Week-6

Genotype likelihoods based analyses - ANGSD

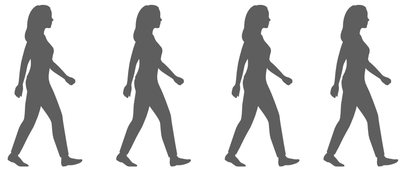
Aim: *Introduction to ANGSD, becoming familiar with probabilistic methods for dataset preparation*

Hands-on: *Running ANGSD on low coverage data*

Pr(R|G)

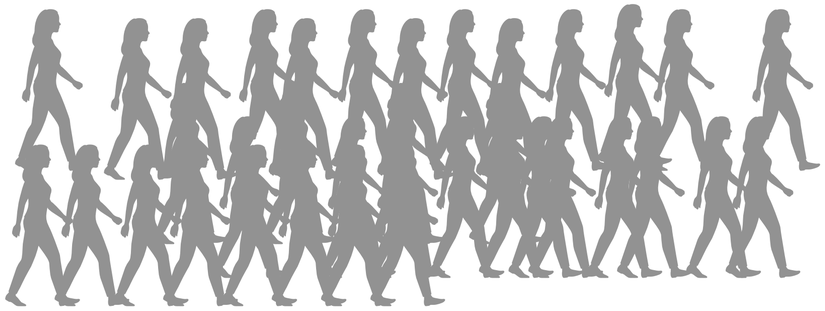
AAGCGGCGTGTGT
AAGCGGCGTGTGT
AAGCGGCGTGTGT
AAGCGCCGTGTGT
AAGCGGCGTGTGT
AAGCGCCGTGTGT
AAGCGCCGTGTGT
AAGCGCCGTGTGT
AAGCGCCGTGTGT

Data Production Strategy & Data processing



n=4

coverage per individual = 80 X



n=800

coverage per individual = 2 X

Alignment (read mapping) -> Filtering -> SNP calling & Genotype calling -> Filtering

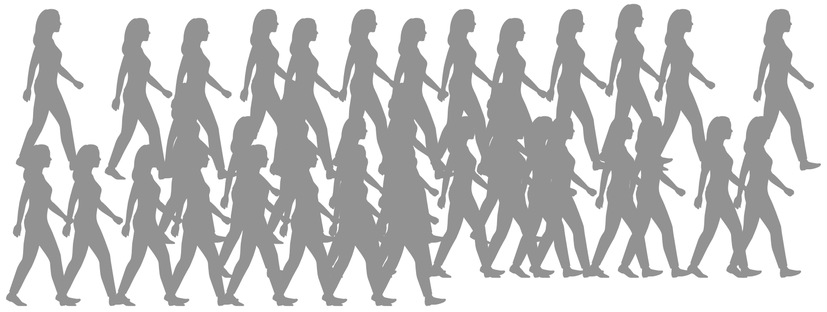
A process of genotype discovery per individual using base calls and quality scores per base

Some analysis can be run even under uncertainty of genotypes

Genotype likelihoods: A probabilistic framework

Traditional genotype&SNP calling -> count alleles per site / use a set of filtering parameters (such as mapping quality, base quality etc…)

Genotype likelihoods: incorporate uncertainty and additional iformation such as LD.



n=800

coverage per
individual = 2 X

Low - Medium coverage sequencing data:

Regular SNP&genotype calling approach

-heterzoygous genotypes

-individual read qualities not incorporated

Probablistic framework -> posterior probability per genotype

A genotype likelihood

Genotype likelihood for genotype G: Pr(X|G)

X = All sequencing read data for a locus in an individual's genome

A genotype w/ the highest posterior probability is selected

Advantage: Provide a measure of statistical uncertainty - more accurate - incorporate LD and allele frequencies

A genotype likelihood

Pr(Xi|G);

Xi: data in read i for an individual and a particular site with genotype G

-> reads are assumed to be independent / free from PCR duplicates and alignment errors

-> base quality recalibration can improve the process

-> estimating error rates from read data per site also improves the process

ANGSD: A software based on genotype likelihoods
http://www.popgen.dk/angsd/index.php/ANGSD

Log in

Page | Discussion | Read | View source | View history | Search

ANGSD overview
Installation
Quick Start/Testdata
Input data
Filters
snpFilters

Population genetics
- SFS Estimation
- Thetas,Tajima,Neutrality test
- (Multi) SFS Estimation
- Direct Ancestry

Population structure
- Admixture
- Fst
- ABBABABA (D-stat)
- ABBABABA (multipop)
- Population branch statistics (pbs)
- PCA
- PCA (sampling approach)
- Linkage disequilibrium

ANGSD: Analysis of next generation Sequencing Data

Latest tar.gz version is (0.934/0.935 on github), see Change_log for changes, and download it here.

