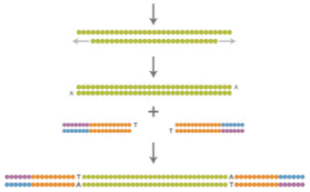
# Personal Genome Analysis

**Variant calling and Examples**

Fabio Navarro, Declan Clarke

Feb. 1 2017

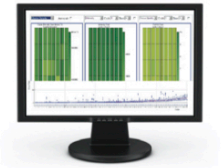
# Where do these reads come from?



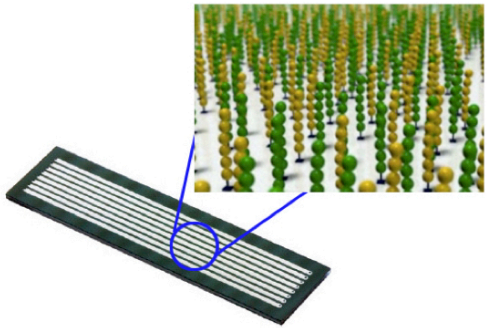
Library Preparation
~2 h [15 min hands-on (Nextera)]
< 6 h [< 3 h hands-on (TruSeq)]

Cluster Generation
~5 h (<10 min hands-on)

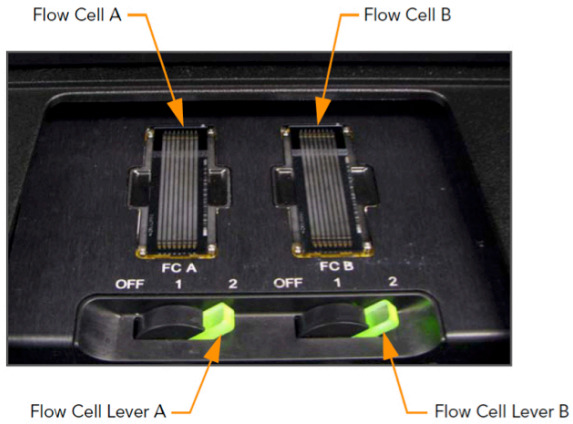
Sequencing by Synthesis
~1.5 to 11 days



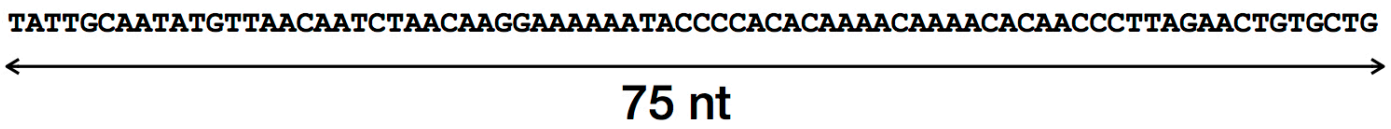
CASAVA
2 days (30 min hands-on)



Flow cell



# How long are the reads?




While there are other technologies that can give longer read lengths, Illumina reads are generally 50 nt - 250 nt

# What do I do with my sequencing reads?


Source: Slate via Noonan

# Genome Variation

TP53 Sequence:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

Single Nucleotide Polymorphism (SNP) – 1nt:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

T or A or C

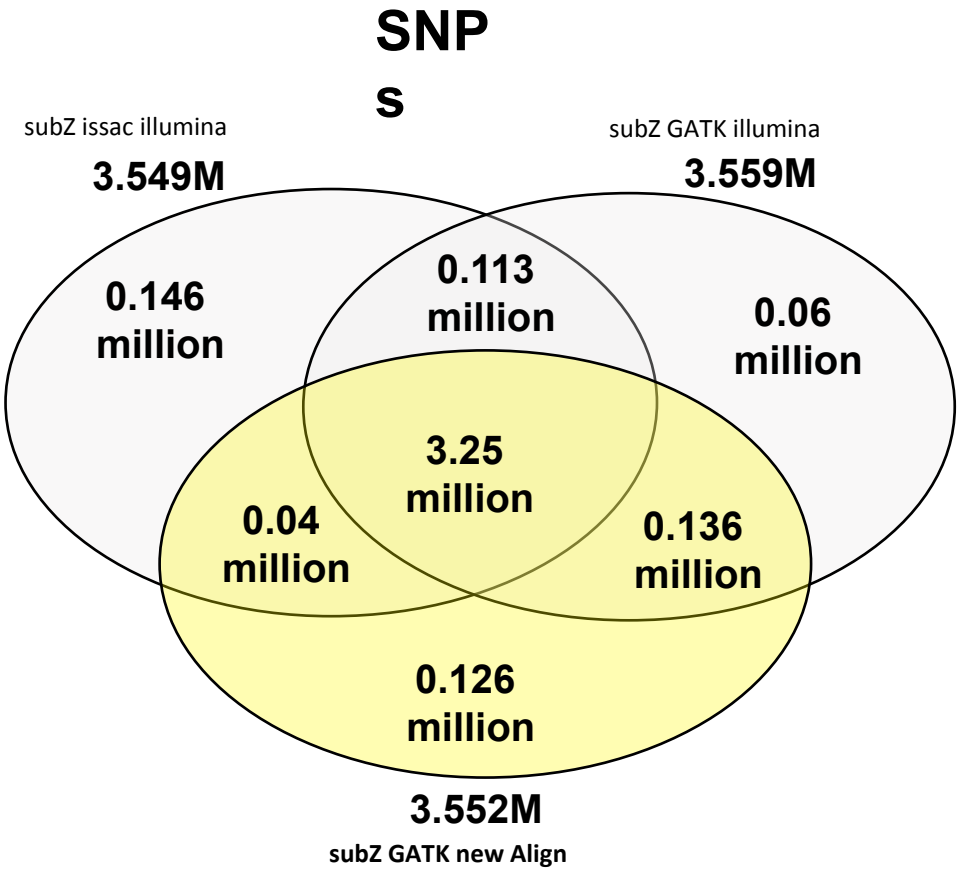
small INsertions and DELetions (INDEL) – 1-10nt:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

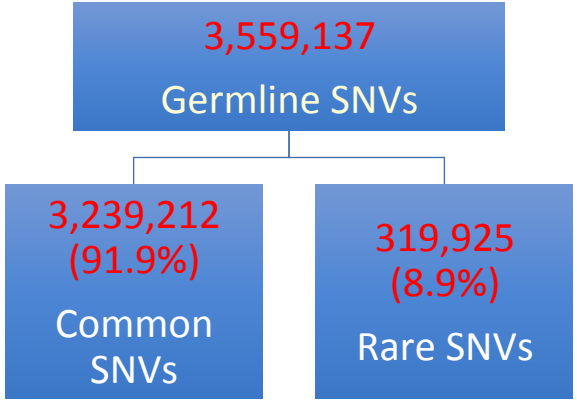
large Structural Variations (SV) - > 100nt:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

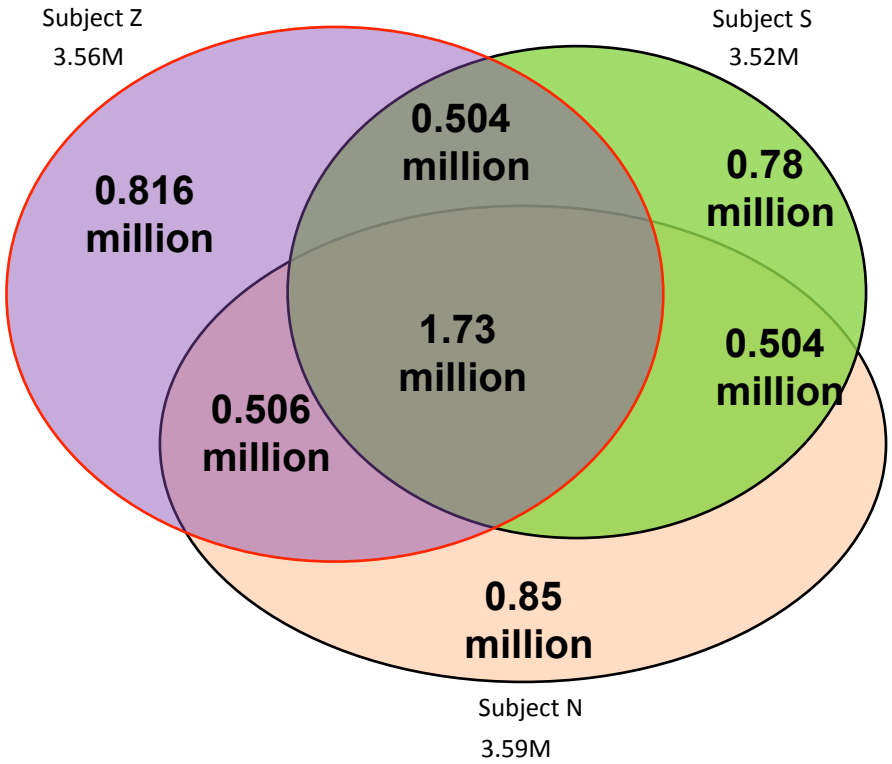
# Comparison of variant calls for subject Z




