Multiple Reference Genomes and Transcriptomes for *A. thaliana*

[Gan et al., Nature, 2011]

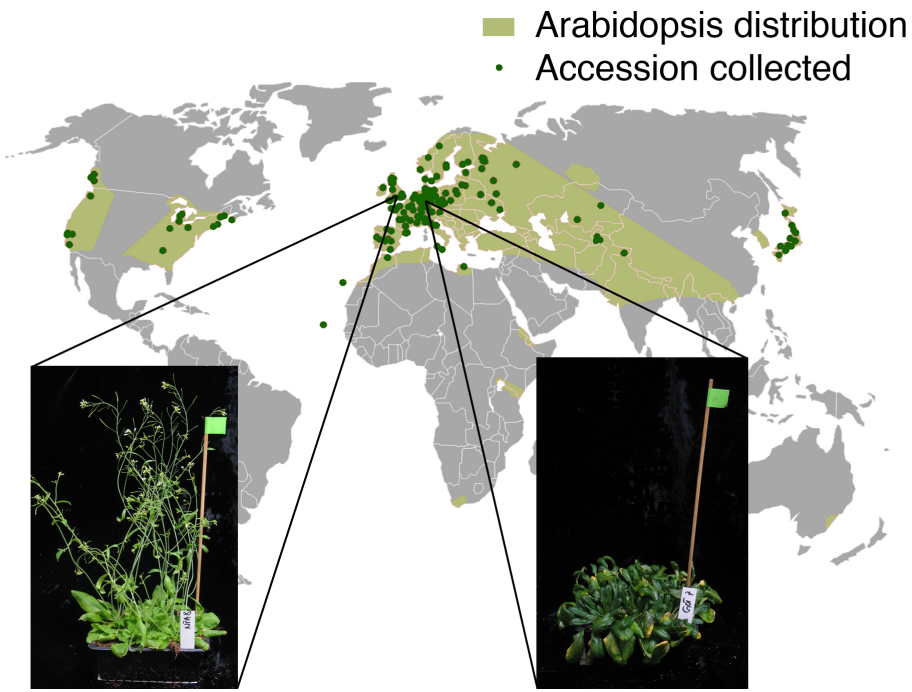
Gunnar Rätsch

Computational Biology Center
Memorial Sloan-Kettering Cancer Center

Memorial Sloan-Kettering Cancer Center

cBio@MSKCC

18+1 Reference Genomes and Transcriptomes

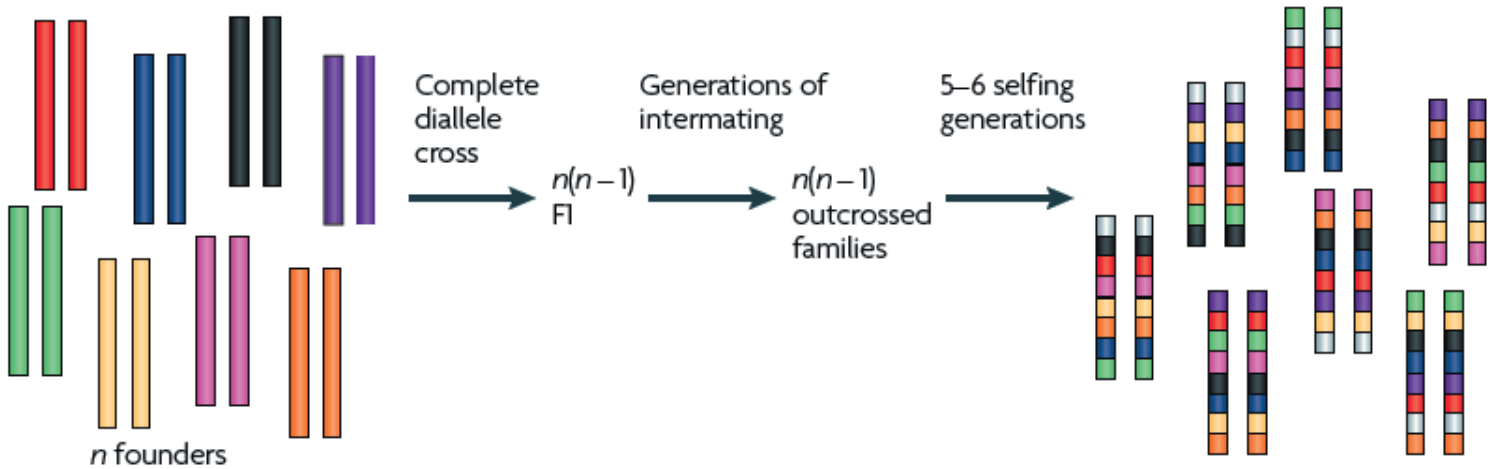


- University of Oxford
  Genome sequencing & reconstruction
  G. Xiangchao, N. Harberd, R. Mott
- University of Utah
  Transcriptome sequencing
  J. Steffen, C. Toomajian, R. Clark
- Max Planck Campus Tübingen
  Reannotation & transcriptome analysis
  O. Stegle, J. Behr, P. Drewe, A. Kahles, R. Bohnert, G. Jean, S. J. Schultheiss, G. Rätsch
- University of Bath (P. Kover), Kansas State University (K. L. Hildebrand, C. Toomajian), EBI (P. Derwent, P. Kersey), The Sainsbury Laboratory (E. Kemen)

[Gan et al., 2011]

Variability & Association Studies

Multiparent Advanced Generation Inter-Cross (MAGIC) lines



- Currently 500 available MAGIC lines [w/ Richard Mott]
- ≈200 natural strains with RNA-seq [w/ Magnus Nordborg]

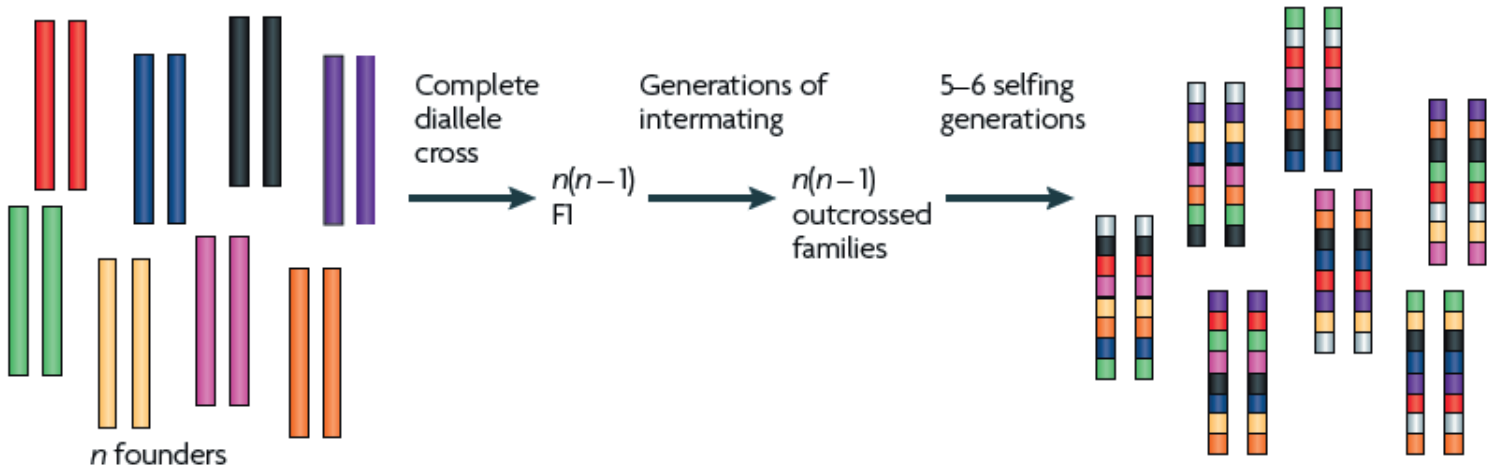
In Progress:

- Associate molecular phenotypes with genotypic variation
- Identify *cis* and *trans*-factors:

Goal: Accurate models $f(\text{DNA}, \quad , \ldots) = \text{RNA}$

Variability & Association Studies

Multiparent Advanced Generation Inter-Cross (MAGIC) lines



- Currently 500 available MAGIC lines [w/ Richard Mott]
- ≈200 natural strains with RNA-seq [w/ Magnus Nordborg]

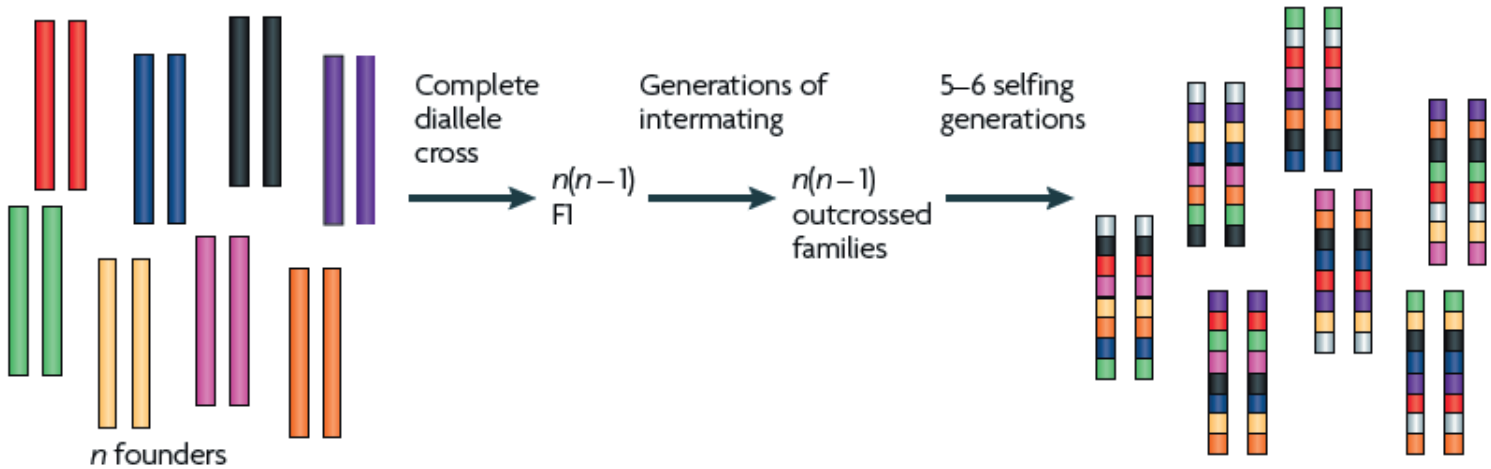
In Progress:

- Associate molecular phenotypes with genotypic variation
- Identify *cis* and *trans*-factors:

Goal: Accurate models $f(\text{DNA}, \quad , \ldots) = \text{RNA}$

Variability & Association Studies

Multiparent Advanced Generation Inter-Cross (MAGIC) lines



- Currently 500 available MAGIC lines [w/ Richard Mott]
- ≈200 natural strains with RNA-seq [w/ Magnus Nordborg]

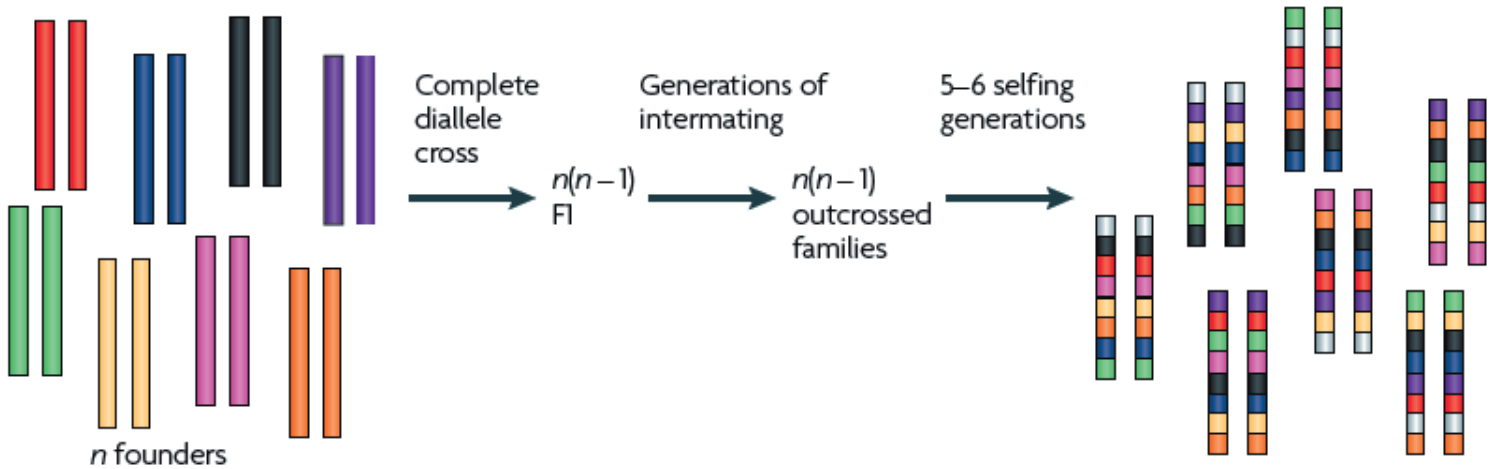
In Progress:

- Associate molecular phenotypes with genotypic variation
- Identify *cis* and *trans*-factors: 1 *cis*/*trans*

Goal: Accurate models $f(\text{DNA}, \quad , \ldots) = \text{RNA}$

Variability & Association Studies

Multiparent Advanced Generation Inter-Cross (MAGIC) lines



- Currently 500 available MAGIC lines [w/ Richard Mott]
- ≈200 natural strains with RNA-seq [w/ Magnus Nordborg]

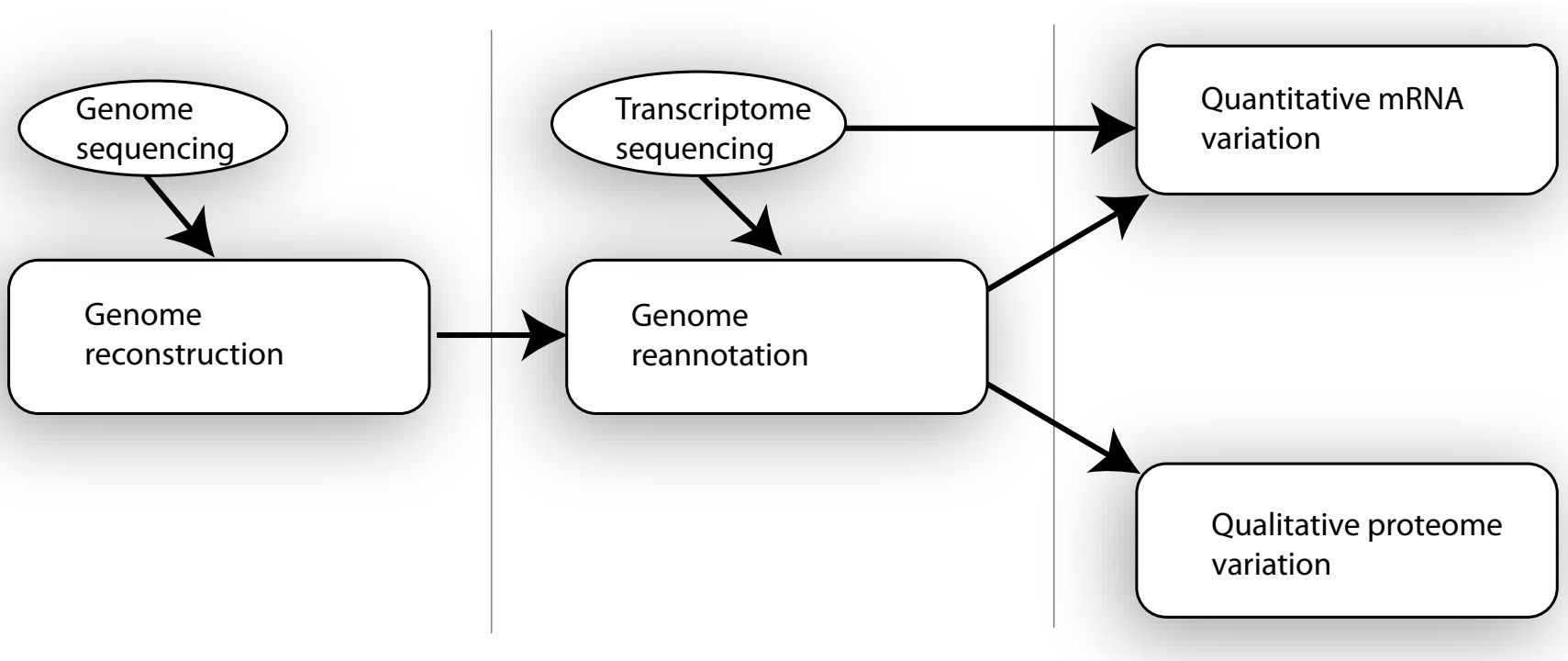
In Progress:

- Associate molecular phenotypes with genotypic variation
- Identify *cis* and *trans*-factors: $\mathbb{1}_{cis/trans}$

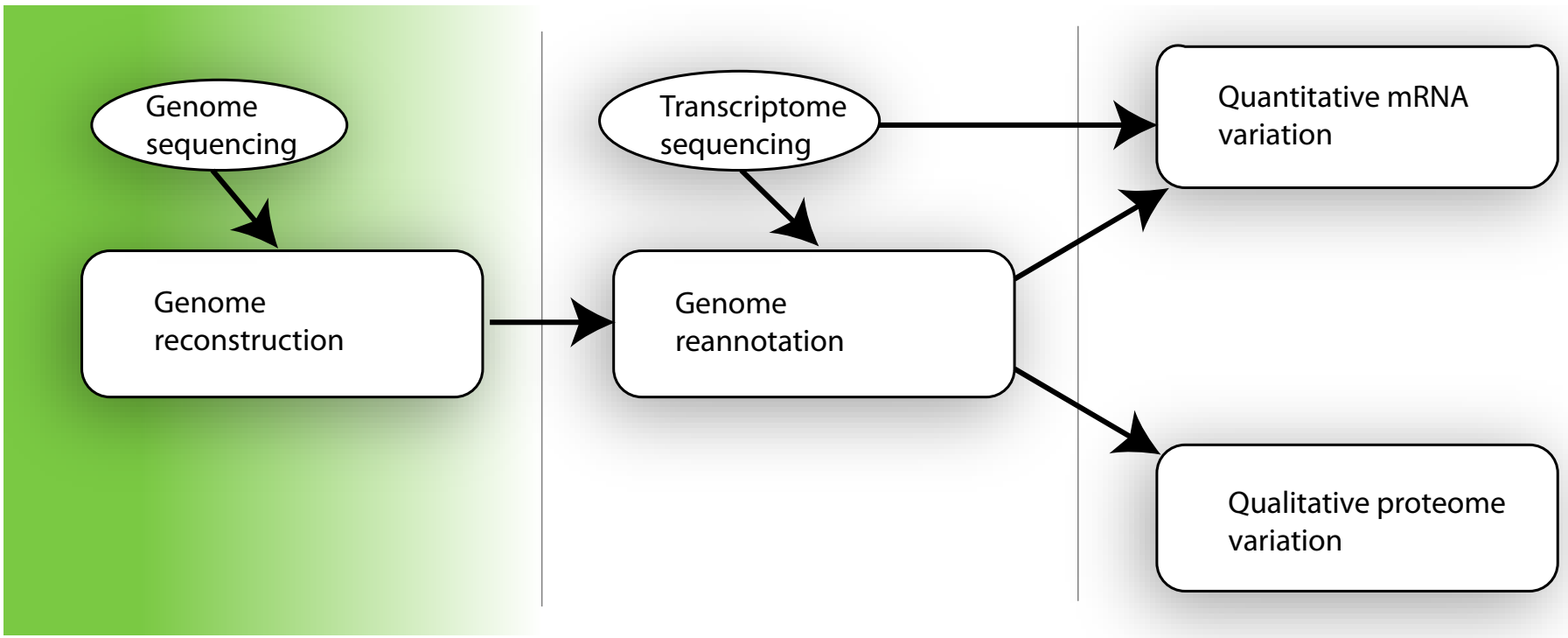
Goal: Accurate models $f(\text{DNA}, \mathbb{1}_{cis/trans}, \ldots) = \text{RNA}$

