imputation and genetic evaluation with sequence data

…and a bit of AI for genomic prediction

Cedric Gondro
Michigan State University
gondroce@msu.edu

NGS genomic prediction
where we seem to be headed

sequence key individuals

impute lower density panels to sequence level

impute low pass data to sequence level

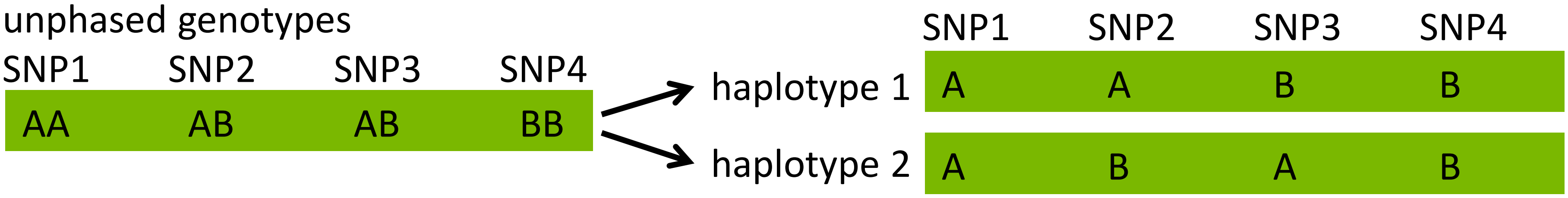
genomic selection at sequence level

computationally intensive

Swine Imputation (SWIM) Server

phasing and imputation 101

phasing – resolve haplotypes



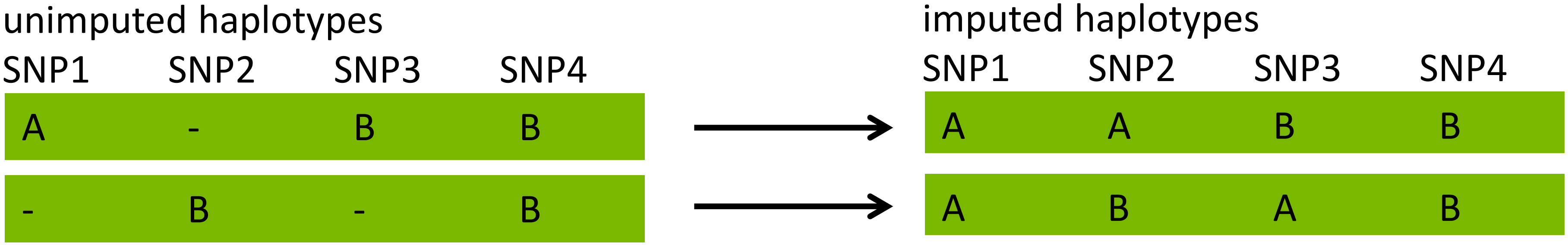
unphased genotypes

SNP1	SNP2	SNP3	SNP4
AA	AB	AB	BB

phased genotypes

	SNP1	SNP2	SNP3	SNP4
haplotype 1	A	A	B	B
haplotype 2	A	B	A	B

imputation – fill in the blanks



unimputed haplotypes

SNP1	SNP2	SNP3	SNP4
A	-	B	B
-	B	-	B

imputed haplotypes

SNP1	SNP2	SNP3	SNP4
A	A	B	B
A	B	A	B

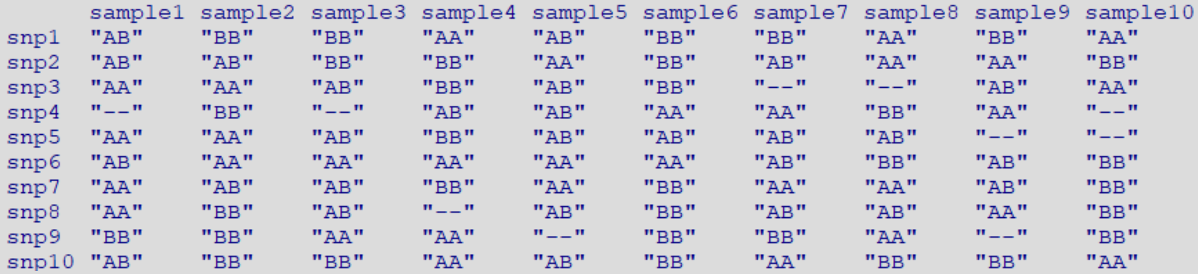
uses

impute randomly missing genotypes

impute genotypes to match different SNP arrays

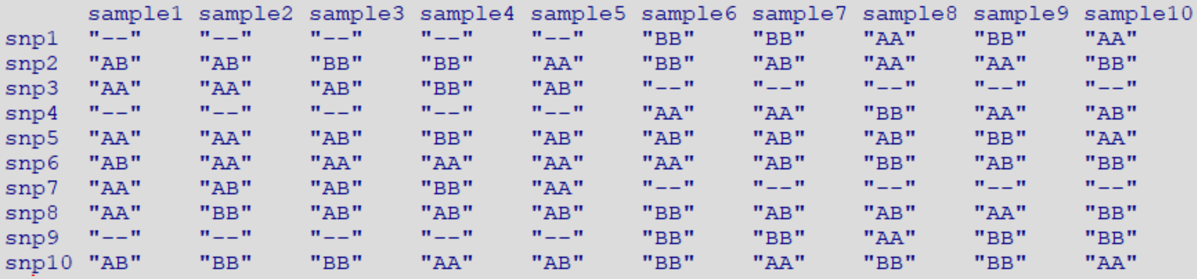
impute genotypes from low-density SNP array to high(er) density SNP array

impute genotypic data from low pass sequencing impute randomly missing genotypes



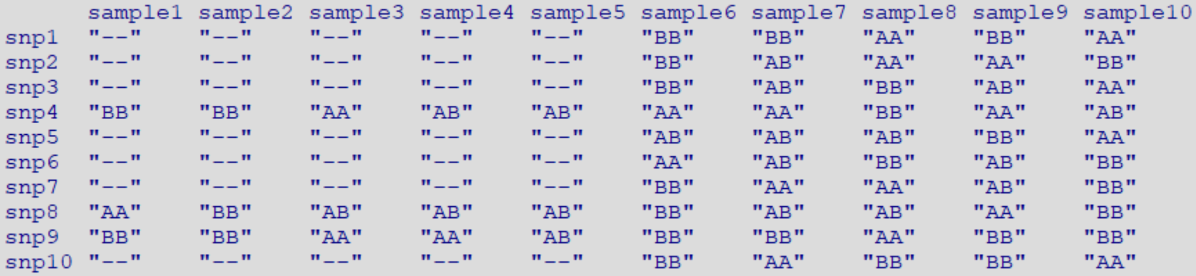
```
      sample1 sample2 sample3 sample4 sample5 sample6 sample7 sample8 sample9 sample10
snp1  "AB"    "BB"    "BB"    "AA"    "AB"    "BB"    "BB"    "AA"    "BB"    "AA"
snp2  "AB"    "AB"    "BB"    "BB"    "AA"    "BB"    "AB"    "AA"    "AA"    "BB"
snp3  "AA"    "AA"    "AB"    "BB"    "AB"    "BB"    "--"    "--"    "AB"    "AA"
snp4  "--"    "BB"    "--"    "AB"    "AB"    "AA"    "AA"    "BB"    "AA"    "--"
snp5  "AA"    "AA"    "AB"    "BB"    "AB"    "AB"    "AB"    "AB"    "--"    "--"
snp6  "AB"    "AA"    "AA"    "AA"    "AA"    "AA"    "AB"    "BB"    "AB"    "BB"
snp7  "AA"    "AB"    "AB"    "BB"    "AA"    "BB"    "AA"    "AA"    "AB"    "BB"
snp8  "AA"    "BB"    "AB"    "--"    "AB"    "BB"    "AB"    "AB"    "AA"    "BB"
snp9  "BB"    "BB"    "AA"    "AA"    "--"    "BB"    "BB"    "AA"    "--"    "BB"
snp10 "AB"    "BB"    "BB"    "AA"    "AB"    "BB"    "AA"    "BB"    "BB"    "AA"
```

impute genotypes to match different SNP arrays



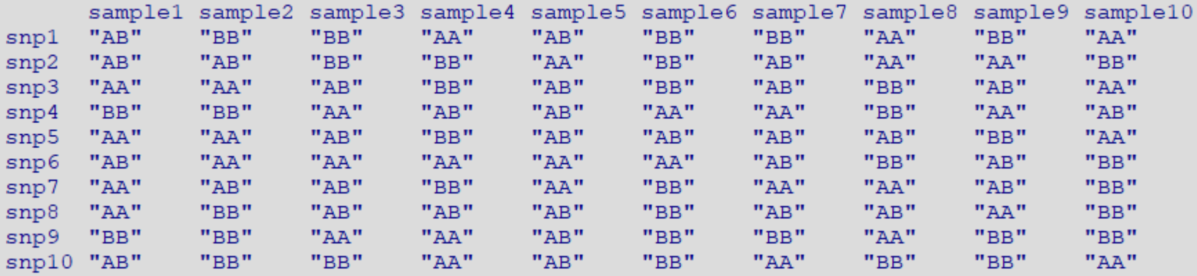
```
      sample1 sample2 sample3 sample4 sample5 sample6 sample7 sample8 sample9 sample10
snp1  "--"    "--"    "--"    "--"    "--"    "BB"    "BB"    "AA"    "BB"    "AA"
snp2  "AB"    "AB"    "BB"    "BB"    "AA"    "BB"    "AB"    "AA"    "AA"    "BB"
snp3  "AA"    "AA"    "AB"    "BB"    "AB"    "--"    "--"    "--"    "--"    "--"
snp4  "--"    "--"    "--"    "--"    "--"    "AA"    "AA"    "BB"    "AA"    "AB"
snp5  "AA"    "AA"    "AB"    "BB"    "AB"    "AB"    "AB"    "AB"    "BB"    "AA"
snp6  "AB"    "AA"    "AA"    "AA"    "AA"    "AA"    "AB"    "BB"    "AB"    "BB"
snp7  "AA"    "AB"    "AB"    "BB"    "AA"    "--"    "--"    "--"    "--"    "--"
snp8  "AA"    "BB"    "AB"    "AB"    "AB"    "BB"    "AB"    "AB"    "AA"    "BB"
snp9  "--"    "--"    "--"    "--"    "--"    "BB"    "BB"    "AA"    "BB"    "BB"
snp10 "AB"    "BB"    "BB"    "AA"    "AB"    "BB"    "AA"    "BB"    "BB"    "AA"
```

impute genotypes from low-density SNP array to high(er) density SNP array



```
      sample1 sample2 sample3 sample4 sample5 sample6 sample7 sample8 sample9 sample10
snp1  "--"    "--"    "--"    "--"    "--"    "BB"    "BB"    "AA"    "BB"    "AA"
snp2  "--"    "--"    "--"    "--"    "--"    "BB"    "AB"    "AA"    "AA"    "BB"
snp3  "--"    "--"    "--"    "--"    "--"    "BB"    "AB"    "BB"    "AB"    "AA"
snp4  "BB"    "BB"    "AA"    "AB"    "AB"    "AA"    "AA"    "BB"    "AA"    "AB"
snp5  "--"    "--"    "--"    "--"    "--"    "AB"    "AB"    "AB"    "BB"    "AA"
snp6  "--"    "--"    "--"    "--"    "--"    "AA"    "AB"    "BB"    "AB"    "BB"
snp7  "--"    "--"    "--"    "--"    "--"    "BB"    "AA"    "AA"    "AB"    "BB"
snp8  "AA"    "BB"    "AB"    "AB"    "AB"    "BB"    "AB"    "AB"    "AA"    "BB"
snp9  "BB"    "BB"    "AA"    "AA"    "AB"    "BB"    "BB"    "AA"    "BB"    "BB"
snp10 "--"    "--"    "--"    "--"    "--"    "BB"    "AA"    "BB"    "BB"    "AA"
```

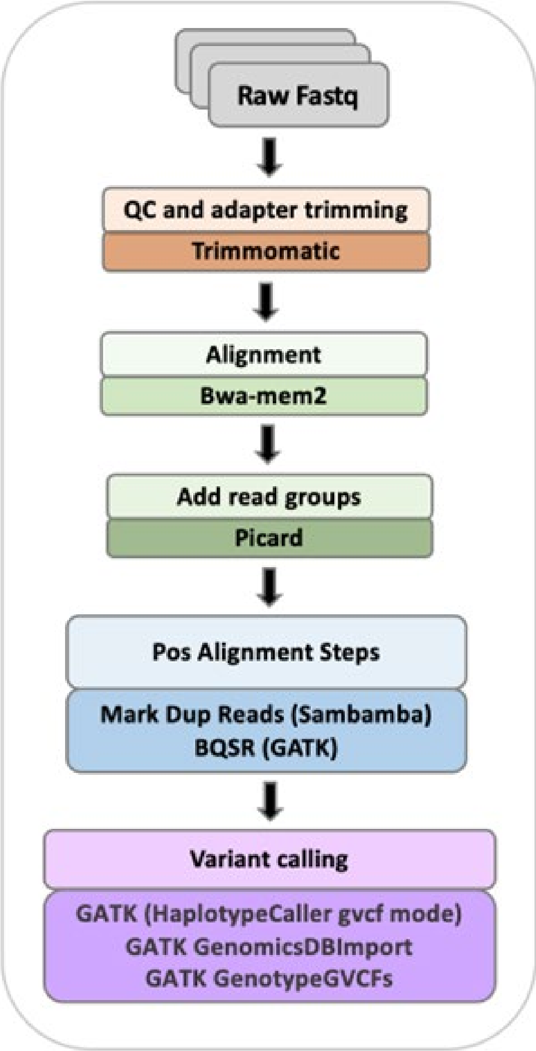
Imputed genotypic data



```
      sample1 sample2 sample3 sample4 sample5 sample6 sample7 sample8 sample9 sample10
snp1  "AB"    "BB"    "BB"    "AA"    "AB"    "BB"    "BB"    "AA"    "BB"    "AA"
snp2  "AB"    "AB"    "BB"    "BB"    "AA"    "BB"    "AB"    "AA"    "AA"    "BB"
snp3  "AA"    "AA"    "AB"    "BB"    "AB"    "BB"    "AB"    "BB"    "AB"    "AA"
snp4  "BB"    "BB"    "AA"    "AB"    "AB"    "AA"    "AA"    "BB"    "AA"    "AB"
snp5  "AA"    "AA"    "AB"    "BB"    "AB"    "AB"    "AB"    "AB"    "BB"    "AA"
snp6  "AB"    "AA"    "AA"    "AA"    "AA"    "AA"    "AB"    "BB"    "AB"    "BB"
snp7  "AA"    "AB"    "AB"    "BB"    "AA"    "BB"    "AA"    "AA"    "AB"    "BB"
snp8  "AA"    "BB"    "AB"    "AB"    "AB"    "BB"    "AB"    "AB"    "AA"    "BB"
snp9  "BB"    "BB"    "AA"    "AA"    "AB"    "BB"    "BB"    "AA"    "BB"    "BB"
snp10 "AB"    "BB"    "BB"    "AA"    "AB"    "BB"    "AA"    "BB"    "BB"    "AA"
```

