Variant detection using Next Generation Sequencing

Tsviya Olender

January 2020

- 4-year-old girl from a Middle-Eastern Arab consanguineous family.

- persistent secretory diarrhea at 18 days of age, which did not improve with different types of diets or medication.

- completely dependent on total parenteral nutrition (TPN) ever since (intravenous administration of nutrition).

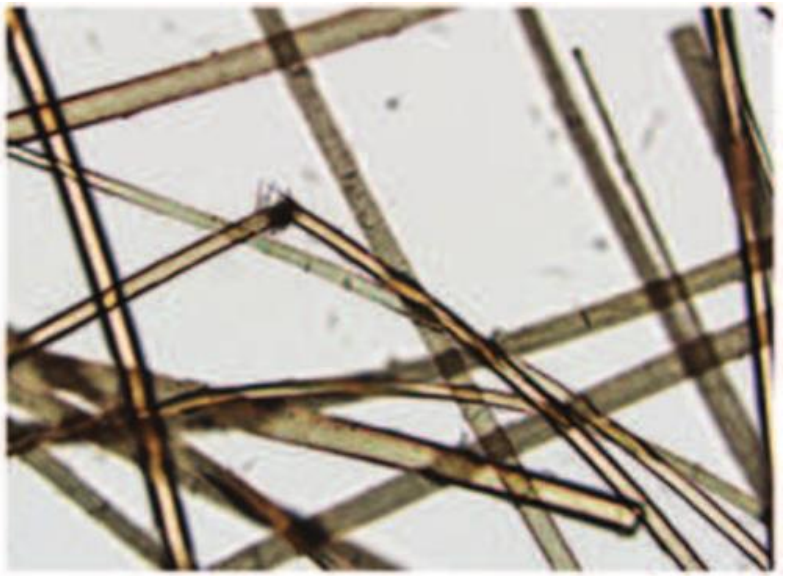
- She had minor dysmorphic features but no obvious developmental delay.

- Various clinical tests including MRI, US, duodenal biopsy, liver profiles, blood tests and many more- were normal

Oz-Levi D. *Clin Genet.* 2015

Exome sequencing discovered a homozygote mutation in the gene *TTC37*, a known gene for trichohepatoenteric syndrome (THES)-
characteristic features include:
intractable diarrhea, growth retardation, facial dysmorphism, in infancy requiring TPN, **hair abnormalities** and immunodepression.

Exome sequencing as a differential diagnosis tool



Oz-Levi D. *Clin Genet.* 2015

NGS in the clinics

- Genetic consultant
- Non-Invasive Prenatal Testing (NIPT) instead of amniotic fluid test.
- Oncogenetics- sequencing of cancer related genes.

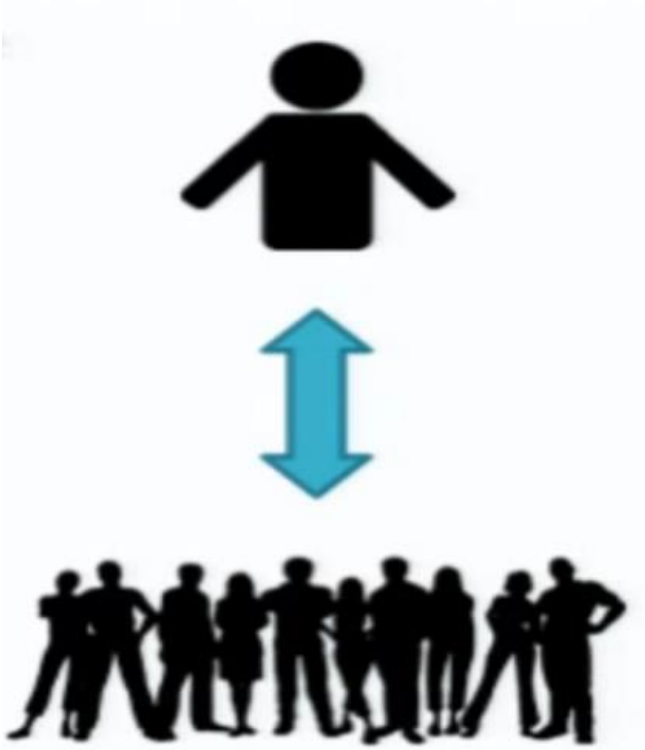
In Israel: these tests are now covered by KUPOT HOLIM for individuals that have high risk for certain diseases.

What is a genetic variation?

A genetic variation is the difference in DNA sequences between individuals within a population (EBI site)

A genetic variation is an alternation in the individual's genome relative to the reference genome (GATK site)

- to what extent the reference genome represents the population
- quality of the reference genome
- How well the reference genome represents the genome that we sequenced



Types of genetic variation

SNPs

ind1 **AACAAGCCA**
ind2 **AACAGGCCA**
ind3 **AACAAGCCA**
ind4 **AACATGCCA**

10M human coding SNPs

√

InDels

ind1 **ACAATCGCCA**
ind2 **ACA---GCCA**
ind3 **ACA---GCCA**
ind4 **ACAATCGCCA**

~800,000 human coding indels

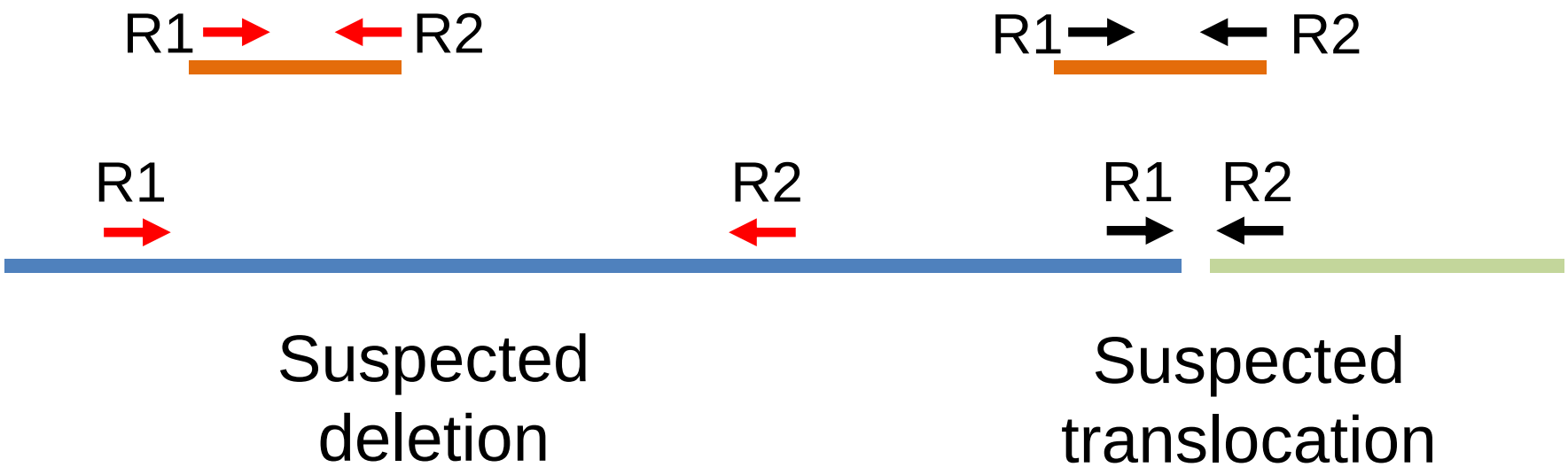
~ half of the read length

Read of 100bp = 50bp indel

Structural variations

Large deletions, duplications, insertions, inversions, translocations

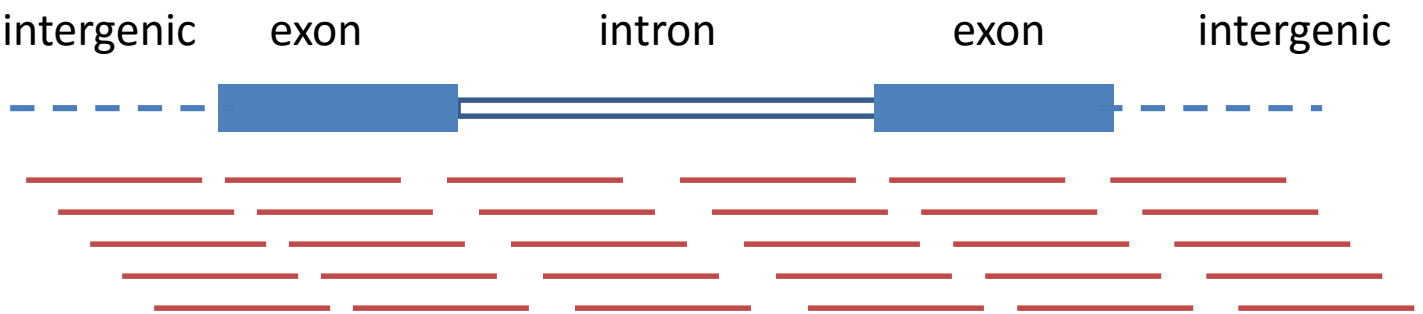
Inferring structural variation from paired-ends



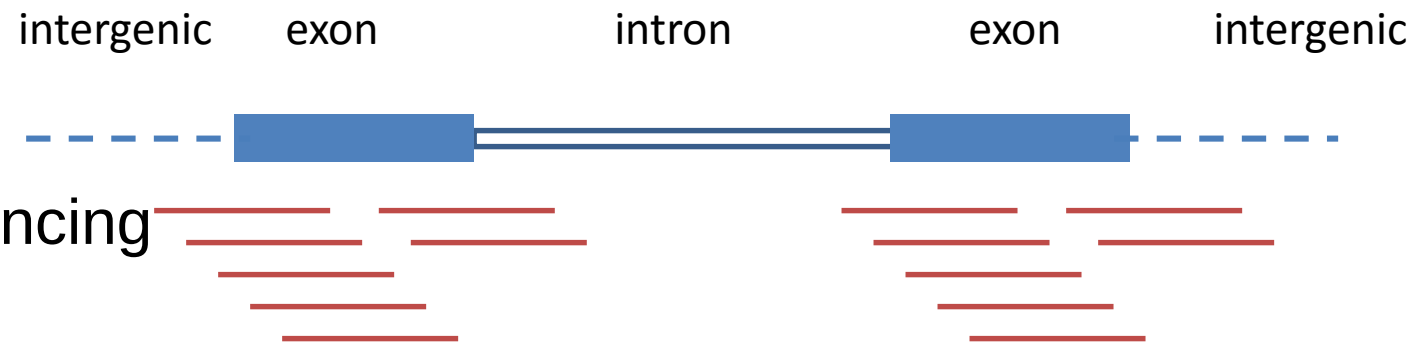
Direct measurement: Oxford nanopore, PacBio

