

Statistics applied to second-generation sequencing data

SECOND-GENERATION SEQUENCING

naturenews

[nature news home](#) [news archive](#) [specials](#) [opinion](#) [features](#) [news blog](#) [events](#)

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

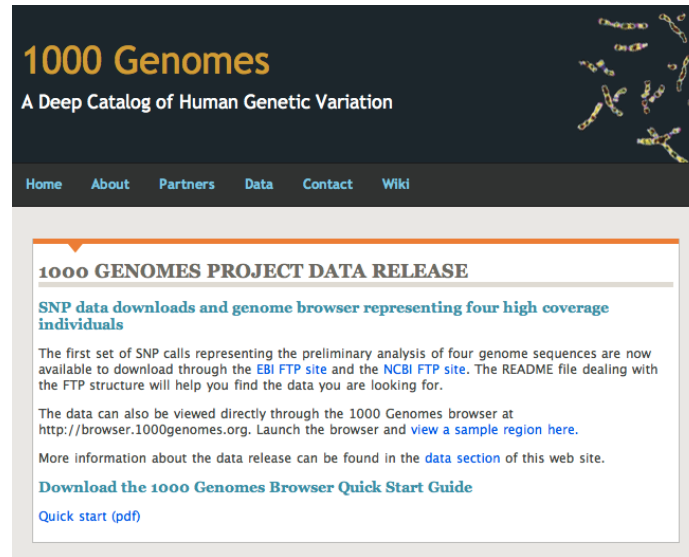


[Send to a Friend](#)

SECOND-GENERATION SEQUENCING

- “Ultra high throughput” DNA sequencing
 - 3 gigabases / week vs.
 - 3 gigabases / 13 years (more or less)

1 0 0 0 GENOMES PROJECT



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](#).

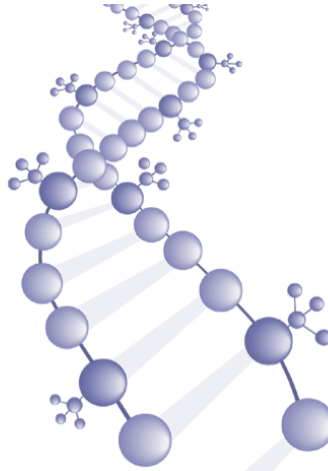
More information about the data release can be found in the [data section](#) of this web site.

Download the 1000 Genomes Browser Quick Start Guide
[Quick start \(pdf\)](#)

Genotyping

HUMAN EPIGENOME PROJECT

HEP
Human
Epigenome
Project



Methylation

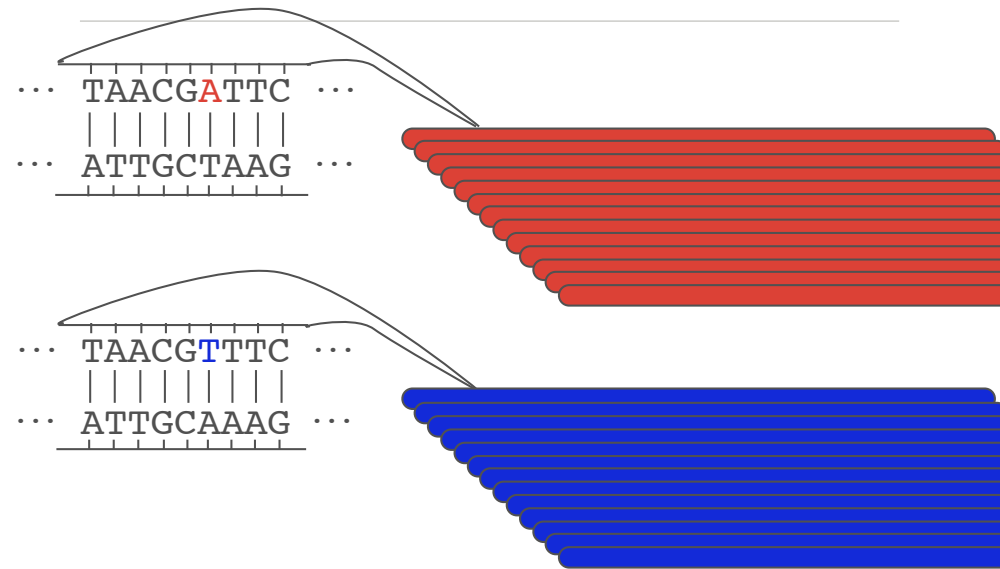


PLATFORMS

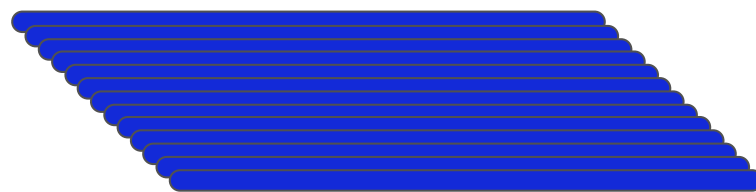
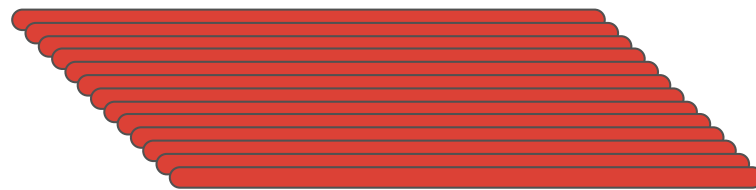


- Millions of short DNA fragments (~36-70 bp) sequenced in parallel

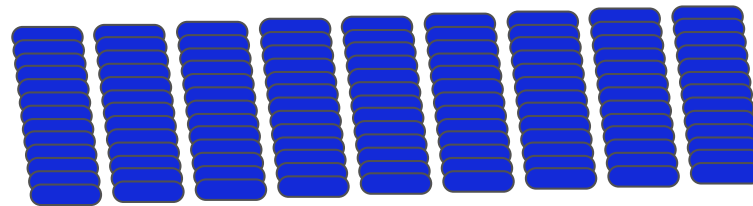
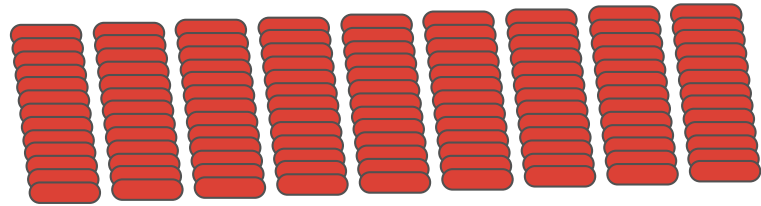
SEC-GEN SEQUENCING



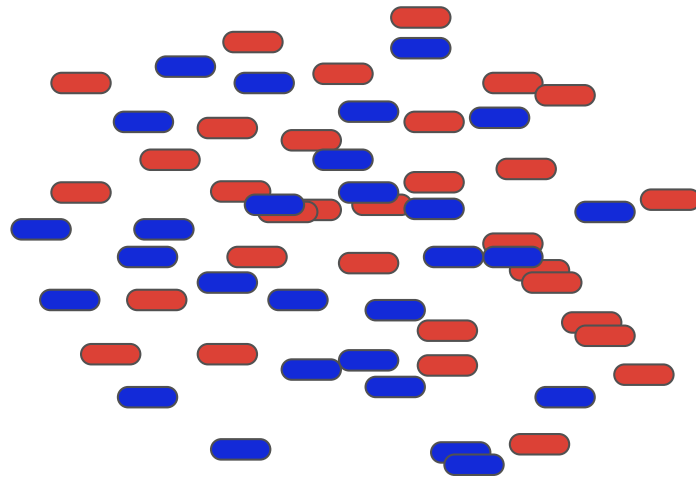
SEC-GEN SEQUENCING



SEC-GEN SEQUENCING



SEC-GEN SEQUENCING

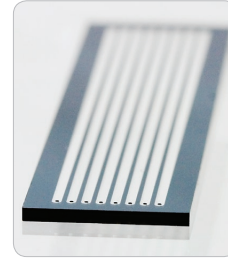


A SET OF SHORT READS

GTTGAGGCTTGCCTTTTGGTACGCTGGACTTTGT
GTACTCGTCGCTGCGTTGAGGCTTGCCTTTTGGT
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTTGCCTTTATGGTACGCTGGACTTTGTAGGATAC
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
GCGTTTATGGTACGCTGGACTTTGTAGGATACCC
GAGGCTTGCCTTTATGGTACGCTGGACTTTGTAGG
GCGTTGAGGCTTGCCTTTATGGTACGCTGGATTTT
CGTTTATGGTACGCTGGACTTTGTAGGATACCCTC
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
GTTTATGGTACGCTGGACTTTGTAGGATACCCTCG
TCTCGTCGCTCGCTGCGTTGAGGCTTGCCTTTA
TGCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTA
GCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTAC
TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
TCGTGCTCGTCGCTGCGTTGAGGCTTGCCTTTTG
CGTCGCTGCGTTGAGGCTTGCCTTTATGGTACGCT
GTTGAGGCTTGCCTTTATGGTACGCTGGGCTTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC

ILLUMINA/SOLEXA

- Eight lanes
- 100 tiles/lane
- ~200K fragments per tile
- ~160M short reads (~50-70 bp) per run



MATCHING

GTTGAGGCTTGCCTTTTGGTACGCTGGACTTTGT
GTACTCGTCGCTGCGTTGAGGCTTGCCTTTTGGT
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTTGCCTTTATGGTACGCTGGACTTTGTAGGATAC
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
GCGTTTATGGTACGCTGGACTTTGTAGGATACCCT
GAGGCTTGCCTTTATGGTACGCTGGACTTTGTAGG
GCGTTGAGGCTTGCCTTTATGGTACGCTGGATTTT
CGTTTATGGTACGCTGGACTTTGTAGGATACCCTC
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
GTTTATGGTACGCTGGACTTTGTAGGATACCCTCG
TCTCGTGCTCGTCGCTGCGTTGAGGCTTGCCTTTA
TGCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTA
GCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTAC
TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
TCGTGCTCGTCGCTGCGTTGAGGCTTGCCTTTTGG
CGTCGCTGCGTTGAGGCTTGCCTTTATGGTACGCT
GTTGAGGCTTGCCTTTATGGTACGCTGGGCTTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTCTCGTGCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTTC

USE BLAST?