30x

2 x 30GB
fastq


Raw Fastq

QC and adapter trimming
Trimmomatic

Alignment
Bwa-mem2

Add read groups
Picard

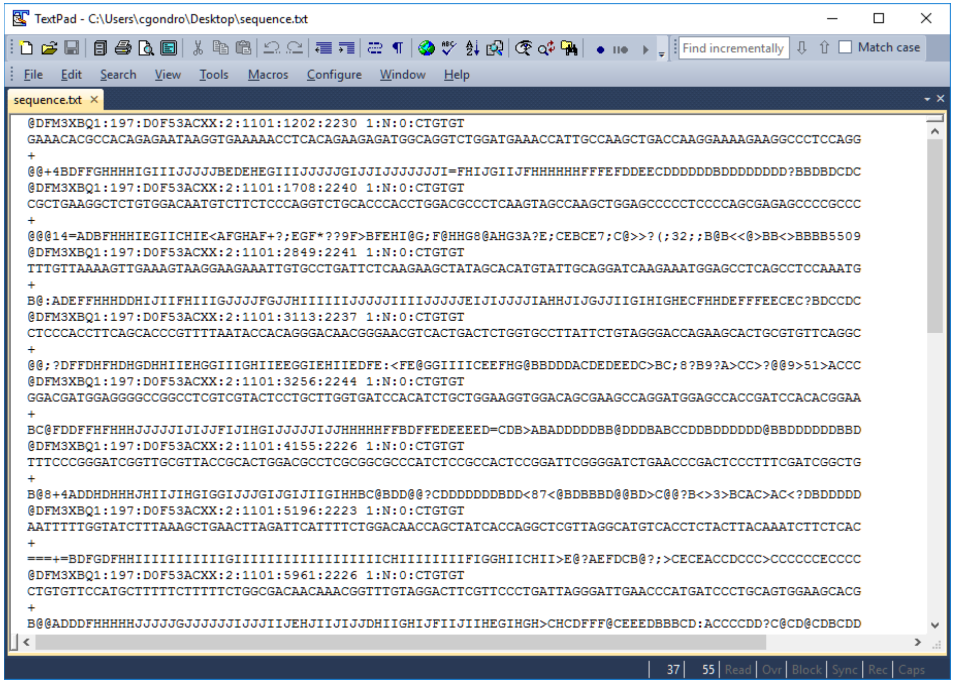
Pos Alignment Steps
Mark Dup Reads (Sambamba)
BQSR (GATK)

Variant calling
GATK (HaplotypeCaller gvcf mode)
GATK GenomicsDBImport
GATK GenotypeGVCFs

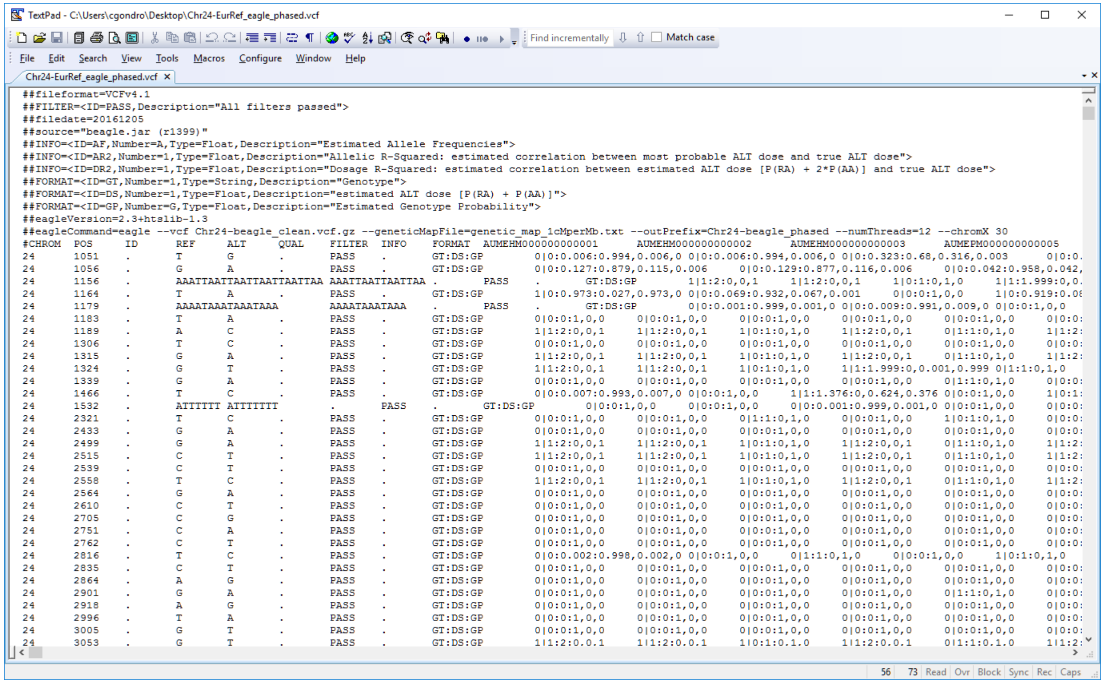
37GB
bam

34GB
gvcf

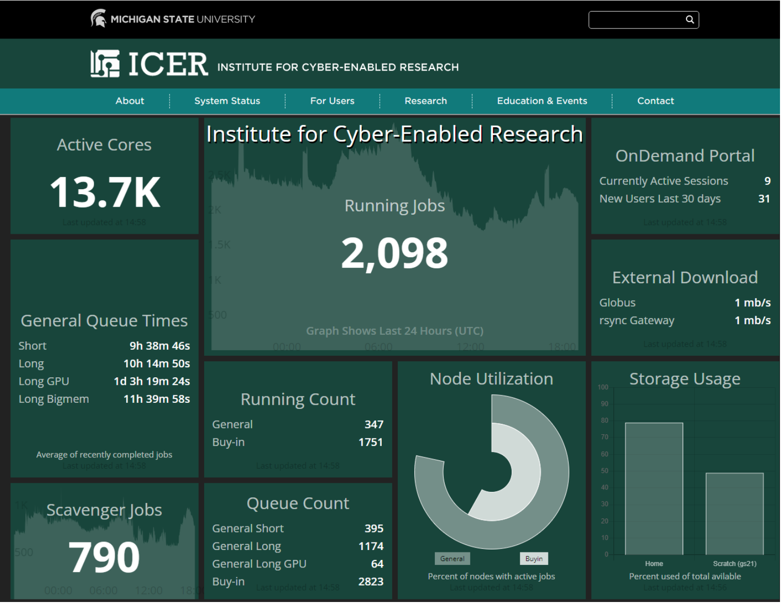
1.5GB vcf – 8 million variants
24MB pgen / 7.5MB bed



```
@DFM3XBQ1:197:D0F53ACXX:2:1101:1202:2230 1:N:0:CTGTGT
GAAACACGCCACAGAGAATAAGGTGAAAAACCTCACAGAAGAGATGGCAGGTCTGGATGAAACCATTGCCAAGCTGACCAAGGAAAAGAAGGCCCTCCAGG
+
@@+4BDFFGHHHHIGIIIJJJJJBEDEHEGIIIJJJJJGIJJIJJJJJJJI=FHIJGIIJFHHHHHHFFFEFDDEECDDDDDDBDDDDDDDD?BBDBDCDC
@DFM3XBQ1:197:D0F53ACXX:2:1101:1708:2240 1:N:0:CTGTGT
CGCTGAAGGCTCTGTGGACAATGTCTTCTCCCAGGTCTGCACCCACCTGGACGCCCTCAAGTAGCCAAGCTGGAGCCCCCTCCCCAGCGAGAGCCCCGCCC
+
@@@14=ADBFHHHIEGIICHIE<AFGHAF+?;EGF*??9F>BFEHI@G;F@HHG8@AHG3A?E;CEBCE7;C@>>?(;32;;B@B<<@>BB<>BBBB5509
@DFM3XBQ1:197:D0F53ACXX:2:1101:2849:2241 1:N:0:CTGTGT
TTTGTTAAAAGTTGAAAGTAAGGAAGAAATTGTGCCTGATTCTCAAGAAGCTATAGCACATGTATTGCAGGATCAAGAAATGGAGCCTCAGCCTCCAAATG
+
B@:ADEFFHHHDDHIJIIFHIIIGJJJJFGJJHIIIIIIJJJJJIIIIJJJJJEIJIJJJJIAHHJIJGJJIIGIHIGHECFHHDEFFFEECEC?BDCCDC
@DFM3XBQ1:197:D0F53ACXX:2:1101:3113:2237 1:N:0:CTGTGT
CTCCCACCTTCAGCACCCGTTTTAATACCACAGGGACAACGGGAACGTCACTGACTCTGGTGCCTTATTCTGTAGGGACCAGAAGCACTGCGTGTTCAGGC
+
@@;?DFFDHFHDHGDHHIIEHGGIIIGHIIEEGGIEHIIEDFE:<FE@GGIIIICEEFHG@BBDDDACDEDEEDC>BC;8?B9?A>CC>?@@9>51>ACCC
@DFM3XBQ1:197:D0F53ACXX:2:1101:3256:2244 1:N:0:CTGTGT
GGACGATGGAGGGGCCGGCCTCGTCGTACTCCTGCTTGGTGATCCACATCTGCTGGAAGGTGGACAGCGAAGCCAGGATGGAGCCACCGATCCACACGGAA
+
BC@FDDFFHFHHHJJJJJIJIJJFIJIHGIJJJJJIJJHHHHHFFBDFFEDEEEED=CDB>ABADDDDDBB@DDDBABCCDDBDDDDDD@BBDDDDDDBBD
@DFM3XBQ1:197:D0F53ACXX:2:1101:4155:2226 1:N:0:CTGTGT
TTTCCCGGGATCGGTTGCGTTACCGCACTGGACGCCTCGCGGCGCCCATCTCCGCCACTCCGGATTCGGGGATCTGAACCCGACTCCCTTTCGATCGGCTG
+
B@8+4ADDHDHHHJHIIJIHGIGGIJJJGIJGIJIIGIHHBC@BDD@@?CDDDDDDDBDD<87<@BDBBBD@@BD>C@@?B<>3>BCAC>AC<?DBDDDDD
@DFM3XBQ1:197:D0F53ACXX:2:1101:5196:2223 1:N:0:CTGTGT
AATTTTTGGTATCTTTAAAGCTGAACTTAGATTCATTTTCTGGACAACCAGCTATCACCAGGCTCGTTAGGCATGTCACCTCTACTTACAAATCTTCTCAC
+
===+=BDFGDFHHIIIIIIIIIIIGIIIIIIIIIIIIIIIIIICHIIIIIIIIFIGGHIICHII>E@?AEFDCB@?;>CECEACCDCCC>CCCCCECCCC
@DFM3XBQ1:197:D0F53ACXX:2:1101:5961:2226 1:N:0:CTGTGT
CTGTGTTCCATGCTTTTTCTTTTTCTGGCGACAACAAACGGTTTGTAGGACTTCGTTCCCTGATTAGGGATTGAACCCATGATCCCTGCAGTGGAAGCACG
+
B@@ADDDFHHHHHJJJJGJJJJJJIJJJIIJEHJIIJIJJDHIIGHIJFIIJIIHEGIHGH>CHCDFFF@CEEEDBBBCD:ACCCCDD?C@CD@CDBCDD
```



```
##fileformat=VCFv4.1
##FILTER=<ID=PASS,Description="All filters passed">
##filedate=20161205
##source="beagle.jar (r1399)"
##INFO=<ID=AF,Number=A,Type=Float,Description="Estimated Allele Frequencies">
##INFO=<ID=AR2,Number=1,Type=Float,Description="Allelic R-Squared: estimated correlation between most probable ALT dose and true ALT dose">
##INFO=<ID=DR2,Number=1,Type=Float,Description="Dosage R-Squared: estimated correlation between estimated ALT dose [P(RA) + 2*P(AA)] and true ALT dose">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DS,Number=1,Type=Float,Description="estimated ALT dose [P(RA) + P(AA)]">
##FORMAT=<ID=GP,Number=G,Type=Float,Description="Estimated Genotype Probability">
##eagleVersion=2.3+htslib-1.3
##eagleCommand=eagle --vcf Chr24-beagle_clean.vcf.gz --geneticMapFile=genetic_map_1cMperMb.txt --outPrefix=Chr24-beagle_phased --numThreads=12 --chromX 30
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  AUMEHM000000000001      AUMEHM000000000002      AUMEHM000000000003      AUMEPM000000000005
24      1051    .       T       G       .       PASS    .       GT:DS:GP        0|0:0.006:0.994,0.006,0 0|0:0.006:0.994,0.006,0 0|0:0.323:0.68,0.316,0.003      0|0:0.
24      1056    .       G       A       .       PASS    .       GT:DS:GP        0|0:0.127:0.879,0.115,0.006     0|0:0.129:0.877,0.116,0.006     0|0:0.042:0.958,0.042,
24      1156    .       AAATTAATTAATTAATTAATTAA AAATTAATTAATTAA .       PASS    .       GT:DS:GP        1|1:2:0,0,1     1|1:2:0,0,1     1|0:1:0,1,0     1|1:1.999:0,0.
24      1164    .       T       A       .       PASS    .       GT:DS:GP        1|0:0.973:0.027,0.973,0 0|0:0.069:0.932,0.067,0.001     0|0:0:1,0,0     1|0:0.919:0.0
24      1179    .       AAAATAAATAAATAAA        AAAATAAATAAA    .       PASS    .       GT:DS:GP        0|0:0.001:0.999,0.001,0 0|0:0.009:0.991,0.009,0 0|0:0:1,0,0
24      1183    .       T       A       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:
24      1189    .       A       C       .       PASS    .       GT:DS:GP        1|1:2:0,0,1     1|1:2:0,0,1     1|0:1:0,1,0     1|1:2:0,0,1     0|1:1:0,1,0     1|1:2:
24      1306    .       T       C       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:
24      1315    .       G       A       .       PASS    .       GT:DS:GP        1|1:2:0,0,1     1|1:2:0,0,1     1|0:1:0,1,0     1|1:2:0,0,1     0|1:1:0,1,0     1|1:2:
24      1324    .       G       T       .       PASS    .       GT:DS:GP        1|1:2:0,0,1     1|1:2:0,0,1     1|0:1:0,1,0     1|1:1.999:0,0.001,0.999 0|1:1:0,1,0
24      1339    .       G       A       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|1:1:0,1,0     0|0:0:
24      1466    .       T       C       .       PASS    .       GT:DS:GP        0|0:0.007:0.993,0.007,0 0|0:0:1,0,0     1|1:1.376:0,0.624,0.376 0|0:0:1,0,0     1|0:1:
24      1532    .       ATTTTTT ATTTTTTT        .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0.001:0.999,0.001,0 0|0:0:1,0,0     0|0:0:
24      2321    .       T       C       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|1:1:0,1,0     0|0:0:1,0,0     1|0:1:0,1,0     0|0:0:
24      2433    .       G       A       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:
24      2499    .       G       A       .       PASS    .       GT:DS:GP        1|1:2:0,0,1     1|1:2:0,0,1     1|0:1:0,1,0     1|1:2:0,0,1     0|1:1:0,1,0     1|1:2:
24      2515    .       C       T       .       PASS    .       GT:DS:GP        1|1:2:0,0,1     1|1:2:0,0,1     1|0:1:0,1,0     1|1:2:0,0,1     0|1:1:0,1,0     1|1:2:
24      2539    .       C       T       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:
24      2558    .       T       C       .       PASS    .       GT:DS:GP        1|1:2:0,0,1     1|1:2:0,0,1     1|0:1:0,1,0     1|1:2:0,0,1     0|1:1:0,1,0     1|1:2:
24      2564    .       G       A       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:
24      2610    .       C       T       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:
24      2705    .       C       G       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:
24      2751    .       C       A       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:
24      2762    .       C       T       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:
24      2816    .       T       C       .       PASS    .       GT:DS:GP        0|0:0.002:0.998,0.002,0 0|0:0:1,0,0     0|1:1:0,1,0     0|0:0:1,0,0     1|0:1:0,1,0
24      2835    .       C       T       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:
24      2864    .       A       G       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:
24      2901    .       G       A       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|1:1:0,1,0     0|0:0:
24      2918    .       A       G       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:
24      2996    .       T       A       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:
24      3005    .       G       T       .       PASS    .       GT:DS:GP        0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:1,0,0     0|0:0:
24      3053    .       G       T       .       PASS    .       GT:DS:GP        1|1:2:0,0,1     1|1:2:0,0,1     1|0:1:0,1,0     1|1:2:0,0,1     0|1:1:0,1,0     1|1:2:
```



