

# INTRODUCTION TO VARIANT ANALYSIS USING 'EXOME SEQUENCING'

Bioinformàtica per a la Recerca Biomèdica

Alex Sánchez Pla

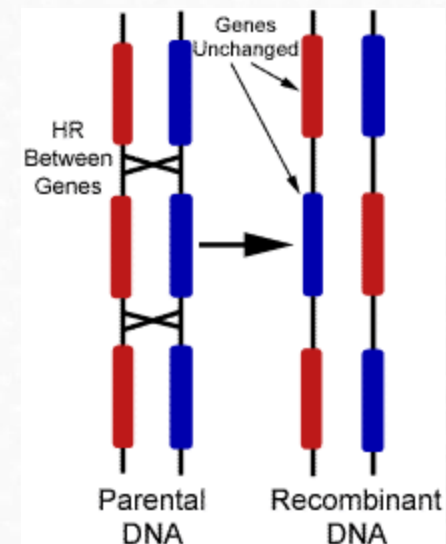
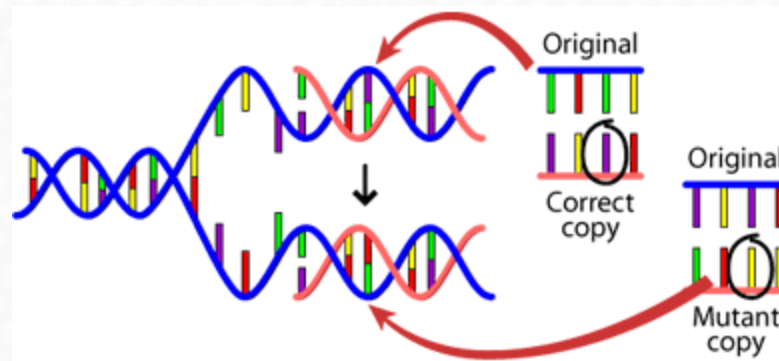
[alex.sanchez@vhir.org](mailto:alex.sanchez@vhir.org)

3/12/2018

- 1.Introduction to human variation**
- 2.Data pre-processing**
- 3.Variant Calling, Filtering and Annotation**
- 4.Downstream Analyses**

# Introduction to human variation

- Genetic variation: is the difference in DNA sequences between individuals.
- Variations occur in
  - Germ cells
  - Somatic cells
- Mutations and recombinations are major sources of variation



# Variants, alleles, haplotypes

- The term **variant** is used to refer *to a specific region of the genome which differs between two genomes*.
- The term **reference allele** refers to the base that is found in the reference genome
  - The reference is just somebody's genome not always the major allele.
  - The **alternative allele** refers to any base, other than the reference
- Alleles at variants close together on the same chromosome tend to occur together more often than is expected by chance. These blocks of alleles are called **haplotypes**.



# Types of genetic variation

- Genetic variation is commonly divided in three forms
  - Single Nucleotide Polymorphism or SNPs
  - Insertion or deletions, also called “indels”
  - Structural variation
    - Copy number variation
    - Chromosomal rearrangement events
- All forms of variations are related with disease but we eill mostly focus on SNPs

# Single nucleotide polymorphisms

- SNPs result from a substitution of a single base-pair

Human 1		
← 33M letters	AGGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTCGA AGGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTCGA	147M letters →
Human 2		
← 33M letters	ACGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTCGA AGGAAAACACGGAGTTGATGCA C AAGCCCCAACATCCAACCTCGA	147M letters →
Human 3		
← 33M letters	AGGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTCGA AGGAAAACACGGAGTTGATGCA C AAGCCCCAACATCCAACCTCGA	147M letters →
Human 4		
← 33M letters	AGGAAAACACGGAGTTGATGCA C AAGCCCCAACATCCAACCTCGA AGGAAAACACGGAGTTGATGCA C AAGCCCCAACATCCAACCTCGA	147M letters →
Human 5		
← 33M letters	ACGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTCGA AGGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTTGA	147M letters →

# Single nucleotide polymorphisms

<b>Human 1</b>		
← 33M letters	AGGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTCGA AGGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTCGA	147M letters →
<b>Human 2</b>		
← 33M letters	ACGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTCGA AGGAAAACACGGAGTTGATGCA C AAGCCCCAACATCCAACCTCGA	147M letters →
<b>Human 3</b>		
← 33M letters	AGGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTCGA AGGAAAACACGGAGTTGATGCA C AAGCCCCAACATCCAACCTCGA	147M letters →
<b>Human 4</b>		
← 33M letters	AGGAAAACACGGAGTTGATGCA C AAGCCCCAACATCCAACCTCGA AGGAAAACACGGAGTTGATGCA C AAGCCCCAACATCCAACCTCGA	147M letters →
<b>Human 5</b>		
← 33M letters	ACGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTCGA AGGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTTGA	147M letters →
<b>SNP1</b> <b>SNP2</b> <b>SNP3?</b>		

