Low-coverage whole-genome re-sequencing

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

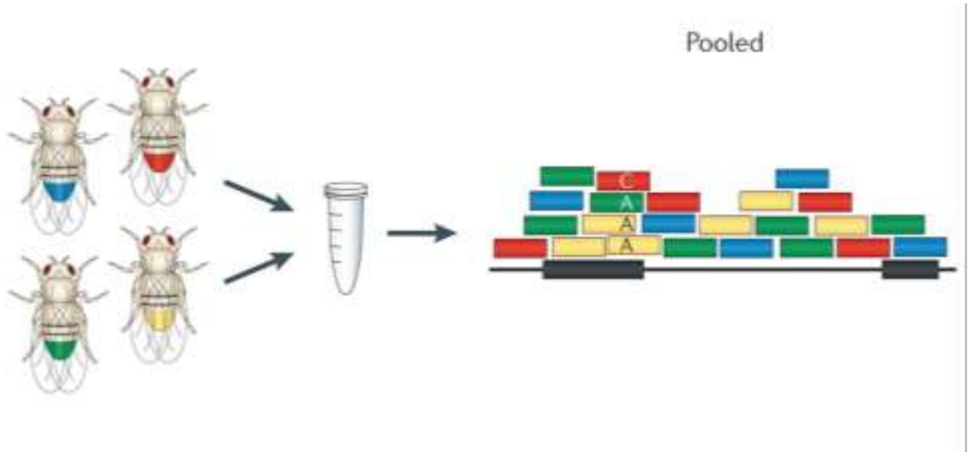
Yes, I have many samples
I want to

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

Reduced-
r ent
(Ra IP-
c ture,e)

I want whole-genome!

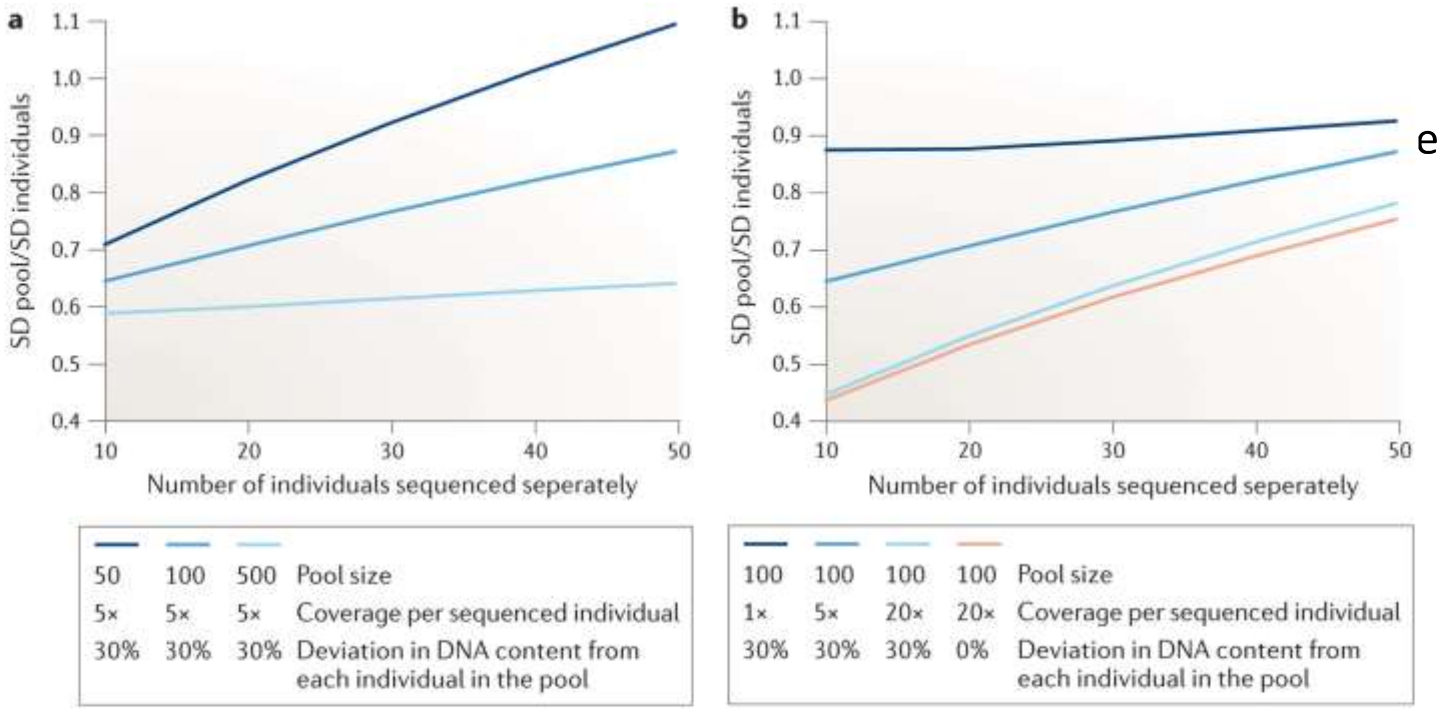
A possible solution:
Pool-seq!



Why using low-coverage data?

(+ low-cost libraries…)

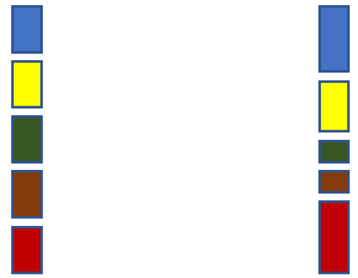
Short note about Pool-seq



Nature Reviews | Genetics

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

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



⇒ Need replicates of pools (>3)

Expected pooling

True pooling

+ 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-~~
~~r ent~~
~~(Ra IP-~~
~~c 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?

