HIGH-THROUGHPUT GENOTYPING WITH SEQUENCING DATA ANALYSIS

Shichen Wang, PhD
Bioinformatics Scientist

Genomics and Bioinformatics Service
Texas A&M University AgriLife Research

Genomics and Bioinformatics Service

Providing Genomics and Bioinformatics Services to the Texas A&M System, Texas, and the World

"How do I start a new sequencing project?"

Questions related to your samples, or the submission process

Search

Welcome

Who We Are

News

Research Highlights

Bioinformatics

Personnel

Publications

FAQ

Contacts

NEXT-GENERATION SEQUENCING, NGS

- DNA sequencing
  - Early methods for reading the DNA
  - High-throughput sequencing (HTS) methods, NGS
    - Sequencing by synthesis (Illumina)
    - Ion semiconductor (Ion Torrent sequencing)
    - GenapSys Sequencing
  - Long read sequencing technology, TGS?
    - **Single-molecule real-time sequencing (Pacific Biosciences)**
    - **Nanopore Sequencing**

NEXT-GENERATION SEQUENCING, NGS

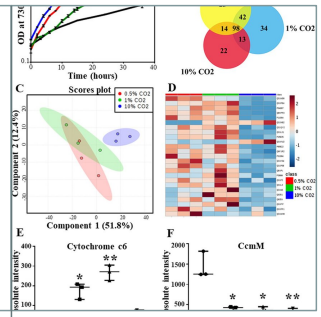
Sample

DNA/RNA extraction

Quality control

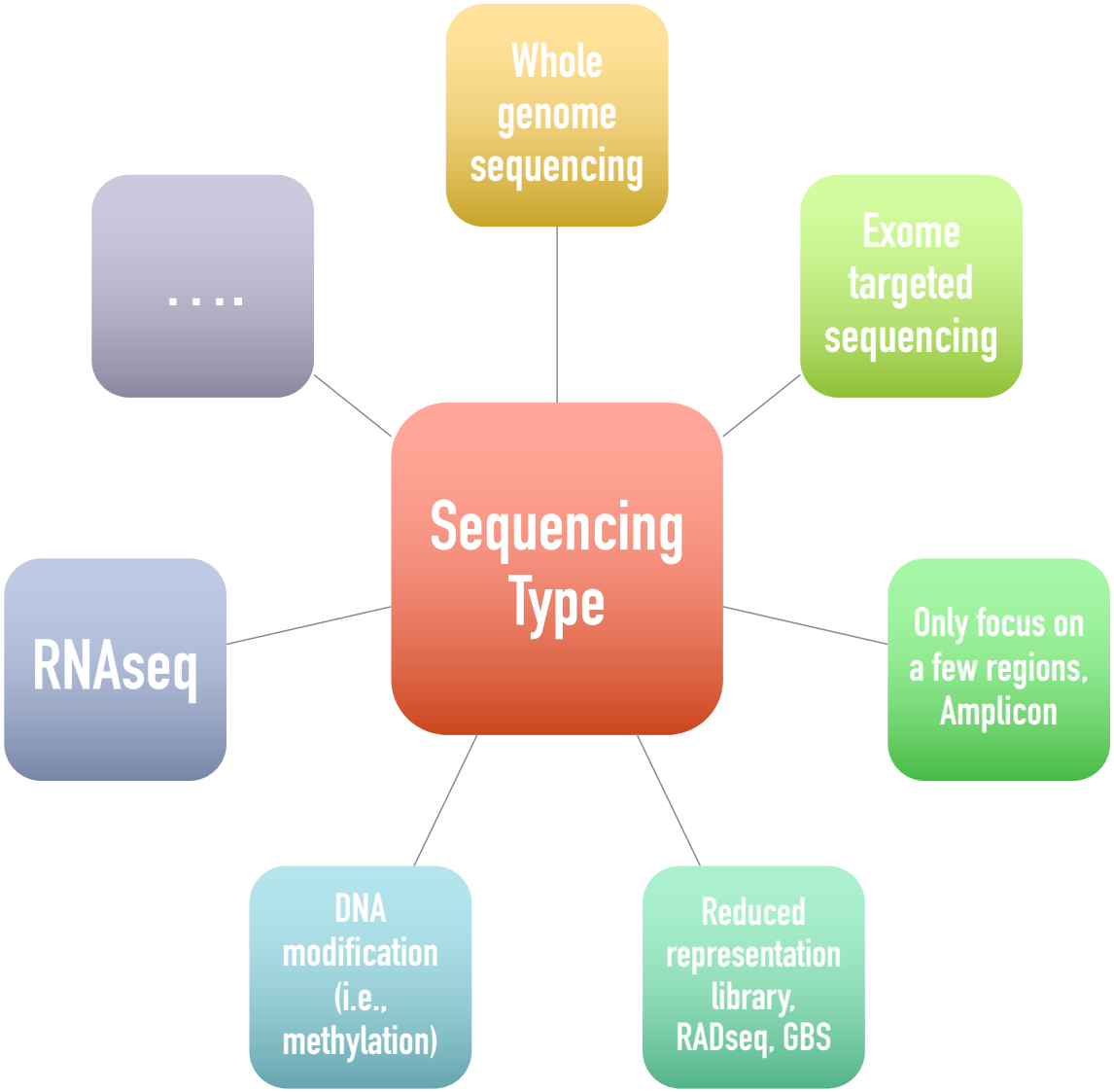
library construction

Sequencing



Bioinformatics analysis

NEXT-GENERATION SEQUENCING, NGS

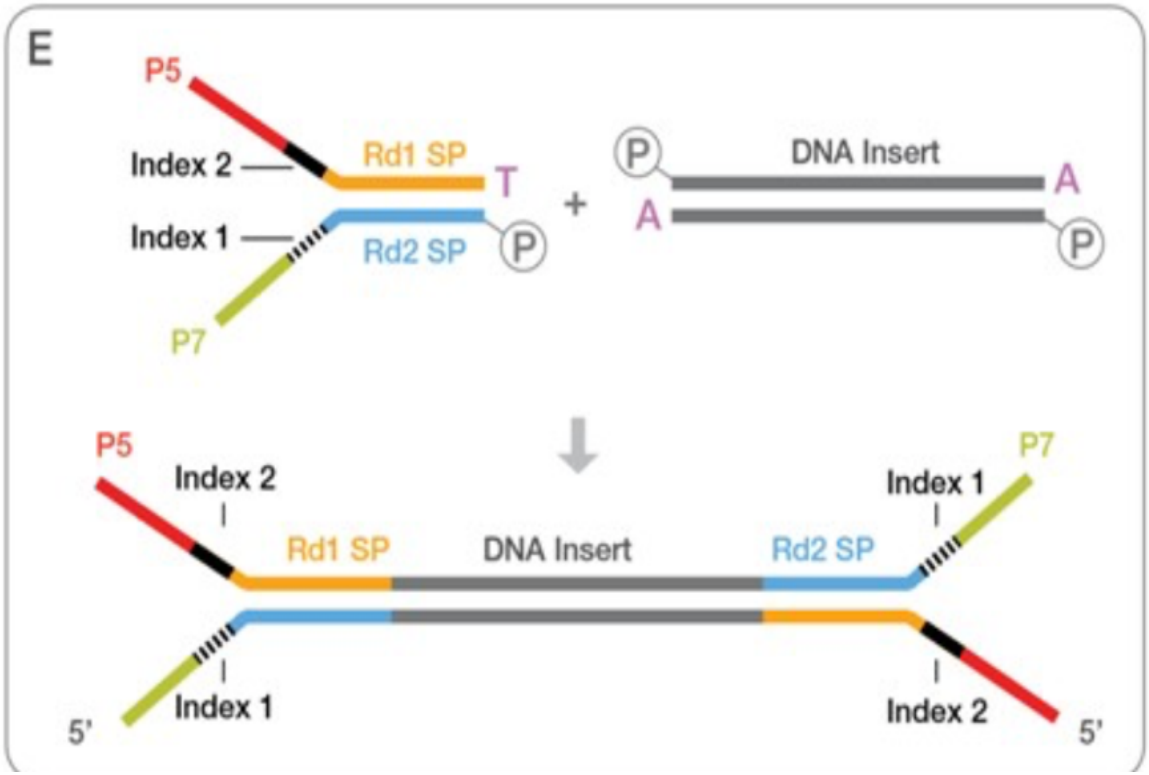


NEXT-GENERATION SEQUENCING, NGS

- **Read length**
  - The sequences were called reads. One read might consist of 75 bp, 100 bp, or more. Longer reads can provide more reliable information about the relative locations of specific base pairs. However, it is usually more expensive to generate longer reads.
- **Single-End or Pair-End**
  - In single-end reading, the sequencer reads a fragment from only one end to the other, generating the sequence of base pairs. In paired-end reading it starts at one read, finishes this direction at the specified read length, and then starts another round of reading from the opposite end of the fragment.
  - Paired-end reading improves the ability to identify the relative positions of various reads in the genome, making it much more effective than single-end reading in resolving structural rearrangements such as gene insertions, deletions, or inversions. It can also improve the assembly of repetitive regions. This degree of accuracy may not be required for all experiments, however, and paired-end reads are more expensive and time-consuming to perform than single-end reads.
- **Coverage depth**
  - The depth of coverage is a measure of the number of times that a specific genomic site is sequenced during a sequencing run.

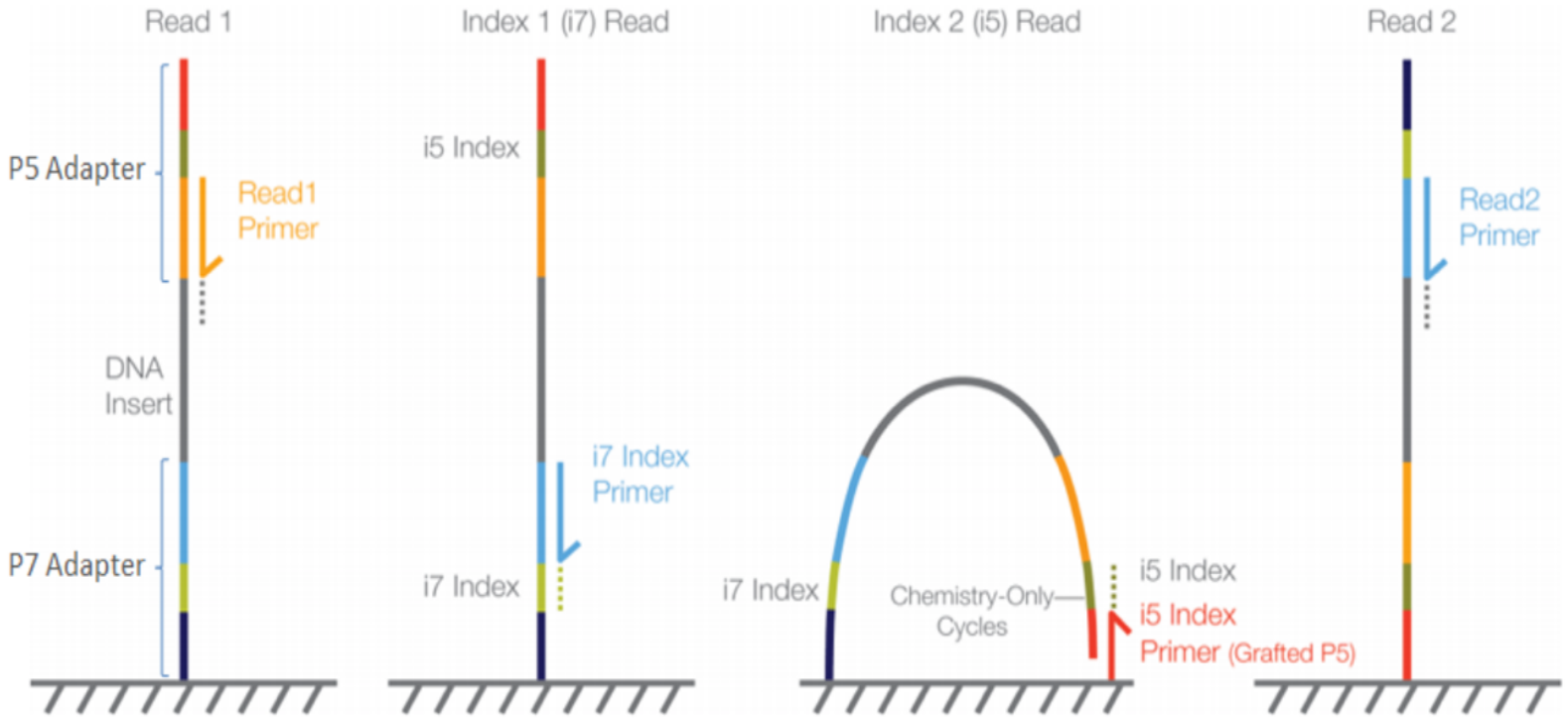
TruSeq library preparation

- Step #4: Ligate adapters containing sequencing primer, indices, flowcell capture site



