

# STATISTICAL METHODS FOR NEXT GENERATION SEQUENCING

<http://www.biostat.jhsph.edu/~khansen/enar2012.html>

Zhijin Wu  
Brown University

Kasper Hansen, Rafael A Irizarry  
Johns Hopkins University

# OUTLINE

---

- Introduction to NGS
- SNP calling and genotyping
- RNA-sequencing
- Hands-on exercise

# INTRODUCTION TO NEXT GENERATION SEQUENCING

RAFAEL A. IRIZARRY

<http://rafalab.org>

Many slides courtesy of:  
Héctor Corrada Bravo and Ben Langmead

# REMEMBER THIS?

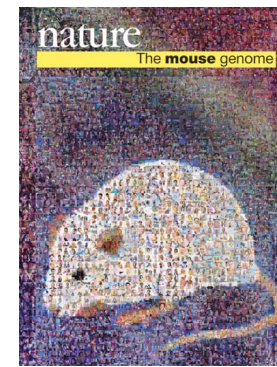
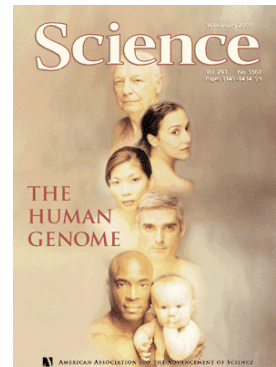
---



*D. melanogaster*, *Science*, 2000



*H. sapiens*, *Nature*, 2000  
and *Science*, 2000



*M. musculus*, *Nature*, 2002



Back then: millions of clones (thousand bps) in 9 months for billions of dollars

Today: billion of short reads (35-100 bps) in a week for thousands of dollars

Claim: Assemble a genome in weeks for less than \$100,000

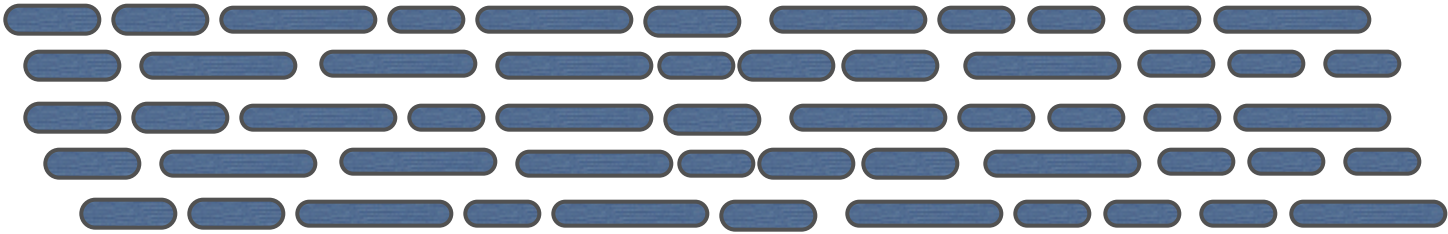
# START WITH DNA (MILLIONS OF COPIES)

---



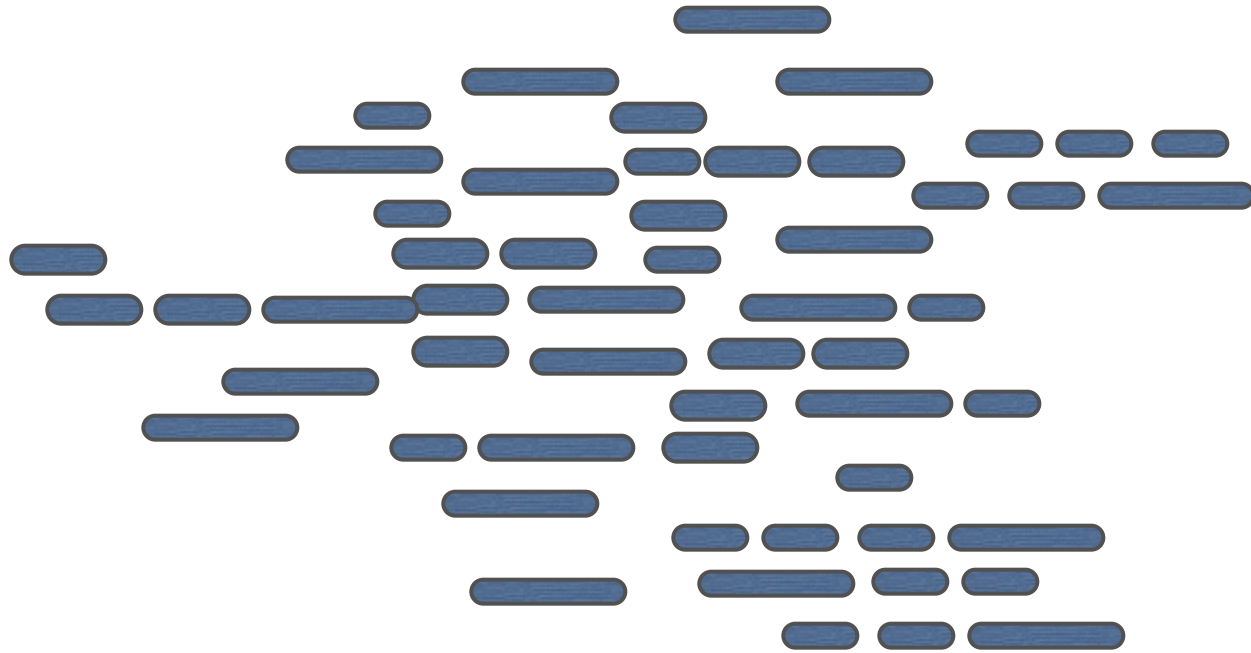
# BREAK IT

---



# PUT IN SEQUENCER

---



# SEQUENCE FIRST 35-400 BPS: CALL THEM “READS”

---

GTTGAGGCTTGCGTTTTTGGTACGCTGGACTTTGT  
GTACTCGTCGCTGCGTTGAGGCTTGCGTTTTTGGT  
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT  
TTGCGTTTTATGGTACGCTGGACTTTGTAGGATACC  
CTTGCGTTTTATGGTACGCTGGACTTTGTAGGATACC  
TTGCGTTTTATGGTACGCTGGACTTTGTAGGATACC  
GCGTTTTATGGTACGCTGGACTTTGTAGGATACCCT  
GAGGCTTGCGTTTTATGGTACGCTGGACTTTGTAGG  
GCGTTGAGGCTTGCGTTTTATGGTACGCTGGATTTT  
CGTTTTATGGTACGCTGGACTTTGTAGGATACCCTC  
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT  
GTTTTATGGTACGCTGGACTTTGTAGGATACCCTCG  
TCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTA  
TGCTCGTCGCTGCGTTGAGGCTTGCGTTTTATGGTA  
GCTCGTCGCTGCGTTGAGGCTTGCGTTTTATGGTAC  
TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT  
TCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTTTG  
CGTCGCTGCGTTGAGGCTTGCGTTTTATGGTACGCT  
GTTGAGGCTTGCGTTTTATGGTACGCTGGGCTTTTT  
TTGCGTTTTATGGTACGCTGGACTTTGTAGGATACC



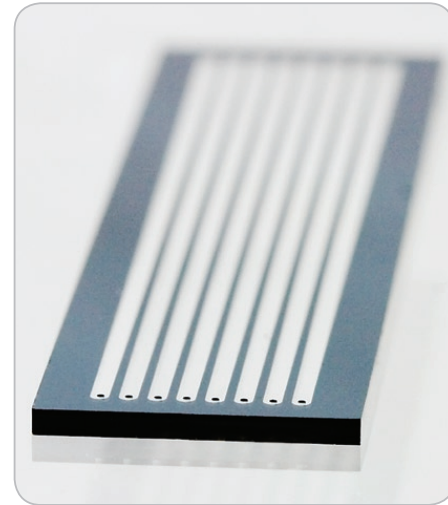
# PLATFORMS

---



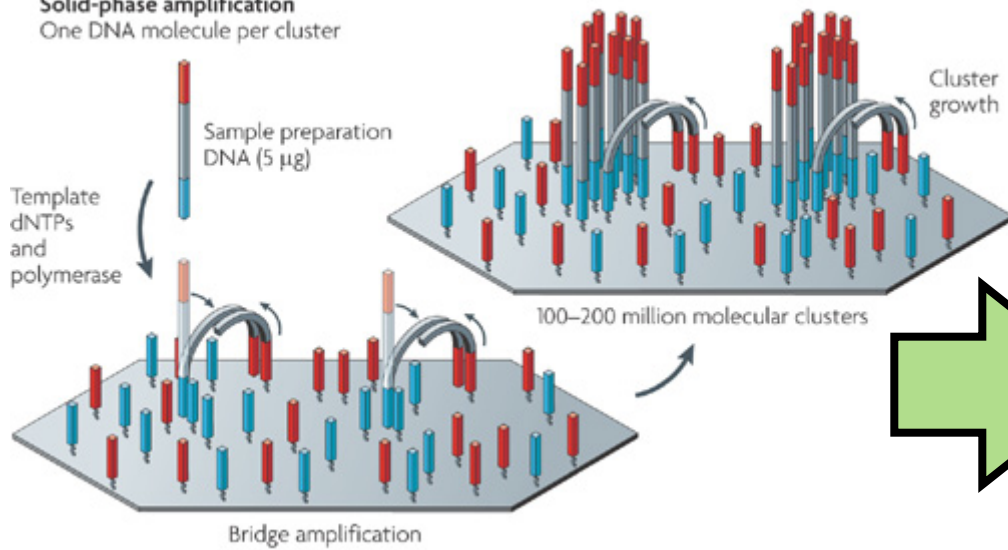
# ILLUMINA/SOLEXA

---



- Eight lanes
- ~160M short reads (~50-70 bp) per lane

**Illumina/Solexa  
Solid-phase amplification**  
One DNA molecule per cluster



Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

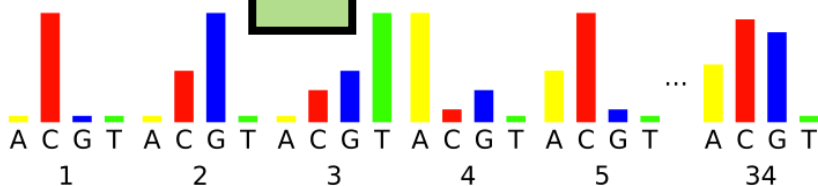
```

@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCGCGGCGCTGNNNNNNNNNNNNNNNN
+
BBBB>A?B@;@BBBBBAA=BA=A%*****
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGNNNNNNNNNNNNNNNN
+
B9B@B<;BAA<@AB9=1-%*****
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNNNNNNNNNNNN
+
A=B7&7:>B@:A>?9;<;>74?%*****
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCGCCAGAAGCAGCAANNNNANTNCTNNNN
+
BBCCCCCBBBCB7CBC=7>+<=>BCBCB%*****
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTATTGTAACTCCGCCTCNNNGNTNAAGNNNN
+
BCC?+<B=?BB5=ABA?B6BBBB4BB?B%*****
@HWI-EAS146:5:1:2:947#0/1
CCCAGGAGAAAGCCATGTTTCAGTTCGAGCGCANNANANCGTANNNN
+
BBB9@?7A7>AAB@>?B=?@.>B7B?%*****
@HWI-EAS146:5:1:2:563#0/1

```

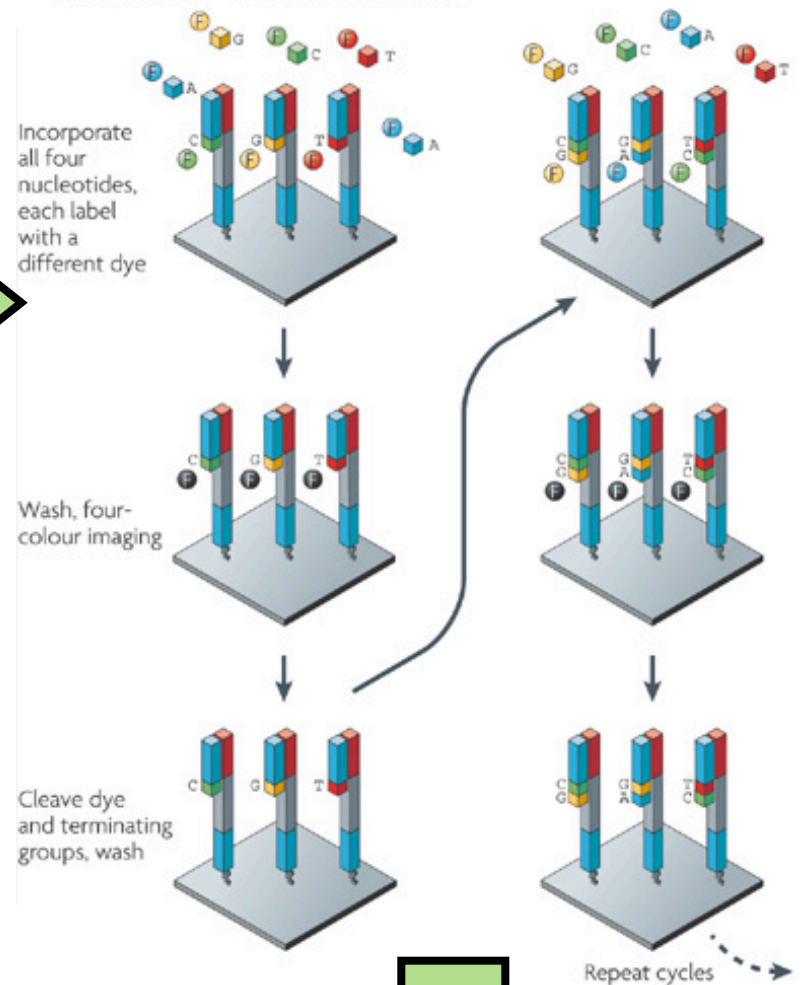
**name**  
**sequence**  
**quality scores**

**x 100s of  
millions**



Source: Whiteford et al. Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics. 2009