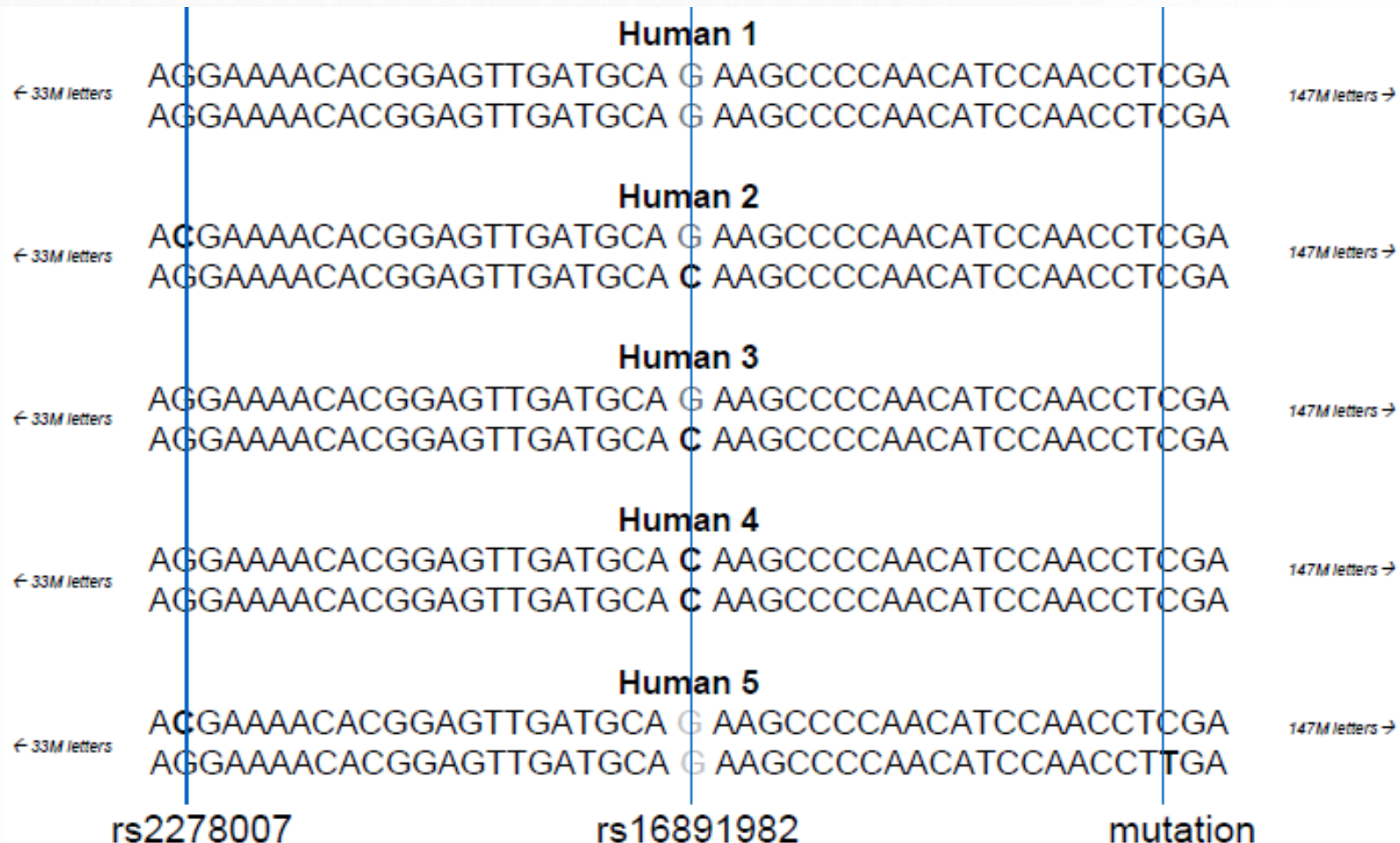
# Single nucleotide polymorphisms

<b>Human 1</b>		
← 33M letters	AGGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTCGA AGGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTCGA	147M letters →
<b>Human 2</b>		
← 33M letters	ACGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTCGA AGGAAAACACGGAGTTGATGCA C AAGCCCCAACATCCAACCTCGA	147M letters →
<b>Human 3</b>		
← 33M letters	AGGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTCGA AGGAAAACACGGAGTTGATGCA C AAGCCCCAACATCCAACCTCGA	147M letters →
<b>Human 4</b>		
← 33M letters	AGGAAAACACGGAGTTGATGCA C AAGCCCCAACATCCAACCTCGA AGGAAAACACGGAGTTGATGCA C AAGCCCCAACATCCAACCTCGA	147M letters →
<b>Human 5</b>		
← 33M letters	ACGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTCGA AGGAAAACACGGAGTTGATGCA G AAGCCCCAACATCCAACCTTGA	147M letters →
<b>SNP1 (1%)</b>		
<b>SNP2 (50%)</b>		
<b>SNP3? (0.1%)</b>		



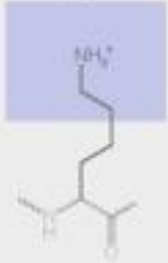
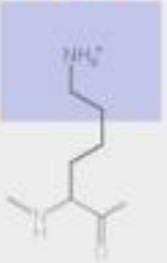
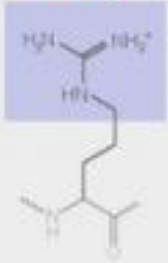
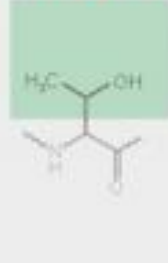
# Single nucleotide polymorphisms

- SNPs: One nucleotide difference occurring in at least 1% of population



# Variants in coding regions

- If a variant falls within a coding region, it can be categorised based on how it would affect the codon it falls within

	Point mutations				
	No mutation	Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	Lys	Lys	STOP	Arg	Thr
					
	basic			polar	

# Insertions/deletions (Frameshift mutations)

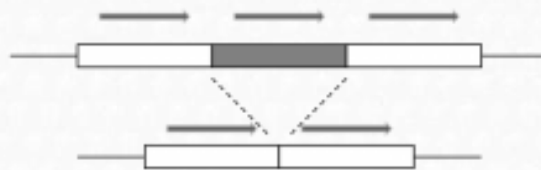
- Insertion or deletion of a single stretch of DNA sequence can range from two to hundreds of base-pairs in length.
- Indels may result in **frameshift mutations**

<b>Normal</b>							
mRNA	AUG	GGG	GCC	AAA	AGU	UAG	UUUG...
polypeptide	Met	Gly	Ala	Lys	Ser	Stop	
<b>Insertion</b>					+U		
					↓		
mRNA	AUG	GGC	GCC	AAA	UAG	UUAGUUUG...	
polypeptide	Met	Gly	Ala	Lys	Stop		
<b>Deletion</b>			-G				
			↓				
mRNA	AUG	GGC	CCA	AAA	GUU	AGU	UUG
polypeptide	Met	Gly	Pro	Lys	Val	Ser	Leu
			Random				

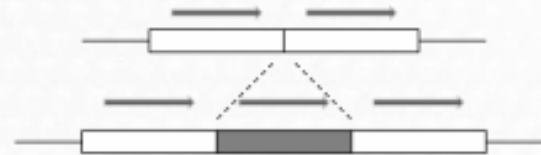
# Structural variation

- Genetic variation that occurs over a “larger” DNA sequence.

