NGS I : VARIANT DETECTION

Javier Perales-Patón
jperales@cnio.es

Fátima Al-Shahrour
[falshahrour@cnio.es]
Elena Piñeiro-Yáñez
[epineiro@cnio.es]
Pedro Fernandes
[pfern@igc.gulbenkian.pt]

cnio
Centro Nacional de Investigaciones Oncológicas

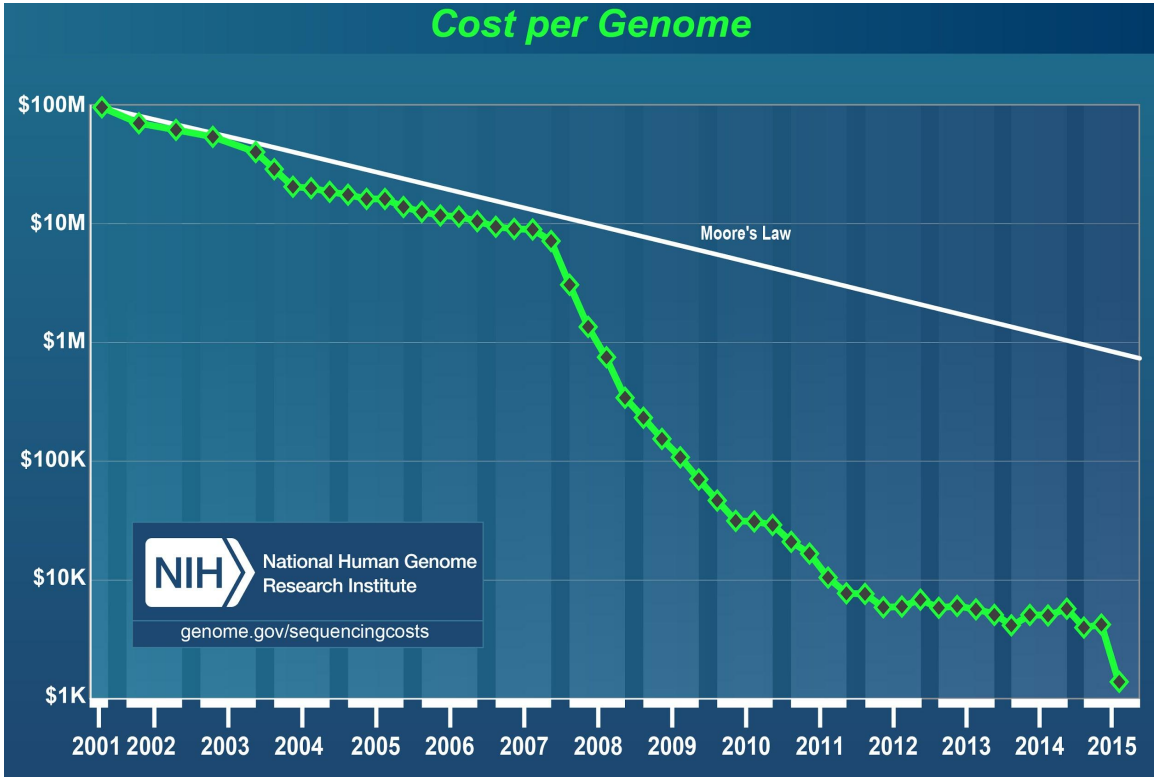
Translational Bioinformatics Unit
CNIO. Madrid, Spain.

INSTITUTO GULBENKIAN DE CIÊNCIA

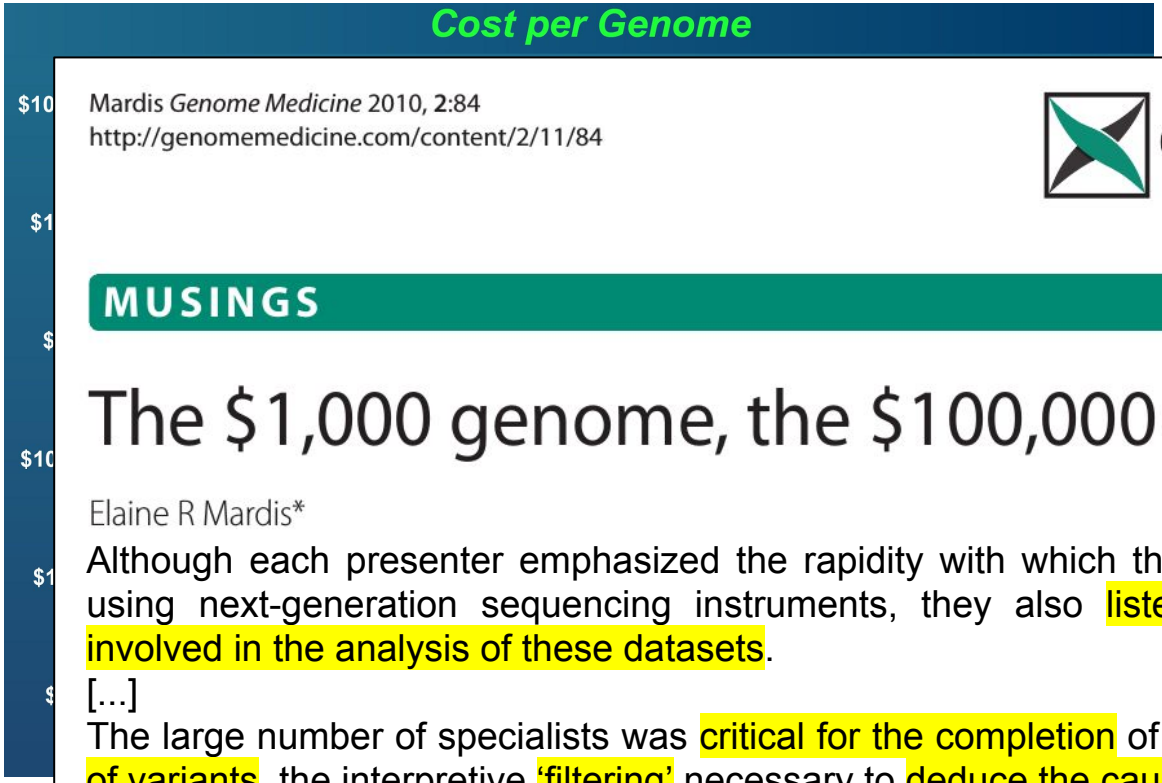
Today

09:30 - 10:00	Introduction to the course and self presentation of the participants. Personalized medicine.
11:30 - 12:30	**NGS I : Variant detection.**
14:00 - 16:00	Playing with the data and the methods.
16:30 - 18:00	Practical : Running the pipeline.

Sequencing cost has been coming down



Sequencing cost has been coming down



Mardis *Genome Medicine* 2010, **2**:84
http://genomemedicine.com/content/2/11/84

Genome Medicine

MUSINGS

The $1,000 genome, the $100,000 analysis?

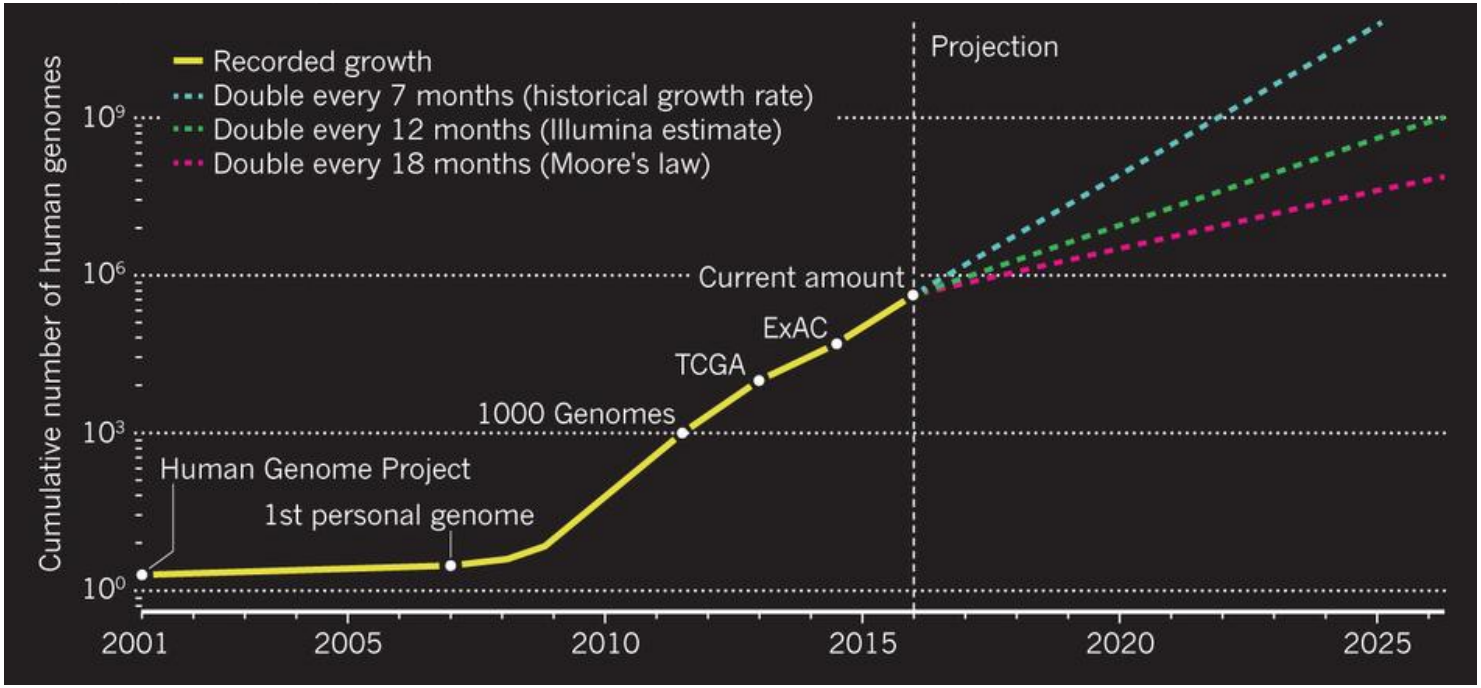
Elaine R Mardis*

Although each presenter emphasized the rapidity with which these data can now be generated using next-generation sequencing instruments, they also listed the large number of people involved in the analysis of these datasets.

[...]

The large number of specialists was critical for the completion of the data analysis, the annotation of variants, the interpretive 'filtering' necessary to deduce the causative or 'actionable' variants, the clinical verification of these variants, and the communication of results and their ramifications to the treating physician, and ultimately to the patient. At the end of the day, although the idea of clinical whole-genome sequencing for diagnosis is exciting and potentially life-changing for these patients, one does wonder how, in the clinical translation required for this practice to become common- place, such a 'dream team' of specialists would be assembled for each case.

DNA sequencing soars



+ 1000 Genomes Project : hundreds of genomes.
+ TCGA : thousands (genome & exomes).
+ ExAC : > 60,000 exomes.

Stephens ZD *et al.* **Big Data: Astronomical or Genomical?**. PLoS Biol. 2015 Jul 7;13(7)
Eisenstein M. **Big data: The power of petabytes**. Nature. 2015 Nov 5;527(7576):S2-4.

1000 Genomes Project

ARTICLE

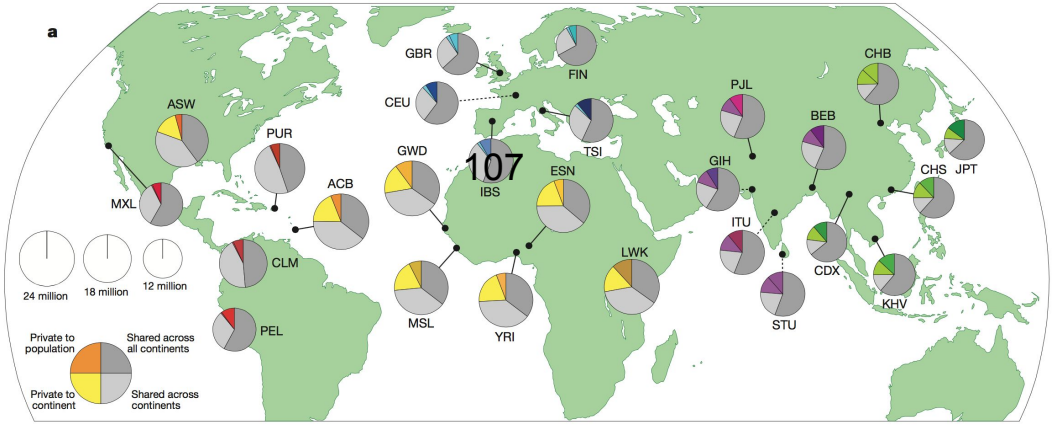
OPEN

doi:10.1038/nature15393

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.



