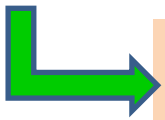
VARIANT DETECTION USING NEXT GENERATION SEQUENCING

Tsviya Olender

June 2017

Applications

- Detection and/or genotyping of disease causing mutations-
Genetic diseases, cancer



Clinial application

- Evolutionary studies

- Detecting RNA modification

- Allelic exclusion

A 20 months old baby

Normal development until 15 mo-o

At age 16-mo she showed a reduction in her vision together with unsteady gait. She was treated with corticosteroids

At age 19-mo she had shown no response to the steroid treatment. She showed rapid progressing neurodegeneration and with possibility of an autoimmune disorder

A genetic analysis of her exome identified a compound heterozygous mutations in the gene SLC52A2 (vitamin B transporter), involved in Brown–Vialetto–Van Laere Syndrome 2.

Following the discovery she was treated with riboflavin, symptoms were improved dramatically.

Petrovski S et. al. Cold Spring Harb Mol Case Stud. 2015

Type of genetic variations

SNPs

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

√

InDels

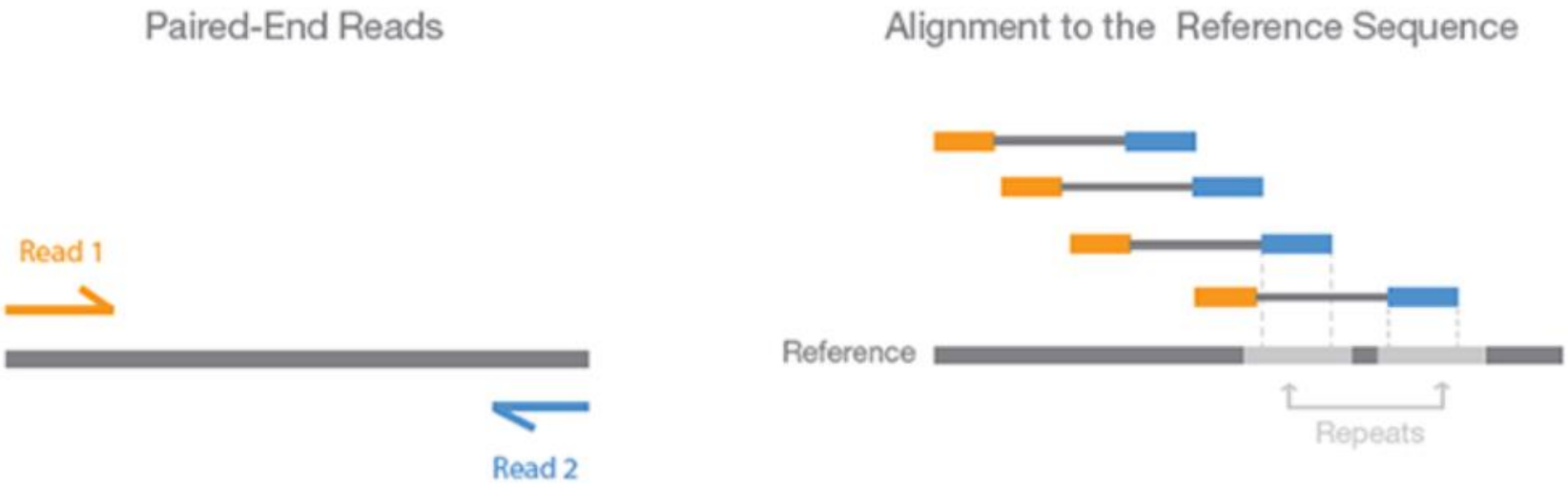
ind1 **ACAATCGCCA**
ind2 **ACA---GCCA**
ind3 **ACA---GCCA**
ind4 **ACAATCGCCA**

~ half of the read length

Structural variations

Deletions, duplications, insertions, inversions, translocations

Paired-End sequencing



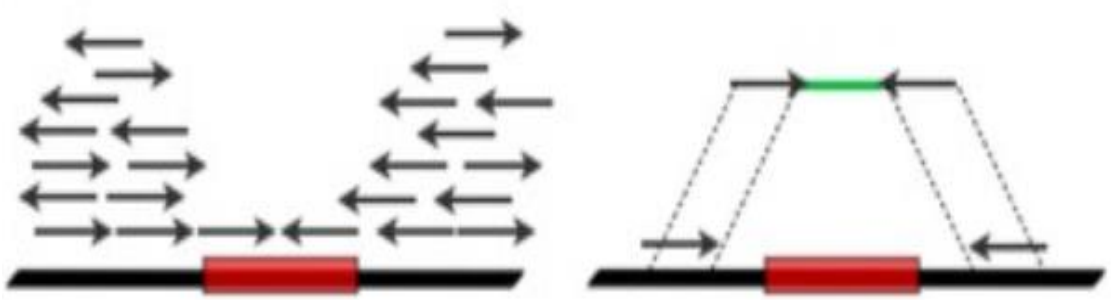
The distance between the paired-end reads is known => accurate alignment

Calling structural variations

- Paired-ends is usually a must

coverage Read pair distance

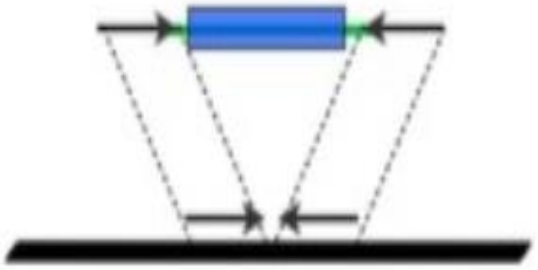
deletion



Reference genome

X

insertion



Reference genome

Lumpy, DELLY2, BreakDancer, Pindel…

Tattini L et al. Front Bioeng Biotechnol. (2015)

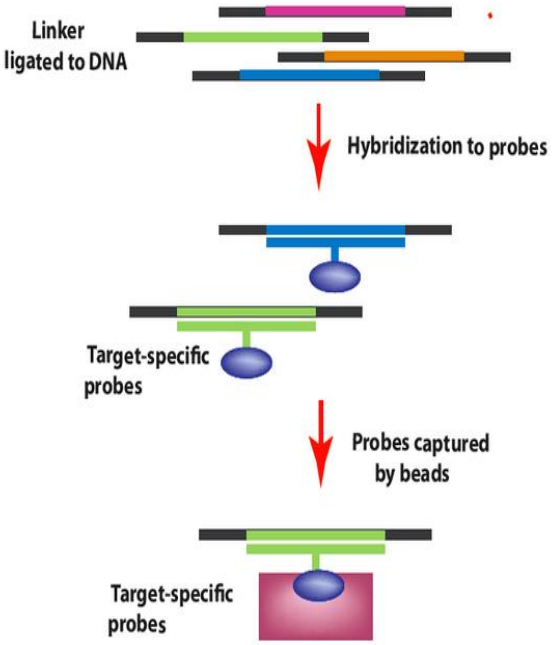
Exception:
Homozygote deletions and gene copy number can be detected by coverage analysis.

PacBio
Allows an access to a broad range of structural variations

Whole genome sequencing versus targeted sequencing

Targeted sequencing-
next-generation DNA sequencing technique focused on amplicons and specific genes.

