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

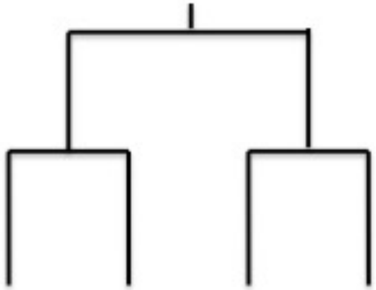
- https://cyberduck.io/download/

Why Unix?

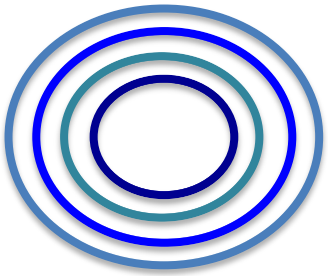
- Most bioinformatics research is performed in a Unix environment
- Allows for easier interactions with text files
- The power of pipes
- Easy to automate repetitive tasks
- Facilitates interfacing with high-performance compute systems

So you want to sequence some bacteria?

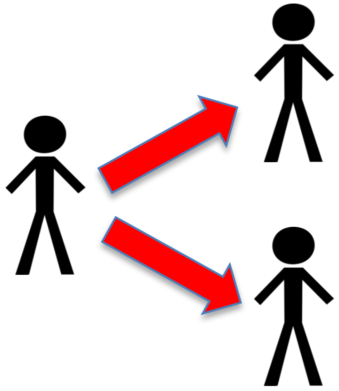
- Microbial phylogenetics



- Comparative genomics

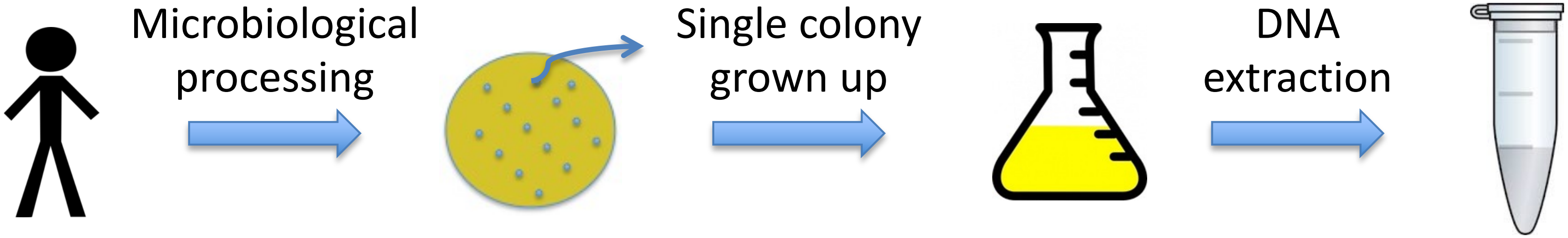


- Genomic epidemiology



DNA and library preparation

1. Sample Preparation

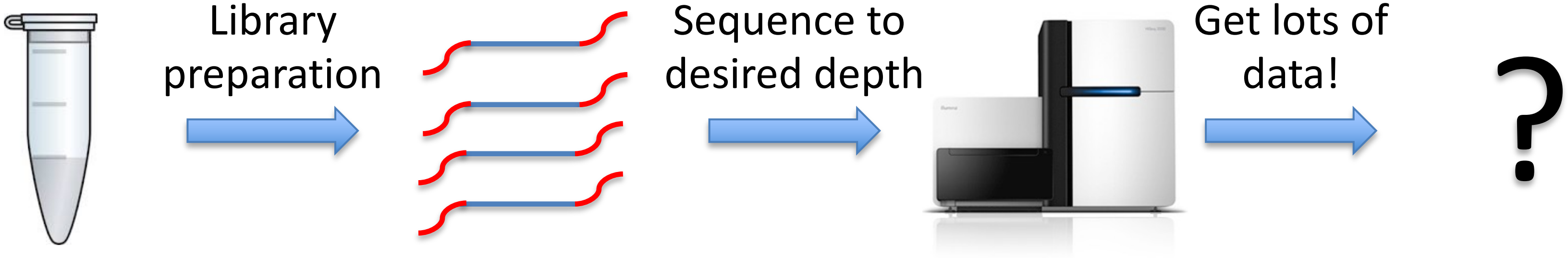


Microbiological processing

Single colony grown up

DNA extraction

2. Sequencing



Library preparation

Sequence to desired depth

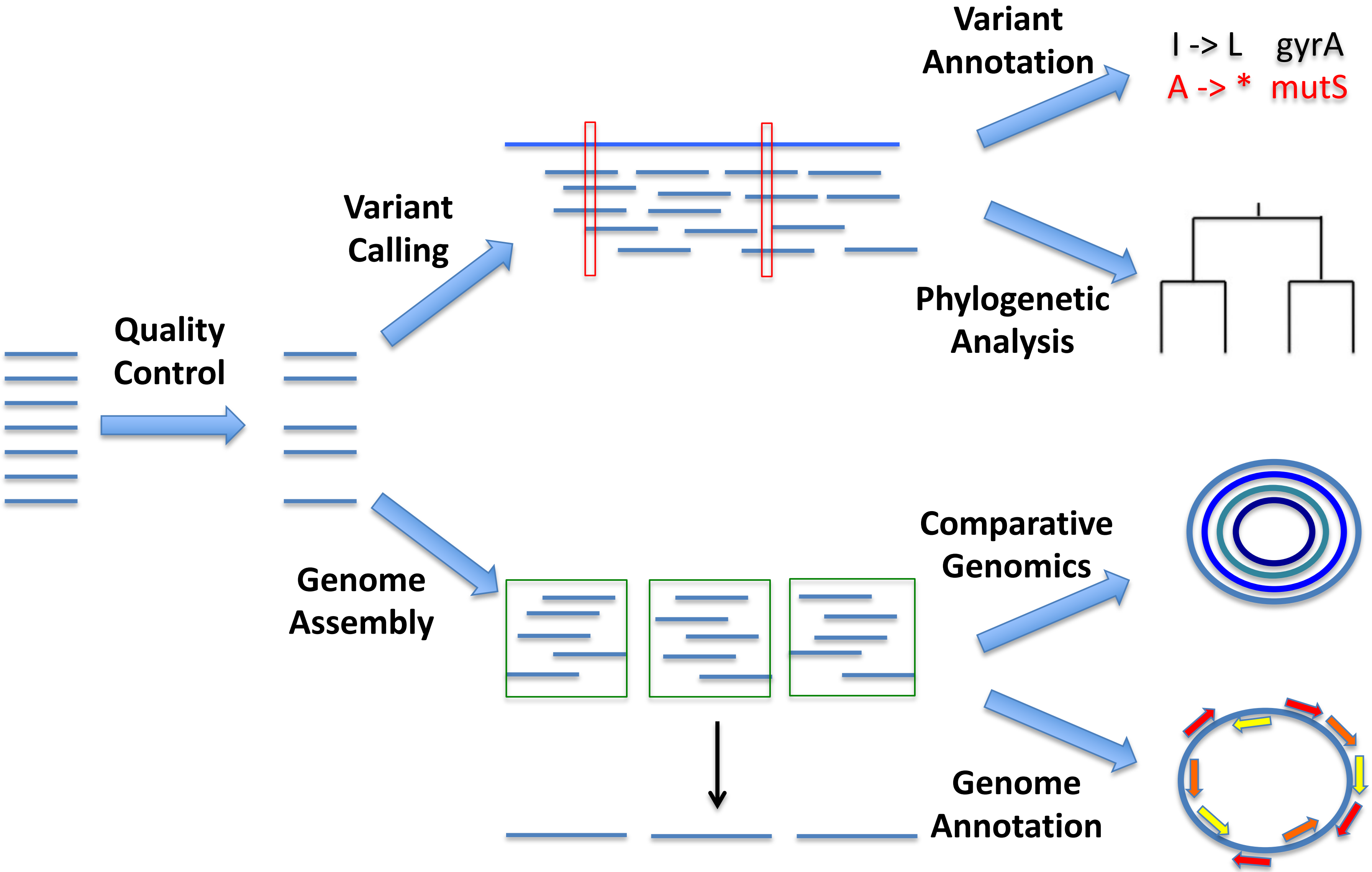
Get lots of data!

