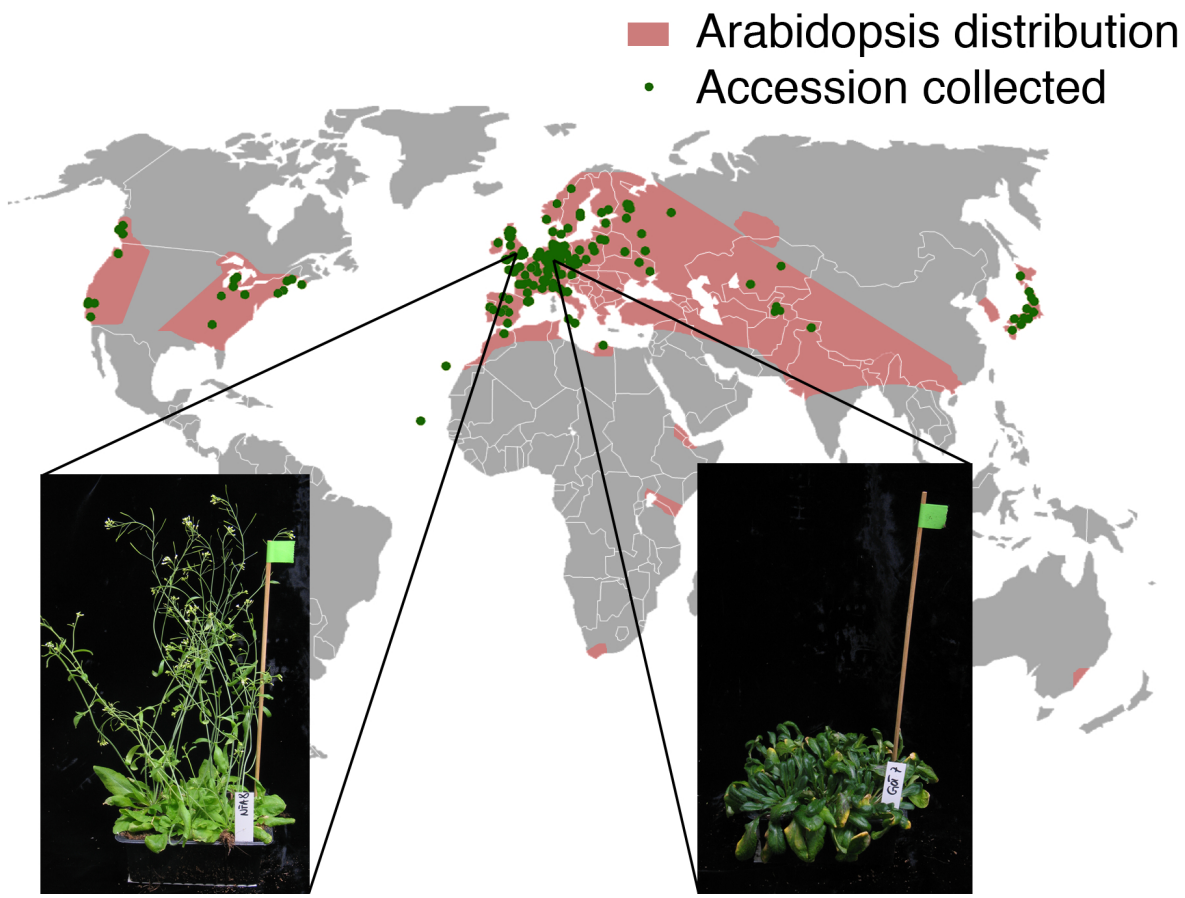
Discovering Common Sequence Variation in *Arabidopsis thaliana*

Gunnar Rätsch

Friedrich Miescher Laboratory
Max Planck Society
Tübingen

ISMB Highlights Track
July 22, 2008

Analysis of Polymorphisms: Introduction



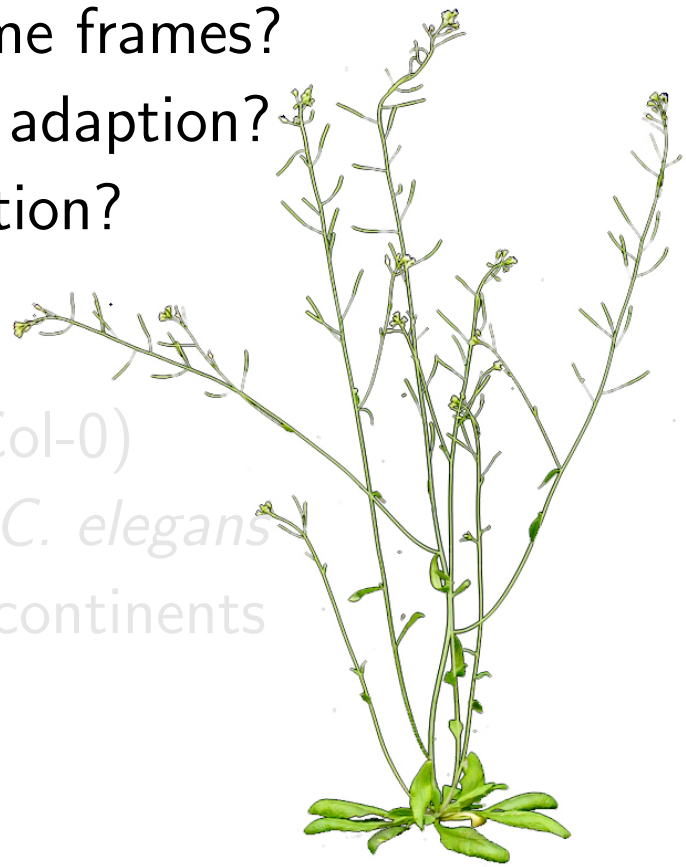
What is the genetic basis of variation?

Modified from Koornneef et al. [2004]

