ANGSD

Analysing low-coverage whole-genome re-sequencing

Club Bioinfo IBIS 26 novembre 2019

Claire Mérot

Louis Bernatchez
Maren Wellenreuther (U. Auckland)

Anne-Laure Ferchaud

Hugo Cayuela
Quentin Rougemont
Eric Normandeau

Why using low-coverage data? *(+ low-cost libraries…)*

Sequencing costs

output=

nb of individuals X genome size X depth of coverage

Linkage map on 1920 progeny!

An Ultra High-Density *Arabidopsis thaliana* Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features

Beth A. Rowan, Darren Heavens, Tatiana R. Feuerborn, Andrew J. Tock, Ian R. Henderson, and Detlef Weigel

Experimental evolution with 6 replicates of 50 ind.

Contrasting genomic shifts underlie parallel phenotypic evolution in response to fishing

Nina O. Therkildsen, Aryn P. Wilder, David O. Conover, Stephan B. Munch, Hannes Baumann, Stephen R. Palumbi

GWAS with > 11,000 whole genomes

Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores

Julian R. Homburger, Cynthia L. Neben, Gilad Mishne, Alicia Y. Zhou, Sekar Kathiresan, Amit V. Khera (BioRxiv)

MOLECULAR ECOLOGY

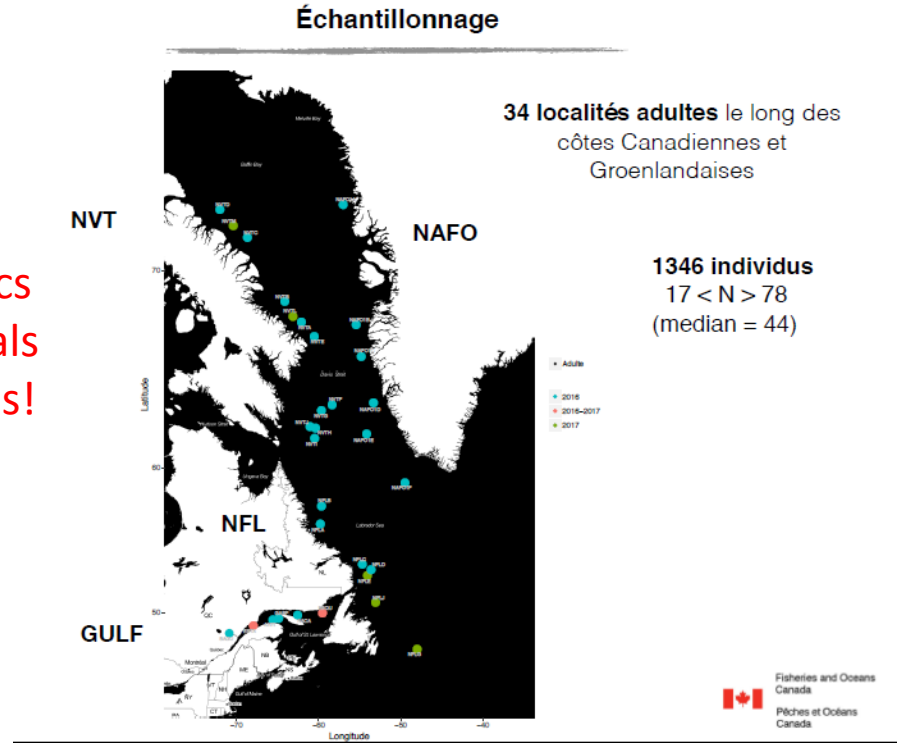
Molecular Ecology (2012) doi: 10.1111/mec.12105

Population genomics based on low coverage sequencing: how low should we go?

C. ALEX BUERKLE* and ZACHARIAH GOMPERT†

*Department of Botany and Program in Ecology, University of Wyoming, Laramie, WY, USA, †Department of Biology, Texas State University, San Marcos, TX, USA

Population genomics with 1346 individuals from 34 populations!



ANGSD : a suite of tools

Korneliussen *et al. BMC Bioinformatics* 2014, **15**:356
http://www.biomedcentral.com/1471-2105/15/356

BMC Bioinformatics

SOFTWARE **Open Access**

ANGSD: Analysis of Next Generation Sequencing Data

Thorfinn Sand Korneliussen[1*], Anders Albrechtsen[2] and Rasmus Nielsen[1,3]

Abstract

Background: High-throughput DNA sequencing technologies are generating vast amounts of data. Fast, flexible and memory efficient implementations are needed in order to facilitate analyses of thousands of samples simultaneously.

Results: We present a multithreaded program suite called ANGSD. This program can calculate various summary statistics, and perform association mapping and population genetic analyses utilizing the full information in next generation sequencing data by working directly on the raw sequencing data or by using genotype likelihoods.

Conclusions: The open source c/c++ program ANGSD is available at http://www.popgen.dk/angsd. The program is tested and validated on GNU/Linux systems. The program facilitates multiple input formats including BAM and imputed beagle genotype probability files. The program allow the user to choose between combinations of existing methods and can perform analysis that is not implemented elsewhere.

Keywords: Next-generation sequencing, Bioinformatics, Population genetics, Association studies

Advantages:

- *Appropriate for low-coverage*
- Flexible inputs
- Multiple methods, filters, etc.
- Large datasets
- Many downstream analyses
- Documentation ok – reactivity Github

Inconvenients:

