AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

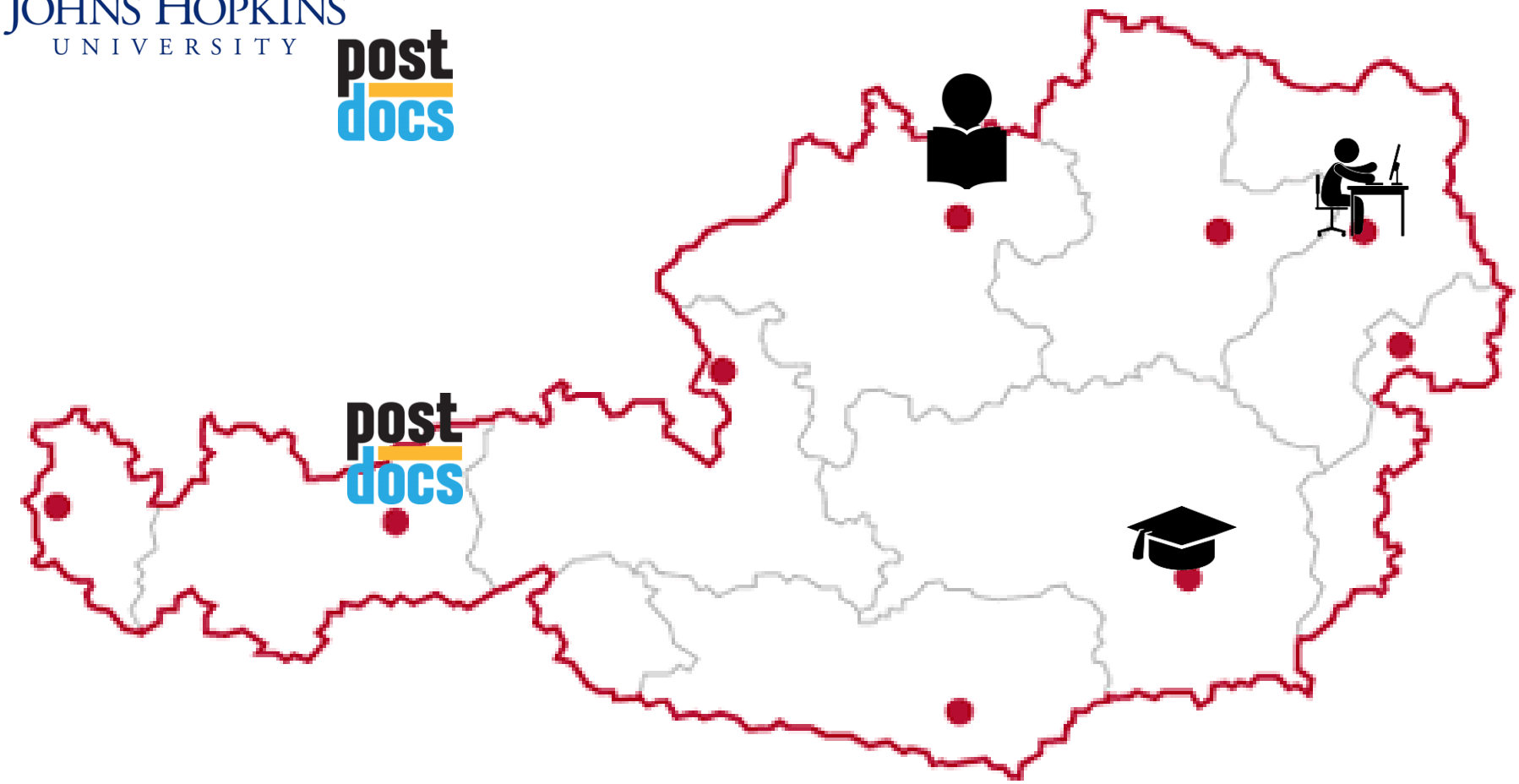
Bioinformatics and Genome Analyses Course

Lecture 1 (Nov 12)

Stephan Pabinger
stephan.pabinger@ait.ac.at

http://pabinger.site44.com

My background



My background

Bioinformatics

- Software development
- Web tools
- Pipeline design

Working with sequencing data

- DNASeq
- RNASeq
- MethylationSeq

Data analysis

- DNA data
- Protein data
- Peptide data
- Immunomics data

What to expect

- Finish with an understanding of major concepts and tools
- Know how to perform variant calling
- Ready to make informed choices about what kind of variant calling tools you may need
- Focus is on Variant Calling and Annotation
  - Alignment refinement
  - Alignment metrics reports
  - Variant calling
  - Variant annotation and filtering
  - Visualization

Resources

Information about practicals can be found here

https://github.com/spabinger/BioinformaticsAndGenomeAnalyses2018

Please make sure that you can access the resource

GIT

GIT

Information

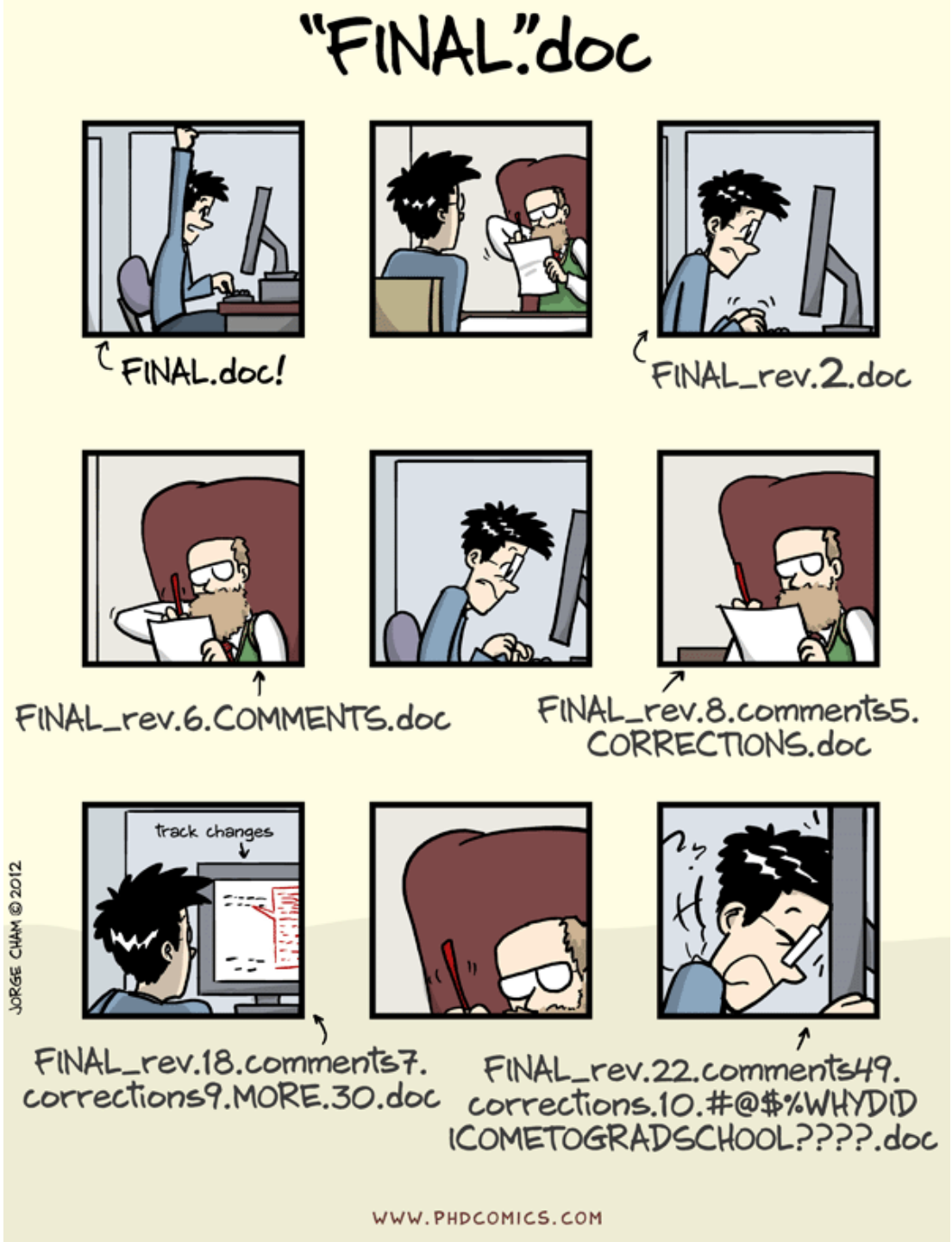
- Distributed revision control system
- Developed by Linus Torvalds (Linux developer)
- Local repositories & remote repositories

Keep track of changes

- Code
- Manuscript
- Presentations
- Data analysis

Master/PhD thesis

Merging collaborators' changes


"FINAL".doc
FINAL.doc!
FINAL_rev.2.doc
FINAL_rev.6.COMMENTS.doc
FINAL_rev.8.comments5.
CORRECTIONS.doc
track changes
FINAL_rev.18.comments7.
corrections9.MORE.30.doc
FINAL_rev.22.comments49.
corrections.10.#@$%WHYDID
ICOMETOGRADSCHOOL????.doc
JORGE CHAM © 2012
WWW.PHDCOMICS.COM

Start with GIT

- Create a new directory
- Open it (cd into it)
- Perform
  `git init`

- Work ….

- Add files you want to store in the repository (locally)
  `git add XY.txt`
  `git add *` (to add everything)

- Commit files
  `git commit –m „Performed first analysis“`

Start with GIT

```
stephan@shaq /tmp $ mkdir super_analysis
stephan@shaq /tmp $ cd super_analysis/
stephan@shaq /tmp/super_analysis $ git init
Initialized empty Git repository in /tmp/super_analysis/.git/
stephan@shaq /tmp/super_analysis $ touch XY.txt
stephan@shaq /tmp/super_analysis $ touch rawfile.txt
stephan@shaq /tmp/super_analysis $ git add XY.txt
stephan@shaq /tmp/super_analysis $ git add *
stephan@shaq /tmp/super_analysis $ git commit -m "Performed first analysis"
[master (root-commit) 915d159] Performed first analysis
 2 files changed, 0 insertions(+), 0 deletions(-)
 create mode 100644 XY.txt
 create mode 100644 rawfile.txt
stephan@shaq /tmp/super_analysis $
```

Features

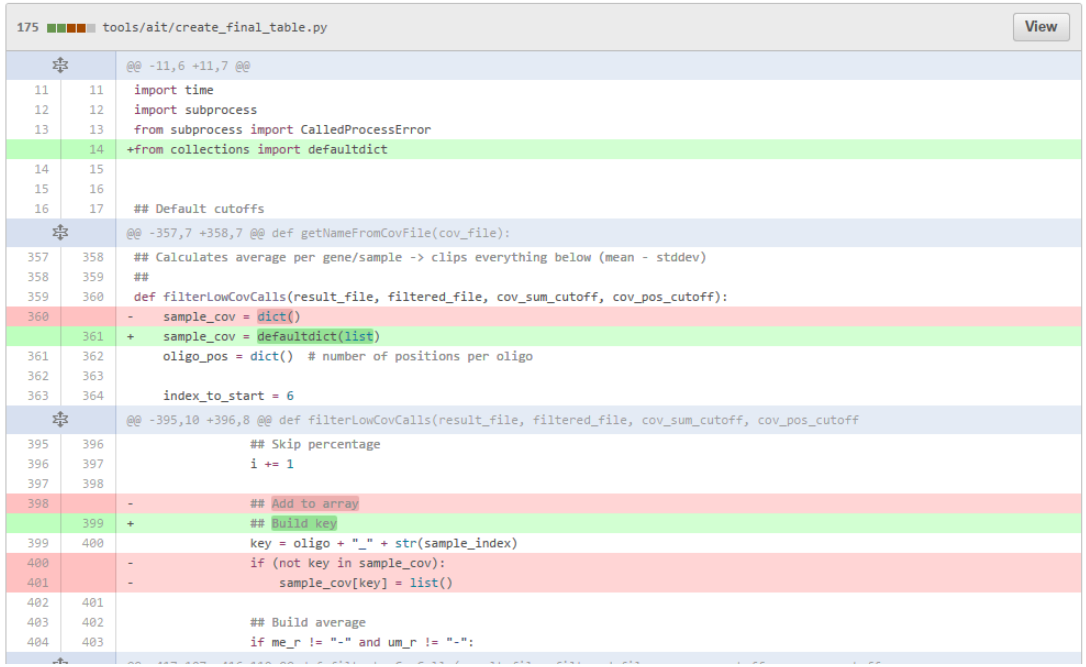
- Status of your project
  `git status`

```
stephan@shaq /tmp/super_analysis $ git status
On branch master
Untracked files:
  (use "git add <file>..." to include in what will be committed)

        nextanalysis.py

nothing added to commit but untracked files present (use "git add" to track)
stephan@shaq /tmp/super_analysis $
```

- History
- Go to a previous version
- Branching (implement a new feature without disrupting main code)
- Merging between different versions/branches



