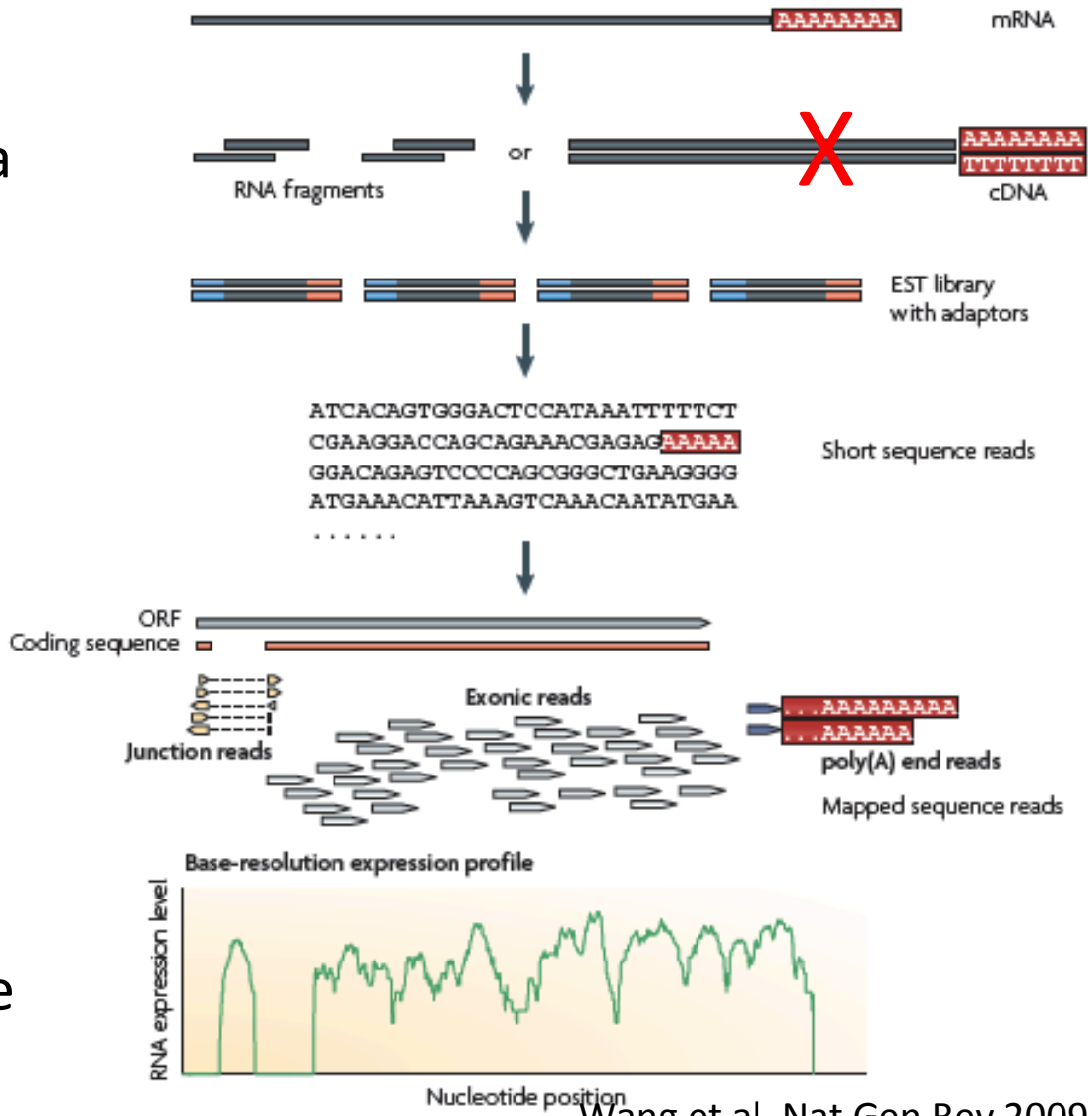
Using RNA-Seq to reveal expression & diversity in maize

C Robin Buell
Michigan State University
Department of Plant Biology
Purdue 8th International Symposium on Statistics,
June 2012

What is RNA-Seq?

- RNA sequencing (RNA-seq)

- Method to sequence RNA via DNA intermediate to:
  - Determine sequence of transcripts (proxy for genome, alternative isoform)
  - Quantitatively assess transcript abundances including allele specific expression
  - Identify variants in genomes (restricted to the transcribed regions)



Wang et al. Nat Gen Rev 2009

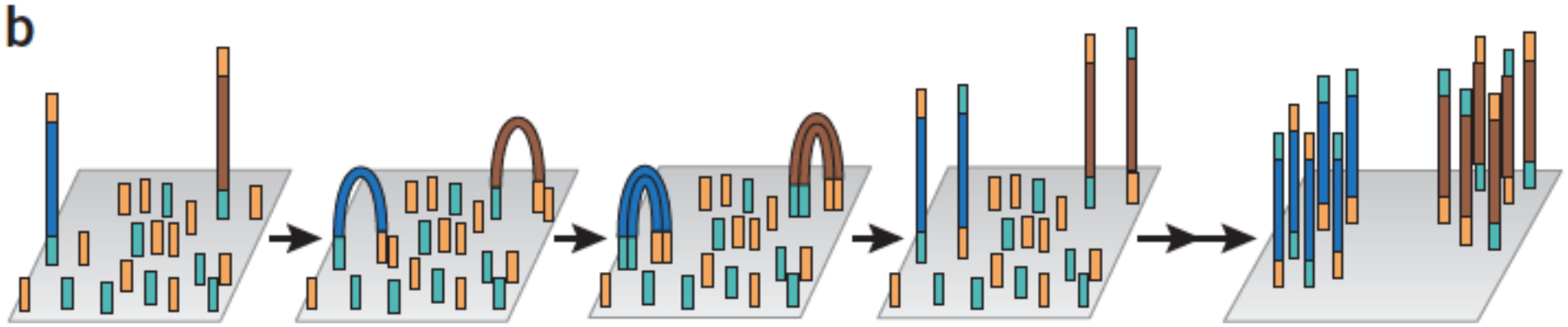
Next Generation Sequencing Platform-Illumina

-Illumina ia the predomoninant platform for next generation sequencing that is being used for RNA-seq

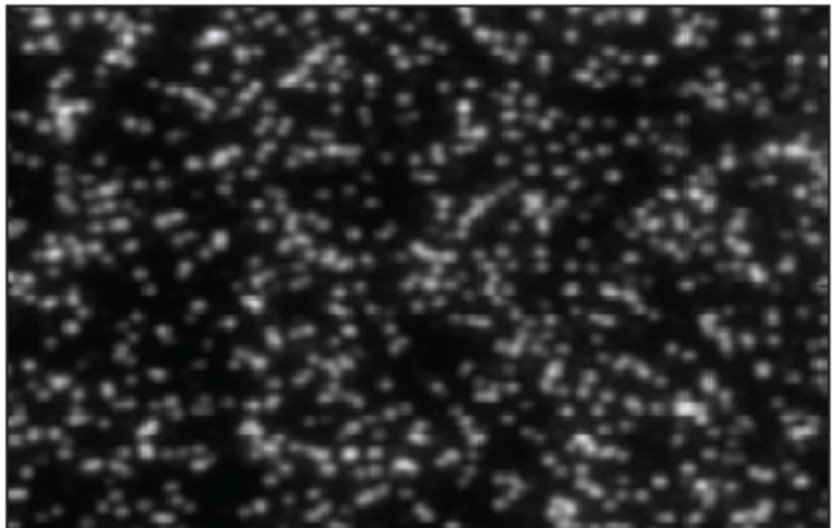
-Other platforms: SoLiD, Ion Torrent, Roche 454, Pacific Biosciences

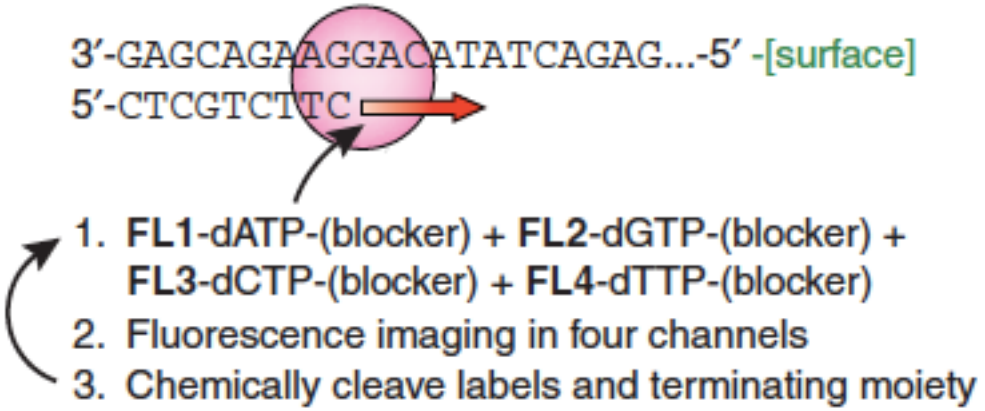
Flow cell: 8 lanes

What happens on the flow cell



Amplification of template on flow cell via bridge PCR (Shendure & Li 2008)





Sequencing-by-Synthesis using fluorescent reversible dye terminators (Shendure & Li 2008)

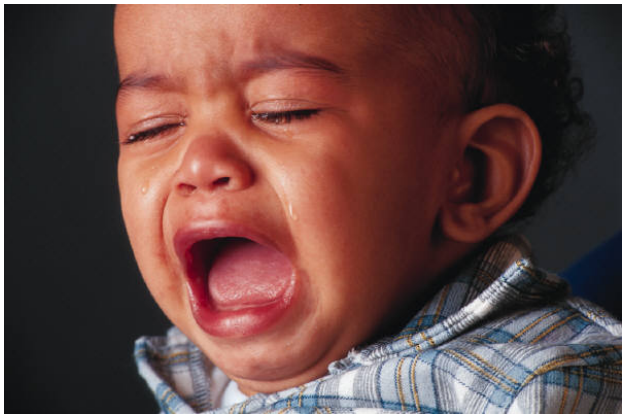
Output: The good, the bad, and the ugly

=> Get TONS of data

1 Lane of Illumina, ~250M paired end sequences, 100 bp

250,000,000 *100 bp = 25 Gb of sequence

The sequenced reads can (do) have errors.

