Analyzing Microbial Communities Using Next Generation Sequencing

PART I: Basic Concepts, Databases and Data types

Rounak Feigelman, Ph.D.
Senior Scientist, Paragon Genomics Inc.

Introduction: Human Microbiome

Seeded at birth

Determines founding microbiota

Impacted by Host Genetics

Variants in genes affect composition

Shaped by Diet

Breastfeeding shapes early microbiome

Affected by lifestyle and aging

Hygiene practices, puberty ,stress exert selective pressure

Aids in digestion

Assists in metabolizing nutrients

Resistance against invasive microbes

Competes for primary nutrition sources

Secretes growth inhibitors

Fortifies immune system

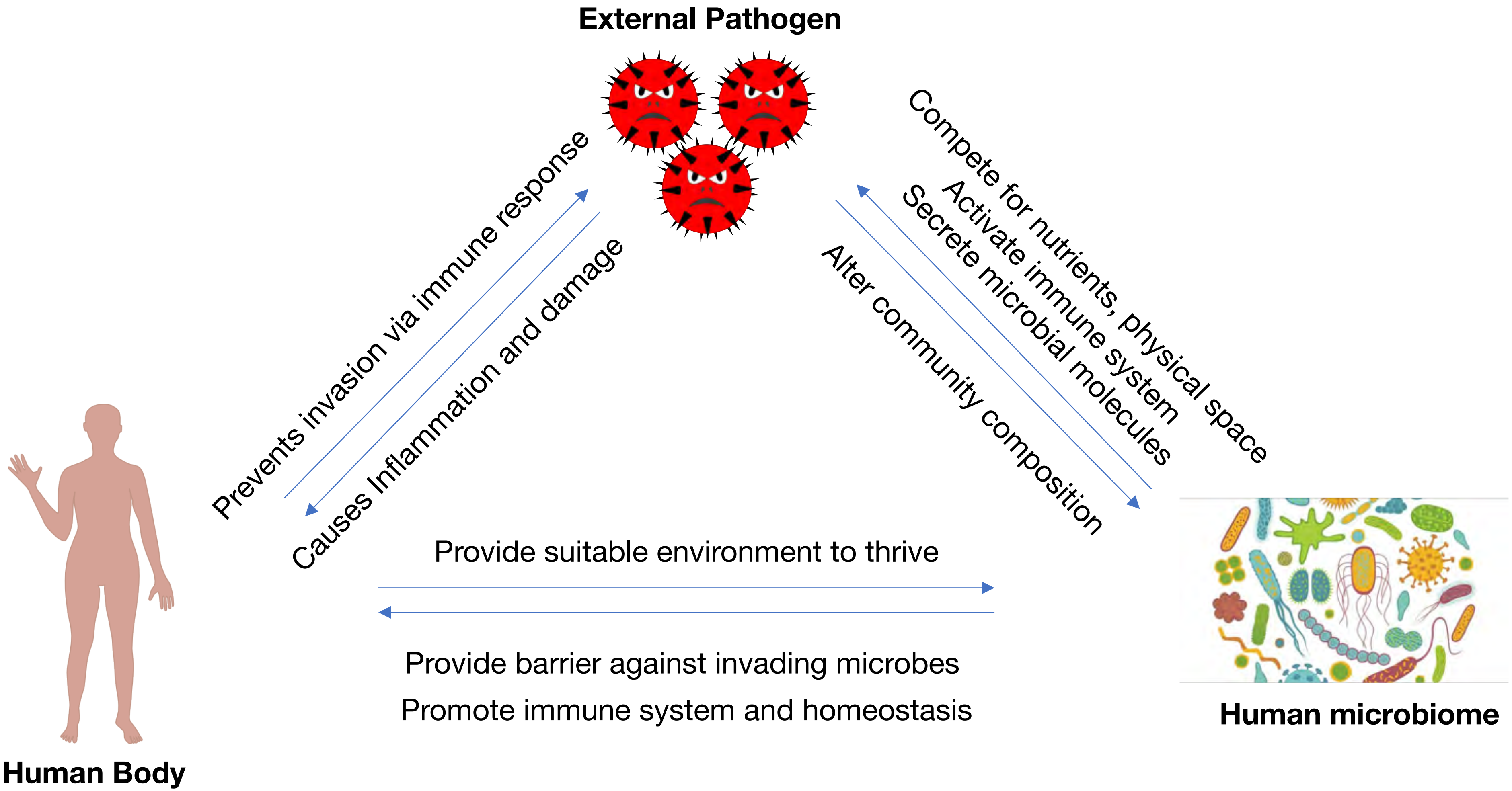
Induces immune response to inhibit colonization

Modulates behavior

Secrete signaling molecules allowing cross talk between gut and brain

Figure adapted from PMID: 26540050

Role of Microbiome in Infection



External Pathogen

Prevents invasion via immune response

Causes Inflammation and damage

Compete for nutrients, physical space

Activate immune system

Secrete microbial molecules

Alter community composition

Provide suitable environment to thrive

Provide barrier against invading microbes

Promote immune system and homeostasis

Human Body

Human microbiome

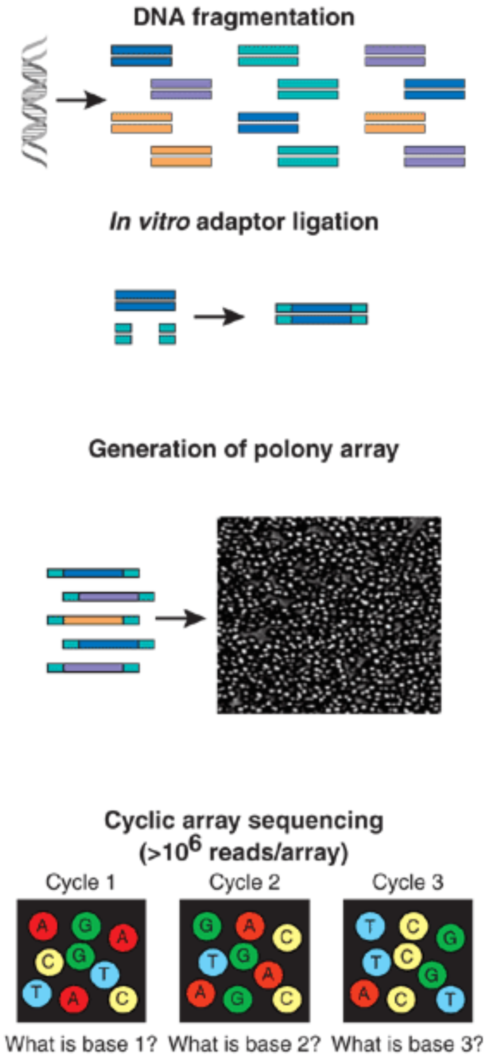
Current Strategies for Pathogen Identification

1. Laboratory culture of biological sample (mucus, stool, etc.)
   + Antibiotic sensitivity
   + Rapid turn around time, molecular diagnostic assays
   - Low detection rate
   - Scales with the no. of pathogens (one bug, one test)
   - Miss slow growing pathogens

2. Next Generation Sequencing diagnostic assays
   + Enables detection of broad range of pathogens, co-infections
   + Enables microbiome characterization
   + Utility in difficult to diagnose cases or immunocompromised patients
   - Data needs analysis and interpretation in clinical context
   - Slow turn around time
   - Require investment in infrastructure for data analysis and storage

Figure adapted from PMID: 18576944

Next Generation Sequencing



Fragments are of 200 – 600 bp length

Adapters are added to each sequence

Adapter ligated sequences are spatially fixed or isolated for clonal amplification

- Feature generation step

Localized clones are read using fluorochromes or change in pH

- Millions of sequences are read in parallel
- Principle is “sequencing by synthesis”
- Signal to noise ratio determines read length

Next Generation Sequencing Technologies

NGS Technology related specifics

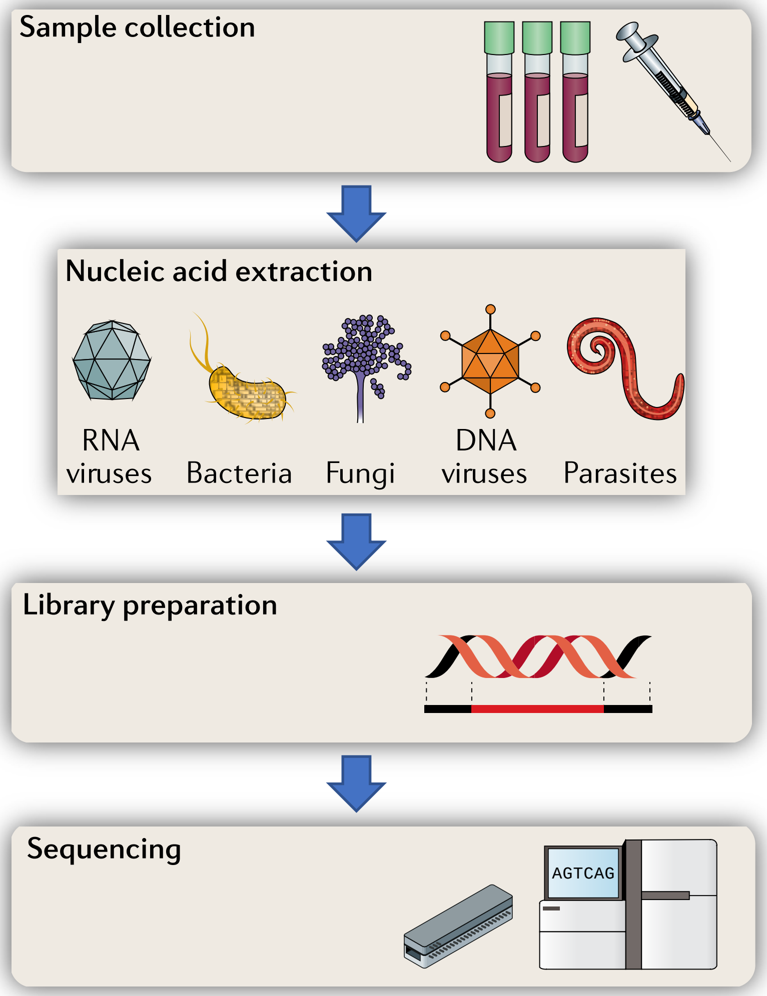
Sequencer	Feature generation	Synthesis mechanism	Read (bp)	Error type
454	Emulsion PCR	Pyrosequencing, PCR	700	Inert-deletion
Illumina	Bridge PCR	Reversible terminators polymerase	150*2	Substitution
SOLiD	Emulsion PCR	ligase	60*2	Substitution
PacBio	Single molecule	Polymerase	1500	Deletion

Next Generation Sequencing Approaches in Clinical Microbiology

Sequencing Method	Potential Application	Type of Data Generated
Amplicon sequencing (universal primer)	Multiplex pathogen detection	16S rRNA gene segments
Amplicon sequencing (targeted primer)	Pathogen identification	Viral genome recovery, variant detection
Capture probe enrichment	Multiplex pathogen detection	Viral genome recovery, variant detection
Untargeted Whole Genome Sequencing (deplete host DNA)	Analyze microbial community	Gene sequences from different members of microbial community.
Untargeted Whole Genome Sequencing (without depletion of host DNA)	Exploratory data	Majority data from host genome with some microbial data

