Agriculture and Agri-Food Canada | Agriculture et Agroalimentaire Canada

Genotype-by-sequence in oat

Nick Tinker, AAFC Ottawa

Canada

Acknowledgements

Coauthors:

Jesse Poland

Eric Jackson

Shiaoman Chao

Gerard Lazo

Becky Oliver

Tinker Lab:

Charlene Wight

Kyle Gardner

Phil Couroux

Jiro Hattori

Biniam Hizbai

Benazir Marquez

Xiaomei Luo

Arsh Singh

CORE:

Rick Jellen, Marty Carson, Howard Rines, Don Obert, Joe Lutz, Irene Shackelford, Abraham Korol, Aaron Beattie, Åsmund Bjørnstad, Mike Bonman, Jean-Luc Jannink, Mark Sorrells, Gina Brown-Guedira, Jennifer Mitchell Fetch, Steve Harrison, Catherine Howarth, Amir Ibrahim, Fred Kolb, Mike McMullen, Paul Murphy, Herb Ohm, Brian Rossnagel, Weikai Yan, Kelci Miclaus, Jordan Hiller, Jeff Maughan, Rachel Redman, Joe Anderson, Emir Islamovic …. And others

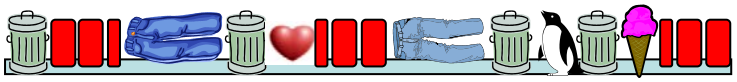
USDA

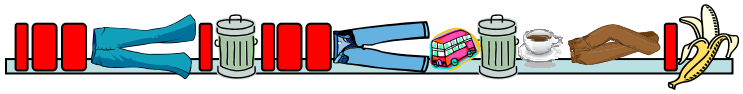
North American Millers' Association NAMA

Prairie Oat Growers Association

Agriculture and Agri-Food Canada

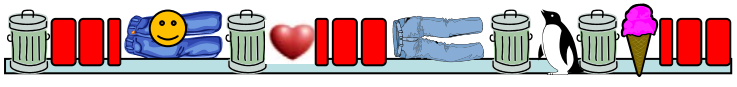
Imagine ….. the complete oat genome!

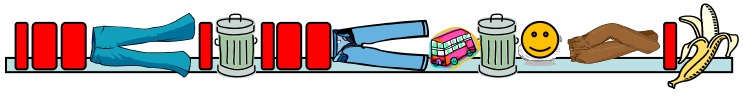


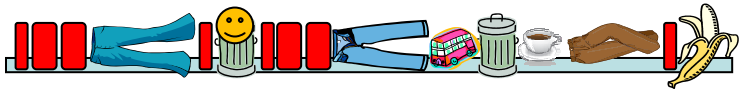


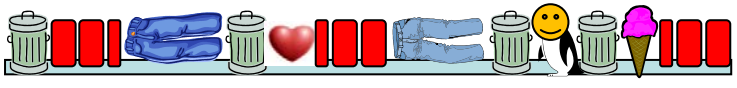
← (this genome stretches across China) →

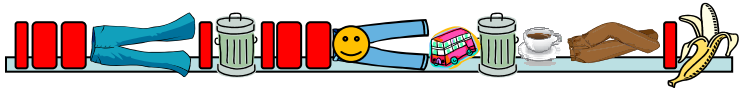
Imagine ….. every complete oat genome !







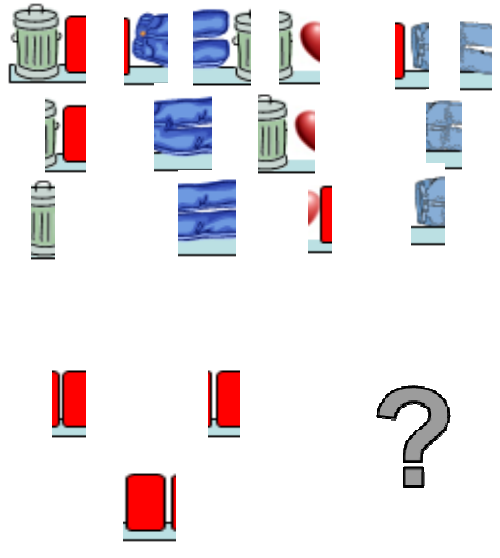




State-of-the-art DNA sequencing

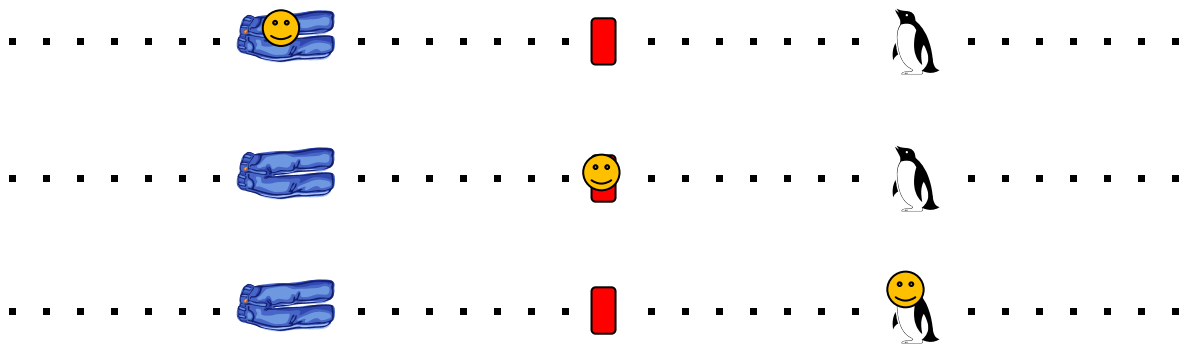
- Make 100's of copies
- Put them through a shredder
- Try to put them back together

e.g. 150x coverage
= 20 billion pieces
x 100 bp
= $100 K



Until then….

