A spatial and temporal map of *C. elegans* gene expression

Gunnar Rätsch

Friedrich Miescher Laboratory

Max Planck Society, Tübingen

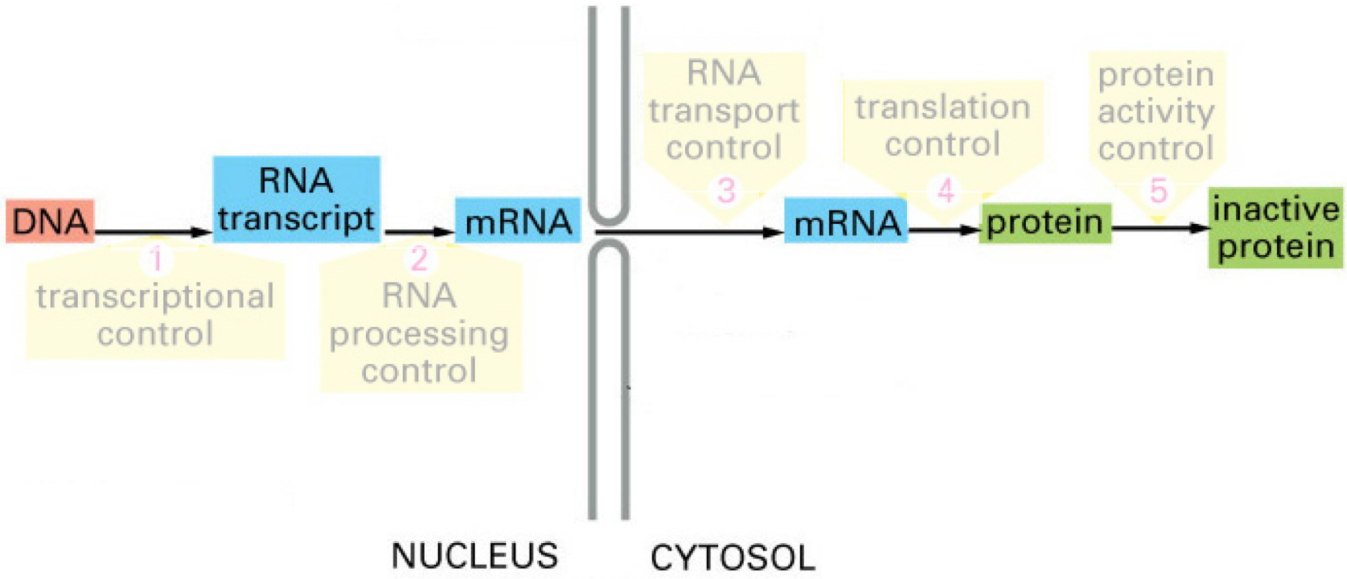
Sloan-Kettering Institute
New York City (August 30, 2011)



Friedrich Miescher Laboratory
of the Max Planck Society



Learning About the Central Dogma



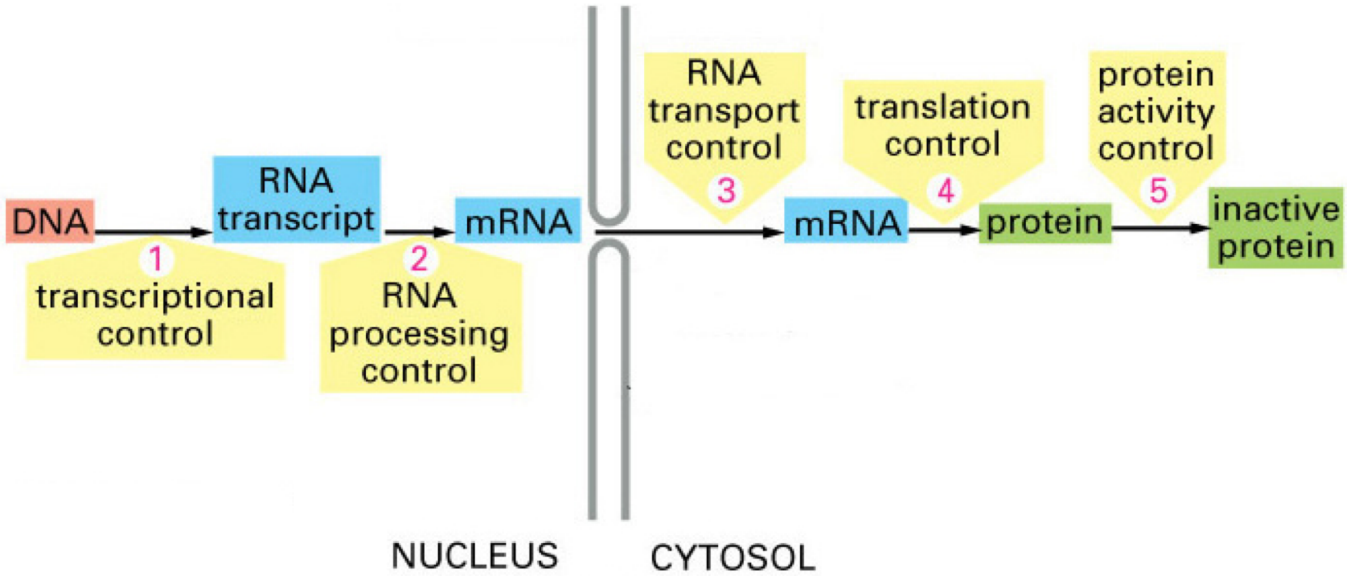
Goal: Learn to predict what these processes accomplish:

- Given the DNA, ..., predict all gene products

$$f(\text{DNA}, \quad) = \text{RNA} \qquad F(\text{RNA}, \quad) = \text{protein}$$

- Estimating f, F amounts to cracking the codes of transcription, epigenetics, splicing, ..., protein interactions, ...

Learning About the Central Dogma



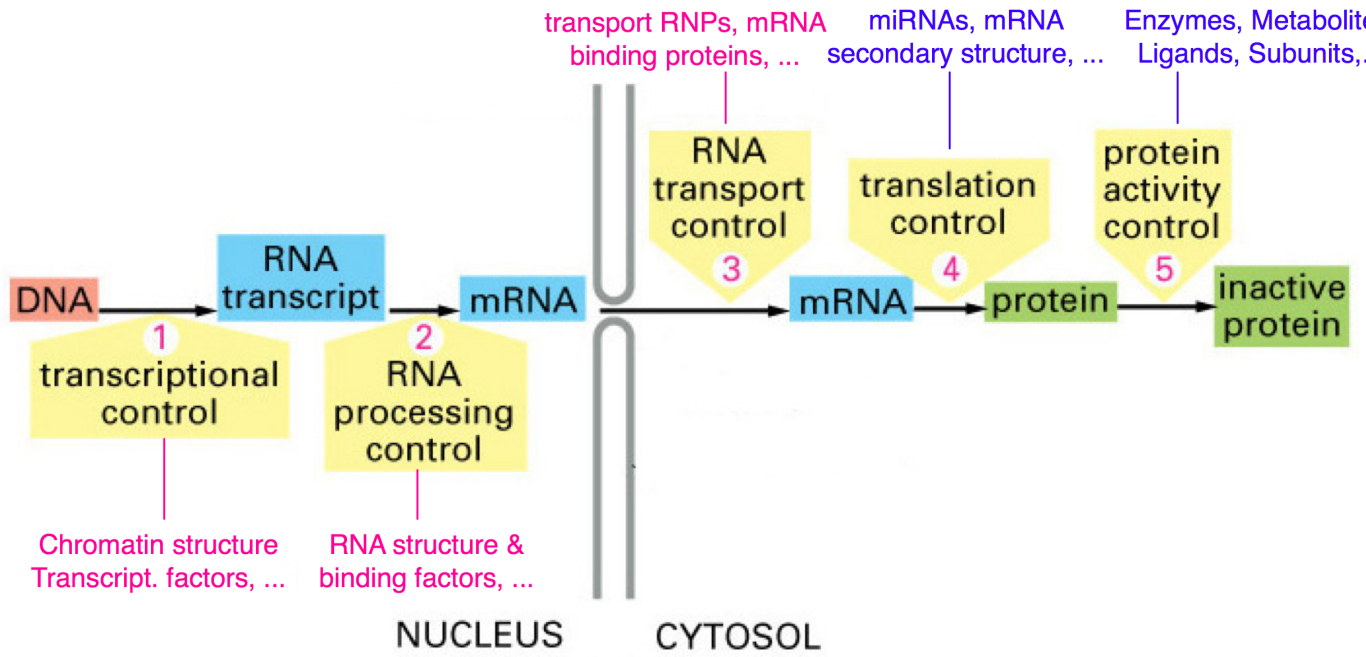
Goal: Learn to predict what these processes accomplish:

- Given the DNA, . . . , predict all gene products

$$f(\text{DNA}, \quad) = \text{RNA} \qquad F(\text{RNA}, \quad) = \text{protein}$$

- Estimating f, F amounts to cracking the codes of transcription, epigenetics, splicing, . . . , protein interactions, . . .

Learning About the Central Dogma



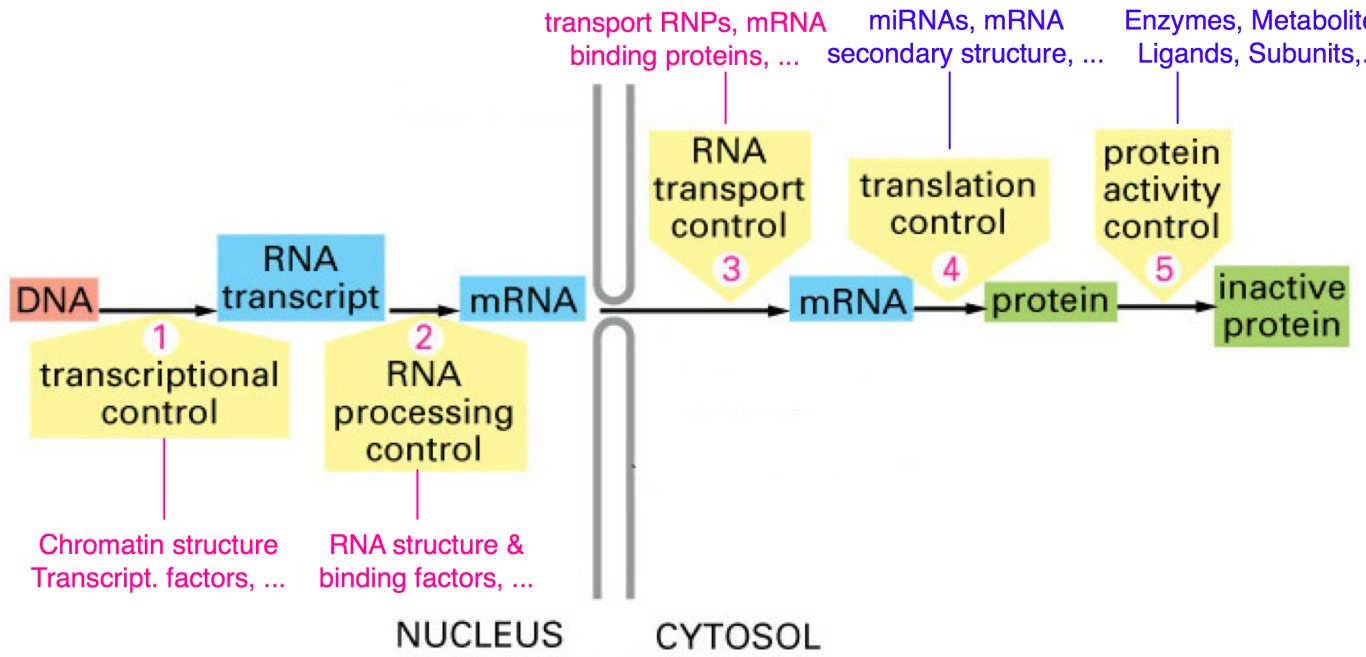
Goal: Learn to predict what these processes accomplish:

- Given the DNA, ..., predict all gene products

$$f(\text{DNA}, \boxed{1\ 2\ 3}) = \text{RNA} \qquad F(\text{RNA}, \boxed{4\ 5}) = \text{protein}$$

- Estimating f, F amounts to cracking the codes of transcription, epigenetics, splicing, ..., protein interactions, ...

Learning About the Central Dogma



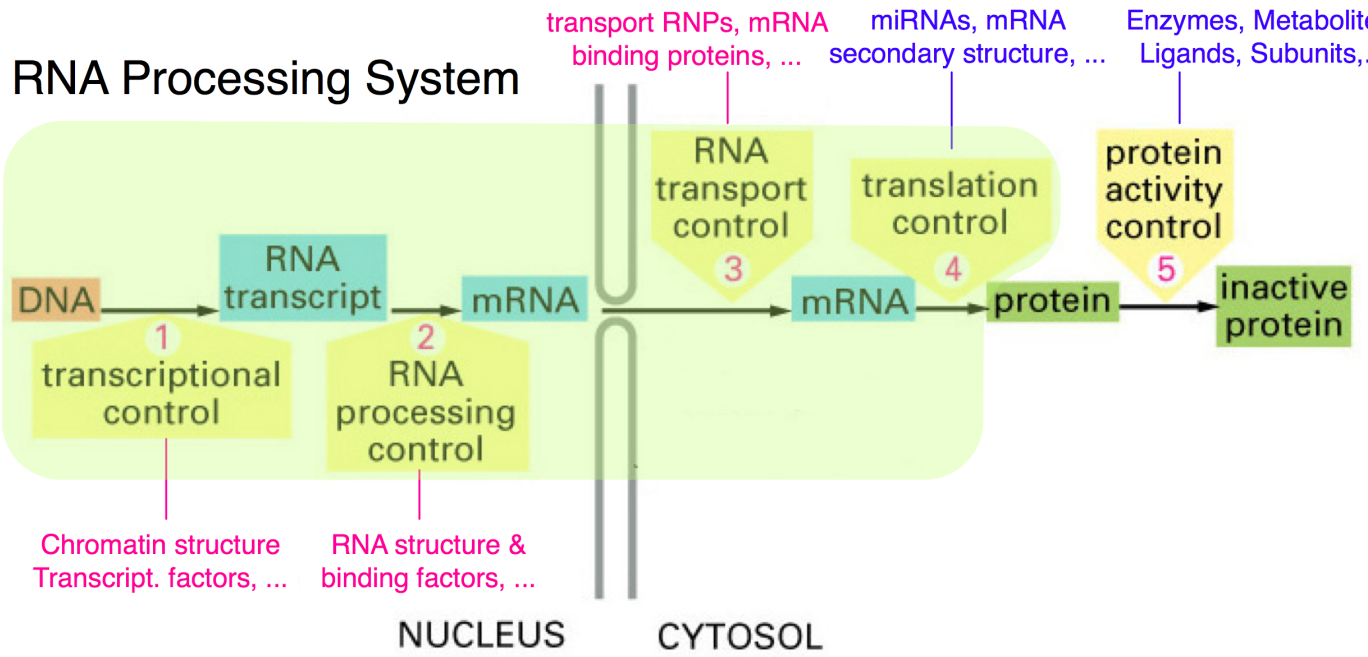
Goal: Learn to predict what these processes accomplish:

- Given the DNA, ..., predict all gene products

$$f(\text{DNA}, \boxed{1\ 2\ 3}) = \text{RNA} \qquad F(\text{RNA}, \boxed{4\ 5}) = \text{protein}$$

- Estimating f, F amounts to cracking the codes of transcription, epigenetics, splicing, ..., protein interactions, ...

Learning About the Central Dogma



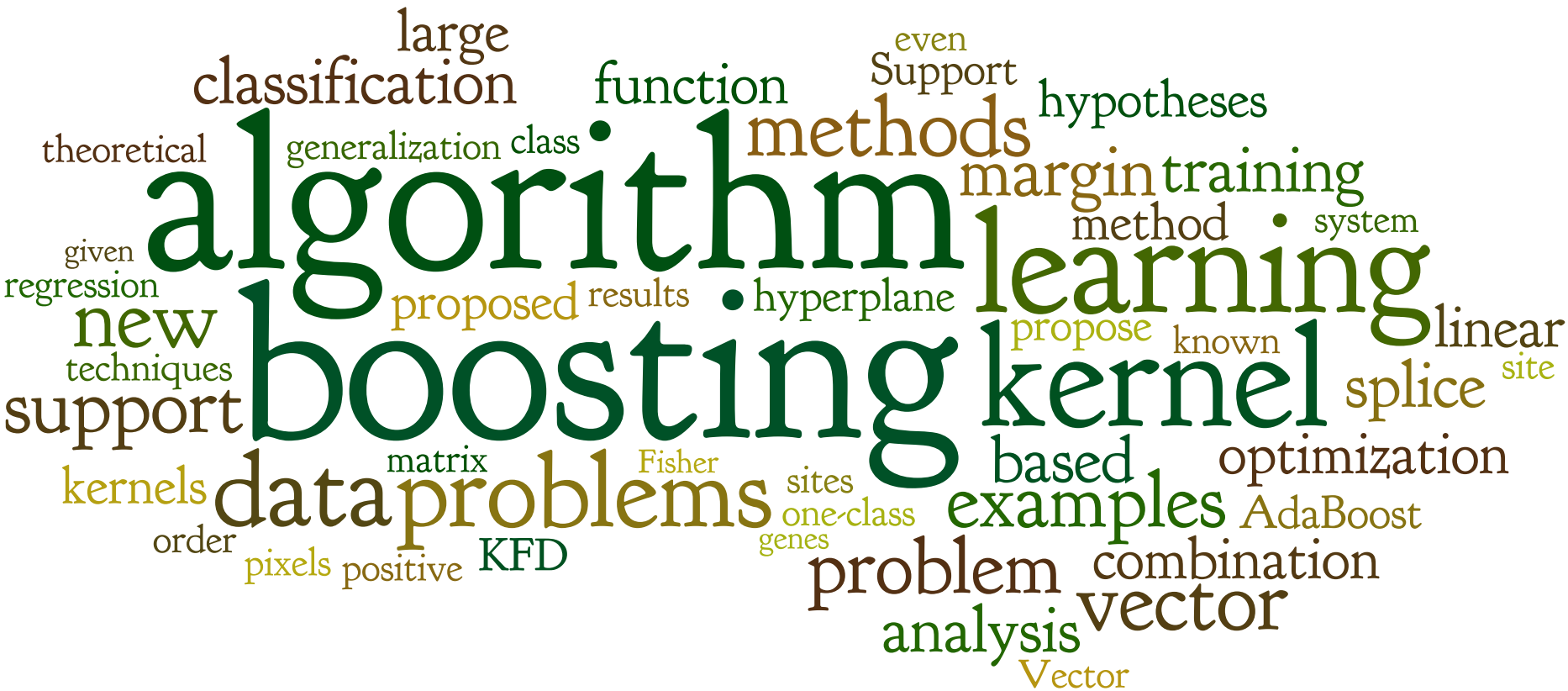
Goal: Learn to predict what these processes accomplish:

- Given the DNA, ..., predict all gene products

$$f(\text{DNA}, \boxed{1}\,\boxed{2}\,\boxed{3}) = \text{RNA} \qquad F(\text{RNA}, \boxed{4}\,\boxed{5}) = \text{protein}$$

- Estimating f, F amounts to cracking the codes of transcription, epigenetics, splicing, ..., protein interactions, ...

Machine Learning
Learning from empirical observations



Frequent words of abstracts from publications 1998-2004.

[wordle.net]

Machine Learning

Learning from empirical observations

Given: Observations of some complex phenomenon

Goal: Learn from data & build predictive models

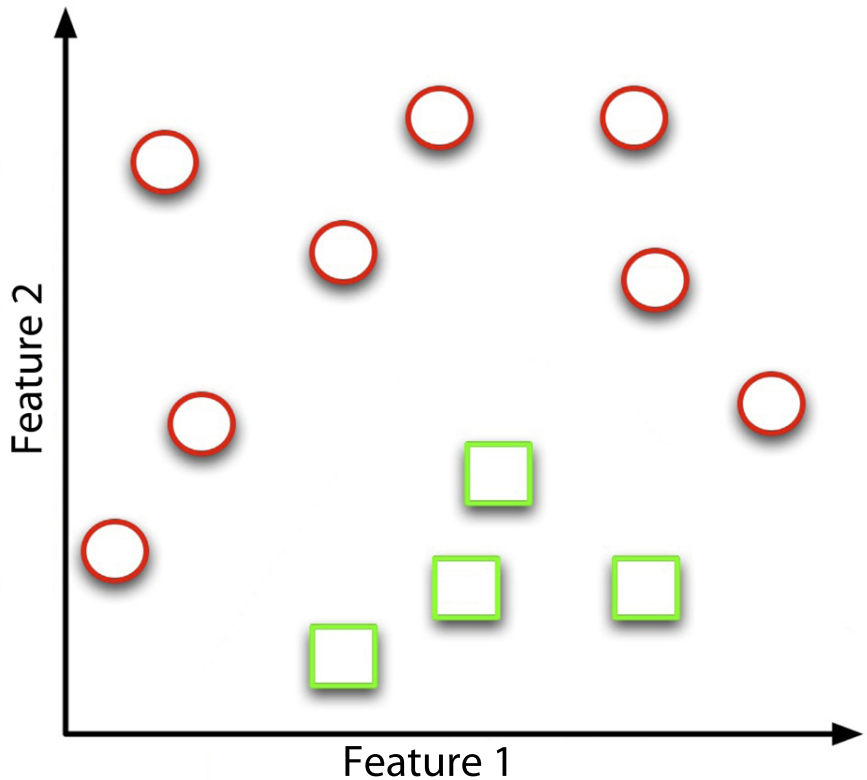
Machine Learning

Learning from empirical observations

Given: Observations of some complex phenomenon

Goal: Learn from data & build predictive models

Example:



Two different classes of observations

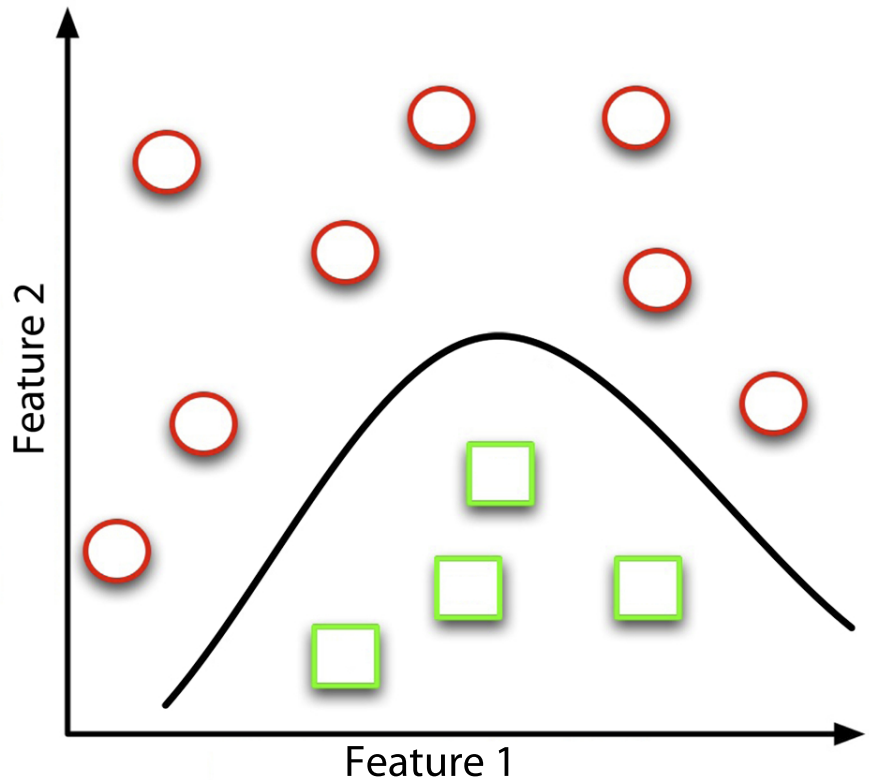
Machine Learning

Learning from empirical observations

Given: Observations of some complex phenomenon

Goal: Learn from data & build predictive models

Example:



Infer separation between classes

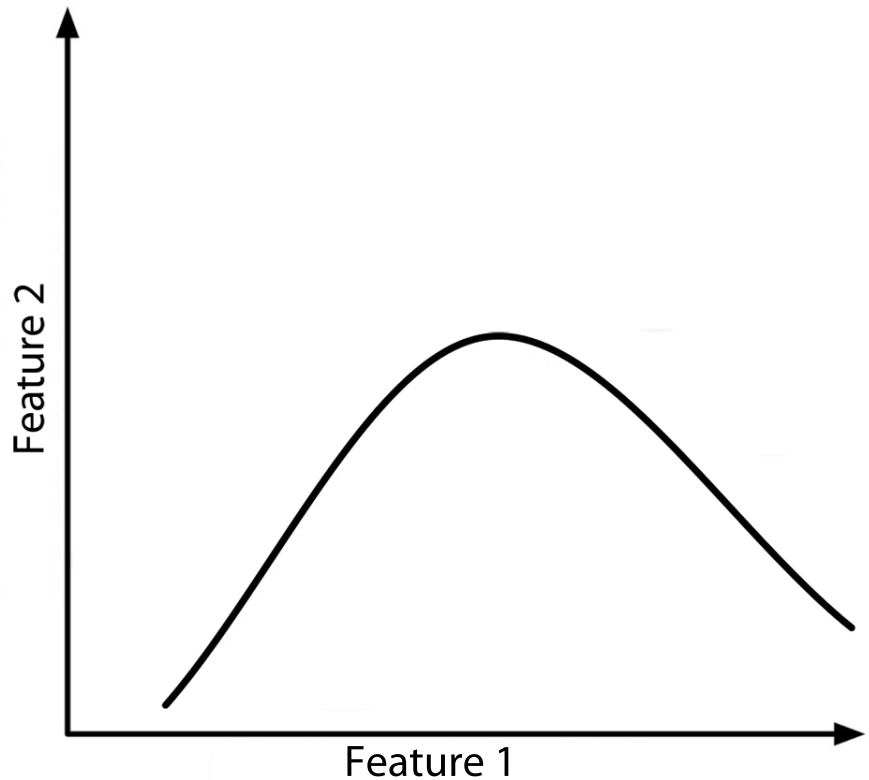
Machine Learning

Learning from empirical observations

Given: Observations of some complex phenomenon

Goal: Learn from data & build predictive models

Example:



Rule for classifying new examples

Machine Learning in Biology
Research Topics



Machine Learning

1. Large scale sequence analysis
   ⇒ *Develop fast and accurate classification methods*
2. Predicting for structured outputs
   ⇒ *Accurate segmentation and structure prediction*

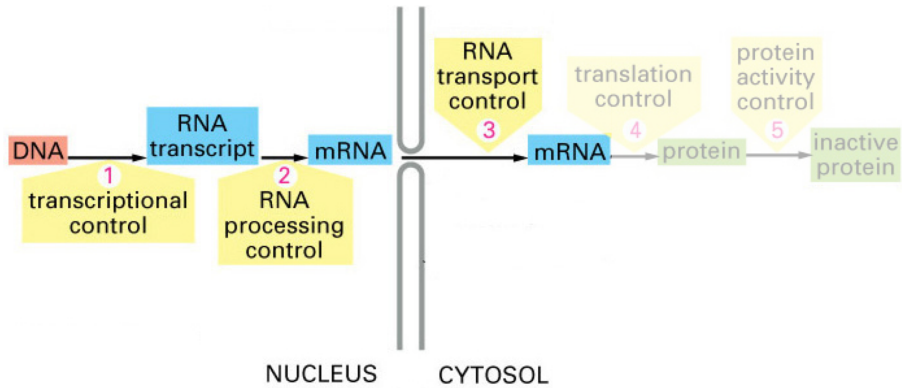
Molecular Biology

3. Transcriptome analysis & genome annotation
   ⇒ *Predict features encoded on DNA*
4. Analysis of polymorphisms
   ⇒ *Discover polymorphisms and associate with phenotypes*
5. Biological networks
   ⇒ *Understand interactions between gene products*

Learning About the Genome

Goal: Estimate f to predict RNAs

$$f(\text{DNA}, \boxed{1\ 2\ 3}) = \text{RNA}$$



Need:

- Good inference method
- Inputs (DNA,)
- Outputs (complete transcriptome)

Intermediate Goal: Fully reconstruct RNAs from experimental data

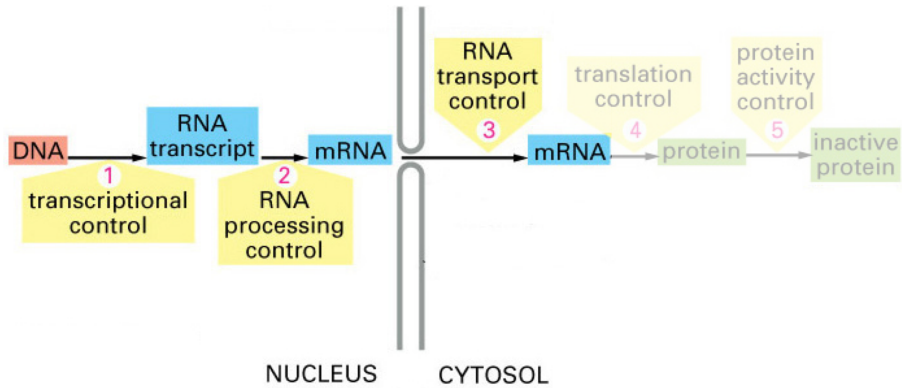
- Whole-genome tiling arrays (2007-2010)
- Deep RNA sequencing (RNA-seq) (2008-)

Learning About the Genome

Goal: Estimate f to predict RNAs

✔

$$f(\text{DNA}, \boxed{1\ 2\ 3}) = \text{RNA}$$



Need:

- Good inference method ✔
- Inputs (DNA,)
- Outputs (complete transcriptome)

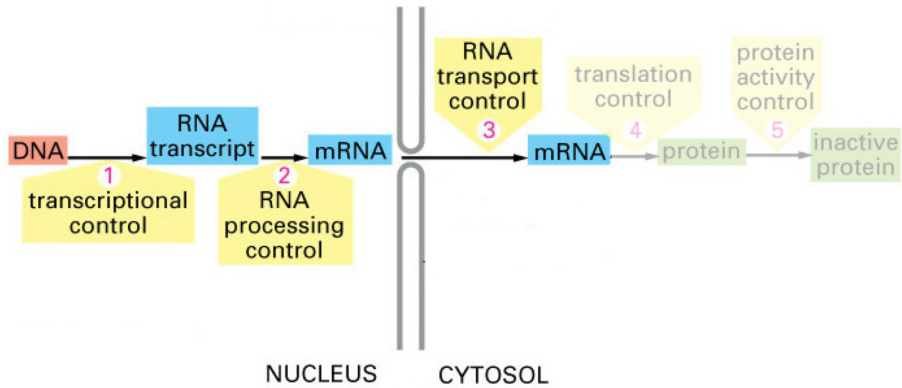
Intermediate Goal: Fully reconstruct RNAs from experimental data

- Whole-genome tiling arrays (2007-2010)
- Deep RNA sequencing (RNA-seq) (2008-)

Learning About the Genome

Goal: Estimate f to predict RNAs

$$f(\text{DNA}, \boxed{1}\boxed{2}\boxed{3}) = \text{RNA}$$



Need:

- Good inference method
- Inputs (DNA, 1 2 3)
- Outputs (complete transcriptome)

Intermediate Goal: Fully reconstruct RNAs from experimental data

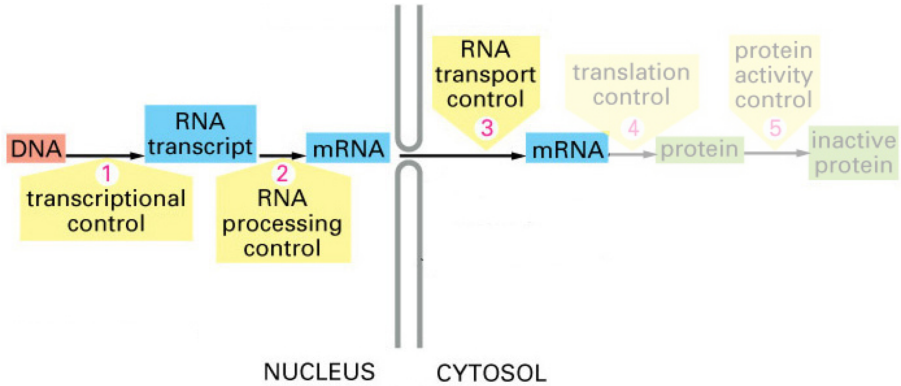
- Whole-genome tiling arrays (2007-2010)
- Deep RNA sequencing (RNA-seq) (2008-)

Learning About the Genome

Goal: Estimate f to predict RNAs

✔ ✔ ?