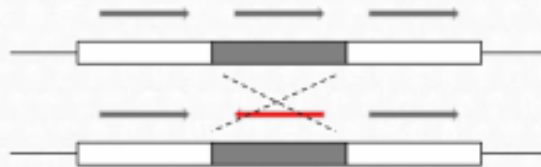
## Structural Variation



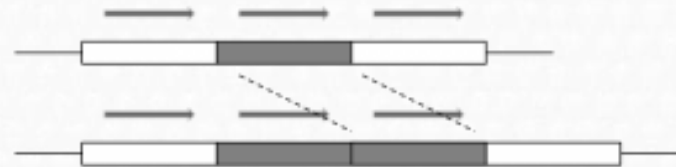
Deletion



Insertion



Inversion



Duplication

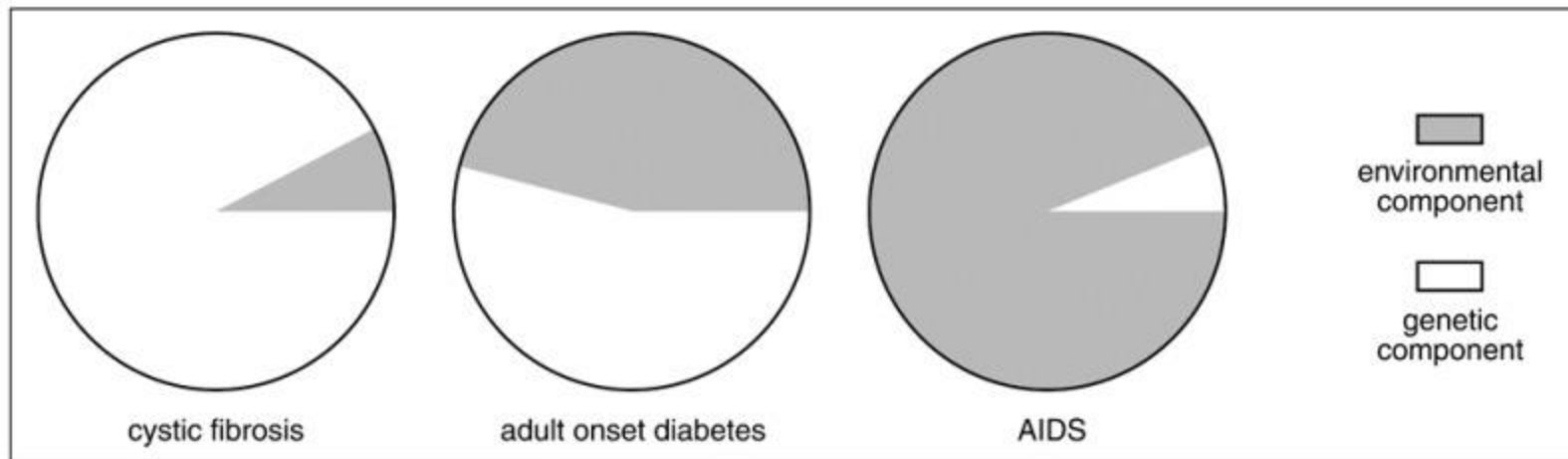


Copy Number Variation



# Effect of sequence variation

- Overall genetic variation may produce distinct effects ranging from innocuous to lethal.
- In many cases the effect will be changes in one or more proteins that will possibly affect the individual's health.

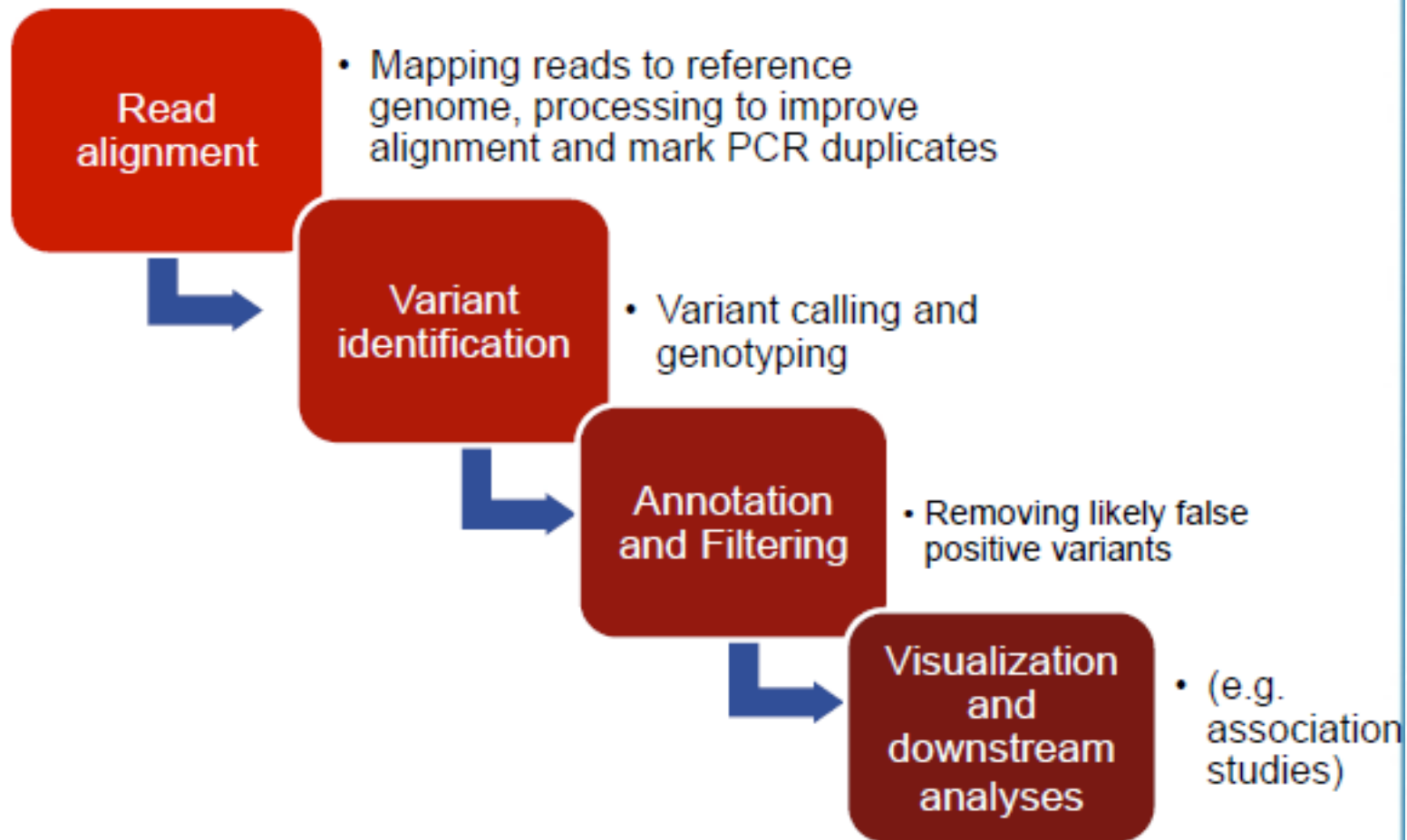


Virtually all human diseases, except perhaps trauma, have a genetic component.