Phase 3	WGS	WExS
Raw bases	89 Tb	18 Tb
Samples	2,504	2,504
Region	Genome	Exome
Mean Depth	8.45x	75x
SNPs	85M	1.5M
Indels	3.6M	22K
Structural Variants	60K	6.5K
Het. Concordance (SNPs)	99.4%	99.8%

http://www.1000genomes.org/about#ProjectSamples ; Phase 1 n=1092 → Phase 3 n=2504

ExAC

ARTICLE

OPEN

doi:10.1038/nature19057

Analysis of protein-coding genetic variation in 60,706 humans

Large-scale reference data sets of human genetic variation are critical for the medical and functional interpretation of DNA sequence changes. Here we describe the aggregation and analysis of high-quality exome (protein-coding region) DNA sequence data for 60,706 individuals of diverse ancestries generated as part of the Exome Aggregation Consortium (ExAC). This catalogue of human genetic diversity contains an average of one variant every eight bases of the exome, and provides direct evidence for the presence of widespread mutational recurrence. We have used this catalogue to calculate objective metrics of pathogenicity for sequence variants, and to identify genes subject to strong selection against various classes of mutation; identifying 3,230 genes with near-complete depletion of predicted protein-truncating variants, with 72% of these genes having no currently established human disease phenotype. Finally, we demonstrate that these data can be used for the efficient filtering of candidate disease-causing variants, and for the discovery of human 'knockout' variants in protein-coding genes.

An average of 85 heterozygous and 35 homozygous Protein-truncating variants per individual. The majority found in any one person are common, and only 2 are singleton.

Blog Post: http://blogs.nature.com/freeassociation/2016/08/joint-calling-of-the-exac-publications.html

Paper : Lek M et al. **Analysis of protein-coding genetic variation in 60,706 humans**. (2016) Nature.

The objective of the Variant Detection:

Identify the most likely **genotype for each genomic position from the individual**.

- - -

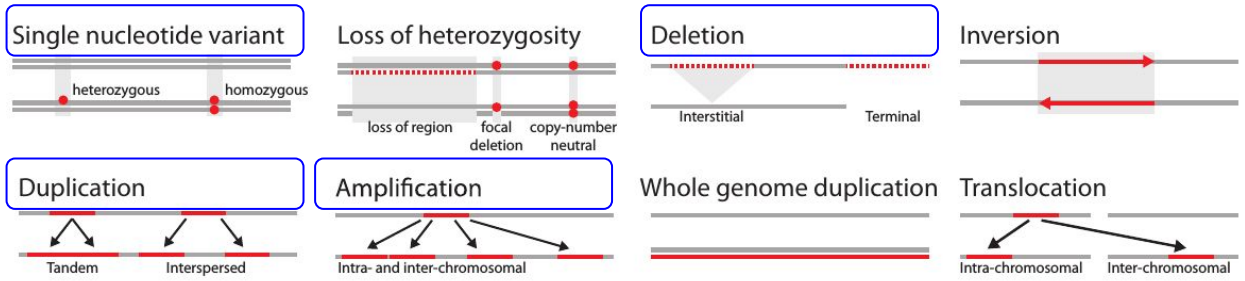
In Cancer genomics, if there is a **matched-normal sample** to be compared against the tumour sample:

+ Identify somatic variants (i.e. only in tumour sample).

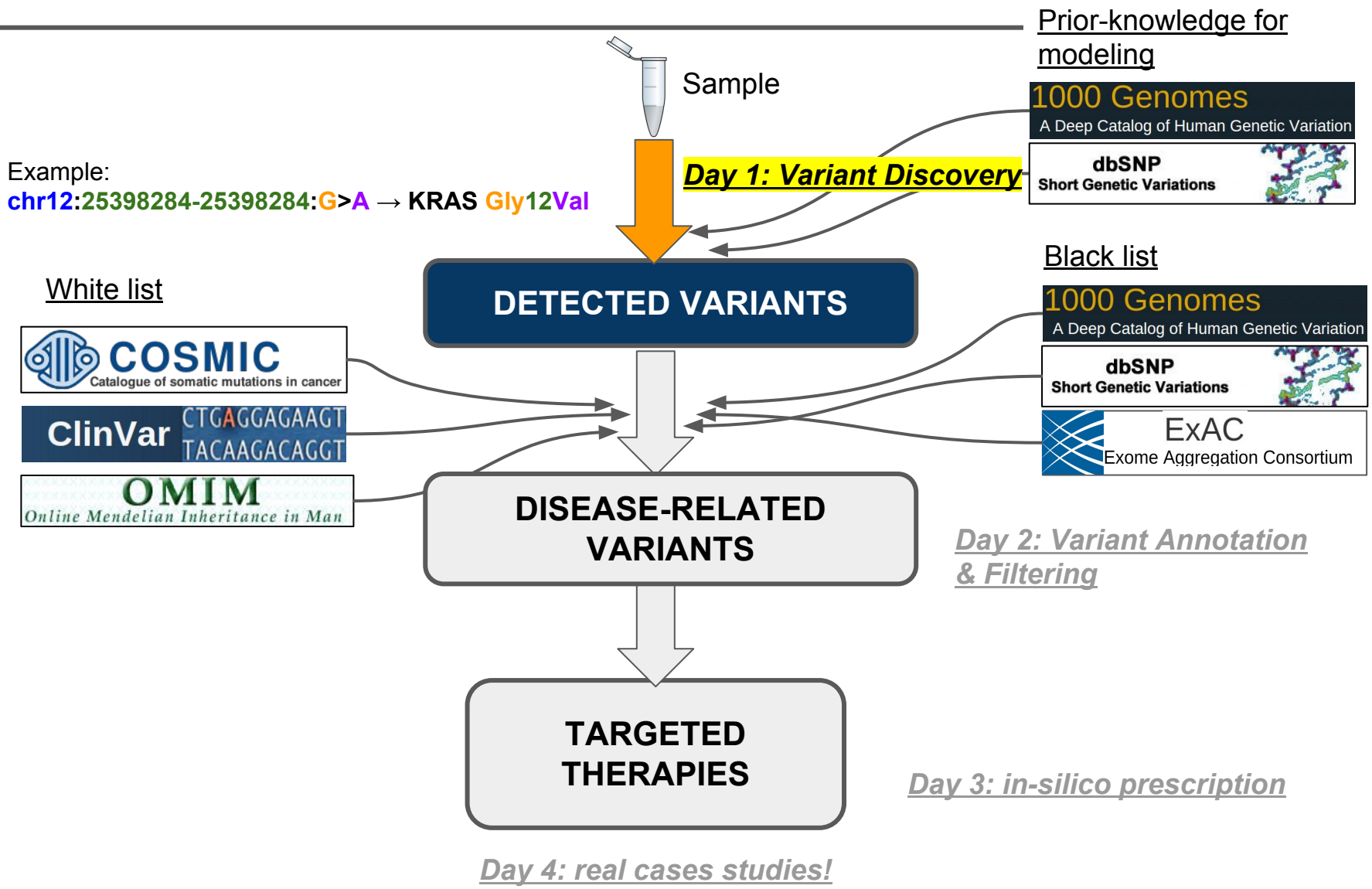
+ Identify copy-number alterations (large genomic aberrations).

Some concepts

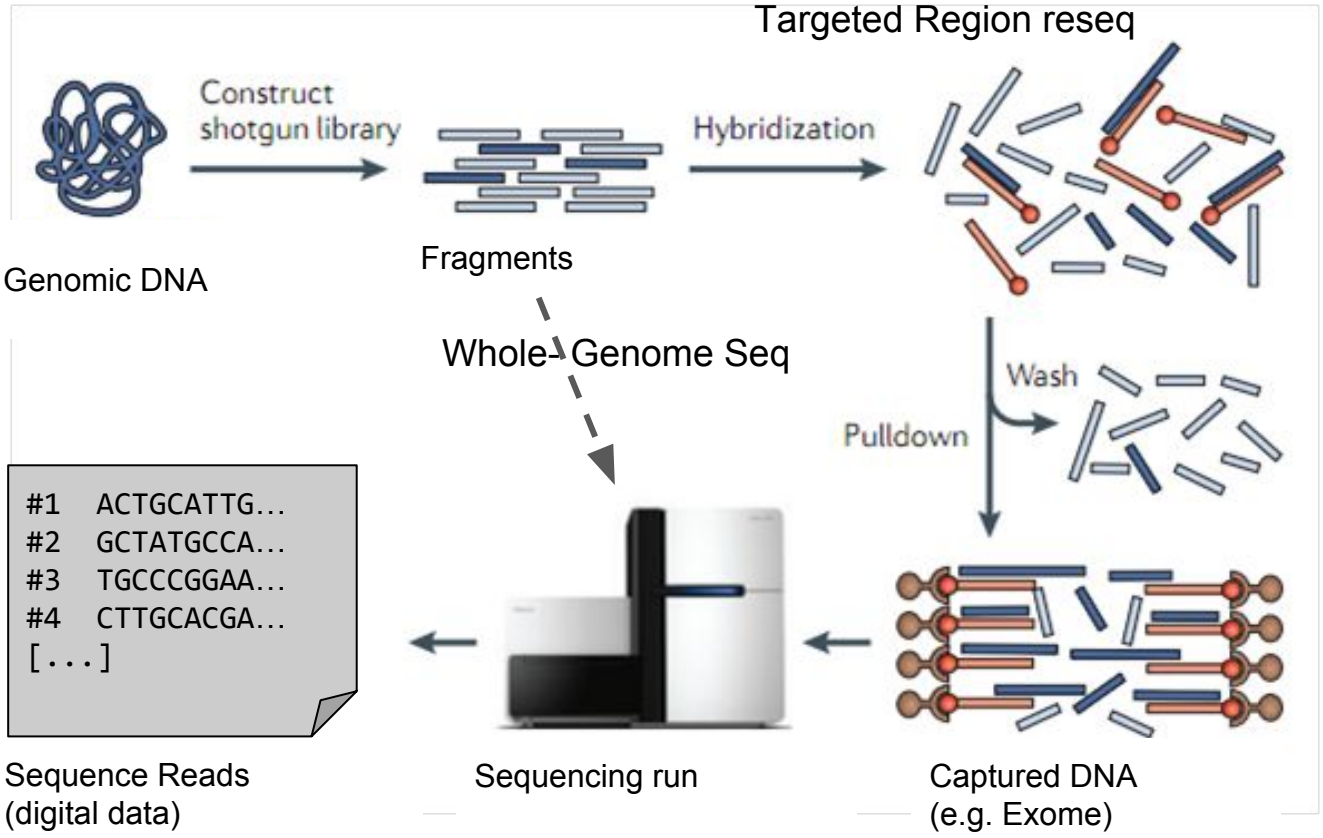
- What is a genetic **variant** ?
  Genetic differences in individuals as compared to a reference genome (built from a population).
  Nomenclature:
  - **First level**: Genomic position and nucleotide change.
    Chromosome Name:Genomic position (coordinates):Reference allele>Alternative allele
    Chr12:25398284-25398284:G>A
- Classes of variants :
  - **Germline** : inherited. E.g. a SNP, or a SNV related to a rare disease.
  - **Somatic** : acquired within a cell lineage. E.g. Cancer mutations.
- Polymorphism
  common variant in a given population (SNP, Single Nucleotide Polymorphism).
  Present in at least 1% in a population.
- Types of genomic variants:



Variants need a context



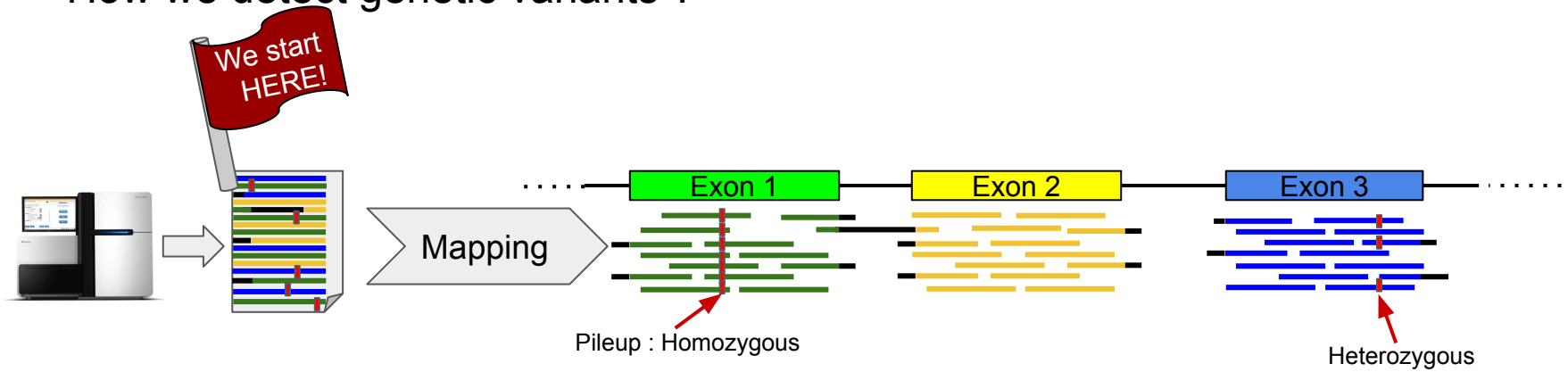
DNA Sequencing data generation



The sequence reads belong from the ends of the original fragment.