Illmina.com

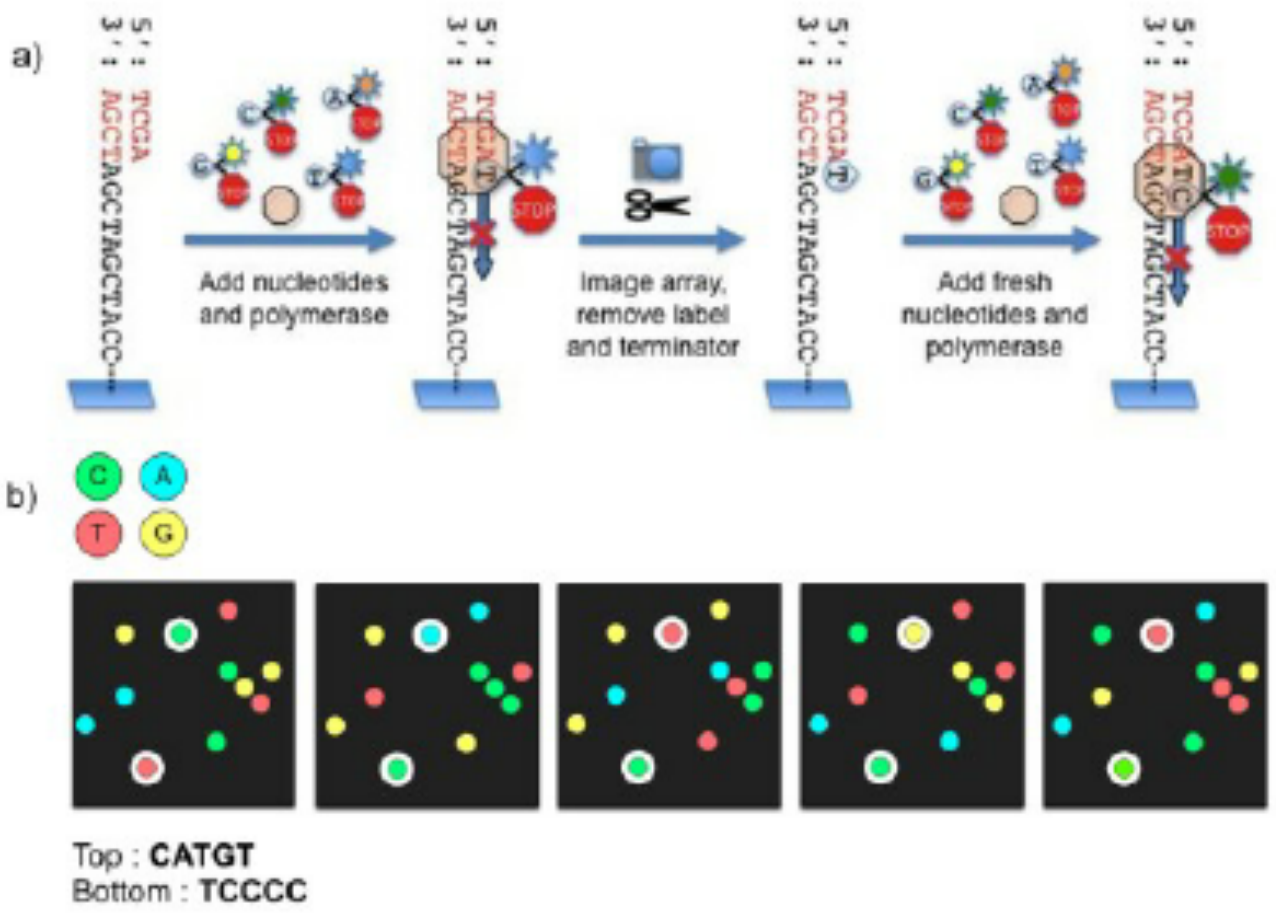
Figure 1. MiSeq, HiSeq 2000/2500 and NovaSeq paired-end flow cell



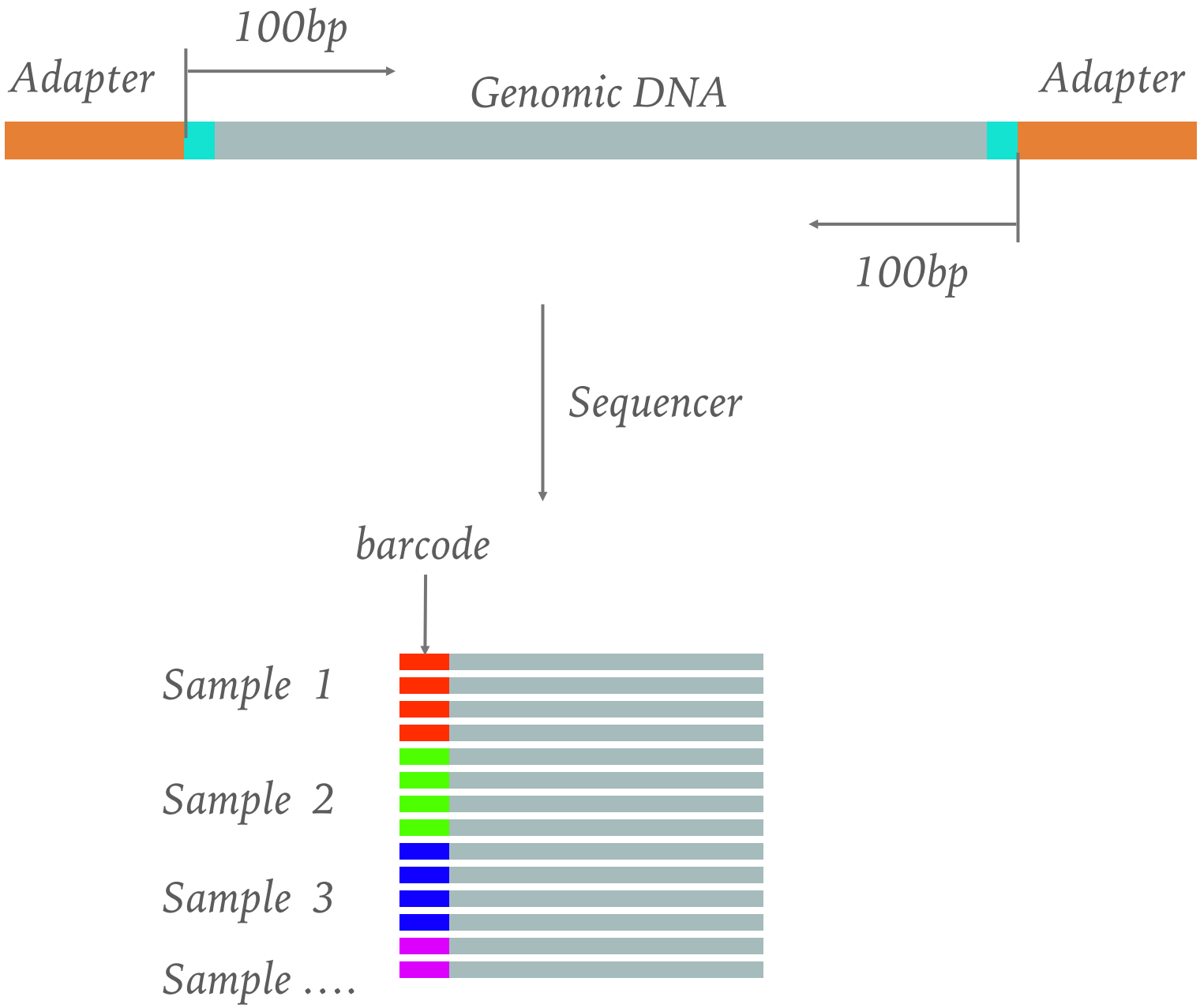
Illmina.com

BIOINFORMATICS DATA ANALYSIS

- What kind of data being generated?



Sequence by Synthesis, Anderson and Schrijver, 2010


100bp
Adapter
Genomic DNA
Adapter
100bp
Sequencer
barcode
Sample 1
Sample 2
Sample 3
Sample ….

BIOINFORMATICS DATA ANALYSIS

- What kind of data being generated?
  - fastq format
    - each read represented by 4 lines

line 1 `@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1`
line 2 `TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGATTTGTTGGGGGAG`
line 3 `+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1`
line 4 `efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd`ddd^dddadd^BB`

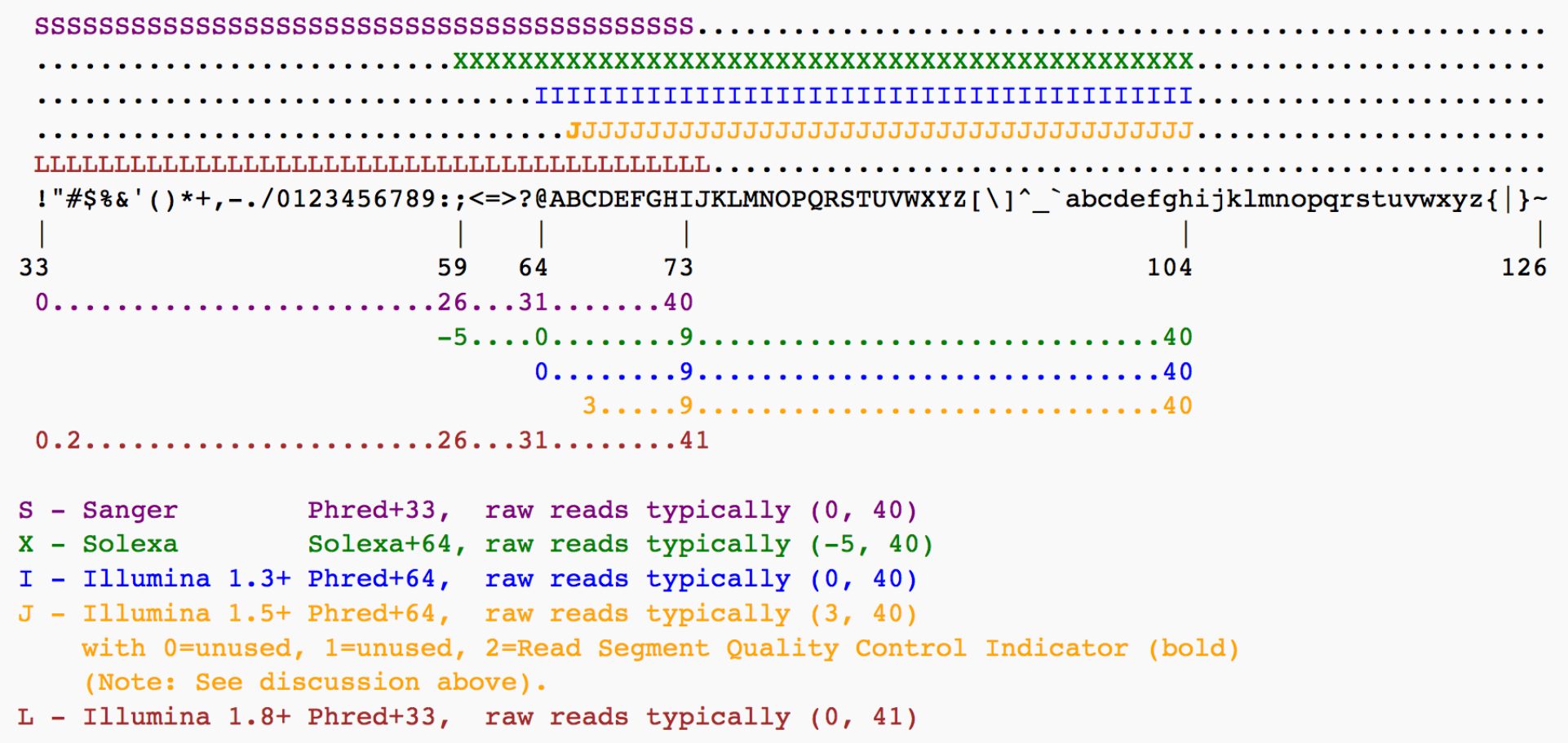
BIOINFORMATICS DATA ANALYSIS

- What kind of data being generated?
  - fastq format

```
[wangsc@login3 ~/fastq_]$ zcat VijayWSC4-Ind7_CAGATC_L004_R1_001.fastq.gz | head
@HISEQ:224:C8PEWACXX:4:1101:1269:1962 1:N:0:CAGATC
AAAAAGAACAAGATTGAGACTAGATCGTGAAGCGAATGTAAGCATACTCACCTTATACCATTCACCATCAAGTCCTGTTTTCCACACGGCTGTGATGACAT
+
BBBFFFFFFFFFFIIIIFBFFFIBBFFFB<0<FFFFFIFIIIFIFBFFFIIIBBFFIIIIFFIFFB<BFFFFF'BBFFFBBBBBBBBB'7B<B<B<BFBBB
```

