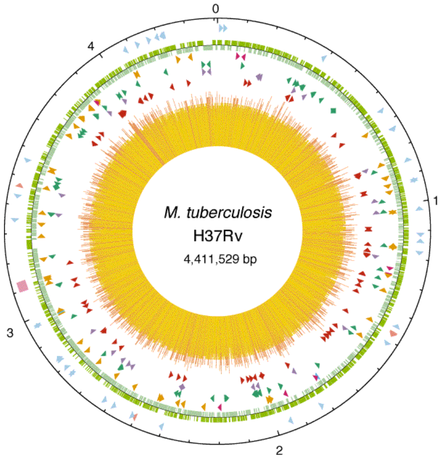
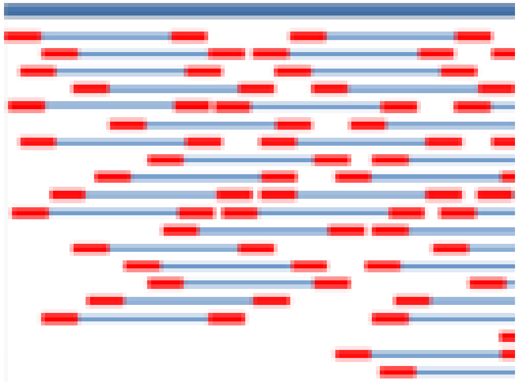
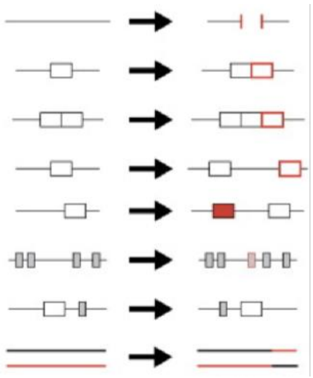
High-throughput sequencing analysis

362 732 Bioinformatics course, MD KKU







Kiatichai Faksri

(Ph.D., Medical Microbiology)

Faculty of Medicine, Khon Kaen University, Thailand

Objectives

1. Know and understand basics in NGS analysis
2. Understand the analysis pipeline of NGS analysis
3. Understand and be able to use the tools for NGS analysis
4. Be able to analyze the NGS from example organism (*M. tuberculosis*) and compare between the two genomes

Outline

- 1. Introduction of High throughput sequencing (HTS) analysis
- 2. Basic terminology
- 3. HTS platforms
- 4. Analysis pipeline
- 5. Technical information

- 6. Practice in HTS analysis (bacteria I)
- 7. Practice in HTS analysis (bacteria II)
- 8. Assignment

1

Introduction in NGS analysis

Genome Sequence Analysis

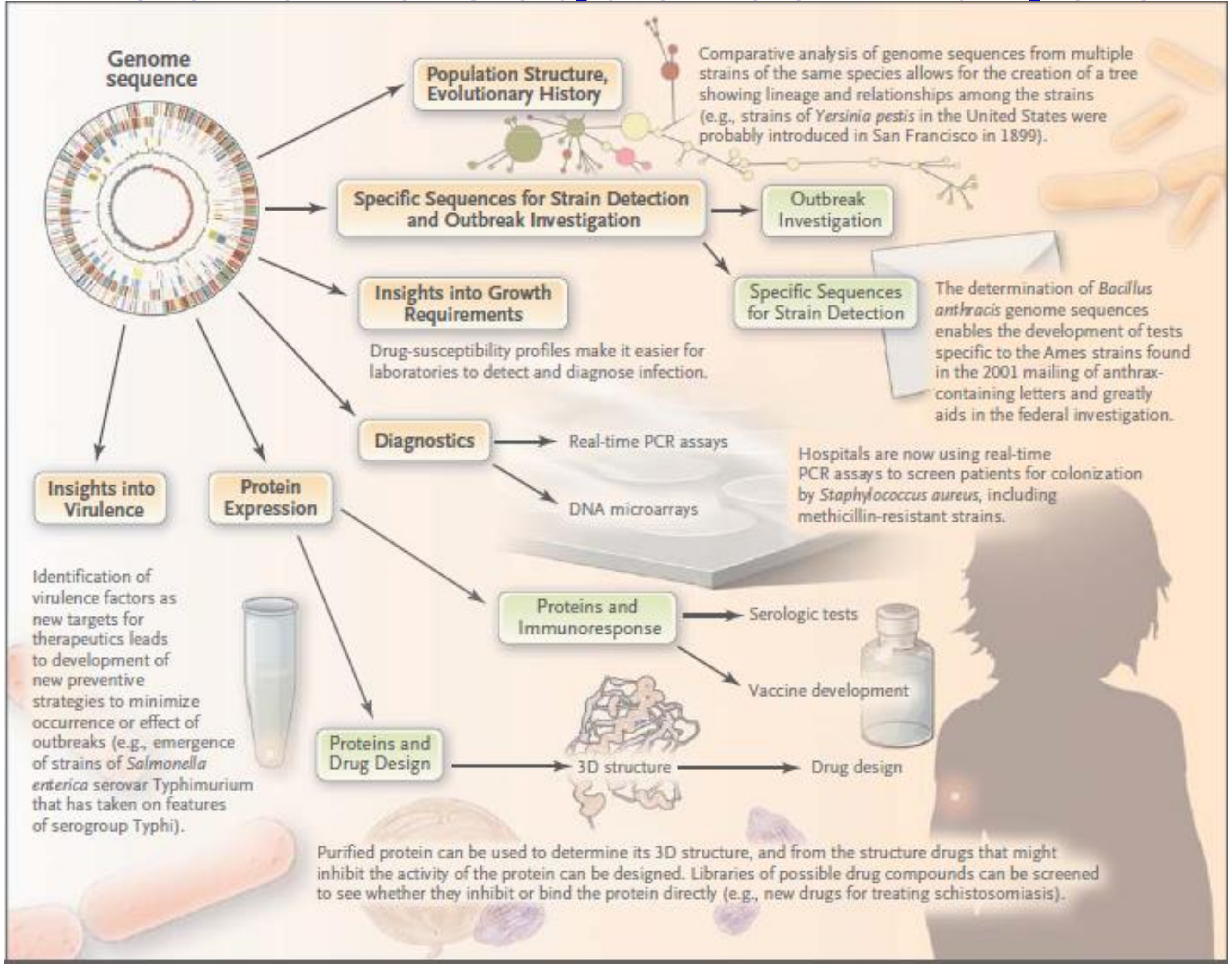

Genome sequence
Population Structure, Evolutionary History
Comparative analysis of genome sequences from multiple strains of the same species allows for the creation of a tree showing lineage and relationships among the strains (e.g., strains of Yersinia pestis in the United States were probably introduced in San Francisco in 1899).
Specific Sequences for Strain Detection and Outbreak Investigation
Outbreak Investigation
Specific Sequences for Strain Detection
The determination of Bacillus anthracis genome sequences enables the development of tests specific to the Ames strains found in the 2001 mailing of anthrax-containing letters and greatly aids in the federal investigation.
Insights into Growth Requirements
Drug-susceptibility profiles make it easier for laboratories to detect and diagnose infection.
Diagnostics
Real-time PCR assays
DNA microarrays
Hospitals are now using real-time PCR assays to screen patients for colonization by Staphylococcus aureus, including methicillin-resistant strains.
Insights into Virulence
Protein Expression
Identification of virulence factors as new targets for therapeutics leads to development of new preventive strategies to minimize occurrence or effect of outbreaks (e.g., emergence of strains of Salmonella enterica serovar Typhimurium that has taken on features of serogroup Typhi).
Proteins and Immunoresponse
Serologic tests
Vaccine development
Proteins and Drug Design
3D structure
Drug design
Purified protein can be used to determine its 3D structure, and from the structure drugs that might inhibit the activity of the protein can be designed. Libraries of possible drug compounds can be screened to see whether they inhibit or bind the protein directly (e.g., new drugs for treating schistosomiasis).

Figure 3. Microbial Genomics and Tool Development.

A genome sequence facilitates the development of a variety of tools and approaches for understanding, manipulating, and mitigating the overall effect of a microbe. The sequence provides insight into the population structure and evolutionary history of a microbe for epidemiologic investigation, information with which to develop new diagnostic tests and cultivation methods, new targets of drug development, and antigens for vaccine development.

N Engl J Med. 2011 Jul 28;365(4):347-57. doi: 10.1056/NEJMra1003071

Analysis purposes

- Detection of novel mutation associated with drug resistance
- Outbreak investigation and transmission analysis
- Differentiation between reinfection vs relapse
- Detection of variants associated with phenotypes
- Phylogentic analysis
- Etc.

Strategies for shotgun sequencing

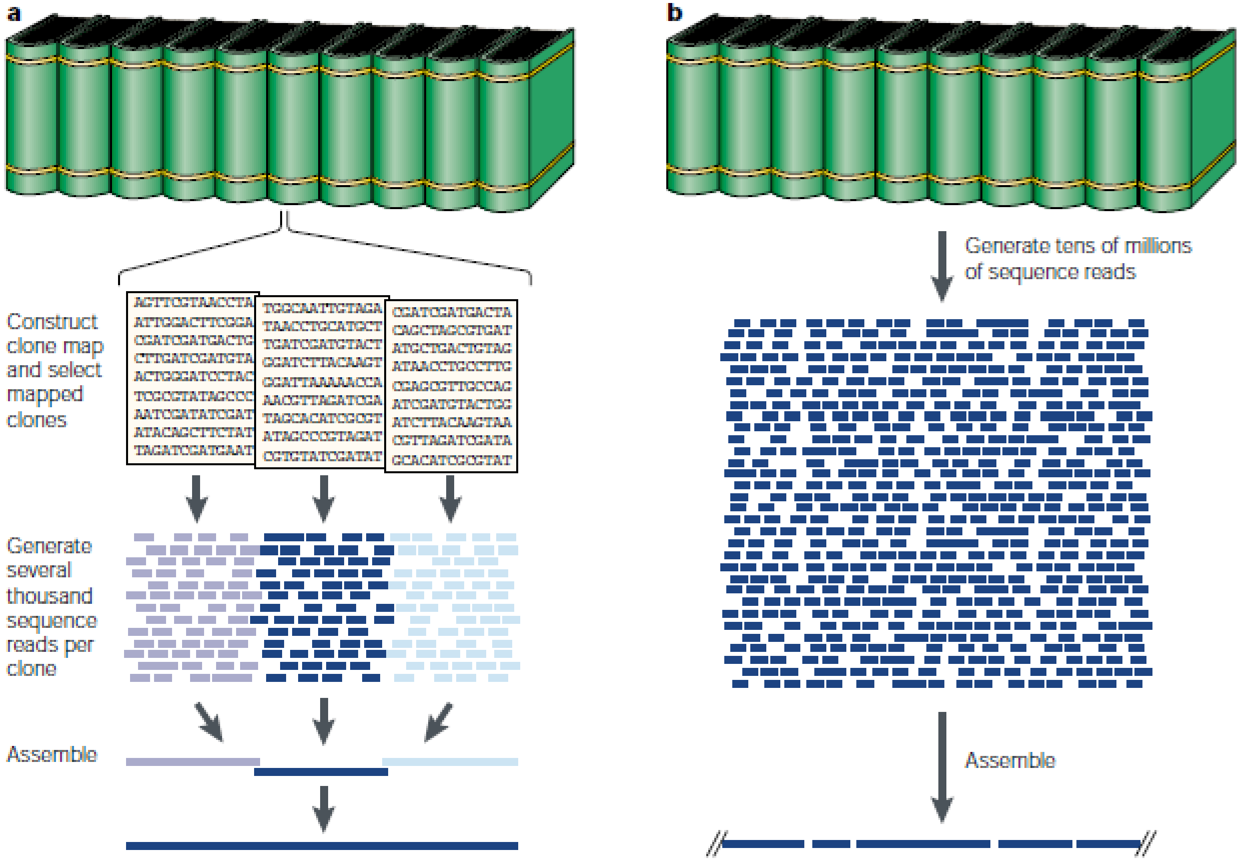
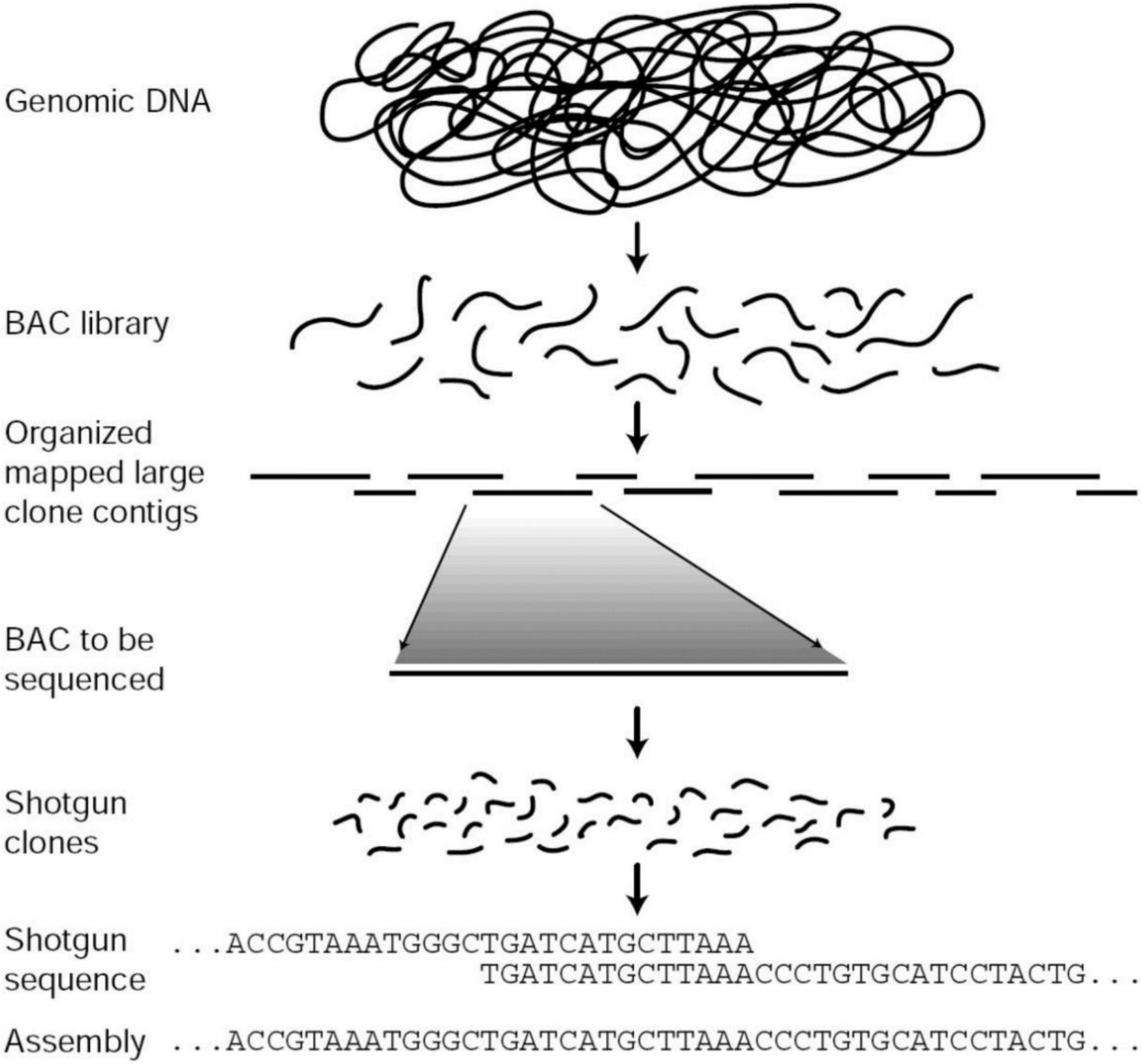


Figure 1 | **Two main shotgun-sequencing strategies. a** | Schematic overview of clone-by-clone shotgun sequencing. A representation of a genome is made by analogy to an encyclopaedia set, with each volume corresponding to an individual chromosome. The construction of clone-based physical maps produces overlapping series of clones (that is, contigs), each of which spans a large, contiguous region of the source genome. Each clone (for example, a bacterial artificial chromosome (BAC)) can be thought of as containing the DNA represented by one page of a volume. For shotgun sequencing, individual mapped clones are subcloned into smaller-insert libraries, from which sequence reads are randomly derived. In the case of BACs, this typically requires the generation of several thousand sequence reads per clone. The resulting sequence data set is then used to assemble the complete sequence of that clone (see FIGS 3,4). **b** | Schematic overview of whole-genome shotgun sequencing. In this case, the mapping phase is skipped and shotgun sequencing proceeds using subclone libraries prepared from the entire genome. Typically, tens of millions of sequence reads are generated and these in turn are subjected to computer-based assembly to generate contiguous sequences of various sizes.

Nat Rev Genet. 2001 Aug;2(8):573-83

Classical shotgun sequencing



Genomic DNA

BAC library

Organized mapped large clone contigs

BAC to be sequenced

Shotgun clones

Shotgun sequence

```
...ACCGTAAATGGGCTGATCATGCTTAAA
                    TGATCATGCTTAAACCCTGTGCATCCTACTG...
```

Assembly

```
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...
```

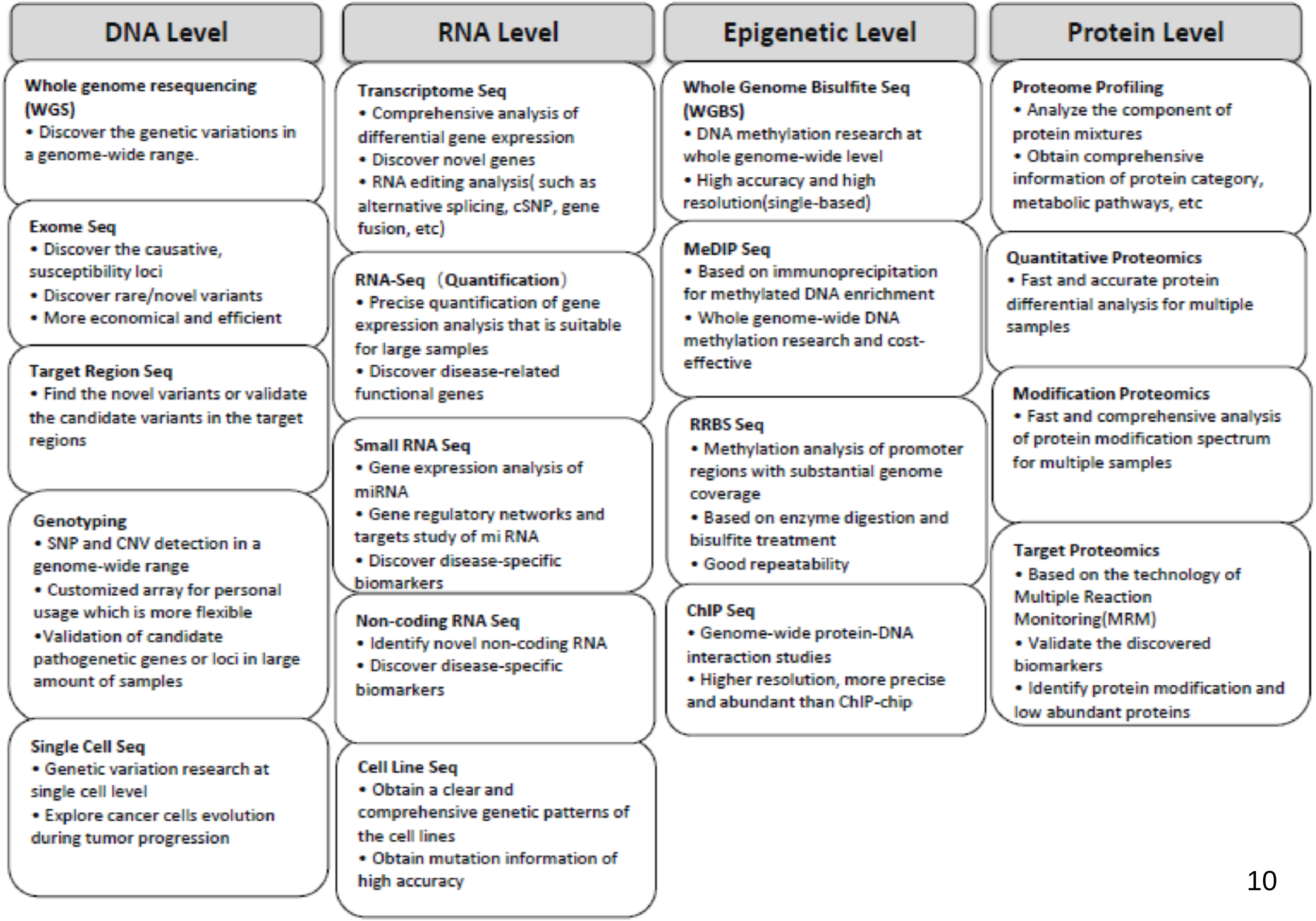
→ Fill the gaps

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 409, 863 (2001).

Sample types

- Cell sample types
  - Single cell
  - Population
  - Communities (Metagenomics)
- Nucleic acid sample type
  - DNA = genomics, epigenetics
  - RNA = transcriptomic (and metatranscriptomics)
- Genome Sizes
  - Whole Genome
  - Targeted, e.g. exome sequencing

Omic analysis



DNA Level

Whole genome resequencing (WGS)
- Discover the genetic variations in a genome-wide range.

