Machine Learning Methods for Discovering Common Sequence Variation in *A. thaliana*

Gunnar Rätsch

Friedrich Miescher Laboratory,
Max Planck Society,
Tübingen

Technical University Berlin
March 31, 2008

Current Projects

Genome Biology

- *Ab initio* gene finding
  - Prediction of alternative splicing
  - Great improvements on *C. elegans*
- Prediction of polymorphisms from genome tiling arrays
  - Prediction of SNPs, deletions & highly polymorphic regions
  - Predicting changes in gene products
- Accurate alignments, RNA secondary structure

Machine Learning

- Improved learning algorithms
  - Large scale (>10 millions of examples)
  - Structured outputs (segmentations, graphs)
- Methods for understanding learning results
  - Helps to understand how biological mechanisms work
- Theoretical analysis of iterative algorithms

Current Projects

Genome Biology

- *Ab initio* gene finding
  - Prediction of alternative splicing
  - Great improvements on *C. elegans*
- Prediction of polymorphisms from genome tiling arrays
  - Prediction of SNPs, deletions & highly polymorphic regions
  - Predicting changes in gene products
- Accurate alignments, RNA secondary structure

Machine Learning

- Improved learning algorithms
  - Large scale (>10 millions of examples)
  - Structured outputs (segmentations, graphs)
- Methods for understanding learning results
  - Helps to understand how biological mechanisms work
- Theoretical analysis of iterative algorithms

Current Projects

Genome Biology

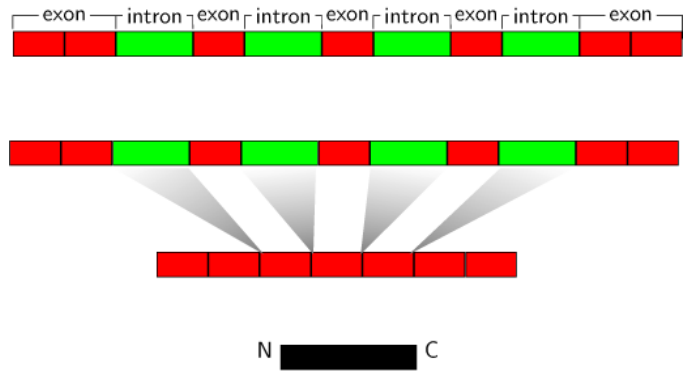
Machine Learning

- *Ab initio* gene finding
  - Prediction of alternative splicing
  - Great improvements on *C. elegans*
- Prediction of polymorphisms from genome tiling arrays
  - Prediction of SNPs, deletions & highly polymorphic regions
  - Predicting changes in gene products
- Accurate alignments, RNA secondary structure
- Improved learning algorithms
  - Large scale (>10 millions of examples)
  - Structured outputs (segmentations, graphs)
- Methods for understanding learning results
  - Helps to understand how biological mechanisms work
- Theoretical analysis of iterative algorithms

Ab initio Gene Finding

From DNA to protein:

- Genes organized in exons & introns
- Transcription: DNA to pre-mRNA
- Splicing: removes introns $\Rightarrow$ mRNA
- Translation: mRNA to protein



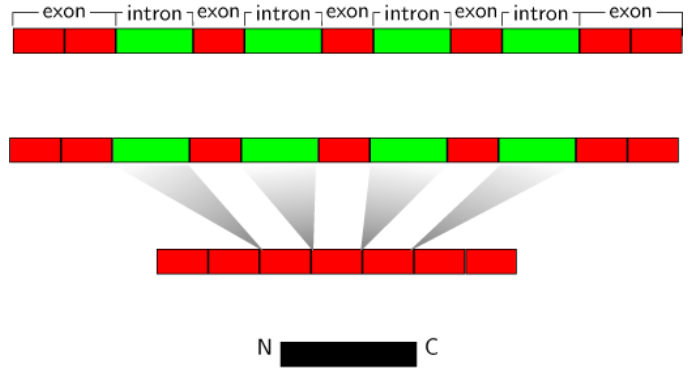
Given the DNA, predict resulting mRNA and protein

- Requires very accurate identification of
  - splice sites, translation & transcription starts & stops
  - sites of regulation (transcription, splicing, etc.)
- Develop methods to integrate single site predictions
  - Novel learning methods for structured outputs

Ab initio Gene Finding

From DNA to protein:

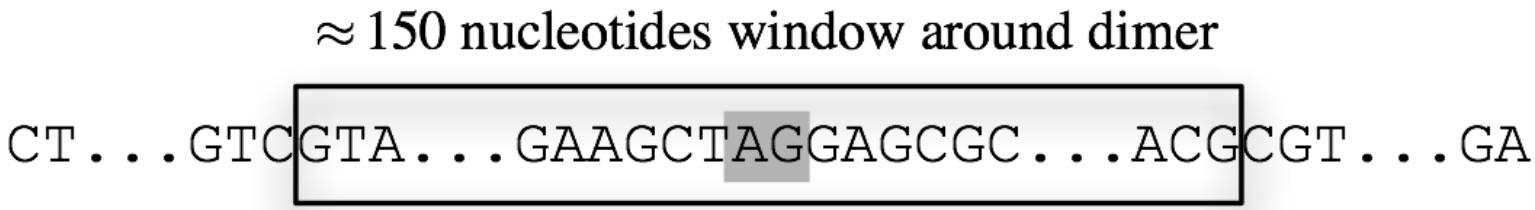
- Genes organized in exons & introns
- Transcription: DNA to pre-mRNA
- Splicing: removes introns $\Rightarrow$ mRNA
- Translation: mRNA to protein



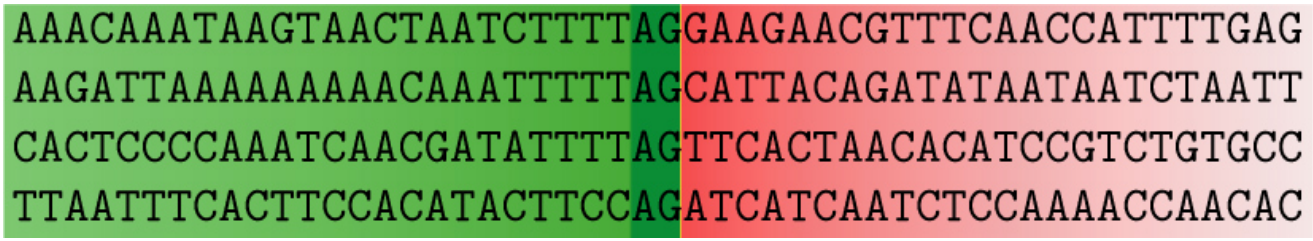
Given the DNA, predict resulting mRNA and protein

- Requires very accurate identification of
  - splice sites, translation & transcription starts & stops
  - sites of regulation (transcription, splicing, etc.)
- Develop methods to integrate single site predictions
  - Novel learning methods for structured outputs

2-Class Splice Site Detection



- True sites: fixed window around a true splice site
- Decoys sites: generated by shifting the window



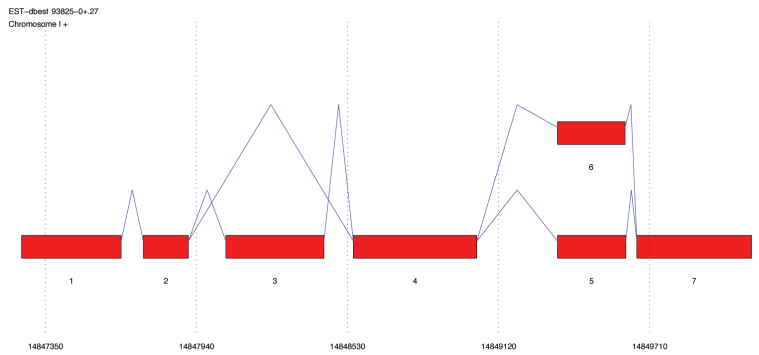
⇒ Very unbalanced problem (1:200)
⇒ Millions of training sequences from EST databases
⇒ Testing on 7 billion test sequences

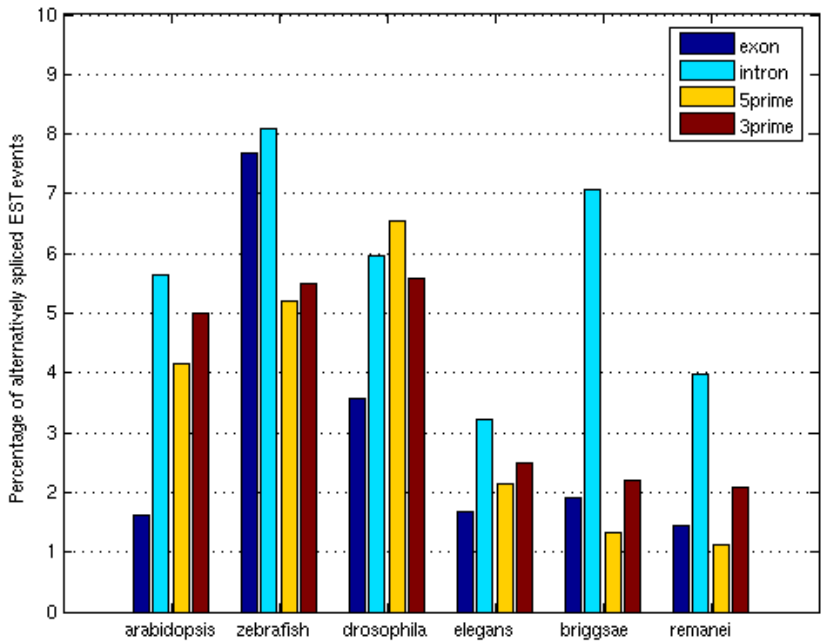
Prediction of Alternative Splicing

Goal: Find sites of alternative splicing conditions and regulating genes

So far: Develop techniques to

- Understand differences between alternative and constitutive splicing
- Predict yet unknown alternative splicing events
- Predict on newly sequenced organisms
- Experimentally verify predictions via RT-PCR.



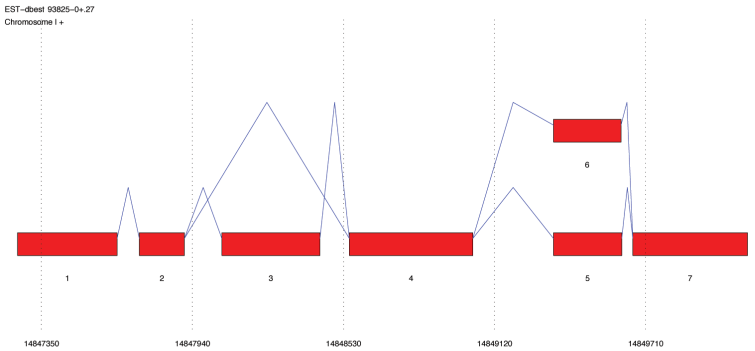


Prediction of Alternative Splicing

Goal: Find sites of alternative splicing conditions and regulating genes

So far: Develop techniques to

- Understand differences between alternative and constitutive splicing
- Predict yet unknown alternative splicing events
- Predict on newly sequenced organisms
- **Experimentally verify predictions via RT-PCR.**

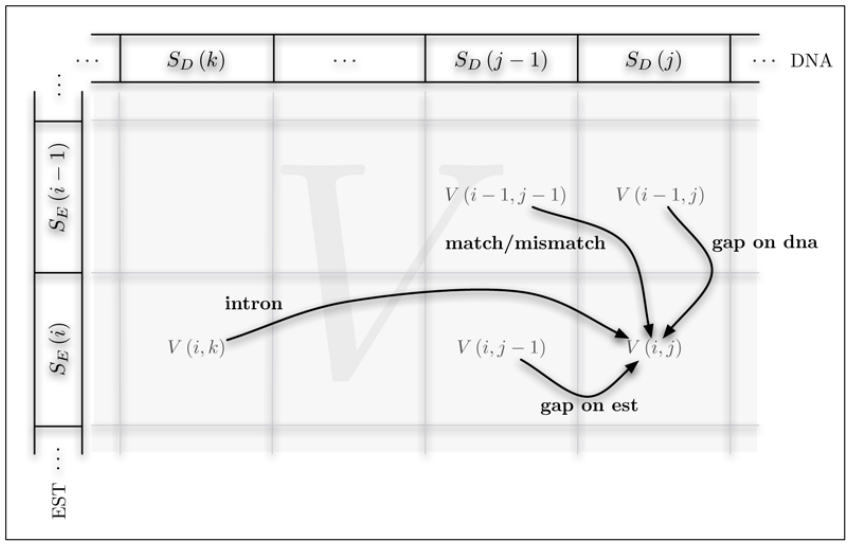




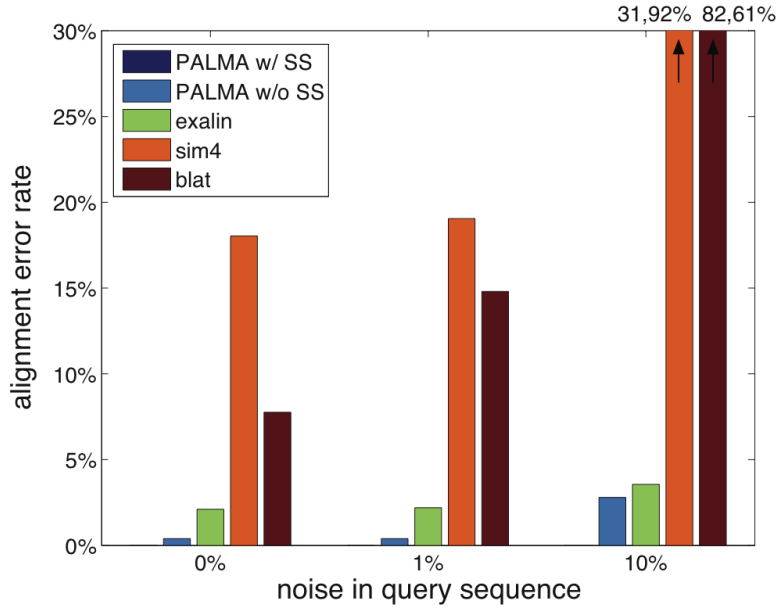
Improved Alignment Algorithms

Alignment approach based on large margins:

- Common alignment techniques
- SVM based splice site predictions
- Convex optimization for optimal combination



Dynamic programming recursion: transition weights used in the dynamic programming algorithm are learned by a large margin algorithm.