e.g. exones (Exome sequencing), panels of disease causing genes or cancer genes…



Cheap, focused of regions of interest, can be applied to whole population

A whole human exome ~300-600$

Whole genome sequencing-
More expensive

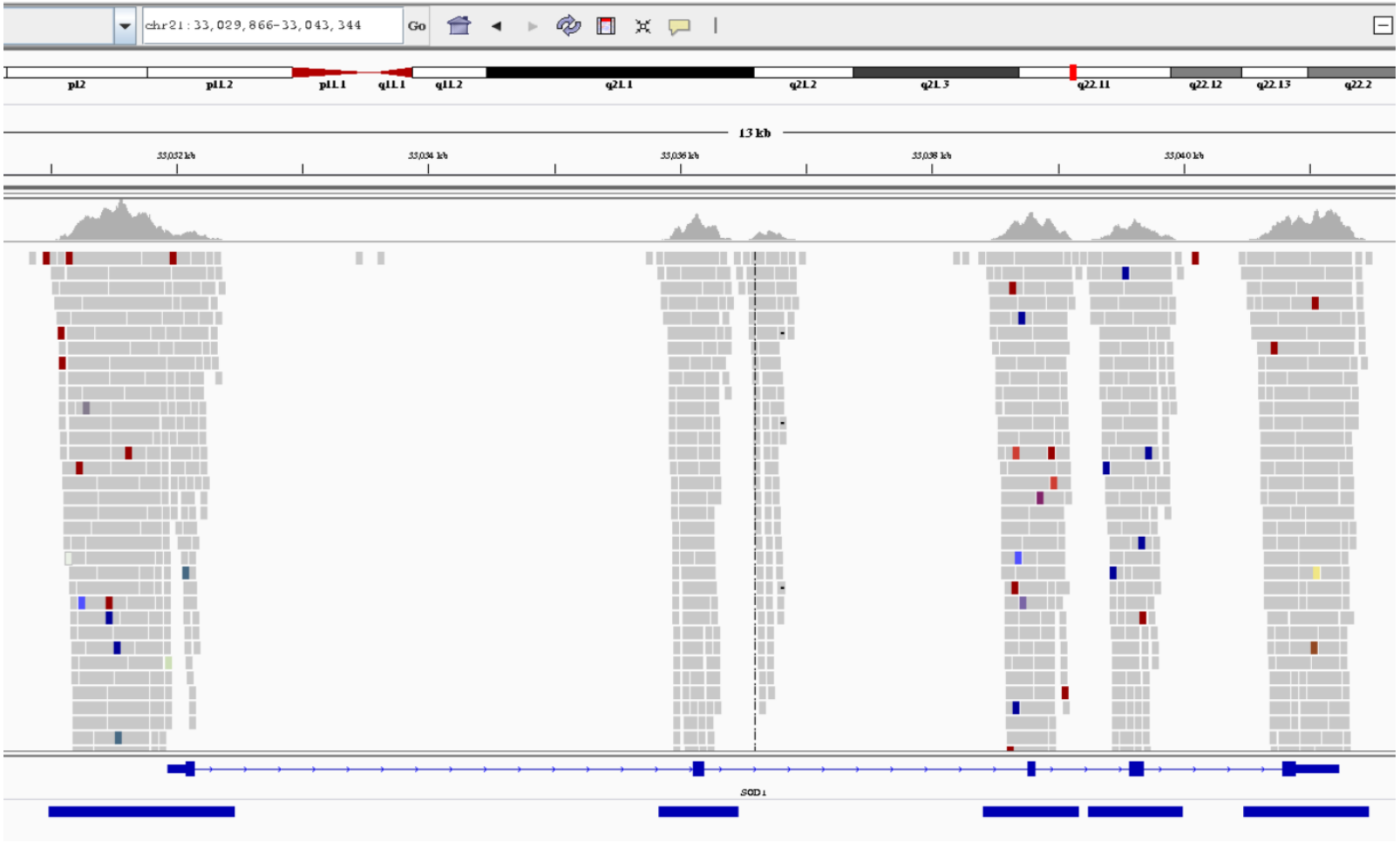
Human genome ~1000$

The false positive calling rate is much lower, the coverage across the genome is uniform.
Quantitative- can be used to call structural variants.

