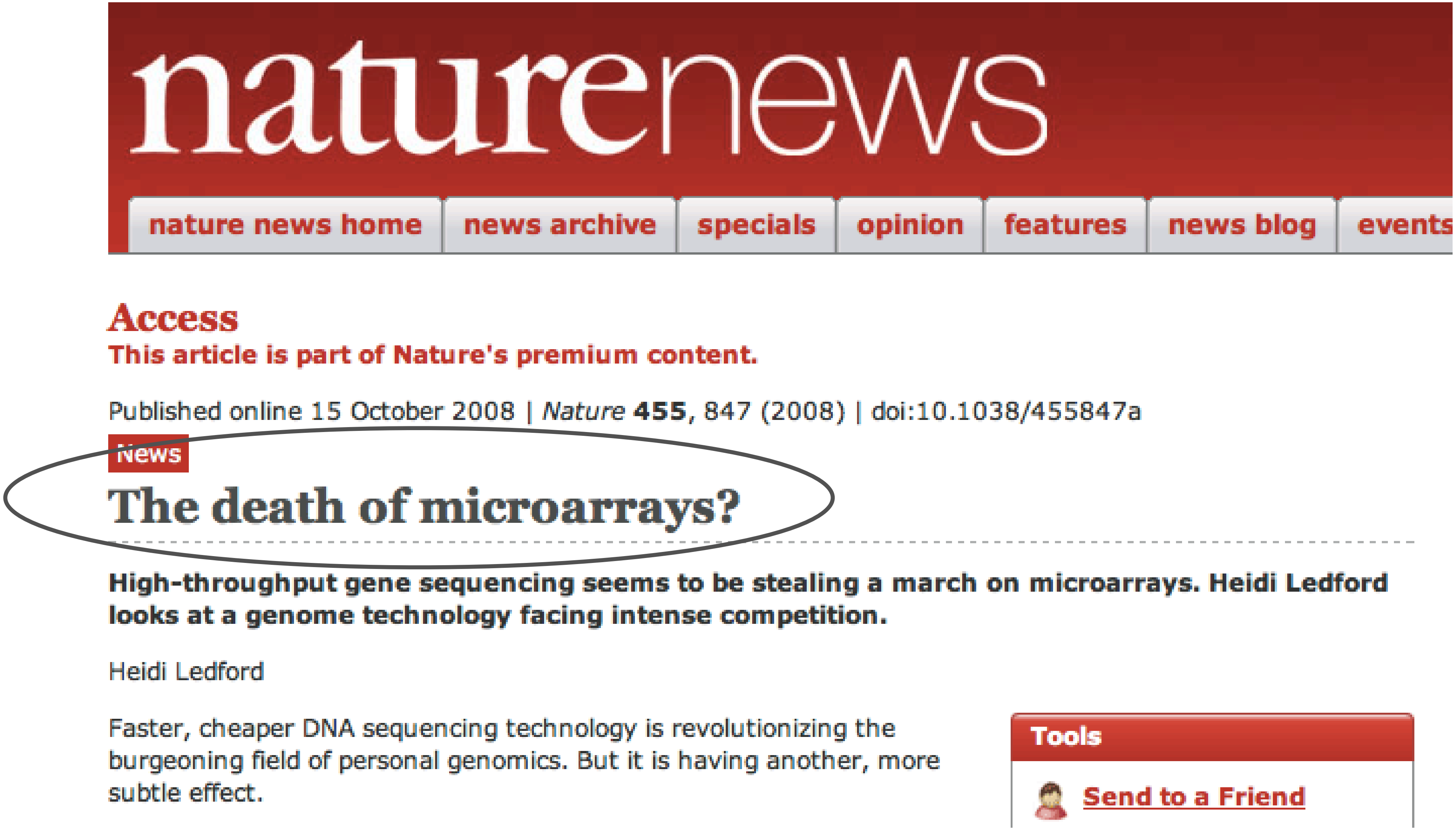
Model-Based Quality Assessment and Base-Calling For Second-Generation Sequencing

Héctor Corrada Bravo & Rafael A. Irizarry
Biostatistics Dept.
Bloomberg School of Public Health
Johns Hopkins University

Second-Generation Sequencing



Access

This article is part of Nature's premium content.

Published online 15 October 2008 | *Nature* **455**, 847 (2008) | doi:10.1038/455847a

News

The death of microarrays?

High-throughput gene sequencing seems to be stealing a march on microarrays. Heidi Ledford looks at a genome technology facing intense competition.

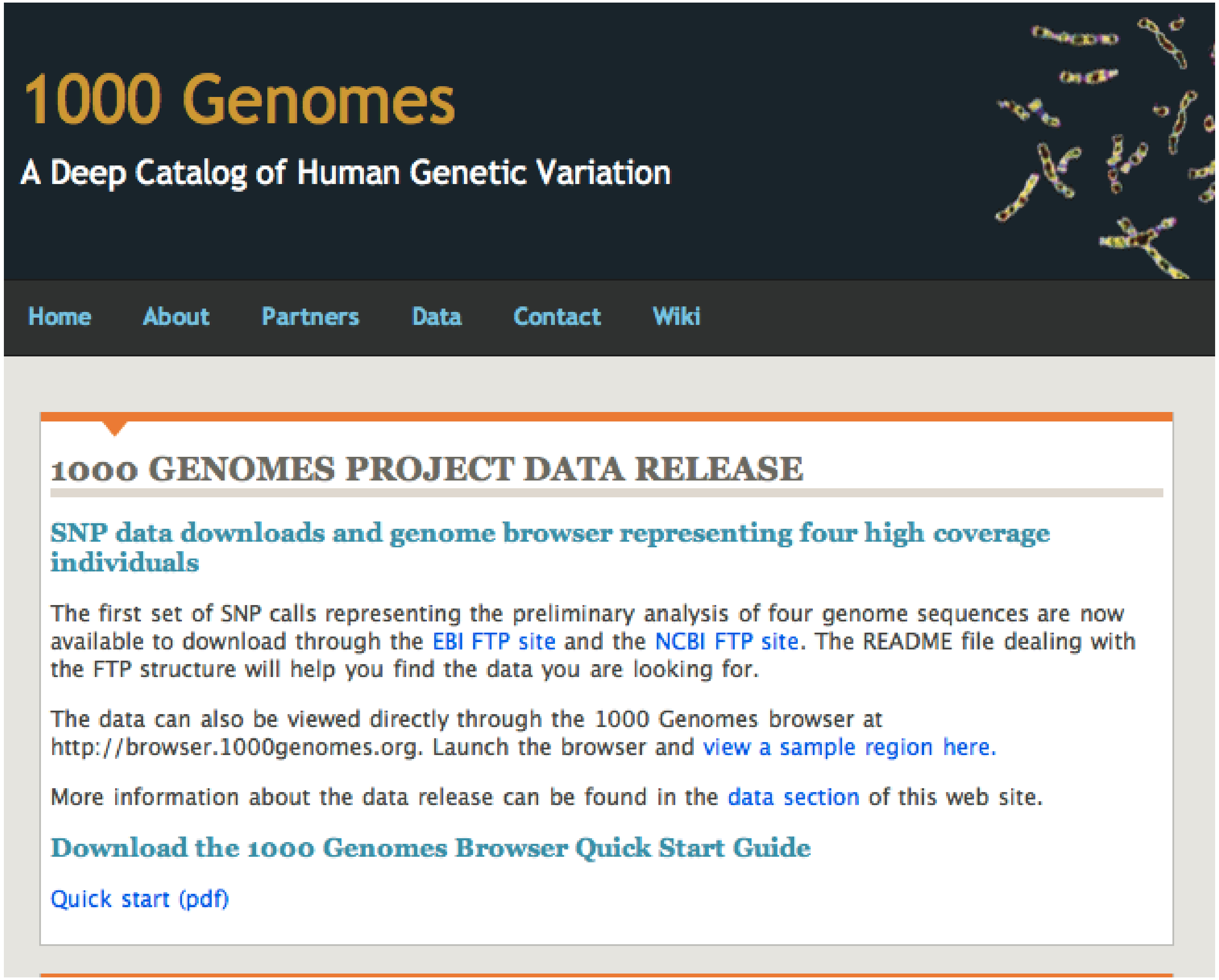
Heidi Ledford

Faster, cheaper DNA sequencing technology is revolutionizing the burgeoning field of personal genomics. But it is having another, more subtle effect.

Second-Generation Sequencing

- "Ultra high throughput" DNA sequencing
  - 3 gigabases / week vs.
  - 3 gigabases / 13 years...

1000 Genomes Project



1000 Genomes

A Deep Catalog of Human Genetic Variation

1000 GENOMES PROJECT DATA RELEASE

SNP data downloads and genome browser representing four high coverage individuals

The first set of SNP calls representing the preliminary analysis of four genome sequences are now available to download through the EBI FTP site and the NCBI FTP site. The README file dealing with the FTP structure will help you find the data you are looking for.

The data can also be viewed directly through the 1000 Genomes browser at http://browser.1000genomes.org. Launch the browser and view a sample region here.

More information about the data release can be found in the data section of this web site.

Download the 1000 Genomes Browser Quick Start Guide

Quick start (pdf)

Platforms

illumina

454 SEQUENCING

Roche

AB Applied Biosystems

- Millions of short DNA fragments (~36-70 bp in Illumina platform) sequenced in parallel

(Third-Generation) Platforms

Oxford NANOPORE™ Technologies

PACIFIC BIOSCIENCES™

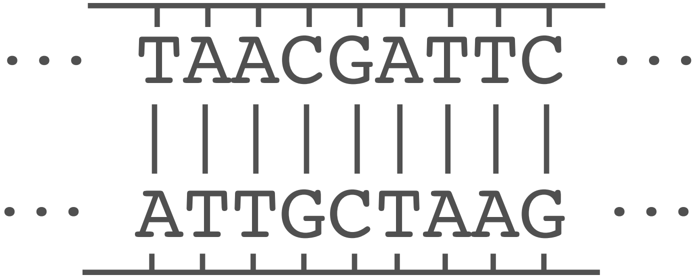
Complete Genomics

- Single-molecule sequencing
    - "the 15-minute genome"

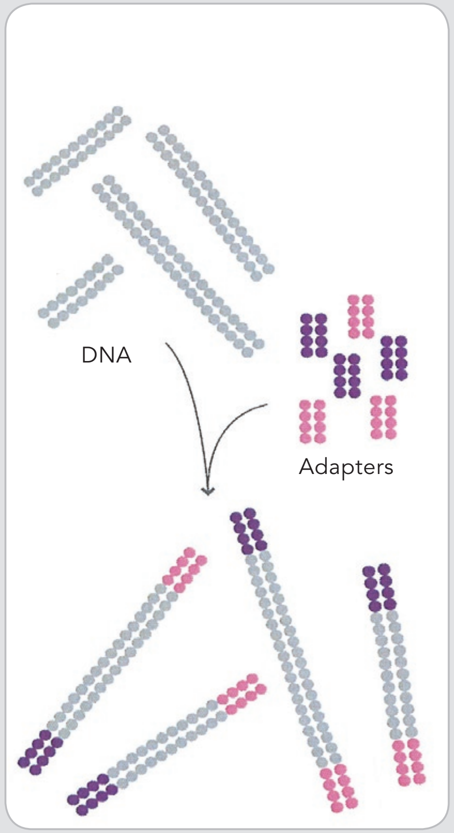
Outline

1. Second-generation sequencing (sec-gen) technology review (Illumina/Solexa)
2. Genotyping w/ sec-gen sequencing
3. Statistical/Computational challenges
4. Model-based base-calling
5. Model-based quality assessment

Illumina/Solexa



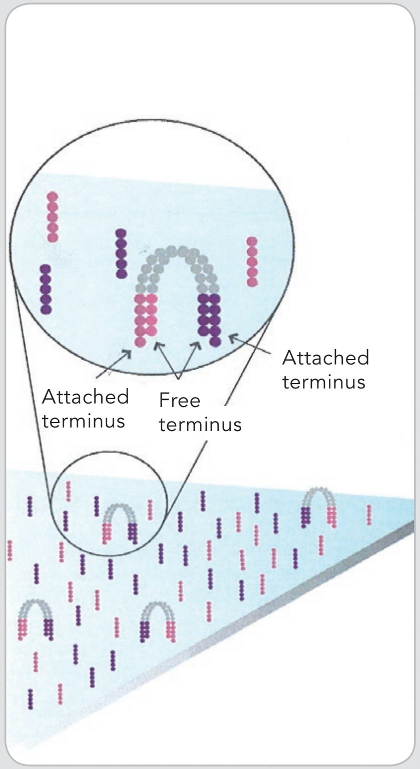
1. PREPARE GENOMIC DNA SAMPLE



Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

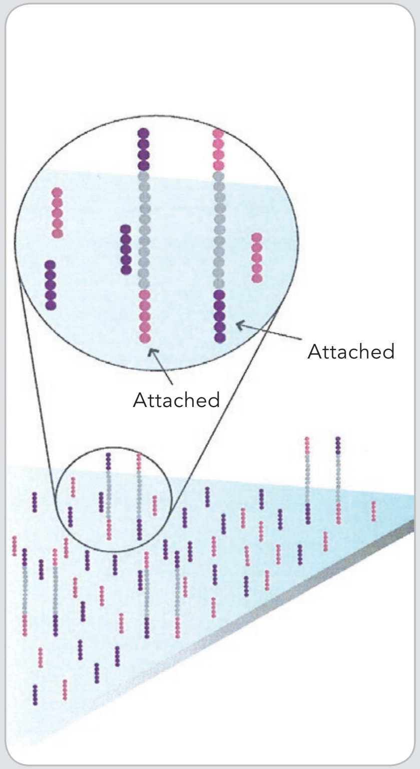
Illumina/Solexa

4. Fragments become double-stranded



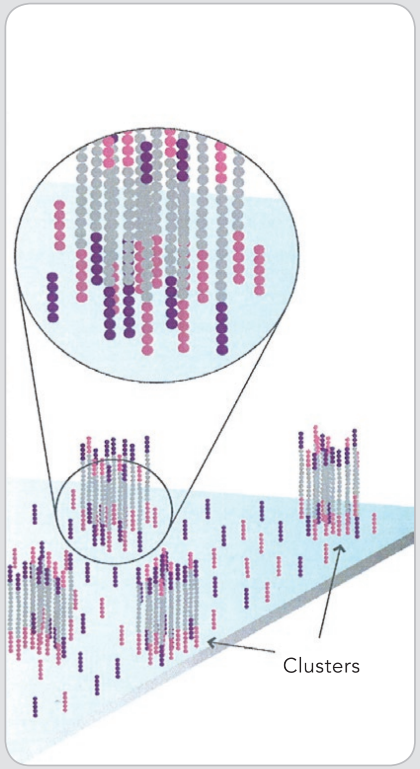
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. Denature the double-stranded molecules



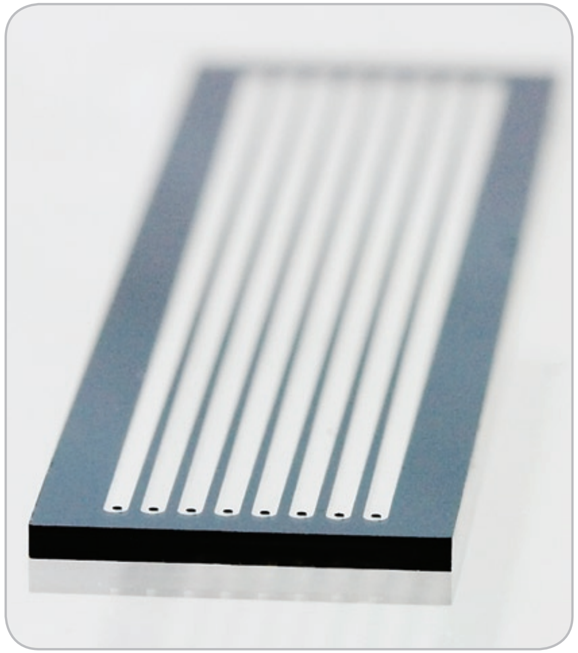
Denaturation leaves single-stranded templates anchored to the substrate.