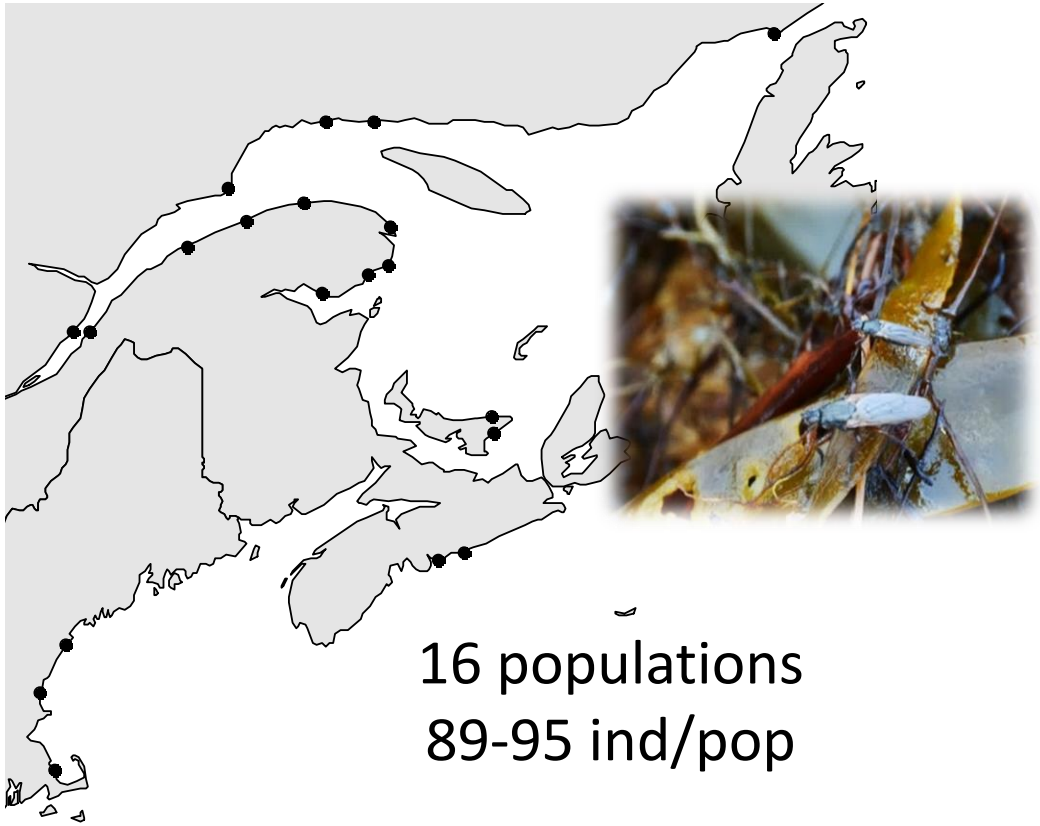
- Demanding for memory/time
- Sometimes update unclear and obscure parameters

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

https://github.com/ANGSD/angsd

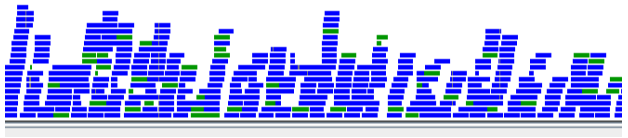
Example for population genomics…

Coelopa frigida



16 populations
89-95 ind/pop

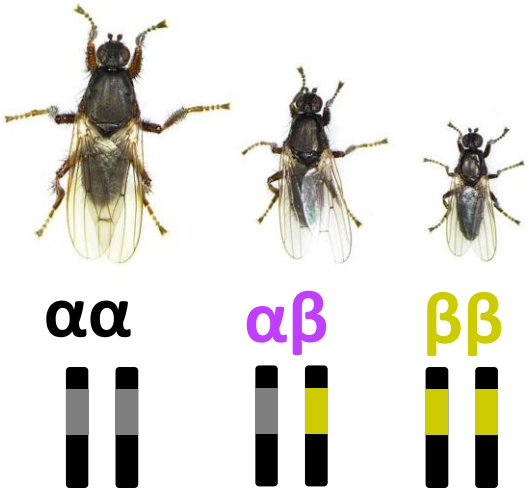
Mean coverage:
1.2x / ind
100x / pop



- Population structure?

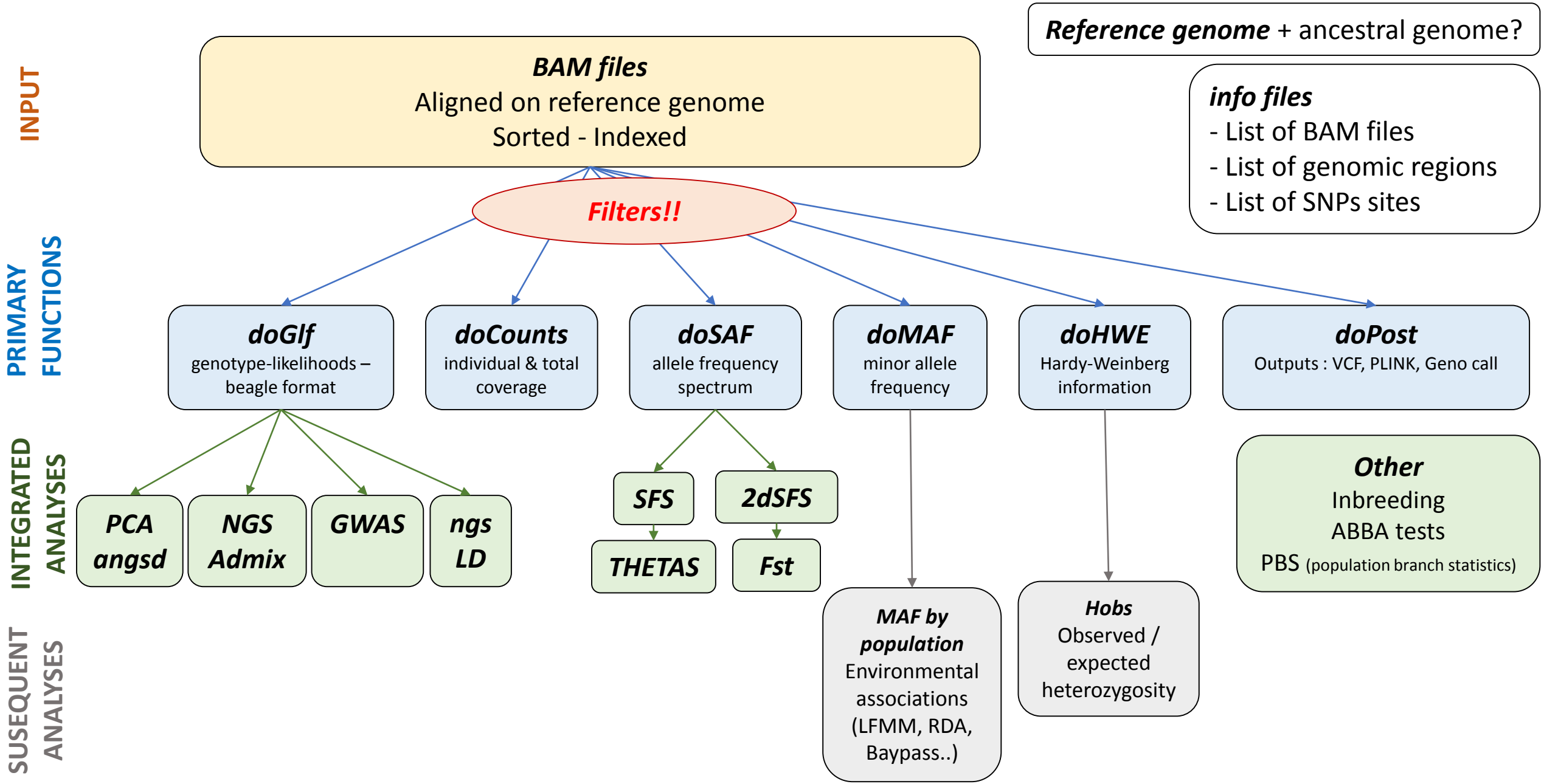
(Geography? Chromosomal inversion?)