- Matching 10,000,000 32 bp reads:

BLAST takes more than 6 months

BLAT takes 2 months

MAQ takes 1 day and half

Bowtie takes 17 minutes

MATCHING

GTGAGGCTTGCCTTTTGGTACGCTGGACTTTGT
GTACTCGTCGCTGCGTTGAGGCTTGCCTTTTGGT
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTTGCCTTTATGGTACGCTGGACTTTGTAGGATAC

Bowtie

An ultrafast memory-efficient short read aligner



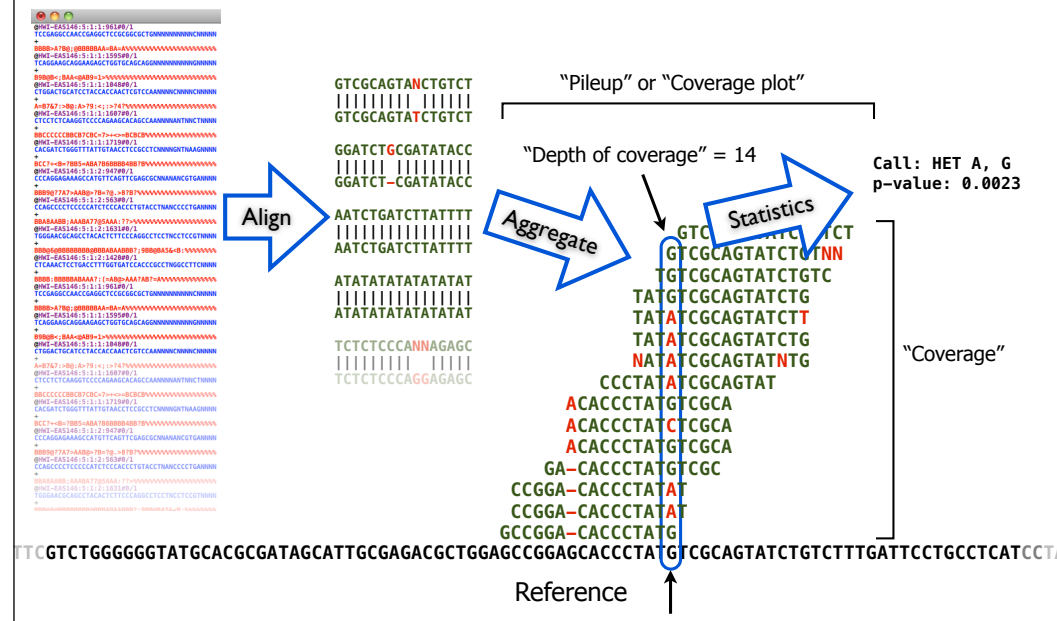
Bowtie is an ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome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



TGCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTA
GCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTAC
TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
TCGTGCTCGTCGCTGCGTTGAGGCTTGCCTTTTGG
CGTCGCTGCGTTGAGGCTTGCCTTTATGGTACGCT
GTTGAGGCTTGCCTTTATGGTACGCTGGGCTTTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTCTCGTGCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTTC

Workflow examples

Variant detection



The diagram illustrates the process of identifying differentially expressed genes using RNA-seq data. It shows two samples, Sample A and Sample B, being aligned, aggregated, and then analyzed for differential expression.

Sample A: The initial data shows multiple reads aligned to a reference sequence. The reads are then aggregated into a single sequence, highlighting the presence of Gene 1. The aggregated sequence for Gene 1 is shown in orange and purple, indicating the presence of the gene in Sample A.

Sample B: Similar to Sample A, the reads are aligned and aggregated. The aggregated sequence for Gene 1 is shown in green, indicating the presence of the gene in Sample B.

Analysis: The aggregated sequences for Gene 1 from both samples are compared. The analysis identifies Gene 1 as differentially expressed, with a p-value of 0.0012.

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

ChIP-seq

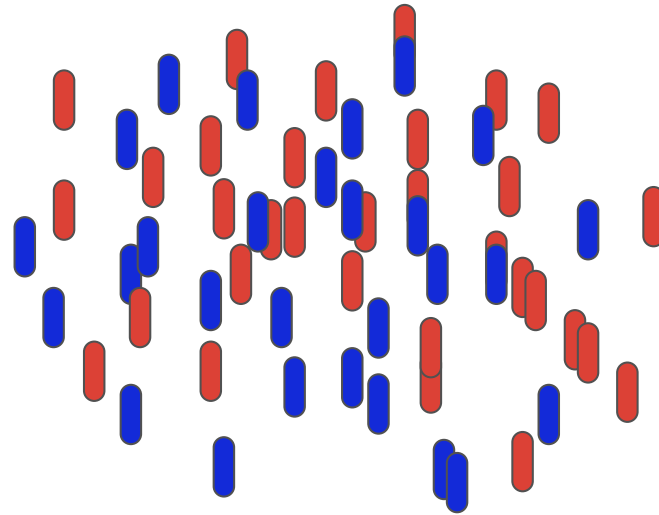


THE END

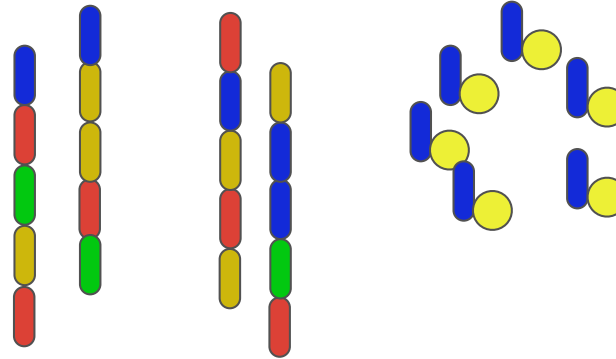
SNPs

GTTGAGGCTTGCCTTTTGGTACGCTGGACTTTGT
GTACTCGTCGCTGCGTTGAGGCTTGCCTTTTGGT
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTTGCCTTTATGGTACGCTGGACTTTGTAGGATAC
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
GCGTTTATGGTACGCTGGACTTTGTAGGATACCCT
GAGGCTTGCCTTTATGGTACGCTGGACTTTGTAGG
GCGTTGAGGCTTGCCTTTATGGTACGCTGGATTTT
CGTTTATGGTACGCTGGACTTTGTAGGATACCCTC
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
GTTTATGGTACGCTGGACTTTGTAGGATACCCTCG
TCTCGTGCCTCGCTGCGTTGAGGCTTGCCTTTA
TGCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTA
GCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTAC
TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
TCGTGCTCGTCGCTGCGTTGAGGCTTGCCTTTT
CGTCGCTGCGTTGAGGCTTGCCTTTATGGTACGCT
GTTGAGGCTTGCCTTTATGGTACGCTGGGCTTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTCTCGTGCCTCGCTGCGTTGAGGCTTGCCTTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTC

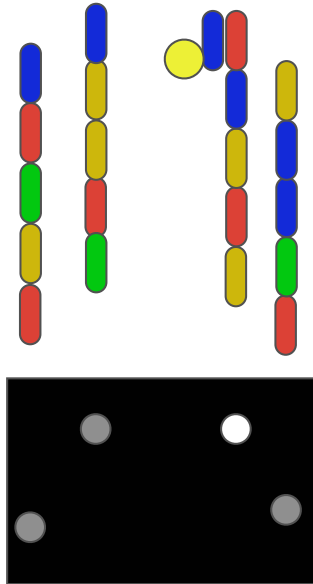
SEC-GEN SEQUENCING



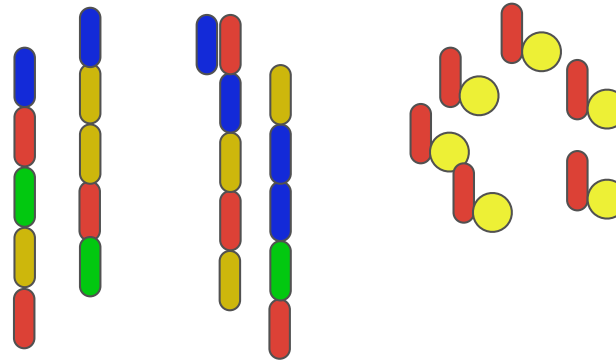
SEC-GEN SEQUENCING



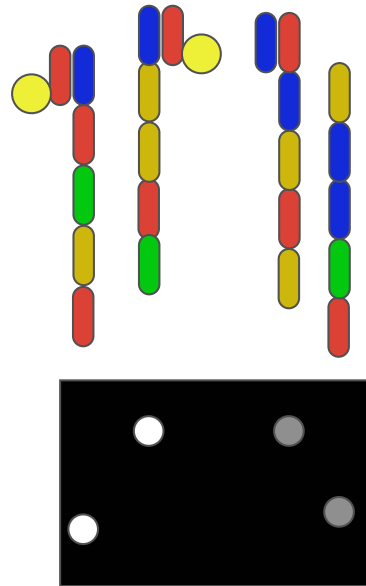
SEC-GEN SEQUENCING



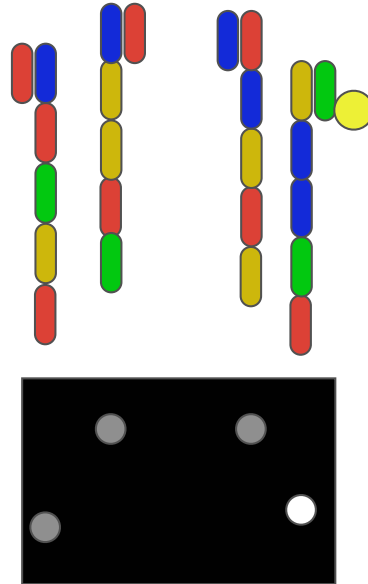
SEC-GEN SEQUENCING



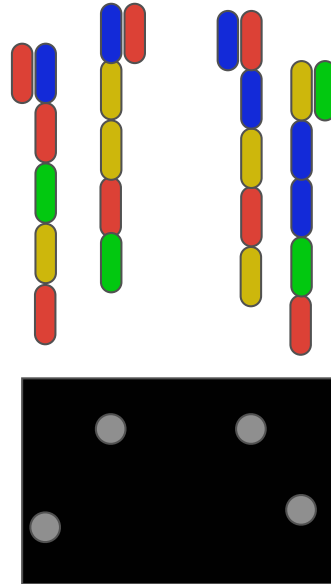
SEC-GEN SEQUENCING



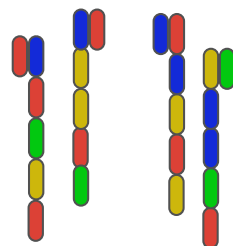
SEC-GEN SEQUENCING



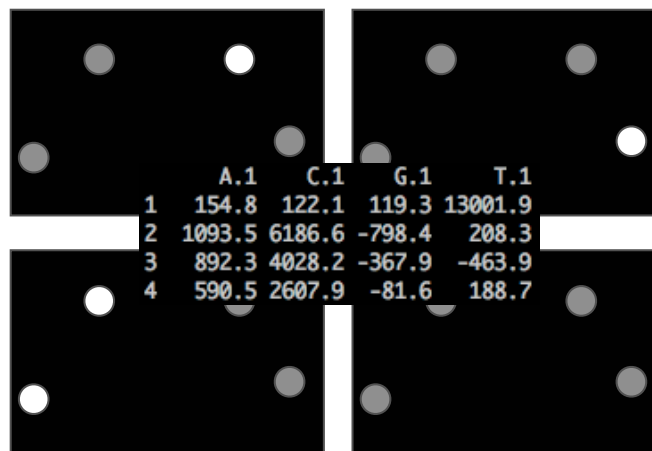
SEC-GEN SEQUENCING



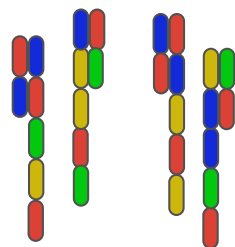
SEC-GEN SEQUENCING



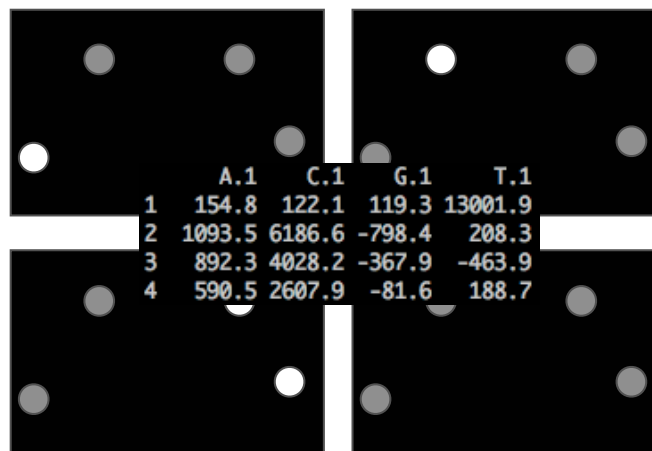
First Cycle



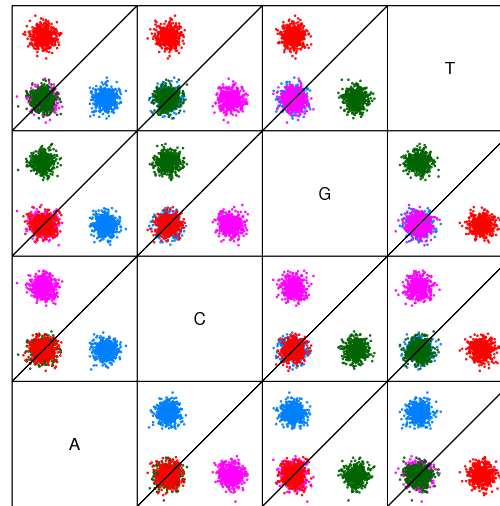
SEC-GEN SEQUENCING



Second Cycle



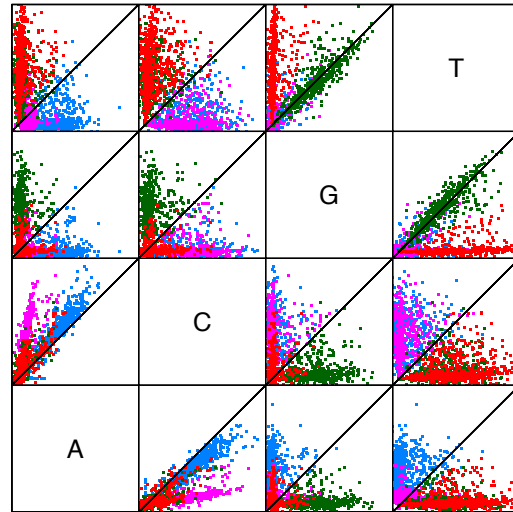
A THOUGHT EXPERIMENT



Four-channel fluorescence intensity, cycle 1

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

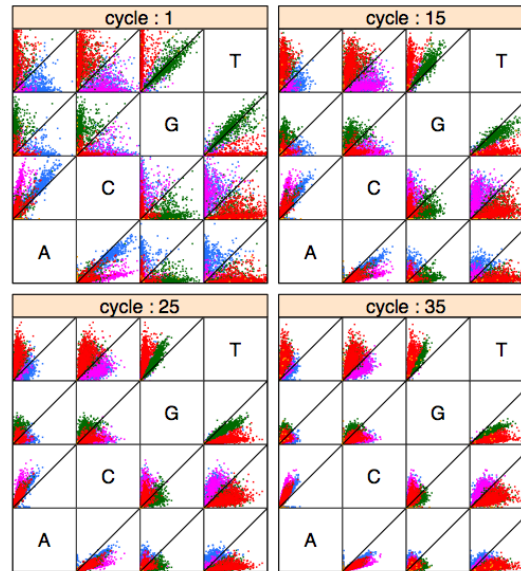
FLUORESCENCE INTENSITY



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

Four-channel fluorescence intensity, cycle 1

FLUORESCENCE INTENSITY



CHALLENGES

- Base-calling is the result of a complicated procedure on noisy measurements
 - Base-calls are not made with the same certainty
- Statistical: How to model this uncertainty?
- Computational: Can we manage this model at sec-gen data scale?

