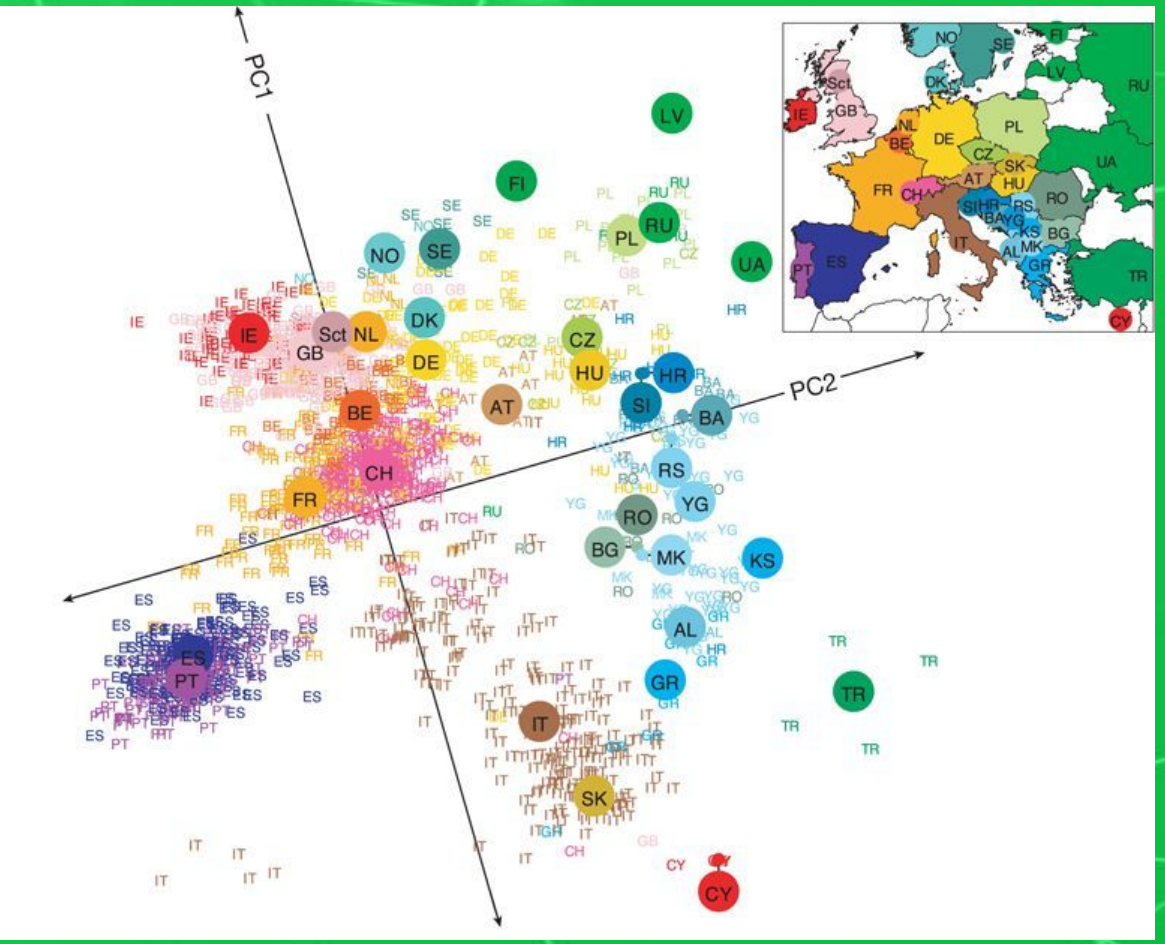
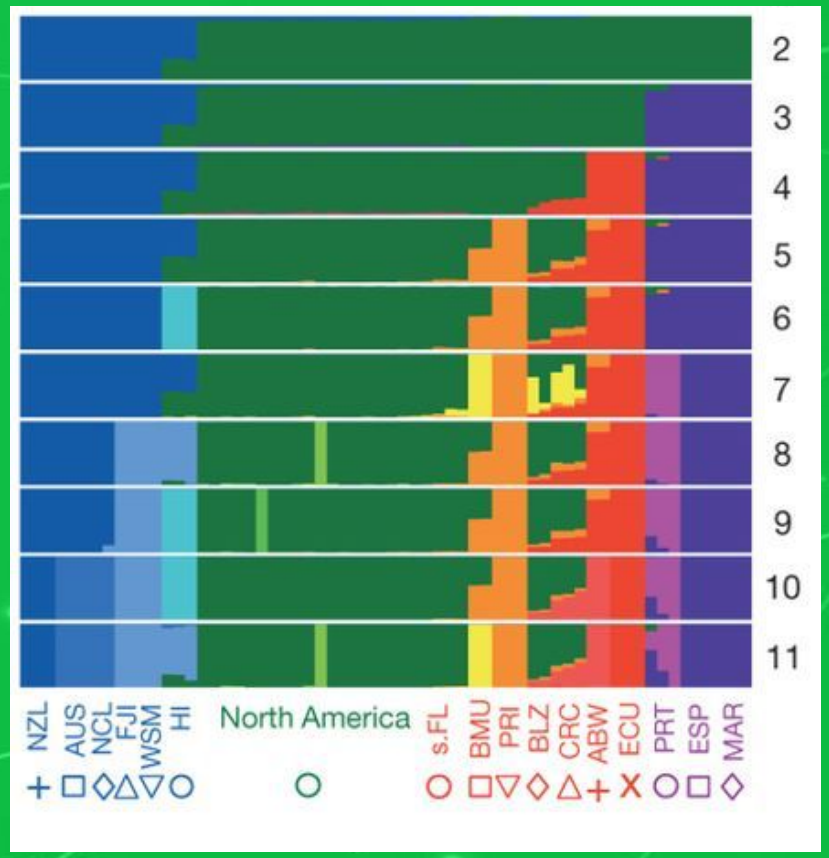
Centre for Genomic Research

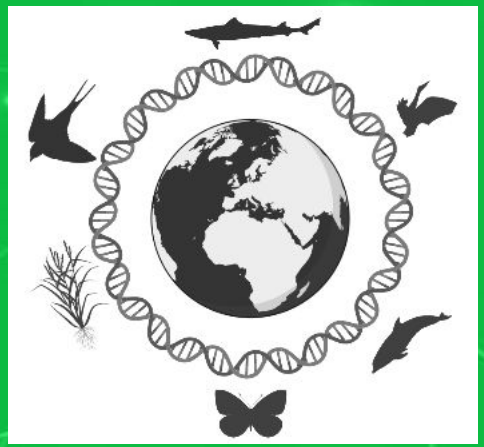
UKRI Natural Environment Research Council

NEOF NERC ENVIRONMENTAL OMICS FACILITY

Population Genomics







UNIVERSITY OF LIVERPOOL

The University Of Sheffield.

UNIVERSITY OF LIVERPOOL

Natural Environment Research Council

Centre for Genomic Research

NEOF

NERC ENVIRONMENTAL OMICS FACILITY

The University Of Sheffield.