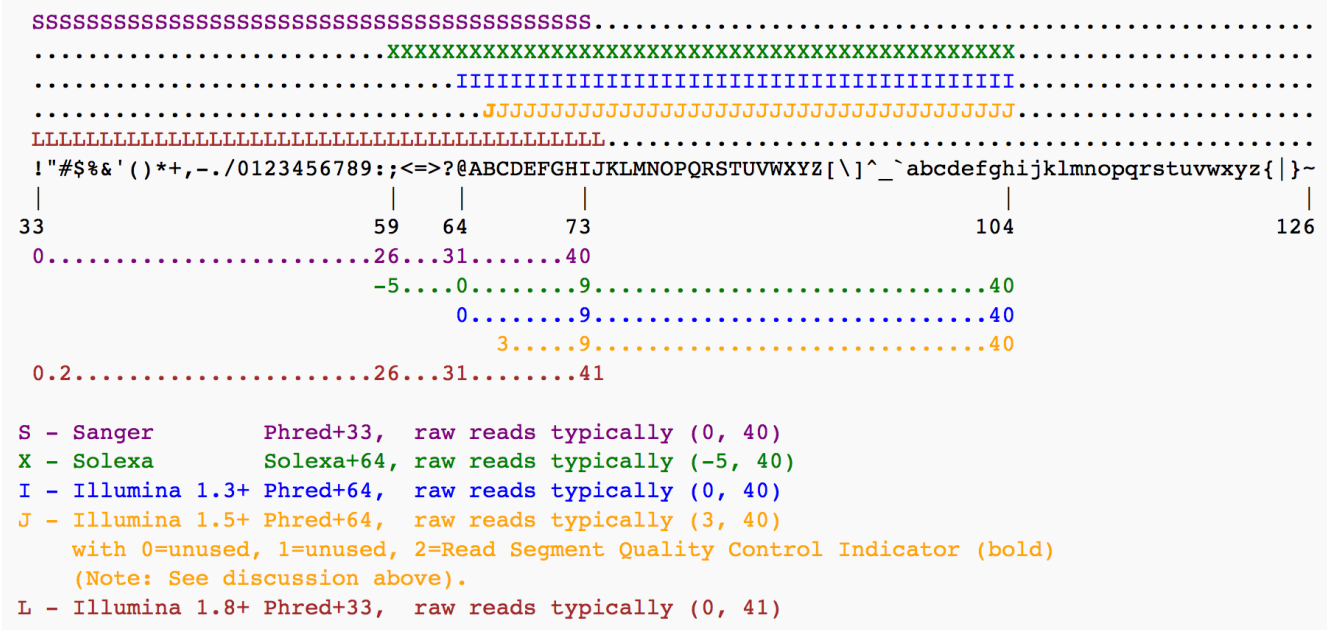
BIOINFORMATICS DATA ANALYSIS

- What kind of data being generated?
  - fastq format



BIOINFORMATICS DATA ANALYSIS

- What kind of data being generated?
  - fastq format



A quality score Q is an integer mapping of p (the probability that the base casting is wrong.

$$Q_{\text{sanger}} = -10 \log_{10} p$$

BIOINFORMATICS DATA ANALYSIS

- File format used by bioinformatics tools
  - Fasta
    - contains a sequence name, a description of the sequence (metadata, sequencer info, annotations, etc.), and the sequence itself

```
>Chr1 CHROMOSOME dumped from ADB: Jun/20/09 14:53; last updated: 2009-02-02
CCCTAAACCCTAAACCCTAAACCCTAAACCTCTGAATCCTTAATCCCTAAATCCCTAAATCTTTAAATCCTACATCCAT
GAATCCCTAAATACCTAATTCCCTAAACCCGAAACCGGTTTCTCTGGTTGAAAATCATTGTGTATATAATGATAATTTT
ATCGTTTTTATGTAATTGCTTATTGTTGTGTGTAGATTTTTTAAAAATATCATTTGAGGTCAATACAAATCCTATTTCT
TGTGGTTTTCTTTCCTTCACTTAGCTATGGATGGTTTATCTTCATTTGTTATATTGGATACAAGCTTTGCTACGATCTA
CATTTGGGAATGTGAGTCTCTTATTGTAACCTTAGGGTTGGTTTATCTCAAGAATCTTATTAATTGTTTGGACTGTTTA
TGTTTGGACATTTATTGTCATTCTTACTCCTTTGTGGAAATGTTTGTTCTATCAATTTATCTTTTGTGGGAAAATTATT
TAGTTGTAGGGATGAAGTCTTTCTTCGTTGTTGTTACGCTTGTCATCTCATCTCTCAATGATATGGGATGGTCCTTTAG
CATTTATTCTGAAGTTCTTCTGCTTGATGATTTTATCCTTAGCCAAAAGGATTGGTGGTTTGAAGACACATCATATCAA
AAAAGCTATCGCCTCGACGATGCTCTATTTCTATCCTTGTAGCACACATTTTGGCACTCAAAAAAGTATTTTTAGATGT
TTGTTTTGCTTCTTTGAAGTAGTTTCTCTTTGCAAAATTCCTCTTTTTTTTAGAGTGATTTGGATGATTCAAGACTTCTC
GGTACTGCAAAGTTCTTCCGCCTGATTAATTATCCATTTTACCTTTGTCGTAGATATTAGGTAATCTGTAAGTCAACTC
ATATACAACTCATAATTTAAAATAAAATTATGATCGACACACGTTTACACATAAAATCTGTAAATCAACTCATATACCC
GTTATTCCCACAATCATATGCTTTCTAAAAGCAAAAGTATATGTCAACAATTGGTTATAAATTATTAGAAGTTTTCCAC
TTATGACTTAAGAACTTGTGAAGCAGAAAGTGGCAACACCCCCCACCTCCCCCCCCCCCCCCCACCCCCCAAATTGAGA
AGTCAATTTTATATAATTTAATCAAATAAATAAGTTTATGGTTAAGAGTTTTTTACTCTCTTTATTTTTCTTTTTCTTT
```

BIOINFORMATICS DATA ANALYSIS

- File format used by bioinformatics tools
  - SAM/BAM/CRAM
    - These formats were introduced to standardize how alignments are reported.

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,254}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

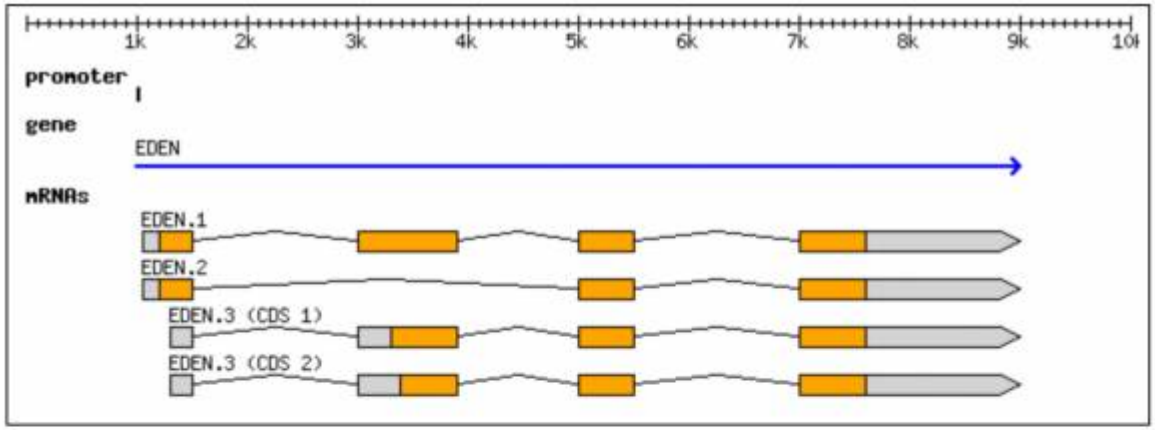
http://samtools.github.io/hts-specs/SAMv1.pdf

BIOINFORMATICS DATA ANALYSIS

- File format used by bioinformatics tools
  - GTF/GFF3
    - tab-delimited text file that holds information any and every feature that can be applied to a nucleic acid or protein sequence. Everything from CDS, microRNAs, binding domains, ORFs, and more can be handled by this format.

BIOINFORMATICS DATA ANALYSIS

- File form
  - GTF/G
    - tab- any and
      evel
      prot cid or
      bind oRNAs,
      this dled by




```
##gff-version 3.2.1
##sequence-region   ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000  1012  .  +  .  Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001
ctg123 . exon            1300  1500  .  +  .  Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .  Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .  Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002
ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002
ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002
ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003
ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003
ctg123 . CDS             7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003
ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003
ctg123 . CDS             5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003
ctg123 . CDS             7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003
```

http://www.sequenceontology.org/gff3.shtml

BIOINFORMATICS DATA ANALYSIS

- File format used by bioinformatics tools
  - VCF/GVCF
    - Variant Calling Format is a tab-delimited text file that is used to describe single nucleotide variants (SNVs) as well as insertions, deletions, and other sequence variations.

```
##fileformat=VCFv4.2
##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        NA00003
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 1/1:43:5:.,.
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   0/0:41:3
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   2/2:35:4
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 0/0:61:2
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       1/1:40:3
```

BIOINFORMATICS DATA ANALYSIS

- File format summary
  - Reference genome, Fasta
  - Gene annotation, GFF/GTF
  - Sequencing data, FASTQ
  - Alignment results, SAM/BAM/CRAM
  - Variations, VCF

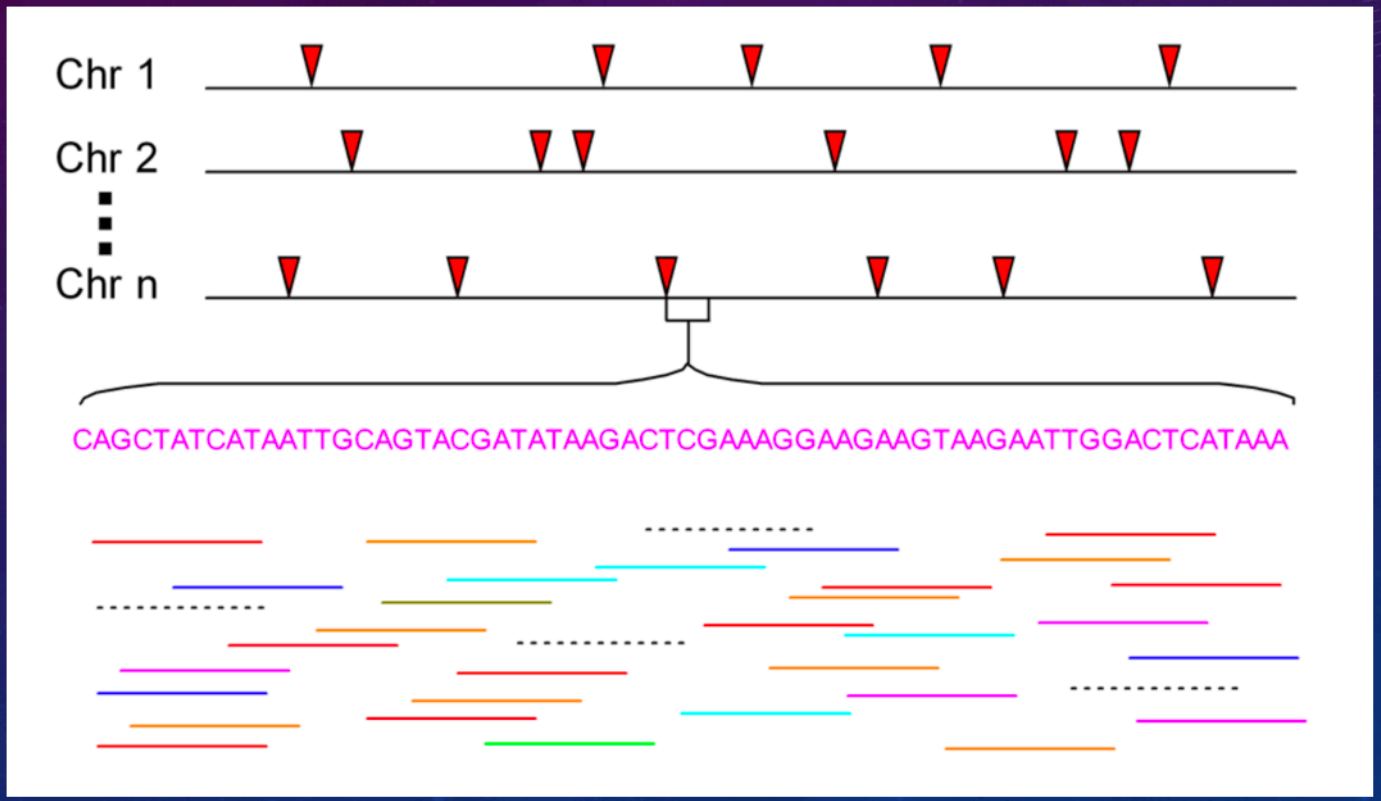
BIOINFORMATICS DATA ANALYSIS

- **What are required for high throughput variation discovery?**
  - Reference genome
    - Reads are mapped to the reference genome, so that we can compare the difference among the samples
  - What if no reference genome available yet?
    - De novo assembly
    - Restriction Associated DNA Sequencing (RADseq)

Restriction Associated DNA Sequencing, RADSeq

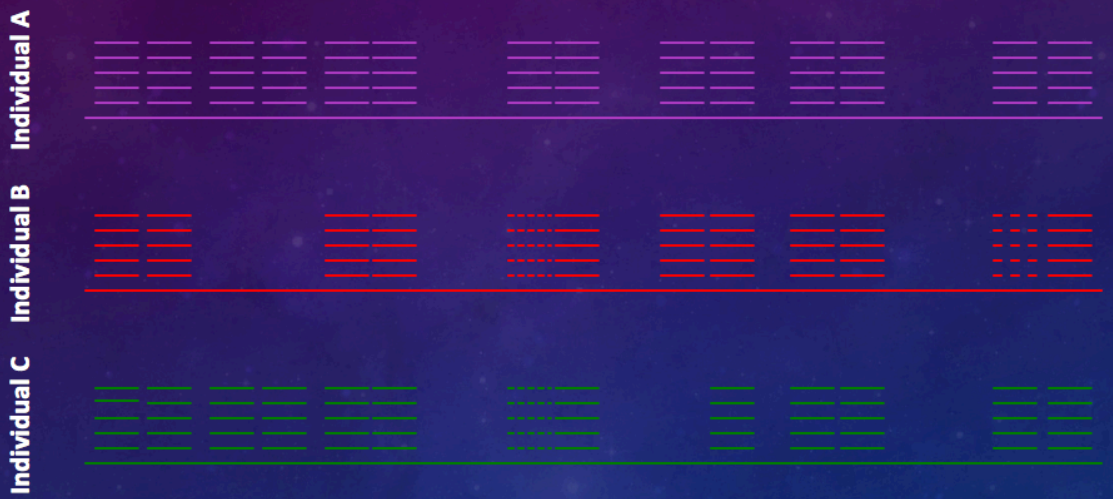
Genotyping By Sequencing, GBS

REDUCED REPRESENTATION GENOME SEQUENCING: RADSEQ/GBS



Lu Fet al. (2013) PLoS Genet 9: e1003215.