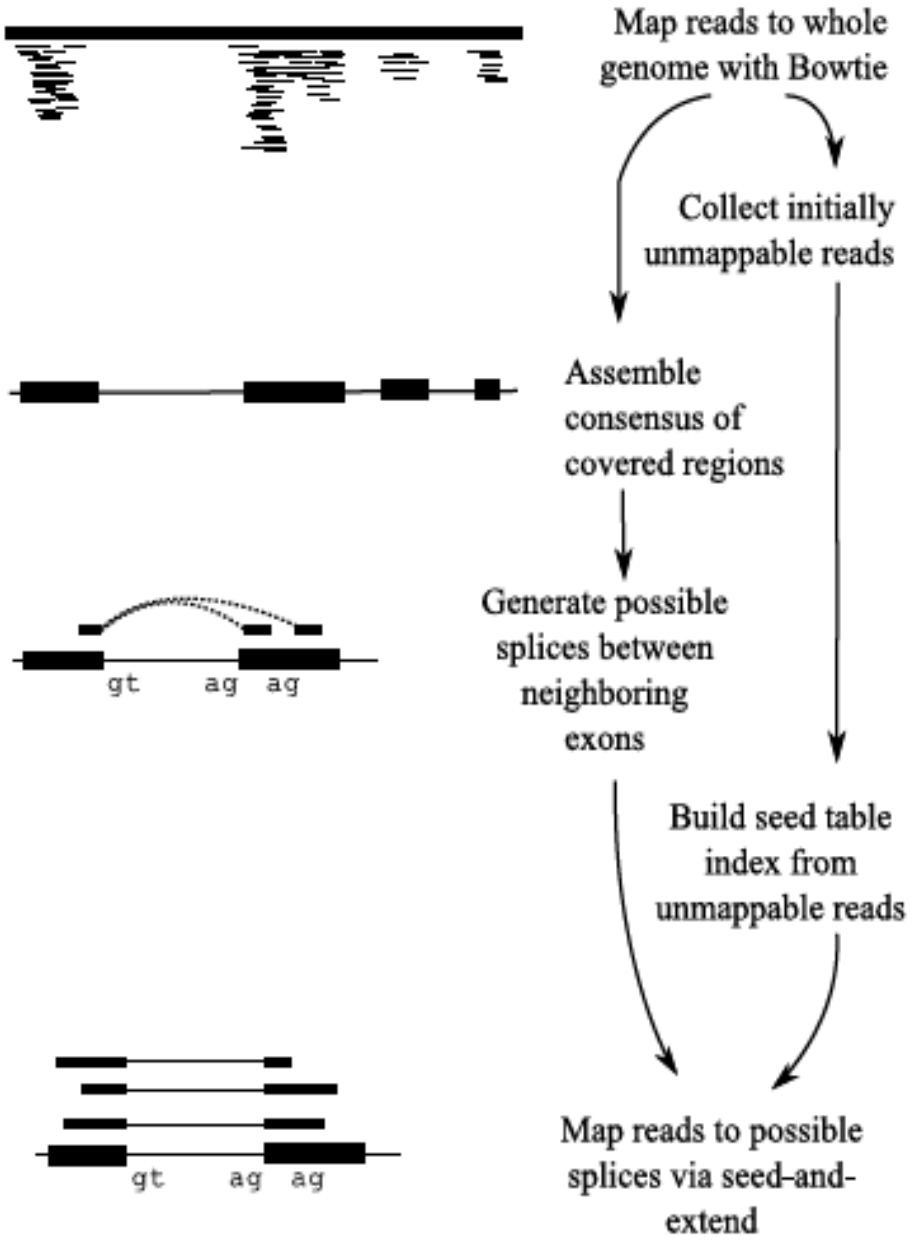


Generate more sequence to compensate
Use quality values to filter sequences
When in doubt, throw it out (probably over-sampled the library anyway)
Use “quality aware” algorithms for analysis
Assume the statistician that wrote the software knows more than you do

Workflow

-Get the reads
-Check the quality of the reads
-Clean the reads
-Map the reads to the genome
-Decide how to penalize multi-mapping reads, mismatches due to polymorphisms between the query RNA and the reference genome (i.e., Single Nucleotide Polymorphisms (SNPs))
-Quantitate the reads = expression abundances

FPKM: Fragments per kb exon model per million reads mapped (normalized for gene length and depth of sequencing in each experiment)
Older papers used **RPKM** (reads instead of fragments)



Trapnell et al. Bioinformatics 2009

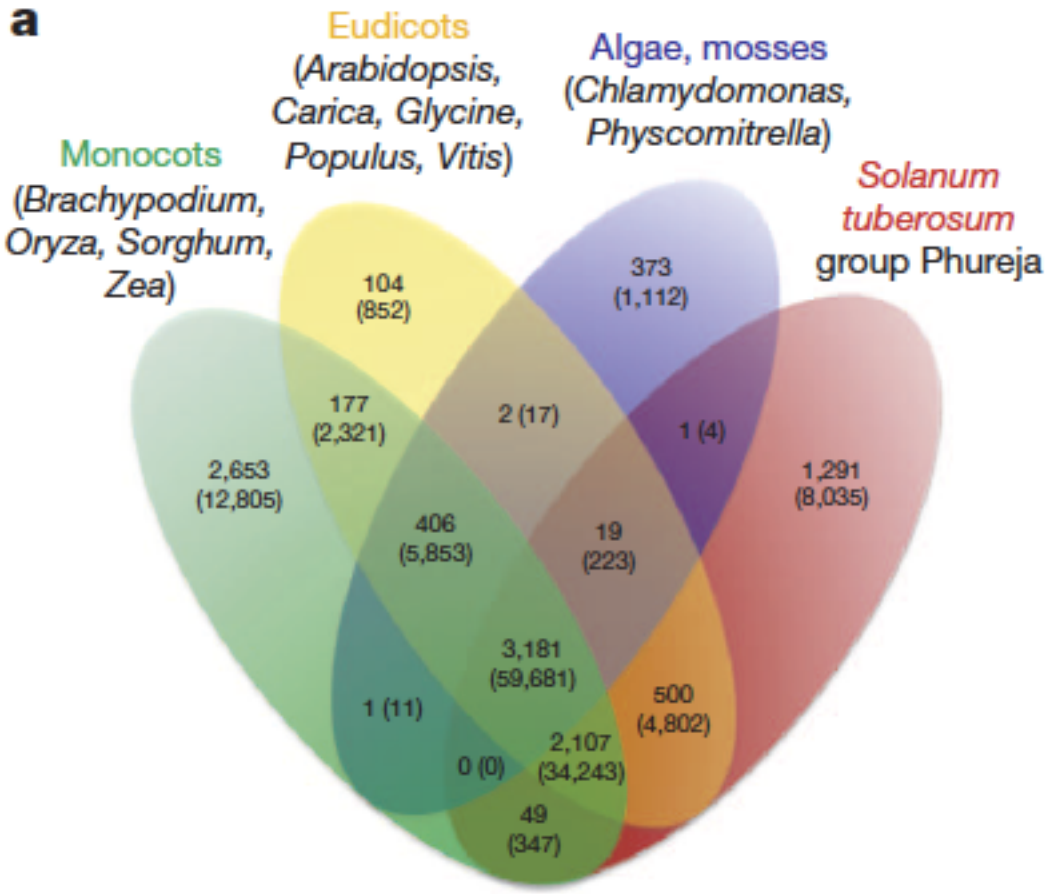
Lineage-specific genes: An enigma in all genomes

-Genome scale comparisons have revealed sets of genes restricted to specific lineages

-Lineage specific genes may be key to phenotypic differences between taxa

-An overwhelming majority of these genes have no known function

-Improve our understanding of the function of lineage specific genes through additional annotation in the form of expression data



Comparison of gene complements in 12 plant species

Potato Genome Sequencing Consortium, Nature 2011

Maize Reproductive Organs

Male: Tassels (top of plant: pollen, anthers)

Female: Ears (cob, silk, ovule)

Seed: Fertilized ovule (seed: whole seed, embryo, endosperm)

Non-reproductive: Leaf (vegetative)

http://www.sciencekids.co.nz/pictures/plants/maizeplantdiagram.html

RNA-Seq Data

Tissue	Number of Purity Filtered Reads*	Number of Mapped Reads	Max FPKM†	Number of Expressed Genes‡	Number of Tissue Restricted Genes§
Leaves	17.1M	14.5M	128,735	21,956	492
Pre-emergence cob	18.3M	16.0M	17,256	23,338	70
Post-emergence cob	18.3M	15.8M	17,784	23,948	44
Silk	18.9M	16.8M	257,614	23,015	138
Ovule	29.9M	26.3M	33,621	25,003	108
Pre-emergence tassel	19.8M	17.0M	24,576	25,165	88
Post-emergence tassel	18.6M	16.1M	35,766	24,984	42
Whole Anthers	27.0M	24.0M	127,465	22,178	96
Pollen	27.9M	25.3M	1,004,070	13,418	206
Seed 5 DAP	19.2M	16.5M	22,966	24,390	37
Seed 10 DAP	26.4M	22.9M	26,927	24,486	105
Embryo 25 DAP	19.9M	17.0M	36,127	22,493	188
Endosperm 25 DAP	23.6M	20.3M	283,698	22,887	251

Davidson et al. Plant Genome 2011

RNA-Seq Data

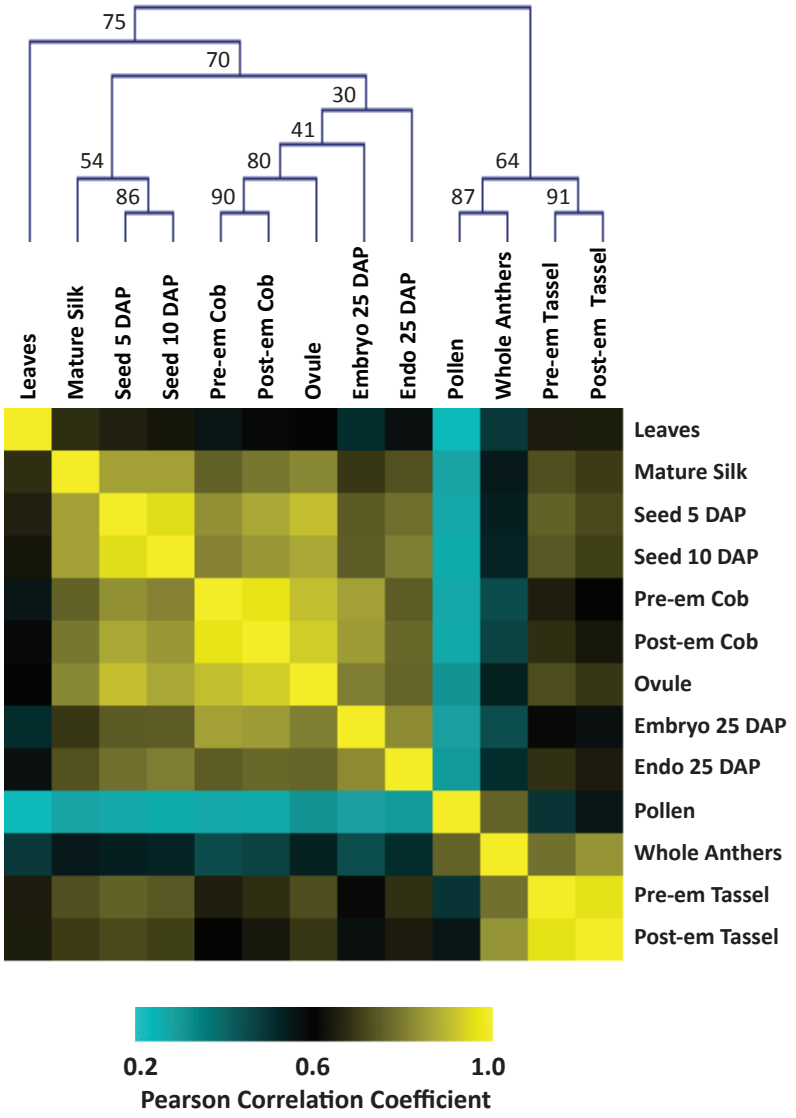
Tissue	Number of Purity Filtered Reads*	Number of Mapped Reads	Max FPKM†	Number of Expressed Genes‡	Number of Tissue Restricted Genes§
Leaves	17.1M	14.5M	128,735	21,956	492
Pre-emergence cob	18.3M	16.0M	17,256	23,338	70
Post-emergence cob	18.3M	15.8M	17,784	23,948	44
Silk	18.9M			015	138
Ovule	29.9M			003	108
Pre-emergence tassel	19.8M			165	88
Post-emergence tassel	18.6M	16.1M	35,766	24,984	42
Whole Anthers	27.0M	24.0M	127,465	22,178	96
Pollen	27.9M	25.3M	1,004,070	13,418	206
Seed 5 DAP	19.2M	16.5M	22,966	24,390	37
Seed 10 DAP	26.4M	22.9M	26,927	24,486	105
Embryo 25 DAP	19.9M	17.0M	36,127	22,493	188
Endosperm 25 DAP	23.6M	20.3M	283,698	22,887	251