Example: targeted sequencing of the human SOD1

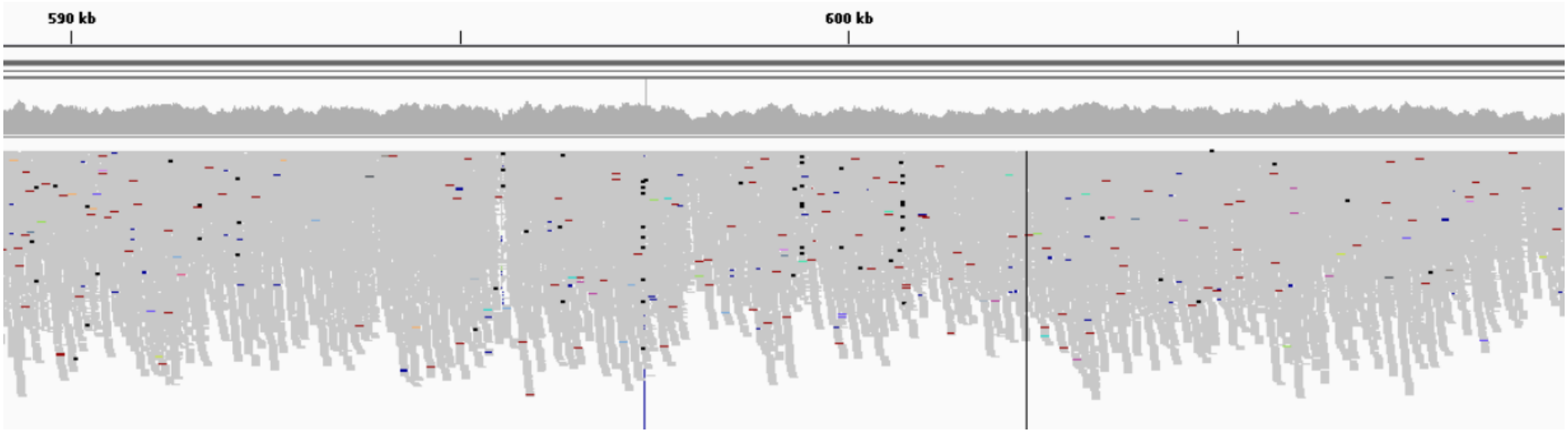
Coverage

Aligned reads



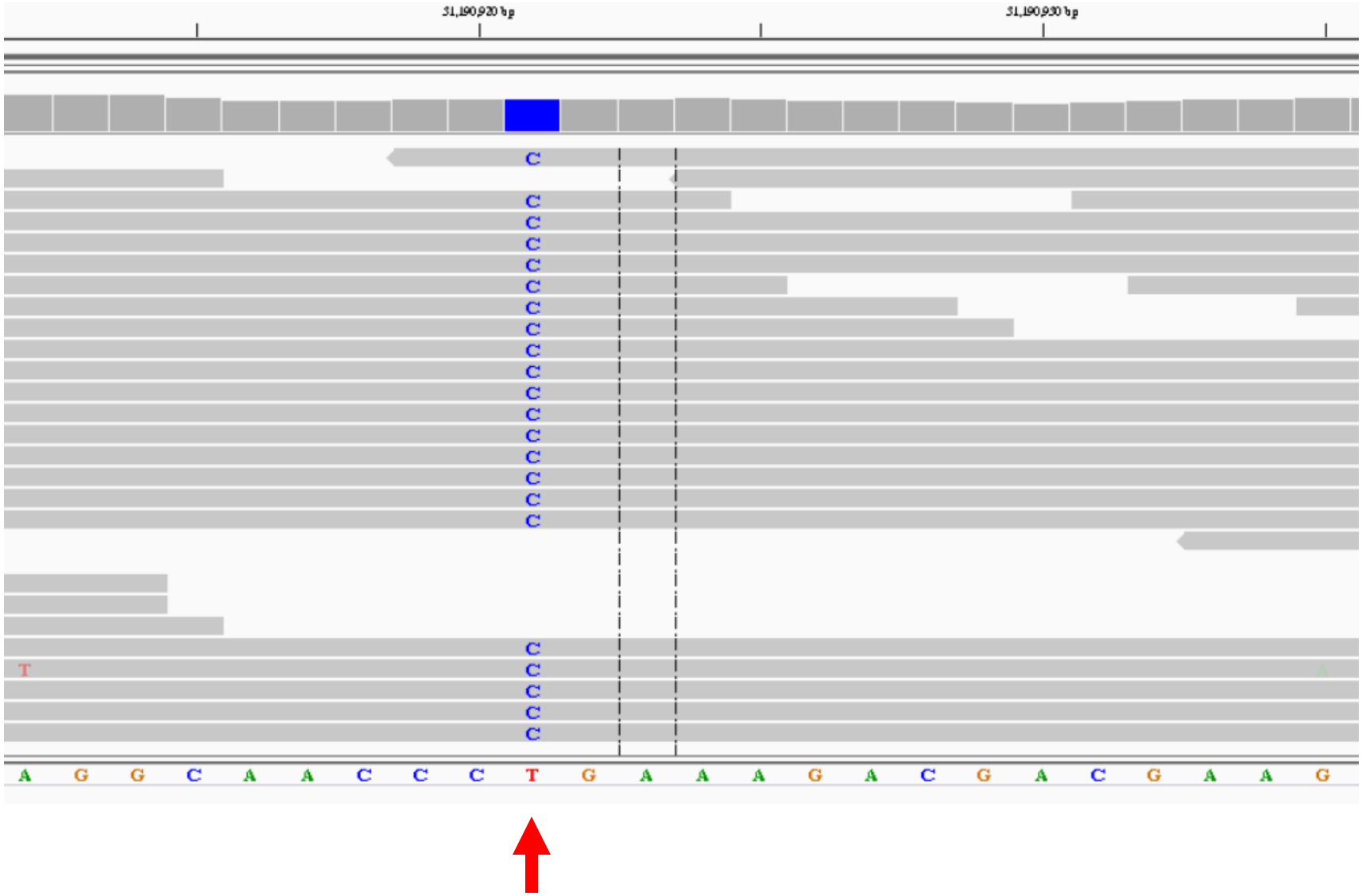
Project of:
Chen Eitan
Eran Hornstein

Example: whole genome sequencing of S. cerevisiae

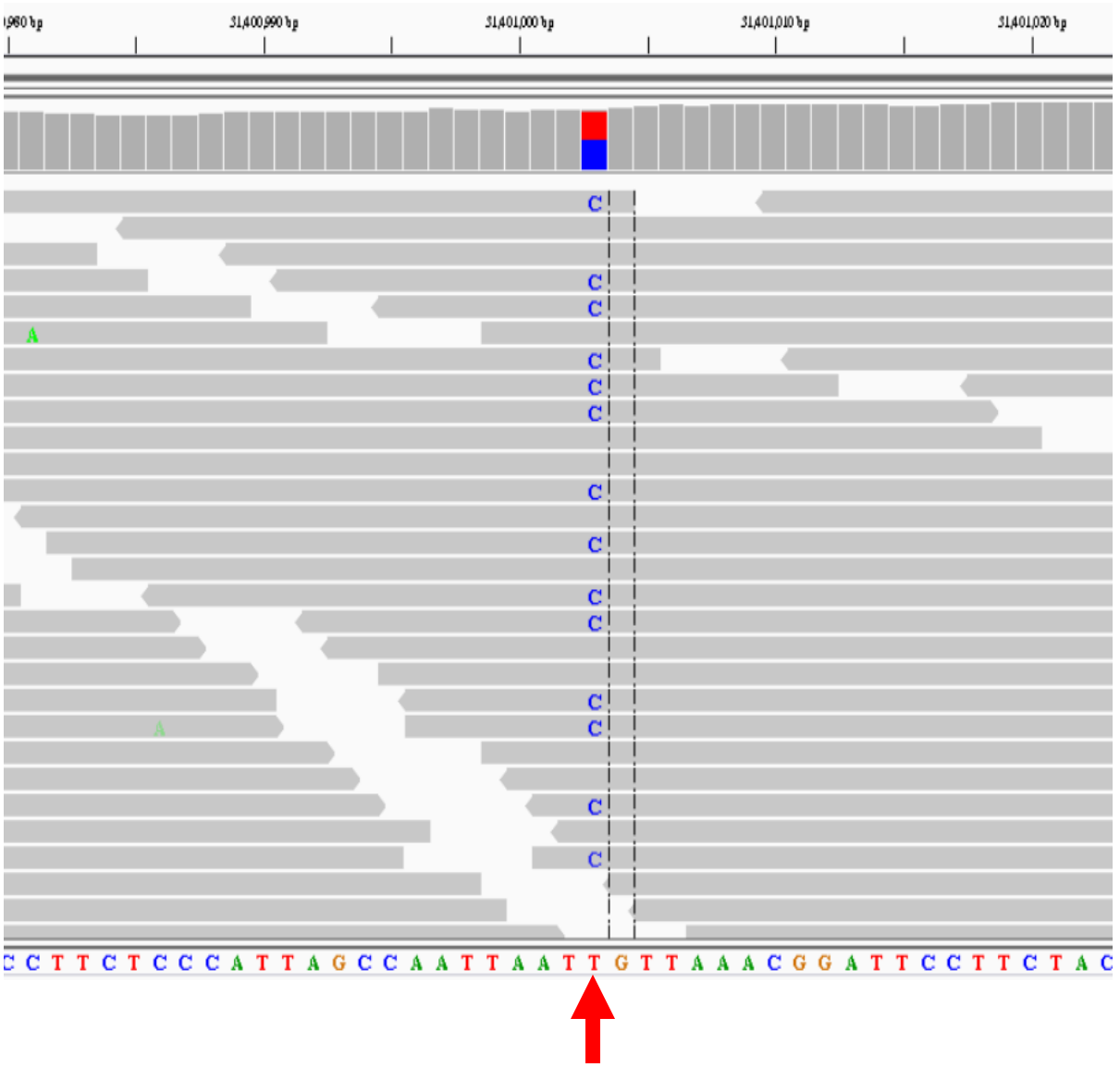


Project of:
Dana Bar Zvi
Naama Barkai

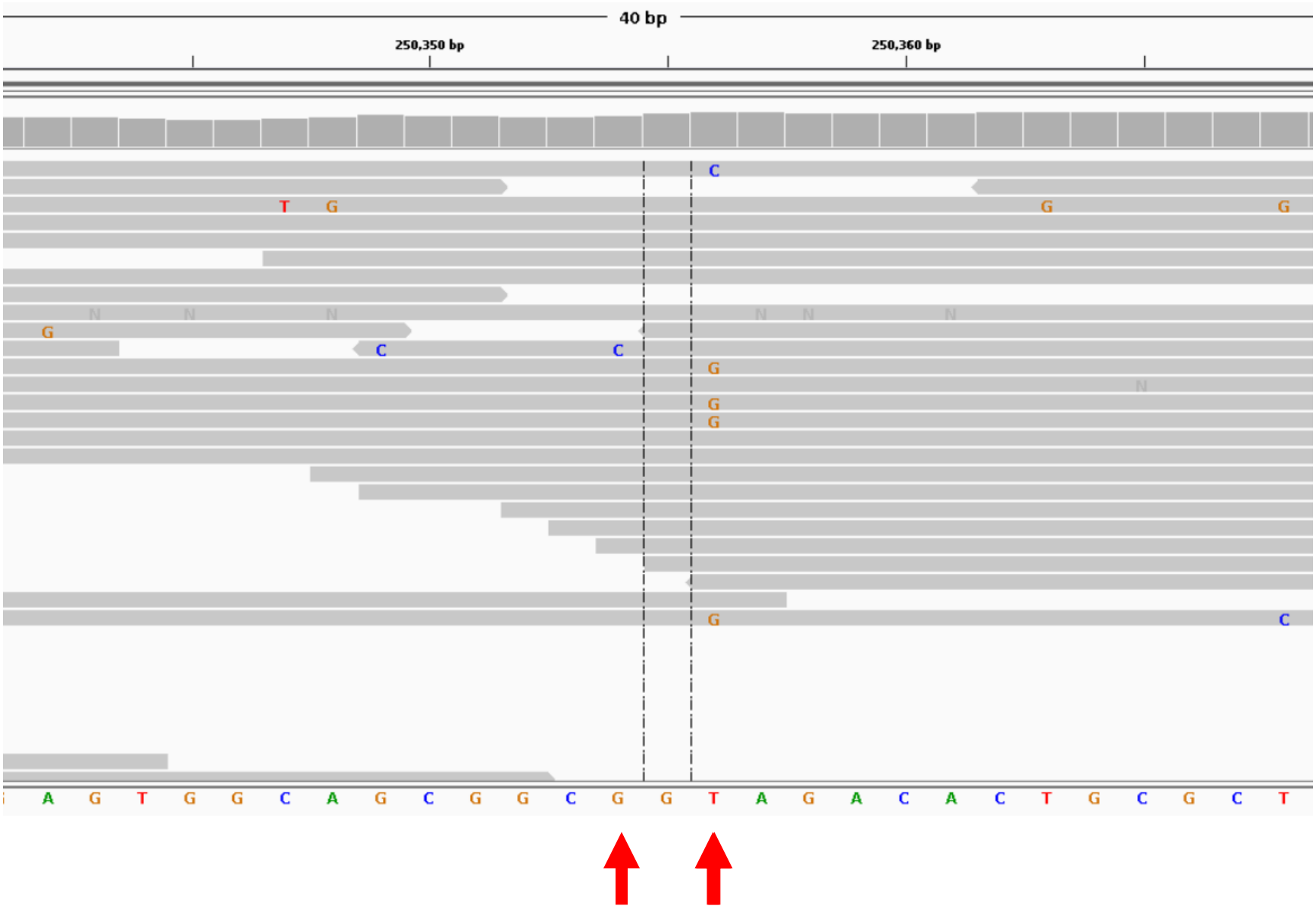
An example of homozygote call



An example of heterozygote call



Sequencing errors lead to false positive calls



Sequencing depth == power to detect variants

Sequencing considerations

- Reduce the false-positive calling rate ➔ High coverage
- A GC bias that is introduced during DNA amplification by PCR

Recommendations:

- Paired-ends sequencing, ~100bp long ➔ allows to achieve the required coverage with more accurate alignment

- The required coverage: depends in your goals.

rule of thumb (for detecting rare variants):

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

Algorithms for variant detection

Samtools

Freebayes

GATK

Varscan

GlfMultiples

Cortex

GATK is the most sensitive software

Warden CD, et al *Peer J.* 2014
Pirooznia M , et al *Hum Genomics*. 2014
Liu X et. al *Plos One*. 2014

gatk

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.

GATK Best Practices

Recommended workflows for variant discovery analysis with GATK

What are the Best Practices?
Analysis phases
Scope and limitations
What is *not* Best Practices?
Beware legacy scripts

Reads-to-variants workflows used at the Broad Institute.

The GATK Best Practices provide step-by-step recommendations for performing variant discovery analysis in high-throughput sequencing (HTS) data. There a several different Best Practices workflows tailored to particular applications depending on the type of variation of interest and the technology employed. The Practices documentation attempts to describe in detail the key principles of the processing and analysis steps required to go from raw reads coming off the sequencing machine, all the way to an appropriately filtered variant callset that can be used in downstream analyses. Wherever we can, we try to provide guidance regarding experimental design, quality control (QC) and pipeline implementation options, but please understand that those are dependent on man factors including sequencing technology and the hardware infrastructure that are at your disposal, so you may need to adapt our recommendations to your specific situation.