**Illumina/Solexa — Reversible terminators**



Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

# NOT JUST ASSEMBLY

---

- Resequencing
- SNP discovery and genotyping
- Variant discovery and quantification
- TF binding sites: ChIP-Seq
- Gene expression: RNA-Seq
- Measuring methylation

# NOT JUST ASSEMBLY



## 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

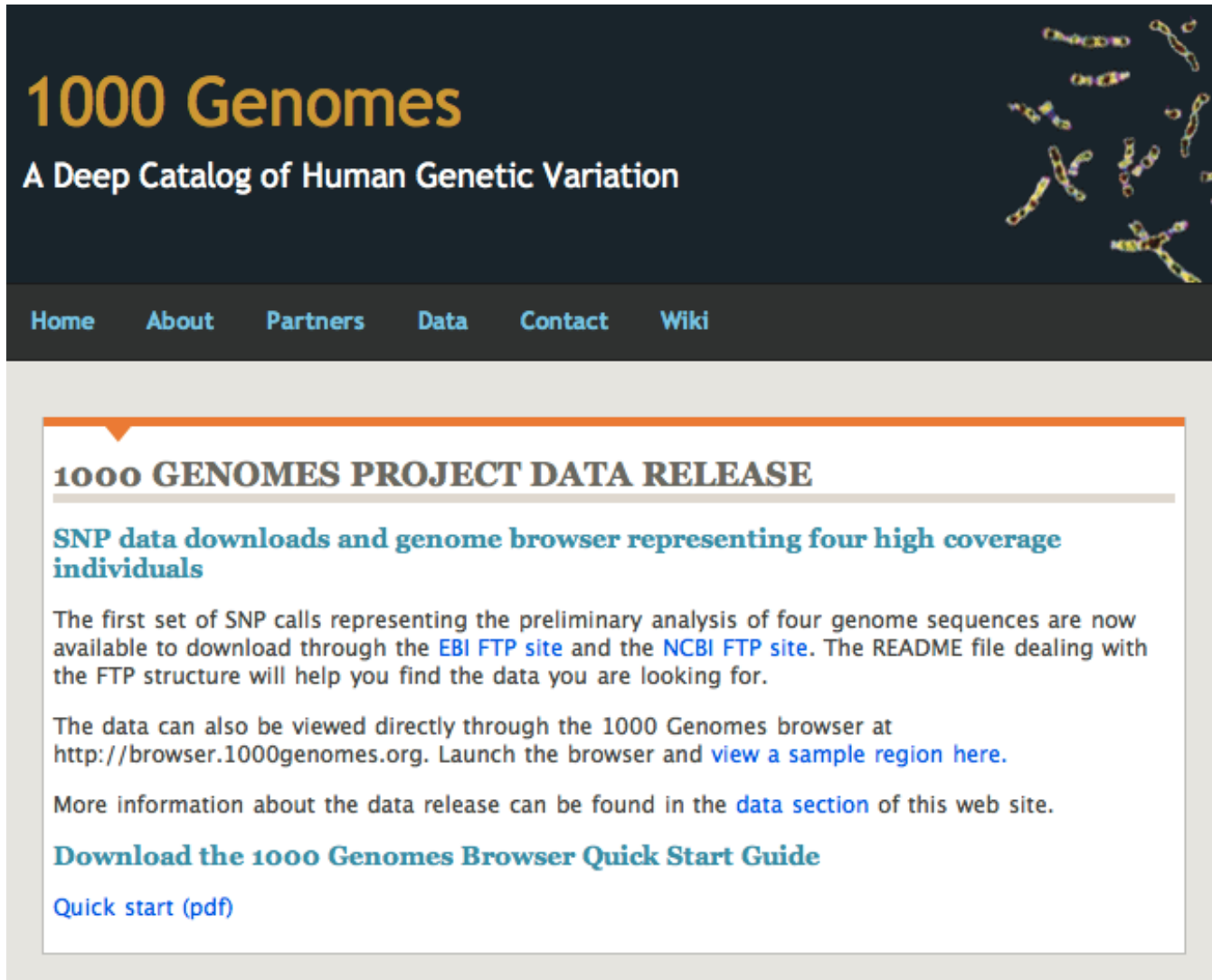
Faster, cheaper DNA sequencing technology is revolutionizing the burgeoning field of personal genomics. But it is having another, more subtle effect.

### Tools



[Send to a Friend](#)

# 1 0 0 0 G E N O M E S P R O J E C T



**1000 Genomes**  
A Deep Catalog of Human Genetic Variation

Home About Partners Data Contact Wiki

## 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 \(pdf\)](#)

*Genotyping*

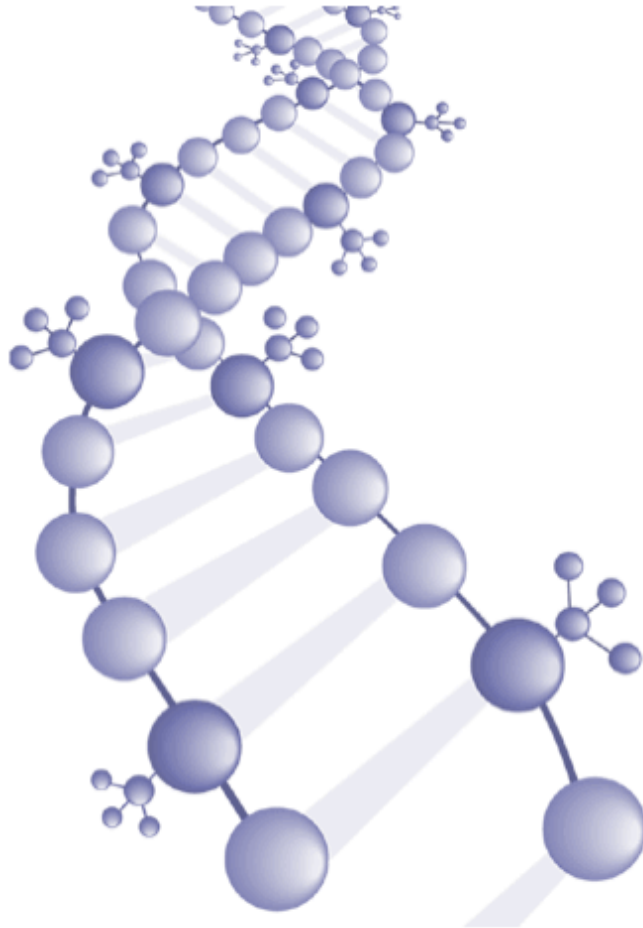
# HUMAN EPIGENOME PROJECT

---

*Methylation*

# HEP

Human  
Epigenome  
Project



# WHAT TO DO WITH ALL THESE SEQUENCES?

---

GTTGAGGCTTGCGTTTTTTGGTACGCTGGACTTTGT  
GTACTCGTCGCTGCGTTGAGGCTTGCGTTTTTTGGT  
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT  
TTGCGTTTTATGGTACGCTGGACTTTGTAGGATACC  
CTTGCGTTTTATGGTACGCTGGACTTTGTAGGATAC  
TTGCGTTTTATGGTACGCTGGACTTTGTAGGATACC  
GCGTTTTATGGTACGCTGGACTTTGTAGGATACCCT  
GAGGCTTGCGTTTTATGGTACGCTGGACTTTGTAGG  
GCGTTGAGGCTTGCGTTTTATGGTACGCTGGATTTT  
CGTTTTATGGTACGCTGGACTTTGTAGGATACCCTC  
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT  
GTTTTATGGTACGCTGGACTTTGTAGGATACCCTCG  
TCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTA  
TGCTCGTCGCTGCGTTGAGGCTTGCGTTTTATGGTA  
GCTCGTCGCTGCGTTGAGGCTTGCGTTTTATGGTAC  
TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT  
TCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTTTTG  
CGTCGCTGCGTTGAGGCTTGCGTTTTATGGTACGCT  
GTTGAGGCTTGCGTTTTATGGTACGCTGGGCTTTTT  
TTGCGTTTTATGGTACGCTGGACTTTGTAGGATACC



# MOST APPS: START BY MATCHING TO REFERENCE

---

GTTGAGGCTTGCCTTTTTGGTACGCTGGACTTTGT  
GTACTCGTCGCTGCCTTGGAGGCTTGCCTTTTTGGT

ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT  
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC  
CTTGCCTTATGGTACGCTGGACTTTGTAGGATAC  
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC  
GCGTTTATGGTACGCTGGACTTTGTAGGATACCCT  
GAGGCTTGCCTTATGGTACGCTGGACTTTGTAGG  
GCGTTGAGGCTTGCCTTATGGTACGCTGGATTTT  
CGTTTATGGTACGCTGGACTTTGTAGGATACCCTC  
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT  
GTTTATGGTACGCTGGACTTTGTAGGATACCCTCG

TCTCGTCGTCGCTGCCTTGGAGGCTTGCCTTTA

TGCTCGTCGCTGCCTTGGAGGCTTGCCTTATGGTA

GCTCGTCGCTGCCTTGGAGGCTTGCCTTATGGTAC

TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT

TCGTGCTCGTCGCTGCCTTGGAGGCTTGCCTTTTTG

CGTCGCTGCCTTGGAGGCTTGCCTTATGGTACGCT

GTTGAGGCTTGCCTTATGGTACGCTGGGCTTTTT

TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC

---

CTCTCGTCGTCGCTGCCTTGGAGGCTTGCCTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTTC

# Variant detection

```
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGCGCTGNNNNNNNNNCNNNN
+
BBBB->A7B@;@BBBBAA=BA=A*****
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGACGAGNNNNNNNGNNNN
+
B9B@< ;BAA<@AB9=1>*****
@HWI-EAS146:5:1:1:1048#0/1
CTGACTGCATCTACCACCACTCGTCCAANNNNCNNNCNNNN
+
A=B7G7:>B@:A>79:<::>74?*****
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCGCCAGAAGCACAGCCAANNANTNCTNNNN
+
BBCCCCCBB7CB7C=7>+<=>=BCBCB*****
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCTCNNGNTHAAGNNNN
+
BCC7+<B=7BB5=ABA?B6BBBB4BB7B*****
@HWI-EAS146:5:1:2:947#0/1
CCAGGAGAAAGCATGTTCAAGTTCGAGGCNNANANCGTANNNN
+
BBB9@77A7>AAB@>7B=7@.>87B?*****
@HWI-EAS146:5:1:2:563#0/1
CCAGCCCTCCCATCTCCACCTGTACCTNANCCCTGANNNN
+
BBABAABB;AAABA77@SAAA:??>*****
@HWI-EAS146:5:1:2:1631#0/1
TGGGAACGCAGCTACACTCTTCCAGGCCTCTNCTCCGTNNNN
+
BBB@<@BBBBBB@BBBBBAABBB7;9BB@BA56<B:*****
@HWI-EAS146:5:1:2:1420#0/1
CTCAAACCTCTGACCTTGGTGATCCACCCGCTNGGCCCTCANN
+
BBBB:BBBBBAAA7:(=A8@>AAA7AB7=A*****
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGCGCTGNNNNNNNNNCNNNN
+
BBBB->A7B@;@BBBBAA=BA=A*****
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGACGAGNNNNNNNGNNNN
+
B9B@< ;BAA<@AB9=1>*****
@HWI-EAS146:5:1:1:1048#0/1
CTGACTGCATCTACCACCACTCGTCCAANNNNCNNNCNNNN
+
A=B7G7:>B@:A>79:<::>74?*****
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCGCCAGAAGCACAGCCAANNANTNCTNNNN
+
BBCCCCCBB7CB7C=7>+<=>=BCBCB*****
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCTCNNGNTHAAGNNNN
+
BCC7+<B=7BB5=ABA?B6BBBB4BB7B*****
@HWI-EAS146:5:1:2:947#0/1
CCAGGAGAAAGCATGTTCAAGTTCGAGGCNNANANCGTANNNN
+
BBB9@77A7>AAB@>7B=7@.>87B?*****
@HWI-EAS146:5:1:2:563#0/1
CCAGCCCTCCCATCTCCACCTGTACCTNANCCCTGANNNN
+
BBABAABB;AAABA77@SAAA:??>*****
@HWI-EAS146:5:1:2:1631#0/1
TGGGAACGCAGCTACACTCTTCCAGGCCTCTNCTCCGTNNNN
+
RRRRRRRRRRRRRRRRRRR&R&R&R7-@RRR&R&R&R-*****
```