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

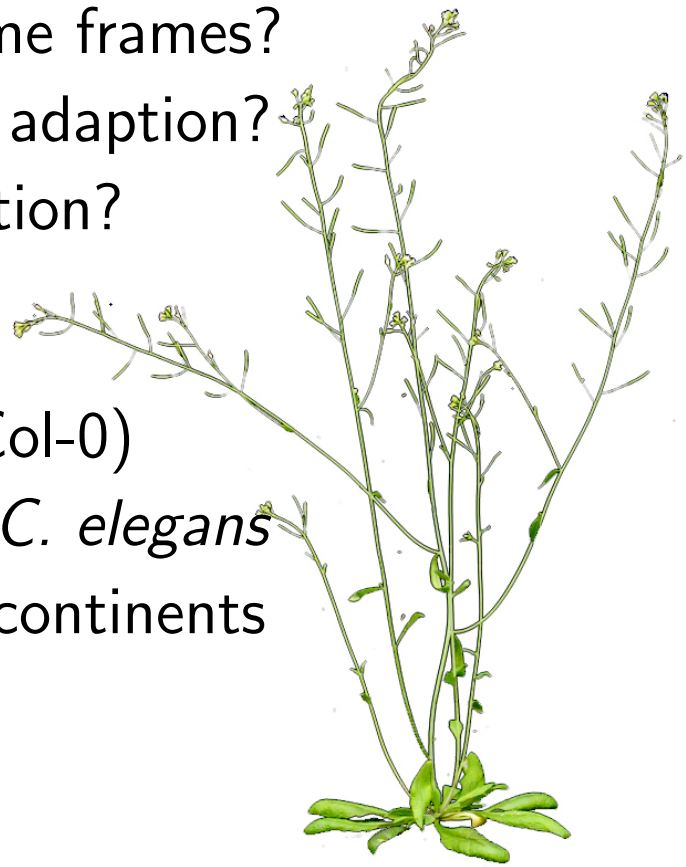
Resequencing of 20 wild strains

- 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

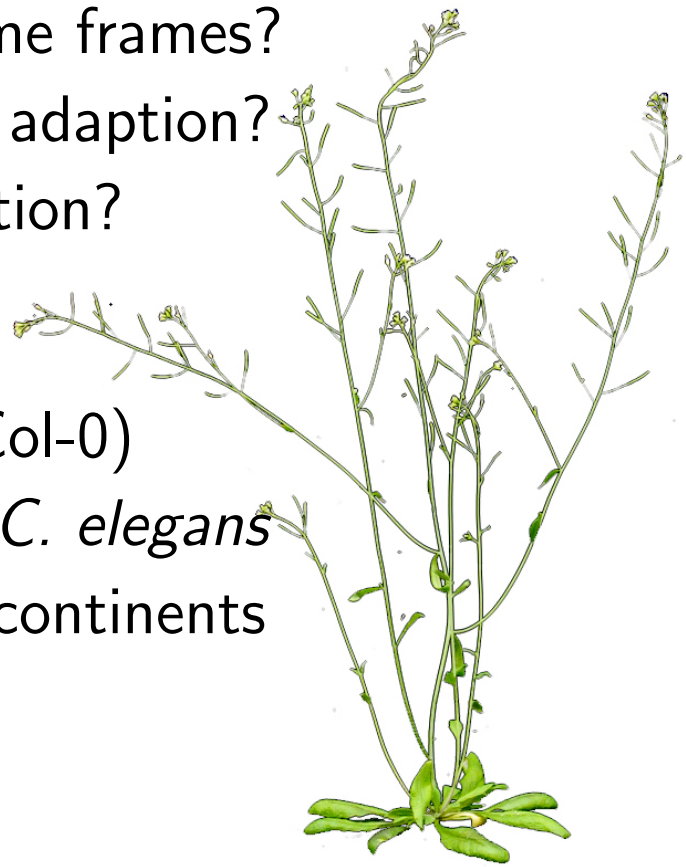
Resequencing of 20 wild strains

- 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

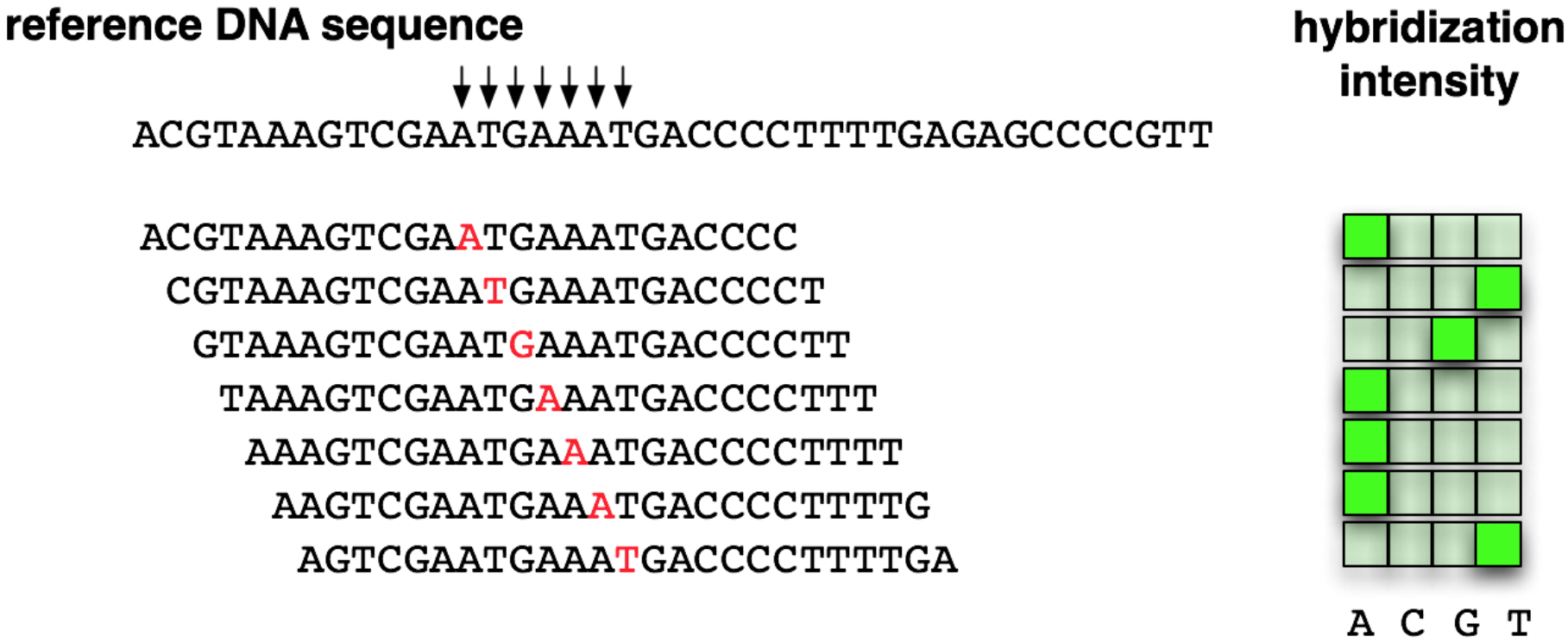
Resequencing of 20 wild strains

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

Resequencing Array Basics I

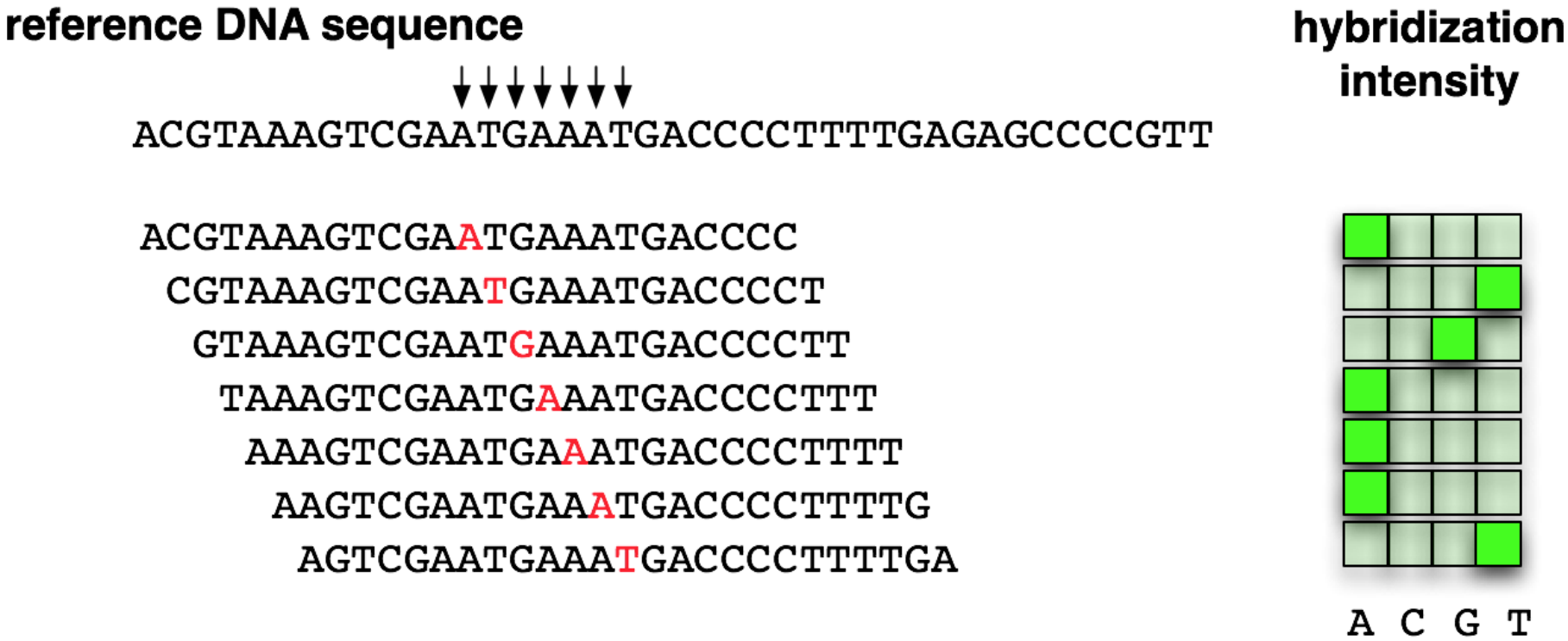


Resequencing Array Basics II



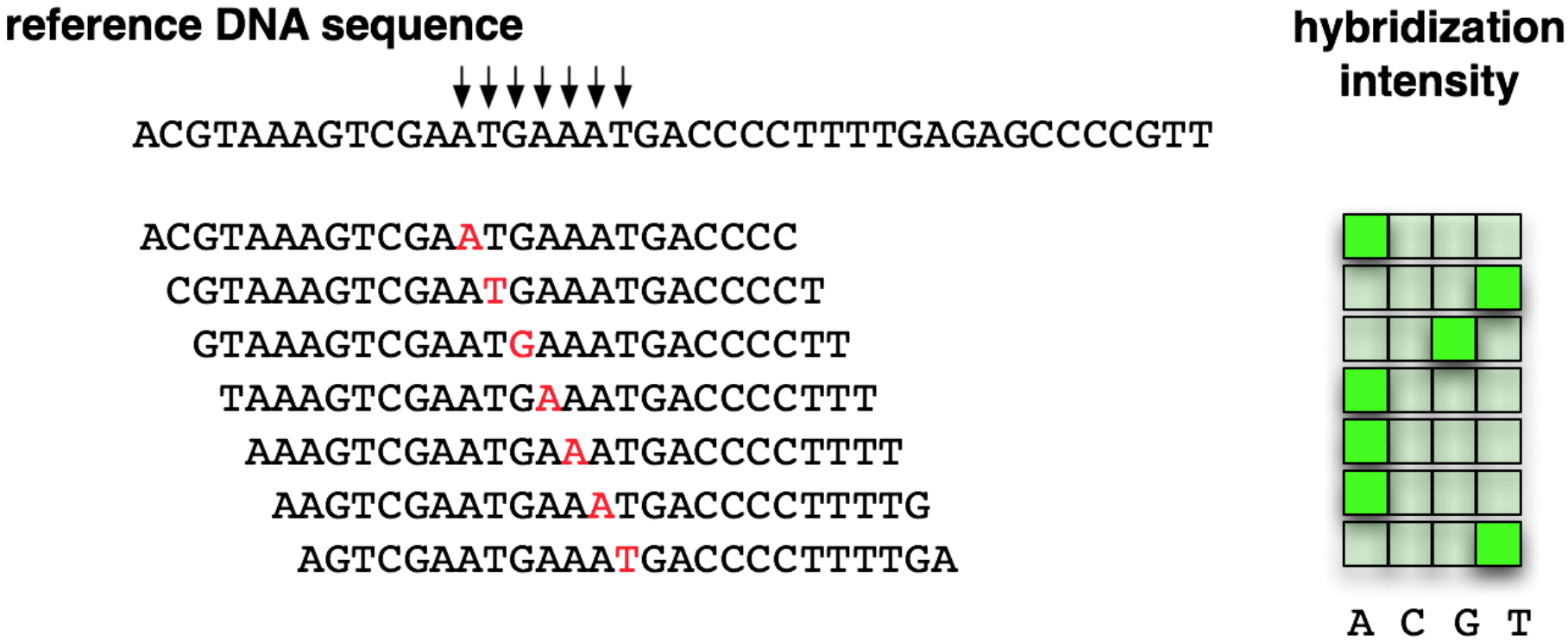
- Each base queried with forward and reverse strand probe quartets
- Nearly 1 billion oligos per accession
- 19+1 accessions surveyed representing worldwide distribution
- **Big datasets!**

Resequencing Array Basics II



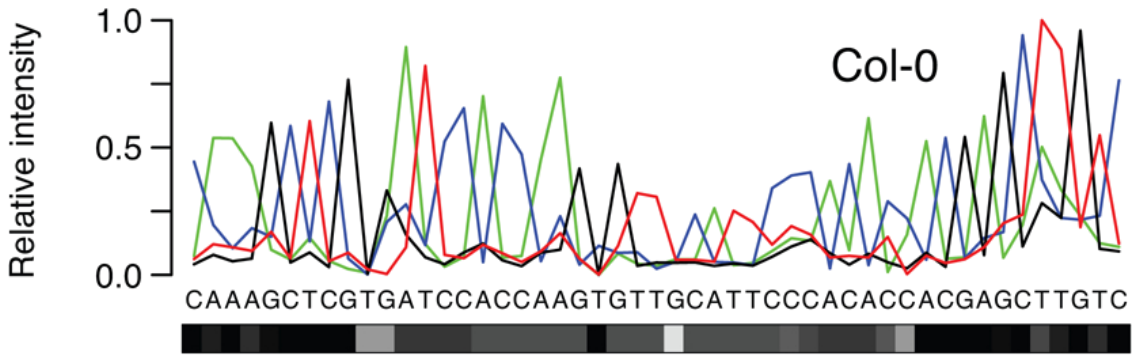
- Each base queried with forward and reverse strand probe quartets
- Nearly 1 billion oligos per accession
- 19+1 accessions surveyed representing worldwide distribution
- Big datasets!

