# Variant calling and more...

José Carbonell-Caballero

GDA
International Course on
Genomic Data Analysis

PRINCIPE FELIPE
CENTRO DE INVESTIGACION

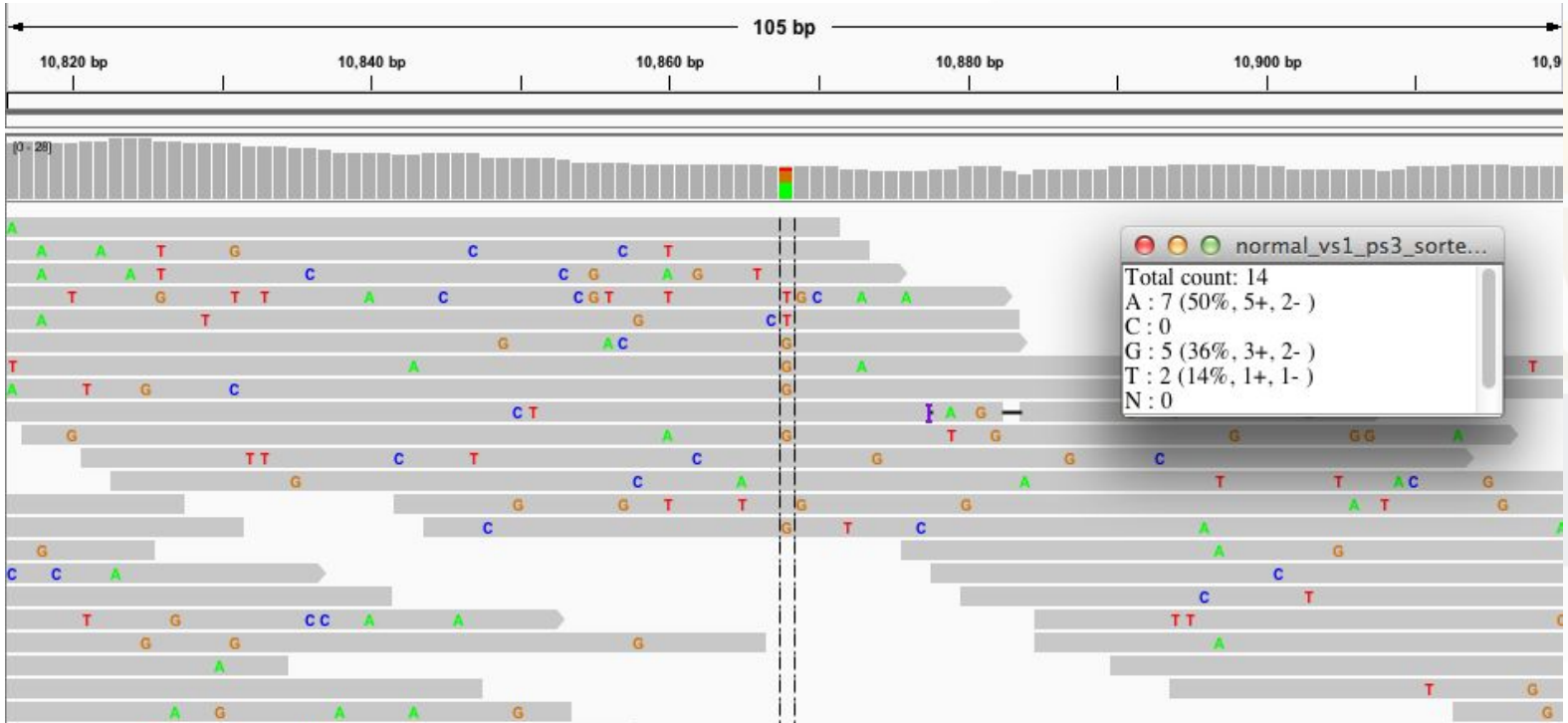
# Index

1. Introduction
2. Tailoring your mapping file
3. Predicting variants
4. Understanding VC output
5. Variant annotation
6. Mosaicism and cellular heterogeneity
7. Exercises

# 1 Introduction

## What is variant calling?

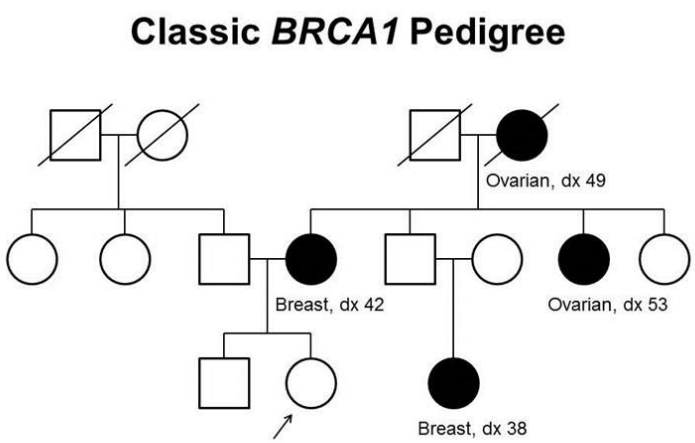
Find (robust) differences between sample genome and reference genome.

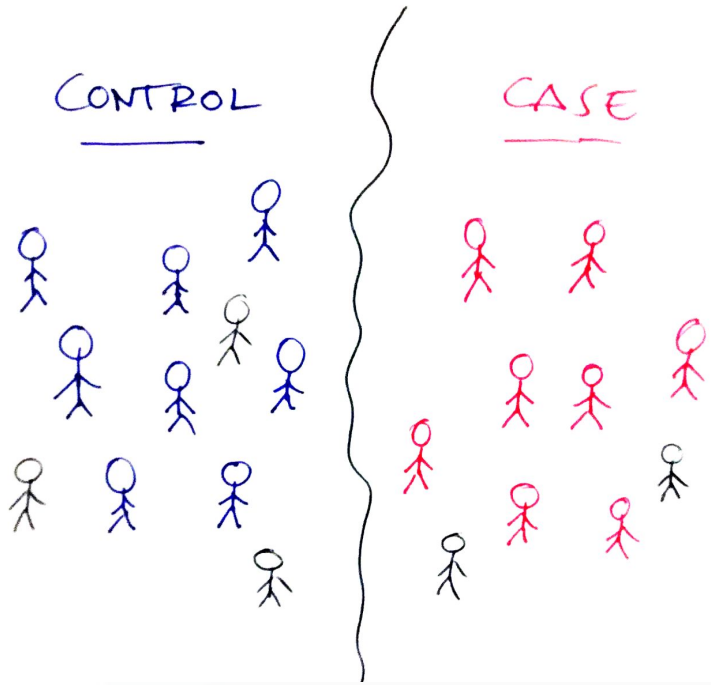


# 1 Introduction

## Why we study variants?

### Variants can explain group or phenotypic differences

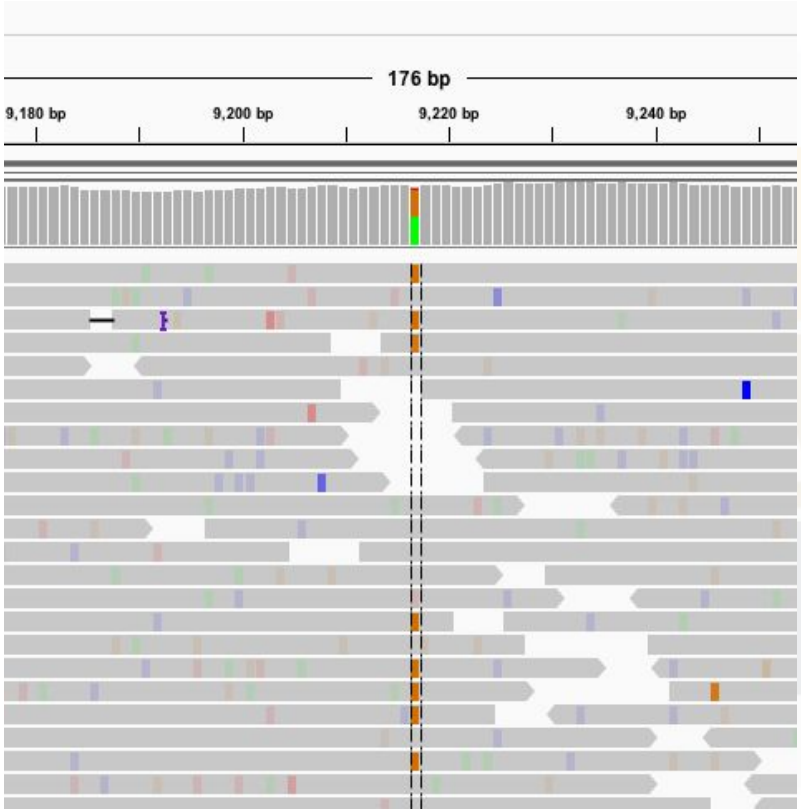


# 1 Introduction

## Dealing with uncertainty

Inherent noise in NGS makes us to face the problem from a **statistical point of view**