Different types of experimental design

Whole genome



Targeted sequencing (gene panels, exome)



synthesized oligos

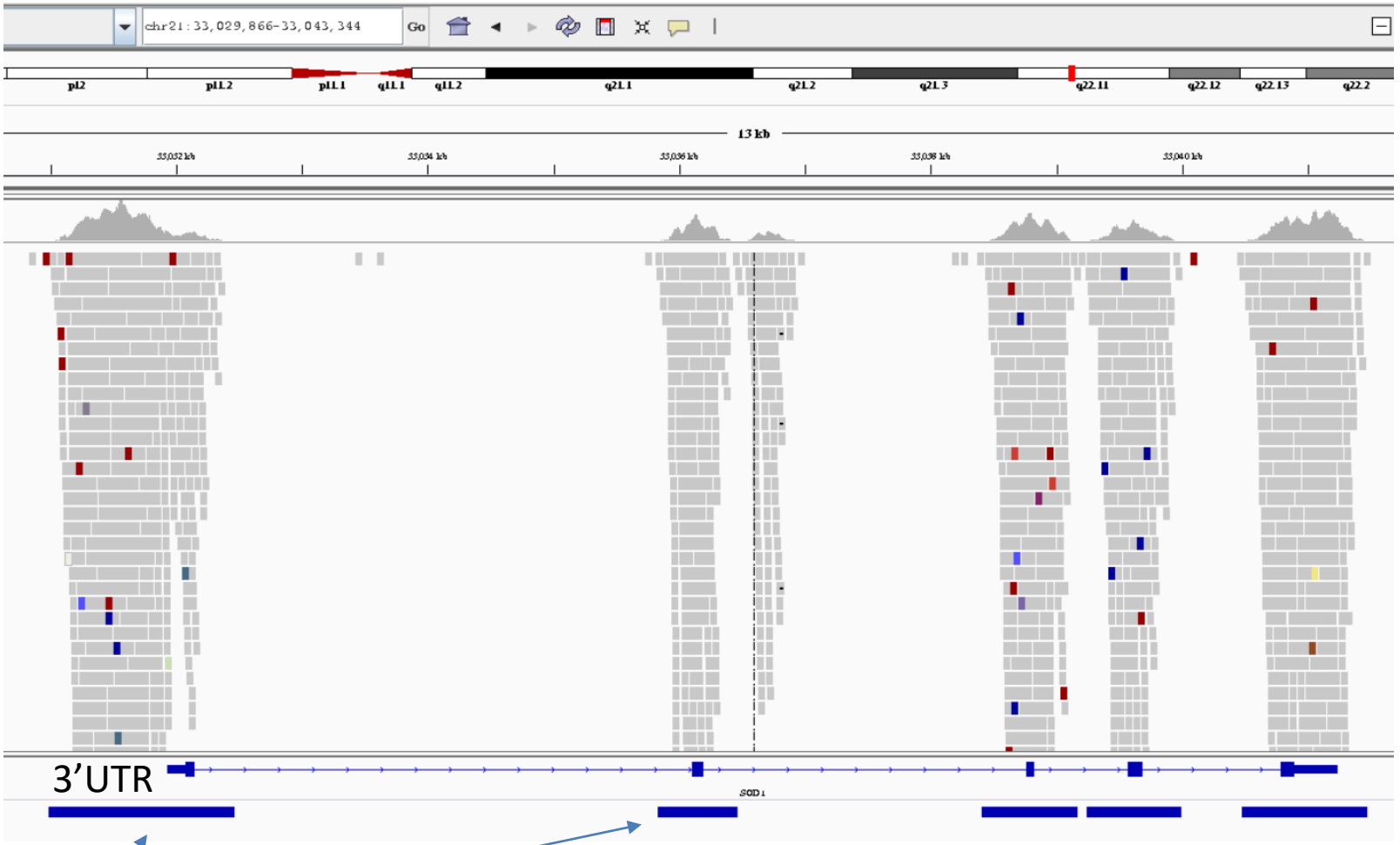
RNA-seq => will not be covered today

Targeted sequencing

Coverage

Aligned reads

SOD1

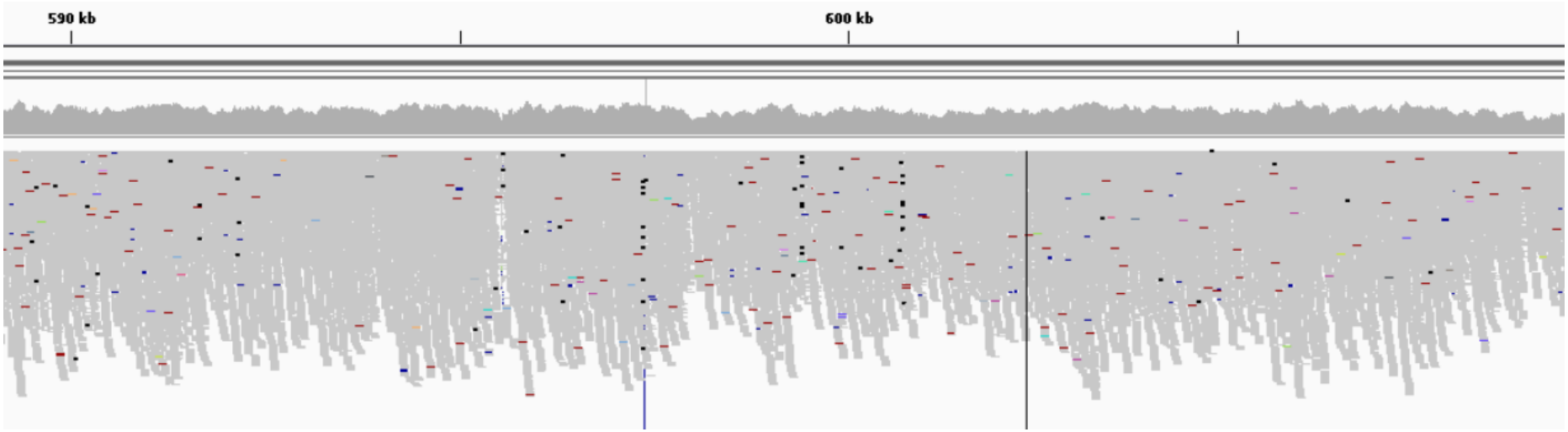


Synthesized oligos

Project of:
Chen Eitan
Eran Hornstein

Example: whole genome sequencing of S. cerevisiae

The coverage is uniform



Non-reference bases are colored, reference bases are grey

Project of:
Dana Bar Zvi
Naama Barkai

Exome sequencing

- The coverage is not uniform
- Structural variations can not be called

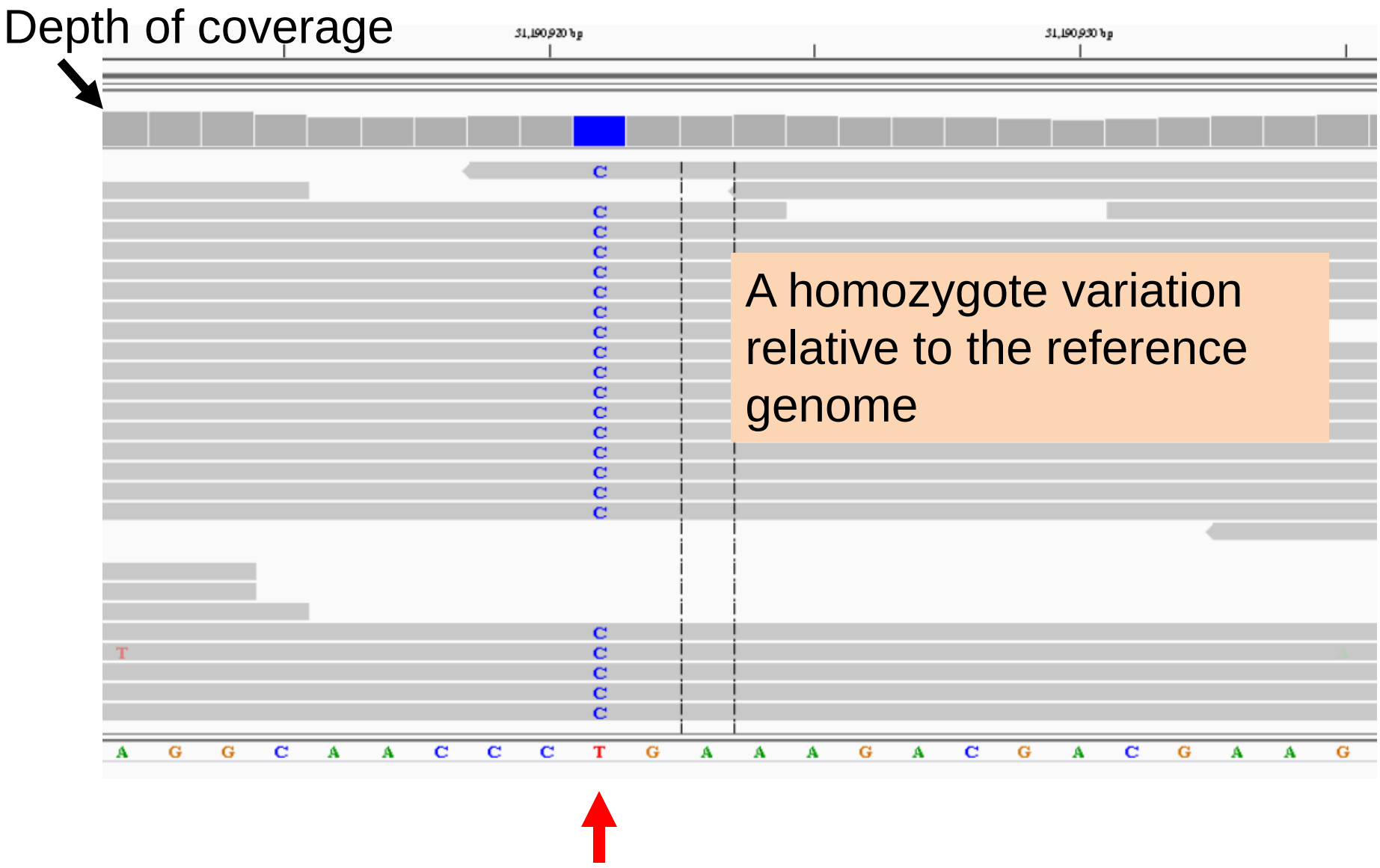
- Cheap (a whole human exome ~250-500$)

Whole genome sequencing

- The false positive calling rate lower
- Uniform coverage.
- Quantitative- good for structural variants.