- Environmental associations?
- Linkage disequilibrium?
- Sex chromosome?
- Demography?



https://github.com/clairemerot/angsd_pipeline

ANGSD overview



ANGSD inputs

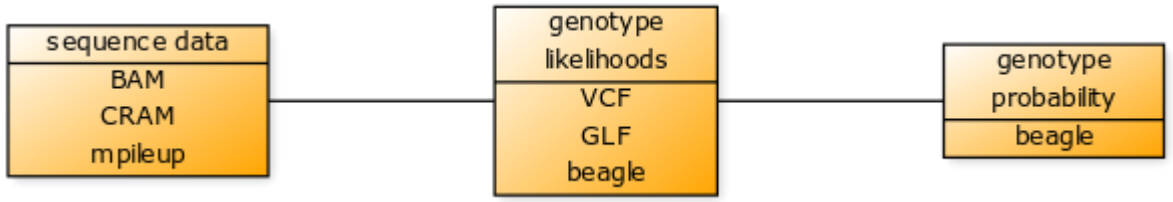
INPUT

BAM files
Aligned on reference genome
Sorted - Indexed

Reference genome + ancestral genome?

info files
- List of BAM files
- List of genomic regions
- List of SNPs sites

The program can also take :



ANGSD inputs

INPUT

BAM files
Aligned on reference genome
Sorted - Indexed

Reference genome + ancestral genome?

info files
- List of BAM files
- List of genomic regions
- List of SNPs sites

- **List of BAM files** is of primary importance!!

-> obtain a saf or a maf by population = give a bam list for the population…

```
../wgs_sample_preparation/09_no_overlap/cfrig_L1_BP16-0001_F_BP_AB_1.no_overlap.bam
../wgs_sample_preparation/09_no_overlap/cfrig_L1_BP16-0002_M_BP_AB_1.no_overlap.bam
../wgs_sample_preparation/09_no_overlap/cfrig_L1_BP16-0004_M_BP_AB_1.no_overlap.bam
```

- **List of genomic regions**

-> to restrain to specific chromosomes/scaffolds : useful for faster analyses!

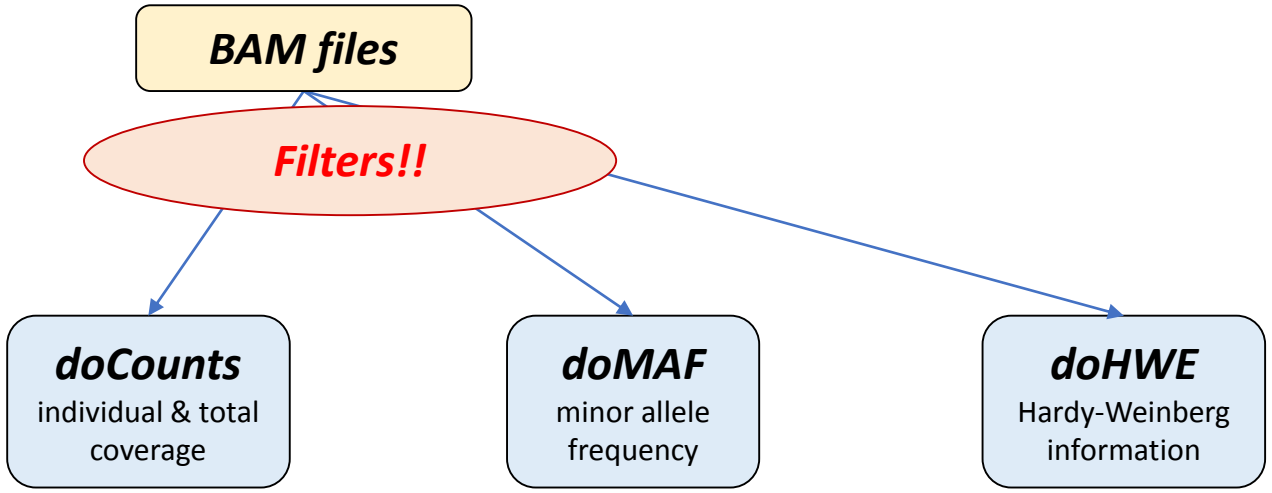
```
LG1
LG2
LG3
LG4
LG5
LG6
scaffold1126
scaffold125
scaffold151
scaffold153
```

- **List of SNPs sites**

-> For instance: get a list of SNPs which pass all filters for all the set of individuals and then restrain maf by pop to this list

```
LG1    3867    T    C
LG1    3870    C    A
LG1    3880    C    G
LG1    7206    G    C
LG1    7207    T    G
LG1    7223    T    C
LG1    7517    C    G
LG1    7520    G    A
```

ANGSD filters



BAM quality:

-remove_bads = 1

-minMapQ = 20

Samtools filters
-baq 1 -C 50

Coverage:
-minInd *Nind*50%*

-setMaxDepth *Nind*3*

-setMinDepthInd *1*

MAF:
-minMaf *0.05*

-SNP_pval *0.00001*
(polymorphic sites)

HW:
-minHWEpval *(sites at HW equilibrium)*

⇒ List of SNPs sites

ANGSD basic code

```
angsd      -b bam.filelist \
           -anc ref.fasta  -ref ref.fasta\
           -rf regions.txt \

           -out folder/output \
           -P $NB_CPU -nQueueSize 50 –underFlowProtect 1 \
```

INPUT

Computation help
(ulimit -S -n 2048)

$NB_CPU Max 8-10 (usually 4-6)
-> risque of fragmenting too much memory and is not very efficient.

Too many Bam files to open
-> « ulimit -S -n 2048 » at the beginning of the script

Splitting by regions (by chromosome ?)
-> saves time for -doSaf

pour une pop a 10-12X, 120 individus et 2Go de genome = 30 jours (pas sur manitou)

par chromo c'était fait en moins de 24h.

