Comprehensive profiling of the androgen receptor in liquid biopsies from castration-resistant prostate cancer reveals novel intra-AR structural variation and splice variant expression patterns

Reference:
De Laere Bram, van Dam Pieter-Jan, Whittington Tom, Mayrhofer Markus, Henao Diaz Emanuel, van den Eynden Gert, Del-Favero Jurgen, van Laere Steven, Drix Luc, Grönberg Henrik, ... Comprehensive profiling of the androgen receptor in liquid biopsies from castration-resistant prostate cancer reveals novel intra-AR structural variation and splice variant expression patterns
Full text (Publisher's DOI): https://doi.org/10.1016/J.EURURO.2017.01.011
To cite this reference: http://hdl.handle.net/10067/1412100151162165141
Comprehensive profiling of the androgen receptor in liquid biopsies from castration-resistant prostate cancer reveals novel intra-AR structural variation and splice variant expression patterns

Bram De Laere
Pieter-Jan van Dam
Tom Whitington
Markus Mayrhofer
Emanuela Henao Diaz
Gert Van den Eynden
Jean Vandebrrok
Jurgen Del-Favero
Steven Van Laere
Luc Dirix
Henrik Grönberg*
Johan Lindberg*

*These authors jointly supervised this work.

Center for Oncological Research (CORE), University of Antwerp, Antwerp, Belgium
Department of Medical Epidemiology and Biostatistics, Science for Life Laboratory, Karolinska Institutet, Stockholm, Sweden.
Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.
Department of Oncology, GZA Hospitals Sint-Augustinus, Antwerp, Belgium
Department of Pathology, GZA Hospitals Sint-Augustinus, Antwerp, Belgium
Multiplicom N.V., Niel, Belgium

*Corresponding author:
Key words: CRPC, AR splice variants, AR-V7, circulating tumour DNA, structural variation, endocrine treatment, androgen receptor.

Word count of the abstract: 299
Word count of text: 2500
Abstract

**Background:** Expression of the androgen receptor (AR) splice variant 7 (AR-V7) is associated with poor response to second-line endocrine therapy in castration-resistant prostate cancer (CRPC). However, a large fraction of the non-responding patients are AR-V7 negative.

**Objective:** To investigate if a comprehensive liquid-biopsy based AR profile may improve patient stratification in the context of second line endocrine therapy.

**Design, settings and participants:** Peripheral blood was collected from patients with CRPC (n=30) prior to initiation of a new line of systemic therapy. We performed circulating tumour DNA profiling by low-pass whole genome sequencing and targeted sequencing of the entire AR gene, including introns. Targeted RNA sequencing was performed on enriched circulating tumour cell fractions to assess the expression levels of seven AR splice variants (ARVs).

**Outcome measurements and statistical analysis:** Somatic variation in AR, including copy-number alterations, structural variation and point mutations was combined with ARV expression patterns and correlated to clinic-pathological parameters.

**Results and limitations:** Collectively, any AR perturbation, including ARVs, was detected in 25/30 patients. Surprisingly, intra-AR structural variation was present in 15/30 patients of whom 14 expressed ARVs. The majority of ARV-positive patients expressed multiple ARVs, where AR-V3 was the most abundantly expressed. Presence of any ARV was associated with progression-free survival of second-line endocrine treatment (hazard ratio: 4.53; 95% CI, 1.424–14.41; \( p = 0.0105 \)). Six out of 17 poor responders were AR-V7 negative, but four carried other AR perturbations.
**Conclusions:** Comprehensive AR profiling, feasible from liquid biopsies, is necessary to increase the understanding of the mechanisms underpinning resistance to endocrine treatment.

**Patients summary:** Alterations in the androgen receptor are associated with endocrine treatment outcome. This study demonstrates that it is possible to identify different types of alterations by simple blood draws. Follow-up studies are needed to determine the effect of such alterations on hormonal therapy.