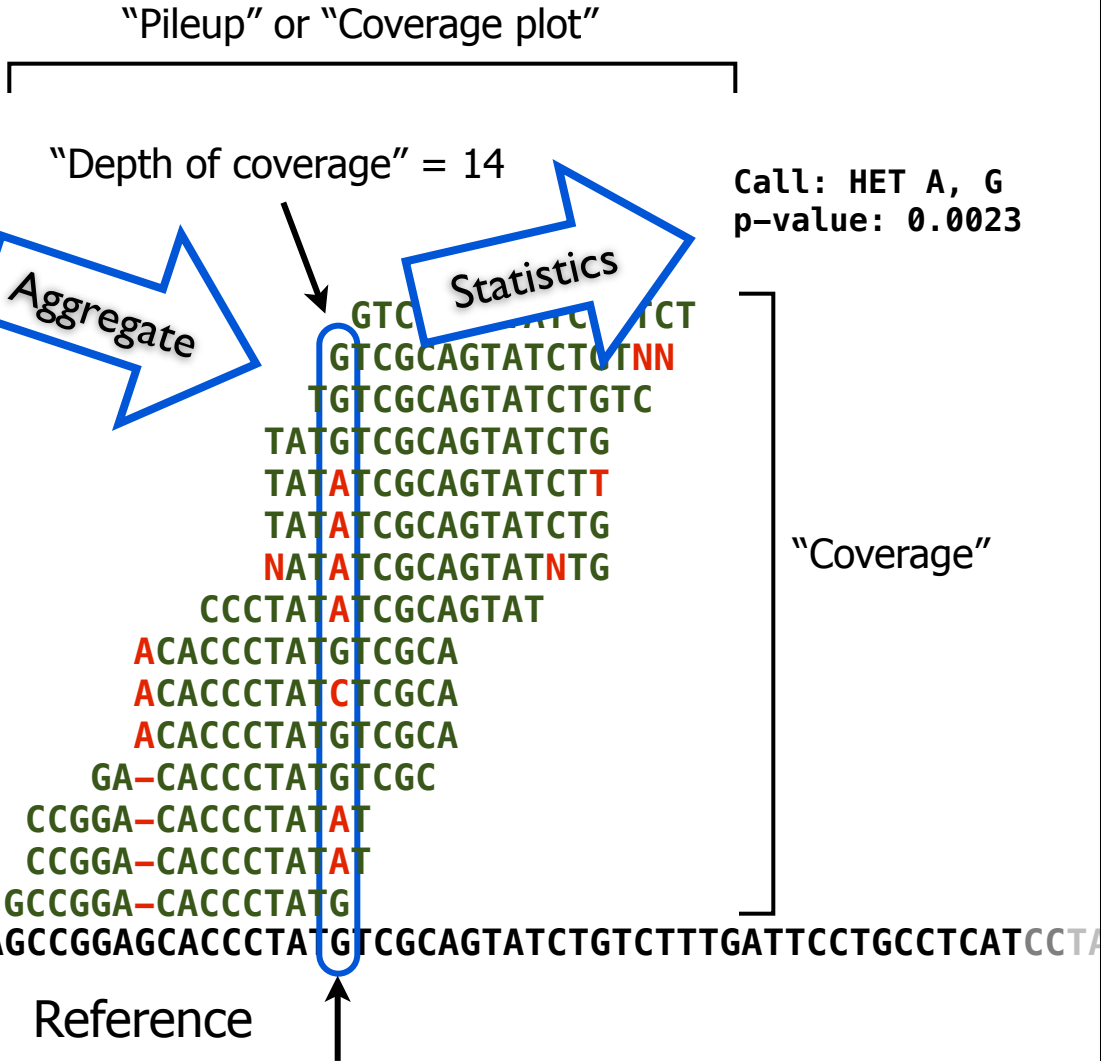
```
GTCGCAGTANCTGTCT
||||||| |||||
GTCGCAGTATCTGTCT

GGATCTGCGATATACC
||||||| |||||
GGATCT-CGATATACC

AATCTGATCTTATTTT
||||||| |||||
AATCTGATCTTATTTT

ATATATATATATATAT
||||||| |||||
ATATATATATATATAT

TCTCTCCCANNAGAGC
||||||| |||||
TCTCTCCCAGGAGAGC
```



TTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCTATGTCGCAGTATCTGTCTTTGATTCTGCCTCATCCTA

# RNA-seq differential expression

```

@HWI-EAS146:5:1:1:961#0/1
TCGAGGCCAACCGAGGTCCGCGCCCTGNNNNNNNNNNNN
+
BBB-17B:BBBBA=BA=AAAAA
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCGAAGAGCTGCTCAGCAGGNNNNNNNNNN
+
BBB-17A:BBBBA=BA=AAAAA
@HWI-EAS146:5:1:1:1595#0/1
TCGAGTCTCCAGCTCCTCAGAGAGCAGCCANNNNNNNN
+
A=B747>B0:A=79<:747
@HWI-EAS146:5:1:1:1887#0/1
CTCTCTCAAGGTCCCGAGAGCAGCCANNNNNNNNNN
+
BBCCCCCBBTCTC=7-->=BCBCPV
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTATTTGTAACCTCCGCTCNNGNNTAAGN
+
BCT7-->=785-ABA788BB848B70V
@HWI-EAS146:5:1:2:947#0/1
CCGAGGAAGACCATTTGAGTCGAGCCNNANACGTGANNN
+
BBB977A7>AAB>7B>7B>=87B7
@HWI-EAS146:5:1:2:953#0/1
CCAGCCCTCCCTCCTCCAGCTCCTACTCTANCCCTGANNN
+
BBBAAB>:AABA7705AAA:77>
@HWI-EAS146:5:1:2:1631#0/1
TGGGAACGAGCTACACTTCCAGGCTCTTCCGCTNN
+
BBB988BB8BB8BB8BABA8BB7:988@A56-B:
@HWI-EAS146:5:1:2:1629#0/1
CTCAAACCTGACTTTGGTGTACCCCGCTGCGCTCN
+
BBB:BBBBA=BA=AAAAA
@HWI-EAS146:5:1:1:961#0/1
TCGAGGCCAACCGAGGTCCGCGCCCTGNNNNNNNNNN
+
BBB-17B:BBBBA=BA=AAAAA
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCGAAGAGCTGCTCAGCAGGNNNNNNNNNN
+
BBB-17A:BBBBA=BA=AAAAA
@HWI-EAS146:5:1:1:1595#0/1
TCGAGTCTCCAGCTCCTCAGAGAGCAGCCANNNNNNNN
+
A=B747>B0:A=79<:747
@HWI-EAS146:5:1:1:1887#0/1
CTCTCTCAAGGTCCCGAGAGCAGCCANNNNNNNNNN
+
BBCCCCCBBTCTC=7-->=BCBCPV
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTATTTGTAACCTCCGCTCNNGNNTAAGN
+
BCT7-->=785-ABA788BB848B70V
@HWI-EAS146:5:1:2:947#0/1
CCGAGGAAGACCATTTGAGTCGAGCCNNANACGTGANNN
+
BBB977A7>AAB>7B>7B>=87B7
@HWI-EAS146:5:1:2:953#0/1
CCAGCCCTCCCTCCTCCAGCTCCTACTCTANCCCTGANNN
+
BBBAAB>:AABA7705AAA:77>
@HWI-EAS146:5:1:2:1631#0/1
TGGGAACGAGCTACACTTCCAGGCTCTTCCGCTNN
+
BBB988BB8BB8BB8BABA8BB7:988@A56-B:
@HWI-EAS146:5:1:2:1629#0/1
CTCAAACCTGACTTTGGTGTACCCCGCTGCGCTCN
+
BBB:BBBBA=BA=AAAAA
@HWI-EAS146:5:1:1:961#0/1
TCGAGGCCAACCGAGGTCCGCGCCCTGNNNNNNNNNN
+
BBB-17B:BBBBA=BA=AAAAA
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCGAAGAGCTGCTCAGCAGGNNNNNNNNNN
+
BBB-17A:BBBBA=BA=AAAAA
@HWI-EAS146:5:1:1:1595#0/1
TCGAGTCTCCAGCTCCTCAGAGAGCAGCCANNNNNNNN
+
A=B747>B0:A=79<:747
@HWI-EAS146:5:1:1:1887#0/1
CTCTCTCAAGGTCCCGAGAGCAGCCANNNNNNNNNN
+
BBCCCCCBBTCTC=7-->=BCBCPV
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTATTTGTAACCTCCGCTCNNGNNTAAGN

```

## Sample A



```

GTCGCAGTANCTGTCT
||||||| |||||
GTCGCAGTATCTGTCT

GGATCTCGCATATACC
||||| |||||
GGATCT-CGATATACC

AATCTGATCTTATTTT
||||||| |||||
AATCTGATCTTATTTT

ATATATATATATATAT
||||||| |||||
ATATATATATATATAT

TCTCTCCANNAGAGC
||||||| |||||
TCTCTCCAGGAGAGC

```



```

GTCGCAGTATCTGTCT
GTCGCAGTATCTGTCT
GTCGCAGTATCTGTCT
GTCGCAGTATCTGTCT
GTCGCAGTATCTGTCT
TGTGCGAGTATCTGTC
TATGTCGCAGTATCTG
TATATCGCAGTATCTG
TATATCGCAGTATCTG
TATATCGCAGTATCTG
CCCTATATCGCAGTAT
AGCACCTATGTCGCA
AGCACCTATATCGCA
AGCACCTATGTCGCA
GAGCACCTATGTCG
CCGGAGCACCTATAT
CCGGAGCACCTATAT
GCCGGAGCACCTATG

```



Gene 1  
differentially  
expressed?: YES  
p-value: 0.0012

Gene 1  
CATTGGTATTTTCGCTGGGGGATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCTATGTCGCAGTATCTGTCTTTGATTTCCTGCCTCATCTATTATTTATCGCACCT

```

@HWI-EAS146:5:1:1:961#0/1
TCGAGGCCAACCGAGGTCCGCGCCCTGNNNNNNNNNN
+
BBB-17B:BBBBA=BA=AAAAA
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCGAAGAGCTGCTCAGCAGGNNNNNNNNNN
+
BBB-17A:BBBBA=BA=AAAAA
@HWI-EAS146:5:1:1:1595#0/1
TCGAGTCTCCAGCTCCTCAGAGAGCAGCCANNNNNNNN
+
A=B747>B0:A=79<:747
@HWI-EAS146:5:1:1:1887#0/1
CTCTCTCAAGGTCCCGAGAGCAGCCANNNNNNNNNN
+
BBCCCCCBBTCTC=7-->=BCBCPV
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTATTTGTAACCTCCGCTCNNGNNTAAGN
+
BCT7-->=785-ABA788BB848B70V
@HWI-EAS146:5:1:2:947#0/1
CCGAGGAAGACCATTTGAGTCGAGCCNNANACGTGANNN
+
BBB977A7>AAB>7B>7B>=87B7
@HWI-EAS146:5:1:2:953#0/1
CCAGCCCTCCCTCCTCCAGCTCCTACTCTANCCCTGANNN
+
BBBAAB>:AABA7705AAA:77>
@HWI-EAS146:5:1:2:1631#0/1
TGGGAACGAGCTACACTTCCAGGCTCTTCCGCTNN
+
BBB988BB8BB8BB8BABA8BB7:988@A56-B:
@HWI-EAS146:5:1:2:1629#0/1
CTCAAACCTGACTTTGGTGTACCCCGCTGCGCTCN
+
BBB:BBBBA=BA=AAAAA
@HWI-EAS146:5:1:1:961#0/1
TCGAGGCCAACCGAGGTCCGCGCCCTGNNNNNNNNNN
+
BBB-17B:BBBBA=BA=AAAAA
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCGAAGAGCTGCTCAGCAGGNNNNNNNNNN
+
BBB-17A:BBBBA=BA=AAAAA
@HWI-EAS146:5:1:1:1595#0/1
TCGAGTCTCCAGCTCCTCAGAGAGCAGCCANNNNNNNN
+
A=B747>B0:A=79<:747
@HWI-EAS146:5:1:1:1887#0/1
CTCTCTCAAGGTCCCGAGAGCAGCCANNNNNNNNNN
+
BBCCCCCBBTCTC=7-->=BCBCPV
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTATTTGTAACCTCCGCTCNNGNNTAAGN

```

## Sample B



```

GTCGCAGTANCTGTCT
||||||| |||||
GTCGCAGTATCTGTCT

GGATCTCGCATATACC
||||| |||||
GGATCT-CGATATACC

AATCTGATCTTATTTT
||||||| |||||
AATCTGATCTTATTTT

ATATATATATATATAT
||||||| |||||
ATATATATATATATAT

TCTCTCCANNAGAGC
||||||| |||||
TCTCTCCAGGAGAGC

```



```

TGTCGCAGTATCTGTC
AGCACCTATGTCGCA
GCCGGAGCACCTATG

```

# ChIP-seq

```

@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNNNCNNNNN
+
BBBB->A7D@;@BBBBAA=BA=A*****
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGACAGGNNNNNNNNNGNNNNN
+
B9B@< ; BAA-<@B9=1>*****
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCTACCACCACTCGTCCAANNNNCNNNCNNNNN
+
A=B767:>B@:A>79:<;>747*****
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCAGAAGCACAGCAANNNNANTNCTNNNN
+
BBCCCCCBBBCB7CB7C=>+<=>=BCBCB*****
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCCTCNNNNGNTHAAGNNN
+
BCC7+<B=7BB5=ABA?B6BBBB4BB7B*****
@HWI-EAS146:5:1:2:947#0/1
CCCAGGAGAAAGCATGTTCAAGTTCGAGGCGNANANCGTANNNN
+
BBB9@77A7>AAB@7B=7@.>87B7*****
@HWI-EAS146:5:1:2:563#0/1
CCAGCCCCCTCCCATCTCCCACTGTACTNANCCCTGANNNN
+
BBABAABB;AAABA77@SAAA:??*****
@HWI-EAS146:5:1:2:1631#0/1
TGGGAACGAGCTACACTCTTCCAGGCTCCTNCTCCGTNNNN
+
BBB@<@BBBBBB@BBBBABAABB7;9BB@BA56<B:*****
@HWI-EAS146:5:1:2:1420#0/1
CTCAACTCTGACCTTTGGTATCCACCGCTNGGCCCTCNNNN
+
BBBB:BBBBBAAAA?:(=A8@>AAA7AB7=A*****
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGGCGCTGNNNNNNNNCNNNN
+
BBBB->A7D@;@BBBBAA=BA=A*****
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGACAGGNNNNNNNNNGNNNNN
+
B9B@< ; BAA-<@B9=1>*****
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCTACCACCACTCGTCCAANNNNCNNNCNNNNN
+
A=B767:>B@:A>79:<;>747*****
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCCAGAAGCACAGCAANNNNANTNCTNNNN
+
BBCCCCCBBBCB7CB7C=>+<=>=BCBCB*****
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTTATTGTAACCTCCGCCTCNNNNGNTHAAGNNN
+
BCC7+<B=7BB5=ABA?B6BBBB4BB7B*****
@HWI-EAS146:5:1:2:947#0/1
CCCAGGAGAAAGCATGTTCAAGTTCGAGGCGNANANCGTANNNN
+
BBB9@77A7>AAB@7B=7@.>87B7*****
@HWI-EAS146:5:1:2:563#0/1
CCAGCCCCCTCCCATCTCCCACTGTACTNANCCCTGANNNN
+
BBABAABB;AAABA77@SAAA:??*****
@HWI-EAS146:5:1:2:1631#0/1
TGGGAACGAGCTACACTCTTCCAGGCTCCTNCTCCGTNNNN
+
RRRRRRRRRRRRRRRRRRRRR?>RRRRR?<R:*****