- Focus on differences among varieties
  - That's what we care about the most
- Order differences by linkage (count recombinations)
- Associate with phenotype (also by linkage)
  - Mapping populations (2-parents, lots of kids)
  - 'Natural' populations (unknown family structure)



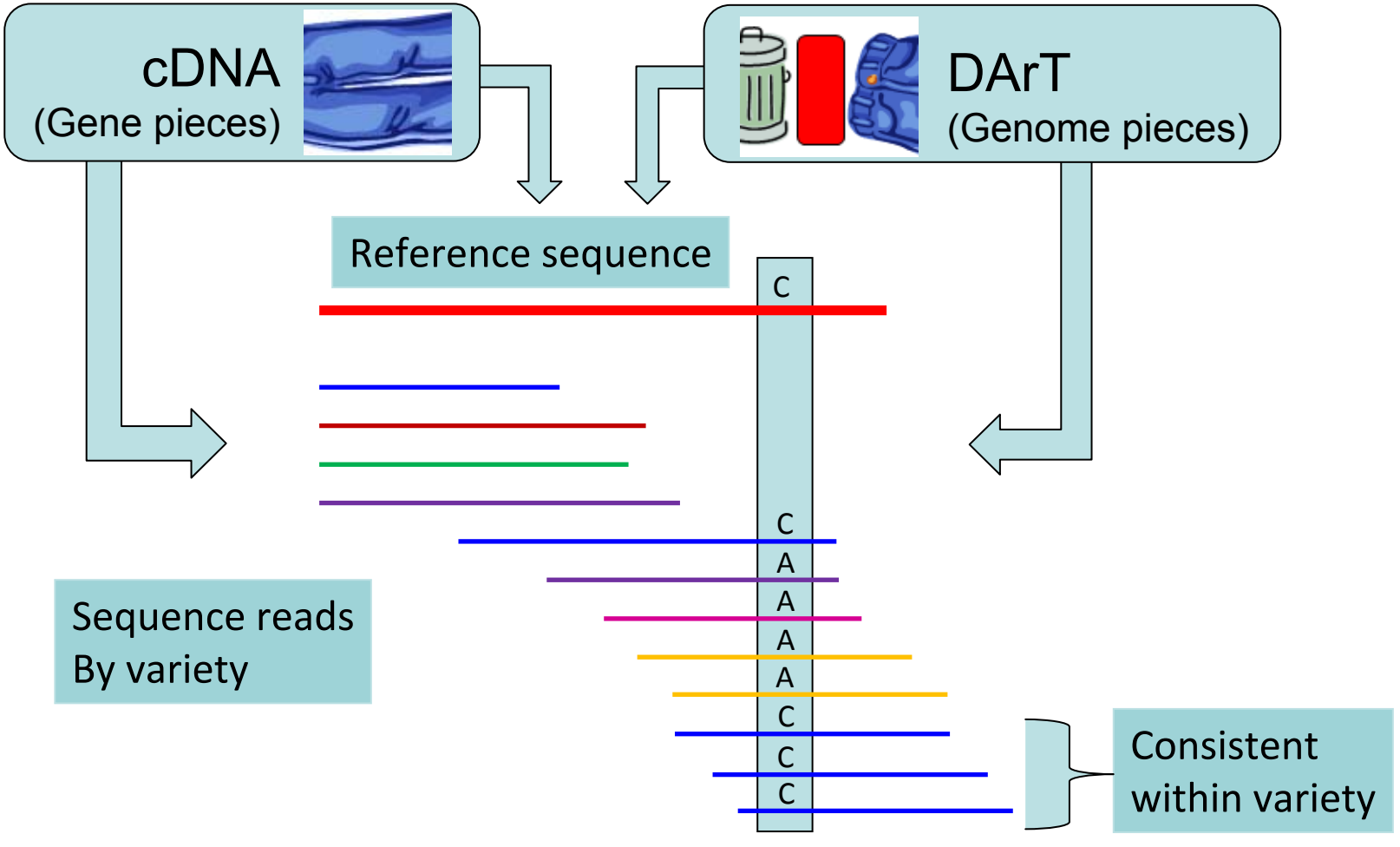
Single Nucleotide Polymorphism (SNP)

- The most common genetic difference

```
    GTACCATGATCGCTAACTGGCATGGCTTACGGCTTGAC
(A) .......T..........G....................
(B) .......G..........G....................
(C) .......G..........A....................
(D) .......T..........A....................
(E) .......G..........G....................
```