?

Illumina sequencing

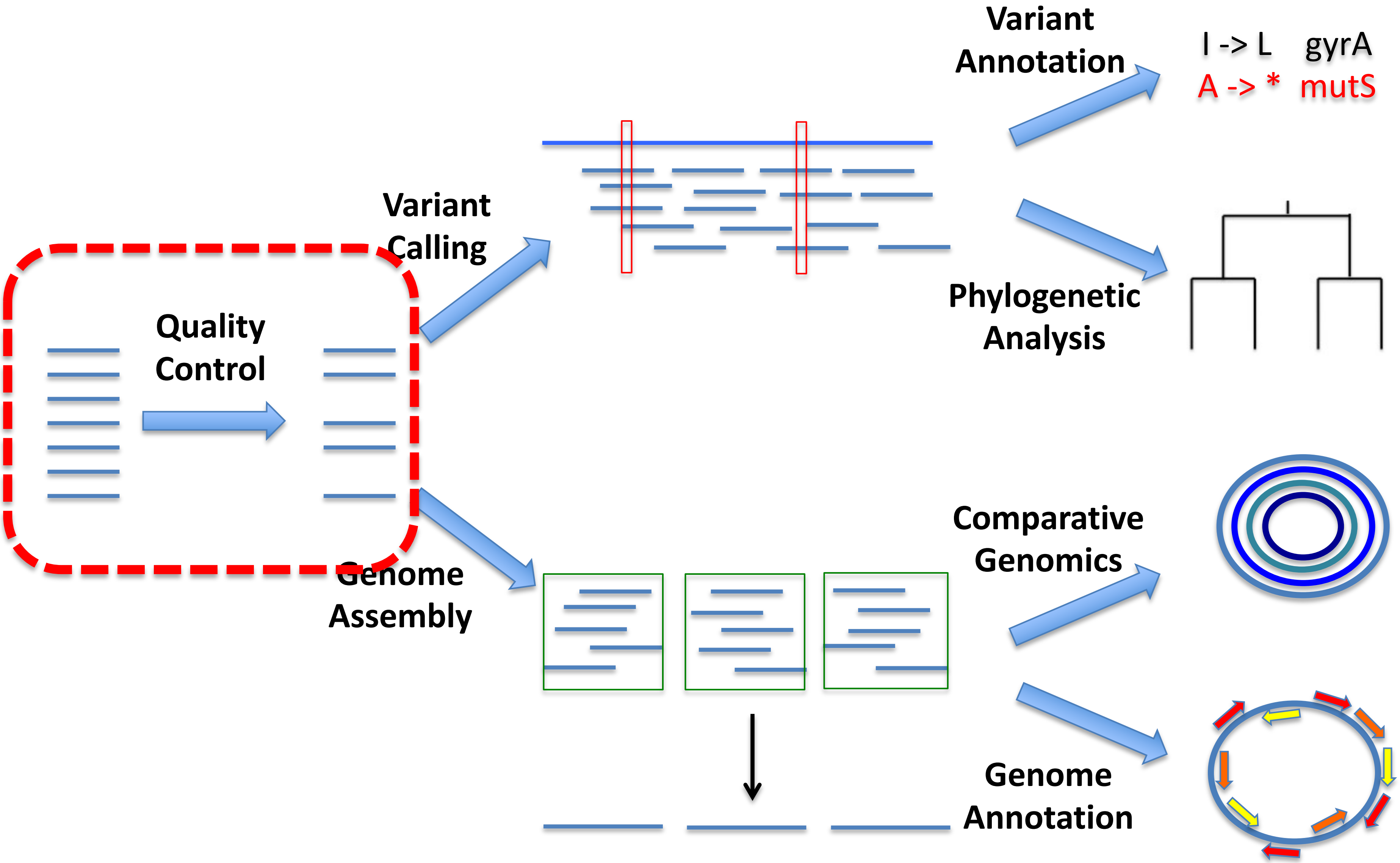
- https://youtu.be/fCd6B5HRaZ8

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
.
.
.
```

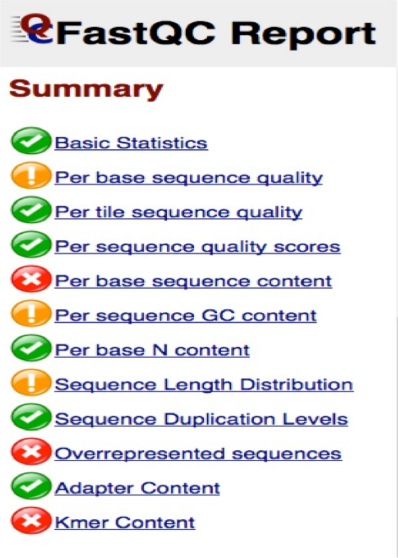
Reverse reads

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

Raw fastq files

FastQC/ Kraken

1. Contaminants
2. Aberrant quality



FastQC Report

Summary

- ✓ Basic Statistics
- ! Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ! Per sequence GC content
- ✓ Per base N content
- ! Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✗ Overrepresented sequences
- ✓ Adapter Content
- ✗ Kmer Content

Trimmomatic

1. Filter reads
2. Trim reads

Forward reads

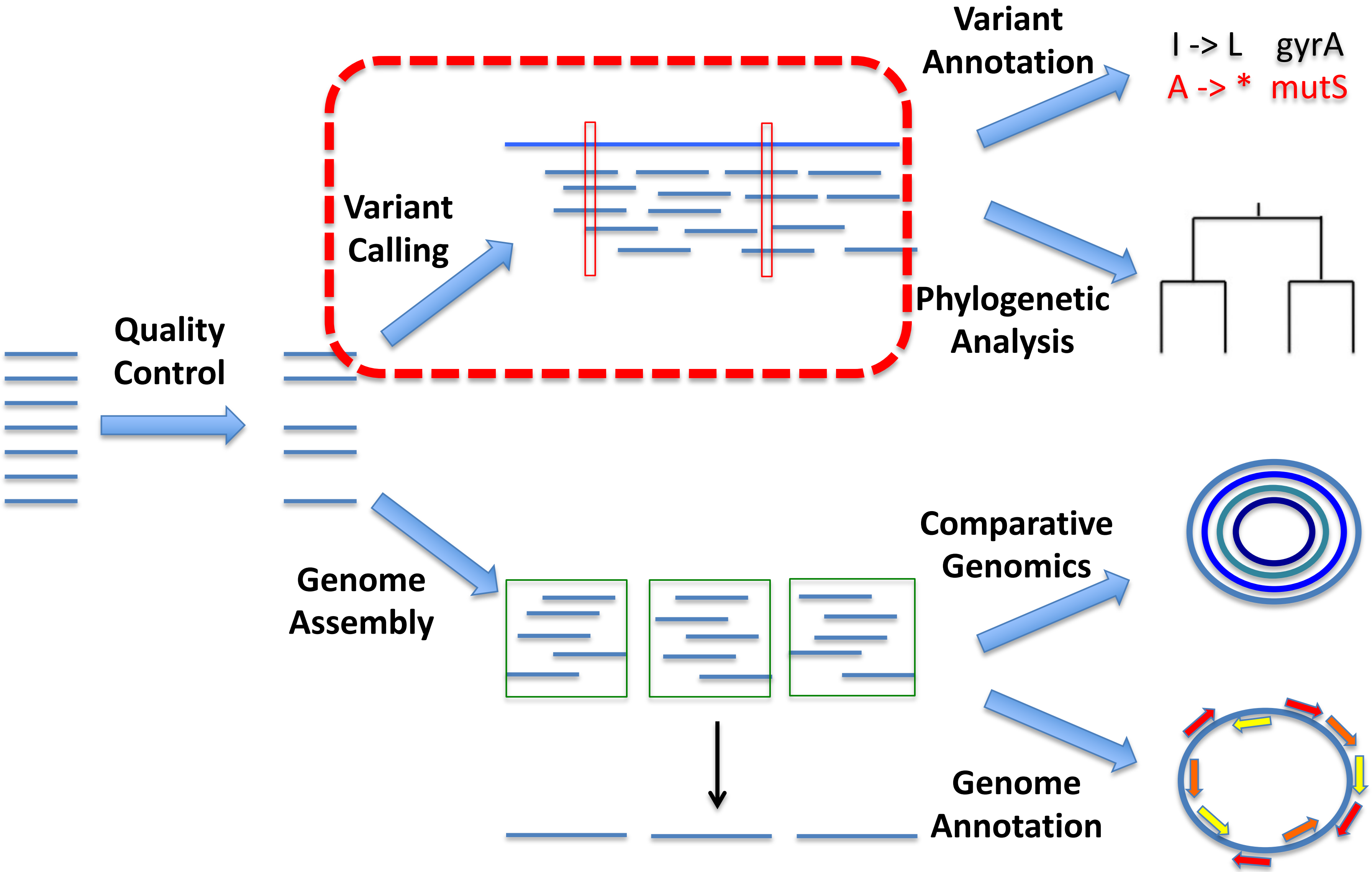
```
@seq1_1
ACTGCACT
+
8-8,,+@+
.
.
.
.
.
.
```

Reverse reads

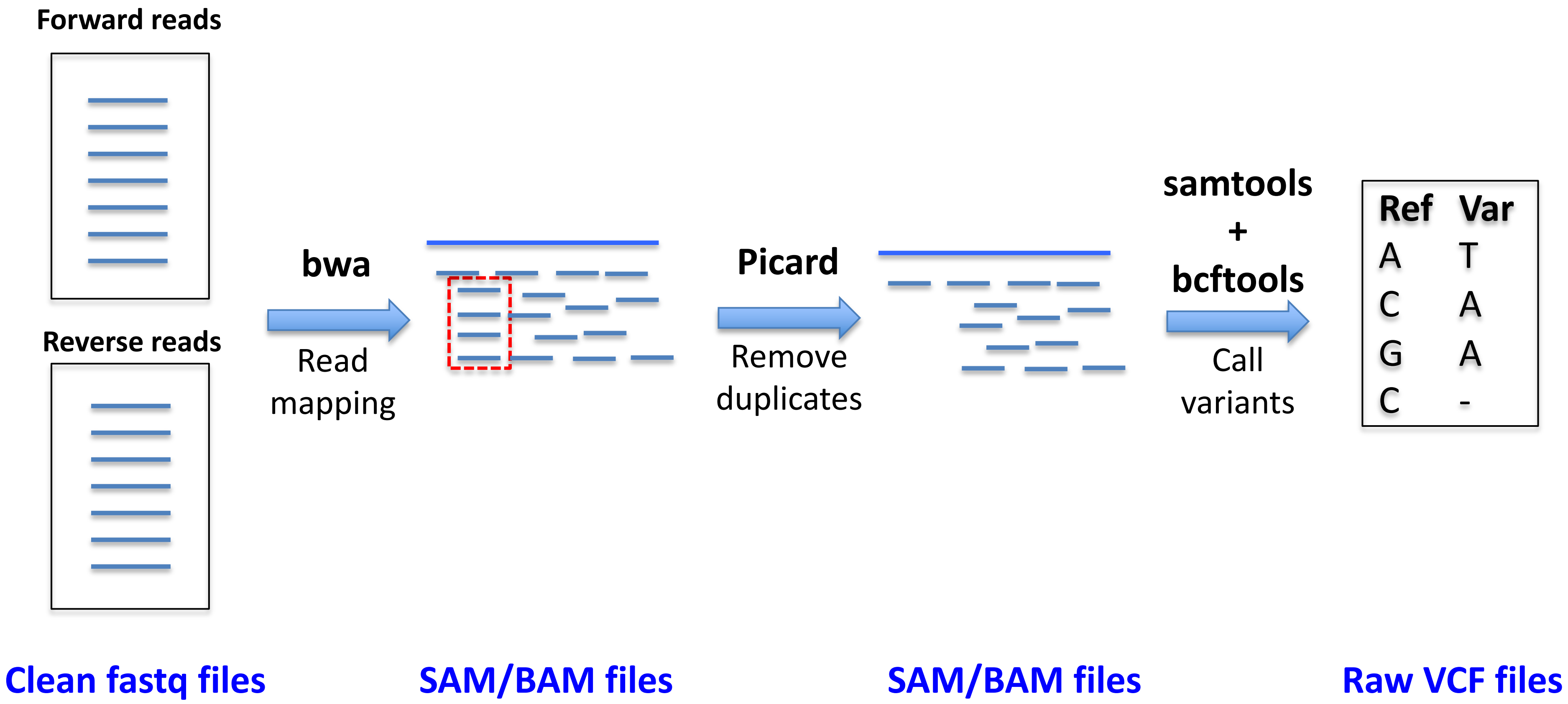
```
@seq1_2
TCTATCGA
+
A<-9BFCFF
.
.
.
.
.
.
```

Clean fastq files

Mile-high view of a genomics pipeline



Variant identification



Forward reads

Reverse reads

bwa

Read mapping

Picard

Remove duplicates

samtools
+
