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

ngs2009 Barcelona Oct. 3 2009

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
```

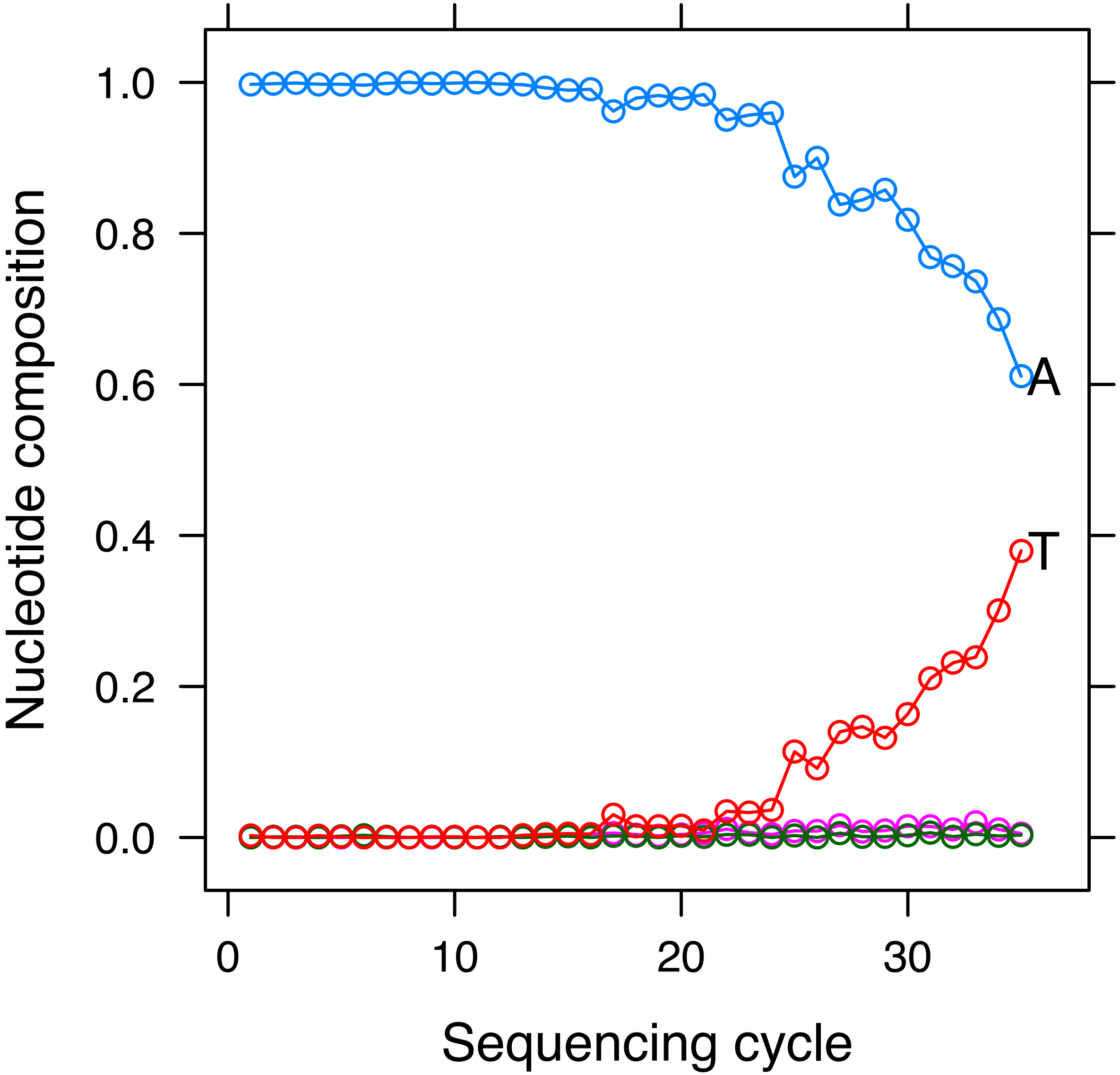
SNPs

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

SNPs

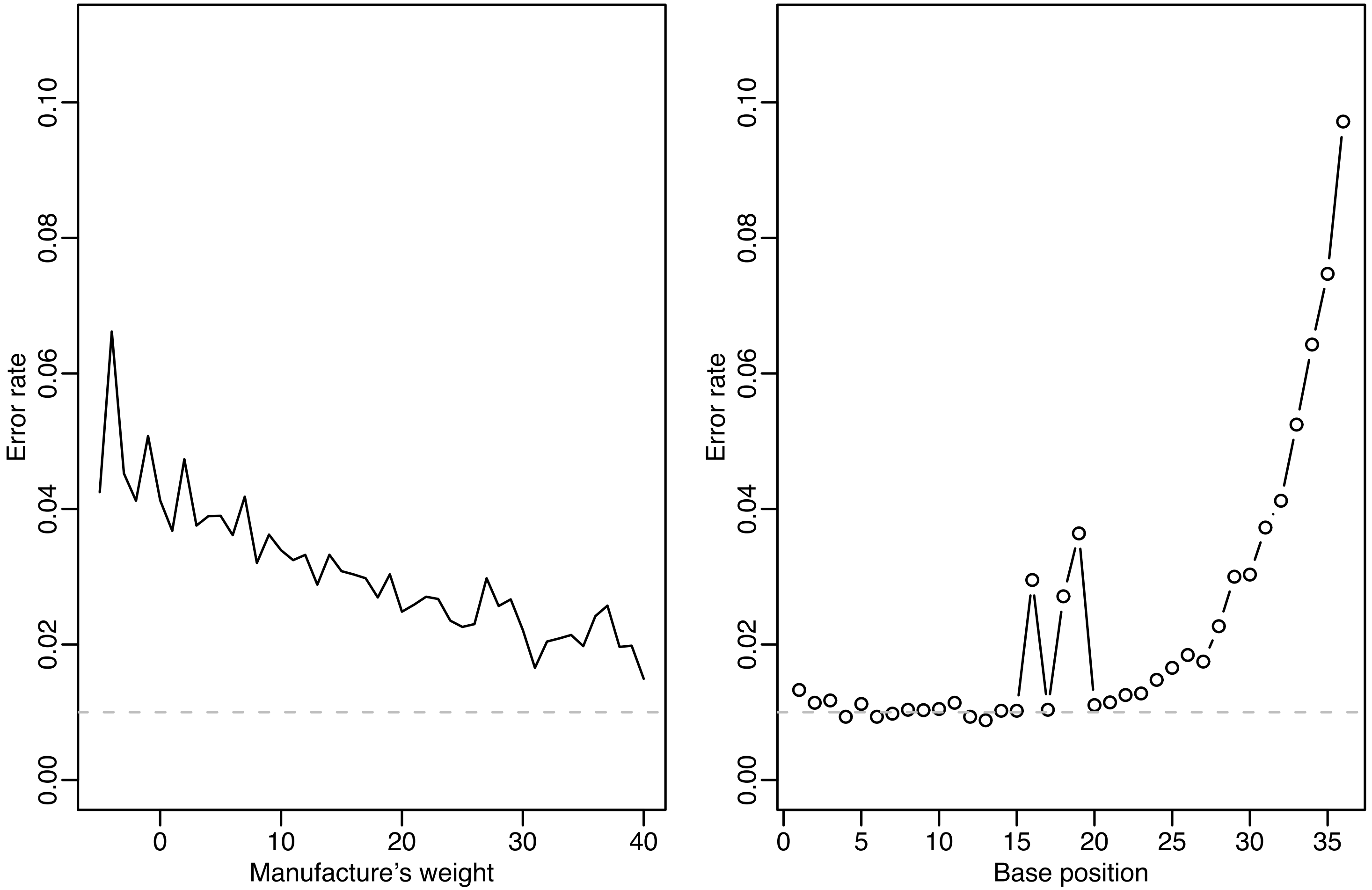
```
 TCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTA
   TCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTTTG
     GTACTCGTCGCTGCGTTGAGGCTTGCGTTTTTGGT
      TGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTA
       GCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTAC
          CGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCT
                 GCGTTGAGGCTTGCGTTTATGGTACGCTGGATTTT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

