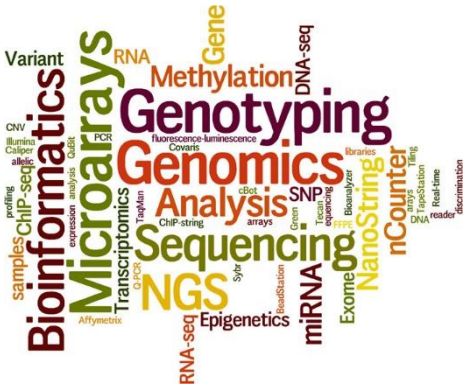
Get to know NGS data

Genome assembly and annotation workshop

7 August 2018

Pumipat Tongyoo, Ph.D.

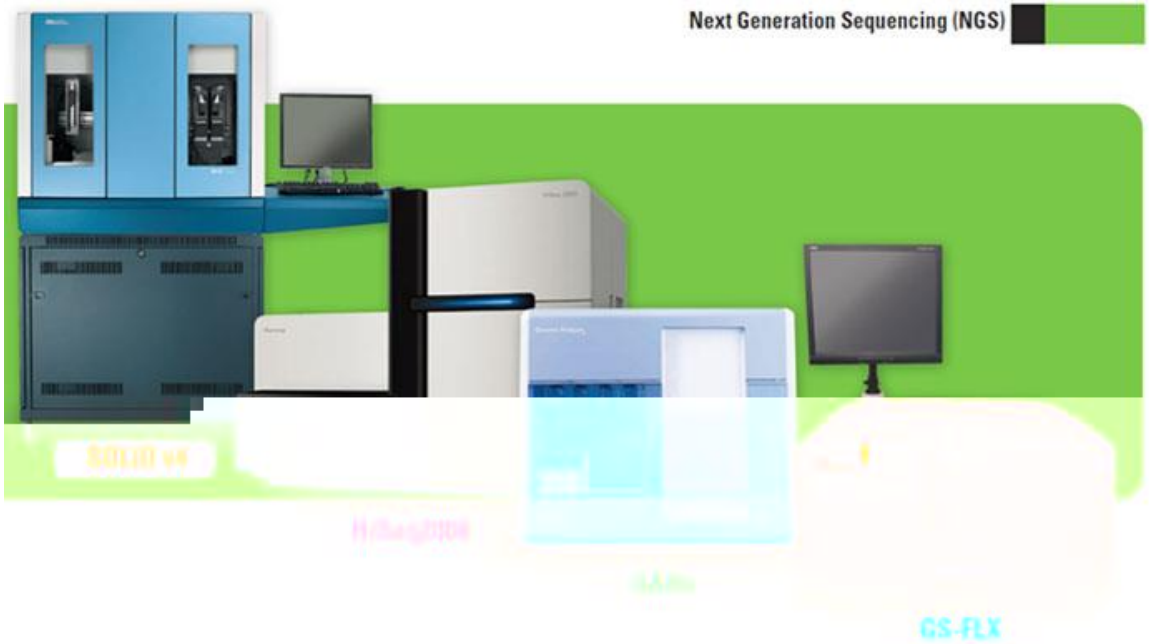
Center of Excellence on Agricultural Biotechnology (AG-BIO)
Center for Agricultural Biotechnology (CAB)
Kasetsart University Kamphaeng Saen Campus



Get the data

NGS raw read

Reference sequence



How to get NGS data?

Sequence your samples

Existing NGS data

- The Sequence Read Archive (SRA)
- ENCODE: Encyclopedia of DNA Elements

Specific resources

- 150 Tomato Genome ReSequencing project
- (http://www.tomatogenome.net/)

Sequence Read Archive (SRA)

An international public resource for NGS data

Operated by the International Nucleotide Sequence Database Collaboration (INSDC)

INSDC partners include

- National Center for Biotechnology Information (NCBI),
- European Bioinformatics Institute (EBI)
- DNA Data Bank of Japan (DDBJ)

Nucleic Acids Res. 2011 Jan; 39(Database issue)

Sequence Read Archive (SRA)

https://www.ncbi.nlm.nih.gov/sra/

NCBI Resources How To Sign in to NCBI

SRA SRA Search

Advanced Help

SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®.

Getting Started

- How to Submit
- Login to SRA
- Login to Submission Portal
- SRA Handbook
- Download Guide
- SRA Fact Sheet (.pdf)

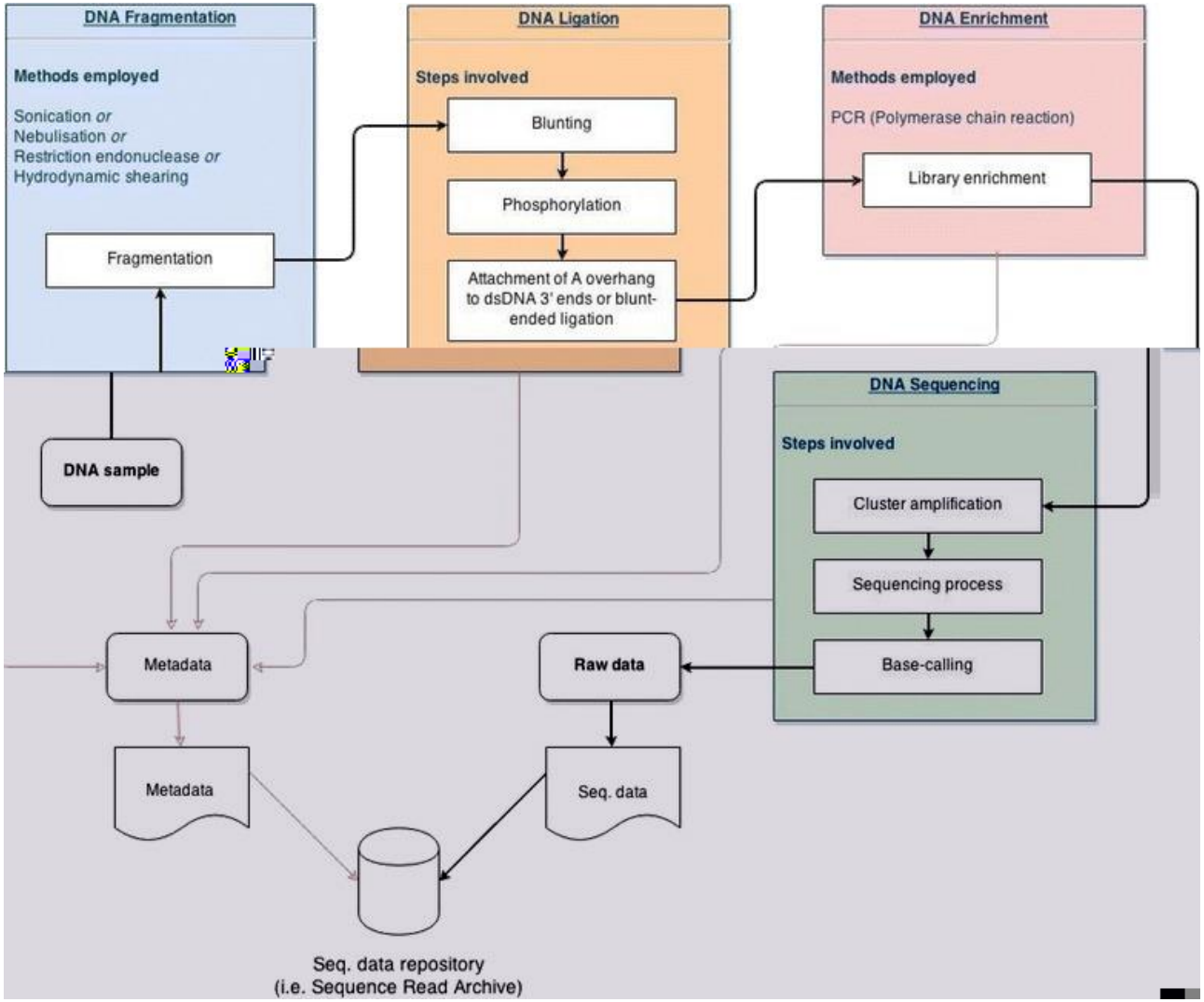
Tools and Software

- Download SRA Toolkit
- SRA Toolkit Documentation
- SRA-BLAST
- SRA Run Browser
- SRA Run Selector

Related Resources

- Submission Portal
- Trace Archive
- dbGaP Home
- BioProject
- BioSample

From NGS to SRA workflow



U.S. National Library of Medicine

Log in

Search NC

tomato

Results found in 37

Literature

Bookshelf

ne loci

abundance

cted

ic and

s

Genetics

ClinVar	0	Human variations of clinical significance
dbGaP	152	Genotype/phenotype interaction studies
dbVar	0	Genome structural variation studies
GTR	0	Genetic testing registry
MedGen	6	Medical genetics literature and links
OMIM	1	Online mendelian inheritance in man
SNP	0	Short genetic variations

Proteins

Conserved Domains	27	Conserved protein domains
Identical Protein Groups	63,067	Protein sequences grouped by identity
Protein	971,212	Protein sequences
Protein Clusters	80	Sequence similarity-based protein clusters
Sparcle	122	Functional categorization of proteins by domain architecture
Structure	215	Experimentally-determined biomolecular structures

Genomes

Genome assembly information

activity screening studies

emical information with structures, formation and links

posited substance and chemical ormation

BioCollections	1	Museum, herbaria, and other bio-repository collections
BioProject	1,022	Biological projects providing data to NCBI
BioSample	12,458	Descriptions of biological source materials
Clone	228,213	Genomic and cDNA clones
Genome	173	Genome sequencing projects by organism
GSS	555,583	Genome survey sequences
Nucleotide	496,689	DNA and RNA sequences
Probe	16,965	Sequence-based probes and primers
SRA	28,823	High-throughput sequence reads
Taxonomy	1	Taxonomic classification and nomenclature

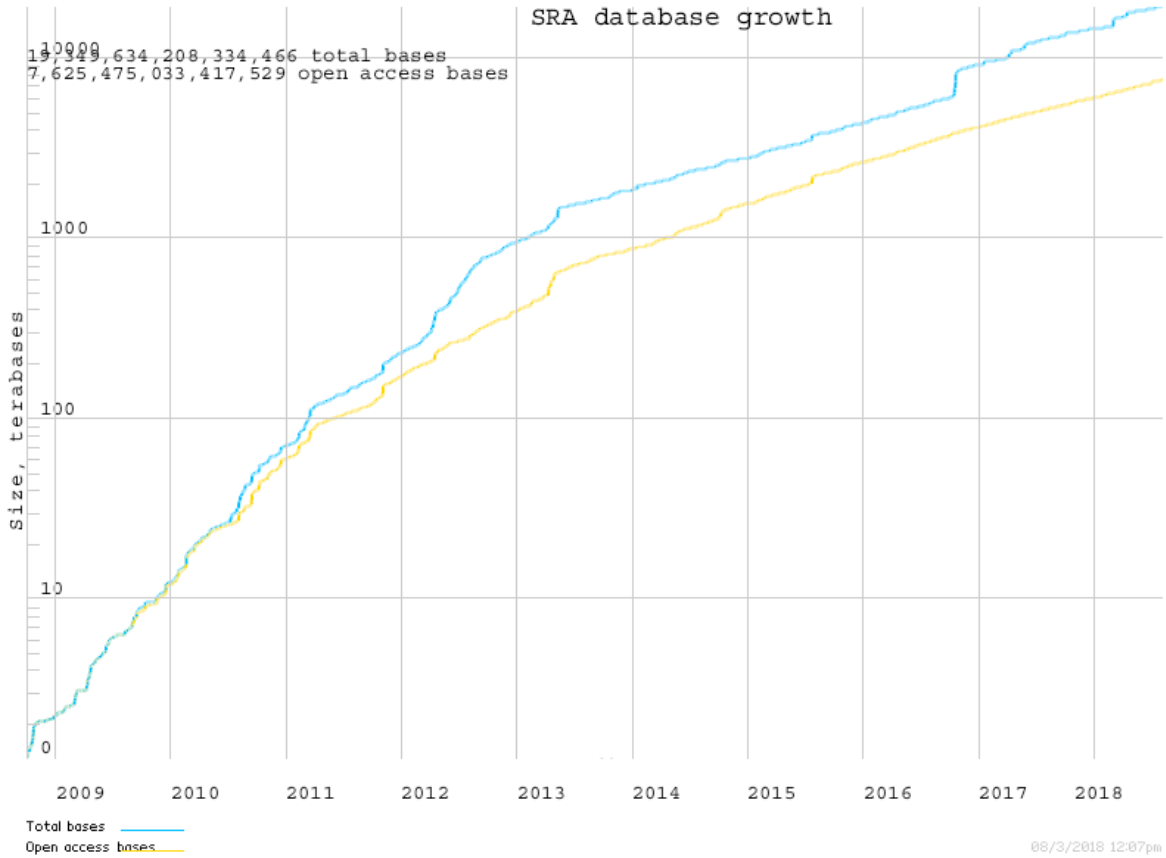
Chemicals

Molecular pathways with links to genes

PubChem BioAssay	1,044	
PubChem Compound	5	
PubChem Substance	166	

The SRA grown rate

- 65% of the SRA was human genomic sequence, 1000 Genome Project
- SRA relies on the NCBI SRA Toolkit



Sequence Read Archive (SRA)

https://www.ebi.ac.uk/ena

EMBL-EBI

Services | Research | Training | About us

ENA
European Nucleotide Archive

sra

Search

Examples: BN000065, histone

Advanced
Sequence

Home | Search & Browse | Submit & Update | Software | About ENA | Support

Search results for *sra*

Show more data from EMBL-EBI

Assembly
Assembly (81)

Sequence
Sequence (Update) (214)
Sequence (Release) (3,759)

Coding
Coding (Update) (478)
Coding (Release) (4,519)

Non-coding
Non-coding (Release) (2)

Read
Experiment (56,362)
Run (52,933)

Analysis
Analysis (74)

Study
Study (165)

Assembly (81 results found)

GCA_003266065.1 ASM326606v1 assembly for Methanosphaera sp. rholeuAM130
View all 81 results

Sequence (Update) (214 results found)

L12012 Chelydra serpentina sex determining sra-7 DNA.
View all 214 results

Sequence (Release) (3,759 results found)