$f(\text{DNA}, \boxed{1}\boxed{2}\boxed{3}) = \text{RNA}$



Need:

- Good inference method ✔
- Inputs (DNA, 1 2 3) ?
- Outputs (complete transcriptome)

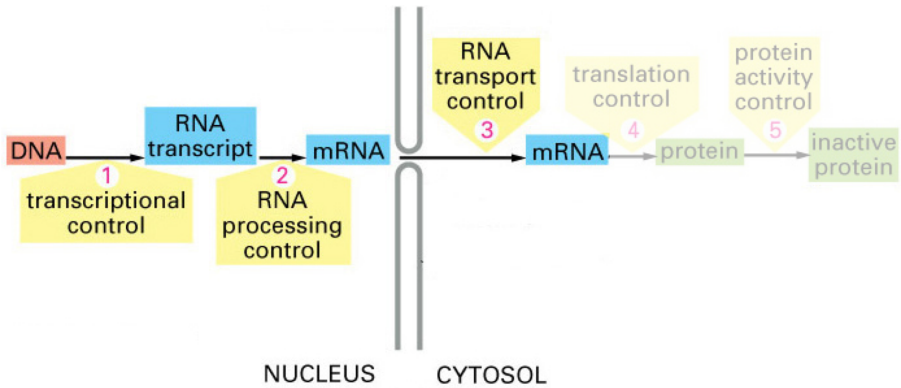
Intermediate Goal: Fully reconstruct RNAs from experimental data

- Whole-genome tiling arrays (2007-2010)
- Deep RNA sequencing (RNA-seq) (2008-)

Learning About the Genome

Goal: Estimate f to predict RNAs

$$f(\text{DNA}, \text{1 2 3}) = \text{RNA}$$



Need:

- Good inference method
- Inputs (DNA, 1 2 3)
- Outputs (complete transcriptome)

Intermediate Goal: Fully reconstruct RNAs from experimental data

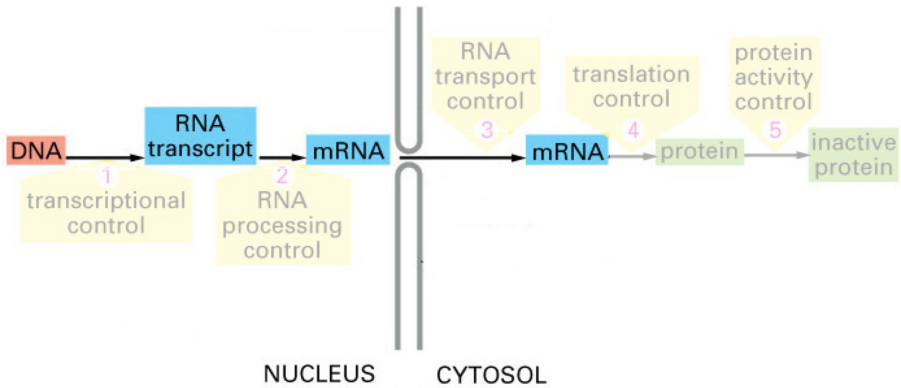
- Whole-genome tiling arrays (2007-2010)
- Deep RNA sequencing (RNA-seq) (2008-)

Learning About the Genome

Goal: Estimate f to predict RNAs

✔ ✔ ? ?

$f(\text{DNA}, \cancel{1\ 2\ 3}) \approx \text{RNA}$



Need:

- Good inference method ✔
- Inputs (DNA, ~~1 2 3~~) ✔
- Outputs (complete transcriptome) ?

Intermediate Goal: Fully reconstruct RNAs from experimental data

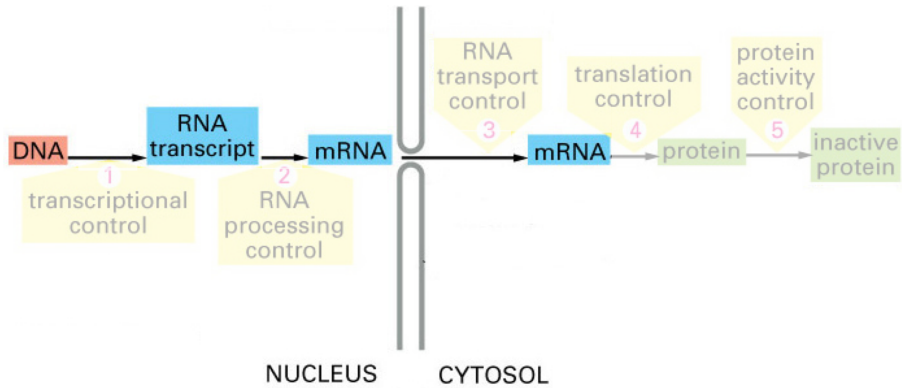
- Whole-genome tiling arrays (2007-2010)
- Deep RNA sequencing (RNA-seq) (2008-)

Learning About the Genome

Goal: Estimate f to predict RNAs

✔ ✔ ? ?

$f(\text{DNA}, \cancel{1\ 2\ 3}) \approx \text{RNA}$



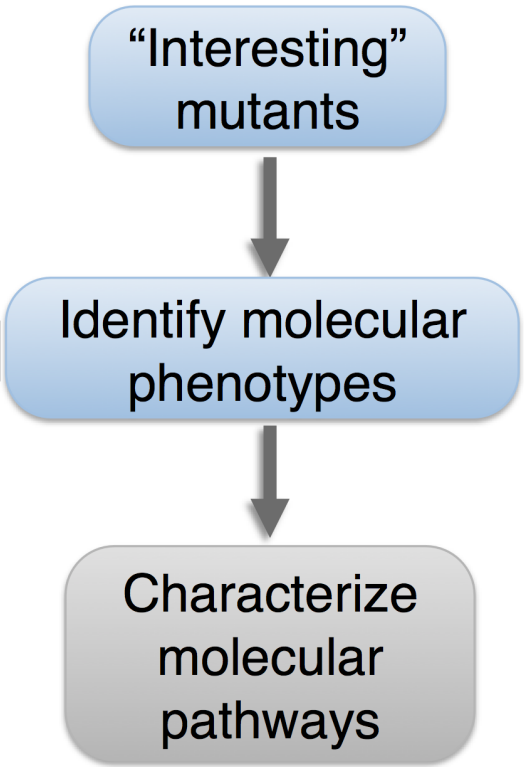
Need:

- Good inference method ✔
- Inputs (DNA, ~~1 2 3~~) ✔
- Outputs (complete transcriptome) ?

Intermediate Goal: Fully reconstruct RNAs from experimental data

- Whole-genome tiling arrays (2007-2010)
- Deep RNA sequencing (RNA-seq) (2008-)

Approaches to Understand Molecular Pathways

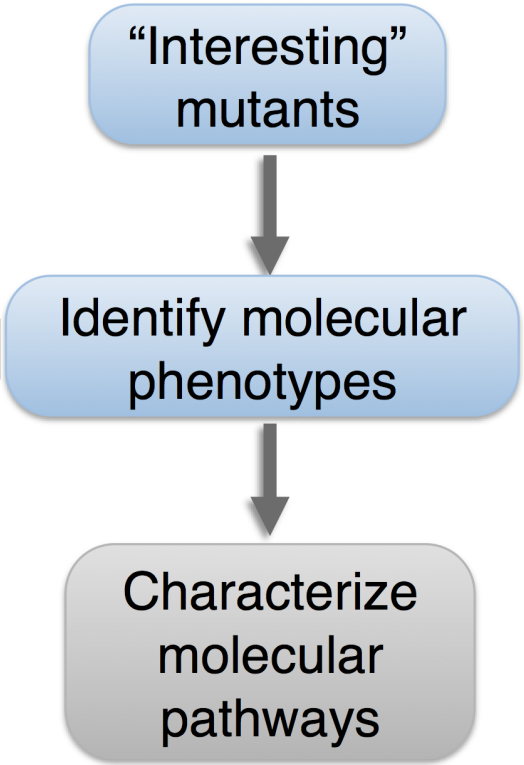


- CBP20/80 [Laubinger et al., 2008]
- sRNA biogenesis (*dcl1-4*) [Laubinger et al., 2010]
- NMD [w/ A. Wachter]
- Exosome [w/ E. Conti]

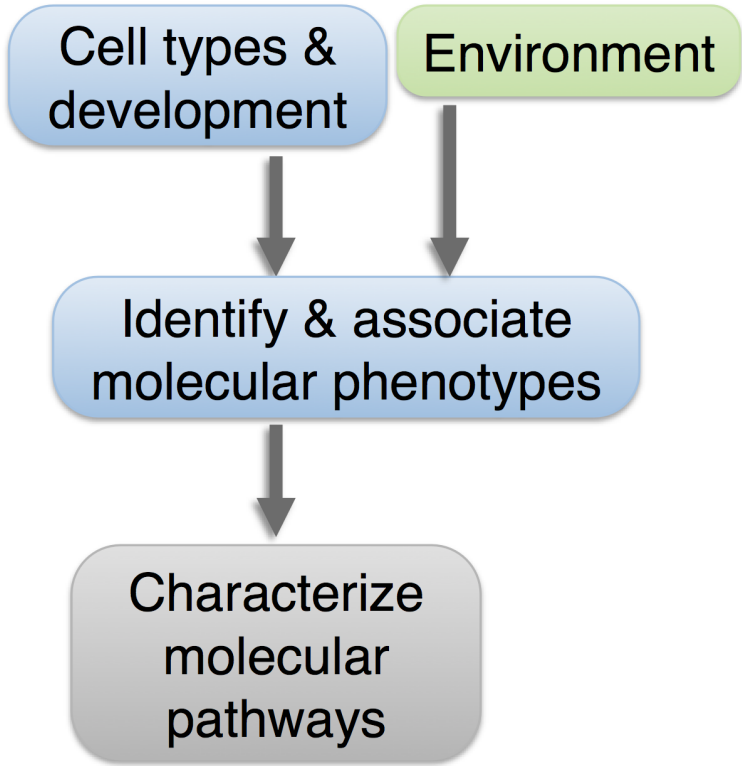
- *A. thaliana* tissue atlas [Laubinger et al., 2008]
- *A. thaliana* stress response [Zeller et al., 2009]
- *C. elegans* tissue/development atlas [Gerstein et al., 2010, Spencer et al., 2011]

- *A. thaliana* genome variation [Clark et al., 2007a]
- *O. sativa* genome variation [McNally et al., 2009]
- *A. thaliana* transcriptome variation [w/ R. Mott & R. Clark]

Approaches to Understand Molecular Pathways



- CBP20/80 [Laubinger et al., 2008]
- sRNA biogenesis (*dcl1-4*) [Laubinger et al., 2010]
- NMD [w/ A. Wachter]
- Exosome [w/ E. Conti]

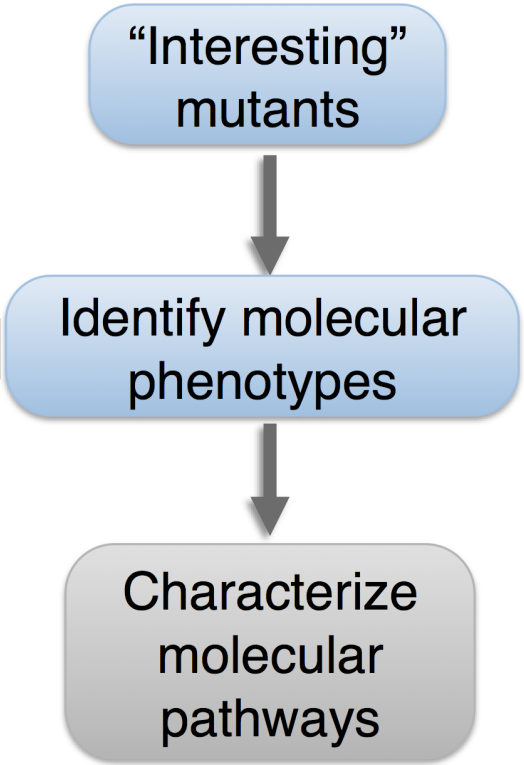


- *A. thaliana* tissue atlas [Laubinger et al., 2008]
- *A. thaliana* stress response [Zeller et al., 2009]
- *C. elegans* tissue/development atlas [Gerstein et al., 2010, Spencer et al., 2011]

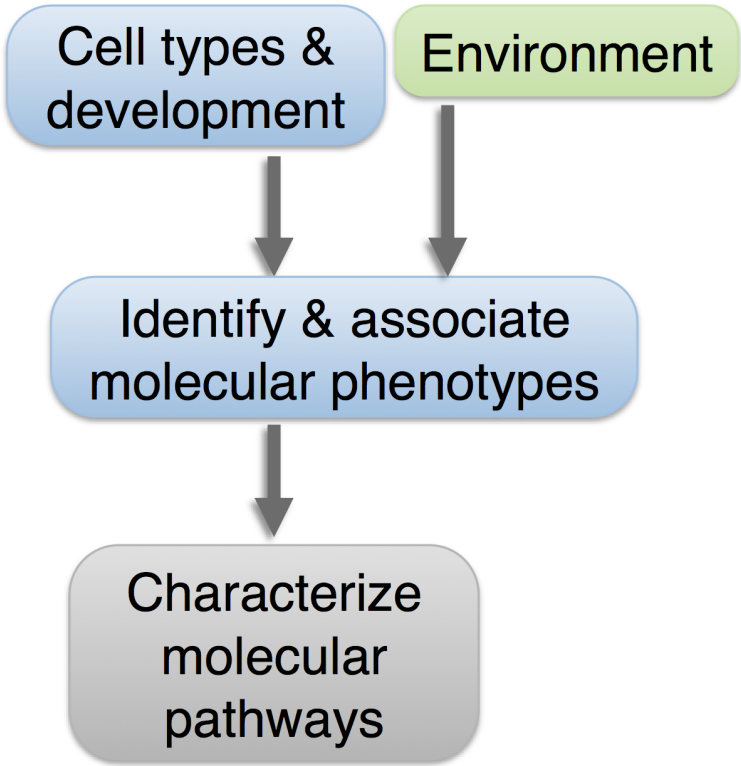
- *A. thaliana* genome variation [Clark et al., 2007a]
- *O. sativa* genome variation [McNally et al., 2009]
- *A. thaliana* transcriptome variation [w/ R. Mott & R. Clark]

Approaches to Understand Molecular Pathways

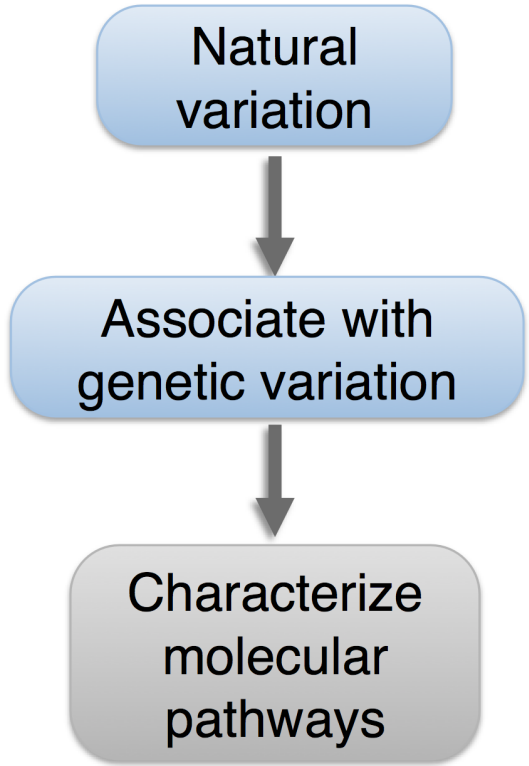




- CBP20/80 [Laubinger et al., 2008]
- sRNA biogenesis (*dcl1-4*) [Laubinger et al., 2010]
- NMD [w/ A. Wachter]
- Exosome [w/ E. Conti]



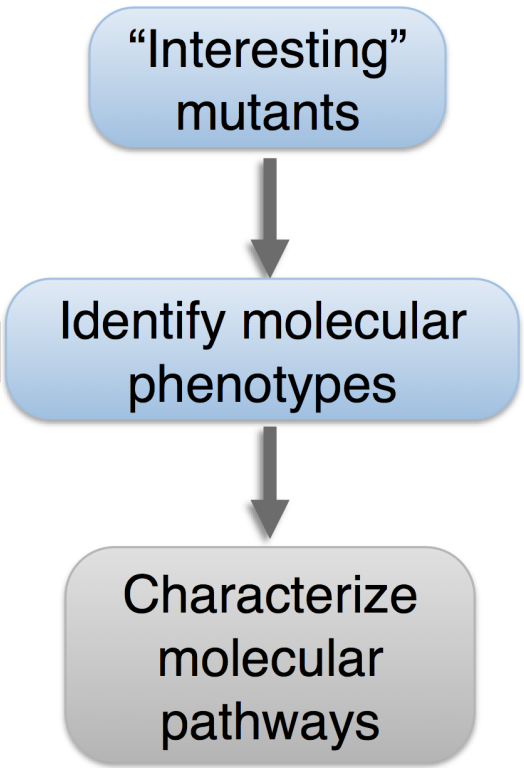
- *A. thaliana* tissue atlas [Laubinger et al., 2008]
- *A. thaliana* stress response [Zeller et al., 2009]
- *C. elegans* tissue/development atlas [Gerstein et al., 2010, Spencer et al., 2011]



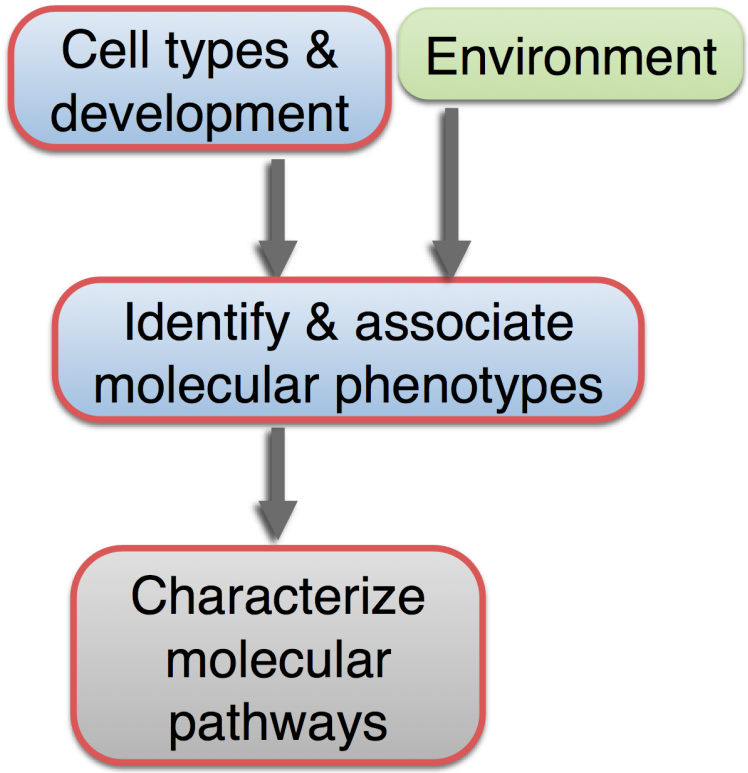
- *A. thaliana* genome variation [Clark et al., 2007a]
- *O. sativa* genome variation [McNally et al., 2009]
- *A. thaliana* transcriptome variation [w/ R. Mott & R. Clark]

Approaches to Understand Molecular Pathways

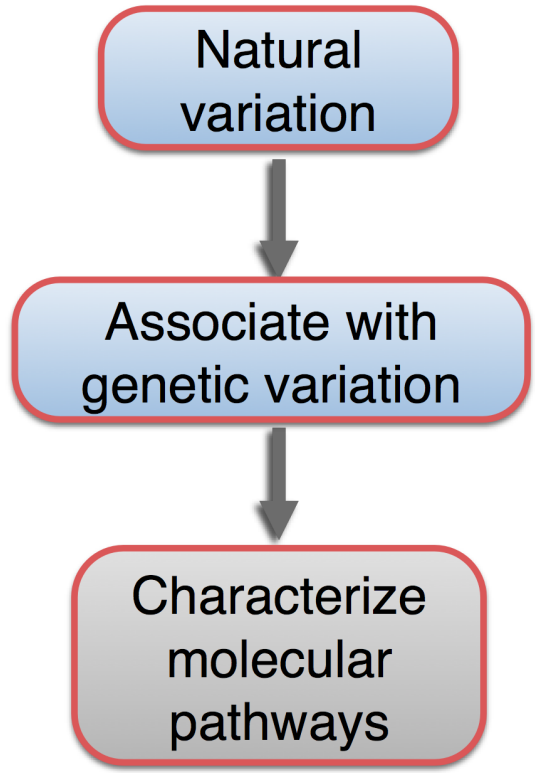




- CBP20/80 [Laubinger et al., 2008]
- sRNA biogenesis (*dcl1-4*) [Laubinger et al., 2010]
- NMD [w/ A. Wachter]
- Exosome [w/ E. Conti]

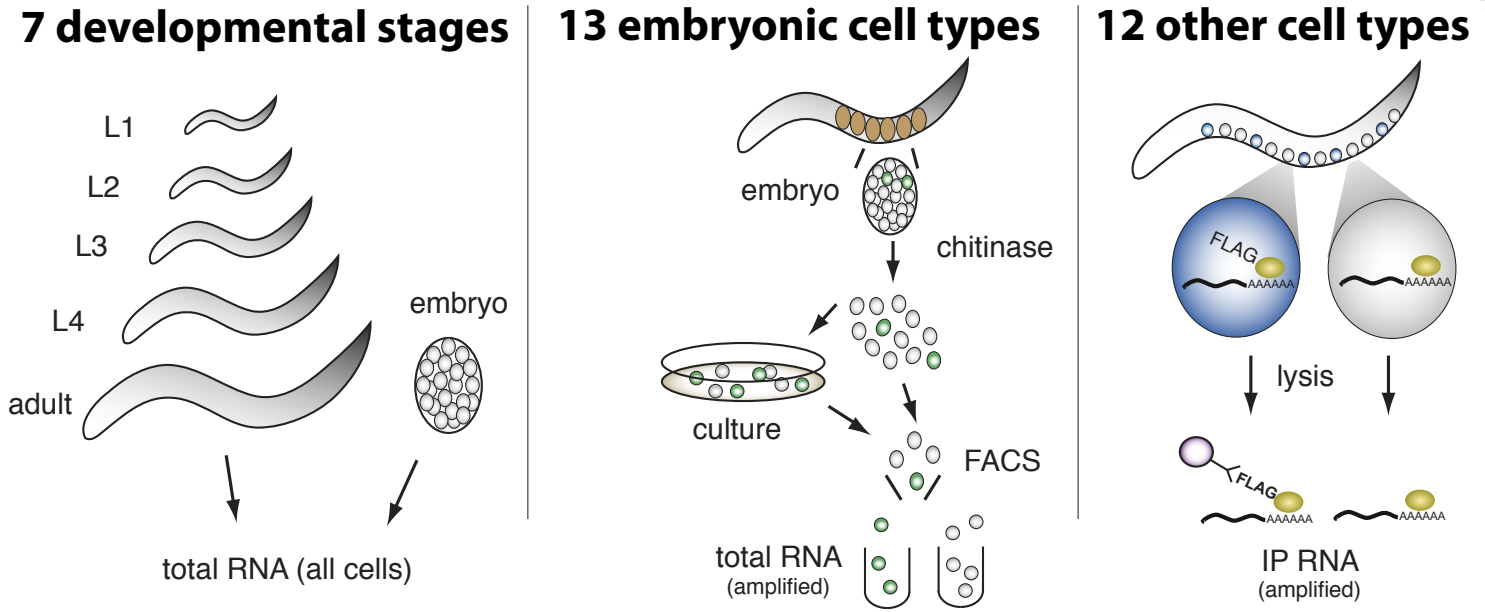


- *A. thaliana* tissue atlas [Laubinger et al., 2008]
- *A. thaliana* stress response [Zeller et al., 2009]
- *C. elegans* tissue/development atlas [Gerstein et al., 2010, Spencer et al., 2011]

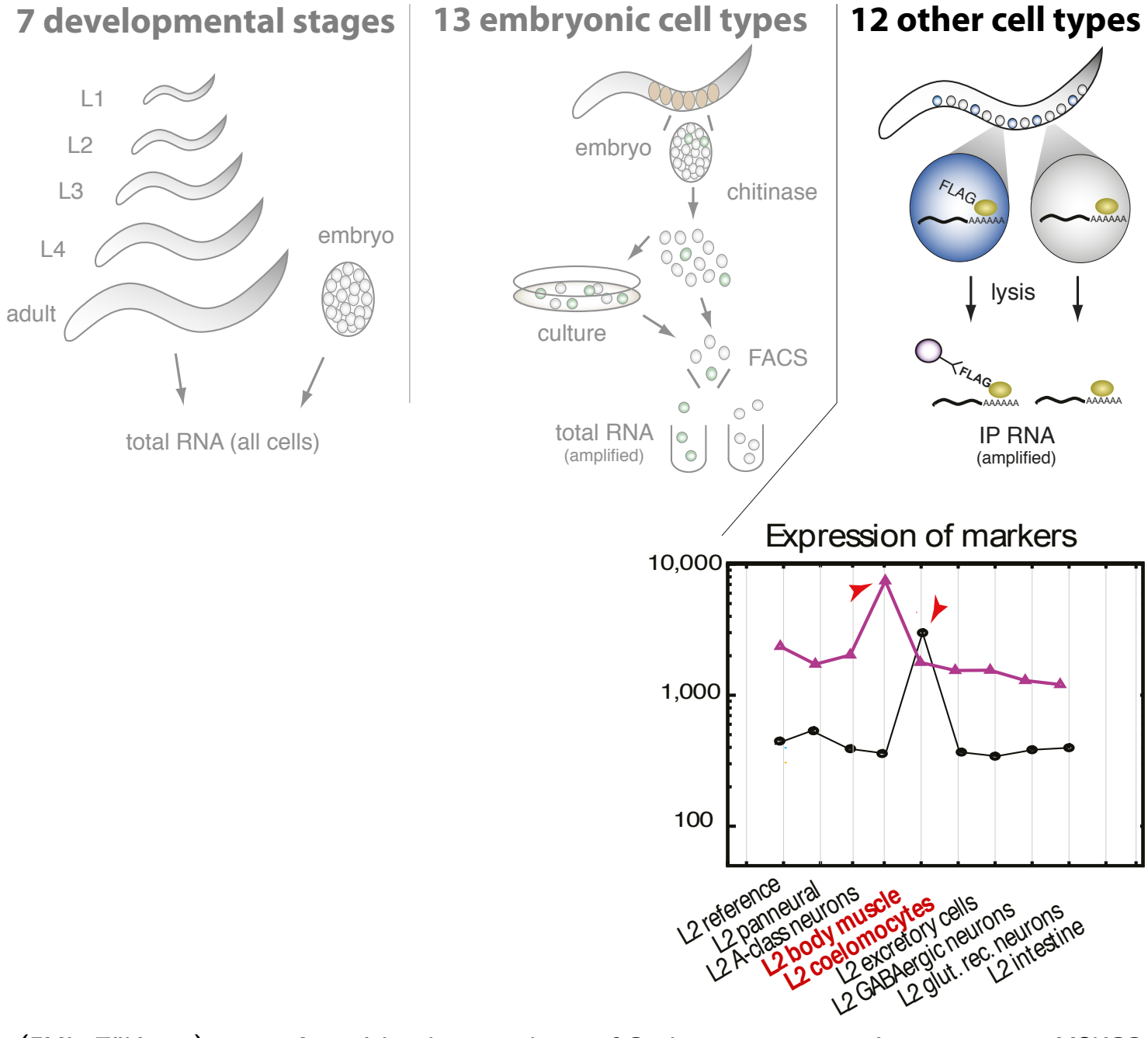


- *A. thaliana* genome variation [Clark et al., 2007a]
- *O. sativa* genome variation [McNally et al., 2009]
- *A. thaliana* transcriptome variation [w/ R. Mott & R. Clark]