SNPs

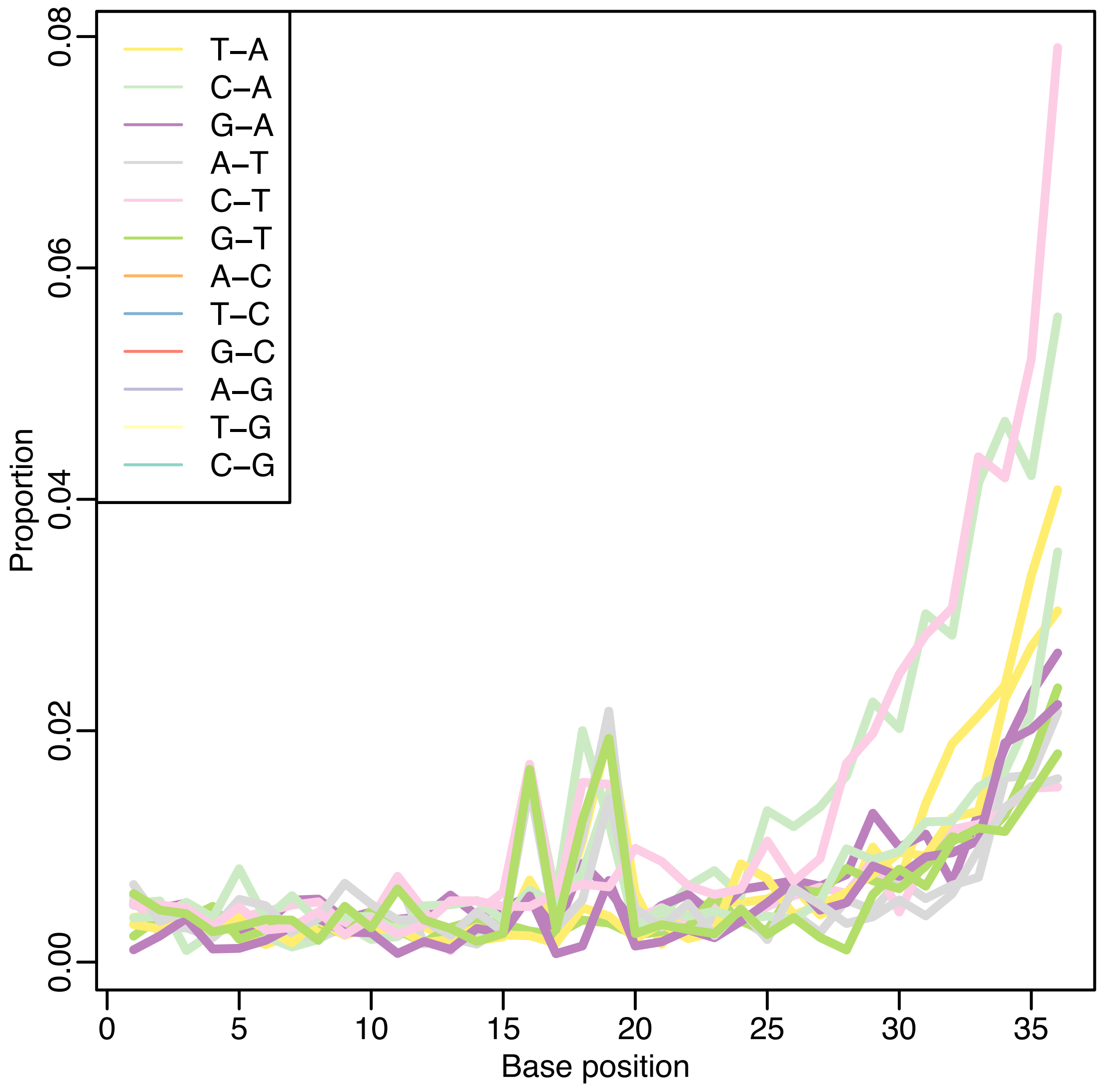


Nucleotide composition
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
Sequencing cycle
A
T

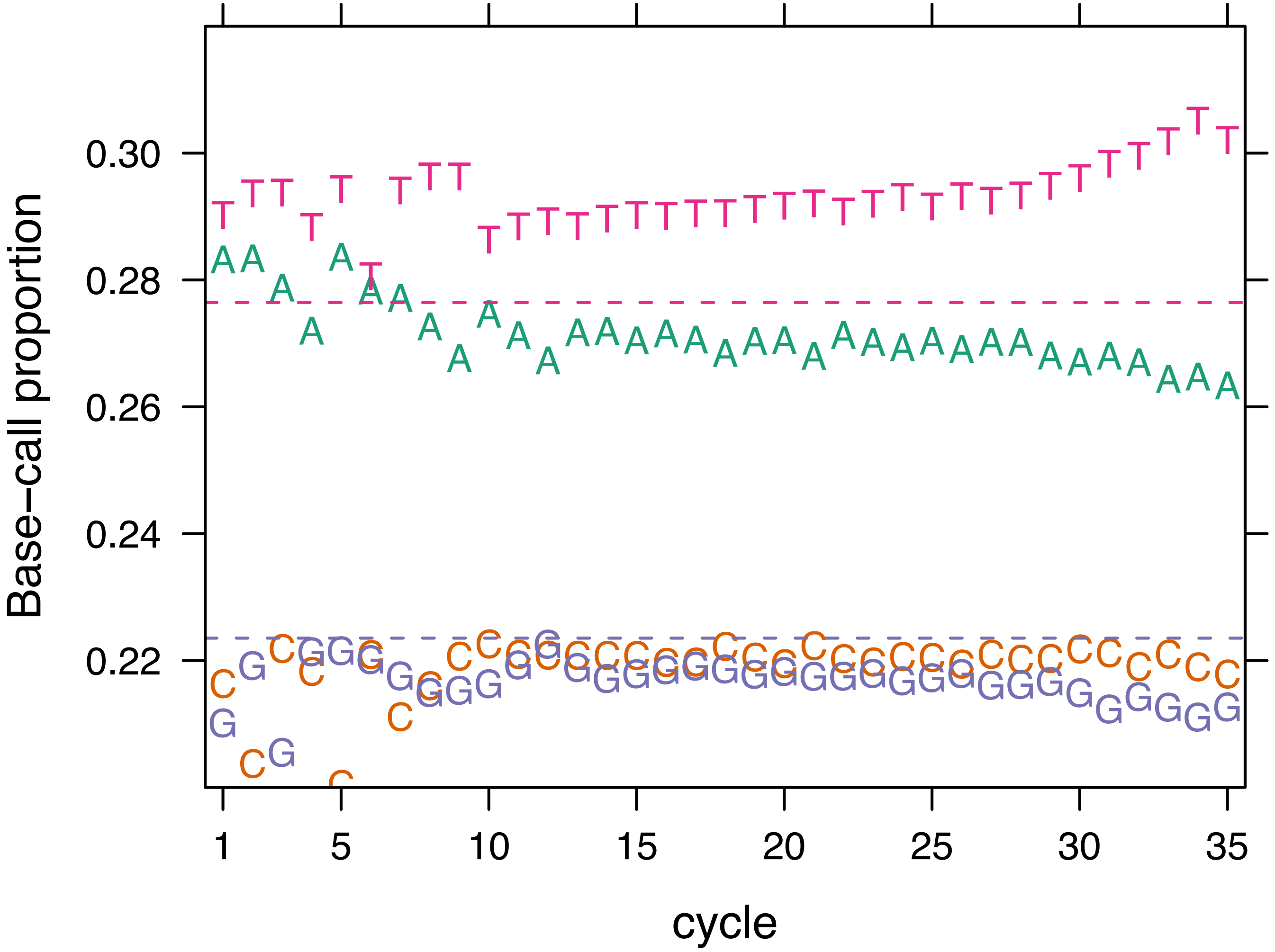
Error Rate and Reported Quality



Systematic Biases



Systematic Biases

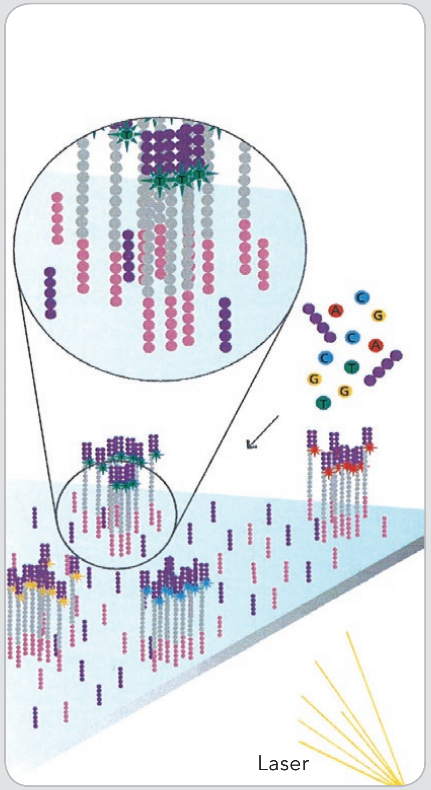


Outline

1. Not all base-calls are equal!
2. Model-based base-calling
3. Model-based quality assessment
4. Results in genotyping pooled samples application

