GBSathon

Benchmarking reproducibility of Genotyping-By-Sequencing analysis workflows through comparison with SNP chip and pedigree data

Rachael Ashby, Rudiger Brauning, Hayley Baird, Ruy Jauregui, Monica Vallender, Aurelie Laugraud, Charles Hefer, Abdul Baten, Paul Maclean, Rayna Anderson, Roger Moraga, Siva Ganesh, Tracey van Stijn, Jeanne Jacobs, Ken Dodds, John McEwan, Shannon Clarke and Andrew Griffiths

agresearch
āta mātai, mātai whetū

agresearch
āta mātai, mātai whetū

Apologies

- Rachael can’t be joining us this week.

agresearch
āta mātai, mātai whetū

AgResearch

- Crown Research Institute
- Delivering cost effective, high-quality, high-throughput solutions to primary industry and more

agresearch
āta mātai, mātai whetū

United in Data, but…

- when working with GBS data we noticed that people have different favourite tools

agresearch
āta mātai, mātai whetū

The challenge

- Find the best GBS analysis workflow
- Convince your peers

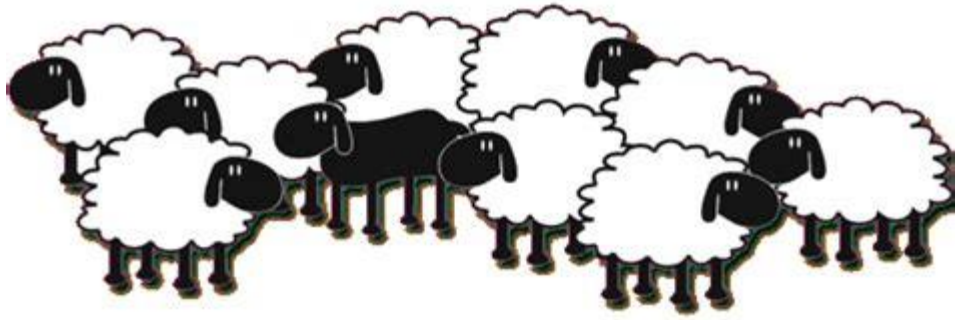
agresearch
āta mātai, mātai whetū

The team

- Lab
  - Hayley Baird, Rayna Anderson, Tracey van Stijn
- Bioinformatics
  - Rachael Ashby, Rudiger Brauning, Ruy Jauregui, Monica Vallender, Aurelie Laugraud, Charles Hefer, Abdul Baten,, Roger Moraga
- Stats
  - Paul Maclean, Siva Ganesh, Ken Dodds
- PI
  - Jeanne Jacobs, John McEwan, Shannon Clarke, Andrew Griffiths

agresearch
āta mātai, mātai whetū

The test species



- Sheep, well behaved diploid species 3Gbp genome
- Previously pooled samples of highly heterozygous species that shall remain unnamed

agresearch
āta mātai, mātai whetū

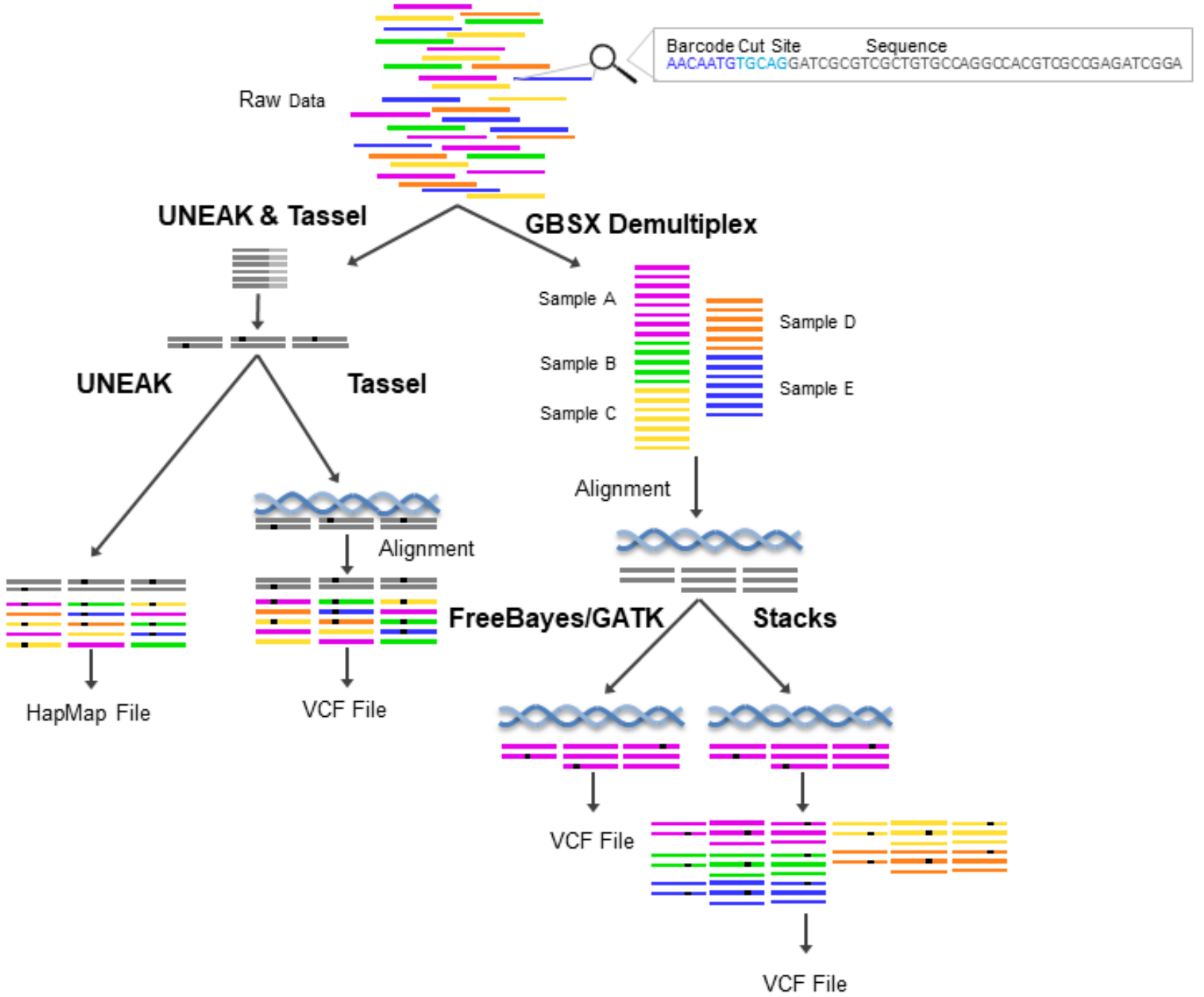
Hold on a minute, what's GBS?