Workflow

Figure adapted from PMID: 30918369



Sample collection

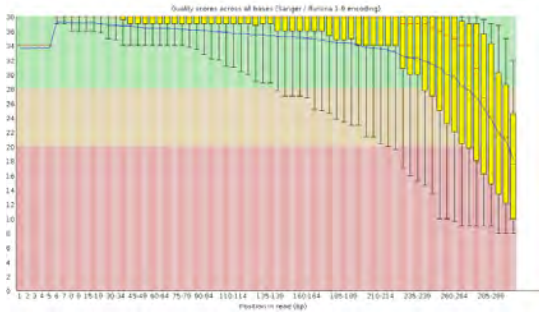
Nucleic acid extraction

Library preparation

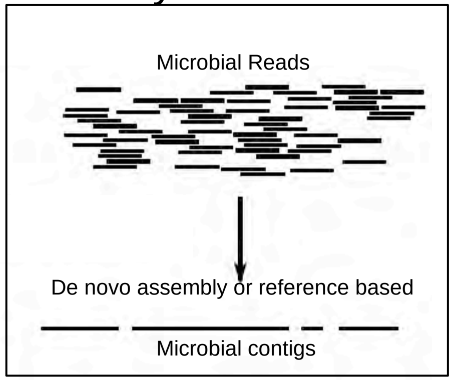
Sequencing

Data Preparation

1. Quality control and read trim



2. Assembly of microbial reads



Data Analysis Workflow

Taxonomic characterization

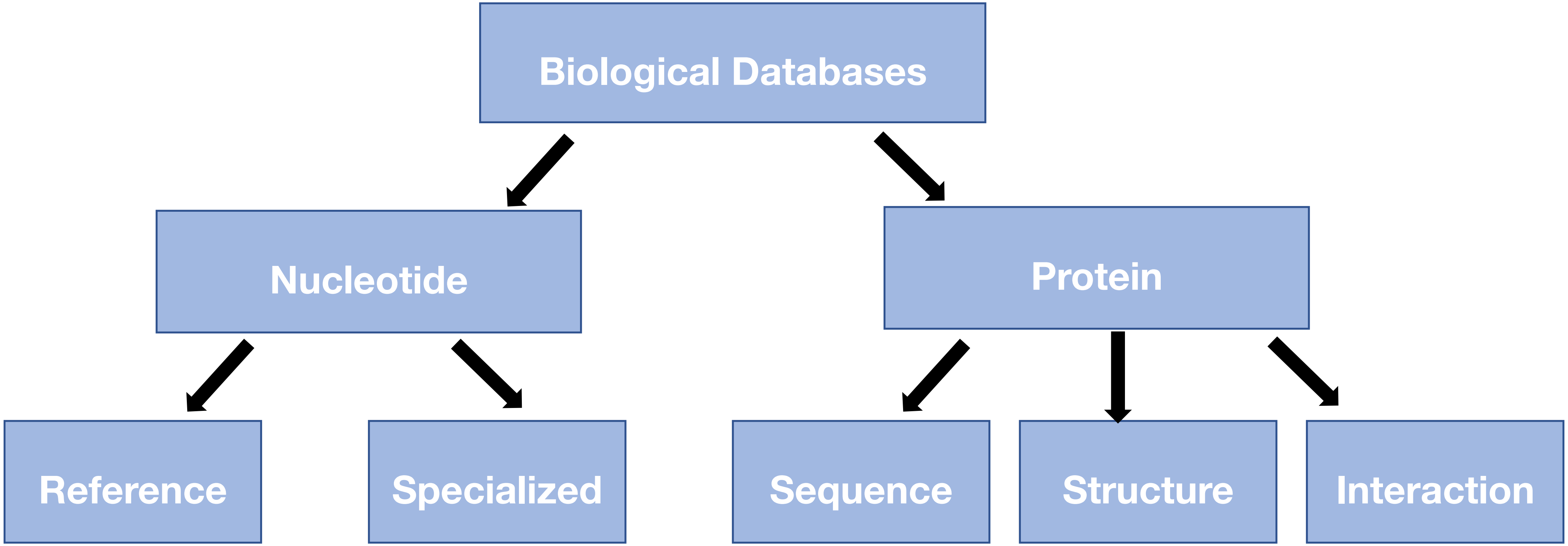
Gene prediction

Gene Annotation

Identify adaptations

Genome reconstruction of “clonal” microbes

Bioinformatics Databases



Biological Databases

Nucleotide

Protein

Reference

Specialized

Sequence

Structure

Interaction

- NCBI
- EMBL
- DNA Data Bank Japan

- ENSEMBL

- UniProt
- PIR
- SwissProt

- PDB

- STRINGdb

Microbial Databases

1. **DNA and protein sequence databases (primary and secondary)**

NCBI
JGI
IMG/M
INTEGRATED MICROBIAL GENOMES & MICROBIOMES
UniProt
PATRIC 3.6.7
PubMLST
Public databases for molecular typing and microbial genome diversity
BacMap
genome atlas
600 kbp
50 kbp
ViPR
Virus Pathogen Resource

2. **Functional databases**

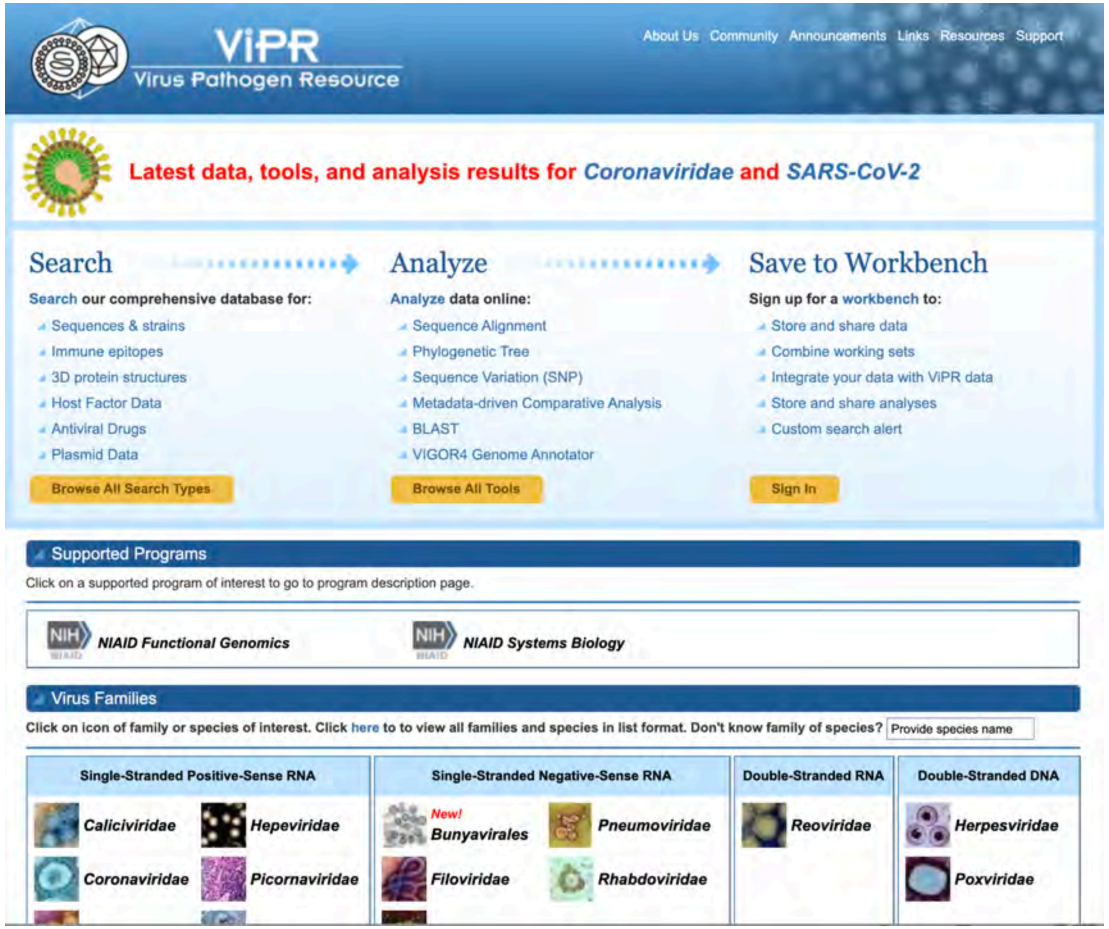
KEGG
Kyoto Encyclopedia of Genes and Genomes
EggNOG 5.0.0
microme
The Comprehensive Antibiotic Resistance Database

Analysis Tools and Software

Taxonomic annotation, gene prediction and functional annotation tools for DNA and protein sequences

Microbial Nucleotide BLAST
GLIMMER
Microbial Gene-Finding System
GeneMark
GENEONTOLOGY
Unifying Biology
ORFfinder
InterPro
Classification of protein families
MG-RAST
metagenomics analysis server
version 4.0.3
cite us

Common Databases



ViPR Virus Pathogen Resource

About Us Community Announcements Links Resources Support

Latest data, tools, and analysis results for *Coronaviridae* and *SARS-CoV-2*

Search

Search our comprehensive database for:

- Sequences & strains
- Immune epitopes
- 3D protein structures
- Host Factor Data
- Antiviral Drugs
- Plasmid Data

Browse All Search Types

Analyze

Analyze data online:

- Sequence Alignment
- Phylogenetic Tree
- Sequence Variation (SNP)
- Metadata-driven Comparative Analysis
- BLAST
- VIGOR4 Genome Annotator

Browse All Tools

Save to Workbench

Sign up for a workbench to:

- Store and share data
- Combine working sets
- Integrate your data with ViPR data
- Store and share analyses
- Custom search alert

Sign In

Supported Programs

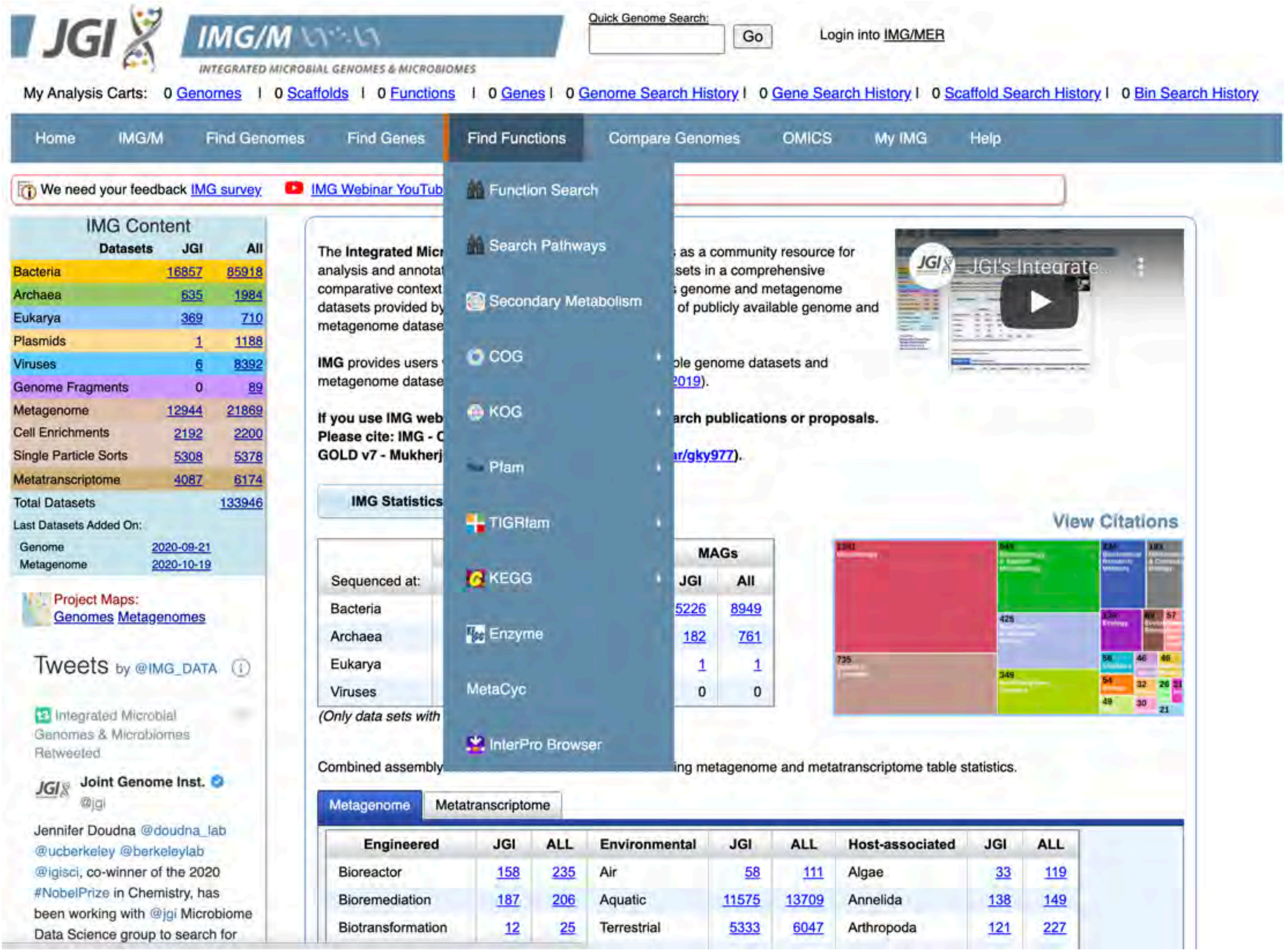
Click on a supported program of interest to go to program description page.

NIAID Functional Genomics NIAID Systems Biology

Virus Families

Click on icon of family or species of interest. Click here to to view all families and species in list format. Don't know family of species? Provide species name

Single-Stranded Positive-Sense RNA		Single-Stranded Negative-Sense RNA		Double-Stranded RNA	Double-Stranded DNA
Caliciviridae	Hepeviridae	New! Bunyavirales	Pneumoviridae	Reoviridae	Herpesviridae
Coronaviridae	Picornaviridae	Filoviridae	Rhabdoviridae		Poxviridae



JGI IMG/M INTEGRATED MICROBIAL GENOMES & MICROBIOMES

Quick Genome Search: Go Login into IMG/MER

My Analysis Carts: 0 Genomes | 0 Scaffolds | 0 Functions | 0 Genes | 0 Genome Search History | 0 Gene Search History | 0 Scaffold Search History | 0 Bin Search History

Home IMG/M Find Genomes Find Genes Find Functions Compare Genomes OMICS My IMG Help

We need your feedback IMG survey IMG Webinar YouTub

Function Search
Search Pathways
Secondary Metabolism
COG
KOG
Pfam
TIGRfam
KEGG
Enzyme
MetaCyc
InterPro Browser

IMG Content

Datasets	JGI	All
Bacteria	16857	85918
Archaea	635	1984
Eukarya	369	710
Plasmids	1	1188
Viruses	6	8392
Genome Fragments	0	89
Metagenome	12944	21869
Cell Enrichments	2192	2200
Single Particle Sorts	5308	5378
Metatranscriptome	4087	6174
Total Datasets		133946

Last Datasets Added On:

Genome	2020-09-21
Metagenome	2020-10-19

Project Maps: Genomes Metagenomes

Tweets by @IMG_DATA

Integrated Microbial Genomes & Microbiomes Retweeted

Joint Genome Inst. @jgi

Jennifer Doudna @doudna_lab @ucberkeley @berkeleylab @igisci, co-winner of the 2020 #NobelPrize in Chemistry, has been working with @jgi Microbiome Data Science group to search for

The **Integrated Micr** … as a community resource for analysis and annotat … sets in a comprehensive comparative context … genome and metagenome datasets provided by … of publicly available genome and metagenome datase

IMG provides users … ole genome datasets and metagenome datase … 2019).

