Statistics applied to second-generation sequencing data

Second-Generation Sequencing

nature news

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

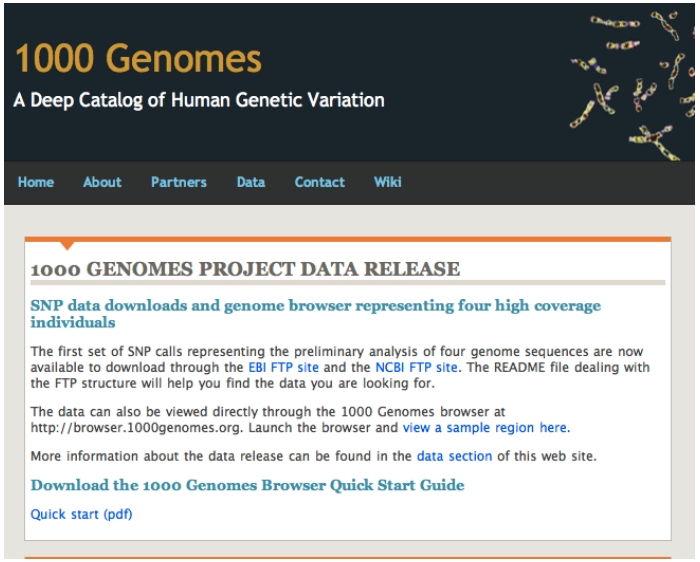
Send to a Friend

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Second-Generation Sequencing

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

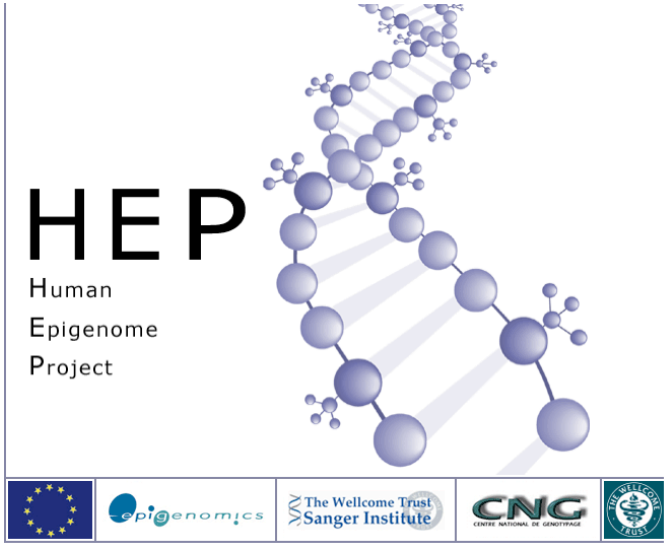
1000 Genomes Project



Genotyping

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Human Epigenome Project



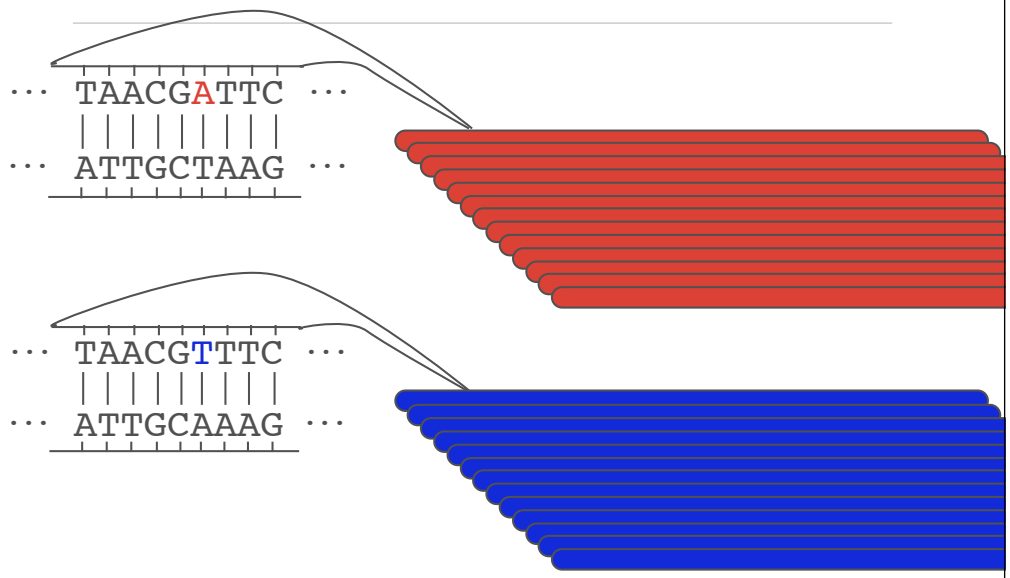
Methylation

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Platforms

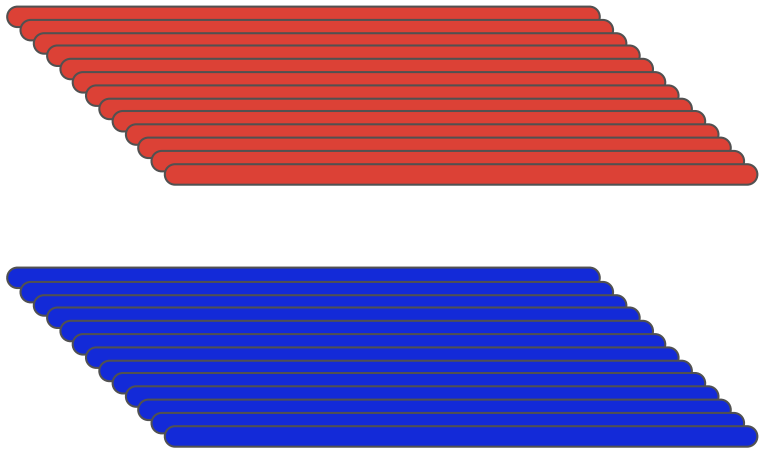
- Millions of short DNA fragments (~36-70 bp) sequenced in parallel