MICHIGAN STATE UNIVERSITY

ICER INSTITUTE FOR CYBER-ENABLED RESEARCH

About | System Status | For Users | Research | Education & Events | Contact

Institute for Cyber-Enabled Research

Active Cores
13.7K
Last updated at 14:58

Running Jobs
2,098
Graph Shows Last 24 Hours (UTC)

OnDemand Portal

Currently Active Sessions	9
New Users Last 30 days	31

Last updated at 14:58

External Download

Globus	1 mb/s
rsync Gateway	1 mb/s

Last updated at 14:58

General Queue Times

Short	9h 38m 46s
Long	10h 14m 50s
Long GPU	1d 3h 19m 24s
Long Bigmem	11h 39m 58s

Average of recently completed jobs
Last updated at 14:58

Running Count

General	347
Buy-in	1751

Last updated at 14:58

Node Utilization
Percent of nodes with active jobs
Last updated at 14:58

Storage Usage
Percent used of total avilable
Last updated at 14:56

Scavenger Jobs
790

Queue Count

General Short	395
General Long	1174
General Long GPU	64
Buy-in	2823

Last updated at 14:58

phasing in *chunks*

split across HPC ~1000 chunks

(in batches of 50k samples)

16 per cores per chunk

160GB RAM per chunk

30 minutes per chunk

=

16000 cores + 160TB RAM

system: 55k cores + 317TB RAM

~21 days on a single machine

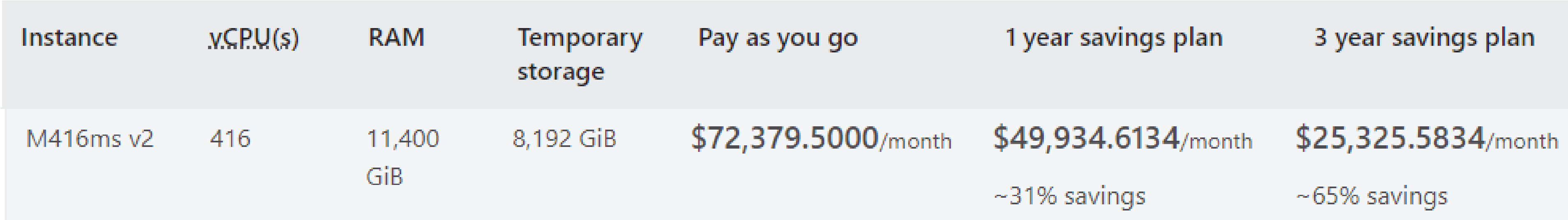
computationally expensive processing and storage

raw data storage

1000 samples (fastq)
2TB @ 1x
60TB @ 30x

memory

30 million variants X 100,000 samples
2-bits – 0.75TB
bytes – 3TB
float – 6TB
double – 12TB



Instance	vCPU(s)	RAM	Temporary storage	Pay as you go	1 year savings plan	3 year savings plan
M416ms v2	416	11,400 GiB	8,192 GiB	$72,379.5000/month	$49,934.6134/month ~31% savings	$25,325.5834/month ~65% savings

real example: 34 million variants and 62,000 thousand samples – 500GB (bed) / 8.5TB (vcf)

considerations

- raw and ready-to-use data storage and what to store
- compute requirements and software
    - parallelization of I/O and processing
    - but still capped by system limits
- smarter programming, approximations (short cuts), dimensionality reduction…

- on the industry side – might only require storing and handling of vcf files, but
    - 70 – 120 million variants across species
    - 10 – 20 million variants within a breed
    - 5 – 10 million after some filtering
    - keep what?
    - how to match data across breeds / organizations?
    - how to revert back – e.g. new assembly?
    - strategy for historical data and seq data – impute up or subset down?

- how good is the imputation?
- how useful is the imputed sequence data?

how good is the imputation?

pattern matching
more patterns -> higher probability of having a match

it's a numbers game

the larger the reference population the better

it's a relationship game

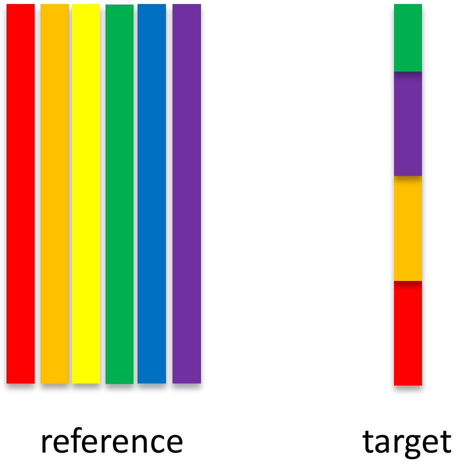
the more connected the reference and target are the better

it's an allele frequency game

the more common an allele is the better

it's a density game

the higher the marker coverage of the target is the better



data

9732 @ 50k
991 @ 700k
224 @ seq

6292 seq from other breeds
136 in common 50k/seq

a bunch of ugly tables

scenarios

50k –> 700k –> seq
50k –> seq
all seq
all seq minus breed of interest
only breed of interest
imputed for imputation

concordance for 136 samples with seq data

cheating

one-step

	AA	AB	BB
AA	97.14	2.65	0.22
AB	1.10	94.85	4.05
BB	0.03	1.07	98.91

two-step

	AA	AB	BB
AA	95.69	3.98	0.33
AB	1.40	91.82	6.78
BB	0.02	1.09	98.89

6292+224

honest

other+interest

6292+0

	AA	AB	BB
AA	76.03	18.97	5.01
AB	12.29	56.95	30.76
BB	0.71	8.09	91.2

6292+88

	AA	AB	BB
AA	90.45	8.84	0.71
AB	5.15	82.15	12.7
BB	0.10	3.23	96.67

0+88

	AA	AB	BB
AA	89.21	9.93	0.86
AB	4.63	85.82	9.54
BB	0.18	4.14	95.68

0+190

	AA	AB	BB
AA	93.69	5.91	0.40
AB	2.47	89.54	8.00
BB	0.04	1.93	98.03

imputed 50k - 9596

	AA	AB	BB
AA	97.11	2.67	0.22
AB	1.11	94.79	4.10
BB	0.03	1.07	98.90

mean concordance

other breeds	0.882
other+interest	0.948
interest (88)	0.945
interest (190)	0.963
imputed	0.983

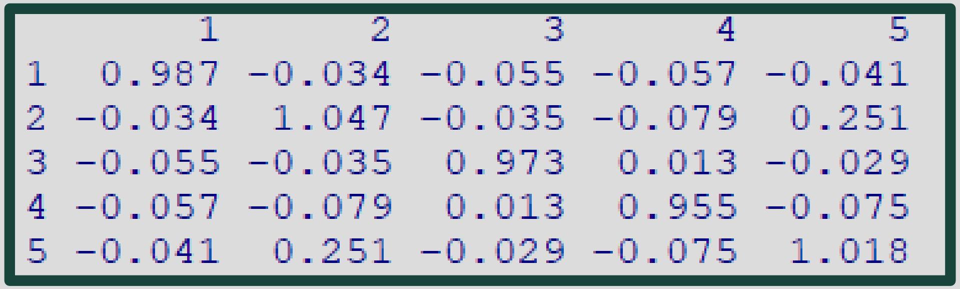
Nawaz et al. 2022

a couple of pretty equations

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda G^{-1} \end{pmatrix} \begin{pmatrix} \hat{b} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

$$G = \frac{M'M}{\sum_{i=1}^{m} 2p_i(1-p_i)}$$

GRM

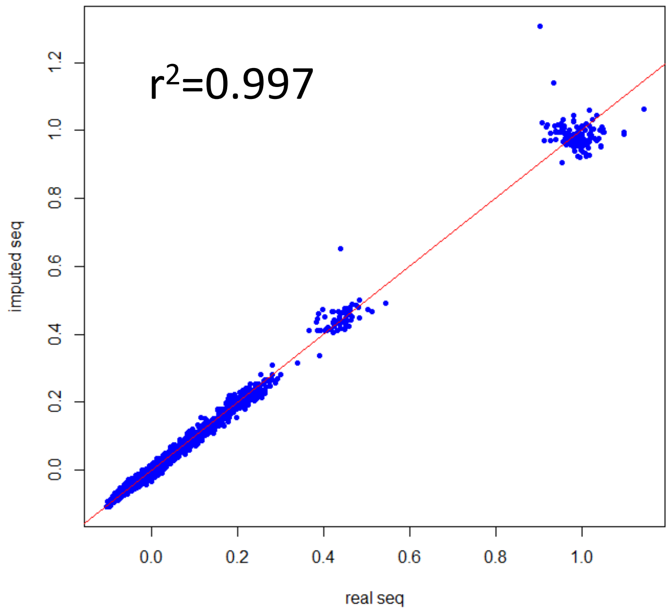


```
         1      2      3      4      5
1  0.987 -0.034 -0.055 -0.057 -0.041
2 -0.034  1.047 -0.035 -0.079  0.251
3 -0.055 -0.035  0.973  0.013 -0.029
4 -0.057 -0.079  0.013  0.955 -0.075
5 -0.041  0.251 -0.029 -0.075  1.018
```