Exome Seq
- Discover the causative, susceptibility loci
- Discover rare/novel variants
- More economical and efficient

Target Region Seq
- Find the novel variants or validate the candidate variants in the target regions

Genotyping
- SNP and CNV detection in a genome-wide range
- Customized array for personal usage which is more flexible
- Validation of candidate pathogenetic genes or loci in large amount of samples

Single Cell Seq
- Genetic variation research at single cell level
- Explore cancer cells evolution during tumor progression

RNA Level

Transcriptome Seq
- Comprehensive analysis of differential gene expression
- Discover novel genes
- RNA editing analysis(such as alternative splicing, cSNP, gene fusion, etc)

RNA-Seq （Quantification）
- Precise quantification of gene expression analysis that is suitable for large samples
- Discover disease-related functional genes

Small RNA Seq
- Gene expression analysis of miRNA
- Gene regulatory networks and targets study of mi RNA
- Discover disease-specific biomarkers

Non-coding RNA Seq
- Identify novel non-coding RNA
- Discover disease-specific biomarkers

Cell Line Seq
- Obtain a clear and comprehensive genetic patterns of the cell lines
- Obtain mutation information of high accuracy

Epigenetic Level

Whole Genome Bisulfite Seq (WGBS)
- DNA methylation research at whole genome-wide level
- High accuracy and high resolution(single-based)

MeDIP Seq
- Based on immunoprecipitation for methylated DNA enrichment
- Whole genome-wide DNA methylation research and cost-effective

RRBS Seq
- Methylation analysis of promoter regions with substantial genome coverage
- Based on enzyme digestion and bisulfite treatment
- Good repeatability

ChIP Seq
- Genome-wide protein-DNA interaction studies
- Higher resolution, more precise and abundant than ChIP-chip

Protein Level

Proteome Profiling
- Analyze the component of protein mixtures
- Obtain comprehensive information of protein category, metabolic pathways, etc

Quantitative Proteomics
- Fast and accurate protein differential analysis for multiple samples

Modification Proteomics
- Fast and comprehensive analysis of protein modification spectrum for multiple samples

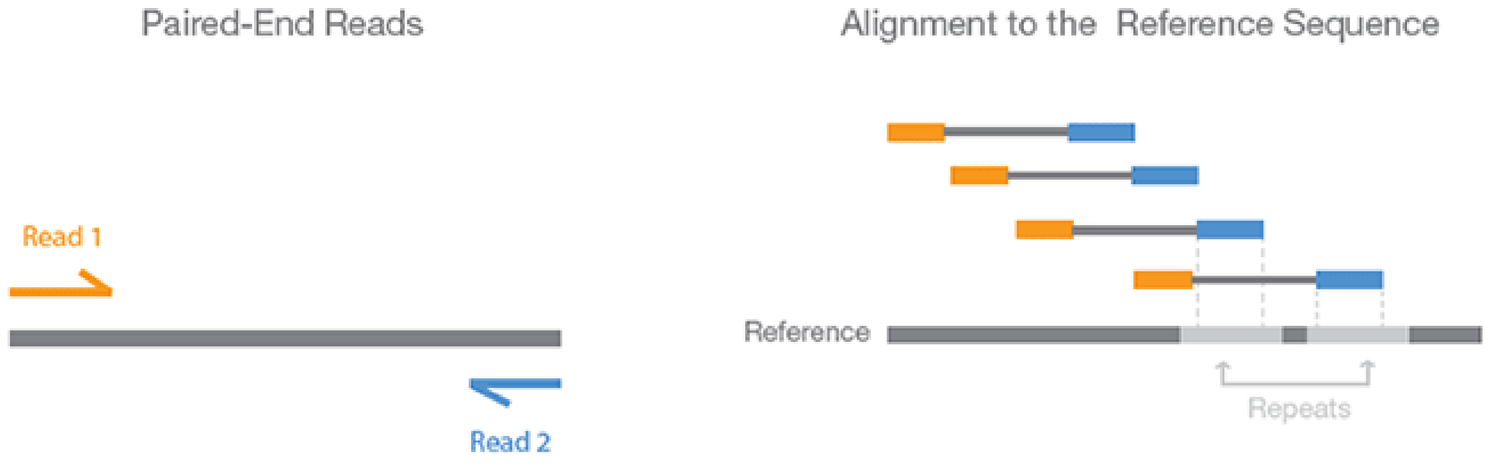
Target Proteomics
- Based on the technology of Multiple Reaction Monitoring(MRM)
- Validate the discovered biomarkers
- Identify protein modification and low abundant proteins

2

Basic terminology

Paired-end read

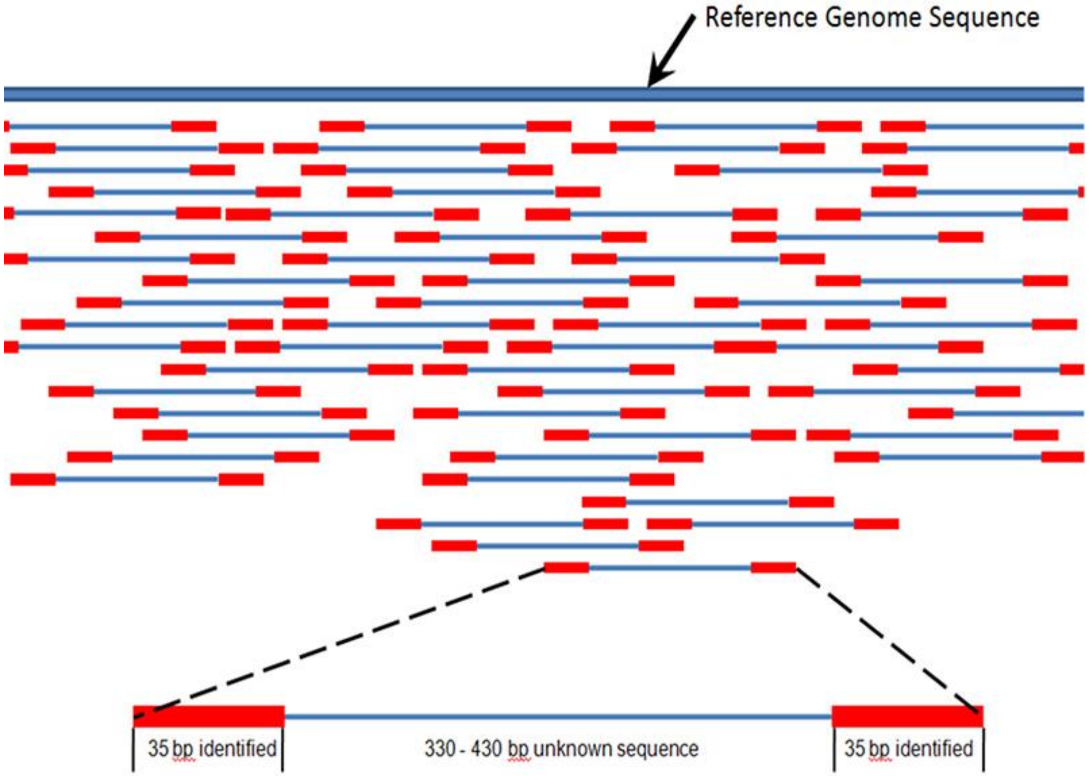
Figure 4. Paired-End Sequencing and Alignment



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

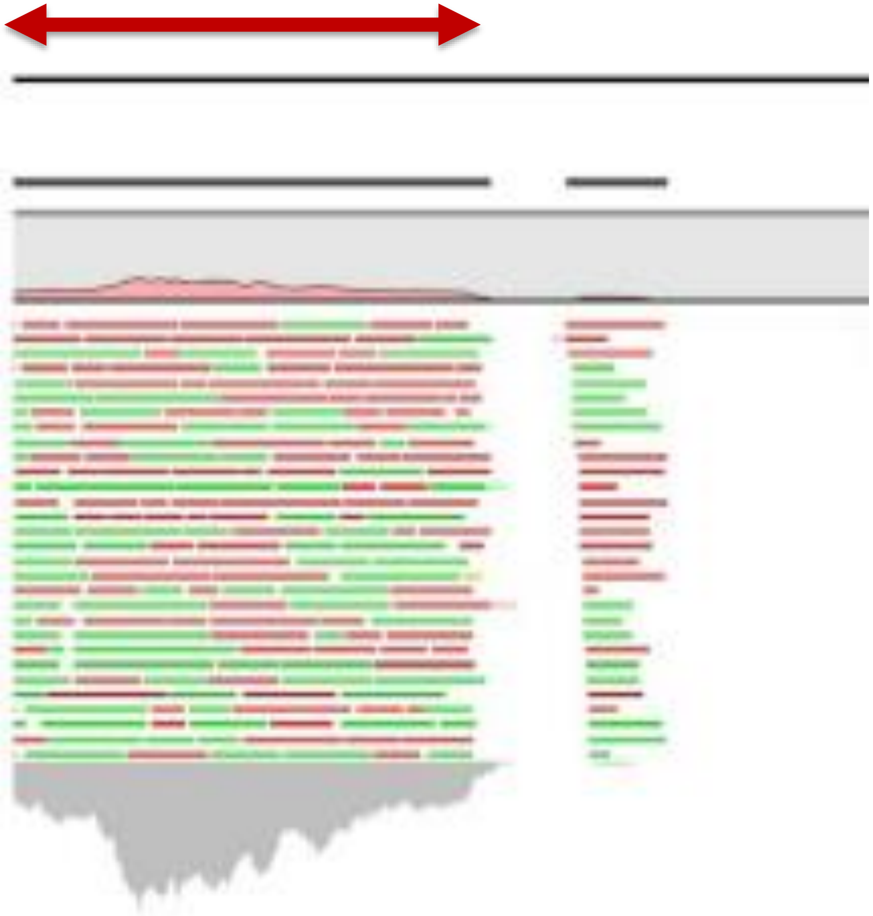
http://www.illumina.com/technology/next-generation-sequencing/paired-end-sequencing_assay.html

Read length

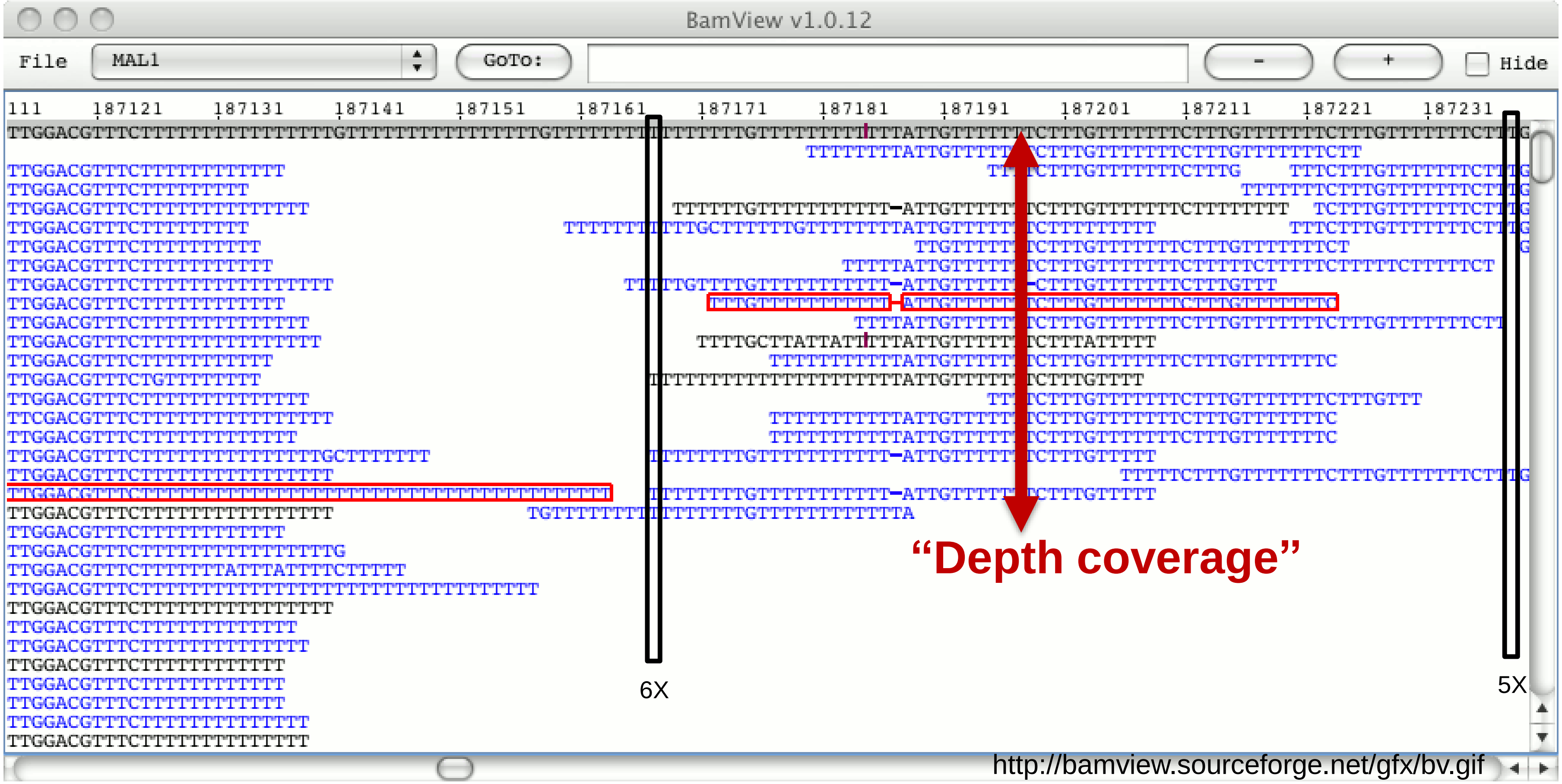


http://img.medscape.com/news/2014/wc_140606_dna_sequencing_mapping_reads_800x600.jpg

Region coverage



Read depth (coverage)

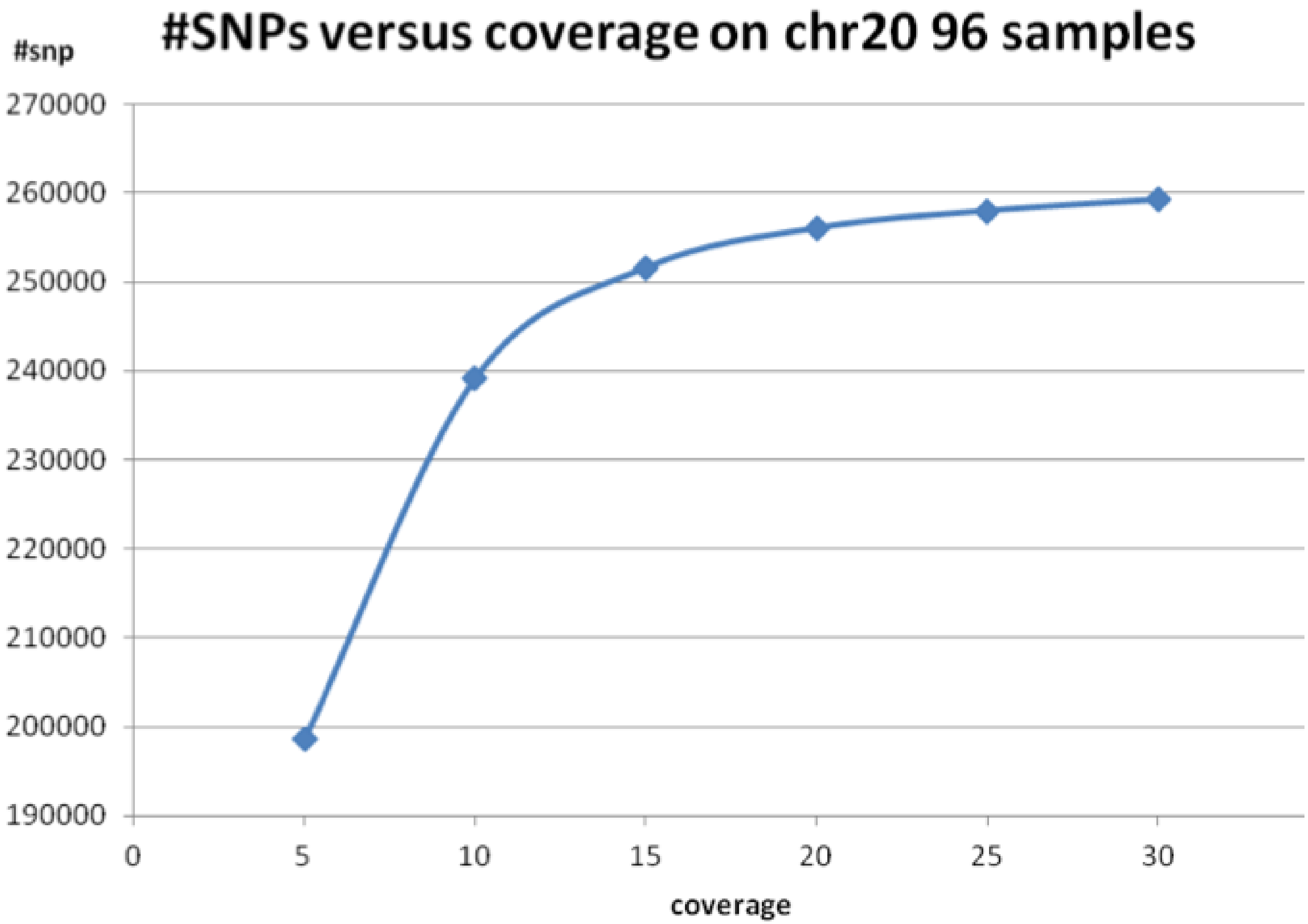


```
111      187121    187131    187141    187151    187161    187171    187181    187191    187201    187211    187221    187231
TTGGACGTTTCTTTTTTTTTTTTTTTTGTTTTTTTTTTTTTTTTGTTTTTTTTTTTTTTTTGTTTTTTTTTTTTATTGTTTTTTTCTTTGTTTTTTTCTTTGTTTTTTTCTTTGTTTTTTTCTTTG
                                                                  TTTTTTTTATTGTTTTTTTCTTTGTTTTTTTCTTTGTTTTTTTCTT
TTGGACGTTTCTTTTTTTTTTTT                                                          TTTTCTTTGTTTTTTTCTTTG    TTTCTTTGTTTTTTTCTTTG
TTGGACGTTTCTTTTTTTTT                                                                                  TTTTTTTCTTTGTTTTTTTCTTTG
TTGGACGTTTCTTTTTTTTTTTTTT                              TTTTTTGTTTTTTTTTTT-ATTGTTTTTTTCTTTGTTTTTTTCTTTTTTTT  TCTTTGTTTTTTTCTTTG
TTGGACGTTTCTTTTTTTTT                          TTTTTTTTTTTGCTTTTTTGTTTTTTTTATTGTTTTTTTCTTTTTTTTT           TTTCTTTGTTTTTTTCTTTG
TTGGACGTTTCTTTTTTTTTT                                                      TTGTTTTTTTCTTTGTTTTTTTCTTTGTTTTTTTCT              G
TTGGACGTTTCTTTTTTTTTTT                                               TTTTTATTGTTTTTTTCTTTGTTTTTTTCTTTTTCTTTTTCTTTTTCTTTTTCT
TTGGACGTTTCTTTTTTTTTTTTTTTT                        TTTTTGTTTTGTTTTTTTTTTT-ATTGTTTTTT-CTTTGTTTTTTTCTTTGTTT
TTGGACGTTTCTTTTTTTTTTTT                                   TTTGTTTTTTTTTTT-ATTGTTTTTTTCTTTGTTTTTTTCTTTGTTTTTTTC
TTGGACGTTTCTTTTTTTTTTTTTT                                             TTTTATTGTTTTTTTCTTTGTTTTTTTCTTTGTTTTTTTCTTTGTTTTTTTCTT
TTGGACGTTTCTTTTTTTTTTTTTTT                               TTTTGCTTATTATTTTTATTGTTTTTTTCTTTATTTTT
TTGGACGTTTCTTTTTTTTTTT                                         TTTTTTTTTTTATTGTTTTTTTCTTTGTTTTTTTCTTTGTTTTTTTC
TTGGACGTTTCTGTTTTTTTT                                TTTTTTTTTTTTTTTTTTTTTATTGTTTTTTTCTTTGTTTT
TTGGACGTTTCTTTTTTTTTTTTTT                                                        TTTTCTTTGTTTTTTTCTTTGTTTTTTTCTTTGTTT
TTCGACGTTTCTTTTTTTTTTTTTTTT                                    TTTTTTTTTTTATTGTTTTTTTCTTTGTTTTTTTCTTTGTTTTTTTC
TTGGACGTTTCTTTTTTTTTTTTT                                       TTTTTTTTTTTATTGTTTTTTTCTTTGTTTTTTTCTTTGTTTTTTTC
TTGGACGTTTCTTTTTTTTTTTTTTTGCTTTTTTT                  TTTTTTTTGTTTTTTTTTTT-ATTGTTTTTTTCTTTGTTTTT
TTGGACGTTTCTTTTTTTTTTTTTTTT                                                                 TTTTTCTTTGTTTTTTTCTTTGTTTTTTTCTTTG
TTGGACGTTTCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT   TTTTTTTTGTTTTTTTTTTT-ATTGTTTTTTTCTTTGTTTTT
TTGGACGTTTCTTTTTTTTTTTTTTTT                TGTTTTTTTTTTTTTTTTGTTTTTTTTTTTTA
TTGGACGTTTCTTTTTTTTTTTT
TTGGACGTTTCTTTTTTTTTTTTTTTTG
TTGGACGTTTCTTTTTTTATTTATTTTCTTTTT
TTGGACGTTTCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT
TTGGACGTTTCTTTTTTTTTTTTTTTT
TTGGACGTTTCTTTTTTTTTTTTT
TTGGACGTTTCTTTTTTTTTTTTTT
TTGGACGTTTCTTTTTTTTTTTT
TTGGACGTTTCTTTTTTTTTTTT
TTGGACGTTTCTTTTTTTTTTTT
TTGGACGTTTCTTTTTTTTTTTTTT
TTGGACGTTTCTTTTTTTTTTTTTT
```

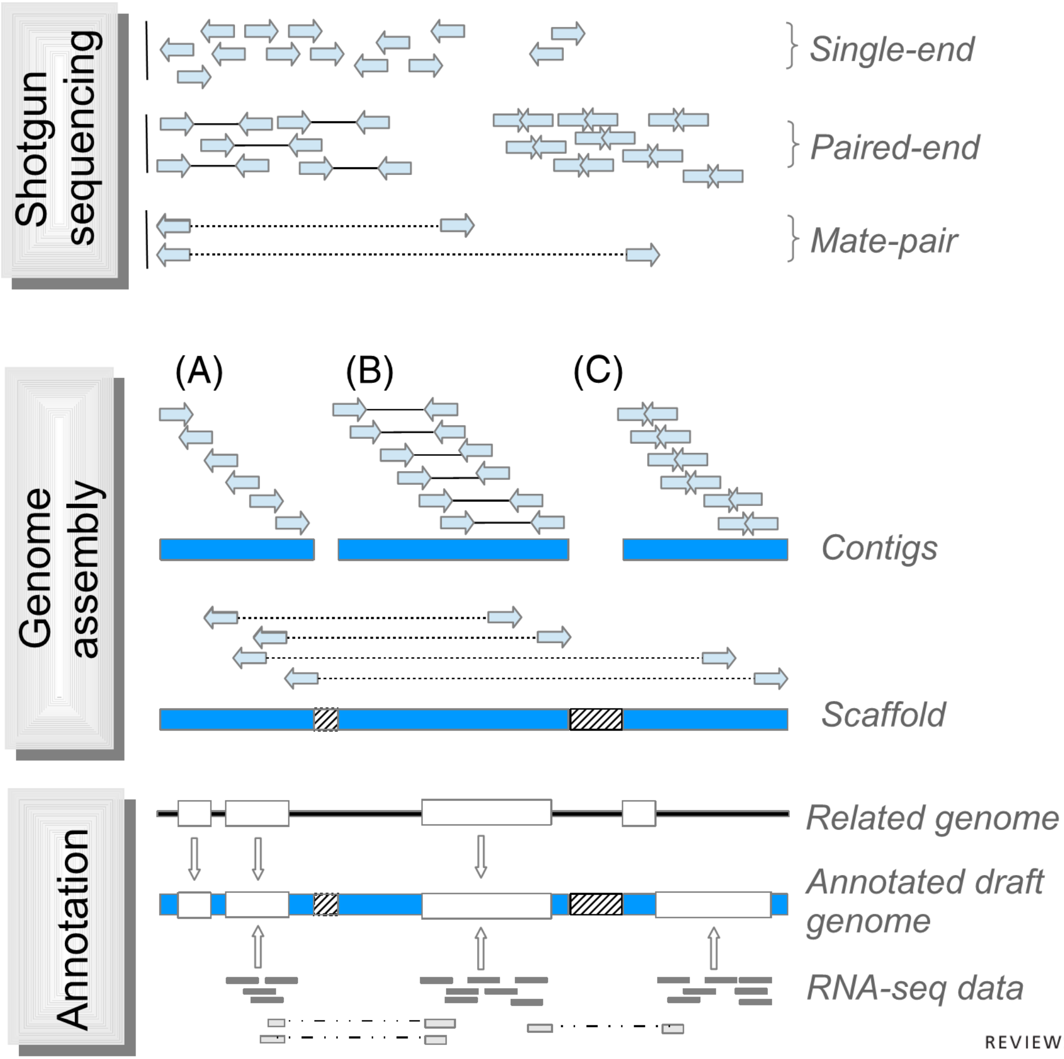
http://bamview.sourceforge.net/gfx/bv.gif

