ANGSD

Analysis of Next Generation Sequencing Data

Major programmer: Thorfinn Korneliusen

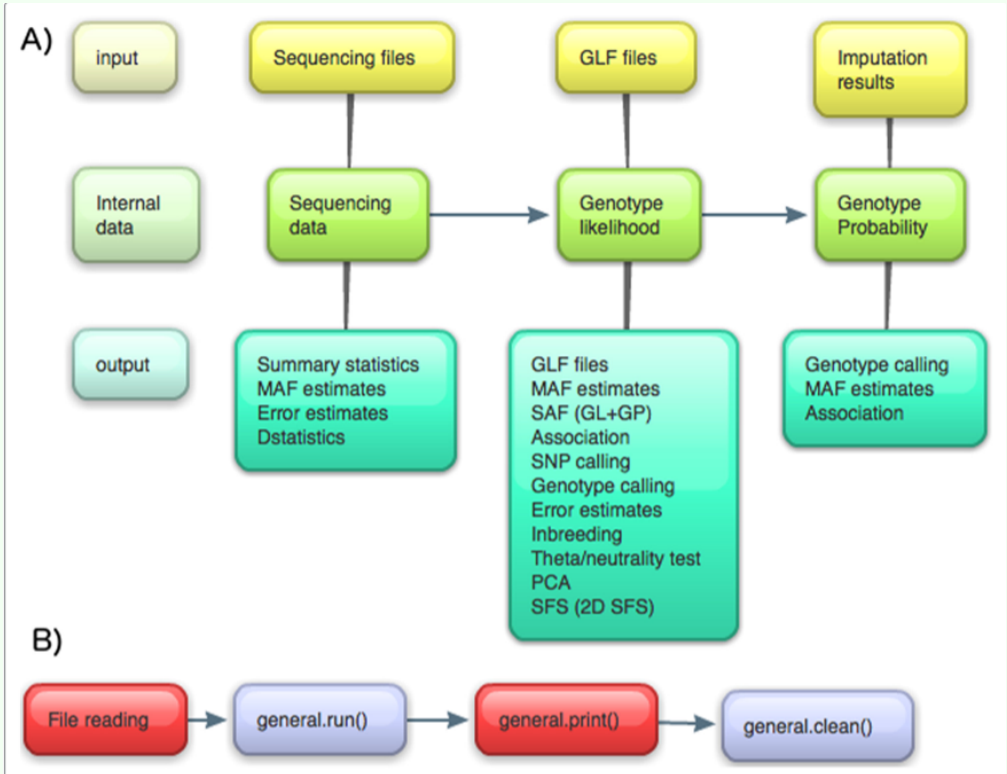


Figure 1 Data formats and call graph. A) Dependency of different data formats and analyses that can be performed in ANGSD. **B)** Simplified call graph. Red nodes indicate areas that are not threaded. With the exception of file readers, all analyses, printing and cleaning is done by objects derived from the abstract base class called **general**

Why ANGSD?

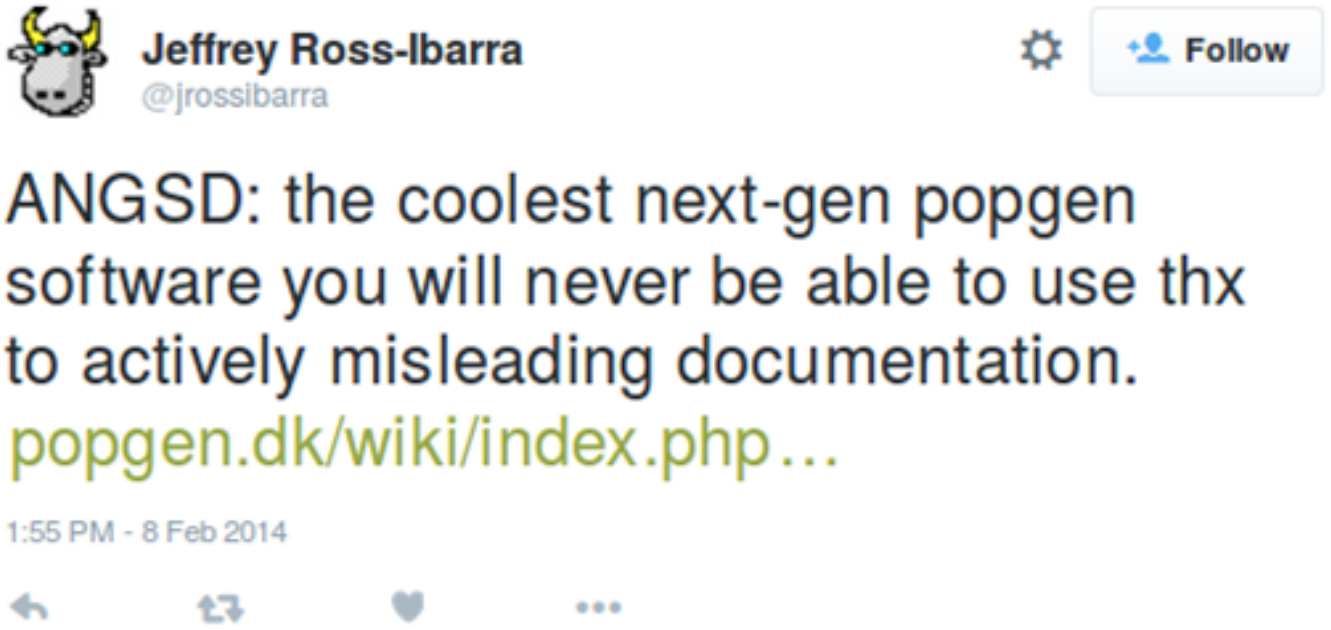
Focus

To perform population or medical genetic analysis on NGS data while taking uncertainly into account even for low depth data

- At the time no other software existed
- Most other NGS software are focused on genotype calling
- Useful as a research development tool
- Somewhat useful for others (not Anders/Thorfinn)