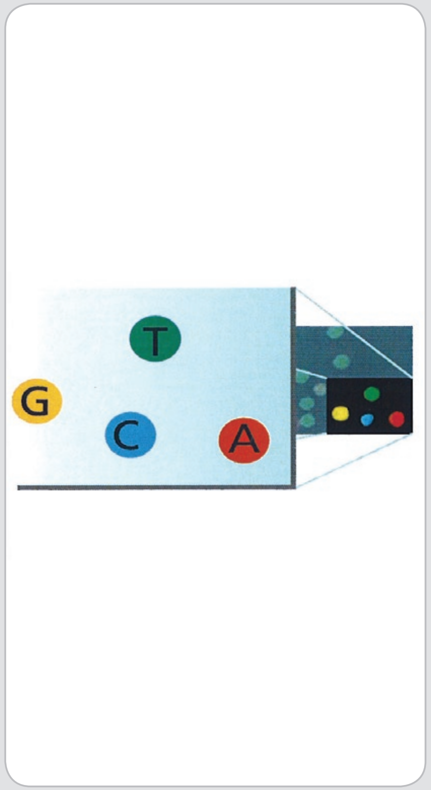
Illumina/Solexa

7. DETERMINE FIRST BASE



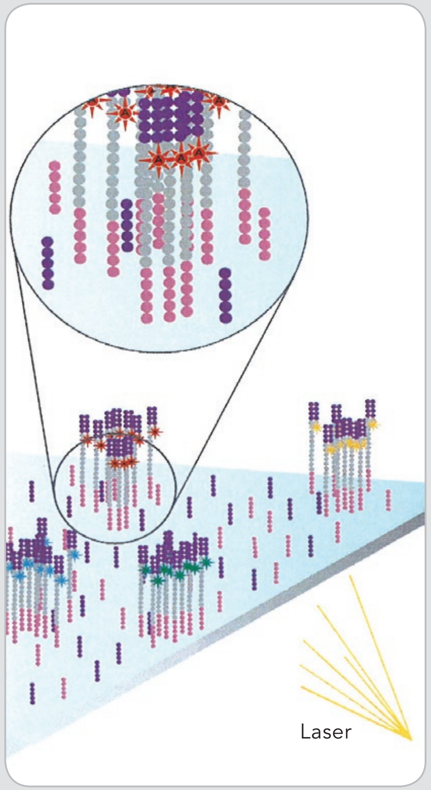
The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

8. IMAGE FIRST BASE



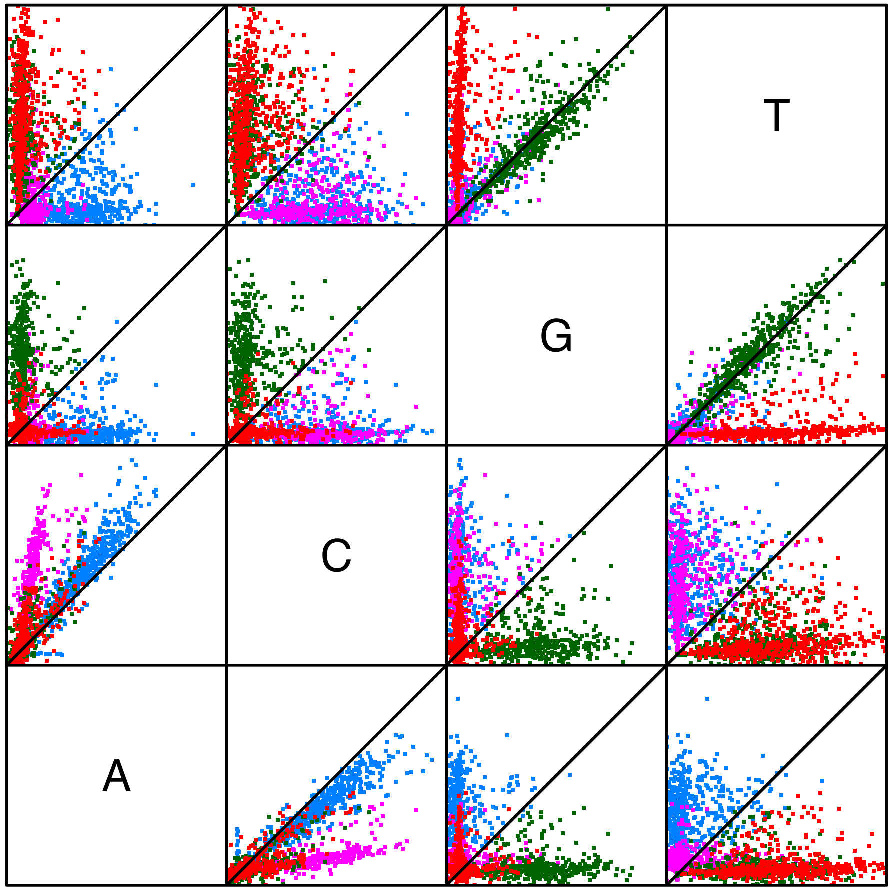
After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