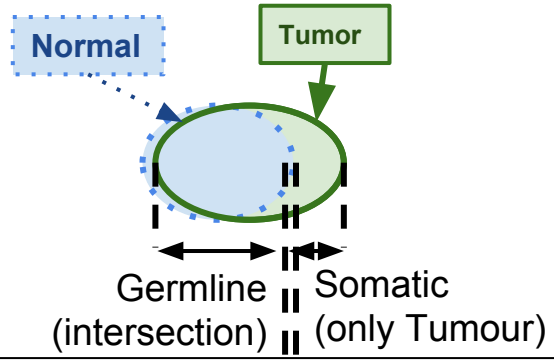
Fundamentals of variant detection

- How we detect genetic variants ?

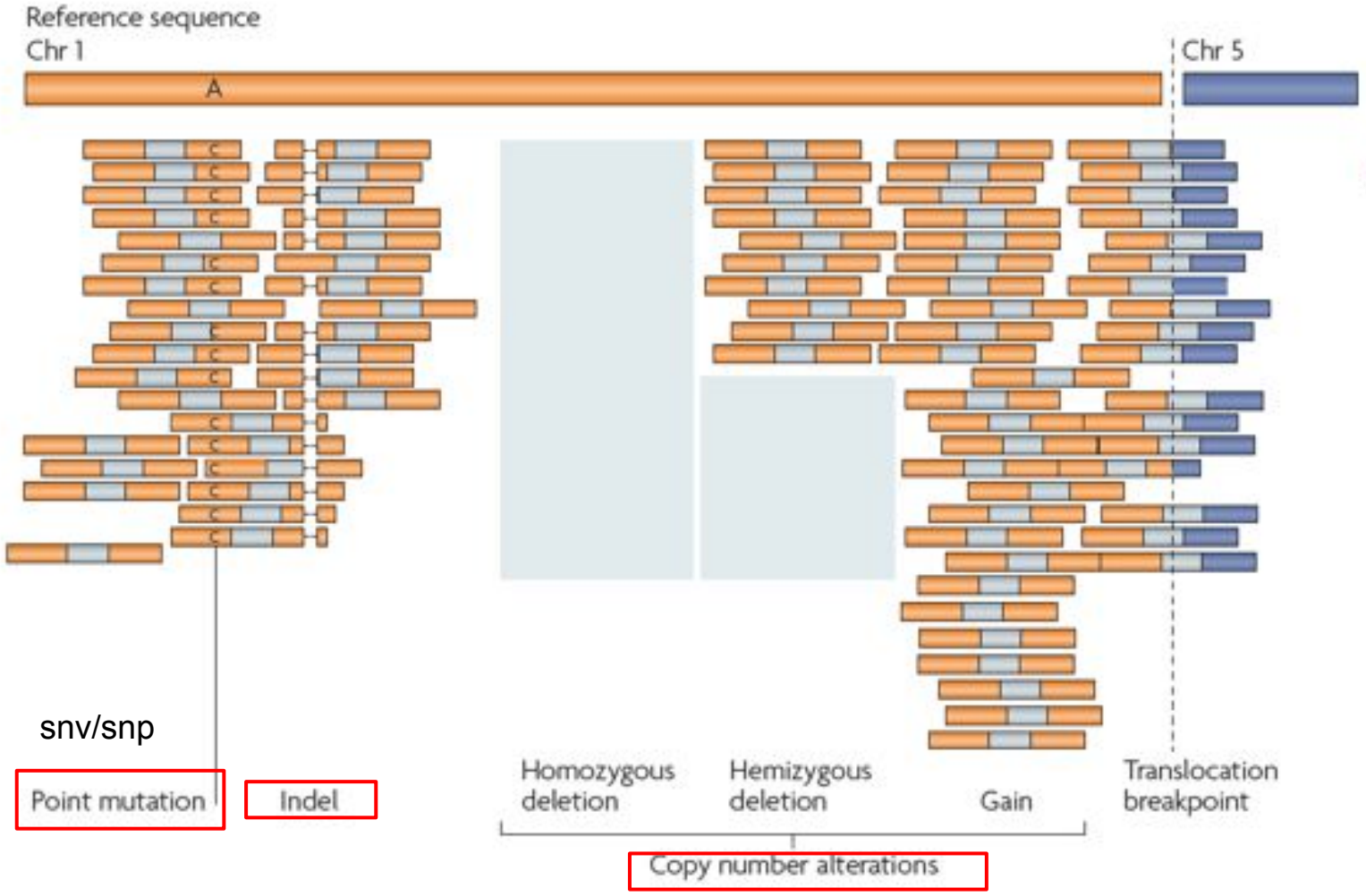


Are these differences true calls?
Statistical method to estimate the most likely genotype.

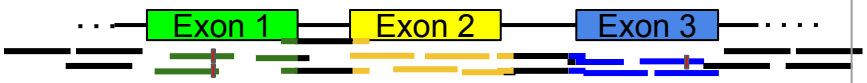
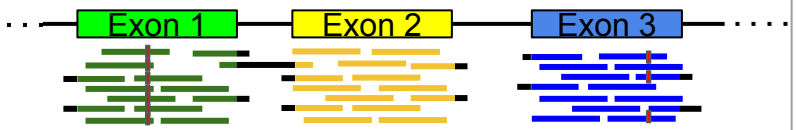
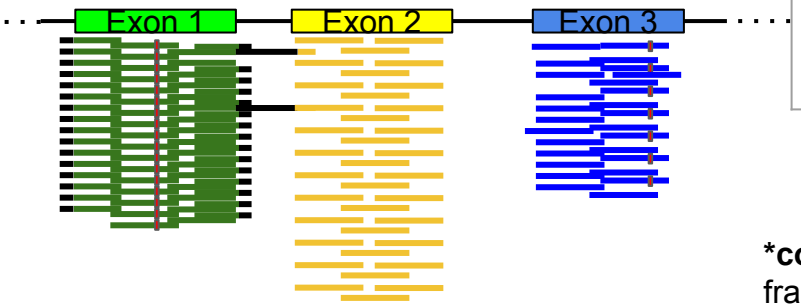
Discern somatic mutations by comparison:



Different types of variants detected by mapping reads

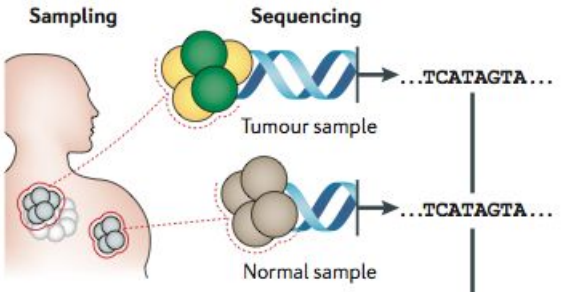


Whole-Genome, and targeted resequencing

	# bp pos seq	Type of variants discovered	Avg Coverage per pos	Cost
Whole-Genome Sequencing 	~100 Gb	**- coding variants*,** intronic and regulatory sites. - **Structural variants** - **CNA** #Variants= 3M - 4M.	30x	High
Whole-Exome Sequencing 	~32Mb 50Mb	- **coding variants***. - Some intronic and regulatory sites. - **CNA** (challenging). **#**Variants= 20k - 60k.	20x - 80x	Low
Panel of genes by amplicon/PCR approach 	*ND*	Depends on the design - Particular **coding variants*** - **CNA** (challenging) # variants = *ND*	1000x - 5000x	Low

***coding variants**: missense, stop gained, stop lost, frameshift, splice region...

Methods for Variant Detection



Several Methods have been published.

Analysis

Variant detection

SNVs and indels
Discover SNVs and small indels using WGS, exome sequencing and RNA-seq data

CNAs, SVs and gene fusions
Uncover large-scale CNAs, SVs and gene fusions using WGS and RNA-seq data

Example tools

VarScan	GATK
SomaticSniper	
Pindel	Strelka
MuTect	Bassovac
JointSNVMix	
BreakDancer	
Genome STRiP	
ChimeraScan	CREST
Hydra	GASV-pro
TIGRA	deFuse

Only WGS

More on CNAs

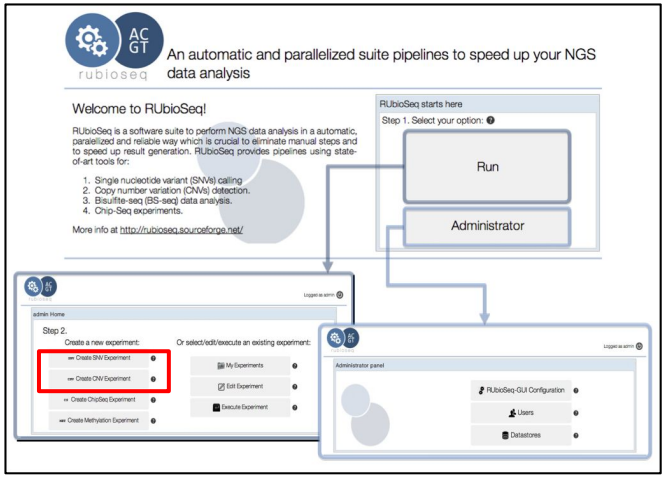
Tool	Year	Language	Paired or pooled data	Segmentation	Feature
ADTEx	2014	Python, R	Both	HMM	Noise reduction Ploidy estimation
CONTRA	2012	Python, R	Both	CBS	GC correction
Control-FREEC	2011	C++, R	Paired	LASSO	GC correction, mappability
EXCAVATOR	2013	Perl, R	Both	HSLM	GC correction, mappability, exon-size correction
ExomeCNV	2011	R	Paired	CBS	GC correction, mappability
Varscan2	2012	Java, Perl, R	Paired	CBS	GC correction

Appropriate methods for Whole-Exome seq

Further reading:

Ding, L., Wendl, M. C., McMichael, J. F. & Raphael, B. J. Nat Rev Genet – (2014). doi:10.1038/nrg3767
Nam J.N. et al. Evaluation of somatic copy number estimation tools for whole-exome sequencing data. Brief. Bioinformatics (2015)

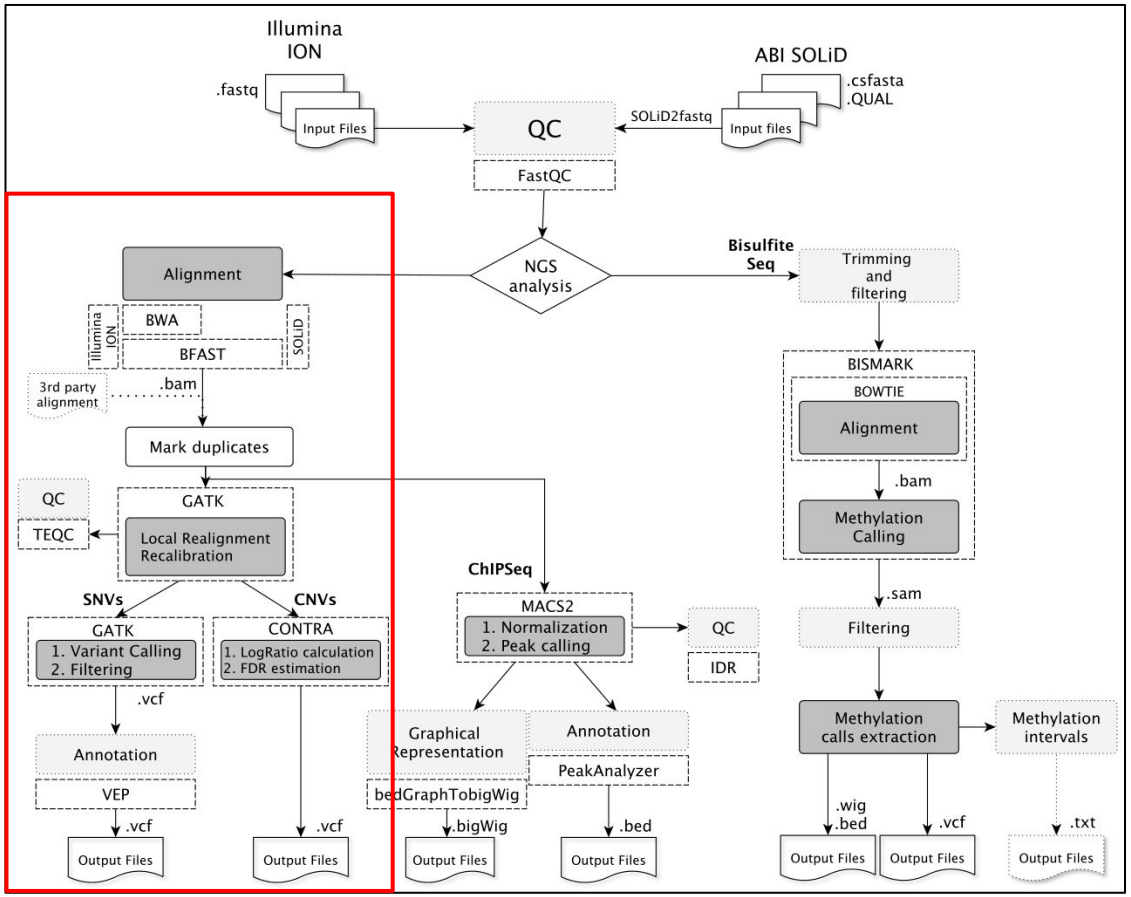
Our proposed Variant Calling Pipeline



Graphic User Interface

rubioseq

http://rubioseq.bioinfo.cnio.es/



workflow schema

Developed by the ***Bioinformatics Unit*** at the **Spanish National Cancer Research Centre** (Madrid, Spain).