- Original approach (146K specific events):
  - Aligner: CASAVA; Variant Caller: Isaac
- Hybrid approach (60K specific events):
  - Aligner: CASAVA; Variant Caller: GATK
- Gold standard approach (126K specific events):
  - Aligner: BWA; Variant Caller: GATK



# Comparison of **SNPs** across three genomes



# Genome Variation

TP53 Sequence:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

Single Nucleotide Polymorphism (SNP) – 1nt:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

T or A or C

small INsertions and DELetions (INDEL) – 1-10nt:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

large Structural Variations (SV) - > 100nt:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

# Genome Variation

TP53 Sequence:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

Single Nucleotide Polymorphism (SNP) – 1nt:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

T or A or C

small INsertions and DELetions (INDEL) – 1-10nt:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

large Structural Variations (SV) - > 100nt:

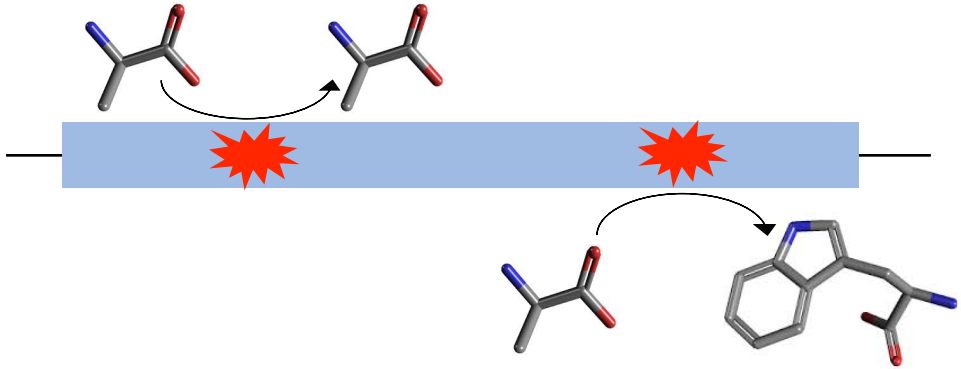
…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

