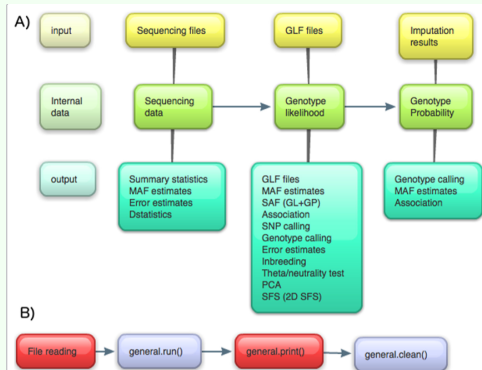


# ANGSD

## 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 clearing is done by objects derived from the abstract base class `AbstractObject`.

# Why ANGSD?

## Focus

To perform population or medical genetic analysis on NGS data while taking uncertainty 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



**Jeffrey Ross-Ibarra**

@jrossibarra



Follow

ANGSD: the coolest next-gen popgen software you will never be able to use thx to actively misleading documentation.  
[popgen.dk/wiki/index.php...](http://popgen.dk/wiki/index.php...)

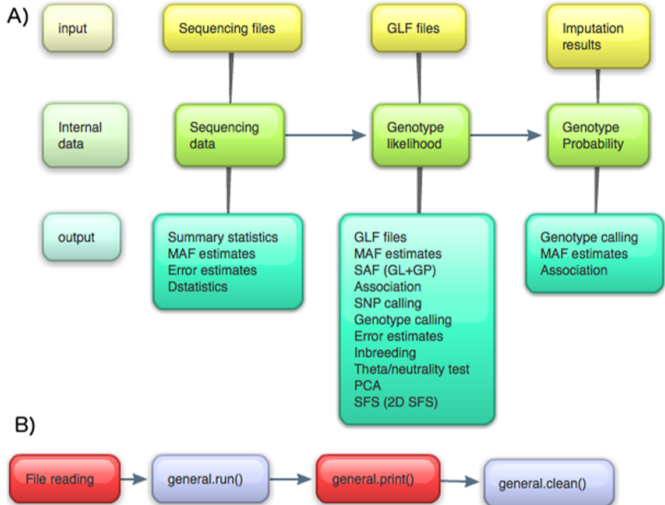
1:55 PM - 8 Feb 2014



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 →  
 $\partial a \partial i$