Rubio-Camarillo *et al*. Comput Methods Programs Biomed. 2017 Jan;138:73-81

Rubio-Camarillo *et al*. Bioinformatics (2013) 29 (13), 1687-1689

What is Crucial in Variant calling

- For clinical practices, the use of **gold standard methods** and **reproducible analysis** are mandatory.

- The analysis is based on the comparison against the **reference genome** :
  A single consensus sequence for the whole genome. It was built up from a high quality set of representative samples of the specie (from different populations).
  It is the first-line comparison during analysis.
  By **Genome Reference Consortium (GRC)** (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/)

  - **Human assemblies (Versions):**
    + **`GRCh37/hg19`** : former version. Released in 2012. It is still the preference for analysis.
    + **`CRCh38/hg38`** : current version (Sep. 2017). Released in 2014. More accurate, comprehensive (includes Haplotypes) and sophisticated.

  "CRCh38 is here now, but still waiting."

- We must know what **regions along the genome** were sequenced in the experiment ? that is, the Sequencing library.

Bundle of files for Variant Detection

1. **Genome Reference** (standard 1000 Genomes, fasta).
2. List of **Target beats or intervals** of genomic regions sequenced by the Library protocol.
3. **dbSNP** (VCF file) for a recent dbSNP release (build 138, it includes the 1000 Genomes).
4. HapMap genotypes and sites VCFs
5. **OMNI 2.5 genotypes for 1000 Genomes samples** (VCF).
6. The current best set of **known indels** to be used for local realignment); use both files:
   - 1000G_phase1.indels.b37.vcf (currently from the 1000 Genomes Phase I indel calls)
   - Mills_and_1000G_gold_standard.indels.b37.sites.vcf

Q: How you can get this bundle of files?

A: you could get them from the **Broad Institute's FTP**

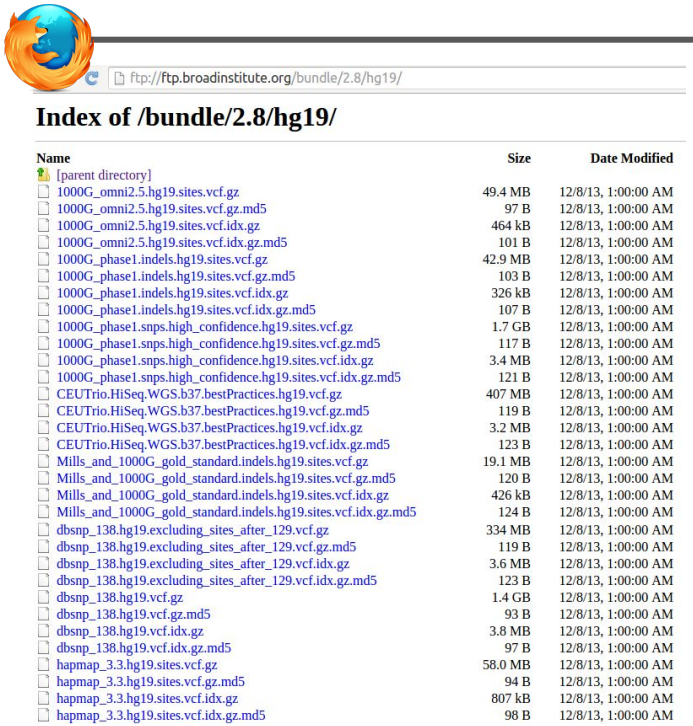
ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg19/

You also need the Intervals file from your NGS provider (Illumina, Ion Torrent,...)

For this workshop, we got these files for you.

More info.:
https://software.broadinstitute.org/gatk/download/bundle

Bundle from the GATK's repository



ftp://ftp.broadinstitute.org/bundle/2.8/hg19/

Index of /bundle/2.8/hg19/

Name	Size	Date Modified
[parent directory]		
1000G_omni2.5.hg19.sites.vcf.gz	49.4 MB	12/8/13, 1:00:00 AM
1000G_omni2.5.hg19.sites.vcf.gz.md5	97 B	12/8/13, 1:00:00 AM
1000G_omni2.5.hg19.sites.vcf.idx.gz	464 kB	12/8/13, 1:00:00 AM
1000G_omni2.5.hg19.sites.vcf.idx.gz.md5	101 B	12/8/13, 1:00:00 AM
1000G_phase1.indels.hg19.sites.vcf.gz	42.9 MB	12/8/13, 1:00:00 AM
1000G_phase1.indels.hg19.sites.vcf.gz.md5	103 B	12/8/13, 1:00:00 AM
1000G_phase1.indels.hg19.sites.vcf.idx.gz	326 kB	12/8/13, 1:00:00 AM
1000G_phase1.indels.hg19.sites.vcf.idx.gz.md5	107 B	12/8/13, 1:00:00 AM
1000G_phase1.snps.high_confidence.hg19.sites.vcf.gz	1.7 GB	12/8/13, 1:00:00 AM
1000G_phase1.snps.high_confidence.hg19.sites.vcf.gz.md5	117 B	12/8/13, 1:00:00 AM
1000G_phase1.snps.high_confidence.hg19.sites.vcf.idx.gz	3.4 MB	12/8/13, 1:00:00 AM
1000G_phase1.snps.high_confidence.hg19.sites.vcf.idx.gz.md5	121 B	12/8/13, 1:00:00 AM
CEUTrio.HiSeq.WGS.b37.bestPractices.hg19.vcf.gz	407 MB	12/8/13, 1:00:00 AM
CEUTrio.HiSeq.WGS.b37.bestPractices.hg19.vcf.gz.md5	119 B	12/8/13, 1:00:00 AM
CEUTrio.HiSeq.WGS.b37.bestPractices.hg19.vcf.idx.gz	3.2 MB	12/8/13, 1:00:00 AM
CEUTrio.HiSeq.WGS.b37.bestPractices.hg19.vcf.idx.gz.md5	123 B	12/8/13, 1:00:00 AM
Mills_and_1000G_gold_standard.indels.hg19.sites.vcf.gz	19.1 MB	12/8/13, 1:00:00 AM
Mills_and_1000G_gold_standard.indels.hg19.sites.vcf.gz.md5	120 B	12/8/13, 1:00:00 AM
Mills_and_1000G_gold_standard.indels.hg19.sites.vcf.idx.gz	426 kB	12/8/13, 1:00:00 AM
Mills_and_1000G_gold_standard.indels.hg19.sites.vcf.idx.gz.md5	124 B	12/8/13, 1:00:00 AM
dbsnp_138.hg19.excluding_sites_after_129.vcf.gz	334 MB	12/8/13, 1:00:00 AM
dbsnp_138.hg19.excluding_sites_after_129.vcf.gz.md5	119 B	12/8/13, 1:00:00 AM
dbsnp_138.hg19.excluding_sites_after_129.vcf.idx.gz	3.6 MB	12/8/13, 1:00:00 AM
dbsnp_138.hg19.excluding_sites_after_129.vcf.idx.gz.md5	123 B	12/8/13, 1:00:00 AM
dbsnp_138.hg19.vcf.gz	1.4 GB	12/8/13, 1:00:00 AM
dbsnp_138.hg19.vcf.gz.md5	93 B	12/8/13, 1:00:00 AM
dbsnp_138.hg19.vcf.idx.gz	3.8 MB	12/8/13, 1:00:00 AM
dbsnp_138.hg19.vcf.idx.gz.md5	97 B	12/8/13, 1:00:00 AM
hapmap_3.3.hg19.sites.vcf.gz	58.0 MB	12/8/13, 1:00:00 AM
hapmap_3.3.hg19.sites.vcf.gz.md5	94 B	12/8/13, 1:00:00 AM
hapmap_3.3.hg19.sites.vcf.idx.gz	807 kB	12/8/13, 1:00:00 AM
hapmap_3.3.hg19.sites.vcf.idx.gz.md5	98 B	12/8/13, 1:00:00 AM

Everything you need for the variant calling is stored in the FTP of the GATK team.
It includes the **human genome reference** too!

There are independent bundles for each assembly of the human genome (hg19, hg38,etc). The difference across them is the coordinate system for the annotations, but it is the same data.

You can download the different bundles from GATK's FTP (Broad Institute) visiting this URL with your Internet Browser:

ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/2.8/

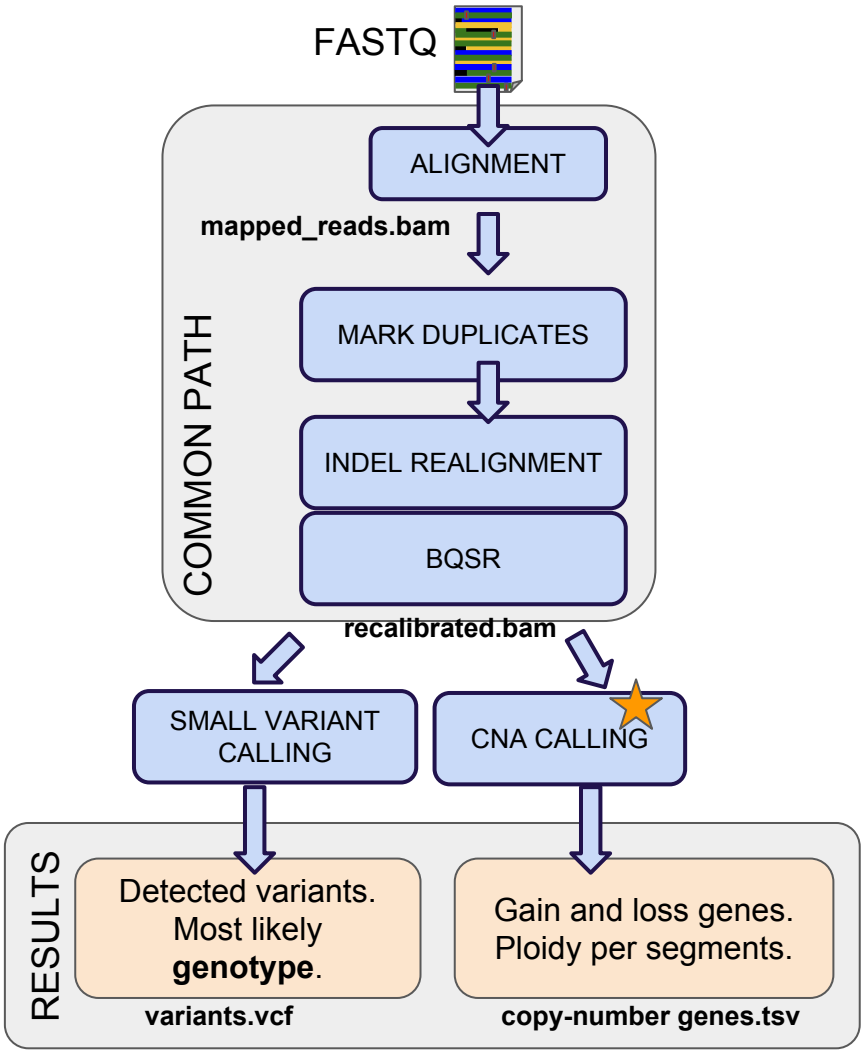
Tip for home: The following UNIX command downloads the whole bundle for **hg19** in one step (~hrs) :

```
$ wget -r -nH --cut-dirs=2 --reject-regex "NA12878|CEUTrio" \

-P /path/to/your_directory/ ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg19/*
```