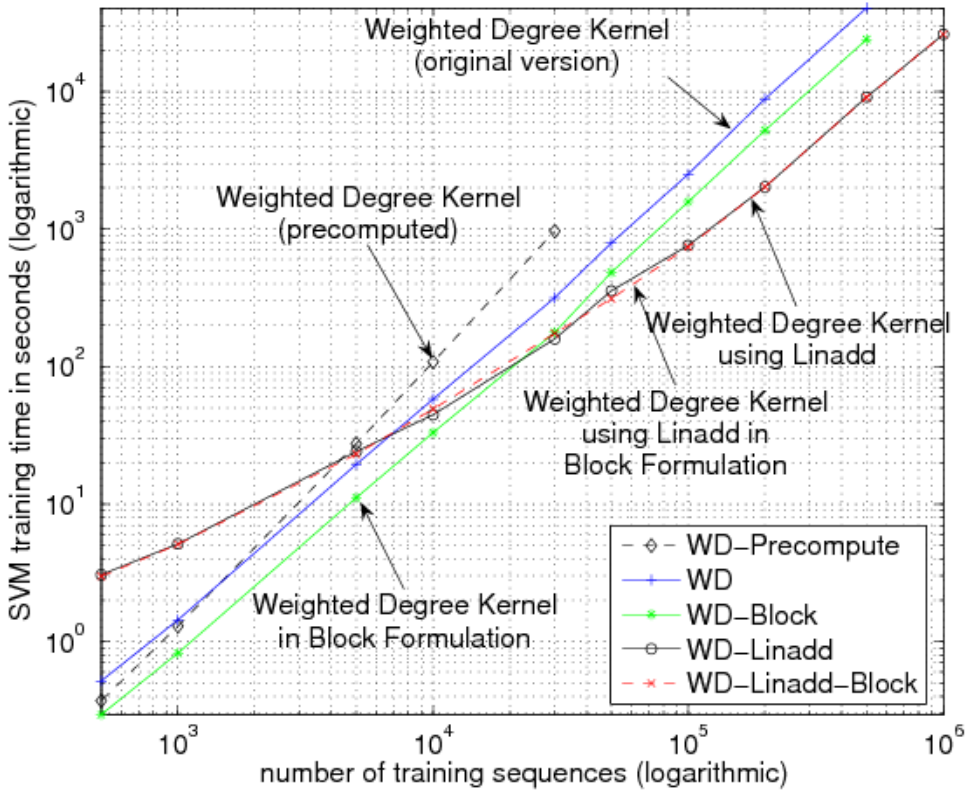
Accuracy of alignment: Error rates of *blat* and *sim4* drastically increase when adding noise. Only *exalin* and *PALMA* have low error rates for noise levels of at most 20%.

Large Scale Kernel Learning

Goal: Extend current SVM training methods to millions of examples

So far: Speed ups obtained for

- String kernels using suitable data structures
- Optimized chunking algorithm for most string kernels
- No need for caching



All implemented in **Shogun** (available under GPL)
(`http://www.shogun-toolbox.org`)

Large Scale Structured Output Learning

- Learn function S_θ to score "structures" $\mathbf{s}$ (linear in $\theta \in \mathbb{R}^d$)
- Idea: Score true structure much higher than all other structures:

$$\begin{aligned} \underset{\xi \in \mathbb{R}^N_+, \theta \in \mathbb{R}^d}{\text{minimize}} \quad & C \sum_{n=1}^{N} \xi_n + \underbrace{\mathbf{P}(\theta)}_{\text{regularizer}} \\ \text{subject to} \quad & S_\theta(\mathbf{s}_n^+) - S_\theta(\mathbf{s}^-) \geq 1 - \xi_n \\ \text{for all} \quad & \mathbf{s}^- \neq \mathbf{s}_n^+ \text{ and } n = 1, \ldots, N \end{aligned}$$

  $\Rightarrow$ Solve problem in a two-stage algorithm

  $\Rightarrow$ Find violating constraints with dynamic programming
- Chose appropriate $\mathbf{P}$
  - Linear $\Rightarrow$ SILP techniques ($\log N/\epsilon^2$)
  - Quadratic $\Rightarrow$ $\mathcal{O}(R^2/\epsilon^2)$ iterations
- Good for: Gene finding, polymorphism detection, inverse alignment problem, secondary structure prediction

Improving Interpretability

Consider Multiple Kernel Learning problem:

$$f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i \mathrm{k}(\mathbf{x}_i, \mathbf{x}) + b \quad \text{with} \quad \mathrm{k}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{K} \beta_k \, \mathrm{k}_k(\mathbf{x}_i, \mathbf{x}_j)$$

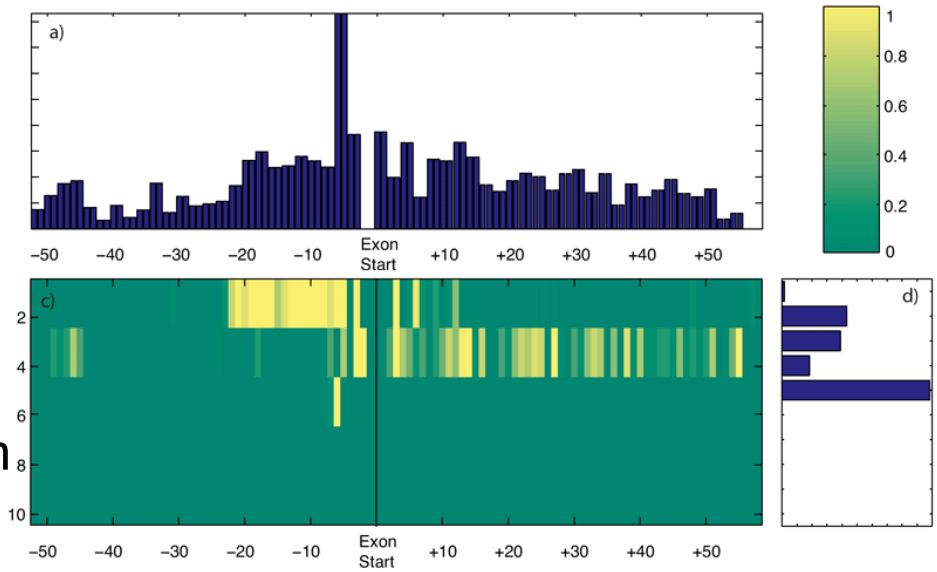
Why?

- Interpret kernel algorithms
- Combine heterogeneous data
- Automated model selection

Contributions:

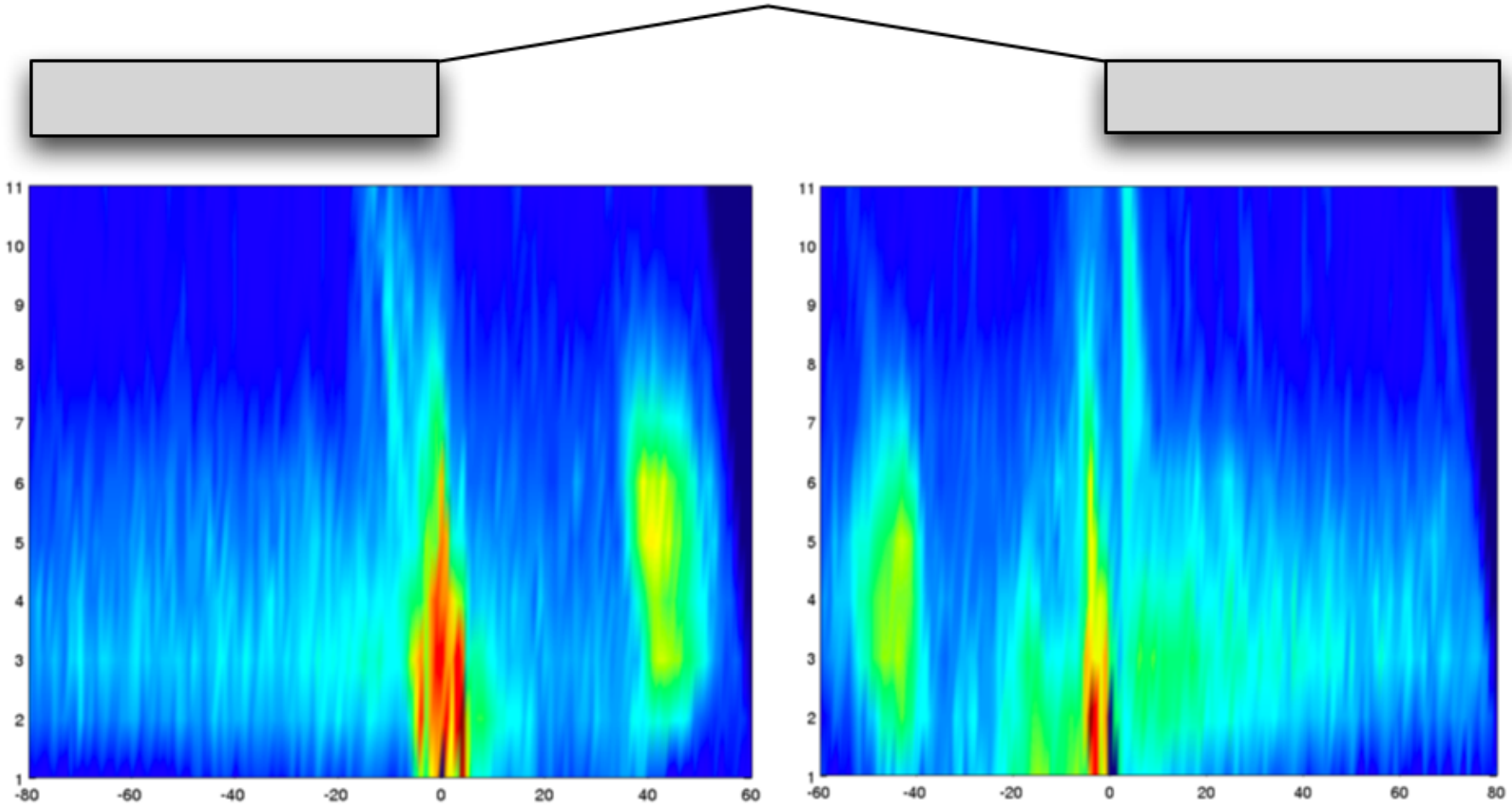
- General formulation
- Efficient algorithms
  - based on single kernel algorithm
  - works with 1 million examples