Resequencing Array Basics II



- Each base queried with forward and reverse strand probe quartets
- Nearly 1 billion oligos per accession
- 19+1 accessions surveyed representing worldwide distribution
- **Big datasets!**

Resequencing Data



Data analysis challenge

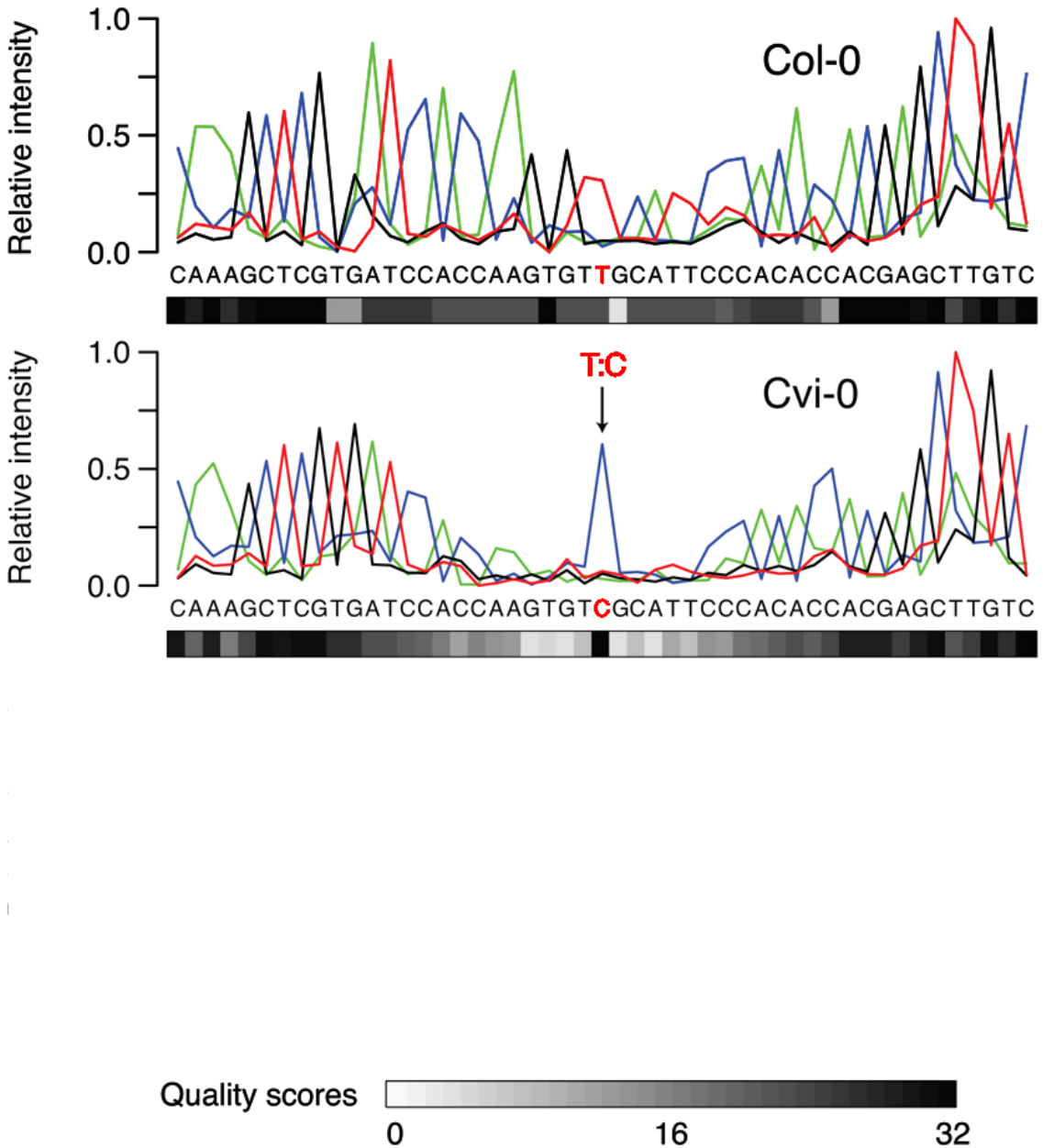
- Hybridization intensities depend on
  - Oligomer
  - Accession
  - Repeats
- Measurement noise
- Identify SNPs

Problematic cases

- Highly polymorphic regions
- Deletions/insertions



Resequencing Data



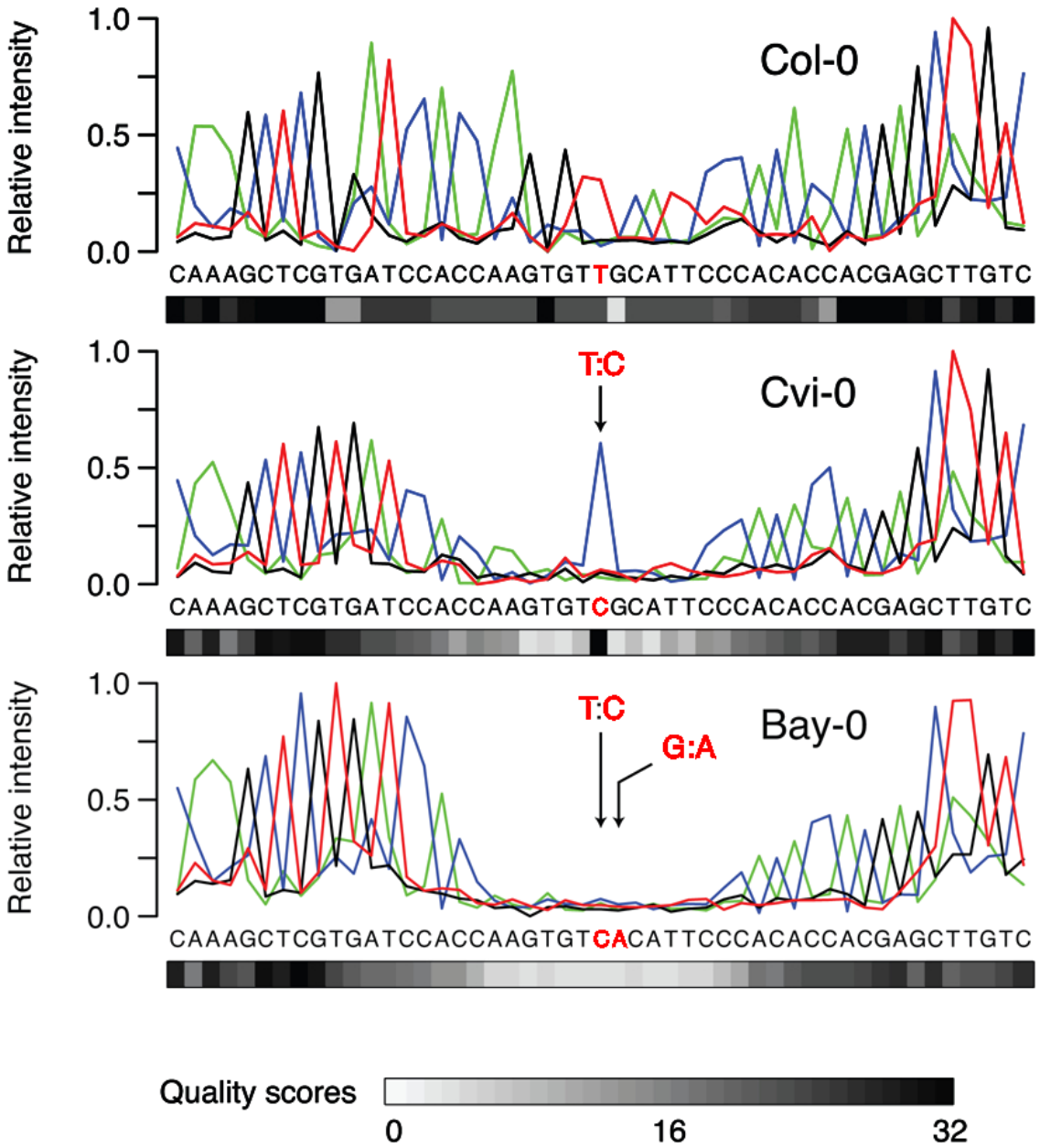
Data analysis challenge

- 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

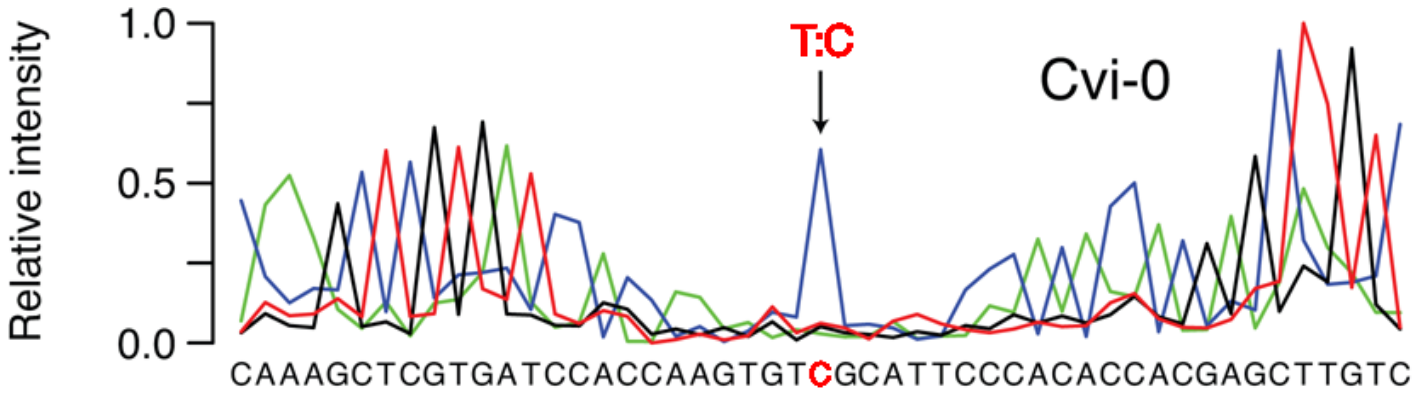
Labeled Data and SVM-Based Classification