Number of bases /Genome length = estimated sequencing depth, e.g. 50X

Significance of sequencing coverage



Simplified illustration of the assembly process



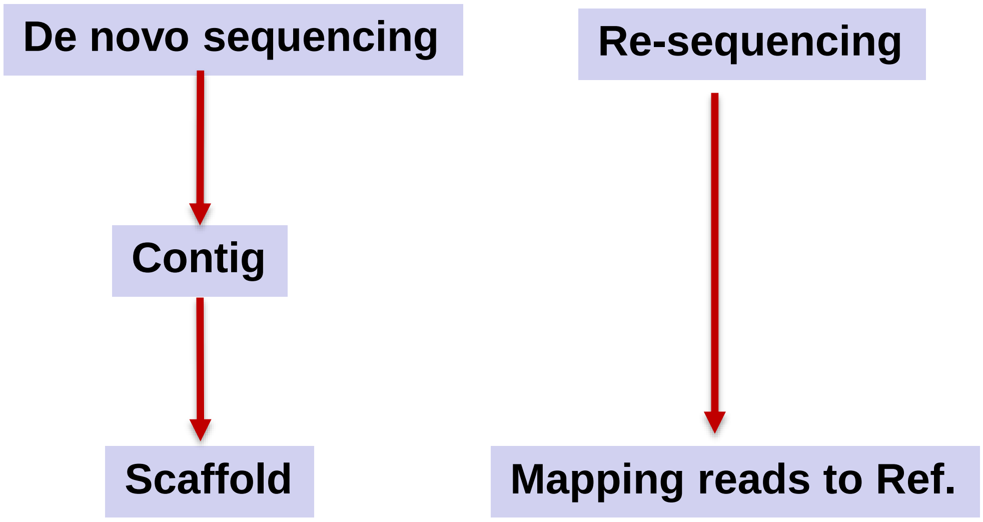
REVIEWS AND SYNTHESIS

A field guide to whole-genome sequencing, assembly and annotation

Robert Ekblom and Jochen B. W. Wolf

Sequencing assembly strategies

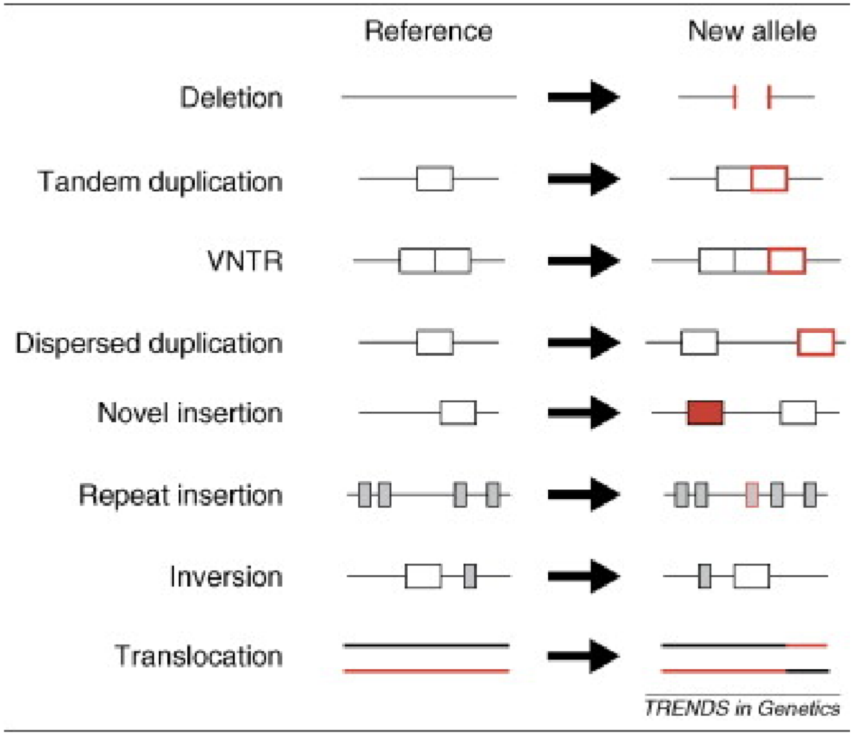
1. Re-sequencing
2. De novo sequencing



Variants

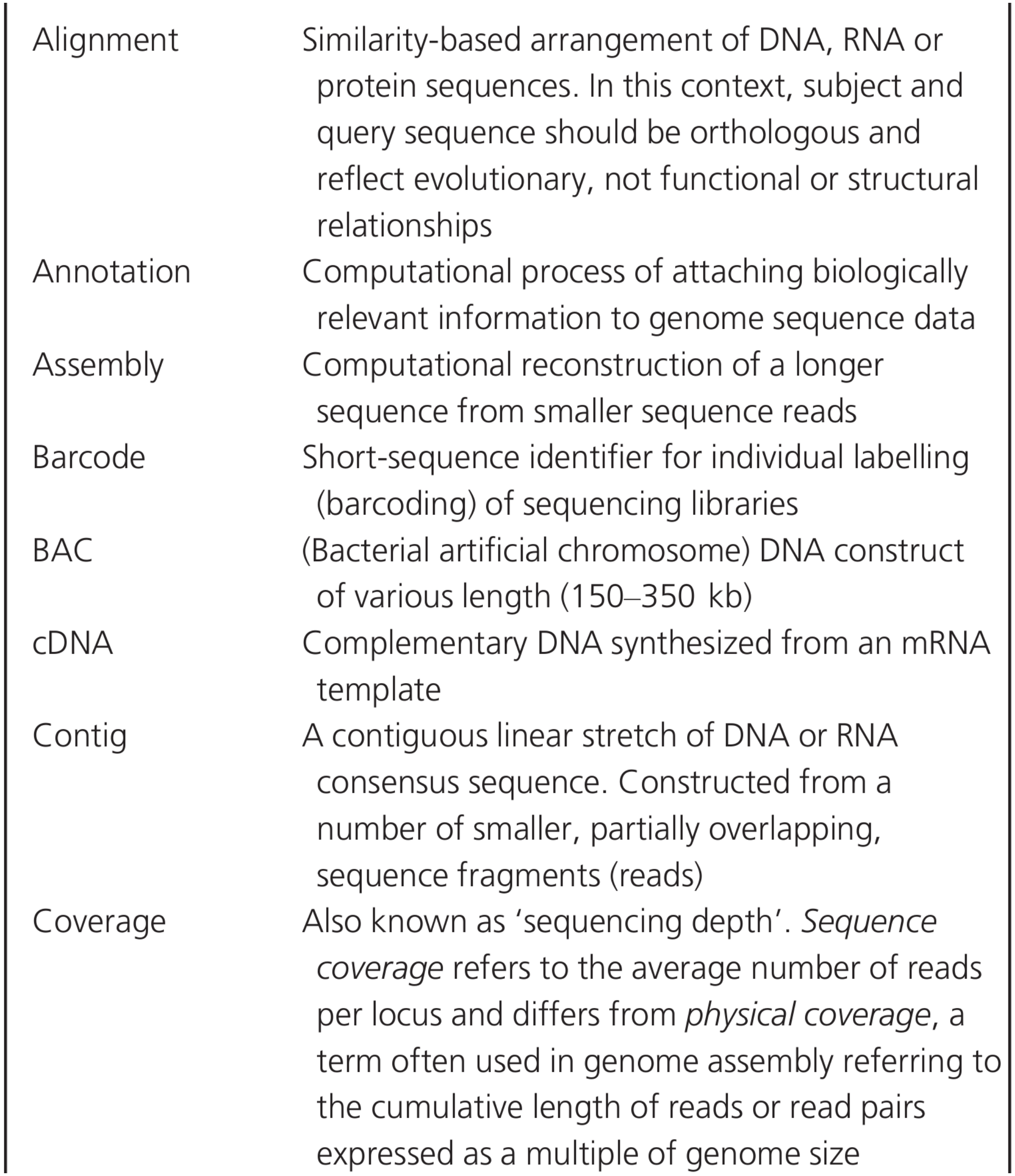
- SNPs
- Indels (small Indels)
- Copy number variants
- Structural variants (large Indels/ inversion/ translocation etc.)

- Problematic region: repetitive region and PE/PEE genes,



Hurles et, Trends in Genetics, 2007

SDL

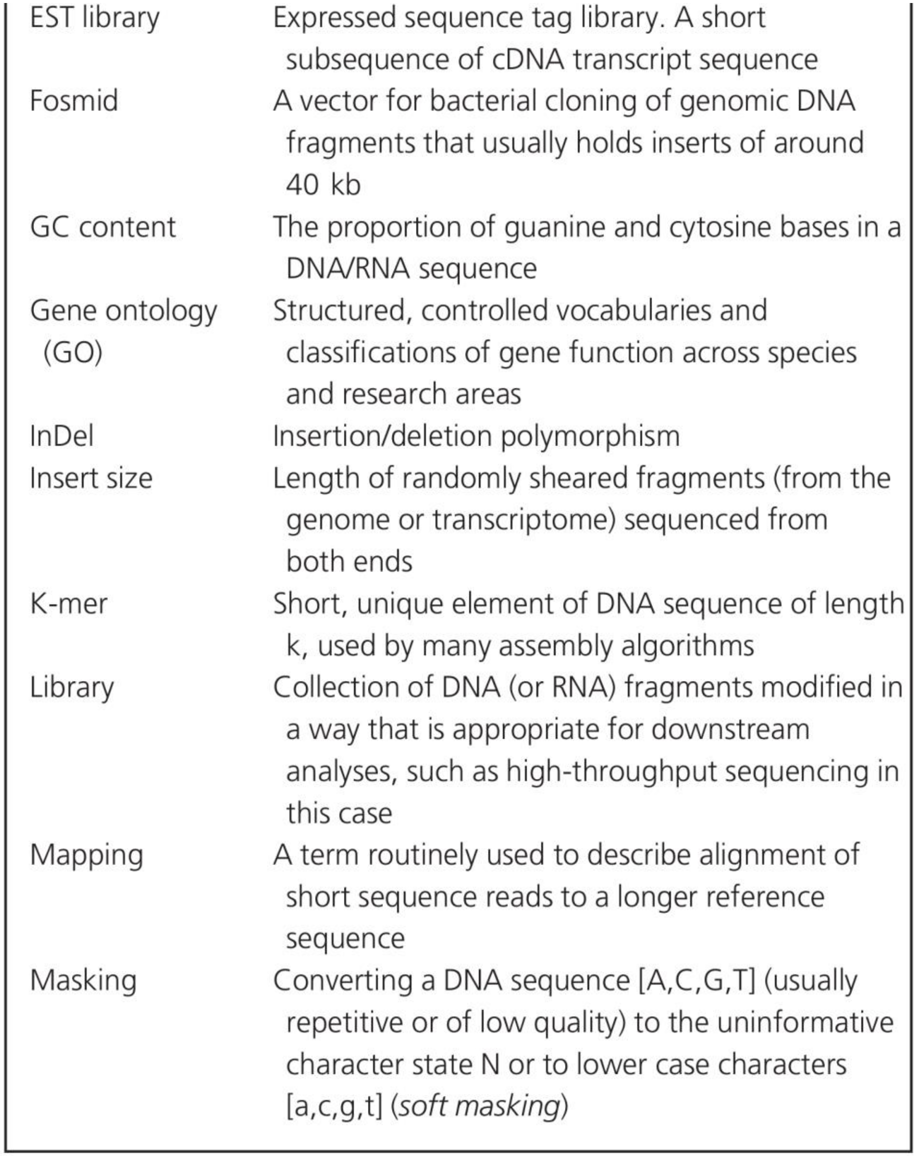


Alignment	Similarity-based arrangement of DNA, RNA or protein sequences. In this context, subject and query sequence should be orthologous and reflect evolutionary, not functional or structural relationships
Annotation	Computational process of attaching biologically relevant information to genome sequence data
Assembly	Computational reconstruction of a longer sequence from smaller sequence reads
Barcode	Short-sequence identifier for individual labelling (barcoding) of sequencing libraries
BAC	(Bacterial artificial chromosome) DNA construct of various length (150–350 kb)
cDNA	Complementary DNA synthesized from an mRNA template
Contig	A contiguous linear stretch of DNA or RNA consensus sequence. Constructed from a number of smaller, partially overlapping, sequence fragments (reads)
Coverage	Also known as ‘sequencing depth’. *Sequence coverage* refers to the average number of reads per locus and differs from *physical coverage*, a term often used in genome assembly referring to the cumulative length of reads or read pairs expressed as a multiple of genome size

Robert Ekblom and Jochen B. W. Wolf. A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications 2014. doi:10.1111/eva.12178

SDL

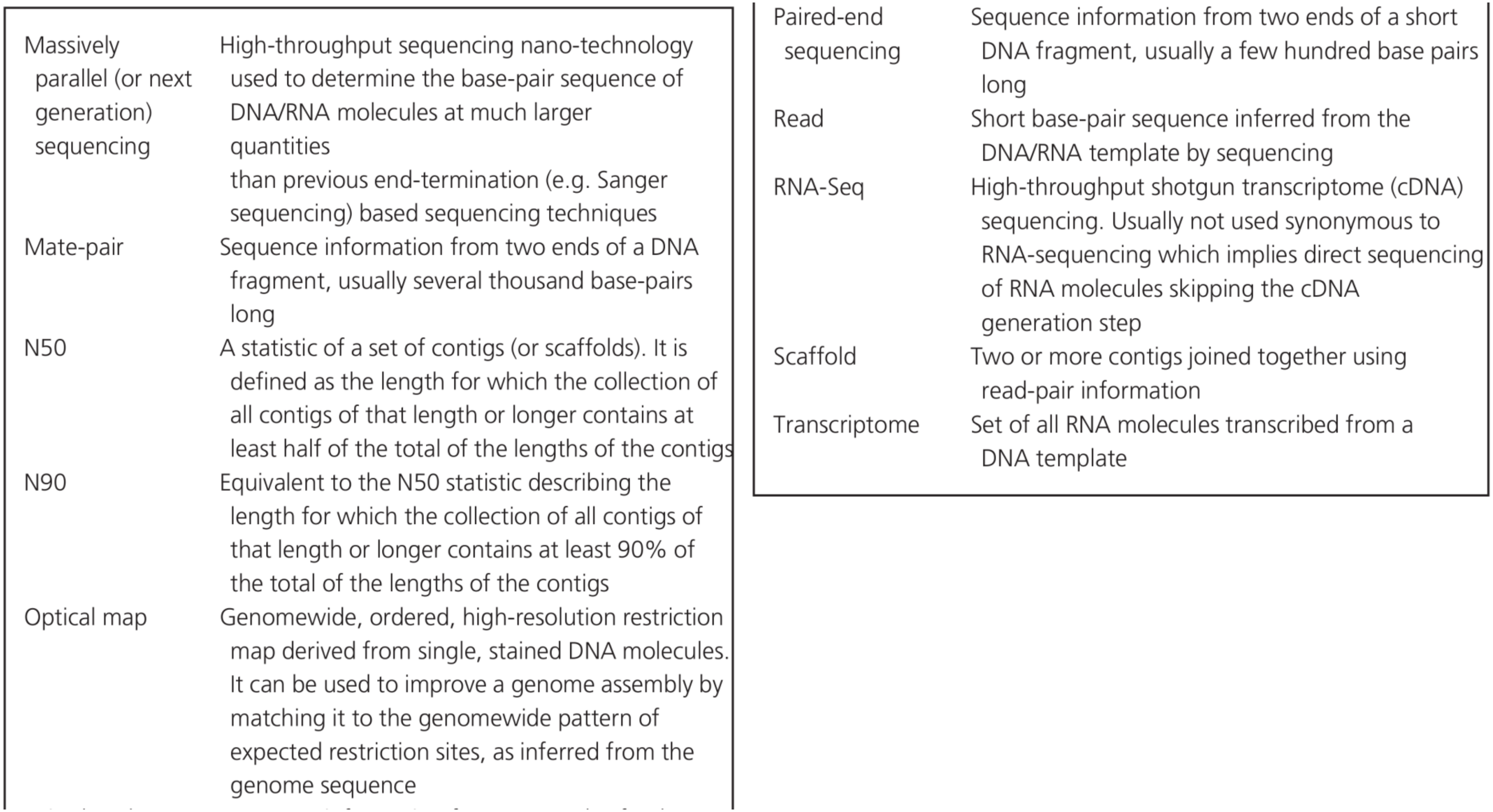
Faculty of Medicine, KKU



EST library	Expressed sequence tag library. A short subsequence of cDNA transcript sequence
Fosmid	A vector for bacterial cloning of genomic DNA fragments that usually holds inserts of around 40 kb
GC content	The proportion of guanine and cytosine bases in a DNA/RNA sequence
Gene ontology (GO)	Structured, controlled vocabularies and classifications of gene function across species and research areas
InDel	Insertion/deletion polymorphism
Insert size	Length of randomly sheared fragments (from the genome or transcriptome) sequenced from both ends
K-mer	Short, unique element of DNA sequence of length k, used by many assembly algorithms
Library	Collection of DNA (or RNA) fragments modified in a way that is appropriate for downstream analyses, such as high-throughput sequencing in this case
Mapping	A term routinely used to describe alignment of short sequence reads to a longer reference sequence
Masking	Converting a DNA sequence [A,C,G,T] (usually repetitive or of low quality) to the uninformative character state N or to lower case characters [a,c,g,t] (*soft masking*)

Robert Ekblom and Jochen B. W. Wolf. A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications 2014. doi:10.1111/eva.12178

SDL



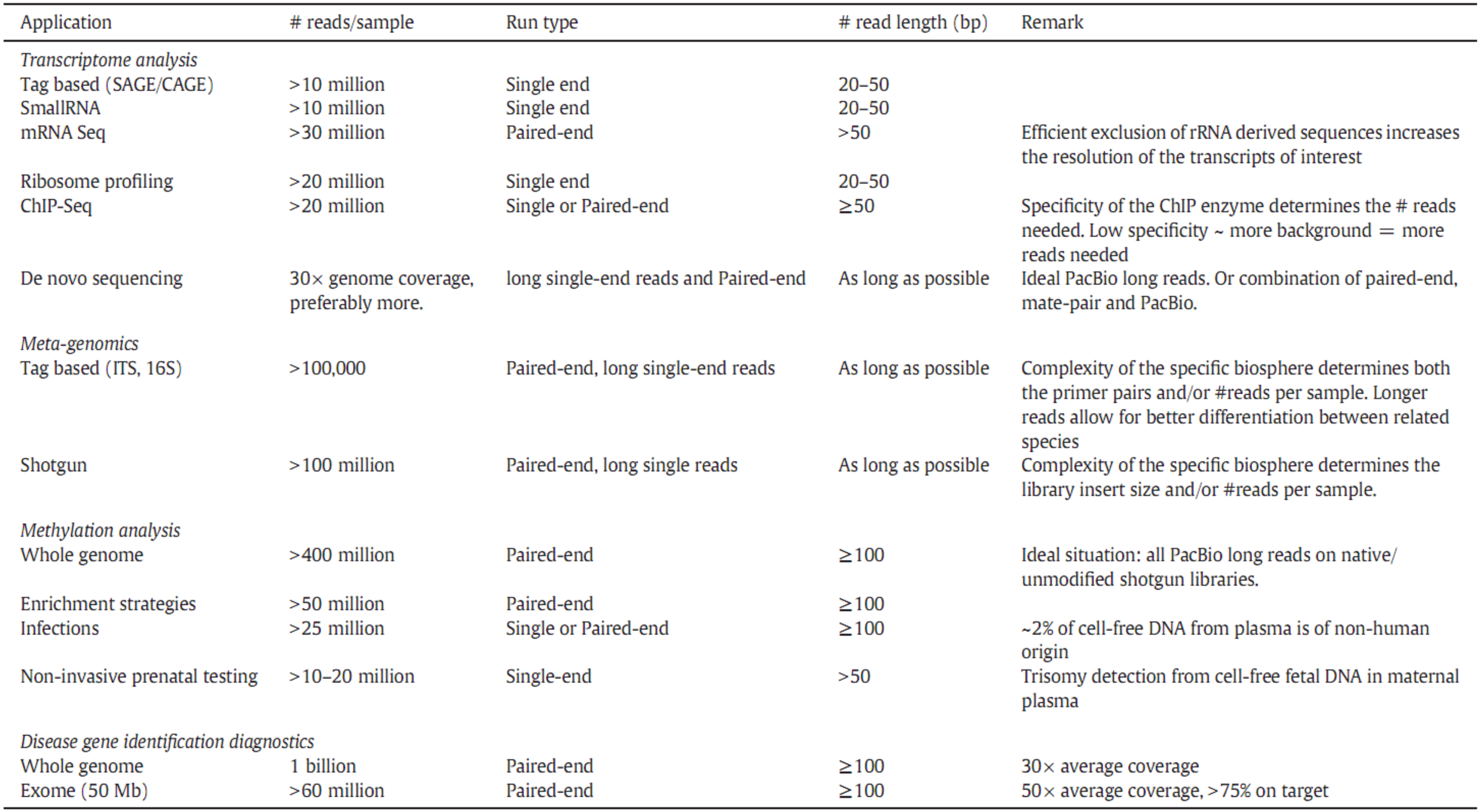
Massively parallel (or next generation) sequencing	High-throughput sequencing nano-technology used to determine the base-pair sequence of DNA/RNA molecules at much larger quantities than previous end-termination (e.g. Sanger sequencing) based sequencing techniques
Mate-pair	Sequence information from two ends of a DNA fragment, usually several thousand base-pairs long
N50	A statistic of a set of contigs (or scaffolds). It is defined as the length for which the collection of all contigs of that length or longer contains at least half of the total of the lengths of the contigs
N90	Equivalent to the N50 statistic describing the length for which the collection of all contigs of that length or longer contains at least 90% of the total of the lengths of the contigs
Optical map	Genomewide, ordered, high-resolution restriction map derived from single, stained DNA molecules. It can be used to improve a genome assembly by matching it to the genomewide pattern of expected restriction sites, as inferred from the genome sequence
Paired-end sequencing	Sequence information from two ends of a short DNA fragment, usually a few hundred base pairs long
Read	Short base-pair sequence inferred from the DNA/RNA template by sequencing
RNA-Seq	High-throughput shotgun transcriptome (cDNA) sequencing. Usually not used synonymous to RNA-sequencing which implies direct sequencing of RNA molecules skipping the cDNA generation step
Scaffold	Two or more contigs joined together using read-pair information
Transcriptome	Set of all RNA molecules transcribed from a DNA template

Robert Ekblom and Jochen B. W. Wolf. A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications 2014. doi:10.1111/eva.12178

3

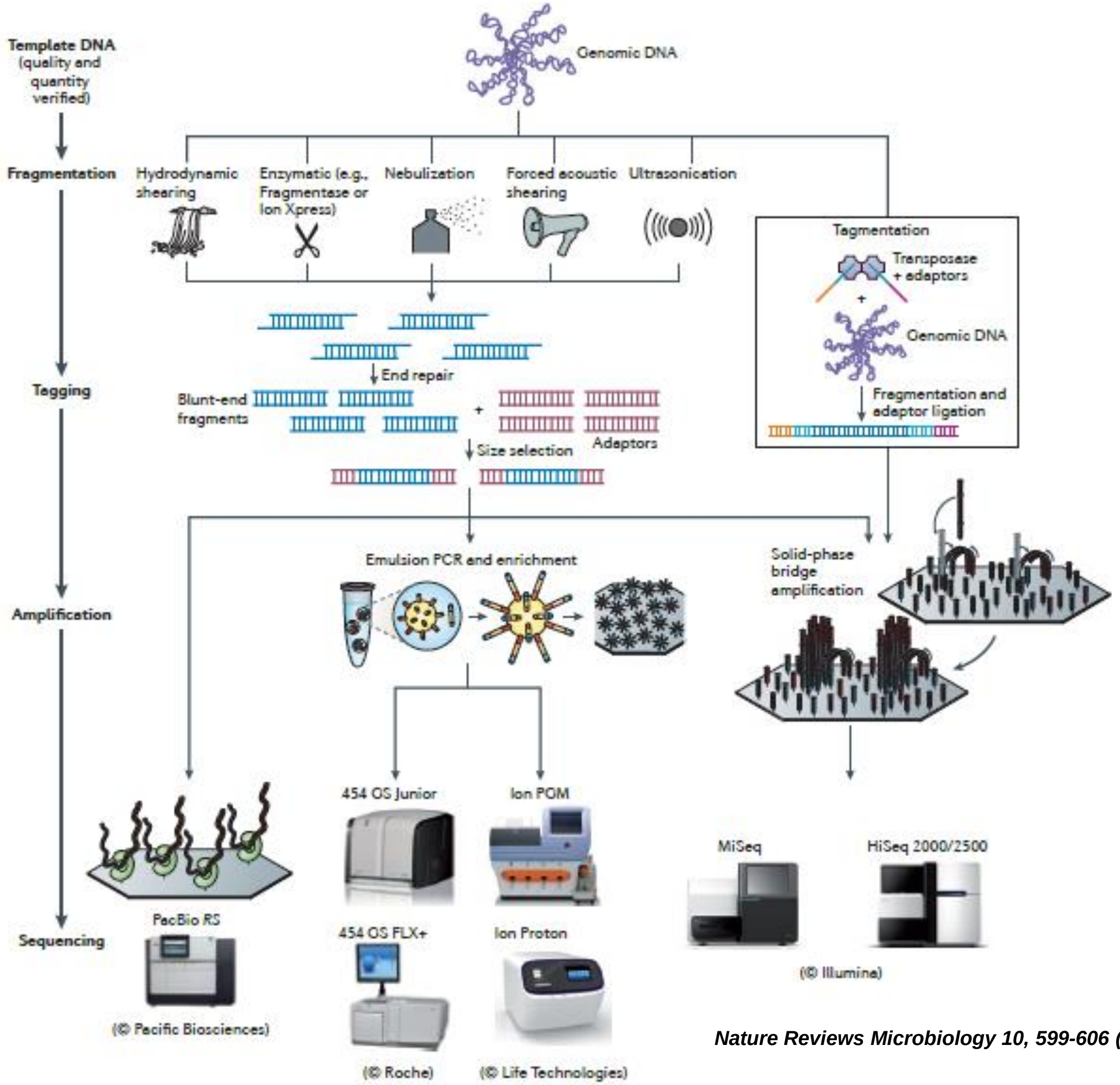
High throughput sequencing platforms

Recommendation for data requirements for a selection of NGS applications



Application	# reads/sample	Run type	# read length (bp)	Remark
Transcriptome analysis				
Tag based (SAGE/CAGE)	>10 million	Single end	20–50	
SmallRNA	>10 million	Single end	20–50	
mRNA Seq	>30 million	Paired-end	>50	Efficient exclusion of rRNA derived sequences increases the resolution of the transcripts of interest
Ribosome profiling	>20 million	Single end	20–50	
ChIP-Seq	>20 million	Single or Paired-end	≥50	Specificity of the ChIP enzyme determines the # reads needed. Low specificity ~ more background = more reads needed
De novo sequencing	30× genome coverage, preferably more.	long single-end reads and Paired-end	As long as possible	Ideal PacBio long reads. Or combination of paired-end, mate-pair and PacBio.
Meta-genomics				
Tag based (ITS, 16S)	>100,000	Paired-end, long single-end reads	As long as possible	Complexity of the specific biosphere determines both the primer pairs and/or #reads per sample. Longer reads allow for better differentiation between related species
Shotgun	>100 million	Paired-end, long single reads	As long as possible	Complexity of the specific biosphere determines the library insert size and/or #reads per sample.
Methylation analysis				
Whole genome	>400 million	Paired-end	≥100	Ideal situation: all PacBio long reads on native/unmodified shotgun libraries.
Enrichment strategies	>50 million	Paired-end	≥100	
Infections	>25 million	Single or Paired-end	≥100	~2% of cell-free DNA from plasma is of non-human origin
Non-invasive prenatal testing	>10–20 million	Single-end	>50	Trisomy detection from cell-free fetal DNA in maternal plasma
Disease gene identification diagnostics				
Whole genome	1 billion	Paired-end	≥100	30× average coverage
Exome (50 Mb)	>60 million	Paired-end	≥100	50× average coverage, >75% on target

High-Throughput Sequencing Technologies



Nature Reviews Microbiology 10, 599-606 (September 2012) | doi:10.1038/nrmicro2850

Figure 1 | **High-throughput sequencing platforms.** The schematic shows the main high-throughput sequencing platforms available to microbiologists today, and the associated sample preparation and template amplification procedures. For full details, see main text. PGM, Personal Genome Machine. The tagmentation schematic is modified, with permission, from REF. 57 © (2010) BioMed Central.