**Introduction**

Prostate cancer is the most common diagnosed cancer in men, with nearly 410,000 diagnoses in Europe each year. Approximately 20-25% will develop metastatic disease, which inevitably will progress to lethal castration-resistant prostate cancer (CRPC). CRPC is characterized by progressive disease under maximal androgen blockade. Nonetheless, a continued targeting of the androgen receptor (AR) has demonstrated that this signalling pathway still remains one of the main drivers of progressive disease, even in the CRPC setting [1]. Besides taxane-based chemotherapy regimens, next generation androgen deprivation therapies, encompassing both the CYP17 inhibitor abiraterone acetate and novel anti-androgens such as enzalutamide have become available. However, up to 20-40% of patients have resistant disease at the start of these second-line AR therapies [2-5].

Various perturbations in AR, e.g. mutations [6-8], amplifications [9-11] and splice variants [12-15], have been associated with resistance to androgen deprivation therapies. The emergence of mutations is affected by the treatment history as individual mutations
have different clinical consequences [16,17]. Amplifications occur in 29-45 % of CRPC patients before start of a new antiandrogen therapy [9,11,18] and increase the expression of AR, which have shown to be associated with resistance to next-generation androgen deprivation therapies [11]. Furthermore, AR splice variants (ARV) can act as constitutively active transcription factors, bypassing the need of activating ligands and therefore stimulating ligand-independent growth and progression of the disease [19-22].

Prostate cancer metastasizes primarily to the bone [2] with low success rates in obtaining adequate material for profiling, even in the research setting [18]. The application of liquid biopsies, in the form of circulating tumour cells (CTCs), circulating tumour DNA (ctDNA) or exosomes has the potential to enable biomarker profiling without access to metastatic tissue. Consequently, AR-V7 has recently been linked to resistance to abiraterone acetate and enzalutamide in multiple studies, all applying various forms of liquid biopsy [12-15].

However, there is an on-going discussion on the discriminatory value of detecting AR-V7 expression [23]. As other AR perturbations have been associated with endocrine treatment outcome, it is likely that a combination test, on both the DNA and RNA level will improve patient stratification. Previous work pioneered by Li and colleagues demonstrated a connection between structural variation within AR and the generation of non-canonical transcripts [24,25]. We hypothesized that at least a subset of CRPC patients may carry relevant intra-AR variation.
Therefore, we performed a pilot study on a selected cohort of patients with CRPC by thoroughly profiling AR on both the DNA and RNA level in liquid biopsies (n=34) from 30 patients. Our profiling combined mutations, copy-number variations (CNVs) and sequencing of the entire AR gene, including introns, in combination with expression information from the full-length AR and seven AR splice variants (i.e. AR45, AR-V1, AR-V2, AR-V3, AR-V5, AR-V7 and AR-V9). The goal of this study was to investigate if a comprehensive AR profile could provide additional information to stratify patients beyond AR-V7 expression in the context of endocrine treatment.

**Patients and methods**

For a detailed description of all materials and methods we refer to the Supplementary Text. In brief, via a non-interventional clinical study, we collected blood samples from chemotherapy pre-treated and treatment–naïve patients with CRPC. Ethical approval was obtained by the institutional review and ethics board of GZA Sint-Augustinus. All patients provided a written informed consent document. Blood collections encompassed samples for germline DNA extraction, CTC enumeration, CTC enrichment and extraction of cell-free DNA from plasma.

ARV expression levels were assessed by performing cDNA synthesis, multiplex exon-junction specific PCR (MASTR, Multiplicom NV) and Illumina sequencing on RNA derived from CellSearch-enriched CTC fractions. DNA based library prep was performed using the ThruPLEX DNA-seq kit (Rubicon Genomics). Low-pass whole genome sequencing (1x50 bp) was performed to enable identification of copy-number alterations. Targeted sequencing was performed by applying the SeqCap EZ system (Roche
Nimblegen) for the detection of point mutation and intra-AR structural variation (2x100 bp). Sequencing was conducted on the Hiseq2500 instrument in rapid mode. Details on sequence data processing and statistical analysis is available in the Supplementary Text. To identify intra-AR structural variation we developed an in-house structural variant calling algorithm, "svcaller", which is publically available (https://github.com/tomwhi/svcaller).

**Results**

From October 2013 through June 2015, liquid biopsies (n=34) were collected from 30 patients with CRPC. Clinic-pathological and radiological features of the cohort are given in Supplementary Table 1. The selected cohort encompasses patients with poor prognosis, with 17/30 (56.7%) patients having M1 disease at initial diagnosis. The goal was to thoroughly investigate the molecular status of AR in the context of endocrine treatment (Fig. 1).