bcftools

Call variants

Ref	**Var**
A	T
C	A
G	A
C	-

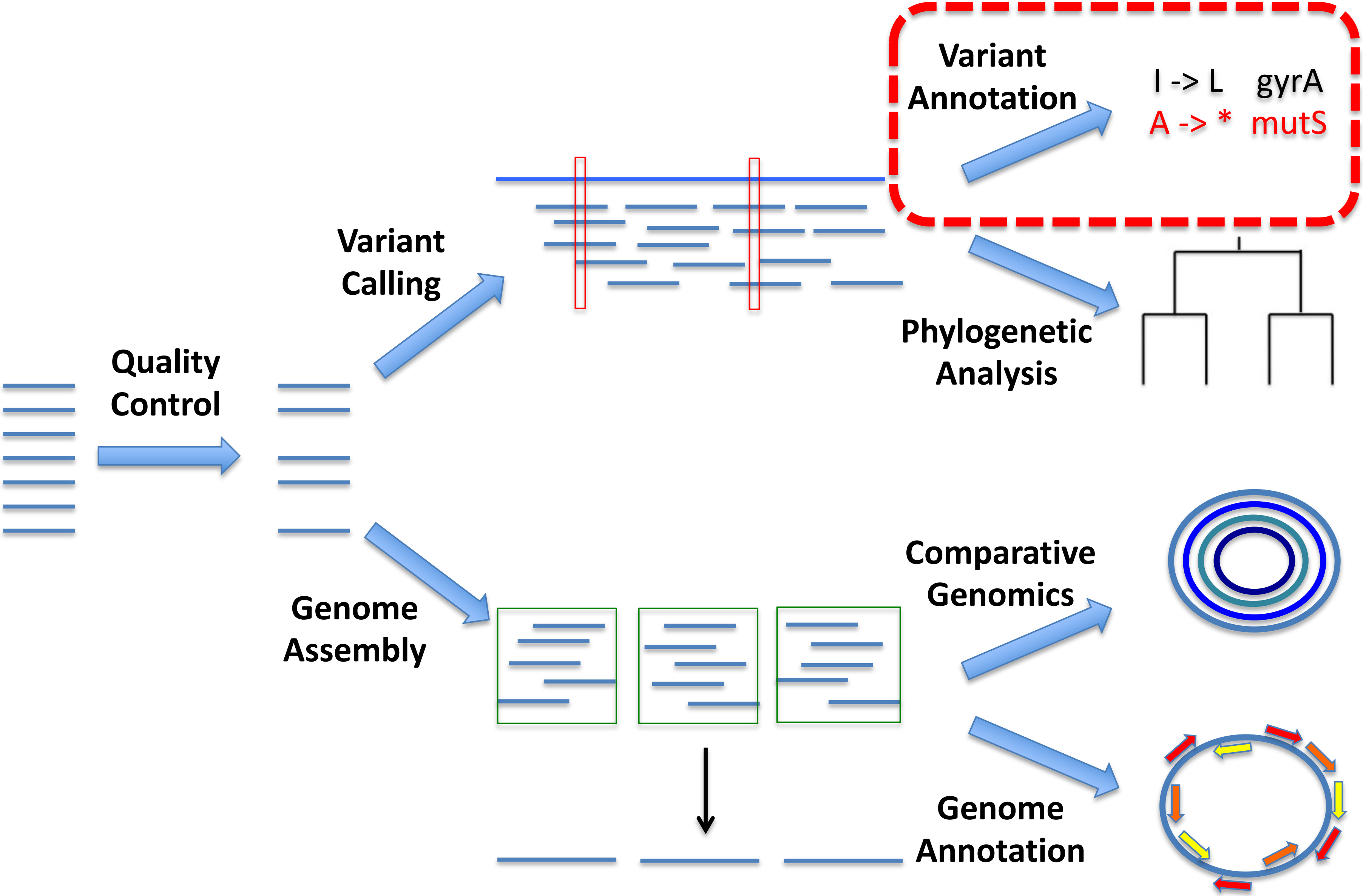
Clean fastq files

SAM/BAM files

SAM/BAM files

Raw VCF files

Mile-high view of a genomics pipeline



Variant Annotation

I -> L gyrA

A -> * mutS

Variant Calling

Phylogenetic Analysis

Quality Control

Comparative Genomics

Genome Assembly

Genome Annotation

Variant filtering and annotation

Ref	Var
A	T
C	A
G	A
C	-

VCF files

gatk

Variant filtering (FQ, MQ, depth)

Ref	Var
A	T
C	A
G	A
C	-

Filtered VCF files

vcftools