~80% of all genes expressed in these 13 tissues

Davidson et al. Plant Genome 2011

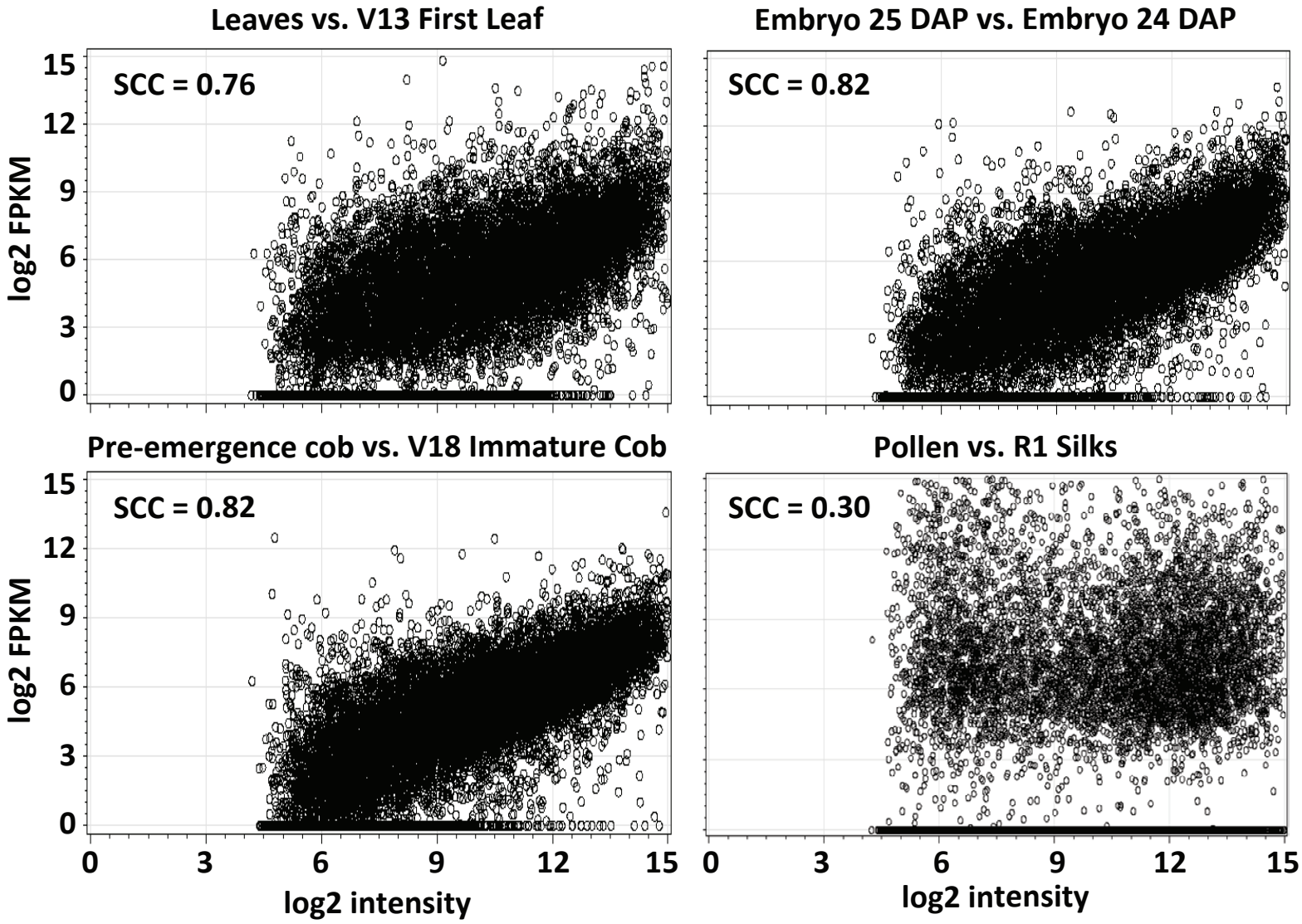
Transcriptome correlations across tissues

-Similar tissues cluster
-Pollen distinct



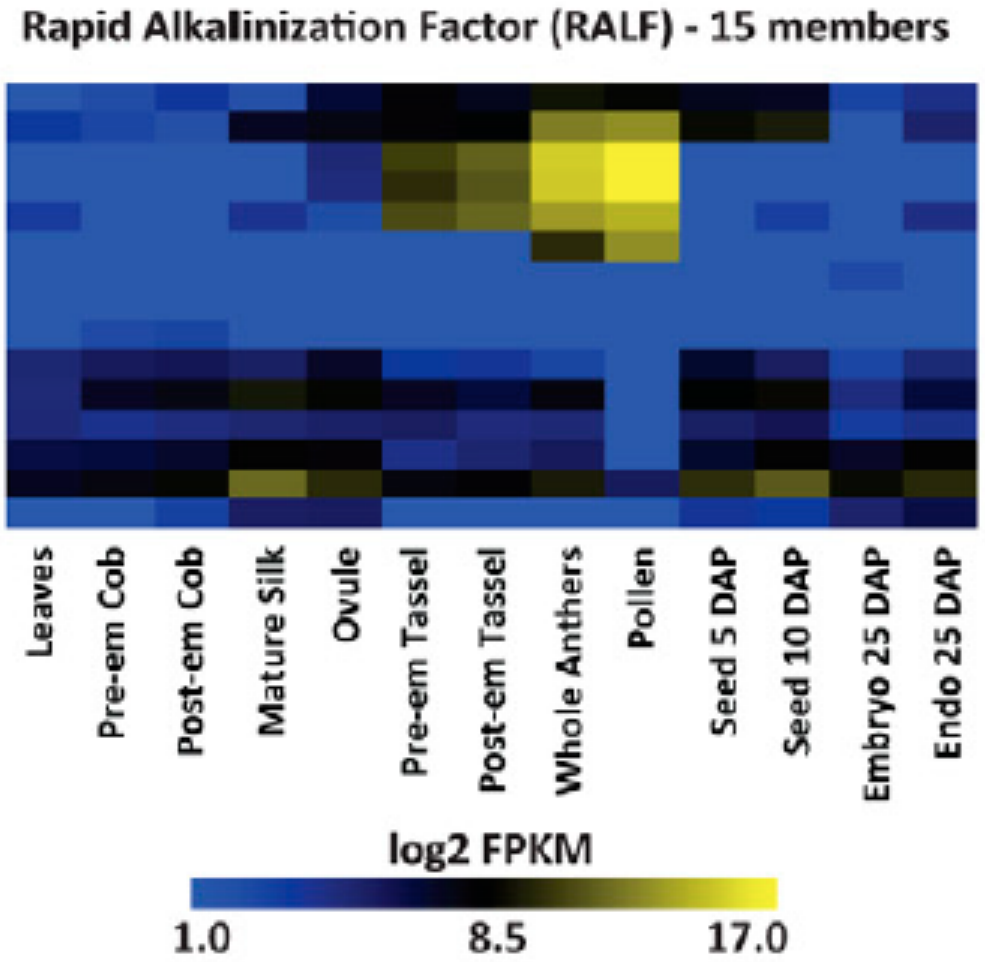
Davidson et al. Plant Genome 2011

How does RNA-seq compare to microarrays?