Cell-free DNA (cfDNA) was successfully extracted from 33 plasma samples and sequencing libraries were constructed. The libraries were subjected to low-pass whole genome sequencing to determine the copy-number status of AR as well as genome-wide somatic CNVs (Supplementary Table 2, Supplementary Fig. 1). Amplifications in AR were detected in 20 out of 30 patients, with high-level amplifications in eleven patients.
Subsequently, targeted sequencing was performed by in-solution based hybridization capture on the same sequencing libraries used for low-pass whole genome sequencing. The target region contained baits complementary to 112 genes (Supplementary Table 3), including all coding exons and non-repetitive intronic regions of AR (Supplementary Fig. 2). The overall average coverage was 1169x (IQR: 904.5x - 2180x) (Supplementary Table 2). Somatic mutations were detected in all profiled samples (Supplementary Fig. 3, Supplementary Table 4). Genes, previously reported to be overrepresented in CRPC compared to primary prostate cancer [18], such as TP53, MLL2, AR, FOXA1 and APC were recurrently mutated. Clonal mutations, adjusted for CNVs, were used to assess the fraction of ctDNA in the cfDNA (average 0.39, range 0.067 – 0.76, Supplementary Table 2). The ctDNA fraction correlated with CTC enumeration by CellSearch \((r = 0.54, p = 0.0037)\).

The use of paired-end sequencing allowed us to identify structural variants, commonly present in prostate cancer [26], through split read and discordant read analysis (Supplementary Fig. 4). A previously published algorithm [27] was applied in parallel with an in-house approach for calling and visualizing variants to minimize false positives. Intra-AR structural variation was detected in 17 out of 33 profiled cfDNA samples, representing all four classes of structural variants as exemplified by selected events in Fig. 2. Patient 4120 carried a deletion overlapping the ligand binding domain (LBD) (Fig. 2A), with independent support provided by calling CNVs based on coverage information (Supplementary Fig. 5A). Patient 4118 carried an inversion, flipping the entire LBD (Fig. 2B). Patient 3843 harboured an ~18 Mb tandem duplication, originating in intron 1 of AR
(Fig. 2C). However, the low-pass whole genome sequencing data from chromosome X suggests a more complex event, where spatially distant regions of AR may have merged (Supplementary Fig. 5B). Additionally, a translocation of patient 4038 removed the LBD, fusing chromosome X to a gene desert of chromosome 16 (Fig. 2D). Surprisingly, the majority of cfDNA samples with intra-AR structural variants harboured multiple events in AR (11/17, Fig. 3), with focal alterations frequently affecting either the cryptic exon region or the LBD.

To determine the expression of AR splice variants (ARVs), directly from CTCs, we developed and validated an RNA-seq approach (Supplementary Fig. 6). To verify presence of CTCs, conventional CellSearch processing was performed to count the number of CTCs per 7.5 ml of blood. In total, 28/34 (82.3%) blood samples had detectable CTCs (median: 72, IQR: 15 - 239). This allowed us to infer expression of full-length AR and seven isoforms (i.e. AR45, AR-V1, AR-V2, AR-V3, AR-V5, AR-V7 and AR-V9). Fourteen out of 15 unique patients with intra-AR structural events expressed ARVs (Supplementary Fig. 7) suggesting a correlation between intra-AR structural variation and expression of ARVs. The only patient with intra-AR structural variation without detectable ARVs was 4175 who rapidly progressed on abiraterone treatment (4.7 months).