Websites

NEOF: https://neof.org.uk/

NERC: https://nerc.ukri.org/

CGR: https://www.liverpool.ac.uk/genomic-research/

Twitter

NEOF: @NERC_EOF

NERC: @NERCscience

CGR: @CGR_UoL

Upcoming workshops

https://neof.org.uk/training/

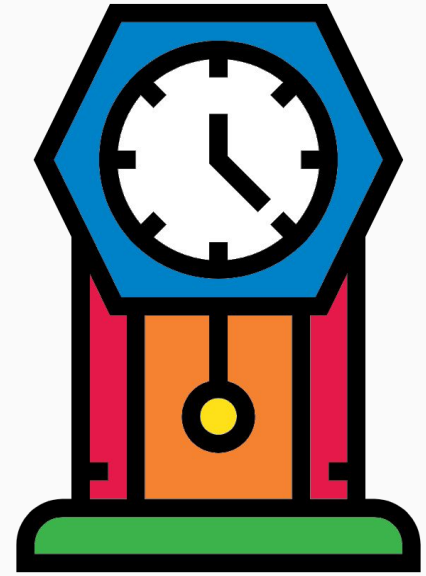
- Bacterial 16S metabarcoding
  - 7th & 9th February 2023
- Metabarcoding for diet analysis and environmental DNA
  - 28th February & 2nd March 2023
- Microbial shotgun metagenomics
  - 21st & 23rd March 2023
- Eukaryote genome assembly
  - 18th & 20th April 2023
- More!

Format & Schedule

This intro
Bookdown
Theory
Practice
Exercises
Optional materials

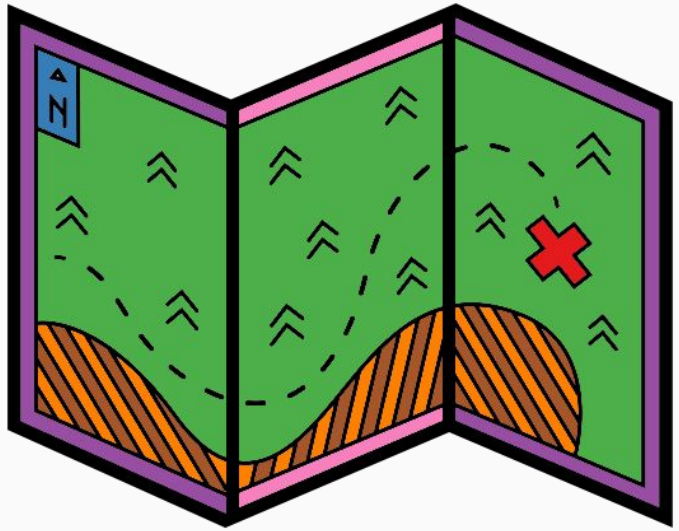
Work at your own pace
We are here to help
Time with breaks in between

- 10:00-11:20
- 11:30-12:30
- 13:30-14:40
- 15:00-16:00



Outline for Today

- SNPs and NGS
- Quality control
- Alignment and SNP calling
- Genetic structure
- Population genetics summary statistics

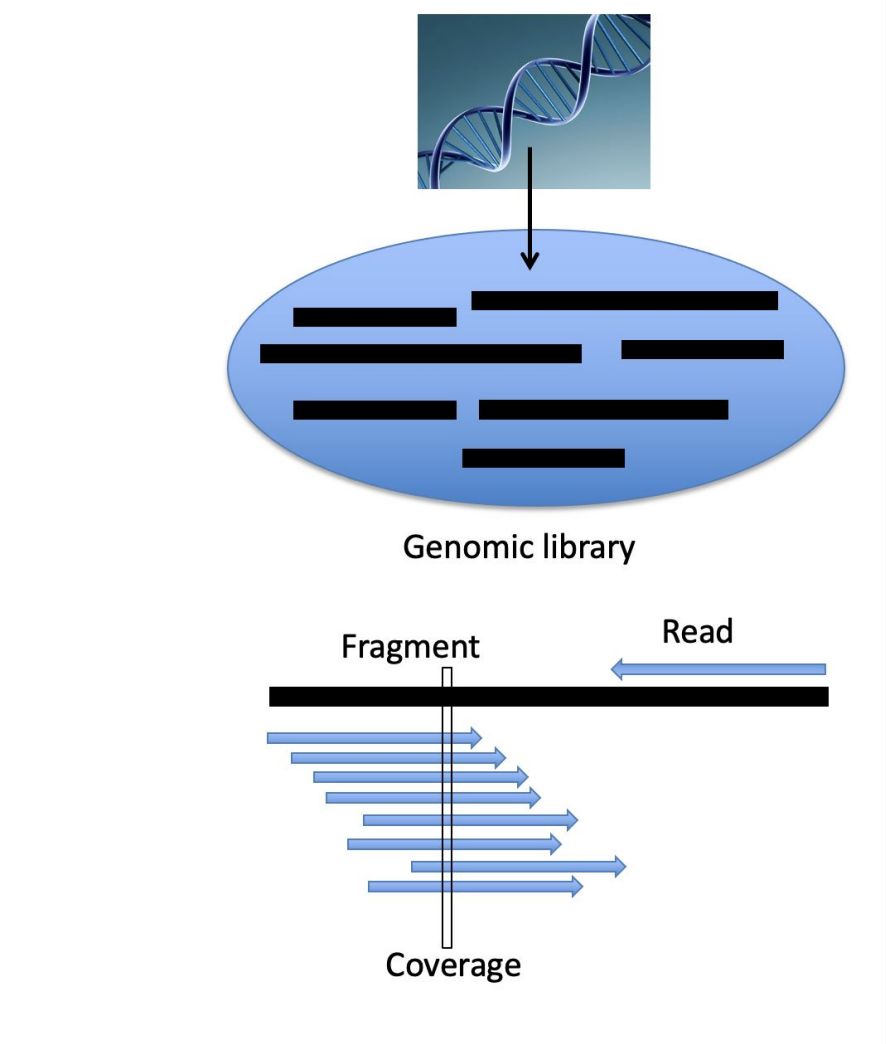


Genome sequencing - Basic concepts

- **Genomic library** – collection of DNA from source organism divided into multiple fragments

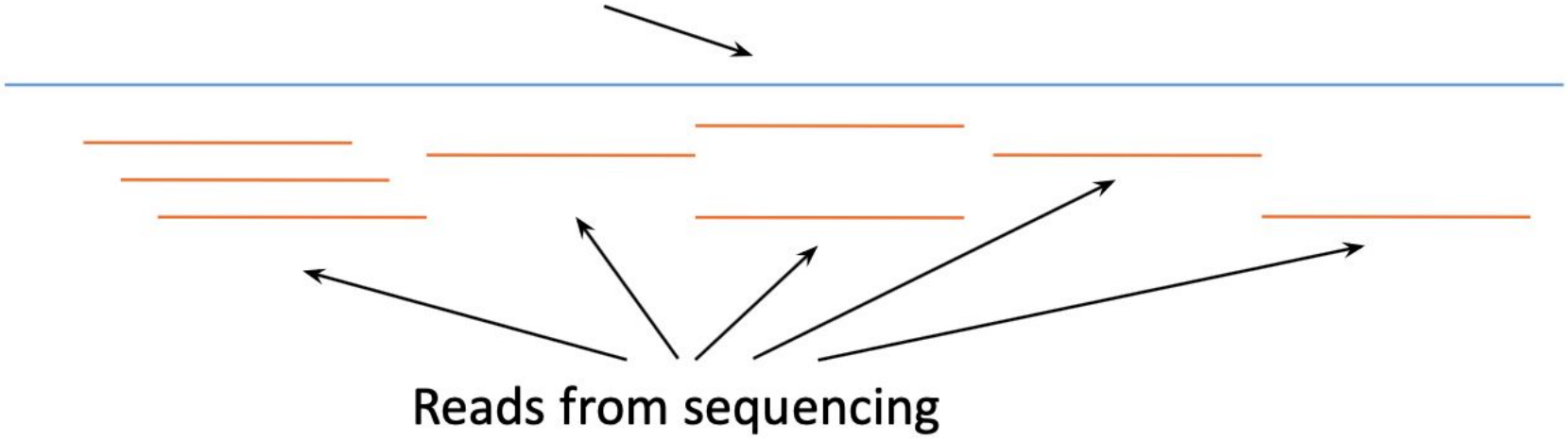
- **Read** – Substrings of genomic sequence output by sequencer

