Revealing Sequence Variation Patterns in Rice with Machine Learning Methods

Regina Bohnert

Friedrich Miescher Laboratory of the Max Planck Society
Tübingen, Germany

July 18, 2008

Motivation

What distinguishes the sequences of subpopulations with different traits?

- Identify sequence variations within one species
- Basis for further evolutionary and functional studies

Genome-wide identification of sequence polymorphisms

- High-density oligonucleotide microarrays for high-throughput resequencing
- Array-based resequencing applied for
  - Human (Hinds et al., *Science*, 2005)
  - *Arabidopsis thaliana* (Clark et al., *Science*, 2007)
  - Mouse (Frazer et al., *Nature*, 2007)
  - Oryza sativa (rice)

Motivation

What distinguishes the sequences of subpopulations with different traits?

- Identify sequence variations within one species
- Basis for further evolutionary and functional studies

Genome-wide identification of sequence polymorphisms

- High-density oligonucleotide microarrays for high-throughput resequencing
- Array-based resequencing applied for
  - Human (Hinds et al., *Science*, 2005)
  - *Arabidopsis thaliana* (Clark et al., *Science*, 2007)
  - Mouse (Frazer et al., *Nature*, 2007)
  - *Oryza sativa* (rice)

Motivation

What distinguishes the sequences of subpopulations with different traits?

- Identify sequence variations within one species
- Basis for further evolutionary and functional studies

Genome-wide identification of sequence polymorphisms

- High-density oligonucleotide microarrays for high-throughput resequencing
- Array-based resequencing applied for
  - Human (Hinds et al., *Science*, 2005)
  - *Arabidopsis thaliana* (Clark et al., *Science*, 2007)
  - Mouse (Frazer et al., *Nature*, 2007)
  - *Oryza sativa* (rice)

Motivation

What distinguishes the sequences of subpopulations with different traits?

- Identify sequence variations within one species
- Basis for further evolutionary and functional studies

Genome-wide identification of sequence polymorphisms

- High-density oligonucleotide microarrays for high-throughput resequencing
- Array-based resequencing applied for
  - Human (Hinds et al., *Science*, 2005)
  - *Arabidopsis thaliana* (Clark et al., *Science*, 2007)
  - Mouse (Frazer et al., *Nature*, 2007)
  - *Oryza sativa* (rice)

Oryza sativa

Prominent model organism

- Most important food source
- Representative of grass family
- Closely related to other cereals
- 372 Mb genome on 12 chr.

Challenges relative to *A. thaliana*

- Different experimental design
- Data not as clean
- No gold standard set of labelled sequences



from Koehler's Medicinal-Plants, 1887

Oryza sativa

Prominent model organism

- Most important food source
- Representative of grass family
- Closely related to other cereals
- 372 Mb genome on 12 chr.

Challenges relative to *A. thaliana*

- Different experimental design
- Data not as clean
- No gold standard set of labelled sequences

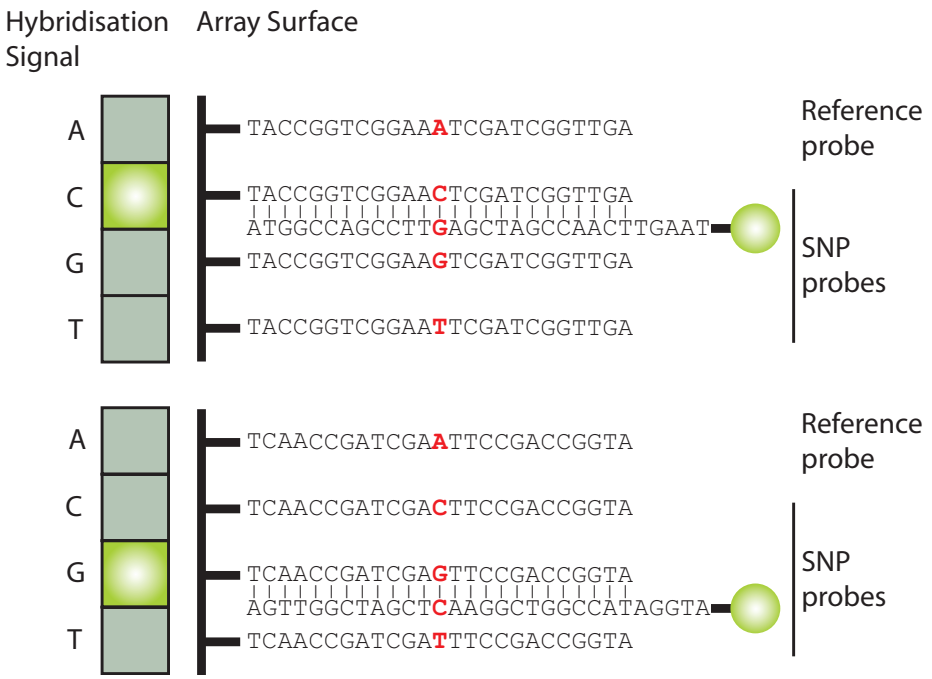




from Koehler's Medicinal-Plants, 1887

The Resequencing Data – Tiling Arrays

Friedrich Miescher Laboratory
of the Max Planck Society



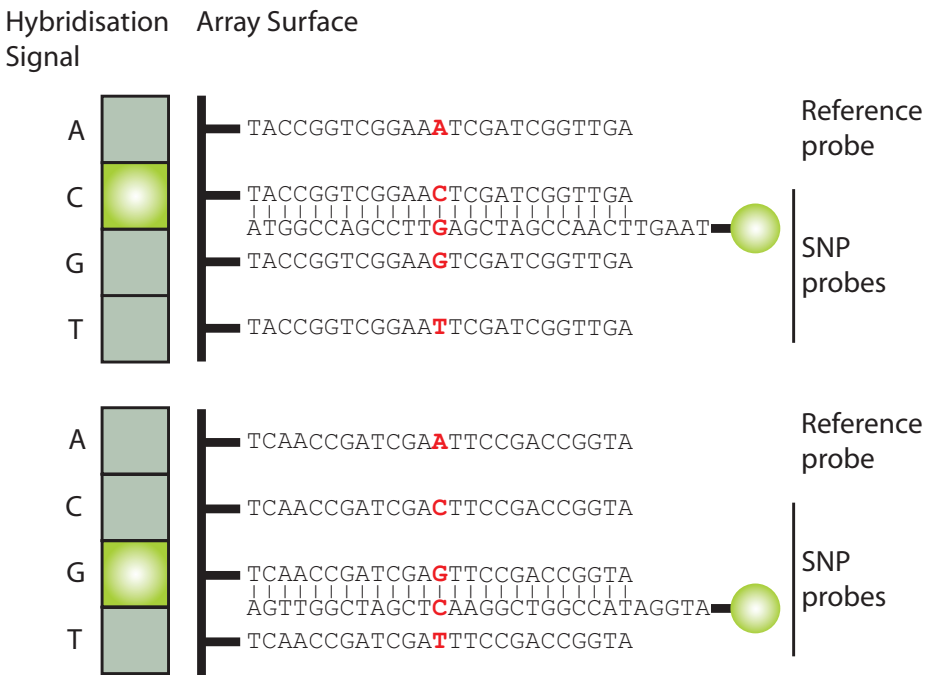
- Oligonucleotides on glass surface
- ssDNA labelled with fluorescence as target DNA

Rice resequencing arrays

- Tiling strategy with 1 bp resolution
- Each base queried with a forward and reverse quartet
- 800 million oligos on 246 arrays for each of the 20 cultivars
  - $\sim$32 % of the genome represented
- Target DNA amplified by long-range PCR