If you use IMG web … arch publications or proposals.
Please cite: IMG - C
GOLD v7 - Mukherj … ar/gky977).

JGI's Integrate…

IMG Statistics

View Citations

Sequenced at:	MAGs JGI	MAGs All
Bacteria	5226	8949
Archaea	182	761
Eukarya	1	1
Viruses	0	0

(Only data sets with

Combined assembly … ing metagenome and metatranscriptome table statistics.

Metagenome Metatranscriptome

Engineered	JGI	ALL	Environmental	JGI	ALL	Host-associated	JGI	ALL
Bioreactor	158	235	Air	58	111	Algae	33	119
Bioremediation	187	206	Aquatic	11575	13709	Annelida	138	149
Biotransformation	12	25	Terrestrial	5333	6047	Arthropoda	121	227

Tools and databases are often integrated

Comprehensive Antibiotic Resistance Database (CARD)

Database to identify antibiotic resistance genes and related information

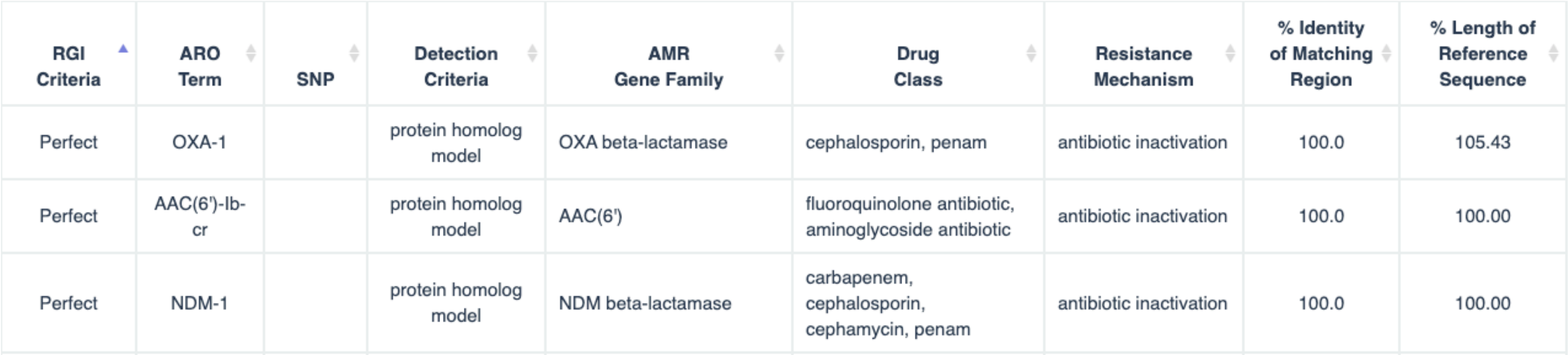
https://card.mcmaster.ca/analyze/rgi

- Accepts DNA or protein sequences
- Performs gene prediction and annotation using third party tools
- Uses curated sequences and detection models to annotate sample resistome

CARD Output

- Interactive sunburst visualizations and tables of predicted resistance genes, gene family, drug class, etc

CARD Result in tabular format



RGI Criteria	ARO Term	SNP	Detection Criteria	AMR Gene Family	Drug Class	Resistance Mechanism	% Identity of Matching Region	% Length of Reference Sequence
Perfect	OXA-1		protein homolog model	OXA beta-lactamase	cephalosporin, penam	antibiotic inactivation	100.0	105.43
Perfect	AAC(6')-Ib-cr		protein homolog model	AAC(6')	fluoroquinolone antibiotic, aminoglycoside antibiotic	antibiotic inactivation	100.0	100.00
Perfect	NDM-1		protein homolog model	NDM beta-lactamase	carbapenem, cephalosporin, cephamycin, penam	antibiotic inactivation	100.0	100.00

https://card.mcmaster.ca/home

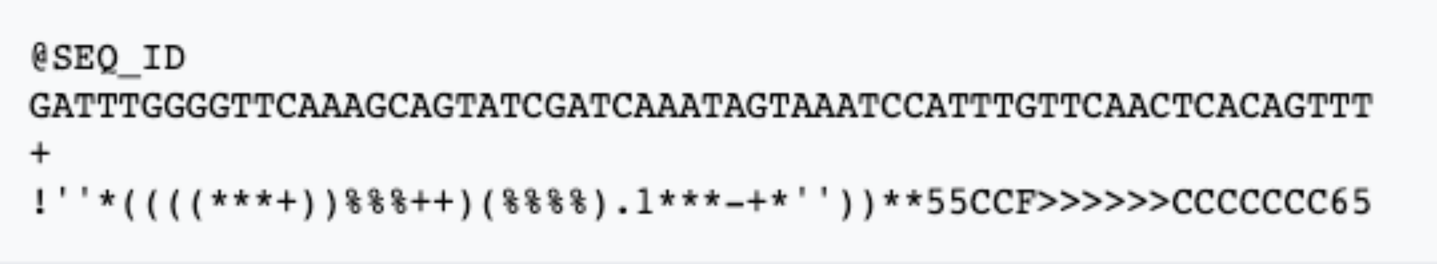
Virulence Factor Database (VFDB)

- Database providing classification of virulence factors present in bacterial pathogens
- http://www.mgc.ac.cn/VFs/main.htm
- Accepts protein or DNA sequence and identifies presence of known virulence factors using sequence similarity
- VFanalyzer for detecting virulence factors in draft or complete genomes

Raw Sequence Data Type

FastQ format

@ Unique identifier
Raw sequence
Optional text
Quality score



```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

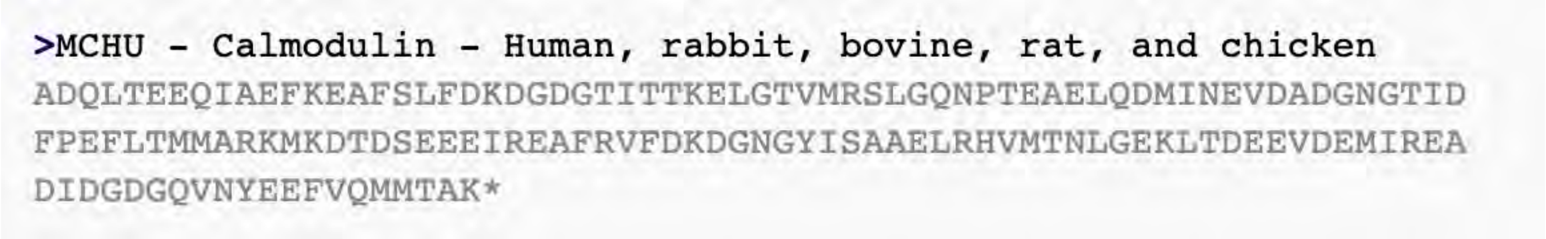
- PHRED quality score encodes the probability of an erroneous call $Q = -10 \log_{10} P$
- Quality score of 30 for a base indicates that the chances of calling this base incorrectly are 1 in 1000
- Encoded in ASCII characters

https://en.wikipedia.org/wiki/FASTQ_format

FASTA Format

- Fasta files normally have extension .fasta, .fas, .fa, .fna, .faa, frn
- Used for nucleotide as well as amino acid sequences

> Header
Sequence



```
>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*
```

https://en.wikipedia.org/wiki/FASTA_form

Sequence Alignment/Map format

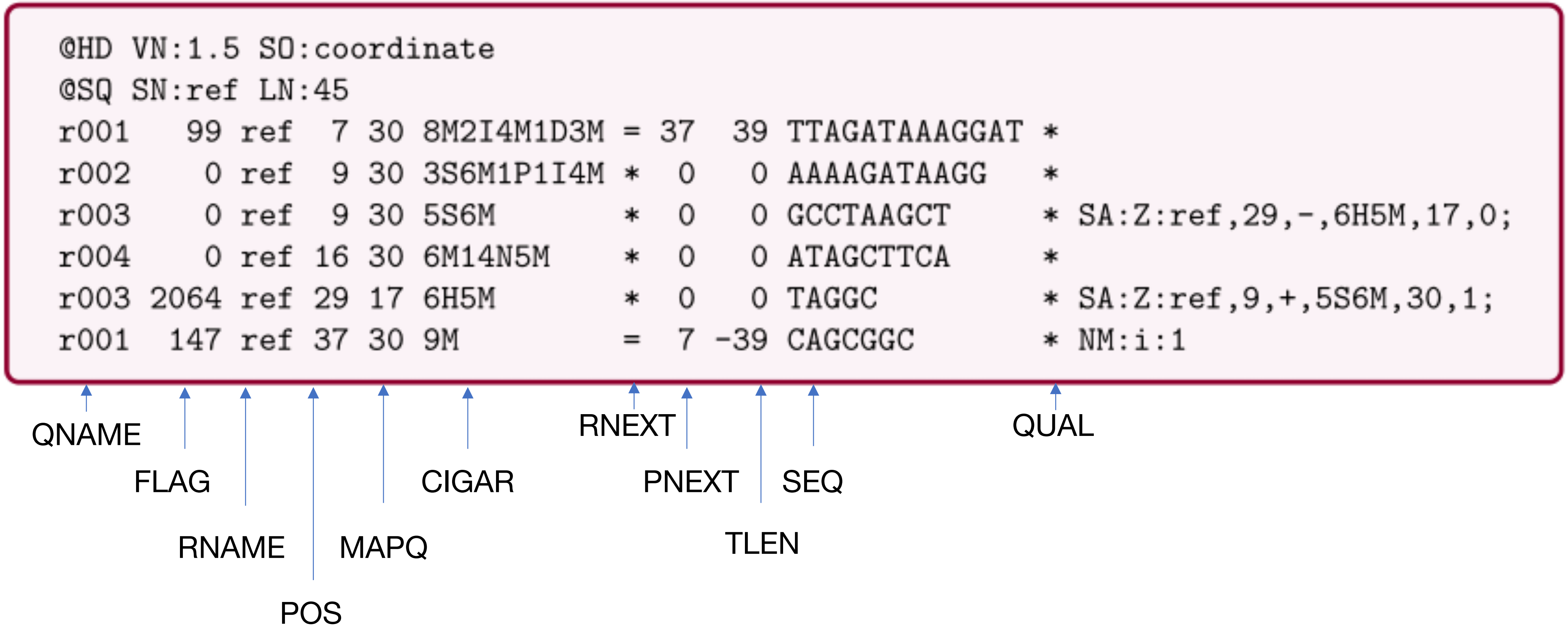
SAM format

- Widely accepted format for storing read alignments against a reference sequence
- Stores read mate pair information
- Reads can be classed by library, sequencer lane
- Binary version of SAM is BAM

Column	Field	Description
1	QNAME	Query Name
2	FLAG	Bit wise flag (Mapped, pairing info)
3	RNAME	Reference name
4	POS	1-based leftmost alignment start, clipped
5	MAPQ	PHRED scaled mapping quality
6	CIGAR	Alignment representation
7	RNEXT	Mate reference information
8	PNEXT	Position of mate
9	TLEN	Observed template length
10	SEQ	Sequence
11	QUAL	PHRED scaled base quality

Image from: https://doi.org/10.3390/info7040056

SAM Format Example



