HIGH-PERFORMANCE BIOLOGICAL COMPUTING
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

H3ABioNet Data Management workshop

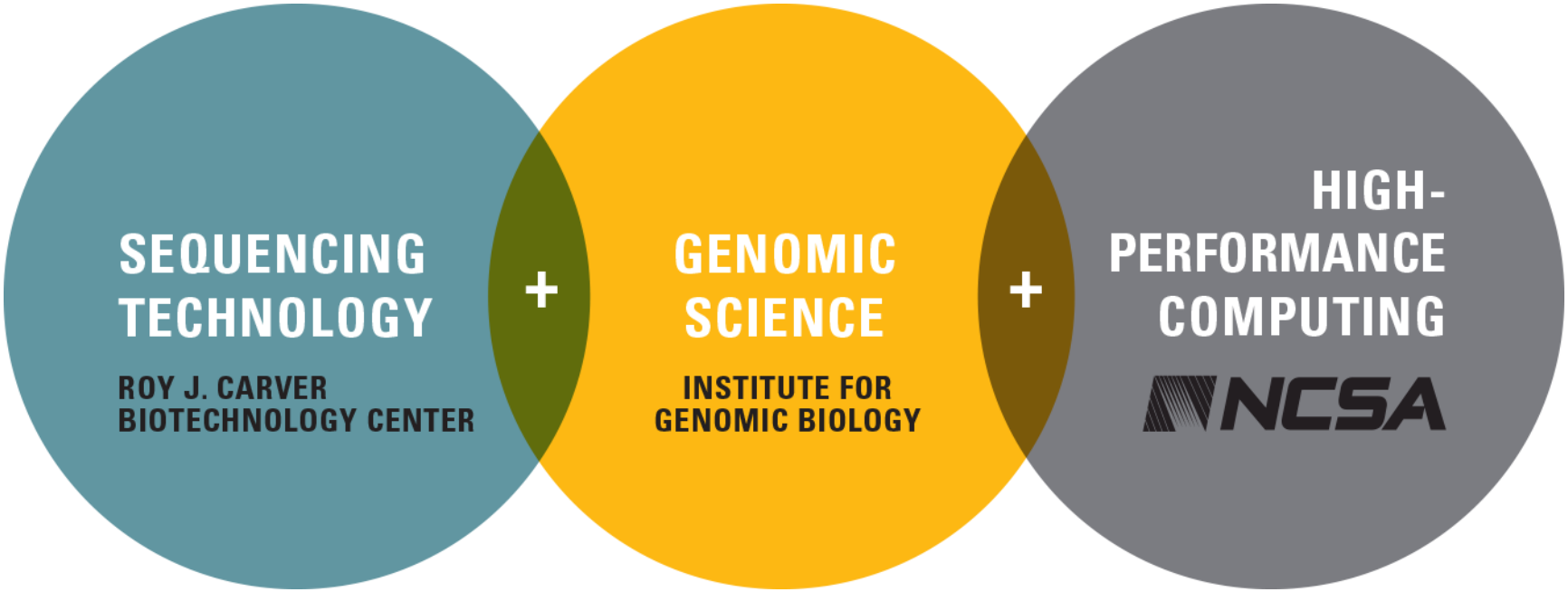
NGS data analysis for Variant Calling

June 5th, 2014

Radhika Khetani, Ph.D.

Technical Lead at HPCBio

University of Illinois, Urbana-Champaign


SEQUENCING TECHNOLOGY
ROY J. CARVER BIOTECHNOLOGY CENTER
+
GENOMIC SCIENCE
INSTITUTE FOR GENOMIC BIOLOGY
+
HIGH-PERFORMANCE COMPUTING
NCSA

Genome sequencing and Variant Calling

Introduction to using NGS for Variant Detection

- Sequencing Technologies, specifically Illumina
- File Formats, FASTQ, SAM, BAM, vcf, bcf
- QC steps
- Variant Calling (data processing)

Tea Break

- Computational Requirements
  - Data Storage
  - Processing Capacity

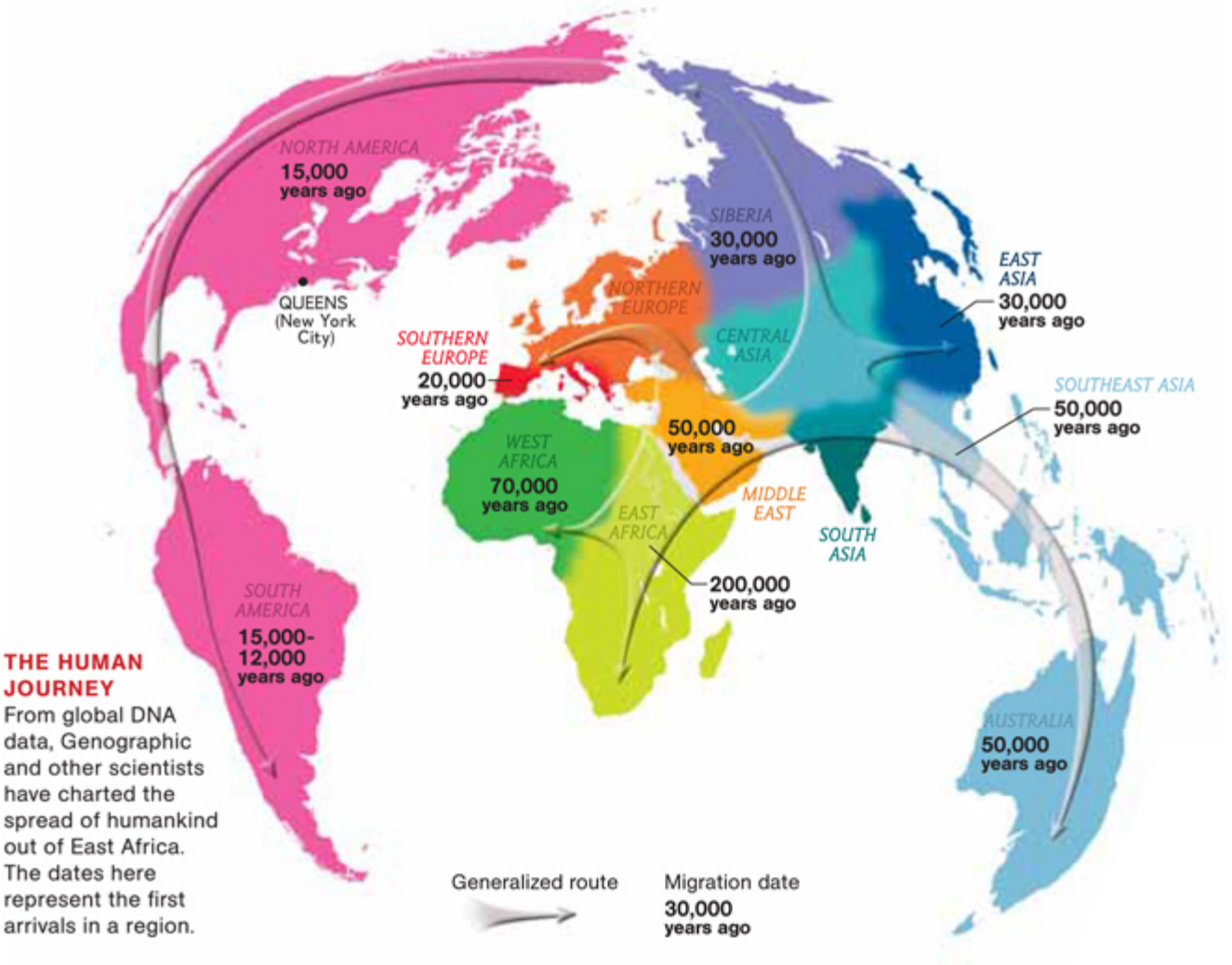
Brief Introduction to using NGS for microbiome analysis

Introduction to using NGS for Variant Detection

Human genomic diversity

- All anatomically modern humans outside Africa descend from a set of relatively small populations that left the continent less than 100,000 years ago. Populations within Africa are much more genetically diverse.
- Until ~500 years ago, there was relatively little admixture between these populations except from events linked to a few large-scale migrations (e.g. invasions of Europe by Central Asians).
- The phenotypic and genotypic diversity seen among these populations stems from two factors: genetic drift and selection based on reproductive fitness. Cultural as well as environmental differences affect traits conferring increased reproductive fitness.
- Extensive genotyping has made it possible to correlate sets of genetic variants (haplotypes) with very specific populations and to reconstruct the ancestry of many living individuals.

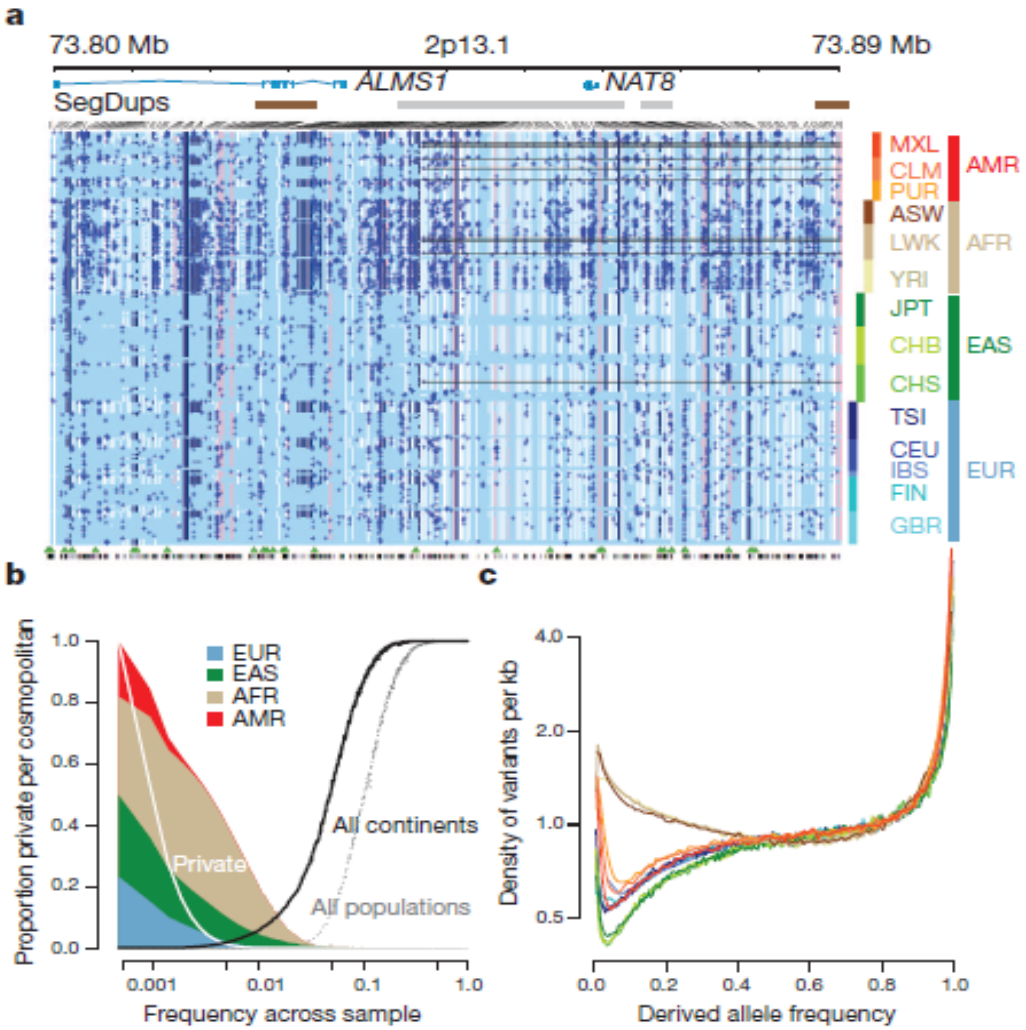
Wandering humans



Modern humans migrated out of Africa, gradually populating the globe in relatively small groups. Current human genetic diversity mirrors the routes and timings of these migrations.

Source: National Geographic, 2009

How diverse are we?



Data from 1000 Genomes project (Nov 2012)

a. Variation across populations (Americas, Africa, East Asia, Europe) in a 90 kb genome interval

b. Frequency of private, continent-specific and population-specific variants

c. Density of variants as a function of their frequency

How do human genomes differ?

Single nucleotide polymorphisms (SNPs):

- At a given position in the genome, some haplotypes carry one nucleotide while others carry another; the vast majority of SNPs are bi-allelic
- It is believed that the vast majority of SNPs present at a minor allele frequency of >5% worldwide have been characterized and deposited in dbSNP, although this may not be true for some African populations

Copy number variants (CNVs):

- Many regions of the genome have been duplicated during evolution, and there are haplotypic differences in copy numbers between individuals; CNVs can range between a few nucleotides and tens of thousands in size

Structural variants

- Regions of an individual haplotype can be inverted, deleted, or translocated relative to the reference genome sequence

△ Most of these variants are not directly pathogenic!

Phenotypic impacts

Most human genomic variants have no phenotypic impacts

Most of those that do have phenotypic impacts are either positively selected (i.e. they confer a reproductive advantage) or neutral

- Typically, they affect traits like height, facial features, hair or skin color, often associated with ethnic origin

Some genomic variants have effects that are deleterious to health

- Most of these are recessive: their effect is observed only if both alleles are affected; these recessive alleles are often associated with specific ethnic groups
- Those that are dominant will either be selected against and disappear, or have effects that minimally impact reproductive fitness (e.g., adult cancer)

This implies that the vast majority of alleles commonly found in the population do not directly cause disease

How to assess genomic diversity?

Methods are available to assess all three major sources of diversity:

SNPs, copy number variants, and structural variants

For SNPs, many different methods have been used:

- Hybridization based, primarily SNP arrays
- Enzyme-based methods, primarily oligonucleotide ligation and RFLP
- Methods measuring physical properties of DNA

For copy number variants, the main methods are hybridization based

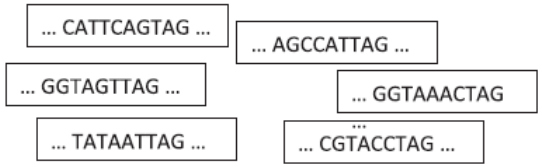
For structural variants, there are no universally accepted methods, but the most reliable ones use partial sequencing of large clones (e.g. fosmids)

High-throughput sequencing should be able to detect all types of variants

Genome Sequencing

Genomic DNA

Next-generation DNA sequencing



millions-billions of ***reads***
~30-1000 nucleotides

Resequencing

Align reads to *reference genome* and identify variants

***De novo* assembly**

Construct genome sequence from overlaps between reads

Benjamin J. Raphael*

Department of Computer Science and Center for Computational Molecular Biology, Brown University, Providence, Rhode Island, United States of America

Genome Resequencing & Variant Detection

- Pros
  - Its per base cost is cheaper than Sanger sequencing
  - It is getting cheaper allowing for large studies to be executed
  - It makes truly Genome-Wide analyses feasible
- Cons
  - The datasets are large and require relatively large computational infrastructure for data storage and processing
  - Some ambiguity in final results (but this can be overcome with stringent methodologies)

Human Reference Genome

- hg19 – most commonly used
- hg38 – the new version released at the end of 2013

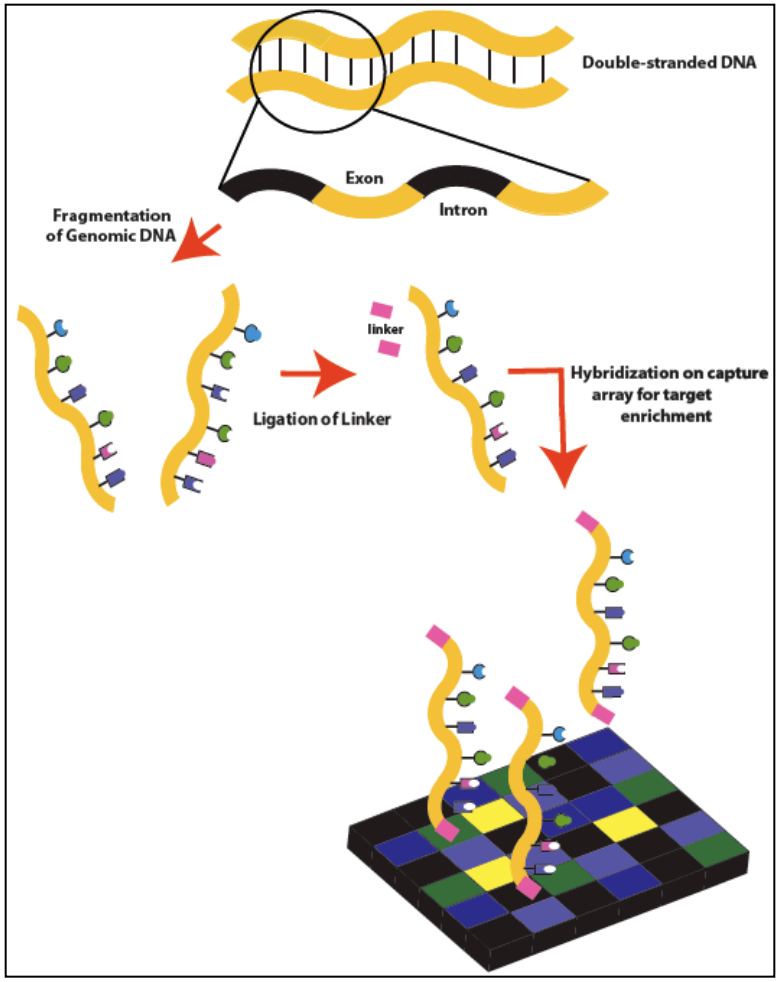
  "GRCh38 is the second major release of the human reference assembly made by the GRC. This release affects chromosome coordinates, includes 261 alternate loci scaffolds and corresponding alignments that provide chromosome context, and replaces centromere gaps with modeled sequence. The GRC resolved 1008 issues."

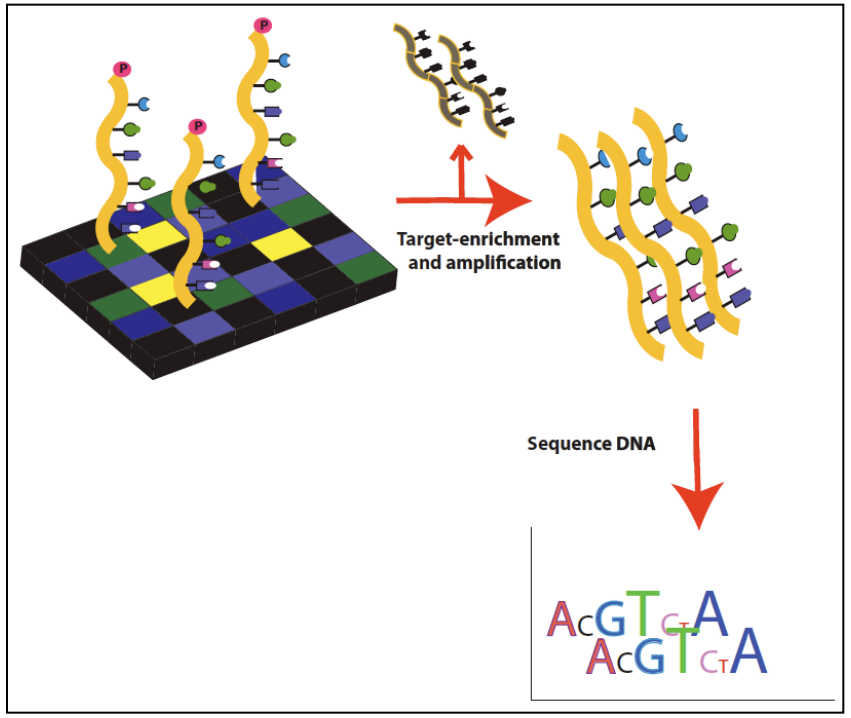
  - http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/

Variant Calling – Types of Data

- Whole Genome Sequencing (WGS)
  - Fragment genomic DNA
  - Sequence all the fragments
- Exome Sequencing
  - Capture DNA pieces that are known to be transcribed (exons) using arrays with sequence similarities
  - Amplify these pieces and sequence them

Exome analysis



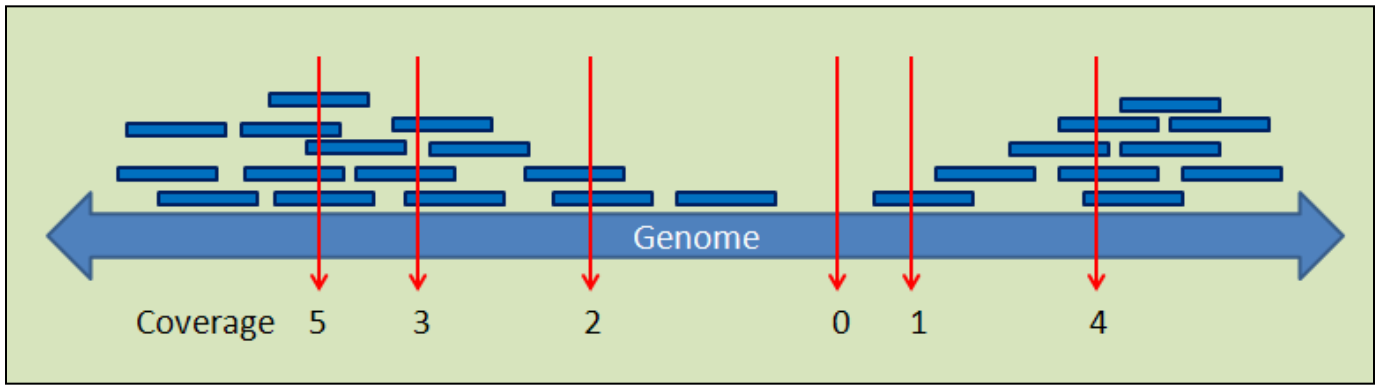


Variant Calling – Types of Data

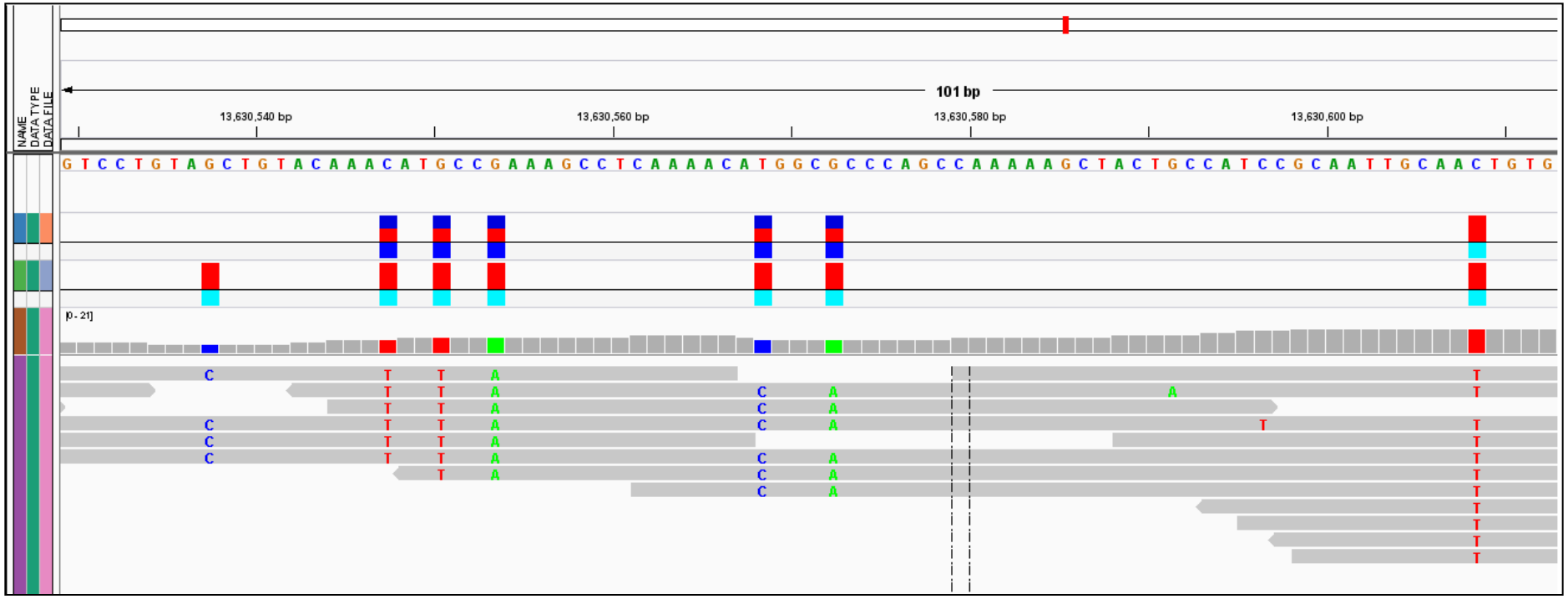
- Whole Genome Sequencing (WGS)
  - Fragment genomic DNA
  - Sequence all the fragments
- Exome Sequencing
  - Capture DNA pieces that are known to be transcribed (exons) using arrays with sequence similarities
  - Amplify these pieces and sequence them
  - Most known exonic regions captured, but not all
  - Smaller dataset with concentrated information
  - Less sequencing necessary to reach the same depth of coverage