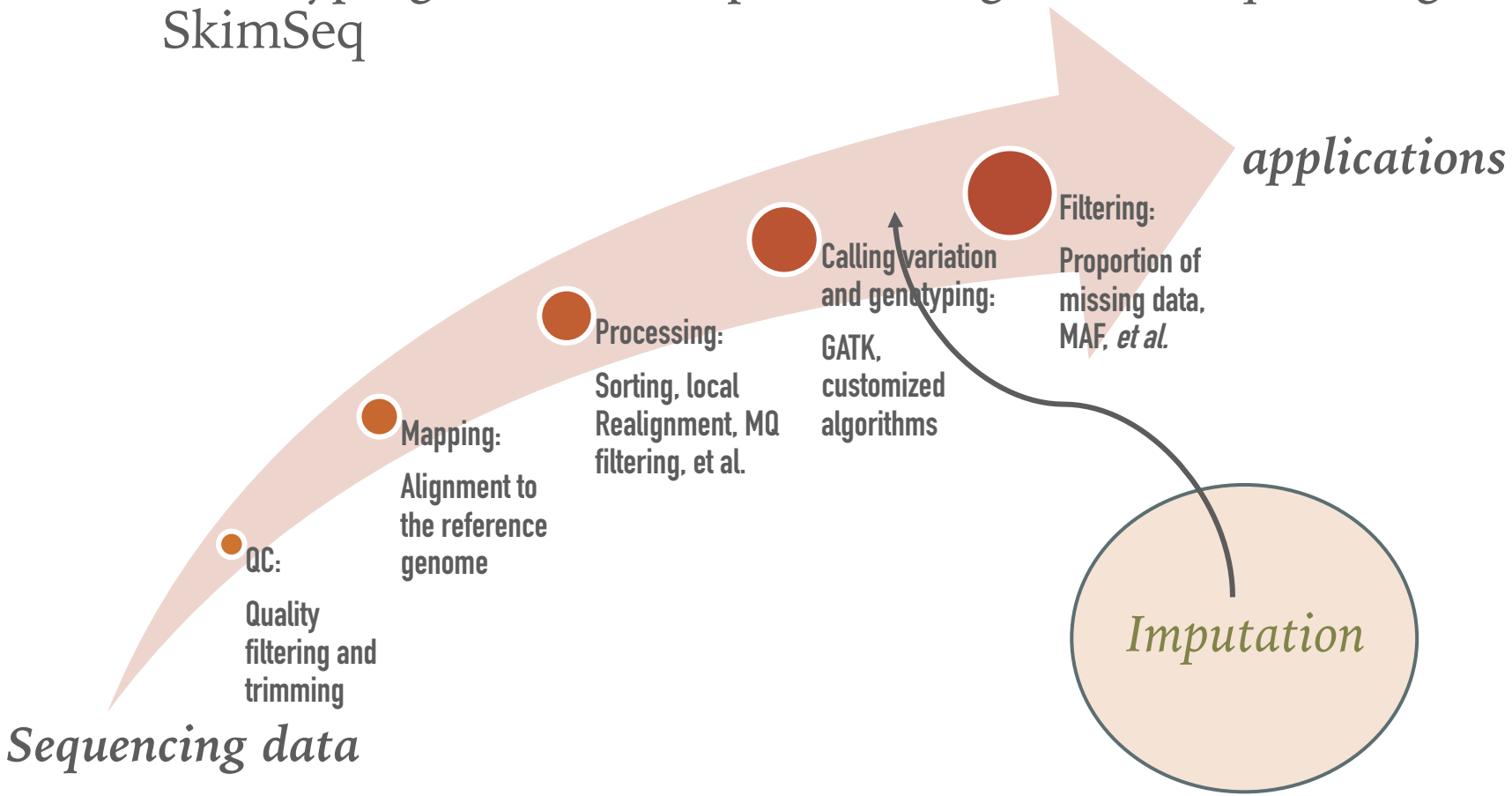
REDUCED REPRESENTATION GENOME SEQUENCING: RADSEQ/GBS



- More even coverage than random shearing
- Requires less data
- Muliplexing, economically feasible

BIOINFORMATICS DATA ANALYSIS

- Requirements
  - If reference genome is available
    - Genotyping with low depth whole genome sequencing, SkimSeq



BIOINFORMATICS DATA ANALYSIS

1. QC: Trimmomtic v0.38

Quality filtering and adapter trimming, ”LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36”

2. Mapping: Bowtie2, v2.3.4.2

Alignment to the reference genome, “--end-to-end –very-sensitive”

3. Processing: SAMTools v1.9 and PICARD v2.18.4

Sorting, local Realignment, MQ filtering (MQ>5), et al.

4. Calling variation and genotyping: GATK v3.5-0-g36282e4, HaplotypeCaller, “-drf BadMate -drf DuplicateRead -U ALLOW_N_CIGAR_READS”

5. Filtering: Only keep variations with <50% of missing and MAF > 5%

6. Imputation with Beagle V4.1

gtgl=$vcf out=$out niterations=10 gprobs=true lowmem=true chrom=$chr

7. Filtered based on the Genotype Probability (GP >= 0.9)

8. Additional filtering (optional):

Proportion of missing data, MAF, *et al.*

BIOINFORMATICS DATA ANALYSIS

- Quality control, trimming
  - Low quality bases
  - Adapter sequences left in the reads
  - Remove reads that are too short
  - Pair-end information
- Tools
  - fastX-toolkit
  - Trimmomatic
  - …

Trimmomatic

http://www.usadellab.org/cms/?page=trimmomatic

Paired End:

Command to execute Trimmomatic

```
java -jar trimmomatic-0.35.jar PE -phred33 input_forward.fq.gz
input_reverse.fq.gz output_forward_paired.fq.gz
output_forward_unpaired.fq.gz output_reverse_paired.fq.gz
output_reverse_unpaired.fq.gz ILLUMINACLIP:TruSeq3-PE.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

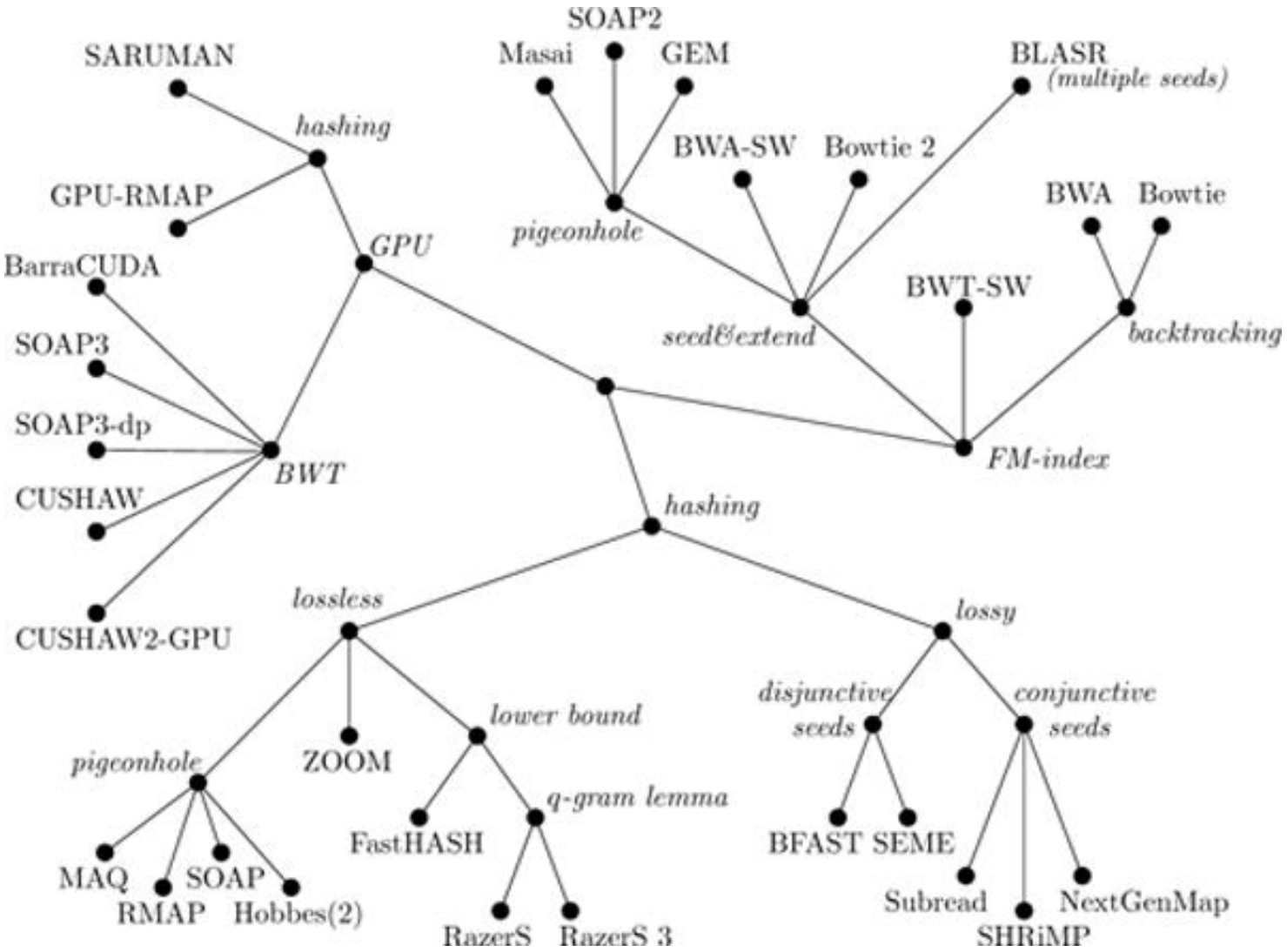
This will perform the following:

- Remove adapters (ILLUMINACLIP:TruSeq3-PE.fa:2:30:10)
- Remove leading low quality or N bases (below quality 3) (LEADING:3)
- Remove trailing low quality or N bases (below quality 3) (TRAILING:3)
- Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15 (SLIDINGWINDOW:4:15)
- Drop reads below the 36 bases long (MINLEN:36)

BIOINFORMATICS DATA ANALYSIS

- Short read alignment
  - Short read alignment is the process of figuring out where in the genome a sequence is from. This is tricky for several reasons:
  - The reference genome is really big. Searching big things is harder than searching small things.
  - You aren't always looking for *exact* matches in the reference genome–or, at least, probably not.
- Alignemnt tools:
  - BWA, BOWTIE2, et al.

BIOINFORMATICS DATA ANALYSIS

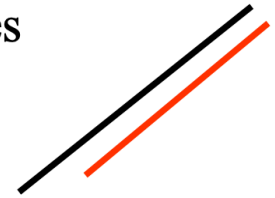


Proc IEEE Inst Electr Electron Eng. 2017 Mar; 105(3): 436–458.

- Sequence alignment

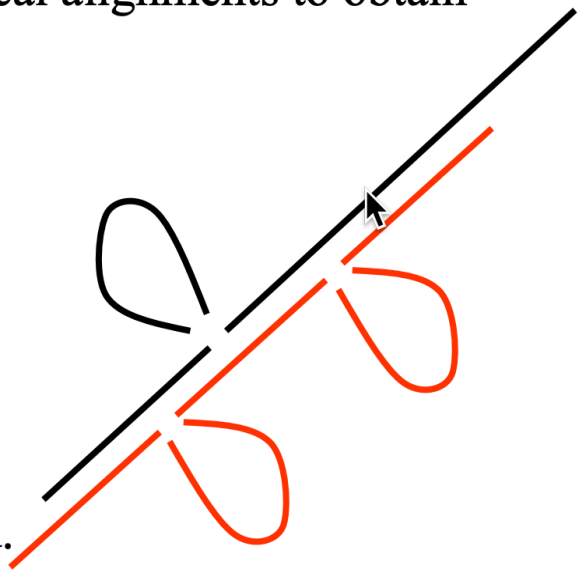
Global and local approaches to aligning sequences

GLOBAL: Attempt to "match" and assess similarity between two entire sequences



LOCAL: Find subsequences of high similarity

… and then possibly "stick" (chain, net, thread) together local alignments to obtain an overall comparison of the original sequences.



The second approach is more meaningful

(especially for long sequences, of different lengths, like whole genomes)

Two protein or DNA sequences are unlikely to present a straightforward overall "match", even if they are closely related.

BIOINFORMATICS DATA ANALYSIS

- Bowtie2, http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
  - ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.
  - It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes.
  - Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB.
  - Bowtie 2 supports gapped, local, and paired-end alignment modes.

BIOINFORMATICS DATA ANALYSIS

- Bowtie2, http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

```
Bowtie 2 version 2.3.4.2 by Ben Langmead (langmea@cs.jhu.edu, www.cs.jhu.edu/~langmea)
Usage:
  bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r> | --interleaved <i>} [-S <sam>]