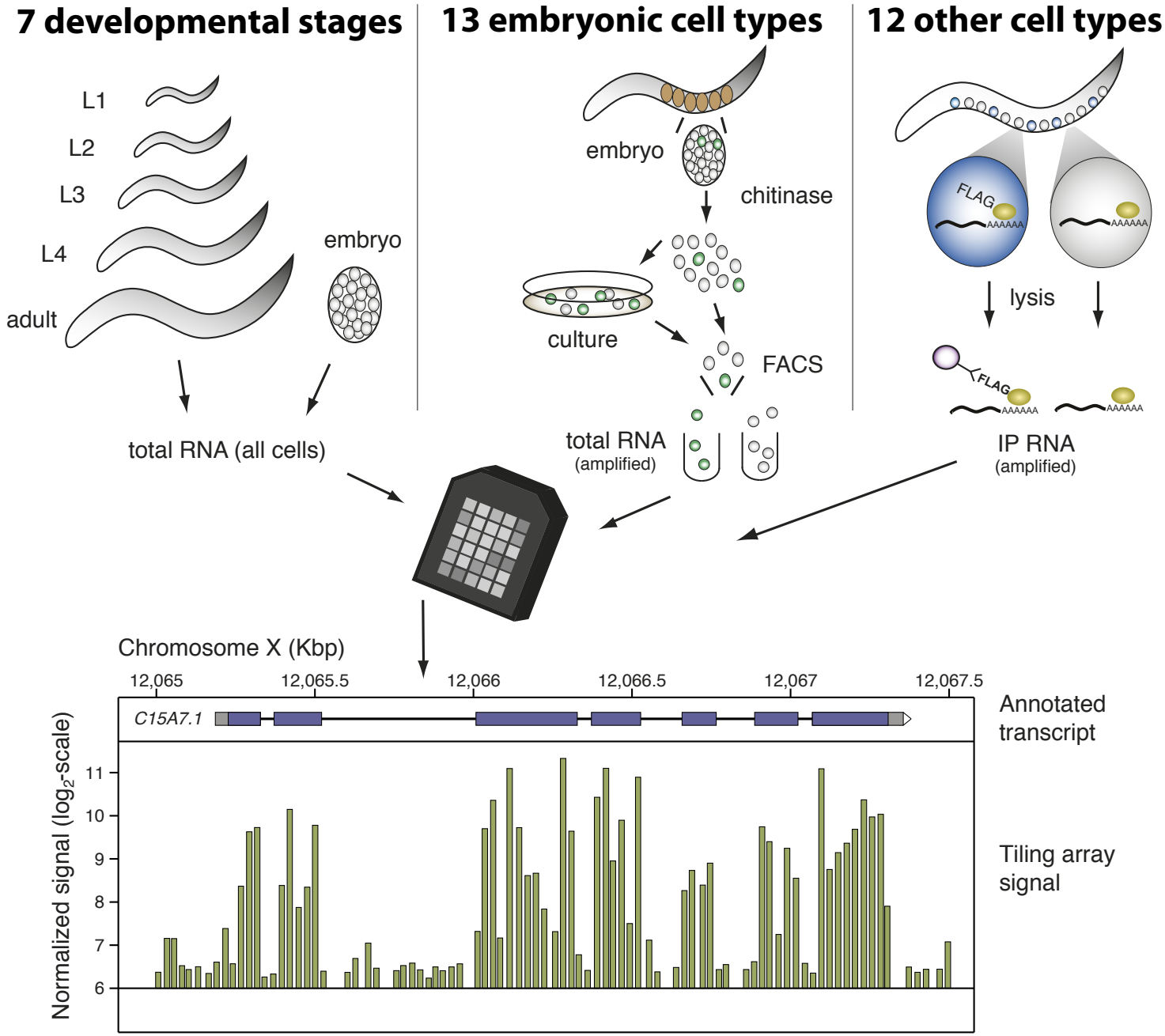
Analysis of C. elegans Transcriptomes



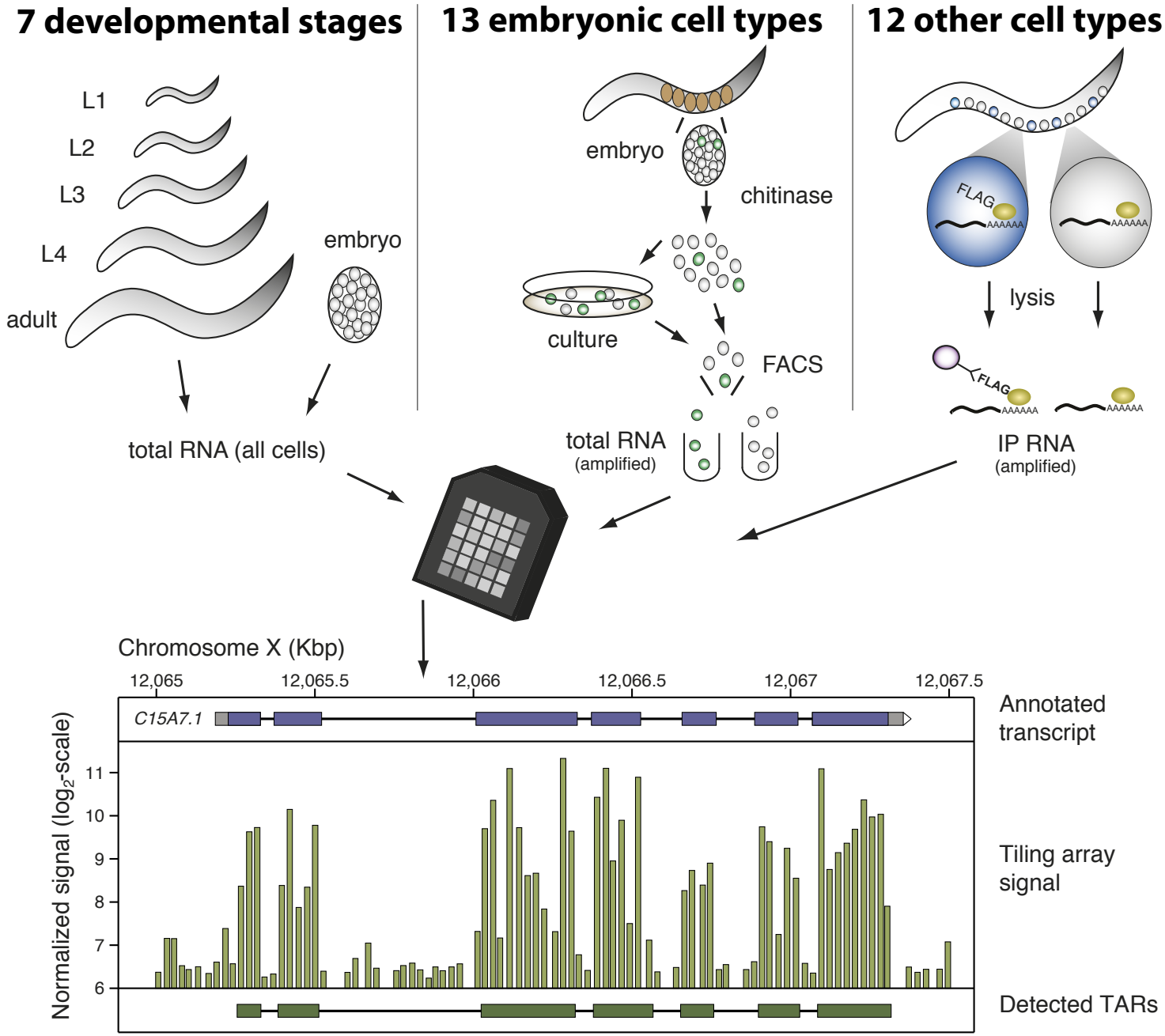
Analysis of C. elegans Transcriptomes



Analysis of C. elegans Transcriptomes



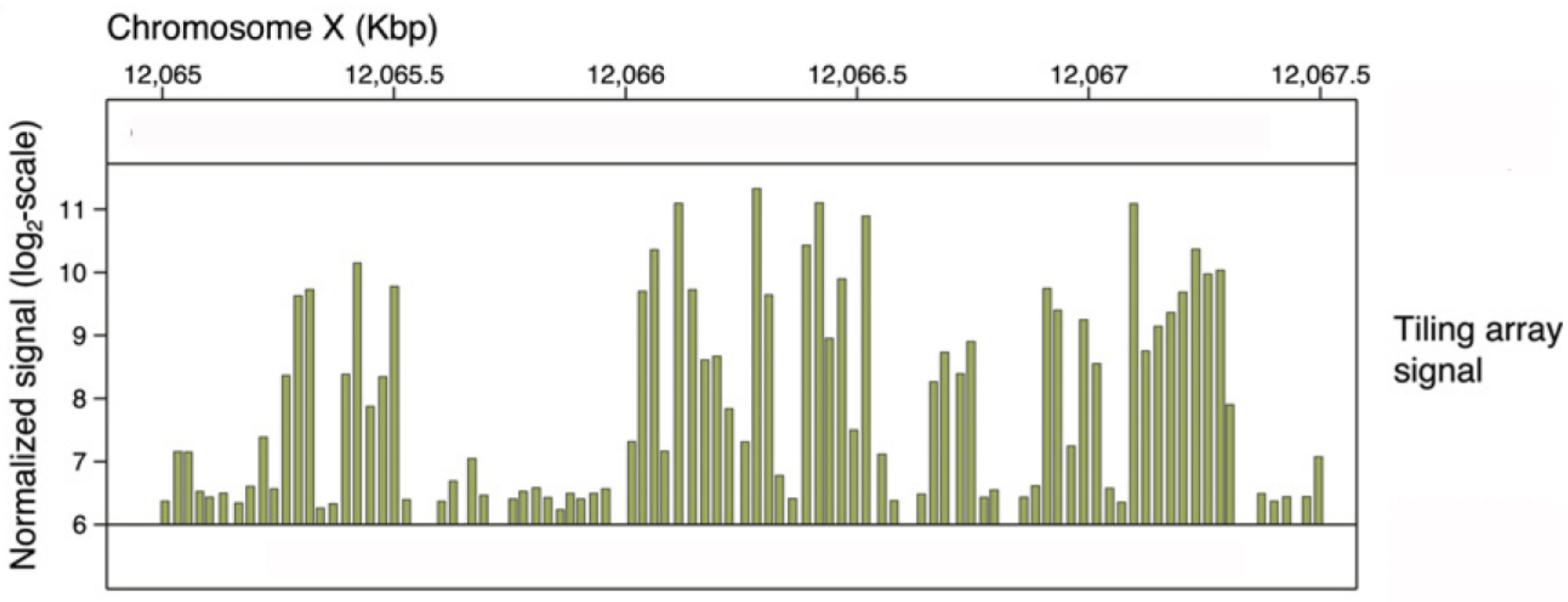
Analysis of C. elegans Transcriptomes



Analysis overview

The expression analysis addressed the following questions:

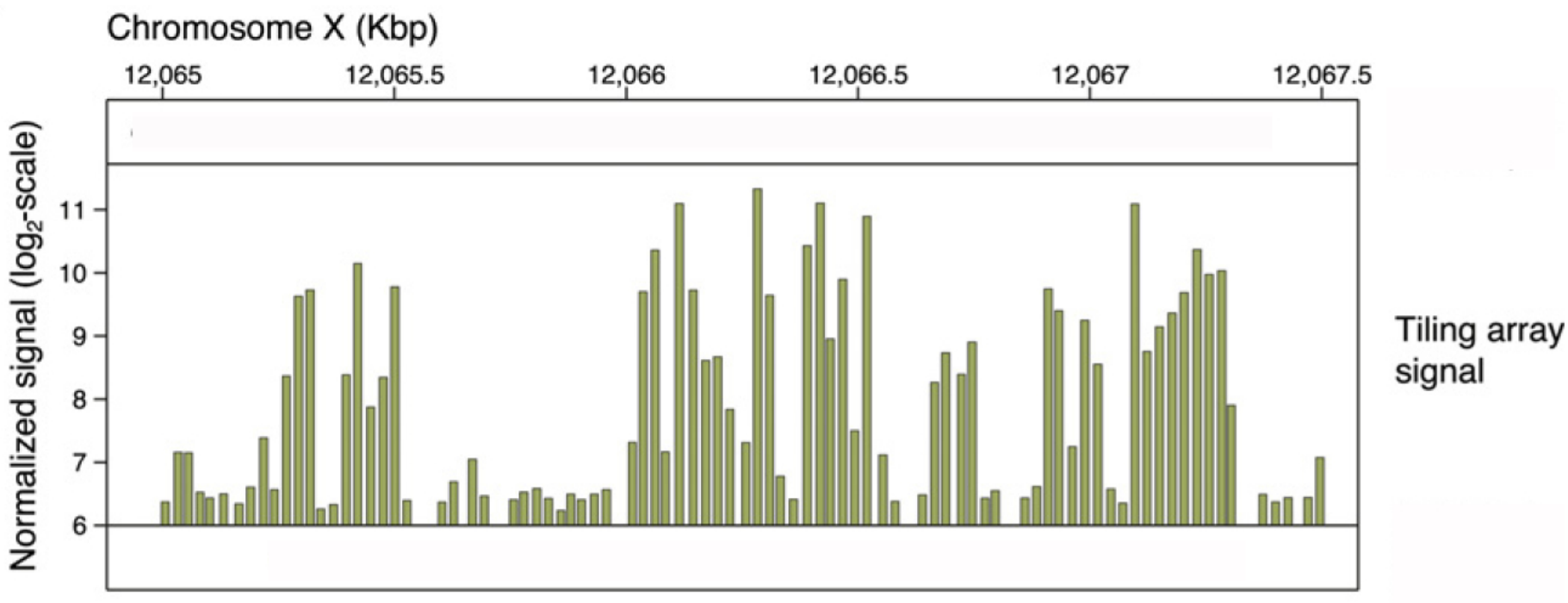
- Which genes are expressed when?
- Are there new genes which are not known up to now?



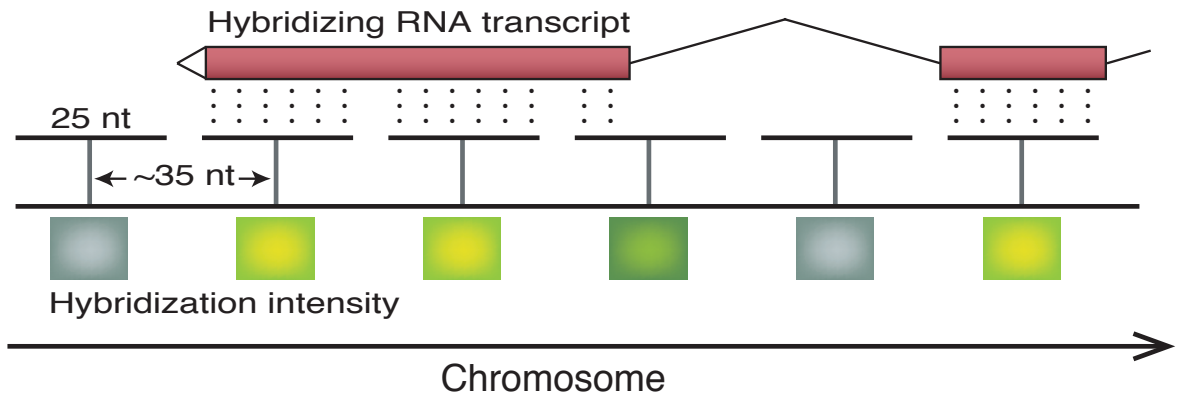
Analysis overview

The expression analysis addressed the following questions:

- Which genes are expressed when?
- Are there new genes which are not known up to now?



Analysis challenges

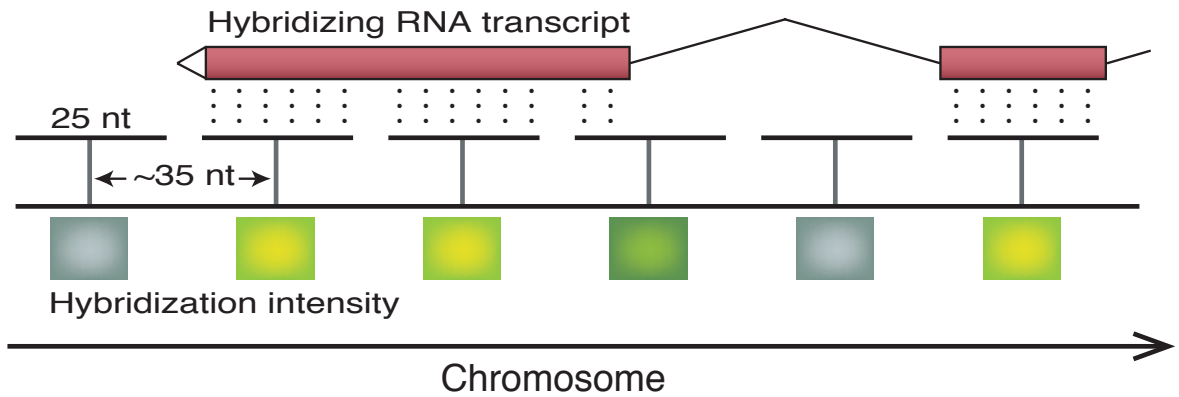


- Cross hybridization
- Different experimental protocols in different labs
- Probe-sequence effects,
- Sample impurity ...

Method developments:

- Improved probe-sequence normalization [Zeller et al., 2008c]
- Accurate segmentation of tiling array signals [Zeller et al., 2008b,c]

Analysis challenges

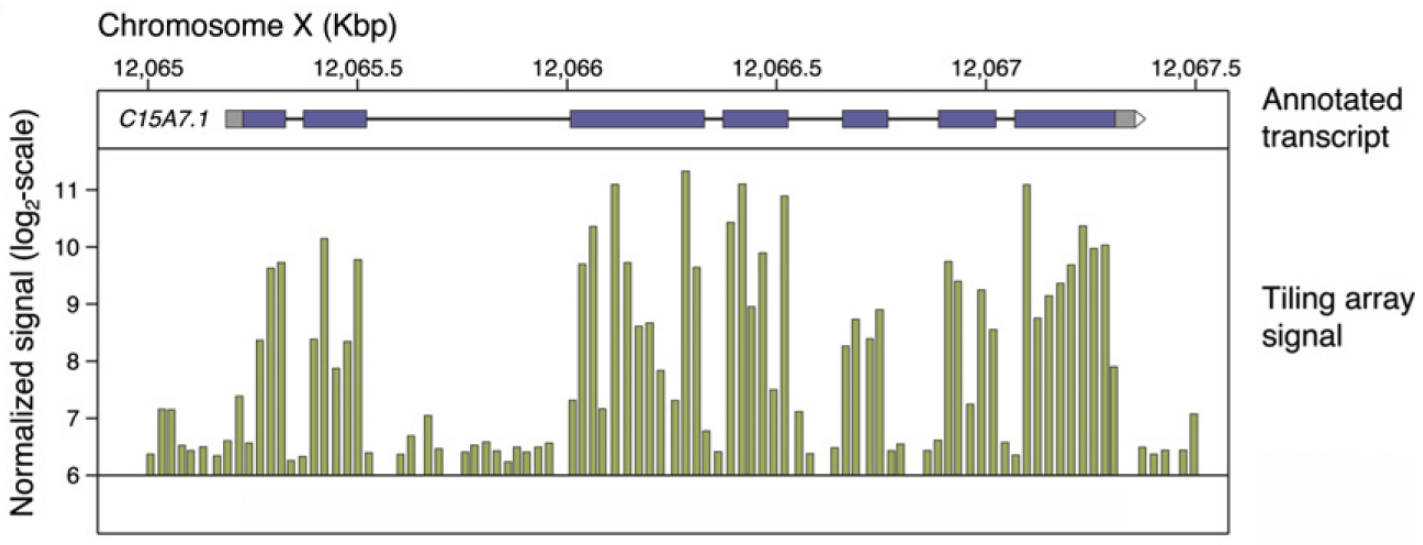


- Cross hybridization
- Different experimental protocols in different labs
- Probe-sequence effects,
- Sample impurity ...

Method developments:

- Improved probe-sequence normalization [Zeller et al., 2008c]
- Accurate segmentation of tiling array signals [Zeller et al., 2008b,c]

Identification of novel TARs



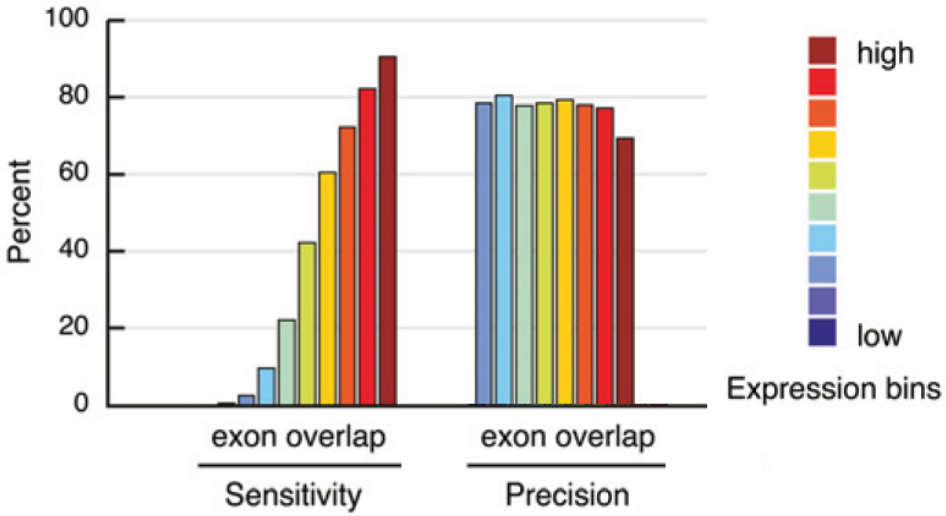
Identification of novel TARs from tiling array data with mSTAD

[Zeller et al., 2008b,c]

- Train the internal parameters of the mSTAD on the hybridization patterns of a few experimentally confirmed genes
- Apply mSTAD to identify novel TARs

Identification of novel TARs

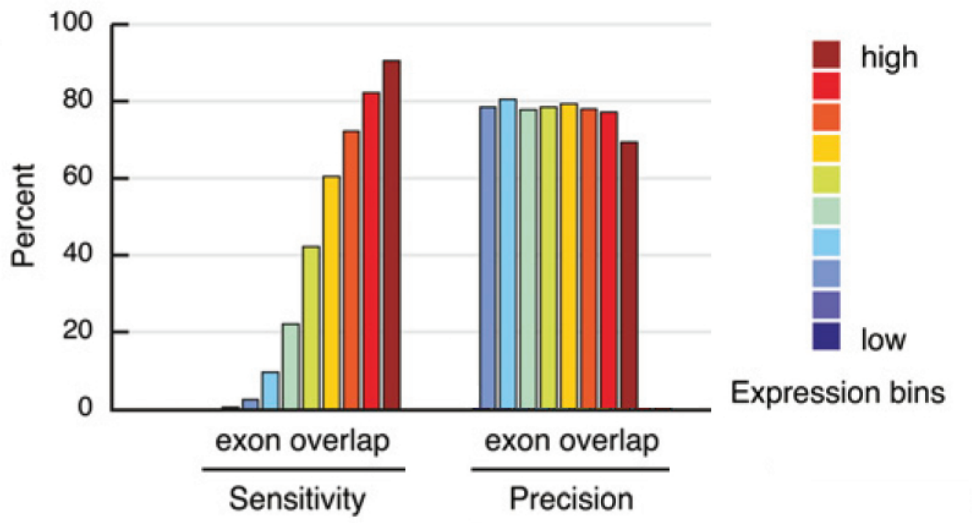
Exon detection accuracy at probe level of mSTAD:

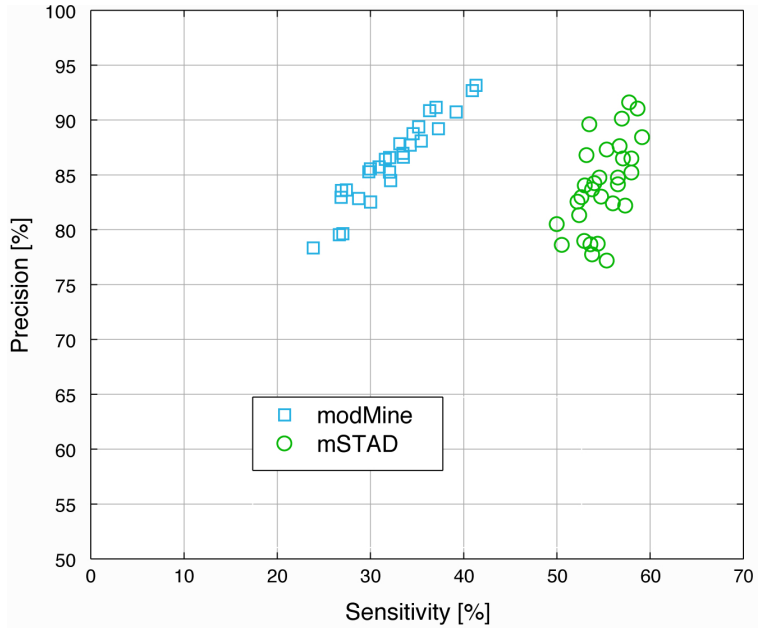


- Evaluation of whole-genome predictions
  - $\Rightarrow$ higher expressed genes are easier to detect
- Method comparison:
  - Modmine [Gerstein et al., Science 2010 Dec 24;330(6012):1775-87]
  - mSTAD approach [Zeller et al., Genome Res., 2011 Feb;21(2):325-41]

Identification of novel TARs

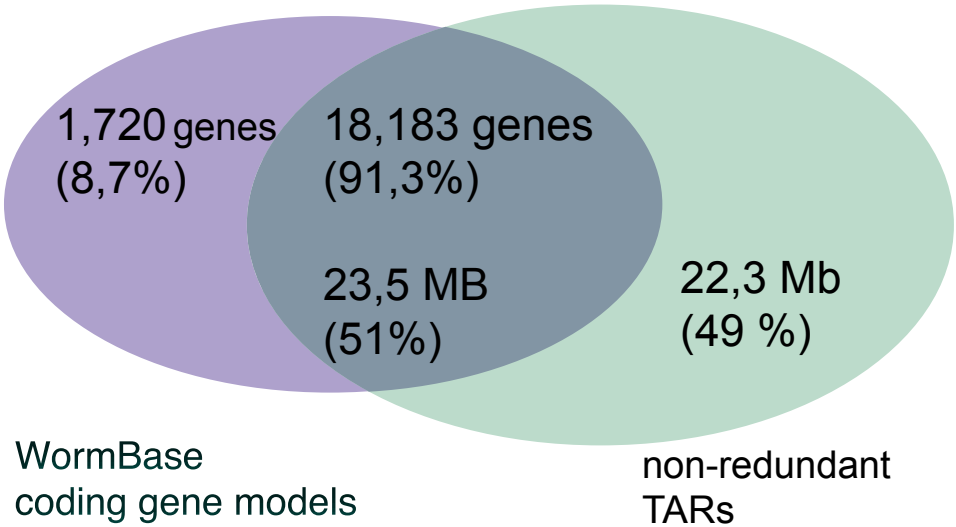
Exon detection accuracy at probe level of mSTAD:





- Evaluation of whole-genome predictions
  - $\Rightarrow$ higher expressed genes are easier to detect
- Method comparison:
  - Modmine [Gerstein et al., Science 2010 Dec 24;330(6012):1775-87]
  - mSTAD approach [Zeller et al., Genome Res., 2011 Feb;21(2):325-41]