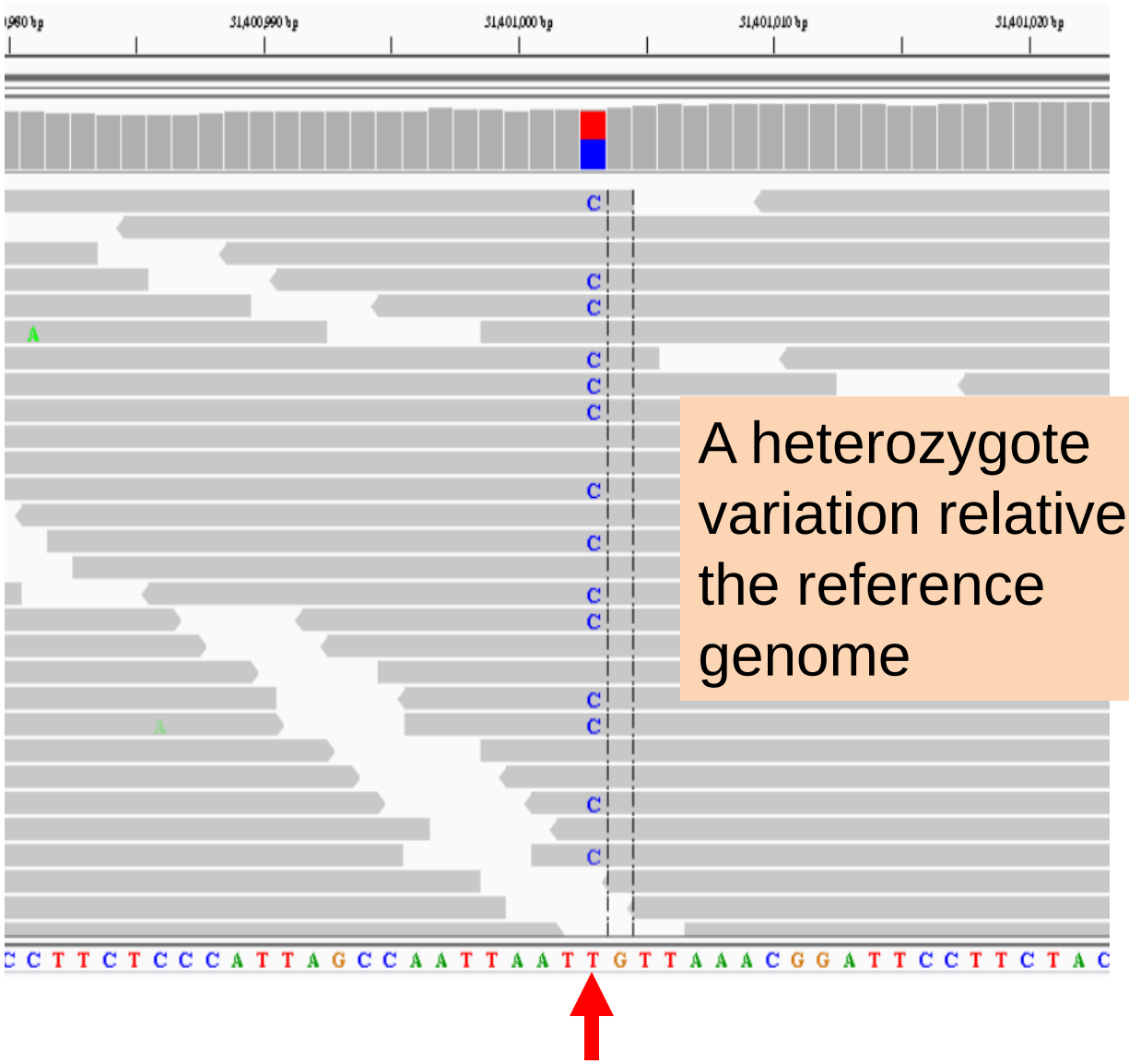
- Expensive -1000$

Example 1: sequencing of a diploid genome



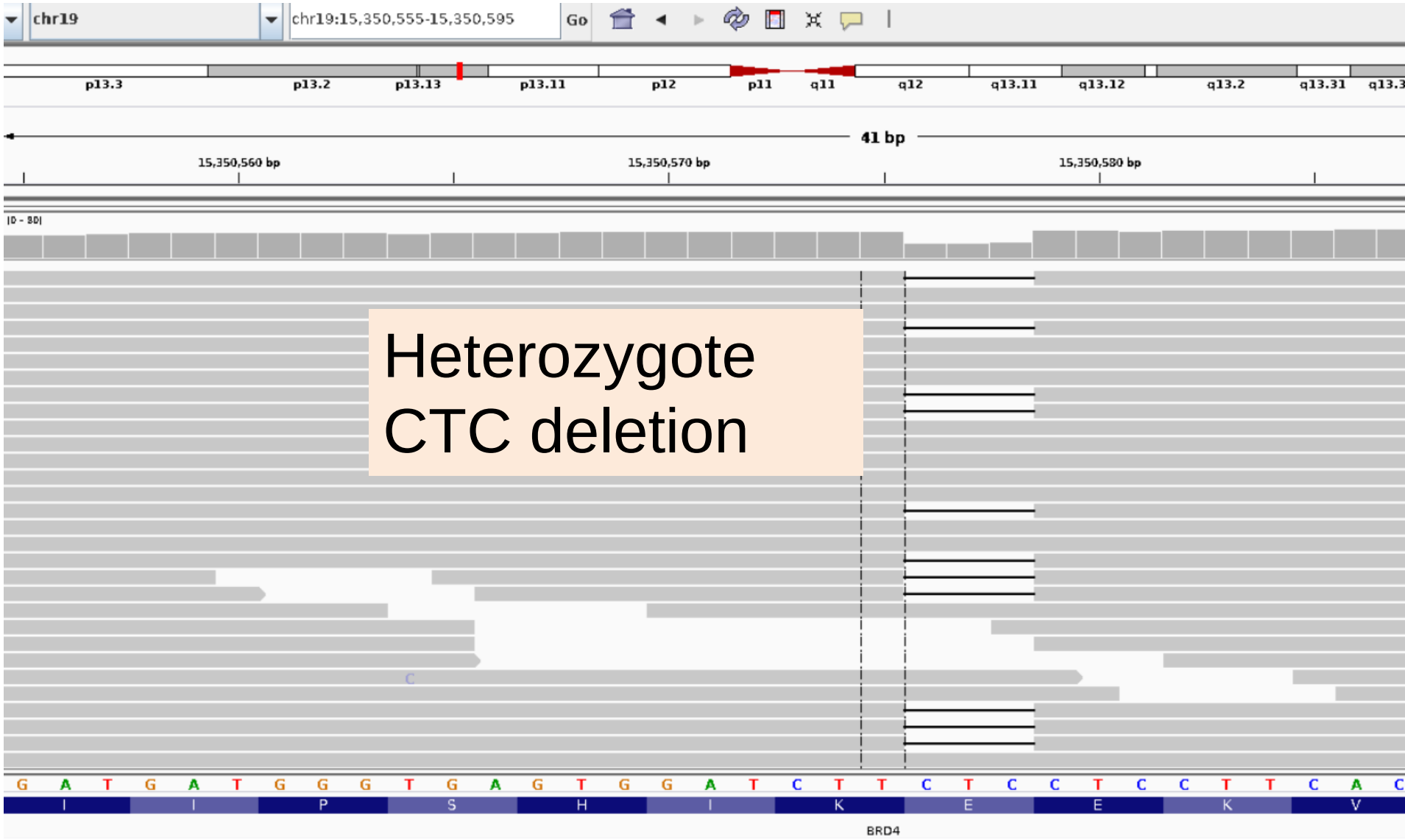
Example 2: sequencing of a diploid genome

Depth of coverage

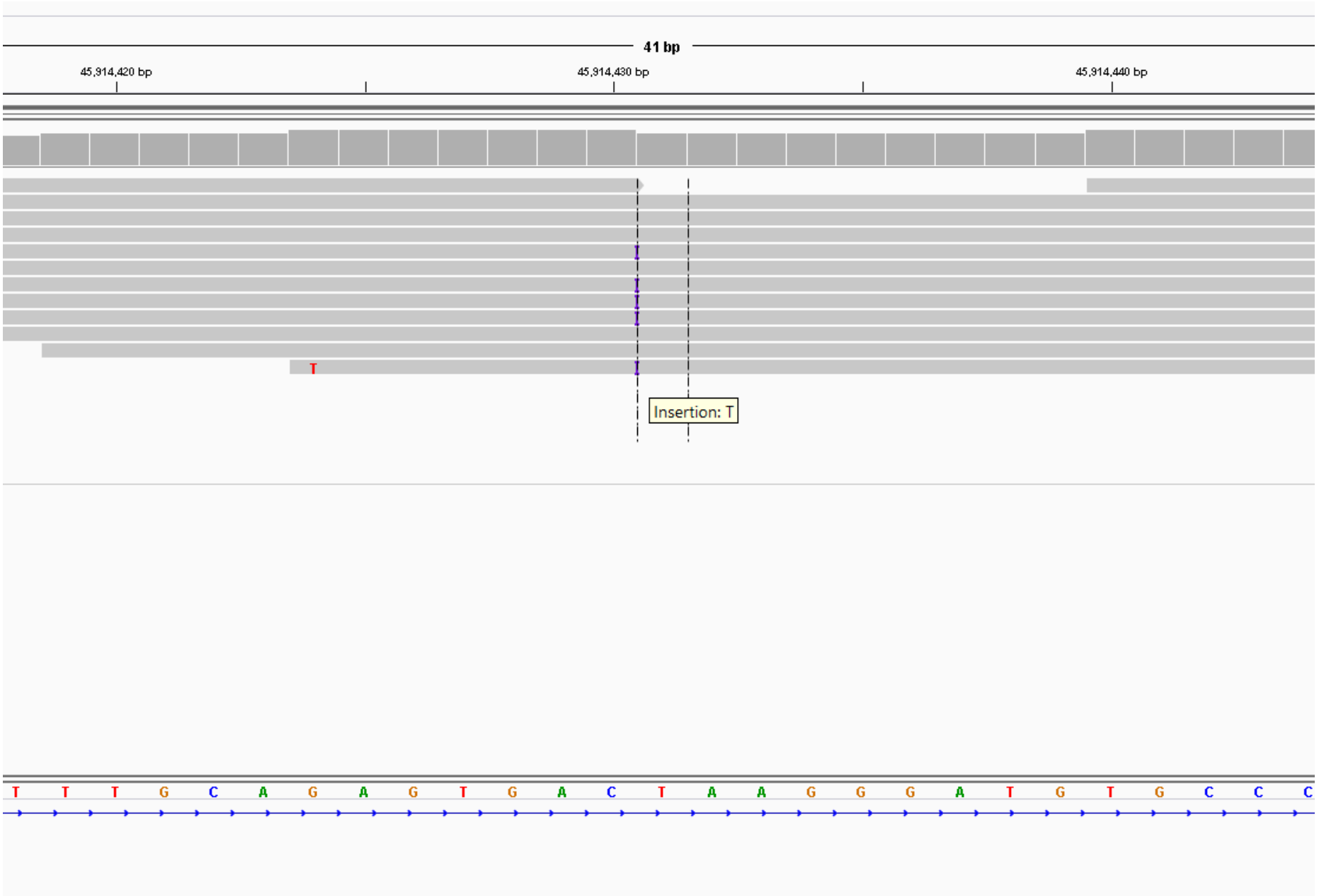


A heterozygote variation relative to the reference genome

Example 3:


chr19
chr19:15,350,555-15,350,595
Go
p13.3
p13.2
p13.13
p13.11
p12
p11
q11
q12
q13.11
q13.12
q13.2
q13.31
41 bp
15,350,560 bp
15,350,570 bp
15,350,580 bp
Heterozygote
CTC deletion
G A T G A T G G G T G A G T G G A T C T T C T C C T C C T T C A C
I I P S H I K E E K V
BRD4

Example3: Heterozygote insertion




40 bp
250,350 bp
250,360 bp
Sequencing error

Variant calling is

- very sensitive to sequencing errors
- very sensitive to alignment errors

Variant calling always suffer from false positive calls

Sequencing considerations

- High coverage
- Accurate alignment

} Reduce the rate of false-positive calls

Recommendations:

- Paired-ends sequencing, ~100bp long
- The required coverage: depends in your goals.

rule of thumb (for detecting rare variants):

20-40X for whole genome
30-50X for targeted sequencing

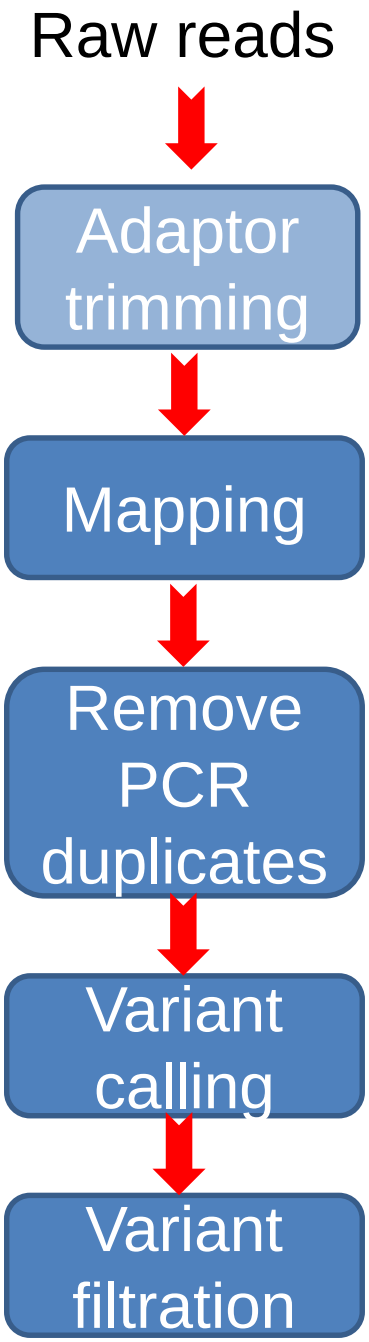
The 1000 genome project

An international effort to establish the most detailed catalogue of human genetic variation. Initiated in 2008