  <bt2-idx>  Index filename prefix (minus trailing .X.bt2).
             NOTE: Bowtie 1 and Bowtie 2 indexes are not compatible.
  <m1>       Files with #1 mates, paired with files in <m2>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <m2>       Files with #2 mates, paired with files in <m1>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <r>        Files with unpaired reads.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <i>        Files with interleaved paired-end FASTQ reads
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <sam>      File for SAM output (default: stdout)

  <m1>, <m2>, <r> can be comma-separated lists (no whitespace) and can be
  specified many times.  E.g. '-U file1.fq,file2.fq -U file3.fq'.
```

BIOINFORMATICS DATA ANALYSIS

- Bowtie2, http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

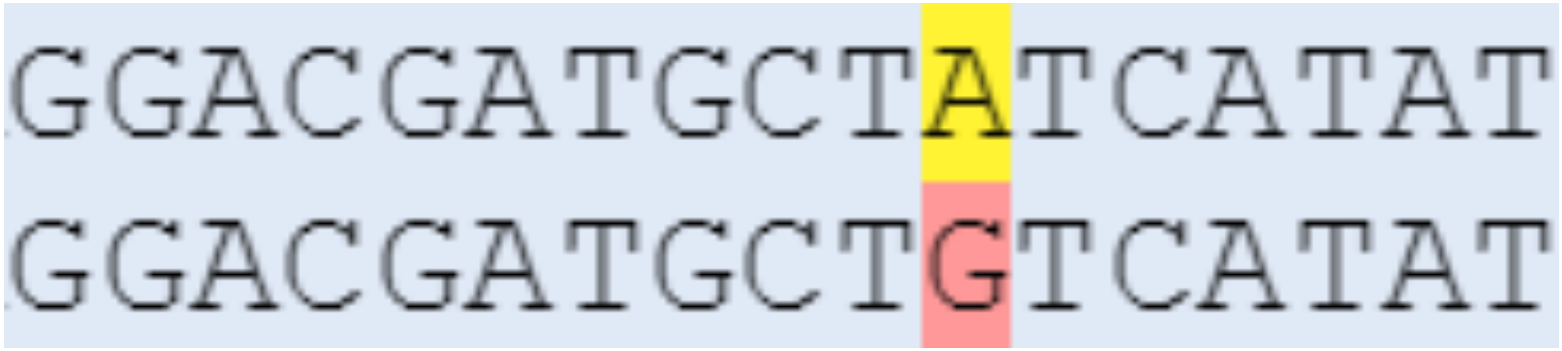
```
Presets:                 Same as:
 For --end-to-end:
  --very-fast            -D 5 -R 1 -N 0 -L 22 -i S,0,2.50
  --fast                 -D 10 -R 2 -N 0 -L 22 -i S,0,2.50
  --sensitive            -D 15 -R 2 -N 0 -L 22 -i S,1,1.15 (default)
  --very-sensitive       -D 20 -R 3 -N 0 -L 20 -i S,1,0.50

 For --local:
  --very-fast-local      -D 5 -R 1 -N 0 -L 25 -i S,1,2.00
  --fast-local           -D 10 -R 2 -N 0 -L 22 -i S,1,1.75
  --sensitive-local      -D 15 -R 2 -N 0 -L 20 -i S,1,0.75 (default)
  --very-sensitive-local -D 20 -R 3 -N 0 -L 20 -i S,1,0.50
```

BIOINFORMATICS DATA ANALYSIS

- Alignment file processing
  - Convert to bam (binary SAM)
  - Sort the BAM files
  - Add the RG (read group) names to BAM if necessary
  - Mark duplicates
  - Local realignment
  - Filtering with MQ (mapping quality)
  - Visualization using the Integrative Genomics Viewer (IGV)

BIOINFORMATICS DATA ANALYSIS

- Variation calling
  - Identifying single nucleotide polymorphisms (SNPs) and small insertions and deletion (indels) from next generation sequencing data.



BIOINFORMATICS DATA ANALYSIS

- Variation calling
  - Identifying single nucleotide polymorphisms (SNPs) and small insertions and deletion (indels) from next generation sequencing data.

