Selection of sequence variants to improve dairy cattle genomic predictions

J. R. O'Connell,[1] M.E. Tooker,[2] D.M. Bickhart,[2] and J.B. Cole,[2] 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

2015 Interbull meeting presentation

- Strategies to choose from millions of imputed sequence (SEQ) variants
  - O’Connell and VanRaden
  - Based on simulated data

2015 simulated data

- 26,984 HOL bulls in U.S. reference population
- 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)

Simulation: REL from 1M, 60K+1M subset

		1 million near QTLs			
Trait	600K	60K+25K	*Difference*	All 1M	*Difference*
1	80.3	85.4	*5.1*	86.7	*6.4*
2	80.1	85.3	*5.2*	87.7	*7.6*
3	80.4	84.9	*4.5*	86.1	*5.7*
4	78.6	83.5	*4.9*	84.8	*6.2*
5	81.2	86.0	*4.8*	87.6	*6.4*
Avg.	80.1	85.0	*4.9*	86.4	*6.3*

1000 Bulls Genome Project

- 1000 Bulls Genome Project is an international SEQ project that seeks to pool resources in order to impute SEQ-derived genetic variants across a wide range of cattle breeds

- To join the project required a minimum of 25 animals sequenced at 10.5× coverage and approval by the project's steering committee
  - USDA contributed 76 bulls (26 Holstein)

Jeff O'Connell

1000 Bulls Genome Project *(continued)*

- SEQ alignment map created according to set specifications and collected from partners

- SAMtools used to identify SNPs and indels and produce genotype probabilities

- Beagle used for imputation

- Project data heavily processed, filtered, and imputed
  - 10% of 60K and HD SNPs missing

SNP vs. SEQ variants

- SNPs
  - At least 2 different nucleotides (A, C, G, or T) observed
  - Previous SNP chips only include these
- SEQ data has many insertions/deletions (indels)
  - Indels can range in length (up 50 bases)
  - Not easily captured by chip technology
  - Calls have lower quality than for SNPs
- Other more complex variant classes (such as copy number variants) were not identified from the raw data

Jeff O'Connell

Methods for HD+, HD+indels (HD+I), 77K

- Current HD chip has 312,614 usable SNPs after removing more than half due to high LD
- HD+: 481,904 candidate SEQ SNPs added
  - 107,471 exonic
  - 9,422 splice variants (same gene, different protein)
  - 35,242 untranslated regions at beginning and end of genes
  - 329,769 SNPs 2kb upstream or 1kb downstream of genes
- HD+I: Also added 249,966 indels in or near genes to HD+
- 77K: Add 17K to current 60K evaluation chip to compare with Wiggans’ 77K selected from HD

