Bacterial Genomics Workshop

Evan Snitkin, Ali Pirani,
Stephanie Thiede, Arianna Miles-Jay,
Kyle Gontjes and Emily Benedict

April 21st – April 23rd 2021

Zoom logistics

- Sessions will be recorded
- Will periodically ask for green check to indicate that we are on the same page
- If you get stuck, put up a red X and we will place you in breakout room with helper
- Please don’t be shy about raising your hand to ask questions (or put them in the chat)

Goals of workshop

- Get an overview of steps in microbial genomics pipeline
- Get exposure to common file formats and terminology in genomics
- Get hands on experience with a set of tools that could compose a genomics pipeline
- Get experience working in a high-performance computing environment

Logistics of the workshop

- We will follow the course website closely (for the most part)

https://github.com/alipirani88/Comparative_Genomics

- The website is extremely rich in detail, beyond what will be covered in the workshop

Format of sessions

- There will be six sessions
  - A Unix/R review and environment setup
  - Four sessions on different aspects of the genomics pipeline
  - An independent work session where you apply all the skills you learned during the week to analyze a microbial genomics dataset from start to finish!

- Each session will work through published datasets (mostly from our lab)

Moving files to/from remote server

- https://cyberduck.io/download/

How do a remote server and compute cluster work?

Why Unix?

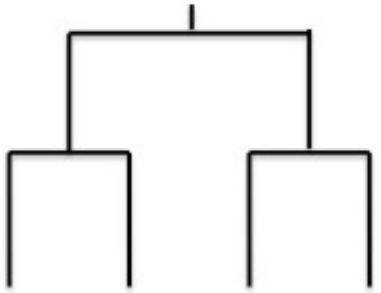
- Most bioinformatics research is performed in a Unix environment
- Allows for easier interactions with text files
- The power of pipes
- Easy to automate repetitive tasks
- Facilitates interfacing with high-performance compute systems

Unix review

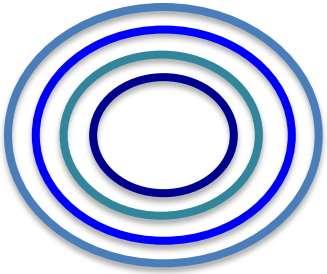
- Moving around
  - ls, pwd, cd
- Directory management
  - mkdir, rmdir
- File management
  - cp, mv, rm
- File viewing/editing
  - less, nano
- Searching files
  - grep, cut, wc, sort, uniq

So you want to sequence some bacteria?