The Resequencing Data – Tiling Arrays

Friedrich Miescher Laboratory
of the Max Planck Society



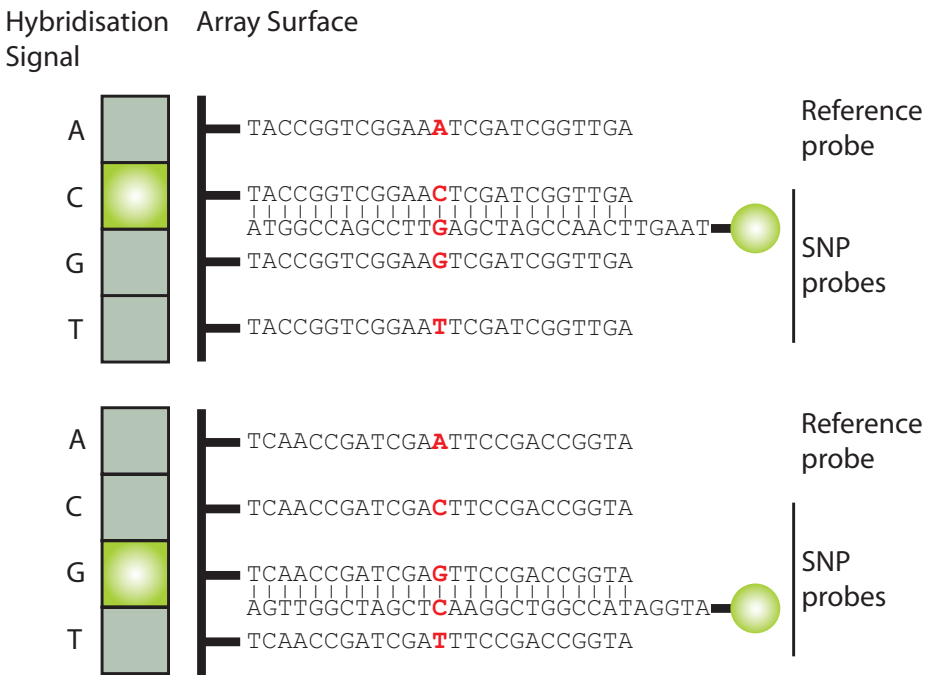
- Oligonucleotides on glass surface
- ssDNA labelled with fluorescence as target DNA

Rice resequencing arrays

- Tiling strategy with 1 bp resolution
- Each base queried with a forward and reverse quartet
- 800 million oligos on 246 arrays for each of the 20 cultivars
  - $\sim$ 32 % of the genome represented
- Target DNA amplified by long-range PCR

The Resequencing Data – Tiling Arrays

Friedrich Miescher Laboratory
of the Max Planck Society



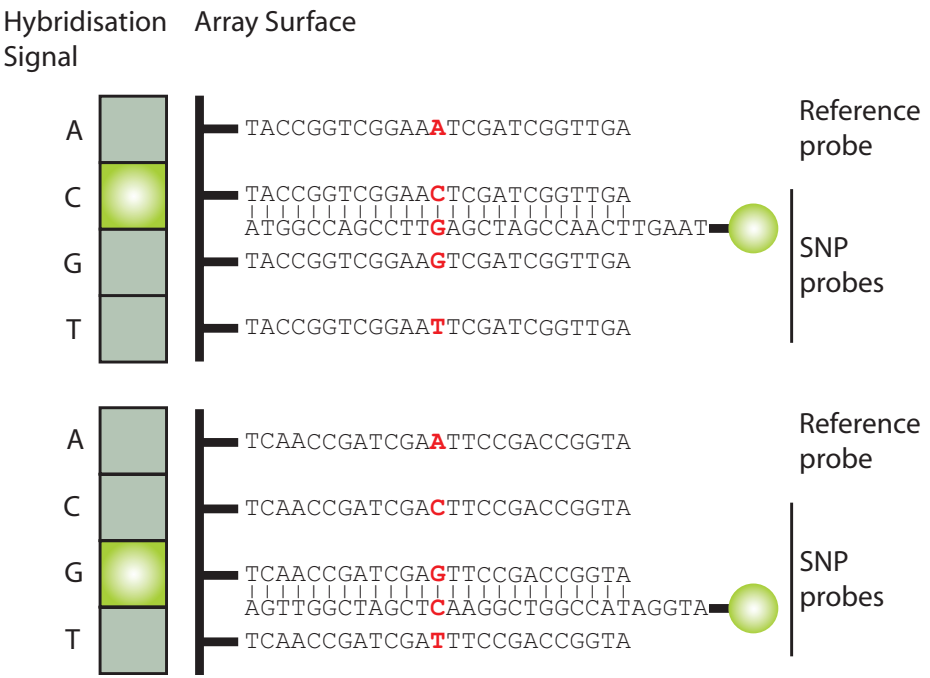
- Oligonucleotides on glass surface
- ssDNA labelled with fluorescence as target DNA

Rice resequencing arrays

- Tiling strategy with 1 bp resolution
- Each base queried with a forward and reverse quartet
- 800 million oligos on 246 arrays for each of the 20 cultivars
  - $\sim$ 32 % of the genome represented
- Target DNA amplified by long-range PCR

The Resequencing Data – Tiling Arrays

Friedrich Miescher Laboratory
of the Max Planck Society

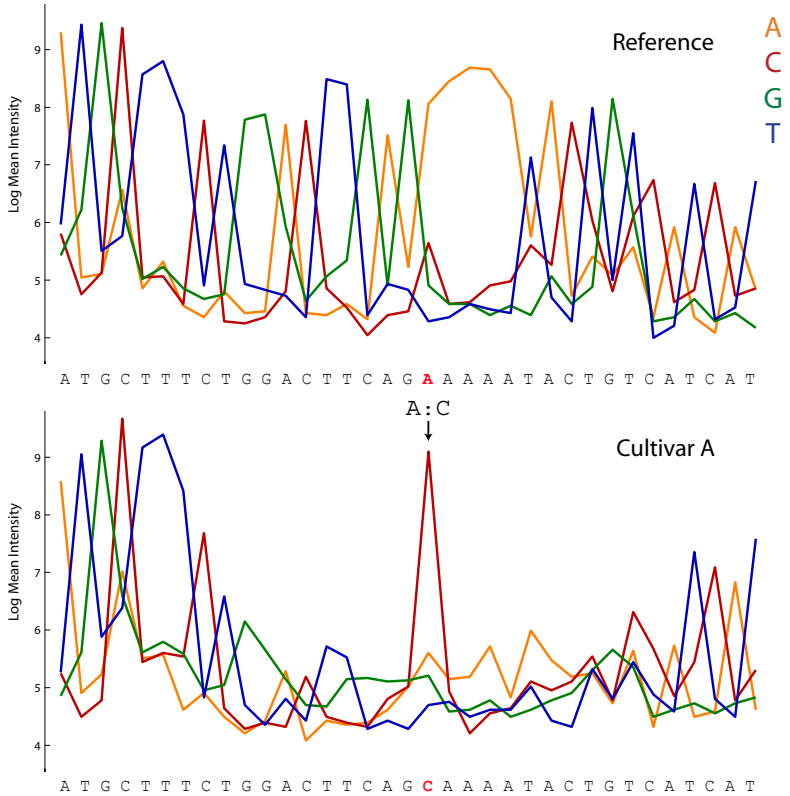


- Oligonucleotides on glass surface
- ssDNA labelled with fluorescence as target DNA

Rice resequencing arrays

- Tiling strategy with 1 bp resolution
- Each base queried with a forward and reverse quartet
- 800 million oligos on 246 arrays for each of the 20 cultivars
  - $\sim$ 32 % of the genome represented
- Target DNA amplified by long-range PCR

Polymorphism Detection



Data analysis challenge

- Hybridisation signal dependent on sequence properties of oligomer, repeats, amplicon, cultivar
- Measurement noise

