Bacterial Genomics Workshop

March 2021

Goals of workshop

- Get an overview of steps in microbial genomics pipeline

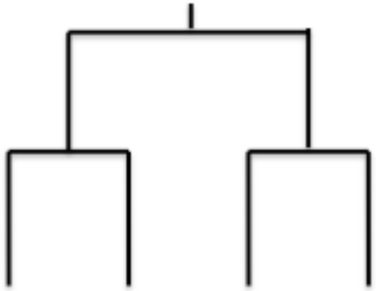
- Get exposure to common file formats and terminology in genomics

- Get hands on experience with a set of tools that could compose a genomics pipeline

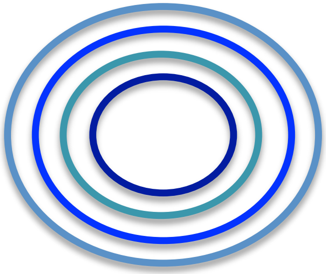
- Get experience working in a high-performance computing environment

So you want to sequence some bacteria?

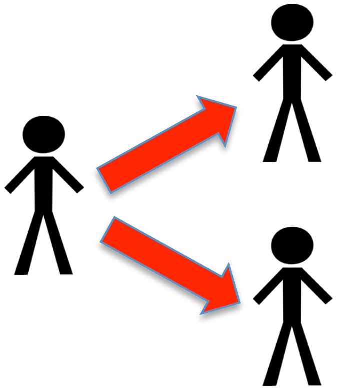
- Microbial phylogenetics



- Comparative genomics

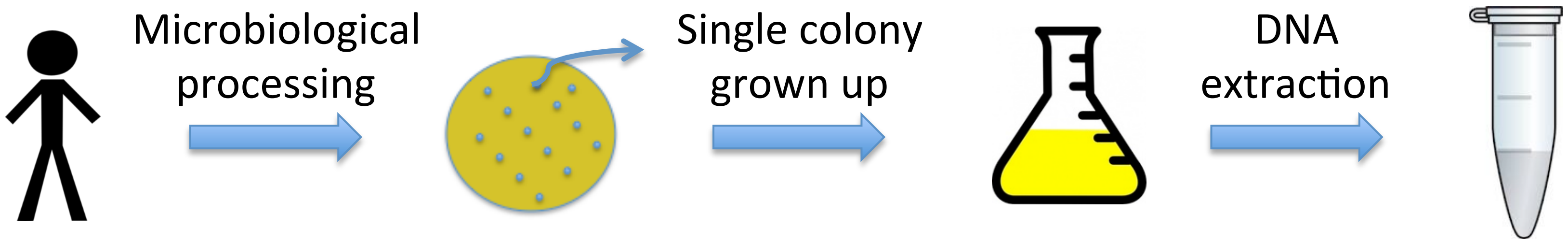


- Population genomics



DNA and library preparation

1. Sample Preparation

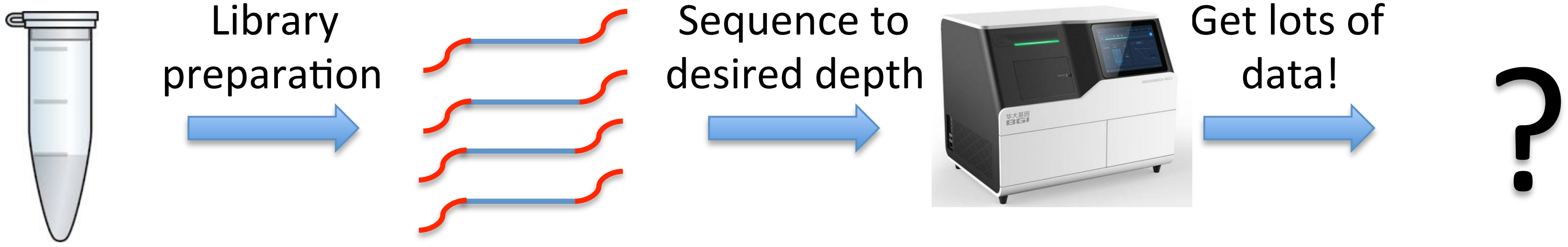


Microbiological processing

Single colony grown up

DNA extraction

2. Sequencing



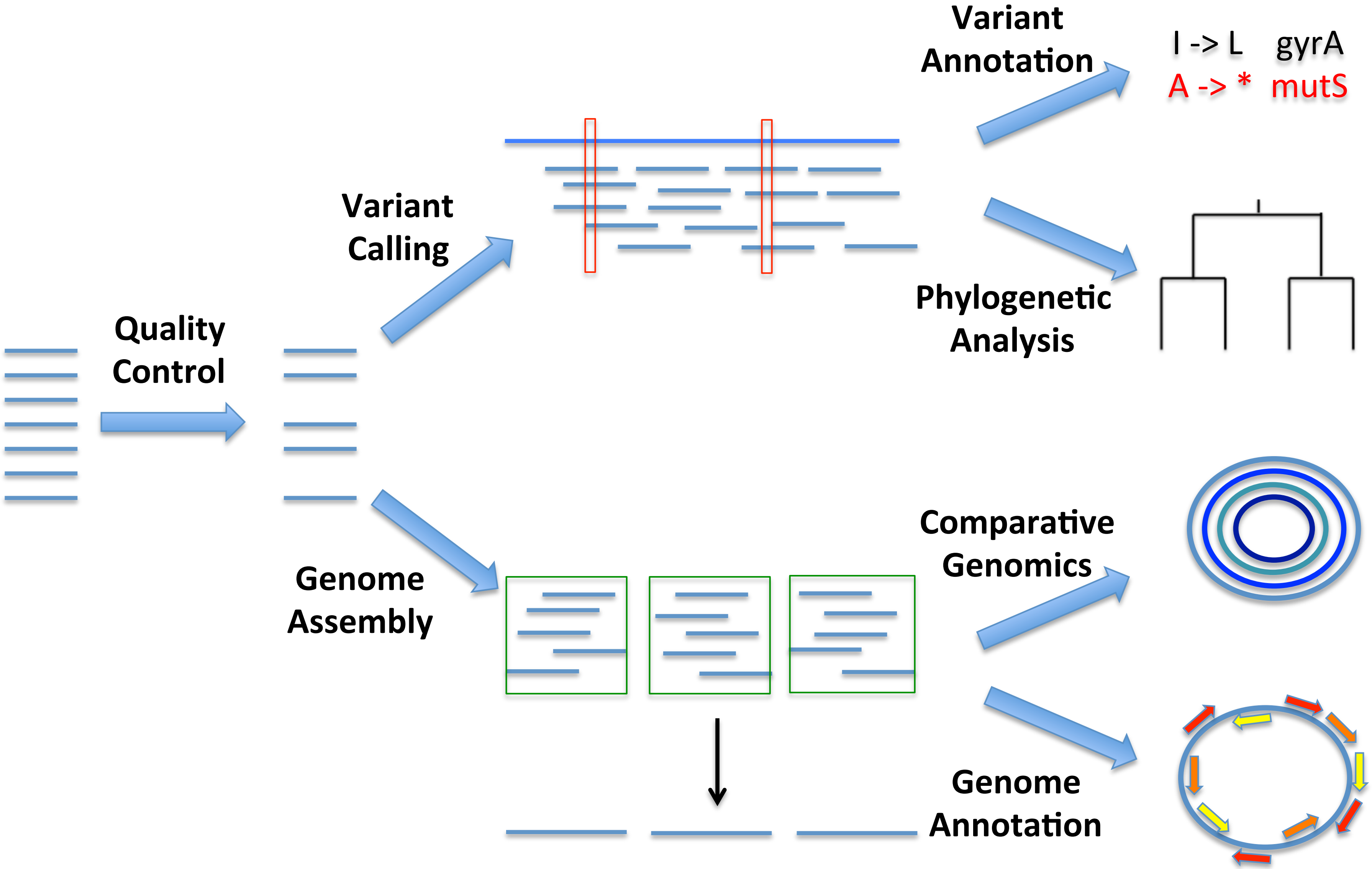
Library preparation

Sequence to desired depth

Get lots of data!

?

Mile-high view of a genomics pipeline



Sequencing quality control

Forward reads

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

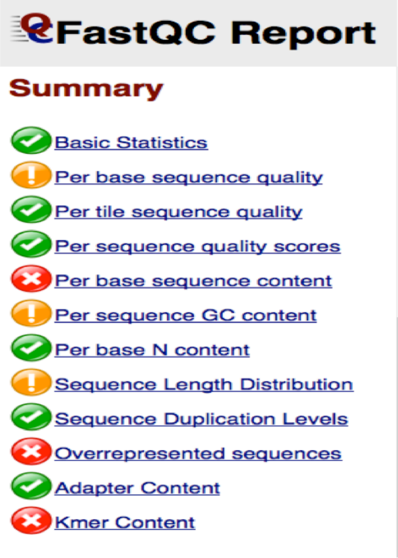
Reverse reads

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

Raw fastq files

FastQC

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

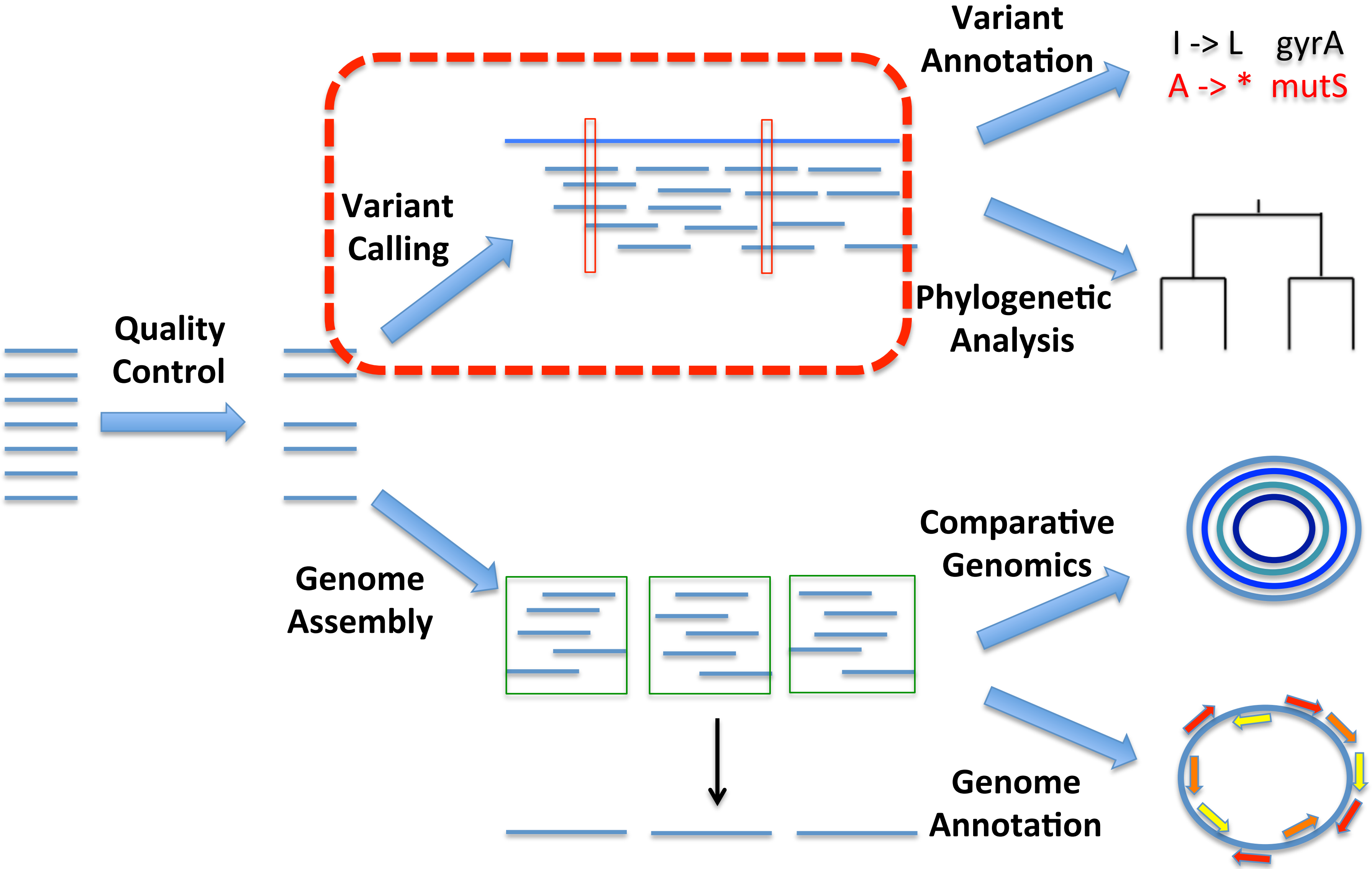
```
@seq1_1
ACTGCACT
+
8-8,,+@+
.
.
.
.
.
.
```

