Bioinformatics: Tools for Genome Analysis
Group Project 1/Team 3
AR Gene, chrX:67,500,001-68,570,000
Matt Cook, Julie Garcia, Hadi Issah
February 23, 2018

1.1) Genes in region chrX:67,500,001-68,570,000

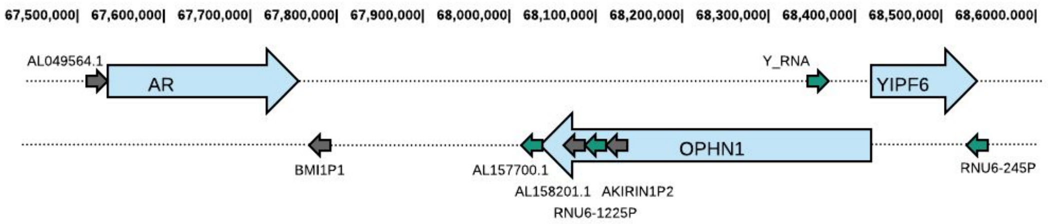


Figure 1. Region chrX:67,500,001-68,570,000 of the h uman chromosome. Arrows show direction of gene transcription. Protein coding genes are in blue (AR, OPHN1, and YIPF6), green genes are non-coding RNA genes, and grey genes are pseudogenes. AR gene transcripts are shown below in Figure 2.

Gene Legend

▭	Protein coding gene
■	Pseudogene
🟩	RNA gene

1.2) AR gene transcripts

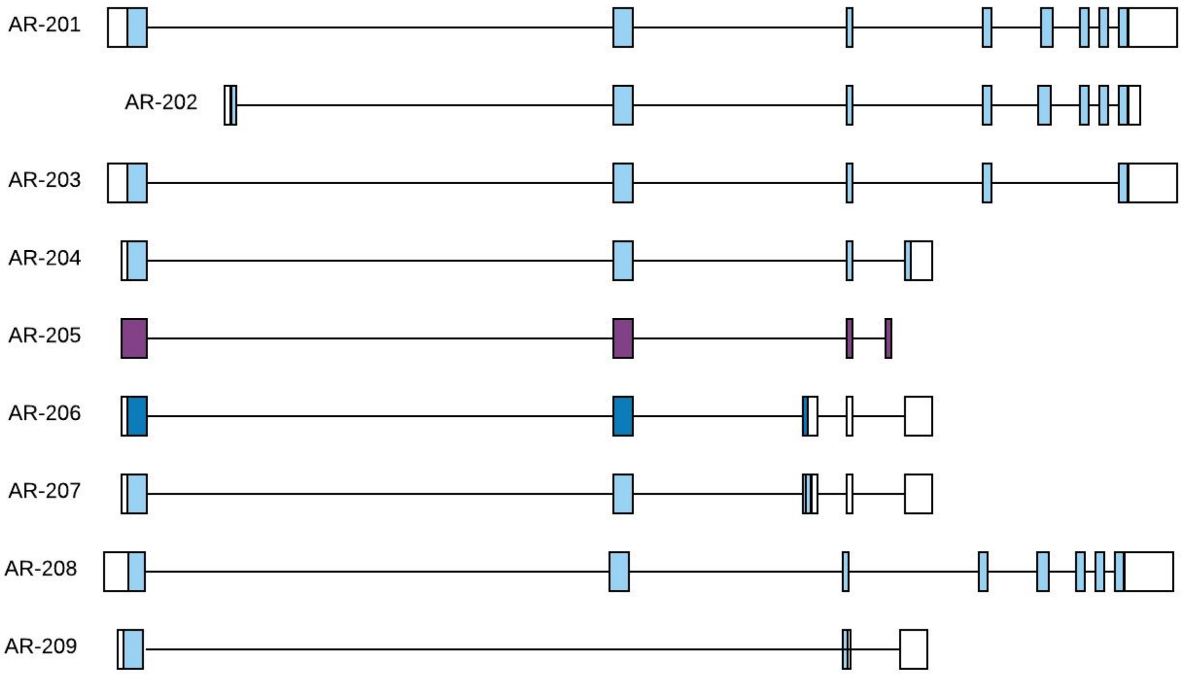


Figure 2. AR Gene has 9 transcripts, 8 protein coding genes, one of those thought to undergo nonsense-mediated decay. One processed transcript with no coding sequence. 5' and 3' UTRs are in white on the ends of the transcripts.

Trancript Legend

🟦 Protein Coding Exons

□ Untranslated Regions

■ Nonsense-mediated decay

🟪 Non-coding exons

1.3) The AR gene product

Androgen Receptor Structure

Androgen receptor (AR) is a nuclear receptor that is activated by binding to testosterone or 5-dihydrotestosterone (Lu et al., 2006). AR is located at Xq11-12 (X:67,544,031-67,730,618) in the human genome and encodes a primary transcript that contains 8 exons. Exon 1 encodes the N-terminal regulatory domain (NTD), exons 2 and 3 the DNA-binding domain (DBD), and exons 4 through 8 the hinge and ligand-binding domain (LBD; Tan et al., 2015; Mangelsdorf et al., 1995). The traditional coding sequence of the transcript is translated into a 919 amino acid protein composed of the 4 domains previously noted. In AR, the NTD is located between residues 1-555, the DBD between residues 555–623, the hinge region between residues 623–665, and the LBD between residues 665–919 (Tan et al., 2015). While it is obvious that the structure of the LBD and DBD greatly affect AR activity, the NTD also contains a number of motifs that affect AR functionality, including the AF1 motif and polyglutamine/polyglycine tracks (Callewaert et al., 2006; Sasaki et al., 2003).

Androgen Receptor Mechanism of Action

Steroid binding to the LBD causes conformational changes within AR. These conformational changes result in the dissociation of chaperone proteins from AR and the exposure of the AR nuclear localization signal (NLS; Cuttress et al., 2008). Importin-α binds to the NLS and interacts with importin-β. Importin-β mediates interactions with the nuclear pore complex that ultimately leads to the translocation of AR from the cytoplasm into the nucleus (Cuttress et al., 2008).

In the nucleus, AR homodimers bind to androgen response elements (ARE) in the promoter regions of a variety of genes (Tan et al., 2015). The AR homodimers typically bind to two hexamers (5’-AGAACA-3’) that are each separated by 3 bases (Shaffer et al. 2004). DNA-bound AR recruits a number of factors that facilitate or co-regulate transcription, including TBP, TFIIF, and CBP (Tan et al., 2015). Kallikrein related peptidase 3 (KLK3) and transmembrane protease, serine 2 (TMPRSS2) are two genes that are known to be regulated by AR (Lin et al., 1999; Cleutjens et al., 1996; Cleutjens et al., 1997). Currently, KLK3 is used as a biomarker in clinical diagnostics - increased levels of KLK3 are indicators of prostate cancer (Tan et al., 2015).

Androgen Receptor and Disease

AR is known to play an important role in male sexual differentiation and in the maintenance of spermatogenesis (Quigley et al., 1995; Patrão et al., 2009). A number of AR variants have been associated with human diseases including complete androgen insensitivity syndrome (CAIS), partial AIS (PAIS), mild AIS (MAIS), and Kennedy’s disease. CAIS is characterized by a complete lack of cellular response to androgens resulting in a loss of male genital development. Generally, individuals with CAIS present phenotypically as female. PAIS can result in various phenotypes of partially masculinized genitalia (Quigley et al., 1995). MAIS is associated with idiopathic azoospermia (lack of spermatogenesis), but male genital differentiation and development are unaffected (Mou and Gui, 2016).

AR activity can be affected by point mutations at specific sites within exons and/or introns, by frameshifts, and by insertions and deletions (McPhaul 2002; Tan et al., 2015). Commonly, CAIS is caused by point mutations in the conserved regions of the LBD or DBD, and rarely in the NTD (Bermúdez de la Vega et al., 2015). An example of a point mutation that has been associated with CAIS is a C>G mutation in exon 2 (position 1752) of the AR transcript. This nonsynonymous mutation leads to a Phe>Leu at position 585 in AR. Entire exon deletions (i.n. exon 2 deletion) have also been associated with CAIS (Shao et al., 2015).

In the human population, it has been shown that there is a variable number of CAG repeats in the polyglutamine tracts found in the AR NTD; the average male has 22 CAG repeats (Edwards et al., 1992; Hardy et al., 1996). A lower repeat number has been strongly associated with prostate cancer while high repeat numbers (38-62) have been associated with the neurodegenerative Kennedy's disease (Hardy et al., 1996; Dejager et al., 2002). It is clear that the AR gene plays a critical role in male sexual development and that AR mutants can have significant impacts on human health and the manifestation of disease.

1.4) Non-coding regions of chrX:67,500,001-68,570,000

Galaxy and data from the UCSC Genome/Table Browser were used to identify conserved non-coding regions close to or within the AR gene. Conserved non-coding regions may facilitate the identification of AR gene functional regulatory elements. First, all conserved regions were identified within this region and then conserved elements within exons were excluded to obtain only those elements within non-coding regions.

Conserved sequences at or in proximity to the AR gene were identified by selecting the Conservation track under the Comparative Genomics group. Three additional tracks were added to the browser: Element Conservation (phastCons/green track), Basewise Conservation (phyloP/blue track), and Multiz Alignments (identifies conserved sequences across selected species). The tracks shown in Figure 3 indicate that the highest peaks are, as expected, under the exons. Although, high peaks are also visible within non-coding sequences.