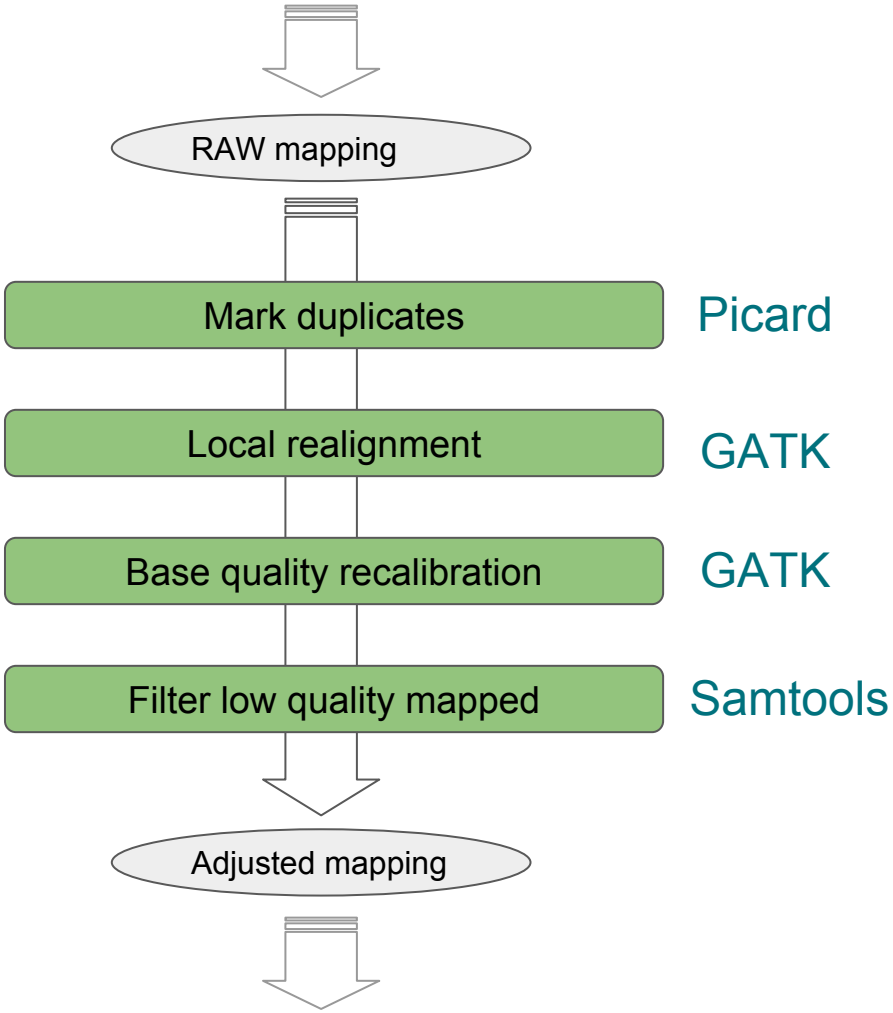



## Where it comes from?

- Random noise
- Sequencer bias
- Primary processing
- Batch effects

# 2 Tailoring your mapping file

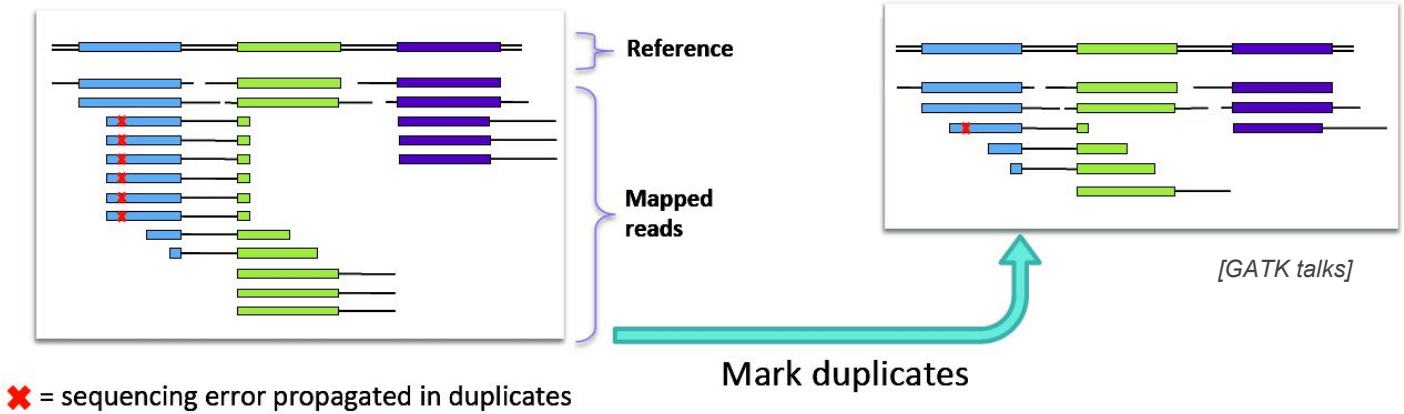
## General view



# 2 Tailoring your mapping file

## Marking PCR duplicates

- Reads from the same DNA template



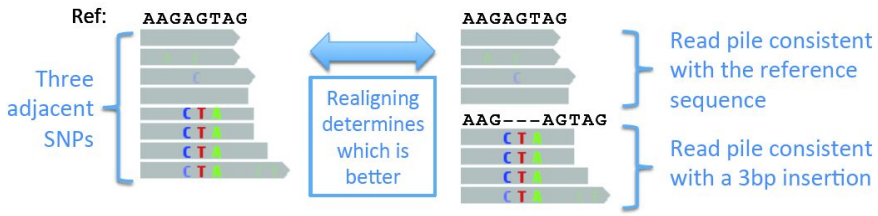

## Implications

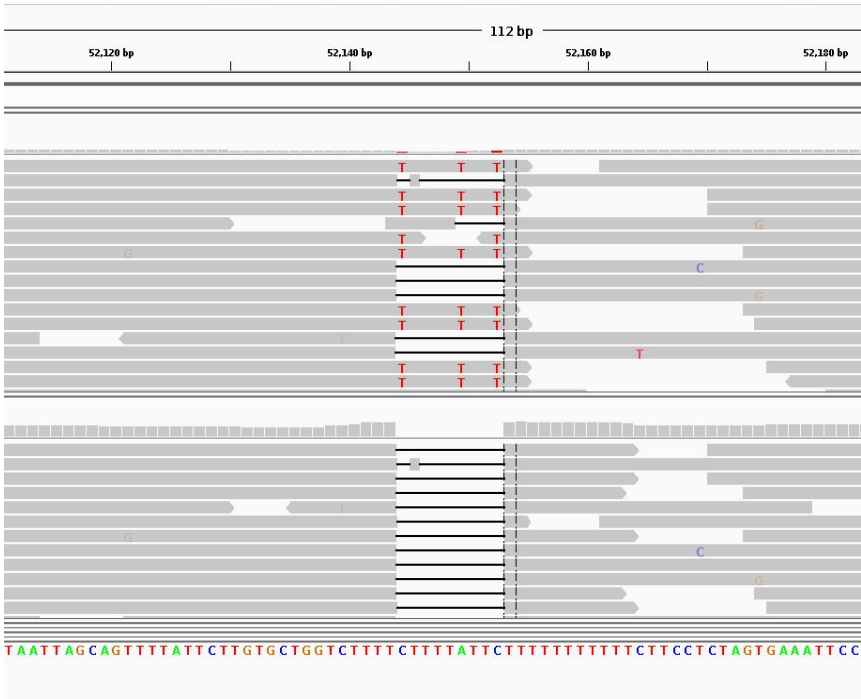
- Allele imbalance
- Common artifacts are supported by several reads

# 2 Tailoring your mapping file

## Local realignment

- Indels can confuse mappers
- Indels + Homopolymers produce ambiguity (several valid solutions)
- FPs are removed when a plausible indel is inserted



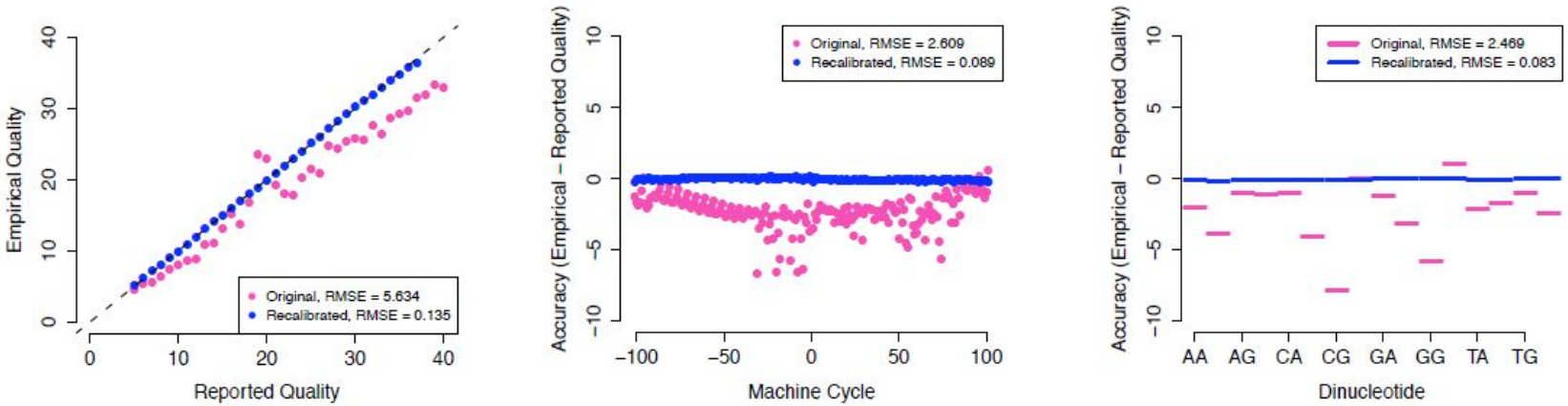

[GATK talks]