High throughput sequencing platforms

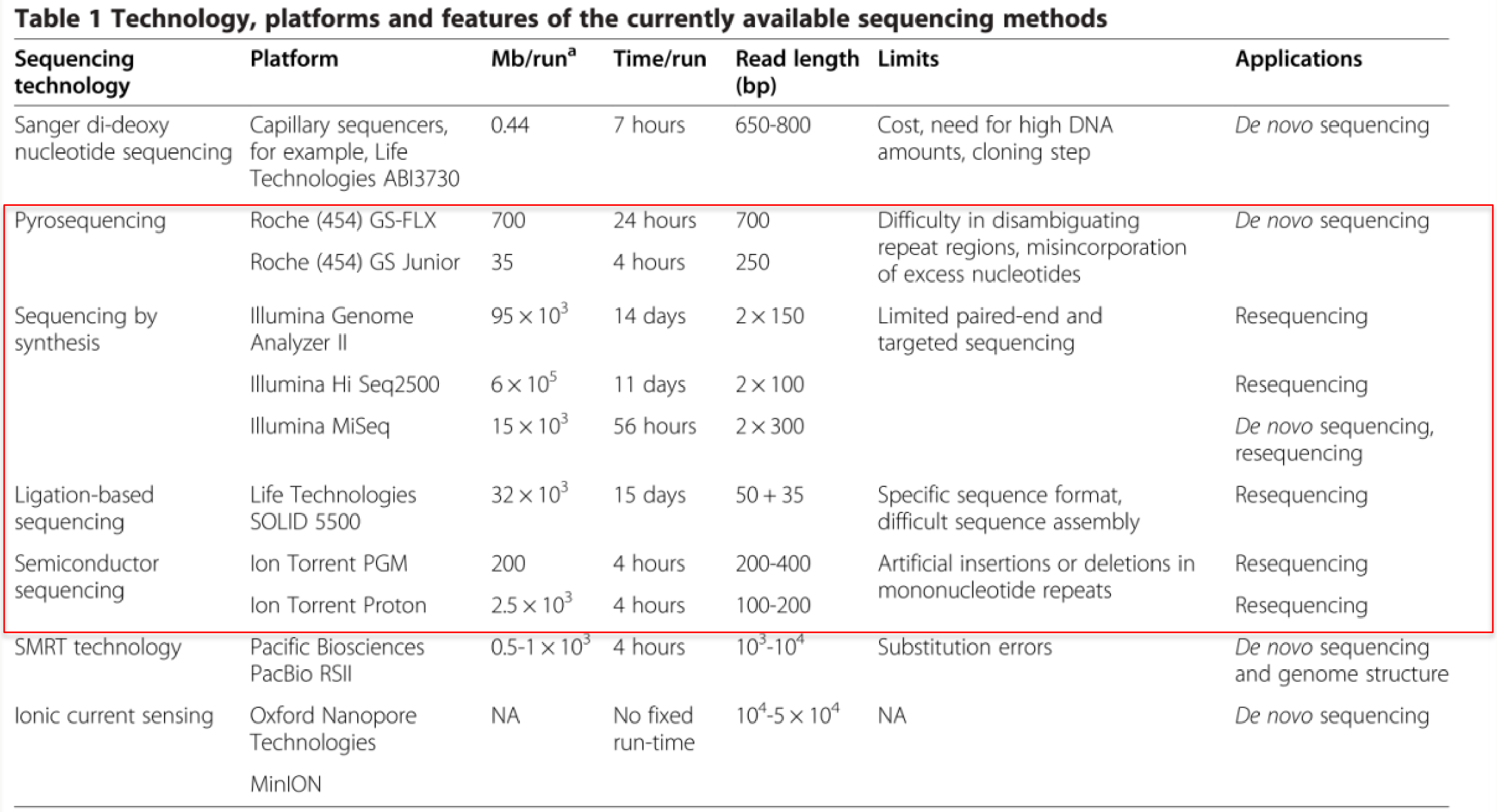
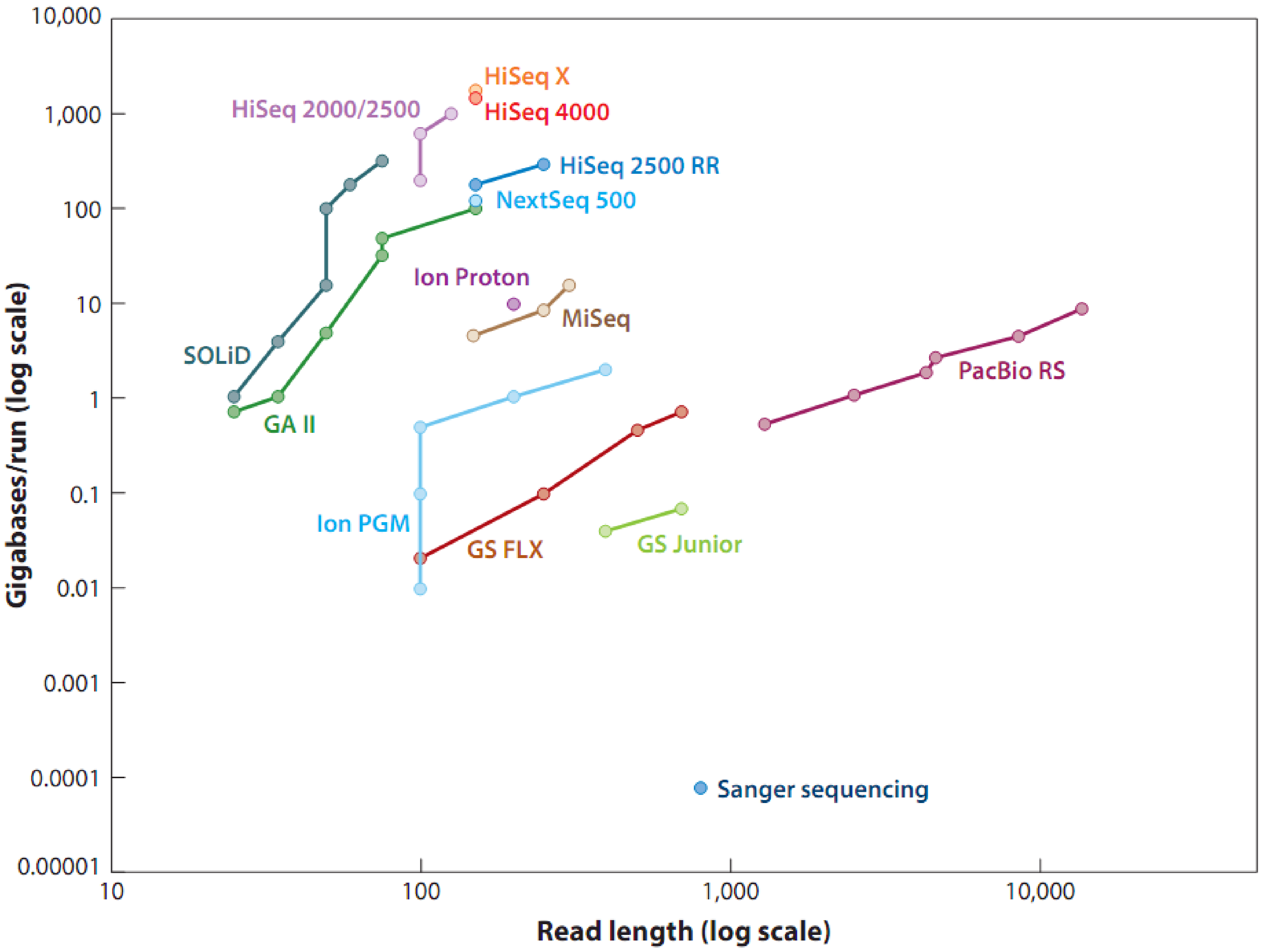
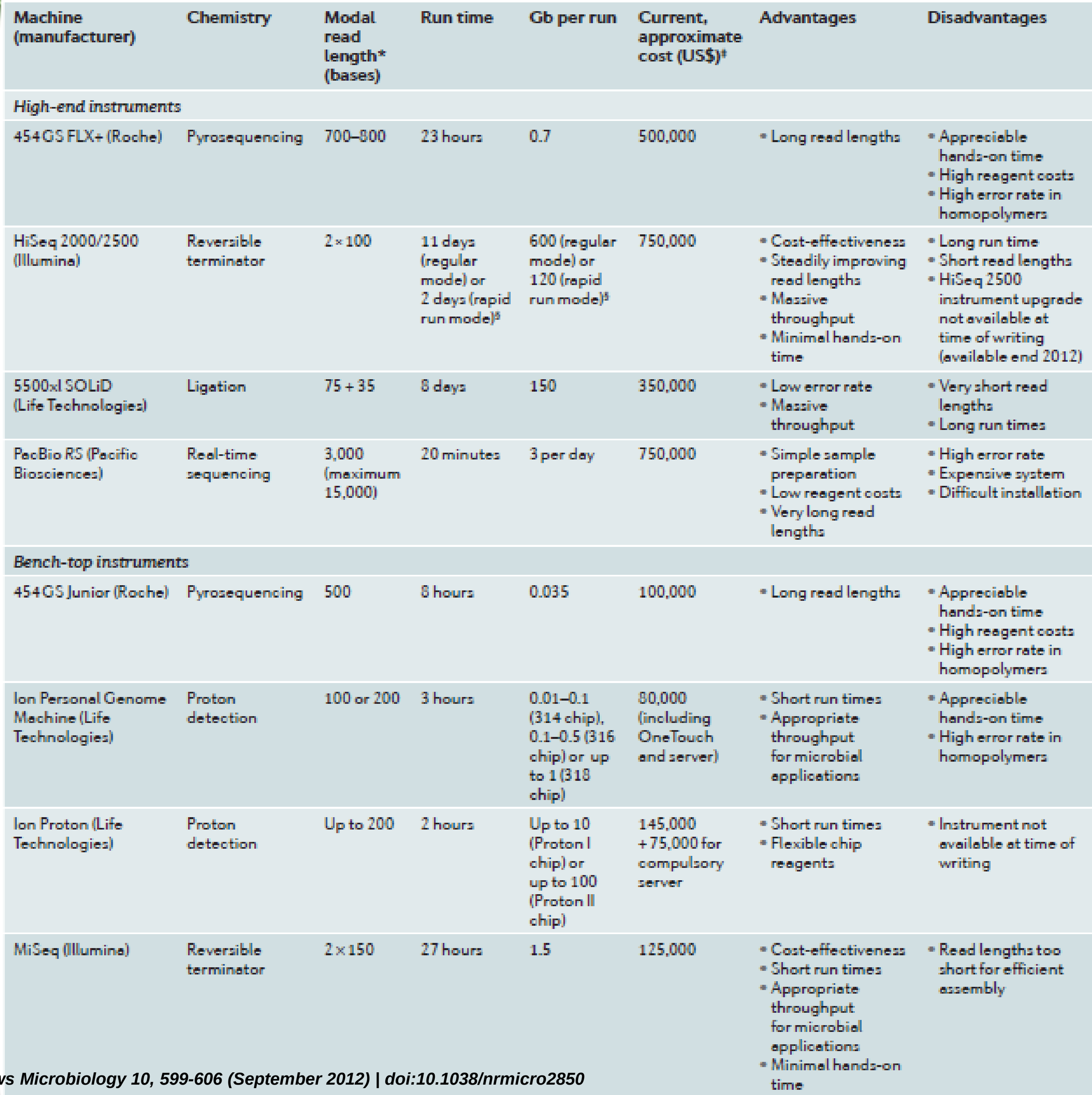


Table 1 Technology, platforms and features of the currently available sequencing methods

Sequencing technology	Platform	Mb/run[a]	Time/run	Read length (bp)	Limits	Applications
Sanger di-deoxy nucleotide sequencing	Capillary sequencers, for example, Life Technologies ABI3730	0.44	7 hours	650-800	Cost, need for high DNA amounts, cloning step	*De novo* sequencing
Pyrosequencing	Roche (454) GS-FLX	700	24 hours	700	Difficulty in disambiguating repeat regions, misincorporation of excess nucleotides	*De novo* sequencing
	Roche (454) GS Junior	35	4 hours	250		
Sequencing by synthesis	Illumina Genome Analyzer II	95×10^3	14 days	2 × 150	Limited paired-end and targeted sequencing	Resequencing
	Illumina Hi Seq2500	6×10^5	11 days	2 × 100		Resequencing
	Illumina MiSeq	15×10^3	56 hours	2 × 300		*De novo* sequencing, resequencing
Ligation-based sequencing	Life Technologies SOLID 5500	32×10^3	15 days	50 + 35	Specific sequence format, difficult sequence assembly	Resequencing
Semiconductor sequencing	Ion Torrent PGM	200	4 hours	200-400	Artificial insertions or deletions in mononucleotide repeats	Resequencing
	Ion Torrent Proton	2.5×10^3	4 hours	100-200		Resequencing
SMRT technology	Pacific Biosciences PacBio RSII	$0.5\text{-}1 \times 10^3$	4 hours	$10^3\text{-}10^4$	Substitution errors	*De novo* sequencing and genome structure
Ionic current sensing	Oxford Nanopore Technologies MinION	NA	No fixed run-time	$10^4\text{-}5 \times 10^4$	NA	*De novo* sequencing

Fournier PE. Et al., Genome Med. 6 (11), 2014

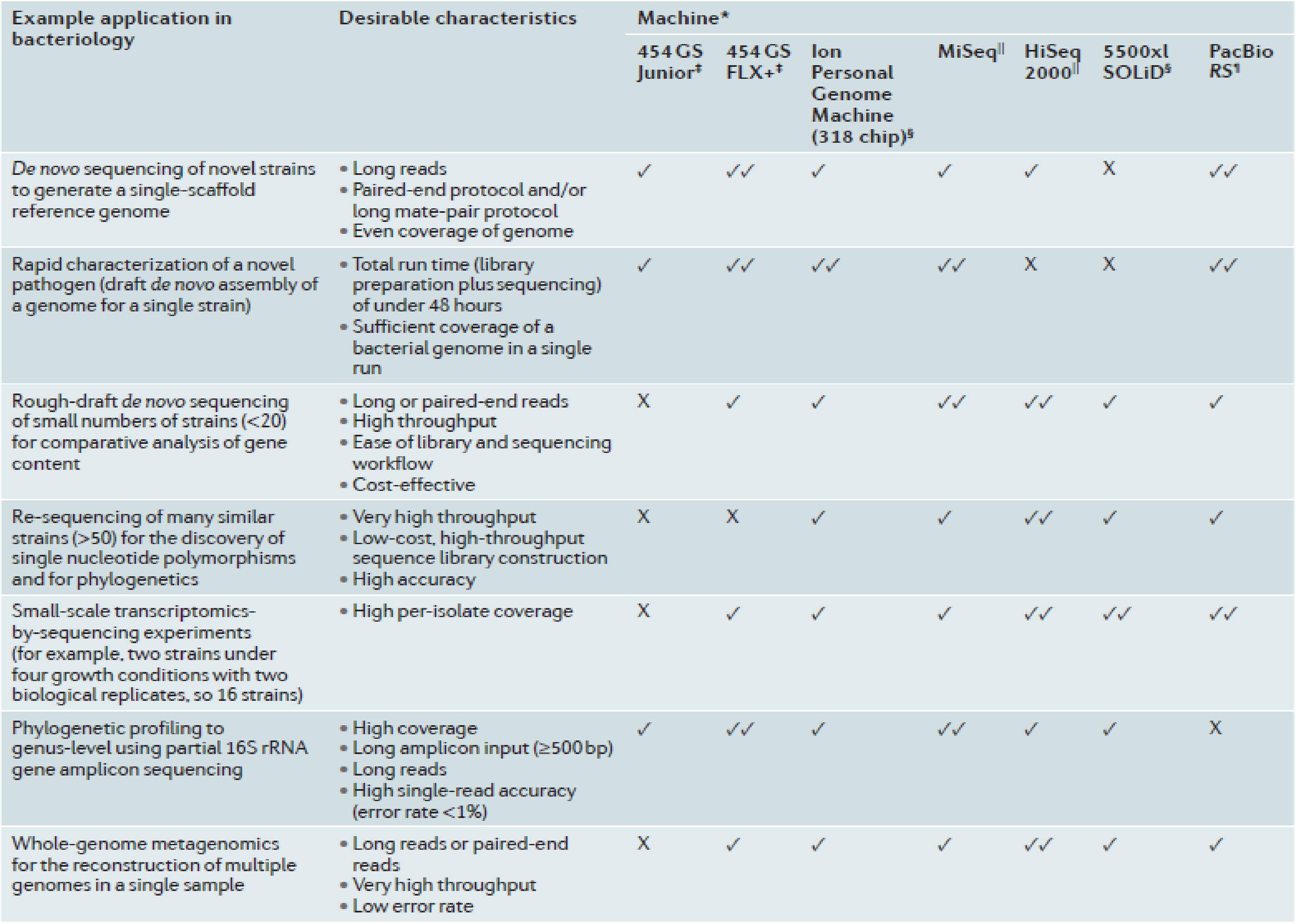

10,000
1,000
100
10
1
0.1
0.01
0.001
0.0001
0.00001
Gigabases/run (log scale)
10
100
1,000
10,000
Read length (log scale)
HiSeq X
HiSeq 4000
HiSeq 2000/2500
HiSeq 2500 RR
NextSeq 500
Ion Proton
MiSeq
SOLiD
GA II
PacBio RS
Ion PGM
GS FLX
GS Junior
Sanger sequencing



Machine (manufacturer)	Chemistry	Modal read length* (bases)	Run time	Gb per run	Current, approximate cost (US$)‡	Advantages	Disadvantages
High-end instruments							
454 GS FLX+ (Roche)	Pyrosequencing	700–800	23 hours	0.7	500,000	• Long read lengths	• Appreciable hands-on time • High reagent costs • High error rate in homopolymers
HiSeq 2000/2500 (Illumina)	Reversible terminator	2 × 100	11 days (regular mode) or 2 days (rapid run mode)§	600 (regular mode) or 120 (rapid run mode)§	750,000	• Cost-effectiveness • Steadily improving read lengths • Massive throughput • Minimal hands-on time	• Long run time • Short read lengths • HiSeq 2500 instrument upgrade not available at time of writing (available end 2012)
5500xl SOLiD (Life Technologies)	Ligation	75 + 35	8 days	150	350,000	• Low error rate • Massive throughput	• Very short read lengths • Long run times
PacBio RS (Pacific Biosciences)	Real-time sequencing	3,000 (maximum 15,000)	20 minutes	3 per day	750,000	• Simple sample preparation • Low reagent costs • Very long read lengths	• High error rate • Expensive system • Difficult installation
Bench-top instruments							
454 GS Junior (Roche)	Pyrosequencing	500	8 hours	0.035	100,000	• Long read lengths	• Appreciable hands-on time • High reagent costs • High error rate in homopolymers
Ion Personal Genome Machine (Life Technologies)	Proton detection	100 or 200	3 hours	0.01–0.1 (314 chip), 0.1–0.5 (316 chip) or up to 1 (318 chip)	80,000 (including OneTouch and server)	• Short run times • Appropriate throughput for microbial applications	• Appreciable hands-on time • High error rate in homopolymers
Ion Proton (Life Technologies)	Proton detection	Up to 200	2 hours	Up to 10 (Proton I chip) or up to 100 (Proton II chip)	145,000 + 75,000 for compulsory server	• Short run times • Flexible chip reagents	• Instrument not available at time of writing
MiSeq (Illumina)	Reversible terminator	2 × 150	27 hours	1.5	125,000	• Cost-effectiveness • Short run times • Appropriate throughput for microbial applications • Minimal hands-on time	• Read lengths too short for efficient assembly

Nature Reviews Microbiology 10, 599-606 (September 2012) | doi:10.1038/nrmicro2850

Applications of Sequencing Technologies



Example application in bacteriology	Desirable characteristics	Machine*						
		454 GS Junior[‡]	454 GS FLX+[‡]	Ion Personal Genome Machine (318 chip)[§]	MiSeq[‖]	HiSeq 2000[‖]	5500xl SOLiD[§]	PacBio RS[¶]
De novo sequencing of novel strains to generate a single-scaffold reference genome	• Long reads • Paired-end protocol and/or long mate-pair protocol • Even coverage of genome	✓	✓✓	✓	✓	✓	X	✓✓
Rapid characterization of a novel pathogen (draft *de novo* assembly of a genome for a single strain)	• Total run time (library preparation plus sequencing) of under 48 hours • Sufficient coverage of a bacterial genome in a single run	✓	✓✓	✓✓	✓✓	X	X	✓✓
Rough-draft *de novo* sequencing of small numbers of strains (<20) for comparative analysis of gene content	• Long or paired-end reads • High throughput • Ease of library and sequencing workflow • Cost-effective	X	✓	✓	✓✓	✓✓	✓	✓
Re-sequencing of many similar strains (>50) for the discovery of single nucleotide polymorphisms and for phylogenetics	• Very high throughput • Low-cost, high-throughput sequence library construction • High accuracy	X	X	✓	✓	✓✓	✓	✓
Small-scale transcriptomics-by-sequencing experiments (for example, two strains under four growth conditions with two biological replicates, so 16 strains)	• High per-isolate coverage	X	✓	✓	✓	✓✓	✓✓	✓✓
Phylogenetic profiling to genus-level using partial 16S rRNA gene amplicon sequencing	• High coverage • Long amplicon input (≥500 bp) • Long reads • High single-read accuracy (error rate <1%)	✓	✓✓	✓	✓✓	✓	✓	X
Whole-genome metagenomics for the reconstruction of multiple genomes in a single sample	• Long reads or paired-end reads • Very high throughput • Low error rate	X	✓	✓	✓	✓✓	✓	✓

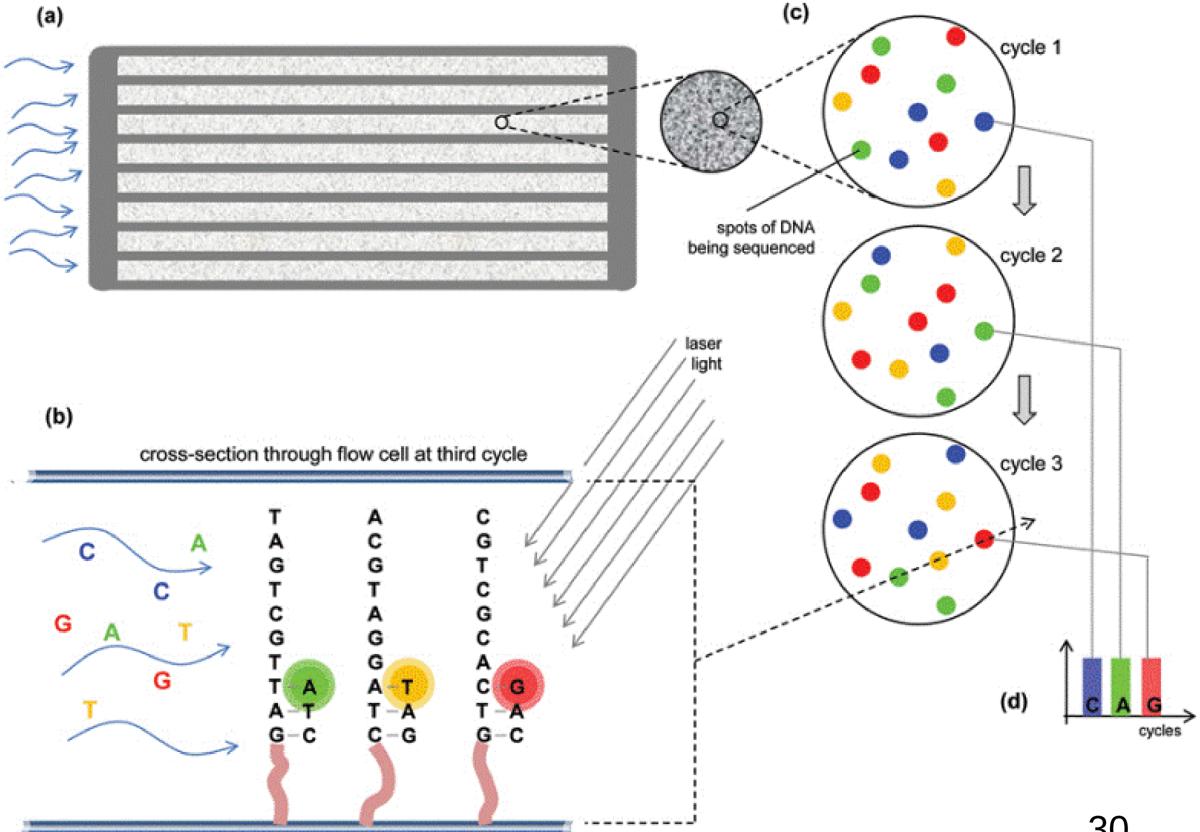
*✓✓, particularly well suited; ✓, suitable; X, not suitable. [‡]From Roche. [§]From Life Technologies. [‖]From Illumina. [¶]From Pacific Biosciences.