```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGAT *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGG   *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCT     * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCA     *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC         * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGC       * NM:i:1
```

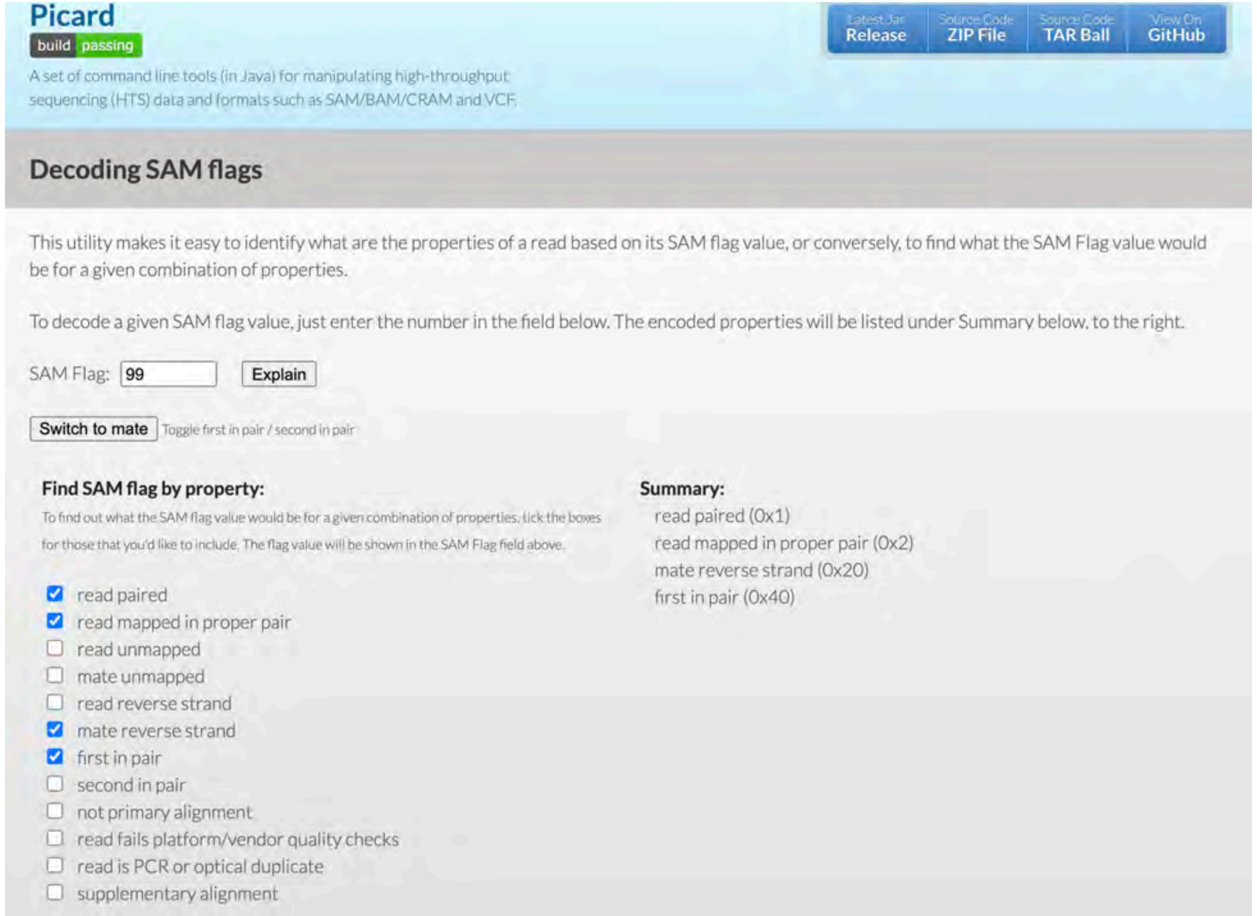
SAM Flags

Bitwise FLAGs

#	Decimal	Description of read
1	1	Read paired
2	2	Read mapped in proper pair
3	4	Read unmapped
4	8	Mate unmapped
5	16	Read reverse strand
6	32	Mate reverse strand
7	64	First in pair
8	128	Second in pair
9	256	Not primary alignment
10	512	Read fails platform/vendor quality checks
11	1024	Read is PCR or optical duplicate
12	2048	Supplementary alignment
Sum	0	

https://www.samformat.info/sam-format-flag

Web based tool for decoding SAM FLAG



A set of command line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.

Decoding SAM flags

This utility makes it easy to identify what are the properties of a read based on its SAM flag value, or conversely, to find what the SAM Flag value would be for a given combination of properties.

To decode a given SAM flag value, just enter the number in the field below. The encoded properties will be listed under Summary below, to the right.

SAM Flag: 99 Explain

Switch to mate Toggle first in pair / second in pair

Find SAM flag by property:

To find out what the SAM flag value would be for a given combination of properties, tick the boxes for those that you'd like to include. The flag value will be shown in the SAM Flag field above.

- ☑ read paired
- ☑ read mapped in proper pair
- □ read unmapped
- □ mate unmapped
- □ read reverse strand
- ☑ mate reverse strand
- ☑ first in pair
- □ second in pair
- □ not primary alignment
- □ read fails platform/vendor quality checks
- □ read is PCR or optical duplicate
- □ supplementary alignment

Summary:

read paired (0x1)
read mapped in proper pair (0x2)
mate reverse strand (0x20)
first in pair (0x40)

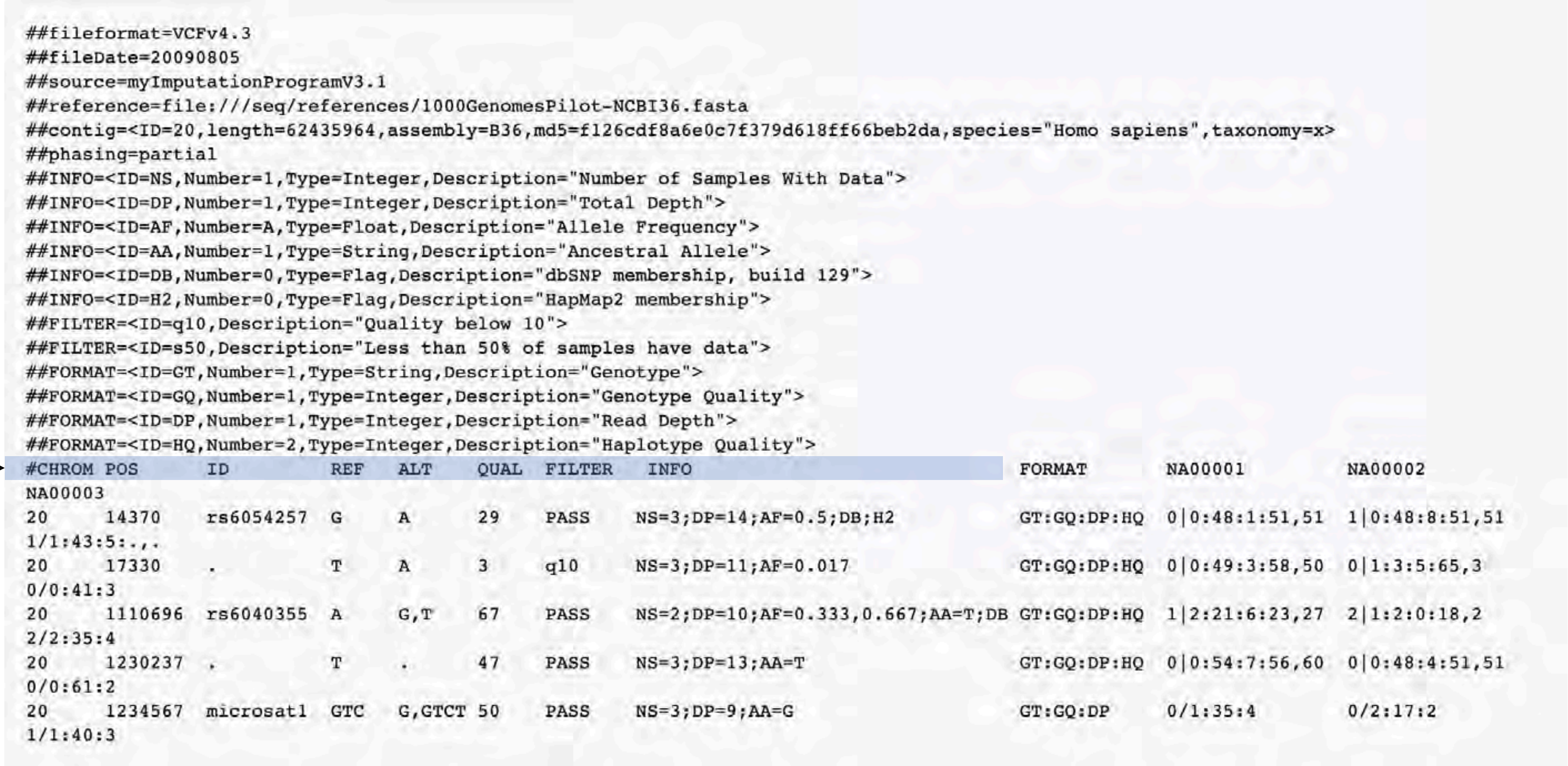
https://broadinstitute.github.io/picard/explain-flags.html

Variant Calling Format

- Used for storing gene sequence variation information
- Contains header section and 8 mandatory columns and unlimited optional columns

Header

8 columns

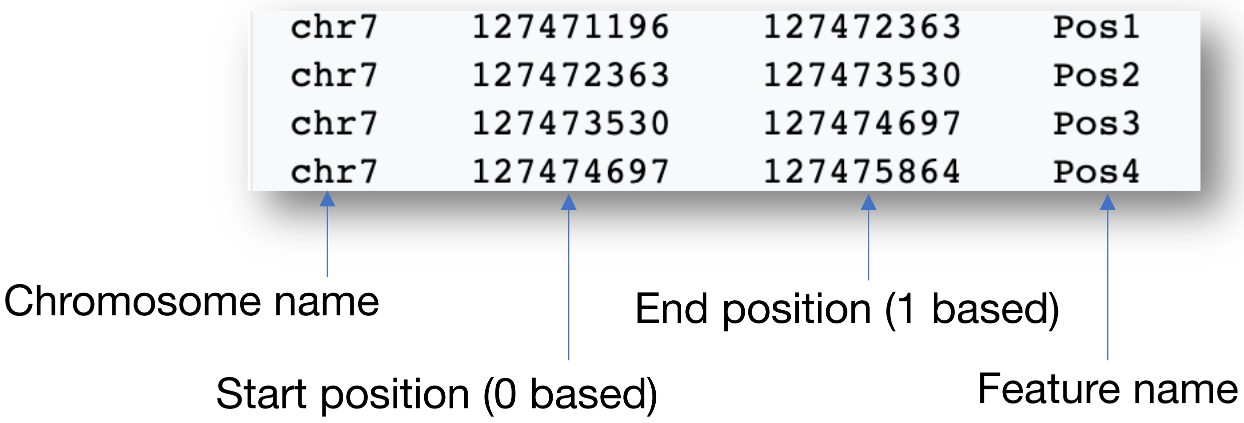


```
##fileformat=VCFv4.3
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number 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
```

BED File Format

Browser **E**xtensible **D**ata

- Used to store annotations on genomic regions
- Requires a minimum of three columns
- File extension is .bed