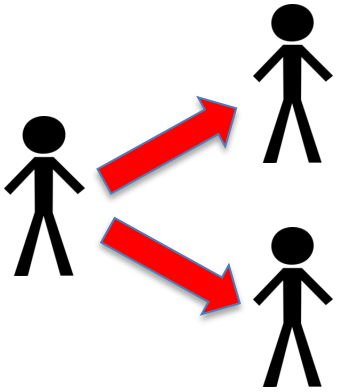
- Microbial phylogenetics



- Comparative genomics

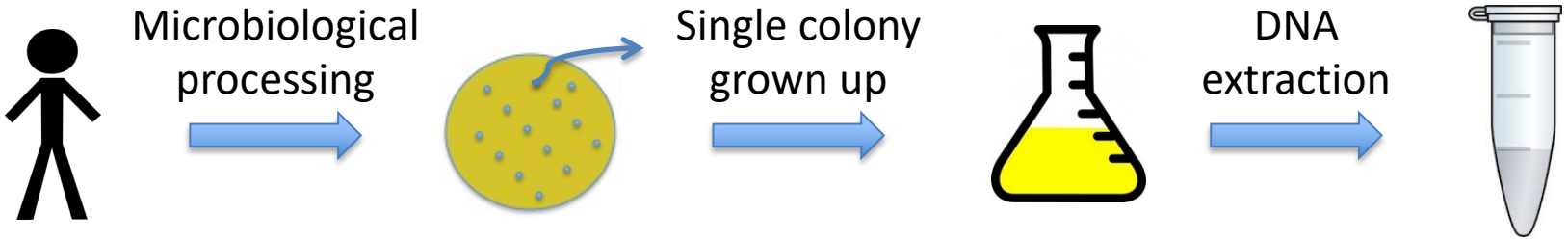


- Genomic epidemiology

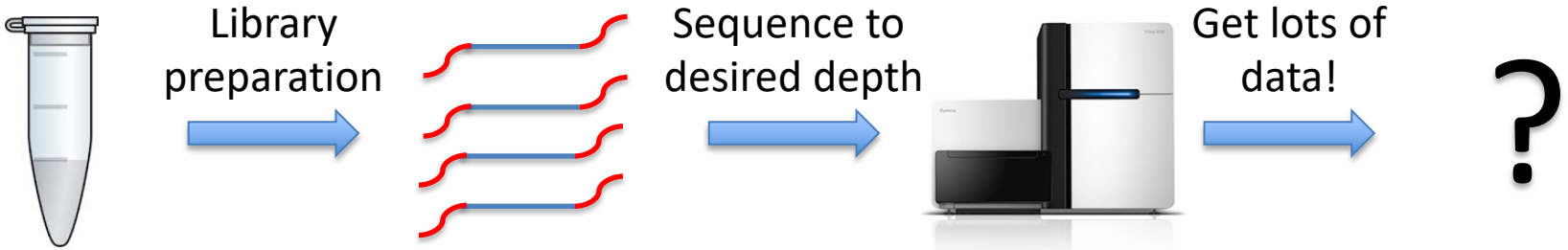


DNA and library preparation

1. Sample Preparation



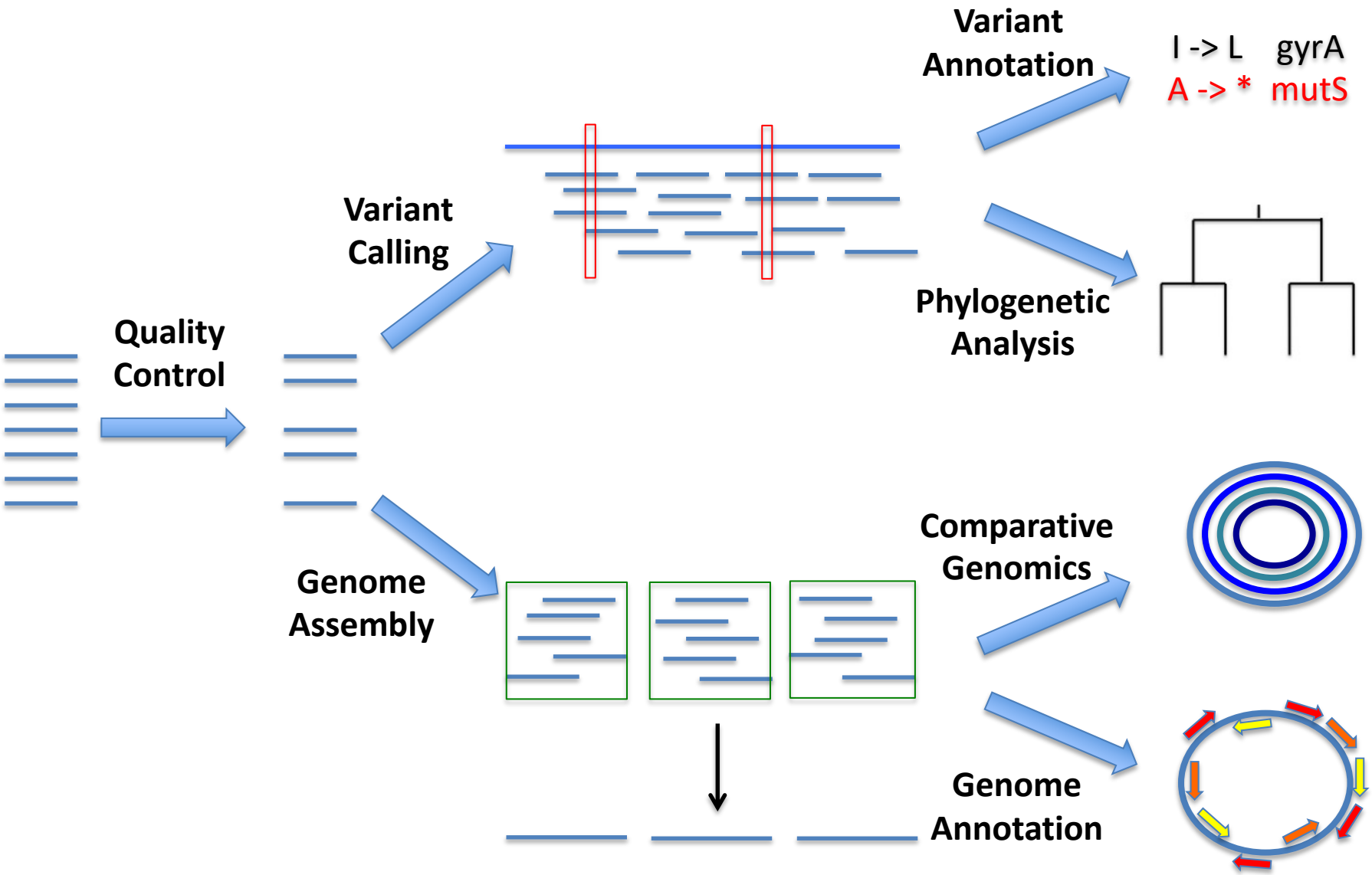
2. Sequencing



Illumina sequencing

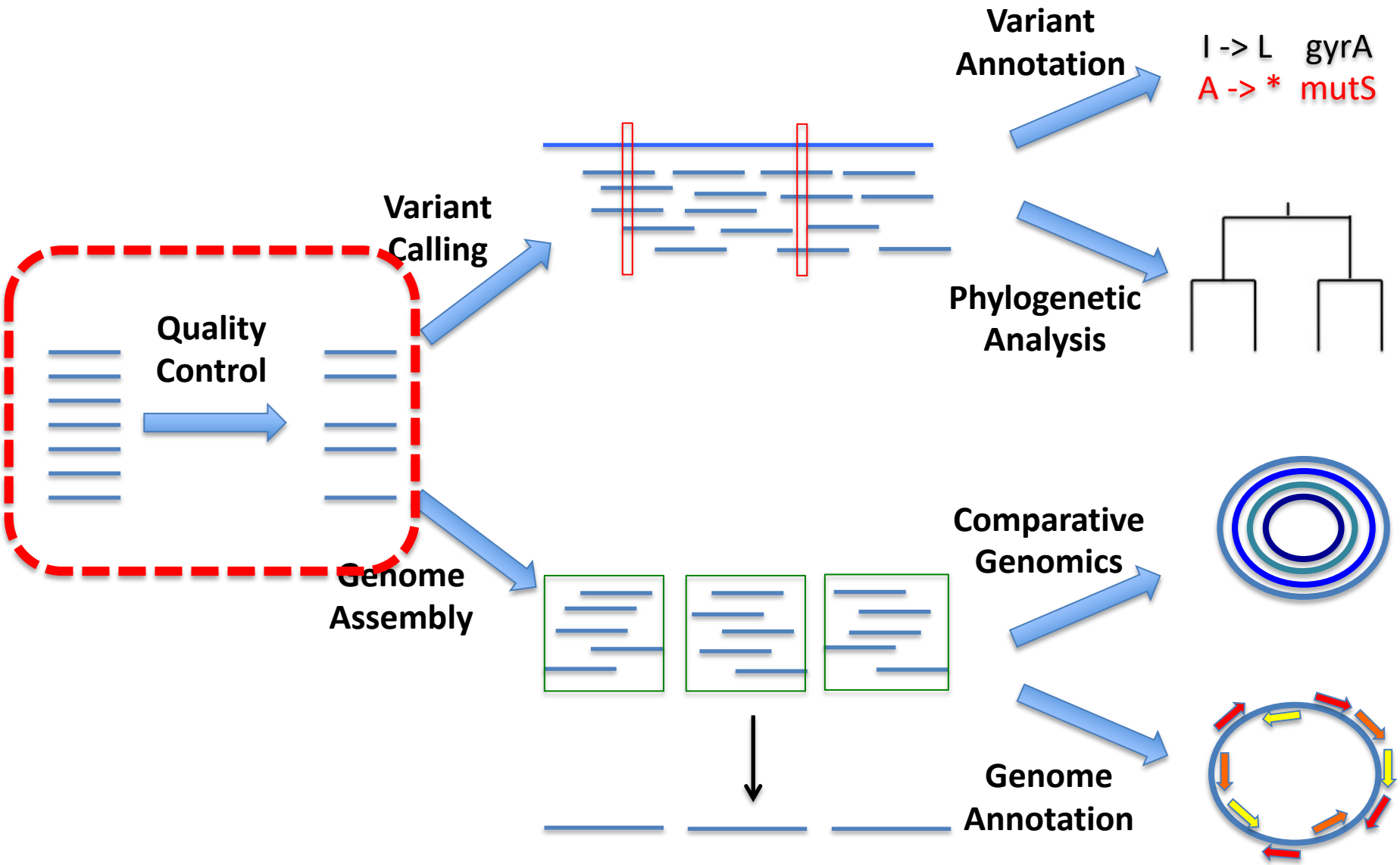
- https://youtu.be/fCd6B5HRaZ8

Mile-high view of a genomics pipeline



Day 1 afternoon – Data QC and variant calling

Mile-high view of a genomics pipeline



Sequencing quality control

Forward reads

```
@seq1_1
ACTGCACT
+
8-8,,+@+
@seq2_1
TGCATCTA
+
@+@E++BF
.
.
.
```