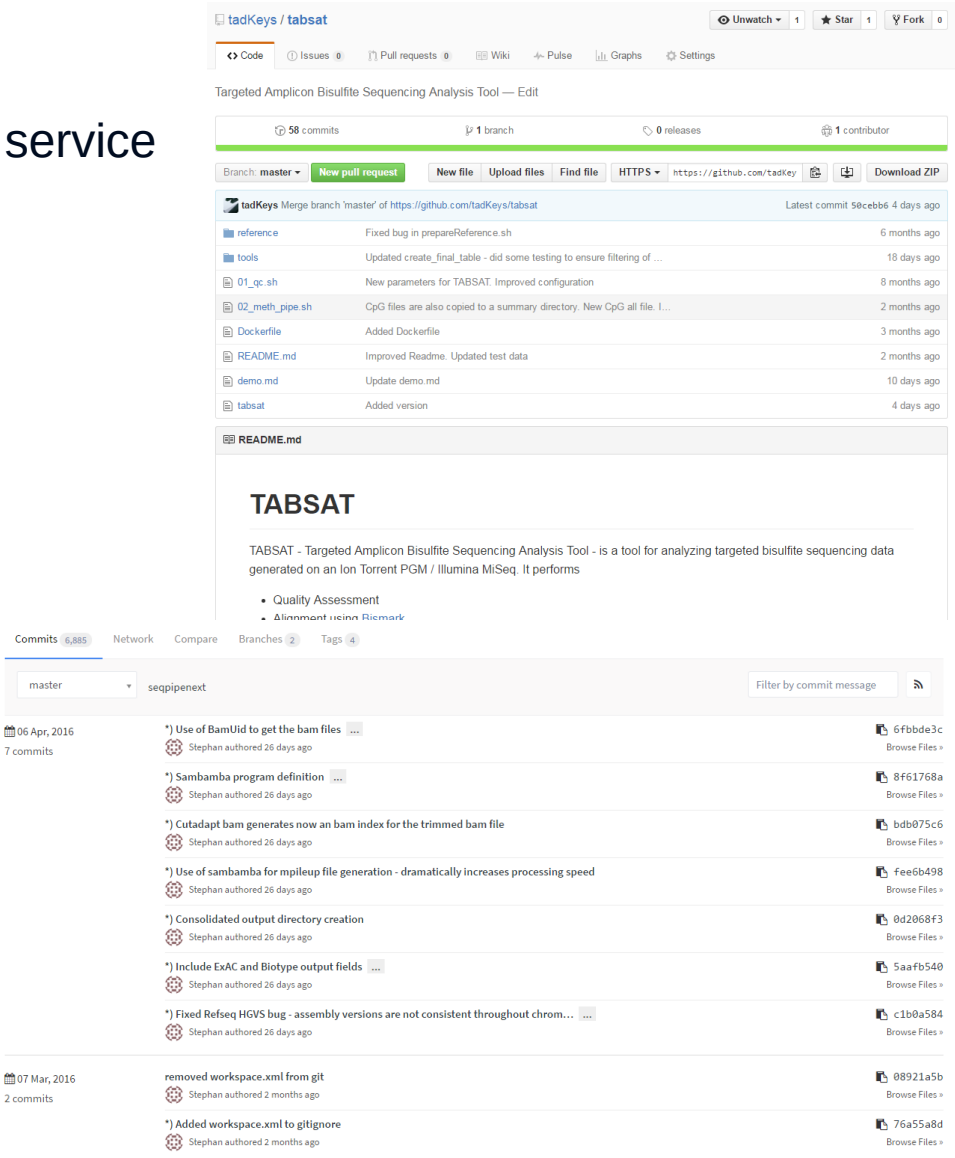
Github - Gitlab

Github

- Web-based Git repository hosting service
- Free to use
- Request for private repositories

Gitlab

- Community version
- Open Source
- Share code, analyses, etc
- Easy to transfer to Github



Remote repositories

- Clone repository
  ```
  git clone username@host:/path/to/repository
  ```

- Push files to remote repository
  ```
  git push
  ```

GIT - terms

- Committed means that the data is safely stored in your local database.
- Modified means that you have changed the file but have not committed it to your database yet.
- Staged means that you have marked a modified file in its current version to go into your next commit snapshot.

GIT – Hints- Resources

- Make lots of commits
- Don‘t commit large files (they should be on the server)

Information

http://rogerdudler.github.io/git-guide/

http://kbroman.org/github_tutorial/

http://nyuccl.org/pages/GitTutorial/

https://swcarpentry.github.io/git-novice/

Windows

- https://git-for-windows.github.io/
  → Graphical user-interface (for init, add, commit, push, compare)

Linux & Mac

- Packages and GUIs available

History

History of Molecular Diagnostics | Sequencing | and more

The Molecular Biology Timeline

1865	Gregor Mendel, Law of Heredity
1866	Johann Miescher, Purification of DNA
1949	Sickle Cell Anemia Mutation

Sickle Cell Anemia

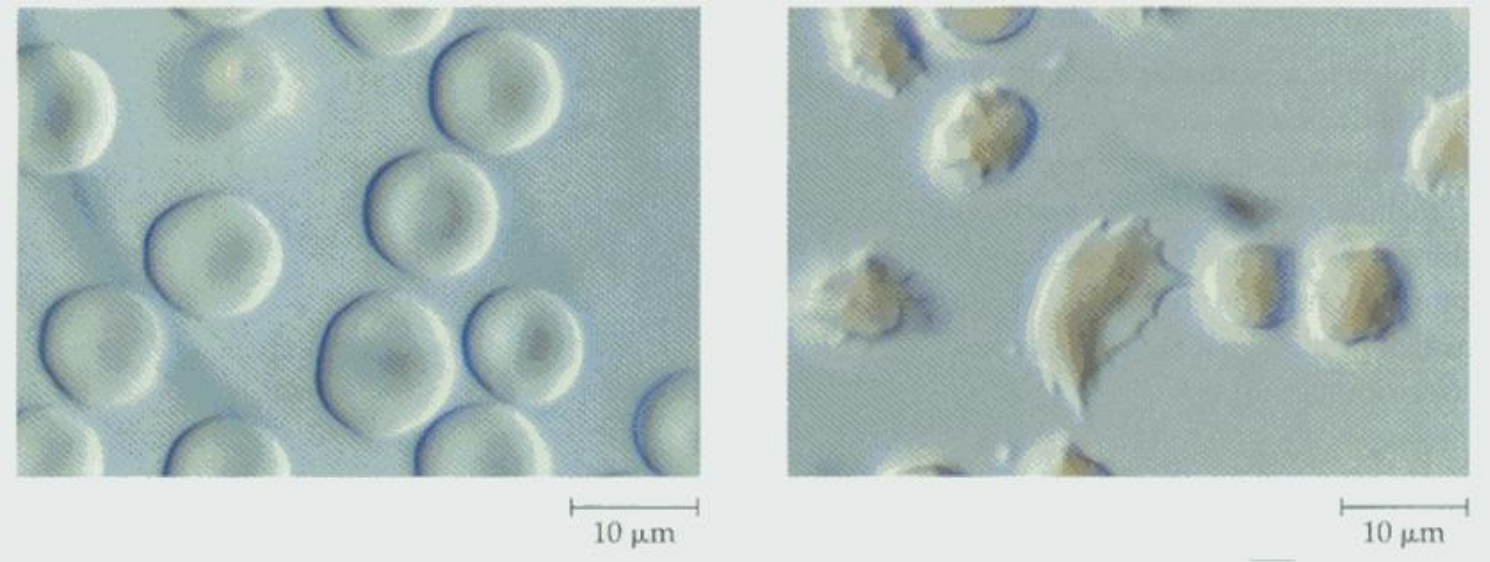
Deformation of red blood cells

Epidemiology

1 out of 500 with African heritage; also common in Asia and Mediterranean region

Genetics

Single point mutation (SNP) → 6. Codon of the β-globin Gens



The Molecular Biology Timeline

1865 Gregor Mendel, Law of Heredity

1866 Johann Miescher, Purification of DNA

1949 Sickle Cell Anemia Mutation

1953 Watson and Crick, Structure of DNA

Discovery of DNA Structure

James Watson & Francis Crick: **DNA structure** - 1953

- First correct double-helix model of DNA structure
- Based on one X-ray diffraction image taken by Rosalind Franklin and Raymond Gosling

equipment, and to Dr. G. E. R. Deacon and the captain and officers of R.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerrard, H., and Jevons, W., *Phil. Mag.*, **40**, 149 (1920).
[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, **5**, 285 (1949).
[3] Von Arx, W. S., Woods Hole Papers in Phys. Oceanog., Meteor., **11** (3) (1956).
[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik. (Stockholm)*, **2** (11) (1905).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

This figure is purely diagrammatic. The two ribbons symbolize the two phosphate–sugar chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the *z*-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical *z*-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[3,4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

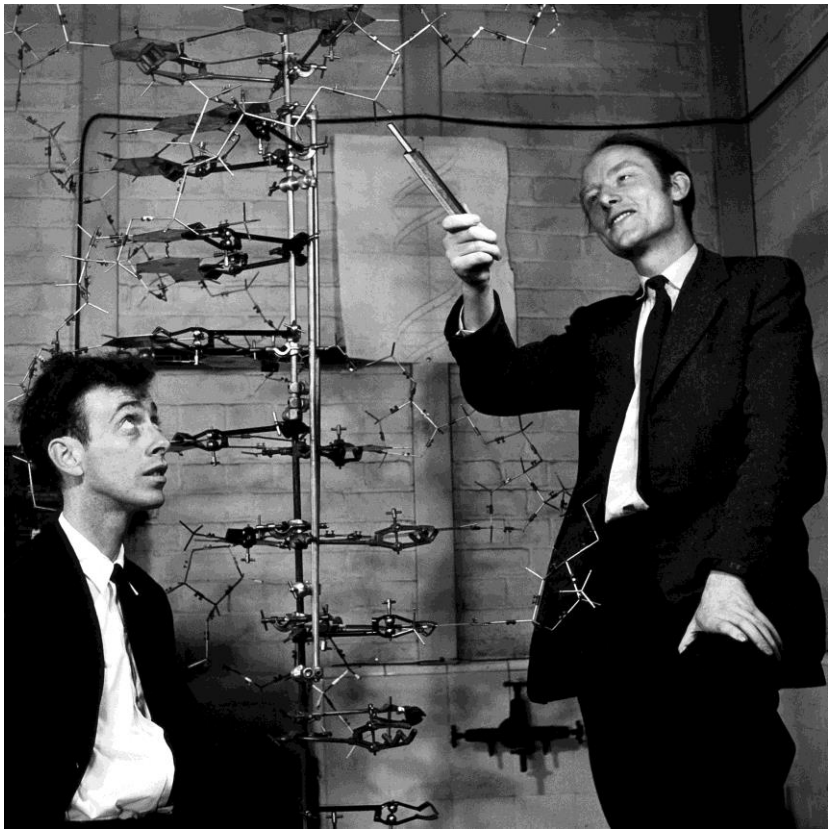
It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at



The Molecular Biology Timeline

1865	Gregor Mendel, Law of Heredity
1866	Johann Miescher, Purification of DNA
1949	Sickle Cell Anemia Mutation
1953	Watson and Crick, Structure of DNA
1970	Recombinant DNA Technology

The Molecular Biology Timeline

1865	Gregor Mendel, Law of Heredity
1866	Johann Miescher, Purification of DNA
1949	Sickle Cell Anemia Mutation
1953	Watson and Crick, Structure of DNA
1970	Recombinant DNA Technology
1977	DNA sequencing

DNA Sequencing - mid 1970s

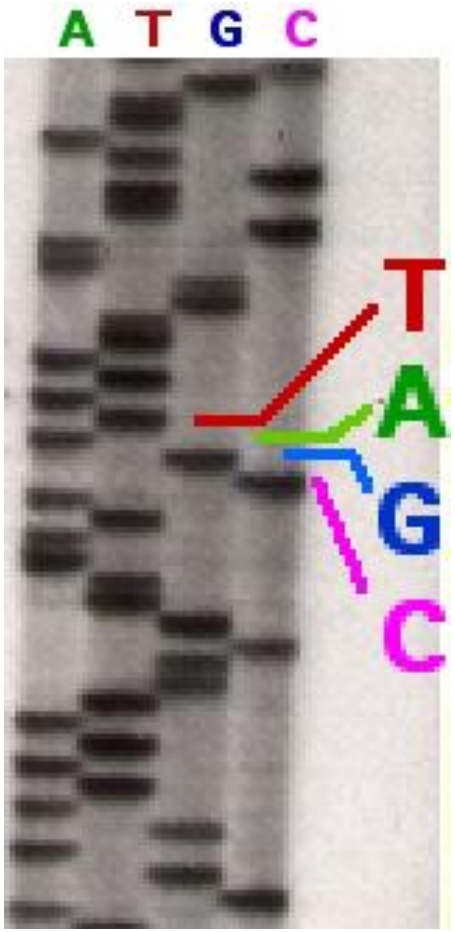
Mutations/Variations can be detected using DNA sequencing