# 2 Tailoring your mapping file

## Base quality recalibration score

- Sequencers do not provide confident qualities (they are biased and innacurate)
  - Reported quality score
  - Position within the read (machine cycle)
  - Adjacent nucleotides

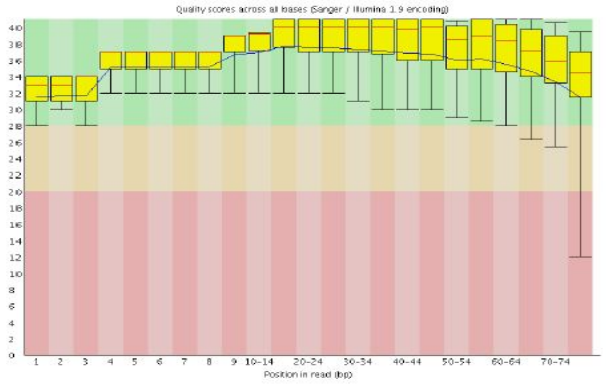
*[GATK talks]*



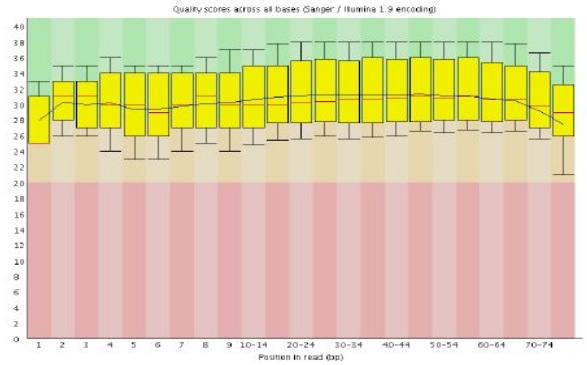
# Tailoring your mapping file

## Base quality recalibration score (BQRS)

Before

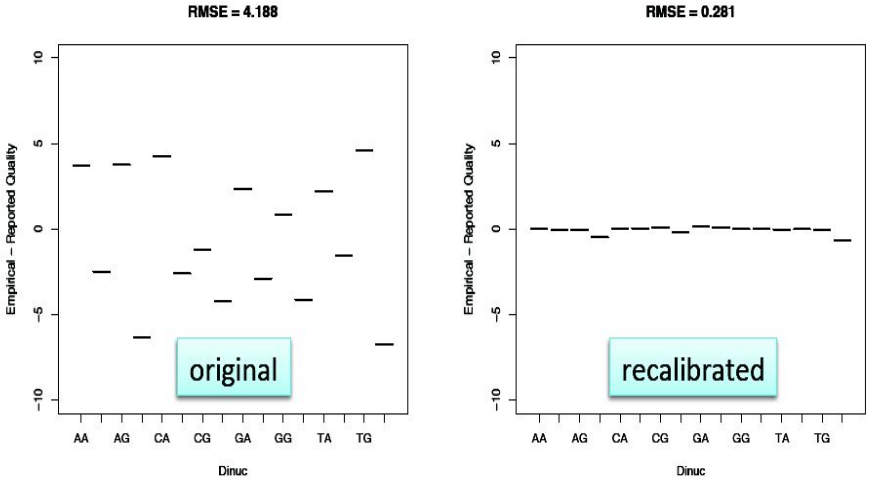



After




*[GATK talks]*

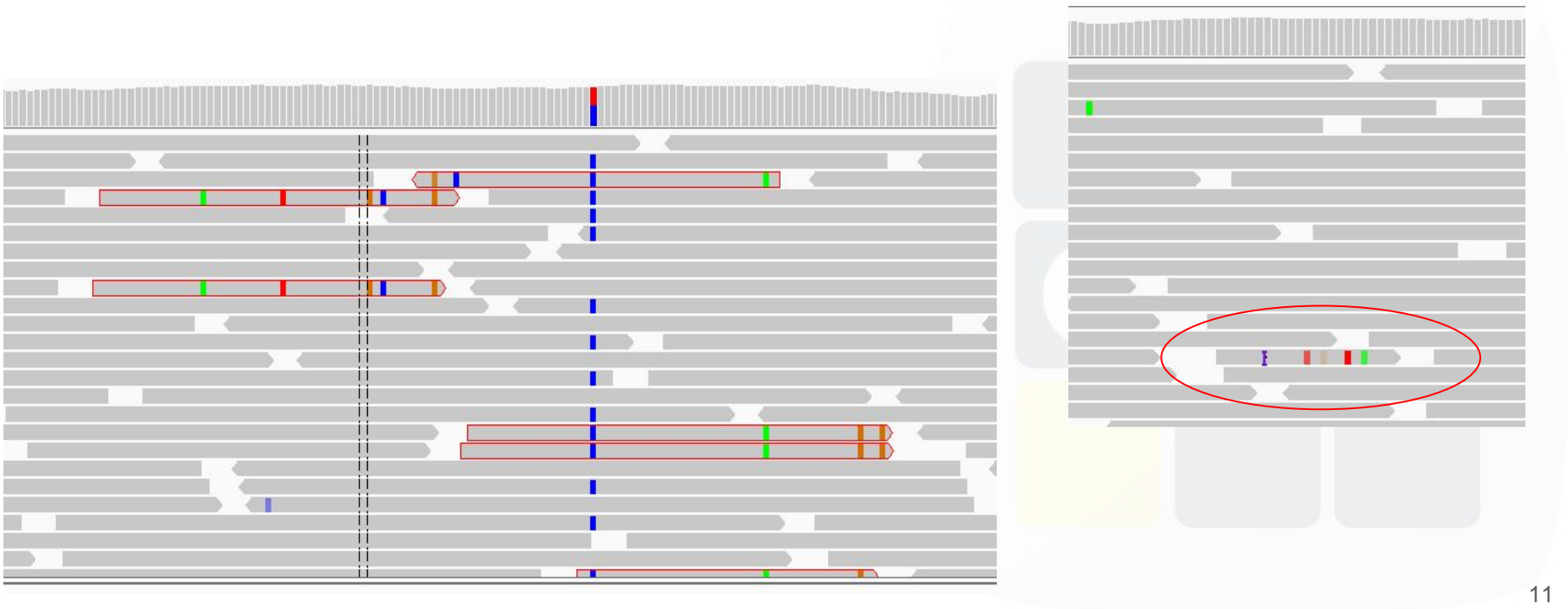
Example of bias: qualities reported depending on nucleotide context



# 2 Tailoring your mapping file

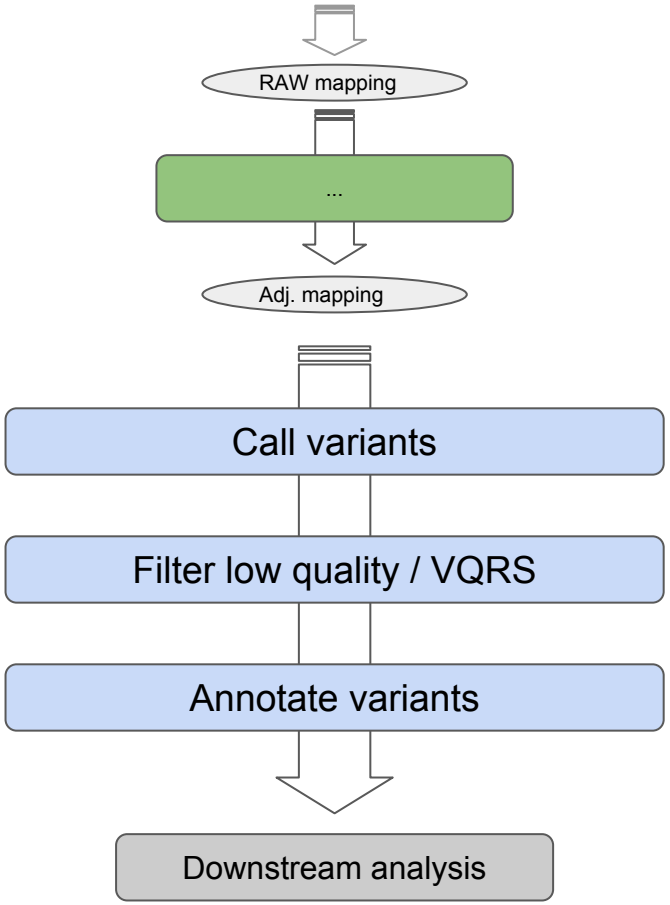
Filter low quality mapped reads

- MQ threshold ~ 0-10 (conditioned to mapper!)
- # of trimmed bases



# 3 Predicting variants

## General view



# 3 Predicting variants

## Several tools

- MAQ
- GATK
- Samtools
- Freebayes
- …

Resource

**Mapping short DNA sequencing reads and calling variants using mapping quality scores**

Heng Li,[1] Jue Ruan,[2] and Richard Durbin[1,3]