Identification of novel TARs

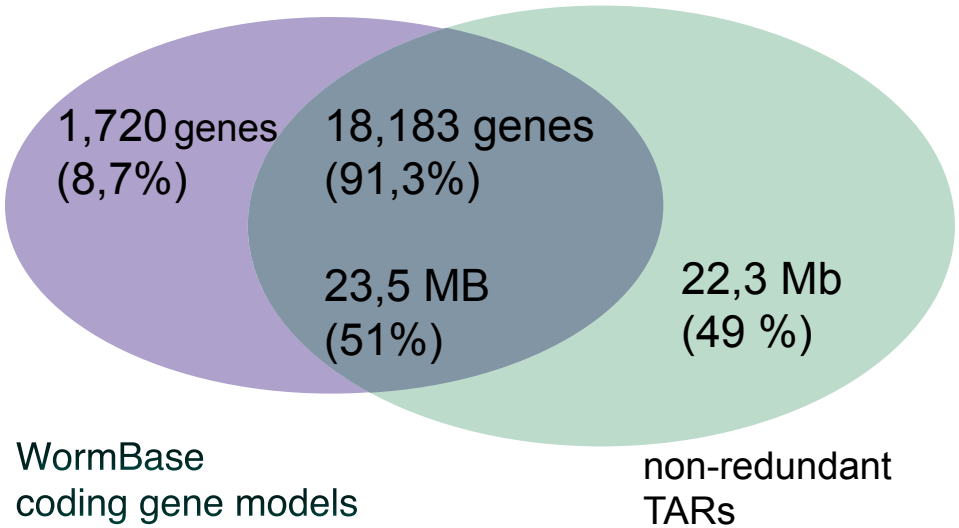


Detected TARs:

- Overlap TARs with coding genes: $\approx 91\%$
- $\approx$ 22 Mb not in protein coding regions
- $\approx$ 14 Mb novel TARs

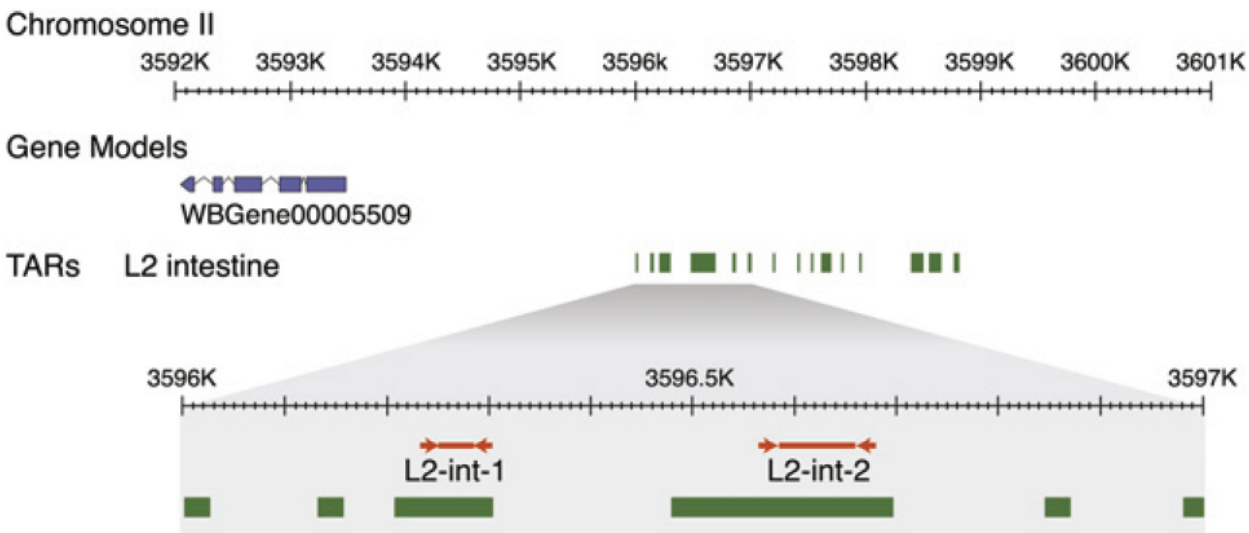
[Spencer et al., 2011]

Identification of novel TARs



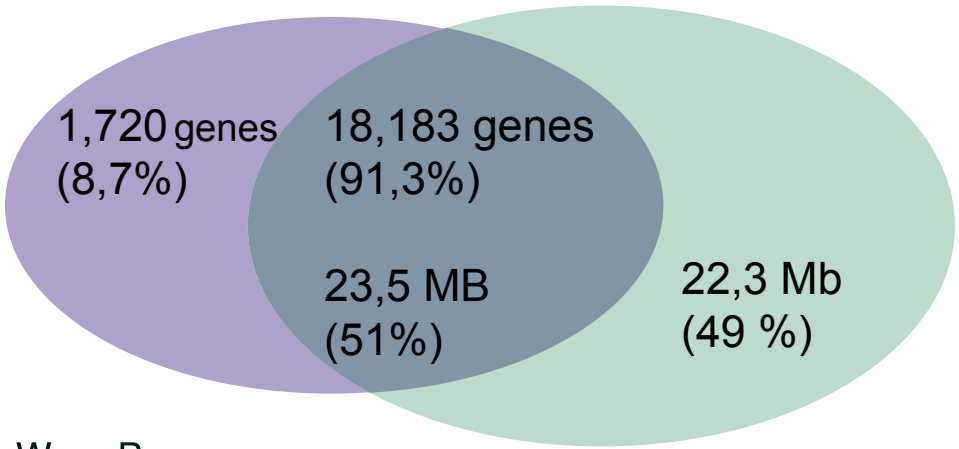
Detected TARs:

- Overlap TARs with coding genes: $\approx 91\%$
- $\approx$ 22 Mb not in protein coding regions
- $\approx$ 14 Mb novel TARs



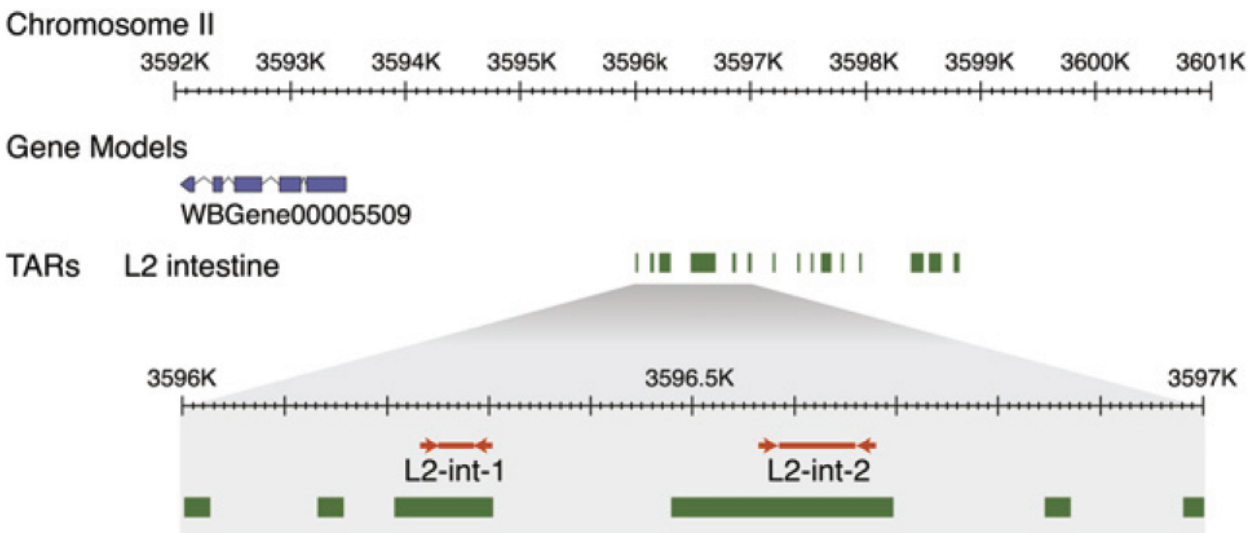
[Spencer et al., 2011]

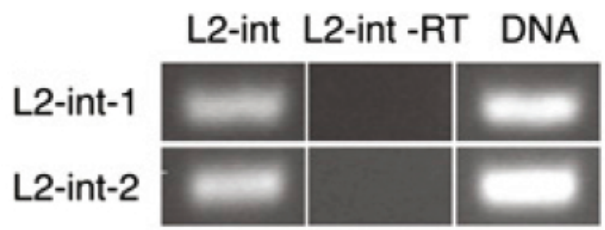
Identification of novel TARs



Detected TARs:

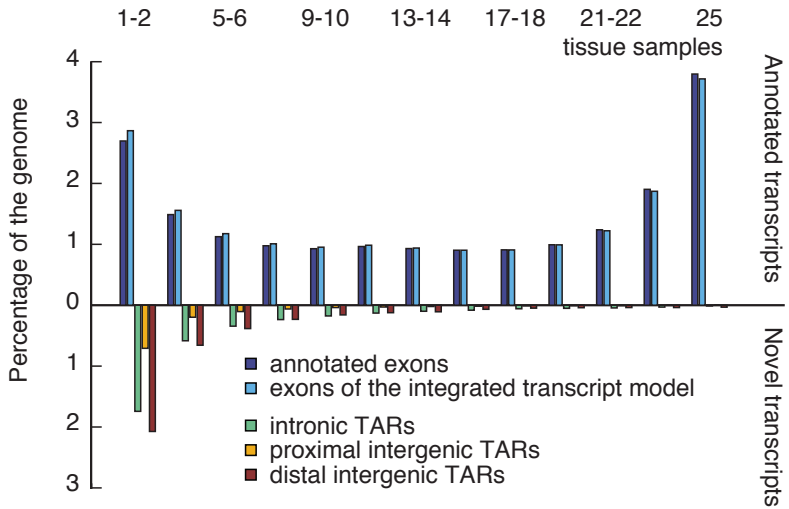
- Overlap TARs with coding genes: $\approx 91\%$
- $\approx$ 22 Mb not in protein coding regions
- $\approx$ 14 Mb novel TARs

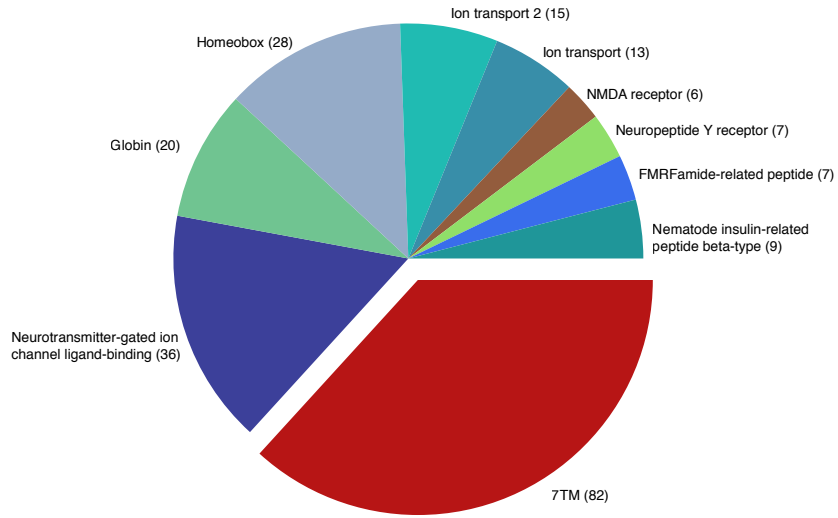




[Spencer et al., 2011]

Which genes are specifically expressed?

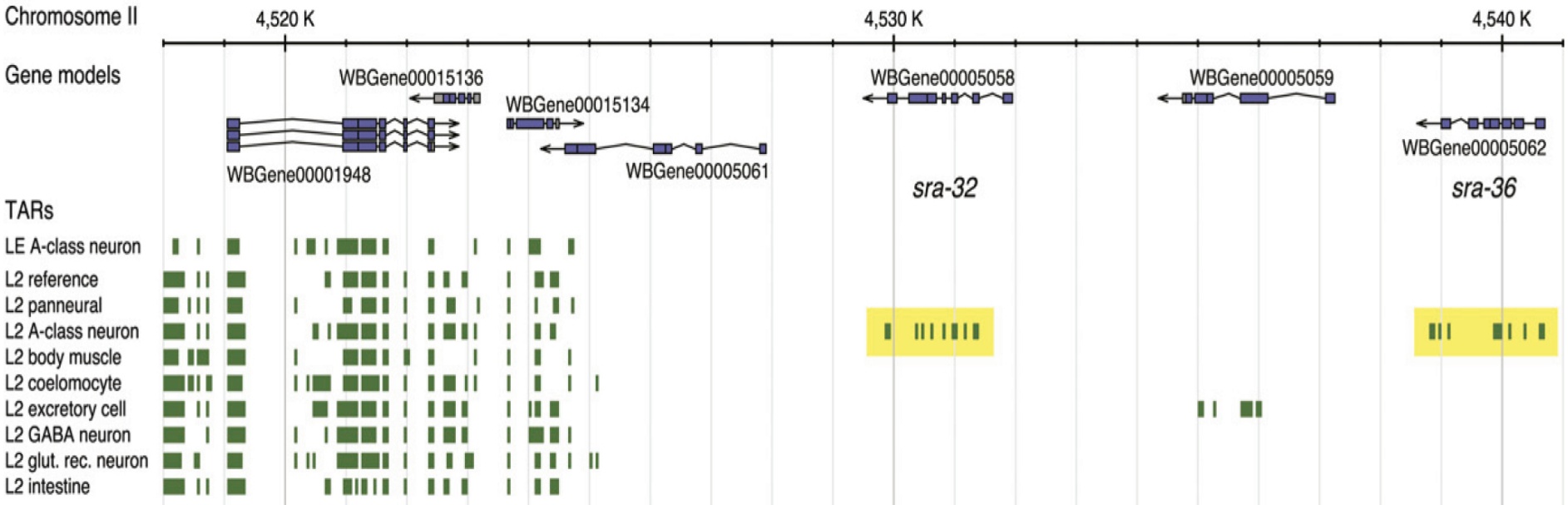




Protein domains enriched in neuronal cells

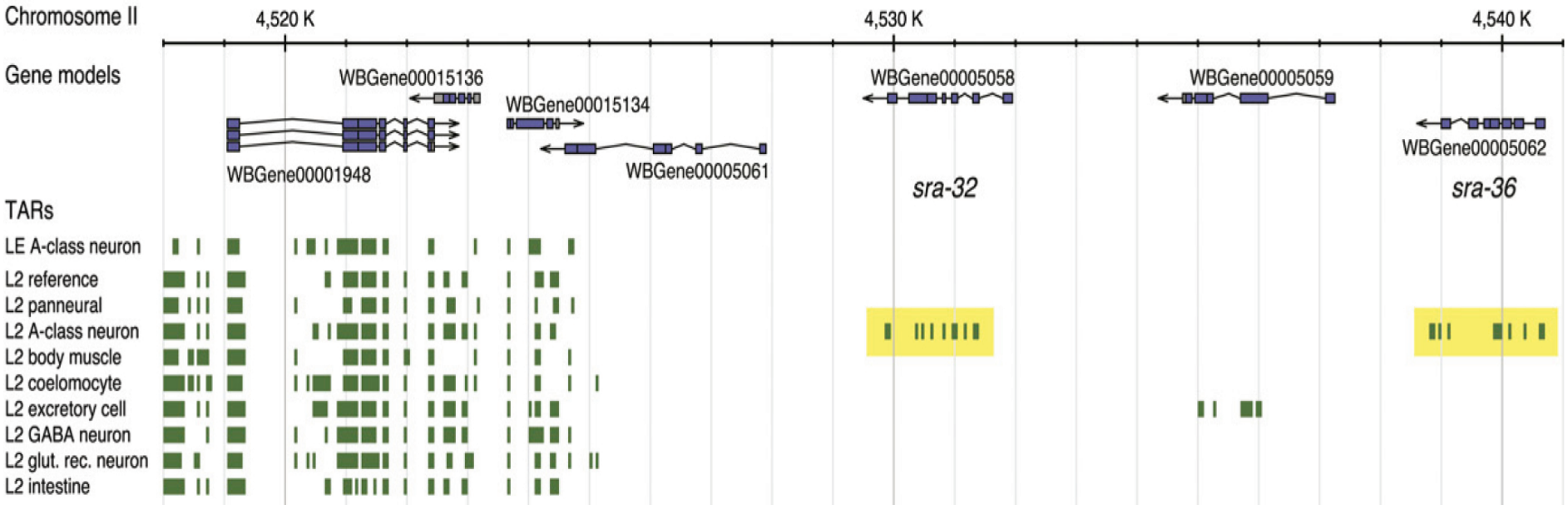
[Spencer et al., 2011]

Which genes are specifically expressed?



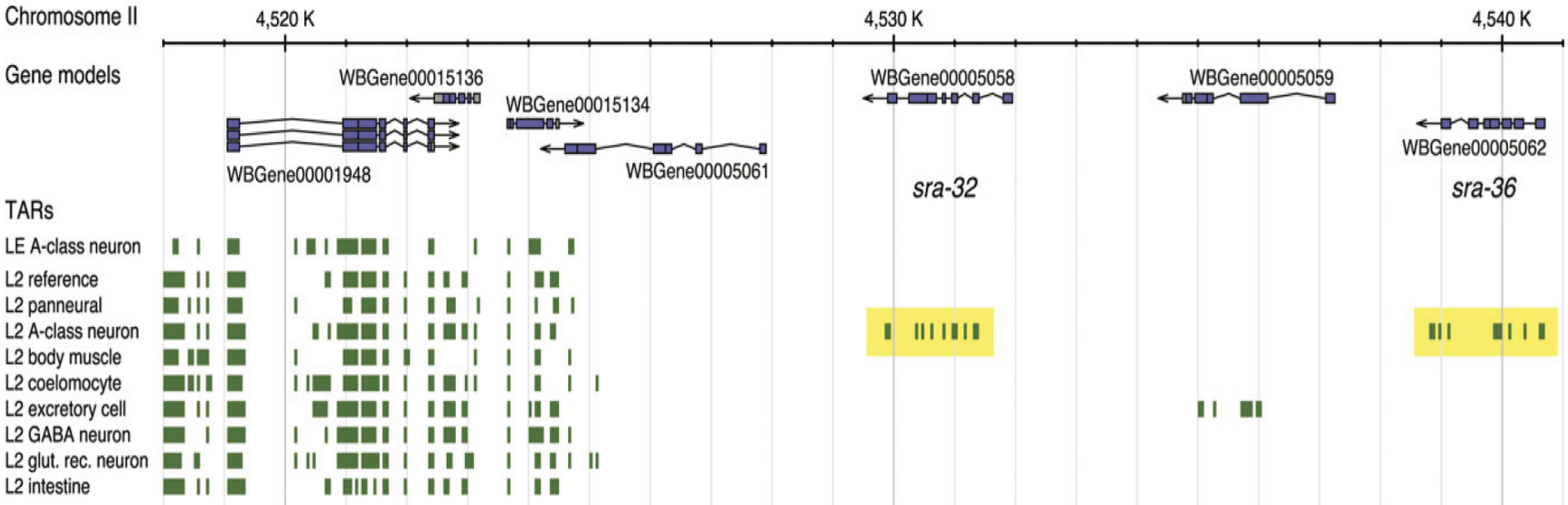
- 7TM-encoding genes selectively enriched in A-type motor neuron
- Under-representation of expressed genes on X chromosome in primordial cells (Z2/Z3 core) [Spencer et al., 2011]
- Guanylate cyclase *gcy-9* in BAG neurons required for CO_2 sensation, delineation of molecular pathway [Hallem et al., 2011]

Which genes are specifically expressed?



- 7TM-encoding genes selectively enriched in A-type motor neuron
- Under-representation of expressed genes on X chromosome in primordial cells (Z2/Z3 core) [Spencer et al., 2011]
- Guanylate cyclase *gcy-9* in BAG neurons required for CO_2 sensation, delineation of molecular pathway [Hallem et al., 2011]

Which genes are specifically expressed?



- 7TM-encoding genes selectively enriched in A-type motor neuron
- Under-representation of expressed genes on X chromosome in primordial cells (Z2/Z3 core) [Spencer et al., 2011]
- Guanylate cyclase *gcy-9* in BAG neurons required for CO_2 sensation, delineation of molecular pathway [Hallem et al., 2011]

Understanding Global Expression Patterns



- Gene expression atlas, many interesting special cases.
- How can we understand global patterns of regulation?

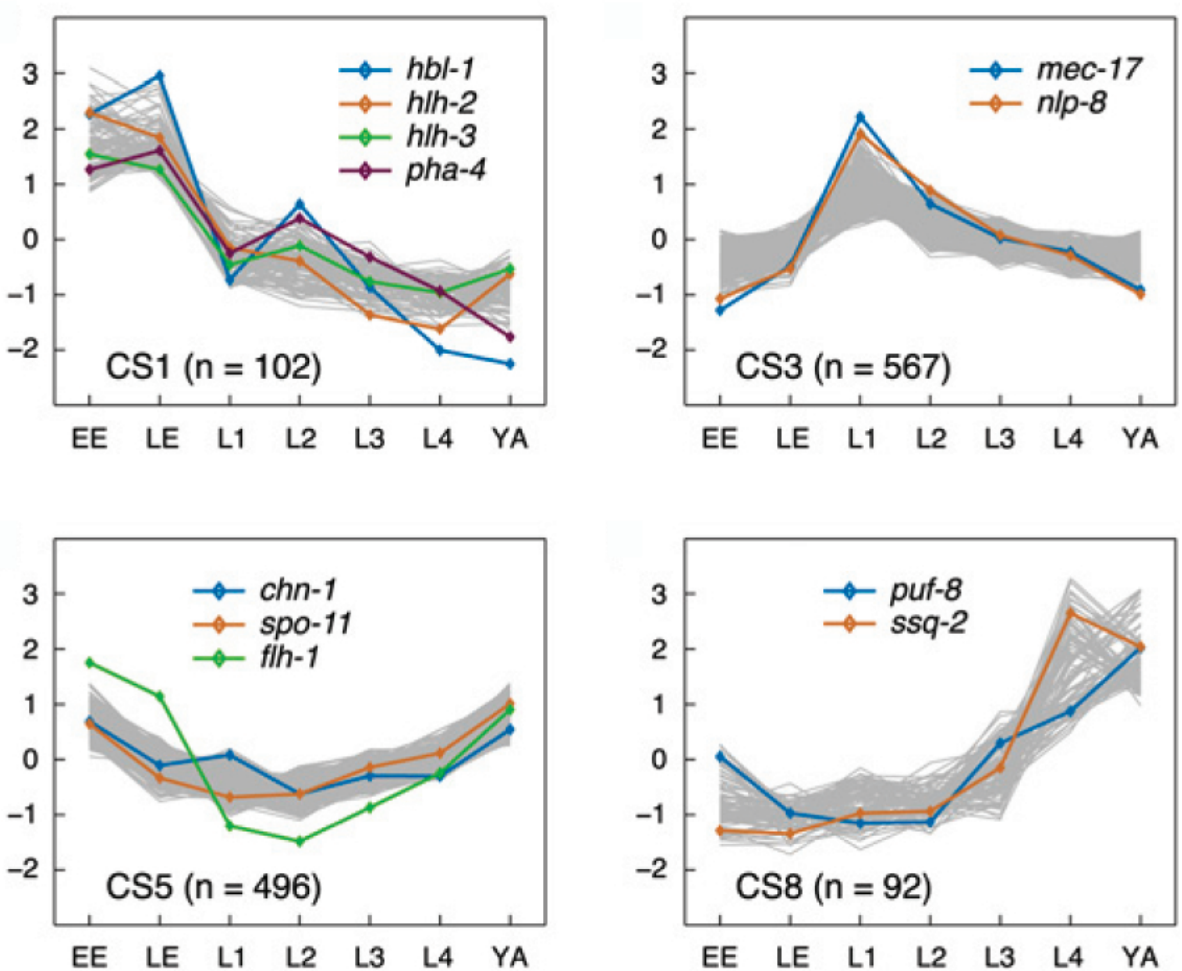
Group genes with similar expression profiles:

1. Clustering via Self-Organizing Maps (SOMs)
2. Analyse gene sets defined by clusters

EE=early embryos stage, LE=late embryos stages, L1-L4=larval stages, YA = young adults

Understanding Global Expression Patterns

- Gene expression atlas, many interesting special cases.
- How can we understand global patterns of regulation?



EE=early embryos stage, LE=late embryos stages, L1-L4=larval stages, YA = young adults

Group genes with similar expression profiles:

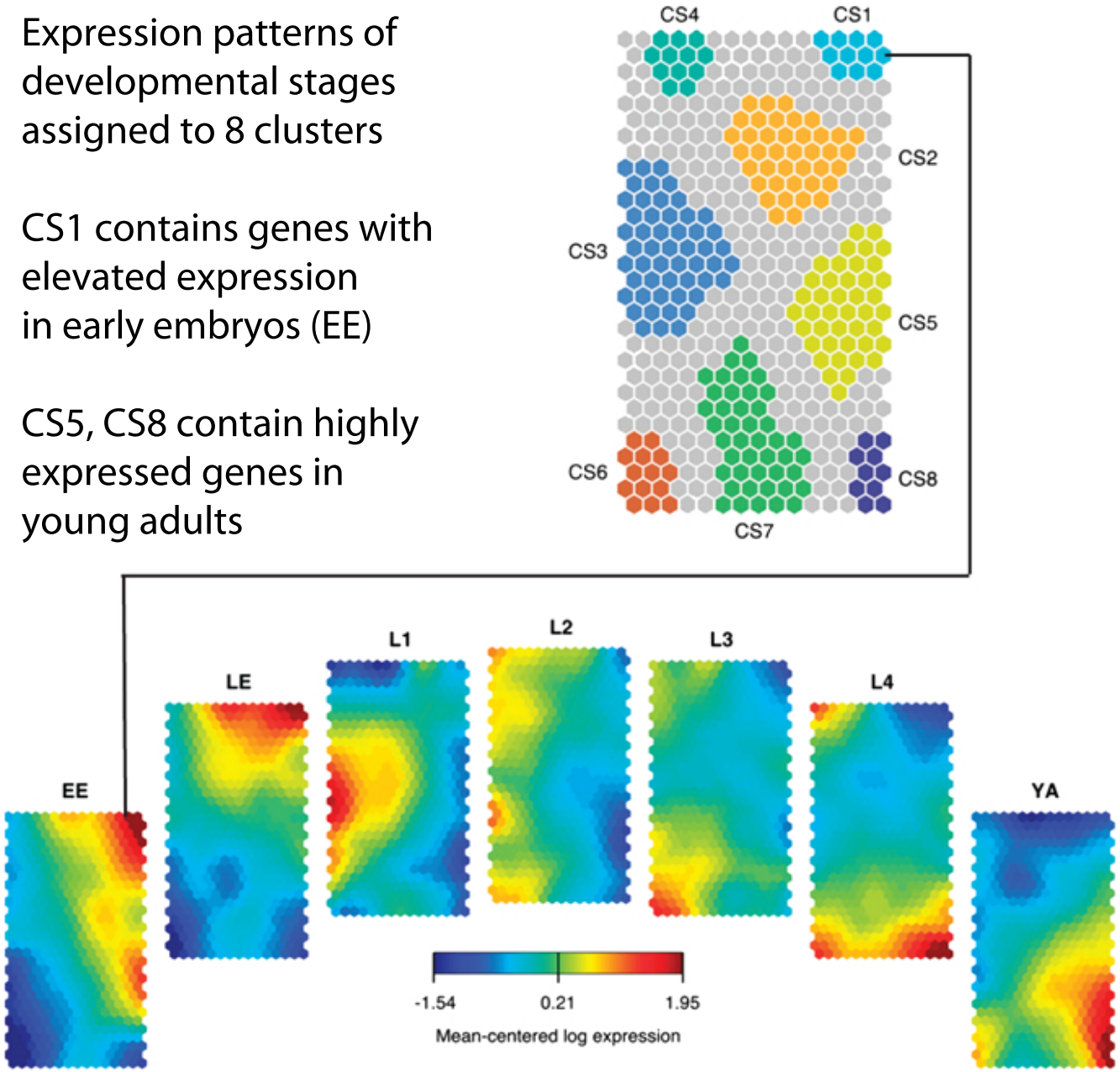
- Clustering via Self-Organizing Maps (SOMs)
- Analyse gene sets defined by clusters

Developmental Stage Specific Clusters

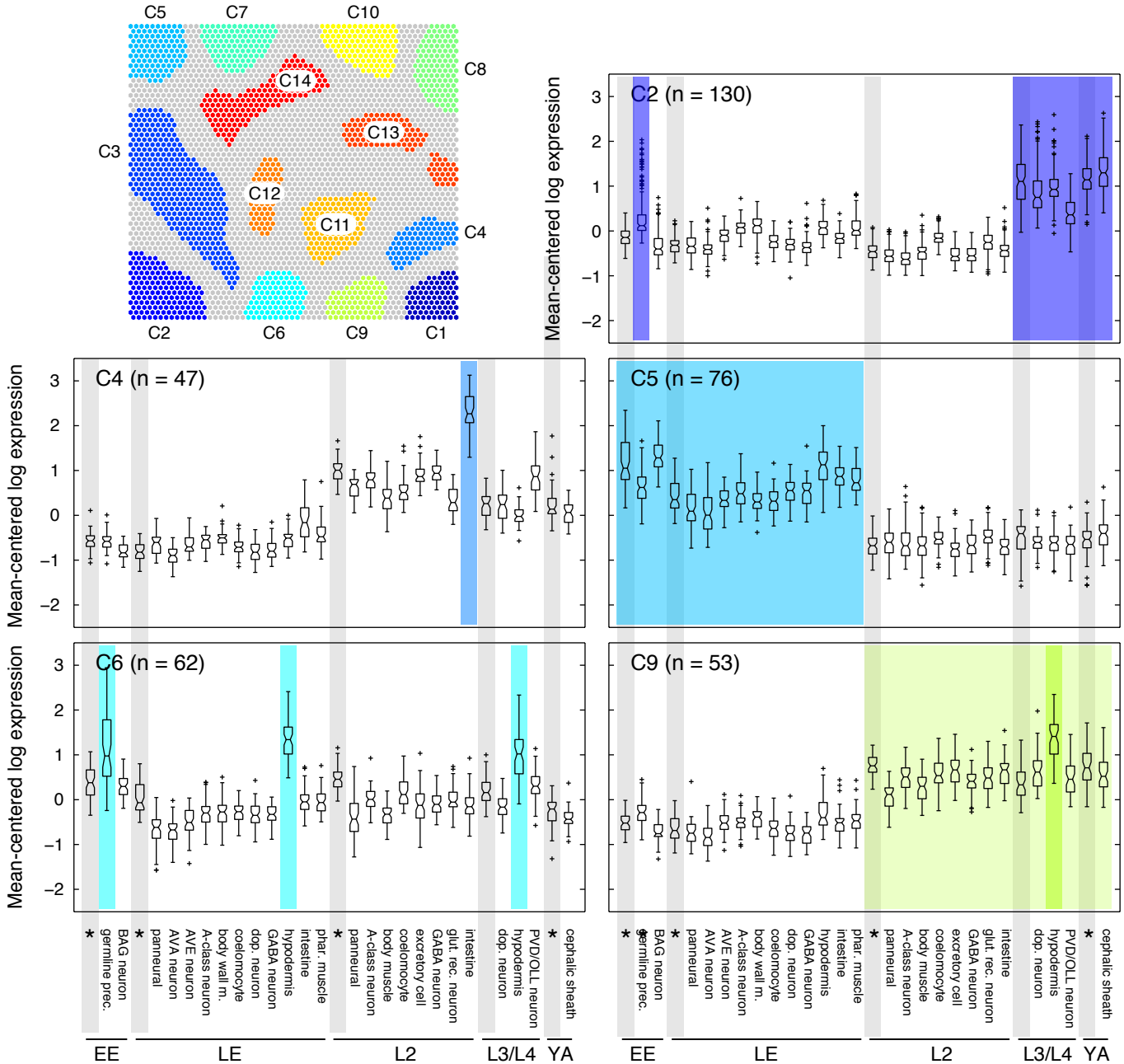
Expression patterns of developmental stages assigned to 8 clusters