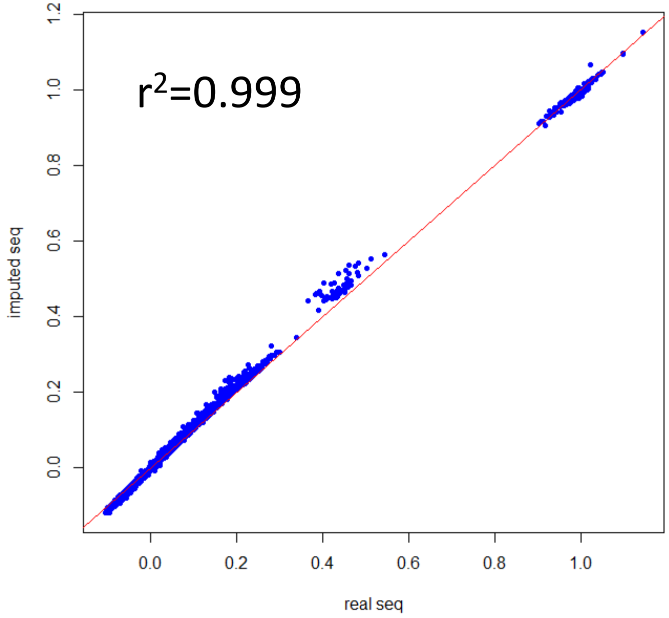
it’s all in the genomic relationship matrix (GRM)

comparing the GRMs

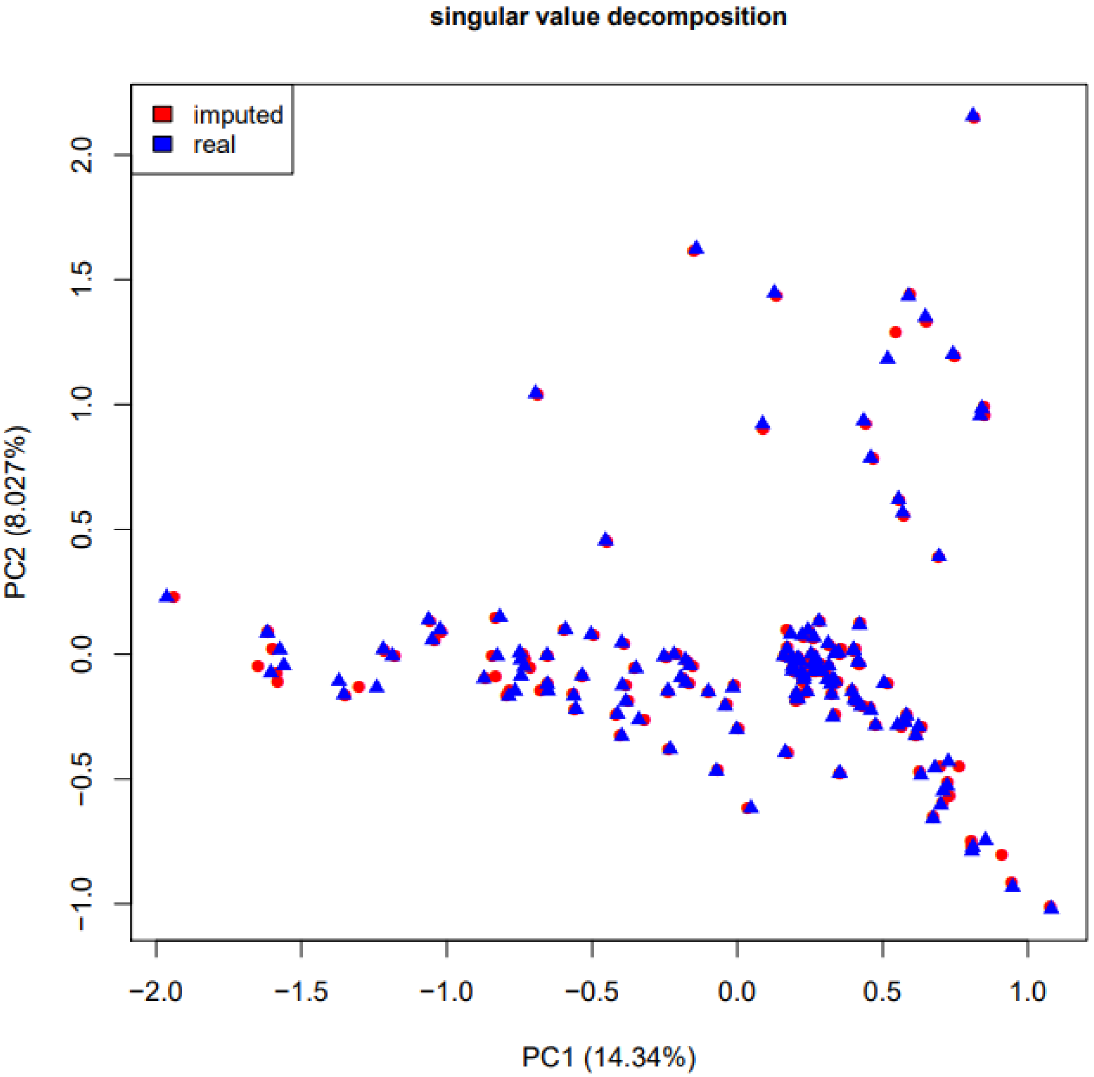
without the breed of interest



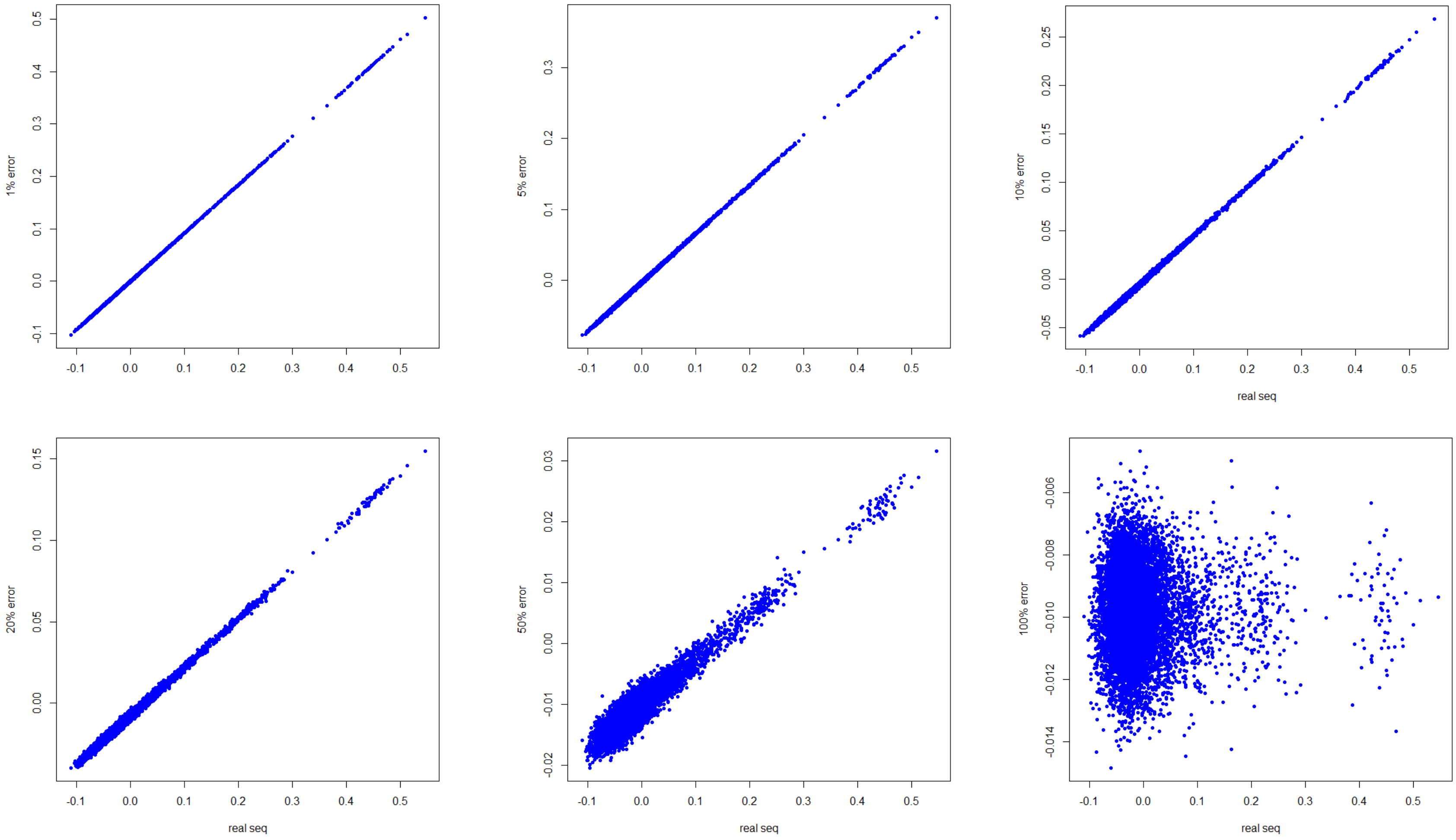
with the breed of interest



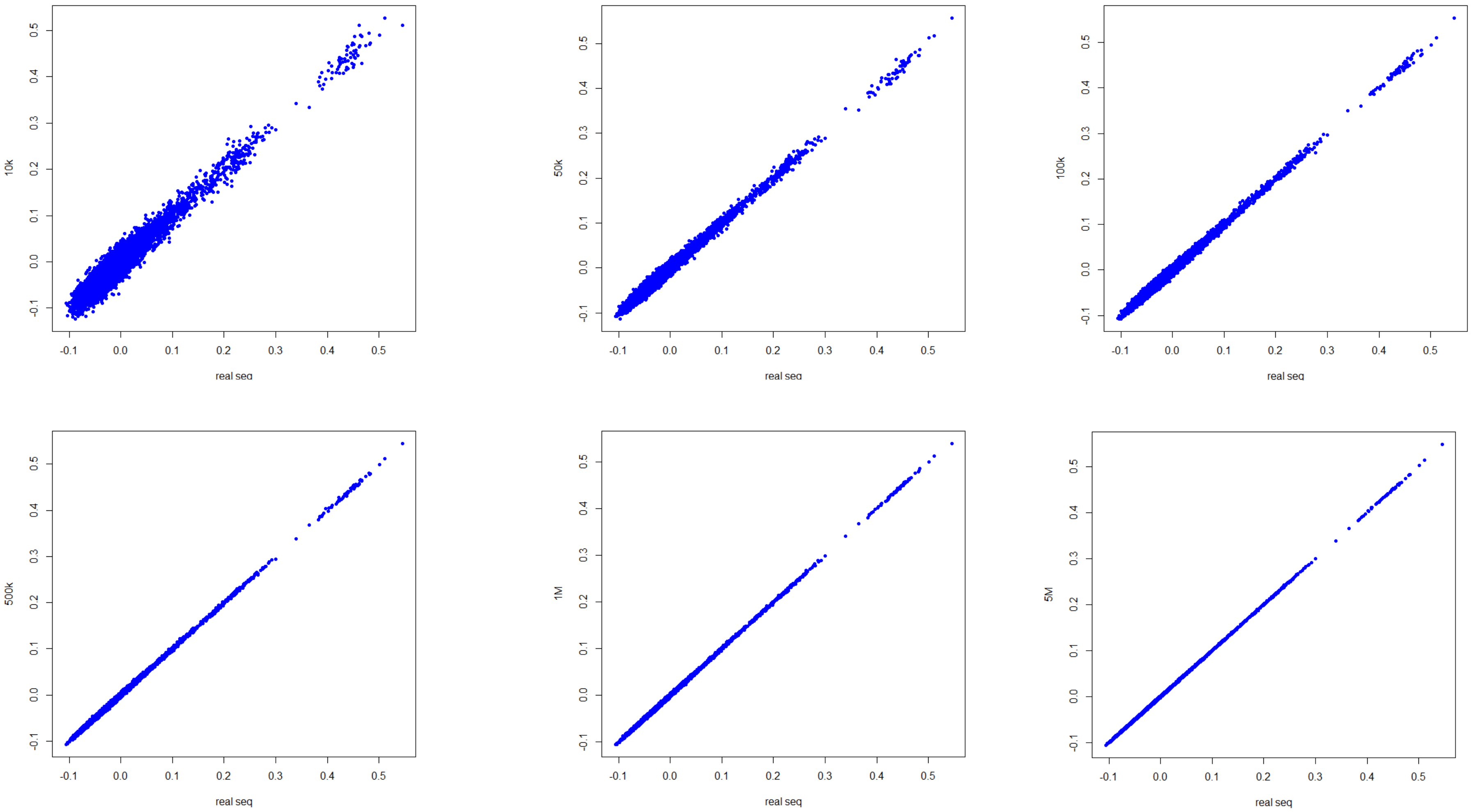
if the GRM does not change the EBVs do not change


singular value decomposition
🟥 imputed
🟦 real
PC2 (8.027%)
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
PC1 (14.34%)

how much do errors actually change the GRM?



changes to the GRM at different SNP panel densities



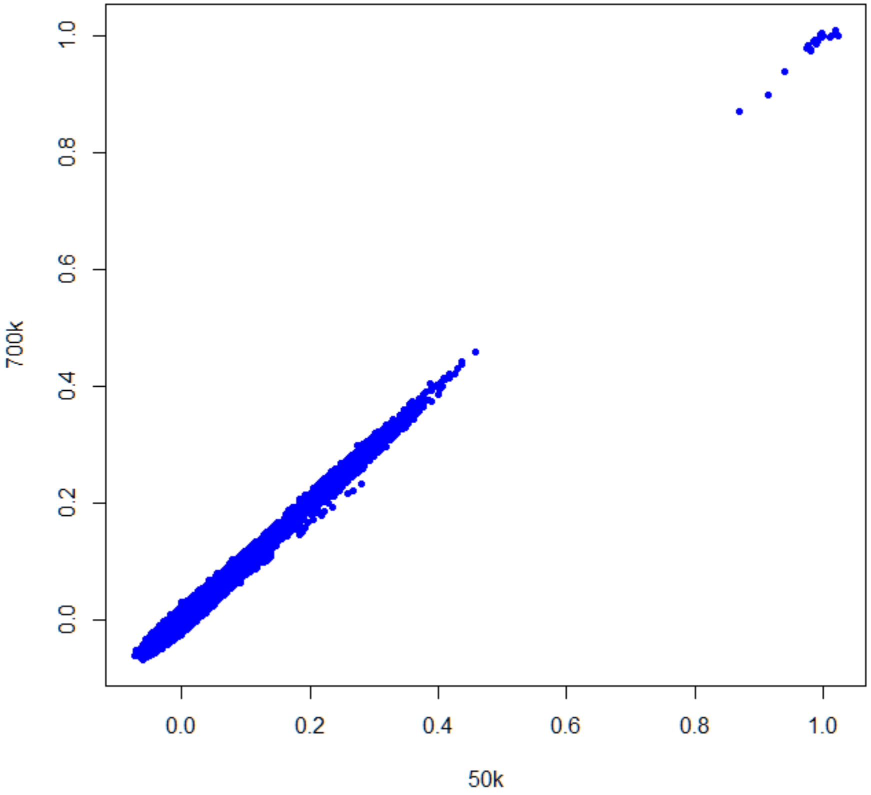
what has changed in the imputed data?

