ROSLIN

Use of low-coverage sequence data for genomic selection

John Hickey, Matthew Cleveland, and Gregor Gorjanc

THE UNIVERSITY of EDINBURGH

BBSRC
bioscience for the future

Why sequence?

- Initial results not exciting
  - Data sets WAY TOO SMALL
- Need millions - not thousands
  - Feasible in larger breeding programs
- Sequence will be useful
  - If enough animals sequenced
    - Phenotypes and RECOMBINATION’S
  - Next generation genetic improvement
    - GS2.0, Genome Editing, Biology

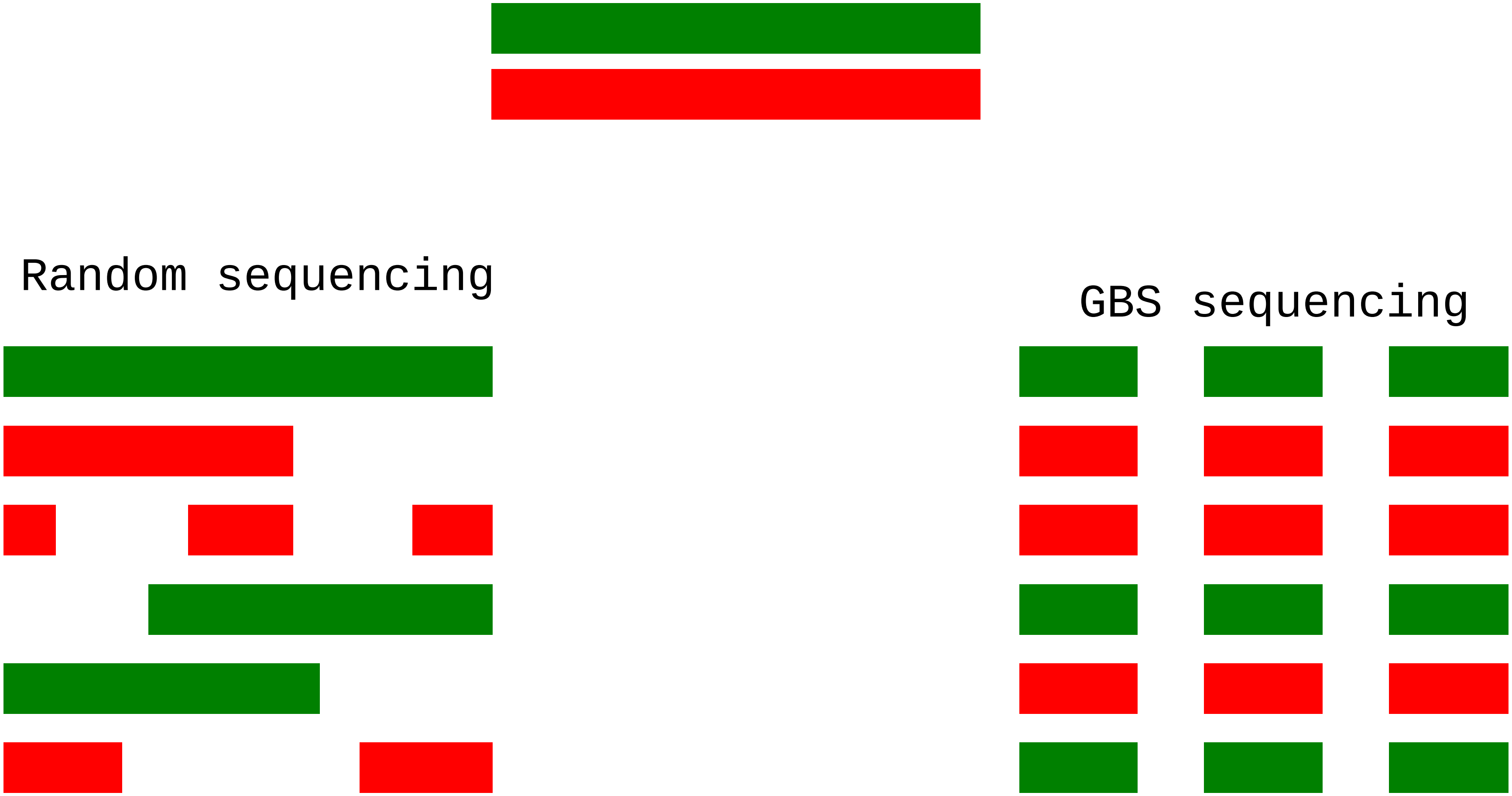
Sequence millions of animals

- Will take time
  - – Approach needs to be competitive with what is currently done but gives long term advantage
  - – Genotype technology needs to be cheap and dense
- High coverage sequence is expensive
- Low coverage sequence is cheaper
  - – Currently lacking infrastructure
    - Imputation and data handling tools, sequencing methods
  - – Is the data competitive currently?