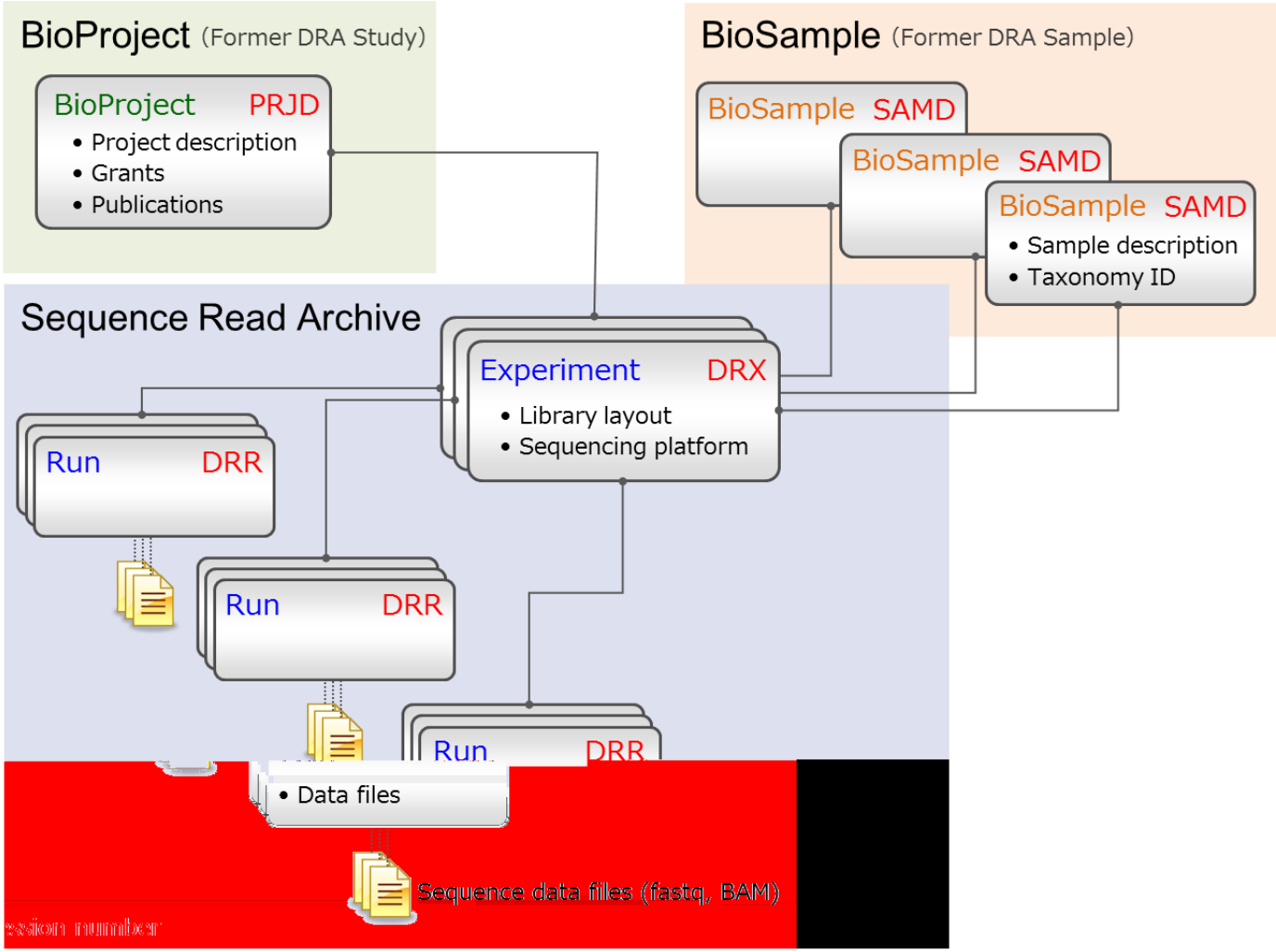
D13179 Escherichia coli sra gene for ribosome-associated protein SRA, complete cds.
View all 3,759 results

Coding (Update) (478 results found)

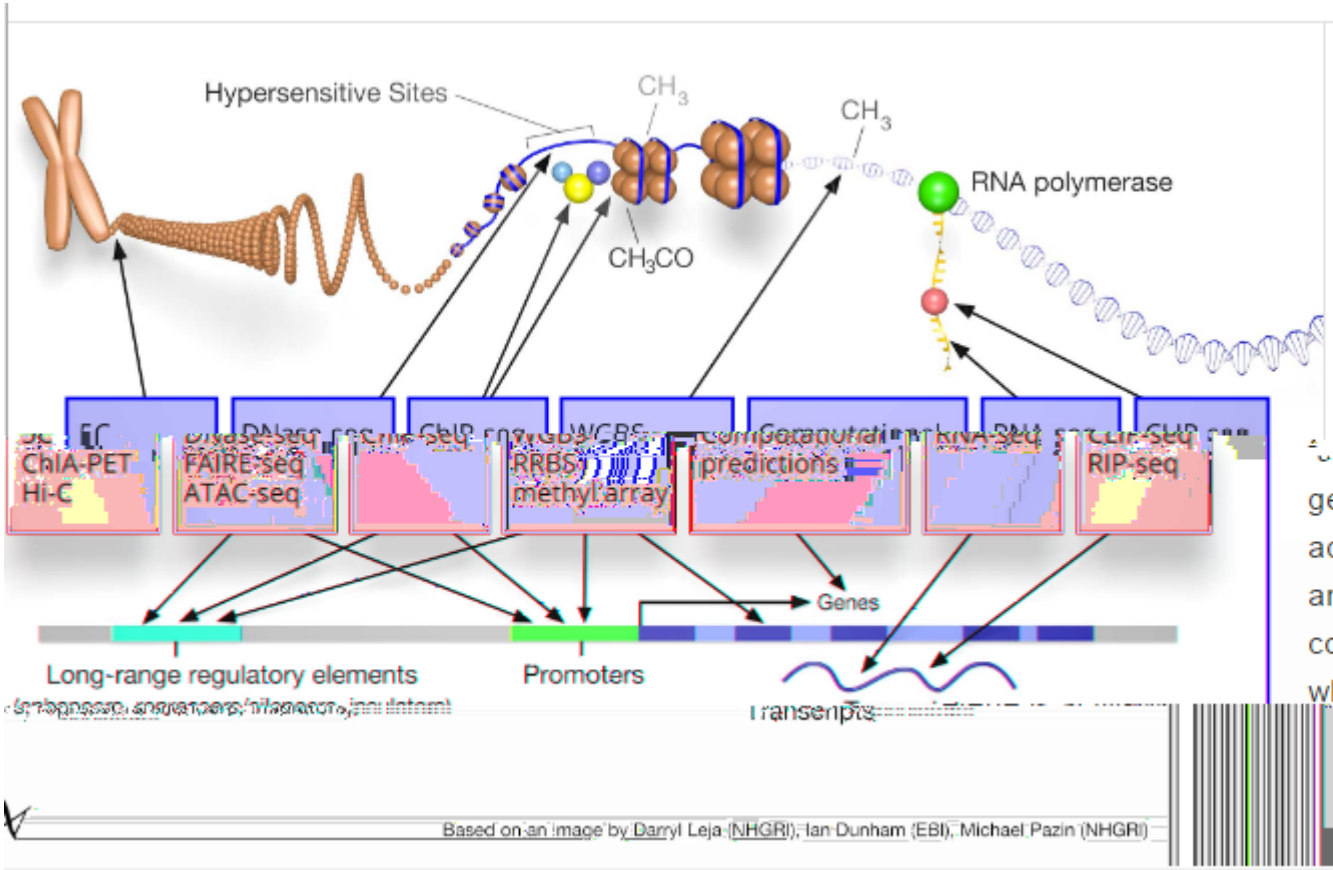
PON60017 Parasponia andersonii SRA-YDG
View all 478 results

DDBJ Sequence Read Archive (DRA)

https://www.ddbj.nig.ac.jp/dra/



ENCODE: Encyclopedia of DNA Elements



The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

Get Started

ENCODE

150 Tomato Genome Resequencing Project

WAGENINGEN UR

Search | Links

Education | Res

Navigation | Read Files | Analysis Files | Portal | Attributes | Publications | Parent Projects

Bulk Download Files (If the downloader app doesn't open, please try using Firefox to launch it.)

Download: 1

Select columns

Showing results 1 - 10 of

Study accession	Sample accession
PRJEB5235	SAMEA2340764
PRJEB5235	SAMEA2340765
PRJEB5235	SAMEA2340766
PRJEB5235	SAMEA2340767
PRJEB5235	SAMEA2340768
PRJEB5235	SAMEA2340769
PRJEB5235	SAMEA2340770

Home
- Selected acc
- Variant brow
- Data Access
- News
- Project Partn
- Project Team
- Ancient acce
- Contact

Select Tracks — Help

My Tracks
- Currently Active
- Recently Used

Category
- 1 Gene models
- 1 Merged VCF
- 1 Reference sequence
- 84 VCF SNPs

Species
- 4 (no data)
- 2 Solanum arcanum
- 2 Solanum chilense
- 2 Solanum chmielewskii
- 1 Solanum corneliomuelleri
- 3 Solanum galapagense
- 7 Solanum habrochaites
- 3 Solanum huaylense
- 54 Solanum lycopersicum
- 2 Solanum neorickii
- 2 Solanum pennellii
- 2 Solanum peruvianum
- 3 Solanum pimpinellifolium

Country
- 33 (no data)
- 1 Argentina
- 2 China
- 1 Columbia
- 1 Costa Rica
- 2 Cuba
- 1 Czechoslovakia
- 9 Ecuador
- 1 German Democratic Republic
- 3 Guatemala
- 2 Italy
- 1 Mexico
- 21 Peru
- 1 Russia
- 1 South Africa
- 1 Spain
- 1 The Netherlands
- 1 Turkey
- 3 USA
- 1 United Kingdom

Genebank
- 8 (no data)
- 4 CGN
- 4 CGN, TGRC
- 5 Craig Lehoullier
- 14 IPK Gatersleben
- 2 PBR
- 1 Rijk Zwaan
- 4 Sand Hill Preservation Center
- 23 TGRC
- 1 TGRC, IPK Gatersleben
- 8 Tomato Growers Supply

Back to browser | Clear All Filters | Contains text | 87 tracks

Label	Species	Cultivar	Collection site	Province/Department	Country	Elevation	Habitat	Reasons	Year	Mating System	Accession	EuSol ID	Other ID	Genebank	Download Link	Data Provider
1	Solanum lycopersicum	Moneymaker	...	...	...	...	...	...	...	Autogamous-SC	PV	several	...	PBR	Download VCF File	Wageningen UR
102	Solanum lycopersicum	...	Makapuu Beach, Oahu	Hawaii	USA	25 (m.a.s.l.)	rocky slope in landscaped area near seashore	Salinity tolerance	2001	Autogamous-SC	LA4133	TR00026	...	TGRC	Download VCF File	Wageningen UR
103	Solanum lycopersicum	...	Santa Cecilia	Napo	Ecuador	500 (m.a.s.l.)	In mixed dooryard garden at river landing	Wagerlogging tolerant	1971	Autogamous-SC	LA1421	TR00027	...	TGRC	Download VCF File	Wageningen UR
104	Solanum galapagense	...	Bartolome	Galapagos Islands	Ecuador	...	...	...	1966	Autogamous-SC	LA1044	TR00029	...	TGRC	Download VCF File	Wageningen UR
105	Solanum lycopersicum	...	Sucua	...	Ecuador	...	Common weed in waste places	...	1969	Autogamous-SC	LA1479	TR00028	...	TGRC	Download VCF File	Wageningen UR
11	Solanum lycopersicum	All Round	...	...	United Kingdom	...	...	Agronomic traits	...	Autogamous-SC	LYC 1365	EA02617	T 187	IPK Gatersleben	Download VCF File	Wageningen UR
12	Solanum lycopersicum	Sonato	...	...	The Netherlands	...	...	Agronomic traits	...	Autogamous-SC	LYC 1969	EA02724	...	IPK Gatersleben	Download VCF File	Wageningen UR
13	Solanum lycopersicum	Cross Country	...	...	Russia	...	...	Agronomic traits	...	Autogamous-SC	LYC 3897	EA03701	T1662	IPK Gatersleben	Download VCF File	Wageningen UR
14	Solanum lycopersicum	Lidi	...	...	German Democratic Republic	...	...	Plant architecture	...	Autogamous-SC	LYC 3476	EA03362	T1203	IPK Gatersleben	Download VCF File	Wageningen UR
15	Solanum lycopersicum	Momotaro (Tough Boy)	...	...	...	...	...	pink beef	...	Autogamous-SC	...	TR00003	...	Rijk Zwaan	Download VCF File	Wageningen UR
16	Solanum lycopersicum	Rote Beere	...	...	...	...	...	brix	...	Autogamous-SC	CGN15464	EA01965	LYC 11; 1.1422	CGN, TGRC	Download VCF File	Wageningen UR
17	Solanum lycopersicum	...	...	...	Cuba	...	...	brix	1986	Autogamous-SC	LYC 3340	EA03306	T1039	IPK Gatersleben	Download VCF File	Wageningen UR
18	Solanum lycopersicum	Dana	...	...	...	...	...	fruit size	...	Autogamous-SC	...	EA01155	...	Sand Hill Preservation Center	Download VCF File	Wageningen UR
19	Solanum lycopersicum	Large Pink	...	...	...	...	...	fruit weight	...	Autogamous-SC	...	EA01049	...	Sand Hill Preservation Center	Download VCF File	Wageningen UR
2	Solanum lycopersicum	Alisa Craig	...	...	...	...	...	old cultivar	...	Autogamous-SC	PV	several, EA06086	...	USDA Geneva, PBR	Download VCF File	Wageningen UR
20	Solanum lycopersicum	...	Pamplona	...	Spain	...	...	fruit weight	1960	Autogamous-SC	LYC 3153	EA03321	T 825, PI 262906	IPK Gatersleben	Download VCF File	Wageningen UR
21	Solanum lycopersicum	Bolivar	...	...	Cuba	...	...	fruit weight	...	Autogamous-SC	LYC 3155	EA3222	T828	IPK Gatersleben	Download VCF File	Wageningen UR
22	Solanum lycopersicum	...	...	...	Columbia	...	...	fruit weight	1938	Autogamous-SC	PI 129097	EA04710	...	USDA Geneva	Download VCF File	Wageningen UR

Metagenomics analysis

Microbial community composition and function insights

EBI Metagenomics

113993 data sets

By selected biomes

Soil (438)

Freshwater (118)

Biome	Project name	Samples	Last updated
	16S amplicon based soil and leaf microbiome survey in Hungarian vineyards	19	02-May-2017
	16S metabarcoding of bacteria associated with cultured strains of the brown alga Ectocarpus sp.	51	12-Jan-2017
	16S rRNA amplicons (V4 region) of bacteria living on and in roots and leaves of Boechera stricta from field experiments in the Rocky Mountains	650	13-Dec-2016
	16S rRNA gene pyrosequnecing- Secondary successional trajectories of structural and catabolic bacterial communities in oil-polluted soil planted with hybrid Poplar	34	12-Jan-2017
	A diverse array of bacteria that inhabit the rhizosphere and different plant organs play a crucial role in plant health and growth.	4	02-Dec-2016
	Accessing and Identification of Novel Environmental Alleles of the ACC Deaminase Domain Region through a Competition Assay	1	02-Dec-2016
	Agroforestry leads to shifts within the gammaproteobacterial microbiome of banana plants cultivated in Central America	48	05-Jan-2017
	Alk B pyrosequencing -Secondary successional trajectories of structural and catabolic bacterial communities in oil-polluted soil planted with hybrid Poplar	34	12-Jan-2017
	AMF from contaminated and uncontaminated rhizosphere soils Metagenome	70	16-May-2016
	Amplicon-based metagenomics analysis of Vitis vinifera L. cv. Corvina grapes and fresh musts	39	08-Sep-2016

www.ebi.ac.uk/metagenomics

Where are reference genome sequences?