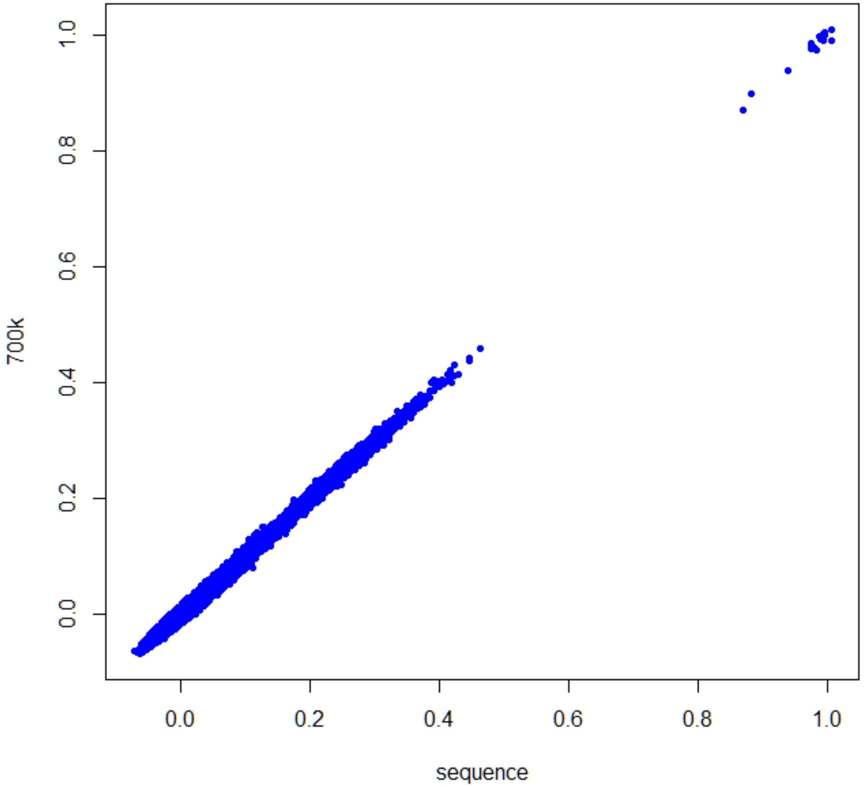
it is the change in G that will change the predictions

G50k x G700k = 0.9884

G700k x Gseq = 0.9950

G50k x Gseq = 0.9839

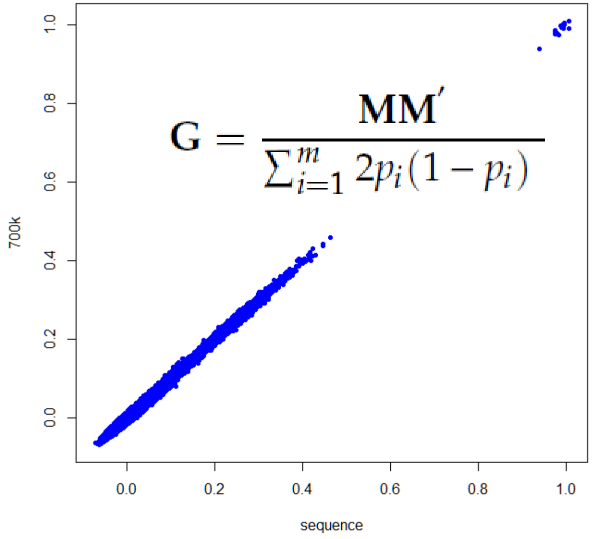




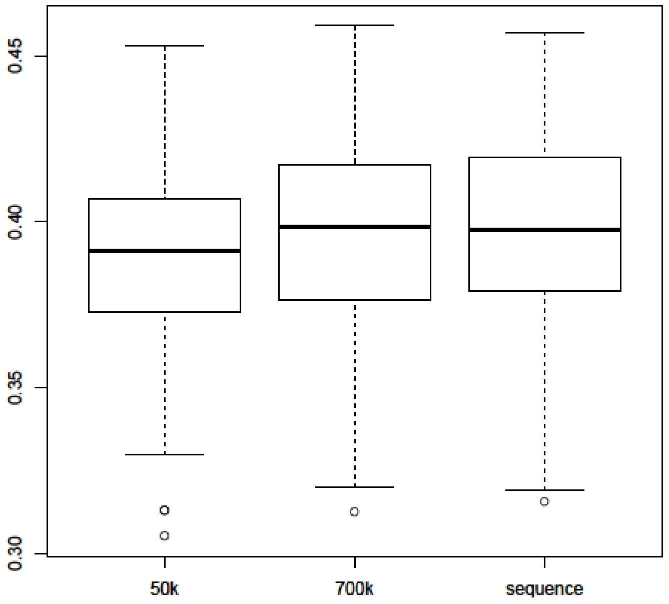
gEBVs from sequence data

limited benefits if business as usual

changes in GRM after 100k are minimal



$$\mathbf{G} = \frac{\mathbf{MM}'}{\sum_{i=1}^{m} 2p_i(1-p_i)}$$



50k → 700k (+1.5%) → sequence (+0.6%)

very small improvement in accuracy

accuracy of prediction	0.389	0.395	0.398
% increase accuracy		1.52	0.63

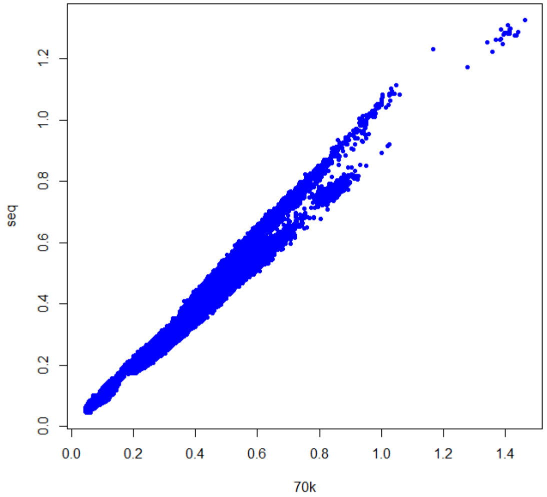
average of 100-fold cross validation | 1800 training | 518 validation

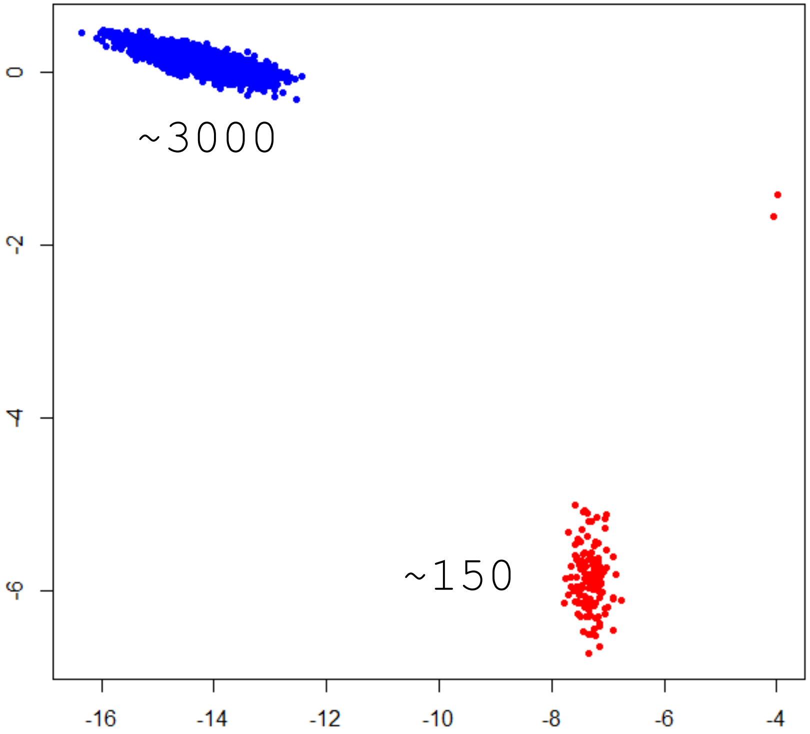
so, what's the point?

larger benefits in multi-breed systems

seq	red	blue
red	0.13	0.07
blue	0.13	0.19

70k	red	blue
red	0.09	0.07
blue	0.05	0.19





seq helps with small sample sizes

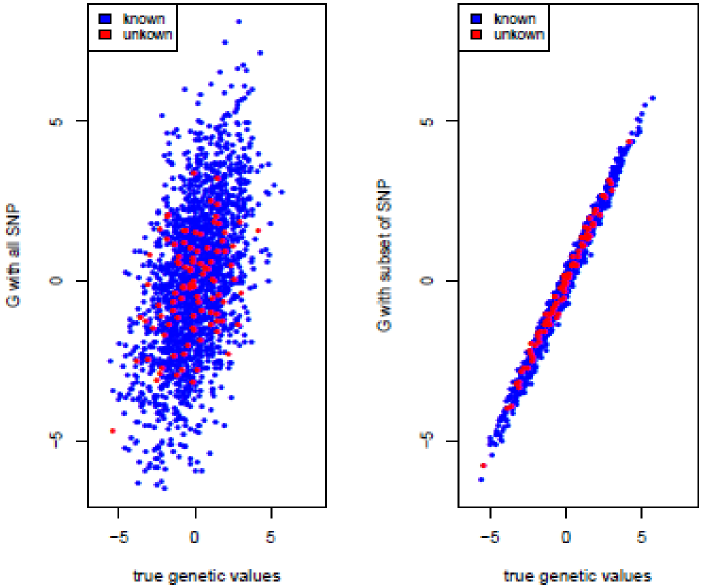
seq helps with crossbreed prediction

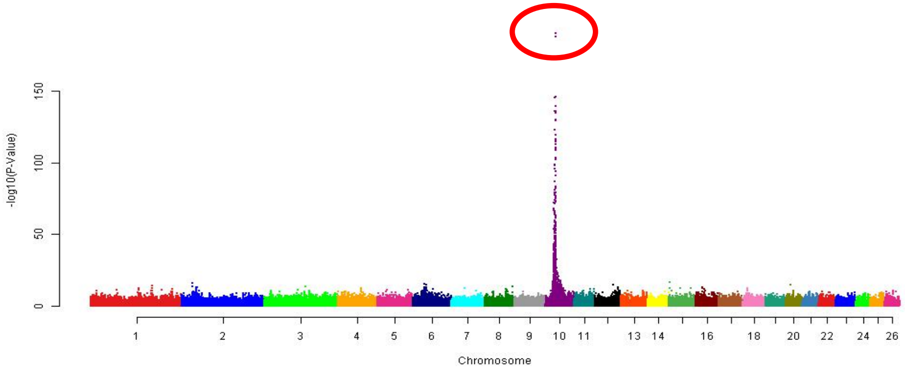
…and a bit of AI for genomic prediction

AI generated image with DALL-E

ideally…

- in a perfect world we would know the true SNP associated to a trait or even better, the functional causal variants
- we would know the variants of large effect but also all the ones with small effects
- and we would use only them for making predictions…