- **Sequencing coverage/depth** - average number of reads representing a given nucleotide in the reconstructed sequence



Next generation (DNA) sequencing

Massive amounts of data, produced very fast…



Read length = number of bases in each orange line

Throughput = number of orange lines produced by each NGS run

Next generation sequencing

Much higher degree of parallelism than Sanger sequencing

Much lower costs

Different platforms differ in terms:

- read lengths
- bp output
- costs of run
- costs of library preparation
- error rates

illumina

PACBIO®

Oxford NANOPORE Technologies

illumina

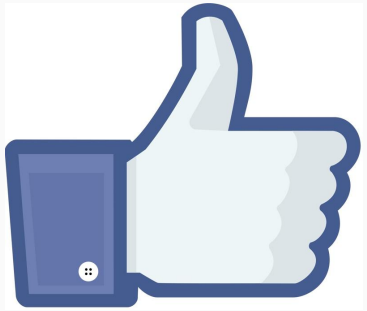
- High output; cheap (£/Mb); low error rate; short reads
- Many tools
- Low Insertion/deletion errors
- de novo genome or transcriptome sequencing, re-sequencing, GBS/capture sequencing

PACBIO®

Oxford NANOPORE Technologies

- 10+ kb reads
- de novo genomes, scaffolding
- Lower quality than Illumina, although has improved
- Can span over long repeats
- De novo assembly, detection of structural variants
- Transcriptomics: full-length transcripts/isoforms
- Epigenetics

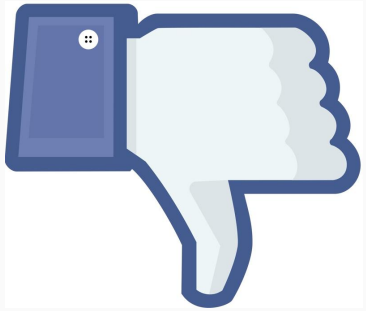
SNPs Benefits



Reference	ACTGACGCATGCATCATGCATGC
SNP	ACTGACGCATGCATCATTCATGC

- Easier to identify and genotype thanks to advances in sequencing technology
- Genome wide
- Co-dominant, well-defined mutation model
- Can genotype in large numbers, so more powerful than fewer but more allelic-rich markers
- Can work with degraded samples

SNPs Disadvantages

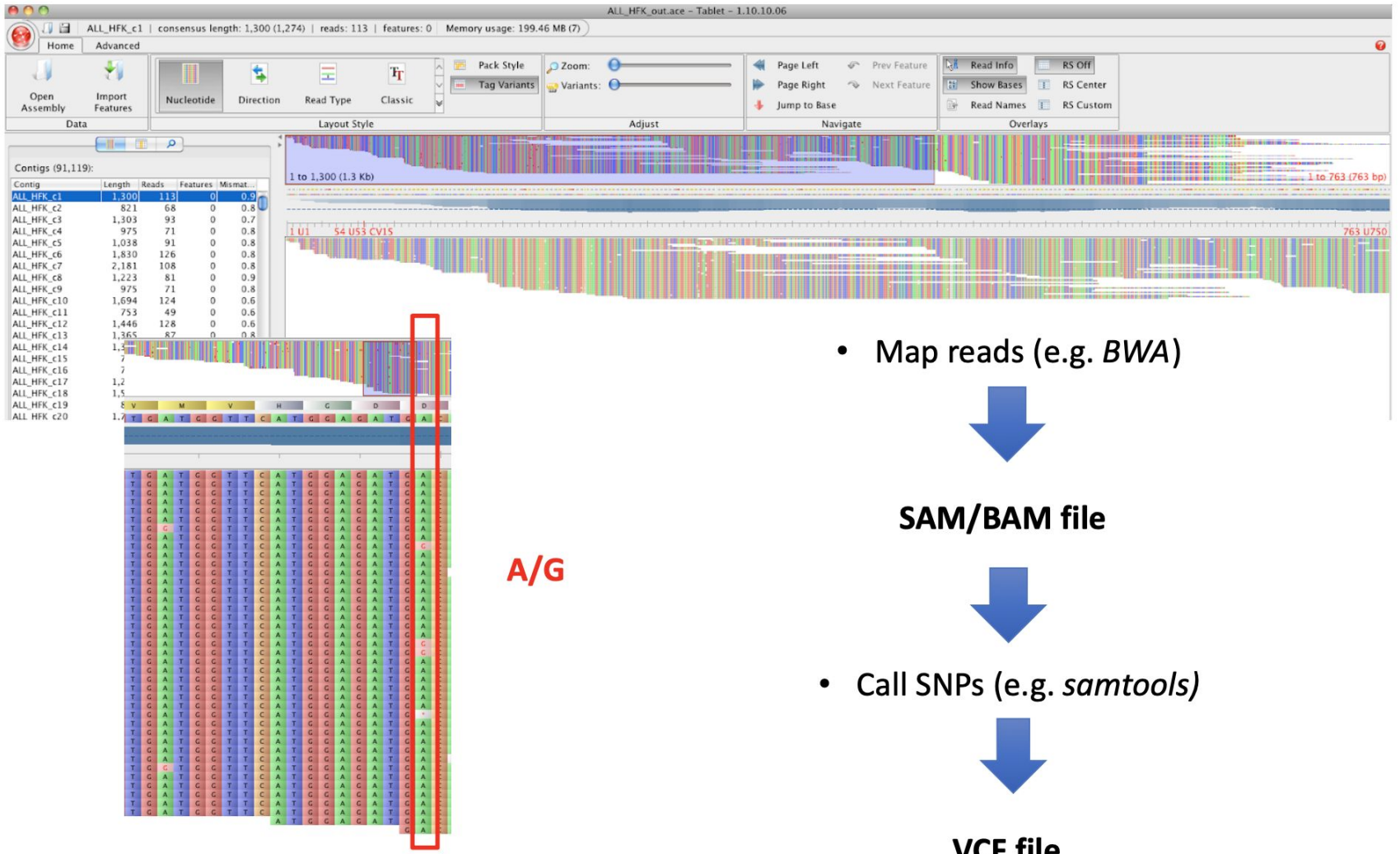


- Can still be quite expensive to generate (depending on the size of the project)
- Need to avoid ascertainment bias - SNP genotyping arrays (use a good number of diverse samples)
- Time-consuming, analysis often requires command line tools and computing server rather than personal computers

generating SNPs

- Genome re-sequencing – can be expensive! Limits number of samples
- GBS/RAD/ddRAD, etc – ‘reduced representation’ = more samples for the same amount of sequencing
- Transcriptome sequencing
- Sequence capture / genome enrichment for known regions / known SNP genotyping – all require prior sequence knowledge to design, e.g. myBaits
- Sequence capture: can use lower quality DNA