```
12001     12011     12021     12031     12041     12051     12061     12071     12081     12091     12101     12111     12121
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
CTGTCCCTGCAGGTCYCTTTCTGCCTGGCTGCTGTCCAGCCCCTCTGTCCCRAACCTGTTGGTTGTTTATCCTGGAGAAAAGGAGTSTCAAAGGAGACAAGGCAGTGTCAGGGTGTAKGTTGGGCTTGATGGTG
CTGTCCC         ctttctgcctggctgctgtccagcccctctgtcccaaacctgttggttgtttatcctggagaaaaggagtgtcaaaggagacaaggcagtgtcagggtgtatgttgggcttgatggtg
CTGTCCCTGCAGGTCTCTTTCTGCCTGGCTGCTGTCCAGCCCCTCTGTCCCAAACCTGTTGGTTGTTTATCCTGGAGAAAAGGAGTCTCAAAGGAGACAAGGCAGTGTC                 TGATGGTG
ctgtccctgcaggtccctttctgcctggctgctgtccagcccctctgtcccaaacctgtcggttgtttatcctggagaaaaggagtctcaaaggagacaaggcagtgtcaggg
                    CTGCCTGGCTGCTGTCCAGCCCCTCTGTCCCAAACCTGTTGGTTGTTTATCCTGGAGAAAAGGAGTGTCAAAGGAGACAAGGCAGTGTCAGGGTGTATGTTGGGCTTGATGGTG
                    CTGCCTGGCTGCTGTCCAGCCCCTCTGTCCCGAACCTGTTGGTTGTTTATCCTGGAGAAAAGGAGTCTCAAAGGAGACAAGGCAGTGTCAGGGTGTAGGTTGGGCTTGATGGTG
                              GCTGTCCAGCCCCTCTGTCCCGAACCTGTTGGTTGTTTATCCTGGAGAAAAGGAGTCTCAAAGGAGACAAGGCAGTGTCAGGGTGTAGGTTGGGCTTGATGGTG
                                        ccctctgtcccgaacctgttggttgtttatcctggagaaaaggagtctcaaaggagacaaggcagtgtcagggtgtaggttgggcttgatggtg
                                                                                                          GTCAGGGTGTAGGTTGGGCTTGATGGTG
```

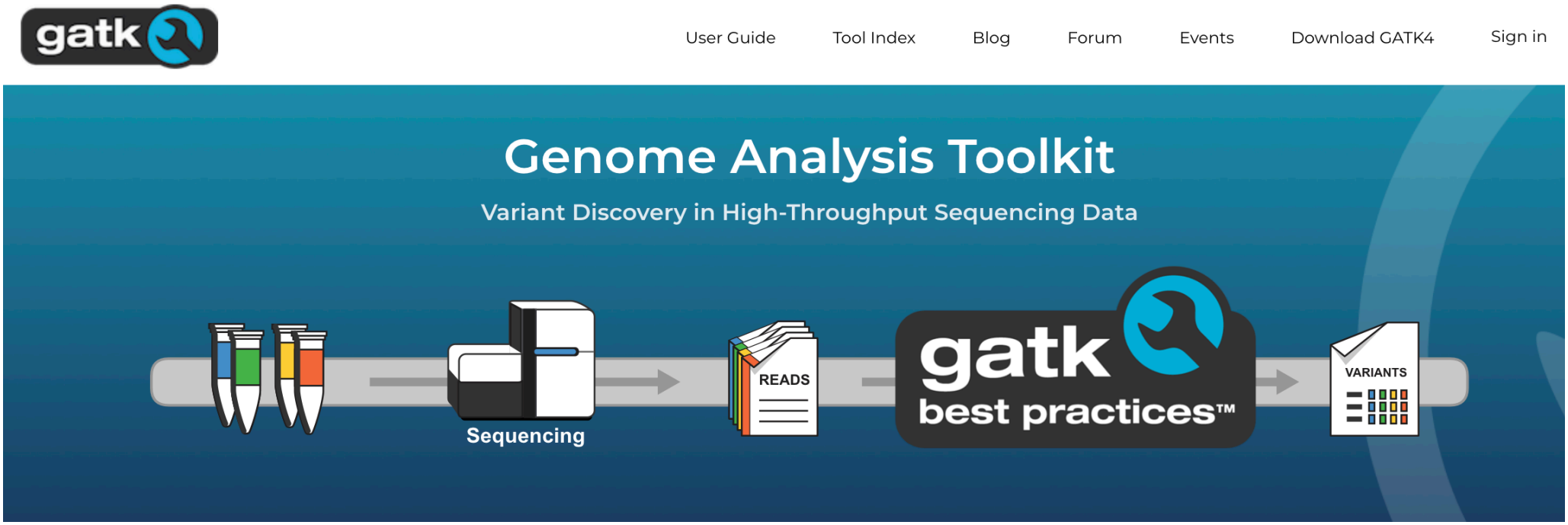
The key challenge with NGS data is distinguishing which mismatches represent real mutations and which are just noise?

BIOINFORMATICS DATA ANALYSIS

- Variation calling
  - Things to consider:
    - How good can the alignment be? Mapping quality (MQ)
      - # of mismatches; the uniqueness of mapping
    - How confidence the variations were called? Vairation quality score.
      - # reads support the calling; Base quality

BIOINFORMATICS DATA ANALYSIS

- Variation calling
  - Genome Analysis Toolkit (GATK)
    - Developed by the Broad Institute
    - Industry Standard for identifying SNPs and indels in germline DNA and RNAseq data
    - In addition to the variant callers themselves, GATK also includes many utilities to perform related tasks such as processing and quality control of high-throughput sequencing data.
- GATK was designed to maximize sensitivity in order to **minimize false negatives**, i.e. failing to identify real variants



Any questions?

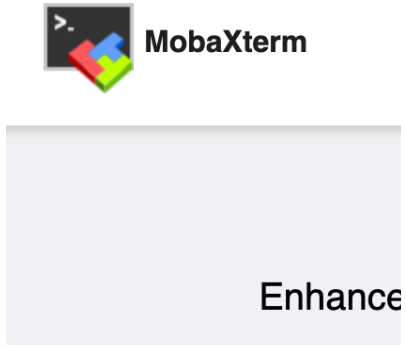
Let’s give it a try!

DEMO OF GBS DATA ANALYSIS USING STACKS ON HPRC

Login to ada:

ssh your-tamu-netid@ada.tamu.edu

Or use MobaXterm



Instructions:

https://github.com/swang8/htg/

Run jobs on working nodes interactively:

https://portal.hprc.tamu.edu/

Blank

DDRAD/GBS DATA ANALYSIS

blank

BIOINFORMATICS DATA ANALYSIS

- **Mainly two types of pipeline**
  - Alignment-based
    - Treat GBS/RADseq data as usual NGS data
  - Clustering-based
    - Cluster the reads that are from the same loci, then discover variations within clusters (i.e. multiple sequence alignment)

ALIGNMENT APPROACH

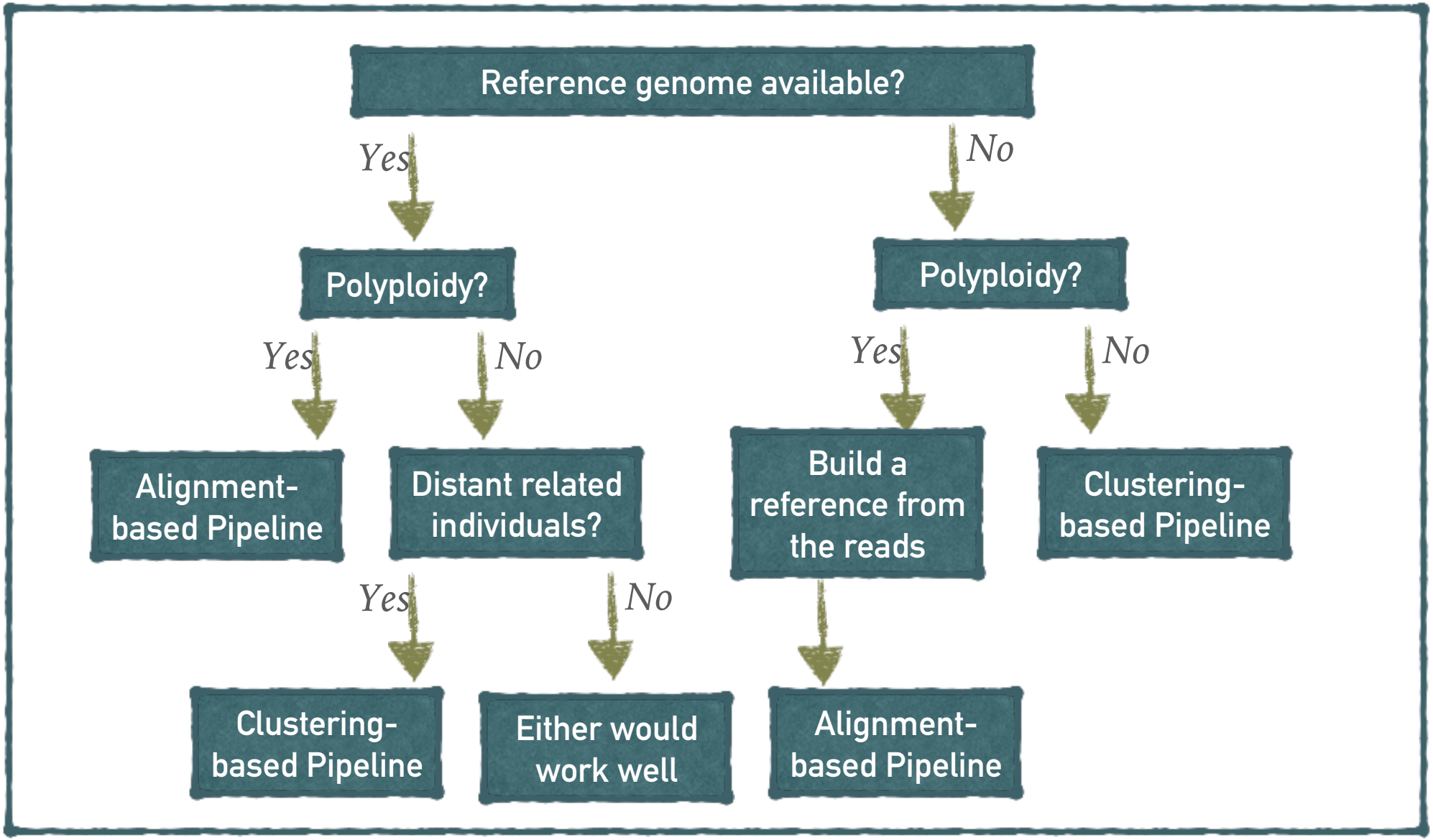
CLUSTERING BASED APPROACH

Clustering based on similarity

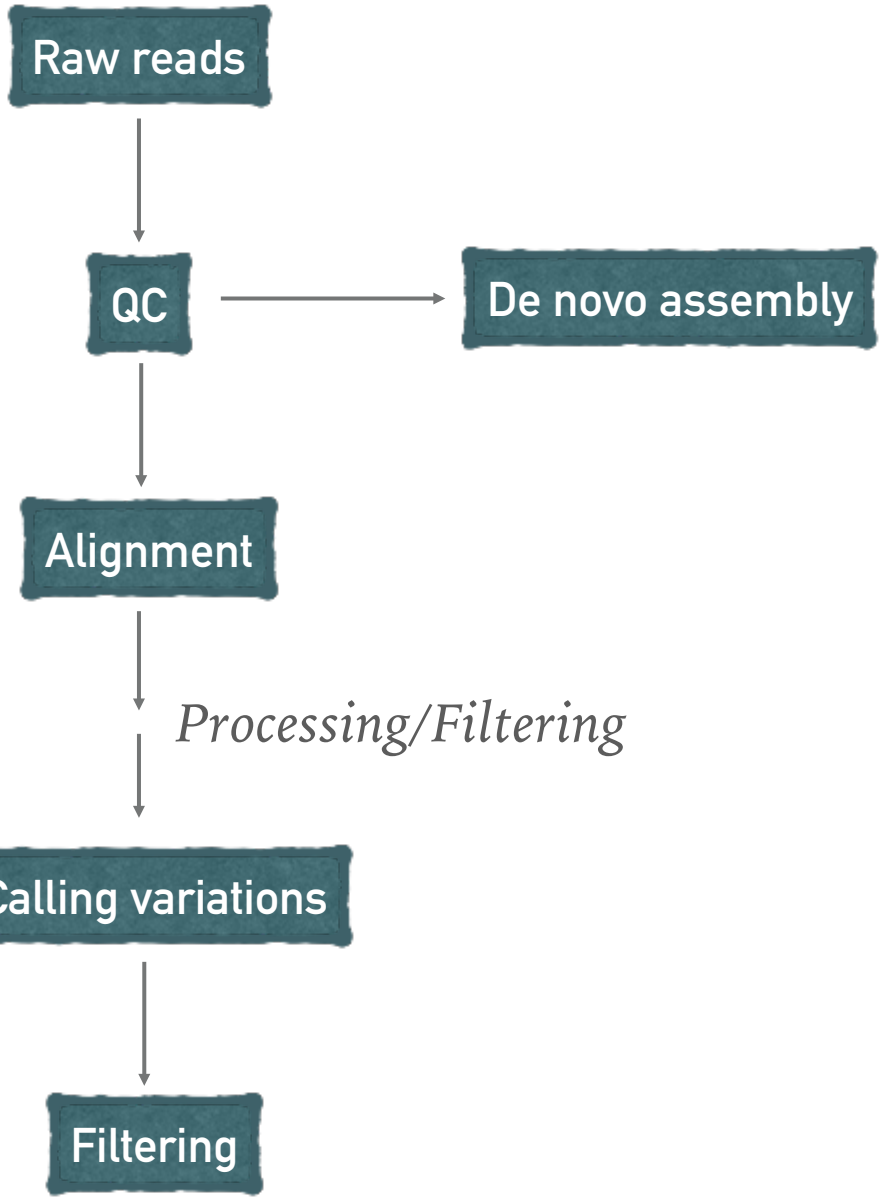