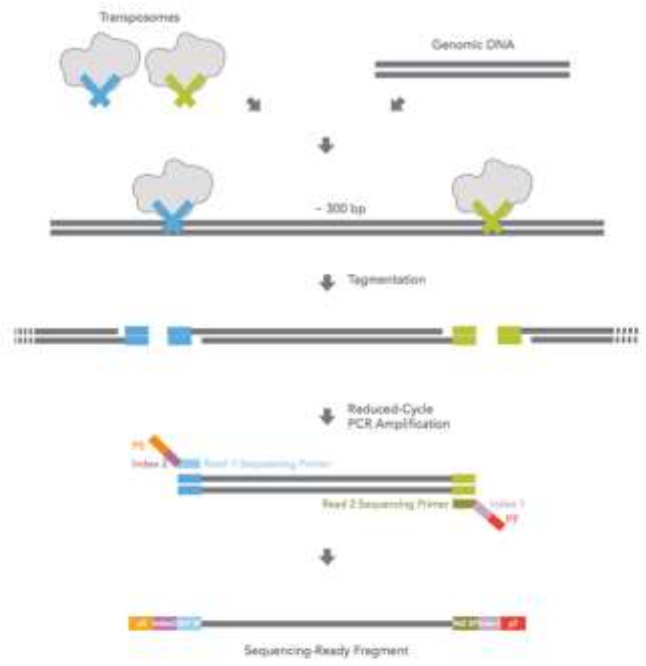
(+ low-cost libraries…)

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.

Why using low-coverage data?

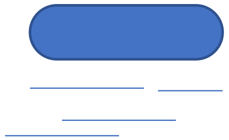
(+ low-cost libraries…)

The matter of genomic complexity

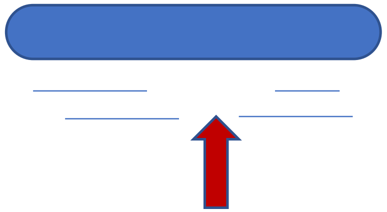
Reduce costs:

USE 1ng of DNA

Same problem with degradated DNA (ancient DNA…)



Small genome -> ok



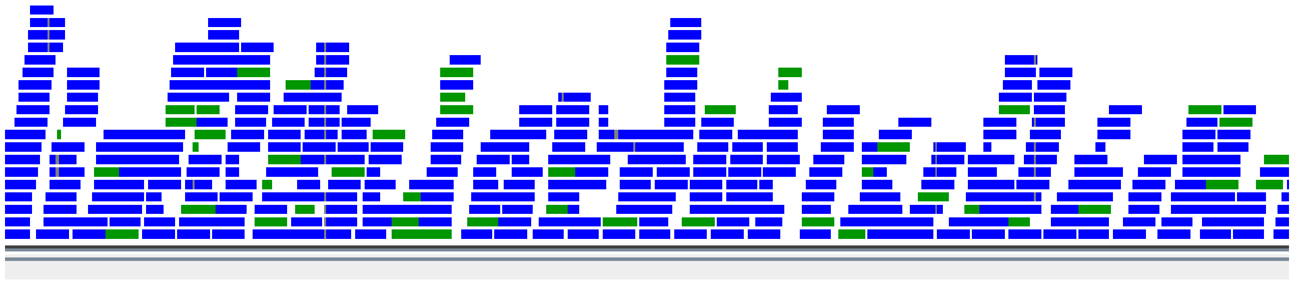
Big genome
 -> Are we subetting too much the DNA and reducing the complexity of what we can sequence?