# How to study & classify SNVs?

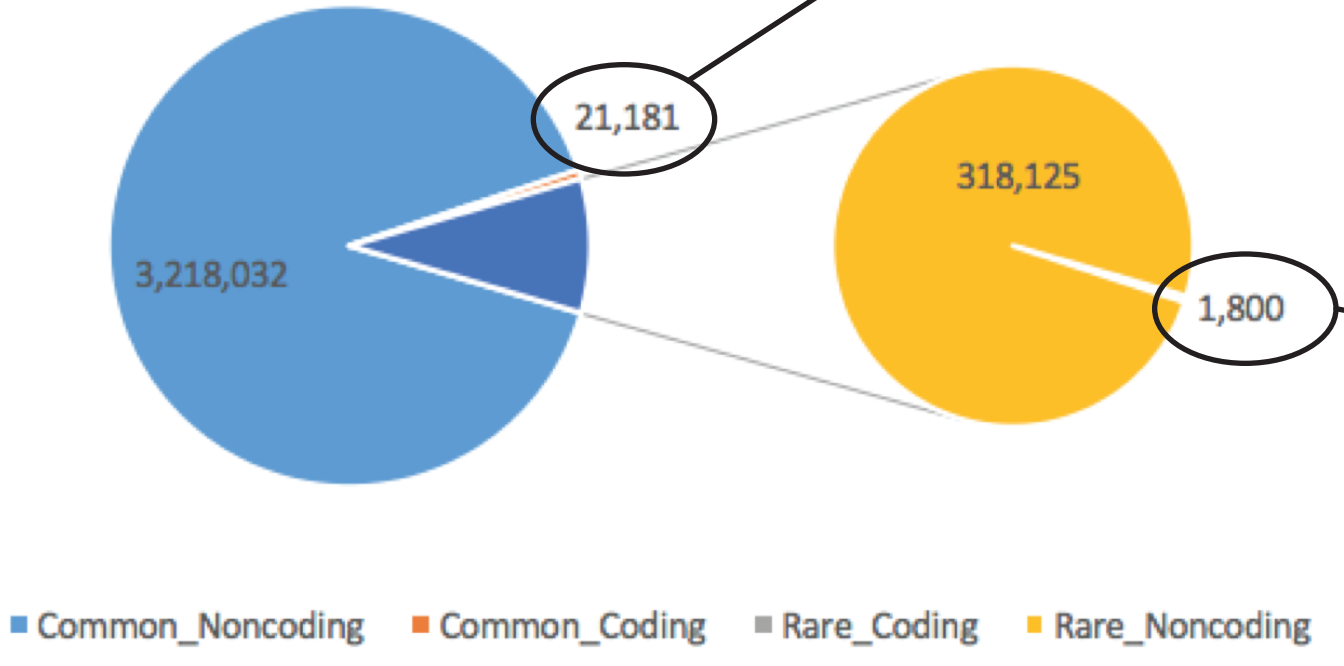
## Coding vs Noncoding

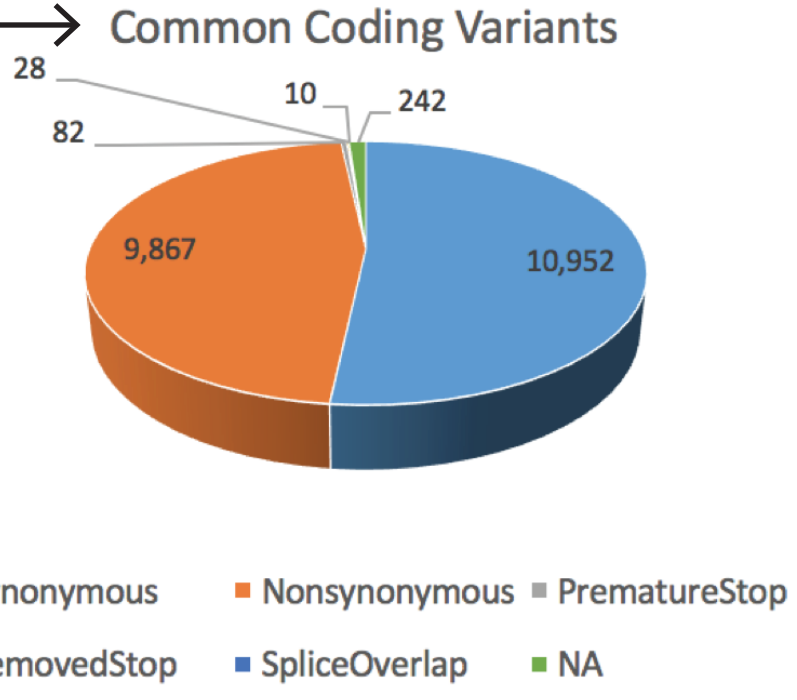

## Rare vs Common SNVs

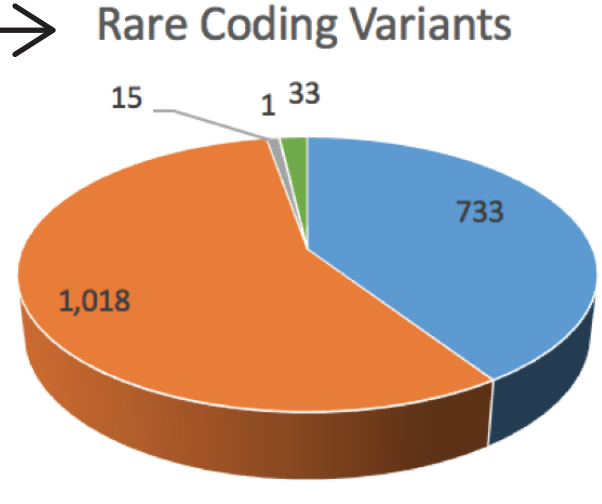
## Synonymous vs. nonynonymous SNVs



# Overview & Coding Variants



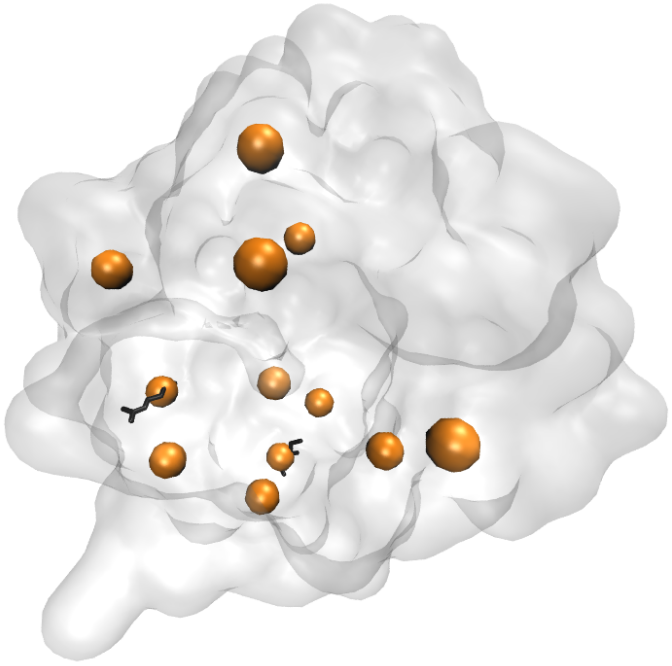
# Rare Non-synonymous Coding Variants

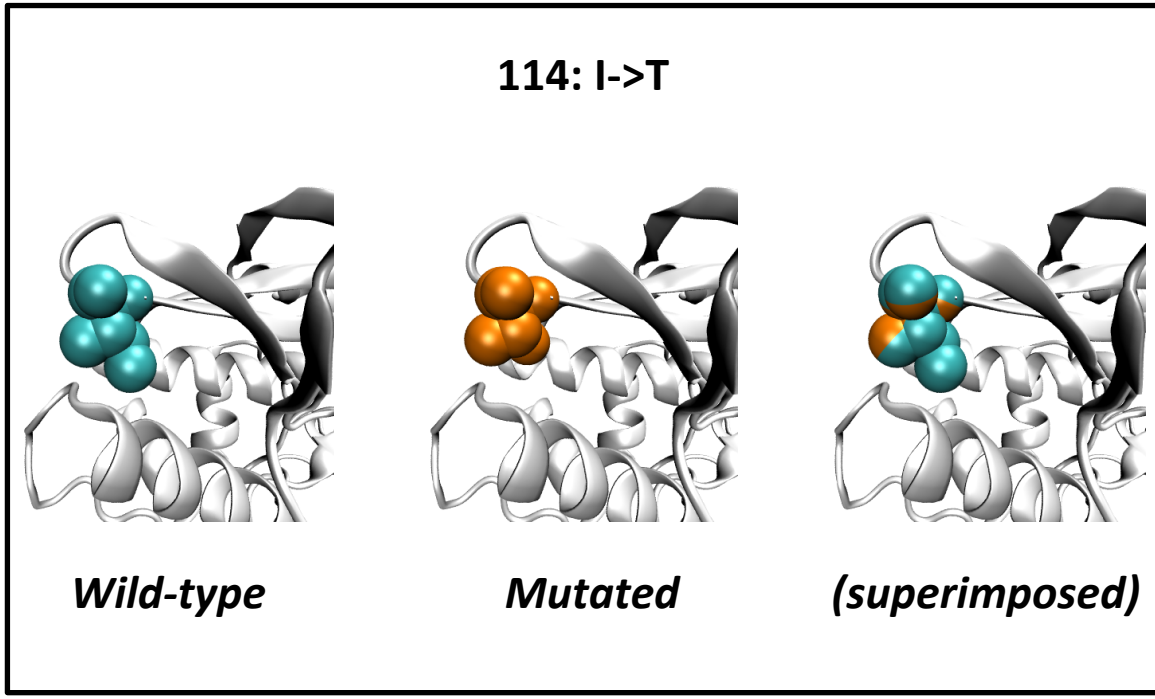
- 1018 SNVs -> 824 target genes

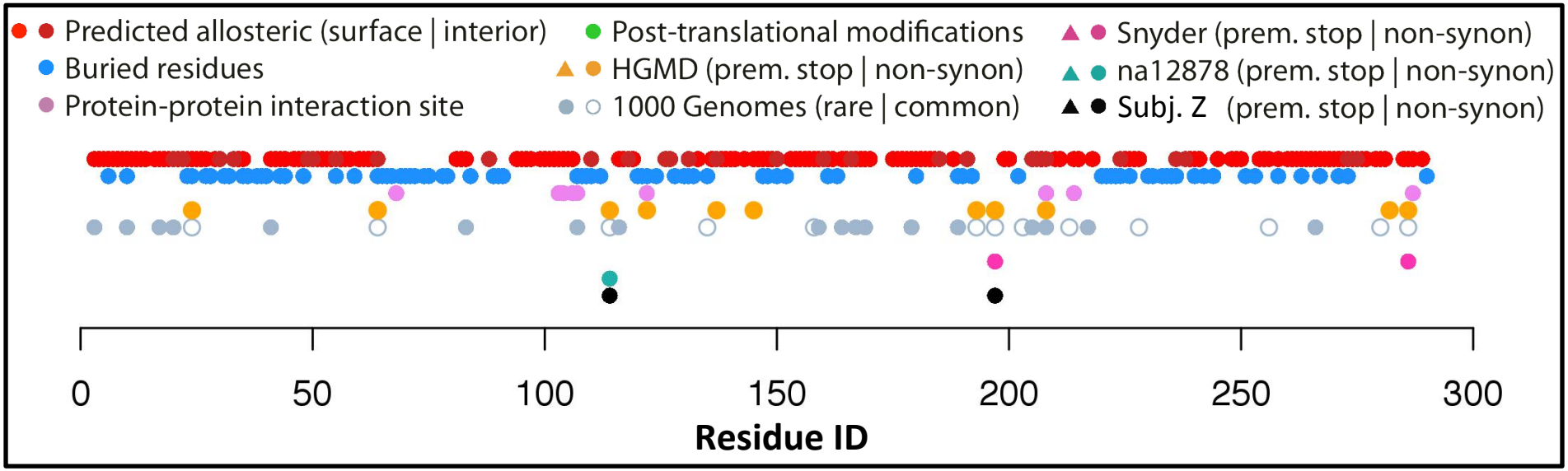
| Gene Annotation | Gene Name |
|---|---|
| Cancer-related | NOTCH2; PDE4DIP; TPR; CRTC3; CDH11; MLLT6; ASXL1; HMGA1; KDM6A |
| DNA repair | RECQL; RAD51; PPM1D; XRCC1; AP1B1; FANCI; PTPRH; RBBP7; SLX4; POLR2A; DCLRE1C; ANKLE1 |
| Cancer & DNA repair | ATM; PMS2; ERCC5 |
| Actionable Gene | ATM; KDM6A; INSR; FOXP4 |

- ATM: Serine/Threonine Kinase; Regulator of p53 and BRCA1; leukemia; ataxia-telangiectasia; breast cancer
- PMS2: Direct p53 effectors; mismatch repair cancer syndrome; colorectal cancer; hereditary nonpolyposis
- ERCC5: Chks in Checkpoint Regulation; DNA Repair; xeroderma pigmentosum
- KDM6A: Transcriptional misregulation in cancer
- INSR: Insulin Receptor; PI3K-Akt signaling pathway; GPCR Pathway; Diabetes mellitus
- FOXP4: Transcriptional repressor that represses lung-specific expression