⇒ Machine learning based SNP calling

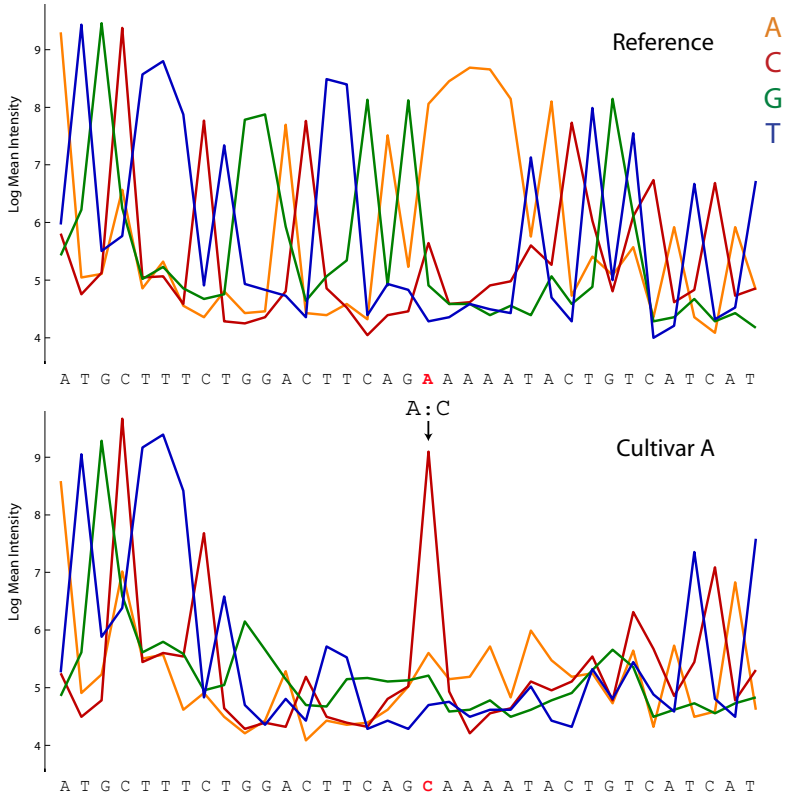
Problematic cases

- Highly polymorphic regions
- Deletions and Insertions

⇒ Margin-based prediction of polymorphic regions

Polymorphism Detection

Friedrich Miescher Laboratory
of the Max Planck Society



Data analysis challenge

- Hybridisation signal dependent on sequence properties of oligomer, repeats, amplicon, cultivar
- Measurement noise

⇒ Machine learning based SNP calling

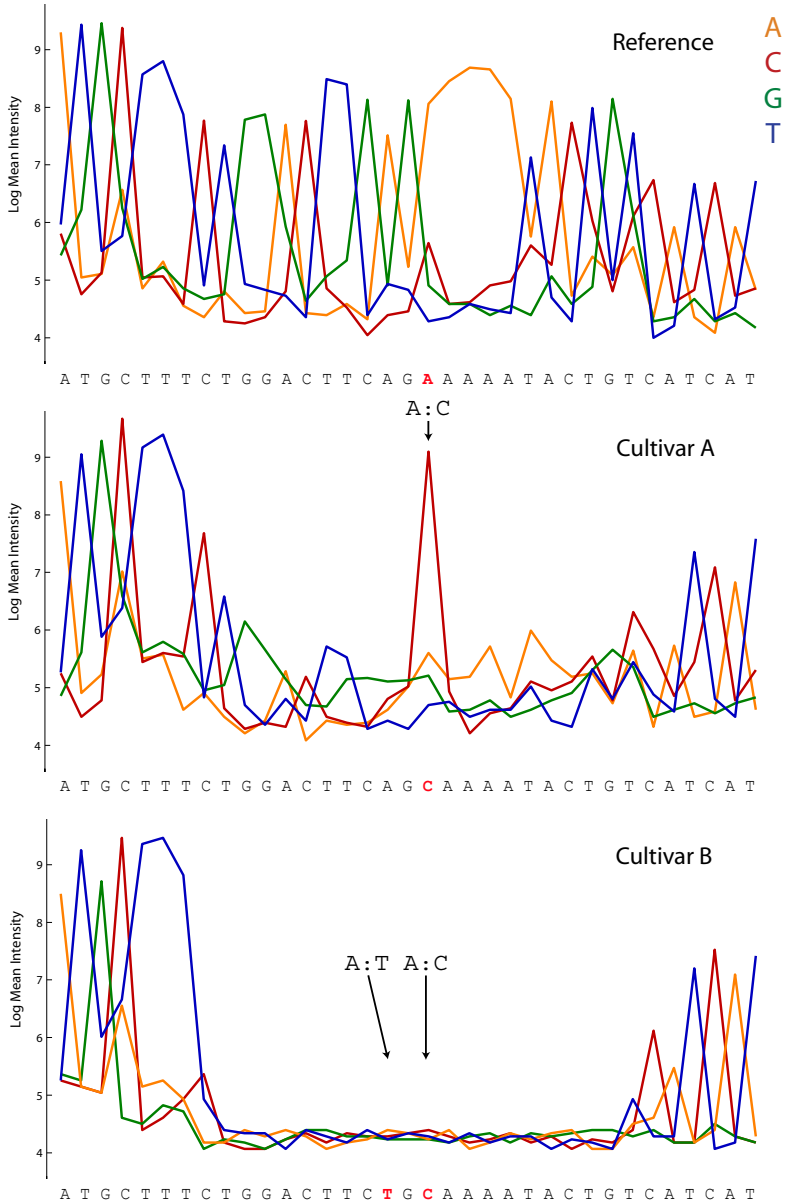
Problematic cases

- Highly polymorphic regions
- Deletions and Insertions

⇒ Margin-based prediction of polymorphic regions

Polymorphism Detection

Friedrich Miescher Laboratory
of the Max Planck Society



Data analysis challenge

- Hybridisation signal dependent on sequence properties of oligomer, repeats, amplicon, cultivar
- Measurement noise

⇒ Machine learning based SNP calling

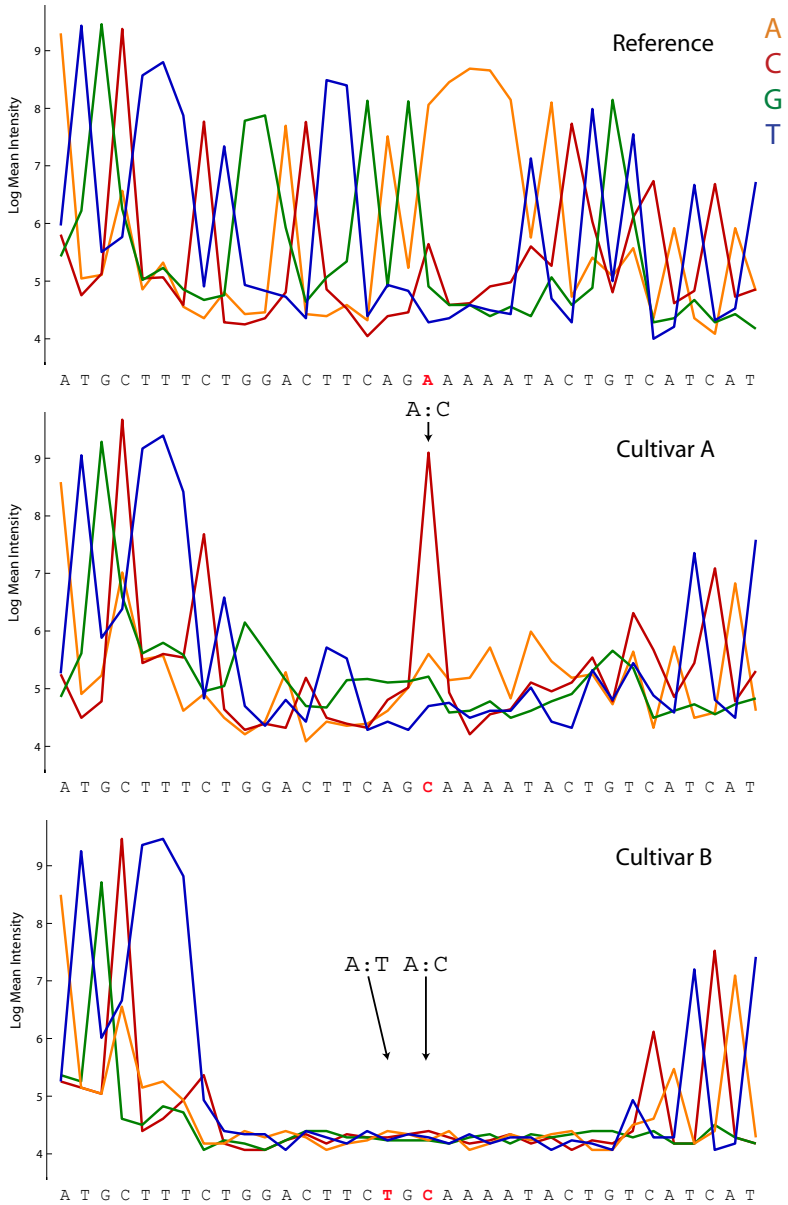
Problematic cases

- Highly polymorphic regions
- Deletions and insertions

⇒ Margin-based prediction of polymorphic regions

Polymorphism Detection

Friedrich Miescher Laboratory
of the Max Planck Society



Data analysis challenge

- Hybridisation signal dependent on sequence properties of oligomer, repeats, amplicon, cultivar
- Measurement noise