# Variant identification and Analysis

- Given the importance and potential effects of variants a relevant aspect in biological studies becomes the **identification** and **analysis** of variants associated with a specific trait of population.
- Variant calling involves comparing a sample sequence, which may be a single gene sequence, a whole exome or a whole genome, and comparing it to a [reference sequence](#).
- Differences between the sample and the reference are identified, which may be single base changes, such as SNPs and indels, or may be larger scale structural variants.

# A typical variant calling pipeline

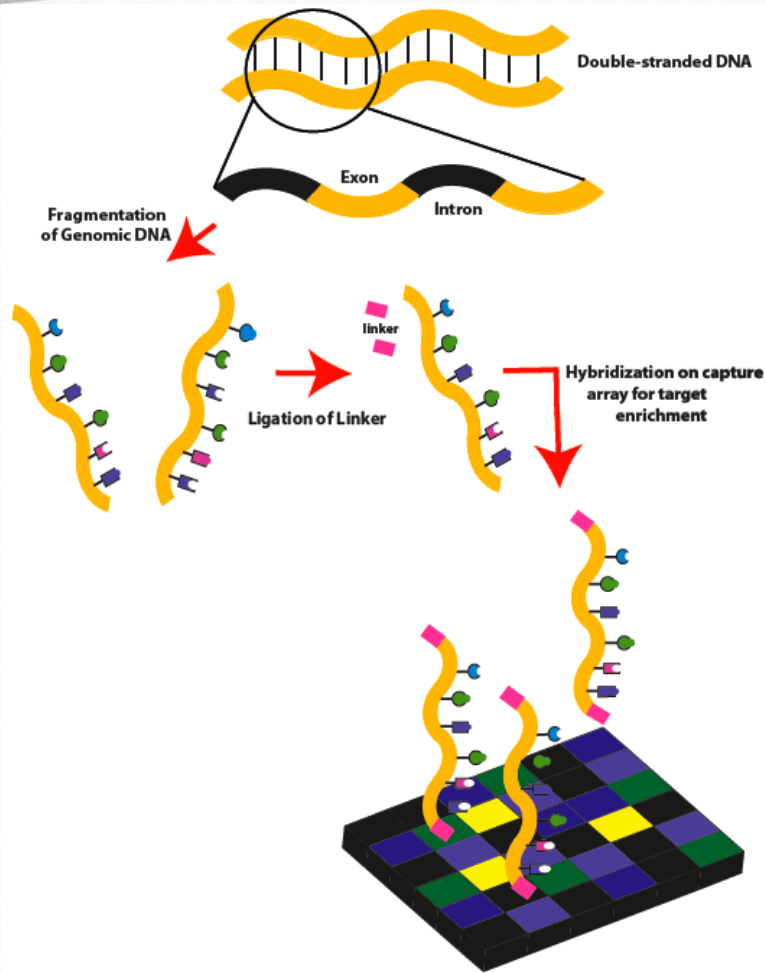


# Exome sequencing

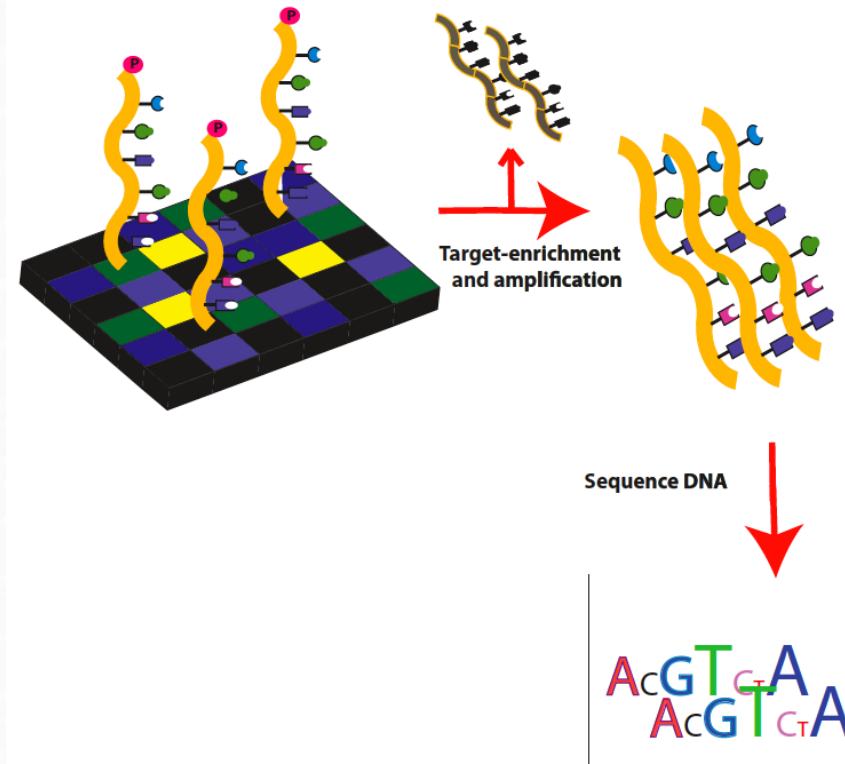
- A genomic technique for sequencing all protein-coding genes in a genome
- It consists of two steps
  1. select only **exons**: the subset of DNA that encodes proteins.
  2. sequence the exonic DNA using any high-throughput DNA sequencing technology
- Humans have
  - 180000 exons
  - 1% of the genome
  - 30 million base pairs