Sanger Sequencing

- Dideoxy DNA sequencing paired with gel electrophoresis
- DNA is 5’ labeled with radioactivity
- Small amount of Dideoxy base added to 4 separate primer extension reactions
- Run on a gel to determine bases at each position by size
- Still considered the gold standard for validating sequencing data

Maxam-Gilbert Sequencing

- Chemical modification and cleavage paired with gel electrophoresis
- DNA is 5’ labeled with radioactivity (gamma-P ATP)
- Exposed to chemical agents that cause specific DNA breaks
- Run on a gel and the pattern reveals which base is at each site



Sanger

The Molecular Biology Timeline

1865	Gregor Mendel, Law of Heredity
1866	Johann Miescher, Purification of DNA
1949	Sickle Cell Anemia Mutation
1953	Watson and Crick, Structure of DNA
1970	Recombinant DNA Technology
1977	DNA sequencing
1985	*In Vitro* Amplification of DNA (PCR)

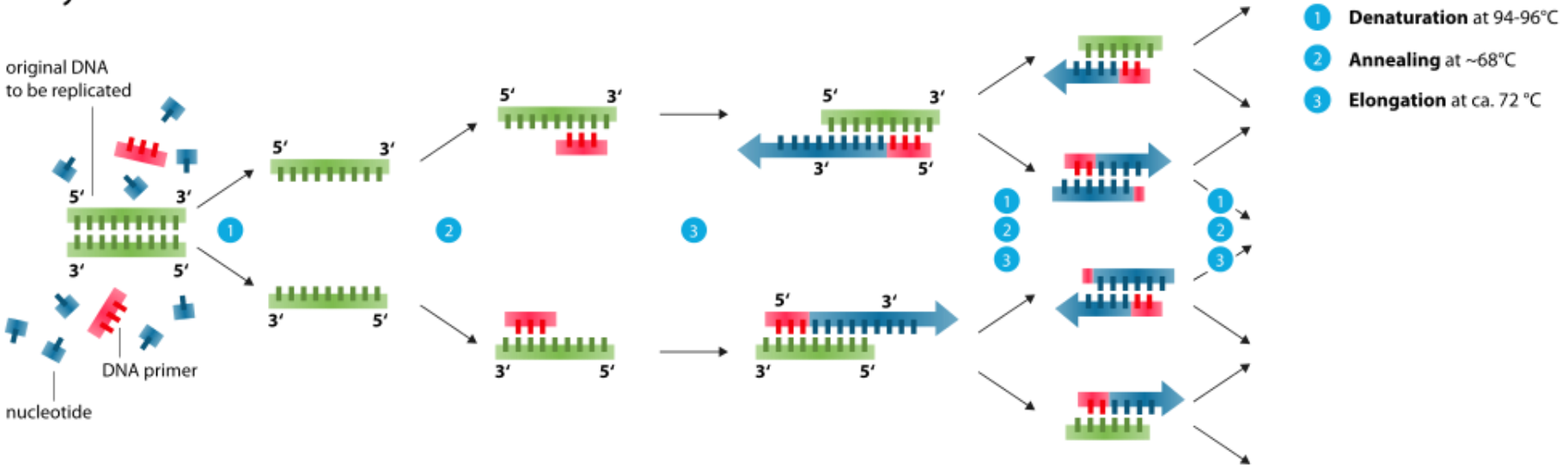
The PCR Revolution

Kary Mullis

- 1983 - Invention of PCR
- 1993 - Received the Noble Prize

Driving late one night, when he had the idea to **use a pair of primers** to **bracket the desired DNA sequence** and to copy it using DNA polymerase

Polymerase chain reaction - PCR



The Molecular Biology Timeline

1865	Gregor Mendel, Law of Heredity
1866	Johann Miescher, Purification of DNA
1949	Sickle Cell Anemia Mutation
1953	Watson and Crick, Structure of DNA
1970	Recombinant DNA Technology
1977	DNA sequencing
1985	*In Vitro* Amplification of DNA (PCR)
2001	The Human Genome Project

Human Genome Project

U.S. Government project coordinated by the Dept. of Energy and NIH

Total cost of over 3 billion $

Goals of the Human Genome Project (1990–2006)

- Identify all of the genes in human DNA
- Determine the sequences of the 3 billion bases that make up human DNA
- Create databases
- Develop tools for data analysis
- Address the ethical, legal, and social issues that arise from genome research

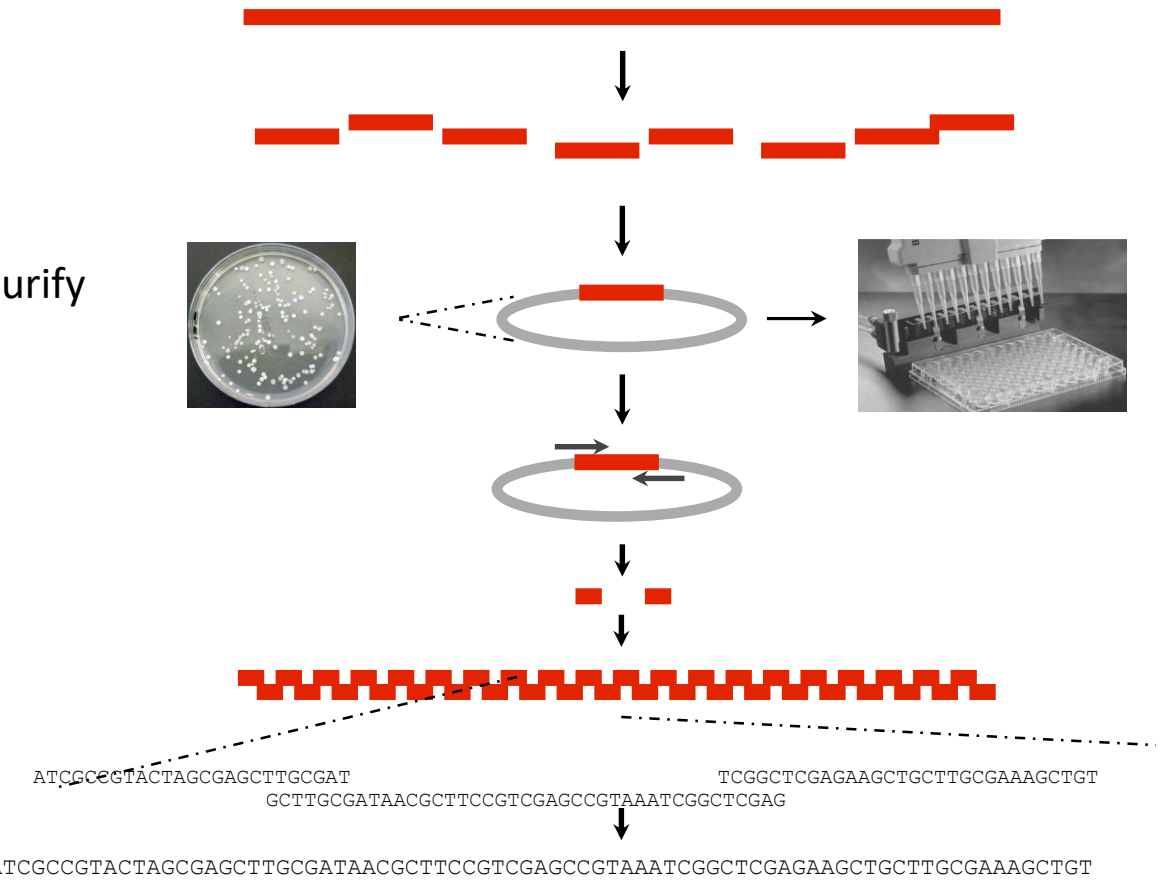
Shotgun genome sequencing (Sanger, 1979)

1) Fragment the genome (or large BAC clones)

2) Clone 2-10kb fragments into plasmids; pick lots of colonies; purify DNA from each

3) Sequence

4) Assemble the genome from overlapping “reads”



1977: Bacteriophage fX 174 (5kb)
1995: H. Influenza (1Mb);
1996: Yeast (12mb);
2000: Drosophila (165Mb);
2002: Human (3Gb)

Technological advances of Sanger sequencing

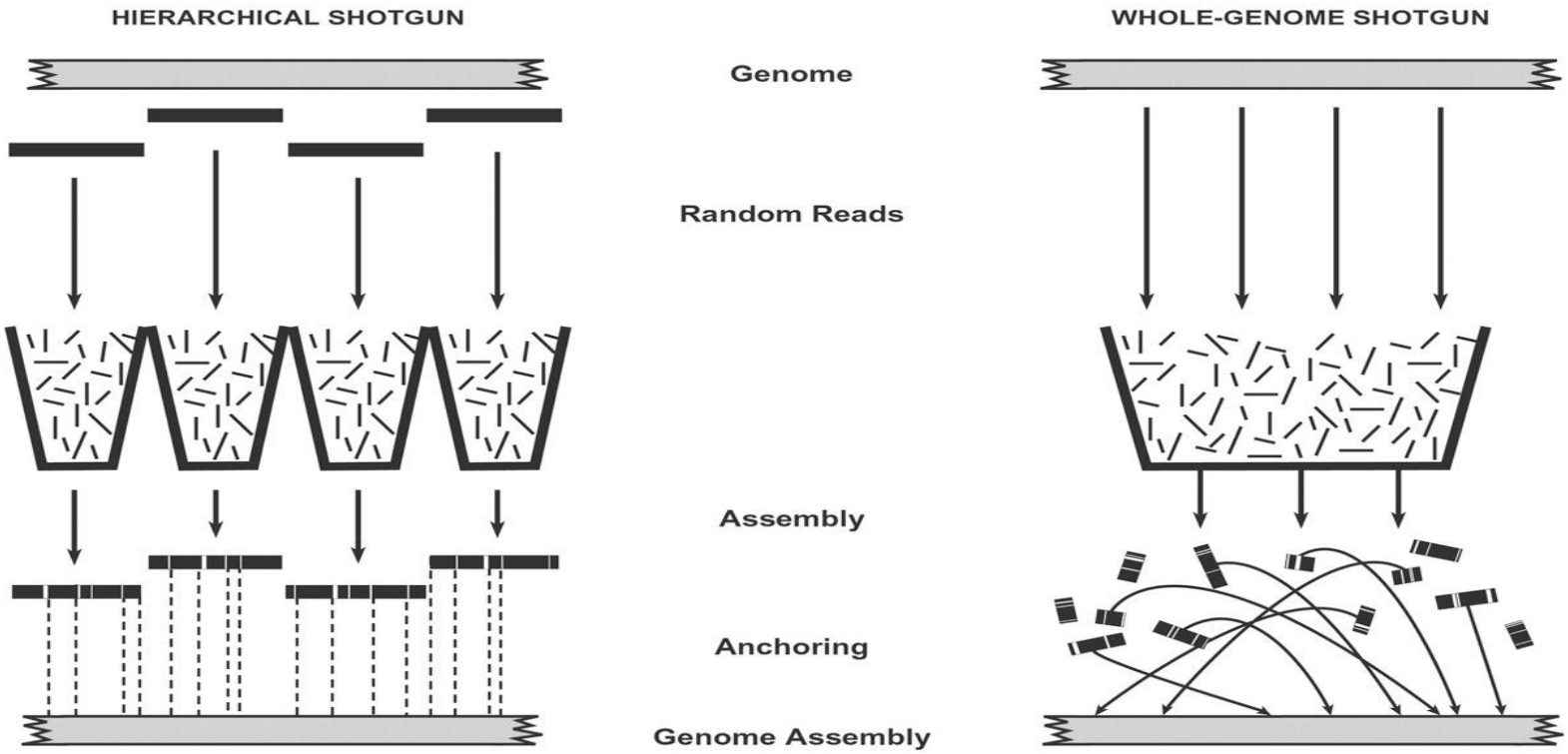
1977: Fred Sanger → 700 bases per day → 118,000 years to sequence the human genome

1985: ABI 370 (first automated sequencer) → 5000 bases per day →16,000 years

1995: ABI 377 (Bigger gels, better chemistry & optics, more sensitive dyes, faster computers) → 19,000 bases per day → 4,400 years

1999: ABI 3700 (96 capillaries, 96 well plates, fluid handling robots) → 400,000 bases per day → 205 years

Human Genome Sequencing



Public (Universities)
1990-2001 (2003)

Celera Corporation
1999-2001 (2003)

Impact on Human Diseases

Novelty

- Discovery of potential **novel molecular markers** of human diseases
- Utility of molecular markers to **develop useful molecular assays** for detection, diagnosis, and prediction of disease outcomes

Monitor diseases more accurately

- Allows for early treatment and better patient care

Determine most appropriate treatment

- Reduces or eliminates unnecessary treatment
- Reduces or eliminates inadequate treatment
- Yields greater cost effectiveness