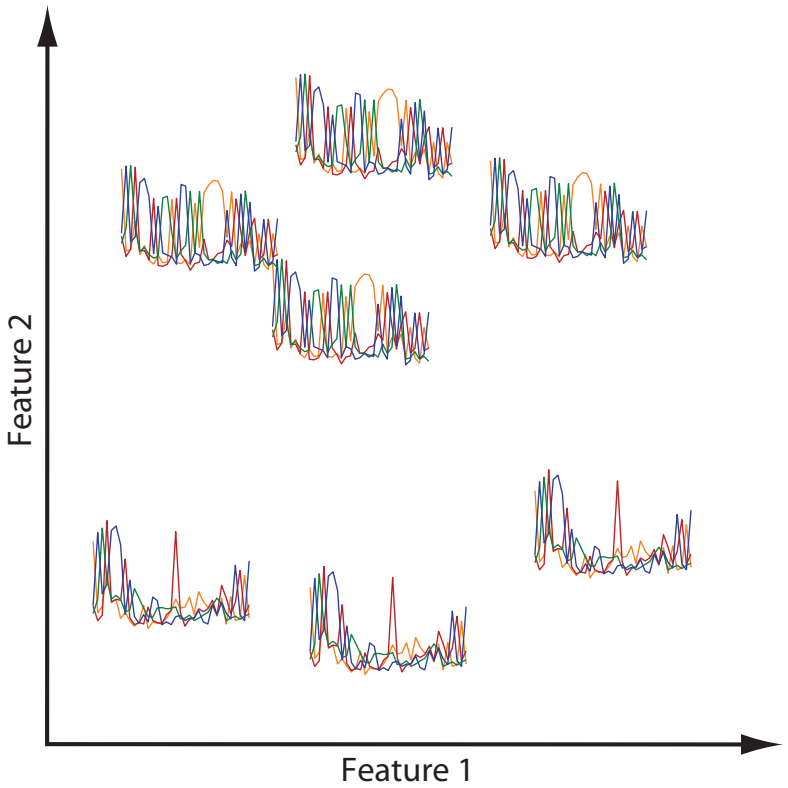
⇒ Machine learning based SNP calling

Problematic cases

- Highly polymorphic regions
- Deletions and insertions

⇒ Margin-based prediction of polymorphic regions

SNP Calling – ML Approach

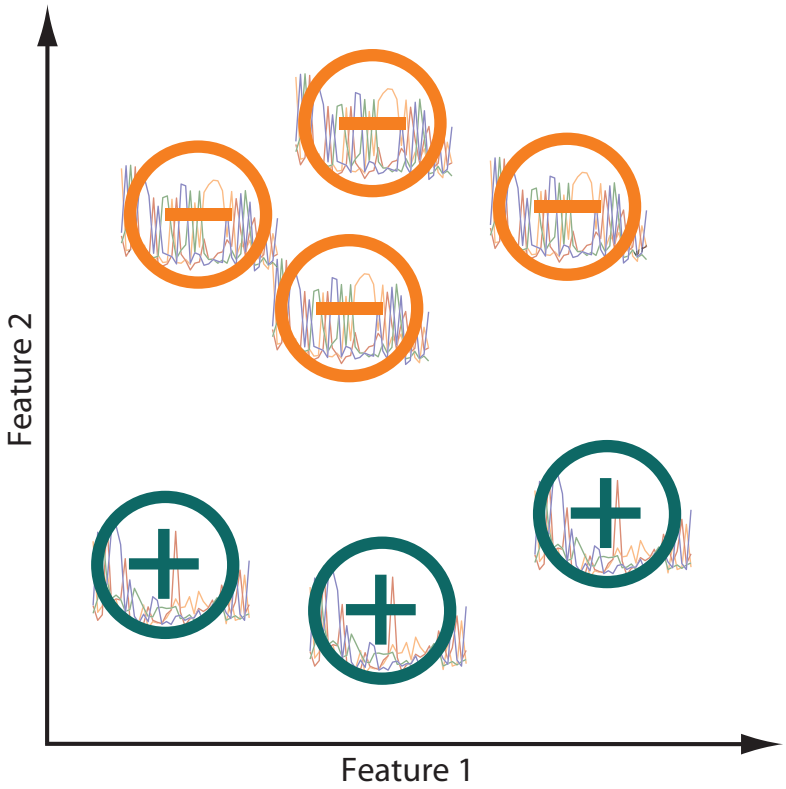


Support Vector Machines (SVM)

- Extract features
  - Array data
  - Sequence
  - Repetitiveness
- Labelled data generated by sequencing of randomly selected fragments
- Apply SVMs using RBF kernel in a two-layered approach
  - 2nd layer exploits information across cultivars

cf. Clark et al., *Science*, 2007

SNP Calling – ML Approach

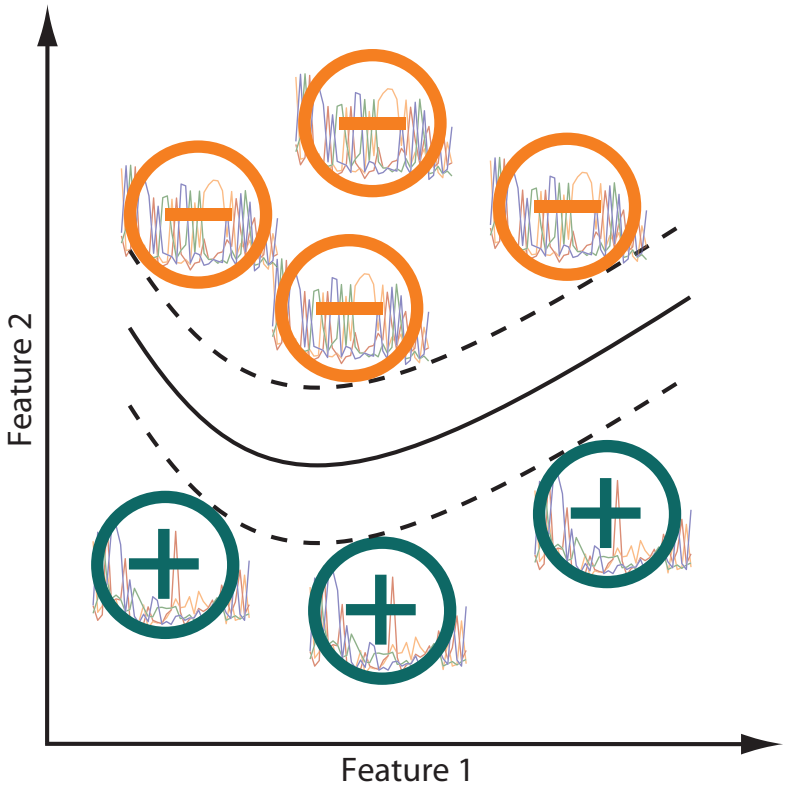


Support Vector Machines (SVM)

- Extract features
  - Array data
  - Sequence
  - Repetitiveness
- Labelled data generated by sequencing of randomly selected fragments
- Apply SVMs using RBF kernel in a two-layered approach
  - 2nd layer exploits information across cultivars

cf. Clark et al., *Science*, 2007

SNP Calling – ML Approach



Support Vector Machines (SVM)

