Towards Empirical Models of the Central Dogma of Molecular Biology

Gunnar Rätsch

Friedrich Miescher Laboratory

Max Planck Society, Tübingen

Institute of Science and Technology, Vienna
July 25, 2011



Friedrich Miescher Laboratory
of the Max Planck Society



Discovery of the Nuclein
(Friedrich Miescher, 1869)





[Dahm, 2005]





Discovery of Nuclein:

- from lymphocyte & salmon
- "multi-basic acid" (≥ 4)

"If one . . . wants to assume that a single substance . . . is the specific cause of fertilization, then one should undoubtedly first and foremost consider nuclein."

[Miescher, 1874]

Discovery of the Nuclein
(Friedrich Miescher, 1869)





[Dahm, 2005]





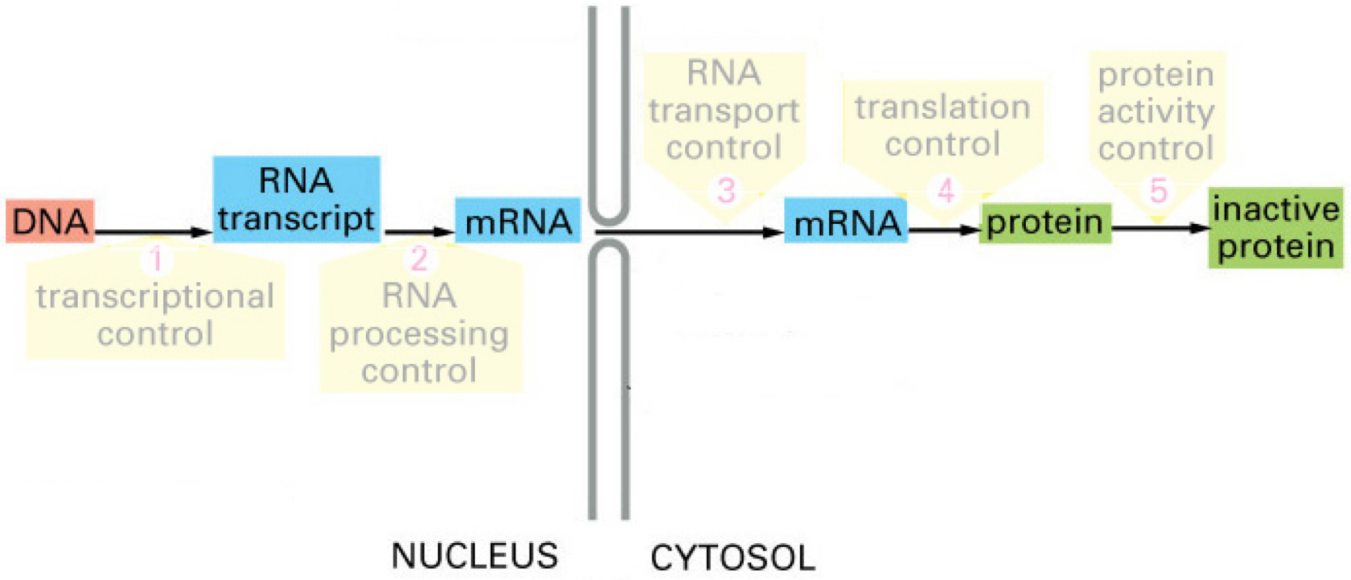
Discovery of Nuclein:

- from lymphocyte & salmon
- "multi-basic acid" (≥ 4)

"If one . . . wants to assume that a single substance . . . is the specific cause of fertilization, then one should undoubtedly first and foremost consider nuclein."

[Miescher, 1874]

Learning About the Central Dogma



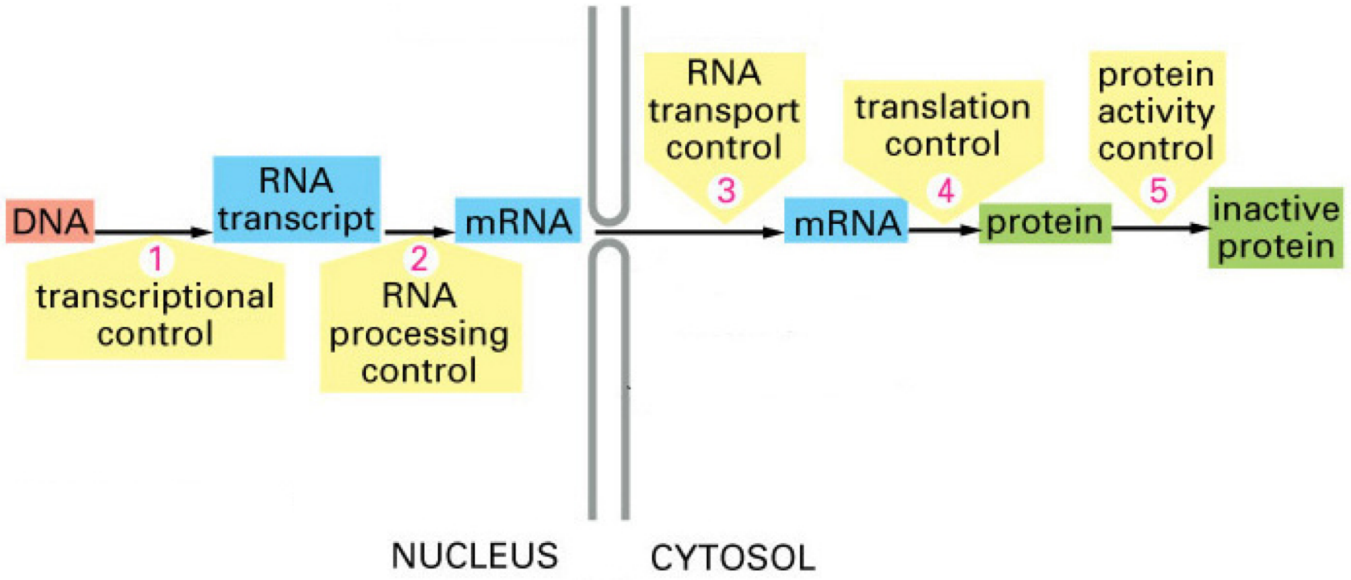
Goal: Learn to predict what these processes accomplish:

- 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



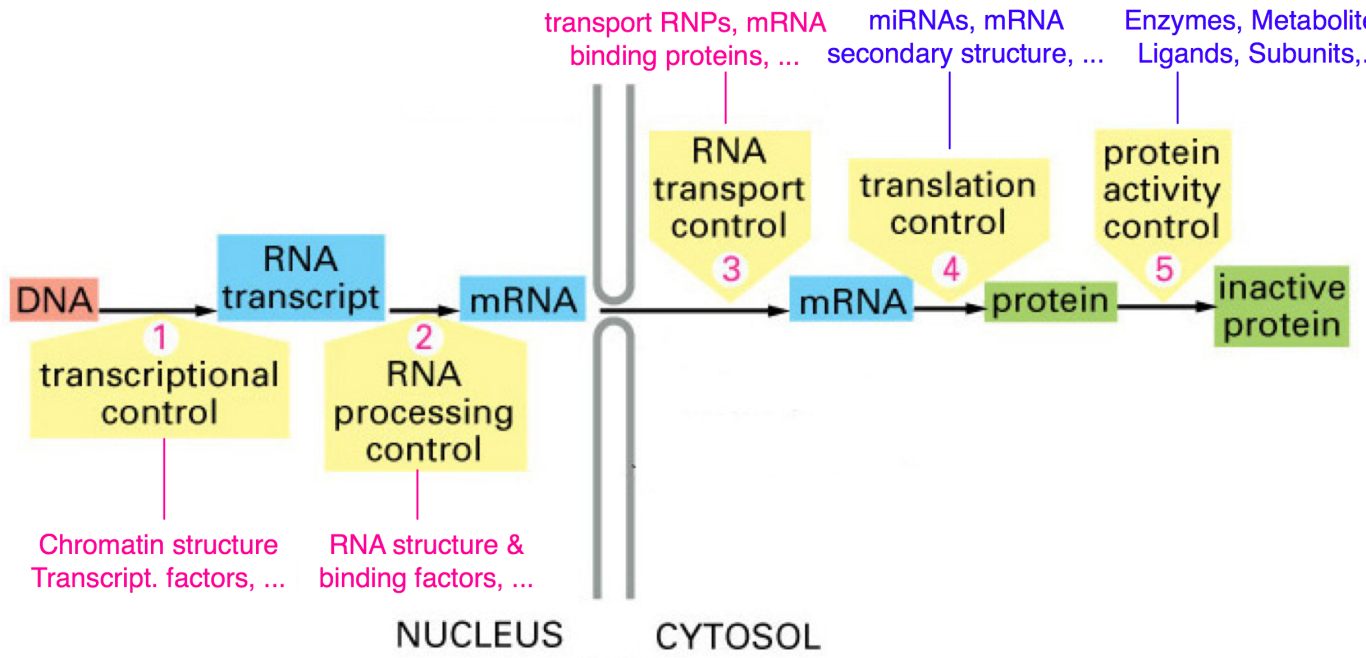
Goal: Learn to predict what these processes accomplish:

- 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



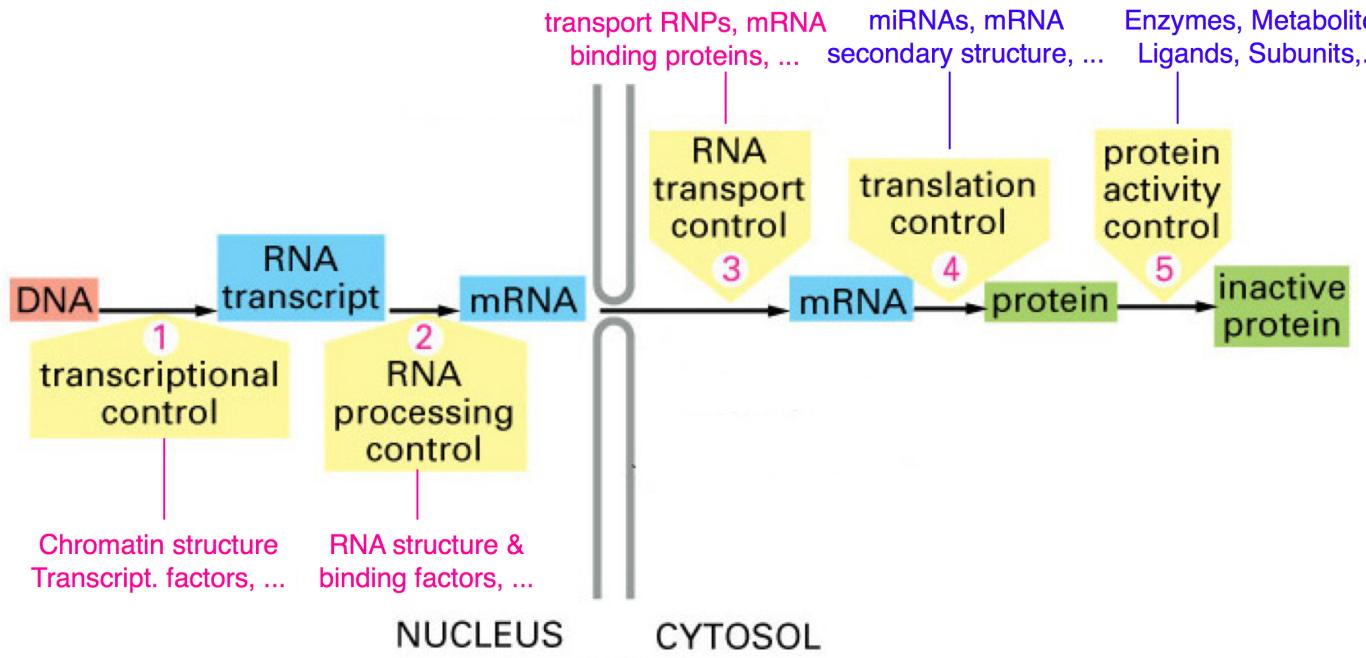
Goal: Learn to predict what these processes accomplish:

- 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



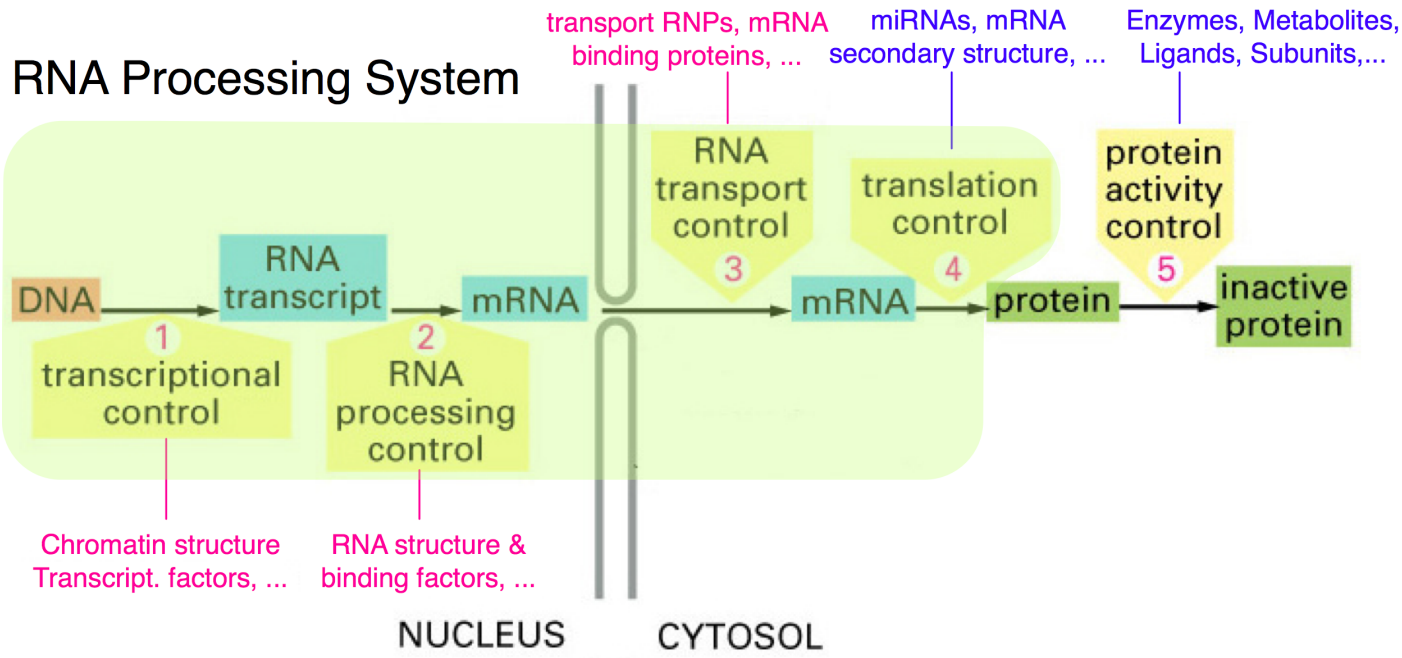
Goal: Learn to predict what these processes accomplish:

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

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

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

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]

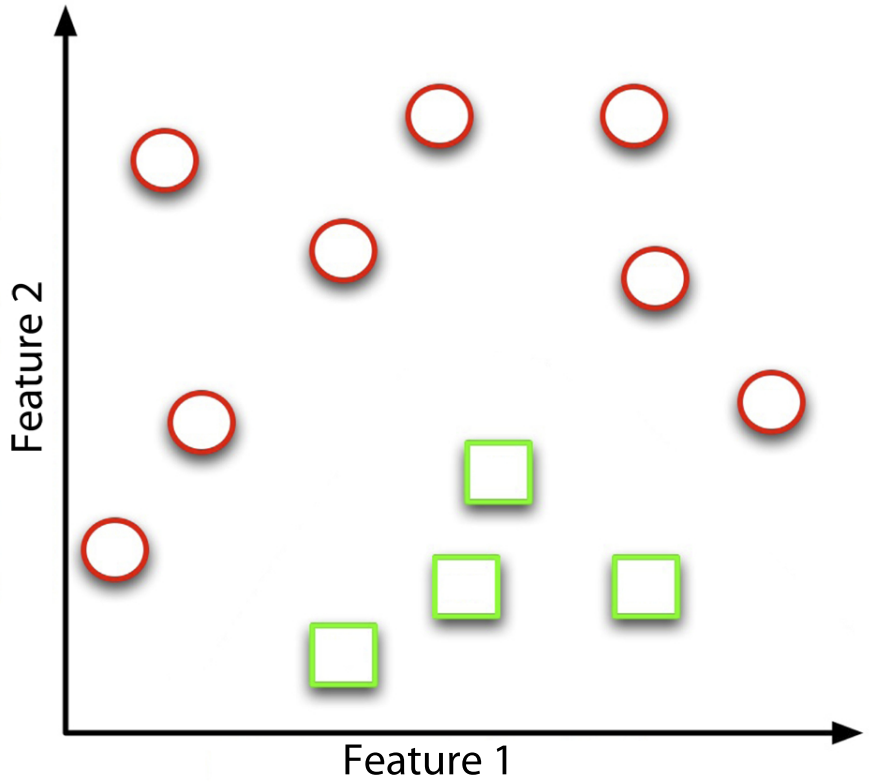
Machine Learning
Learning from empirical observations

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:

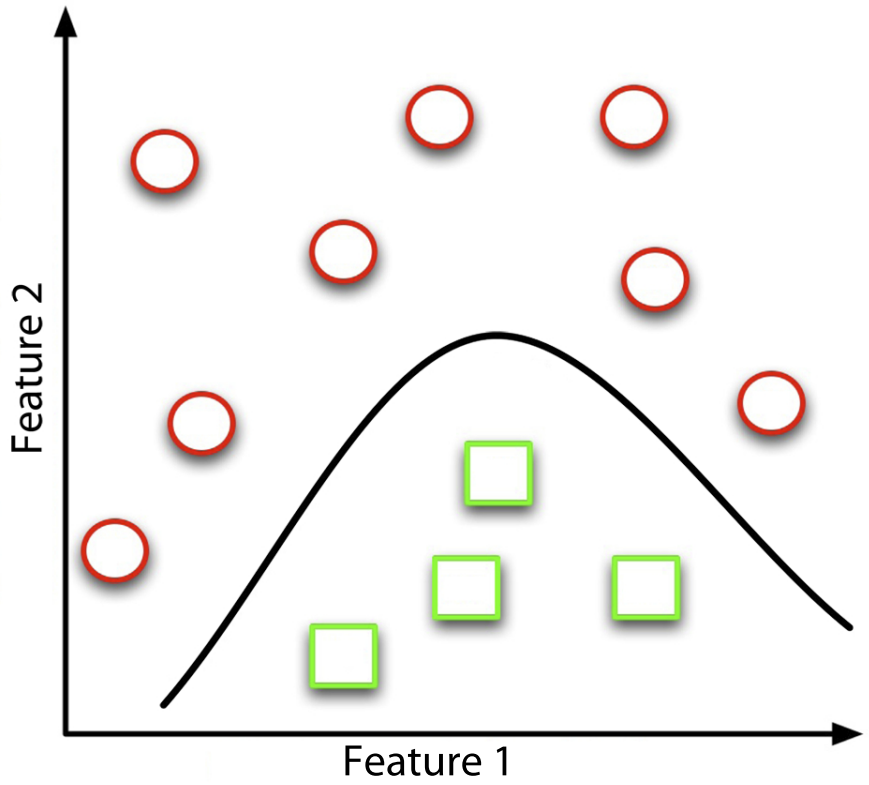


Two different classes of observations

Machine Learning
Learning from empirical observations

Given: Observations of some complex phenomenon
Goal: Learn from data & build predictive models

Example:

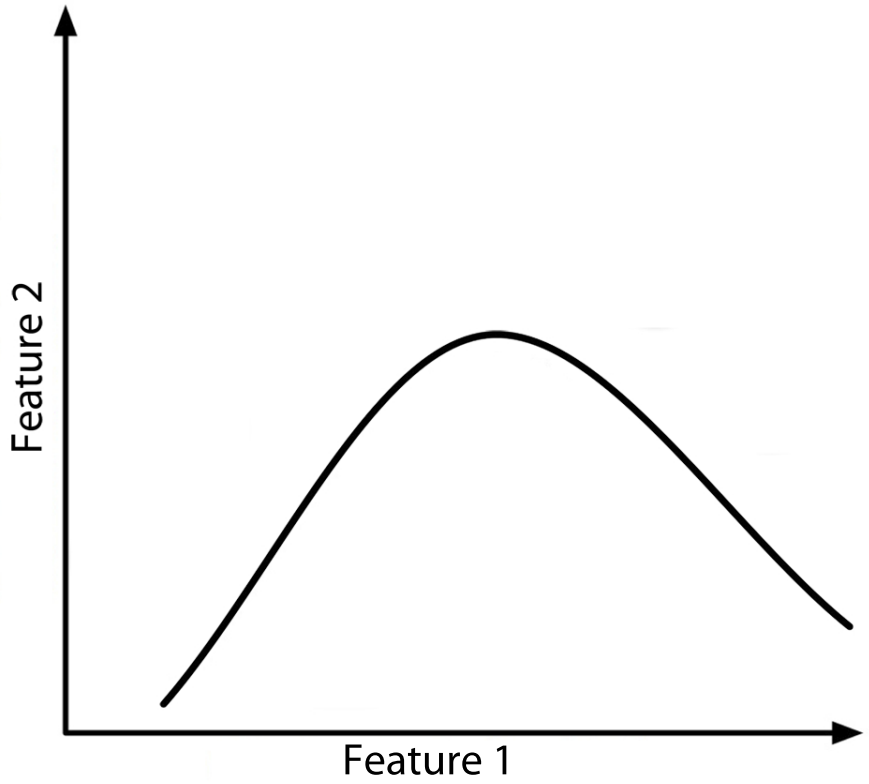


Infer separation between classes

Machine Learning
Learning from empirical observations

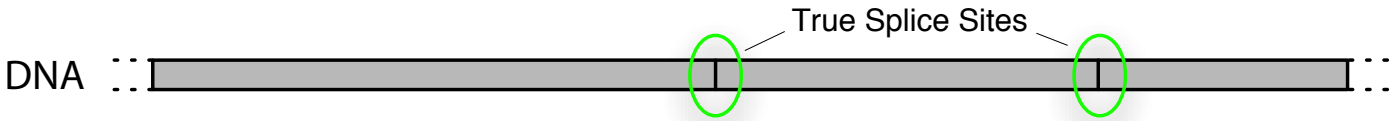
Given: Observations of some complex phenomenon
Goal: Learn from data & build predictive models

Example:



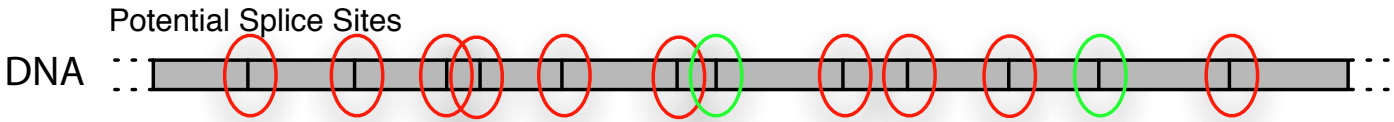
Rule for classifying new examples

Example: Splice Site Recognition



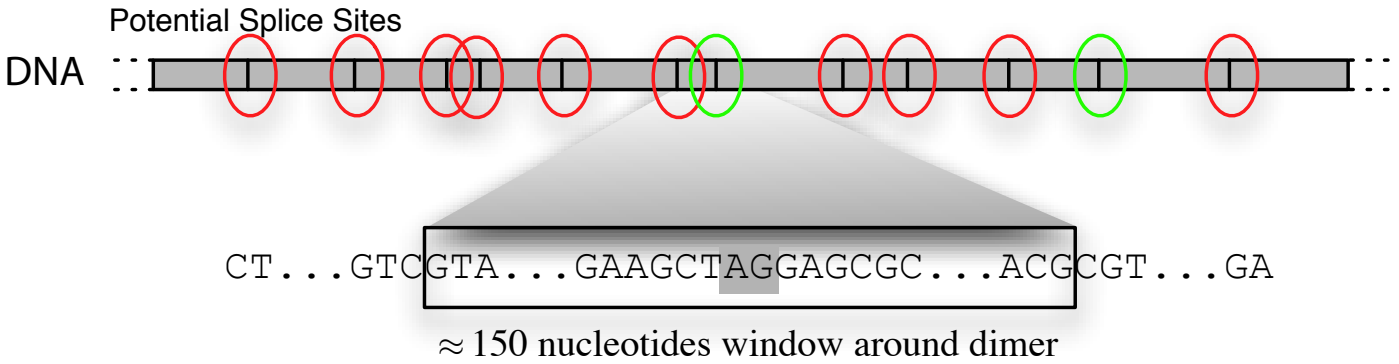
Basic idea:

Example: Splice Site Recognition



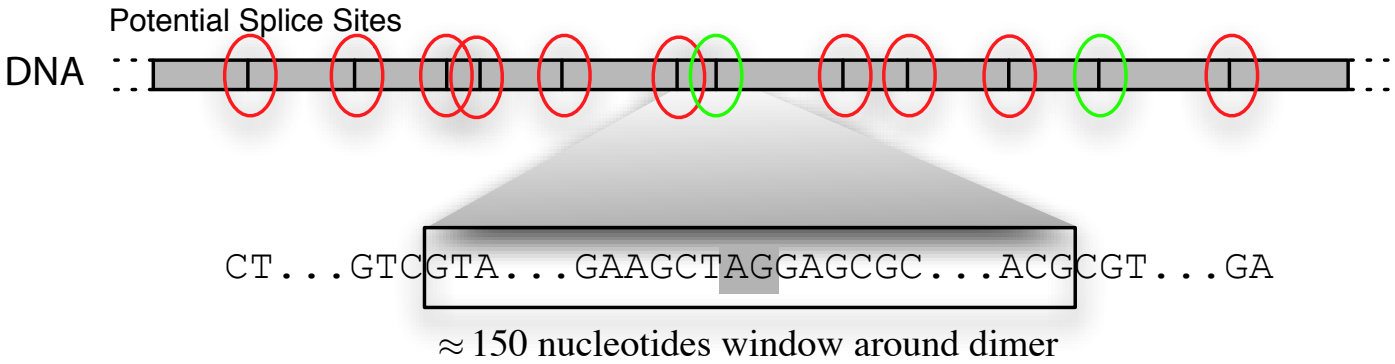
Basic idea:

Example: Splice Site Recognition

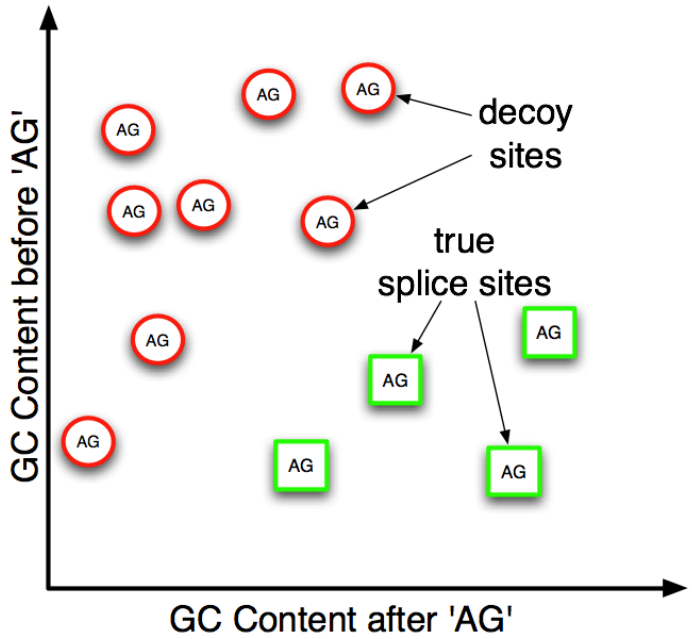


Basic idea:

Example: Splice Site Recognition



Basic idea:



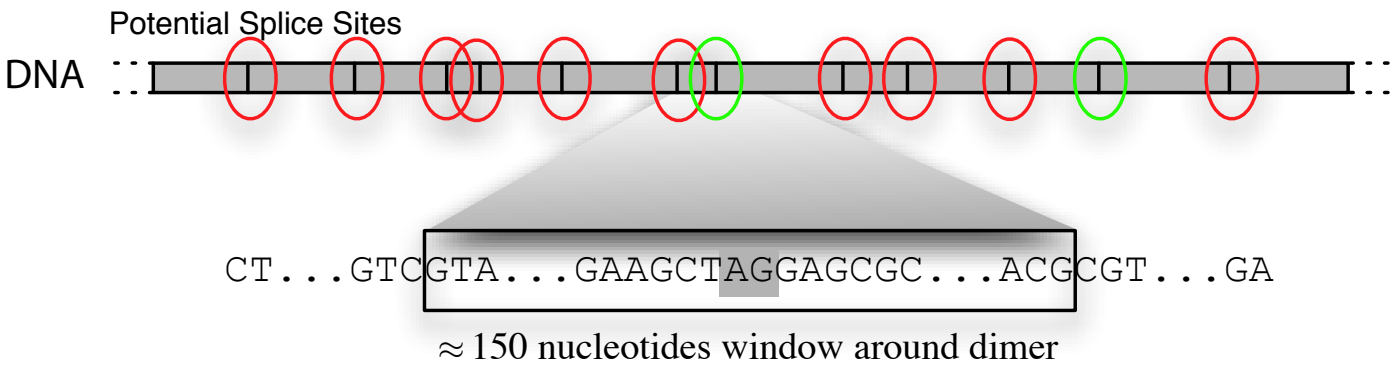
For instance, exploit that exons have higher GC content

or

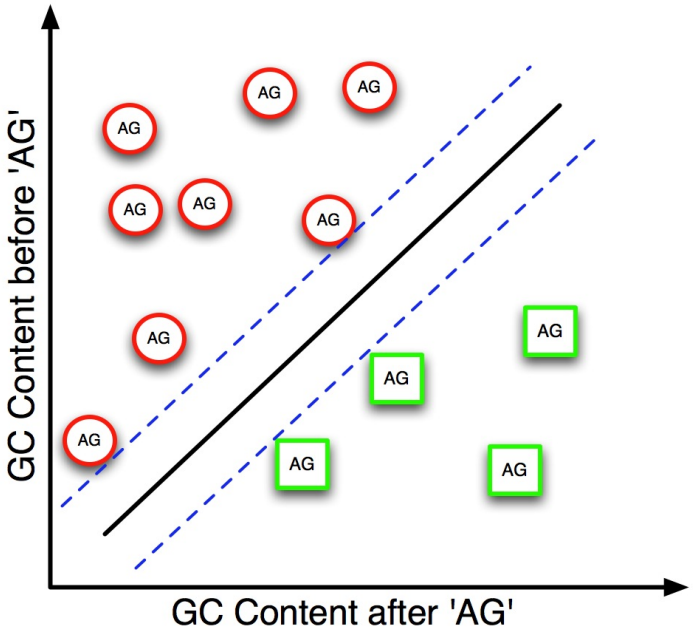
that specific motifs appear near splice sites.