⇓ splice site recognition task ⇓

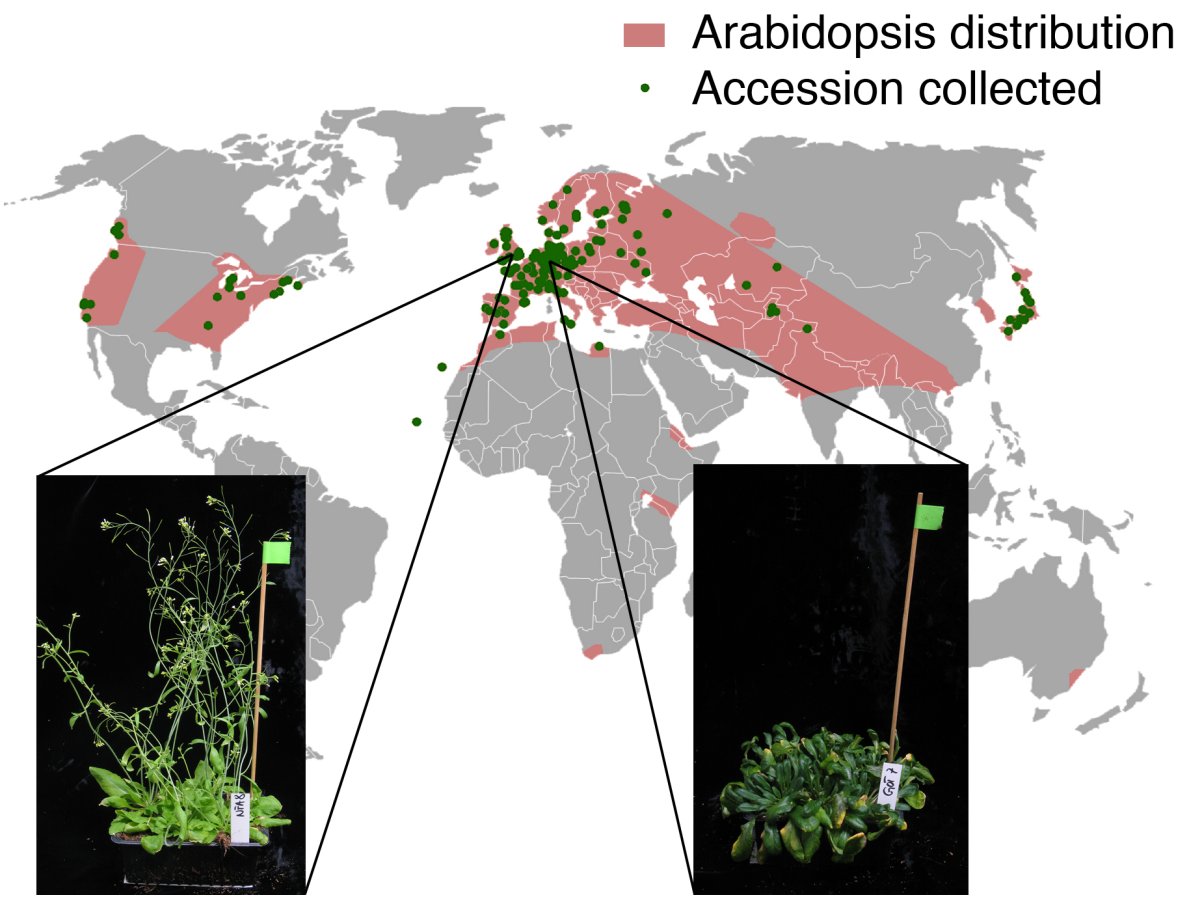


Improving Interpretability II

Example: Splice Site Detection in *C. elegans*



Analysis of Polymorphisms: Introduction



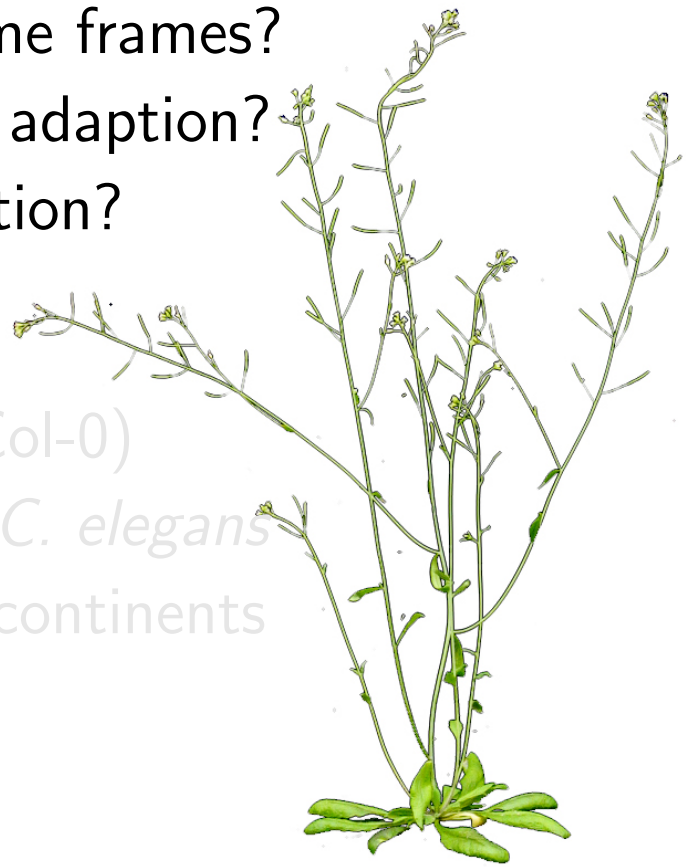
What is the genetic basis of variation?

Modified from Koornneef et al. 2004. *Ann. Rev. Plant Biology*, 55, 141-172.

Introduction

Questions:

- What sequence changes occur in short time frames?
- Which polymorphisms and genes underlie adaption?
- What are the consequences for gene function?



Arabidopsis thaliana:

- 119 Mb finished euchromatic sequence (Col-0)
- Resources comparable to *Drosophila* and *C. elegans*
- Collections of >1000 wild strains from 3 continents
- Strains are largely homozygous

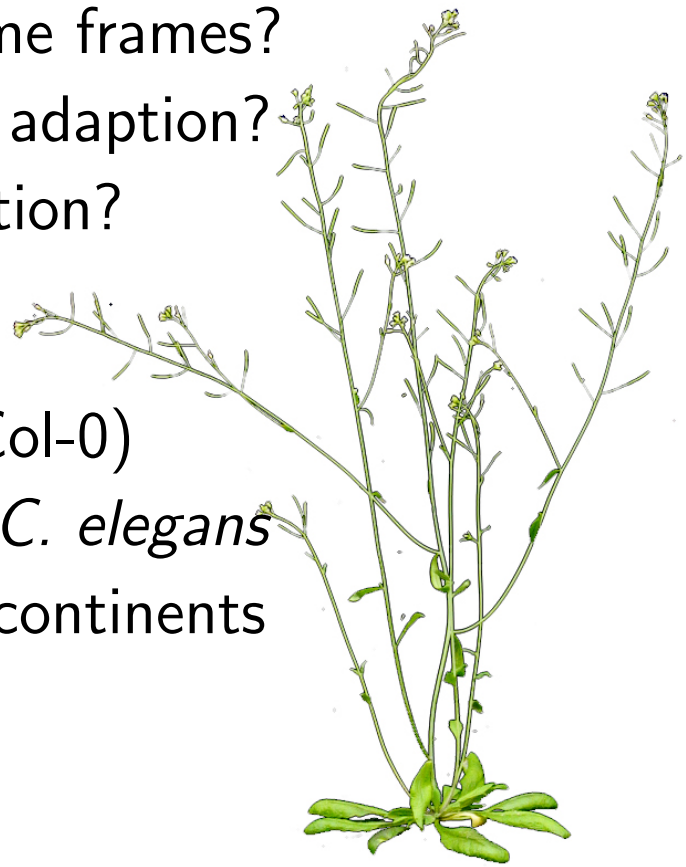
Resequencing of 20 wild strains

- Genome-wide identification of sequence polymorphisms
- High-density oligo-nucleotide arrays for high-throughput resequencing

Introduction

Questions:

- What sequence changes occur in short time frames?
- Which polymorphisms and genes underlie adaption?
- What are the consequences for gene function?



Arabidopsis thaliana:

- 119 Mb finished euchromatic sequence (Col-0)
- Resources comparable to *Drosophila* and *C. elegans*
- Collections of >1000 wild strains from 3 continents
- Strains are largely homozygous

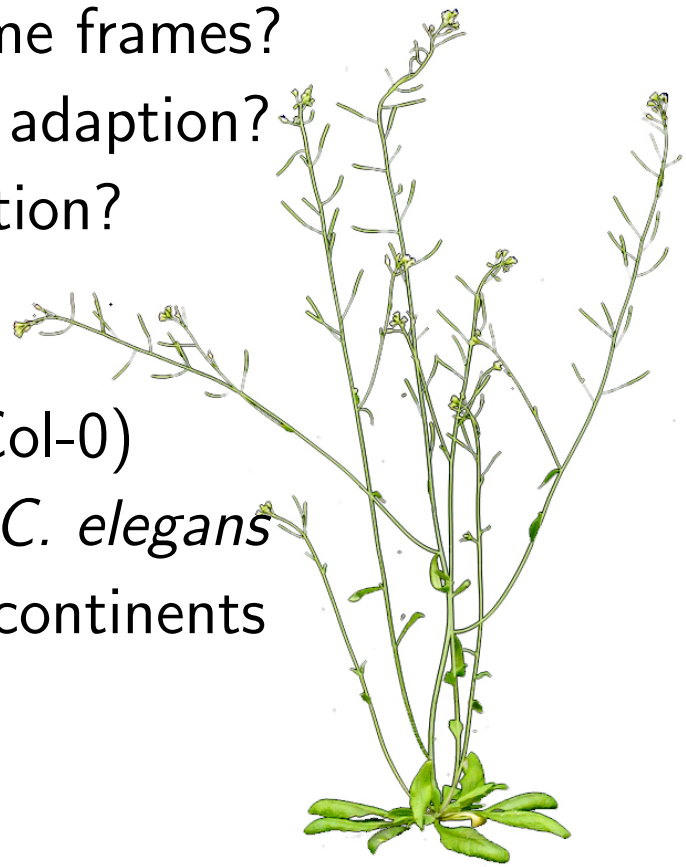
Resequencing of 20 wild strains

- Genome-wide identification of sequence polymorphisms
- High-density oligo-nucleotide arrays for high-throughput resequencing

Introduction

Questions:

- What sequence changes occur in short time frames?
- Which polymorphisms and genes underlie adaption?
- What are the consequences for gene function?



Arabidopsis thaliana:

- 119 Mb finished euchromatic sequence (Col-0)
- Resources comparable to *Drosophila* and *C. elegans*
- Collections of >1000 wild strains from 3 continents
- Strains are largely homozygous

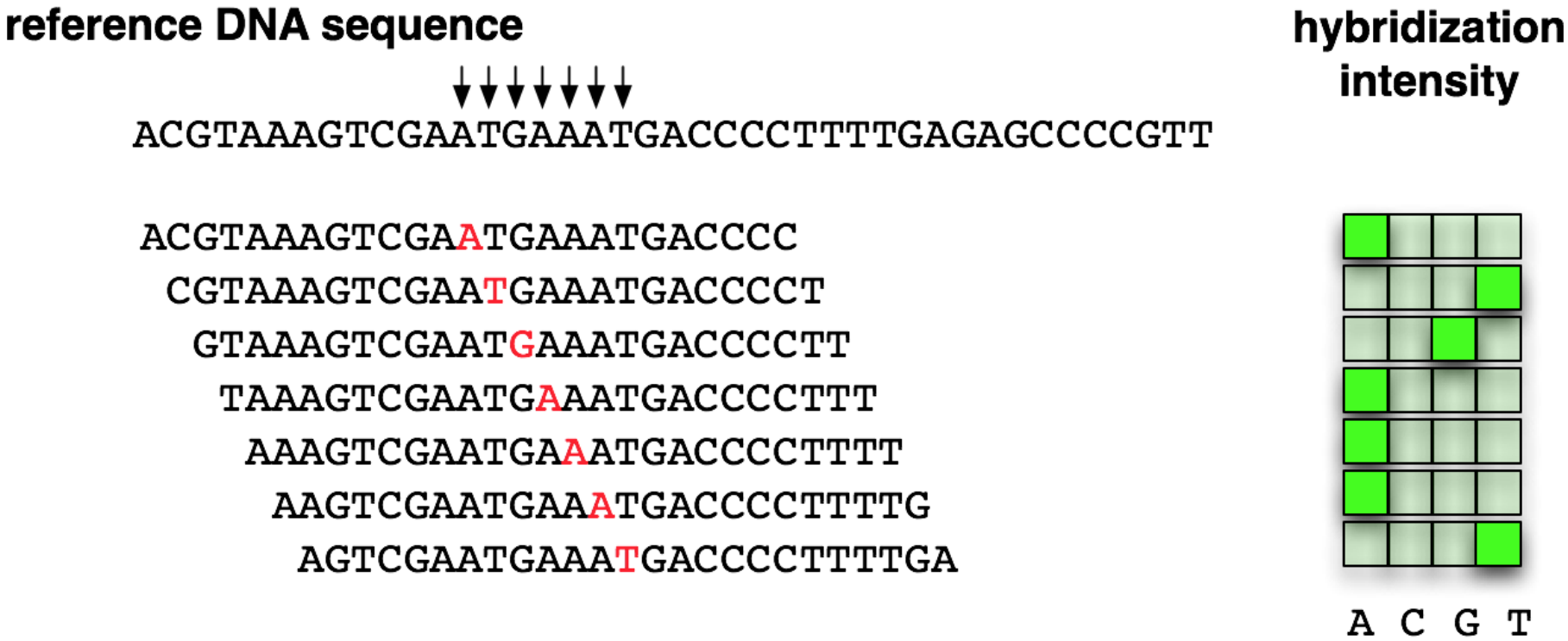
Resequencing of 20 wild strains

- Genome-wide identification of sequence polymorphisms
- High-density oligo-nucleotide arrays for high-throughput resequencing

