Low-coverage whole-genome re-sequencing
ANGSD

Claire Mérot, Anna Tigano & Anne-Laure Ferchaud
Physalia Courses
September 2020

Why using low-coverage data?

(+ low-cost libraries…)

Sequencing costs

output=

nb of individuals X genome size X depth of coverage

Reduced-representation sequencing (Rad-seq, SNP-chip, exome capture,etc)

Why using low-coverage data?

(+ low-cost libraries…)

Sequencing costs

output=

nb of individuals X genome size X depth of coverage

Yes, I have many samples

I want to

- cover a large geographic zone
- study different ecological conditions
- keep statistical power to analyse phenotypes
- have good inference of population parameters…

Reduced-

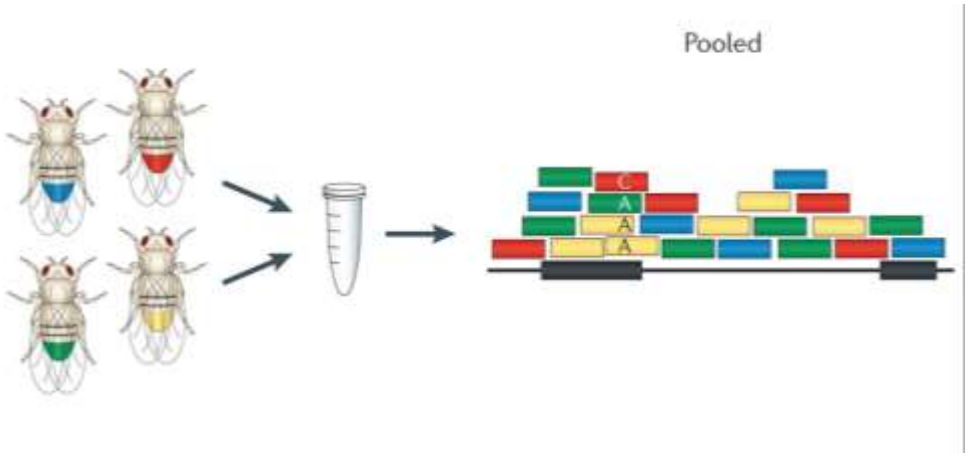
(Ra IP-

ture,e

I want whole-genome!

A possible solution:

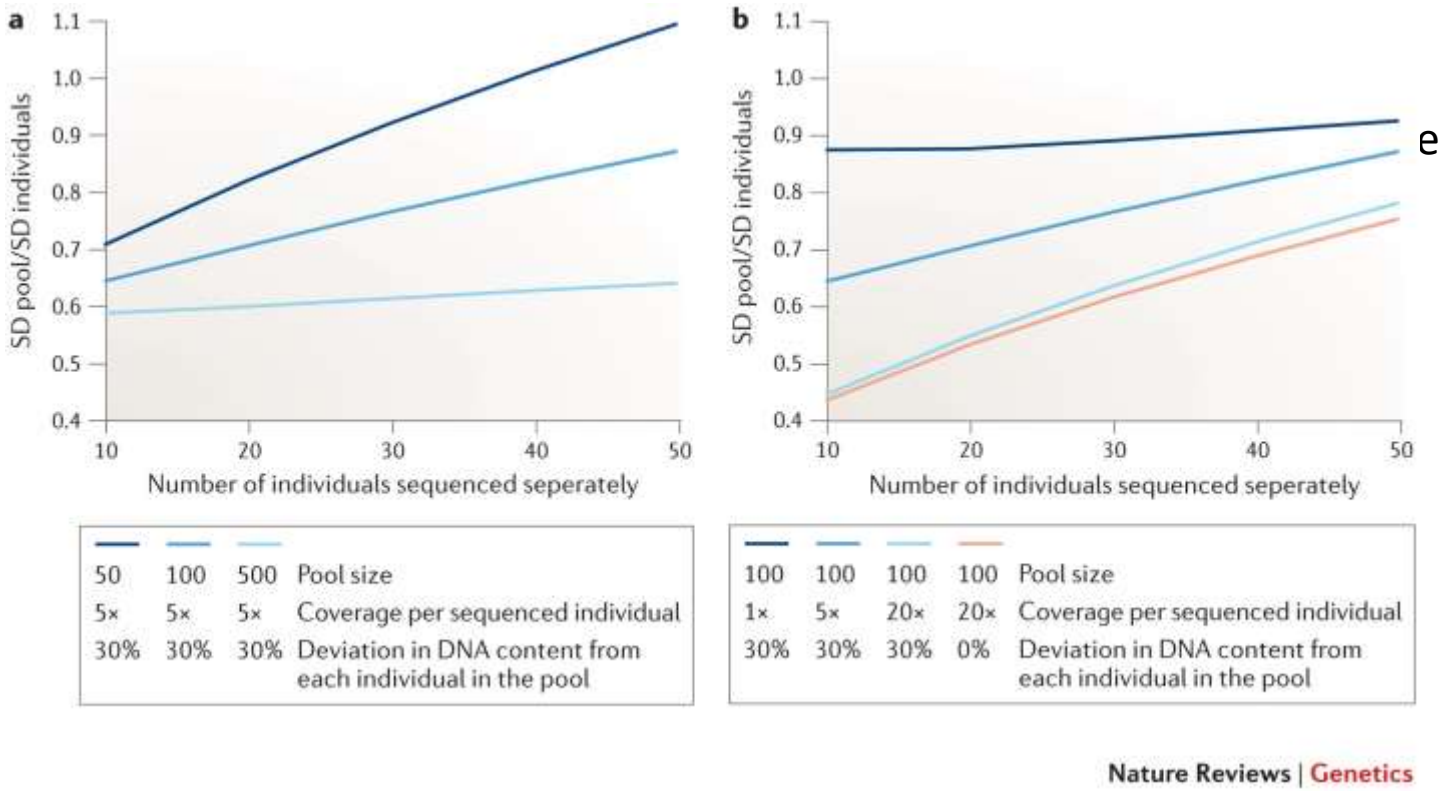
Pool-seq!



Why using low-coverage data?

(+ low-cost libraries…)

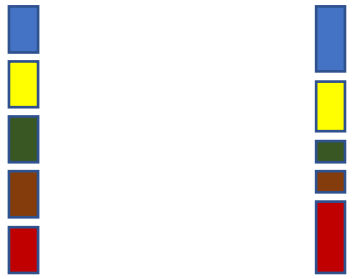
Short note about Pool-seq



e

Minimum number in a pool: 40
Minimum coverage: 50x

⇒ Pool-seq is a cost-effective strategy for many applications but:



Expected pooling

True pooling

⇒ Need replicates of pools (>3)

+ problems if contamination by one misassigned individual
+ difficulties dues to CNV

Schlötterer C, Tobler R, Kofler R, Nolte V. Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nat Rev Genet*. 2014;15(11):749-763. https://doi.org/10.1038/nrg3803

Why using low-coverage data?

(+ low-cost libraries…)

Sequencing costs

output=

nb of individuals X genome size X depth of coverage

Yes, I have many samples

I want to

- cover a large geographic zone
- study different ecological conditions
- keep statistical power to analyse phenotypes
- have good inference of population parameters…

Reduced-

(Ra IP-

ture,e

I want whole-genome!