```



```

GTCGCAGTANCTGTCT
||||||| |||||
GTCGCAGTATCTGTCT

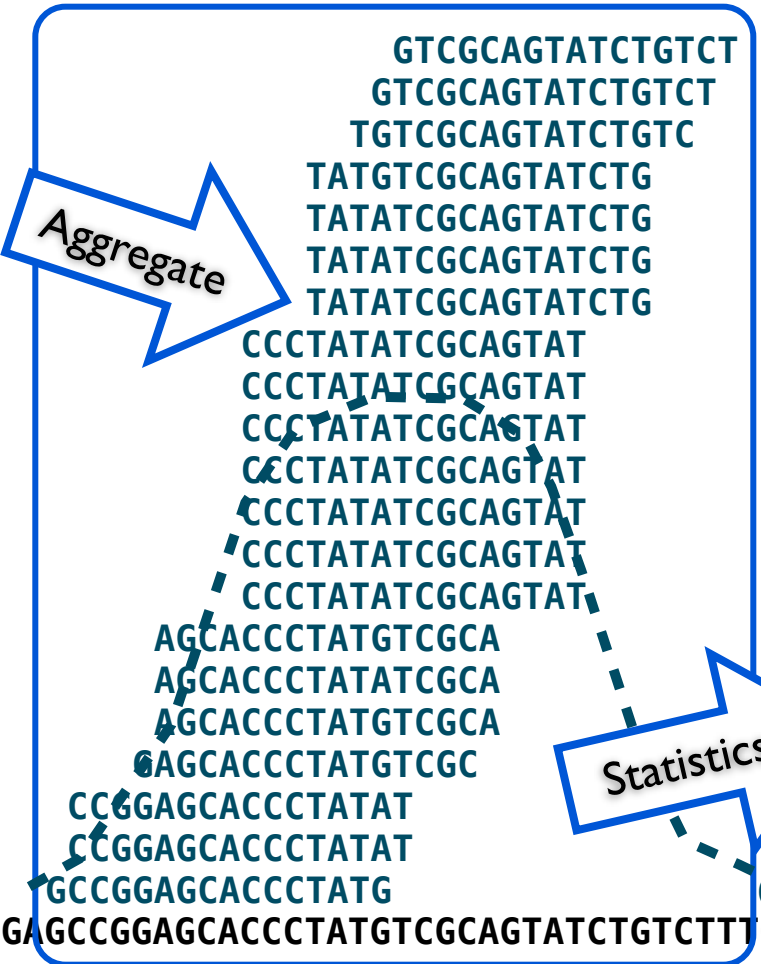
GGATCTGCGATATACC
||||| |||||
GGATCT-CGATATACC

AATCTGATCTTATTTT
|||||||
AATCTGATCTTATTTT

ATATATATATATATAT
|||||||
ATATATATATATATAT

TCTCTCCCANNAGAGC
||||||| |||||
TCTCTCCAGGAGAGC

```



Binding occurs here  
p-value:  
0.0023

GATAGCATTGCGAGAC  
TATGCACGCGATAGCA  
TTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGA  
GCC&GAGCACCCTATGTCGCAGTATCTGTCTTT  
GATTCCCTGCCTCATCC

Reference

# MATCHING REVISTED

---

GTTGAGGCTTGCCTTTTTGGTACGCTGGACTTTGT  
GTACTCGTCGCTGCCTTGGAGGCTTGCCTTTTTGGT

ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT  
TTGCGTTTTATGGTACGCTGGACTTTGTAGGATACC  
CTTGCCTTTATGGTACGCTGGACTTTGTAGGATAC  
TTGCGTTTTATGGTACGCTGGACTTTGTAGGATACC  
GCGTTTTATGGTACGCTGGACTTTGTAGGATACCCT  
GAGGCTTGCCTTTATGGTACGCTGGACTTTGTAGG  
GCGTTGAGGCTTGCCTTTATGGTACGCTGGATTTT

CGTTTTATGGTACGCTGGACTTTGTAGGATACCCTC  
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT  
GTTTTATGGTACGCTGGACTTTGTAGGATACCCTCG

TCTCGTCGTCGCTGCCTTGGAGGCTTGCCTTTA  
TGCTCGTCGCTGCCTTGGAGGCTTGCCTTTATGGTA  
GCTCGTCGCTGCCTTGGAGGCTTGCCTTTATGGTAC

TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT  
TCGTGCTCGTCGCTGCCTTGGAGGCTTGCCTTTTTG  
CGTCGCTGCCTTGGAGGCTTGCCTTTATGGTACGCT

GTTGAGGCTTGCCTTTATGGTACGCTGGGCTTTTT  
TTGCGTTTTATGGTACGCTGGACTTTGTAGGATACC

---

CTCTCGTCGTCGCTGCCTTGGAGGCTTGCCTTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTTC

# MATCHING 10,000,000 32 BPS READS

---

- BLAST takes more than 6 months
- BLAT takes 2 months
- MAQ takes 1 day and half
- Bowtie takes 17 minutes

# MATCHING

GTTGAGGCTTGCCTTTTTGGTACGCTGGACTTTGT  
GTACTCGTCGCTGCGTTGAGGCTTGCCTTTTTGGT  
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT  
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC  
CTTGCCTTTATGGTACGCTGGACTTTGTAGGATAC

**Bowtie**

An ultrafast memory-efficient short read aligner



JOHNS HOPKINS  
BLOOMBERG  
SCHOOL of PUBLIC HEALTH

**Bowtie** is an ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome (2.9 GB for paired-end).



TGCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTA  
GCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTAC  
TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT  
TCGTGCTCGTCGCTGCGTTGAGGCTTGCCTTTTTG  
CGTCGCTGCGTTGAGGCTTGCCTTTATGGTACGCT  
GTTGAGGCTTGCCTTTATGGTACGCTGGGCTTTTT  
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC

CTCTCGTGCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTTC

# Mapping

Take a read:

```
CTCAAACCTCTGACCTTTGGTGATCCACCCGCTNGGCCTTC
```

And a reference sequence:

```
>MT dna:chromosome chromosome:GRCh37:MT:1:16569:1  
GATCACAGGTCATCACCCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT  
CGTCTGGGGGGTATGCACGGGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCCTATGTC  
GCAGTATCTGTCTTTGATTCCTGCCTCATCTATTATTTATCGCACCTACGTTCAATATT  
ACAGGCGAACATACTTACTAAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA  
ACAATTGAATGTCGACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA  
AACCCCTCTCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAACCCCAAAA  
ACAAAGAACCTAACACCAGCTAACCCAGATTTCAAATTTTATCTTTTGGCGGTATGCAC  
TTTTAACAGTCAACCCCACTAACACATTATTTCCCTCCCACTCCATACTACTAAT  
CTCATCAATACAACCCCGCCATCCTACCCAGCACACACACACCGCTGTAAACCCATA  
CCCCGAACCAACAAACCCCGGAAACAGACGCGGACAGCTTTATCTAGCTTAGCTCCTCAA  
GCAATACACTGACCCCTCAAACCTCTGGATTTGGATCCACCAGCGCTTGGCCTAACT  
CTAGCCTTTCTATTAGCTCTTAACTAATTAACATGATCAACATCCCTCCCTCCTCACTGAT  
TCACCCCTCAAATCACCAGATCAAAAGGAAACAAGCATCAAGCACGCGAATGCAGCTC  
AAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTATTAACCTTAGCAATAA  
ACGAAAGTTAACTAAGCTATAACCCAGGGTGGTCAATTTCTGCCAGCCACCCG  
GGTACACAGATTAACCCCAAGTCAATAGAAGCCGCGTAAAGAGTGTTTTAGATCACCC  
TCCCAATAAAAGCTAAAACCTCACCTGAGTTGTAAGAAACTCCAGTTGACACAAAATAGAC  
TACGAAAGTGGCTTTAACAATATCTGAACACACAATAGCTAAGACCCAACTGGGATTAGA  
TACCCCACTATGCTTAGCCCTAAACCTCAACAGTTAAATCAACAAAACCTGCTCGCCAGAA  
CACTACGAGCCACAGCTTAAAACCTCAAAGGACCTGGCGGTGCTTCATATCCCTTAGAGG  
AGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACCTCTTGCTCAGCCTATATA  
CCGCCATCTCAGCAAACCTGATGAAGGTACAAGTAAGCGCAAGTACCCACGTAAG  
ACGTTAGGTCAAGGTGTAGCCATGAGGTGGCAAGAAATGGGCTACATTTTCTACC  
AAAACCTACGATAGCCCTTATGAACTTAAGGTCGAAGGTGGATTTAGCAGTAACCTAAG  
AGTAGAGTGCTTAGTTGAACAGGCGCTGAAGCGGTACACACCCGCTGACCTCTCTC  
AAGTATCTTCAAGGACATTAACTAAAACCCCTAGCCATTATATAGAGGAGACAAGT  
CGTAACTCAAACCTCTGCCTTTGGTGATCCACCCGCTTGGCCTACCTGCATAATGAAG  
AAGCACCCAACTTACACTTAGGATTTCAACTTAACTTAACTTAACTTAACTTAACTTAA  
GCCCCAACCCACTCCACCTTACTACCAGACAACTTAGCCAAACCATTTACCCAAATAA  
AGTATAGGCGATAGAATTAAGAACTGGCGCAATAGATATAGTACCGCAAGGAAAGATG  
AAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTTCTGCATAATGAA  
TTAACTAGAAATACTTTGCAAGGAGAGCCAAAGCTAAGACCCCGAAACCAGACGAGCT
```

How do we determine the read's point of origin with respect to the reference?

Answer: sequence similarity

Hypothesis 1:



Hypothesis 2:



Which hypothesis is better?

Say hypothesis 2 is correct. Why are there still mismatches and gaps?