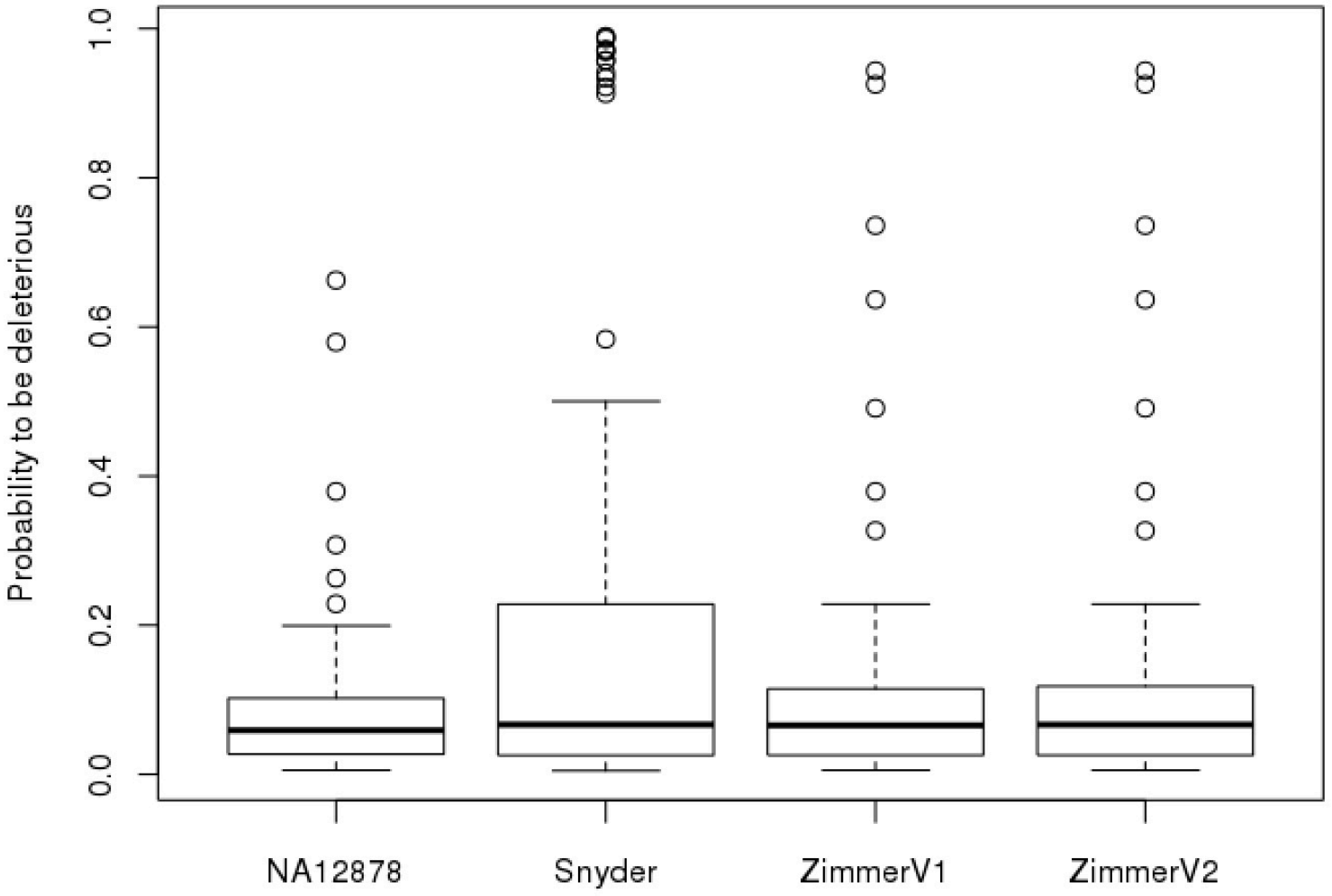
# Arylamine N-acetyltransferase (PDB: 2PFR_A ; gene: NAT2)



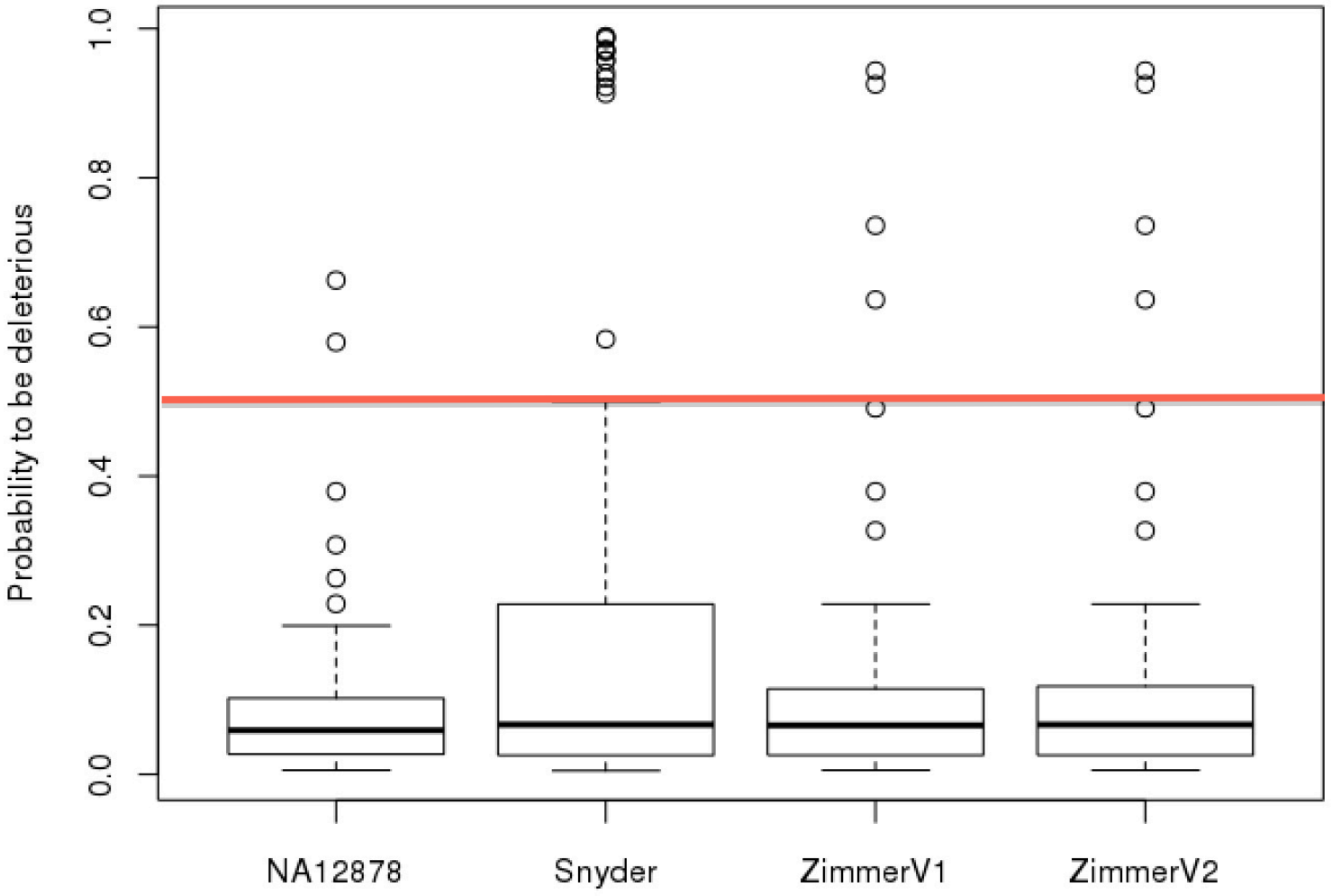


# LOF variants



# LOF variants



# LoF variants that are predicted to be the most deleterious (along with their associated genes)

## Subject Z

No disease associations in OMIM (but CCDC47 is associated with Schizophrenia)

| chr | pos | ref | alt | gene | Score | genotype | Gene function |
|---|---|---|---|---|---|---|---|
| 6 | 17606162 | C | T | FAM8A1 | 0.94365 | 0/1 | Unknown, Autism related ? Pubmed: 22495306 |
| 6 | 155577717 | T | A | TIAM2 | 0.63655 | 0/1 | Cell migration |
| 17 | 61829719 | A | C | CCDC47 | 0.92540 | 0/1 | unknown |
| 19 | 759925 | C | A | MISP | 0.73605 | 0/1 | Mitotic spindle positioning |

## Snyder

| chr | pos | ref | alt | gene | Score | genotype | OMIM |
|---|---|---|---|---|---|---|---|
| 2 | 44079970 | C | A | ABCG8 | 0.92190 | 0/1 | Sitosterolemia |
| 2 | 215854316 | T | A | ABCA12 | 0.97240 | 0/1 | Ichthyosis |
| 2 | 216240022 | G | T | FN1 | 0.98975 | 0/1 | fibronectin deficiency |
| 9 | 111718091 | G | T | CTNNAL1 | 0.98845 | 0/1 | |
| 9 | 130635074 | G | T | AK1 | 0.96915 | 0/1 | Hemolytic anemia |
| 10 | 29581479 | C | A | LYZL1 | 0.58365 | 0/1 | |
| 11 | 64056777 | C | A | GPR137 | 0.94075 | 0/1 | |
| 12 | 18800840 | G | T | PIK3C2G | 0.95735 | 0/1 | |
| 12 | 122400030 | C | A | WDR66 | 0.93380 | 0/1 | |
| 14 | 71570264 | C | A | PCNX | 0.98635 | 0/1 | |
| 15 | 68504073 | G | T | CLN6 | 0.97080 | 0/1 | Ceroid lipofuscinosiss |
| 15 | 93007504 | C | A | ST8SIA2 | 0.91290 | 0/1 | |
| 20 | 5157344 | C | A | CDS2 | 0.95755 | 0/1 | |

# Enrichment of genes affected by LoF SNVs in SubjectZ

Significant representation in **olfactory genes**!