Nature Reviews Microbiology 10, 599-606 (September 2012) | doi:10.1038/nrmicro2850

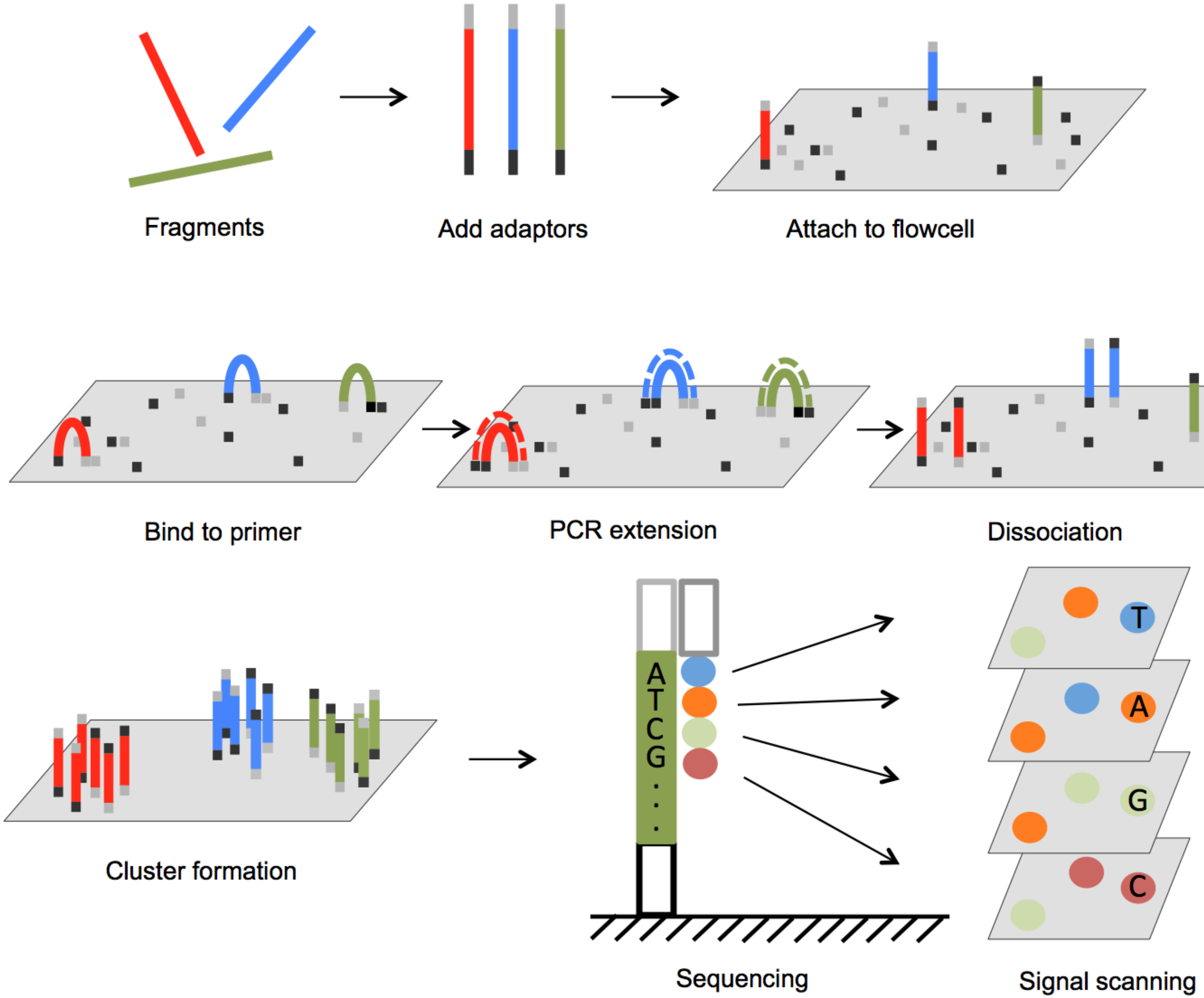
Illumina sequencing

- Illumina Hiseq 2500 sequencing
  - Paired-end reads (2 x 250 bp)
  - 1 isolate = 2 files

- Read length
- Paired-end
- Sequencing read depth
- Sequencing error rate
- Cost
- Others



http://www.historyofnimr.org.uk/files/2015/04/illumina-large.gif



http://www.intechopen.com/source/html/49419/media/image2.png

Animation of Illumina sequencing platform

- https://www.youtube.com/watch?v=HMyCqWhwB8E

4

Analysis pipeline

Simple analysis pipeline

.fastq

Steps	Purposes	Example tools
-QC	(sequencing read checking)	FastQC
-Trimming	(to remove unwanted region of read)	Trimmomatics
-Mapping	(Map the raw reads to ref. e.g. H37Rv)	BWA MEM
-Sam > Bam	(BAM is smaller)	Samtools
-Sorting BAM file	(co-ordinate sort to genome)	Samtools
-Indexing	(data structuring for strings)	Samtools
-Realignment	(decrease mapping error)	GATK
-Stat report	(see info of mapping & parameters)	Samtools/ GATK
-Variant calling	(call the variant)	Samtools
-Variant filtering	(filter low quality variants)	Samtools
-Varian annotation	(to annotate the variant to the ref.)	snpEff

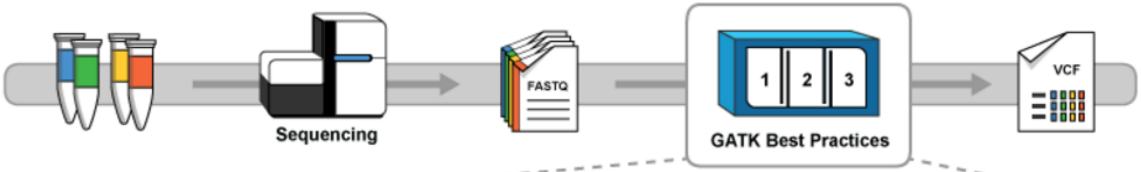
Sideline analysis

-Phylogenetic analysis	MEGA
-Variant comparisons	*Manual*
-Others	

results

Not include structural variants (SVs)

Not include Denovo assembly



1. PRE-PROCESSING

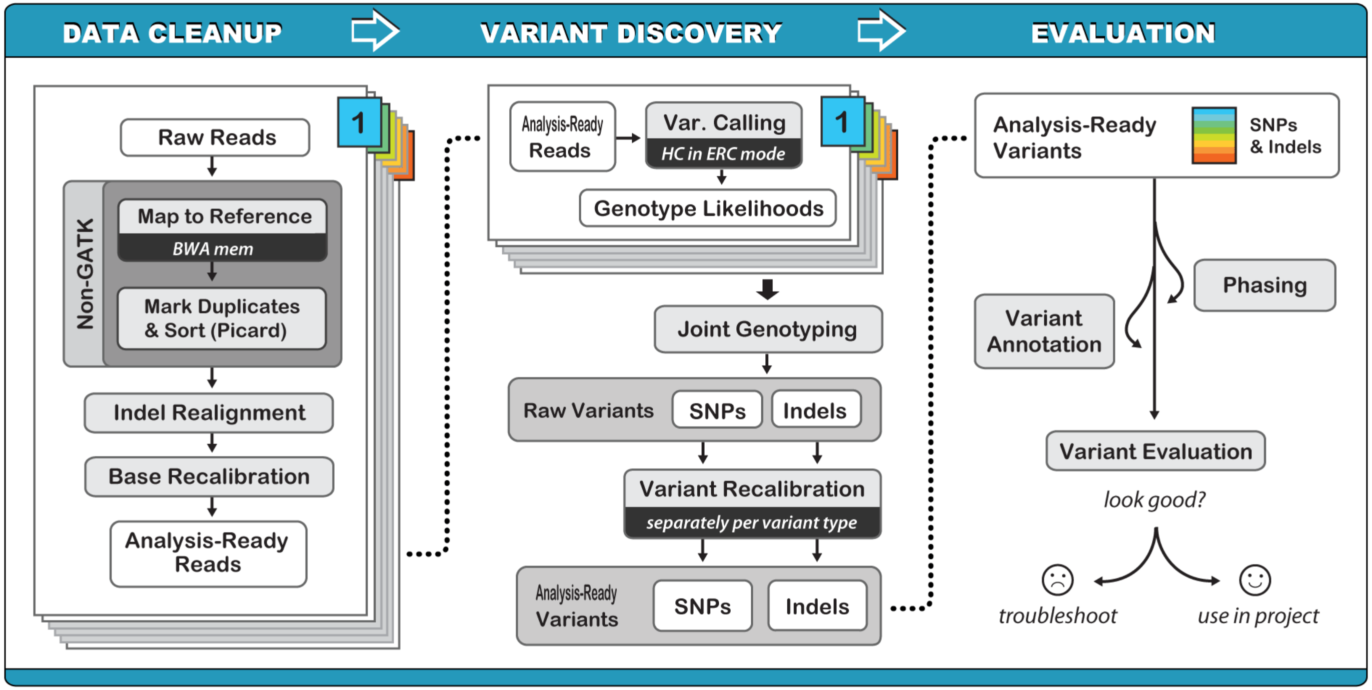
Pre-processing starts from raw sequence data, either in FASTQ or uBAM format, and produces analysis-ready BAM files. Processing steps include alignment to a reference genome as well as some data cleanup operations to correct for technical biases and make the data suitable for analysis.

2. VARIANT DISCOVERY

Variant Discovery starts from analysis-ready BAM files and produces a callset in VCF format. Processing involves identifying sites where one or more individuals display possible genomic variation, and applying filtering methods appropriate to the experimental design.

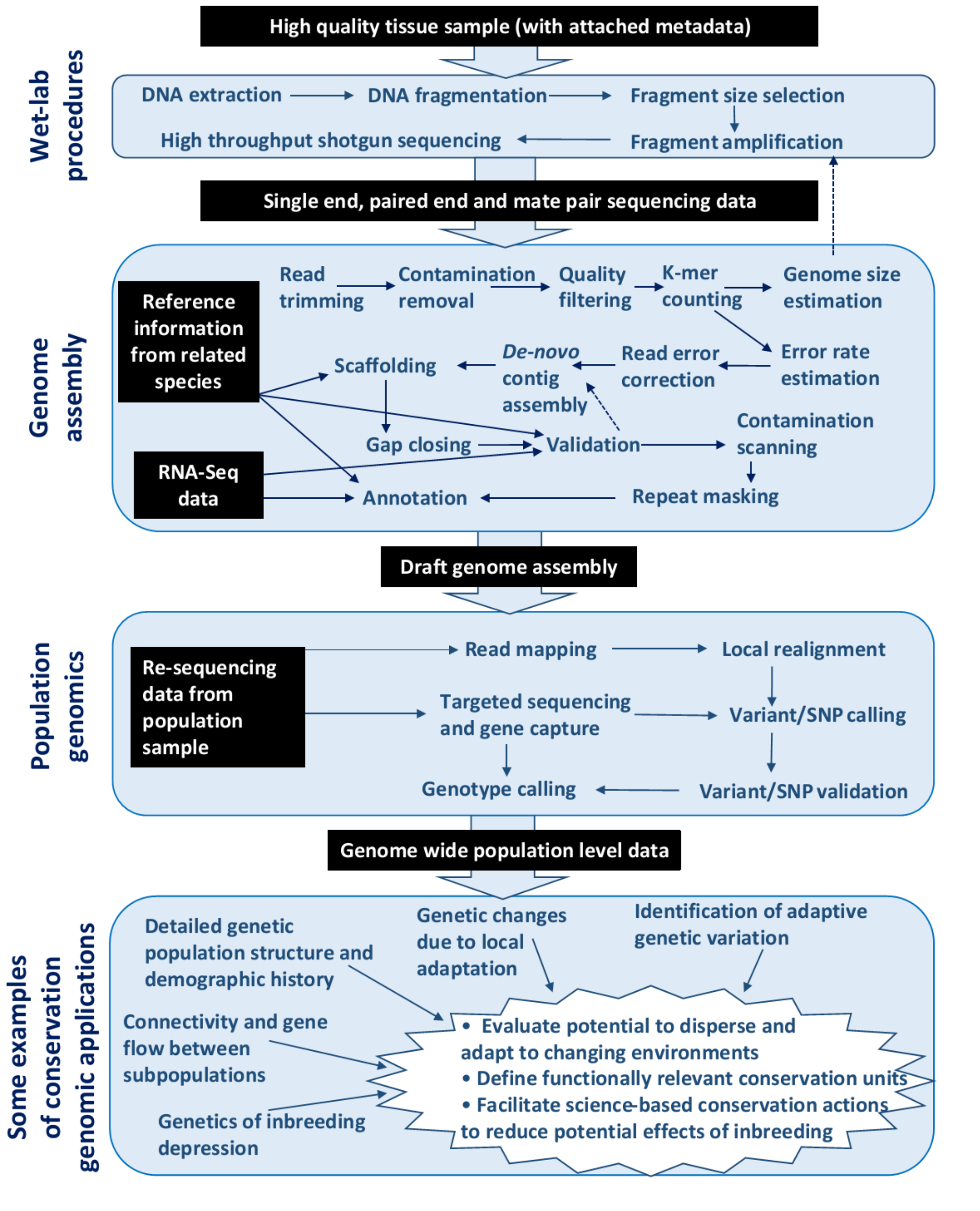
3. CALLSET REFINEMENT

Callset Refinement starts and ends with a VCF callset. Processing involves using meta-data to assess and improve genotyping accuracy, attach additional information and evaluate the overall quality of the callset.



https://www.broadinstitute.org/gatk/guide/best-practices

Workflow of a typical whole-genome sequencing analysis



High quality tissue sample (with attached metadata)

Wet-lab procedures

DNA extraction → DNA fragmentation → Fragment size selection → Fragment amplification → High throughput shotgun sequencing

Single end, paired end and mate pair sequencing data

Genome assembly

Read trimming → Contamination removal → Quality filtering → K-mer counting → Genome size estimation

Error rate estimation → Read error correction → *De-novo* contig assembly → Scaffolding → Gap closing → Validation → Contamination scanning → Repeat masking → Annotation

Reference information from related species

RNA-Seq data

Draft genome assembly

Population genomics

Re-sequencing data from population sample

Read mapping → Local realignment → Variant/SNP calling

Targeted sequencing and gene capture → Variant/SNP calling

Genotype calling

Variant/SNP validation

Genome wide population level data

Some examples of conservation genomic applications

Detailed genetic population structure and demographic history

Genetic changes due to local adaptation

Identification of adaptive genetic variation

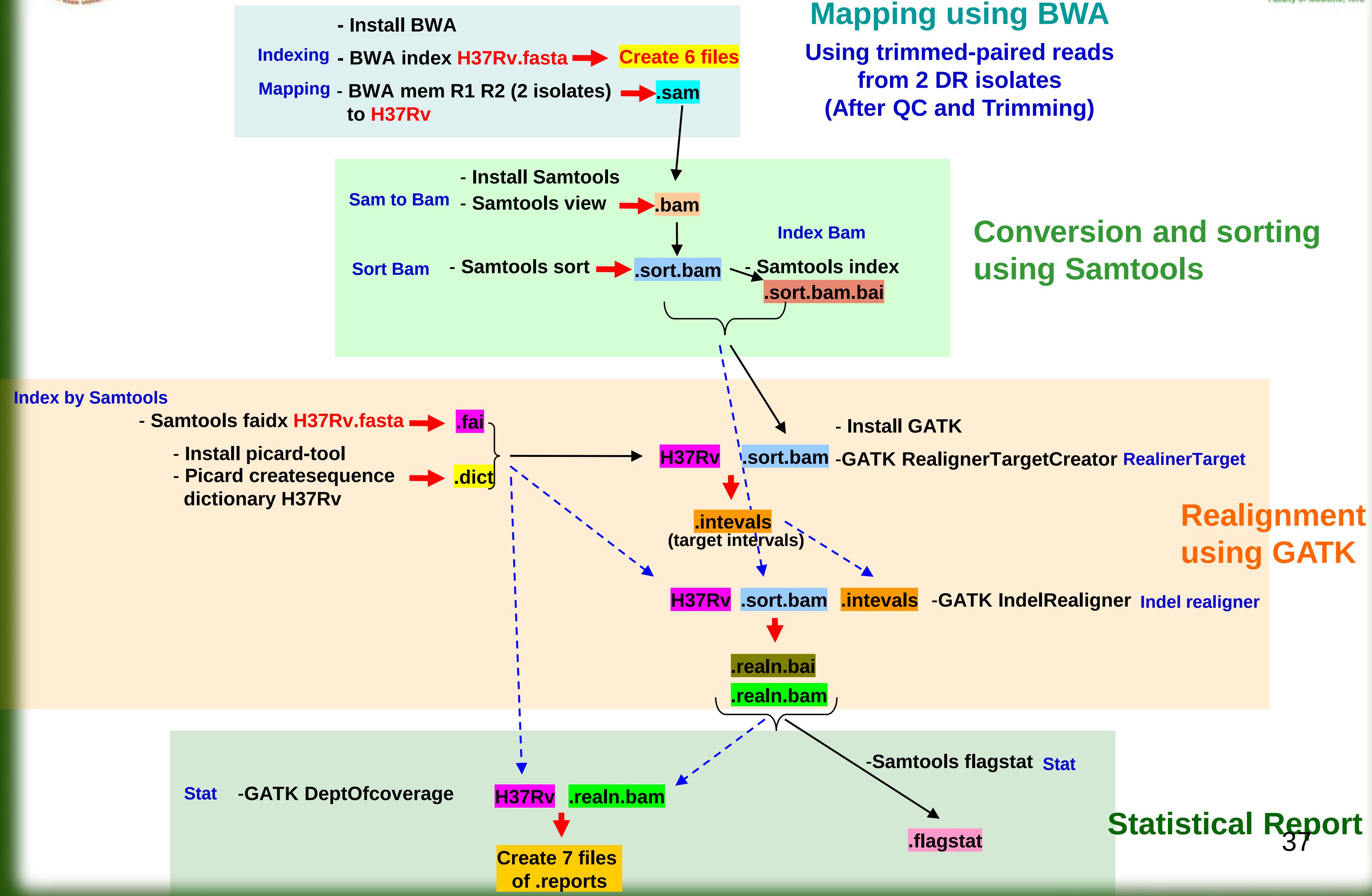
Connectivity and gene flow between subpopulations

Genetics of inbreeding depression

- Evaluate potential to disperse and adapt to changing environments
- Define functionally relevant conservation units
- Facilitate science-based conservation actions to reduce potential effects of inbreeding

Robert Ekblom and Jochen B. W. Wolf, Evolutionary Applications 2014

Analysis pipeline of NGS Illumina data (.fastq)



Mapping using BWA

Using trimmed-paired reads from 2 DR isolates (After QC and Trimming)

- Install BWA

Indexing - BWA index H37Rv.fasta → Create 6 files

Mapping - BWA mem R1 R2 (2 isolates) to H37Rv → .sam

Conversion and sorting using Samtools

- Install Samtools

Sam to Bam - Samtools view → .bam

Sort Bam - Samtools sort → .sort.bam

Index Bam - Samtools index → .sort.bam.bai

Realignment using GATK

Index by Samtools

- Samtools faidx H37Rv.fasta → .fai

- Install picard-tool
- Picard createsequence dictionary H37Rv → .dict

- Install GATK

H37Rv .sort.bam -GATK RealignerTargetCreator RealinerTarget

.intevals (target intervals)

H37Rv .sort.bam .intevals -GATK IndelRealigner Indel realigner

.realn.bai

.realn.bam

Statistical Report

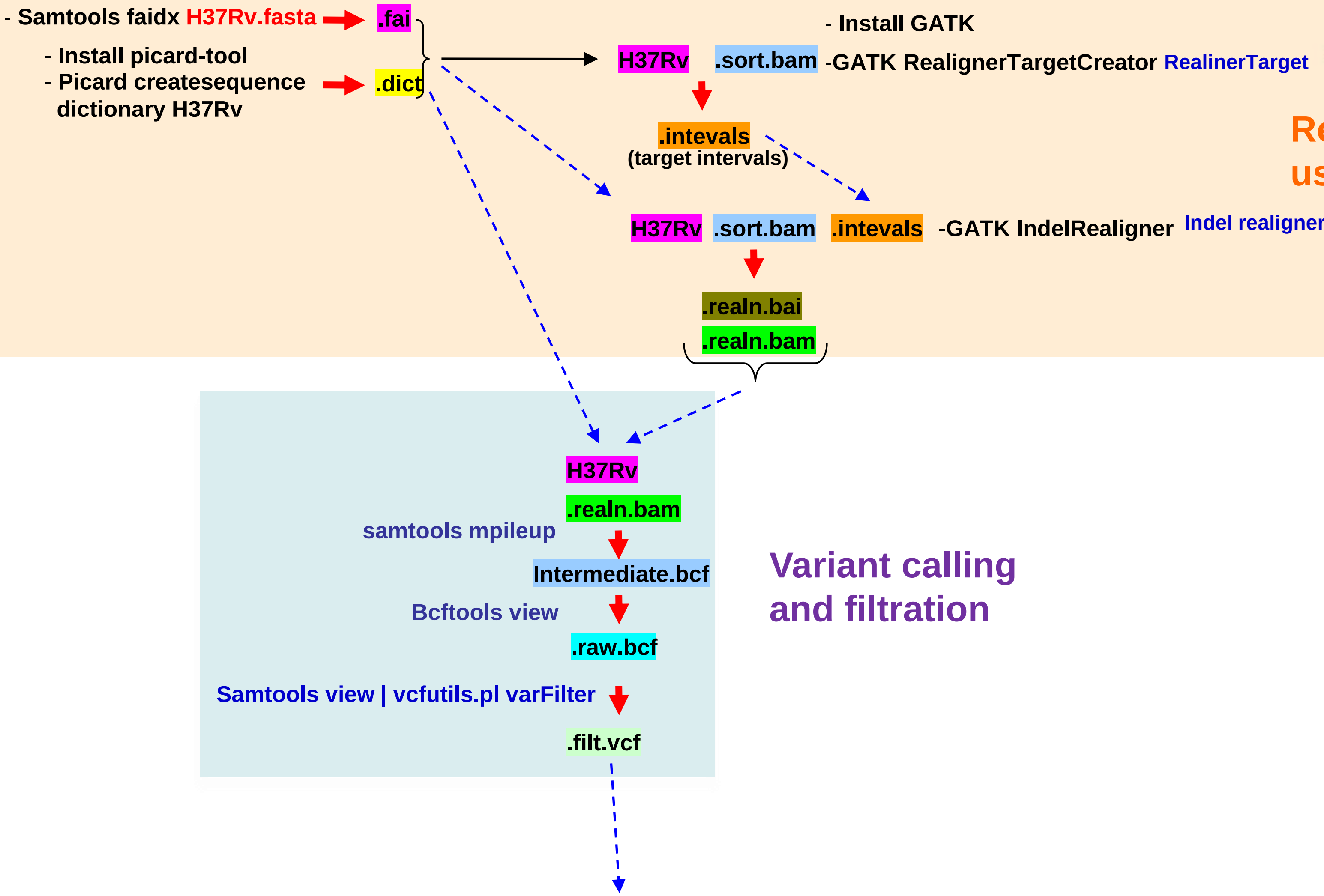
Stat -GATK DeptOfcoverage

H37Rv .realn.bam

Create 7 files of .reports

-Samtools flagstat Stat

.flagstat



Index by Samtools

- Samtools faidx H37Rv.fasta → .fai
- Install picard-tool
- Picard createsequence dictionary H37Rv → .dict