9. DETERMINE SECOND BASE

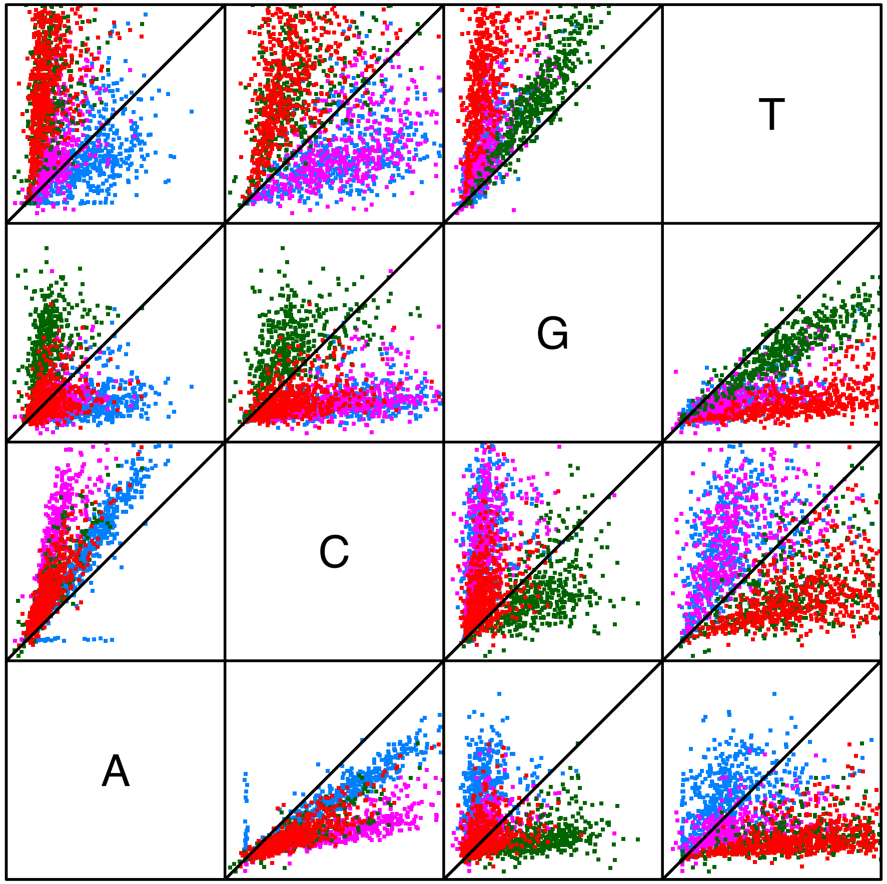


The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

Fluorescence Intensity



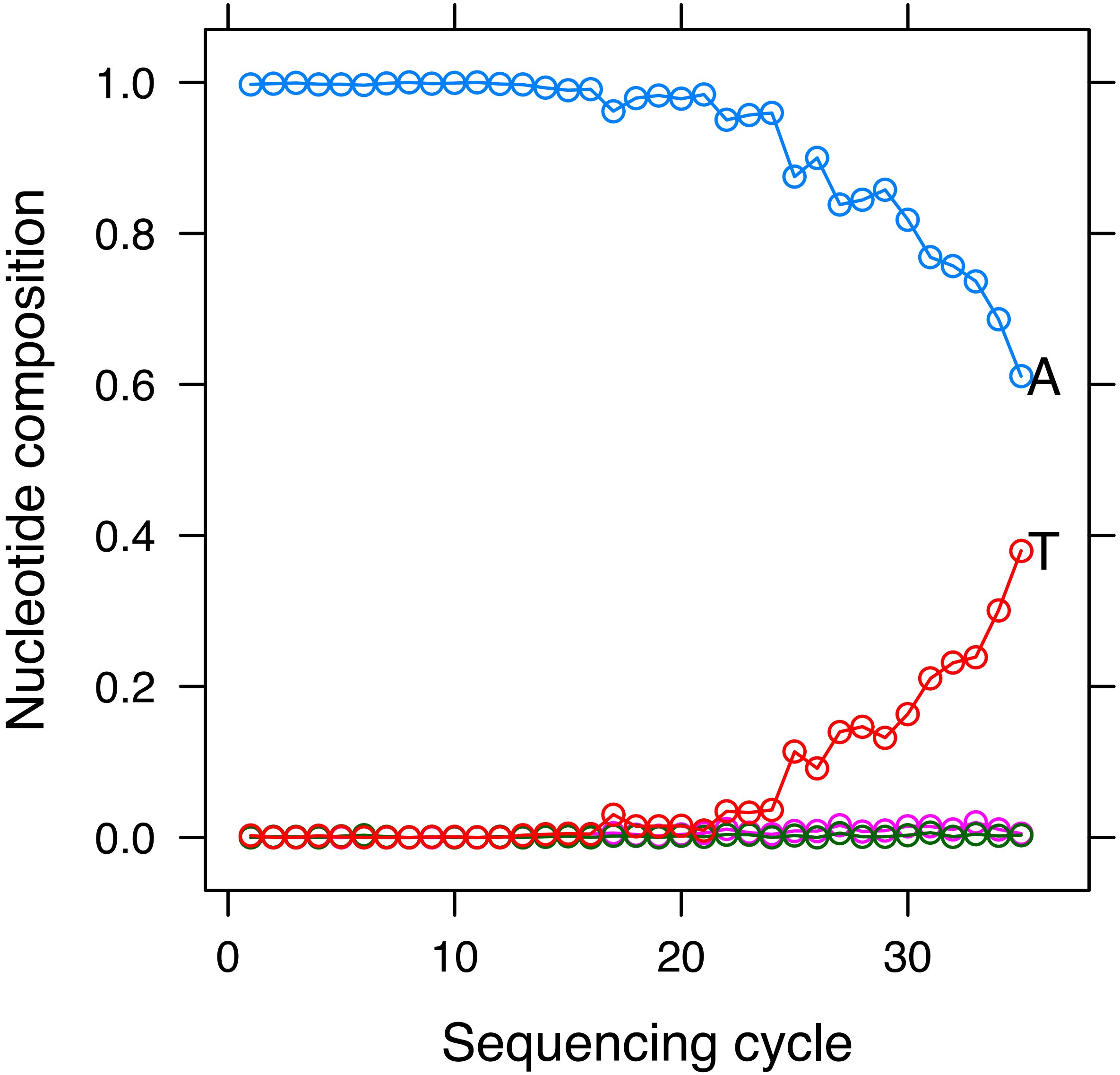
Four–channel fluorescence intensity, cycle 1