18+1 Reference Genomes and Transcriptomes

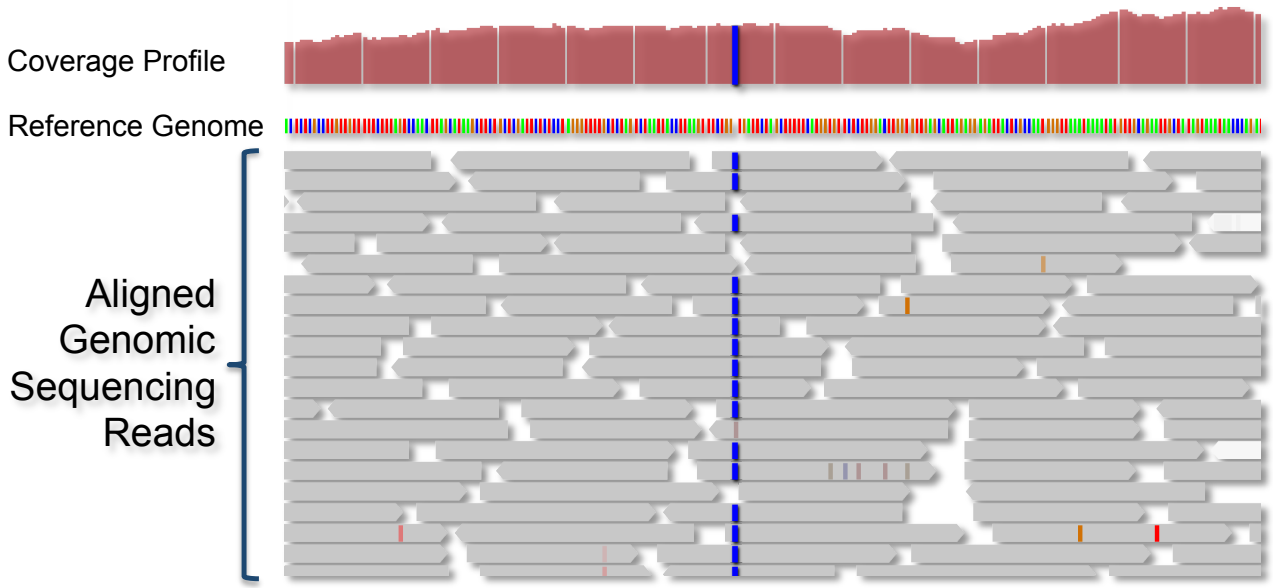
Goal: Characterize genome, transcriptome and proteome variation in 19 strains of the plant *Arabidopsis thaliana*



18+1 Reference Genomes and Transcriptomes

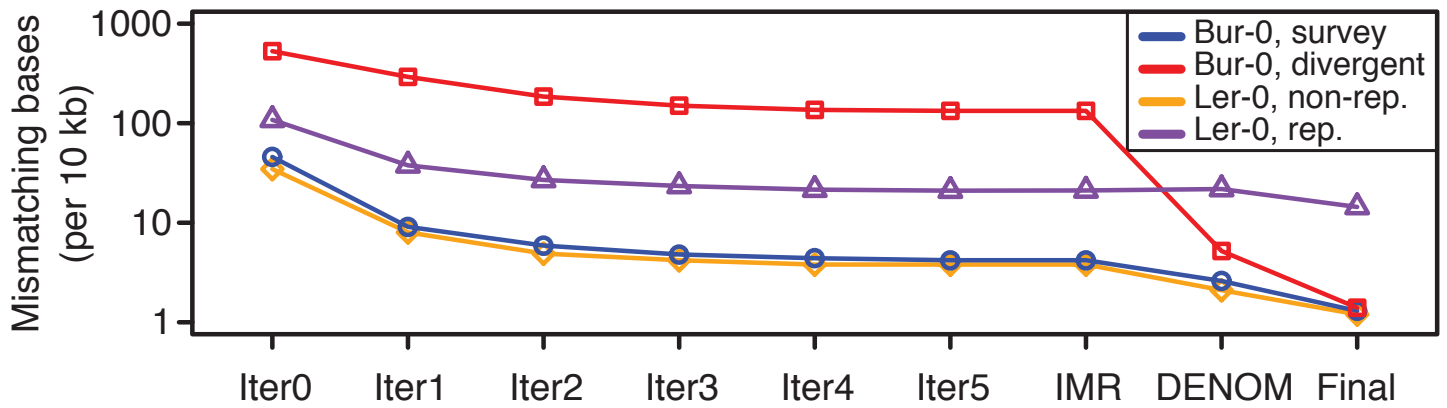


Sequencing 18 Arabidopsis Genomes



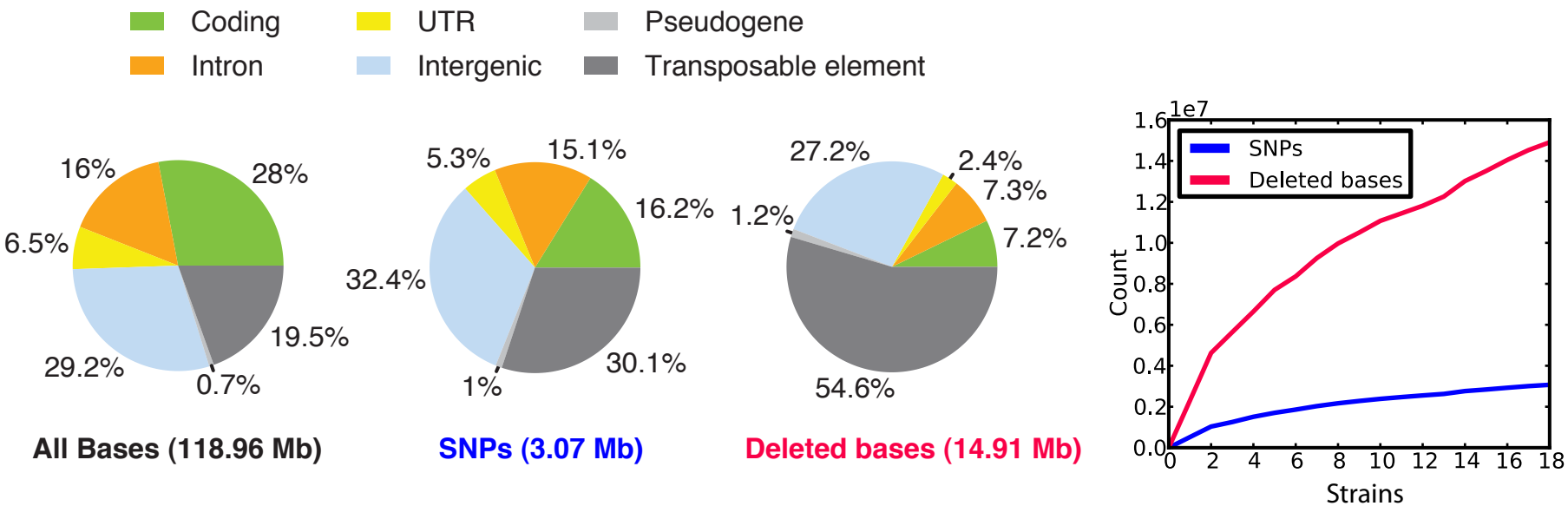
- ≈40x coverage with Illumina paired end reads
- Multiple insert sizes (200 & 400 bp, read lengths 32-51bp)
- Iterative read mapping against the reference genome
- Full *de novo* assembly with SOAP and integration of contigs
  $\Rightarrow \leq 1$ error / 10kb $\Rightarrow$ sufficient for genome reannotations
- Extensive genomic variation (3Mbp SNPs, 15Mbp indels)

Sequencing 18 Arabidopsis Genomes



- ≈40x coverage with Illumina paired end reads
- Multiple insert sizes (200 & 400 bp, read lengths 32-51bp)
- Iterative read mapping against the reference genome
- Full *de novo* assembly with SOAP and integration of contigs
  $\Rightarrow \leq 1$ error / 10kb $\Rightarrow$ sufficient for genome reannotations
- Extensive genomic variation (3Mbp SNPs, 15Mbp indels)

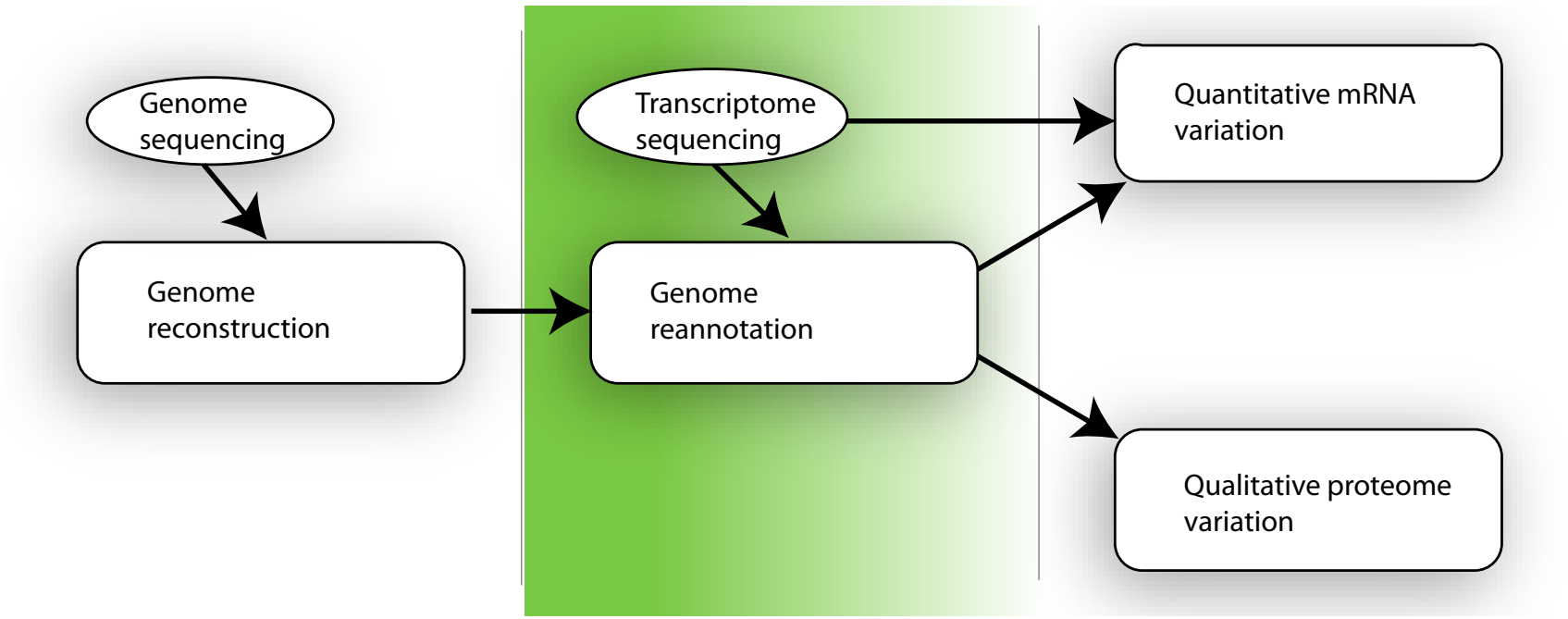
Sequencing 18 Arabidopsis Genomes



- ≈40x coverage with Illumina paired end reads
- Multiple insert sizes (200 & 400 bp, read lengths 32-51bp)
- Iterative read mapping against the reference genome
- Full *de novo* assembly with SOAP and integration of contigs
  $\Rightarrow \leq$ 1 error / 10kb $\Rightarrow$ sufficient for genome reannotations
- Extensive genomic variation (3Mbp SNPs, 15Mbp indels)

18+1 Reference Genomes and Transcriptomes

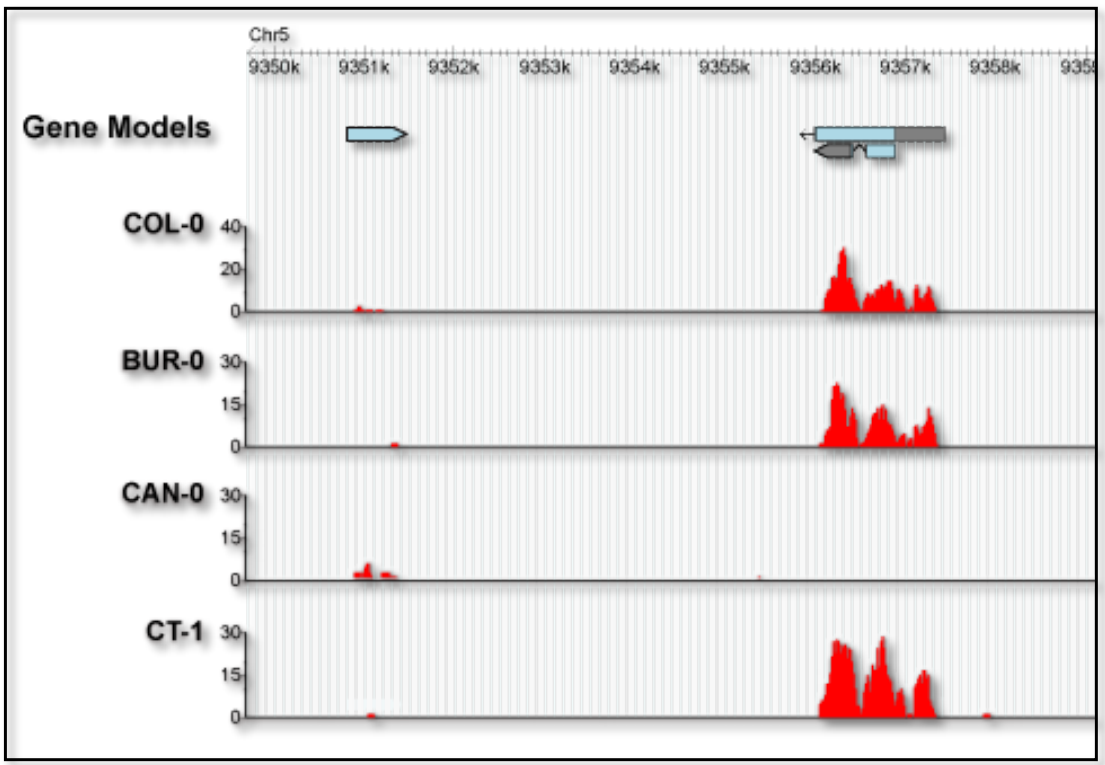
Goal: Characterize genome, transcriptome and proteome variation in 19 strains of the plant *Arabidopsis thaliana*



Transcriptome Sequencing

RNA-seq (Illumina, 76bp) of founder accessions:

- Non-strand-specific, 13 million reads in two replicates
- Strand-specific, 75 million reads for validation



- Gene expression estimates
- Genome reannotation

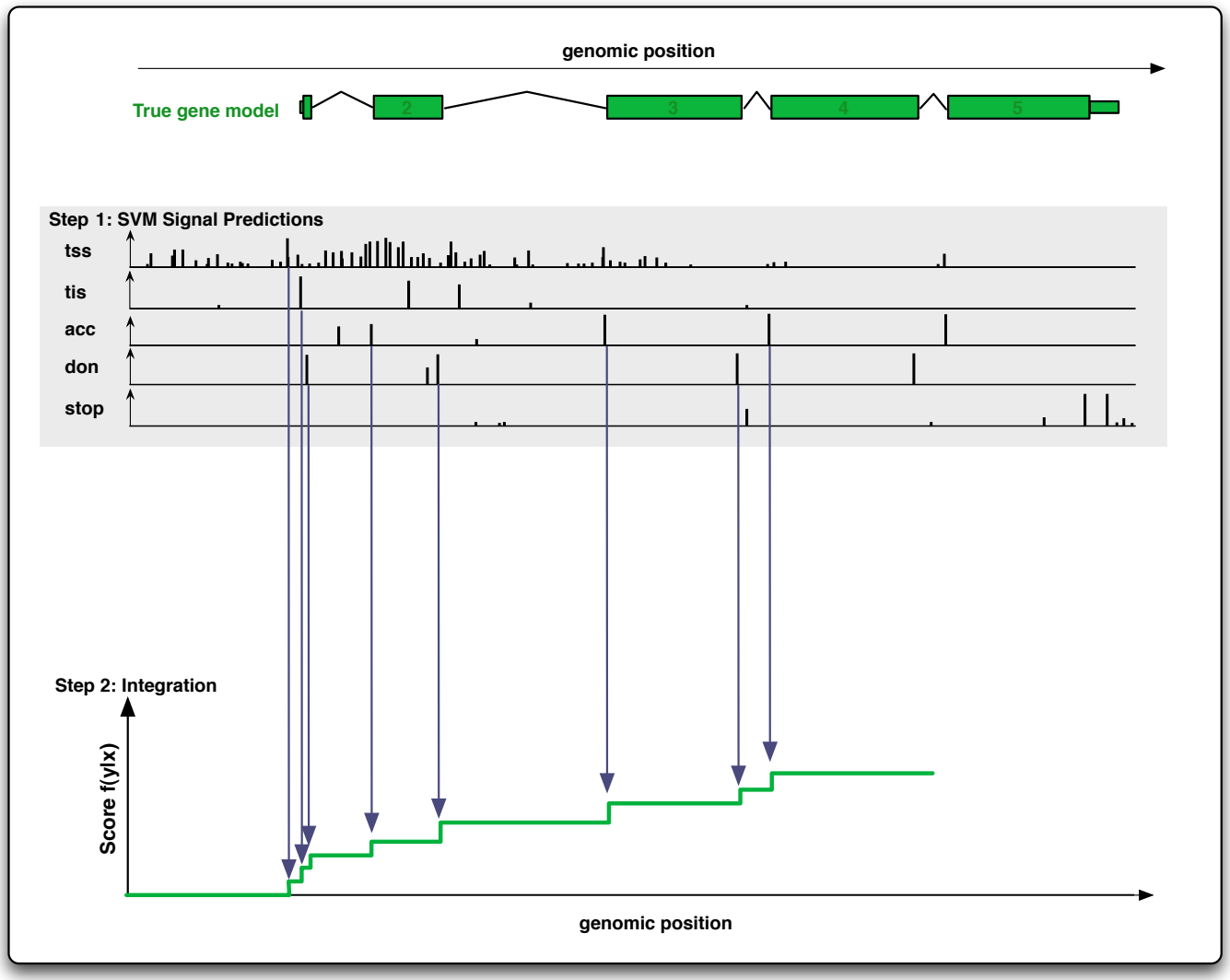
Genome Reannotation

mGene [Schweikert et al., 2009] & mGene.ngs [Behr et al., 2010, 2012, i.p.]