More info.:
http://gatkforums.broadinstitute.org/discussion/1213/whats-in-the-resource-bundle-and-how-can-i-get-it

Point mutations and CNV Calling Process



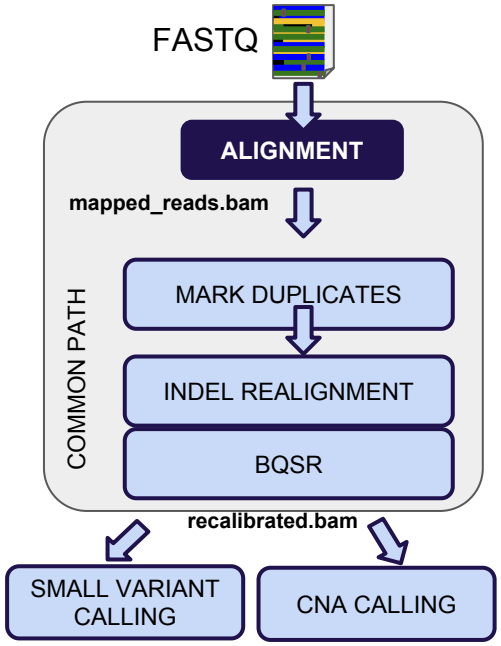
★ It requires **tumour and matched normal** sample (or a **panel of normals)**

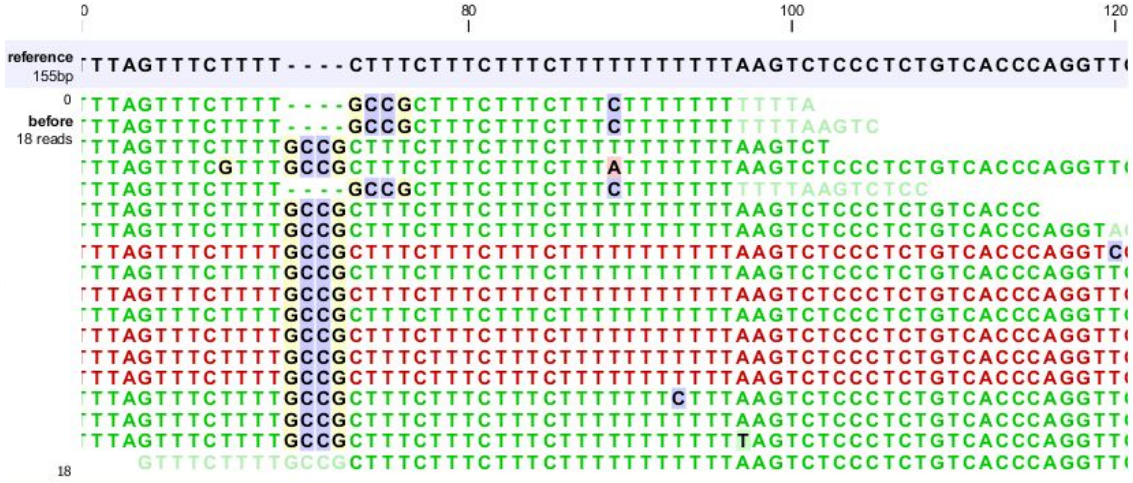
Variant Calling Pipeline

1. Alignment

rubioseq

WORKFLOW:



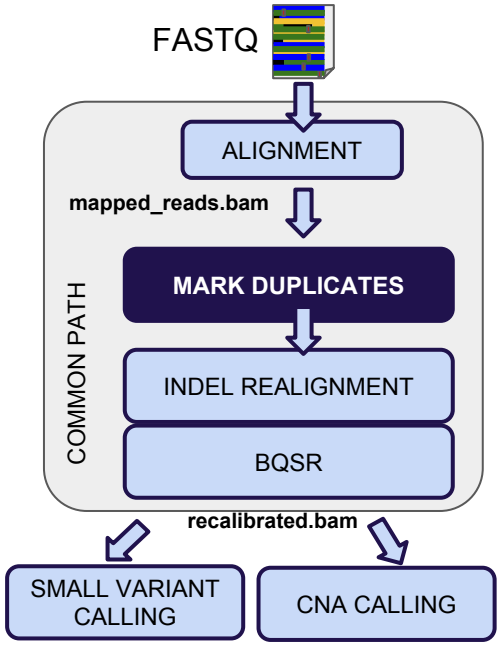


METHOD: by **BWA & Bfast+BWA**
https://github.com/lh3/bwa#citing-bwa
http://sourceforge.net/projects/bfast/files/bfast%2Bbwa/

- Fast mapping on the reference genome by creating indexes. It is computationally intensive, but it is done only once.
- Search for candidate sites to align a given read by using seeds (fragments of a read).

2. Mark duplicates

WORKFLOW:



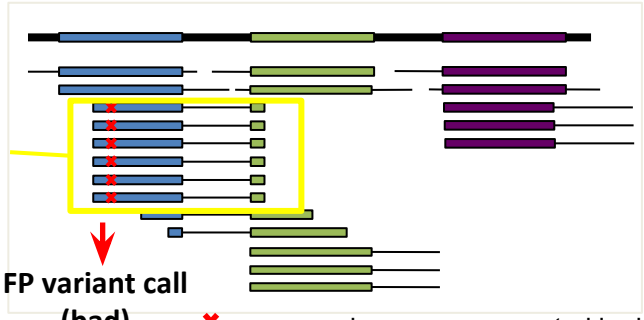
Under the hood:

- Duplicates derive from PCR amplification (library preparation): one fragment is sequenced multiple times.
- An error at the beginning of the PCR (first steps) is propagated.

- Therefore, duplicates are **worthless** for the analysis: *Duplicates are source of False Positives calls while only provide redundancy.*

Solution: retrieve the best one, discard the duplicates:

Duplicates share the **same alignment properties** : sequence, start and end positions



MARK DUPLICATES

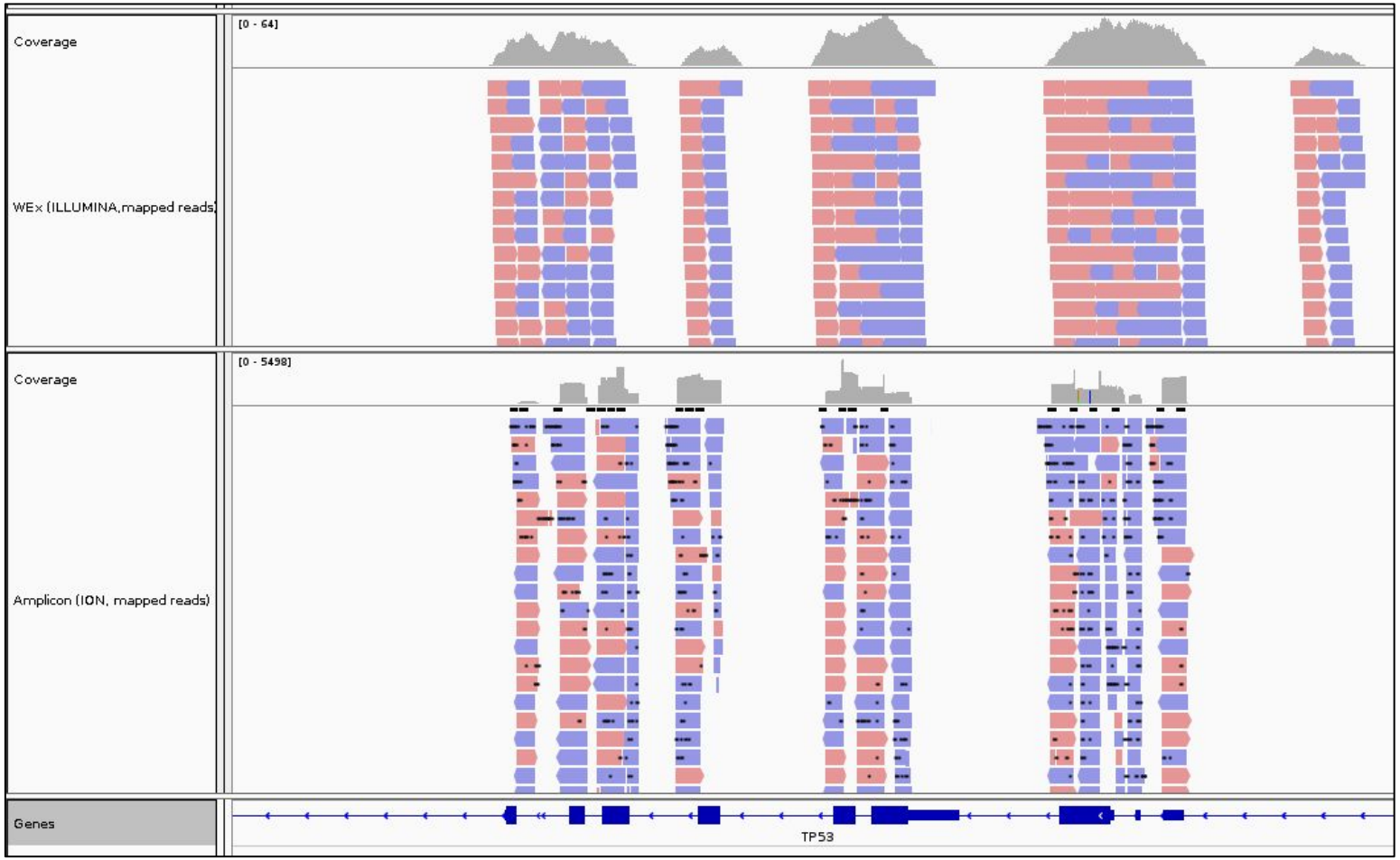
After marking duplicates, the variant caller will only see :

… and thus be more likely to make the right call

METHOD: by **Picard-tools**
http://broadinstitute.github.io/picard/
(alternatives : **samtools**)

Adapted from GATK

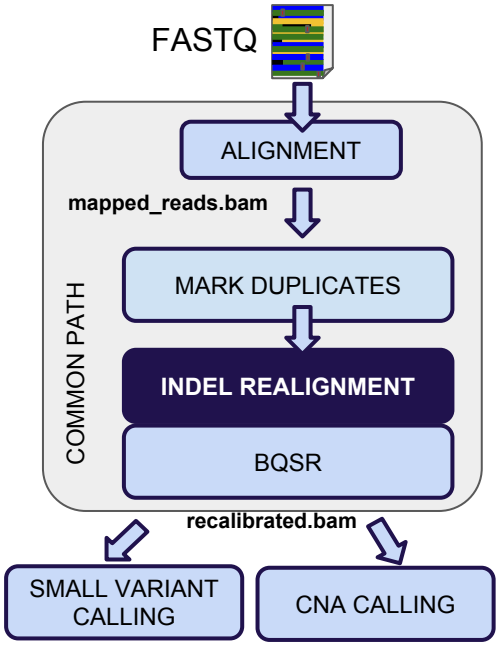
2. Mark duplicates: WEx Vs. Amplicon



WARNING: **Do NOT mark duplicates in data derived from** amplicon techniques (**Ion Torrent**).
More info.: http://gatkforums.broadinstitute.org/discussion/5847/remove-duplicates-from-targetted-sequencing-using-amplicon-approach

3. Indel realignment

WORKFLOW:



Algorithms align reads very fast with high accuracy, but not perfectly.

