Strategies to choose from millions of imputed sequence variants

J. R. O'Connell[1] and P. M. VanRaden[2]

[1]University of Maryland School of Medicine, Baltimore, MD, USA

[2]Animal Genomics and Improvement Laboratory, Agricultural Research Service, USDA, Beltsville, MD, USA

joconnel@medicine.umaryland.edu

Introduction

- Whole-genome sequencing costs are rapidly declining
- Large numbers of animals will be sequenced in the next few years
- How will WGS be used in genomic evaluations?

Applications of WGS data

- Develop reference panels for imputation using SNP chip
- Discover QTLs
- Discover rare and de novo mutations with deleterious effects for surveillance
- Build better SNP chips

WGS implementation considerations

- Cost of higher density SNP chips to industry
- Imputation cost for weekly/monthly genomic evaluations
- Redesigning chip incurs cost of generating new data for imputation
  - Build off the 60K chip

Project goal: Prepare for WGS

- Develop programs to simulate genotypes and QTLs
- Develop programs to process sequence data
- Compare strategies for variant selection
- Predict gains in reliability
- Determine computational footprint for large populations

Data

- 26,984 HOL bulls in U.S. reference population
- 112,905 animals in the pedigree
- 30 million simulated variants, 10,000 QTLs
- 30 equal-length chromosomes (100 Mbases)
- 3 different chip densities (HD, MD, LD)
- 5 independent traits (same QTL locations)