CS1 contains genes with elevated expression in early embryos (EE)

CS5, CS8 contain highly expressed genes in young adults



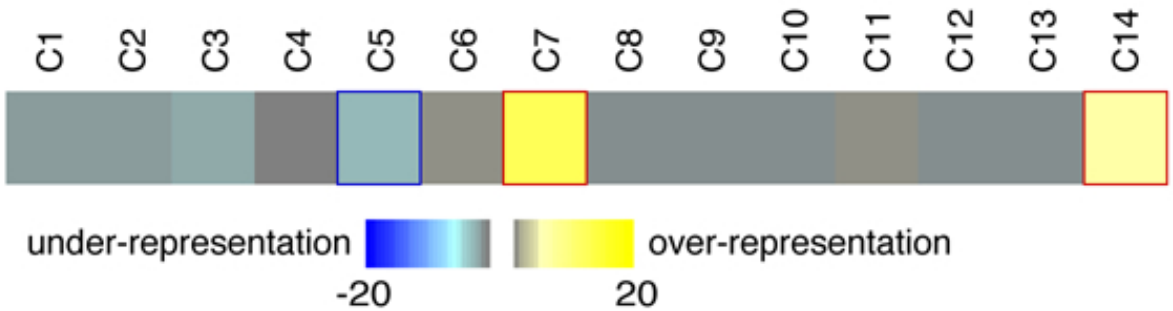
Cell-type Specific Clusters



Understanding Global Expression Patterns

Sequence motifs over- or underrepresented in SOM clusters
$\Rightarrow$ putative regulatory elements

Example: Motif related to vertebrate transcription factor (C/EBPalpha)

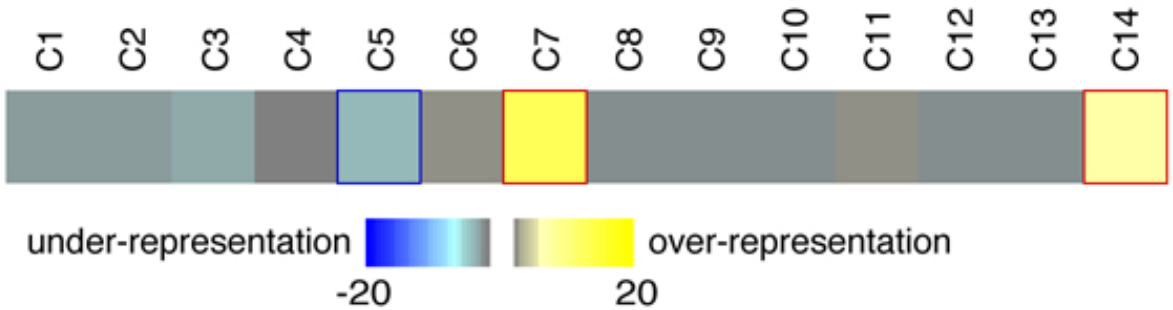




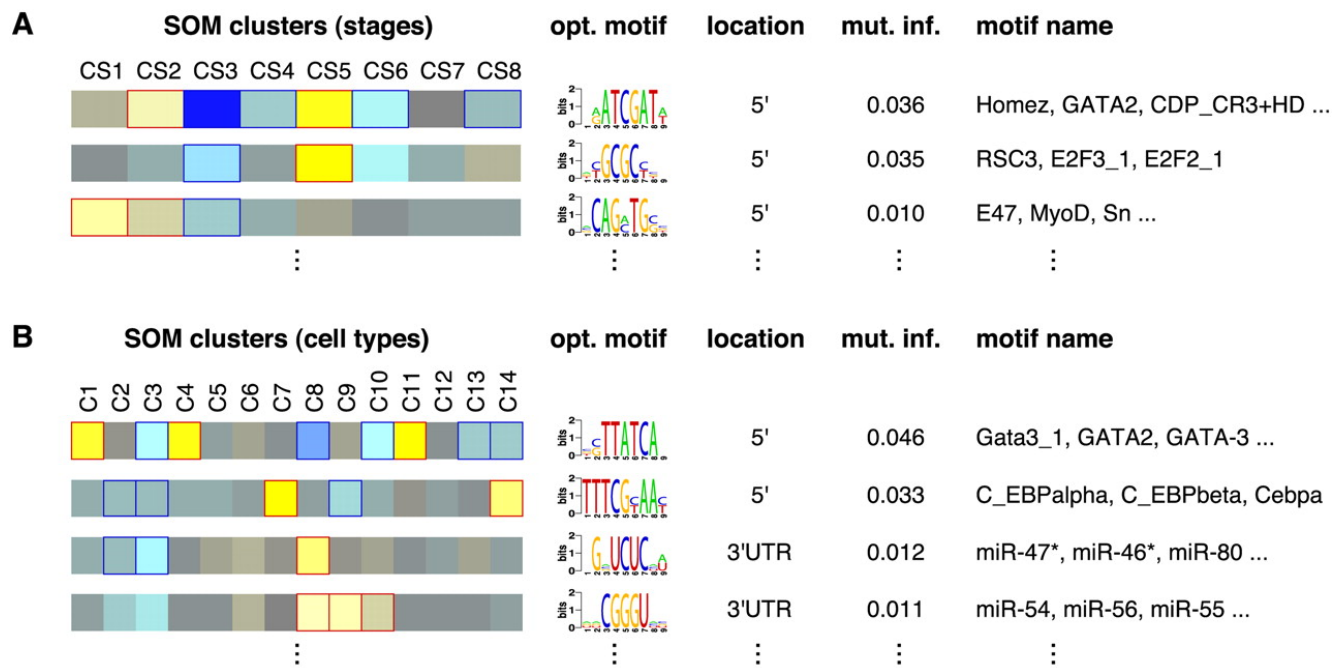
Understanding Global Expression Patterns

Sequence motifs over- or underrepresented in SOM clusters
$\Rightarrow$ putative regulatory elements

Example: Motif related to vertebrate transcription factor (C/EBPalpha)



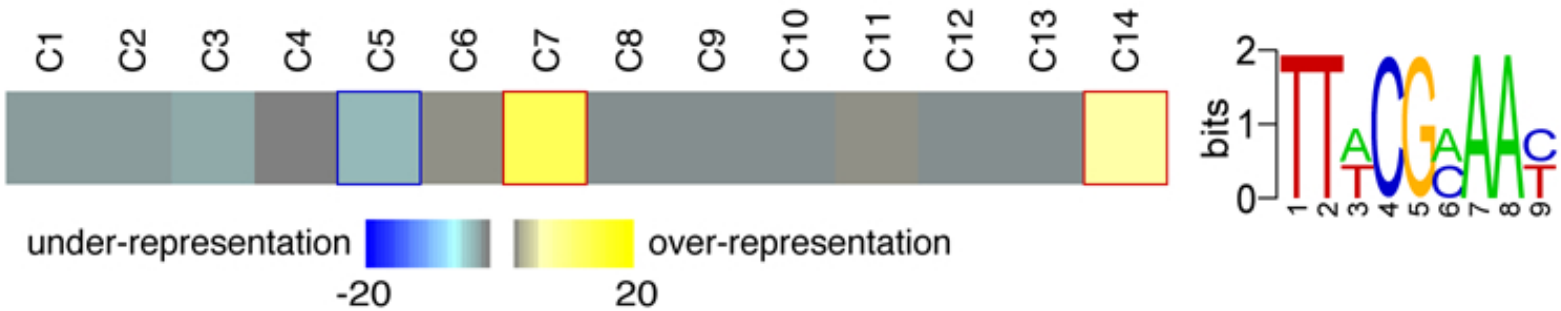




Understanding Global Expression Patterns

Sequence motifs over- or underrepresented in SOM clusters
$\Rightarrow$ putative regulatory elements

Example: Motif related to vertebrate transcription factor (C/EBPalpha)



Next steps:

- *cis*-regulatory modules for *predicting* gene expression [Schultheiss et al., 2009]
- Analyse sub-genic features (exons/introns/UTRs) [Eichner et al., 2011]
- Higher genomic resolution by RNA-seq (in prep.)
- Higher temporal & spatial resolution by single cell analysis

Digestion



We can

- reconstruct RNAs from experimental data (arrays/RNA-seq) and
- predict RNA transcripts from sequence and other *cis* information.

Limitations:

- Have incomplete annotations, imperfect inference algorithms.
- Ignore *trans*-effects, dynamics, spatial information .

Need:

$$f(DNA, \quad , \quad) = RNA$$

$$F(RNA, \quad , \quad) = protein$$

$$f'(DNA, \quad , \quad) = d\,RNA/d\,t$$

$$F'(RNA, \quad , \quad) = d\,protein/d\,t$$

Digestion

We can

- reconstruct RNAs from experimental data (arrays/RNA-seq) and
- predict RNA transcripts from sequence and other *cis* information.

Limitations:

- Have incomplete annotations, imperfect inference algorithms.
- Ignore *trans*-effects, dynamics, spatial information.

Need:

$$f(\text{DNA}, \boxed{1}_{cis/trans}, \boxed{2}_{cis/trans}) = \text{RNA}$$

$$F(\text{RNA}, \boxed{4}, \boxed{5}) = \text{protein}$$

$$f'(\text{DNA}, \quad , \quad) = d\,\text{RNA}/d\,t$$

$$F'(\text{RNA}, \quad , \quad) = d\,\text{protein}/d\,t$$

Digestion

We can

- reconstruct RNAs from experimental data (arrays/RNA-seq) and
- predict RNA transcripts from sequence and other *cis* information.

Limitations:

- Have incomplete annotations, imperfect inference algorithms.
- Ignore *trans*-effects, dynamics, spatial information.

Need:

$$f(\text{DNA}, \boxed{1}_{cis/trans}, \boxed{2}_{cis/trans}) = \text{RNA}$$

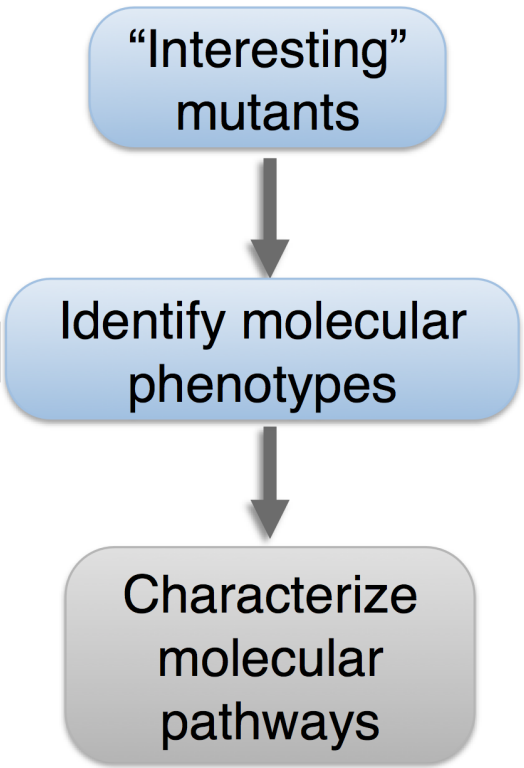
$$F(\text{RNA}, \boxed{4}, \boxed{5}) = \text{protein}$$

$$f'(\text{DNA}, \boxed{1}_{cis/trans}, \boxed{2}_{cis/trans}) = d\,\text{RNA}/d\,\text{t}$$

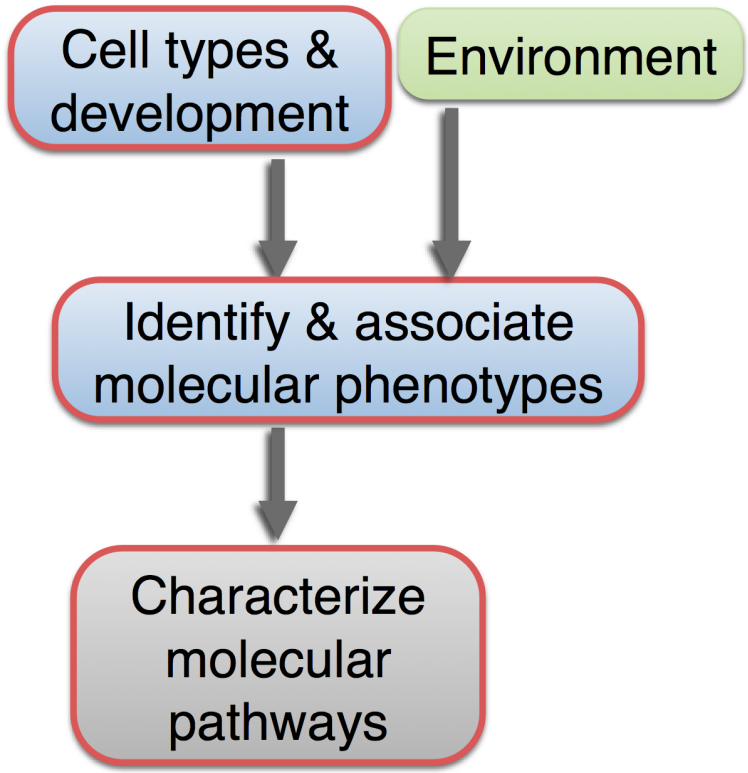
$$F'(\text{RNA}, \boxed{4}, \boxed{5}) = d\,\text{protein}/d\,\text{t}$$

Approaches to Understand Molecular Pathways

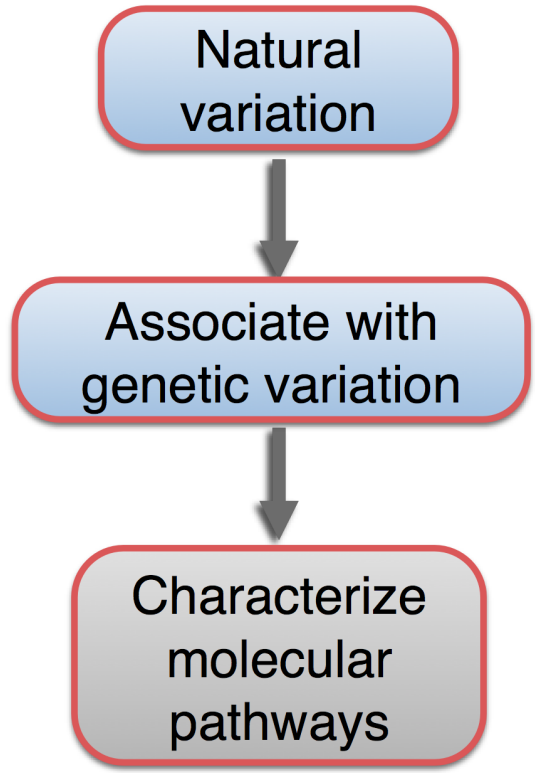




- CBP20/80 [Laubinger et al., 2008]
- sRNA biogenesis (*dcl1-4*) [Laubinger et al., 2010]
- NMD [w/ A. Wachter]
- Exosome [w/ E. Conti]



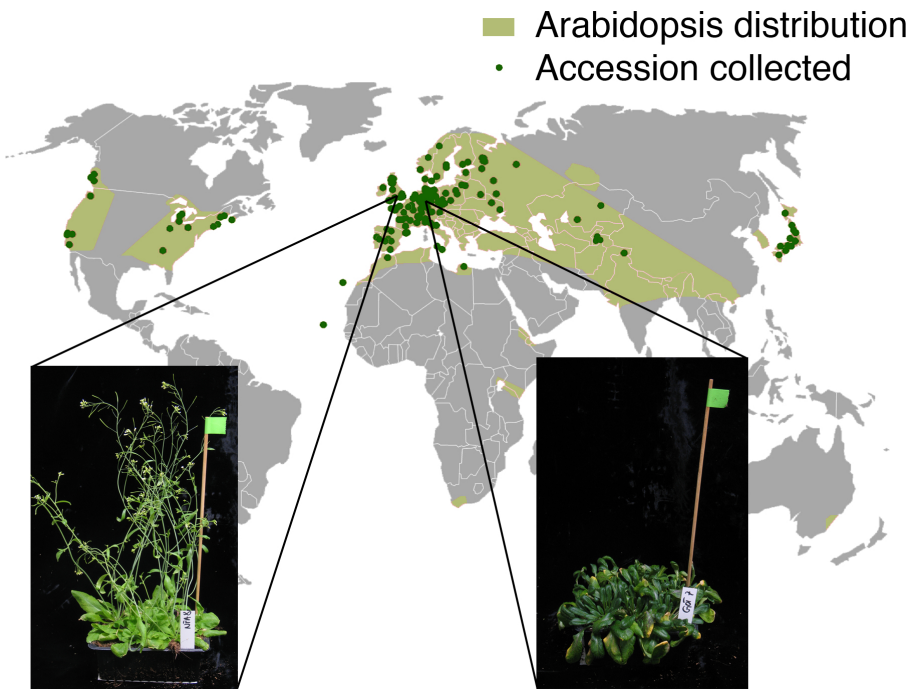
- *A. thaliana* tissue atlas [Laubinger et al., 2008]
- *A. thaliana* stress response [Zeller et al., 2009]
- *C. elegans* tissue/development atlas [Gerstein et al., 2010, Spencer et al., 2011]



- *A. thaliana* genome variation [Clark et al., 2007a]
- *O. sativa* genome variation [McNally et al., 2009]
- *A. thaliana* transcriptome variation [w/ R. Mott & R. Clark]

Multiple reference genomes and transcriptomes for A. thaliana

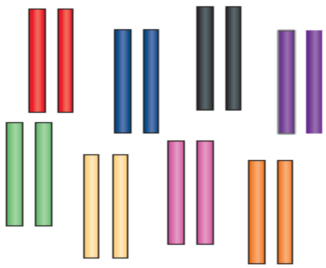
G. Xiangchao, et al., G. Rätsch, R. Mott, *Nature* (August 28, 2011)



- 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 & 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)

18 +1 reference genomes and transcriptomes for A. thaliana

Multiparent Advanced Generation Inter-Cross (MAGIC) lines



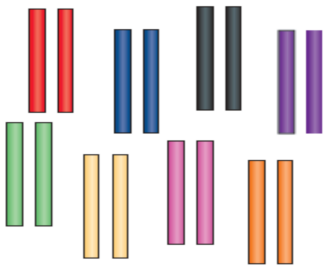
n founders

- Eventually 1200 lines
- Imputation of magic lines genomes based on founders
- Powerful QTL mapping population

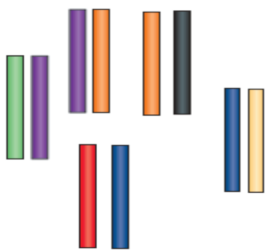
This work: Genome and transcriptome analysis of the MAGIC lines founders

18 +1 reference genomes and transcriptomes for A. thaliana

Multiparent Advanced Generation Inter-Cross (MAGIC) lines



n founders



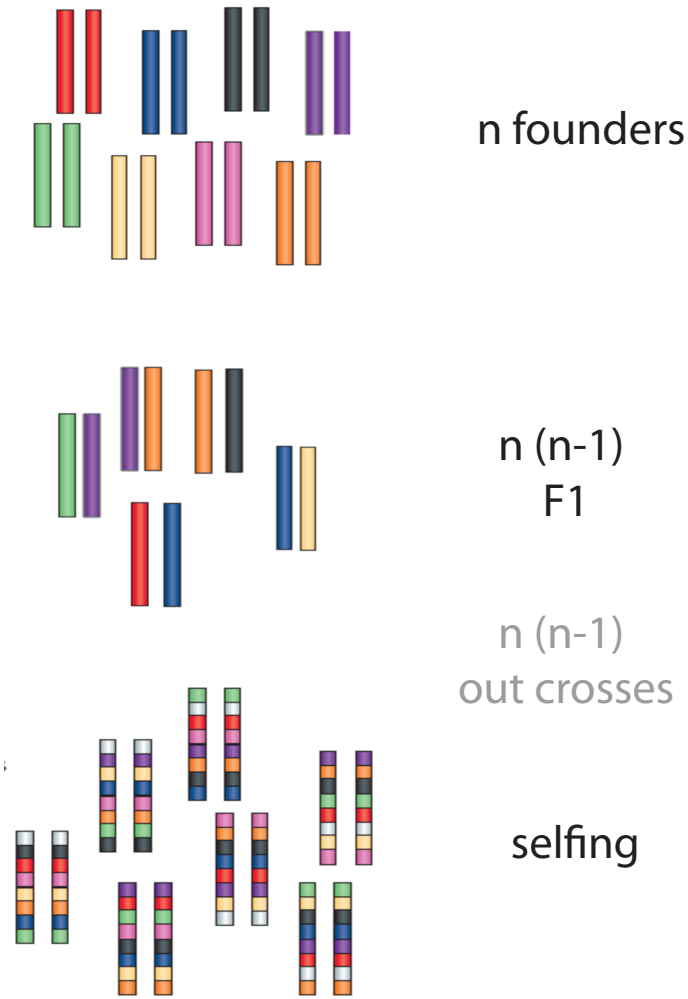
n (n-1)
F1

- Eventually 1200 lines
- Imputation of magic lines genomes based on founders
- Powerful QTL mapping population

This work: Genome and transcriptome analysis of the MAGIC lines founders

18 +1 reference genomes and transcriptomes for A. thaliana

Multiparent Advanced Generation Inter-Cross (MAGIC) lines

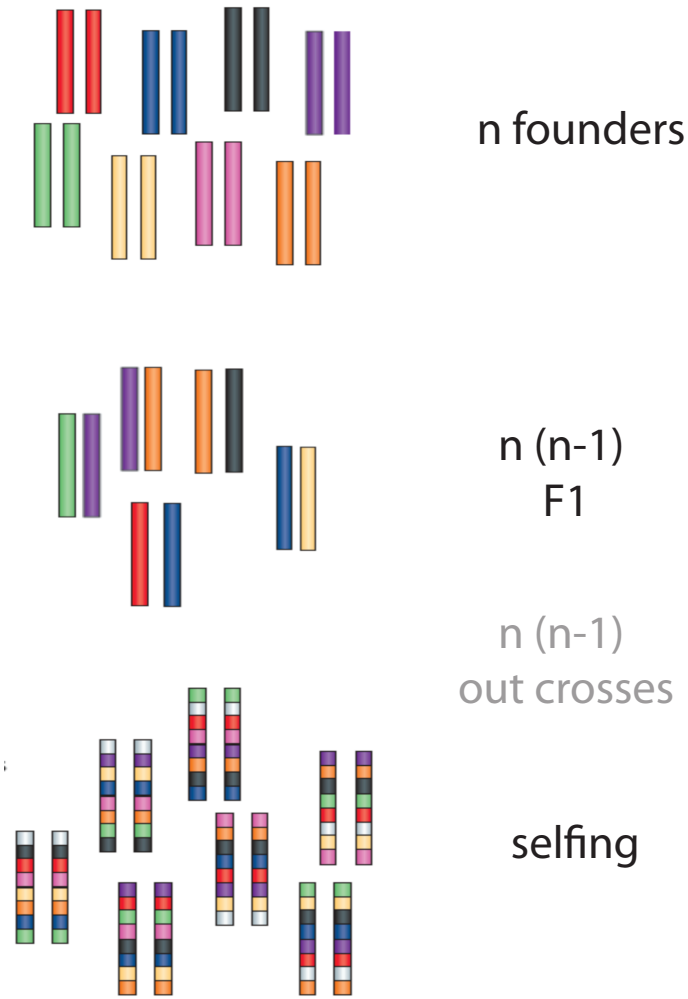


- Eventually 1200 lines
- Imputation of magic lines genomes based on founders
- Powerful QTL mapping population

This work: Genome and transcriptome analysis of the MAGIC lines founders

18 +1 reference genomes and transcriptomes for A. thaliana

Multiparent Advanced Generation Inter-Cross (MAGIC) lines

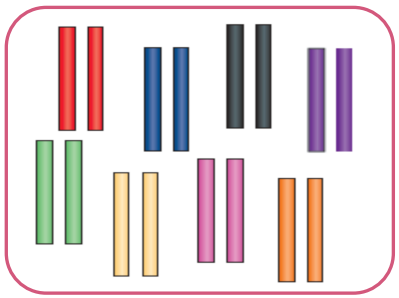


- Eventually 1200 lines
- Imputation of magic lines genomes based on founders
- Powerful QTL mapping population

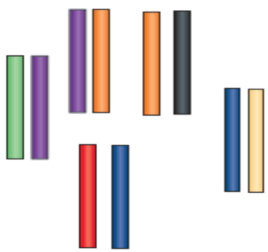
This work: Genome and transcriptome analysis of the MAGIC lines founders