Labeled training set

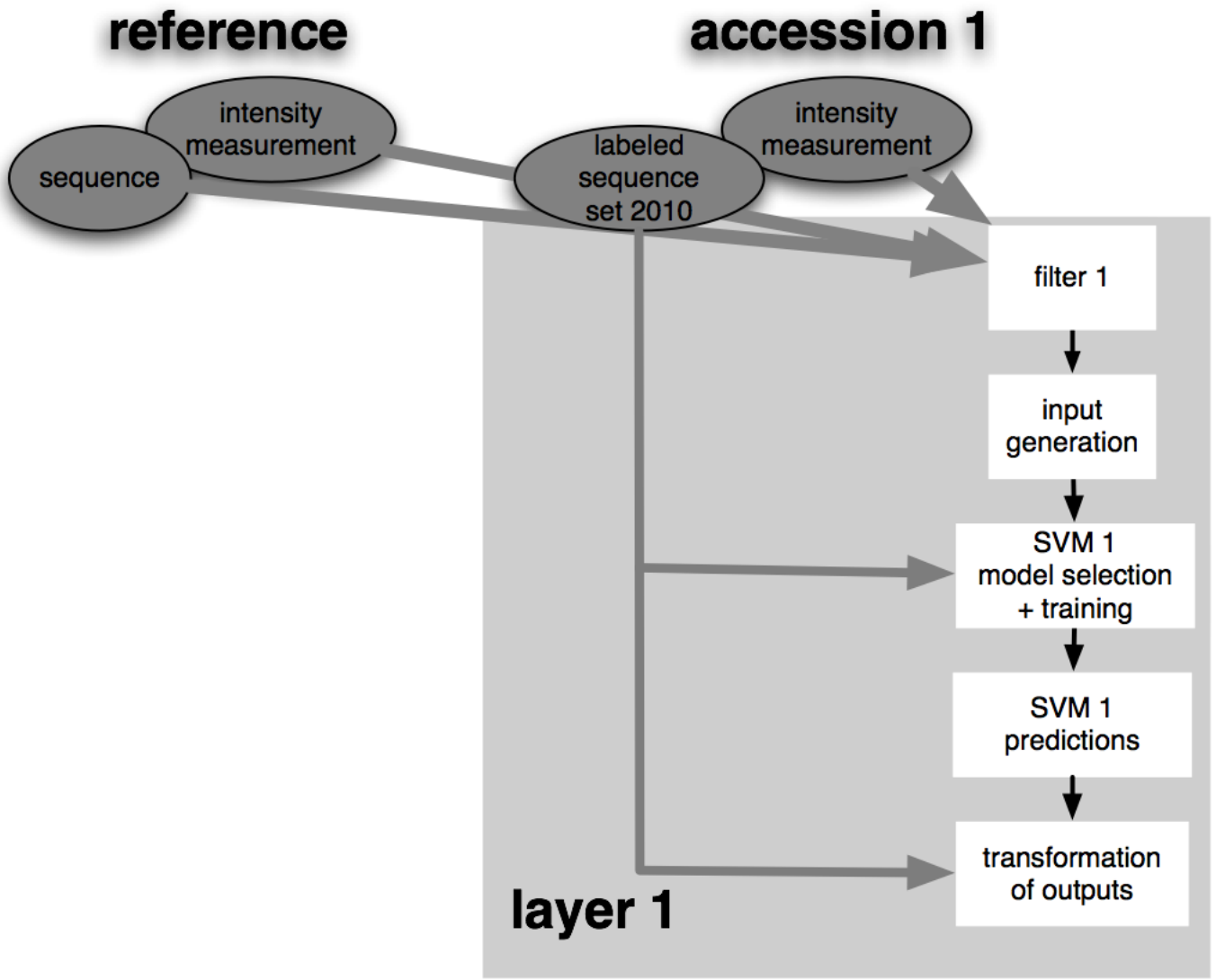
- 1,213 fragments of length $\approx$ 550bp for 19 accessions
- $\approx$2,700 known SNPs/accession [Nordborg et al., 2005]
- $\approx$400 indel polymorphisms/accession

Classification Method

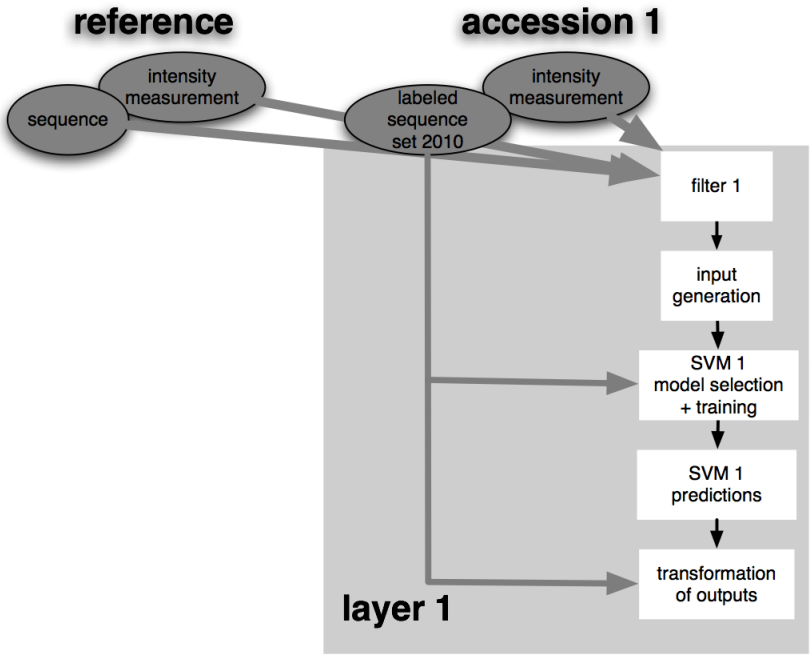
- Classification using SVMs with 302 features (ML)
- Two-layered approach: including cross-accession features
- Comparison with Perlegen's method (MB) [Hinds et al., 2005]



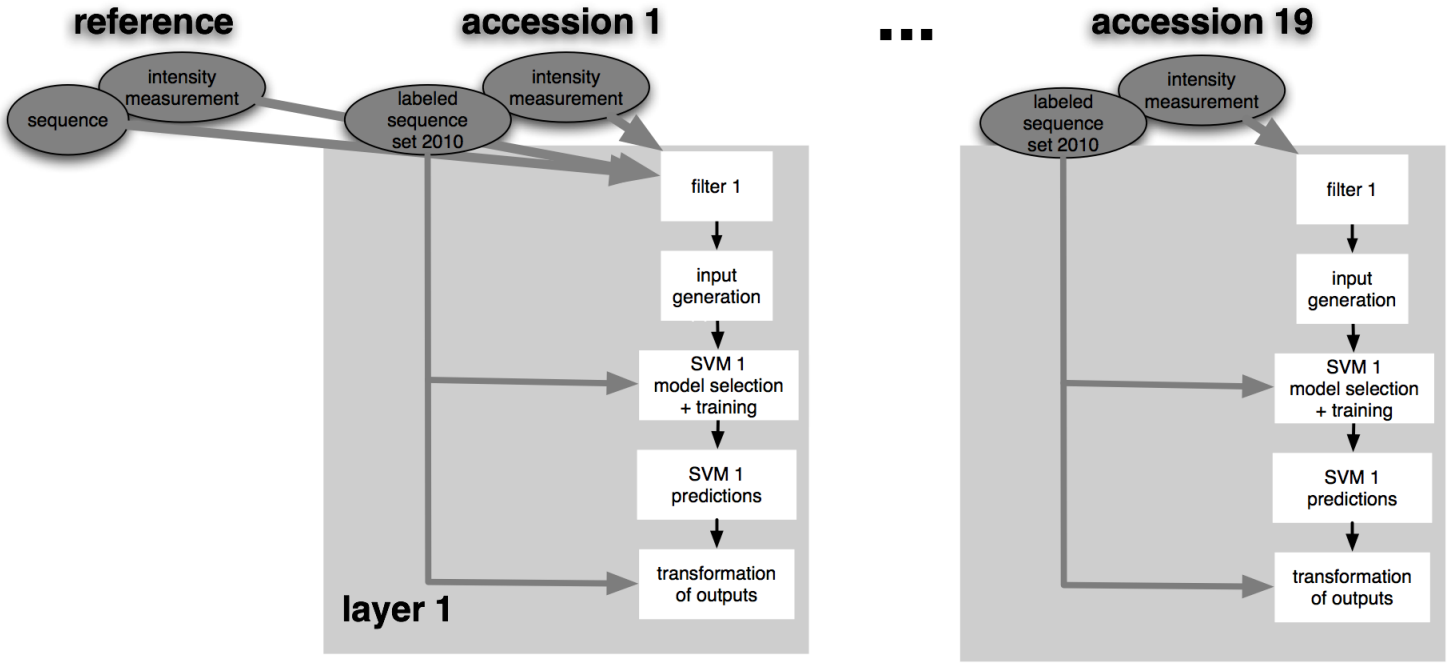
2-Layered Architecture for Inter-Strain Integration