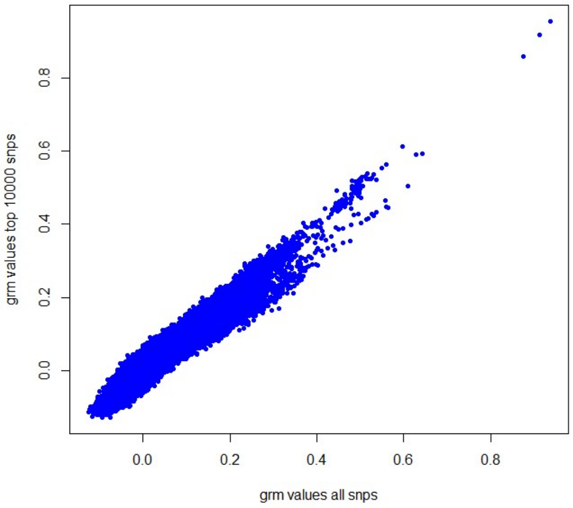


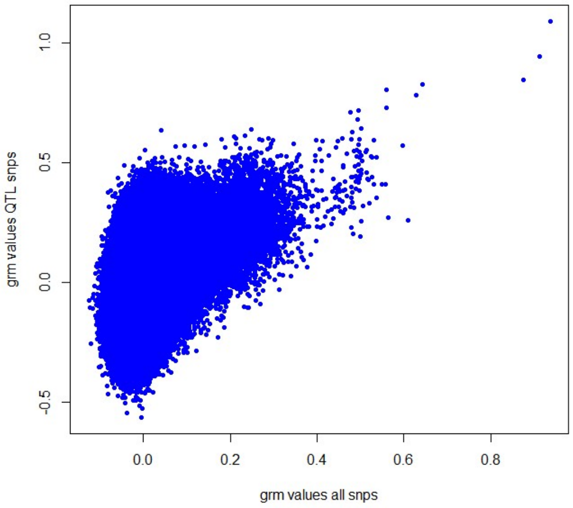
Use *trait G* instead of G

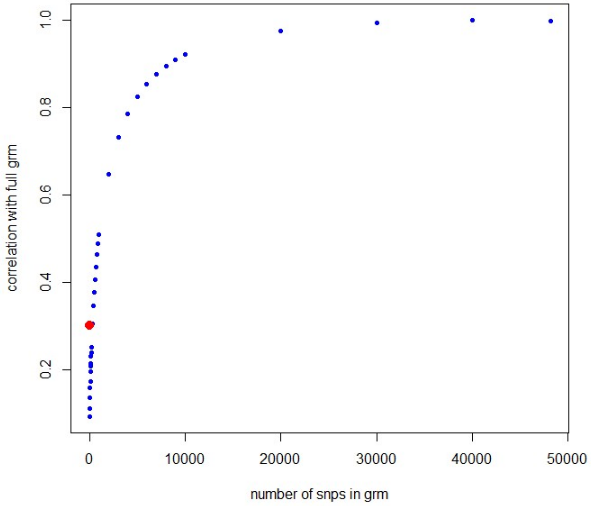
trait relationship matrix

gBLUP using only *functional* markers

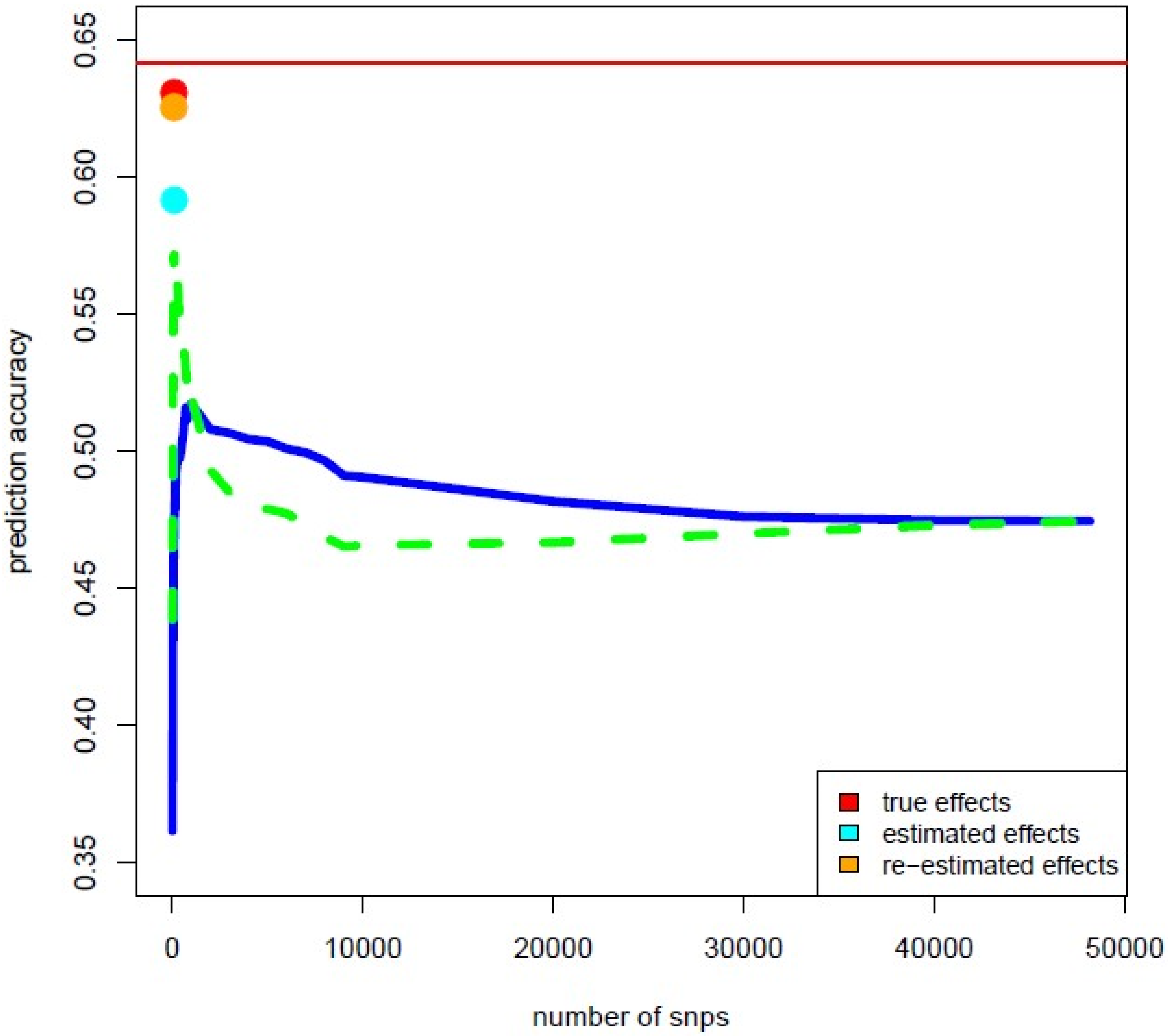
genomic prediction as a feature selection problem



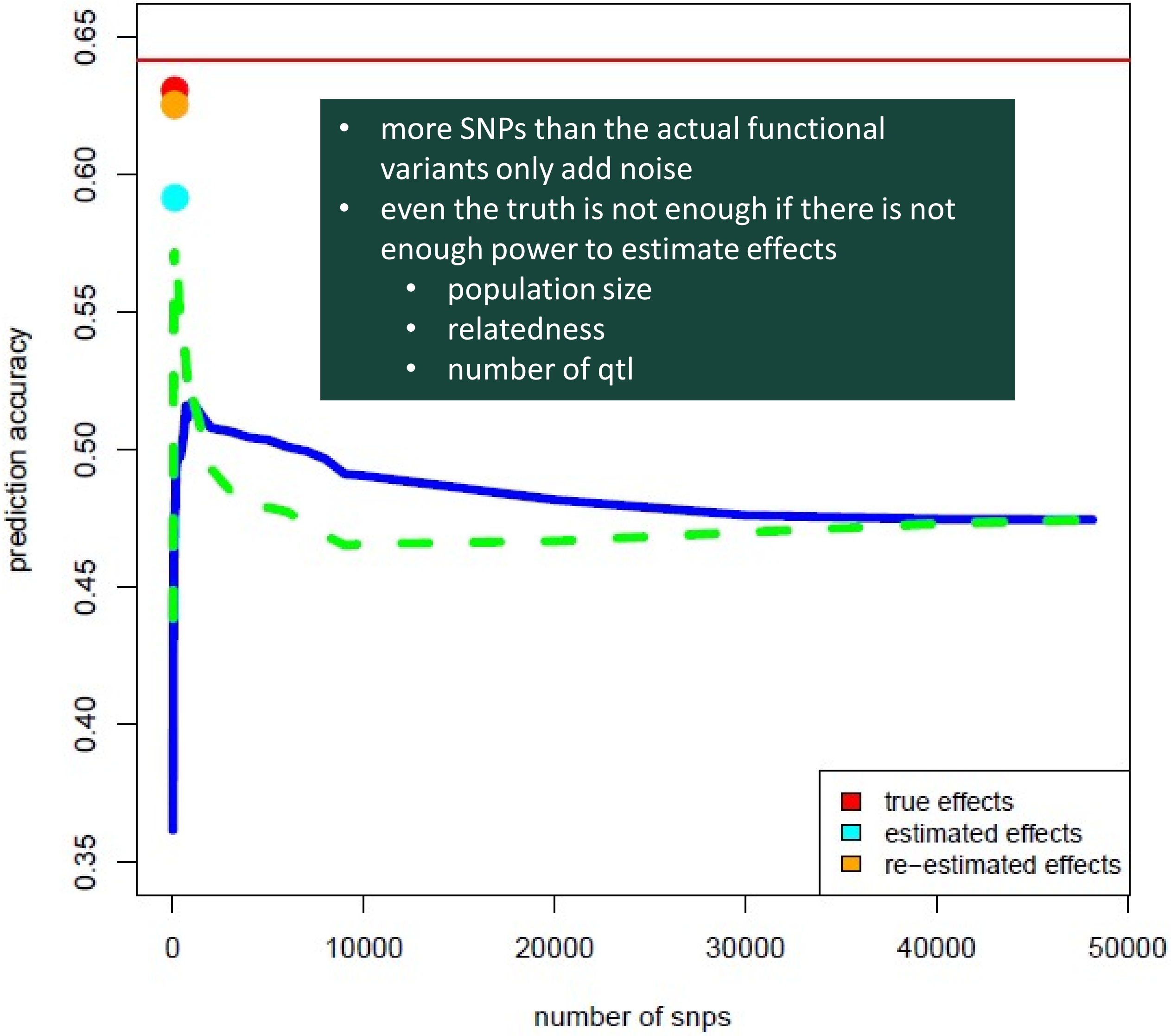




the 'real' (unknown) grm is very different from the full grm



spurious SNPs just add noise to the prediction



- more SNPs than the actual functional variants only add noise
- even the truth is not enough if there is not enough power to estimate effects
  - population size
  - relatedness
  - number of qtl

spurious SNPs just add noise to the prediction

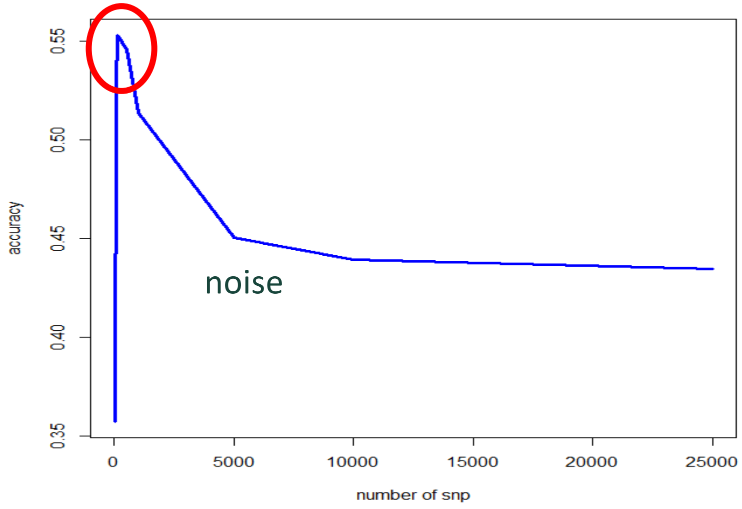
iterative weighted gblup with local search

- split population into 3 parts – training, internal testing and external testing
- perform weighted gBLUP and iterate until the weights converge
- find a rough number of SNP to use based on accuracy of sorted SNP
- test every SNP and check if it improves/worsens prediction accuracy in internal testing set – remove non-informative SNP
- refit final SNP set with gBLUP
- evaluate on external testing data

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z+\lambda G^{-1} \end{pmatrix} \begin{pmatrix} \hat{b} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

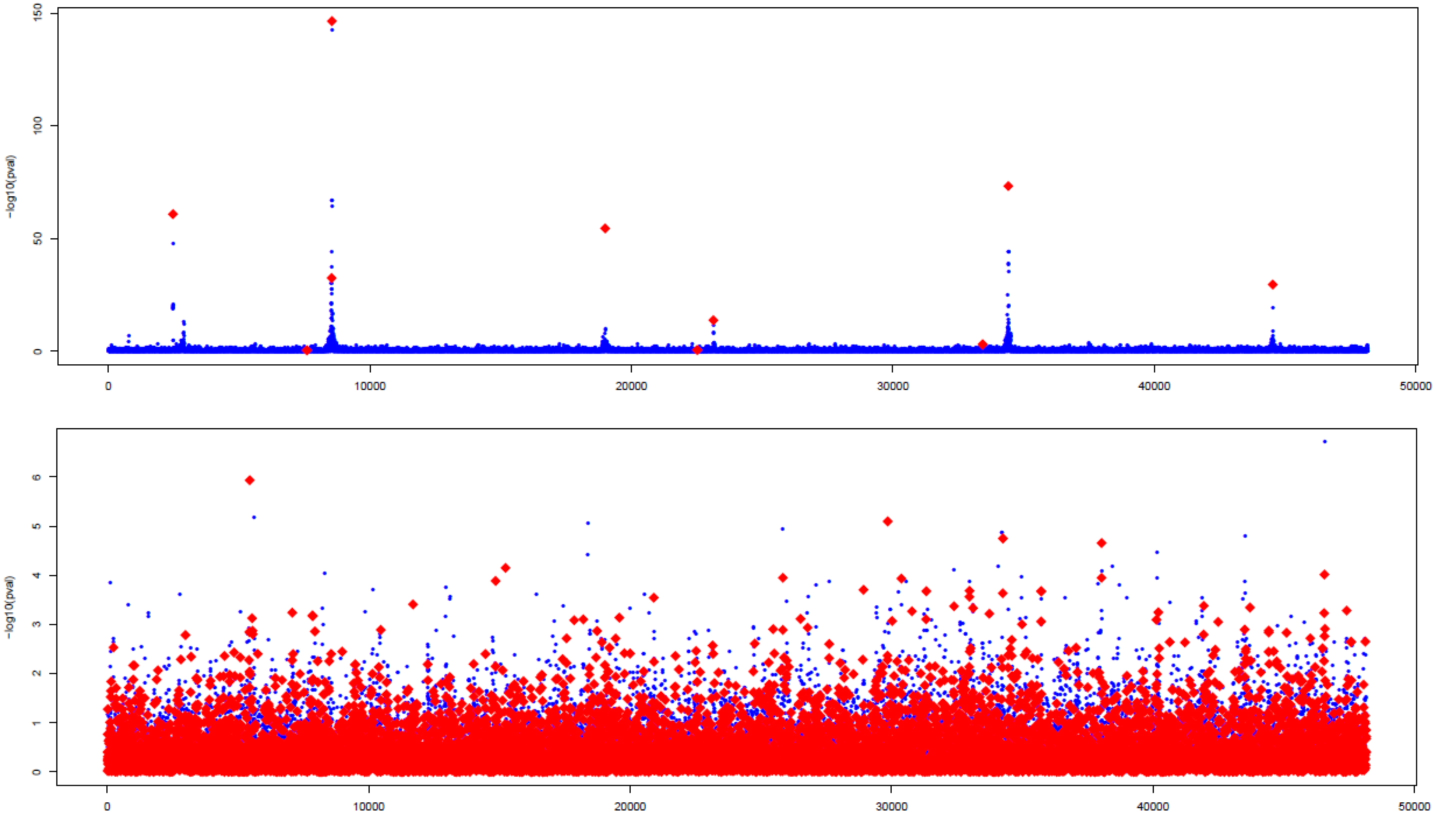
$$\mathbf{G} = \frac{\mathbf{MM}'}{\sum_{i=1}^{m} 2p_i(1-p_i)}$$