Reverse reads

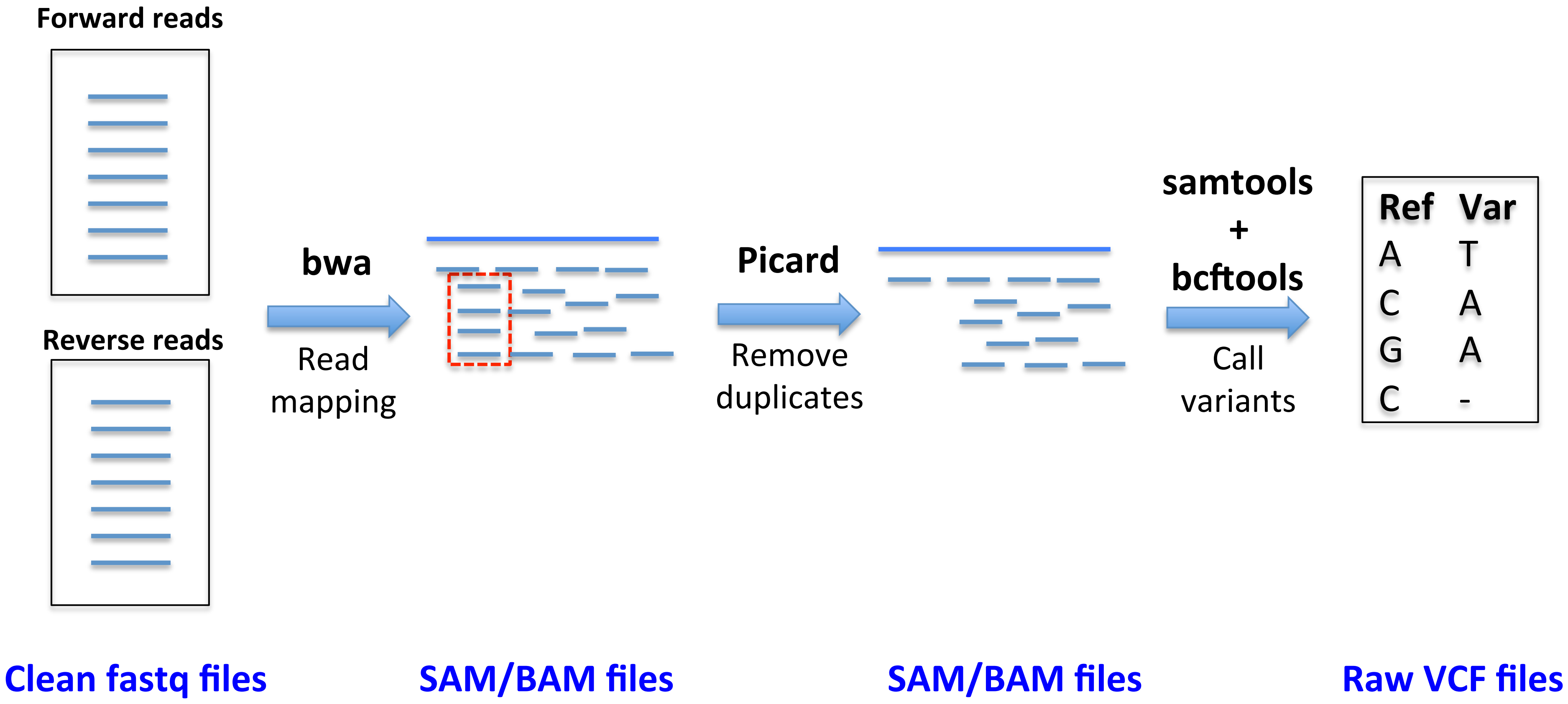
```
@seq1_2
TCTATCGA
+
A<-9BFCFF
.
.
.
.
.
.
```