Indel removal

Ref	Var
A	T
C	A
G	A
C	-

Filtered VCF files

snpEff

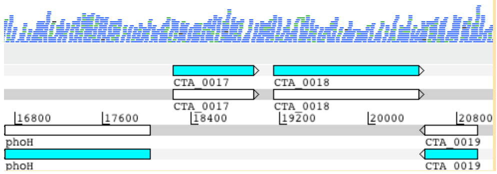
Variant annotation

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

Annotated VCF files

IGV

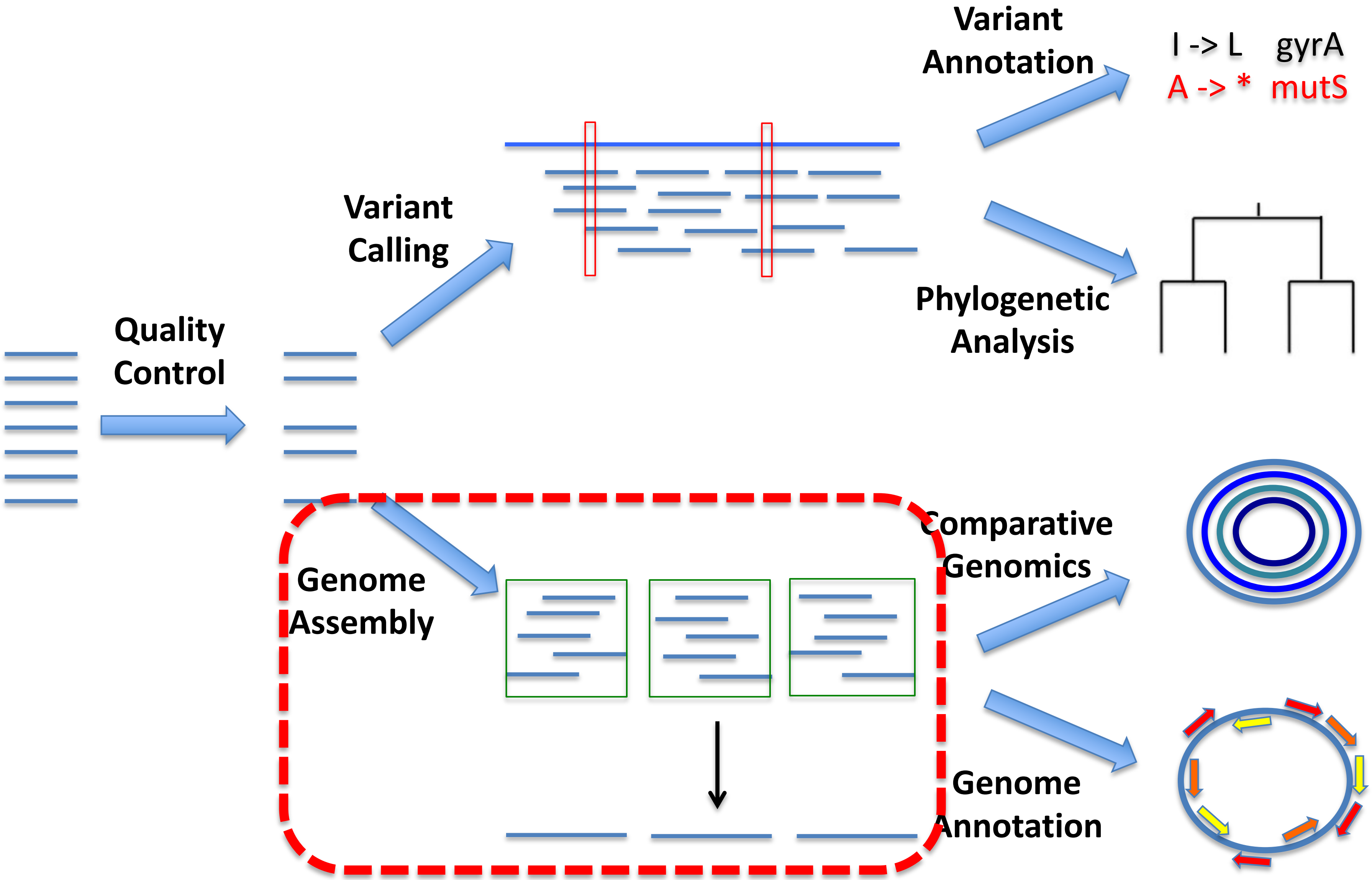
Visualization



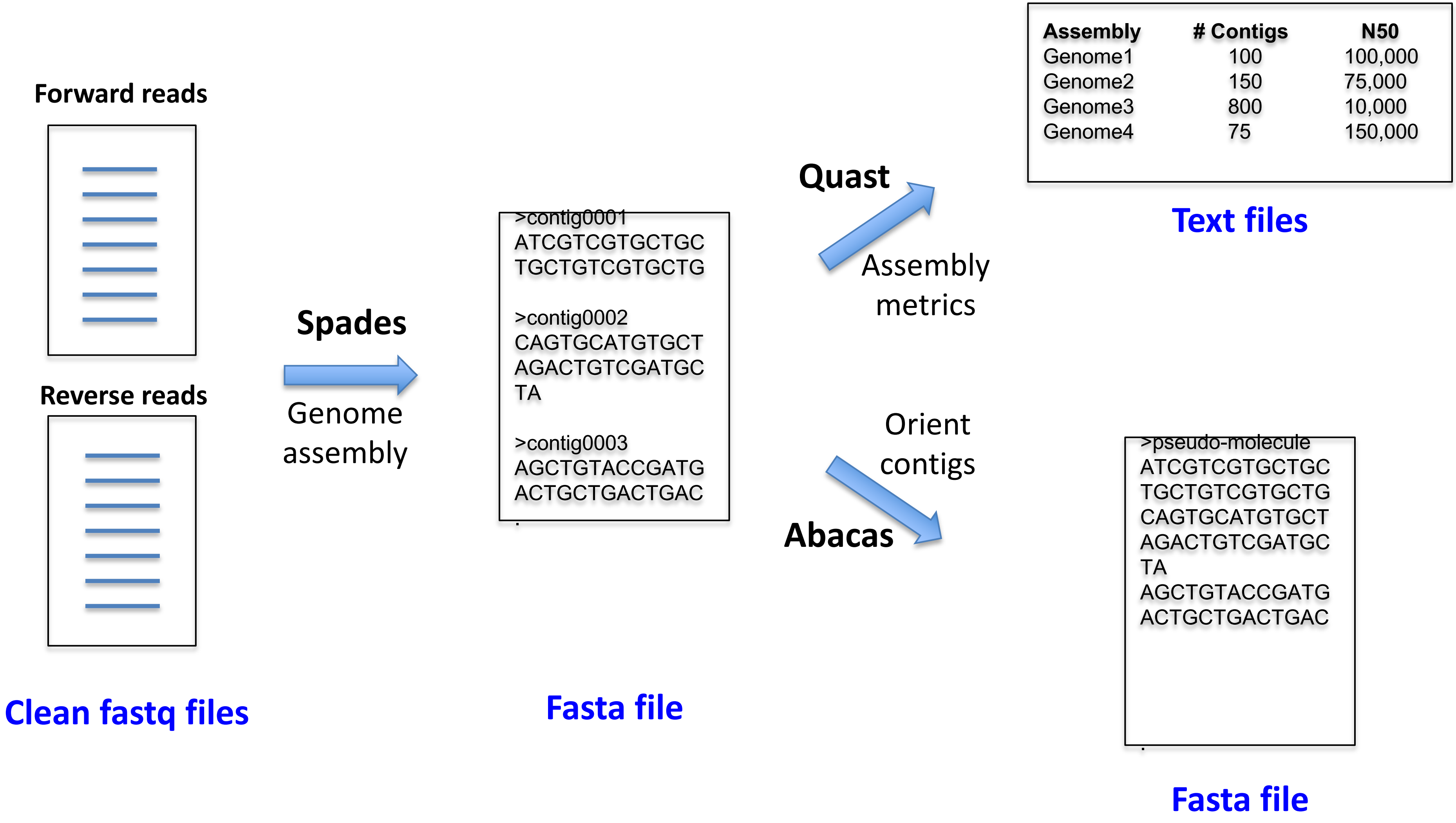
VCF, BAM, BAI, fasta files

Day 2 morning – Genome assembly and annotation

Mile-high view of a genomics pipeline



Genome assembly



Forward reads

Reverse reads

Clean fastq files

Spades

Genome assembly