Low-coverage sequence

- Reduced representation of the genome
- Only sequence a portion of the genome with only a few reads
- Uses restriction enzymes and multiplexing
- The portion and number of reads can be controlled by the user
- Higher cost gives higher quality

Low-coverage sequence data



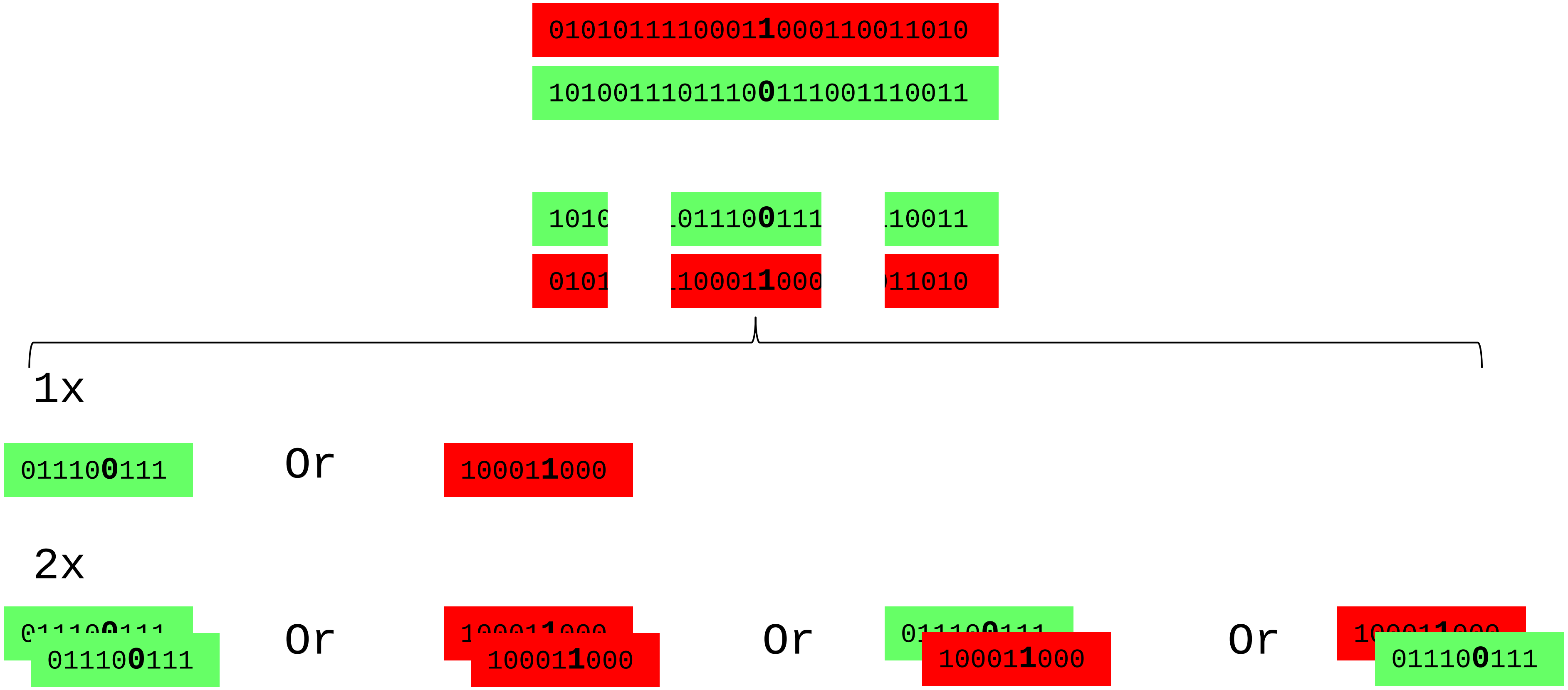
Random sequencing

GBS sequencing

Cutting enzymes

GBS at different *x*

- *x* = the number of reads at a position
  - – Sampled from a Poisson distribution
  - – 1*x* = 1 read, 2*x* = 2 reads, etc.