6. Complete amplification



Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Illumina/Solexa



- Eight lanes
- 330 tiles/lane
- ~30K fragments per tile
- ~80M short sequences per run

A Set of Short Reads

```
GTTGAGGCTTGCGTTTTTGGTACGCTGGACTTTGT
GTACTCGTCGCTGCGTTGAGGCTTGCGTTTTTGGT
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTTGCGTTTATGGTACGCTGGACTTTGTAGGATAC
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
GCGTTTATGGTACGCTGGACTTTGTAGGATACCCT
GAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGG
GCGTTGAGGCTTGCGTTTATGGTACGCTGGATTTT
CGTTTATGGTACGCTGGACTTTGTAGGATACCCTC
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
GTTTATGGTACGCTGGACTTTGTAGGATACCCTCG
TCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTA
TGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTA
GCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTAC
TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
TCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTTTG
CGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCT
GTTGAGGCTTGCGTTTATGGTACGCTGGGCTTTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
```

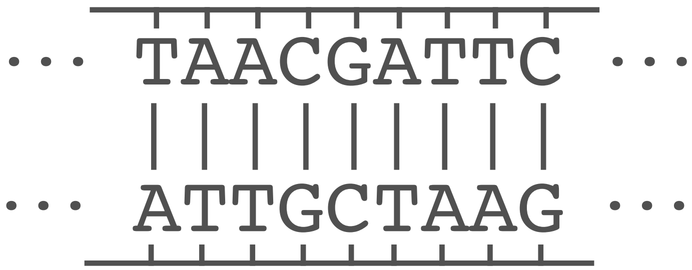
Matching

```
                   GTTGAGGCTTGCGTTTTTGGTACGCTGGACTTTGT
     GTACTCGTCGCTGCGTTGAGGCTTGCGTTTTTGGT
                                   ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
                           TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
                          CTTGCGTTTATGGTACGCTGGACTTTGTAGGATAC
                           TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
                             GCGTTTATGGTACGCTGGACTTTGTAGGATACCCT
                      GAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGG
                 GCGTTGAGGCTTGCGTTTATGGTACGCTGGATTTT
                              CGTTTATGGTACGCTGGACTTTGTAGGATACCCTC
                                   ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
                               GTTTATGGTACGCTGGACTTTGTAGGATACCCTCG
 TCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTA
      TGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTA
       GCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTAC
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
   TCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTTTG
          CGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCT
                   GTTGAGGCTTGCGTTTATGGTACGCTGGGCTTTTT
                           TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTTC
```

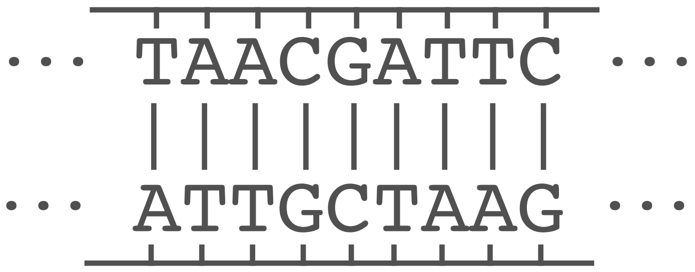
Applications

- *de novo* sequencing, resequencing
- Genotyping, copy number variation
- RNA-seq, microRNA-seq: transcriptome analysis
- ChIP-seq: transcription factor binding sites
- Methyl-seq: methylation detection

Genotyping

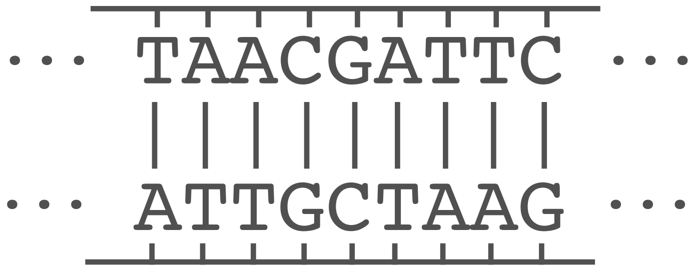


Genotyping



... TAACGATTC ...

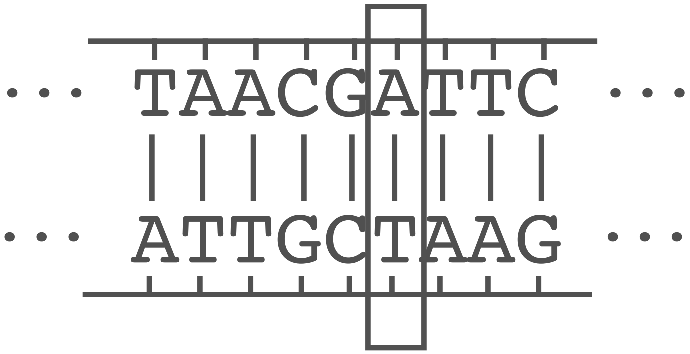
... ATTGCTAAG ...

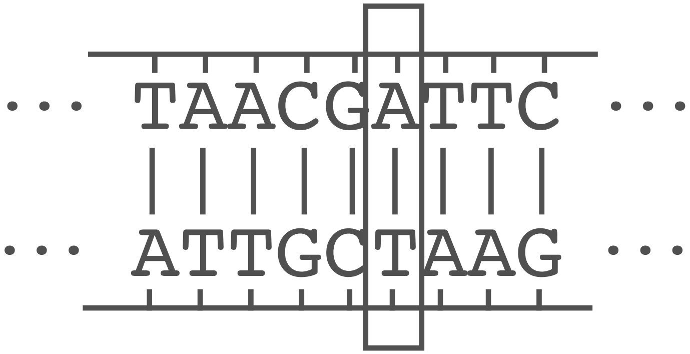


... TAACGATTC ...

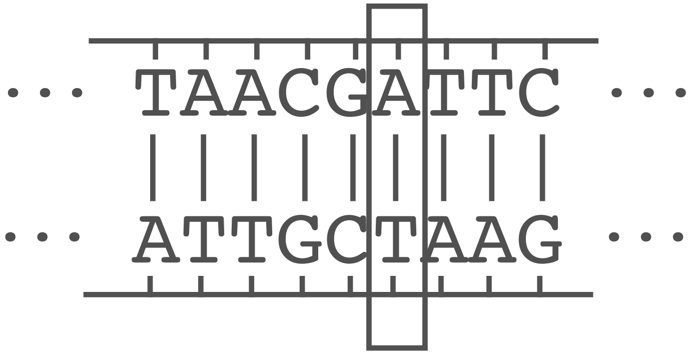
... ATTGCTAAG ...

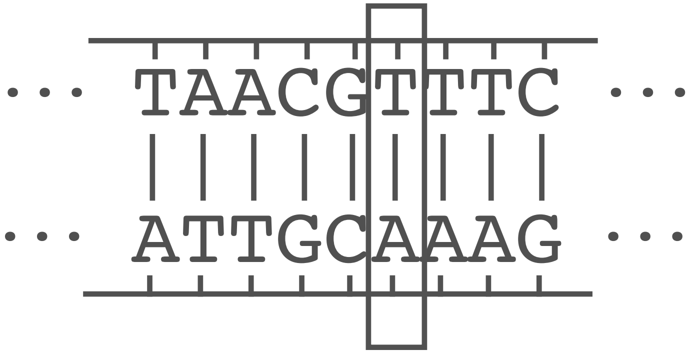
Genotyping

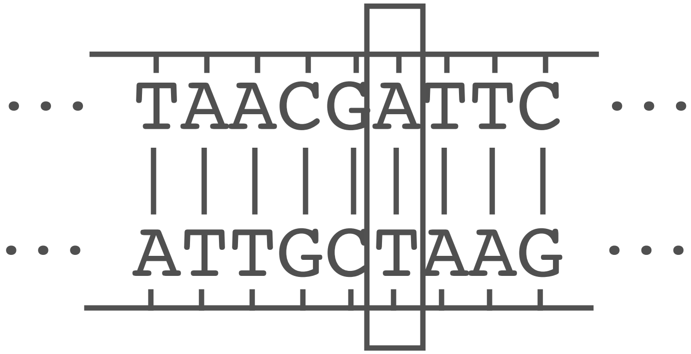


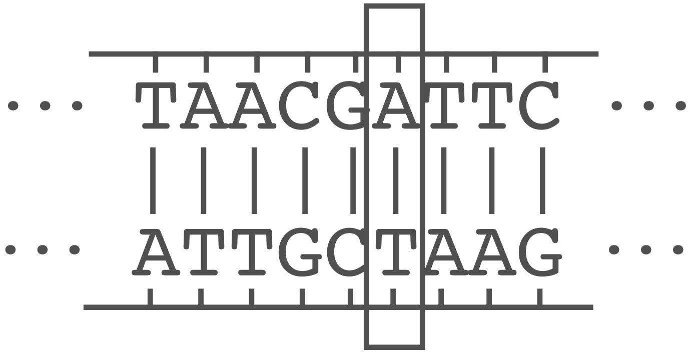


Genotyping

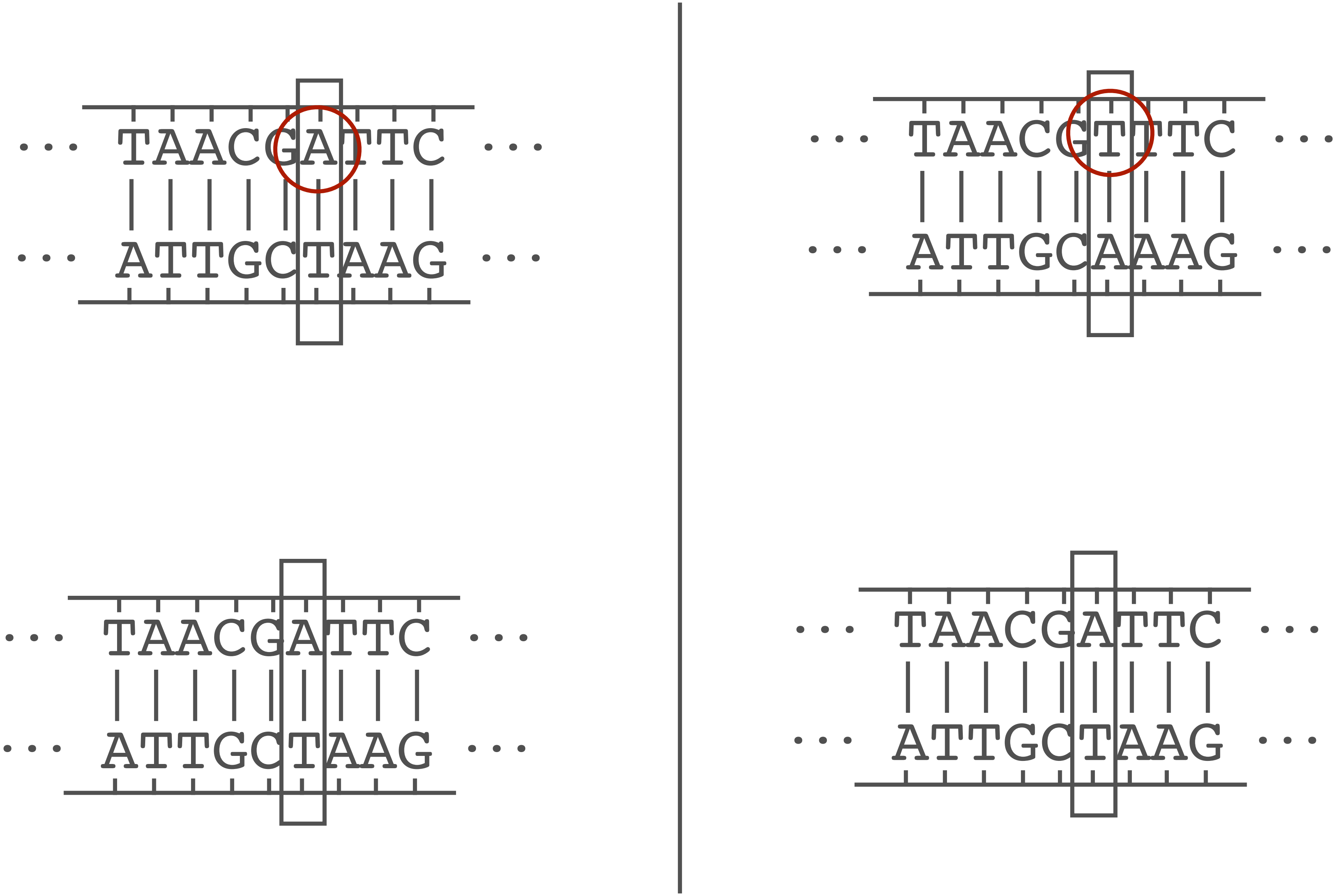

... TAACGATTC ...
... ATTGCTAAG ...


... TAACGTTTC ...
... ATTGCAAAG ...