GERMLINE		SOMATIC	
SNPS & INDELS	COPY NUMBER	SNVS & INDELS	COPY NUMBER
EXOME/PANEL + WGS BWA + HaplotypeCaller GVCF	EXOME/PANEL In development	EXOME/PANEL + WGS BWA + MuTect	EXOME/PANEL BWA + CalISegments
RNASEQ STAR + HaplotypeCaller	WHOLE GENOME In development	EXOME/PANEL + WGS BWA + MuTect2 BETA	WHOLE GENOME In development

BROAD INSTITUTE
© Broad Institute 2016

Search

Geraldine_VdAuwera

Folks, I'm sorry but I can't respond to questions on this page. Please post your questions in the public forum.

About

Username Geraldine_VdAuwera Location Cambridge, MA Joined July 2012 Visits 5,358 Last Active 1:57AM Broadie Points 7,487 Badges 25

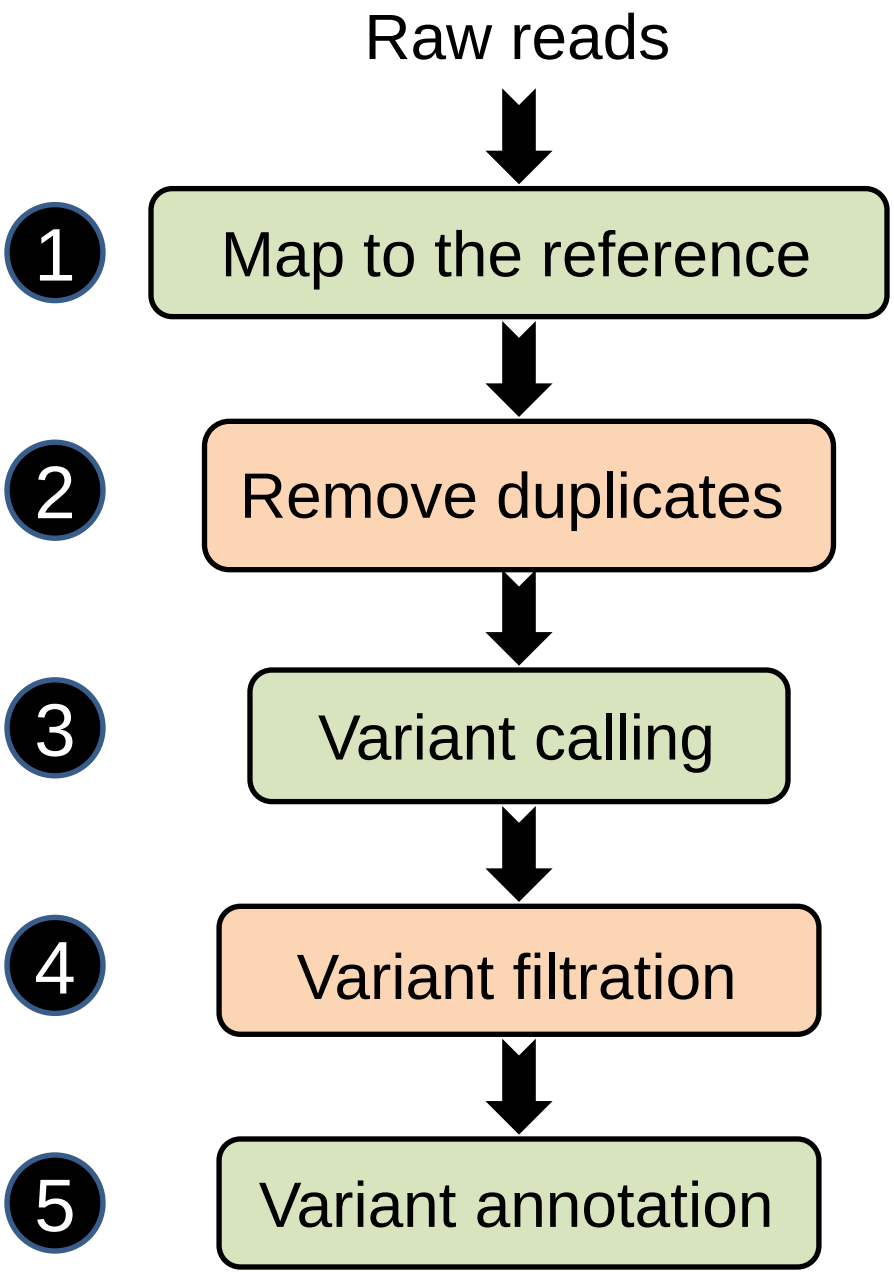
Reactions

21	704
Disagree	Like

- Activity
- Discussions 513
- Comments 11249

In case you have questions ask Geraldine

A typical work-flow for variant calling



1 Map to the reference

Most popular: **BWA -Burrows-Wheeler Aligner**
Developed by: Heng Li at the Sanger Institute

- Allows indels and mismatches in the alignment, suitable for variant calling

- Do not apply adapter trimming, and low quality filtration!!

Other algorithm: Novoalign

2

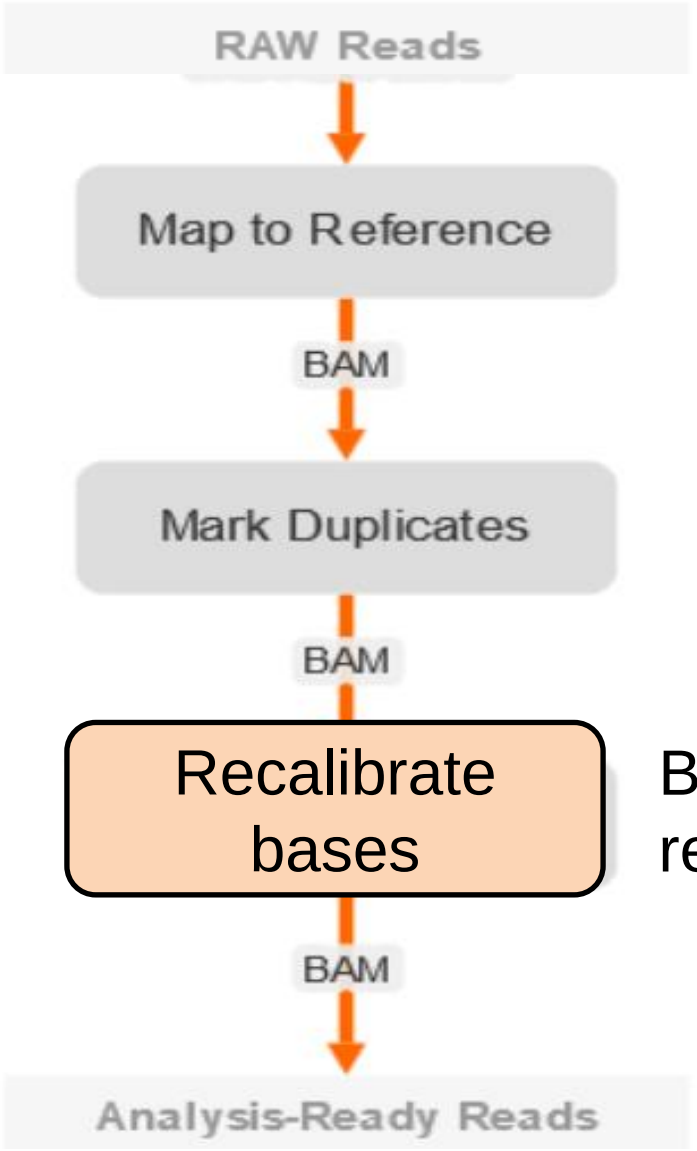
Mark/Remove duplicates

PCR duplicates-
multiple PCR products from the same template molecule binding on the flowcell.

Duplicate reads are not informative and should not be counted as additional evidence for or against a putative variant.