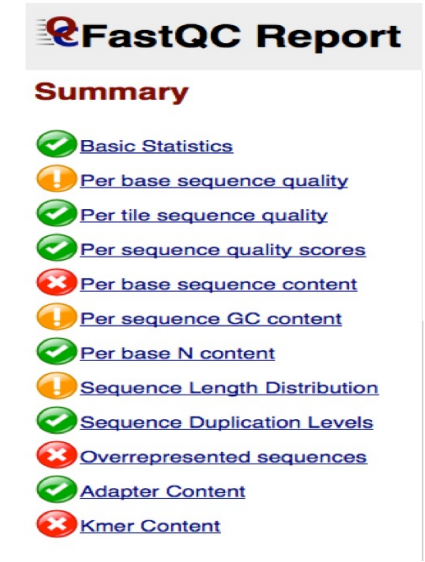
Reverse reads

```
@seq1_2
TCTATCGA
+
A<-9BFCFF
@seq2_2
CTAGTTAA
+
**>D7?7=.
.
.
.
```

Raw fastq files

FastQC/ Kraken

1. Contaminants
2. Aberrant quality



Trimmomatic

1. Filter reads
2. Trim reads

Forward reads

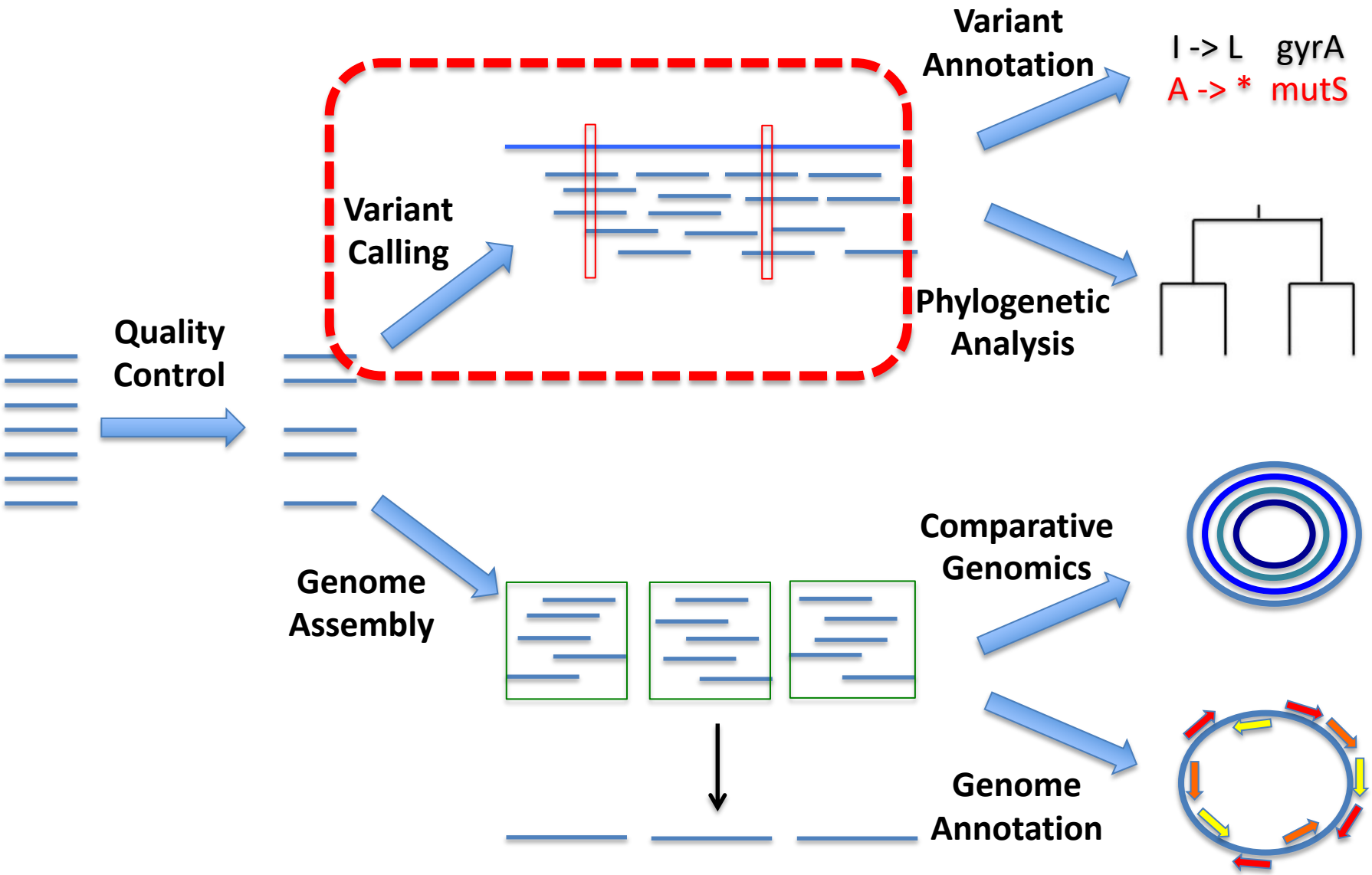
```
@seq1_1
ACTGCACT
+
8-8,,+@+
.
.
.
.
.
.
```