Include (many) different individuals

Run test librairies to adjust protocol to the study system

Run test sequencing lanes

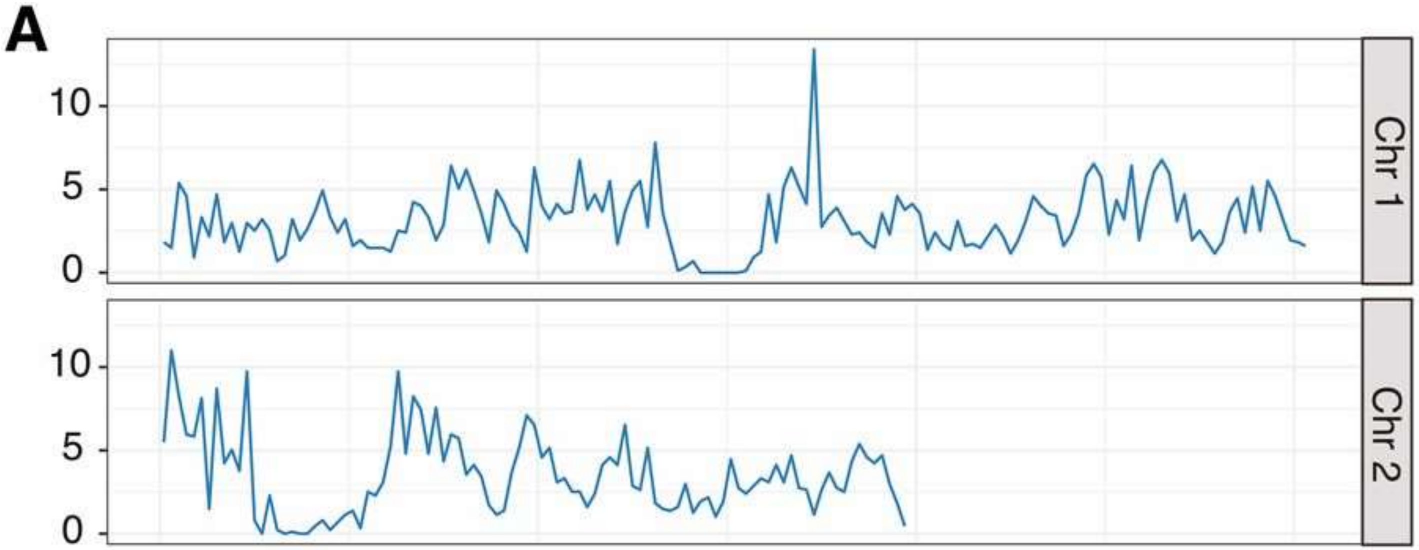
⇒ Evaluate coverage along the genome



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

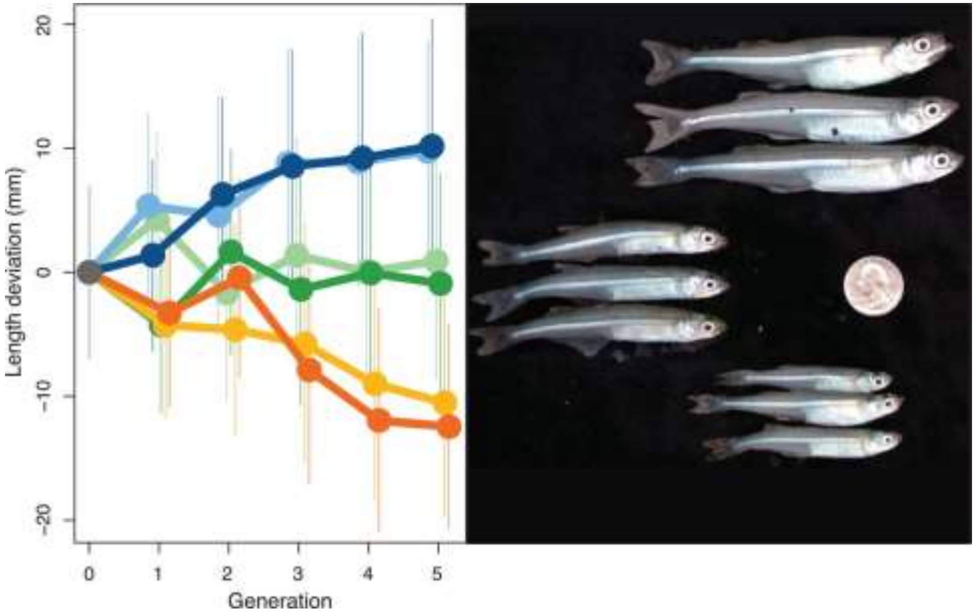
⇒ Super fine-scale resolution of crossing-over thanks to

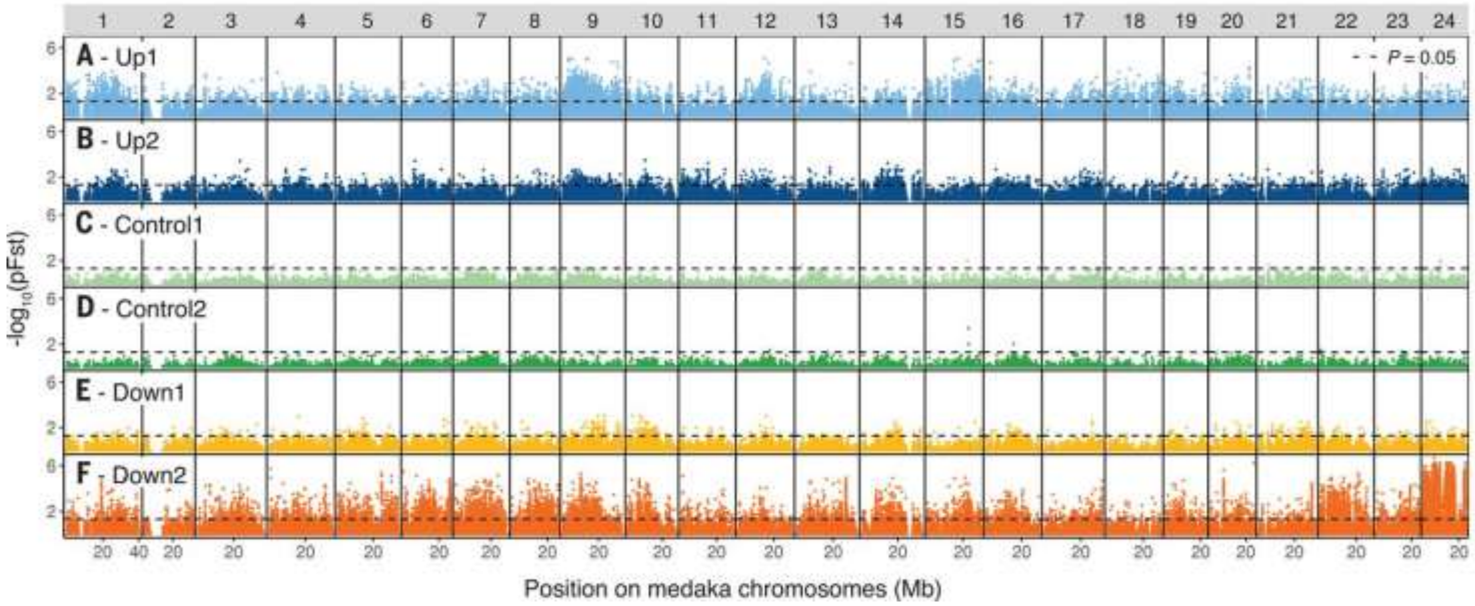
- Many recombination events (large family)
- Very dense markers (whole-genome!)