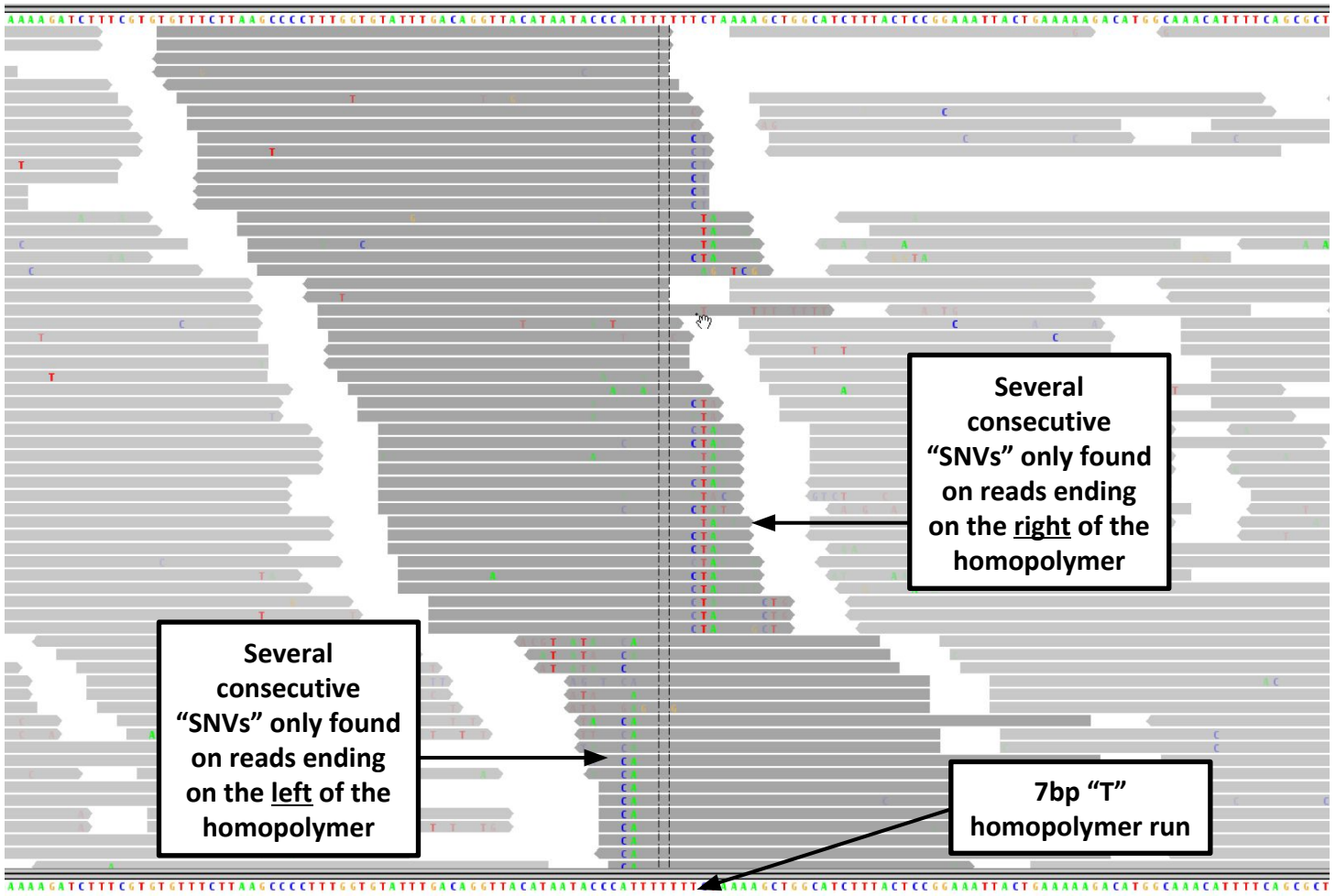
> *During alignment, penalties on mismatches are much cheaper than gaps (indels). Aligners will tend to choose Mismatches at the beginning, and locate indels in the rest.*

Also, there are sometimes multiple solution (alignments) for a given read. Aligners choose one randomly.

Variant calling requires the most perfect alignment as possible to avoid False Positives.

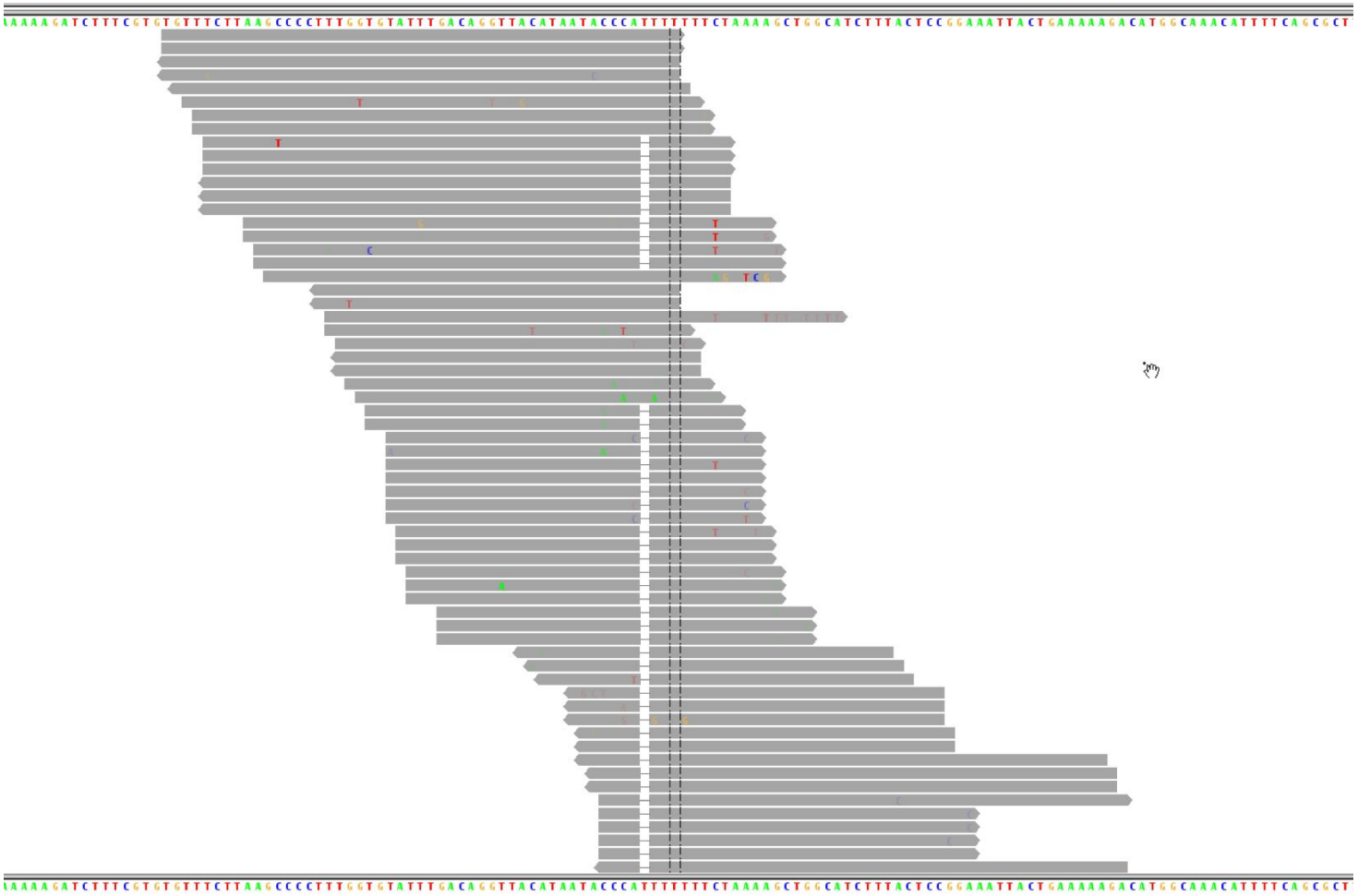
METHOD: by **GATK**
https://www.broadinstitute.org/gatk/gatkdocs/org_broadinstitute_gatk_tools_walkers_indels_IndelRealigner.php

3. Indel realignment



Taken from GATK team

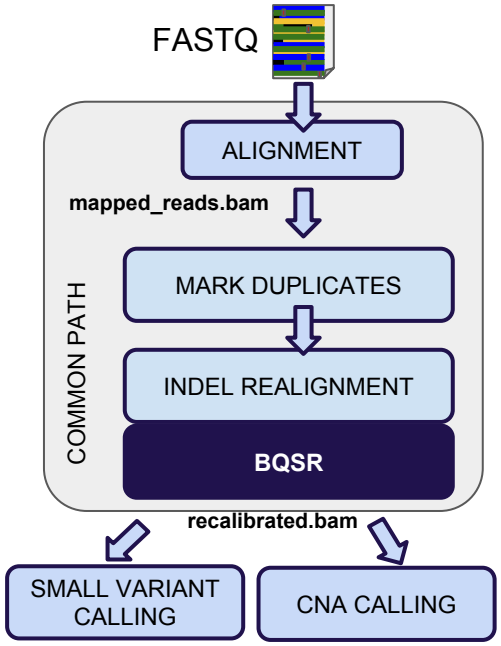
3. Indel realignment



Taken from GATK team

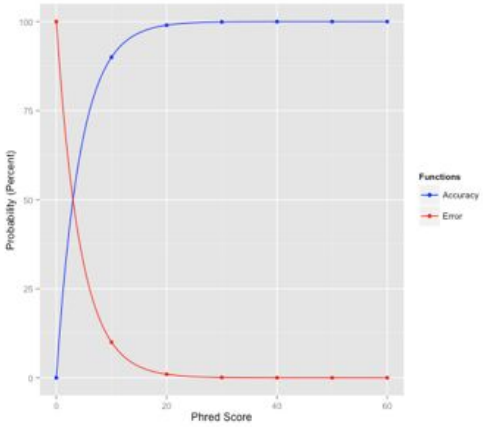
4. Base Quality Score Recalibration

WORKFLOW:

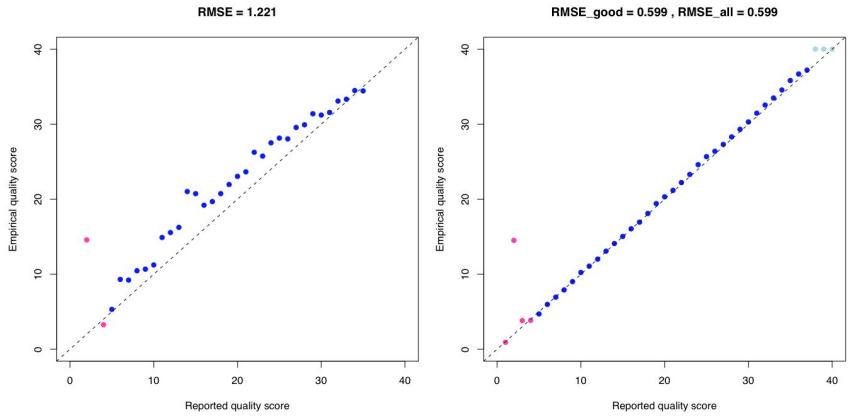


Phred Quality : each position of the sequence has its particular **base Quality score**.

The individual quality measures are NOT very important during the alignment step (mapping), but crucial during Variant calling.



Different NGS technologies have their particular bias in QS depending on the context. **They** could correct **empirically** these biases.



Original

After BQSR recalibration

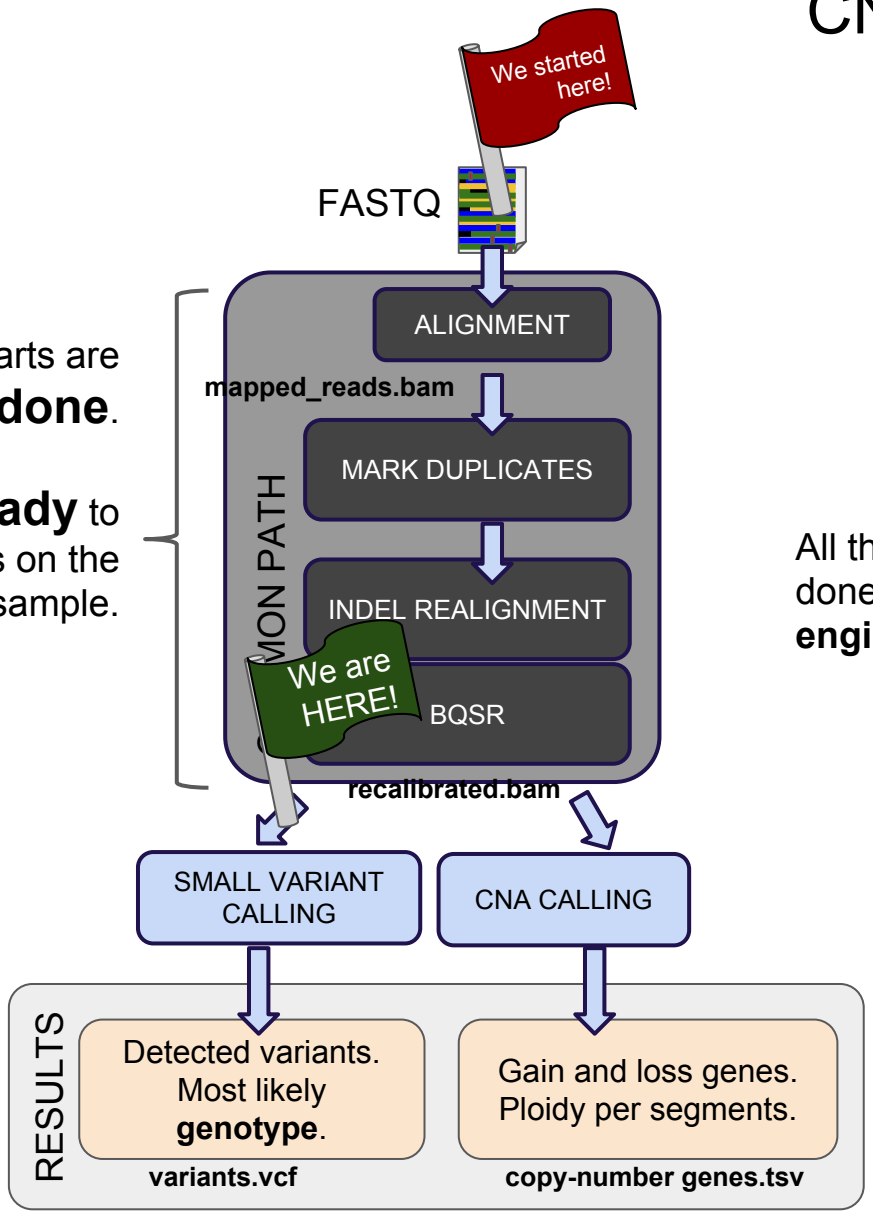
METHOD: by **GATK**
http://gatkforums.broadinstitute.org/discussion/44/base-quality-score-recalibration-bqsr

Point mutations and CNV Calling Process

rubioseq

The first two parts are **done**.