[1]The Wellcome Trust Sanger Institute, Hinxton CB10 1SA, United Kingdom; [2]Beijing Genomics Institute, Chinese Academy of Science, Beijing 100029, China

BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 21 2011, pages 2987–2993 doi:10.1093/bioinformatics/btr509

Sequence analysis Advance Access publication September 8, 2011

**A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data**

Heng Li

Medical Population Genetics Program, Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA

Associate Editor: Jeffrey Barrett

Resource

**The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data**

Aaron McKenna,[1] Matthew Hanna,[1] Eric Banks,[1] Andrey Sivachenko,[1] Kristian Cibulskis,[1] Andrew Kernytsky,[1] Kiran Garimella,[1] David Altshuler,[1,2] Stacey Gabriel,[1] Mark Daly,[1,2] and Mark A. DePristo[1,3]

[1]Program in Medical and Population Genetics, The Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA; [2]Center for Human Genetic Research, Massachusetts General Hospital, Richard B. Simches Research Center, Boston, Massachusetts 02114, USA

**Haplotype-based variant detection from short-read sequencing**

Erik Garrison and Gabor Marth

July 24, 2012

**Abstract**

The direct detection of haplotypes from short-read DNA sequencing data requires changes to existing small-variant detection methods. Here, we develop a Bayesian statistical framework which is capable of modeling multiallelic loci in sets of individuals with non-uniform copy number. We then describe our implementation of this framework in a haplotype-based variant detector, FreeBayes.

# 3 Predicting variants

gatk

Home | About | Guide | Blog | Forum | Download

Search

The Genome Analysis Toolkit or GATK is a software package for analysis of high-throughput sequencing data, developed by the Data Science and Data Engineering group at the Broad Institute. The toolkit offers a wide variety of tools, with a primary focus on variant discovery and genotyping as well as strong emphasis on data quality assurance. Its robust architecture, powerful processing engine and high-performance computing features make it capable of taking on projects of any size.

Learn more »

**About**
Overview of the tools and development team

**Guide**
Detailed documentation, guidelines and tutorials

**Forum**
Ask here for help with questions and bug reports

**Blog**
Announcements and progress updates

**Latest version: 3.5**
Release Notes

Download now

Licensing information

BROAD INSTITUTE

© Broad Institute 2015

# 3 Predicting variants

## GATK

- SNV and Indel calling are of special interest

Bayesian model

$$\Pr\{G|D\} = \frac{\overbrace{\Pr\{G\}}^{\text{Prior of the genotype}}\ \overbrace{\Pr\{D|G\}}^{\text{Likelihood of the genotype}}}{\Sigma_i \Pr\{G_i\}\Pr\{D|G_i\}}, \text{ [Bayes' rule]}$$

$$\Pr\{D|G\} = \prod_j \left(\frac{\Pr\{D_j|H_1\}}{2} + \frac{\Pr\{D_j|H_2\}}{2}\right) \text{ where } \overbrace{G = H_1 H_2}^{\text{Diploid assumption}}$$

$$\Pr\{D|H\} \text{ is the haploid likelihood function}$$

# 3 Predicting variants

## GATK

- But, it is a framework (with tons of tools) to manage VCF and BAM files

**1. PRE-PROCESSING**

Pre-processing starts from raw sequence data, either in FASTQ or uBAM format, and produces analysis-ready BAM files. Processing steps include alignment to a reference genome as well as some data cleanup operations to correct for technical biases and make the data suitable for analysis.

**2. VARIANT DISCOVERY**

Variant Discovery starts from analysis-ready BAM files and produces a callset in VCF format. Processing involves identifying sites where one or more individuals display possible genomic variation, and applying filtering methods appropriate to the experimental design.

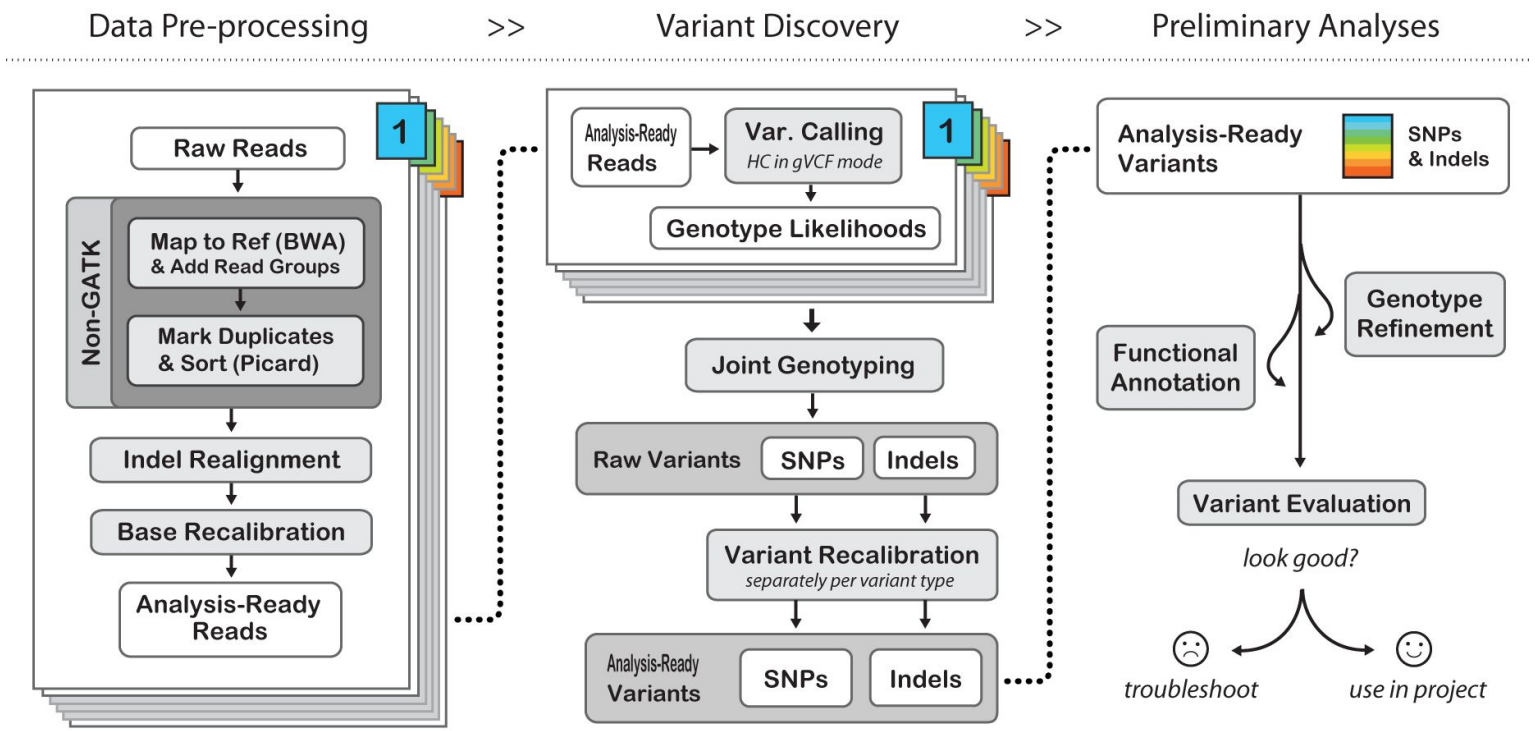
**3. CALLSET REFINEMENT**

Callset Refinement starts and ends with a VCF callset. Processing involves using meta-data to assess and improve genotyping accuracy, attach additional information and evaluate the overall quality of the callset.

# 3 Predicting variants

## GATK

- They promoted “Best practices in NGS processing”