```
>contig0001
ATCGTCGTGCTGC
TGCTGTCGTGCTG

>contig0002
CAGTGCATGTGCT
AGACTGTCGATGC
TA

>contig0003
AGCTGTACCGATG
ACTGCTGACTGAC
.
```

Fasta file

Quast

Assembly metrics

Assembly	**# Contigs**	**N50**
Genome1	100	100,000
Genome2	150	75,000
Genome3	800	10,000
Genome4	75	150,000

Text files

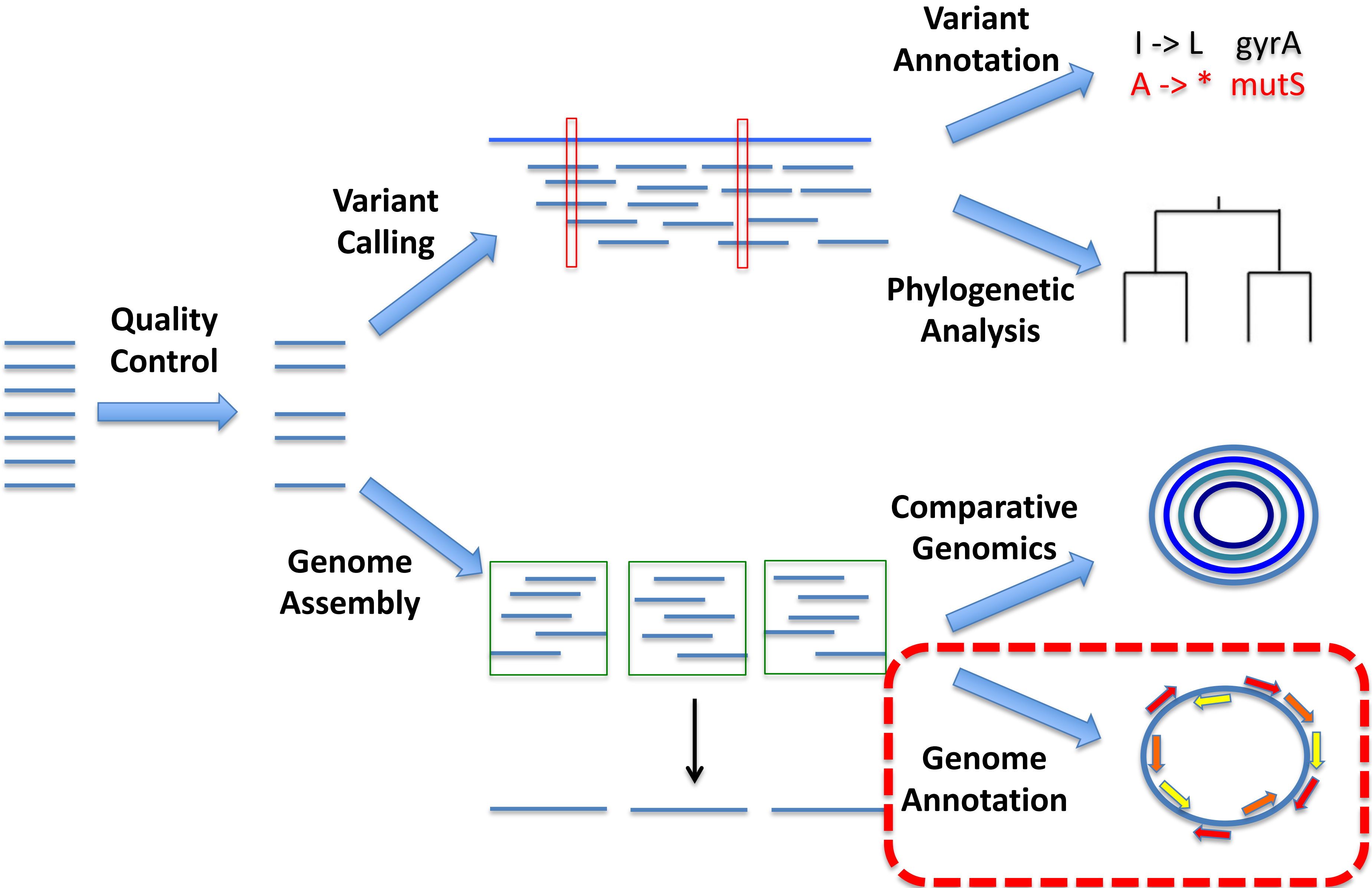
Orient contigs

Abacas

```
>pseudo-molecule
ATCGTCGTGCTGC
TGCTGTCGTGCTG
CAGTGCATGTGCT
AGACTGTCGATGC
TA
AGCTGTACCGATG
ACTGCTGACTGAC
.
```

Fasta file

Mile-high view of a genomics pipeline

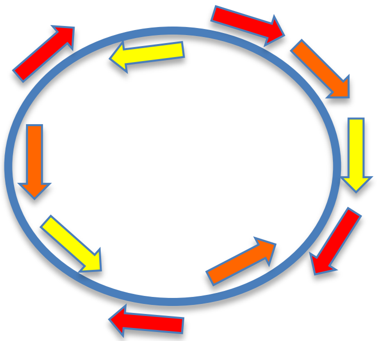


Genome annotation