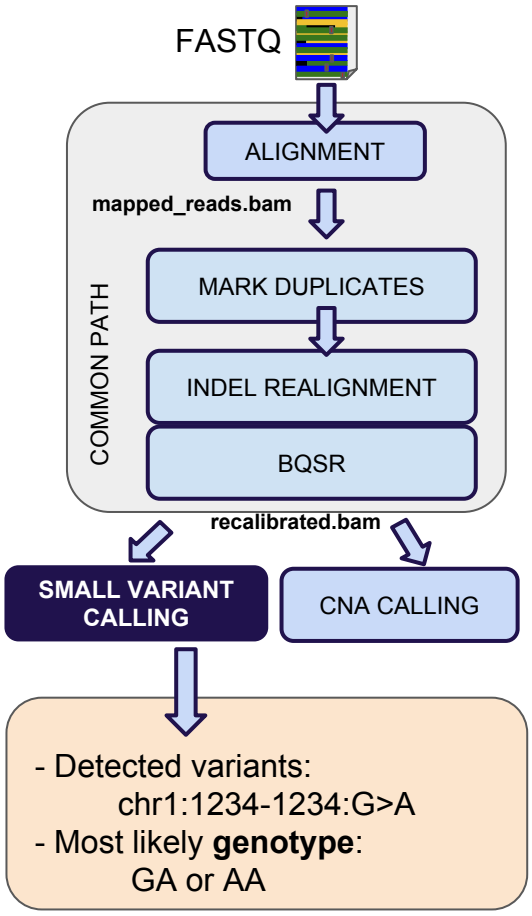
We are **ready** to discover variants on the sample.



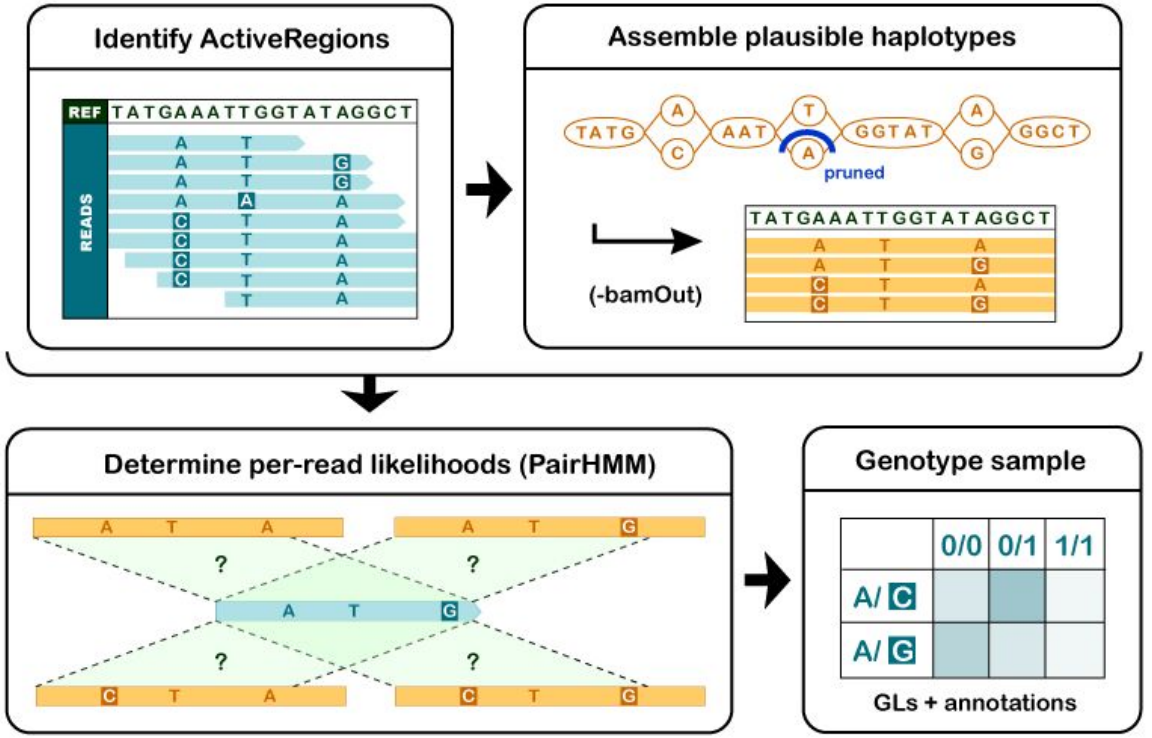
All these steps are **automatically** done by the pipeline (**RUbioSeq's engine**).

5. GATK Variant Calling Process : SNV & Indels

WORKFLOW:



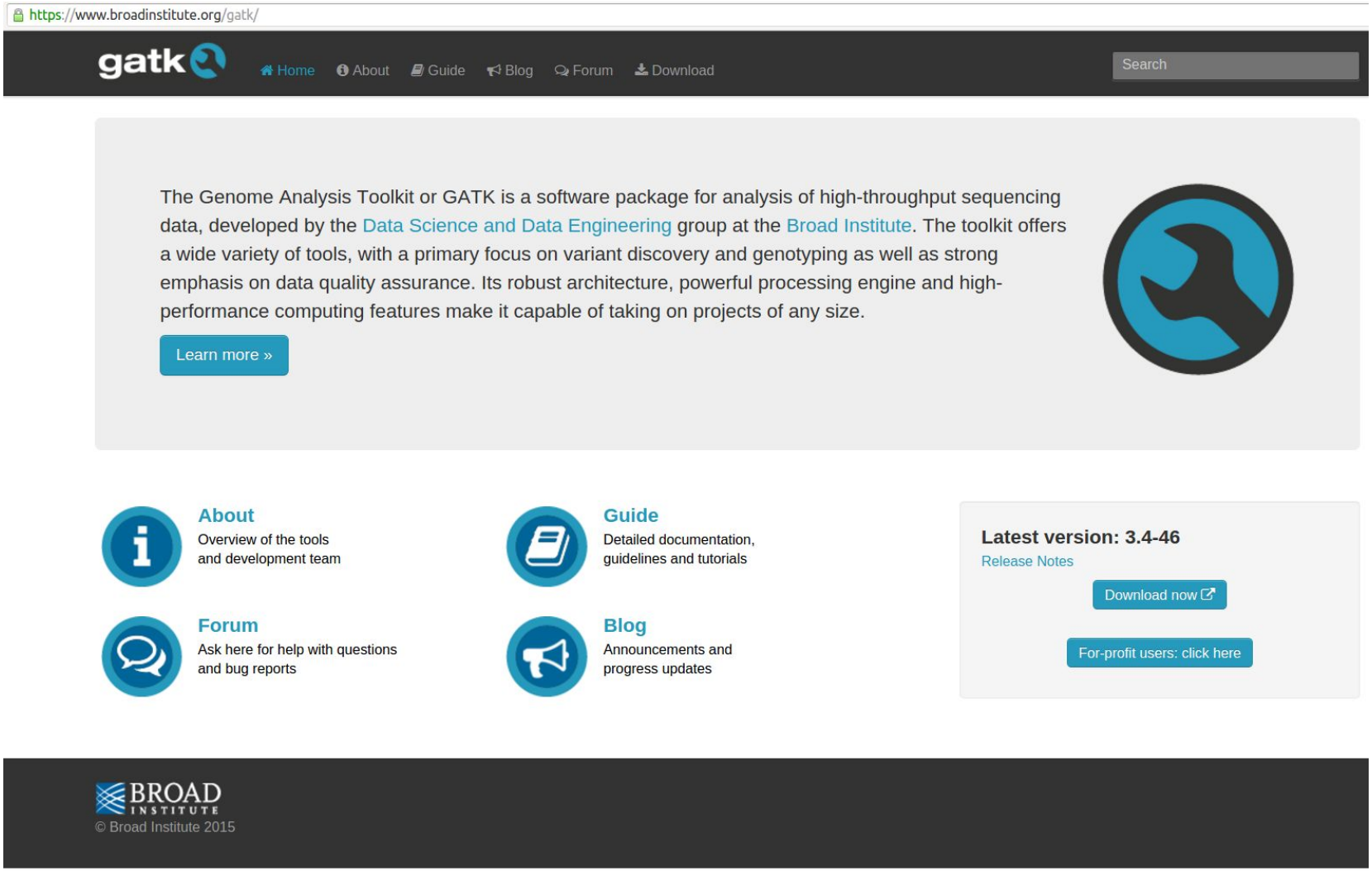
Haplotype Caller (new, included in **RUbioSeq v3.8.1)** : Variant calling based on the calculation of genotype likelihoods:



Assumptions: Diploid genome (2n).
Limitation: Allele freq > 20%.

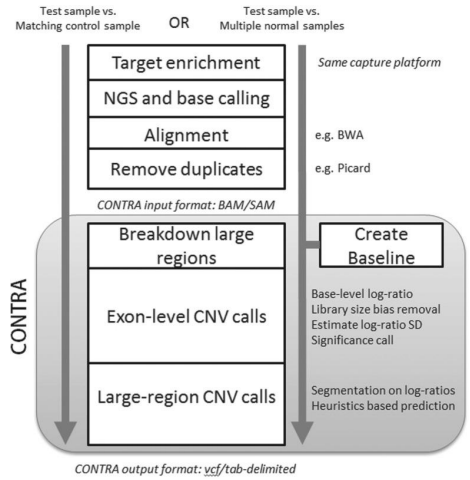
Further reading:
http://gatkforums.broadinstitute.org/discussion/4148/hc-overview-how-the-haplotypecaller-works
HC steps 1-4: https://software.broadinstitute.org/gatk/documentation/topic?name=methods

GATK is in active development



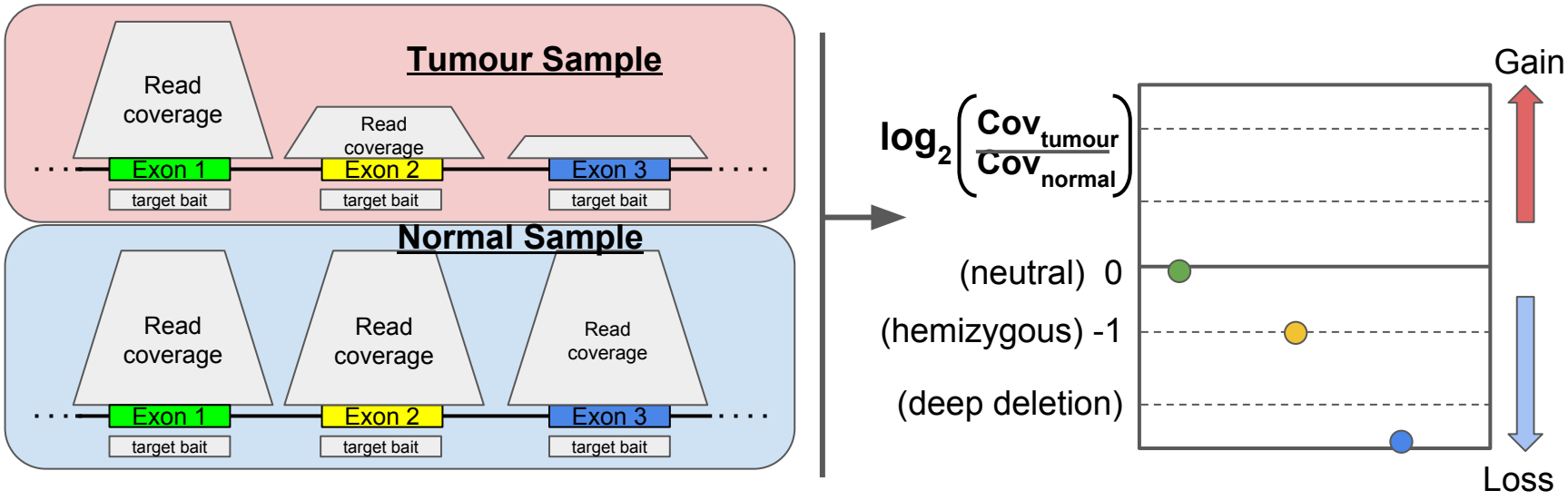
GATK is in active development

6. CNA Variant Calling



- **Normalization**: Split large regions. GC-content bias, unbalanced library size effect on log-ratios.
- Read-depth coverage & log2 CN ratio are corrected.
- Significance:
  Assumption: log2-transformed coverage fits a normal distribution:

$$RLR \sim N(\mu_d, \sigma_d)$$; Two-tailed P-value. multiple testing correction (FDR).