- SNP = SNP …. no matter how you find it !
  - "Old" non-sequence-based methods (AFLP, DArT)
  - Discover by sequence / assay by design
  - Discover and assay by sequencing (GBS)

SNP – discover by sequence, assay by design



CORE – Illumina SNP array

cDNA
(Gene pieces)

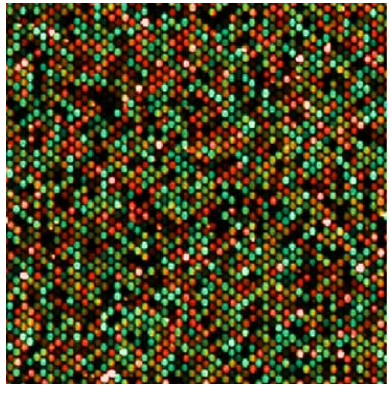
- 20 varieties
- 9 million reads
- 18,000 templates

DArT
(Genome pieces)

- 25 varieties
- 4 million reads
- 12,000 templates

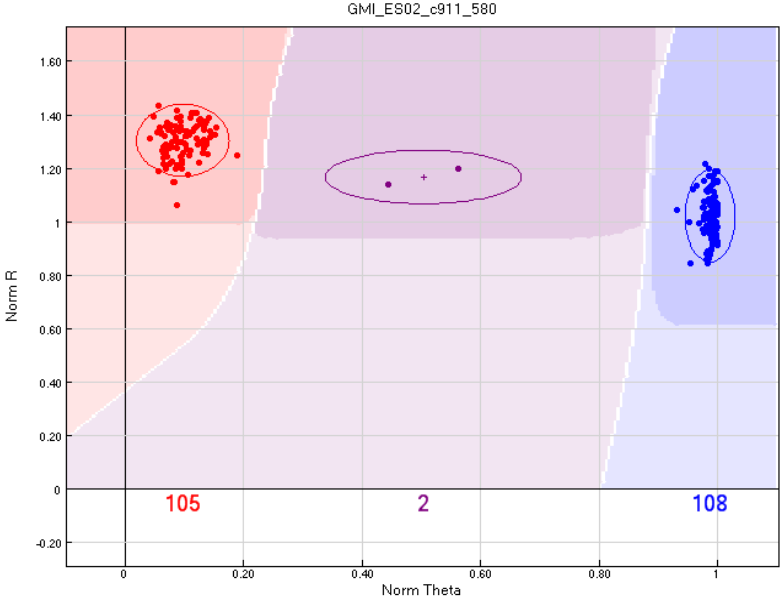
…**TGATCGCTA[G/T]CTGGCATGGCT**……

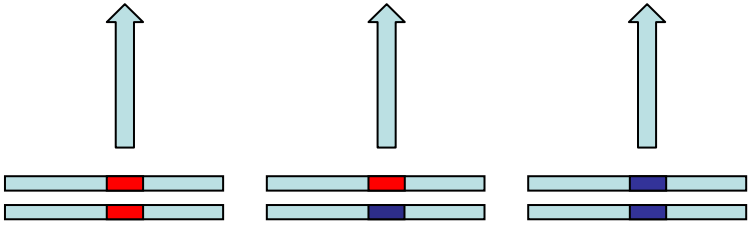
- 80,000 predicted SNPs
- 4600 tested SNPs
- 2300 validated SNPs (Golden Gate)

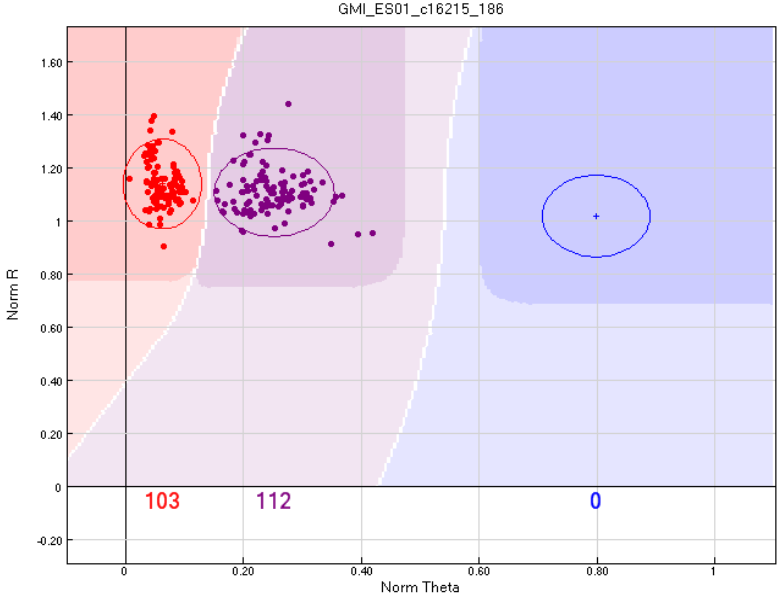


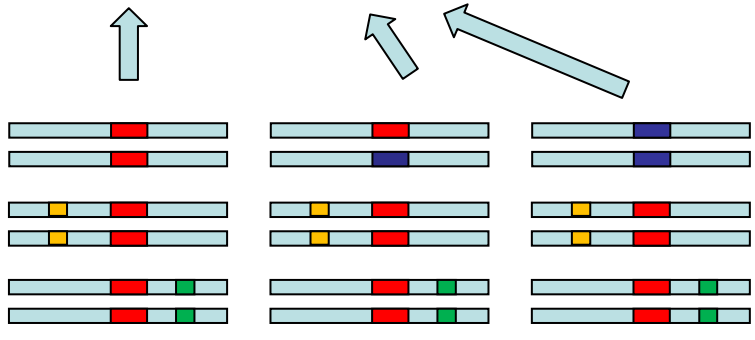
- 6000 SNPS in progress (Infinium)
- (we estimate 4000 will work)