Algorithms:
Picard (most popular), samblaster, samtools

GATK pre-processing step

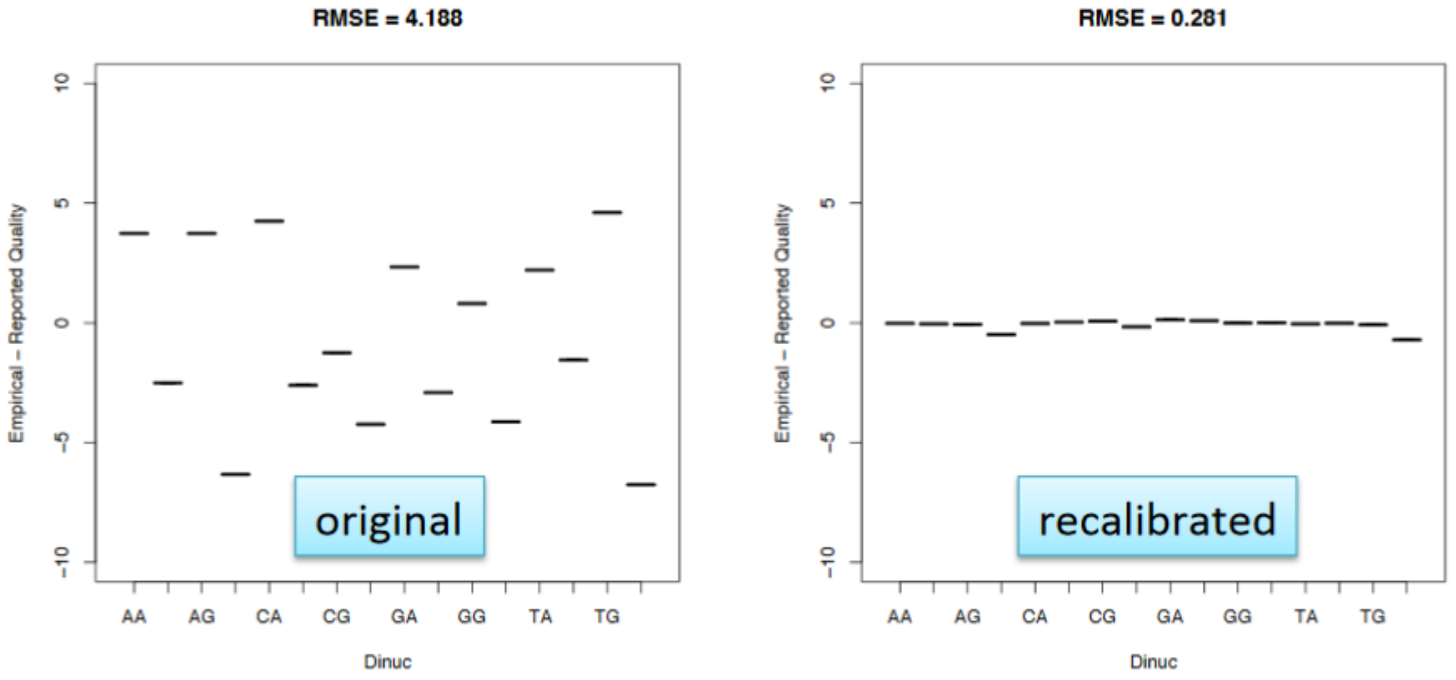


Base quality score recalibration (**BQSR**)

Quality scores emitted by sequencing machines are **biased** and **inaccurate**

- Quality scores are critical for all downstream analysis
- Systematic biases are a major contributor to bad calls

Example of bias: qualities reported depending on nucleotide context



BQSR method identifies bias and applies correction

Base recalibration phase

- Model the error modes

Requires an accurate database of polymorphisms for your organism… human??

Tian *et al. BMC Bioinformatics* (2016) 17:403
DOI 10.1186/s12859-016-1279-z

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Impact of post-alignment processing in variant discovery from whole exome data

Shulan Tian[1], Huihuang Yan[1], Michael Kalmbach[2] and Susan L. Slager[1*]

BQSR had negligible effect on INDEL calling and generally reduced sensitivity for SNP calling depends on caller, coverage and level of divergence.

My recommendation: apply BQSR only when possible, and when working with GATK

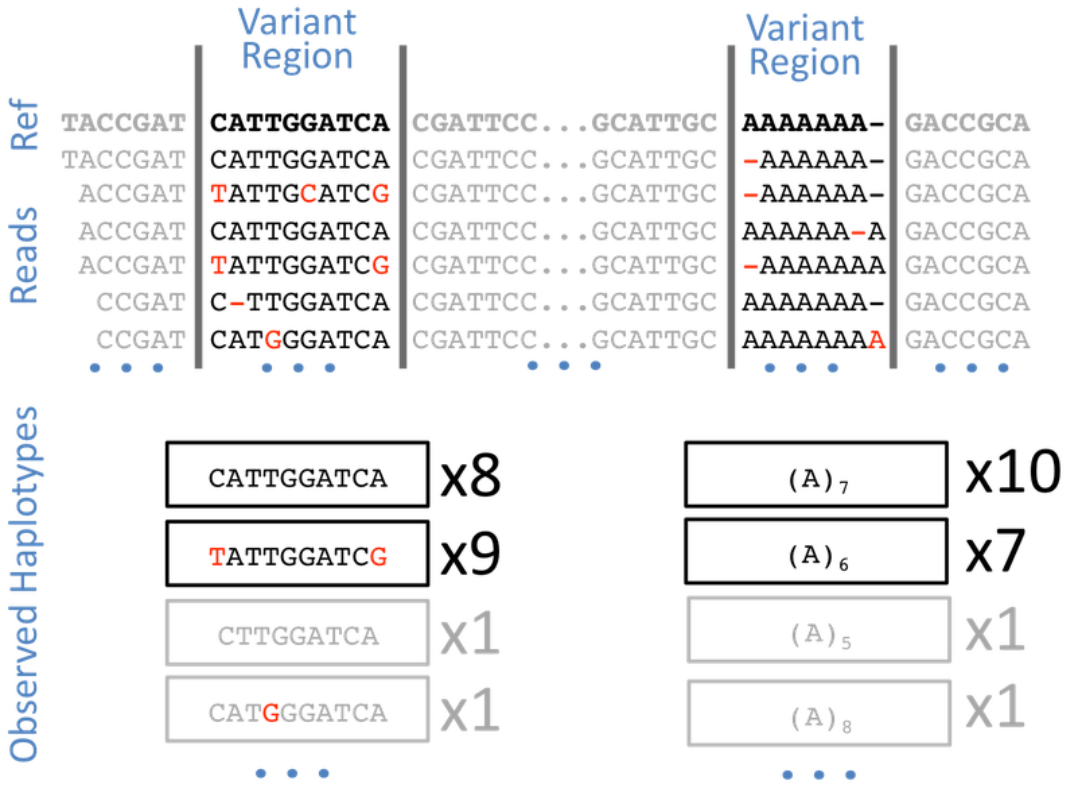
QC metrics

How many reads were aligned to the genome?
What was the coverage along the sequenced genome?

3

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.

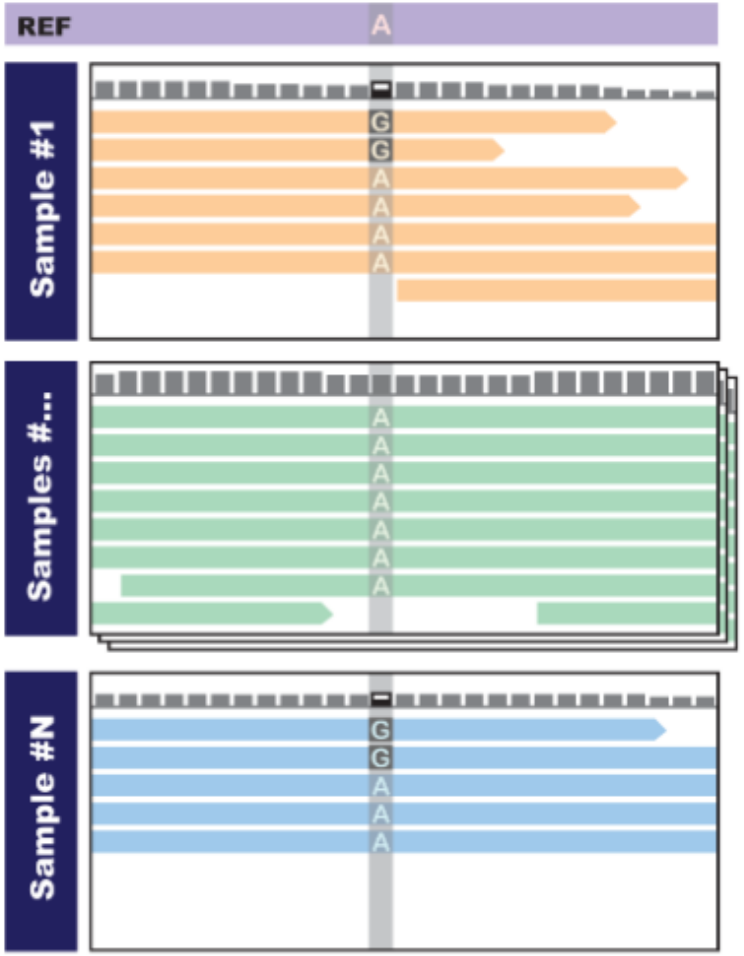
```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=.,Type=Float,Description="Allele Frequency">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
#CHROM  POS     ID      REF  ALT   QUAL  FILTER   INFO              FORMAT    IND1
chr2    4370  rs6057    G    A     29      .      DP=13;AF=0.5      GT:GQ     1/1:43
chr4    7330    .       T    A      3     q10     DP=12;AF=0.017    GT:GQ     0/1:41
```

Joint calling

analyzing cohorts rather than single individual 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



Sample1

chr10	10	A	C	0/1
chr11	200	T	A	1/1
chr11	800	A	G	0/1

Sample2

chr10	10	A	C	0/1
chr12	24	G	A	1/1
chr16	30	A	G	0/1

Joint calling

chr10	10	A	C	0/1	0/1
chr11	200	T	A	1/1	0/0
chr11	800	A	G	0/1	0/0
chr12	24	G	A	0/0	1/1
chr16	30	A	G	0/0	0/1