Resequencing Array Basics I

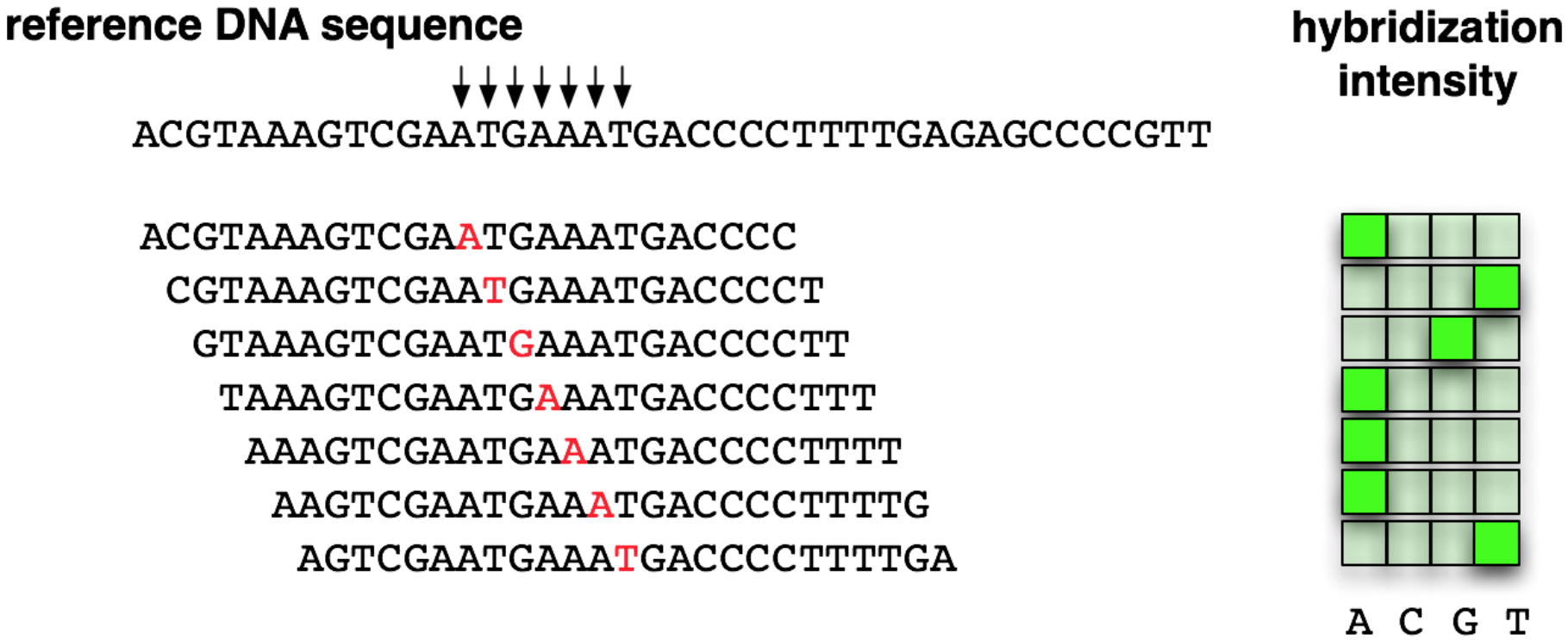


Resequencing Array Basics II



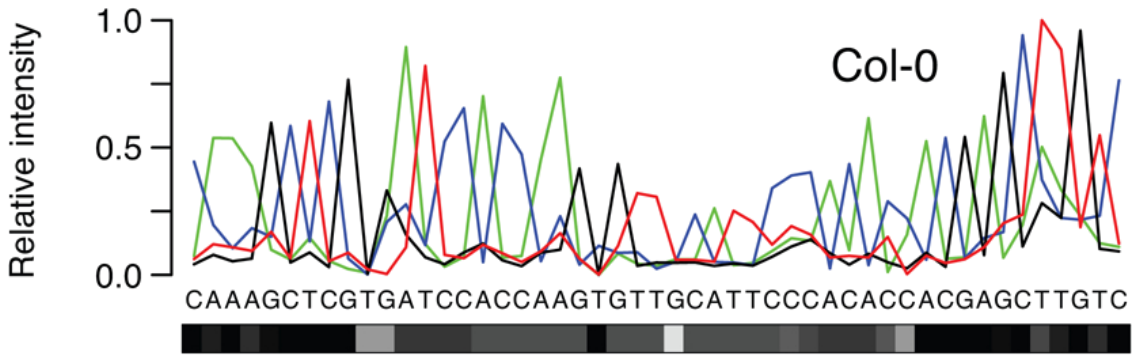
- >99.99% of bases represented
- Each base queried with forward and reverse strand probe quartets
- Nearly 1 billion oligos per accession
- 19+1 accessions surveyed representing worldwide distribution

Resequencing Array Basics II



- >99.99% of bases represented
- Each base queried with forward and reverse strand probe quartets
- Nearly 1 billion oligos per accession
- 19+1 accessions surveyed representing worldwide distribution

Resequencing Data



Data analysis challenge

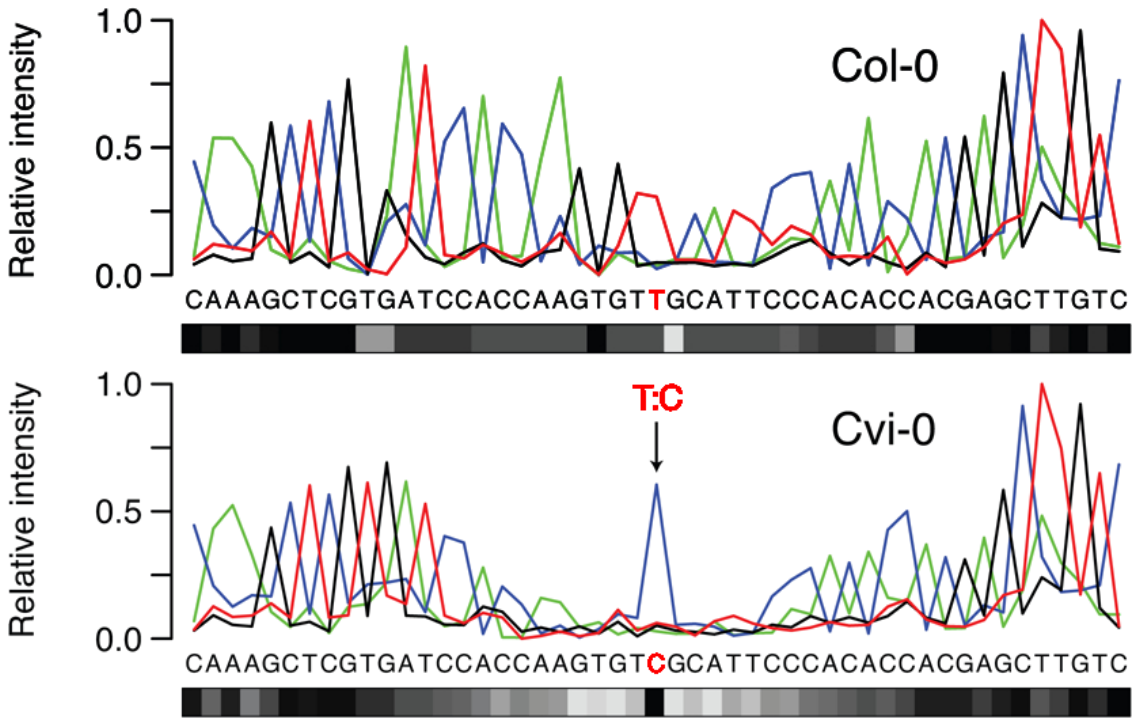
- Hybridization intensities depend on
  - Oligomer
  - Accession
  - Repeats
- Measurement noise
- Identify SNPs

Problematic cases

- Highly polymorphic regions
- Deletions/insertions



Resequencing Data





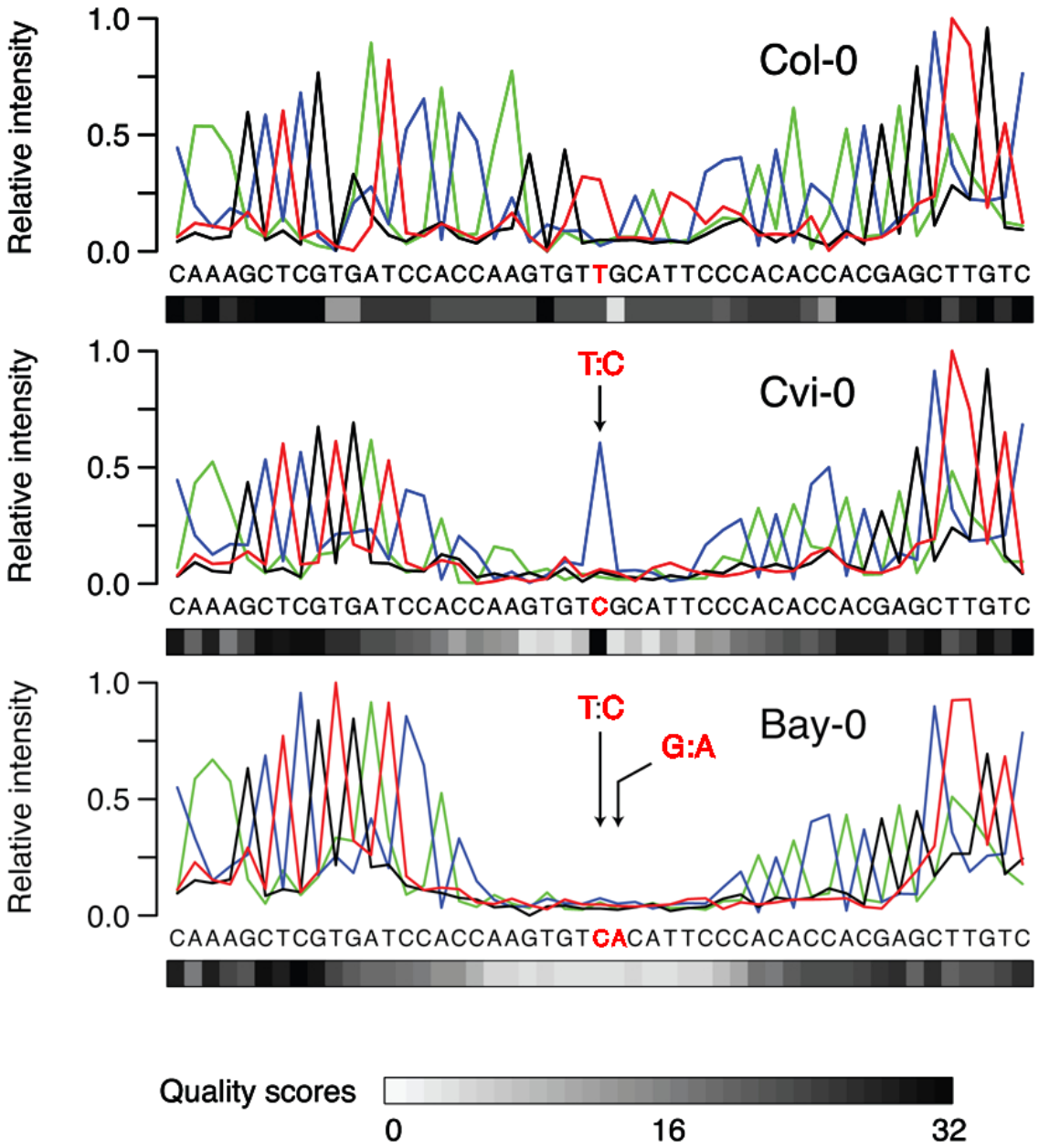
Data analysis challenge

- Hybridization intensities depend on
  - Oligomer
  - Accession
  - Repeats
- Measurement noise
- Identify SNPs

Problematic cases

- Highly polymorphic regions
- Deletions/insertions

Resequencing Data



Data analysis challenge

- Hybridization intensities depend on
  - Oligomer
  - Accession
  - Repeats
- Measurement noise
- Identify SNPs

Problematic cases

- Highly polymorphic regions
- Deletions/insertions

Labeled data and training

Labeled training set

- 1,213 fragments of length $\approx$ 550bp for 19 accessions
- sampled by PCR and dideoxy sequencing
- $\approx$2,700 known SNPs/accession (Nordborg et al., PLoS Biol., 2005)
- $\approx$400 indel polymorphisms/accession

Training

- Classification using Support Vector Machines with 302 features
- two-layered approach; including cross-accesion features in second layer
- Out-of-sample evaluation and prediction on whole genome
- Comparison with Perlegen's model based method (Hinds et al., Science, 2005)