Genome Studio Software (example SNPs)



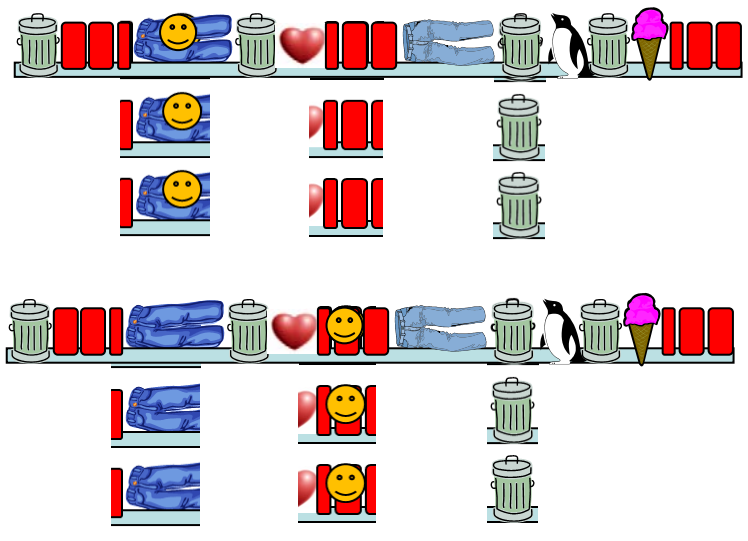






Genotype by sequence (GBS) - concept

- Discover and assay SNPs by direct sequencing
- Similar to SNP discovery for planned assay
  - But much larger numbers of and sequences
- Based on subset of genome (enzyme / amplify)



Variety A

Variety B

GBS details

OPEN ACCESS Freely available online

PLoS one

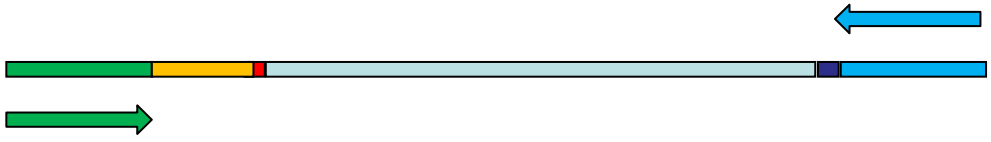
Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach

Jesse A. Poland[1,2*], Patrick J. Brown[3], Mark E. Sorrells[4], Jean-Luc Jannink[4,5]

1. Digest with *PstI* & *MspI*

2. Ligate sequencing adapters + variety-specific barcode

3. Mix together (multiplex), amplify



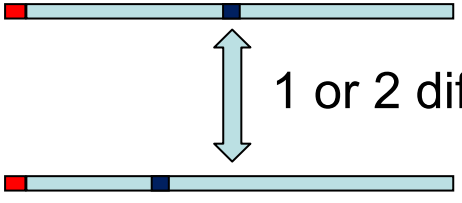
GBS - details

3 billion sequenced fragments

5. Trim barcode, trim to 64 bases, keep track of variety

6. Identify all unique tags, count in each variety (Tag x Taxa)

7. Match tag pairs, call SNPs (across full data set)

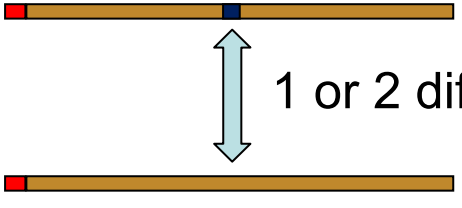


Varieties: A, C, E

Varieties: B, D, H

Is a SNP

100,000



Varieties: A, C, D, H

Varieties: B, C, D, E

Not a SNP

GBS - caveats

- Missing data 10% to 70%
  - Depends on sequencing depth (“plexity”)
  - Depends on how many SNPs you call
  - e.g. to get 95% complete, I could only call 2400 SNPs
- Bioinformatics is under development
  - Different methods give different SNP sets
- Massive storage and computing requirements
- Data sets too large for some software
  - More markers and more samples
  - Have not yet managed a consensus map

DArT *vs* SNP *vs* GBS

	DArT	SNP	GBS
Assay cost per sample	$50	$68	$20 *
Markers across all taxa	1500	4000 *	40 k → 100 k
Markers per population	300	800	4 k → 20 k
Missing data	>5 %	< 1%	10% → 50%
Co-dominant	0%	> 25%	100% *
Genes / orthology	20%	100%	5%
Duplicate loci (map inconsistently)	> 5%	< 3 %	?