Reverse reads

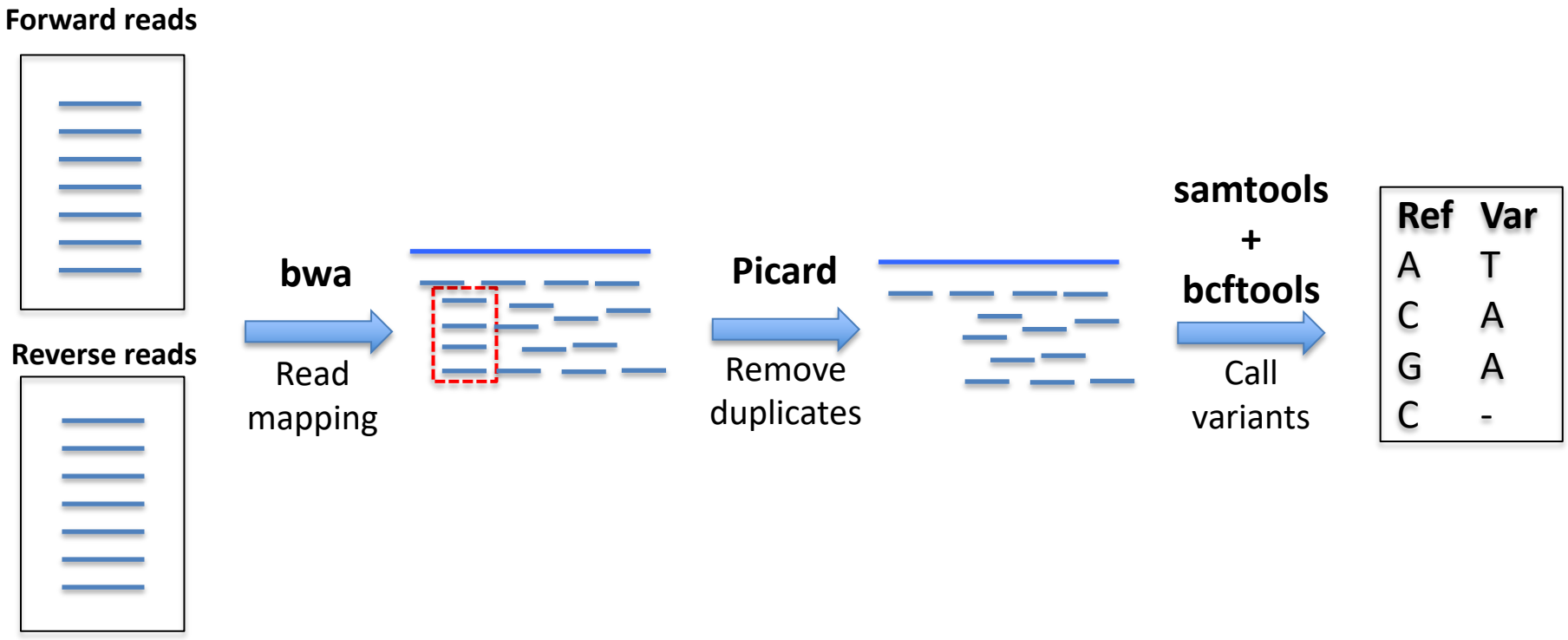
```
@seq1_2
TCTATCGA
+
A<-9BFCFF
.
.
.
.
.
.
```

Clean fastq files

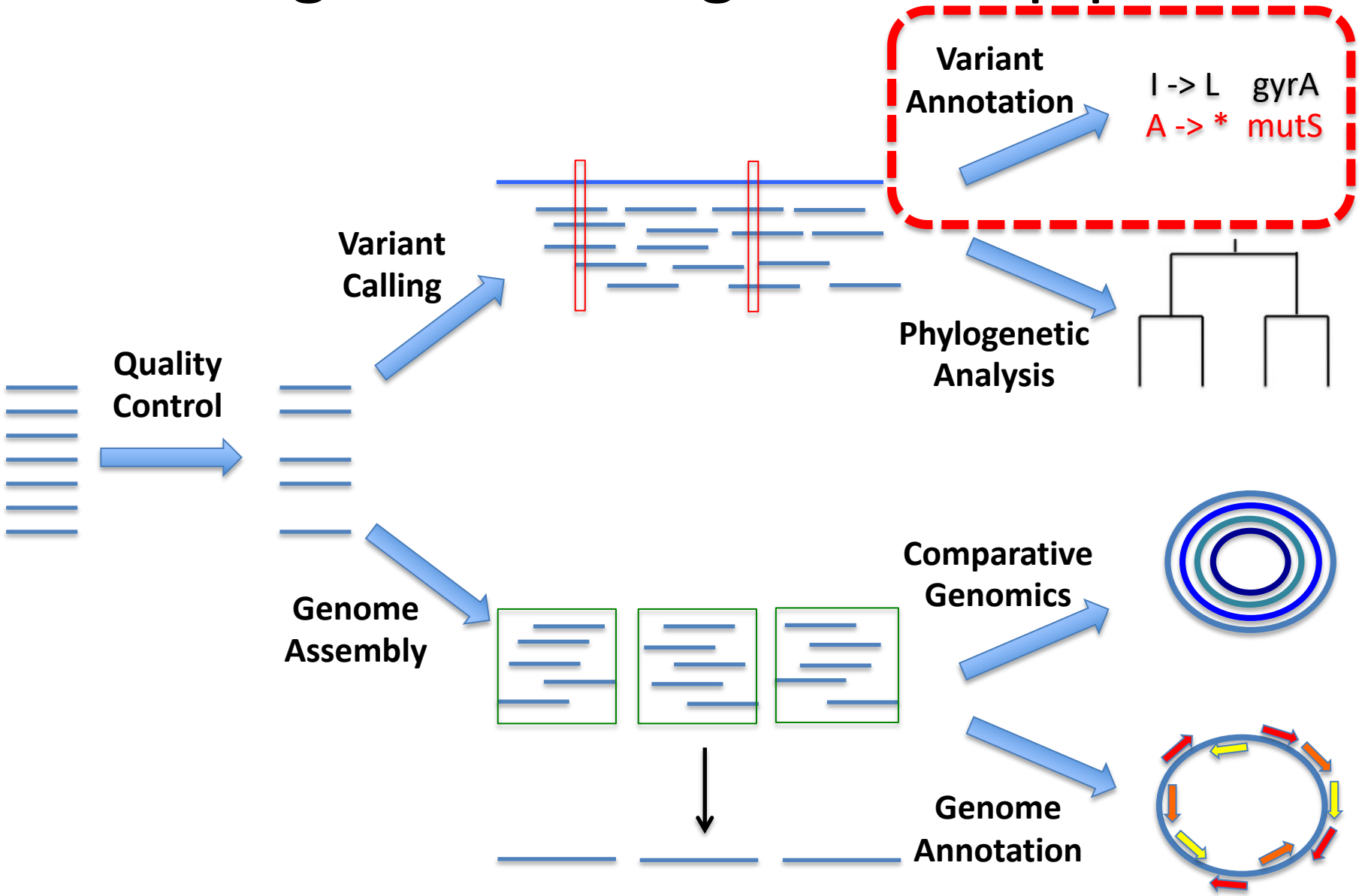
Mile-high view of a genomics pipeline



Variant identification



Mile-high view of a genomics pipeline



Variant filtering and annotation

Ref	Var
A	T
C	A
G	A
C	-

gatk

Variant filtering (FQ, MQ, depth)