Sec-gen Sequencing

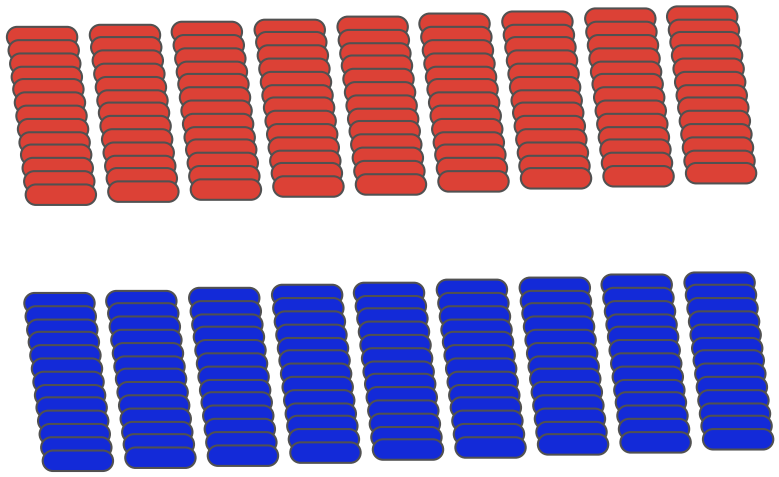


JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

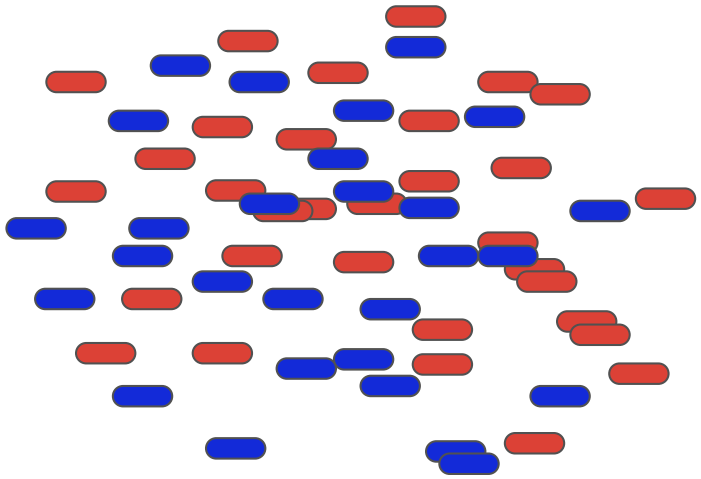
Sec-gen Sequencing



Sec-gen Sequencing



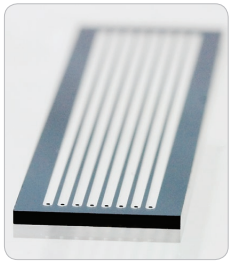
Sec-gen Sequencing



A Set of Short Reads

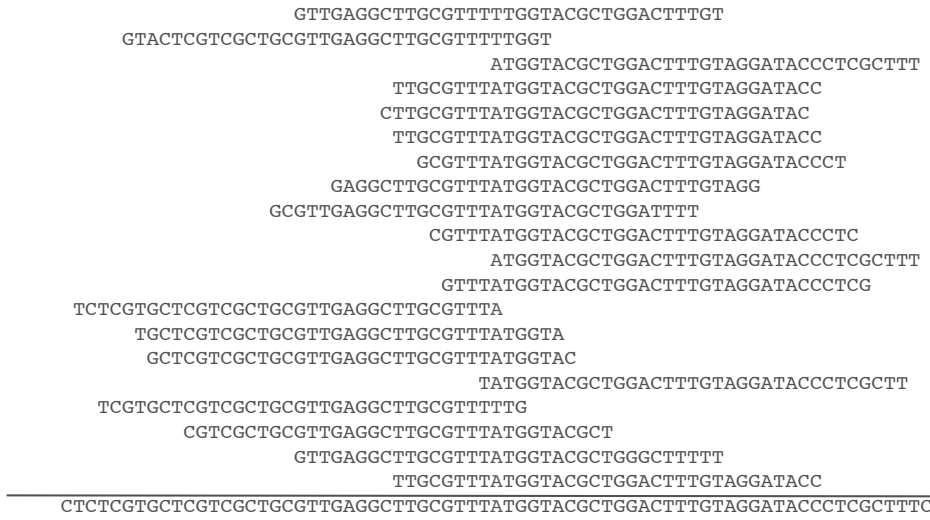
```
GTTGAGGCTTGCGTTTTTGGTACGCTGGACTTTGT
GTACTCGTCGCTGCGTTGAGGCTTGCGTTTTTGGT
ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
CTTGCGTTTATGGTACGCTGGACTTTGTAGGATAC
TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
GCGTTTATGGTACGCTGGACTTTGTAGGATACCCT
GAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGG
GCGTTGAGGCTTGCGTTTATGGTACGCTGGATTTT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

Illumina/Solexa



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

Matching



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

```
                        GTTGAGGCTTGCGTTTTTGGTACGCTGGACTTTGT
  GTACTCGTCGCTGCGTTGAGGCTTGCGTTTTTGGT
                                              ATGGTACGCTGGACTTTGTAGGATACCCTCGCTTT
                                 TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
                                CTTGCGTTTATGGTACGCTGGACTTTGTAGGATAC
```

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).

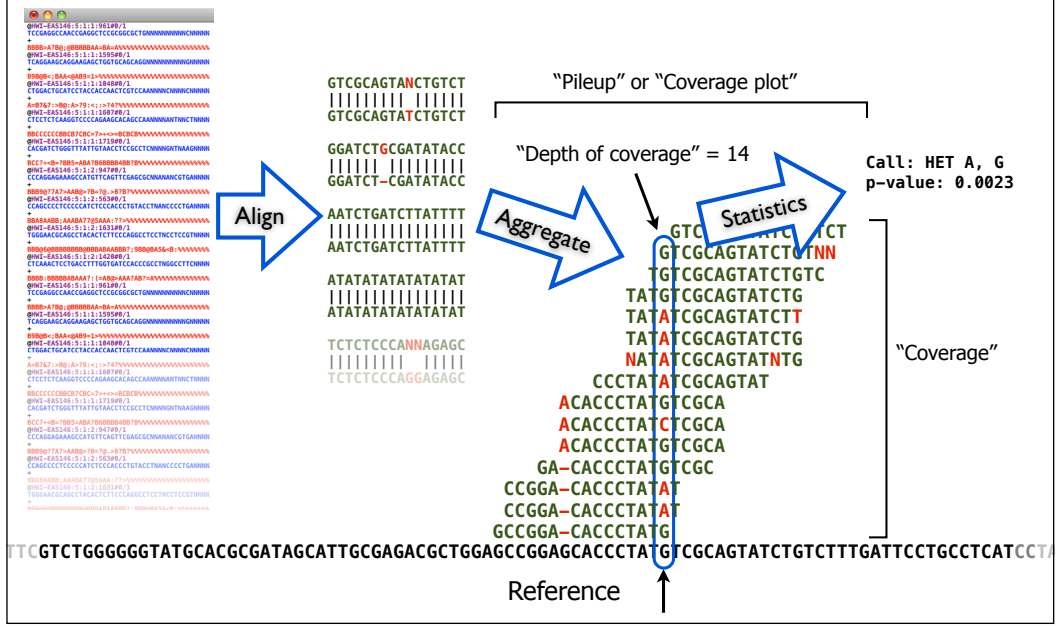
OSI certified