Are they useful ?

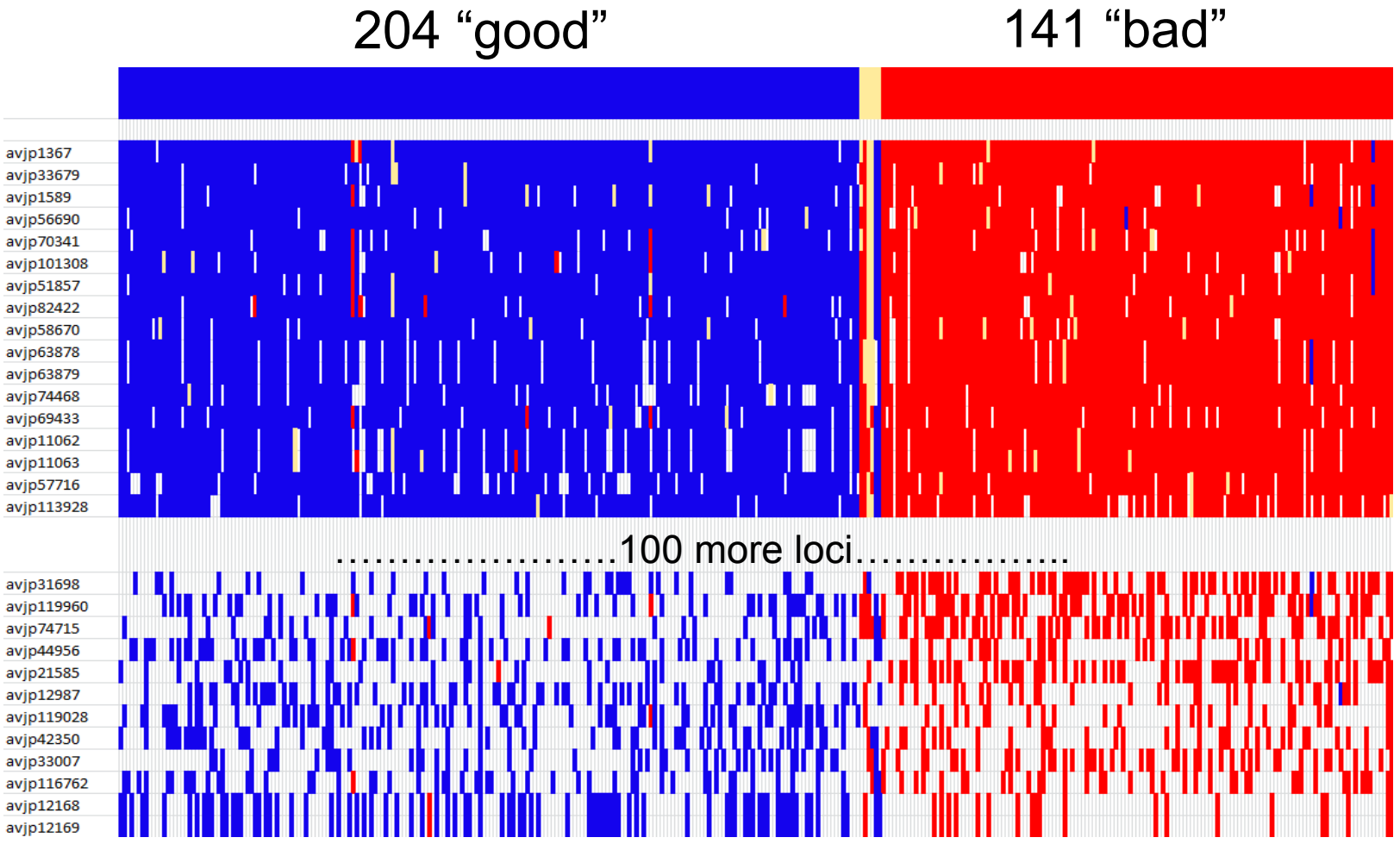
- Already have more GBS data than anything else:
  - Not just more loci..... more varieties too
  - DArT: 350 diversity + 4 bi-parental populations
  - SNP: 108 diversity + 6 bp-pop (400)
  - GBS: 738 diversity + 8 bp-pop (700) + 16 iso-lines
- 10 x more likely to find **[marker – QTL]** ?
- But missing data …
  - Mapping difficulty ?
  - Association artefacts ?
  - MAS predictions ?