- Install GATK

H37Rv .sort.bam -GATK RealignerTargetCreator RealinerTarget

.intevals (target intervals)

H37Rv .sort.bam .intevals -GATK IndelRealigner Indel realigner

.realn.bai

.realn.bam

Realignme
using GATK

H37Rv

.realn.bam

samtools mpileup

Intermediate.bcf

Bcftools view

.raw.bcf

Samtools view | vcfutils.pl varFilter

.filt.vcf

Variant calling and filtration

Variant annotation and evaluation

Can do multiple files too, what is the different?
Should do separate or combined isolates

So, it should be separate because 1. good coverage already 2. easier to compare later 3. reflect the actual situation
It should be combine for multiple sequence alignment and comparisons

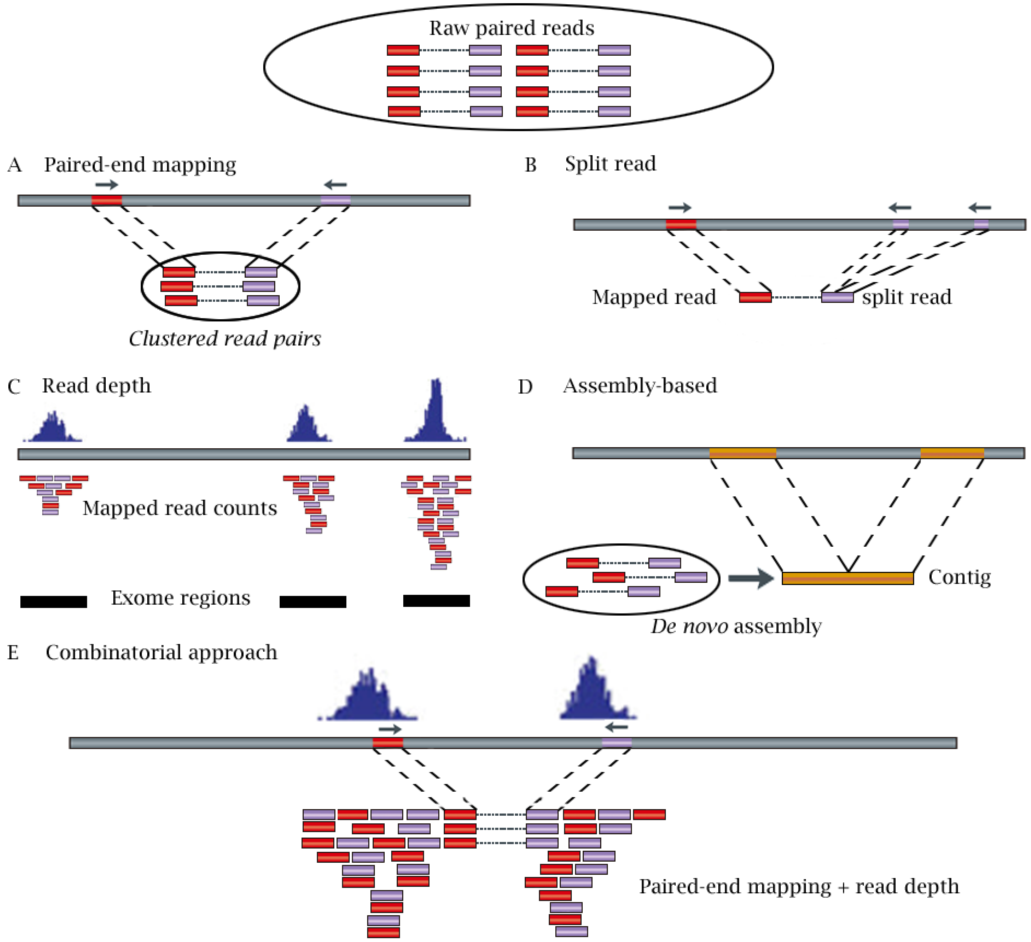
5

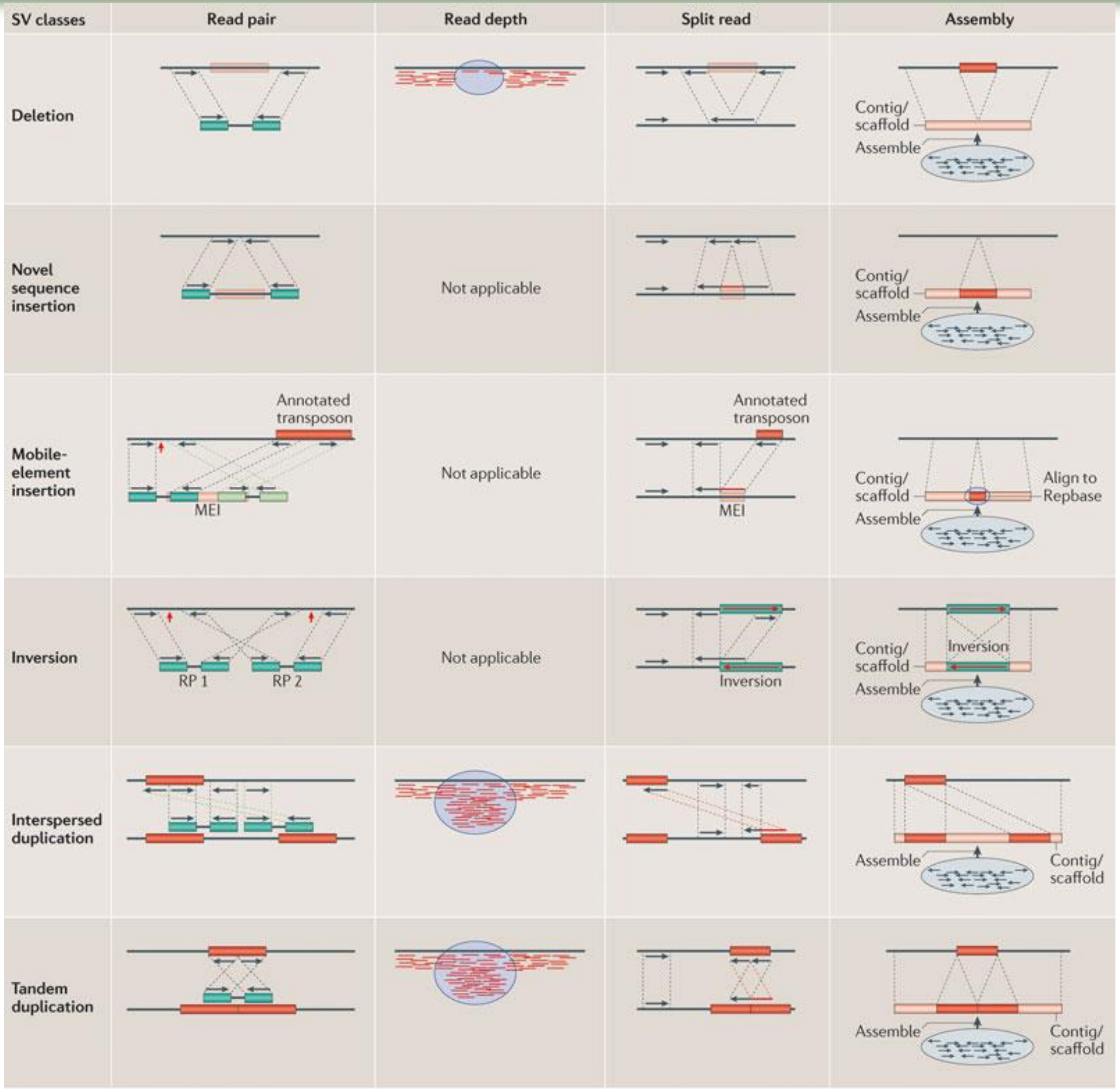
Technical information

Consideration

- **The tools that used in the analysis pipeline for WGS analysis of bacteria were mostly adopted from human genome project.**

Related terms





http://www.nature.com/nrg/journal/v12/n5/images/nrg2958-f2.jpg

Nature Reviews | Genetics

Heterozygous SNP

- The coverage that support both Ref vs variant allele
- Whether or not it should be excluded, depend on biology
- Inside the vcf file can see the number

Position	Ref	A	C	G	T	N	Total
1							
2							
3							
.							
.							
235	A	23	26				49

Source of false variant

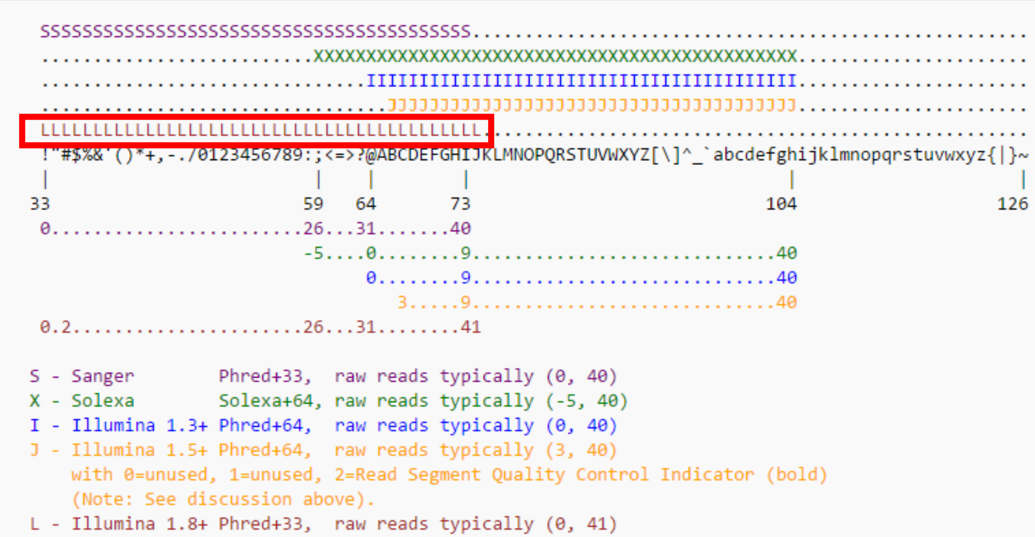
- Sequencing error: missing data at particular region
- Mapping error
- Calling error: genotyping error (e.g. heterozygous)
- Software: parameters
- Monomorphic SNP: unique to specific population
- Reference: the reference error

Fastq file format

- Each entry in a FASTQ file consists of four lines:
  - Sequence identifier
  - Sequence
  - Quality score identifier line (consisting of a +)
  - Quality score

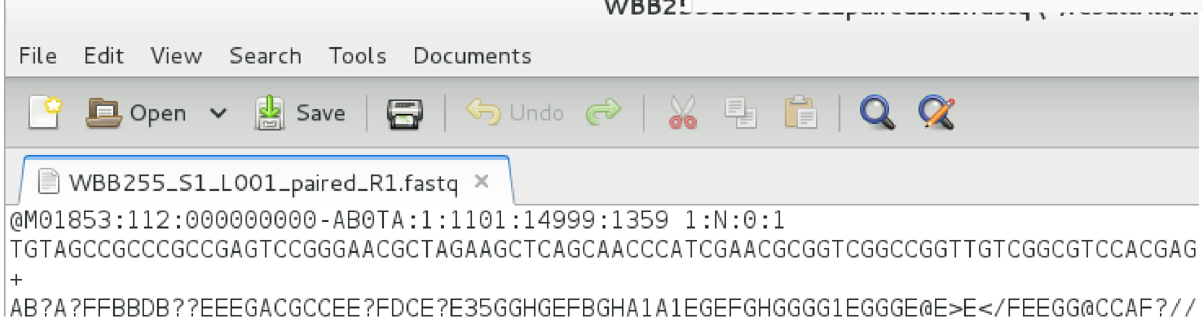
NOTES:

1. If there are 1,000 raw sequence reads then there will be 4x1,000 = 4,000 lines in the FASTQ file.
2. There should be the same number of sequences/lines in the corresponding FASTQ files if the sequencing run is paired-end.



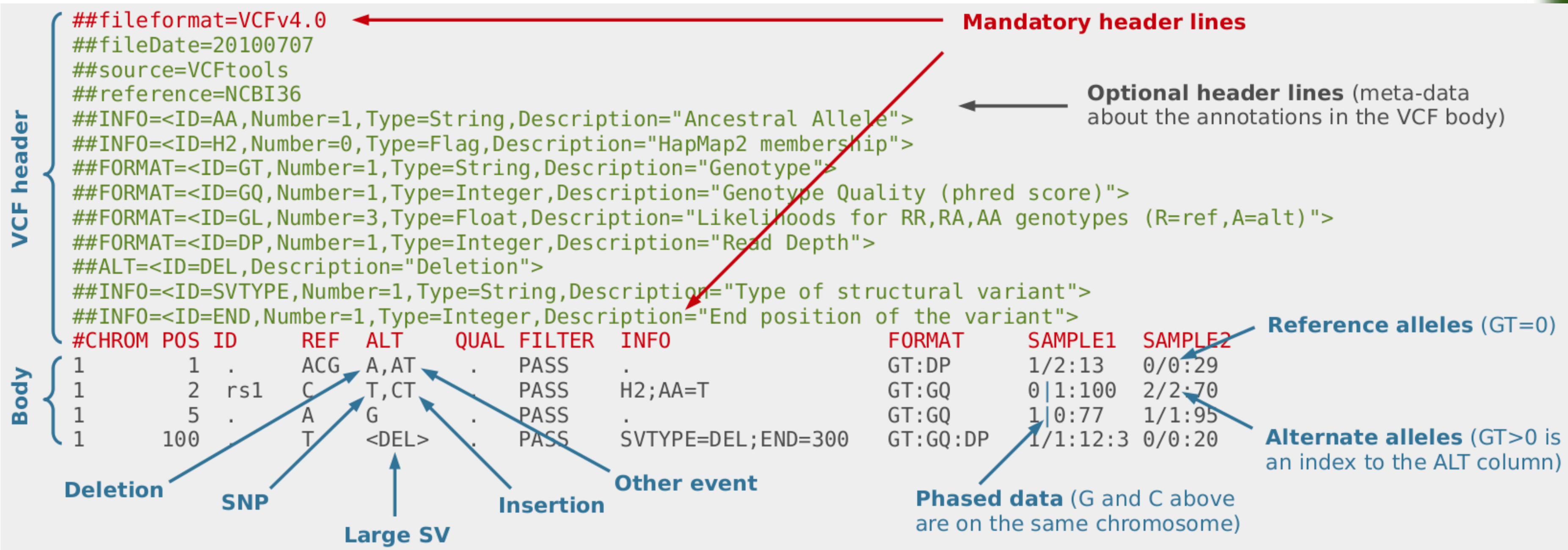
```
SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
.................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
|                         |    |        |                              |                     |
33                        59   64       73                            104                   126
0........................26...31.......40
                         -5....0........9.............................40
                               0........9.............................40
                                  3.....9.............................40
0.2......................26...31........41

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```



```
1 @M01853:112:000000000-AB0TA:1:1101:14999:1359 1:N:0:1
2 TGTAGCCGCCCGCCGAGTCCGGGAACGCTAGAAGCTCAGCAACCCATCGAACGCGGTCGGCCGGTTGTCGGCGTCCACGAG
3 +
4 AB?A?FFBBDB??EEEGACGCCEE?FDCE?E35GGHGEFBGHA1A1EGEFGHGGGG1EGGGE@E>E</FEEGG@CCAF?//
```

VCF format



```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID   REF  ALT    QUAL FILTER  INFO                 FORMAT     SAMPLE1   SAMPLE2
1        1 .    ACG  A,AT   .    PASS    .                    GT:DP      1/2:13    0/0:29
1        2 rs1  C    T,CT   .    PASS    H2;AA=T              GT:GQ      0|1:100   2/2:70
1        5 .    A    G      .    PASS    .                    GT:GQ      1|0:77    1/1:95
1      100 .    T    <DEL>  .    PASS    SVTYPE=DEL;END=300   GT:GQ:DP   1/1:12:3  0/0:20
```

VCF header

Body

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Deletion

SNP

Large SV

Insertion

Other event

Phased data (G and C above are on the same chromosome)

Data processing

- Linux based software (open source) (lacking of basic can make you headache)

Data Analysis

- R programming

New analyzer should aware

- Practice for Linux and how to use Terminal
- Symbolic and option of the bash language
- Typing: beware a case sensitive letter
- Beware extra “space”, “tab”, “enter” that cause error to the command line

6

Practice in NGS analysis I

Need to prepare

1. Copy all files (28 Gb) to your computer
2. Install VirtualBox
3. Try to lunch the VM and setting share folder
4. Send me the e-mail (kiatichai@kku.ac.th) “asking for the code”

VirtualBox

To run the command line tools, we use Unix based OS

- Linux
- Mac OSX

64 bit OS system

Window !?!?

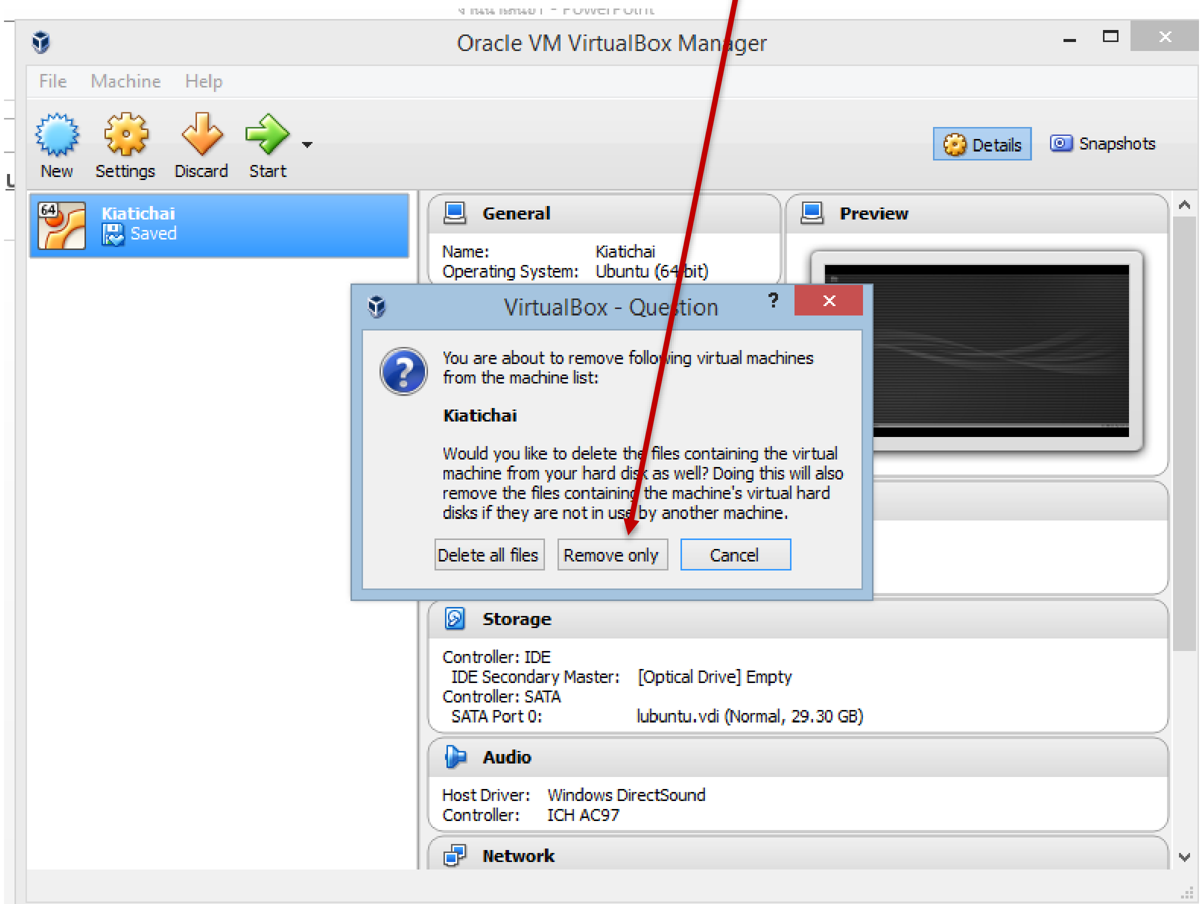
- We will use virtual machine (VirtualBox) based on linux OS
- All necessary software are preinstalled (normally you install by yourself)
- There are codes that can be copy and paste
  (normally you type yourself and there are options to consider)

Install Oracle VM VirtualBox >> Create VM by go to NEW
>> Select Ubantu 64 bit >> Set System (CPU and Ram) >> Set Share folder

Note There is a way to do “share folder” between host (Window) and VM (Ubantu), (you do it later)

Be careful!!!

When you want to delete the old set up
DO NOT select "delete all file"
As all file will be gone.
Just choose remove only



Name: Kiatichai
Operating System: Ubuntu (64-bit)

VirtualBox - Question

You are about to remove following virtual machines from the machine list:

Kiatichai

Would you like to delete the files containing the virtual machine from your hard disk as well? Doing this will also remove the files containing the machine's virtual hard disks if they are not in use by another machine.

Delete all files | Remove only | Cancel

Controller: IDE
IDE Secondary Master: [Optical Drive] Empty
Controller: SATA
SATA Port 0: lubuntu.vdi (Normal, 29.30 GB)

Host Driver: Windows DirectSound
Controller: ICH AC97

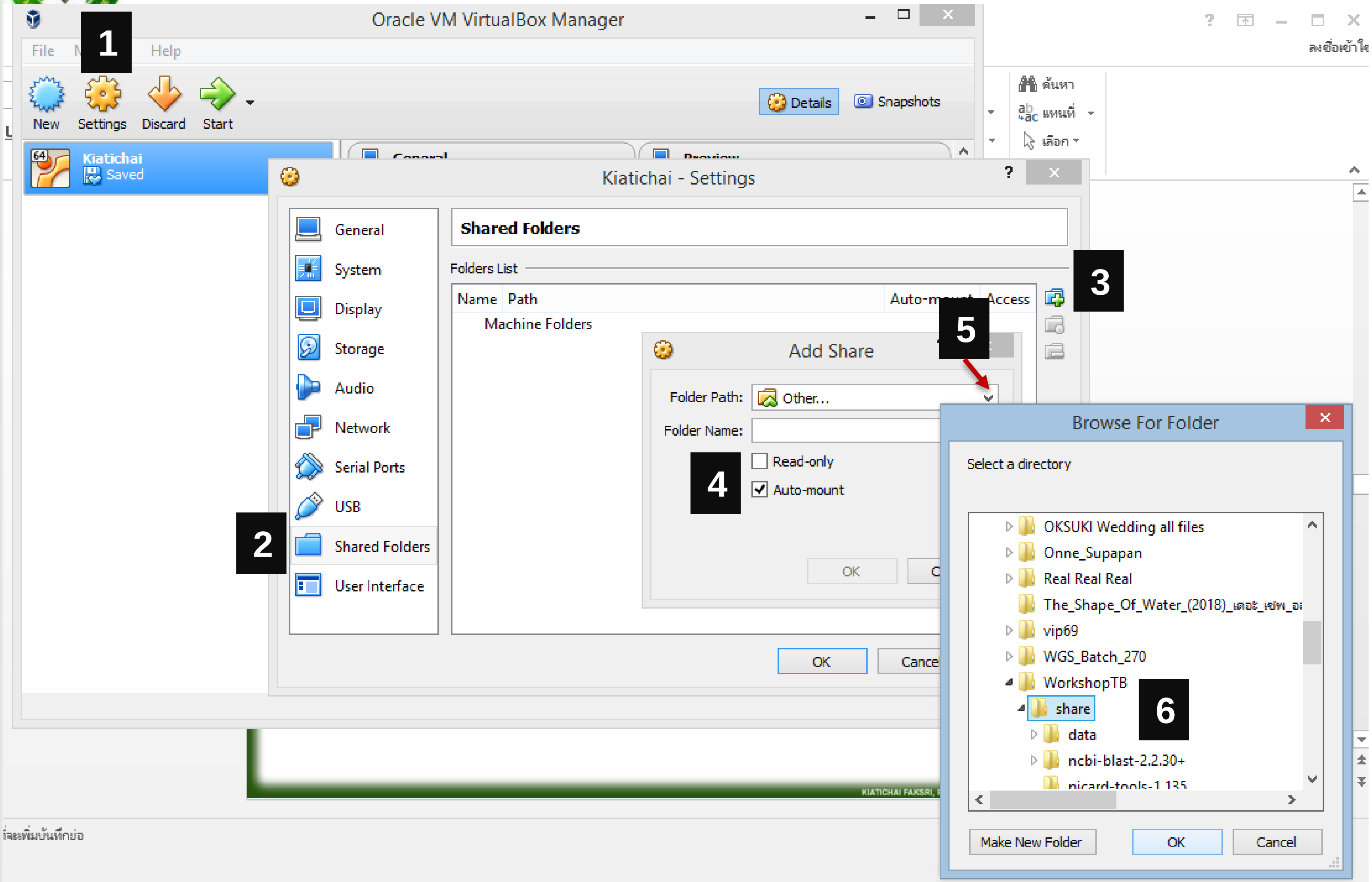
Making share folder for VM

Share folder: please follow the step by step guide from

http://www.htpcbeginner.com/setup-virtualbox-shared-folders-linux-windows/

- Select both automount and permanent
- Select
- Select the Folder location>>>> /media/sf_Share

Making share folder for VM



Select Setting >> 2. Select "Share Folder" >>> 3. Add new share folder >>
4. Tick "Auto-mount" >> 5. Click drop down list of "Folder path" >>
6. Select the "share" folder inside the "Workshop TB" where you copied your file >> Then click OK

Location of necessary file

- /usr/bin/ = for all miscellaneous software , e.g. samtools, bwa
- /usr/local/bin = for fastqc, trimmomatics, GenomAnalysisTK.jar, vcfutils.pl
- /home/user/program = TBprofiler, SpoTyping, snfEFF, piecard

Tested strains

- There are 2 *Mtb* strains,
- Illumina Miseq, paired, 250 bp read length, 50X read depth

1.) TB1_259: (TB1_259_R1.fastq/ TB1_259_R2.fastq)
2.) TB2_260: (TB2_260 _R1.fastq/ TB2_260 _R2.fastq)

Basic for using Terminal (Bash)

- File location and patch
- List the file = “ls”, “ls –alh”
- Go to particular directory by “cd”
- See the file by “ls”
- Hidden file = ./
- Refresh the terminal screen = Ctrl + l
- Others, you will know it when you do it, might stuck! but you can google it

Simple analysis pipeline

.fastq

Steps	Purposes	Example tools
-QC	(sequencing read checking)	FastQC
-Trimming	(to remove unwanted region of read)	Trimmomatics
-Mapping	(Map the raw reads to ref. e.g. H37Rv)	BWA MEM
-Sam > Bam	(BAM is smaller)	Samtools
-Sorting BAM file	(co-ordinate sort to genome)	Samtools
-Indexing	(data structuring for strings)	Samtools
-Realignment	(decrease mapping error)	GATK
-Stat report	(see info of mapping & parameters)	Samtools/ GATK
-Variant calling	(call the variant)	Samtools
-Variant filtering	(filter low quality variants)	Samtools
-Varian annotation	(to annotate the variant to the ref.)	snpEff

Sideline analysis

-Phylogenetic analysis	MEGA
-Variant comparisons	*Manual*
-Others	

results

Not include structural variants (SVs)

Not include Denovo assembly

1 QC: fastQ

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

To see option