Labeled data and training

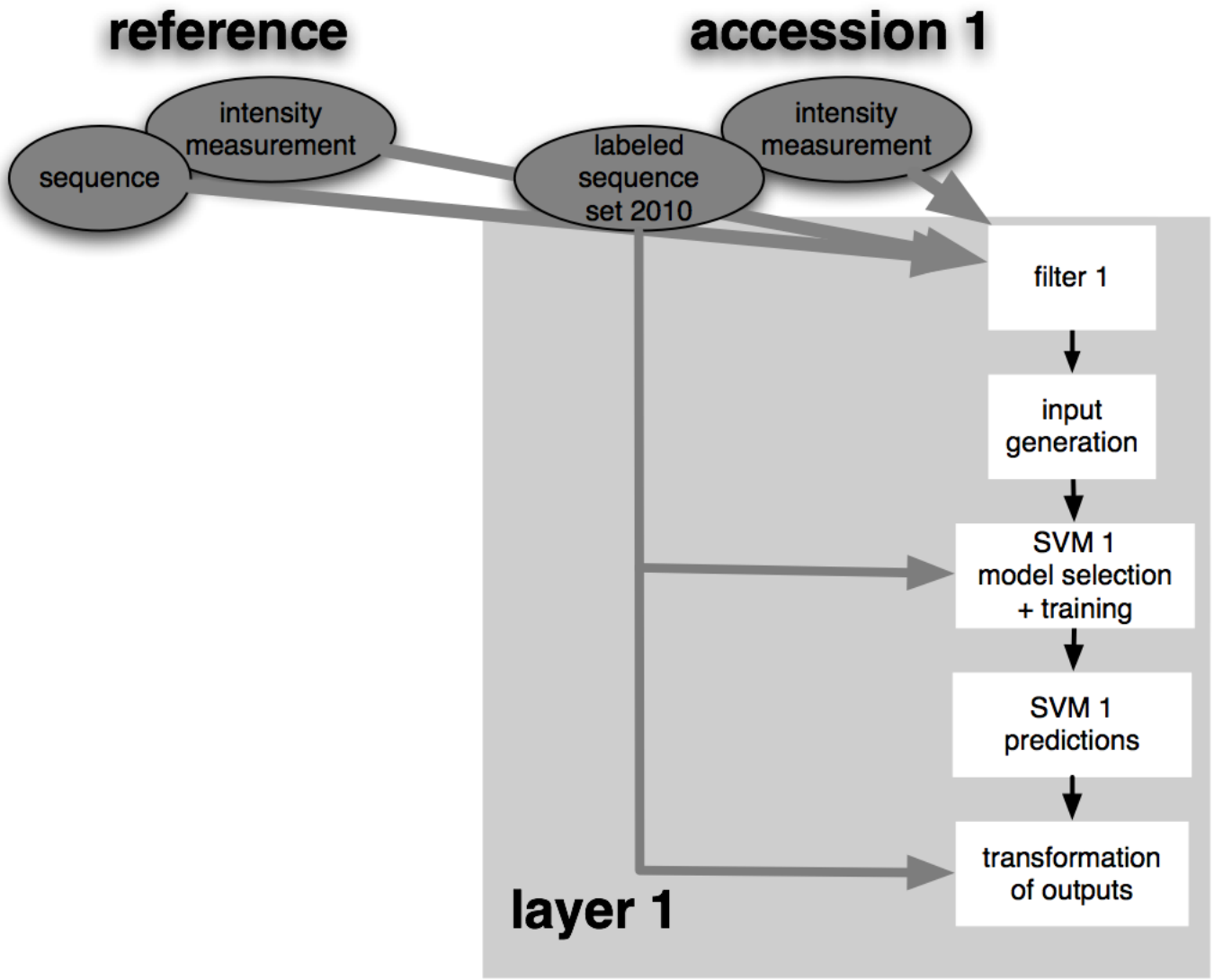
Labeled training set

- 1,213 fragments of length $\approx$ 550bp for 19 accessions
- sampled by PCR and dideoxy sequencing
- $\approx$2,700 known SNPs/accession (Nordborg et al., PLoS Biol., 2005)
- $\approx$400 indel polymorphisms/accession

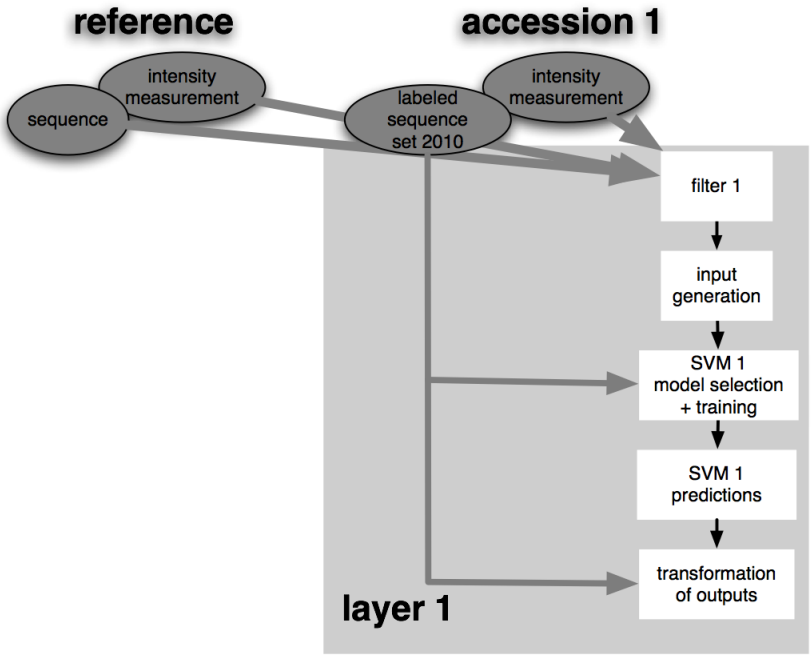
Training

- Classification using Support Vector Machines with 302 features
- two-layered approach; including cross-accesion features in second layer
- Out-of-sample evaluation and prediction on whole genome
- Comparison with Perlegen's model based method (Hinds et al., Science, 2005)

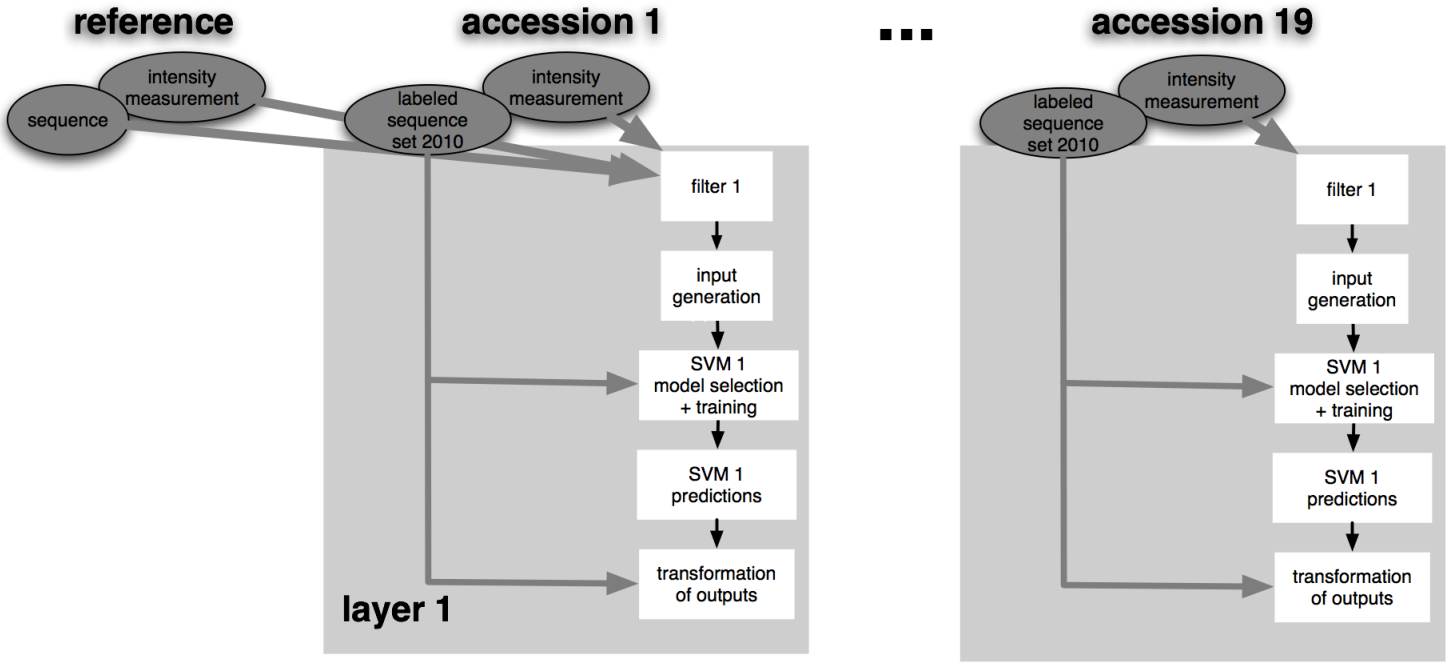
2-Layered Architecture for Inter-Strain Integration



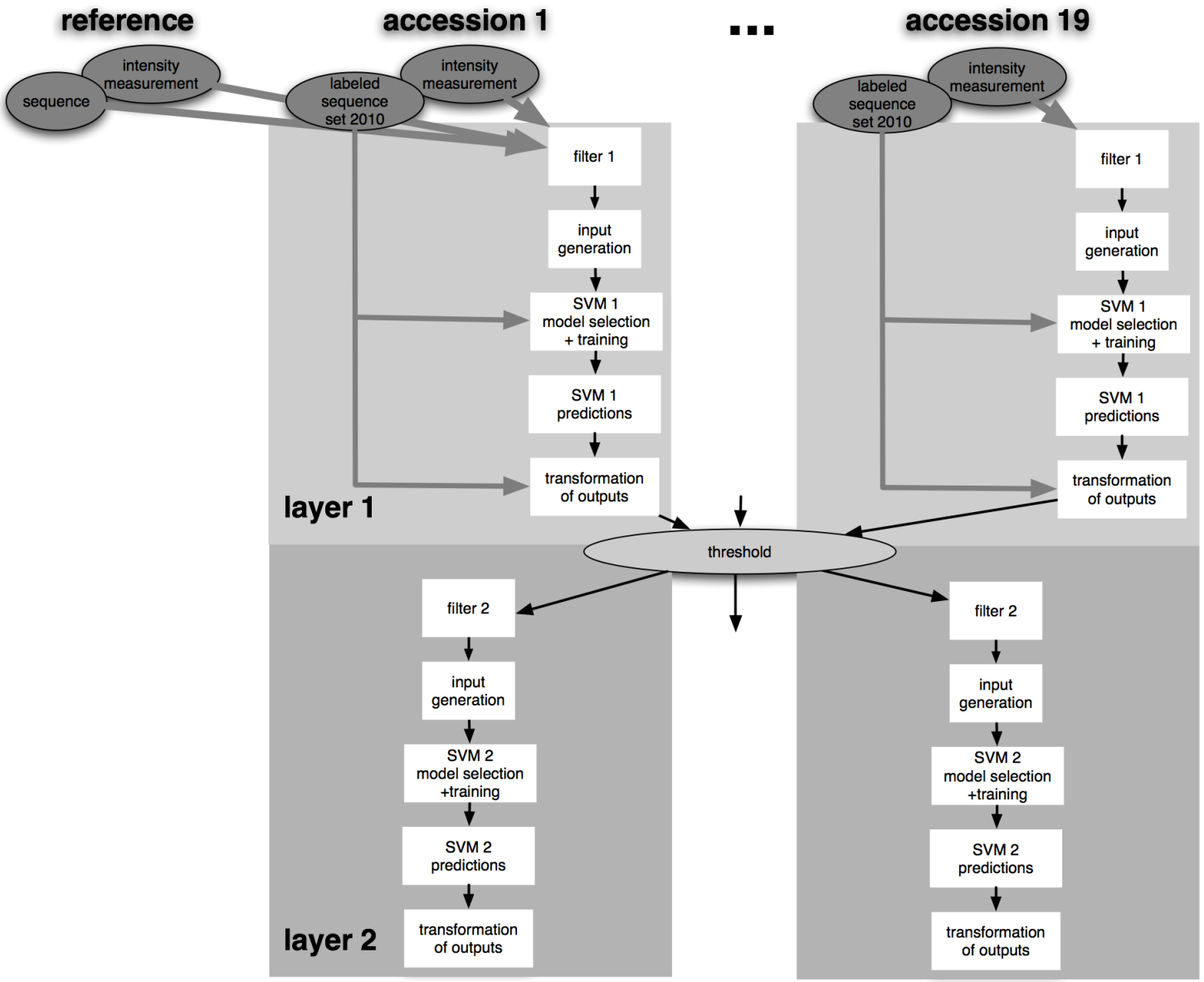
2-Layered Architecture for Inter-Strain Integration