Simulated MAS using GBS

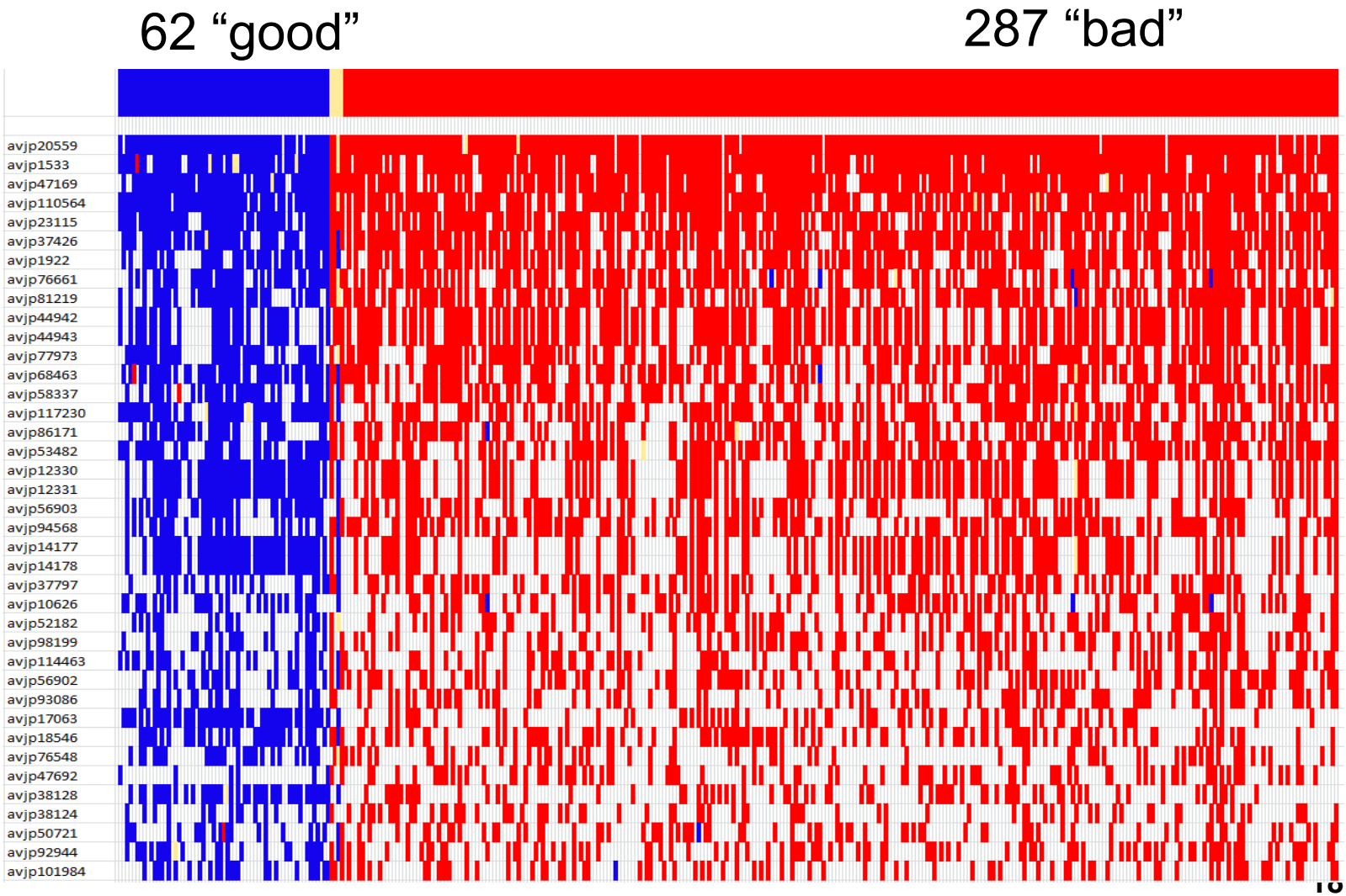
- Scenario:
  - 4 target loci, simulate with random GBS loci
  - Discover markers by association in odd # lines
  - Predict genotypes in even # lines

Marker	f(A)	f(B)	f(H)	Chr
avjp100014	568	782	20	1C
avjp100585	339	966	18	16A
avjp100261	150	1188	9	9D
avjp108144	1196	143	8	7C_17A

Predictive markers (target locus avjp100014)



Predictive markers (target locus avjp100261)