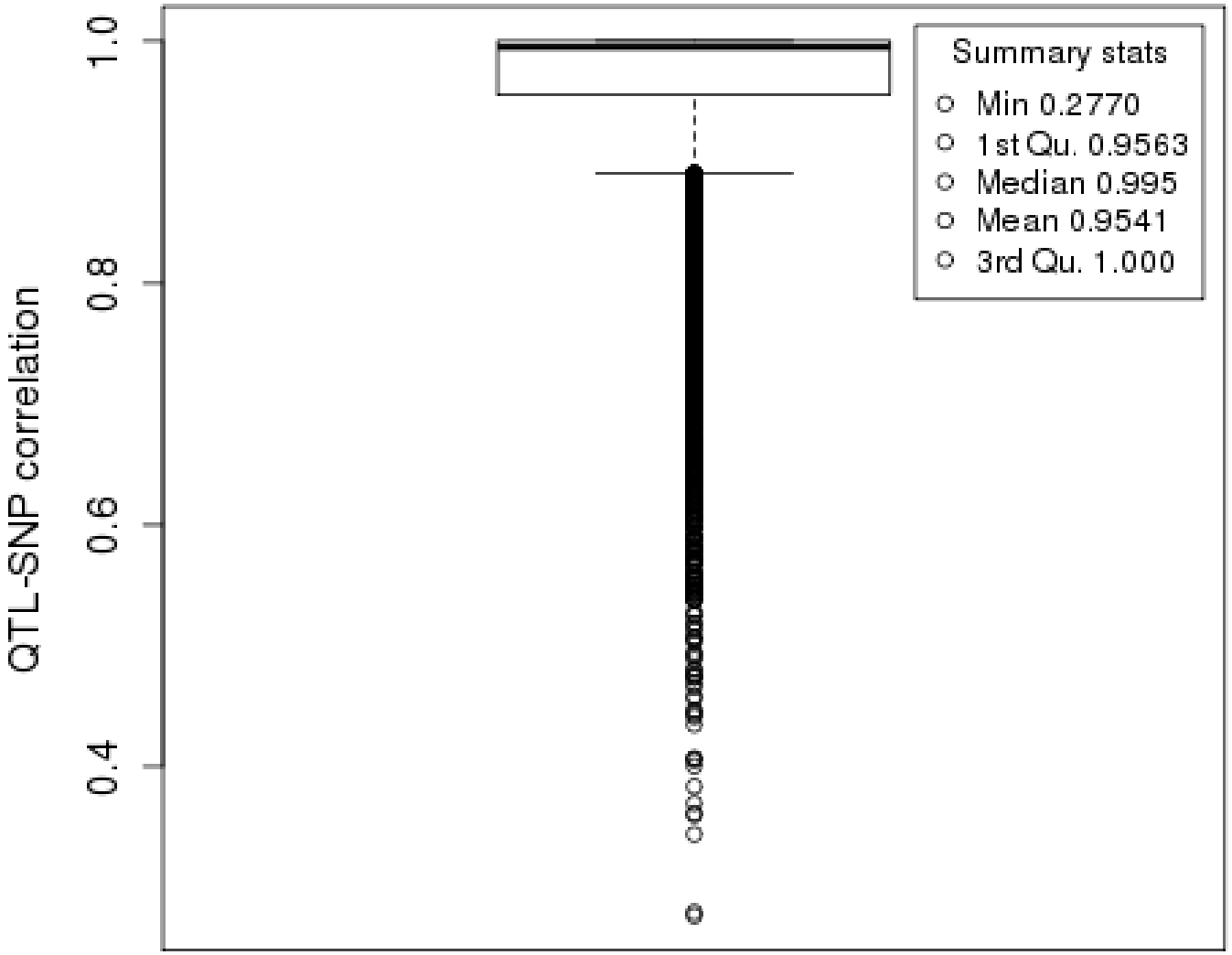
Genotype simulation and pruning

- Simulate 30 million genotypes plus QTLs in 1000 bulls using genosim.f90
- Remove loci with MAF <0.01 or linkage disequilibrium >0.95
- Keep 500K loci near QTLs (genes) + 600K chip + 60K chip
- 8,403,858 loci remain

QTL effect distribution (% variance)

Top QTLs	Trait 1	2	3	4	5	Avg
1	3.7	13	3.9	3.5	5.6	5.9
10	20	34	21	23	29	25.4
100	57	63	59	57	62	59.6
1,000	90	92	92	92	93	91.8
10,000	100	100	100	100	100	100.0

QTL correlation to 8M SNP set



Bull genotypes: Imputation findhap.f90

Bulls	Genotype density	Variants
1,000	Sequenced pruned to	8,403,858
773	High density	600,000
24,863	Medium density	60,000
348	Low density	12,000
26,984	Total bulls, all imputed to	8,403,858
17,896	Phenotyped (old)	DYD
9,088	Validation (young)	True BV

Computation required

Step	Proc-essors	Time (hours)	Memory (Gbyte)	Disk (Gbyte)
Simulate 30M	1	56	210	32
Prune linkage	10	1	27	10
Impute 8M	20	38	13	220

Study1: Select 25K from 600K chip

- Compare 60K and 600K REL using same data
- Choose best markers from 600K, add to 60K
  - Select largest 5,000 for each of 5 traits
  - Multiple regression: By effect size, effect variance
  - GWA: By *p*-value significance
  - Merge and remove duplicates, adding ~23K