Great reviews from the scientific community

Twitter



They actually make a wrapper for ANGSD

`https://github.com/mojaveazure/angsd-wrapper`

Input and output

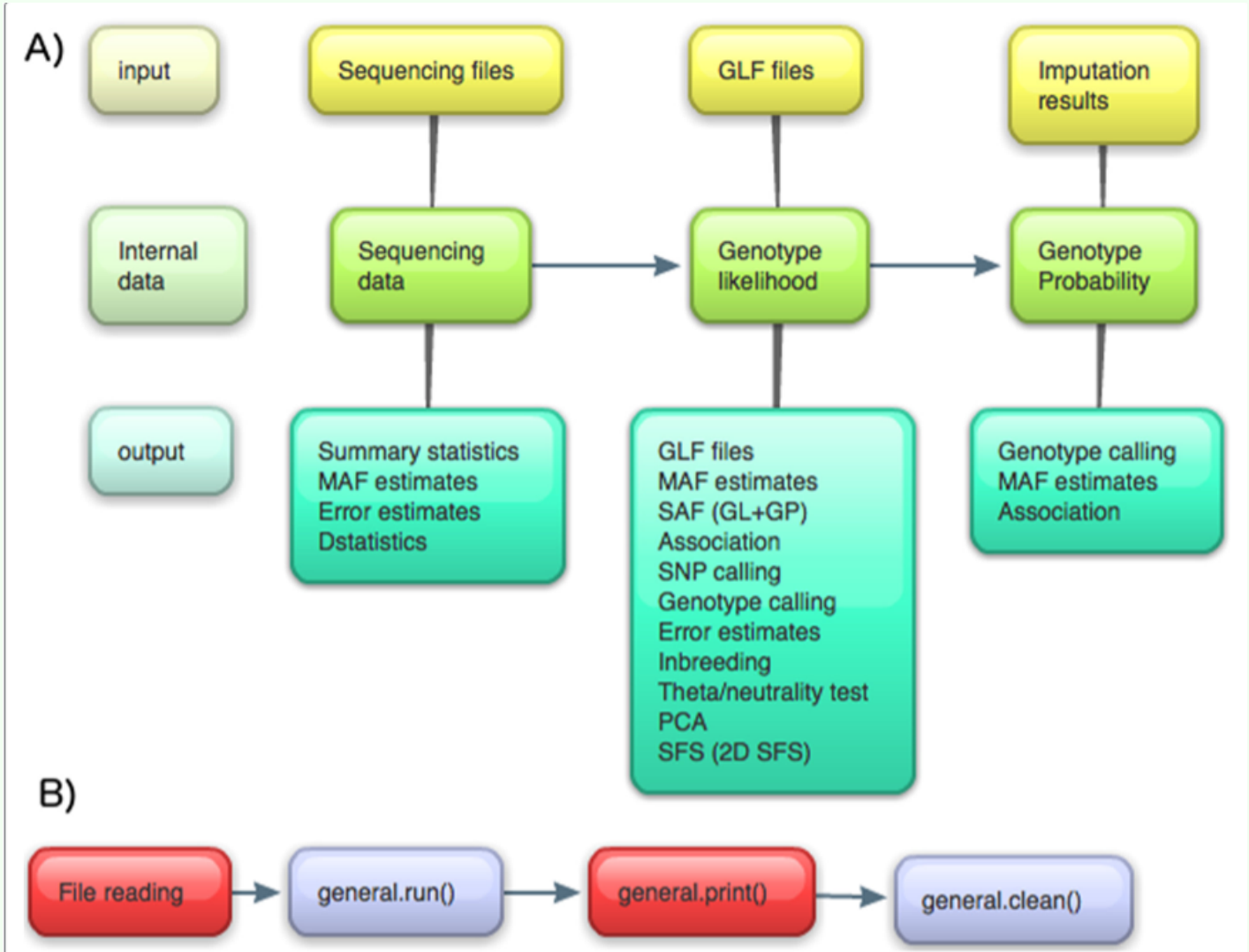


Figure 1 Data formats and call graph. A) Dependency of different data formats and analyses that can be performed in ANGSD. **B)** Simplified call graph. Red nodes indicate areas that are not threaded. With the exception of file readers, all analyses, printing and cleaning is done by objects derived from the abstract base class called `general`.