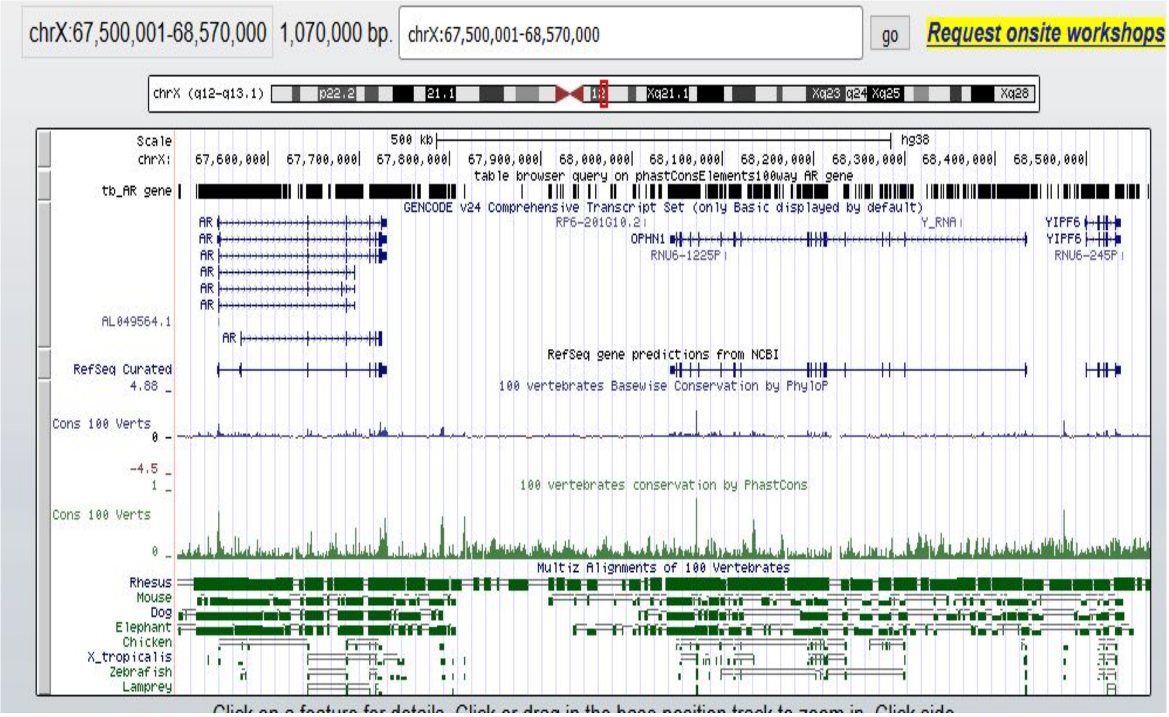
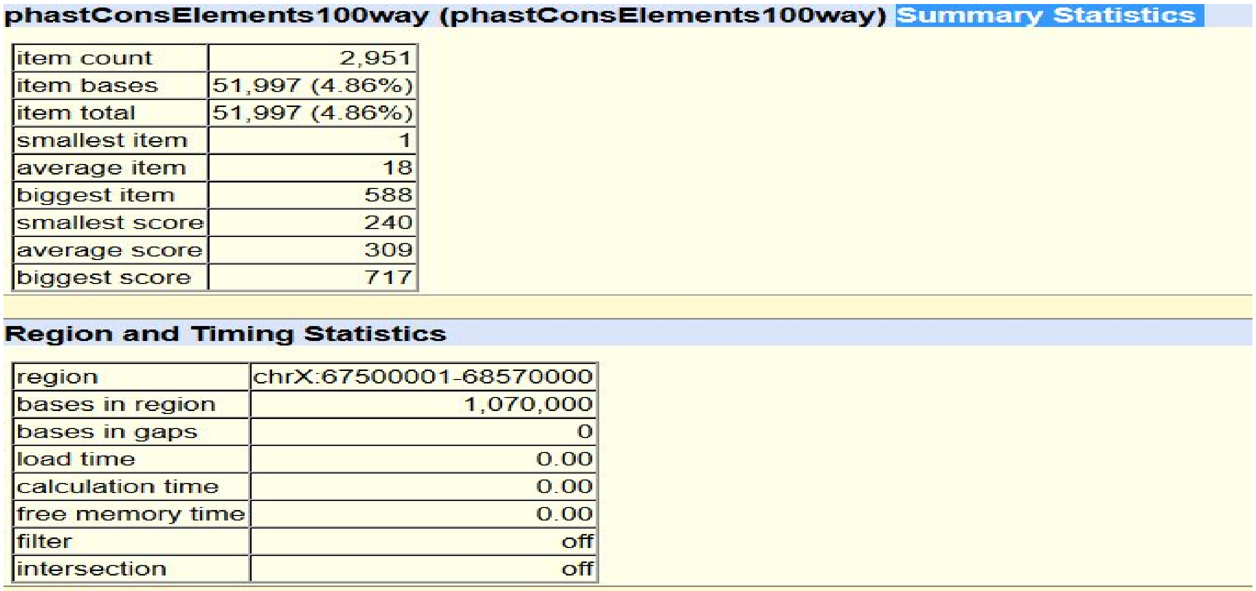

chrX:67,500,001-68,570,000
1,070,000 bp.
chrX:67,500,001-68,570,000
go
Request onsite workshops
chrX (q12-q13.1)
p22.2
21.1
13
Xq21.1
Xq23
q24
Xq25
Xq28
Scale
500 kb
hg38
chrX:
67,600,000
67,700,000
67,800,000
67,900,000
68,000,000
68,100,000
68,200,000
68,300,000
68,400,000
68,500,000
table browser query on phastConsElements100way AR gene
tb_AR gene
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
AR
RP6-201G10.2
OPHN1
RNU6-1225P
Y_RNA
YIPF6
RNU6-245P
AL049564.1
RefSeq gene predictions from NCBI
RefSeq Curated
4.88
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
0
-4.5
1
100 vertebrates conservation by PhastCons
Cons 100 Verts
0
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Click on a feature for details. Click or drag in the base position track to zoom in. Click side

Figure 3. Conserved genomic sequences in close proximity to and within the AR gene.

A custom track, tb_AR gene, identified 2,951 conserved elements scattered in coding and non-coding sequences (Figure 4). These conserved elements indicated by lod scores can be viewed using full mode in the browser.



phastConsElements100way (phastConsElements100way) Summary Statistics

item count	2,951
item bases	51,997 (4.86%)
item total	51,997 (4.86%)
smallest item	1
average item	18
biggest item	588
smallest score	240
average score	309
biggest score	717

Region and Timing Statistics

region	chrX:67500001-68570000
bases in region	1,070,000
bases in gaps	0
load time	0.00
calculation time	0.00
free memory time	0.00
filter	off
intersection	off

Figure 4: Conserved elements summary statistics.

To obtain the conserved elements of the non-coding regions, conserved elements within exons were excluded by completing the following:

1. Pick Table Browser, then choose group Comparative Genomics, track Conservation, and phastConsElements100way in the table section.
2. Return to Table Browser page.
3. Click the create button beside intersection to intersect the data from this track with another track.
4. Select group mRNA and EST and track Human mRNAs and table Human mRNAs (all_mrna).
5. Select the radio button for All Most Conserved records that have no overlap with Human mRNA.
6. Select the radio button for All phastConsElements100way records that have no overlap with Human mRNAs, and then click Submit.

These actions reduced the elements under consideration from 2,951 to 2,242 in the track labeled tb_AR gene (Figure 5).