# More on variants and base-calling

# SNPs

GTTGAGGCTTGCCTTTT**T**TGGTACGCTGGACTTTGT  
GTACTCGTCGCTGCGTTGAGGCTTGCCTTTT**T**TGGT

**A**TGGTACGCTGGACTTTGTAGGATACCCTCGCTTT

TTGCGTTT**A**TGGTACGCTGGACTTTGTAGGATACC

CTTGCCTTT**A**TGGTACGCTGGACTTTGTAGGATACC

TTGCGTTT**A**TGGTACGCTGGACTTTGTAGGATACC

GCGTTT**A**TGGTACGCTGGACTTTGTAGGATACCCT

GAGGCTTGCCTTT**A**TGGTACGCTGGACTTTGTAGG

GCGTTGAGGCTTGCCTTT**A**TGGTACGCTGGATTTT

CGTTT**A**TGGTACGCTGGACTTTGTAGGATACCCTC

**A**TGGTACGCTGGACTTTGTAGGATACCCTCGCTTT

GTTT**A**TGGTACGCTGGACTTTGTAGGATACCCTCG

TCTCGTGCTCGTCGCTGCGTTGAGGCTTGCCTTT**A**

TGCTCGTCGCTGCGTTGAGGCTTGCCTTT**A**TGGTA

GCTCGTCGCTGCGTTGAGGCTTGCCTTT**A**TGGTAC

**T**A**T**G**G**T**A**C**G**C**T**G**G**A**C**T**T**T**G**T**A**G**G**A**T**A**C**C**T**C**G**C**T**T

TCGTGCTCGTCGCTGCGTTGAGGCTTGCCTTT**T**TG

CGTCGCTGCGTTGAGGCTTGCCTTT**A**TGGTACGCT

GTTGAGGCTTGCCTTT**A**TGGTACGCTGGGCTTTTT

TTGCGTTT**A**TGGTACGCTGGACTTTGTAGGATACC

---

CTCTCGTGCTCGTCGCTGCGTTGAGGCTTGCCTTT**A**TGGTACGCTGGACTTTGTAGGATACCCTCGCTTTC

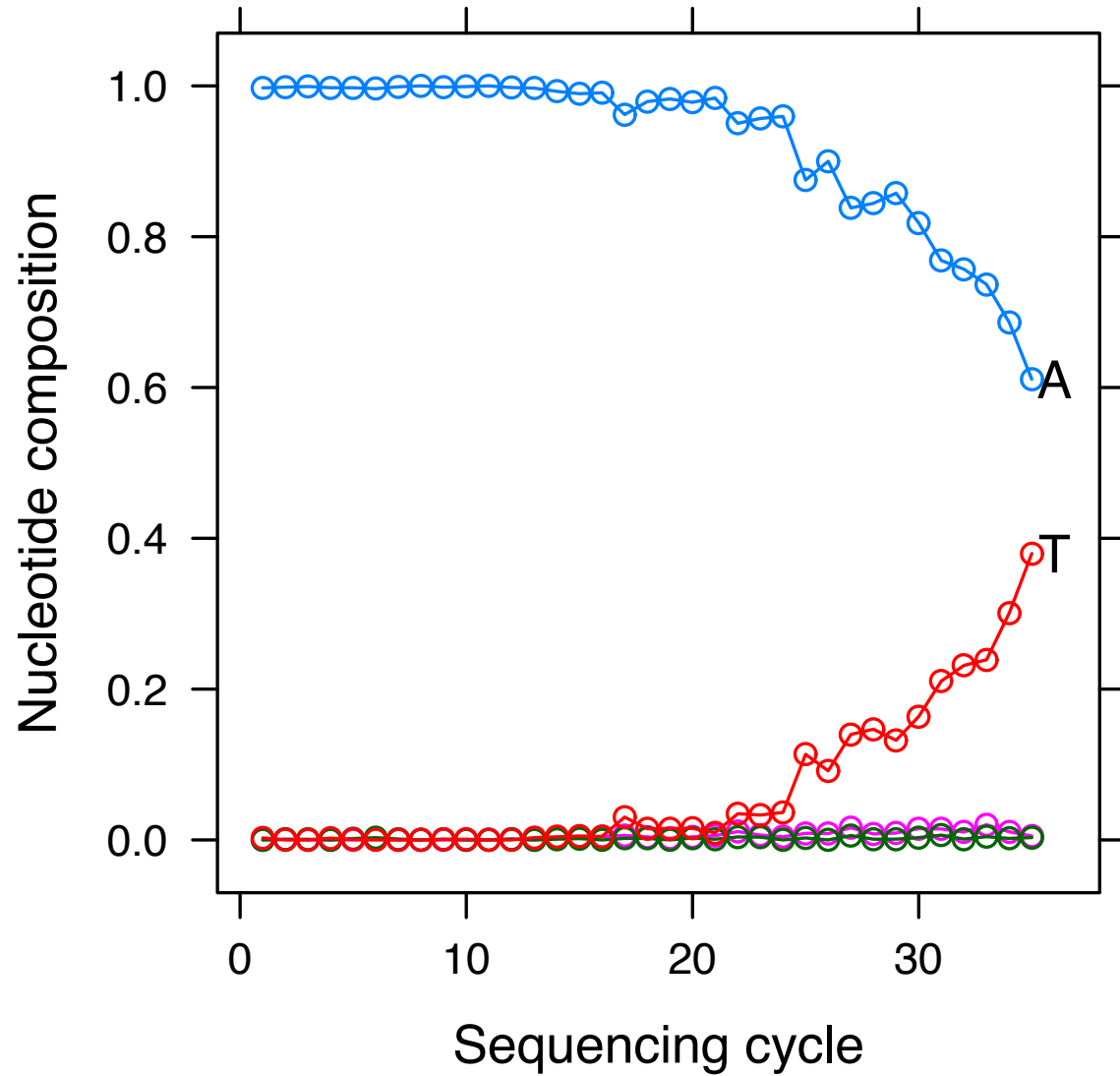
# SNPs

TCTCGTGCCTCGTCGCTGCGTTGAGGCTTGCCTTTA  
TCGTGCTCGTCGCTGCGTTGAGGCTTGCCTTTTGTG  
GTACTCGTCGCTGCGTTGAGGCTTGCCTTTTGTGGT  
TGCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTA  
GCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTAC  
CGTCGCTGCGTTGAGGCTTGCCTTTATGGTACGCT  
GCGTTGAGGCTTGCCTTTATGGTACGCTGGATTTT  
GTTGAGGCTTGCCTTTTGGTACGCTGGACTTTGT  
GTTGAGGCTTGCCTTTATGGTACGCTGGGCTTTTT  
GAGGCTTGCCTTTATGGTACGCTGGACTTTGTAGG  
CTTGCCTTTATGGTACGCTGGACTTTGTAGGATAC  
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC  
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC  
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC  
GCGTTTATGGTACGCTGGACTTTGTAGGATACCCT  
CGTTTATGGTACGCTGGACTTTGTAGGATACCCTC  
GTTTATGGTACGCTGGACTTTGTAGGATACCCTCG  
TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT  
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT  
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT

---

CTCTCGTGCCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTTC

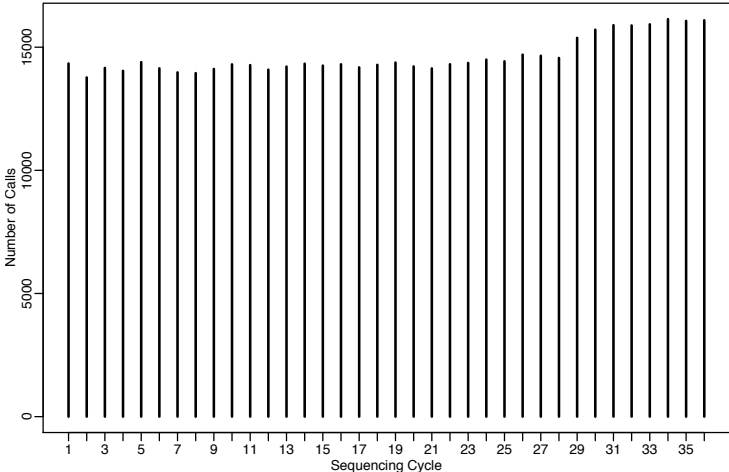
# All Reads



# 1000 Genomes Data

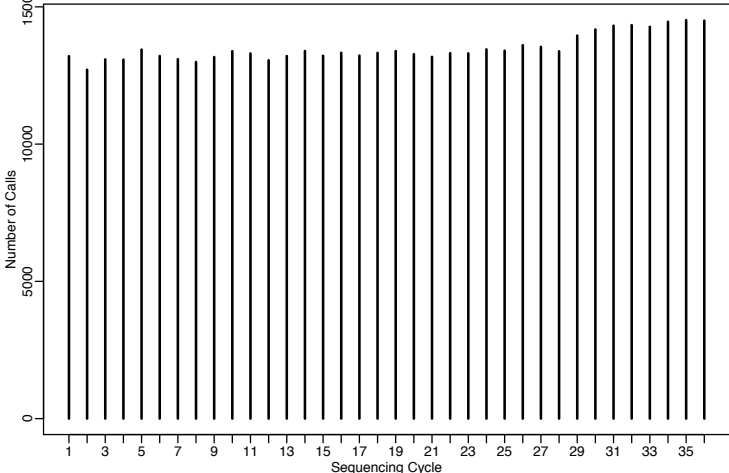
## SNPs in dbSNP

Sample NA19238, UCSC Loci, n=48864



All data

Sample NA19238, UCSC Loci, n=42691



Filtered:  
snpg>=20,  
nreads<=360

Here we aggregate reads and record cycle at which variant appears

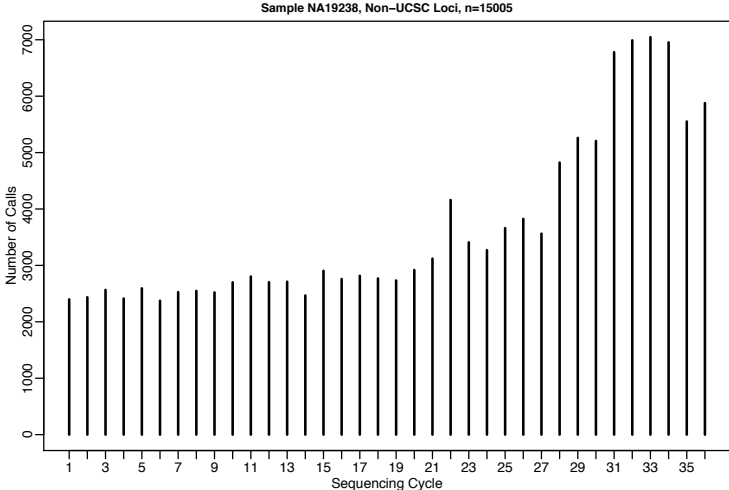
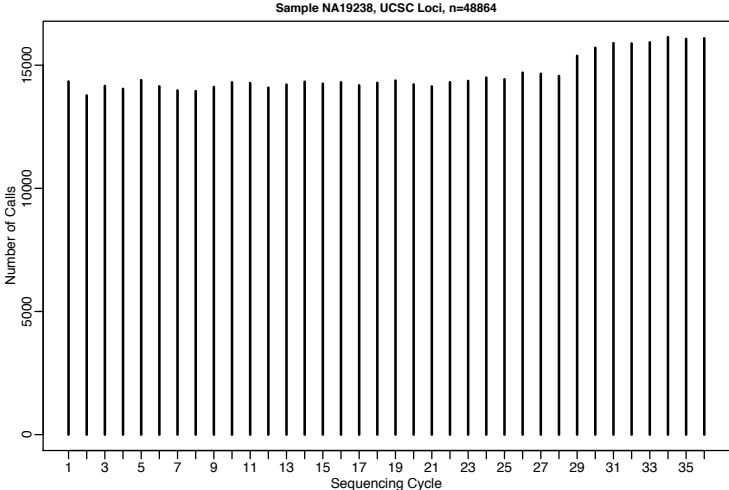
Cycle

# 1000 Genomes Data

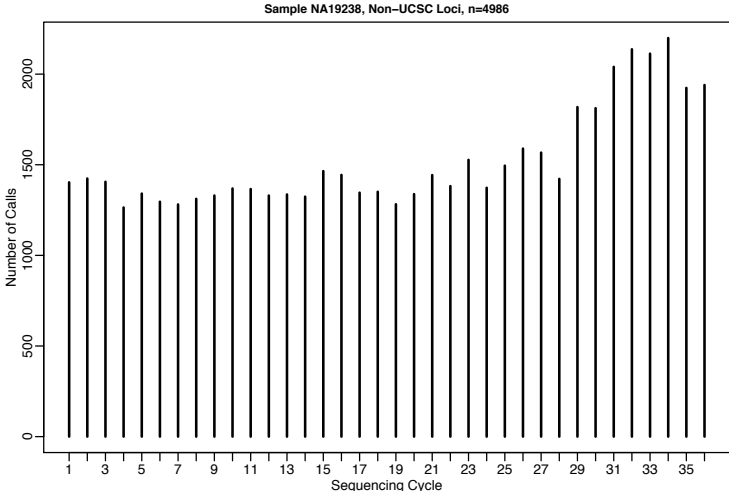
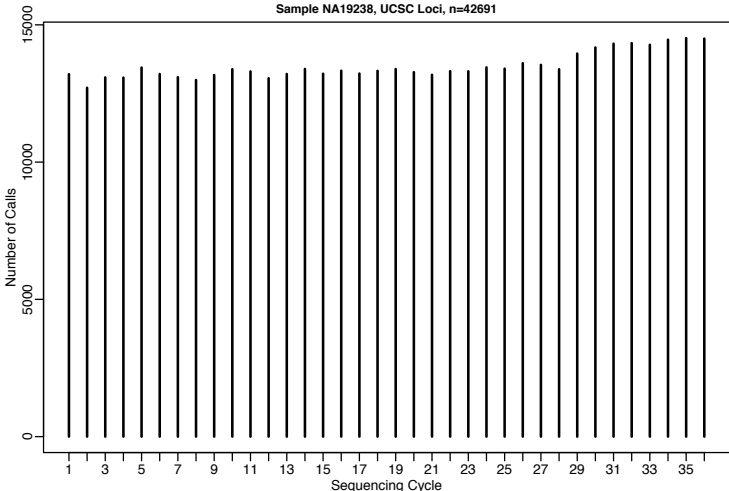
### SNPs in dbSNP

### Novel SNPs

All data



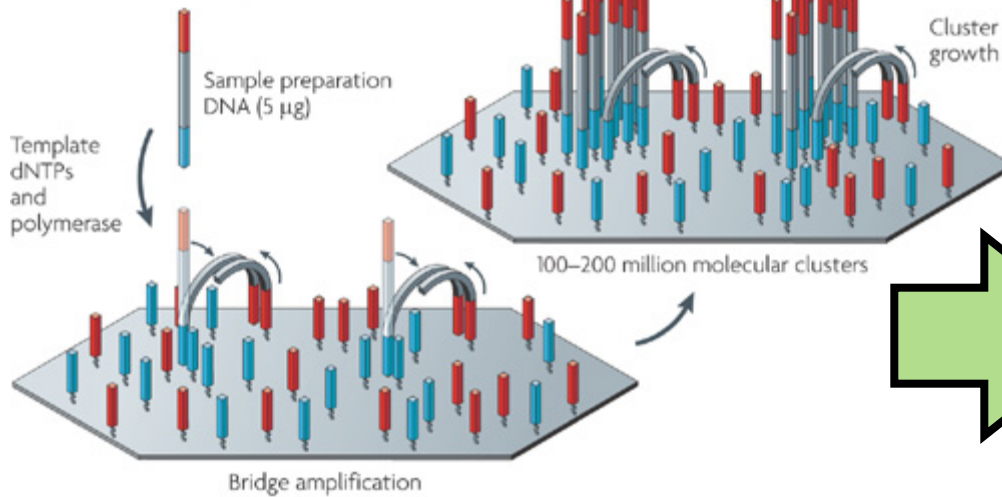
Filtered:  
snpg>=20,  
nreads<=360



Cycle

**What is causing this?**

**Illumina/Solexa  
Solid-phase amplification**  
One DNA molecule per cluster



Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

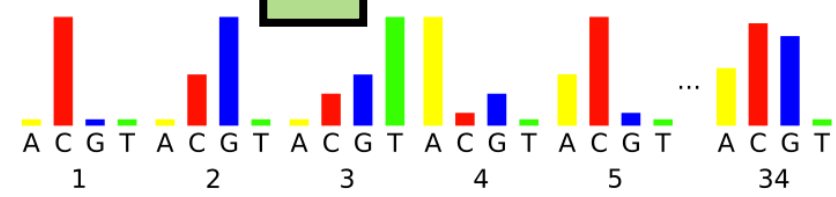
```

@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCGCGGGCCTGNNNNNNNNNNNNNNNN
+
BBBB>A?B@;@BBBBBAA-BA-A%*****
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGCAGCAGNNNNNNNNNNNGNNNN
+
B9B@B<;BAA<@AB9-1>%*****
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCAACTCGTCCAANNNNNNNNNNNNN
+
A=B7&7:>B@:A>?9;<;>74?%*****
@HWI-EAS146:5:1:1:1607#0/1
CTCCTCTCAAGGTCCTCCAGAACGACAGCAANNNNANTNCTNNN
+
BBCCCCCBBBCB7CBC-7>+<=>BCBCB%*****
@HWI-EAS146:5:1:1:1719#0/1
CACGATCTGGGTTATTGTAACCTCCGCCTCANNNGNTNAAGNNN
+
BCC?+<B=?BB5-ABA?B6BBBB4BB?B%*****
@HWI-EAS146:5:1:2:947#0/1
CCGAGGAGAAAGCCATGTTTCAGTTCGAGCGCANNANANCGTANN
+
BBB9@?7A7>AAB@>?B=?@.>B7B?%*****
@HWI-EAS146:5:1:2:563#0/1