[Sonnenburg et al., 2007]

Example: Splice Site Recognition



Basic idea:



In practice: Use one feature per possible substring (e.g. $\leq$20) at all positions

$150 \cdot (4^1 + \ldots + 4^{20}) \approx 2 \cdot 10^{14}$ features

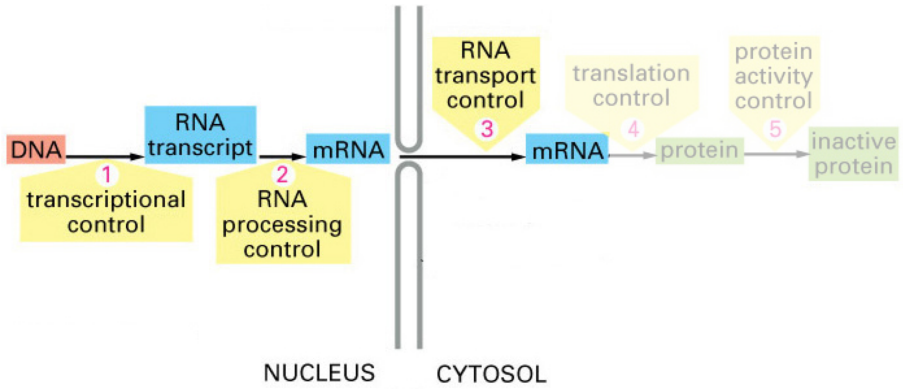
World records in splice site and transcription start site recognition.

www.shogun-toolbox.org

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

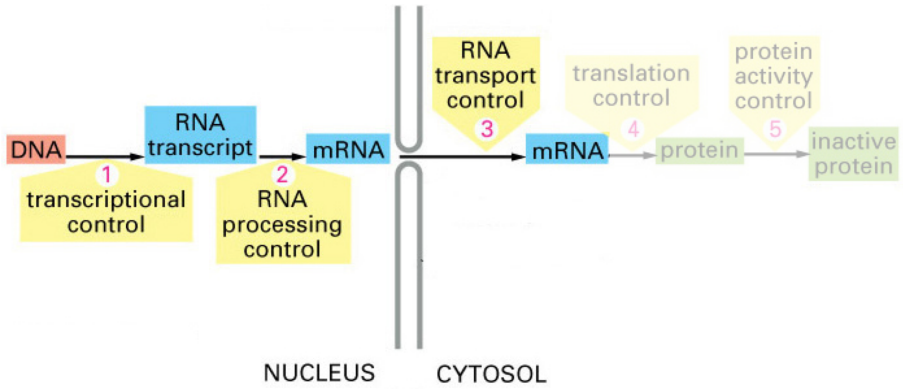
Solution: Deep RNA sequencing (RNA-seq)

$$f(\text{DNA, RNA-seq}) = \text{RNA}$$

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)

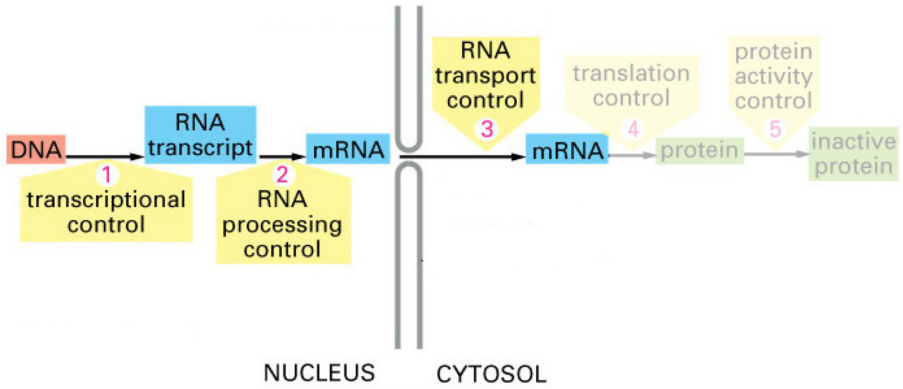
Intermediate Goal: Fully reconstruct RNAs from experimental data
Solution: Deep RNA sequencing (RNA-seq)

$$f(\text{DNA}, \text{RNA-seq}) = \text{RNA}$$

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

Solution: Deep RNA sequencing (RNA-seq)

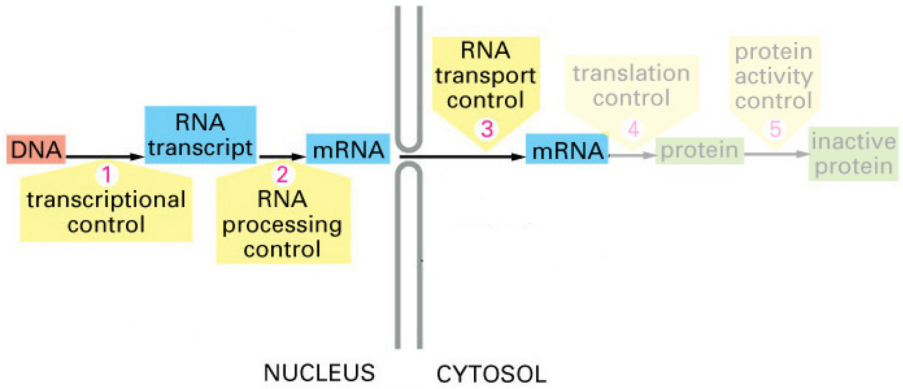
$$f(\text{DNA}, \text{RNA-seq}) = \text{RNA}$$

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

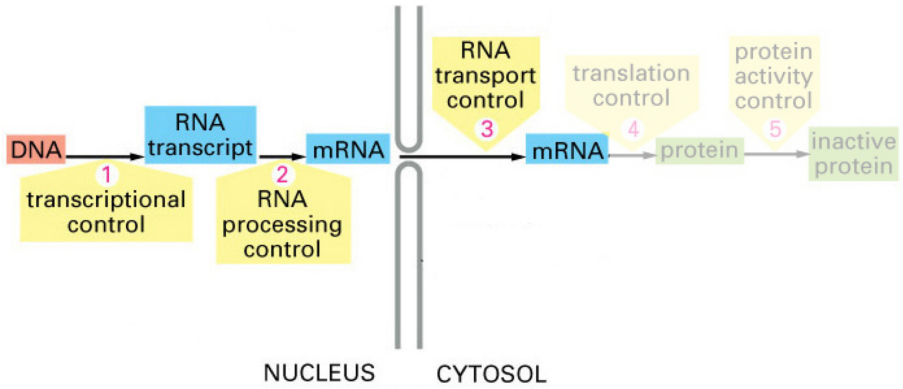
Solution: Deep RNA sequencing (RNA-seq)

$f(\text{DNA}, \text{RNA-seq}) = \text{RNA}$

Learning About the Genome

Goal: Estimate f to predict RNAs

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



Need:

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

Intermediate Goal: Fully reconstruct RNAs from experimental data

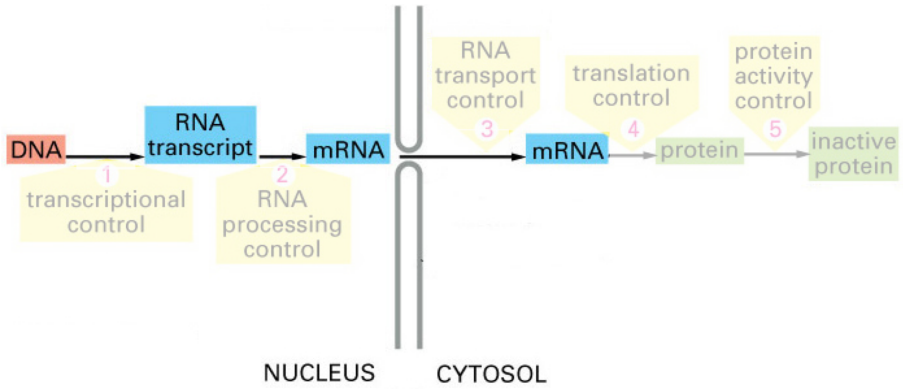
Solution: Deep RNA sequencing (RNA-seq)

$$f(\text{DNA, RNA-seq}) = \text{RNA}$$

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, $\cancel{1\ 2\ 3}$)
- Outputs (complete transcriptome)

Intermediate Goal: Fully reconstruct RNAs from experimental data

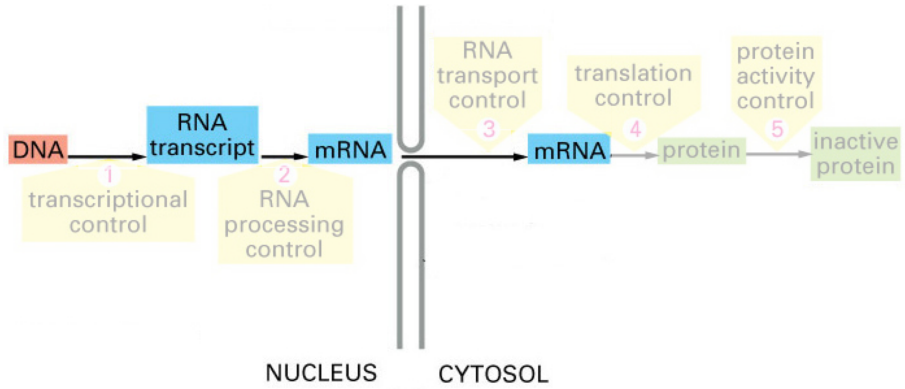
Solution: Deep RNA sequencing (RNA-seq)

$$f(\text{DNA, RNA-seq}) = \text{RNA}$$

Learning About the Genome

Goal: Estimate f to predict RNAs

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



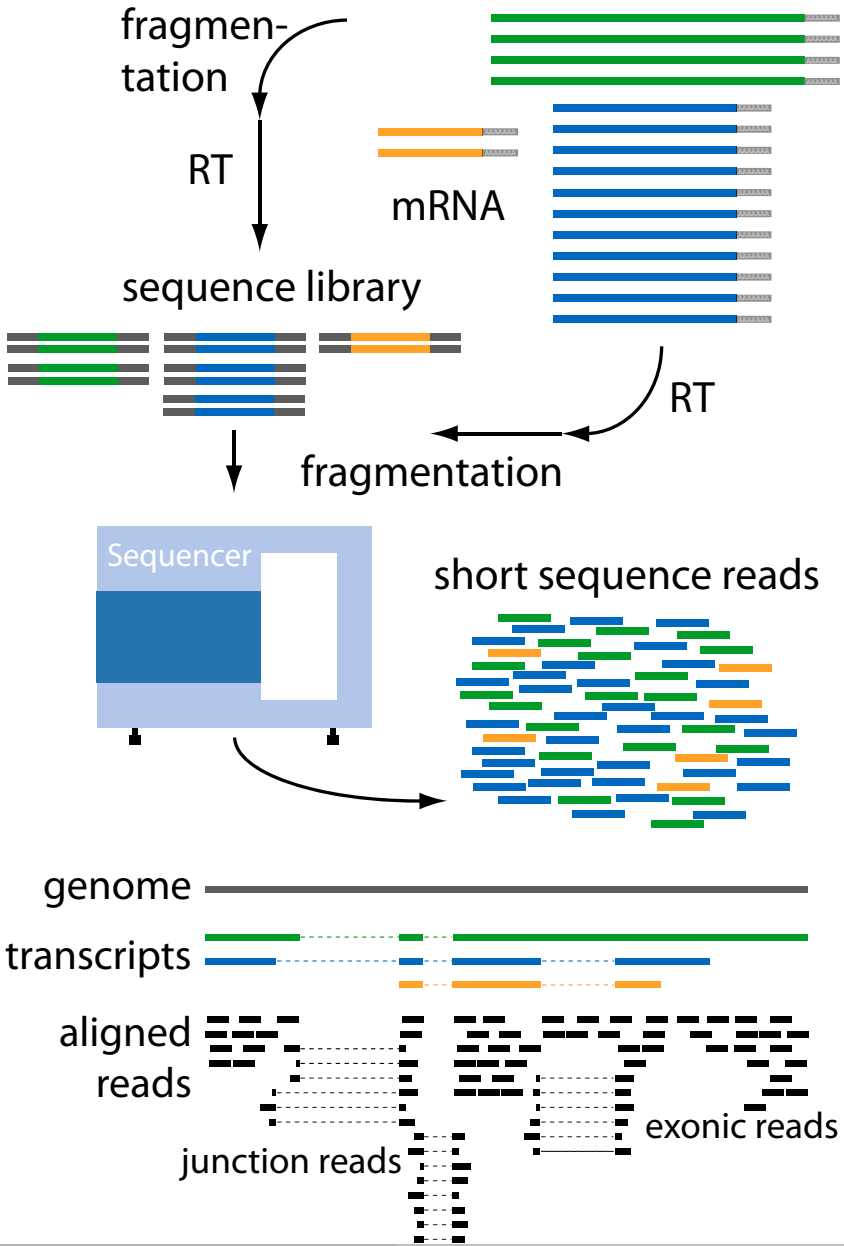
Need:

- Good inference method
- Inputs (DNA, $\cancel{1\ 2\ 3}$)
- Outputs (complete transcriptome)

Intermediate Goal: Fully reconstruct RNAs from experimental data
Solution: Deep RNA sequencing (RNA-seq)

$$f(\text{DNA, RNA-seq}) = \text{RNA}$$

Deep RNA Sequencing (RNA-seq)

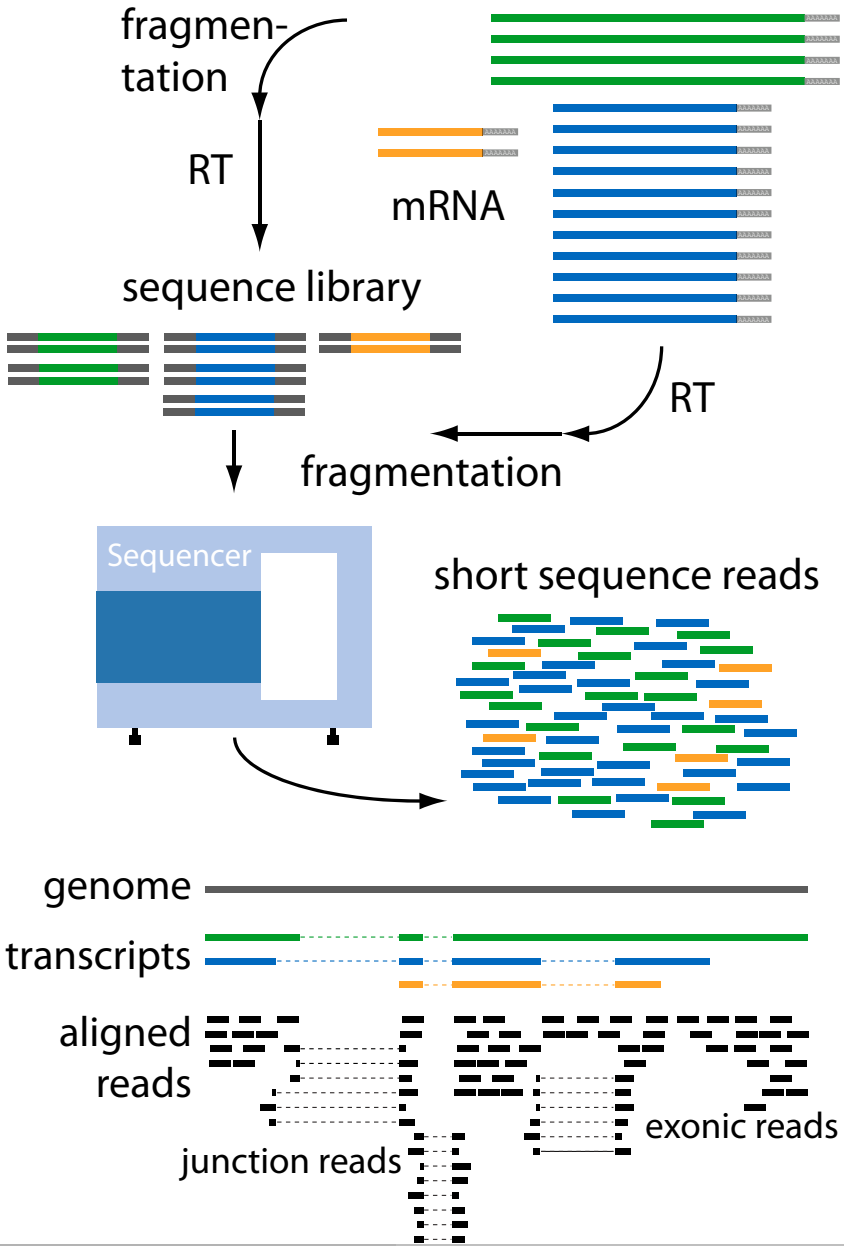


Sequencers produce $\approx 10^9$ reads of length 50-150nt (100-500Gbp) per run.

How to reconstruct RNAs?

How to quantify RNAs?

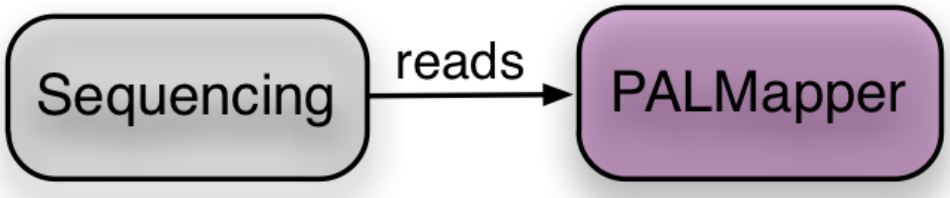
Deep RNA Sequencing (RNA-seq)



Sequencers produce $\approx 10^9$ reads of length 50-150nt (100-500Gbp) per run.

How to reconstruct RNAs?
How to quantify RNAs?

RNA-seq Analysis Steps

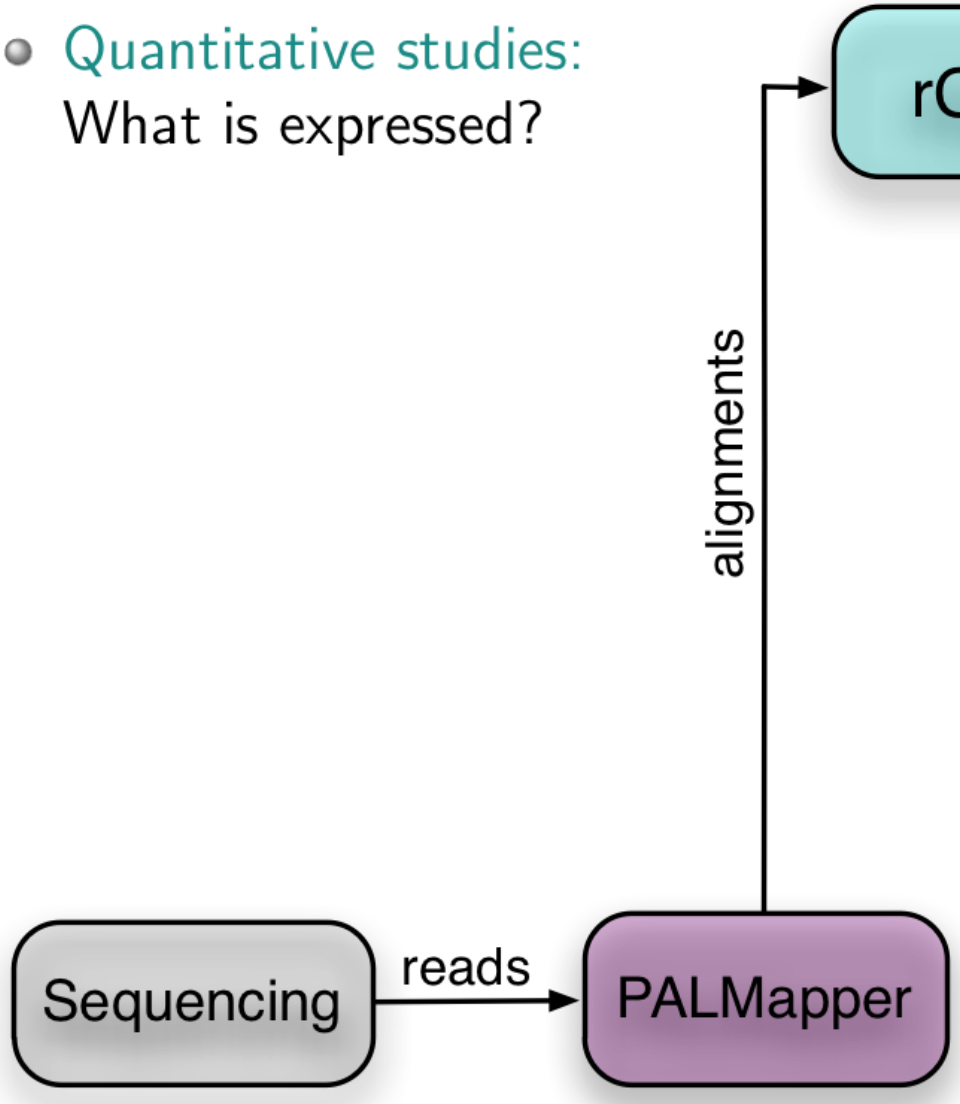


Accurate spliced alignments

[De Bona et al., 2008, Jean et al., 2010]

RNA-seq Analysis Steps

- Quantitative studies:
  What is expressed?

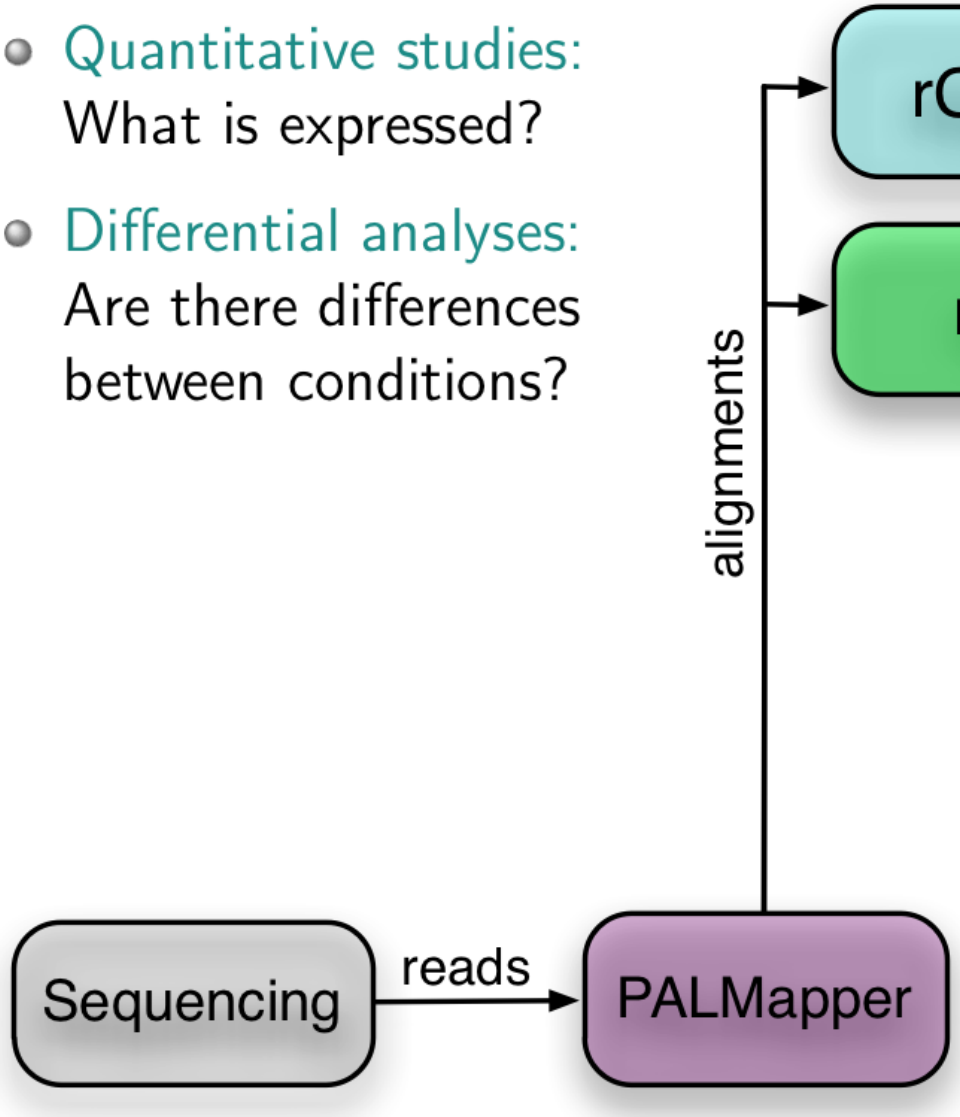


Isoform quantitation and bias modeling
[Bohnert et al., 2009, 2010]

Accurate spliced alignments
[De Bona et al., 2008, Jean et al., 2010]

RNA-seq Analysis Steps

- Quantitative studies:
  What is expressed?
- Differential analyses:
  Are there differences between conditions?



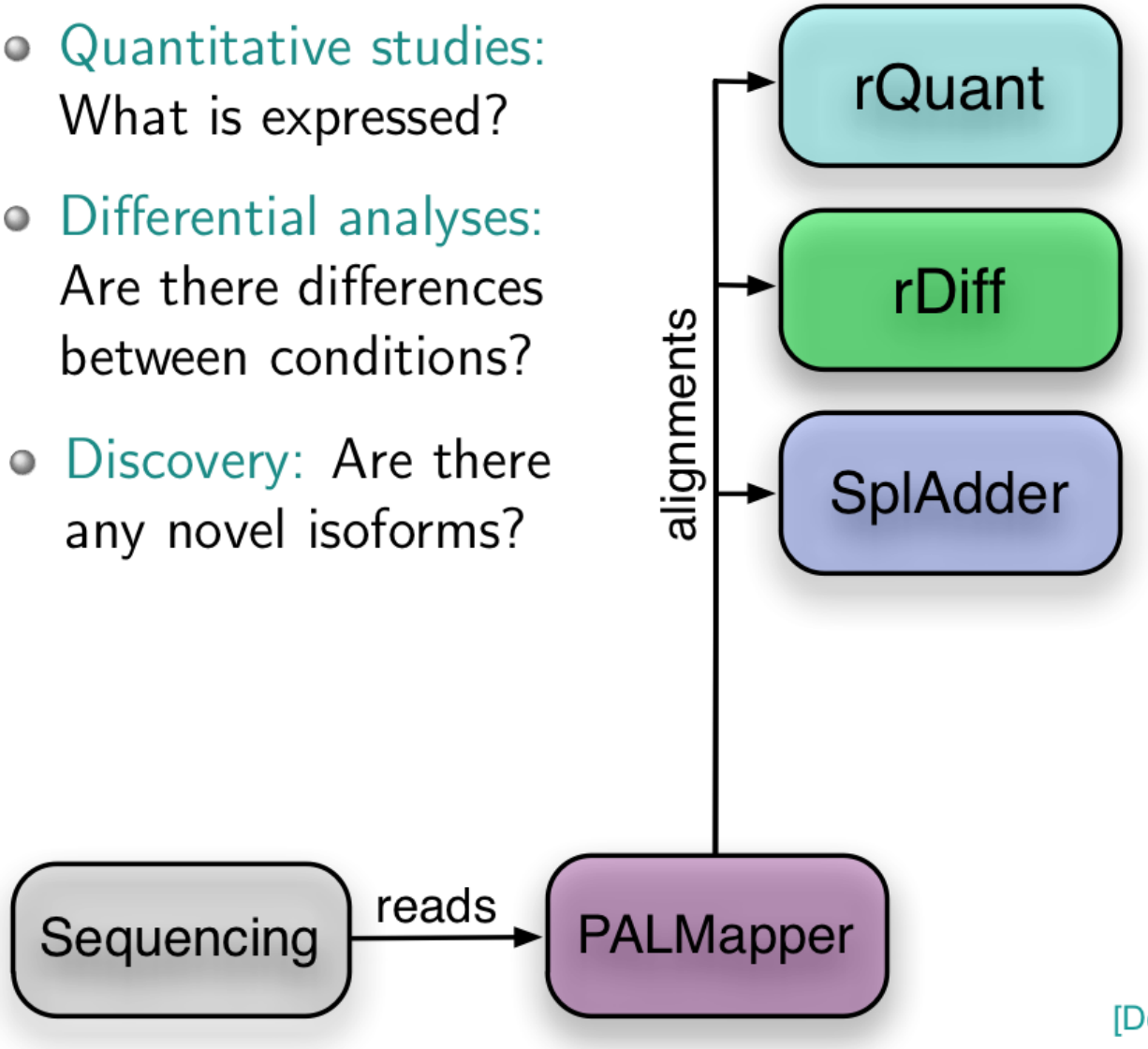
Isoform quantitation and bias modeling
[Bohnert et al., 2009, 2010]

Tests for differential isoform expression
[Stegle et al., 2010, 2011 i.p.]

Accurate spliced alignments
[De Bona et al., 2008, Jean et al., 2010]

RNA-seq Analysis Steps

- Quantitative studies: What is expressed?
- Differential analyses: Are there differences between conditions?
- Discovery: Are there any novel isoforms?



Isoform quantitation and bias modeling
[Bohnert et al., 2009, 2010]

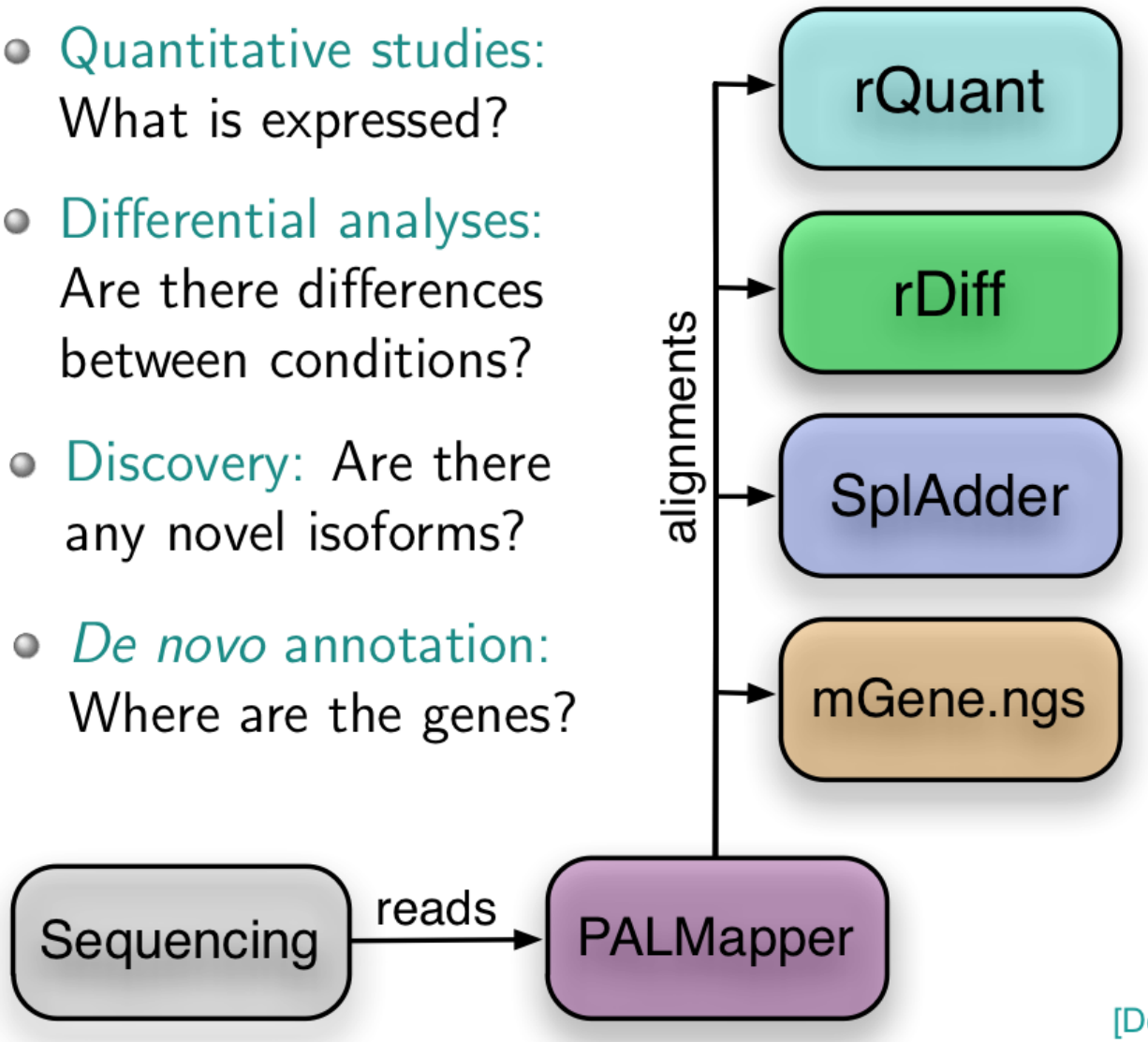
Tests for differential isoform expression
[Stegle et al., 2010, 2011 i.p.]

Graph construction & sampling
[Behr et al., 2011 i.p.]

Accurate spliced alignments
[De Bona et al., 2008, Jean et al., 2010]

RNA-seq Analysis Steps

- Quantitative studies: What is expressed?
- Differential analyses: Are there differences between conditions?
- Discovery: Are there any novel isoforms?
- *De novo* annotation: Where are the genes?



Isoform quantitation and bias modeling
[Bohnert et al., 2009, 2010]

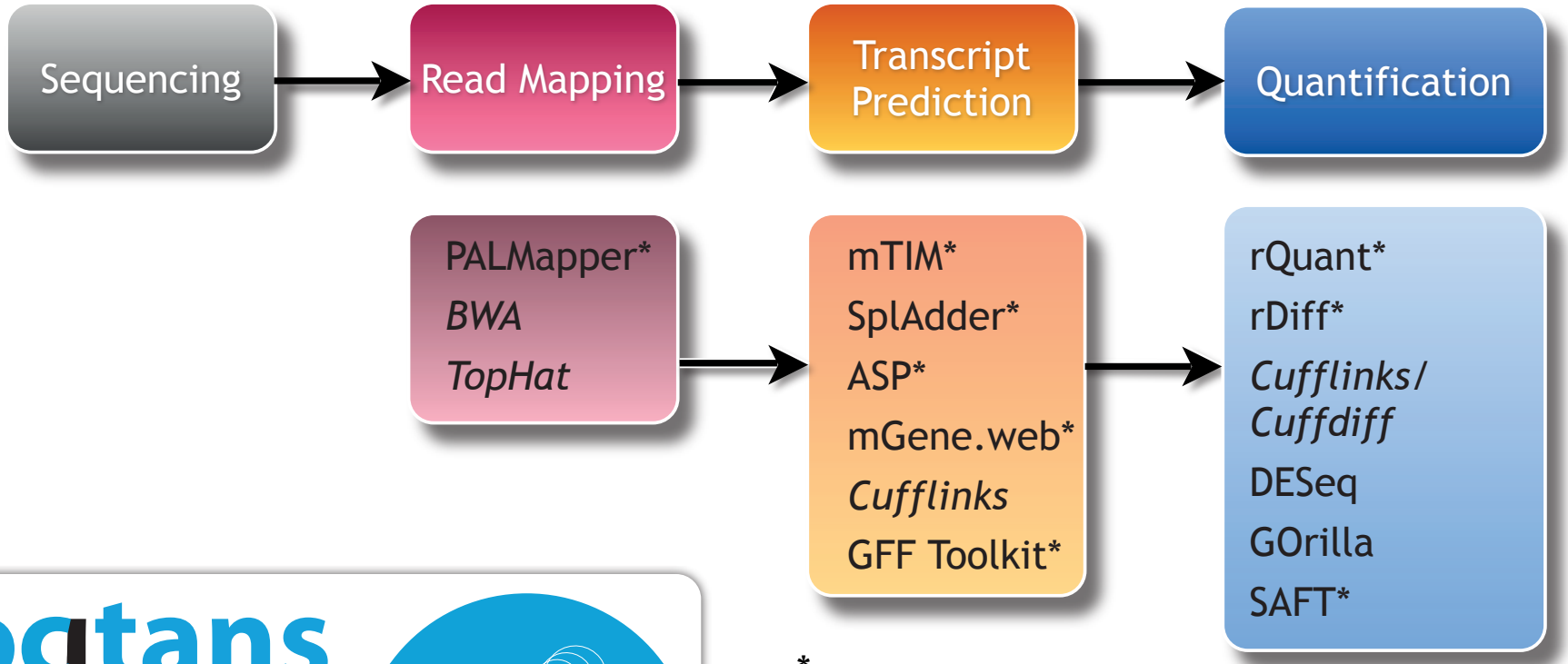
Tests for differential isoform expression
[Stegle et al., 2010, 2011 i.p.]

Graph construction & sampling
[Behr et al., 2011 i.p.]

Gene finding with RNA-seq evidence
[Behr et al., 2010, 2011 i.p.]

Accurate spliced alignments
[De Bona et al., 2008, Jean et al., 2010]

Oqtans: Galaxy-based Transcriptome Analysis



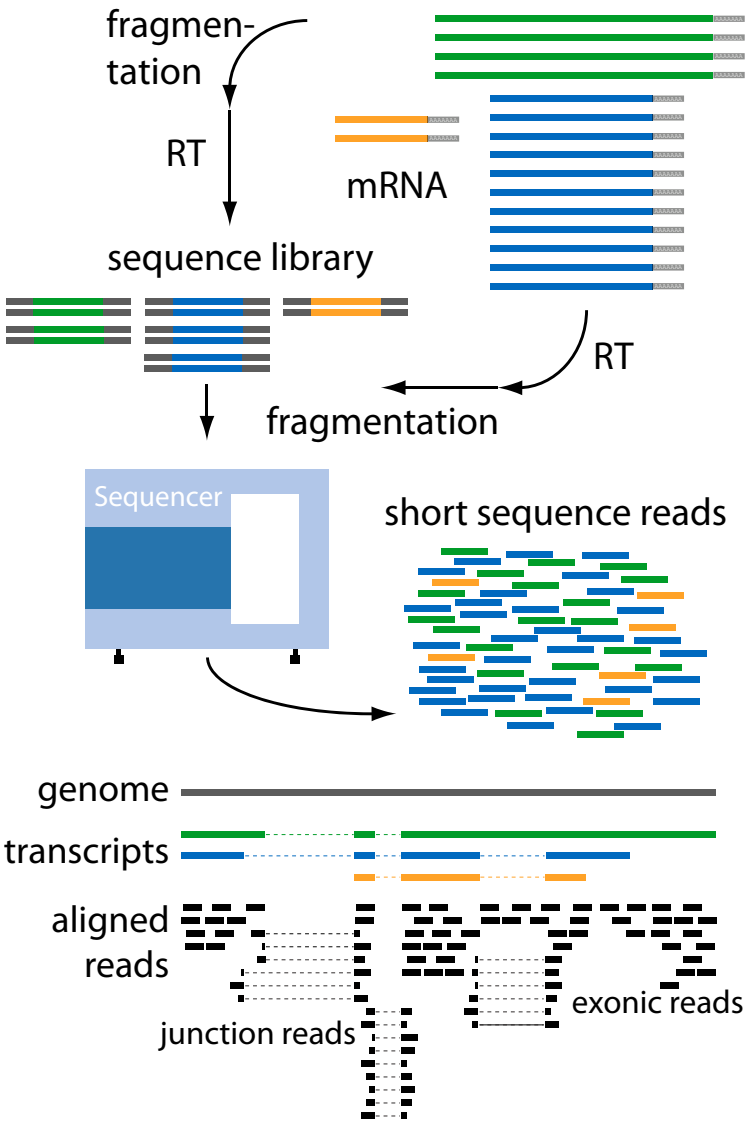
* *developed by the MLB Rätsch Research Group, FML, Tübingen*

oqtans

online quantitative transcript analysis

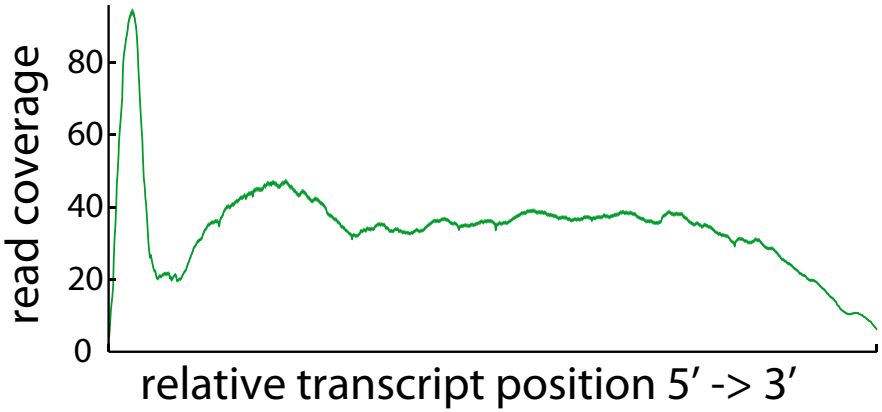
www.oqtans.org

RNA-seq Quantitation and Biases



Biases due to ...