Next we performed ARV analysis on CTC fractions (n=30) from 26 patients. In 18/26 patients we collected blood at the start (i.e. baseline) of abiraterone or enzalutamide, with 4 cases also having blood sampling at progressive disease (Supplementary Fig. 8). In the
other 8/26 patients we only collected blood at the moment the patient was progressive on abiraterone or enzalutamide and a new agent had to initiated. Patients were stratified in three categories, based on time to treatment failure (Fig. 4). The average number of reads from RNA sequencing was 3672x (IQR: 1272 – 5970) (Supplementary Table 5). AR splice variants were detected in 17/30 (56.6%) analysed samples, resulting in 15/26 (57.7%) patients being ARV-positive, of whom 13/15 (86.7%) had less than 6 months benefit from their therapy (Fisher exact: $p = 0.0115$). ARV expression was heterogeneous with 10/15 (66.7%) patients expressing several ARVs. AR-V7 was the most frequently occurring splice variant (12/15 patients), followed by AR-V3 (11/15), AR45 (10/15), ARV9 (6/15), ARV1 (5/15), ARV2 (3/15) and ARV5 (3/15). However, AR-V3 had a 3.5 fold higher median expression compared to AR-V7 (Wilcox signed rank, $p = 0.0029$) (Supplementary Fig. 9). In addition, two poorly responding patients were AR-V7 negative whilst expressing AR-V3.

Baseline ARV data were available in 18 patients (abiraterone (n=16) and enzalutamide (n=2)). Patient 4042 discontinued treatment and was left out of the analysis. At 10-12 weeks, a PSA response rate of 30% and 50% was achieved in 8/17 (47%) and 5/17 (29.4%) patients, respectively. The PSA response rates at 10-12 weeks in patients with (n=8) and without (n=9) any ARV detected were not statistically different (Fisher exact test: $p = 0.15$ and $p = 0.29$, respectively) (Fig. 5A). The median follow-up time for progression-free survival was 5.7 months (range, 0.25 to 18.3). At the time of analysis, one patient was still receiving abiraterone. Presence of an AR splice variant was
associated with progression-free survival, as measured by time to no longer clinically benefitting (hazard ratio: 4.53; 95% CI, 1.42–14.41; \( p = 0.0105 \)) (Fig. 5B).
**Discussion**

We performed comprehensive profiling of AR in liquid biopsies obtained from 30 CRPC patients. By simultaneously analysing CNVs, point mutations, intra-AR structural variation and splice variants we present three key findings: 1) intra-AR structural variation is prevalent and questions the ubiquitous assumption that non-canonical AR transcripts arise merely as a consequence of alternative splicing; 2) the majority of splice variant positive patients express multiple variants with AR-V3 being most abundantly expressed; 3) comprehensive AR profiling will be necessary to increase the understanding of the molecular mechanisms underpinning resistance to endocrine treatment.

To investigate the presence of intra-AR structural variation we performed targeted sequencing of the entire AR gene, including non-repetitive intronic regions. Structural variation was detected in 15/30 profiled patients and the clustering of focal events towards the 3’ regions of AR was striking (Fig. 3). In turn, 14/15 patients harbouring intra-AR structural variation also expressed splice variants. The only intra-AR structural variation-positive patient that was splice variant-negative rapidly progressed on abiraterone treatment. Although we cannot demonstrate a causal relationship between intra-AR structural variation and expression of non-canonical AR transcripts, previous pioneering work on disease models have demonstrated an association between intra-AR genomic variation and the generation of truncated AR transcripts [24,25,28].
The majority of cfDNA samples with intra-AR variation harboured multiple structural variants in AR (11/17). Events were detected that are highly likely to generate a non-functional version of the androgen receptor. Due to the evolutionary pressure of endocrine treatment, diverse somatic versions of AR will emerge. In the context of AR amplifications, multiple versions of AR may exist within the same cell with non-functional versions existing in parallel to truncated and full-length copies of AR. Interestingly, multiple structural variants were also detected in 4120-P-2015352, harbouring no amplification (Fig. 3). Therefore, we suggest a model, where multiple versions of AR may emerge either within the same clone or between different clones as a consequence of endocrine treatment (Supplementary Fig. 10).