Input formats

Sequencing data

- Bam
- Cram
- mpileup

Example

BAM $->$ ANGSD $->$ BEAGLE $->$ ANGSD $->$ Association

Genotype likelihoods

- Beagle
- glfV3
- tglf
- others

Example

MSMS $->$ mpileup $->$ ANGSD $->$ SFS $->$ ∂a∂i

Genotype (posterior) probability

- Beagle

Example

Cram $->$ SNPtools $->$GL $->$ ANGSD $->$ NGSadmix

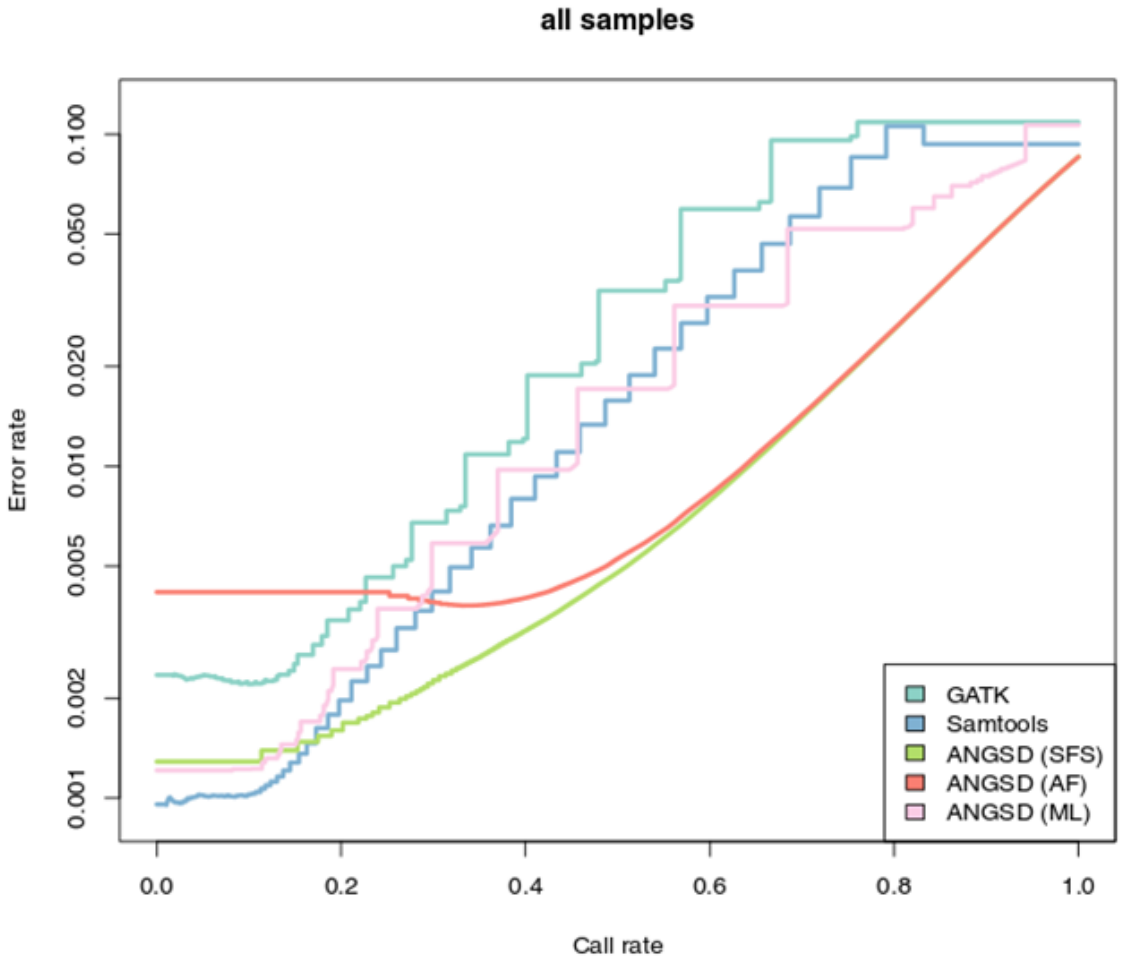
Analysis

Analysis	Basis	Reference
Contamination estimates based on the X-chromosomes	BC	[19][b]
Type specific error estimation estimated by simultaneously estimating allele frequencies and genotype likelihoods	GL	[10]
Type specific error estimation based on an outgroup and a high quality genome	BC	[20][ab]
Genotype likelihoods (GL) (diploids)	BC/Seq	[6,8,10,15]
Allele frequencies for a site	BC/GL/GP	[21][b] [10]
SNP discovery (LRT) used for rejecting that the allele frequency is different from zero	GL	[10]
Genotype posteriors (GP) can be used for calling genotypes by specifying a cutoff	GL/SAF	[9,10]
Sample allele frequencies (SAF) the probability of all read data given the sample allele frequency	GL/GP	[9][b]
Population differentiation statistics F_{st}	SAF	[14][ac]
Population structure via principle components analysis **(PCA)**	GP	[14][ac]
Admixture analysis (NGSadmix) NGS data	GL	[22][ab]
Detection of ancient admixture **ABBA-BABA/d-statistics**	BC	[20][b]
Estimation of **SFS (1D)**	SAF	[9][ab]
Estimation of **SFS (2D)**	SAF	
Selection scans, Neutrality tests (e.g θ's and Tajima's D)	SAF	[12][ab]
Estimation of individual and site-wise **Inbreeding** coefficients. Also MAF and GP estimation for inbreed individuals	GL	[13][abc]
Allele frequency based association for case/control data)	GL	[10]
Association score test in a generalized linear model framework for both quantitative and case/control data while allowing for additional covariates	GL-GP	[11][b]