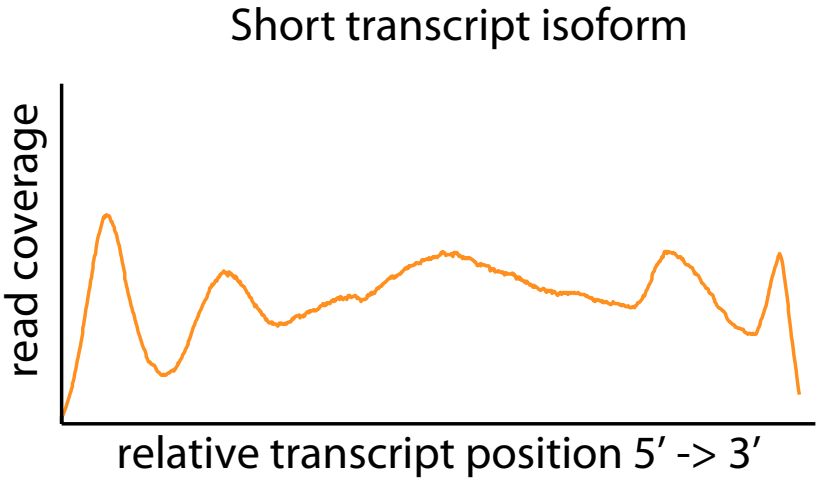
- cDNA library construction
- Sequencing
- Read mapping

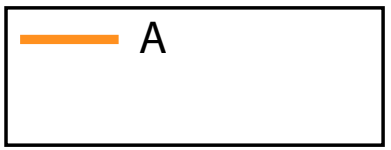


(average over annotated transcripts of length ≈1kb for the *C. elegans* SRX001872 dataset)

The rQuant Approach

[Bohnert & Rätsch, 2010]

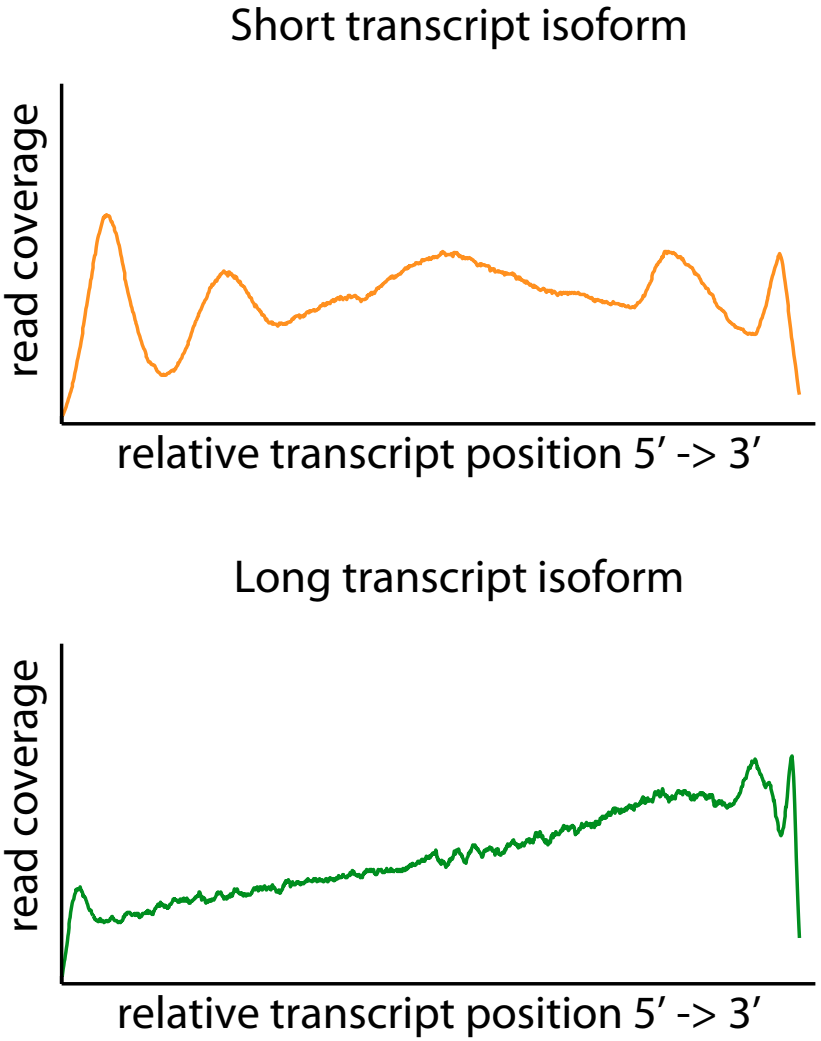




$$M_p = w_A A_p + w_B B_p \quad \Rightarrow \quad \min_{w_A, w_B} \sum_{p=1}^{P} \ell(M_p, R_p)$$

The rQuant Approach

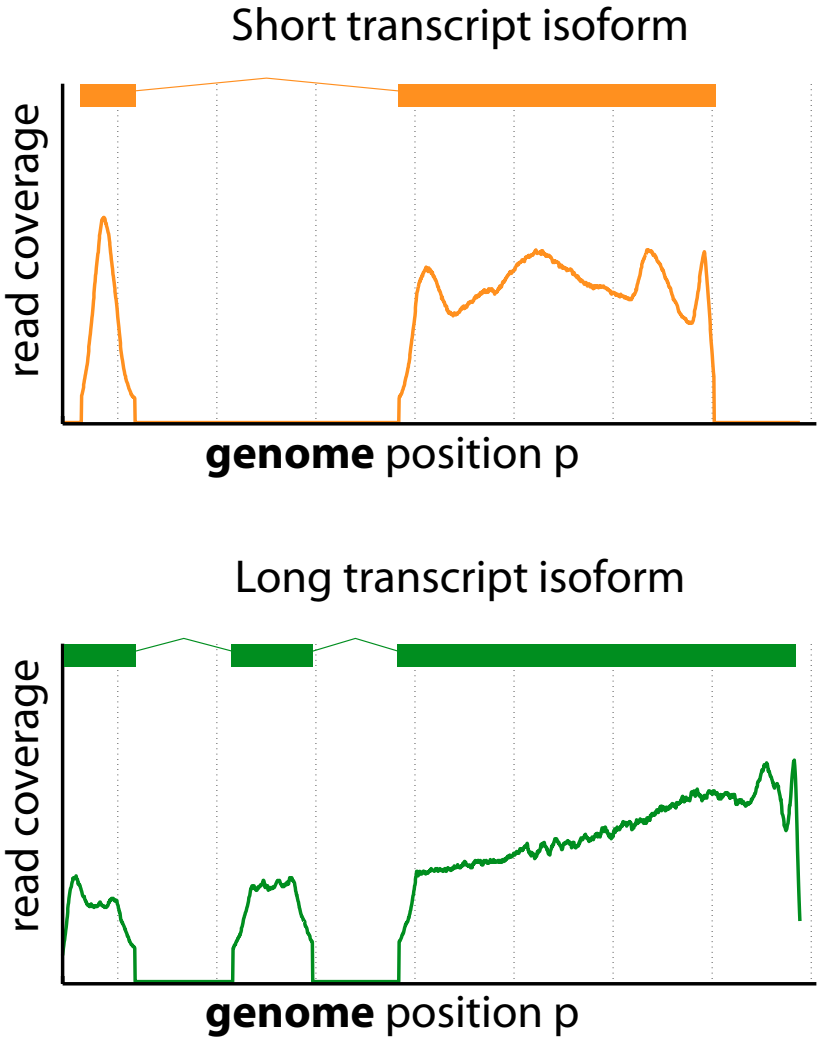
[Bohnert & Rätsch, 2010]



$$M_p = w_A A_p + w_B B_p \quad \Rightarrow \quad \min_{w_A, w_B} \sum_{p=1}^{P} \ell(M_p, R_p)$$

The rQuant Approach

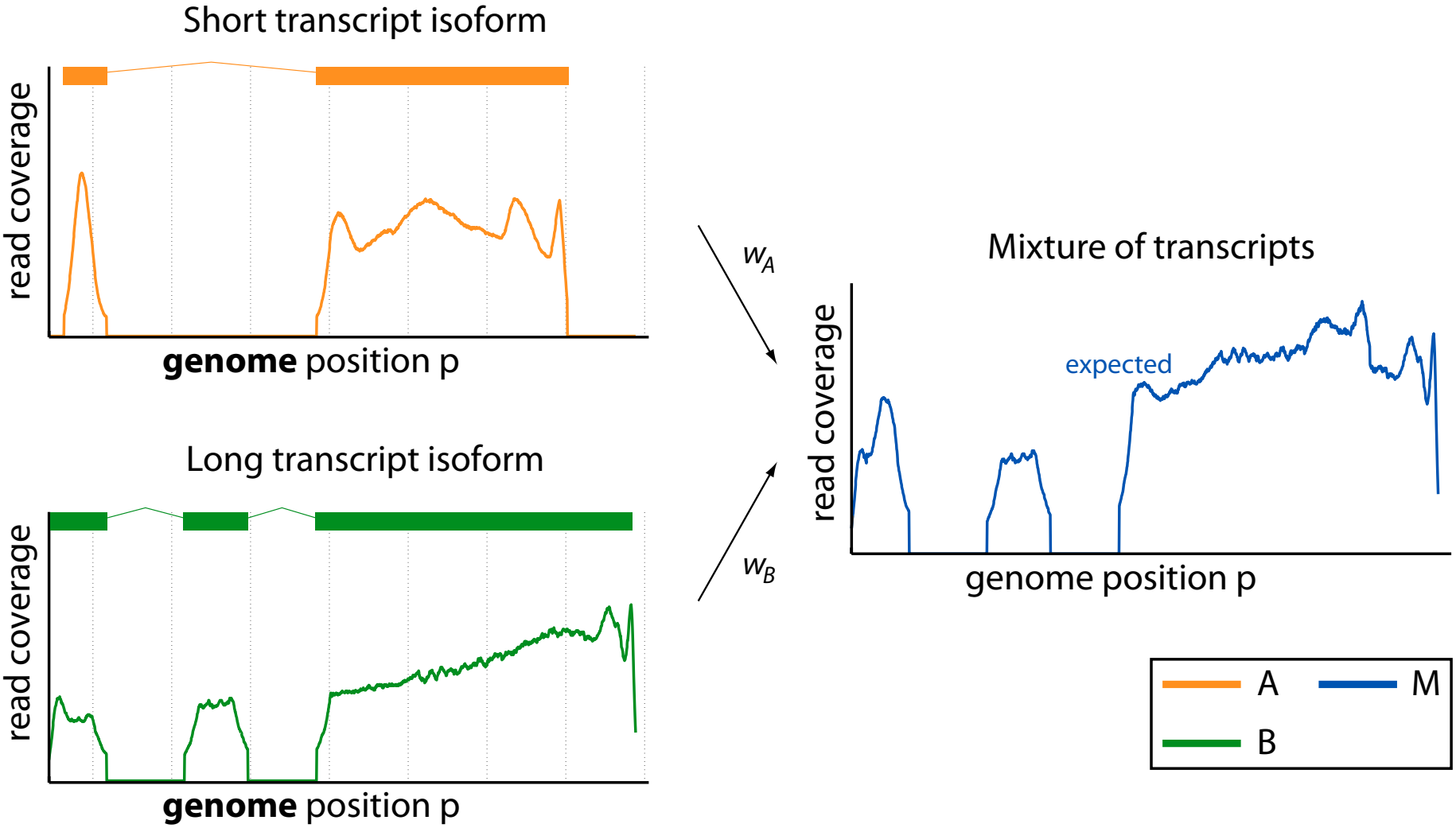
[Bohnert & Rätsch, 2010]



$$M_p = w_A A_p + w_B B_p \quad \Rightarrow \quad \min_{w_A, w_B} \sum_{p=1}^{P} \ell(M_p, R_p)$$

The rQuant Approach

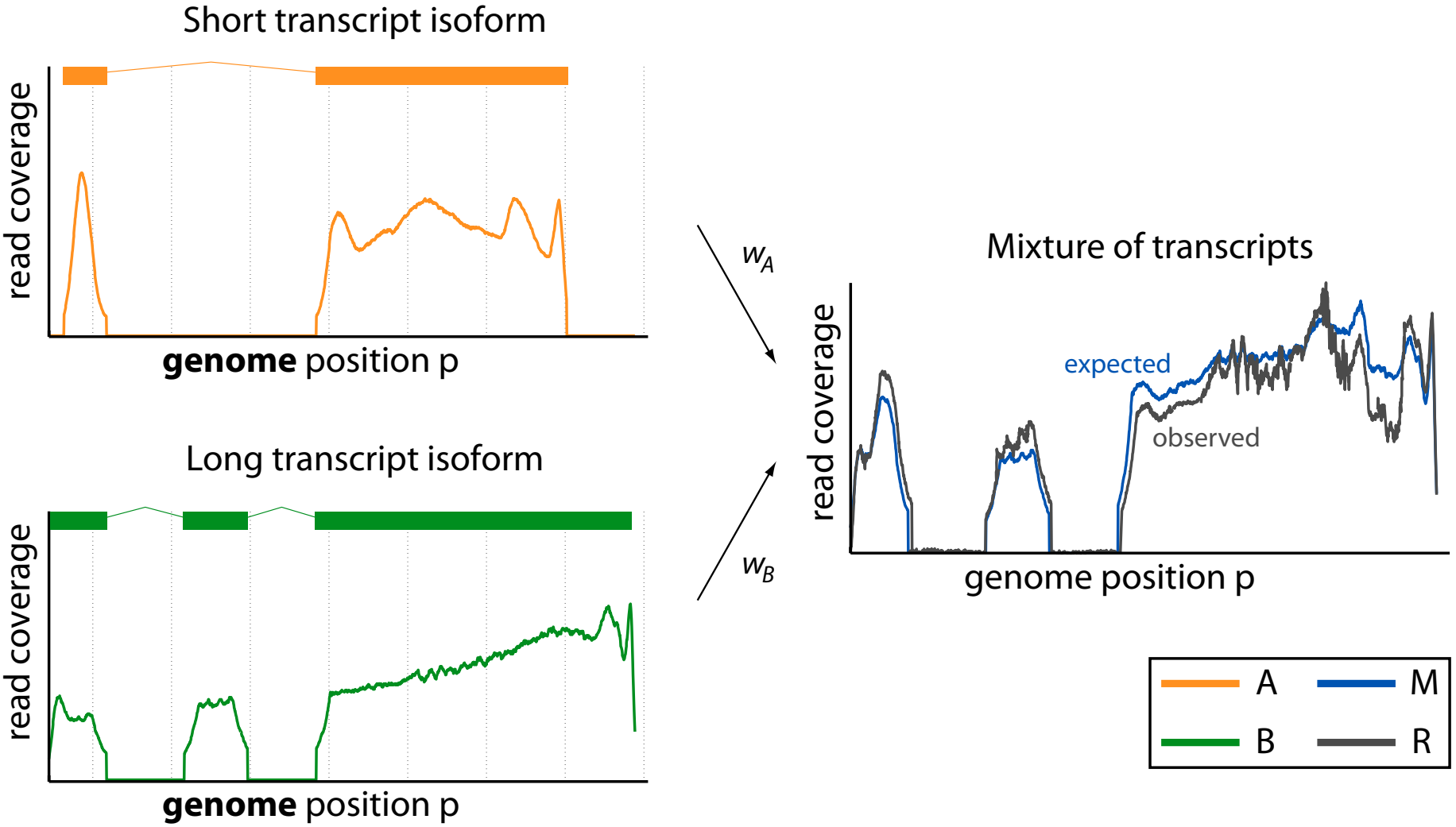
[Bohnert & Rätsch, 2010]



$$M_p = w_A A_p + w_B B_p \quad \Rightarrow \quad \min_{w_A, w_B} \sum_{p=1}^{P} \ell(M_p, R_p)$$

The rQuant Approach

[Bohnert & Rätsch, 2010]



$$M_p = w_A A_p + w_B B_p \quad \Rightarrow \quad \min_{w_A, w_B} \sum_{p=1}^{P} \ell\left(M_p, R_p\right)$$

Transcriptome Reconstruction with RNA-seq

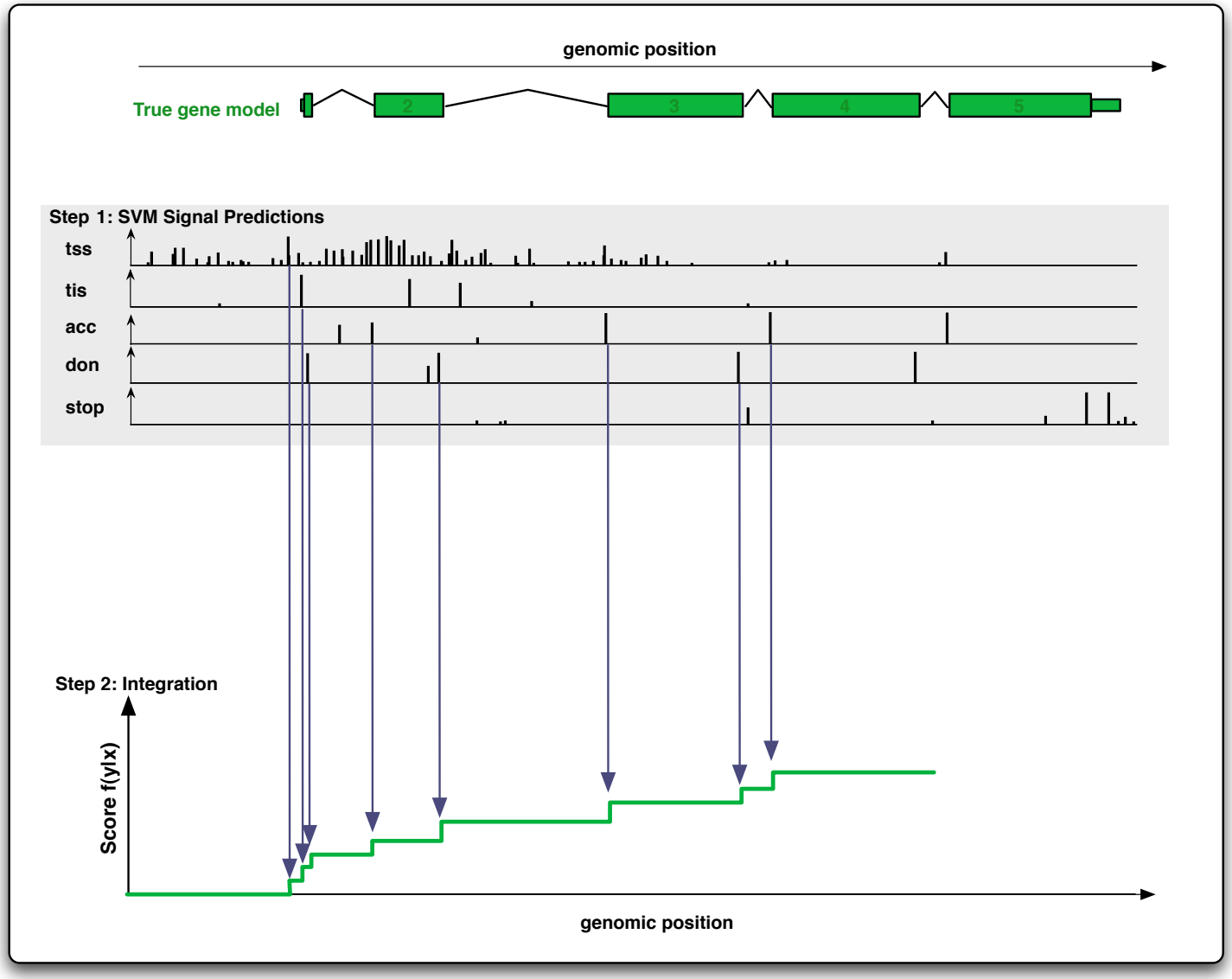
[Behr et al., 2010, 2011, i.p.]



Based on algorithms in [Altun et al., 2003, Rätsch and Sonnenburg, 2007]

Transcriptome Reconstruction with RNA-seq

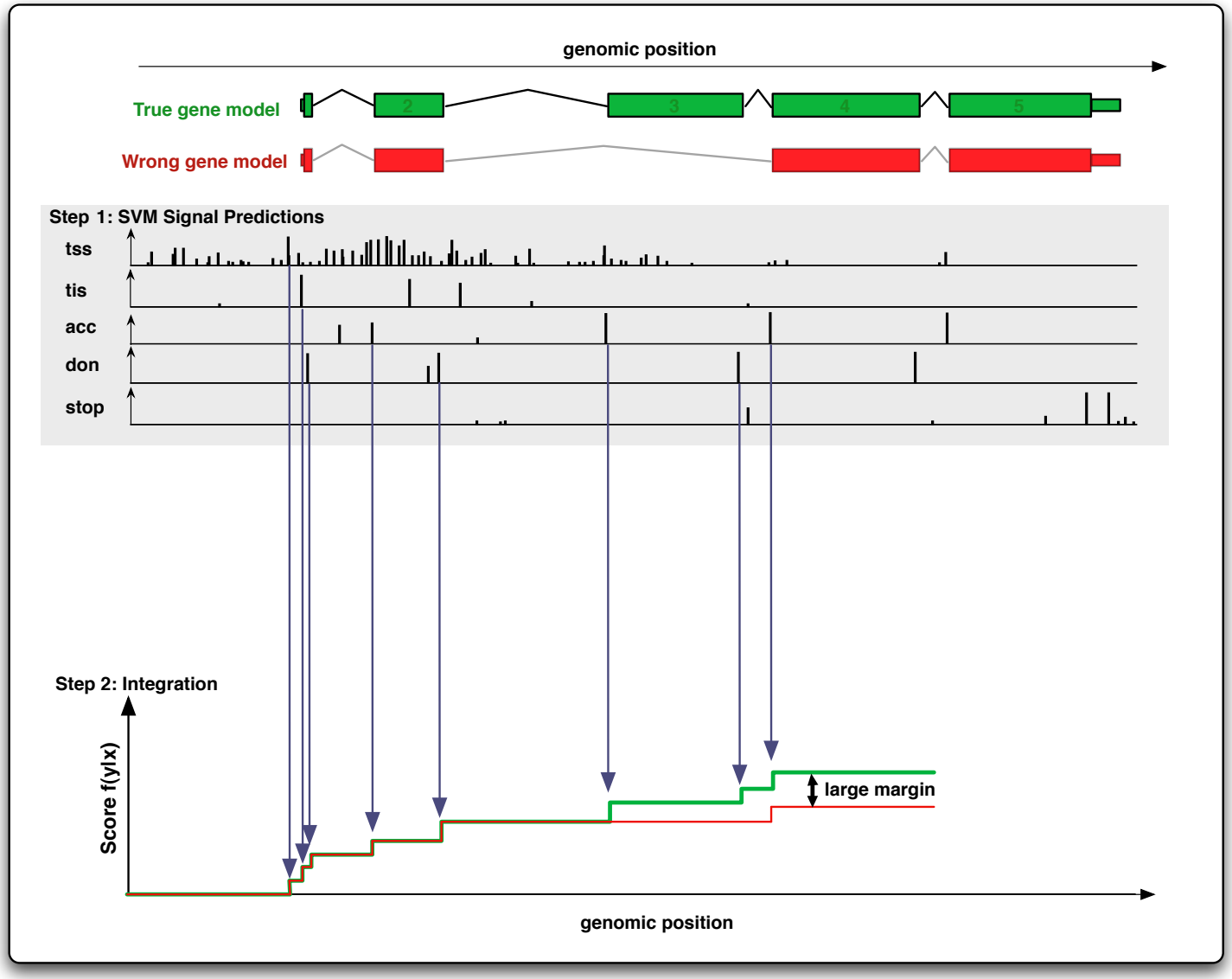
[Behr et al., 2010, 2011, i.p.]



Based on algorithms in [Altun et al., 2003, Rätsch and Sonnenburg, 2007]

Transcriptome Reconstruction with RNA-seq

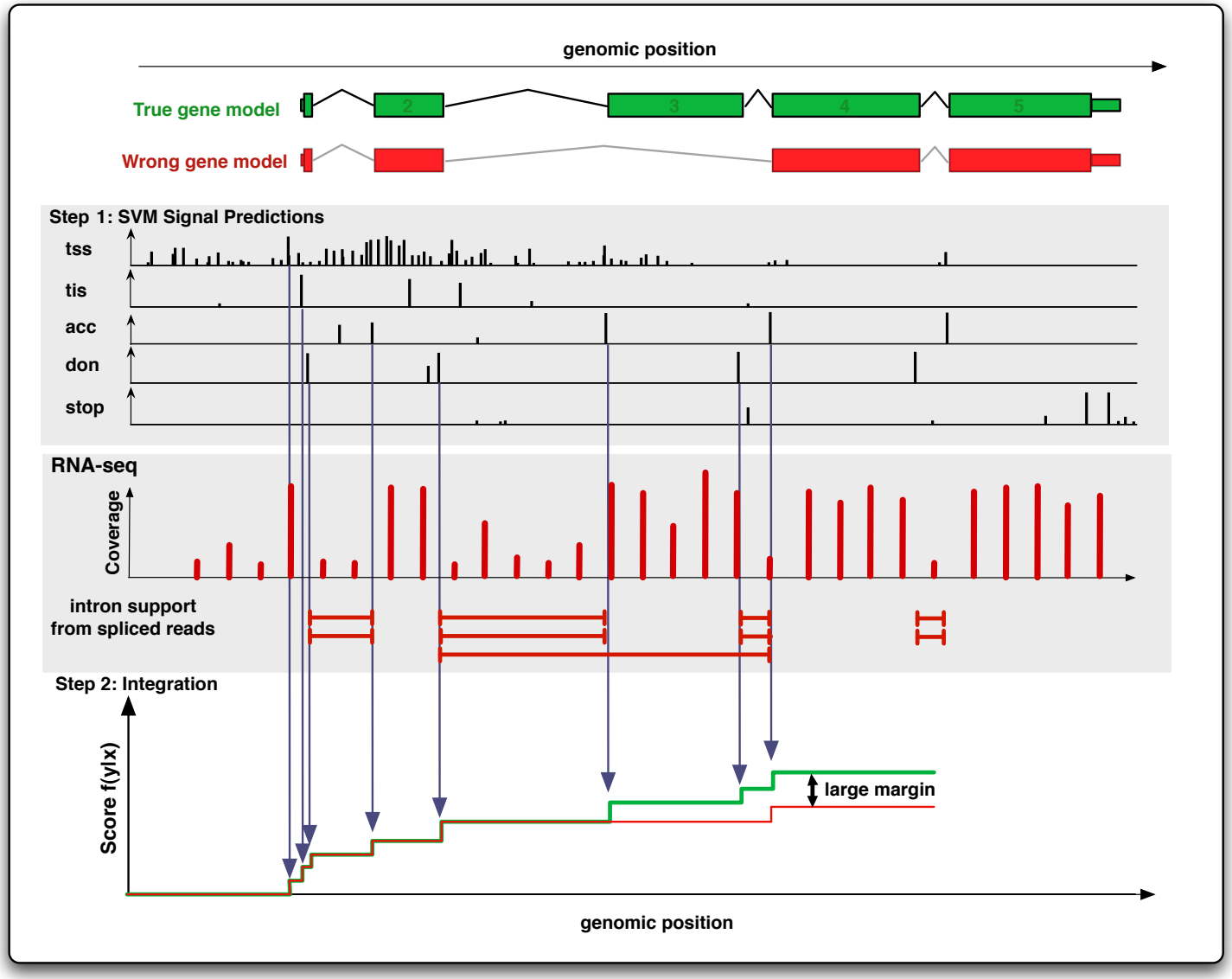
[Behr et al., 2010, 2011, i.p.]



Based on algorithms in [Altun et al., 2003, Rätsch and Sonnenburg, 2007]

Transcriptome Reconstruction with RNA-seq

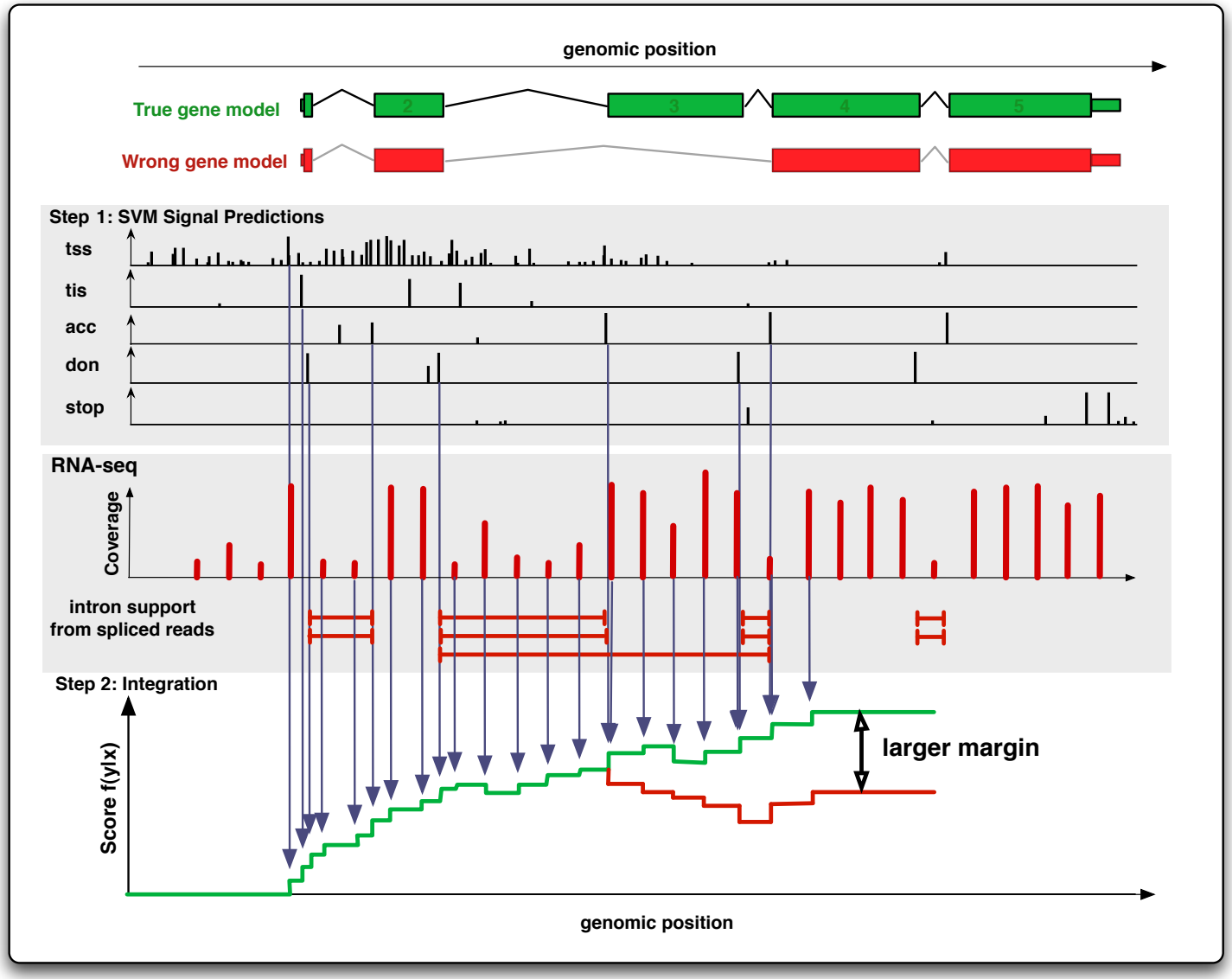
[Behr et al., 2010, 2011, i.p.]



Based on algorithms in [Altun et al., 2003, Rätsch and Sonnenburg, 2007]

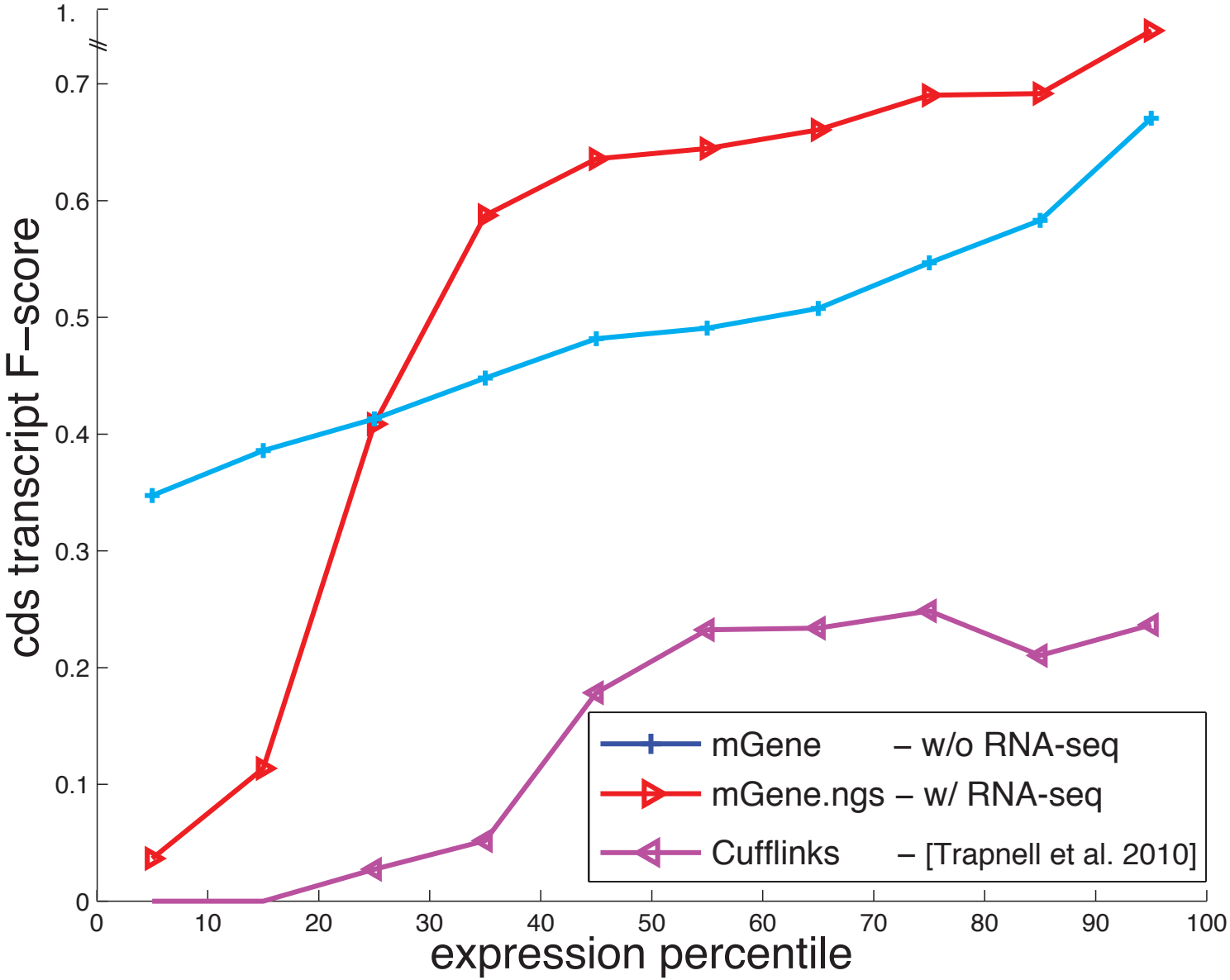
Transcriptome Reconstruction with RNA-seq

[Behr et al., 2010, 2011, i.p.]