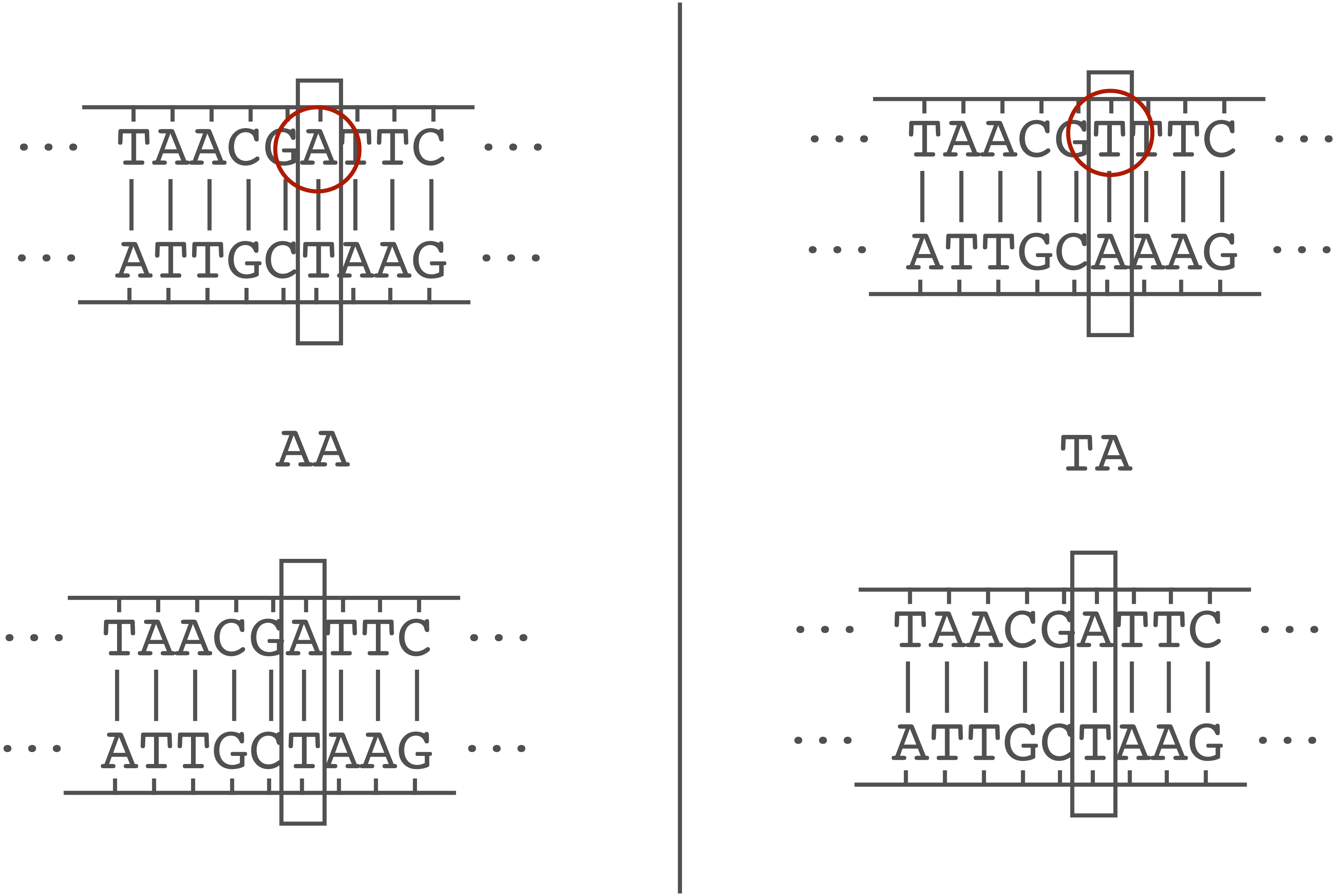

... TAACGATTC ...
... ATTGCTAAG ...


... TAACGATTC ...
... ATTGCTAAG ...

Genotyping


... TAACGATTC ...
... ATTGCTAAG ...
... TAACGTTTC ...
... ATTGCAAAG ...
... TAACGATTC ...
... ATTGCTAAG ...
... TAACGATTC ...
... ATTGCTAAG ...

Genotyping



... TAACGATTC ...

... ATTGCTAAG ...

AA

... TAACGATTC ...

... ATTGCTAAG ...

... TAACGTTTC ...

... ATTGCAAAG ...

TA

... TAACGATTC ...

... ATTGCTAAG ...

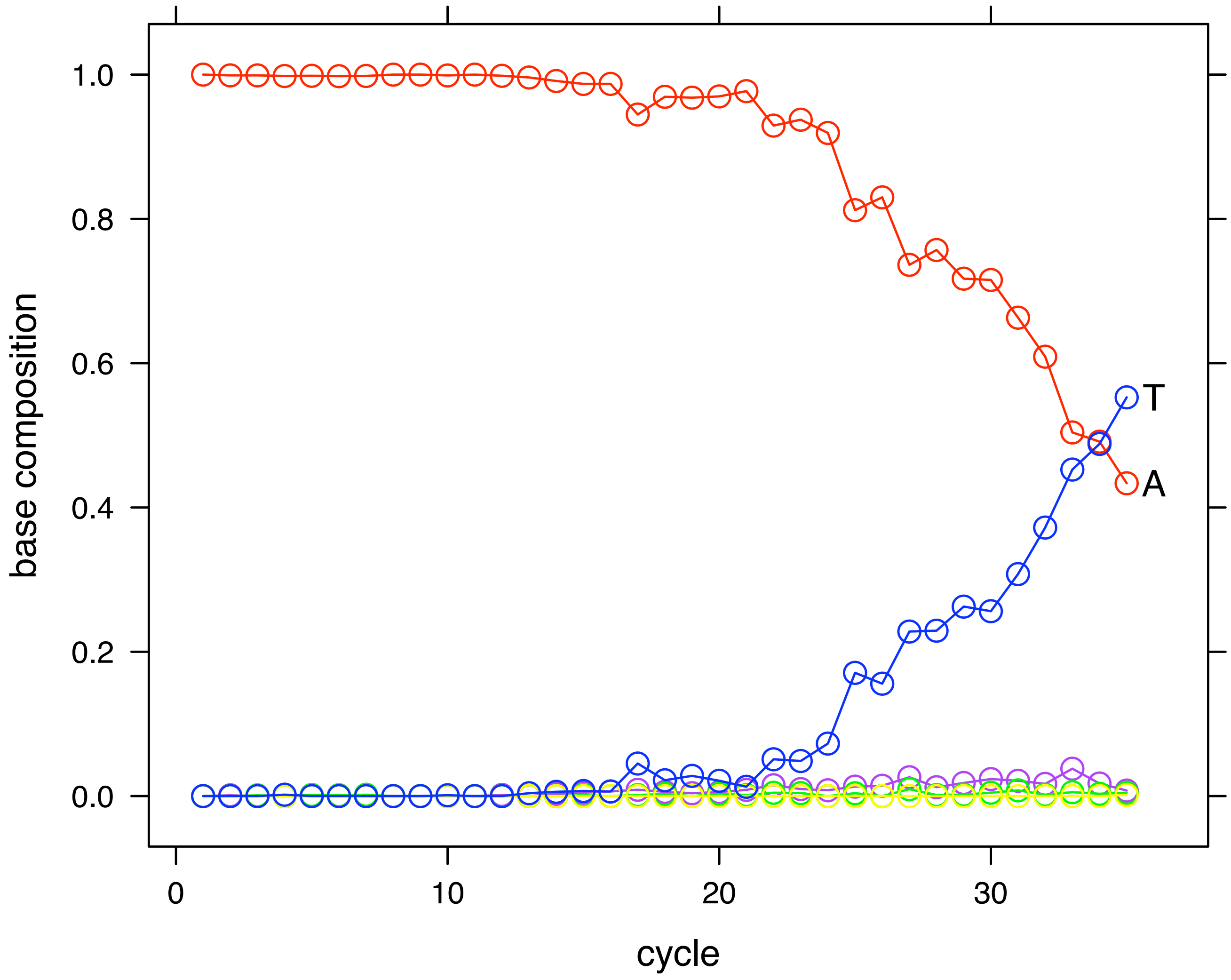
SNPs

```
                   GTTGAGGCTTGCGTTTTTGGTACGCTGGACTTTGT
     GTACTCGTCGCTGCGTTGAGGCTTGCGTTTTTGGT
                                   ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
                           TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
                          CTTGCGTTTATGGTACGCTGGACTTTGTAGGATAC
                           TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
                             GCGTTTATGGTACGCTGGACTTTGTAGGATACCCT
                      GAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGG
                 GCGTTGAGGCTTGCGTTTATGGTACGCTGGATTTT
                              CGTTTATGGTACGCTGGACTTTGTAGGATACCCTC
                                   ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
                               GTTTATGGTACGCTGGACTTTGTAGGATACCCTCG
 TCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTA
      TGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTA
       GCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTAC
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
   TCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTTTG
          CGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCT
                   GTTGAGGCTTGCGTTTATGGTACGCTGGGCTTTTT
                           TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTTC
```

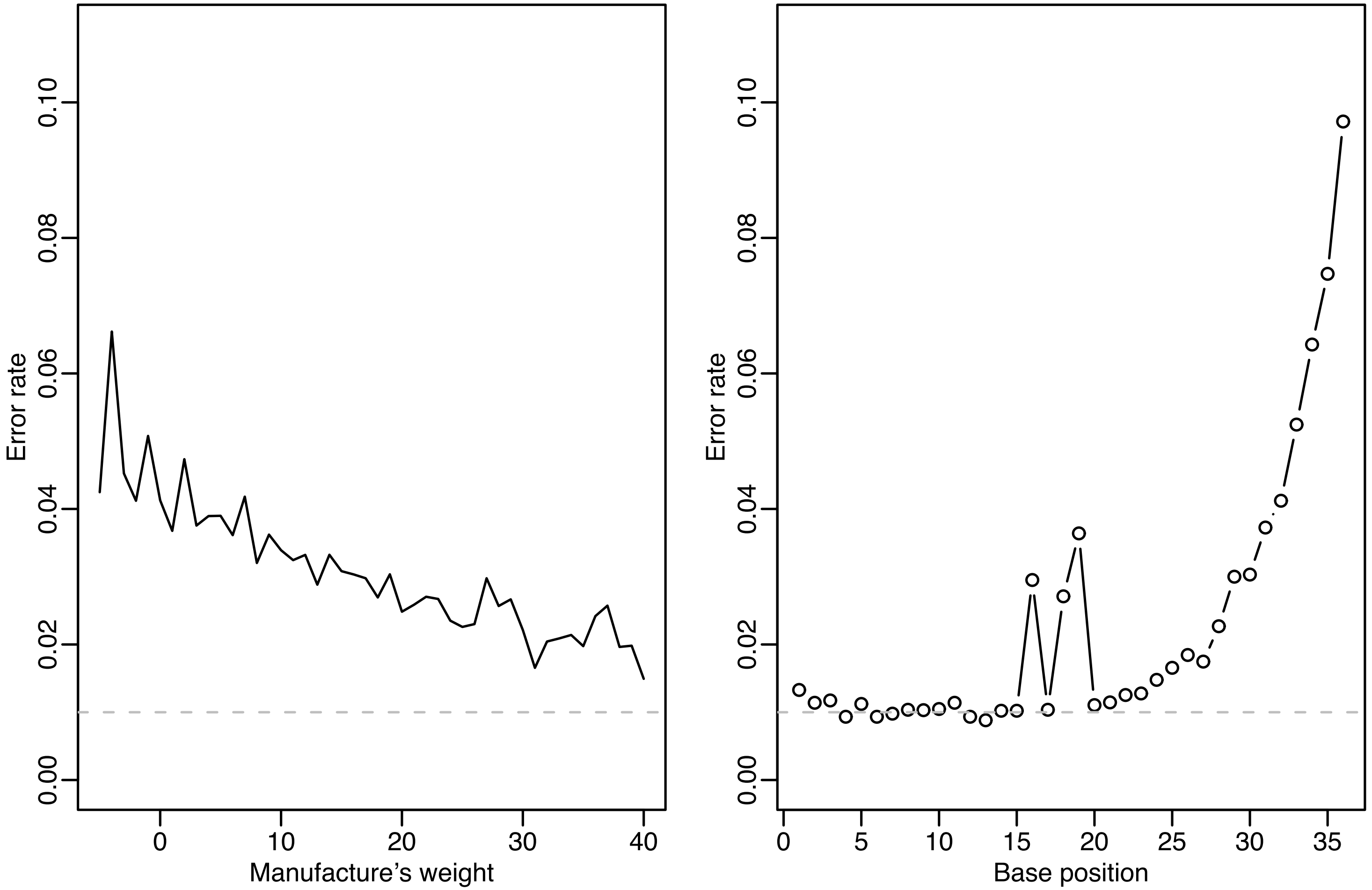
SNPs

```
 TCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTA
   TCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTTTG
     GTACTCGTCGCTGCGTTGAGGCTTGCGTTTTTGGT
      TGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTA
       GCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTAC
          CGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCT
                 GCGTTGAGGCTTGCGTTTATGGTACGCTGGATTTT
                   GTTGAGGCTTGCGTTTTTGGTACGCTGGACTTTGT
                   GTTGAGGCTTGCGTTTATGGTACGCTGGGCTTTTT
                      GAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGG
                          CTTGCGTTTATGGTACGCTGGACTTTGTAGGATAC
                           TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
                           TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
                           TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
                             GCGTTTATGGTACGCTGGACTTTGTAGGATACCCT
                              CGTTTATGGTACGCTGGACTTTGTAGGATACCCTC
                               GTTTATGGTACGCTGGACTTTGTAGGATACCCTCG
                                  TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
                                   ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
                                   ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
CTCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTTC
```