Clean fastq files

Mile-high view of a genomics pipeline



Variant identification



Forward reads

Reverse reads

Clean fastq files

bwa

Read mapping

SAM/BAM files

Picard

Remove duplicates

SAM/BAM files

samtools
+
bcftools

Call variants

Ref	Var
A	T
C	A
G	A
C	-

Raw VCF files

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

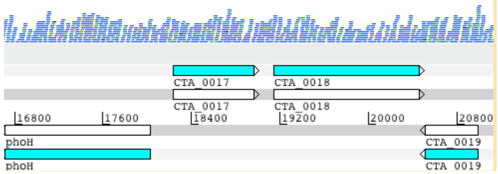
Variant annotation

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

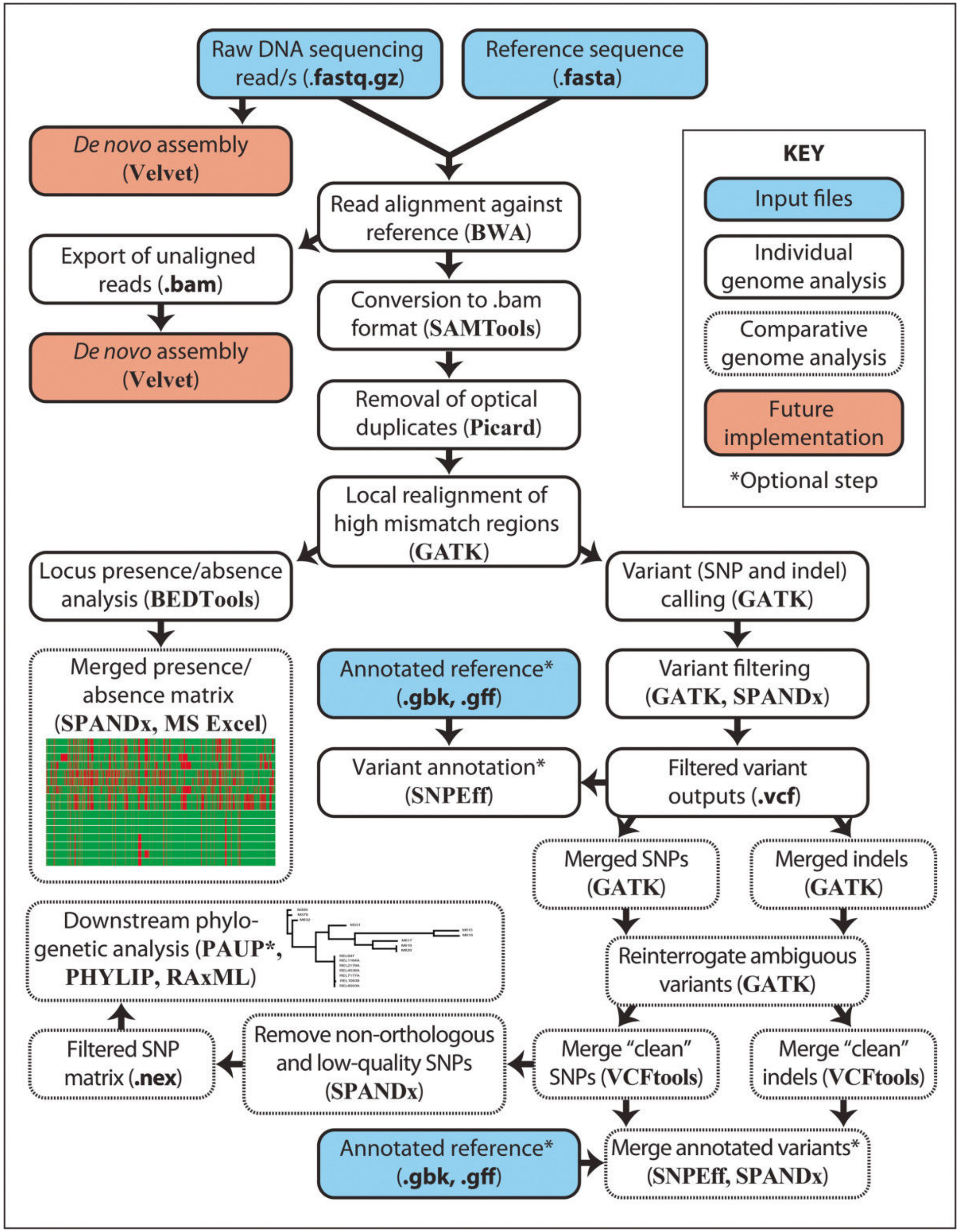
Annotated VCF files

Visualization

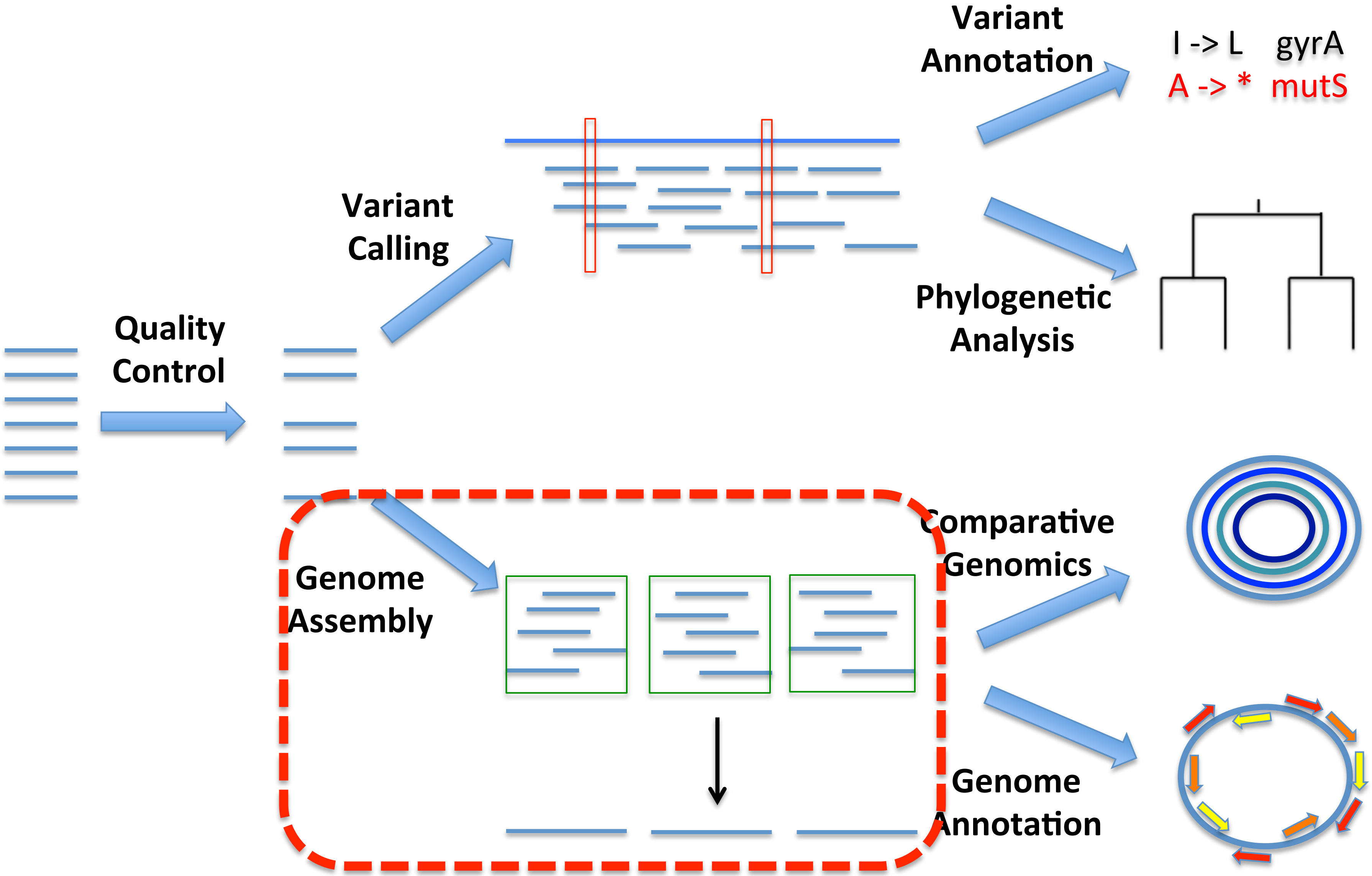
Artemis



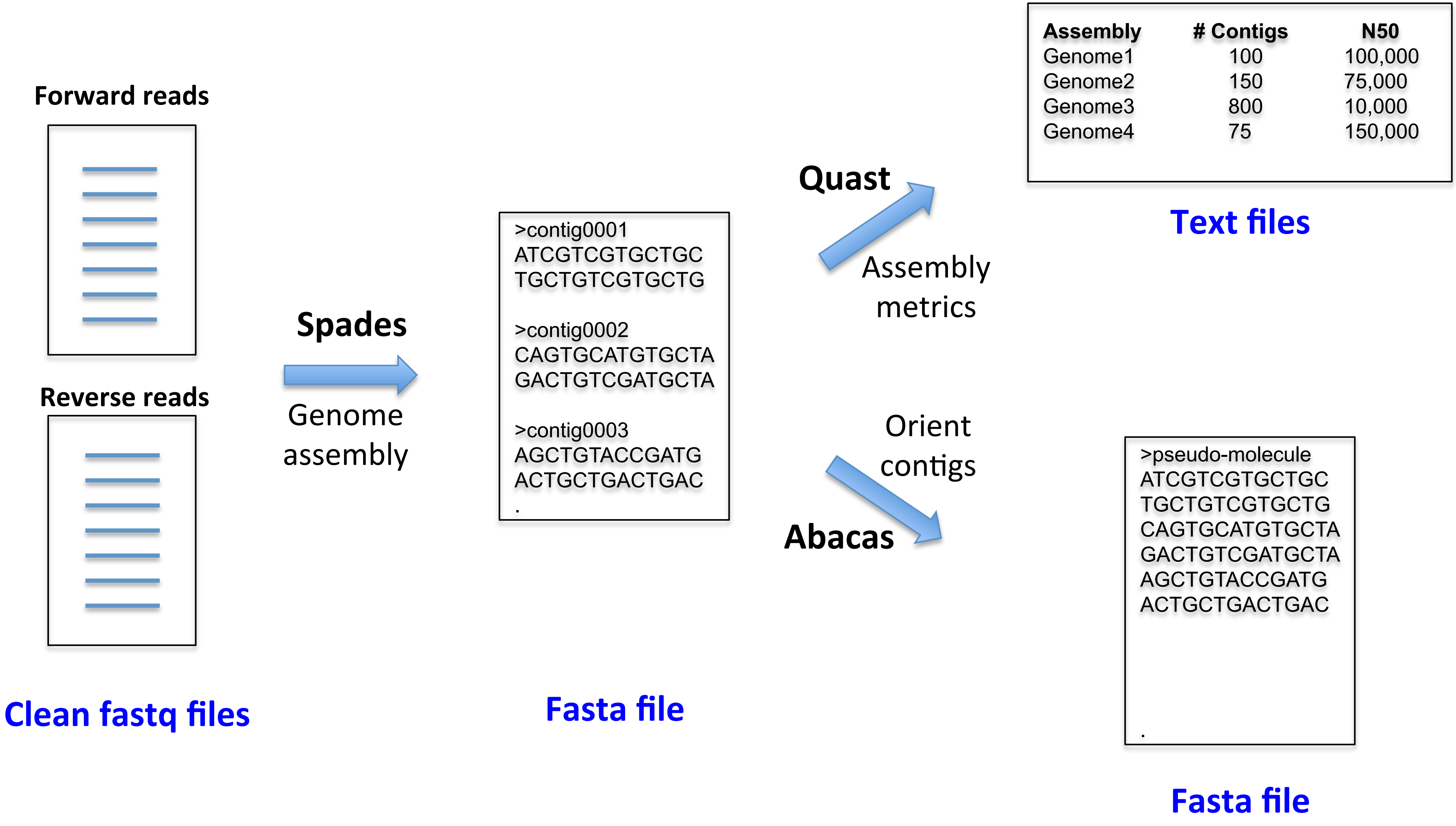
VCF, BAM, BAI, fasta files


Raw DNA sequencing read/s (.fastq.gz)
Reference sequence (.fasta)
De novo assembly (Velvet)
Read alignment against reference (BWA)
Export of unaligned reads (.bam)
Conversion to .bam format (SAMTools)
De novo assembly (Velvet)
Removal of optical duplicates (Picard)
Local realignment of high mismatch regions (GATK)
Locus presence/absence analysis (BEDTools)
Variant (SNP and indel) calling (GATK)
Merged presence/ absence matrix (SPANDx, MS Excel)
Annotated reference* (.gbk, .gff)
Variant filtering (GATK, SPANDx)
Variant annotation* (SNPEff)
Filtered variant outputs (.vcf)
Merged SNPs (GATK)
Merged indels (GATK)
Downstream phylogenetic analysis (PAUP*, PHYLIP, RAxML)
M300
M379
M632
M331
M615
M618
M617
M619
M620
REL607
REL1164A
REL2179A
REL4536A
REL7177A
REL10938
REL8593A
Reinterrogate ambiguous variants (GATK)
Filtered SNP matrix (.nex)
Remove non-orthologous and low-quality SNPs (SPANDx)
Merge "clean" SNPs (VCFtools)
Merge "clean" indels (VCFtools)
Annotated reference* (.gbk, .gff)
Merge annotated variants* (SNPEff, SPANDx)
KEY
Input files
Individual genome analysis
Comparative genome analysis
Future implementation
*Optional step