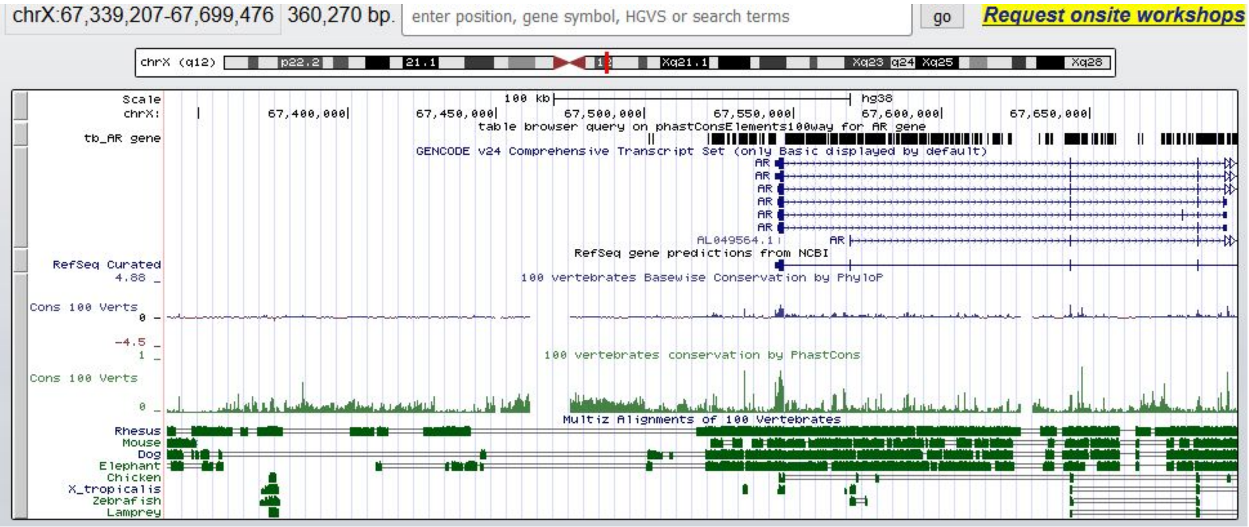
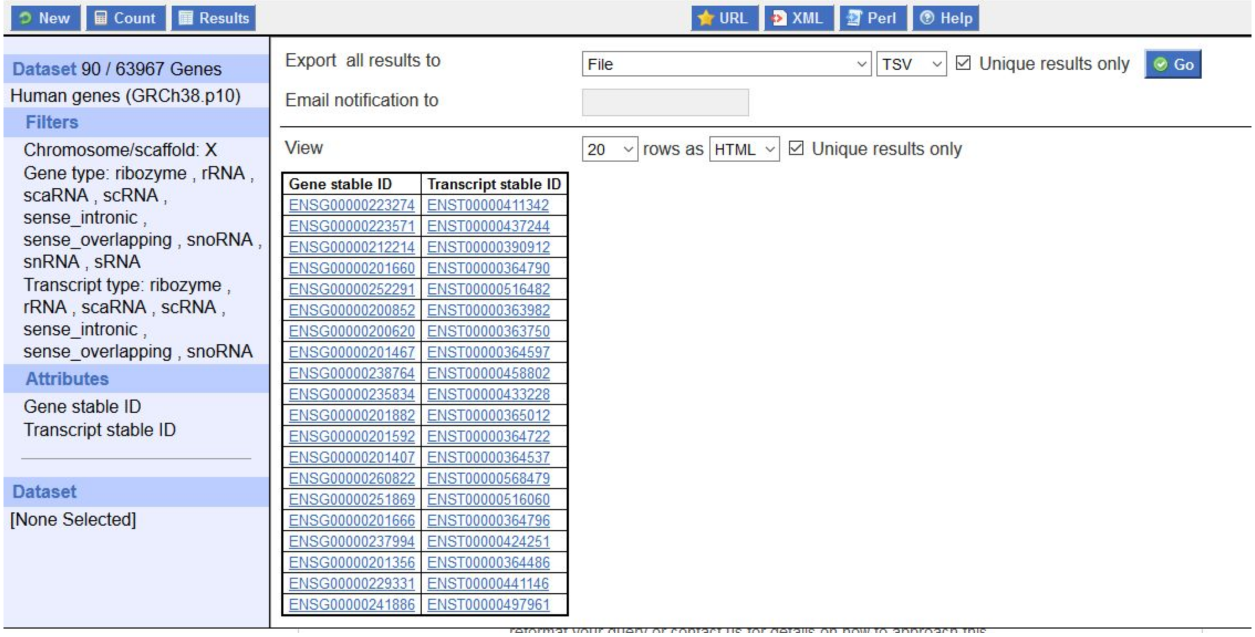

chrX:67,339,207-67,699,476
360,270 bp.
enter position, gene symbol, HGVS or search terms
go
Request onsite workshops
chrX (q12)
p22.2
21.1
Xq21.1
Xq23
q24
Xq25
Xq28
Scale
chrX:
100 kb
hg38
67,400,000
67,450,000
67,500,000
67,550,000
67,600,000
67,650,000
table browser query on phastConsElements100way for AR gene
tb_AR gene
GENCODE v24 Comprehensive Transcript Set (only Basic displayed by default)
AR
AL049564.1
RefSeq gene predictions from NCBI
RefSeq Curated
4.88
100 vertebrates Basewise Conservation by PhyloP
Cons 100 Verts
0
-4.5
1
100 vertebrates conservation by PhastCons
Cons 100 Verts
0
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey

Figure 5. The conserved non-coding regions of chrX:67,500,001-68,570,000. These are conserved non-coding elements that have no overlap with Human mRNAs (tb_AR gene track).

Most conserved non-coding regions are transcribed into functional non-coding RNA molecules (e.g. transfer RNA, ribosomal RNA, and regulatory RNAs). They also act as transcriptional and translational regulators of protein-coding sequences, scaffold attachment regions, origins of DNA replication, centromeres and telomeres. Ensembl was use with the filters shown below to locate 90 non-coding functional RNA transcripts for the AR gene (Figure 6).



Dataset 90 / 63967 Genes
Human genes (GRCh38.p10)

Filters

Chromosome/scaffold: X
Gene type: ribozyme , rRNA , scaRNA , scRNA , sense_intronic , sense_overlapping , snoRNA , snRNA , sRNA
Transcript type: ribozyme , rRNA , scaRNA , scRNA , sense_intronic , sense_overlapping , snoRNA

Attributes

Gene stable ID
Transcript stable ID

Dataset

[None Selected]

Export all results to: File, TSV, ☑ Unique results only

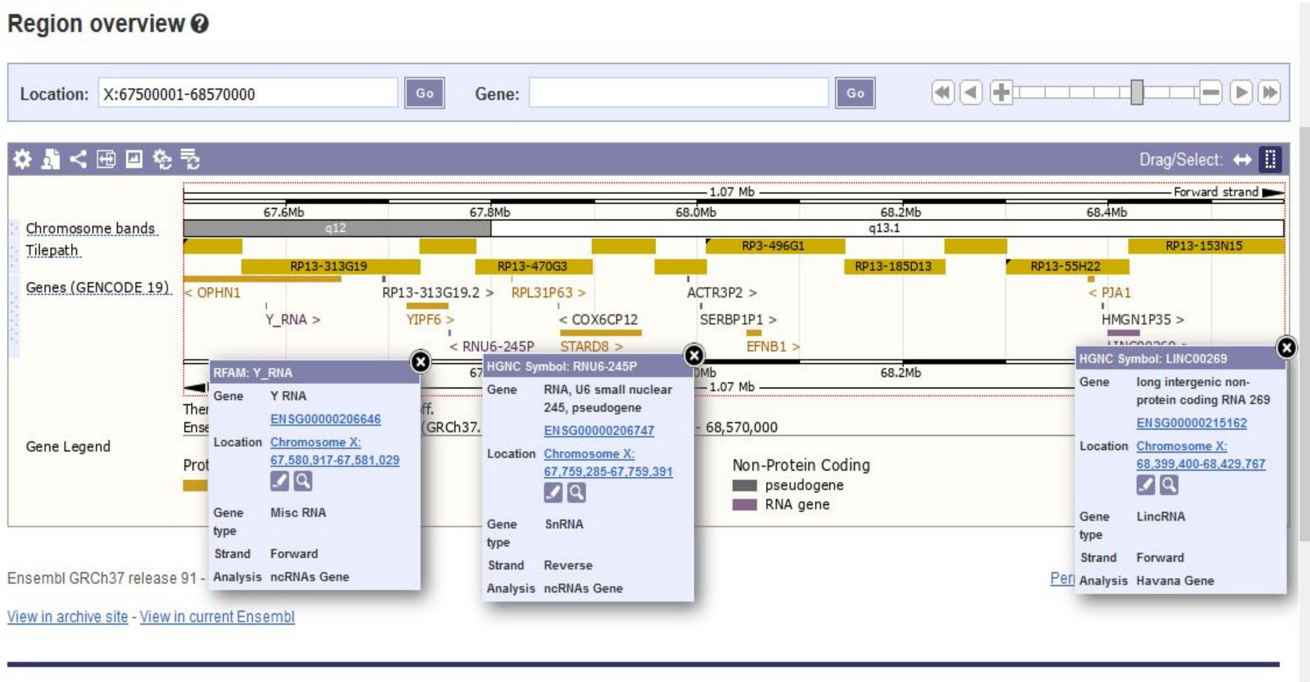
Email notification to:

View: 20 rows as HTML ☑ Unique results only

Gene stable ID	Transcript stable ID
ENSG00000223274	ENST00000411342
ENSG00000223571	ENST00000437244
ENSG00000212214	ENST00000390912
ENSG00000201660	ENST00000364790
ENSG00000252291	ENST00000516482
ENSG00000200852	ENST00000363982
ENSG00000200620	ENST00000363750
ENSG00000201467	ENST00000364597
ENSG00000238764	ENST00000458802
ENSG00000235834	ENST00000433228
ENSG00000201882	ENST00000365012
ENSG00000201592	ENST00000364722
ENSG00000201407	ENST00000364537
ENSG00000260822	ENST00000568479
ENSG00000251869	ENST00000516060
ENSG00000201666	ENST00000364796
ENSG00000237994	ENST00000424251
ENSG00000201356	ENST00000364486
ENSG00000229331	ENST00000441146
ENSG00000241886	ENST00000497961

Figure 6. Non-coding functional RNA transcript for the AR gene in BioMart.

The non-coding regions in our chromosome region of interest (chrX:67,500,001-68,570,000) are shown below. Three genes code for ncRNAs and five are pseudogenes (Figure 7).