INTENSITY MODEL

- log intensity read i , cycle j , channel c

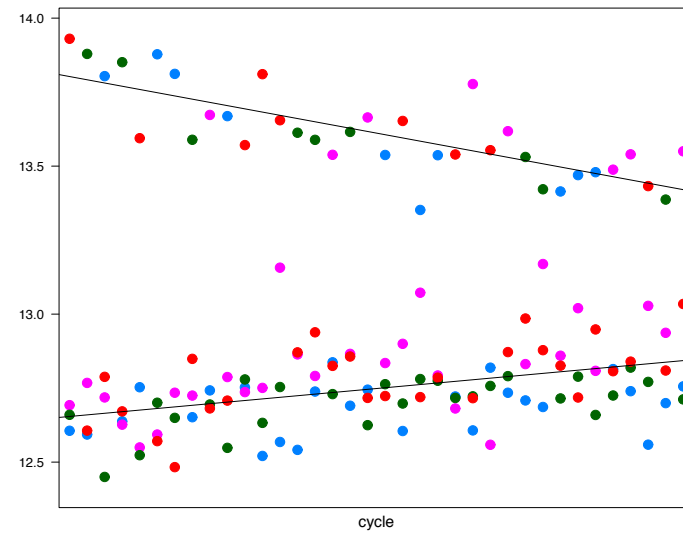
$$u_{ijc} = \frac{\Delta_{ijc}}{(1 - \Delta_{ijc})} (\mu_{cj\alpha} + x_j^T \alpha_i + \epsilon_{ijc}^\alpha) +$$

$$(1 - \Delta_{ijc}) (\mu_{cj\beta} + x_j^T \beta_i + \epsilon_{ijc}^\beta)$$

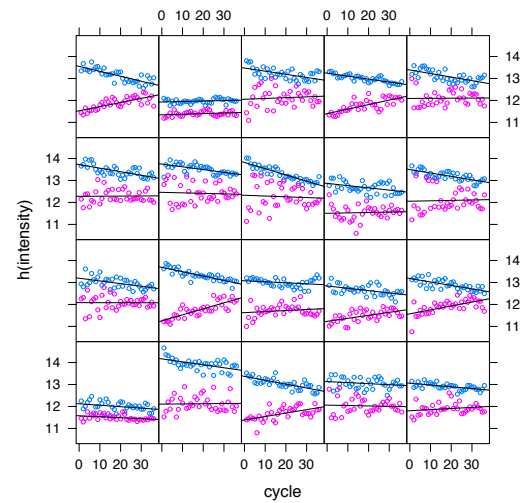
- indicators of nucleotide identity, read i , pos. j

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

INTENSITY MODEL (READ-CYCLE EFFECT)



READ & CYCLE EFFECTS



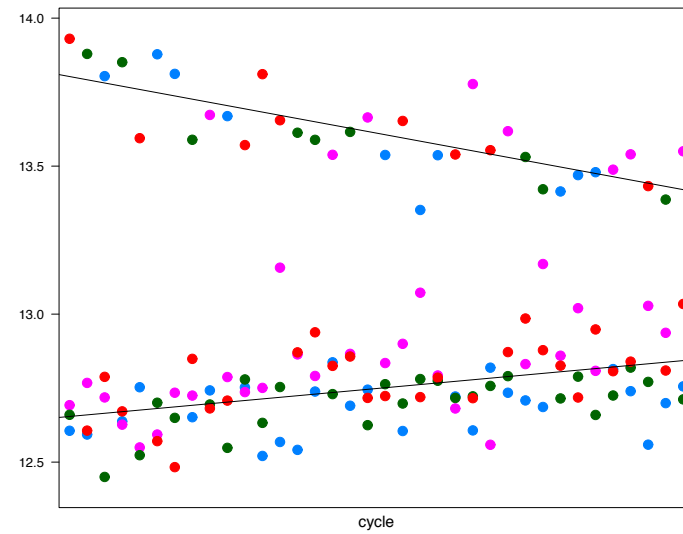
INTENSITY MODEL

- log intensity read i , cycle j , channel c

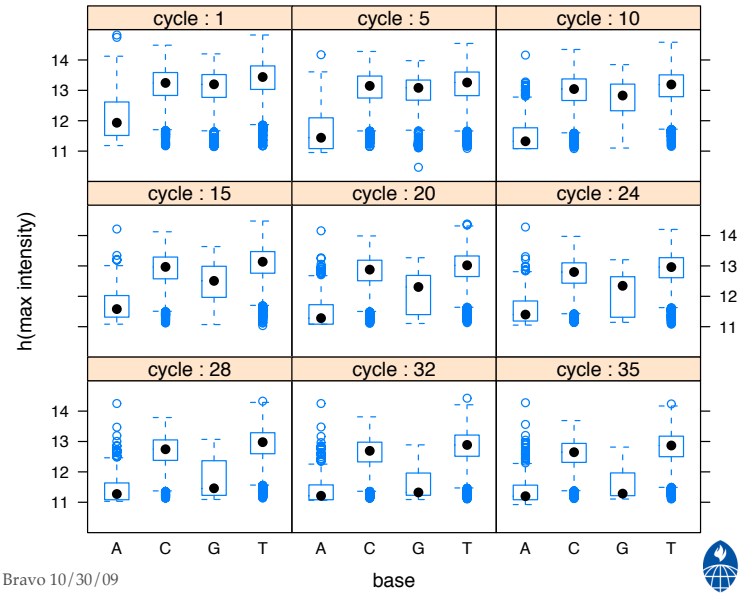
$$u_{ijc} = \Delta_{ijc}(\mu_{cj\alpha} + \underline{x_j^T \alpha_i} + \epsilon_{ijc}^\alpha) + (1 - \Delta_{ijc})(\mu_{cj\beta} + \underline{x_j^T \beta_i} + \epsilon_{ijc}^\beta)$$

- read-specific linear models

INTENSITY MODEL (READ-CYCLE EFFECT)



INTENSITY MODEL (BASE-CYCLE EFFECT)



INTENSITY MODEL

- log intensity read i , cycle j , channel c

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

- base-cycle effects

INTENSITY MODEL

- log intensity read i , cycle j , channel c

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

- random noise

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

INTENSITY MODEL

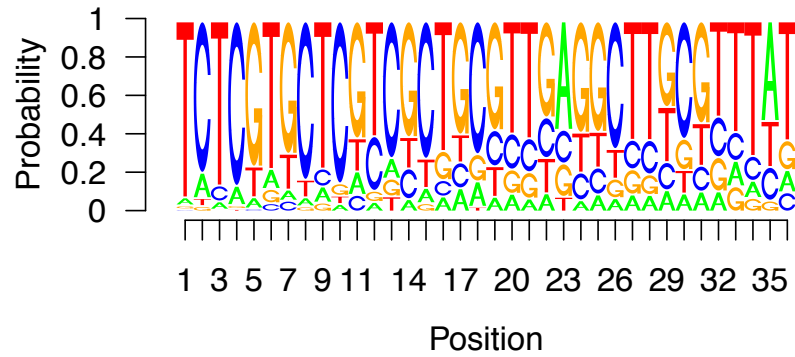
- log intensity read i , cycle j , channel c

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

- estimate parameters with EM algorithm
- which also estimates

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

BASE IDENTITY PROBABILITY PROFILES



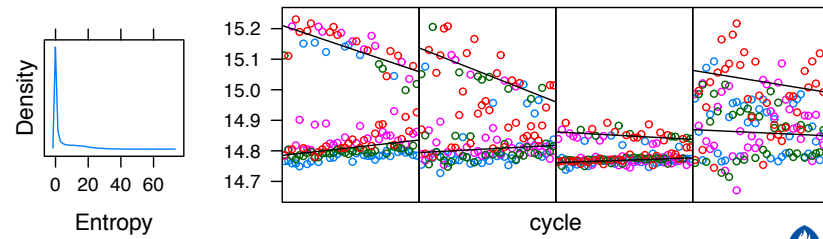
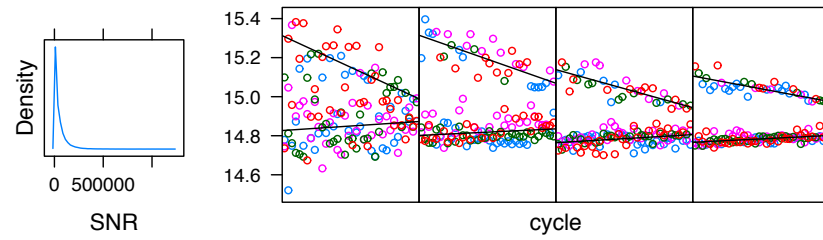
QUALITY METRICS