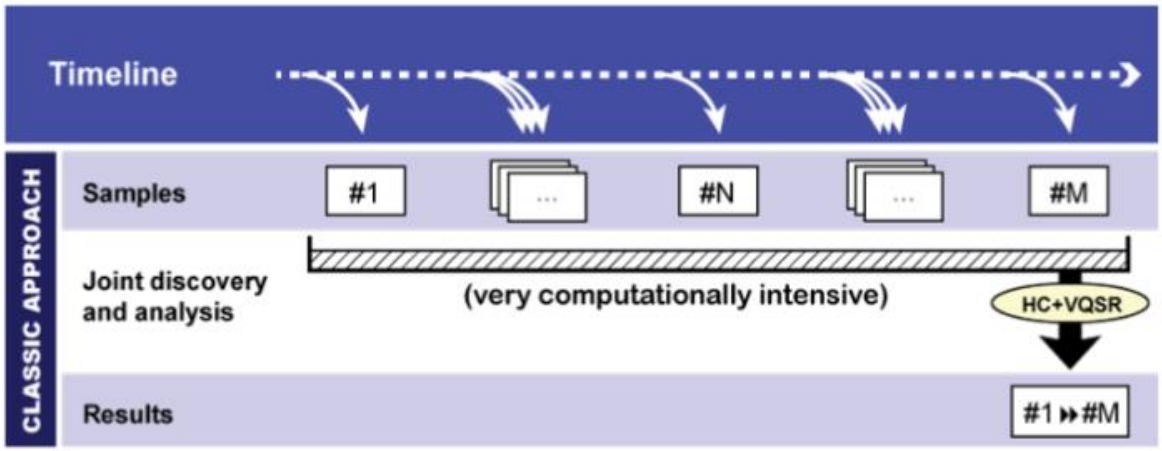
ExAC- the Exome Aggregation Consortium

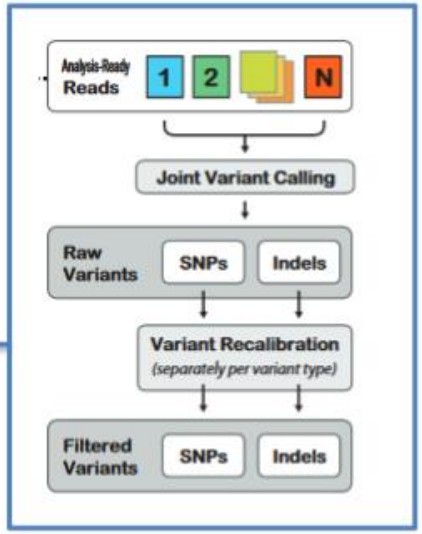
- A catalogue of human variations in the protein-coding regions of the genome

- The catalogue was established using **joint calling of ~65,000 individuals** from across 21 research studies.

The N+1 problem

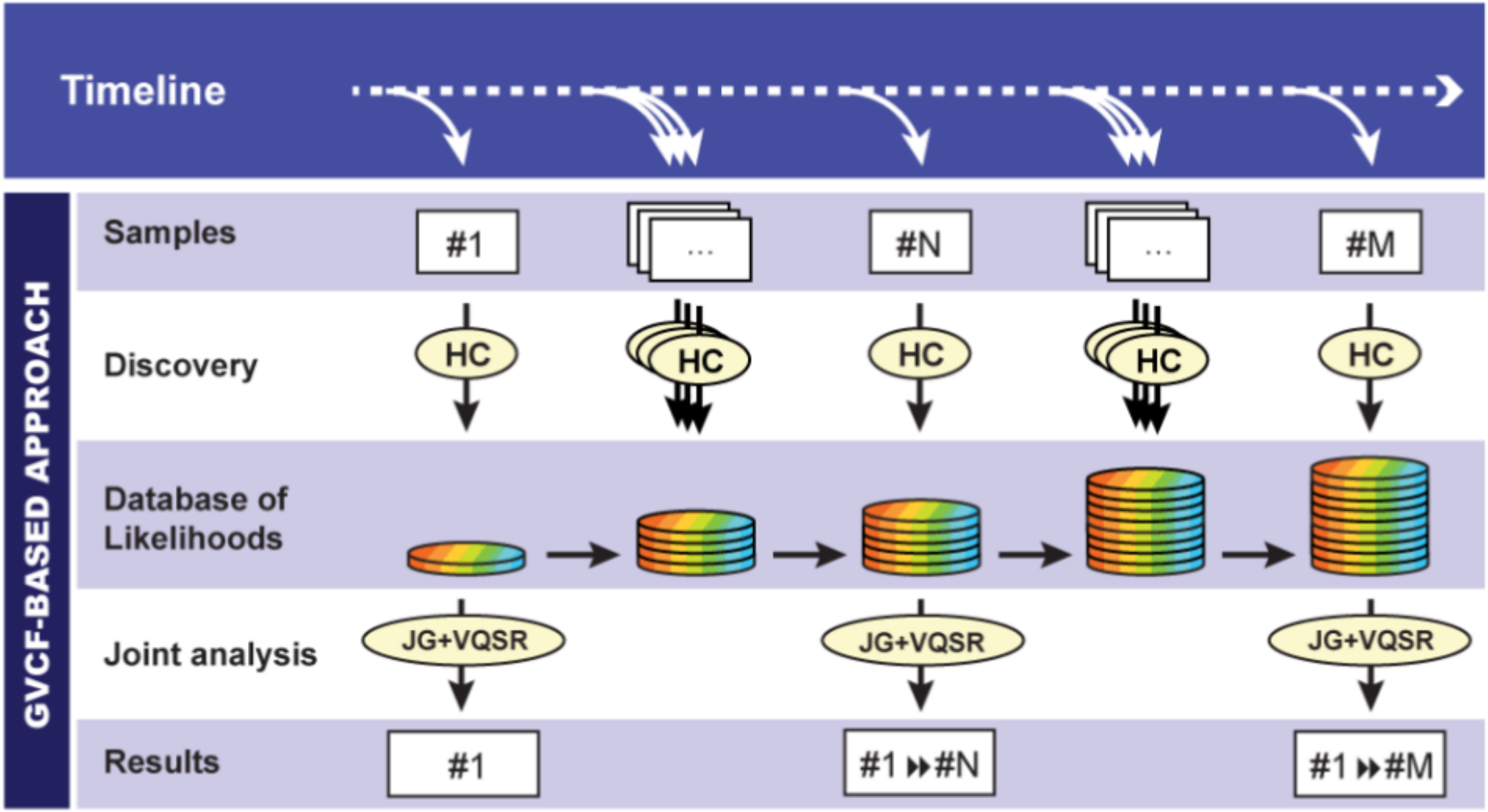
Monkol Lek, 2013





Conceptual solution for the N+1 problem

Genome VCF file == gVCF



Genome VCF (gVCF) Vs. VCF

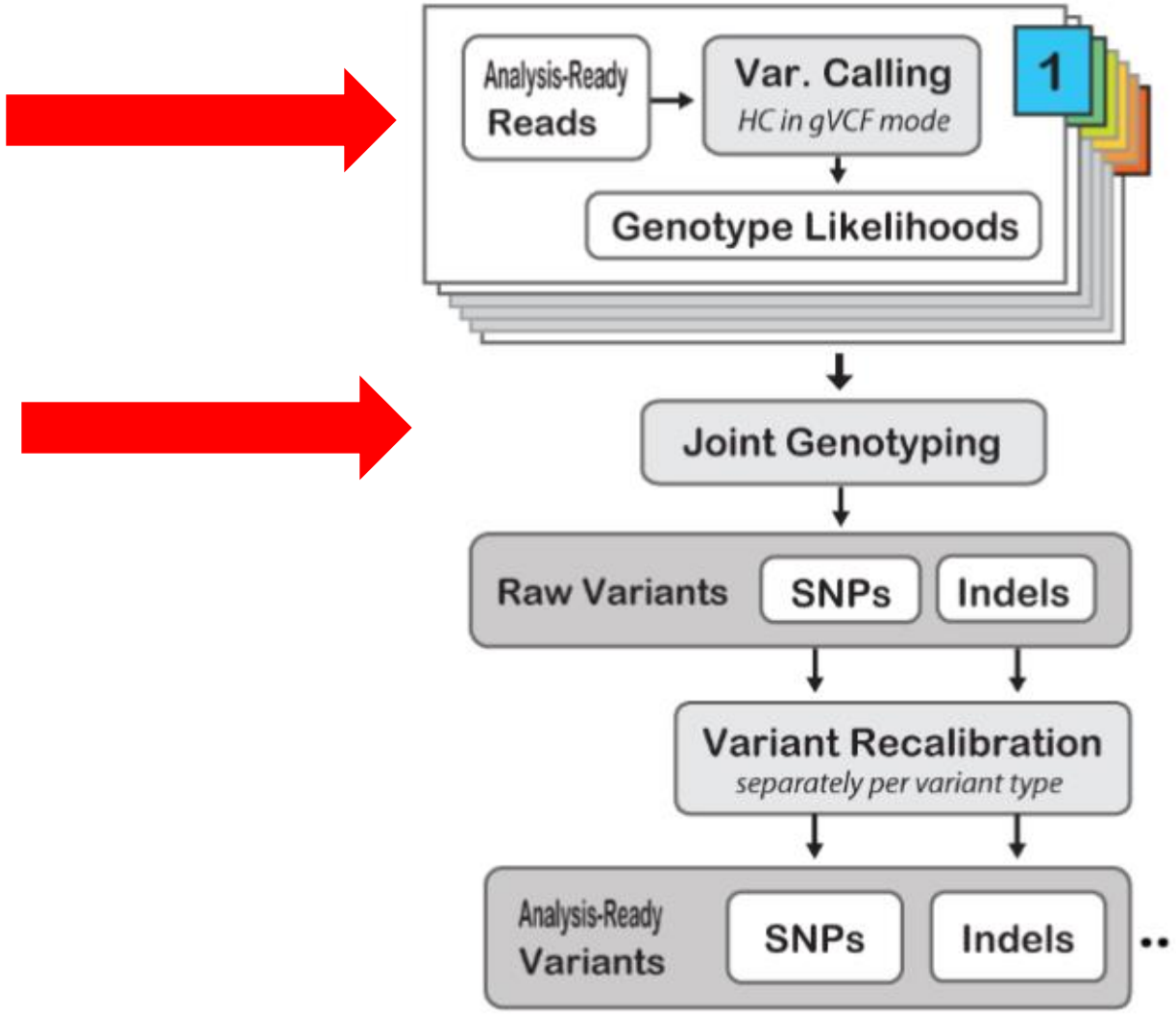
- gVCF file has records for all sites in the genome, whether there is a variant call there or not.
- The records in a gVCF include an accurate estimation of how confident we are in the determination that the sites are homozygous-reference or not.
- Saves a lot of computation time.
- Variant callers that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