RFAM: Y_RNA

Gene	Y RNA ENSG00000206646
Location	Chromosome X: 67,580,917-67,581,029
Gene type	Misc RNA
Strand	Forward
Analysis	ncRNAs Gene

HGNC Symbol: RNU6-245P

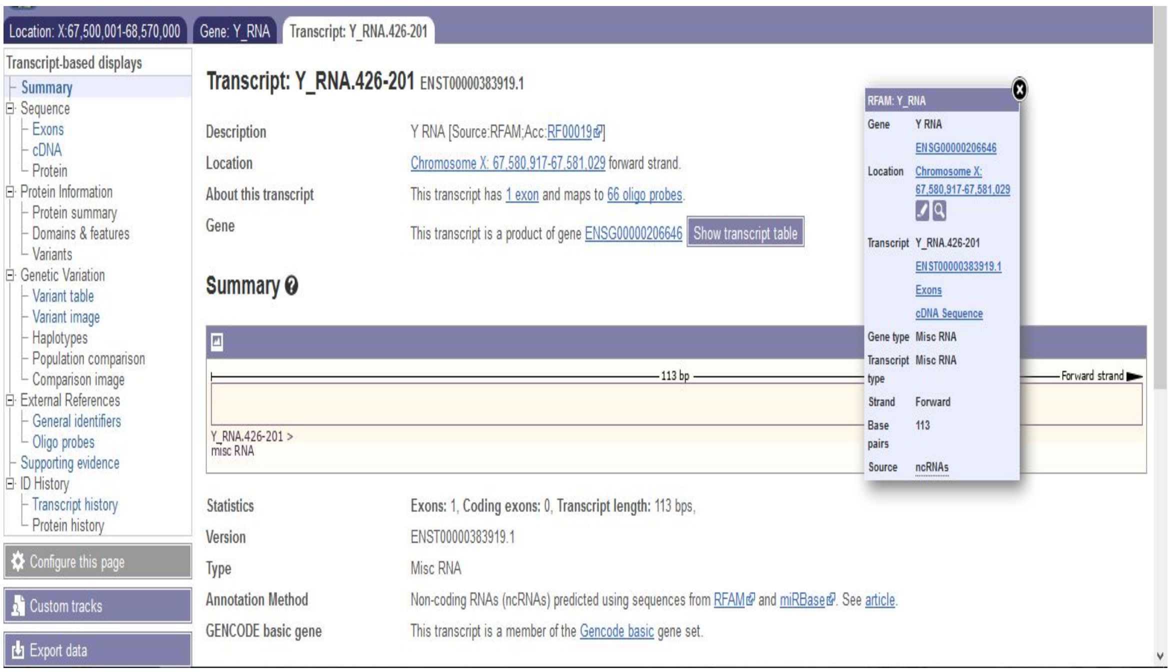
Gene	RNA, U6 small nuclear 245, pseudogene ENSG00000206747
Location	Chromosome X: 67,759,285-67,759,391
Gene type	SnRNA
Strand	Reverse
Analysis	ncRNAs Gene

HGNC Symbol: LINC00269

Gene	long intergenic non-protein coding RNA 269 ENSG00000215162
Location	Chromosome X: 68,399,400-68,429,767
Gene type	LincRNA
Strand	Forward
Analysis	Havana Gene

Figure 7. Known non-coding RNAs and pseudogenes in the region chrX:67,500,001-68,570,000.

Details on one of the transcript for the non-coding gene region Y_RNA is shown below. The transcript is a non-coding RNA (ncRNA) with the following statistics: Exons: 1, Coding exons: 0, Transcript length: 113 bp (Figure 8).



Transcript: Y_RNA.426-201 ENST00000383919.1

Description	Y RNA [Source:RFAM;Acc:RF00019]
Location	Chromosome X: 67,580,917-67,581,029 forward strand.
About this transcript	This transcript has 1 exon and maps to 66 oligo probes.
Gene	This transcript is a product of gene ENSG00000206646 Show transcript table

Summary

Statistics	Exons: 1, Coding exons: 0, Transcript length: 113 bps,
Version	ENST00000383919.1
Type	Misc RNA
Annotation Method	Non-coding RNAs (ncRNAs) predicted using sequences from RFAM and miRBase. See article.
GENCODE basic gene	This transcript is a member of the Gencode basic gene set.

RFAM: Y_RNA

Gene	Y RNA ENSG00000206646
Location	Chromosome X: 67,580,917-67,581,029
Transcript	Y_RNA.426-201 ENST00000383919.1 Exons cDNA Sequence
Gene type	Misc RNA
Transcript type	Misc RNA
Strand	Forward
Base pairs	113
Source	ncRNAs

Figure 8. Non-coding gene Y_RNA in region chrX:67,500,001-68,570,000.

2.1) SNP locations in AR gene transcripts

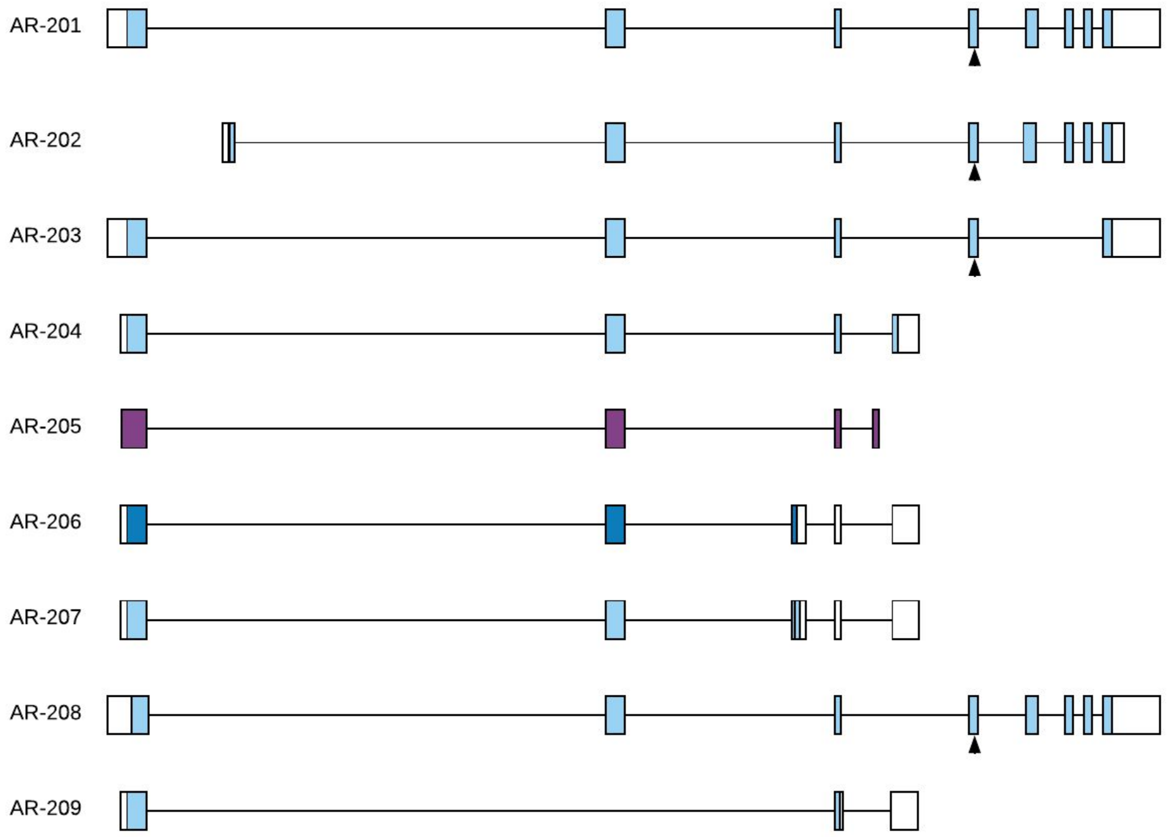


Figure 3. Missense SNP (rs1085307962) location in AR gene transcripts. A total of 4 AR variant transcripts may contain this SNP in exon 4. This SNP was identified by Lou and Gui (2016) as having potential implications in idiopathic azoospermia. AR-201 and AR-203 SNP location c.1888C>T, p.R630W; AR-202 SNP location c.292C>T, p.R98W; AR-208 SNP location c.1318C>T, p.R440W.

2.2) Phenotype Comparison

The exon 4 C>T (Arg630Trp) mutation in AR has been linked to MAIS and specifically idiopathic azoospermia (Mou and Gui, 2016). Due to the differences in polarity, charge, and size between the amino acid residues Arg and Trp, it is believed that this missense mutation affects AR secondary structure (ClinVar, 2018). These structural changes may affect AR ligand binding, activation, and/or DNA binding leading to MAIS. A number of other allelic AR variants have been found to be or are likely to be pathogenic. A total of 198 potentially pathogenic AR variants have been identified by ClinVar including AR variants with deletions, insertions, frameshifts, and missense, nonsense, and splice site mutations. Here, a brief overview of some of these variants and their associated conditions are presented.

McPaul et al. (1992) found that in the individuals they studied, 90% of the mutations they found were located within 35% of LBD (specifically the regions between amino acids 726 and 772 and between 826 and 864). For example, they found that a missense mutation (c.2323C>T) in exon 6 resulted in an Arg772Cys substitution in AR. This change in a conserved region of the LBD resulted in CAIS. Similarly, Sai et al. (1990) found that a nonsense mutation in exon 4 (c.2157G>A; p.Trp719Ter) that is believed to nearly completely remove the LBD from AR caused CAIS. Lastly, Vilchis et al. (2003) identified a 7 bp deletion and 11 bp insertion in exon 5 that resulted in a frameshift mutation (c.2281_2287 delAGGATGCinsTTCGCCCCTGA; p.Arg761Phefs). This frameshift mutation lead to the incorporation of a stop codon 27 bp downstream of the insertion. The researchers believed that this indel was the result of a slipped-strand mispairing mechanism. Ultimately, this mutation truncated the AR LBD and caused CAIS.

While the majority of pathogenic AR mutations are found in the exons, a few have been found in the introns. Brüggenwirth et al. (1997) found a T>A mutation in intron 2 located 11 bp upstream of exon 3. They found that this mutation affected splicing of transcripts and resulted in a 69 bp insertion into the AR transcript. The AR variant produced was unable to effectively bind DNA and resulted in PAIS. Sammarco et al. (2000) also identified a mutation that affected splice sites and resulted in PAIS. A G>T mutation 5 bp upstream of the intron/exon 6 splice site (IVS6, G-T, +5) was identified as the culprit. This mutation caused the inclusion of intron 6 into AR transcripts and resulted in a defective AR LBD.