- RRS 1-10% of the genome depending on RE (PstI-MspI)
- Multiplexed samples
- Genotypes
- Like a custom SNP chip, fitting your samples perfectly
- Range of downstream applications (breeding values, parentage test, population structure, …)

agresearch
āta mātai, mātai whetū

Pipelines

- Different pipelines, with/without reference, NR/full depth, full length/64bp, 1 or more SNPs per tag, WGS vs RRS, Bayesian (prior knowledge) or not



agresearch
āta mātai, mātai whetū

Tools, an acquired taste

- Funny requirements like restrictions around chromosome names
- User-friendly?
- Compute expense (time, threads, RAM)
- Conda packages

pipeline	max RAM [GB]	compute time [dd]
bwa	9	10.20
GBSX	1	1.50
freebayes	35	4.00
GATK	14	23.00
Tassel5	49	0.04
Tassel3	10	0.04
Stacks	25	0.80
Samtools	2	2.00
UNEAK	26	0.17

agresearch
āta mātai, mātai whetū

Tools, some surprises…

- ‘unsupported’ GT
- Indels
- N SNPs
- duplicate positions, …

agresearch
āta mātai, mātai whetū

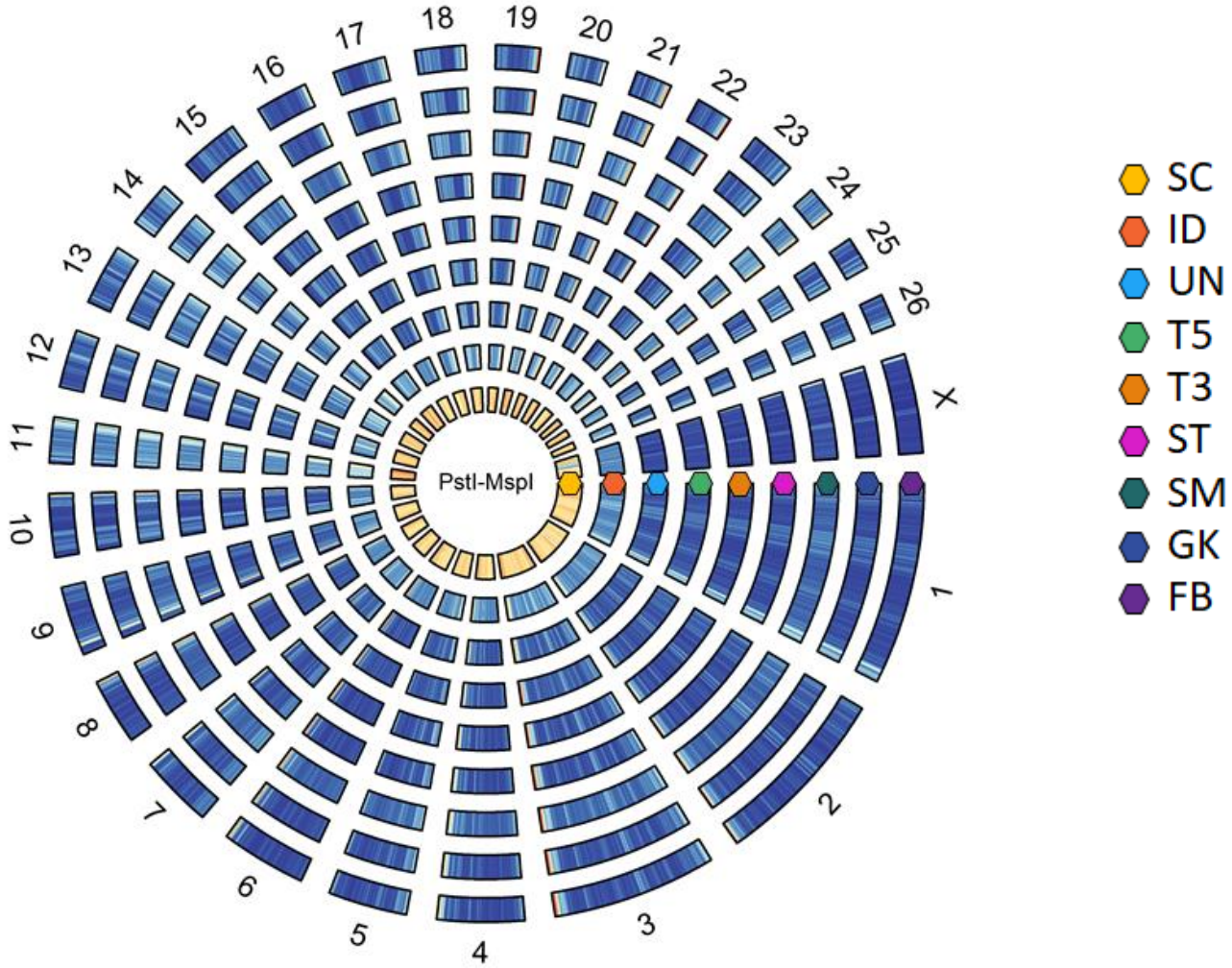
How to compare?

	SNP-calling Pipeline						
Statistics	STACKS	GATK	Freebayes	SM	Tassel3	Tassel5	UNEAK
Number of SNPs	210 044	338 692	188 131	461 506	264 136	210 779	111 053
Call rate	0.58	0.69	0.89	0.58	0.91	0.80	0.69
Mean sample depth	5.18	5.26	10.78	4.46	13.29	8.74	6.30
Mean co-call rate (sample pairs)	0.44	0.59	0.83	0.49	0.86	0.71	0.59
Mean Inbreeding	0.26	0.31	0.10	0.09	0.21	0.12	0.08