Jeff O’Connell

Edits to 39 million variants

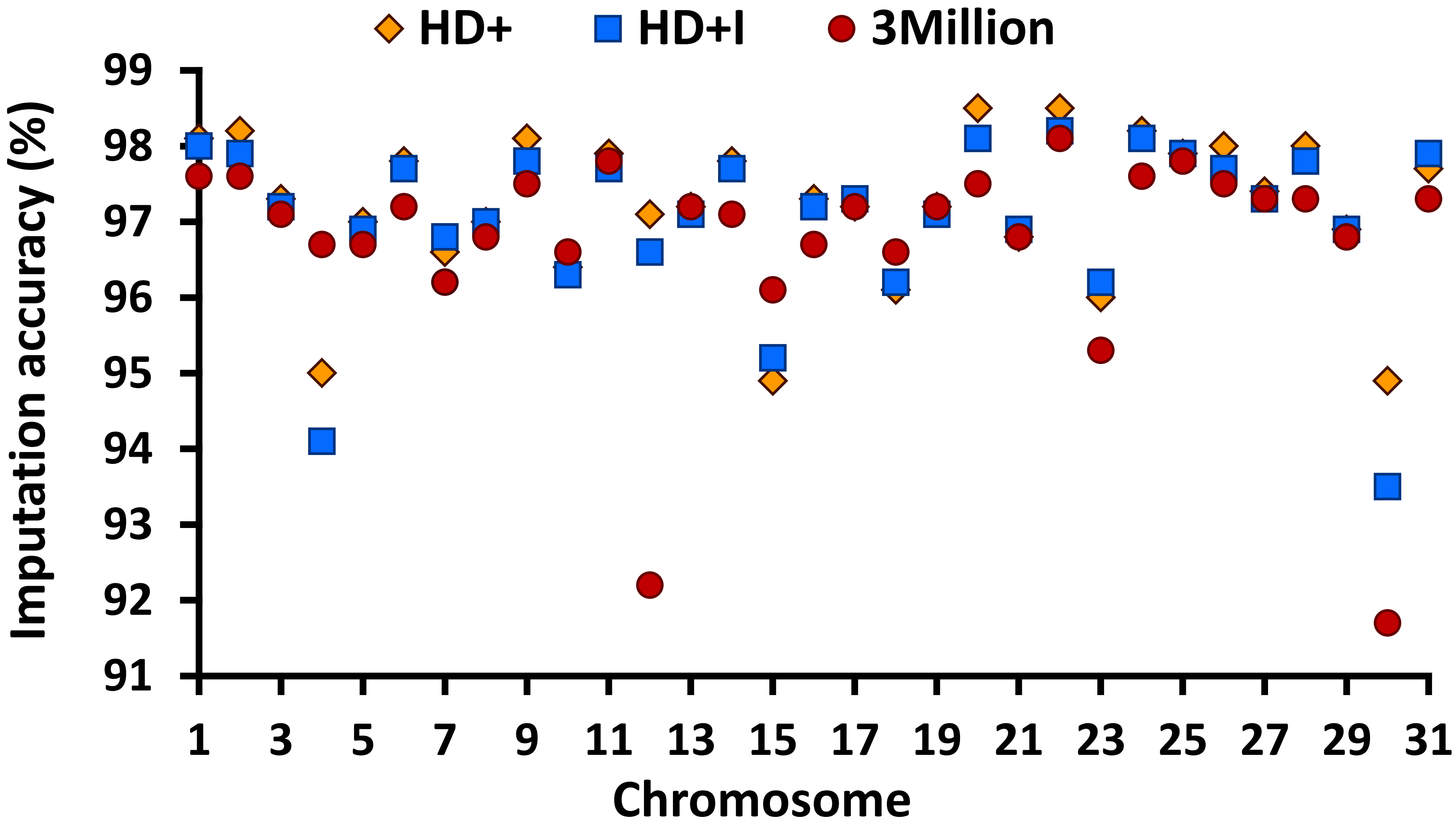
Edit	Number removed
Remove MAF < 0.01	20M
Remove for LD > 0.95	13M
Total removed	33M
Total remaining	6M
Imputation	
Remove for imputation accuracy	3M

Methods for imputation

- Imputation quality assessment
  - Select 40 of 440 SEQ Holsteins
  - Reduce to HD
  - Impute to SEQ
  - Compare with original SEQ
- HD imputed genotypes for 26,970 progeny-tested Holstein bulls
- Findhap designed for equally spaced markers, but SEQ-selected markers are bunched near genes