Mutations in AR have also been linked to prostate cancer. Gaddipati et al. (1994) identified a missense mutation (c.2632A>G; p.Thr878Ala) in the LBD that is thought to decrease ligand binding specificity. Thus, this AR variant could be stimulated by a number of other steroid hormones and this in turn may provide a growth advantage to cancer cells.

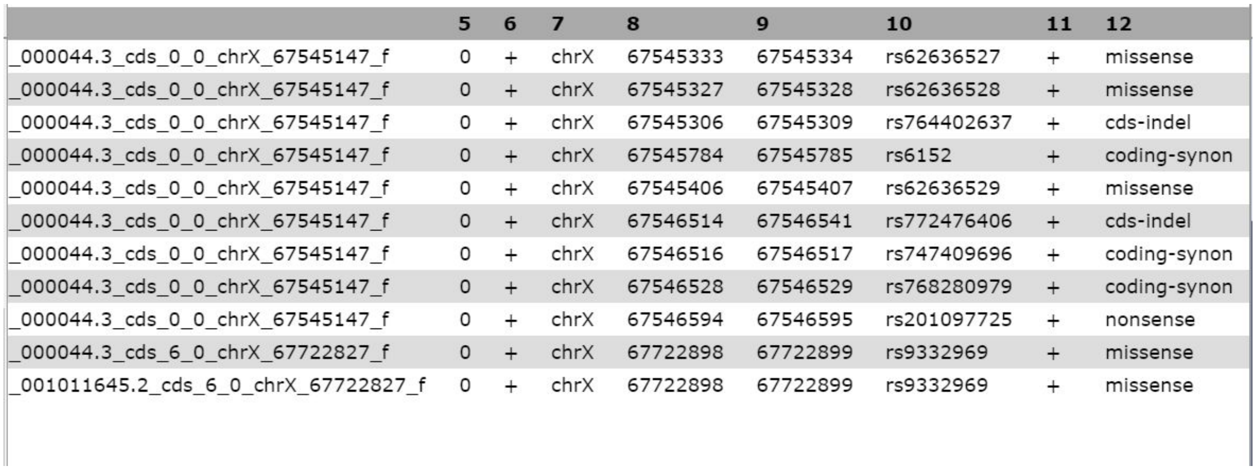
The expansion of the CAG polyglutamine track (>38 repeats) has been linked to Kennedy's disease. La Spada et al. (1991) were the first to make the association between the expansion and the neurodegenerative disease. This expansion occurs in the first exon of the AR transcript in the NTD and results in an in-frame AR variant.

It is apparent that there is a number of AR variants with pathogenic phenotypes and that these phenotypes are dependant on the type and location of genomic AR mutations. Mutations, like deletions and frameshifts, that ablate AR functional domains and hence activity appear most likely to have significant downstream effects. Our mutant appears to have a much more limited effect on AR functionality than many of the other mutants noted above. Thus, the phenotypic presentation of disease in our variant is much less noticeable than in conditions like PAIS and CAIS.

2.3) SNPs and CNVs in Galaxy, Biomart, IGV

SNPs in the AR Gene

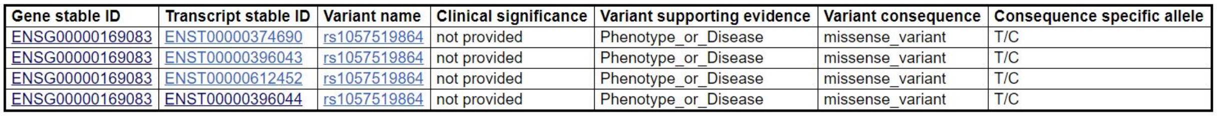
First we used Galaxy and the UCSC Main Table Browser in order to locate SNPs in the AR Gene overlapping with coding exons. When selecting All SNPs, there were 16,033 SNPs overlapping with the AR gene. Choosing Common SNPs(150) we get 307 SNP results and when selecting Flagged SNPs or clinically relevant SNPs there were 65 results. We then joined the coding exon data with the 307 common SNPs to find SNPs to most likely affect the resulting protein. There were 11 results shown in the table below.



	5	6	7	8	9	10	11	12
_000044.3_cds_0_0_chrX_67545147_f	0	+	chrX	67545333	67545334	rs62636527	+	missense
_000044.3_cds_0_0_chrX_67545147_f	0	+	chrX	67545327	67545328	rs62636528	+	missense
_000044.3_cds_0_0_chrX_67545147_f	0	+	chrX	67545306	67545309	rs764402637	+	cds-indel
_000044.3_cds_0_0_chrX_67545147_f	0	+	chrX	67545784	67545785	rs6152	+	coding-synon
_000044.3_cds_0_0_chrX_67545147_f	0	+	chrX	67545406	67545407	rs62636529	+	missense
_000044.3_cds_0_0_chrX_67545147_f	0	+	chrX	67546514	67546541	rs772476406	+	cds-indel
_000044.3_cds_0_0_chrX_67545147_f	0	+	chrX	67546516	67546517	rs747409696	+	coding-synon
_000044.3_cds_0_0_chrX_67545147_f	0	+	chrX	67546528	67546529	rs768280979	+	coding-synon
_000044.3_cds_0_0_chrX_67545147_f	0	+	chrX	67546594	67546595	rs201097725	+	nonsense
_000044.3_cds_6_0_chrX_67722827_f	0	+	chrX	67722898	67722899	rs9332969	+	missense
_001011645.2_cds_6_0_chrX_67722827_f	0	+	chrX	67722898	67722899	rs9332969	+	missense

Figure 9. Results of joining 307 common SNPs in the AR gene region, with the coding exon region of the AR gene. This resulted in 11 common SNPs in the coding region (6 missense, 2 cds-indel, and 3 coding-synon).

In BioMart, first we searched for all variants from all sources overlapping with the AR Gene. Searching the Human genes (GRCh38.p10) dataset, filtering by chromosome X, and the AR Gene (Ensembl Gene Stable ID = ENSG00000169083), then selecting Somatic Variant attributes (Variant Name) provided 5456 results. Most of these results came from the Catalogue Of Somatic Mutations In Cancer (COSMIC) database. Further filtering to only retrieve those with supporting clinical evidence (Source = ClinVar) we found one SNP of clinical significance shown in the table below that affects 4 of the transcripts of the AR gene. This SNP is a missense (T/C) variant that is linked to a phenotype (Figure 10).



Gene stable ID	Transcript stable ID	Variant name	Clinical significance	Variant supporting evidence	Variant consequence	Consequence specific allele
ENSG00000169083	ENST00000374690	rs1057519864	not provided	Phenotype_or_Disease	missense_variant	T/C
ENSG00000169083	ENST00000396043	rs1057519864	not provided	Phenotype_or_Disease	missense_variant	T/C
ENSG00000169083	ENST00000612452	rs1057519864	not provided	Phenotype_or_Disease	missense_variant	T/C
ENSG00000169083	ENST00000396044	rs1057519864	not provided	Phenotype_or_Disease	missense_variant	T/C

Figure 10. BioMart SNPs of clinical significance in the AR gene region.

In IGV, we loaded all SNPs from dbSNP in the AR region. Numerous SNPs in the AR gene region can be seen the figure below (Figure 11).