# Exome sequencing workflow



1: Target enrichment



2: Sequencing

# Exome sequencing applications

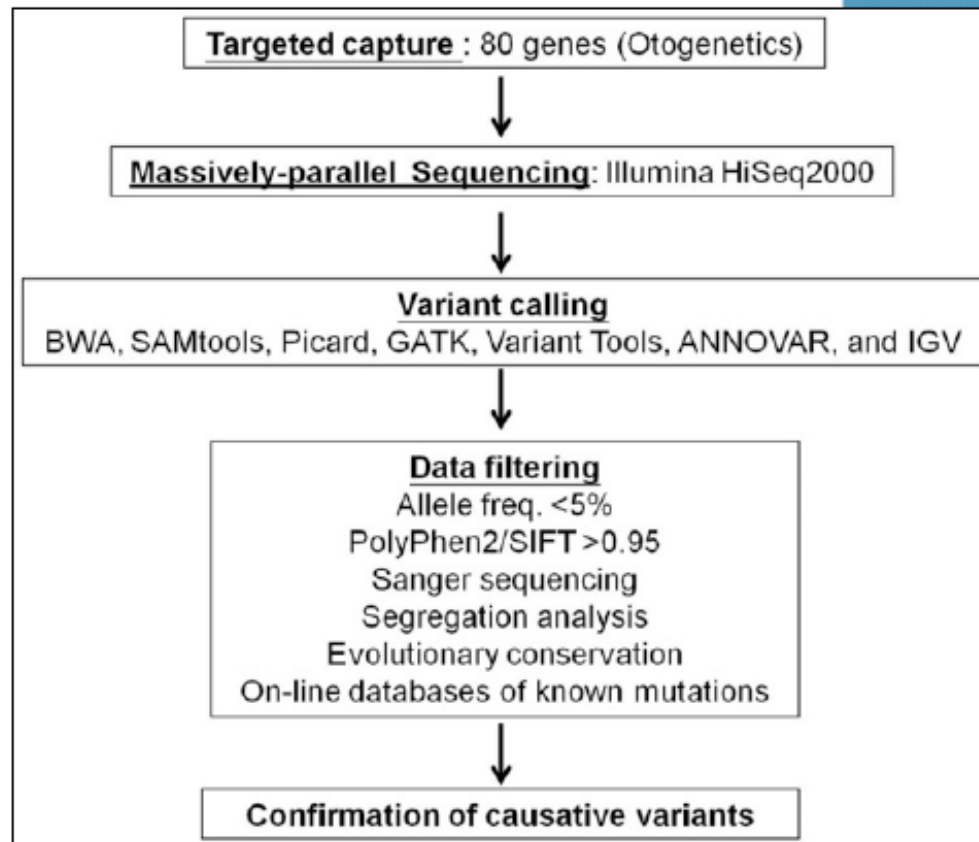
PLoS One. 2013;8(2):e57369. doi: 10.1371/journal.pone.0057369. Epub 2013 Feb 22.

## Application of massively parallel sequencing to genetic diagnosis in multiplex families with idiopathic sensorineural hearing impairment.

Wu CC, Lin YH, Lu YC, Chen PJ, Yang WS, Hsu CJ, Chen PL.

Department of Otolaryngology, National Taiwan University Hospital, Taipei, Taiwan ; Department of Medical Genetics, National Taiwan University Hospital, Taipei, Taiwan.

Analysis  
Workflow →



## Goals:

- Identify variant bases, genotype likelihood and allele frequency while avoiding instrument noise.
- Essentially a three step process:
  - Carry out WG or WE sequencing to create [FASTQ](#) files.
  - Align the sequences to a reference creating [BAM](#) or [CRAM](#) files.
  - Identify where the aligned reads differ from the reference genome and write to a [VCF](#) file with genotypes assigned to each sample.

# Not free from difficulties

- Base calling errors
  - Different types of errors that vary by technology, sequence cycle and sequence context
- Low coverage sequencing
  - Lack of sequence from two chromosomes of a diploid individual at a site
- Inaccurate mapping
  - Aligned reads should be reported with mapping quality score



# Pre-processing reads for variant analysis

- Trimming:

- Unless quality of read data is poor, low quality end trimming of short reads (e.g. Illumina) prior to mapping is usually not required.
- If adapters are present in a high number of reads, these could be trimmed to improve mapping.

- *Recommended tools: Prinseq, Btrim64, FastQC*

- Marking duplicates:

- It is not required to remove duplicate reads prior to mapping but instead it is recommended to mark duplicates after the alignment.

- *Recommended tool: Picard.*

- A library that is composed mainly of PCR duplicates could produce inaccurate variant calling.

- <http://www.biomedcentral.com/1471-2164/13/S8/S8>

# SNP Calling Methods

- Early SNP callers and some commercial packages use a simple method of counting reads for each allele that have passed a mapping quality threshold. **\*\* This is not good enough in particular when coverage is low.**
- It is best to use advanced SNP callers which add more statistics for more accurate variant calling of low coverage datasets and indels.

Bayesian  
model

$$\begin{aligned} \Pr\{G|D\} &= \frac{\overbrace{\Pr\{G\}}^{\text{Prior of the genotype}} \overbrace{\Pr\{D|G\}}^{\text{Likelihood of the genotype}}}{\sum_i \Pr\{G_i\} \Pr\{D|G_i\}}, \text{ [Bayes' rule]} \\ \Pr\{D|G\} &= \prod_j \left( \frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2} \right) \text{ where } \overbrace{G = H_1 H_2}^{\text{Diploid assumption}} \\ \Pr\{D|H\} &\text{ is the haploid likelihood function} \end{aligned}$$

# SNP and Genotype calling tools

- There is a plethora of tools for each step of the process



Volume 15, Issue 2  
March 2014

## A survey of tools for variant analysis of next-generation genome sequencing data

Stephan Pabinger, Andreas Dander, Maria Fischer, Rene Snajder, Michael Sperk, Mirjana Efremova, Birgit Krabichler, Michael R. Speicher, Johannes Zschocke, Zlatko Trajanoski

*Briefings in Bioinformatics*, Volume 15, Issue 2, 1 March 2014, Pages 256–278,  
<https://doi.org/10.1093/bib/bbs086>

*Slightly outdated but a good overview*

# VCF format

```
[HEADER LINES]
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA12878
chr1 873762 . T G 5231.78 PASS [ANNOTATIONS] GT:AD:DP:GQ:PL 0/1:173,141:282:99:255,0,255
chr1 877664 rs3828047 A G 3931.66 PASS [ANNOTATIONS] GT:AD:DP:GQ:PL 1/1:0,105:94:99:255,255,0
chr1 899282 rs28548431 C T 71.77 PASS [ANNOTATIONS] GT:AD:DP:GQ:PL 0/1:1,3:4:25:92:103,0,26
chr1 974165 rs9442391 T C 29.84 LowQual [ANNOTATIONS] GT:AD:DP:GQ:PL 0/1:14,4:14:60:91:61,0,255
```