Reduce patient morbidity and mortality

Further projects

First version of HG

Venter *et al.*,

Lander *et al.*,

Published "end-version" of HG

Start 1000 Genomes Project
-) lists genetic variations

1990 — 2001 — 2003 — 2008 — 2010

Start Human Genome Project
-) complete HG
-) 3.000.000.000 bp
-) 20 Institutes
-) 2.500 Scientists

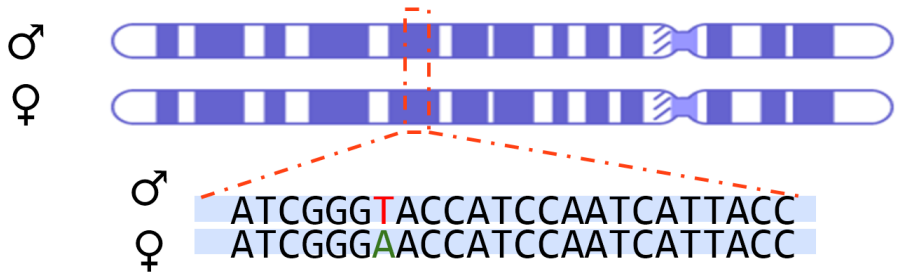
Start ENCODE Project
-) Encyclopedia of DNA Elements
-) functional elements of DNA

ENCODE Project
-) Data available
-) 1000 Genomes Browser

Terms

Genetics Terms

Humans are diploid: Our genome is comprised of a paternal and a maternal "haplotype“
→ form our "genotype"



Gene: A hereditary unit consisting of a sequence of DNA that occupies a specific location on a chromosome and determines a particular characteristic in an organism

Trait: A distinguishing feature, a genetically determined characteristic or condition

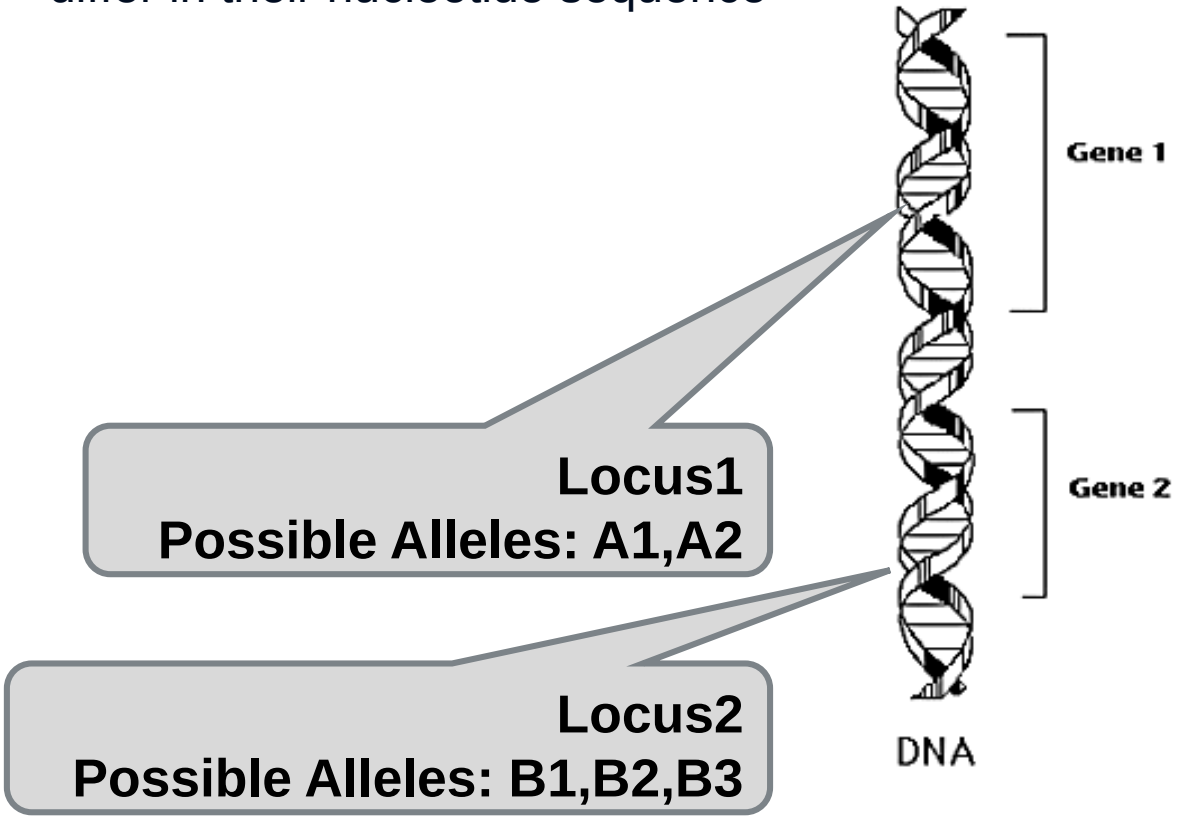
Genotype: Genetic makeup, distinguished from the physical appearance

Phenotype: The observable physical or biochemical characteristics as determined by both genetic makeup and environment

Genetics terms

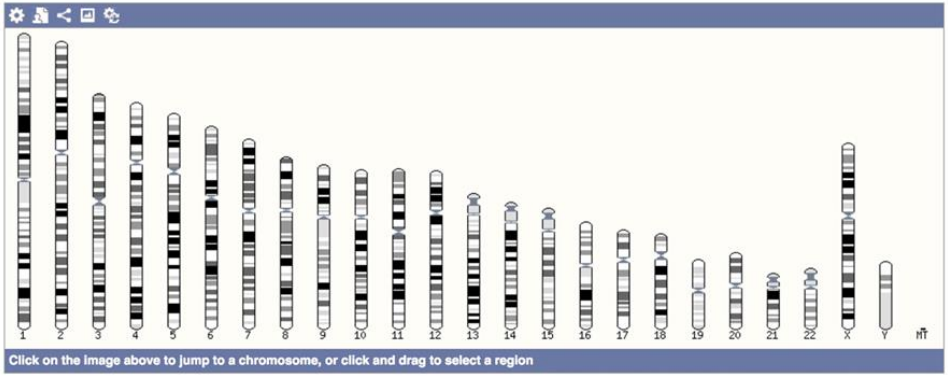
Locus location of a *gene/marker* on the chromosome

Allele one **variant form** of a gene/marker at a particular locus differ in their nucleotide sequence



The human genome - basic stats

- ~ 3 billion base pairs (haploid)
- ~ 20,000 protein coding genes
- ~ 200,000 coding transcripts (isoforms of a gene that each encode a distinct protein product)



Summary

Assembly	GRCh38.p7 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.22, Dec 2013
Database version	87.38
Base Pairs	3,547,762,741
Golden Path Length	3,096,649,726
Genebuild by	Ensembl
Genebuild method	Full genebuild
Genebuild started	Jan 2014
Genebuild released	Jul 2014
Genebuild last updated/patched	Jun 2016
Gencode version	GENCODE 25

Gene counts (Primary assembly)

Coding genes	20,441 (incl 526 readthrough)
Non coding genes	22,219
Small non coding genes	5,052
Long non coding genes	14,727 (incl 214 readthrough)
Misc non coding genes	2,222
Pseudogenes	14,606 (incl 5 readthrough)
Gene transcripts	198,002

http://uswest.ensembl.org/Homo_sapiens/Location/Genome

Forms of genetic variations

Single nucleotide substitution
Replacement of one nucleotide with another

(Recap) Tandem repeat
ATTCG ATTCG ATTCG

Microsatellites or mini-satellites
These tandem repeats often present high levels of inter- and intra-specific polymorphism

Deletions or insertions
Loss or addition of one or more nucleotides

Structural variations
Changes in chromosome number, segmental rearrangements and deletions

Genetics Terms

Polymorphism

- Variations in DNA sequence (substitutions, deletions, insertion, etc) that are present at a frequency greater than 1% in a population.
- Have a WEAK EFFECT or NO EFFECT at all.
- Ancient and COMMON

Mutation

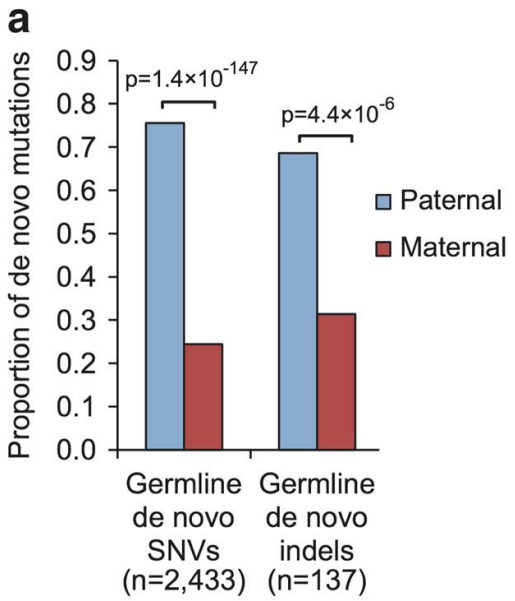
- Variations in DNA sequence (substitutions, deletions, etc) that are present at a frequency lower than 1% in a population.
- Can produce a gain of function and a loss of function.
- Recent and RARE.

De-novo mutations (DNM)

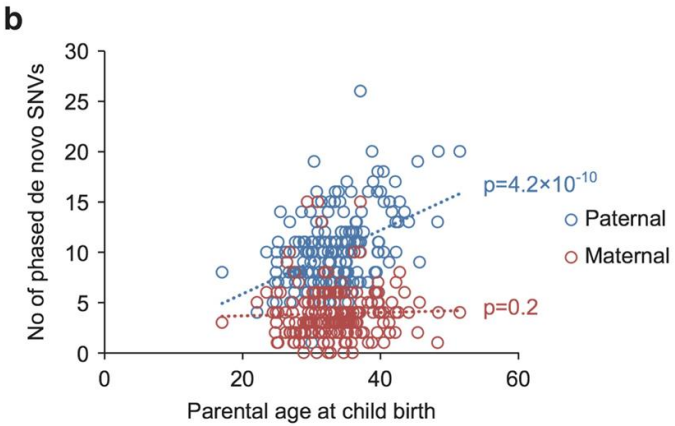
New mutation occurs in father's or mother's germ cell

- Roach et al. (2010) Science
  28 new mutations (per haploid genome) -> 56 diploid genome
- Nachman et al. (2000) Genetics
  175 mutations

DNMs are more likely to occur in the paternal germline & correlate with paternal age



2 new DNMs per year of paternal age (Kong et al. 2012, *Nature*)



Yuen et al. (2016) *Nature Genomic Medicine*

Glossary and Definitions

Homozygote
An organism with two identical alleles

Heterozygote
An organism with two different alleles

Hemizygote
Having only one copy of a gene
→ males are hemizygous for most genes on the sex chromosomes

Dominant trait
A trait that shows in a heterozygote

Recessive trait
A trait that is hidden in a heterozygote

Mutations

Humans

In human beings, **99.9% bases are same**

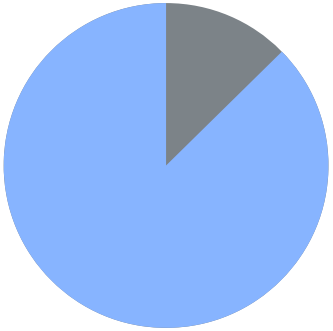
Remaining 0.1% makes a person unique

- Different attributes / characteristics / traits
- How a person looks
- Diseases

These variations can be:

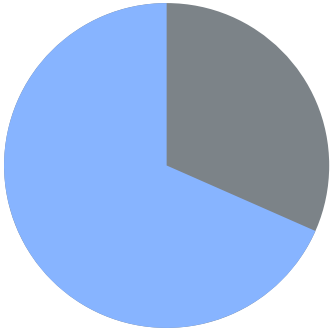
- Harmless (change in phenotype)
- Harmful (diabetes, cancer, heart disease, Huntington's disease, hemophilia)
- Latent (e.g. susceptibility to heart attack; variations found in coding and regulatory regions, are not harmful on their own, and the change in each gene only becomes apparent under certain conditions)

Genetic variations cause inherited diseases



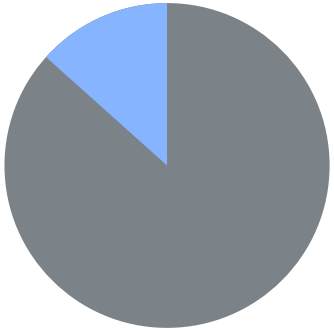
Genetic diseases

- Cystic fibrosis
- Down syndrome
- Sickle cell disease
- Turner syndrome



Complex Diseases

- Alzheimer disease
- Cardiovascular disease
- Diabetes (type 2)
- Parkinson Disease