- Extract features
  - Array data
  - Sequence
  - Repetitiveness
- Labelled data generated by sequencing of randomly selected fragments
- Apply SVMs using RBF kernel in a two-layered approach
  - 2nd layer exploits information across cultivars

cf. Clark et al., *Science*, 2007

SNP Calling – Results

Predicted SNPs

- $\sim$ 1.2 M MB SNP calls
- $\sim$ 1.3 M ML SNP calls
- $\sim 760,000$ SNPs in MB$\cap$ML at $\sim 160,000$ positions

MB: Model based approach by Perlegen Sciences

ML: Proposed machine learning approach

SNP Calling – Results

Predicted SNPs

- $\sim$ 1.2 M MB SNP calls
- $\sim$ 1.3 M ML SNP calls
- $\sim 760,000$ SNPs in MB$\cap$ML at $\sim 160,000$ positions

	Recall	Precision
MB	14 %	91 %
ML	21 %	92 %
MB$\cap$ML	11 %	97 %

MB: Model based approach by Perlegen Sciences

ML: Proposed machine learning approach

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{P}$$

Visit Poster M09 and P28

SNP Calling – Results

Predicted SNPs

- $\sim$ 1.2 M MB SNP calls
- $\sim$ 1.3 M ML SNP calls
- $\sim 760,000$ SNPs in MB$\cap$ML at $\sim 160,000$ positions

	Recall	Precision
MB	14 %	91 %
ML	21 %	92 %
MB$\cap$ML	11 %	97 %

MB: Model based approach by Perlegen Sciences

ML: Proposed machine learning approach

$$\text{Precision} = \frac{TP}{TP+FP}, \text{Recall} = \frac{TP}{P}$$

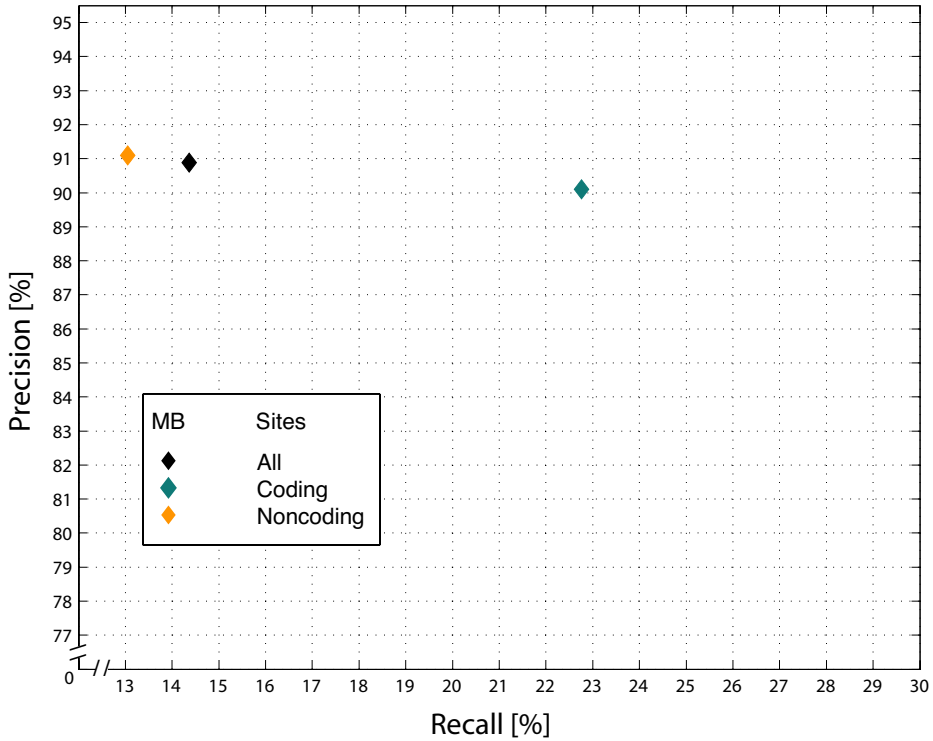
Visit Poster M09 and P28

SNP Calling – Results

Predicted SNPs

- $\sim$ 1.2 M MB SNP calls
- $\sim$ 1.3 M ML SNP calls
- $\sim 760,000$ SNPs in MB$\cap$ML at $\sim 160,000$ positions

	Recall	Precision
MB	14 %	91 %
ML	21 %	92 %
MB$\cap$ML	11 %	97 %



MB: Model based approach by Perlegen Sciences

ML: Proposed machine learning approach

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{P}}$$

Visit Poster M09 and P28

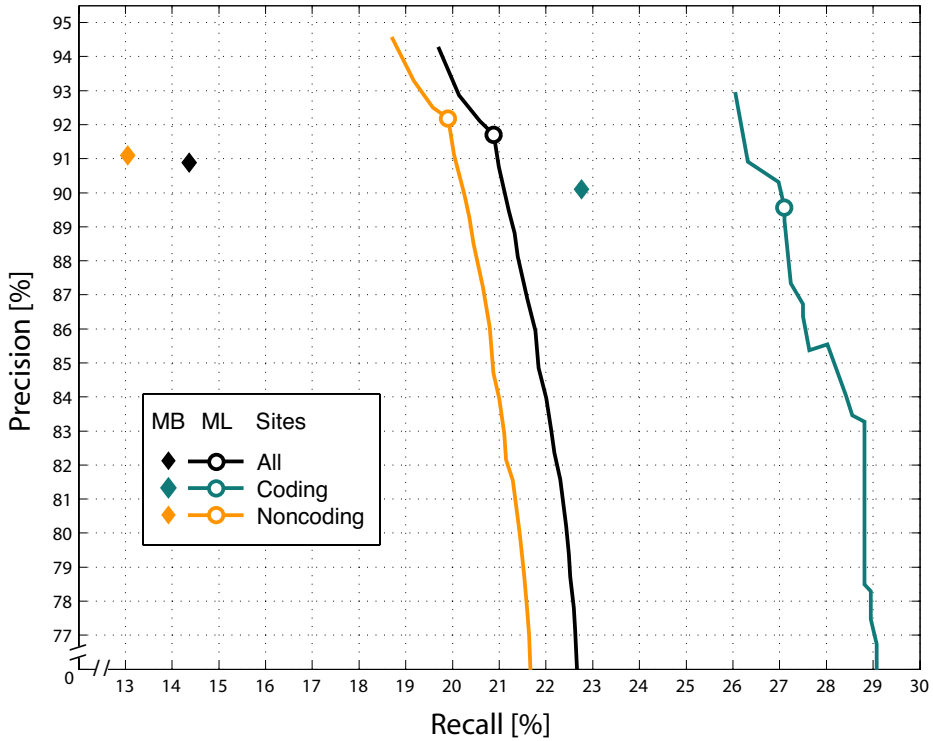
SNP Calling – Results

Friedrich Miescher Laboratory
of the Max Planck Society

Predicted SNPs

- $\sim$ 1.2 M MB SNP calls
- $\sim$ 1.3 M ML SNP calls
- $\sim 760,000$ SNPs in MB$\cap$ML at $\sim 160,000$ positions