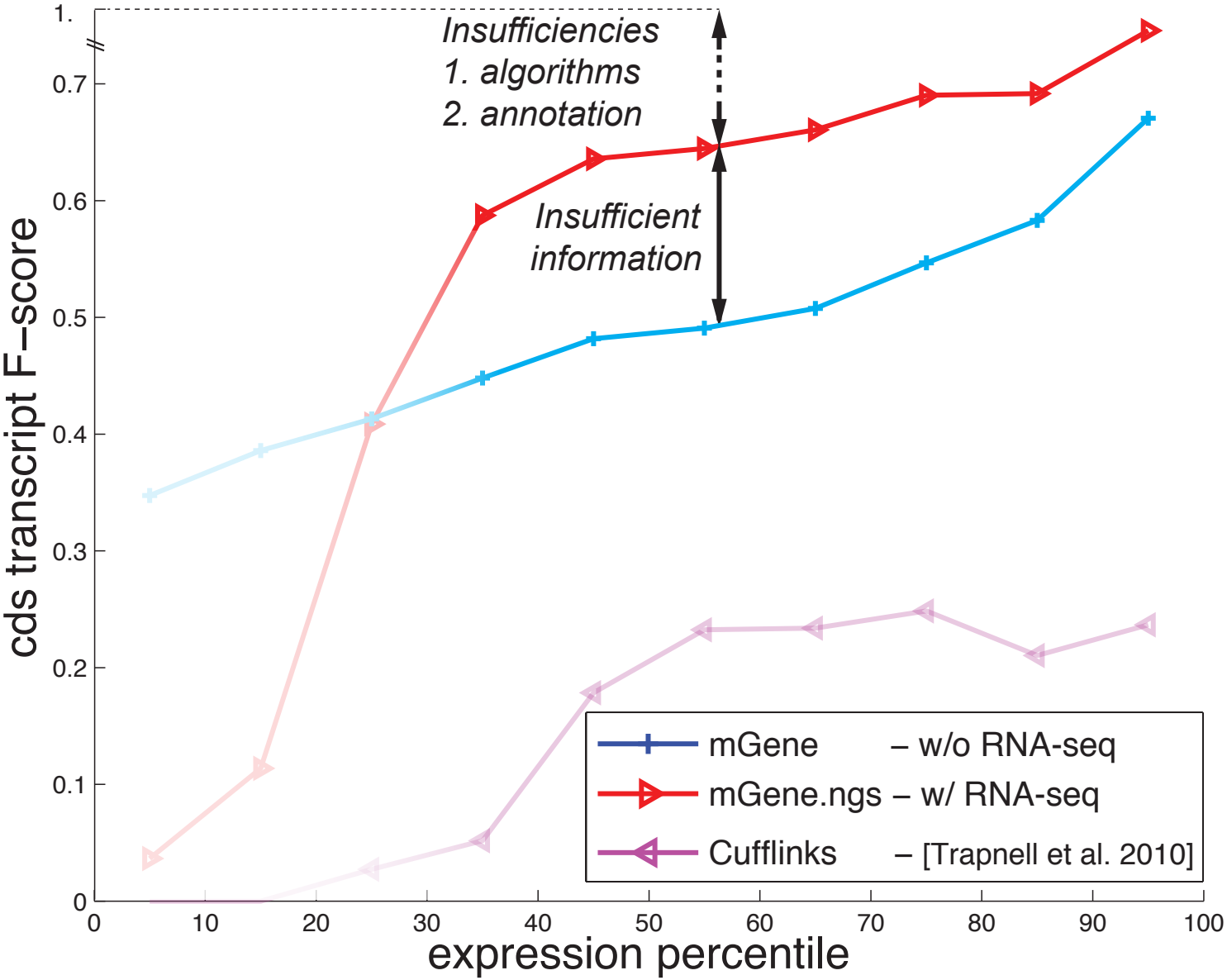
Based on algorithms in [Altun et al., 2003, Rätsch and Sonnenburg, 2007]

Transcriptome Reconstruction Accuracy



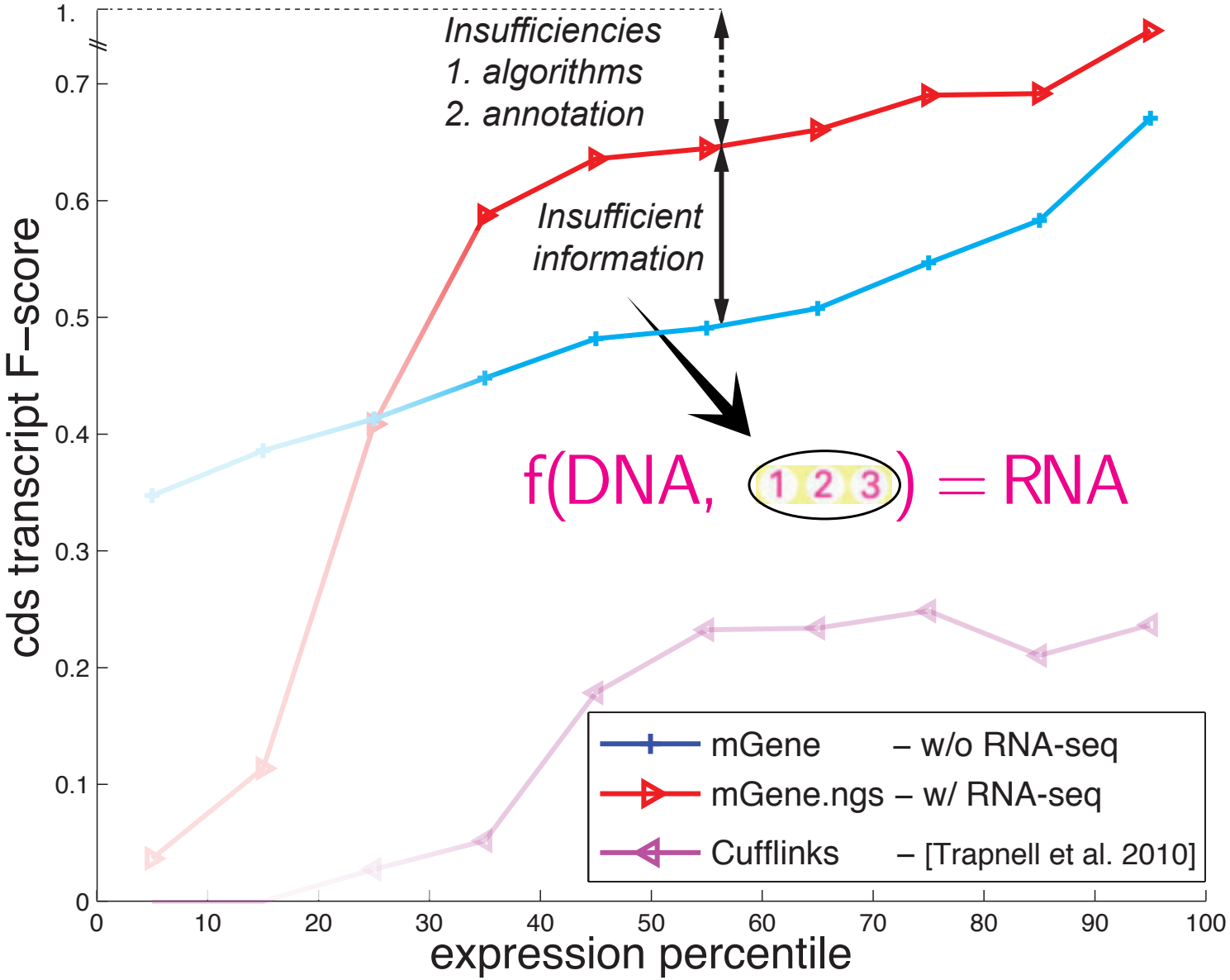
C. elegans, [Behr et al., 2011]

Transcriptome Reconstruction Accuracy



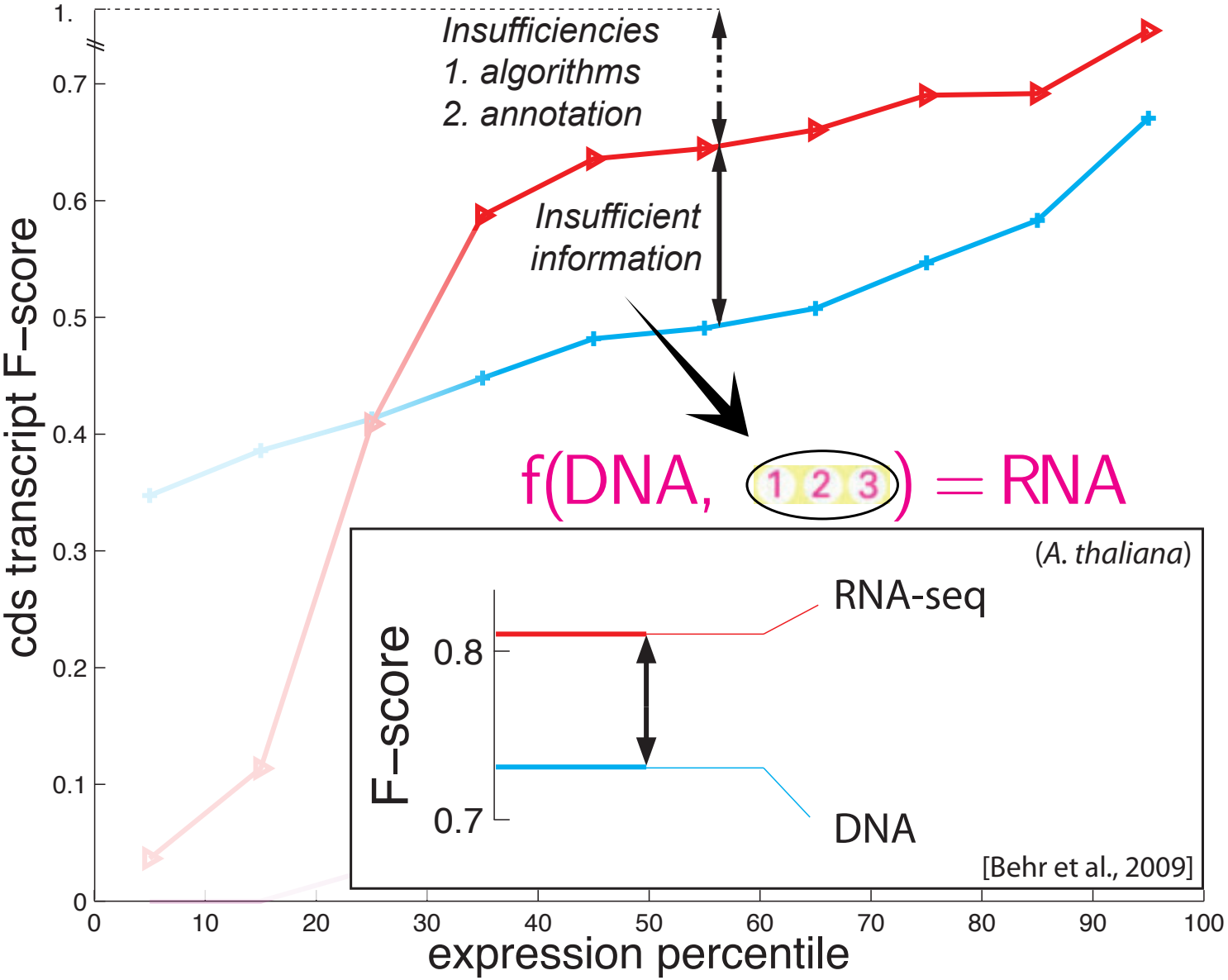
C. elegans, [Behr et al., 2011]

Transcriptome Reconstruction Accuracy



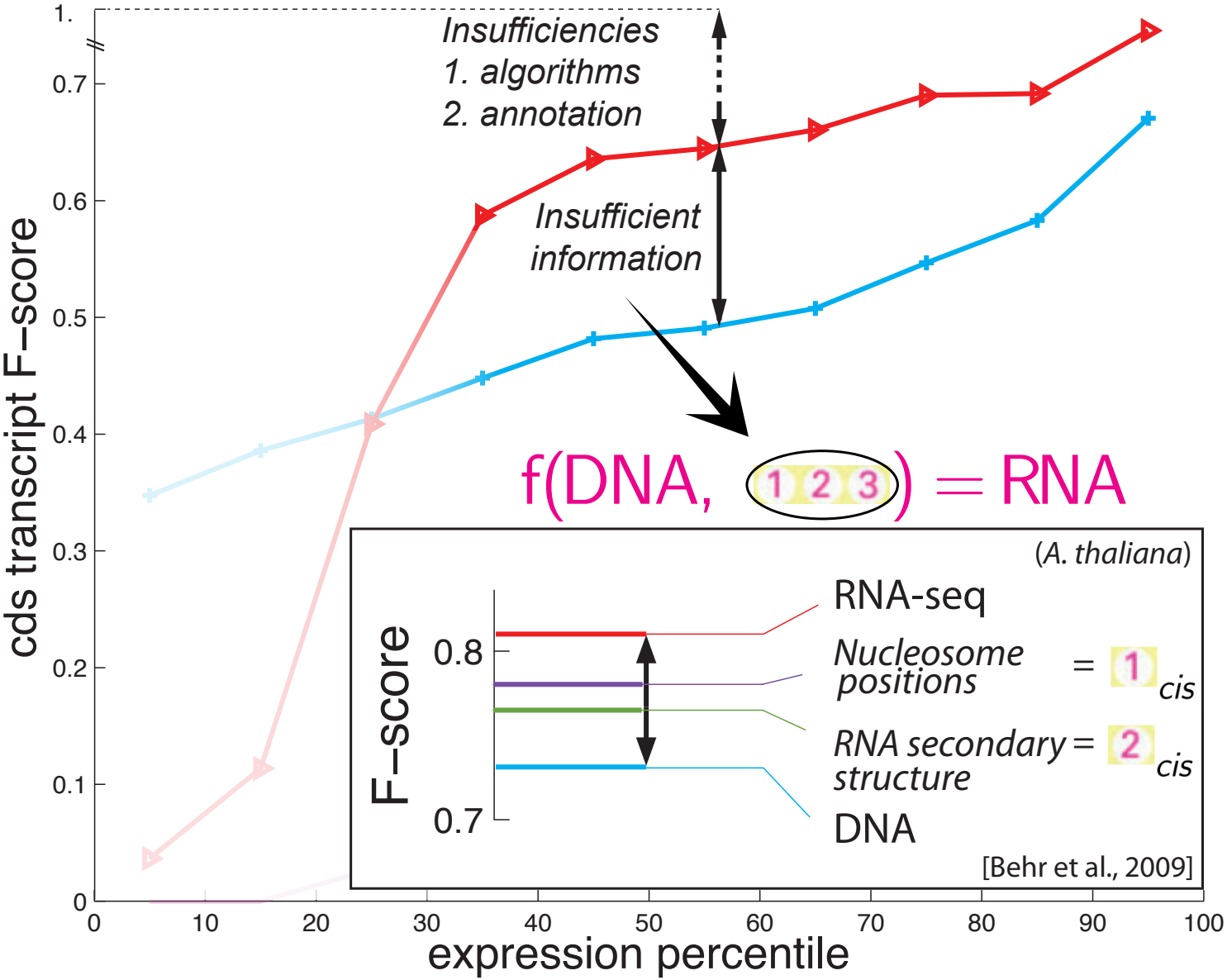
C. elegans, [Behr et al., 2011]

Transcriptome Reconstruction Accuracy



C. elegans, [Behr et al., 2011]

Transcriptome Reconstruction Accuracy



C. elegans, [Behr et al., 2011]

Digestion

We can

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

Limitations:

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

Need:

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

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

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

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

Digestion

We can

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

Limitations:

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

Need:

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

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

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

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

Digestion

We can

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

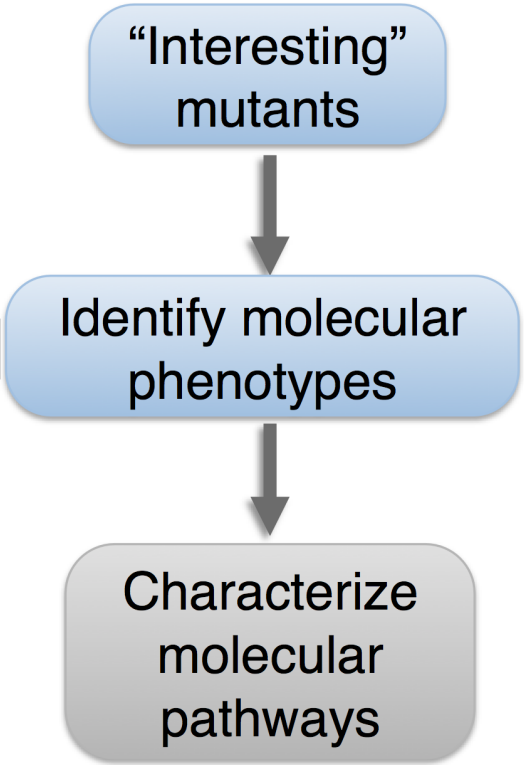
Limitations:

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

Need:

$$\begin{aligned} f(\text{DNA}, \boxed{1}_{cis/trans}, \boxed{2}_{cis/trans}) &= \text{RNA} \\ F(\text{RNA}, \boxed{4}, \boxed{5}) &= \text{protein} \\ f'(\text{DNA}, \boxed{1}_{cis/trans}, \boxed{2}_{cis/trans}) &= d\,\text{RNA}/d\,\text{t} \\ F'(\text{RNA}, \boxed{4}, \boxed{5}) &= d\,\text{protein}/d\,\text{t} \end{aligned}$$

Approaches to Understanding Molecular Pathways

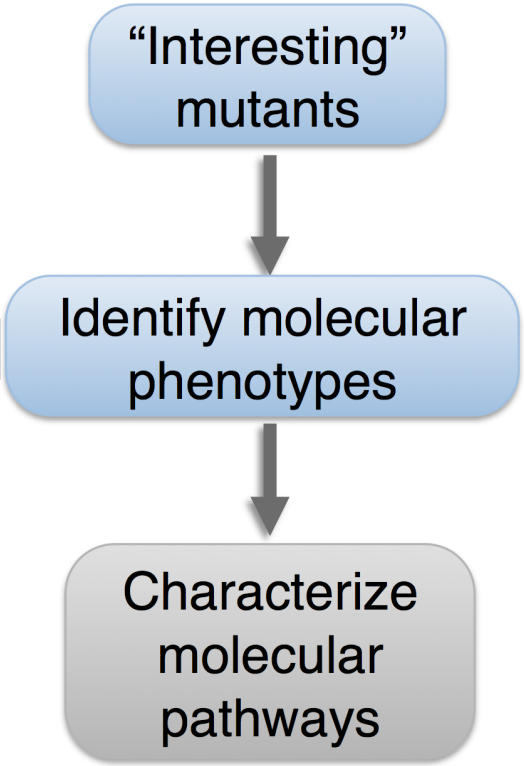


- CBP20/80 [Laubinger et al., 2008]
- sRNA biogenesis (*dcl1-4*) [Laubinger et al., 2010]
- NMD [w/ A. Wachter]
- Exosome [w/ E. Conti]

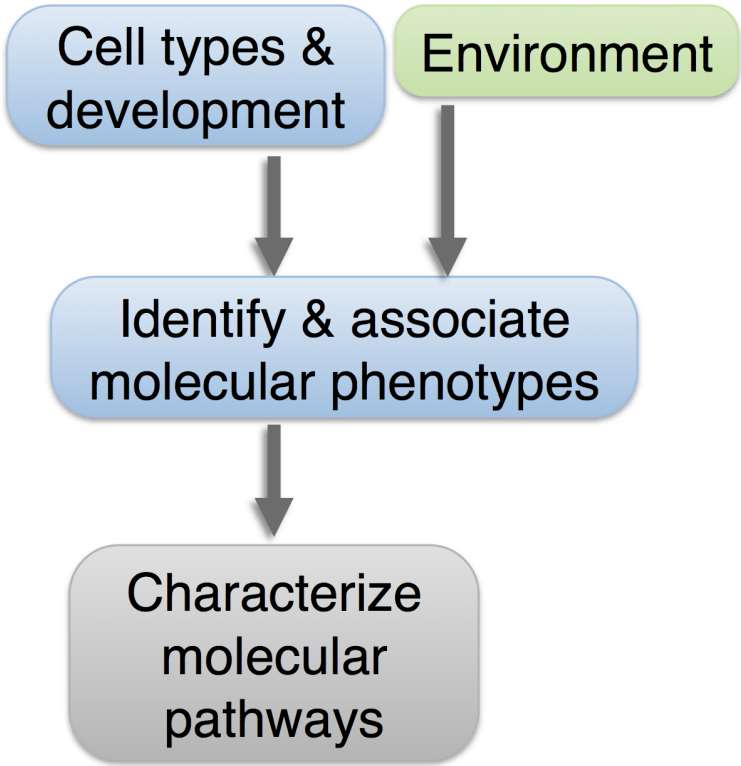
- *A. thaliana* tissue atlas [Laubinger et al., 2008]
- *A. thaliana* stress & light response [Zeller et al., 2009, Hartmann et al., 2011, i.P.]
- *C. elegans* tissue/development atlas [Gerstein et al., 2010, Spencer et al., 2011]

- *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 Understanding Molecular Pathways



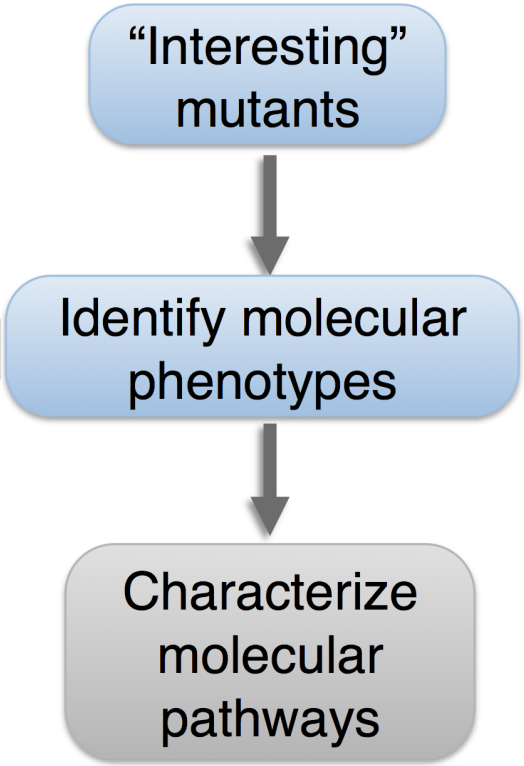
- CBP20/80 [Laubinger et al., 2008]
- sRNA biogenesis (*dcl1-4*) [Laubinger et al., 2010]
- NMD [w/ A. Wachter]
- Exosome [w/ E. Conti]



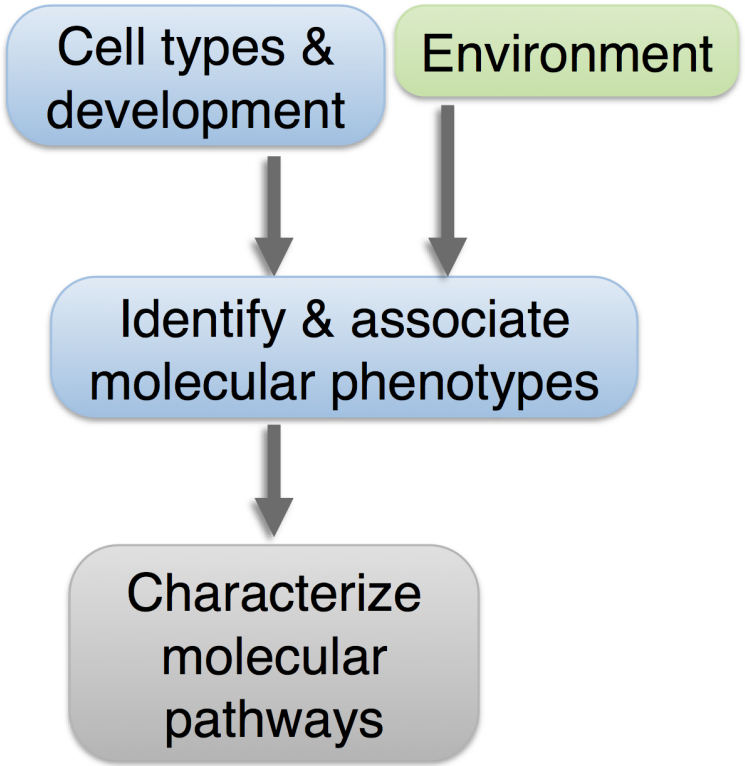
- *A. thaliana* tissue atlas [Laubinger et al., 2008]
- *A. thaliana* stress & light response [Zeller et al., 2009, Hartmann et al., 2011, i.P.]
- *C. elegans* tissue/development atlas [Gerstein et al., 2010, Spencer et al., 2011]

- *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]

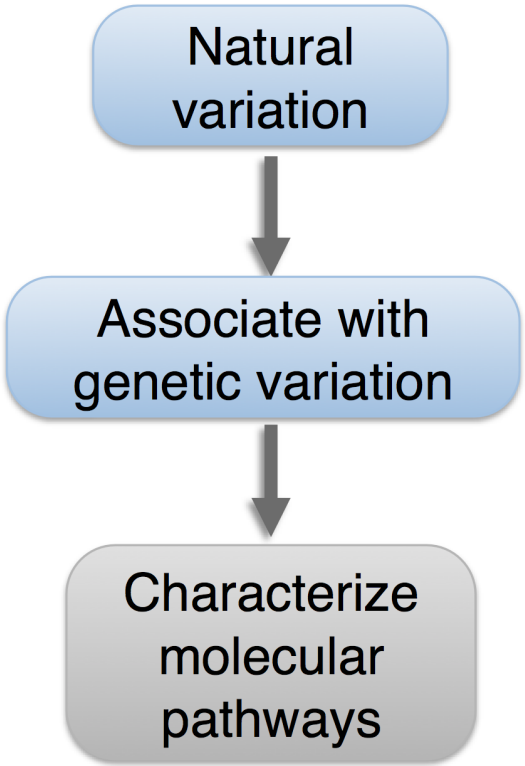
Approaches to Understanding Molecular Pathways



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

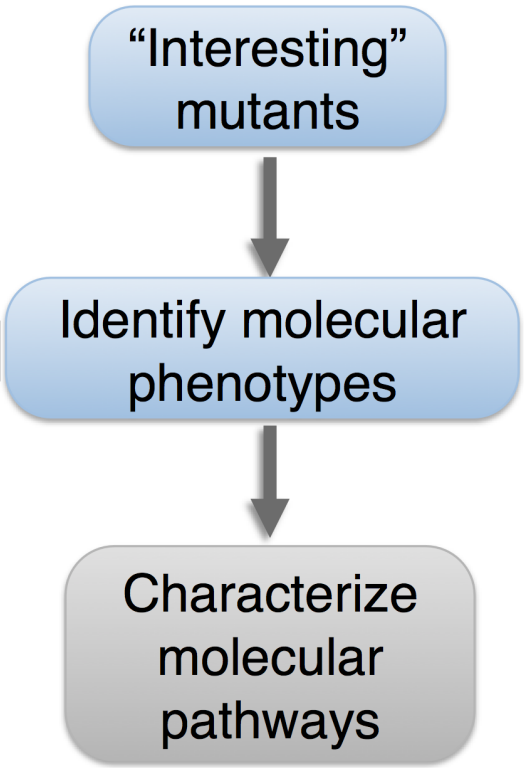


- *A. thaliana* tissue atlas [Laubinger et al., 2008]
- *A. thaliana* stress & light response [Zeller et al., 2009, Hartmann et al., 2011, i.P.]
- *C. elegans* tissue/development atlas [Gerstein et al., 2010, Spencer et al., 2011]

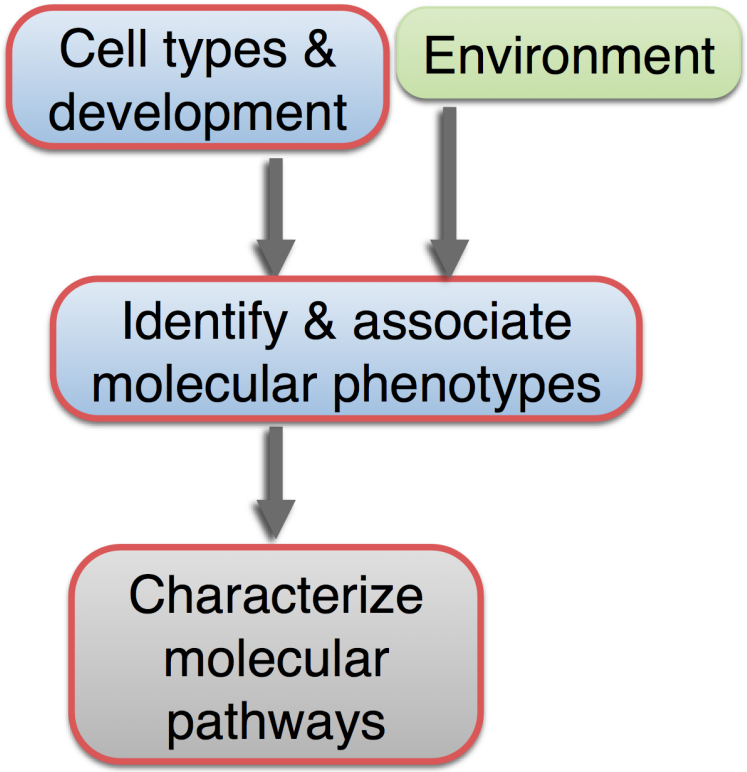


- *A. thaliana* genome variation [Clark et al., 2007a]
- *O. sativa* genome variation [McNally et al., 2009]
- *A. thaliana* transcriptome variation [w/ R. Mott & R. Clark]

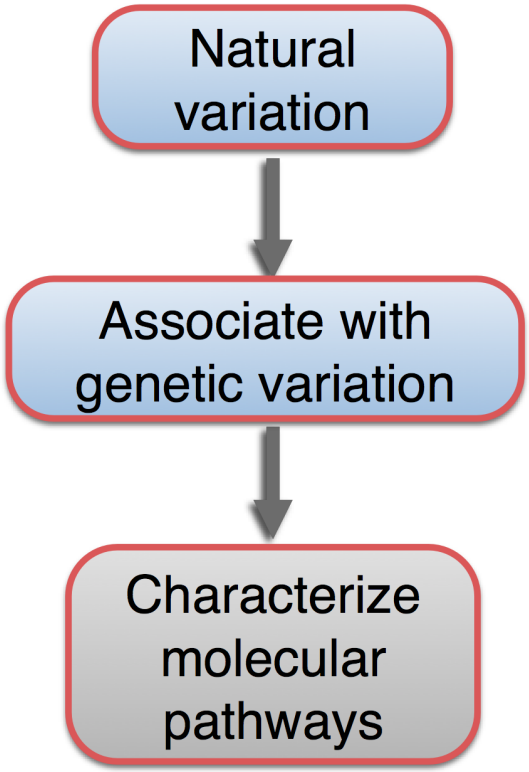
Approaches to Understanding Molecular Pathways



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

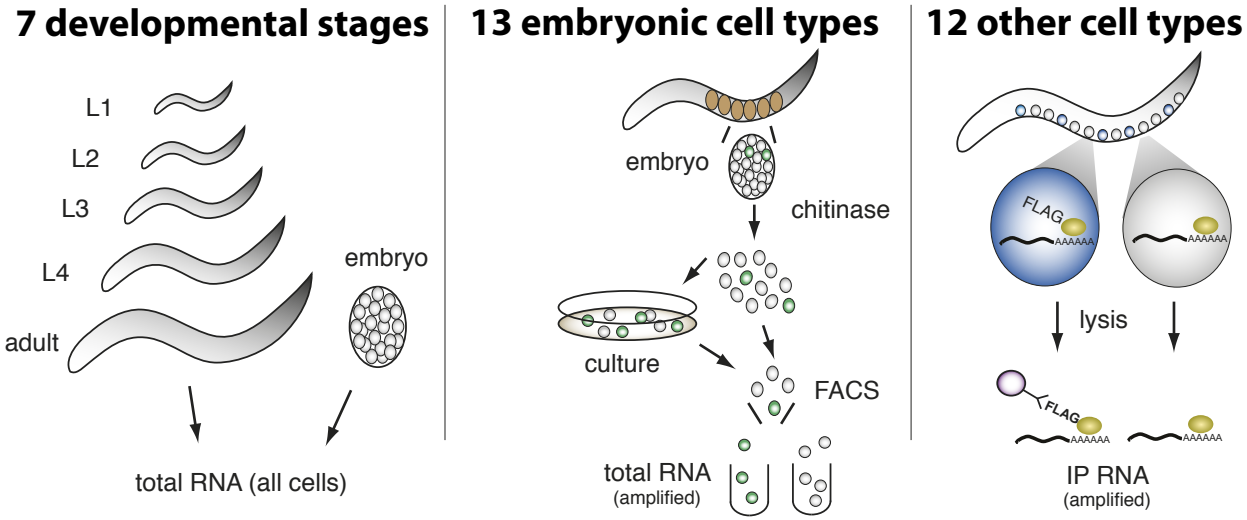


- *A. thaliana* tissue atlas [Laubinger et al., 2008]
- *A. thaliana* stress & light response [Zeller et al., 2009, Hartmann et al., 2011, i.P.]
- *C. elegans* tissue/development atlas [Gerstein et al., 2010, Spencer et al., 2011]



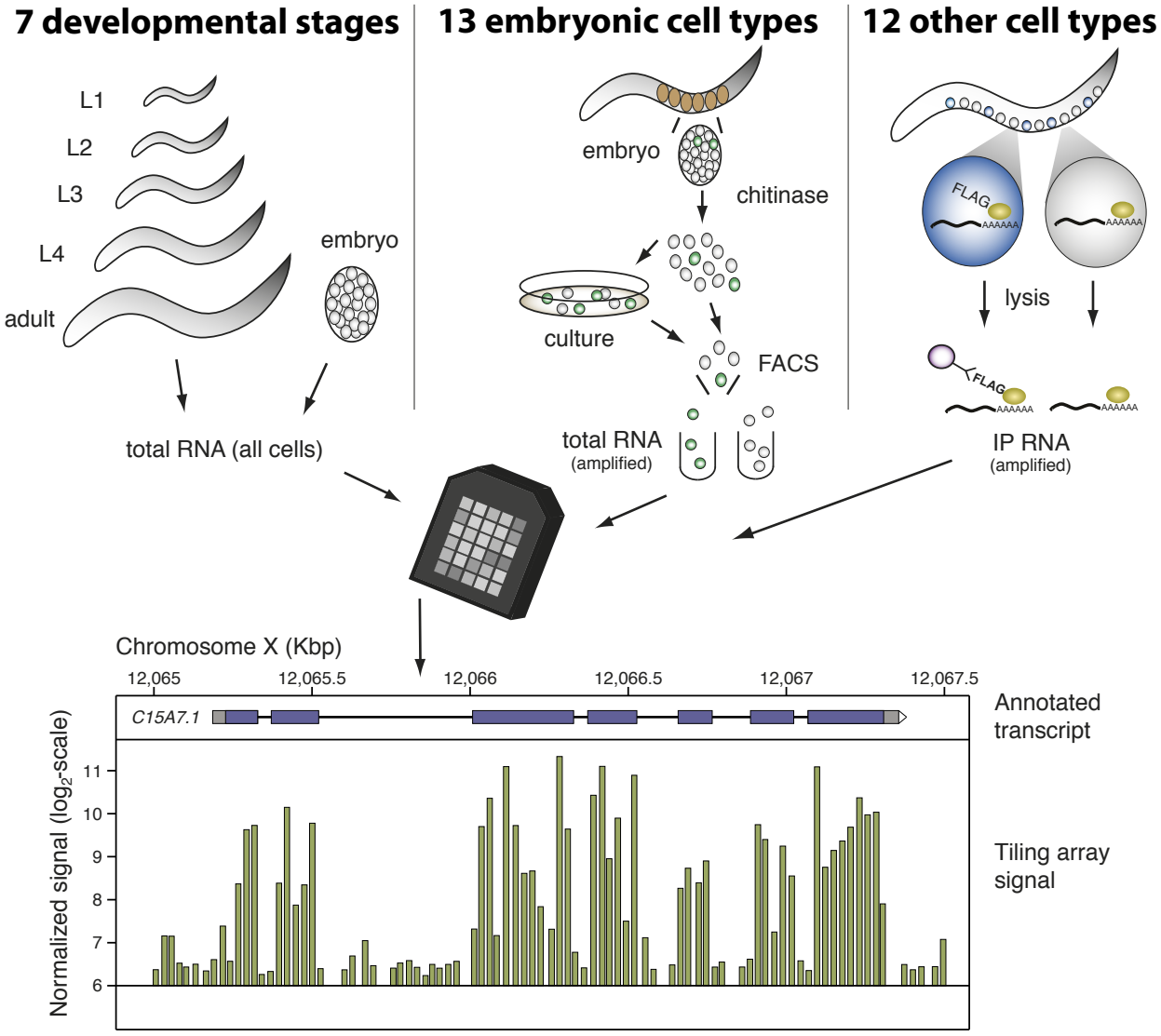
- *A. thaliana* genome variation [Clark et al., 2007a]
- *O. sativa* genome variation [McNally et al., 2009]
- *A. thaliana* transcriptome variation [w/ R. Mott & R. Clark]