## Genotype (posterior) probability

- Beagle

## Example

Cram → SNPtools  
→ GL → ANGSD →  
NGSadmix

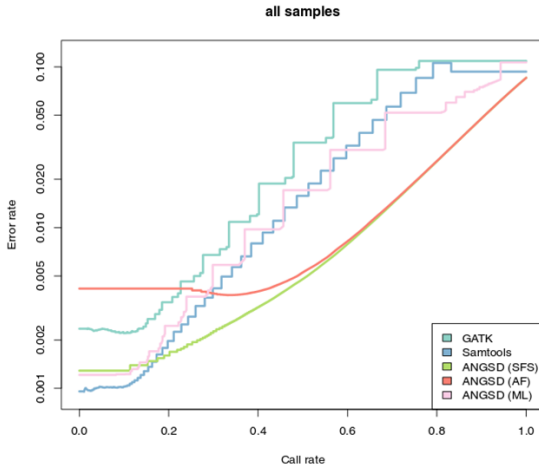
# Analysis

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

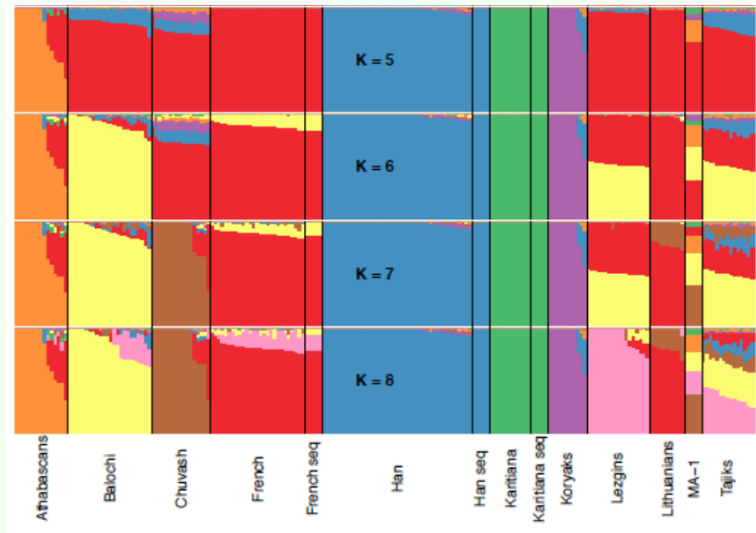
## Where ANGSD does less well

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

Its not bad there are just better options

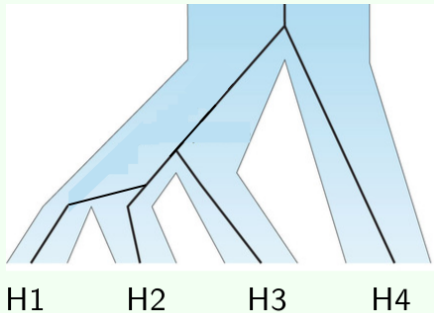


## Common use - NGSadmix



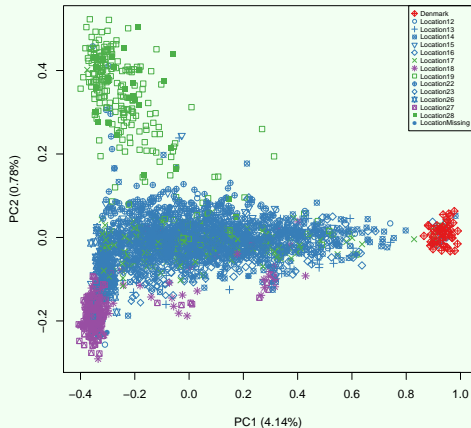


## Common use - D-stat/ABBABABA

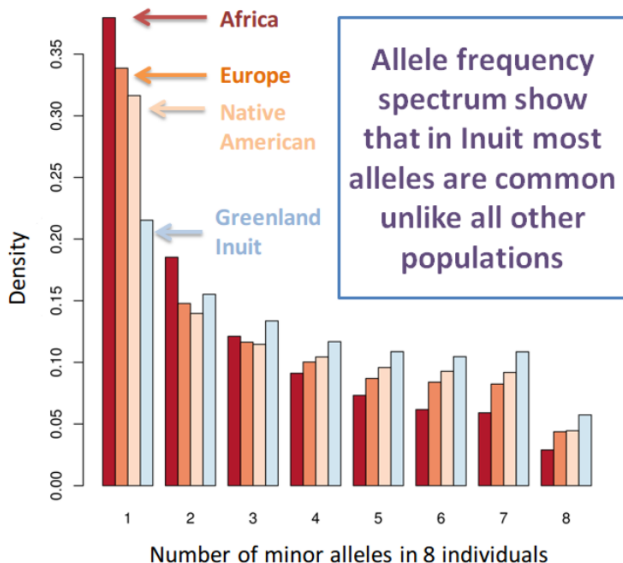


# Common use - MDS/PCA

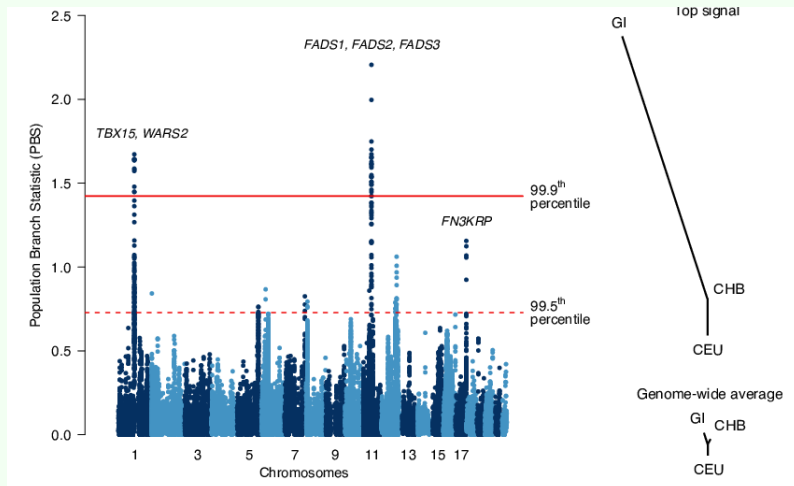