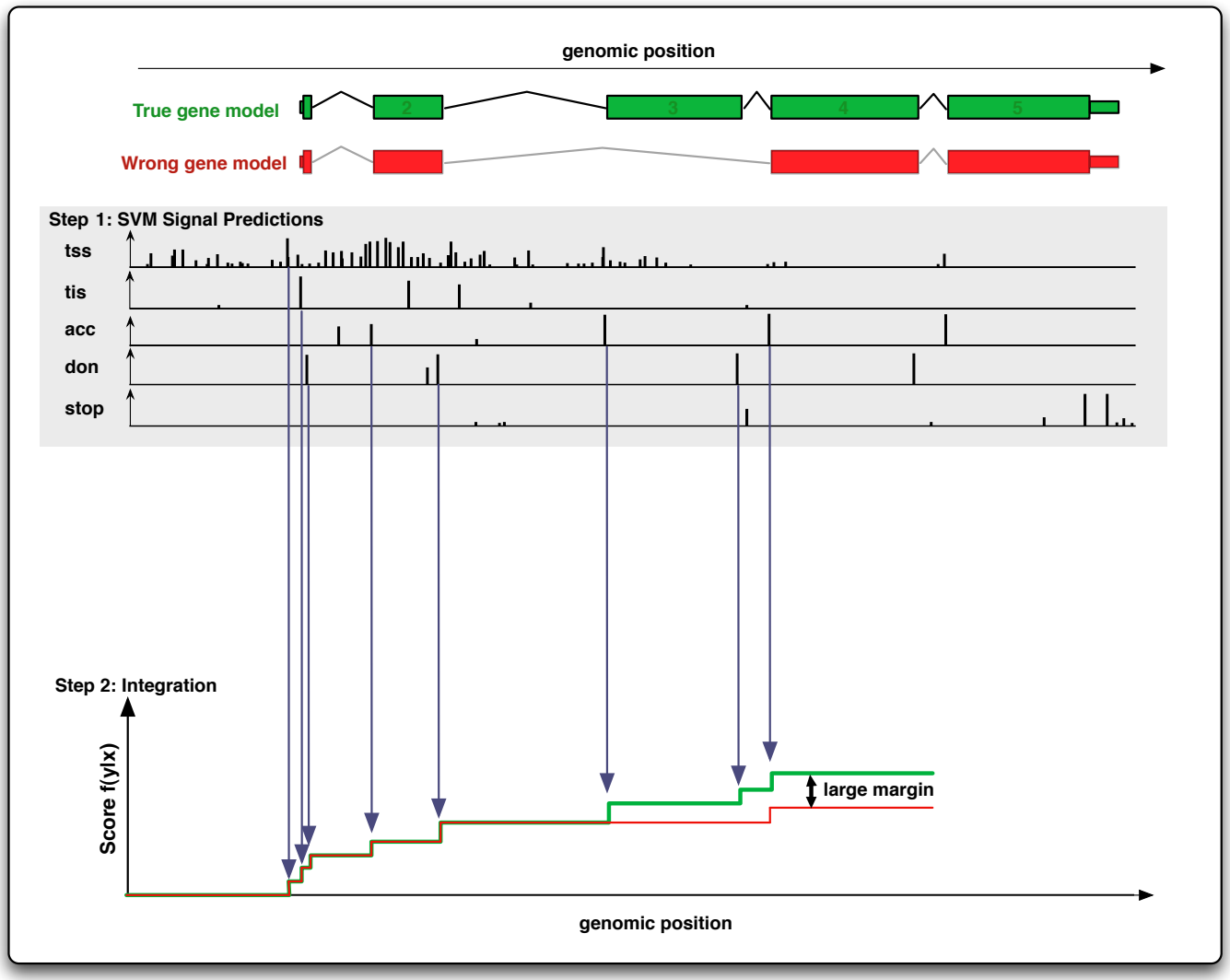
Genome Reannotation

mGene [Schweikert et al., 2009] & mGene.ngs [Behr et al., 2010, 2012, i.p.]



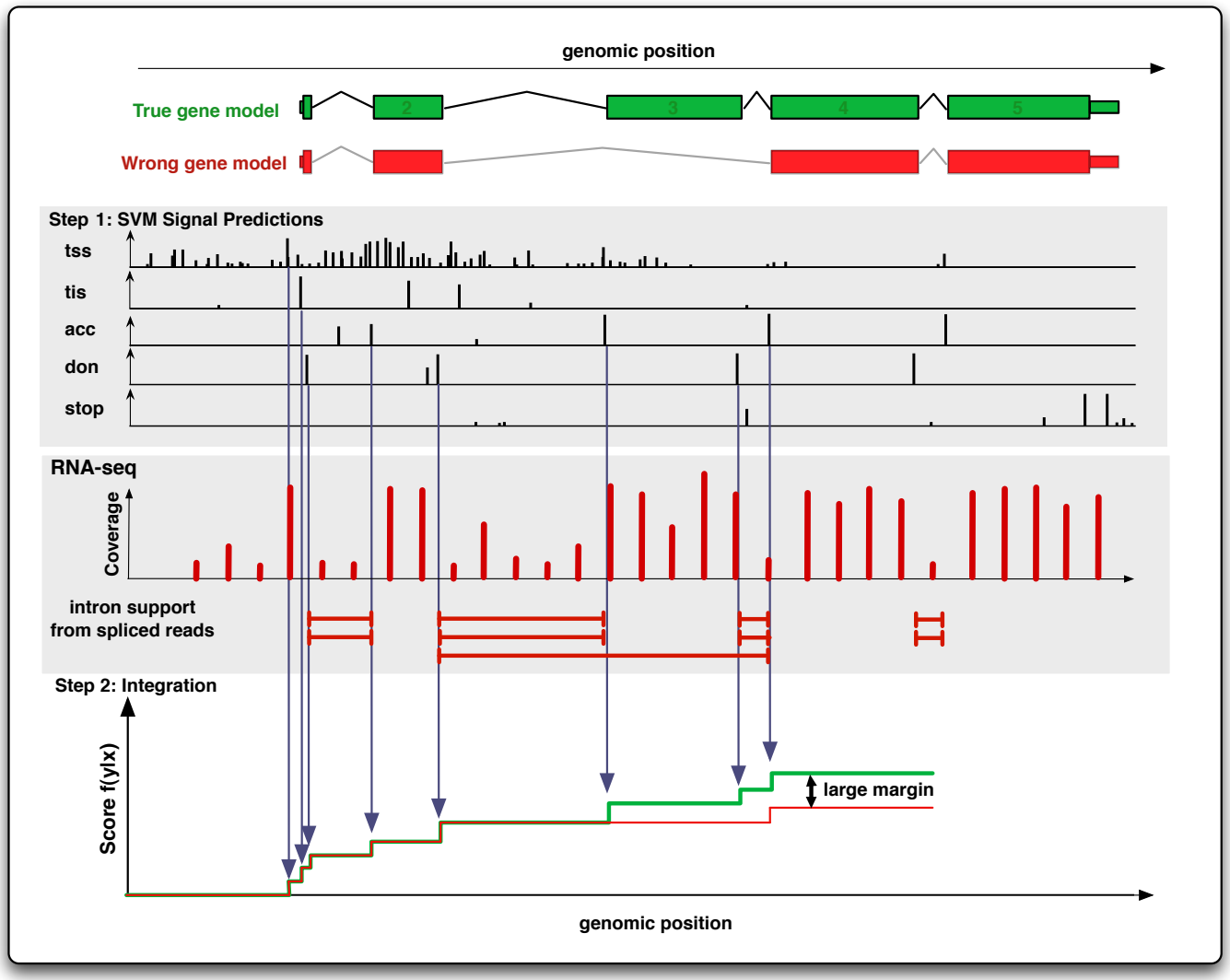
Genome Reannotation

mGene [Schweikert et al., 2009] & mGene.ngs [Behr et al., 2010, 2012, i.p.]



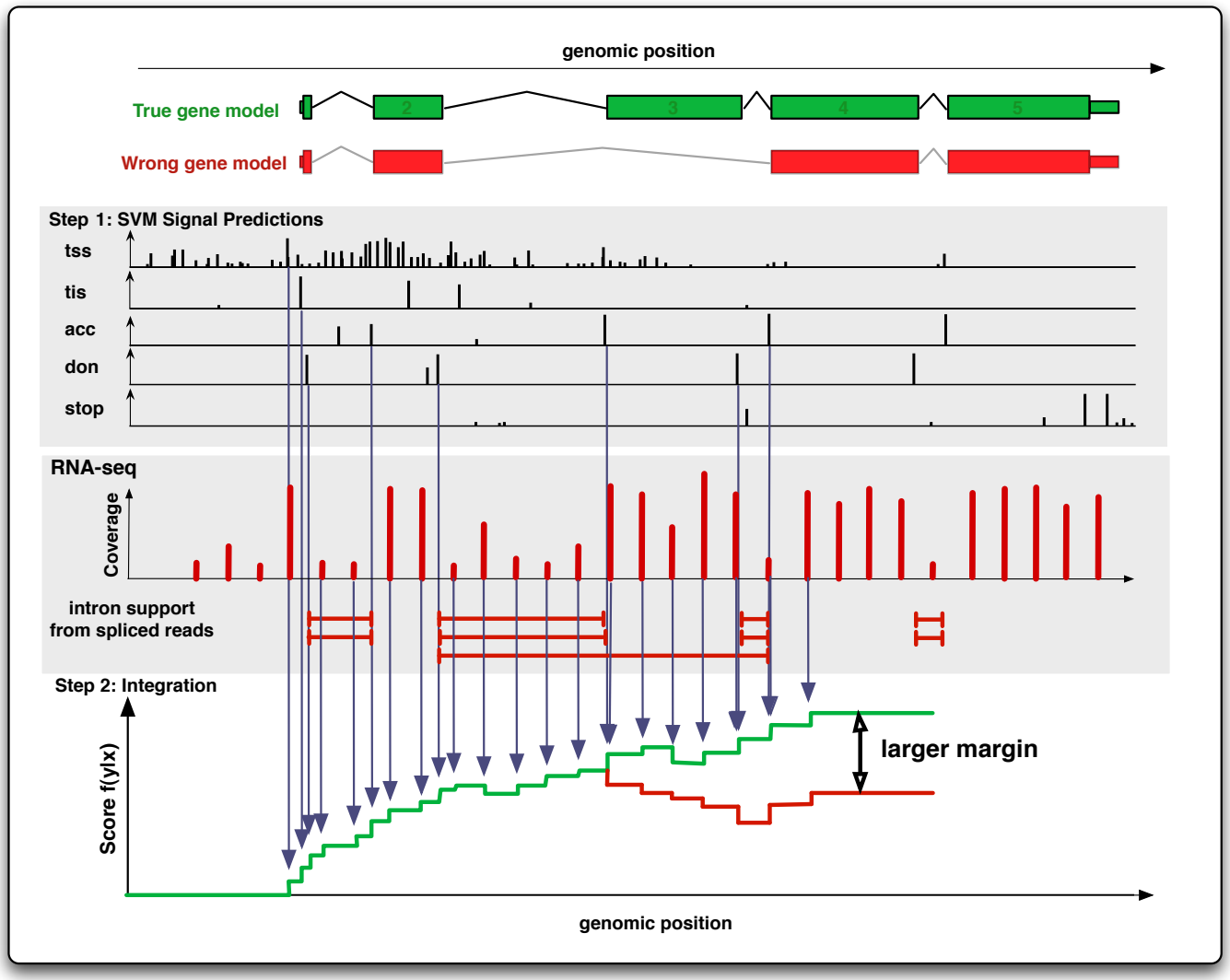
Genome Reannotation

mGene [Schweikert et al., 2009] & mGene.ngs [Behr et al., 2010, 2012, i.p.]



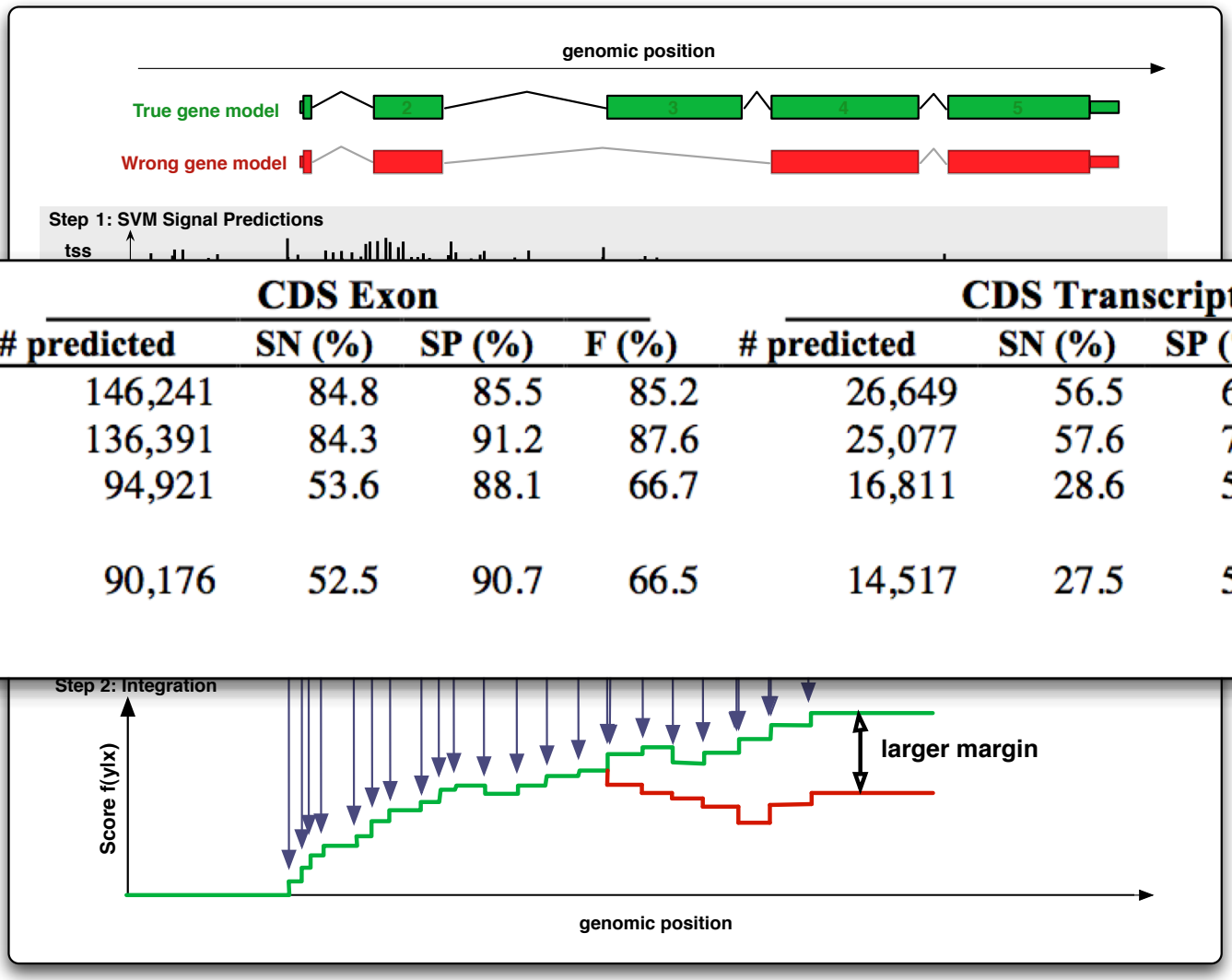
Genome Reannotation

mGene [Schweikert et al., 2009] & mGene.ngs [Behr et al., 2010, 2012, i.p.]



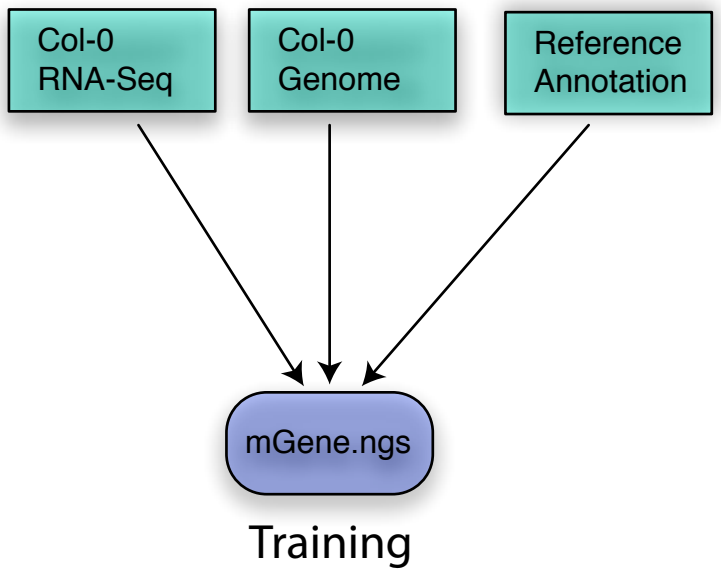
Genome Reannotation

mGene [Schweikert et al., 2009] & mGene.ngs [Behr et al., 2010, 2012, i.p.]