Generating SNPs from sequence data - Basic concepts



A/G

- Map reads (e.g. *BWA*)

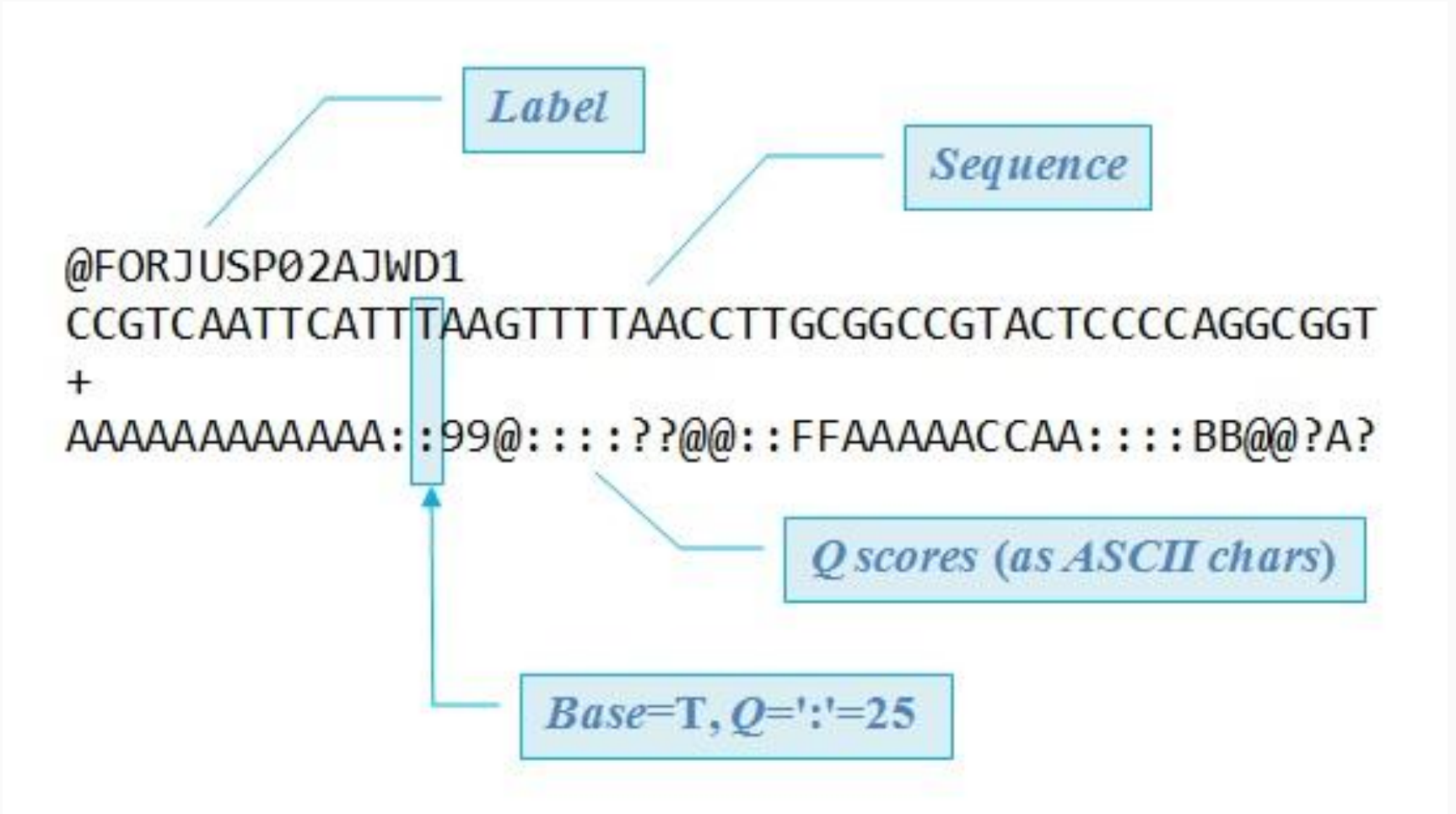
SAM/BAM file

- Call SNPs (e.g. *samtools)*

VCF file

Tablet alignment viewer: https://ics.hutton.ac.uk/tablet/

Fastq files – raw sequence data



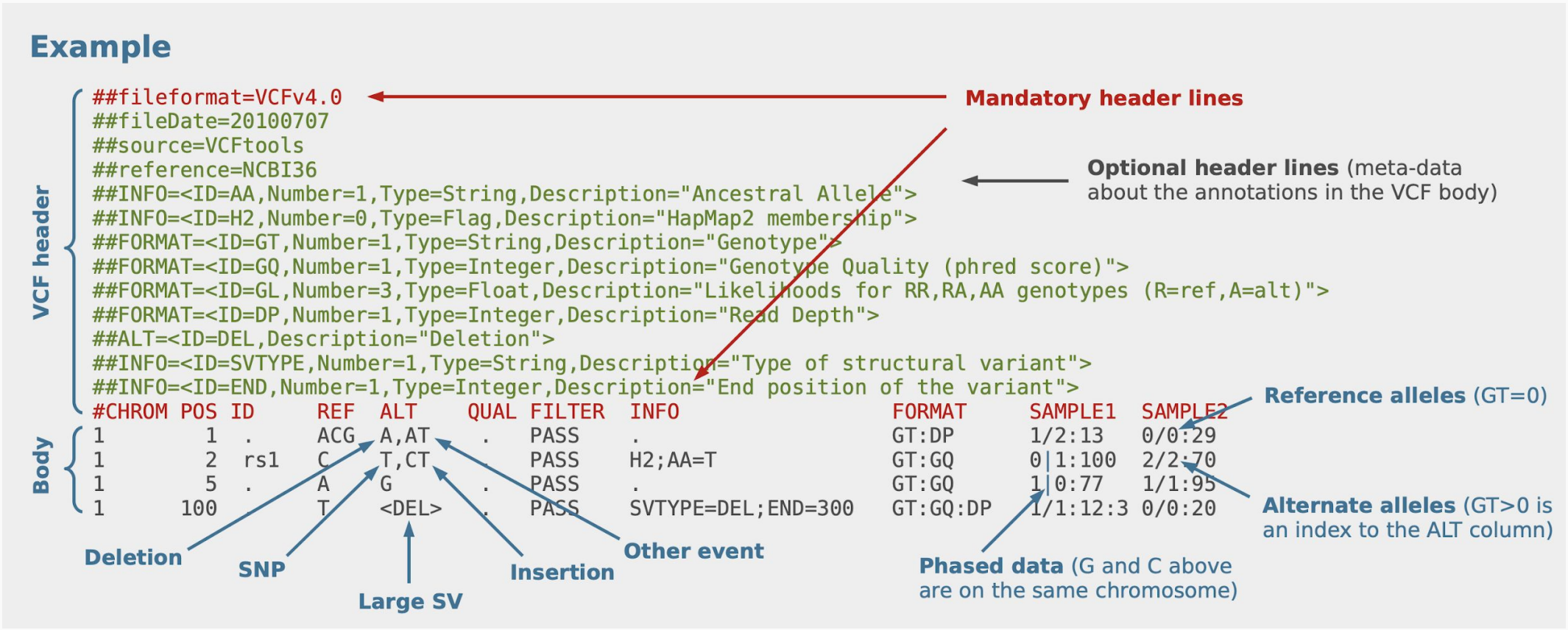
https://drive5.com/usearch/manual/fastq_files.html