D36–D42 Nucleic Acids Research, 2013, Vol. 41, Database issue *Published online 27 November 2012*
doi:10.1093/nar/gks1195

GenBank

Dennis A. Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell and Eric W. Sayers*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received September 28, 2012; Revised and Accepted October 29, 2012

ABSTRACT

GenBank® (http://www.ncbi.nlm.nih.gov) is a comprehensive database that contains publicly available nucleotide sequences for almost 300,000 formally described species. These sequences are obtained primarily through submissions from individual laboratories and batch submissions from large-scale sequencing projects, including whole-genome shotgun (WGS) and environmental sampling projects. Most submissions are made using the web-based Bankit or standalone Sequin programs, and GenBank staff assigns accession numbers upon data receipt. Daily data exchange with the European Nucleotide Archive (ENA) and the DNA

sequence (GSS), whole-genome shotgun (WGS) and other high-throughput data from sequencing centres. The U.S. Patent and Trademark Office also contributes sequences from issued patents. GenBank participates with the European Molecular Biology Laboratory-Nucleotide Sequence Database (EMBL-Bank), part of the European Nucleotide Archive (ENA), and the DNA Data Bank of Japan (DDBJ) as a partner in the International Nucleotide Sequence Database Collaboration (INSDC). The INSDC partners exchange data daily to ensure that a uniform and comprehensive collection of sequence information is available worldwide. NCBI makes the GenBank data available at no cost over the Internet, through FTP and a wide range of web-based retrieval and analysis services (4).

NCBI Resources How To

Genome

Genome rice[orgn]

Create alert Limits Advanced

Oryza sativa (rice)
Reference genome: Oryza sativa Japonica Group (assembly Build 4.0)
Download sequences in FASTA format for **genome**, **transcript**, **protein**
Download genome annotation in **GFF**, **GenBank** or **tabular** format
BLAST against Oryza sativa **genome**, **transcript**, **protein**
All 21 genomes for species:
Browse the **list**
Download sequence and annotation from **RefSeq** or **GenBank**

Display Settings: Overview Send to:

Organism Overview ; Genome Assembly and Annotation report [21] ; Plasmid Annotation Report [2] ; Organelle Annotation Report [7] ID: 10

Oryza sativa (rice)

Oryza sativa Organism overview

Lineage: Eukaryota[2171]; Viridiplantae[233]; Streptophyta[204]; Embryophyta[203]; Tracheophyta[201]; Spermatophyta[199]; Magnoliophyta[194]; Liliopsida[43]; Poales[32]; Poaceae[31]; BOP clade[22]; Oryzoideae[15]; Oryzeae[15]; Oryzinae[14]; Oryza[13]; Oryza sativa[1]

Rice is one of the most important food crops in the world and feeds more people than any other crop. Rice belongs to the genus *Oryza* which includes approximately 24 species. They are widely distributed growing in different habitats and different soil types. They show differences in plant growth, yield, pest and disease resistance, stress tolerance More...

Summary

Sequence data: genome assemblies: 21; sequence reads: 1596 (See Genome Assembly and Annotation report)
Statistics: median total length (Mb): 362.279
median protein count: 36376
median GC%: 43.3432

Publications

1. Indica rice genome assembly, annotation and mining of blast disease resistance genes. Mahesh HB, et al. BMC Genomics 2016 Mar 16
2. Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. Sakai H, et al. Plant Cell Physiol 2013 Feb

f the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Kawahara Y, et 013 Feb 6

3. Improvement of al. Rice (N Y) 2

Sequence alignment

"Most common but one of the most powerful process in Bioinformatics"

nucleotide or amino acid residues

Global alignment :

Suite for similar sequences

" span the entire length of all query sequences"

Local alignment

"identify regions of similarity within long sequences

that are often widely divergent overall"

Local vs. Global Alignment

- Global Alignment

```
--T—-CC-C-AGT—-TATGT-CAGGGGACACG—A-GCATGCAGA-GAC
  |  || |  ||  | | | |||    || | | |  | ||||   |
AATTGCCGCC-GTCGT-T-TTCAG----CA-GTTATG—T-CAGAT--C
```

- Local Alignment—better alignment to find conserved segment

```
                    TCCCAGTTATGTCAGGGGACACGAGCATGCAGAGAC
                       ||||||||||||
AATTGCCGCCGTCGTTTTCAGCAGTTATGTCAGATC
```

Sequence alignment

A consequence of functional

Structural

Evolutionary

A common ancestor, mismatches can be interpreted as point mutation and gaps as indels introduced in one or both lineages in the time since they diverged from one another.

FASTA format

Text based format

```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAKKADRLAAEG
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKDPNKPEHK
IPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQLDSKLTL
MGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKTEDFAAEVAAQL
>SEQUENCE_2
SATVSEINSETDFVAKNDQFIALTKDTTAHIQSNSLQSVEELHSSTINGVKFEEYLKSQI
ATIGENLVVRRFATLKAGANGVVNGYIHTNGRVGVVIAAACDSAEVASKSRDLLRQICMH
```

Common identifier

Database	Format
GenBank	`gb|accession|locus`
EMBL Data Library	`emb|accession|locus`
DDBJ, DNA Database of Japan	`dbj|accession|locus`
NBRF PIR	`pir||entry`
Protein Research Foundation	`prf||name`
SWISS-PROT	`sp|accession|entry name`
Brookhaven Protein Data Bank	`pdb|entry|chain`
Patents	`pat|country|number`
GenInfo Backbone Id	`bbs|number`
General database identifier	`gnl|database|identifier`
NCBI Reference Sequence	`ref|accession|locus`
Local Sequence identifier	`lcl|identifier`

Fasta file extension

Extension	Meaning	Notes
fasta	generic fasta	Any generic fasta file. Other extensions can be fas, fa, seq, fsa
fna	fasta nucleic acid	Used generically to specify nucleic acids.
ffn	FASTA nucleotide of gene regions	Contains coding regions for a genome.
faa	fasta amino acid	Contains amino acids. A multiple protein fasta file can have the more specific extension mpfa.
frn	FASTA non-coding RNA	Contains non-coding RNA regions for a genome, in DNA alphabet e.g. tRNA, rRNA

“Extract a small part from long sequence”

How to do alignment ?

BLAST

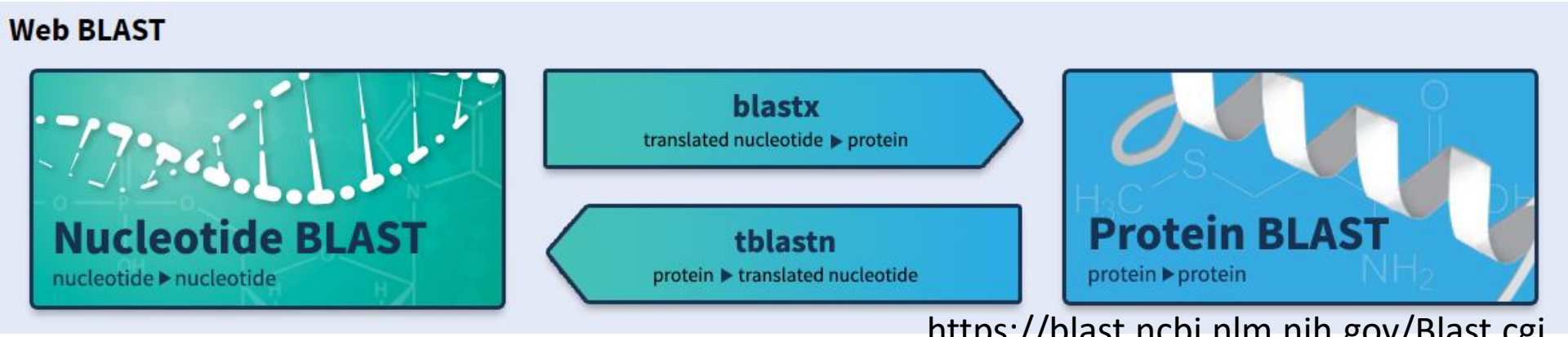
PREDICTED: Sus scrofa uncharacterized LOC100736873 (LOC100736873), mRNA

Sequence ID: ref|XM_005668591.1| Length: 1294 Number of Matches: 1

Range 1: 137 to 282 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
222 bits(120)	1e-54	138/146(95%)	3/146(2%)	Plus/Plus

```
Query  38   CAGAGGAGCCAACGGTGCTGATCTGGTTTGTTCTGGACAAAAGGAcccccc-gccc-cc  95
            |||| |||||||||||||||||||||||||||||||||||||||||||||||||  ||| ||
Sbjct  137  CAGAAGAGCCAACGGTGCTGATCTGGTTTGTTCTGGACAAAAGGACCCCCCCCCCCCACC  196

Query  96   g-cccGCCACTGCCAAGCCCAAACTTCACAGCGACACGTGGGACGAAAGCAGCCGGGCCC  154
            | |||||||| || ||||||||||||||||||||| |||||||||||||||||||||||||
Sbjct  197  GCCCCGCCACCGCAAAGCCCAAACTTCACAGCGACGCGTGGGACGAAAGCAGCCGGGCCC  256

Query  155  CCGCCCTGCCGCCGCGCCCAGCCCGT  180
            ||||||||||||||||||||||||||
Sbjct  257  CCGCCCTGCCGCCGCGCCCAGCCCGT  282
```



https://blast.ncbi.nlm.nih.gov/Blast.cgi

Command-line BLAST+

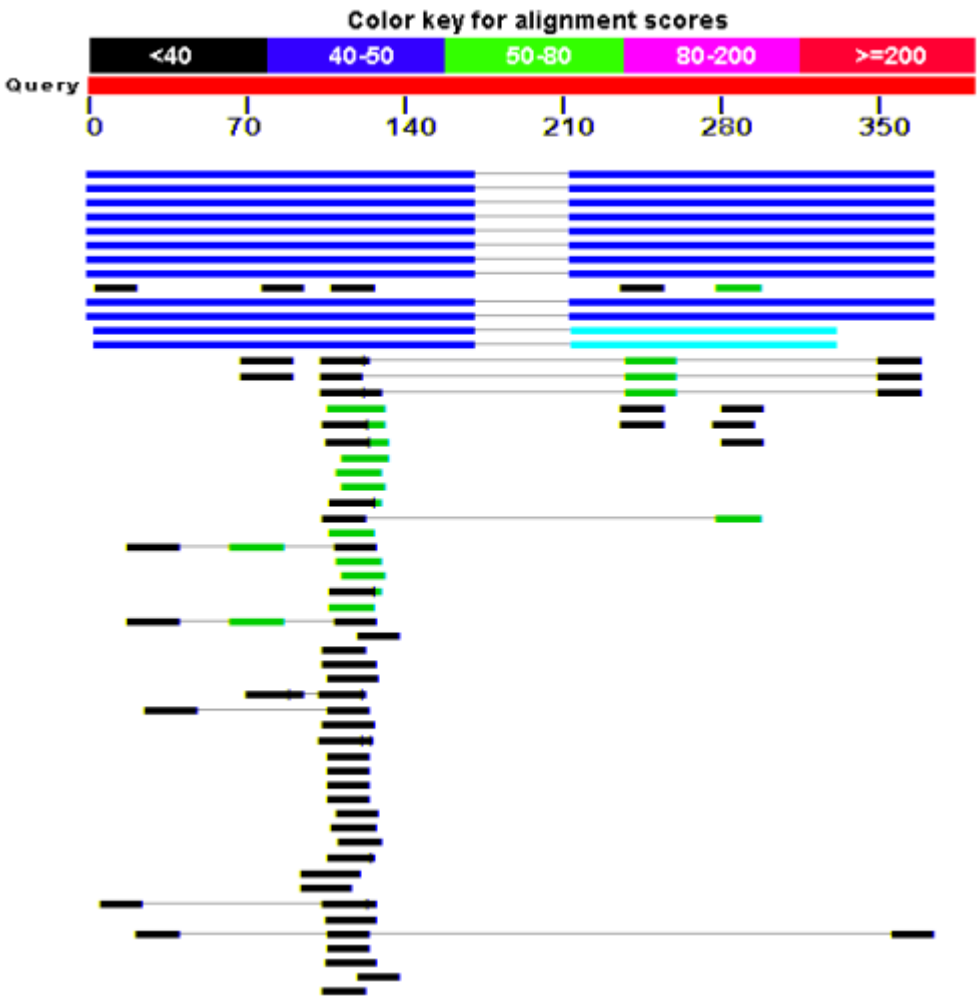
ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

Custom database using set of nucleotides/proteins

```
[pumie@agcipher aew]$ makeblastdb --help
USAGE
  makeblastdb [-h] [-help] [-in
    -dbtype molecule_type [-tit
    [-hash_index] [-mask_data m
    [-mask_desc mask_algo_descr
    [-gi_mask_name gi_based_mas
    [-max_file_sz number_of_byt
    [-taxid_map TaxIDMapFile] [

DESCRIPTION
   Application to create BLAST

Use '-help' to print detailed d
================================
```

```
[pumie@agcipher aew]$ blastn
BLAST query/options error: Either a BLAST database or subject sequence(s) must be specified
Please refer to the BLAST+ user manual.
[pumie@agcipher aew]$ blastn -h
USAGE
  blastn [-h] [-help] [-import_search_strategy filename]
    [-export_search_strategy filename] [-task task_name] [-db database_name]
    [-dbsize num_letters] [-gilist filename] [-seqidlist filename]
    [-negative_gilist filename] [-entrez_query entrez_query]
    [-db_soft_mask filtering_algorithm] [-db_hard_mask filtering_algorithm]
    [-subject subject_input_file] [-subject_loc range] [-query input_file]
    [-out output_file] [-evalue evalue] [-word_size int_value]
    [-gapopen open_penalty] [-gapextend extend_penalty]
    [-perc_identity float_value] [-qcov_hsp_perc float_value]
    [-max_hsps int_value] [-xdrop_ungap float_value] [-xdrop_gap float_value]
    [-xdrop_gap_final float_value] [-searchsp int_value]
    [-sum_stats bool_value] [-penalty penalty] [-reward reward] [-no_greedy]
    [-min_raw_gapped_score int_value] [-template_type type]
    [-template_length int_value] [-dust DUST_options]
    [-filtering_db filtering_database]
    [-window_masker_taxid window_masker_taxid]
    [-window_masker_db window_masker_db] [-soft_masking soft_masking]
    [-ungapped] [-culling_limit int_value] [-best_hit_overhang float_value]
    [-best_hit_score_edge float_value] [-window_size int_value]
    [-off_diagonal_range int_value] [-use_index boolean] [-index_name string]
    [-lcase_masking] [-query_loc range] [-strand strand] [-parse_deflines]
    [-outfmt format] [-show_gis] [-num_descriptions int_value]
    [-num_alignments int_value] [-line_length line_length] [-html]
    [-max_target_seqs num_sequences] [-num_threads int_value] [-remote]
    [-version]

DESCRIPTION
   Nucleotide-Nucleotide BLAST 2.6.0+
```

BLAST output



Bit score (S)

Expect value (E-value)

$$E = m\,n\,2^{-S'}$$

Nucleotide :1E-3

Amino : 1E-6

Identities

Gaps

Strand

Causes for sequence (dis)similarity

mutation: a nucleotide at a certain location is replaced by ***an ther*** nucleotide (e.g.: A**T**A → A**G**A)

insertion: at a certain location one new nucleotide is inserted inbetween two existing nucleotides (e.g.: AA → A**G**A)

deletion: at a certain location one existing nucleotide is deleted (e.g.: AC**T**G → AC**-**G)

indel: an **in**sertion or a **del**etion

Sequence alignment for NGS data

Alignment; “**mapping**”

Re-sequencing

Reference sequence -> Genome

Detecting variation in samples

Allow mismatch alignment

```
GCTGATGTGCCGCCTCACTTCGGTGGTGAGGTG  Reference sequence
 CTGATGTGCCGCCTCACTTCGGTGGT        Short read 1
  TGATGTGCCGCCTCACTACGGTGGTG       Short read 2
   GATGTGCCGCCTCACTTCGGTGGTGA      Short read 3
GCTGATGTGCCGCCTCACTACGGTG          Short read 4
GCTGATGTGCCGCCTCACTACGGTG          Short read 5
```

Alignment types (NGS applications)

.

sequence all DNA from an organism and map it to the appropriate reference sequence, to find genetic variation.