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

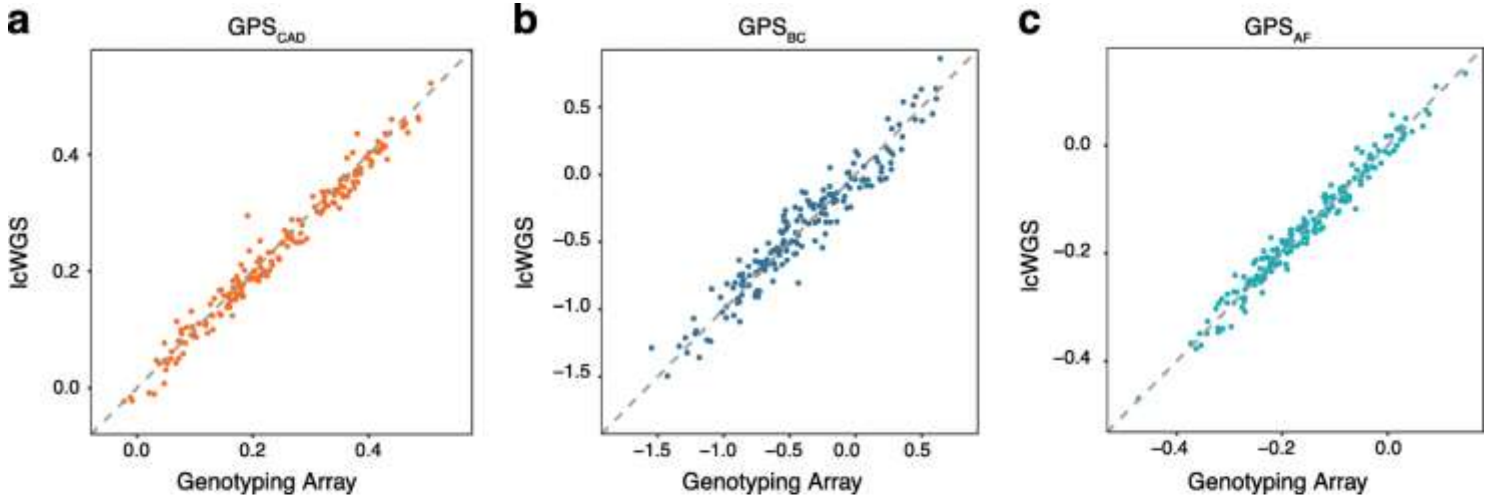
⇒ No bias due to pooling contrary to pool-seq

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"

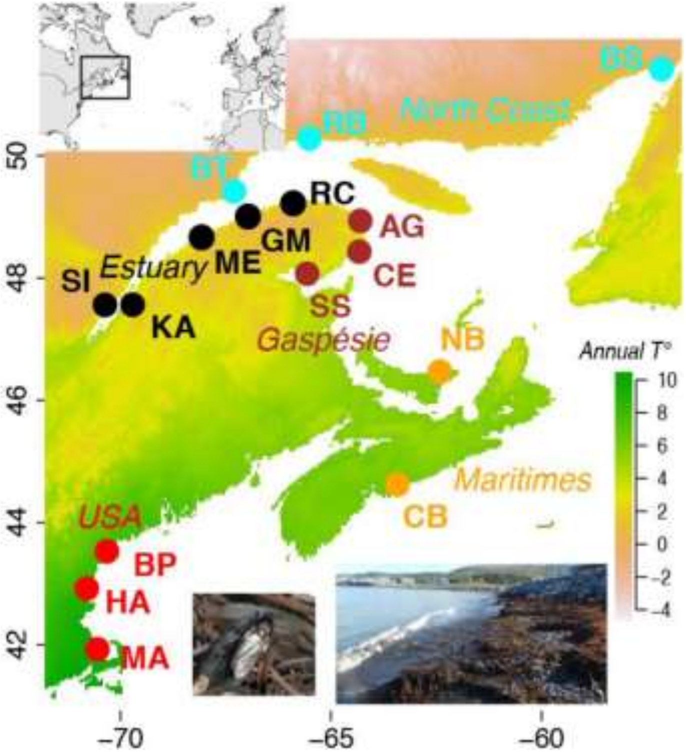
Homburger, J. R., Neben, C. L., Mishne, G., Zhou, A. Y., Kathiresan, S., & Khera, A. V. (2019). Low coverage whole genome sequencing enables accurate assessment of common variants and calculation of genome-wide polygenic scores. *Genome medicine*, *11*(1), 1-12.
https://doi.org/10.1186/s13073-019-0682-2

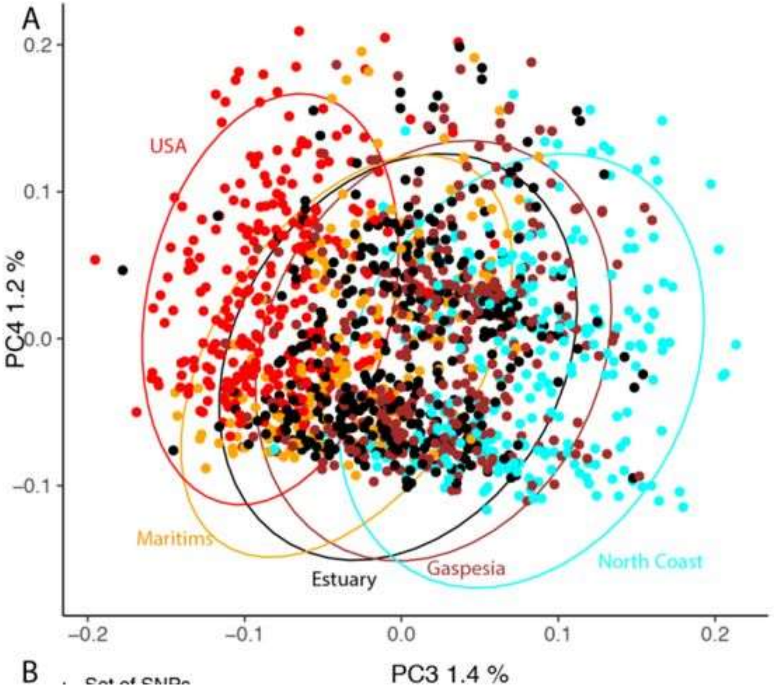
⇒ An efficient alternative to SNParray

Why using low-coverage data?