Variant Calling – "Coverage" & "Depth"

- Coverage – What % of the genome sequence is represented in the sequencing data
- Depth of coverage – How many times is every base in the genome represented (on average)



Variant Calling – Depth of coverage



Variant Calling – Depth of coverage

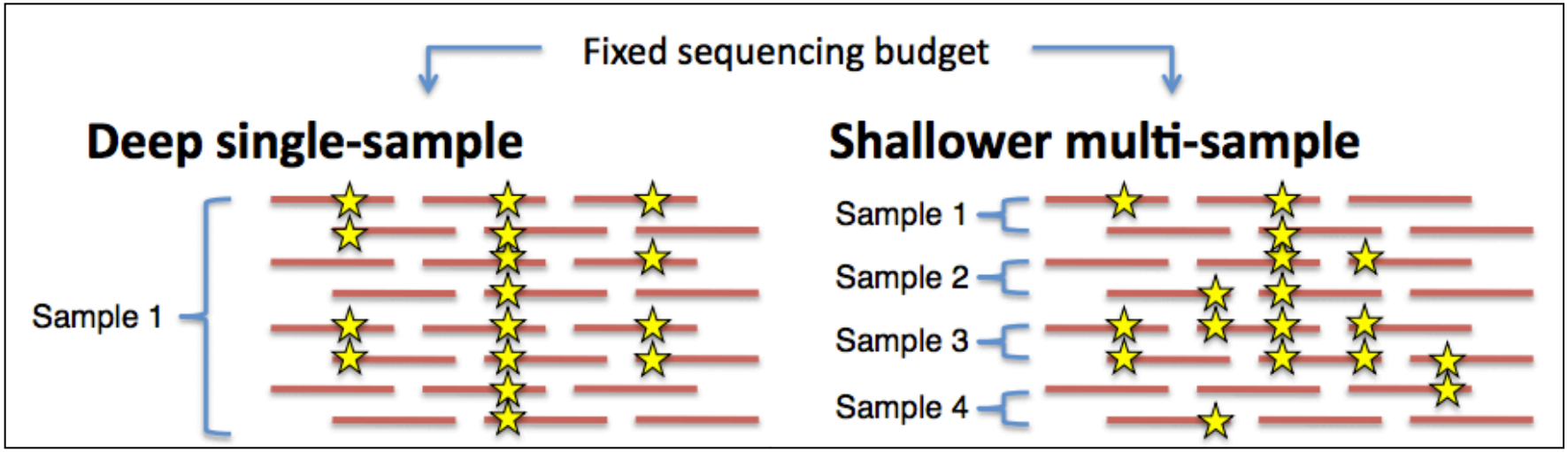
For WGS

- Haploid genome size => 3.2 Giga base pairs (3.2 billion)
- 50x coverage => ~160 Gbp

For Exome Sequencing

- Exome size => 33 Mega base pairs (33 million bases)
- 50x coverage => ~1.65 Gbp
- About 100 times smaller than WGS
- Depending on your method of capture, this number can vary

Variant Calling – Depth of coverage



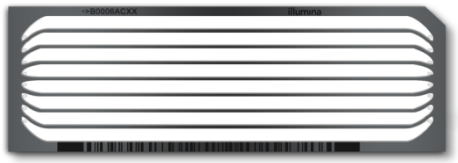
Sequencing Technologies

(Illumina)

HiSeq 2500 Sequencing System

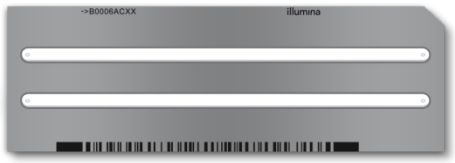
HIGH OUTPUT

50 – 600Gb
2 – 11 days
2 x 100bp max



RAPID RUN

10 – 180Gb
7 – 40 hours
2 x 150bp max



Larger projects, fewer runs

Smaller projects, quick results

MiSeq v3 Sequencing System

Flowcell (1 lane)

Reads: 250nt-300nt in length

Yield per run:
25 to 50 million paired-reads

Applications:
16s rRNA
Sequencing of small genomes
(bacteria, fosmids, BACs, virus)
Targeted sequencing (exome capture)
de novo transcriptome assembly

Turnaround time: ****FAST****

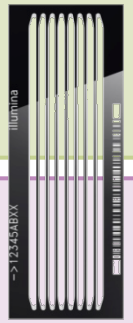
Illumina Sequencing Workflow

1 Library Preparation

- Fragment DNA
- Repair ends
- Add A overhang
- Ligate adapters
- Purify

2 Cluster Generation

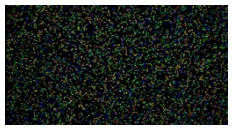
- Hybridize to flow cell
- Extend hybridized template
- Perform bridge amplification
- Prepare flow cell for sequencing



3 Sequencing

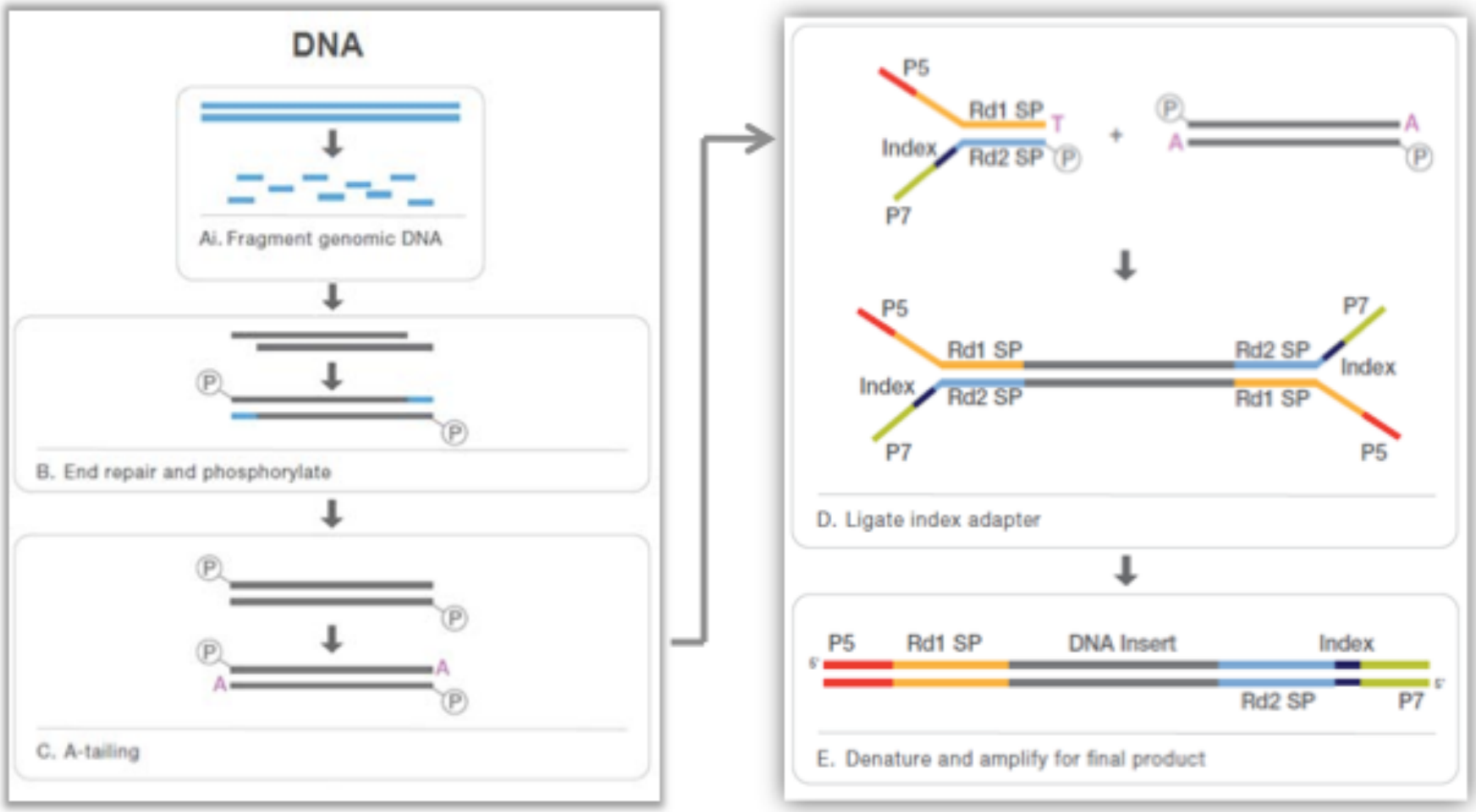
- Perform sequencing
- Generate base calls

4 Data Analysis



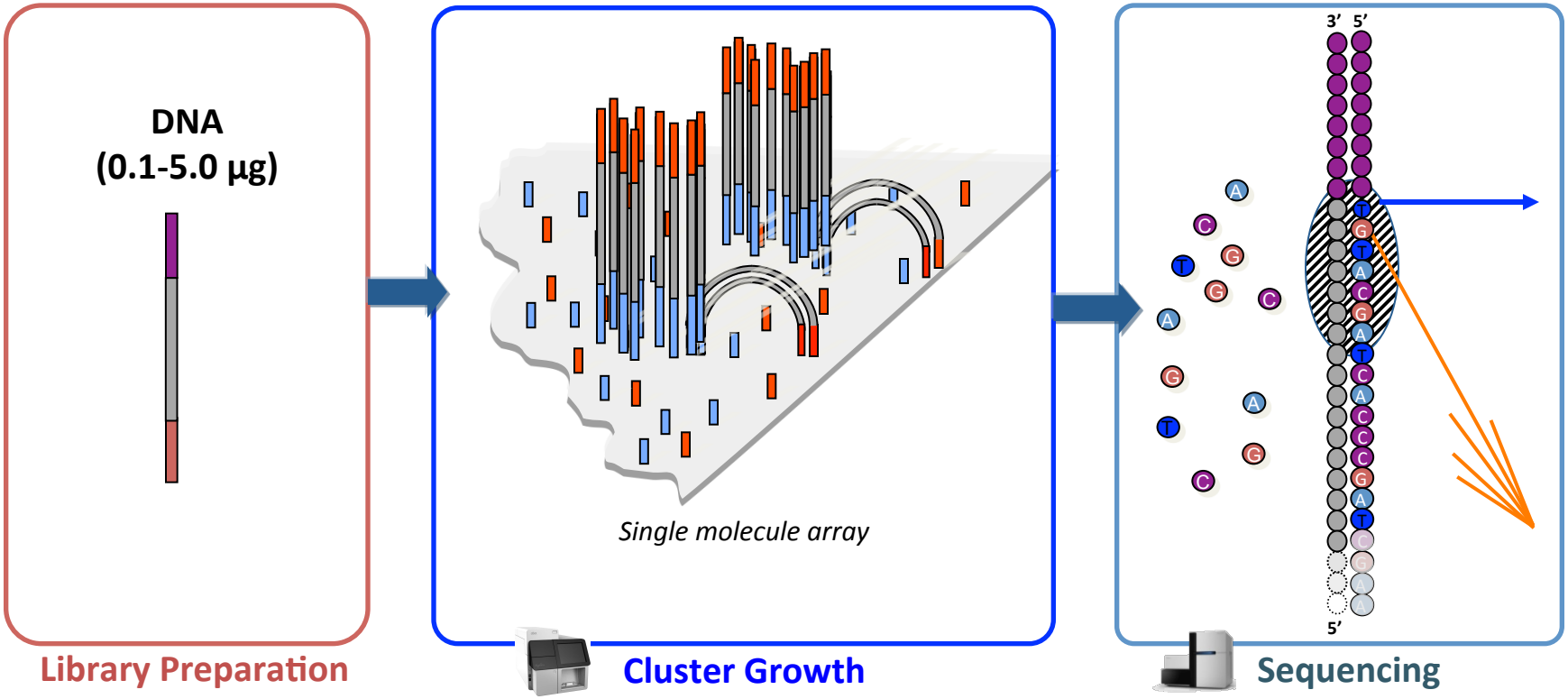
- Images
- Intensities
- Reads
- Alignments

Libraries = ds DNA ligated to “Y” adaptors



Illumina Sequencing Technology: Reads and BarCoding

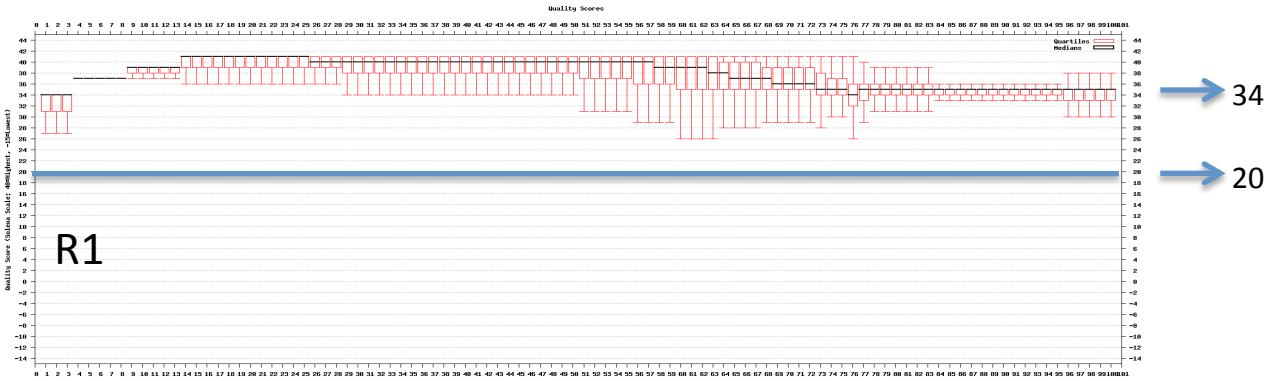
Illumina Sequencing Technology Workflow

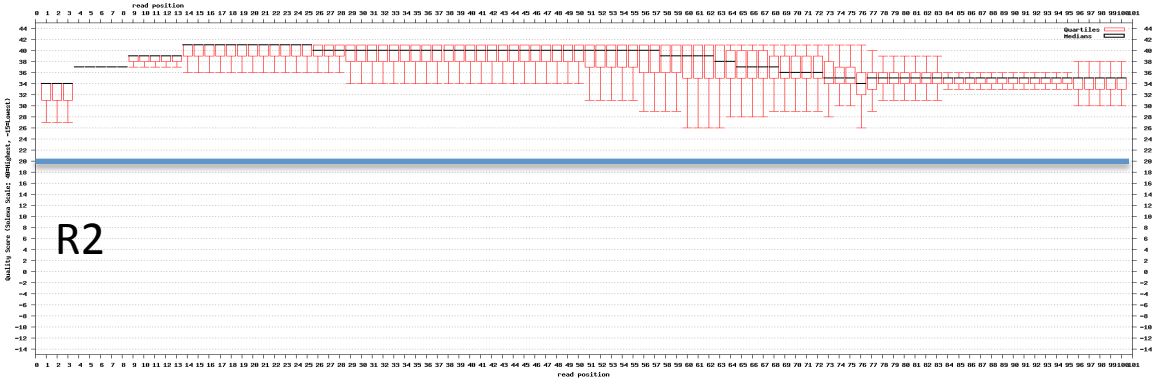


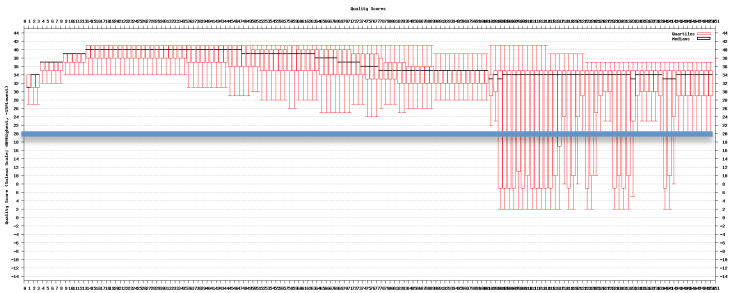
"Phred" quality (Q) scores

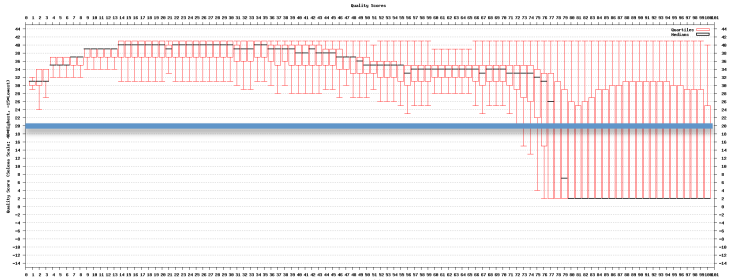
- Each base call is associated with a quality score (Q)
- $Q = -10 \times log_{10}(P)$, where P is the probability that a base call is erroneous
  - A Q score of 20 => 1:100 chance that the base is called incorrectly
  - A Q score of 30 => 1:1000 chance …
- It is generally believed that the Illumina Q scores are accurate

"Phred" quality (Q) scores









Variant Calling – Depth of coverage

For WGS

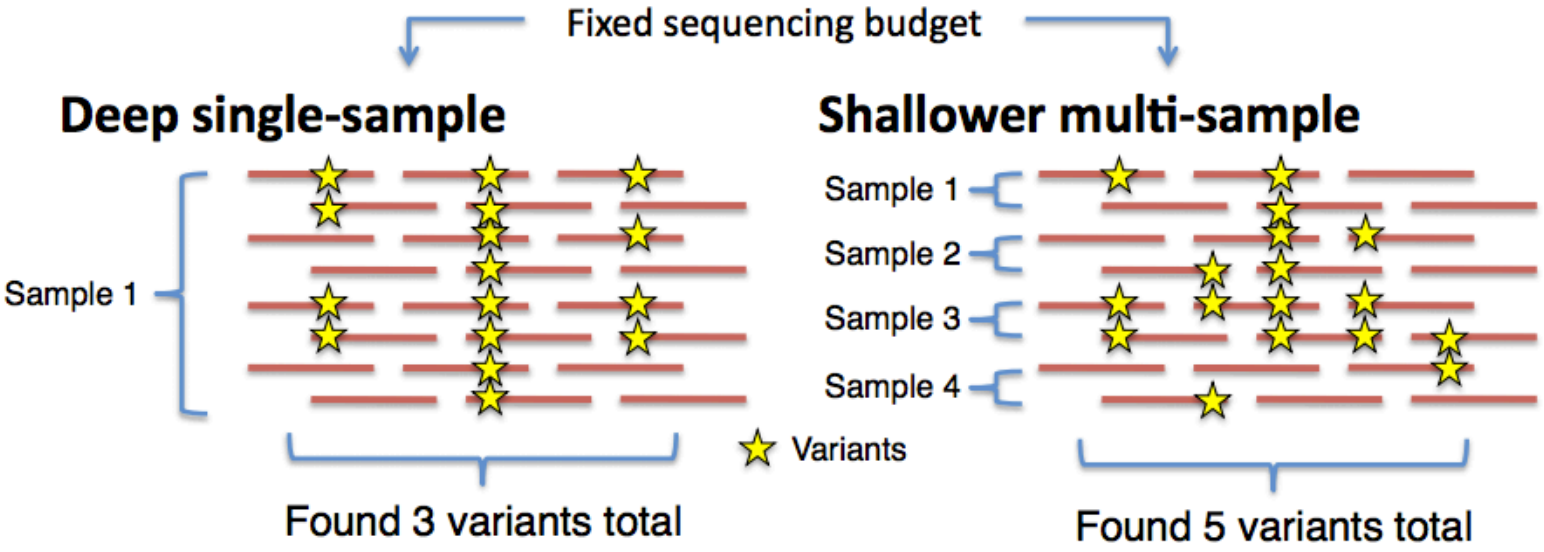
- Haploid genome size => 3.2 Giga base pairs (3.2 billion)
- 50x coverage => ~160 Gbp
- Assuming 100 nucleotide Paired-End reads this is equivalent to 800 million paired reads
- **~5 lanes of Illumina Hi-Seq per sample for WGS**

Variant Calling – Depth of coverage

For Exome Sequencing

- Exome size => 33 Mega base pairs (33 million bases)
- 50x coverage => ~1.65 Gbp
- Assuming 100 nucleotide Paired-End reads this is equivalent to 80 million paired reads
- **<1 lane of Illumina Hi-Seq per sample for exome sequencing**

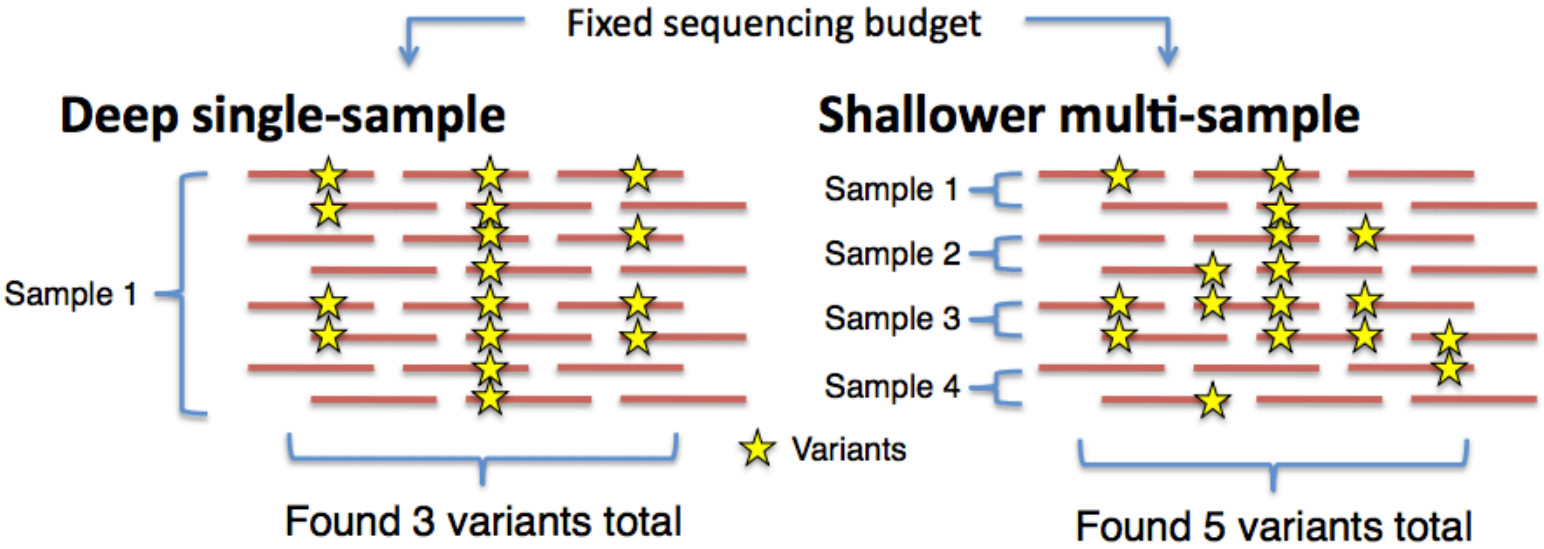
Variant Calling – Depth of coverage



- Higher sensitivity for variants in the sample
- More accurate genotyping per sample
- Cost: no information about other samples

- Sensitivity dependent on frequency of variation
- Worse genotyping
- More total variants discovered

Variant Calling – Depth of coverage



- Higher sensitivity for variants in the sample
- Sensitivity dependent on frequency of variation

Sensitivity is the measure of true variants being identified

other samples

discovered

File Formats

(NGS)

Formats associated with Variant Detection

Input: FASTQ - Raw sequence (potentially billions of small strings)

Output: VCF - A human 'diff' file

Intermediary files:

- **FASTA**
- **SAM/BAM**
- **Optional ones, depending on your needs:**
  - **Known variants (VCF)**
  - **Pedigree information (PED)**
  - **Genotyping information (SnpEff database)**

Formats: FASTA

```
>unique_sequence_ID My sequence is pretty cool
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAAGAATAC
```

Deceptively simple format (e.g. there is no standard)

However in general:

- Header line, starts with '>',
- followed **directly** by an ID,
- … and an optional description (separated by a space)

Files can be fairly large (genomes)

Formats: FASTA

E.g. a read

```
>unique_sequence_ID My sequence is pretty cool
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAAGAATAC
```