18 +1 reference genomes and transcriptomes for A. thaliana

Multiparent Advanced Generation Inter-Cross (MAGIC) lines

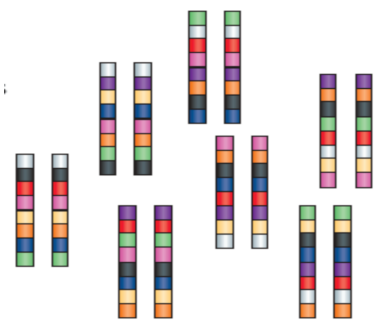


n founders



n (n-1)
F1

n (n-1)
out crosses

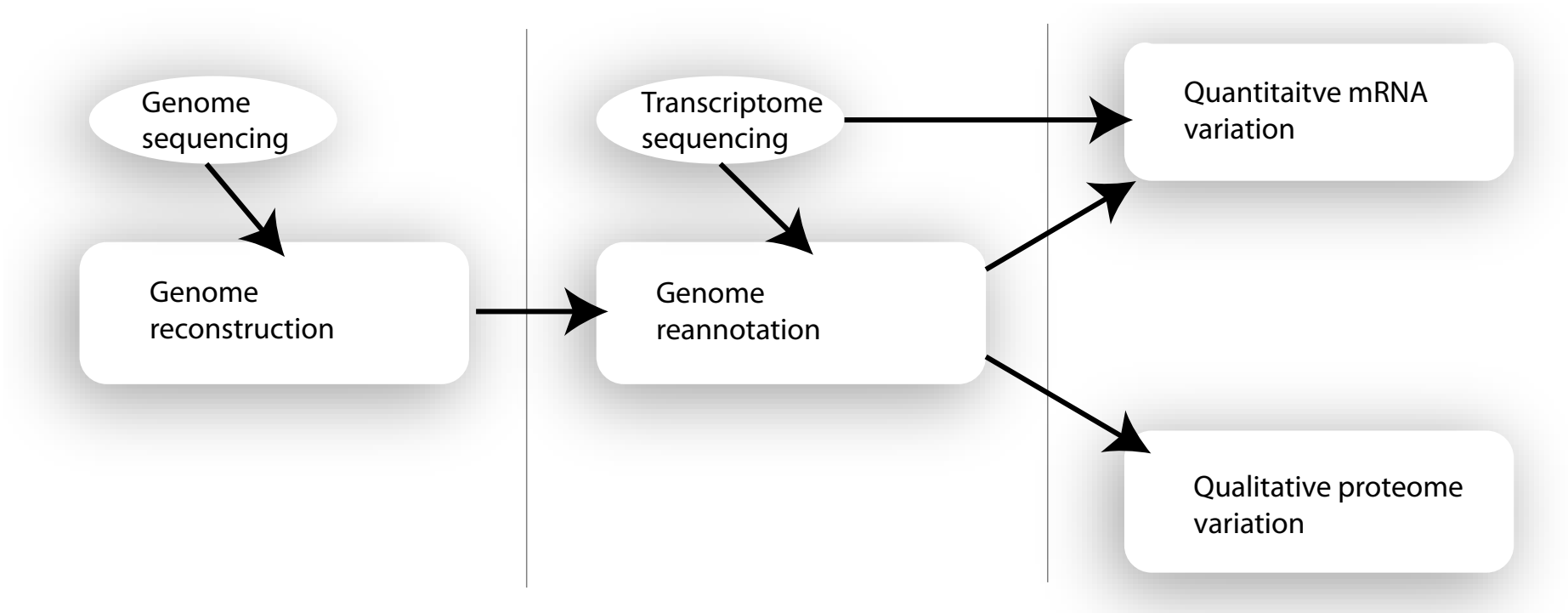


selfing

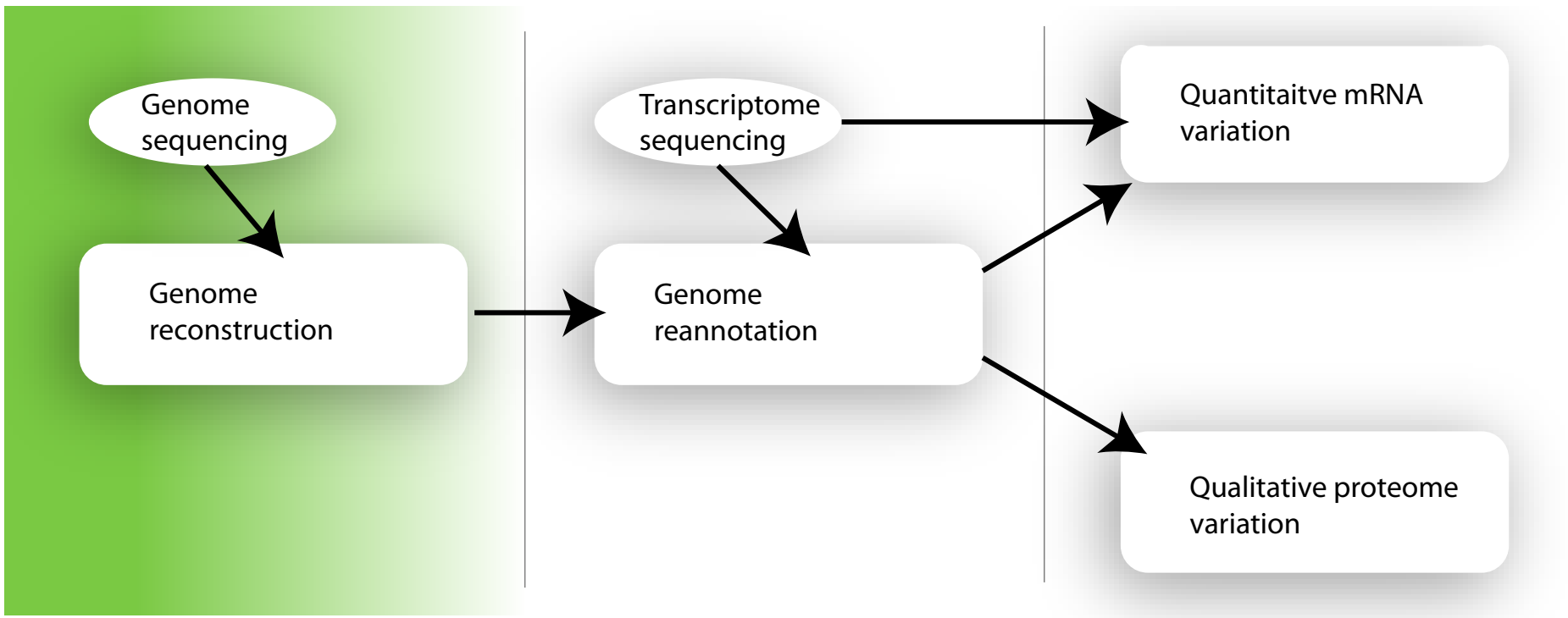
- Eventually 1200 lines
- Imputation of magic lines genomes based on founders
- Powerful QTL mapping population

This work: Genome and transcriptome analysis of the MAGIC lines founders

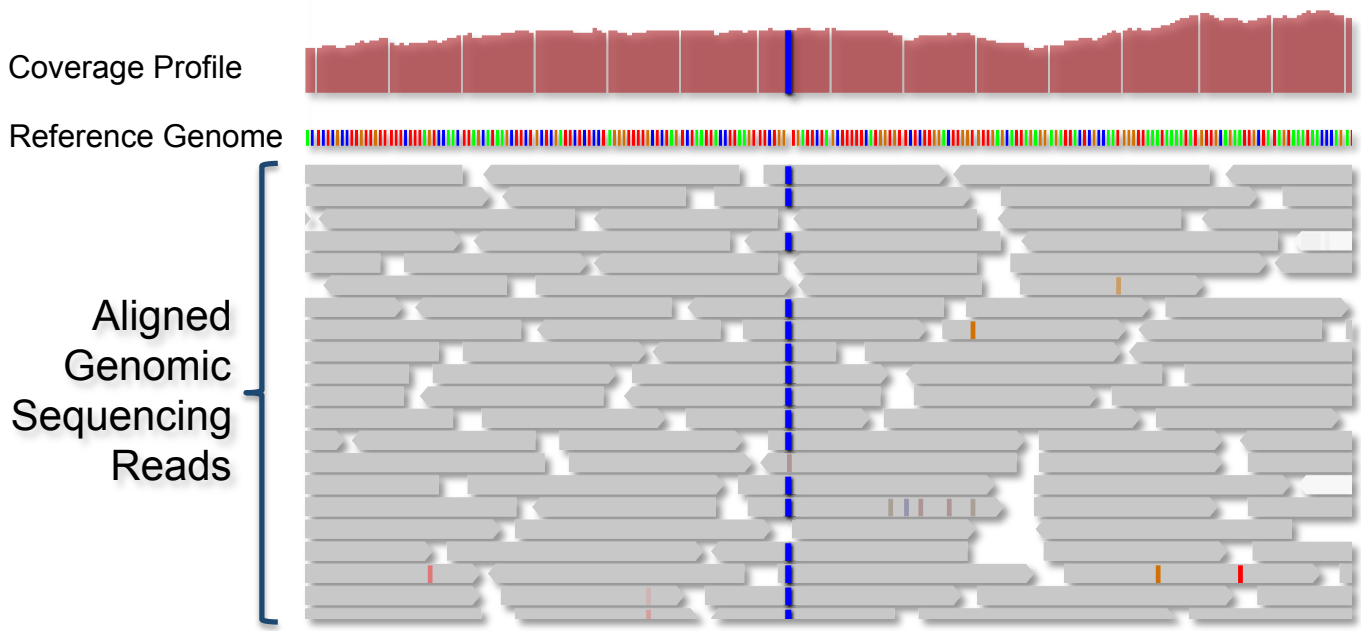
Outline


Genome sequencing
Transcriptome sequencing
Quantitaitve mRNA variation
Genome reconstruction
Genome reannotation
Qualitative proteome variation

Outline


Genome sequencing
Genome reconstruction
Transcriptome sequencing
Genome reannotation
Quantitaitve mRNA variation
Qualitative proteome variation

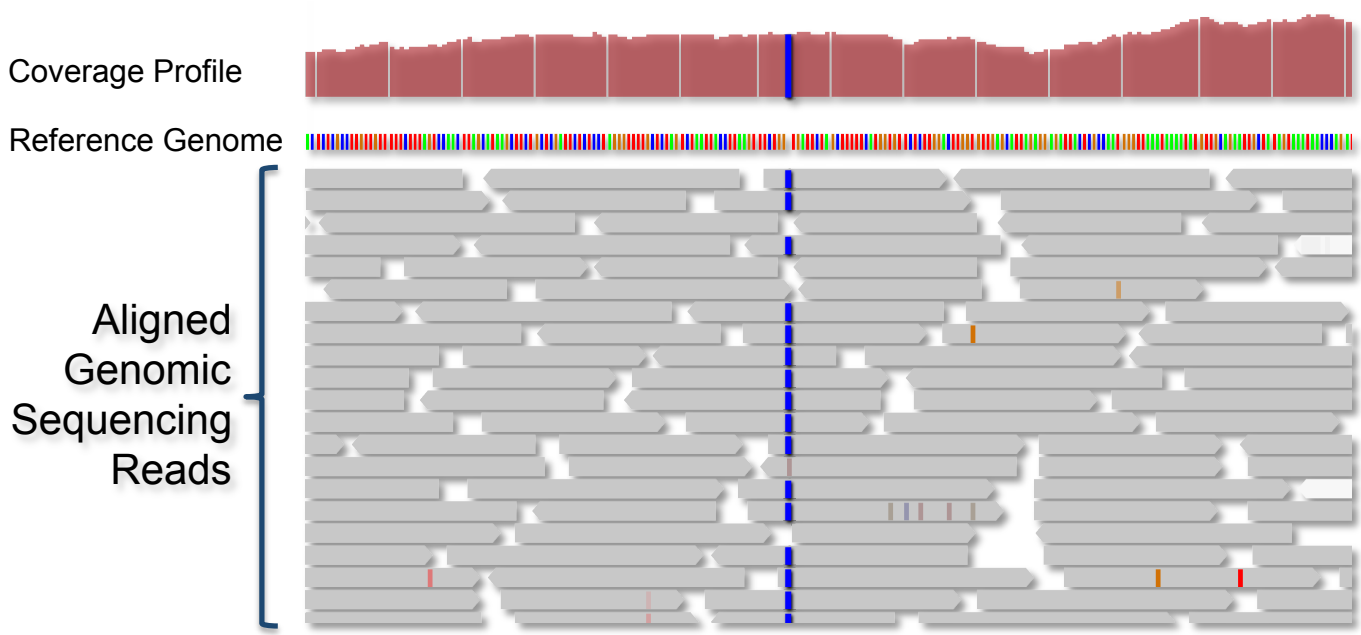
Sequencing 18 Arabidopsis genomes



- ≈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
- Extensive genomic variation (3Mbp SNPs, 15Mbp indels)

$\Rightarrow \leq 1$ error / 10kb $\Rightarrow$ good enough for genome reannotations

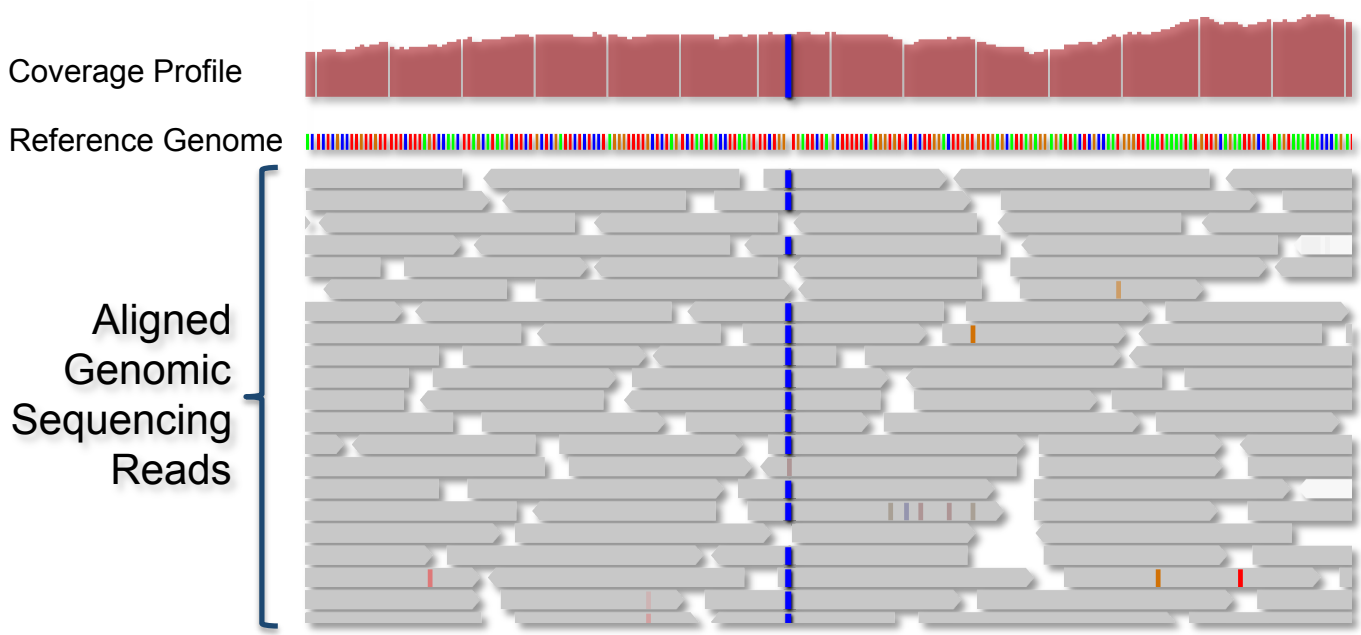
Sequencing 18 Arabidopsis genomes



- ≈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
- Extensive genomic variation (3Mbp SNPs, 15Mbp indels)

$\Rightarrow \leq 1$ error / 10kb $\Rightarrow$ good enough for genome reannotations

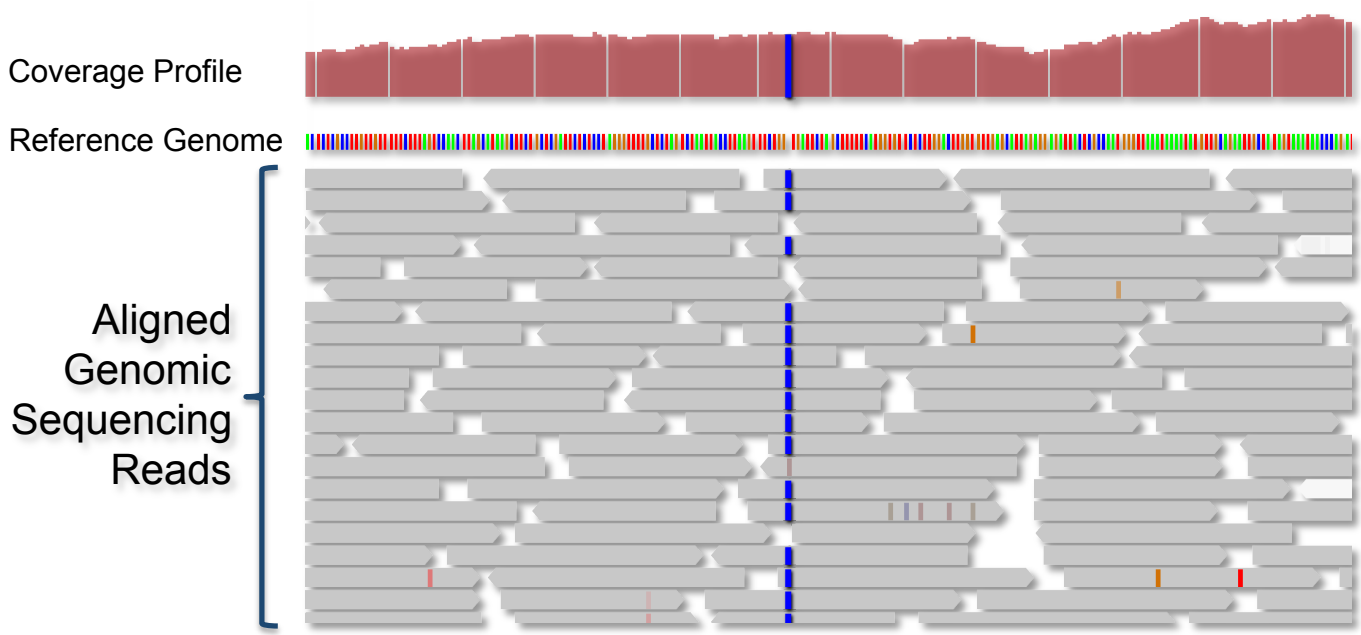
Sequencing 18 Arabidopsis genomes



- ≈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
- Extensive genomic variation (3Mbp SNPs, 15Mbp indels)

$\Rightarrow \leq 1$ error / 10kb $\Rightarrow$ good enough for genome reannotations

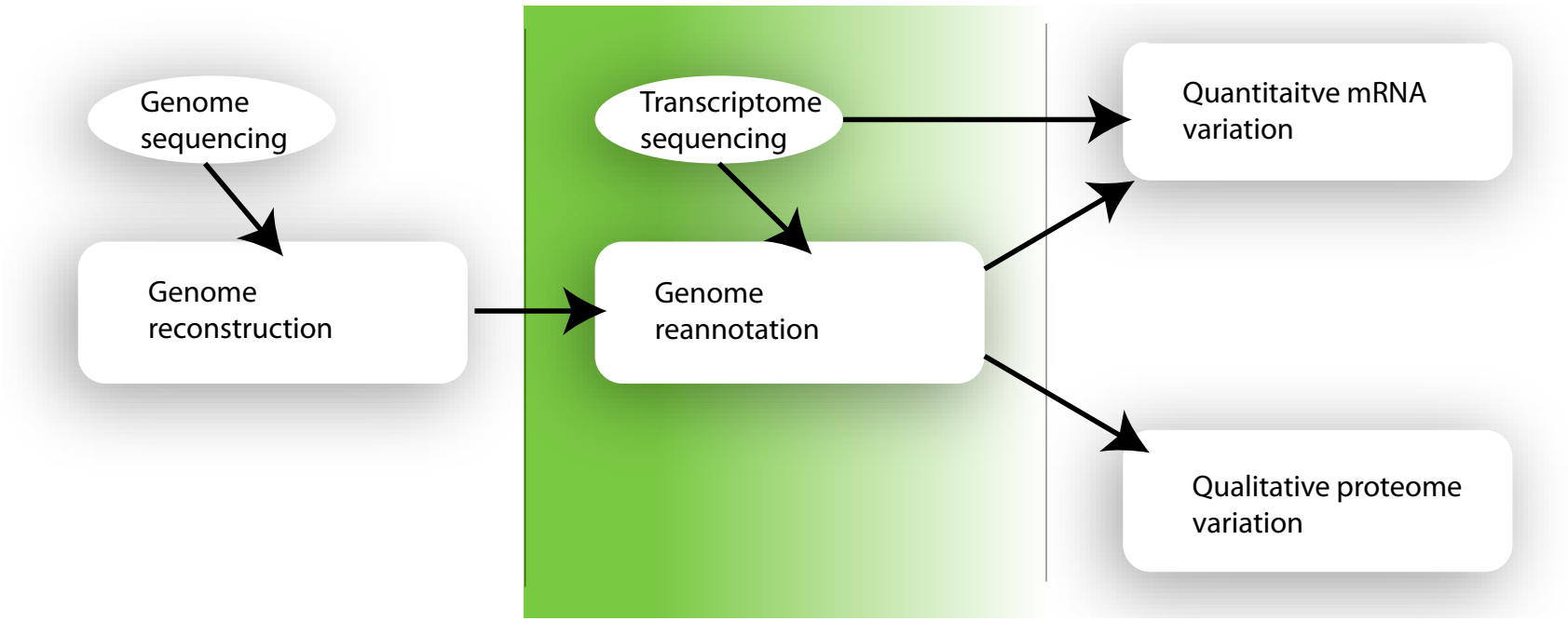
Sequencing 18 Arabidopsis genomes



- ≈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
- Extensive genomic variation (3Mbp SNPs, 15Mbp indels)

$\Rightarrow \leq 1$ error / 10kb $\Rightarrow$ good enough for genome reannotations

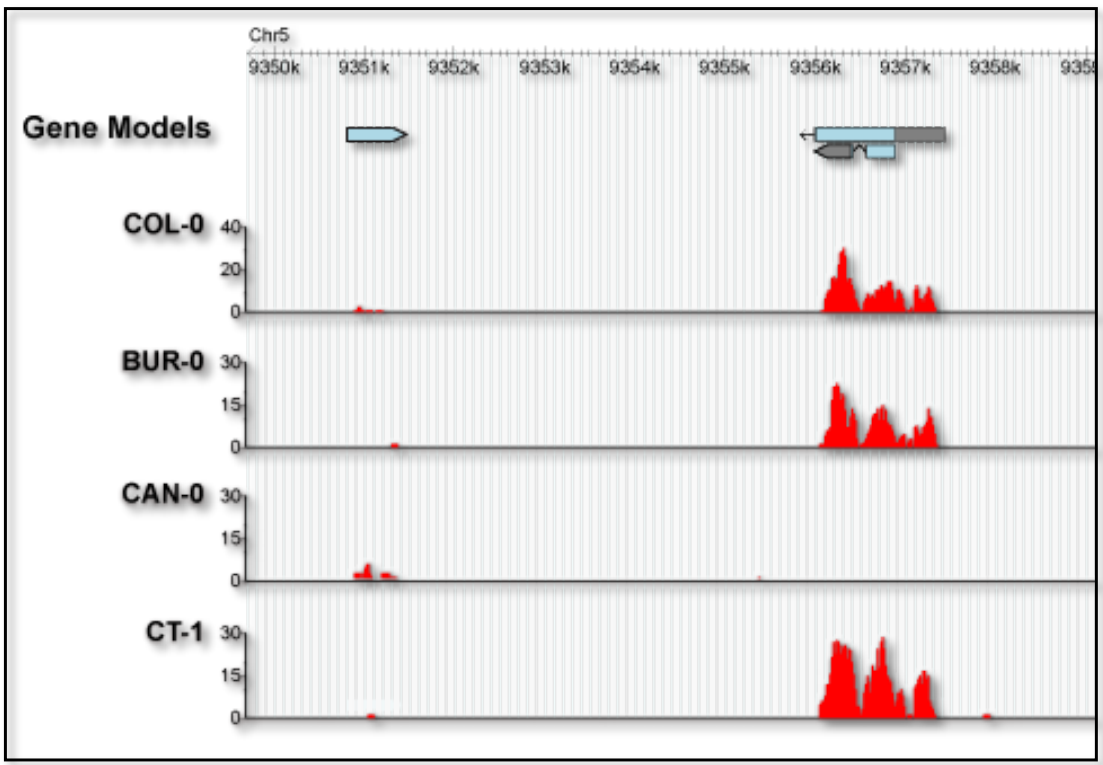
Outline


Genome sequencing
Transcriptome sequencing
Quantitaitve mRNA variation
Genome reconstruction
Genome reannotation
Qualitative proteome variation

Transcriptome Sequencing

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

- Strand-specific, 50 million reads per strain
- Non-strand-specific, 13 million reads in two replicates

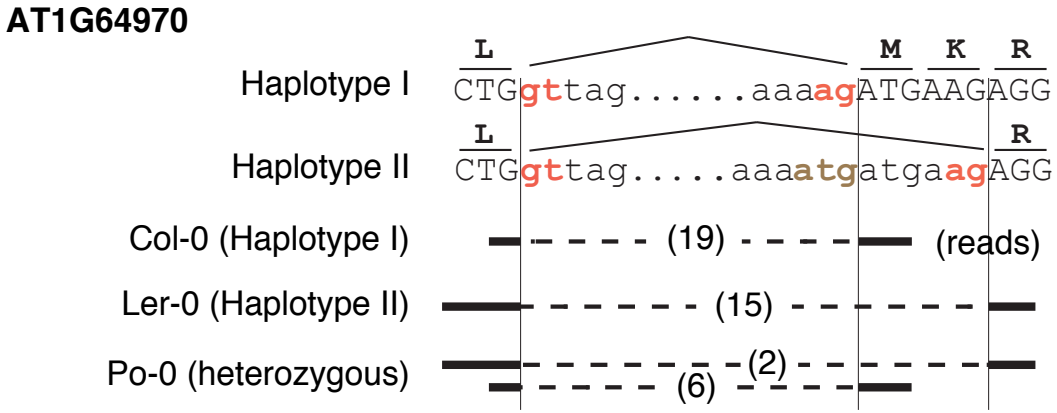


- Validation of genome sequence
- Genome reannotation

Transcriptome Sequencing

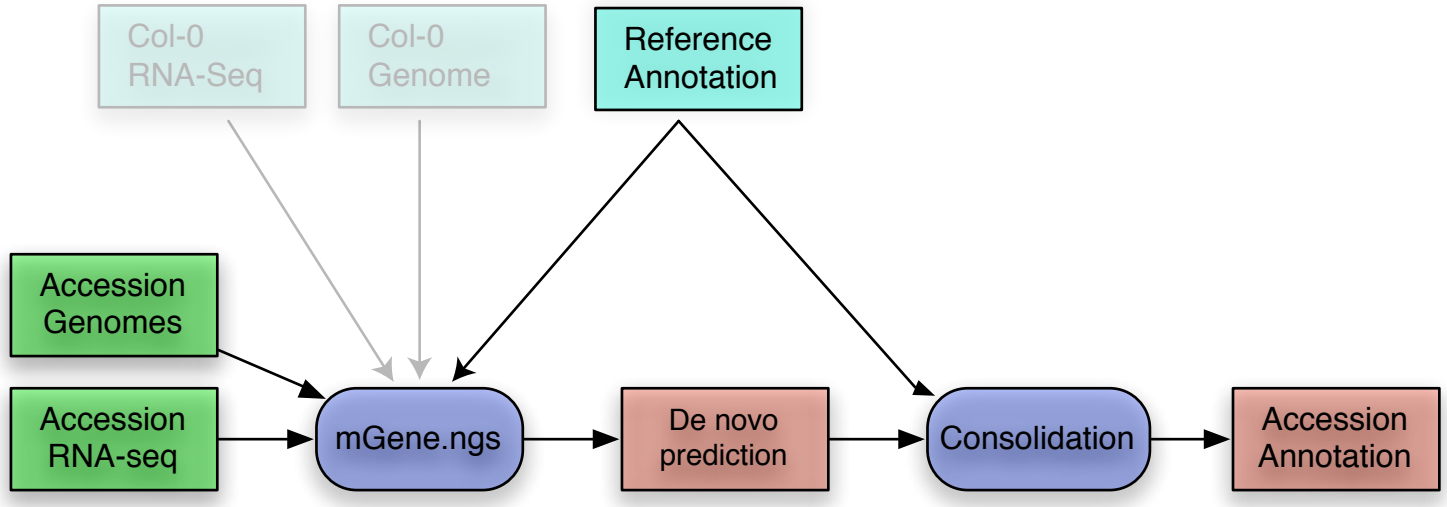
RNA-seq (Illumina, 76bp) of founder accessions:

- Strand-specific, 50 million reads per strain
- Non-strand-specific, 13 million reads in two replicates



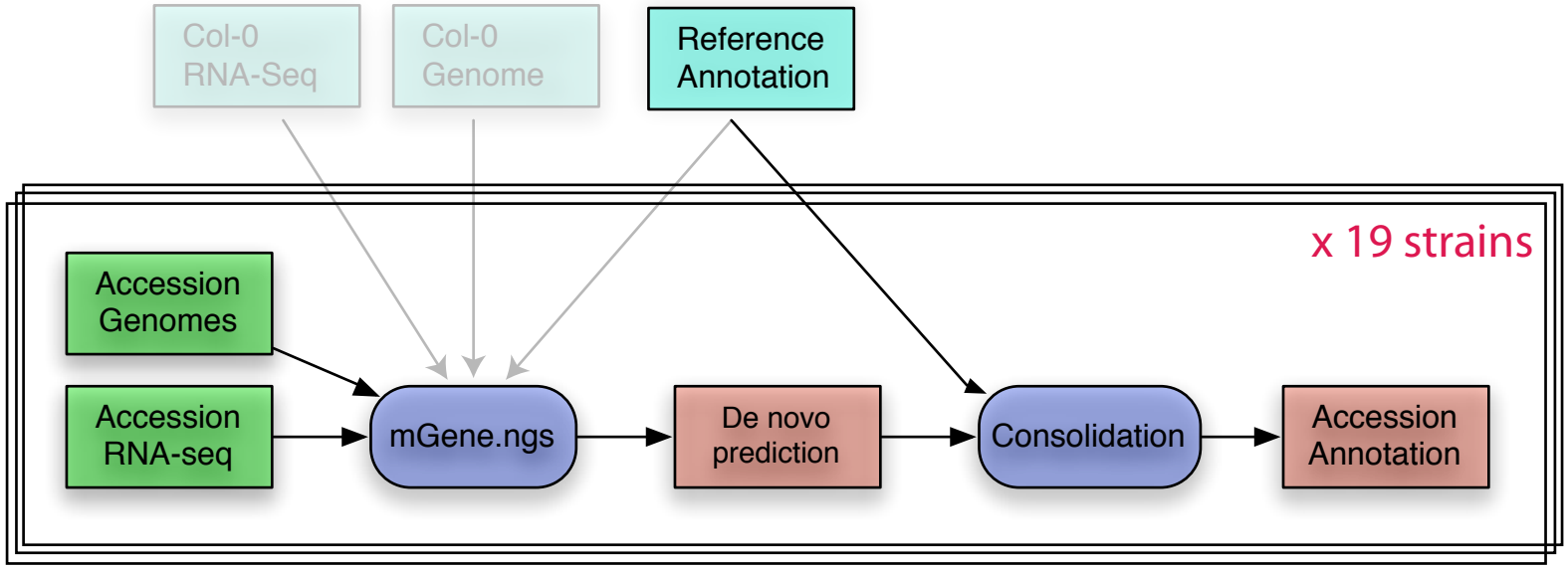
- Validation of genome sequence
- **Genome reannotation**