```

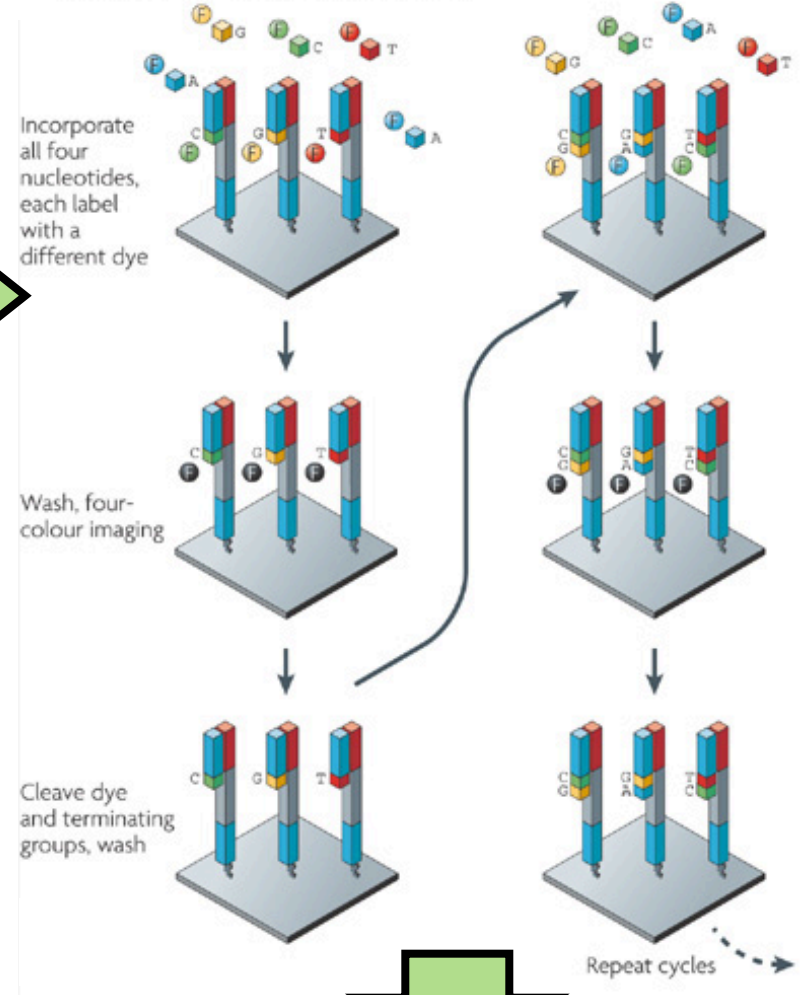
name  
sequence  
quality scores

x 100s of  
millions



Source: Whiteford et al. Swift: primary data analysis for the Illumina Solexa sequencing platform. Bioinformatics. 2009

**Illumina/Solexa — Reversible terminators**



Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

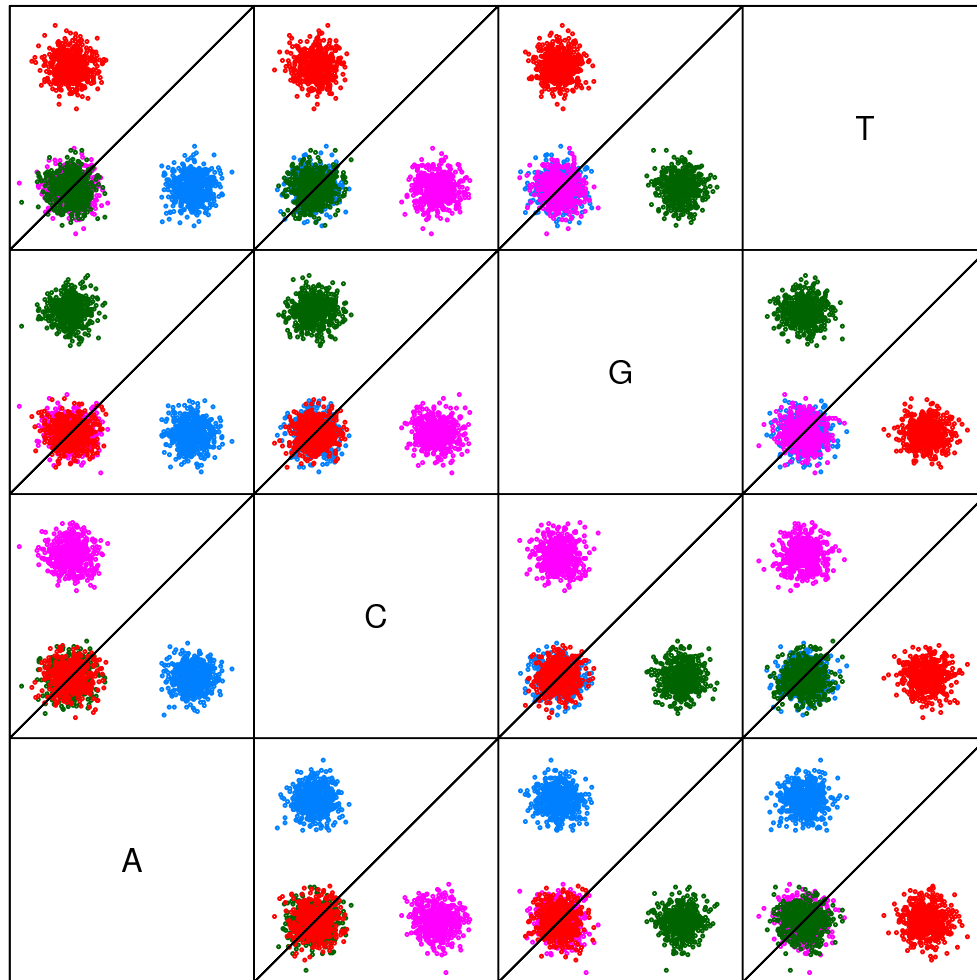
(slide courtesy of Ben Langmead)