$$\mathbf{G}^* = \frac{\mathbf{MDM}'}{\sum_{i=1}^{m} 2p_i(1-p_i)}$$

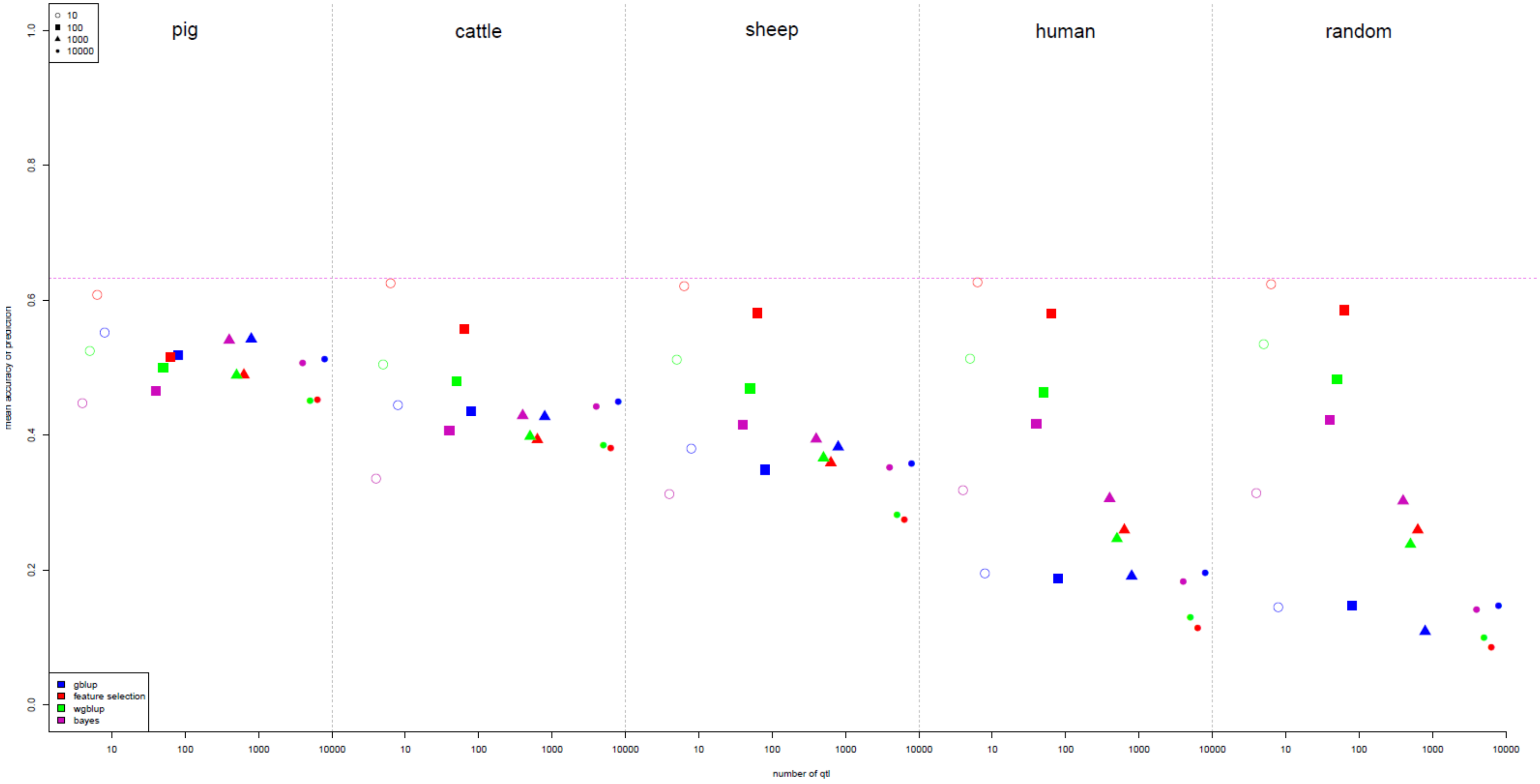


if we get this right:
accuracies should hold across generations
can combine multiple breeds and crosses
costs can be reduced
computational burden can be reduced

signal to noise ratio of a trait – *genetic architecture*

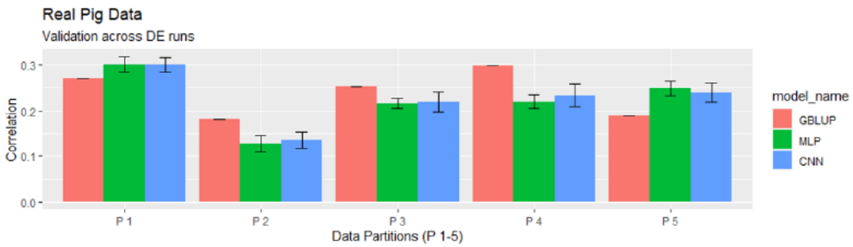


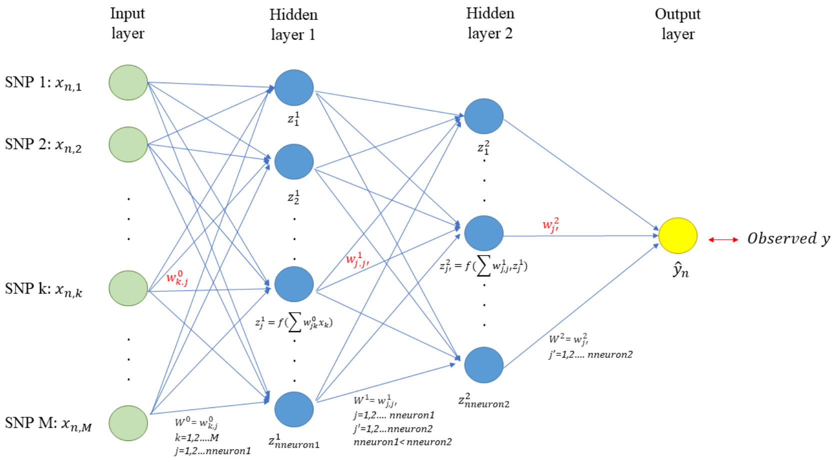
methods comparison

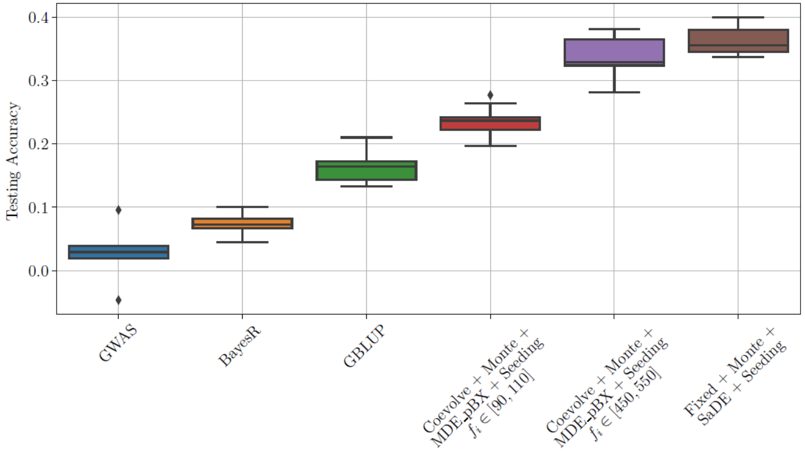


prediction is a function of sample size, genetic architecture, relatedness

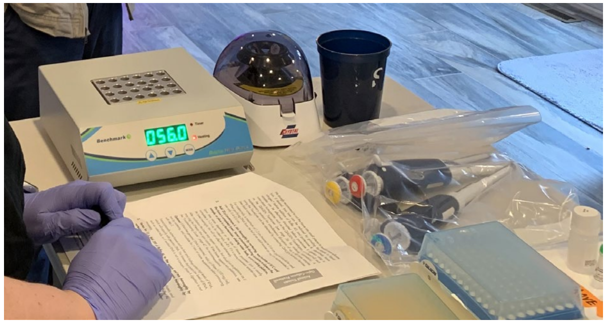
machine learning for genomic prediction MLP, CNN, DE, XGboost