Founder Genome Reannotation



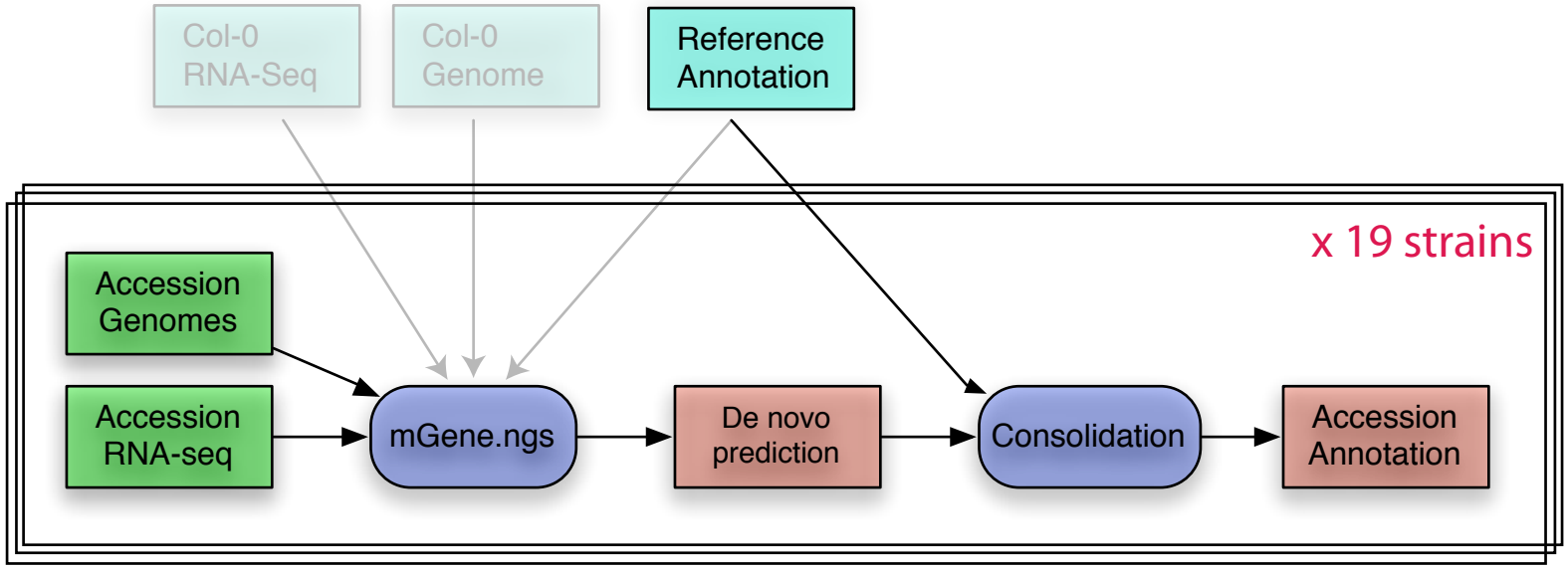
- Training on col-0 reference
- Strain-specific predictions, accounting for strain coordinate system
- $>$ 8,000 genes with new transcript annotations
- Per accession: $\approx$ 1,500 new transcripts, $\approx$ 650 new introns, $\approx$ 500 new TIS/STOP sites

Founder Genome Reannotation



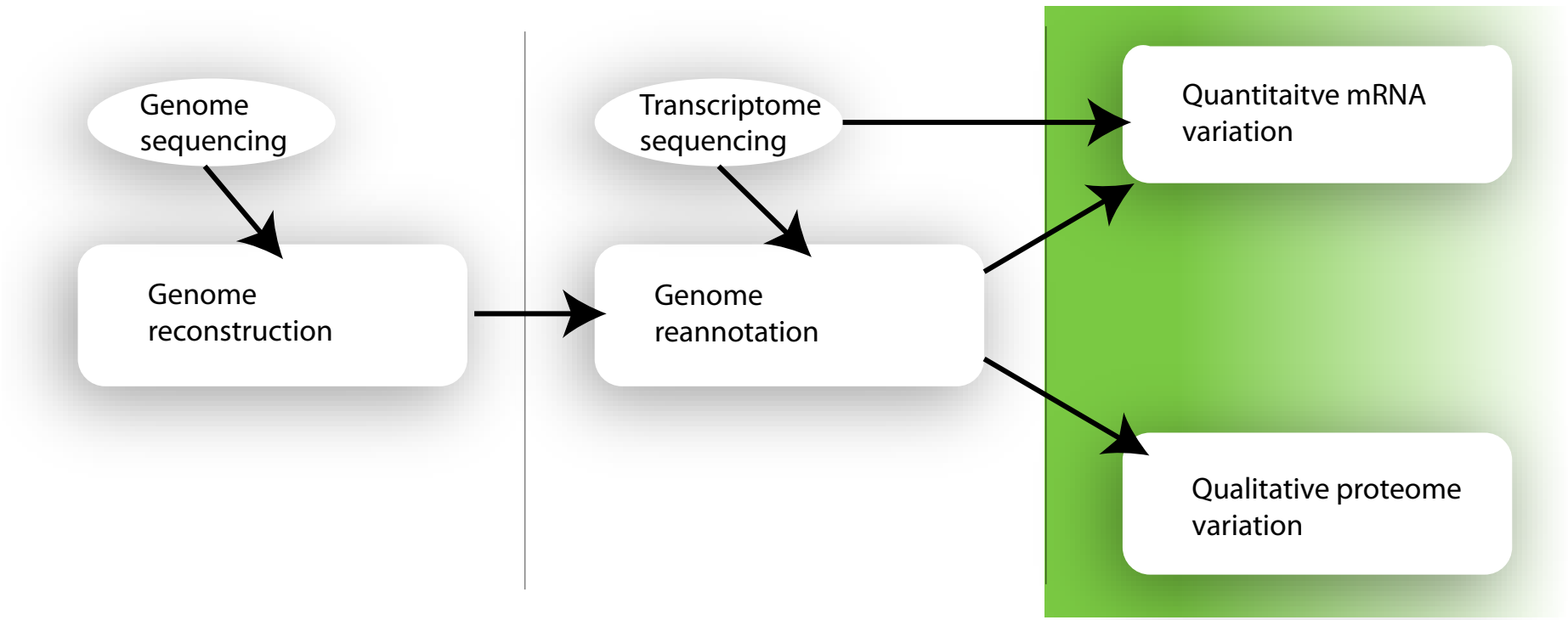
- Training on col-0 reference
- Strain-specific predictions, accounting for strain coordinate system
- $>$ 8,000 genes with new transcript annotations
- Per accession: $\approx$ 1,500 new transcripts, $\approx$ 650 new introns, $\approx$ 500 new TIS/STOP sites

Founder Genome Reannotation



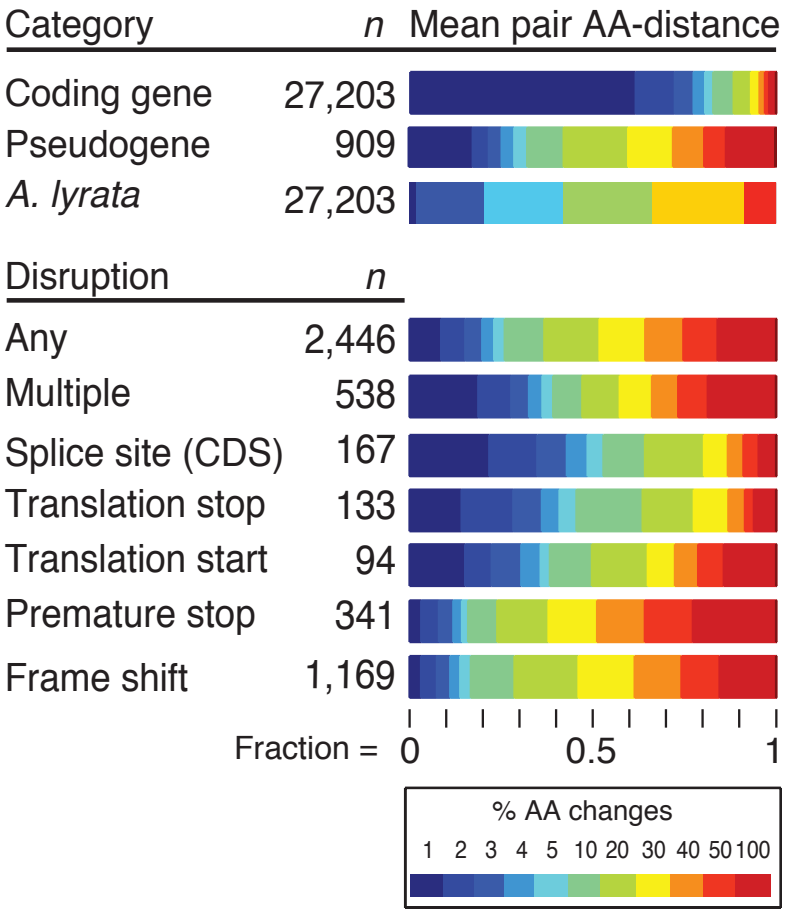
- Training on col-0 reference
- Strain-specific predictions, accounting for strain coordinate system
- $>$ 8,000 genes with new transcript annotations
- Per accession: $\approx$ 1,500 new transcripts, $\approx$ 650 new introns, $\approx$ 500 new TIS/STOP sites

Outline


Genome sequencing
Transcriptome sequencing
Quantitaitve mRNA variation
Genome reconstruction
Genome reannotation
Qualitative proteome variation

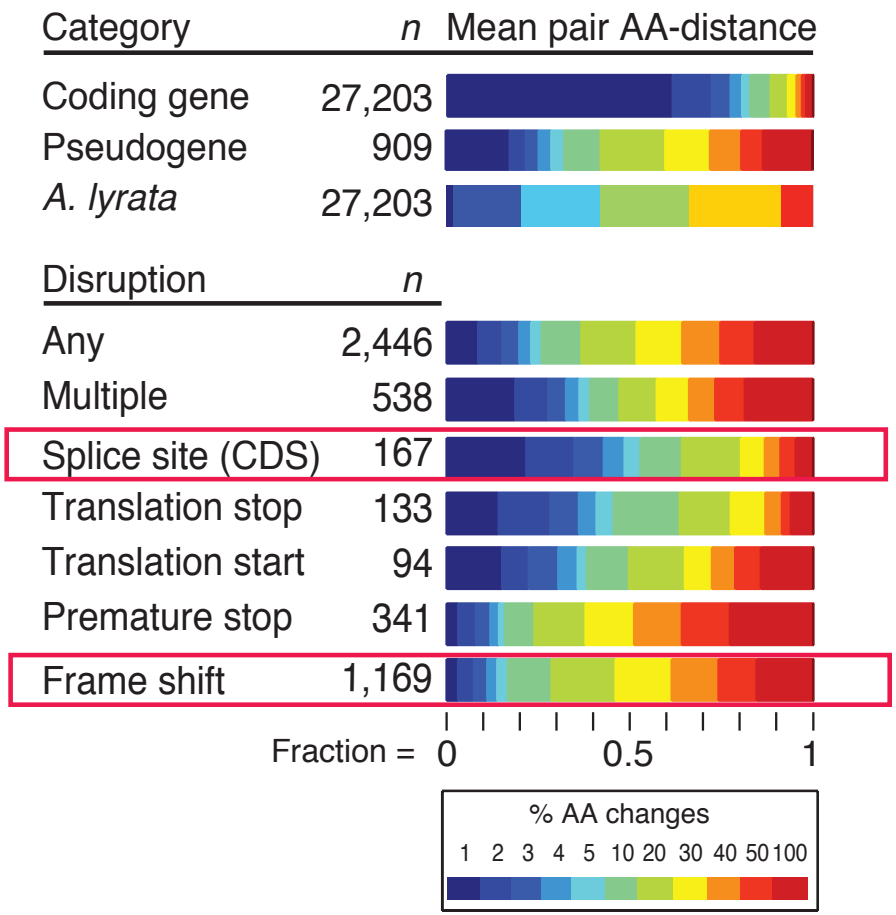
Amino acid variability

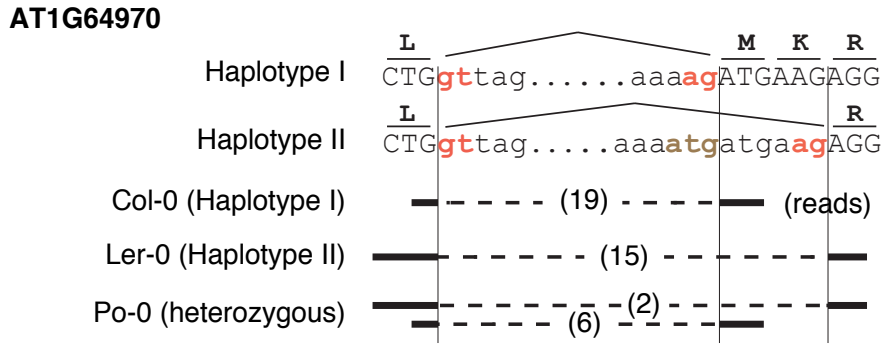
Large effects changes



Amino acid variability

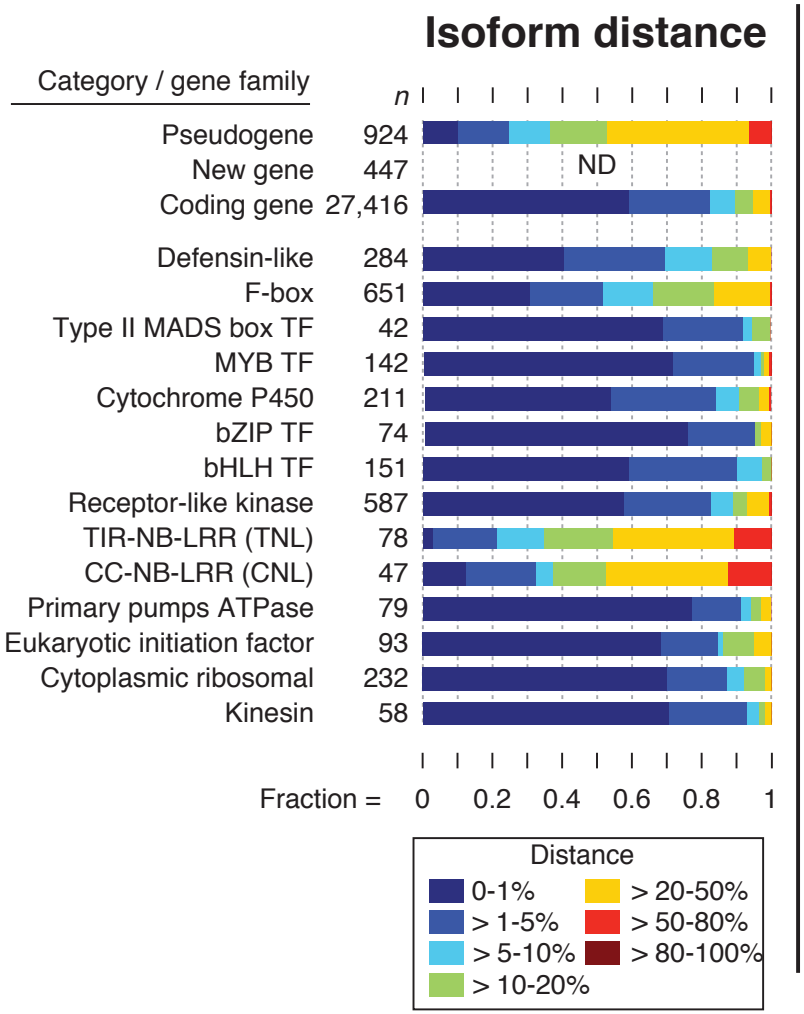
Large effects changes





Proteome Variation

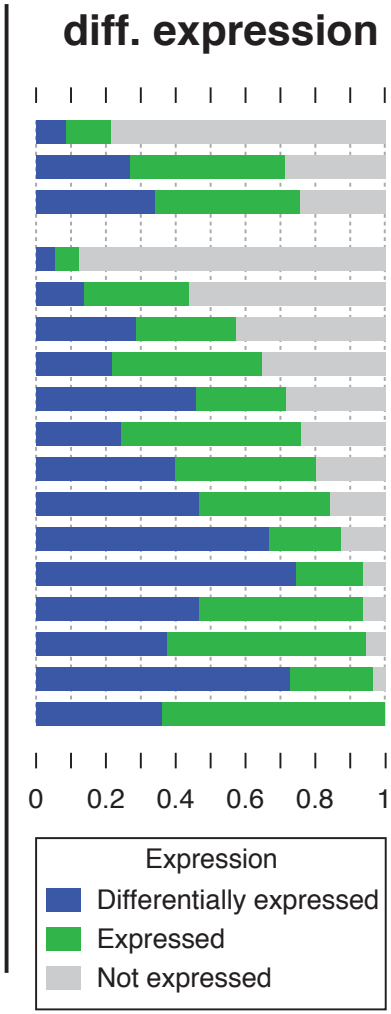
Variability by gene family



Gene expression variability

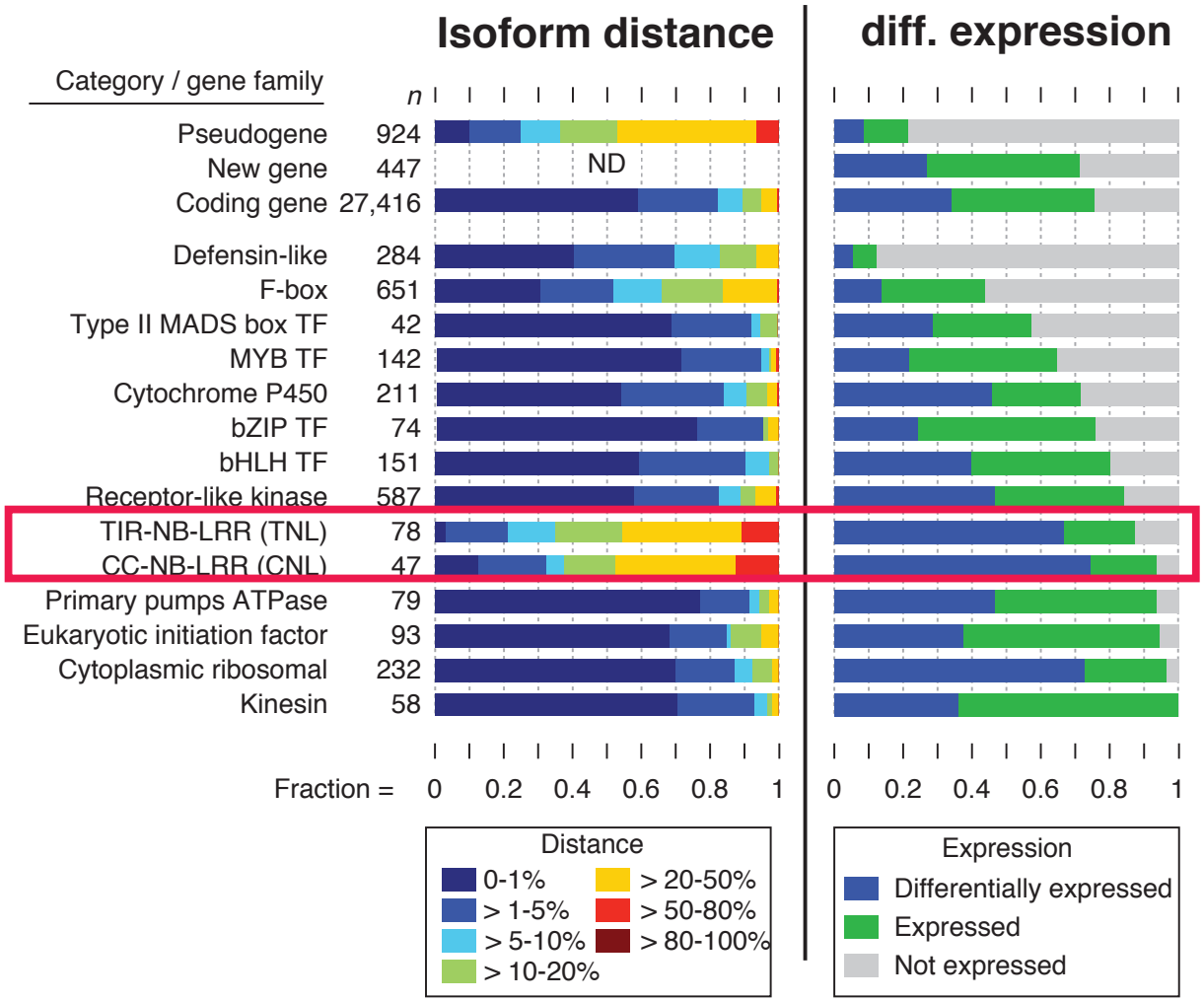
Category / gene family	*n*
Pseudogene	924
New gene	447
Coding gene	27,416
Defensin-like	284
F-box	651
Type II MADS box TF	42
MYB TF	142
Cytochrome P450	211
bZIP TF	74
bHLH TF	151
Receptor-like kinase	587
TIR-NB-LRR (TNL)	78
CC-NB-LRR (CNL)	47
Primary pumps ATPase	79
Eukaryotic initiation factor	93
Cytoplasmic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

Fraction =



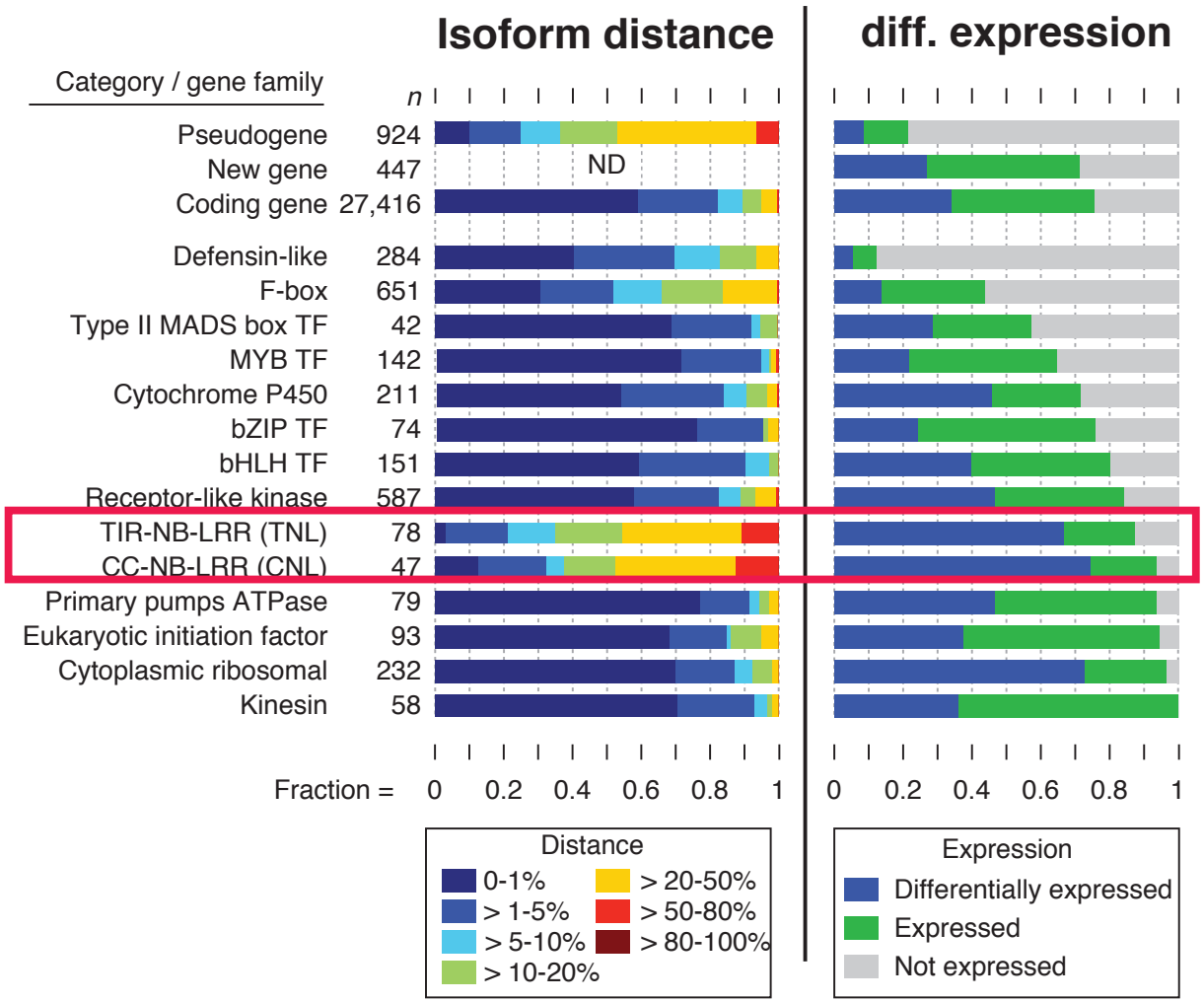
- ≈19,000 genes with detectable expression
  - 9,360 differentially expressed genes
  - 1,712 with fold-change ≥ 10
  - 308 with fold-change ≥ 50
- Causes of differential expression?
  - Structural: large indels, gene models
  - CNVs
  - *cis*/*trans*-effects

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?
  - Structural: large indels, gene models
  - CNVs
  - *cis*/*trans*-effects

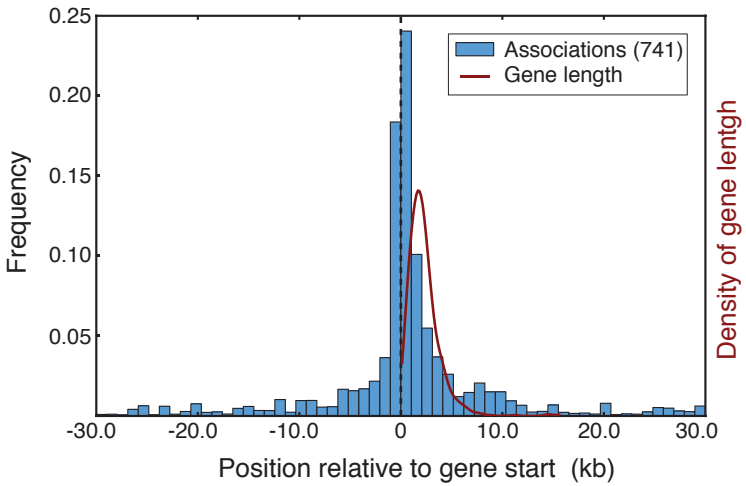
Gene expression variability



- ≈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?
  - Structural: large indels, gene models
  - CNVs
  - *cis*/*trans*-effects