Jeff O'Connell

Imputation accuracy

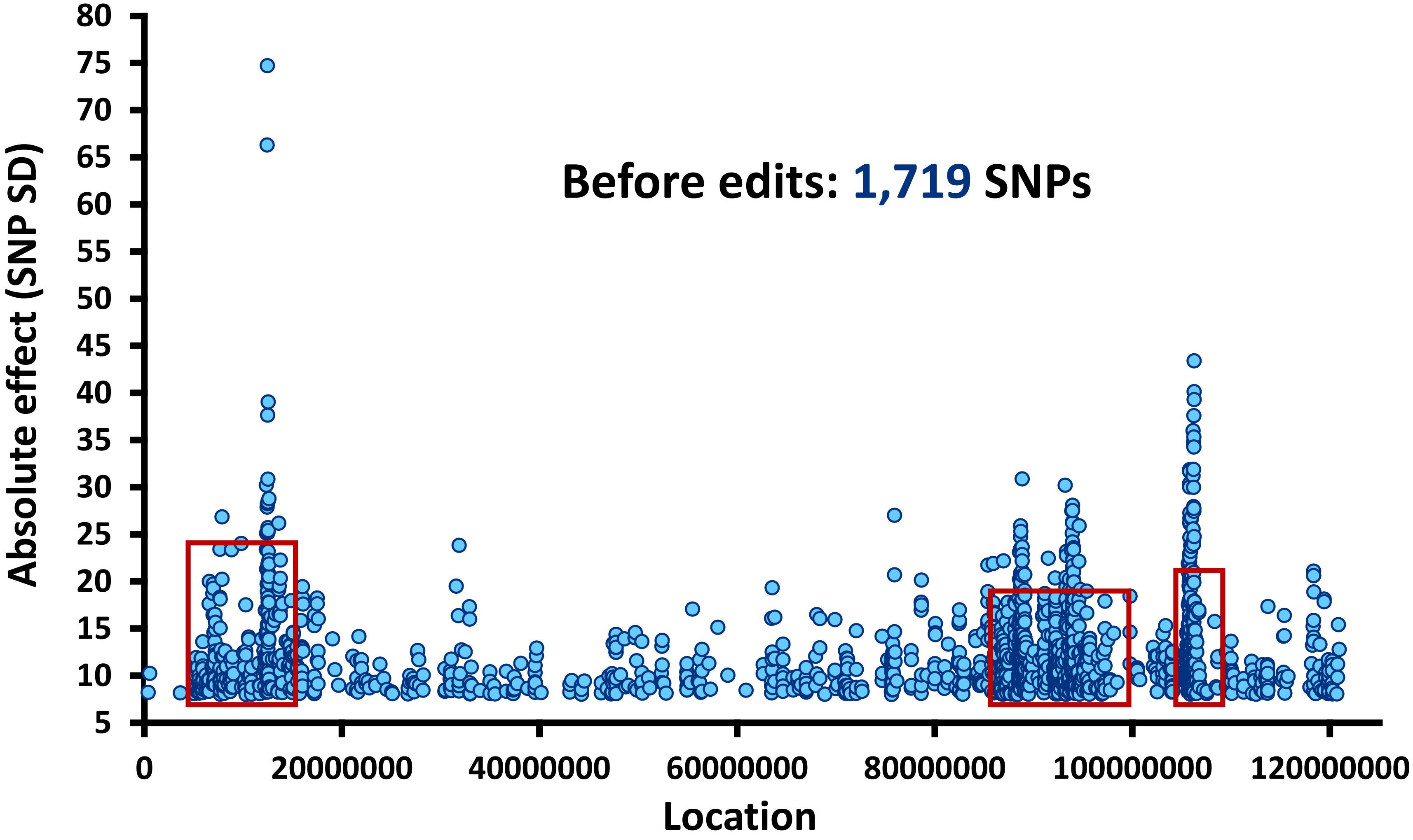


Selecting the best SEQ variants

- Developing field – no “gold” standard as to best way to select variants
- 77K chip
  - HD+ results used to choose 1,000 variants with large effects for each of the 33 traits
  - Reduce 33,000 to 17,000
    - SNPs near *DGAT1* and other QTL
    - 60K chip
    - Duplicate SNPs that effect multiple traits