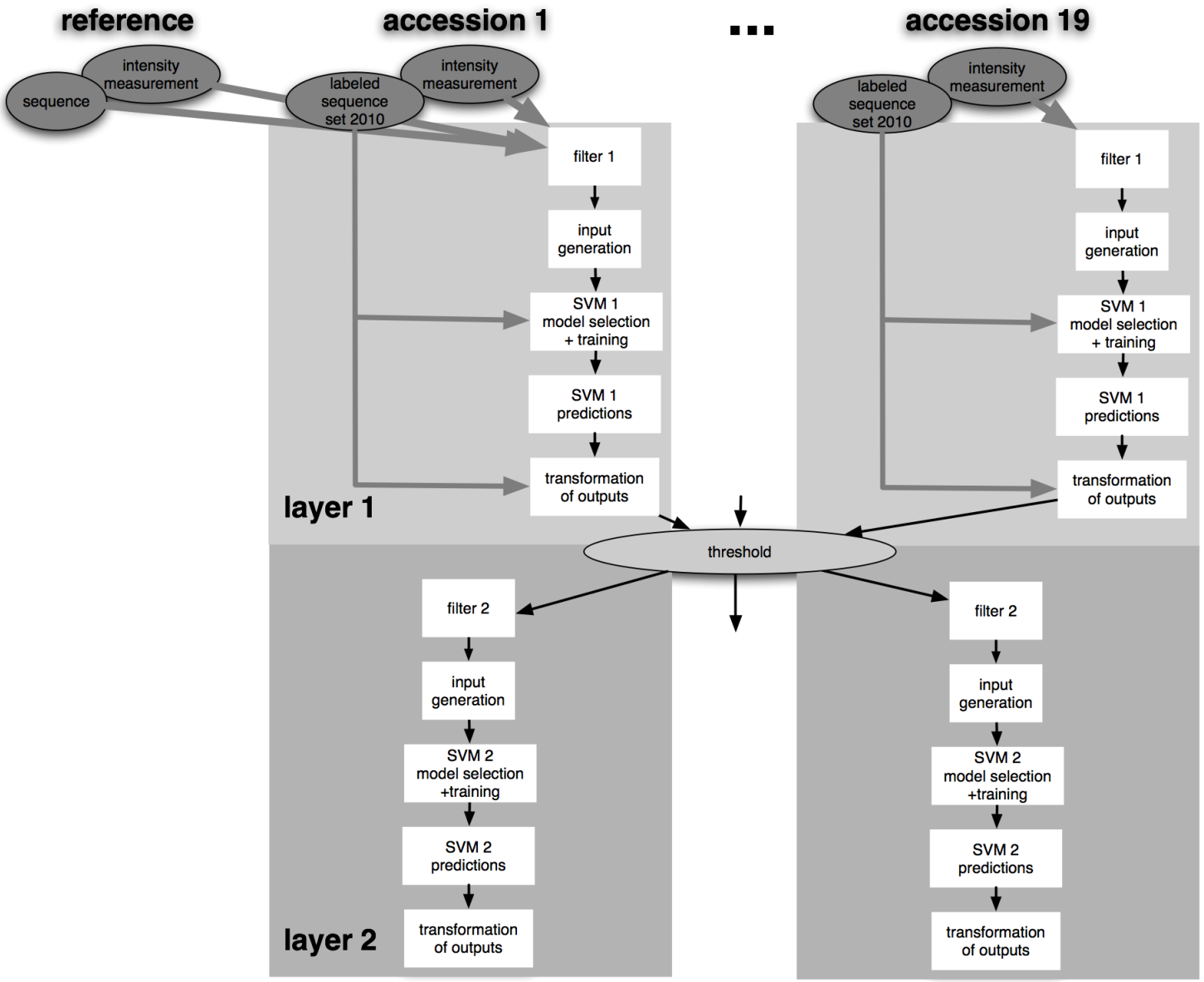
2-Layered Architecture for Inter-Strain Integration



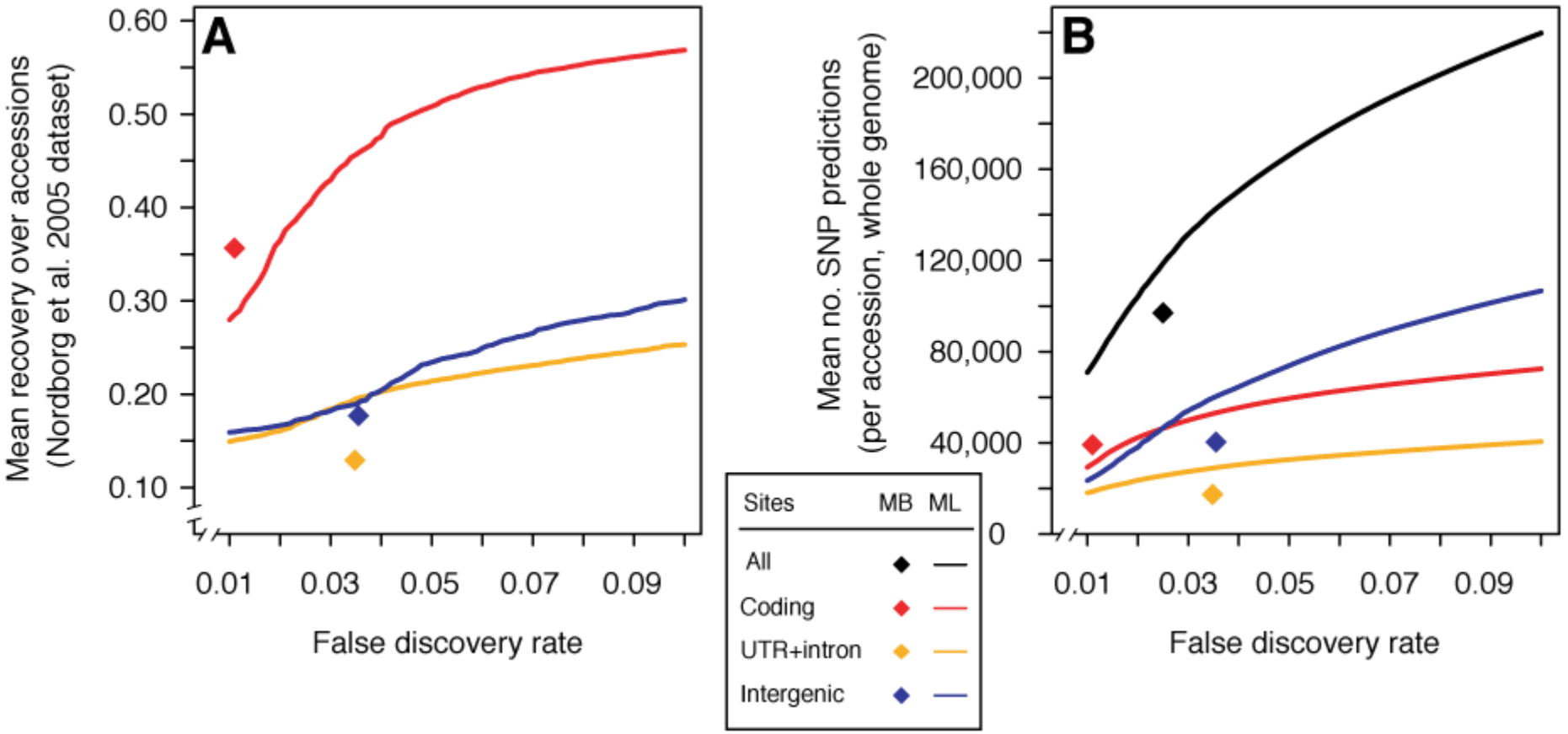
2-Layered Architecture for Inter-Strain Integration



2-Layered Architecture for Inter-Strain Integration

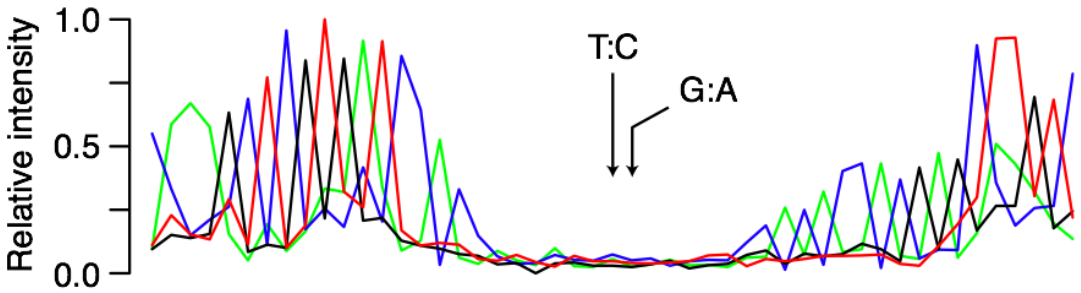


Application to SNP discovery



[Clark et al., 2007]