E.g. a chromosome

```
>Group10 gi|323388978|ref|NC_007079.3| Amel_4.5, whole genome shotgun sequence
TAATTTATATATCTATTTTTTTTATTAAAAAATTTATATTTTTGTTAAAATTTTATTTGATTAGAAATAT
TTTTACTATTGTTCATTAATCGTTAATTAAAGATAGCACAGCACATGTAAGAATTCTAGGTCATGCGAAA
TTAAAAATTAAAAATATTCATATTTCTATAATAATTAAATTATTGTTTTAATTTAAGTAAAAAAATTTCT
AAGAAATCAAAAATTTGTTGTAATATTGAAACAAAATTTTGTTGTCTGCTTTTTATAGTAACTAATAAAT
ATTTAATAAAAAATTACTTTATTTAATATTTTATAATAAATCAAATTGTCCAATTTGAAATTTATTTTAT
CACTAAAAATATCTTTATTATAGTCAATATTTTTTGTTAGGTTTAAATAATTGTTAAAATTAGAAAATGA
TCGATATTTTCAAATAGTACGTTTAACTAATACTTAAGTGAAAGGTAAAGCGGTTATTTAAAATATTGAT
TTATAATATTCGTGACATAATATATTTATAAATAGATTATATATATATATATACATCAAAATATTATACG
AGAACTAGAAAATATTACAGATGCAAAATAAATTAAATTTTGTAAATGTTACAGAATTAAAAATCGAAGT
```

Formats: FASTQ

```
@unique_sequence_ID
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAAGAATAC
+
=-(DD--DDD/DD5:*1B3&)-B6+8@+1(DDB:DD07/DB&3((+:?=8*D+DDD+B)*)B.8CDBDD4DDD@@D
```

May be ‘raw’ data (straight from seq pipeline) or processed (trimmed for various reasons)

Can hold 100’s of millions of records **per sample**

Files can be very large (100’s of GB) apiece

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
12345678910……40
```

Formats: SAM/BAM

SAM – Sequence Alignment/Map format

- SAM file format stores alignment information
- Normally converted into BAM (text format is mostly useless for analysis)

Specification: http://samtools.sourceforge.net/SAM1.pdf

Contains FASTQ reads, quality information, alignment information, other information about samples (meta data) etc.

Files are typically very large: Many 100's of GB or more

Formats: SAM/BAM

BAM – BGZF compressed SAM format

- May be unsorted, or sorted by sequence name or genome coordinates
- May be accompanied by an index file (.bai)
- Makes the alignment information easily accessible to downstream applications
- Relatively simple format makes it easy to extract specific features, e.g. genomic locations
- BAM is the compressed/binary version of SAM and is not human readable. Uses a specialized compression algorithm optimized for indexing and record retrieval (bgzip)

Files are typically very large: 1/5 of SAM, but still very large

Formats: VCF/BCF

VCF (Variant Call Format)

BCF – direct bgzip-compressed VCF format

Specification:

From the 1000 Genomes Project

http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41

Formats: VCF