Jeff O’Connell

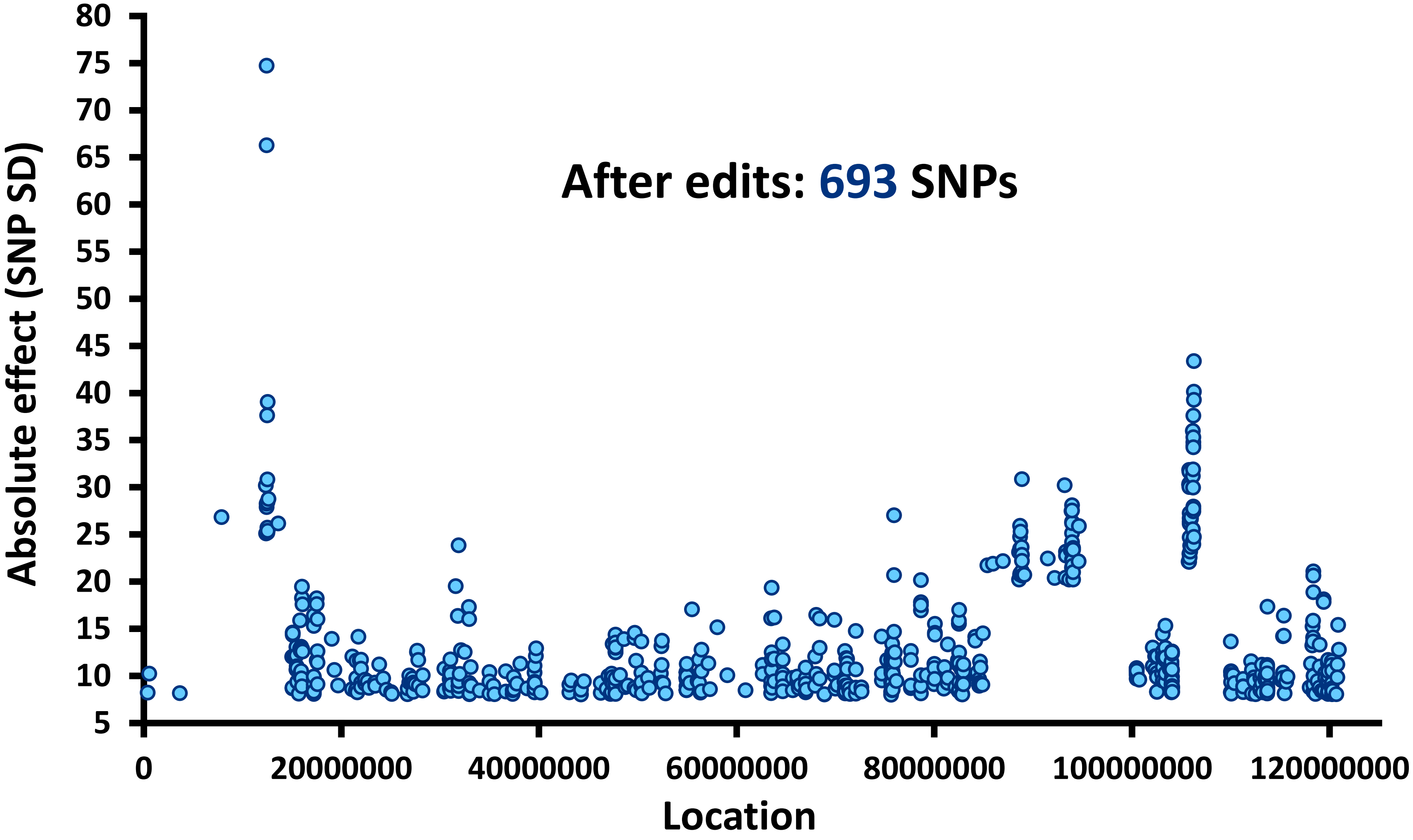
Chr 5 net merit SNP selection example



Before edits: 1,719 SNPs

Jeff O'Connell

Chr 5 net merit SNP selection example



Jeff O'Connell

Gains in REL

	HD + candidate SNPs					60K + selected		
Trait	HD only	HD + 482K	*Differ-ence*	HD + indels	*Differ-ence*	60K only	60K+ 17K	*Differ-ence*
Milk	34.1	33.9	*–0.2*	33.9	*–0.2*	34.3	35.7	*1.4*
Fat	33.7	34.0	*0.3*	33.4	*–0.3*	34.3	35.1	*0.8*
Protein	27.9	27.0	*–0.9*	26.7	*–1.2*	27.5	28.2	*0.7*
Fat %	49.2	52.7	*3.5*	52.4	*3.2*	52.9	54.8	*1.9*
Protein %	42.1	41.6	*0.5*	43.0	*0.9*	41.6	44.3	*2.7*