Four–channel fluorescence intensity, cycle 25

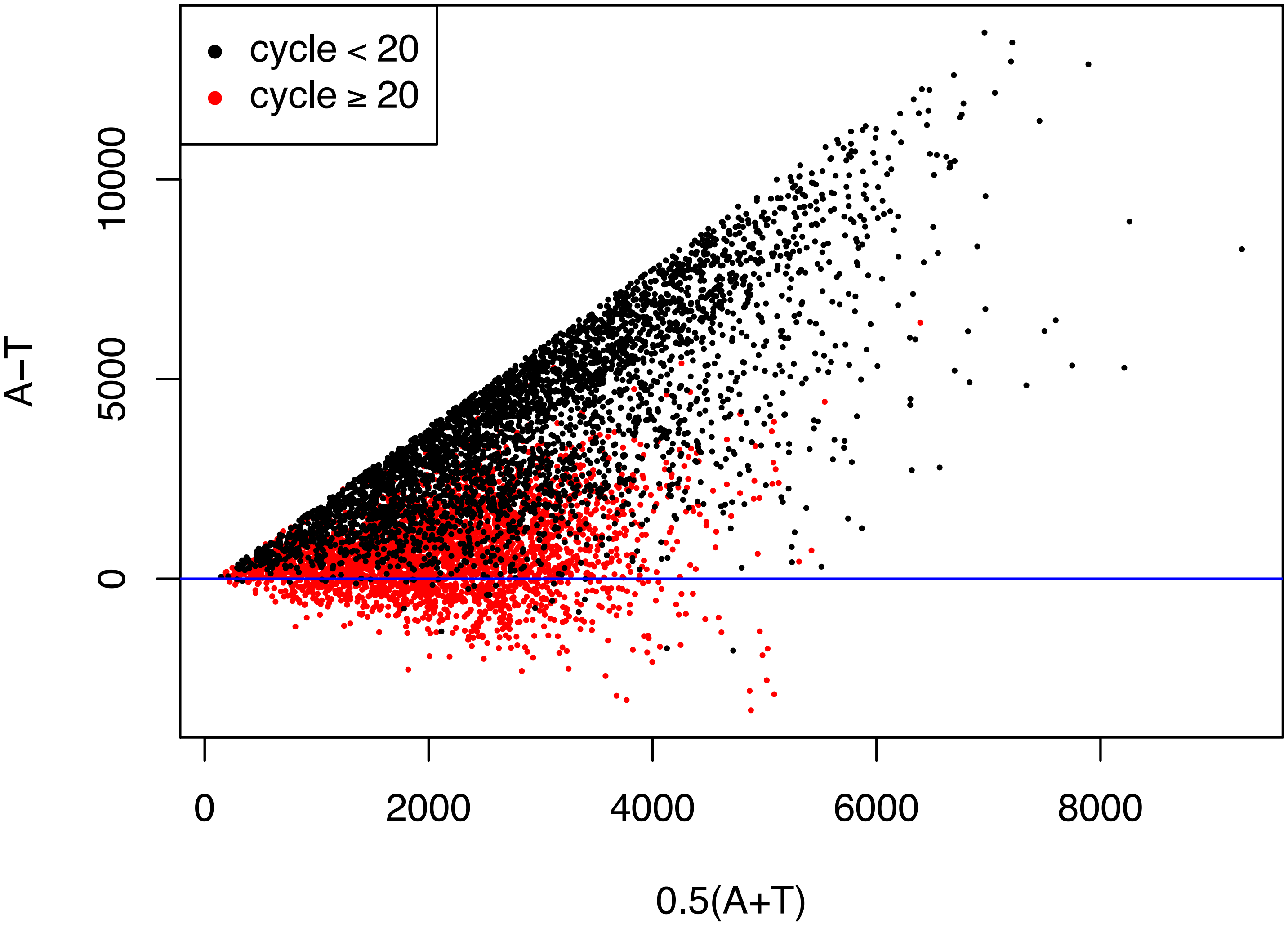
Color coded by call made: A, C, G, T

SNPs



Nucleotide composition
1.0
0.8
0.6
0.4
0.2
0.0
A
T
0
10
20
30
Sequencing cycle

SNP Intensities

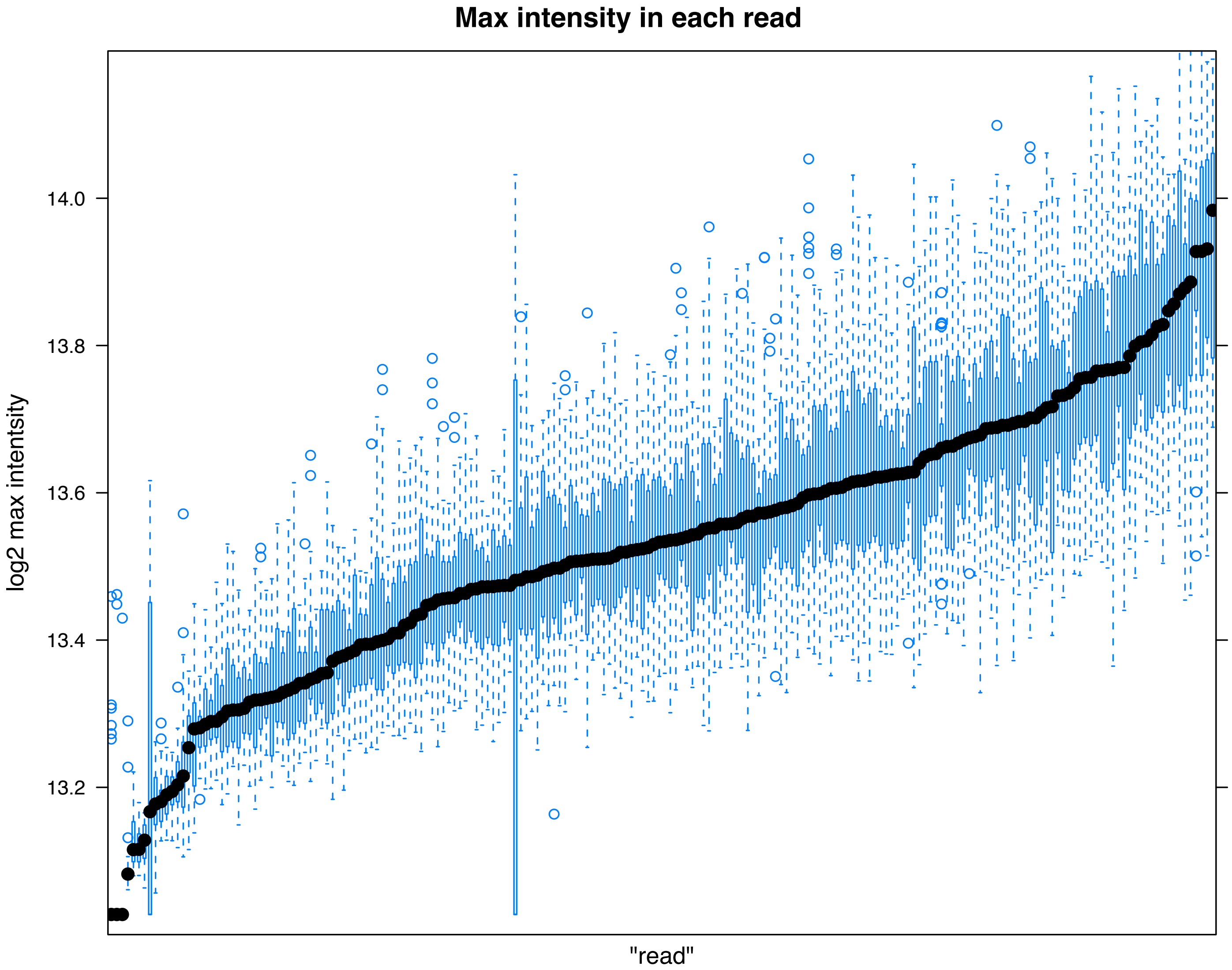


Challenges

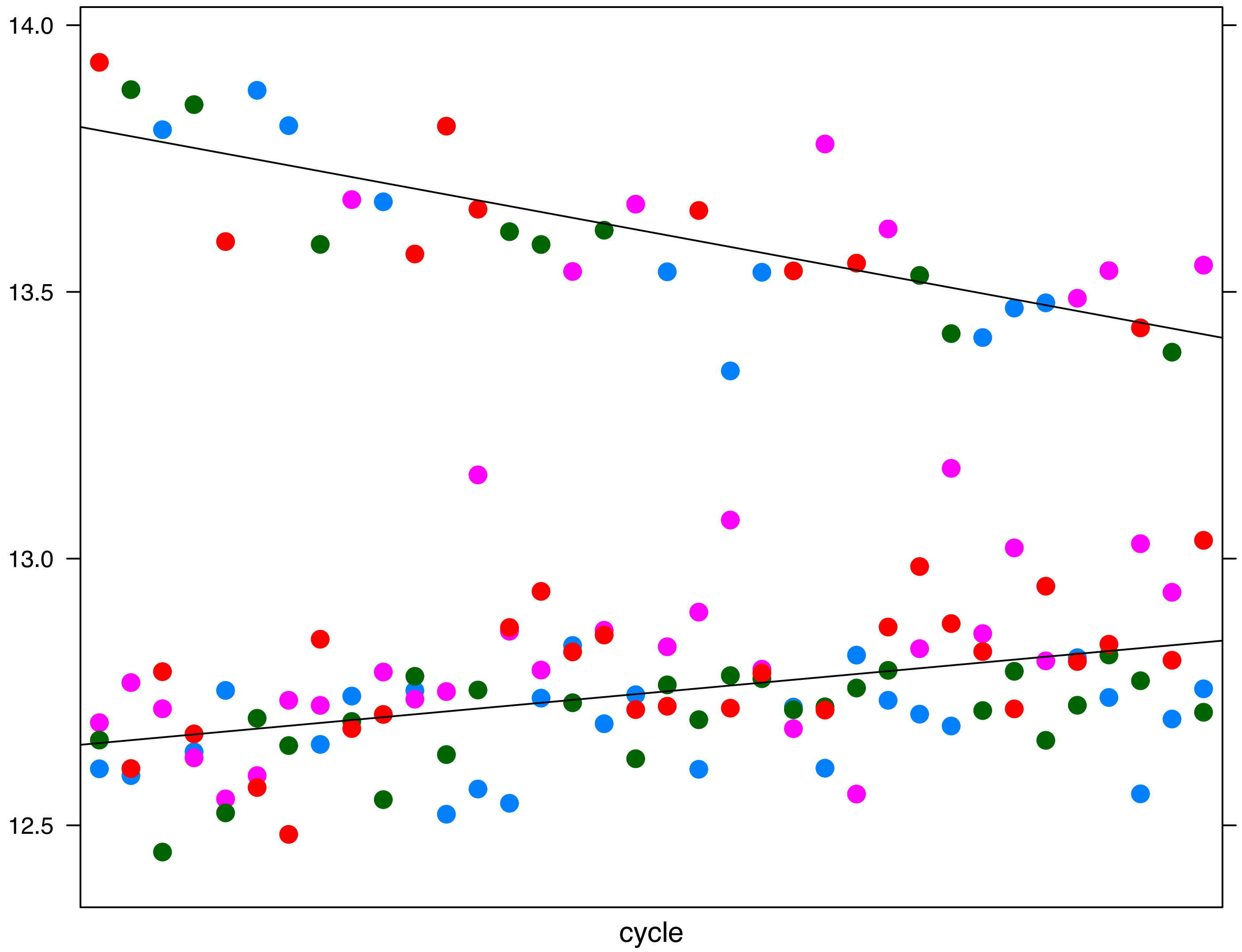
- Base-calling is the result of a complicated procedure on noisy data
- Not all base-calls are made with the same certainty
- Statistical: What is the proper way of modeling this (un)certainty?
- Computational: Can we use this model at sec-gen data scale?