They aimed to sequence 1000 individuals and finished with many more

They applied only 4X coverage

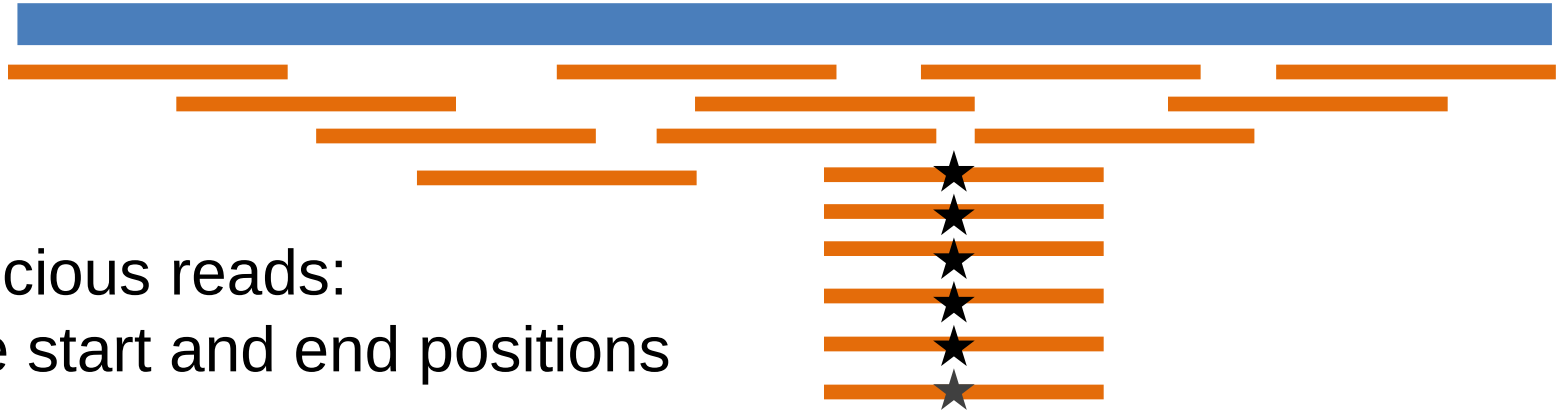
A typical work-flow



	RNAseq	Variant calling
Adaptor trimming	√	X
Mapping	bowtie	Bwa- allows mismatches
Remove PCR duplicates	X	√
Variant calling	X	e.g GATK
Variant filtration	X	e.g vcftools

Mark/Remove duplicates

A method to identify duplicated reads



Suspicious reads:
Same start and end positions

PCR artifacts

To avoid getting a lot of evidences from sequencing errors

Algorithms:
Picard (most popular), samblaster, samtools

Algorithms for variant detection

Samtools

Freebayes

GATK

Varscan

GlfMultiples

Cortex

GATK
Genome Analysis ToolKit

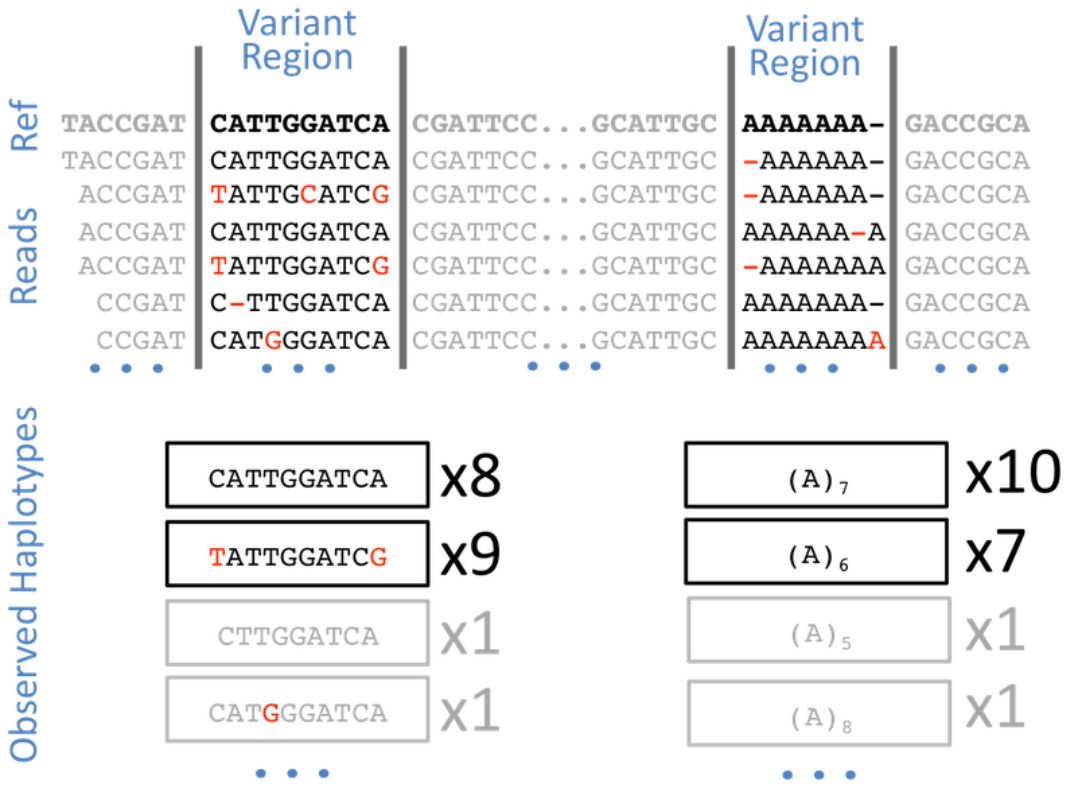
Considered as industrial gold standard
Sensitive, low rate of false positive

Disadvantages:

Extremely slow. Computationally heavy.
Not easy to implement.
Not compatible with other tools.
Optimized to work on human data.

Variant calling

GATK (haplotype caller) and other algorithms (freebayes) use local de-novo assembly for variant calling.



Output- The Likelihoods of all possible genotypes at all sites

The VCF file format

- The Variant Call Format (VCF) file is the most accepted format to store information of genetic variations.

```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=GRCh37
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
#CHROM  POS     ID      REF  ALT   QUAL  FILTER   INFO      FORMAT    IND1
chr2    4370    rs6057   G    A     29      .      DP=13     GT:GQ     1/1:43
chr4    7330    .        T    A     3      q10     DP=12     GT:GQ     0/1:41
```

How to call the genotype
Reference allele= 0
Alternate allele = 1

0/1 => hetrozygote
1/1 => homozygote
0/0 =>?

Variant filtration

- The initial variant calling **always** contains a non negligible amount of false positive calls.

- It is important to **filter** the results based on the quality parameters in the VCF files.

Parameters to be used for variant filtration:

Qual=quality of SNP
DP= depth of coverage
GQ= genotype quality

Tools: VCFtools, bcftools and more

Joint calling

A genetic variation is a characteristic of population, but we measure the alternation of the individual's genome relative to the reference genome

In most studies, more than one genome is being sequenced.

Joint calling-
determines the genotypes of all individuals, in all variant positions.

We call variations relative to the reference genome

0/1	hetrozygote
1/1	homozygote

Child

chr2:29451793	T->A	0/1
chr6:49494505	G->C	1/1
chr8:1842627	T->C	0/1
chr12:88483184	A->G	0/1

Inherited from the mother. What is the father genotype?

mother

chr2:29451793	T->A	0/1
chr6:49494505	G->C	0/1

father

chr6:49494505	G->C	0/1
chr12:88483184	A->G	0/1

After joint calling

		Child	Mother	Father
chr2:29451793	T->A	0/1	0/1	0/0
chr6:49494505	G->C	1/1	0/1	0/1
chr8:1842627	T->C	0/1	0/0	./.
chr12:88483184	A->G	0/1	./.	0/1

Missing genotype

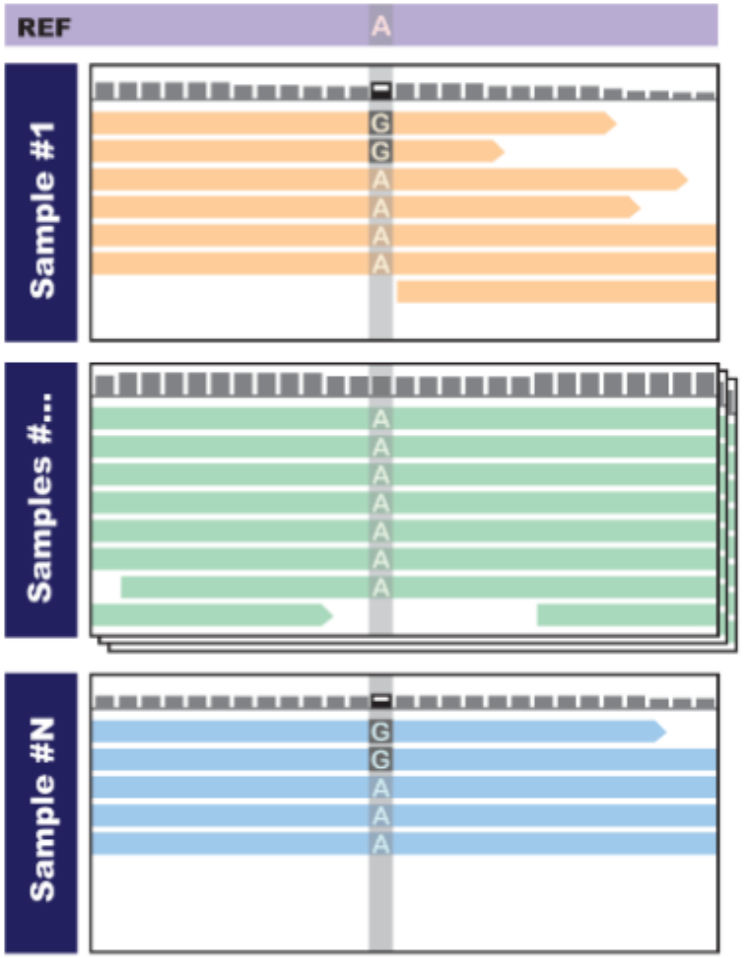
Inherited? De-novo?
Uncertain

0/1	heterozygote
1/1	homozygote
0/0	reference homozygote
./.	missing info

Joint calling empowers analysis

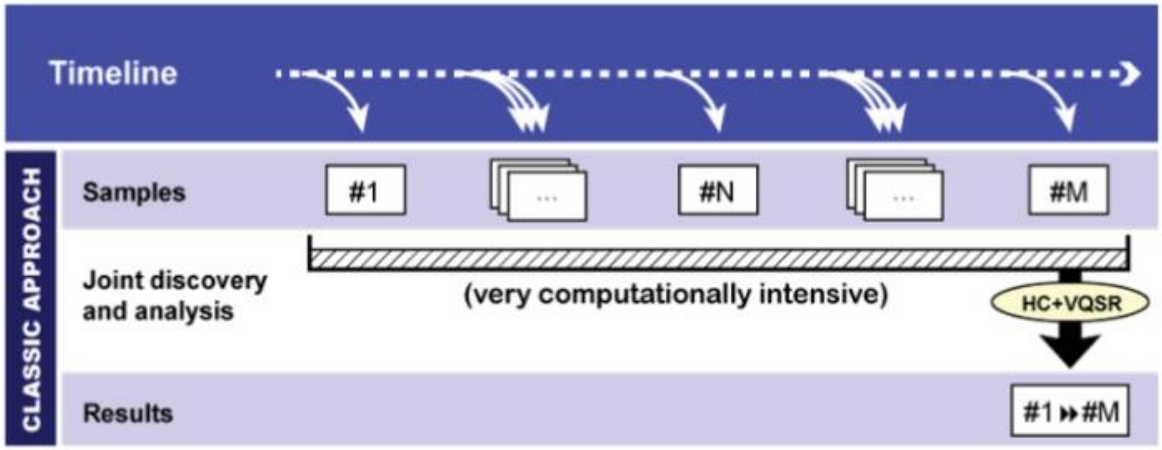
- If we analyze Sample #1 or Sample #N alone we are not confident that the variant is real