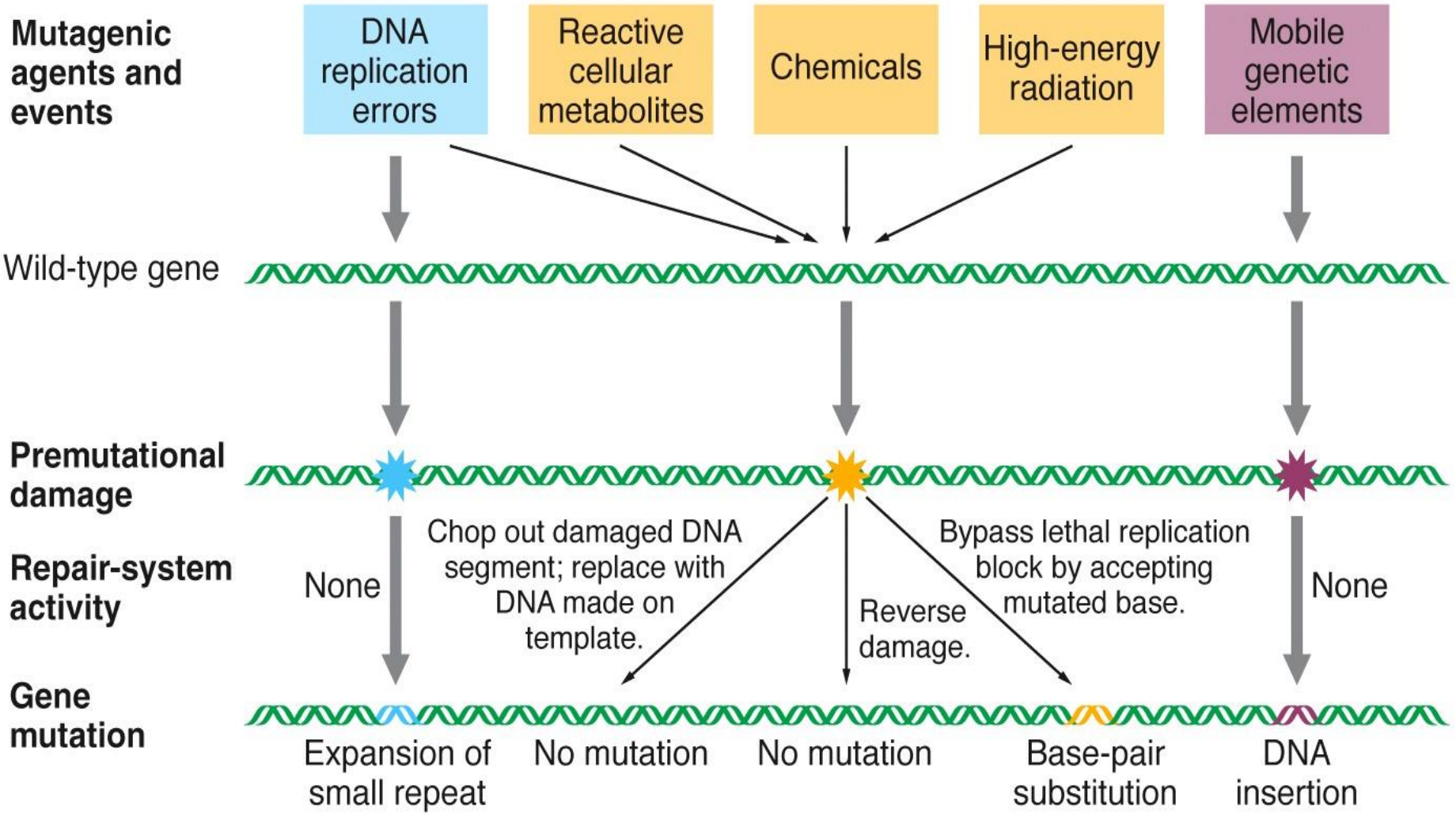


Environmental Diseases

- Influenza
- Hepatitis
- Measles

Genetically predisposed to a particular condition

Causes of gene mutations



Consequences of mutations

Most mutations are neutral

- 97% DNA neither codes for protein or RNA, nor indirectly affects gene function
- A new variant in the 1.5% coding regions may not result in a change in amino acid
- Variants that change amino acid may not affect function

Certain mutations have functional effect and even cause disease

- Gain-of-function mutations often produce dominant disorders
- Loss-of-function mutations result in recessive disease

Sequencing

Second Generation Sequencing

- Developed to **increase throughput of Sanger sequencing**
- Can sequence **many molecules in parallel**
  - Does not require homogenous input
  - DNA sequenced as clusters or in nanowells
  - Single machine can sequence 3-10 Billion independent DNA fragments at once
  - Single Sanger Sequencer maxes out at 1152 reactions per machine

- Time from DNA to genome reduced from 10 years to 1 day!

Disadvantages of 2nd Generation Tech

Rely on amplification to create libraries and clusters

- All polymerases have an inherent error rate (10^{-6}-10^{-7})
- Errors introduced every 10 million to 100 million bases
- Secondary validation of variants is key

GC bias

- PCR bias against GC rich sequences
- Exome capture bias against GC rich sequences

Short reads can miss large structural variations

- Genome Translocations and inversions likely will be missed
- Require significant read depth at break points for these variations to be detected

Trouble detecting small insertions and deletions

- Short reads computationally hard to align and call

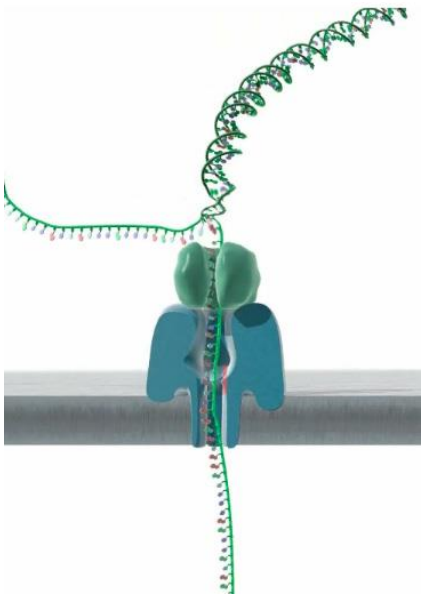
Third Generation Sequencing

Single molecule sequencing

- Less complex sample prep
- Much longer read length (1-100kb)
- Many technical hurdles with very high error rates
- Expensive

Sequencing by synthesis - Pacific Biosciences

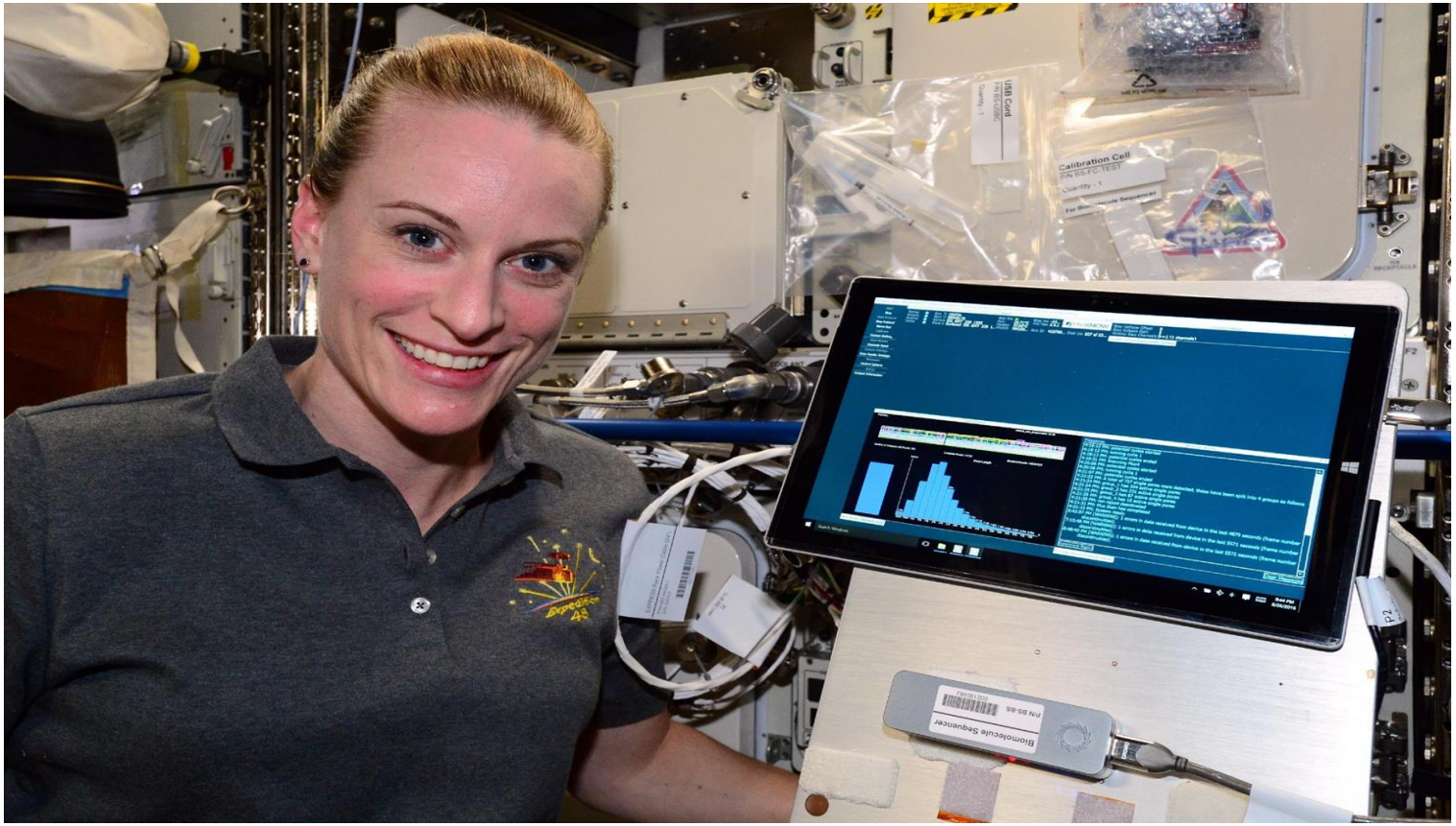
- 1 DNA molecule and 1 polymerase in each well (zero-mode waveguide)
- 4 different marker on the phosphate of the nucleotide
  → Polymerase interacts; marked freed
- No “theoretical” limit to DNA fragment length



Direct sequencing by passing DNA through a nanopore

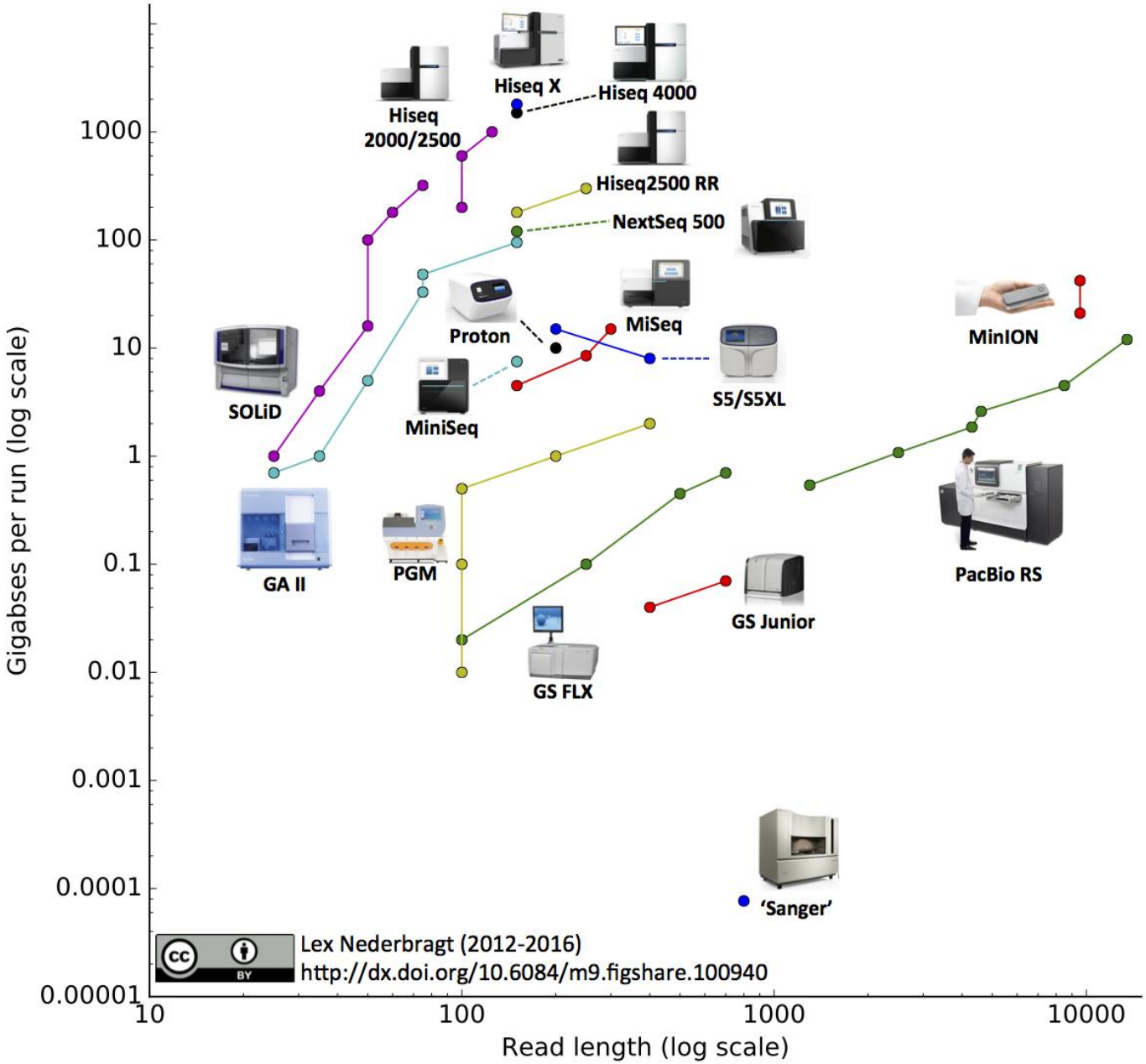
- Bases fed through a membrane bound nanopore
- Ionic difference between both sides of the membrane