GATK, Isaac (of Illumina).

GATK model for joint calling

Generate gVCF file per sample

Join samples



Variant filtration/calibration

- 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.

Variant filtration

Soft filter

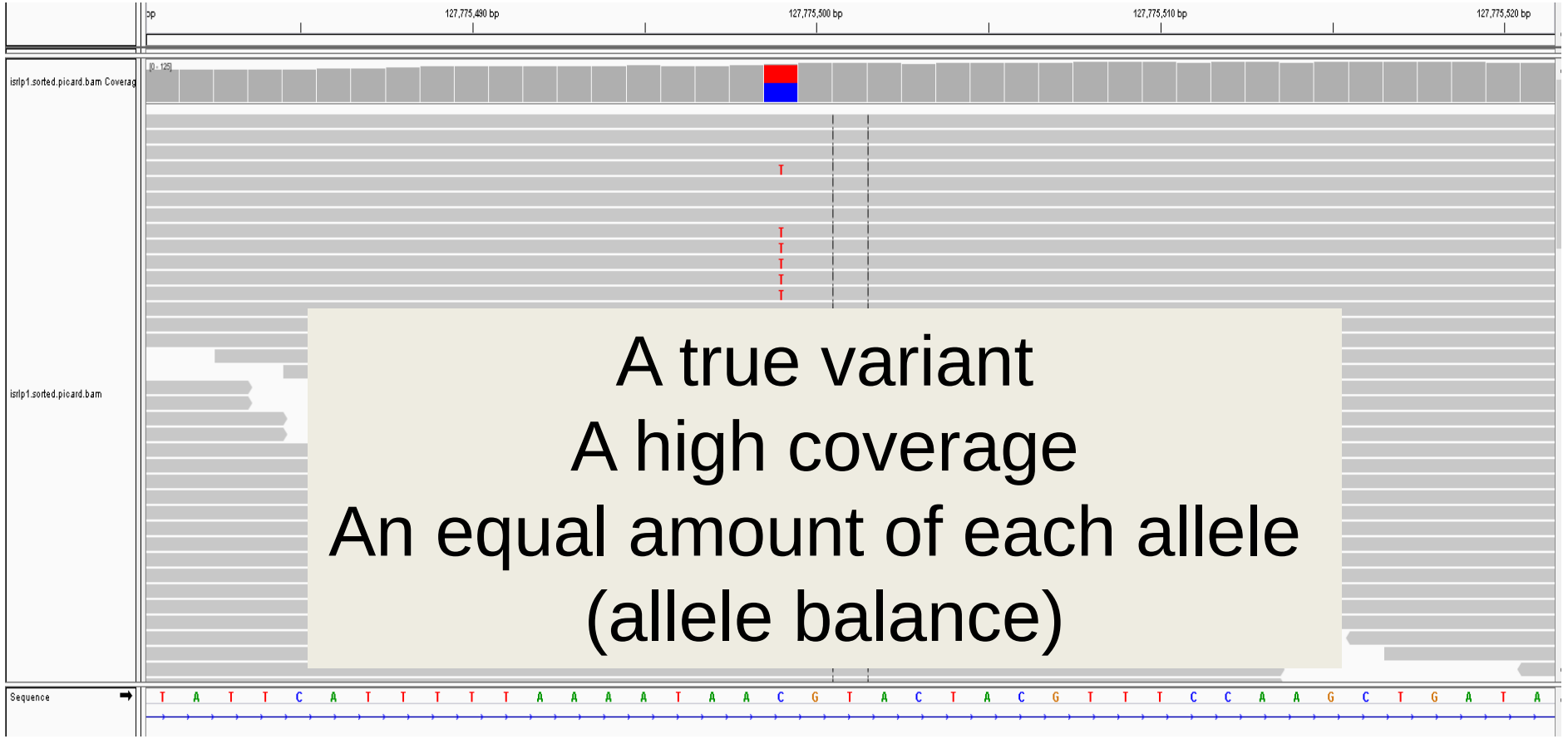
The VQSR (GATK):

- A machine learning algorithm that learns from each dataset what is the annotation profile of good variants vs. bad variants.
- Requires a large data set, and a reliable database of genetic variants. (human??).