```
##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=<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=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##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        NA
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 1/
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   0/
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   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 0/
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       1/
```

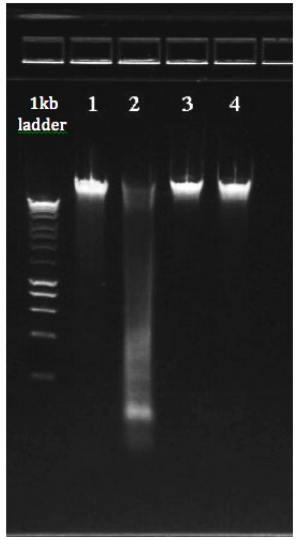
QC steps to consider for NGS-based Variant Calling

QC steps

- During and after library prep
  - Is the quality and amount of genomic DNA reasonable?
  - Is the quality and amount of prepared library good?
- During sequencing and immediately after
  - Are there too many or too few clusters?
  - Is the sequencing proceeding as expected?
- Before data processing
  - Is the data quality good?
  - If not, can getting rid of low quality reads or bases help?

QC - During and after library prep

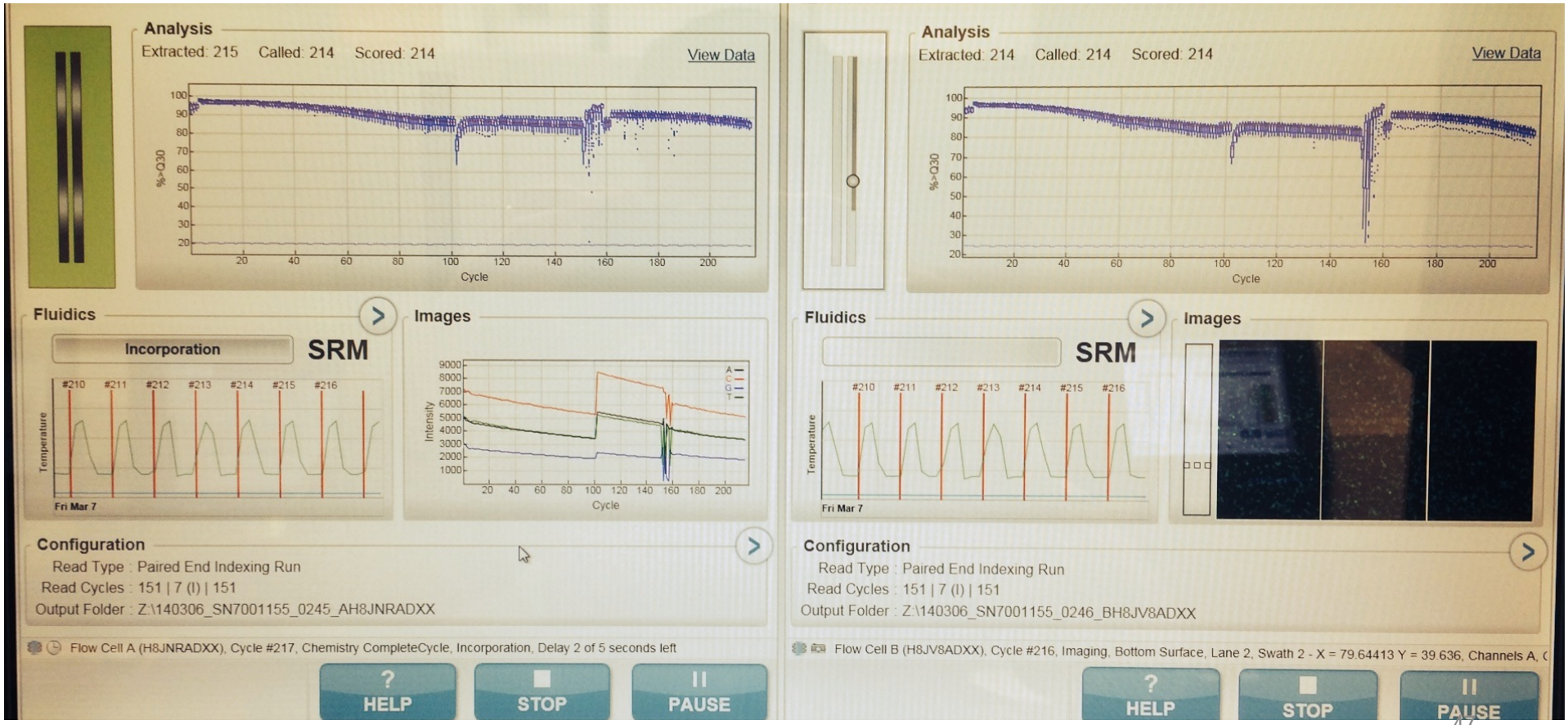
- Is the quality and amount of genomic DNA reasonable?
  - A 1% agarose gel can be run to check quality
  - For estimating DNA amount, a nanodrop (spectrophotometric method) can often be inaccurate due to various reasons and the recommendation is to use the Qubit (fluorometric method)
- Is the quality and amount of prepared library good?
  - Perform a Bioanalyzer run to double check the size
  - Perform a Qubit DNA assay estimate the quantity



QC - During sequencing and immediately after

- Are there too many or too few clusters?
  - Perform one cycle of sequencing to test if all 8 lanes of the flow cell have a good number of clusters (“Goldilocks” effect)
  - This will impact final data quality!
- Is the sequencing proceeding as expected?
  - Monitor the stats on the monitor of the machine every few hours to ensure there are no issues with the runs

QC - During sequencing and immediately after

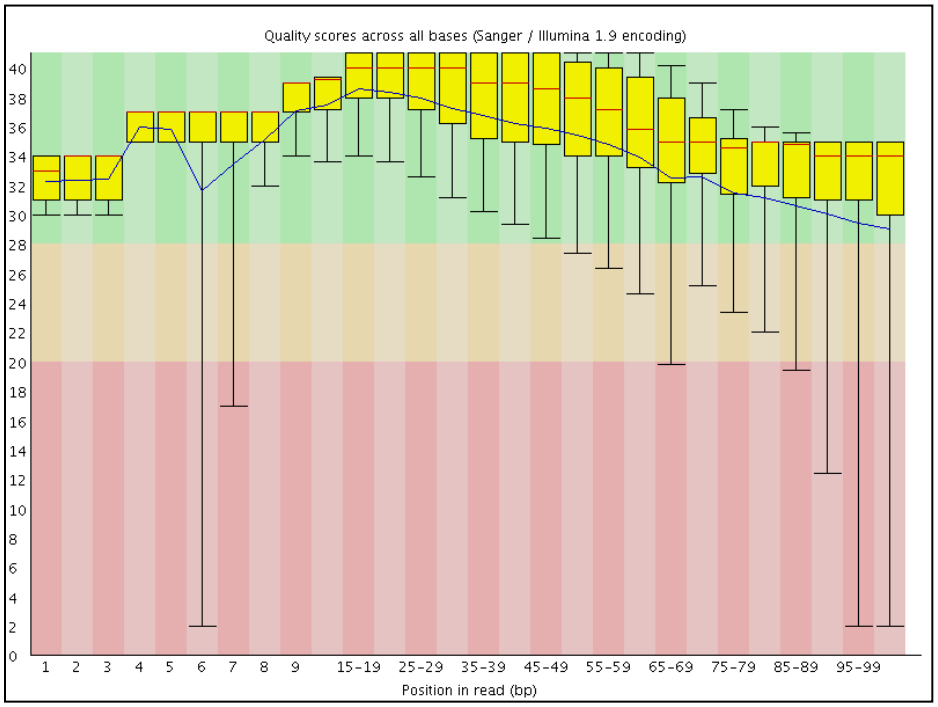


QC - Before data processing

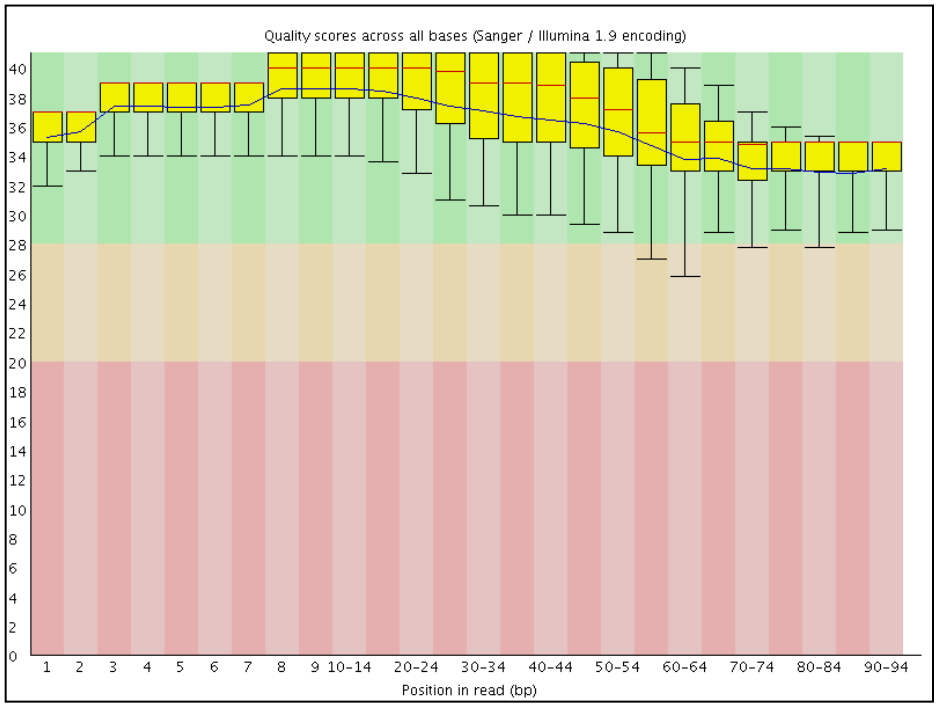
- FastQC to check quality scores and other metrics of the FASTQ data file
- Trimmomatic to remove low quality bases from either end and choose to keep only reads with enough nucleotides remaining
- Trimmomatic to remove any leftover adaptor sequences
- FastQC to check the metrics after trimming

QC - Before data processing

Before quality trimming



After quality trimming



Variant Calling Data Processing Steps

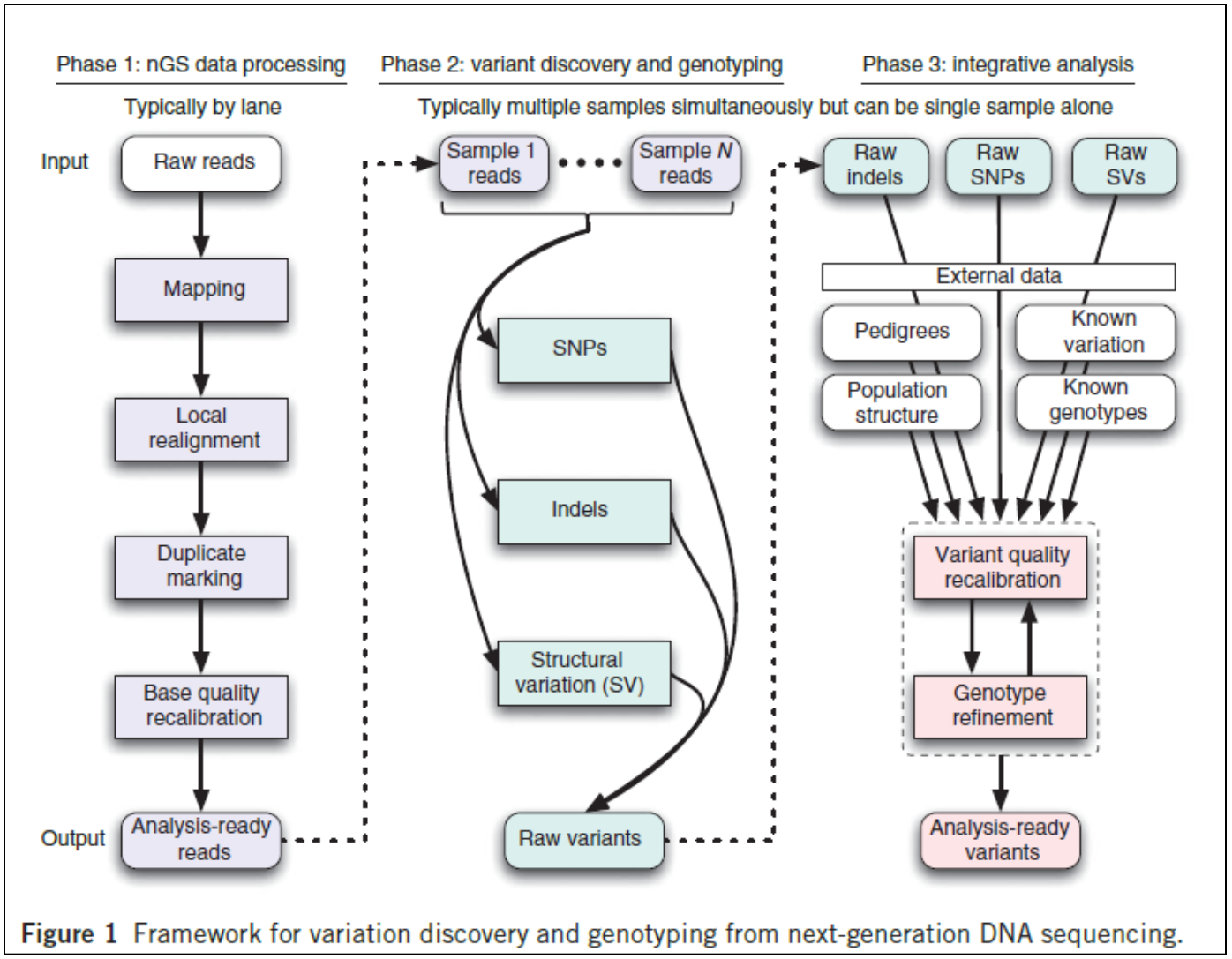


Figure 1 Framework for variation discovery and genotyping from next-generation DNA sequencing.

De Pristo *et al.* Nature Genetics May 2011

Calling variants with the GATK

- GATK

*"**The Genome Analysis Toolkit** or GATK is a software package developed at the Broad Institute to analyse next-generation resequencing data. 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."*

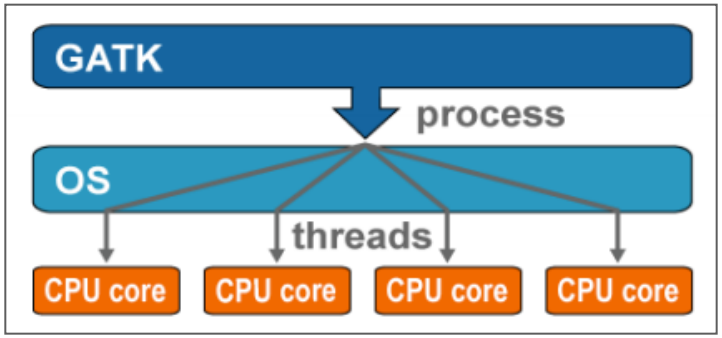
- http://www.broadinstitute.org/gatk/

Calling variants with the GATK

- GATK provides a good infrastructure and a guide to best practices to be employed for variant calling
- It utilizes several open-source tools at various steps, along with GATK-specific tools and scripts
- Can use both exome and WGS data for variant calling
- GATK takes advantage of the concept of parallel computing to speed up the pipeline

Calling variants with the GATK

Parallelism

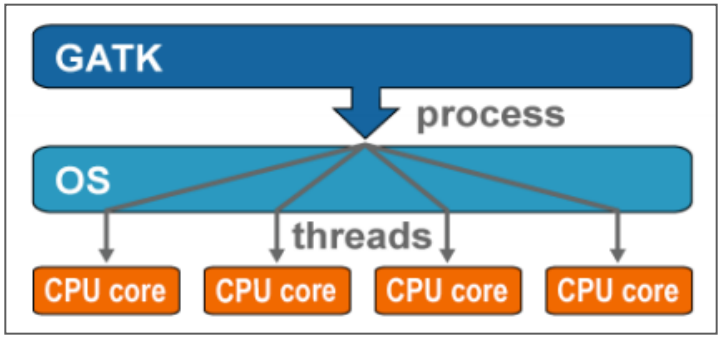


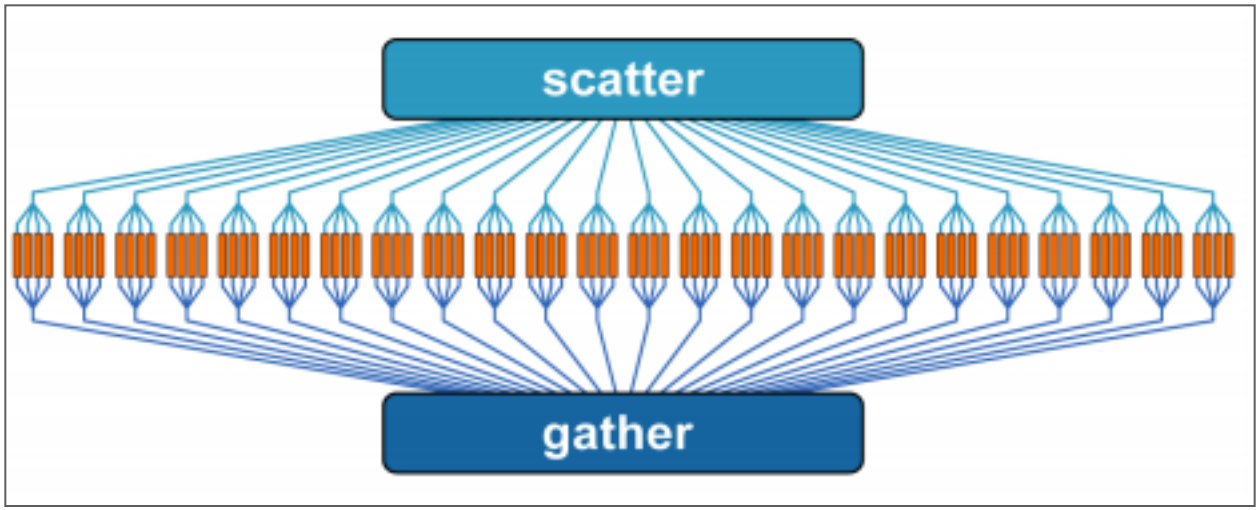
Example of a Node on a cluster (UNIX)

- 1 Dell PowerEdge R620 Node
- 24 Intel Xeon E5-2697 @ 2.7GHz CPU Cores
- 384 Giga Bytes of RAM

Calling variants with the GATK

Parallelism

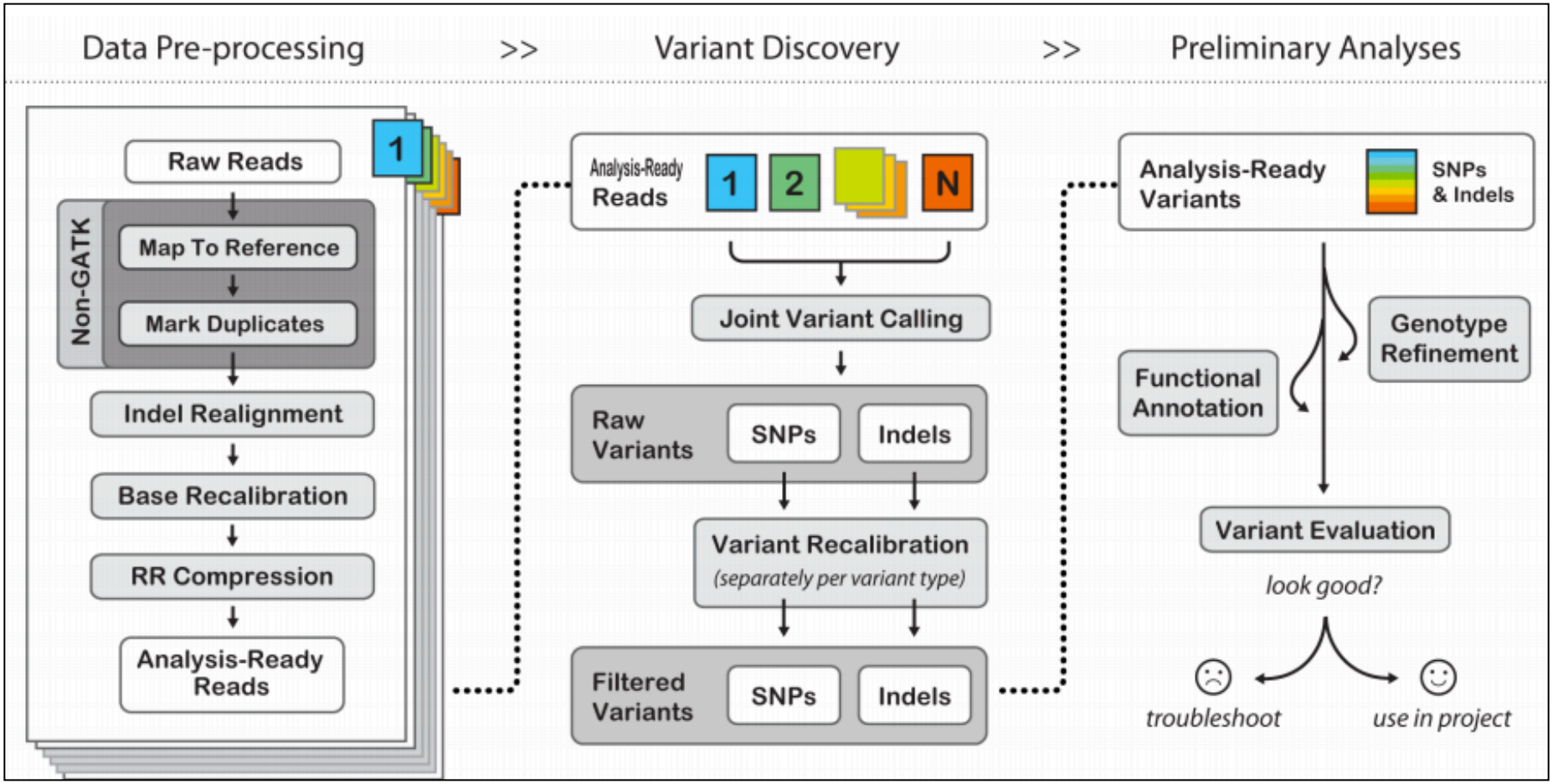




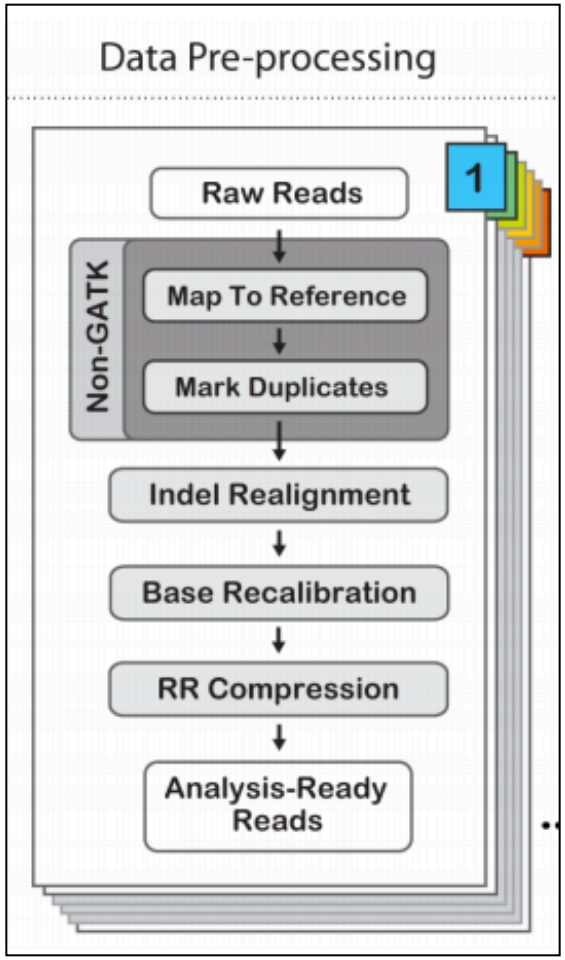
Calling variants with the GATK

- GATK provides a good infrastructure and a guide to best practices to be employed for variant calling
- It utilizes several open-source tools at various steps, along with GATK-specific tools and scripts
- Can use both exome and WGS data for variant calling
- GATK takes advantage of the concept of parallel computing to speed up the pipeline
  - You can implement this type of a set up outside of the GATK constraints

Calling variants with the GATK

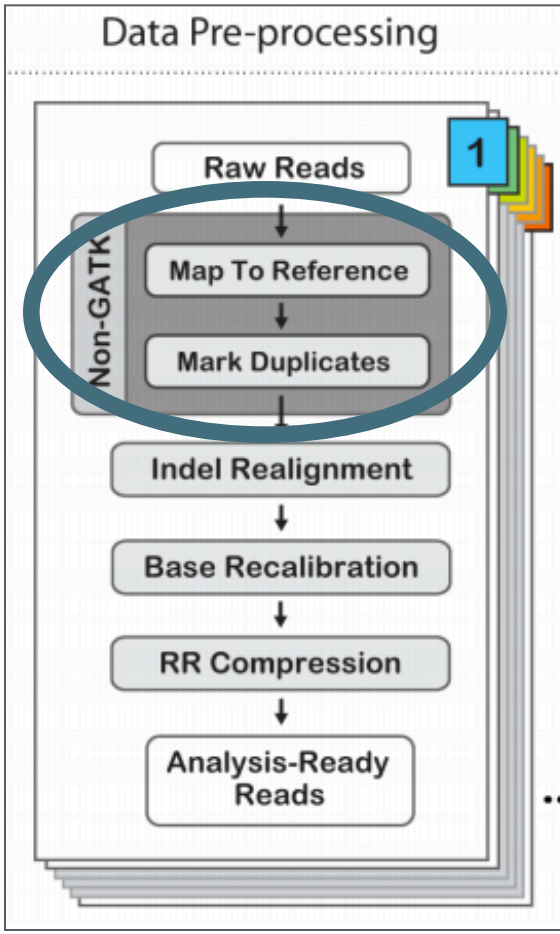


Calling variants with the GATK



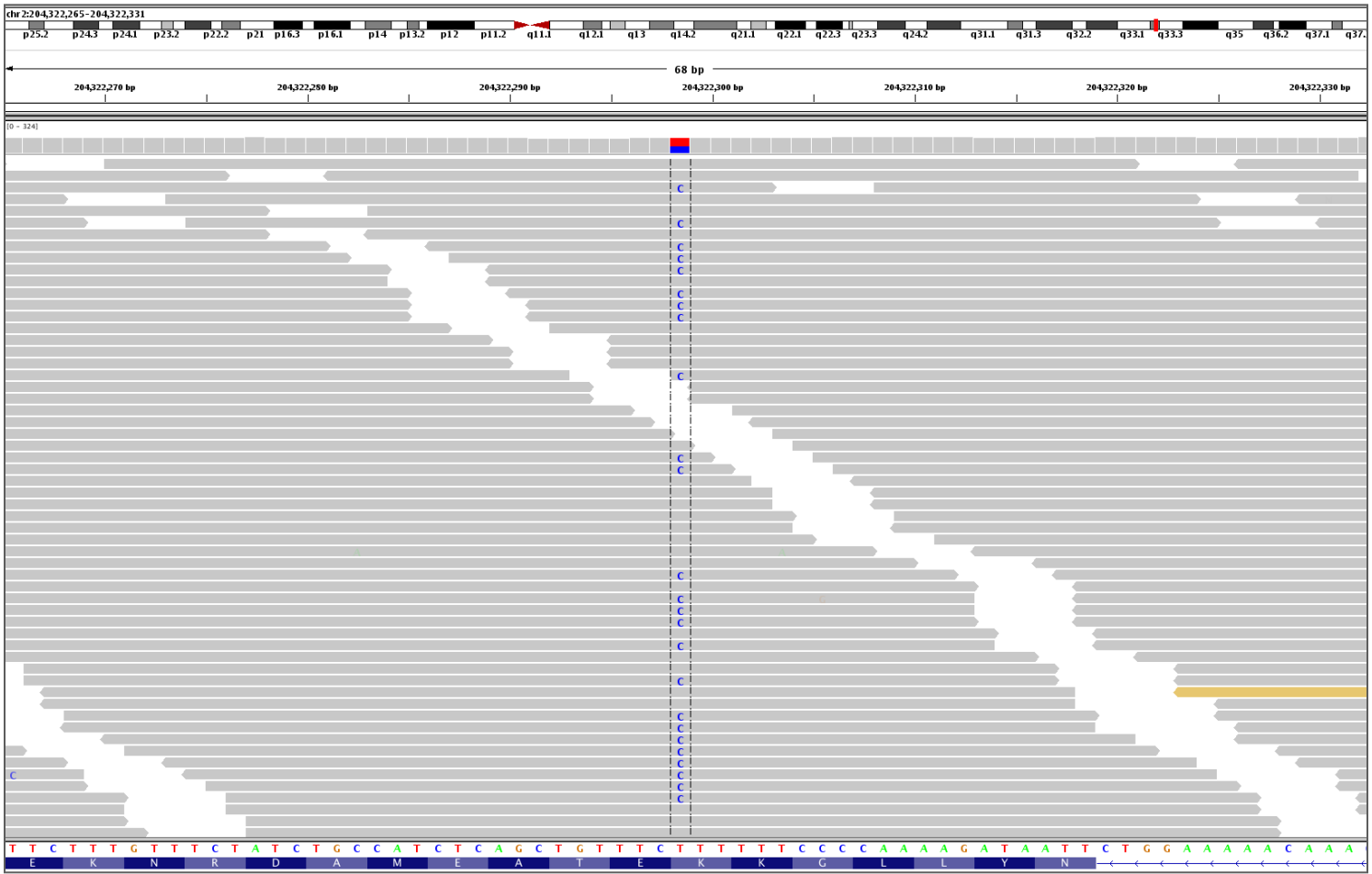
Calling variants with the GATK

- Mapping or Aligning raw reads to reference genome is usually done with **BWA** (Burrows-Wheeler Aligner)
- Duplicates are marked using **Picard**
- Very important steps that set up the quality of the variant calling
- Tools used for these steps are external to GATK



Calling variants with the GATK

Mapping



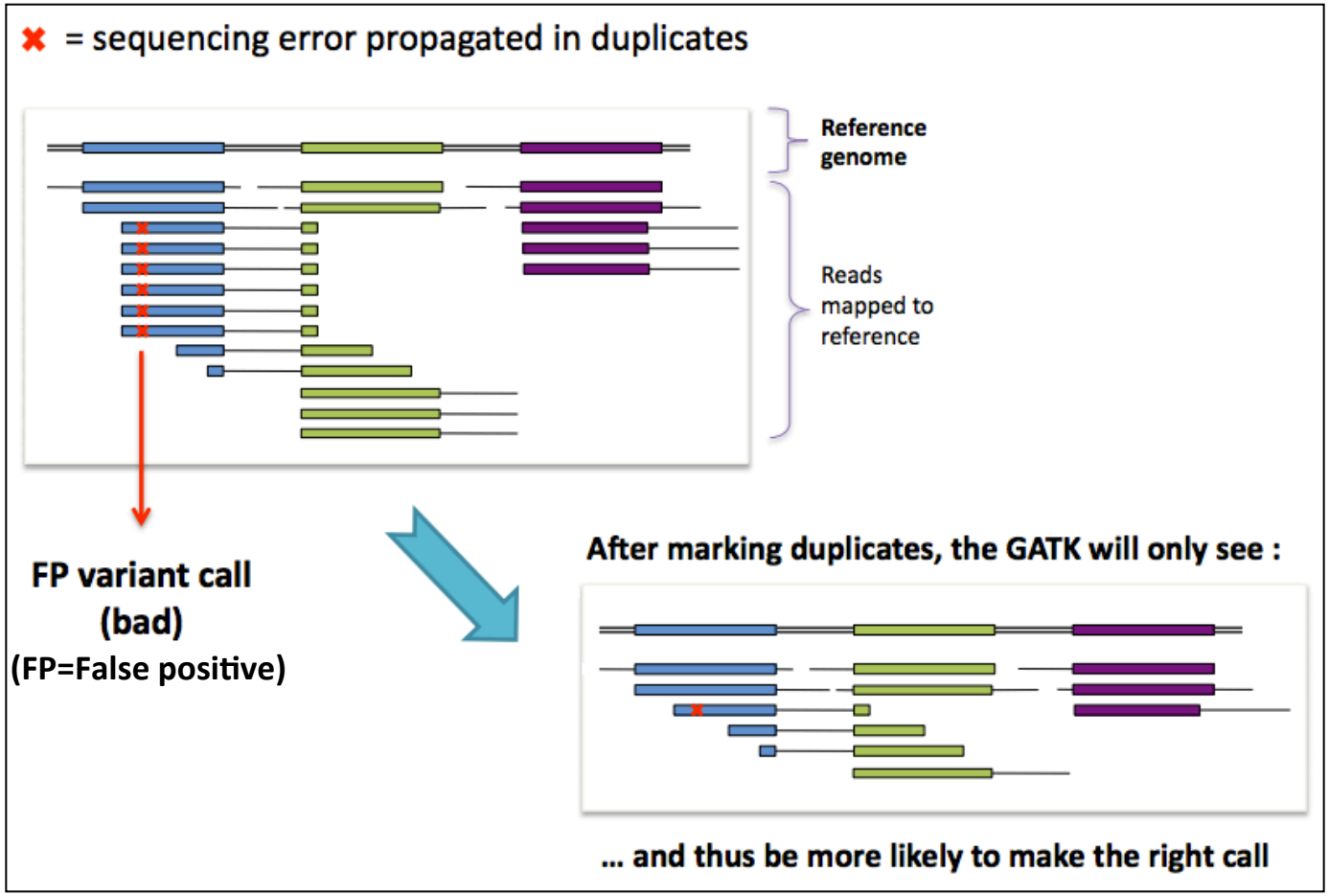
Calling variants with the GATK

Mapping

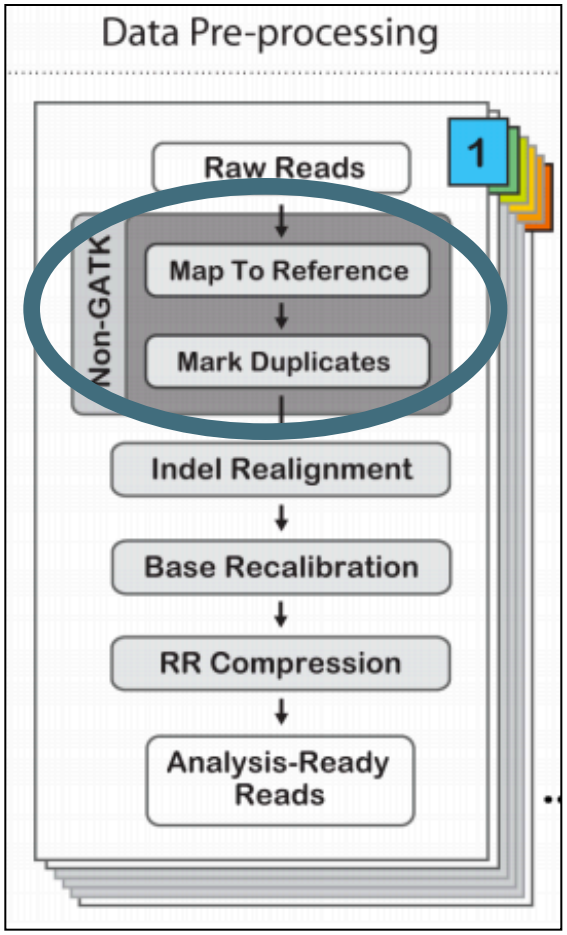
- Theoretically this is a simple step to determine where the read matches the reference genome
- But, there are several issues to be considered in practice