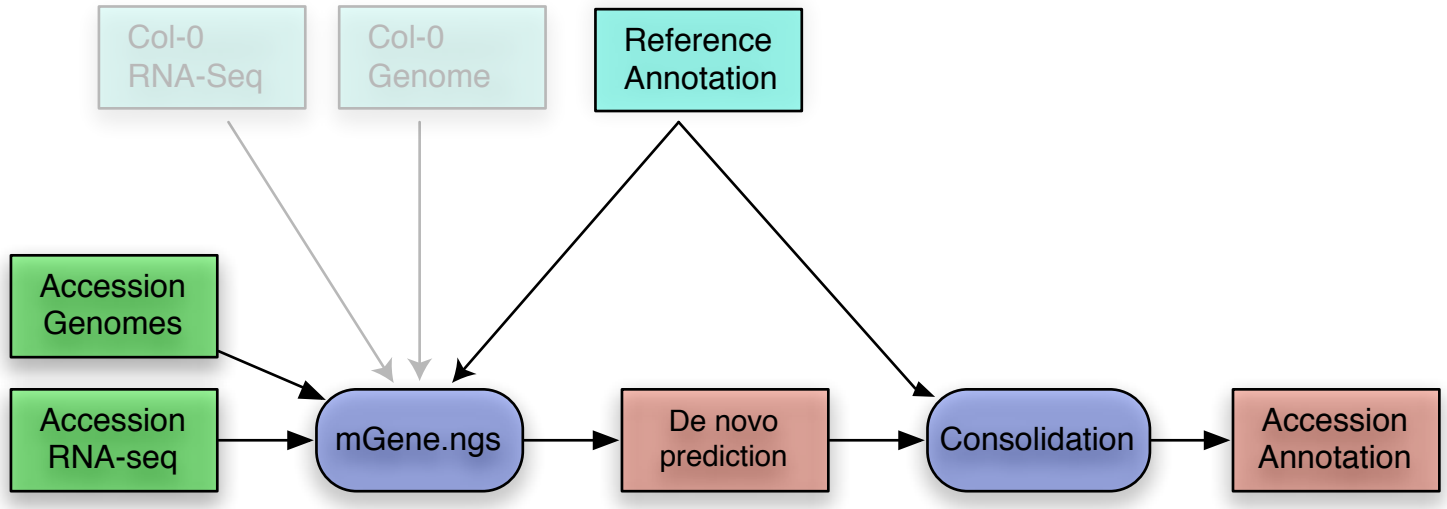
	CDS Exon				CDS Transcript			
	# predicted	SN (%)	SP (%)	F (%)	# predicted	SN (%)	SP (%)	F (%)
mGene	146,241	84.8	85.5	85.2	26,649	56.5	63.0	59.6
mGene.ngs	136,391	84.3	91.2	87.6	25,077	57.6	75.5	65.2
Cufflinks	94,921	53.6	88.1	66.7	16,811	28.6	52.0	36.9
Cufflinks ORF≥100	90,176	52.5	90.7	66.5	14,517	27.5	58.8	37.5
Cufflinks ORF≥300								

Genome Reannotation



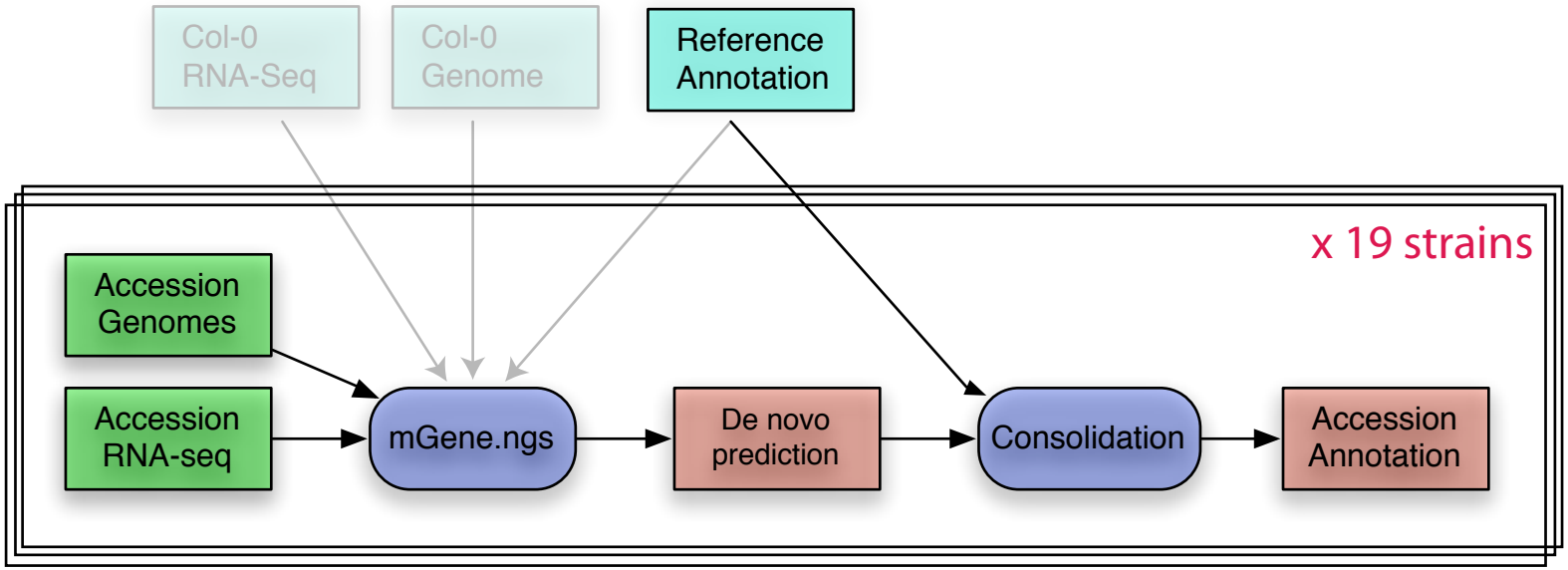
- Training on col-0 reference
- Strain-specific predictions and coordinate system

Genome Reannotation



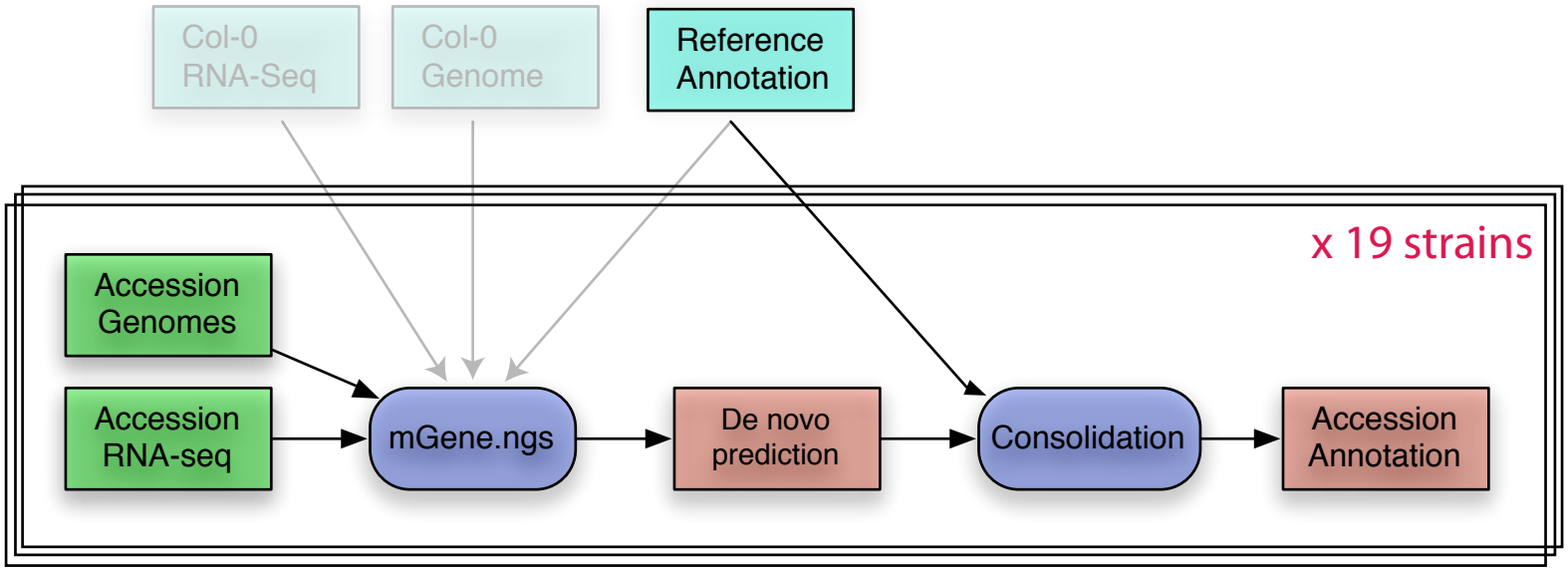
- Training on col-0 reference
- Strain-specific predictions and coordinate system

Genome Reannotation



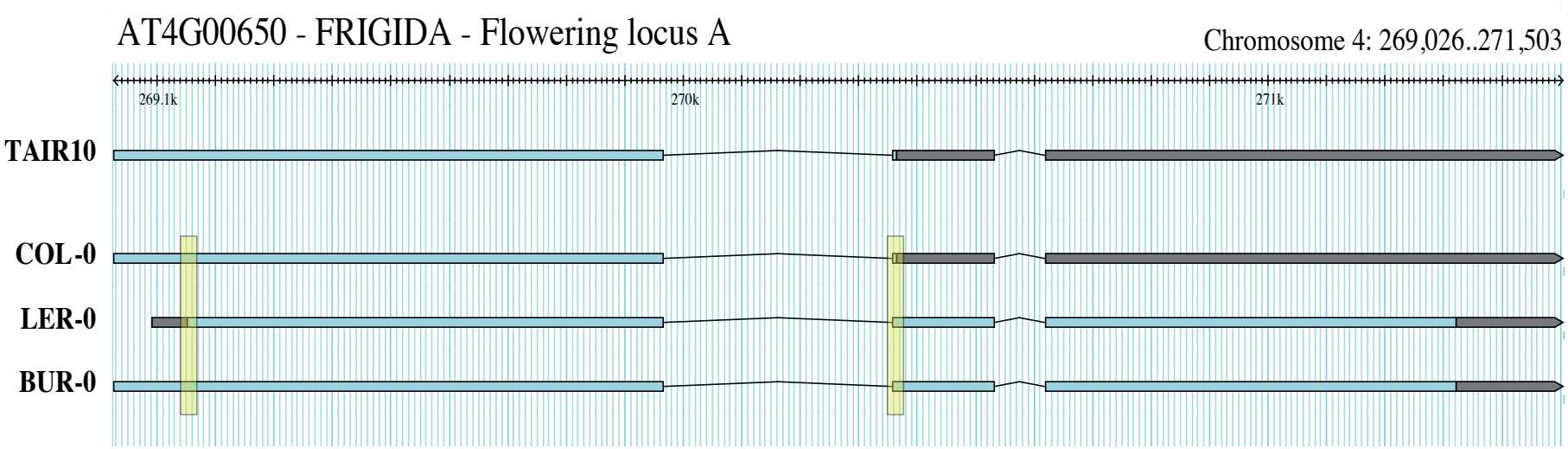
- Training on col-0 reference
- Strain-specific predictions and coordinate system

Genome Reannotation

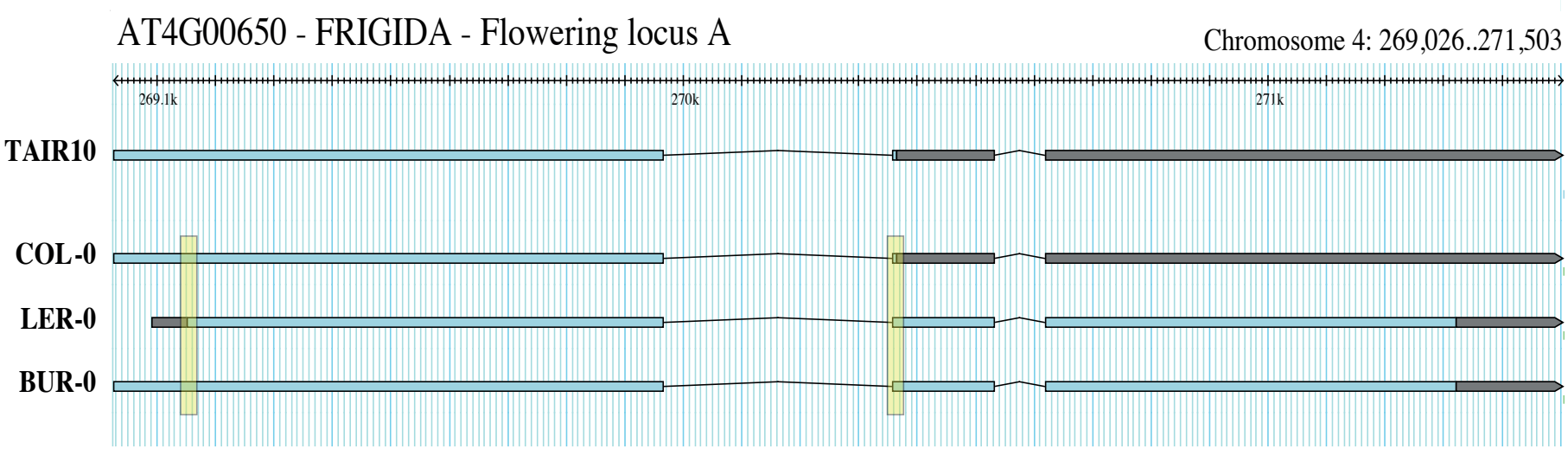


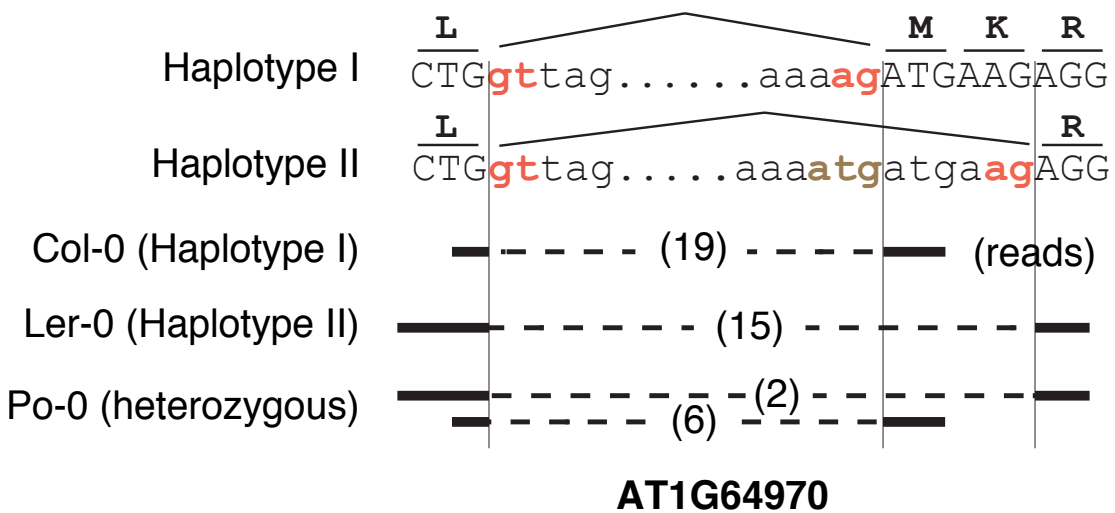
	Total		Novel	
Type	Per accession	RNA-seq confirmed (%)	Per accession	RNA-seq confirmed (%)
Genes	33,197	62.7	319	88.4
Transcripts	42,338	59.9	2,316	84.2
Introns	127,640	83.3	1,345	72.1
Start codons	33,264	n.a.	503	n.a.
Stop codons	33,720	n.a.	528	n.a.
Intron retentions	1,192	78.1	873	76.5
Exon skips	80	80.5	38	76.7