Structural variation was not detected in all splice variant positive liquid biopsies. It has been shown that in the context of AR amplification, increased elongation rates leads to generation of splice variants, which suggests a connection to other processes than intra-AR structural variation [19]. However, as this is a retrospective study, we generated DNA sequencing libraries from ~1.25 ml available plasma aliquots. Low DNA input would potentially limit the power to detect structural variation, especially as baits applied for enrichment are biased towards the reference genome (Supplementary Fig. 11). Indeed, to achieve 95% sensitivity to detect all variants, including subclonal events, >10,000X coverage would be needed (Supplementary Fig. 12). As expected, there was a clear coverage bias, where samples with detectable intra-AR had higher coverage than negative samples (Supplementary Fig. 13). Therefore, it is almost certain that undetected intra-AR variation was due to sequence coverage. In addition, as repetitive regions exist in AR,
which are not possible to enrich and sequence, variants may be missed if: 1) both ends of the structural variant reside within repetitive regions; 2) one end lies within repetitive DNA and the other is outside AR. Of note, all intra-AR structural variants detected here were classified as somatic, except the 3.7 kb intron 1 deletion detected in 3883-P-2013569, which was present in the germline DNA. At this stage we have no information on family history, however this is an interesting observation for future investigations.

Of the patients treated with abiraterone or enzalutamide 15/26 (57.7%) were ARV-positive, with nearly all of these patients being resistant (13/15) or showing a moderate response (2/15). Within these patients AR-V7 was the most prevalent splice variant. However, not all poor responders expressed AR-V7. By accounting for both baseline and progression samples, six out of 17 poor responders were AR-V7 negative. Poor responding AR-V7 negative patients was recently reported by Antonarakis et al. and Sher et al., with 6/23 and 65/81 resistant patients being AR-V7 negative, respectively [12,13]. Notably, two AR-V7 negative patients expressed the constitutively active AR-V3 isoform [22,29], which was detected at higher expression levels compared to AR-V7. Two other patients with ARV-negative disease, and poor outcome carried AR-LBD mutations or intra-AR structural variation. No AR perturbations were detected for 4173-P-2014436 and 4072-P-2014101. However, the average targeted DNA sequence coverage for these two samples was too low (212x and 80x coverage, respectively) for robust detection of somatic variation.
During the submission process of this manuscript, Henzler and colleagues published concurrent findings on intra-AR structural variation in CRPC [30]. Structural variation was prevalent in both studies, revealing an AR perturbation landscape spanning cancers with multiple alterations to occasions where no variation could be detected. However, the analysis of Henzler et al analysis was performed on metastatic tissue obtained from rapid autopsy. We performed the analysis directly on liquid biopsies, a prerequisite for future investigations performing comprehensive analysis of AR due to the difficulty of obtaining adequate biopsy material from metastatic sites.

Future studies are now needed to address which molecular mechanisms that are truly responsible for the resistance to endocrine treatment and how to best interrogate these mechanisms to optimize patient stratification. A prospective multi-centric study is currently on-going which will address these questions.

**Conclusions**

Comprehensive AR profiling, possible to perform in liquid biopsies, is necessary to reveal the complexity of the AR signalling processes underpinning resistance to endocrine treatment.
Figure legends

**Fig. 1 – Multilevel analysis of liquid biopsies.** Circulating tumour DNA and AR splice variant expression was analysed from liquid biopsies. **A)** Cell-free DNA was extracted from plasma. Library prep was subsequently performed, preparing the cell-free DNA for Illumina sequencing. Illumina adapters are displayed with yellow or blue colours. Each DNA library was used both for targeted sequencing and low-pass whole genome sequencing. Targeted sequencing was applied to: 1) detect mutations in genes commonly mutated in prostate cancer; 2) investigate the presence of intra-AR structural variation. Low-pass whole genome sequencing was performed to identify copy-number alterations throughout the whole genome. **B)** EpCAM-positive CTCs were enriched on the CellSearch platform. Total RNA was extracted and AR exon-junction RNAseq was performed to determine AR splice variant expression. **C)** To verify the presence of CTCs, the same enrichment was performed as in B), but cells were labelled, stained and counted.