.

For large genomes (e.g., human), capture just the exomic DNA before sequencing.

.

Mapping can be done either to the full reference sequence, or to a special "transcriptome reference".

.

Multiple sequence alignment (MSA)

used in identifying conserved sequence regions

establishing evolutionary relationships by constructing phylogenetic trees

"alignment space“

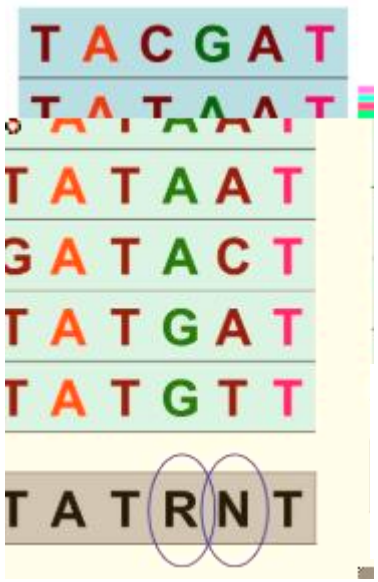
computationally expensive in both time and memory

Aim of MSA

Grouping samples

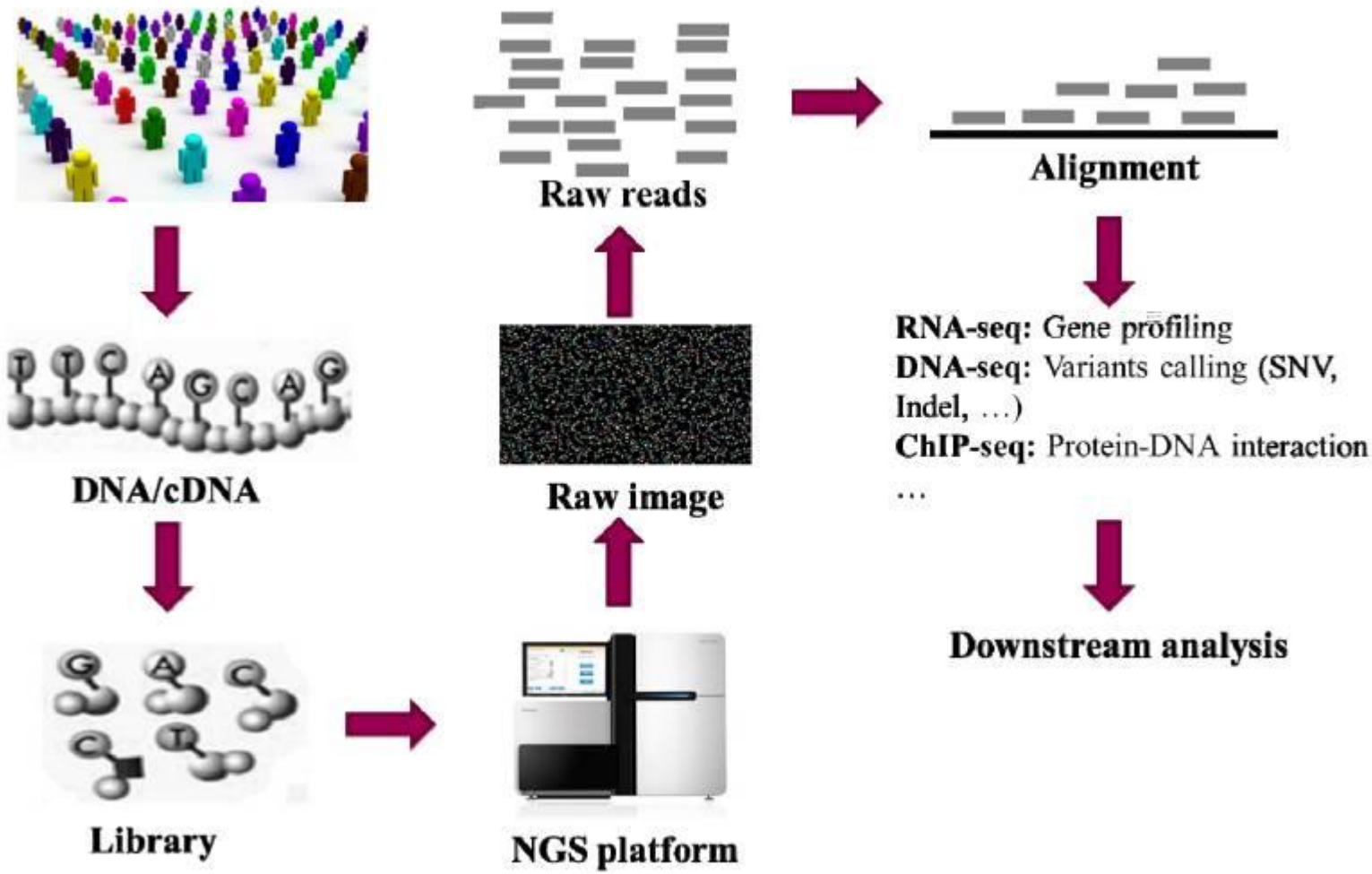
Consensus sequence

```
sp|P35547|282-314      CLVCSDEASGCHYGVLTCGSCKVFFKRAVEG---------QH------------------ 33
sp|O46567|421-486      CLVCSDEASGCHYGVLTCGSCKVFFKRAVEG---------QHNYLCAGRNDCIIDKIRRK 51
sp|Q9N1U3|426-491      CLVCSDEASGCHYGVLTCGSCKVFFKRAVEG---------QHNYLCAGRNDCIIDKIRRK 51
sp|P06536|440-505      CLVCSDEASGCHYGVLTCGSCKVFFKRAVEG---------QHNYLCAGRNDCIIDKIRRK 51
sp|P04150|421-486      CLVCSDEASGCHYGVLTCGSCKVFFKRAVEG---------QHNYLCAGRNDCIIDKIRRK 51
sp|P4983|387-461       CLVCSDEASGCHYGVLTCGSCKVFFKRAVEGWRARQNTDGQHNYLCAGRNDCIIDKIRRK 60
                       ******************************         **

sp|P35547|282-314      ---------------
sp|O46567|421-486      NCPACRYRK------ 60
sp|Q9N1U3|426-491      NCPACRYRKCLQAGM 66
sp|P06536|440-505      NCPACRYRKCLQAGM 66
sp|P04150|421-486      NCPACRYRKCLQAGM 66
sp|P4983|387-461       NCPACRFRKCLQAGM 75
```



Y = Pyrimidine
R = Purine
N = Any nucleotide

A brief flow chart of genetic studies using NGS



Ye H, et al. Pharmaceutics. 2015 Nov 23;7(4):523-41

Get to know NGS

File format: FAST


```
@SRR038845.3 H
CAACGAGTTCACAC
+SRR038845.3 H
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
@SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
BBCBBBBBB@@BAB?BBBBCBC>BBBAA8>BBBAA@
```

sequence

quality

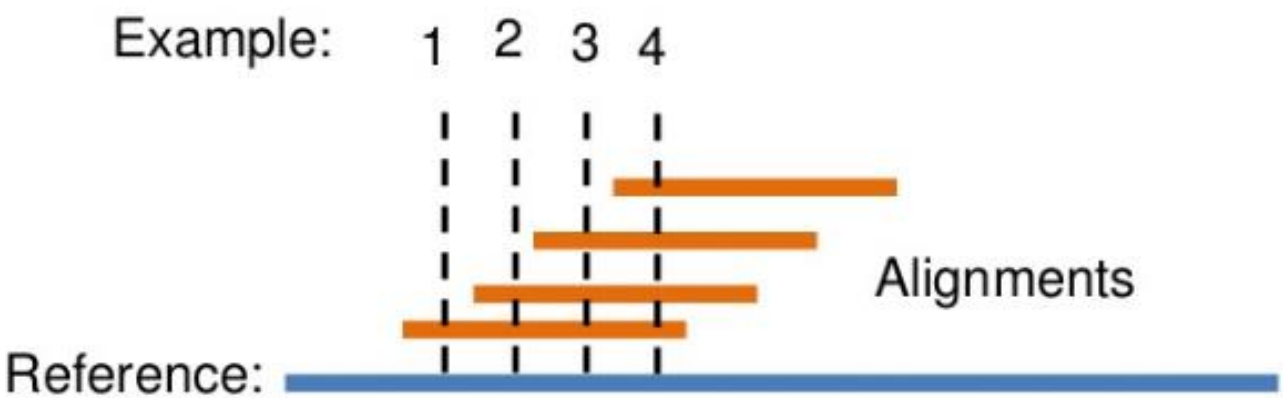
one read

NCBI Sequence Read Archive

A format for NGS read (FASTQ + quality)

Get to know NGS

Coverage/Depth



"Coverage" is simply the average number of reads that overlap each true base in genome.

Get to know NGS

What is a **base quality**?

- Give a base calling error information.
- The first is the standard Sanger known as Phred quality (Q)

Base Quality (Q)	P_{error}(obs. base)	Base call accuracy
3	50 %	50%
5	32 %	68%
10	10 %	90%
20	1 %	99%
30	0.1 %	99.9%
40	0.01 %	99.99%

Q scores and ASCII characters

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

ASCII_BASE=64 Old Illumina

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	64 @	11	0.07943	75 K	22	0.00631	86 V	33	0.00050	97 a
1	0.79433	65 A	12	0.06310	76 L	23	0.00501	87 W	34	0.00040	98 b
2	0.63096	66 B	13	0.05012	77 M	24	0.00398	88 X	35	0.00032	99 c
3	0.50119	67 C	14	0.03981	78 N	25	0.00316	89 Y	36	0.00025	100 d
4	0.39811	68 D	15	0.03162	79 O	26	0.00251	90 Z	37	0.00020	101 e
5	0.31623	69 E	16	0.02512	80 P	27	0.00200	91 [	38	0.00016	102 f
6	0.25119	70 F	17	0.01995	81 Q	28	0.00158	92 \	39	0.00013	103 g
7	0.19953	71 G	18	0.01585	82 R	29	0.00126	93]	40	0.00010	104 h
8	0.15849	72 H	19	0.01259	83 S	30	0.00100	94 ^	41	0.00008	105 i
9	0.12589	73 I	20	0.01000	84 T	31	0.00079	95 _	42	0.00006	106 j
10	0.10000	74 J	21	0.00794	85 U	32	0.00063	96 `			

How to check NGS data quality

FASTQC

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

A Java Runtime Environment

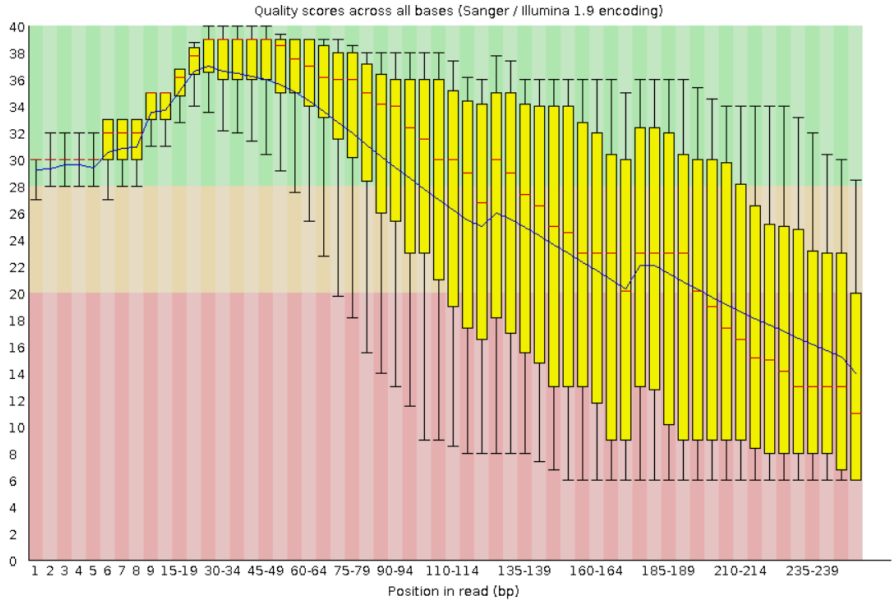
The main functions of FastQC are

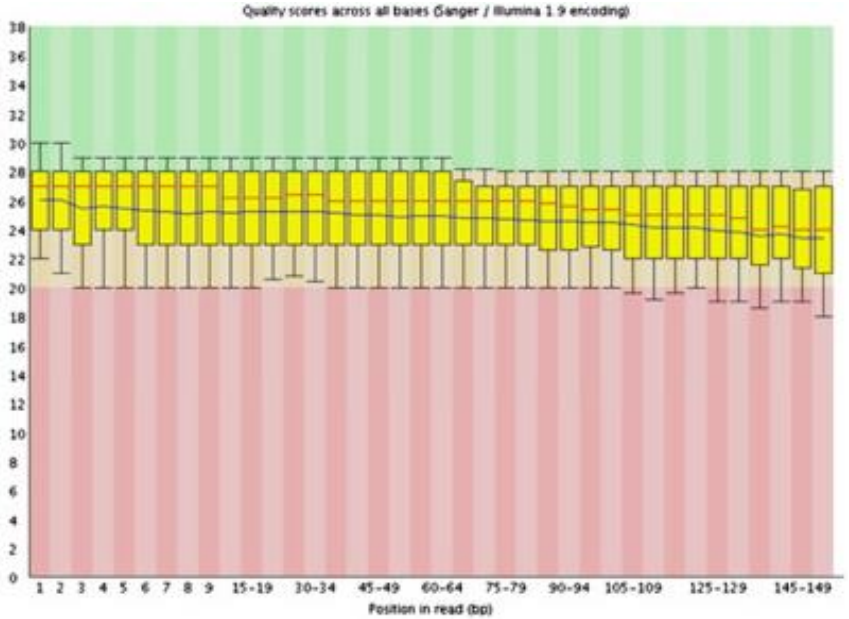
- Import of data from BAM, SAM or FastQ files (any variant)
- Quick overview
- Summary graphs and tables
- HTML based permanent report
- Offline operation

Average Q scores is a bad idea

Q scores in read	Avg. Q	Expected number of errors
140 x Q35 + 10 x Q2	33	**6.4 !**
150 x Q25	25	0.5