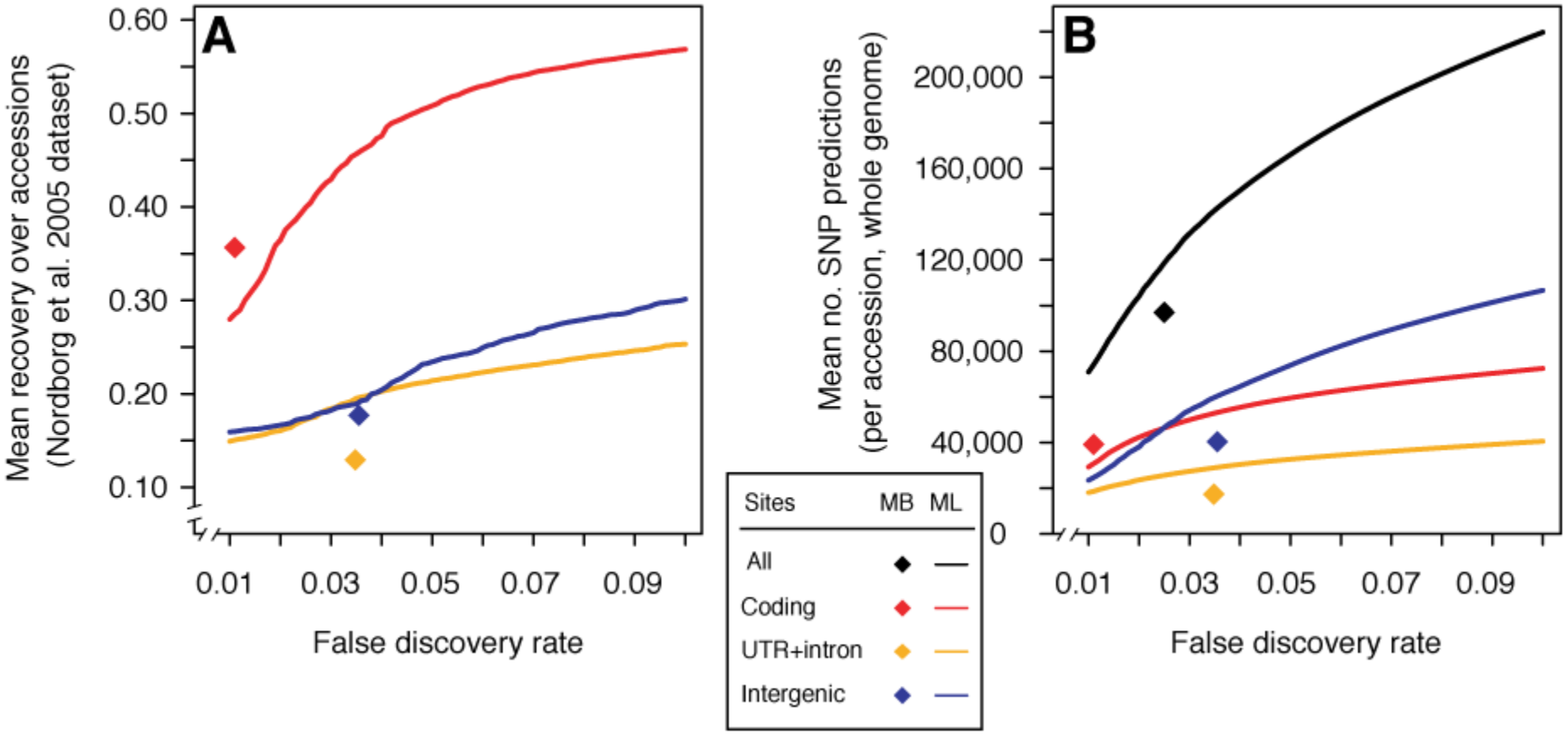
2-Layered Architecture for Inter-Strain Integration



2-Layered Architecture for Inter-Strain Integration



Application to SNP discovery



(Clark et al., Science, 2007)

A Relevant Fitness Trait: Biomass

Late-onset necrosis



Rmx-A180 (US)



Uod-7 (Austria)



Tamm-27 (Finland)



Lov-5 (Sweden)

(Slide provided by Detlef Weigel)

A Relevant Fitness Trait: Biomass



Rmx-A180
(US)



Uod-7
(Austria)



Tamm-27
(Finland)

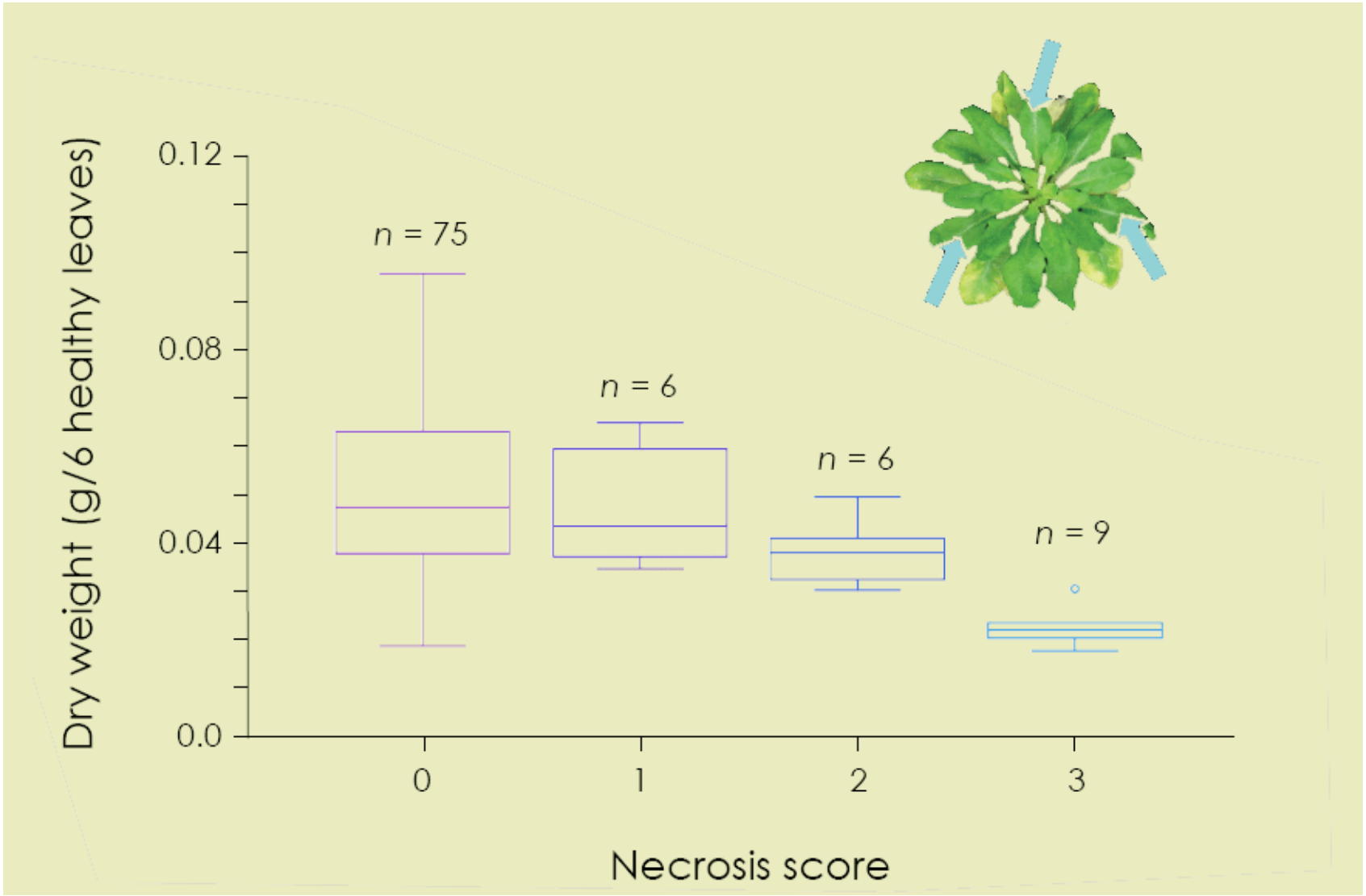
Late-onset
necrosis



Lov-5
(Sweden)

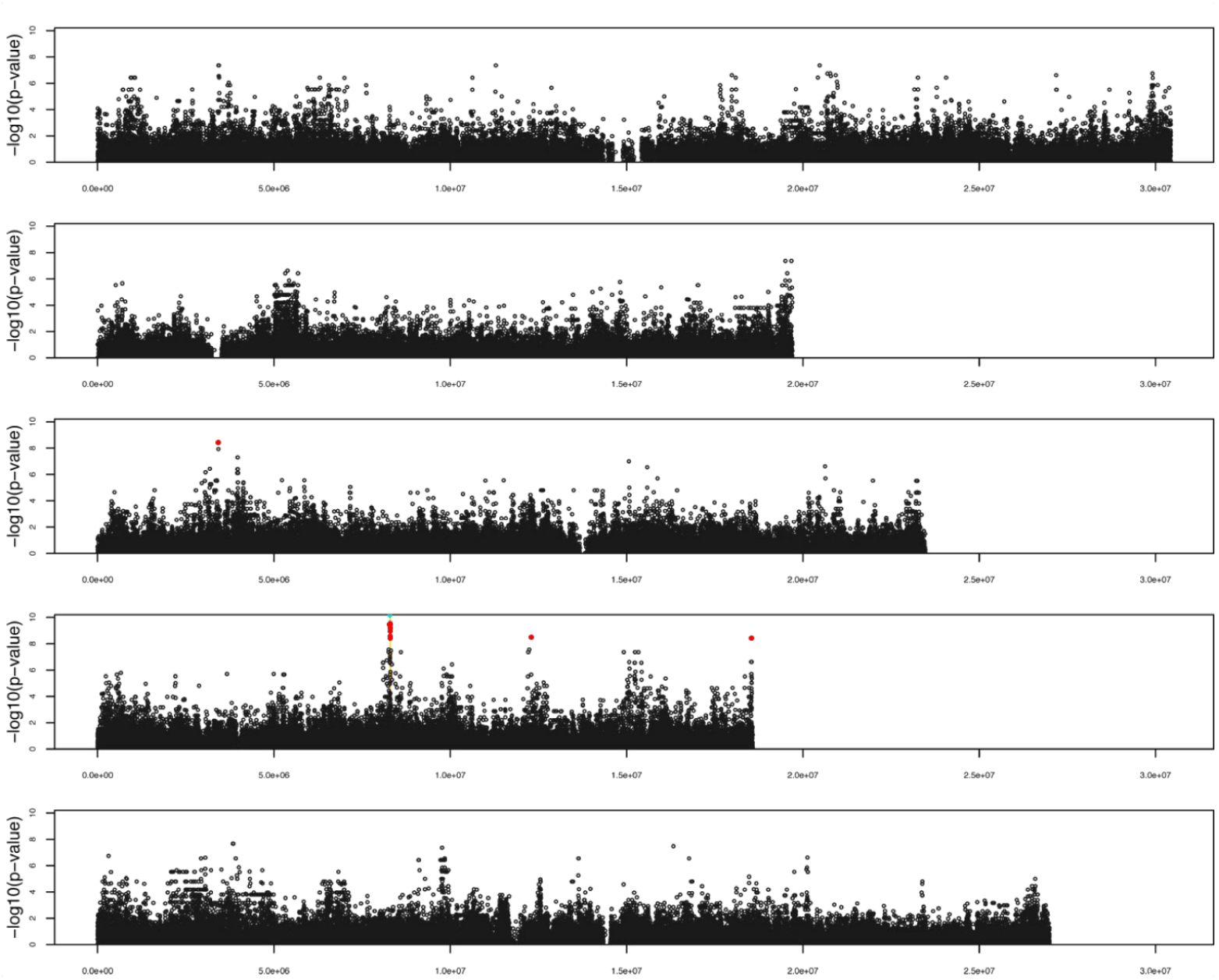
(Slide provided by Detlef Weigel)

Correlation Between Leaf Mass and Necrosis Score



(Slide provided by Detlef Weigel)

Genome-Wide Association Mapping of Necrosis

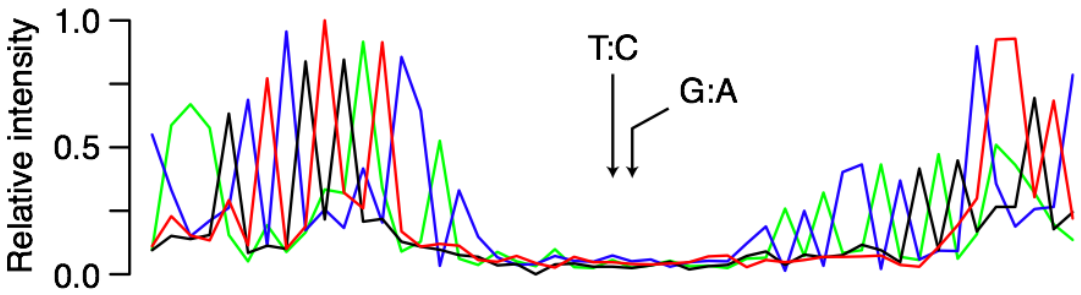


p-value from Cruscal-Wallis test

red: $p < 10^{-8}$

(Slide provided by Detlef Weigel)