SAM/BAM files – alignment data

```
HISEQ:131:C7HUDANXX:1:1101:4678:1130 1   163 2   Hmel216016 3    194941 4 60 5     125M 6  = 7     194952 8 136 9   AGCTCCTCGA
GAATGGAATAAAGGATCAGCCATCAGGCAAGCTGGCCAAGAGAGGTATTATCTCATCTGGTCTGTCTCATGGTATTGGCTTGTCACATGGTATTGGTGTGTCCCATGGCATCGGT     CC
CCCGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGFGGGGGGGGGGGGGGGGGFGGGEGGGGGGGGGGGGGGGGGGGG
G       XT:A:U  NM:i:1  SM:i:37 AM:i:37 X0:i:1  X1:i:0  XM:i:1  XO:i:0  XG:i:0  MD:Z:124C0
```

Column	Name	Description
1	QNAME	Query sequence name
2	FLAG	A bitwise FLAG representing things about the read, e.g. if mapped and/or paired
3	RNAME	Reference contig name
4	POS	Mapping start position
5	MAPQ	Mapping quality
6	CIGAR	Describes which positions match or any insertions or deletions compared to the reference
7	RNEXT	Reference contig name for the paired read
8	PNEXT	Position for the paired read
9	TLEN	Template sequence length
10	SEQ	Sequence
11	QUAL	Sequence base qualities

VCF files

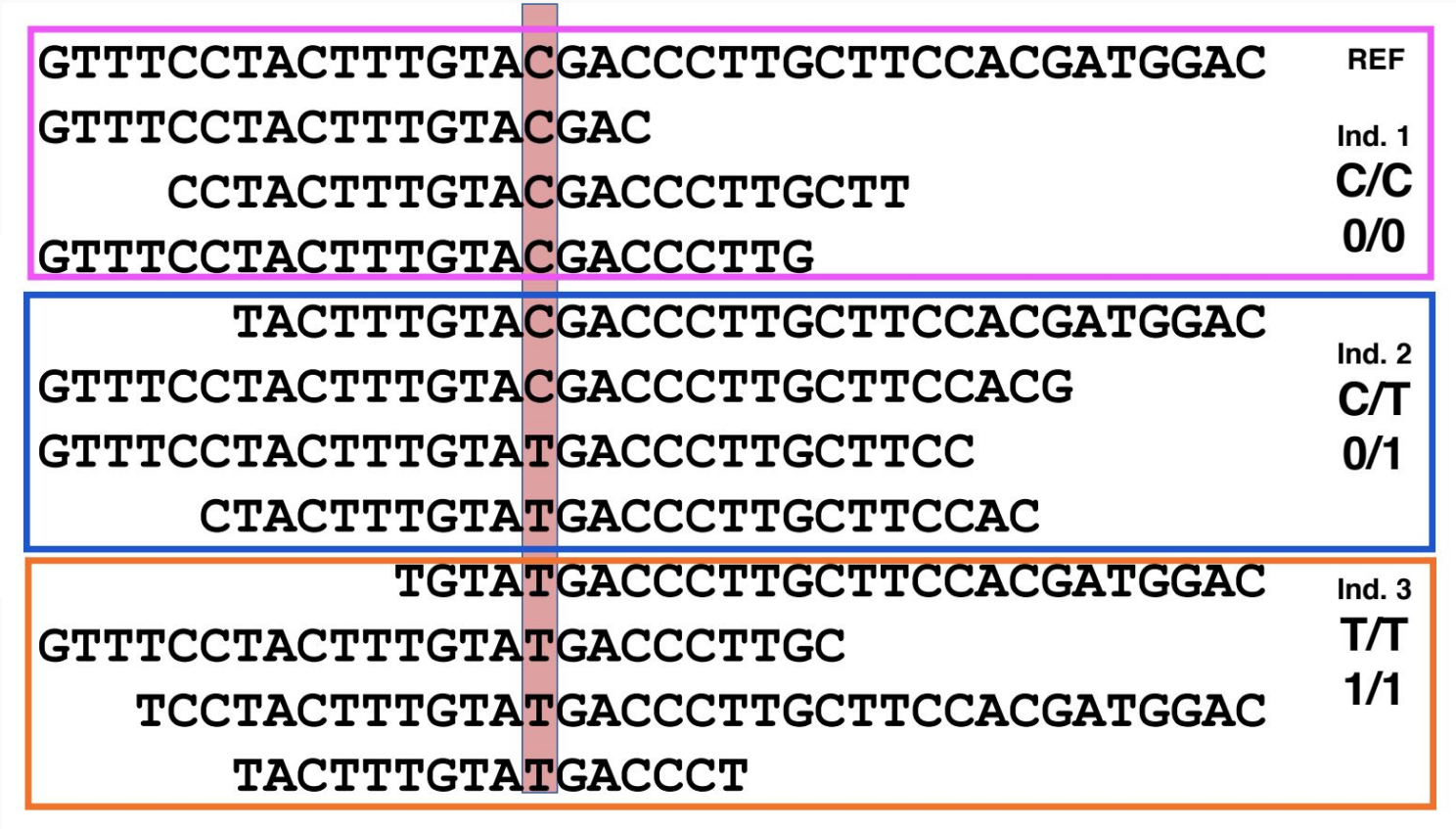


http://vcftools.sourceforge.net/VCF-poster.pdf

VCF files

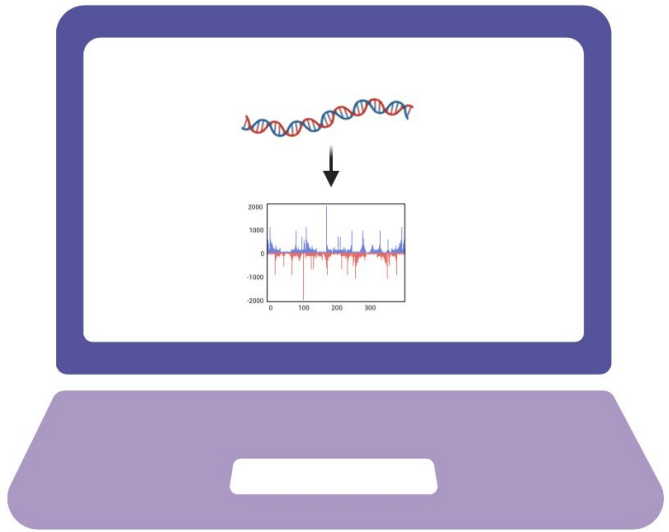
SNPs

Alignment	*VCF representation*		
ACGT	POS	REF	ALT
ATGT	2	C	T



Analysis

- Population genetics/genomics: structure, gene flow, pop history
- Assignment: species/hybrids, parentage/relatedness, loci involved in adaptation
- Mapping: link genotype & phenotype. QTL, GWAS. Functional annotation