Per base sequence quality



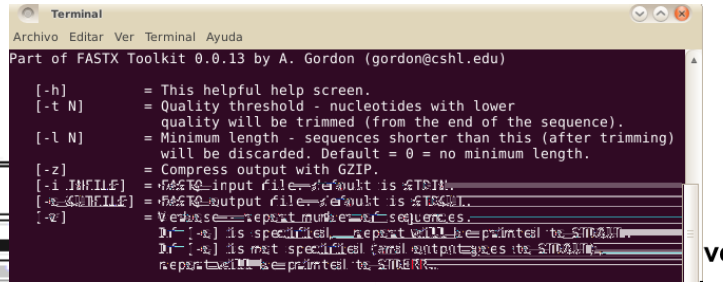


QC and sequence manipulation

FASTX-Toolkit Short-Reads FASTA/FASTQ files preprocessing suite

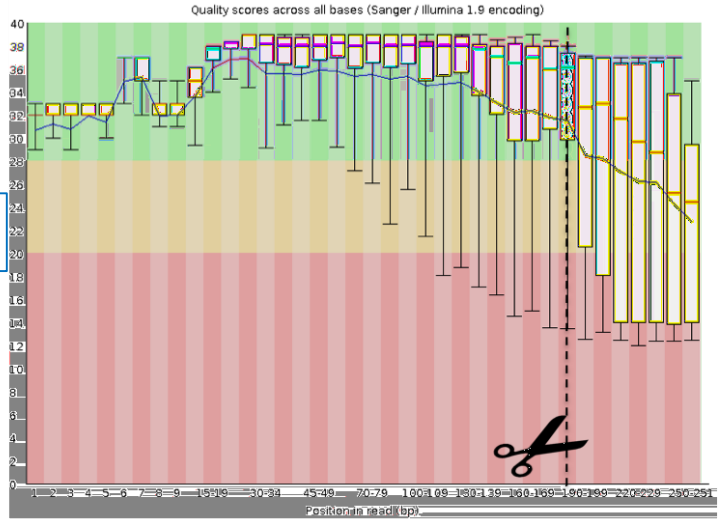
Command Line Arguments

- FASTQ-to-FASTA
- FASTQ/A Quality Statistics
- FASTQ Quality chart
- [illegible]
- [illegible]
- [illegible]
- [illegible]
- [illegible]
- [illegible]
- [illegible]



verse Run

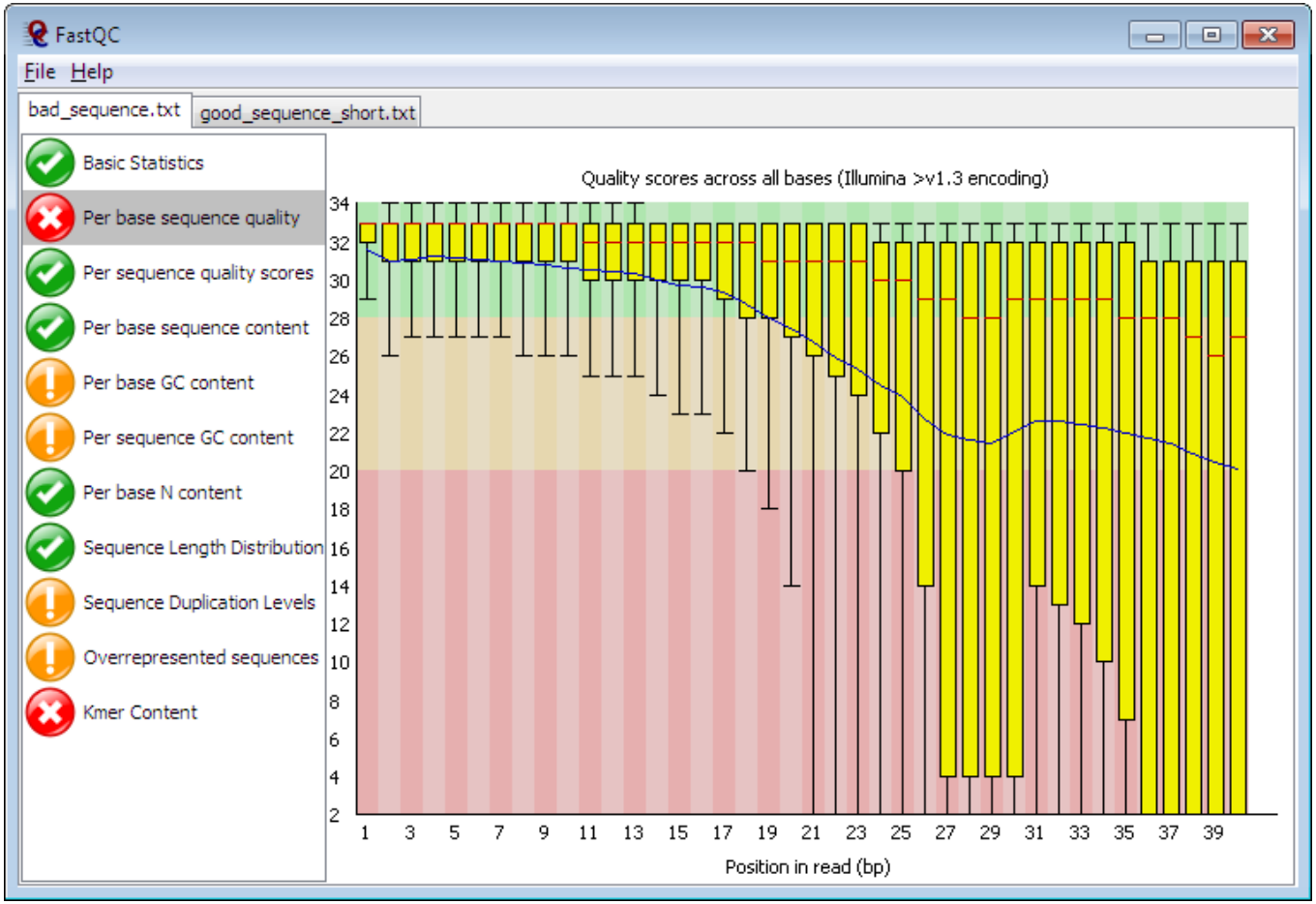
fastx_trimmer



Joshi NA, Fass JN. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software].

QC and sequence manipulation

FASTQC



FASTQC

FastQC Report

Good data

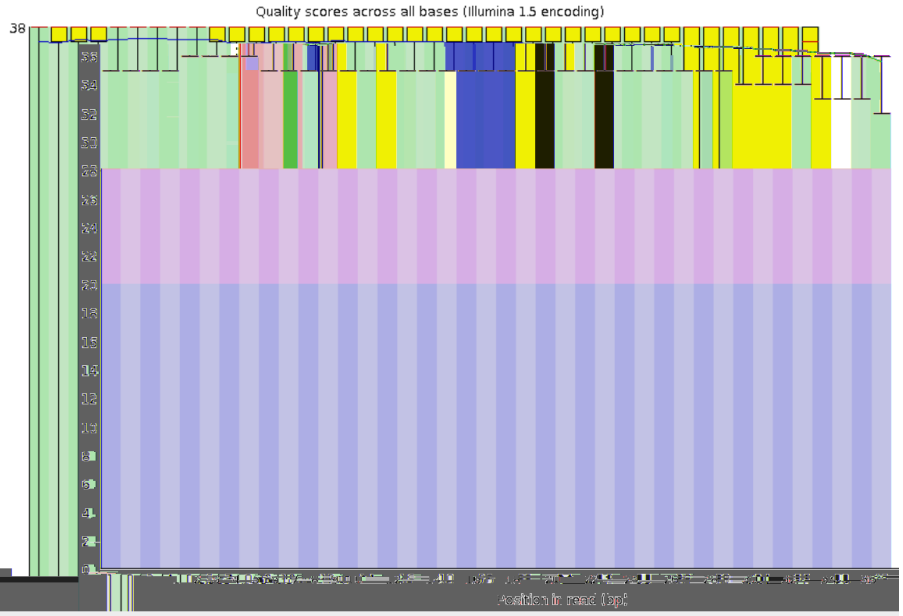
Summary

- 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

Basic Statistics

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Per base sequence quality



FASTQC

FastQC Report

Bad data

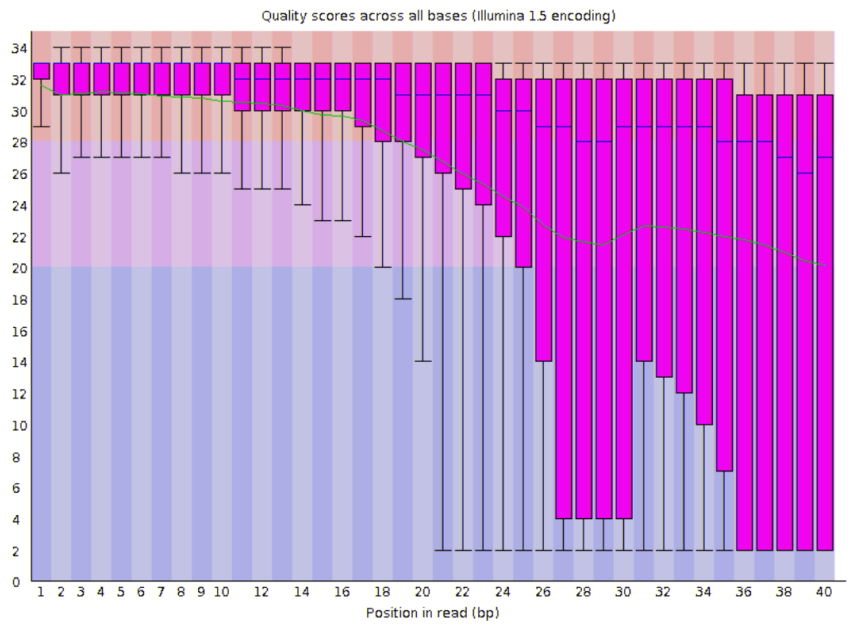
Summary

- 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

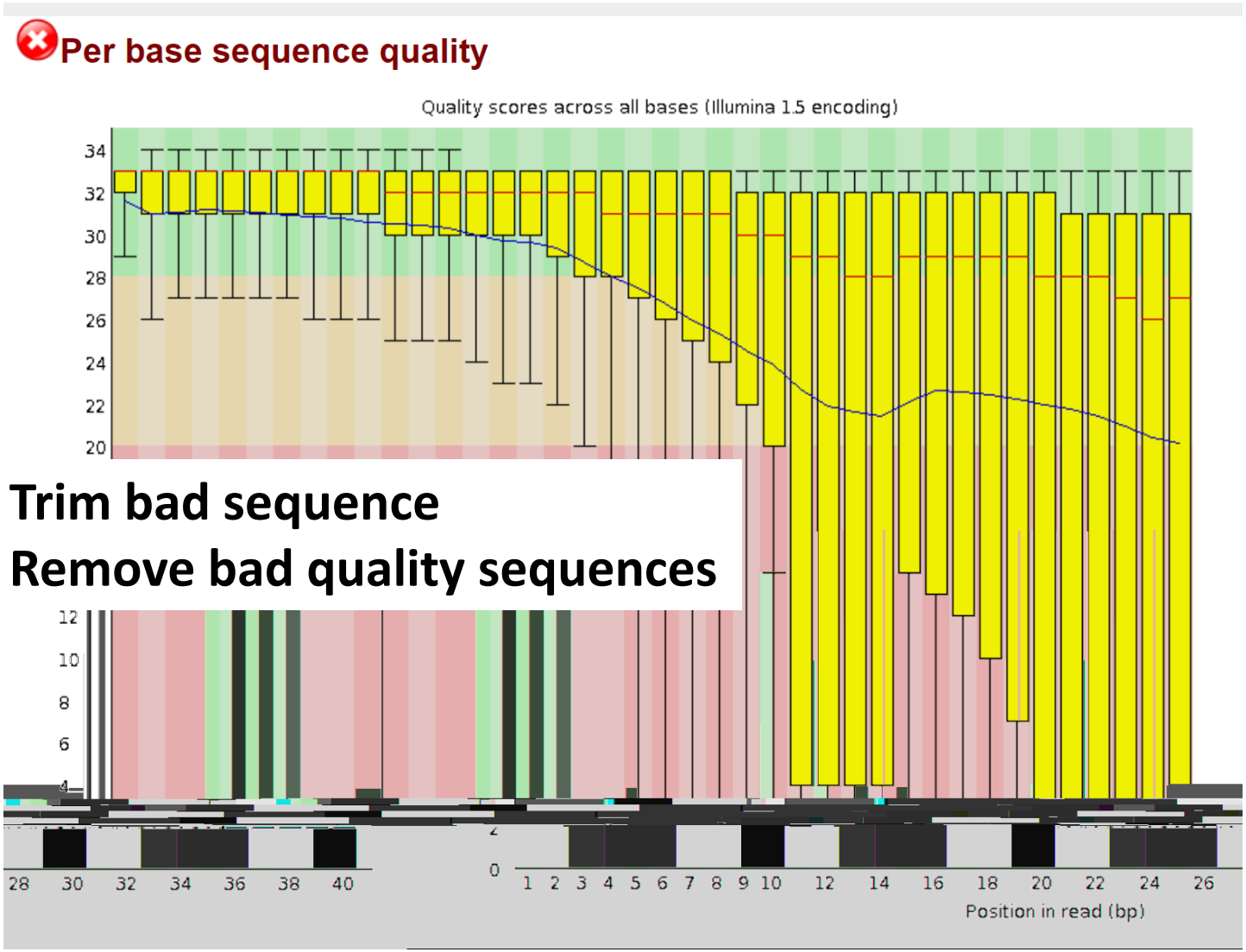
Basic Statistics

Measure	Value
Filename	bad_sequence.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	395288
Sequences flagged as poor quality	0
Sequence length	40
%GC	47