ANGSD basic code

```
angsd    -b bam.filelist \
         -anc ref.fasta  -ref ref.fasta\
         -rf regions.txt \

         -out folder/output \
         -P $NB_CPU -nQueueSize 50 –underFlowProtect 1 \

         -GL 2 \
         -doMajorMinor 1  \
         -doSaf 1 –doMaf 1 – do Glf 1 -doHWE 1 –doCounts 1 –doPost 1\

         -remove_bads 1 \
         -minMapQ 20 \

         -minInd 50 \
         -setMaxDepth 300 \

         -minMaf 0.05
```

INPUT

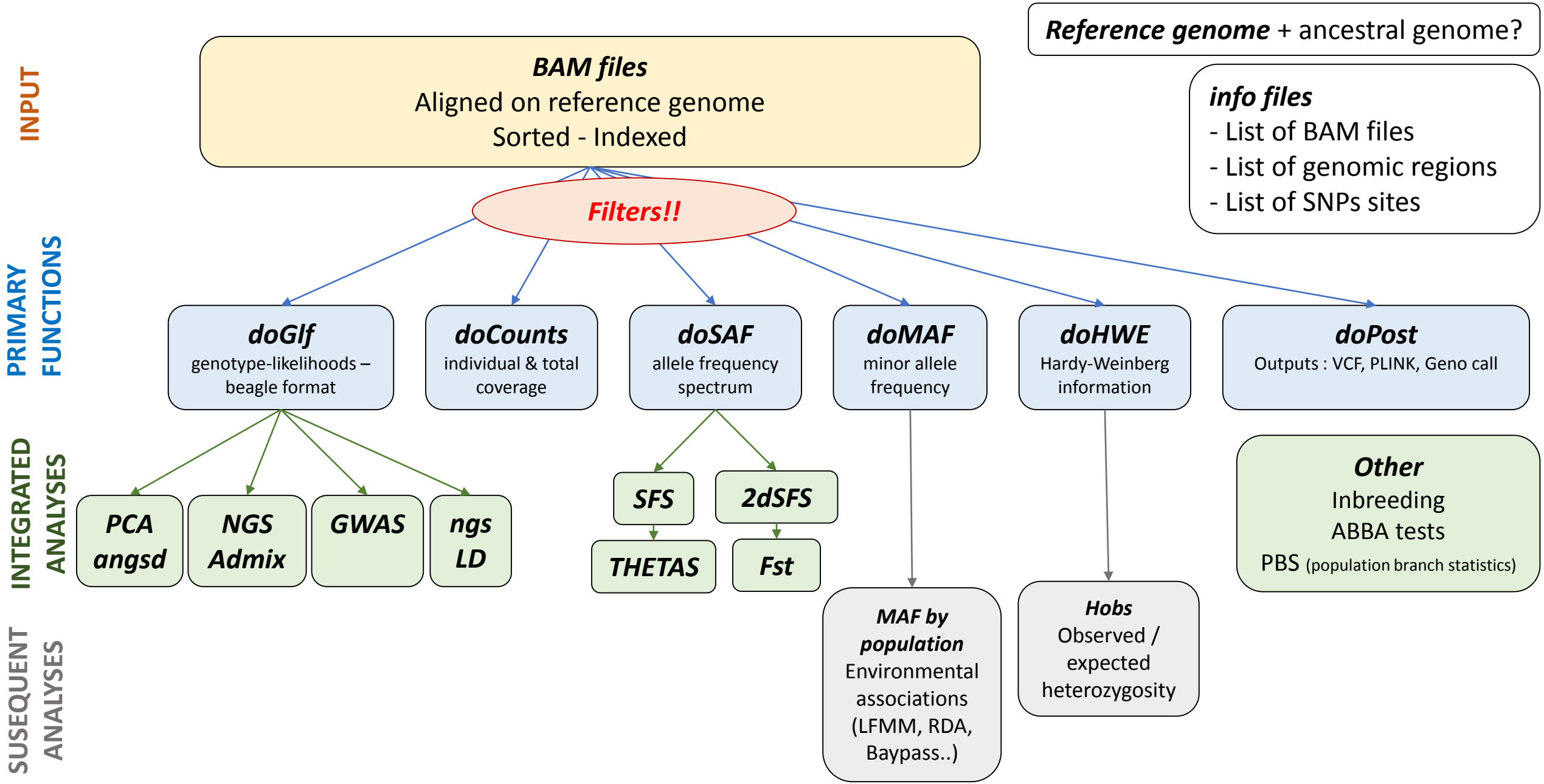
Computation help
(ulimit -S -n 2048)

Choose the underlying model :
GL 1 = samtools; GL 2 = GATK

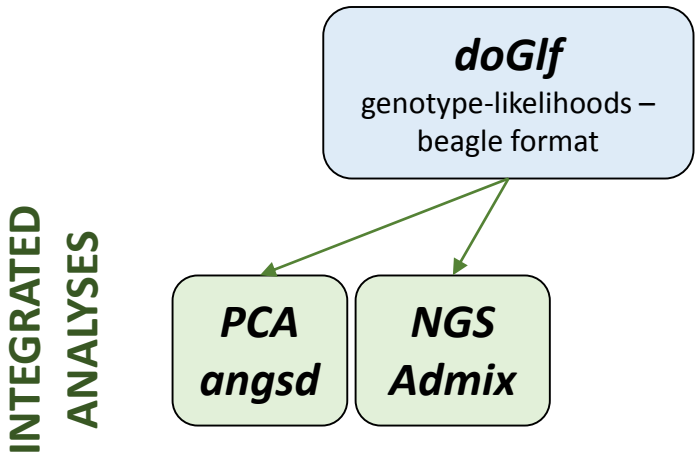
& the basic analysis to run

Filters!!