Identification of Highly Polymorphic Regions



Results

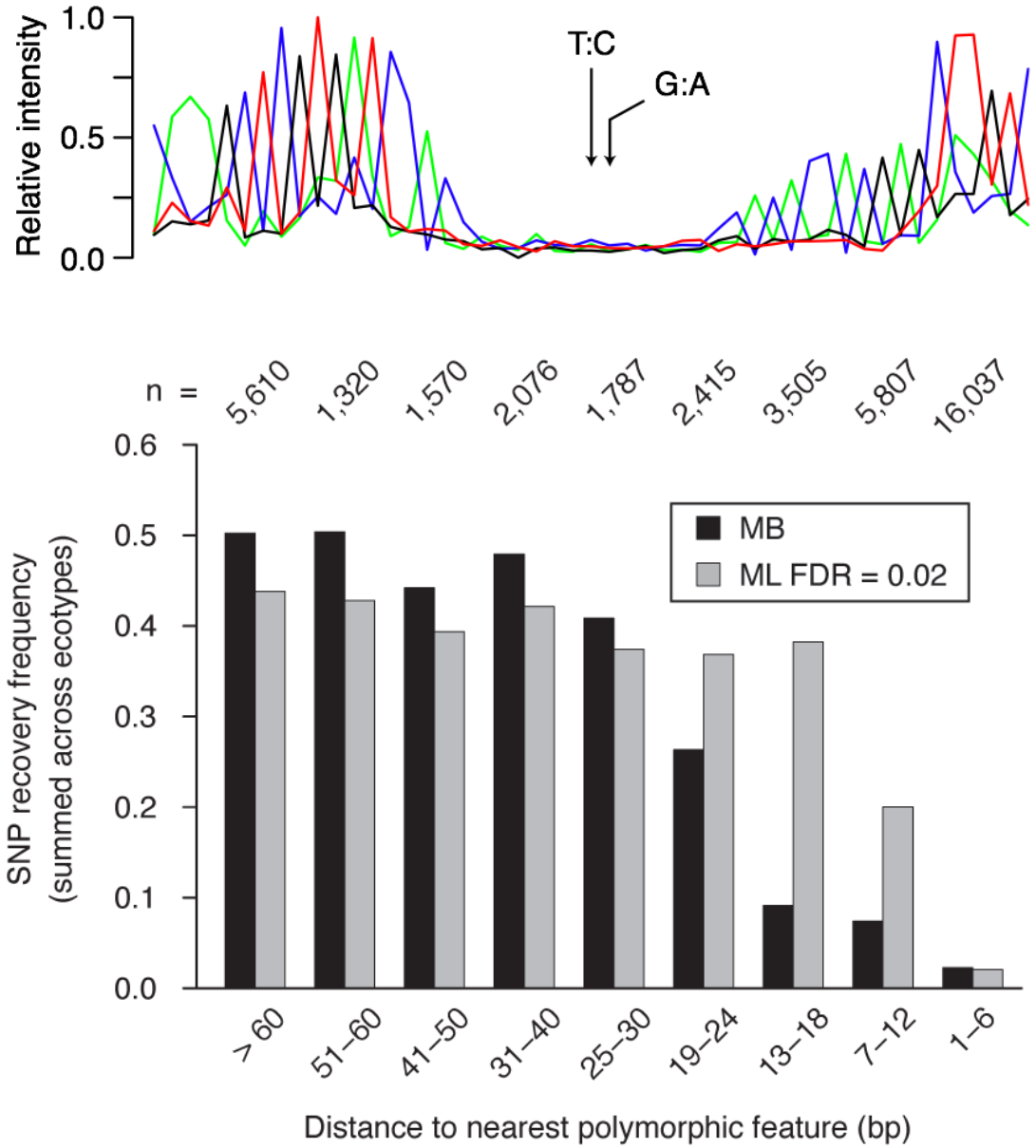
- Performance drops, when other SNPs are in vicinity (1-20nt)
- Least predicted SNPs in highly polymorphic regions!
- ML more sensitive

New Approach

- Learn to segment into polymorphic and conserved regions

$\Rightarrow$ **P**olymorphic **R**egion **P**rediction (PRP)

Identification of Highly Polymorphic Regions



Results

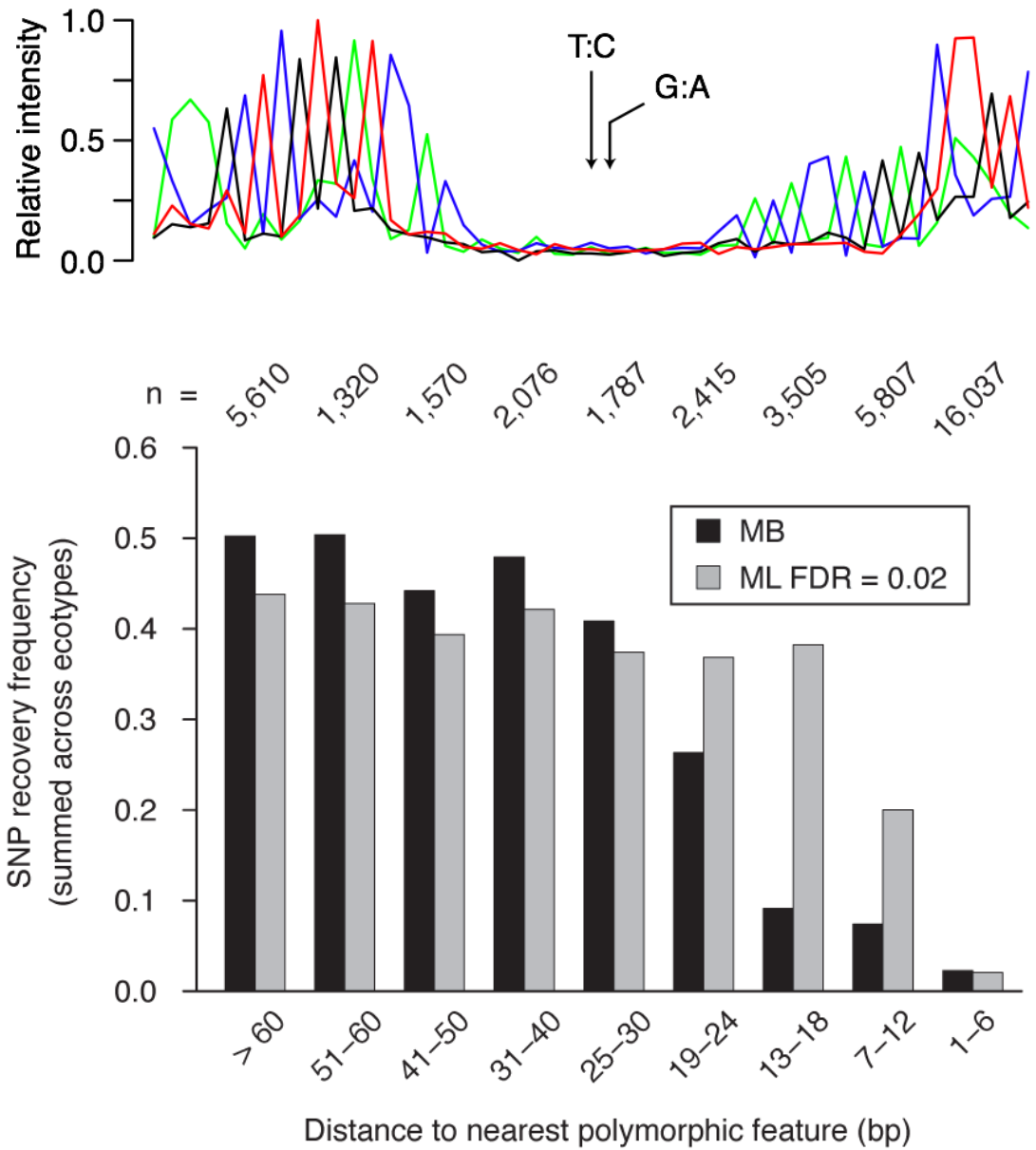
- Performance drops, when other SNPs are in vicinity (1-20nt)
- Least predicted SNPs in highly polymorphic regions!
- ML more sensitive

New Approach

- Learn to segment into polymorphic and conserved regions

$\Rightarrow$ **P**olymorphic **R**egion **P**rediction (PRP)

Identification of Highly Polymorphic Regions



Results

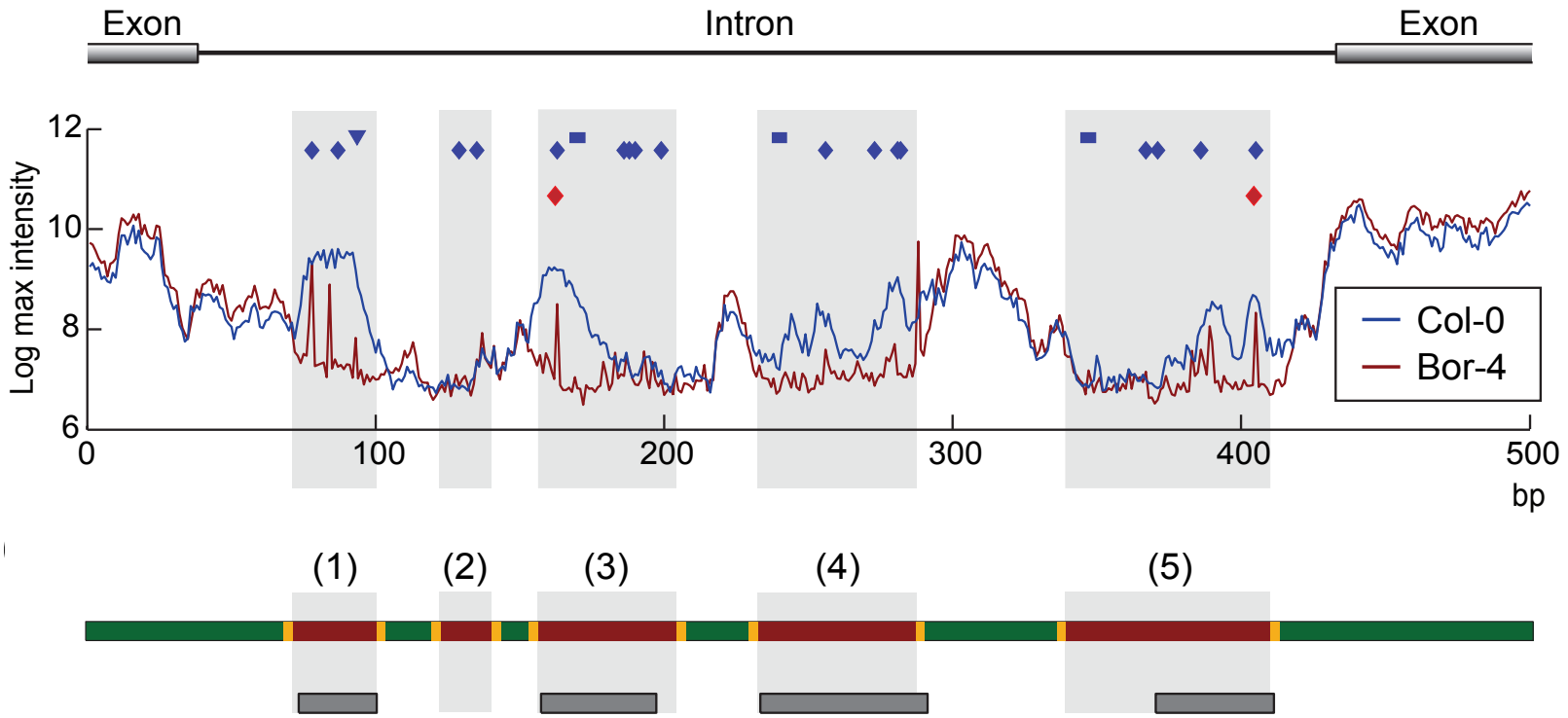
- Performance drops, when other SNPs are in vicinity (1-20nt)
- Least predicted SNPs in highly polymorphic regions!
- ML more sensitive