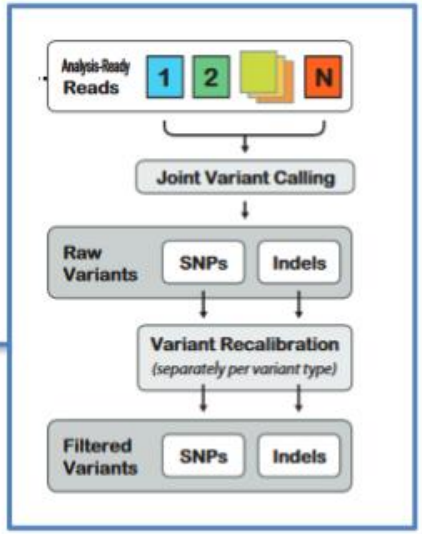
- If we see both samples then we are more confident that there is real variation at this site in the cohort



The N+1 problem

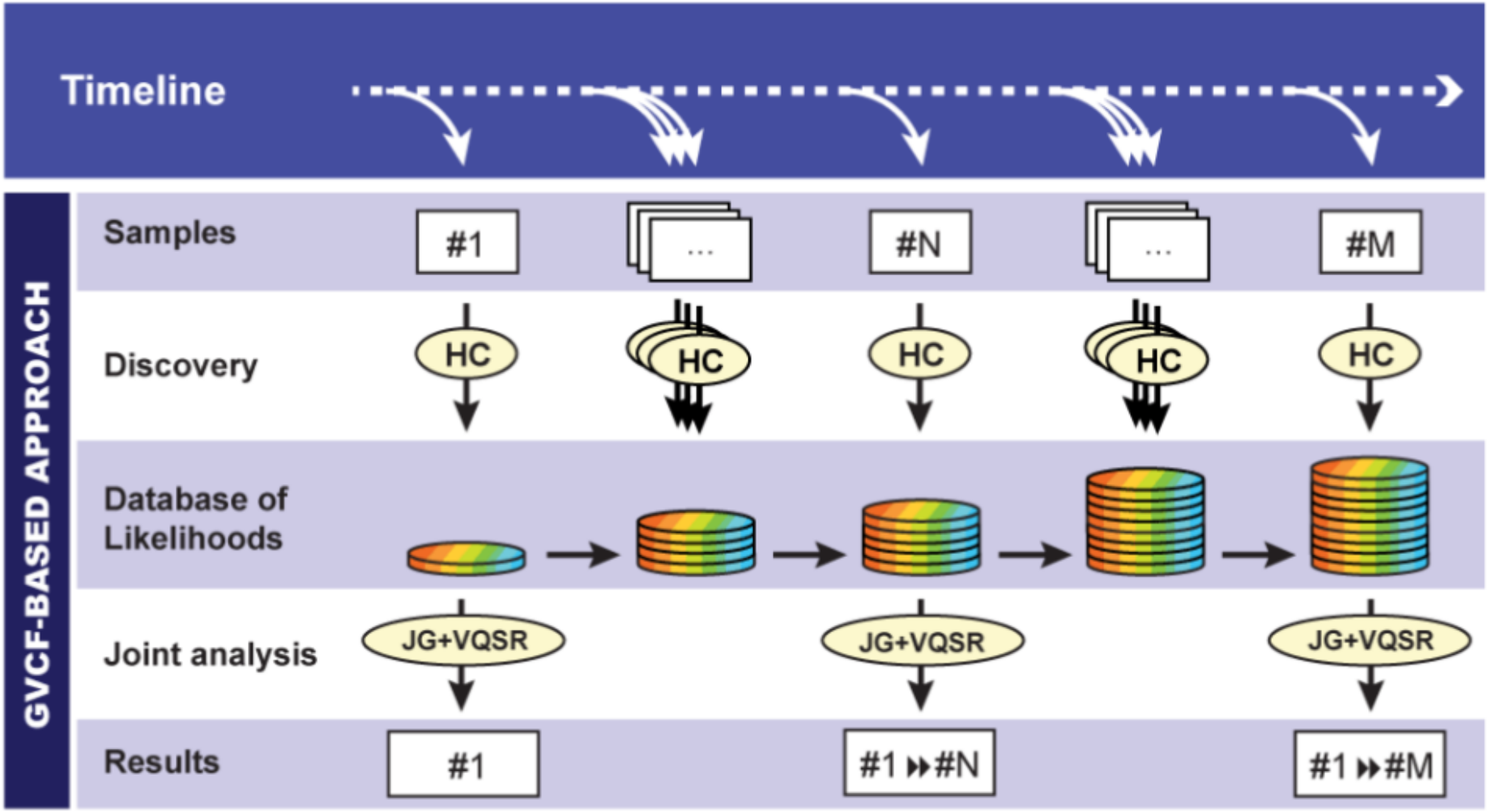
Monkol Lek, 2013





Conceptual solution for the N+1 problem

Genome VCF file == gVCF



Genome VCF (gVCF) Vs. VCF

- gVCF file has records for all sites in the genome, whether there is a variant call there or not.
- The records in a gVCF include an accurate estimation of how confident we are in the determination that the sites are homozygous-reference or not.
- Saves a lot of computation time.
- Variant callers that work with the gVCF model:

GATK, Isaac (of Illumina).

gnomAD- Genome Aggregation Database

- A catalogue of human variations
- The catalogue was established using

joint calling of ~138,000 human individuals

http://gnomad.broadinstitute.org/variant/7-116335828-G-C

Variant calling

- Sensitive to sequencing errors
- Sensitive to alignment errors
- Requires high coverage, paired-ends sequencing

RNAseq

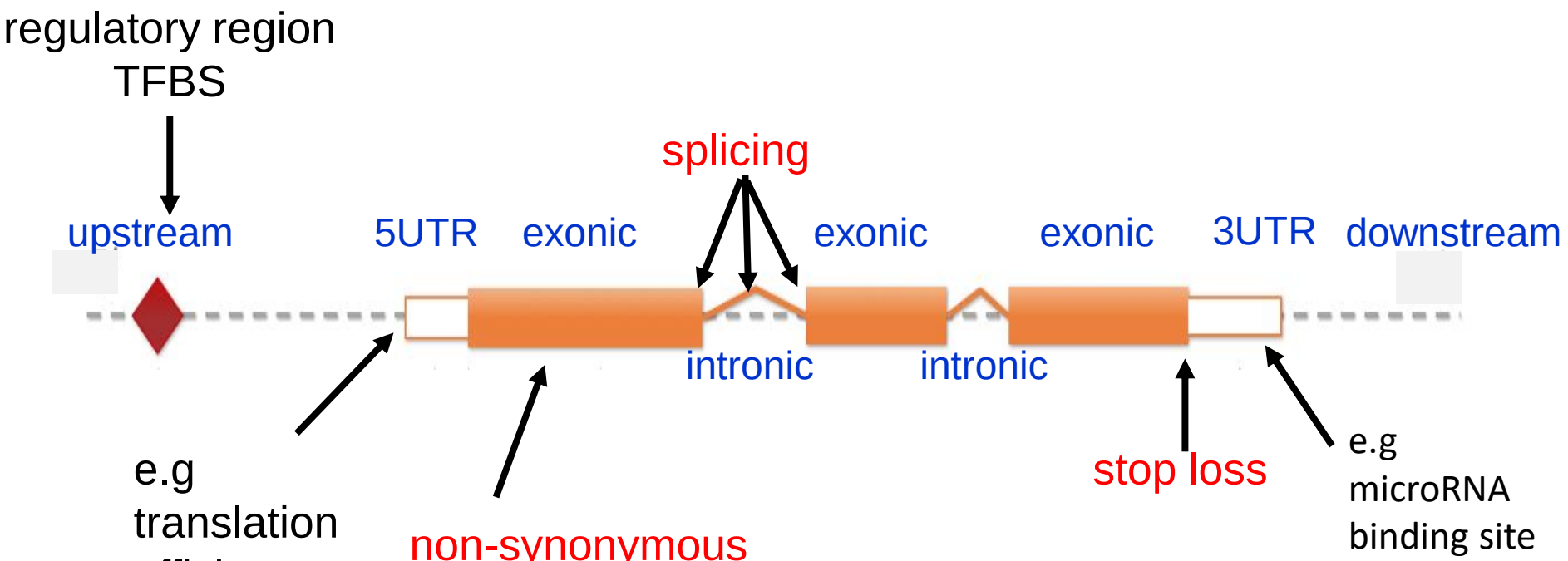
- Less sensitive to sequencing errors
- Less sensitive to alignment errors
- Moderate coverage, single-end sequencing will do

In most cases (99.9%) variations are called via automatic pipelines, except for the variant filtration step

Variant interpretation
(takes 99% of the research time)

Variant annotation

Assigning biological meaning to each variant



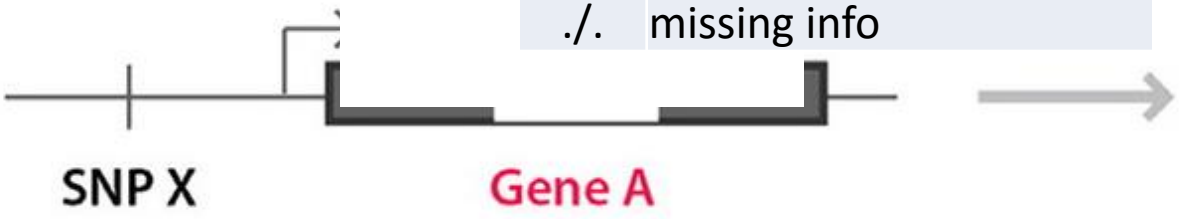
SNPeff,
Variant annotation intergrator,
VEP, annovar