Identification of Highly Polymorphic Regions



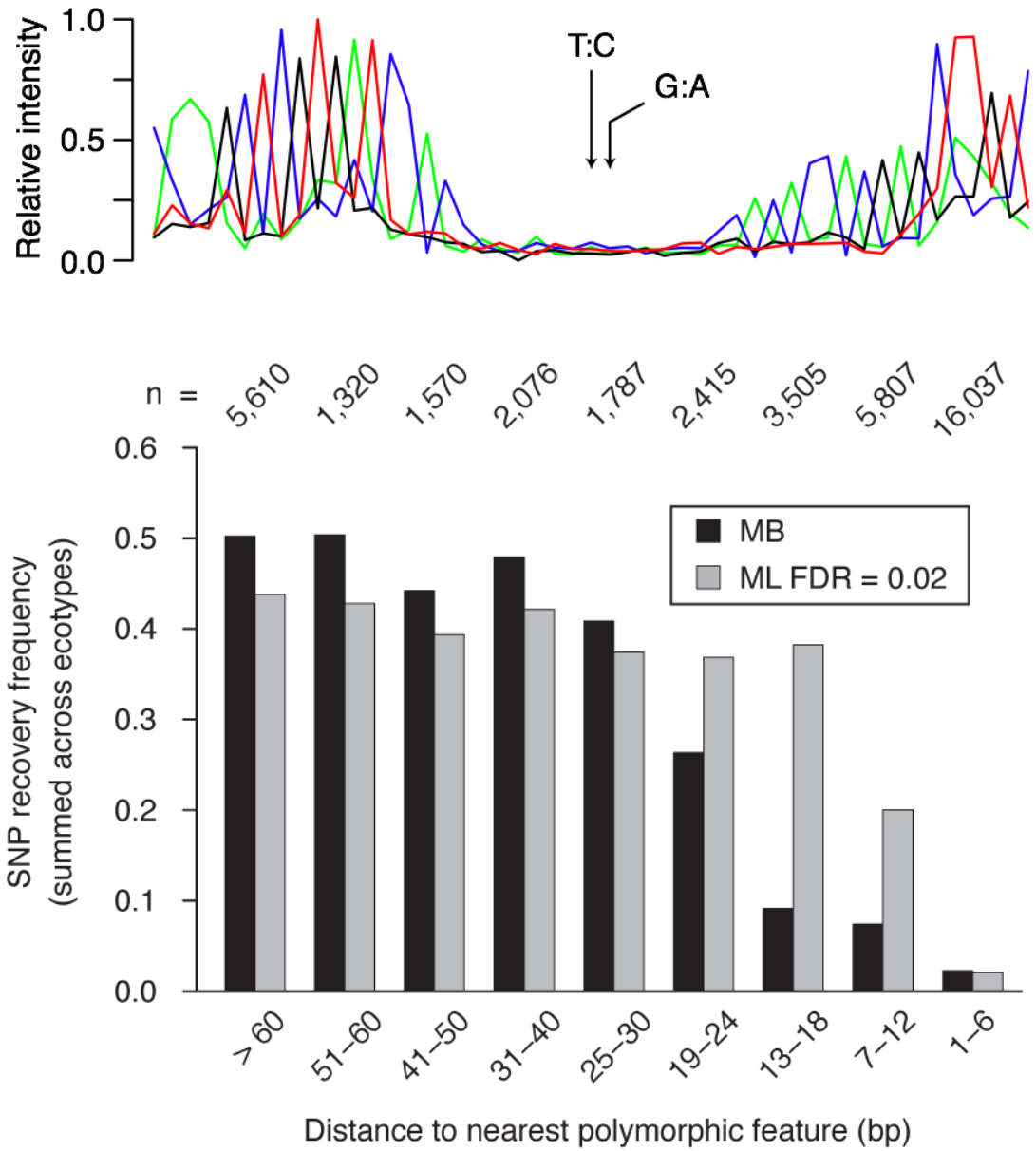
Results

- Performance drops, when other SNPs are in vicinity (1-20nt)
- Least predicted SNPs in highly polymorphic regions!
- ML more sensitive

New Approach

- **P**olymorphic **R**egion **P**rediction (PRP)
- Use HMSVM for segmenting the sequence

Identification of Highly Polymorphic Regions



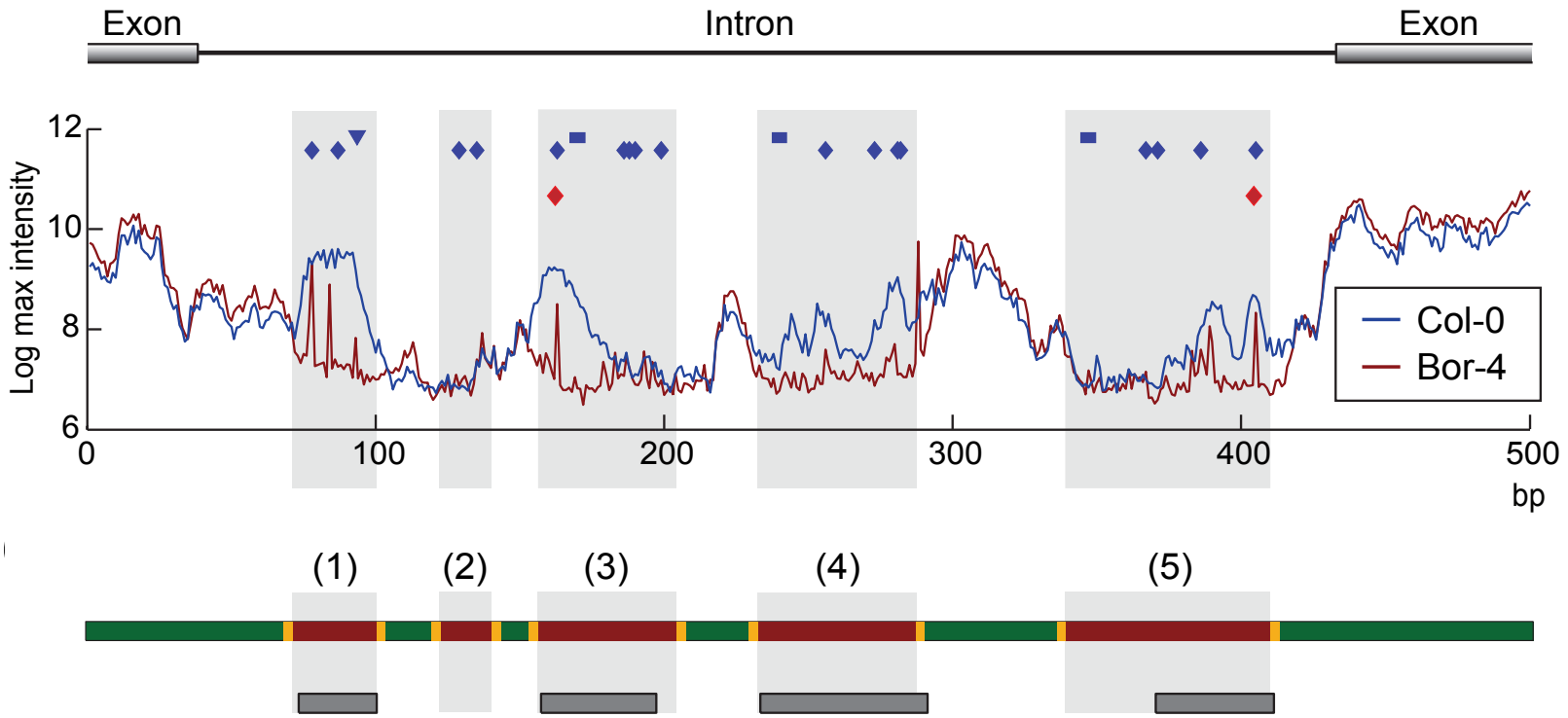
Results

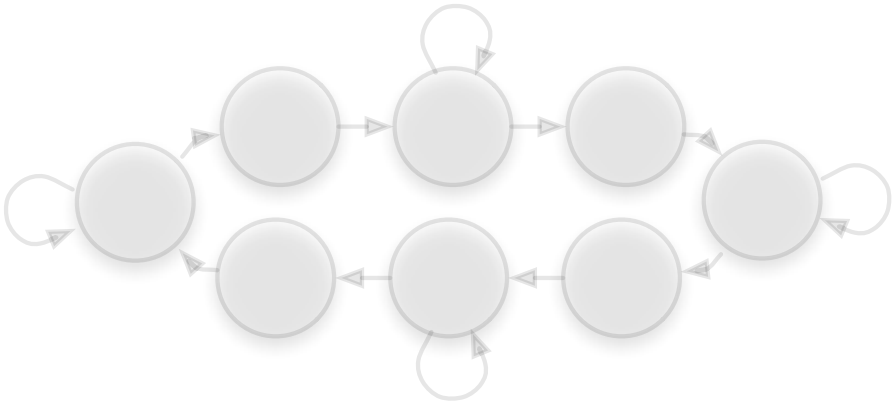
- Performance drops, when other SNPs are in vicinity (1-20nt)
- Least predicted SNPs in highly polymorphic regions!
- ML more sensitive

New Approach

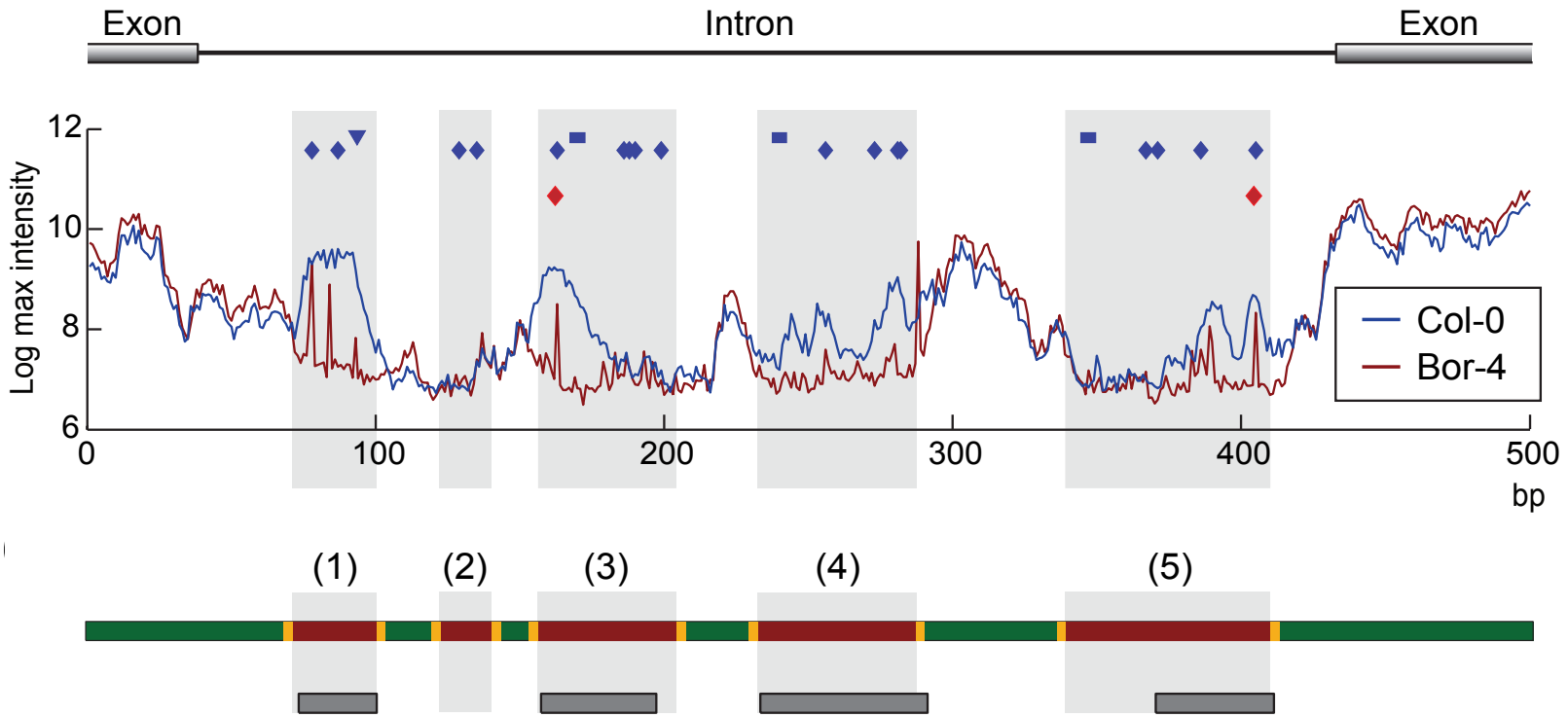
- **P**olymorphic **R**egion **P**rediction (PRP)
- Use HMSVM for segmenting the sequence

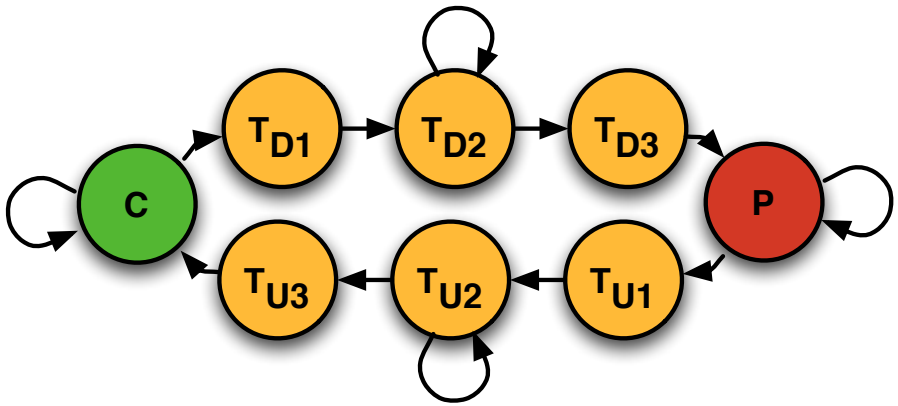
Modeling polymorphic regions





Modeling polymorphic regions





The Learning Problem

Given a sequence of observations (features) $x \in X$
We want to learn a function:

$$f : X \to Y$$

which yields a **label sequence (or path)** $\pi \in Y$
(of equal length: $|x| = |y|$).