ANGSD overview



ANGSD : using Genotype likelihoods



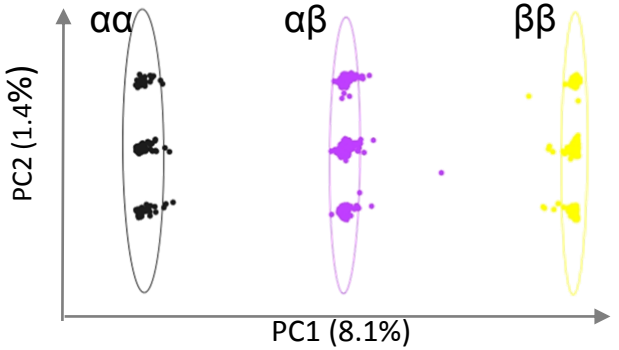
marker	allele1	allele2	Ind0	Ind0	Ind0	Ind1	Ind1	Ind1
LG1_3867	3	1	0.799992	0.200008	0.000000	0.333333	0.333333	0.333333
LG1_3870	1	0	0.799985	0.200015	0.000000	0.333333	0.333333	0.333333
LG1_3880	1	2	0.000000	0.200015	0.799985	0.333333	0.333333	0.333333
LG1_7206	2	1	0.888863	0.111137	0.000000	0.333333	0.333333	0.333333
LG1_7207	3	2	0.666649	0.333333	0.000018	0.333333	0.333333	0.333333

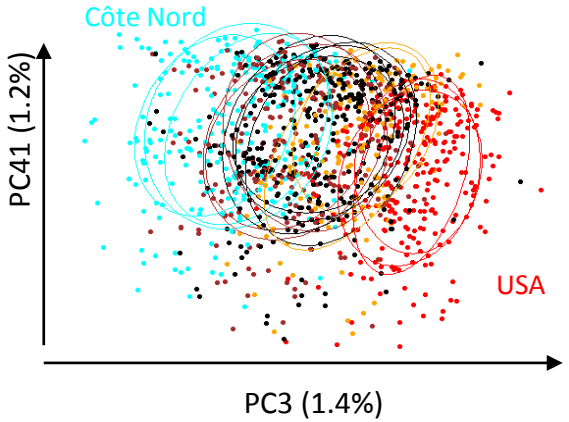
Filters: Only polymorphic sites maf >0,05 (0,10-0,20)

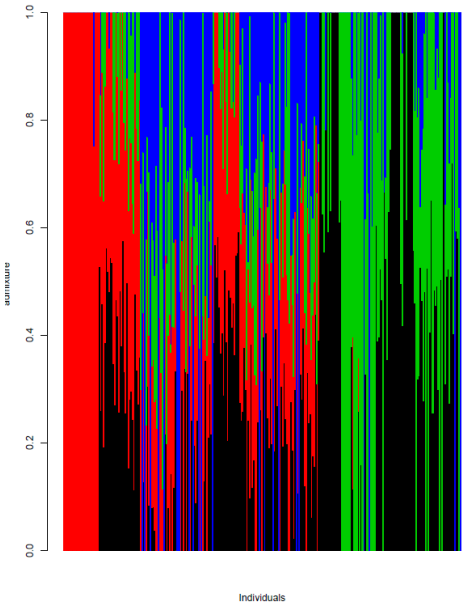
Explore genetic structure within the population

Meisner, J., & Albrechtsen, A. (2018). Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, *210*(2), 719-731.

Skotte, L., Korneliussen, T. S., & Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. *Genetics*, *195*(3), 693-702.







ANGSD : using Genotype likelihoods

INTEGRATED ANALYSES

doGlf
genotype-likelihoods – beagle format

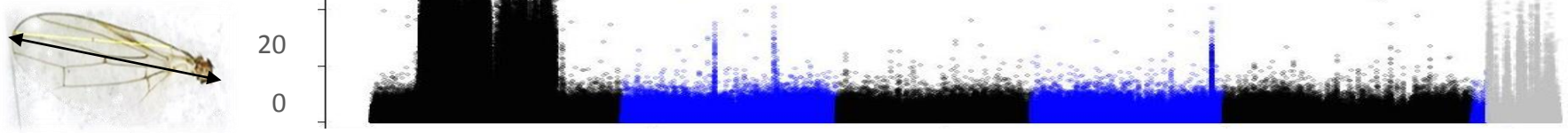
GWAS

marker	allele1	allele2	Ind0	Ind0	Ind0	Ind1	Ind1	Ind1
LG1_3867	3	1	0.799992	0.200008	0.000000	0.333333	0.333333	0.333333
LG1_3870	1	0	0.799985	0.200015	0.000000	0.333333	0.333333	0.333333
LG1_3880	1	2	0.000000	0.200015	0.799985	0.333333	0.333333	0.333333
LG1_7206	2	1	0.888863	0.111137	0.000000	0.333333	0.333333	0.333333
LG1_7207	3	2	0.666649	0.333333	0.000018	0.333333	0.333333	0.333333

Filters: Only polymorphic sites maf >0,05 (0,10-0,20)

Explore genotype-phenotype associations

GWAS



Jørsboe, E., & Albrechtsen, A. (2019). A Genotype Likelihood Framework for GWAS with Low Depth Sequencing Data from Admixed Individuals. *bioRxiv*, 786384.

ANGSD : using Genotype likelihoods

INTEGRATED ANALYSES

doGlf
genotype-likelihoods – beagle format

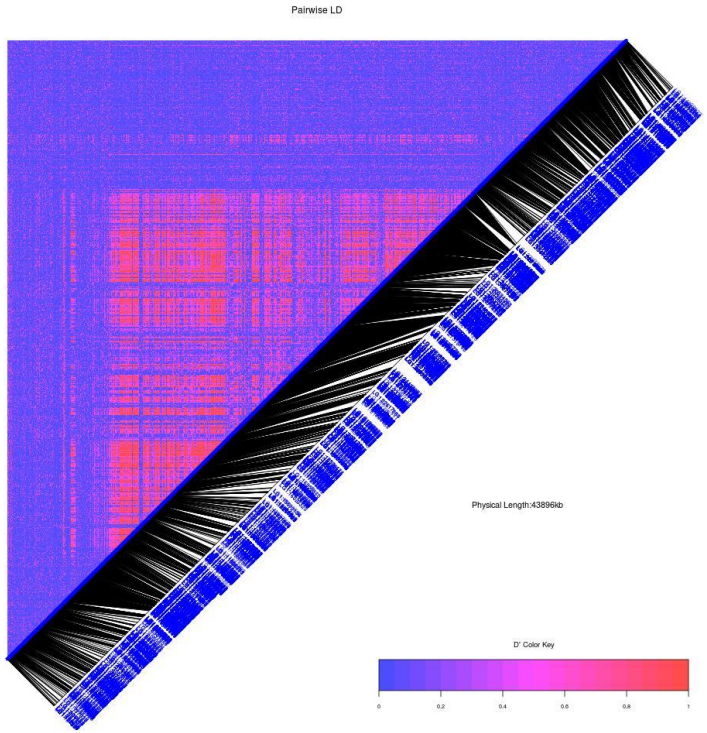
ngs LD

marker	allele1	allele2	Ind0	Ind0	Ind0	Ind1	Ind1	Ind1
LG1_3867	3	1	0.799992	0.200008	0.000000	0.333333	0.333333	0.333333
LG1_3870	1	0	0.799985	0.200015	0.000000	0.333333	0.333333	0.333333
LG1_3880	1	2	0.000000	0.200015	0.799985	0.333333	0.333333	0.333333
LG1_7206	2	1	0.888863	0.111137	0.000000	0.333333	0.333333	0.333333
LG1_7207	3	2	0.666649	0.333333	0.000018	0.333333	0.333333	0.333333