  - Mismatches due to a variant or a sequencing error
  - A read mapping to more than one location (repeats)
  - Mapping Quality of the read depends on these factors
- The FASTQ files are often “chunked” into smaller files for this step, and remerged after alignment

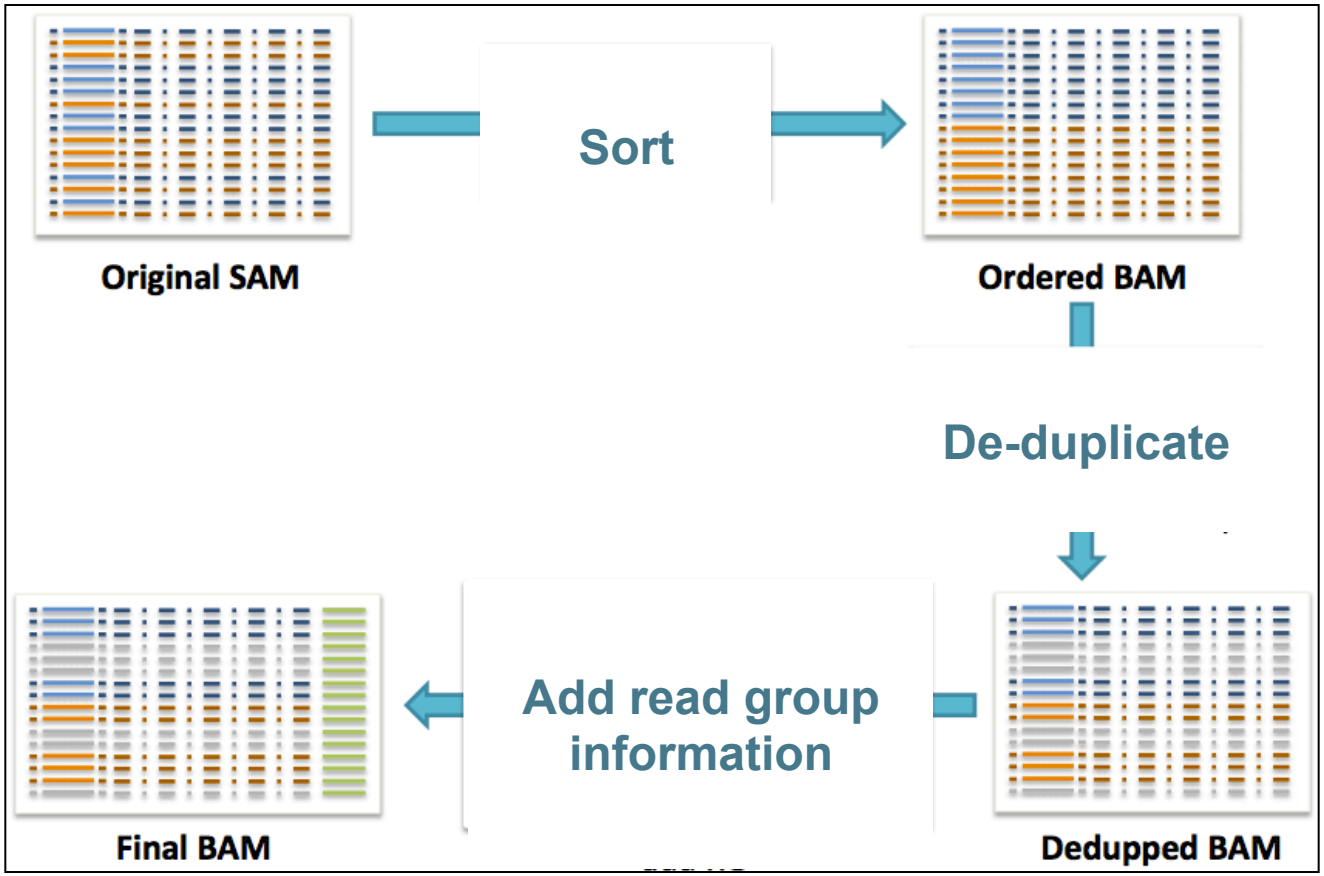
Calling variants with the GATK

Marking duplicates (de-duplicating)



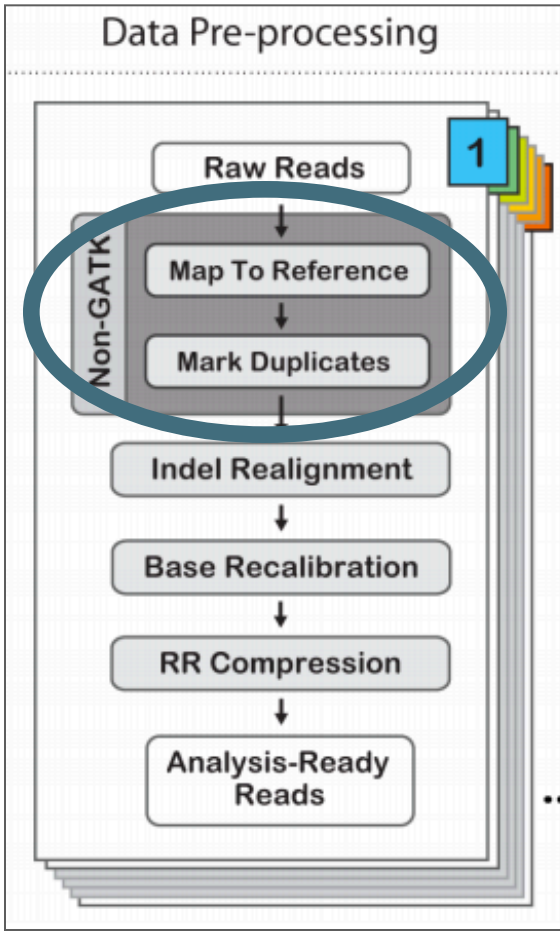
Calling variants with the GATK



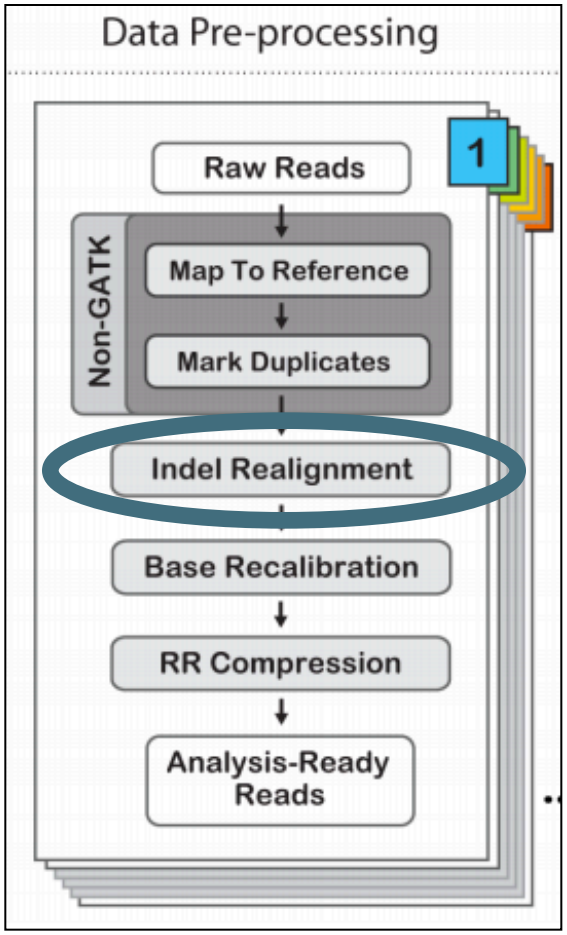


Calling variants with the GATK

- These steps are computationally expensive, and data are usually split into smaller “chunks” prior to mapping and marking duplicates.
- If multiple samples are being processed, these steps are performed separately for each sample
- These steps set up the stage for good quality calls
  - All later steps assume that reads are placed in the right location and represent that region of the genome
  - Duplicates originate mostly from DNA prep methods and cause biases that skew variant calling results



Calling variants with the GATK



Indel Realignment

There are 2 steps to the realignment process:

1. Determining (small) suspicious intervals which are likely in need of realignment
2. Running the re-aligner over those intervals

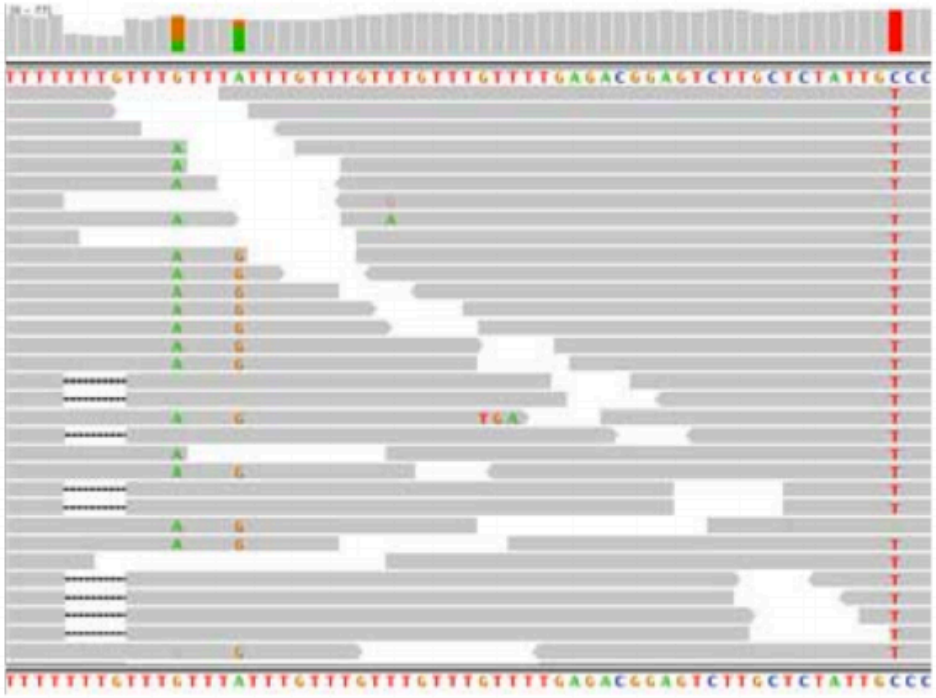
Can use known sites to aid in the realignment

- All samples can be merged together and then separated by chromosome and data for each chromosome can be processed separately from this point on

Calling variants with the GATK

Indel Realignment

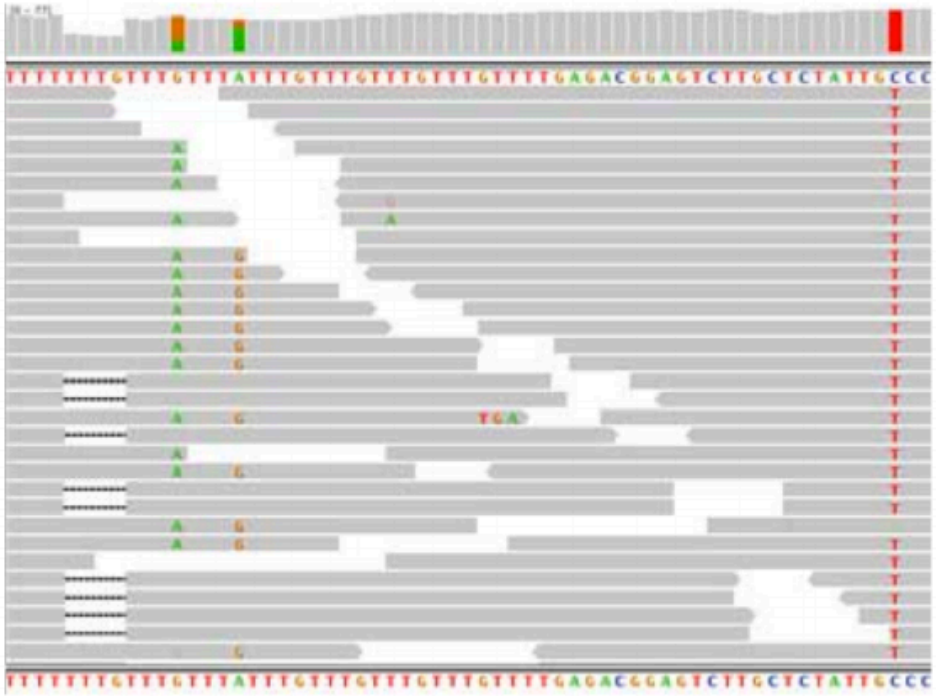
Before realignment



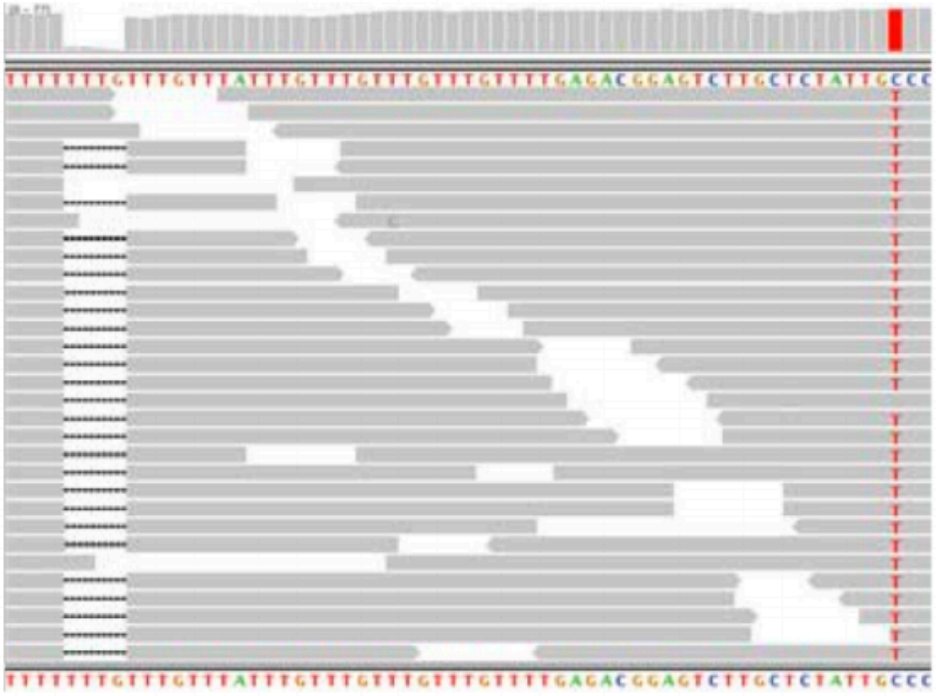
Calling variants with the GATK

Indel Realignment

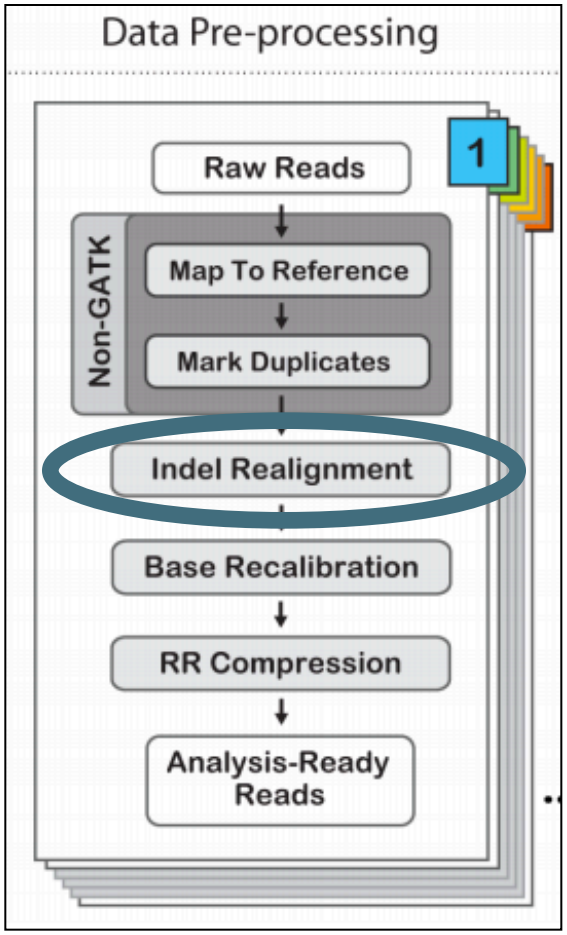
Before realignment



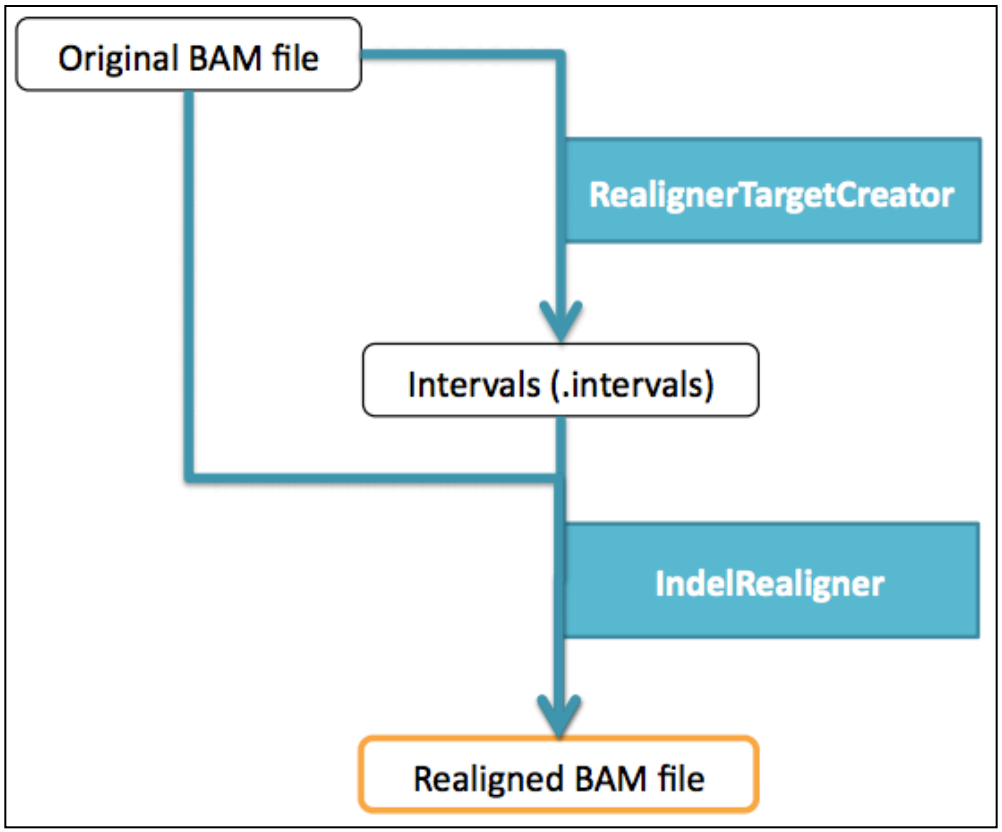
After realignment



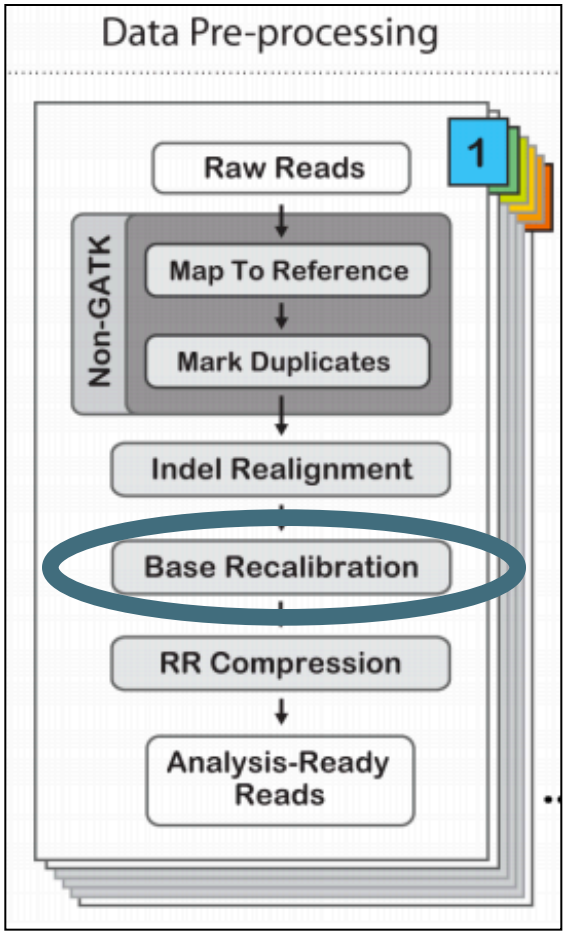
Calling variants with the GATK



Indel Realignment



Calling variants with the GATK

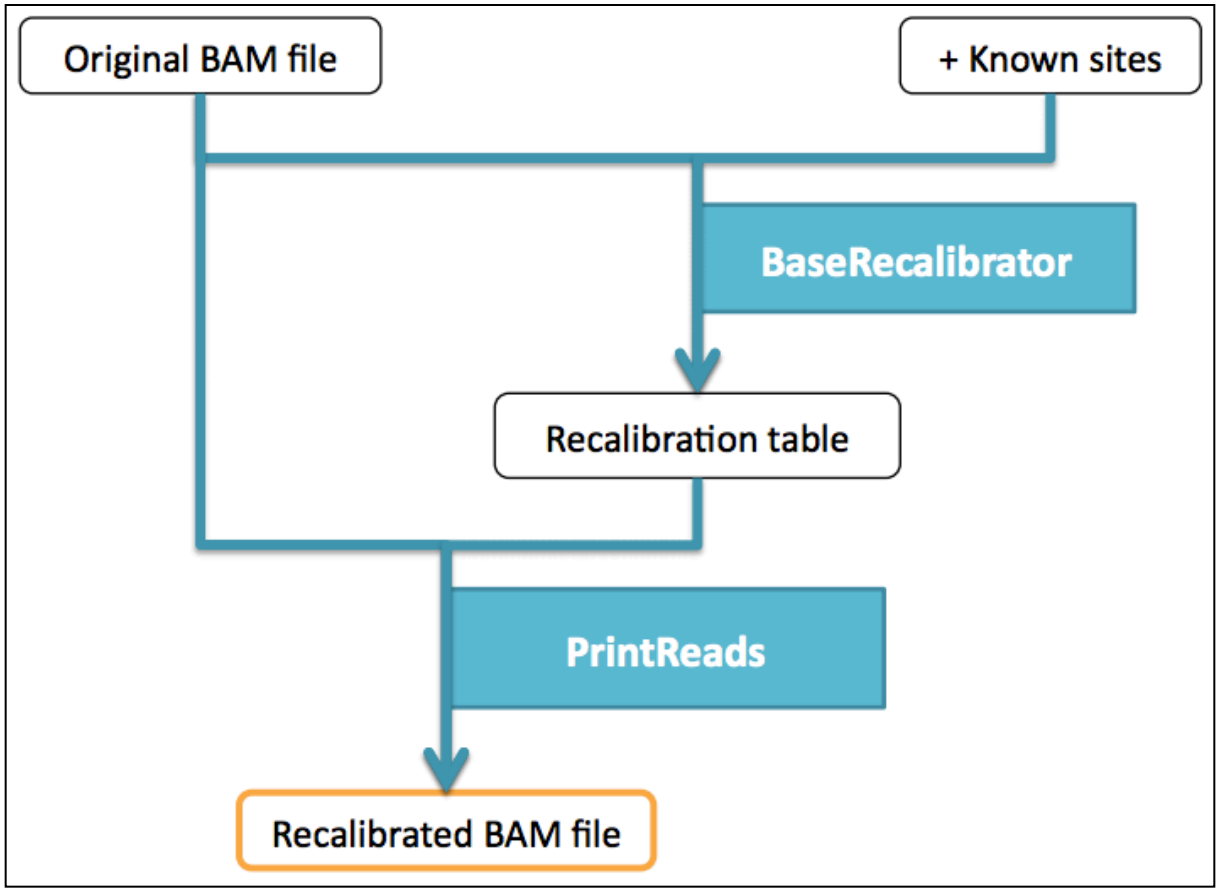


Base Recalibration

- Start with reads aligned to reference genome (.bam file)
- Estimate the likelihood of a biased quality score
- Following criteria are considered for estimating bias:
  - Reported quality score
  - Machine cycle on sequencer
  - Dinucleotide context
  - Down-weighting or remove duplicate clones
- Remove the estimated bias

Calling variants with the GATK

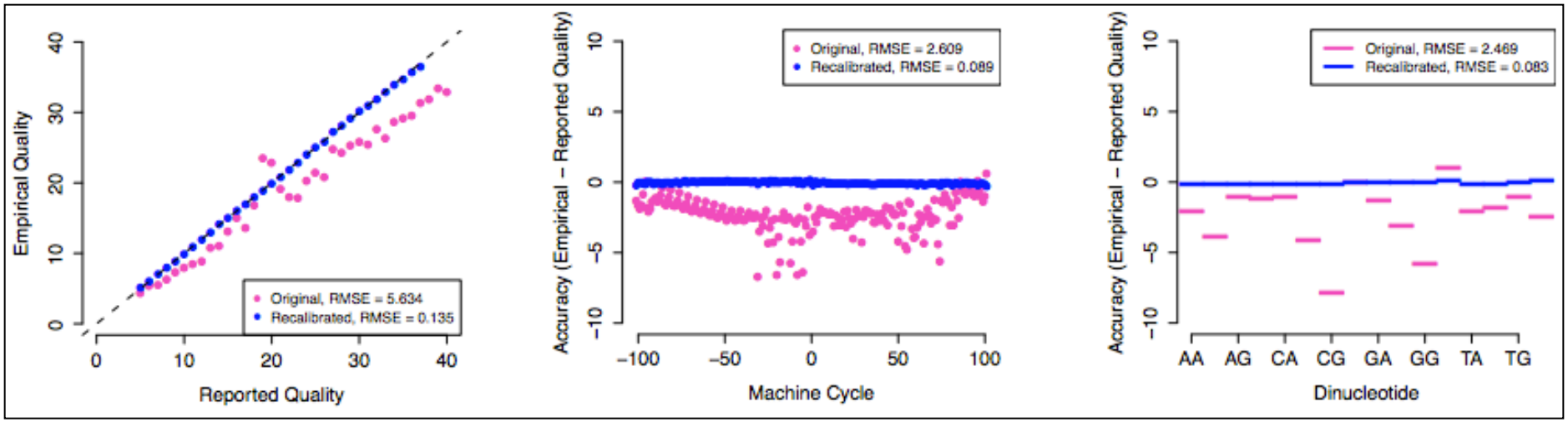
Base Recalibration



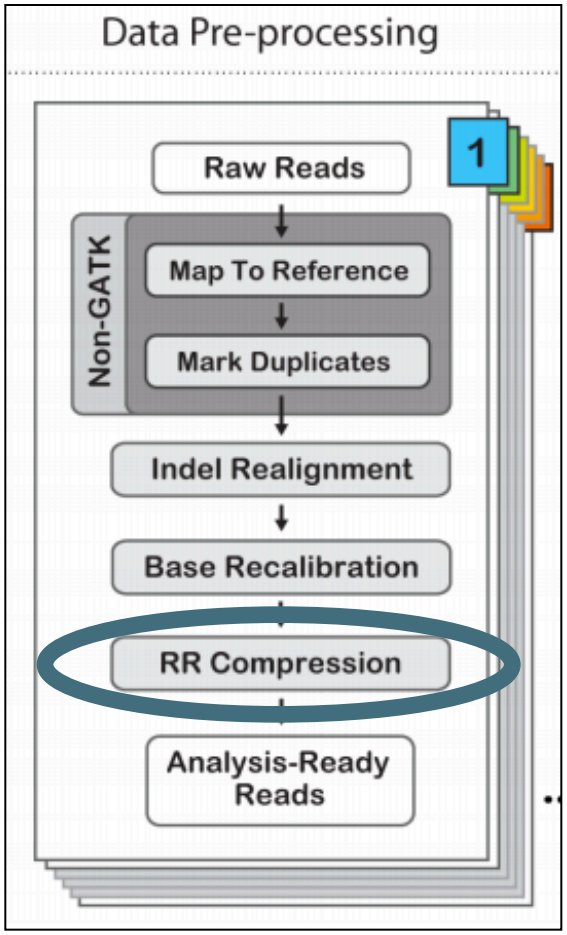
Calling variants with the GATK

Base Recalibration

The step removes and systematic biases the creep in during data generation and the previous data processing steps



Calling variants with the GATK

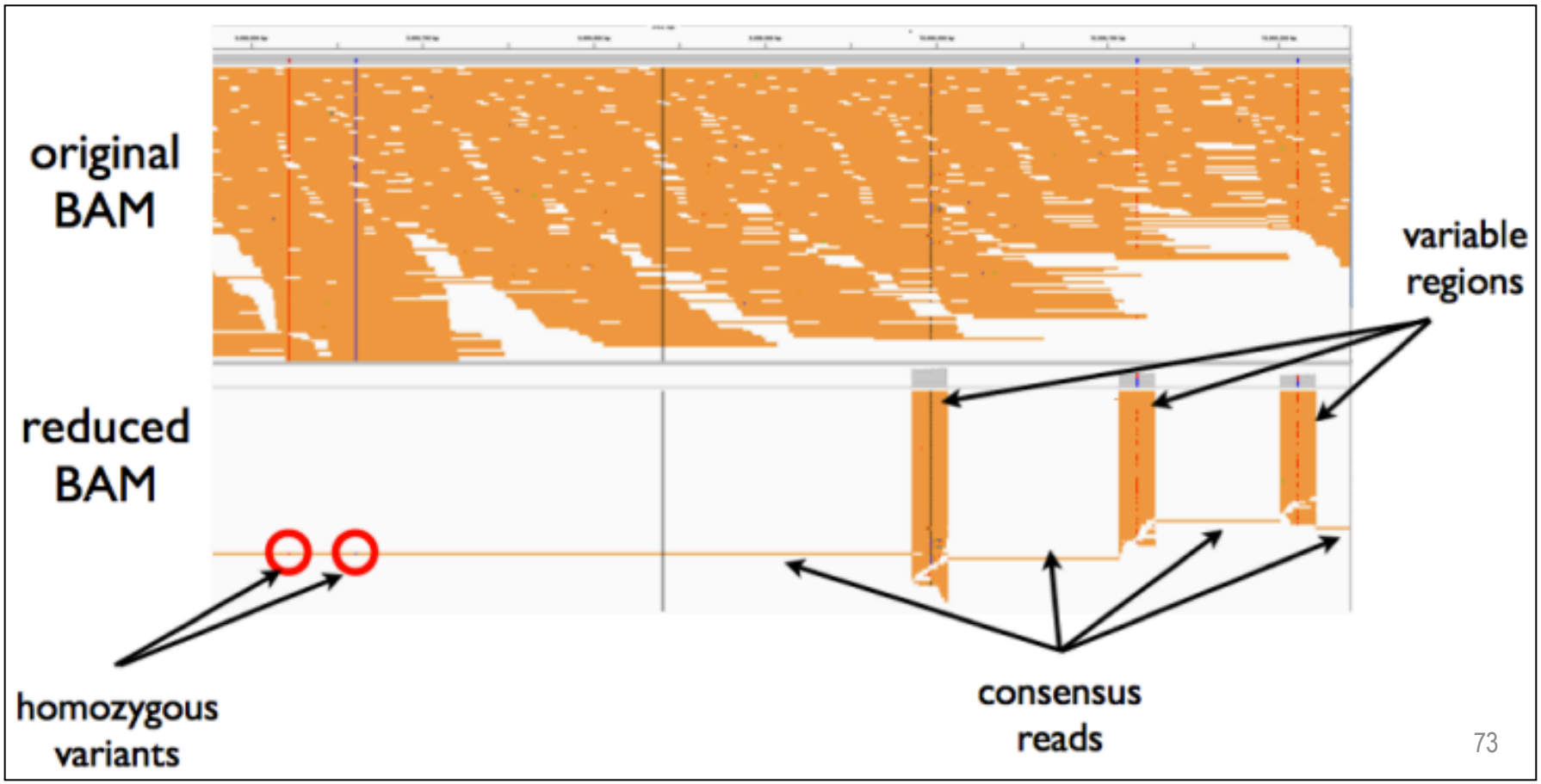


Reduce Reads Compression

- Reduce the size of the BAM file by removing non-essential information
- Distinguish between consensus and variable regions, and remove consensus information
- Down-sample coverage in variable regions
- A set of samples are co-reduced to ensure consistency (and hence, effective comparisons) when working with multiple samples

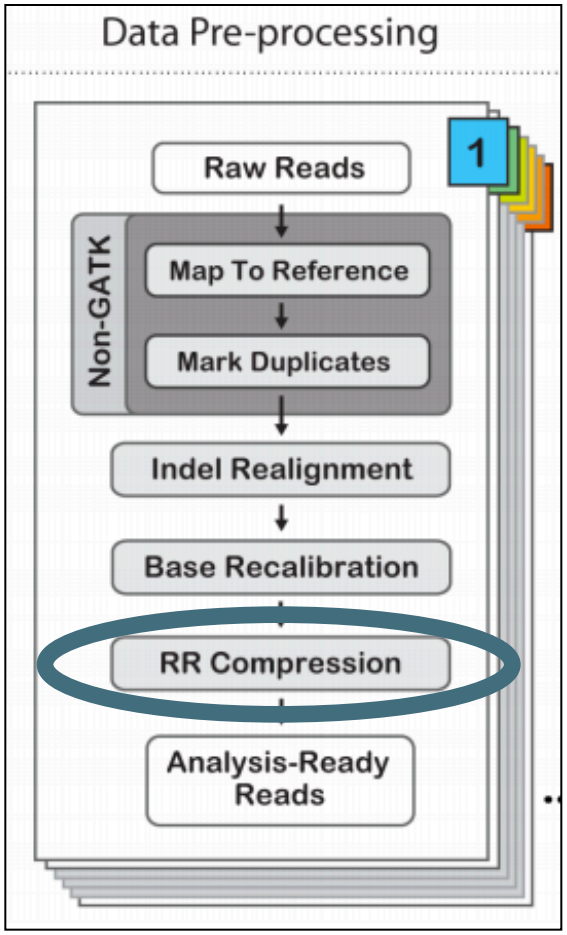
Calling variants with the GATK

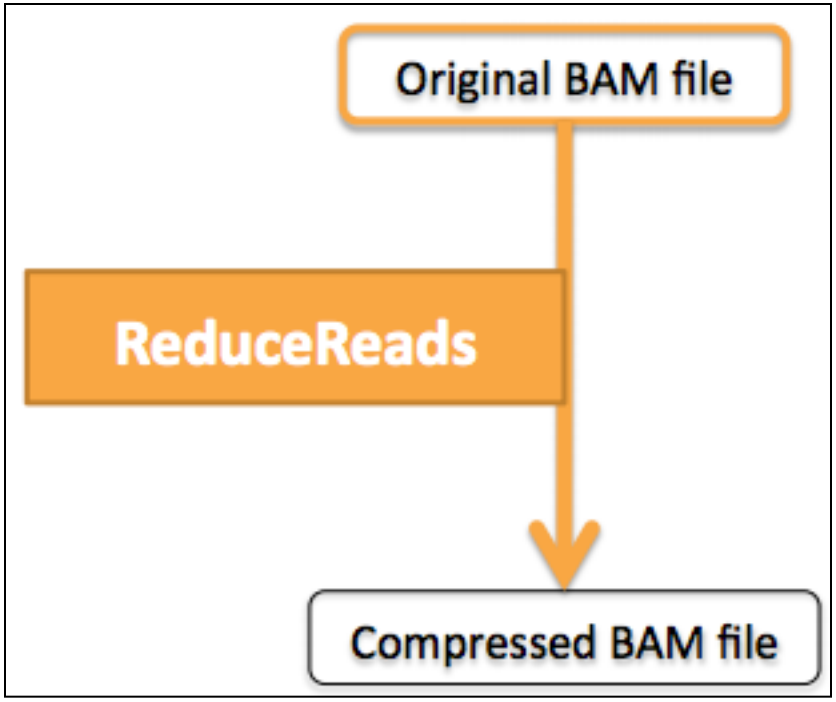
Reduce Reads Compression



Calling variants with the GATK

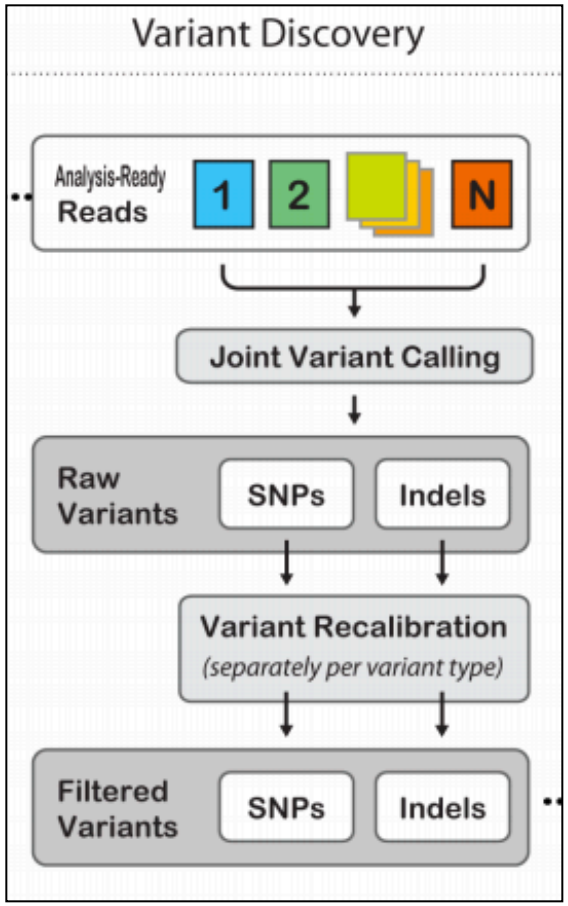
Reduce Reads Compression





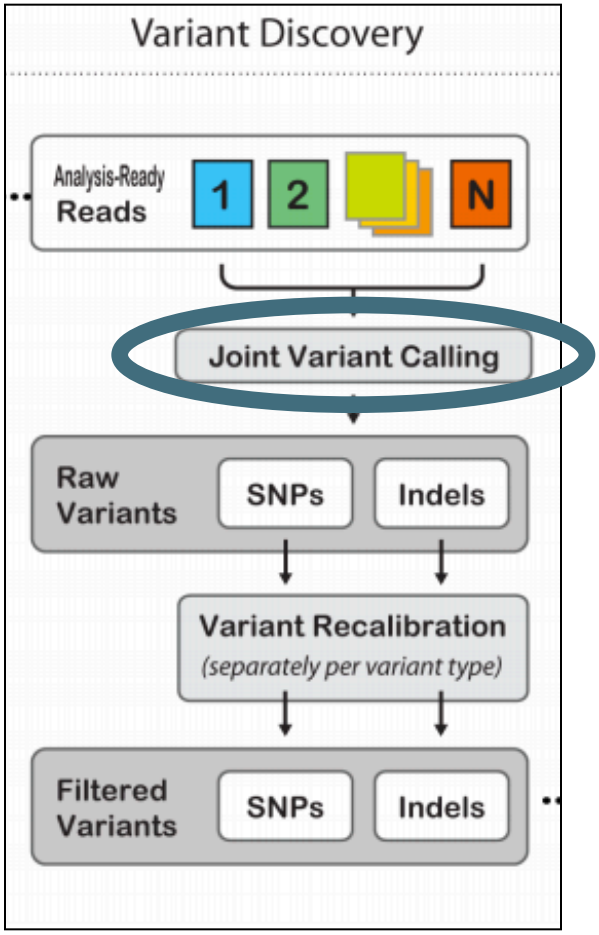
Calling variants with the GATK

Calling Variants (finally)!!



Calling variants with the GATK

Calling Variants



- Discovery of real variants buried in the noise
- Several steps have been taken to reduce the noise with both SNPs and indels
- The BAM files going into this portion of the pipeline are “cleaner” and reduced

- All samples can be merged together and then separated by chromosome and data for each chromosome can be processed separately from this point on

Calling variants with the GATK

Calling Variants

- UnifiedGenotyper

  Call SNPs and indels separately by considering each variant locus independently

  - Accepts any ploidy
  - Pooled calling
  - High sample numbers

- HaplotypeCaller

  Call SNPs, indels, and some SVs simultaneously by performing a local *de-novo* assembly

  - More accurate, especially for indels
  - Will eventually replace UG

Calling variants with the GATK

Unified Genotyper (UG) -

- UG calls SNPs and indels separately by considering each variant locus independently
- Currently, this program runs faster than HaplotypeCaller
- UG is going to be phased out in favor of the HaplotypeCaller

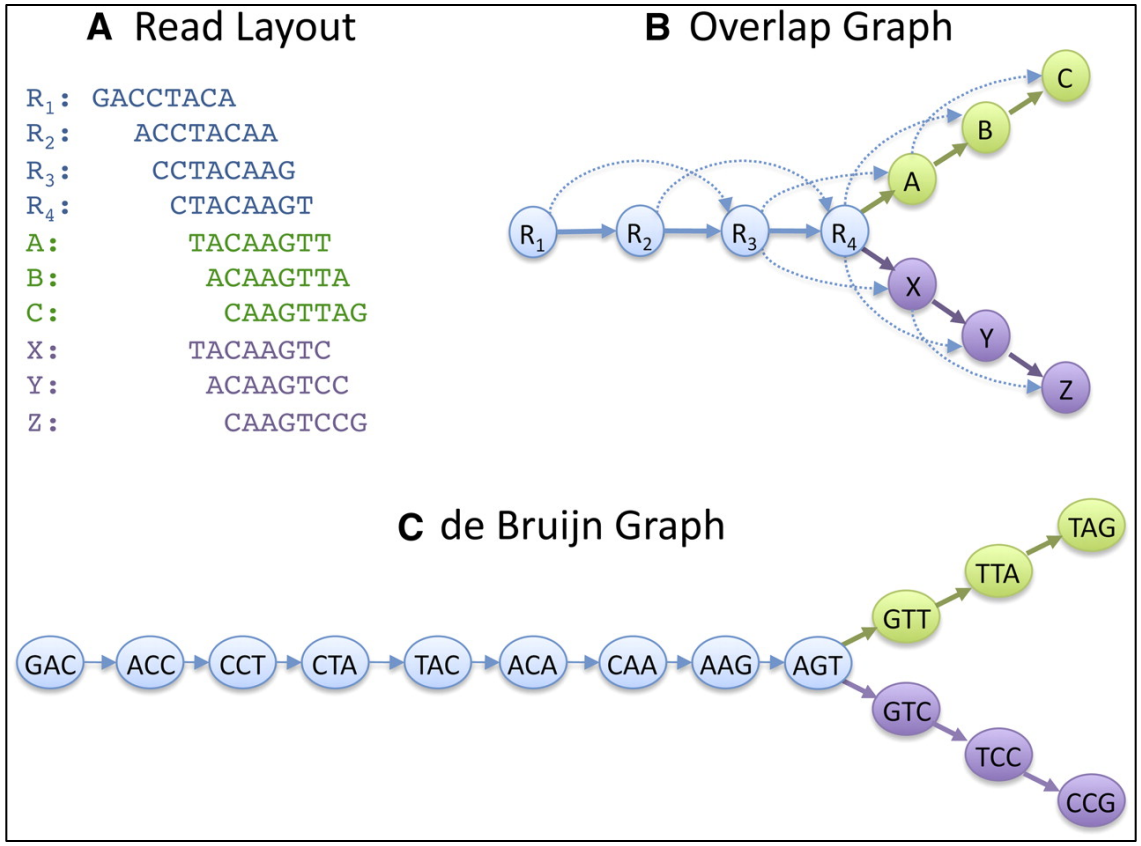
Calling variants with the GATK

HaplotypeCaller -

» Call SNPs, indels, and some SVs simultaneously by performing a local de-novo assembly (deBruijn graph-based assembly)

Calling variants with the GATK

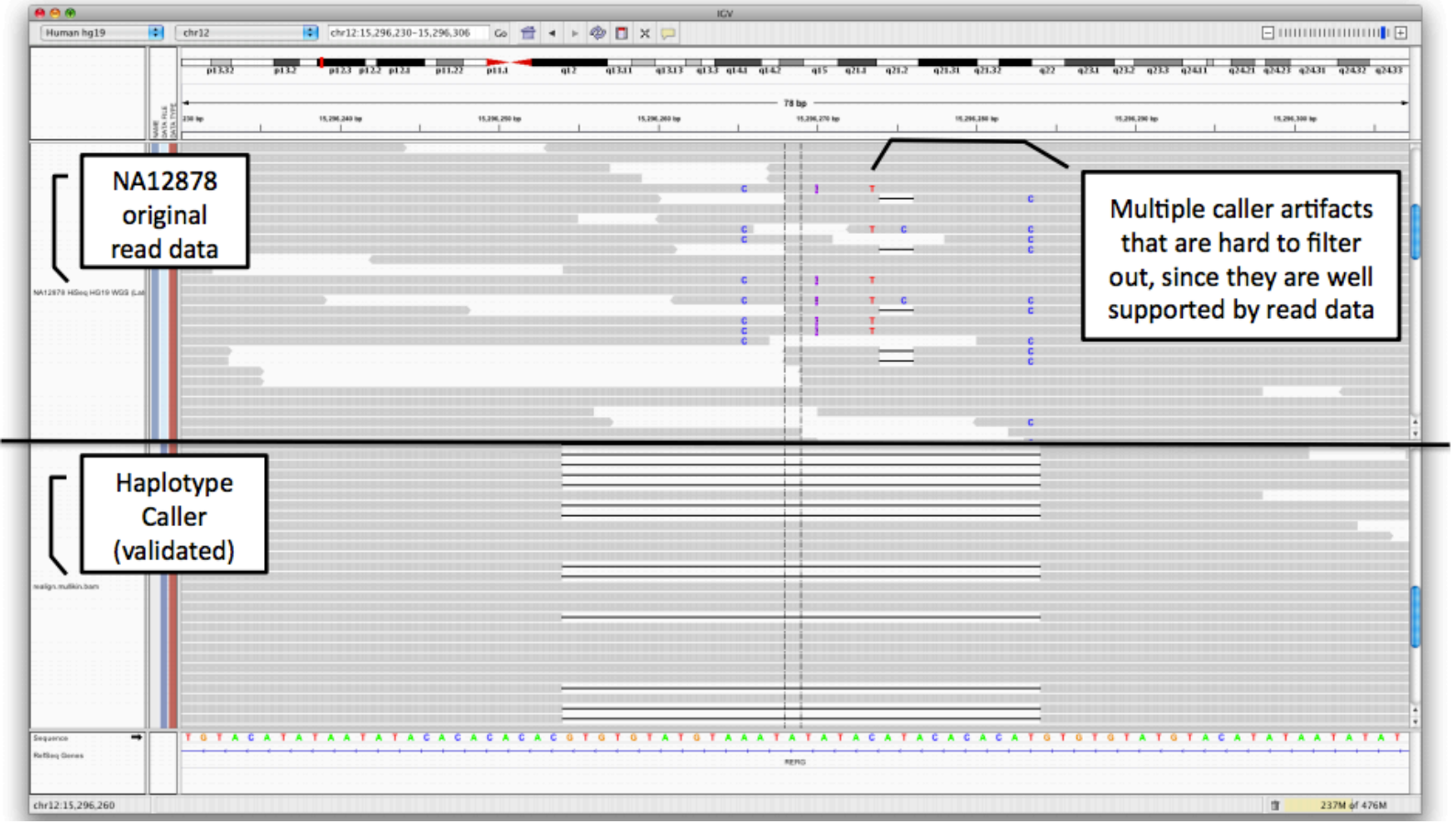
HaplotypeCaller



Schatz M C et al. Genome Res. 2010;20:1165-1173

Calling variants with the GATK

HaplotypeCaller

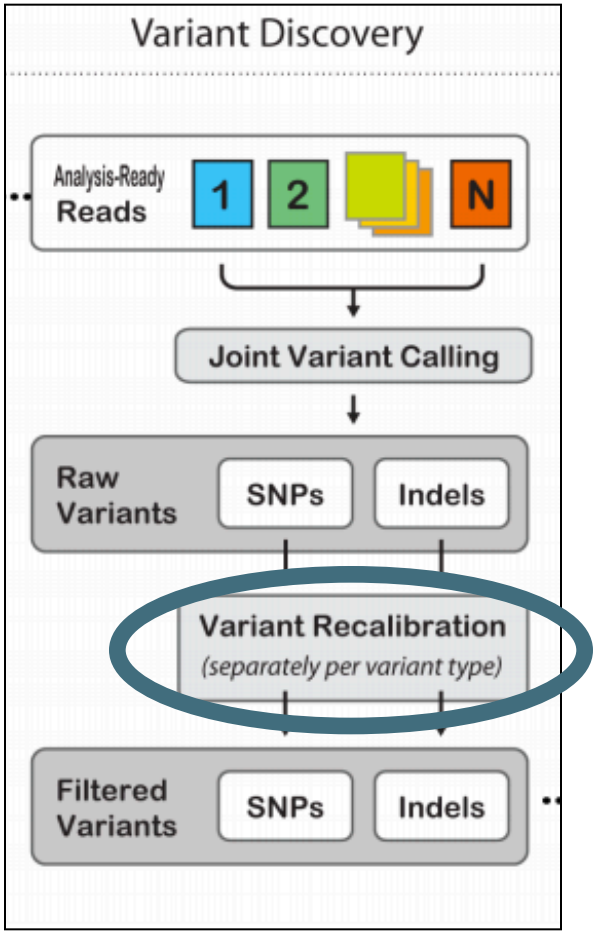


Short read alignment

Assembled data re-alignment

Calling variants with the GATK

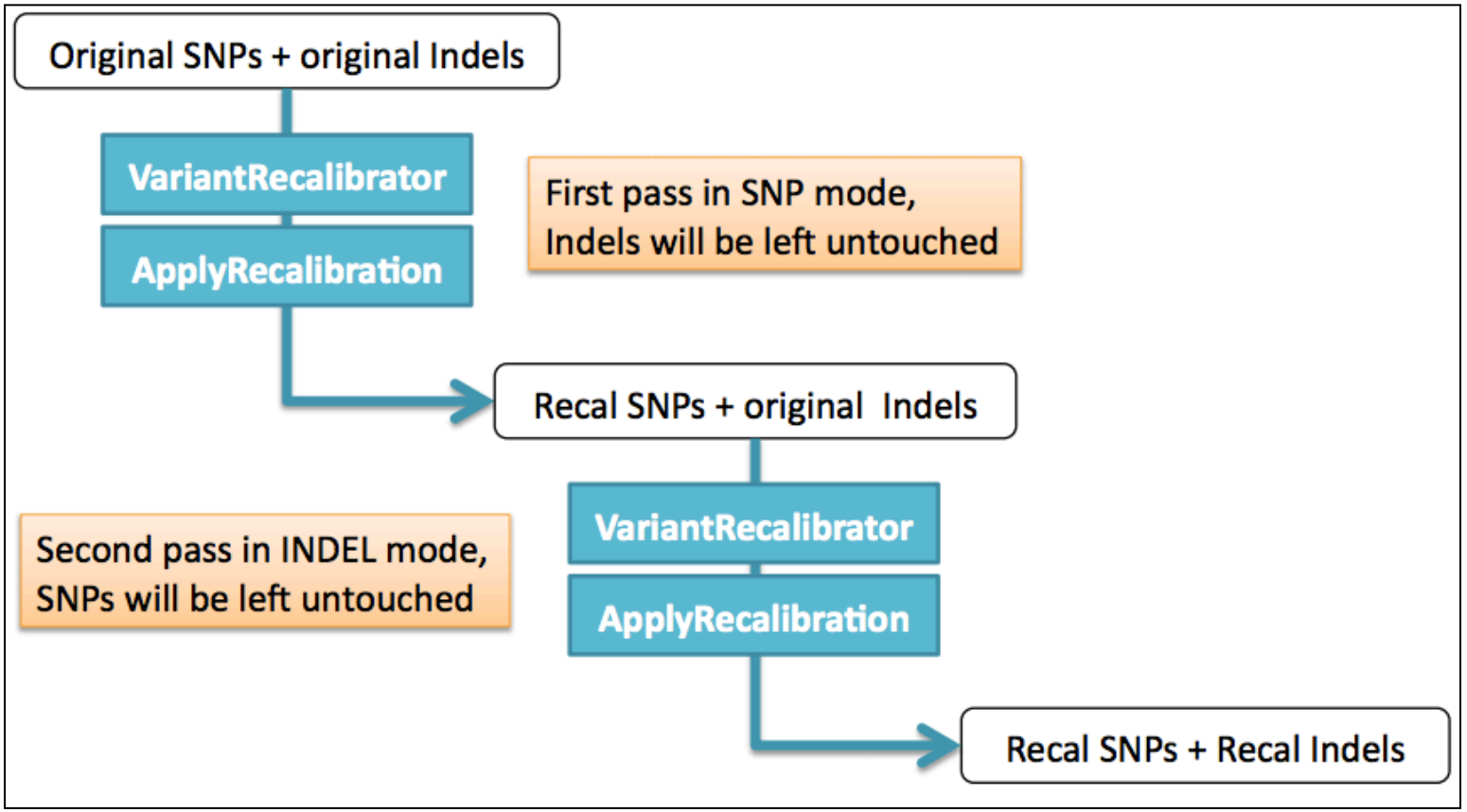