Analysis of C. elegans Transcriptomes



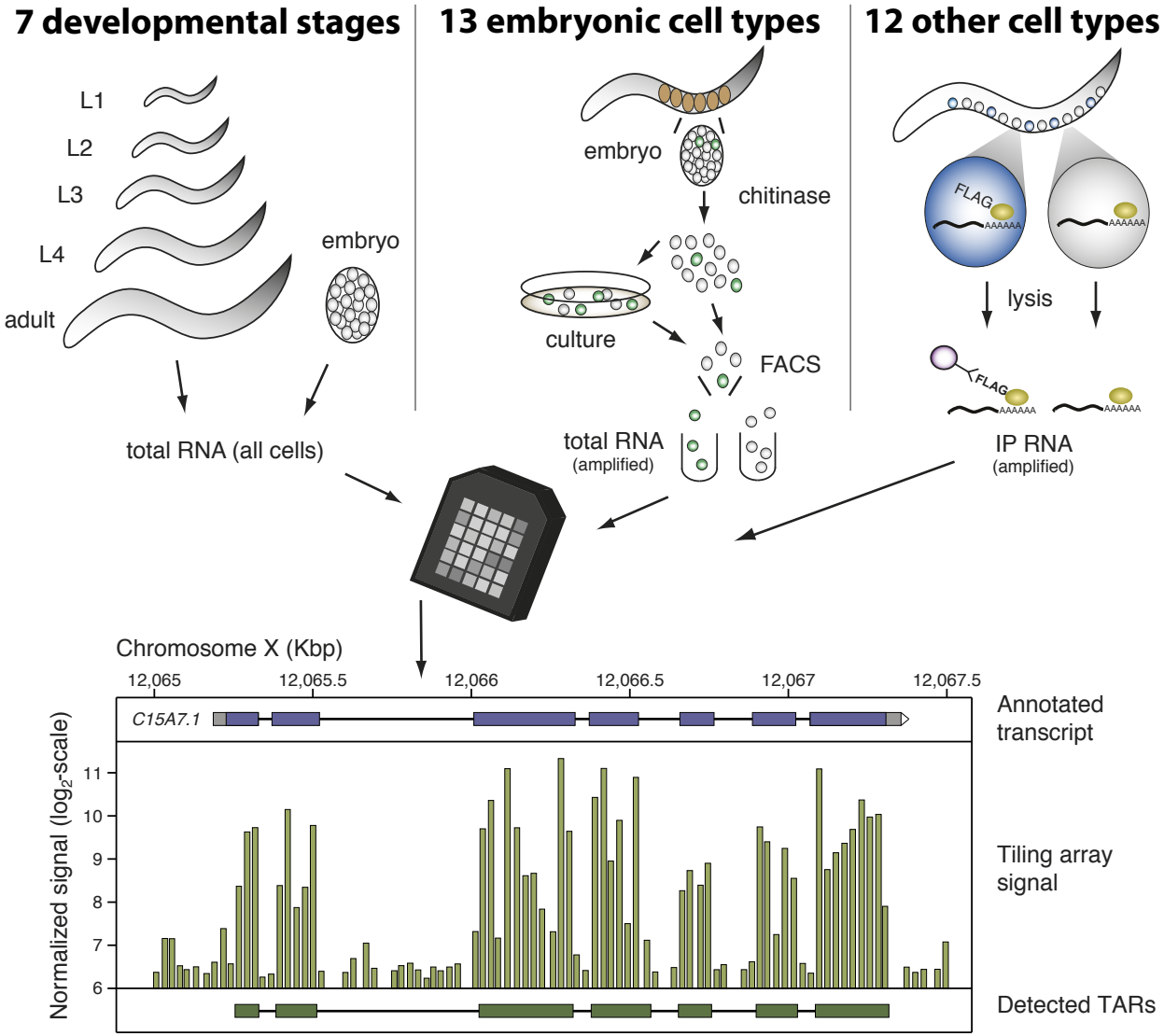
[Spencer et al., 2011]

Analysis of C. elegans Transcriptomes



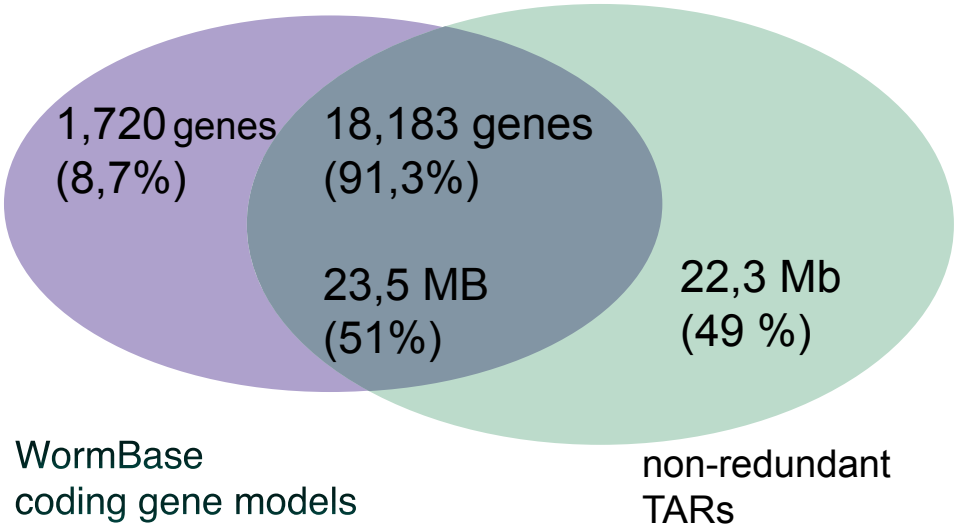
[Spencer et al., 2011]

Analysis of C. elegans Transcriptomes



[Spencer et al., 2011]

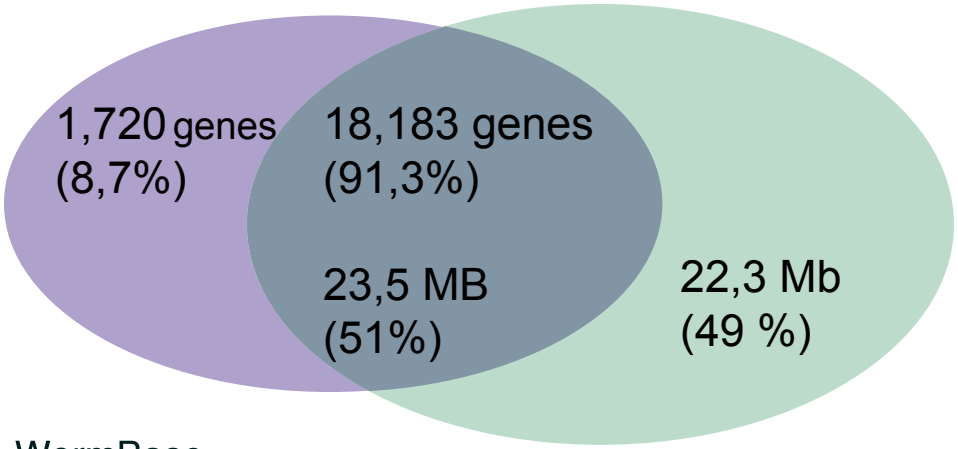
Identification of novel TARs



Detected TARs:

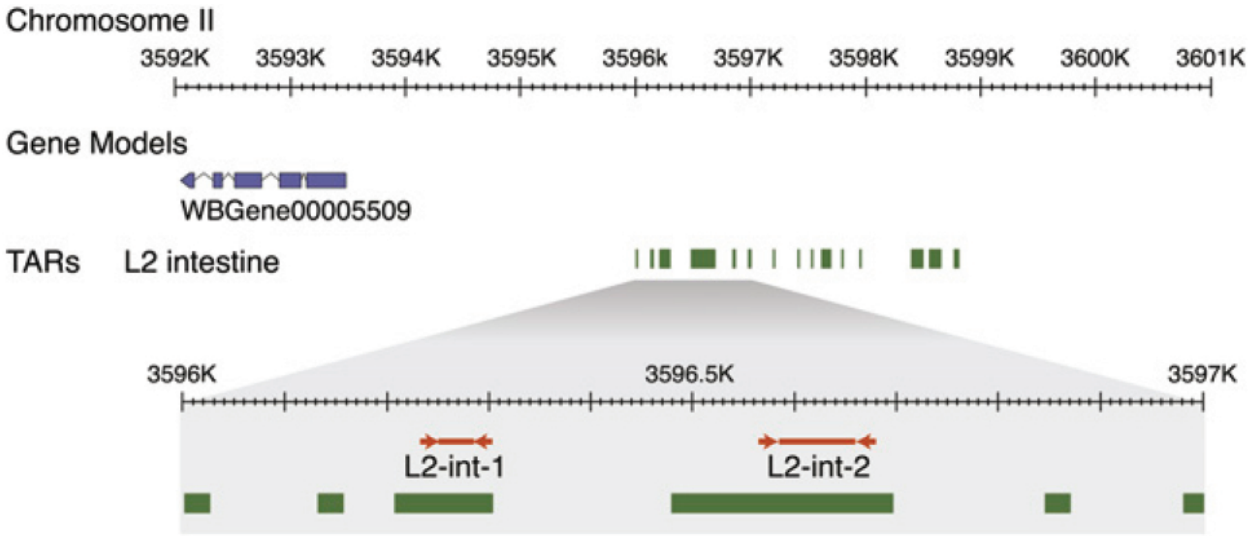
- Overlap TARs with coding genes: $\approx 91\%$
- $\approx$ 22 Mb not in protein coding regions
- $\approx$ 14 Mb novel TARs

Identification of novel TARs

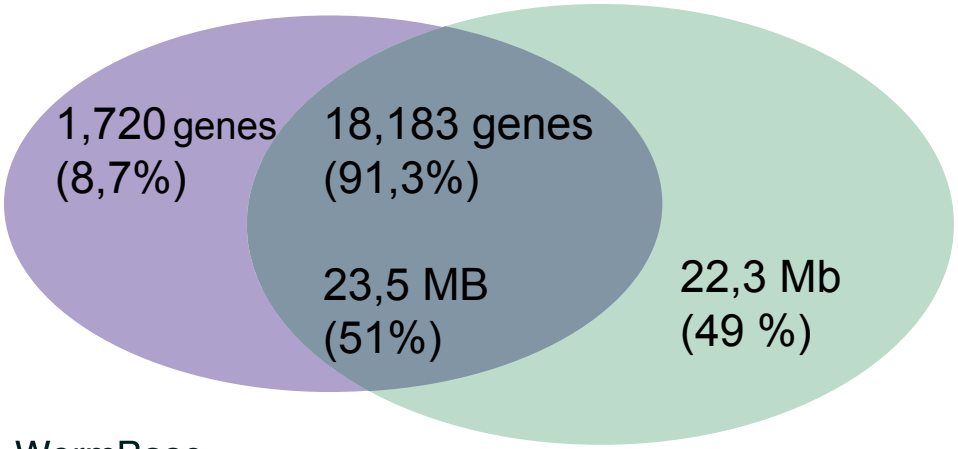


Detected TARs:

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

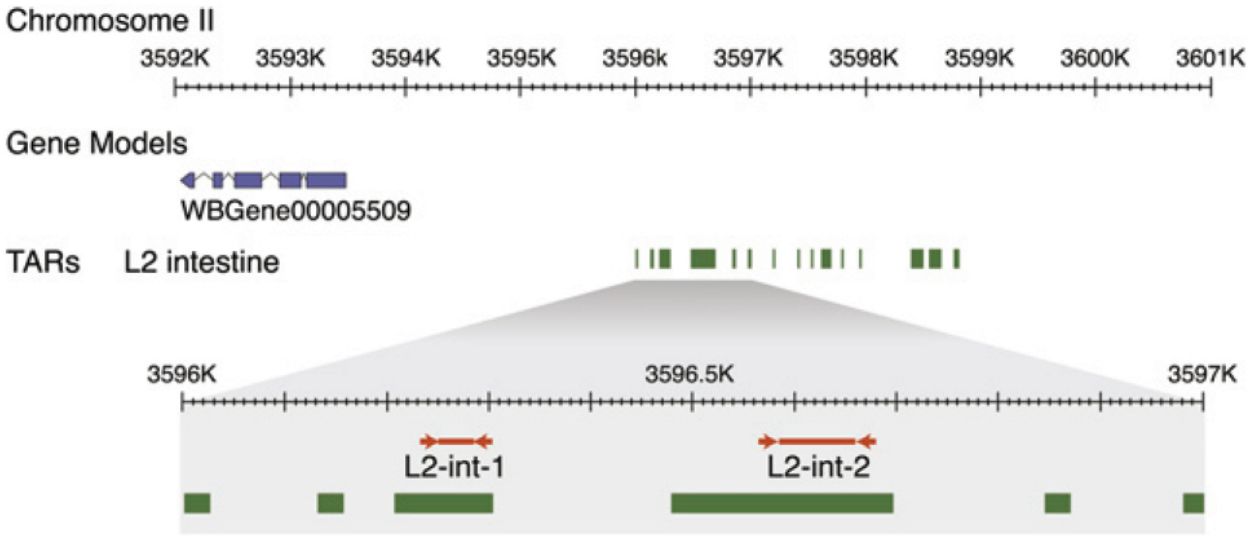


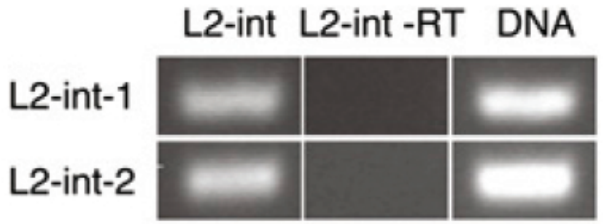
Identification of novel TARs



Detected TARs:

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





Understanding Global Expression Patterns

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

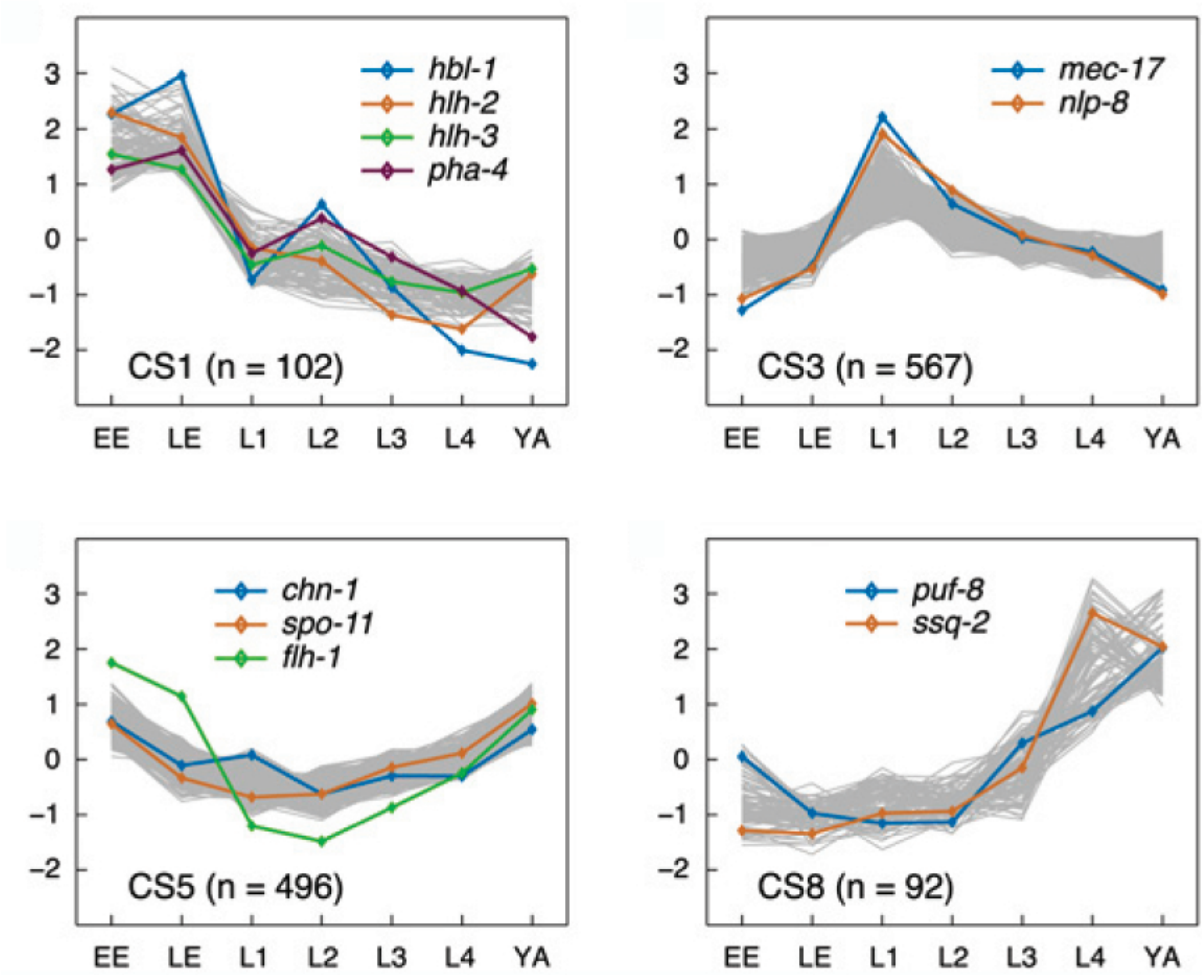
Group genes with similar expression profiles:

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:

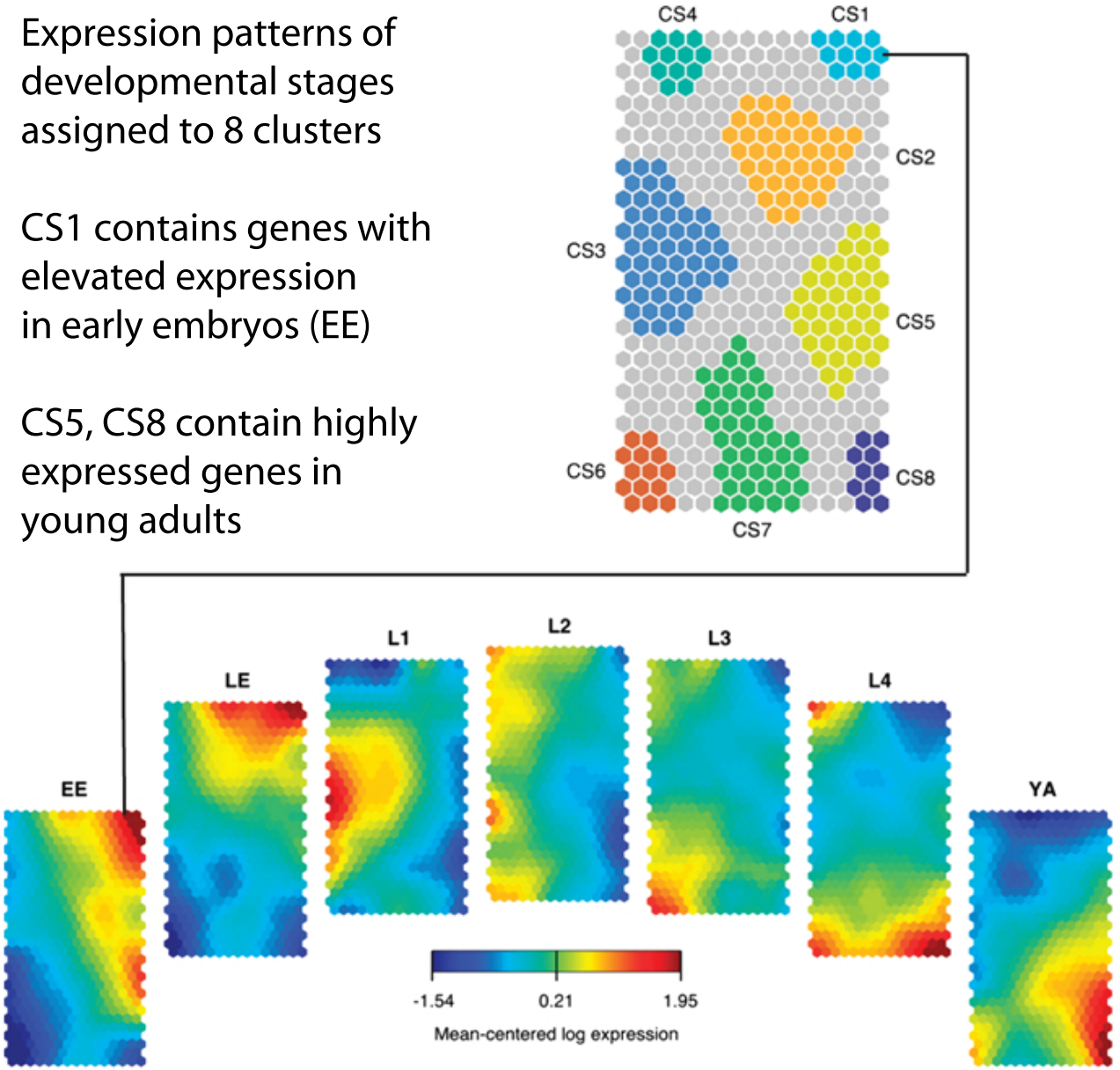
- Clustering via Self-Organizing Maps (SOMs)
- Analyse gene sets defined by clusters

Understanding Global Expression Patterns

Expression patterns of developmental stages assigned to 8 clusters

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

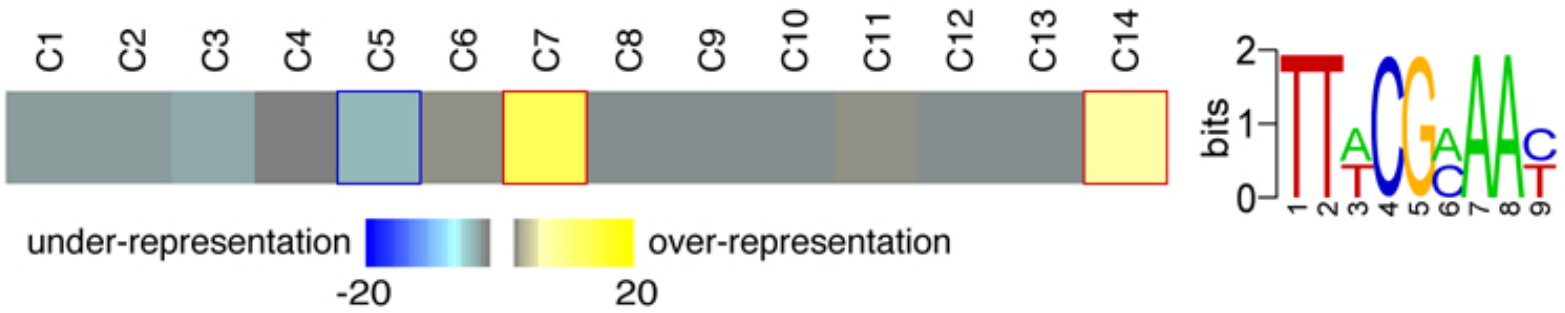
CS5, CS8 contain highly expressed genes in young adults



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)



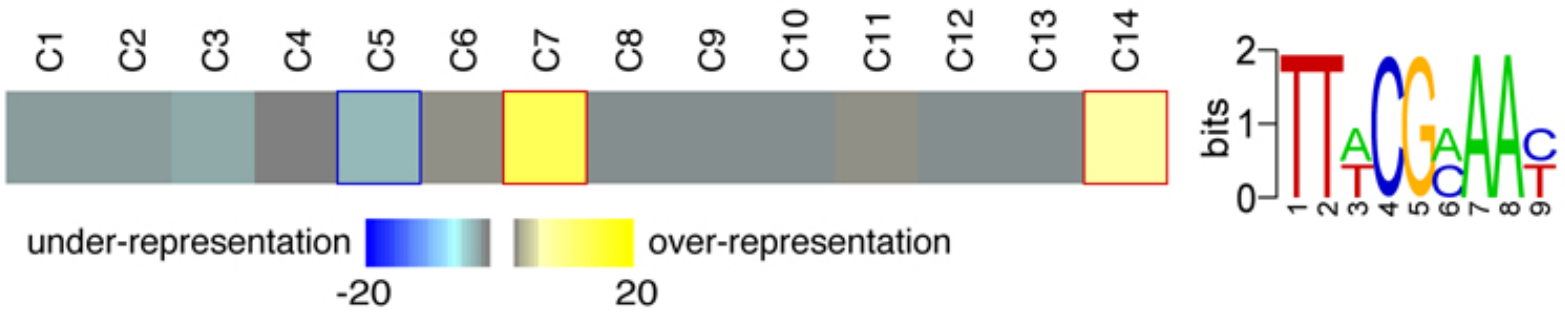
Next steps:

- Analyse sub-genic features (as for *A. thaliana*) [Eichner et al., 2011]
- Higher genomic resolution by RNA-seq (in prep.)
- Higher temporal & spatial resolution by single cell analysis

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)

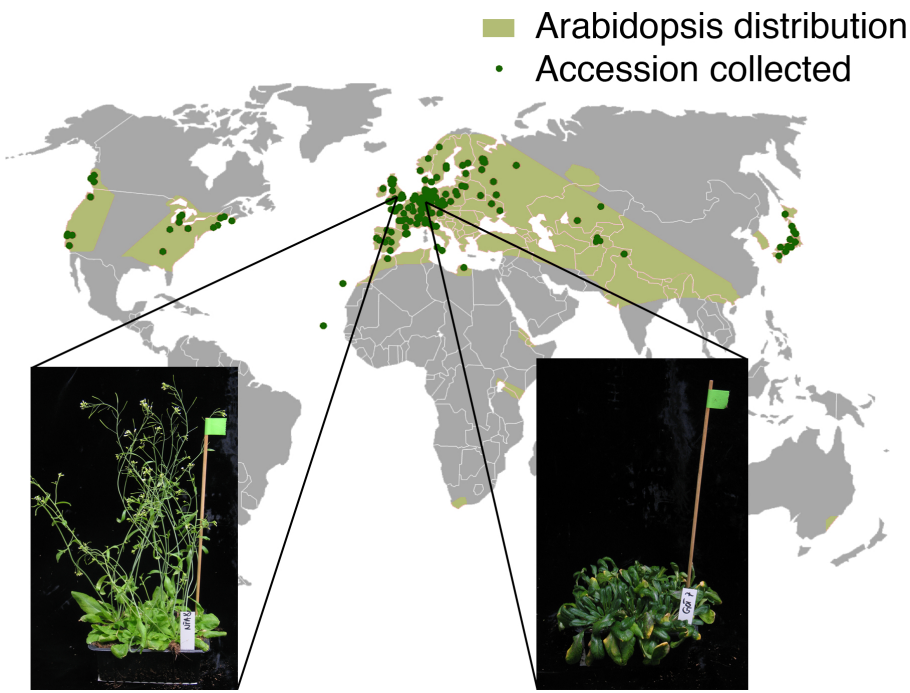


Next steps:

- Analyse sub-genic features (as for *A. thaliana*) [Eichner et al., 2011]
- Higher genomic resolution by RNA-seq (in prep.)
- Higher temporal & spatial resolution by single cell analysis

Multiple reference genomes and transcriptomes for A. thaliana

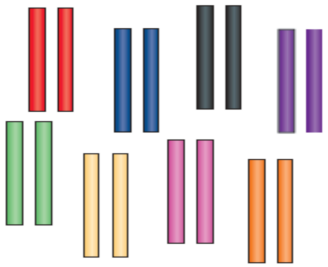
G. Xiangchao, et al., G. Rätsch, R. Mott, Nature (accepted in principle)



- 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



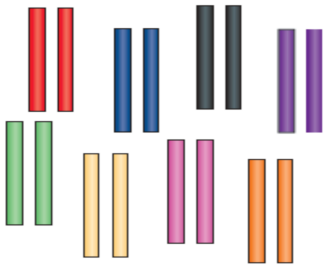
n founders

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

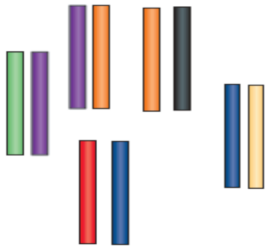
This talk: Genome and transcriptome analysis of the MAGIC lines founders

18 +1 reference genomes and transcriptomes for A. thaliana

Multiparent Advanced Generation Inter-Cross (MAGIC) lines



n founders



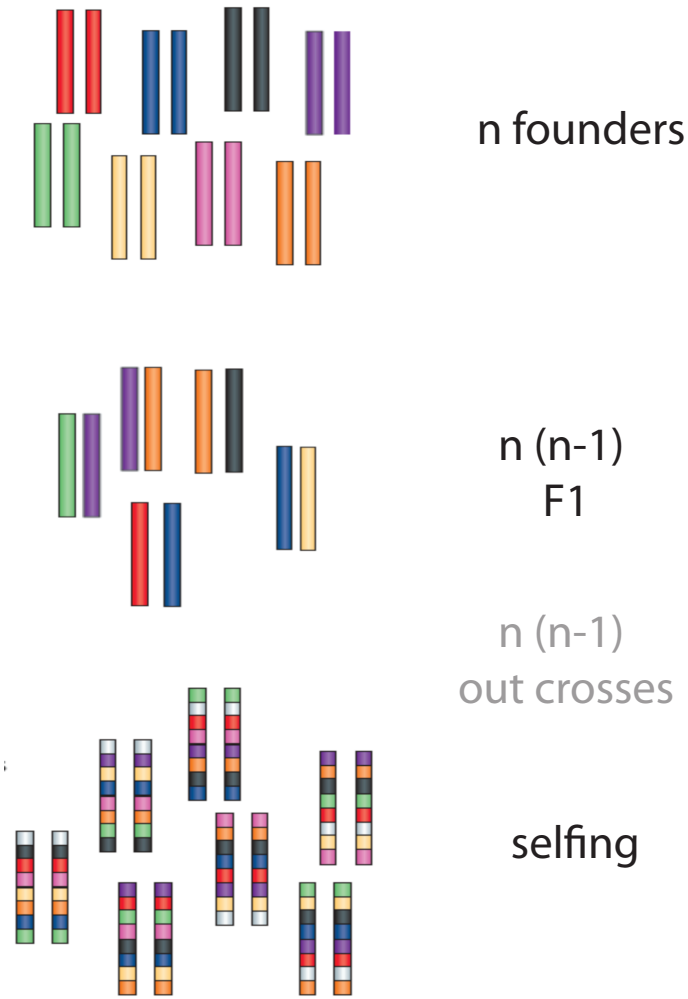
n (n-1)
F1

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

This talk: 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

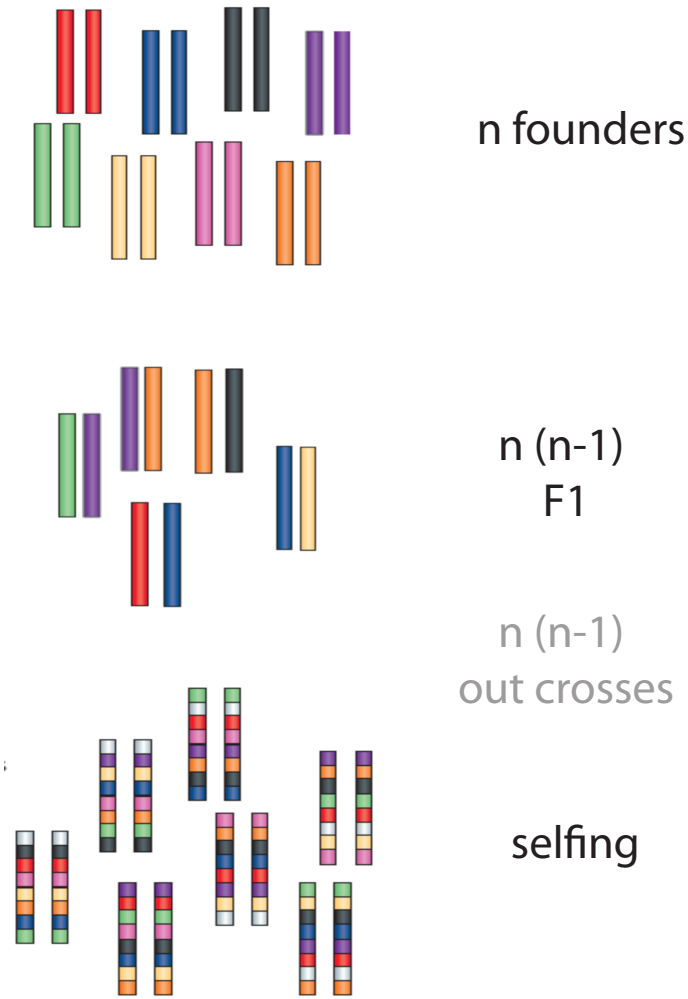


- Eventually 1200 lines
- Imputation of magic lines genomes based on founders
- Powerful QTL mapping population

This talk: 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

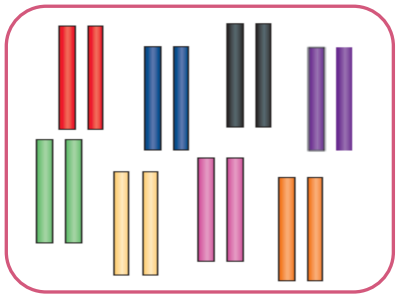


- Eventually 1200 lines
- Imputation of magic lines genomes based on founders
- Powerful QTL mapping population

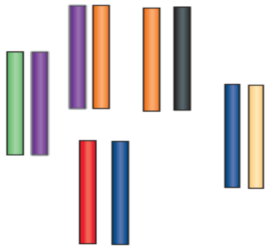
This talk: Genome and transcriptome analysis of the MAGIC lines founders