Per base sequence quality



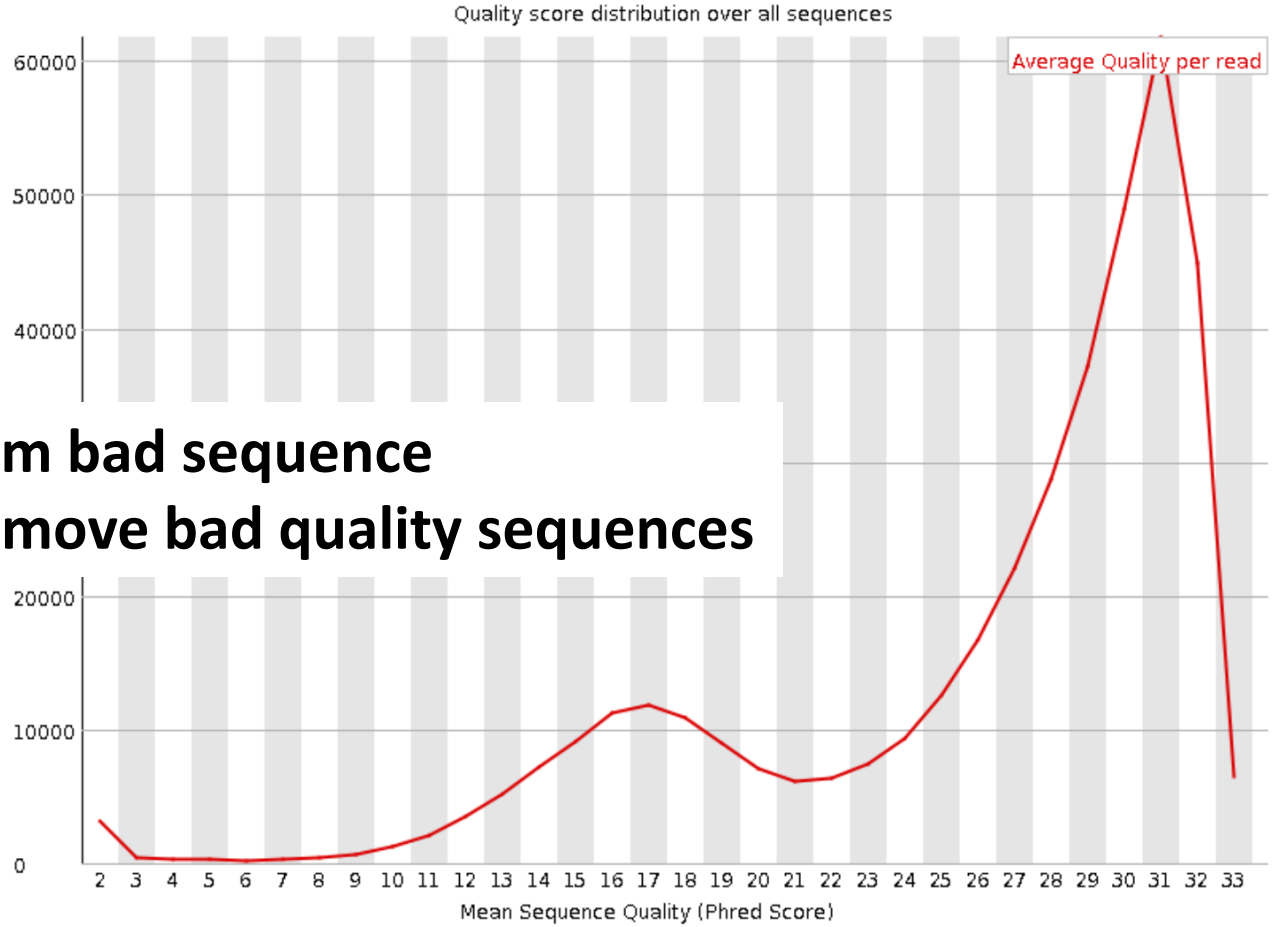
FASTQC



- **Trim bad sequence**
- **Remove bad quality sequences**

FASTQC

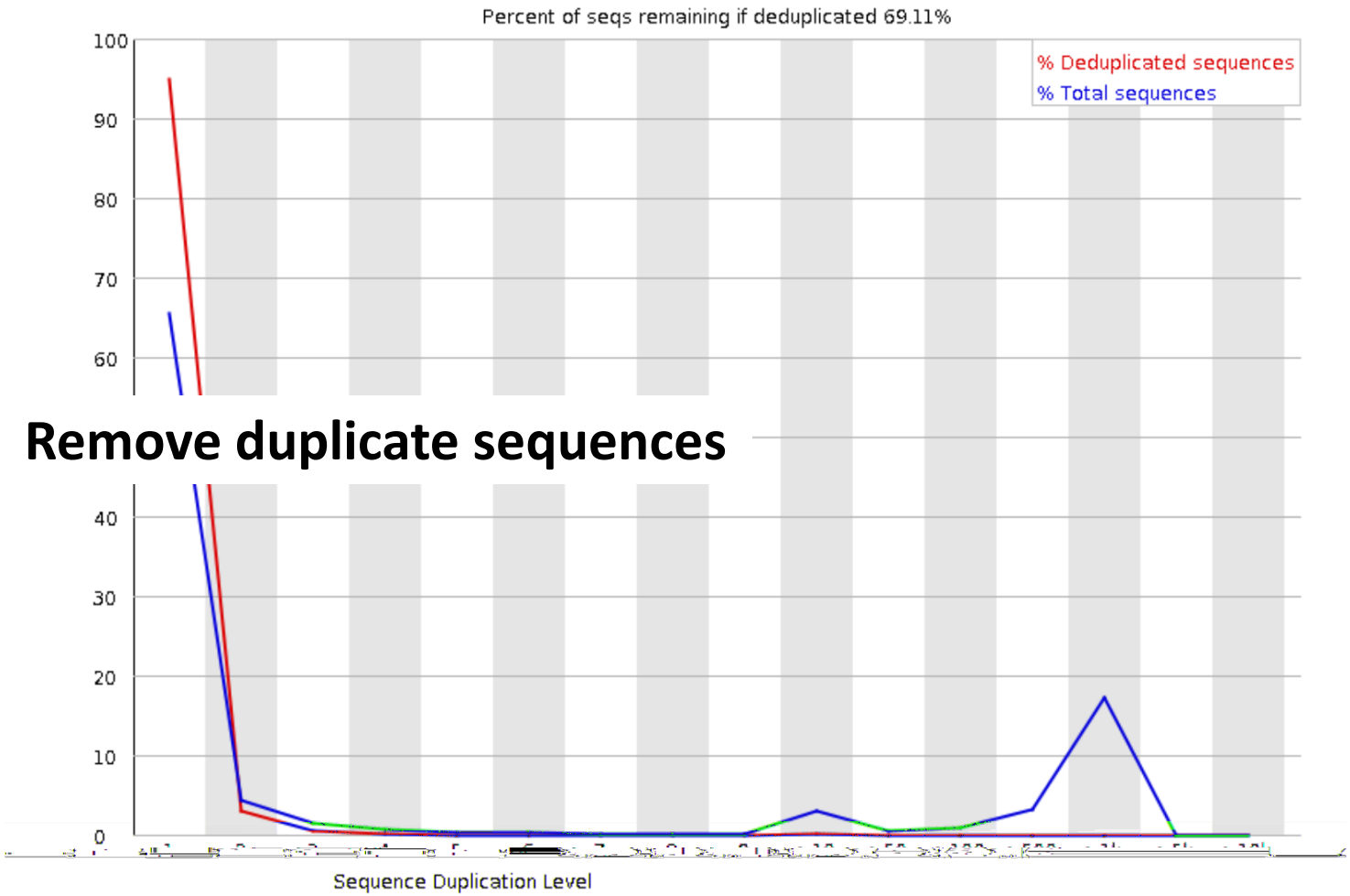
Per sequence quality scores



- **Trim bad sequence**
- **Remove bad quality sequences**

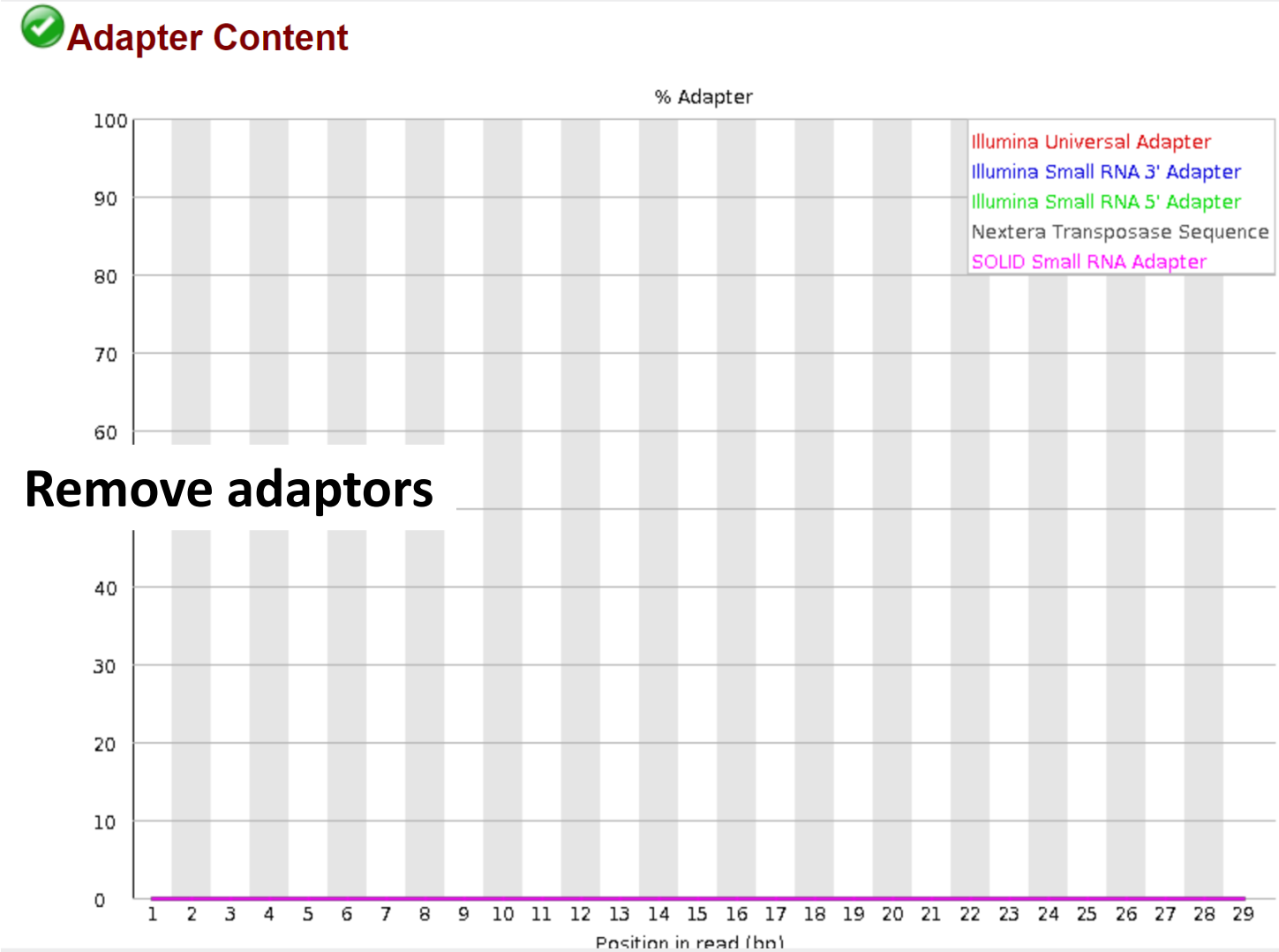
FASTQC

Sequence Duplication Levels



- **Remove duplicate sequences**

FASTQC

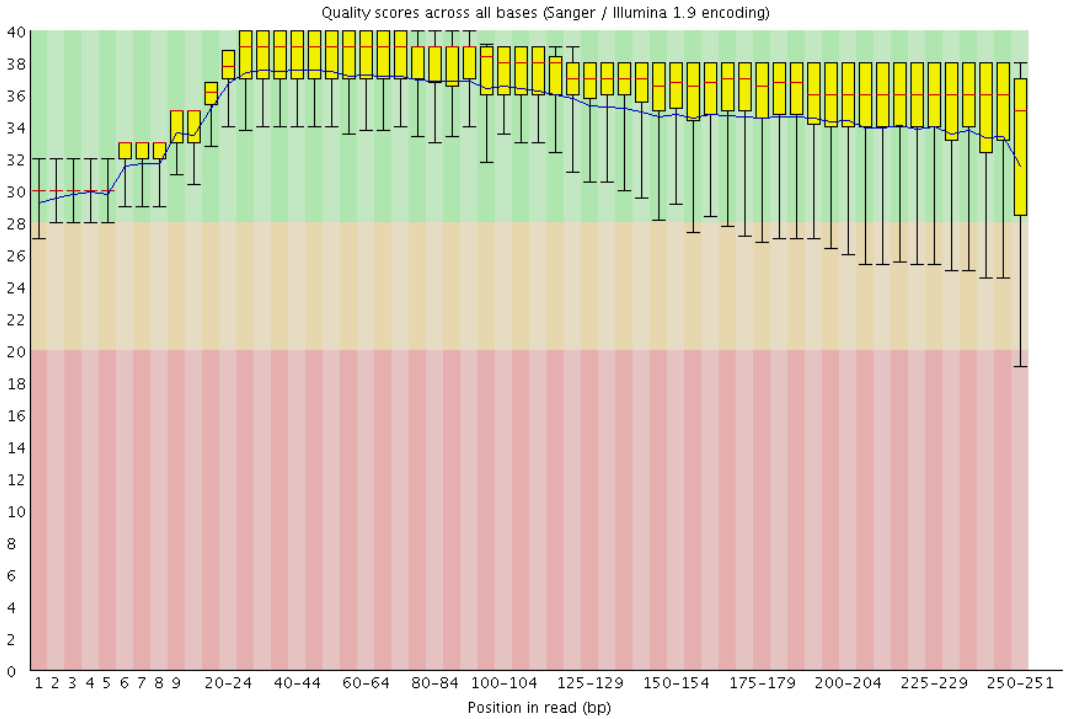


- **Remove adaptors**

QC and sequence manipulation

Sickle A windowed adaptive trimming tool for FASTQ files using quality

Available at https://github.com/najoshi/sickle.



Joshi NA, Fass JN. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software].

Alignment; “**mapping**”

Re-sequencing

Reference sequence -> Genome

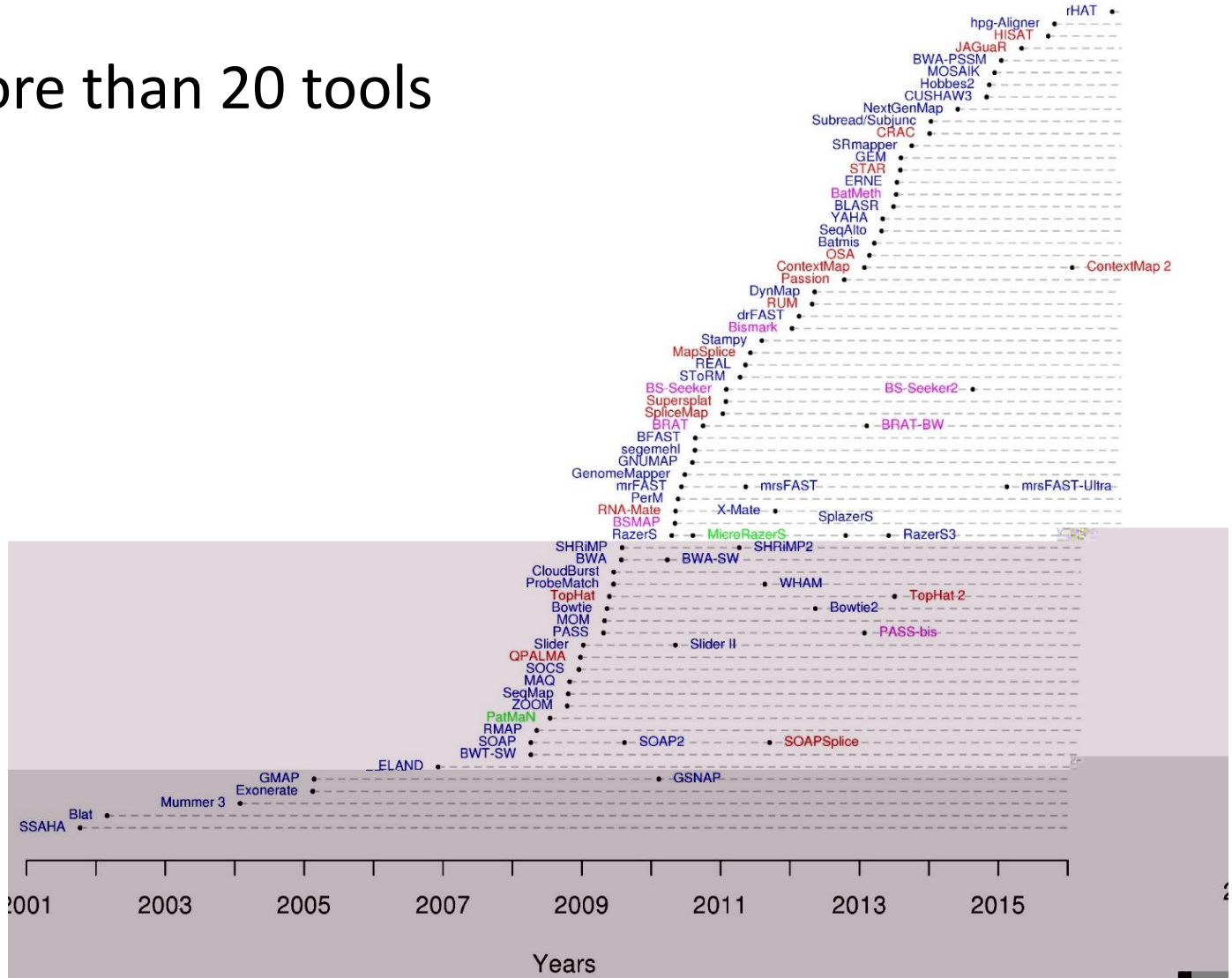
Detecting variation in samples

Allow mismatch alignment

```
GCTGATGTGCCGCCTCACTTCGGTGGTGAGGTG  Reference sequence
 CTGATGTGCCGCCTCACTTCGGTGGT        Short read 1
  TGATGTGCCGCCTCACTACGGTGGTG       Short read 2
   GATGTGCCGCCTCACTTCGGTGGTGA      Short read 3
GCTGATGTGCCGCCTCACTACGGTG          Short read 4
GCTGATGTGCCGCCTCACTACGGTG          Short read 5
```