- Gold standard (SNP chip) & biology (pedigree)
- End results (GRM)

agresearch
āta mātai, mātai whetū

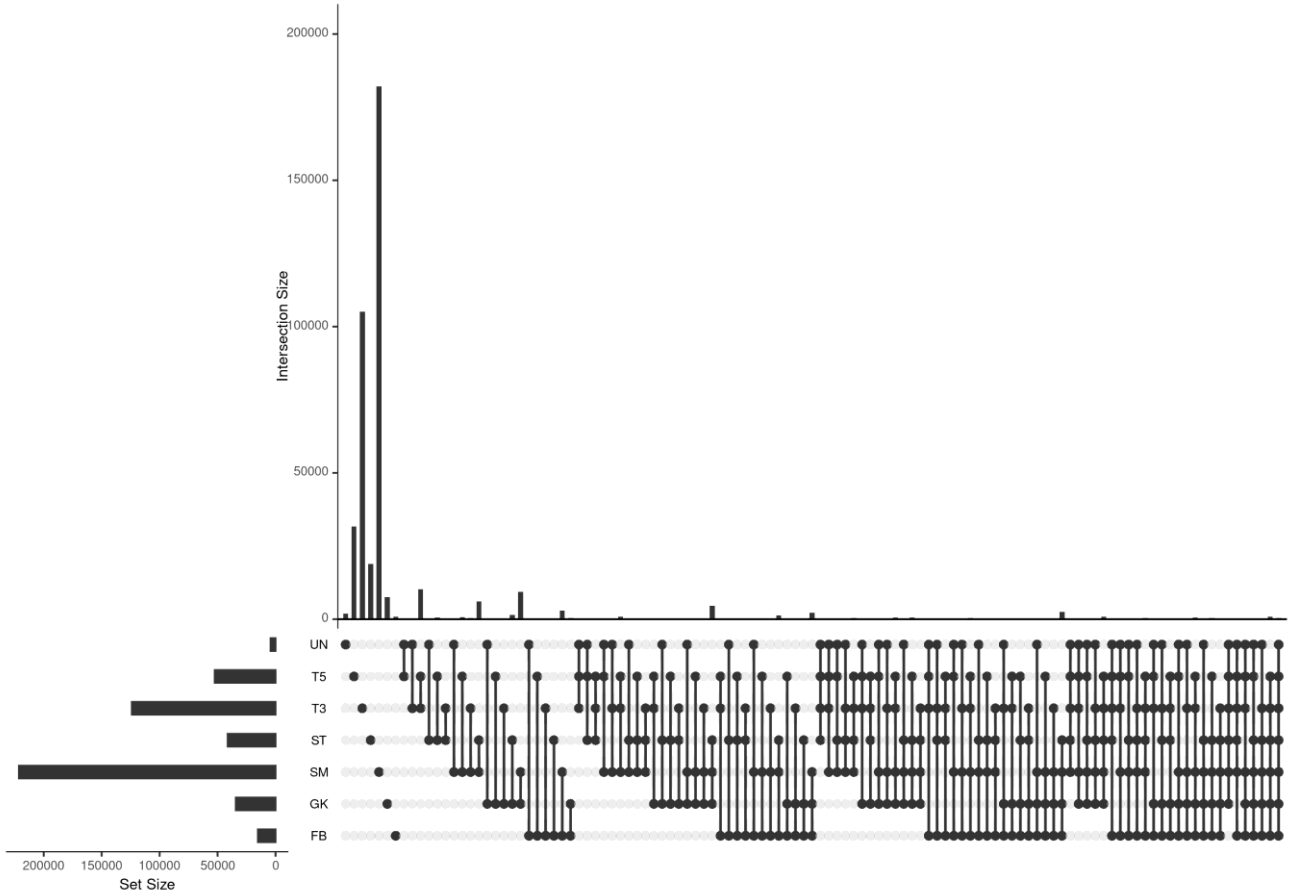
Relative genomic coverage



agresearch
āta mātai, mātai whetū

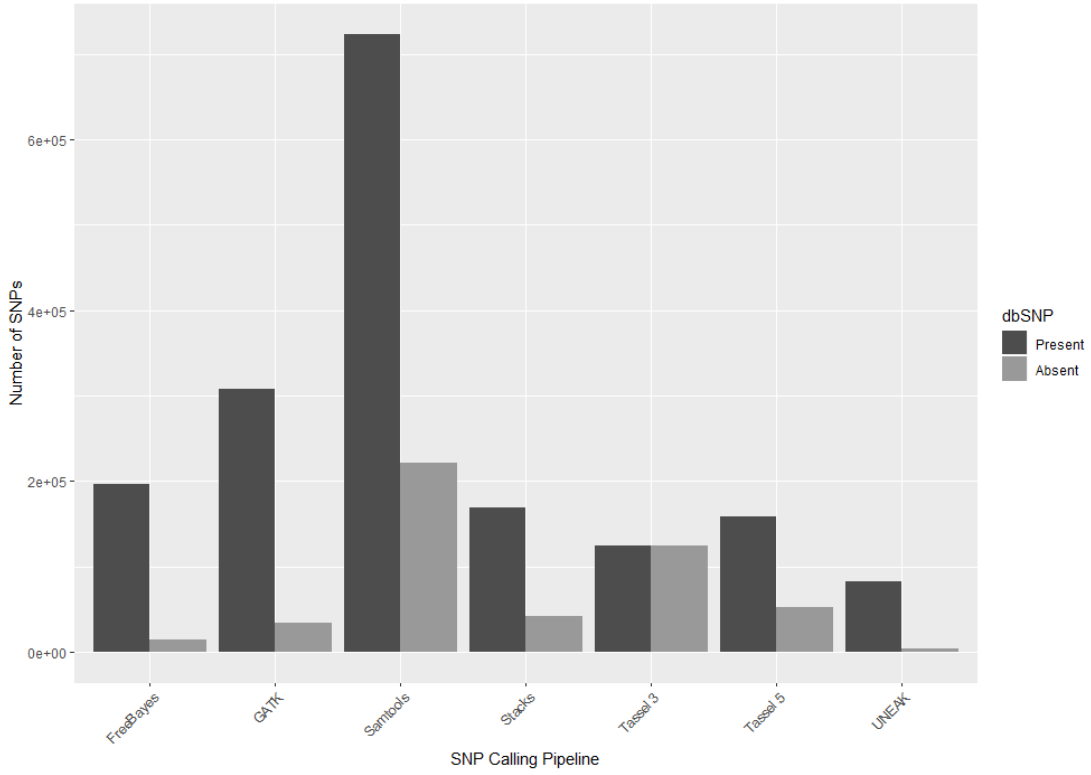
SNP sets