Mile-high view of a genomics pipeline



Genome assembly



Forward reads

Reverse reads

Clean fastq files

Spades

Genome assembly

```
>contig0001
ATCGTCGTGCTGC
TGCTGTCGTGCTG

>contig0002
CAGTGCATGTGCTA
GACTGTCGATGCTA

>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

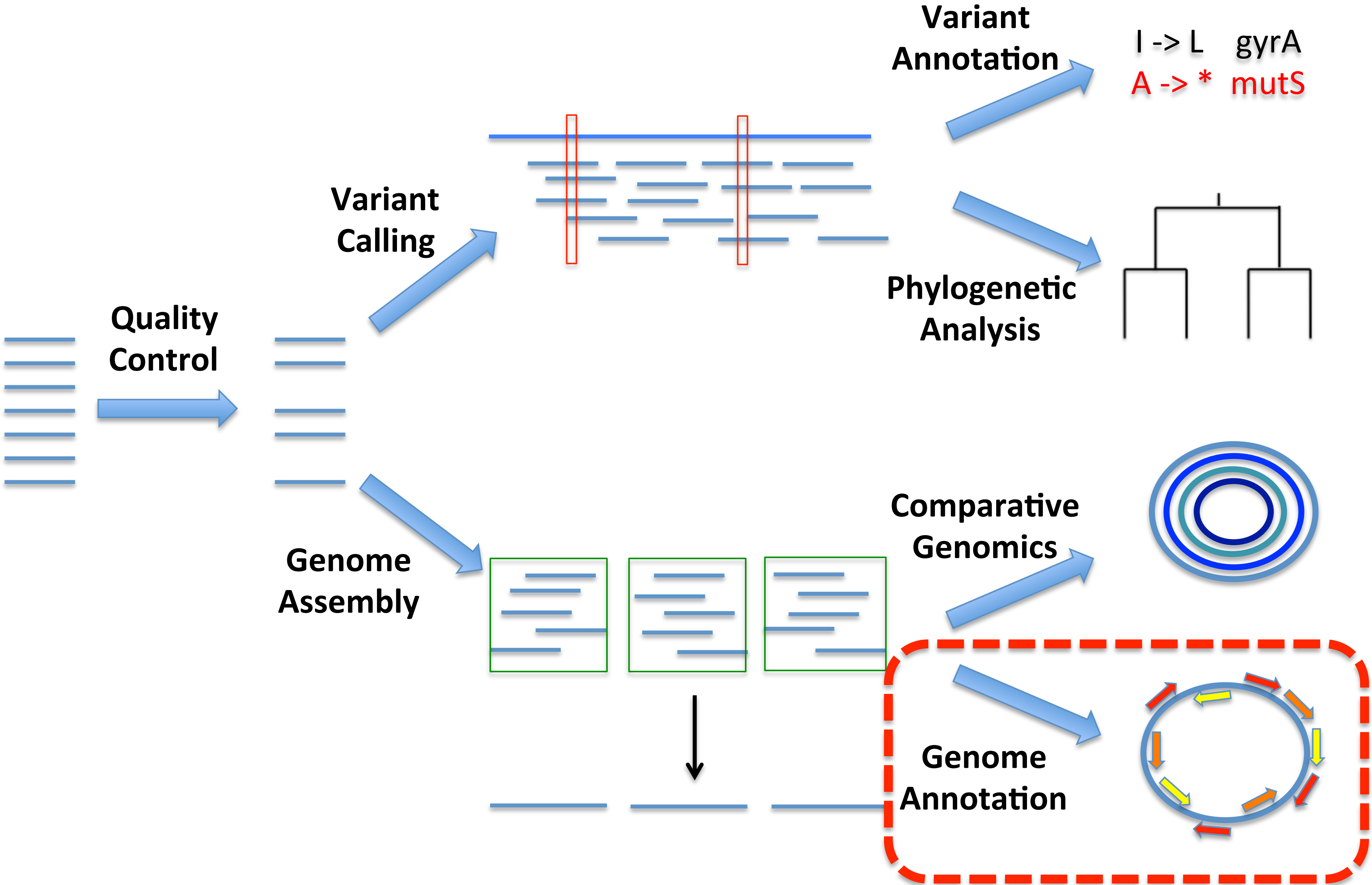
Orient contigs

Abacas

```
>pseudo-molecule
ATCGTCGTGCTGC
TGCTGTCGTGCTG
CAGTGCATGTGCTA
GACTGTCGATGCTA
AGCTGTACCGATG
ACTGCTGACTGAC



.
```

Fasta file

Mile-high view of a genomics pipeline

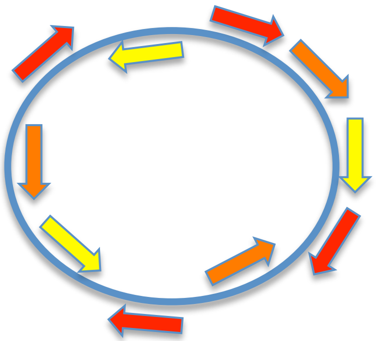


Genome annotation