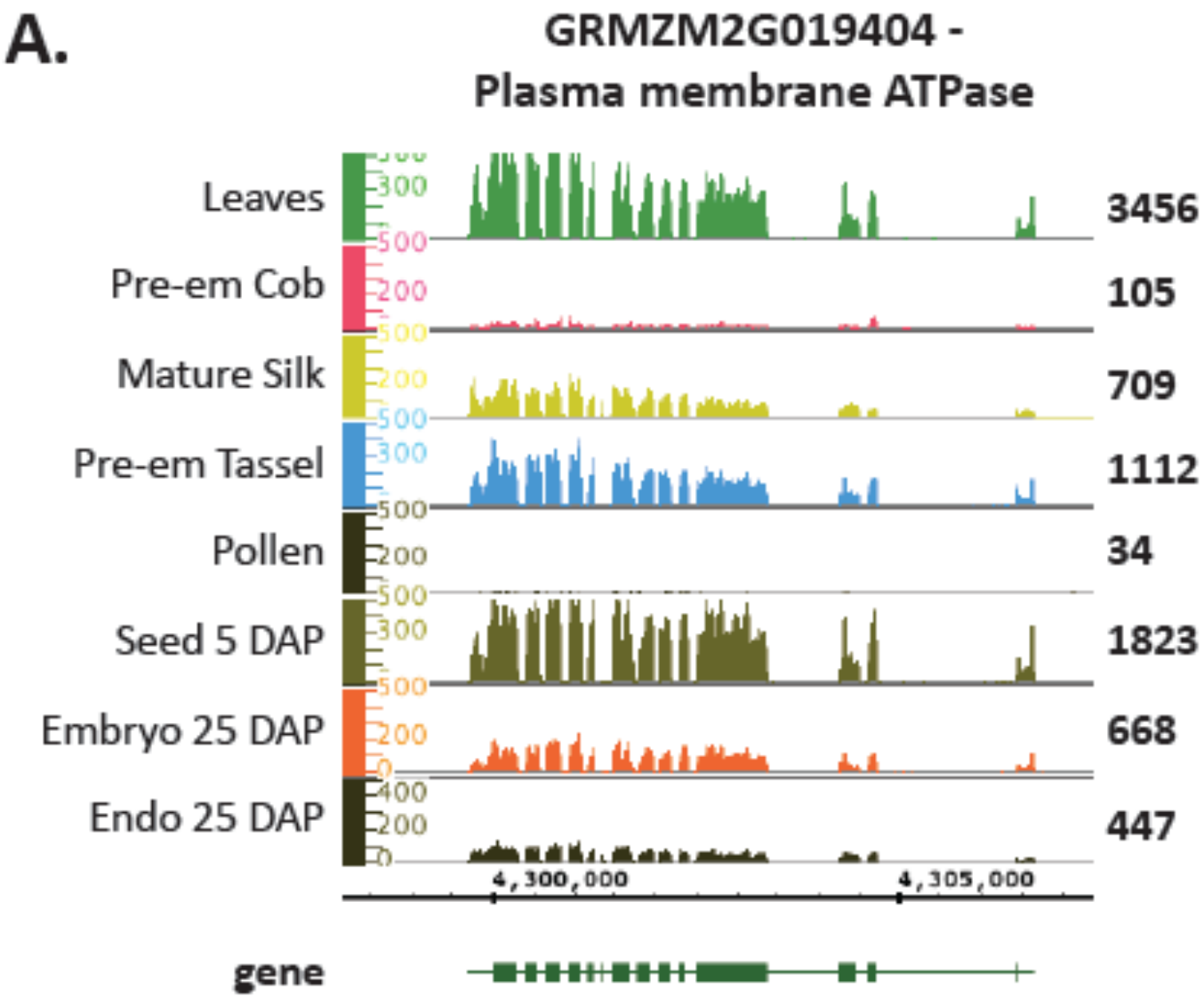
Davidson et al. Plant Genome 2011

Value of RNA-seq over microarrays



Davidson et al. Plant Genome 2011

RNA-seq resolves gene structure

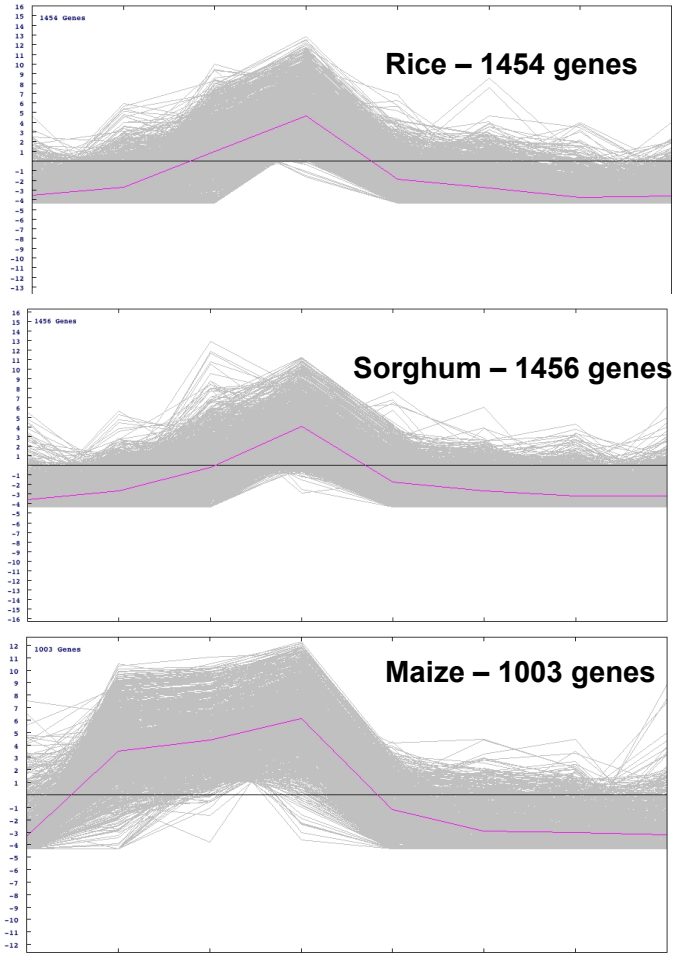


Davidson et al. Plant Genome 2011

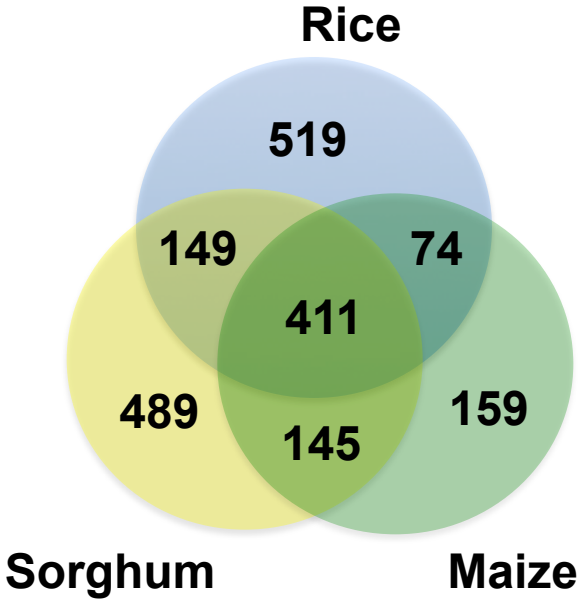
Shared Expression Patterns of Orthologous Genes

Male specific expression

Lvs Stg1 Stg2 Anth Pist Sd5 Sd10 Emb25



Orthologous Groups (protein level clustering)



k-means clustering, k = 8
8 core tissues
log2FPKM > 0.5 across libraries

Improving use of maize as a biofuel feedstock

Corn stover is an important source of lignocellulosic biomass in the short term and can be used as a model C4 grass for improvement of dedicated bioenergy grasses in the long term

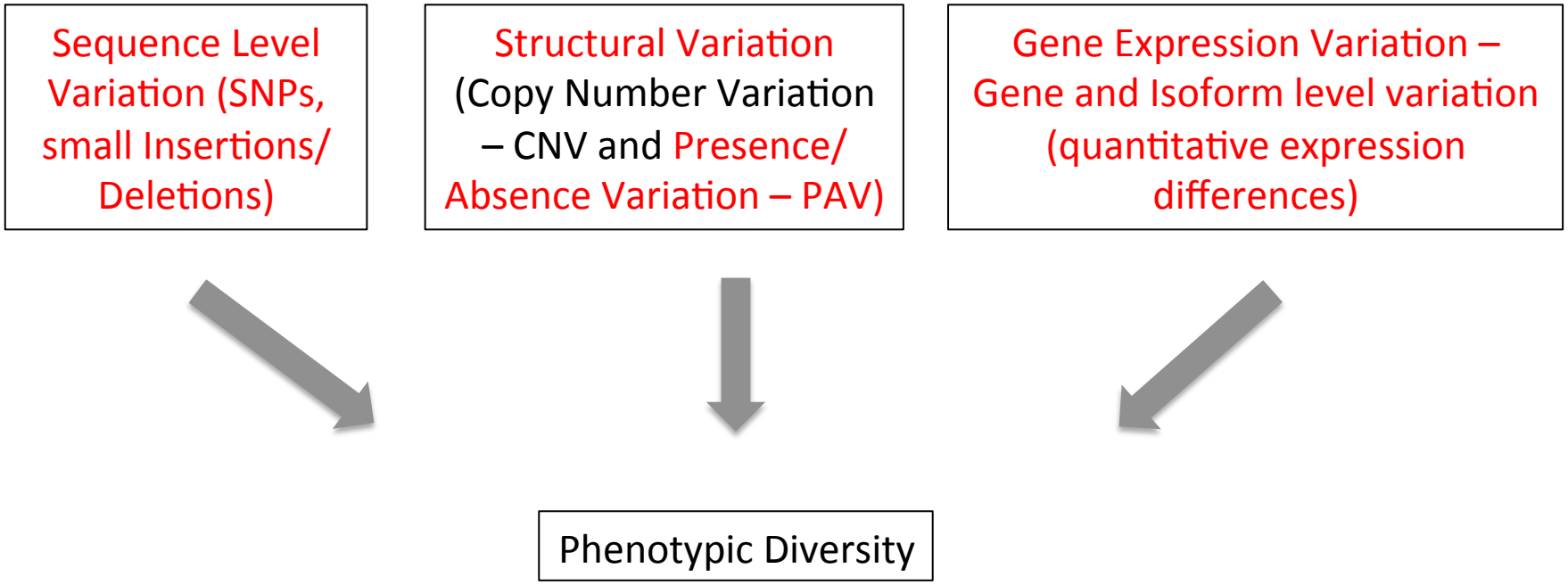
Collaboration with the Kaeppler/de Leon groups at University of Wisconsin (maize geneticists/ breeders)

Goal is to identify genes (and more specifically alleles) for improved biomass yield and composition through linkage mapping, gene expression, and linkage disequilibrium mapping

Developed diversity panel of maize inbred lines adapted to Wisconsin=> phenotype, genotype

stalks, leaves, husks, cobs, tassels

Underlying Causes of Phenotypic Diversity

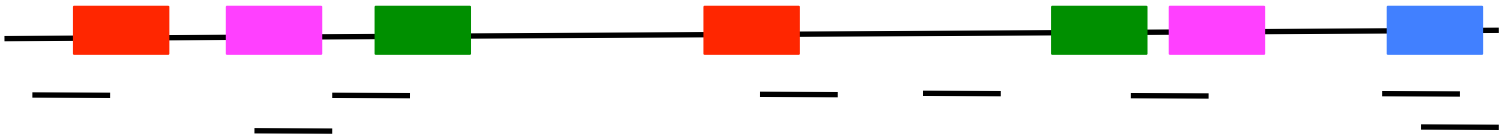


Underlying genetic variation that can be evaluated with RNA-seq

Utility of RNAseq for Variant Detection

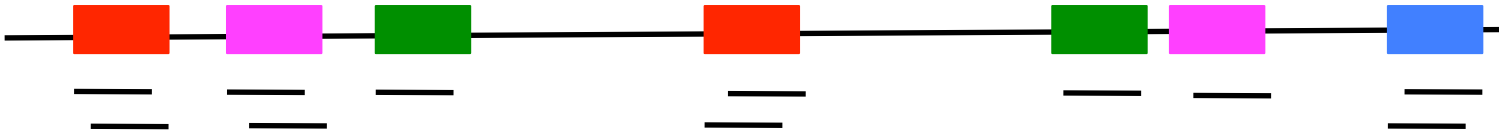
Multiple Copies of Genes
Extensive repetitive intergenic sequence