**Fig. 2 – Different classes of intra-AR structural variation.** **A)** 4120-P-2015352 carried a focal deletion overlapping the ligand-binding domain (leftmost panel). Individual reads from each read-pair supporting the event are connected by a grey line and displayed in either orange- (forward orientation) or blue (reverse orientation) colour. The 5'- and 3' regions containing reads spanning the event are shown as black outlined boxes connected with a dashed line. The structure of AR is displayed at the bottom of the panel in dark blue with cryptic exons coloured red. The middle- and rightmost panels contain a zoomed
in view of the 5’- and 3’-regions harbouring reads supporting the structural event. The bases of reads mapping partially to the other end of the structural event are displayed to denote mismatches to the reference genome. The cyan colour indicates regions with complementary sequence to the partially mapped reads. B) 4118-P-2014611 harboured an inversion affecting the ligand-binding domain. C) 3843-P-2013537 presented with a tandem duplication. The 5’-end was located in intron 1 of AR and the 3’-end ~18 Mb upstream in an intergenic region. Colours as for 4120, except for other genes displayed in dark grey. D) 4038-P-2014253 presented with a translocation removing the ligand-binding domain. The ideograms (leftmost panel) at the top display the locations of the 5’- and 3’ end of the translocation with the non-fused regions of the chromosomes greyed out. Nomenclature of sample IDs separated by dashes: patient ID; letter P denoting a prostate cancer patient; unique biobank identifier.

**Fig. 3 – Overview of Intra-AR structural variation.** Each panel display intra-AR variation from one profiled cfDNA sample except the bottom panel. The bottom panel was used to visualize the locations of the coding exons in AR and the baited regions that were enriched and sequenced. The horizontal lines display the start and/or the stop of a structural event. Structural variation was coloured according to the right legend. Intra-AR relevant annotations were given for the top panel. Brackets mark cfDNA samples originating from the same patient, sampled at different time points. Abbreviations: CE, cryptic exon; LBD, ligand binding domain; DEL, deletion; DUP, tandem duplication; INV, inversion; TRA, translocation.
**Fig. 4** – Multi-level AR profiling in patients with CRPC (n=26) at baseline or at progression on abiraterone or enzalutamide. Samples are grouped according to time on therapy. CTC panel: the number of CTCs is expressed per 7.5mL of blood. * denote aborted samples during CellSearch. CNV panel: AR copy number stratified according to amplification status. Intra-AR panel: structural variants across the AR gene. Complex rearrangements denote multiple overlapping variant types within the particular region. AR-LBD MUT panel: hotspot mutations within the ligand-binding domain of AR. Black dots denote hotspot variants detected in full length AR transcripts during RNA sequencing. Bottom two AR-V panels provide the qualitative and quantitative overview of ARV expression. Brackets mark samples coming from the same patient.

**Fig. 5** – Androgen receptor splice variants, PSA response and progression-free survival on hormonal therapy. A) Waterfall plots of prostate-specific antigen (PSA) responses after 10-12 weeks on therapy. Dashed and full lines represent 30% and 50% increase and decreases. * PSA increase of 528%. B) Kaplan-Meier analysis of progression-free survival (time to no longer clinically benefitting (NLCB)) stratified according to ARV presence in CTCs at baseline. *p*-value is calculated via log-rank test.

**Supplementary Fig. 1** – Copy-number status of AR. Low-pass whole genome sequencing was performed to infer copy-number status of AR. Top panel: 3542-P-2014235 carried a complex amplification on chromosome X affecting AR and neighbouring regions. Middle panel: 3949-P-2014061 harboured a focal amplification event affecting one region, including AR. Bottom panel: No amplification could be
detected on chromosome X for 4120-P-2015352. Y-axis: log2 copy-number ratio; X-axis: position on chromosome X. Green points: Binned regions on chromosome X. Horizontal lines: Segmentation of binned regions used to infer copy-number alterations. Dashed vertical lines: location of AR. Colours: Grey, copy number neutral; dark red, moderately amplified; bright red, high level amplification; dark blue, – deletion.

Supplementary Fig. 2 – AR structure and baited regions. The structure of AR, displaying the non-repetitive regions possible to profile using in-solution targeted capture. Bait regions denote regions captured and subsequently profiled. CE – cryptic exon.

Supplementary Fig. 3 – Circulating tumour DNA mutational landscape. The mutations and small indels detected directly from cell-free DNA. X-axis: gene names sorted according to number of detected mutations among all samples. Y-axis: Cell-free DNA samples profiled. Type of mutation is coloured according to the right legend. Only mutations with potentially protein altering function are displayed. Brackets mark cell-free DNA samples originating from the same patient, sampled at different time points. Note: AR Mutations within and outside hotspots are reported here.