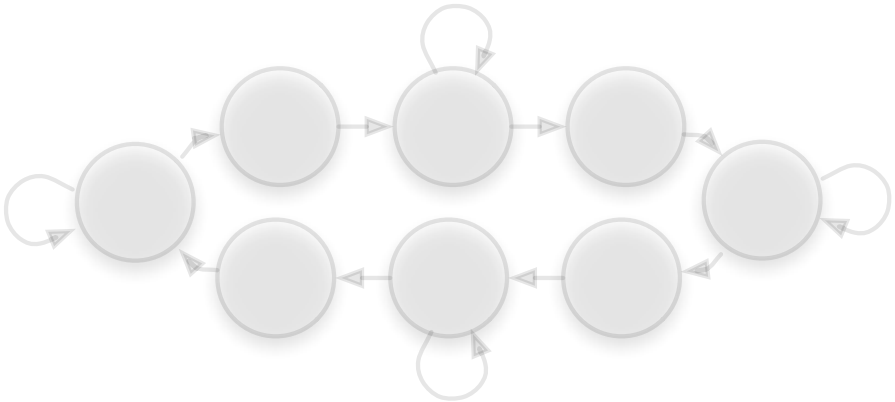
New Approach

- Learn to segment into polymorphic and conserved regions

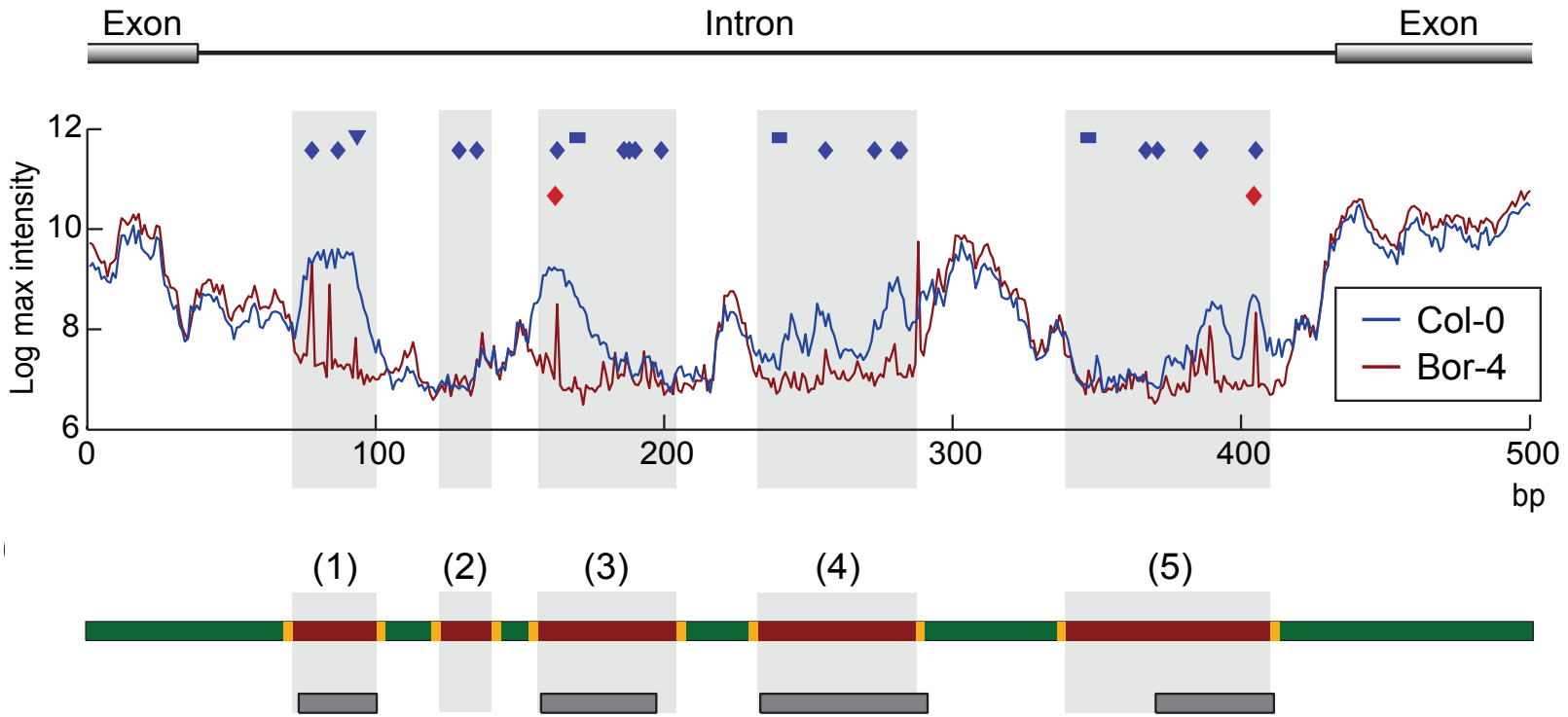
$\Rightarrow$ **P**olymorphic **R**egion **P**rediction (PRP)

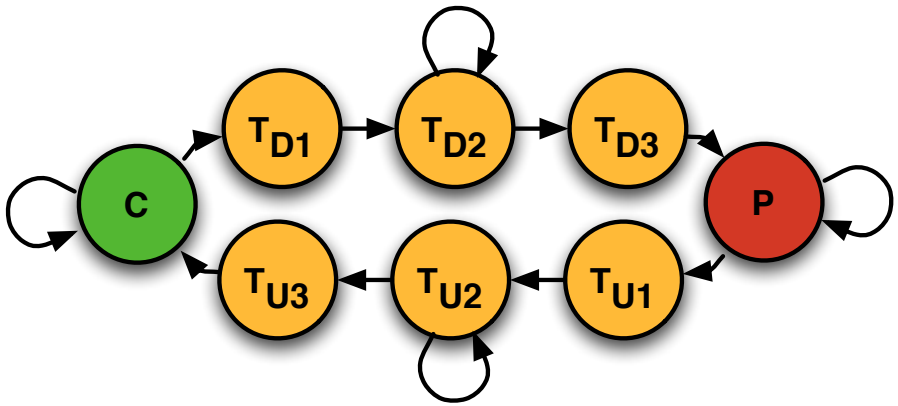
Modeling polymorphic regions





Modeling polymorphic regions





The Learning Problem

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

$$f : X \rightarrow Y$$

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

Employ a function

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

to be used as:

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

[Altun et al., 2003]

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 $\overline{\pi}$.

Solve Linear Optimization Problem

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

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

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

- ξ: slack variables
- Ω: linear regularization term

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 $\overline{\pi}$.

Solve Linear Optimization Problem

$$\begin{aligned} \min_{\boldsymbol{\theta},\, \boldsymbol{\xi} \geq \mathbf{0}} \quad & \frac{1}{n} \sum_{i=1}^{n} \xi^{(i)} + C\, \Omega(\boldsymbol{\theta}) \\ \text{s.t.} \quad & F_\theta(\mathbf{x}^{(i)}, \boldsymbol{\pi}^{(i)}) - F_\theta(\mathbf{x}^{(i)}, \overline{\boldsymbol{\pi}}) \;\geq\; 1 - \xi^{(i)} \\ & \forall \overline{\boldsymbol{\pi}} \neq \boldsymbol{\pi}^{(i)} \quad \forall i = 1, \ldots, n, \end{aligned}$$

- $\boldsymbol{\xi}$: slack variables
- Ω: linear regularization term

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., 2008]

Evaluation

Count prediction as

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

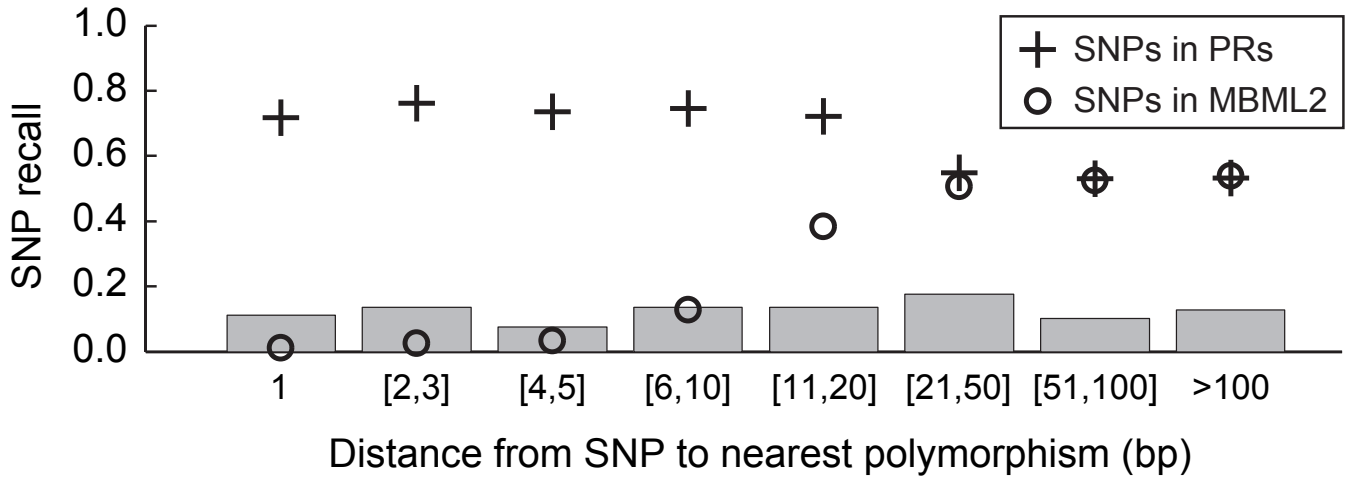
Count known polymorphic region as

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

56% Sensitivity, 90% Specificity

[Zeller et al., 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., 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