Another solution:

Low-coverage whole-genome resequencing

+ cheap librairies

Key reference (for simulations of coverage variation)

Alex Buerkle, C. and Gompert, Z. (2013), Population genomics based on low coverage sequencing: how low should we go?. Mol Ecol, 22: 3028-3035. doi:10.1111/mec.12105

Why using low-coverage data?

(+ low-cost libraries…)

Minimize sequencing costs…

But what about library preparation?

The idea of the protocole:

- Cheap library preparation (<10$)
- Using Nextera tagmentation process with small volumes of enzyme (and small amount of DNA)
- Individual barcodes (384 combinations with Nextera)

Key references (for protocole)

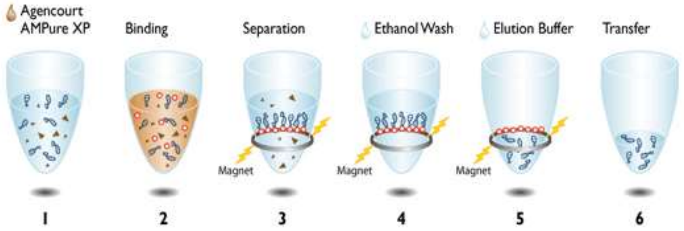
Baym M, Kryazhimskiy S, Lieberman TD et al. (2015) Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One*, 10, e0128036.

Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular Ecology Resources*, *17*(2), 194-208.

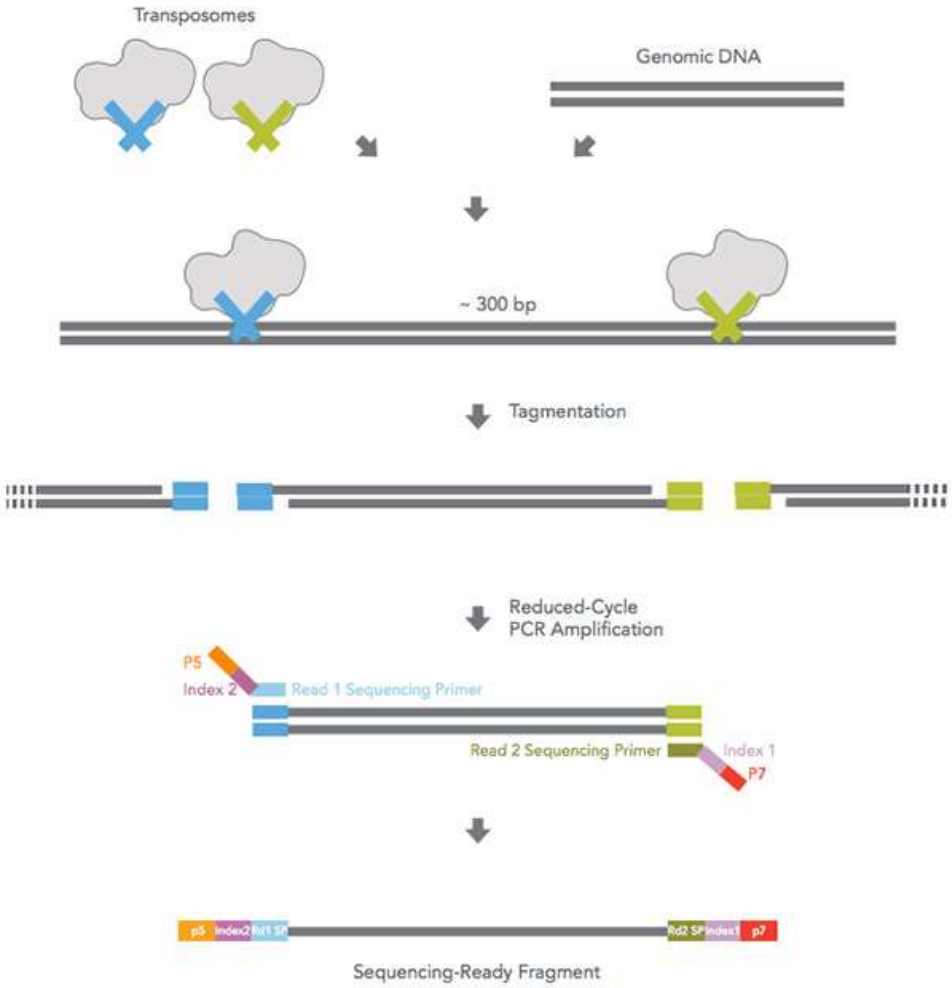
General idea of the library protocol

- Based on Nextera libraries (2 days/96 ind?)

1. Prepare/quantify DNA (the longest)
2. Cut DNA = *Tagmentation* (30 min)
3. Add Barcode & amplify DNA= *PCR* (2h)
4. Size-selection & cleaning (45 min)



5. QC & Quantification for pooling (? ½ day?)



Why using low-coverage data?

(+ low-cost libraries…)

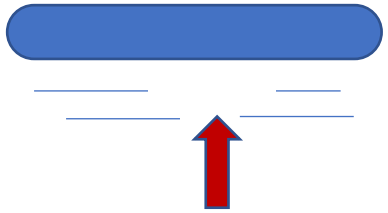
The matter of genomic complexity

Reduce costs:

USE 1ng of DNA

Same problem with degradated DNA (ancient DNA…)

Small genome -> ok



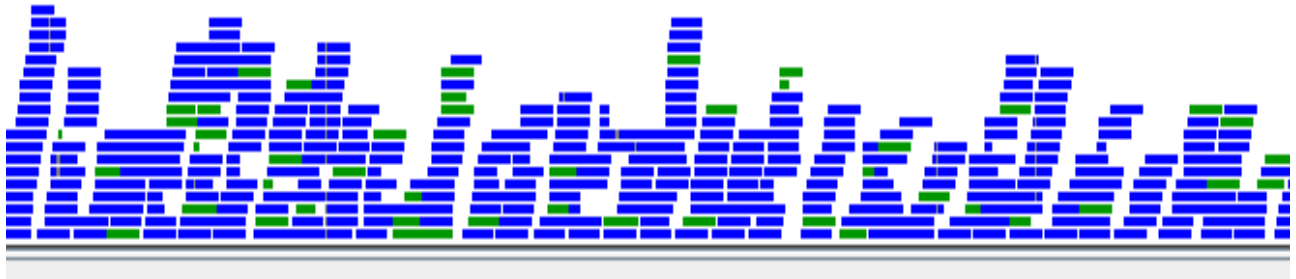
Big genome
-> Are we subetting too much the DNA and reducing the complexity of what we can sequence?