Timeline of NGS read aligners

More than 20 tools



Read mapping/alignment

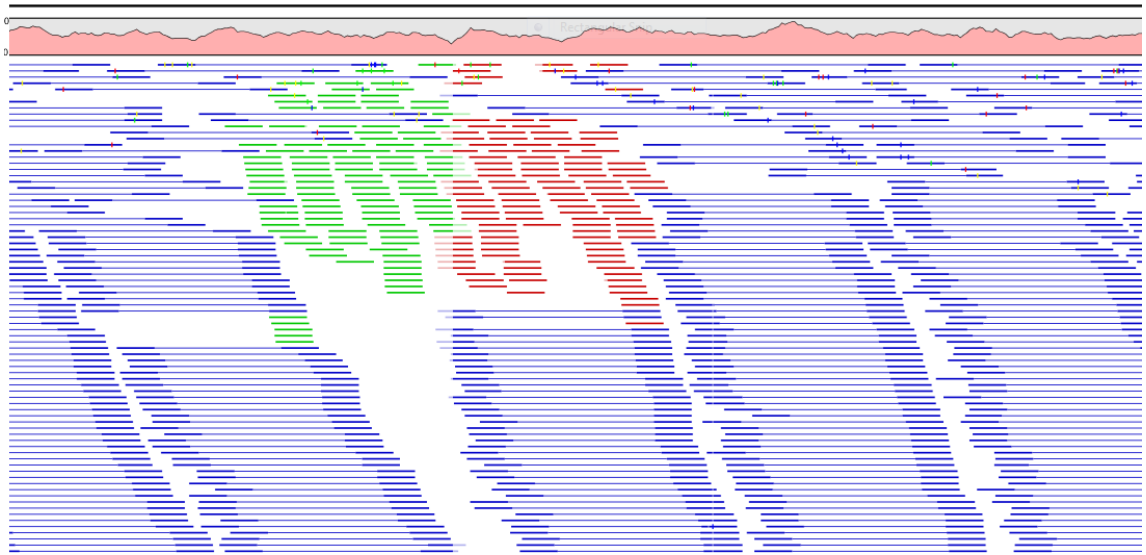
Bowtie

BWA / Burrows-Wheeler Aligner

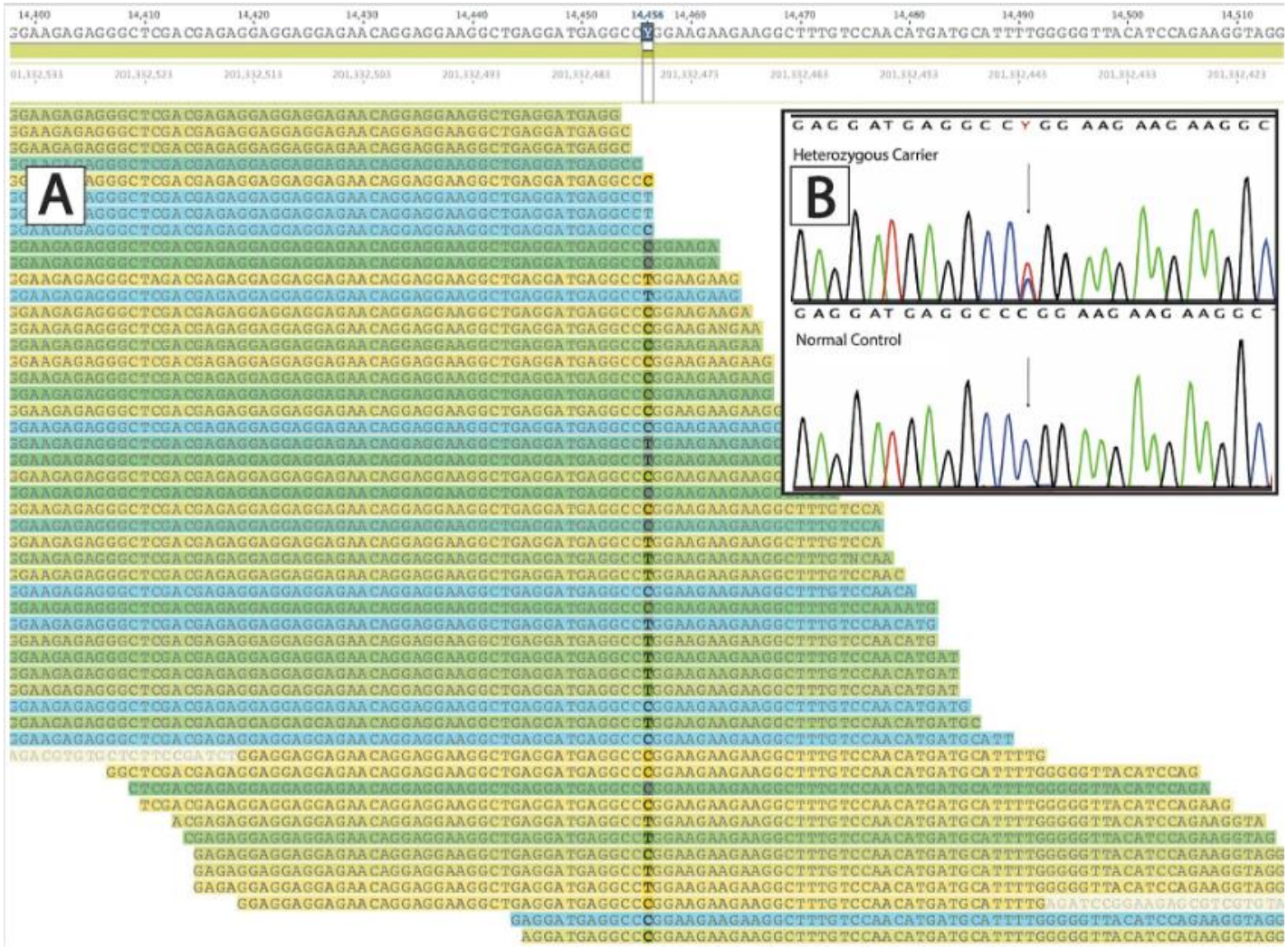
MAQ Mapping and Assembly with Quality

BFAST

SOAP



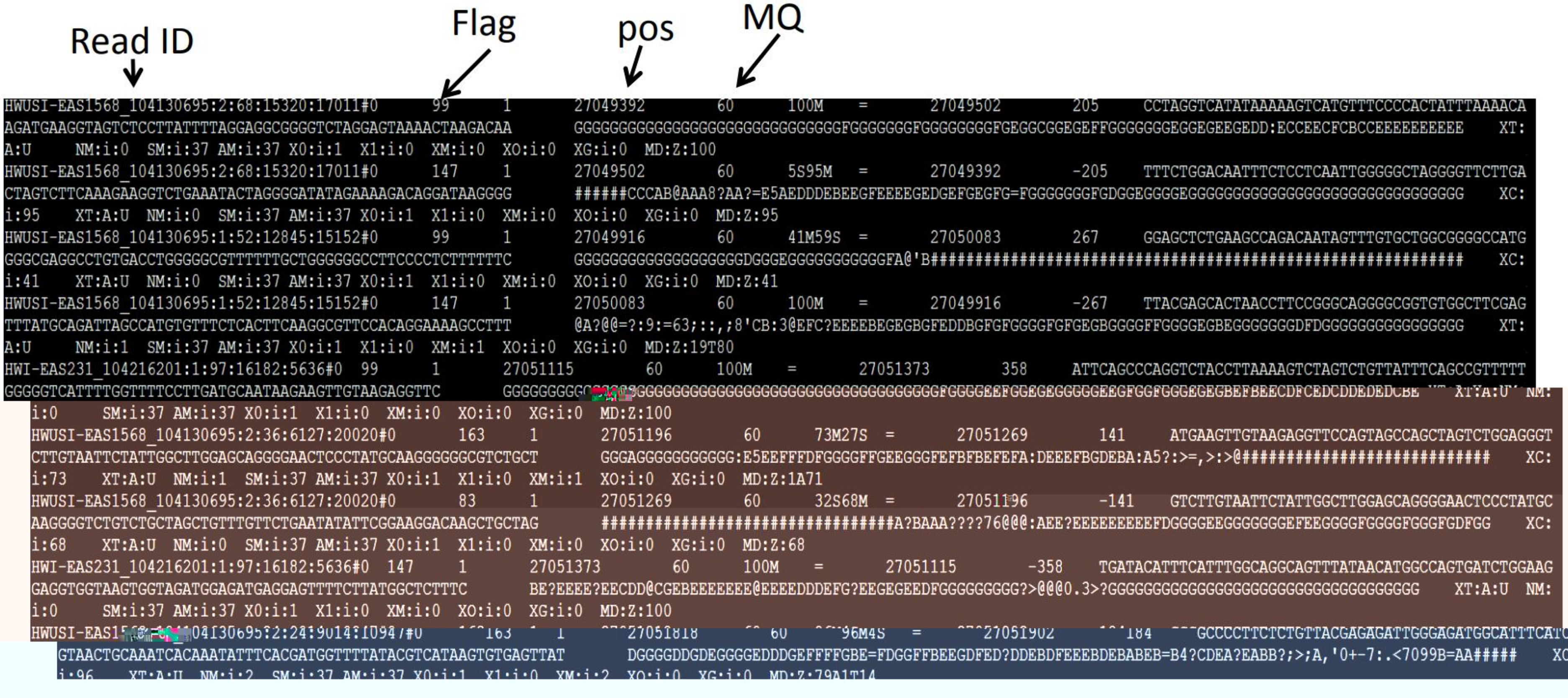
Aligned reads



NGS alignment format

The SAM/BAM format

“samtools”



```
HWUSI-EAS1568_104130695:2:68:15320:17011#0      99      1       27049392        60      100M    =       27049502        205     CCTAGGTCATATAAAAAGTCATGTTTCCCCACTATTTAAAACA
AGATGAAGGTAGTCTCCTTATTTTAGGAGGCGGGGTCTAGGAGTAAAACTAAGACAA       GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGFGGGGGGGGFGEGGCGGEGEFFGGGGGGGEGGEGEEGEDD:ECCEECFCBCCEEEEEEEEEE    XT:
A:U     NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:100
HWUSI-EAS1568_104130695:2:68:15320:17011#0      147     1       27049502        60      5S95M   =       27049392        -205    TTTCTGGACAATTTCTCCTCAATTGGGGGCTAGGGGTTCTTGA
CTAGTCTTCAAAGAAGGTCTGAAATACTAGGGGATATAGAAAAGACAGGATAAGGGG       ######CCCAB@AAA8?AA?=E5AEDDDEBEEGFEEEEGEDGEFGEGFG=FGGGGGGGFGDGGEGGGGEGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG    XC:
i:95    XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:95
HWUSI-EAS1568_104130695:1:52:12845:15152#0      99      1       27049916        60      41M59S  =       27050083        267     GGAGCTCTGAAGCCAGACAATAGTTTGTGCTGGCGGGGCCATG
GGGCGAGGCCTGTGACCTGGGGGCGTTTTTTGCTGGGGGGCCTTCCCCTCTTTTTTC       GGGGGGGGGGGGGGGGGGGDGGGEGGGGGGGGGGGFA@'B############################################################    XC:
i:41    XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:41
HWUSI-EAS1568_104130695:1:52:12845:15152#0      147     1       27050083        60      100M    =       27049916        -267    TTACGAGCACTAACCTTCCGGGCAGGGGCGGTGTGGCTTCGAG
TTTATGCAGATTAGCCATGTGTTTCTCACTTCAAGGCGTTCCACAGGAAAAGCCTTT       @A?@@=?:9:=63;::,;8'CB:3@EFC?EEEEBEGEGBGFEDDBGFGFGGGGFGFGEGBGGGGFFGGGGEGBEGGGGGGGDFDGGGGGGGGGGGGGGG    XT:
A:U     NM:i:1  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:1  XO:i:0  XG:i:0  MD:Z:19T80
HWI-EAS231_104216201:1:97:16182:5636#0  99      1       27051115        60      100M    =       27051373        358     ATTCAGCCCAGGTCTACCTTAAAAGTCTAGTCTGTTATTTCAGCCGTTTTT
GGGGGTCATTTTGGTTTTCCTTGATGCAATAAGAAGTTGTAAGAGGTTC       GGGGGGGGG[illegible]GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGEEFGGEGEGGGGGEEGFGGFGGGEGEGBEFBEECDFCEDCDDEDEDCBE    XT:A:U  NM:
i:0     SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:100
HWUSI-EAS1568_104130695:2:36:6127:20020#0       163     1       27051196        60      73M27S  =       27051269        141     ATGAAGTTGTAAGAGGTTCCAGTAGCCAGCTAGTCTGGAGGGT
CTTGTAATTCTATTGGCTTGGAGCAGGGGAACTCCCTATGCAAGGGGGGCGTCTGCT       GGGAGGGGGGGGGGG:E5EEFFFDFGGGGFFGEEGGGFEFBFBEFEFA:DEEEFBGDEBA:A5?:>=,>:>@############################    XC:
i:73    XT:A:U  NM:i:1  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:1  XO:i:0  XG:i:0  MD:Z:1A71
HWUSI-EAS1568_104130695:2:36:6127:20020#0       83      1       27051269        60      32S68M  =       27051196        -141    GTCTTGTAATTCTATTGGCTTGGAGCAGGGGAACTCCCTATGC
AAGGGGTCTGTCTGCTAGCTGTTTGTTCTGAATATATTCGGAAGGACAAGCTGCTAG       ################################A?BAAA????76@@@:AEE?EEEEEEEEEFDGGGGEEGGGGGGGEFEEGGGGFGGGGFGGGFGDFGG    XC:
i:68    XT:A:U  NM:i:0  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:68
HWI-EAS231_104216201:1:97:16182:5636#0  147     1       27051373        60      100M    =       27051115        -358    TGATACATTTCATTTGGCAGGCAGTTTATAACATGGCCAGTGATCTGGAAG
GAGGTGGTAAGTGGTAGATGGAGATGAGGAGTTTTCTTATGGCTCTTTC       BE?EEEE?EECDD@CGEBEEEEEEEE@EEEEDDDEFG?EEGEGEEDFGGGGGGGGG?>@@@0.3>?GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG    XT:A:U  NM:
i:0     SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:100
HWUSI-EAS1[illegible]04130695:2:24:9014:10947#0       163     1       27051818        60      96M4S   =       27051902        184     GCCCCTTCTCTGTTACGAGAGATTGGAGATGGCATTTCAT
GTAACTGCAAATCACAAATATTTCACGATGGTTTTATACGTCATAAGTGTGAGTTAT       DGGGGDDGDEGGGGEDDDGEFFFFGBE=FDGGFFBEEGDFED?DDEBDFEEEBDEBABEB=B4?CDEA?EABB?;>;A,'0+-7:.<7099B=AA#####    X
i:96    XT:A:U  NM:i:2  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:2  XO:i:0  XG:i:0  MD:Z:79A1T14
```

Get to know NGS

Alignment format: SAM/BAM

SAM stands for Sequence Alignment/Map format.