# 4 Understanding VC output

## VCF format

Misc

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
```

Info

```
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
```

Filter

```
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
```

Format

```
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
```

| #CHROM | POS | ID | REF | ALT | QUAL | FILTER | INFO | FORMAT | NA00001 | NA00002 |
|---|---|---|---|---|---|---|---|---|---|---|
| 20 | 14370 | rs6054257 | G | A | 29 | PASS | NS=3;DP=14;AF=0.5;DB;H2 | GT:GQ:DP:HQ | 0\|0:48:1:51,51 | 1\|0:48:8:51,51 |
| 20 | 17330 | . | T | A | 3 | q10 | NS=3;DP=11;AF=0.017 | GT:GQ:DP:HQ | 0\|0:49:3:58,50 | 0\|1:3:5:65,3 |
| 20 | 1110696 | rs6040355 | A | G,T | 67 | PASS | NS=2;DP=10;AF=0.333,0.667;AA=T;DB | GT:GQ:DP:HQ | 1\|2:21:6:23,27 | 2\|1:2:0:18,2 |
| 20 | 1230237 | . | T | . | 47 | PASS | NS=3;DP=13;AA=T | GT:GQ:DP:HQ | 0\|0:54:7:56,60 | 0\|0:48:4:51,51 |
| 20 | 1234567 | microsat1 | GTC | G,GTCT | 50 | PASS | NS=3;DP=9;AA=G | GT:GQ:DP | 0/1:35:4 | 0/2:17:2 |

variant summary

summary of quality

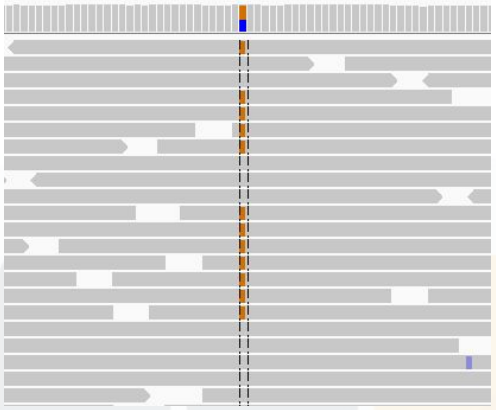
Sample summary and tool dependent inferred parameters

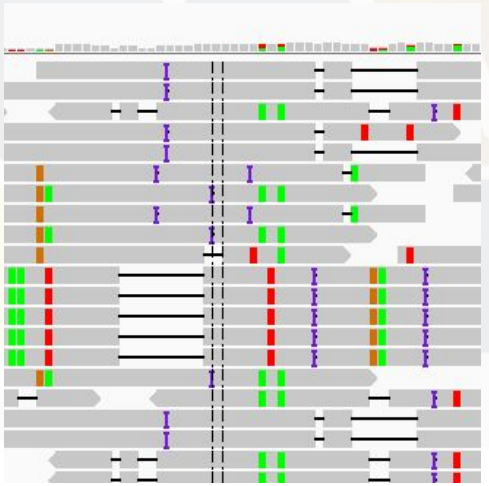
Sample detail

# 4 Understanding VC output

## How to (hard) filter variants for SNV/Indels

- Depth (DP) ?/?
- QualByDepth (QD) < 2/2
- RMSMappingQuality (MQ) < 40.0/-
- FisherStrand (FS) > 60.0/200.0
- HaplotypeScore > 13.0/-
- MQRankSumTest < -12.5/-
- ReadPosRankSumTest < -8.0/-20.0



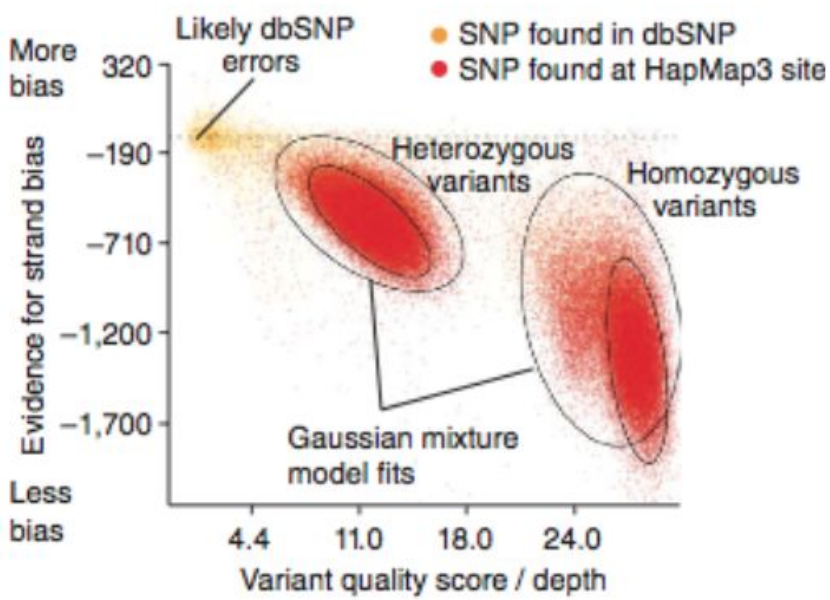


# 4 Understanding VC output

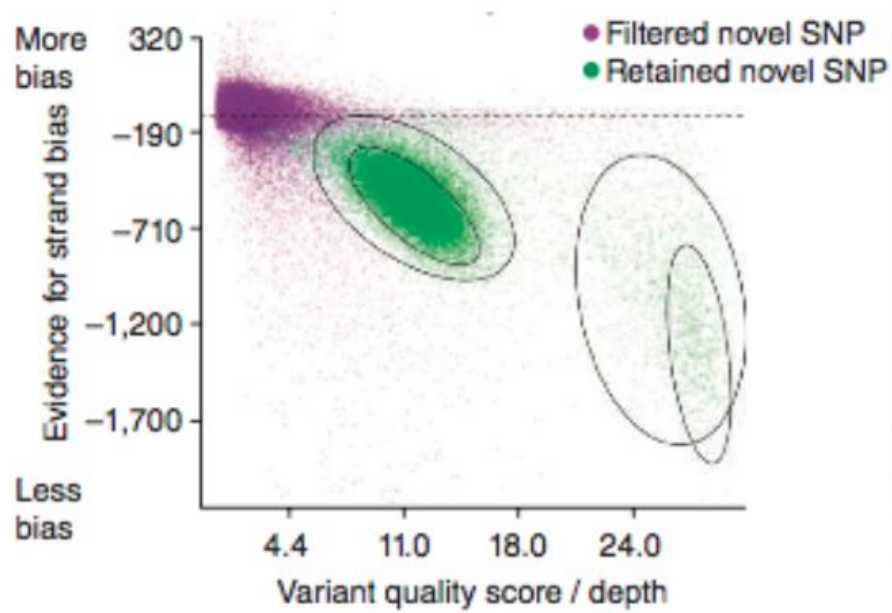
## Variant Quality Recalibration Score (VQRS)

- We can use together quality scores in a model
- We need a set of true (known) variants (e.g. dbSNP)

(1) Train model using HapMap




(2) Apply model to callset



# 5 Variant annotation

## How we enrich our prediction?

- Mutation effect
  - Annovar
  - SnpEff

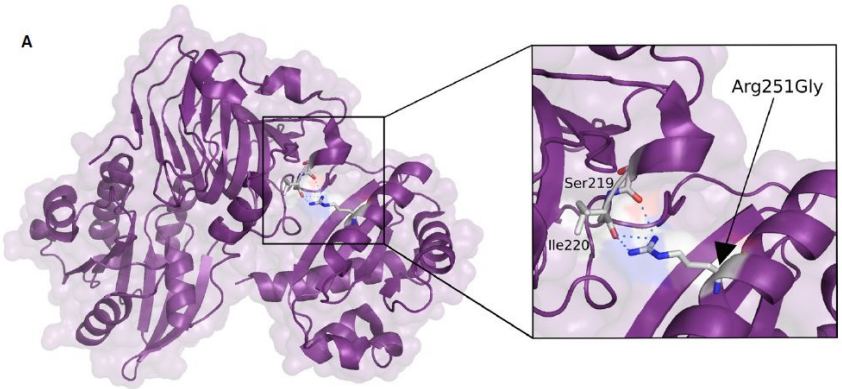
**Type**

| Type (alphabetical order) | Count | Percent |
|---|---|---|
| DOWNSTREAM | 2,093 | 1.766% |
| INTERGENIC | 26,314 | 22.204% |
| INTRAGENIC | 78 | 0.066% |
| INTRON | 54,238 | 45.767% |
| NON_SYNONYMOUS_CODING | 237 | 0.2% |
| NON_SYNONYMOUS_START | 1 | 0.001% |