Ref	Var
A	T
C	A
G	A
C	-

vcftools

Indel removal

Ref	Var
A	T
C	A
G	A
C	-

snpEff

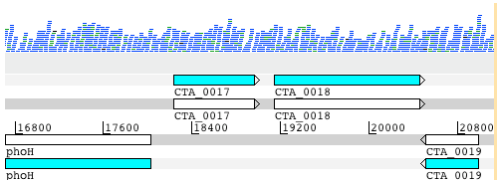
Variant annotation

Ref	Var	Ann
A	T	M->P/HIGH
C	A	I->L/LOW
G	A	
C	-	

Annotated VCF files

Visualization

IGV



VCF files

Filtered VCF files

Filtered VCF files

VCF, BAM, BAI, fasta files

Identifying antibiotic resistance mutations

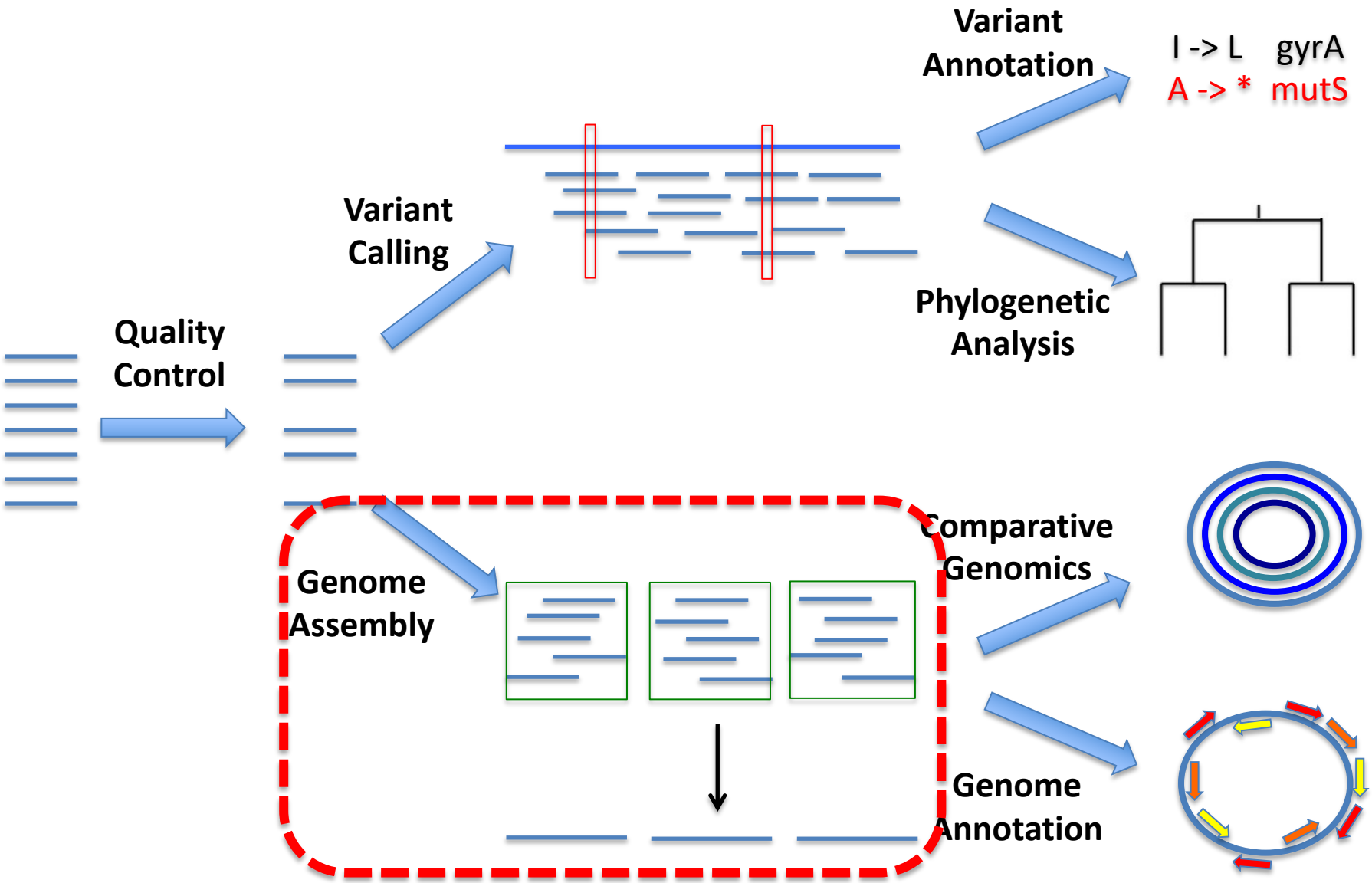
1. Colistin resistance in Klebsiella pneumoniae
    1. Distant reference genome
    2. Use biological knowledge to hone in on variant of interest

2. Daptomycin resistance in VRE
    1. Reference is susceptible ancestor from same patient
    2. Examine the small number of individual variants to identify putative causal mutation

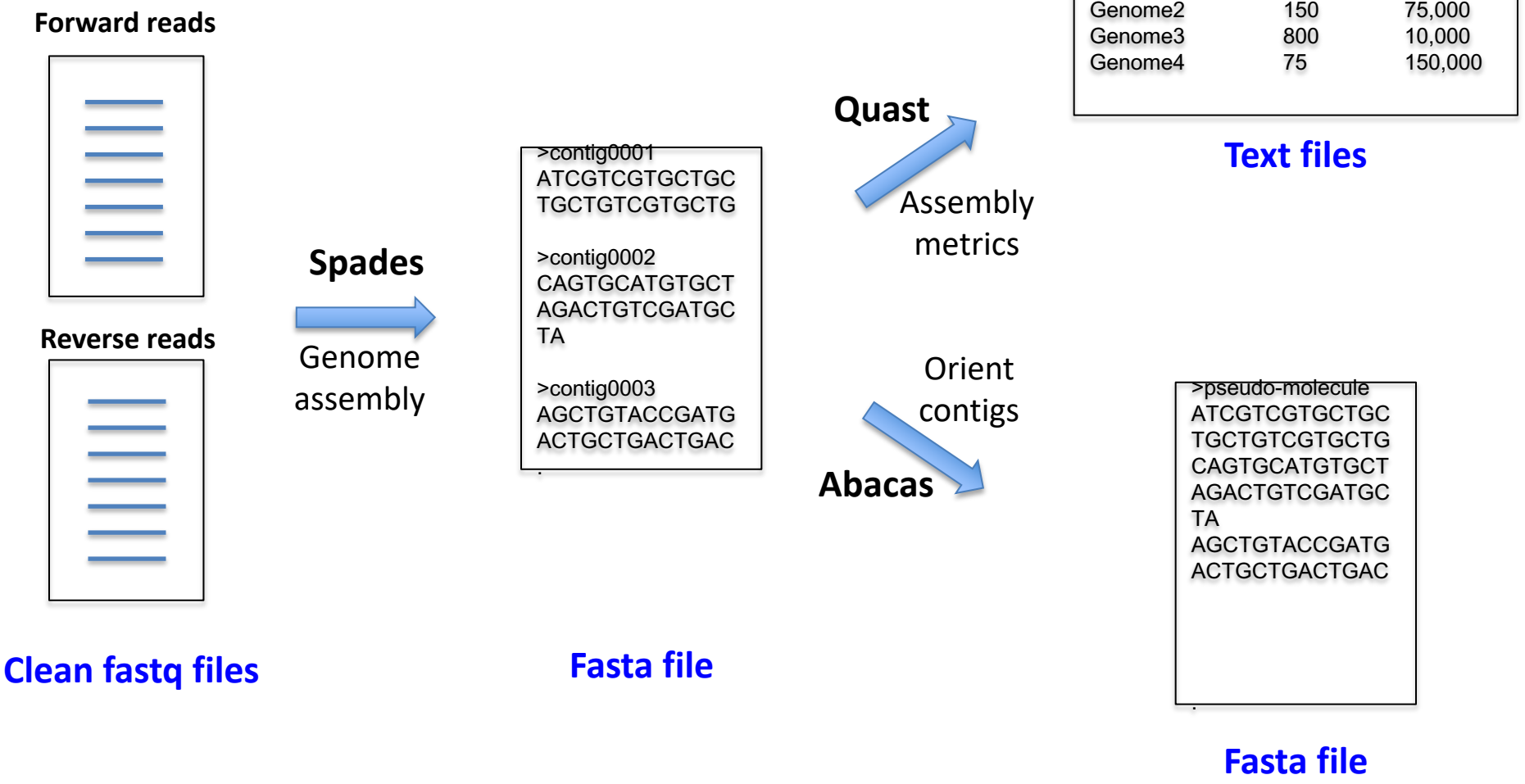
3. Colistin resistance in Acinetobacter
    1. Have resistant and susceptible isolate from same patient, but they are quite different
    2. Use biological knowledge to identify putative resistance variants that occur in resistant isolate, but not susceptible ancestor

