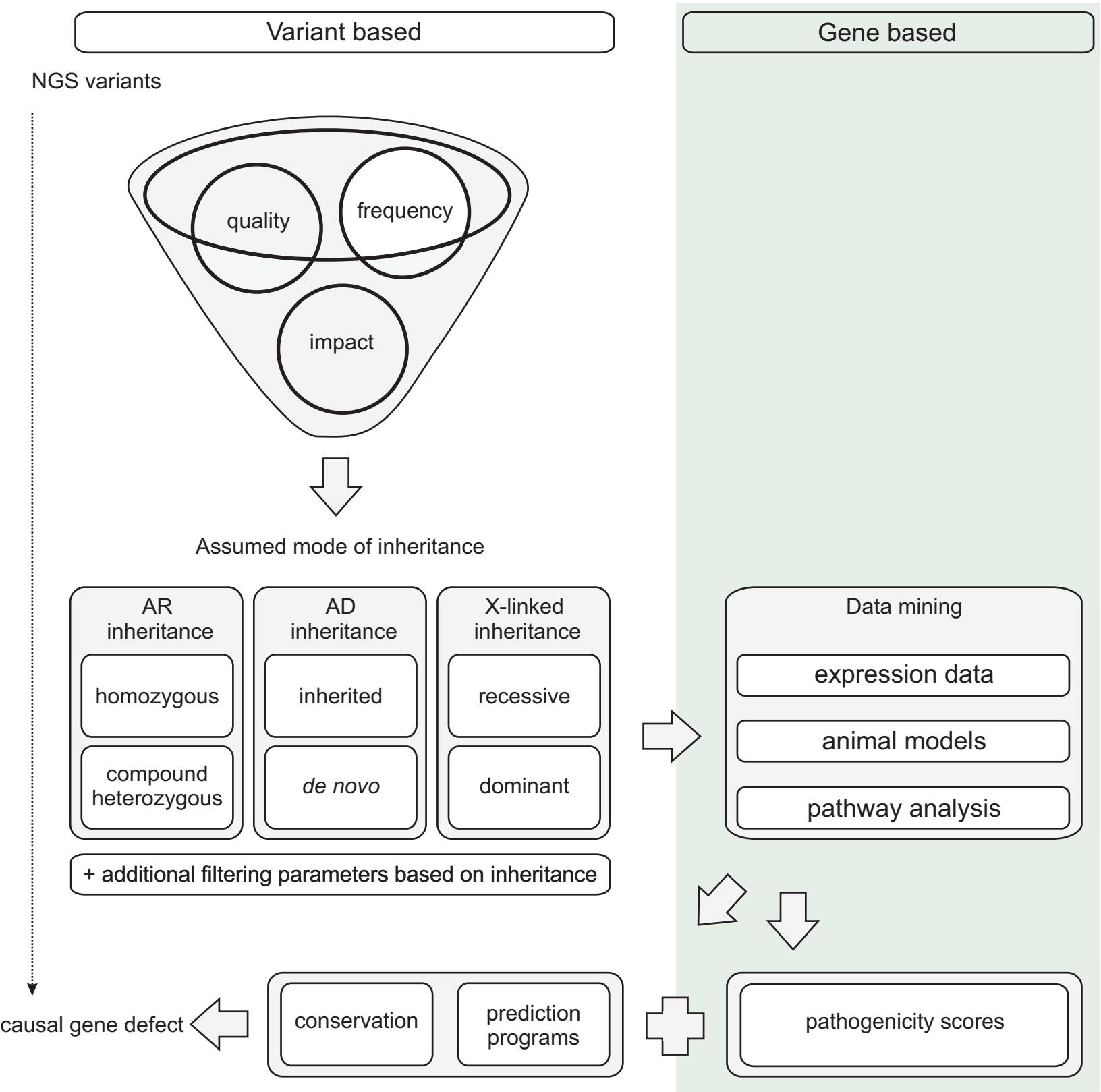
Table 1: Overview of 35 studies using NGS technologies to identify causal gene defects in presumed Mendelian epilepsy disorders, ordered by the NGS methodology and including their sample size, investigated phenotype, starting strategy and findings.