| SPLICE_SITE_DONOR | 4 | 0.003% |
| START_GAINED | 57 | 0.048% |
| STOP_GAINED | 3 | 0.003% |
| STOP_LOST | 1 | 0.001% |
| SYNONYMOUS_CODING | 378 | 0.319% |
| TRANSCRIPT | 32,163 | 27.14% |
| UPSTREAM | 2,102 | 1.774% |
| UTR_3_PRIME | 690 | 0.582% |
| UTR_5_PRIME | 149 | 0.126% |

**Region**

| Type (alphabetical order) | Count | Percent |
|---|---|---|
| DOWNSTREAM | 2,093 | 1.766% |
| EXON | 620 | 0.523% |
| INTERGENIC | 26,314 | 22.204% |
| INTRON | 54,238 | 45.767% |
| NONE | 32,241 | 27.206% |
| SPLICE_SITE_DONOR | 4 | 0.003% |
| UPSTREAM | 2,102 | 1.774% |
| UTR_3_PRIME | 690 | 0.582% |
| UTR_5_PRIME | 206 | 0.174% |

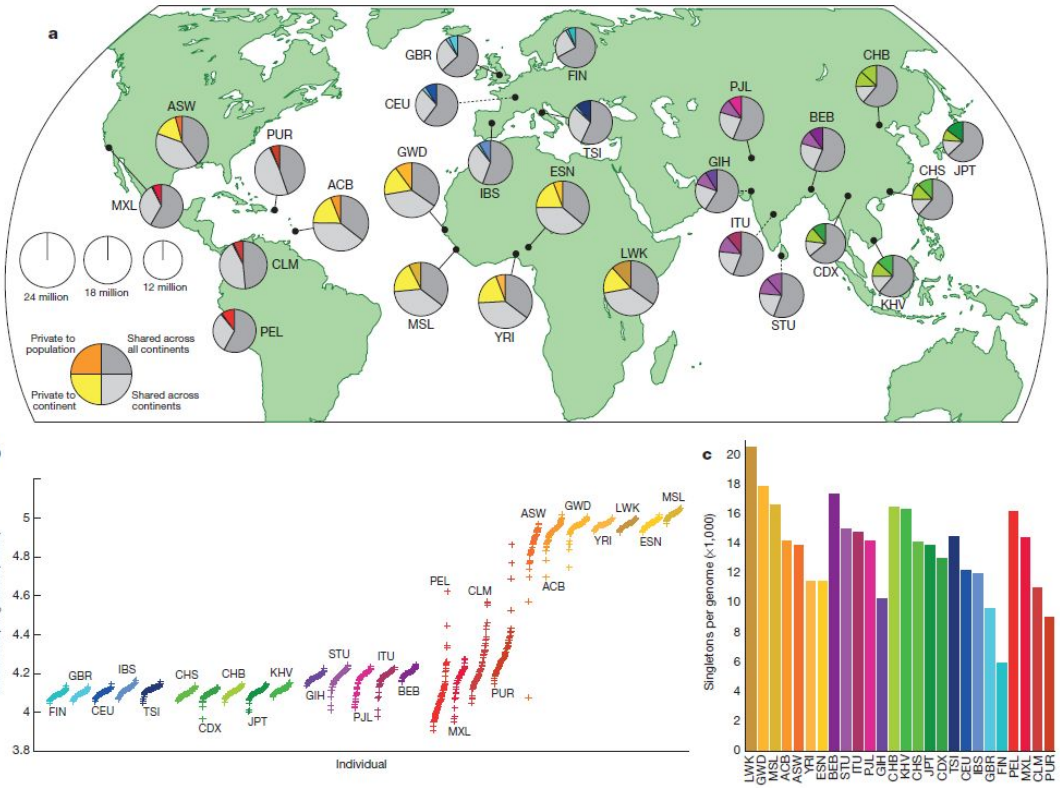
- Loss/Gain of function in protein?
  - SIFT
  - Polyphen
  - CAAD



# 5 Variant annotation

How we enrich our prediction?

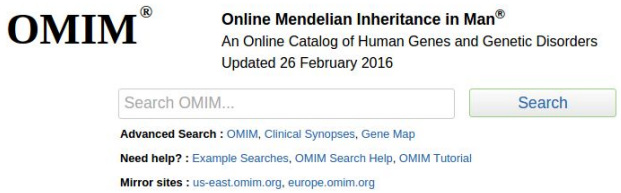
- Allele frequency in (almost) healthy populations
  - 1000K
  - dbSNP
  - ExAC

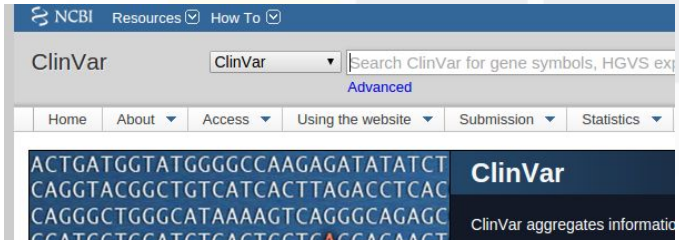


# 5 Variant annotation

## How we enrich our prediction?

- Variant allele in pathological mutation DBs
  - COSMIC (Catalogue of somatic mutations in cancer)
    - http://cancer.sanger.ac.uk/cosmic
  - HGMD (Human Gene Mutation Database)
    - http://www.hgmd.cf.ac.uk/
  - OMIM (Online Mendelian inheritance in men)
    - http://www.omim.org/
  - ClinVar (NCBI clinical variants repository)
    - http://www.ncbi.nlm.nih.gov/clinvar/

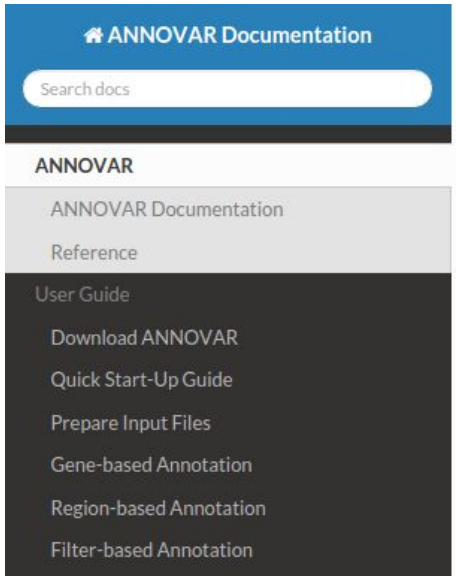


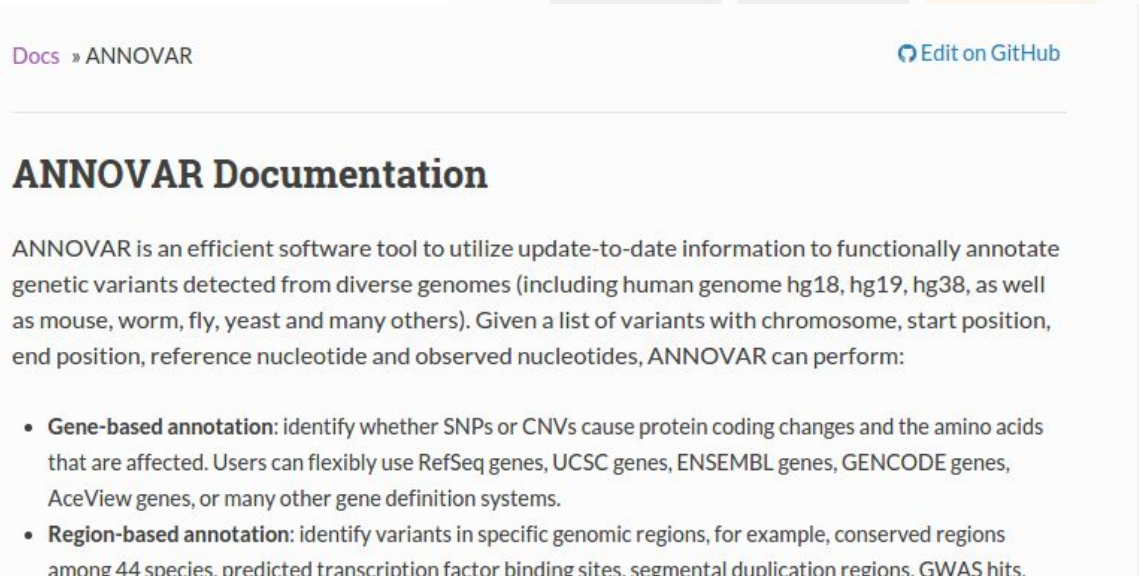



# 5 Variant annotation

## Command line tools

- Annovar
  - Region-based annotator
  - Highly flexible (precomputed scores for almost anything you want to annotate to a variant)

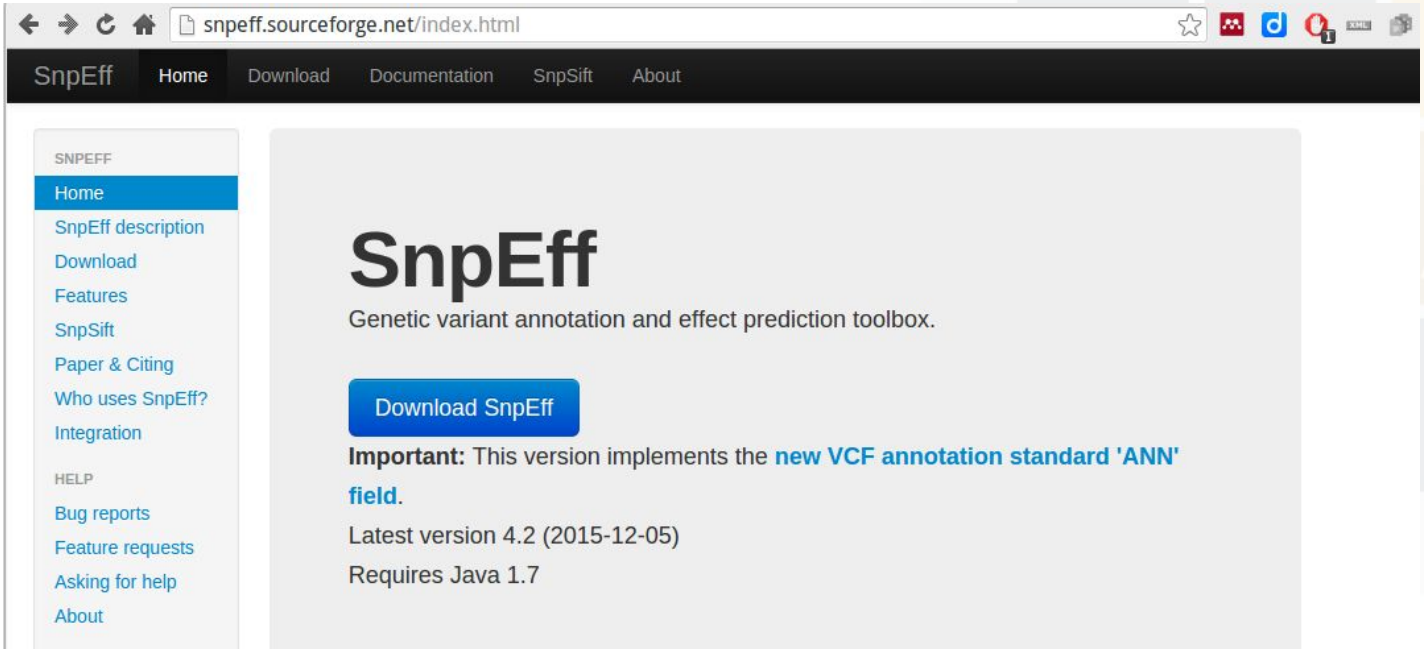


# 5 Variant annotation

## Command line tools

- SnpEff
  - Less flexible than Annovar
  - But faster than Annovar

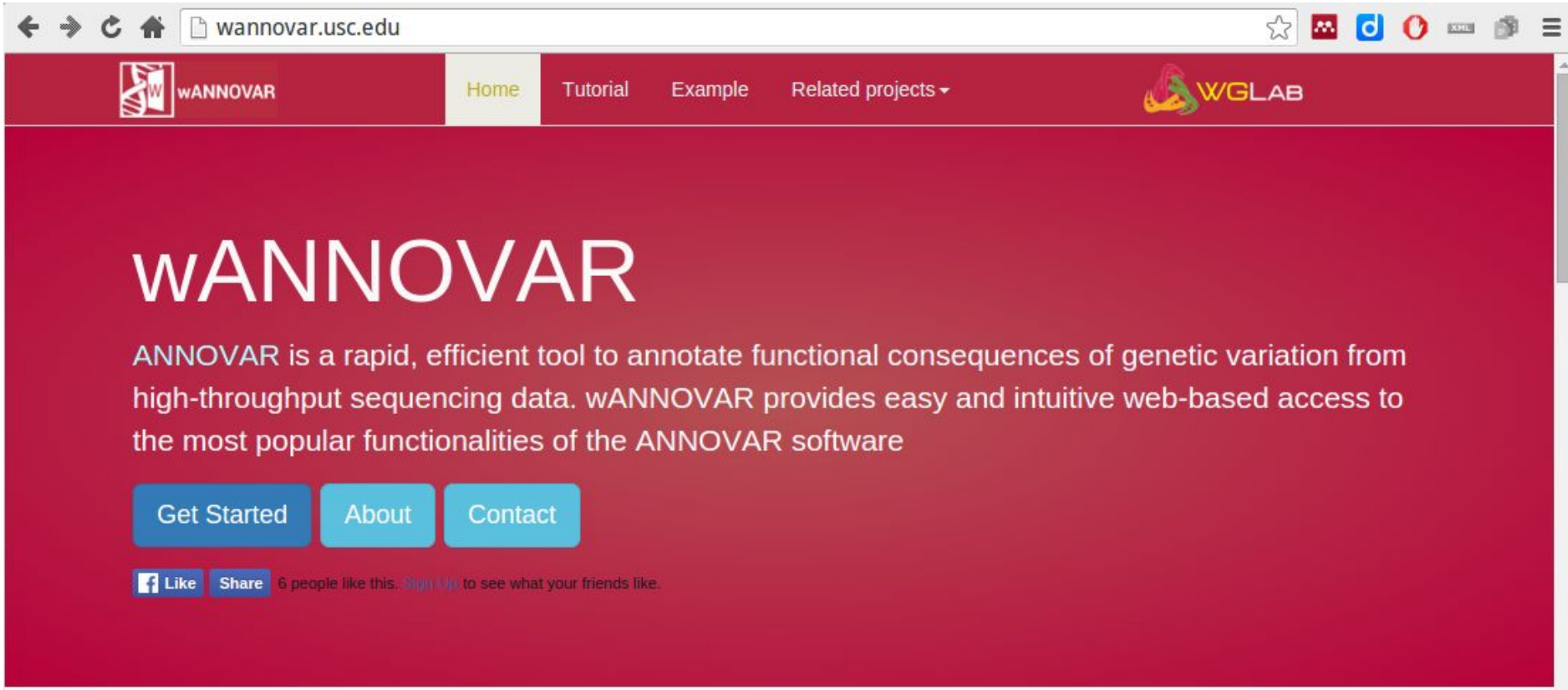


# 5 Variant annotation

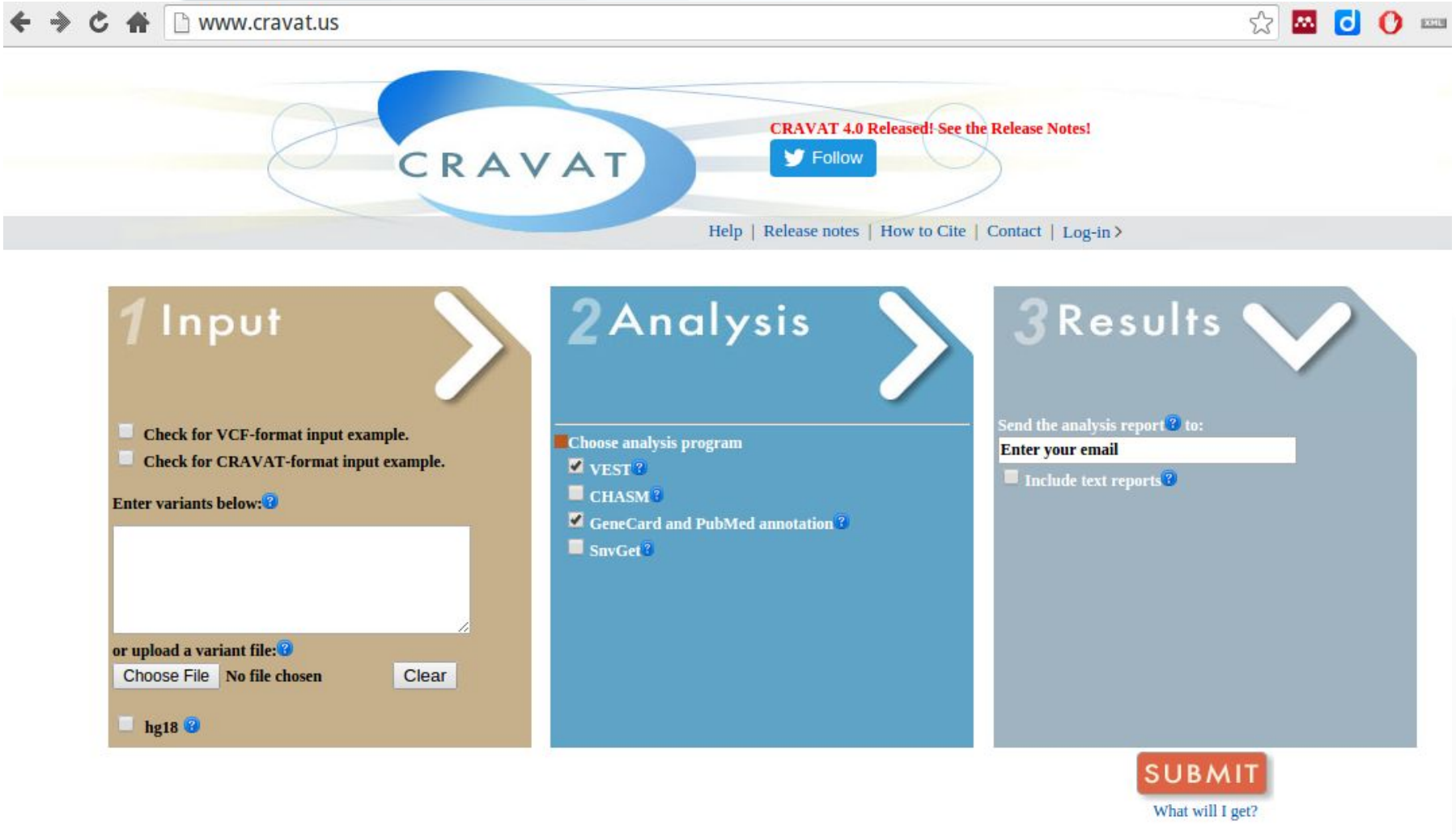
## Webtools

- Easy to use
- Useful for beginners in informatics
- But not so useful for (relatively) large VCFs
- Small number of candidates
- They cannot be include in a bioinformatics pipeline

# 5 Variant annotation



# 5 Variant annotation



# 5 Variant annotation

https://www.broadinstitute.org/oncotator/

Oncotator