```
$ fastqc -h
```

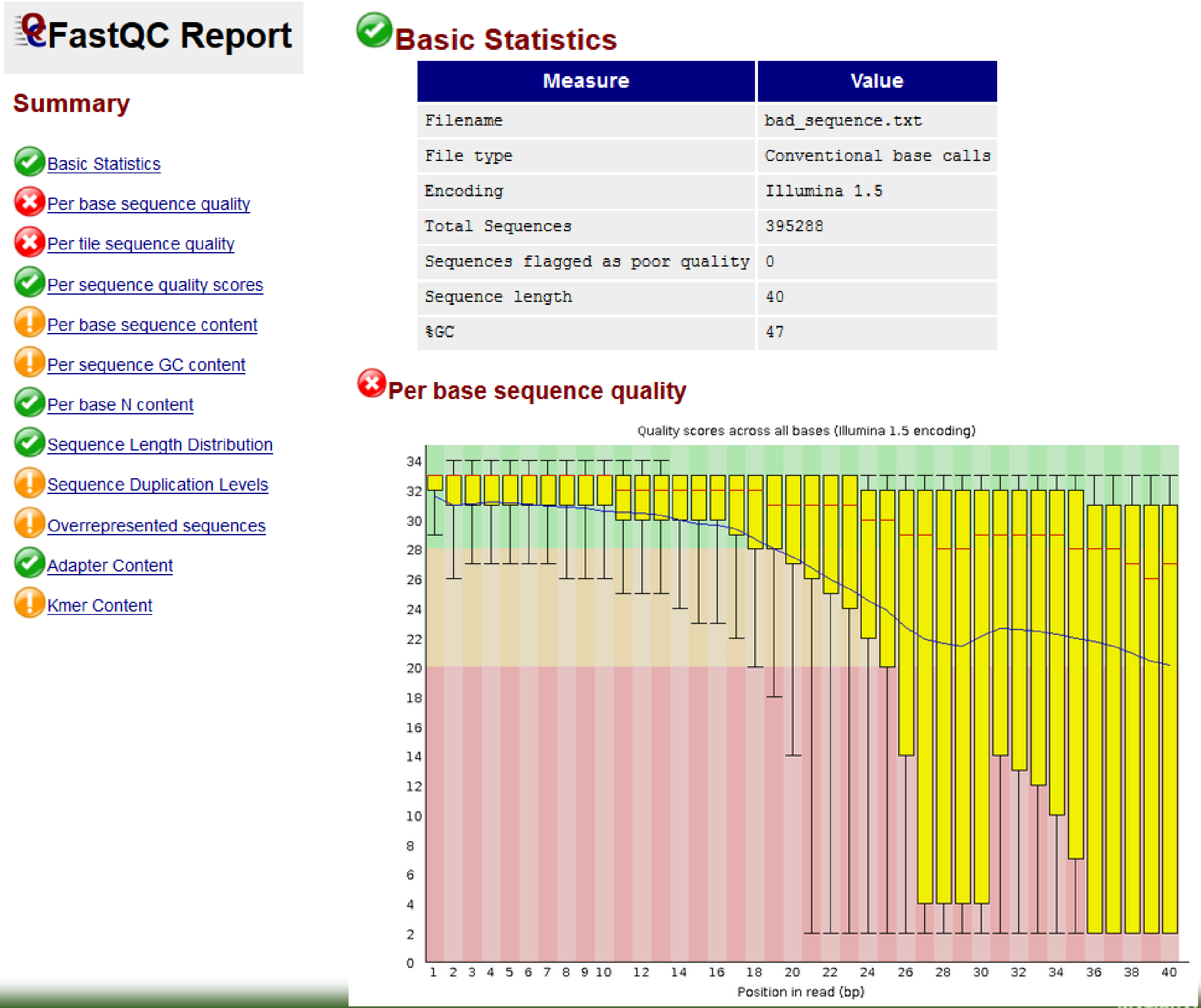
To run

```
$ fastqc ~/data/fastq/TB*.fastq  -o ~/result/fastqc_result/
```

There are 2 file “.html” + “.zip” for each fastq (8 files in total)
See the result of each read (R1 and R2) of both strains in “.html file”
Is that OK?

Note: You can compare the fastqc result before vs after trimming (do it later yourself)

Quality Control: Bad Illumina Data



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

✓ 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

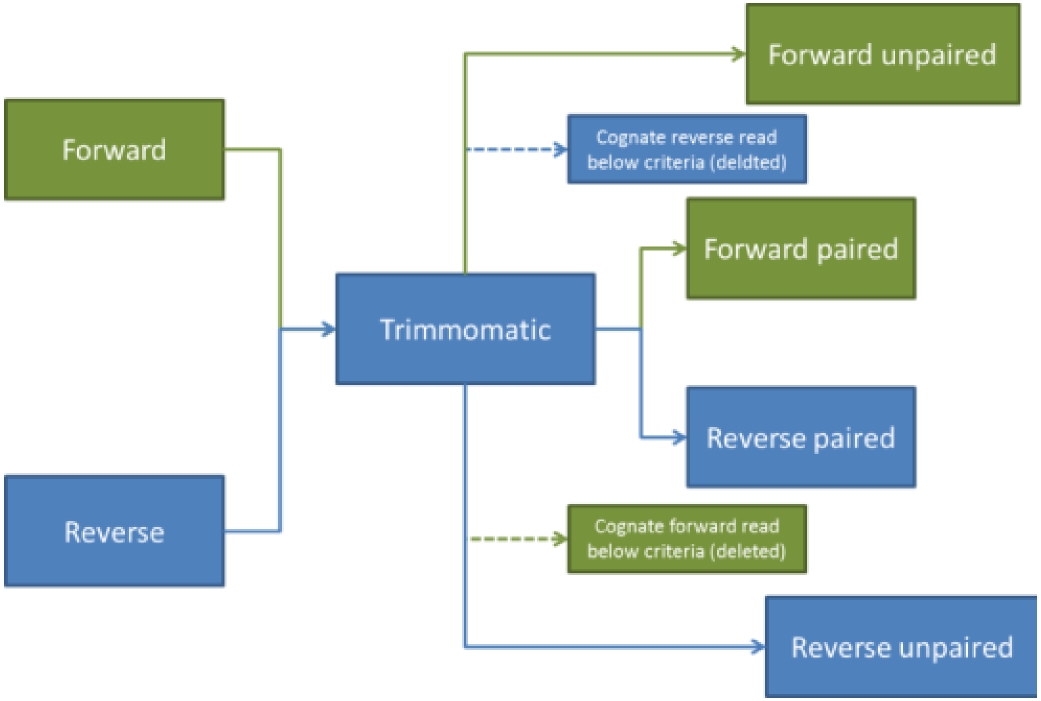
✗ Per base sequence quality

2 Trimming: Trimmomatic

http://www.usadellab.org/cms/index.php?page=trimmomatic

- LEADING:3 = cut the base at the start of read when the score below 3
- TRAILING:3 = cut the base at the end of read when the score below 3
- SLIDINGWINDOW:4:15 = cut sliding window of 4 bps when the average score below 15
- MINLEN:75 = Exclude read below 75 bp
- TOPHRED33/64: for changing the offset of the quality score to the preferred format
- HEADCROP: cut specific length of start of read
- CROP: 230 cut the read to specific length
- ILLUMINACLIP: exclude the adaptor

Stringency is matter for the quality (but some data will be lost)


Forward
Reverse
Trimmomatic
Forward unpaired
Cognate reverse read below criteria (deldted)
Forward paired
Reverse paired
Cognate forward read below criteria (deleted)
Reverse unpaired

Flow of reads in Trimmomatic Paired End mode

To run (for TB1_259 strain)

```
$ java -jar /usr/local/bin/trimmomatic-0.35.jar PE -phred33
~/data/fastq/TB1_259_R1.fastq ~/data/fastq/TB1_259_R2.fastq
~/result/trimming_result/TB1_259_pair_R1.fastq ~/result/trimming_result/TB1_259_unpair_R1.fastq
~/result/trimming_result/TB1_259_pair_R2.fastq ~/result/trimming_result/TB1_259_unpair_R2.fastq
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:75
```

See the result = 4 files for 1 strain: 2 paired + 2 unpaired reads

Do for the 2nd strain: TB2_260 strain

```
$ java -jar /usr/local/bin/trimmomatic-0.35.jar PE -phred33
~/data/fastq/TB2_260_R1.fastq ~/data/fastq/TB2_260_R2.fastq
~/result/trimming_result/TB2_260_pair_R1.fastq ~/result/trimming_result/TB2_260_unpair_R1.fastq
~/result/trimming_result/TB2_260_pair_R2.fastq ~/result/trimming_result/TB2_260_unpair_R2.fastq
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:75
```

See the result = 8 files of the 2 strains
You can see that >90% (see output message) of the paired reads are survived

So, you can select only paired reads for the downstream analysis

Note: You can do batch (many strain on one run), you can find the option to do it later

Commonly Used “short-read” DNA Mappers

- BWA
  - http://bio-bwa.sourceforge.net/
- Bowtie2
  - http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
- MAQ
  - http://maq.sourceforge.net/
- SMALT
  - https://www.sanger.ac.uk/resources/software/smalt/
- NovoAlign
  - http://www.novocraft.com/main/page.php?s=novoalign

3 Mapping: BWA MEM

http://bio-bwa.sourceforge.net/

3.1 Indexing the reference strain (for later mapping with tested stains)

Skipped

```
$ bwa index ~/ref/refBWA/h37rv_sequence.fasta
```

5 additional files will be added

However, h37Rv reference was indexed (**No need to do this**)

Note: Bowtie2 is another popular software for mapping but
– BWA: (BWT) is fast, detect small indels and good sensitivity

3.2 mapping the trimmed read to the indexed H37Rv reference)

```
$ bwa mem ~/ref/refBWA/h37rv_sequence.fasta
~/result/trimming_result/TB1_259_pair_R1.fastq
~/result/trimming_result/TB1_259_pair_R2.fastq
> ~/result/mapping_result/TB1_259.sam
-R '@RG\tID:TB1_259\tSM:Mtb\tSW:bwa'
```

-R: option to add the read group name

Wait for a while, it take time for mapping

You will get TB1_259.sam in your result folder

Do again for TB2_260 strain

```
$ bwa mem ~/ref/refBWA/h37rv_sequence.fasta
~/result/trimming_result/TB2_260_pair_R1.fastq
~/result/trimming_result/TB2_260_pair_R2.fastq
> ~/result/mapping_result/TB2_260.sam
-R '@RG\tID:TB2_260\tSM:Mtb\tSW:bwa'
```

Now you get 2 SAM file for 2 strains

7

Practice in NGS analysis II

4 Re-organization of mapped reads

4.1 SAM to BAM conversion (to save space)

```
$ samtools view -bS ~/result/mapping_result/TB1_259.sam
-o ~/result/mapping_result/TB1_259.bam
```

- Convert into the compressed files (binary format)
- Compressed around 600 Mb (for Mtb) into 180 Mb, per isolate
- For downstream analysis, you can delete SAM and keep BAM (to save space)

Do again for TB2_260 strain

```
$ samtools view -bS ~/result/mapping_result/TB2_260.sam
-o ~/result/mapping_result/TB2_260.bam
```

4.2 Sorting the BAM files

```
$ samtools sort ~/result/mapping_result/TB1_259.bam
~/result/mapping_result/TB1_259.sort
```

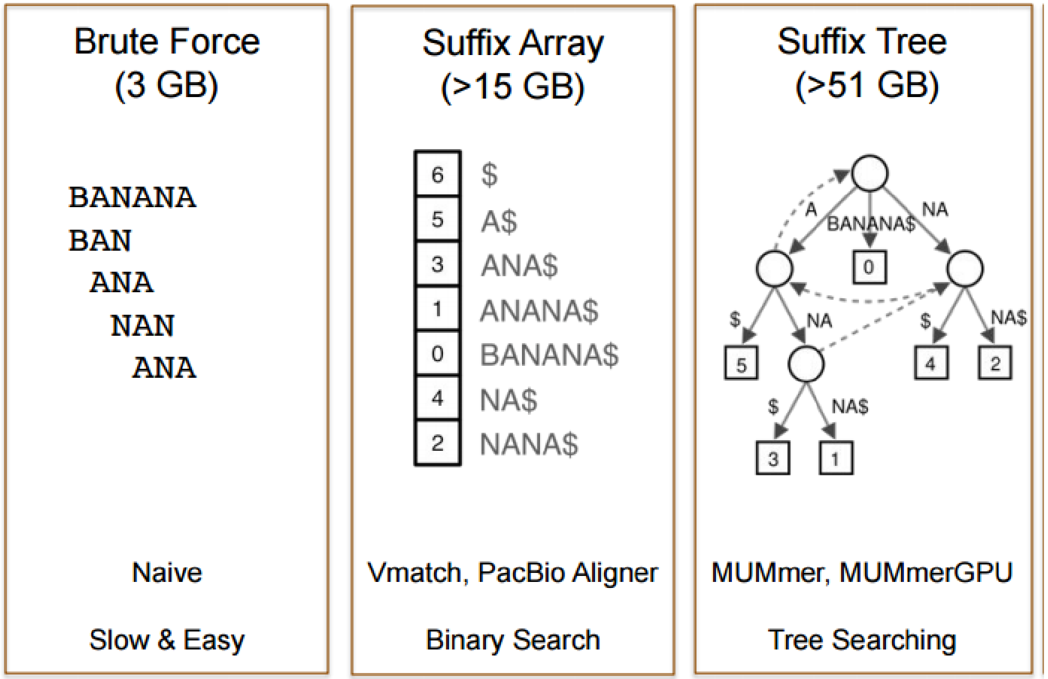
- The output will be ".sort.bam"

Do again for TB2_260 strain

```
$ samtools sort ~/result/mapping_result/TB2_260.bam
~/result/mapping_result/TB2_260.sort
```

Definitions

Indexing = preprocessing data for faster access e.g. suffix array/tree
Sorting = sort according to the genomic position
Mapping = map to the reference
Mapping software = BWA (MEM), BOWTIE2
Variant calling = GATK, Samtools



Brute Force
(3 GB)

```
BANANA
BAN
 ANA
  NAN
   ANA
```

Naive

Slow & Easy

Suffix Array
(>15 GB)

6	$
5	A$
3	ANA$
1	ANANA$
0	BANANA$
4	NA$
2	NANA$

Vmatch, PacBio Aligner

Binary Search

Suffix Tree
(>51 GB)

MUMmer, MUMmerGPU

Tree Searching

4.3 Indexing the sorted BAM file

```
$ samtools index ~/result/mapping_result/TB1_259.sort.bam
```

- The output will be “.sort.bam.bai”

Do again for TB2_260 strain

```
$ samtools index ~/result/mapping_result/TB2_260.sort.bam
```

5 Realignment

5.1 Indexing the reference strain using samtool is required for GATK Realigner (pre-indexed, so skip this step)

Skipped

additional files will be added

However, h37Rv reference was indexed by samtools (**No need to do this**)

h37rv_sequence.fai

```
samtools faidx ~/ref/refSamtool/h37rv_sequence.fasta
```

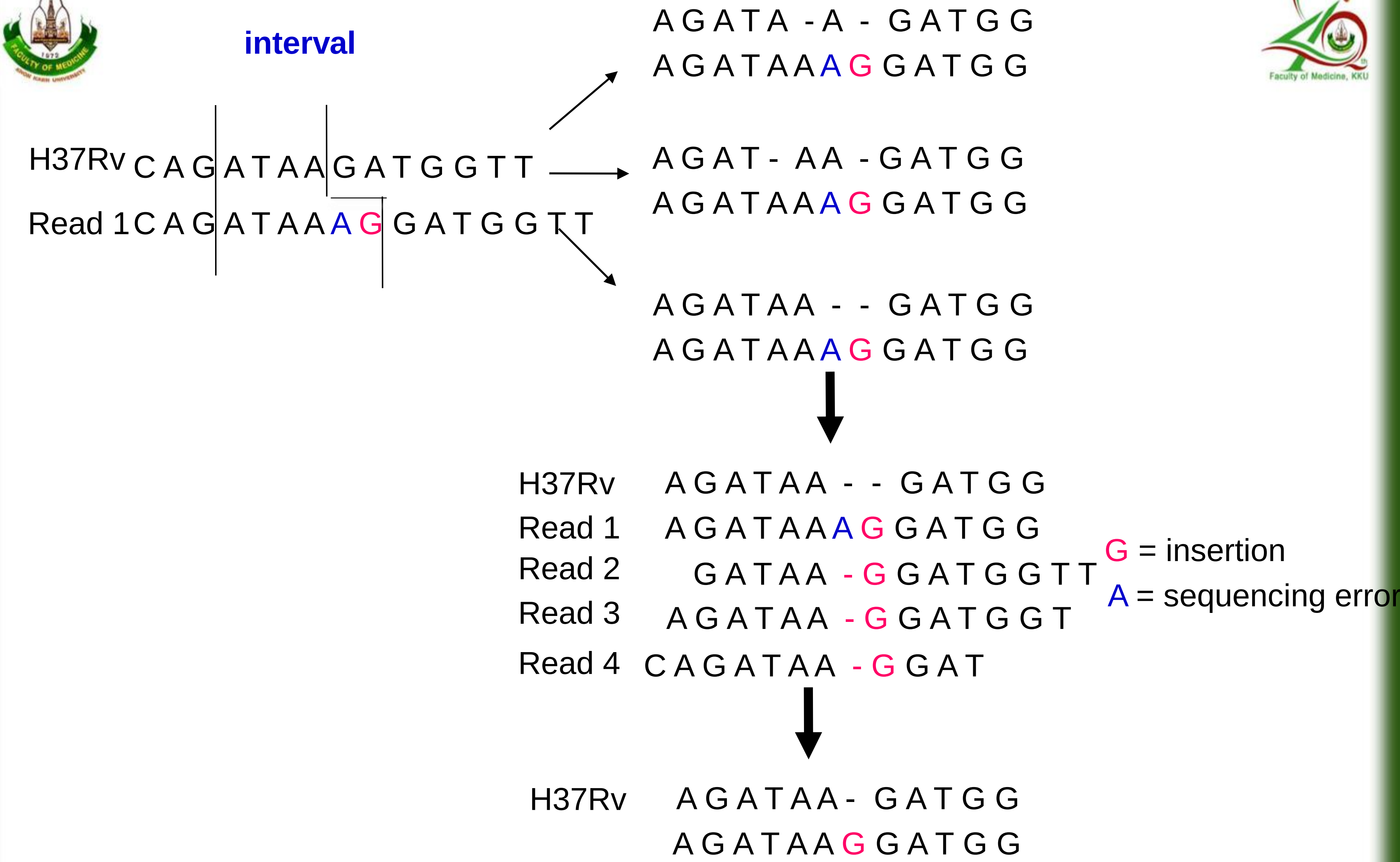
h37rv_sequence.dict

```
picard CreateSequenceDictionary  R= ~/ref/refSamtool/h37rv_sequence.fasta  O=
h37rv_sequence.fasta.dict
```

Meaning; A sequence dictionary contains the sequence name, sequence length, genome assembly identifier, and other information about sequences.

Local Realignment around Indels

- The algorithms that are used in the initial mapping step tend to produce various types of artifacts. For example, reads that align on the edges of indels often get mapped with mismatching bases that might look like evidence for SNPs, but are actually mapping artifacts. The realignment process identifies the most consistent placement of the reads relative to the indel in order to clean up these artifacts. It occurs in two steps: first the program identifies intervals that need to be realigned, then in the second step it determines the optimal consensus sequence and performs the actual realignment of reads.



interval

```
H37Rv  CAGATAAGATGGTT
Read 1 CAGATAAAGGATGGTT
```

```
AGATA-A-GATGG
AGATAAAGGATGG

AGAT-AA-GATGG
AGATAAAGGATGG

AGATAA--GATGG
AGATAAAGGATGG
```

```
H37Rv   AGATAA--GATGG
Read 1  AGATAAAGGATGG
Read 2    GATAA-GGATGGTT
Read 3  AGATAA-GGATGGT
Read 4 CAGATAA-GGAT
```

G = insertion

A = sequencing error