NGS method	Study population	Phenotype	Starting strategy	Molecular diagnosis	Gene(s)	Ref.
WES	trio	EE	*de novo*	known gene - known phenotype	*SCN8A*	36
WES	264 trios	IS, LGS	*de novo*	known gene - known phenotype	*SCN1A, STXBP1, CDKL5, SCN8A, SCN2A*	8
				known gene - novel phenotype	*CACNA1A, CHD2, FLNA, GABRA1, GRIN2B*	
				novel gene - novel phenotype	*ALG13, GABRB3, DNM1, HDAC4*	
				candidate gene	*IQSEC2, MTOR, NEDD4L, HNRNPU*	
WES	10 trios	EE	*de novo*	known gene - known phenotype	*SCN1A, CDKL5, EEF1A2*	37
				candidate gene	*KCNH5, CLCN4, ARHGEF15*	
WES	index case	EE	*de novo*	known gene - known phenotype	*SCN2A*	38
WES	2 trios	EE	*de novo*	known gene - known phenotype	*EEF1A2*	39
WES	trio	EE	*de novo*	known gene - novel phenotype	*KCNT1*	40
WES	9 trios	DS	*de novo*	known gene - novel phenotype	*CHD2*	14
WES	trio	EE	*de novo*	known gene - novel phenotype	*SCN2A*	41
WES	quartet	Epilepsy + autism	*de novo*	known gene - novel phenotype	*KCND2*	42
WES	7 index cases + 5 trios	EE, OS	*de novo*	known gene - novel phenotype	*CASK*	43,44
				novel gene - novel phenotype	*GNAO1*	
WES	3 trios	MMPSI	*de novo*	novel gene - novel phenotype	*KCNT1*	45
WES	index case	MMPSI	AR - compound heterozygous	known gene - novel phenotype	*TBC1D24*	46
WES	sibs	EE	AR - compound heterozygous	known gene - novel phenotype	*PIGO*	47
WES	sibs	LGS	AR - compound heterozygous	novel gene - novel phenotype	*KPNA7*	48
WES	sibs	EE	AR - compound heterozygous	novel gene - novel phenotype	*TNK2*	49
WES	17 index cases	DOORS	AR - compound heterozygous & AR - homozygous	known gene - novel phenotype	*TBC1D24*	50
WES	2 index cases	EE	AR - compound heterozygous & AR - homozygous	novel gene - novel phenotype	*SZT2*	51
WES	2 index cases	MMPSI	AR – homozygous	known gene - novel phenotype	*SLC25A22*	52
WES	index case	EE	AR – homozygous	novel gene - novel phenotype	*CACNA2D2*	53

