ANGSD
Analysis of Next Generation Sequencing Data

Major programmer: Thorfinn Korneliusen
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

They actually make a wrapper for ANGSD
https://github.com/mojaveazure/angsd-wrapper
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 called `general`.
## Input formats

### Sequencing data
- Bam
- Cram
- mpileup

### Genotype likelihoods
- Beagle
- glfV3
- tglf
- others

### Genotype (posterior) probability
- Beagle

---

### Example

**BAM** → **ANGSD** → **BEAGLE** → **ANGSD** → **Association**

**MSMS** → **mpileup** → **ANGSD** → **SFS** → **∂a∂i**

**Cram** → **SNPtools** → **GL** → **ANGSD** → **NGSadmix**
## Analysis

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

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

It's not bad there are just better options
Common use - NGSadmix

1

1Raghaven et. al Nature 2014
Common use - D-stat/ABBABABABA
Allele frequency spectrum show that in Inuit most alleles are common unlike all other populations.
SFS - selection scans - theta/Tajima/Fst/PBS
Conclusion

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

---

2 Rasmussen et al. Nature 2014
Common use - Error rate estimation
Common use - contamination

Polymorphic site in contaminant population

AGTATGCTAAAGGTAACCTAGGCGAGTGCTA Consensus sequence
TACTAAAGGTAACCTAGGCG
CTAAGGTAACCTAGGCGA
TAAGGTAATAGGCGAG
Error
AGGTAATAGGCGAG
AAGGTAATAGGCGACTG
AGGGTAGGTAGGCGACTGCTA

Flanking sites

Contaminant reads

Contaminant reads
Common use - relatedness
### 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)