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

$$H_{ij} = - \sum_c z_{ijc} \log_2 z_{ijc}$$

$$H_i = - \sum_{cj} z_{ijc} \log_2 z_{ijc}$$

QUALITY METRICS



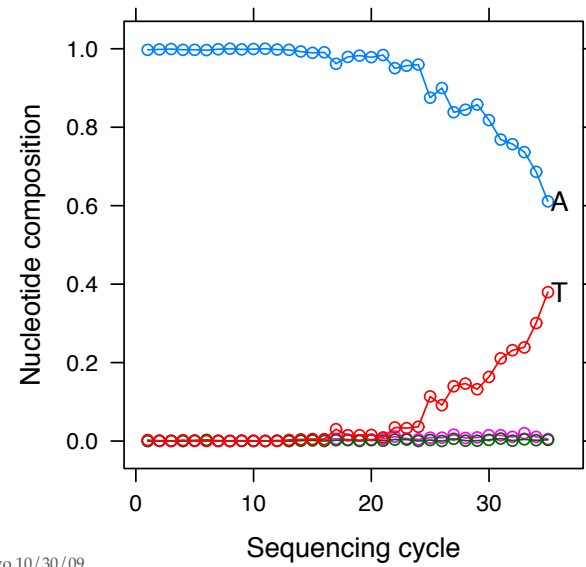
SNPs

GTTGAGGCTTGCCTTTTGGTACGCTGGACTTTGT
GTACTCGTCGCTGCGTTGAGGCTTGCCTTTTGGT
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTTGCCTTTATGGTACGCTGGACTTTGTAGGATAC
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
GCGTTTATGGTACGCTGGACTTTGTAGGATACCCT
GAGGCTTGCCTTTATGGTACGCTGGACTTTGTAGG
GCGTTGAGGCTTGCCTTTATGGTACGCTGGATTTT
CGTTTATGGTACGCTGGACTTTGTAGGATACCCTC
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
GTTTATGGTACGCTGGACTTTGTAGGATACCCTCG
TCTCGTGCCTCGCTGCGTTGAGGCTTGCCTTTA
TGCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTA
GCTCGTCGCTGCGTTGAGGCTTGCCTTTATGGTAC
TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
TCGTGCTCGTCGCTGCGTTGAGGCTTGCCTTTTGTG
CGTCGCTGCGTTGAGGCTTGCCTTTATGGTACGCT
GTTGAGGCTTGCCTTTATGGTACGCTGGGCTTTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTCTCGTGCCTCGCTGCGTTGAGGCTTGCCTTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTTC

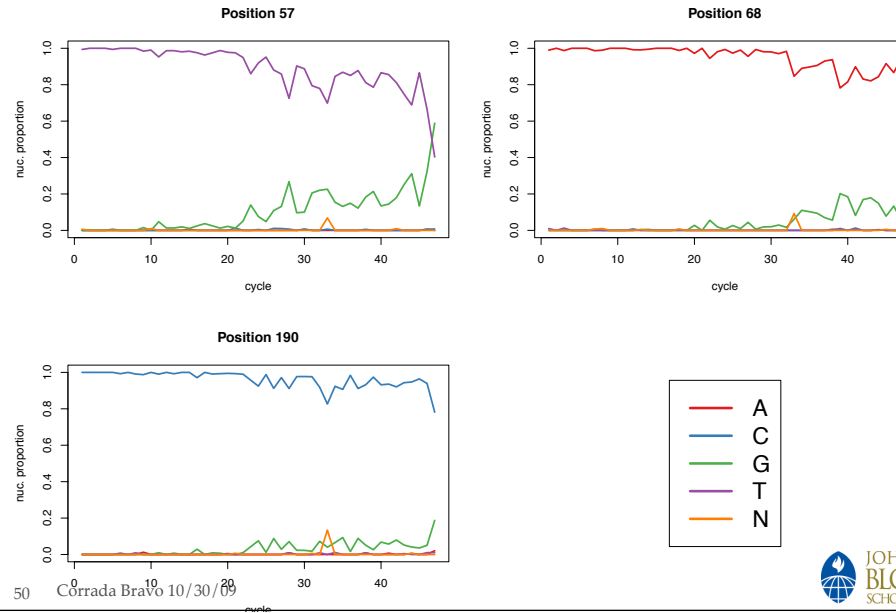
SNPs

```
TCTCGTGCCTCGCTGCGTTGAGGCTTGCCTTTA
TCGTGCTCGCTCGCTGCGTTGAGGCTTGCCTTTTG
GTACTCGTGCCTGCGTTGAGGCTTGCCTTTTGGT
TGCTCGTGCCTGCGTTGAGGCTTGCCTTTATGGTA
GCTCGTGCCTGCGTTGAGGCTTGCCTTTATGGTAC
CGTCGCTGCGTTGAGGCTTGCCTTTATGGTACGCT
GCGTTGAGGCTTGCCTTTATGGTACGCTGGATTTT
GTTGAGGCTTGCCTTTTGGTACGCTGGACTTTGT
GTTGAGGCTTGCCTTTATGGTACGCTGGGCTTTT
GAGGCTTGCCTTTATGGTACGCTGGACTTTGTAGG
CTTGCGTTTATGGTACGCTGGACTTTGTAGGATAC
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
GCGTTTATGGTACGCTGGACTTTGTAGGATACCTT
CGTTTATGGTACGCTGGACTTTGTAGGATACCTT
GTTTATGGTACGCTGGACTTTGTAGGATACCTT
TATGGTACGCTGGACTTTGTAGGATACCTTGCCTT
ATGGTACGCTGGACTTTGTAGGATACCTTGCCTTT
ATGGTACGCTGGACTTTGTAGGATACCTTGCCTTT
CTCTCGTGCCTCGCTGCGTTGAGGCTTGCCTTTATGGTACGCTGGACTTTGTAGGATACCTTGCCTTTC
```

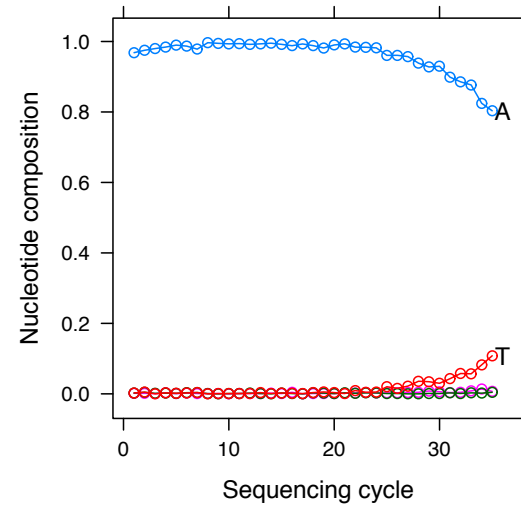
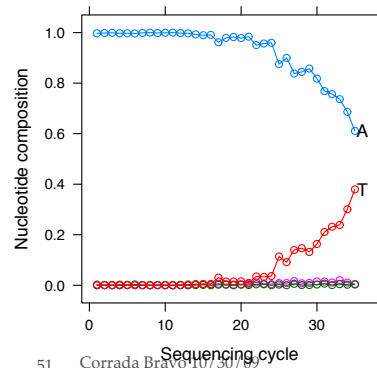
SNPs



SNPs



SNPs



From reads to evidence

```
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGCGCTGNNNNNNNNNNNNNNNN
+
BBB>A7B;@BBBBA=BA=AAAAAAAAAAAAAAAAAAAAAAAA
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGACGAGNNNNNNNNNNNNNNNN
+
B9BQ<;BAA=AB9=1>AAAAAAAAAAAAAAAAAAAAAAAA
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCACTGTCCAANNNNNNNNNNNNNN
+
A=B7&7>B;A=79<;>747AAAAAAAAAAAAAAAAAAAAAAAA
@HWI-EAS146:5:1:1:1687#0/1
CTCTCTCAAGGTCCCAGAAGCAGCAGCAANNNNNNNNNNNNNN
+
BBCCCCCBBB7CB=7>+<=>=BCBCBAAAAAAAAAAAAAAAA
@HWI-EAS146:5:1:1:1719#0/1
CAGATCTGGGTTATTGTAACTCCGCCTCNNNNNNNNNNNNNN
+
BCC7+<B=7BB=ABA7BBB8BB7B7AAAAAAAAAAAAAAAA
@HWI-EAS146:5:1:2:947#0/1
CCAGGAGAAAGCATGTTCAGTTGAGCGCNNANANGTGANNNN
+
BBBQ77A7>AAB>7B=7B>7B77AAAAAAAAAAAAAAAA
@HWI-EAS146:5:1:2:963#0/1
CCAGCCCTCCCTCATCTCCACCTGTACCTNANCCCTGANNNN
+
BBABAAB;AABA77@5AAA;??>AAAAAAAAAAAAAAAA
@HWI-EAS146:5:1:2:1631#0/1
TGGGAACGAGCTACACTCTCCAGGCTCTCTNCTCCGTNNNN
+
BBB@6@BBBDBB@BBBABAABBB7;9BB@BA5<B:~
@HWI-EAS146:5:1:2:1420#0/1
CTCAACTCCTGACCTTTGGTGATCCACCCCTTGGCCTTCNNNN
+
BBB;BBBBAABAA7; (=A@>AAA7AB7=AAAAAAAAAAAA
@HWI-EAS146:5:1:1:961#0/1
TCCGAGGCCAACCGAGGCTCCGCGCGCTGNNNNNNNNNNNNNN
+
BBB>A7B;@BBBBA=BA=AAAAAAAAAAAAAAAAAAAAAAAA
@HWI-EAS146:5:1:1:1595#0/1
TCAGGAAGCAGGAAGAGCTGGTGACGAGNNNNNNNNNNNNNN
+
B9BQ<;BAA=AB9=1>AAAAAAAAAAAAAAAAAAAAAAAA
@HWI-EAS146:5:1:1:1048#0/1
CTGGACTGCATCCTACCACCACTGTCCAANNNNNNNNNNNN
+
A=B7&7>B;A=79<;>747AAAAAAAAAAAAAAAAAAAAAAAA
@HWI-EAS146:5:1:1:1687#0/1
CTCTCTCAAGGTCCCAGAAGCAGCAGCAANNNNNNNNNNNN
+
BBCCCCCBBB7CB=7>+<=>=BCBCBAAAAAAAAAAAAAAAA
```