```
chr7    127471196    127472363    Pos1
chr7    127472363    127473530    Pos2
chr7    127473530    127474697    Pos3
chr7    127474697    127475864    Pos4
```

Chromosome name

Start position (0 based)

End position (1 based)

Feature name

Format Conversion Tools

- Analysis tools need input in different formats
- EMBOSS seqret is web-based tool for file format conversion https://www.ebi.ac.uk/Tools/sfc/emboss_seqret/

FASTQ
FASTA

- EMBOSS provides comprehensive set of web-based tools and databases for performing complex analysis https://www.ebi.ac.uk/services

Summary

1. Human Microbiome
2. Next Generation Sequencing (NGS) principle and applications
3. Workflow for a typical metagenomics project
4. Bioinformatics databases , MGI , CARD, VFDB
5. Bioinformatics data types, FASTQ, SAM, BED

Thank You

Analyzing Microbial Communities Using Next Generation Sequencing

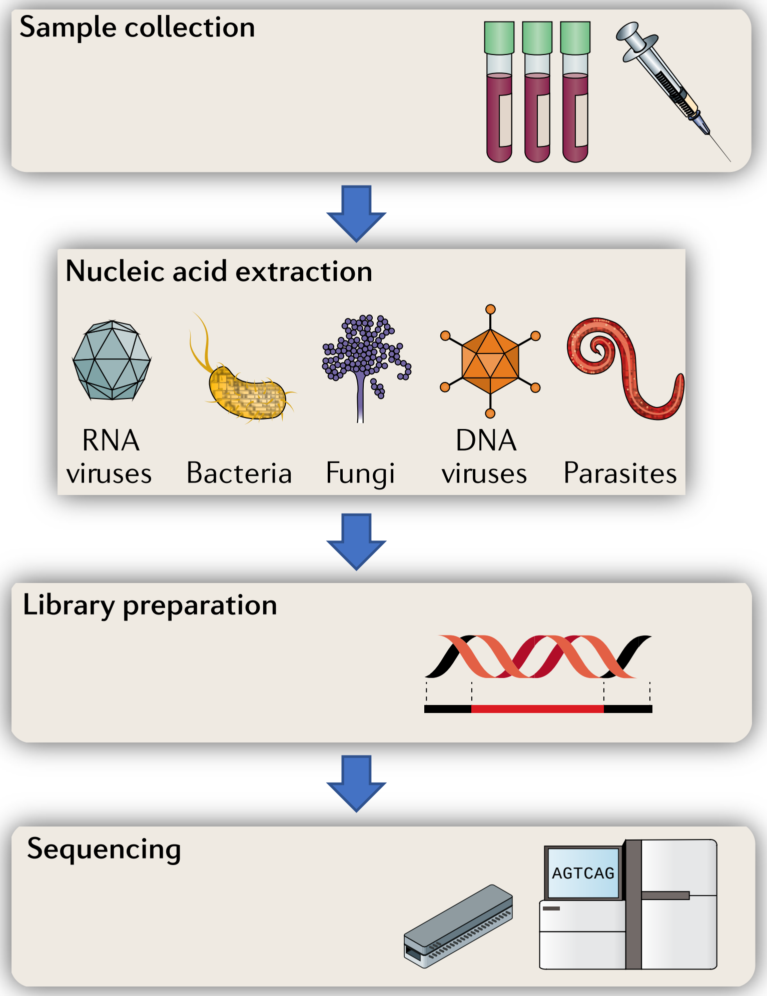
PART II: Workflow, Tools and Application

Rounak Feigelman, Ph.D.

Senior Scientist, Paragon Genomics Inc.

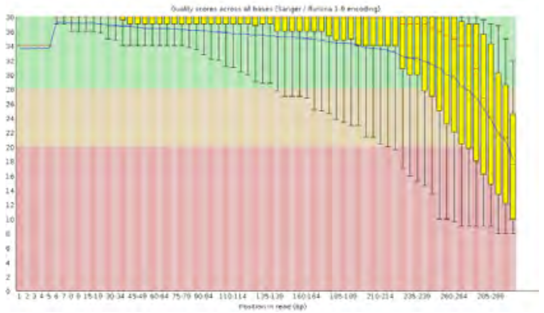
Figure adapted from PMID: 30918369

Recap: Workflow

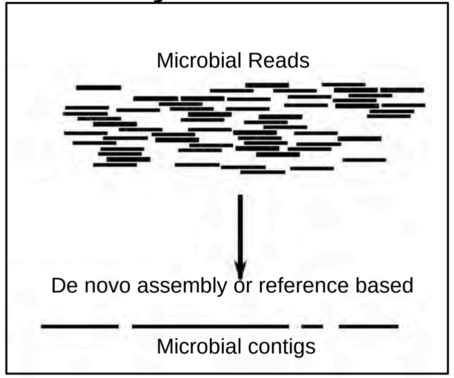


Data Preparation

1. Quality control and read trim



2. Assembly of microbial reads



Data Analysis Workflow

Taxonomic characterization

Gene prediction

Gene Annotation

Identify adaptations

Genome reconstruction of “clonal” microbes

Data Quality Assessment

FASTQC

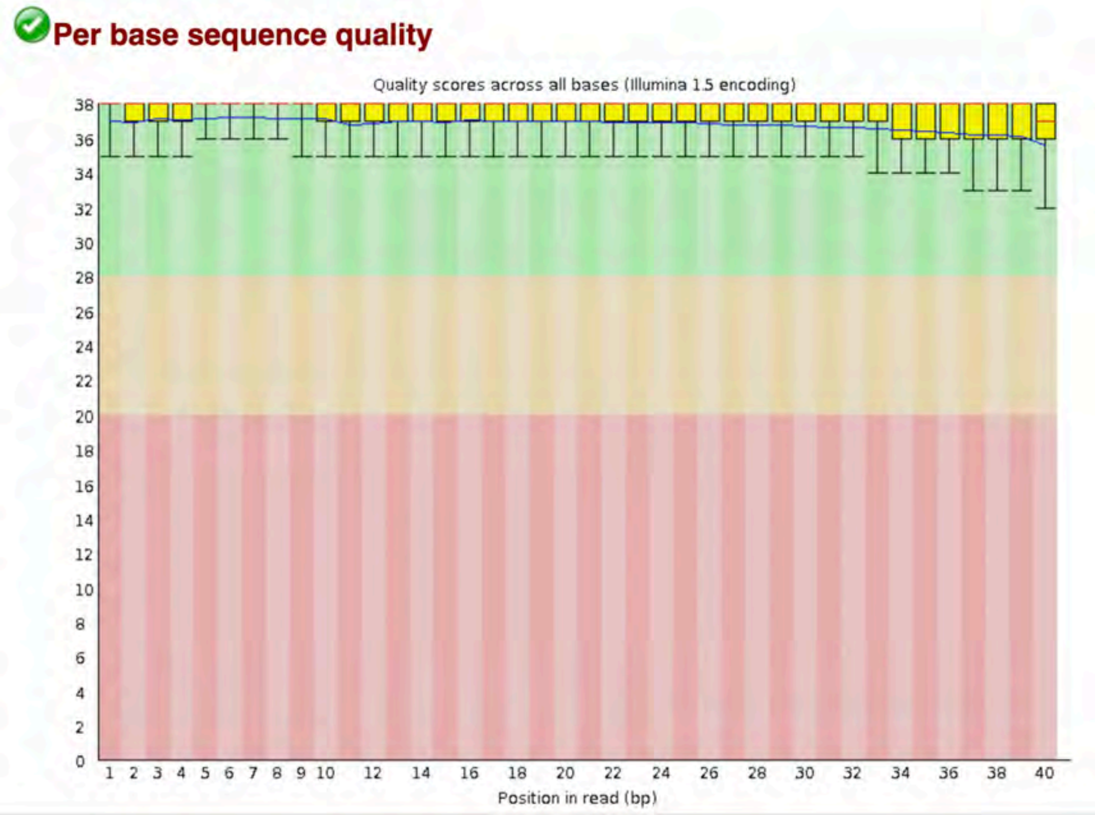
- Open source tool designed to identify issues with sequencing data
- Accepts raw sequencing data in FASTQ format
- Runs multiple analysis and reports pass/warning/fail
- Graphical output

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

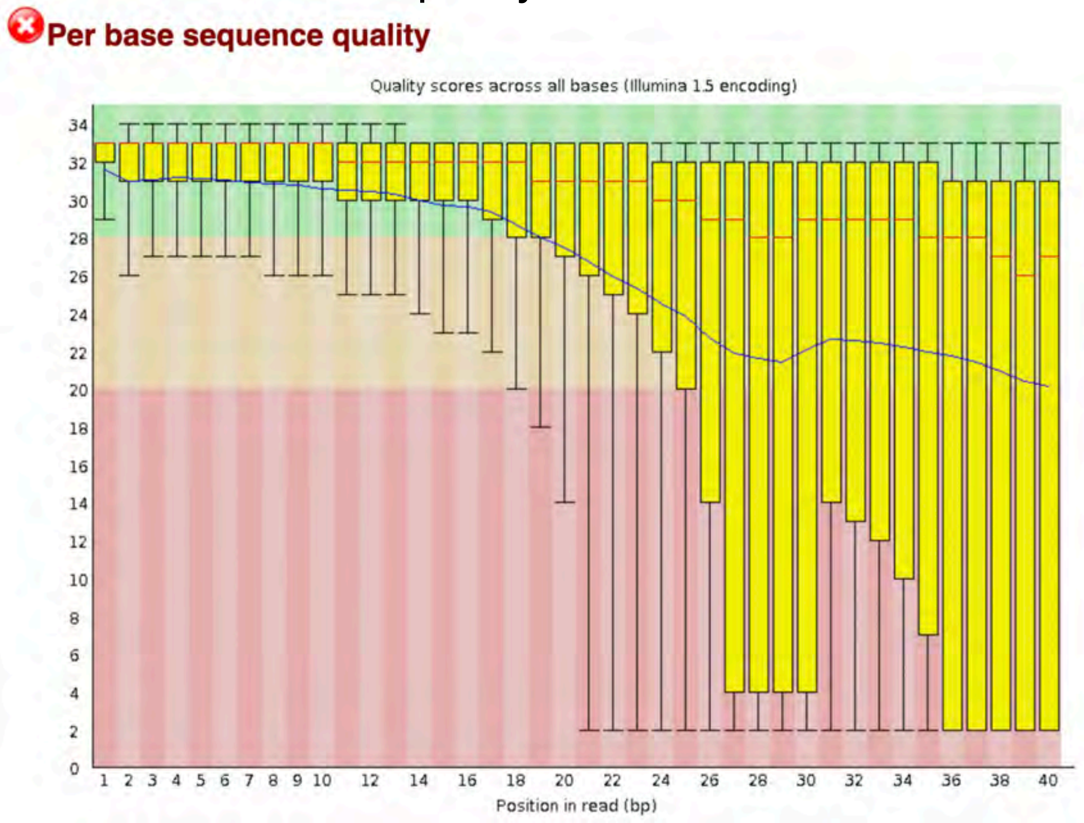
Images: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FASTQC Output

Good quality Illumina data



Poor quality Illumina data



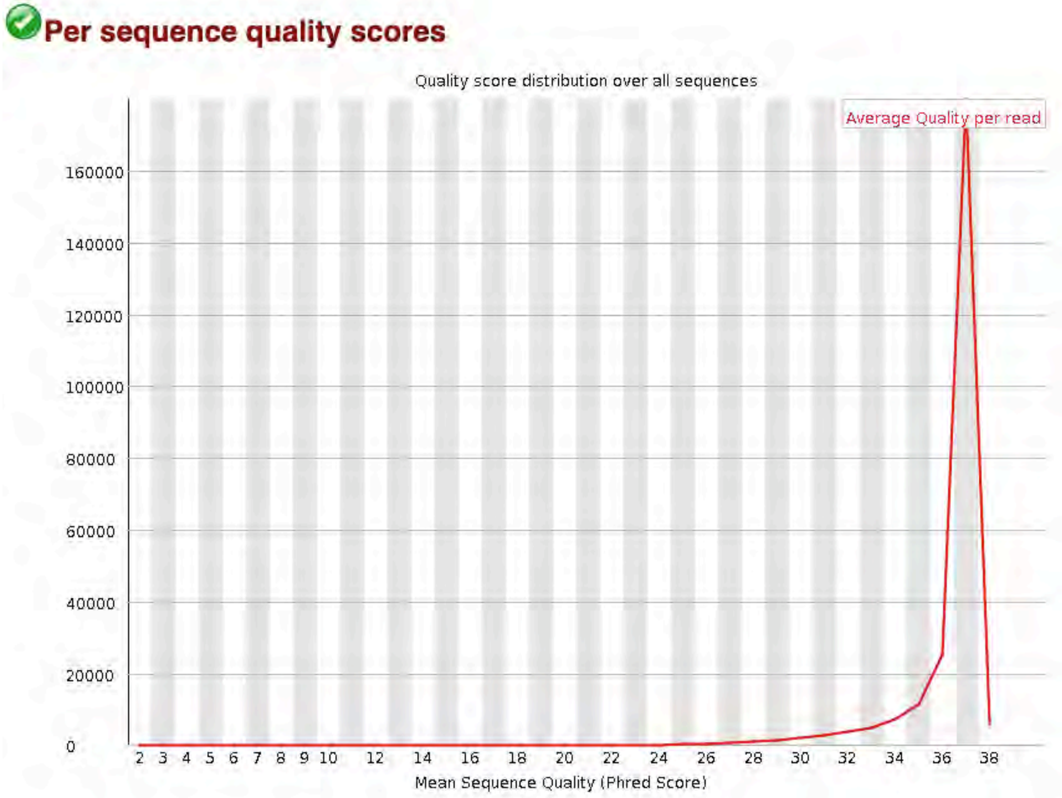
Phred scores drop towards the end of reads

WorkFlow → Quality Control → Alignment → Assembly → Annotation

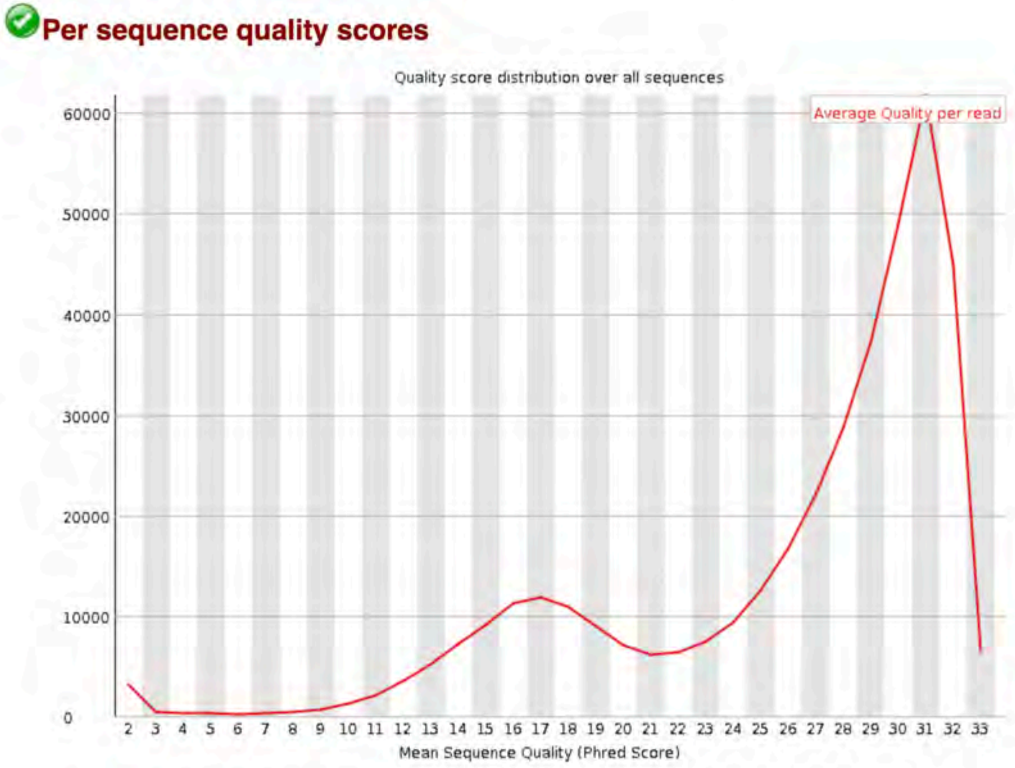
Images: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FASTQC Output

Good quality Illumina data

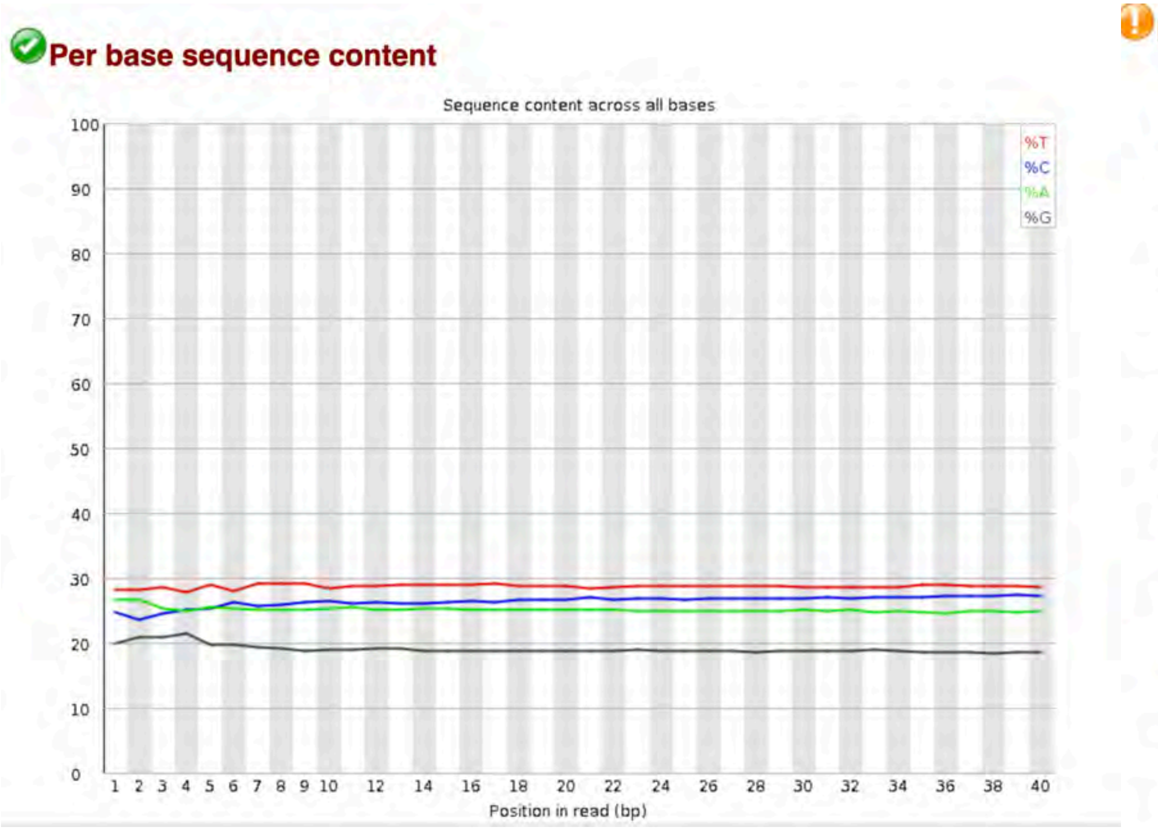


Poor quality Illumina data

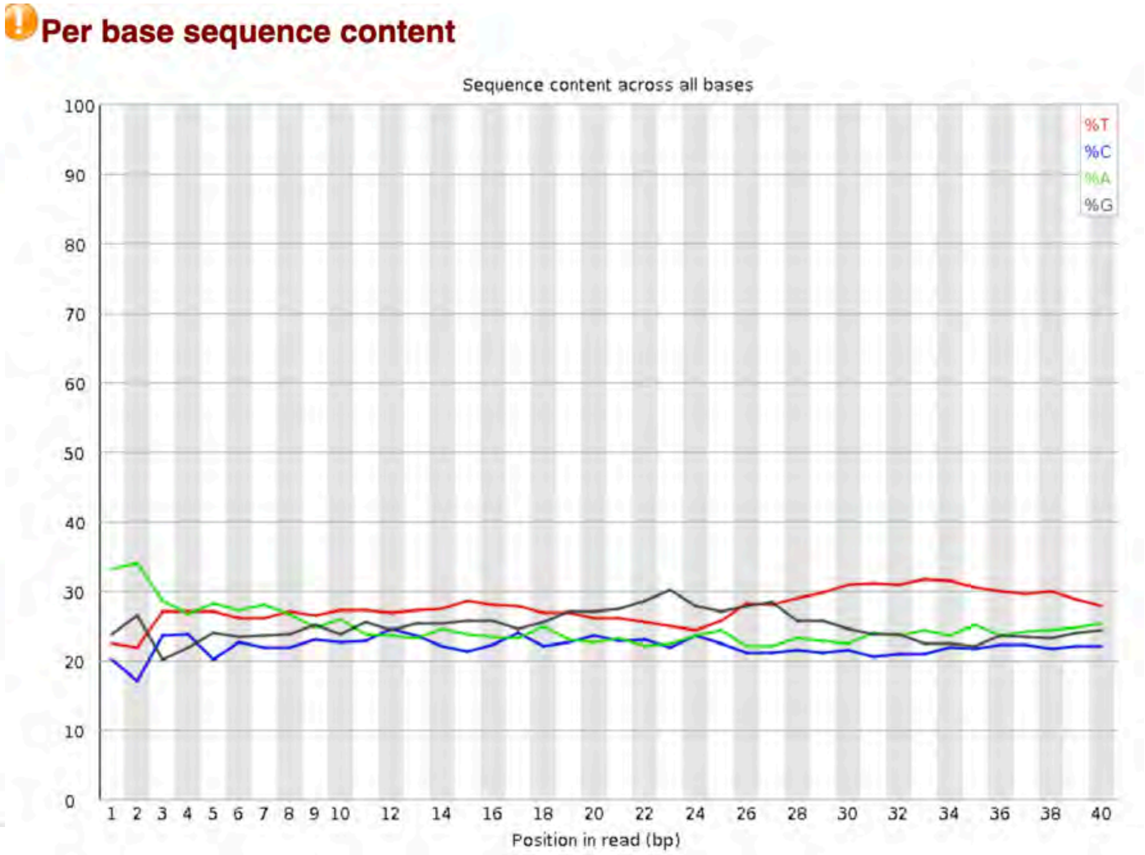


Images: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FASTQC Output



✓ Per base sequence content

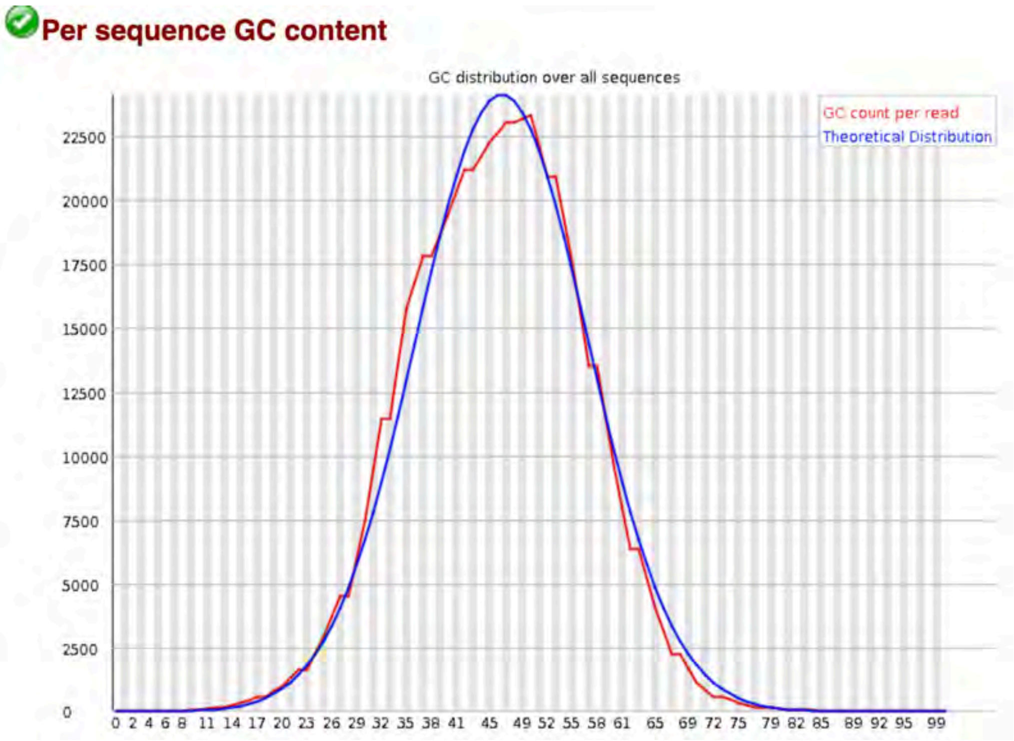


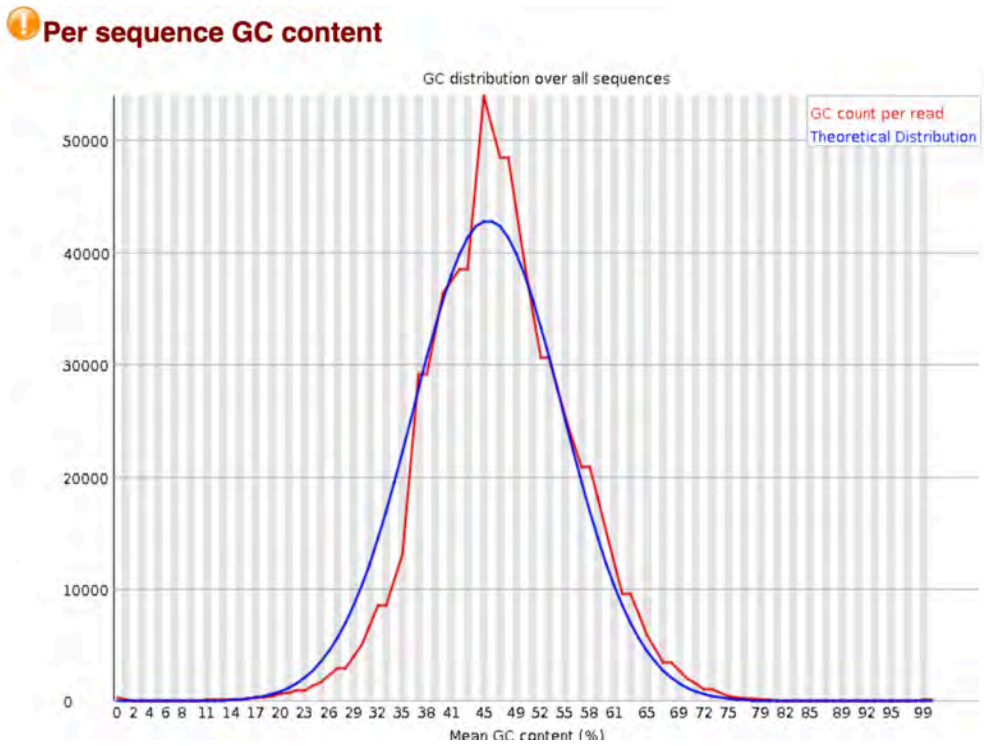
! Per base sequence content

Per base sequence content helps identify bias in sequence composition

WorkFlow → Quality Control → Alignment → Assembly → Annotation

FASTQC Output