```
H37Rv  AGATAA-GATGG
       AGATAAGGATGG
```

Consensus calling: from the reads (multiple sequence realignment) according to coverage of that particular interval

5.2 Create Realigner Target (RealignerTargetCreator)

```
$ java -Xmx4g -jar ~/program/GenomeAnalysisTK.jar
-T RealignerTargetCreator -R ~/ref/refSamtool/h37rv_sequence.fasta
-I ~/result/mapping_result/TB1_259.sort.bam
-o ~/result/mapping_result/TB1_259.sort.bam.intervals
```

- The output will be “bam.intervals”

Do again for TB2_260 strain

```
$ java -Xmx4g -jar ~/program/GenomeAnalysisTK.jar
-T RealignerTargetCreator –R ~/ref/refSamtool/h37rv_sequence.fasta
-I ~/result/mapping_result/TB2_260.sort.bam
-o ~/result/mapping_result/TB2_260.sort.bam.intervals
```

5.3 to do realignment (IndelRealigner)

```
$ java -Xmx4g -jar ~/program/GenomeAnalysisTK.jar -T IndelRealigner
-R ~/ref/refSamtool/h37rv_sequence.fasta
-I ~/result/mapping_result/TB1_259.sort.bam
-targetIntervals ~/result/mapping_result/TB1_259.sort.bam.intervals
-o ~/result/mapping_result/TB1_259.realn.bam
```

- The 2 output files will be “.realn.bam” + “.realn.bai”

Do again for TB2_260 strain

```
$ java -Xmx4g -jar ~/program/GenomeAnalysisTK.jar -T IndelRealigner
-R ~/ref/refSamtool/h37rv_sequence.fasta
-I ~/result/mapping_result/TB2_260.sort.bam
-targetIntervals ~/result/mapping_result/TB2_260.sort.bam.intervals
-o ~/result/mapping_result/TB2_260.realn.bam
```

6 Stat report

This step is just provide you the QC information of mapping, the output is not necessary for the downstream analysis

6.1 Coverage report by GATK

```
$ java -Xmx4g -jar ~/program/GenomeAnalysisTK.jar
-T DepthOfCoverage -R ~/ref/refSamtool/h37rv_sequence.fasta
-I ~/result/mapping_result/TB1_259.realn.bam
-o ~/result/stat_result/TB1_259.realn.bam.report
```

-This step might take time, e.g. > 5 min (if it take too long, you can do it back home)
-There are 7 report files, the ".bam.report" tell coverage or read depth of each position

Do again for TB2_260 strain

```
$ java -Xmx4g -jar ~/program/GenomeAnalysisTK.jar
-T DepthOfCoverage -R ~/ref/refSamtool/h37rv_sequence.fasta
-I ~/result/mapping_result/TB2_260.realn.bam
-o ~/result/stat_result/TB2_260.realn.bam.report
```

6.2 Flagstat report by SAMtools

```
$ samtools flagstat ~/result/mapping_result/TB1_259.realn.bam
> ~/result/stat_result/TB1_259.realn.bam.flagstat
```

-There is 1 output file: tell the % of mapped reads and mapped paired-reads

Do again for TB2_260 strain

```
$ samtools flagstat ~/result/mapping_result/TB2_260.realn.bam
> ~/result/stat_result/TB2_260.realn.bam.flagstat
```

7 Variant calling using SAMtools

7.1 variant calling using SAMtools

- You can use other tool (such as GATK) to call the variant
- The intersect set of variants between SAMtools and GATK might be used.
- For this practice, we will use only SAMtools for variant calling

```
$ samtools mpileup -B -Q 20 -d 2000 -C 50 -ugf
~/ref/refSamtool/h37rv_sequence.fasta
~/result/mapping_result/TB1_259.realn.bam | bcftools view -bvcg - >
~/result/calling_result/TB1_259.raw.bcf
```

- You can see the meaning of the option by "samtools mpileup", and then enter
- You can also see option of "bcftools view"
- The output = ".raw.bcf " file contain un-filtered variants in BCF (binary) format

Do again for TB2_260 strain

This step also take time!

```
$ samtools mpileup -B -Q 20 -d 2000 -C 50 -ugf
~/ref/refSamtool/h37rv_sequence.fasta
~/result/mapping_result/TB2_260.realn.bam | bcftools view -bvcg - >
~/result/calling_result/TB2_260.raw.bcf
```

7.2 BCF to VCF conversion

This step is just simply converse BCF to VCF file format
So, the output = un-filter variants in the VCF format

```
$ bcftools view ~/result/calling_result/TB1_259.raw.bcf >
~/result/calling_result/TB1_259.raw.vcf
```

Do again for TB2_260 strain

```
$ bcftools view ~/result/calling_result/TB2_260.raw.bcf >
~/result/calling_result/TB2_260.raw.vcf
```

7.3 Variant filtration

Then, we want to filter some low quality variants
Notably, the sensitivity to detect the variants might decrease, depending on the stringency

```
$ vcfutils.pl varFilter -d 10 -D 2000 -Q 20 ~/result/calling_result/TB1_259.raw.vcf >
~/result/calling_result/TB1_259.filt.vcf
```

- You can see the meaning of the option by “vcfutils.pl varFilter ”, and then enter
- Here, we exclude the variants with <20 mapping quality and <10 read depth

Do again for TB2_260 strain

```
$ vcfutils.pl varFilter -d 10 -D 2000 -Q 20 ~/result/calling_result/TB2_260.raw.vcf
> ~/result/calling_result/TB2_260.filt.vcf
```

8 Variant annotation using snpEff

- Too much detail of variant annotation using snpEFF can be covered in this practice, recommend to read the manual yourself
- The objective of this practice is get the concept to use the software
- snfEff required database (H37rv ref.), need to prepare this before running the program, please refer to the manual (the step is skipped, as it pre-installed in VM)

Not all SNPs you want to know the annotation,
so you can selectively see the annotation of SNP of interest,
e.g. after SNPs comparisons between strains

8.1 change the heading of the vcf file to compatible with the databases

- This step is the trouble shooting step. If you do not change the header of the vcf file, when you run the snpEff in step 2.8.2, the error message will pop up.
- If you face the trouble, You can find the suggestion in the discussion forum, google it!
- You can change the header manually but it is not practical.

Here we will use a short script to edit the header.

go to the location of your vcf file

```
$ cd ~/result/calling_result/
```

change the header of the vcf file

```
$ ls *.vcf | while read line; do sed 's/gi|448814763|ref|NC_000962.3|/NC_000962/'
$line > rewrite.$line; done
```

The output of the 2 strains = “rewrite………… filt.vcf”

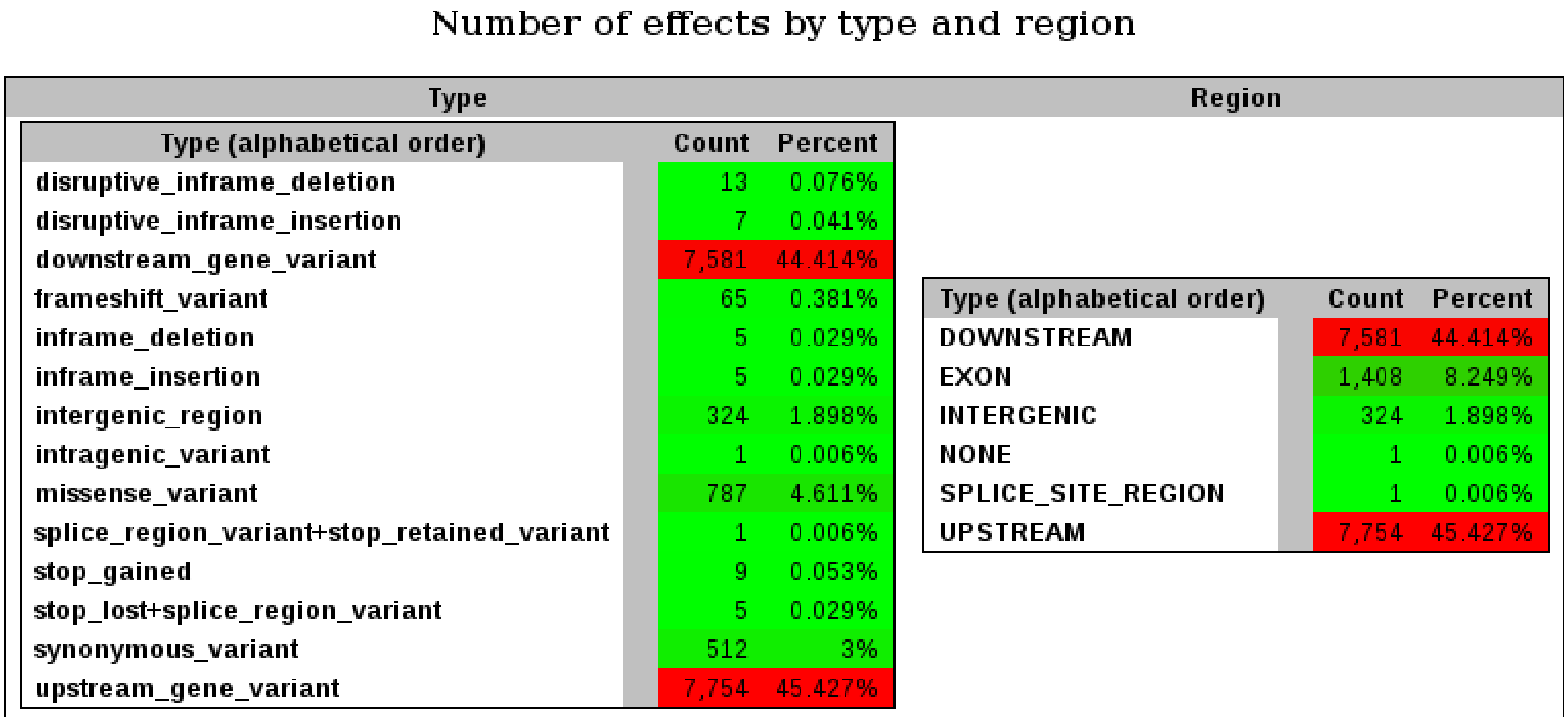
8.2 running the annotation using snfEFF

```
$ for A in ~/result/calling_result/rewrite*.vcf;
do
B=${A%.vcf}.ann.vcf;
C=${A%.vcf}.ann.html;
java -jar ~/program/snpEff/snpEff.jar m_tuberculosis_H37Rv $A > $B -s $C;
done
```

* This is the example of “loop command” to run multiple strains in one command

- There are 3 outputs = “.filt.ann.vcf” + “.html” + “.genes.txt”
- You can see the content inside the each of output files
- you can see the additional information from the annotated file (compare to the un-annotated filt.vcf one)
- Again, too much detail can be covered for the annotation output, please read the manual of snpEFF yourself.

Example output from SnpEff



Number of effects by type and region

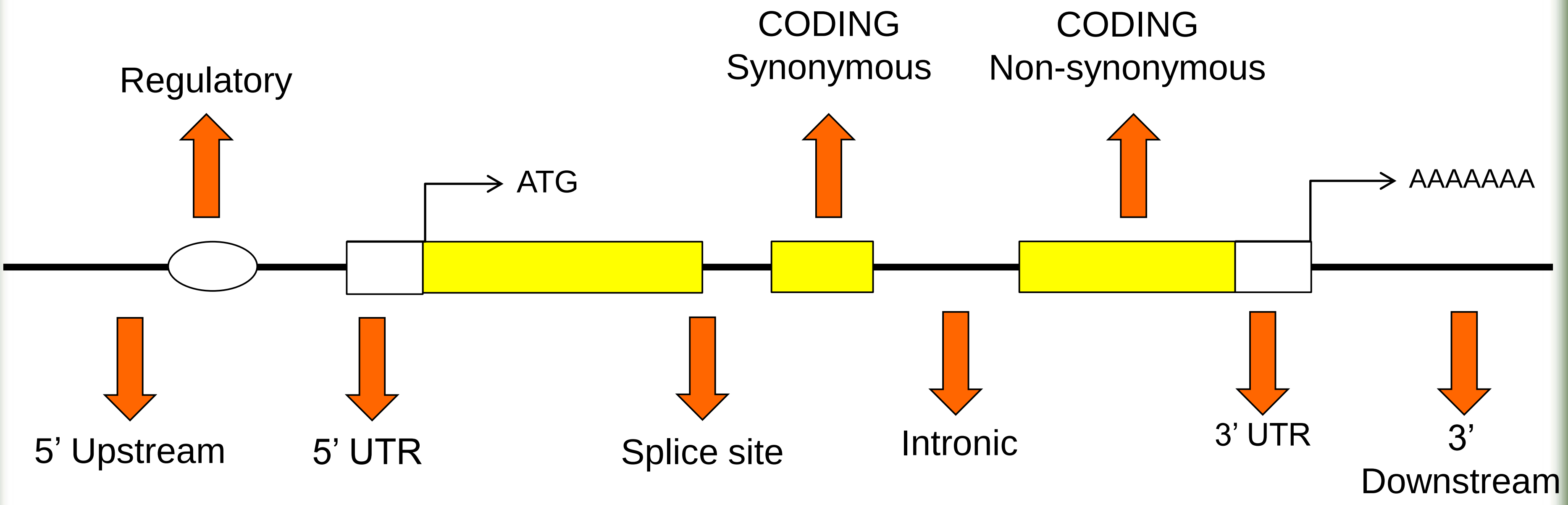
Type

Type (alphabetical order)	Count	Percent
disruptive_inframe_deletion	13	0.076%
disruptive_inframe_insertion	7	0.041%
downstream_gene_variant	7,581	44.414%
frameshift_variant	65	0.381%
inframe_deletion	5	0.029%
inframe_insertion	5	0.029%
intergenic_region	324	1.898%
intragenic_variant	1	0.006%
missense_variant	787	4.611%
splice_region_variant+stop_retained_variant	1	0.006%
stop_gained	9	0.053%
stop_lost+splice_region_variant	5	0.029%
synonymous_variant	512	3%
upstream_gene_variant	7,754	45.427%

Region

Type (alphabetical order)	Count	Percent
DOWNSTREAM	7,581	44.414%
EXON	1,408	8.249%
INTERGENIC	324	1.898%
NONE	1	0.006%
SPLICE_SITE_REGION	1	0.006%
UPSTREAM	7,754	45.427%

Variation annotation : functional consequences



Laura Clarke, 2013

9 Sideline analysis (variant comparison)

#9.1 Union of the 2 strains

```
$ java -Xmx2g -jar ~/program/GenomeAnalysisTK.jar -T CombineVariants
--genotypemergeoption UNIQUIFY -R ~/ref/refSamtool/h37rv_sequence.fasta
-V:TB1 ~/result/calling_result/TB1_259.filt.vcf
-V:TB2 ~/result/calling_result/TB2_260.filt.vcf
-o ~/result/sideline_result/unionTB1_TB2.filt.vcf
```

There are 2 output files = .vcf + .idx
You can see the union set of variants between the 2 strains from “.vcf” file

#9.2 Intersect of the 2 strains

```
$ java -Xmx2g -jar ~/program/GenomeAnalysisTK.jar -T SelectVariants -R
~/ref/refSamtool/h37rv_sequence.fasta -V:IntersectTB1_TB2
~/result/sideline_result/unionTB1_TB2.filt.vcf -select 'set == "Intersection";' -o
~/result/sideline_result/intersectTB1_TB2.filt.vcf
```

You can see the intersection set of variants between the 2 strains from “.vcf” file

#9.3 Unique for TB1_259 strain

```
$ java -Xmx2g -jar ~/program/GenomeAnalysisTK.jar -T SelectVariants
-R ~/ref/refSamtool/h37rv_sequence.fasta -V:uniqueTB1
~/result/sideline_result/unionTB1_TB2.filt.vcf -select 'set == "TB1";' -o
~/result/sideline_result/uniqueTB1.filt.vcf
```

#9.4 Unique for TB2_260 strain

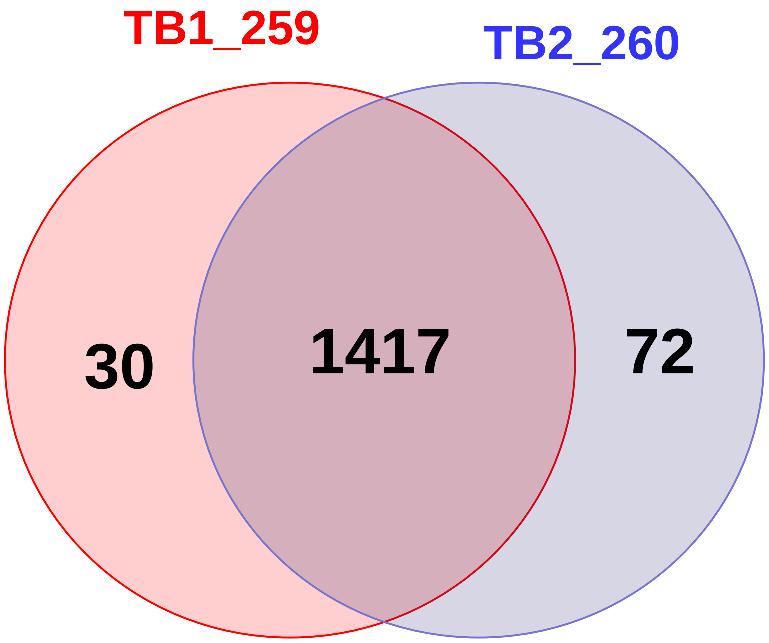
```
$ java -Xmx2g -jar ~/program/GenomeAnalysisTK.jar -T SelectVariants
-R ~/ref/refSamtool/h37rv_sequence.fasta -V:uniqueTB2
~/result/sideline_result/unionTB1_TB2.filt.vcf -select 'set == "TB2";' -o
~/result/sideline_result/uniqueTB2.filt.vcf
```

To count the number of variant (line) you can use the command below

```
$ grep -vc '^$' ~/result/calling_result/rewrite.TB1_259.filt.vcf
```

This is count all lines, so excluding the header line = number of variant lines

Venn diagram comparing the number of variants between the 2 strains

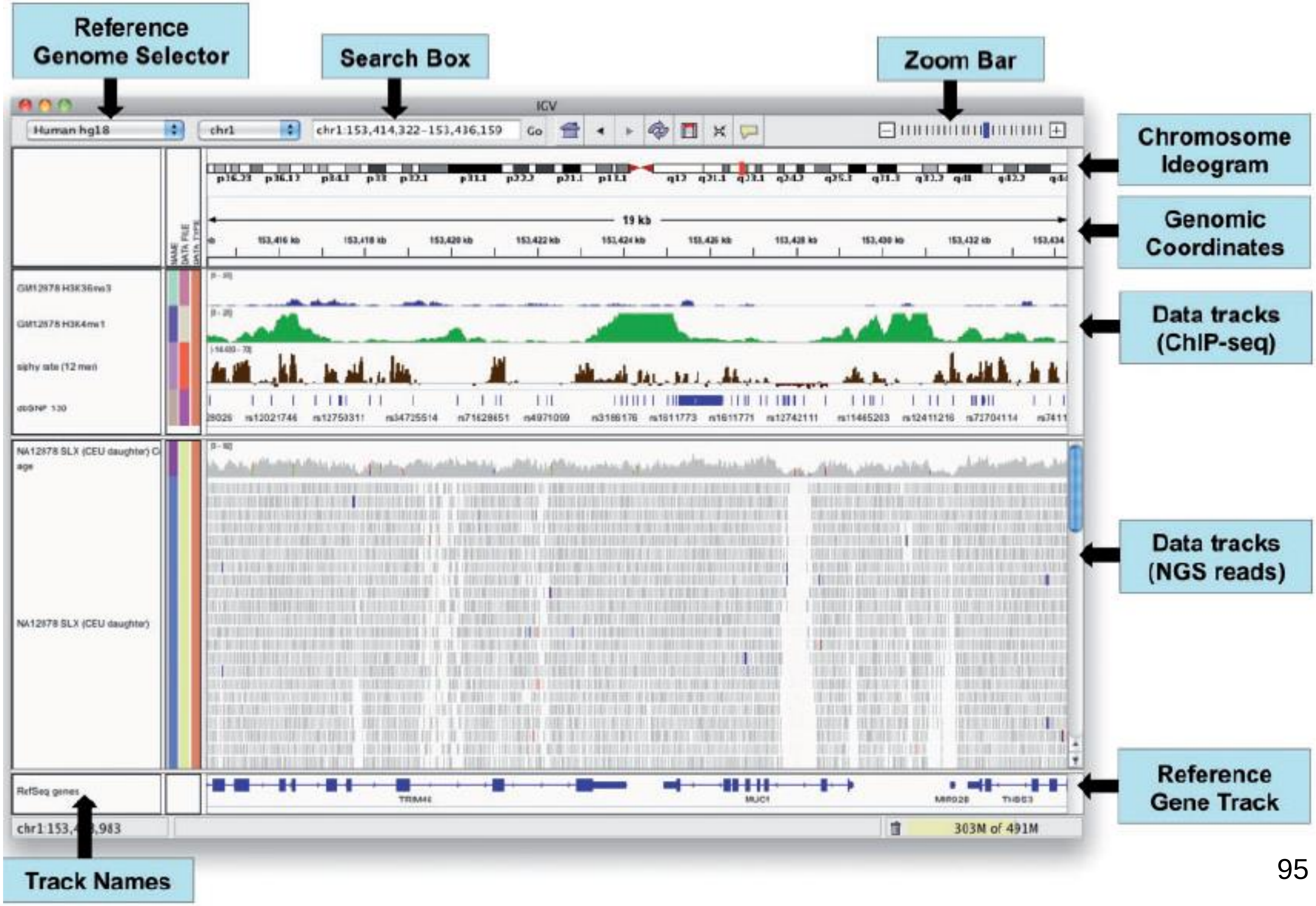


- Number of variant include both SNPs and Indel
- The number of variant can be further filtered, e.g. excluding of
    - SNPs with < 20% of read depth to support
    - heterozygous SNPs with allelic frequencies <75 %

Visualization

- Integrative Genomics Viewer (IGV)
  - http://www.broadinstitute.org/igv/
- Artemis/ACT
  - http://www.sanger.ac.uk/resources/software/artemistview (samtools)
- BAMView
  - http://bamview.sourceforge.net/
- Tablet
  - http://bioinf.scri.ac.uk/tablet/

Visualization - IGV



Reference Genome Selector

Search Box

Zoom Bar

Chromosome Ideogram

Genomic Coordinates

Data tracks (ChIP-seq)

Data tracks (NGS reads)

Reference Gene Track

Track Names

BRIEFINGS IN BIOINFORMATICS. VOL 14. NO 2. 178 ^192

8

Assignment

Assignments

4 genome of *M. tuberculosis:* 6 pairs

1. Mtb 1 versus Mtb 2
2. Mtb 1 versus Mtb 3
3. Mtb 1 versus Mtb 4
4. Mtb 2 versus Mtb 3
5. Mtb 2 versus Mtb 4
6. Mtb 3 versus Mtb 4

Each pair match to the successive number of student ID from the registration
Copy from your friend will be punished by 50% subtraction of the actual score.

Please copy only 4 files (~300 MB) corresponding to your assignment

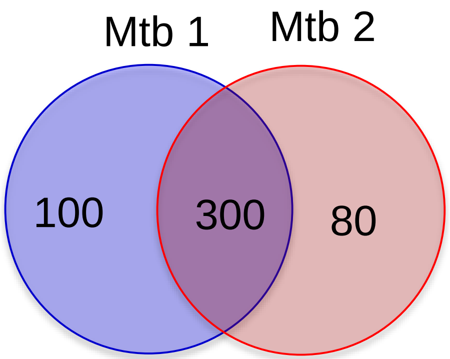
From each comparative analysis of 2 strains you received

- Submit the code for each step
- Submit the Venn diagram describing the number of SNP from the comparative analysis and Stat of each genome such as to total read count, read depth, mapped read etc.
- OPTION: Submit the function of the unique SNPs from each strain

one code txt file (file name = Mtb1 vs Mtb 2_your Name, e.g. 1vs2_code_Orawee Kaewprasert)
one MS Word for Venn diagram (file name, similar = 1vs2_Venn_Orawee Kaewprasert)
Option for description of function of unique SNP (annotated SNPs)

Within 2 weeks after the lectures

1vs2_Venn_Orawee Kaewprasert



Stats	Mtb1	Mtb2
Total read count of raw fastq file		
%Mapped reads		
Read depth Etc.		

1vs2_code_Orawee Kaewprasert

Code file

Thank you for your attention