```
          TGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTA
           GCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTAC
                                              TATGGTACGCTGGACTTTGTAGGATACCCTCGCTT
      TCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTTTG
             CGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCT
                        GTTGAGGCTTGCGTTTATGGTACGCTGGGCTTTTT
                                 TTGCGTTTATGGTACGCTGGACTTTGTAGGATACC
  CTCTCGTGCTCGTCGCTGCGTTGAGGCTTGCGTTTATGGTACGCTGGACTTTGTAGGATACCCTCGCTTTC
```

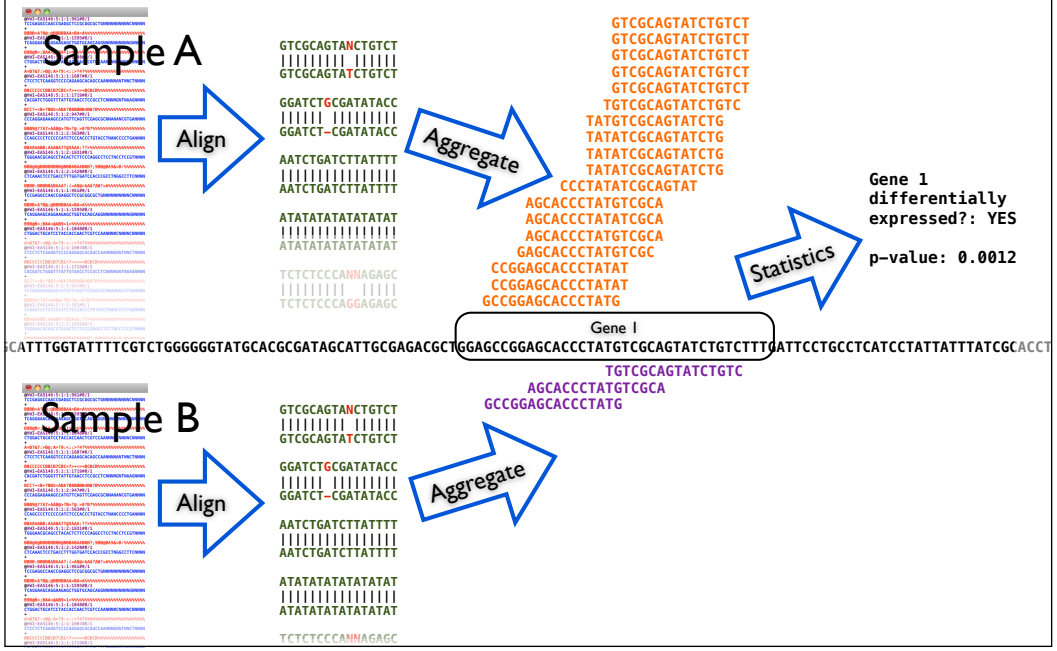

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Workflow examples

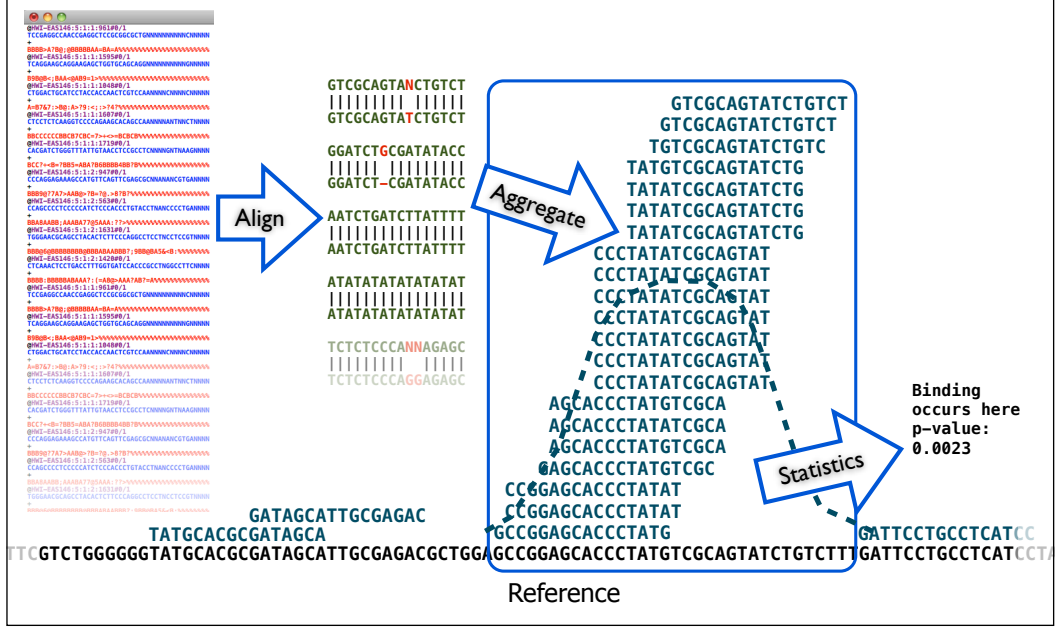
Variant detection



RNA-seq differential expression

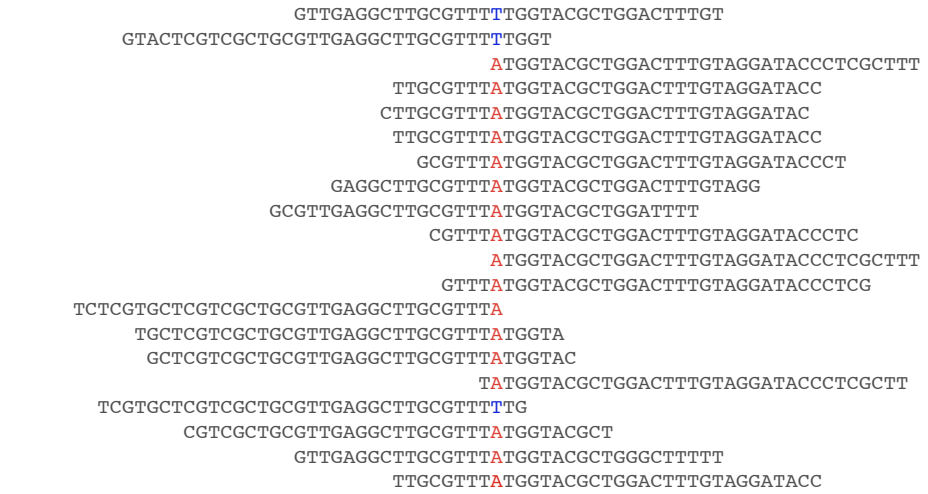


ChIP-seq

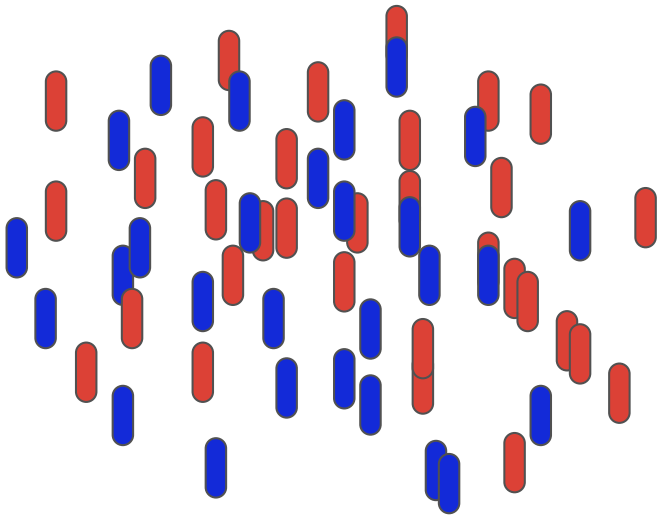

Align
Aggregate
Statistics
Binding occurs here
p-value: 0.0023
Reference

The End

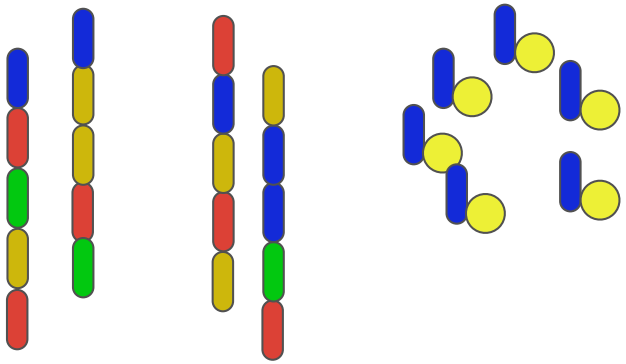
SNPs



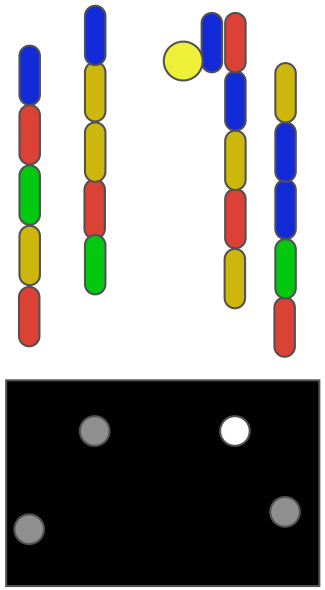
Sec-gen Sequencing



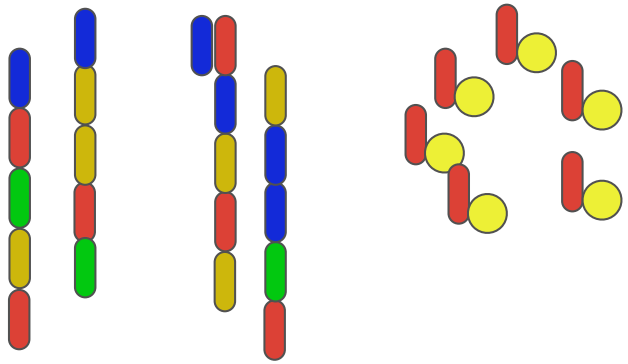
Sec-gen Sequencing



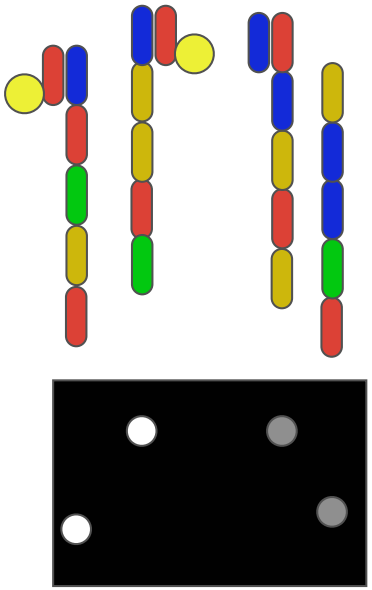
Sec-gen Sequencing



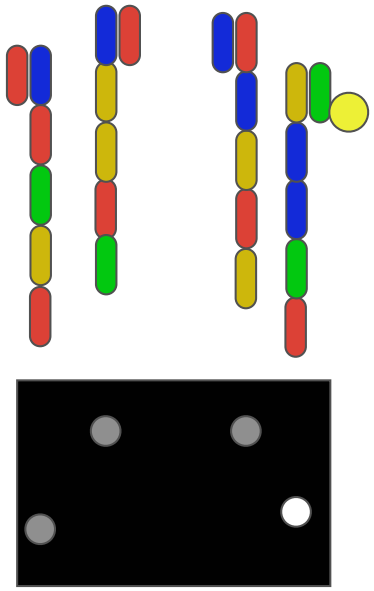
Sec-gen Sequencing



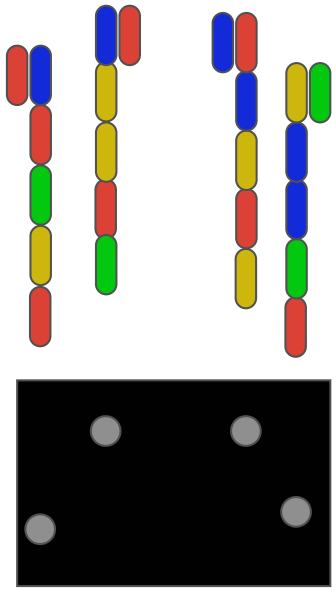
Sec-gen Sequencing



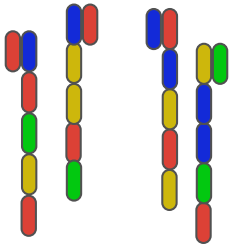
Sec-gen Sequencing



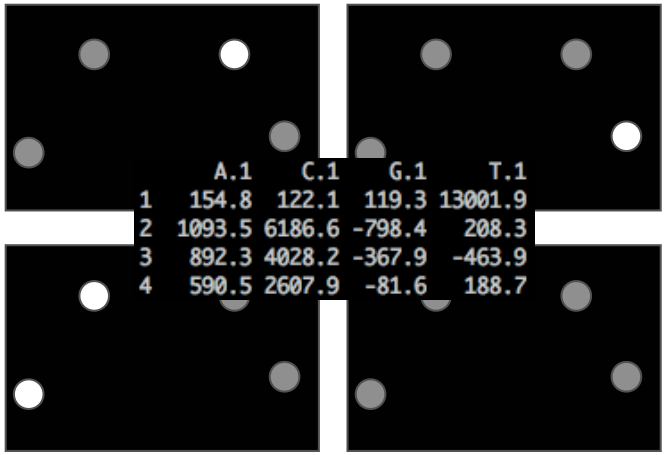
Sec-gen Sequencing



Sec-gen Sequencing

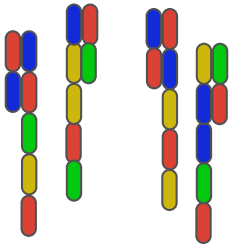


First Cycle

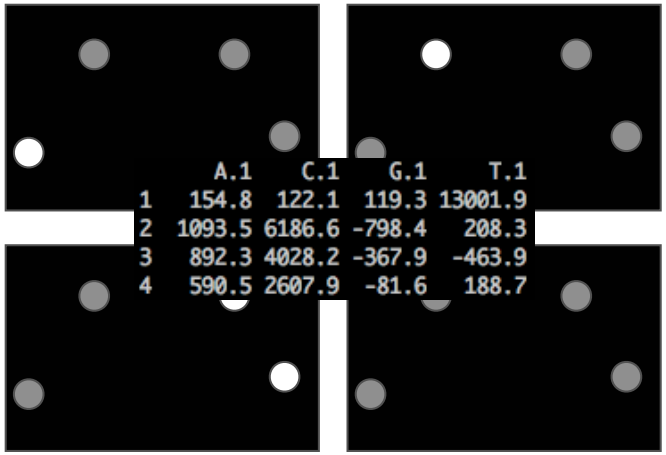


JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Sec-gen Sequencing

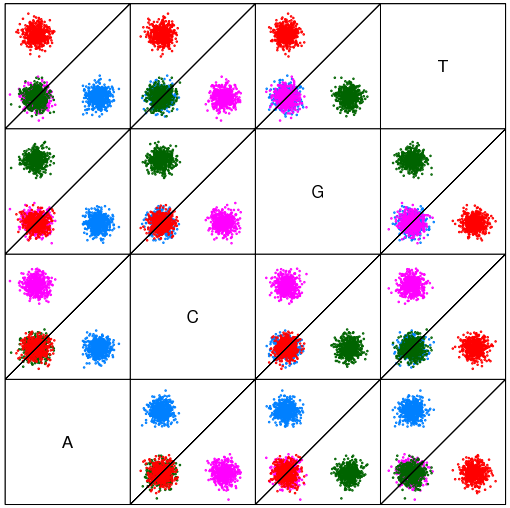