Jeff O'Connell

Gains in REL *(continued)*

	HD + candidate SNPs					60K + selected		
Trait	HD only	HD + 482K	*Differ-ence*	HD + indels	*Differ-ence*	60K only	60K+ 17K	*Differ-ence*
PL	36.1	33.9	*–0.3*	36.4	*0.3*	35.6	38.2	*2.6*
SCS	35.9	34.0	*0.2*	37.1	*1.2*	35.1	37.0	*1.9*
DPR	30.8	27.0	*–0.8*	31.2	*0.4*	29.0	33.0	*4.0*
CCR	28.7	52.7	*–0.6*	28.8	*0.1*	28.9	31.8	*2.9*
HCR	19.0	41.6	*1.3*	19.7	*0.7*	20.5	21.5	*1.0*

Gains in REL *(continued)*

	HD + candidate SNPs					60K + selected		
Trait	HD only	HD + 482K	*Difference*	HD + indels	*Difference*	60K only	60K+ 17K	*Difference*
Final score	24.7	25.5	*0.8*	25.8	*1.1*	24.6	27.8	*3.2*
Stature	30.4	32.4	*2.0*	32.8	*2.4*	30.3	34.7	*4.3*
Strength	29.9	31.8	*1.9*	31.8	*1.9*	29.9	34.5	*4.6*
Dairy form	33.8	35.3	*1.5*	35.8	*2.0*	35.0	38.2	*3.2*
Net merit	23.8	24.3	*0.5*	24.4	*0.6*	23.4	24.7	*1.3*

Overall gains in REL

	HD + candidate SNPs		
Trait group	HD + 482K	HD + indels	60K + 17K
Production	0.6	0.5	1.5
Health	–0.1	0.5	2.5
Calving	–0.6	–1.8	3.3
Type	1.0	0.8	3.2
All traits	0.6	0.5	2.7

Jeff O'Connell

Summary

- 39M sequenced genotypes from 444 Holsteins edited to 6M
- Imputed 6M to 26,970 reference bulls then edited to 3M
- Added gene-centric loci to HD chip to create HD+ and HD+I
- Estimated effect sizes using 2012 data
- Selected 17K SNPs to add to 60K
- Compared HD+ and HDII to HD and 77K to 60K using 2016 data

Jeff O'Connell

Summary

- HD+ and HD+I candidate approach
  - Negative REL differences
  - Prior variance spread thinner
  - Indels have less accurate calls
- 77K chip selection approach
  - Difference in REL always positive
  - Average REL gain of 2.7 percentage points across traits
  - Best performance

Sequence data – the future?

- The 1000 Bulls Genome Project run5 – 1500 bulls
  - Unfiltered data on 70M variants available
- The 1000 Bulls Genome Project run6 – 3000 bulls
  - Number of Holsteins?
  - Release date?
- Additional independent SEQ projects underway
- Better reference assembly
- Resources to collect data and generate independent call sets