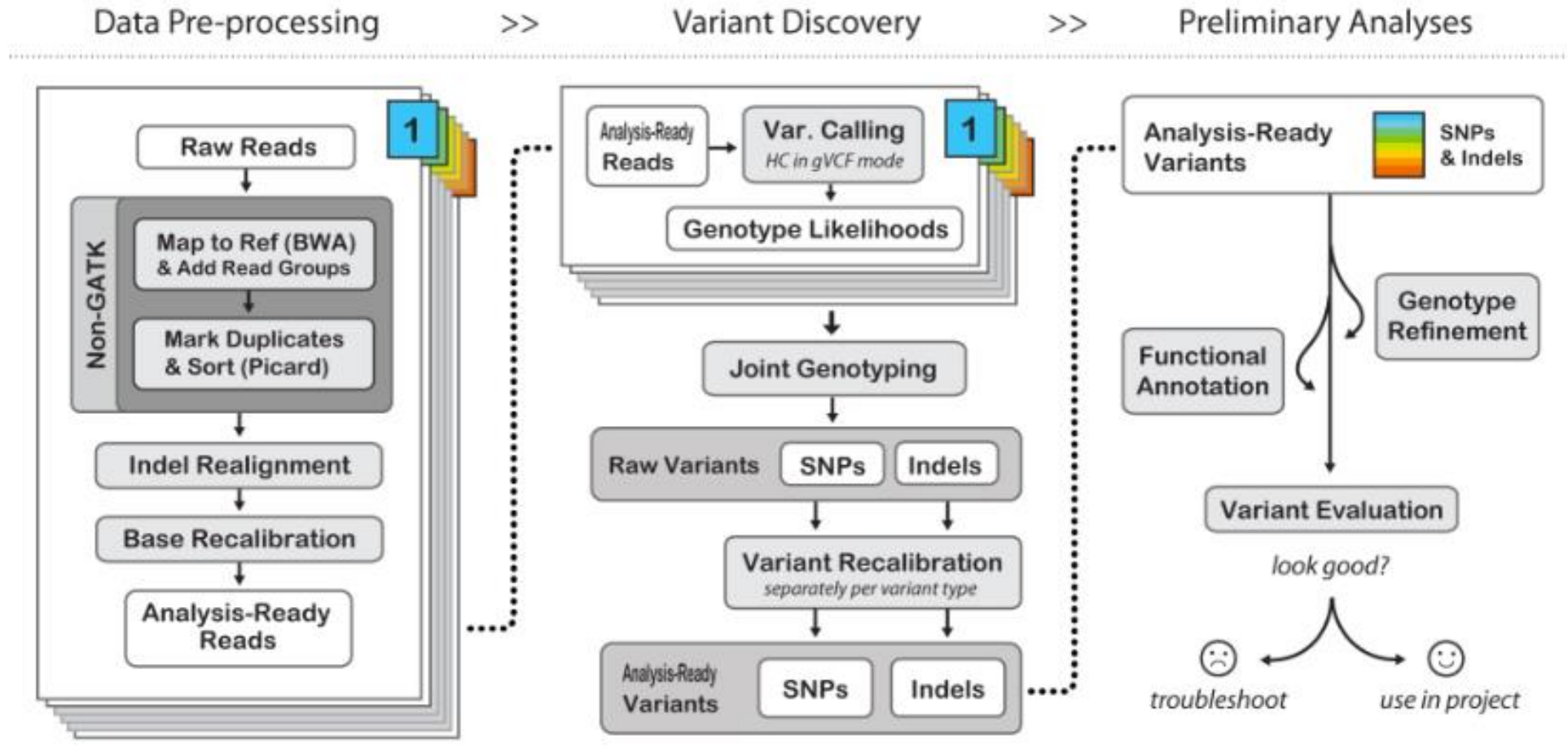
## How variation is represented in a VCF

Each line represents one variant (here everything is a SNP, but some could be indels or CNVs) as well as the genotype of our sample, NA12878, at that variant. I've chosen these four variants because they each represent an important aspect in interpreting a VCF file:

- chr1:873762 is a novel T/G polymorphism, found with very high confidence (QUAL = 5231.78).
- chr1:877664 is a known A/G SNP (rs3828047), found with very high confidence (QUAL = 3931.66)
- chr1:899282 is a known C/T SNP (rs28548431), but has a relative low confidence (QUAL = 71.77)
- chr1:974165 is a known T/C SNP but we have so little evidence for this variant in our data that although we write out a record for it (book keeping, really) our statistical evidence is so low that we filter the record out as a bad site "LowQual".