## Categories Affected by ***Non-Synonymous*** SNVs

| Sublist | Category | Term | RT | Genes | Count | % | P-Value | Benjamini |
|---|---|---|---|---|---|---|---|---|
| | SP_PIR_KEYWORDS | polymorphism | RT | | 556 | 79.2 | 5.1E-32 | 2.5E-29 |
| | SP_PIR_KEYWORDS | alternative splicing | RT | | 338 | 48.1 | 3.8E-8 | 9.3E-6 |
| | GOTERM_BP_FAT | cellular component morphogenesis | RT | | 30 | 4.3 | 1.7E-4 | 6.6E-2 |
| | PIR_SUPERFAMILY | PIRSF003152:G protein-coupled olfactory receptor, class II | RT | | 26 | 3.7 | 2.8E-4 | 3.2E-2 |
| | PIR_SUPERFAMILY | PIRSF800006:rhodopsin-like G protein-coupled receptors | RT | | 41 | 5.8 | 3.1E-4 | 2.3E-2 |
| | GOTERM_BP_FAT | sensory perception of smell | RT | | 31 | 4.4 | 3.1E-4 | 9.7E-2 |
| | SP_PIR_KEYWORDS | coiled coil | RT | | 102 | 14.5 | 3.2E-4 | 1.9E-2 |
| | GOTERM_BP_FAT | cell morphogenesis | RT | | 27 | 3.8 | 3.7E-4 | 1.0E-1 |
| | GOTERM_BP_FAT | sensory perception of chemical stimulus | RT | | 33 | 4.7 | 4.0E-4 | 9.3E-2 |
| | SP_PIR_KEYWORDS | olfaction | RT | | 30 | 4.3 | 4.0E-4 | 2.1E-2 |
| | KEGG_PATHWAY | Olfactory transduction | RT | | 27 | 3.8 | 4.3E-4 | 2.8E-2 |
| | KEGG_PATHWAY | Antigen processing and presentation | RT | | 11 | 1.6 | 4.7E-4 | 2.1E-2 |
| | PIR_SUPERFAMILY | PIRSF005491:tumor associated protein MAGE | RT | | 6 | 0.9 | 5.2E-4 | 2.9E-2 |

## Categories Affected by ***Premature Stop*** SNVs

| Sublist | Category | Term | RT | Genes | Count | % | P-Value | Benjamini |
|---|---|---|---|---|---|---|---|---|
| | PIR_SUPERFAMILY | PIRSF800006:rhodopsin-like G protein-coupled receptors | RT | | 12 | 14.0 | 1.3E-5 | 3.2E-4 |
| | GOTERM_MF_FAT | olfactory receptor activity | RT | | 10 | 11.6 | 4.0E-5 | 5.3E-3 |
| | GOTERM_BP_FAT | sensory perception of smell | RT | | 10 | 11.6 | 5.4E-5 | 1.7E-2 |
| | SP_PIR_KEYWORDS | olfaction | RT | | 10 | 11.6 | 7.3E-5 | 9.2E-3 |
| | GOTERM_BP_FAT | sensory perception of chemical stimulus | RT | | 10 | 11.6 | 1.2E-4 | 1.9E-2 |
| | INTERPRO | Olfactory receptor | RT | | 10 | 11.6 | 1.4E-4 | 2.4E-2 |
| | PIR_SUPERFAMILY | PIRSF003152:G protein-coupled olfactory receptor, class II | RT | | 8 | 9.3 | 2.1E-4 | 2.5E-3 |
| | KEGG_PATHWAY | Olfactory transduction | RT | | 9 | 10.5 | 4.2E-4 | 1.7E-2 |
| | INTERPRO | GPCR, rhodopsin-like superfamily | RT | | 12 | 14.0 | 4.3E-4 | 3.7E-2 |
| | INTERPRO | 7TM GPCR, rhodopsin-like | RT | | 12 | 14.0 | 4.4E-4 | 2.5E-2 |
| | SP_PIR_KEYWORDS | g-protein coupled receptor | RT | | 12 | 14.0 | 8.2E-4 | 5.1E-2 |
| | SP_PIR_KEYWORDS | sensory transduction | RT | | 10 | 11.6 | 1.1E-3 | 4.7E-2 |
| | GOTERM_BP_FAT | sensory perception | RT | | 11 | 12.8 | 1.4E-3 | 1.4E-1 |
| | SP_PIR_KEYWORDS | transducer | RT | | 12 | 14.0 | 1.4E-3 | 4.4E-2 |
| | GOTERM_BP_FAT | G-protein coupled receptor protein signaling pathway | RT | | 13 | 15.1 | 1.6E-3 | 1.2E-1 |

# Genome Variation

TP53 Sequence:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

Single Nucleotide Polymorphism (SNP) – 1nt:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

T or A or C

small INsertions and DELetions (INDEL) – 1-10nt:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

large Structural Variations (SV) - > 100nt:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

# Genome Variation

TP53 Sequence:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

Single Nucleotide Polymorphism (SNP) – 1nt:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

T or A or C

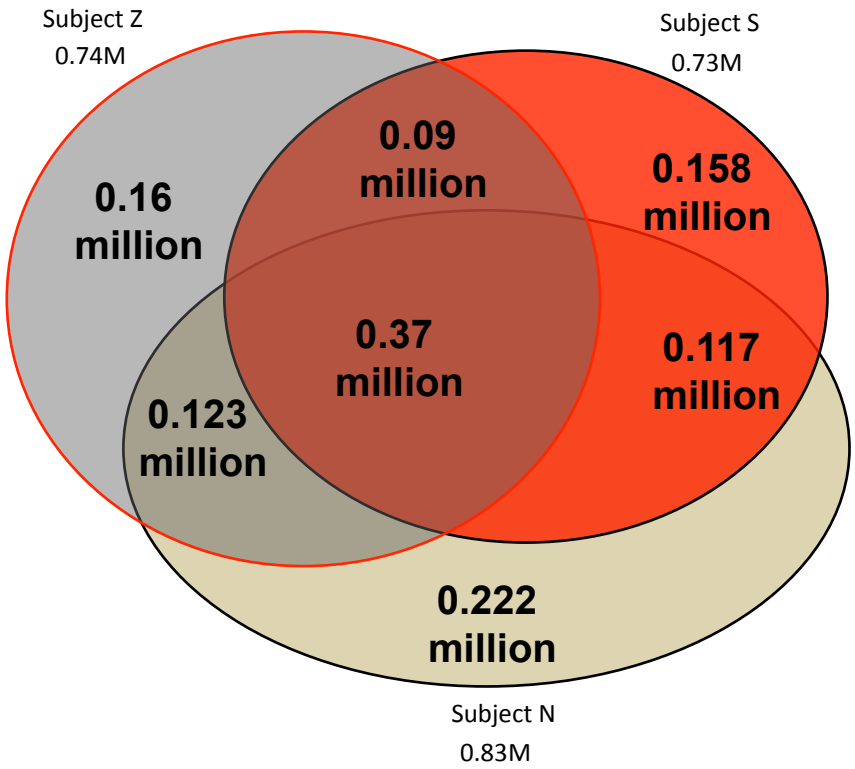
small INsertions and DELetions (INDEL) – 1-10nt:

…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

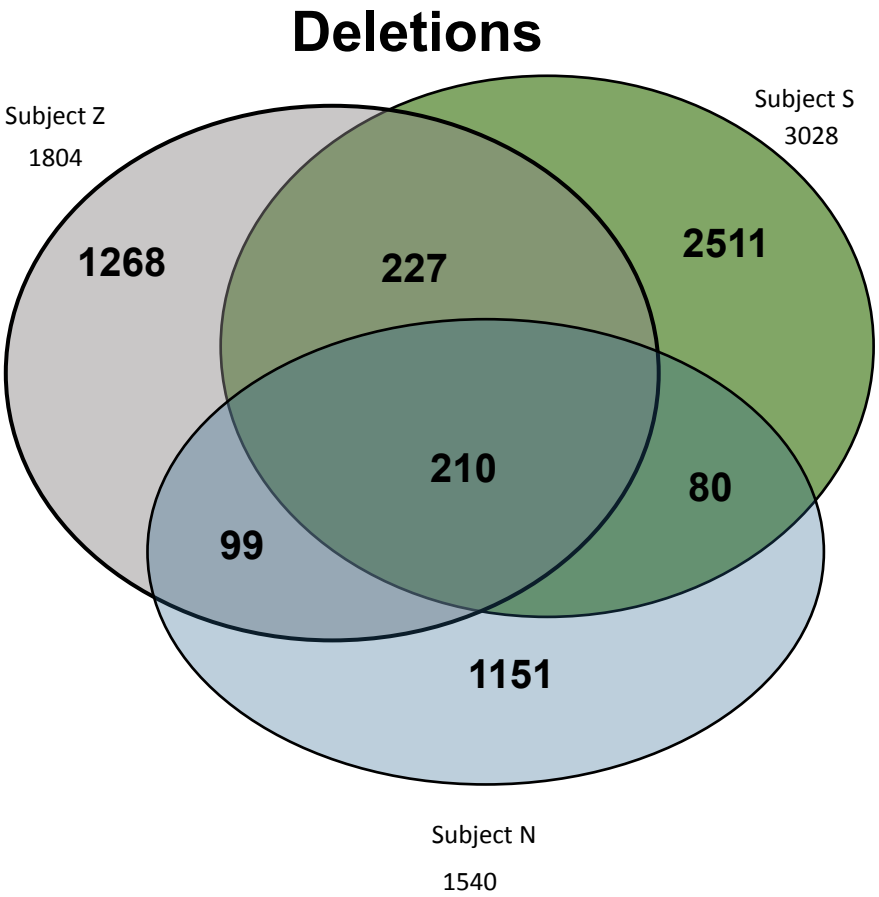
large Structural Variations (SV) - > 100nt:

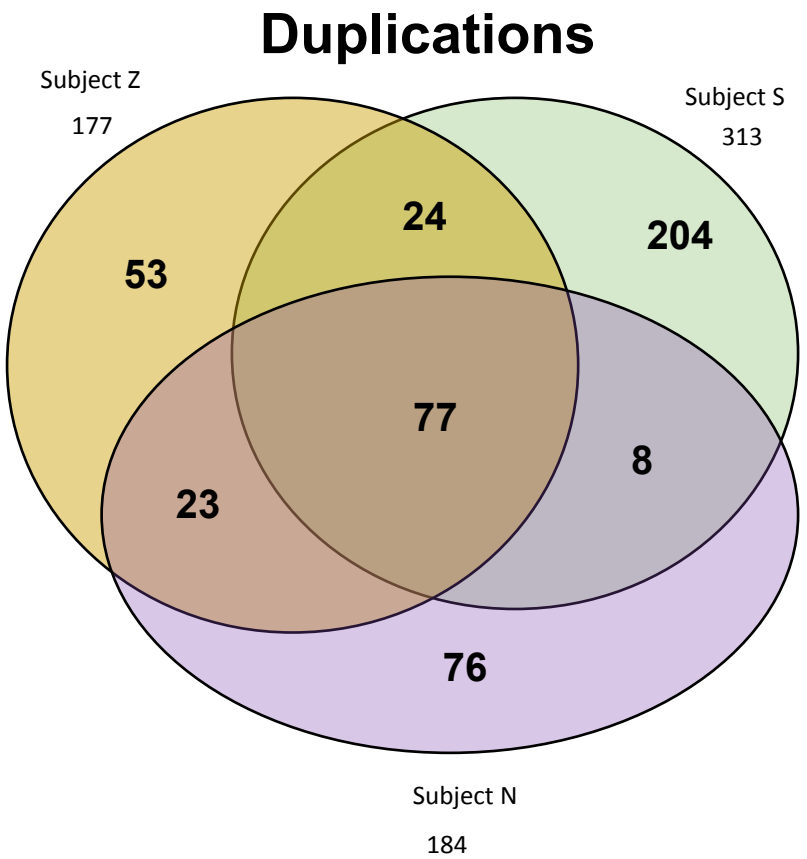
…GGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTT…

# Comparison of **INDELs** across three genomes



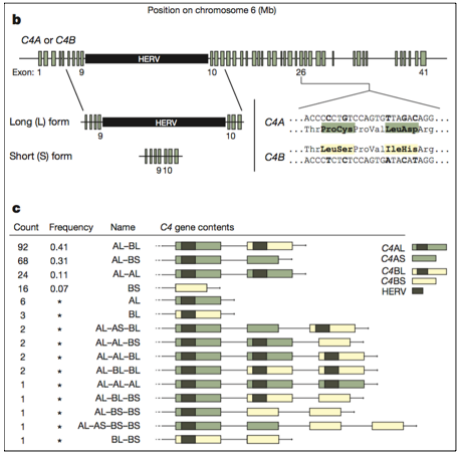
# Consensus **Structural Variations** across three genomes

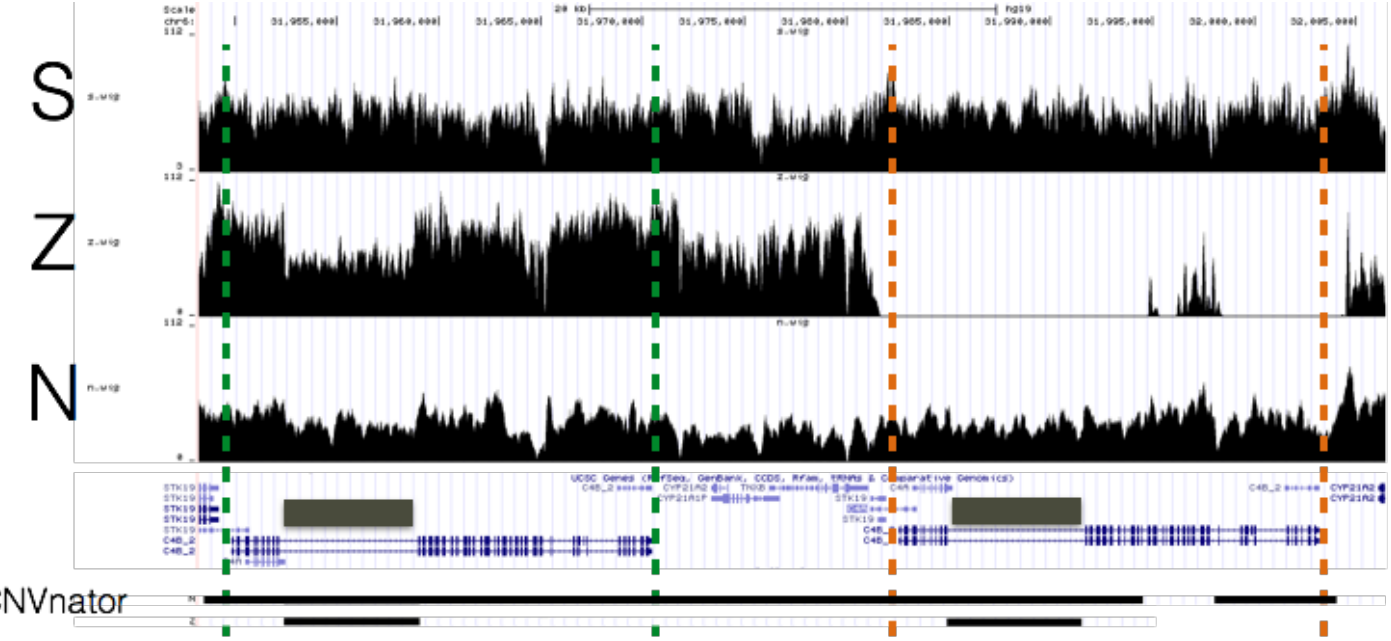


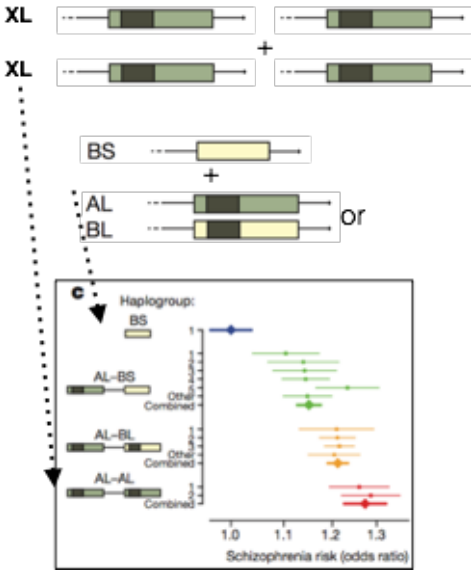



# Schizophrenia risk from complex variation of complement component 4

Aswin Sekar[1,2,3], Allison R. Bialas[4,5], Heather de Rivera[1,2], Avery Davis[1,2], Timothy R. Hammond[4], Nolan Kamitaki[1,2], Katherine Tooley[1,2], Jessy Presumey[5], Matthew Baum[1,2,3,4], Vanessa Van Doren[1], Giulio Genovese[1,2], Samuel A. Rose[2], Robert E. Handsaker[1,2], Schizophrenia Working Group of the Psychiatric Genomics Consortium*, Mark J. Daly[2,6], Michael C. Carroll[5], Beth Stevens[2,4] & Steven A. McCarroll[1,2]