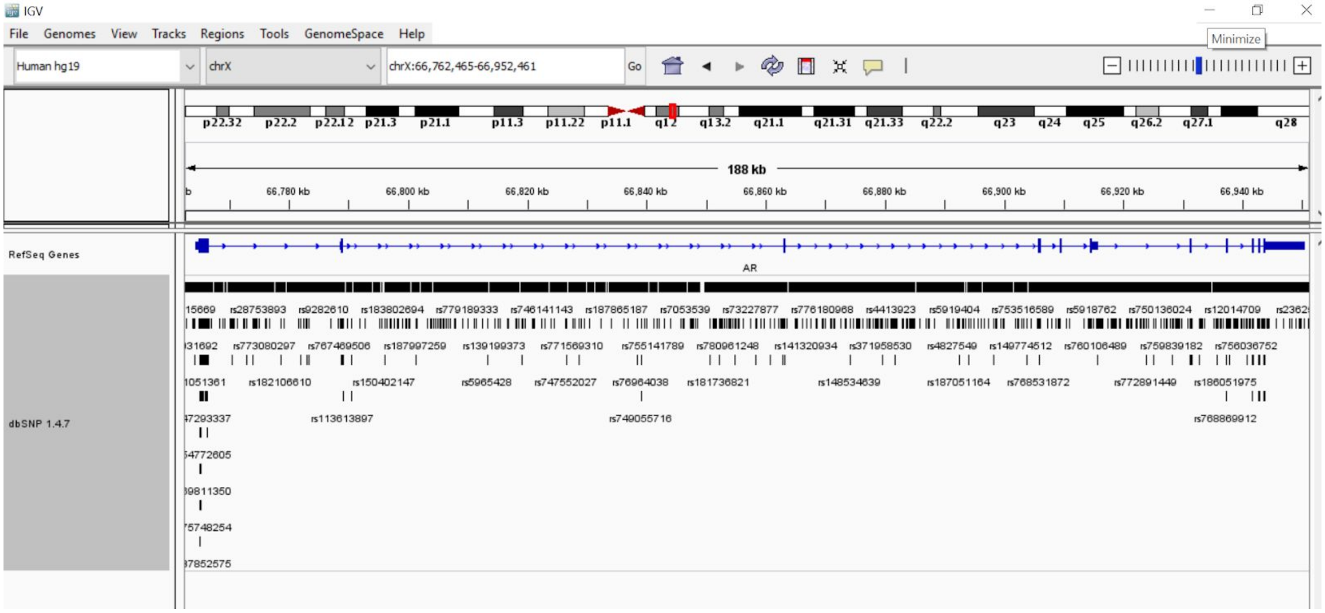
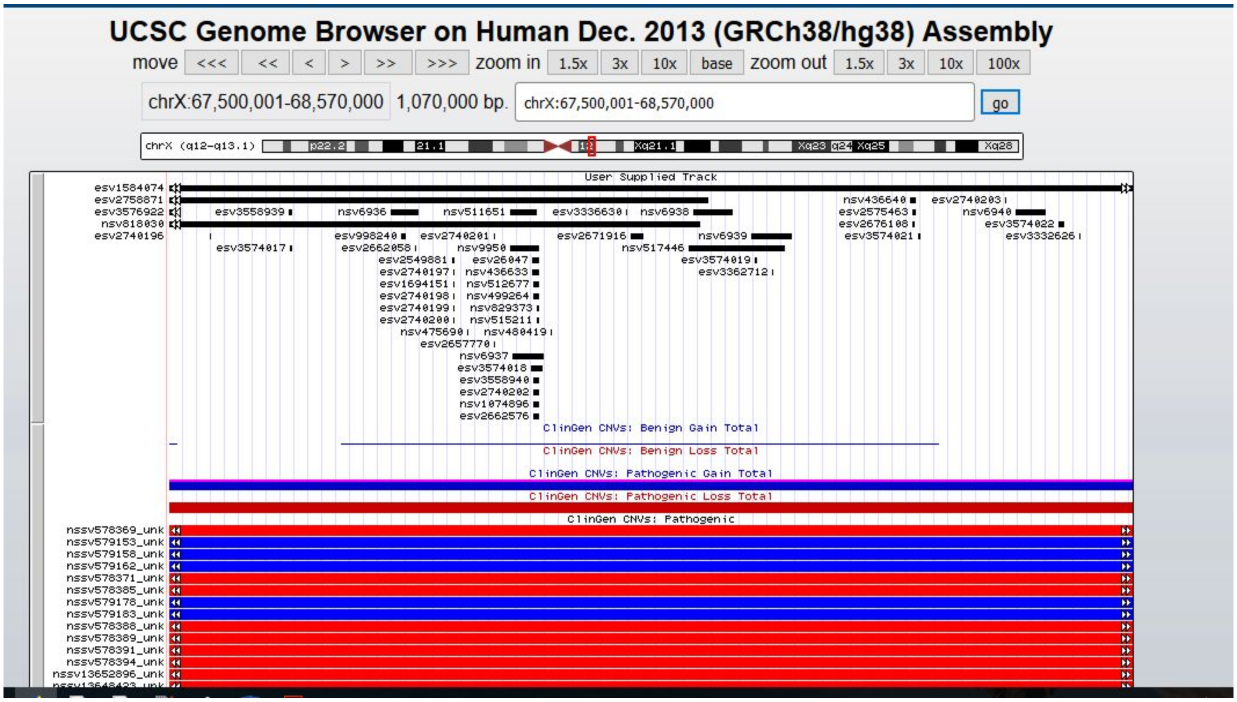


Figure 11. SNPs overlapping with the AR gene, shown in IGV.

In summary, the UCSC Main Table Browser gave the most results of SNP data (over 16,000), which included 11 common SNPs that were joined with exon boundaries to find the most significant known variations in this gene. These were mostly missense variations, with a few coding sequence synonymous variants, and three indels in the CDS. Of these 11, most of them were reported to be associated with Androgen Resistance Syndrome or PAIS. In BioMart, there were thousands of SNPs from COSMIC, which were submitted as associated with various types of tumors. The one SNP pulled from BioMart that was located in the dbSNP database (rs1057519864) is a missense variant, located on the forward strand. It overlaps with 4 of the AR transcripts and is associated with the development of prostate cancer.

CNVs in the AR Gene

CNVs in the AR gene were mapped in Galaxy using the DGV Struct Var track on the UCSC Table Browser. The DGV Struct Var track displays large structural variants such as CNVs, inversions and indels. Using the dvgMerged table, we found 49 CNV regions for the AR gene which was exported as the attached BED file CNVs_on_AR_gene_Galaxy_part_1. A partial view of this BED file is shown below. The black bold lines show the 49 CNV regions associated with the AR location chrX:67,500,001-68,570,000 (Figure 12).



UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chrX:67,500,001-68,570,000 1,070,000 bp.

chrX:67,500,001-68,570,000

Figure 12. Somatic Copy Number Variations in the region of chrX:67,500,001-68,570,000.

Using DGV Supp Var (dvg Supporting) gave 947 CNV regions saved as bed file CNVs_on_AR_galaxy_part2. Filtering was done to retrieve only CNVs supporting clinical evidence using ClinGen CNVs and ClinVar Variants as filters.

From the IGV server, using Cancer Genome Atlas and TCGA Broad GDAC track, we chose the Firehose Standard Data and then expand Broad Firehose Standard Data Run: 2016_01_28 for human genome assembly hg19. We selected prostate cancer primary tumor (PRAD-TP) as our cancer of choice, selected only somatic copy number variations (CNV minus germline), and selected chrX:67,500,001-68,570,000 as our chromosome view. A partial list of CNVs in this region shown below (Figure 13).

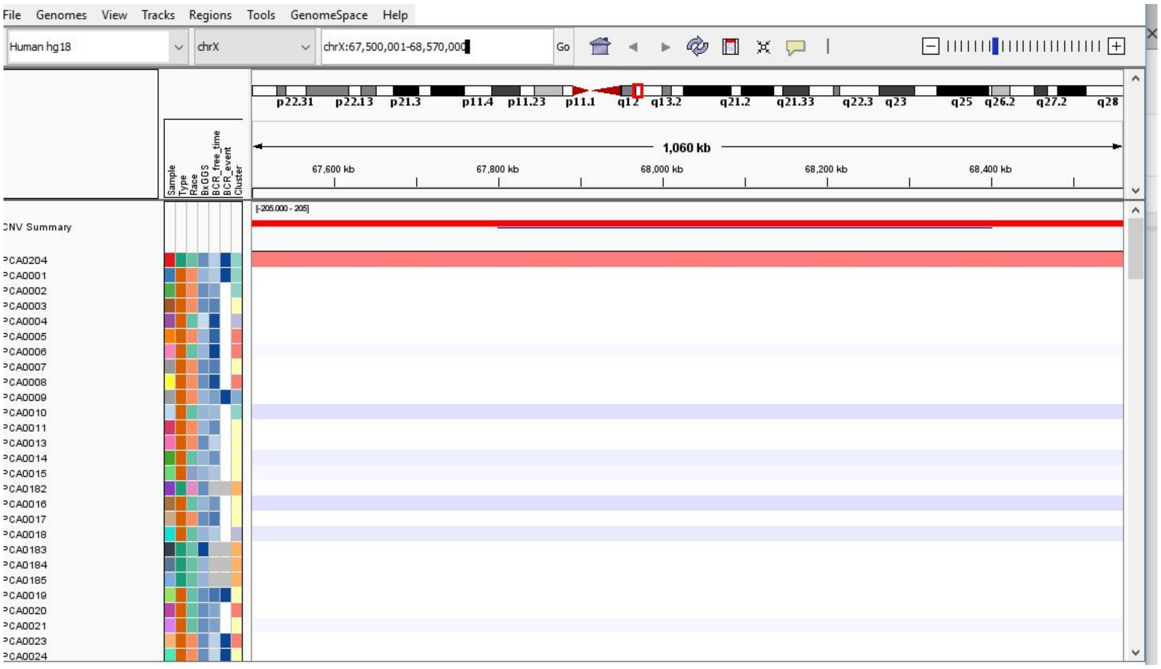


Figure 13. Copy number variations in region chrX:67,500,001-68,570,000, viewed in IGV in the TCGA Broad GDAC track with PRAD-TP dataset loaded.

Searching for the AR gene in IGV, we noticed that this region was slightly shifted on the genome at position chrX:66,678,599-66,869,186 as opposed to the location of chrX:67,544,623-67,730,619 in UCSC Genome Browser and Ensembl. Loading the TCGA prostate cancer dataset (PRAD-TP) for only somatic mutations (minus germline) overlapping with the AR gene, we found the following results. The primary tumor dataset did not show many significant copy number variations across samples. Only 2% of individuals showed a gain in copy number in the AR gene region and 1% showed a loss (Figure 14).