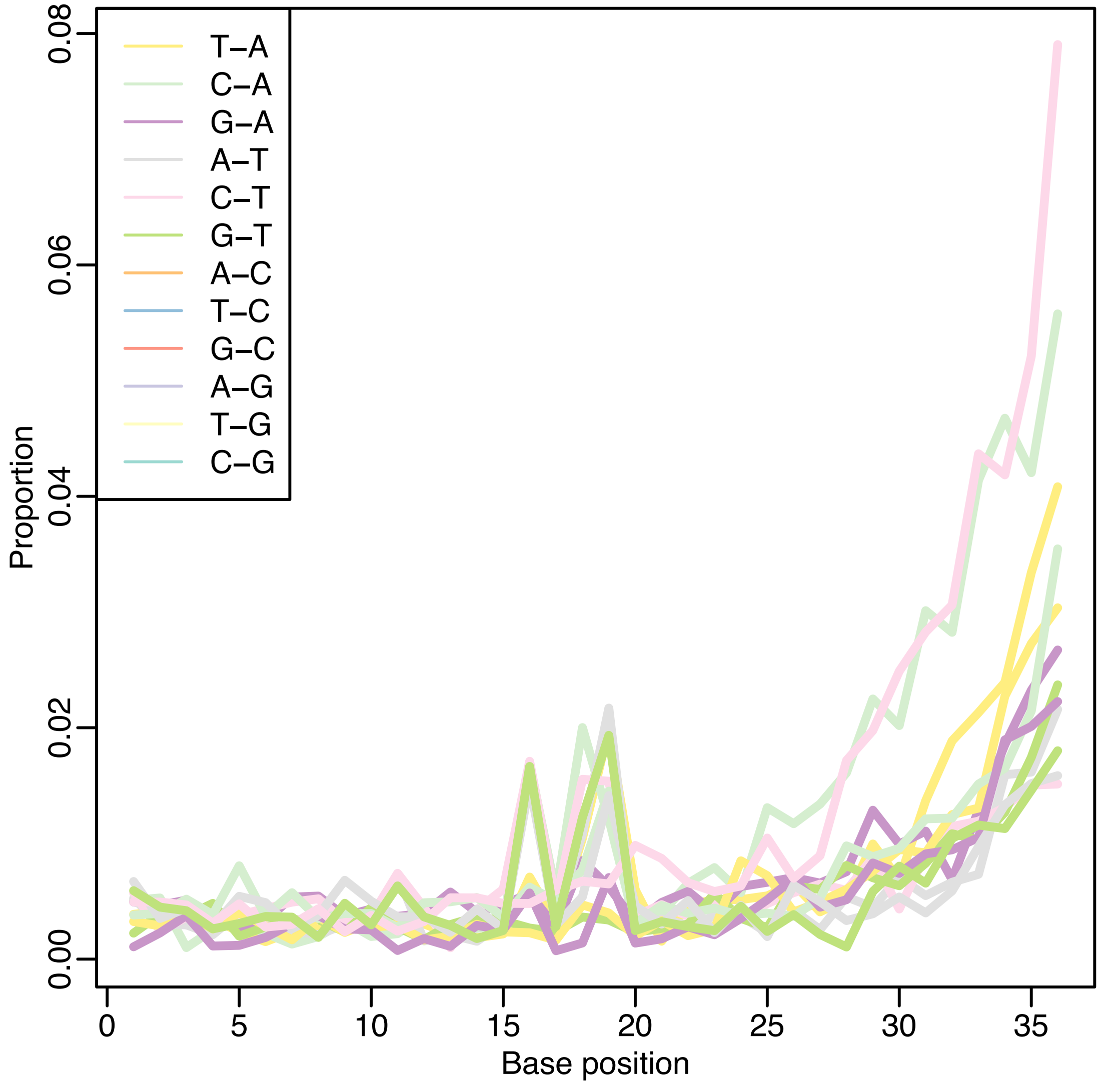
SNPs



Error Rate and Reported Quality

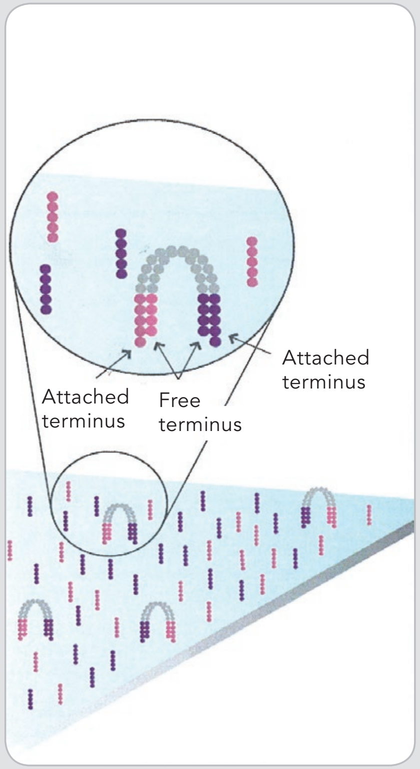


Systematic Biases



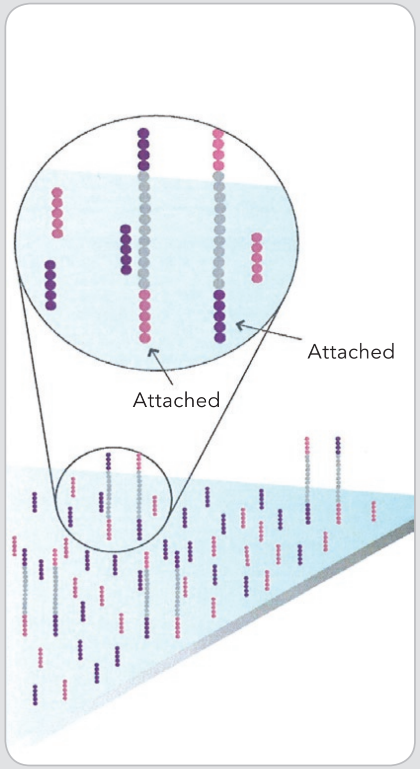
Illumina/Solexa

4. FRAGMENTS BECOME DOUBLE-STRANDED



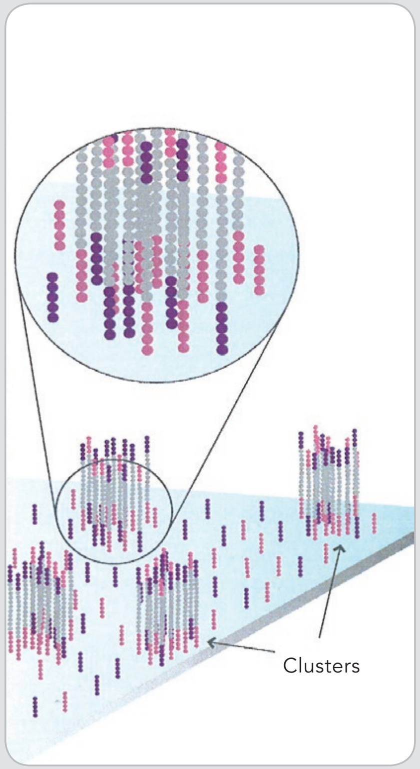
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES



Denaturation leaves single-stranded templates anchored to the substrate.

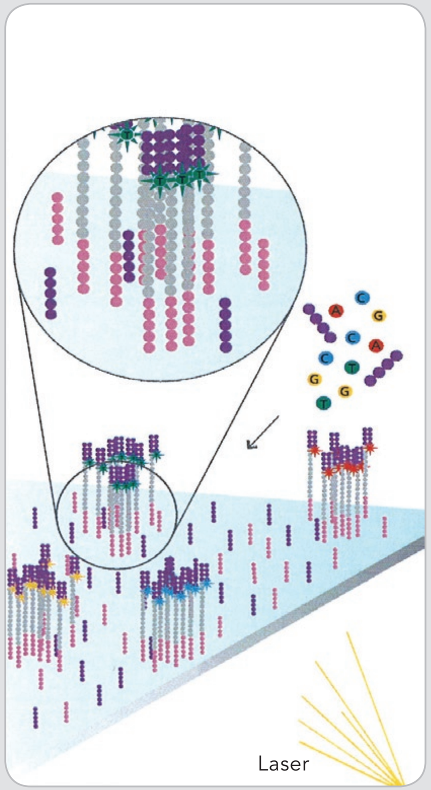
6. COMPLETE AMPLIFICATION



Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

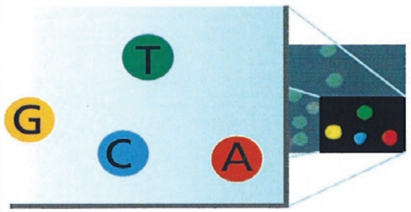
Illumina/Solexa

7. DETERMINE FIRST BASE



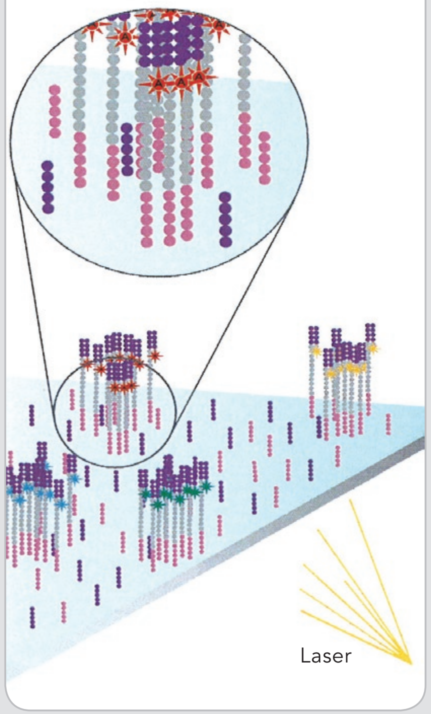
The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

8. IMAGE FIRST BASE



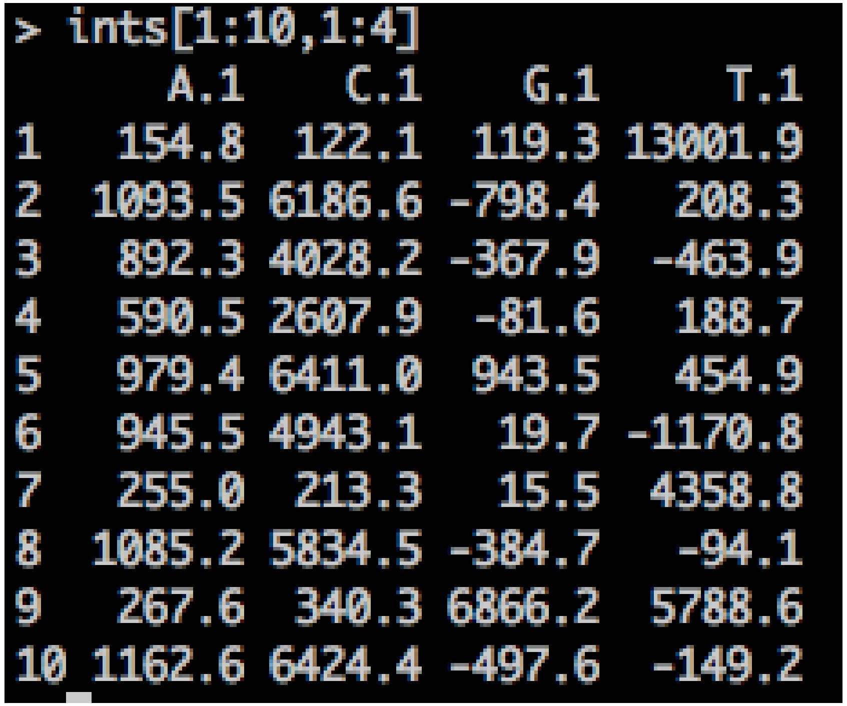
After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

9. DETERMINE SECOND BASE



The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

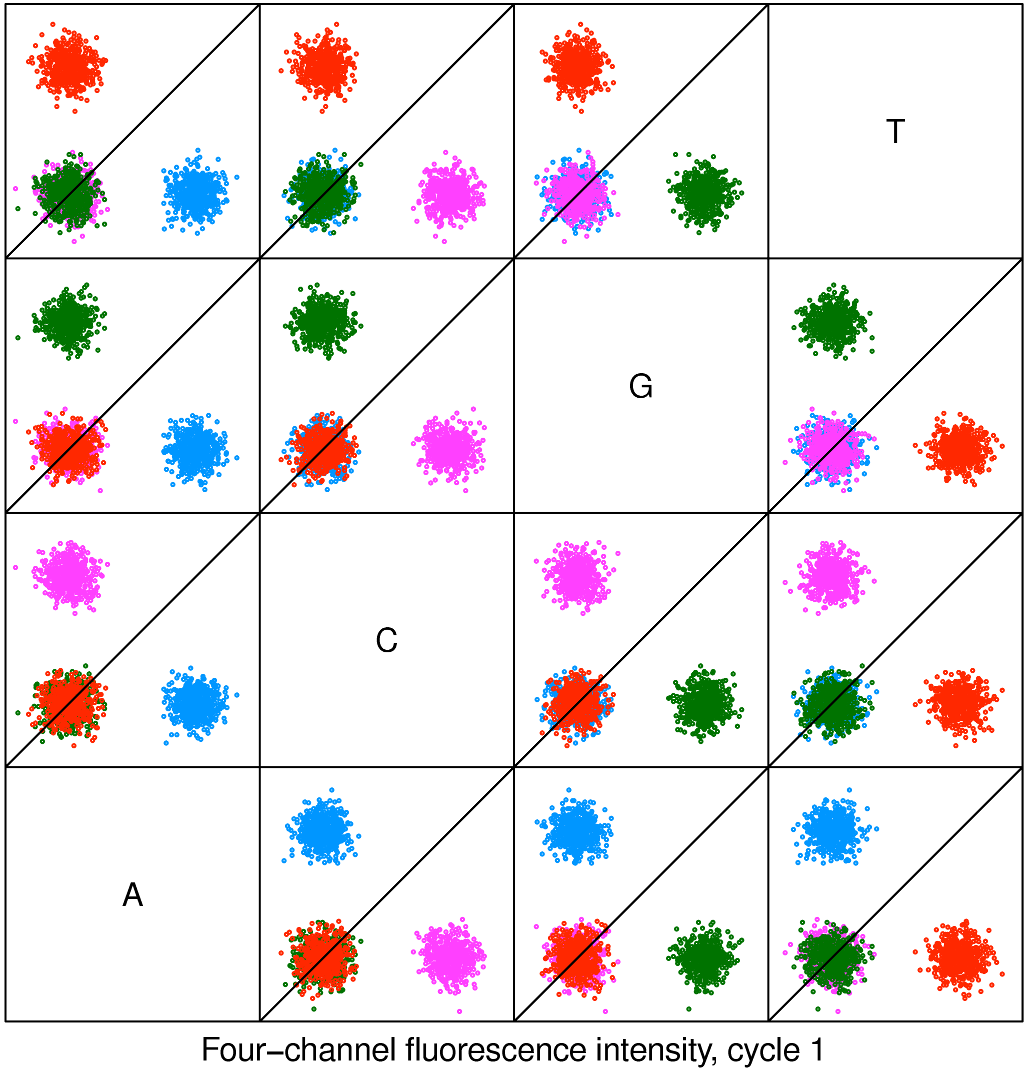
Fluorescence Intensity



```
> ints[1:10,1:4]
       A.1    C.1    G.1     T.1
1    154.8  122.1  119.3 13001.9
2   1093.5 6186.6 -798.4   208.3
3    892.3 4028.2 -367.9  -463.9
4    590.5 2607.9  -81.6   188.7
5    979.4 6411.0  943.5   454.9
6    945.5 4943.1   19.7 -1170.8
7    255.0  213.3   15.5  4358.8
8   1085.2 5834.5 -384.7   -94.1
9    267.6  340.3 6866.2  5788.6
10  1162.6 6424.4 -497.6  -149.2
```