```
>pseudo-molecule
ATCGTCGTGCTGC
TGCTGTCGTGCTG
CAGTGCATGTGCTA
GACTGTCGATGCTA
AGCTGTACCGATG
ACTGCTGACTGAC



.
```

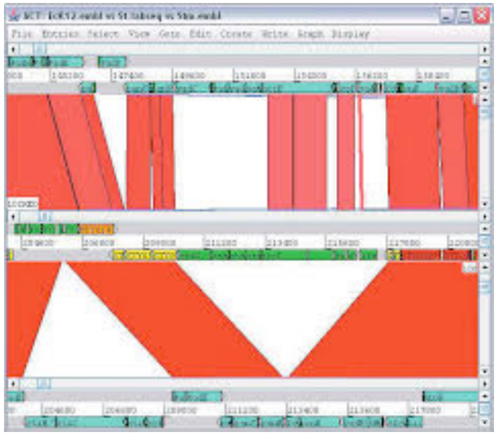
Prokka

1) Gene finding
2) Basic annotation



ACT

Visualization

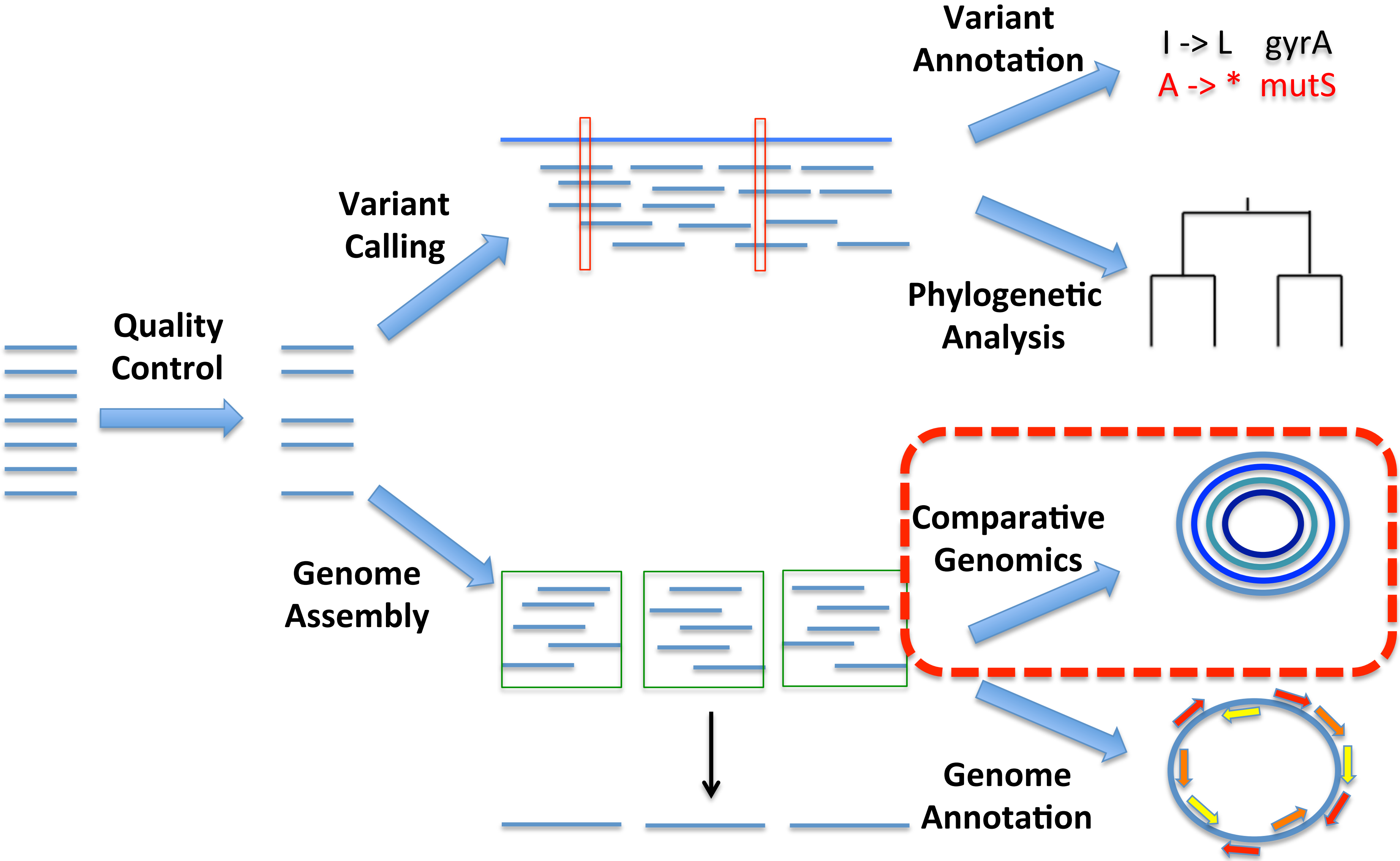

File Entries Select View Goto Edit Create Write Graph Display

Fasta file

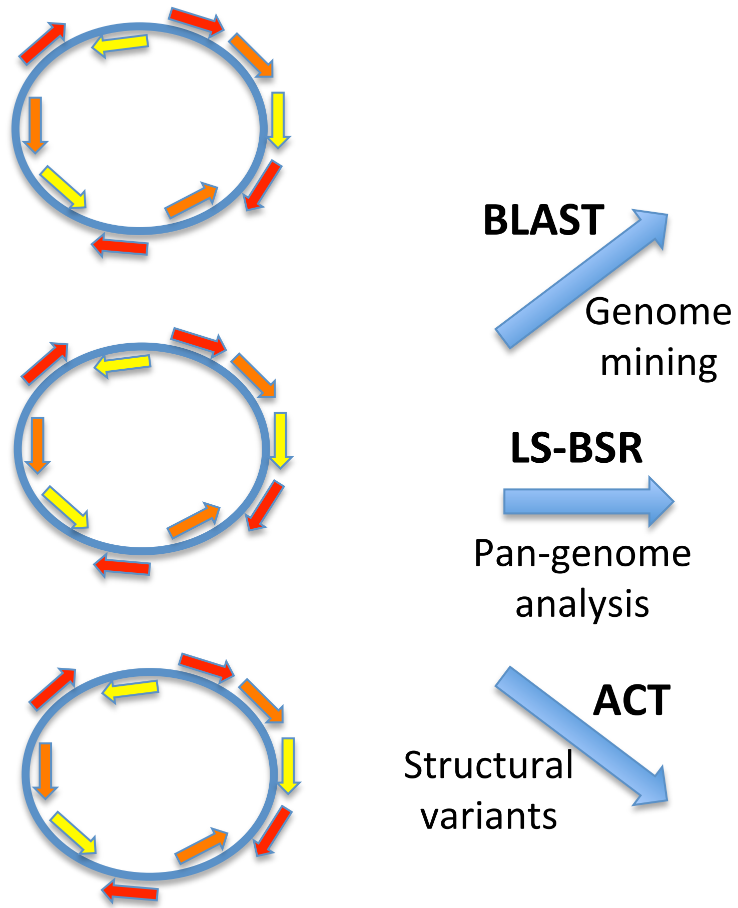
Genbank file

Genbank files, alignment files

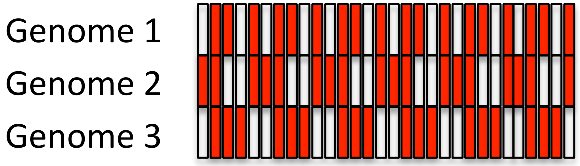
Mile-high view of a genomics pipeline