[Corrada Bravo, Irizarry. To appear, *Biometrics,* 2009]

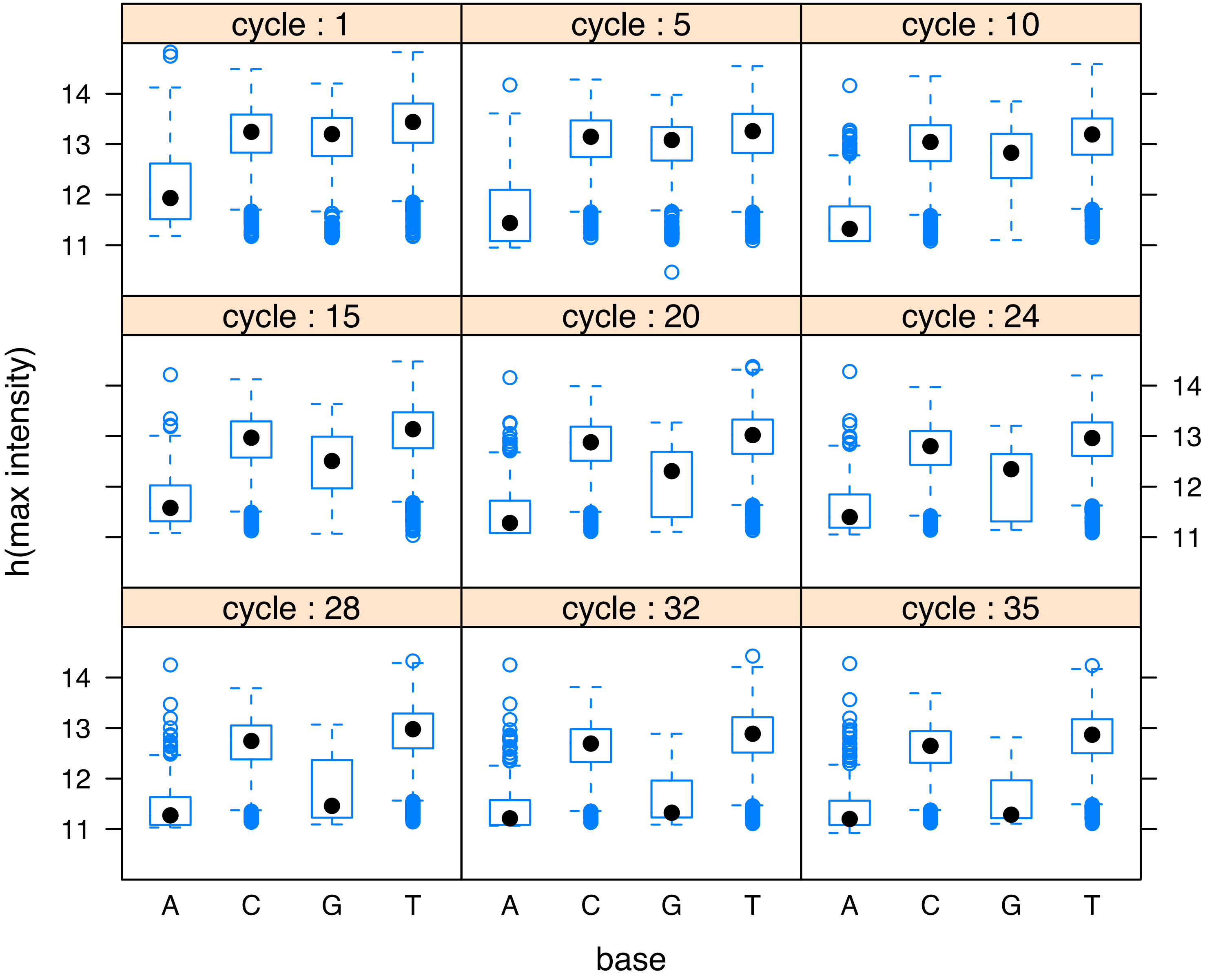
The Read Effect



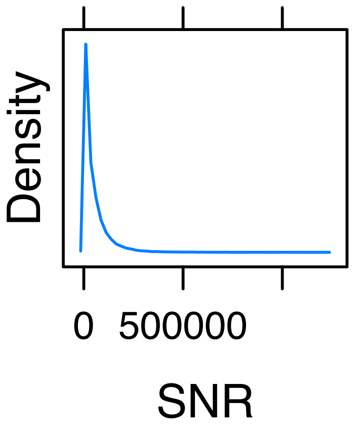
Intensity Model

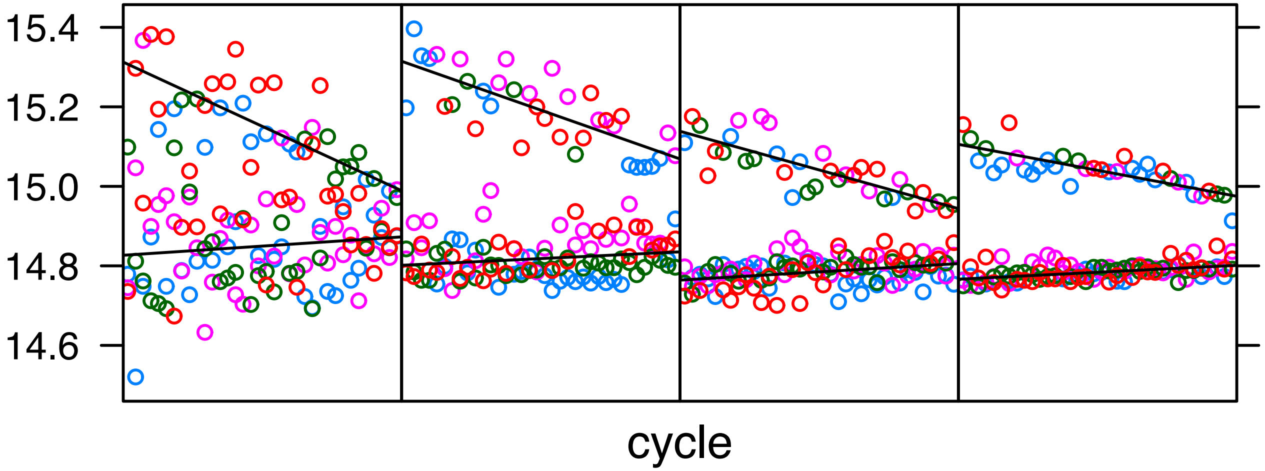


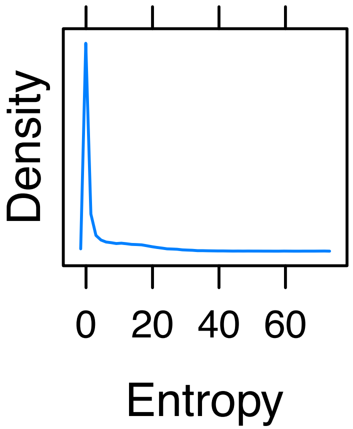
Base-Cycle Effect

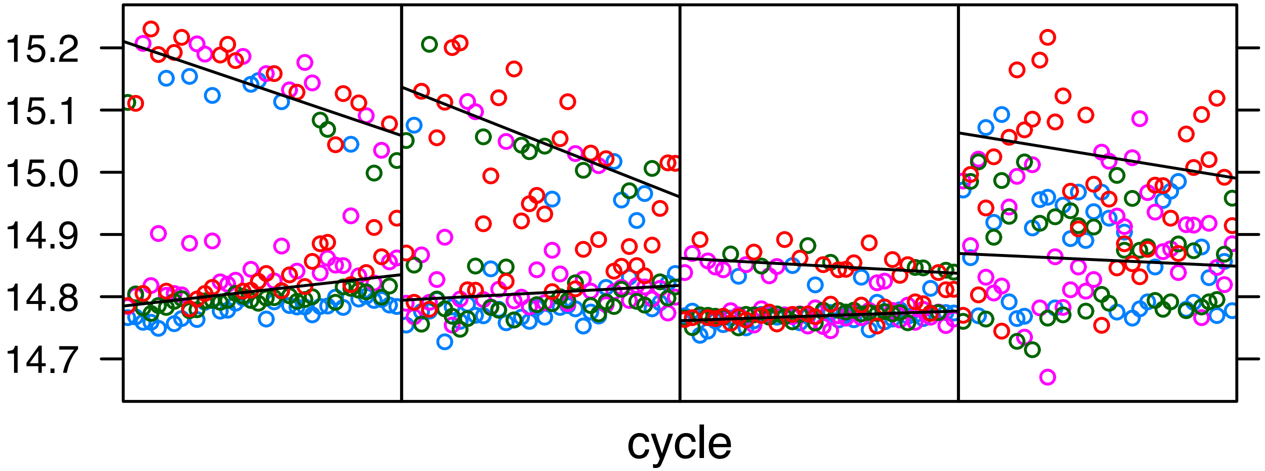


Quality Metrics









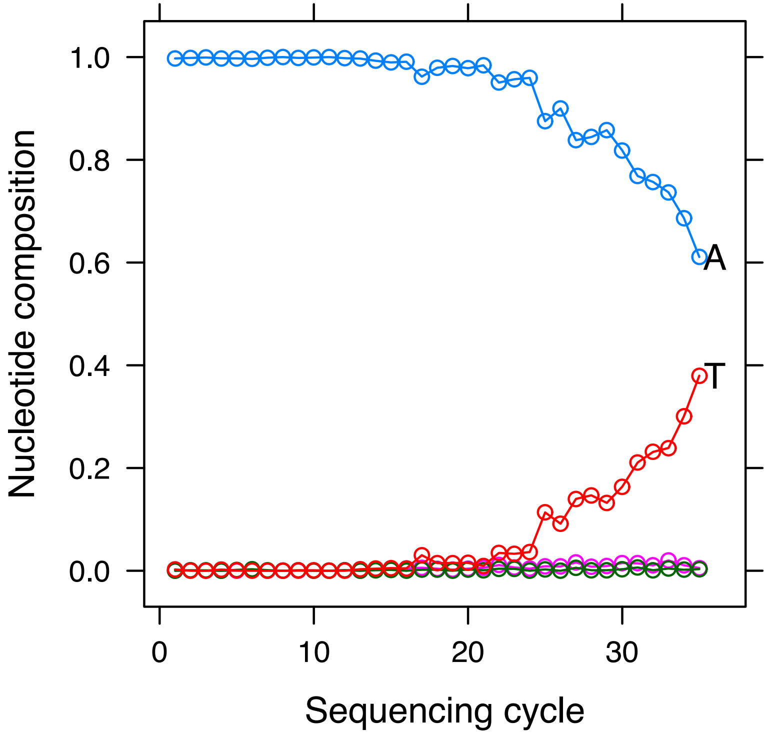
Yield & Accuracy

	Bustard	Seraphim	%increase
Total mapped reads	5,096,667	5,686,797	11.5
0 mismatch	4,332,125	4,645,492	7.2
1 mismatch	514,635	688,880	33.8
2 mismatch	141,421	235,035	66.2

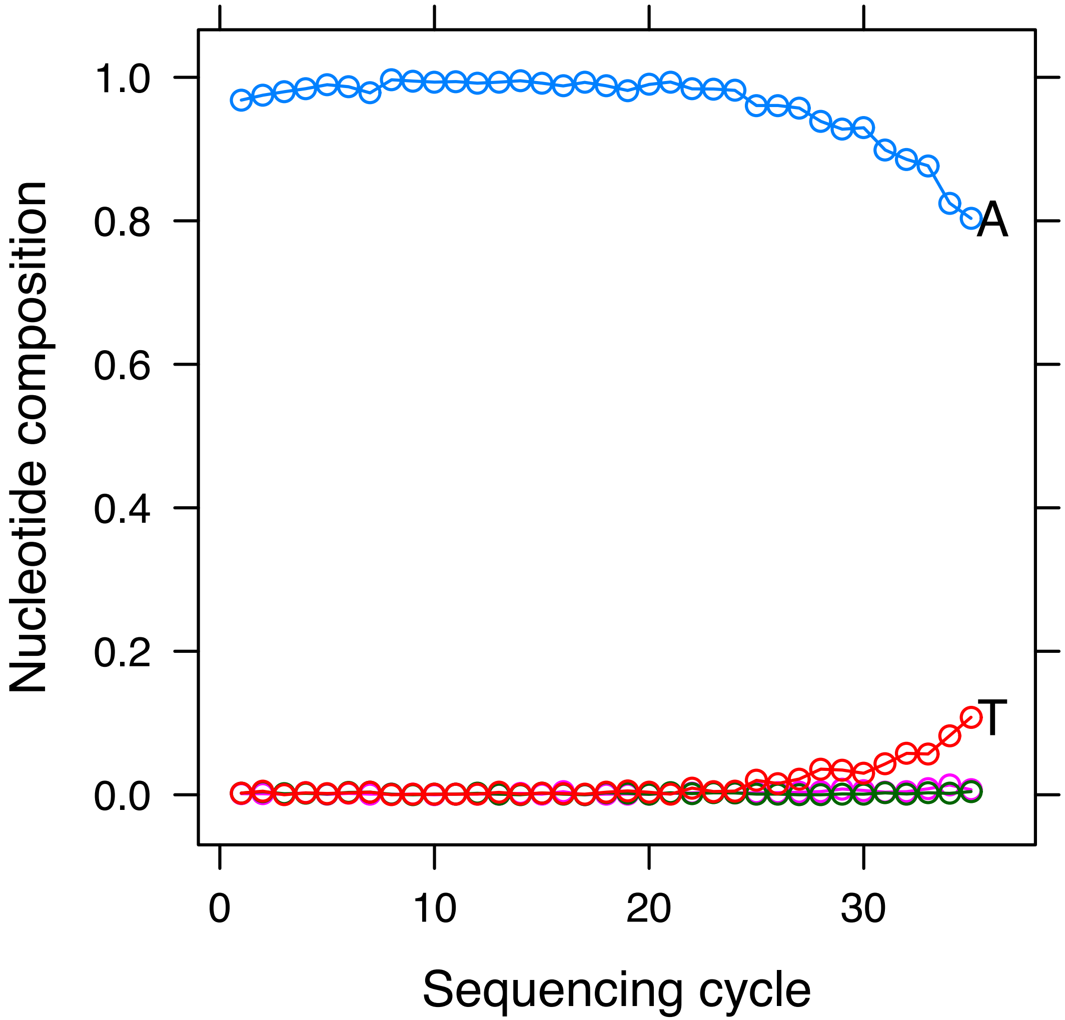
SNPs

- Running MAQ pipeline, number of high quality SNPs (MAQ quality greater than 100)
  - Solexa: 37
  - Seraphim (us): 10
- 70% fewer false positives
  - some of the remaining look real!

SNPs



Nucleotide composition
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
Sequencing cycle
A
T



Nucleotide composition
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
Sequencing cycle
A
T

Genotyping Pooled Samples

- Pilot study for variant discovery in pooled samples
- Targeted sequencing of ~20 exons in GRIP2
- Multiplexed reads (12 multiplex pools), 40 patients per pool (!!)