# Before Reads There were Intensities

```
> 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
```

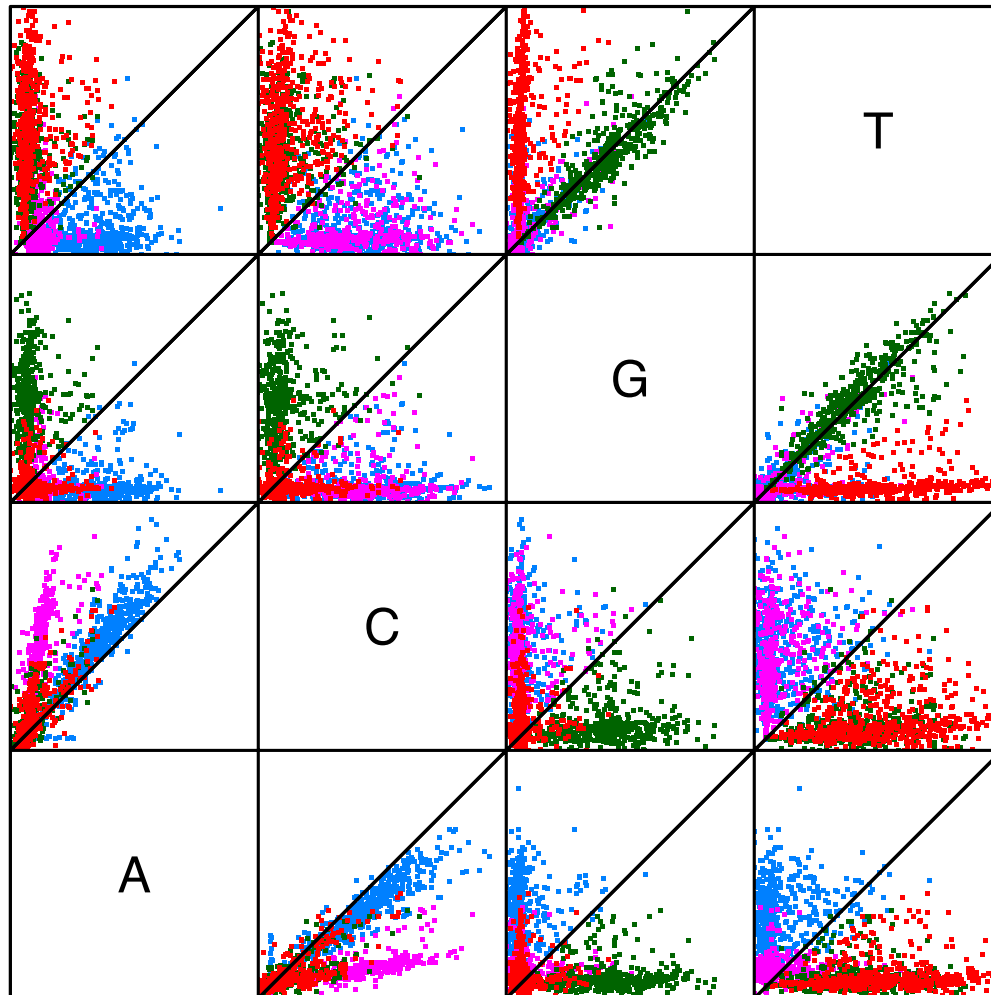
# We Want to See This



Four-channel fluorescence intensity, cycle 1

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

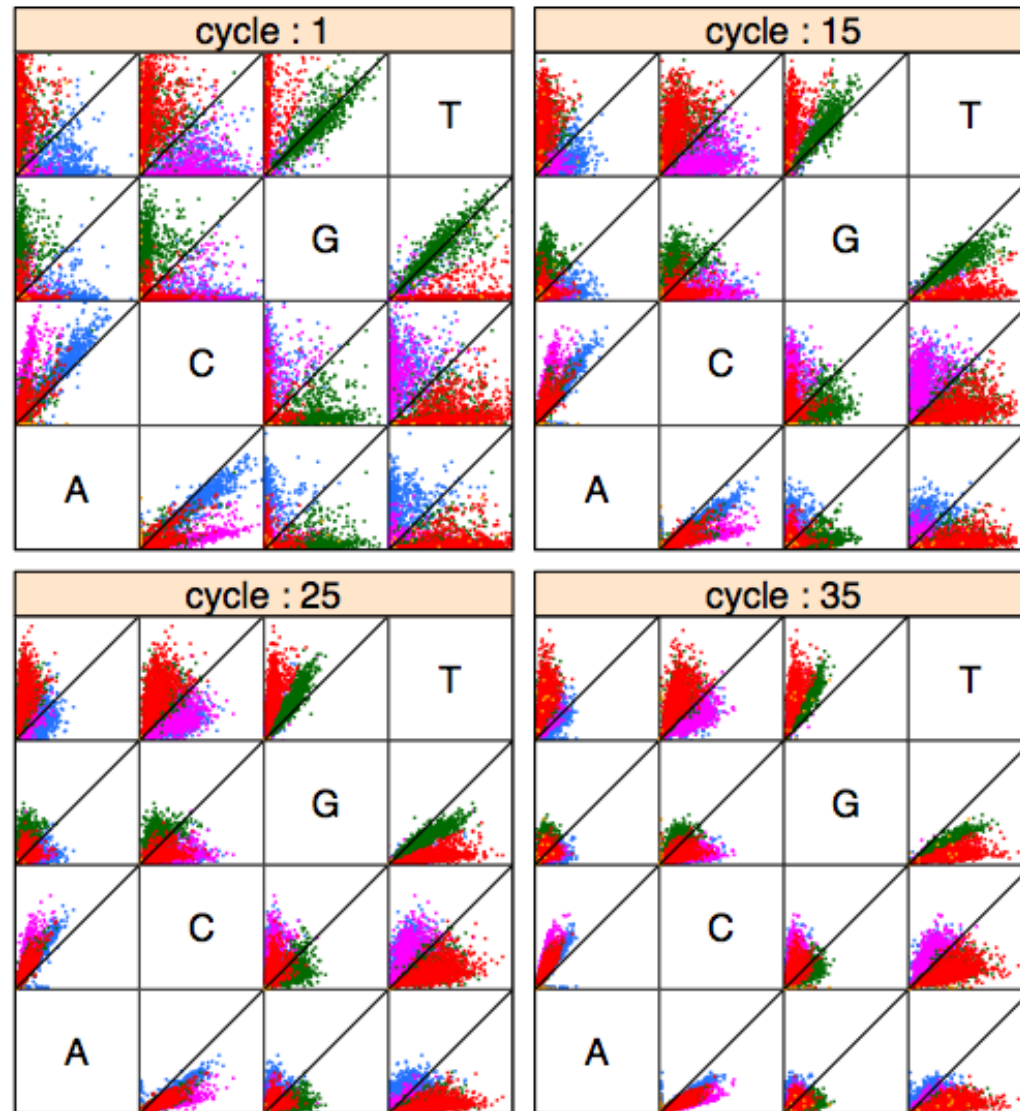
# But See This



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

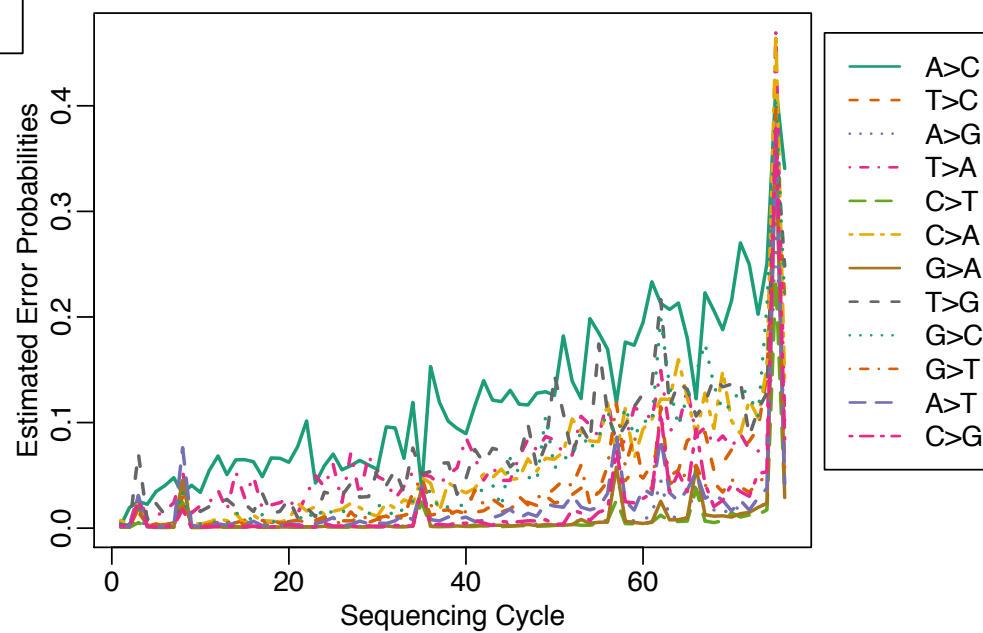
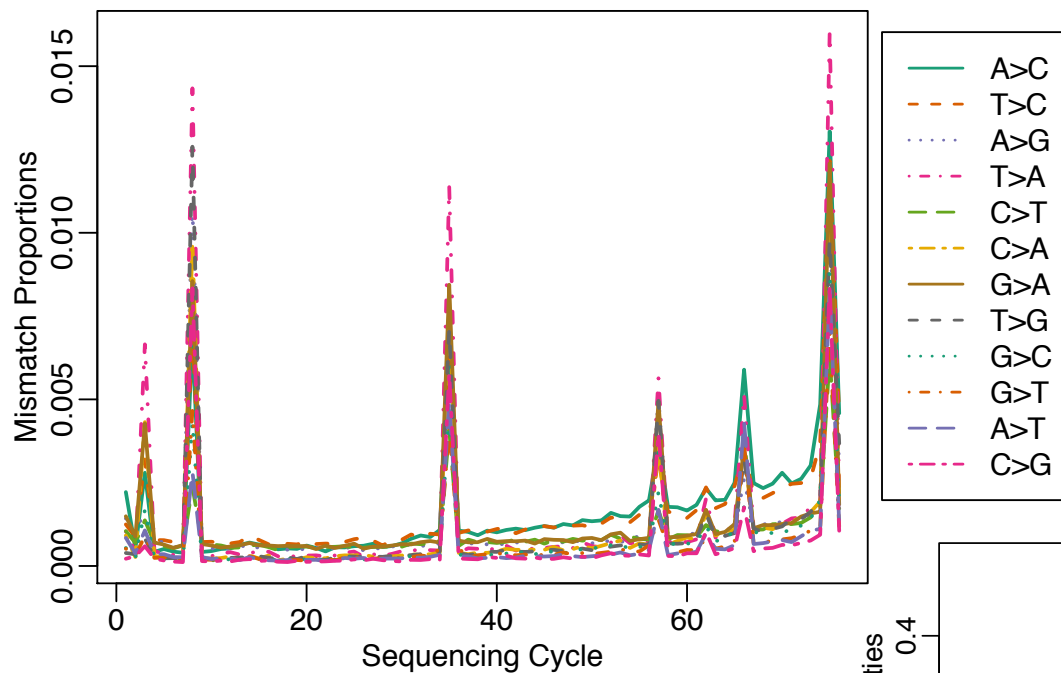
Four-channel fluorescence intensity, cycle 1

# Gets Worse for higher cycles

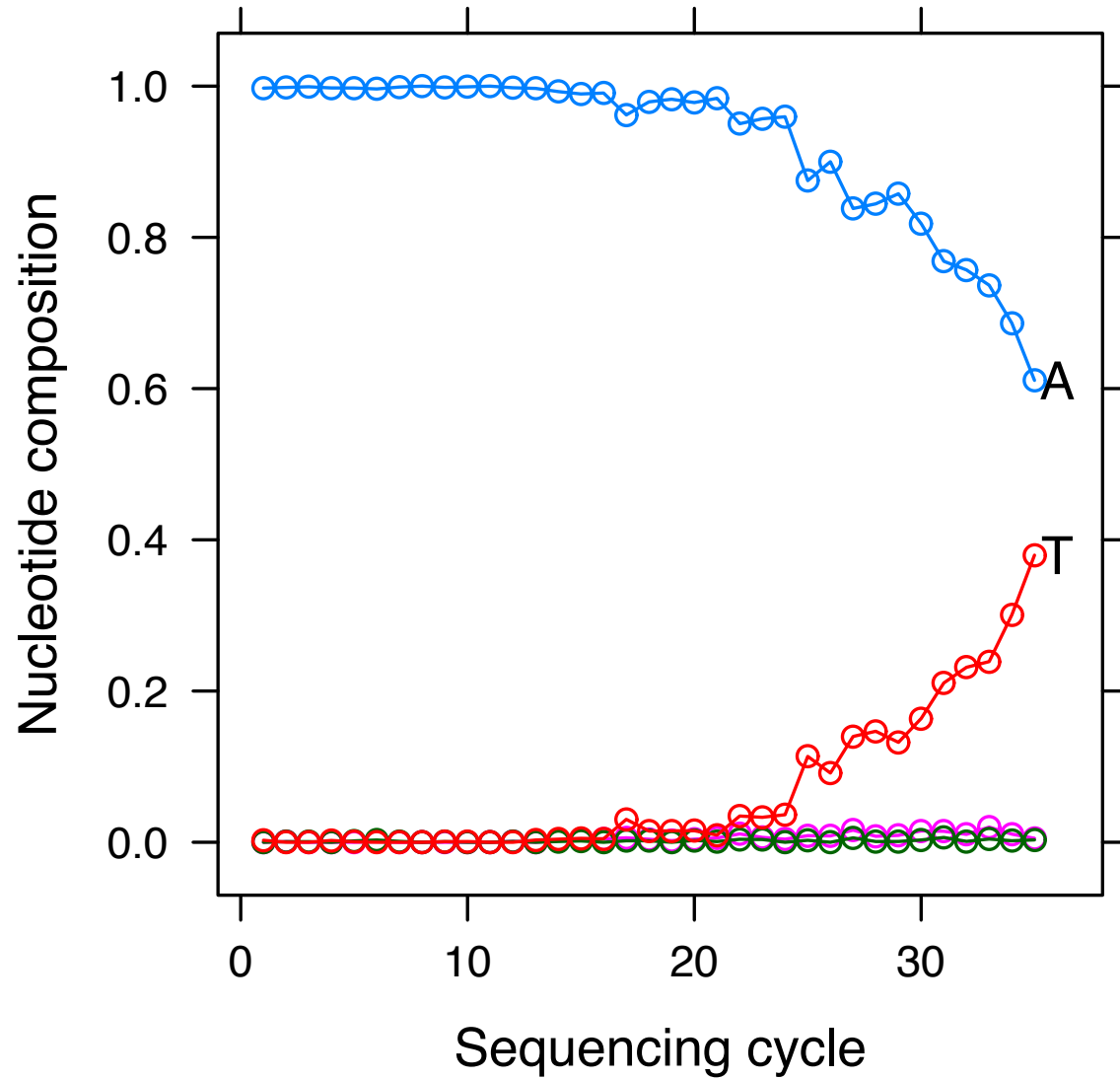


Four-channel fluorescence intensity

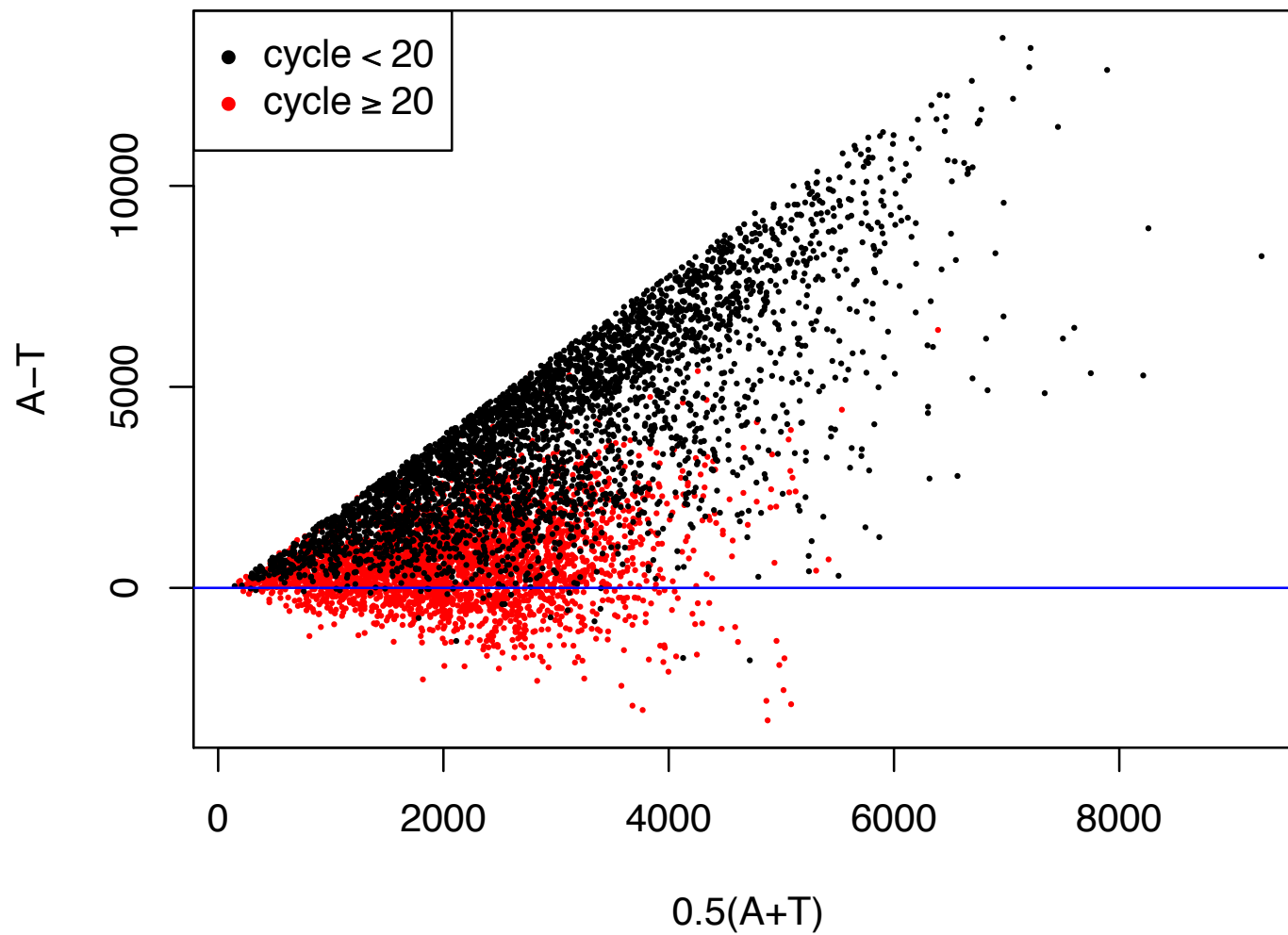
# Error Rate and Reported Quality



# Remember This?



# Bias Explained



# Base Calling

1) Rougemont et al. Probabilistic base calling of Solexa sequencing data. BMC Bioinformatics (2008)

2) Erlich et al. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. Nat Methods (2008)

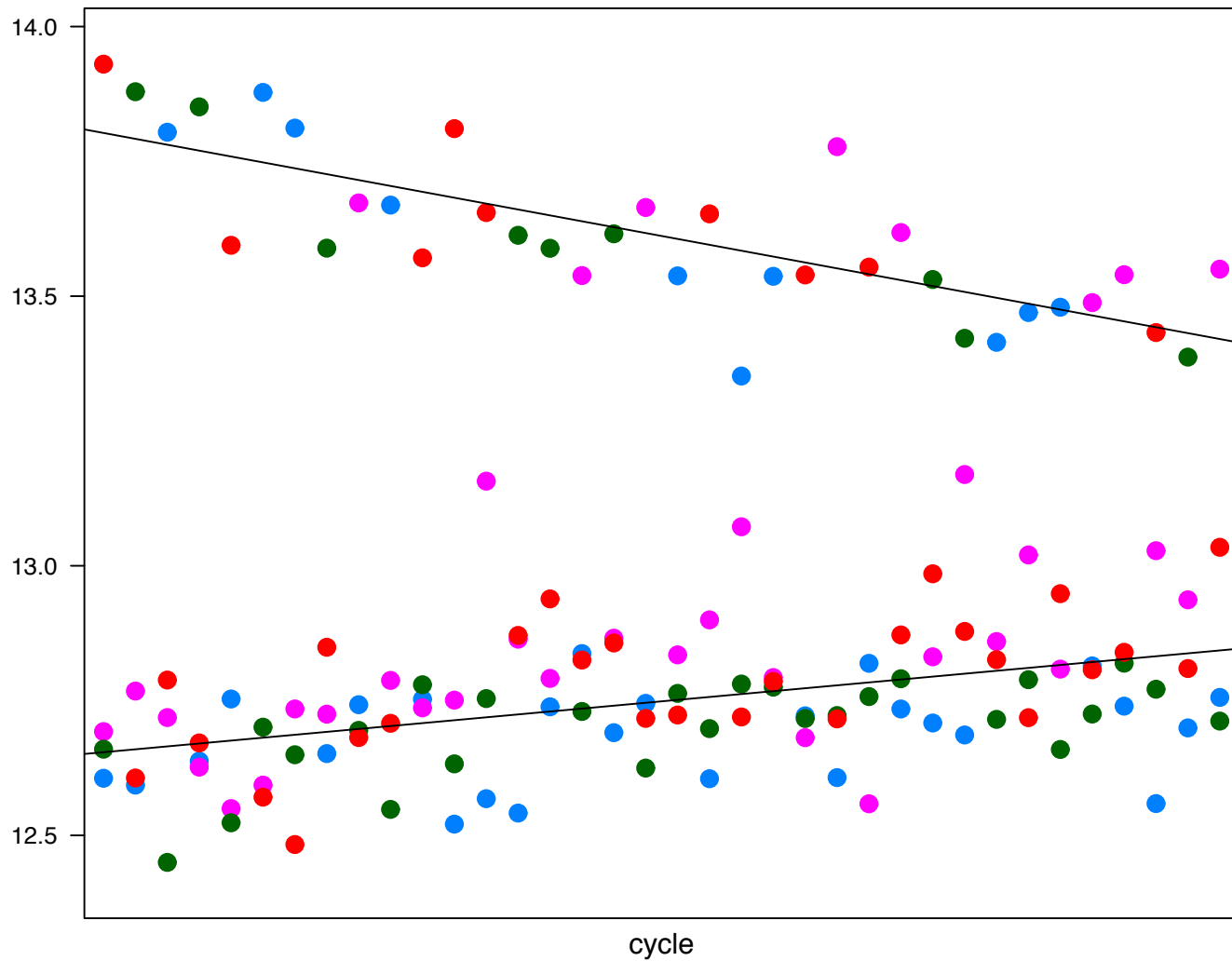
3) Kao et al. BayesCall: A model-based base-calling algorithm for high-throughput short-read sequencing. Genome Res (2009)

