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 adaptation?
- 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

Reference DNA sequence:

ACGTAAGTCGAAATGAAATGACCCCTTTTGAGAGCCCCGTT

Reference call:

ACGTAAGTCGAAATGACCC

SNP:

ACGTAAGTCGAGTGAAATGACCC

TGCATTTTCAGCTCACTTACTGGGAAAACCTTC

ACGTAAGTCGATTGAAATGACCC

Hybridization intensity:

A

C

G

T
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!
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
  - Repetition
- 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

Quality scores

Gunnar Rätsch (MPI, Tübingen)
Discovering Common Sequence Variation in A. thaliana
ISMB'08 Highlight
Labeled Data and SVM-Based Classification

Labeled training set

- 1,213 fragments of length $\approx 550\text{bp}$ 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

**reference**
- sequence
- intensity measurement
- labeled sequence set 2010

**accession 1**
- intensity measurement
- filter 1
- input generation
- SVM 1 model selection + training
- SVM 1 predictions
- transformation of outputs

**layer 1**
2-Layered Architecture for Inter-Strain Integration

Reference

Accession 1

Layer 1

Sequence

Intensity measurement

Labeled sequence set 2010

Filter 1

Input generation

SVM 1 model selection + training

SVM 1 predictions

Transformation of outputs
2-Layered Architecture for Inter-Strain Integration

Gunnar Rätsch (MPI, Tübingen)

Discovering Common Sequence Variation in *A. thaliana*

ISMB'08 Highlight
2-Layered Architecture for Inter-Strain Integration

Gunnar Rätsch (MPI, Tübingen)
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

⇒ Polymorphic Region Prediction (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

⇒ Polymorphic Region Prediction (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

⇒ Polymorphic Region Prediction (PRP)
Modeling polymorphic regions

Log max intensity

bp

Exon
Intron
Exon

D
A
B
C

Col-0
Bor-4

(1)
(2)
(3)
(4)
(5)

Transition Labels

Not polymorphic

PR Predictions

MBML2 SNP

Known polymorphisms

Insertion

Deletion

Gunnar Rätsch (MPI, Tübingen) Discovering Common Sequence Variation in A. thaliana ISMB’08 Highlight
Modeling polymorphic regions

![Graphical representation of genomic data showing polymorphic regions. The graph includes a line plot with two lines, one for Col-0 and one for Bor-4, across different genomic positions. The positions are labeled (1) to (5) with corresponding transition labels and predictions.]

Transitions:
- (1): Not polymorphic
- (2): MBML2 SNP
- (3): Known polymorphisms
- (4): Insertion
- (5): Deletion
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

\[
F_\Theta : X \times Y \rightarrow \mathbb{R} \quad \text{(path scoring)}
\]

to be used as:

\[
f(x) = \arg \max_{\pi \in Y} F(x, \pi) \quad \text{(Viterbi decoding)}
\]

[Altun et al., 2003]
Large Margin Separation

\[ F(x^i, \pi^i) - F(x^i, \bar{\pi}) \gg 0 \quad \forall \bar{\pi} \neq \pi^i \quad \forall i \]

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

Solve Linear Optimization Problem

\[
\begin{align*}
\min_{\theta, \xi \geq 0} & \quad \frac{1}{n} \sum_{i=1}^{n} \xi^{(i)} + C \Omega(\theta) \\
\text{s.t.} & \quad F_\theta(x^{(i)}, \pi^{(i)}) - F_\theta(x^{(i)}, \bar{\pi}) \geq 1 - \xi^{(i)} \\
& \quad \forall \bar{\pi} \neq \pi^{(i)} \quad \forall i = 1, \ldots, n,
\end{align*}
\]

- \( \xi \): slack variables
- \( \Omega \): linear regularization term
Large Margin Separation

\[ F(x^i, \pi^i) - F(x^i, \bar{\pi}) \gg 0 \quad \forall \bar{\pi} \neq \pi^i \quad \forall i \]

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

Solve Linear Optimization Problem

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

s.t. \[ F_\theta(x^{(i)}, \pi^{(i)}) - F_\theta(x^{(i)}, \bar{\pi}) \geq 1 - \xi^{(i)} \]
\[ \forall \bar{\pi} \neq \pi^{(i)} \quad \forall i = 1, \ldots, n, \]

- \( \xi \): slack variables
- \( \Omega \): 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):

<table>
<thead>
<tr>
<th></th>
<th>SNP calling (MB+ML)</th>
<th>Region predictor</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPs</td>
<td>∼32%</td>
<td>∼65%</td>
</tr>
<tr>
<td>Deletions (per base)</td>
<td>∼53%</td>
<td></td>
</tr>
<tr>
<td>Insertions</td>
<td>∼39%</td>
<td></td>
</tr>
</tbody>
</table>

[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 >8,000 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

Major-effect changes by gene category (%)

<table>
<thead>
<tr>
<th>Gene family</th>
<th>(n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cytoplasmic ribosomal</td>
<td>230</td>
</tr>
<tr>
<td>MYB transcription factor</td>
<td>131</td>
</tr>
<tr>
<td>bHLH transcription factor</td>
<td>158</td>
</tr>
<tr>
<td>Glycosyltransferase related</td>
<td>277</td>
</tr>
<tr>
<td>Organic solute cotransporter</td>
<td>357</td>
</tr>
<tr>
<td>Acyl lipid metabolism</td>
<td>603</td>
</tr>
<tr>
<td>EF-hand containing</td>
<td>187</td>
</tr>
<tr>
<td>Glycoside hydrolase</td>
<td>303</td>
</tr>
<tr>
<td>Cytochrome P450</td>
<td>215</td>
</tr>
<tr>
<td>RING</td>
<td>463</td>
</tr>
<tr>
<td>Receptor-like kinase</td>
<td>613</td>
</tr>
<tr>
<td>F-box</td>
<td>660</td>
</tr>
<tr>
<td>NB-LRR</td>
<td>125</td>
</tr>
</tbody>
</table>

- **Large-effect SNP**
- **Coding region in PRs**

Gunnar Rätsch (MPI, Tübingen)  Discovering Common Sequence Variation in *A. thaliana*  ISMB’08 Highlight
Conclusions

- Created inventory for polymorphisms in *Arabidopsis thaliana*
- Highly polymorphic: $\approx 0.5\%$ in SNPs, $\approx 27\%$ 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, ...
Conclusions

- Created inventory for polymorphisms in *Arabidopsis thaliana*
  - Highly polymorphic: \( \approx 0.5\% \) in SNPs, \( \approx 27\% \) 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, ...
Conclusions

- Created inventory for polymorphisms in *Arabidopsis thaliana*
  - Highly polymorphic: \( \approx 0.5\% \) in SNPs, \( \approx 27\% \) 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, ...
Conclusions

- Created inventory for polymorphisms in *Arabidopsis thaliana*
  - Highly polymorphic: $\approx 0.5\%$ in SNPs, $\approx 27\%$ 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 Developmental Biology
- Richard Clark  
- Stephan Ossowski  
- Korbinian Schneeberger  
- Norman Warthmann  
- Detlef Weigel

### MPI for Biological Cybernetics
- Bernhard Schölkopf  
- Gabriele Schweikert

### The Salk Institute, CA
- Paul Shinn  
- Joseph Ecker

### ZBIT, University of Tübingen
- Daniel Huson

### USC, CA
- Magnus Nordberg

### 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

Coding regions are underrepresented for PRPs while transposons are overrepresented.

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