Variant Quality Score Recalibration



- The variant calling process is relatively permissive, and produces many false positives
- The variant recalibration workflow compares properties of novel predicted variants to those of variants known to exist in the population (from dbSNP database)
- Basically, this step filters out likely false positives producing a high-quality set

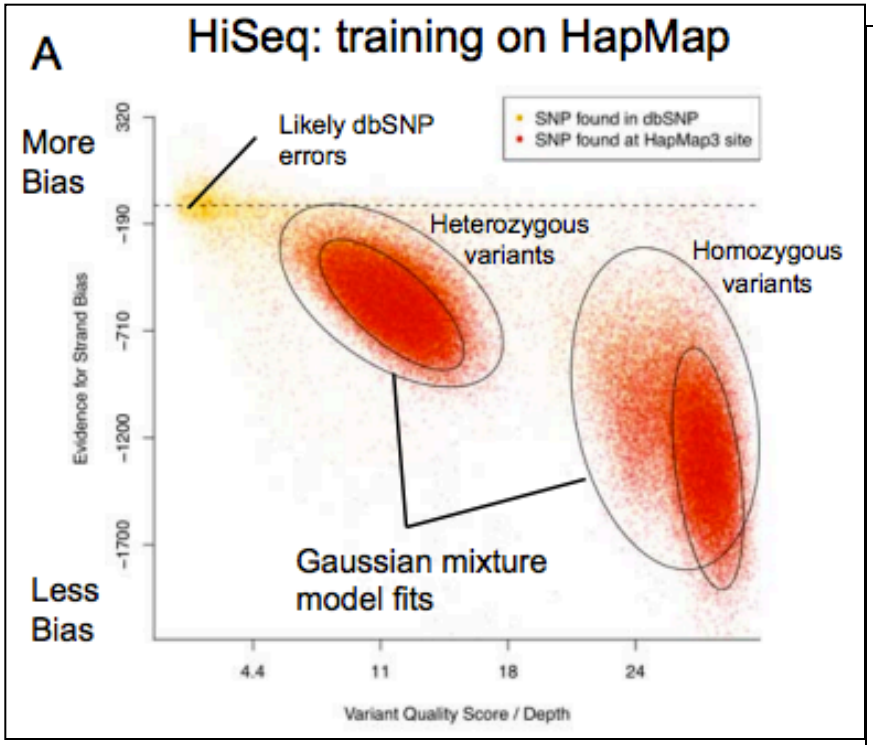
Calling variants with the GATK

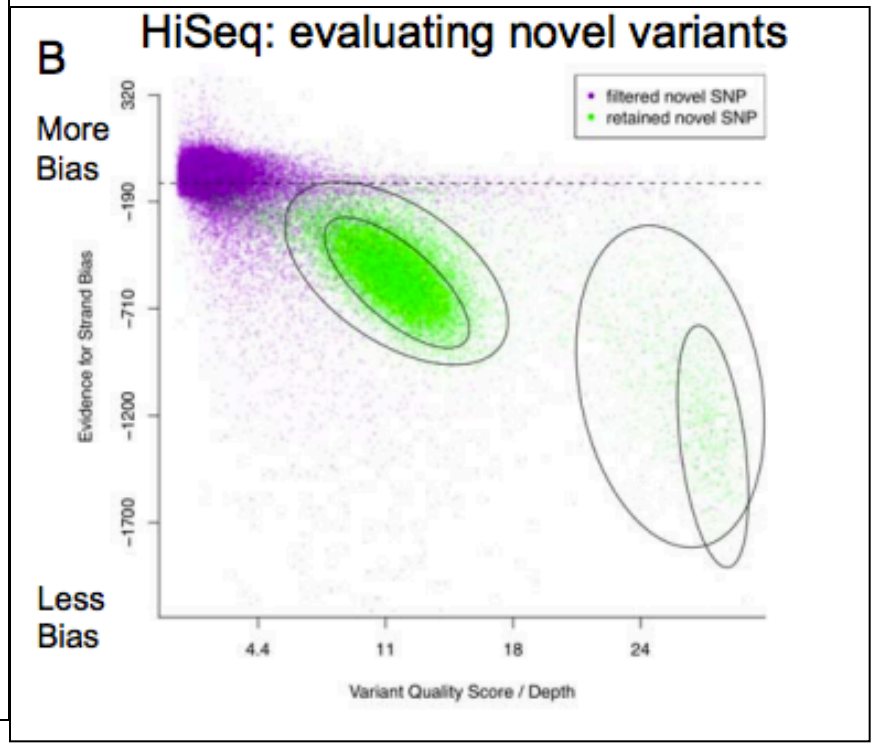
Variant Quality Score Recalibration



Calling variants with the GATK

Variant Quality Score Recalibration





Catalogs of human genetic variation

The 1000 Genomes Project

- http://www.1000genomes.org/
- SNPs and structural variants
- genomes of about 2500 unidentified people from about 25 populations around the world will be sequenced using NGS technologies

HapMap

- http://hapmap.ncbi.nlm.nih.gov/
- identify and catalog genetic similarities and differences

dbSNP

- http://www.ncbi.nlm.nih.gov/snp/
- Database of SNPs and multiple small-scale variations that include indels, microsatellites, and non-polymorphic variants

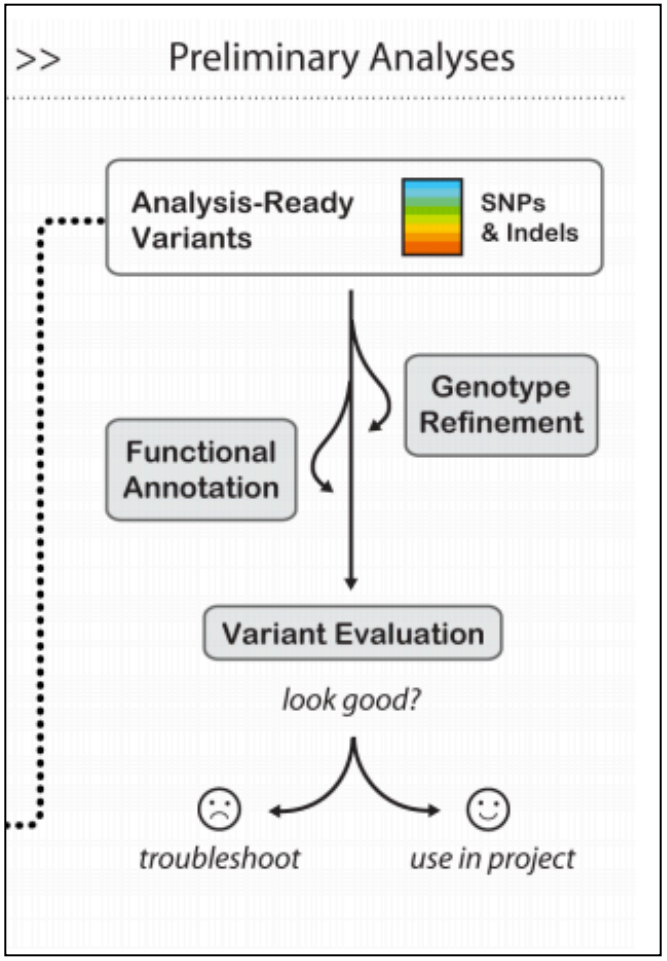
COSMIC

- http://www.sanger.ac.uk/genetics/CGP/cosmic/
- Catalog of Somatic Mutations in Cancer

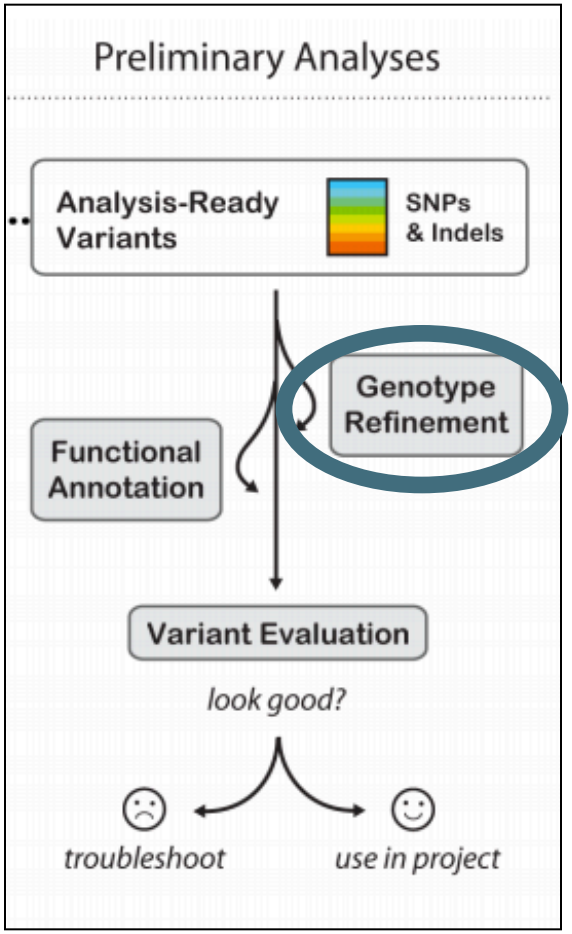
TCGA

- http://cancergenome.nih.gov/
- The Cancer Genome Atlas researchers are mapping the genetic changes in 20 selected cancers

Variant Calling Data Processing Steps



Calling variants with the GATK

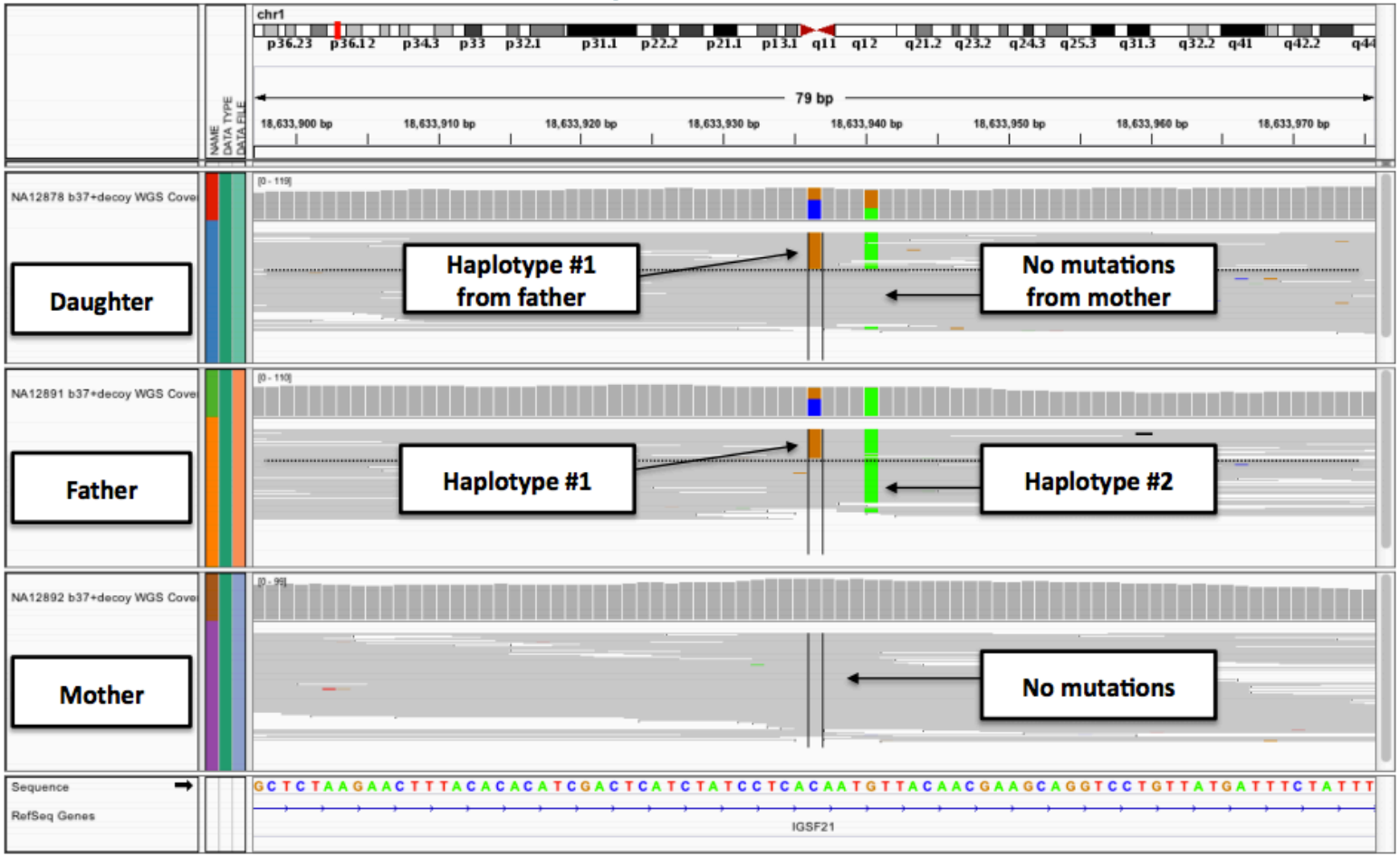


Genotype Refinement

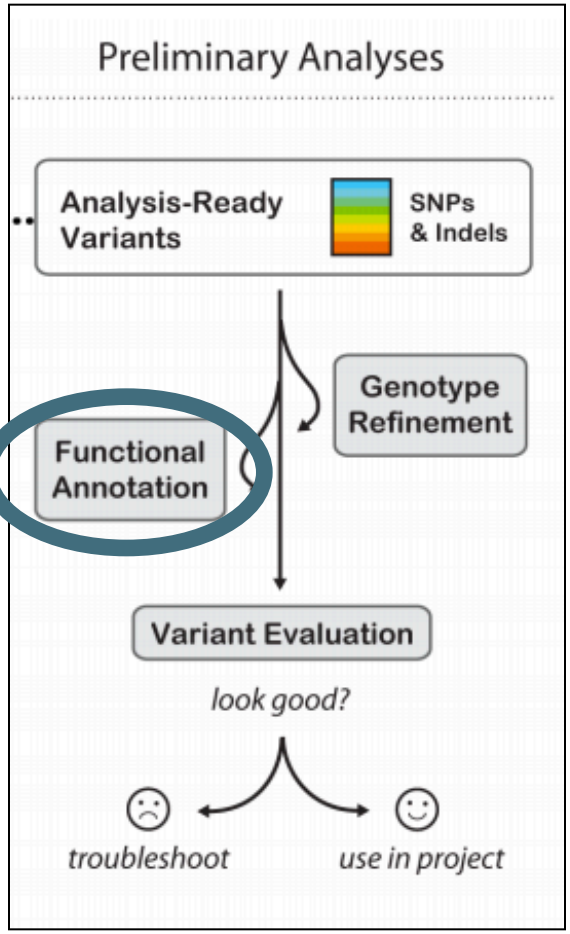
- Improve the genotype assignments and inferring haplotypes for your samples
- Infer phasing information based on population analyses using familial information or random population information
- Critical in population genetics studies to determine haplotype structure

Calling variants with the GATK

Genotype Refinement



Calling variants with the GATK



Functional Annotation

- Which gene is affected?
- Is the change in a coding or non-coding region?
- Does the mutation create *synonymous* or a *non-synonymous* change?

Calling variants with the GATK

Functional Annotation

snpEff (non-GATK)

Add functional annotations to a set of variants

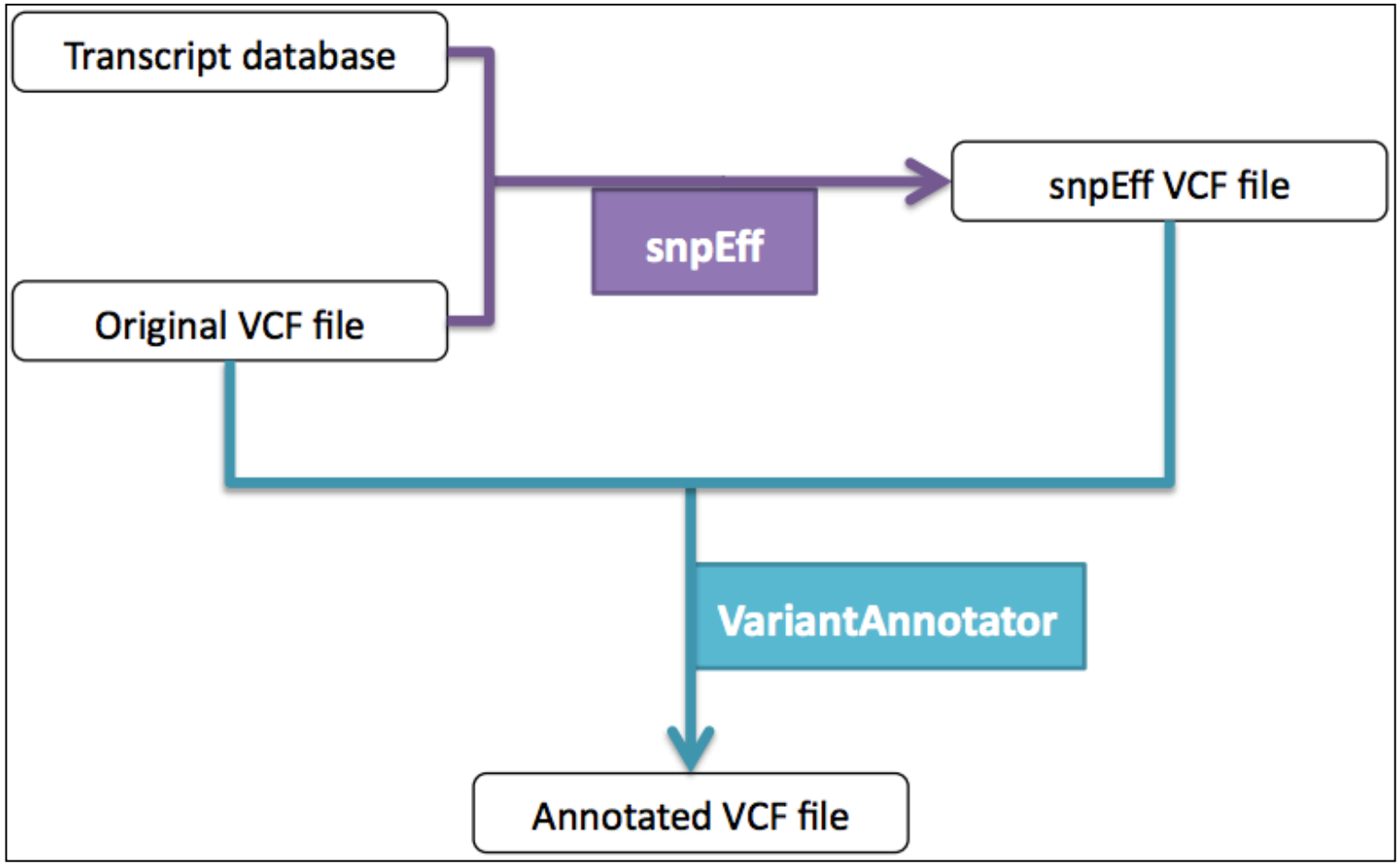
SnpEff annotates and predicts the effects of variants on genes (such as amino acid changes).

SnpEff annotation gives the following information:

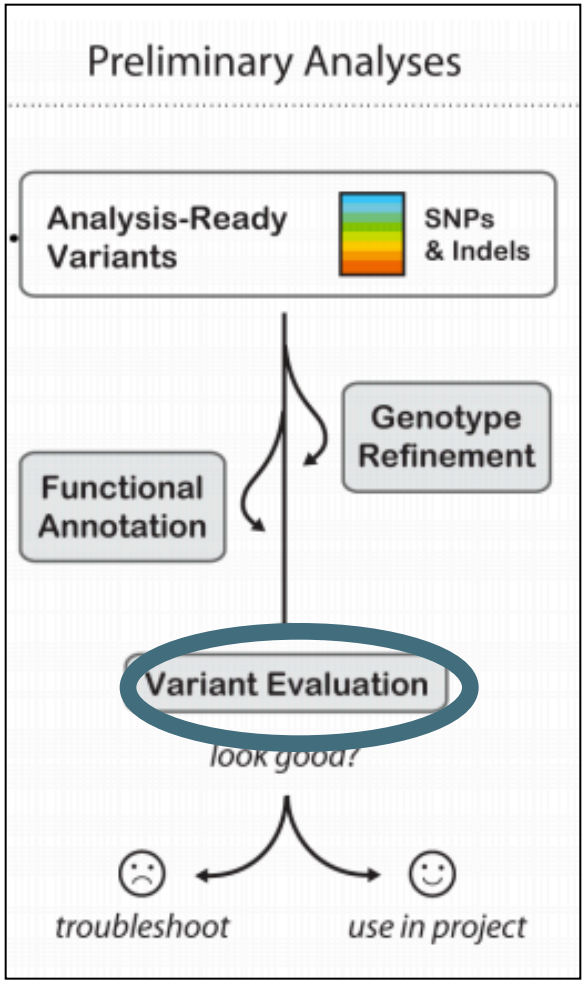
- Is the variant genic or intergenic, exonic or intronic, in a UTR?
- Change caused by variant is synonymous or non-synonymous?

Calling variants with the GATK

Functional Annotation



Calling variants with the GATK



Variant Evaluation

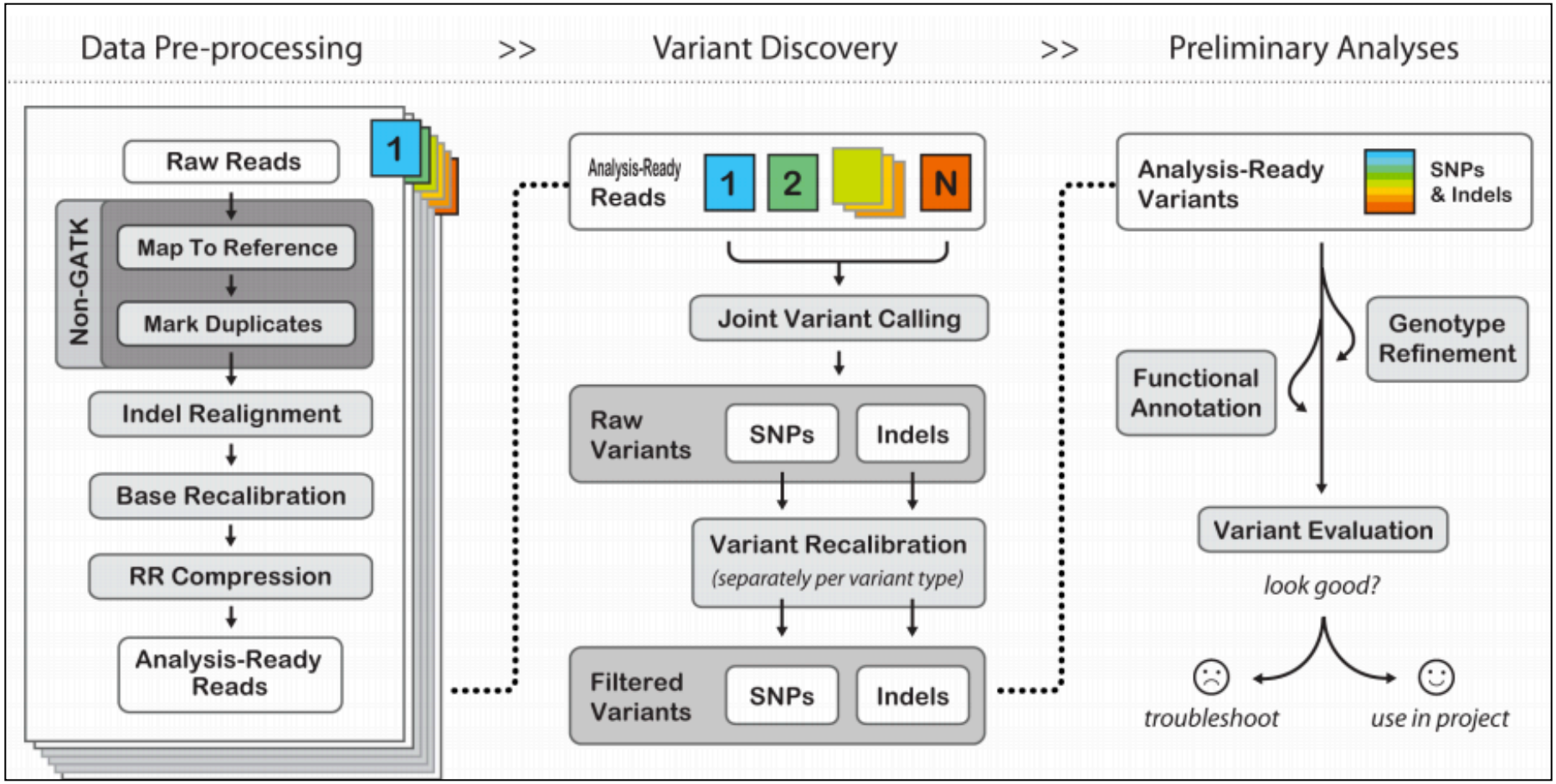
- When compared to known variant databases how do the basic statistics compare?

It is very important to compare apples to apples at this step:

  - Compare to a matched dataset
  - Pick the database or a subset of a database derived from the population closest to your population of interest

- How many of them are unique between the samples or groups?

Variant Calling Data Processing Steps

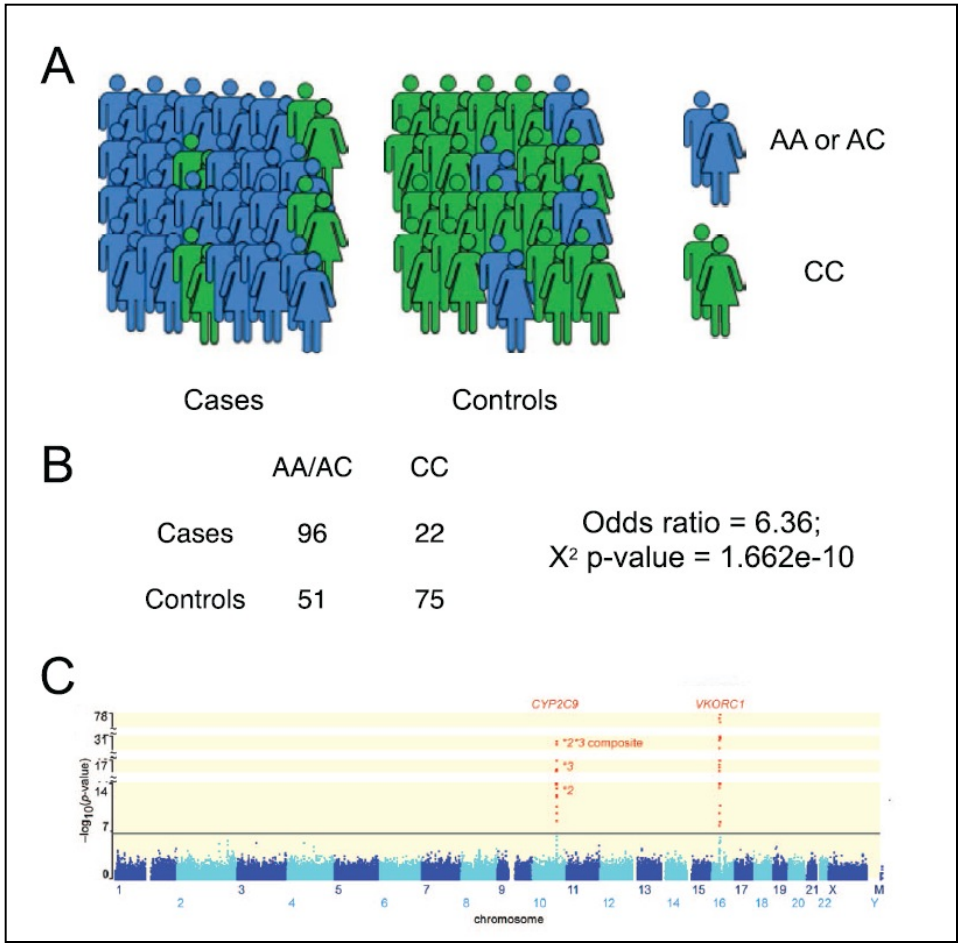


A complex puzzle…

Variant Calling – Interpretation

- Fundamental problem in biology: how does genotype inform phenotype, and what other factors (e.g., environmental, epigenetic) are involved?
- For some phenotypes, e.g., those linked to ethnic differences, or highly penetrant Mendelian traits, we can predict phenotype from genotype quite accurately
- For many “complex” traits where we know that there is a strong inherited component (e.g., from twin and family studies), we still have a ways to go
- Two common approaches:
  - Genome-wide association studies (GWAS)
  - Integrated analyses

GWAS – basic principles



(A) In a case / control study, genotypes are determined for all cases and controls

(B) For each allelic variant, the distribution of alleles between cases and control groups is measured, and deviation from a random distribution calculated

(C) The X^2 p-values and positions in the genome for each of the measured loci are displayed in a Manhattan plot. The Figure shows two genomic regions enriched for variants with highly significant distribution biases that putatively contain causal variants for the trait being analyzed.

Konrad J. Karczewski[1,2], Roxana Daneshjou[2,3], Russ B. Altman[2,3]*

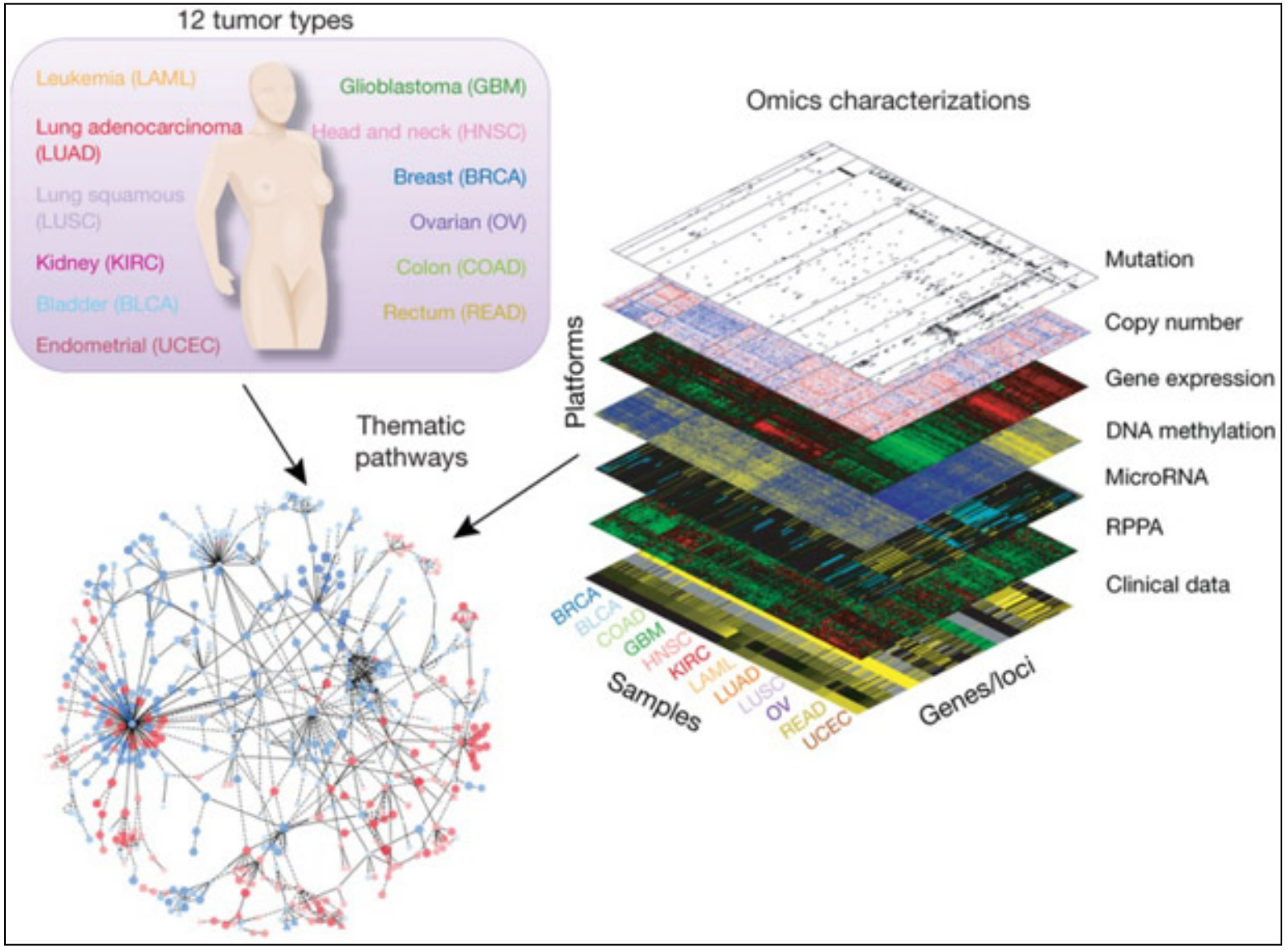
1 Program in Biomedical Informatics, Stanford University, Stanford, California, United States of America, **2** Department of Genetics, Stanford University, Stanford, California, United States of America, **3** Department of Medicine, Stanford University, Stanford, California, United States of America

GWAS – extension

Detection of epistatic interactions:

- Assumption is that phenotype is influenced by more than one allelic variant, and that the effects are synergistic (more than additive)
- Ideally involves the calculation of the effect sizes of all combinations of observed alleles, and a comparison to their individual effects