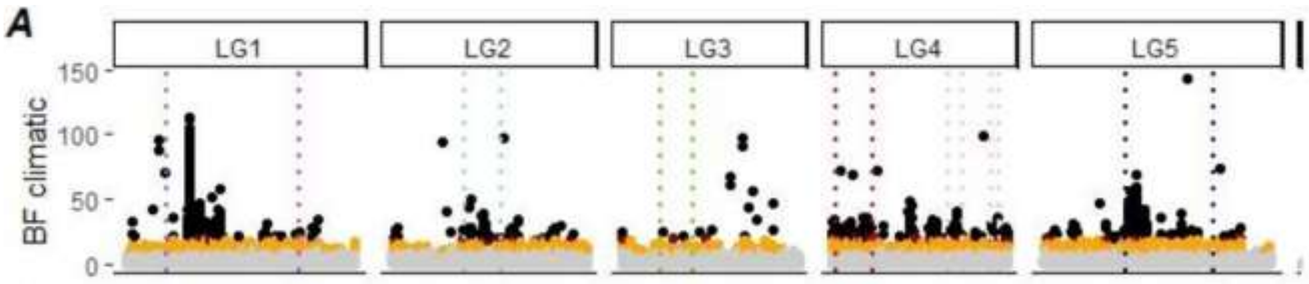
(+ low-cost libraries…)

Population genomics with ~1, 500 flies





⇒ Population structure, environmental associations, inversion detection



Locally adaptive inversions modulate genetic variation at different geographic scales in a seaweed fly
Claire Mérot, Emma Berdan, Hugo Cayuela, Haig Djambazian, Anne-Laure Ferchaud, Martin Laporte, Eric Normandeau, Jiannis Ragoussis, Maren Wellenreuther, Louis Bernatchez
bioRxiv 2020.12.28.424584; doi: https://doi.org/10.1101/2020.12.28.424584

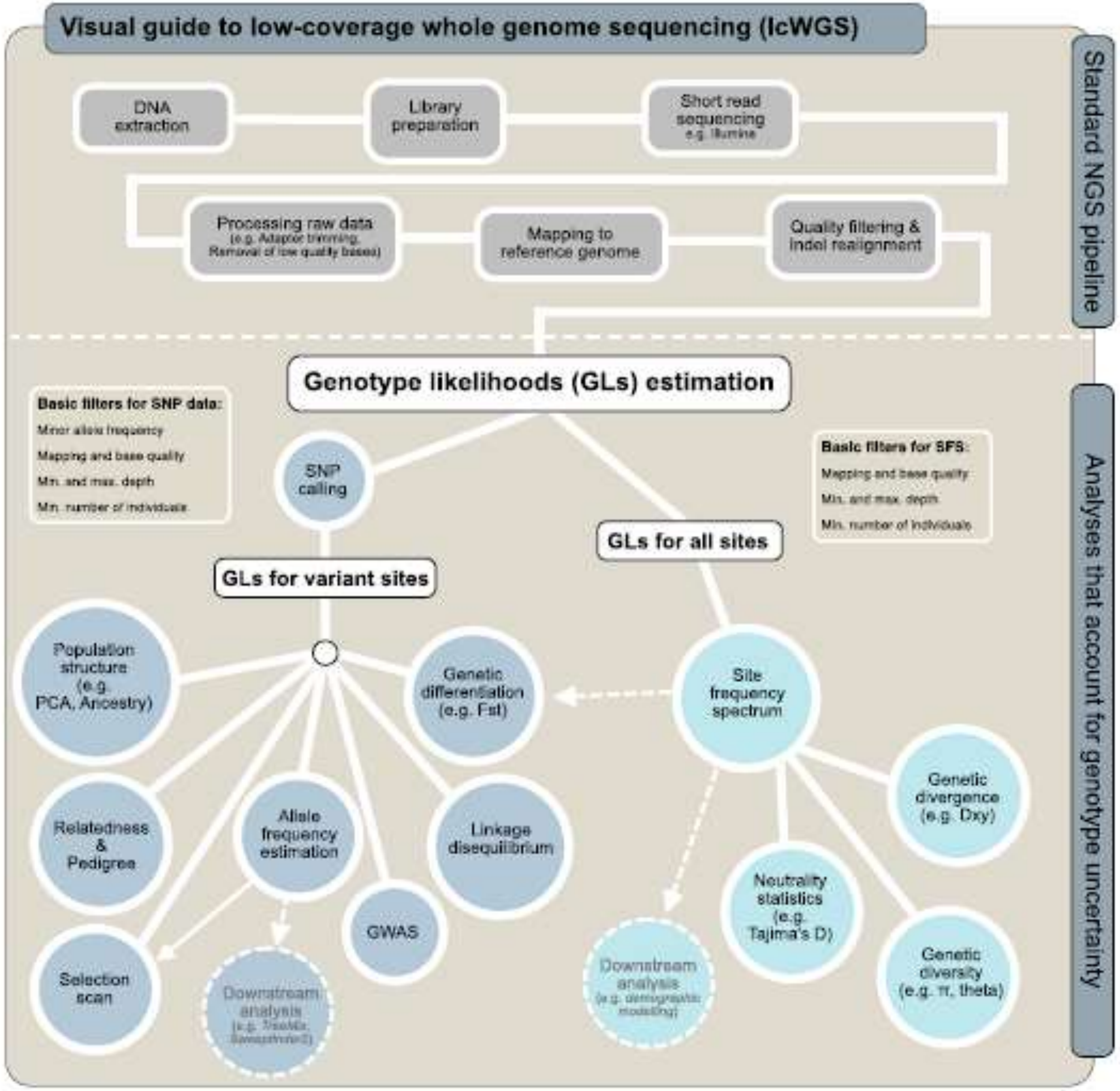
How to use low-coverage data?

Runyang Nicolas Lou, Arne Jacobs, Aryn Wilder, et al.