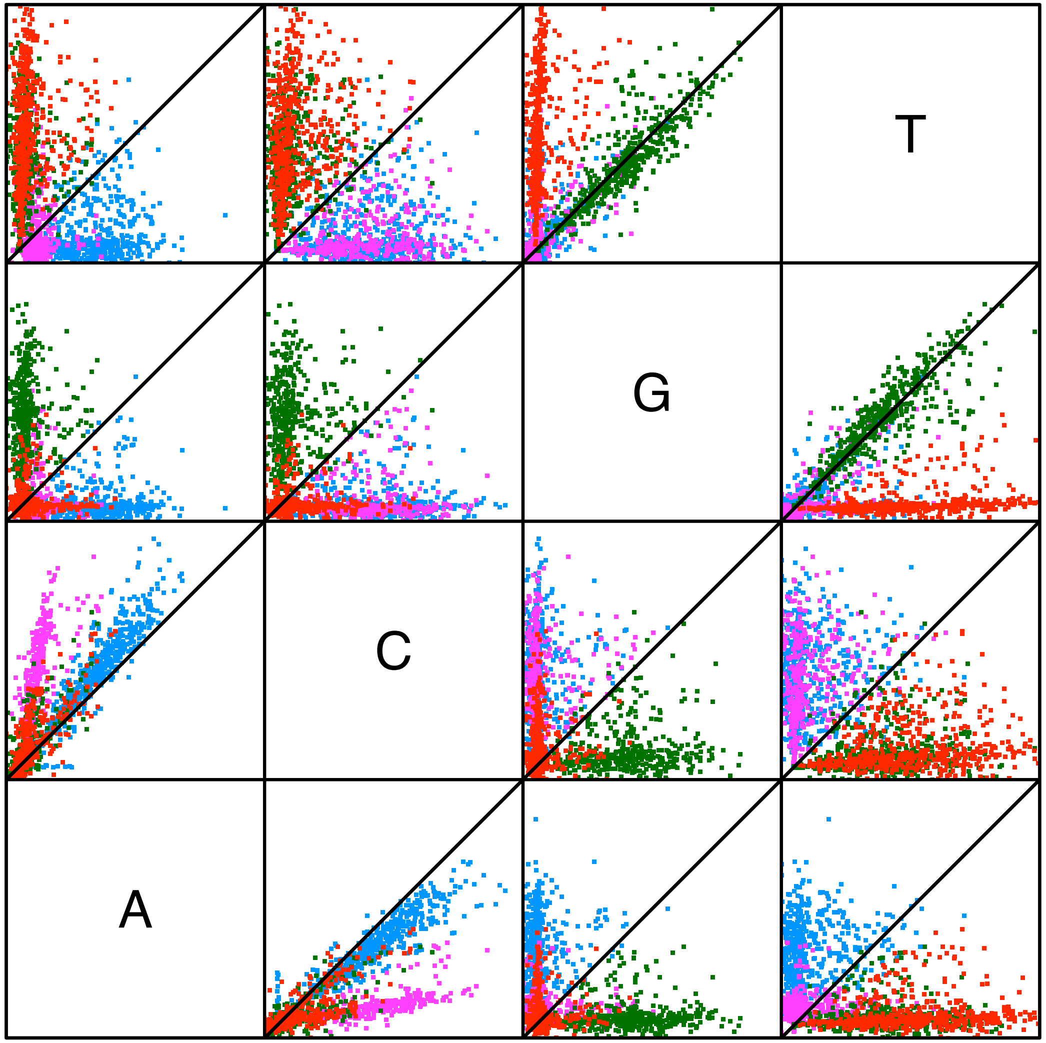
- For read *n,* cycle *i,* we observe an intensity vector of size 4

A Thought Experiment



Color coded by call made: A, C, G, T

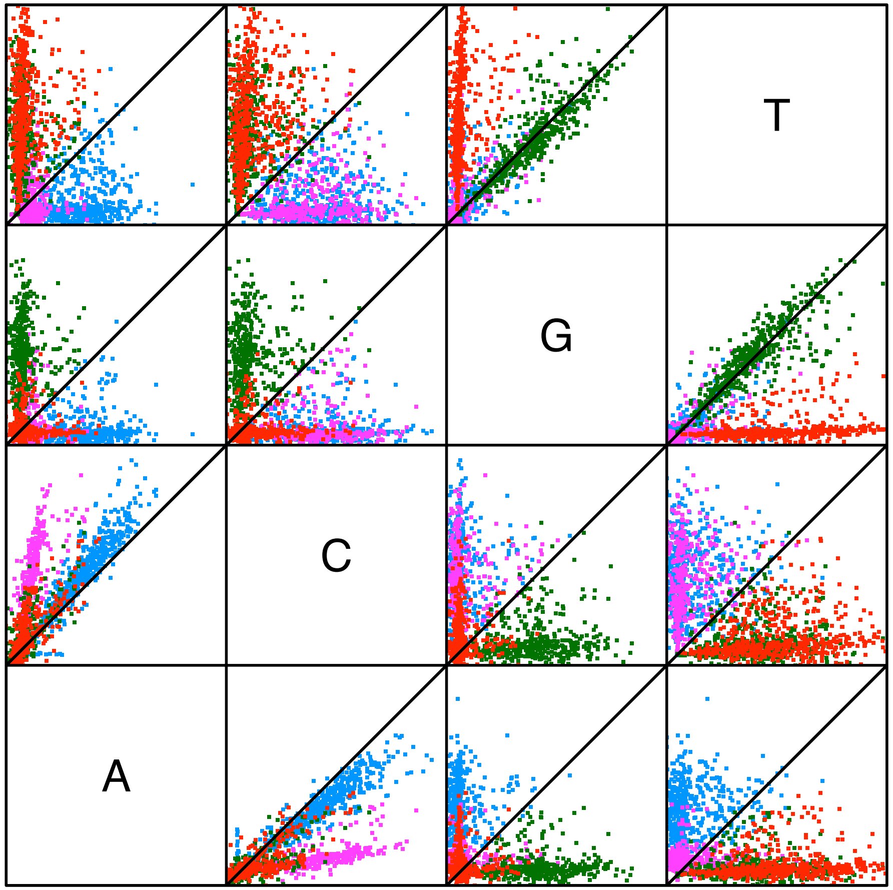
Fluorescence Intensity



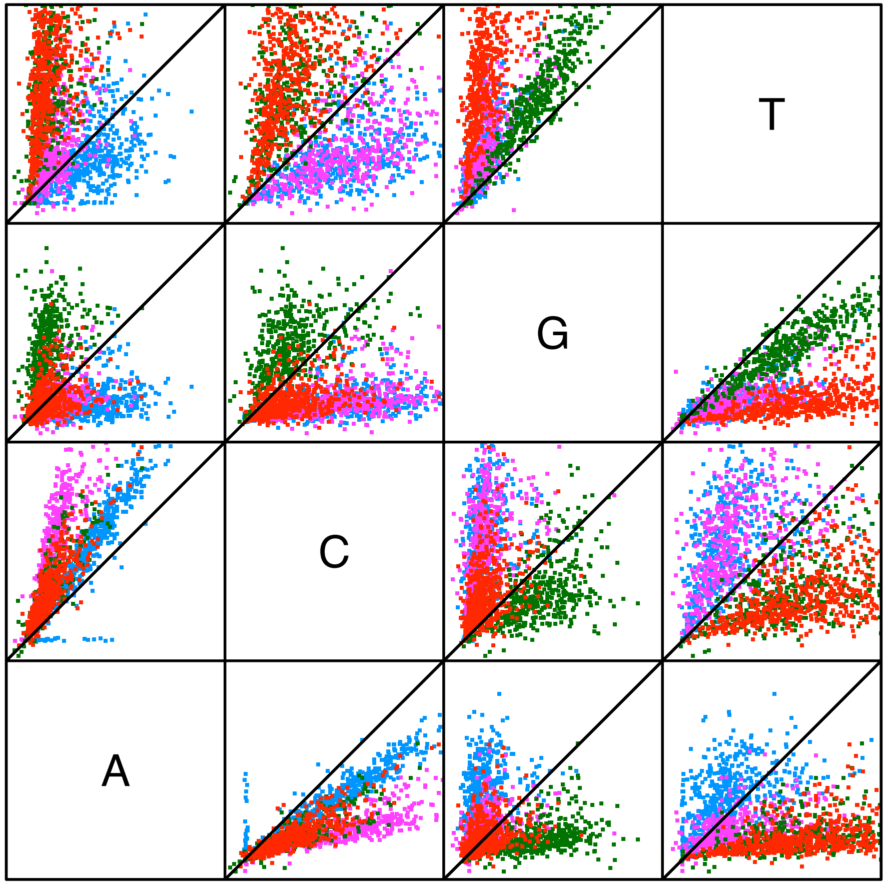
Four–channel fluorescence intensity, cycle 1

Color coded by call made: A, C, G, T

Fluorescence Intensity



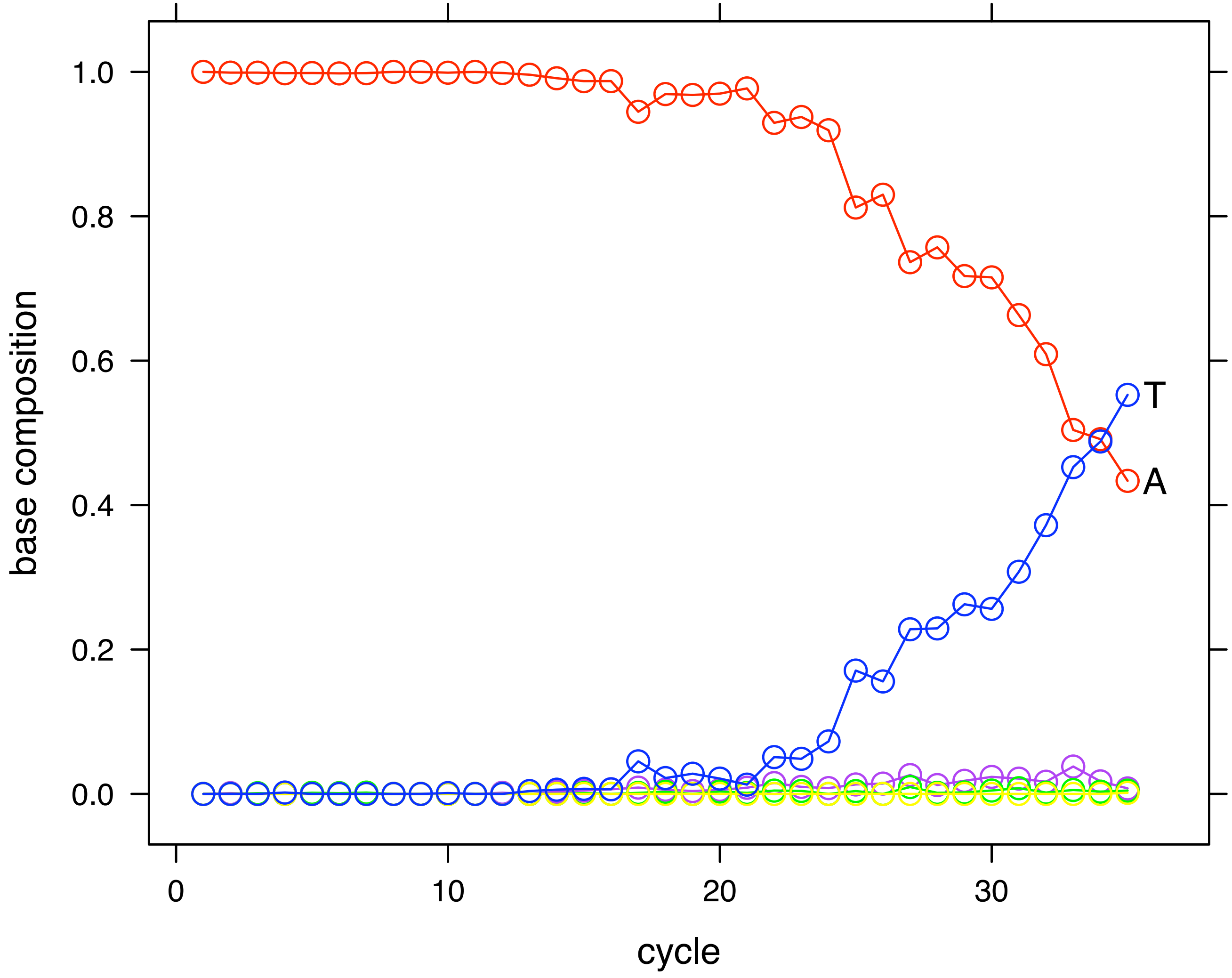
Four–channel fluorescence intensity, cycle 1



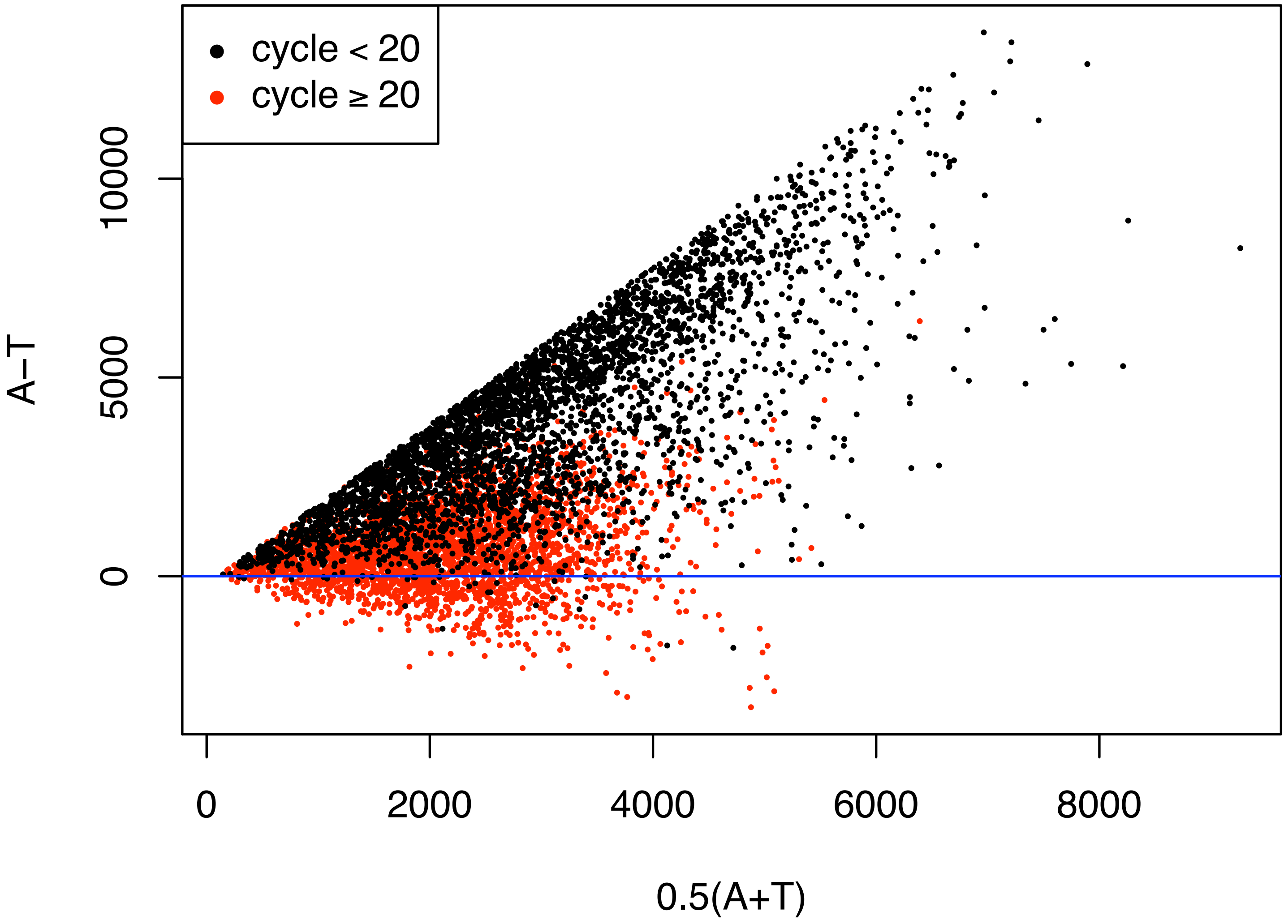
Four–channel fluorescence intensity, cycle 25