Simulated MAS using GBS

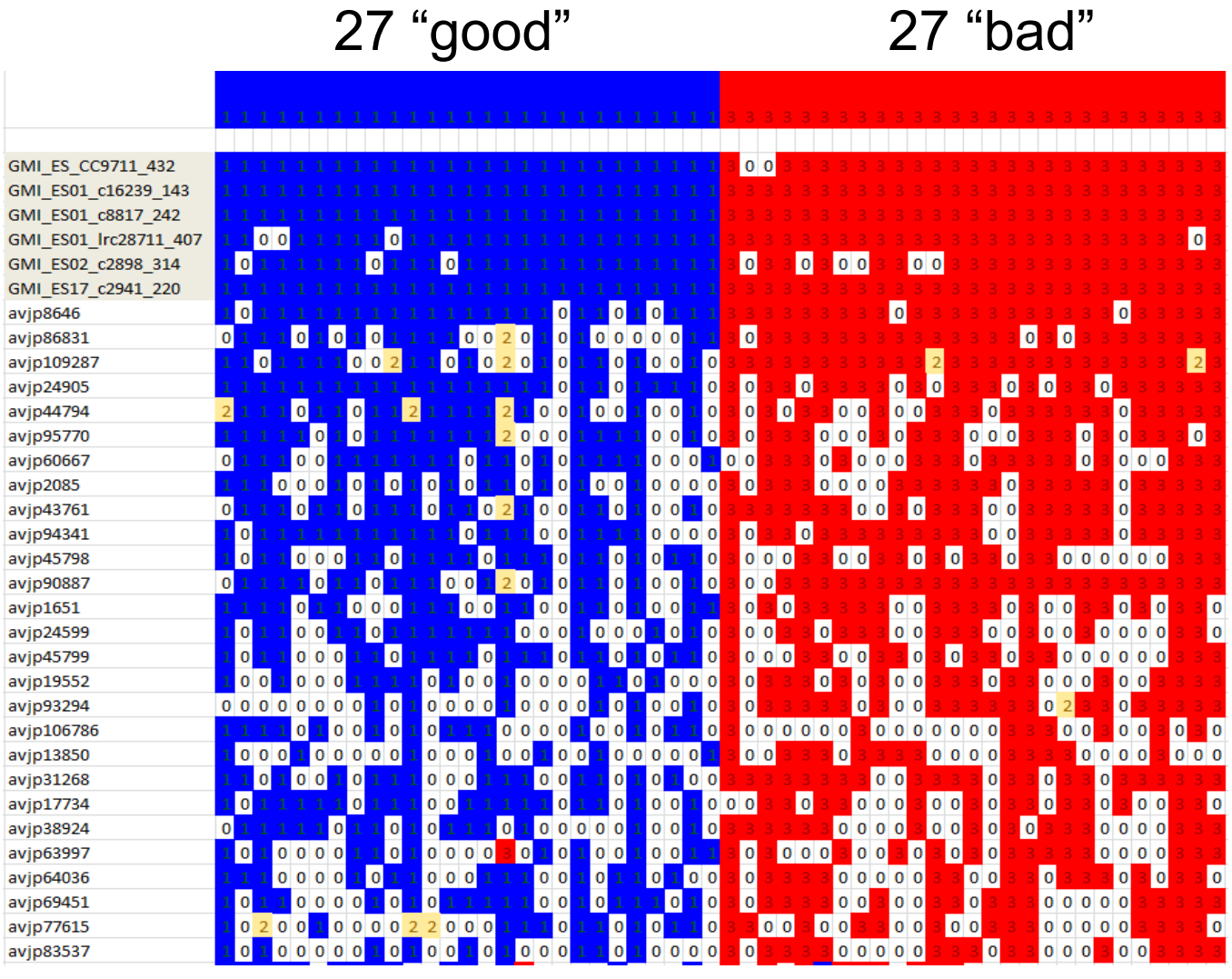
- Surprise!
  - Many loci in (near) perfect LD with a “QTL”
  - Bias in genome regions sampled by GBS ?
  - Too much good data ?
- Test predictions of array-based SNPs
  - Independently discovered
  - Most based on expressed genes
  - BUT…. SNP data only available for 108 varieties

Simulated QTLs based on Illumina SNPs

- 4 random Illumina cDNA SNPs (targets)
- Fit model with 54 odd # lines
- Test model with 54 even # lines

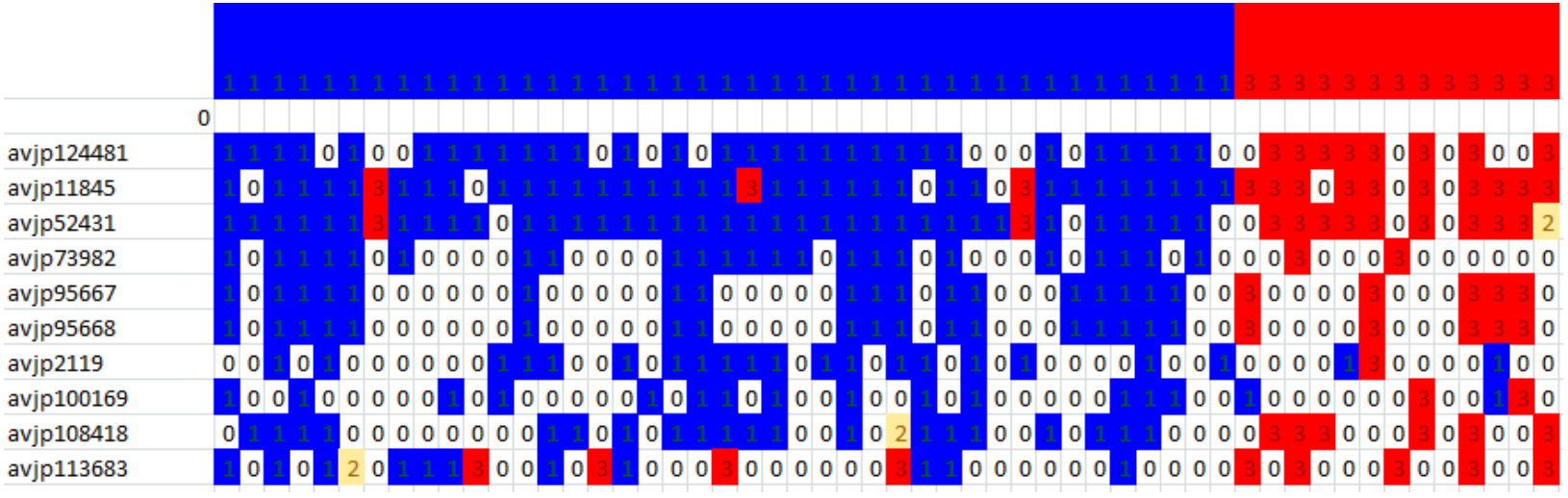
Marker	f(A)	f(B)	Chr
GMI_ES01_c10033_104	56	52	1C
GMI_ES15_c10388_464	57	51	18D
GMI_ES01_c796_180	84	24	19A
GMI_ES15_c8064_341	25	83	3C