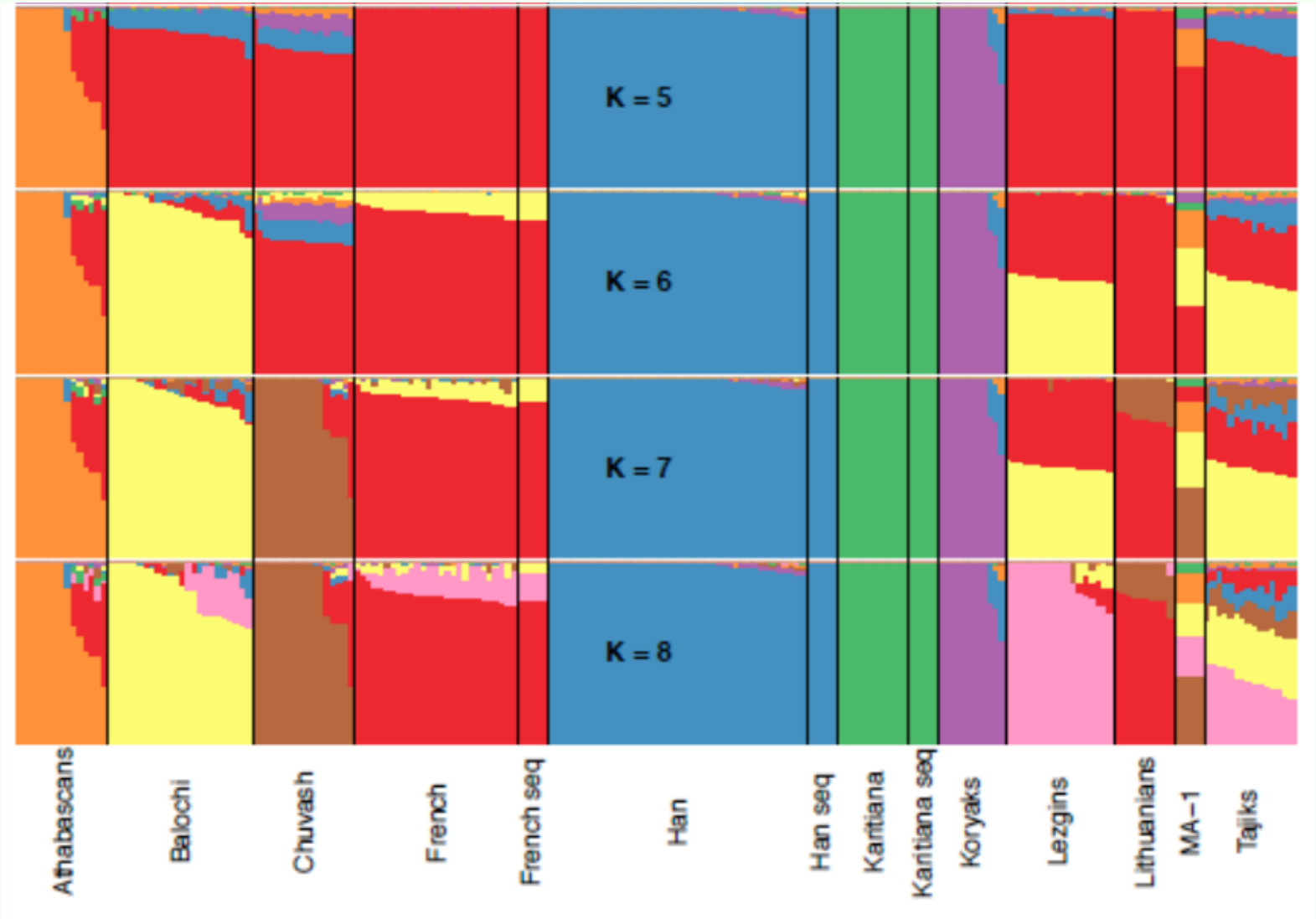
Where ANGSD does less well

- freeBayes/GATK/Samtools are better at SNP calling and genotype calling
- ANGSD does not including indels in ANY analysis