RADSEQ/GBS DATA ANALYSIS

- Features of the two types of pipelines
  - Alignment-based
    - require reference
    - compute intensive
    - More accurate
    - Imputation might be easier
  - Clustering-based
    - reference not required
    - Reduce the computation cost by clustering
    - Might lead to large false positive, or removing too many variation with stringent filtering criteria


Alignment
Clustering


Reference genome available?
Yes
No
Polyploidy?
Polyploidy?
Yes
No
Yes
No
Alignment-based Pipeline
Distant related individuals?
Build a reference from the reads
Clustering-based Pipeline
Yes
No
Clustering-based Pipeline
Either would work well
Alignment-based Pipeline

General NGS analysis workflow for variation discovery

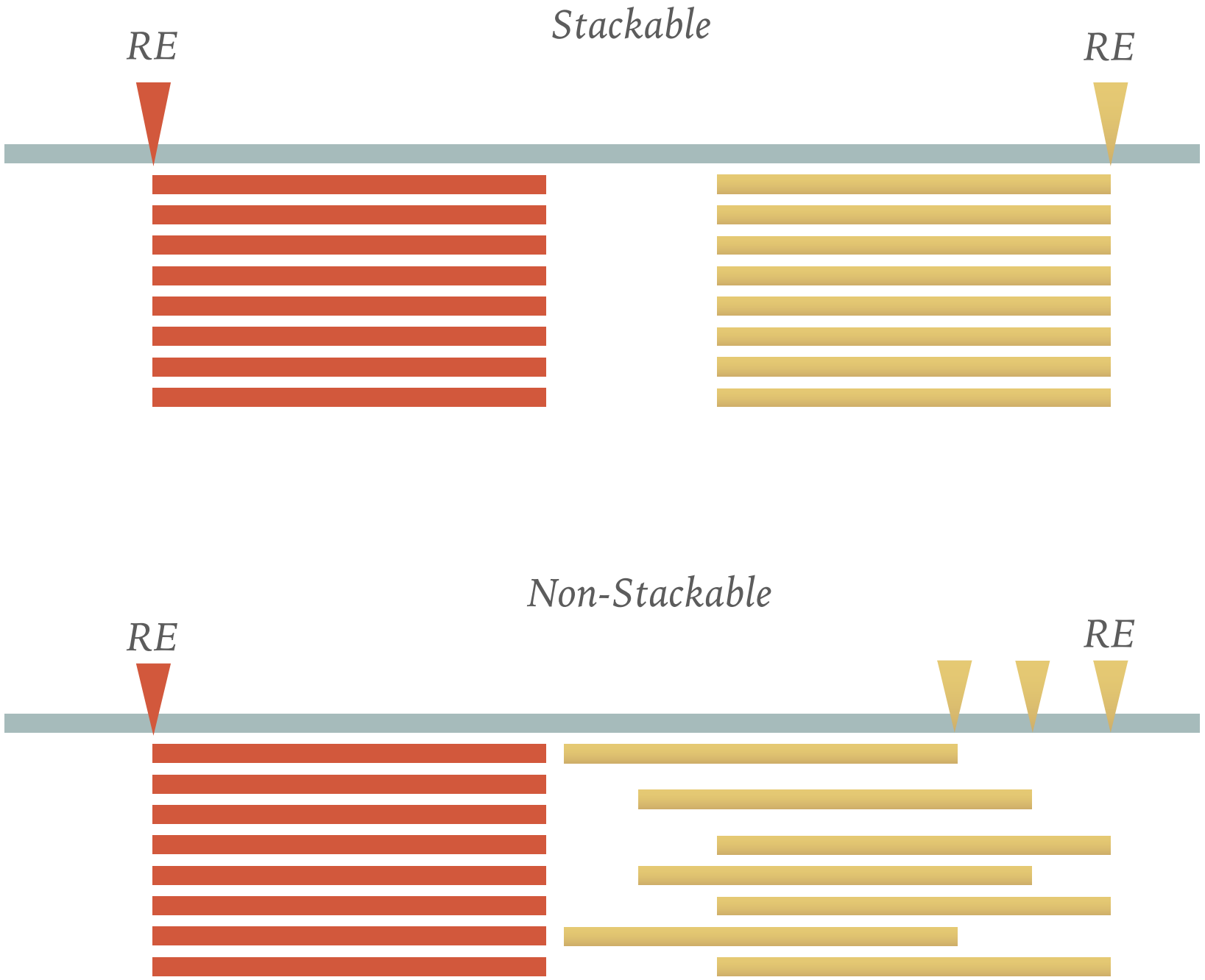


RADSEQ/GBS DATA ANALYSIS PIPELINE

Pipeline/Progarm	Alignment	Clustering	Comment
Stacks	Y	Y	
TASSEL-GBS	Y	Y	Trim reads
UNEAK	N	Y	Trim reads
PyRAD	N	Y	
dDocent	Y	N	
AftrRAD	N	Y	

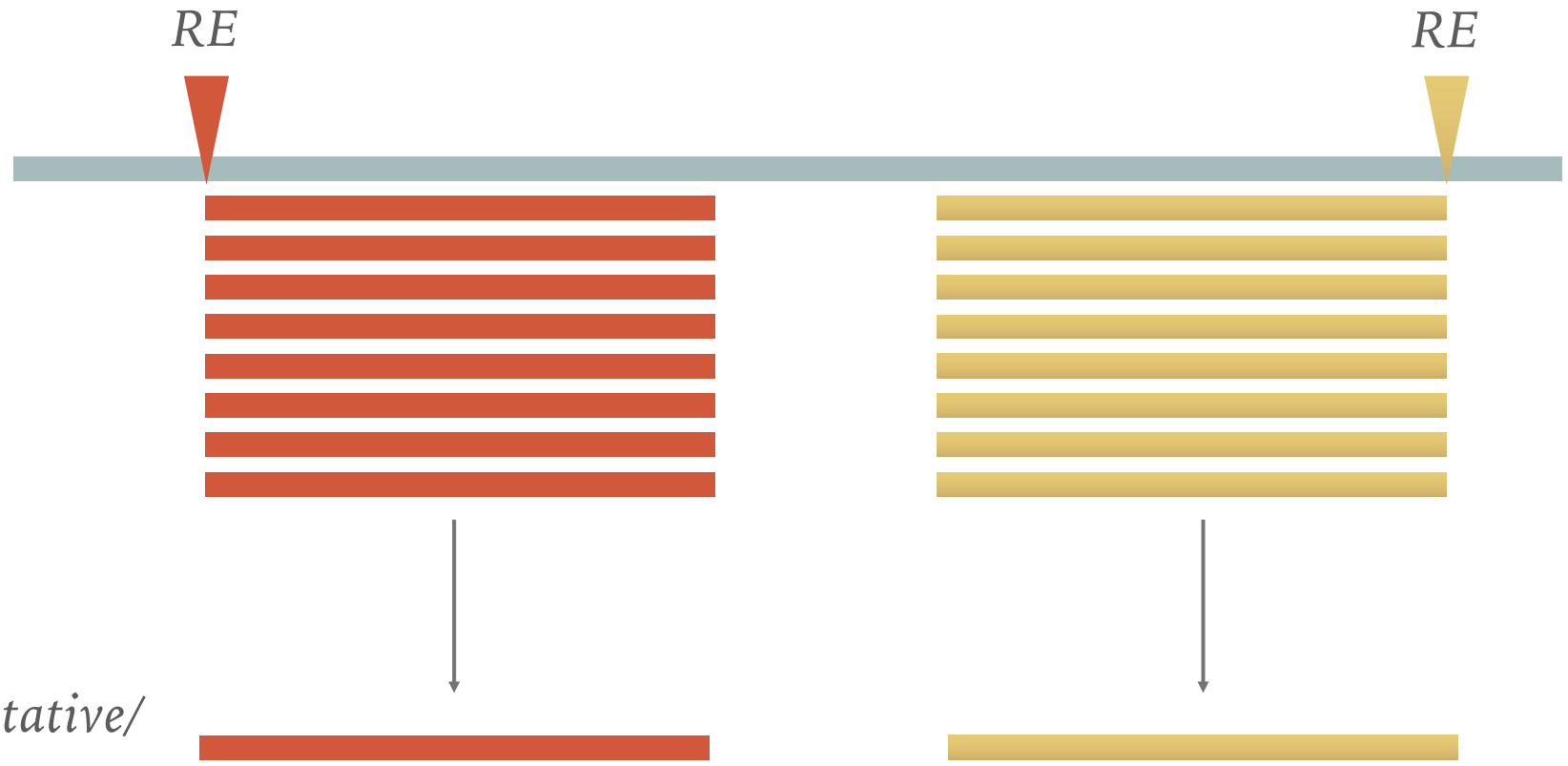
Stacks

- Designed to work with short read (max 1024bp)
- Uniform length of reads
  - Ideal for Illumina
  - For Ion Torrent platform, reads would need to be truncated to a particular length
- Deal with most of the RADseq/GSB protocols
- *Stacks* is designed to process data that *stacks* together:
  - In the case of double-digest RAD, both the single-end and paired-end read are anchored by a restriction enzyme and can be assembled as independent loci;
  - In cases such as with the RAD protocol, where the molecules are sheared and the paired-end therefore does not stack-up, cannot be directly used.


Stackable
RE
RE
Non-Stackable
RE
RE

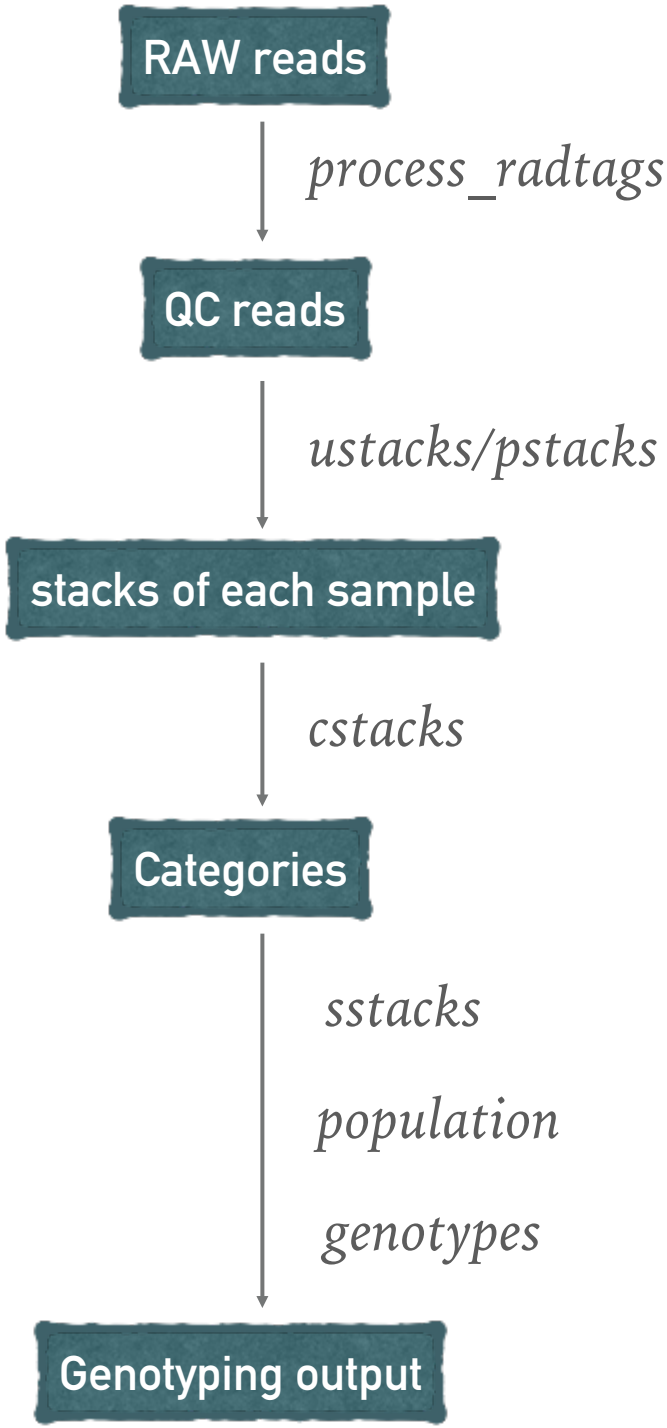
WHY CLUSTERING OR STACKING REDUCE THE COST OF COMPUTATION

TASSEL-GBS:

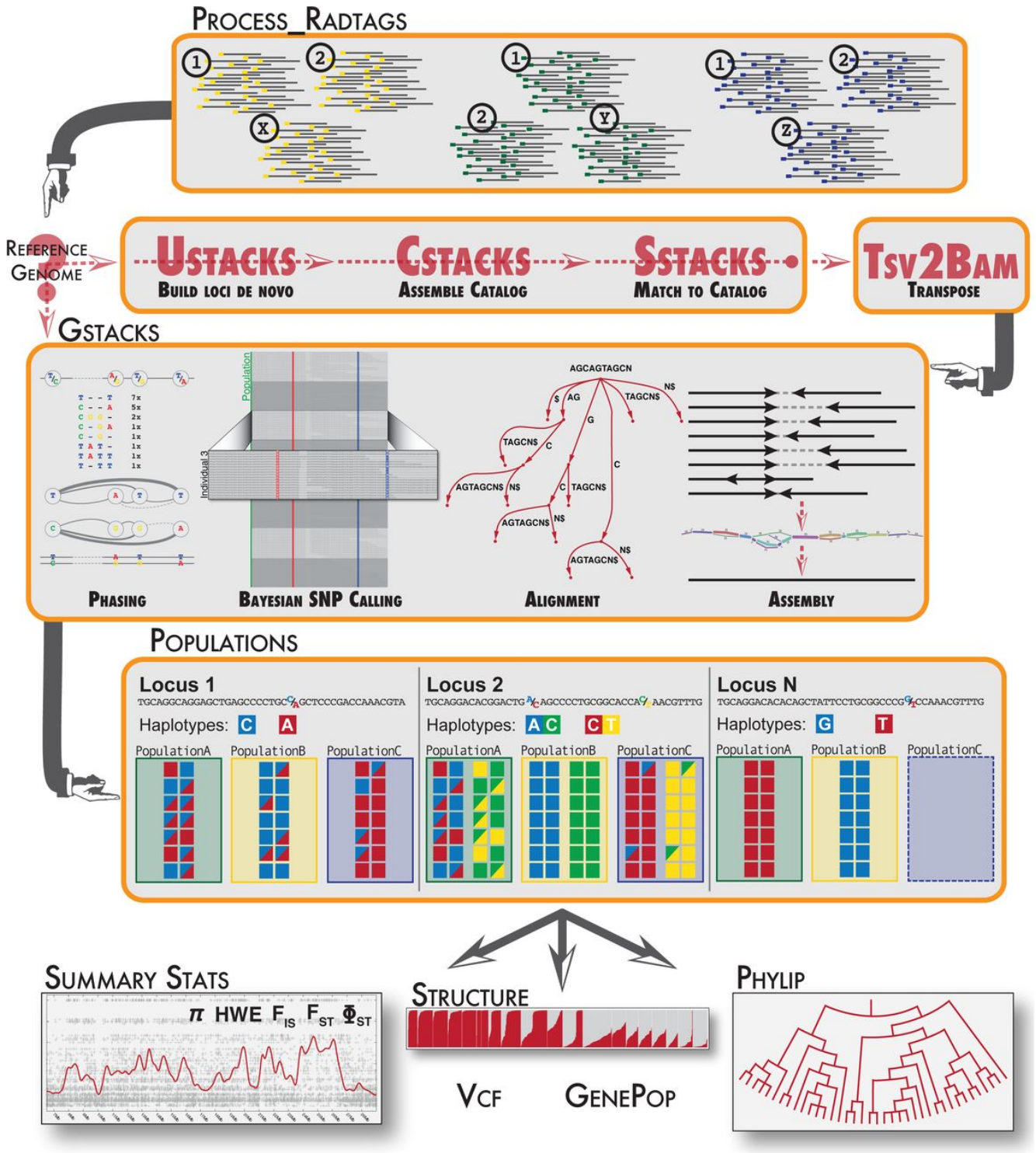


Instead of trying to map eight reads separately, we may just take one as representative.

Stacks V1 workflow



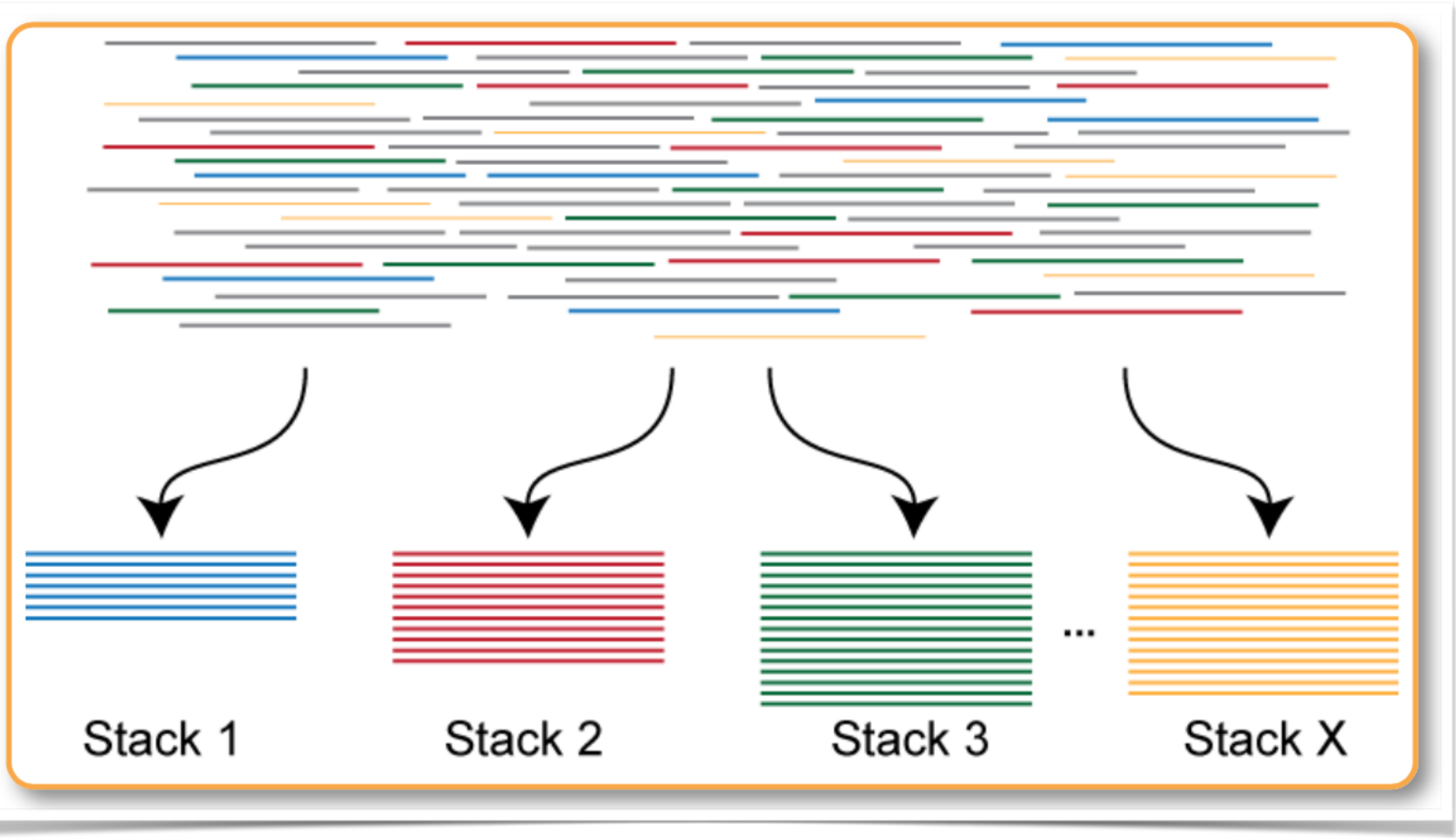
Stacks 2 workflow



MAJOR PARAMETERS

Parameter Description	`denovo_map.pl` Parameter	Pipeline component	Component Parameter	Default Value
Minimum stack depth / minimum depth of coverage	`-m`	`ustacks`	`-m`	3
Distance allowed between stacks	`-M`	`ustacks`	`-M`	2
Distance allowed between catalog loci	`-n`	`cstacks`	`-n`	0

1. MINIMUM STACK DEPTH *-m 3*



1. MINIMUM STACK DEPTH *-m 3*

1 If set to a value of 3 then three or more identical reads must be found to consider those reads a stack. If a stack is formed with only two reads, then those reads are set aside (**secondary reads**) and a stack is not constructed.

2 If this parameter is set too low, then reads with convergent sequencing errors are likely to be erroneously labled as stacks.

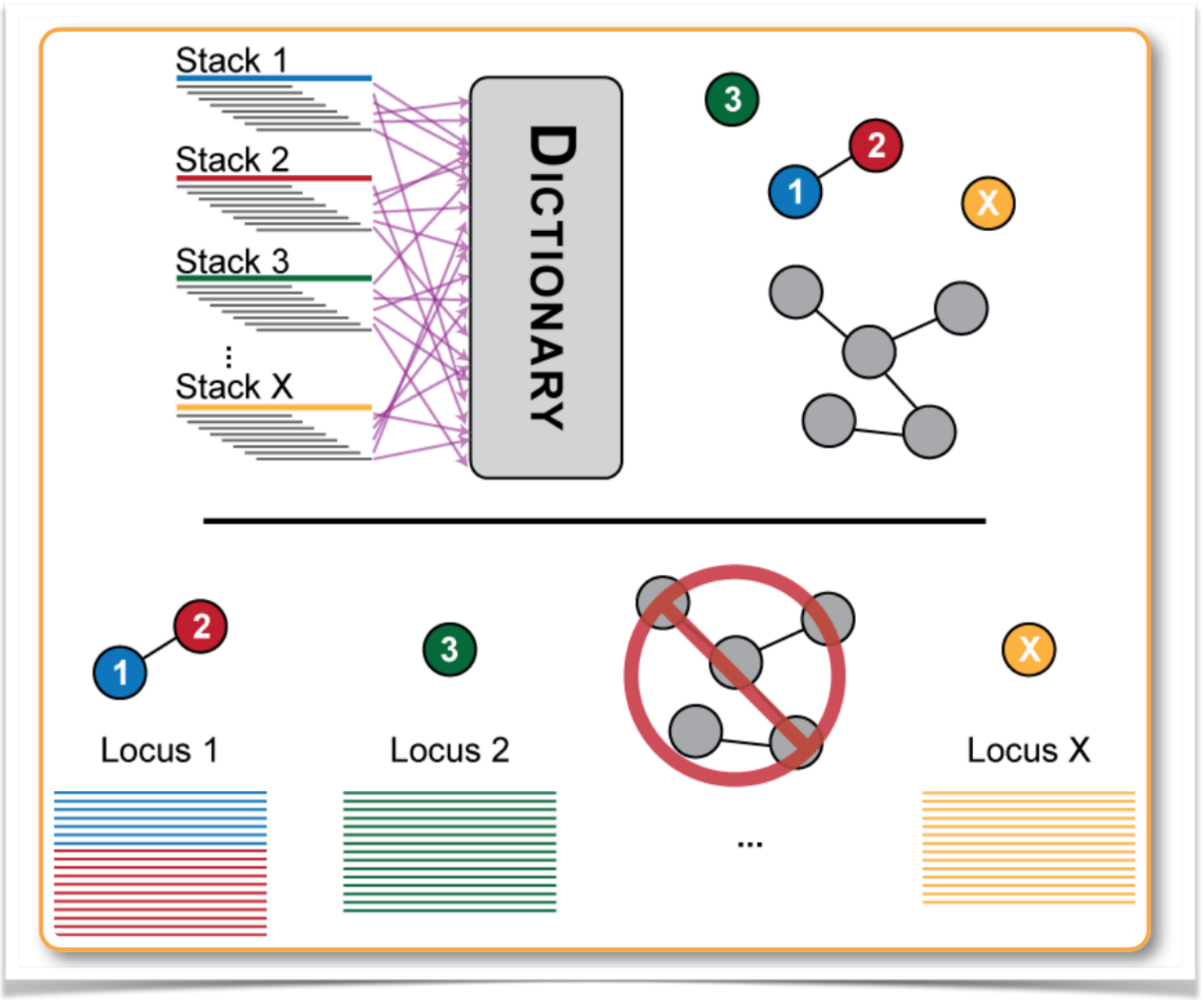
3 If this parameter too high, then true alleles will not be recorded and will drop out of the analysis.

4 If you have low sequencing depth for your samples, you will have to set this parameter to a relatively low value. Conversely, if you have very high sequencing coverage, you will want to increase this parameter.