Include (many) different individuals

Run test librairies to adjust protocol to the study system

Run test sequencing lanes

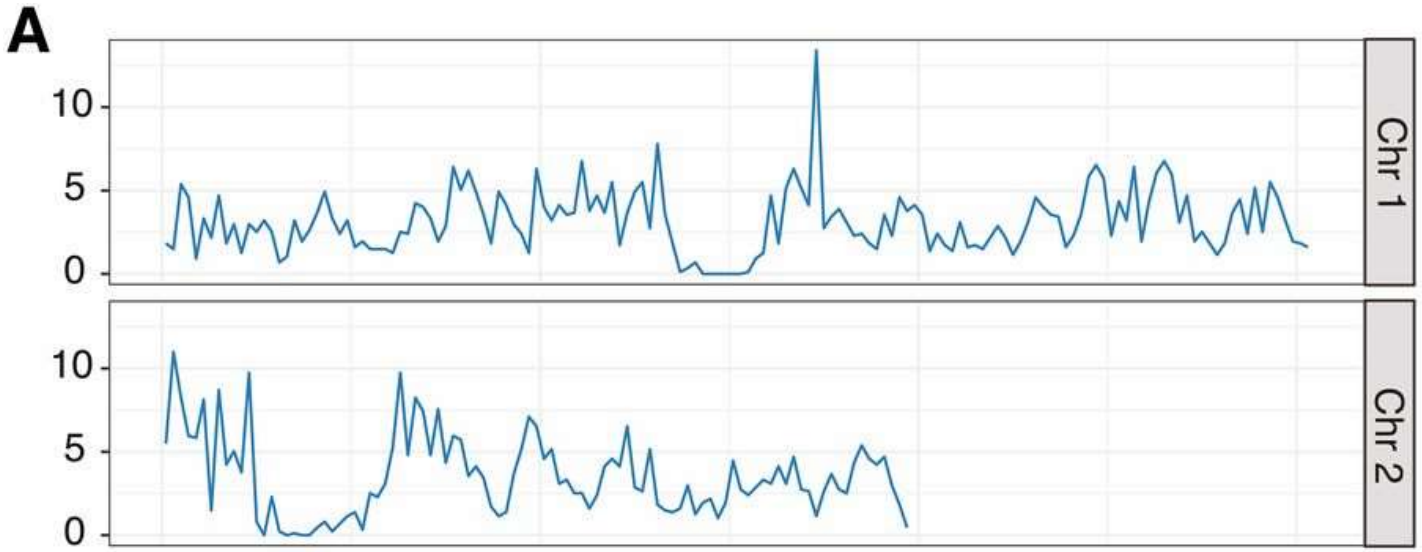
⇒ Evaluate coverage along the genome



Why using low-coverage data?

(+ low-cost libraries…)

Linkage map on 1920 progeny in Arabidopsis thaliana



Rowan, B. A., Heavens, D., Feuerborn, T. R., Tock, A. J., Henderson, I. R., & Weigel, D. (2019). An ultra high-density Arabidopsis thaliana crossover map that refines the influences of structural variation and epigenetic features. *Genetics*, *213*(3), 771-787. https://doi.org/10.1534/genetics.119.302406

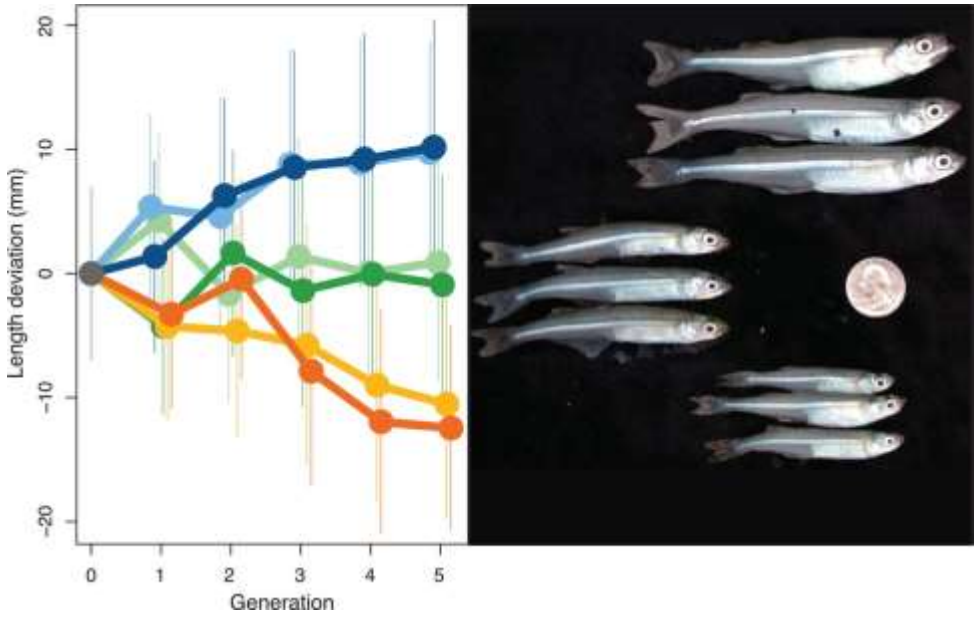
⇒ Super fine-scale resolution of crossing-over thanks to

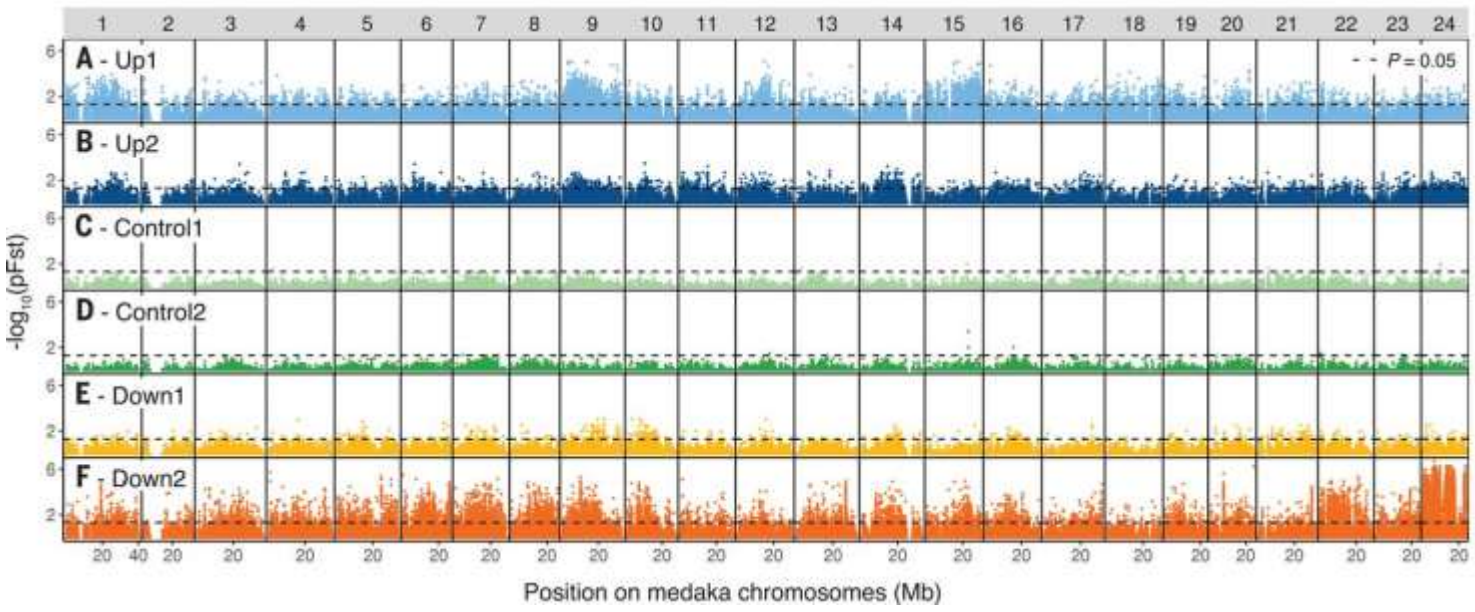
- Many recombination events (large family)
- Very dense markers (whole-genome!)