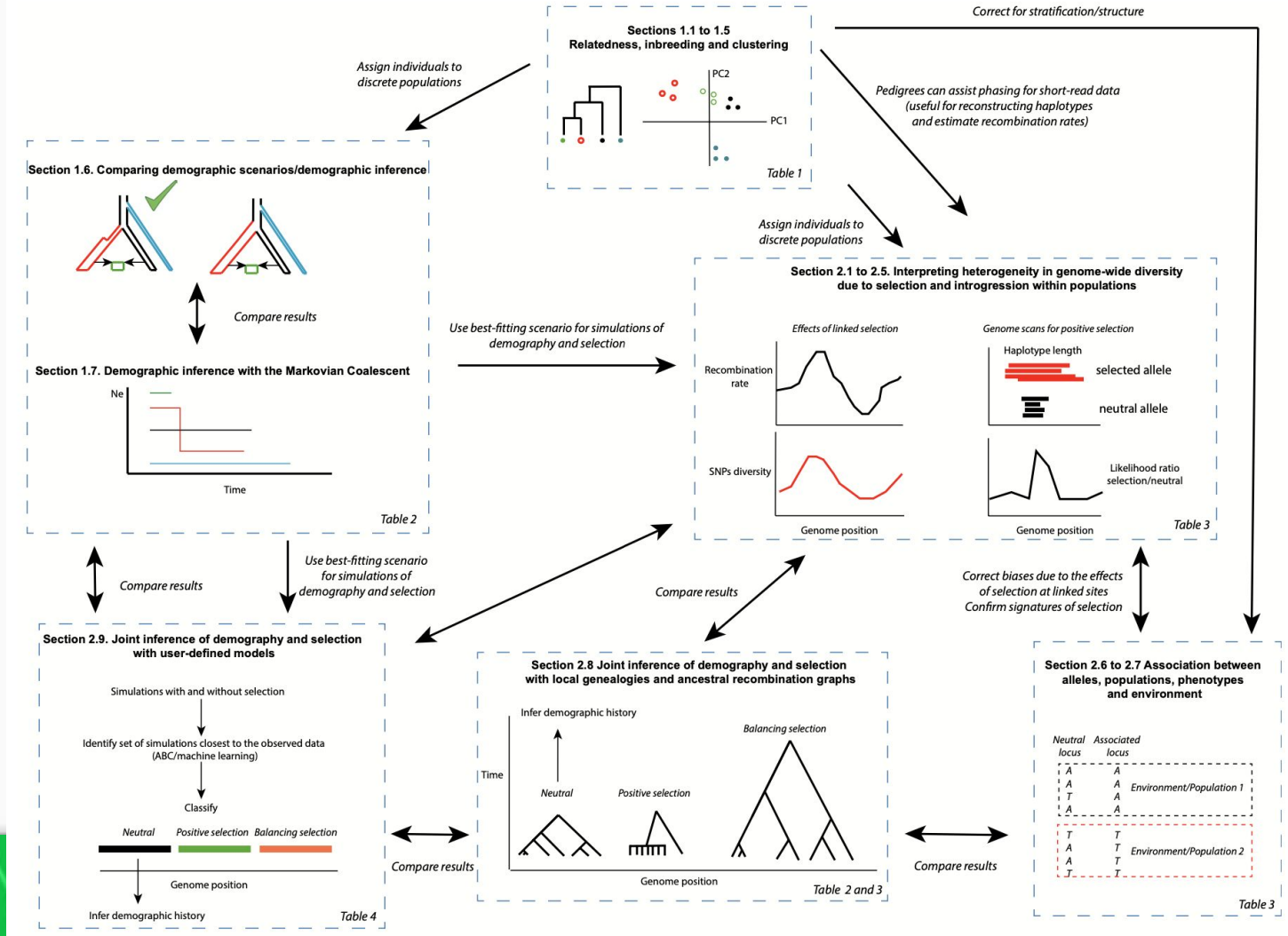


SPECIAL ISSUE

MOLECULAR ECOLOGY WILEY

An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes

Yann X. C. Bourgeois[1] | Ben H. Warren[2]



SPECIAL ISSUE

MOLECULAR ECOLOGY WILEY

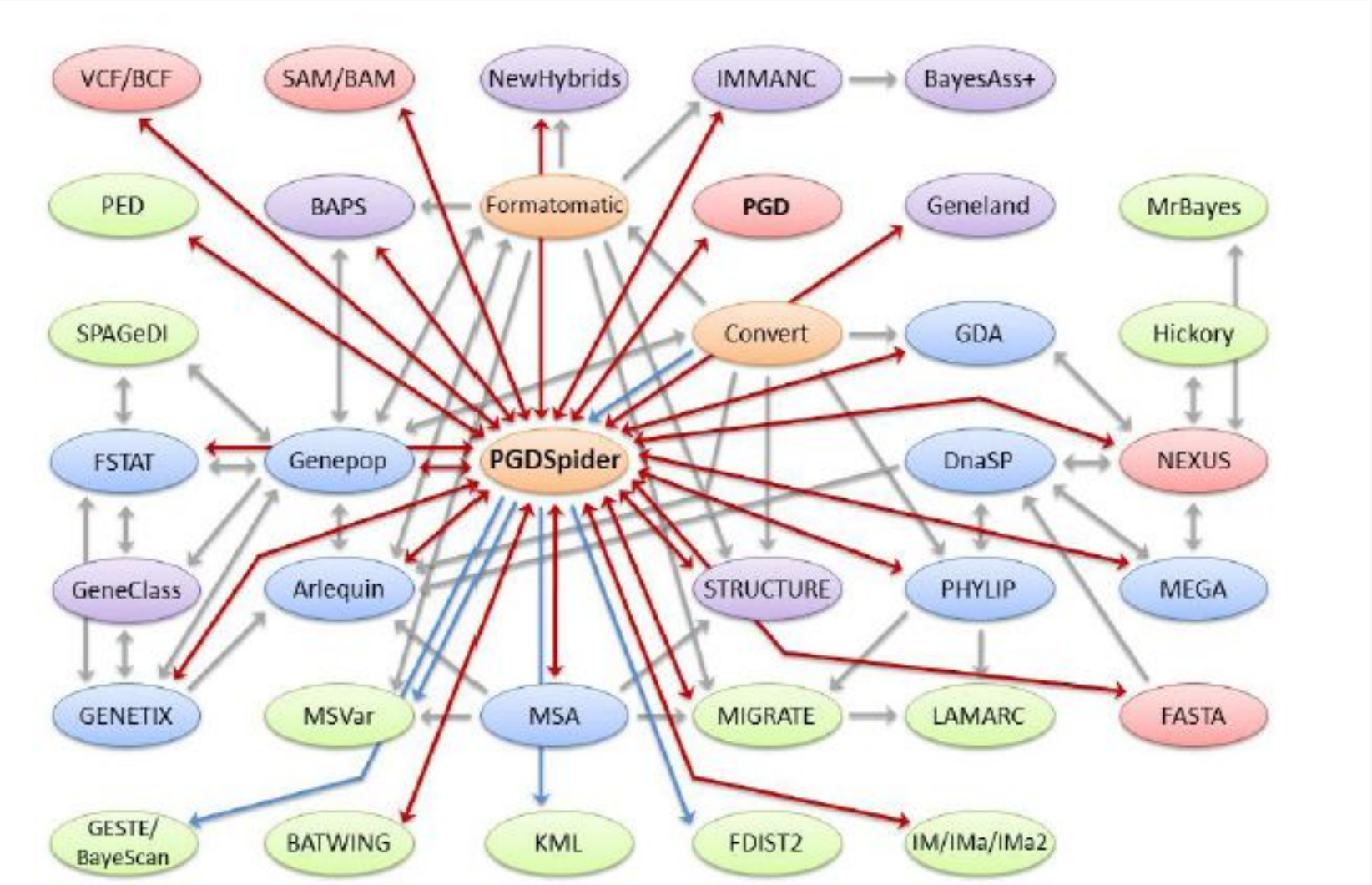
An overview of current population genomics methods for the analysis of whole-genome resequencing data in eukaryotes