Day 2 morning – Genome assembly and annotation

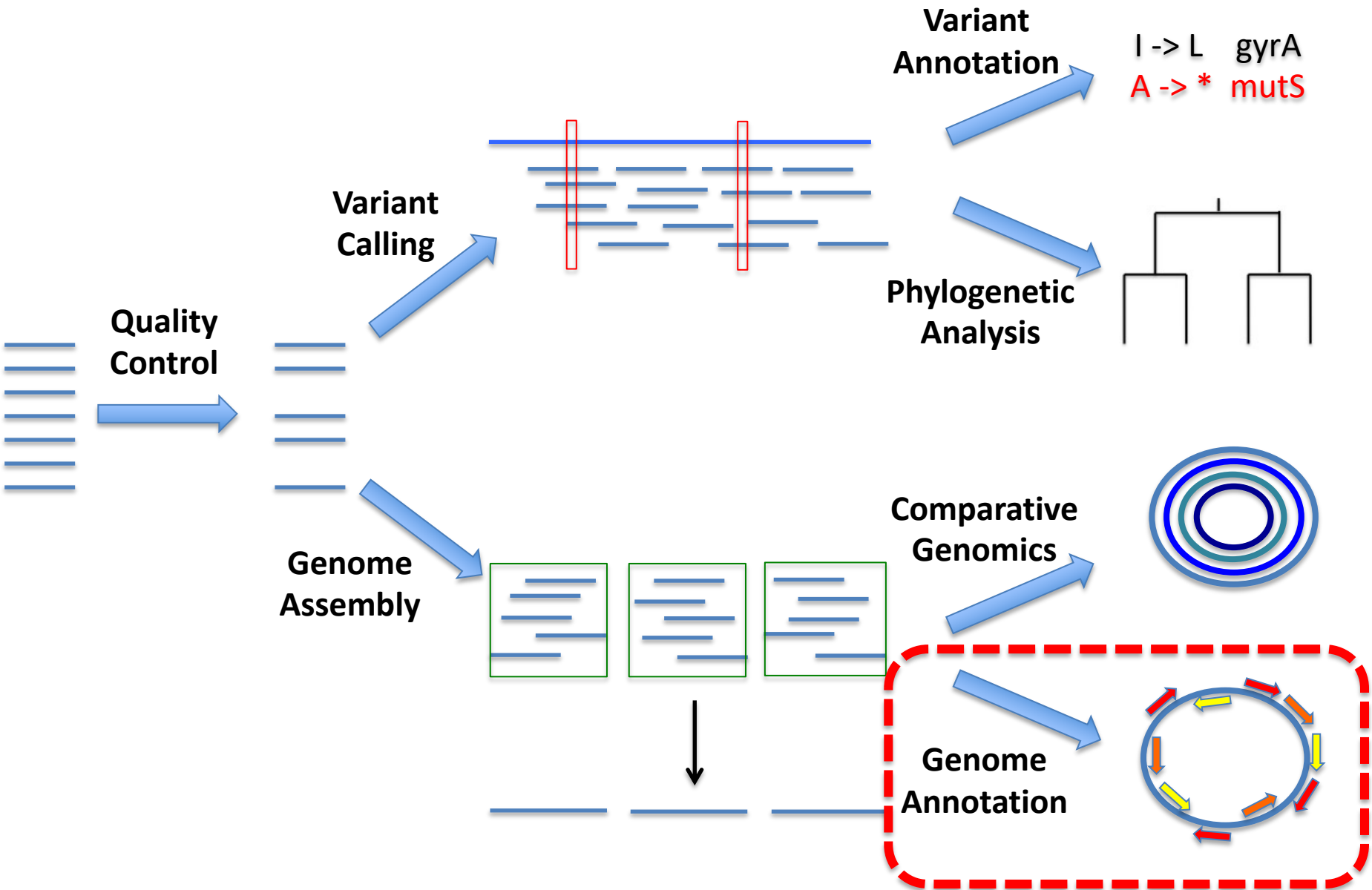
Mile-high view of a genomics pipeline



Genome assembly



Mile-high view of a genomics pipeline

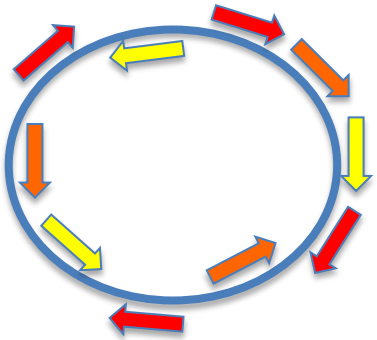


Genome annotation

>pseudo-molecule
ATCGTCGTGCTGC
TGCTGTCGTGCTG
CAGTGCATGTGCT
AGACTGTCGATGC
TA
AGCTGTACCGATG
ACTGCTGACTGAC

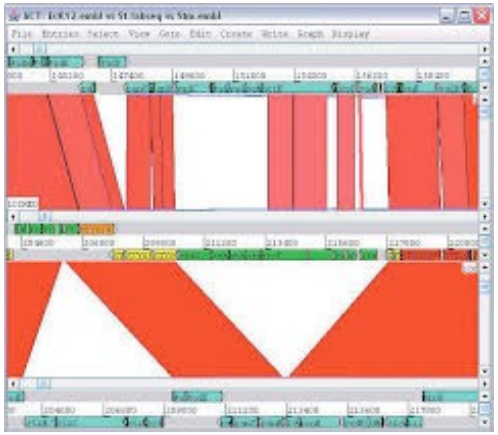
Prokka

1) Gene finding
2) Basic annotation



ACT

Visualization



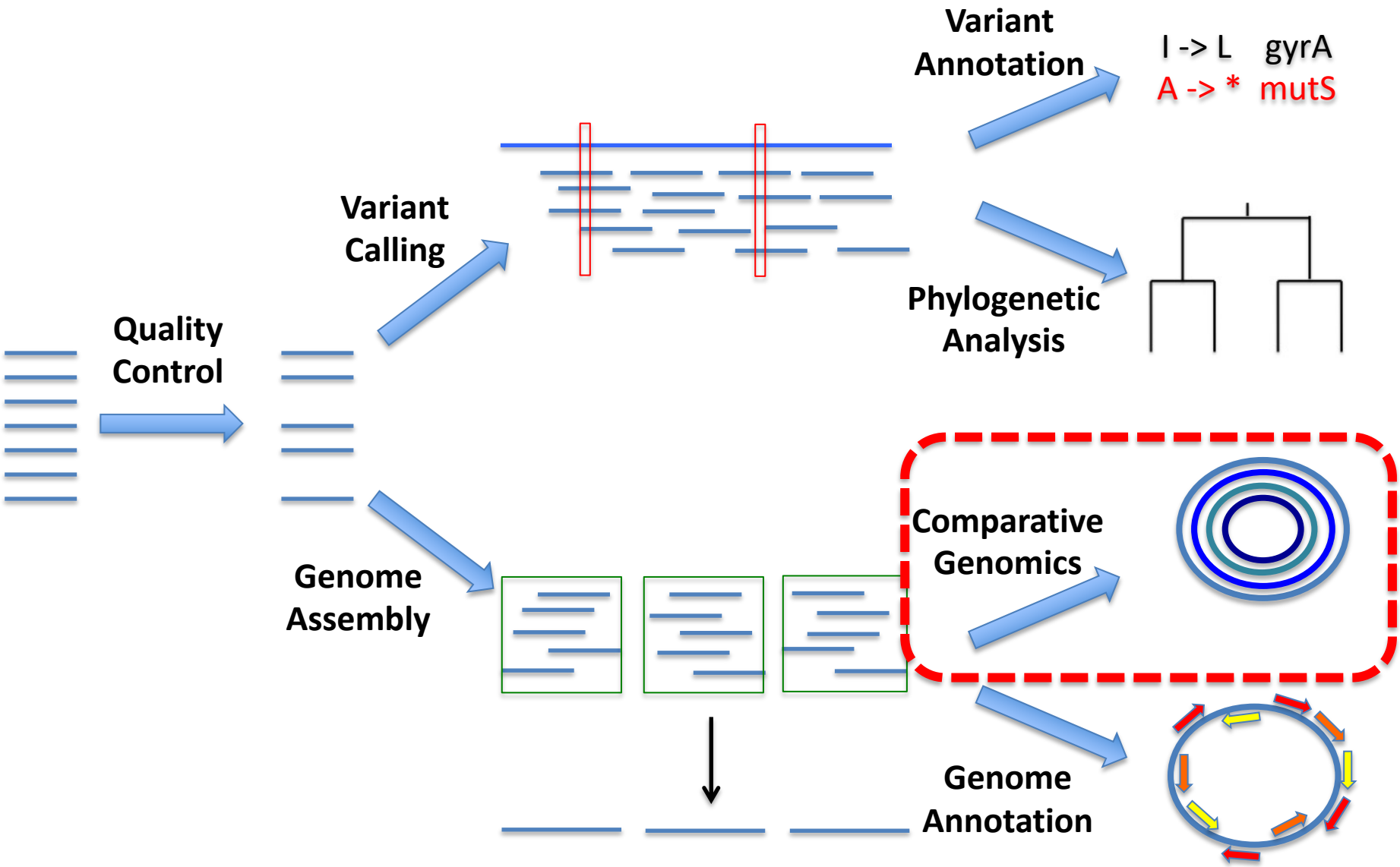
Fasta file