5 If you have a high error rate in your sequencing lane, then you are likely to see convergent sequencing or PCR errors (errors that occur independently at the same nucleotide position in the same read) and should increase the minimum stack depth.

2. DISTANCE ALLOWED BETWEEN STACKS

-M 2



2. DISTANCE ALLOWED BETWEEN STACKS *-M 2*

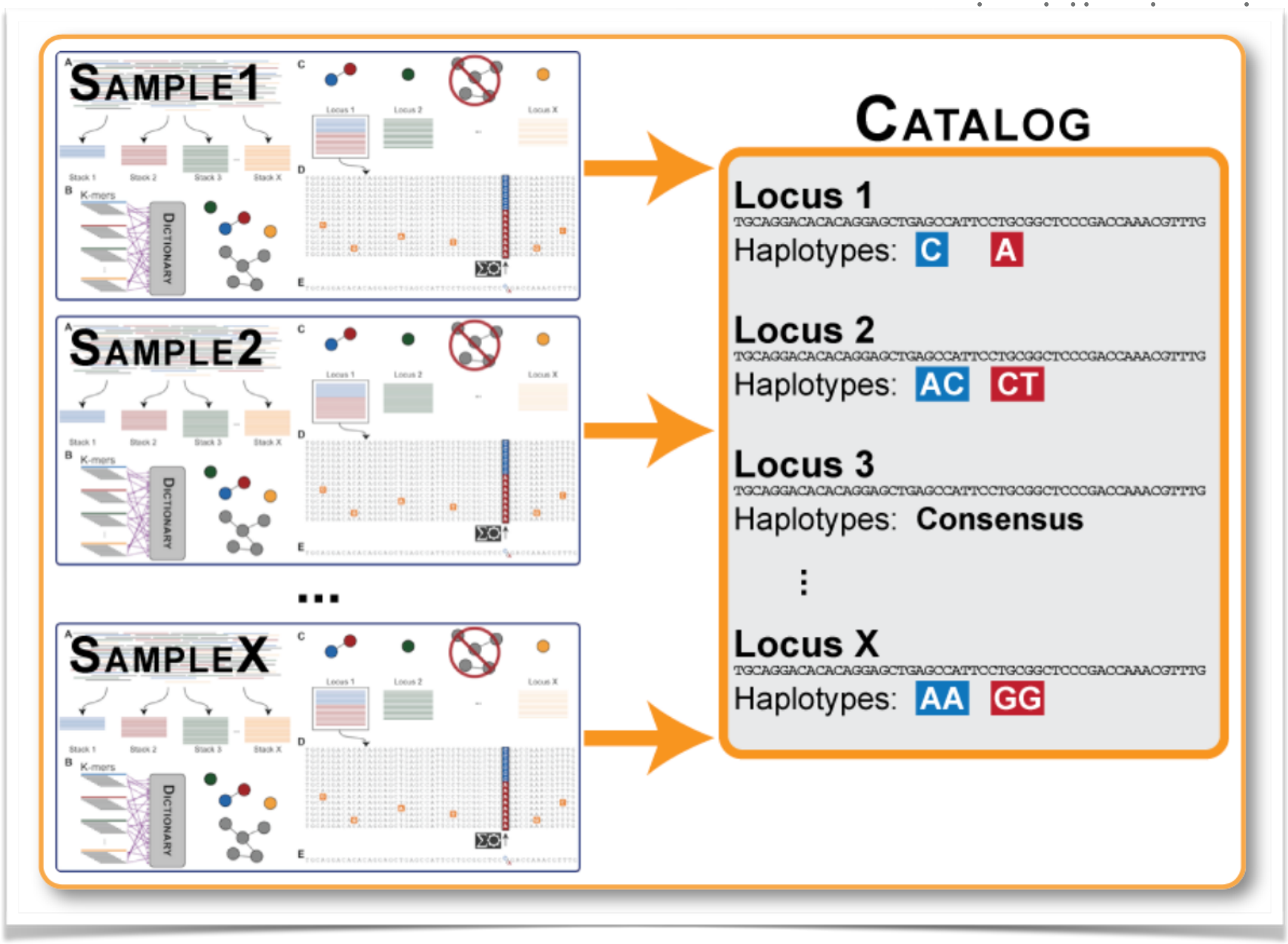
1 If you set this parameter too low, then some loci will fail to be reconstructed. This means the SNPs contained in that locus will not be identified and this locus will appear as two loci to the remainder of the pipeline.

2 Setting this parameter too high will allow repetitive sequence to chain together in to large, nonsensical loci. For example, if stack A is one nucleotide apart from stack B, which is one nucleotide apart from stack C, which is one nucleotide apart from stack D, then A, B, C, and D will be merged into a locus despite A and D being four nucleotides apart. These loci are not useful to the pipeline and at several points the pipeline will try to detect these and set them aside.

3 You will want to experiment with several different values of this parameter to see how many polymorphic loci you can construct.

3. DISTANCE BETWEEN CATALOG LOCI

-n 0



3. DISTANCE BETWEEN CATALOG LOCI *-n 0*

1 If you set this parameter too low, then some loci will fail to be reconstructed. This means the SNPs contained in that locus will not be identified and this locus will appear as two loci to the remainder of the pipeline.

2 Setting this parameter too high will allow repetitive sequence to chain together in to large, nonsensical loci. For example, if stack A is one nucleotide apart from stack B, which is one nucleotide apart from stack C, which is one nucleotide apart from stack D, then A, B, C, and D will be merged into a locus despite A and D being four nucleotides apart. These loci are not useful to the pipeline and at several points the pipeline will try to detect these and set them aside.

3 You will want to experiment with several different values of this parameter to see how many polymorphic loci you can construct.

OPTIMIZE THE PARAMETERS

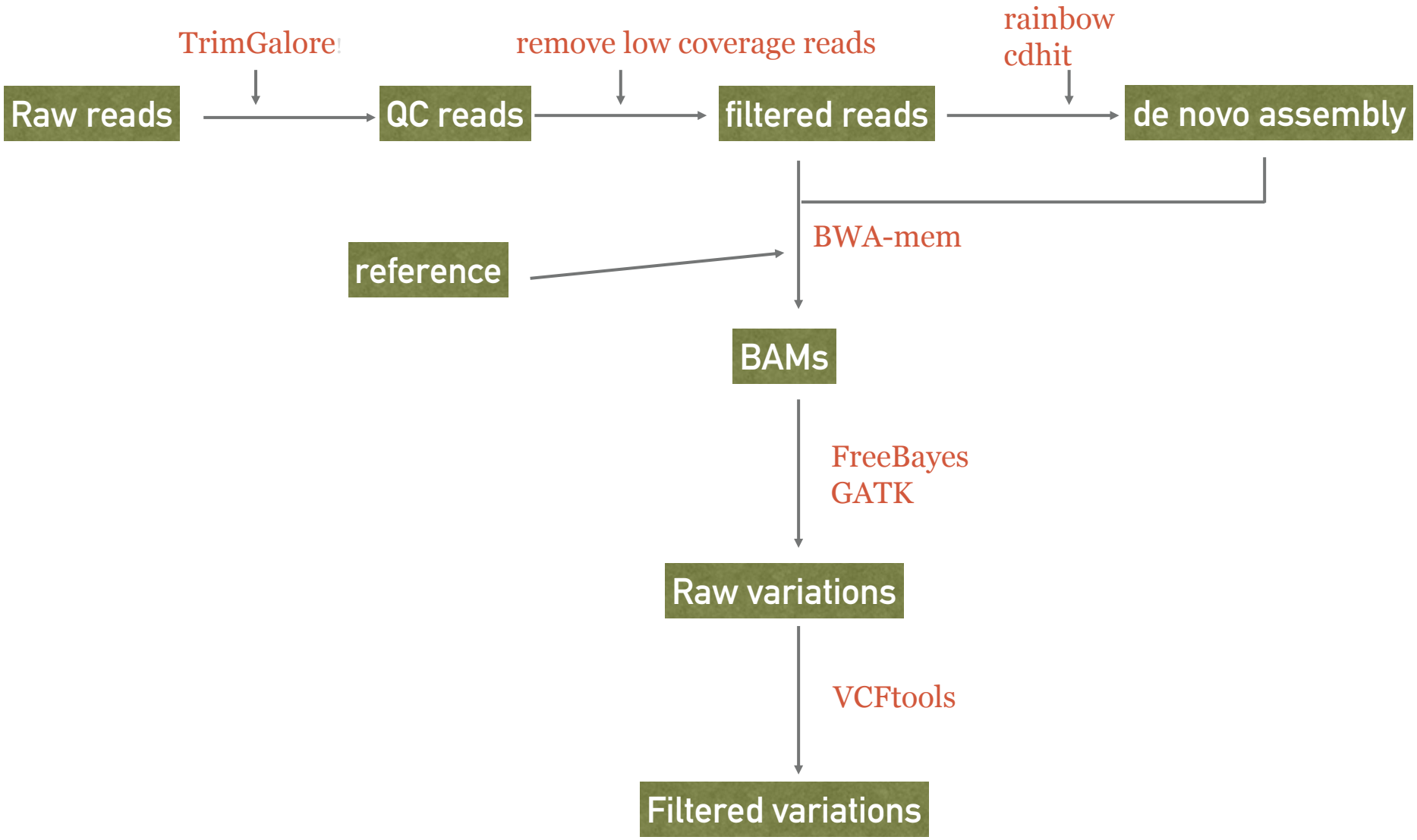
- How to optimize the parameters for my project?
  - Simulation
    - With reference genome available, simulate RADseq/GBS reads from the reference genome with predefined SNPs;
    - Call SNPs with different set of parameters, pick the one with the lowest FP and/ high TP.
  - Generate SNPs for multiple sets of parameters, then check the SNP accuracy

DDOCENT PIPELINE

- dDocent relies almost entirely on third party software to complete every step of the analysis pipeline..

FreeBayes	https://github.com/ekg/freebayes
STACKS	http://creskolab.uoregon.edu/stacks/
PEAR	http://sco.h-its.org/exelixis/web/software/pear/
Trimmomatic	http://www.usadellab.org/cms/?page=trimmomatic
Mawk	http://invisible-island.net/mawk/
BWA	http://bio-bwa.sourceforge.net
SAMtools	http://samtools.sourceforge.net
VCFtools v.1.11**	http://vcftools.sourceforge.net/index.html
rainbow	http://sourceforge.net/projects/bio-rainbow/files/
seqtk	https://github.com/lh3/seqtk
CD-HIT	http://weizhong-lab.ucsd.edu/cd-hit/

DDOCENT PIPELINE



DDOCENT PIPELINE