0101011110001**1**000110011010

1010011101110**0**111001110011

1010 011100**0**111 10011

0101 10001**1**000 11010

1x

01110**0**111 Or 10001**1**000

2x

01110**0**111 01110**0**111 Or 10001**1**000 10001**1**000 Or 01110**0**111 10001**1**000 Or 10001**1**000 01110**0**111

Power of low-coverage

Simulated data

- Coalescent simulator to generate historical events
  - Final generation has Ne of 100
- Drop haplotypes through pedigree
  - 2 generations
  - 500, 1000, 5000, or 10000 animals per generation
- 4 marker densities/enzymes
  - 3k, 10k, 60k, 300k
- Sample GBS from Poisson
- Trait
  - $h^2 = 0.35$
  - 10,000 QTL additive effects from normal distribution

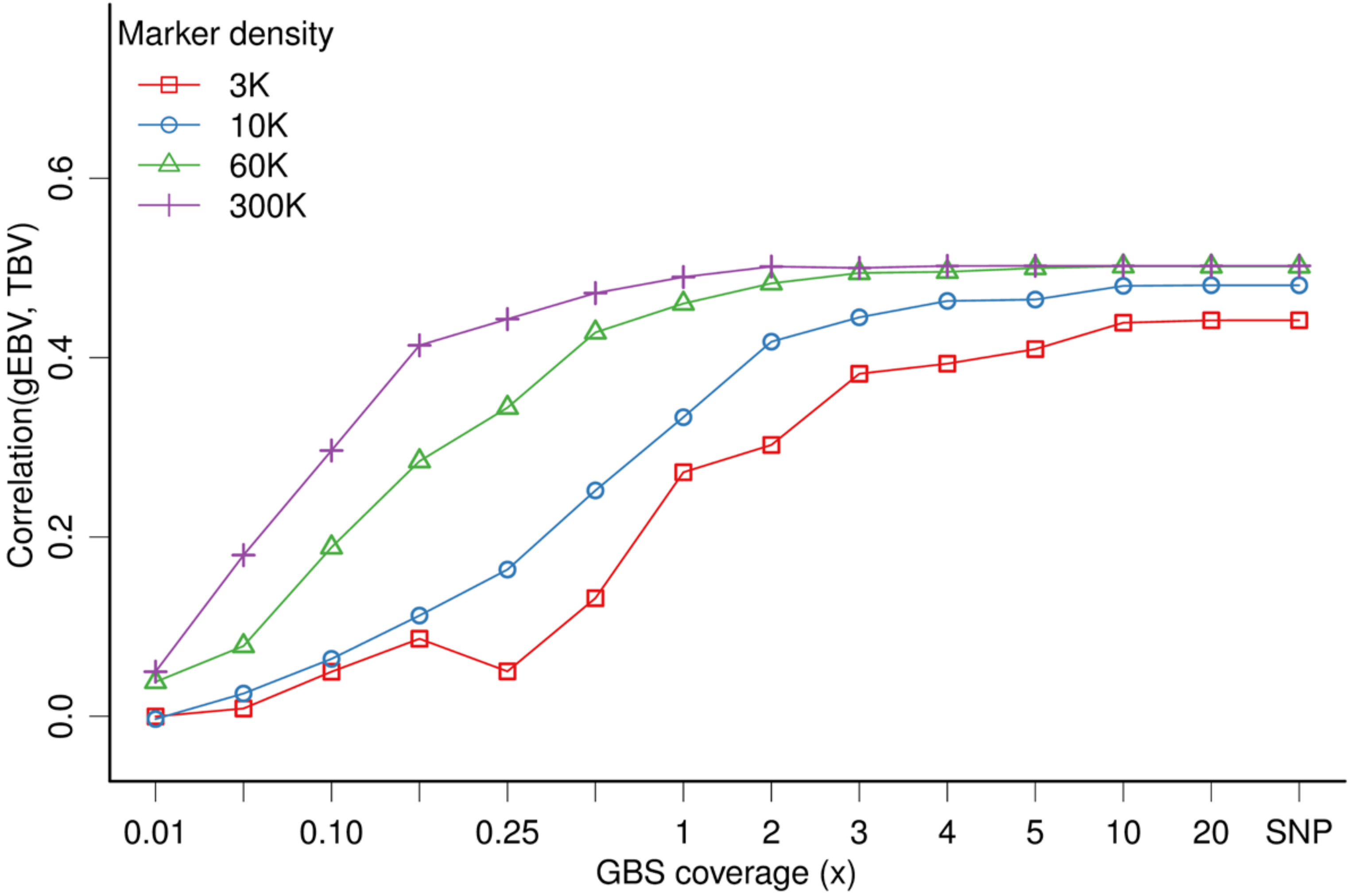
Simulated data

- Train in generation 1
- Predict in generation 2
- Very close relationships
- Ridge regression
- No imputation