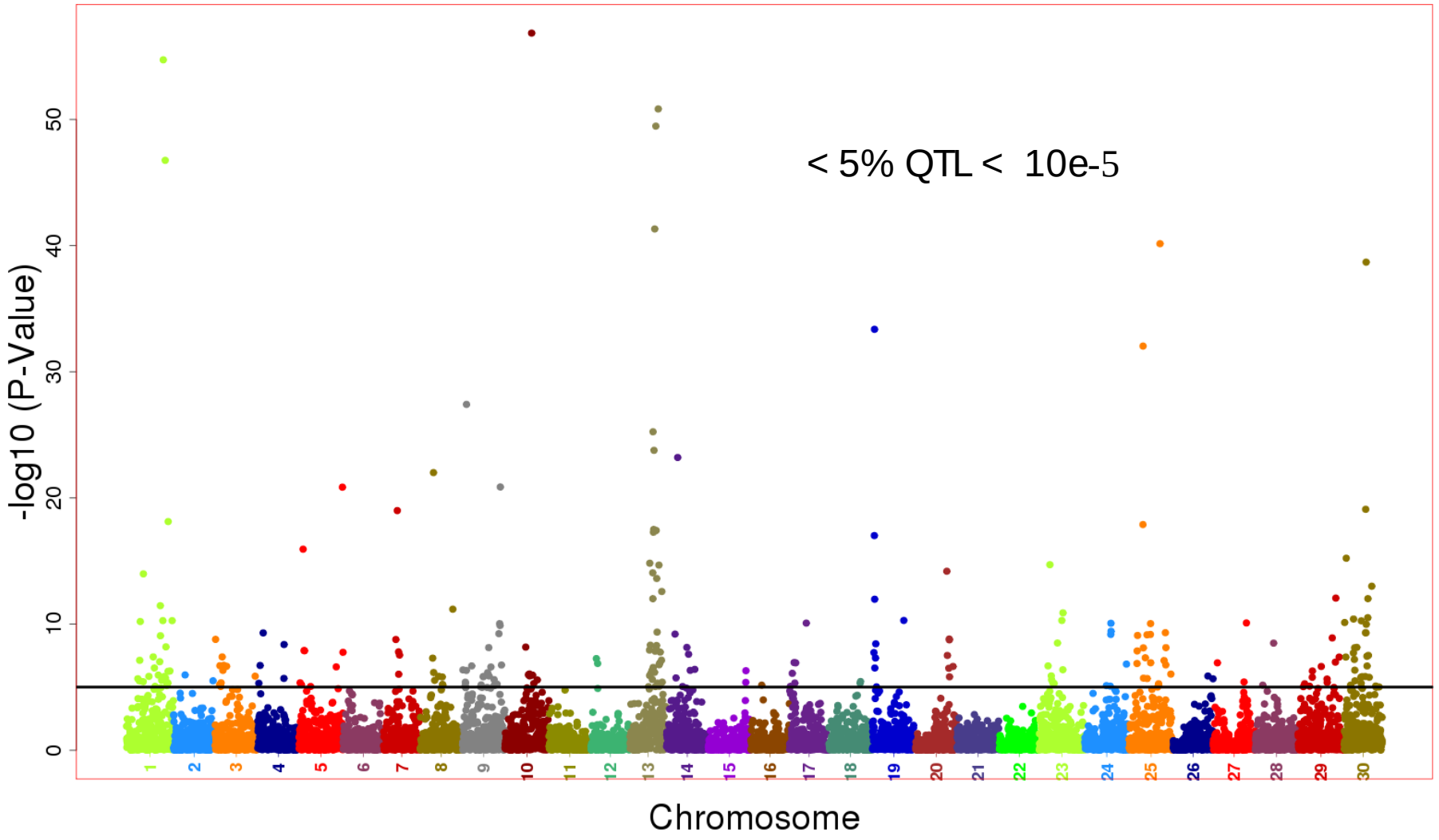
Results: REL from PA, 60K, 85K, 600K

			60K + 25K			
Trait	PA	60K	GWA	Size	Var	600K
1	24.4	77.9	79.2	81.6	81.3	80.3
2	31.2	77.9	79.3	81.4	81.2	80.1
3	32.7	78.3	79.5	81.3	81.5	80.4
4	23.3	76.6	77.7	80.2	79.8	78.6
5	30.4	78.3	80.0	82.5	82.2	81.2
Avg	28.4	77.8	79.1	81.4	81.2	80.1

Genome-wide association

- Single SNP regression within mixed model that accounts for the polygenic effect using pedigree kinship
  - $Y = Xb + SNP + a + e$
- Multiple testing adjustment leads to stringent cutoffs for significance
- Focus on estimation rather than prediction

GWA trait 2 imputed QTLs



Study 2: Select sequence variants

- 8M – search everywhere
- 1M – search by bioinformatics using 2.5kb window on QTL
- QTLs – perfect functional knowledge
- Include 60K chip for imputation, but assign all or most prior variance to the selected variants
  - Selection of variants may affect optimal shape/ choice of prior distribution

REL from 1M, 60K+1M subset, QTLs

		1 million near QTLs		Include QTLs	
Trait	600K	60K+25K	All 1M	60K+10K	10K
1	80.3	85.4	86.7	84.6	87.2
2	80.1	85.3	87.7	84.9	87.7
3	80.4	84.9	86.1	85.0	87.8
4	78.6	83.5	84.8	82.9	85.9
5	81.2	86.0	87.6	85.2	87.5
Avg	80.1	85.0	86.4	84.5	87.2

Computation required

Step	Proc-essors	Time (hours)	Memory (Gbyte)	Disk (Gbyte)
Simulate 30M	1	56	210	32
Prune linkage	10	1	27	10
Impute 8M	20	38	13	220
Predict 1M	5	22	20	<1
Select 25K from 8M	30	0.5	<1	<1

Conclusions

- Selection of variants requires several steps
- Potential gains from sequence are large
- Still requires very large reference population
- Methods scalable to more sequences
- Bioinformatics can narrow the search
- GWA is scalable to larger data but multiple regression give higher reliability

Current directions

- Evaluate several additional SNP selection criteria including priors for GWA
- Apply methods to real data from 1000 Bulls Project
- Evaluate bovine bioinformatics resources for functional annotation

Acknowledgments

- JRO supported by USDA SCA 58-45-14-070-1