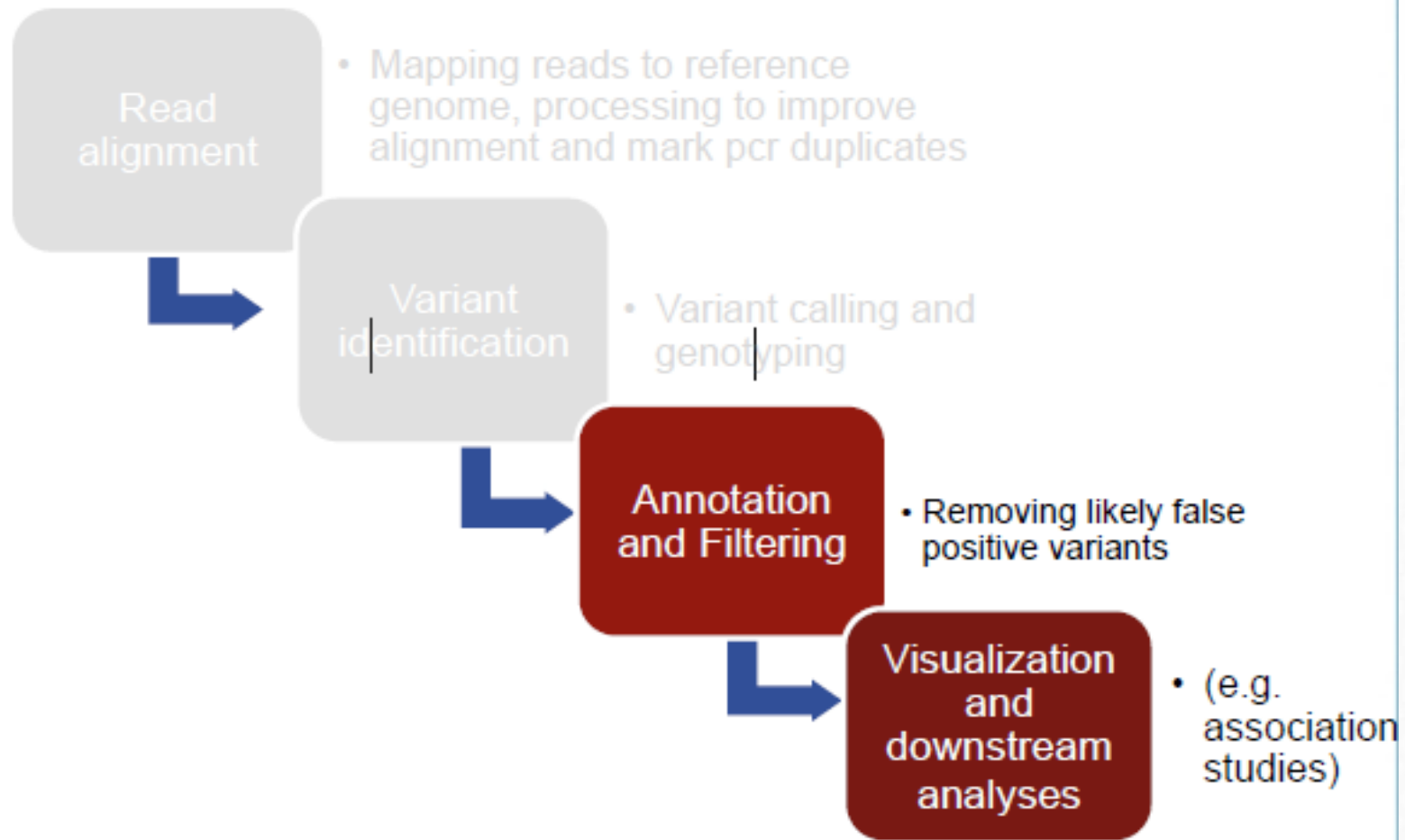
# GATK pipeline for variant calling



# From the output of thousands of variants which ones should you consider?

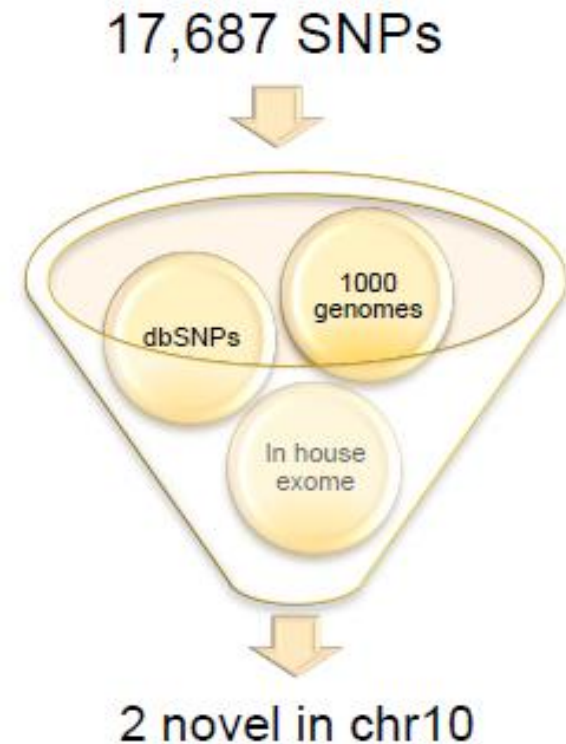
- Various important considerations
  - Is the variant call of good quality?
  - If you expect a rare mutation, is the variant commonly found in the general population?
- What is the predicted effect of the variant?
  - Non-synonymous
  - Detrimental for function

# A typical variant calling pipeline



# Filtering variants

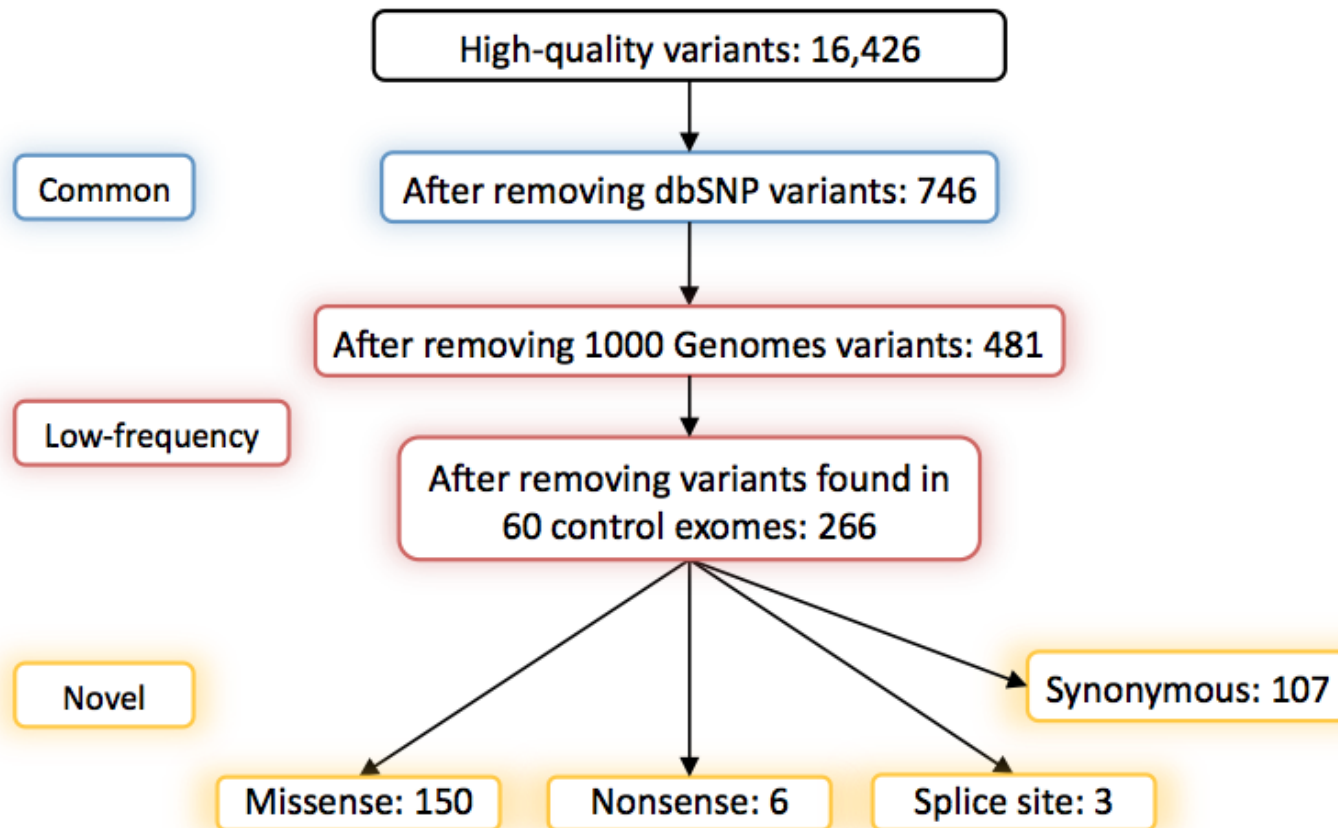
- The initial set of variants is usually filtered extensively in hopes of removing false positives.
- More filtering can include public data or custom filters based on in-house data.