Simulated data

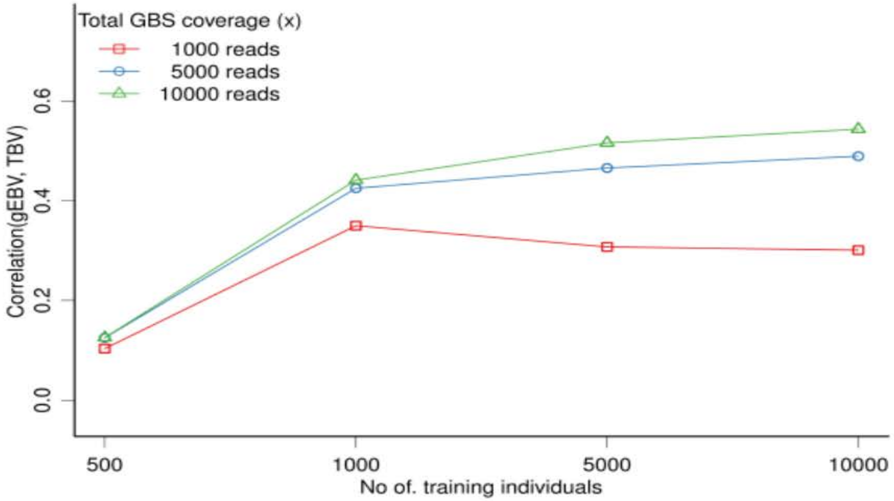
- Many questions could be asked
  - Power of GBS for genomic selection
    - Different densities/enzymes, different *x*/multiplex
  - Effect of using GBS in training population
  - Effect of using GBS in prediction population