18 +1 reference genomes and transcriptomes for A. thaliana

Multiparent Advanced Generation Inter-Cross (MAGIC) lines

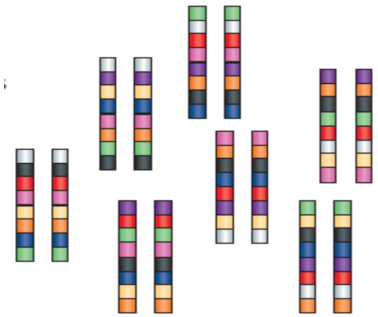


n founders



n (n-1)
F1

n (n-1)
out crosses

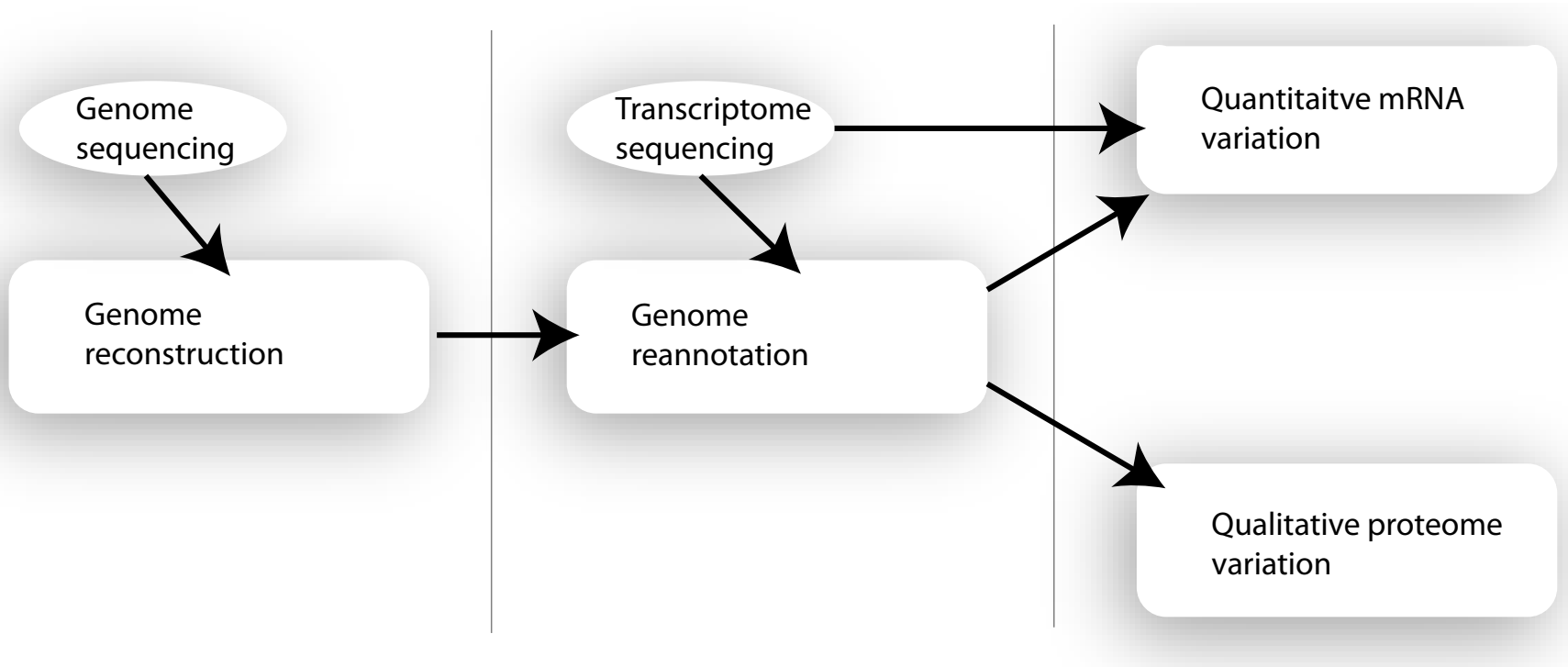


selfing

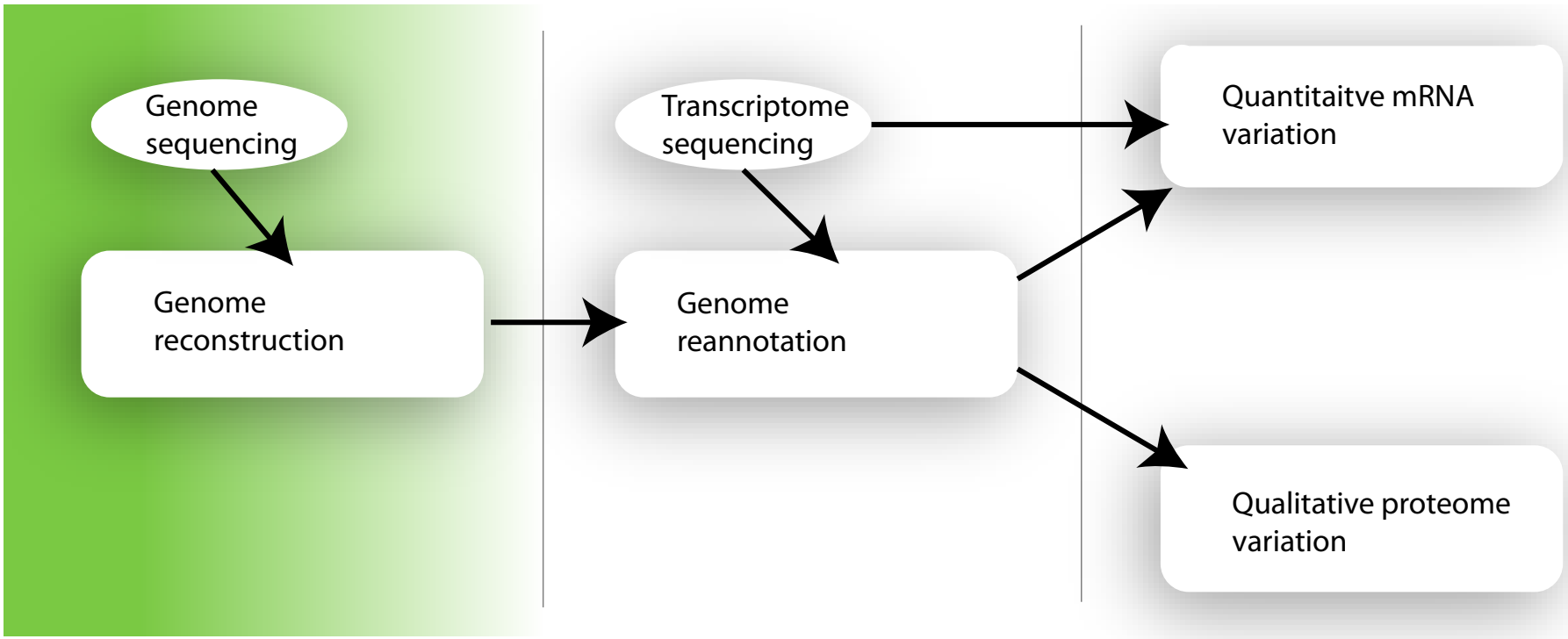
- Eventually 1200 lines
- Imputation of magic lines genomes based on founders
- Powerful QTL mapping population

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

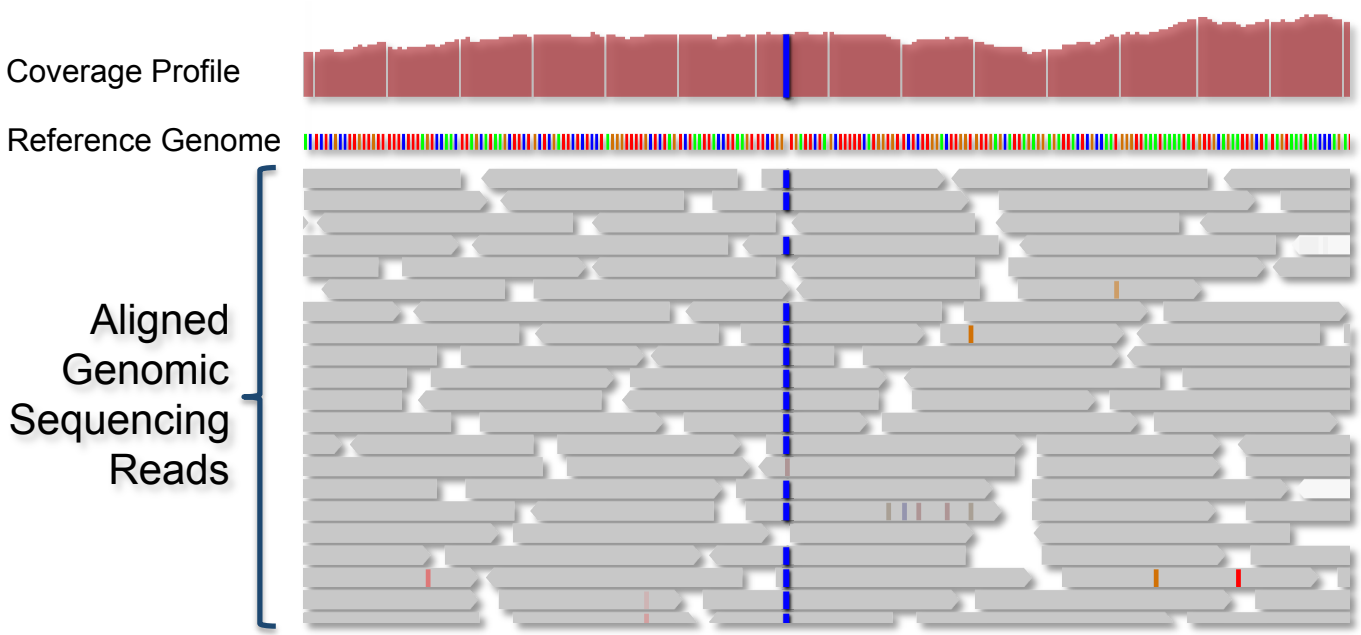
Outline


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

Outline


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

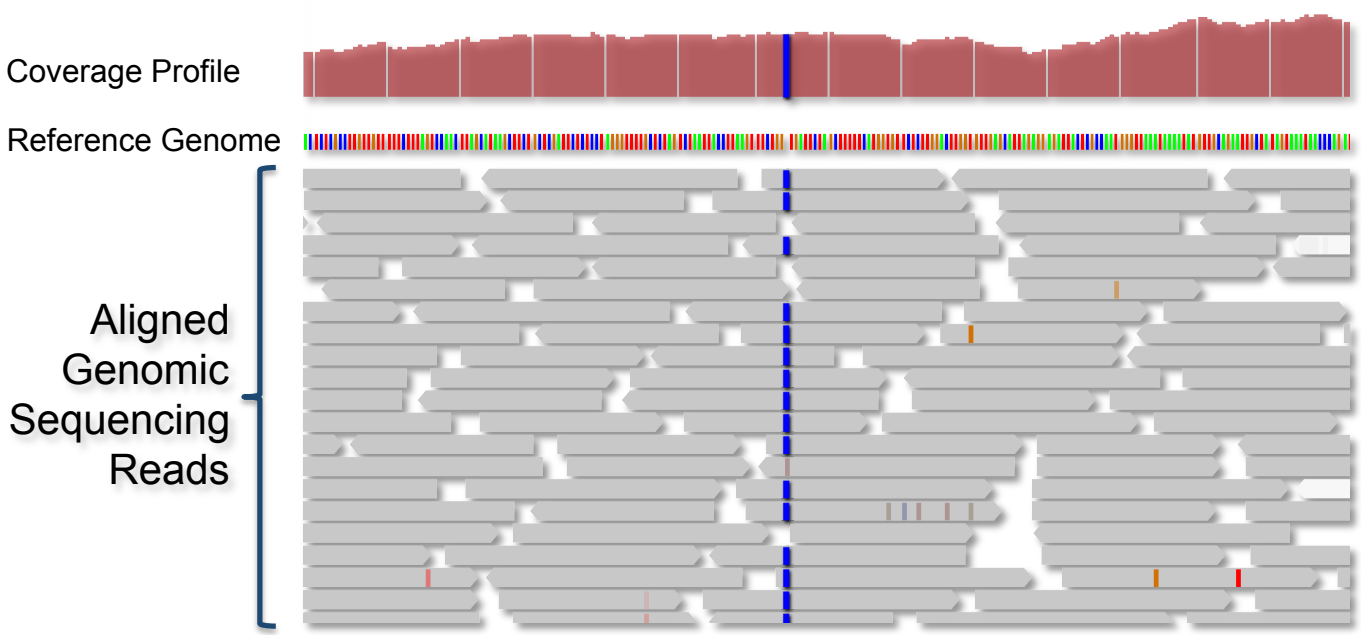
Sequencing 18 Arabidopsis genomes



- ≈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

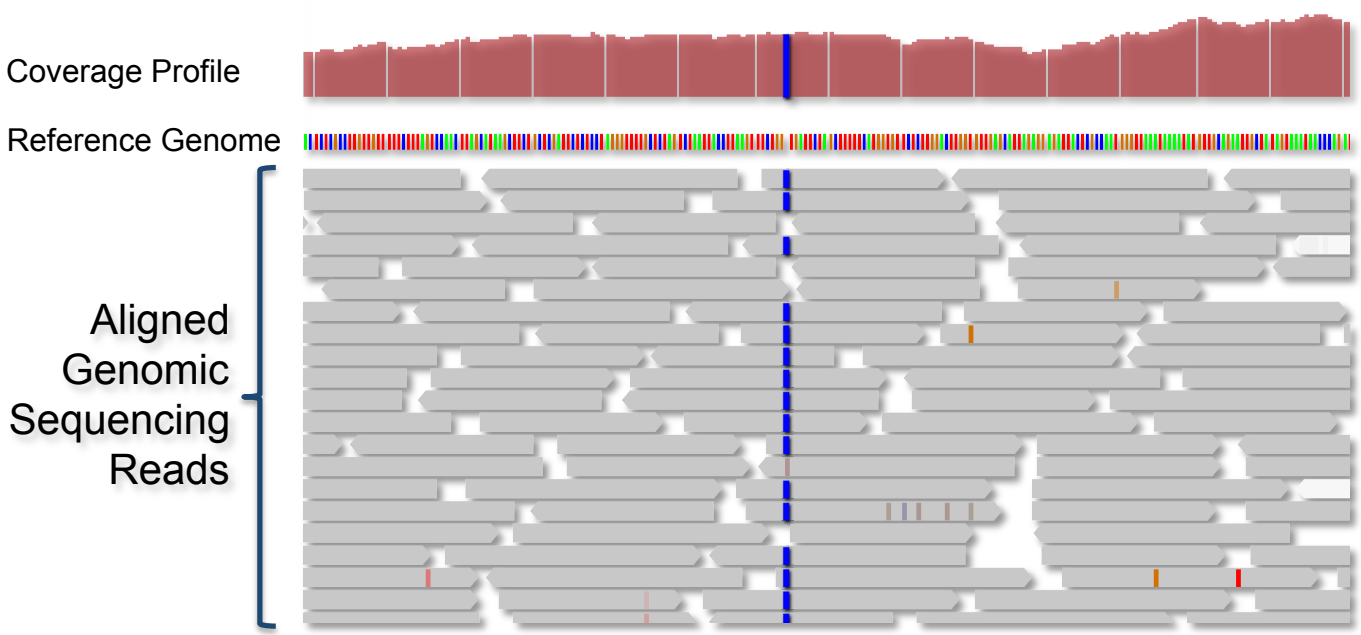
Sequencing 18 Arabidopsis genomes



- ≈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

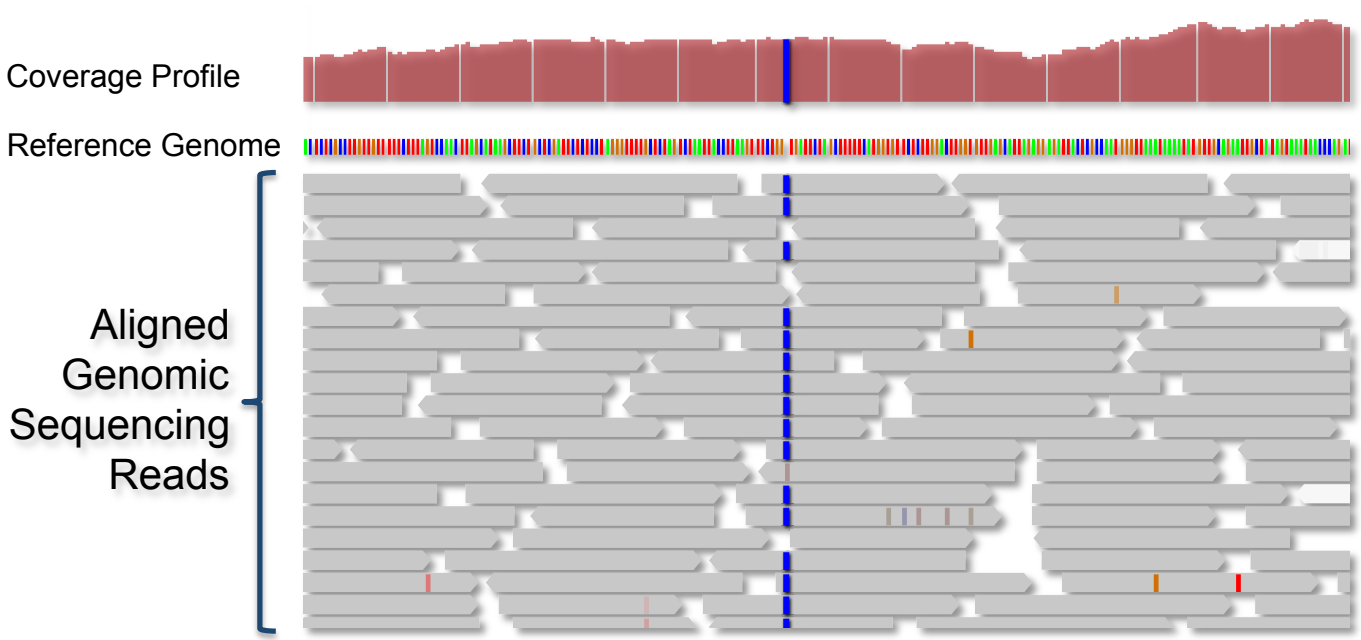
Sequencing 18 Arabidopsis genomes



- ≈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

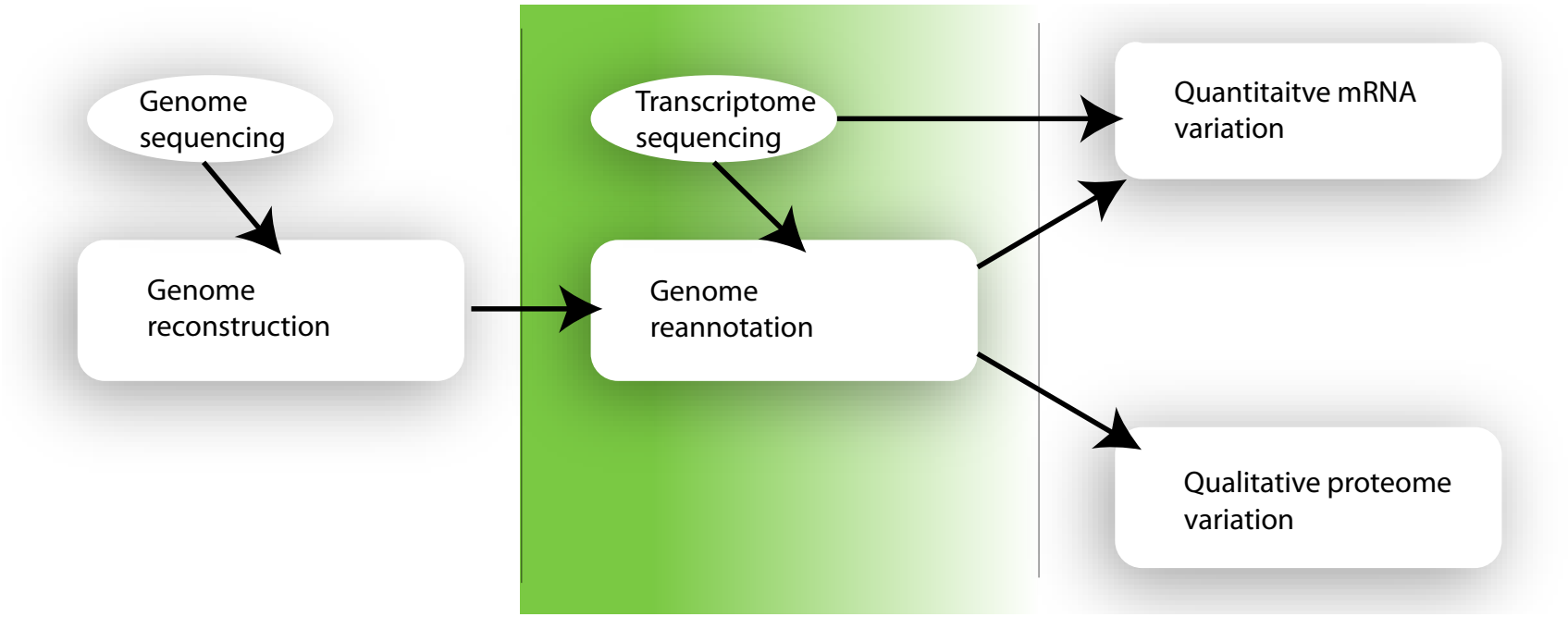
Sequencing 18 Arabidopsis genomes



- ≈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

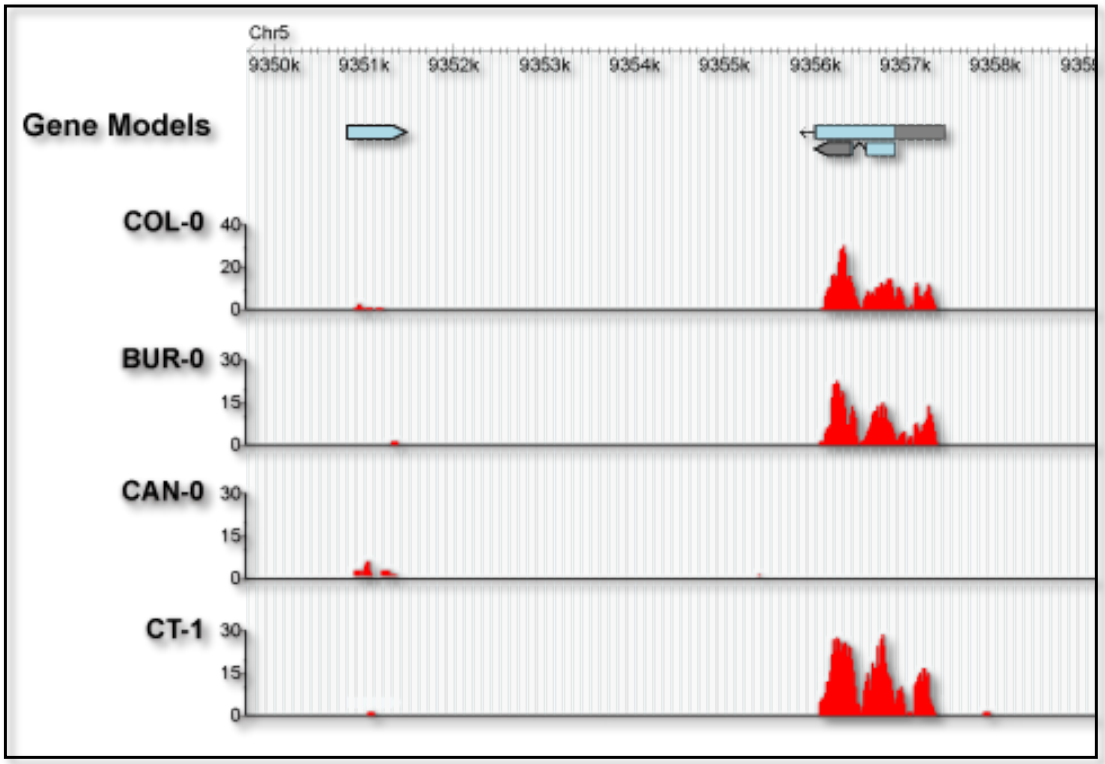
Outline


Genome sequencing
Genome reconstruction
Transcriptome sequencing
Genome reannotation
Quantitaitve mRNA variation
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

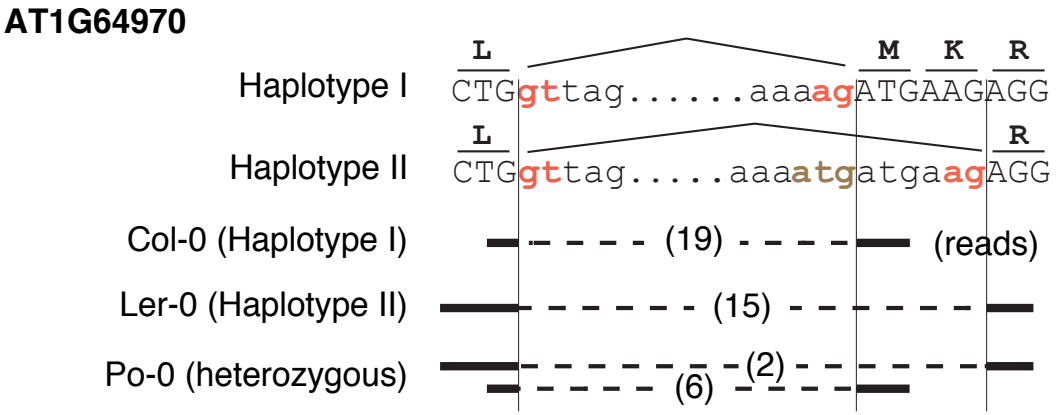


- Validation of genome sequence
- Genome reannotation

Transcriptome Sequencing

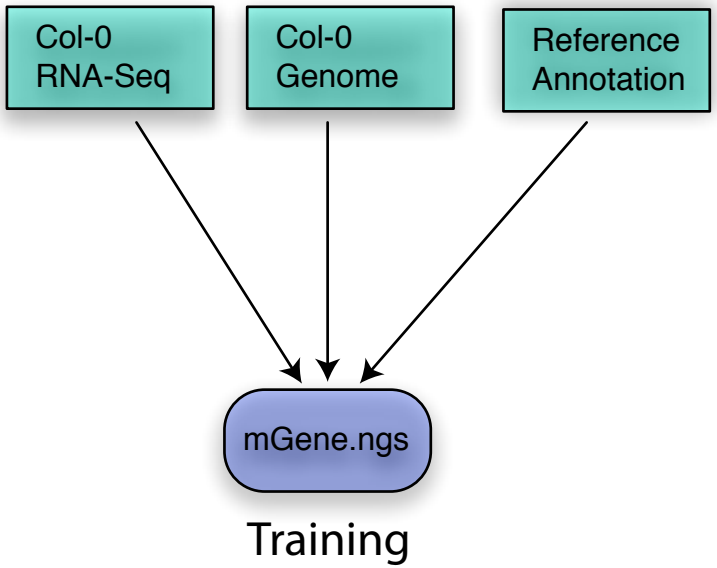
RNA-seq (Illumina, 76bp) of founder accessions:

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



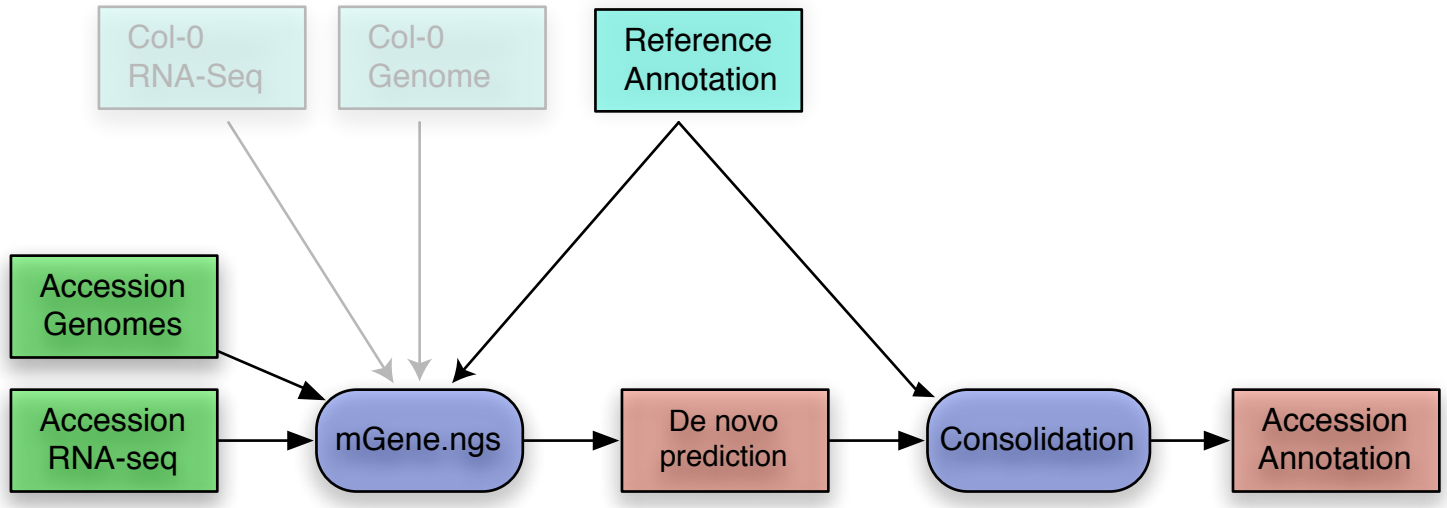
- Validation of genome sequence
- **Genome reannotation**

Genome Reannotation



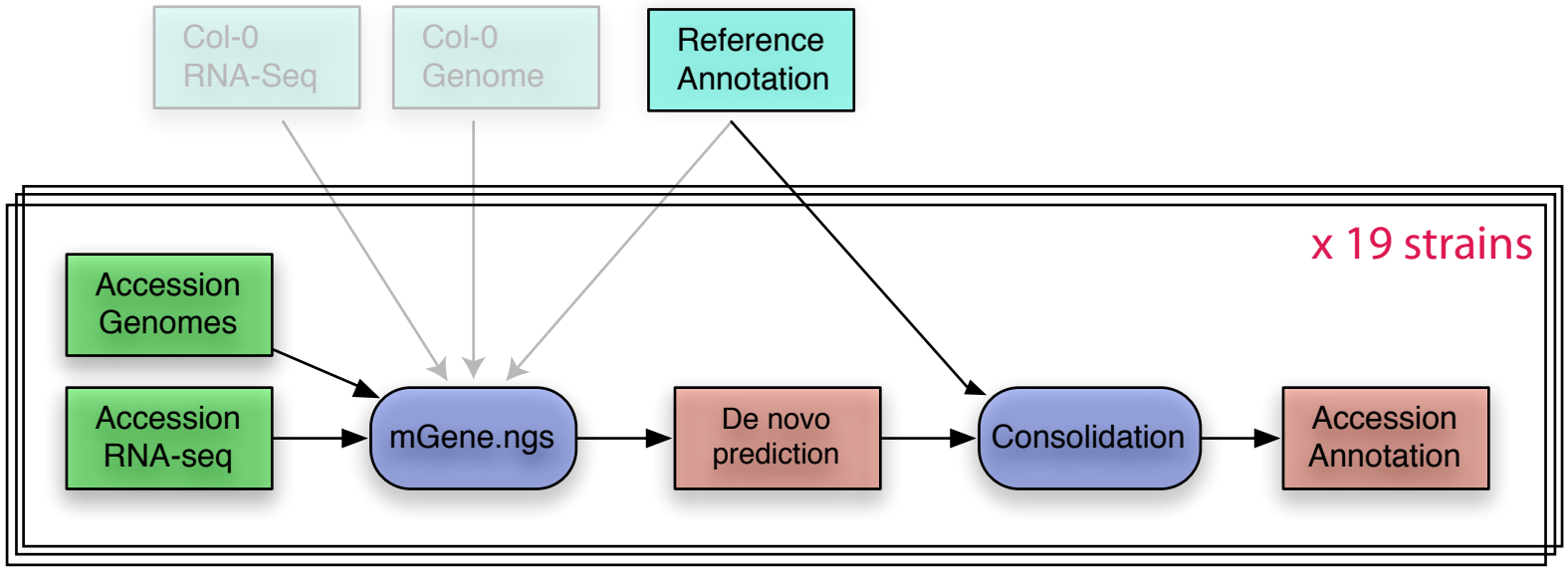
- Training on col-0 reference
- Strain-specific predictions, accounting for strain coordinate system
- $>$ 8,000 genes with new transcript annotations
- Per accession: $\approx$ 1,500 new transcripts, $\approx$ 650 new introns, $\approx$ 500 new TIS/STOP sites

Genome Reannotation



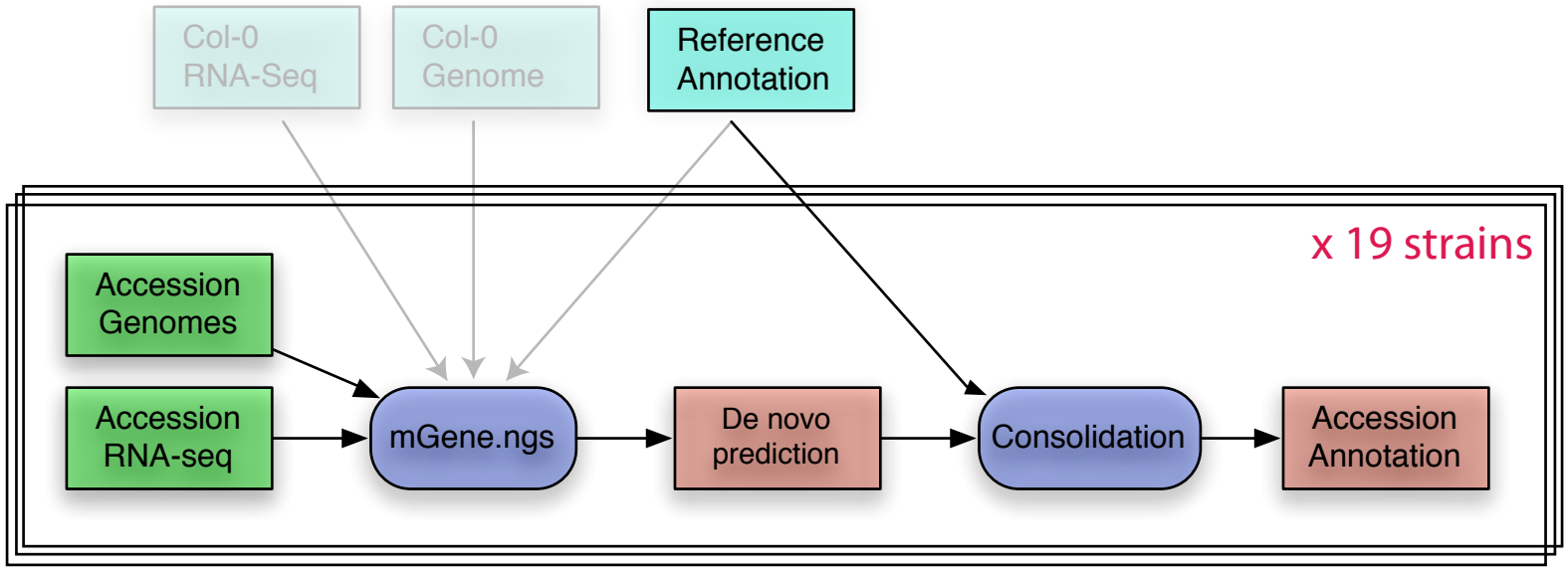
- Training on col-0 reference
- Strain-specific predictions, accounting for strain coordinate system
- $>$ 8,000 genes with new transcript annotations
- Per accession: $\approx$ 1,500 new transcripts, $\approx$ 650 new introns, $\approx$ 500 new TIS/STOP sites

Genome Reannotation



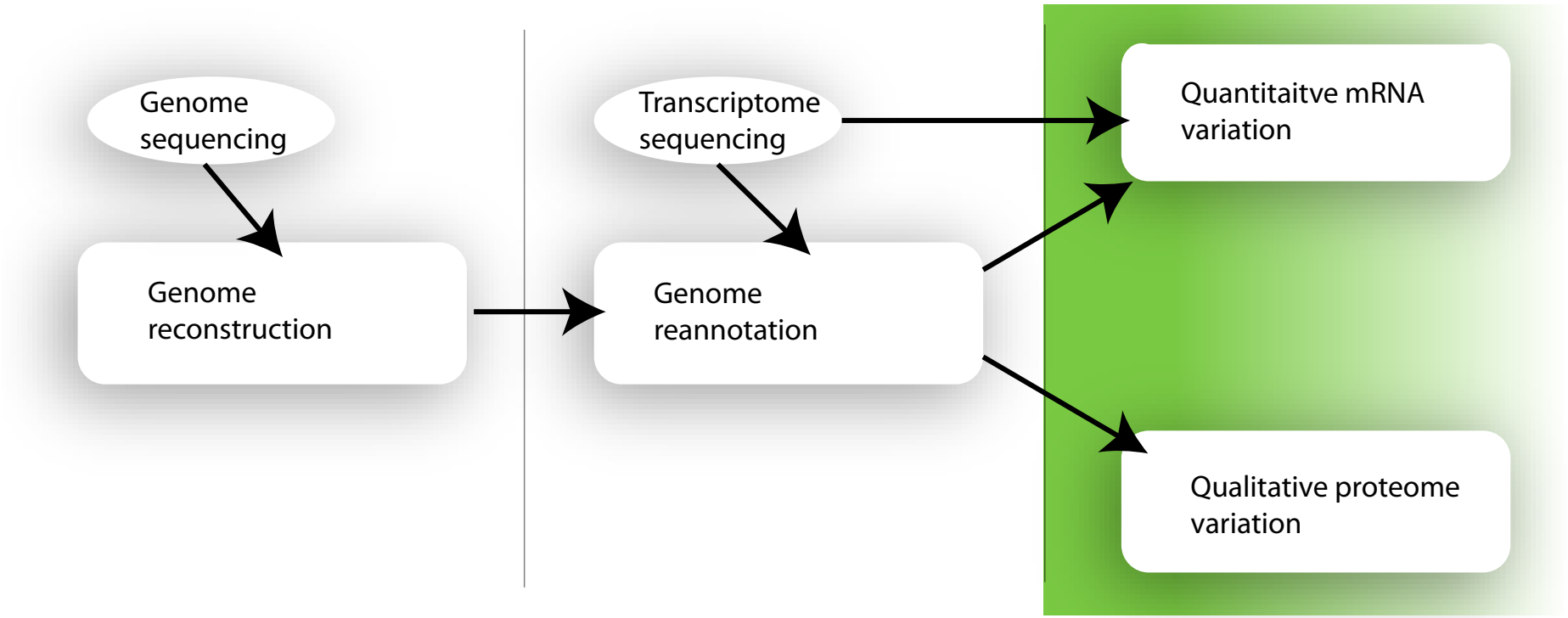
- Training on col-0 reference
- Strain-specific predictions, accounting for strain coordinate system
- $>$ 8,000 genes with new transcript annotations
- Per accession: $\approx$ 1,500 new transcripts, $\approx$ 650 new introns, $\approx$ 500 new TIS/STOP sites