Its not bad there are just better options

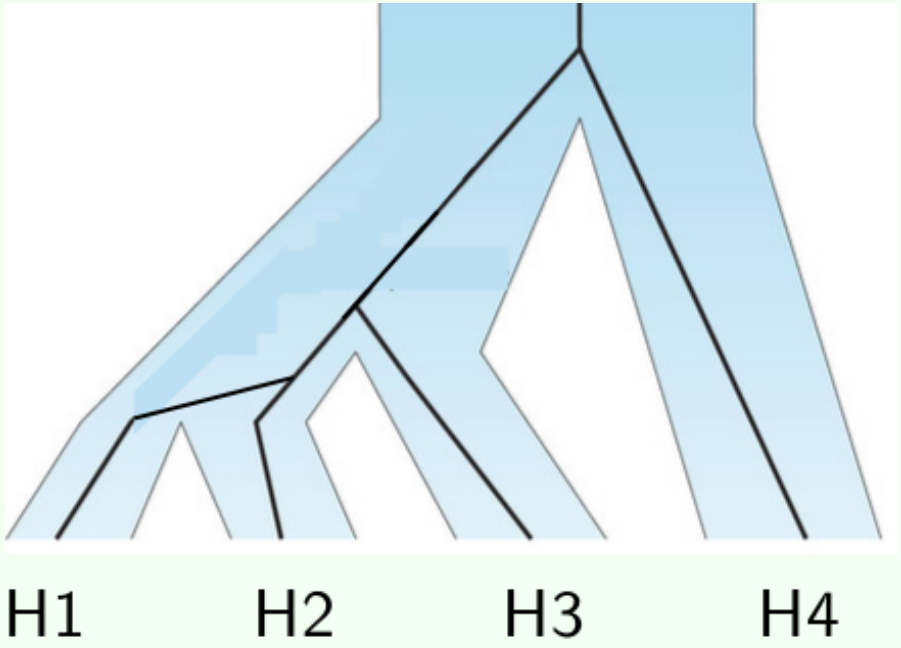


Common use - NGSadmix

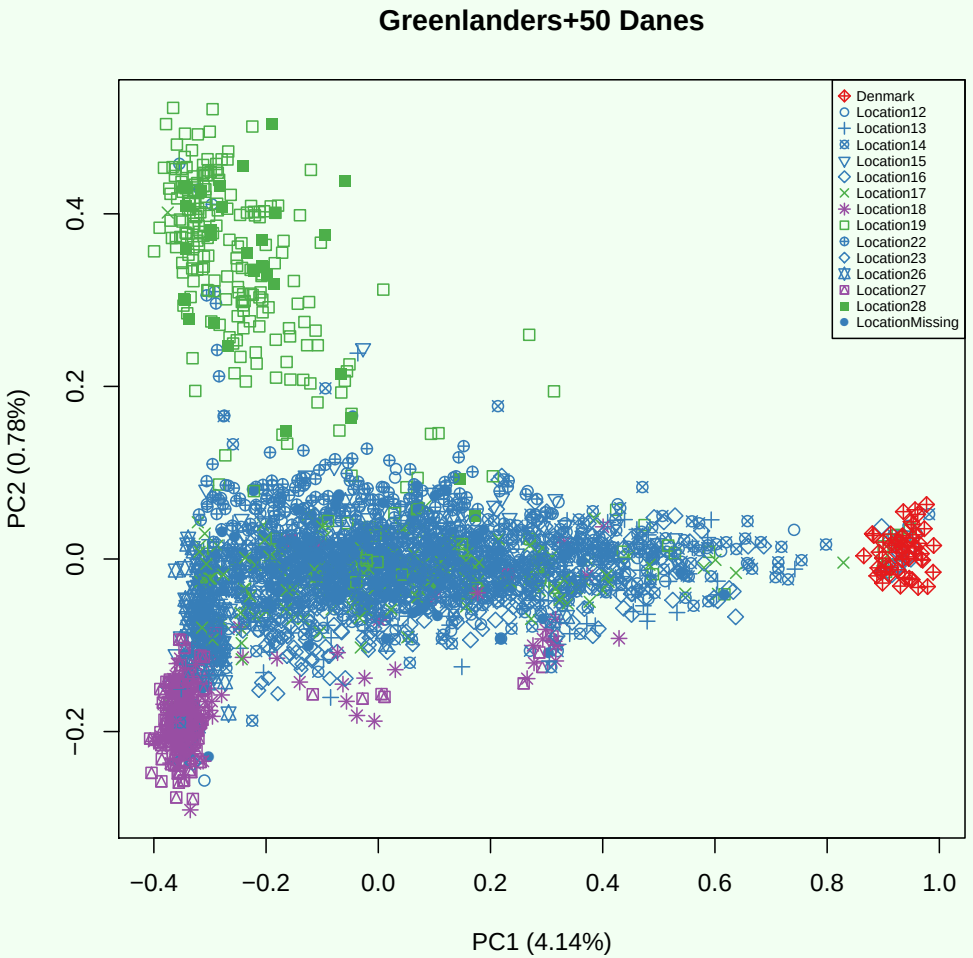


[1] Raghaven *et. al* Nature 2014

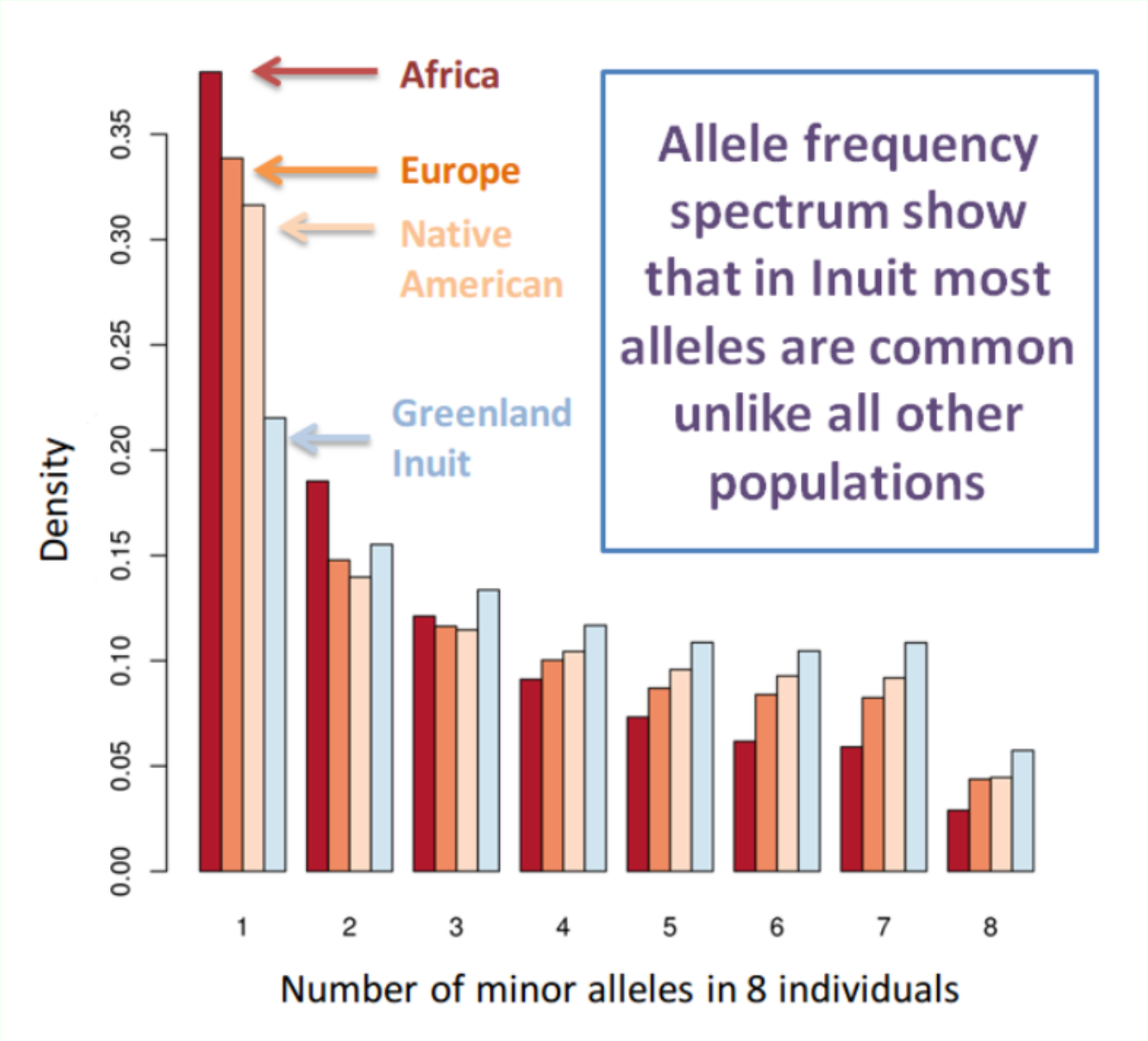