	Recall	Precision
MB	14 %	91 %
ML	21 %	92 %
MB$\cap$ML	11 %	97 %



MB: Model based approach by Perlegen Sciences

ML: Proposed machine learning approach

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{P}}$$

Visit Poster M09 and P28

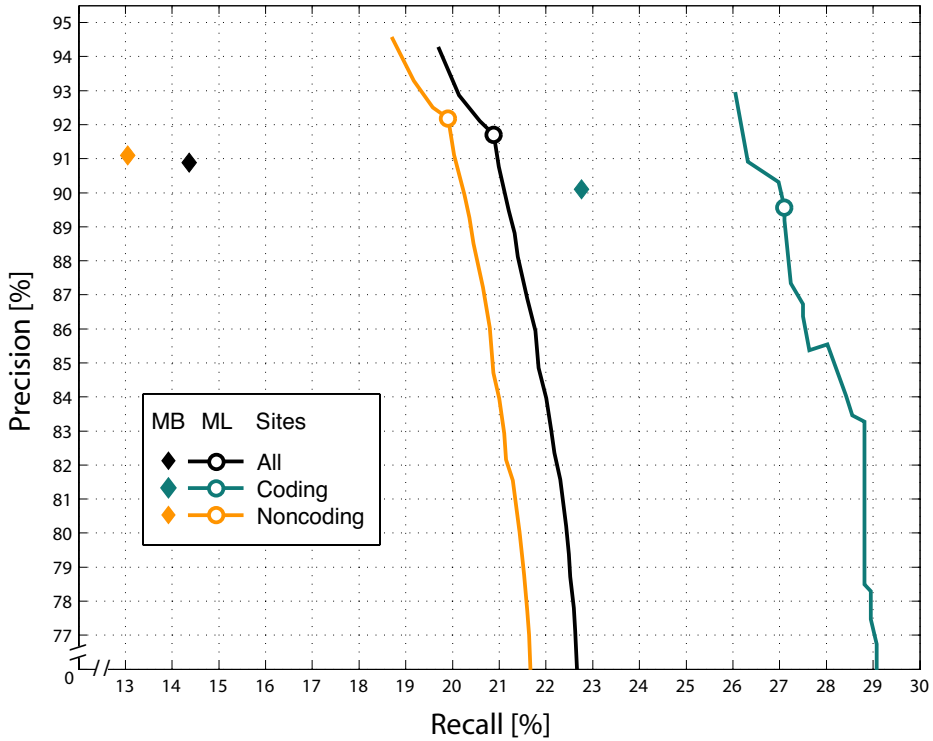
SNP Calling – Results

Friedrich Miescher Laboratory
of the Max Planck Society

Predicted SNPs

- $\sim$ 1.2 M MB SNP calls
- $\sim$ 1.3 M ML SNP calls
- $\sim 760,000$ SNPs in MB$\cap$ML at $\sim 160,000$ positions

	Recall	Precision
MB	14 %	91 %
ML	21 %	92 %
MB$\cap$ML	11 %	97 %



MB: Model based approach by Perlegen Sciences

ML: Proposed machine learning approach

$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{Recall} = \frac{TP}{P}$$

Visit Poster M09 and P28

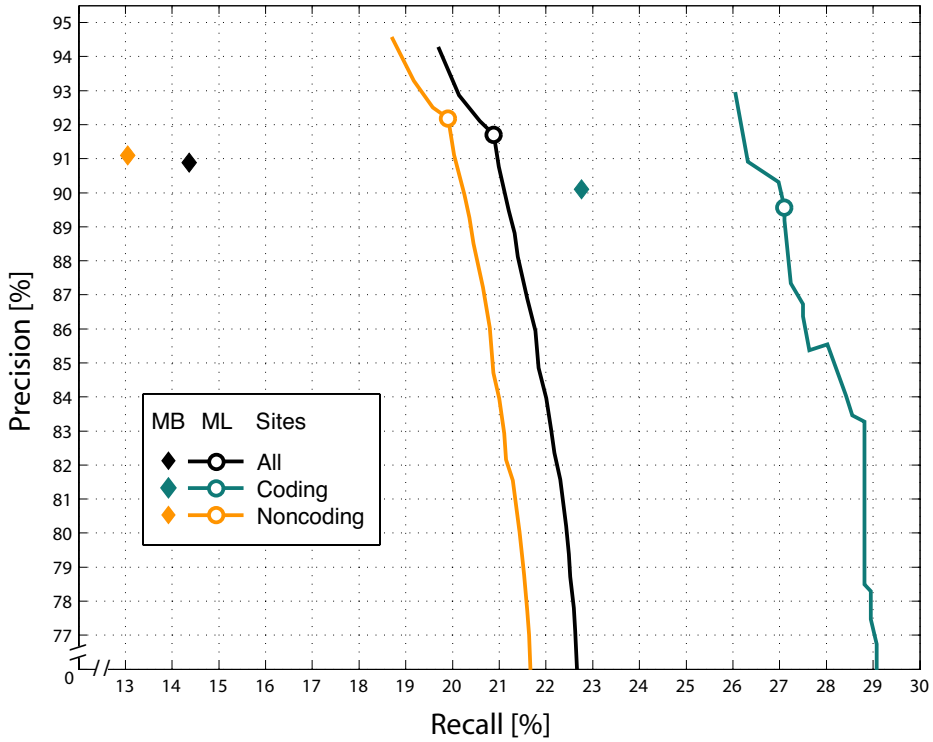
SNP Calling – Results

Friedrich Miescher Laboratory
of the Max Planck Society

Predicted SNPs

- $\sim$ 1.2 M MB SNP calls
- $\sim$ 1.3 M ML SNP calls
- $\sim 760,000$ SNPs in MB$\cap$ML at $\sim 160,000$ positions