- Computationally very challenging depending on the number of variants

Integrated Analyses



Source: The Cancer Genome Atlas, TCGA


it's tea time!

Genome sequencing and Variant Calling

- Introduction to using NGS for Variant Detection
- Sequencing Technologies, specifically Illumina
- File Formats, FASTQ, SAM, BAM, vcf, bcf
- QC steps
- Variant Calling (data processing)

Tea Break

- Computational Requirements
  - Data Storage
  - Processing Capacity

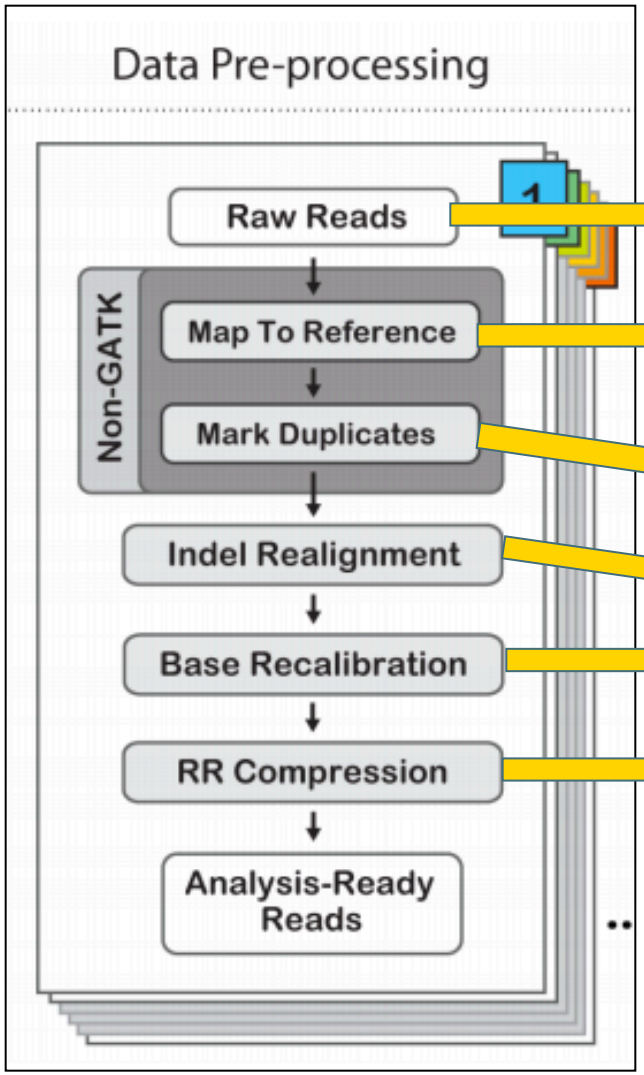
Brief Introduction to using NGS for microbiome analysis

Computational Requirements

Data Storage Requirements

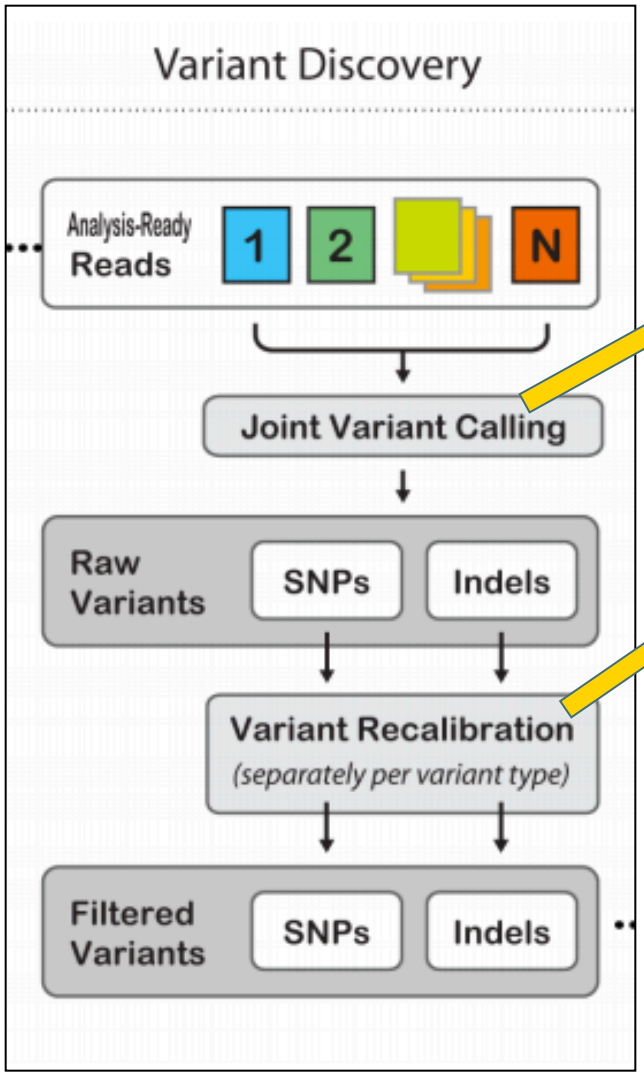
Processing Capacity Required

Data Storage Requirements



Input files	Output files	Cumulative Size
	PE FASTQ ~10-15X coverage	~100GB
PE FASTQ	SAM	~300GB
Large SAM	Sorted BAM, de-dupped BAM, BAM with read groups added	~450GB
Large BAM	Realigned BAM	~600GB
Realigned BAM	BAM with recalibrated quality	~750GB
BAM with recalibrated quality	Reduced BAM	~810GB

Data Storage Requirements



Input files	Output files	Cumulative Size
Reduced BAM	Small vcf files (text files)	~820GB
Small vcf files (text files)	More small vcf files (text files)	**~825GB**

Data Storage Requirements

Grand total for the first 2 parts of the pipeline: > 0.8 TB

Oh wait, that's per sample for 15x coverage!!

And that's only the first phase of the pipeline ***and*** it is a middle-case scenario for WGS data!!

50x Exome = 30GB raw FASTQ + everything else

50x WGS = 350GB raw FASTQ + everything else

Data Storage Requirements

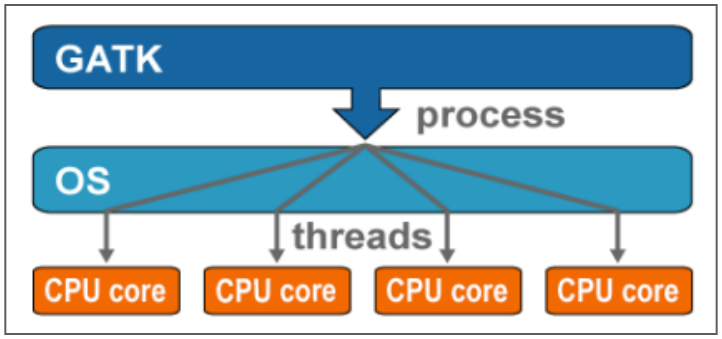
- Storage and data backup plans should be ready before you get the raw FASTQ data from the sequencing center
- Maintain a good documentation of steps as well as data organization; this is especially crucial for a large population analysis
- Remove any files that are easily re-creatable; e.g. once you have the BAM file that has been realigned and recalibrated for quality scores, you can maybe afford to lose the 2 previous versions of the BAM files *(I'd save the original BAM)*
- Compress any "compressable" files, e.g. vcf to bcf or SAM to BAM
- Compress (tar) everything prior to archiving (long-term storage)

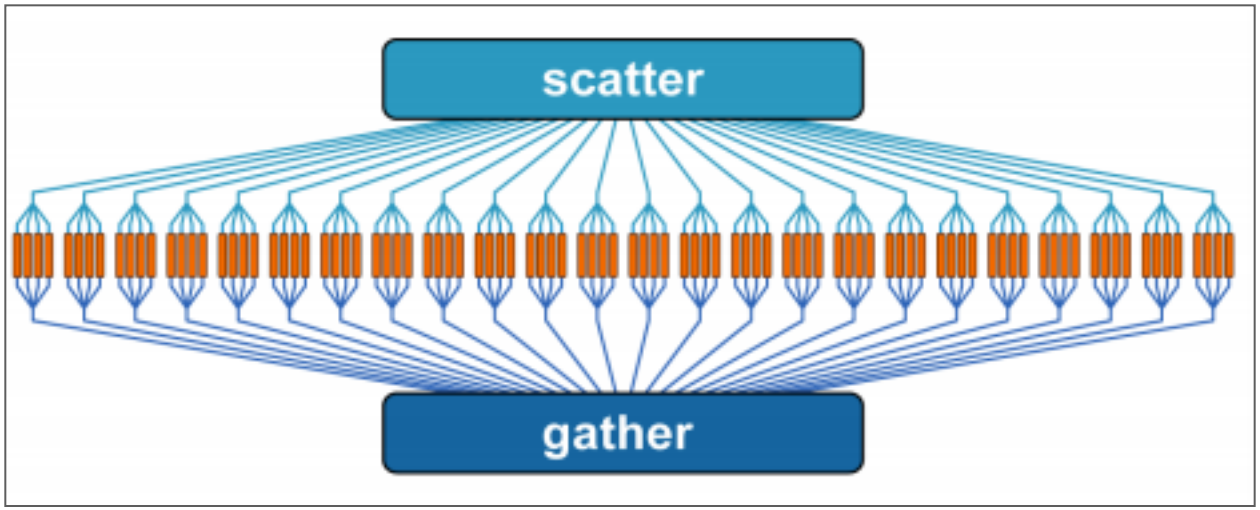
Computational Requirements

Data Storage Requirements

Processing Capacity Required

Parallelism





Parallelism

- Some software can be run in a “multi-threaded” mode, wherein the parallelization is built in (multiple cores)
- Parallelization when feasible is great, but this is not always the case. For some steps there is no efficient way to parallelize…
- There are ways external to the software that can be used to optimize efficiency and are not mutually exclusive
  - Example 1 - align smaller chunks of the fastq files to the genome simultaneously
  - Example 2 - separate out all the aligned data by chromosome, and run the downstream analysis per chromosome

Processing Requirements

Step	Hours on 1 core 50x WGS	Hours on 1 core 50x Exome
QC	10	0.5
Alignment (parallelizable)	320	55 to 111 (depending on the aligner used)
Sorting, de-dupping, read group addition (partly parallelizable)	35	2
Indel realignment + Recalibration (***Forcibly* parallelizable, by separating into chromosomes**)	69	6
Variant Calling (UG) (***Forcibly* parallelizable, by separating into chromosomes**)	40	5

Processing Requirements

Step	Hours on 1 core 50x WGS	Hours on 1 core 50x Exome
QC	10	0.5
Alignment (parallelizable)	320	55 to 111 (depending on the aligner used)
Sorting, c		2
Indel realignment + Recalibration (***Forcibly*** **parallelizable, by separating into chromosomes**)	69	6
Variant Calling (UG) (***Forcibly*** **parallelizable, by separating into chromosomes**)	40	5

When using GATK or a similar pipeline and making use of parallelization for efficient processing, the memory requirements are fairly low, 10 GB per process is deemed to be enough.

A brief introduction to using NGS for microbiome analysis

Microbiome (in the human context)

- For every single human cell there are at least 10 microbial cells in or on our bodies making up ~500 grams of your body weight
- Joshua Lederberg coined the term "***microbiome**, to signify the ecological community of commensal, symbiotic, and pathogenic microorganisms that literally share our body space*" (Scientist 2001)
- Prior to high-throughput sequencing, microbiome analysis was restricted to microorganisms that could be cultured in the lab
- Sampling has been performed from various parts of the human body, on the surface and from within the body cavity

Microbiome (in the human context) (contd.)

Why study the microbiome?

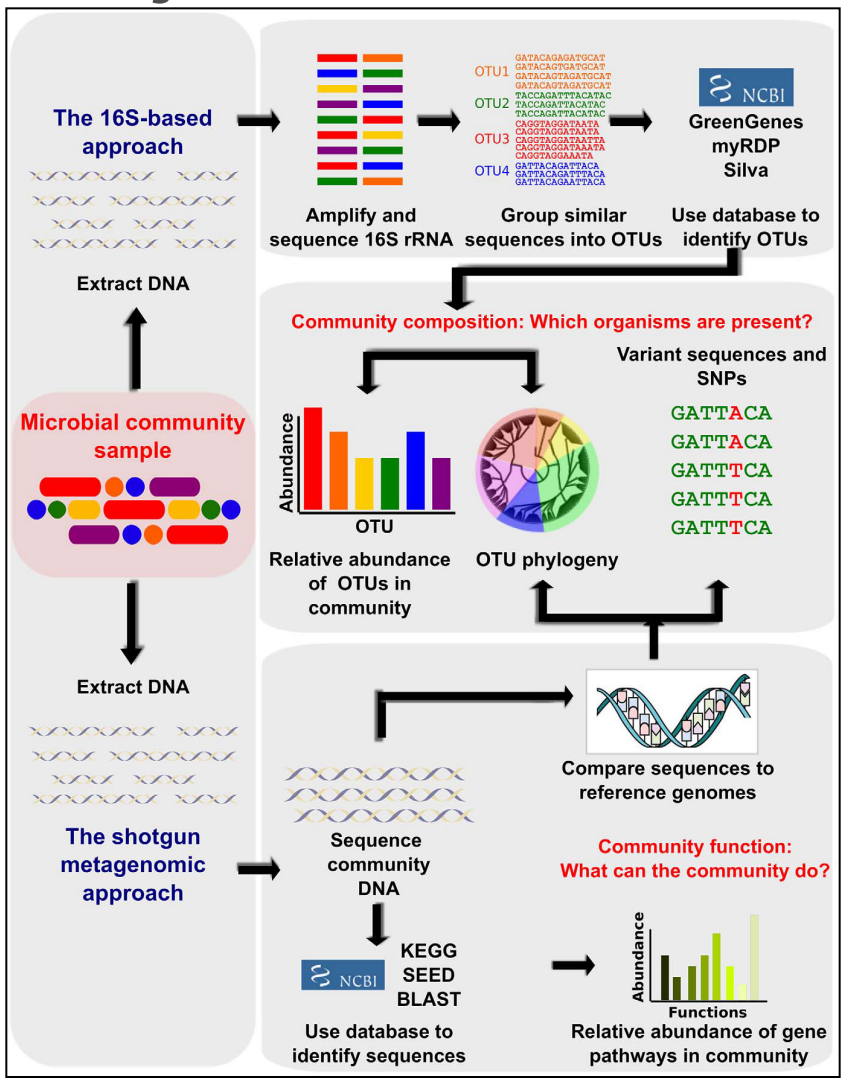
- To understand the contribution of this large population of cells on the human body
- To study if and how it impacts disease states
- To assess the impact of environmental factors on the microbiota and downstream phenotypes

Methods to study the Microbiome

To study a specific microbiome (e.g. intestinal, vaginal, feet etc.), you can isolate and study the DNA or the RNA from the environmental samples