GBS – same coverage T+P

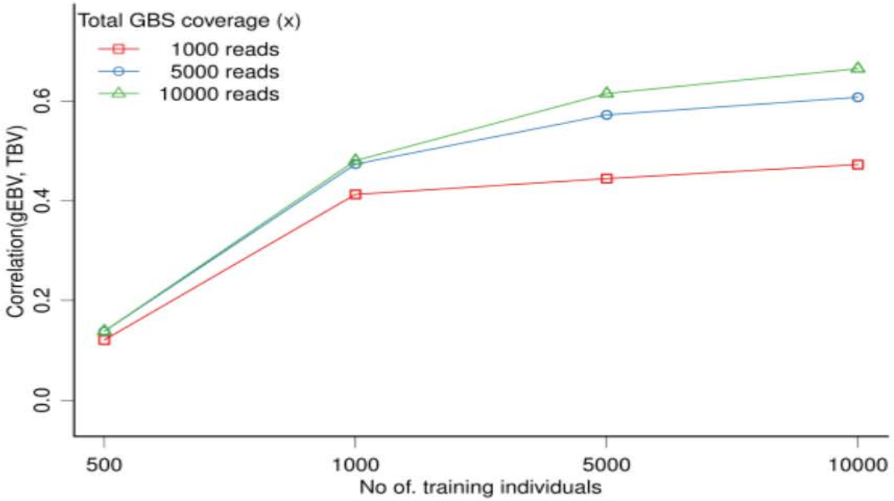


GBS – Total reads in training

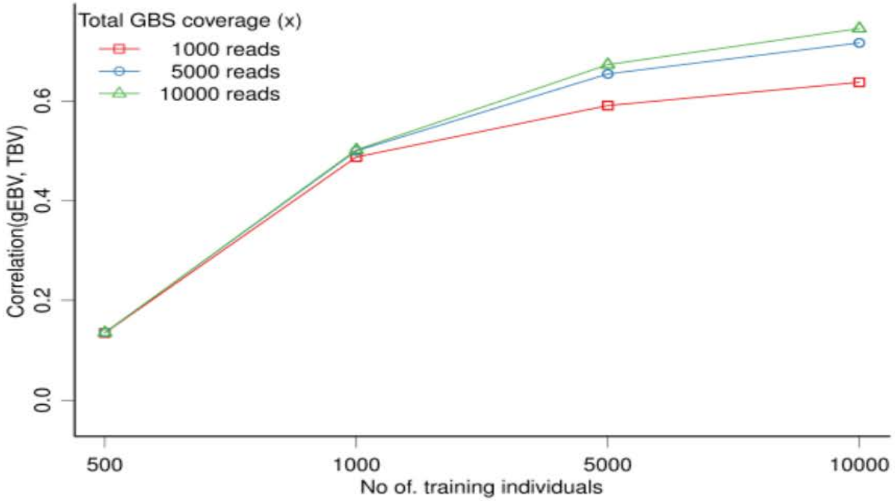
3k



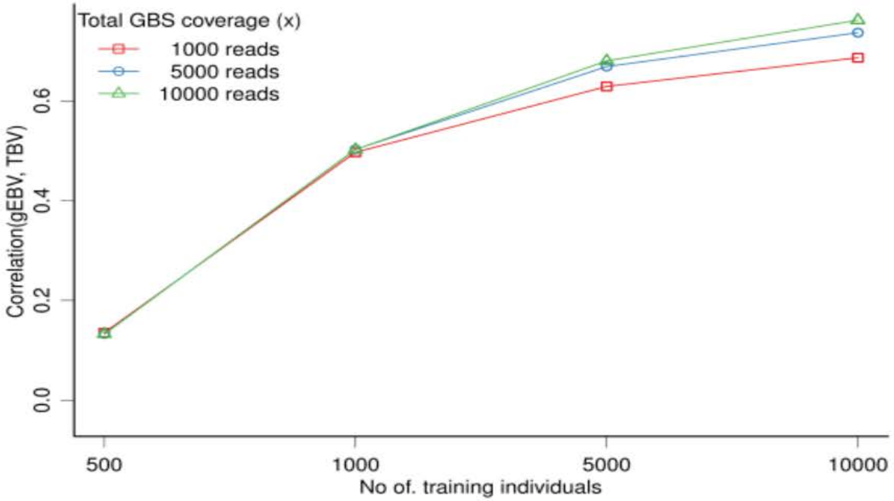
10k



60k

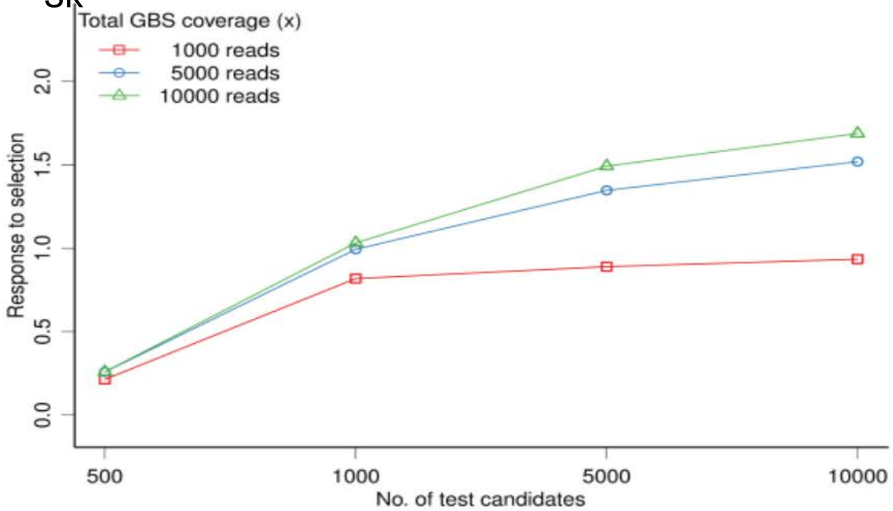


300k

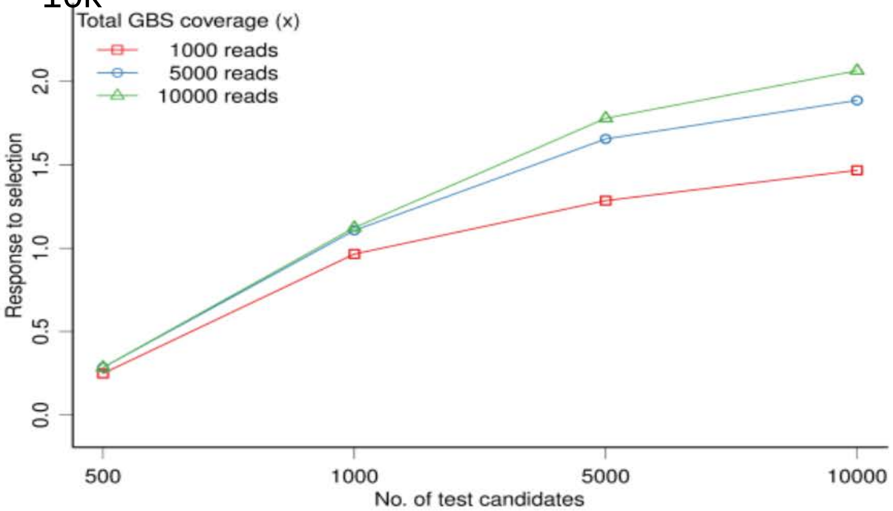


GBS – Total reads in prediction

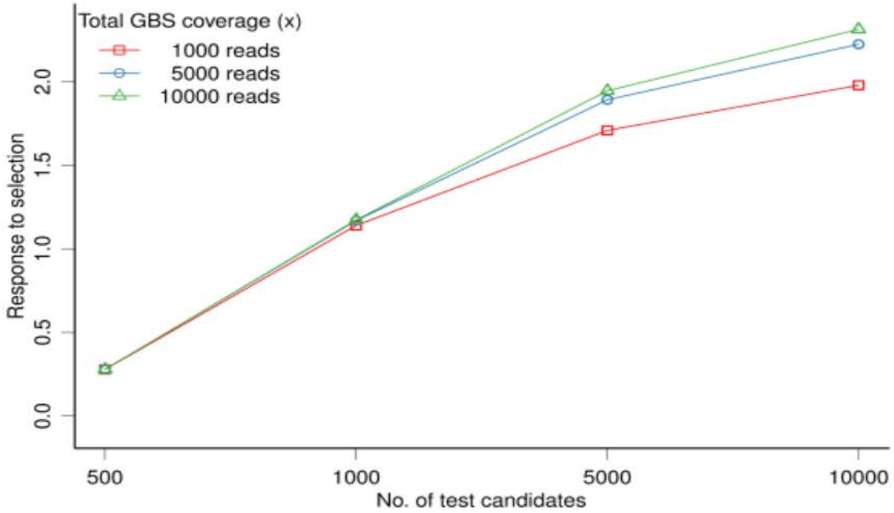
3k