Whole Genome Sequencing



~30% of reads are informative
2.3 Gb of sequence

RNAseq

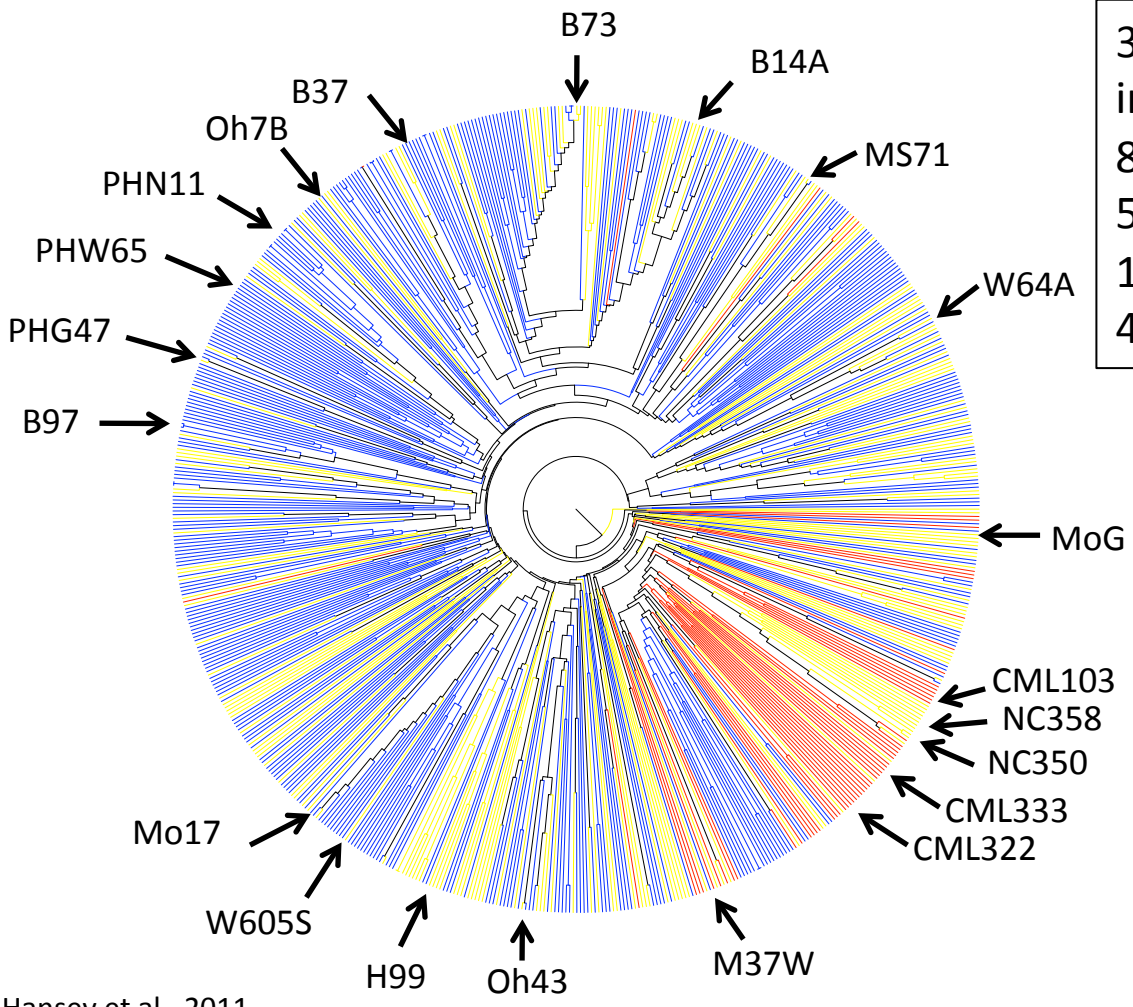


~60% of reads are informative
97 Mb of sequence

Utility of RNAseq for Variant Detection

- RNA-seq limitations
  - Genes/alleles must be expressed in the tissue used to detect variants
    - Seedling tissue has a high percentage of genes (66%) expressed (Sekhon and Lin et al., 2011)
    - Genotypes in this study are highly homozygous removing concerns of allele specific expression

The 21 Genotypes Used in the Study



Adapted from Hansey et al., 2011

- 3 Stiff Stalk Lines (SSS, including B73)
- 8 Non Stiff Stalk Lines (NSS)
- 5 Tropical Lines
- 1 Iodent Line
- 4 Unclassified Lines

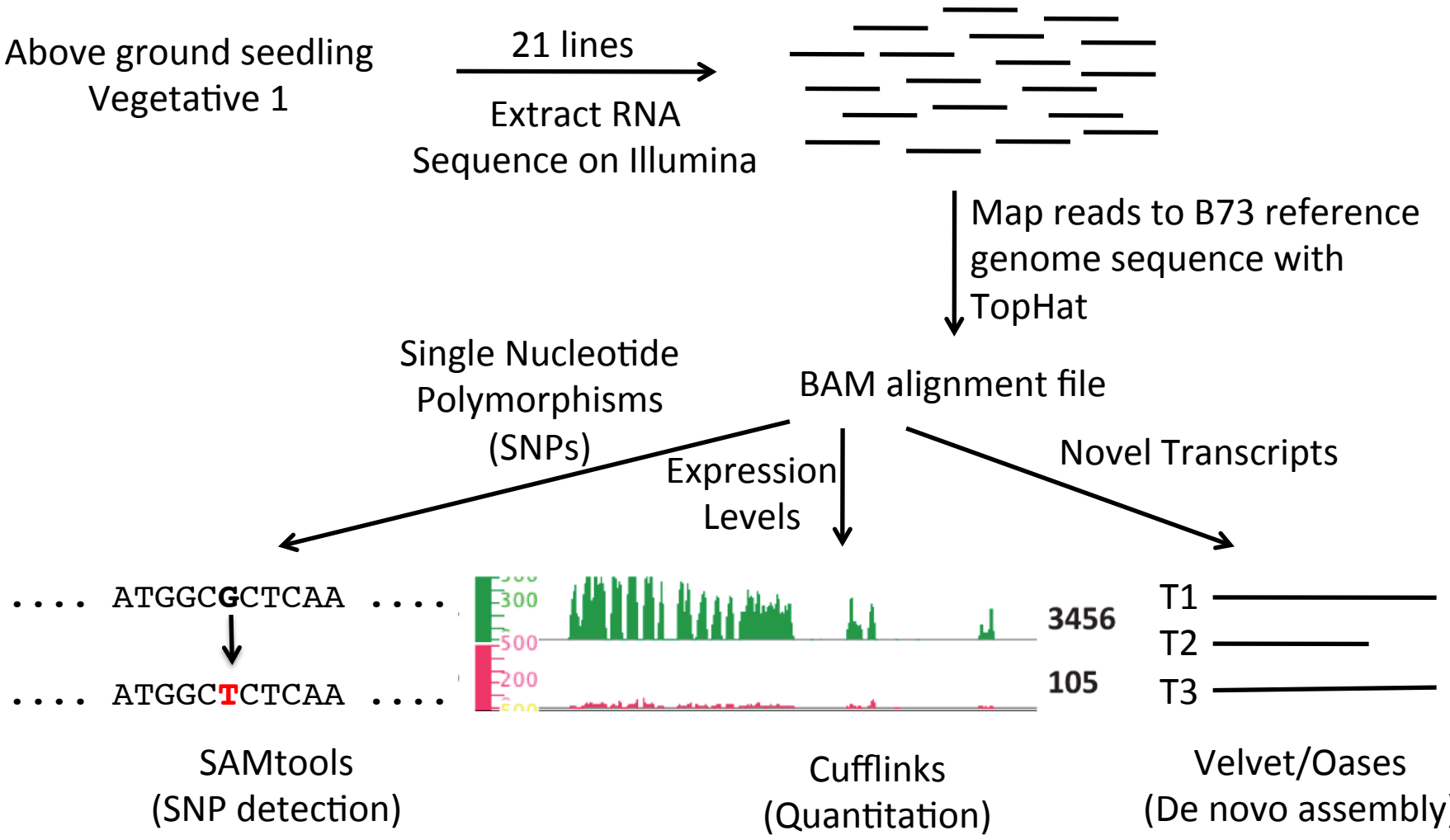
http://www.bsb.unimi.it/plant_genetics.htm

Unique to the Wisconsin Diversity Panel

Unique to the Goodman-Buckler Diversity Panel

Common to Both Diversity Panels

Analysis Methods



SNP Variant Detection Summary

Number of Genotypes	Number of SNPs	Number of Genes
21	53,094	9,498
20	34,510	2,551
19	27,135	1,509
18	23,785	1,159
17	21,467	934
16	19,745	806
15	17,984	689
14	16,914	656
13	16,252	589
12	15,785	557
11	15,174	544
10	14,893	524
9	14,483	524
8	14,265	524
7	13,769	504
6	12,612	458
5	11,484	419
4	8,359	385
Total	351,710	22,830