```
>pseudo-molecule
ATCGTCGTGCTGC
TGCTGTCGTGCTG
CAGTGCATGTGCT
AGACTGTCGATGC
TA
AGCTGTACCGATG
ACTGCTGACTGAC
```

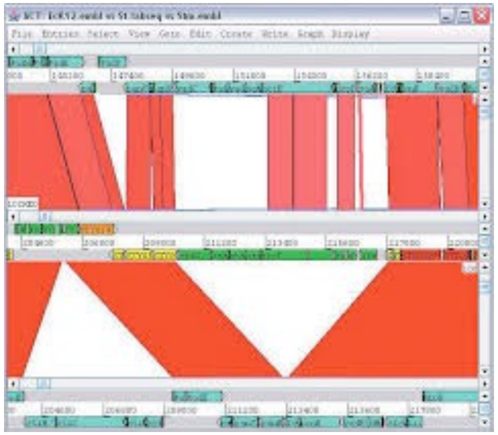
Prokka

1) Gene finding
2) Basic annotation



ACT

Visualization



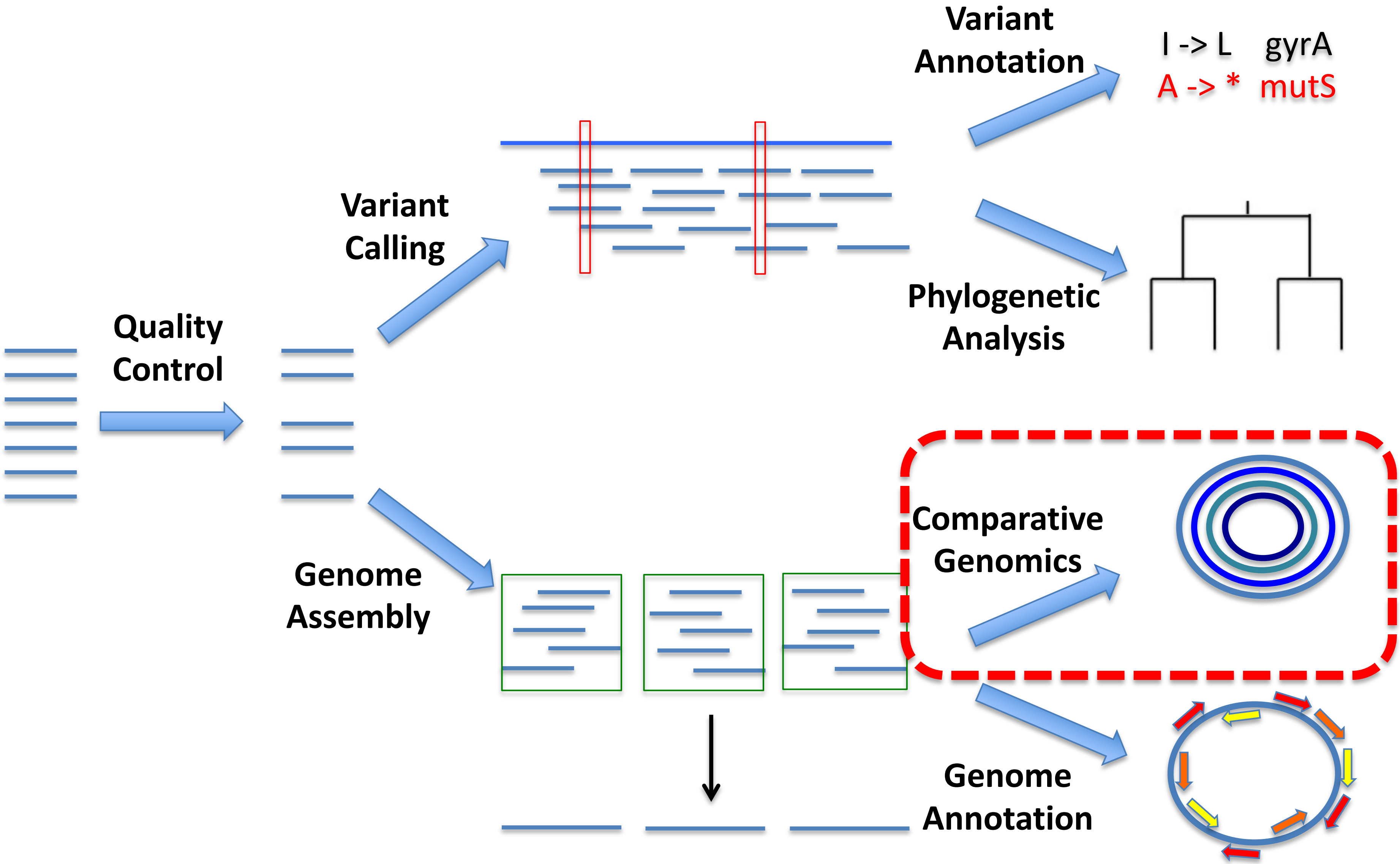
Fasta file

Genbank file

Genbank files, alignment files

Day 2 afternoon– Comparative genomics

Mile-high view of a genomics pipeline



Quality Control

Variant Calling

Genome Assembly

Variant Annotation

I -> L gyrA

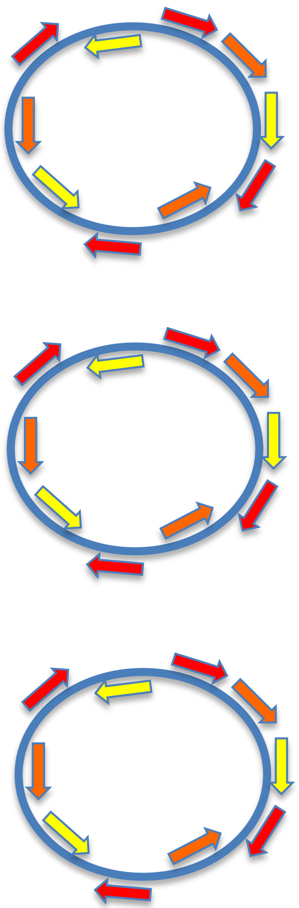
A -> * mutS

Phylogenetic Analysis

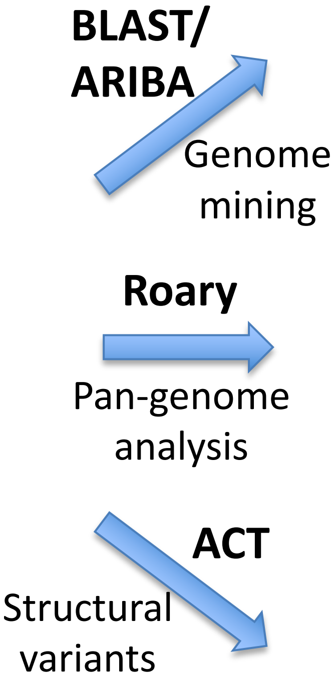
Comparative Genomics

Genome Annotation

Comparative genomics



Fasta, genbank and/or pep



BLAST/ ARIBA

Genome mining

Roary

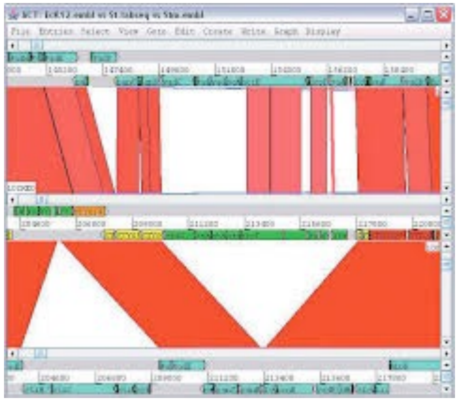
Pan-genome analysis

ACT

Structural variants

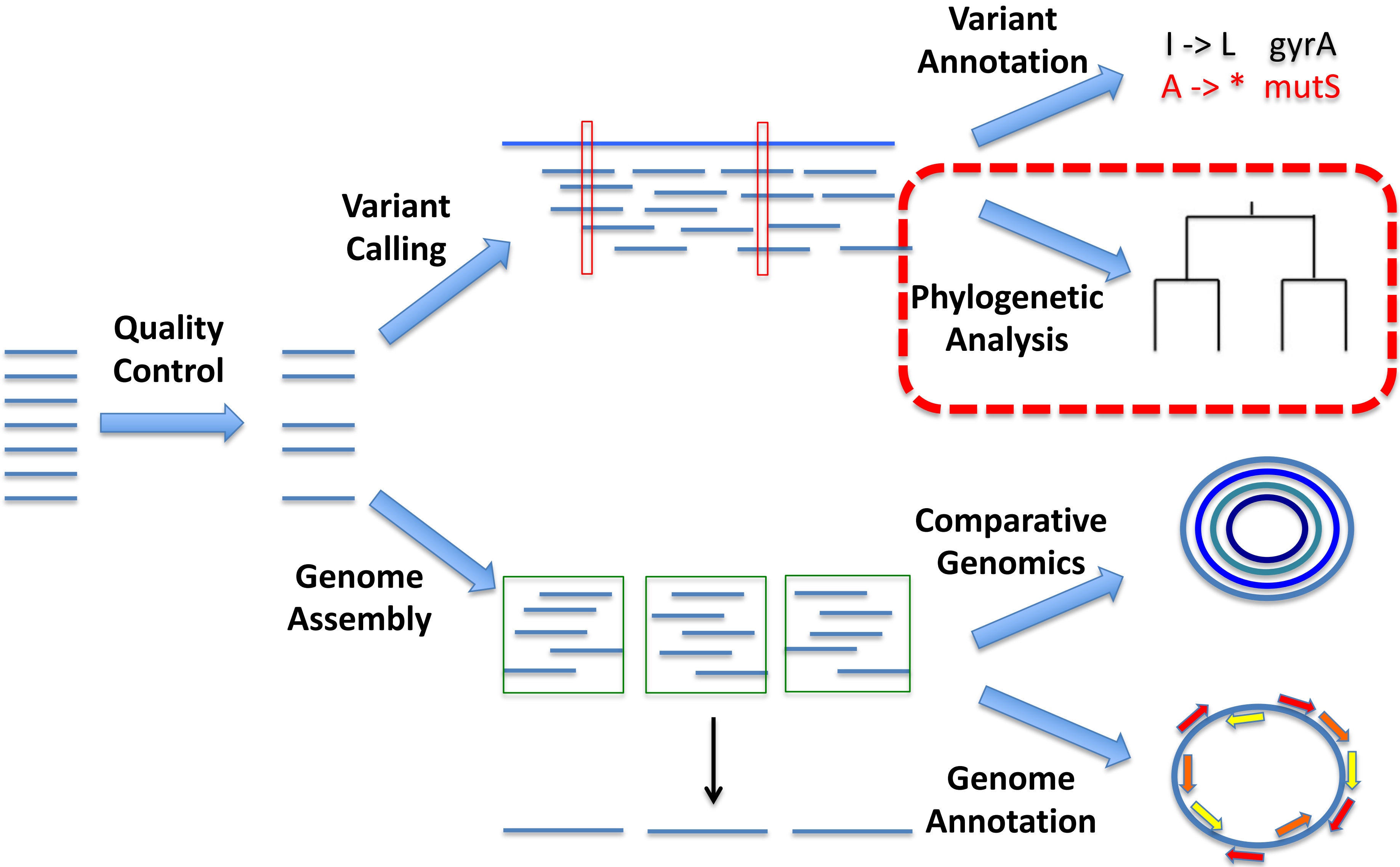
	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5