SAM = text, BAM = binary

Read name | Flag | Reference Position | Quality | CIGAR | Mate Position

```
SRR013667.1 99 19 8882171 60 76M = 8882214 119
NCCAGCAGCCATAACTGGAATGGGAAATAAACACTATGTTCAAAGCAGAGAAAAT
AGGAGTGTGCAATAGACTTAT
#>A@BABAAAAADDEGCEFDHDEDBCFDBCDBCBDCEACB>AC@CDB@>>CB?>B
A:D?9>8AB685C26091:77
```

Bases

Base Qualities

Post-alignment manipulation

Samtools (http://samtools.sourceforge.net/)

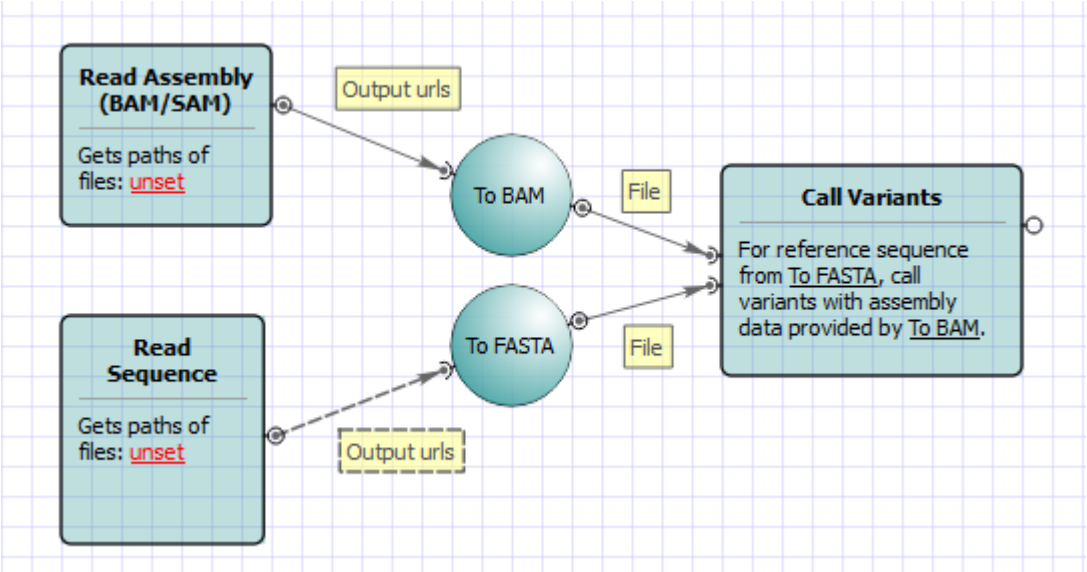
Is flexible

Is simple

Indexed by genomic position

Is compact in file size

Save memory



SNP Discovery: Goal

```
GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACAATGTC
GTTACTGTCGTTGTAATgCTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACAATGTC
GTTACTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTGGTAATACTCCACaATGTC
GTTACTGTCGTTGTAATACTCCACaATGTC
GTTAaTGTCGTTGTAATACTCCACGATGTC
GTTACTGTCGTTGTAcTACTCCACGATGTC
GTTACTGTCGTTGTAATACTCCACaATGTC
```

sequencing errors

SNP

SNP calling with samtools pipeline

Samtools (http://samtools.sourceforge.net/)

Is flexible

Is simple

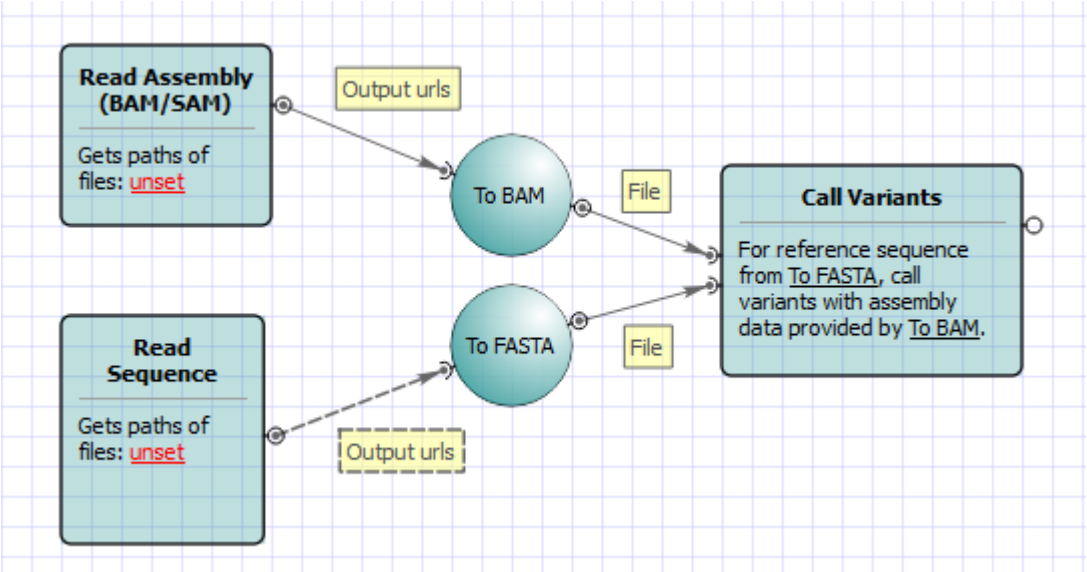
Indexed by genomic position

Is compact in file size

Save memory

mpipeup

bcftools



Variant Calling format: VCF

##fileformat=VCFv4.0

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003

samples | Combined depth | Allele frequency | In dbSNP? | In HapMap2?

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:.,.

Genotype | Genotype quality | Read depth | Haplotype qualities

SNP manipulation & filtering

Vcftools (vcftools.github.io)

Filter out specific variants
Compare files
Summarize variants
Convert to different file types
Validate and merge files
Create intersections and subsets of variants

Other popular tools

GATK: for selecting variants

R and Bioconductor: VariantFiltering

snpEff: SNP annotation

Hapmap (“Haplotype Map”)

International HapMap Consortium (2001)

1	2
rs#	alleles

- `rs#` cont
- `alleles`

[redacted]

3	4	5	6	7	8	9	10	11	12
rs#	alleles	chrom	pos	strand	assembly#	center	protLSID	assayLSID	panel	QCcode	33-16	38-11	42
PZB00859.1	A/C	1	157104	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	CC	CC
PZA01271.1	C/G	1	1947984	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	GG	CC
PZA03613.2	G/T	1	2914066	+	AGPv1	Panzea	NA	NA	maize282	NA	GG	GG	GG
PZA03613.1	A/T	1	2914171	+	AGPv1	Panzea	NA	NA	maize282	NA	TT	TT	TT
PZA03614.2	A/G	1	2915078	+	AGPv1	Panzea	NA	NA	maize282	NA	GG	GG	GG
PZA03614.1	A/T	1	2915242	+	AGPv1	Panzea	NA	NA	maize282	NA	TT	TT	TT
PZA00258.3	C/G	1	2973508	+	AGPv1	Panzea	NA	NA	maize282	NA	GG	CC	CC
PZA02962.13	A/T	1	3205252	+	AGPv1	Panzea	NA	NA	maize282	NA	TT	TT	TT
PZA02962.14	C/G	1	3205262	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	CC	CC
PZA00599.25	C/T	1	3206090	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	TT	CC
PZA02129.1	C/T	1	3706018	+	AGPv1	Panzea	NA	NA	maize282	NA	TT	CC	CC
PZA00393.1	C/T	1	4175293	+	AGPv1	Panzea	NA	NA	maize282	NA	TT	TT	TT
PZA02869.8	C/T	1	4429897	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	TT	CC
PZA02869.4	C/G	1	4429927	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	CC	CC
PZA02869.2	C/T	1	4430055	+	AGPv1	Panzea	NA	NA	maize282	NA	NN	TT	TT
PZA02032.1	A/T	1	4490461	+	AGPv1	Panzea	NA	NA	maize282	NA	AA	TT	AA
zagl1.5	A/T	1	4835434	+	AGPv1	Panzea	NA	NA	maize282	NA	AA	NN	AA
zagl1.2	A/C	1	4835558	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	CC	CC
zagl1.6	C/T	1	4835658	+	AGPv1	Panzea	NA	NA	maize282	NA	TT	TT	TT
PZD00081.2	C/T	1	4836542	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	CC	CC
zagl1.1	A/C	1	4912526	+	AGPv1	Panzea	NA	NA	maize282	NA	AA	AA	AA
PZB00919.1	A/C	1	5353319	+	AGPv1	Panzea	NA	NA	maize282	NA	CC	CC	CC
PZB00919.2	G/T	1	5353655	+	AGPv1	Panzea	NA	NA	maize282	NA	GG	GG	GG

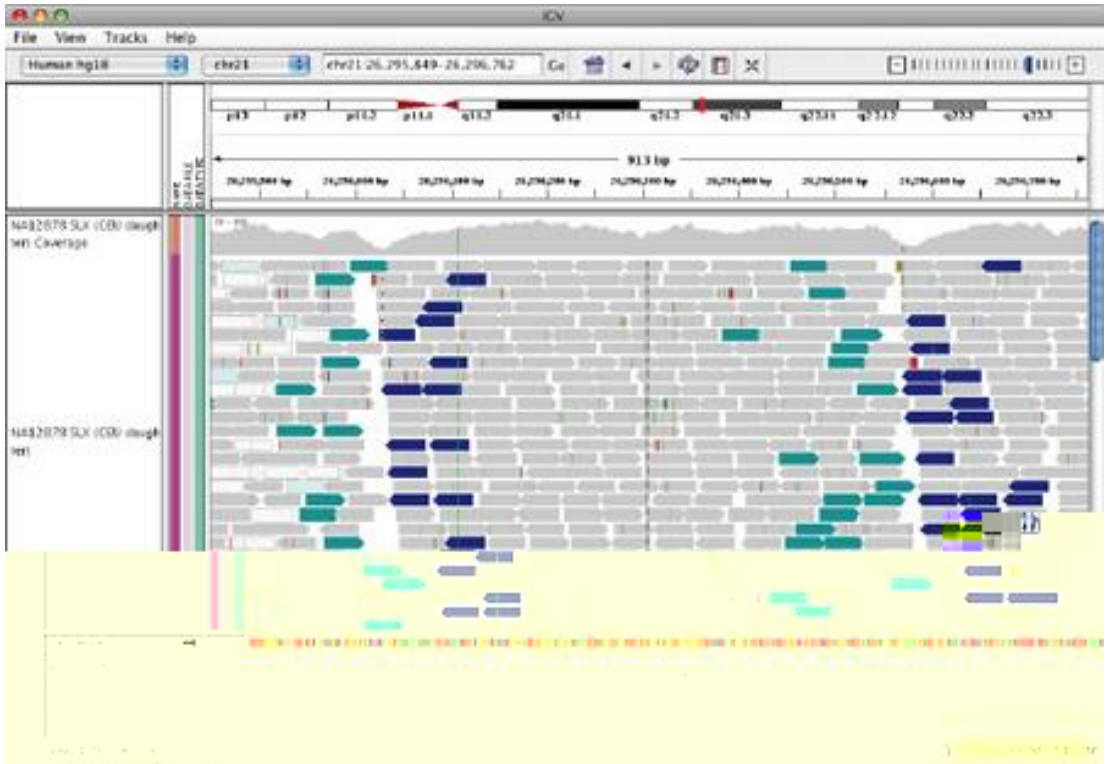
Post-alignment manipulation

Viewing the alignment

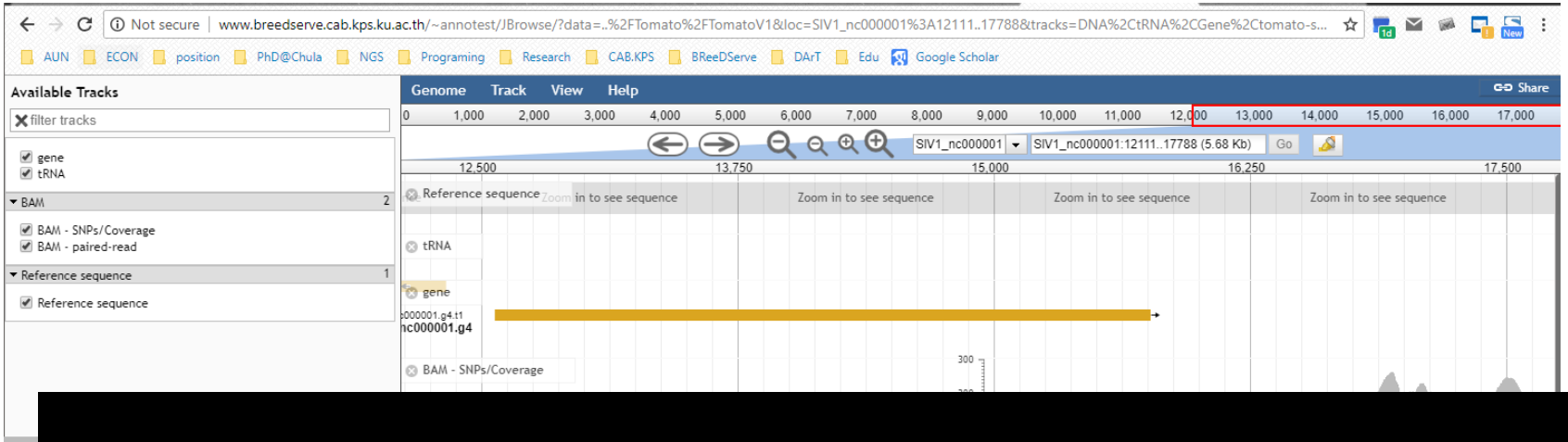
Integrative Genomics Viewer

SAM to BAM

samtools tview



jBrowse



[gwasviewer]

Galaxy https://usegalaxy.org/

Galaxy is an open source

web-based platform for data intensive biological research

NGS data analysis

Offline environmental

Tool Shed – flexible for installing tools to Galaxy

Thank you for your attention