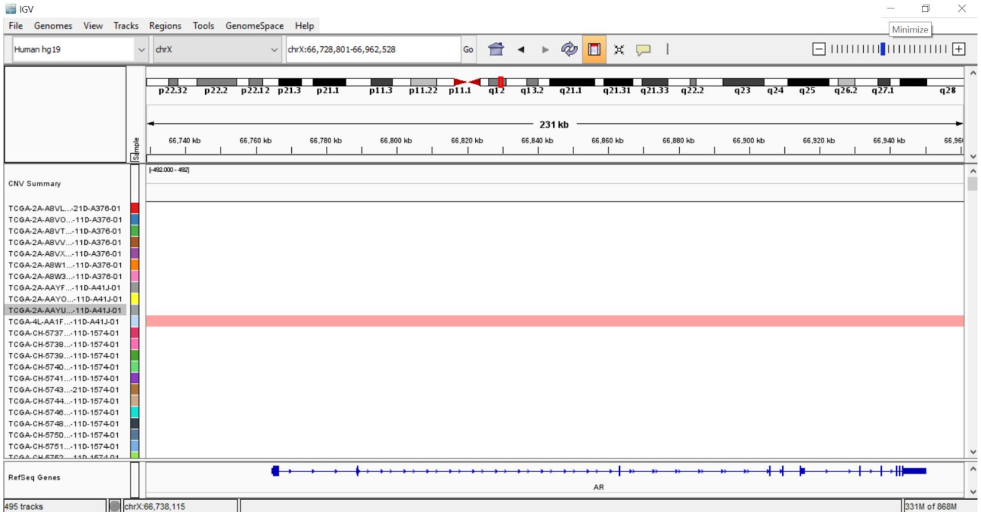
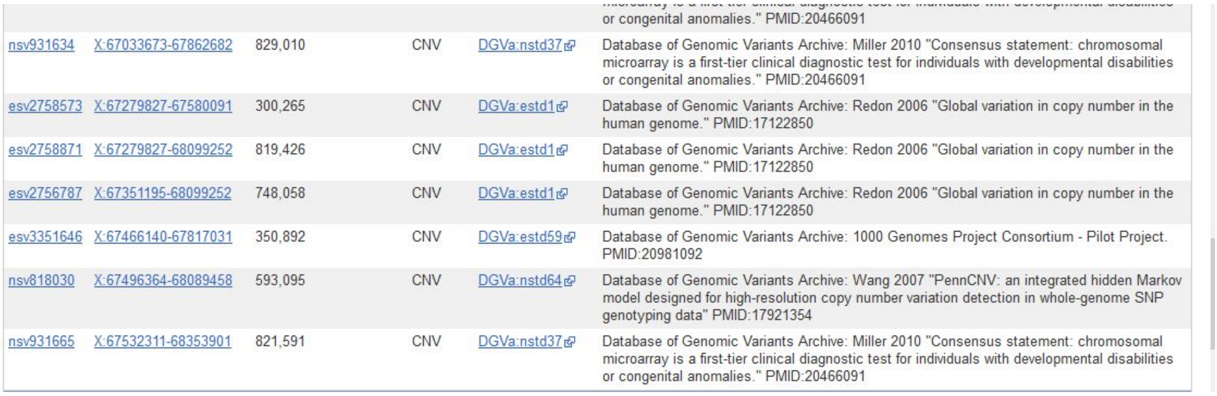

IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Minimize
Human hg19
chrX
chrX:66,728,801-66,962,528
Go
p22.32 p22.2 p22.12 p21.3 p21.1 p11.3 p11.22 p11.1 q12 q13.2 q21.1 q21.31 q21.33 q22.2 q23 q24 q25 q26.2 q27.1 q28
231 kb
66,740 kb
66,760 kb
66,780 kb
66,800 kb
66,820 kb
66,840 kb
66,860 kb
66,880 kb
66,900 kb
66,920 kb
66,940 kb
Sample
[-492.000 - 492]
CNV Summary
TCGA-2A-A8VL...-21D-A376-01
TCGA-2A-A8VO...-11D-A376-01
TCGA-2A-A8VT...-11D-A376-01
TCGA-2A-A8VV...-11D-A376-01
TCGA-2A-A8VX...-11D-A376-01
TCGA-2A-A8W1...-11D-A376-01
TCGA-2A-A8W3...-11D-A376-01
TCGA-2A-AAYF...-11D-A41J-01
TCGA-2A-AAYO...-11D-A41J-01
TCGA-2A-AAYU...-11D-A41J-01
TCGA-4L-AA1F...-11D-A41J-01
TCGA-CH-5737...-11D-1574-01
TCGA-CH-5738...-11D-1574-01
TCGA-CH-5739...-11D-1574-01
TCGA-CH-5740...-11D-1574-01
TCGA-CH-5741...-11D-1574-01
TCGA-CH-5743...-21D-1574-01
TCGA-CH-5744...-11D-1574-01
TCGA-CH-5746...-11D-1574-01
TCGA-CH-5748...-11D-1574-01
TCGA-CH-5750...-11D-1574-01
TCGA-CH-5751...-11D-1574-01
RefSeq Genes
AR
495 tracks
chrX:66,738,115
331M of 868M

Figure 14. Copy number variations in AR gene region, viewed in IGV in the TCGA Broad GDAC track with PRAD-TP dataset loaded (2% gain, 1% loss, overwhelmingly no variation).

Searching again using the metastatic prostate cancer dataset (PRAD-TM), no significant copy number results were found as this was a very limited dataset.

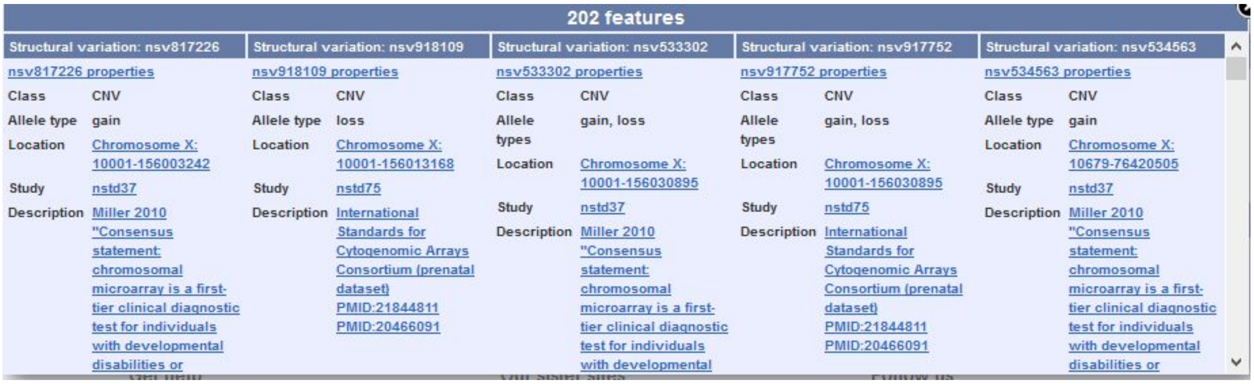
The ensembl Human(GRCh38.p10) dataset was queried in Biomart with the AR gene used as a filter to obtain structural variants (specifically CNVs). Some CNVs found are listed below. These structural variants were imported from the DGVa dataset (Figure 15).



					or congenital anomalies." PMID:20466091
nsv931634	X:67033673-67862682	829,010	CNV	DGVa:nstd37	Database of Genomic Variants Archive: Miller 2010 "Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies." PMID:20466091
esv2758573	X:67279827-67580091	300,265	CNV	DGVa:estd1	Database of Genomic Variants Archive: Redon 2006 "Global variation in copy number in the human genome." PMID:17122850
esv2758871	X:67279827-68099252	819,426	CNV	DGVa:estd1	Database of Genomic Variants Archive: Redon 2006 "Global variation in copy number in the human genome." PMID:17122850
esv2756787	X:67351195-68099252	748,058	CNV	DGVa:estd1	Database of Genomic Variants Archive: Redon 2006 "Global variation in copy number in the human genome." PMID:17122850
esv3351646	X:67466140-67817031	350,892	CNV	DGVa:estd59	Database of Genomic Variants Archive: 1000 Genomes Project Consortium - Pilot Project. PMID:20981092
nsv818030	X:67496364-68089458	593,095	CNV	DGVa:nstd64	Database of Genomic Variants Archive: Wang 2007 "PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data" PMID:17921354
nsv931665	X:67532311-68353901	821,591	CNV	DGVa:nstd37	Database of Genomic Variants Archive: Miller 2010 "Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies." PMID:20466091

Figure 15. Copy number variations in the AR gene region. Results from BioMart.

We then zoomed in on the Ensembl results for region chrX:67,084,528-68,063,864 to obtain the results below. 202 CNVs were found for the AR gene region highlighted (Figure 16). The deep purple line represents the CNVs of the AR gene for the chromosomal location. Clicking and zooming on it gives the details on all CNVs located in this region (Figure 17).