hard filter

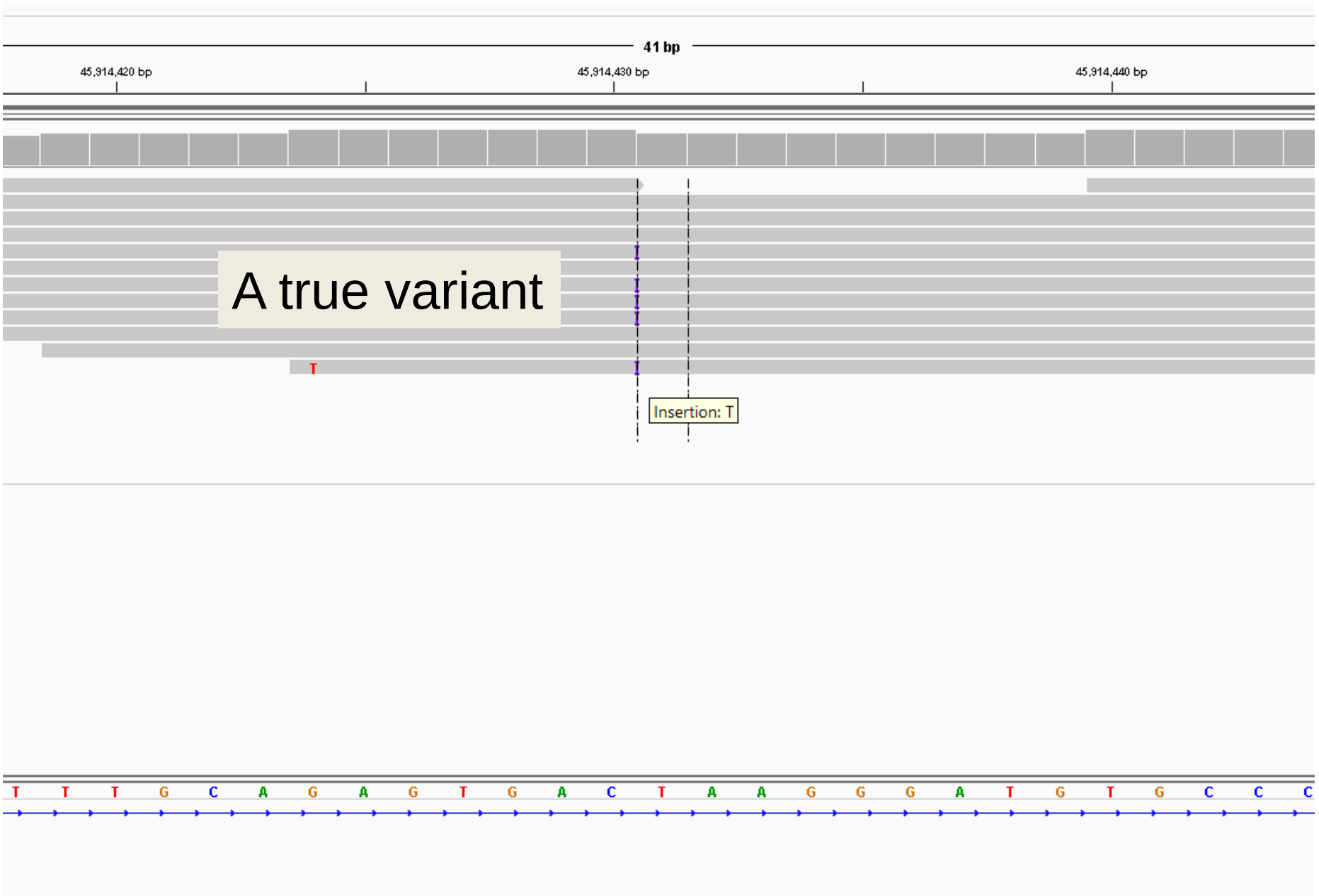
Use set criteria to filter the data

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



Example2:
Homozygote deletion of A

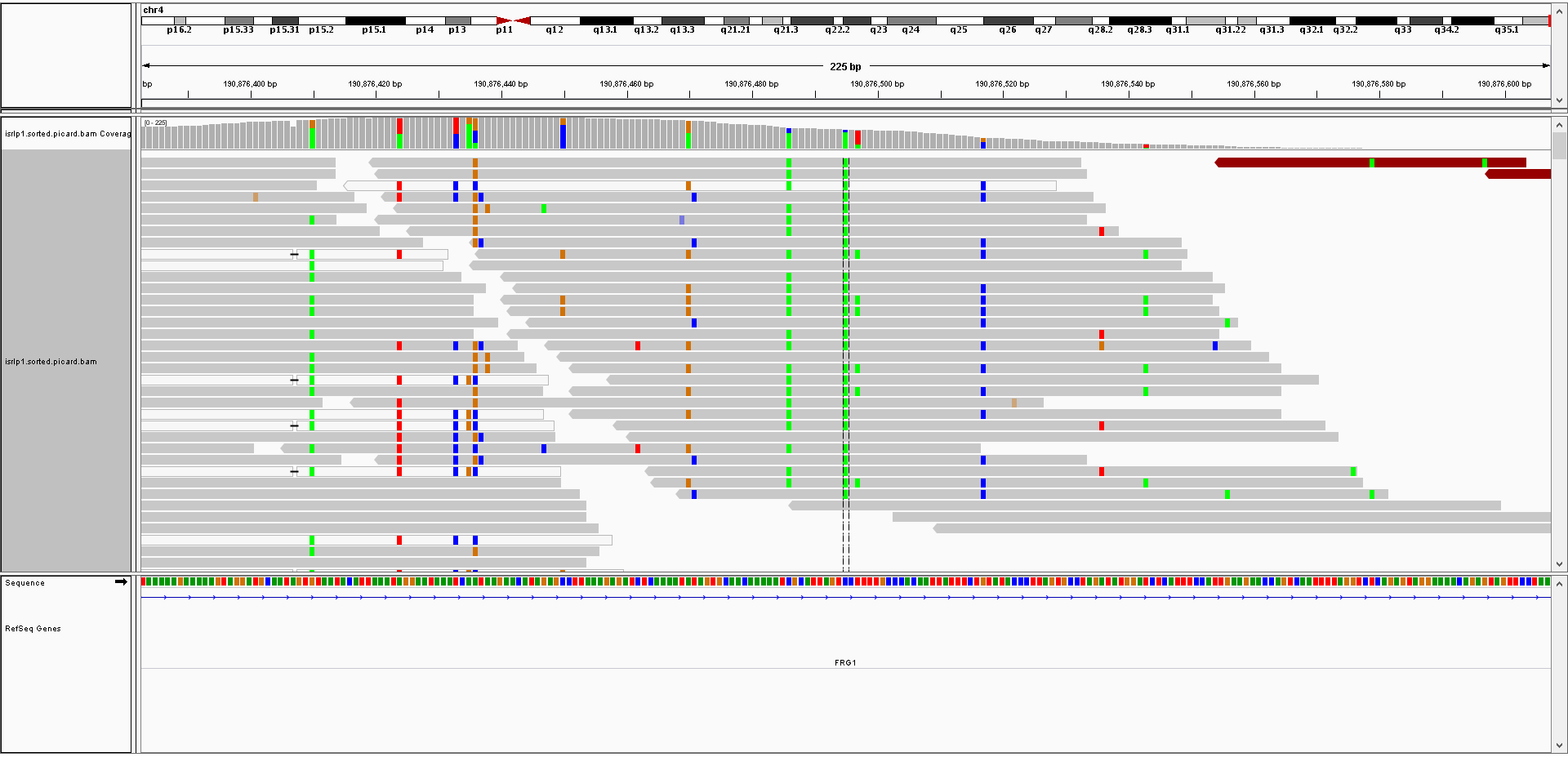
Example3: Hetrozygote insertion



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

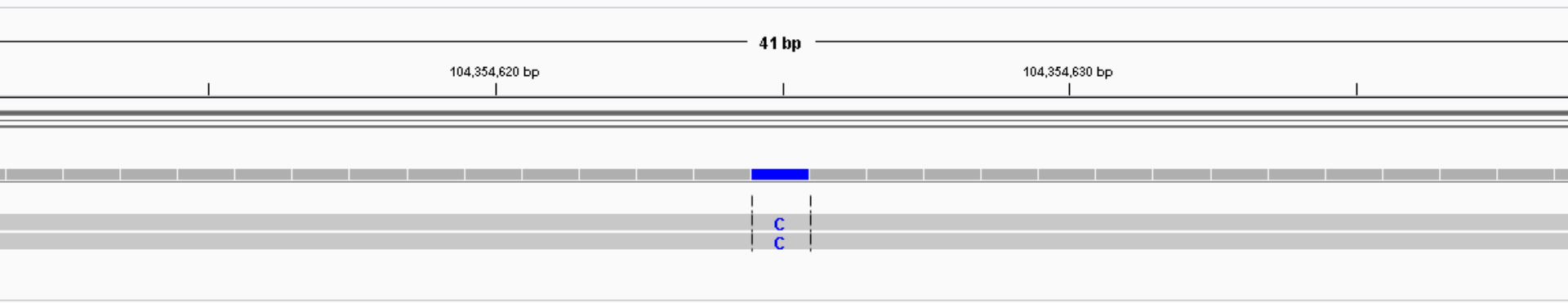
False!!
All reads are the same direction ➔ strand bias
The quality of the variation is only 3

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

Rules for variant filtration

1. The variant is called from both directions (strand bias)

2. The number of reads that support the variation versus the reads that support the reference call

3. The location of the variant in the read- variations in the 3’ end of the read are less reliable

4. The quality of the global read alignment. A true variation is more reliable when the genomic region has a good alignment

Parameters to be used for variant filtration:

Qual=quality of SNP

DP= depth of coverage

GQ= genotype quality

FS = strand bias

Also depends in the specific algorithm that was used to read variants

Tools that can be used to apply filtration:

- Vcftools
- Bcftools
- GATK
- vcflib

Variant annotation

Assigning biological meaning to each variant

- consequence on the protein sequence
(e.g. stop gained, missense, stop lost, frameshift)

- pathogenicity

- Abundance

SNPeff, Variant annotation intergrator (UCSC),
Variant Effect Predictor (ensembl), annovar

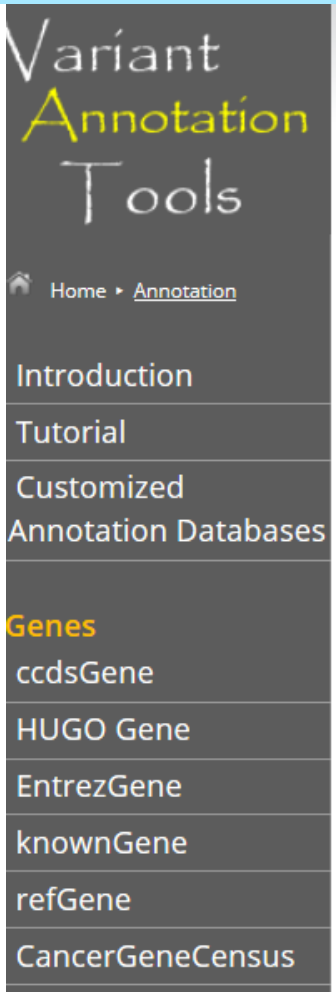
How to assess pathogenicity of genetic variants

- Evolutionary conservation.

- Prediction programs
(sift, polyphen, and more)

Tool	Numerical value	Qualitative prediction
PolyPhen	0	"probably damaging"
PolyPhen	1	"possibly damaging"
PolyPhen	2	"benign"
PolyPhen	3	"unknown"
SIFT	0	"tolerated"
SIFT	1	"deleterious"

dbSNFP- aggregates prediction scores of 11 algorithms



dbNSFP

On this page... (hide)

1. dbNSFP
2. dbNSFP_gene
3. dbNSFP_light

dbNSFP is an annotation database for non-synonymous SNPs assembled by Xiaoming Liu from the Univ
you can use the **dbNSFP** database or **dbNSFP-light** (a version with fewer features) - see which features are

1. dbNSFP

There can be frequent changes of name and their meanings of the fields across versions
them.

For the latest version dbNSFP 2.4

Population frequency

Human:
Exac
1000 Genomes Project
NHLBI

variant prioritization

Exomiser
VarElect
…..