WES	sib + quartet	EE	AR – homozygous	novel gene - novel phenotype	*QARS*	[17]
WES (+ linkage data)	3x 2 family members	ADNFLE	AD- heterozygous	known gene - novel phenotype	*KCNT1*	[54]
WES (+ linkage data)	2 index cases	FFEVF	AD- heterozygous	novel gene - novel phenotype	*DEPDC5*	[55]
WES (+ linkage data)	2 index cases	FFEVF	AD- heterozygous	novel gene - novel phenotype	*DEPDC5*	[56]
WES (+ linkage data)	trio	LGS	AR – homozygous	novel gene - novel phenotype	*CACNA2D2*	[57]
WES (+ linkage data)	index case	EE	AR – homozygous	novel gene - novel phenotype	*NECAP1*	[58]
Targeted resequencing linkage region	index case	Focal epilepsy + ID	AR - homozygous	known gene - novel phenotype	*TBC1D24*	[59]
Targeted resequencing linkage region	2 index cases	Cortical myoclonus and epilepsy	AD- heterozygous	novel gene - novel phenotype	*ADRA2B*	[60]
Targeted resequencing linkage region	2 index cases	ICCA	AD- heterozygous	novel gene - novel phenotype	*PRRT2*	[61]
Targeted resequencing gene panel	500 index cases	EE	Candidate approach	known gene - novel phenotype	*MEF2C, MBD5, GABRG2*	
				novel gene - novel phenotype	*SYNGAP1, CHD2*	[62]
				candidate gene	*HNRNPU*	
Targeted resequencing gene panel	357 index cases	epilepsy	Candidate approach	known gene - known phenotype	*SCN1A, SCN2A, STXBP1, KCNJ10, KCTD7, KCNQ3, ARHGEF9, SMS, TPP1, MFSD8*	[64,65]
				known gene - novel phenotype	*GRIN2A, GRIN2B*	
Targeted resequencing gene panel + WES	8 index cases + 18 trios	IS	Candidate approach	known gene - known phenotype	*STXBP1, CASK, ALG13, PNPO, ADSL*	[9]
WGS	6 trios	EE	*de novo*	known gene - known phenotype	*KCNQ2, SCN2A*	
				known gene - novel phenotype	*KCNT1, PIGQ*	[68]
				candidate gene	*CBL, CSNK1G1*	
WGS	index case	OS-like	*de novo*	known gene - novel phenotype	*SCN2A*	[63]
WGS	quartet	EE	*de novo*	novel gene - novel phenotype	*SCN8A*	[66]
WGS	6 family members	ICCA	AD- heterozygous	known gene - novel phenotype	*PRRT2*	[67]

AR: autosomal recessive; AD: autosomal dominant; ID: intellectual disability; LGS: Lennox-Gastaut Syndrome; EE: epileptic encephalopathy; MMPSI: malignant migrating partial seizures in infancy; DOORS: deafness, onychdystrophy, osteodystrophy, and mental retardation Syndrome; ICCA: infantile convulsions and choreoathetosis syndrome; FFEVF: familial focal epilepsy with variable foci; ADNFLE: autosomal dominant nocturnal frontal lobe epilepsy; DS: Dravet syndrome; IS: infantile spasms; OS: Ohtahara syndrome; NGS: next generation sequencing; WES: whole exome sequencing; WGS: whole genome sequencing


Study design determinants
Different modes of inheritance
Personal choice of NGS technology
START
* Index patient with assumed genetic etiology
DETERMINE STRATEGY
Pinpoint most likely inheritance pattern
Optional
Prescreen index case
Identification of variant in known gene or good candidate(s)
No
Yes
Only 1 affected individual in family
No
> 1 affected family member in different generations of the family
Yes
AD - heterozygous
No
Known consanguinity
Yes
AR - homozygous
No
Phenotyping
Mild
Reduced penetrance
Severe
AR - compound heterozygous
Yes
Phenotyping
Mild
Severe
Specific phenotype
Yes
No
Re-check *
De novo - heterozygous
Several similar cases
Candidate approach
X-linked, mosaïc, AR, ...
Different inheritance pattern
Copy number variations
Technical errors
Consider alternative hypotheses and possible mishaps
No
DNA availability of relatives of index patient
or
Father
Mother
Affected sib(s)
Unaffected sib(s)
Other affected family members
Other unaffected family members
Other index cases
Sequence additional family members
Choice of sequencing platform depending on availability and expertise
Identification of good candidate variant(s) according to the assumed inheritance pattern
No
Yes
Validate segregation of variant(s) in additional family members
Yes
Molecular diagnosis for patient/family


Variant based
Gene based
NGS variants
quality
frequency
impact
Assumed mode of inheritance
AR inheritance
homozygous
compound heterozygous
AD inheritance
inherited
de novo
X-linked inheritance
recessive
dominant
Data mining
expression data
animal models
pathway analysis
+ additional filtering parameters based on inheritance
causal gene defect
conservation
prediction programs
pathogenicity scores