Supplementary Fig. 4 – Schematic display of structural variant detection. Top panel: The regions A, B and C are directly adjacent in the reference genome. Region B is deleted in the tumour genome and sequence data is generated by sequencing the DNA of the tumour. Paired end sequencing is applied which generates sequencing data 100 bp
from each end of each DNA fragment, directed inward (arrows). The dashed line denotes unknown sequence from each sequenced DNA fragment. Subsequently, mapping is performed to the reference genome to determine the location of each read from each read-pair. As region B exist in the reference genome, read-pairs supporting the deletion will: 1) harbour unexpected large distance between read-pairs, visualized through angled dashed lines 2) partially map to one end of the structural event, visualized by shaded arrows. Bottom panel: As top panel but displaying an inversion. The reads of read-pairs supporting the event now point in the same direction with unexpectedly large distance.

**Supplementary Fig. 5 – Copy-number alterations for 4120 and 3843.** A) The targeted sequencing data was applied to infer intra-AR deletions for 4120-P-2015352. The AR exons are displayed to visualize the region of AR affected by the deletion. CE – cryptic exon. B) Low-pass whole genome sequencing was performed to infer copy-number alterations on the X chromosome for 3843-P-2013537. Vertical solid lines mark the start and stop of AR. The arrows denote the 5’ and the 3’ end of the tandem duplication. Y-axis: log2 copy-number ratio. X-axis: position on chromosome X. Green points: Binned regions on chromosome X. Horizontal lines: Segmentation of binned regions used to infer copy-number alterations. Colours: Grey, copy number neutral; dark red, moderately amplified; bright red, high level amplification; dark blue, – deletion.

**Supplementary Fig. 6 – Development and validation of a targeted RNA-Seq assay for ARV expression analysis.** A) ARV qRT-PCR and targeted RNA-Seq assay design, with primer and probes against unique exon-cryptic exon-specific junctions. B) Representative result from ARV
sequencing in 22Rv1 demonstrating coupled reads (see red tracking) between exon 3 and sequences within the CE region in intron 3. CE – cryptic exon. C) RNA-seq validation by qRT-PCR for full-length and splice variants in enriched CTC fractions. r denotes Pearson’s correlation coefficient.

**Supplementary Fig. 7 – Multi-level AR profiling in patients harbouring structural variants.**

CTC panel: the number of CTCs is expressed per 7.5mL of blood. * denote aborted samples. CNV panel: AR copy number stratified according to amplification status. Intra-AR panel: structural variants across the AR gene. Complex rearrangements denote multiple overlapping variant types within the particular region. Bottom two AR-V panels provide the qualitative and quantitative overview of ARV expression. Brackets mark samples coming from the same patient.

**Supplementary Fig. 8 – Changes in AR splice and structural variants in patients with pre- and post-abiraterone samples.** A) Patient 3542 had 129 circulating tumour cells (CTCs) at baseline, which were positive for AR splice variants (ARV). Throughout the course of therapy CTCs and number of ARV transcripts increased. AR gene amplification and complex structural variants were only inferable in plasma at progressive disease. B) Patient 3885 demonstrated an increase in CTC number and ARV expression during treatment. AR was highly amplified in both samples, with more complex intragenic rearrangements at progressive disease. C) Patient 4070 demonstrated a decrease in number of CTCs, with absence of ARV expression. AR remained moderately amplified, without any structural variants detected pre- and post-treatment. D) Patient 4174 demonstrated low level ARV expression at the start of therapy, which was undetected at progressive disease. The AR copy number status shifted from a high-level to
moderately amplified state, with an increase in the number of tandem-duplicated structural events downstream of exon 1. For each patient, the first and last PSA measurement represents baseline and progression levels, respectively.

**Supplementary Fig. 9 – AR splice variant abundance across ARV-expressing patients.**
Boxplot analysis of the number of detected ARV transcripts per 1000 sequenced reads.

**Supplementary Fig. 10 – Proposed model of intra-AR structural variation.** Due to the evolutionary pressure of endocrine treatment, pre-existing or spontaneously emerging clones, expressing non-canonical versions of the androgen receptor will outgrow the competition. The expressed, non-canonical transcripts are generated as a consequence of intra-AR variation. In the presence of AR amplifications, full-length, truncated and non-functional versions of the androgen receptor may exist within the same cell. As the sequencing data is mapped to the reference genome, multiple variants will be visualized over AR, as detected in our data and displayed in Fig. 2. In the absence of AR amplifications, independent clones may harbour different versions of AR. Of note, both scenarios may occur simultaneously in the same patient.