Nanopore – very portable



Kate Rubins sequencing DNA on the ISS

High throughput sequencing



Properties of different technologies

Instrument	Amplification	Run time	Millions of Reads/run	Bases / read	Gbp/run
Illumina MiSeq	BridgePCR	5-55h	1-22	50-600	0.3-13.2
Illumina NextSeq 500	BridgePCR	11-30h	130-400	75-300	19.5-120
Illumina HiSeq 2500	BridgePCR	10h - 11days	300-2000	50-300	15-500
Ion Torrent - PGM	emPCR	2-7h	0.475-4.75	200-400	0.095-1.9
Ion Torrent - Proton	emPCR	4-6h	70-500	175	12.25-87.5
Pacific Biosciences RS II	None	2 hrs.	0,03	3000	0,09
Oxford Nanopore MinION (forecast)	None	≤6 hrs.	0,1	9000	0,9

http://www.molecularecologist.com/next-gen-fieldguide-2016/

Error Rates

Instrument	Primary Errors	Single-pass Error Rate (%)	Final Error Rate (%)
Illumina	Substitutions	~0.1	~0.1
Ion Torrent	INDELs	~1	~1
Oxford Nanopore	Deletions	≥4	4
PacBio RS	INDELs	~13	≤1

Flavors of Sequencing

Whole Genome Sequencing

- Obtain whole blood or tissue sample
- Create sequencing libraries of all DNA fragments

Whole Exome Sequencing

- Utilizes a selection protocol
- Attach complimentary RNA or DNA strands to beads
- Fish out **only** coding DNA sequences
- Create sequencing libraries from enriched DNA
- Reduces cost and analysis time

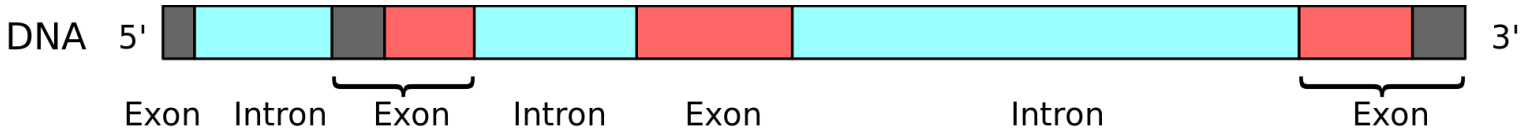
Custom Capture

- Same protocol as Exome sequencing
- Only target desired DNA sequences

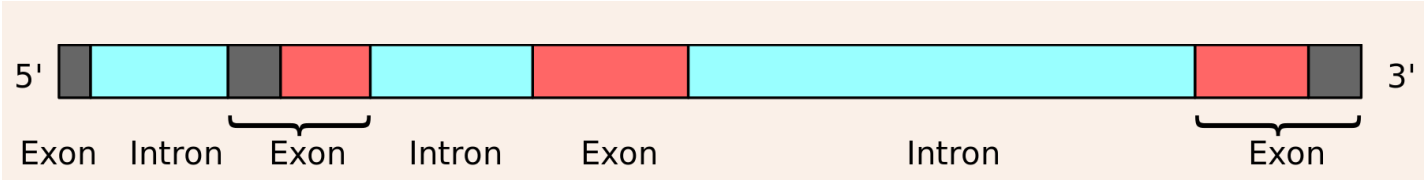
Amplicon Sequencing

- Use PCR to amplify target DNA
- Sequence amplified DNA (Amplicon)

Sequencing Techniques

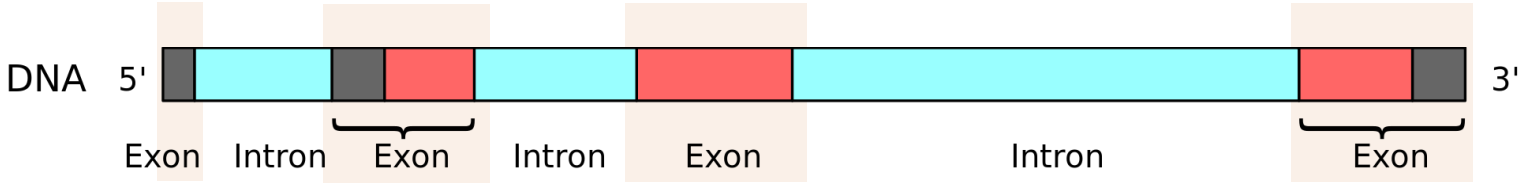


Sequencing Techniques



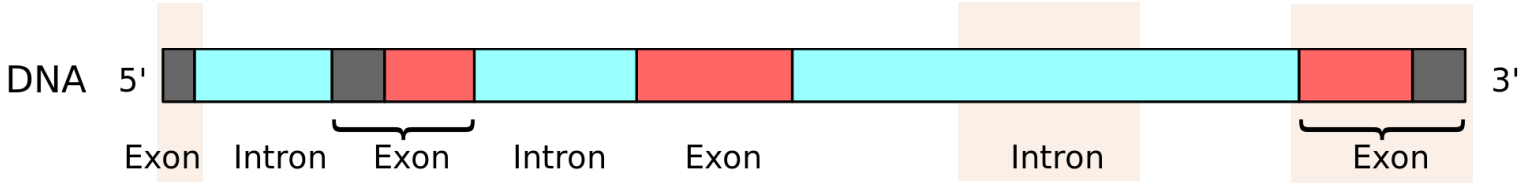
- Whole genome sequencing

Sequencing Techniques



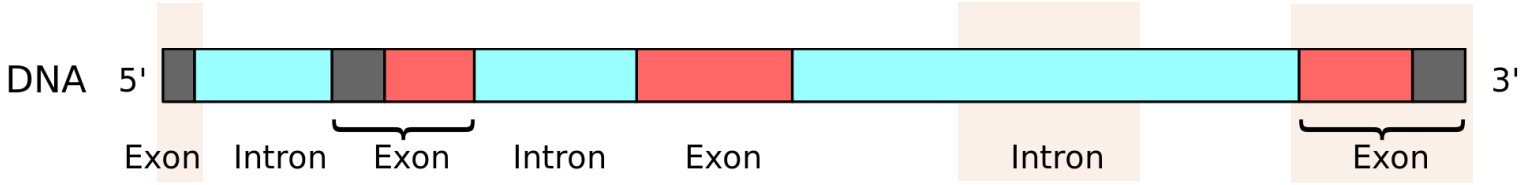
- Whole genome sequencing
- Whole exome sequencing
- Custom capture

Sequencing Techniques



- Whole genome sequencing
- Whole exome sequencing
- Custom capture
- Amplicon sequencing

Sequencing Techniques



- Whole genome sequencing
- Whole exome sequencing
- Custom capture
- Amplicon sequencing

What is the best technology for my use-case?

- Biological/clinical question?
- Number of samples?
- Cost?
- Future strategies?

NGS **Data analysis**

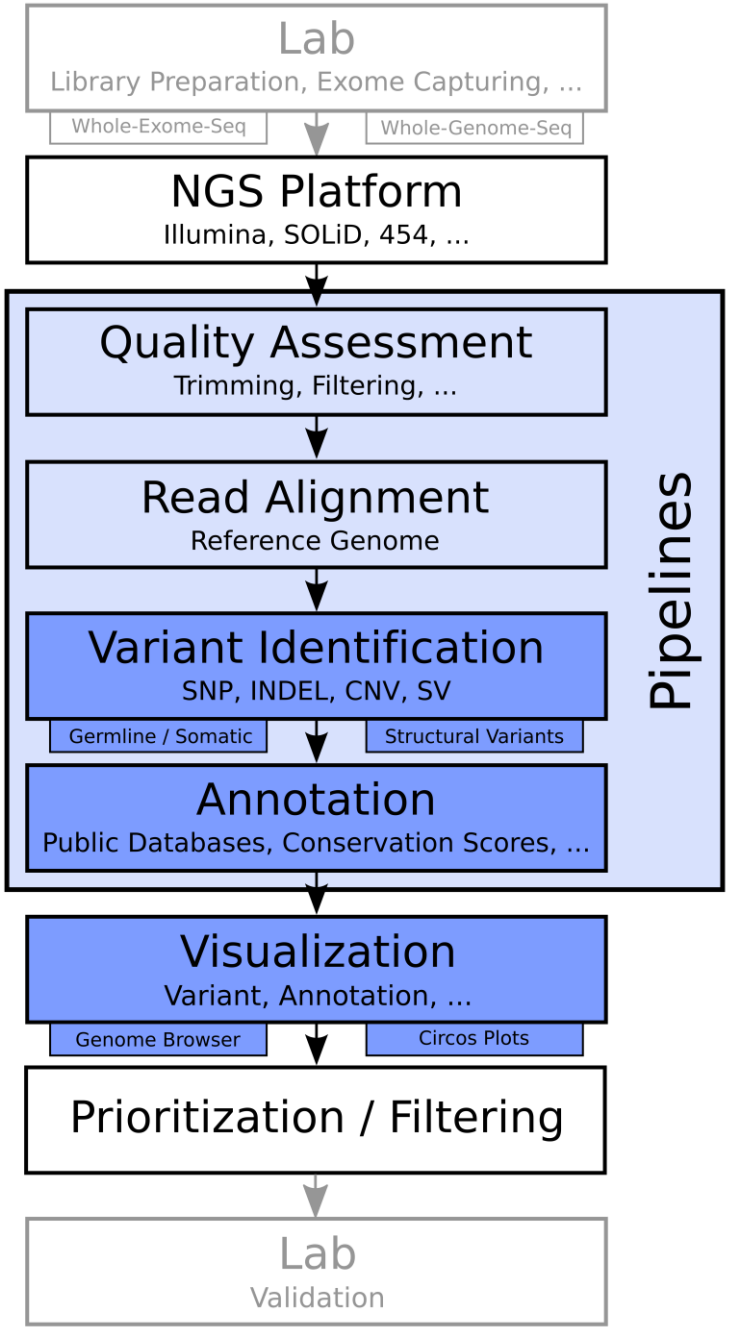
- Absolutely the largest roadblock for next generation sequencing
- Terabytes of data are useless if we can't efficiently analyze the data
- How long should data be kept?
  - Depends on application
  - Human Diagnostic sequencing?
  - Research sequencing?
- Where should data be kept and processed?
  - Local or Cloud (Amazon, etc)?
  - Cost of infrastructure vs cost of cloud service
  - Security issues
- Future
  - Cloud based solutions will become more attractive

Clinical Sequencing Costs (high X)

	Whole Genome	Exome	Custom Capture	Amplicon
Size (GB)	100	12.5	0.13-1	0.03-0.13
Preparation	$400	$200	$80	$40
Sequencing	$4,300	$400	$12-100	$1-12
Data Processing/Storage	$350	$200	$50	$25
Clincal Review	$5,000-10,000	$2,000-6000	$700-2000	$400-900
Total	$10,000-15,000	$2,800-6,800	$1,000-2,000	$500-1,000

DNA sequencing costs are falling, but analysis and clinical review cost will likely remain stable

Workflow overview

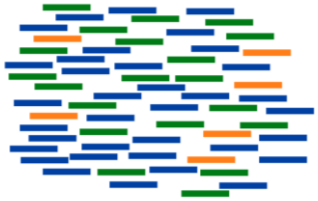

Lab
Library Preparation, Exome Capturing, ...
Whole-Exome-Seq
Whole-Genome-Seq
NGS Platform
Illumina, SOLiD, 454, ...
Quality Assessment
Trimming, Filtering, ...
Read Alignment
Reference Genome
Variant Identification
SNP, INDEL, CNV, SV
Germline / Somatic
Structural Variants
Annotation
Public Databases, Conservation Scores, ...
Pipelines
Visualization
Variant, Annotation, ...
Genome Browser
Circos Plots
Prioritization / Filtering
Lab
Validation