Genome Reannotation



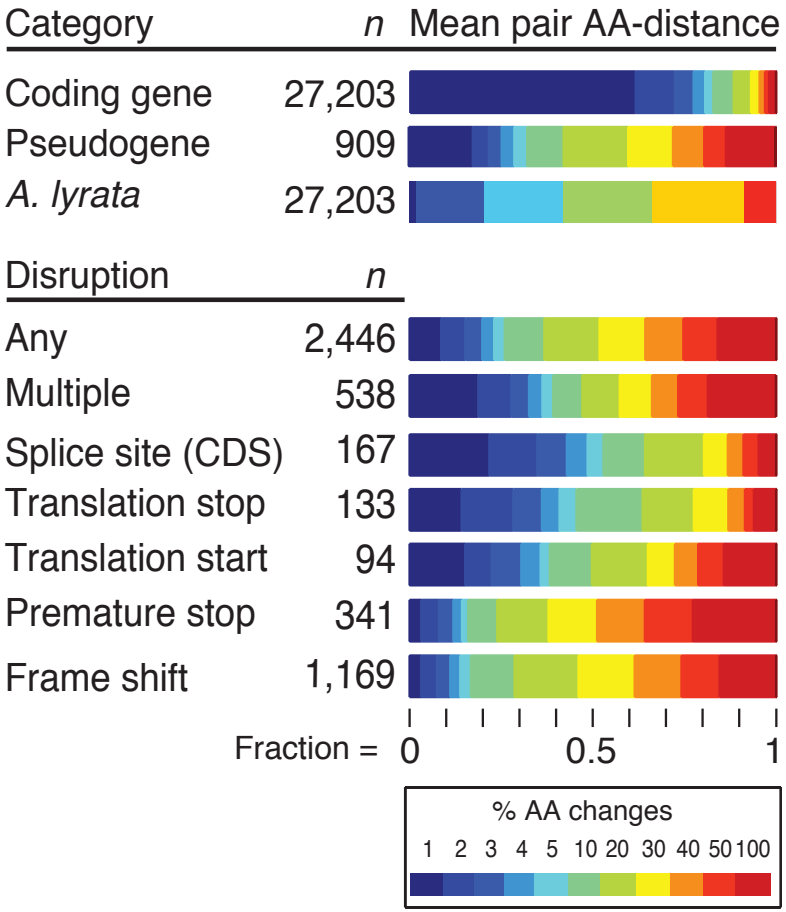
- Training on col-0 reference
- Strain-specific predictions, accounting for strain coordinate system
- $>$ 8,000 genes with new transcript annotations
- Per accession: $\approx$ 1,500 new transcripts, $\approx$ 650 new introns, $\approx$ 500 new TIS/STOP sites

Outline


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

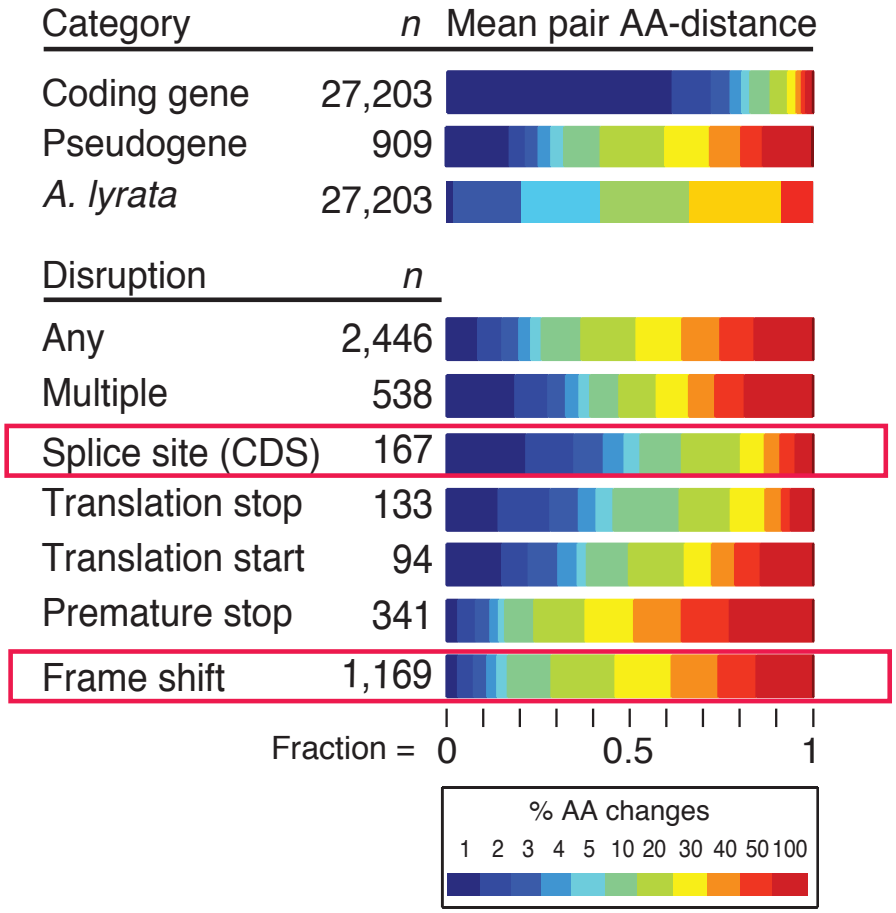
Amino acid variability

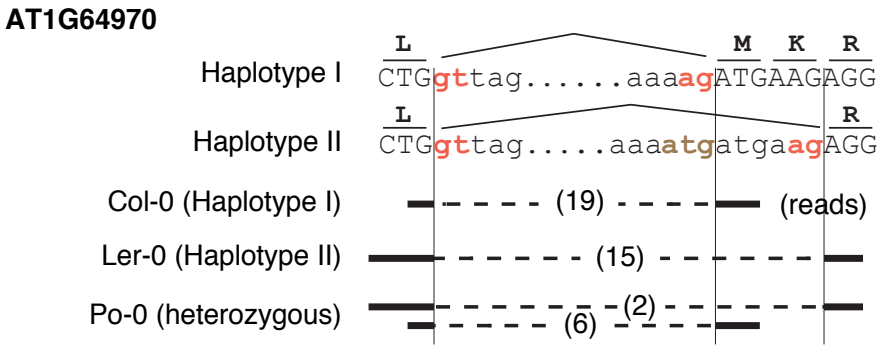
Large effects changes



Amino acid variability

Large effects changes

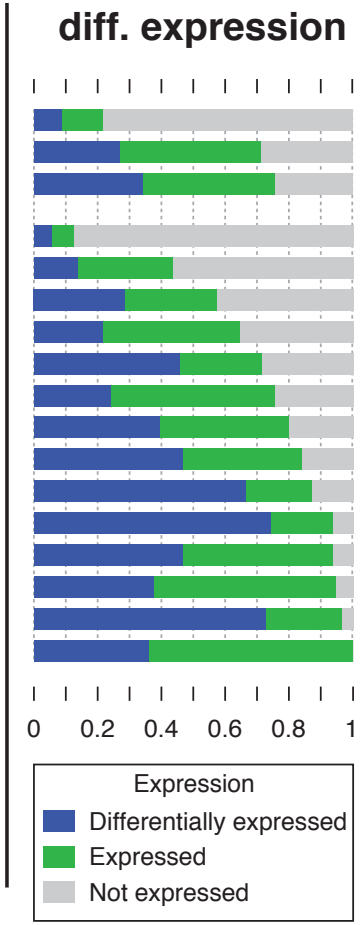




Gene expression variability

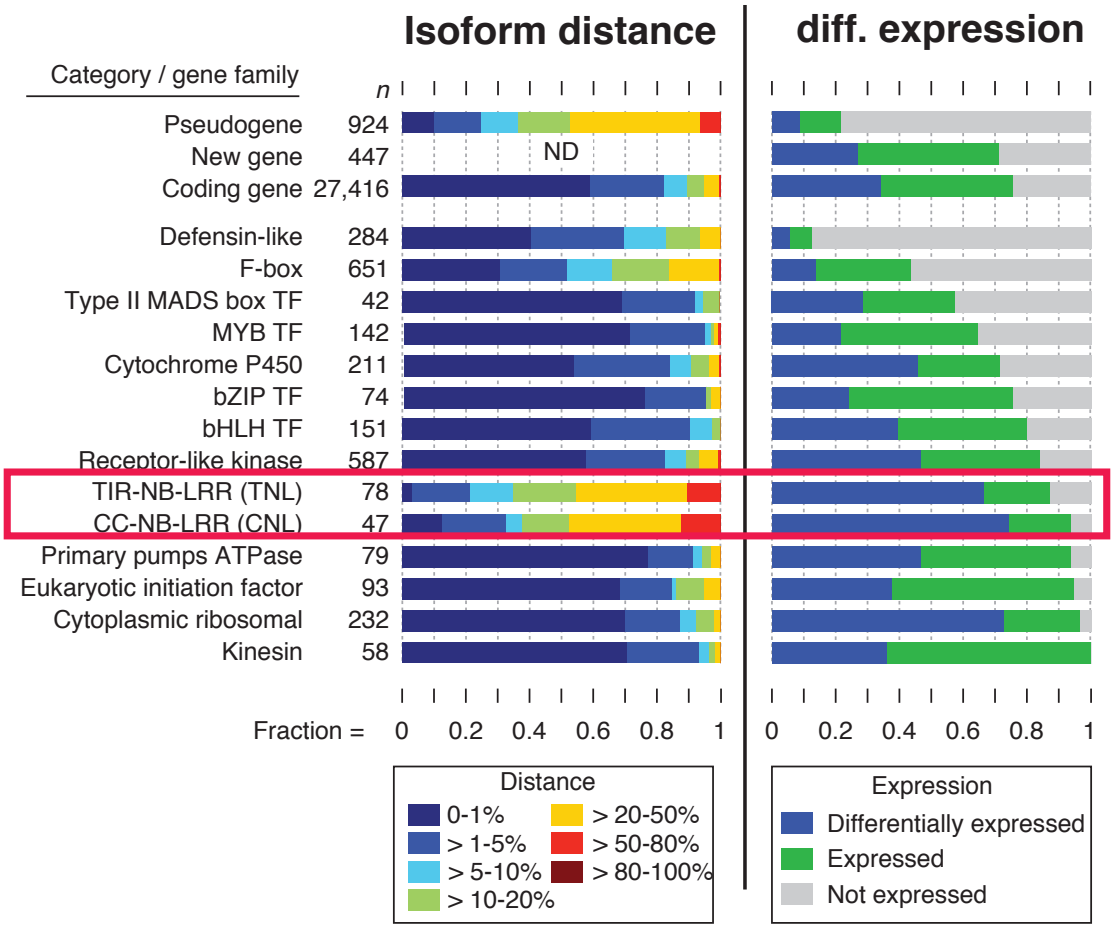
Category / gene family	*n*
Pseudogene	924
New gene	447
Coding gene	27,416
Defensin-like	284
F-box	651
Type II MADS box TF	42
MYB TF	142
Cytochrome P450	211
bZIP TF	74
bHLH TF	151
Receptor-like kinase	587
TIR-NB-LRR (TNL)	78
CC-NB-LRR (CNL)	47
Primary pumps ATPase	79
Eukaryotic initiation factor	93
Cytoplasmic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

Fraction =



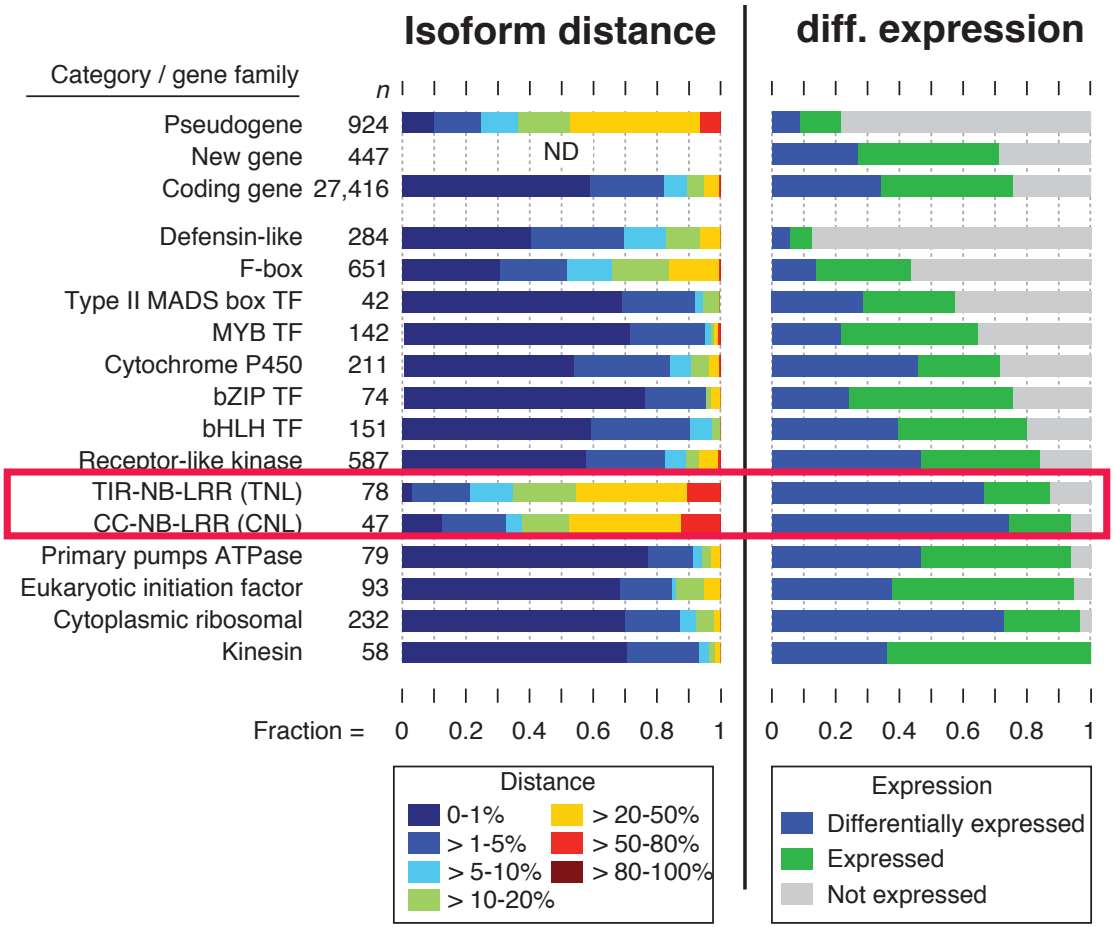
- ≈19,000 genes with detectable expression
  - 9,360 differentially expressed genes
  - 1,712 with fold-change ≥ 10
  - 308 with fold-change ≥ 50
- Causes of differential expression?

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?

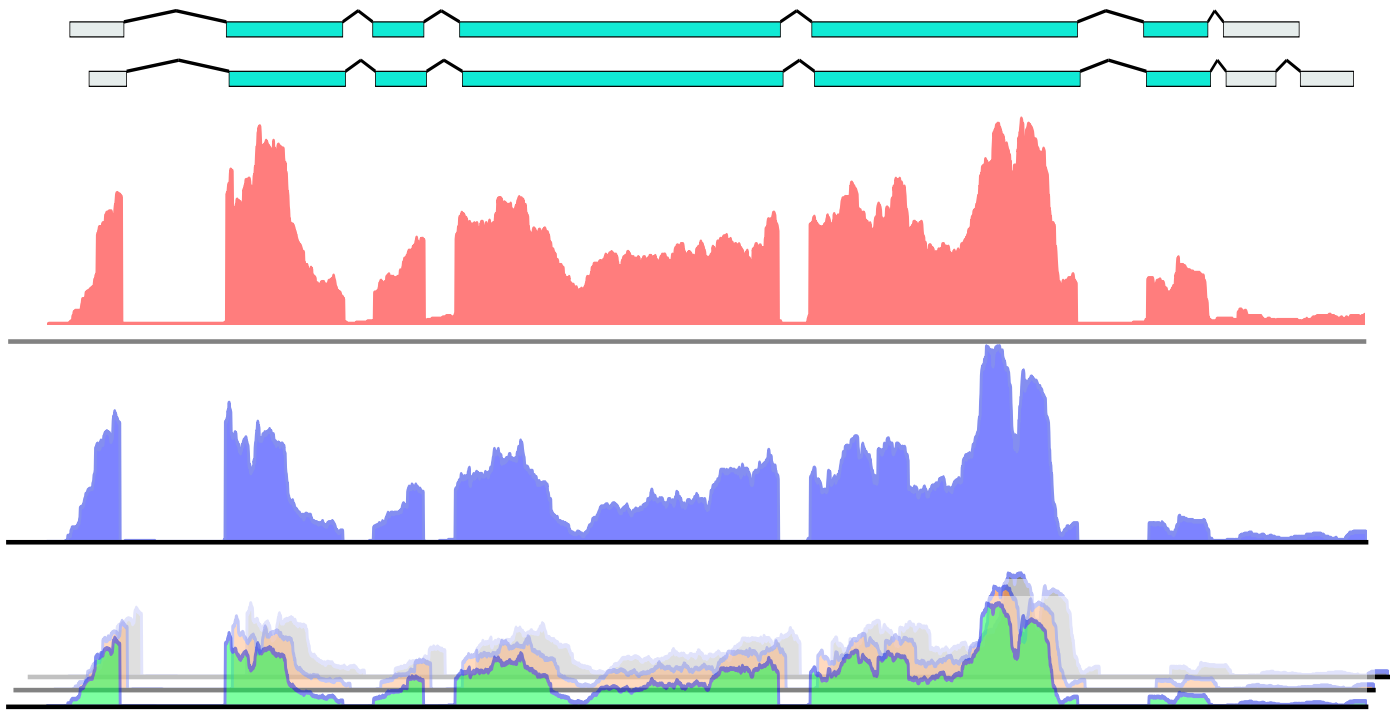
Gene expression variability



- ≈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

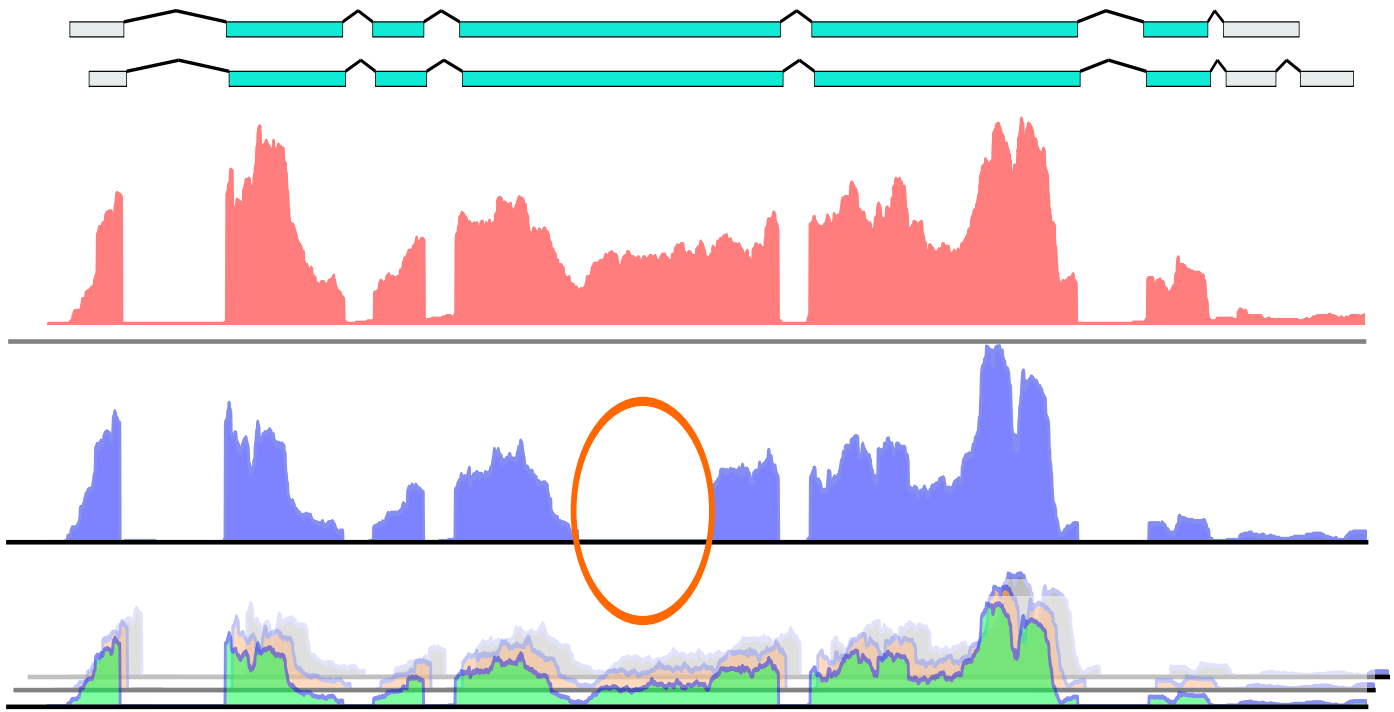
Structural Variation



[DESeq, Anders 2010 rDiff, Stegle & Drewe et al. 2010]

Gene expression variability

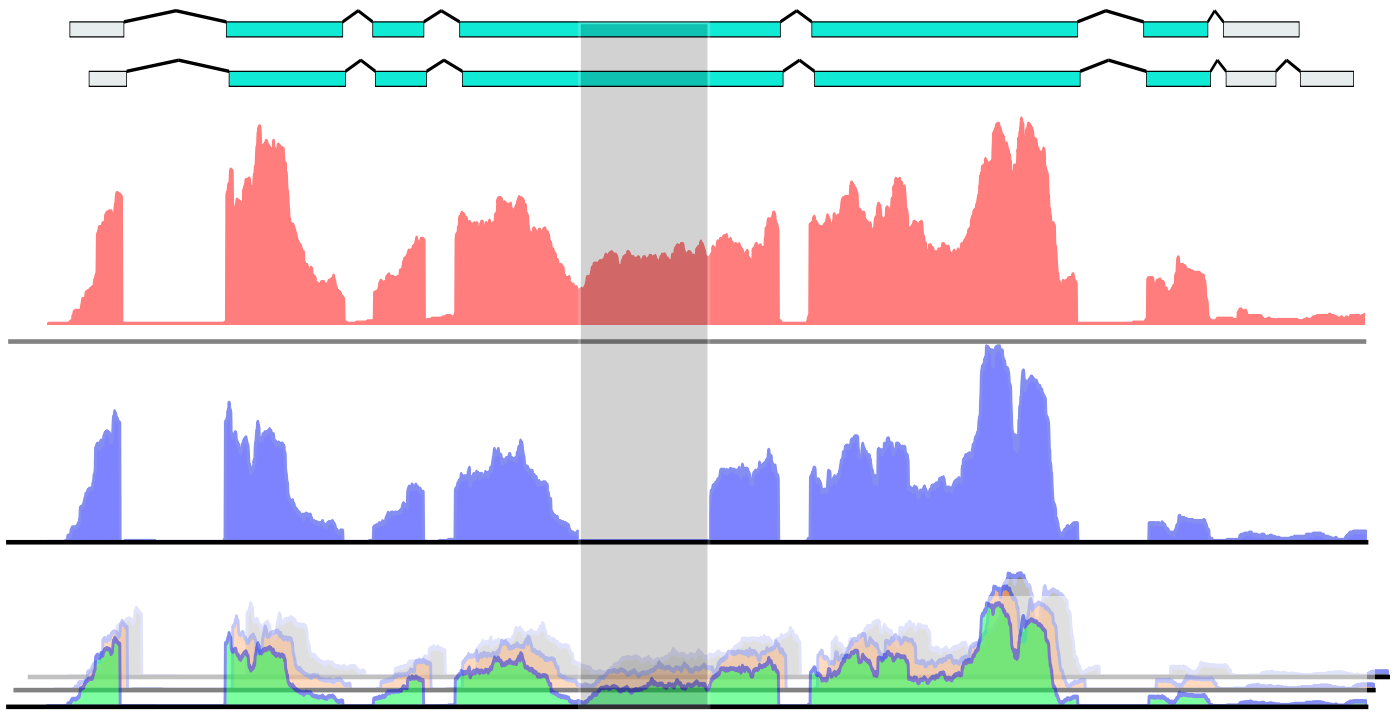
Structural Variation



Strain-specific deletion (genome reconstruction)

Gene expression variability

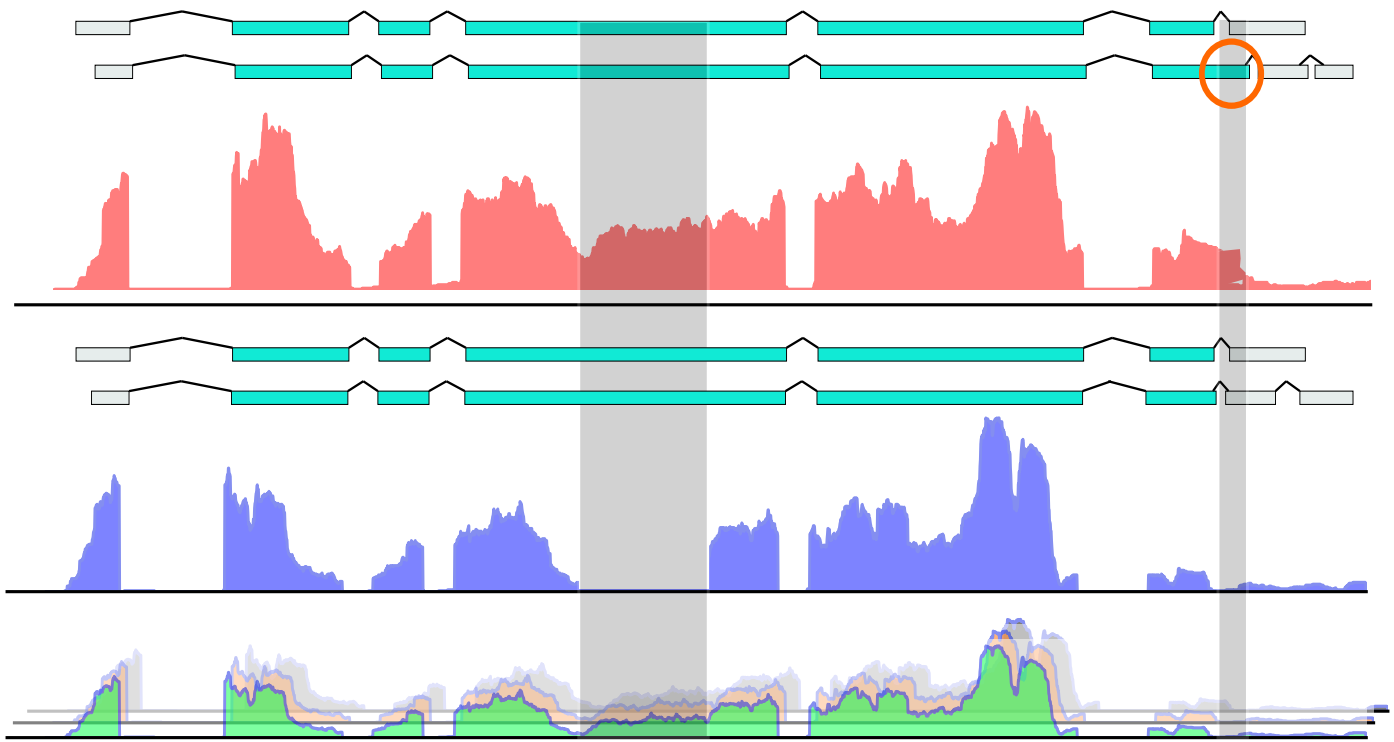
Structural Variation



Strain-specific deletion (genome reconstruction)

Gene expression variability

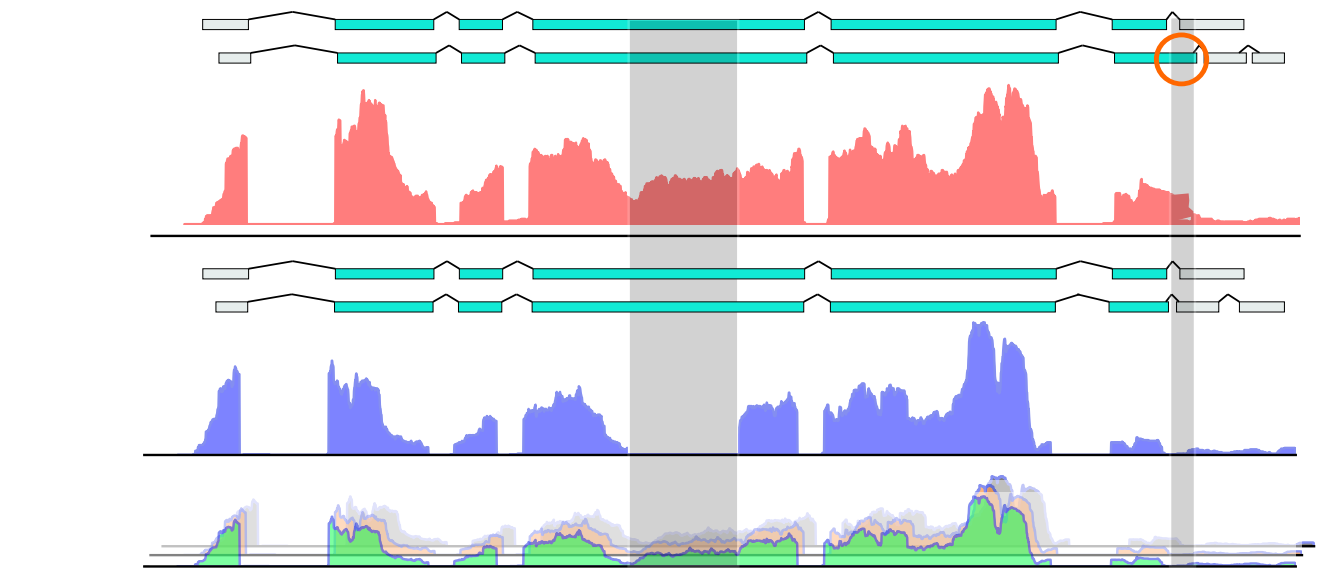
Structural Variation

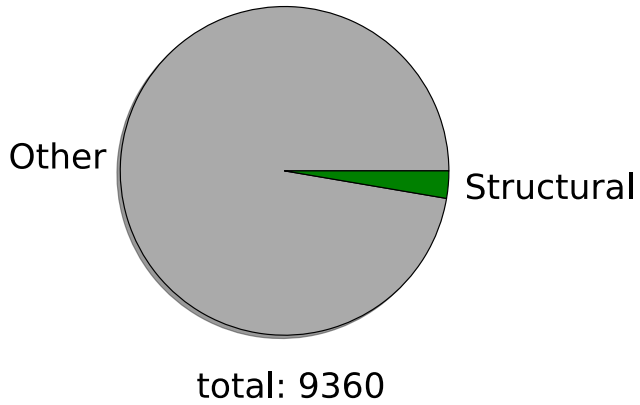


Strain-specific gene model change (genome reannotation)