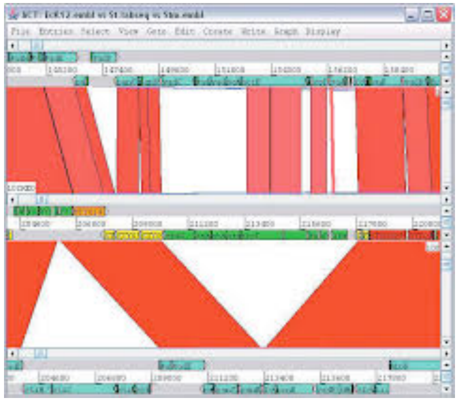


Comparative genomics



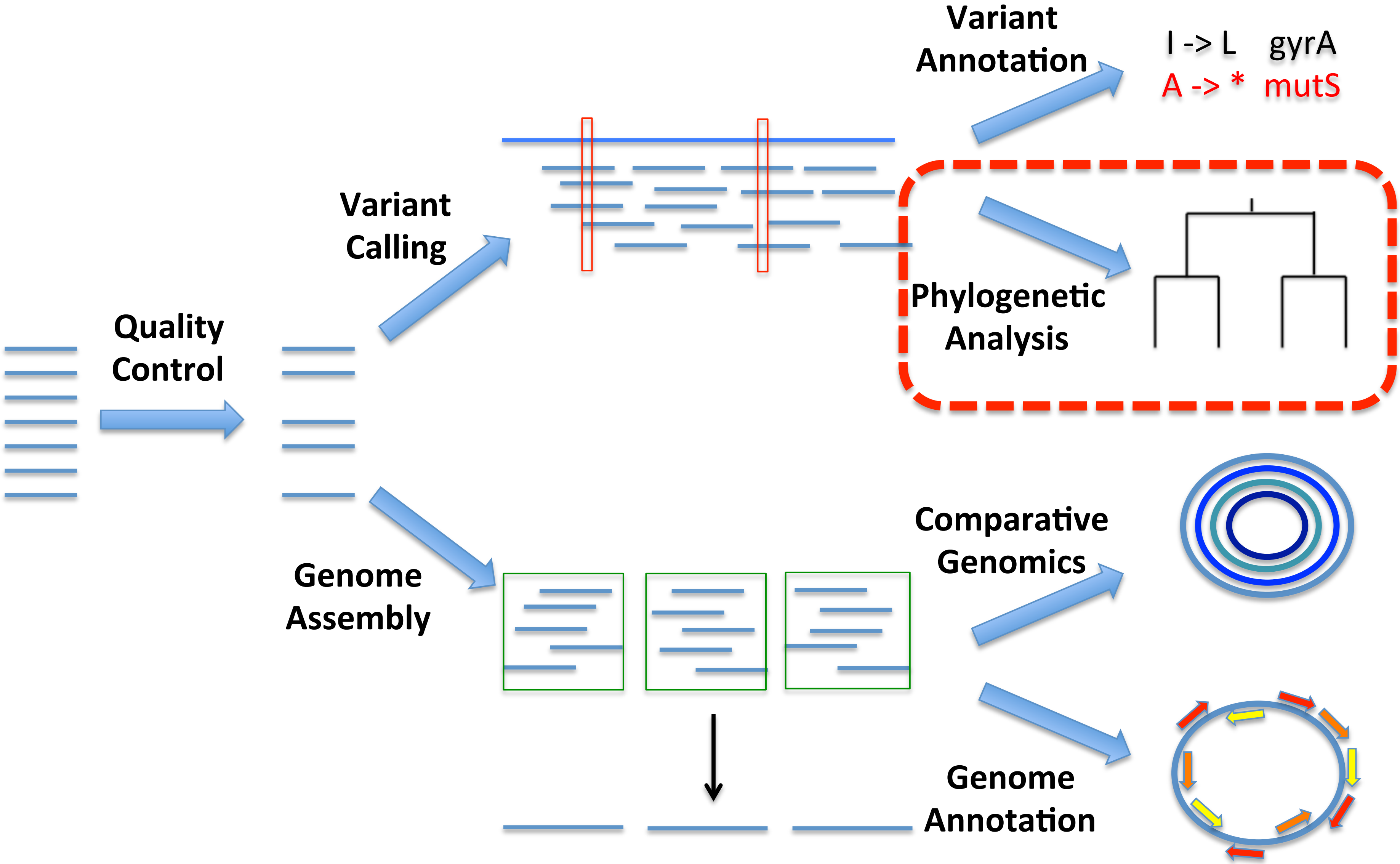
	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5
Genome 1	⬜	🟥	🟥	⬜	🟥
Genome 2	🟥	🟥	⬜	⬜	🟥
Genome 3	⬜	🟥	🟥	🟥	⬜





Fasta, genbank and/or pep

Mile-high view of a genomics pipeline



Quality Control

Variant Calling

Variant Annotation

I -> L gyrA

A -> * mutS

Phylogenetic Analysis

Genome Assembly

Comparative Genomics

Genome Annotation

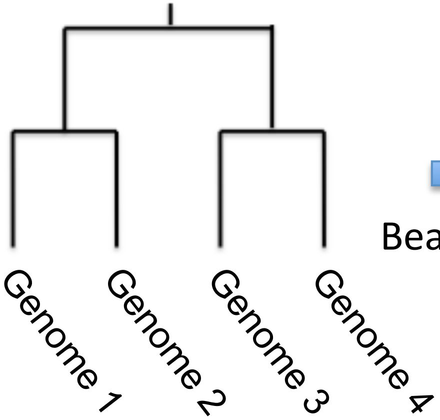
Phylogenetics