Common use - D-stat/ABBABABA



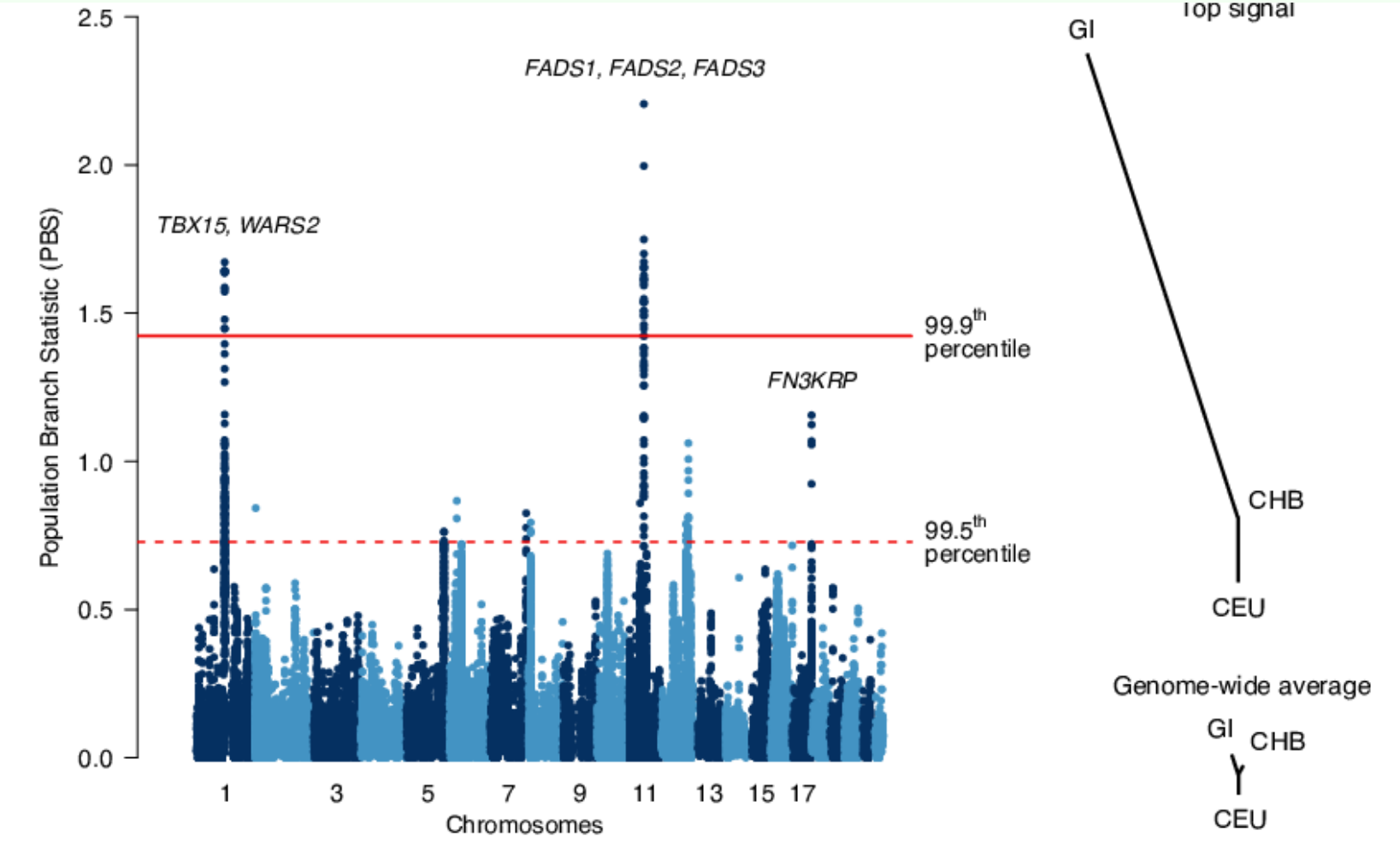
Common use - MDS/PCA



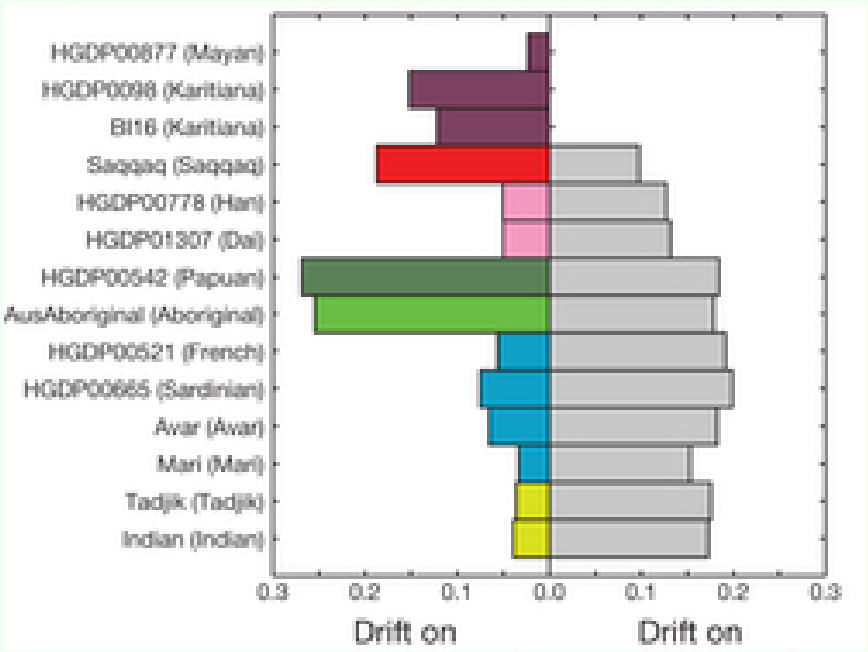
Common use - SFS

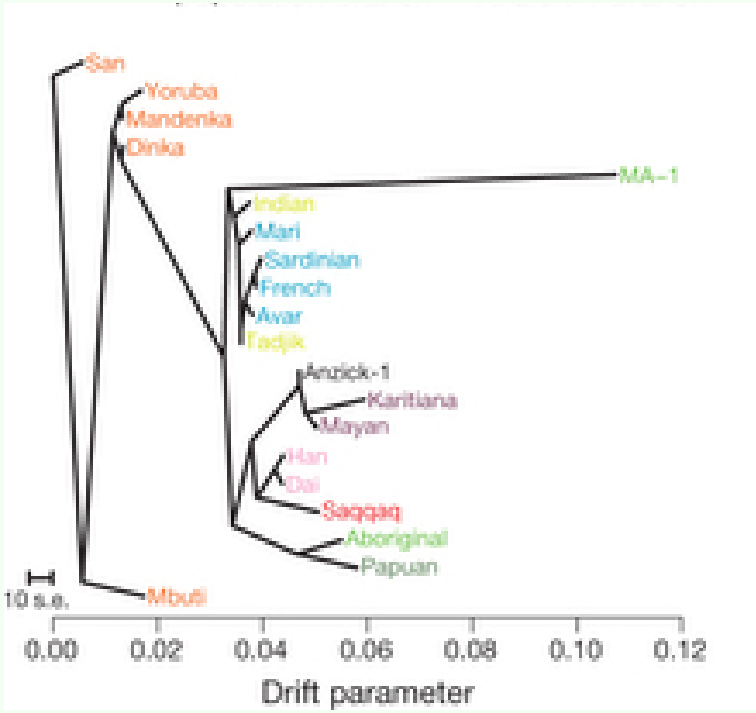