4) Corrada Bravo and Irizarry. Model-Based Quality Assessment and Base-Calling for Second-Generation Sequencing Data. Biometrics (2009)

5) Cokus et al. Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature (2009)



# Intensity Model



# Intensity Model

log intensity read  $i$ , cycle  $j$ , channel  $c$

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

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

log intensity read  $i$ , cycle  $j$ , channel  $c$

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

read-specific linear models

# Intensity Model

log intensity read  $i$ , cycle  $j$ , channel  $c$

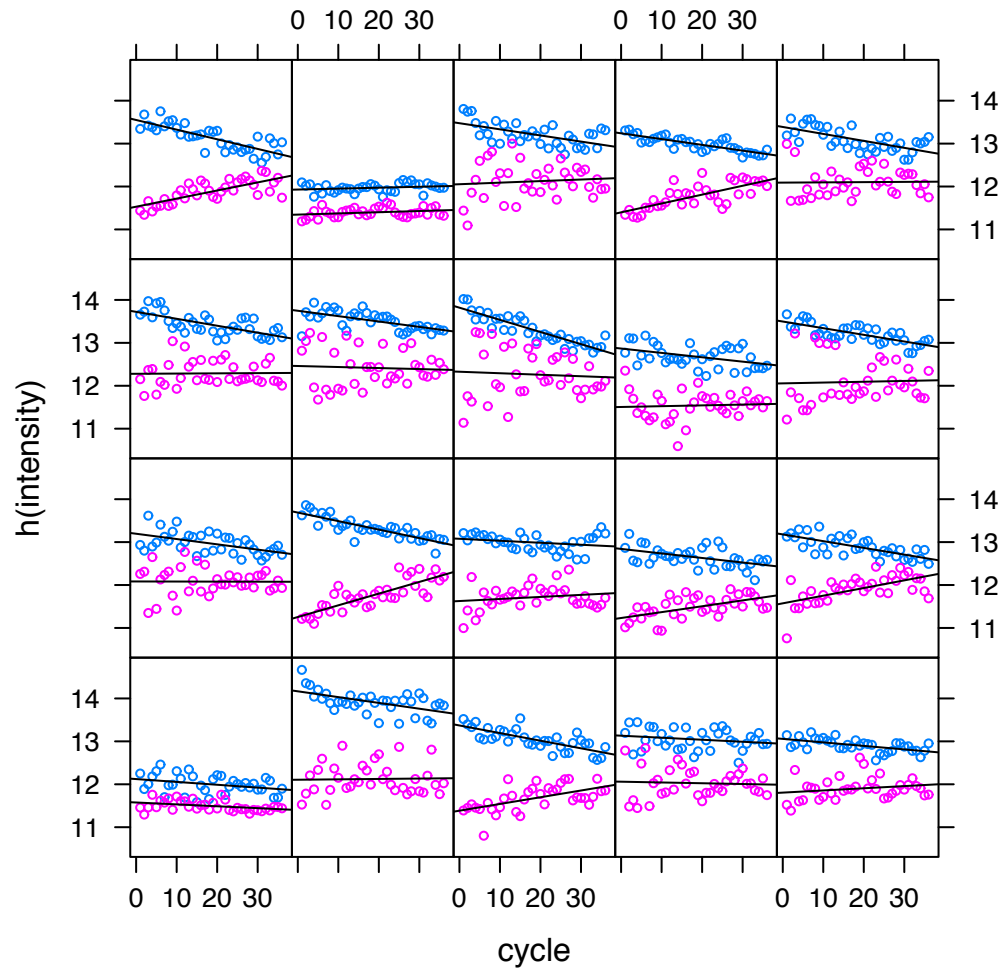
$$u_{ijc} = \Delta_{ijc}(\mu_{cj\alpha} + x_j^T \alpha_i + \underline{\epsilon_{ijc}^\alpha}) + \\ (1 - \Delta_{ijc})(\mu_{cj\beta} + x_j^T \beta_i + \underline{\epsilon_{ijc}^\beta})$$

measurement error

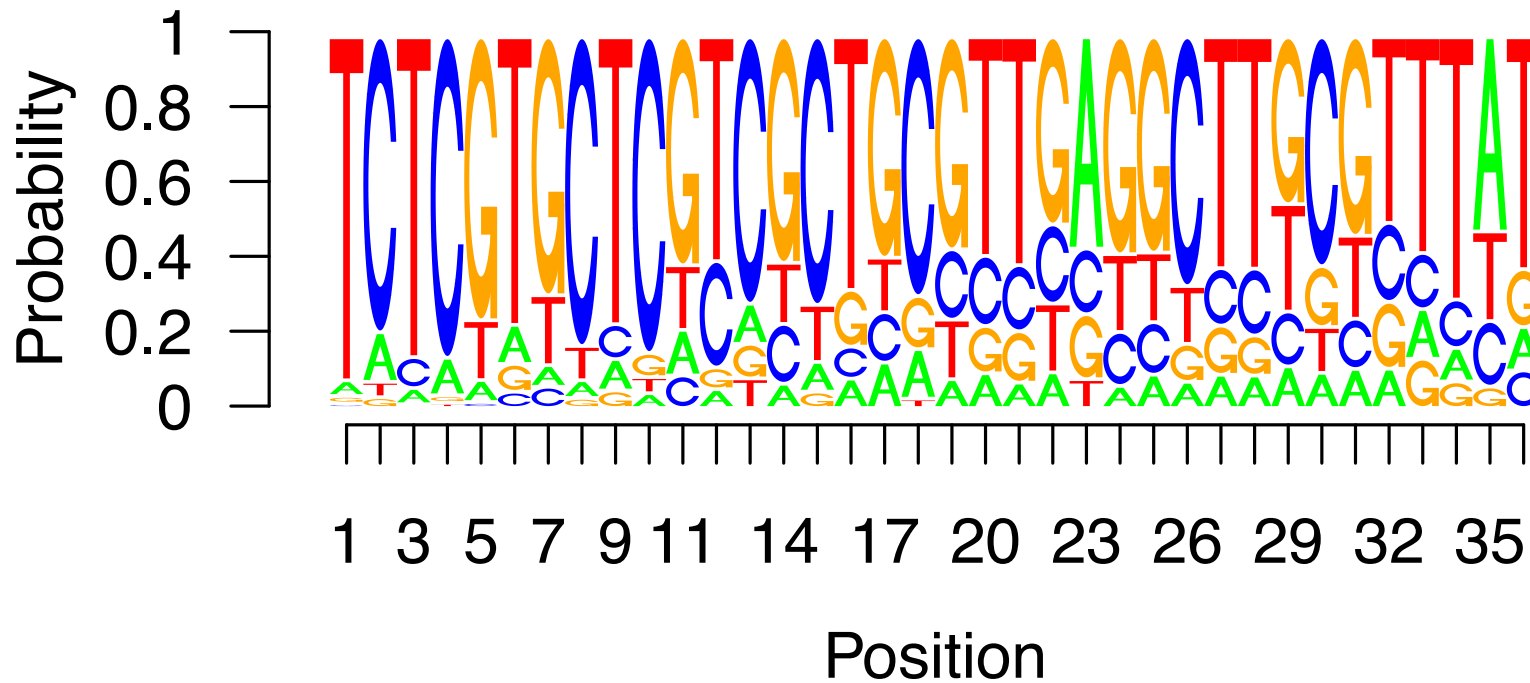
$$\epsilon_{ijc}^\alpha \sim N(0, \sigma_{\alpha i}^2)$$

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

# Read & Cycle Effects

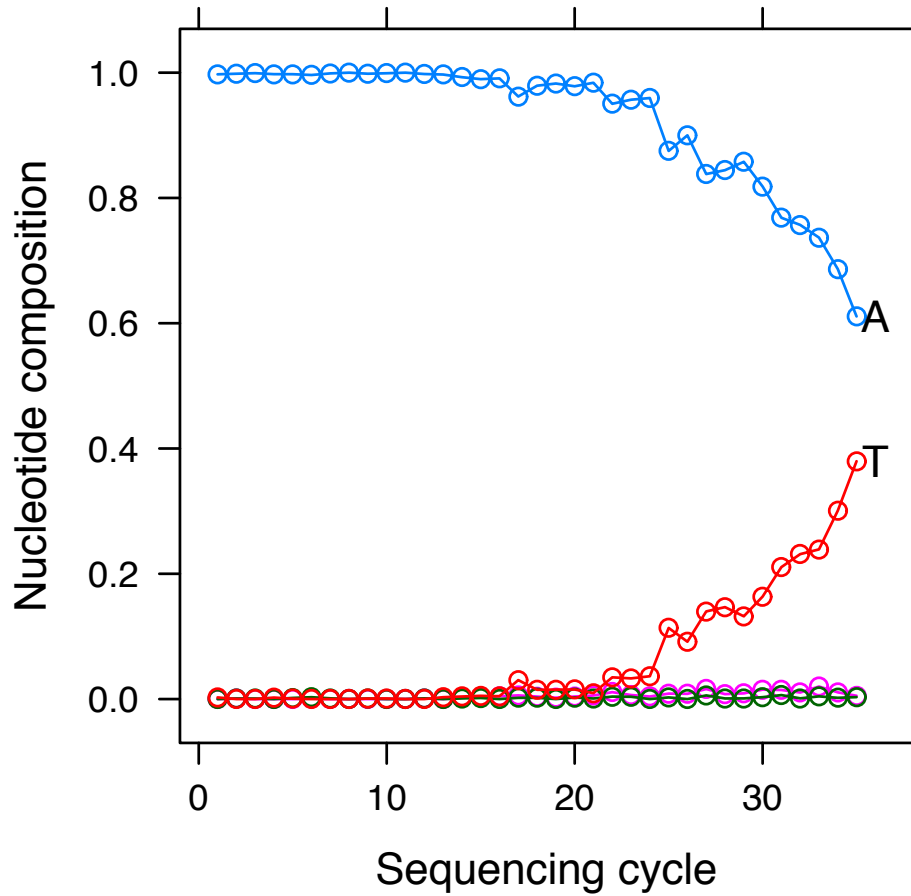


# Base Identity Probability Profiles

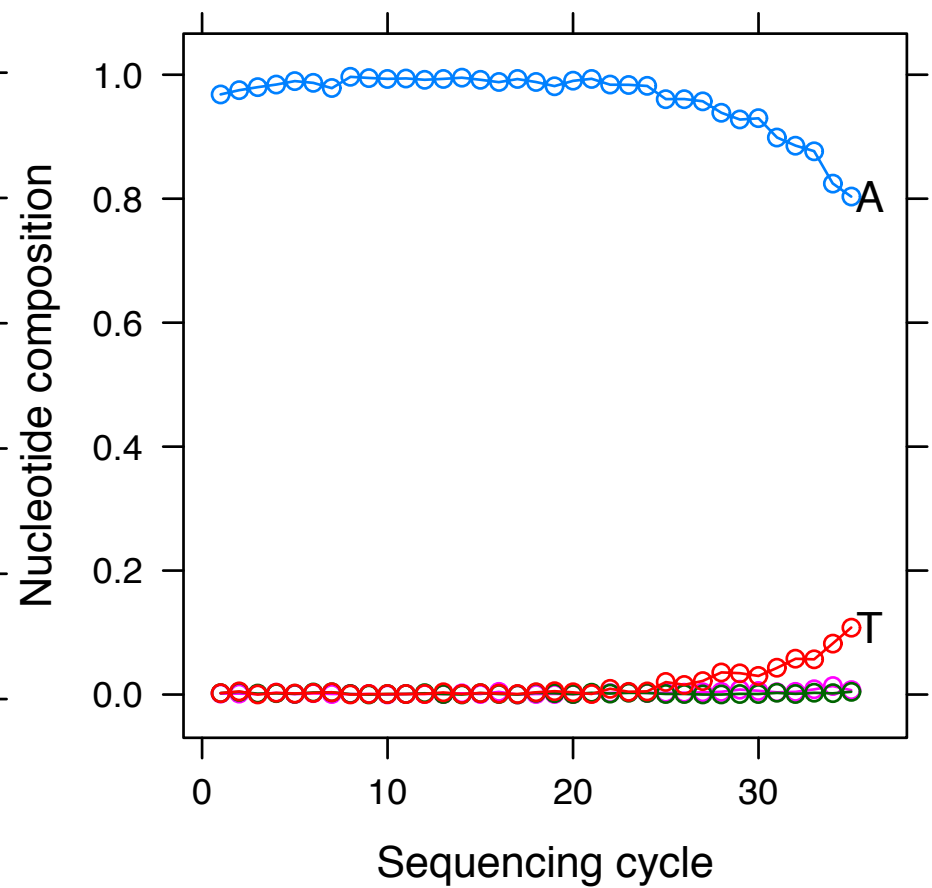


# Before And After

## Solexa (Default)



## Srfim (Statistical Approach)



**The End**