**Supplementary Fig. 11 – Bait design in the context of structural variation.** Baits are designed to capture DNA from non-repetitive unique regions of the genome. If the DNA of interest matches the reference genome (left panel) or harbour mutations (middle panel) it will be captured with high efficiency. DNA fragments supporting structural variation is
likely to be captured with low efficiency if the DNA fragment aligns poorly to the bait as a consequence of the structural variant.

**Supplementary Fig. 12 – Theoretical calculation of sequence coverage required to detect intra-AR structural variation.** Calculations were performed in R, to determine the AR X-fold coverage required to detect structural variants with varying prevalence, in terms of the fraction of all reads supporting the event. First, a linear model was fit over the observed samples, with average coverage as the dependent variable and number of read fragments as the predictor variable. For each observed structural variant event identified by svcaller, the fraction of total reads supporting the event was computed. This value \( p \) was then used to compute the minimum number of reads \( N \) required to detect such an event with 95% sensitivity. Here, event detection was defined by observing a number of reads \( n \) supporting the event \( \geq 3 \). \( n \) was assumed to follow a binomial distribution with parameters \( p \) and \( N \). The linear model relating number of read fragments to average AR coverage was then used to convert \( N \) to a coverage value. The resulting minimum required coverage and \( p \) values were then plotted.

**Supplementary Fig. 13 – Coverage and sensitivity to detect intra-AR variation.** Top-panel: Horizontal bars denote sequence coverage for each profiled cfDNA library. Cyan coloured points mark the coverage needed to detect each structural variant with 95% sensitivity. The cyan points are visualized for the sample in which it was detected. Bottom panel: The bottom panel display the cumulative fraction of all detected structural
variants in relation to coverage needed for 95% sensitivity. The coloured inset bars denote the range of coverage for intra-AR positive (green) and negative (red) samples.

**Supplementary Fig. 14 – Comparison of svcaller and Delly output.** A) Svcaller identified three tandem duplication events for 4213-P-2015142 in which at least one event terminus occurred within the AR region, with all three events also detected by Delly. Tandem duplication event calls generated by svcaller are denoted by light blue dashed lines, with soft-clipped sequence support denoted with solid vertical lines at the event termini. Tandem duplication event calls generated by Delly are denoted by dark blue rectangles. AR and other gene annotations are displayed in light grey at the bottom of the panel whilst the chromosome X ideogram and corresponding selected genomic region is indicated at the top of the panel. Individual read-pairs retained by the tandem duplication event filter are shown, connected by a grey line and displayed in either orange- (forward orientation) or blue (reverse orientation) colour. B) Svcaller identified seven inversion events for 3542-P-2014235 in which at least one event terminus occurred with the AR region, with all seven events also detected by Delly. Colours as for A), although read pairs displayed are in this instance those retained by the inversion event filter.

**Table legends**

**Supplementary Table 1** – Patient and Sample Characteristics

**Supplementary Table 2** – Basic sequencing metrics, AR copy-number status and circulating tumour DNA fraction for each cell-free DNA sample.

**Supplementary Table 3** – HG19 bait coordinates for targeted sequencing.
**Supplementary Table 4** – Somatic mutations detected in the circulating tumour DNA by targeted sequencing.

**Supplementary Table 5** – AR and ARV RNA sequencing library sizes (expressed as number of BWA-mem mapped reads)

**Supplementary Table 6** – Primer and hydrolysis probe sequences for targeted AR and ARV sequencing and qRT-PCR

**Acknowledgements**

This study was supported by AstraZeneca-Karolinska Insitutet Joint Research Program in Translational Science (grant number 2360/12); the Swedish Research Council (grant number K2010-70X-20430-04-3); the Swedish Cancer Foundation (grant number 09-0677); The Belgian Foundation Against Cancer (grant number C/2014/227) and the University of Antwerp Proof-Of-Concept Grant (I14/022). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
References