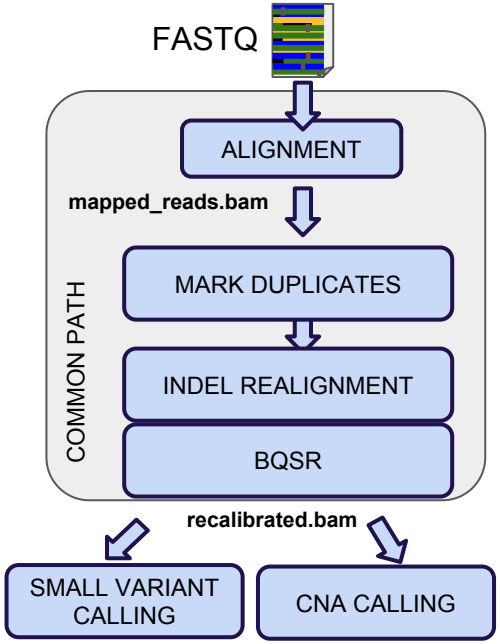
Li J et al. CONTRA: copy number analysis for targeted resequencing. (2012) Bioinformatics

What have we learnt?

Main concepts:
- Variants.
- How to detect them.
- Differences between platforms.

Requirements:
- Files.
- Methods

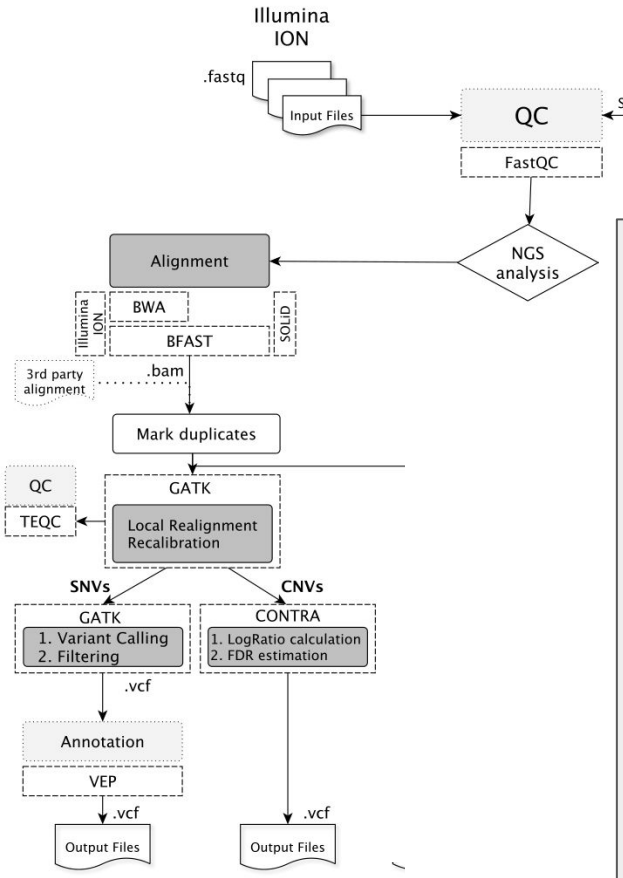
step-by-step WORKFLOW



Questions?

Thanks to Miriam Rubio-Camarillo & Gonzalo Gómez for the support and development of RUbioSeq.

Configuration files



It is already installed.

Let's configure it!

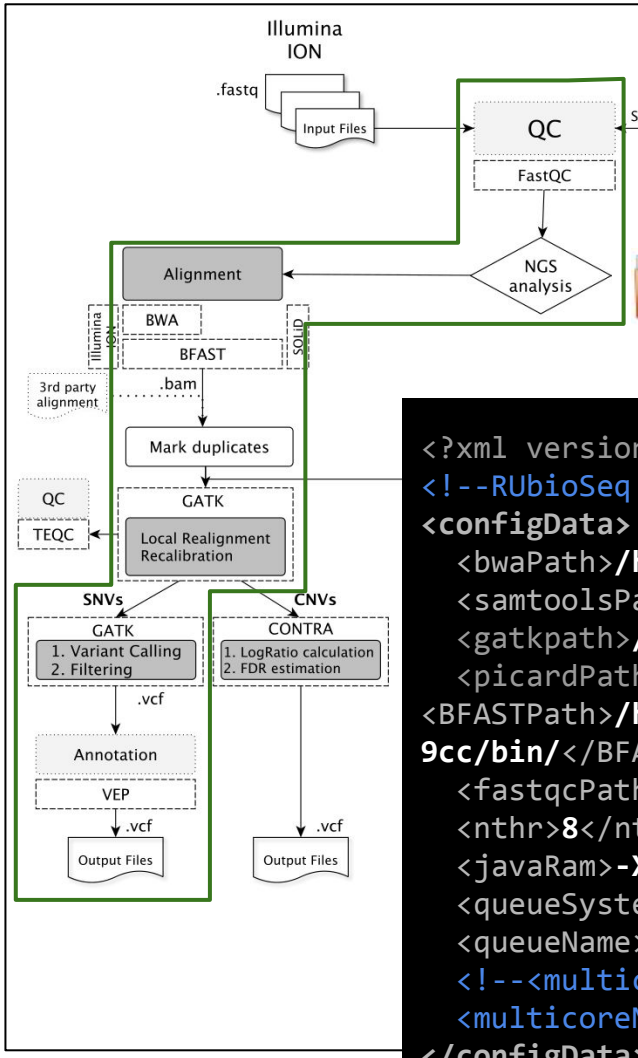
- What will we do?

Configure the pipeline for the detection of:

- Small variants (SNV + indels)
- Copy-Number Variants

First hands-on :: Configure RUbioSeq+

Configuration for Small variant analysis



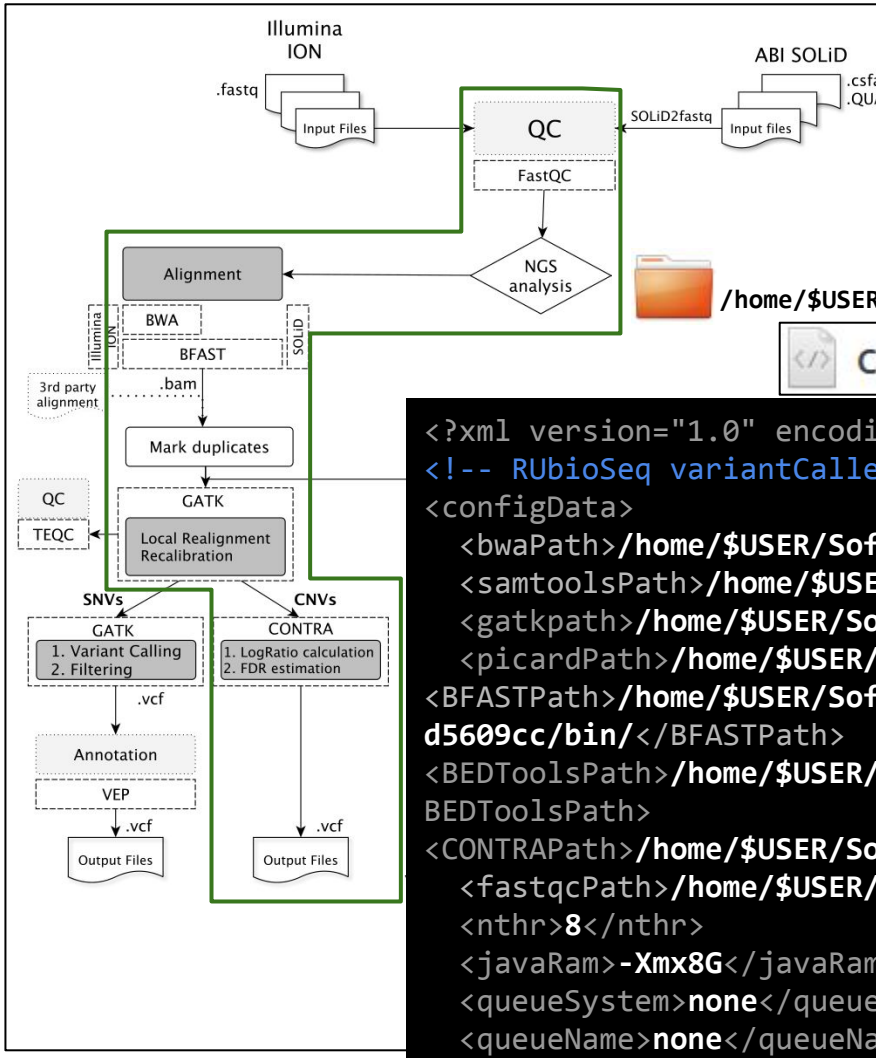
/home/$USER/Software/RUbioSeq3.8.1/variantCalling/config/snv

configProgramPaths.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<!--RUbioSeq variantCaller CONFIG FILE-->
<configData>
  <bwaPath>/home/$USER/Software/bwa-0.7.10/</bwaPath>
  <samtoolsPath>/home/$USER/Software/samtools-0.1.19/</samtoolsPath>
  <gatkpath>/home/$USER/Software/GATK-3.4-0/GenomeAnalysisTK-3.4-0/</gatkpath>
  <picardPath>/home/$USER/Software/picard-tools-1.107/</picardPath>
<BFASTPath>/home/$USER/Software/bfast-bwa-ed42c18ea7f48af862935be52f1c072b1d560
9cc/bin/</BFASTPath>
  <fastqcPath>/home/$USER/Software/FastQC/</fastqcPath>
  <nthr>8</nthr>
  <javaRam>-Xmx8G</javaRam>
  <queueSystem>none</queueSystem>
  <queueName>none</queueName>
  <!--<multicoreName>multicore</multicoreName>
  <multicoreNumber>4</multicoreNumber>-->
</configData>
```

First hands-on :: Configure RUbioSeq+

Configuration for CNV analysis



/home/$USER/Software/RUbioSeq3.8.1/variantCalling/config/cnv

configProgramPaths.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<!-- RUbioSeq variantCaller CONFIG FILE -->
<configData>
  <bwaPath>/home/$USER/Software/bwa-0.7.10/</bwaPath>
  <samtoolsPath>/home/$USER/Software/samtools-0.1.19/</samtoolsPath>
  <gatkpath>/home/$USER/Software/GATK-3.1-1-g07a4bf8/</gatkpath>
  <picardPath>/home/$USER/Software/picard-tools-1.107/</picardPath>
<BFASTPath>/home/$USER/Software/bfast-bwa-ed42c18ea7f48af862935be52f1c072b1
d5609cc/bin/</BFASTPath>
<BEDToolsPath>/home/$USER/Software/CONTRA.v2.0.3/BEDTools-Version-2.11.2/</
BEDToolsPath>
<CONTRAPath>/home/$USER/Software/CONTRA.v2.0.3/</CONTRAPath>
  <fastqcPath>/home/$USER/Software/FastQC/</fastqcPath>
  <nthr>8</nthr>
  <javaRam>-Xmx8G</javaRam>
  <queueSystem>none</queueSystem>
  <queueName>none</queueName>
</configData>
```