Home Download News Help About

## What is Oncotator?

Oncotator is a web application for annotating human genomic point mutations and indels with data relevant to cancer researchers. Annotations are aggregated from the following resources:

### Genomic Annotations

- Gene, transcript, and functional consequence annotations using GENCODE for hg19.
- Reference sequence around a variant.
- GC content around a variant.
- Human DNA Repair Gene annotations from Wood et al.

### Protein Annotations

- Site-specific protein annotations from UniProt.
- Functional impact predictions from dbNSFP.

### Cancer Variant Annotations

- Observed cancer mutation frequency annotations from COSMIC.
- Cancer gene and mutation annotations from the Cancer GenCensus.
- Overlapping mutations from the Cancer Cell Line Encyclopedia.
- Cancer gene annotations from the Familial Cancer Database.
- Cancer variant annotations from ClinVar.

### Non-Cancer Variant Annotations

- Common SNP annotations from dbSNP.
- Variant annotations from 1000 Genomes.
- Variant annotations from NHLBI GO Exome Sequencing Project (ESP).

## Get Started

Upload your input data below. Please note, only GRCh37/hg19 coordinates are supported and there is a maximum of 5,000 alterations per input file. If you need additional help, please visit our help page or feel free to contact us directly.

### Upload Mutations

Choose File No file chosen

Submit

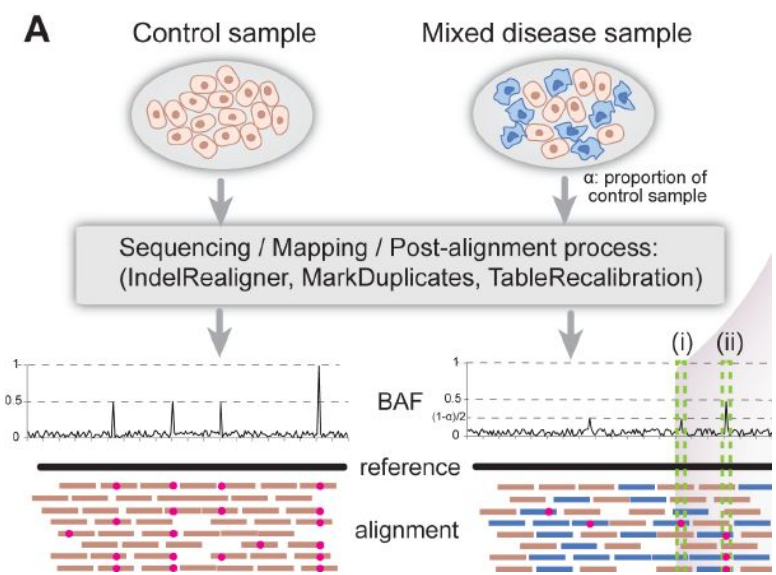
Upload TSV file of mutations in Oncotator format.

### Paste Mutations

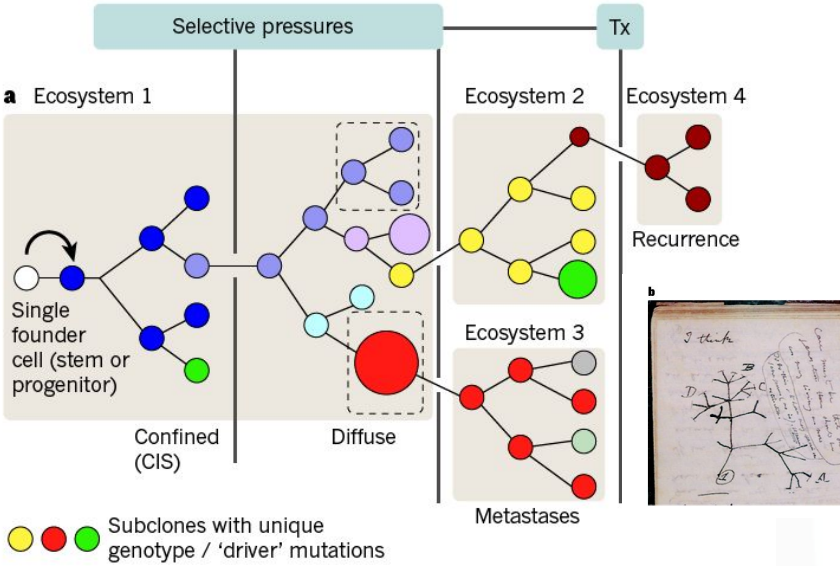
# 6 Mosaicism and cellular heterogeneity

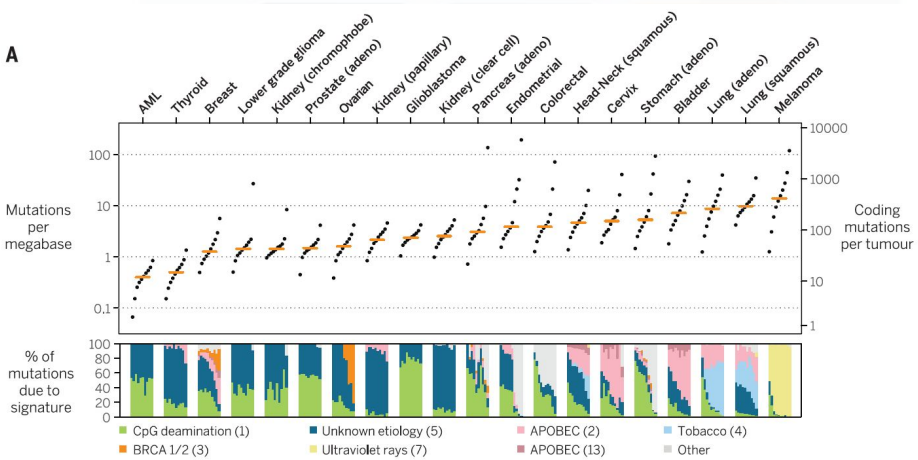
## Cancer genomics

- Huge amount of alterations
- Tumors are not homogeneous
- Germline models are not valid




Kim, S., et al (2013). Virmid: accurate detection of somatic mutations with sample impurity inference. Genome Biology, 14(8), R90. doi:10.1186/gb-2013-14-8-r90



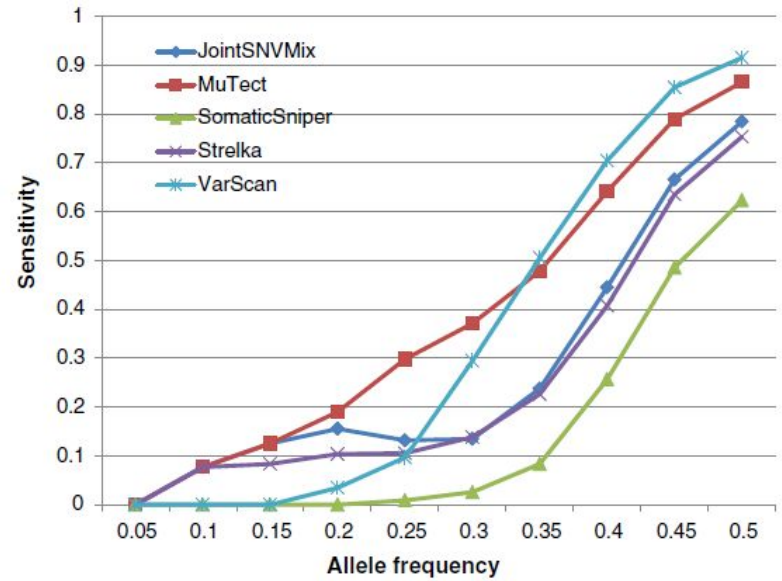




Campbell, P. J. (2015). Somatic mutation in cancer and normal cells, 349(6255)

# 6 Mosaicism and cellular heterogeneity

## Cancer genomics

- Somatic variant calling
  - SNVmix
  - Varscan
  - **Mutect**
  - Virmid
  - Strelka




Xu et al. BMC Genomics 2014, 15:244