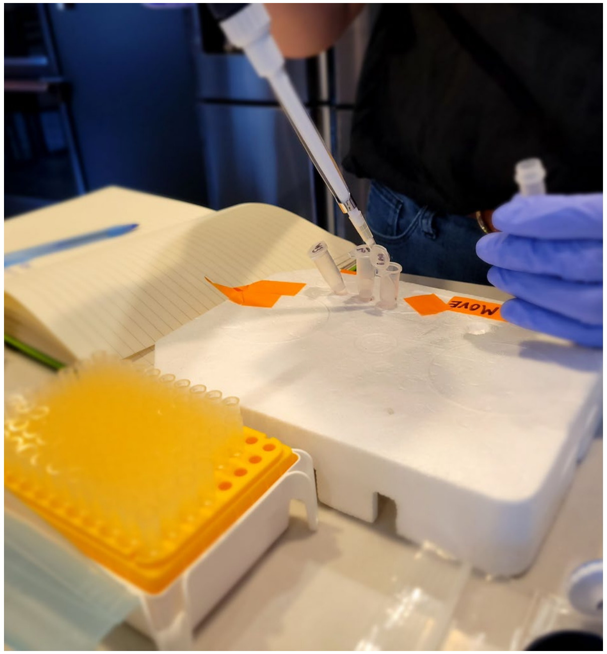


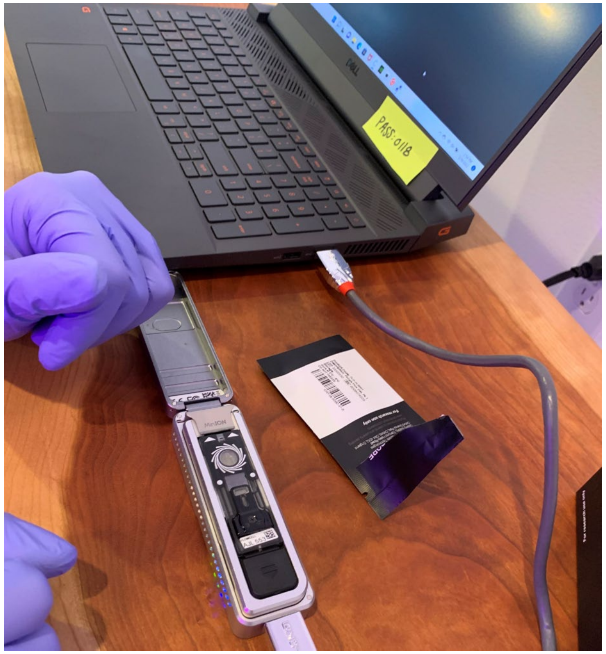


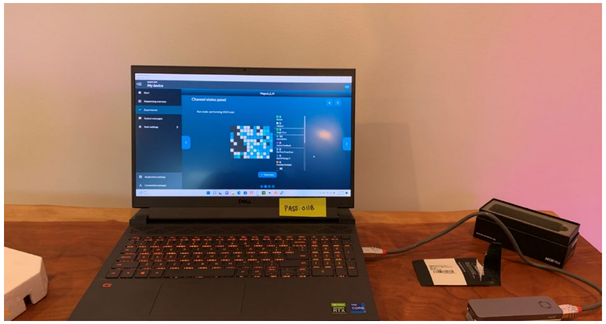


one of these days in the future…





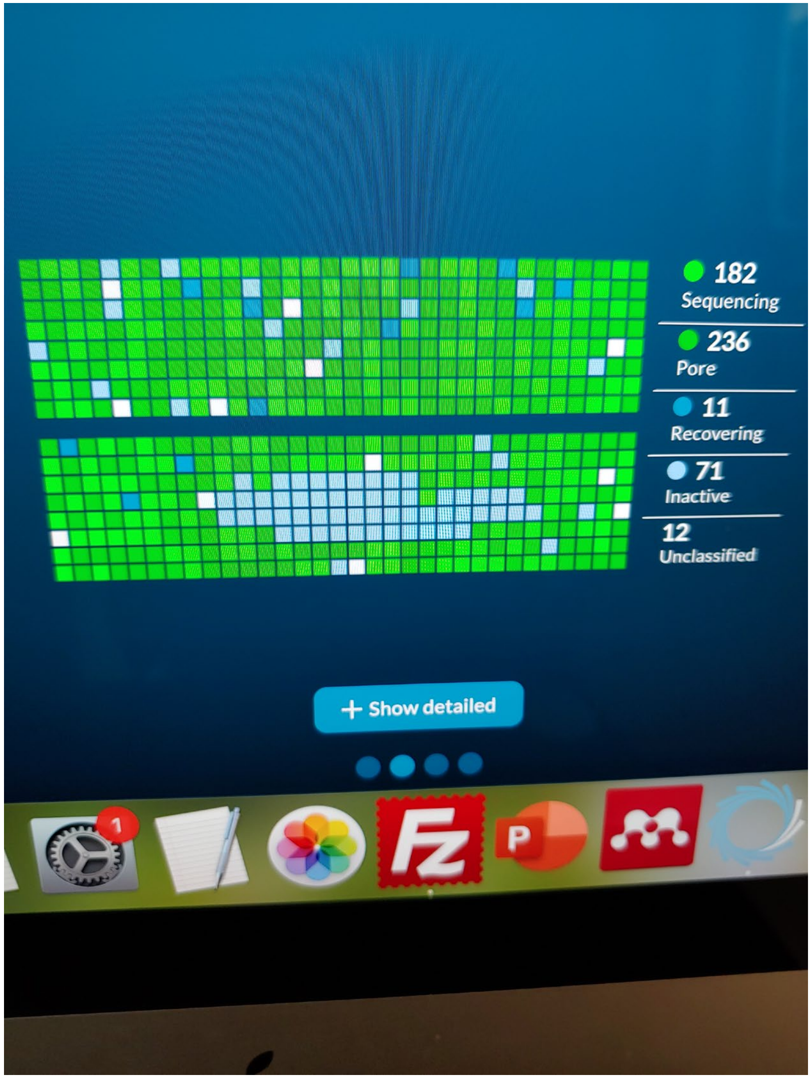




home sequencing

the future is kind of already here, just maybe a tad less glamorous

Nanopore MinION



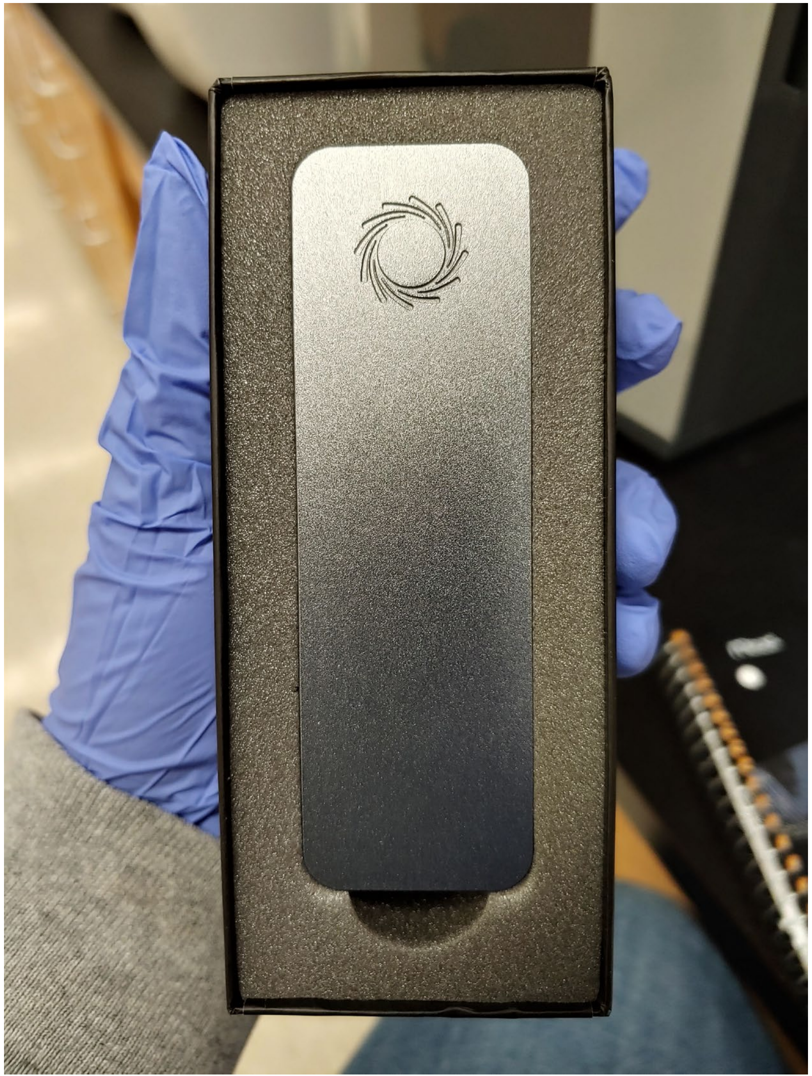
● 182 Sequencing

● 236 Pore

● 11 Recovering

● 71 Inactive

12 Unclassified



Ostrovski

don’t need to send samples to a lab for genotyping anymore

- portable sequencer – pocket sized, USB connection, 87g
- can produce long and ultra-long reads

applications and limitations

onsite sequencing without a lab or specialized personnel

farm

determine parentage, ***breed composition***, test for recessives and estimate breeding values

turnover time from sample to knowledge of less than four hours (?)

disease testing

positive/negative results in a couple of hours

supply chain

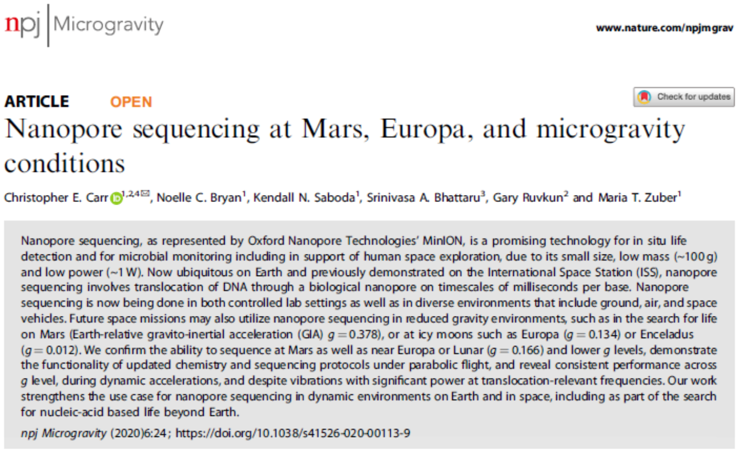
origin of product can be regulated/certified on site by DNA testing (breed, provenance…)

food safety

rapidly traced back through the supply chain by matching the DNA signature of the contaminated product with sequences stored in databases

cons

- takes some practice
- reagents not stable at room temperature, short shelf life
- still need to perform DNA extraction
- prices not yet competitive with lab genotyping
- data structures need to be in place for analyses
- great for a few samples but does not scale up



npj | Microgravity

www.nature.com/npjmgrav

ARTICLE OPEN

Check for updates

Nanopore sequencing at Mars, Europa, and microgravity conditions

Christopher E. Carr[1,2,4], Noelle C. Bryan[1], Kendall N. Saboda[1], Srinivasa A. Bhattaru[3], Gary Ruvkun[2] and Maria T. Zuber[1]

Nanopore sequencing, as represented by Oxford Nanopore Technologies' MinION, is a promising technology for in situ life detection and for microbial monitoring including in support of human space exploration, due to its small size, low mass (~100 g) and low power (~1 W). Now ubiquitous on Earth and previously demonstrated on the International Space Station (ISS), nanopore sequencing involves translocation of DNA through a biological nanopore on timescales of milliseconds per base. Nanopore sequencing is now being done in both controlled lab settings as well as in diverse environments that include ground, air, and space vehicles. Future space missions may also utilize nanopore sequencing in reduced gravity environments, such as in the search for life on Mars (Earth-relative gravito-inertial acceleration (GIA) $g = 0.378$), or at icy moons such as Europa ($g = 0.134$) or Enceladus ($g = 0.012$). We confirm the ability to sequence at Mars as well as near Europa or Lunar ($g = 0.166$) and lower g levels, demonstrate the functionality of updated chemistry and sequencing protocols under parabolic flight, and reveal consistent performance across g level, during dynamic accelerations, and despite vibrations with significant power at translocation-relevant frequencies. Our work strengthens the use case for nanopore sequencing in dynamic environments on Earth and in space, including as part of the search for nucleic-acid based life beyond Earth.

npj Microgravity (2020)6:24; https://doi.org/10.1038/s41526-020-00113-9

acknowledgements

Dion Detterer
Beatriz Cuyabano
Rodrigo Savegnago
Ken Reid
Ian Whalen
Hawlader Al-Mamun
Yasir Nawaz
Salman Ali
Jacob Newsted
Stephen Kelly
Hanna Ostrovski
Andre Nascimento
Andrea Romero
Junjie Han
Penda Ndiaye
Gabriel Rovere
Bayode Makanjuola

Paul Kwan
Juan Steibel
Wolfgang Banzhaf

USDA
United States Department of Agriculture
National Institute of Food and Agriculture

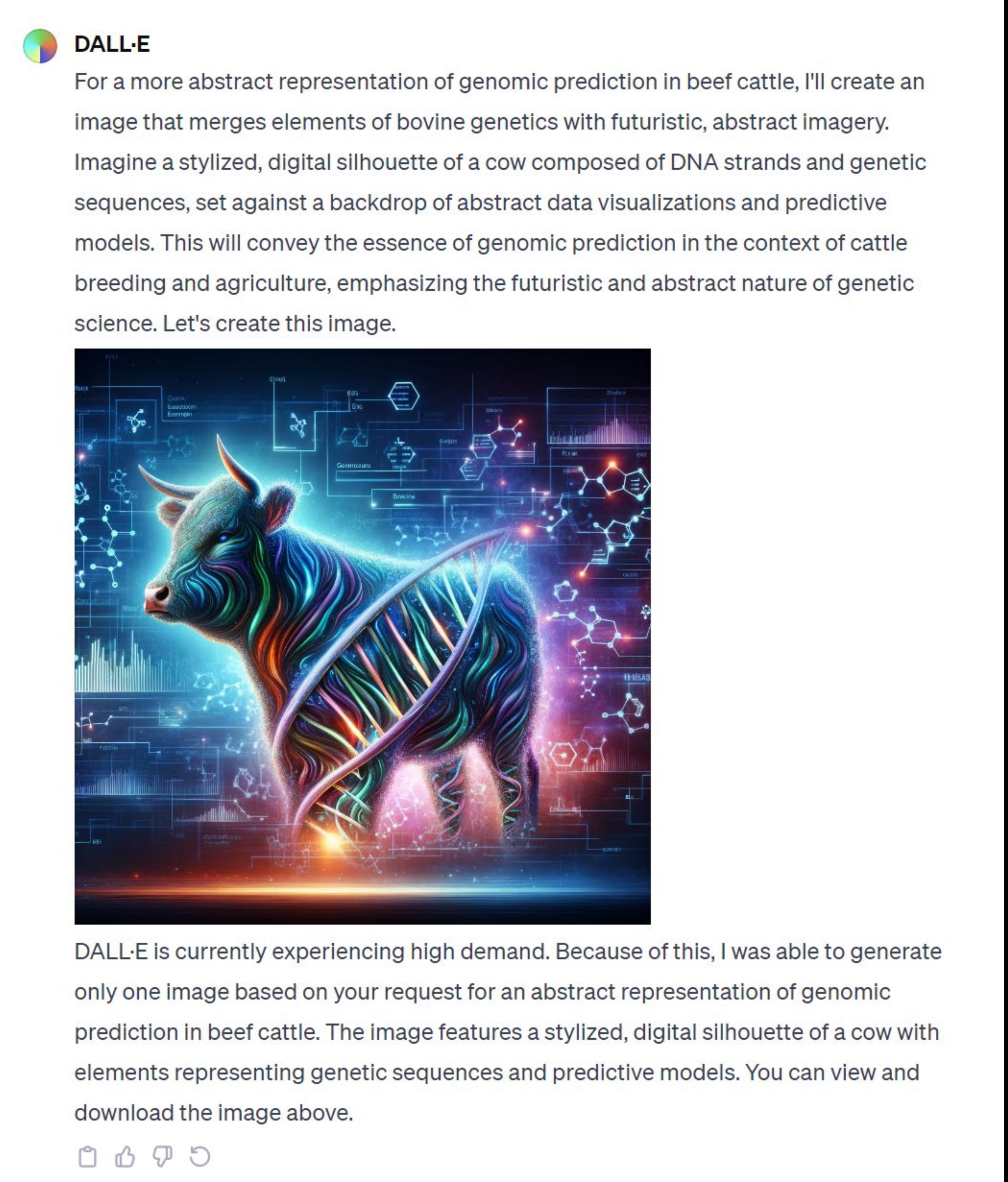
RDA
RURAL DEVELOPMENT ADMINISTRATION
National Institute of Animal Science

BEACON
An NSF Center for the Study of Evolution in Action

zoetis

questions?

Cedric Gondro
Professor of Computational and Quantitative Genomics
Michigan State University
Department of Animal Science
gondroce@msu.edu



DALL·E

For a more abstract representation of genomic prediction in beef cattle, I'll create an image that merges elements of bovine genetics with futuristic, abstract imagery. Imagine a stylized, digital silhouette of a cow composed of DNA strands and genetic sequences, set against a backdrop of abstract data visualizations and predictive models. This will convey the essence of genomic prediction in the context of cattle breeding and agriculture, emphasizing the futuristic and abstract nature of genetic science. Let's create this image.

DALL·E is currently experiencing high demand. Because of this, I was able to generate only one image based on your request for an abstract representation of genomic prediction in beef cattle. The image features a stylized, digital silhouette of a cow with elements representing genetic sequences and predictive models. You can view and download the image above.