The GTEx project:
54 tissues from 1000 individuals

Cis-eQTL

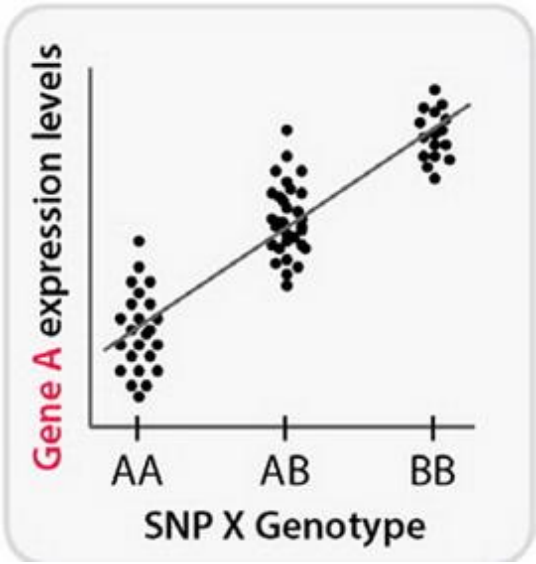
SNP X has an effect on loc

0/1	heterozygote
1/1	homozygote
0/0	reference homozygote
./.	missing info



located in promoter region or enhancer

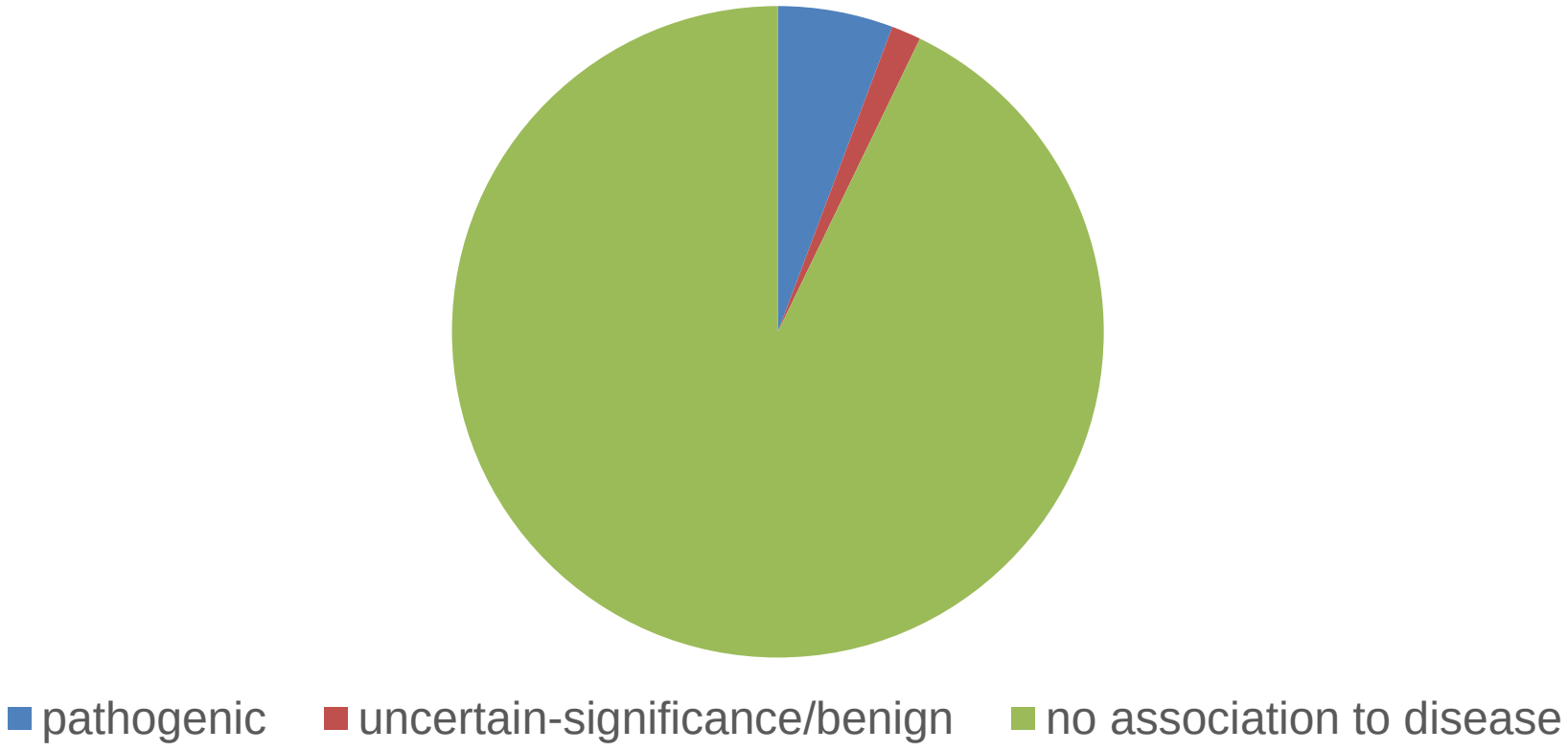
RNAseq expression level



How to assess variant's pathogenicity

Stop-gain variations in the human genome (dbSNP)

Total: 337,763 variations



How to assess pathogenicity of missense variants

SIFT: Sorting Intolerant From Tolerant
Predicts whether an **amino acid substitution affects protein function** based on sequence homology and the physical properties of amino acids.

PolyPhen: predicts the possible impact of an amino acid substitution on the protein function, using **phylogenetic properties** and **structural information** characterizing the substitution.

dbSNFP- database for non-synonymous SNP prediction programs

- prediction scores of 29 prediction algorithms
- conservation scores from 9 algorithms

Frequency in the population

An important measure in the way to ping-point the relevant variation

How common is the phenotype that we study?

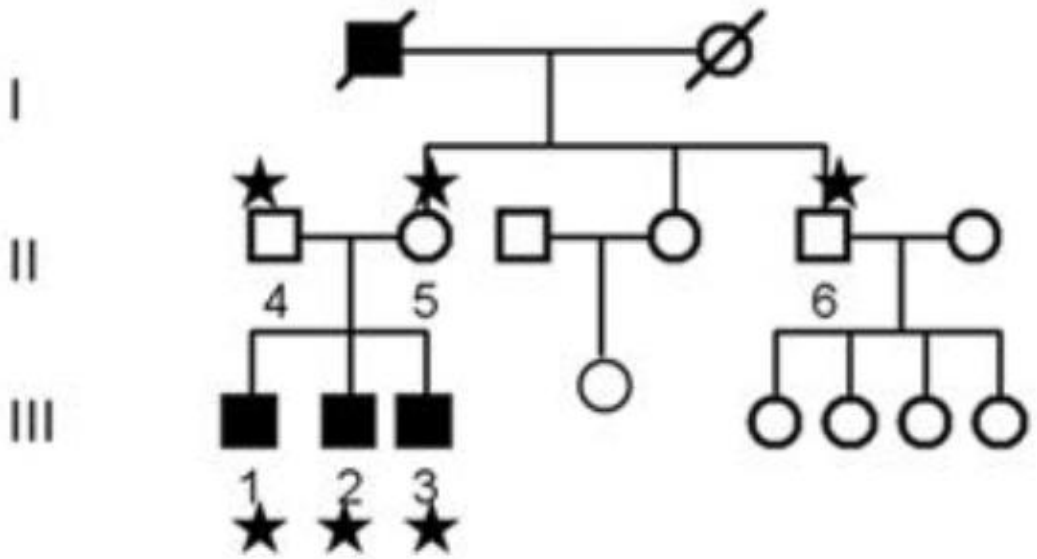
Rare variants <1% in the population

Rare genetic disease

common variants > 5% in the population

Common genotypes (e.g. diabetics)

- Common phenotypes might be a result of rare variations, in the same gene
- Every genome harbors many rare benign variations



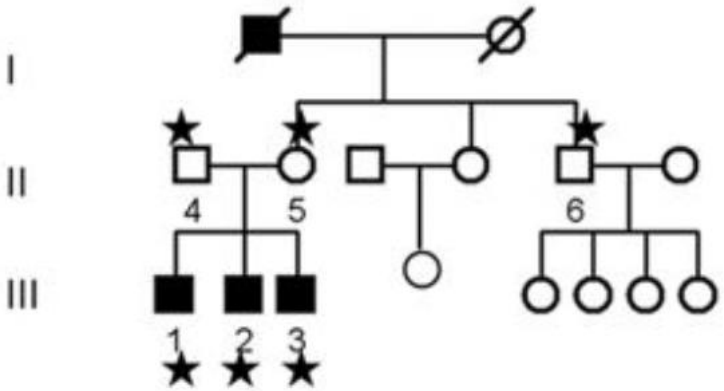
★ Exome sequencing

After variant calling (including joint calling) and filtration we obtained ~80,000 variations

What next?

Alkelai A. et. al. Eur J Hum Genet. 2017 Dec;25(12):1377-1387, Previously: Ester Feldmesser

Congenital anosmia- absence of sense of smell from birth



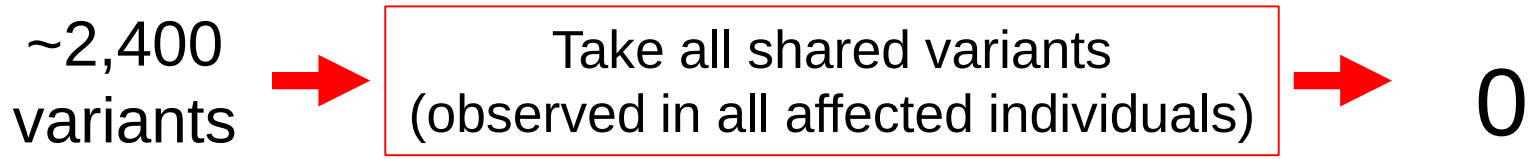
What next?

~80,000 variants → Annotation pipeline → Filter:
Coding variations that change the protein

Rare in the population >1%

→ ~2,400 variants

Filter based on genetic model



Genetic models:

Recessive => shared homozygote variants
X-linked => a shared mutation on chr. X
Compund => 2 hetrozygote mutations in the same gene, shared

Genetic model: of dominant with partial penetrance

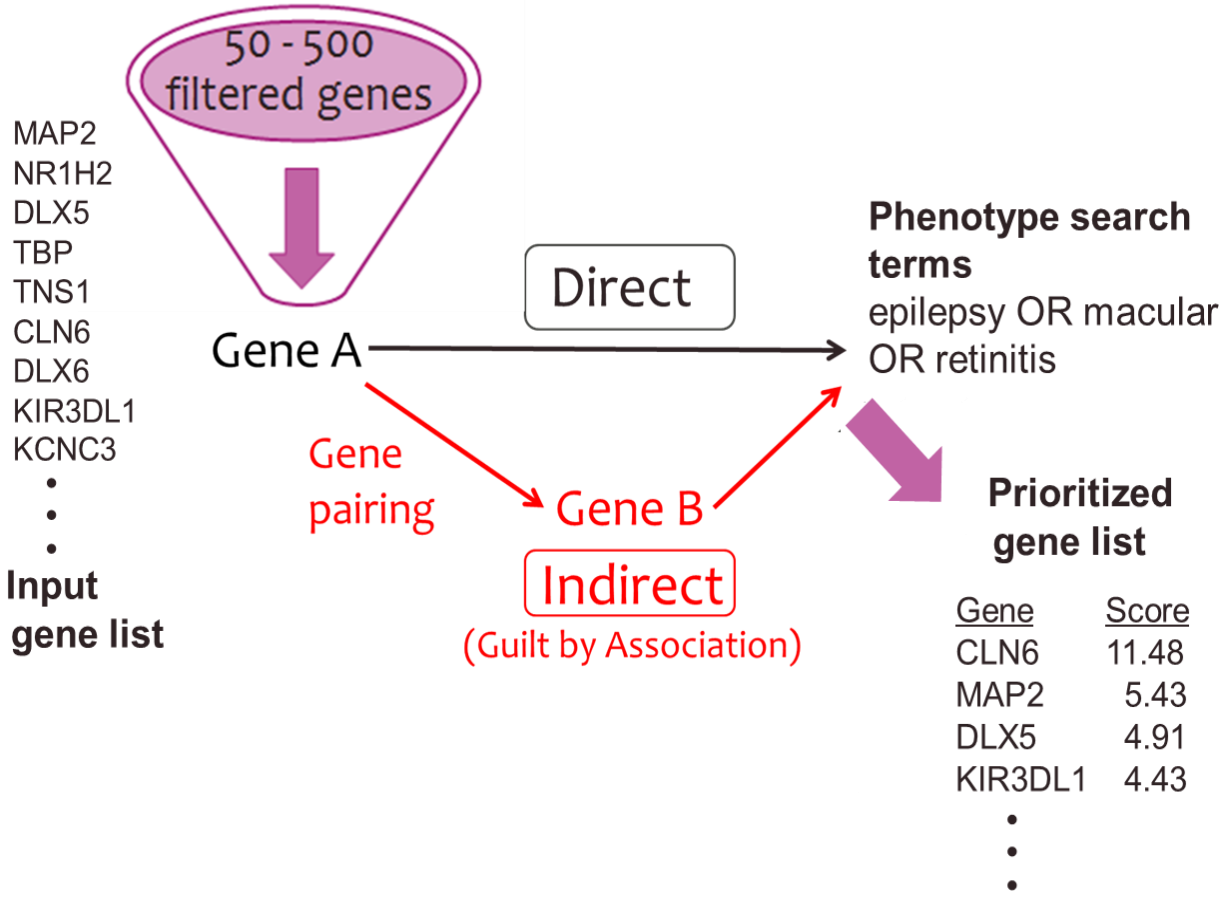
All patients must be heterozygote to the mutation