	Recall	Precision
MB	14 %	91 %
ML	21 %	92 %
MB$\cap$ML	11 %	97 %



MB: Model based approach by Perlegen Sciences

ML: Proposed machine learning approach

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{P}}$$

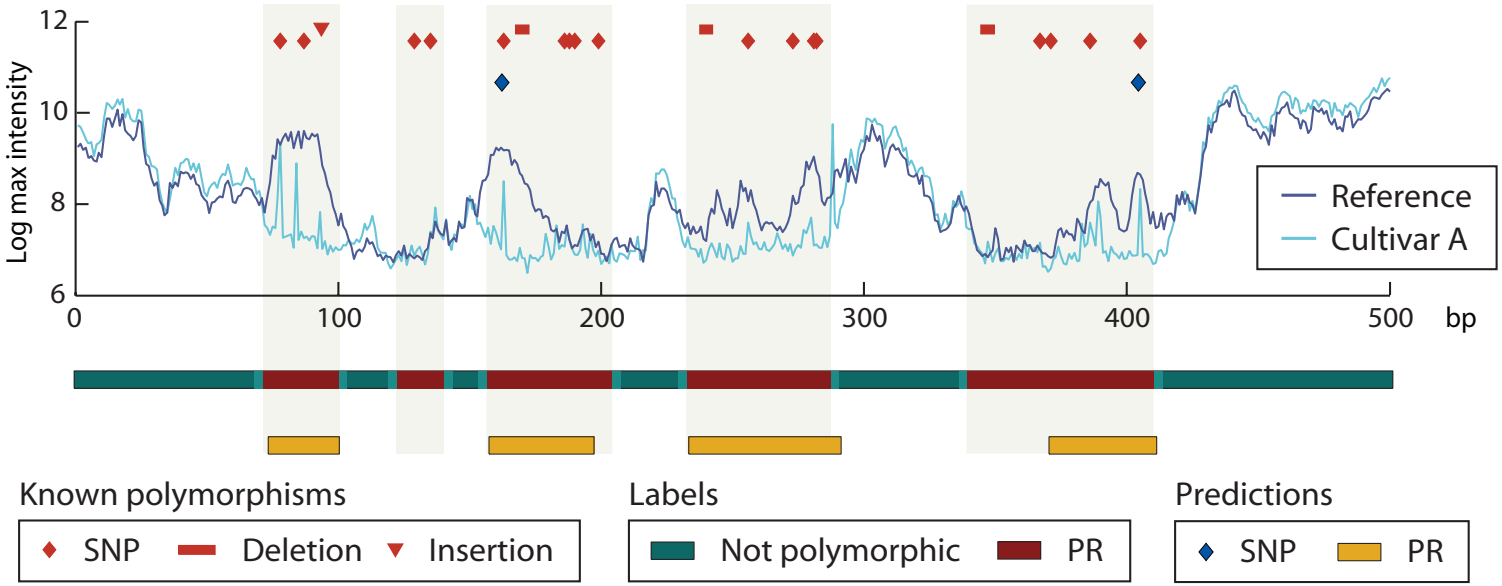
Visit Poster M09 and P28

SNP Calling – Database

http://irfgc.irri.org/cgi-bin/gbrowse/oryzasnp10/



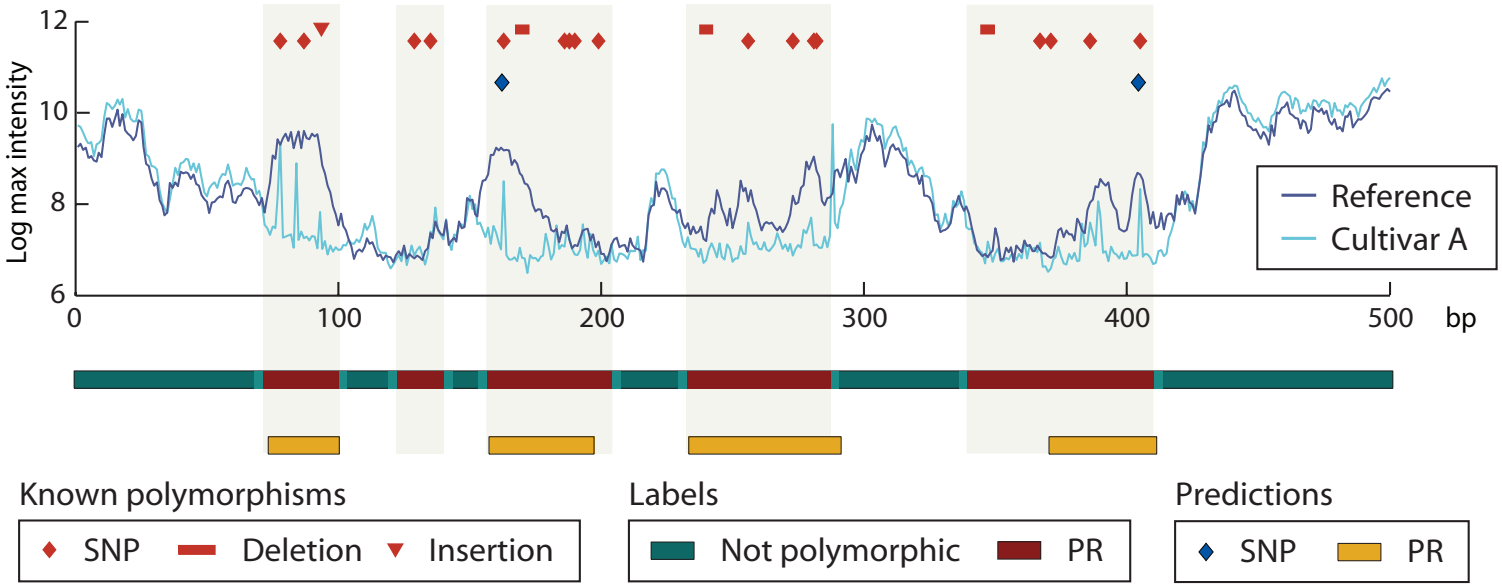
Predicting Polymorphic Regions



- Difficulties when SNPs occur in vicinity
- Approach: Predict genomic segments with label sequence learning algorithm
  - Conserved and
  - **Polymorphic regions (PRs)**

cf. Zeller et al., *Genome Research*, 2008

Predicting Polymorphic Regions

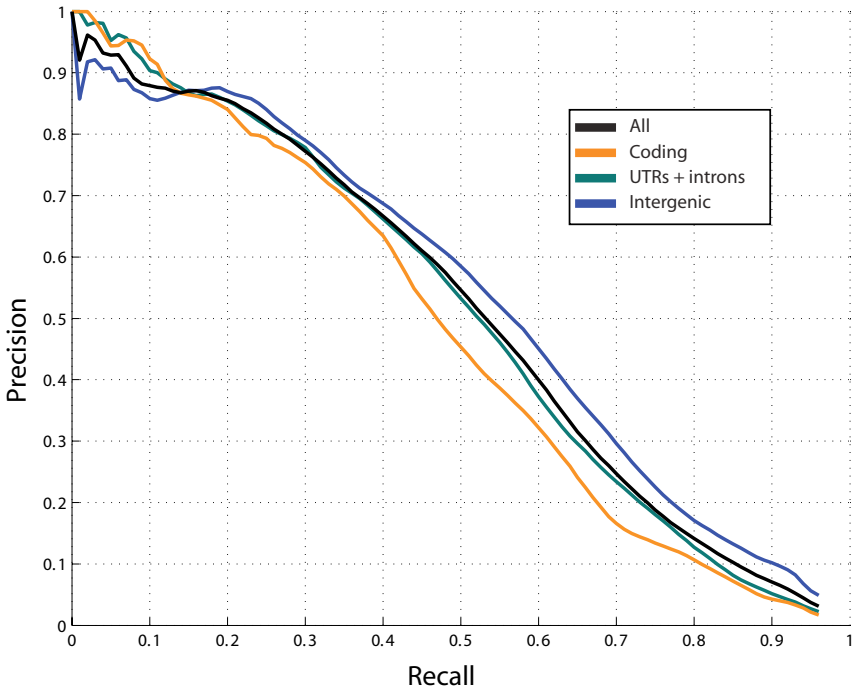


- Difficulties when SNPs occur in vicinity
- Approach: Predict genomic segments with label sequence learning algorithm
  - Conserved and
  - **Polymorphic regions (PRs)**

cf. Zeller et al., *Genome Research*, 2008

Predicting PRs – Results

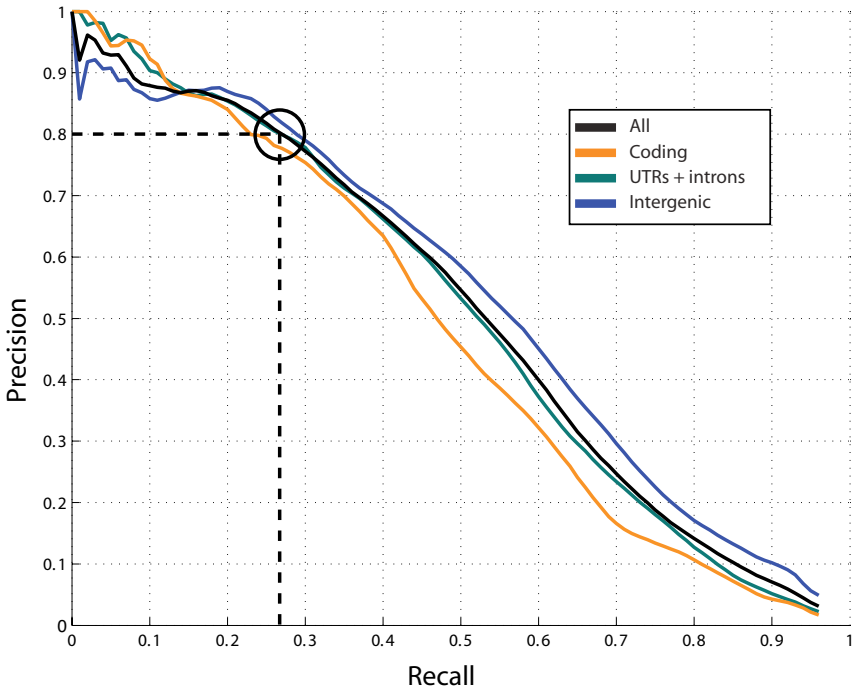
- Between $\sim 65,000$ and $\sim 203,000$ PRs predicted per cultivar
- ~ 27 % recall at a precision of 80 %
- Between 1.7 % and 5.1 % of the genome covered



$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \ \text{Recall} = \frac{\text{TP}}{\text{P}}$$