Genome 1	⬜	🟥	🟥	⬜	🟥
Genome 2	🟥	🟥	⬜	⬜	🟥
Genome 3	⬜	🟥	🟥	🟥	⬜

Genome 1	⬜	🟥	🟥	⬜	🟥	⬜	🟥	🟥	⬜	🟥	⬜	🟥	🟥	⬜	🟥	⬜	🟥	🟥	⬜	🟥	⬜	🟥	🟥	⬜	🟥	⬜	🟥	🟥	⬜	🟥
Genome 2	🟥	🟥	⬜	⬜	🟥	🟥	🟥	⬜	⬜	🟥	🟥	🟥	⬜	⬜	🟥	🟥	🟥	⬜	⬜	🟥	🟥	🟥	⬜	⬜	🟥	🟥	🟥	⬜	⬜	🟥
Genome 3	⬜	🟥	🟥	🟥	⬜	⬜	🟥	🟥	🟥	⬜	⬜	🟥	🟥	🟥	⬜	⬜	🟥	🟥	🟥	⬜	⬜	🟥	🟥	🟥	⬜	⬜	🟥	🟥	🟥	⬜



Day 3 morning – Basic phylogenetic analysis

Mile-high view of a genomics pipeline



Quality Control

Variant Calling

Variant Annotation

Phylogenetic Analysis

Genome Assembly

Comparative Genomics

Genome Annotation

Phylogenetics

>Genome 1
ATCGTCGTGCTGC
TGCTGTCGTGCTG

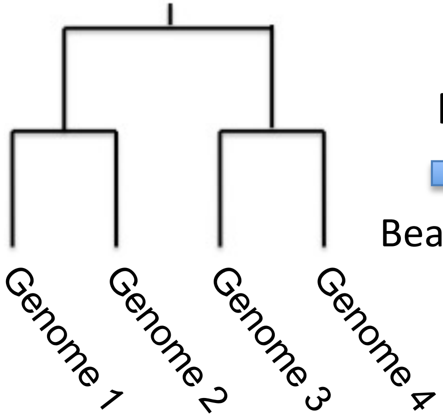
>Genome 2
CAGTGCATGTGCT
AGACTGTCGATGC
TA

>Genome 3
AGCTGTACCGATG
ACTGCTGACTGAC

Multi-fasta file

ape

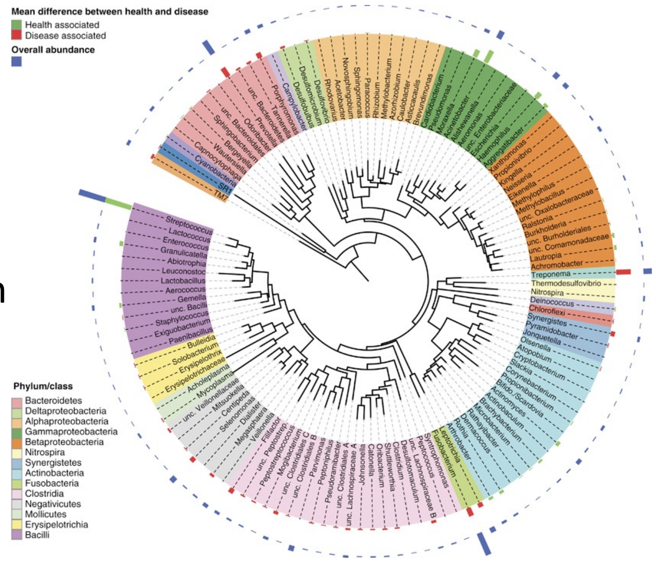
Tree construction



Nexus file

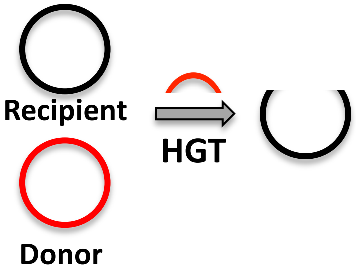
R/iTOL

Beautification



Gubbins

Recombination filtering



Gubbins/ape

Tree construction