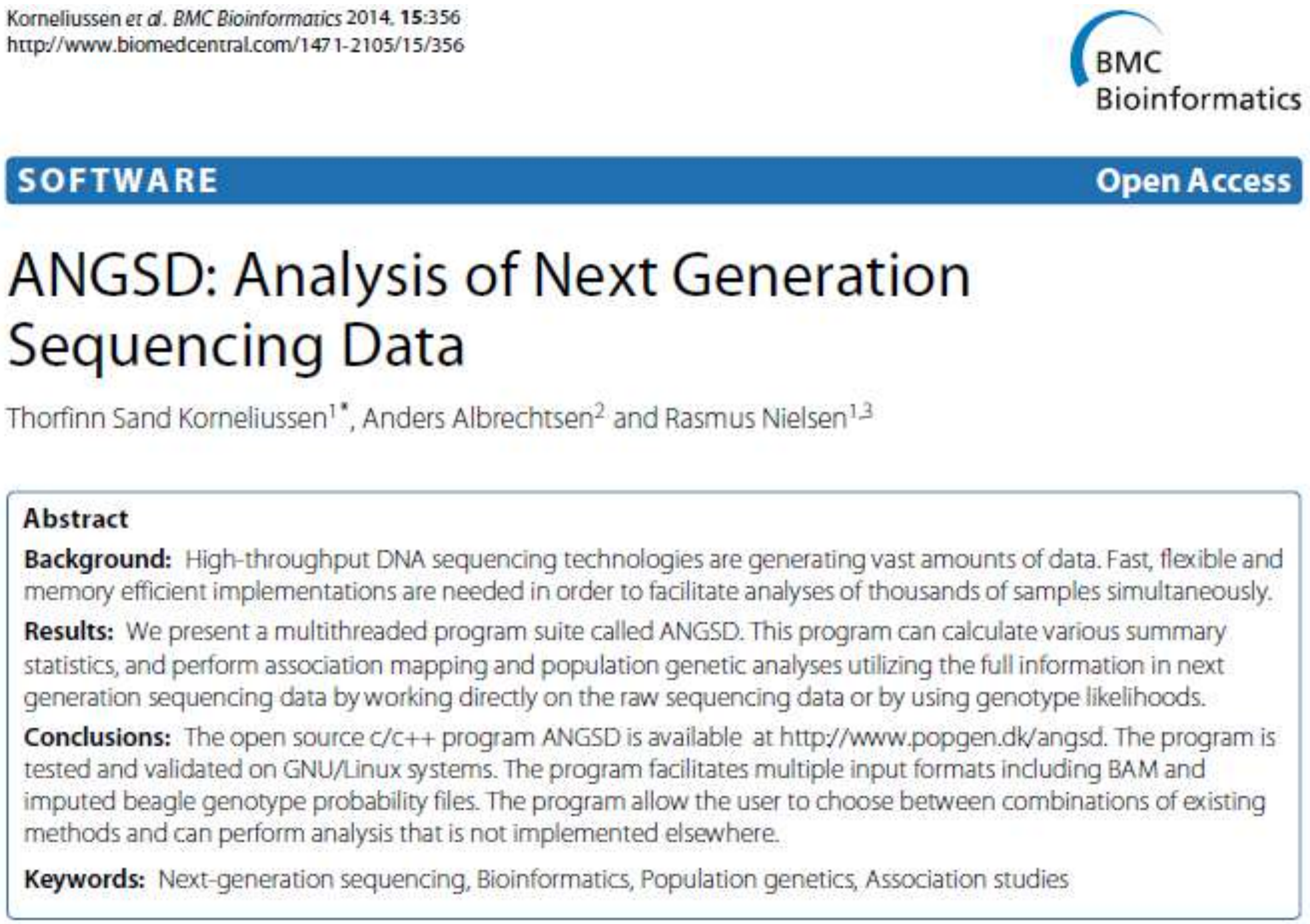
A beginner's guide to low-coverage whole genome sequencing for population genomics.

Authorea. April 21, 2021.

DOI: 10.22541/au.160689616.68843086/v3


Visual guide to low-coverage whole genome sequencing (lcWGS)
Standard NGS pipeline
DNA extraction
Library preparation
Short read sequencing e.g. Illumina
Processing raw data (e.g. Adapter trimming, Removal of low quality bases)
Mapping to reference genome
Quality filtering & Indel realignment
Analyses that account for genotype uncertainty
Genotype likelihoods (GLs) estimation
Basic filters for SNP data:
Minor allele frequency
Mapping and base quality
Min. and max. depth
Min. number of individuals
SNP calling
Basic filters for SFS:
Mapping and base quality
Min. and max. depth
Min. number of individuals
GLs for all sites
GLs for variant sites
Population structure (e.g. PCA, Ancestry)
Genetic differentiation (e.g. Fst)
Site frequency spectrum
Genetic divergence (e.g. Dxy)
Relatedness & Pedigree
Allele frequency estimation
Linkage disequilibrium
GWAS
Neutrality statistics (e.g. Tajima's D)
Genetic diversity (e.g. π, theta)
Selection scan
Downstream analysis (e.g. TreeMix, [illegible])
Downstream analysis (e.g. demographic modeling)

How to use low-coverage data?