# Removing common and low frequency to get the novel SNPs

## Variants found in Individual II.4



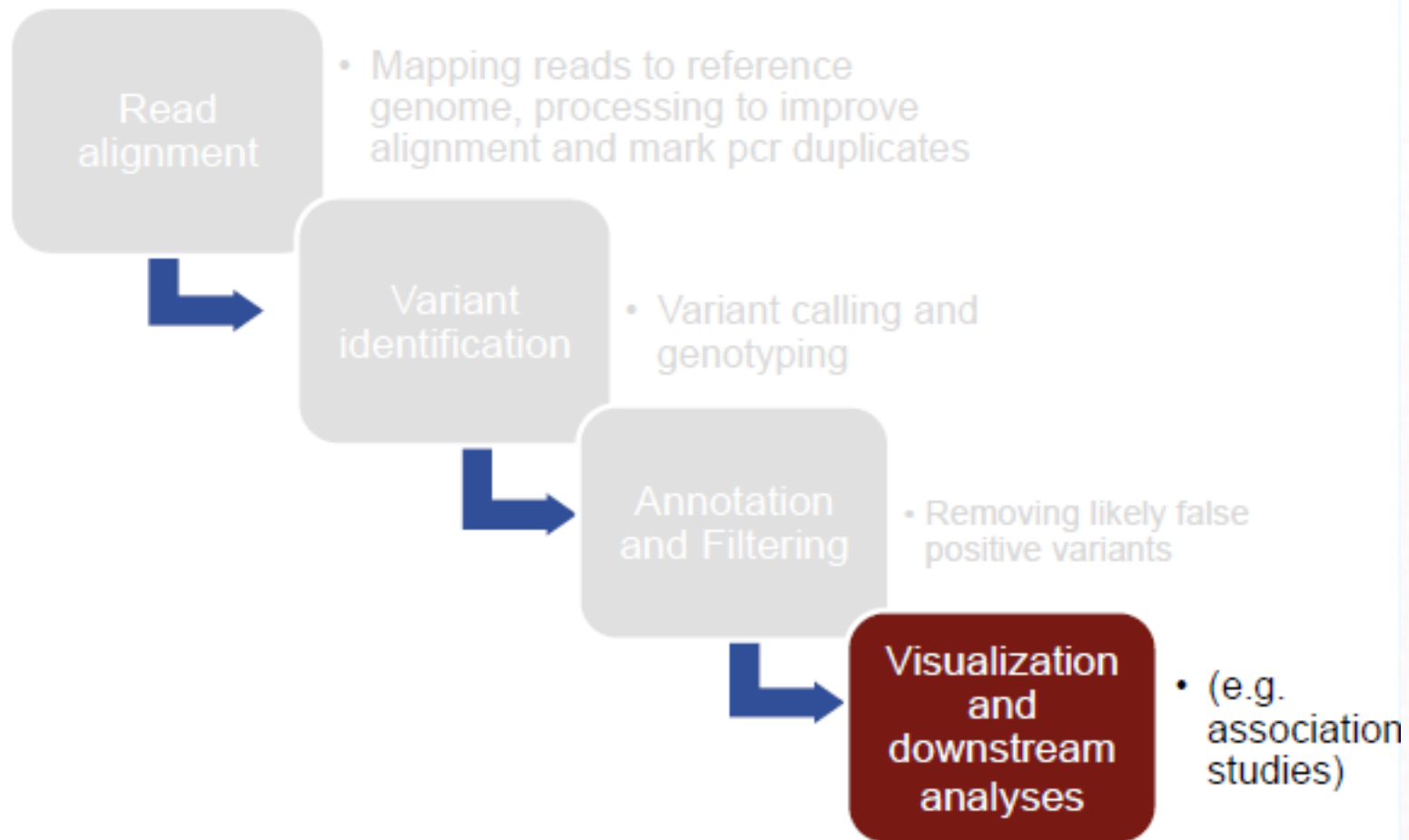
# Filtering variants

- Most exome studies will then filter common variants (>1%) such as those in dbSNP database
  - Good tools for filtering and annotation: Annovar, SNPeff
  - Reports genes at or near variants as well as the type (nonsynonymous coding, UTR, splicing, etc.)
- To filter by variant consequence, use SIFT and Polyphen2.
  - These are available via Annovar or ENSEMBL VEP
  - [http://www.ensembl.org/Homo\\_sapiens/Tools/VEP](http://www.ensembl.org/Homo_sapiens/Tools/VEP)
- For a more flexible annotation workflow, explore GEMINI
  - <http://gemini.readthedocs.org/en/latest/>

# Evaluating variant consequence

- **SIFT predictions** - SIFT predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids.
- **PolyPhen predictions** - PolyPhen is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations.
- **Condel consensus predictions** - Condel computes a weighed average of the scores (WAS) of several computational tools aimed at classifying missense mutations as likely deleterious or likely neutral. The VEP currently presents a Condel WAS from SIFT and PolyPhen.

# A typical variant calling pipeline





# Viewing reads in browser

- If your genome is available via the UCSC genome browser <http://genome.ucsc.edu/>, import bam format file to the UCSC genome browser by hosting the file on a server and providing the link.
- If your genome is not in UCSC, use another browser such as IGV <http://www.broadinstitute.org/igv/> , or IGB <http://bioviz.org/igb/>
  - Import genome (fasta)
  - Import annotations (gff3 or bed format)
  - Import data (bam)

**IT'S YOUR TURN NOW!**