Why using low-coverage data?

(+ low-cost libraries…)

Experimental selection with 6 replicates of 50 individuals





Therkildsen, N. O., Wilder, A. P., Conover, D. O., Munch, S. B., Baumann, H., & Palumbi, S. R. (2019). Contrasting genomic shifts underlie parallel phenotypic evolution in response to fishing. *Science*, *365*(6452), 487-490. https://doi.org/10.1126/science.aaw7271

⇒ Genome-wide evolution of allelic frequencies

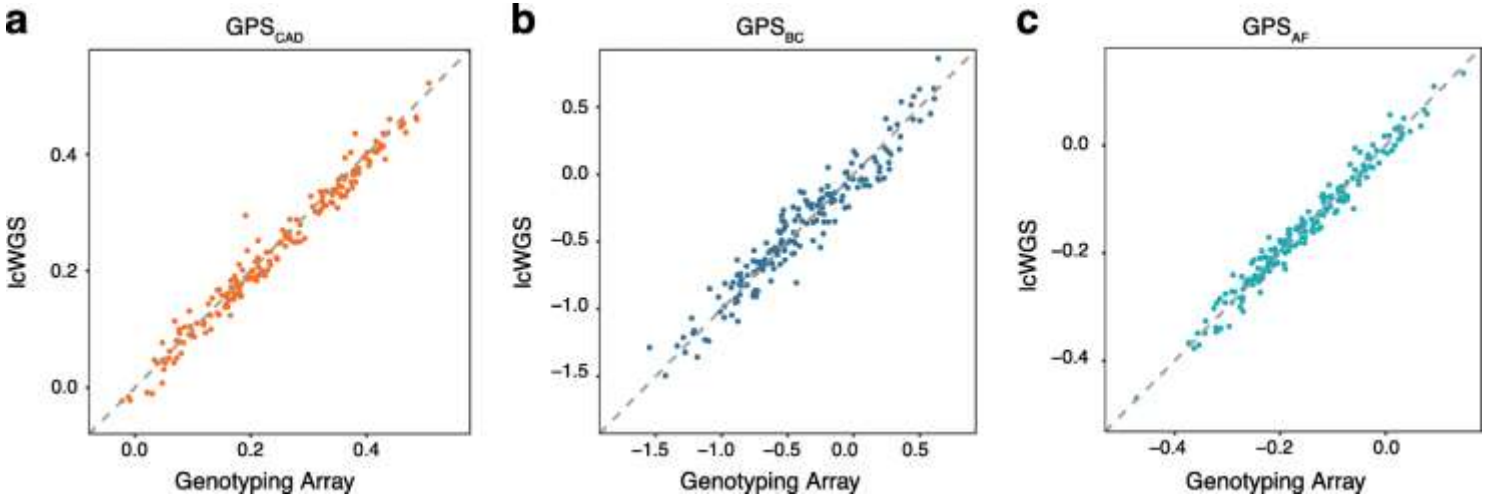
⇒ No bias due to pooling contrary to pool-seq

Why using low-coverage data?

(+ low-cost libraries…)

GWAS with > 11,000 whole genomes in humans

Genome-polygenic scores



"lcWGS provides comparable imputation accuracy while also overcoming the ascertainment bias inherent to variant selection in genotyping array design"

Homburger, J. R., Neben, C. L., Mishne, G., Zhou, A. Y., Kathiresan, S., & Khera, A. V. (2019). Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. *Genome medicine*, *11*(1), 1-12. https://doi.org/10.1186/s13073-019-0682-2

⇒ An efficient alternative to SNParray

Why using low-coverage data?

(+ low-cost libraries…)

Pros and Cons

- Relatively cheap
- Keep individual information
- Cover the whole genome
- Genotype likelihood based methods are now well-developped

- Hard-calling of genotype is not possible
- Population-level analysis: need to be able to gather samples (at least 30-50 per pop)
- Check heterogeneity of coverage along the genome
- Need reference

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*
- All whole-genome data
- Flexible inputs
- Multiple methods, filters, etc.
- Large datasets
- Many downstream analyses
- Documentation ok – reactivity Github

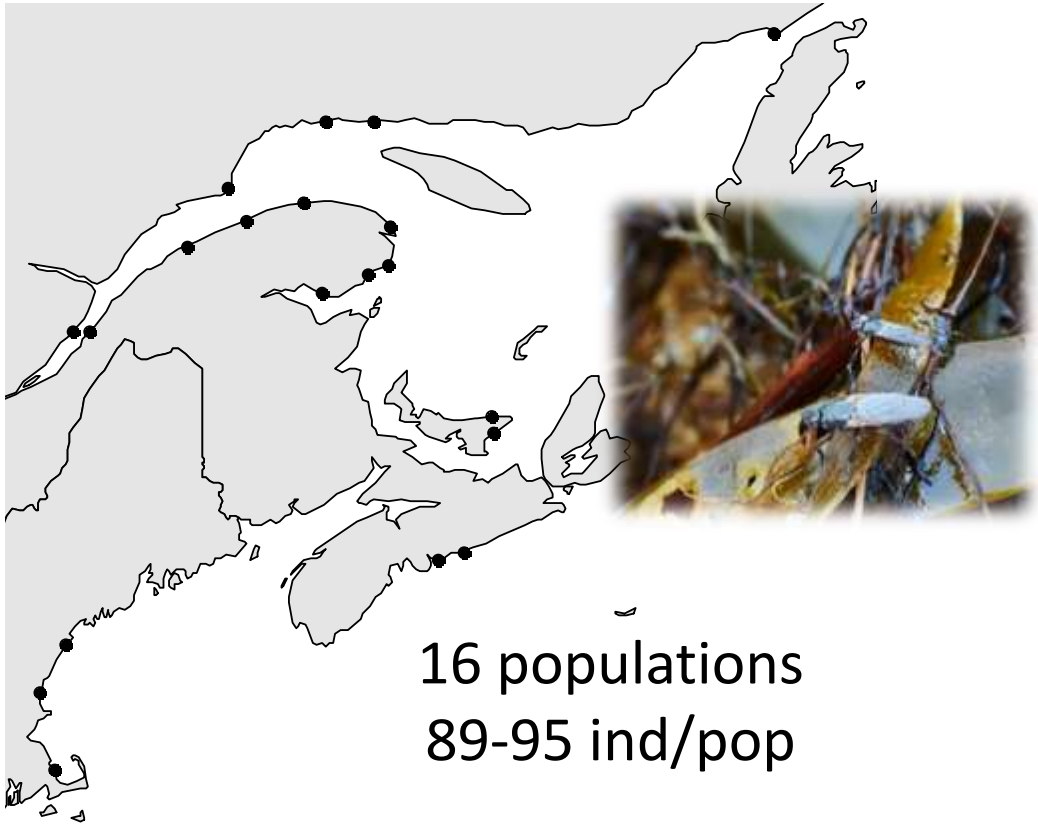
Inconvenients:

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

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

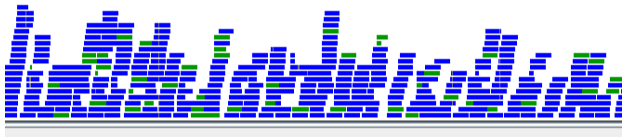
https://github.com/ANGSD/angsd

Example for population genomics…
Coelopa frigida

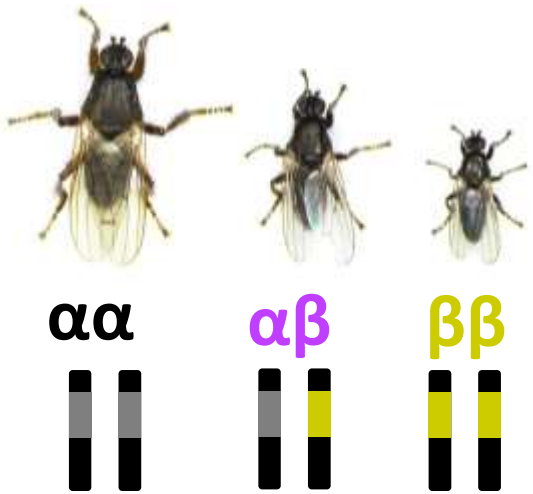


16 populations
89-95 ind/pop

Mean coverage:
1.2x / ind
100x / pop

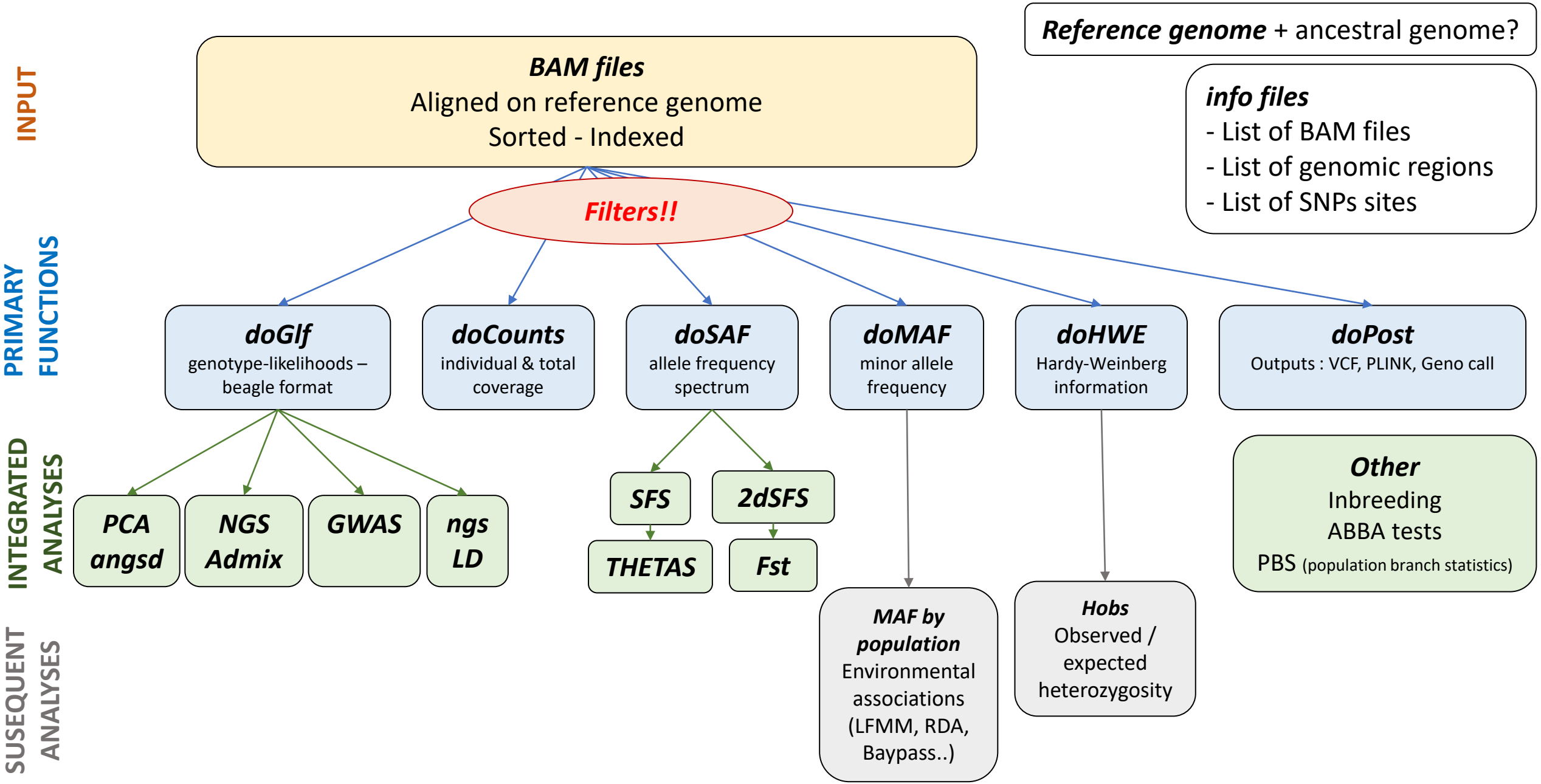


- Population structure?
(Geography? Chromosomal inversion?)
- Environmental associations?
- Linkage disequilibrium?
- Sex chromosome?
- Demography?