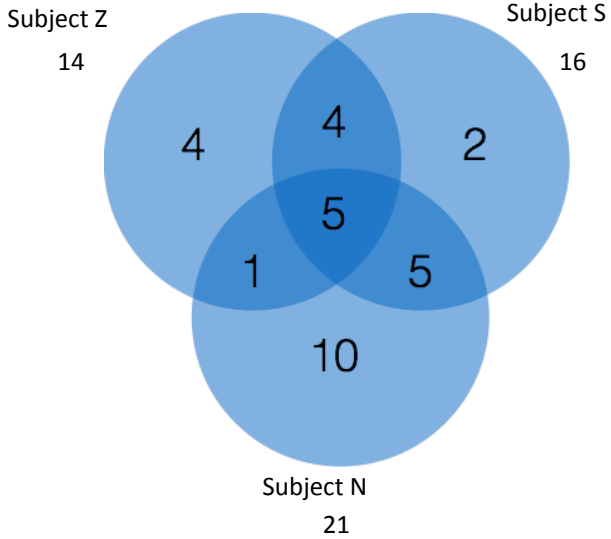



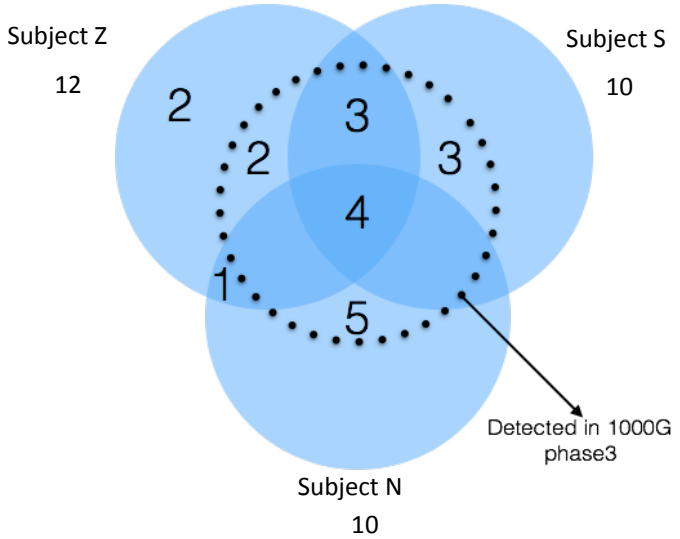
# Processed Pseudogene Copy Number Variation

| | Pseudogenes | Processed pseudogenes | Human specific processed pseudogenes |
|---|---|---|---|
| Human | ~14,000 | 7,831 | 127 |