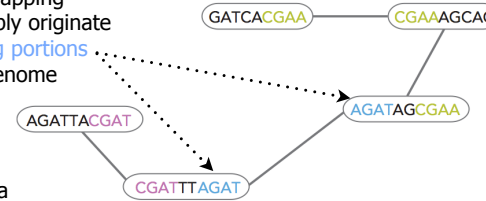

From reads to evidence

2. *de novo*

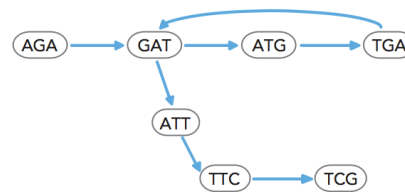
Assume nothing! - let reads tell us everything

Reads with overlapping sequence probably originate from overlapping portions of the subject genome

Encode overlap relationships as a graph



Source: De Novo Assembly Using Illumina Reads. Illumina. 2010



The full genome sequence
is a "tour" of the graph

Source: De Novo Assembly Using Illumina Reads. Illumina. 2010

http://www.illumina.com/Documents/products/technotes/technote_denovo_assembly.pdf

[illegible]

Mapping

Take a read:

```
CTCAAACTCCTGACCTTTGGTGATCCACCCGCTTNGGCCTTC
```

And a reference sequence:

```
>MT dna:chromosome chromosome:GRCh37:MT:1:16569:1
GATCACAGGTCTATCACCTATTAACTCACTACGGGAGCTCTCATGATTGGTATTT
CGTCTGGGGGATGACCGGATAGCATTGCGAGACGCTGGAGCCGGAGCACCTATGTC
GCAGTATCTGCTTTGATTCTGCTCATCTATTATTTATCGACCTACGTTCAATATT
ACAGGCCAACATACTACTAAAGTGTGTTAATTAATTATGCTTTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATATAACAAAAATTTCCACCA
AACCCCCCTCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAGAACCTTAACACAGCTTAACAGATTCAAAATTTATCTTTGGGCTATGCAC
TTTTAAGAGTCAACCCCAACTAACACATTATTTCCCTCCCACTCCCACTACTAAT
CTCATCAATACAAACCCCGCCATCTACCCAGCACACACACCGCTGCTAACCCATA
CCCGAACCAACCAACCCGACAGACAGCCGACCTTATCTAGCTTACCTGCTCAAA
GCAATACACTGACCCCTCAAACTCCTGGATTTTGGATCCACCCAGCGCTTGGCTAA
CTAGCCTTTCTATTAGCTTAACTAAATTAACATGACATGACATGACATGACATGAC
TCACCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGCAAGCAATGCAGCTC
AAAACGCTTAGCTAGCCACACACCCACGGGAACAGCAGTGATTAACTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACAAACCCAGGGTTGGTCAATTTCTGTCAGCCACCGC
GGTCAACGATTAAACCAAGTCAATAGAAGCCGGCTAAAGAGTGTTTAGATCACCCCT
TCCCAATAAAGCTAAACTCAGCTGAGTTGTAAATAACTCAAGTTGACACAAATAGAC
TACGAAAGTGGCTTAACTATCTGAACACACAATAGCTAAGACCAAACTGGGATTAGA
TACCCCACTATGCTAGCCCTAACTCAACAGTTAAATCAACAAACTGCTGCCAGAA
CACTACGAGCCACAGCTTAAACTCAAGGACCTGGCGGTCTTATATCCCTCTAGAGG
AGCTGTTCTGTAATGATAACCCGATCAACCTCACCACTCTTGTCTCAGCCTATAA
CCGCCATCTTCAAGCAACCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCAGTAAG
ACGTTAGGTCAGGTGTAGCCATGAGGTGGCAAGAATGGGCTACATTTTCTACCCAG
AAAACCTACGATAGCCCTTATGAACTTAAGGGTCGAAGGTGGATTAGCAGTAACTAAG
AGTAGAGTGCTTAGTTGAACAGGGCCCTGAAGCGCTACACACCCGCTGCTCCTCTC
AAGTATAGTGAAGGAGATTAAGGATTAAGGGGATTAAGGATTAAGGATTAAGGAT
CGTAACTCAAACTCCTGCTTTGGTGATCCACCCGCTTGGCTTACGATAAATGAAG
AAGCACTCACTCTGCTTAACTTAAGCTTAAGCTTAAGCTTAAGCTTAAGCTTAAGCT
GCCCAAACCACTCCACCTTACTACAGACAACTTAGCCAAACCAATTAACCAATAA
AGTATAGCGGATAGAAATGAACCTGGCGCAATAGATATAGTACCGCAAGGGAAGATG
AAAAATTAAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTCTGCAATAATGAA
TTAACTAGAAATAACTTTGCAAGGAGAGCAAGCTAAGACCCCGAAACAGACGAGCT
```

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

Answer: sequence similarity

Hypothesis 1:

```
Read
CTCAAAAGACCTGACCTTTGGTGATCCACCC-----GCCTNGGCCTTC
||||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
CTCAAACTCCTGATTTTG--GATCCACCCAGCTGGCCTTGCCCTAA
Reference
```

Hypothesis 2:

```
Read
CTCAAACTCCTGACCTTTGGTGATCCACCCGCTTNGGCCTTC
||||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| ||| |||
CTCAAACTCCTG--CCTTTGGTGATCCACCCGCTTGCCCTAC
Reference
```

Which hypothesis is better?

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

Mapping

Read

CTCAAACCTCTGACCTTTGGTGATCCACCCGCTTNGGCCTTC

Reference

```
>MT daa:chromosome chromosome:GRCh37:MT:1:16569:1
GATCACAGGTCTATCACCTATTAACTCACTACGGGAGCTCTCATGATTGGTATTTT
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCTATGTC
GCAGTATCTGCTTTGATTCTCTGCTCATCTATTATTTATCGACCTACGTTCAATATT
ACAGGCGAACATACTTACTAAAGTGTGTTAATTAAATTAATGCTTTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCACTTTCCACACAGACATCAACAAAAAATTTCCACCA
AACCCCCCTCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCTTAACACAGCCTAACAGATTTCAAATTTTATCTTTGGCGGTATGCAC
TTTAAACAGTACCCCCCACTAACACATTATTTTCCCTCCCACTCCCATACTACTAAT
CTCATCAATACACCCCCGCTATCTACCCAGCAGACACACCGCTGCTAACCCATA
CCCCGAACCAACCAAAACCCCAAGACACCCCCCAAGTTTATGTAGCTTACCTCTCAAA
GCAATACACTGACCCGCTCAAACTCTGGATTTTGTGATCCACCCAGCGCTTGGCCTAA
CTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCGTTCCAGTGAGT
TCACCTCTAAATCACCACGATCAAAGGAACAAGCATCAAGCAGCGACGATGCAGCTC
AAAACGCTTAGCCTAGCCACACCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACCTAACCCAGGGTTGGTCAATTTCTGTCAGCCACCGC
GGTCACACGATTAAACCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTTAGATCACCCCT
TCCCAATAAAGCTAAAACTCACCTGAGTTGTAATAAACTCCAGTTGACACAAAATAGAC
TACGAAAGTGGCTTTAAACATATCTGAACACACAATAGCTAAGACCCAACTGGGATTAGA
TACCCCACTATGCTAGCCCTAAACCTCAACAGTTAAATCAACAAAATGCTCGCAGAA
CACTACGAGCACAAGCTTAAACTCAAGGACCTGGCGGTGCTTATATCCTCTAGAGG
AGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACTCTTGTCTAGCCTATATA
CCGCCATCTTCAGCAACCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCAGTAAAG
ACGTTAGTCAAGGTGTAGCCCATGAGGTGGCAAGAATGGGCTACATTTTCTACCCAG
AAAACACGATAGCCCTTATGAACTTAAGGGTGAAGGTGGATTAGCAGTAAACTAAG
AGTAGAGTGCTTAGTTGAACAGGGCCCTGAAGCGGTACACACCGCCGCTCACCTCTC
AAGTATACTTCAAGGACATTTAACTAAACCCCTACGCATTATATAGAGGAGACAAGT
CGTAACCTCAAACTCTGGCCTTTGGTGATCCACCCGCTTGGCCTACCTGCATAATGAA
AAGCACCCTTACACTTAGGAGATTTCACCTTAACCTTGACCGCTCTGAGCTAAACCTA
GCCCCAAACCTCTCACCTTACTACAGCAACCTTAGCACAACATTATCCCAATTA
AGTATAGGCGATAGAATTGAACCTGGCGCAATAGATATAGTACCGAAGGGAAGATG
AAAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTTCTGCATAATGAA
TTAACTAGAAATAACTTTGCAAGGAGAGCAAGCTAAGACCCCGAAACAGACGAGCT
ACCTAAGAACAGCTAAAAGAGCACACCGCTCTATGTAGCAAAATAGTGGGAAGATTATA
GGTAGAGGCGACAAACCTACCGAGCTGGTGATAGCTGGTTGTCCAAGATAGAATCTAG
```

This is an **alignment**:

```
Read
CTCAAGACCTGACCTTTGGTGATCCACCC-----GCCTNGGCCTTC
||||| ||| ||||| ||||| ||||| ||||| ||||| |||||
CTCAAACTCCTGGATTTG--GATCCACCCAGCTGGCCTTGGCCTAA
Reference
```

Software programs that compare reads to references and find alignments are **aligners**.