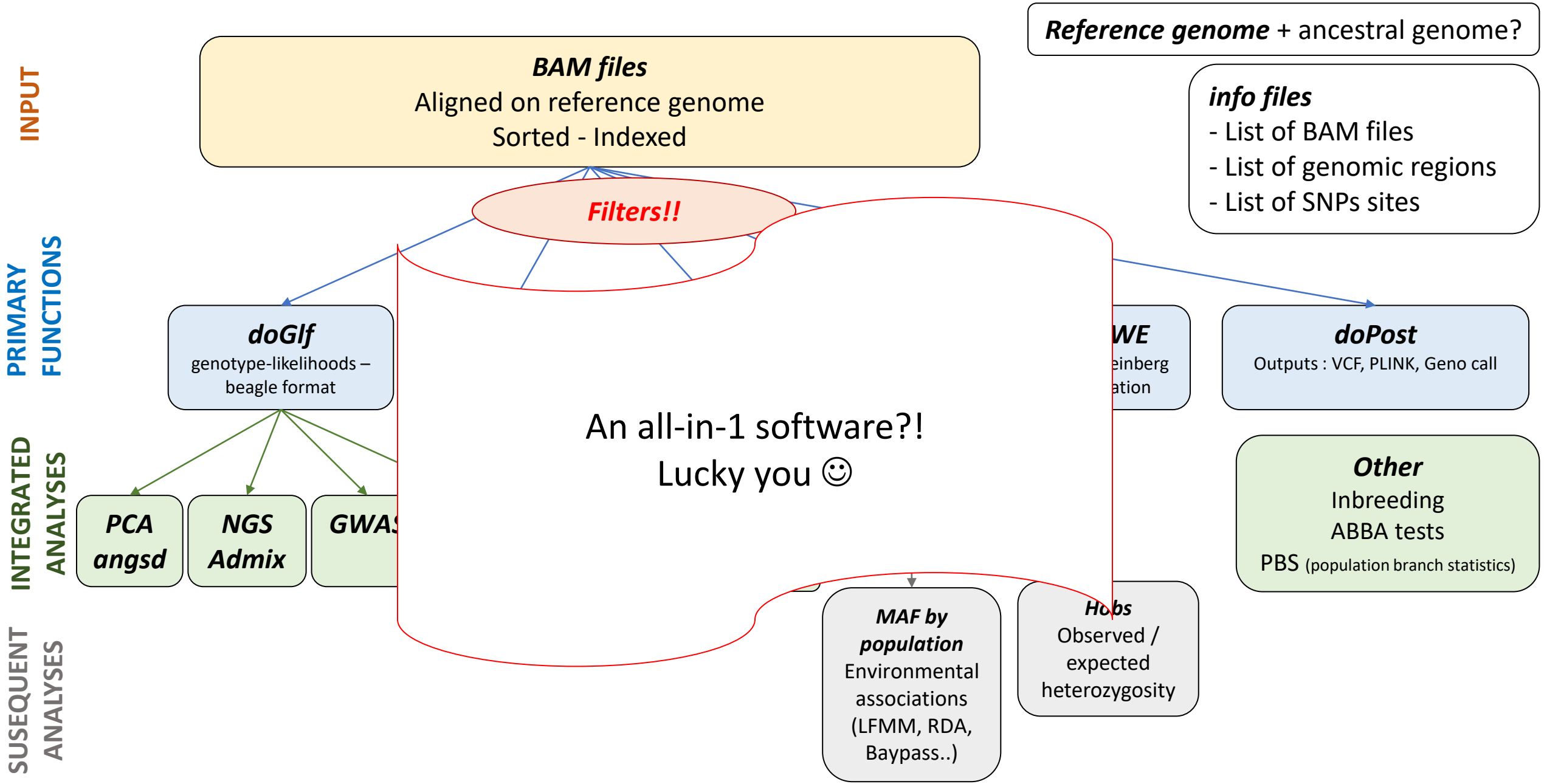


https://github.com/clairemerot/angsd_pipeline

ANGSD overview



ANGSD overview



ANGSD filters

BAM files

Filters!!

BAM quality

Coverage:
Minimum number of individuals
Minimum depth
Maximum depth

Minor Allele Frequency:
Minimum frequency
Probability of being a polymorphic site

Hardy-Weinberg:
sites at HW equilibrium

⇒ List of SNPs sites
⇒ Variants / invariants?

ANGSD : using Genotype likelihoods

INTEGRATED ANALYSES

doGlf
genotype-likelihoods – beagle format

PCA angsd

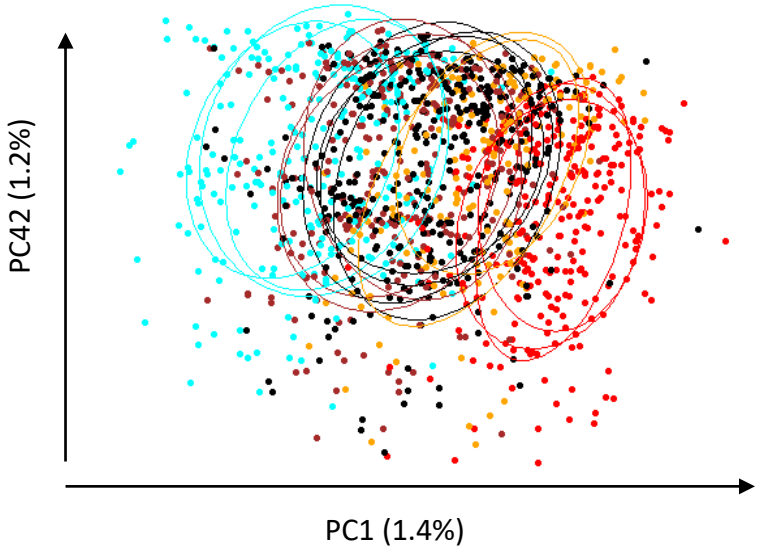
NGS Admix

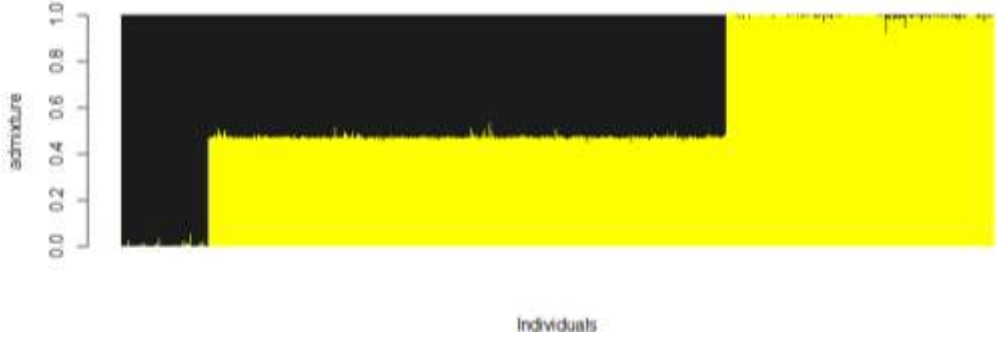
marker	allele1	allele2	Ind0	Ind0	Ind0	Ind1	Ind1	Ind1
LG1_3867	3	1	0.799992	0.200008	0.000000	0.333333	0.333333	0.333333
LG1_3870	1	0	0.799985	0.200015	0.000000	0.333333	0.333333	0.333333
LG1_3880	1	2	0.000000	0.200015	0.799985	0.333333	0.333333	0.333333
LG1_7206	2	1	0.888863	0.111137	0.000000	0.333333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