SFS - selction scans - theta/Tajima/Fst/PBS



SFS - test for continuation





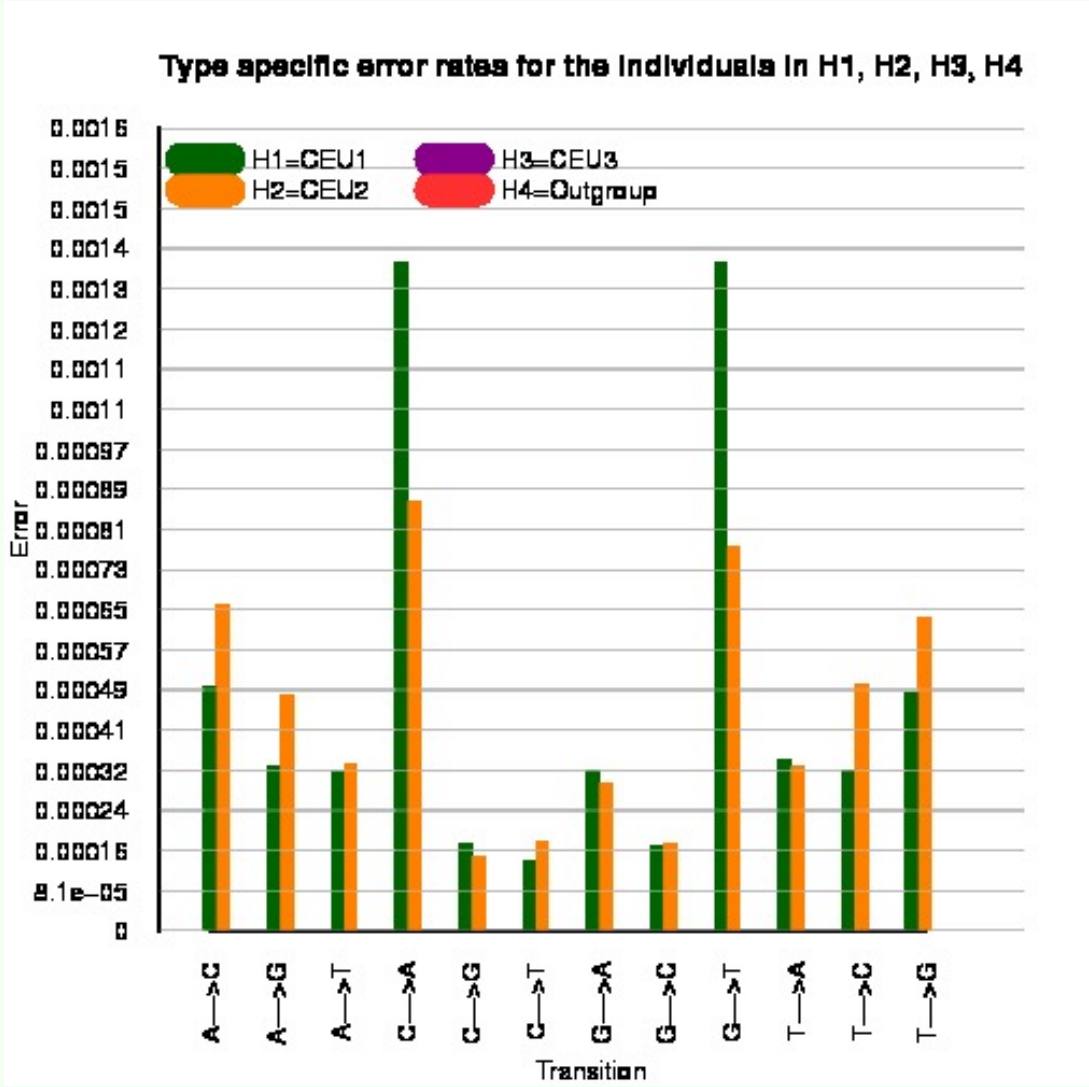
Conclusion

The ancient clovis native american is a direct ancestor to most modern native americans