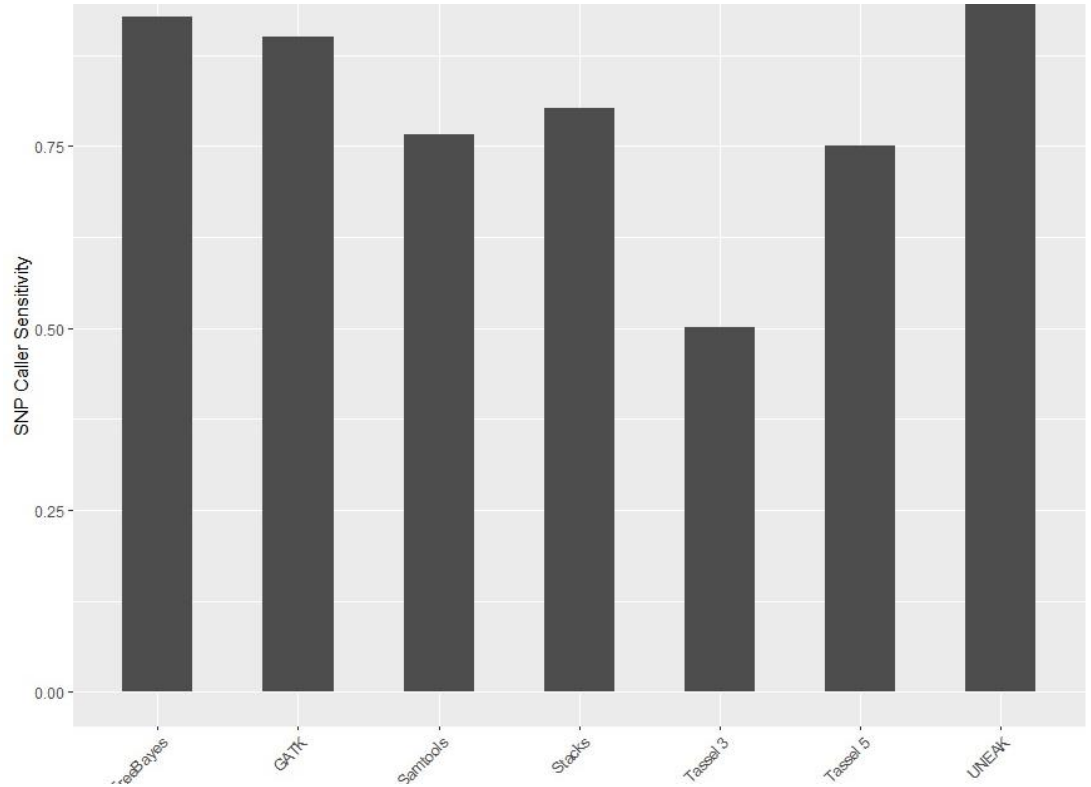
- Upset plots, dbSNP



agresearch
āta mātai, mātai whetū

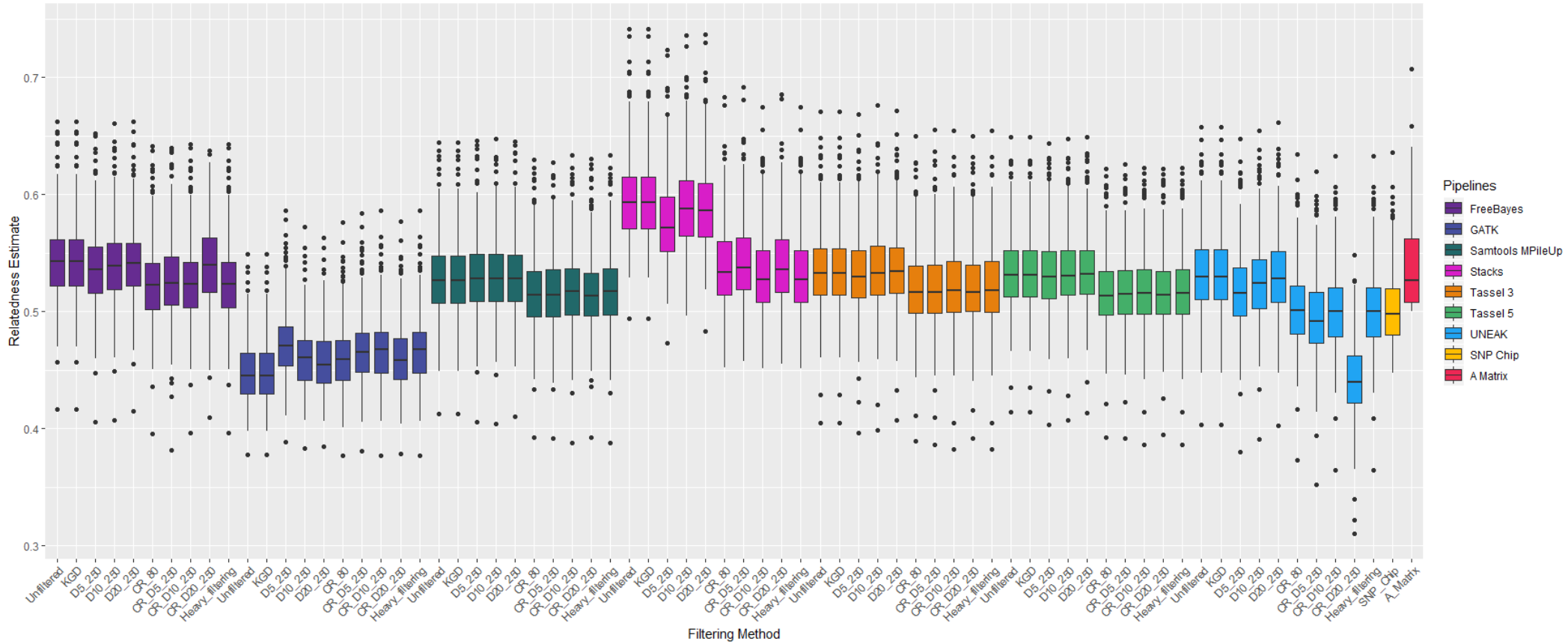
SNP sets vs dbSNP





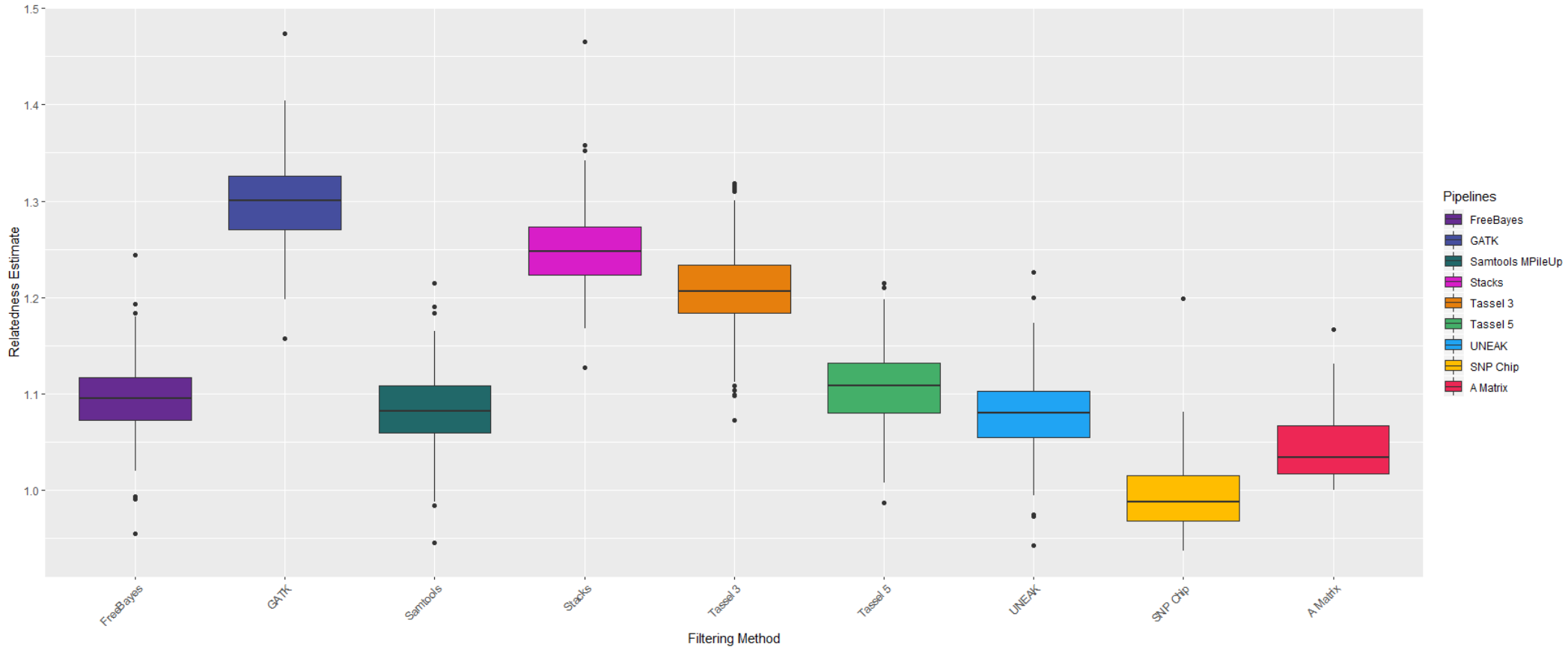
agresearch
āta mātai, mātai whetū

Relatedness, dams, etc.