```
>Genome 1
ATCGTCGTGCTGC
TGCTGTCGTGCTG

>Genome 2
CAGTGCATGTGCTA
GACTGTCGATGCTA

>Genome 3
AGCTGTACCGATG
ACTGCTGACTGAC
.
```

Multi-fasta file

Orthofiner
iqtree

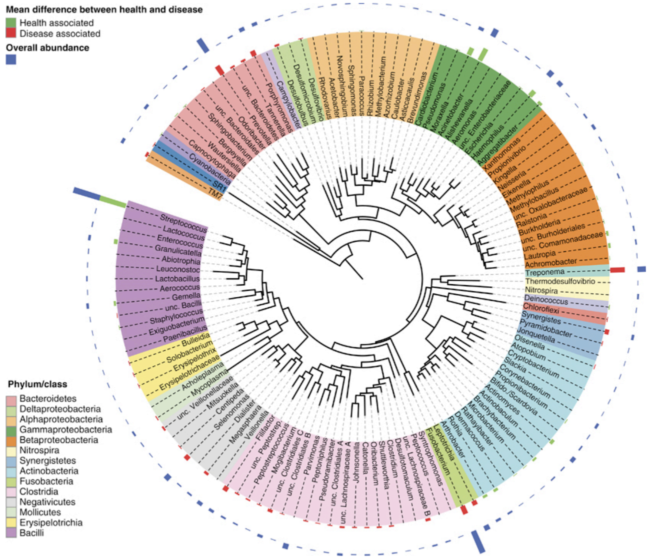
Tree construction



Nexus file

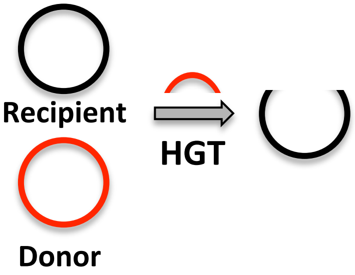
iTOL

Beautification



Gubbins

Recombination filtering



Seaview /ape

Tree construction