An individual can carry the mutation without showing the phenotype

~240 mutations

Use variant prioritizer

NGS interpretation using VarElect

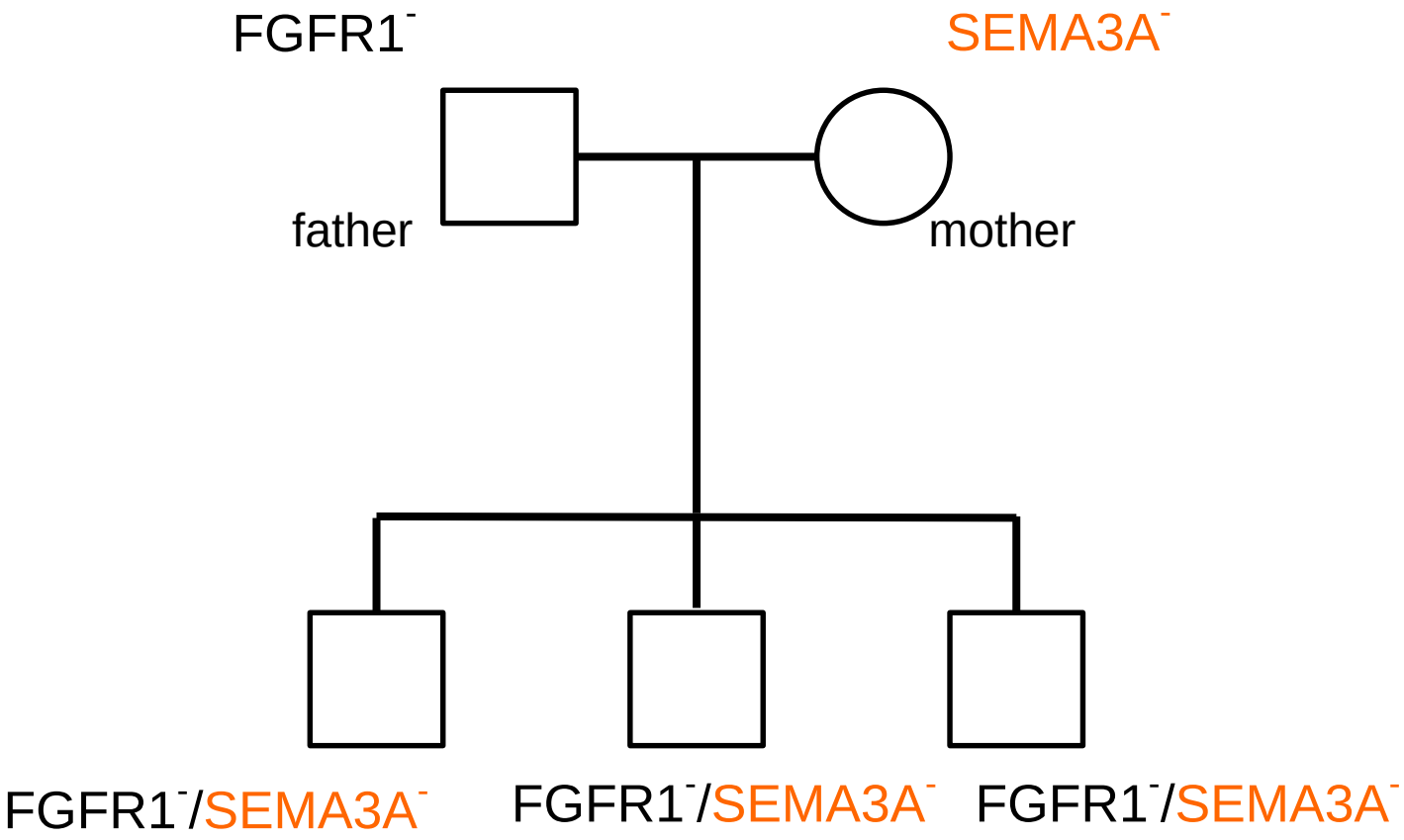


GeneCardsSuite

VarElect
NGS PHENOTYPER

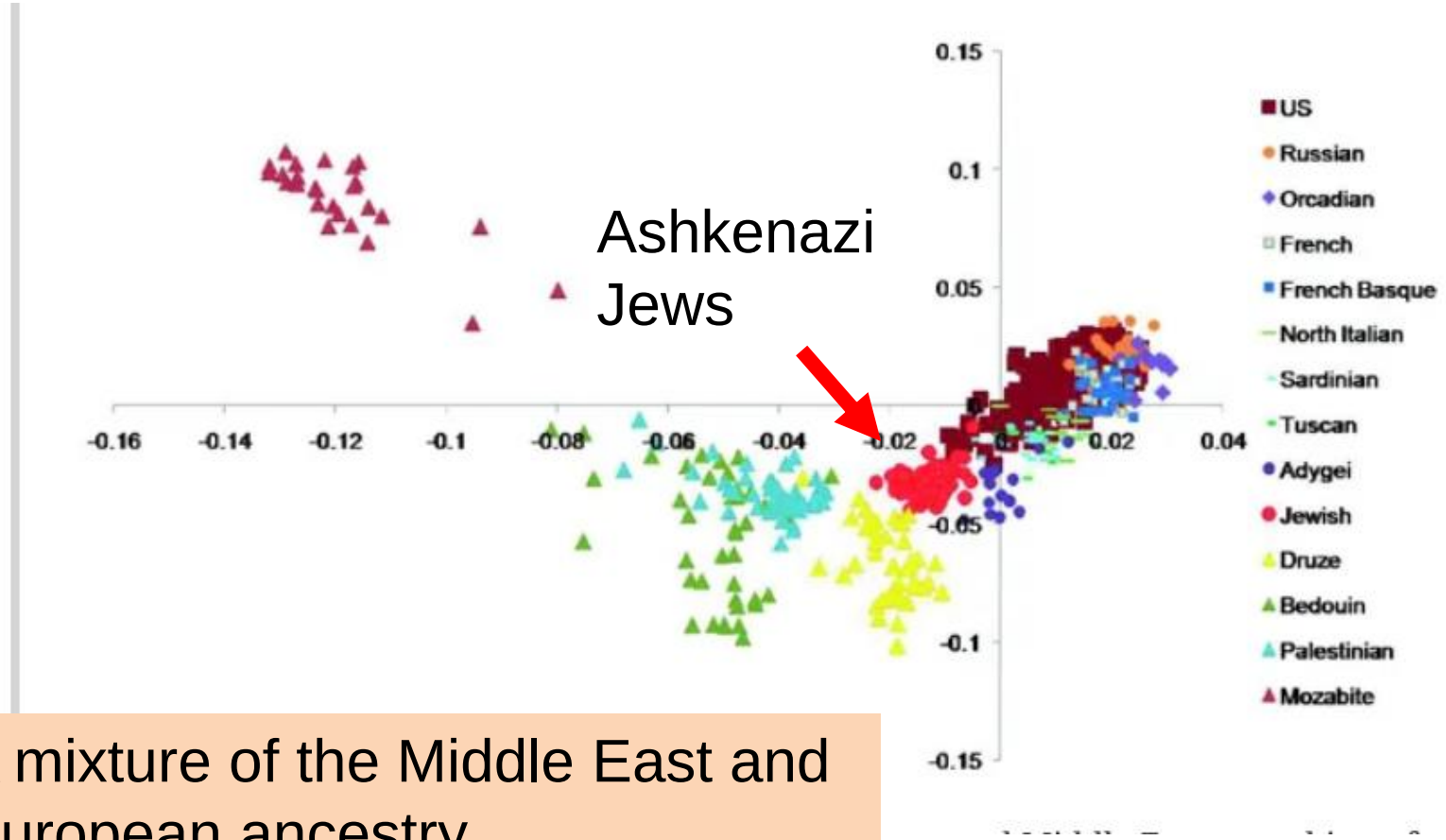
Stelzer et al2016

A di-genic mode of inheritance



The Ashkenazi Jewish genome

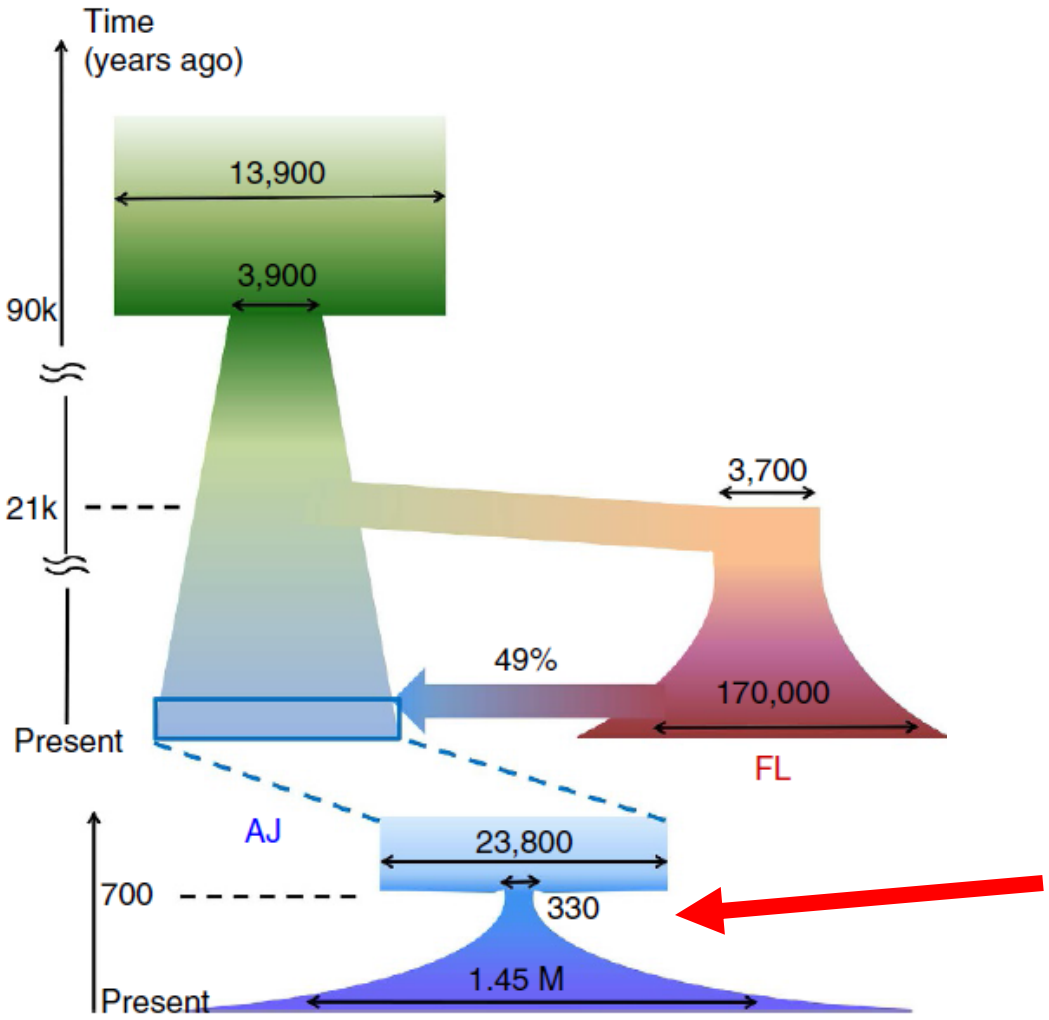
Whole genome sequencing of 128 healthy Ashkenazi Jews