Second Cycle



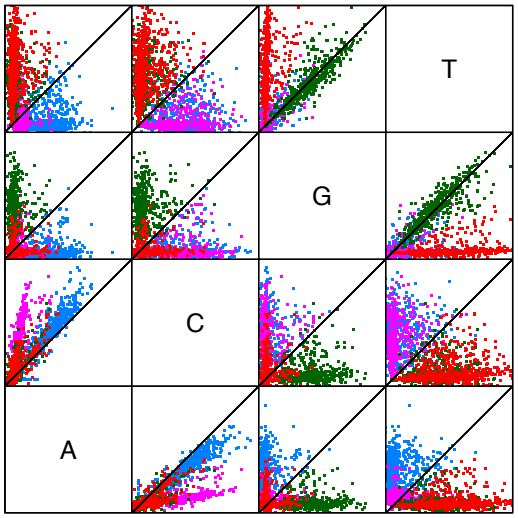
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

A Thought Experiment



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

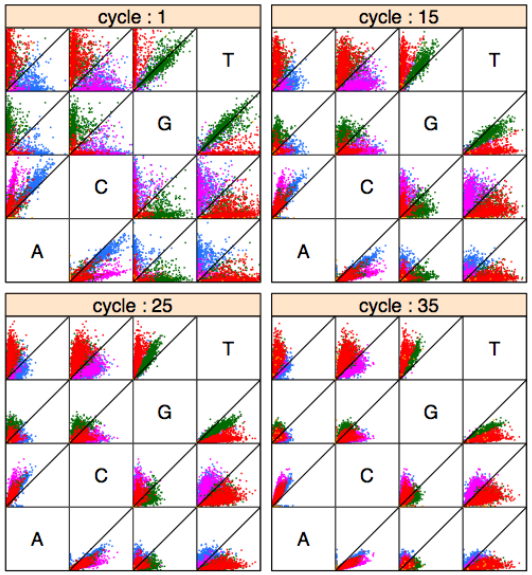
Fluorescence Intensity



Four–channel fluorescence intensity, cycle 1

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

Fluorescence Intensity



Four−channel fluorescence intensity

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

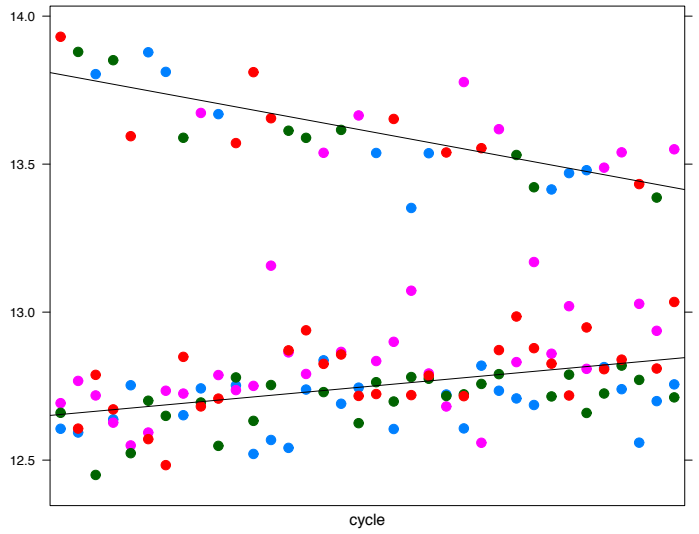
- log intensity read *i*, cycle *j*, channel *c*

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

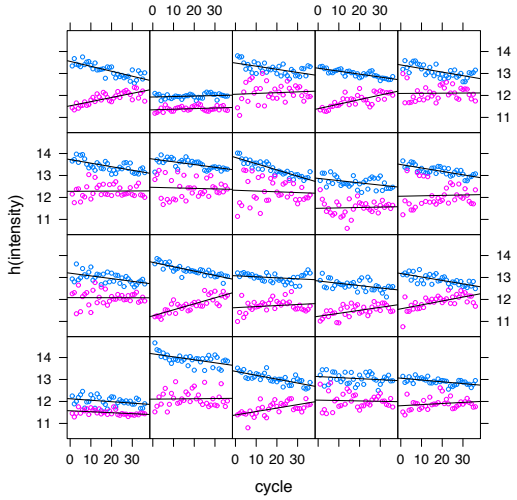
- indicators of nucleotide identity, read *i*, pos. *j*

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

Intensity Model (read-cycle effect)



Read & Cycle Effects



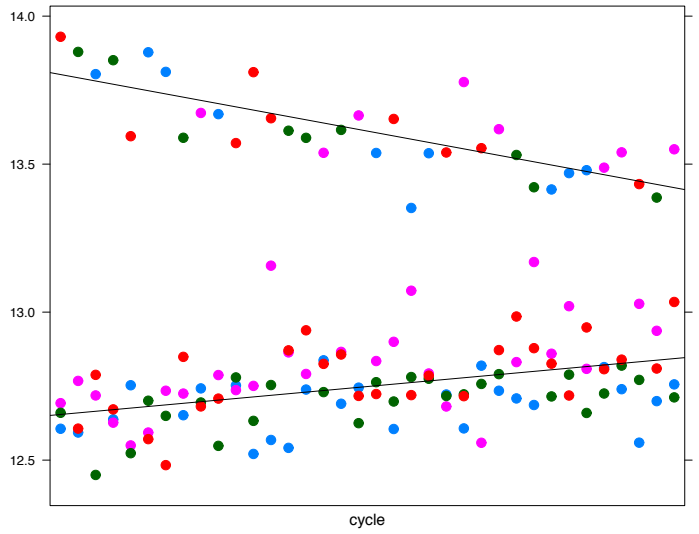
Intensity Model

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

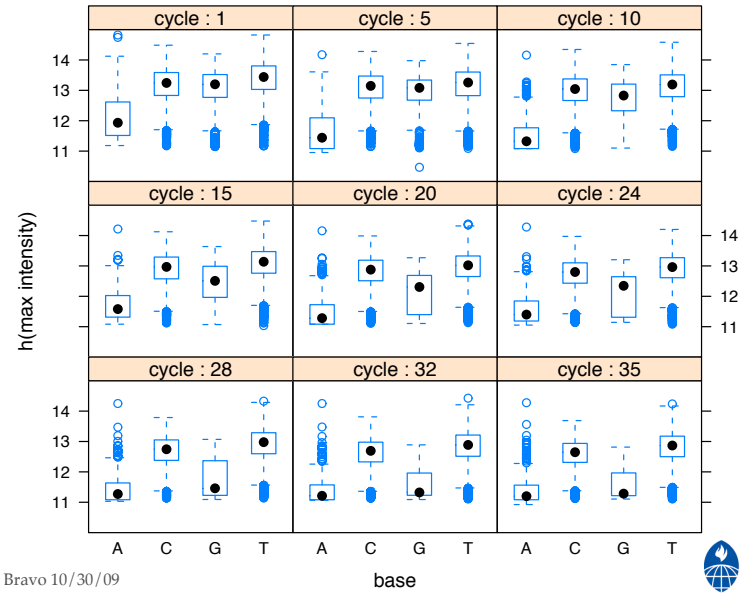
$$u_{ijc} = \Delta_{ijc}(\mu_{cj\alpha} + \underline{x_j^T \alpha_i} + \epsilon_{ijc}^{\alpha}) + (1 - \Delta_{ijc})(\mu_{cj\beta} + \underline{x_j^T \beta_i} + \epsilon_{ijc}^{\beta})$$

- 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*

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

- random noise

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

Intensity Model

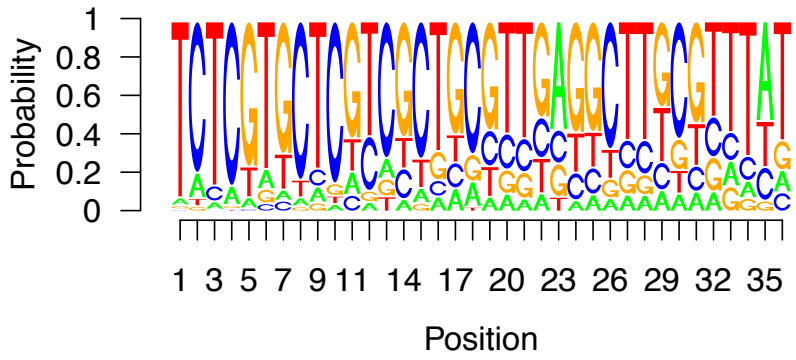
- log intensity read *i*, cycle *j*, channel *c*

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

- estimate parameters with EM algorithm
- which also estimates

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

Base Identity Probability Profiles



JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

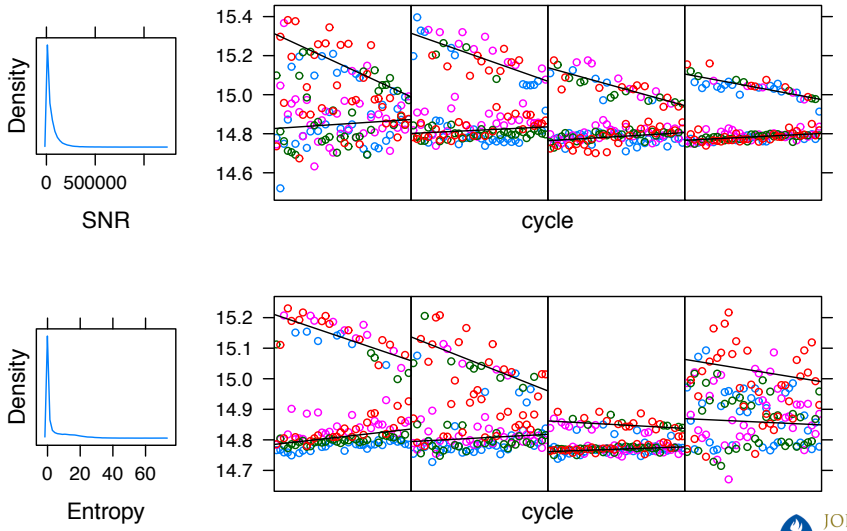
Quality Metrics

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

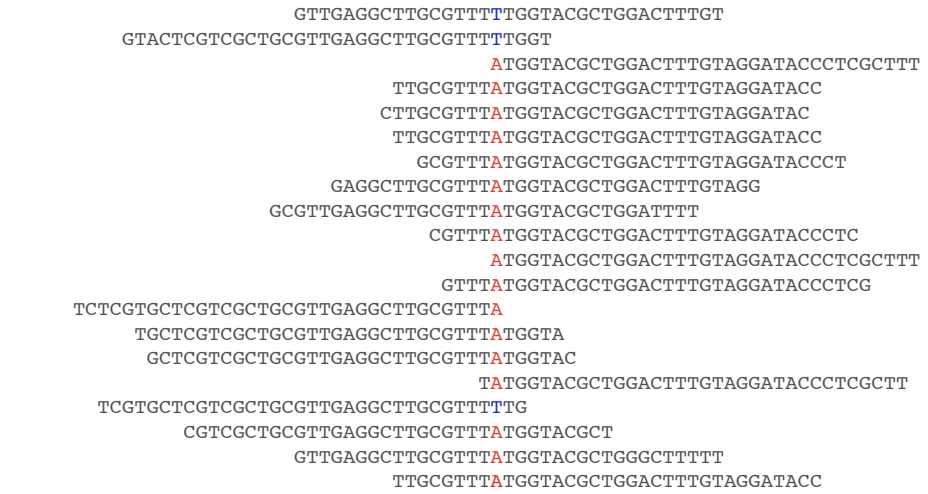
$$H_{ij} = -\sum_{c} z_{ijc} \log_2 z_{ijc}$$

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

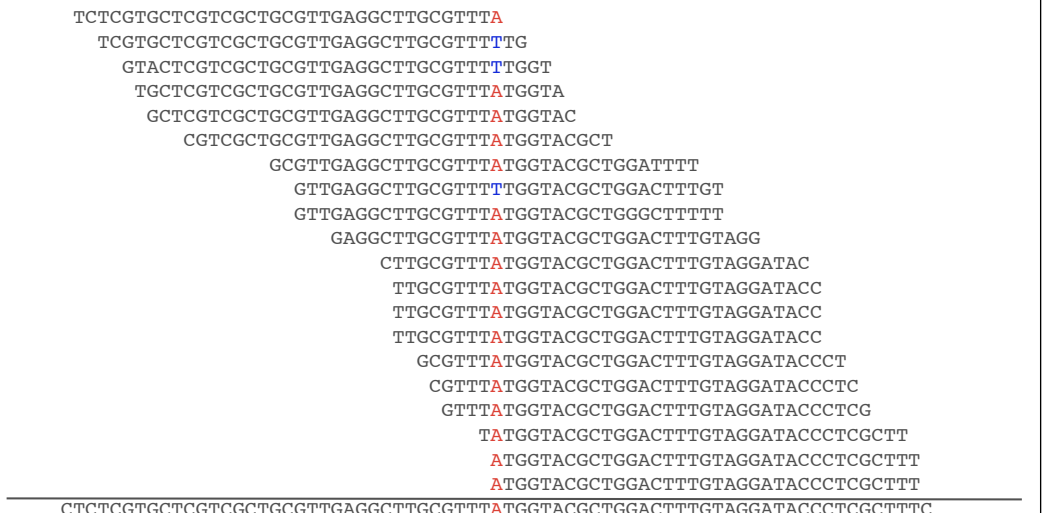
Quality Metrics



SNPs

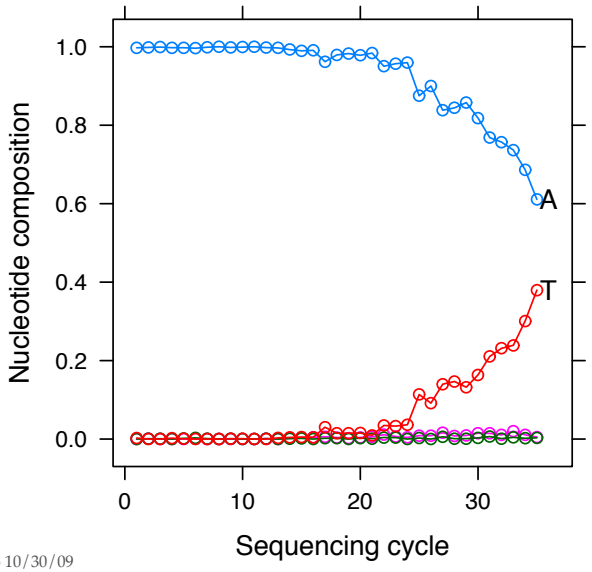


SNPs



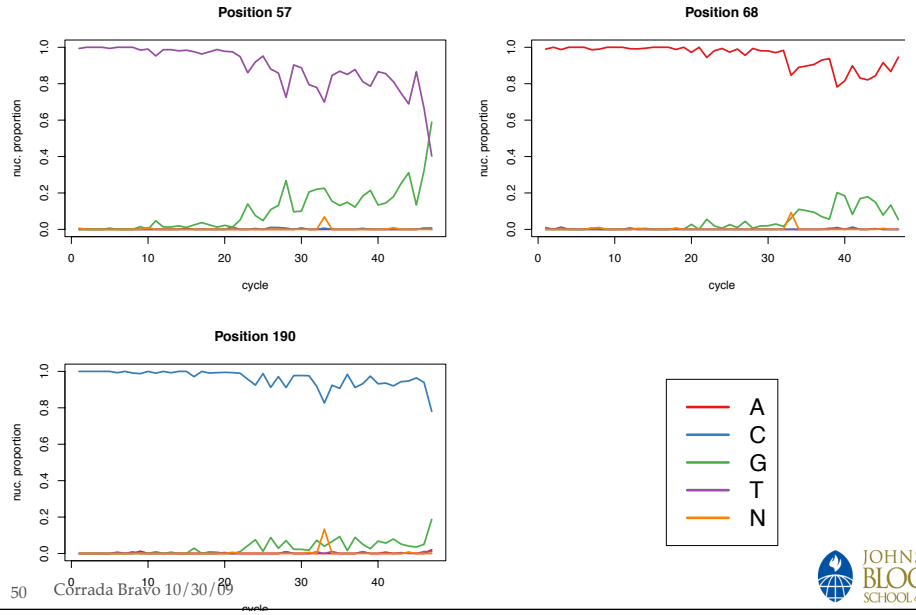
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

SNPs



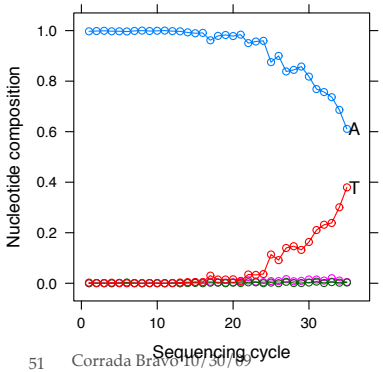
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

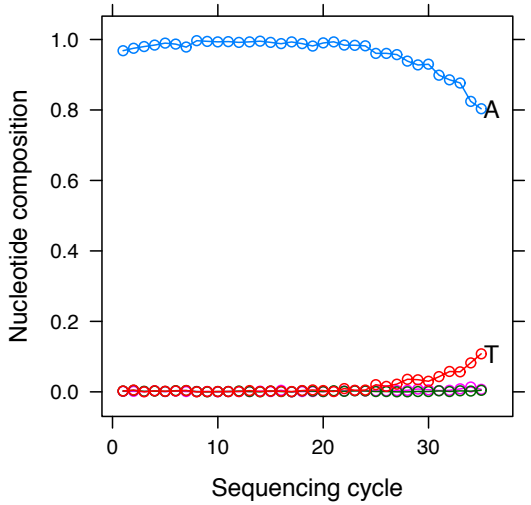
SNPs

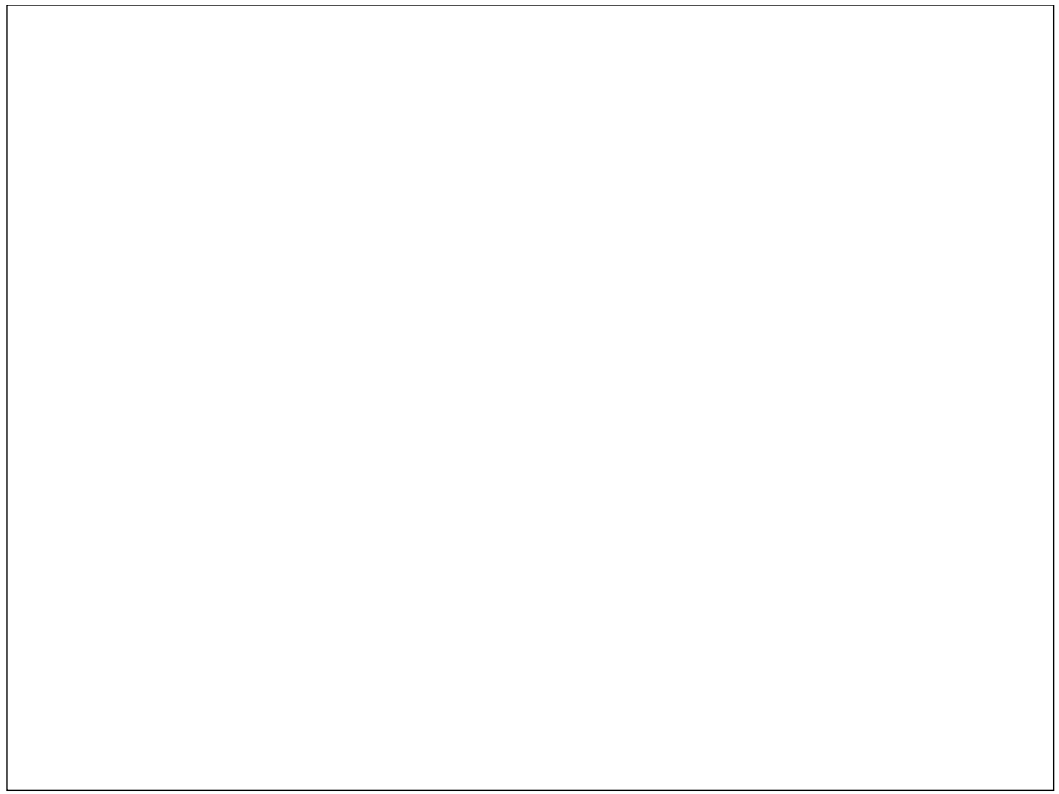


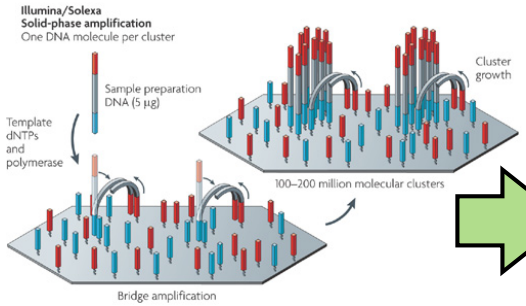
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

SNPs





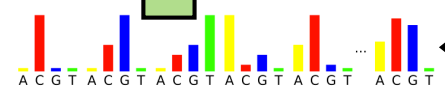




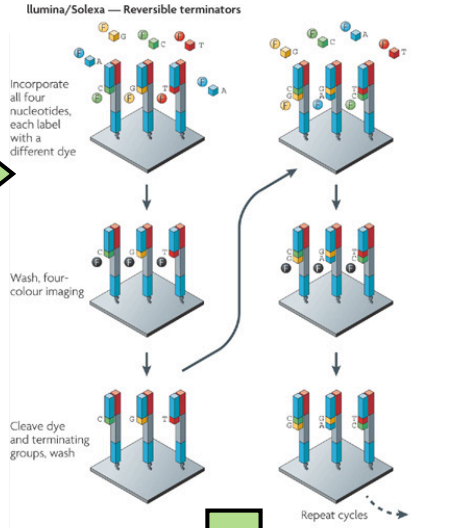
Source: Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010

name
sequence
quality scores

x 100s of millions



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



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

From reads to evidence

From reads to evidence

1. Comparative

Sequence-wise, individuals of a species are nearly identical

Well curated, annotated "reference" genomes exist

D. melanogaster, *Science*, 2000 | *H. sapiens*, *Nature*, 2000 and *Science*, 2000 | *M. musculus*, *Nature*, 2002

Idea: "Map" reads to their point of origin with respect to a reference, then study differences

From reads to evidence

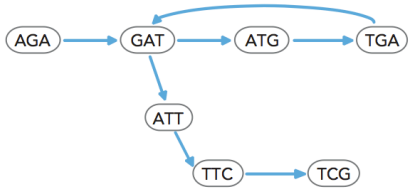
2. *de novo*

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

Mapping

Take a read:

```
CTCAAACTCCTGACCTTTGGTGATCCACCCGCCTNGGCCTTC
```

And a reference sequence:

```
>MT dna:chromosome chromosome:GRCh37:MT:1:16569:1
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC
GCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT
ACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCAC
TTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAAT
CTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATA
CCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAA
GCAATACACTGACCCGCTCAAACTCCTGGATTTTGGATCCACCCAGCGCCTTGGCCTAAA
CTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCCGTTCCAGTGAGT
TCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGCAGCAATGCAGCTC
AAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACCGC
GGTCACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTTTAGATCACCCCC
TCCCCAATAAAGCTAAAACTCACCTGAGTTGTAAAAAACTCCAGTTGACACAAAATAGAC
TACGAAAGTGGCTTTAACATATCTGAACACACAATAGCTAAGACCCAAACTGGGATTAGA
TACCCCACTATGCTTAGCCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAA
CACTACGAGCCACAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATCCCTCTAGAGG
AGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACCTCTTGCTCAGCCTATATA
CCGCCATCTTCAGCAAACCCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCACGTAAAG
ACGTTAGGTCAAGGTGTAGCCCATGAGGTGGCAAGAAATGGGCTACATTTTCTACCCCAG
AAAACTACGATAGCCCTTATGAAACTTAAGGGTCGAAGGTGGATTTAGCAGTAAACTAAG
AGTAGAGTGCTTAGTTGAACAGGGCCCTGAAGCGCGTACACACCGCCCGTCACCCTCCTC
AAGTATACTTCAAAGGACATTTAACTAAAACGGGTAGGGATTTATATAGAGGAGACAAGT
CGTAACCTCAAACTCCTGCCTTTGGTGATCCACCCGCCTTGGCCTACCTGCATAATGAAG
AAGCACCCAACTTACACTTAGGAGATTTCAACTTAACTTGACCGCTCTGAGCTAAACCTA
GCCCCAAACCCACTCCACCTTACTACCAGACAACCTTAGCCAAACCATTTACCCAAATAA
AGTATAGGCGATAGAAATTGAAACCTGGCGCAATAGATATAGTACCGCAAGGGAAAGATG
AAAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTTCTGCATAATGAA
TTAACTAGAAATAACTTTGCAAGGAGAGCCAAAGCTAAGACCCCCGAAACCAGACGAGCT
```

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

Answer: sequence similarity

Hypothesis 1:

```
Read
CTCAAAGACCTGACCTTTGGTGATCCACCC-----GCCTNGGCCTTC
||||||  ||||   ||||  ||||||||||     |||| |||||
CTCAAACTCCTGGATTTTG--GATCCACCCAGCTGGCCTTGGCCTAA
Reference
```

Hypothesis 2:

```
Read
CTCAAACTCCTGACCTTTGGTGATCCACCCGCCTNGGCCTTC
|||||||||||| ||||||||||||||||||||| ||||| |
CTCAAACTCCTG-CCTTTGGTGATCCACCCGCCTTGGCCTAC
Reference
```

Which hypothesis is better?

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

Mapping

Read

CTCAAACTCCTGACCTTTGGTGATCCACCCGCCTNGGCCTTC

Reference

>MT dna:chromosome chromosome:GRCh37:MT:1:16569:1
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC
GCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT
ACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCAC
TTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAAT
CTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATA
CCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAA
GCAATACACTGACCCGCTCAAACTCCTGGATTTTGTGATCCACCCAGCGCCTTGGCCTAA
CTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCCGTTCCAGTGAGT
TCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGCAGCAATGCAGCTC
AAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACCGC
GGTCACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTTTAGATCACCCCC
TCCCCAATAAAGCTAAAACTCACCTGAGTTGTAAAAAACTCCAGTTGACACAAAATAGAC
TACGAAAGTGGCTTTAACATATCTGAACACACAATAGCTAAGACCCAAACTGGGATTAGA
TACCCCACTATGCTTAGCCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAA
CACTACGAGCCACAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATCCCTCTAGAGG
AGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACCTCTTGCTCAGCCTATATA
CCGCCATCTTCAGCAAACCCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCACGTAAAG
ACGTTAGGTCAAGGTGTAGCCCATGAGGTGGCAAGAAATGGGCTACATTTTCTACCCCAG
AAAACTACGATAGCCCTTATGAAACTTAAGGGTCGAAGGTGGATTTAGCAGTAAACTAAG
AGTAGAGTGCTTAGTTGAACAGGGCCCTGAAGCGCGTACACACCGCCCGTCACCCTCCTC
AAGTATACTTCAAAGGACATTTAACTAAAACCCCTACGCATTTATATAGAGGAGACAAGT
CGTAACCTCAAACTCCTGGCCTTTGGTGATCCACCCGCCTTGGCCTACCTGCATAATGAA
AAGCACCCAACTTACACTTAGGAGATTTCAACTTAACTTGACCGCTCTGAGCTAAACCTA
GCCCCAAACCCACTCCACCTTACTACCAGACAACCTTAGCCAAACCATTTACCCAAATAA
AGTATAGGCGATAGAAATTGAAACCTGGCGCAATAGATATAGTACCGCAAGGGAAAGATG
AAAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTTCTGCATAATGAA
TTAACTAGAAATAACTTTGCAAGGAGAGCCAAAGCTAAGACCCCCGAAACCAGACGAGCT
ACCTAAGAACAGCTAAAAGAGCACACCCGTCTATGTAGCAAAATAGTGGGAAGATTTATA
GGTAGAGGCGACAAACCTACCGAGCCTGGTGATAGCTGGTTGTCCAAGATAGAATCTTAG

This is an **alignment**:

Read

```
CTCAAAGACCTGACCTTTGGTGATCCACCC-----GCCTNGGCCTTC
||||||  ||||   ||||  |||||||||     |||| |||||
CTCAAACTCCTGGATTTTG--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:

CTCAAACTCCTGACCTTTGGTGATCCA

And a reference sequence:

```
>MT dna:chromosome chromosome:GRCh37:MT:1:16569:1
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC
GCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT
ACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCAC
TTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAAT
CTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATA
CCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAA
GCAATACACTGACCCGCTCAAACTCCTGGATTTTGTGATCCACCCAGCGCCTTGGCCTAA
CTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCCGTTCCAGTGAGT
TCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGCAGCAATGCAGCTC
AAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACCGC
GGTCACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTTTAGATCACCCCC
TCCCCAATAAAGCTAAAACTCACCTGAGTTGTAAAAAACTCCAGTTGACACAAAATAGAC
TACGAAAGTGGCTTTAACATATCTGAACACACAATAGCTAAGACCCAAACTGGGATTAGA
TACCCCACTATGCTTAGCCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAA
CACTACGAGCCACAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATCCCTCTAGAGG
AGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACCTCTTGCTCAGCCTATATA
CCGCCATCTTCAGCAAACCCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCACGTAAAG
ACGTTAGGTCAAGGTGTAGCCCATGAGGTGGCAAGAAATGGGCTACATTTTCTACCCCAG
AAAACTACGATAGCCCTTATGAAACTTAAGGGTCGAAGGTGGATTTAGCAGTAAACTAAG
AGTAGAGTGCTTAGTTGAACAGGGCCCTGAAGCGCGTACACACCGCCCGTCACCCTCCTC
AAGTATACTTCAAAGGACATTTAACTAAAACCCCTACGCATTTATATAGAGGAGACAAGT
CGTAACCTCAAACTCCTGGCCTTTGGTGATCCACCCGCCTTGGCCTACCTGCATAATGAA
AAGCACCCAACTTACACTTAGGAGATTTCAACTTAACTTGACCGCTCTGAGCTAAACCTA
GCCCCAAACCCACTCCACCTTACTACCAGACAACCTTAGCCAAACCATTTACCCAAATAA
AGTATAGGCGATAGAAATTGAAACCTGGCGCAATAGATATAGTACCGCAAGGGAAAGATG
AAAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTTCTGCATAATGAA
TTAACTAGAAATAACTTTGCAAGGAGAGCCAAAGCTAAGACCCCCGAAACCAGACGAGCT
```

Which hypothesis is better?