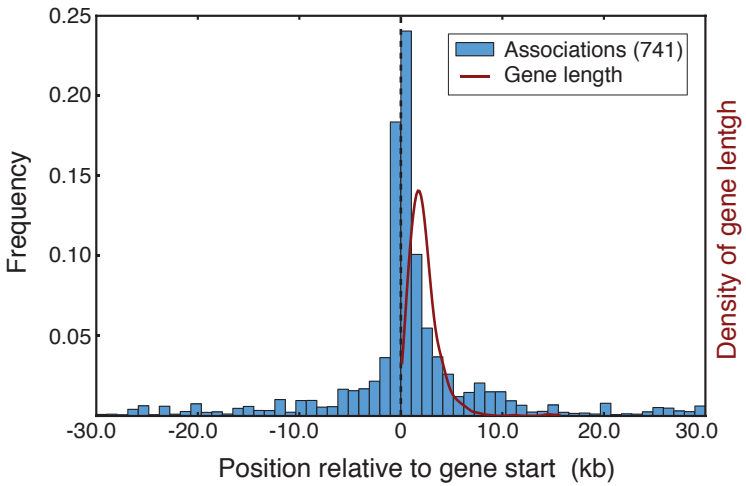
Gene expression variability

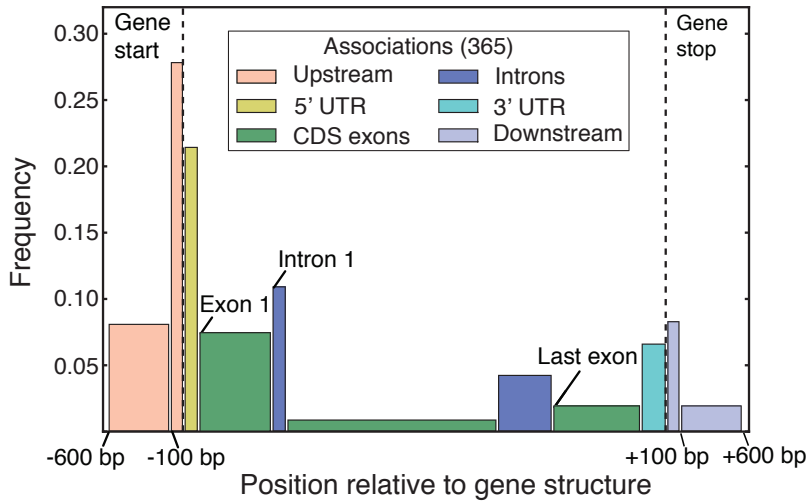
Mappable *cis* eQTLs



Gene expression variability

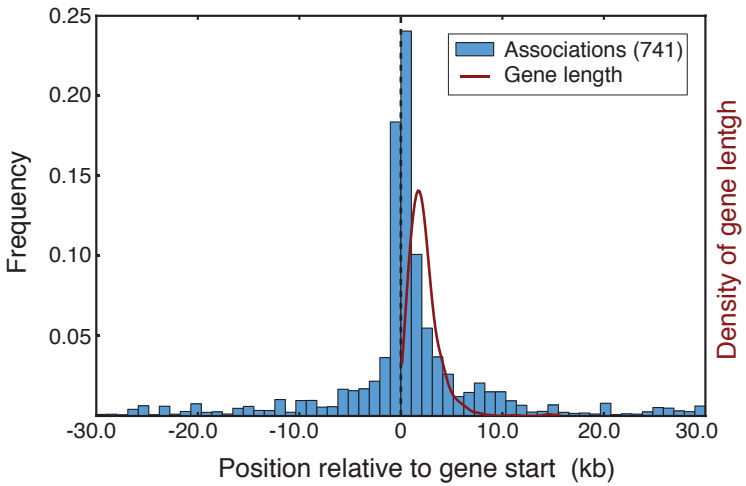
Mappable *cis* eQTLs

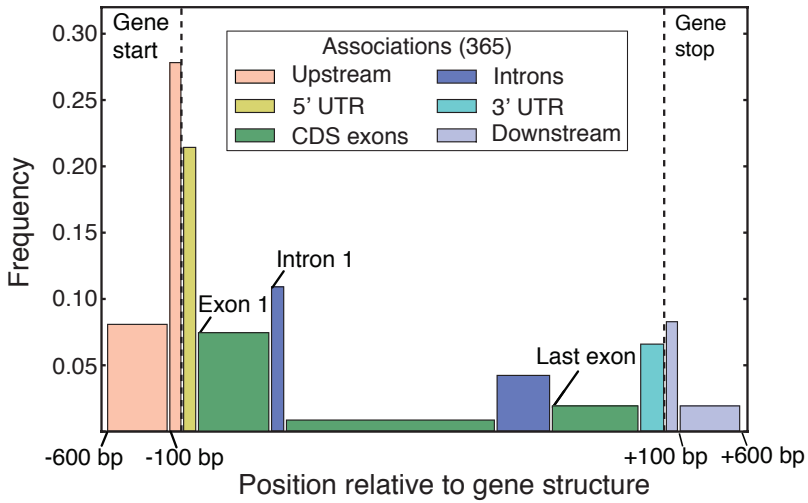


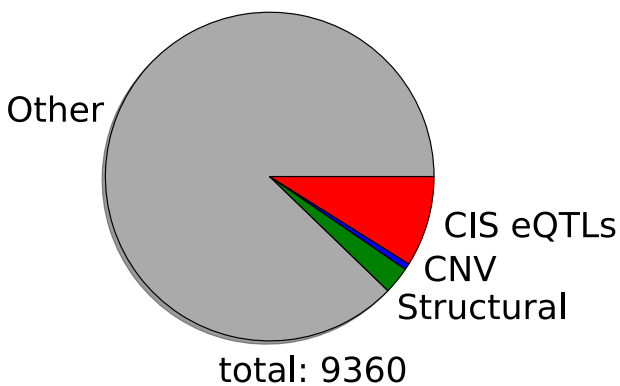


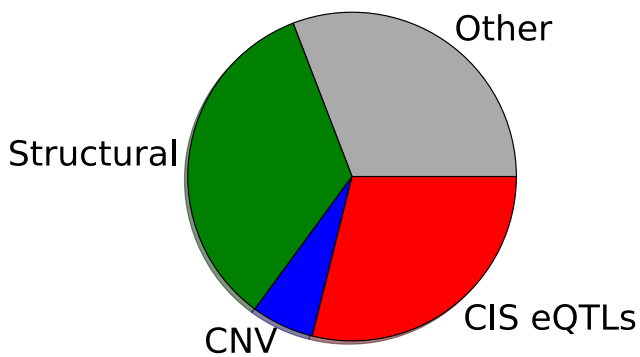
Gene expression variability

Mappable *cis* eQTLs





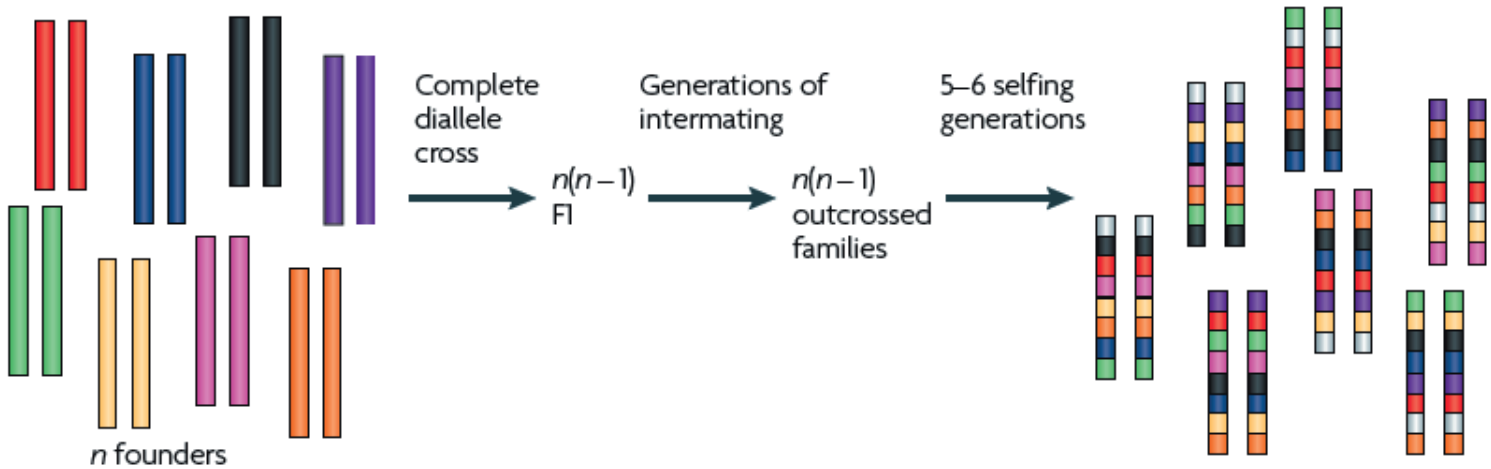




(FC > 50): 308

MAGIC Lines & Association Studies

Multiparent Advanced Generation Inter-Cross (MAGIC) lines



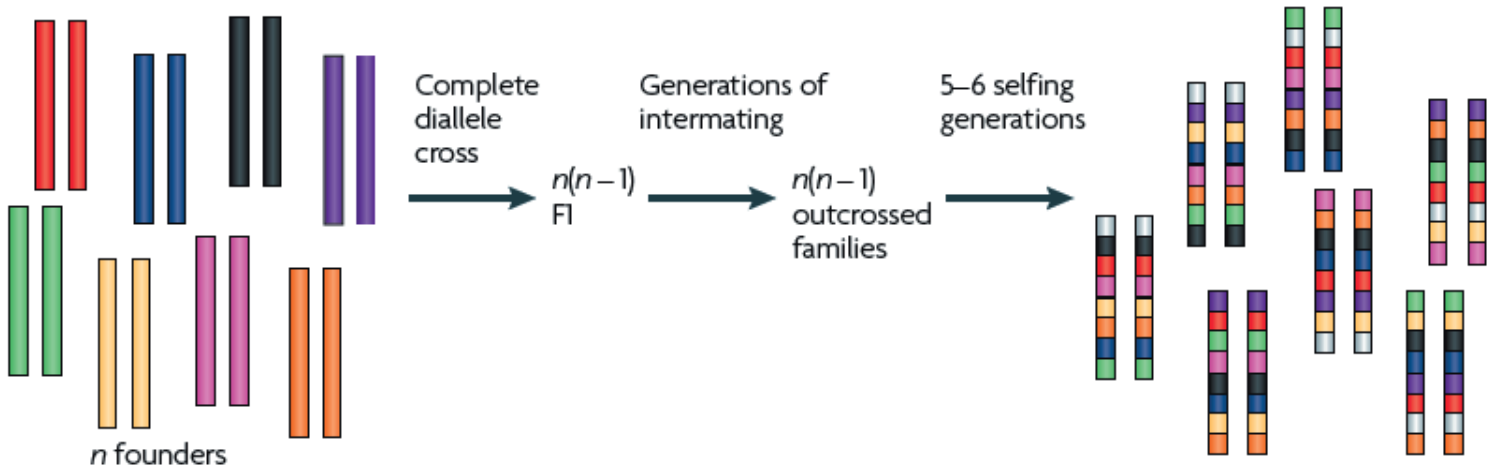
- Currently 500 available MAGIC lines [w/ Richard Mott]
- ≈200 natural strains with RNA-seq [w/ Magnus Nordborg]

Goals:

- Associate molecular phenotypes with genotypic variation
- Identify *trans*-factors:
- Extend models: f(DNA, , = RNA

MAGIC Lines & Association Studies

Multiparent Advanced Generation Inter-Cross (MAGIC) lines



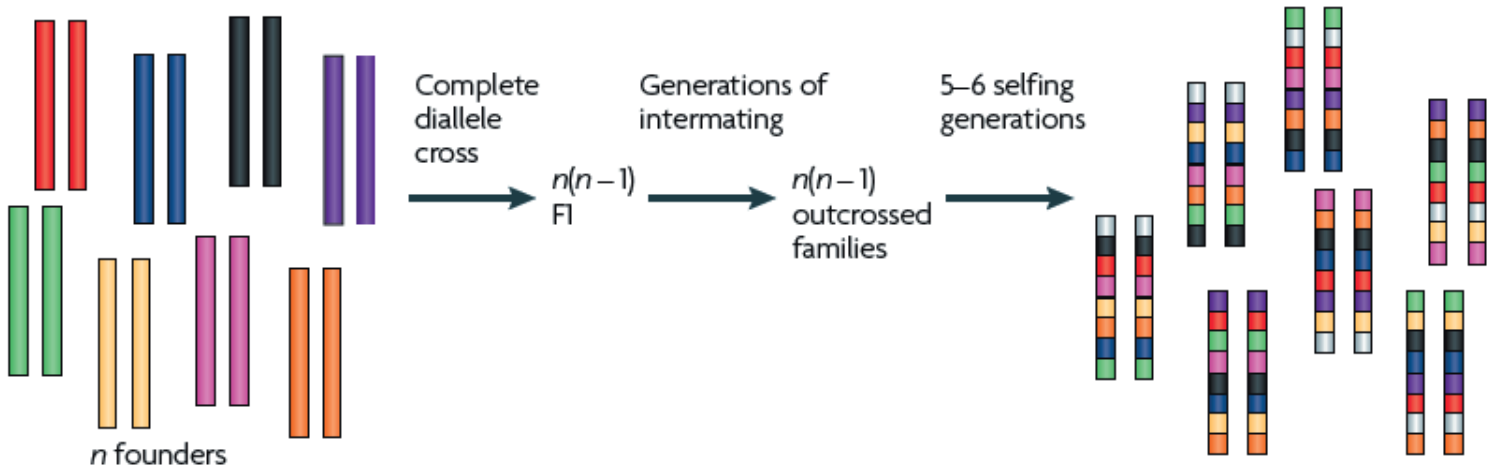
- Currently 500 available MAGIC lines [w/ Richard Mott]
- ≈200 natural strains with RNA-seq [w/ Magnus Nordborg]

Goals:

- Associate molecular phenotypes with genotypic variation
- Identify *trans*-factors:
- Extend models: f(DNA, , = RNA

MAGIC Lines & 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]

Goals:

- Associate molecular phenotypes with genotypic variation
- Identify *trans*-factors: 1_{trans}, 2_{trans}
- Extend models: $f(DNA, 1_{cis/trans}, 2_{cis/trans}) = RNA$

Summary and Outlook



Observe qualitative and quantitative transcriptome variation

- during development,
- under different external conditions
- with different genotypes

Extract cues by associating transcriptome variation with

- expression of factors
- genotypes/polymorphisms

Computational Modeling

- using yet incomplete knowledge
- candidates of *trans*-acting factors
- sequence information

Goal: Build a predictive system of the RNA processing system

Summary and Outlook

Observe qualitative and quantitative transcriptome variation

- during development,
- under different external conditions
- with different genotypes

Extract cues by associating transcriptome variation with

- expression of factors
- genotypes/polymorphisms

Computational Modeling

- using yet incomplete knowledge
- candidates of *trans*-acting factors
- sequence information

Goal: Build a predictive system of the RNA processing system

Summary and Outlook



Observe qualitative and quantitative transcriptome variation

- during development,
- under different external conditions
- with different genotypes

Extract cues by associating transcriptome variation with

- expression of factors
- genotypes/polymorphisms

Computational Modeling

- using yet incomplete knowledge
- candidates of *trans*-acting factors
- sequence information

Goal: Build a predictive system of the RNA processing system

Summary and Outlook



Observe qualitative and quantitative transcriptome variation

- during development,
- under different external conditions
- with different genotypes

Extract cues by associating transcriptome variation with

- expression of factors
- genotypes/polymorphisms

Computational Modeling

- using yet incomplete knowledge
- candidates of *trans*-acting factors
- sequence information

Goal: Build a predictive system of the RNA processing system

Acknowledgements

Method collaborators:

1. Bernhard Schölkopf, MPI Tü (Kernel & Transfer learning)
2. **Oliver Stegle, MPI Tü** (QTLs, statistical tests for RNA-seq)
3. Oliver Kohlbacher, Uni Tübingen (Learning for Immunoinformatics)

Biology collaborators:

1. Detlef Weigel, MPI Tü (Annotation of plant genomes, ...)
2. **David Miller, Vanderbilt U** (*C. elegans* transcriptome analyses)
3. **Richard Mott, U Oxford** (Genomes of *A. thaliana* strains)
4. **Richard Clark, U Utah** (Transcriptomes of *A. thaliana* strains)
5. Andreas Wachter, ZMBP Tü (Light-regulation in *A. thaliana*)
6. Rita Gross-Hardt, ZMBP Tü (Splicing in *A. thaliana)*
7. Andre Fischer, ENI, Göttingen (Transcriptomes in aging mice)
8. Elena Conti, MPI Munich (Analysis of Exosome mutants)
9. Dan Bumberger, MPI Tü (3D reconstruction of worms)

Thank you for attention!

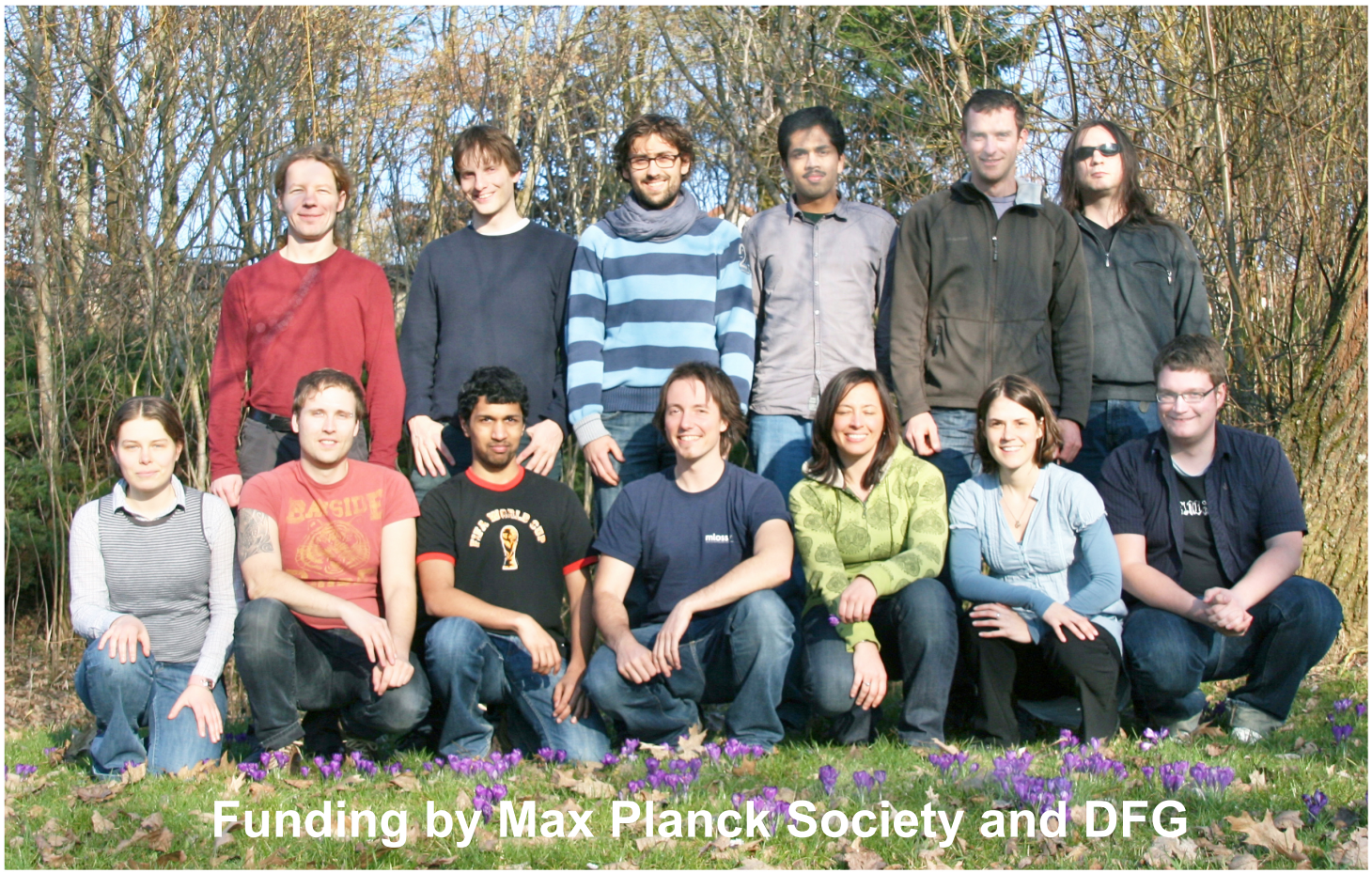
Acknowledgements

Method collaborators:

1. Bernhard Schölkopf, MPI Tü (Kernel & Transfer learning)
2. **O** RNA-seq)
3. Ol ormatics)

Biolog

1. De nes, . . .)
2. **D** analyses)
3. **Ri** strains)
4. **Ri** strains)
5. Ar *thaliana*)
6. Ri *thaliana)*
7. Ar ng mice)
8. El mutants)
9. Da f worms)



Thank you for attention!

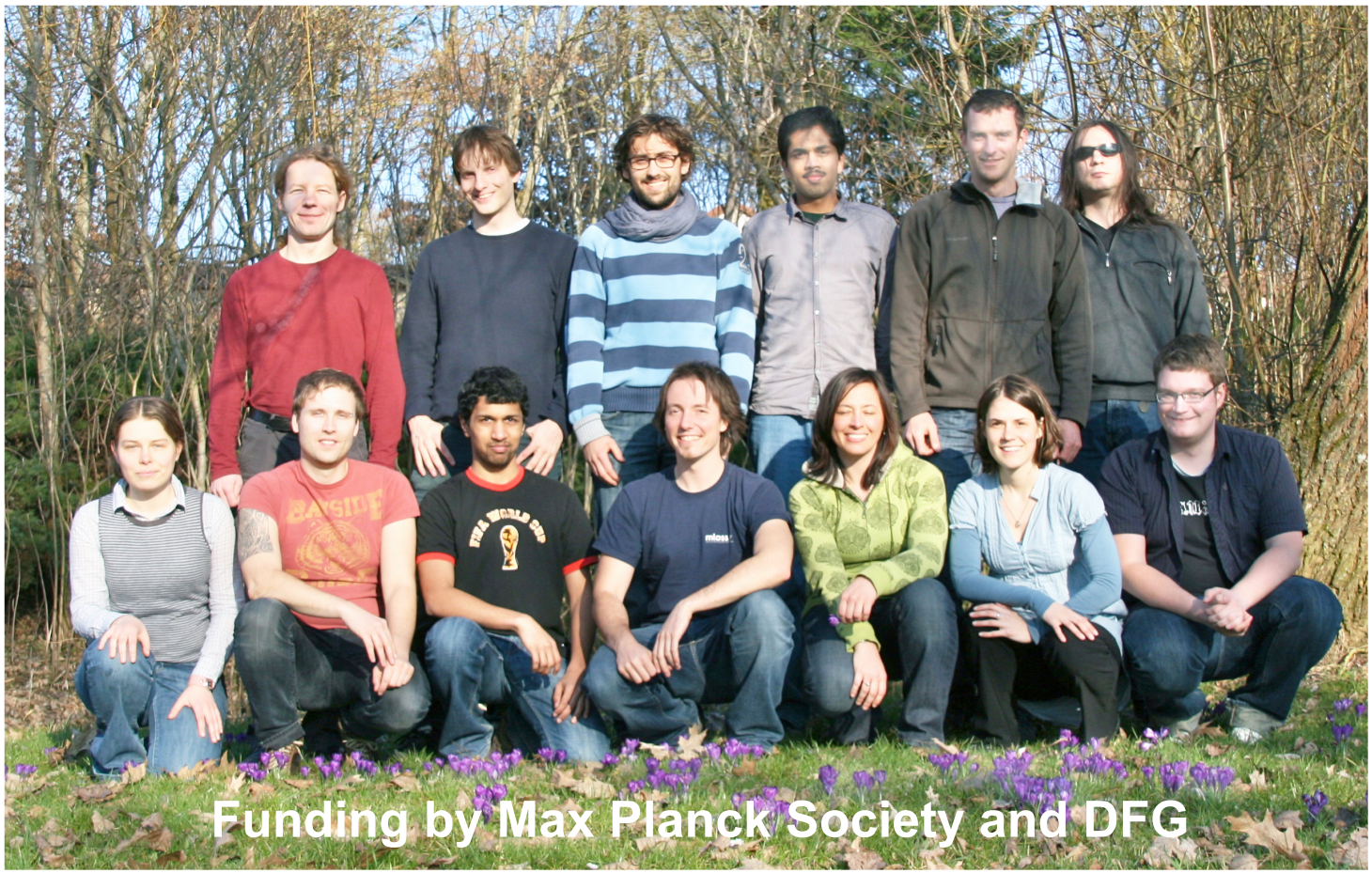
Acknowledgements

Method collaborators:

1. Bernhard Schölkopf, MPI Tü (Kernel & Transfer learning)
2. **O** RNA-seq)
3. Ol ormatics)

Biolog

1. De nes, . . .)
2. **D** analyses)
3. **Ri** strains)
4. **Ri** strains)
5. Ar *thaliana*)
6. Ri *thaliana)*
7. Ar ng mice)
8. El mutants)
9. Da f worms)



Thank you for attention!

References I

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`.

Richard M Clark, Gabriele Schweikert, Christopher Toomajian, Stephan Ossowski, Georg Zeller, Paul Shinn, Norman Warthmann, Tina T Hu, Glenn Fu, David A Hinds, Huaming Chen, Kelly A Frazer, Daniel H Huson, Bernhard Schölkopf, Magnus Nordborg, Gunnar Rätsch, Joseph R Ecker, and Detlef Weigel. Common sequence polymorphisms shaping genetic diversity in arabidopsis thaliana. *Science*, 317(5836):338–42, Jul 2007a. doi: 10.1126/science.1138632.

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, 2007b. ISSN 1095-9203 (Electronic). doi: 10.1126/science.1138632.

Johannes Eichner, Georg Zeller, Sascha Laubinger, and Gunnar Rätsch. Support vector machines-based identification of alternative splicing in arabidopsis thaliana from whole-genome tiling arrays. *BMC Bioinformatics*, 12:55, 2011. doi: 10.1186/1471-2105-12-55.

References II

Mark B Gerstein, Zhi John Lu, Eric L Van Nostrand, Chao Cheng, Bradley I Arshinoff, Tao Liu, Kevin Y Yip, Rebecca Robilotto, Andreas Rechtsteiner, Kohta Ikegami, Pedro Alves, Aurelien Chateigner, Marc Perry, Mitzi Morris, Raymond K Auerbach, Xin Feng, Jing Leng, Anne Vielle, Wei Niu, Kahn Rhrissorrakrai, Ashish Agarwal, Roger P Alexander, Galt Barber, Cathleen M Brdlik, Jennifer Brennan, Jeremy Jean Brouillet, Adrian Carr, Ming-Sin Cheung, Hiram Clawson, Sergio Contrino, Luke O Dannenberg, Abby F Dernburg, Arshad Desai, Lindsay Dick, Andréa C Dosé, Jiang Du, Thea Egelhofer, Sevinc Ercan, Ghia Euskirchen, Brent Ewing, Elise A Feingold, Reto Gassmann, Peter J Good, Phil Green, Francois Gullier, Michelle Gutwein, Mark S Guyer, Lukas Habegger, Ting Han, Jorja G Henikoff, Stefan R Henz, Angie Hinrichs, Heather Holster, Tony Hyman, A Leo Iniguez, Judith Janette, Morten Jensen, Masaomi Kato, W James Kent, Ellen Kephart, Vishal Khivansara, Ekta Khurana, John K Kim, Paulina Kolasinska-Zwierz, Eric C Lai, Isabel Latorre, Amber Leahey, Suzanna Lewis, Paul Lloyd, Lucas Lochovsky, Rebecca F Lowdon, Yaniv Lubling, Rachel Lyne, Michael MacCoss, Sebastian D Mackowiak, Marco Mangone, Sheldon McKay, Desirea Mecenas, Gennifer Merrihew, David M Miller, 3rd, Andrew Muroyama, John I Murray, Siew-Loon Ooi, Hoang Pham, Taryn Phippen, Elicia A Preston, Nikolaus Rajewsky, Gunnar Rätsch, Heidi Rosenbaum, Joel Rozowsky, Kim Rutherford, Peter Ruzanov, Mihail Sarov, Rajkumar Sasidharan, Andrea Sboner, Paul Scheid, Eran Segal, Hyunjin Shin, Chong Shou, Frank J Slack, Cindie Slightam, Richard Smith, William C Spencer, E O Stinson, Scott Taing, Teruaki Takasaki, Dionne Vafeados, Ksenia Voronina, Guilin Wang, Nicole L Washington, Christina M Whittle, Beijing Wu, Koon-Kiu Yan, Georg Zeller, Zheng Zha, Mei Zhong, Xingliang Zhou, modENCODE Consortium, Julie Ahringer, Susan Strome, Kristin C Gunsalus, Gos Micklem, X Shirley Liu, Valerie Reinke, Stuart K Kim, LaDeana W Hillier,

References III

Steven Henikoff, Fabio Piano, Michael Snyder, Lincoln Stein, Jason D Lieb, and Robert H Waterston. Integrative analysis of the caenorhabditis elegans genome by the modencode project. *Science*, 330(6012):1775–87, Dec 2010. doi: 10.1126/science.1196914.

Elissa A Hallem, W Clay Spencer, Rebecca D McWhirter, Georg Zeller, Stefan R Henz, Gunnar Rätsch, David M Miller, 3rd, H Robert Horvitz, Paul W Sternberg, and Niels Ringstad. Receptor-type guanylate cyclase is required for carbon dioxide sensation by caenorhabditis elegans. *Proc Natl Acad Sci U S A*, 108(1):254–9, Jan 2011. doi: 10.1073/pnas.1017354108.

Sascha Laubinger, Timo Sachsenberg, Georg Zeller, Wolfgang Busch, Jan U Lohmann, Gunnar Rätsch, and Detlef Weigel. Dual roles of the nuclear cap-binding complex and serrate in pre-mrna splicing and microrna processing in arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 105(25):8795–800, Jun 2008. doi: 10.1073/pnas.0802493105.

Sascha Laubinger, Georg Zeller, Stefan R Henz, Sabine Buechel, Timo Sachsenberg, Jia-Wei Wang, Gunnar Rätsch, and Detlef Weigel. Global effects of the small rna biogenesis machinery on the arabidopsis thaliana transcriptome. *Proc Natl Acad Sci U S A*, 107(41): 17466–73, Oct 2010. doi: 10.1073/pnas.1012891107.

References IV

Kenneth L McNally, Kevin L Childs, Regina Bohnert, Rebecca M Davidson, Keyan Zhao, Victor J Ulat, Georg Zeller, Richard M Clark, Douglas R Hoen, Thomas E Bureau, Renee Stokowski, Dennis G Ballinger, Kelly A Frazer, David R Cox, Badri Padhukasahasram, Carlos D Bustamante, Detlef Weigel, David J Mackill, Richard M Bruskiewich, Gunnar Rätsch, C Robin Buell, Hei Leung, and Jan E Leach. Genomewide snp variation reveals relationships among landraces and modern varieties of rice. *Proc Natl Acad Sci U S A*, 106 (30):12273–8, Jul 2009. doi: 10.1073/pnas.0900992106.

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.

Sebastian J Schultheiss, Wolfgang Busch, Jan U Lohmann, Oliver Kohlbacher, and Gunnar Rätsch. Kirmes: kernel-based identification of regulatory modules in euchromatic sequences. *Bioinformatics*, 25(16):2126–33, Aug 2009. doi: 10.1093/bioinformatics/btp278.

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.

References V

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.

W Clay Spencer, Georg Zeller, Joseph D Watson, Stefan R Henz, Kathie L Watkins, Rebecca D McWhirter, Sarah Petersen, Vipin T Sreedharan, Christian Widmer, Jeanyoung Jo, Valerie Reinke, Lisa Petrella, Susan Strome, Stephen E Von Stetina, Menachem Katz, Shai Shaham, Gunnar Rätsch, and David M Miller, 3rd. A spatial and temporal map of c. elegans gene expression. *Genome Res*, 21(2):325–41, Feb 2011. doi: 10.1101/gr.114595.110.

Oliver Stegle, Leopold Parts, Richard Durbin, and John Winn. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS Comput Biol*, 6(5):e1000770, May 2010. doi: 10.1371/journal.pcbi.1000770.

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, 2008a. ISSN 1088-9051 (Print). doi: 10.1101/gr.070169.107.

G Zeller, S Henz, C Widmer, T Sachsenberg, G Rätsch, D Weigel, and S Laubinger. Stress-induced changes in the arabidopsis thaliana transcriptome analyzed using whole genome tiling arrays. *Plant J*, Feb 2009. doi: 10.1111/j.1365-313X.2009.03835.x.

Georg Zeller, Richard M Clark, Korbinian Schneeberger, Anja Bohlen, Detlef Weigel, and Gunnar Rätsch. Detecting polymorphic regions in arabidopsis thaliana with resequencing microarrays. *Genome Res*, 18(6):918–29, Jun 2008b. doi: 10.1101/gr.070169.107.

References VI

Georg Zeller, Stefan R Henz, Sascha Laubinger, Detlef Weigel, and Gunnar Rätsch. Transcript normalization and segmentation of tiling array data. *Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing*, pages 527–38, Jan 2008c.

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.