Predictive markers (locus GMI_ES01_c10033_104)



Predictive markers (locus GMI_ES15_c8064_341)

41 “good” 13 “bad”



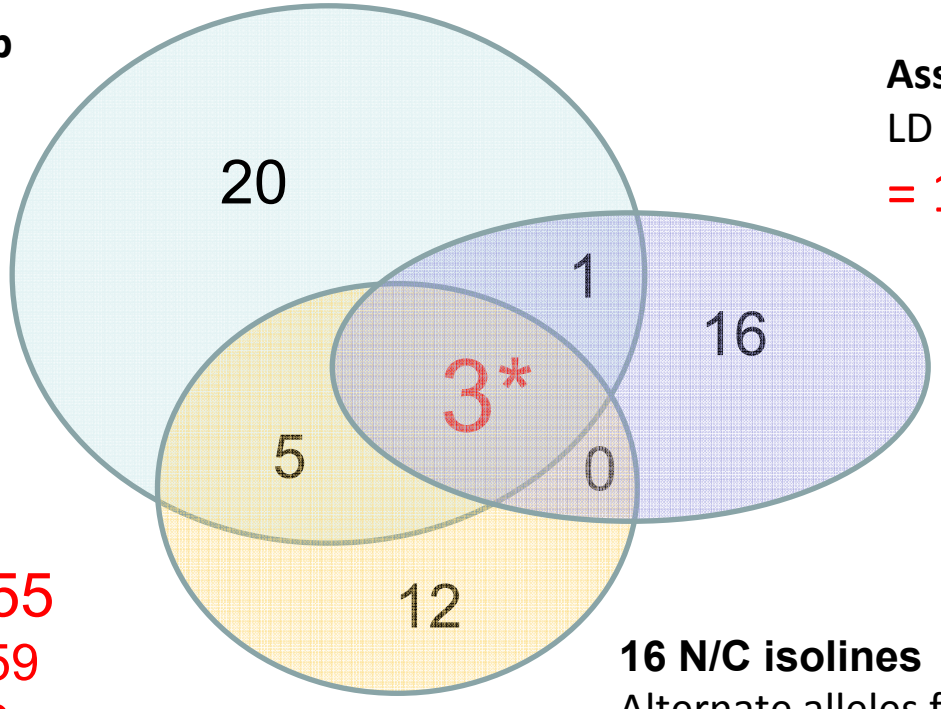
Case study with naked (hulless) oat

- Other genes (N2,N3,N4) or will we just find N1 ?
  - N1 was Mapped in Terra x Marion
  - Poorly placed by comparative mapping
- GBS data:
  - Diversity lines (*rare trait)
    - I only had phenotypes for 100 covered + 20 naked
  - 100 Terra x Marion progeny
  - 8 pairs of naked / covered iso-lines
    - OT253/Marion, from F6 heterozygotes
    - Developed by Solomon Kibite

Look at all pairwise linkages and LD's

N1 ←→ GBS-tag

TxM Map
R < 0.05
= 28



Association
LD R2 > 50%
= 18

*avjp23455
avjp123459
avjp17890

16 N/C isolines
Alternate alleles fixed
= 20

N1 predictions

- LD analysis based on partial data:
  - "training set" (the only ones I had data for at the time)
  - 20 naked / 100 covered
- Based on "avjp23455" we predicted:
  - 12 more naked lines (remaining 600 = covered)
  - All were correct!

Training set	
95Ab13050	OT253
Boudrias	Provena
Bullion	Racoon
Gehl	Salomon
IOI033	Shadow
IOI108	Terra
IOI114	VAO-44
IOI150	Vao-48
IOI191	VAO-51
Navaro	VAO--58

Predicted Naked Lines
98Ab7265
FL03184-FLID-B-S1
FL04178-FLID-B-S-2
HLA05AB1-34
IL02-10836
IL03-7936
LA02012-S-B-139-S2-B-S2-B-S2
LA0210SBSBSBSB-S1
LA03066SBS-L1
Lennon
Nudist
Zuton

Conclusions

- GBS will be exceptionally efficient for tagging genes
  - Bi-parental AND association mapping
- Excellent potential for MAS
- SNP array or custom assays are better for
  - Genome analysis, comparative genomics
  - Critical genotyping, well-characterized targets
- Work required
  - Streamline informatics
  - Build GBS/SNP based consensus map
  - Evaluate consistency of map position
  - Evaluate genomic selection

First Announcement: **web:** aowc.ca

American Oat Workers Conference

Ottawa Canada, 2014

Sun. July 13 to Wed. July 16, 2014

Ottawa Marriott Hotel
100 Kent Street, Ottawa
Ontario K1P 5R7 Canada