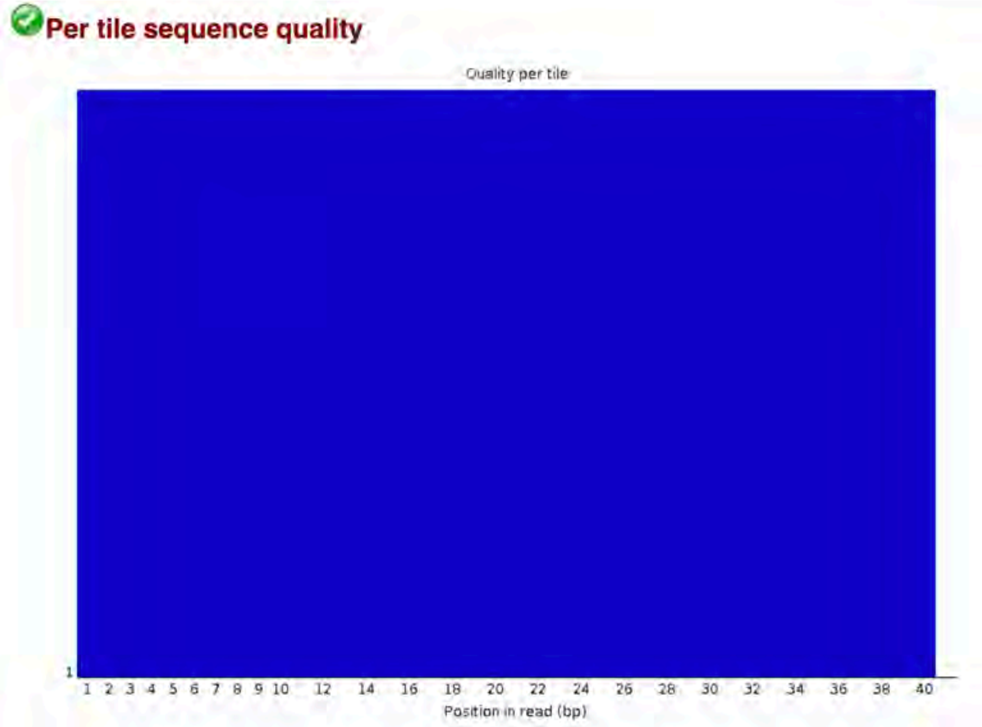

Per sequence GC content


Per sequence GC content

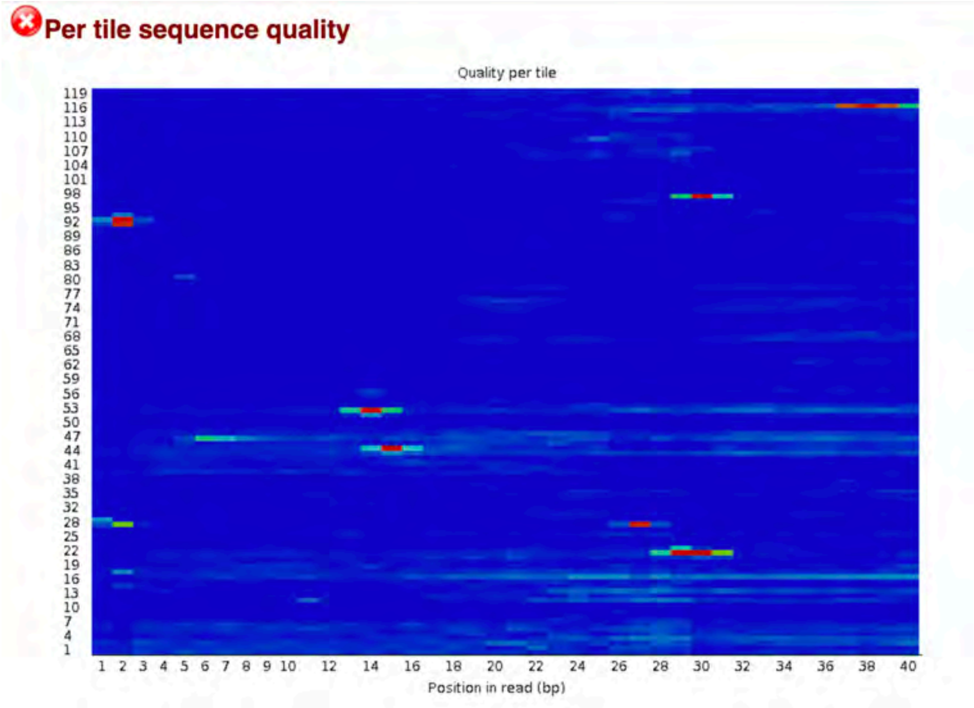
Images: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

FASTQC Output : Illumina Specific

- Deviation from average quality score at each flowcell tile
- Red indicates lower than average
- Tiles showing consistently poor quality indicate issue with the flowcell lane such as debris



Per tile sequence quality



Per tile sequence quality

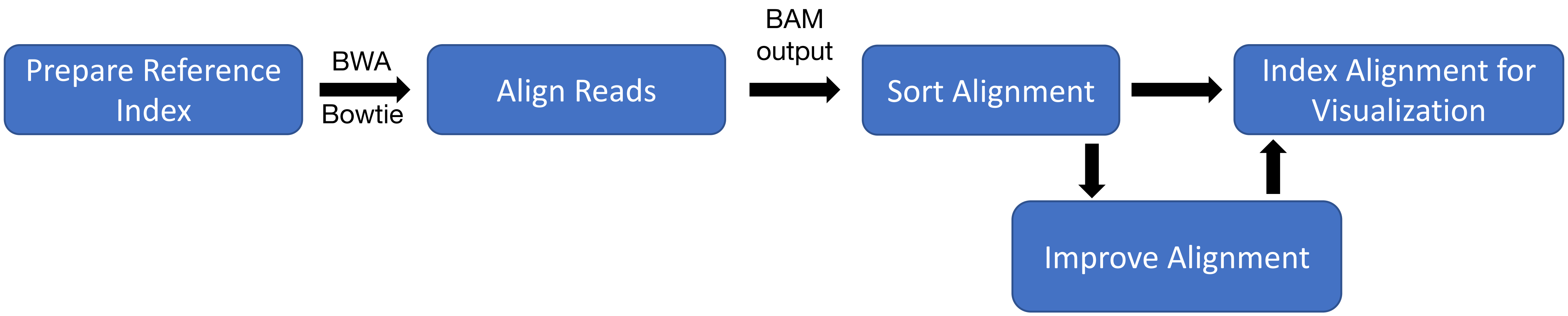
FASTQC Report

Additional reporting includes

- Ambiguous nucleotide content per base
- Sequence duplication levels
- Overrepresented sequences
- Adapter content

Alignment Workflow

Datatype	Application	Use Case
Amplicon based sequencing	Map reads to reference of intended target	Microbe detection and variant analysis
Untargeted WGS sequencing	Map reads to host genome	Filter out reads of non-microbial origin



https://gatk.broadinstitute.org/hc/en-us

https://broadinstitute.github.io/picard/

Alignment Improvement

GATK and **Picard** tools are most widely used for improving alignments

1. Realignment around insertion/deletion
2. Base quality recalibration
3. Library duplicate removal
    - When multiple PCR products from same template molecule bind to the flowcell, PCR duplicates are sequenced
    - Duplicates can result in false variant calls

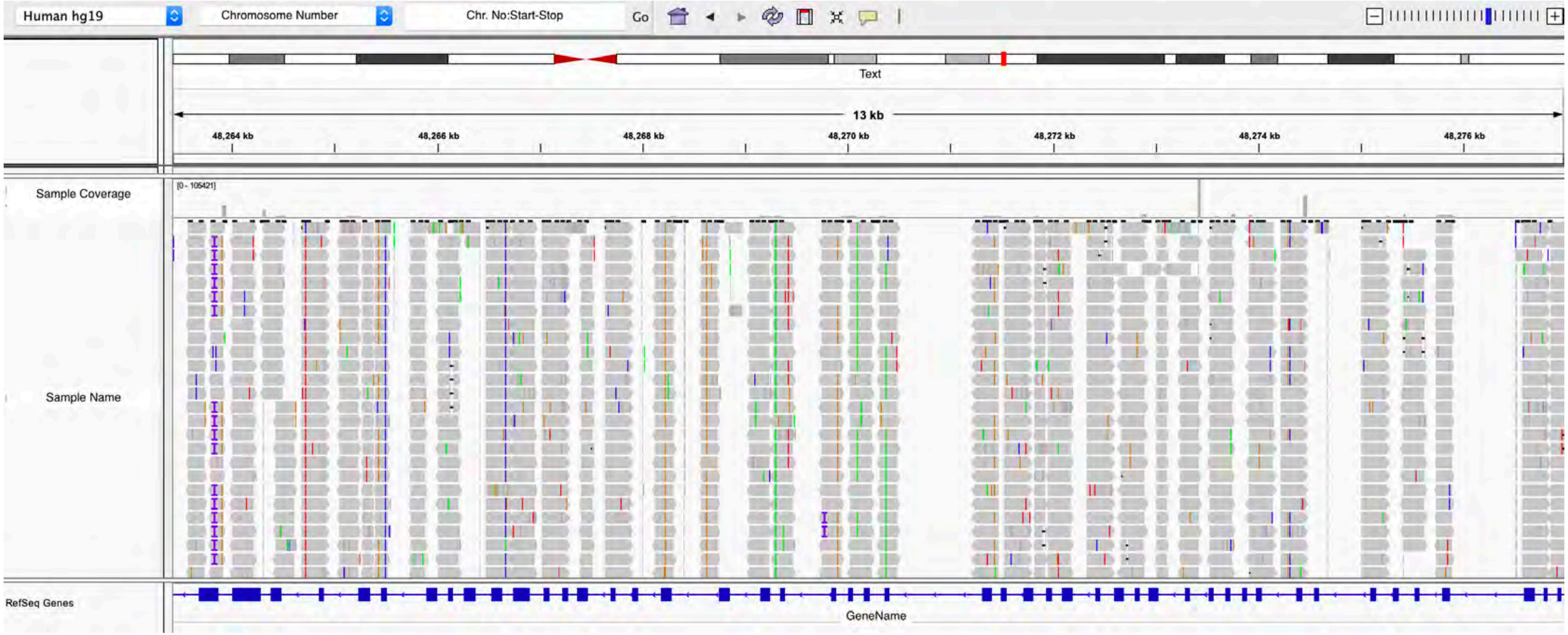
Alignment Tools

1. Burrows-Wheeler Alignment Tool
    - Performs local alignment
    - Used for mapping against a large reference
    - Seeds alignment and extends to in both directions
    - http://bio-bwa.sourceforge.net/bwa.shtml
2. Bowtie2
    - http://bowtie-bio.sourceforge.net/bowtie2/index.shtml

Samtools is used to post process SAM and BAM formats, http://htslib.org

Alignment Visualization Tool

Integrative Genomics Viewer : http://www.broadinstitute.org/igv/download



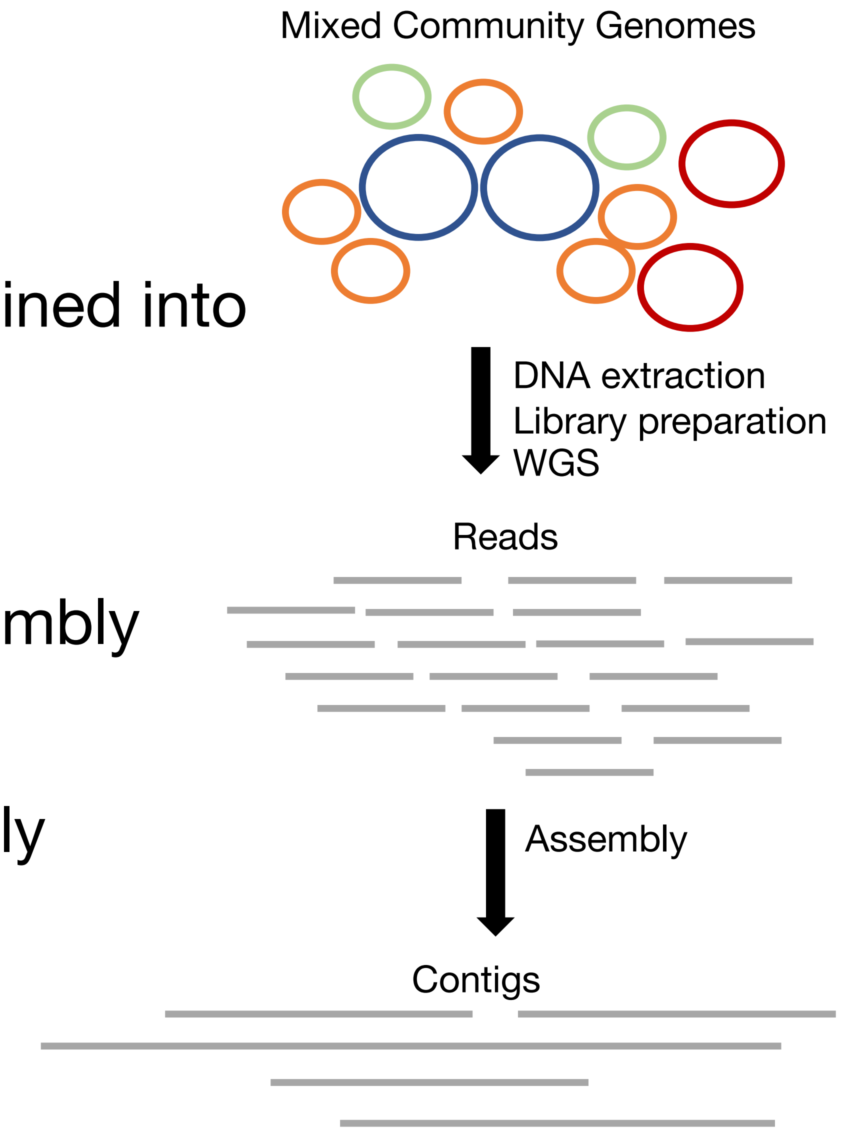
Assembly

- Overlapping reads from a genomic region are combined into contiguous sequence, known as **contigs**

- Two approaches: Reference based or De novo assembly

- Metagenomic assemblers perform de novo assembly

- Available tools
  - metaSPAdes https://cab.spbu.ru/software/spades/
  - MetaVelvet https://www.ebi.ac.uk/~zerbino/velvet/



Mixed Community Genomes

DNA extraction
Library preparation
WGS

Reads

Assembly

Contigs

Metagenomic vs Isolate Assembly

Metagenomic Assembly

a) Bacterial species are mixtures of strains in a mixed community sample

b) Abundance of each species is variable resulting in uneven coverage of each genome

c) Metagenome assembled genomes (MAGs) are composite representative genomes of multiple strains