http://www.biomedcentral.com/1471-2164/15/244

BMC Genomics

RESEARCH ARTICLE Open Access

Comparison of somatic mutation calling methods in amplicon and whole exome sequence data

Huilei Xu, John DiCarlo, Ravi Vijaya Satya, Quan Peng and Yexun Wang*

Kim and Speed BMC Bioinformatics 2013, 14:189
http://www.biomedcentral.com/1471-2105/14/189

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Comparing somatic mutation-callers: beyond Venn diagrams

Su Yeon Kim[1*] and Terence P Speed[1,2*]

Wang et al. Genome Medicine 2013, 5:91
http://genomemedicine.com/content/5/10/91

Genome Medicine

RESEARCH Open Access

Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers

Qingguo Wang[1], Peilin Jia[1,2], Fei Li[3], Haiquan Chen[4,5], Hongbin Ji[3], Donald Hucks[6], Kimberly Brown Dahlman[6,7], William Pao[6,8*] and Zhongming Zhao[1,2,7,9*]

# References

## Some references

- GATK talks https://www.broadinstitute.org/gatk/guide/presentations
- GATK best practices https://www.broadinstitute.org/gatk/guide/best-practices
- SIFT http://sift.jcvi.org/
- Polyphen2 http://genetics.bwh.harvard.edu/pph2/
- Annovar http://annovar.openbioinformatics.org/en/latest/

## Some recommended papers


- DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., … Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–8. doi:10.1038/ng.806
- Wang, Q., Jia, P., Li, F., Chen, H., Ji, H., Hucks, D., … Zhao, Z. (2013). Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Medicine*, *5*(10), 91. doi:10.1186/gm495
- Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., … Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74. doi:10.1038/nature15393
- Hui Yang, Kai Wang (2015). Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nature Protocols 10, 1556–1566. doi:10.1038/nprot.2015.105
- Kumar P, Henikoff S, Ng PC. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073-81. doi: 10.1038/nprot.2009.86
- Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. doi:10.1093/bioinformatics/btr509
- Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., … Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, *31*(3), 213–9. doi:10.1038/nbt.2514
- Exac. (2015). Analysis of protein-coding genetic variation in 60,706 humans. *Bioarxiv*, 1–26.

# Now it’s time to practice!