Mapping and QC

SAM - Sequence Alignment/Map Format

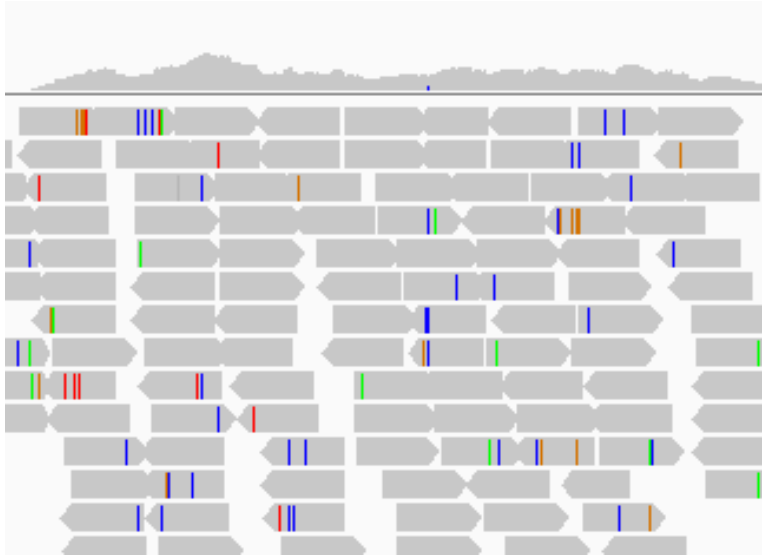
Mapping - Principle

Short sequence reads

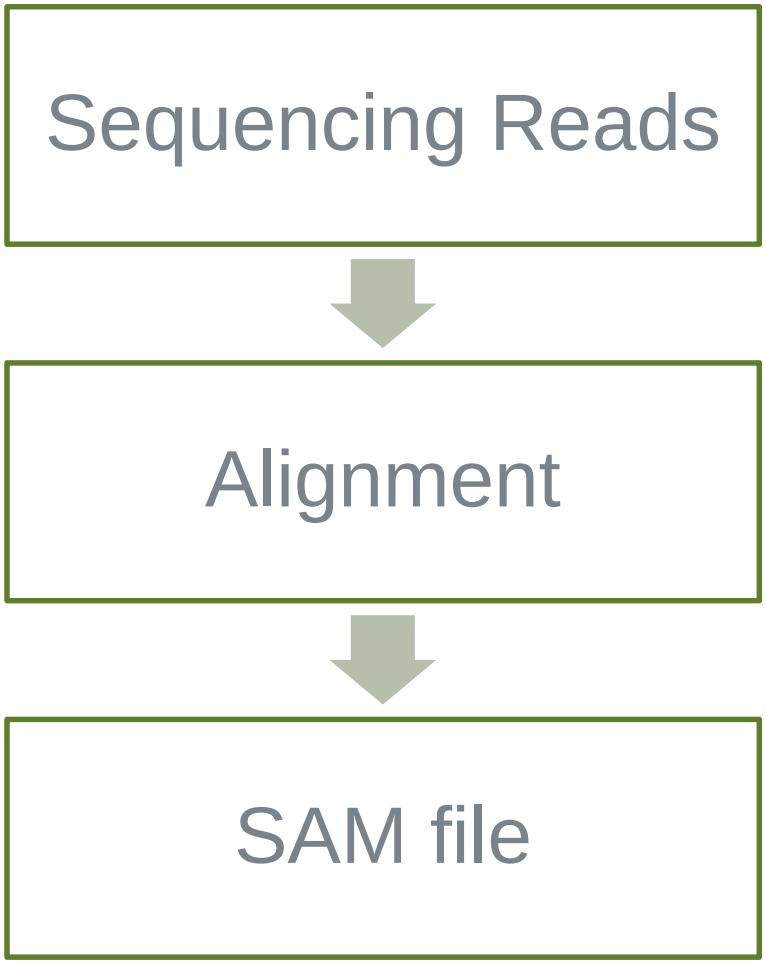


Alignment

Reference



Adapted from http://raetschlab.org//members/research/transcriptomics/images/RNA-Sequencing.png


Sequencing Reads
Alignment
SAM file

Sequence **mapping** versus **alignment**

Mapping: (quickly) find the best possible loci to which a sequence could be aligned

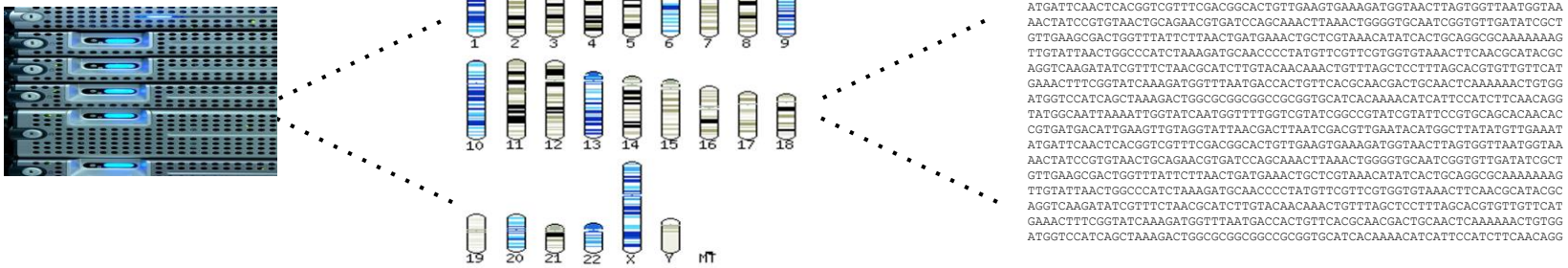
Alignment: for each locus to which a sequence can be mapped, determine the optimal base by base alignment of the query sequence to the reference sequence

Paired-end sequencing

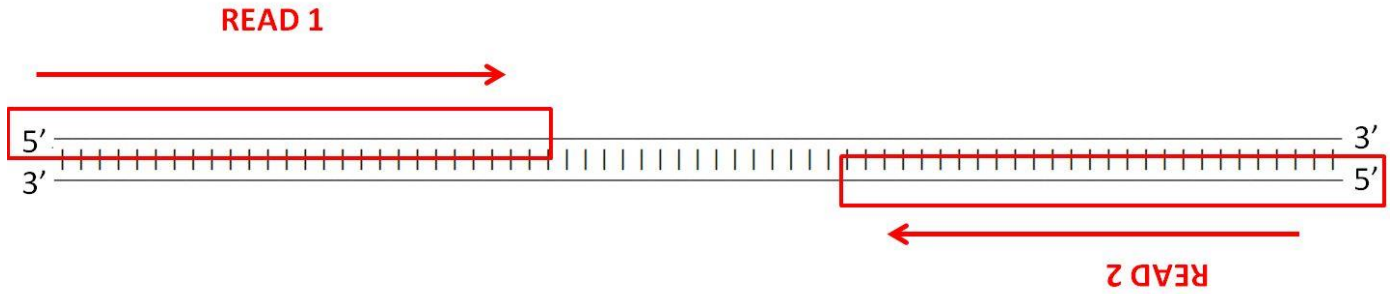
1) A read-pair

5’ GGTGTACGAATAGTTTCCTTTTACACTCCTTGACCATCCTAGC ----------//---------- GGACTGAAACTTCATCTGTCTTTATAGATATGCGTGCAGCAGC 5’

2) A reference genome on a computer



3) Alignment of the read-pair to the reference genome gives coordinates describing where in the human genome the read-pair came from



SAM file format

The Sequence Alignment/Map (SAM) Format and SAMtools
Heng Li et al. Bioinformatics, 2009

Tab-delimited text file

Output of most alignment programs

- Header section
- Alignment section
- 11 Required columns
- Optional fields

SAM Header

- Header lines start with @ symbol
- Always at top of file
- Contain lots of information about what was mapped, what it was mapped to, and how (metadata)
  - The version information for the SAM/BAM file
  - Whether or not and how the file is sorted
  - Information about the reference sequences
  - Any processing that was used to generate the various reads in the file
  - Software version

1.4 The alignment section: mandatory fields

In the SAM format, each alignment line typically represents the linear alignment of a segment. Each line has 11 mandatory fields. These fields always appear in the same order and must be present, but their values can be '0' or '*' (depending on the field) if the corresponding information is unavailable. The following table gives an overview of the mandatory fields in the SAM format:

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0, 2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0, 2^{31}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0, 2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next read
8	PNEXT	Int	$[0, 2^{31}-1]$	Position of the mate/next read
9	TLEN	Int	$[-2^{31}+1, 2^{31}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

Example SAM file

AUSTRIAN INSTITUTE OF TECHNOLOGY

```
@PG     ID:bwa_sam      PN:bwa  PP:bwa_aln_fastq        VN:0.5.9-r16    CL:bwa sampe -a 1050 -r $rg_line -f $sam_file $reference_fasta $sai_file(s) $fastq_file(s)
@PG     ID:sam_to_fixed_bam     PN:samtools     PP:bwa_sam      VN:0.1.17 (r973:277)    CL:samtools view -bSu $sam_file | samtools sort -n -o - samtools_nsort_tmp | s
ort_tmp | samtools fillmd -u - $reference_fasta > $fixed_bam_file
@PG     ID:gatk_target_interval_creator PN:GenomeAnalysisTK     PP:sam_to_fixed_bam     VN:1.2-29-g0acaf2d      CL:java $jvm_args -jar GenomeAnalysisTK.jar -T Realig
dels_file(s)
@PG     ID:bam_realignment_around_known_indels  PN:GenomeAnalysisTK     PP:gatk_target_interval_creator VN:1.2-29-g0acaf2d      CL:java $jvm_args -jar GenomeAnalysis
bam_file -targetIntervals $intervals_file -known $known_indels_file(s) -LOD 0.4 -model KNOWNS_ONLY -compress 0 --disable_bam_indexing
@PG     ID:bam_count_covariates PN:GenomeAnalysisTK     PP:bam_realignment_around_known_indels  VN:1.2-29-g0acaf2d      CL:java $jvm_args -jar GenomeAnalysisTK.jar -
recal_data.csv -knownSites $known_sites_file(s) -l INFO -L '1;2;3;4;5;6;7;8;9;10;11;12;13;14;15;16;17;18;19;20;21;22;X;Y;MT' -cov ReadGroupCovariate -cov QualityScor
@PG     ID:bam_recalibrate_quality_scores       PN:GenomeAnalysisTK     PP:bam_count_covariates VN:1.2-29-g0acaf2d      CL:java $jvm_args -jar GenomeAnalysisTK.jar -
.csv -I $bam_file -o $recalibrated_bam_file -l INFO -compress 0 --disable_bam_indexing
@PG     ID:bam_calculate_bq     PN:samtools     PP:bam_recalibrate_quality_scores       VN:0.1.17 (r973:277)    CL:samtools calmd -Erb $bam_file $reference_fasta > $
@PG     ID:bam_merge    PN:picard       PP:bam_calculate_bq     VN:1.53 CL:java $jvm_args -jar MergeSamFiles.jar INPUT=$bam_file(s) OUTPUT=$merged_bam VALIDATION_STR
@PG     ID:bam_mark_duplicates  PN:picard       PP:bam_merge    VN:1.53 CL:java $jvm_args -jar MarkDuplicates.jar INPUT=$bam_file OUTPUT=$markdup_bam_file ASSUME_SOR
@PG     ID:bam_merge.1  PN:picard       PP:bam_mark_duplicates  VN:1.53 CL:java $jvm_args -jar MergeSamFiles.jar INPUT=$bam_file(s) OUTPUT=$merged_bam VALIDATION_STR
@CO     $known_indels_file(s) = ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_mapping_resources/ALL.wgs.indels_mills_devine_hg19_leftAligned_co
@CO     $known_indels_file(s) .= ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_mapping_resources/ALL.wgs.low_coverage_vqsr.20101123.indels.site
@CO     $known_sites_file(s) = ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_mapping_resources/ALL.wgs.dbsnp.build135.snps.sites.vcf.gz
SRR070823.24480225      163     11      60464   0       100M    =       60720   355     CCATGGTTAAACATAAATGCAAATATGTTAATGTTTACTGAATAACTTATCTGTGCCAAGTGGTGTATTAATGATTC
MNKJKMKNJNNNMGNDNNGNGGKKGNKNKKJNMILLLLCKLLFKKLBLHJGFHED    X0:i:3  X1:i:6  MD:Z:100        RG:Z:SRR070823  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@I@@@@@@@@@@@@
@@@@@@@@@@@@@GI
SRR070823.24480518      1187    11      60464   0       100M    =       60720   355     CCATGGTTAAACATAAATGCAAATATGTTAATGTTTACTGAATAACTTATCTGTGCCAAGTGGTGTATTAATGATTC
DJKBEIJBEJFIJCJGKEGKGCHIG@HIHDHNLKCJLIAKLLIKMIJLFBEE?EB    X0:i:3  X1:i:6  MD:Z:100        RG:Z:SRR070823  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@E@@@@@@@@@@@
@@@@@@@@@@@@@GG
SRR070823.24480225      83      11      60720   0       100M    =       60464   -355    TATGGAGTTTGATGTTATGTCAGGGTAATTACATGATTATATAATTAACAGGTTTCTTTTTAAATCAGCTATATCAA
GHHKKJJJJGJJJJJGHHGGJJJJGGGGGJGGGJIHGGHGGGGGHIEJIIEDCF6    X0:i:9  X1:i:0  MD:Z:100        RG:Z:SRR070823  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@@@@@@@@@@@@@
@@@@@@@@@@@@@@@
SRR070823.24480518      1107    11      60720   0       100M    =       60464   -355    TATGGAGTTTGATGTTATGTCAGGGTAATTACATGATTATATAATTAACAGGTTTCTTTTTAAATCAGCTATATCAA
=EGJJF>?@DBADDD>BBB<AD:@2B@BABBC>B;::::E@DBDG;6E5=A??@6    X0:i:9  X1:i:0  MD:Z:100        RG:Z:SRR070823  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@@@@@@@@@@@@@
@@@@@@@@@@@@@@@
SRR070531.23281260      99      11      61942   0       100M    =       62035   192     TCCATGCCGCTTTGATGACAGGATAATATGTAAGCTTTTCTATATTTCAGAAACTATATGACATGACGAAAAGTAAA
JJMLKHLIJMJHJKNHFGKNHMGGMJHJFGIGIINOLNIKMNENNIFILKGGHEA    X0:i:8  X1:i:1  MD:Z:100        RG:Z:SRR070531  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@@@@@@@@@@@@@
@@@@@@@@@@@@@@@
SRR070531.23281260      147     11      62035   0       100M    =       61942   -192    GGAGGTATCCTGAATTGACTGAGAAATAAGGAGGTATTCCACAGAGAATATAAATAAACATATACTTAGTGTTCAAG
MKKJKJKKCJKKJNKJJJJMLJMLLLLIMJMMLJJIJIIJIIFIKKKKHGJECG6    X0:i:8  X1:i:2  MD:Z:100        RG:Z:SRR070531  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:CB@@@@@@@@@@@
@@@@@@@@@@@@@@@
SRR070823.17243685      99      11      62388   0       100M    =       62452   163     CCTTGCCAATTGTGTTCTCCTTTATTTCCTGCTGGATATGACCACTGTCCTTCCATTGCATTGTATGTGTTTTTTAA
MGHMMMKIMLKHHJEKKFFDHJEKDEIDCCDGF?IEGJHIDEHJIGE?GAFDCCC    X0:i:10 X1:i:0  MD:Z:100        RG:Z:SRR070823  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@E@@@@@@@@@@@
@@@@@@@@@@@@@@@
SRR070823.17243685      147     11      62452   0       100M    =       62388   -163    ATGTGTTTTTTAATAGACTTTAATGGTTCTCAAGTGATGCATTATTTAGTTTGGTTCTTTGAAACTTATATAAATGA
MJNMMMLJMLJNLLIMKKJFLJGGJIJJJLLJJGLIJKKKJKHLKIKHJGJHDA6    X0:i:10 X1:i:0  MD:Z:100        RG:Z:SRR070823  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@@@@@@@@@@@@@
@@@@@@@@@@@@@@@
SRR070823.1968642       163     11      62725   0       100M    =       62918   292     TTTTTAACCATTACAAACAATGCTGCTATGAACATTCTTTTGTAAATCACCTGGTTCATATGTGCAAGATATCCTCT
LKMLMNNNNHKMMKJHNHNNMMNHHKKNNKNLNNNMGHIFIEIIIGJDHHGCCBD    X0:i:9  X1:i:1  MD:Z:100        RG:Z:SRR070823  AM:i:0  NM:i:0  SM:i:0  MQ:i:0  XT:A:R  BQ:Z:@EDDB@@@@@@@@@
@@@@@@@@@@@@AAF
```