Yann X. C. Bourgeois[1] | Ben H. Warren[2]

TABLE 1 Summary of methods dedicated to data description and assessing population structure. VCF: variant call format (see Danecek et al., 2011)

Software	Class of method	Purpose	Specifics	Issues and warnings	Link	Reference
SNMF	Clustering and characterizing admixture	Grouping individuals in clusters maximizing Hardy–Weinberg (HW) equilibrium and LD between loci	Fast (30× than ADMIXTURE)	Still slow computation time for very large data sets	http://membres-timc.imag.fr/Olivier.Francois/snmf/index.htm	Frichot et al. (2014)
STRUCTURE	Clustering and characterizing admixture	Grouping individuals in clusters maximizing HW equilibrium and LD between loci	User-friendly interface. Bayesian inference	Not suited for large whole genomes. Requires specific input format. Might be used on a small set of high-quality markers for small genomes	http://pritchardlab.stanford.edu/structure.html	Pritchard et al. (2000)
FASTSTRUCTURE	Clustering and characterizing admixture	Grouping individuals in clusters maximizing HW equilibrium and LD between loci	~100× faster than STRUCTURE	Approximate inference of the original STRUCTURE model	http://rajanil.github.io/fastStructure/	Raj et al. (2014)
ADMIXTURE	Clustering and characterizing admixture	Grouping individuals in clusters maximizing HW equilibrium and LD between loci	Maximum likelihood, faster than STRUCTURE. Can handle sex-linked markers	Often slower than its counterparts	https://www.genetics.ucla.edu/software/admixture/index.html	Alexander and Novembre (2009)
FINESTRUCTURE/GLOBETROTTER	Clustering and characterizing admixture	Chromosome painting, admixture and clustering	Estimates time since admixture, fast, set of scripts to facilitate analysis	Relies on STRUCTURE and FASTSTRUCTURE assumptions. Requires phased data	http://paintmychromosomes.com/	Hellenthal et al. (2014)
PCADMIX	Clustering and characterizing admixture	Chromosome painting	Fast, uses HMM to smooth out windows and limit noise due to low-confidence ancestry	Requires a priori definition of ancestral populations and phased haplotypes	https://sites.google.com/site/pcadmix/	Brisbin et al. (2012)

PGDSpider – data conversion tool



http://www.cmpg.unibe.ch/software/PGDSpider/

illumina

We will be working with different datasets generated using Illumina Sequencing for the two days of the workshop

Tuesday

- You will generate VCF files from raw sequencing data
  - Fastq -> BAM -> VCF
- Analysis
  - Plotting genetic structure
  - Summary stats

Thursday

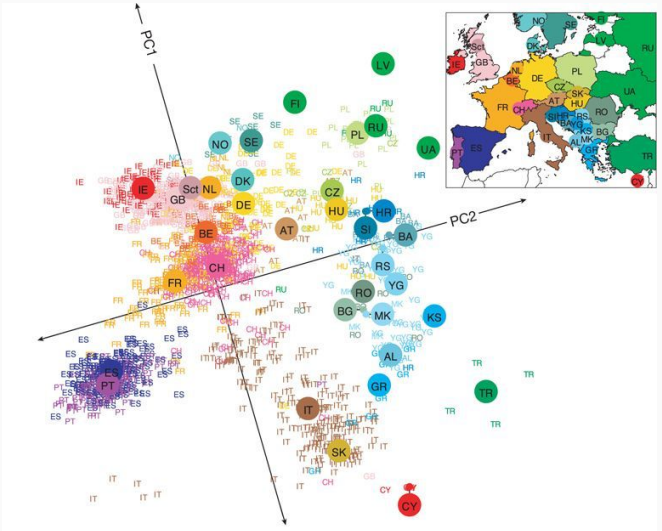
Analysis continued…

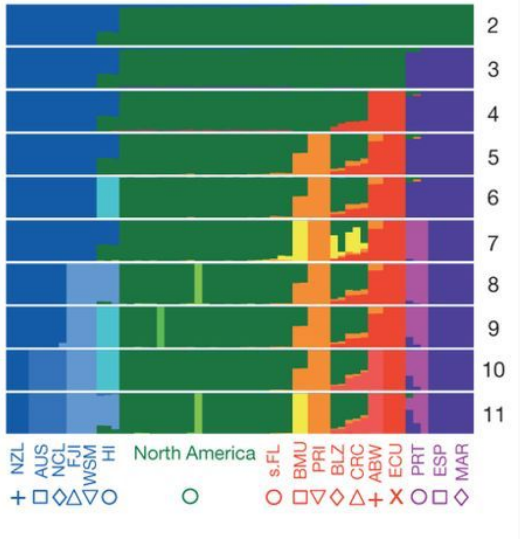
- Detecting selection
- GWAS

Methods to infer structure, ancestry and admixture

Model free – PCA (EIGENSOFT, PCAdmix, prcomp in R)
Model based – Maximum likelihood (e.g. NGSadmix - Expectation-Maximization algorithm) or Bayesian models (e.g. STRUCTURE - Markov chain Monte Carlo)

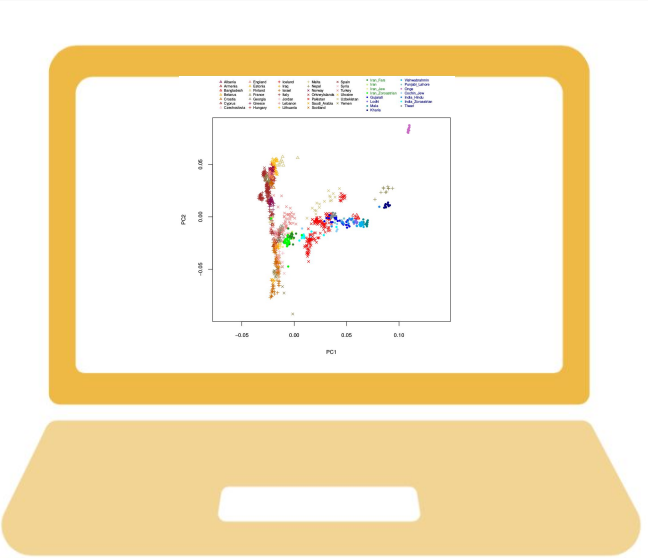
Genotype calls (ADMIXTURE, STRUCTURE, fastSTRUCTURE)
Genotype likelihoods for NGS data (NGSadmix, entropy)