Filters: Only polymorphic sites maf >0,05 (0,10-0,20)

Explore Linkage disequilibrium

Chr I



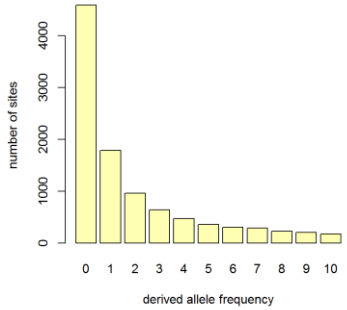
Fox, E. A., Wright, A. E., Fumagalli, M., & Vieira, F. G. (2019). ngsLD: evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics*.

ANGSD : Allele frequency spectrums & statistics

INTEGRATED ANALYSES

No MAF filters for thetas

doSAF
allele frequency spectrum



SFS

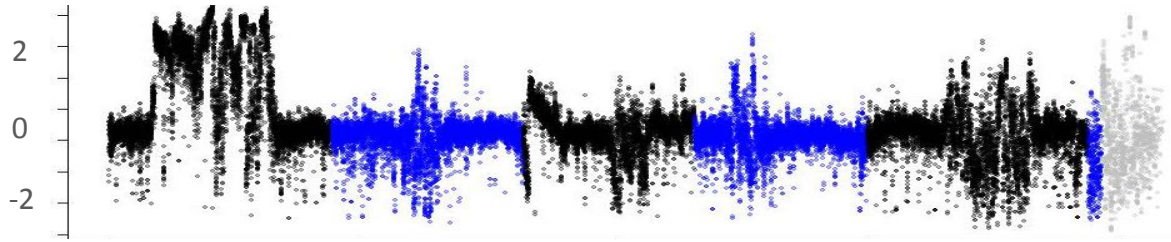
THETAS

RealSFS
Reduce the number of sites
(nsites =1,000,000)
Lots of memory

Thetas Watterson
Thetas diversity
Tajima's D…

Output by sliding-windows…

Tajima's D



#(indexStart,indexStop)(firstPos_withData,lastPos_withData)(WinStart,WinStop)	Chr	WinCenter	tW	tP	tF	tH	tL	Tajima	fuf	fud
(176,1131)(7450,35760)(5000,30000)	LG1	17500	14.219130	7.847303	21.363536	3.136890	5.492096	-1.386729	-1.656608	-1.371548
(342,1131)(15474,35760)(10000,35000)	LG1	22500	12.209059	7.268703	17.587680	3.100106	5.184405	-1.242102	-1.442608	-1.162469
(342,1293)(15474,43874)(15000,40000)	LG1	27500	15.588463	10.059668	21.122273	4.422301	7.240984	-1.102375	-1.251050	-0.987698
(526,1313)(22758,48244)(20000,45000)	LG1	32500	11.937200	7.151281	17.655682	3.843268	5.497275	-1.229086	-1.497065	-1.257446
(869,1318)(25128,55089)(25000,50000)	LG1	37500	8.536802	6.020949	10.987299	2.797213	4.409081	-0.883283	-0.936547	-0.691591
(1131,1318)(35760,55089)(30000,55000)	LG1	42500	3.961965	3.301488	4.171383	1.518896	2.410192	-0.460141	-0.296214	-0.099487
(1131,1394)(35760,63970)(35000,60000)	LG1	47500	4.797212	3.978684	5.261946	1.693481	2.836082	-0.483032	-0.379722	-0.195108
(1293,1569)(43874,77361)(40000,65000)	LG1	52500	4.853182	3.257536	7.174833	0.923866	2.090701	-0.932091	-1.149144	-0.967295

Korneliussen, T. S., Moltke, I., Albrechtsen, A., & Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC bioinformatics*, *14*(1), 289.

ANGSD : Allele frequency spectrums & statistics

No MAF filters for demography

doSAF
allele frequency spectrum

SAF on pop 1
SAF on pop 2

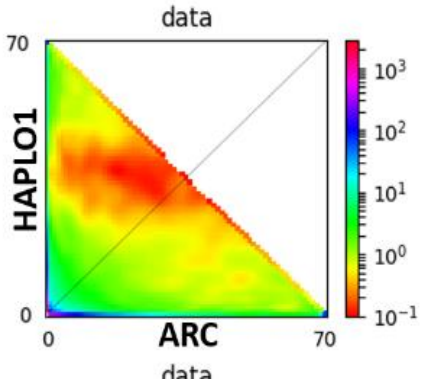
RealSFS
Reduce the number of sites…
Lots of memory

RealSFS **2dSFS saf1 vs saf 2 (option – fold 1)**

Fold if
ancestral genome = reference genome…

INTEGRATED ANALYSES

2dSFS



Demography
(dadi, fastsimcoal, ABC)

Warmuth VM & Ellegren H. (2019) Genotype-free estimation of allele frequencies reduces bias and improves demographic inference from RADSeq data. Molecular Ecology Ressources. *19*(3), 586-596.

⇒ Better estimation of models & parameters with SFS from ANGSD than from SNPs calling through GATK (except if coverage > 100x!)

ANGSD : Allele frequency spectrums & statistics

MAF filters for FST?

doSAF
allele frequency spectrum

2dSFS

Fst

INTEGRATED ANALYSES

SAF on pop 1
SAF on pop 2

RealSFS **2dSFS saf1 vs saf 2 (option – fold 1)**

FST across all genome

	BS	RB	BT	SI	KA	ME	GM	RC	AG	CE	SS	NB	CB	BP	HA	MA
BS	0	0.009	0.009	0.014	0.015	0.014	0.014	0.014	0.013	0.014	0.016	0.018	0.018	0.026	0.028	0.034
RB	0	0.008	0.017	0.017	0.017	0.017	0.017	0.015	0.016	0.018	0.023	0.024	0.029	0.031	0.037	
BT	0	0.013	0.012	0.014	0.014	0.013	0.012	0.013	0.014	0.018	0.02	0.023	0.025	0.031		
SI	0	0.008	0.007	0.008	0.007	0.008	0.008	0.008	0.009	0.01	0.013	0.014	0.02			
KA	0	0.009	0.01	0.009	0.008	0.009	0.009	0.01	0.012	0.013	0.015	0.02				
ME	0	0.008	0.007	0.008	0.008	0.009	0.01	0.01	0.012	0.014	0.02					
GM	0	0.008	0.009	0.008	0.008	0.011	0.012	0.014	0.017	0.021						
RC	0	0.007	0.007	0.008	0.01	0.011	0.014	0.016	0.02							
AG	0	0.007	0.008	0.01	0.012	0.014	0.016	0.021								
CE	0	0.007	0.01	0.012	0.014	0.016	0.02									
SS	0	0.01	0.012	0.013	0.016	0.019										
NB	0	0.008	0.013	0.016	0.021											
CB	0	0.015	0.018	0.024												
BP	0	0.007	0.012													
HA	0	0.01														
MA	0															