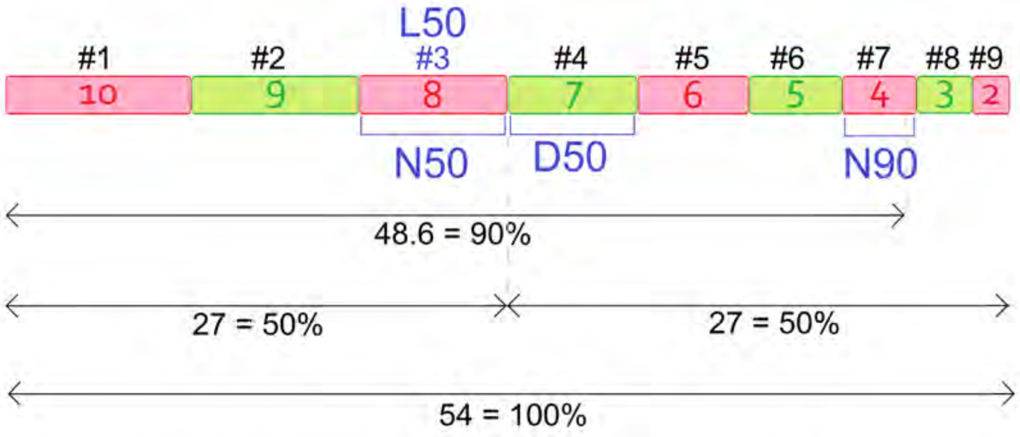
Isolate Assembly

a) Sample is clonal in nature, little to no diversity is expected

b) Coverage is assumed to to uniform across genome

c) Isolate genomes are more accurate representatives of the strain

Assembly Metrics



- N**50**

  50% of the assembly is in contigs of equal or longer length

- L**50**

  Smallest number of fragments that contain **50**% of the assembly

- Min, Max and Mean contig length, number of contigs

Image : https://en.wikipedia.org/wiki/N50,_L50,_and_related_statistics

Functional Annotation

- Gene prediction is performed on assembled sequences “contigs”

- Open reading frames are identified

- PRODIGAL Gene Prediction Software https://github.com/hyattpd/Prodigal

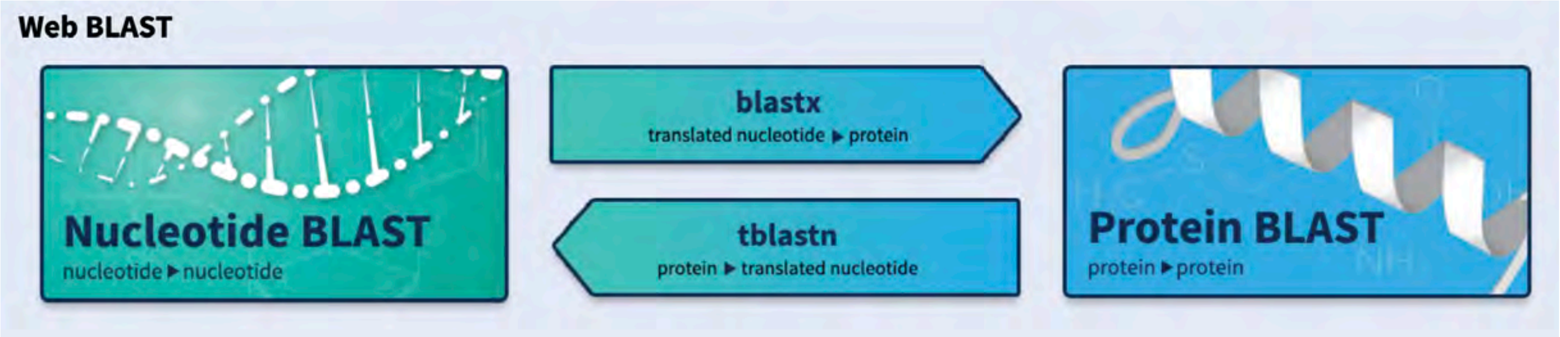
    - Predicts prokaryotic protein coding genes using unsupervised machine learning algorithm
    - Suitable for finished, draft genome or metagenomes
    - Able to detect partial open reading frames that run over contig edges

Functional Annotation

- Prodigal identifies protein coding genes but does not annotate its product
- **Prokka** performs annotation by comparing the predicted gene with high quality protein database of known function and transfer the annotation
- Along with high quality protein sequence databases it uses domain specific databases and models of protein families for annotation

https://github.com/tseemann/prokka

BLAST



Basic **l**ocal **a**lignment **s**earch **t**ool

- BWA and Bowtie work best with lowly divergent sequences
- BLAST is optimized for identifying homology (shared ancestry)
- Used for annotating DNA as well as protein sequences
- Web based and standalone version available https://blast.ncbi.nlm.nih.gov/Blast.cgi

BLAST Search

What is the goal of search? – Identify appropriate database for search

- Identify potential homologs in a particular species - species specific database
- Determine whether these sequences are found in any species – Genbank, RefSeq
- Determine whether sequences contains any coding functional domains - Pfam

Tool	Query Type	Database Type
BLASTn	DNA	DNA
BLASTp	Protein	Protein
blastx	DNA	Protein
tblastn	Protein	DNA
tblastx	DNA	DNA

BLAST Example

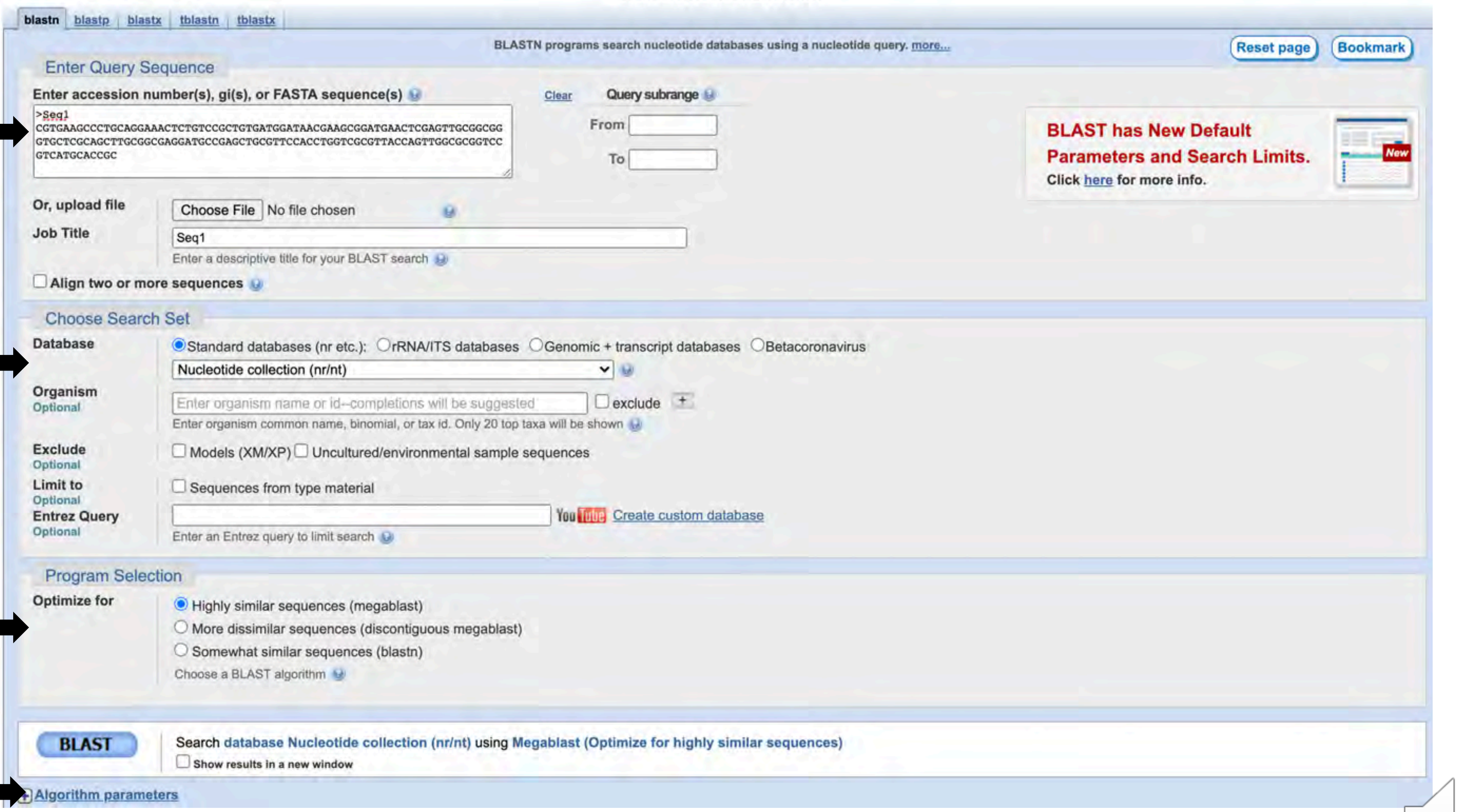
Identify sequence recovered from sputum of a Cystic Fibrosis patient

Enter Query Sequence

Select Reference Database

Select Algorithm for Search

Fine-tune Search Parameters



BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
>Seq1
CGTGAAGCCCTGCAGGAAACTCTGTCCGCTGTGATGGATAACGAAGCGGATGAACTCGAGTTGCGGCGG
GTGCTCGCAGCTTGCGGCGAGGATGCCGAGCTGCGTTCCACCTGGTCGCGTTACCAGTTGGCGCGGTCC
GTCATGCACCGC
```