Genome Reannotation

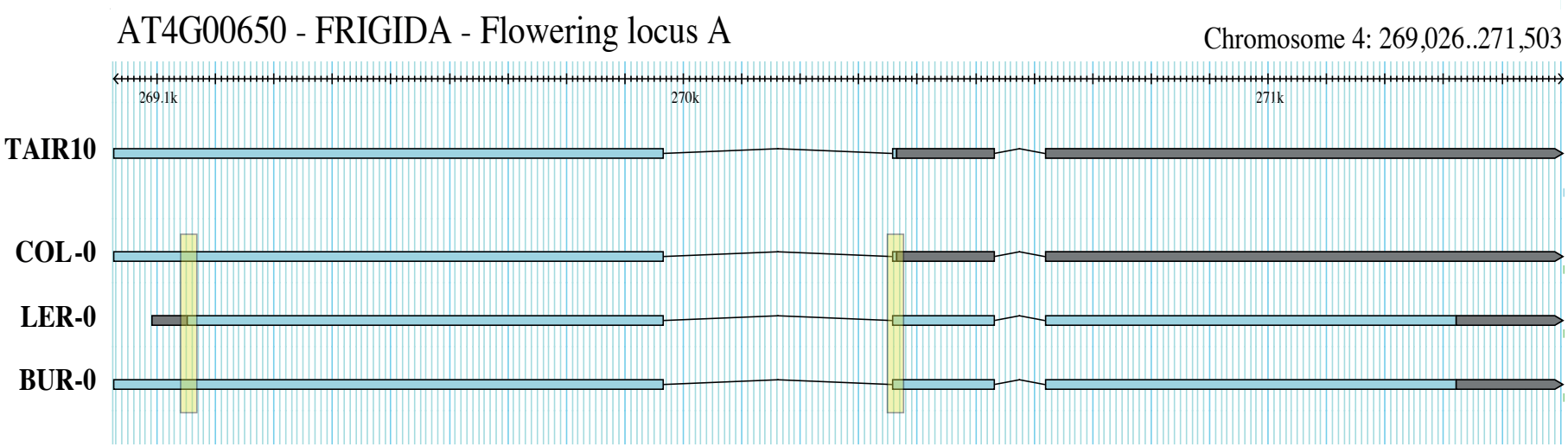


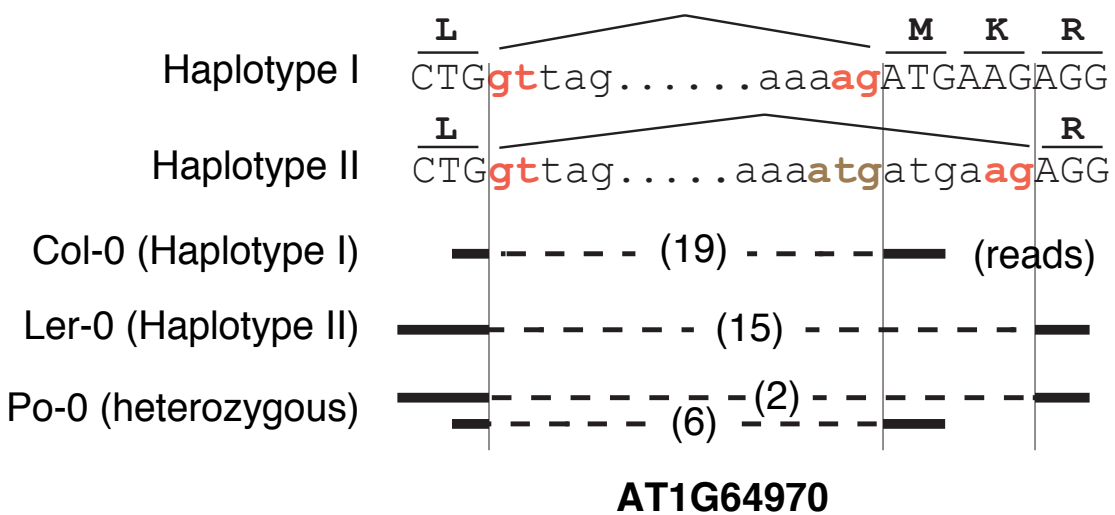
Genome Reannotation

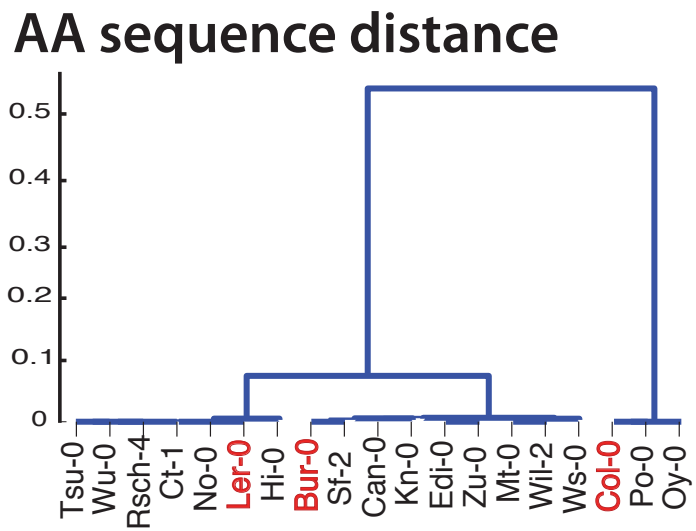




Genome Reannotation

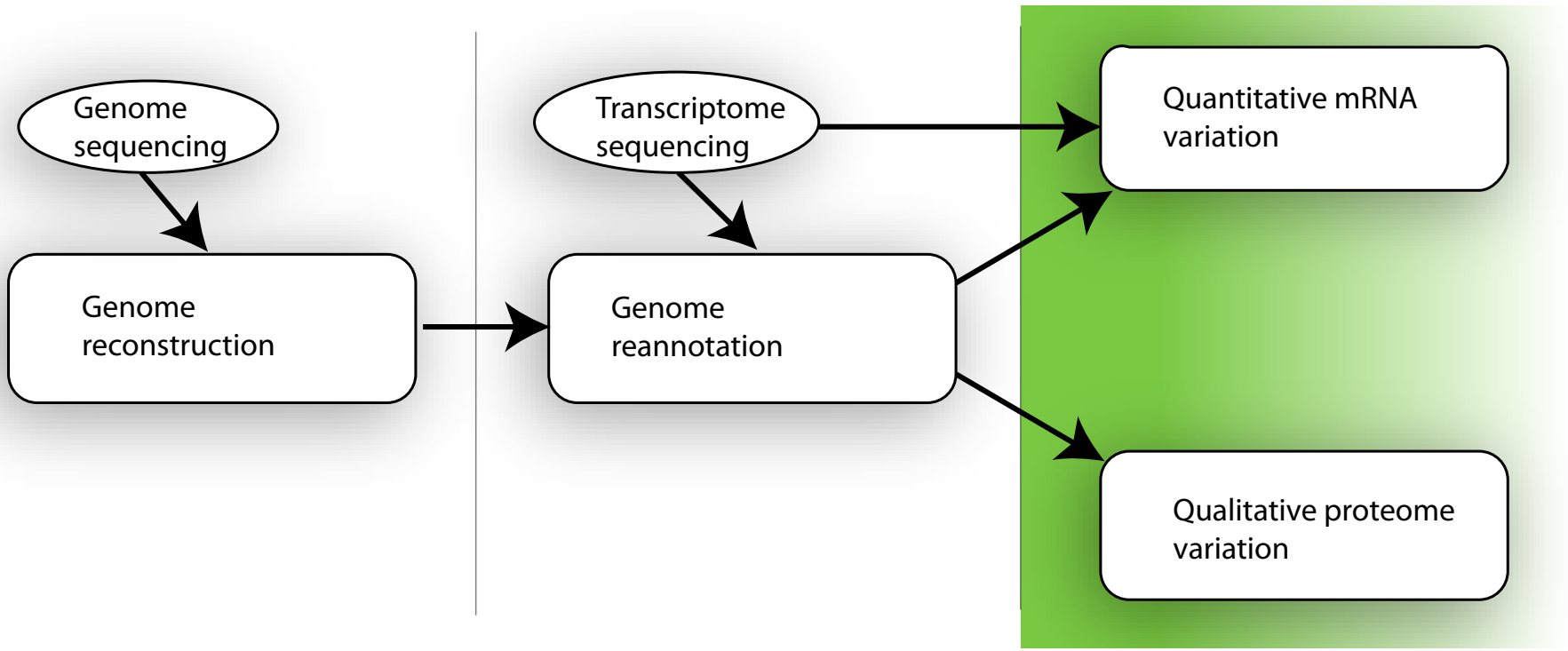




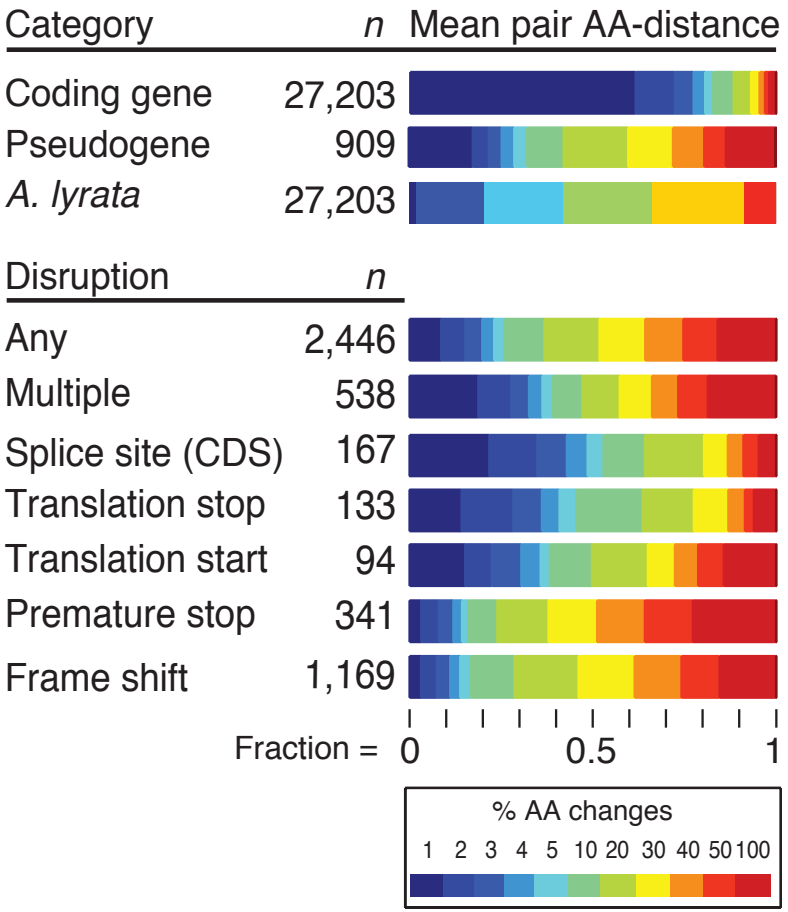


18+1 Reference Genomes and Transcriptomes

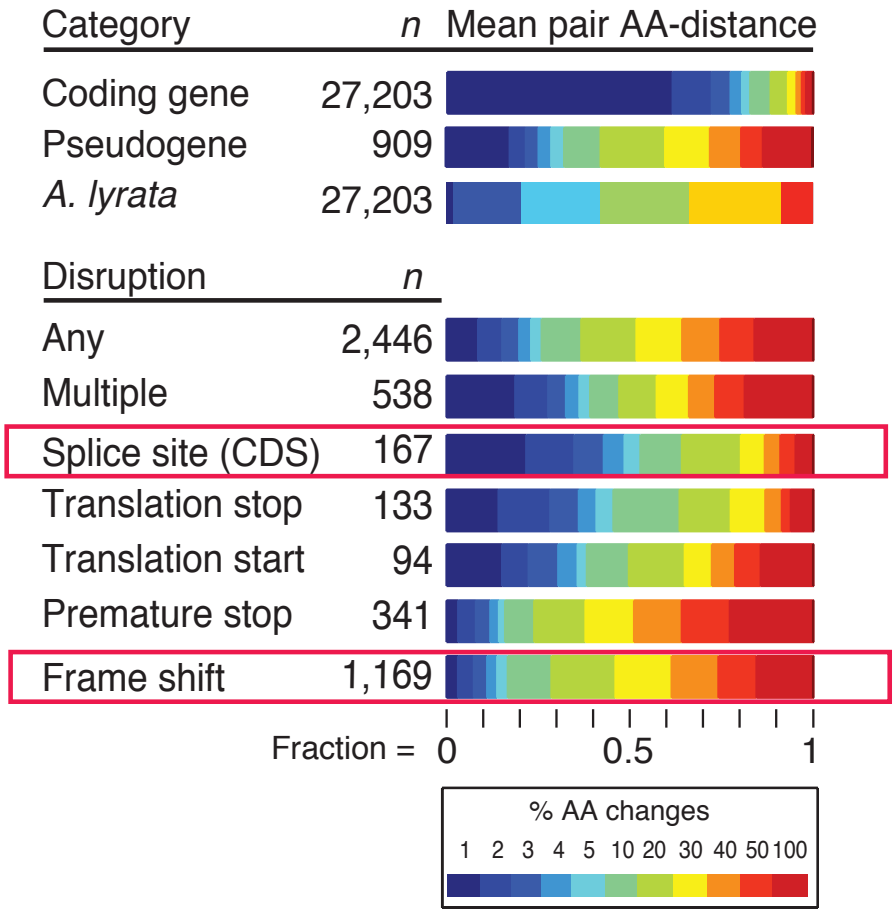
Goal: Characterize genome, transcriptome and proteome variation in 19 strains of the plant *Arabidopsis thaliana*