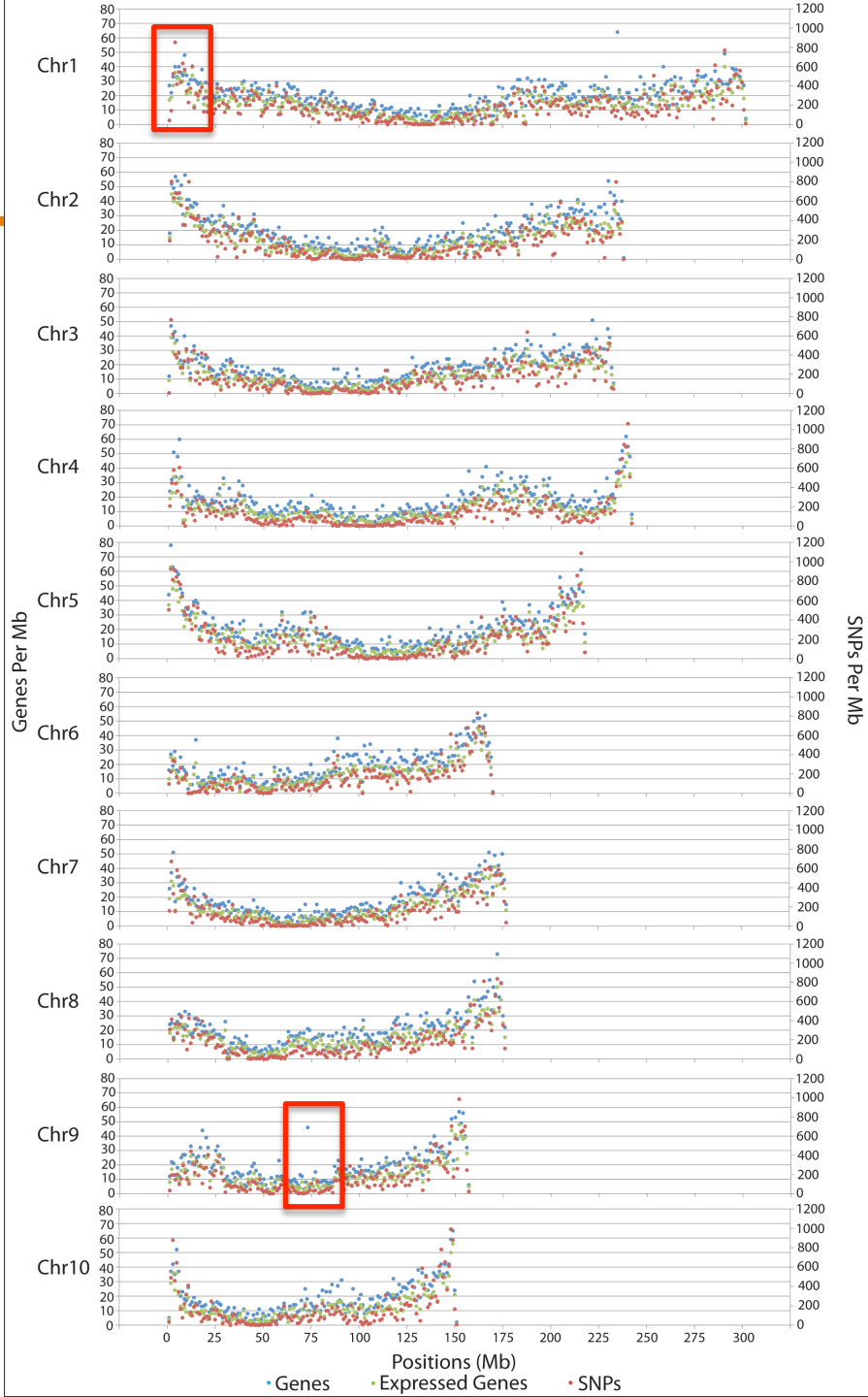
197,720 SNPs in 17,149 genes

Hansey et al., 2012 PLoS ONE

SNP Distribution

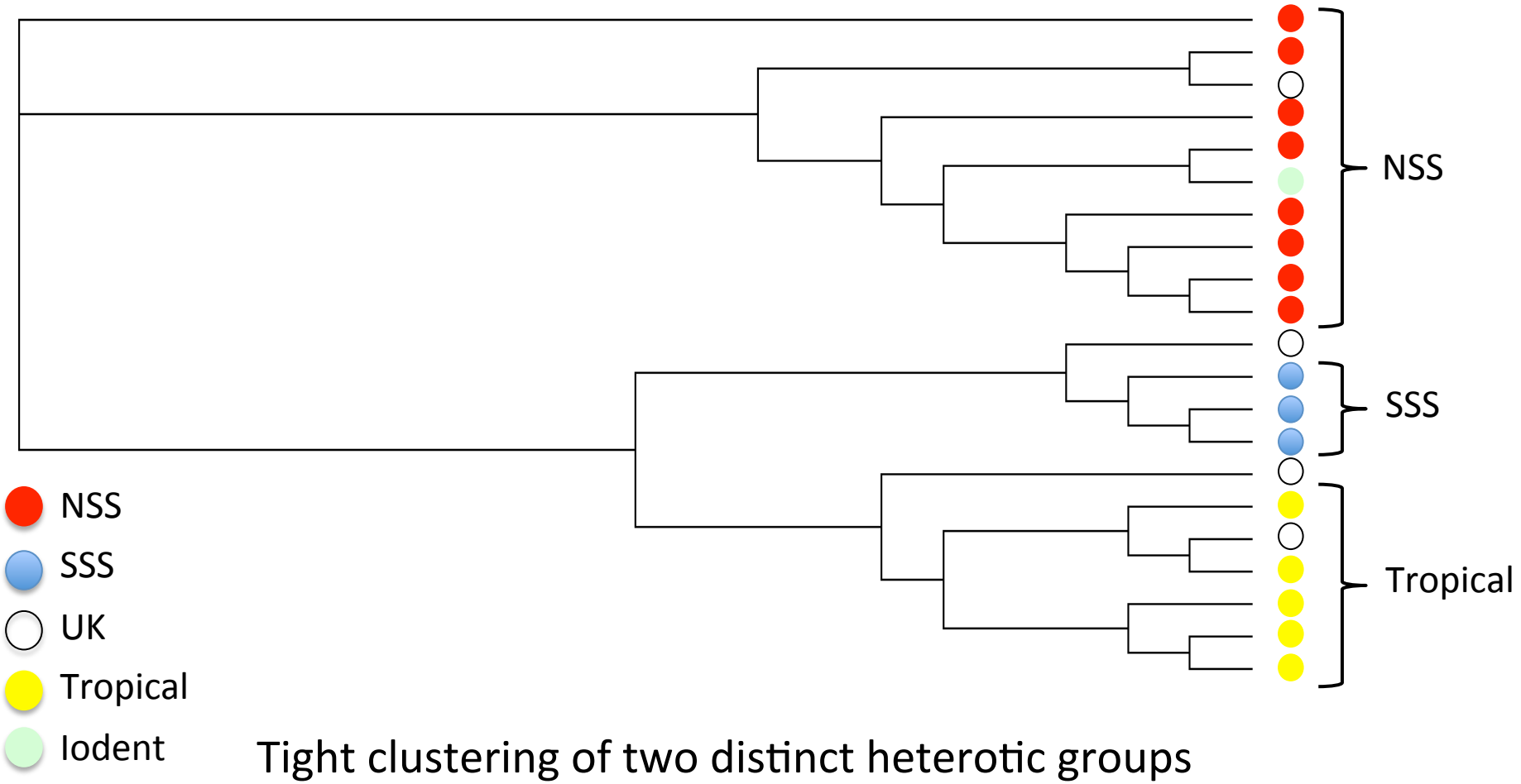
-Good distribution of SNPs across genome

-Some regions where there is high (low) SNP density



Hansey et al., 2012 PLoS ONE

Clustering Based on SNPs

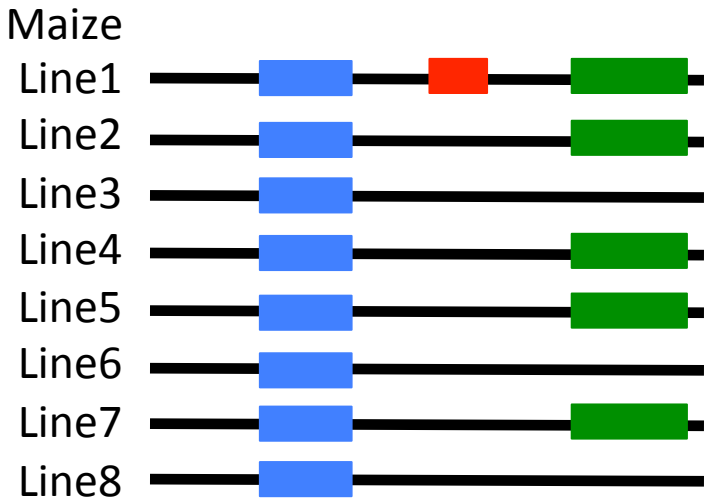


Tight clustering of two distinct heterotic groups and exotic lines is evident with SNP genetic markers

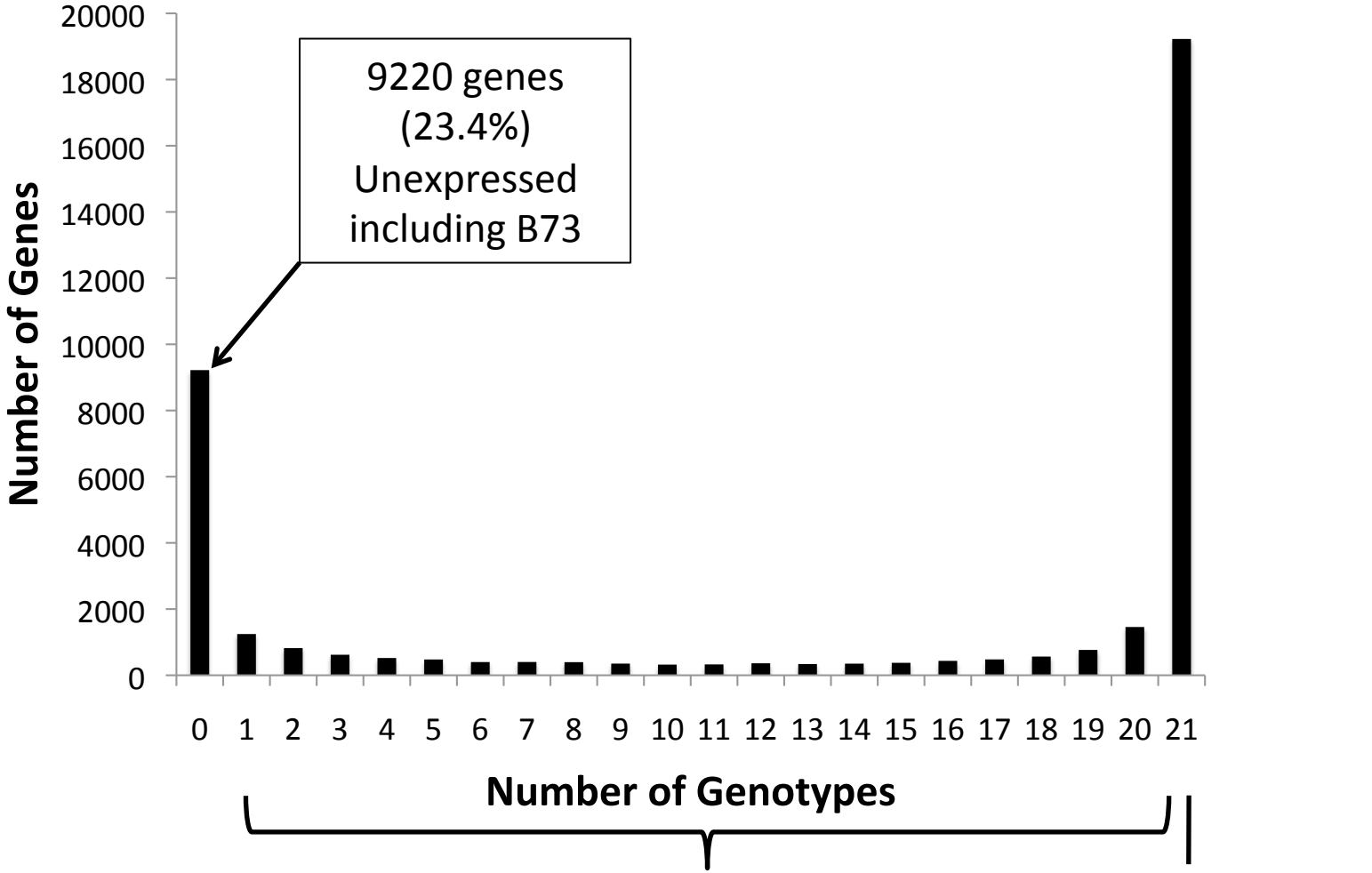
Hansey et al., 2012 PLoS ONE

Pan Genome and Transcriptome

- Pan genome – full complement of genes in a species
- Core genome – genes present in all individuals
- Dispensable genome – genes found in only a subset of the individuals
  - Dispensable and unique genome
- Pan, core and dispensable transcriptome