Predicting PRs – Results

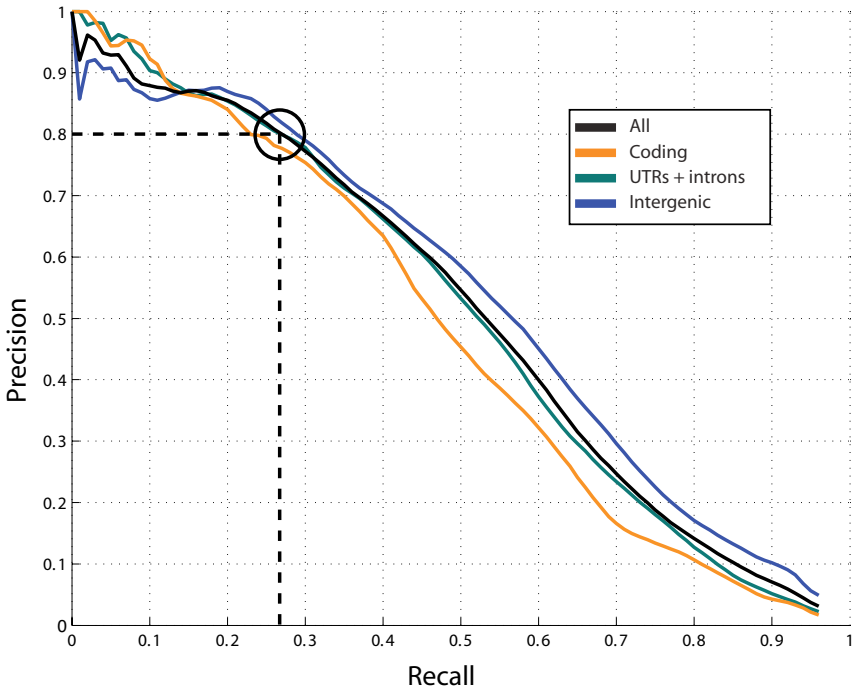
- Between $\sim 65,000$ and $\sim 203,000$ PRs predicted per cultivar
- ~ 27 % recall at a precision of 80 %
- Between 1.7 % and 5.1 % of the genome covered



$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{P}}$$

Predicting PRs – Results

- Between $\sim 65,000$ and $\sim 203,000$ PRs predicted per cultivar
- ~ 27 % recall at a precision of 80 %
- Between 1.7 % and 5.1 % of the genome covered

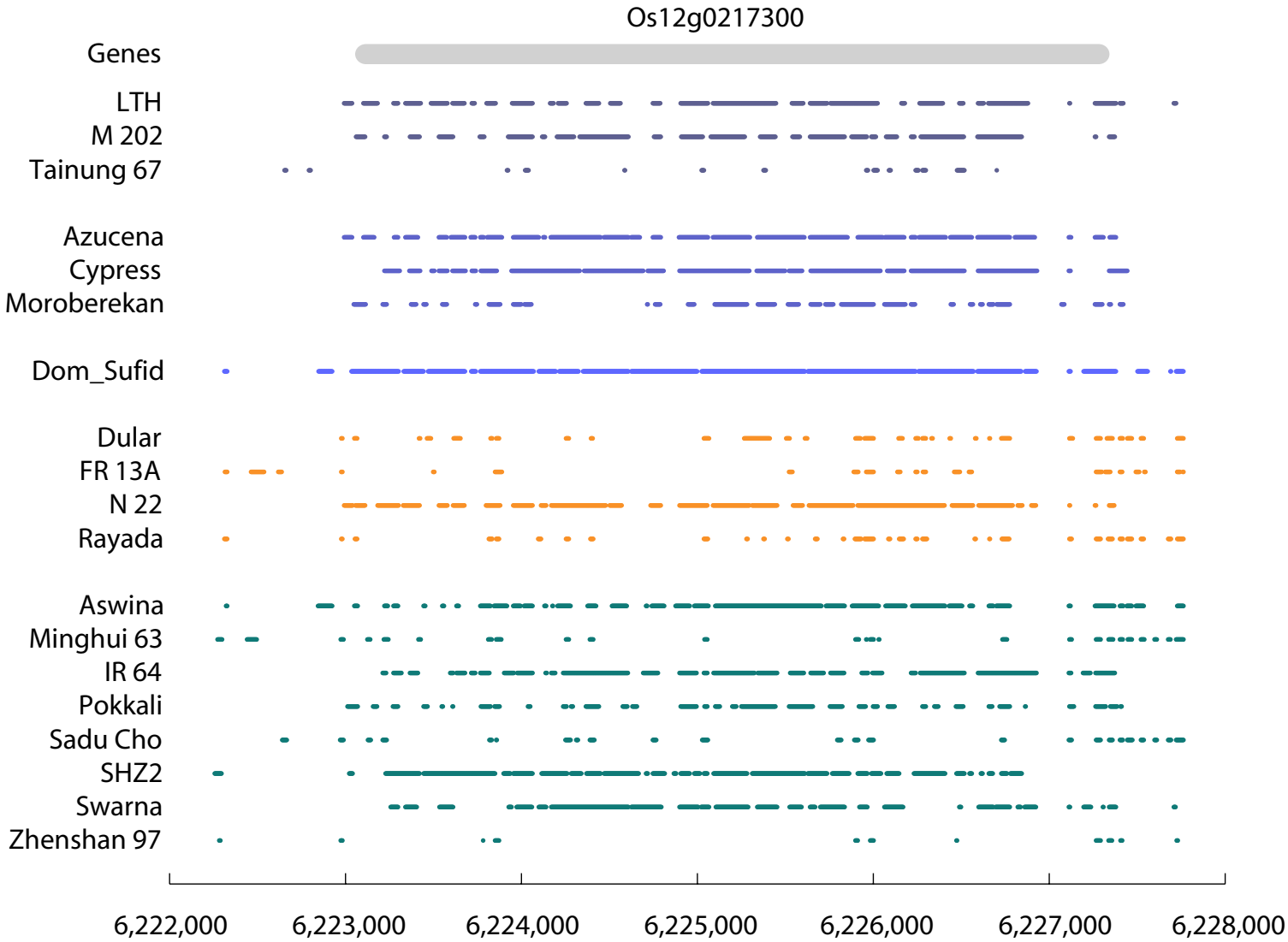


$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \ \text{Recall} = \frac{\text{TP}}{\text{P}}$$

Long Deletion Example

Friedrich Miescher Laboratory
of the Max Planck Society

Disease resistance protein SlVe1 precursor on chromosome 12



Conclusions

- Created the first whole-genome inventory of polymorphisms for rice
  - Highly polymorphic $\sim$ 0.2 % in SNPs, $\sim$ 2.4 % in PRPs
- Intersection of MB and ML calls provides highly reliable SNP predictions
  - Used to genotype many more rice cultivars
- Polymorphic region predictions
  - Important for more detailed analyses (e.g. dideoxy sequencing)
  - Useful for primer design to increase PCR success rates

Conclusions

- Created the first whole-genome inventory of polymorphisms for rice
  - Highly polymorphic $\sim$ 0.2 % in SNPs, $\sim$ 2.4 % in PRPs
- Intersection of MB and ML calls provides highly reliable SNP predictions
  - Used to genotype many more rice cultivars
- Polymorphic region predictions
  - Important for more detailed analyses (e.g. dideoxy sequencing)
  - Useful for primer design to increase PCR success rates

Conclusions

- Created the first whole-genome inventory of polymorphisms for rice
  - Highly polymorphic ~ 0.2 % in SNPs, ~ 2.4 % in PRPs
- Intersection of MB and ML calls provides highly reliable SNP predictions
  - Used to genotype many more rice cultivars
- Polymorphic region predictions
  - Important for more detailed analyses (e.g. dideoxy sequencing)
  - Useful for primer design to increase PCR success rates

Acknowledgements

Friedrich Miescher Laboratory
of the Max Planck Society

Friedrich Miescher Laboratory

Gunnar Rätsch

Georg Zeller

Gabriele Schweikert

MPI for Developmental Biology

Detlef Weigel

Richard Clark

Michigan State University, USA

Robin Buell

Kevin Childs

Perlegen Sciences, USA

Renee Stokowski

Dennis Ballinger

Kelly Frazer

David Cox

IRRI, The Philippines

Kenneth McNally

Victor Ulat

Hei Leung

Colorado State University, USA

Jan Leach