Query subrange

From

To

BLAST has New Default Parameters and Search Limits.
Click here for more info.

Or, upload file: Choose File No file chosen

Job Title: Seq1

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ○ Betacoronavirus

Nucleotide collection (nr/nt)

Organism
Optional

Enter organism name or id--completions will be suggested □ exclude

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude
Optional

□ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to
Optional

□ Sequences from type material

Entrez Query
Optional

Enter an Entrez query to limit search

Program Selection

Optimize for

● Highly similar sequences (megablast)
○ More dissimilar sequences (discontiguous megablast)
○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

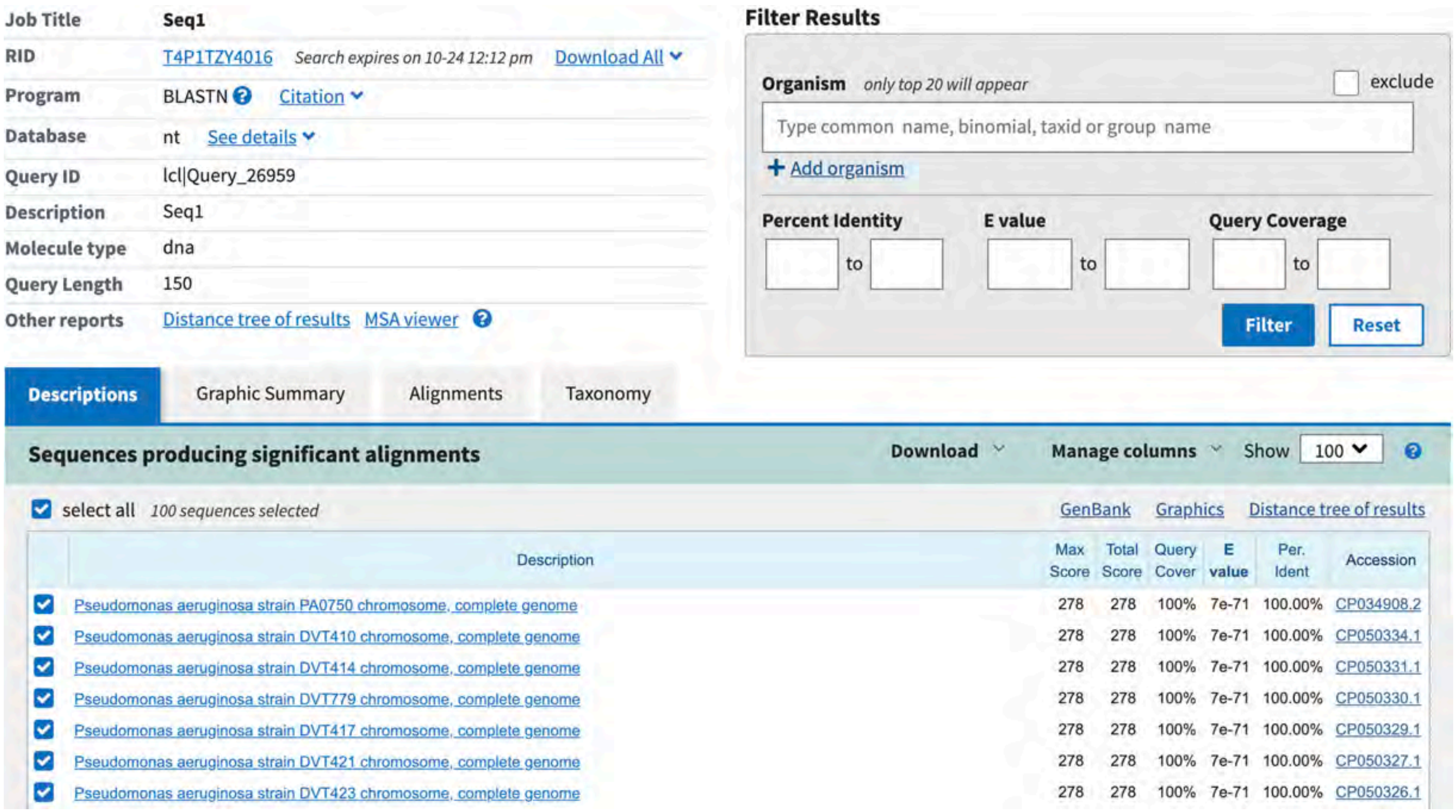
Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

□ Show results in a new window

+ Algorithm parameters

WorkFlow | Quality Control | Alignment | Assembly | Annotation

BLAST Example



Job Title	Seq1
RID	T4P1TZY4016 *Search expires on 10-24 12:12 pm* Download All
Program	BLASTN Citation
Database	nt See details
Query ID	lcl|Query_26959
Description	Seq1
Molecule type	dna
Query Length	150
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* □ exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to **E value** to **Query Coverage** to

Sequences producing significant alignments

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Pseudomonas aeruginosa strain PA0750 chromosome, complete genome	278	278	100%	7e-71	100.00%	CP034908.2
☑	Pseudomonas aeruginosa strain DVT410 chromosome, complete genome	278	278	100%	7e-71	100.00%	CP050334.1
☑	Pseudomonas aeruginosa strain DVT414 chromosome, complete genome	278	278	100%	7e-71	100.00%	CP050331.1
☑	Pseudomonas aeruginosa strain DVT779 chromosome, complete genome	278	278	100%	7e-71	100.00%	CP050330.1
☑	Pseudomonas aeruginosa strain DVT417 chromosome, complete genome	278	278	100%	7e-71	100.00%	CP050329.1
☑	Pseudomonas aeruginosa strain DVT421 chromosome, complete genome	278	278	100%	7e-71	100.00%	CP050327.1
☑	Pseudomonas aeruginosa strain DVT423 chromosome, complete genome	278	278	100%	7e-71	100.00%	CP050326.1

E value

- Number of hits expected to see by chance when searching the database
- Dependent on database size
- Small e values values indicate high confidence in match

Summary

- Quality Control, FASTQC
- Alignment workflow and tools
- Assembly principles and metrics
- Annotation tools and examples, BLAST

Thank You