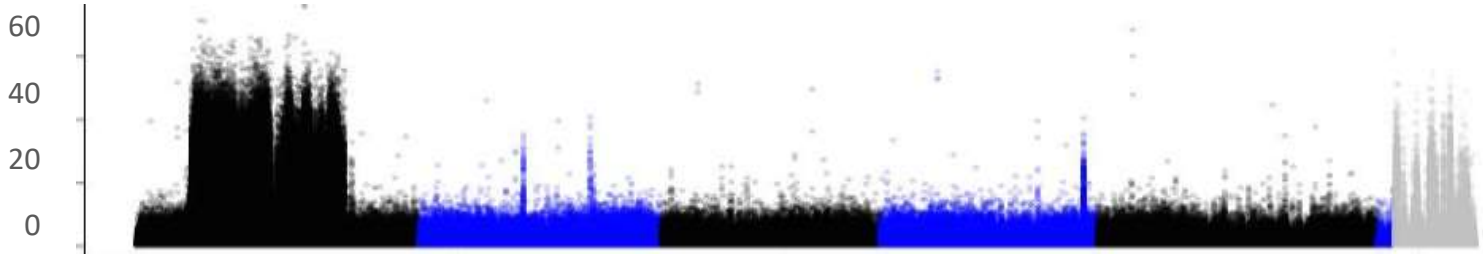
GWAS

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

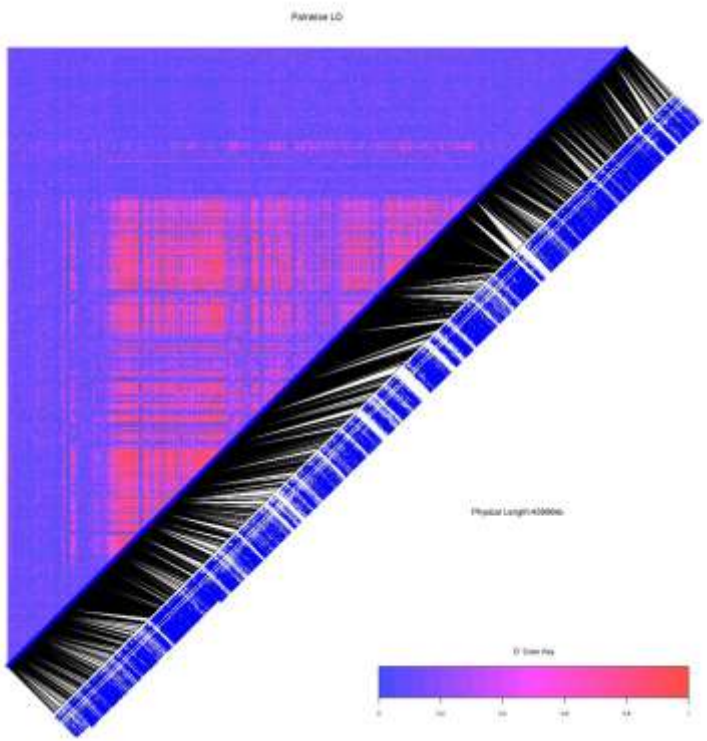
ngs LD

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)

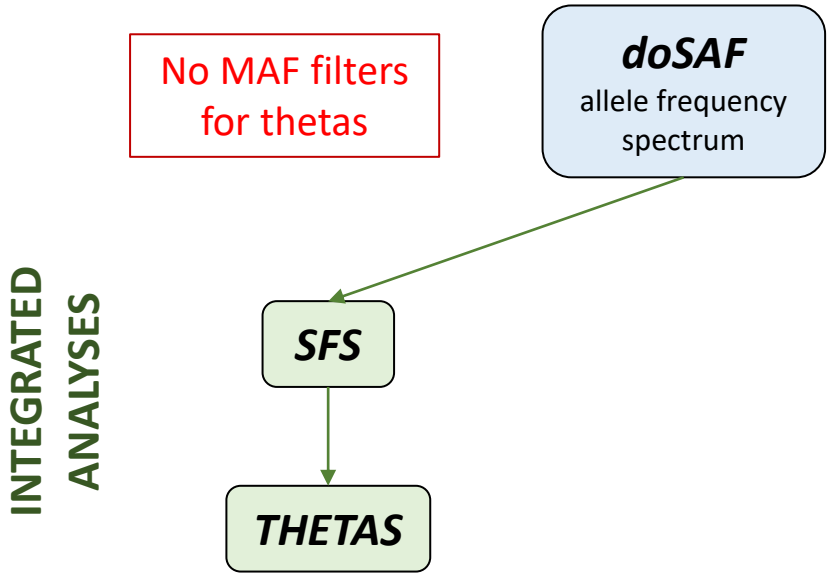
Explore Linkage disequilibrium

Chr I

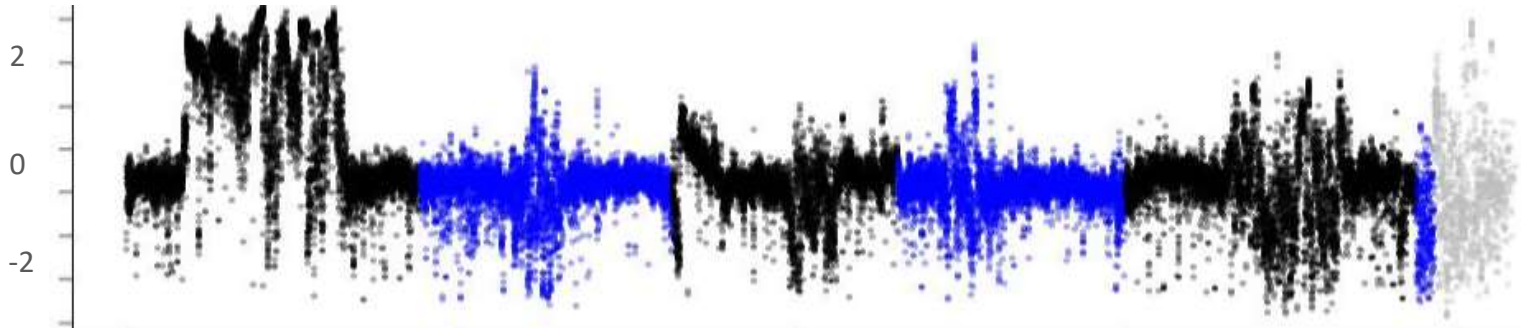


Fox, E. A., Wright, A. E., Fumagalli, M., & Vieira, F. G. (2019). ngsLD: evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics*.

ANGSD : Allele frequency spectrums & statistics



Tajima's D



Thetas Watterson
Thetas diversity
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

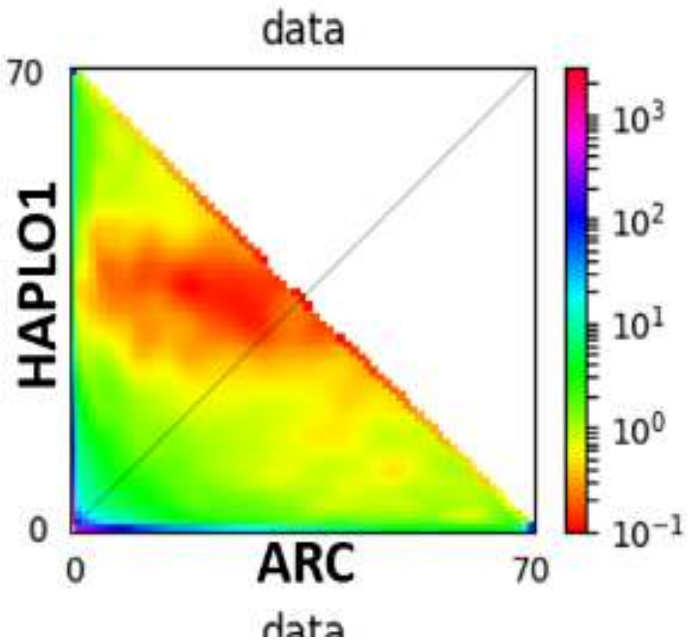
No MAF filters for demography

INTEGRATED ANALYSES

doSAF
allele frequency spectrum

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?