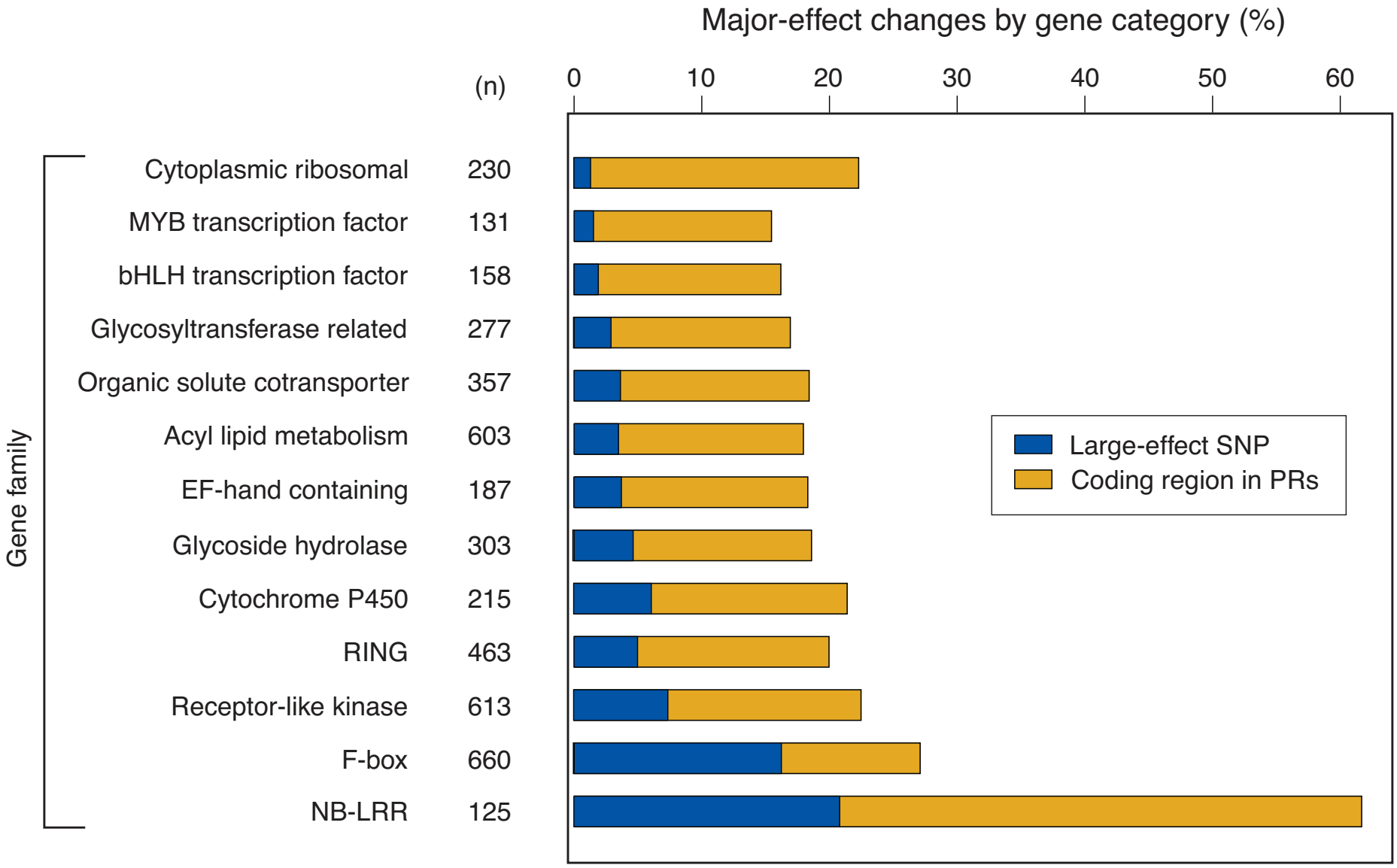
- 110,000 amino acid changes
- >1,200 premature stops
- 200 stop codon changes
- >150 initiation methionine changes
- >400 splice site changes
- Major changes in 573 genes validated by dideoxy sequencing*

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

Distribution of Major Effects



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

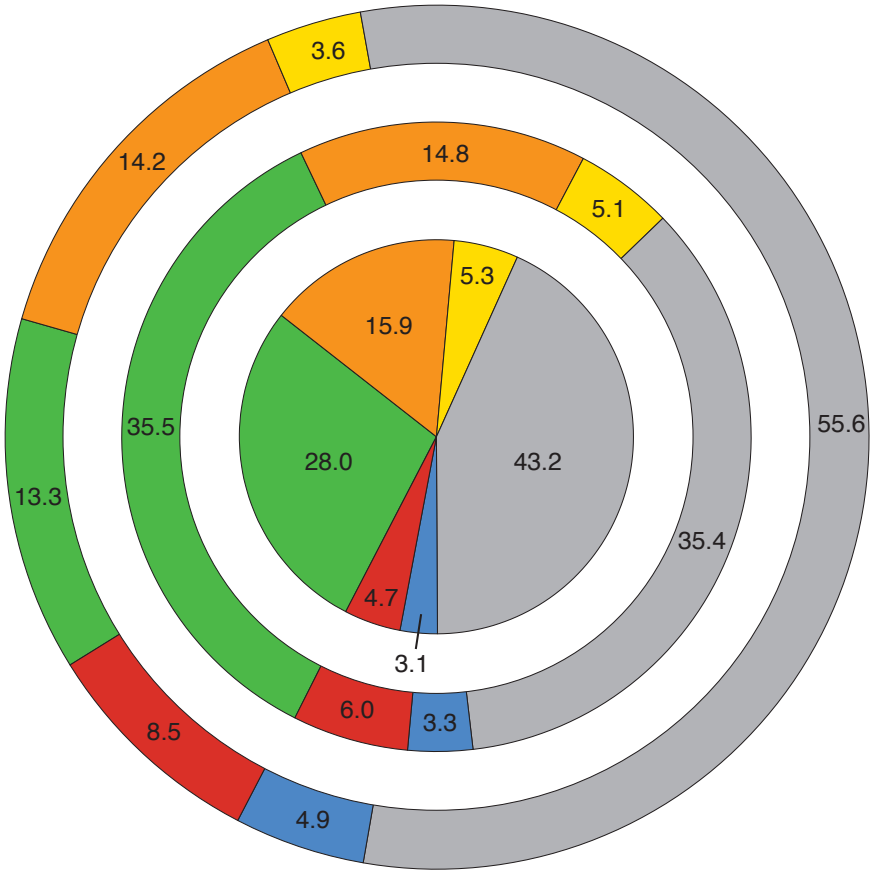
- Kelly Frazer

Questions?

Advertisements:

- Positions available (1 postdoc, 1 programmer, 1 PhD student)
  `http://www.fml.mpg.de/raetsch/offers`
- Slides with references available online
  `http://www.fml.mpg.de/raetsch/lectures`
- SVM Tutorial in PLoS Computational Biology
  `http://svmcompbio.tuebingen.mpg.de`
- Large Scale Learning Toolbox
  `http://shogun-toolbox.org`
- NIPS Workshop on Machine Learning in Computational Biology
  (invited speakers Thomas Lengauer, Steven Brenner and Aviv Regev)
  `http://www.mlcb.org`

Polymorphism Distribution



Sequence type	SNP no.	bp in PSPs
Coding	230,186	3,828,824
Intron	96,069	4,090,306
Untranslated RNA	32,771	1,029,948
Intergenic	229,335	16,046,419
Pseudogene	21,552	1,413,369
Transposon	38,657	2,462,213

Inner circle: genomic bases
Inner ring: SNP predictions
Outer ring: bases in PRPs

Coding regions are underrepresented for PRPs while transposons are overrepresented.

[Clark et al., 2007, Zeller et al., 2008]

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.

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

D.A. Hinds, L.L. Stuve, G.B. Nilsen, E. Halperin, E. Eskin, D.G. Ballinger, K.A. Frazer, and D.R. Cox. Whole-genome patterns of common DNA variation in three human populations. *Science*, 307(5712):1072–1079, 2005. URL `http://www.sciencemag.org/cgi/content/abstract/307/5712/1072`.

References II

D.A. Hinds, A.P. Kloek, M. Jen, X.Y. Chen, and K.A. Frazer. Common deletions and snps are in linkage disequilibrium in the human genome. *Nat. Genet.*, 38(1):82–85, 2006. ISSN 1061-4036. URL http://www.nature.com/ng/journal/v38/n1/abs/ng1695.html.

M. Koornneef, C. Alonso-Blanco, and D. Vreugdenhil. Naturally occurring genetic variation in Arabidopsis thaliana. *Annu. Rev. Plant Biol.*, 55(1): 141–172, 2004. URL http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.arplant.55.031903.141605.

M. Nordborg, T.T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, H. Zheng, E. Bakker, P. Calabrese, J. Gladstone, and R. Goyal et al. The pattern of polymorphism in Arabidopsis thaliana. *PLoS Biology*, 3(7): 1289–1299, 2005. URL http://dx.doi.org/10.1371/journal.pbio.0030196.

References III

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*, volume 19, pages 1161–1168, Cambridge, MA, 2007. MIT Press. URL `http://www.fml.tuebingen.mpg.de/raetsch/projects/HSMSVM`.

S. Sonnenburg, G. Rätsch, and K. Rieck. *Large Scale Kernel Machines*, chapter Large Scale Learning with String Kernels. MIT Press, 2007.

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.