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

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

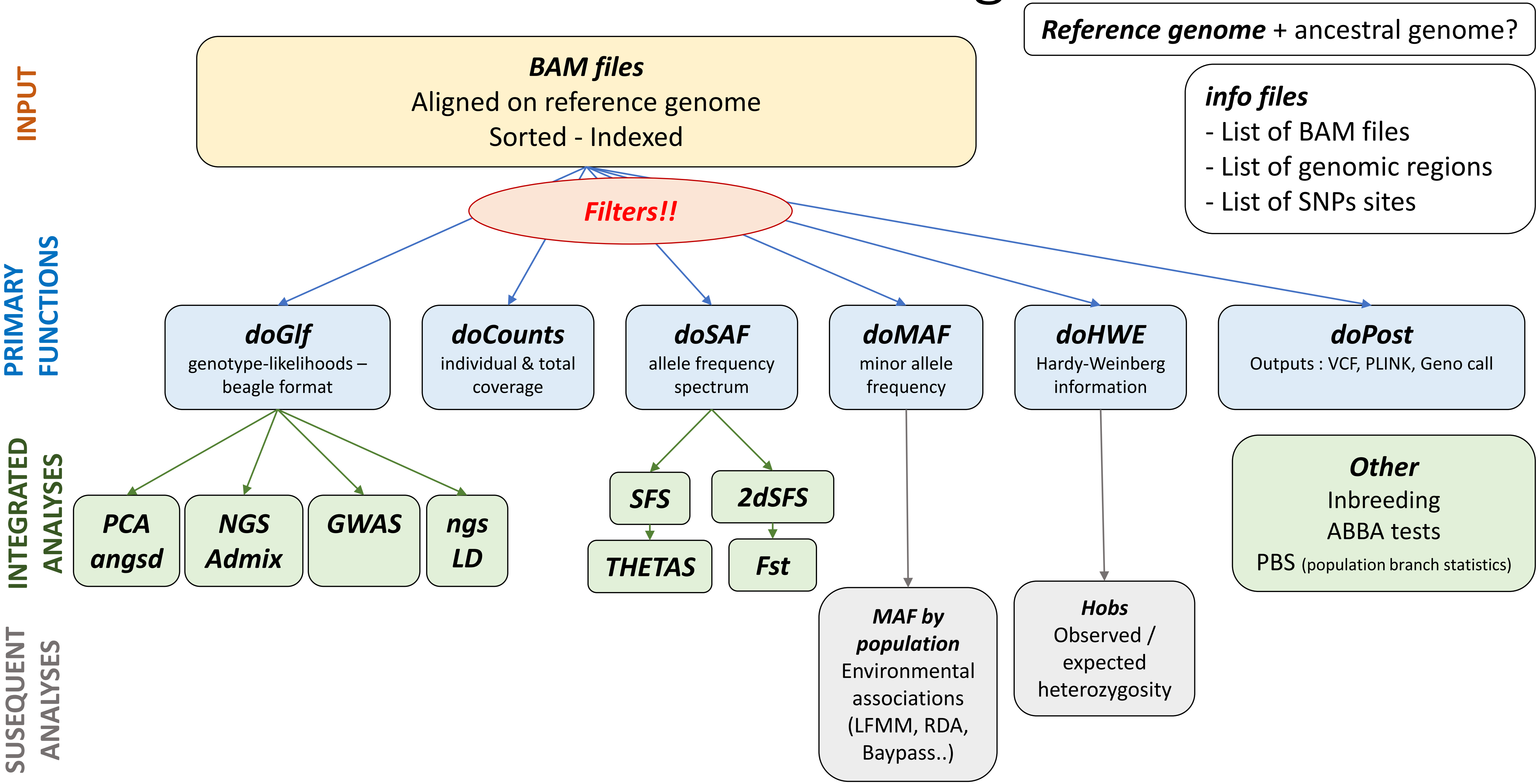
Inconvenients:

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

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

https://github.com/ANGSD/angsd

How to use low-coverage data?



Reference genome + ancestral genome?

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

INPUT

BAM files
Aligned on reference genome
Sorted - Indexed

Filters!!

PRIMARY FUNCTIONS

doGlf
genotype-likelihoods – beagle format

doCounts
individual & total coverage

doSAF
allele frequency spectrum

doMAF
minor allele frequency

doHWE
Hardy-Weinberg information

doPost
Outputs : VCF, PLINK, Geno call

INTEGRATED ANALYSES

PCA angsd

NGS Admix

GWAS

ngs LD

SFS

THETAS

2dSFS

Fst

Other
Inbreeding
ABBA tests
PBS (population branch statistics)

SUSEQUENT ANALYSES

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

Hobs
Observed / expected heterozygosity

How to use low-coverage data?

Reference genome + ancestral genome?