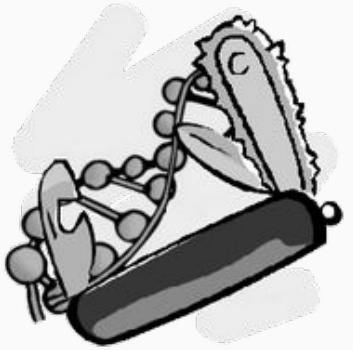
Software

EIGENSOFT – Patterson et al. (2006)
https://www.hsph.harvard.edu/alkes-price/software/
PCAdmix – Brisbin et al. (2012)
https://sites.google.com/site/pcadmix/
prcomp – François et al. (2010), Rishinwar et al. (2015)
https://stat.ethz.ch/R-manual/R-devel/library/stats/html/prcomp.html
ADMIXTURE – Alexander et al. (2009)
http://software.genetics.ucla.edu/admixture/
NGSadmix – Skotte et al. (2013)
http://www.popgen.dk/software/index.php/NgsAdmix
STRUCTURE – Pritchard et al (2000)
https://web.stanford.edu/group/pritchardlab/structure.html
fastSTRUCTURE – Raj et al. (2014)
https://rajanil.github.io/fastStructure/
Entropy – Gompert et al. (2014)
https://datadryad.org/resource/doi:10.5061/dryad.pq93h



PopGenome: An Efficient Swiss Army Knife for Population Genomic Analyses in R

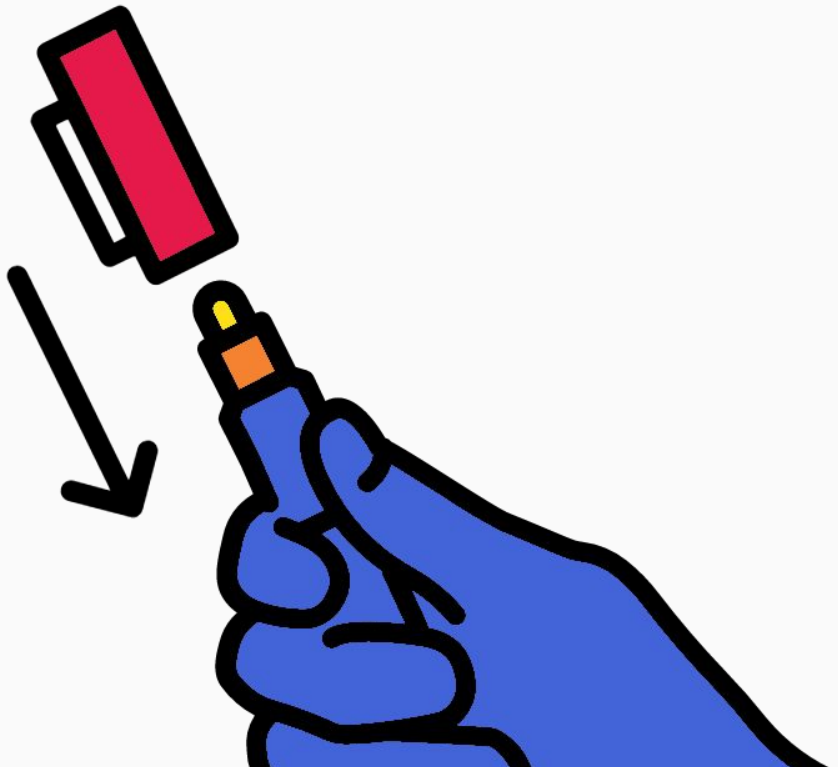
Bastian Pfeifer,[1] Ulrich Wittelsbürger,[1] Sebastian E. Ramos-Onsins,[2] and Martin J. Lercher*[,1,3]



- An Efficient Swiss Army Knife for Population Genomic Analyses in R
  - Read data in a variety of input formats
  - Implement a comprehensive range of population genetics/genomics analyses and statistics
  - Read associated annotation files and allow to systematically select regions of interest
  - Able to analyse individual loci, multiple loci, and sliding windows
  - Open source and be easily extendable by the scientific community to incorporate new types of analyses
  - Integrated with powerful numerical and graphical capabilities
  - Platform independent

Recap

Sequencing technologies
SNP data
Common file types
Analysis



Reminders and Tips

Work at your own pace

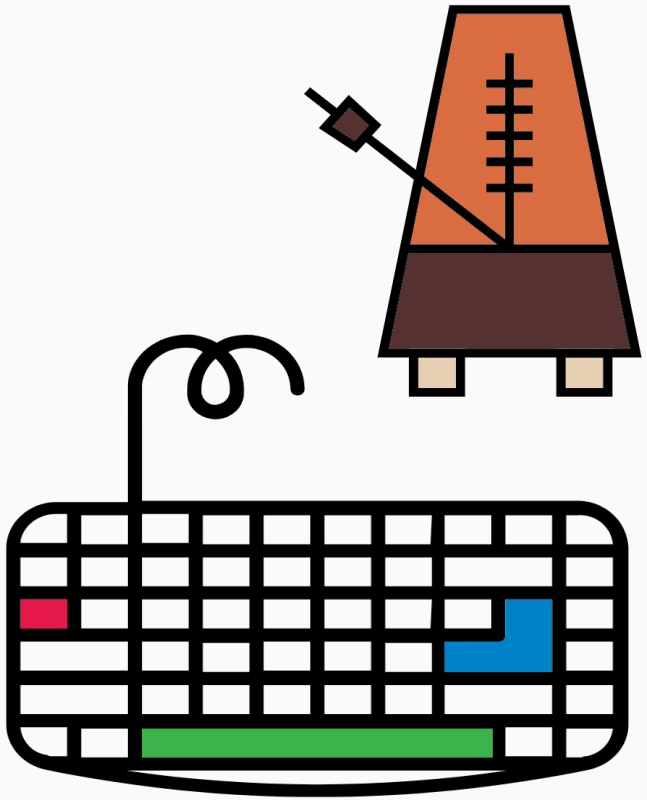
Typos

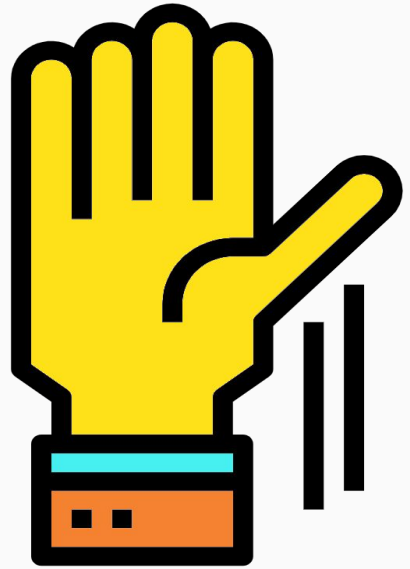
Ask questions

Breaks are important

Tab, space, and enter

Answers in back of bookdown





During sessions

Zoom - Ask via microphone if no question currently being asked/answered

Slack - Ask questions via the channel or ask to go into a zoom breakout room with one of us

WebVNC - We can connect to your webVNC to see and help with issues.

Breakout rooms upon request

Thank you!

Questions?