Color coded by call made: A, C, G, T

SNPs



SNP Intensities



Challenges

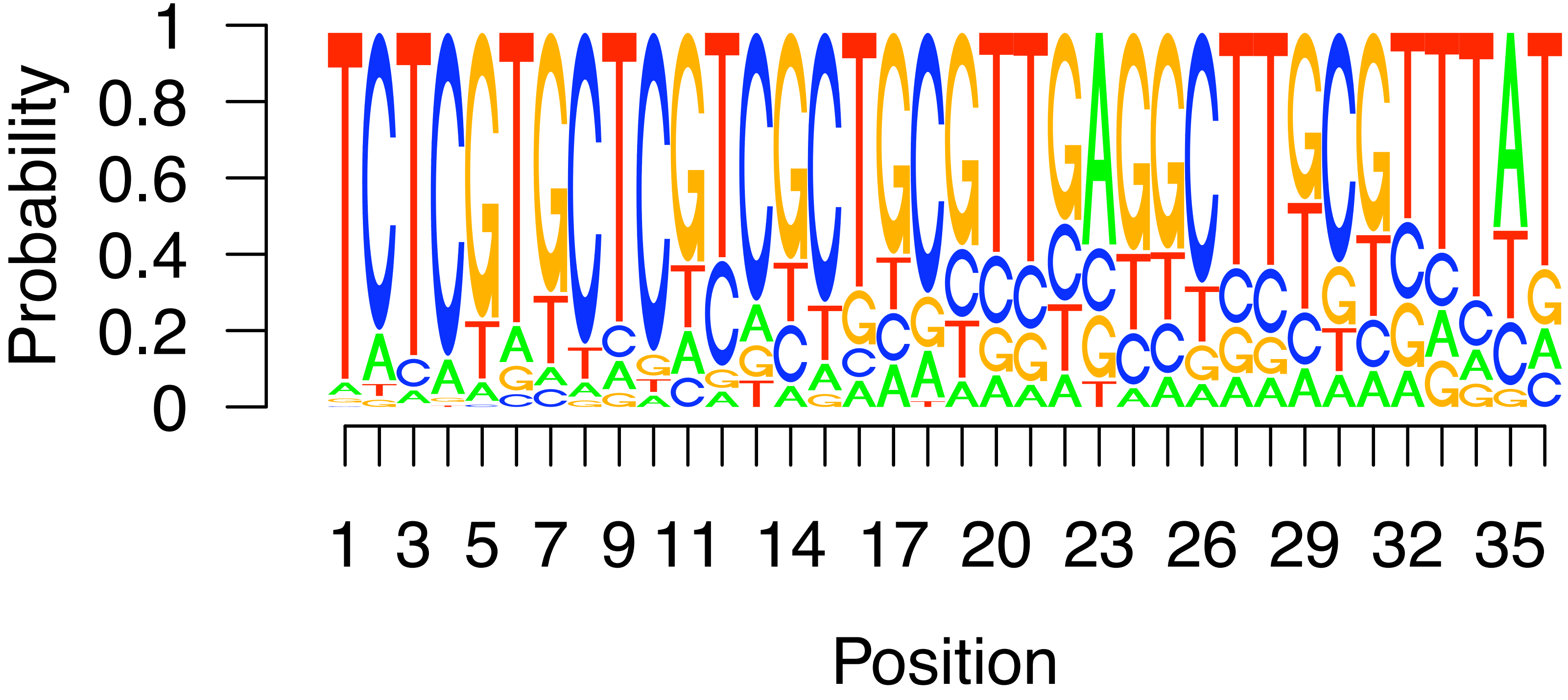
- Base-calling is the result of a complicated procedure on noisy data
- Not all base-calls are made with the same certainty
- Statistical: What is the proper way of modeling this uncertainty?
- Computational: Can we use this model at sec-gen data scale?

Capturing Uncertainty

- For read *n,* we observe over *k* cycles, a *4-by-k* matrix of intensities y_n
- Genome is a set of candidates $\Theta \subseteq \{A, C, G, T\}^k$
- Denote the "true" *k*-mer in genome sequenced by read *n* as $\tilde{\theta} \in \Theta$
- Probability profile is given by

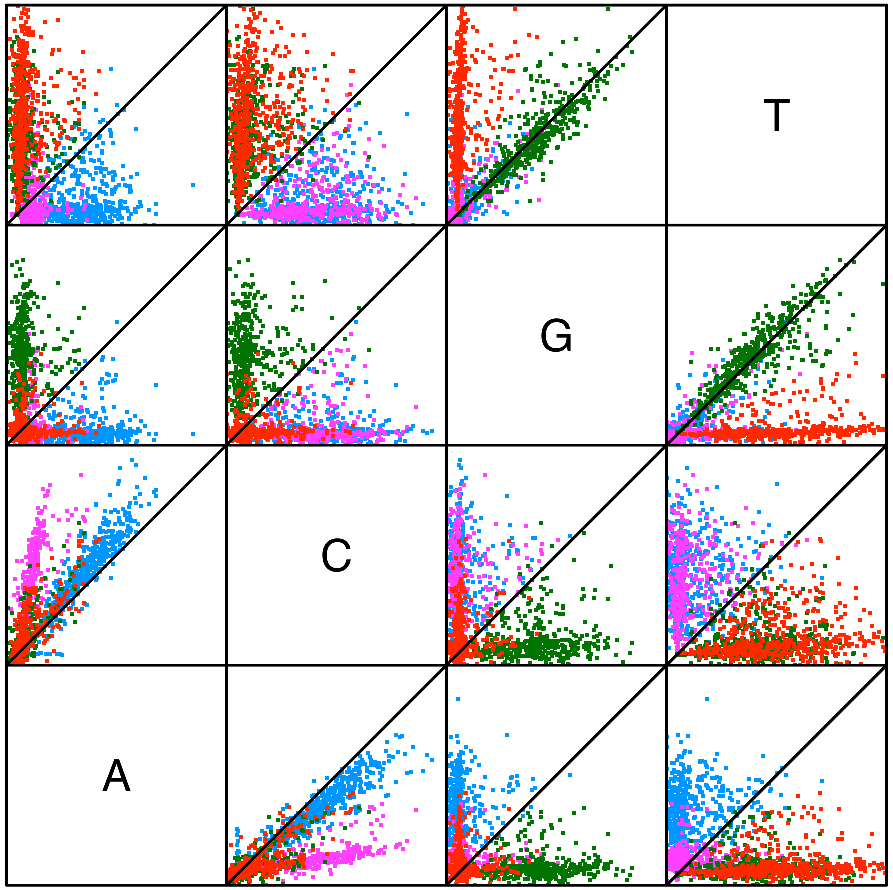
$$\Pr(\theta = \tilde{\theta} | y)$$

Base Identity Probability Profiles

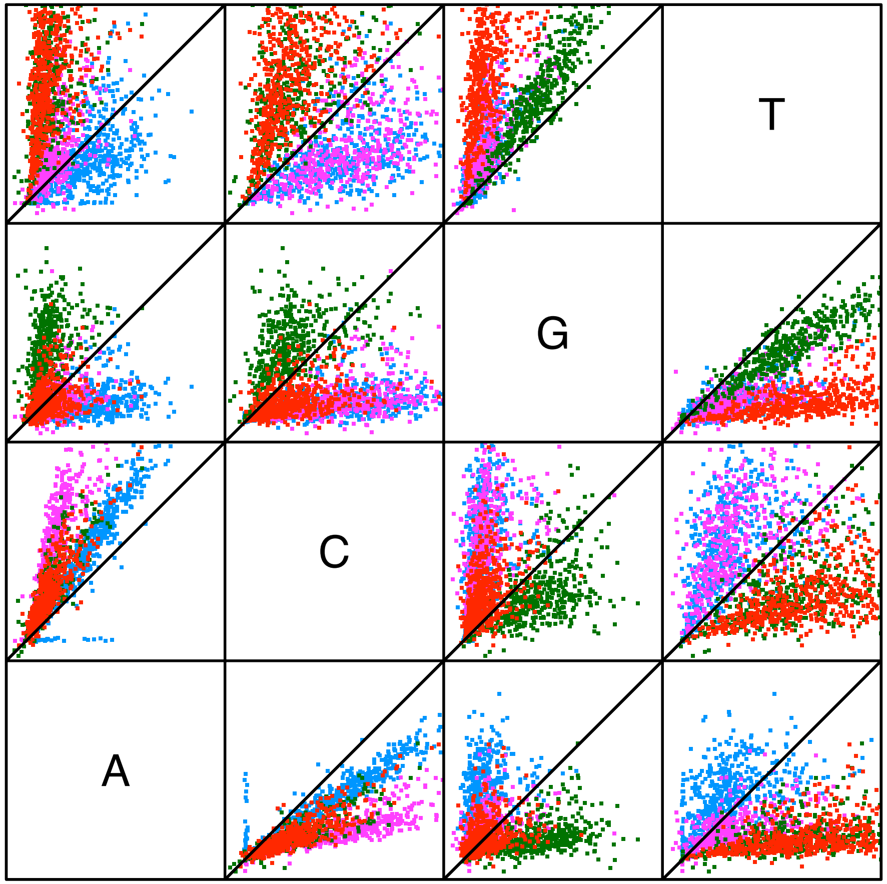


Getting Probability Profiles

Fluorescence Intensity



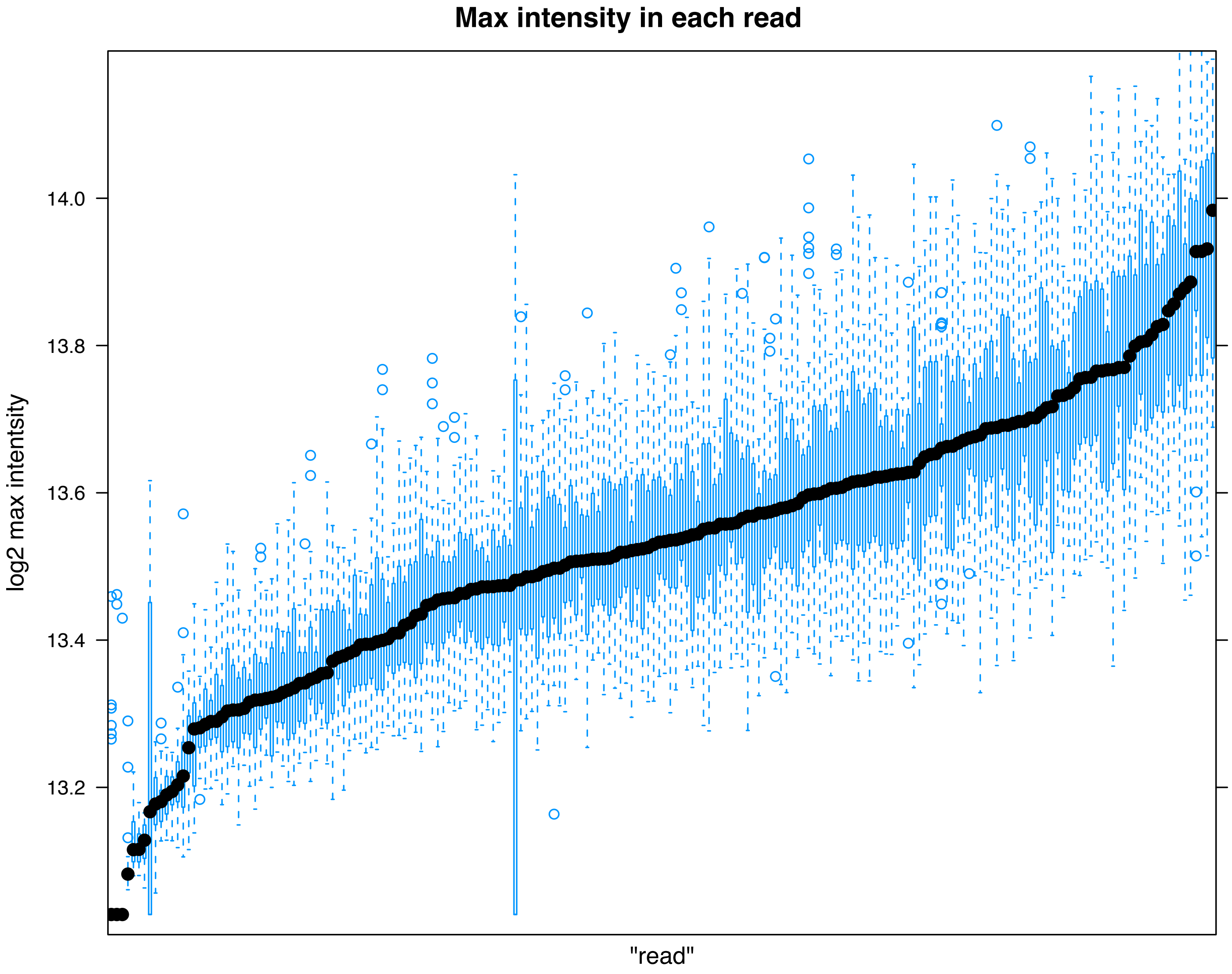
Four–channel fluorescence intensity, cycle 1