agresearch
āta mātai, mātai whetū

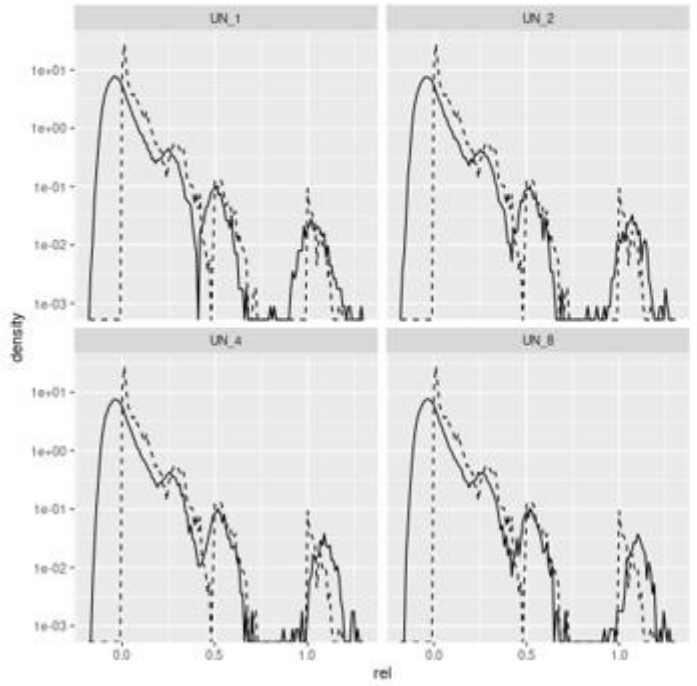
Inbreeding



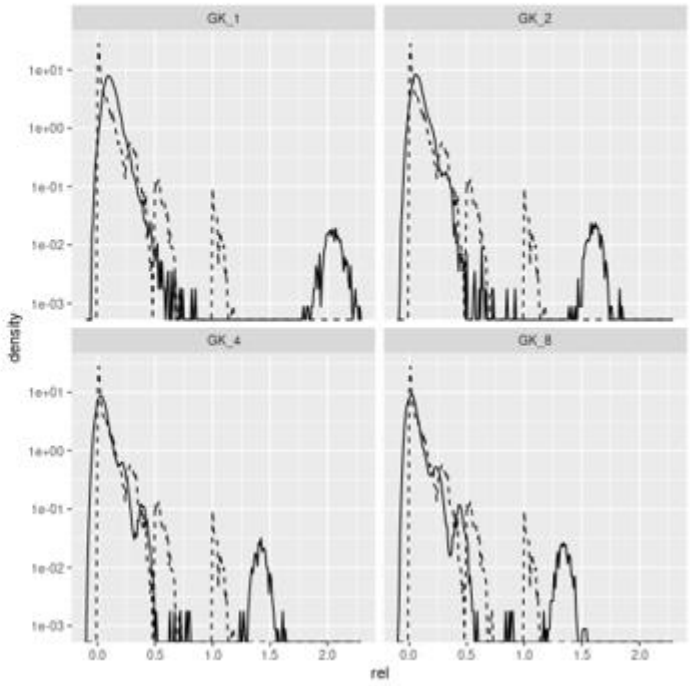
agresearch
āta mātai, mātai whetū

Relatedness matrix, plotted

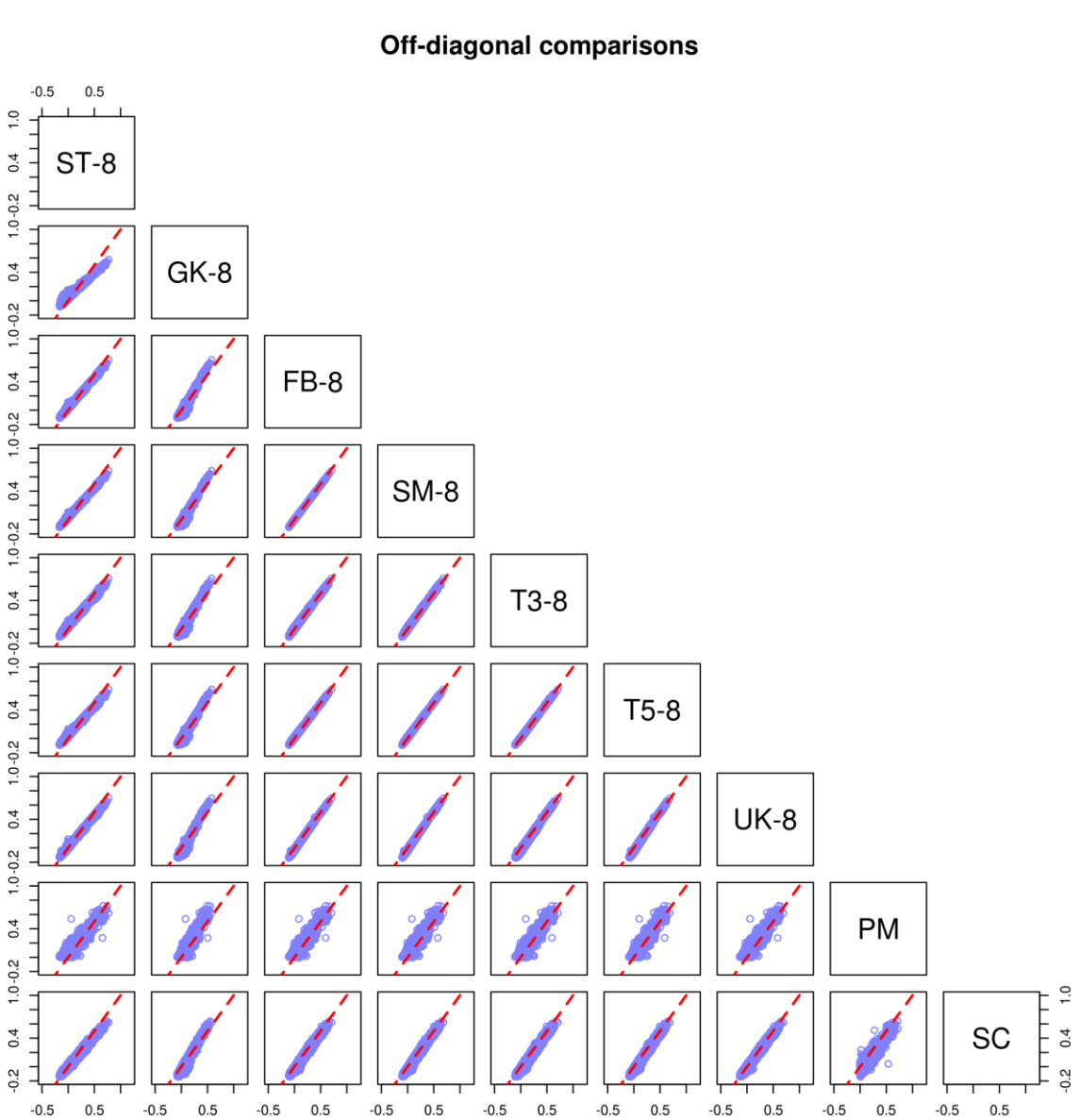
UN

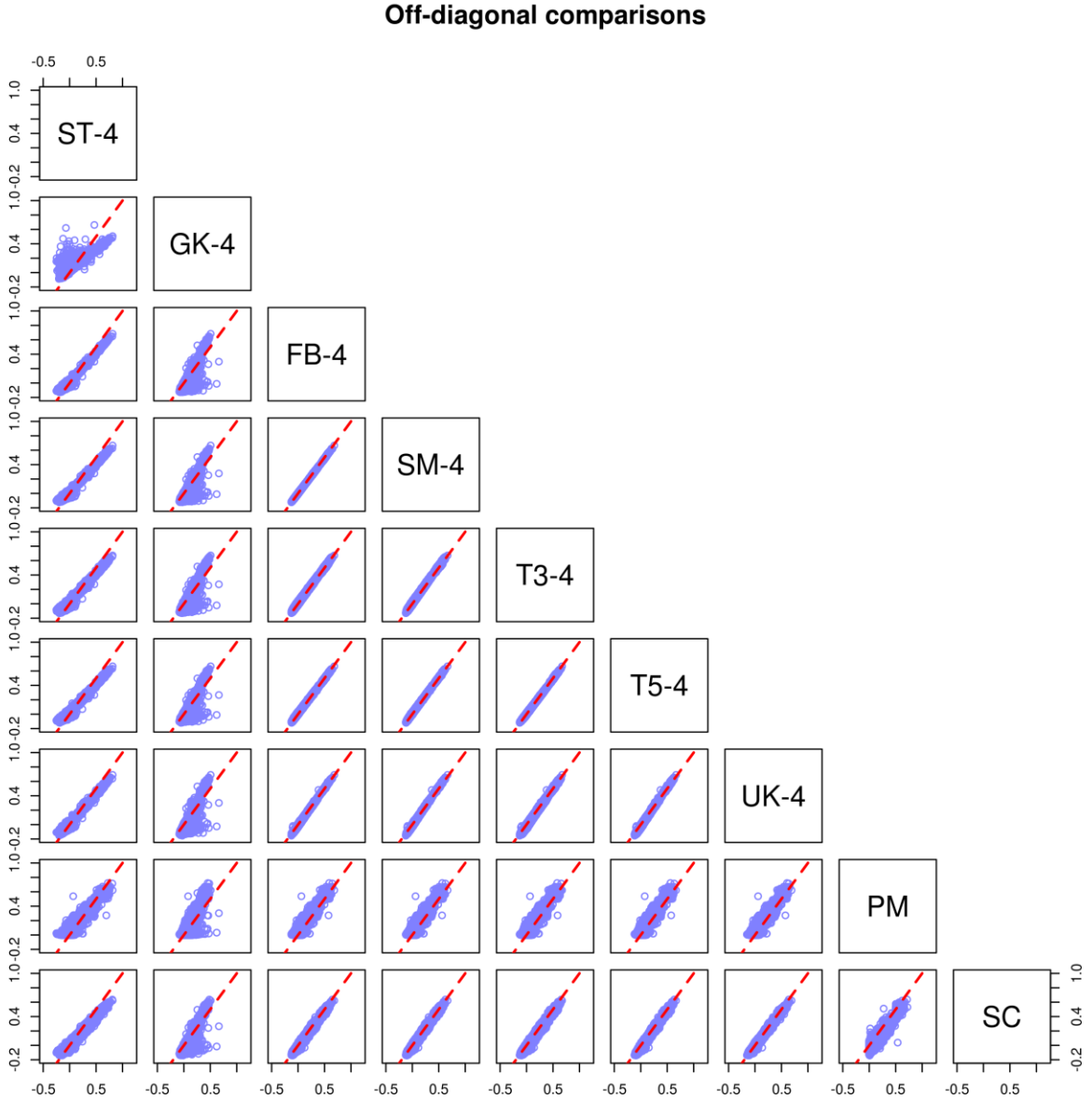


GK