ANGSD

ANGSD is a software for analyzing next generation sequencing data. The software can handle a number of different input types from mapped reads to imputed genotype probabilities. Most methods take genotype uncertainty into account instead of basing the analysis on called genotypes. This is especially useful for low and medium depth data. The software is written in C++ and has been used on large sample sizes.

This program is not for manipulating BAM/CRAM files, but solely a tool to perform various kinds of analysis. We recommend the excellent program SAMtools for outputting and modifying bamfiles.

ANGSD is also on github: https://github.com/ANGSD/angsd

Synopsis

```
./angsd [OPTIONS]
```

example of allele frequency estimated from genotype likelihoods with bam files as input using 10 threads

```
./angsd -out outFileName -bam bam.filelist -GL 1 -doMaf 1 -doMajorMinor 1 -nThreads 10
```

Platform

The program is developed on tested on a Linux system with gcc compiler. It compiles on OSX with clang, but OSX is not really that tested.

ANGSD: A software based on genotype likelihoods
http://www.popgen.dk/angsd/index.php/ANGSD

angsd is installed in our server:

$: angsd

```
   -> angsd version: 0.935-53-gf475f10 (htslib: 1.14-7-g1d79f44) build(Nov 15 2021 21:14:34)
   -> /usr/local/sw/angsd/angsd
   -> No '-out' argument given, output files will be called 'angsdput'

   -> angsd version: 0.935-53-gf475f10 (htslib: 1.14-7-g1d79f44) build(Nov 15 2021 21:14:32)
   -> Please use the website "http://www.popgen.dk/angsd" as reference
   -> Use -nThreads or -P for number of threads allocated to the program
Overview of methods:
   -GL   Estimate genotype likelihoods
   -doCounts Calculate various counts statistics
   -doAsso     Perform association study
   -doMaf   Estimate allele frequencies
   -doError  Estimate the type specific error rates
   -doAncError  Estimate the errorrate based on perfect fastas
   -HWE_pval Est inbreedning per site or use as filter
   -doGeno     Call genotypes
   -doFasta  Generate a fasta for a BAM file
   -doAbbababa  Perform an ABBA-BABA test
   -sites   Analyse specific sites (can force major/minor)
   -doSaf   Estimate the SFS and/or neutrality tests genotype calling
   -doHetPlas  Estimate hetplasmy by calculating a pooled haploid frequency

   Below are options that can be usefull
   -bam     Options relating to bam reading
   -doMajorMinor   Infer the major/minor using different approaches
   -ref/-anc Read reference or ancestral genome
   -doSNPstat  Calculate various SNPstat
   -cigstat  Printout CIGAR stat across readlength
   many others

Output files:
    In general the specific analysis outputs specific files, but we support basic bcf output
   -doBcf   Wrapper around -dopost -domajorminor -dofreq -gl -dovcf docounts
For information of specific options type:
   ./angsd METHODNAME eg
      ./angsd -GL
      ./angsd -doMaf
      ./angsd -doAsso etc
      ./angsd sites for information about indexing -sites files
Examples:
   Estimate MAF for bam files in 'list'
      './angsd -bam list -GL 2 -doMaf 2 -out RES -doMajorMinor 1'
```

ANGSD: A software based on genotype likelihoods
http://www.popgen.dk/angsd/index.php/ANGSD

bam files: data/week1_2/bams/
make a bamlist: bamlist.txt

```
/DATA/BIN784/data/week1_2/bams/Ash129_2019.merged.hs37d5.fa.cons.90perc.bam
/DATA/BIN784/data/week1_2/bams/Ash133.merged.hs37d5.fa.cons.90perc.bam
```

#we have already indexed the bams, skip this step:

while read i; do samtools index ${i}; done < bamlist.txt

reference genome: /DATA/share/ref/hsa/hs37d5.fa

```
angsd -b bamlist.txt -GL 1 -doMajorMinor 1 -doMaf 2
```

http://www.popgen.dk/angsd/index.php/Genotype_Likelihoods

http://www.popgen.dk/angsd/index.php/Major_Minor

http://www.popgen.dk/angsd/index.php/Allele_Frequencies