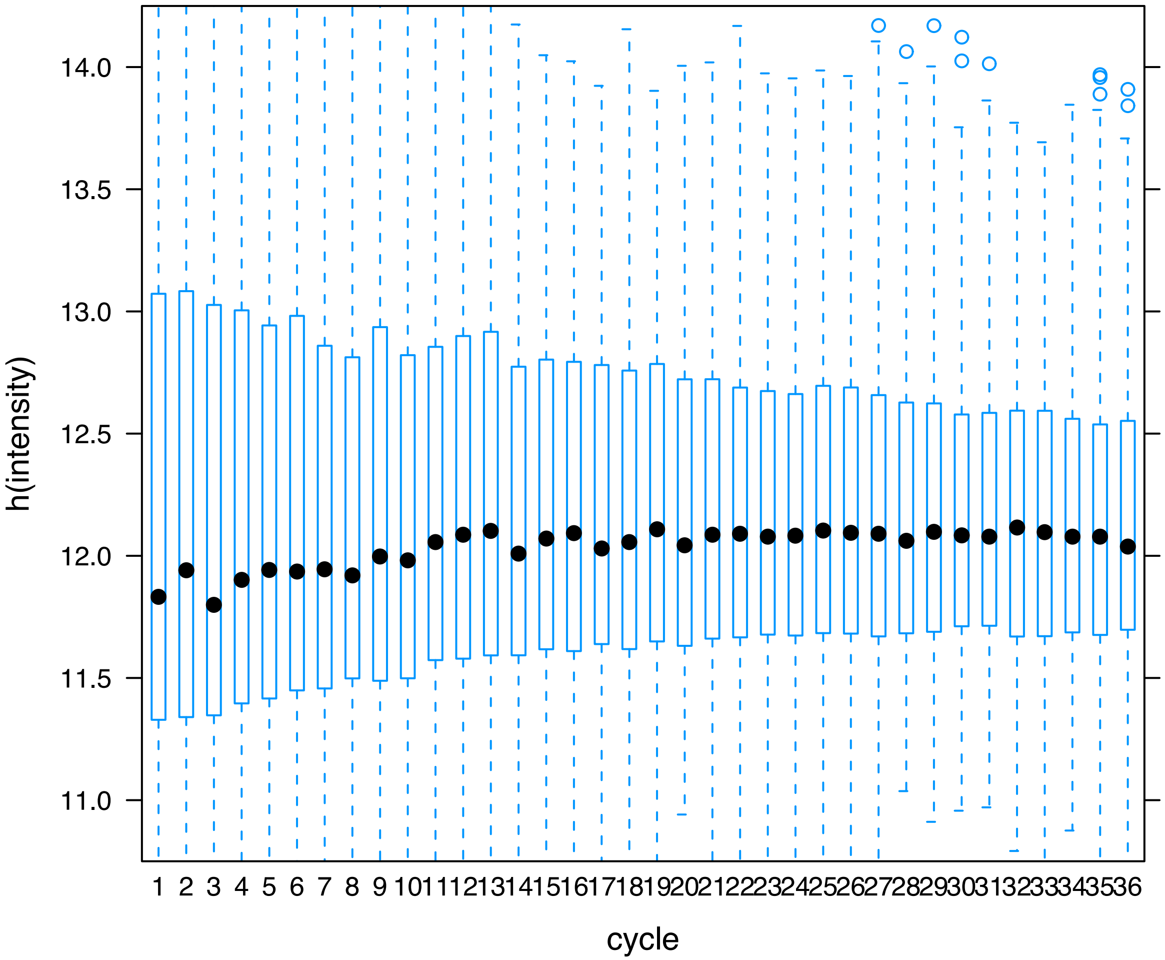
Four–channel fluorescence intensity, cycle 25

Color coded by call made: A, C, G, T

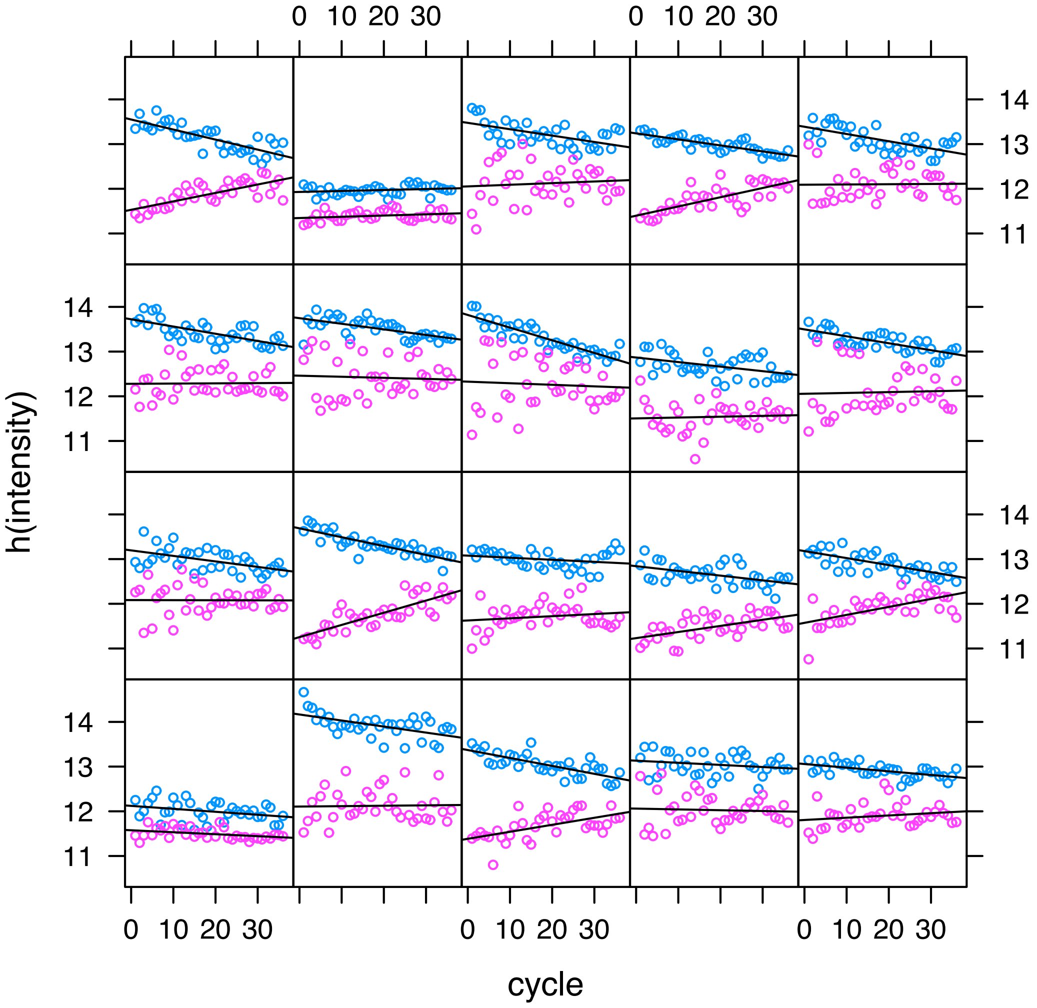
The Read Effect



The Cycle Effect



Read & Cycle Effects



Intensity Model

- We use the following model for read *i,* cycle *j*:

$$h(y_{ij}) = Mu_{ij}$$

  - started log transform: $h(y_{ij})$

Intensity Model

- We use the following model for read *i*, cycle *j*:

$$h(y_{ij}) = Mu_{ij}$$

  - started log transform: $h(y_{ij})$
  - cross-talk matrix

$$M = \begin{bmatrix} 1 & m_{AC} & m_{AG} & m_{AT} \\ m_{CA} & 1 & m_{CG} & m_{CT} \\ m_{GA} & m_{GC} & 1 & m_{GT} \\ m_{TA} & m_{TC} & m_{TG} & 1 \end{bmatrix}$$

Intensity Model

- We use the following model for read *i*, cycle *j*:

$$h(y_{ij}) = M u_{ij}$$

  - actual log intensity read *i*, cycle *j*, channel *c*

$$u_{ijc} = \Delta_{ijc}(x_j^T \alpha_i + \epsilon_{ijc}^{\alpha}) + (1 - \Delta_{ijc})(x_j^T \beta_i + \epsilon_{ijc}^{\beta})$$

Intensity Model

- We use the following model for read *i*, cycle *j*:

$$h(y_{ij}) = Mu_{ij}$$

  - actual log intensity read *i*, cycle *j*, channel *c*

$$u_{ijc} = \Delta_{ijc}(\underline{x_j^T \alpha_i + \epsilon_{ijc}^{\alpha}}) + (1 - \Delta_{ijc})(\underline{x_j^T \beta_i + \epsilon_{ijc}^{\beta}})$$

  - read-specific linear models

$$\epsilon_{ijc}^{\alpha} \sim N(0, \sigma_{\alpha i}^2) \qquad \epsilon_{ijc}^{\beta} \sim N(0, \sigma_{\beta i}^2)$$

Intensity Model

- We use the following model for read *i*, cycle *j*:

$$h(y_{ij}) = M u_{ij}$$

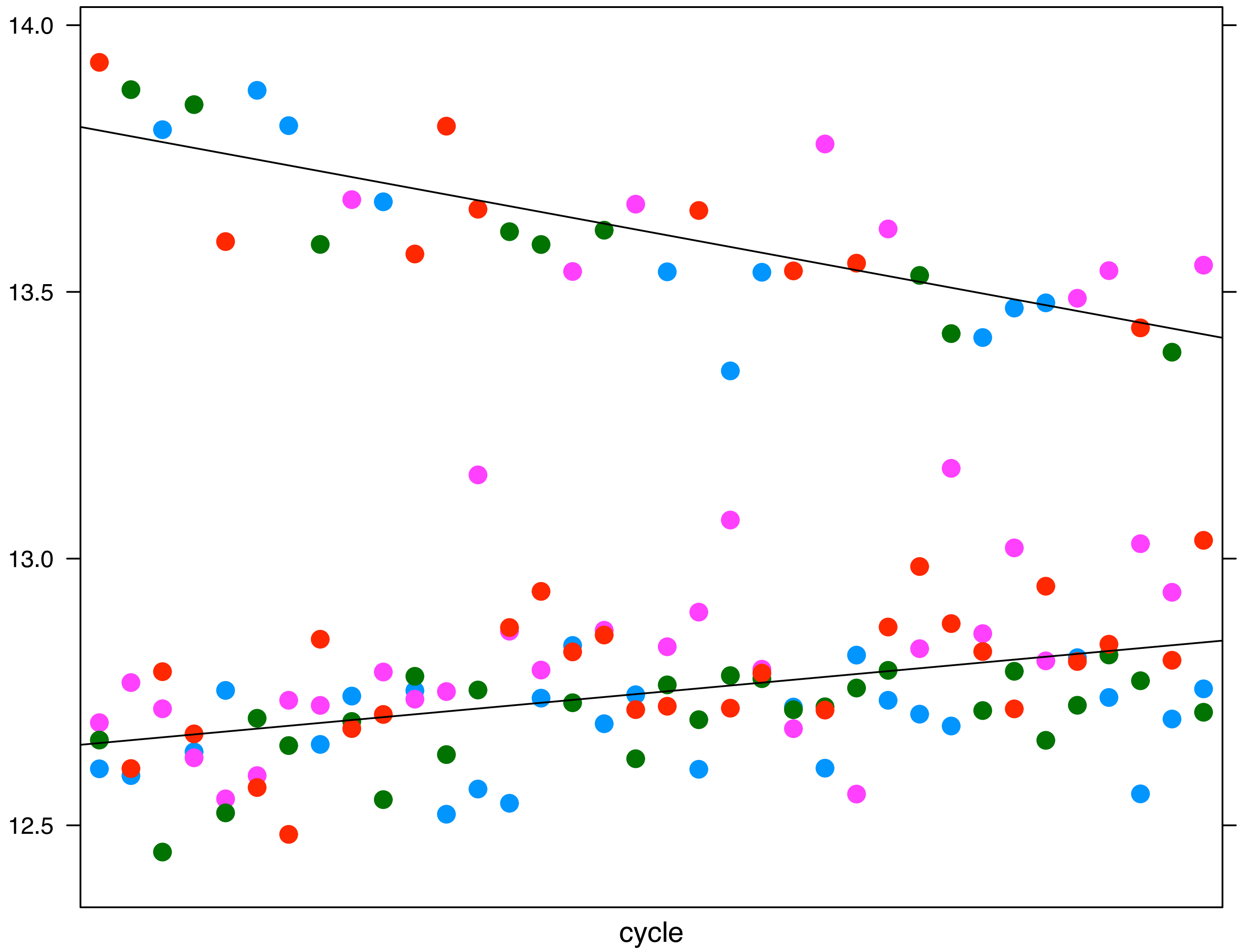
  - actual log intensity read *i*, cycle *j*, channel *c*

$$u_{ijc} = \Delta_{ijc}(x_j^T \alpha_i + \epsilon_{ijc}^{\alpha}) + (1 - \Delta_{ijc})(x_j^T \beta_i + \epsilon_{ijc}^{\beta})$$

  - indicators of nucleotide identity, read *i*, pos. *j*

$$\Delta_{ijc} = \begin{cases} 1 & \text{if } c \text{ is the nucleotide in read } i \text{ position } j \\ 0 & \text{otherwise} \end{cases}$$

Intensity Model



Intensity Model

- We use the following model for read *i*, cycle *j*:

$$h(y_{ij}) = M u_{ij}$$

  - actual log intensity read *i*, cycle *j*, channel *c*

$$u_{ijc} = \Delta_{ijc}(x_j^T \alpha_i + \epsilon_{ijc}^{\alpha}) + (1 - \Delta_{ijc})(x_j^T \beta_i + \epsilon_{ijc}^{\beta})$$

  - get Maximum Likelihood estimates with EM algorithm, also estimates

$$z_{ijc} := \mathrm{E}\{\Delta_{ijc} = 1 | u_{ij}\} = P(\Delta_{ijc} = 1 | u_{ij})$$

Intensity Model

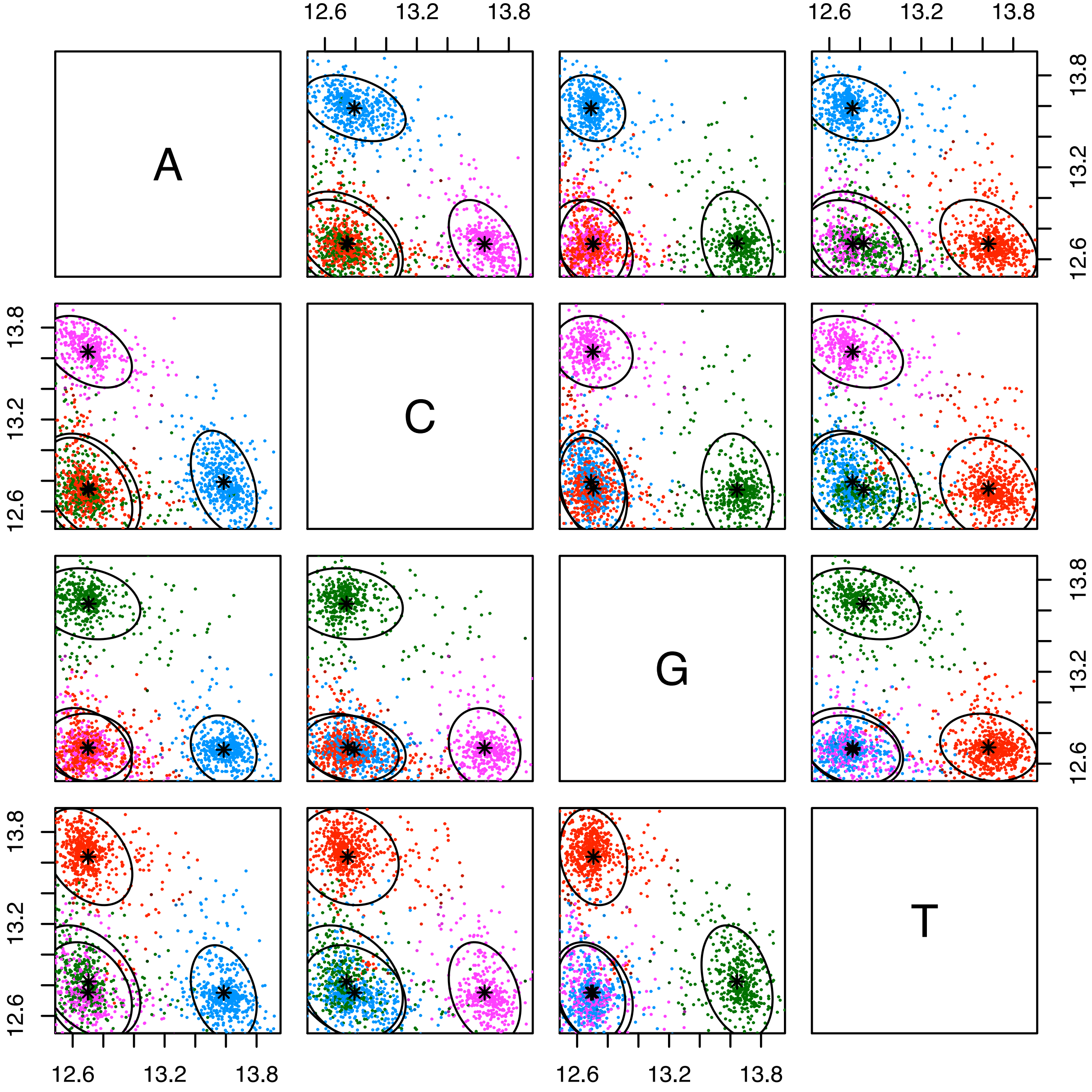
- EM-algorithm also estimates

$$z_{ijc} := \mathrm{E}\{\Delta_{ijc} = 1 | u_{ij}\} = P(\Delta_{ijc} = 1 | u_{ij})$$