Genbank file

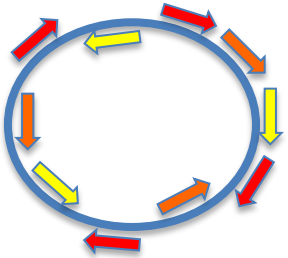
Genbank files, alignment files

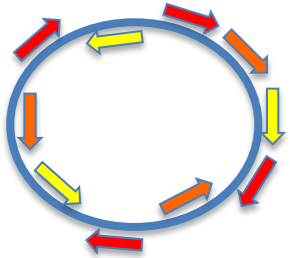
Day 2 afternoon– Comparative genomics

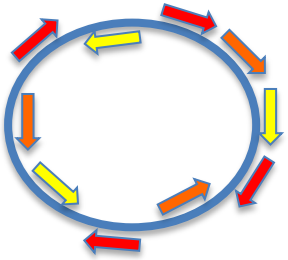
Mile-high view of a genomics pipeline



Comparative genomics



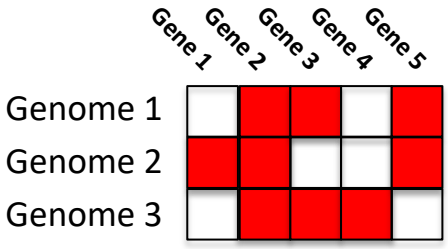




Fasta, genbank and/or pep

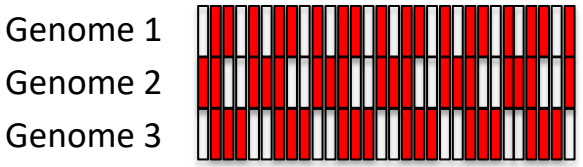
BLAST/ ARIBA

Genome mining



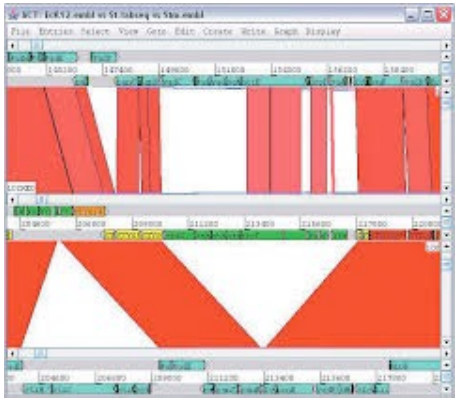
Roary

Pan-genome analysis



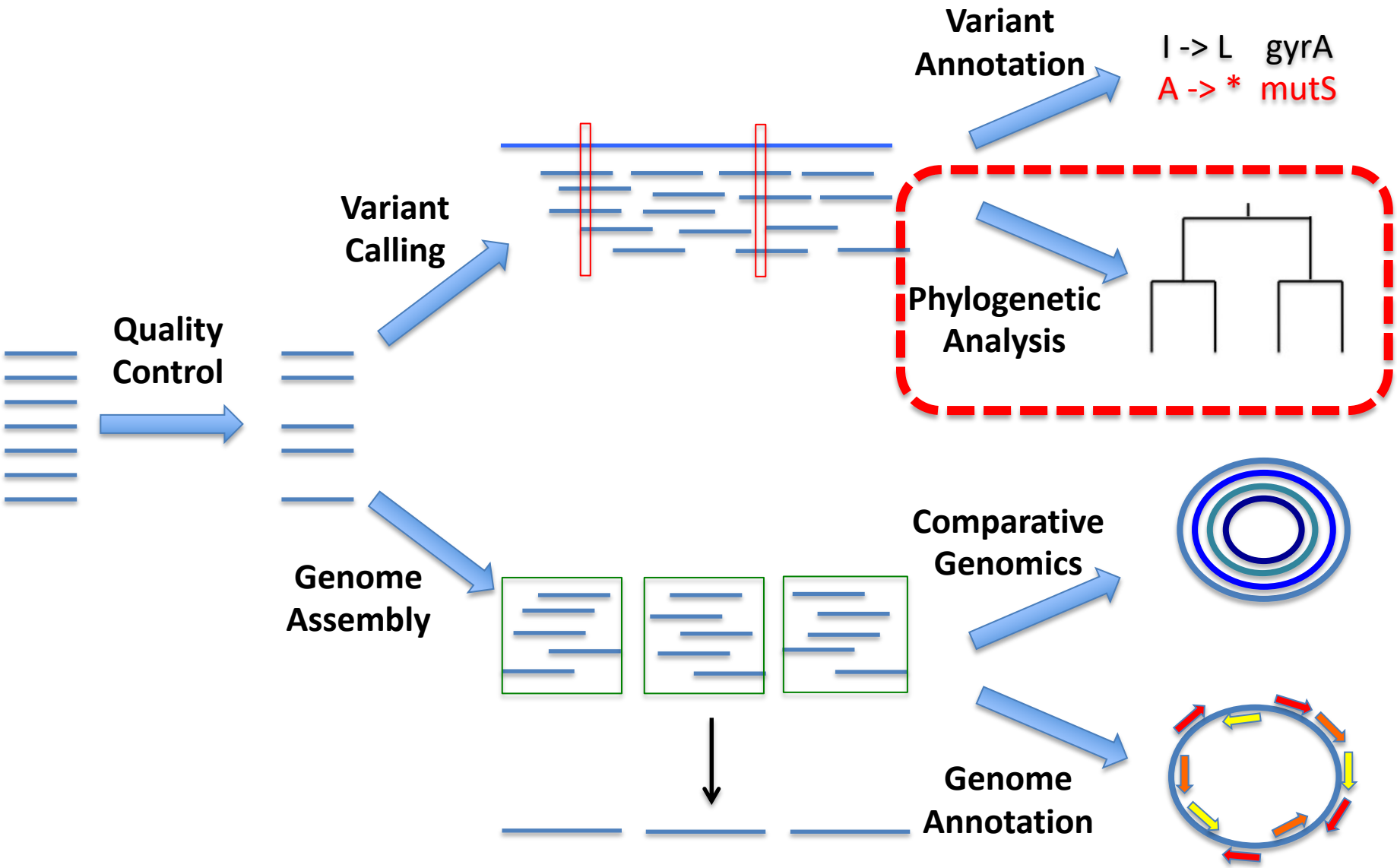