Greenlanders+50 Danes



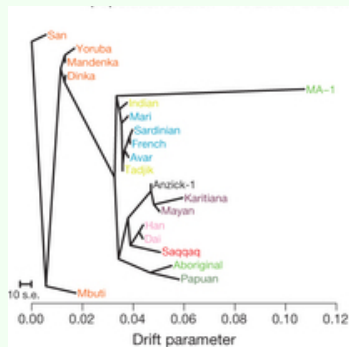
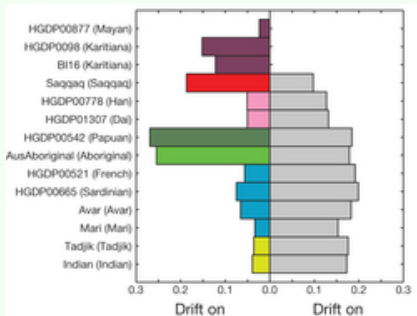
## Common use - SFS



## SFS - selection scans - theta/Tajima/Fst/PBS



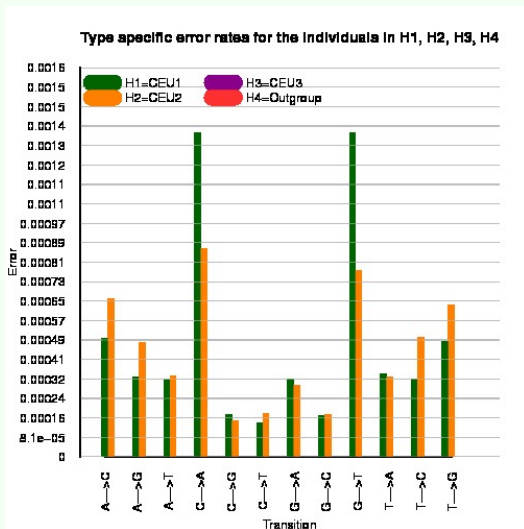
## SFS - test for continuation



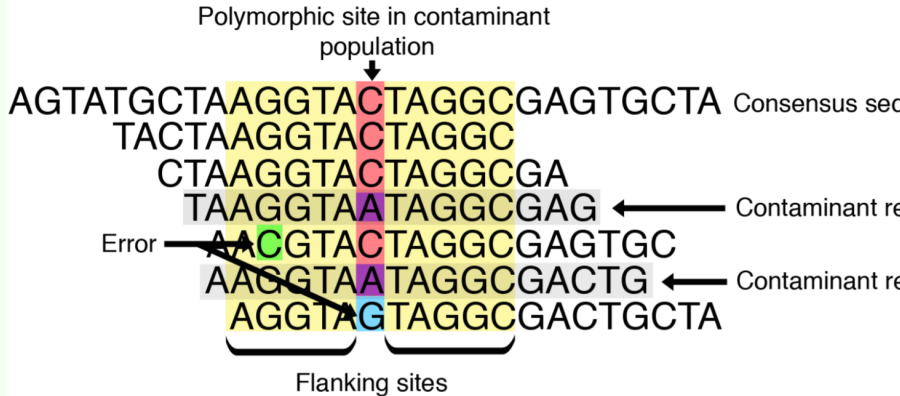
### Conclusion

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

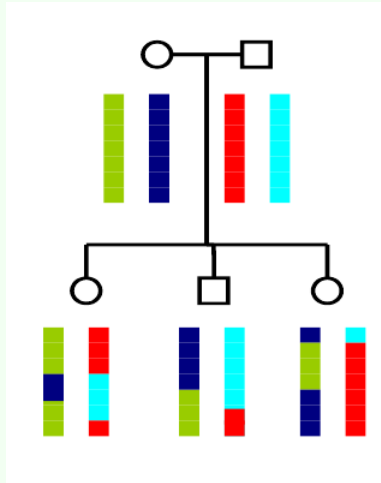
## Common use - Error rate estimation



## Common use - contamination



## Common use - relatedness





## exercises

### Data from 1000 Genomes

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