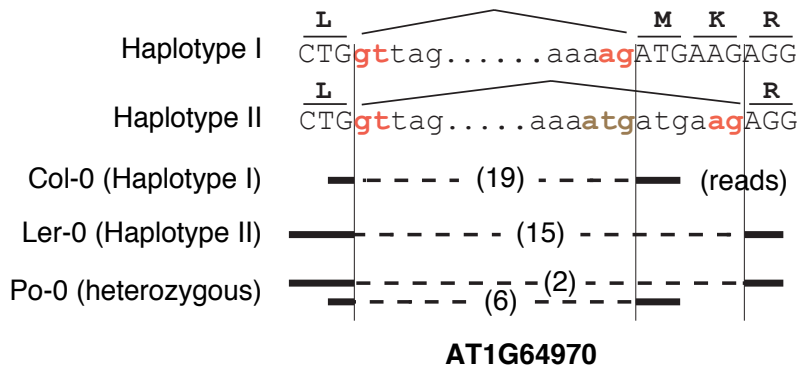


Amino Acid Sequence Variability

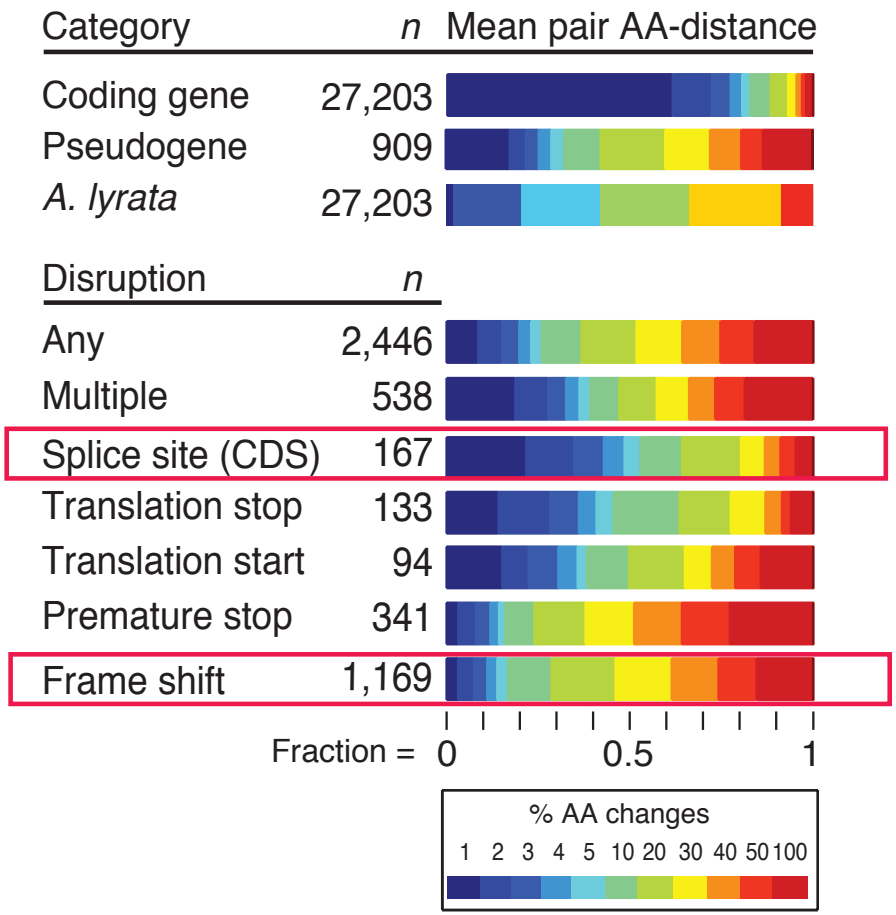


Amino Acid Sequence Variability

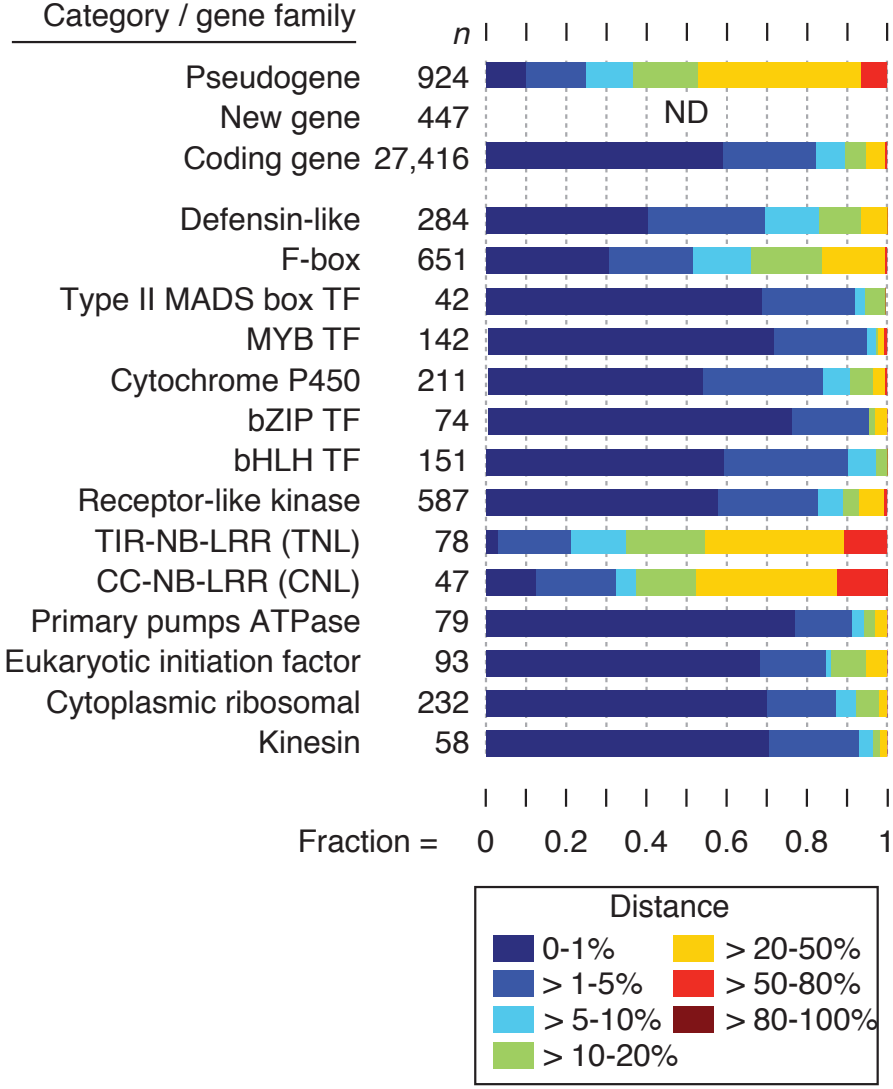




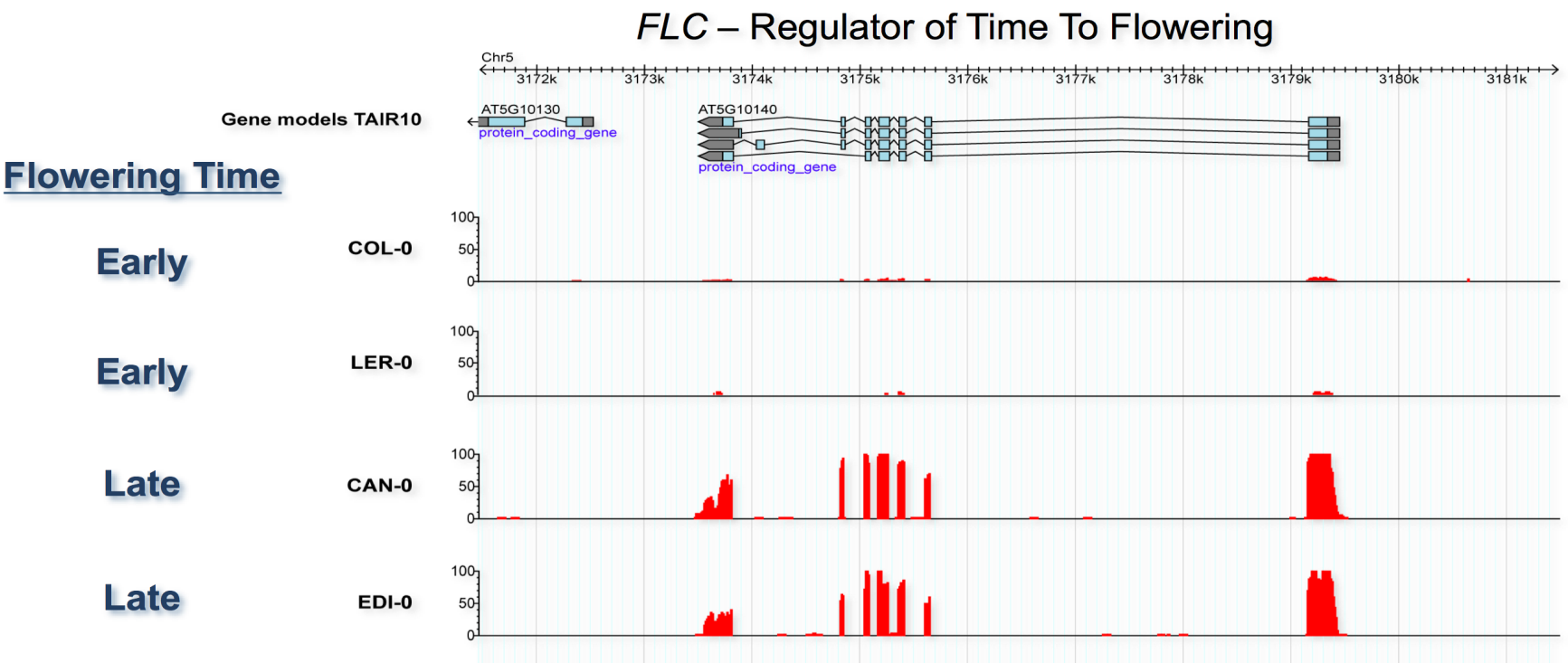
Amino Acid Sequence Variability



Isoform variability



Gene Expression Variability

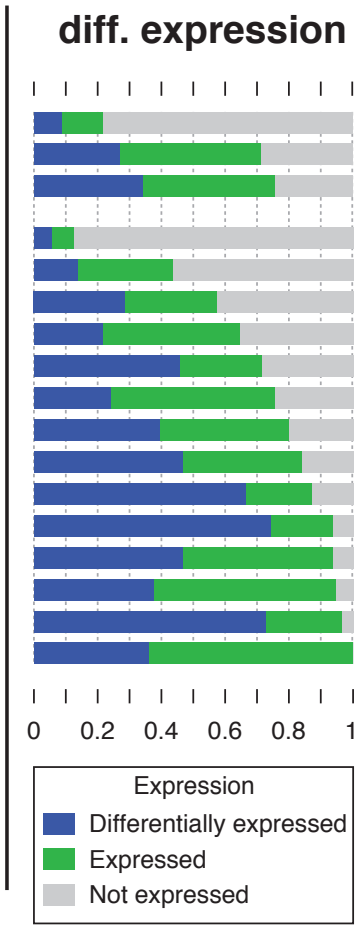


- Mapping of RNA-seq reads against private reference genomes
- ≈19,000 genes with detectable expression

Gene Expression Variability

Category / gene family	*n*
Pseudogene	924
New gene	447
Coding gene	27,416
Defensin-like	284
F-box	651
Type II MADS box TF	42
MYB TF	142
Cytochrome P450	211
bZIP TF	74
bHLH TF	151
Receptor-like kinase	587
TIR-NB-LRR (TNL)	78
CC-NB-LRR (CNL)	47
Primary pumps ATPase	79
Eukaryotic initiation factor	93
Cytoplasmic ribosomal	232
Kinesin	58

Fraction =

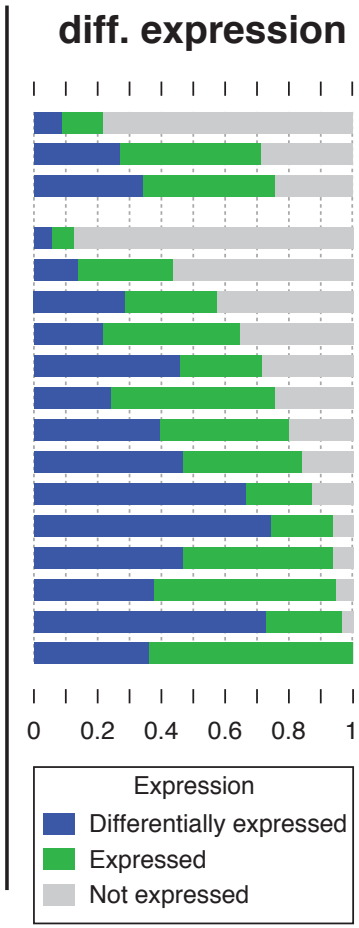


- ≈19,000 genes with detectable expression
  - 9,360 differentially expressed genes
  - 1,712 with fold-change ≥ 10
  - 308 with fold-change ≥ 50
- Causes of differential expression?

Gene Expression Variability

Category / gene family	*n*
Pseudogene	924
New gene	447
Coding gene	27,416
Defensin-like	284
F-box	651
Type II MADS box TF	42
MYB TF	142
Cytochrome P450	211
bZIP TF	74
bHLH TF	151
Receptor-like kinase	587
TIR-NB-LRR (TNL)	78
CC-NB-LRR (CNL)	47
Primary pumps ATPase	79
Eukaryotic initiation factor	93
Cytoplasmic ribosomal	232
Kinesin	58

Fraction =

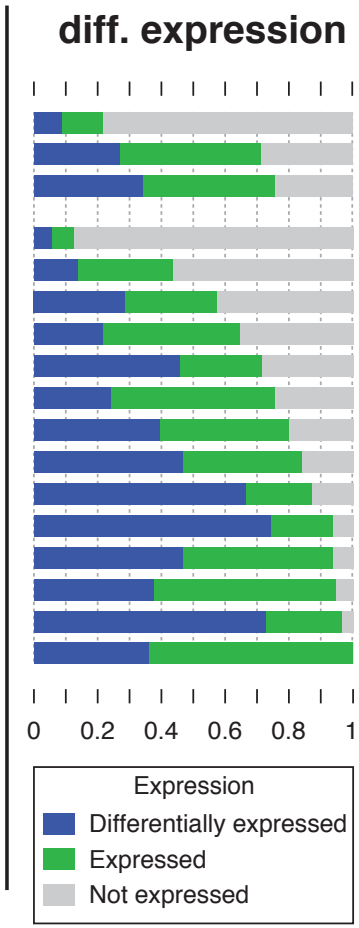


- ≈19,000 genes with detectable expression
  - 9,360 differentially expressed genes
  - 1,712 with fold-change ≥ 10
  - 308 with fold-change ≥ 50
- Causes of differential expression?

Gene Expression Variability

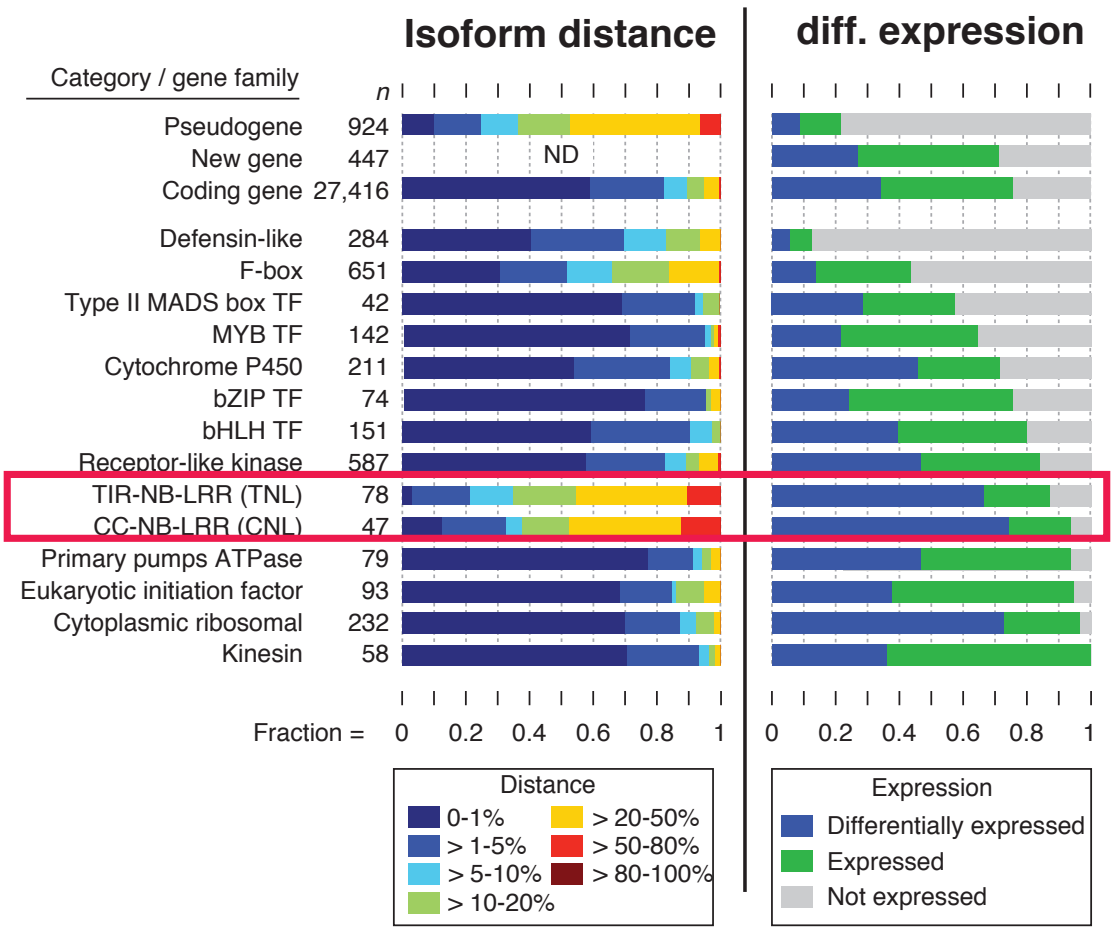
Category / gene family	*n*
Pseudogene	924
New gene	447
Coding gene	27,416
Defensin-like	284
F-box	651
Type II MADS box TF	42
MYB TF	142
Cytochrome P450	211
bZIP TF	74
bHLH TF	151
Receptor-like kinase	587
TIR-NB-LRR (TNL)	78
CC-NB-LRR (CNL)	47
Primary pumps ATPase	79
Eukaryotic initiation factor	93
Cytoplasmic ribosomal	232
Kinesin	58

Fraction =



- ≈19,000 genes with detectable expression
  - 9,360 differentially expressed genes
  - 1,712 with fold-change ≥ 10
  - 308 with fold-change ≥ 50
- Causes of differential expression?

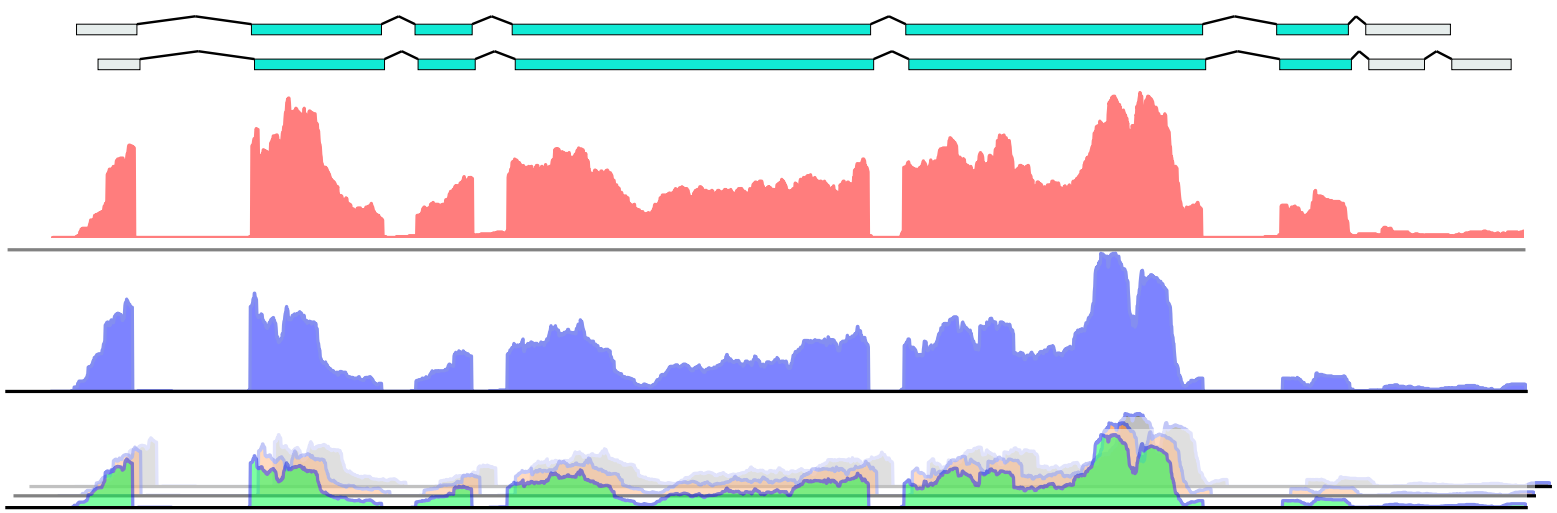
Gene Expression Variability



- ≈19,000 genes with detectable expression
  - 9,360 differentially expressed genes
  - 1,712 with fold-change ≥ 10
  - 308 with fold-change ≥ 50
- **Causes of differential expression?**

Gene Expression Variability

Structural Variation

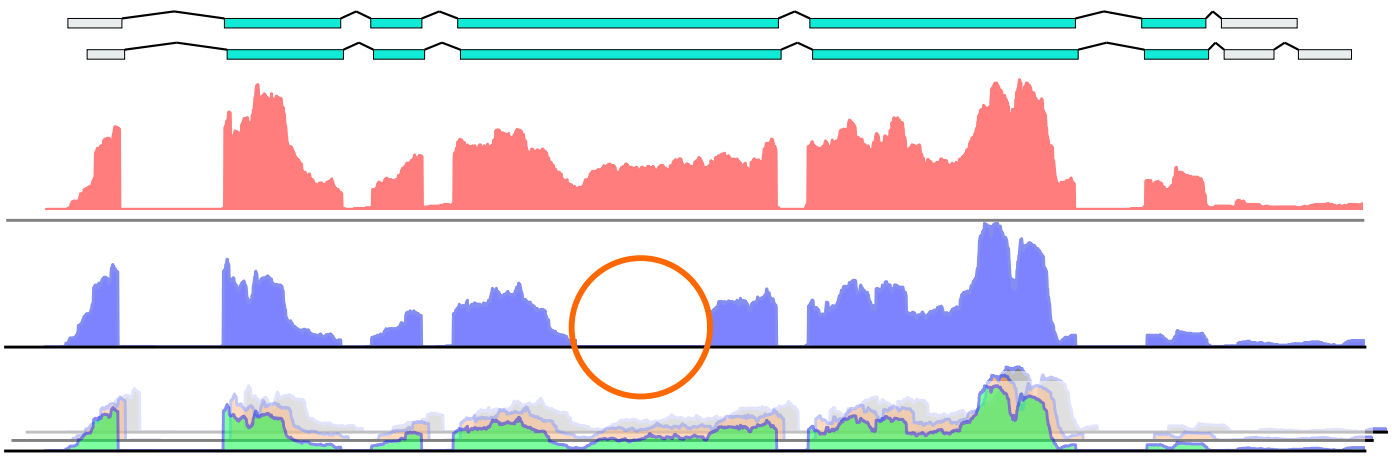


[DESeq, Anders & Huber, 2010]

[rDiff, Stegle & Drewe et al. 2010]

Gene Expression Variability

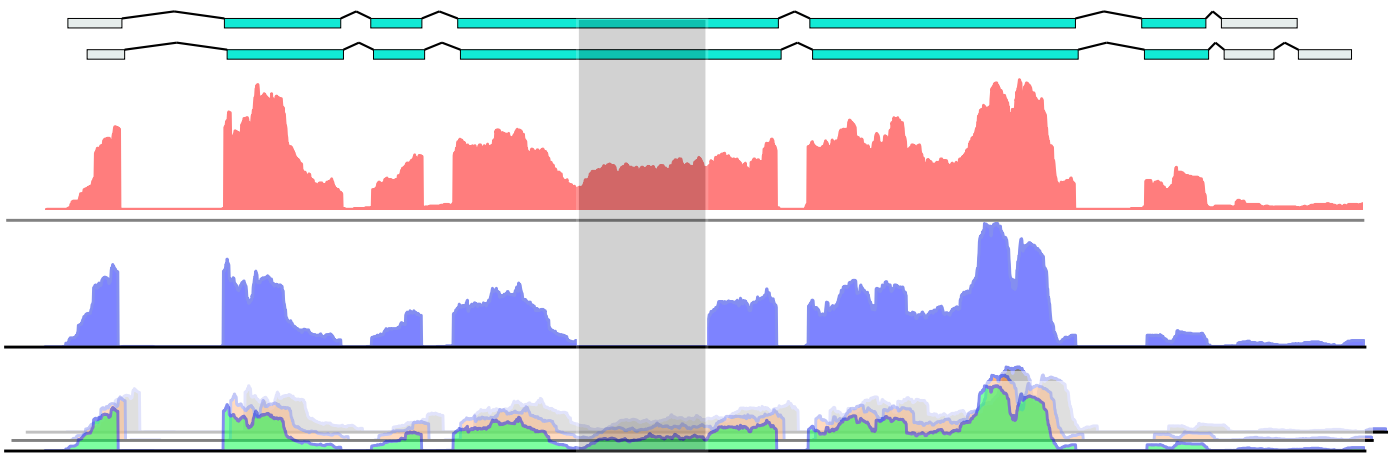
Structural Variation



Strain-specific deletion (genome reconstruction)