info files
- List of BAM files

INPUT

A

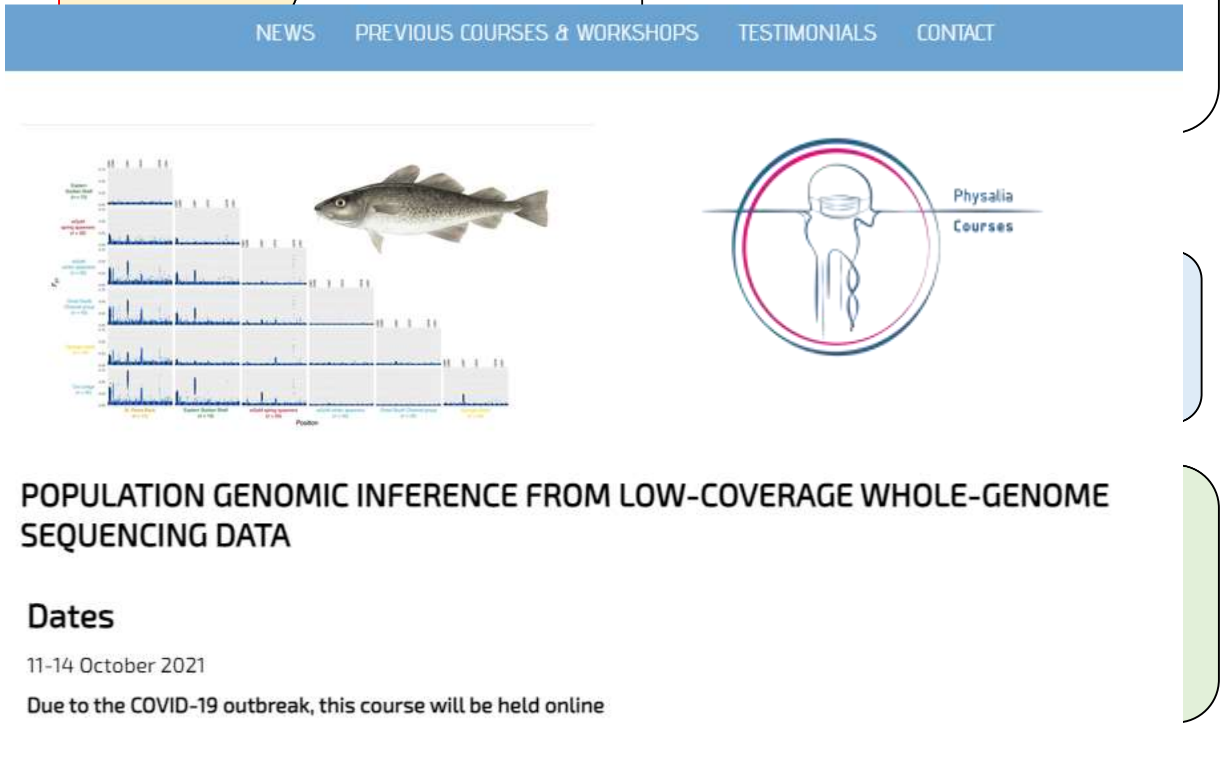
PRIMARY

INTEGRATED

- An all-in-1 software?! Lucky you ☺
  - Pop structure: PCA, Admixture, Fst
    - Allelic frequencies
  - GWAS, H-W test, coverage, statistics (Taj D, Pi, etc)

- Help:

https://github.com/clairemerot/angsd_pipeline
https://github.com/nt246/lcwgs-guide-tutorial



POPULATION GENOMIC INFERENCE FROM LOW-COVERAGE WHOLE-GENOME SEQUENCING DATA

Dates

11-14 October 2021

Due to the COVID-19 outbreak, this course will be held online

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

Observed / expected heterozygosity

SUSEQUENT
AN

Pros/Cons low-coverage WGS?

(+ low-cost libraries…)

Pros and Cons

+

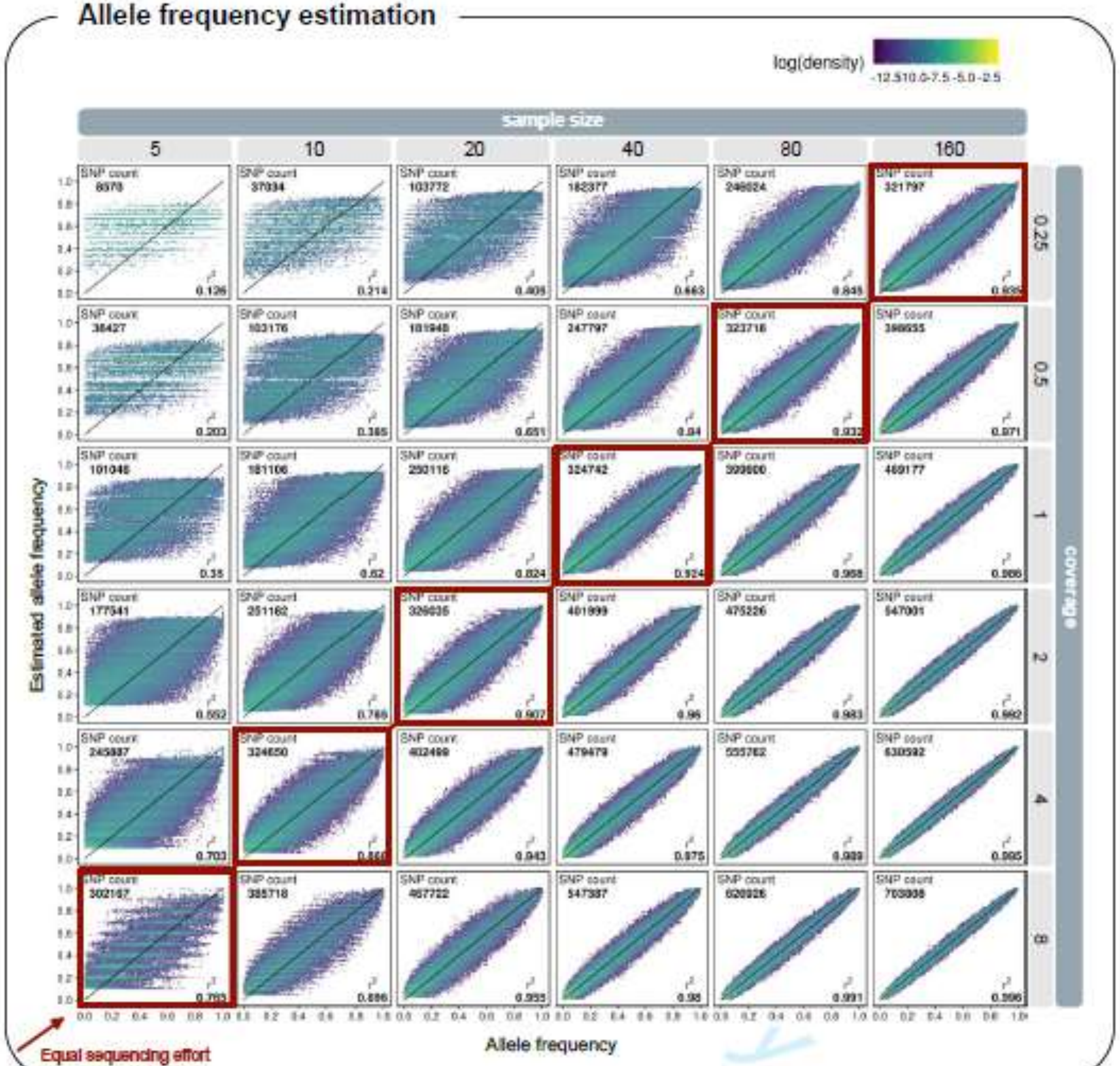
- Relatively cheap
- Keep individual information
- Cover the whole genome
- Genotype likelihood based methods are now well-developped

−

- 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

When not do lcWGS?

Too few samples



When not do lcWGS?

???