Hypothesis 1:

Read

```
CTCAAACTCCTGACCTTTGGTGATCCA
|||||||||||| ||||||||||||||
CTCAAACTCCTGCCCTTTGGTGATCCA
```

Reference

Hypothesis 2:

Read

```
CTCAAACTCCTGACCTTTGGTGATCCA
|||||||||||||||||| ||||||||
CTCAAACTCCTGACCTTTCGTGATCCA
```

Reference

Is there any way to break the tie?

Mapping

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:

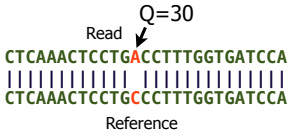
CTCAAACTCCTGACCTTTGGTGATCCA

And a reference sequence:

```
>MT dna:chromosome chromosome:GRCh37:MT:1:16569:1
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC
GCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT
ACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCAC
TTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAAT
CTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATA
CCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAA
GCAATACACTGACCCGCTCAAACTCCTGGATTTTGTGATCCACCCAGCGCCTTGGCCTAA
CTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCCGTTCCAGTGAGT
TCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGCAGCAATGCAGCTC
AAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACCGC
GGTCACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTTTAGATCACCCCC
TCCCCAATAAAGCTAAAACTCACCTGAGTTGTAAAAAACTCCAGTTGACACAAAATAGAC
TACGAAAGTGGCTTTAACATATCTGAACACACAATAGCTAAGACCCAAACTGGGATTAGA
TACCCCACTATGCTTAGCCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAA
CACTACGAGCCACAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATCCCTCTAGAGG
AGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACCTCTTGCTCAGCCTATATA
CCGCCATCTTCAGCAAACCCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCACGTAAAG
ACGTTAGGTCAAGGTGTAGCCCATGAGGTGGCAAGAAATGGGCTACATTTTCTACCCCAG
AAAACTACGATAGCCCTTATGAAACTTAAGGGTCGAAGGTGGATTTAGCAGTAAACTAAG
AGTAGAGTGCTTAGTTGAACAGGGCCCTGAAGCGCGTACACACCGCCCGTCACCCTCCTC
AAGTATACTTCAAAGGACATTTAACTAAAACCCCTACGCATTTATATAGAGGAGACAAGT
CGTAACCTCAAACTCCTGGCCTTTGGTGATCCACCCGCCTTGGCCTACCTGCATAATGAA
AAGCACCCAACTTACACTTAGGAGATTTCAACTTAACTTGACCGCTCTGAGCTAAACCTA
GCCCCAAACCCACTCCACCTTACTACCAGACAACCTTAGCCAAACCATTTACCCAAATAA
AGTATAGGCGATAGAAATTGAAACCTGGCGCAATAGATATAGTACCGCAAGGGAAAGATG
AAAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTTCTGCATAATGAA
TTAACTAGAAATAACTTTGCAAGGAGAGCCAAAGCTAAGACCCCCGAAACCAGACGAGCT
```

Which hypothesis is better?

Hypothesis 1:



Hypothesis 2:

Mapping

Take a read:

CTCAAACTCCTGACCTTTGGTGATCCA

And a reference sequence:

```
>MT dna:chromosome chromosome:GRCh37:MT:1:16569:1
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC
GCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT
ACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCAC
TTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAAT
CTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATA
CCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAA
GCAATACACTGACCCGCTCAAACTCCTGGATTTTGTGATCCACCCAGCGCCTTGGCCTAA
CTAGCCTTTCTATTAGCTCTTAGTAAGATTACACATGCAAGCATCCCCGTTCCAGTGAGT
TCACCCTCTAAATCACCACGATCAAAAGGAACAAGCATCAAGCACGCAGCAATGCAGCTC
AAAACGCTTAGCCTAGCCACACCCCCACGGGAAACAGCAGTGATTAACCTTTAGCAATAA
ACGAAAGTTTAACTAAGCTATACTAACCCCAGGGTTGGTCAATTTCGTGCCAGCCACCGC
GGTCACACGATTAACCCAAGTCAATAGAAGCCGGCGTAAAGAGTGTTTTAGATCACCCCC
TCCCCAATAAAGCTAAAACTCACCTGAGTTGTAAAAAACTCCAGTTGACACAAAATAGAC
TACGAAAGTGGCTTTAACATATCTGAACACACAATAGCTAAGACCCAAACTGGGATTAGA
TACCCCACTATGCTTAGCCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAA
CACTACGAGCCACAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATCCCTCTAGAGG
AGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACCTCTTGCTCAGCCTATATA
CCGCCATCTTCAGCAAACCCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCACGTAAAG
ACGTTAGGTCAAGGTGTAGCCCATGAGGTGGCAAGAAATGGGCTACATTTTCTACCCCAG
AAAACTACGATAGCCCTTATGAAACTTAAGGGTCGAAGGTGGATTTAGCAGTAAACTAAG
AGTAGAGTGCTTAGTTGAACAGGGCCCTGAAGCGCGTACACACCGCCCGTCACCCTCCTC
AAGTATACTTCAAAGGACATTTAACTAAAACCCCTACGCATTTATATAGAGGAGACAAGT
CGTAACCTCAAACTCCTGGCCTTTGGTGATCCACCCGCCTTGGCCTACCTGCATAATGAA
AAGCACCCAACTTACACTTAGGAGATTTCAACTTAACTTGACCGCTCTGAGCTAAACCTA
GCCCCAAACCCACTCCACCTTACTACCAGACAACCTTAGCCAAACCATTTACCCAAATAA
AGTATAGGCGATAGAAATTGAAACCTGGCGCAATAGATATAGTACCGCAAGGGAAAGATG
AAAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTTCTGCATAATGAA
TTAACTAGAAATAACTTTGCAAGGAGAGCCAAAGCTAAGACCCCCGAAACCAGACGAGCT
```

Which hypothesis is better?

Hypothesis 1:

Read

```
CTCAAACTCCTGACCTTTGGTGATCCA
|||||||||||| ||||||||||||||
CTCAAACTCCTG-CCTTTGGTGATCCA
```

Reference

Hypothesis 2:

Read

```
CTCAAACTCCTGACCTTTGGTGATCCA
|||||||||||||||||| ||||||||
CTCAAACTCCTGACCTTTCGTGATCCA
```

Reference

Is there any way to break the tie?

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

Mapping

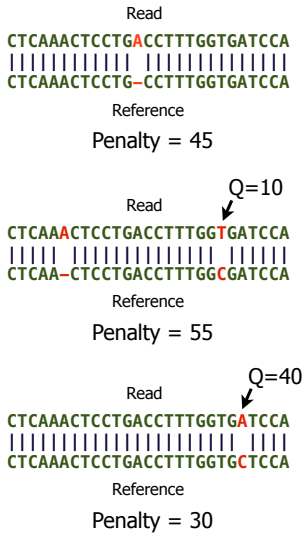
Aligners can employ **penalties** to account for the relative probability of seeing different dissimilarities

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 _\textrm{gap}& \equiv -10 \log _{10} (P_\textrm{gap}) \\
&= -10 \log _{10} (0.00005) \\
&\approx 45
\end{align*}

\begin{align*}
Pen _\textrm{mm}& \equiv \arg\min(-10 \log _{10} (P_\textrm{miscall}), -10 \log _
{10} (P_\textrm{SNP})) \\
&= \arg\min(Q, -10 \log _{10} (0.001)) \\
&= \arg\min(Q, 30) \\
\end{align*}
```