Alignment is computationally difficult because references (e.g. human) are **very long** (more than 1M times longer than what's shown to the left) and sequencers produce data very rapidly, e.g. up to **25 billion bases per day** in 2010.

Sequencing throughput increases by ~5x per year, whereas computers get faster at a rate closer to ~2x every 2 years.

Mapping

Take a read:

CTCAAACCTCTGACCTTTGGTGATCCA

And a reference sequence:

```
>MT dna:chromosome chromosome:GRCh37:MT:1:16569:1
GATCACAGGTCTATCACCCCTATTAAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTATGCACGGGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCCTATGTC
GCAGTATCTGCTTTGATTCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT
ACAGGCCAACACTACTTACTAAAGTGTGTTAATTAATTAACTGCTTAGGCATATAATAA
ACAATTGAATGTCTGCACAGCCACTTTCACACAGACATCATACAAAAAATTTCCACCA
AACCCCCCTCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCTTAACACAGCCTAACAGATTTCAAATTTTATCTTTTGGCGGTATGCAC
TTTTAAGAGTCACCCCCAACTAACACATTATTTCCCTCCCACTCCCATACTACTAAT
CTCATCAATACAAACCCCGCCATCTACCCAGCACACACACCGCTGCTAACCCCA
CCCGAACCAACCAACCCCAAGACACCCCAACAGTTTATGATGCTTACCTCCTCAA
GCAATACACTGACCCGCTCAAACTCTGGATTTTGTGATCCACCCAGCGCTTGGCTAA
CTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCGTTCCAGTGAGT
TCACCCCTCTAAATCACCAGCATCAAAGGAACAAGCATCAAGCAGCGCAATGCAAGCTC
AAACGCTTAGCCTAGCCACACACCCACGGGAACAGCAGTGATTAACTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACTAACCCCAAGGTTGGTCAATTTCTGTCAGCCACCGC
GGTCAACGATTAAACCAAGTCAATAGAAGCGGCGTAAGAGTGTTTTAGATCACCC
TCCCAATAAAGCTAAACTCACCTGAGTTGTAAAAAACTCCAGTTGACACAAATAGAC
TACGAAAGTGGCTTAACTATCTGAACACACAATAGCTAAGACCAAACTGGGATTAGA
TACCCCACTATGCTTAGCCCTAACTCAACAGTTAAATCAACAAACTGCTGCCAGAA
CACTACGAGCCACAGCTTAAACTCAAAGGACCTGGCGGTGCTTATATCCCTCTAGAGG
AGCCTGTTCTGTAATCGATAAACCCGATCAACCTCACACCTCTTGTCTCAGCCTATATA
CCGCCATCTTCAGCAACCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCAGTAAAG
ACGTTAGGTCAGGTGTAGCCATGAGGTGGCAAGAATGGGCTACATTTTCTACCCAG
AAAACTACGATAGCCCTTATGAACTTAAGGGTCGAAGGTGGATTAGCAGTAACTAAG
AGTAGAGTGTCTAGTTGAACAGGGCCCTGAAGCGGTACACACCGCCGTCACCTCTC
AAGTATACCTCAAGGACATTAACATAAACCCCTACGCTTTATATAGAGGAGACAAAGT
CGTAACCTCAAACTCTGGCTTTGGTGATCCACCGCTTGGGCTACCTGCATAATGAA
AAGCACCACCACTTACCTTAGGAGATTCAACTTAACTTGACGCTCTGAGCTAACTA
GCCCAAAACCACTCCACCTTACTACCAAGACAACTTAGCACAACCTTACCAATAA
AGTATAGCGATAGAAATTGAACTGGCGCAATAGATAGTACCGCAAGGGAAAGATG
AAAAATTATAACCAAGCATATAAGCAAGGACTAACCCCTATACCTCTGCTAATGAA
TTAAGTAGAAATACTTTGCAAGGAGAGCAAGCTAAGACCCCGAAACAGACGAGCT
```

Which hypothesis is better?

Hypothesis 1:

Read
CTCAAACCTCTGACCTTTGGTGATCCA
|||||
Reference
CTCAAACCTCTGACCTTTGGTGATCCA

Hypothesis 2:

Read
CTCAAACCTCTGACCTTTGGTGATCCA
|||||
Reference
CTCAAACCTCTGACCTTTGGTGATCCA

Is there any way to break the tie?

Mapping

@H1-EAS146:5:1:1:9639/1
 TCGAGGCGCAACCGAGGCTCGCGGCGTCNNNNNNNNNNNNNNNN
 +
 BBB@A7BB:GBBBBAAA-BA=AA/1
 @H1-EAS146:5:1:1:139590/1
 TCAGGAAGCAGGAAGCTGCTGACGAGGNNNNNNNNNNNNNNNN
 +
 B9BB@:BAA@AB9-1=1:10480/1
 @H1-EAS146:5:1:1:10480/1
 CTGCAGTCGATCTCAACCAACTCGTCCAAANNCHNNNNCHNNNN
 +
 A@B767:~BB@:A~79:~1:~7474/
 @H1-EAS146:5:1:1:160780/1
 CTCCTCTCAAGGTCCCGAGAGACGACCAANNNNANNNTNCTNNNN
 +
 BACGCCCCCBGBTCB?CB=?-?@-?BCB@B?
 @H1-EAS146:5:1:1:160780/1
 CACCTCTCTGATTTGTAAATCCCTCGCTCCTNNNGNTTAAAGNNNN
 +
 C?7?-B-7BBS-ABA-7BBSBBB-AB87B?
 @H1-EAS146:5:1:1:29479/1
 CCGAGGAAGAAGCATCTGAGTTCGAGGCGNNNANNCCTGAGNNNN
 +
 BBBS977A7-AB@B-7B?-7@B7?
 @H1-EAS146:5:1:1:25639/1
 CAGCGCCCTCCCGATCTCCACCGCTACTCTNACCCCTGAGNNNN
 +
 BBBAABAB:AAAB77@5AA:77?
 @H1-EAS146:5:1:1:26131/1
 TGGGAGCAGCGGCTCAACTCTTCGAGGCTCTCTCCCTGCTNNNN
 +
 BBBBGGBBBBAABBB@BBBBAABBB7;9BB@A5@-~?~~~~~
 @H1-EAS146:5:1:1:24280/1
 TCAAACTCTGAGGCTTTGATGATCACCCTGACCCGCTTNGGCTCTNNNN
 +
 BBBB:BBBBAABAA7:-(A@B@-AA7A87-A@~~~~~
 @H1-EAS146:5:1:1:24280/1
 TCGAGGACGACGAGGCTCGCGGCTCGNNNNNNNNNNNNNNNN
 +
 BBB@-A7BB:GBBBBAAA-BA=AA/1
 @H1-EAS146:5:1:1:139590/1
 TCAGGAAGCAGGAAGCTGCTGACGAGGNNNNNNNNNNNNNNNN
 +
 B9BB@:BAA@AB9-1=1:10480/1
 @H1-EAS146:5:1:1:10480/1
 CTGCAGTCGATCTCAACCAACTCGTCCAAANNCHNNNNCHNNNN
 +
 A@B767:~BB@:A~79:~1:~7474/
 @H1-EAS146:5:1:1:160780/1
 CTCCTCTCAAGGTCCCGAGAGACGACCAANNNNANNNTNCTNNNN
 +
 BACGCCCCCBGBTCB?CB=?-?@-?BCB@B?
 @H1-EAS146:5:1:1:160780/1
 CACCTCTCTGATTTGTAAATCCCTCGCTCCTNNNGNTTAAAGNNNN

- Recall that reads come with per-cycle quality values (in red)

In FASTQ format (left), qualities are encoded as ASCII characters like B, = or %, but really they're integers [0, 40]

A quality value Q is a function of the probability P that the sequencing machine called the wrong base:

$$Q = -10 \cdot \log_{10}(P)$$

Q = 10: 1 in 10 chance that base was miscalled

Q = 20: 1 in 100 chance

Q = 30: 1 in 1000 chance

Higher is "better."