ACT

Structural variants



Day 3 morning – Basic phylogenetic analysis

Mile-high view of a genomics pipeline



Phylogenetics

>Genome 1
ATCGTCGTGCTGC
TGCTGTCGTGCTG

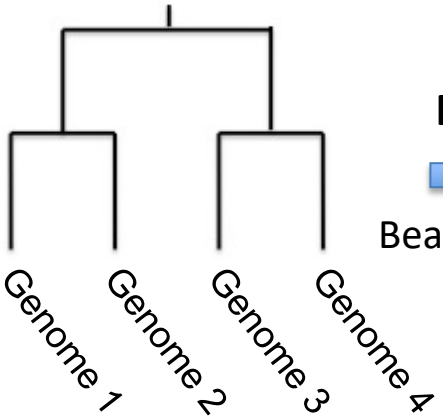
>Genome 2
CAGTGCATGTGCT
AGACTGTCGATGC
TA

>Genome 3
AGCTGTACCGATG
ACTGCTGACTGAC

Multi-fasta file

ape

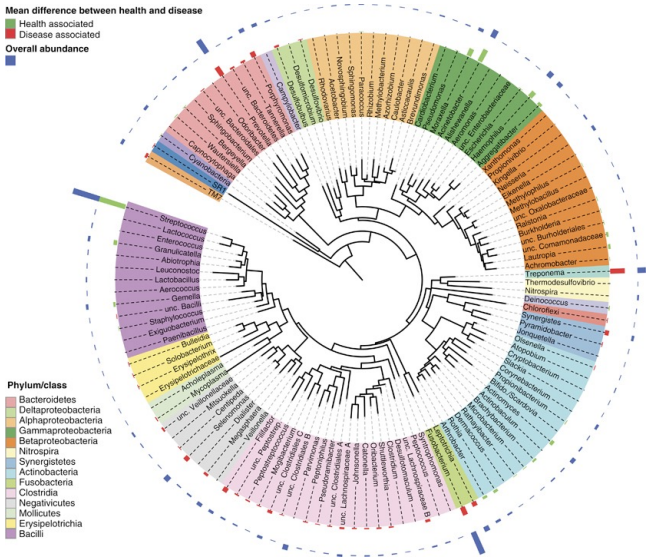
Tree construction

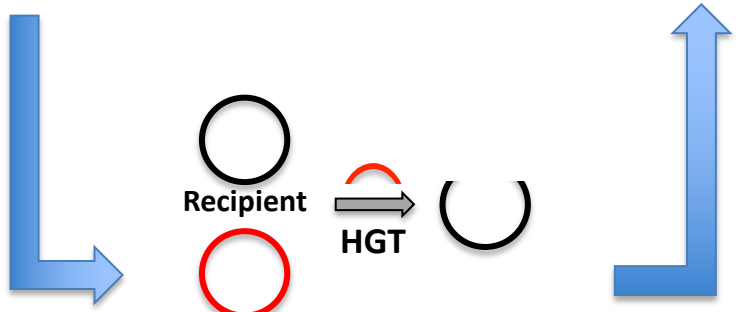


Nexus file

R/iTOL

Beautification





Gubbins

Recombination filtering

Gubbins/ape

Tree construction