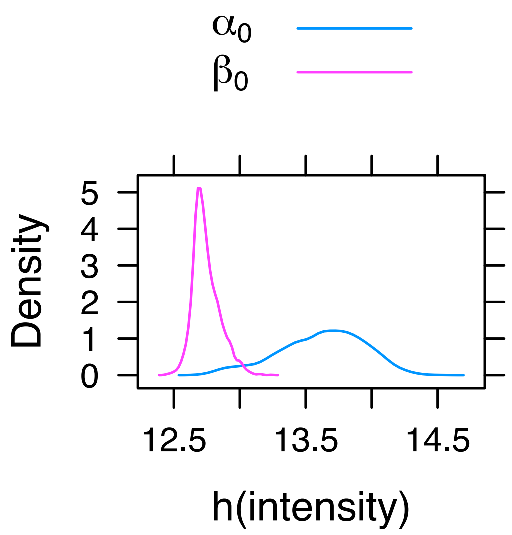
Intensity Model

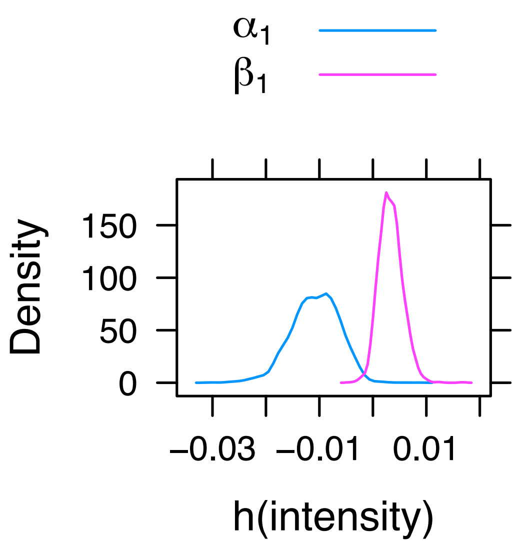
- After removing effects, we use a standard normal mixture clustering model
- Initizalized by probability profiles estimated by effects model (z_{ijc})
- Clustering refines probability profiles from effects model by drawing from other reads and cycles

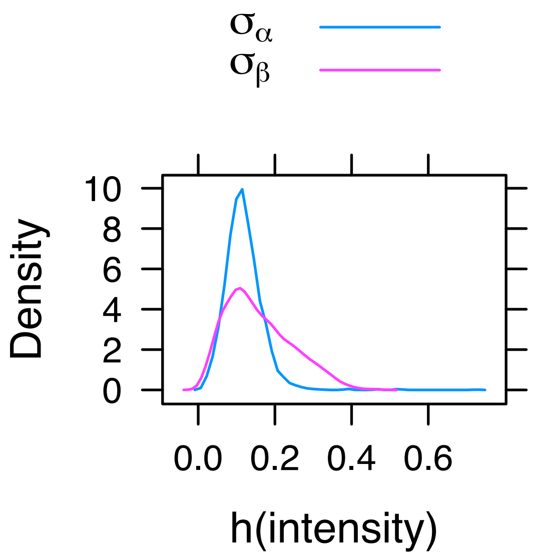
Intensity Model



Model Estimates







Quality Metrics

1. Entropy: Certainty according to probability profiles in each read position

Quality Metrics

1. Entropy: Certainty according to probability profiles in each read position

$$H_{ij} = -\sum_{c} z_{ijc} \log z_{ijc}$$

Quality Metrics

1. Entropy: Certainty according to probability profiles in each read position

$$H_{ij} = -\sum_{c} z_{ijc} \log z_{ijc}$$

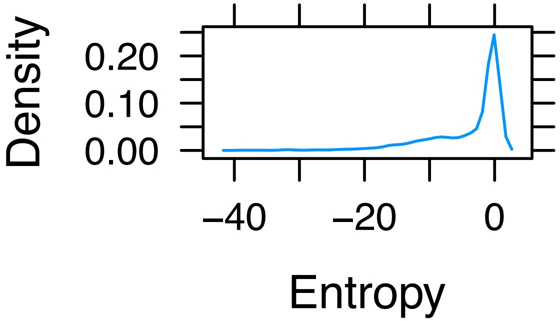
$$H_{i} = -\sum_{jc} z_{ijc} \log z_{ijc}$$

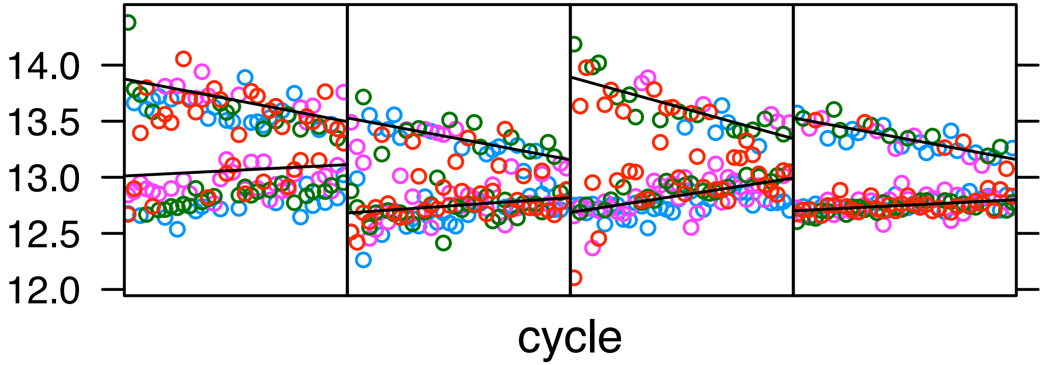
Quality Metrics

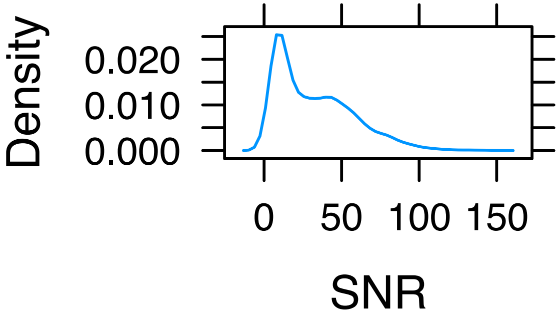
1. Entropy: Certainty according to probability profiles in each read position
2. SNR: How easy is it to distinguish signal and noise linear models?

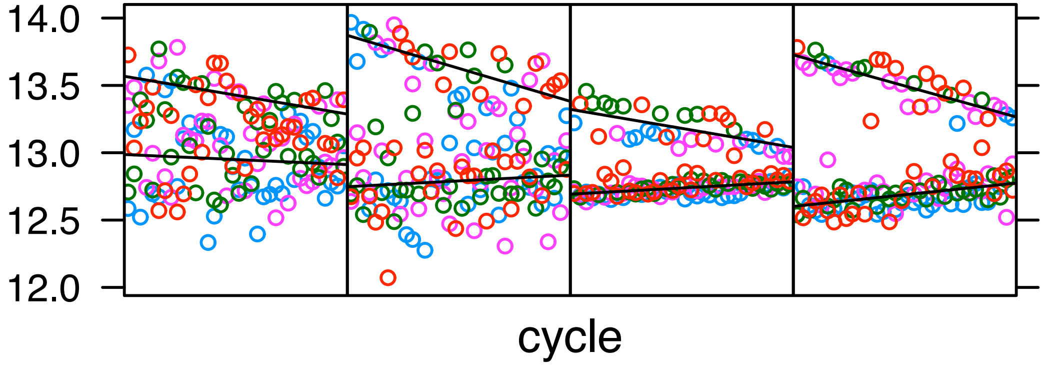
$$SNR_i = \frac{1/N\|X(\alpha_i - \beta_i)\|_2^2}{1/2(\sigma_{\alpha i}^2 + \sigma_{\beta i}^2)}$$

Quality Metrics

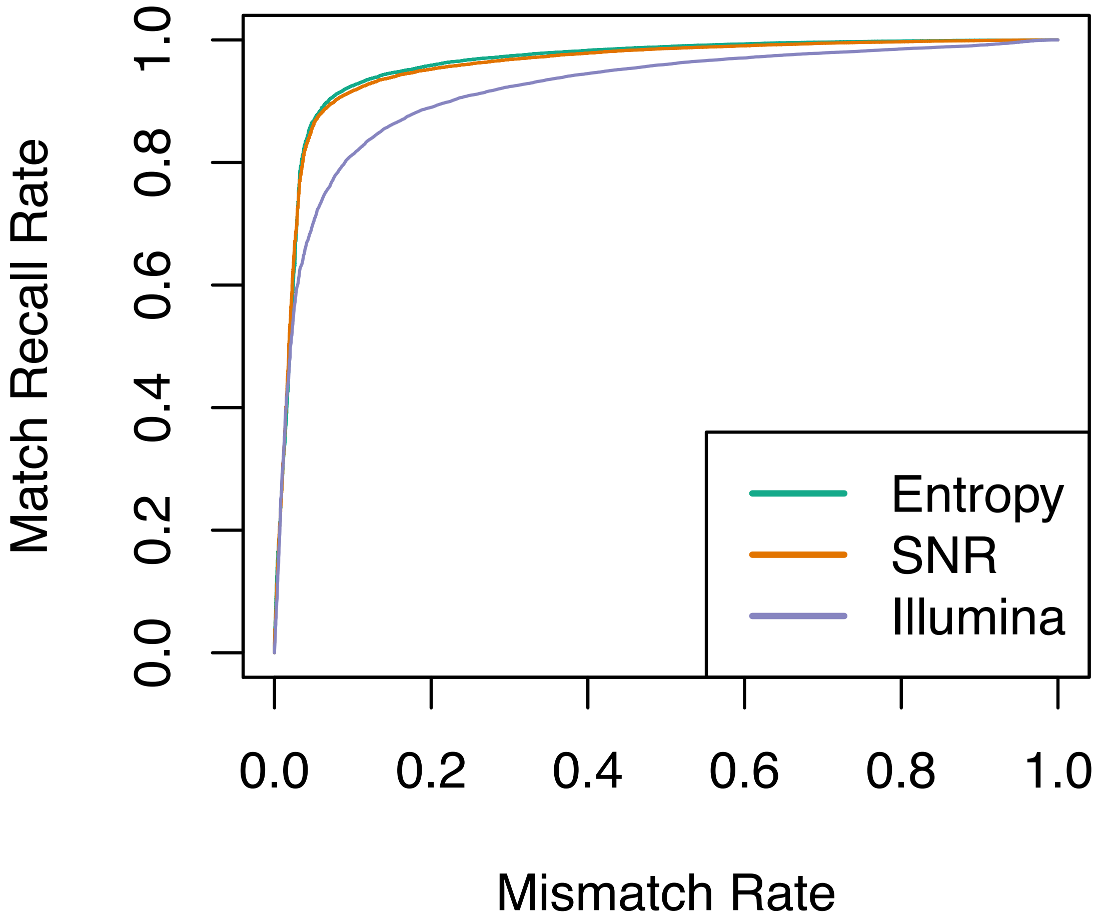








Quality Metrics



Genotyping

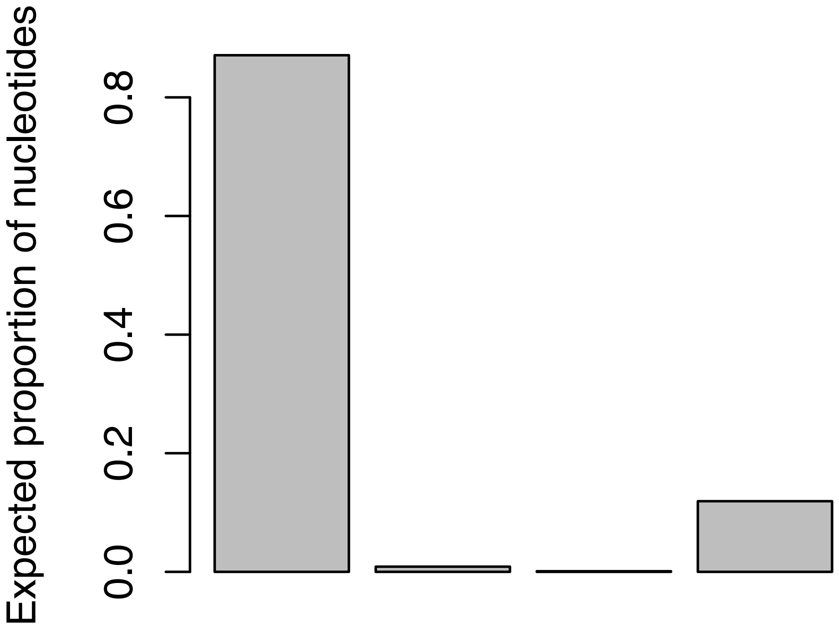
- A very simple solution: get expected proportion of nucleotides at each position

Genotyping

- Use expected proportion of each nucleotide at genomic position

$$T_{jc} = \sum_i z_{ijc}$$

Genotyping



Computational Challenges

- Efficient model estimation (robust estimates of effects use linear programming, fast clustering)
- Parallel computation
- Storage & retrieval
- Matching

Conclusion

- Described model-based solution to handle uncertainty inherent in sec-gen data analysis
- Particularily important for genotyping
- Now the fun starts...

Thanks!