Qs are estimated by the sequencer's software and aren't necessarily accurate

Mapping

Take a read:

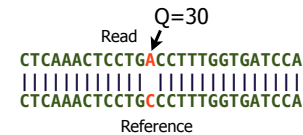
CTCAAACCTCTGACCTTTGGTGATCCA

And a reference sequence:

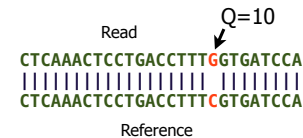
```
>MT dna:chromosome chromosome:GRCh37:MT:1:16569:1
GATCACAGGTCTATCACCCCTATTAAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTATGCACGGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCCTATGTC
GCAGTATCTGCTTTGATTCTGCCTCATCTATTATTTATCGCACCTACGTTCAATATT
ACAGGCCAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCACACAGACATCATAAACAAAAATTTCCACCA
AACCCCCCTCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCTTAACACAGCCTAACAGATTTCAAATTTTATCTTTTGGGGTATGCAC
TTTTAAGAGTCACCCCAACTAACACATATTTTCCCTCCCACTCCCATACTACTAAT
CTCATCAATACAAACCCCGCCATCTACCCAGCACACACACCGCTGCTAACCCATA
CCCGAACCAACCAACCCCAAGACACCCCAAGTTTATGTAGCTTACCTCCTCAA
GCAATACACTGACCCGCTCAAACTCTGGATTTTGTGATCCACCCAGCGCTTGGCTAA
CTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCGTTCCAGTGAGT
TCACCTCTAAATCACCAGCATCAAAGGAACAAGCATCAAGCACGCAAGCAATGCAGCTC
AAACGCTTAGCCTAGCCACACACCCACGGGAACAGCAGTGATTAACTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACTAACCCCAAGGTTGGTCAATTTCTGTCAGCCACCGC
GGTCAACGATTAAACCAAGTCAATAGAAGCGGCGTAAGAGTGTTTATAGATCACCC
TCCCAATAAAGCTAAACTCACCTGAGTTGTAAATAACTCCAGTTGACACAAATAGAC
TACGAAAGTGGCTTAACATATCTGAACACACAATAGCTAAGACCCTAACTGGGATTAGA
TACCCCACTATGCTAGCCCTAACTCAACAGTTAAATCAACAAACTGCTGCGCAGAA
CACTACGAGCCACAGCTTAAACTCAAAGGACCTGGCGGTCTTATATCCCTCTAGAGG
AGCCTGTTCTGTAATCGATAAACCCGATCAACCTCACACCTCTTGTCTCAGCCTATATA
CCGCCATCTTCAGCAACCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCAGTAAAG
ACGTTAGGTCAGGTGTAGCCATGAGGTGGCAAGAAATGGGCTACATTTCTACCCAG
AAAACCTACGATAGCCCTTATGAACTTAAGGGTGAAGGTGGATTAGCAGTAAACTAAG
AGTAGAGTGCTTAGTTGAACAGGGCCCTGAAGCGGTACACACCGCCGTCACCTCTCTC
AAGTATACCTCAAGGACATTAACATAAACCCCTACGCTTTATATAGAGGAGACAAAGT
CGTAACCTCAAACTCTGGCTTTGGTGATCCACCCGCTTGGGCTACCTGCATAATGAA
AAGCAACCACTTACCTTAGGAGATTCAACTTAACTTGACGCTCTGAGCTAAACCTA
GCCCAAAACCACTCCACCTTACTACCAAGACAACTTAGCCAAACCTTTACCAATAA
AGTATAGCGGATAGAAATTGAACTTGGCGCAATAGATATAGTACCGCAAGGGAAAGATG
AAAAATTATAACCAAGCATATAATAGCAAGGACTAACCCCTATACCTCTGCTAATGAA
TTAACTAGAAATAACTTTGCAAGGAGAGCAAGCTAAGACCCCGAAACCAAGCAGAGCT
```

Which hypothesis is better?

Hypothesis 1:



Hypothesis 2:



Mapping

Take a read:

CTCAAACCTCTGACCTTTGGTGATCCA

And a reference sequence:

```
>MT dna:chromosome chromosome:GRCh37:MT:1:16569:1
GATCACAGGTCTATCACCTATTAACCACTCACGGGAGCTCTCCATGATTGGTATTTT
CGTCTGGGGGGTATGCACGGATAGCATTGCGAGACGCTGGAGCCGGAGCACCTATGTC
GCAGTATCTGCTTTGATTCTGCTCATCTTATTTATCGCACCTACGTTCAATATT
ACAGGCCAACATACTTACTAAAGTGTGTTAATTAATTAACTGCTTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCACACAGACATCATACAAAAAATTTCCACCA
AACCCCCCTCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCTTAACACAGCTTAACAGATTCAAAATTTATCTTTTGGGGTATGCAC
TTTTAAGAGTCAACCCCAACTAACACATTTTCCCTCCCACTCCCATACTACTAAT
CTCATCAATACAAACCCCGCCATCTACCCAGCACACACACCGCTGCTAACCCATA
CCCCGAACCAACCAACCCCAAGACACCCCAACAGTTTATGATGCTTACCTCTCAA
GCAATACACTGACCCGCTCAAACTCTGGATTTTGTGATCCACCCAGCGCTTGGCTAA
CTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCGTTCCAGTGAGT
TCACCCCTCTAAATCACCAGCATCAAAGGAACAAGCATCAAGCACGCAATGCAAGCTC
AAAACGCTTAGCCTAGCCACACACCCACGGGAACAGCAGTGATTAACTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACTAACCCAGGGTTGGTCAATTTCTGTCAGCCACCGC
GGTCAACGATTAAACCAAGTCAATAGAAGCCGGCGTAAGAGTGTTTTAGATCACCCC
TCCCAATAAAGCTAAACTCACCTGAGTTGTAAAAAACTCCAGTTGACACAAATAGAC
TACGAAAGTGGCTTAACTATCTGAACACACAATAGCTAAGACCAAACTGGGATTAGA
TACCCCACTATGCTTAGCCCTAACTCAACAGTTAAATCAACAAACTGCTGCCAGAA
CACTACGAGCCACAGCTTAAACTCAAAGGACCTGGCGGTGCTTATATCCCTCTAGAGG
AGCCTGTTCTGTAATCGATAACCCGATCAACCTCACACCTCTTGTCTCAGCTATATA
CCGCCATCTTCAGCAACCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCAGTAAAG
ACGTTAGGTCAGGTTAGCCCATGAGGTGGCAAGAAATGGGCTACATTTCTACCCAG
AAAACCTACGATAGCCCTTATGAACTTAAGGGTCGAAGGTGGATTAGCAGTAACTAAG
AGTAGAGTGTCTAGTTGAACAGGGCCCTGAAGCGCTACACACCGCCGTCACCTCTCTC
AAGTATACCTCAAGGACATTAACATAAACCCCTACGCTTTATATAGGAGGAGACAGT
CGTAACCTCAAACTCTGGCTTTGGTGATCCACCGCTTGGCTACCTGCATAATGAA
AAGCACCCCACTTACCTTAGGAGATTCAACTTAACTTGACCGCTCTGAGCTAACTA
GCCCAAACTCACTTCACTTACTACAGACAACTTAGCCAAACCTTATCCCAATAA
AGTATAGGCGATAGAAATGAACCTGGCGCAATAGATATAGTACCGCAAGGGAAGATG
AAAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTCTGCATAATGAA
TTAACTAGAAATAACTTTGCAAGGAGAGCCAAAGCTAAGACCCCGAAACAGACGAGCT
```

Which hypothesis is better?

Hypothesis 1:

Read
CTCAAACCTCTGACCTTTGGTGATCCA
|||||
CTCAAACCTCTG-CCTTTGGTGATCCA
Reference

Hypothesis 2:

Read
CTCAAACCTCTGACCTTTGGTGATCCA
|||||
CTCAAACCTCTGACCTTTGGTGATCCA
Reference

Is there any way to break the tie?

Hint: In Illumina sequencing, sequencing errors
almost never manifest as gaps

Mapping

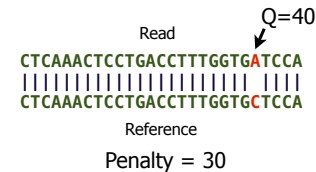
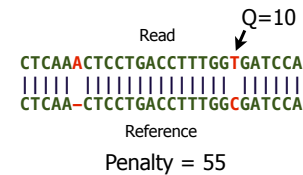
Aligners can employ **penalties** to account for the relative probability of seeing different dissimilarities

Estimates vary, but small gaps ("indels") occur in humans at 1 in ~10-100K positions.

$$\begin{aligned} Pen_{\text{gap}} &\equiv -10 \log_{10}(P_{\text{gap}}) \\ &= -10 \log_{10}(0.00005) \\ &\approx 45 \end{aligned}$$

SNPs occur in humans at 1 in ~1K positions, but depending on Q, sequencing error may be more likely

$$\begin{aligned} Pen_{\text{mm}} &\equiv \arg \min(-10 \log_{10}(P_{\text{miscall}}), -10 \log_{10}(P_{\text{SNP}})) \\ &= \arg \min(Q, -10 \log_{10}(0.001)) \\ &= \arg \min(Q, 30) \end{aligned}$$



```
\begin{align*}
Pen_{\text{gap}} &\equiv -10 \log_{10}(P_{\text{gap}}) \\
&= -10 \log_{10}(0.00005) \\
&\approx 45
\end{align*}
```

```
\begin{align*}
Pen_{\text{mm}} &\equiv \arg \min(-10 \log_{10}(P_{\text{miscall}}), -10 \log_{10}(P_{\text{SNP}})) \\
&= \arg \min(Q, -10 \log_{10}(0.001)) \\
&= \arg \min(Q, 30)
\end{align*}
```