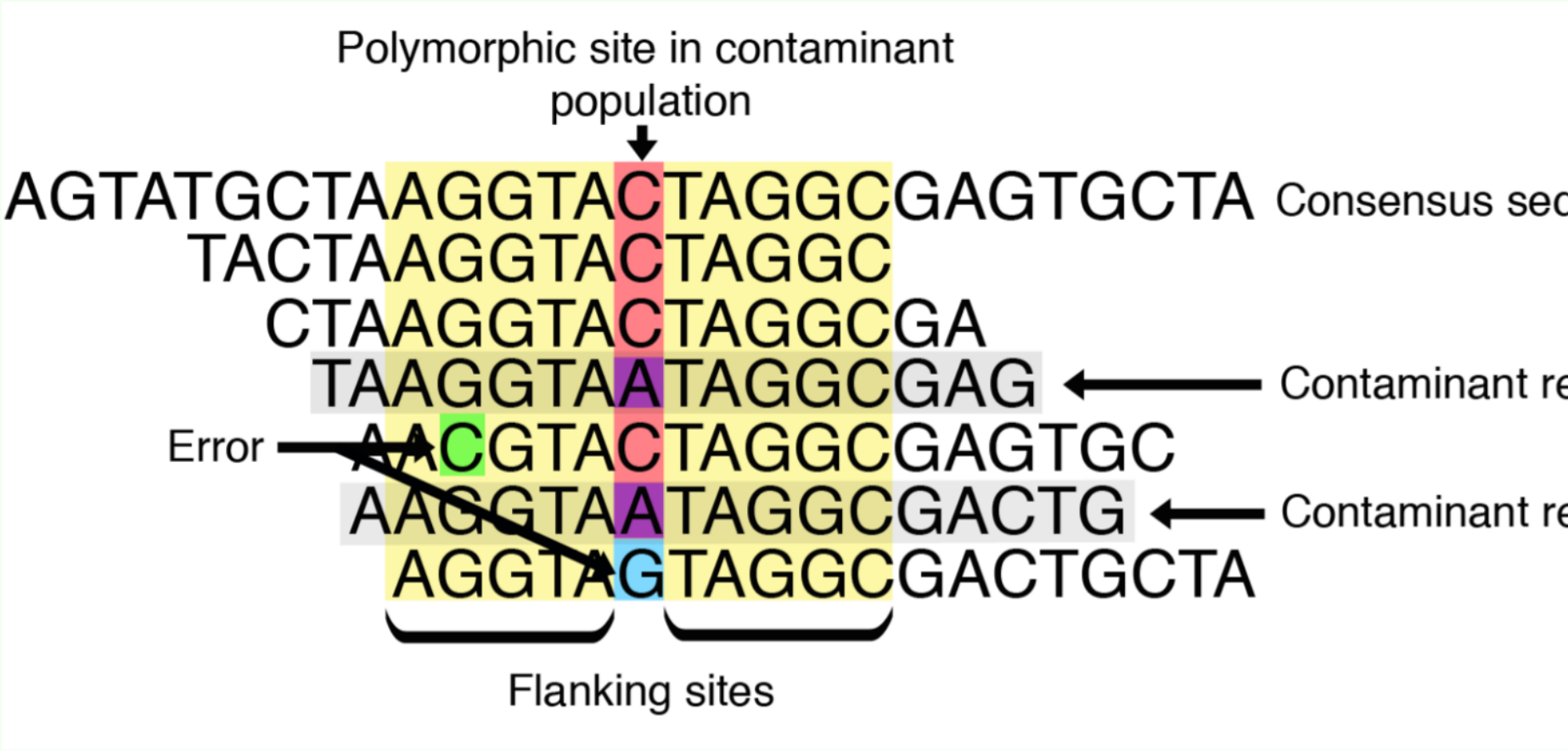
2

[2] Rasmussen et al. Nature 2014

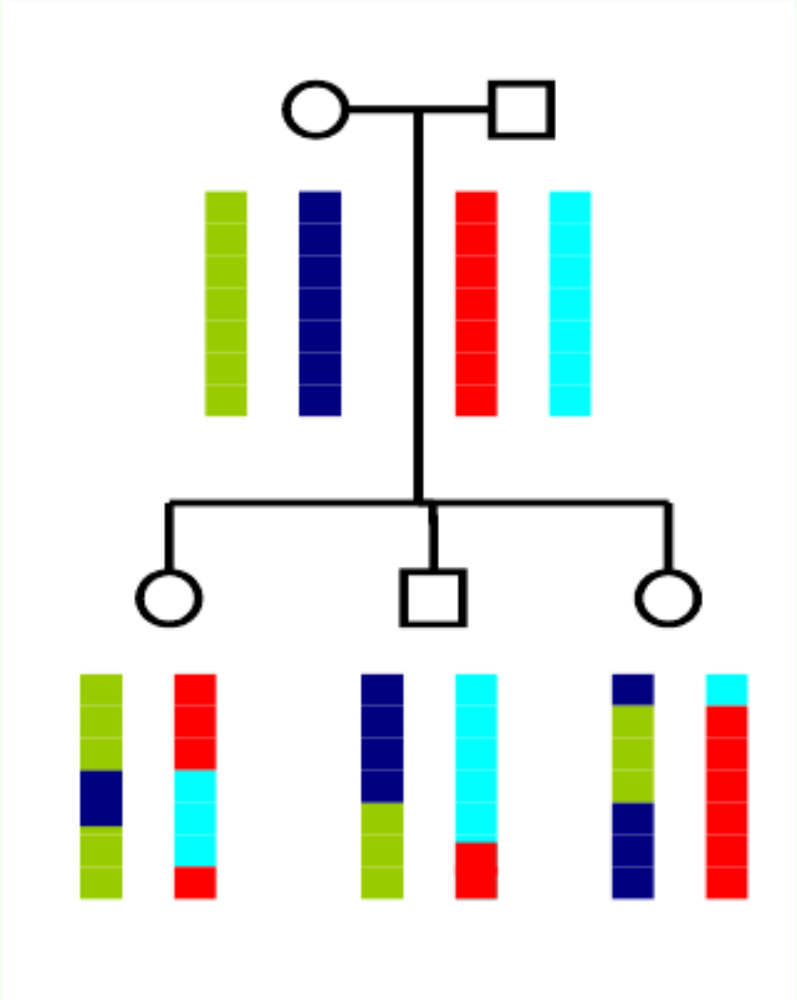
Common use - Error rate estimation



Common use - contamation



Common use - relatedness



exercises

Data from 1000 Genomes

- 2500 individuals sequenced at low/medium depth (3-8X)
- mulitple populations

Reduced genomes for admixture/pca

- 22 100k regions (one for each autosome)
- 50,000 SNP genotype likelihoods (multiple pop)
- 100,000 SNP genotype likelihoods (europeans)

Reduced genomes for SFS

- 22 100k regions (one for each autosome)
- 1Mb region on chr5
- 3 x 10 individuals from
- African(YRI), European (CEU), East Asian (JPT)