## Pseudogene absence

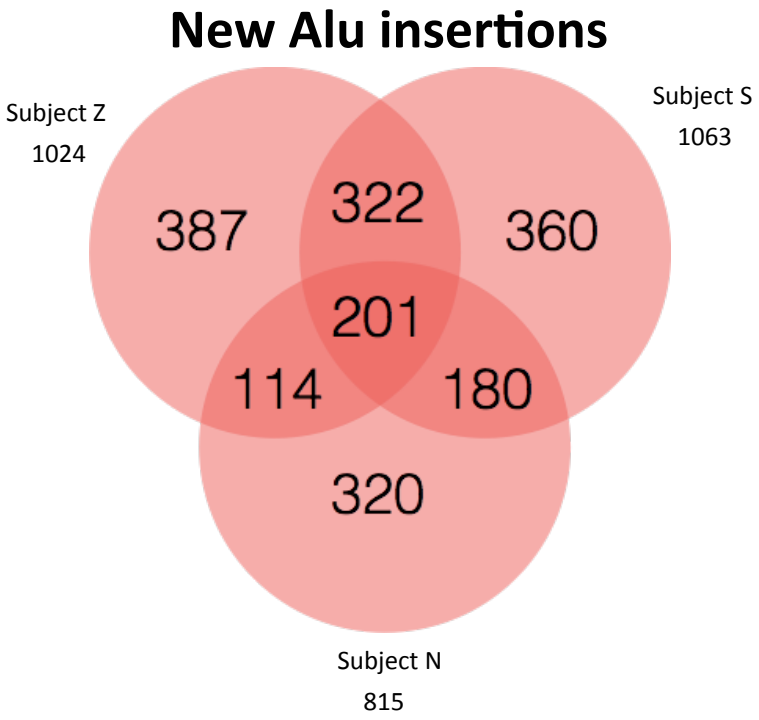



## Pseudogene insertion



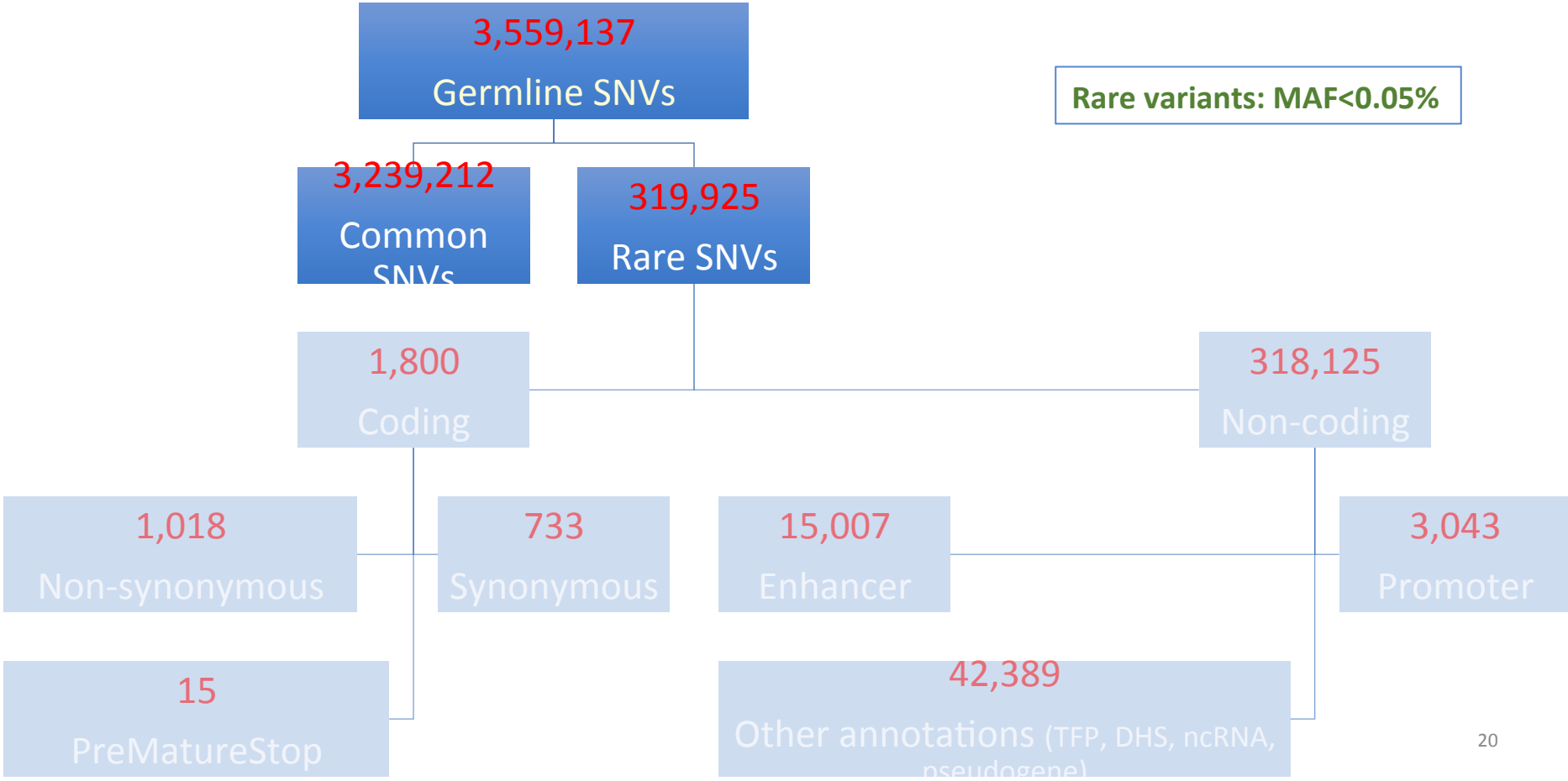
# ALU variation

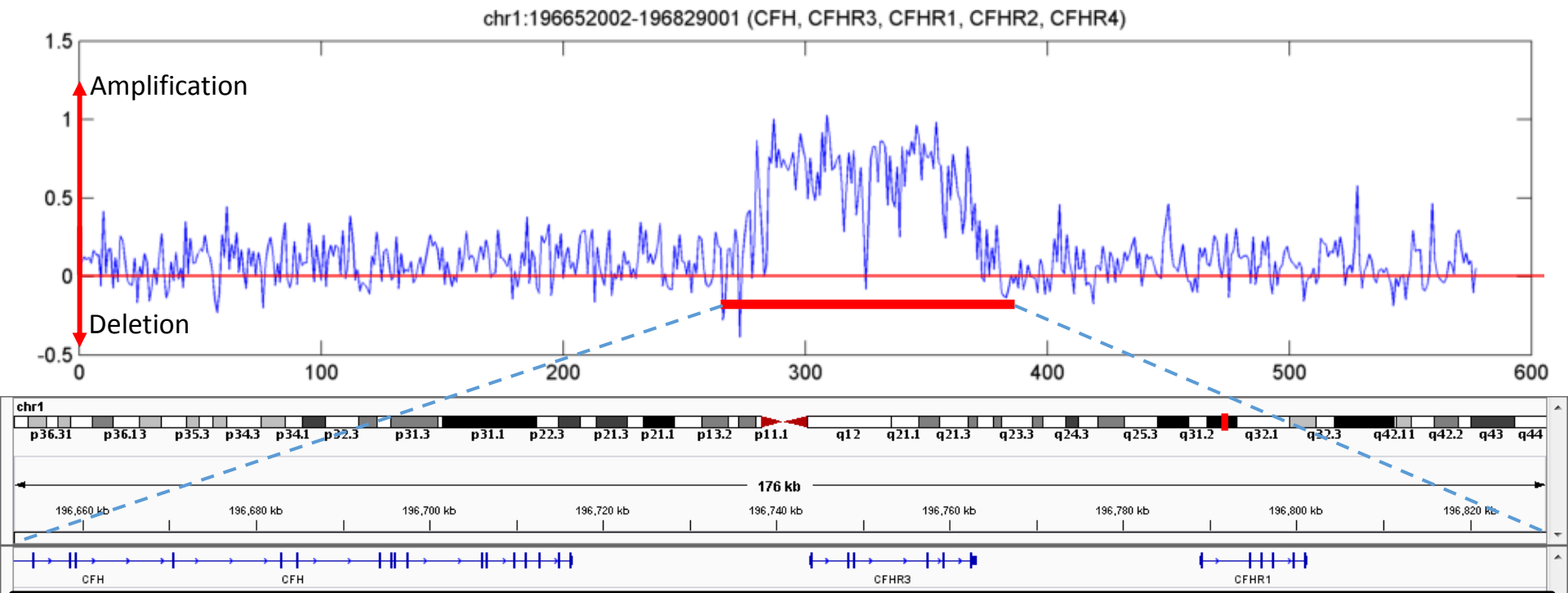
| | # of Alu in the genome | AluY |
|---|---|---|
| **Human** | 1,238,995 | 146,308 |



# Supplementary Slides

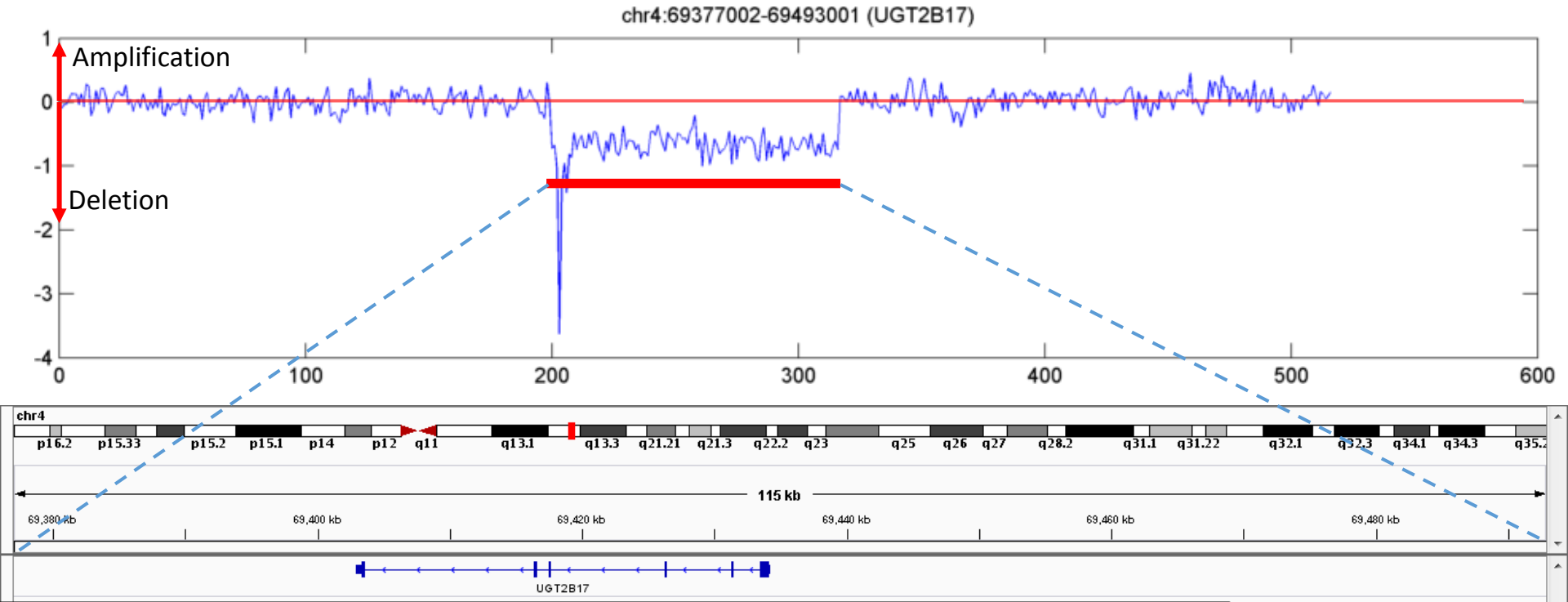
# Subject Z - SNPs frequency






# A common *CFH* haplotype, with deletion of *CFHR1* and *CFHR3*, is associated with lower risk of age-related macular degeneration

Anne E Hughes[1], Nick Orr[1], Hossein Esfandiary[1], Martha Diaz-Torres[2], Timothy Goodship[2] & Usha Chakravarthy[3]

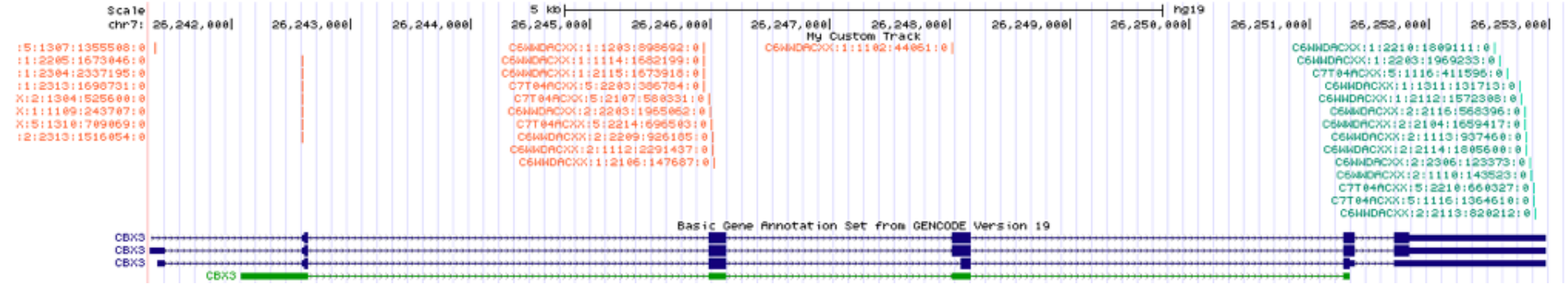



# Deletion Polymorphism of UDP-Glucuronosyltransferase 2B17 and Risk of Prostate Cancer in African American and Caucasian Men

can American controls, respectively. When all subjects were considered, a significant association was found between the *UGT2B17* deletion polymorphism and prostate cancer risk

# Pseudogene CNV – Example I

## CBX3 Parental gene




## CBX3 Insertion Point