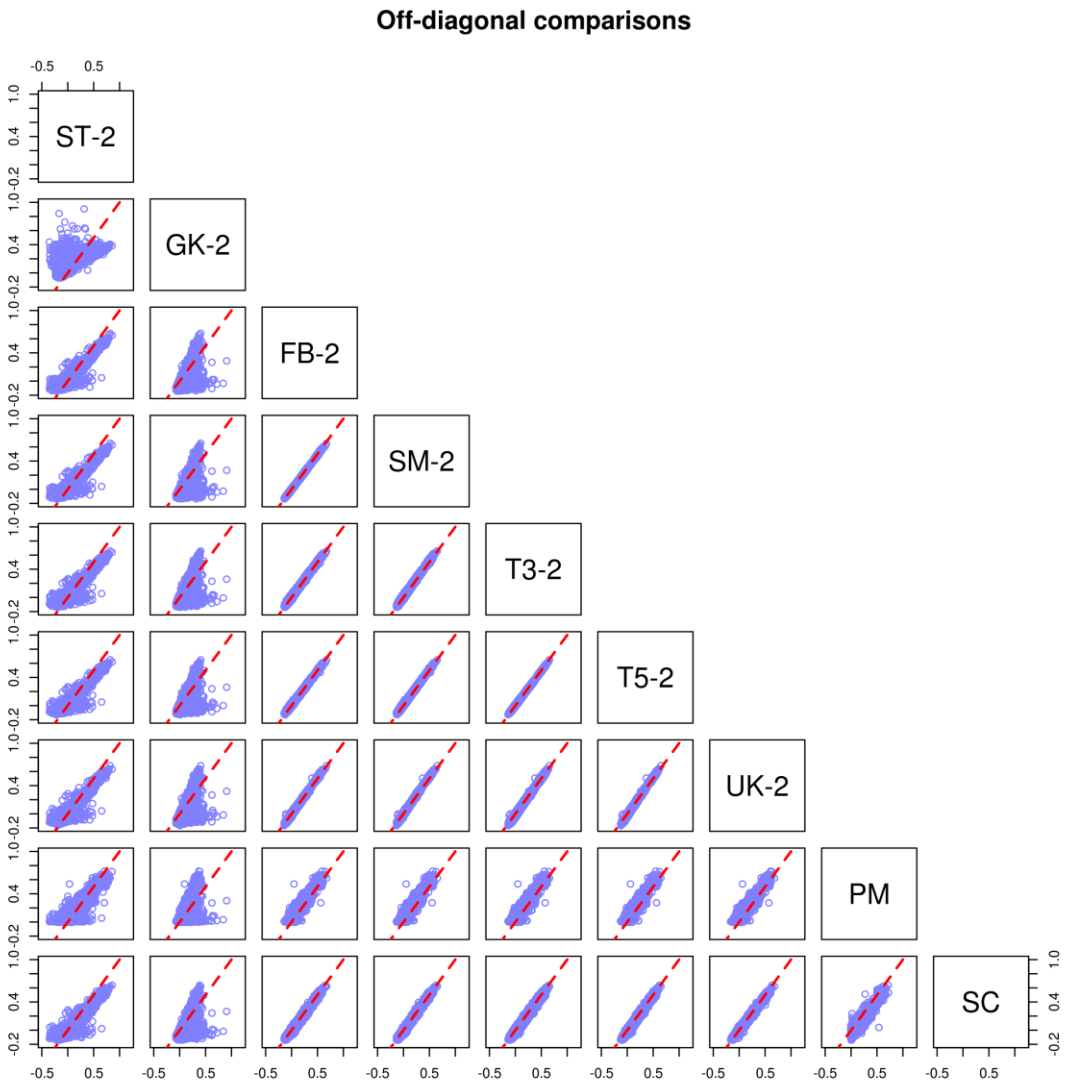


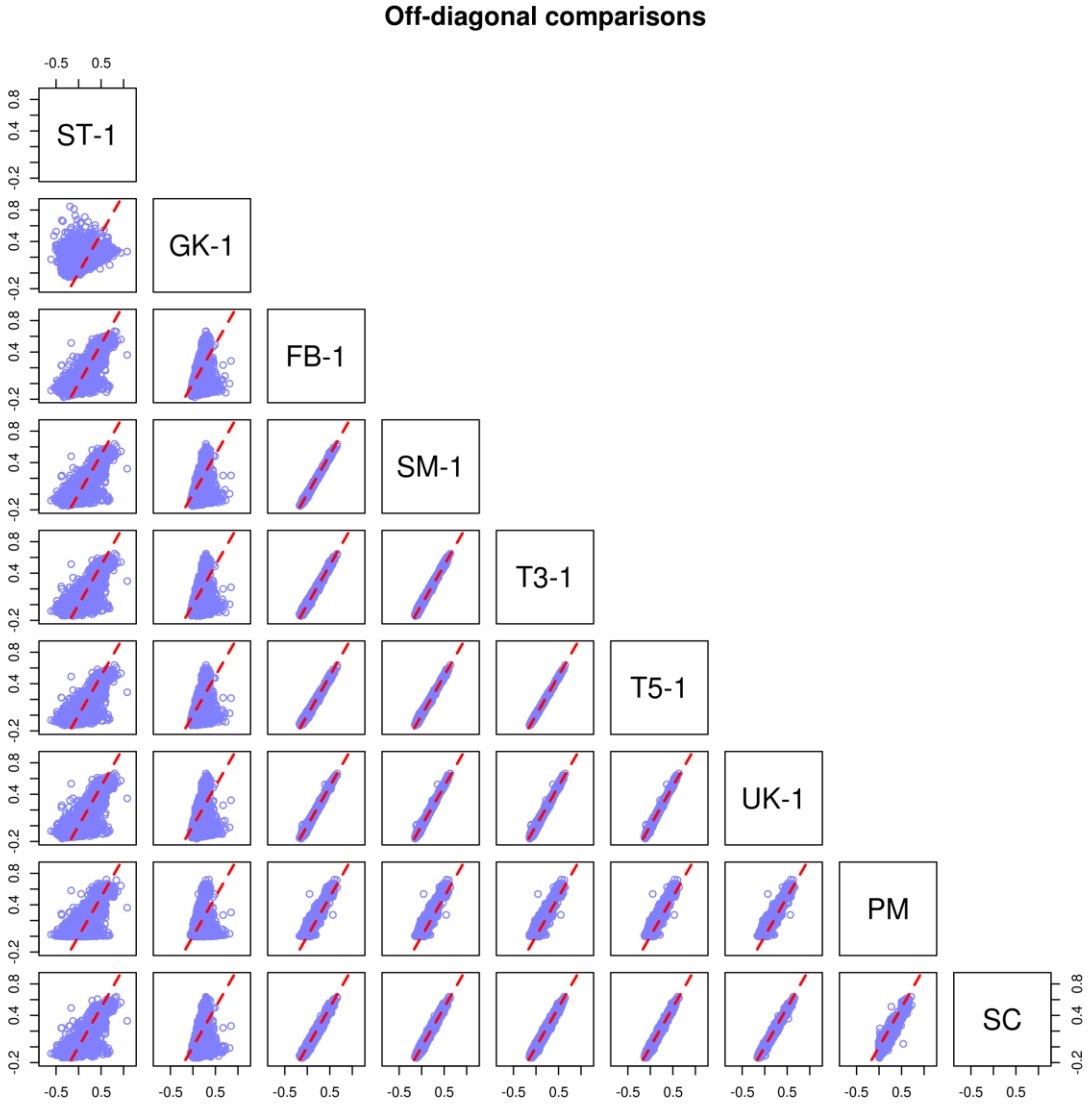
Lowering the depth





Lowering the depth





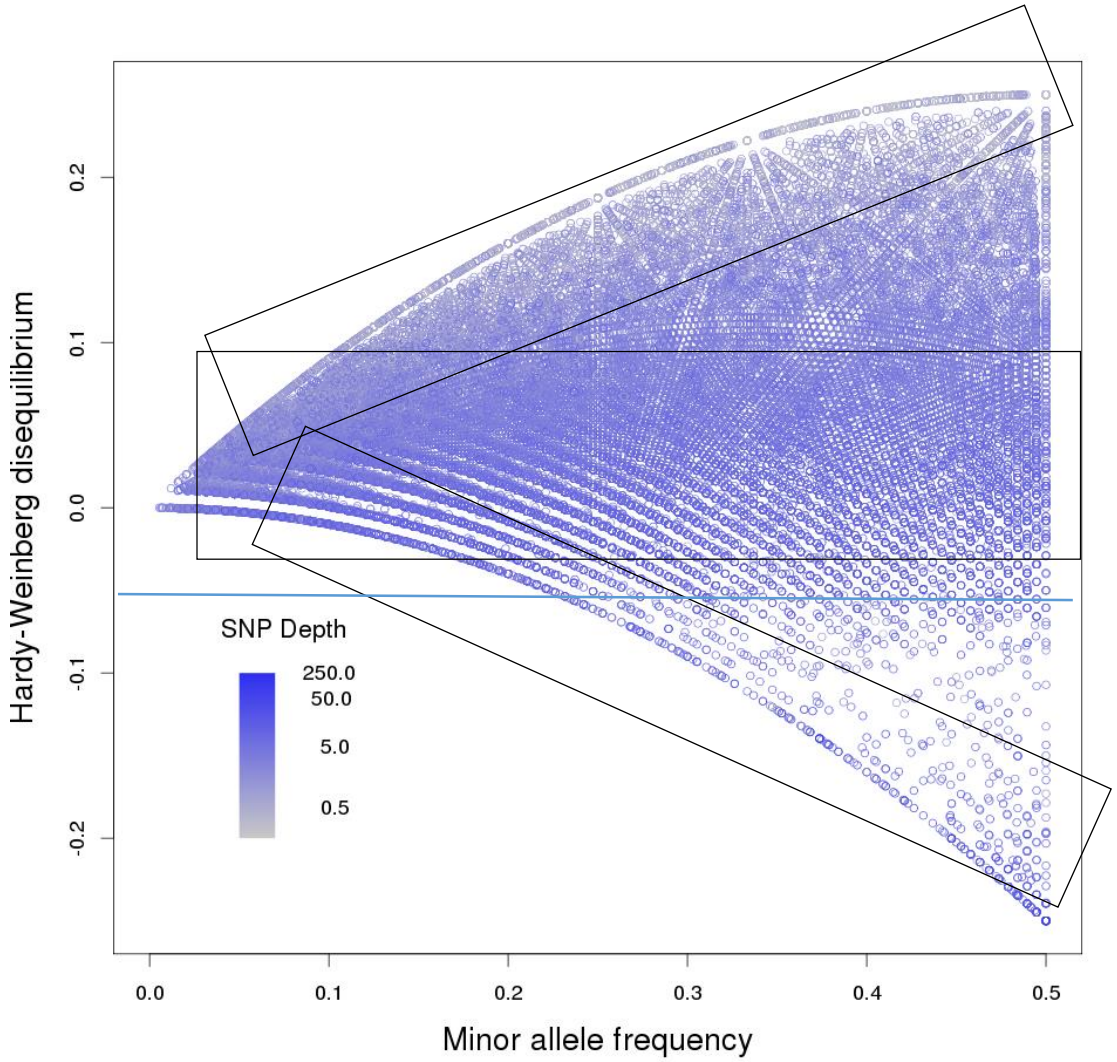
agresearch
āta mātai, mātai whetū

Low depth can work!

- Our tricks:
- problematic absolute GT, therefore use RA to get probabilistic genotypes
- don’t require callrate of .8, instead create relationship matrix stepwise by doing all possible pairwise comparisons (more SNPs can be utilized)
- HW filter to get rid of repetitive sequence
- Finplot

agresearch
āta mātai, mātai whetū

Sheep example



Low depth
appear homozygous

Intermediate depth
in HWE

remove SNPs below line

High depth
maximum heterozygosity
Possibly duplication regions

agresearch
āta mātai, mātai whetū

Lessons learned

- naming convention
- version control
- thoroughly check results before moving on
- gold standard
- biology

- Thanks!
- More on biorxiv soon….