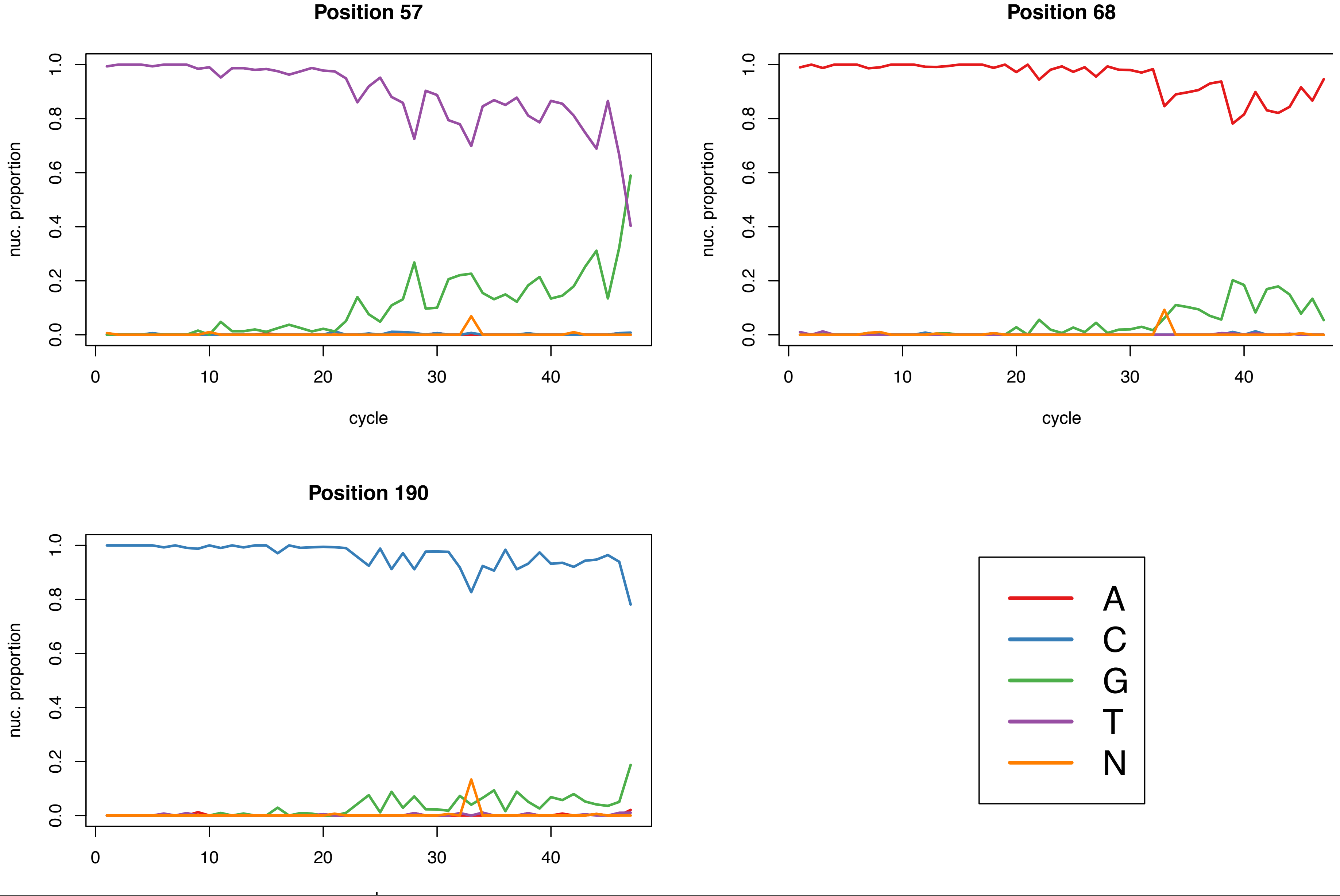
Pilot Study Analysis

1. One lane of Illumina GAII
2. Primary analysis by 1.3 Pipeline
3. Matched to GRIP2 exons with Bowtie

   Average coverage ~15x per allele

4. Pooled SNP calling by MAQ (quality over 185)

Pilot Study (exon 1)



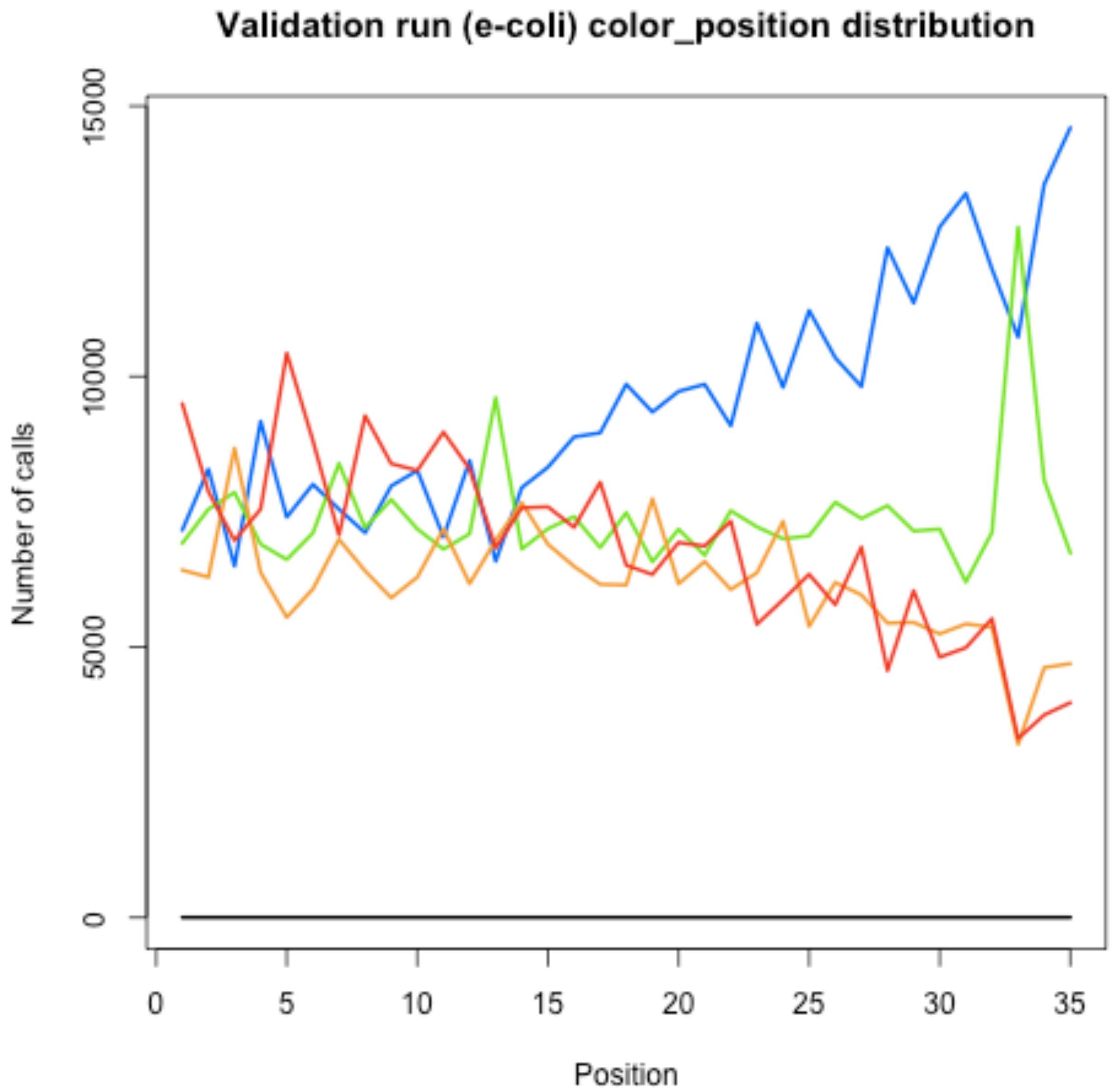
Pilot Study Result

- **201** SNPs called (MAQ quality over 185)
  - includes 19/20 known variants for these GRIP2 exons
- With our base-calls and log-entropy quality:
  - 5% increase in total matches
  - **80** SNPs called by MAQ
    - includes 18/20 known variants
- Verification: Under way

More to come...

- Matching w/ probability profiles
- Genotyping from matched probability profiles
- Extension to SOLiD platform

SOLiD


Validation run (e-coli) color_position distribution
Number of calls
0
5000
10000
15000
Position
0
5
10
15
20
25
30
35

Conclusion

- Described model-based solution to handle uncertainty inherent in sec-gen data analysis
- Particularly important for genotyping
- Improved base-calling performance with interpretable model parameters (QA)

Acknowledgements

JHU Biostatistics

Hao Wu

Harris Jaffee

Ben Langmead

JHU Medical Institutions

Sarah Wheelan

Sarven Sabunciyan

JHU Institute of Genetic Medicine

Tao Wang

Abby Adamczyk

Tejas Niranjani

Positions available: postdocs and programmers

genomics@jhu.edu

www.genomics.jhu.edu