INTEGRATED ANALYSES

doSAF
allele frequency spectrum

↓

2dSFS

↓

Fst

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

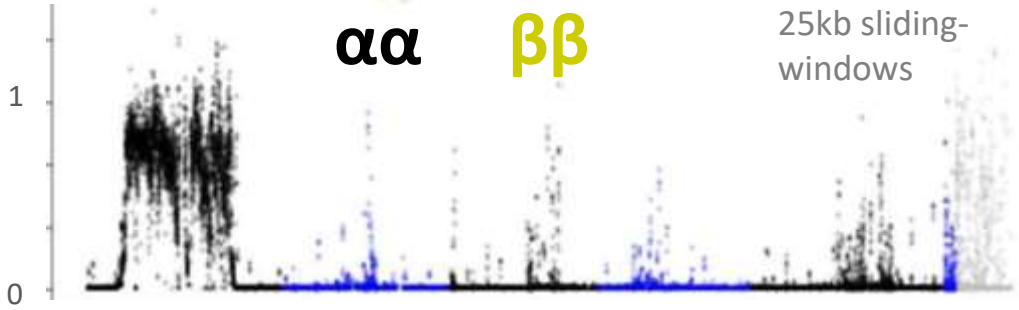
North

South

Pairwise FST between populations

FST

αα ββ



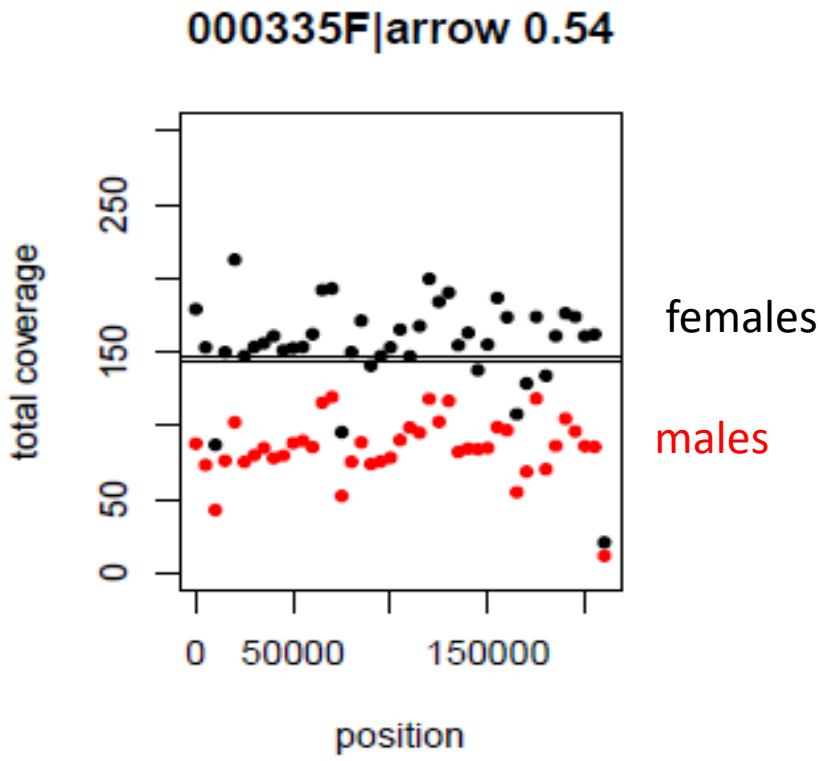
25kb sliding-windows

Output by sliding-windows…

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



Check homogeneity along the genome and between samples

ANGSD : Minor allele frequency

doMAF
minor allele frequency

MAF by population
Environmental associations (LFMM, RDA, Baypass..)

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

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)

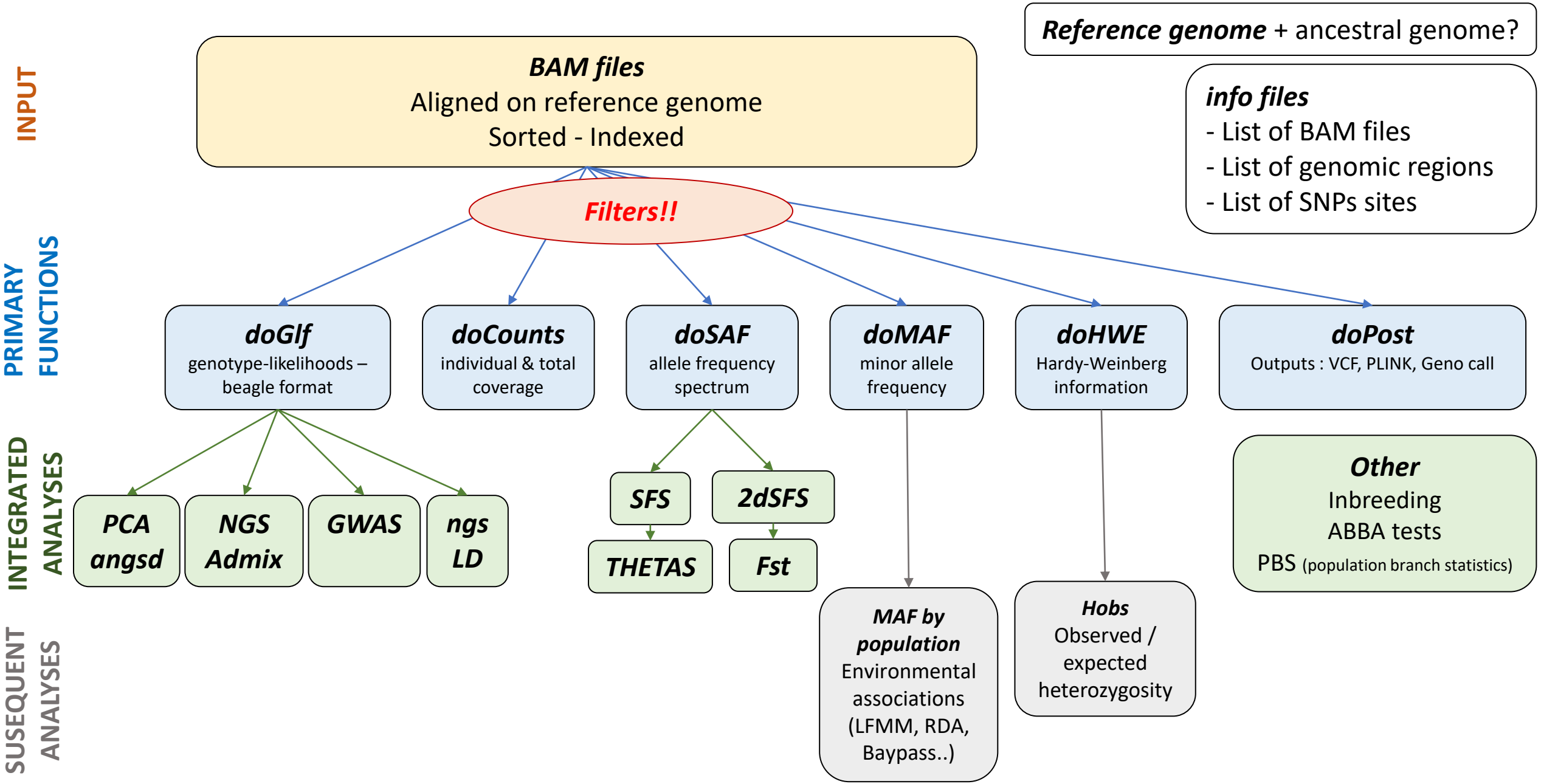
POSSIBILITY: use SITES list of filtered SNPs + Maj/Min info

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

Environmental associations
(baypass, lfmm, RDA)

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??