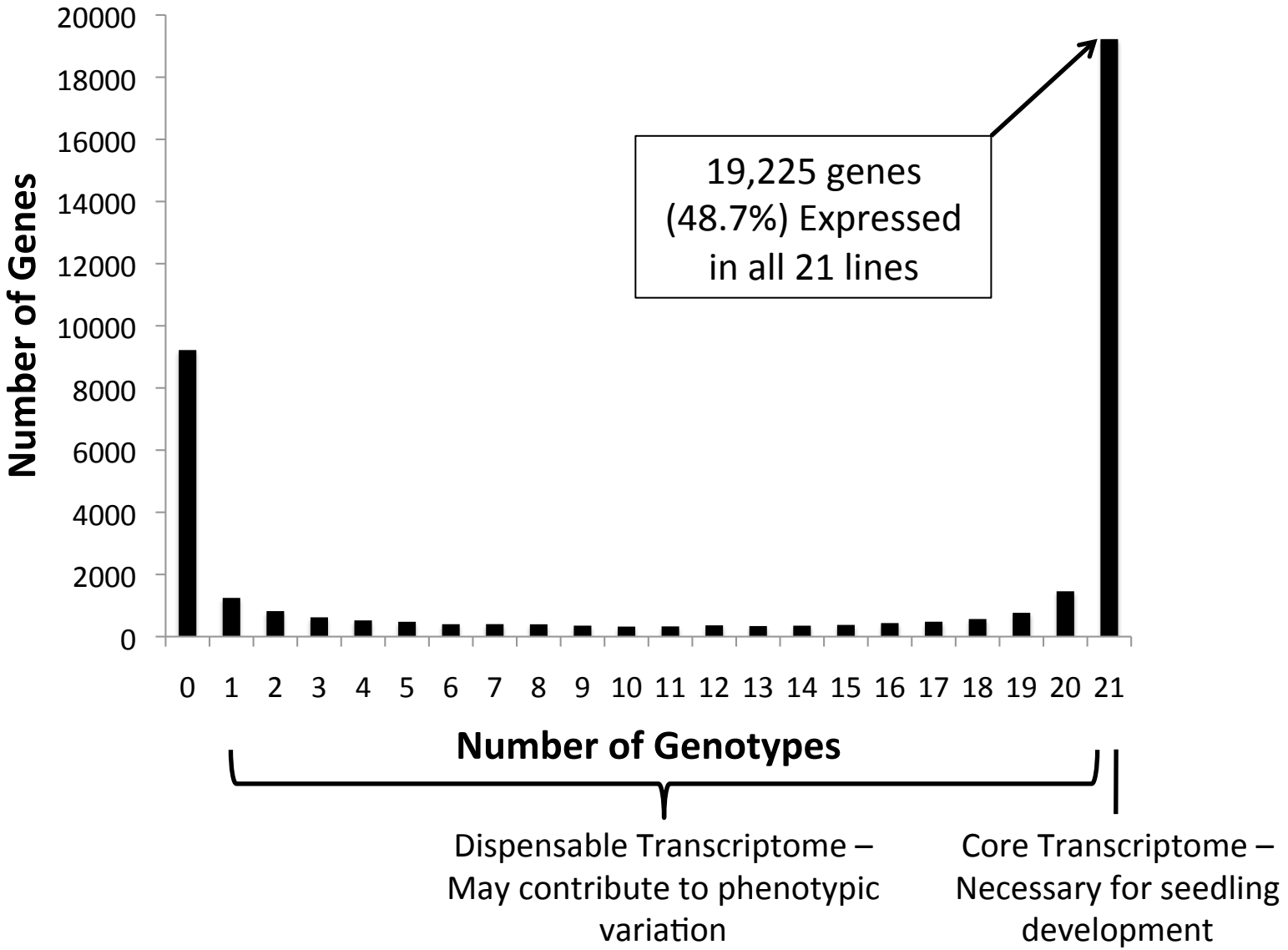


Quantitative Pan Transcriptome



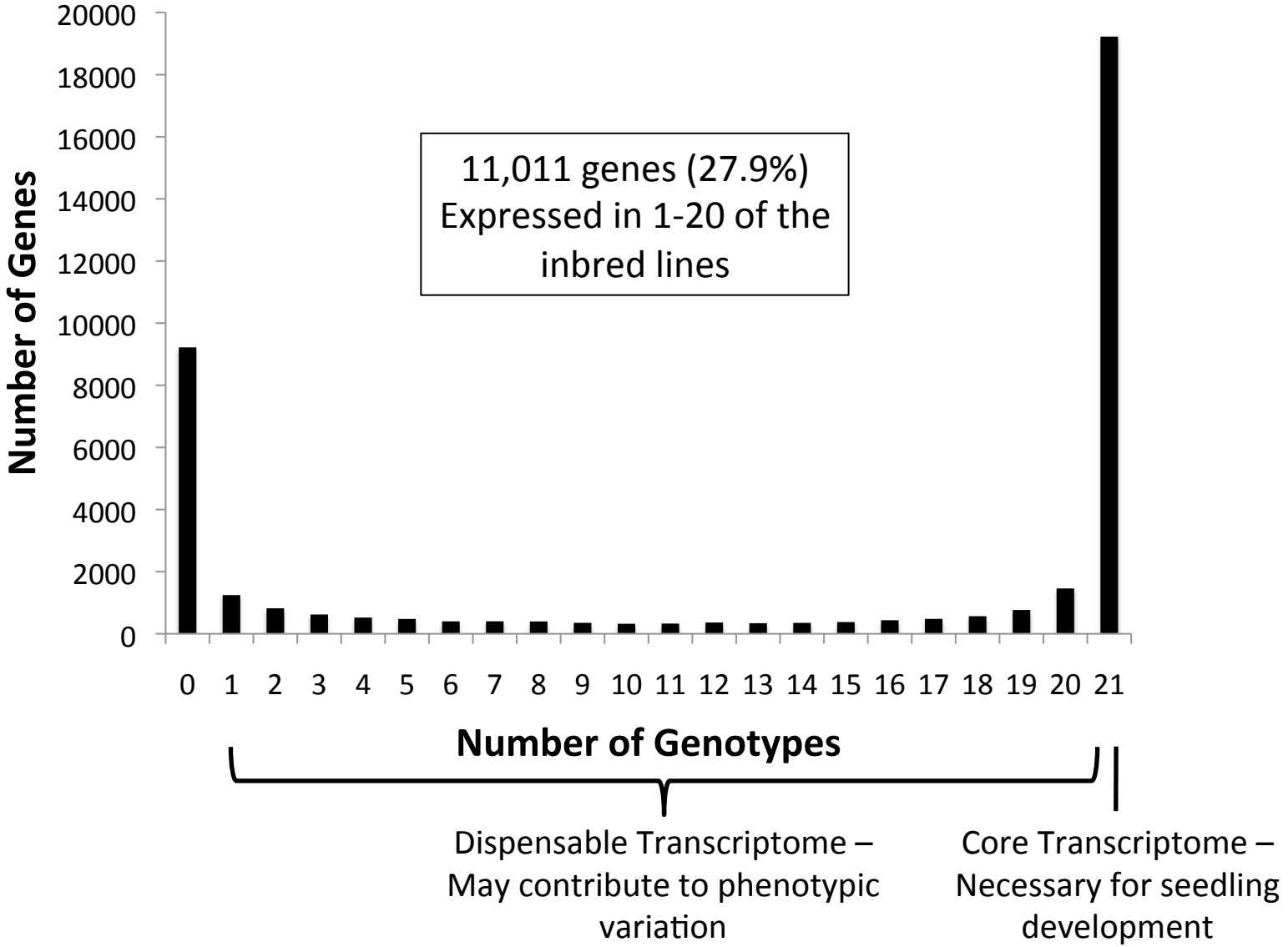
Hansey et al., 2012 PLoS ONE

Quantitative Pan Transcriptome



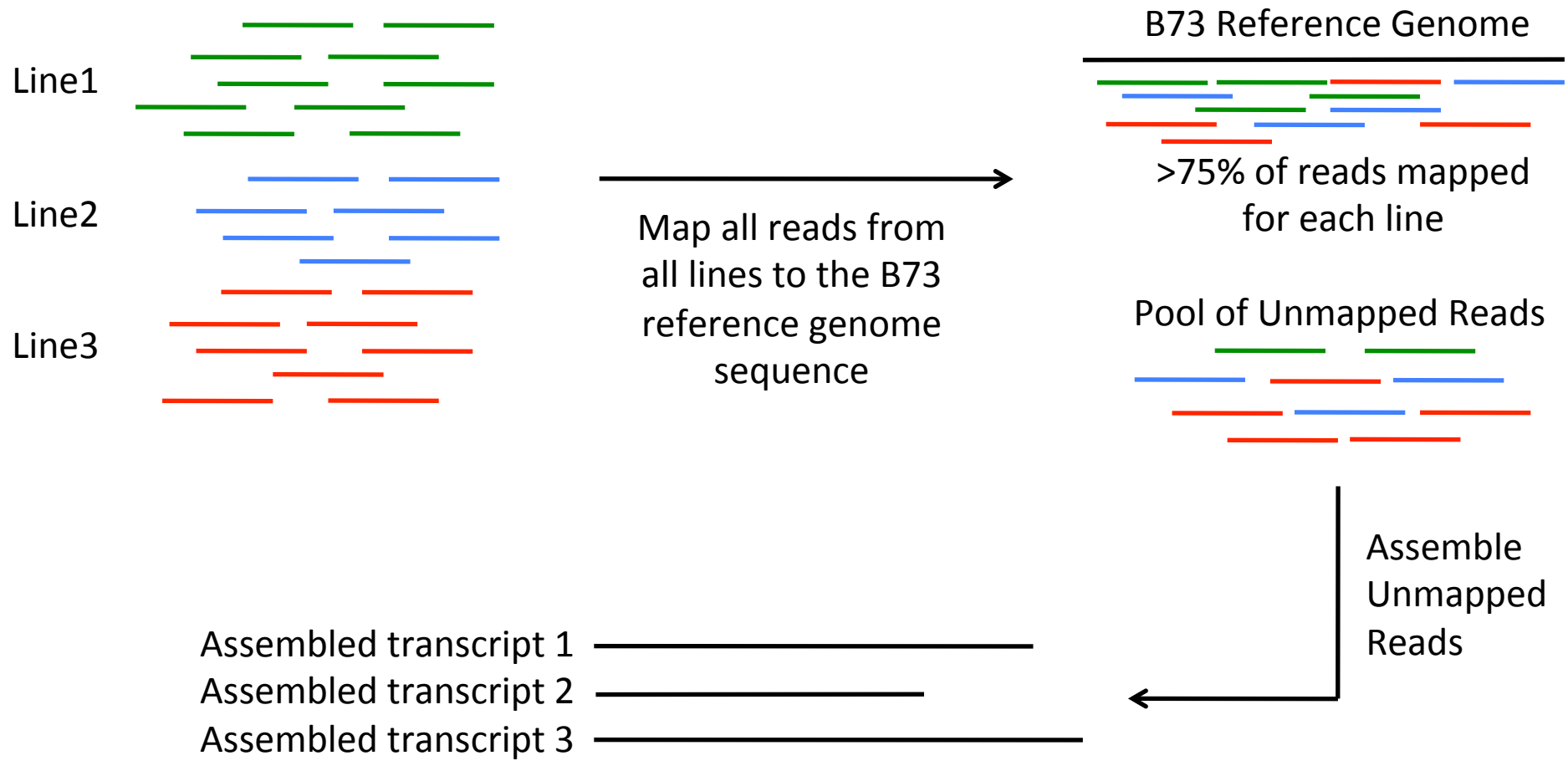
Hansey et al., 2012 PLoS ONE

Quantitative Pan Transcriptome



Hansey et al., 2012 PLoS ONE

Novel Transcript Discovery



Transcript Assembly

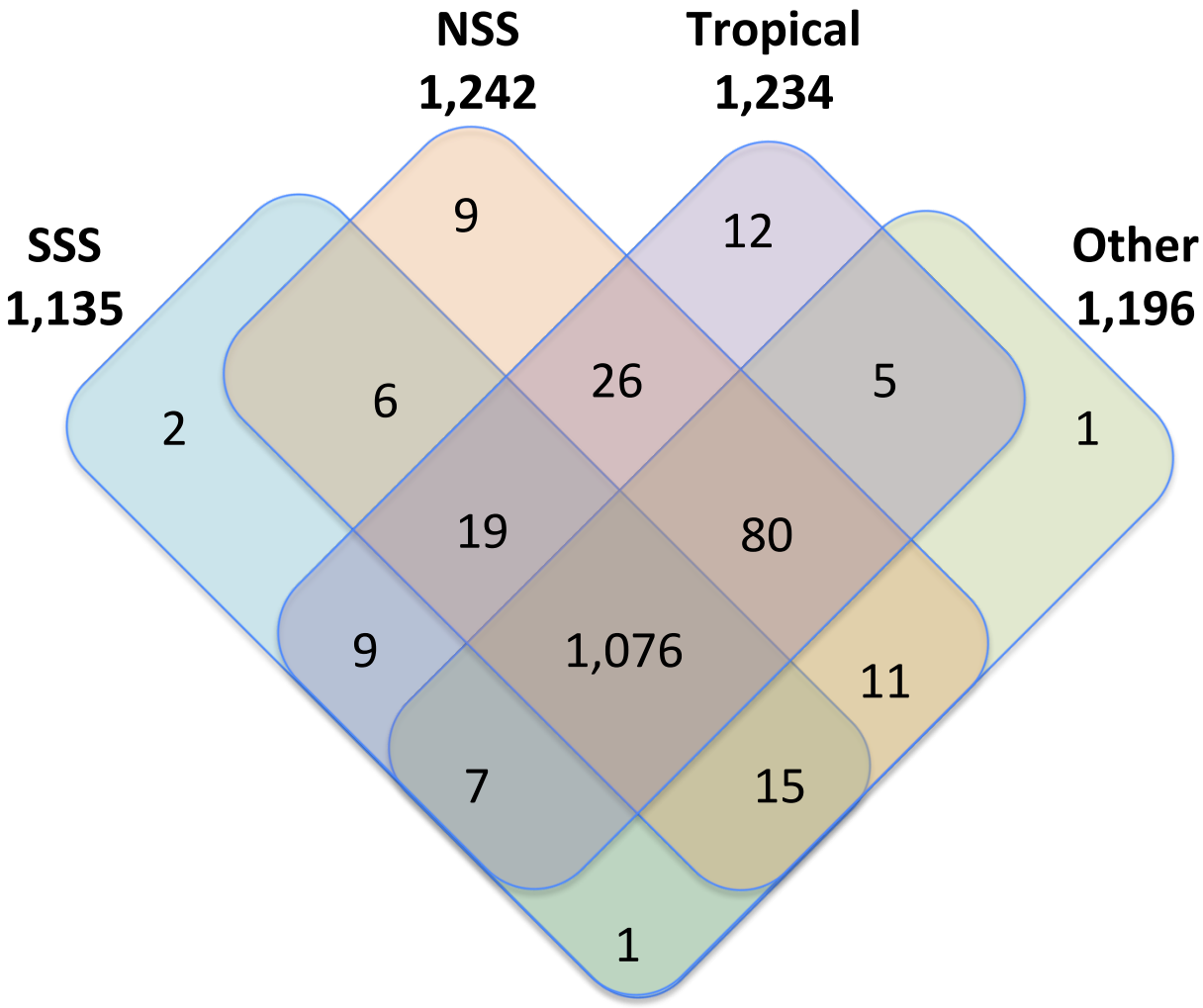
- Assembled 4,701 unique loci
- N50 contig of 725 bp
- 1,321 high confidence transcripts after mapping back to the reference sequence
- Computationally predicted the presence/absence variation (PAV) of the assembled transcripts in each inbred line by mapping reads to the assembled transcripts
- RT-PCR validation of computational predictions (87.5%)

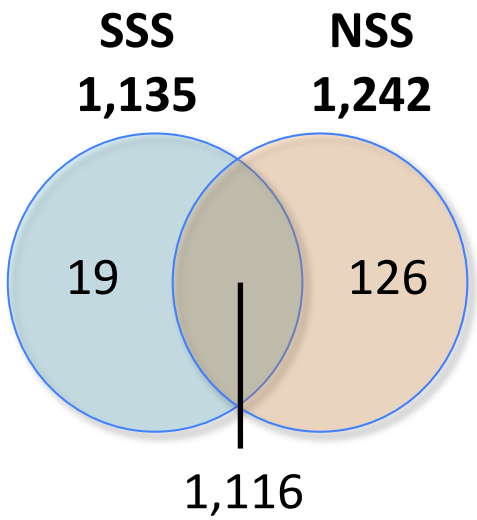
Hansey et al., 2012 PLoS ONE

Novel Transcript Assembly

- 564 transcripts expressed in all 21 lines and missing from the B73 reference sequence
- 715 transcripts with transcript presence/ absence variants that could reflect genomic presence/absence variants

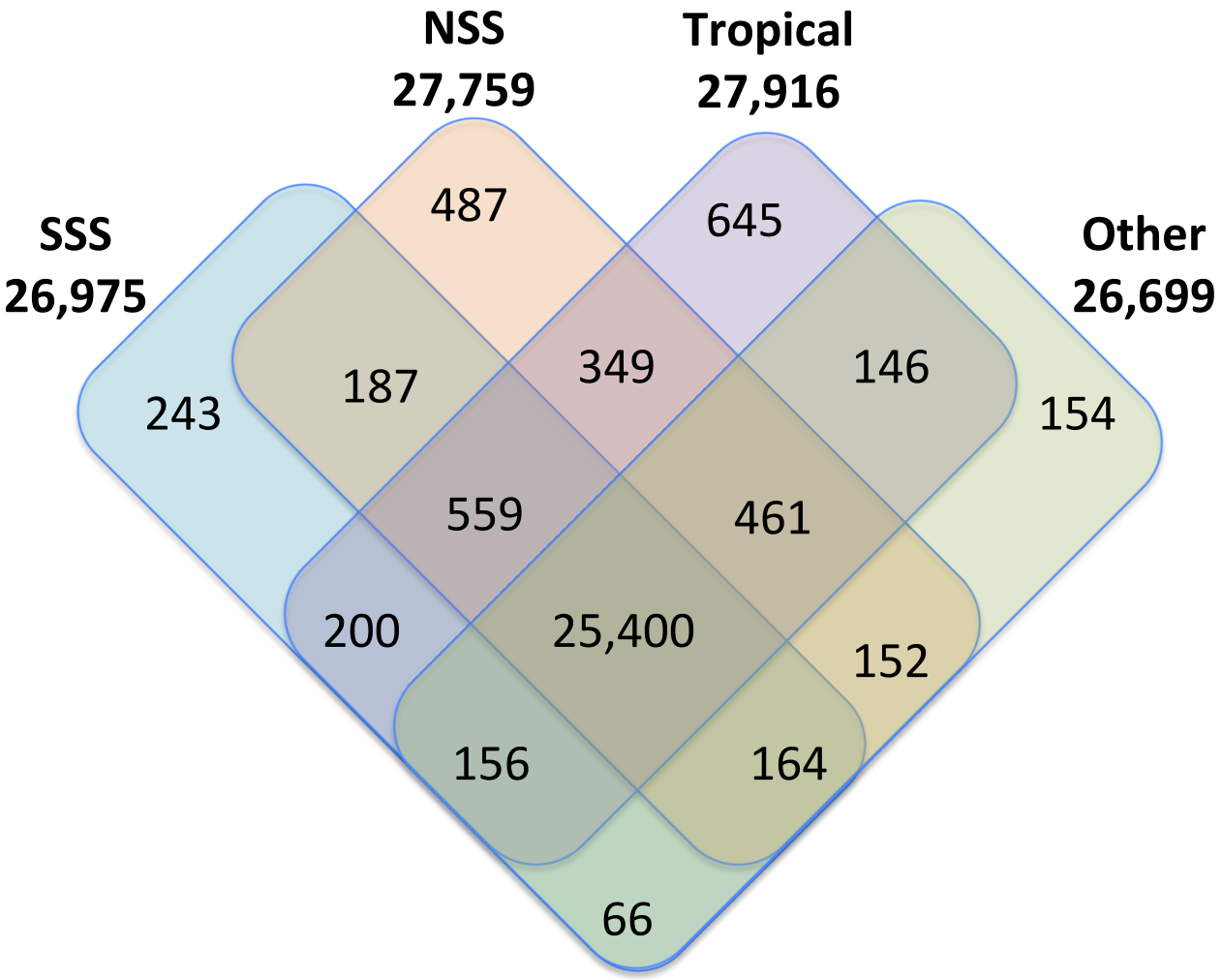
Are Novel Presence/Absence Transcripts Associated with Heterosis?

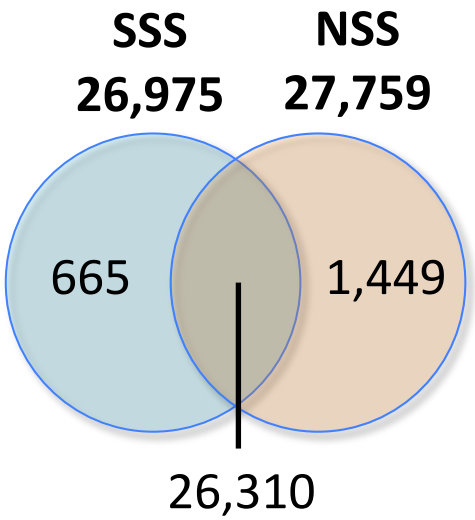




Hansey et al., 2012 PLoS ONE

Annotated and Novel Transcripts Associated with Heterotic Groups





Hansey et al., 2012 PLoS ONE

Conclusions

- RNA-seq is a robust, rapid, and inexpensive method to:
  - Measure expression abundances across a core set of tissues
  - Improve our annotation of genomes
  - identify SNPs in genic regions in crop species with large, complex, repetitive genomes
- Using *de novo* assembly, we discovered novel sequences previously unidentified in maize
- We identified a core set of essential genes, as well as a set of genes that are dispensable to the maize seedling transcriptome and may be contributing to phenotypic variation
- The structural variation observed at the genome level in maize between inbred lines in opposite heterotic groups extends to the transcriptome
- While additional research is needed to definitively implicate allelic, structural, and transcriptome level variation in heterosis, this study provides growing evidence to the involvement of all of these levels of variation in heterosis

Acknowledgements

- Michigan State University
  - Candice Hansey (Hirsch)
  - Rebecca Davidson
  - Brieanne Vaillancourt
  - Kevin Childs
  - Malali Gowda
  - Ning Jiang
- University of Wisconsin
  - Natalia de Leon
  - Shawn Kaeppler
  - Rajandeep Sekhon
- Funding
  - DOE Great Lakes Bioenergy Research Center

  - USDA NIFA