- DNA-based methods
  - Taxonomic diversity by sequencing variable 16S regions
  - Shotgun sequencing of the whole *metagenome*
    - Functional information, what genes are enriched in the microbiome?
    - Taxonomic diversity
- RNA-based method
  - Shotgun sequencing of the whole *metatranscriptome*
    - More direct functional information about gene expression

Two major DNA-based methods



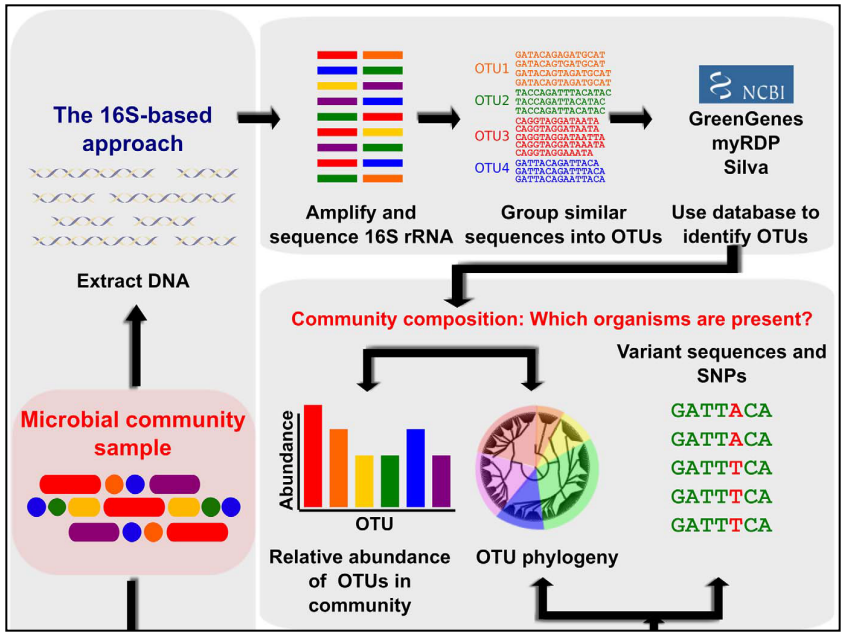
Morgan XC, Huttenhower C (2012)
Chapter 12: Human Microbiome Analysis.
PLoS Comput Biol 8(12)

Two major DNA-based methods

To study the DNA isolated from environmental samples to assess the diversity of microbial community in that environment

1. Isolate and sequence the DNA that encodes 16S ribosomal RNA
   - 16S ribosomal RNA (DNA) is very well conserved among bacterial and archaeal species, with small sections of “hypervariable” regions (e.g. V4, V6 etc.)
   - Because of the highly conserved areas outside of these hypervariable regions, common or universal primers can be designed to amplify the variable DNA
   - The sequence of the hypervariable regions can be utilized to characterize the sequence into taxonomic groups
   - Thus the basic output of such an analysis is OTUs or Operational Taxonomic Units or phylotypes
   - Well-developed methods available for this analysis

16S-based approach



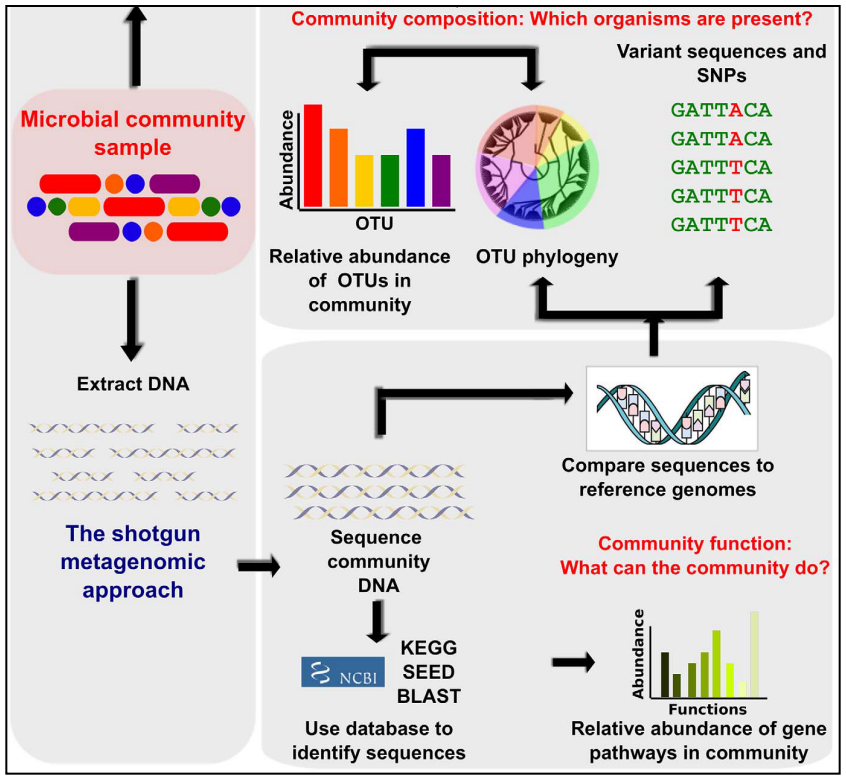
Morgan XC, Huttenhower C (2012)
Chapter 12: Human Microbiome Analysis.
PLoS Comput Biol 8(12)

Two major DNA-based methods

To study the DNA isolated from environmental samples to assess the diversity of microbial community in that environment

2. Shotgun sequencing of the whole metagenome
   - True metagenomics, since you are potentially looking at the whole genomes of the microbial community in the sample
   - This method also enables taxonomic diversity analysis
   - Often environmental constraints result in selective metabolic processes being enriched in a given environment and this method can enable gathering functional information of this nature
   - Methods are still being developed and there are many opinions about right and wrong

Shotgun sequencing – metagenomics

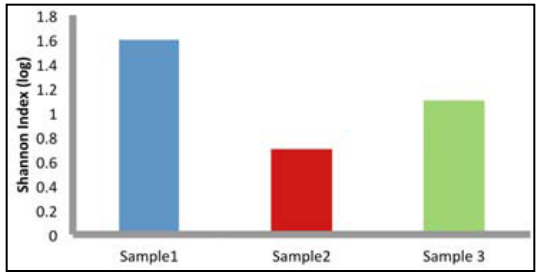


Morgan XC, Huttenhower C (2012)
Chapter 12: Human Microbiome Analysis.
PLoS Comput Biol 8(12)

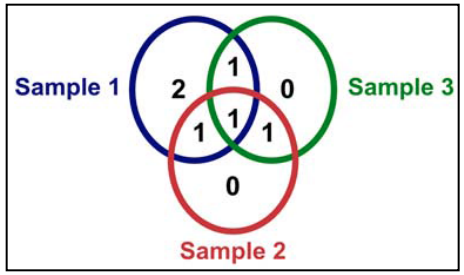
- Filter out host DNA by alignment to the host genome [lots of low-memory processors]
- 2 alternatives for filtered short reads
  1. Align the remaining short-reads to various databases to identify taxonomic and genic information [lots of low-memory processors]
  2. Alternatively, assemble the short reads into longer pieces or contigs before aligning to the databases. [at least one high-memory processor + lots of low-memory processors]

Two major DNA-based methods

- Broad taxonomic classifications can be made with both the methods
- Diversity plots of OTU composition/profiles provide key insight-
  - Alpha or within-sample diversity



  - Beta or between-sample diversity



- The whole-genome shotgun approach will also provide genic, i.e. functional information (metabolic process enrichment, etc.)

Metatranscriptomics

To study the transcriptome (RNA) isolated from environmental samples.
Shotgun sequencing of the metatranscriptome

» How to sequence the RNA component of a sample?

- Treat the sample extremely carefully to prevent degradation (GIGO)
- Library preparation involves isolating the RNA you are interested in (removing anything that looks like ribosomal RNA)
- Convert the RNA to DNA using reverse transcription
- Proceed with making a dsDNA library from the complementary DNA (cDNA)
- Sequence as usual

Metatranscriptomics

To study the transcriptome (RNA) isolated from environmental samples.
Shotgun sequencing of the metatranscriptome

- Filter out host RNA by alignment to the host genome
- Filter out any remaining rRNA by alignment to databases like SILVA
- 2 alternatives for filtered short reads (similar to DNA)
  1. Align the remaining short-reads to various databases to identify genic information
  2. Alternatively, assemble the short reads into longer pieces or contigs before aligning to the genic databases.
- Deduce the role of the microbiome based on the transcripts expressed

Methods to study the Microbiome

To study a specific microbiome (e.g. intestinal, vaginal, feet etc.), you can isolate and study the DNA or the RNA from the environmental samples

- DNA-based methods
  - Taxonomic diversity by sequencing variable 16S regions
  - Shotgun sequencing of the whole *metagenome*
    - Functional information, what genes are enriched in the microbiome?
    - Taxonomic diversity
- RNA-based method
  - Shotgun sequencing of the whole *metatranscriptome*
    - More direct functional information about gene expression

Acknowledgements

The GATK team for sharing their materials generously

Dr. C. Victor Jongeneel

Dr. Christopher J. Fields

Dr. Liudmila S. Mainzer

Dr. Alvaro G. Hernandez

Dr. Chris L. Wright

Daniel Davidson