Gene expression variability

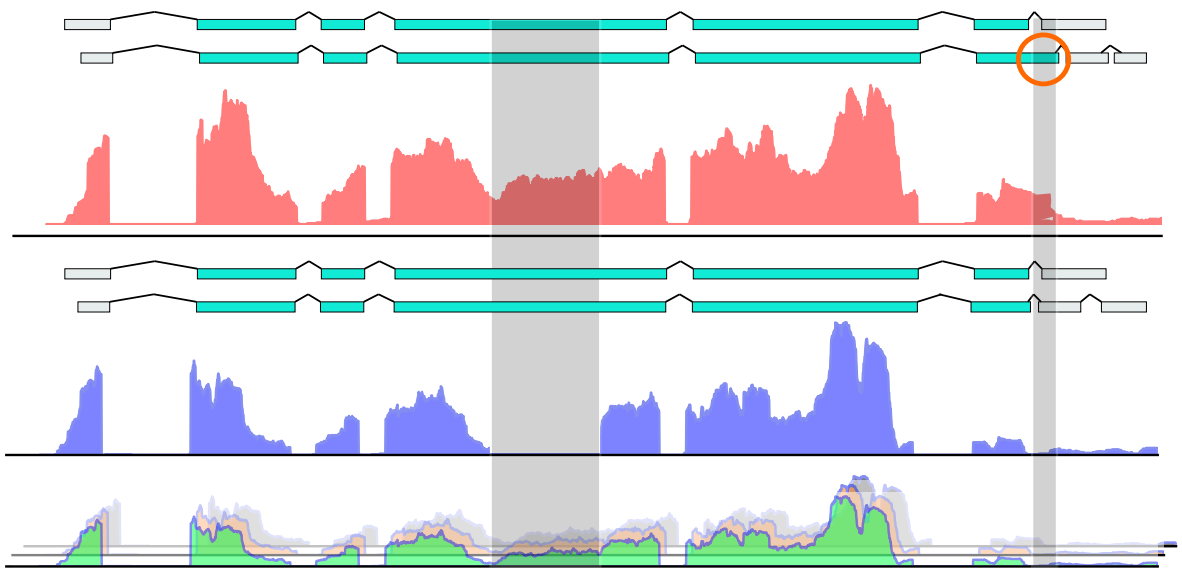
Structural Variation

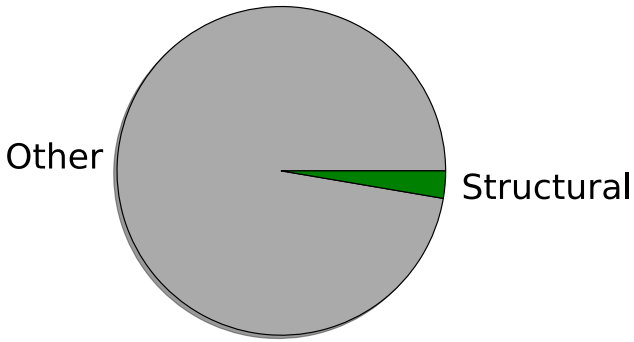




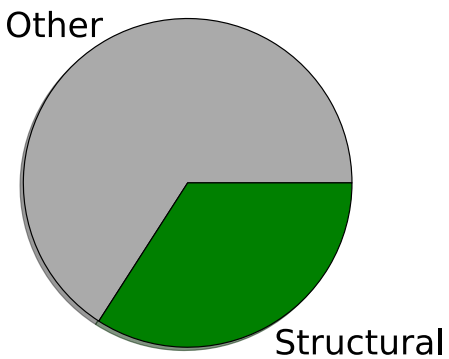
Gene expression variability

Structural Variation





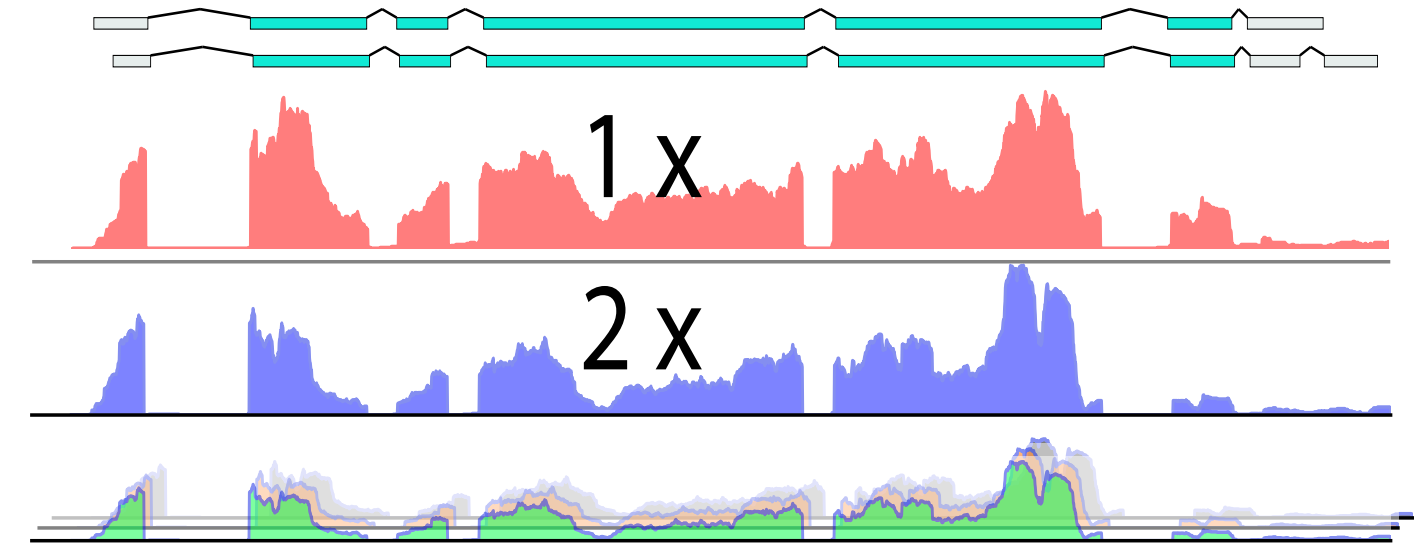
total: 9360

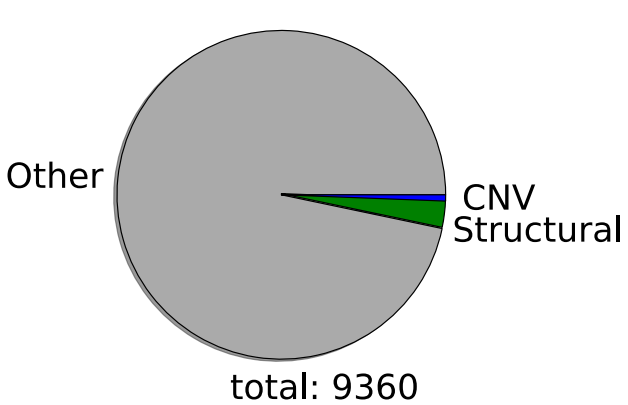


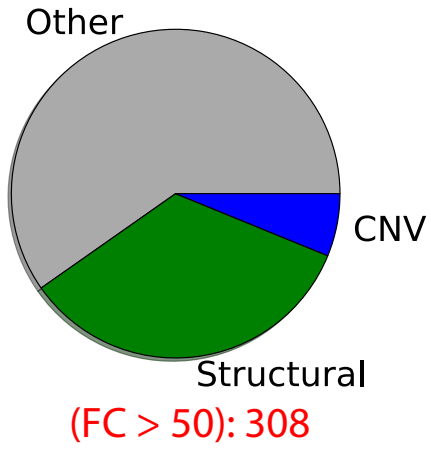
(FC > 50): 308

Gene expression variability

Copy Number Variation

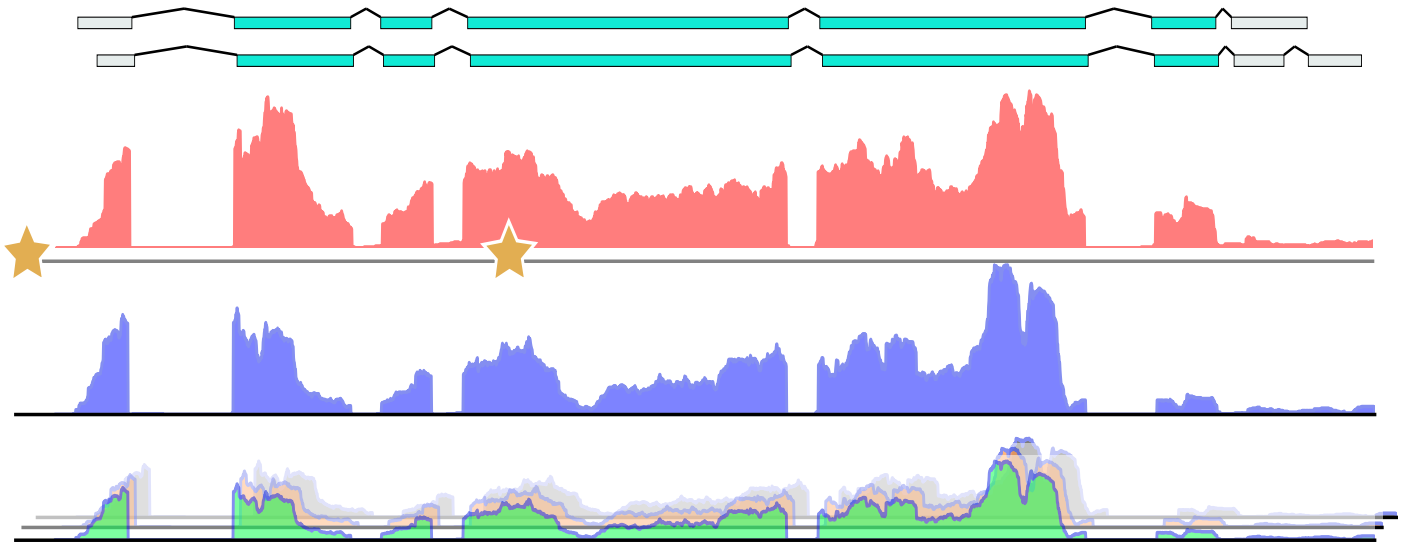






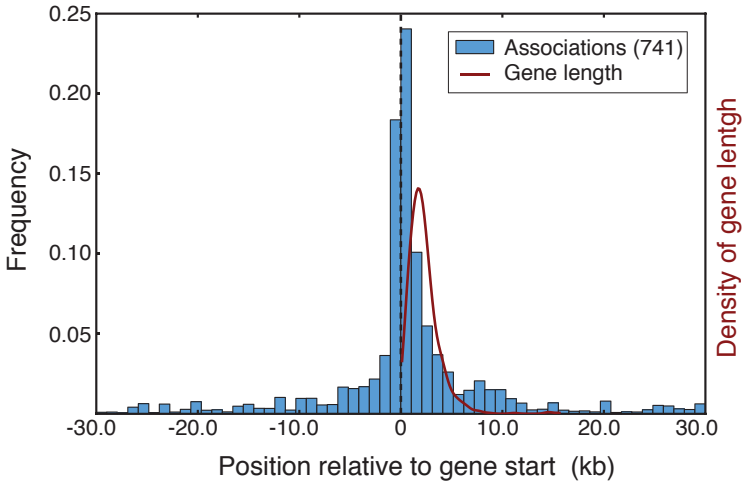
Gene expression variability

Mappable *cis* eQTLs



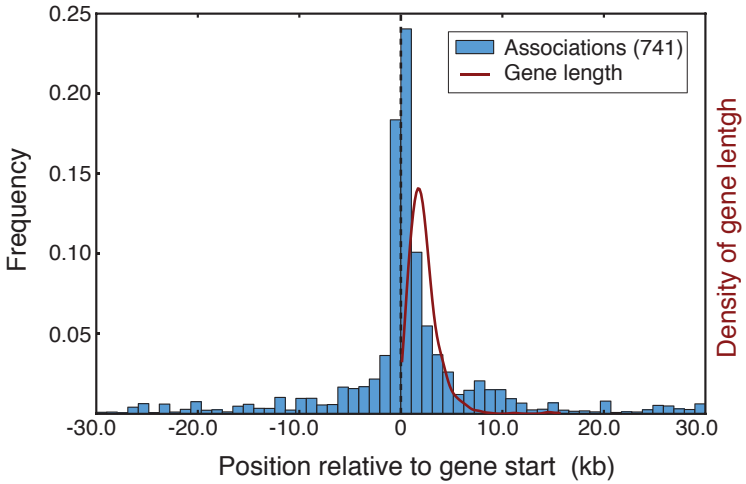
Gene expression variability

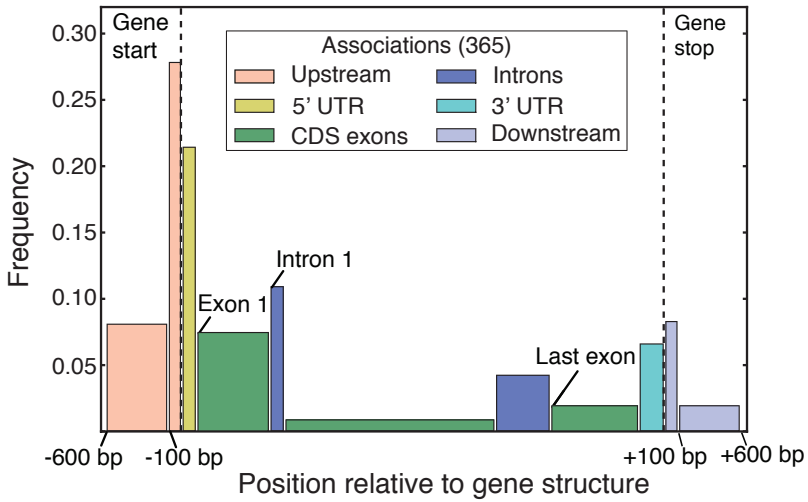
Mappable *cis* eQTLs



Gene expression variability

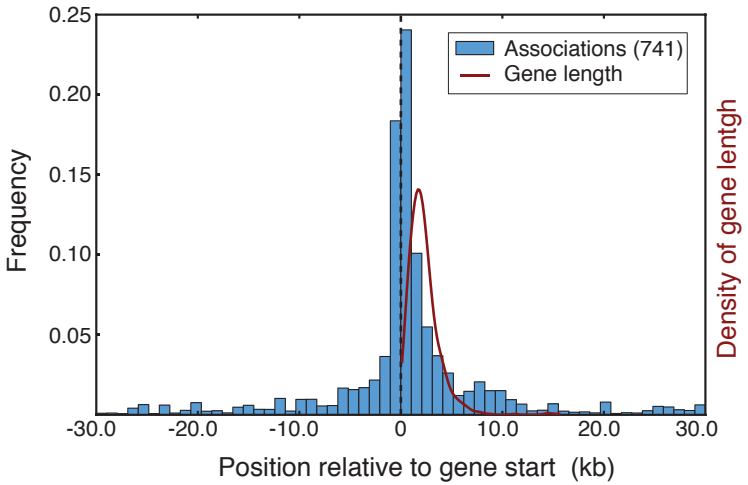
Mappable *cis* eQTLs

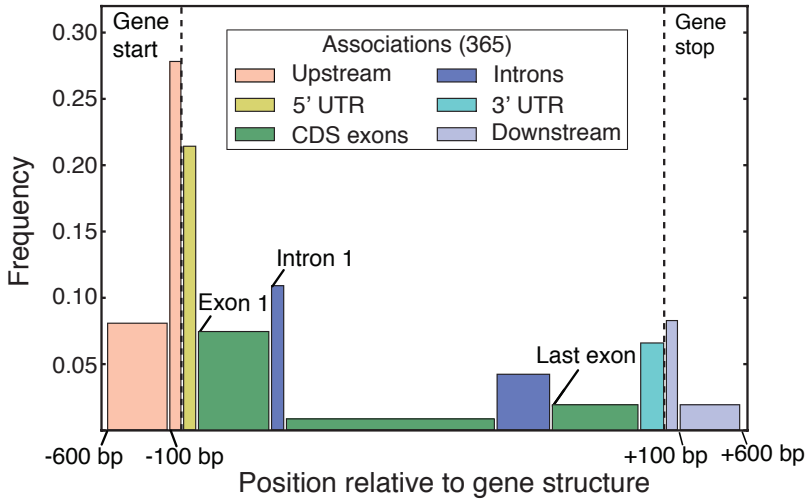


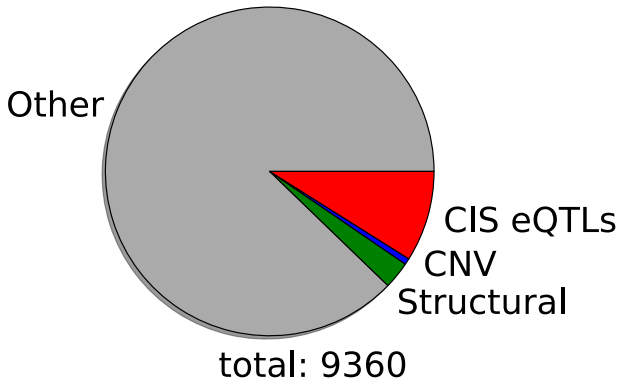


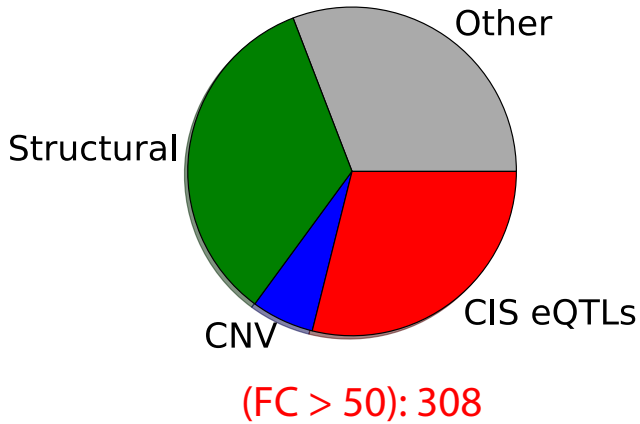
Gene expression variability

Mappable *cis* eQTLs



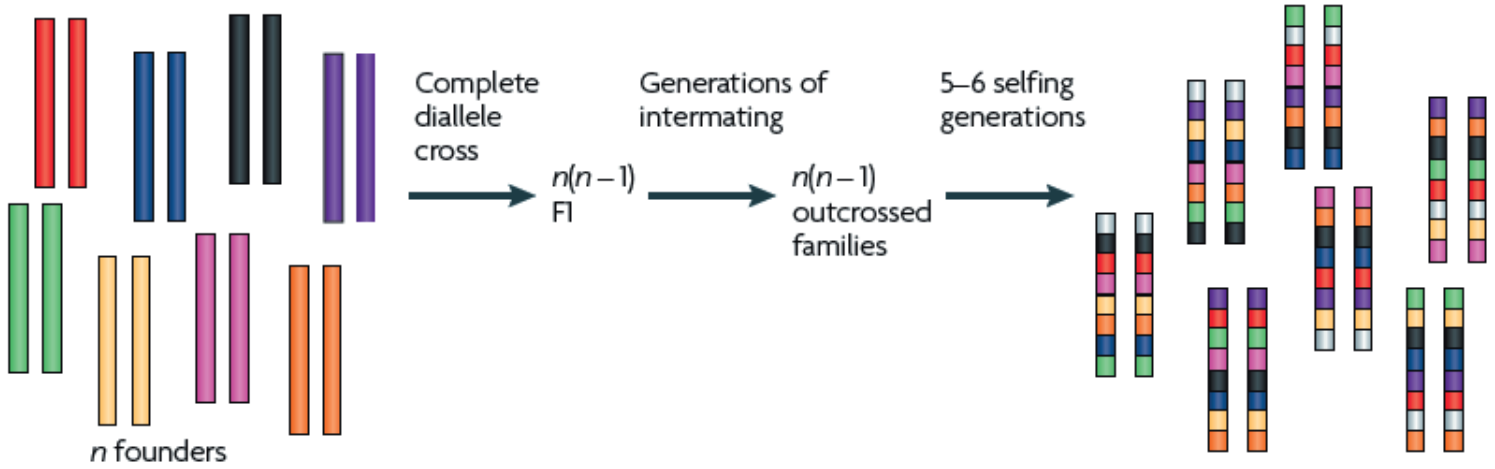






MAGIC Lines & Association Studies

Multiparent Advanced Generation Inter-Cross (MAGIC) lines



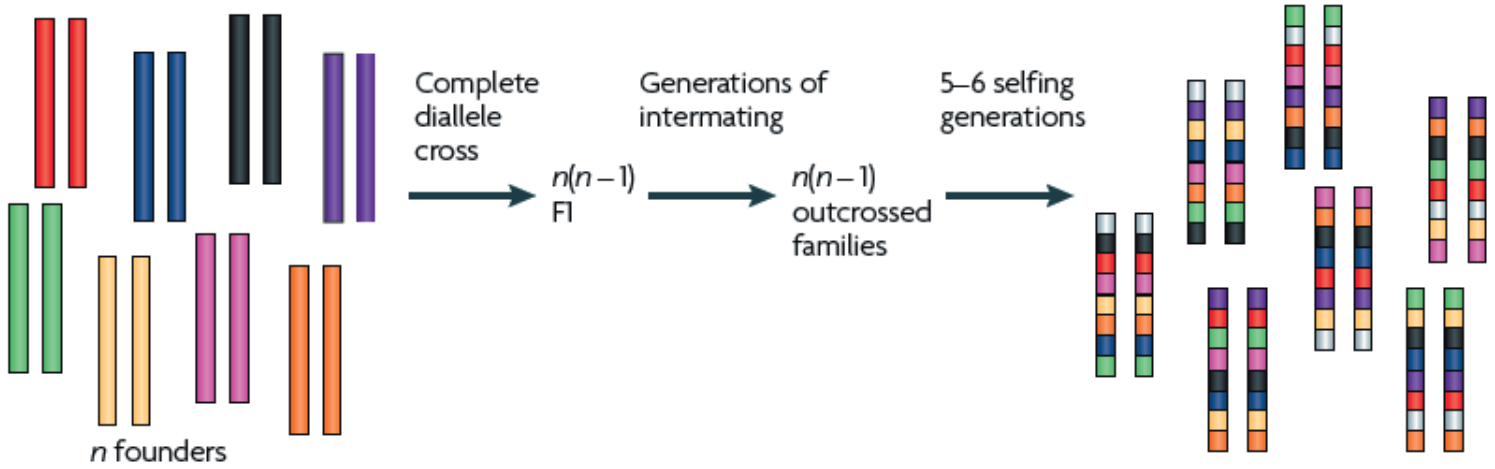
- Currently 500 available MAGIC lines [w/ Richard Mott]
- ≈200 natural strains with RNA-seq [w/ Magnus Nordborg]

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



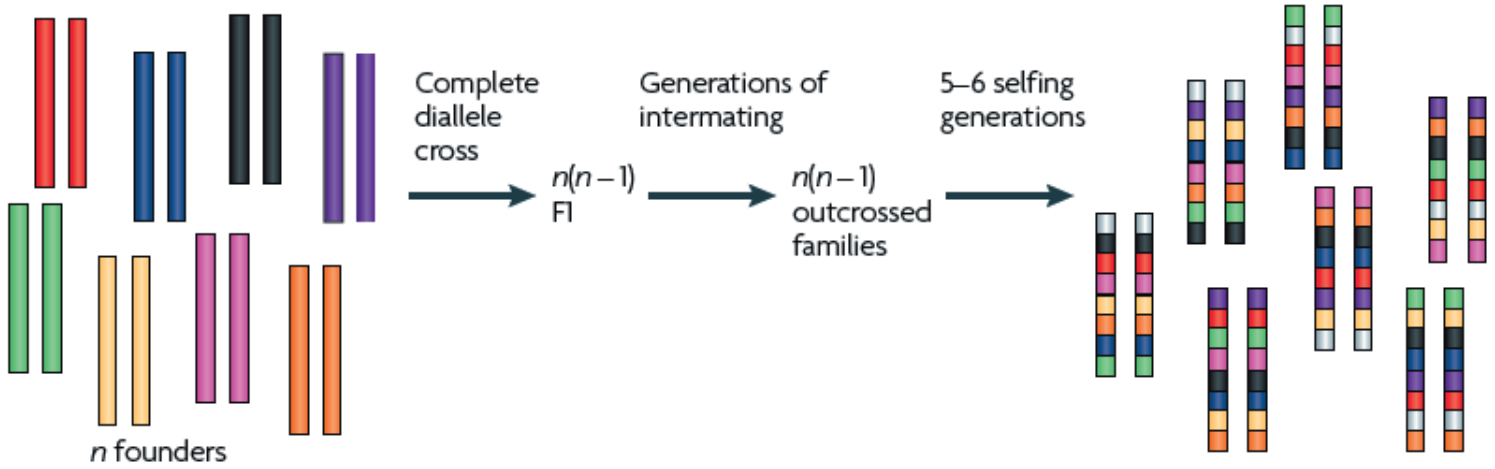
- Currently 500 available MAGIC lines [w/ Richard Mott]
- ≈200 natural strains with RNA-seq [w/ Magnus Nordborg]

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

Summary & Conclusions

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 combining

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

Goal: Develop/extend predictive models of the central dogma

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

Summary & Conclusions

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 combining

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

Goal: Develop/extend predictive models of the central dogma

f(DNA,) = RNA F(RNA,) = protein

Summary & Conclusions

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 combining

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

Goal: Develop/extend predictive models of the central dogma

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

Summary & Conclusions

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 combining

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

Goal: Develop/extend predictive models of the central dogma

f(DNA, 1 2 3) = RNA F(RNA, 4 5) = protein

Acknowledgements

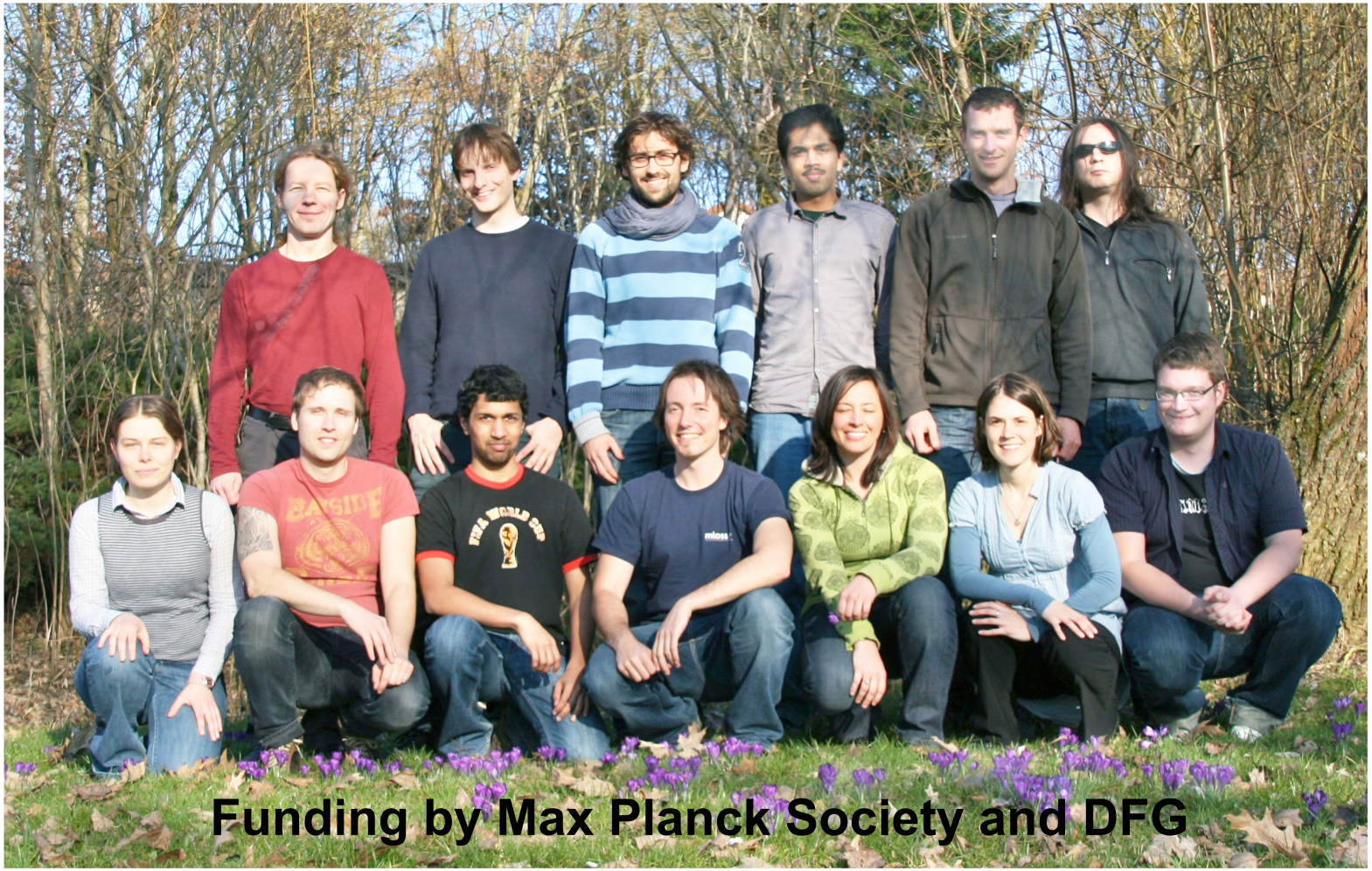
Method collaborators:

1. **Bernhard Schölkopf, MPI Tü** (Kernel & Transfer learning)
2. **Oliver Stegle, MPI Tü** (QTLs, statistical tests for RNA-seq)
3. **Klaus-Robert Müller, TU Berlin** (Large scale sequence classification)
4. Manfred Warmuth, UCSC (Analysis of boosting algorithms)
5. Oliver Kohlbacher, Uni Tübingen (Learning for Immunoinformatics)

Biology collaborators:

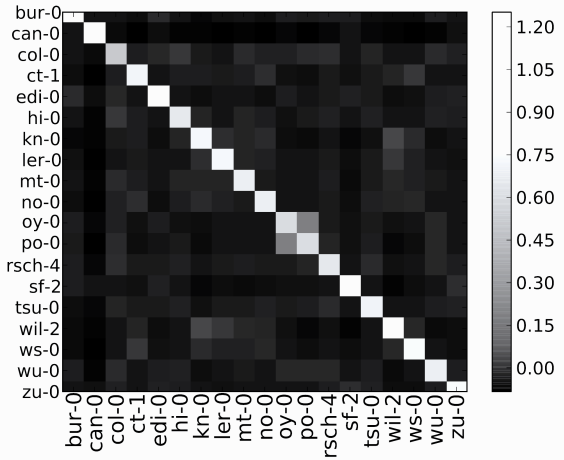
1. **Richard Mott, U Oxford** (Genomes of *A. thaliana* strains)
2. **Richard Clark, U Utah** (Transcriptomes of *A. thaliana* strains)
3. Dan Bumbarger, MPI Tü (3D reconstruction of worms)
4. **David Miller, Vanderbilt U** (*C. elegans* transcriptome analyses)
5. Detlef Weigel, MPI Tü (Annotation of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

Acknowledgements



Gene expression variability

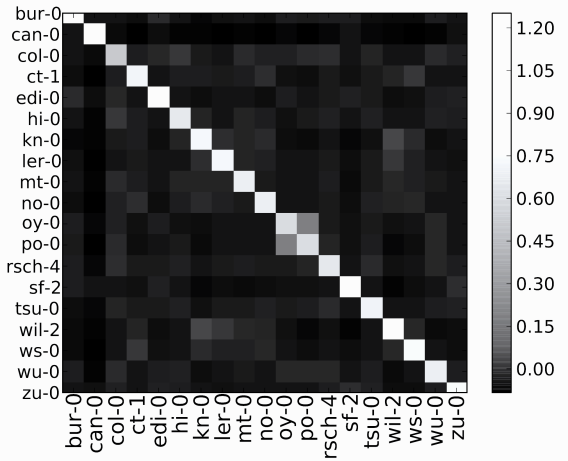
Population structure & unknown confounding



PEER, Stegle & Parts et al. 2010

Gene expression variability

Population structure & unknown confounding



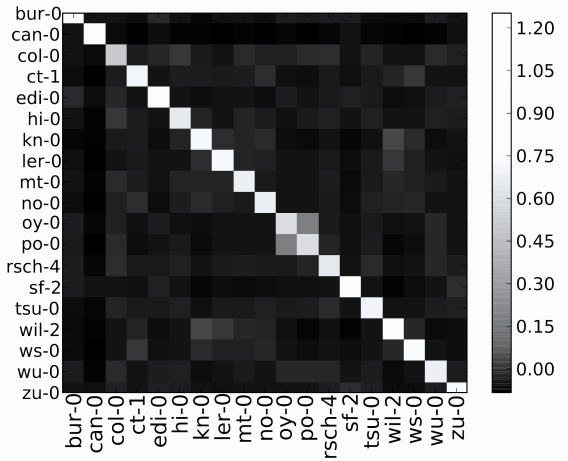
Variance component model:

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

PEER, Stegle & Parts et al. 2010

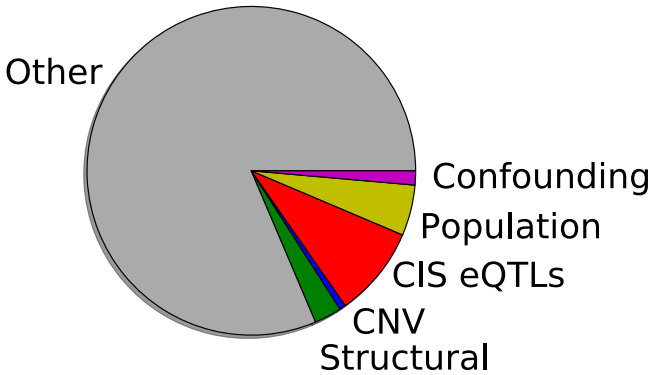
Gene expression variability

Population structure & unknown confounding

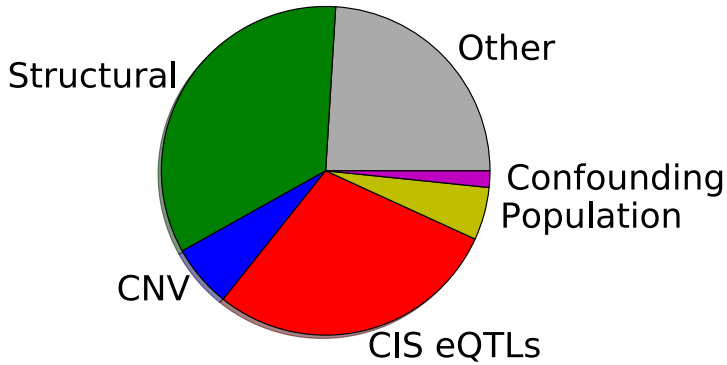


Variance component model:

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



total: 9360

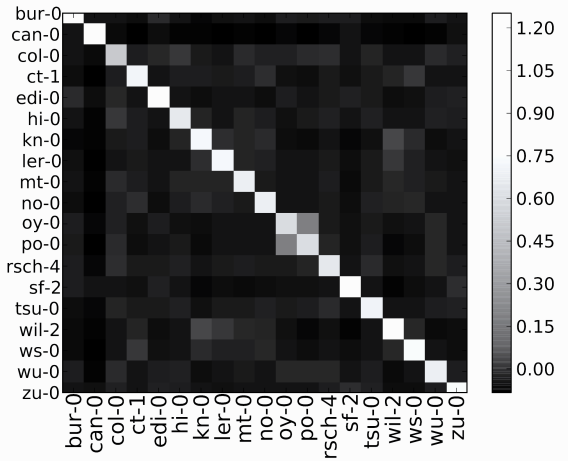


(FC > 50): 308

PEER, Stegle & Parts et al. 2010

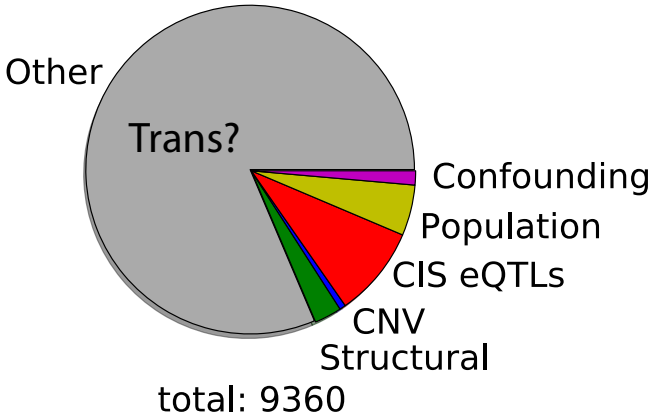
Gene expression variability

Population structure & unknown confounding

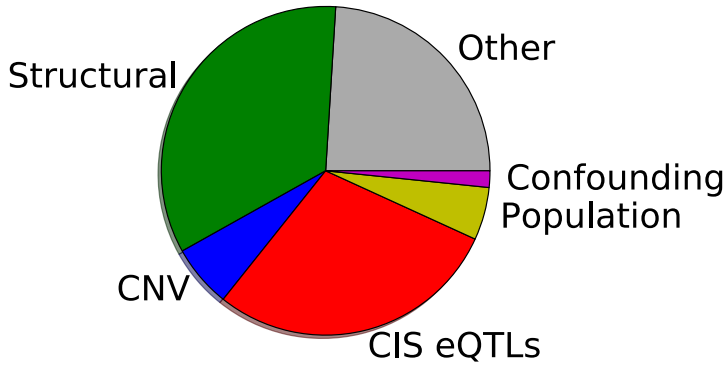


Variance component model:

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



total: 9360



(FC > 50): 308

PEER, Stegle & Parts et al. 2010

References I

Y. Altun, I. Tsochantaridis, and T. Hofmann. Hidden Markov Support Vector Machines. In *Proc. 20th Int. Conf. Mach. Learn.*, pages 3–10, 2003.

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`.

J Behr, R Bohnert, L Hartmann, VT Sreedharan, G Zeller, G Schweikert, and **G Rätsch**. *De novo* RNA-seq-based genome annotation with mGene.ngs. We developed a genome annotation technique that combines genomic features with RNA-seq features for *de novo* genome annotation. The material has been presented at the Genome Informatics meeting in Hinxton in September 2010. In preparation for *Genome Biology*., 2011.

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.

References II

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.

R Dahm. Friedrich miescher and the discovery of dna. *Developmental Biology*, 278(2):274–288, 2005.

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.

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

References III

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, 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.

L. Hartmann, P. Drewe, G. Rätsch, and A. Wachter. Light induced transcriptome changes in arabidopsis thaliana. in preparation., 2011, i.P.

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.

References IV

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.

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.

F. Miescher. Die spermatozoen einiger wirbelthiere. ein beitrag zur histochemie. *Verh. Nat.forsch. Ges. Basel*, 1874.

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 and S. Sonnenburg. Large scale hidden semi-markov svms. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems (NIPS'06)*, volume 19, pages 1161–1168, Cambridge, MA, 2007. MIT Press. URL `http://www.fml.tuebingen.mpg.de/raetsch/projects/HSMSVM`.

References V

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

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.

S Sonnenburg, G Schweikert, P Philips, J Behr, and G Rätsch. Accurate splice site prediction using support vector machines. *BMC Bioinformatics*, 8 Suppl 10:S7, 2007. ISSN 1471-2105 (Electronic). doi: 10.1186/1471-2105-8-S10-S7.

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.

References VI

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.

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.

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.