2 FLAG

- Bitwise
- Picard (online) – explain flags

Examples

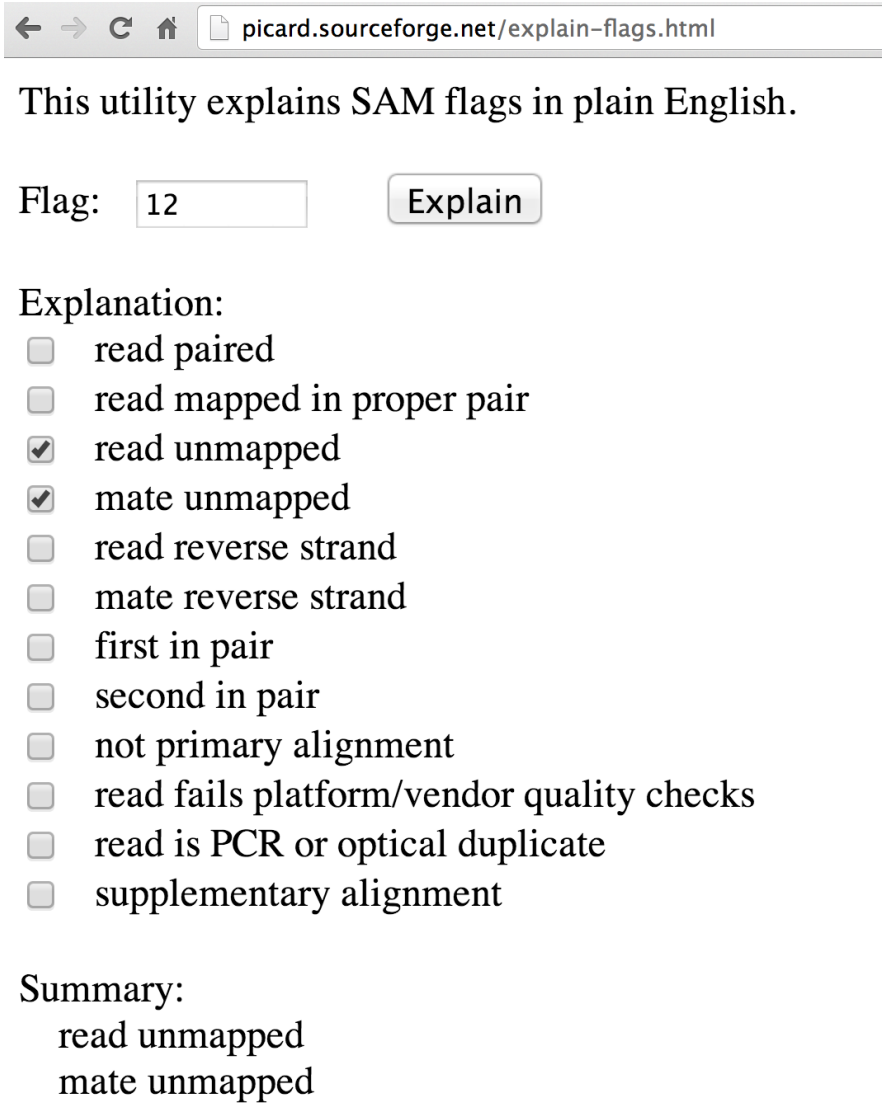
1 = 0001 -> PE read
4 = 0100 -> Unmappable
5 = 0101 -> Unmapped PE

2 FLAG

- Bitwise
- Picard (online) – explain flags

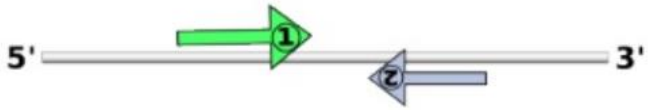
Examples

1 = 0001 -> PE read
4 = 0100 -> Unmappable
5 = 0101 -> Unmapped PE



2 FLAG

Reads mapped in proper pair

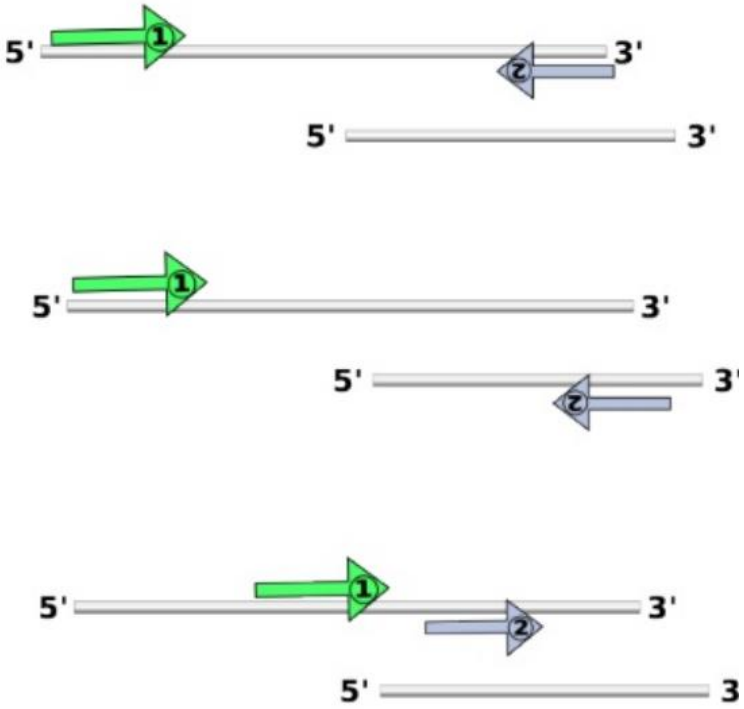


Flag: 67 Explain

Explanation:

- [x] read paired
- [x] read mapped in proper pair
- [ ] read unmapped
- [ ] mate unmapped
- [ ] read reverse strand
- [ ] mate reverse strand
- [x] first in pair

Reads not mapped in proper pair



3 RNAME & 4 POS & 5 MAPQ

AUSTRIAN INSTITUTE OF TECHNOLOGY

RNAME

- Reference name
  FASTA sequence name (e.g.: chr4)

POS

- Mapping position
  leftmost position in reference (e.g.: 142345)
  ! Reverse strand

MAPQ

- Mapping quality
- Phred score
- Depends on mapping program

6 CIGAR

Used as a compact way to represent sequence alignment

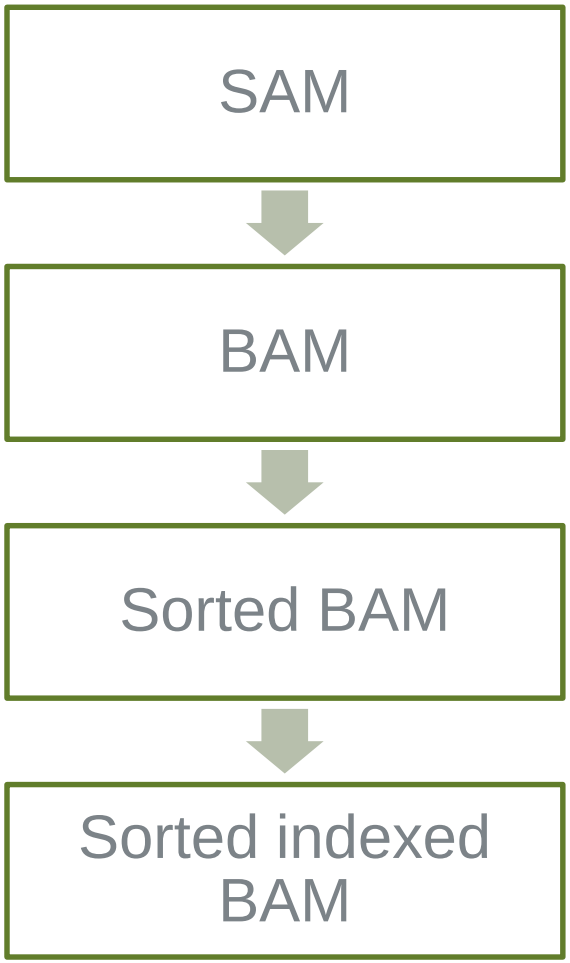
Read `ACGC-TGCAGTTATATAGG`
Ref `ACTCAGTG--GT`
Cigar 4M1D3M2I2M7S

?

Op	BAM	Description
`M`	0	alignment match (can be a sequence match or mismatch)
`I`	1	insertion to the reference
`D`	2	deletion from the reference
`N`	3	skipped region from the reference
`S`	4	soft clipping (clipped sequences present in SEQ)
`H`	5	hard clipping (clipped sequences NOT present in SEQ)
`P`	6	padding (silent deletion from padded reference)
`=`	7	sequence match
`X`	8	sequence mismatch

BAM – Binary SAM

SAM / BAM



Convert (view)

sort

index

BAM

SAM

- Information on the alignment of each read
- Optimized for readability and sequential access

BAM (Binary SAM)

- Compressed -> saves disk space; with BGZF (Blocked GNU Zip Format) - a variant of GZIP
- Can be sorted & indexed - quick viewing/searching (bigger than GZIP files)
- Cannot be read without a tool (samtools)

uBAM

- unmapped BAM → compress FASTQ files

CRAM

- Better lossless compression than BAM
- Cramtools for conversion from/to BAM
- http://www.ebi.ac.uk/ena/about/cram_toolkit

Practicals – Day 1