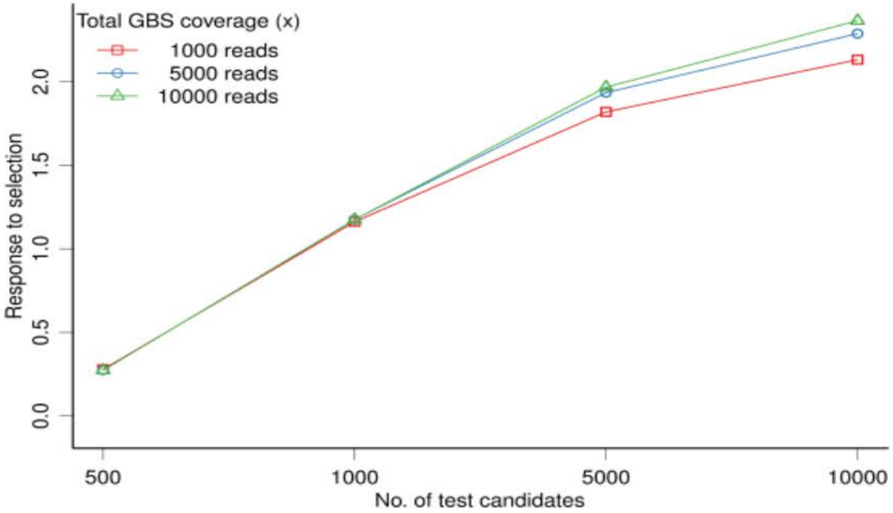
10k



60k



300k



Conclusions

- GBS is competitive in the short term
  - Large training sets with poor quality genotyping are better
  - Large numbers of selection candidates with poor quality genotyping are better
- With imputation things will be better
- In the longer term low-coverage sequence data can be used to generate massive data sets