North

South

Pairwise FST between populations

FST

αα

ββ

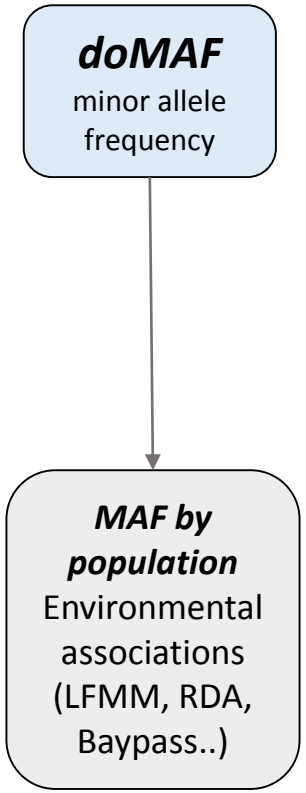
25kb sliding-windows

1

0

Output by sliding-windows…

ANGSD : Minor allele frequency



MAF for population BP

chr	pos	maj	min	anc	maf	nInd
LG1	3867	T	C	T	0.258300	50
LG1	3870	C	A	C	0.242971	50
LG1	3880	C	G	G	0.375692	52
LG1	7517	C	G	C	0.070817	45
LG1	7520	G	A	G	0.088480	46

Join with R

Chr_pos	BP	BS	BT	CB
LG1_8758	0.016149	0.033839	0.015712	0.040306
LG1_22838	0.12912	0.0989	0.117701	0.123505
LG1_25197	0.069546	0.160342	0.210446	0.073502
LG1_39818	0.162017	0.149856	0.143678	0.228882
LG1_80251	0.114682	0.069471	0.10154	0.087802
LG1_91603	0.047935	0.094046	0.081615	0.026046
LG1_92586	0.126451	0.118993	0.068226	0.052894
LG1_92914	0.293357	0.082381	0.199689	0.288091
LG1_94101	0.084773	0.092265	0.026972	0.053312

By POP: need do re-do doMAF on each group
(provide specific bam list & additional filter by pop?)

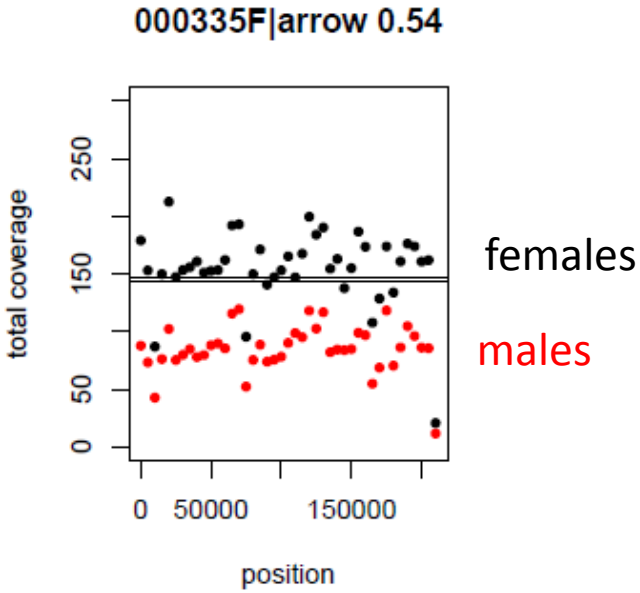
CAUTION: Ensure the same allele is called Major/Minor
(no option –doMajorMinor 1)

POSSIBILITY: use SITES list of filtered SNPs + Maj/Min info

ANGSD : Coverage

doCounts
individual & total coverage

chr	pos	totDepth	ind0	ind1	ind2	ind3	ind4
LG1	3867	1317	2	0	0	0	0
LG1	3870	1373	2	0	0	0	1
LG1	3880	1456	2	0	0	0	1
LG1	7206	1313	3	0	3	5	1
LG1	7207	1302	1	0	3	5	1
LG1	7223	1308	2	0	4	4	1