Gene Expression Variability

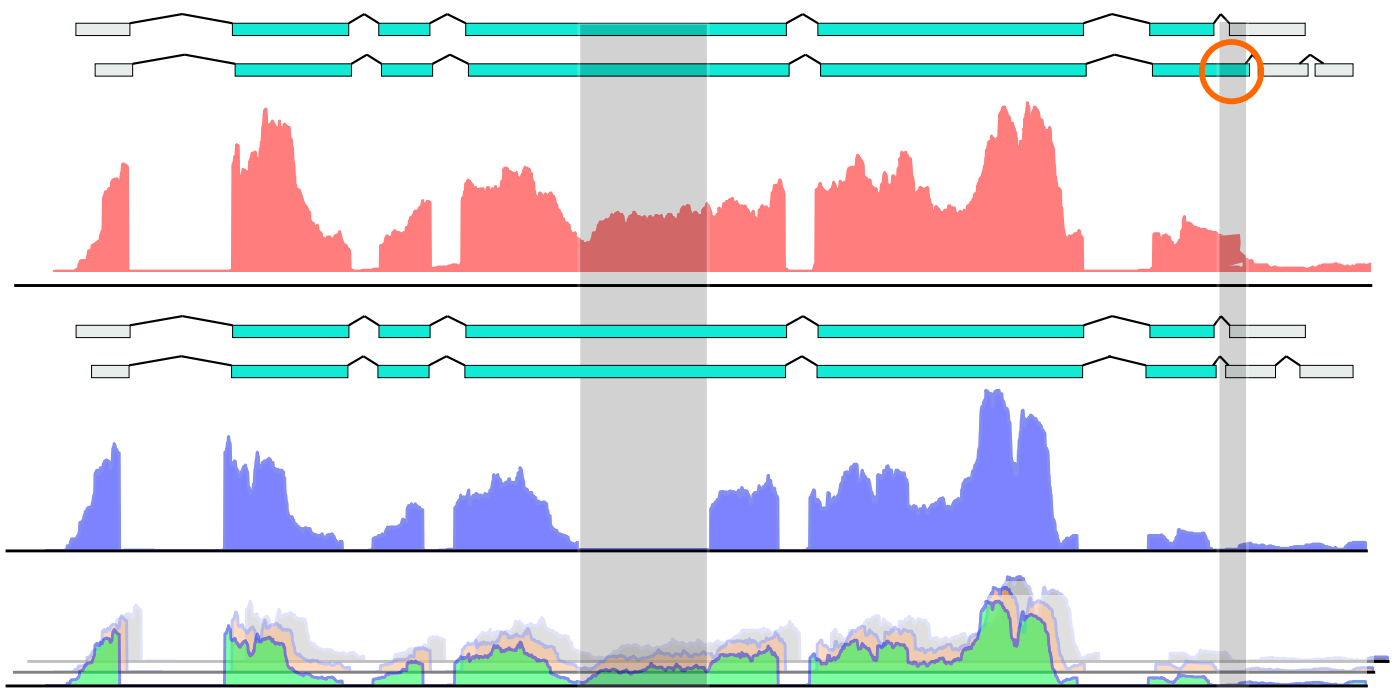
Structural Variation



Strain-specific deletion (genome reconstruction)

Gene Expression Variability

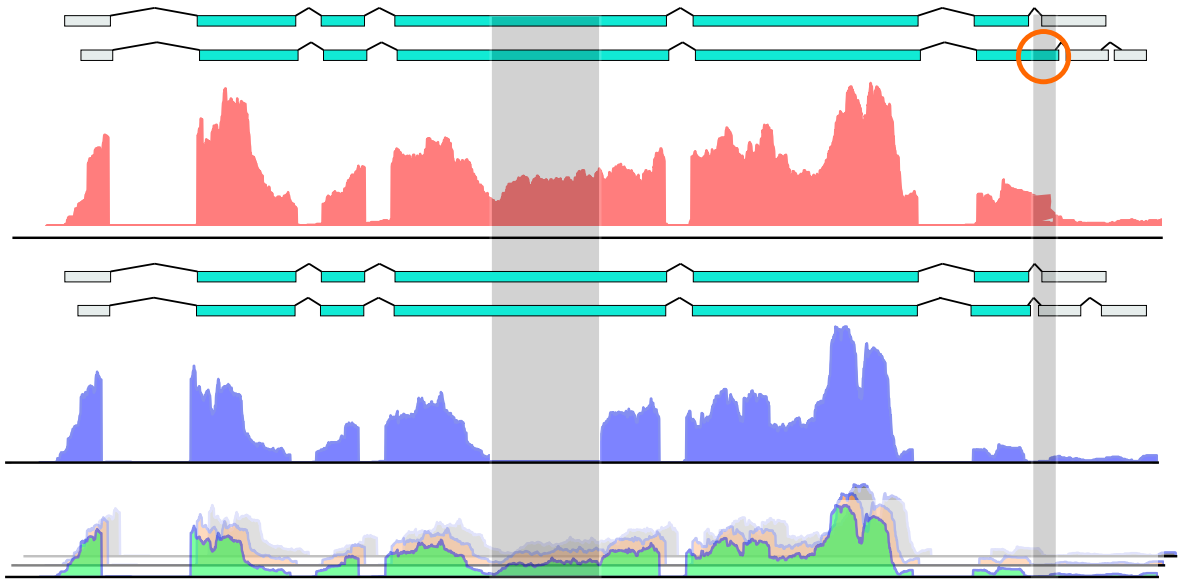
Structural Variation

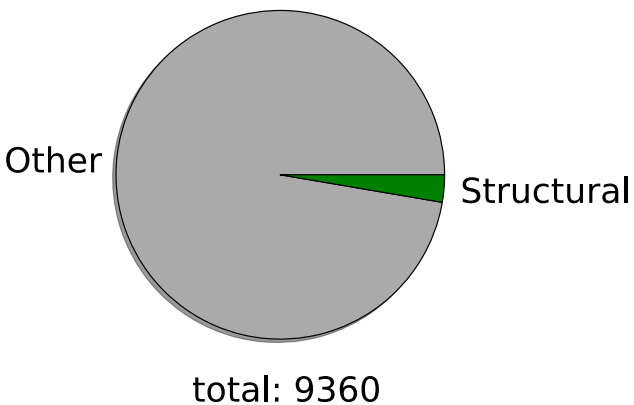


Strain-specific gene model change (genome reannotation)