202 features

Structural variation: nsv817226
nsv817226 properties
Class CNV
Allele type gain
Location Chromosome X: 10001-156003242
Study nstd37
Description Miller 2010 "Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or

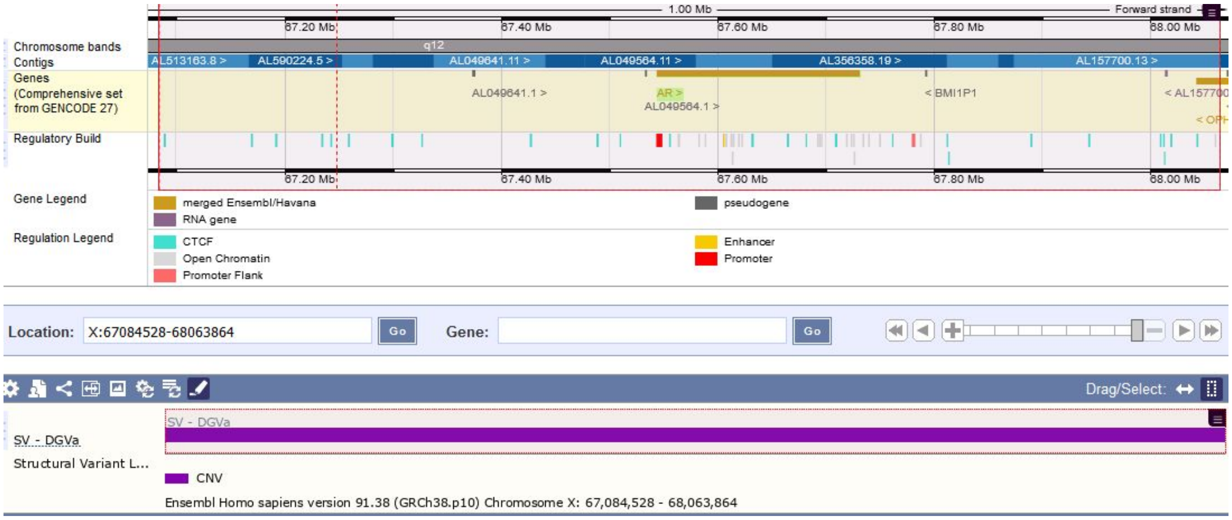
Structural variation: nsv918109
nsv918109 properties
Class CNV
Allele type loss
Location Chromosome X: 10001-156013168
Study nstd75
Description International Standards for Cytogenomic Arrays Consortium (prenatal dataset) PMID:21844811 PMID:20466091

Structural variation: nsv533302
nsv533302 properties
Class CNV
Allele types gain, loss
Location Chromosome X: 10001-156030895
Study nstd37
Description Miller 2010 "Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental

Structural variation: nsv917752
nsv917752 properties
Class CNV
Allele types gain, loss
Location Chromosome X: 10001-156030895
Study nstd75
Description International Standards for Cytogenomic Arrays Consortium (prenatal dataset) PMID:21844811 PMID:20466091

Structural variation: nsv534563
nsv534563 properties
Class CNV
Allele type gain
Location Chromosome X: 10679-76420505
Study nstd37
Description Miller 2010 "Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or

Figure 16. Subset of 202 copy number variations in the AR gene region.



Location: X:67084528-68063864

Ensembl Homo sapiens version 91.38 (GRCh38.p10) Chromosome X: 67,084,528 - 68,063,864

Figure 17. Zoomed in DGVs in AR gene region.

In Summary, the UCSC Main Table Browser gave more consistent results and covered our chromosome area of interest chrX:67,500,001-68,570,000 completely, ensuring that we locate all relevant CNVs associated with this region. The other browsers gave minimal results due to the limited nature of the dataset used in the case of IGV and the fact that it was difficult to generate data consistent with the coordinates of our gene region chrX:67,500,001-68,570,000 when using biomart.

References:

1. Bermúdez de la Vega JA, Fernández-Cancio M, Bernal S, Audí L. Complete androgen insensitivity syndrome associated with male gender identity or female precocious puberty in the same family. Sex Dev. 2015;9(2):75-9.
2. Brüggenwirth HT, Boehmer AL, Ramnarain S, Verleun-Mooijman MC, Satijn DP, Trapman J, Grootegoed JA, Brinkmann AO. Molecular analysis of the androgen-receptor gene in a family with receptor-positive partial androgen insensitivity: an unusual type of intronic mutation. Am J Hum Genet. 1997 Nov;61(5):1067-77.
3. Callewaert L, Van Tilborgh N, Claessens F. Interplay between two hormone-independent activation domains in the androgen receptor. Cancer Res 2006; 66: 543–53.
4. Cleutjens KB, van Eekelen CC, van der Korput HA, Brinkmann AO, Trapman J. Two androgen response regions cooperate in steroid hormone regulated activity of the prostate-specific antigen promoter. J Biol Chem. 1996 Mar 15;271(11):6379-88.
5. Cleutjens KB, van der Korput HA, van Eekelen CC, van Rooij HC, Faber PW, Trapman J. An androgen response element in a far upstream enhancer region is essential for high, androgen-regulated activity of the prostate-specific antigen promoter. Mol Endocrinol. 1997 Feb;11(2):148-61.
6. ClinVar [online]. NCBI: ClinVar; 2018 Feb 13 (cited 2018 Feb 21). Available from https://www.ncbi.nlm.nih.gov/clinvar/variation/427107/#clinical-assertions.
7. Cutress ML, Whitaker HC, Mills IG, Stewart M, Neal DE. Structural basis for the nuclear import of the human androgen receptor. J Cell Sci. 2008 Apr 1;121(Pt 7):957-68.
8. Dejager S, Bry-Gauillard H, Bruckert E, Eymard B, Salachas F, LeGuern E, Tardieu S, Chadarevian R, Giral P, Turpin G. A comprehensive endocrine description of Kennedy's disease revealing androgen insensitivity linked to CAG repeat length. J Clin Endocrinol Metab. 2002 Aug;87(8):3893-901.
9. Edwards A, Hammond HA, Jin L, Caskey CT, Chakraborty R. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics. 1992 Feb;12(2):241-53.
10. Gaddipati JP, McLeod DG, Heidenberg HB, Sesterhenn IA, Finger MJ, Moul JW, Srivastava S. Frequent detection of codon 877 mutation in the androgen receptor gene in advanced prostate cancers. Cancer Res. 54: 2861-2864, 1994.
11. Hardy DO, Scher HI, Bogenreider T, Sabbatini P, Zhang ZF, Nanus DM, Catterall JF. Androgen receptor CAG repeat lengths in prostate cancer: correlation with age of onset. J Clin Endocrinol Metab. 1996 Dec;81(12):4400-5.
12. La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy. Nature 352: 77-79, 1991.
13. Lin B, Ferguson C, White JT, Wang S, Vessella R, True LD, Hood L, Nelson PS. Prostate-localized and androgen-regulated expression of the membrane-bound serine protease TMPRSS2. Cancer Res. 1999 Sep 1;59(17):4180-4
14. Lu NZ1 Wardell SE, Burnstein KL, Defranco D, Fuller PJ, Giguere V, Hochberg RB, McKay L, Renoir JM, Weigel NL, Wilson EM, McDonnell DP, Cidlowski JA. International

Union of Pharmacology. LXV. The pharmacology and classification of the nuclear receptor superfamily: glucocorticoid, mineralocorticoid, progesterone, and androgen receptors. Pharmacol Rev. 2006 Dec;58(4):782-97.

15. Mangelsdorf DJ, Thummel C, Beato M, Herrlich P, Schütz G, Umesono K, Blumberg B, Kastner P, Mark M, Chambon P, Evans RM. The nuclear receptor superfamily: the second decade. Cell. 1995 Dec 15;83(6):835-9.
16. McPhaul MJ. Androgen receptor mutations and androgen insensitivity. Mol Cell Endocrinol. 2002 Dec 30;198(1-2):61-7.
17. McPhaul MJ, Marcelli M, Zoppi S, Wilson CM, Griffin JE, Wilson JD Mutations in the ligand-binding domain of the androgen receptor gene cluster in two regions of the gene. J. Clin. Invest. 90: 2097-2101, 1992.
18. Mou L, Gui Y. A novel variant of androgen receptor is associated with idiopathic azoospermia. Mol Med Rep. 2016 Oct; 14(4): 2915–2920.
19. Patrão MT, Silva EJ, Avellar MC. Androgens and the male reproductive tract: an overview of classical roles and current perspectives. Arq Bras Endocrinol Metabol. 2009 Nov;53(8):934-45.
20. Quigley CA, De Bellis A, Marschke KB, el-Awady MK, Wilson EM, French FS. Androgen receptor defects: historical, clinical, and molecular perspectives. Endocr Rev 1995; 16: 271–321.
21. Sai T, Seino S, Chang C, Trifiro M, Pinsky L, Mhatre A, Kaufman M, Lambert B, Trapman J, Brinkmann AO, Rosenfield RL, Liao S. An exonic point mutation of the androgen receptor gene in a family with complete androgen insensitivity. Am. J. Hum. Genet. 46: 1095-1100.
22. Sammarco I, Grimaldi P, Rossi P, Cappa M, Moretti C, Frajese G, Geremia R. Novel point mutation in the splice donor site of exon-intron junction 6 of the androgen receptor gene in a patient with partial androgen insensitivity syndrome. J Clin Endocrinol Metab. 2000 Sep;85(9):3256-61.
23. Sasaki M, Kaneuchi M, Sakuragi N, Fujimoto S, Carroll PR, Dahiya R. The polyglycine and polyglutamine repeats in the androgen receptor gene in Japanese and Caucasian populations. Biochem Biophys Res Commun 2003; 312: 1244–7.
24. Shaffer PL, Jivan A, Dollins DE, Claessens F, Gewirth DT. Structural basis of androgen receptor binding to selective androgen response elements. Proc Natl Acad Sci U S A 2004; 101: 4758–63.
25. Shao J, Hou J, Li B, Li D, Zhang N, Wang X. Different types of androgen receptor mutations in patients with complete androgen insensitivity syndrome. Intractable Rare Dis Res. 2015 Feb;4(1):54-9.
26. Tan MH, Li J, Xu HE, Melcher K, Yong EL. Androgen receptor: structure, role in prostate cancer and drug discovery. Acta Pharmacol Sin. 2015 Jan;36(1):3-23.
27. Vilchis F, Ramos L, Kofman-Alfaro S, Zenteno JC, Mendez JP, Chavez B. Extreme androgen resistance in a kindred with a novel insertion/deletion mutation in exon 5 of the androgen receptor gene. J. Hum. Genet. 48: 346-351, 2003.
28. https://genome.ucsc.edu/cgi-bin/hgTables