Employ a function

$$F_\Theta : X \times Y \to \mathbf{R} \qquad \text{(path scoring)}$$

with which

$$f(x) = \arg\max_{\pi \in Y} F(x, \pi) \qquad \text{(Viterbi decoding)}$$

(Altun, ICML, 2003)

The Learning Problem

The discriminant function F_Θ

$$F_\Theta(x,\pi) = \sum_{p=1}^{L} \sum_{j,k} [[\pi_p = k]]\, g_{j,k}(x_{j,p}) + \phi(\pi_{p-1}, \pi_p)$$

- $g_{j,k}$: feature scoring contributions, for each pair of features $j \in 1,...,M$ and states $k \in 1,...,S$
- ϕ: transition scores

Large Margin Separation

$$F(x^i,\pi^i) - F(x^i,\overline{\pi}) \gg 0 \quad \forall \overline{\pi} \neq \pi^i \quad \forall i$$

between the correct labeling π^i
and any other (wrong) labeling $\overline{\pi}$.

The Learning Problem

The discriminant function F_Θ

$$F_\Theta(x,\pi) = \sum_{p=1}^{L} \sum_{j,k} [[\pi_p = k]]\, g_{j,k}(x_{j,p}) + \phi(\pi_{p-1}, \pi_p)$$

- $g_{j,k}$: feature scoring contributions, for each pair of features $j \in 1,...,M$ and states $k \in 1,...,S$
- ϕ: transition scores

Large Margin Separation

$$F(x^i,\pi^i) - F(x^i,\overline{\pi}) \gg 0 \quad \forall \overline{\pi} \neq \pi^i \quad \forall i$$

between the correct labeling π^i
and any other (wrong) labeling $\overline{\pi}$.

Training: Solve Linear Program

$$\min_{\theta,\, \xi \geq 0} \quad \frac{1}{n}\sum_{i=1}^{n} \xi^{(i)} + C\,\Omega(\theta)$$

$$\text{s.t.} \quad F_\theta(\mathbf{x}^{(i)}, \pi^{(i)}) - F_\theta(\mathbf{x}^{(i)}, \overline{\pi}) \;\geq\; 1 - \xi^{(i)}$$

$$\forall \overline{\pi} \neq \pi^{(i)} \quad \forall i = 1, \ldots, n,$$

- ξ: slack variables
- Ω: linear regularization term

Exponentially number of margin constraints !

- employ column generation technique

Evaluation

count prediction as

- true positive (TP), if it overlaps by $\geq 75\%$
- false positive (FP), else.

count known polymorphic region as

- true discovery (TD), if polymorphisms covered, or $\geq 75\%$ included in prediction
- false negative (FN), else.

56% Sensitivity, 90% Specificity

(Zeller et al., Genome Research, 2008)

Evaluation

count prediction as

- true positive (TP), if it overlaps by $\geq 75\%$
- false positive (FP), else.

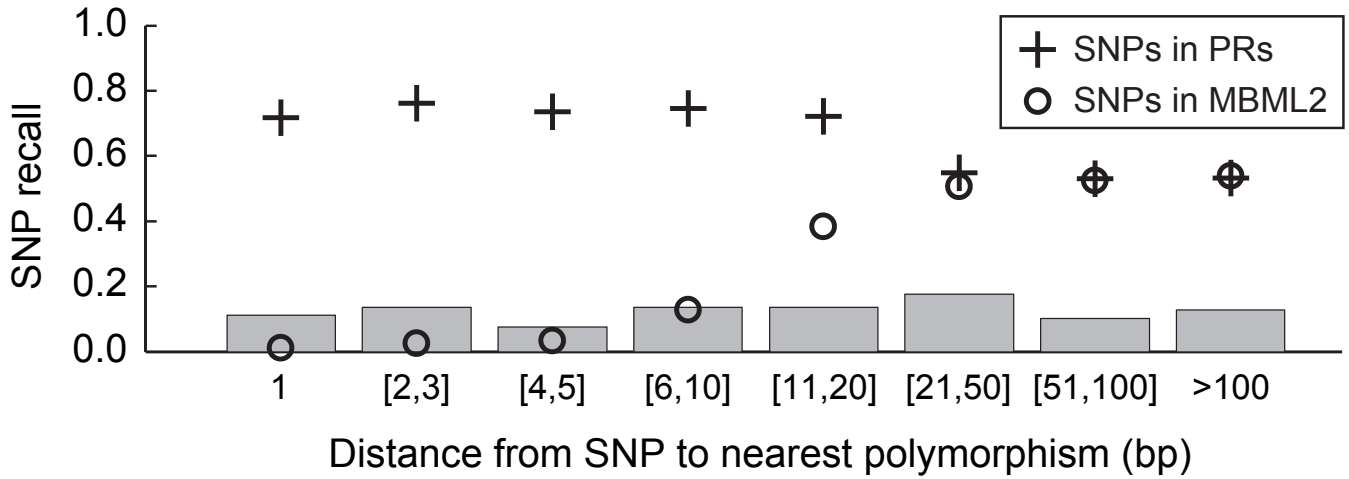
count known polymorphic region as

- true discovery (TD), if polymorphisms covered, or $\geq 75\%$ included in prediction
- false negative (FN), else.

56% Sensitivity, 90% Specificity

(Zeller et al., Genome Research, 2008)

Complementing SNP Calls



Fraction of called/covered polymorphisms (test set):

	SNP calling (MB+ML)	Region predictor
SNPs	∼32%	∼65%
Deletions (per base)		∼53%
Insertions		∼39%

(Zeller et al., Genome Research, 2008)

Effects on Genes

SNPs (MB+ML) leading to consensus changes

- 110,000 amino acid changes
- >1,200 premature stops
- 200 stop codon changes
- >150 initiation methionine changes
- >400 splice site changes
- Major changes in 573 genes validated by dideoxy sequencing*

Polymorphic Regions Predictions

- Overlap coding regions of >26 000 genes in at least one accession
- 25% overlap in >8000 genes, 90% overlap in >500 genes
- 122 coding sequence deletions validated by dideoxy sequencing*

* Verified in collaboration with laboratory of Joseph Ecker

Effects on Genes

SNPs (MB+ML) leading to consensus changes

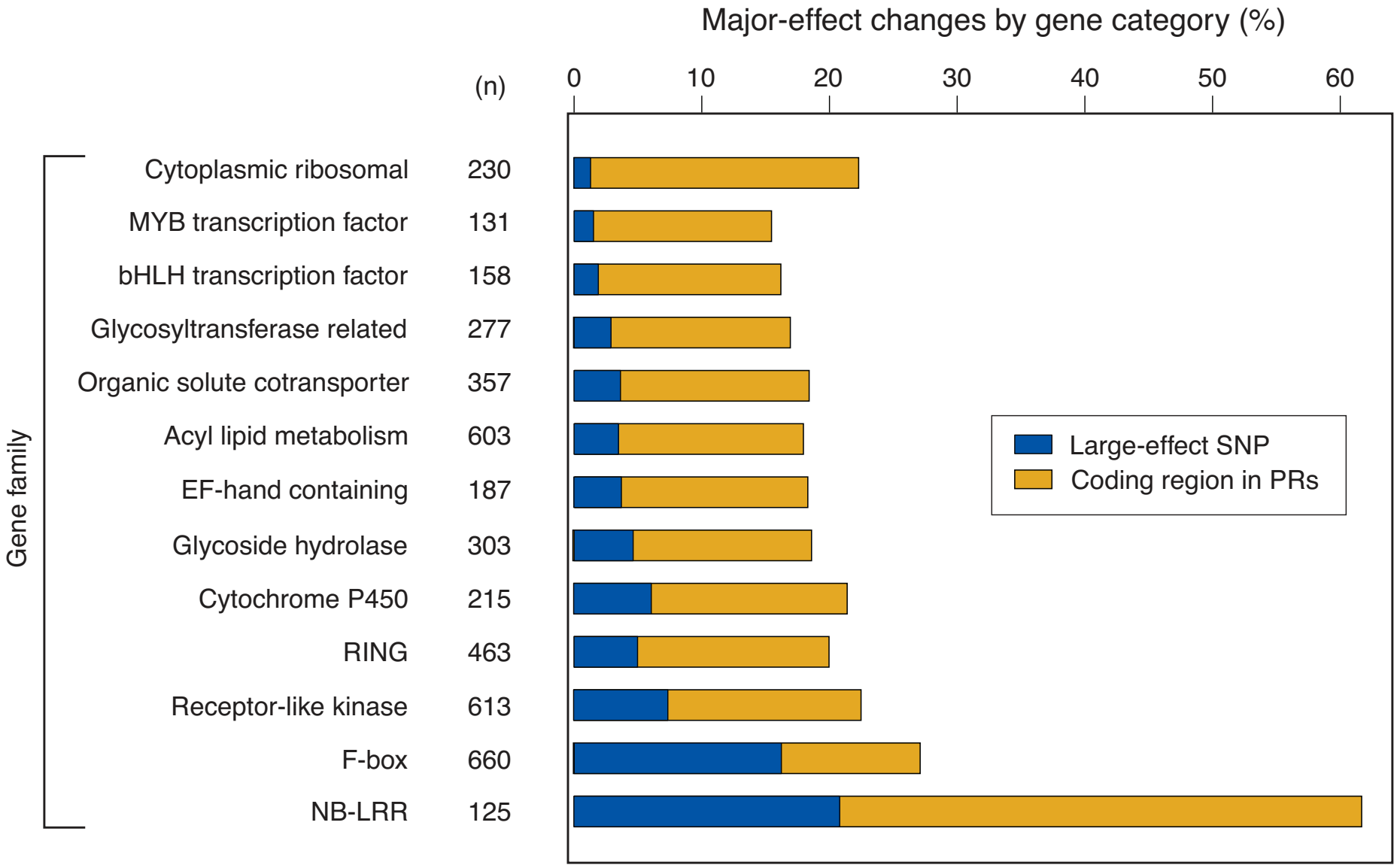
- 110,000 amino acid changes
- >1,200 premature stops
- 200 stop codon changes
- >150 initiation methionine changes
- >400 splice site changes
- Major changes in 573 genes validated by dideoxy sequencing*

Polymorphic Regions Predictions

- Overlap coding regions of >26 000 genes in at least one accession
- 25% overlap in >8000 genes, 90% overlap in >500 genes
- 122 coding sequence deletions validated by dideoxy sequencing*

* Verified in collaboration with laboratory of Joseph Ecker

Major Effects Distribution



Conclusions

- Created inventory for polymorphisms in *Arabidopsis thaliana*
  - Highly polymorphic: ≈0.5% in SNPs, ≈27% in PRPs
- Accurate polymorphic region predictions
  - Important for further analysis (e.g. dideoxy sequencing)
- Large number of major effect changes
  - Overrepresented in R genes, F-box genes, Receptor-like kinases
- Application to other genomes (rice, mouse, human)
- Study variations using
  - mRNA tiling arrays
  - next generation sequencing technology

  ⇒ Expression levels, splicing, . . .

Conclusions

- Created inventory for polymorphisms in *Arabidopsis thaliana*
  - Highly polymorphic: ≈0.5% in SNPs, ≈27% in PRPs
- Accurate polymorphic region predictions
  - Important for further analysis (e.g. dideoxy sequencing)
- Large number of major effect changes
  - Overrepresented in R genes, F-box genes, Receptor-like kinases
- Application to other genomes (rice, mouse, human)
- Study variations using
  - mRNA tiling arrays
  - next generation sequencing technology

  ⇒ Expression levels, splicing, . . .

Conclusions

- Created inventory for polymorphisms in *Arabidopsis thaliana*
  - Highly polymorphic: ≈0.5% in SNPs, ≈27% in PRPs
- Accurate polymorphic region predictions
  - Important for further analysis (e.g. dideoxy sequencing)
- Large number of major effect changes
  - Overrepresented in R genes, F-box genes, Receptor-like kinases
- Application to other genomes (rice, mouse, human)
- Study variations using
  - mRNA tiling arrays
  - next generation sequencing technology

  ⇒ Expression levels, splicing, . . .

Conclusions

- Created inventory for polymorphisms in *Arabidopsis thaliana*
  - Highly polymorphic: ≈0.5% in SNPs, ≈27% in PRPs
- Accurate polymorphic region predictions
  - Important for further analysis (e.g. dideoxy sequencing)
- Large number of major effect changes
  - Overrepresented in R genes, F-box genes, Receptor-like kinases
- Application to other genomes (rice, mouse, human)
- Study variations using
  - mRNA tiling arrays
  - next generation sequencing technology

  $\Rightarrow$ Expression levels, splicing, ...

Acknowledgments

Friedrich Miescher Laboratory
- Georg Zeller
- Anja Bohlen

MPI for Biological Cybernetics
- Bernhard Schölkopf
- Gabriele Schweikert

The Salk Institute, CA
- Paul Shinn
- Joseph Ecker

USC, CA
- Magnus Nordberg

MPI for Developmental Biology
- Richard Clark
- Stephan Ossowski
- Korbinian Schneeberger
- Norman Warthmann
- Detlef Weigel

ZBIT, University of Tübingen
- Daniel Huson

Perlegen Sciences, CA
- Kelly Frazer

Any Questions?



Any Questions?