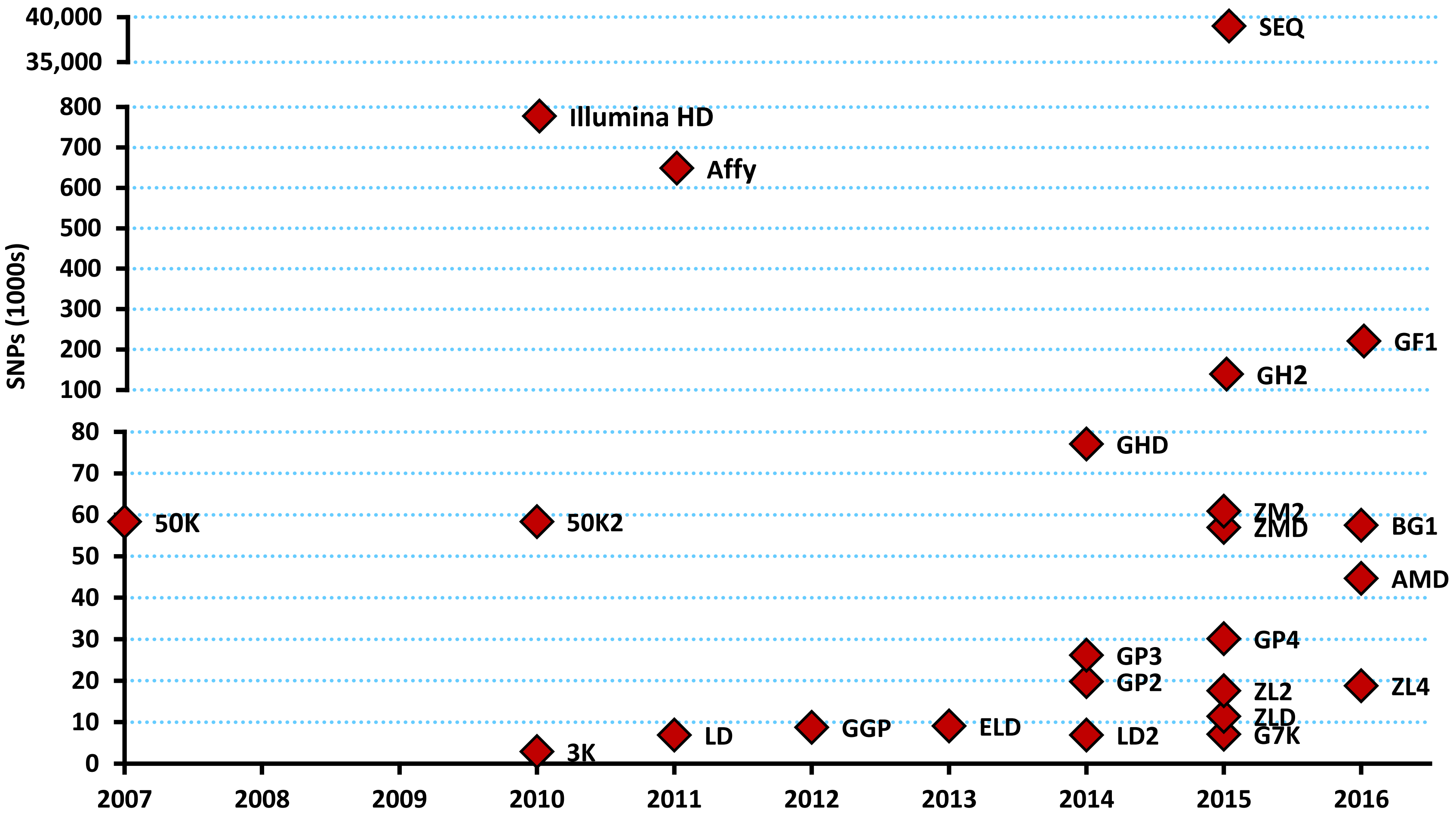
Jeff O'Connell

SNP selection – the future?

- Fixed or variable number for each trait
- GWA, multiple regression or other methods to estimate effect size
- Bioinformatics
  - Gene expression, proteomics, methylation, chromatin structure to find [e,m,me,p]QTLs
  - Prioritize non-genic SNPs and SNPs in LD groups
- Functional data
  - Difficult and expensive to go from correlation to causality

Jeff O'Connell

Integration of SNP selection into genomics



Jeff O'Connell

Integration of SNP selection into genomics

- Different chips designers will choose different SNPs
- Low density chips will not able to include all SNPs
- SEQ SNPs may not perform on chip
- Need sufficient number of chips to power imputation
- Timing of sequencing, SNP selection and chip design
- Evaluating performance of SNPs for future designs

Jeff O'Connell

Wrapping it up

- Acknowledgments
  - JRO supported by USDA SCA 58-45-14-070-1
- Slides available at https://aipl.arsusda.gov/publish/present.htm
- Stay tuned for updates next year!