A mixture of the Middle East and European ancestry

Nature communication, 2014

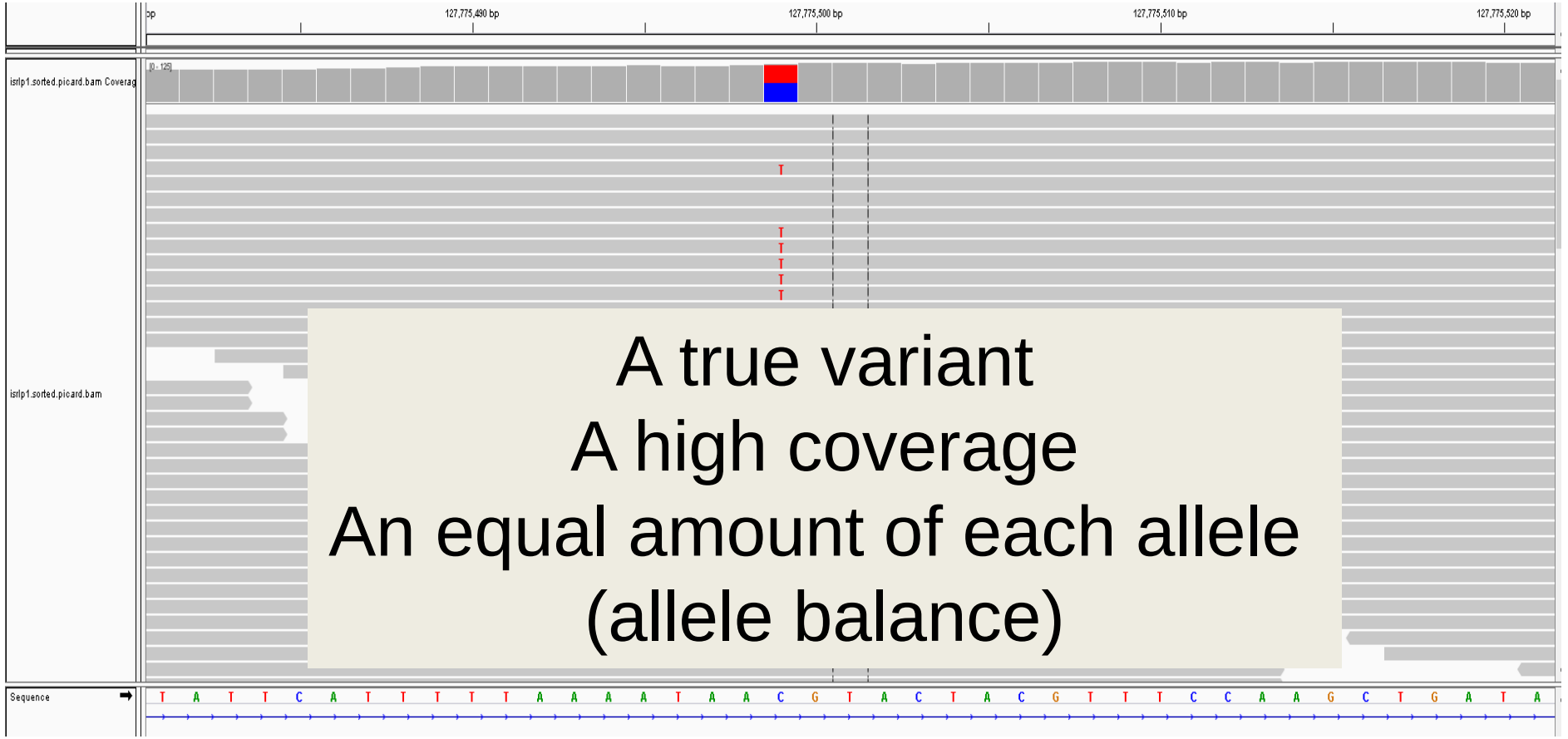
Reconstructing the AJ population history



A genetic bottleneck (catastrophe) that took place ~700 years ago, where only ~300 AJ were left

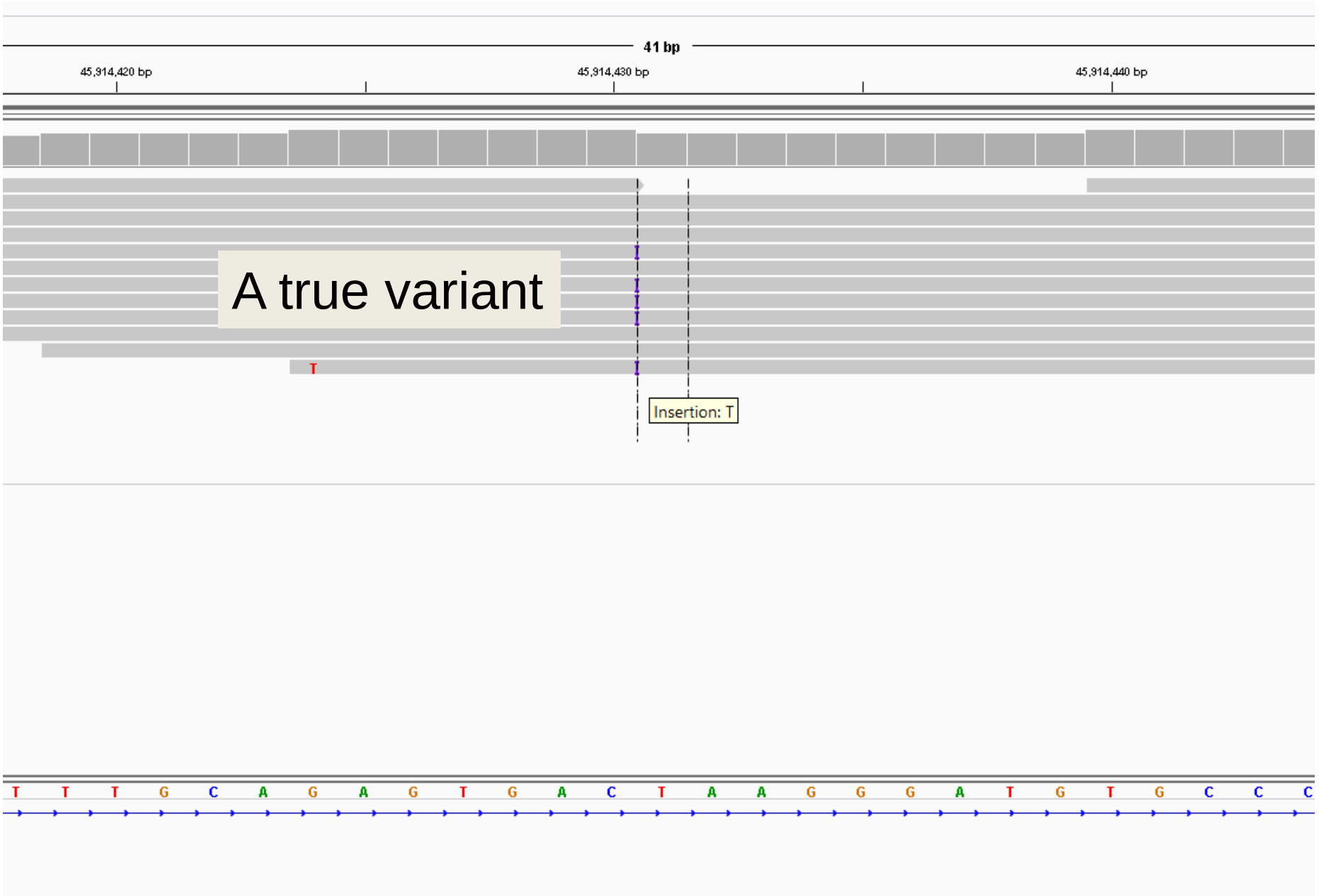
quiz

Example1:
A heterozygote change from C -> T
(54 calls A, 53 calls T)



Example2:
Homozygote deletion of A

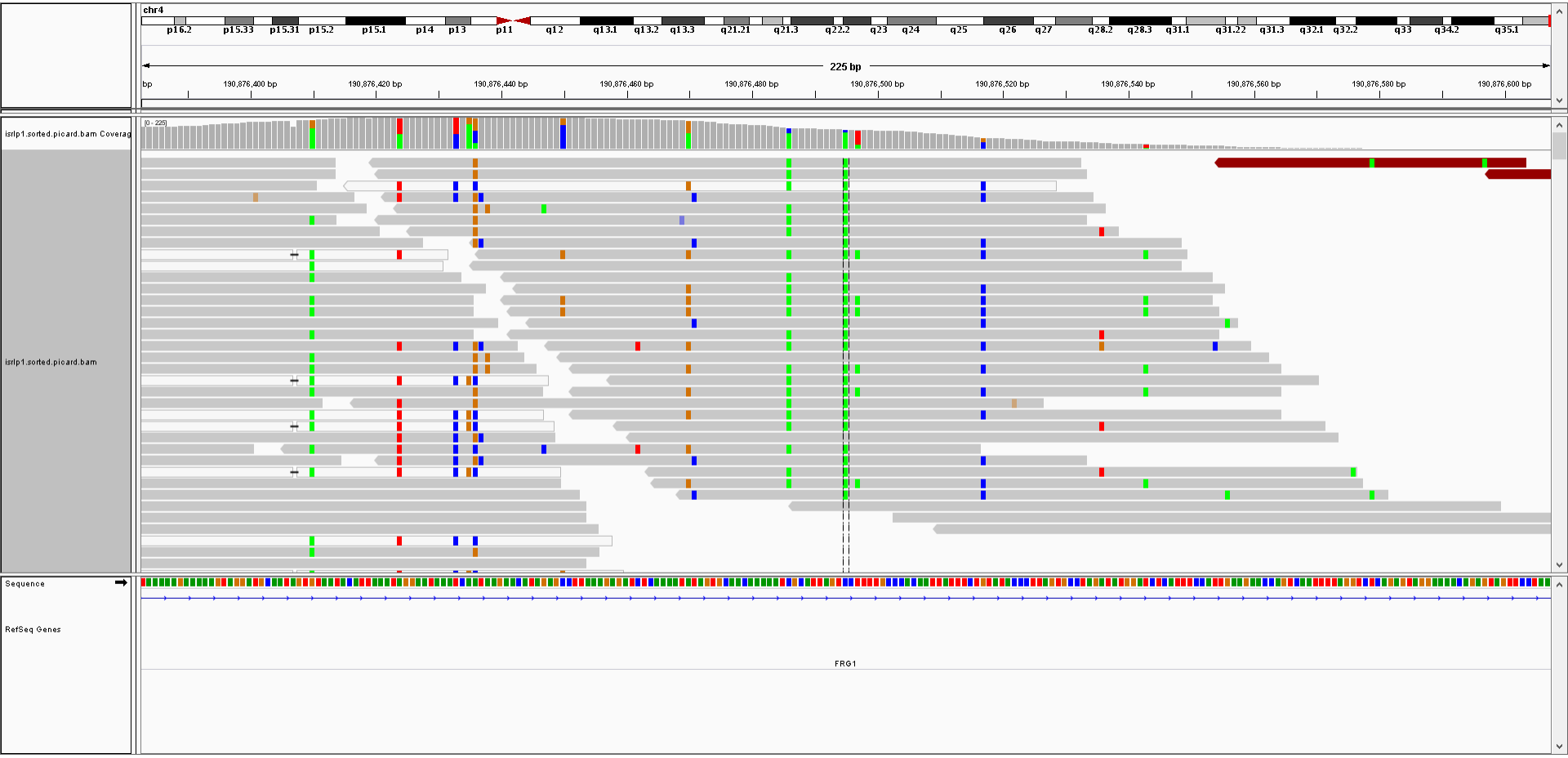
Example3: Hetrozygote insertion



Example4:
A change from C -> A
(116 calls A, 18 calls C)
True or False?

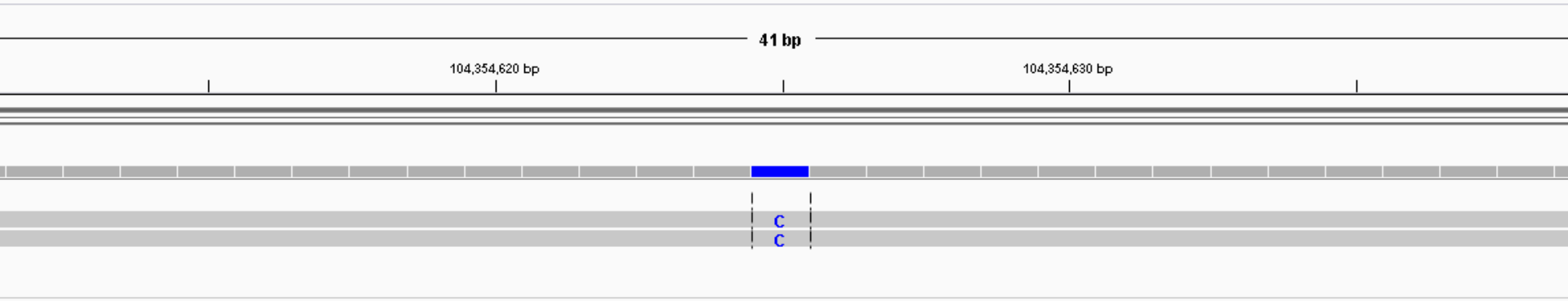
False!!
All reads are the same direction ➔ strand bias
The quality of the variation is only 3

Zoom out of the same region



Example4: hetrozygote deletion
True or False?

Example 5:
Both SNP quality and genotype quality are high



Was called as heterozygote to the reference
Low coverage positions have a higher degree of false positive