Gene Expression Variability

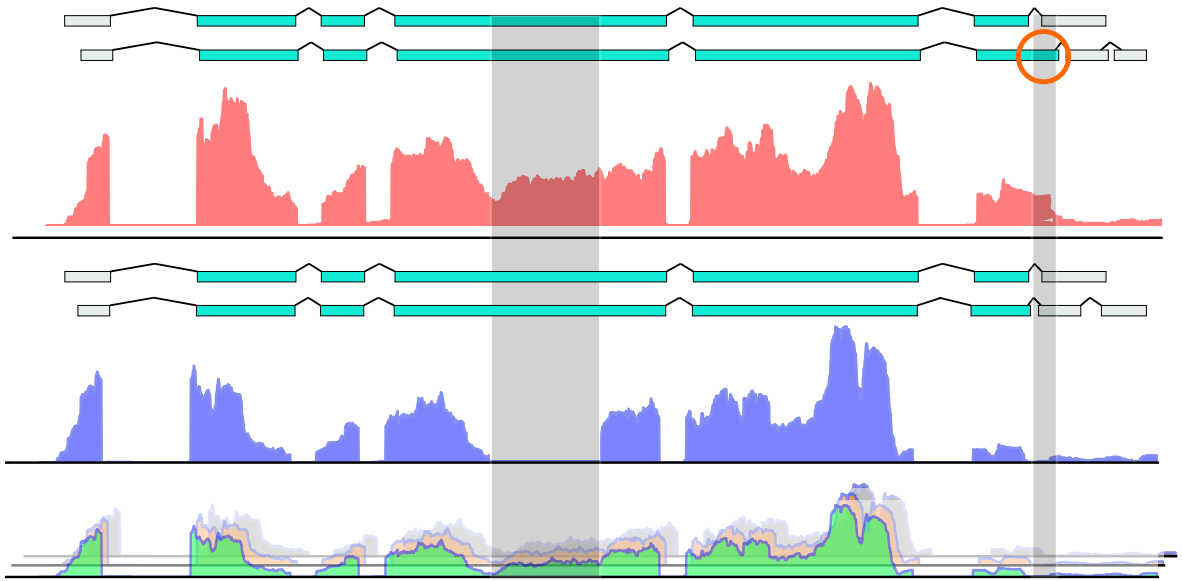
Structural Variation

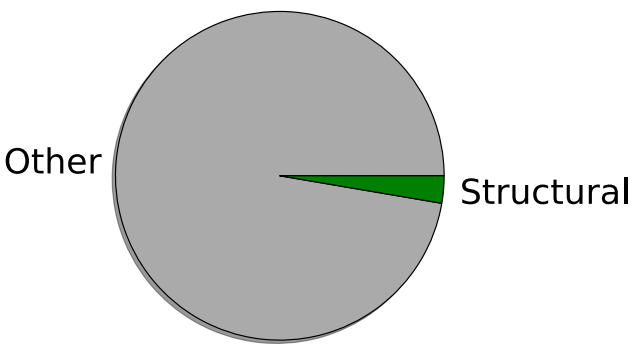




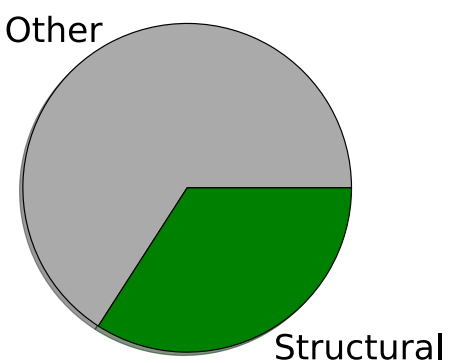
Gene Expression Variability

Structural Variation





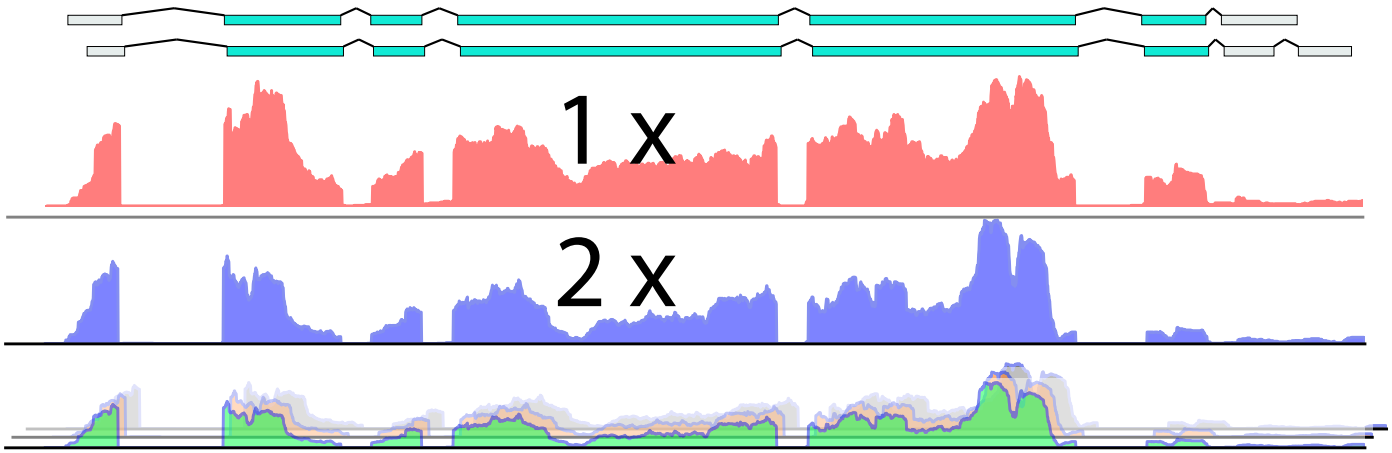
total: 9360

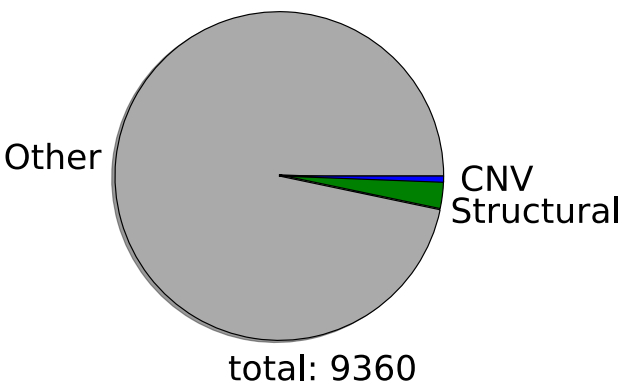


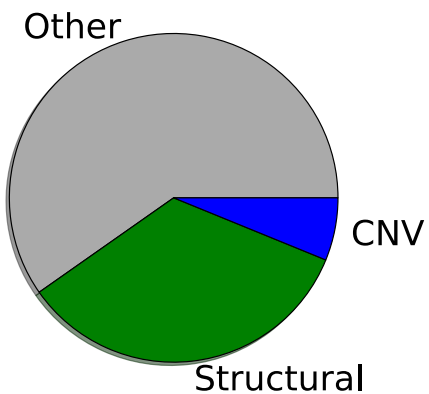
(FC > 50): 308

Gene expression variability

Copy Number Variation



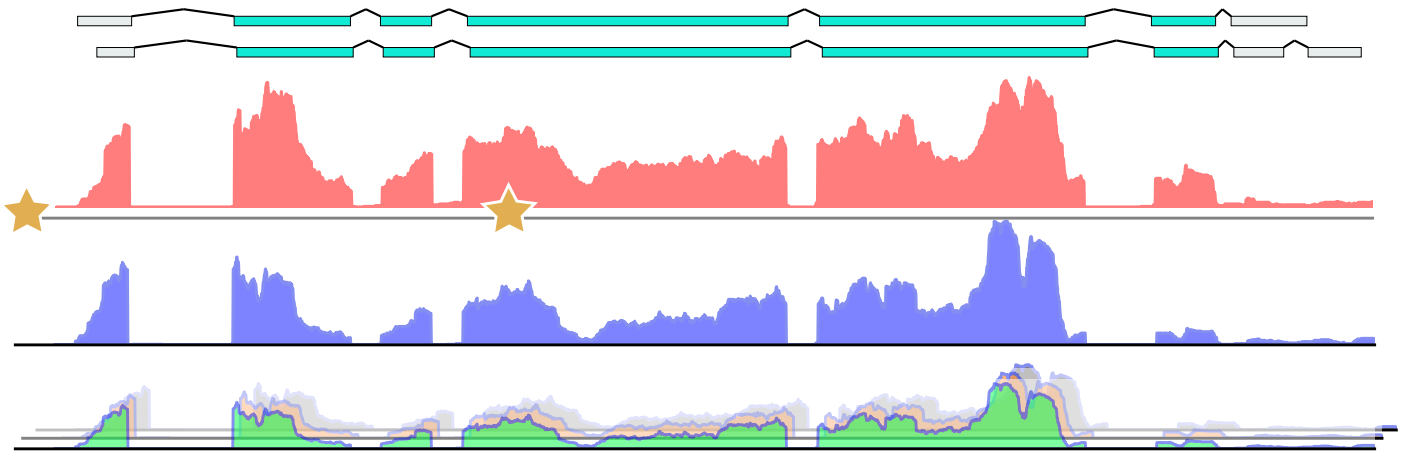




(FC > 50): 308

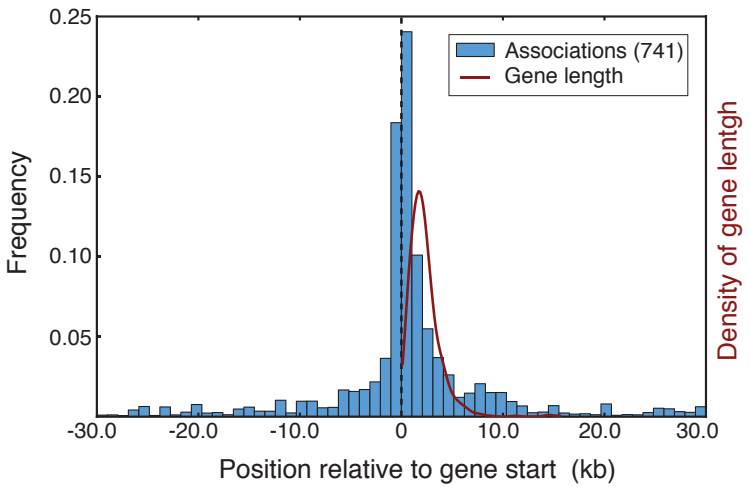
Gene expression variability

Mappable cis eQTLs



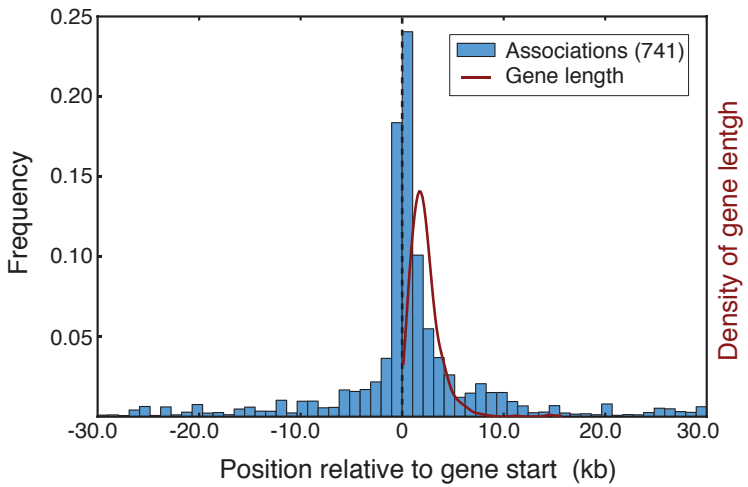
Gene expression variability

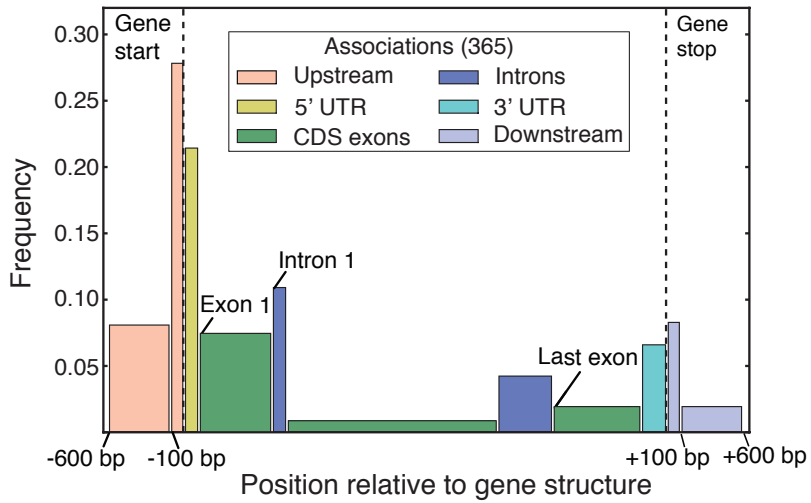
Mappable cis eQTLs



Gene expression variability

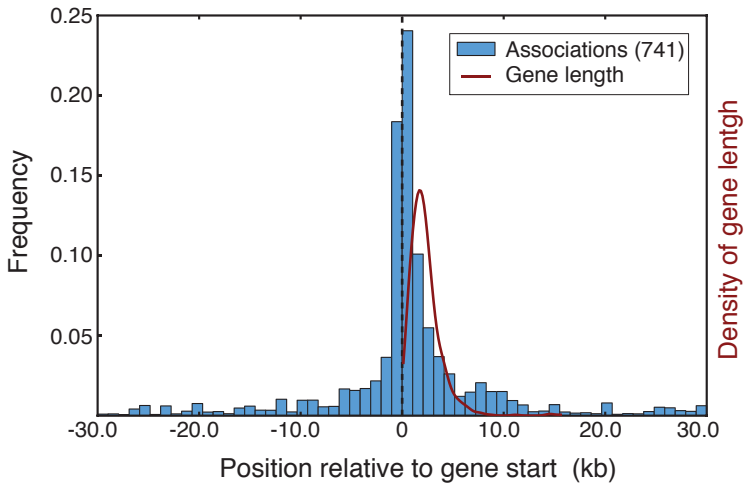
Mappable cis eQTLs

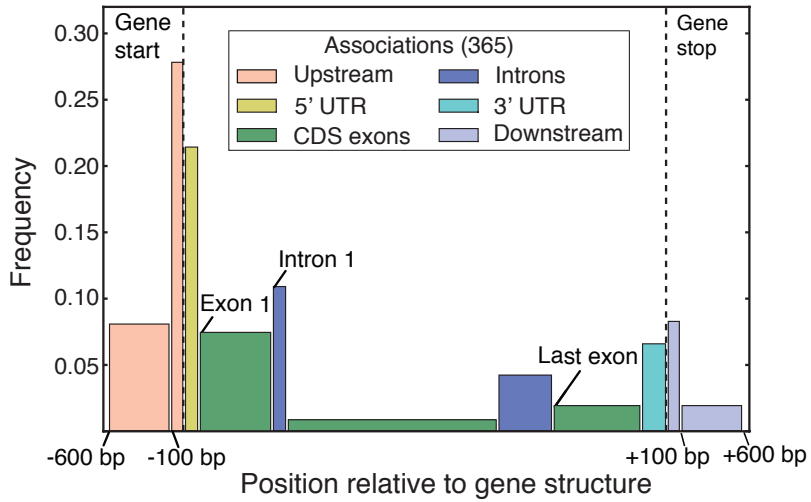


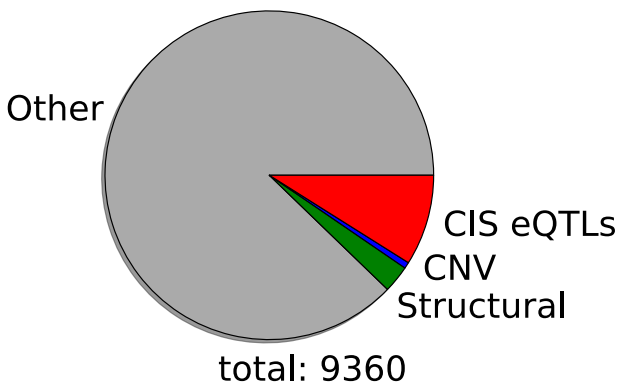


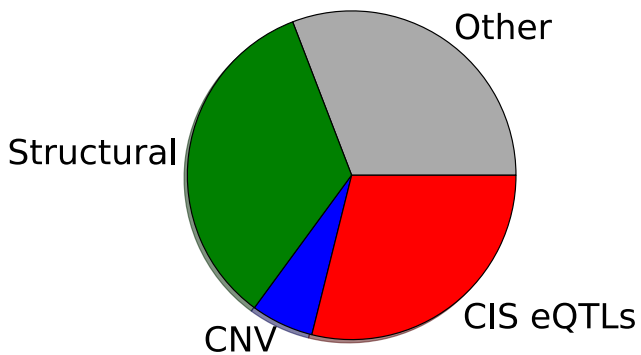
Gene expression variability

Mappable cis eQTLs



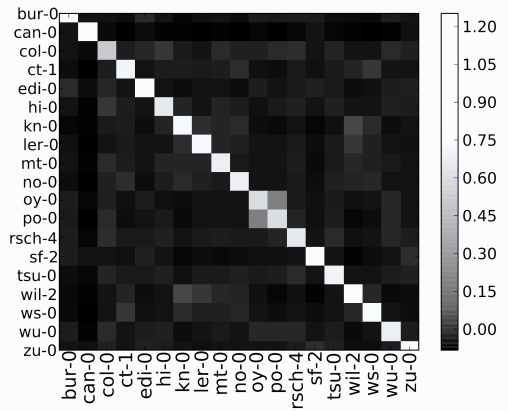






(FC > 50): 308

Gene expression variability
Population structure & unknown confounding

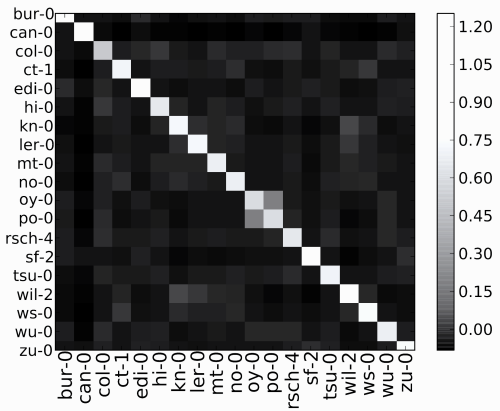


Variance component model

$$\mathbf{y}_g \sim \mathcal{N}(0,\ \underbrace{\alpha_p^2\mathbf{K}_p}_{\text{population}} + \underbrace{\alpha_c^2\mathbf{K}_c}_{\text{confounding}} + \underbrace{\alpha_s^2\mathbf{K}_s}_{\text{cis}} + \sigma^2\mathbf{I})$$

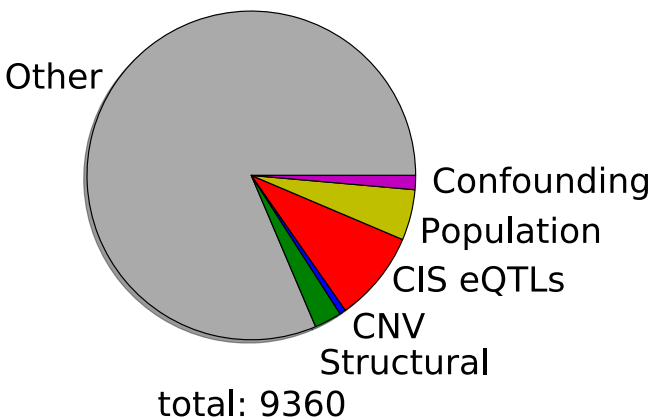
Stegle & Parts et al. 2010
Gan et al., Nature, 2011

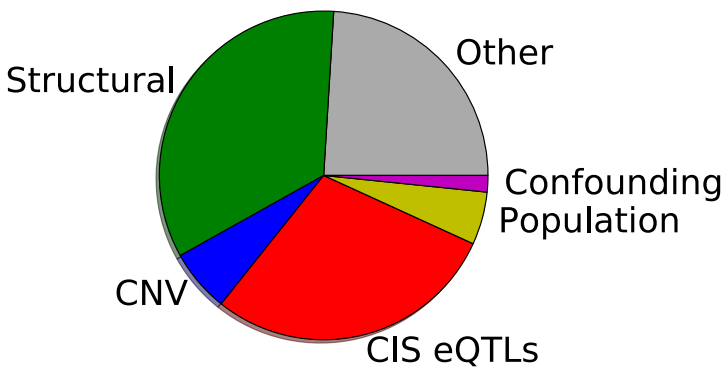
Gene expression variability
Population structure & unknown confounding



Variance component model

$$\mathbf{y}_g \sim \mathcal{N}(0,\ \underbrace{\alpha_p^2\mathbf{K}_p}_{\text{population}} + \underbrace{\alpha_c^2\mathbf{K}_c}_{\text{confounding}} + \underbrace{\alpha_s^2\mathbf{K}_s}_{\text{cis}} + \sigma^2\mathbf{I})$$





Stegle & Parts et al. 2010
Gan et al., Nature, 2011

Conclusions and future directions

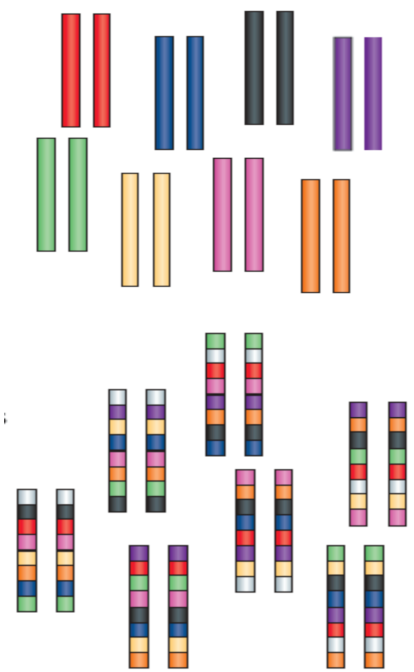
Catalogue of variation for 19 *Arabidopsis* MAGIC Lines founders:

- Accurate genome **reconstruction**
- De novo genome **reannotation**
- Detailed dissection of **qualitative** and **quantitative transcriptome** variation

Conclusions and future directions

Catalogue of variation for 19 *Arabidopsis* MAGIC Lines founders:

- Accurate genome **reconstruction**
- De novo genome **reannotation**
- Detailed dissection of **qualitative** and **quantitative** **transcriptome** variation

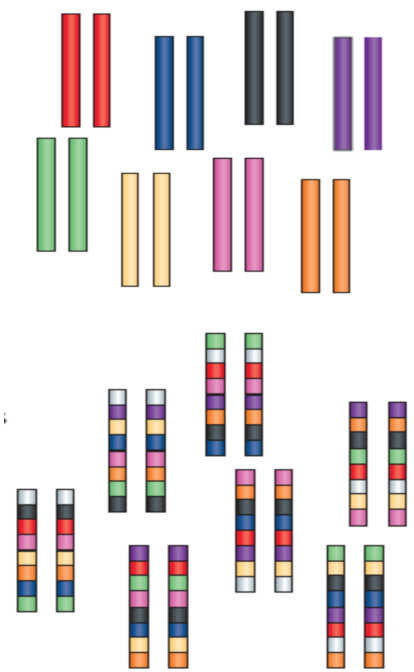


1. This study
   - Investigate magic founders
   - Analysis of *cis* variation
2. Bioinformatic tools to handle "personal" genomes and annotations
3. Ongoing/Future
   - Investigate magic lines themselves
   - Analysis of *trans* variation
   - Integration of methylation info

Conclusions and future directions

Catalogue of variation for 19 *Arabidopsis* MAGIC Lines founders:

- Accurate genome **reconstruction**
- De novo genome **reannotation**
- Detailed dissection of **qualitative** and **quantitative** **transcriptome** variation

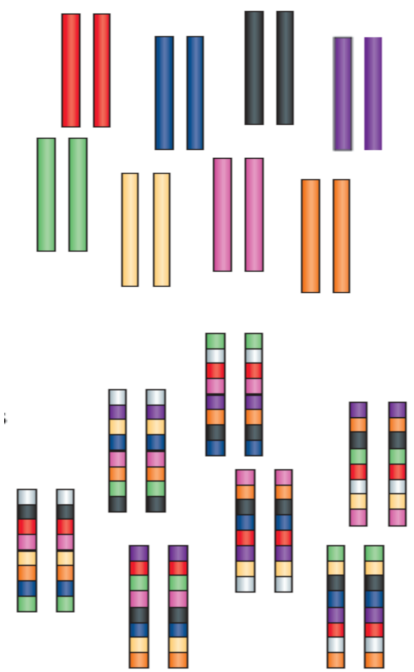


1. This study
   - Investigate magic founders
   - Analysis of *cis* variation
2. Bioinformatic tools to handle "personal" genomes and annotations
3. Ongoing/Future
   - Investigate magic lines themselves
   - Analysis of *trans* variation
   - Integration of methylation info

Conclusions and future directions

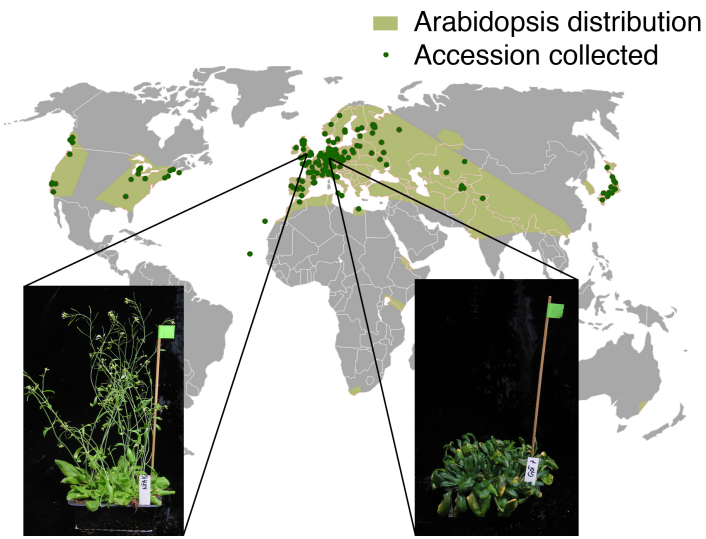
Catalogue of variation for 19 *Arabidopsis* MAGIC Lines founders:

- Accurate genome **reconstruction**
- De novo genome **reannotation**
- Detailed dissection of **qualitative** and **quantitative** **transcriptome** variation



1. This study
   - Investigate magic founders
   - Analysis of *cis* variation
2. Bioinformatic tools to handle "personal" genomes and annotations
3. Ongoing/Future
   - Investigate magic lines themselves
   - Analysis of *trans* variation
   - Integration of methylation info

Thank You!



More information and data:

`http://bioweb.me/19g`

Talk slides:

`http://raetschlab.org/lectures`

Looking for postdocs

`http://raetschlab.org/offers`

[Gan et al., Nature, 2011]

- University of Oxford
  Genome sequencing & reconstruction
  G. Xiangchao, N. Harberd, R. Mott
- University of Utah
  Transcriptome sequencing
  J. Steffen, C. Toomajian, R. Clark
- Max Planck Campus Tübingen
  Reannotation & transcriptomes
  O. Stegle, J. Behr, P. Drewe, A. Kahles, R. Bohnert, G. Jean, S. J. Schultheiss, G. Rätsch
- University of Bath: P. Kover
- Kansas State U: K. L. Hildebrand
- EBI: P. Derwent, P. Kersey
- The Sainsbury Laboratory: E. Kemen

References

J. Behr, G. Schweikert, J. Cao, F. De Bona, G. Zeller, S. Laubinger, S. Ossowski, K. Schneeberger, D. Weigel, and G. Rätsch. Rna-seq and tiling arrays for improved gene finding. Oral presentation at the CSHL Genome Informatics Meeting, September 2008. URL `http://www.fml.tuebingen.mpg.de/raetsch/lectures/RaetschGenomeInformatics08.pdf`.

RM Clark, G Schweikert, C Toomajian, S Ossowski, G Zeller, P Shinn, N Warthmann, TT Hu, G Fu, DA Hinds, H Chen, KA Frazer, DH Huson, B Schölkopf, M Nordborg, G Rätsch, JR Ecker, and D Weigel. Common sequence polymorphisms shaping genetic diversity in arabidopsis thaliana. *Science*, 317(5836):338–342, 2007. ISSN 1095-9203 (Electronic). doi: 10.1126/science.1138632.

Xiangchao Gan, Oliver Stegle, Jonas Behr, Joshua G Steffen, Philipp Drewe, Katie L Hildebrand, Rune Lyngsoe, Sebastian J Schultheiss, Edward J Osborne, Vipin T Sreedharan, André Kahles, Regina Bohnert, Géraldine Jean, Paul Derwent, Paul Kersey, Eric J Belfield, Nicholas P Harberd, Eric Kemen, Christopher Toomajian, Paula X Kover, Richard M Clark, Gunnar Rätsch, and Richard Mott. Multiple reference genomes and transcriptomes for arabidopsis thaliana. *Nature*, 477(7365):419–23, Sep 2011. doi: 10.1038/nature10414.

G. Rätsch and S. Sonnenburg. Accurate splice site detection for *Caenorhabditis elegans*. In K. Tsuda B. Schoelkopf and J.-P. Vert, editors, *Kernel Methods in Computational Biology*. MIT Press, 2004.

G. Rätsch, S. Sonnenburg, and B. Schölkopf. RASE: recognition of alternatively spliced exons in *C. elegans*. *Bioinformatics*, 21(Suppl. 1):i369–i377, June 2005.

References

Gabriele Schweikert, Alexander Zien, Georg Zeller, Jonas Behr, Christoph Dieterich, Cheng Soon Ong, Petra Philips, Fabio De Bona, Lisa Hartmann, Anja Bohlen, Nina Krüger, Sören Sonnenburg, and Gunnar Rätsch. mgene: Accurate svm-based gene finding with an application to nematode genomes. *Genome Research*, 2009. URL `http://genome.cshlp.org/content/early/2009/06/29/gr.090597.108.full.pdf+html`. Advance access June 29, 2009.

S. Sonnenburg, G. Rätsch, A. Jagota, and K.-R. Müller. New methods for splice-site recognition. In *Proc. International Conference on Artificial Neural Networks*, 2002.

Sören Sonnenburg, Alexander Zien, and Gunnar Rätsch. ARTS: Accurate Recognition of Transcription Starts in Human. *Bioinformatics*, 22(14):e472–480, 2006.

G Zeller, RM Clark, K Schneeberger, A Bohlen, D Weigel, and G Ratsch. Detecting polymorphic regions in arabidopsis thaliana with resequencing microarrays. *Genome Res*, 18 (6):918–929, 2008. ISSN 1088-9051 (Print). doi: 10.1101/gr.070169.107.

A. Zien, G. Rätsch, S. Mika, B. Schölkopf, T. Lengauer, and K.-R. Müller. Engineering Support Vector Machine Kernels That Recognize Translation Initiation Sites. *BioInformatics*, 16(9): 799–807, September 2000.