females

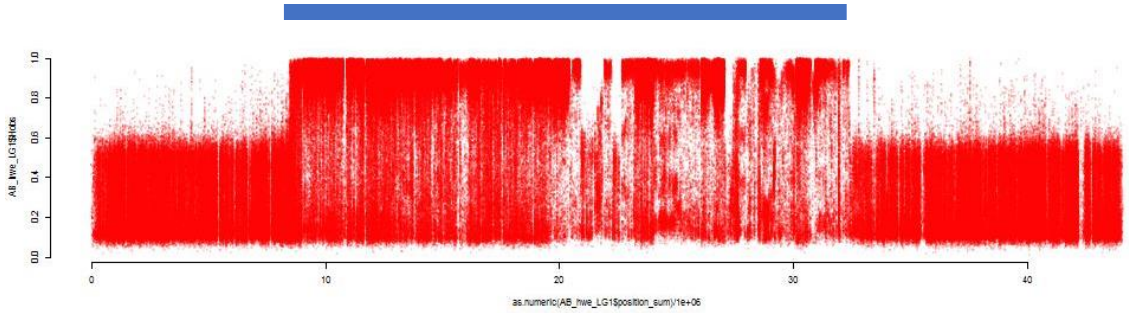
males

ANGSD : Hardy-Weinberg

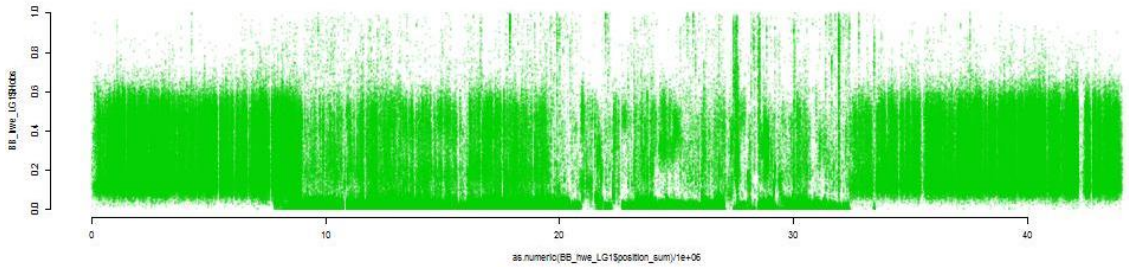
doHWE
Hardy-Weinberg information

Hobs
Observed / expected heterozygosity

Chromo	Position	Major	Minor	hweFreq	Freq	F	LRT	p-value
LG1	3867	T	C	0.368173	0.368488	0.013077	1.286494e-02	9.096946e-01
LG1	3870	C	A	0.346484	0.346097	-0.021122	4.025702e-02	8.409789e-01
LG1	3880	C	G	0.362878	0.362898	-0.000763	7.411434e-05	9.931311e-01
LG1	7206	G	C	0.125864	0.125804	-0.008976	6.477322e-03	9.358541e-01
LG1	7207	T	G	0.078733	0.078949	-0.084340	1.687467e+00	1.939352e-01
LG1	7223	T	C	0.462544	0.462309	-0.076438	5.069208e-01	4.764749e-01
LG1	7517	C	G	0.132875	0.132105	-0.043676	1.178875e-01	7.313371e-01
LG1	7520	G	A	0.105394	0.104367	-0.097492	7.774986e-01	3.779072e-01
LG1	8758	G	A	0.054450	0.054123	-0.050850	2.729722e-01	6.013450e-01

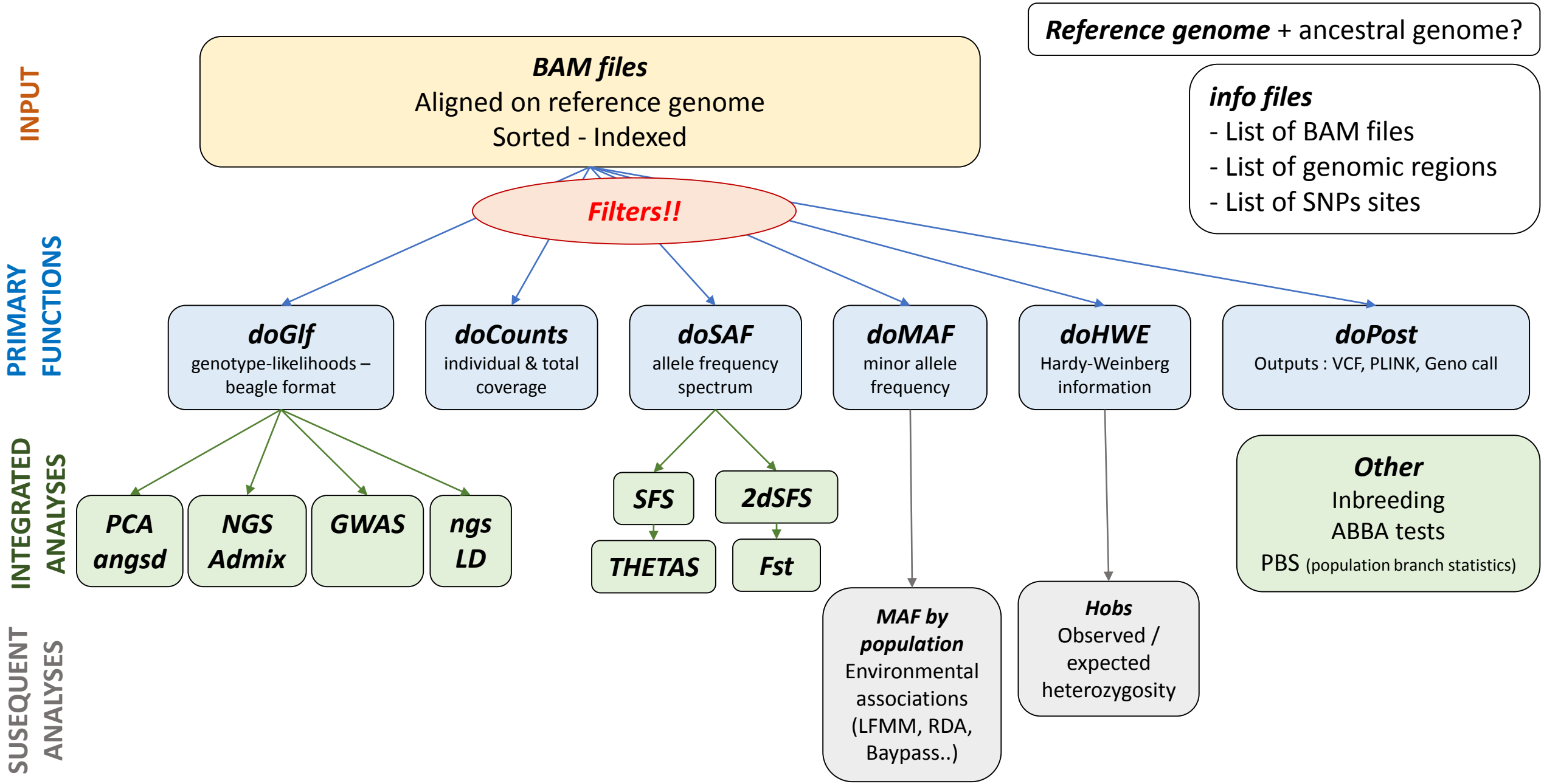


Inversion heterozygotes = excess Hobs



Inversion homozygotes = deficit Hobs

ANGSD overview



ANGSD/low-coverage: to conclude…

-> ANGSD is quite straightforward at the beginning…
BUT subtilities in filters, functions, datasets : be careful!

-> ANGSD can be long to run/demanding in memory :
try splitting by region
try splitting the different steps (e. g. ANGSD – RealSFS)

-> Gathers plenty of analyses + diverse input/output :
All in 1!

-> Takes into account uncertainty due to low coverage
(is known to perform well on higher coverage too.)

-> Other tools that you know to deal with low-coverage data??

Thanks…

Louis Bernatchez
M. Wellenreuther (U. Auckland)

A-L. Ferchaud

E. Normandeau
Q. Rougemont
H. Cayuela

IBIS
Bioinformatic Platform

ANGSD is up-to-date on Manitou/Katak

Thanks for your attention!

Thanks to Club Bioinfo!