Introduction to second-generation sequencing

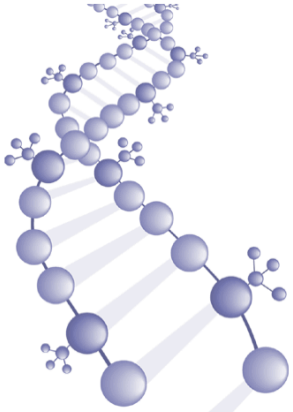
CMSC858B Spring 2012
Many slides courtesy of Ben Langmead

Second-Generation Sequencing

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Human Epigenome Project

HEP
Human
Epigenome
Project

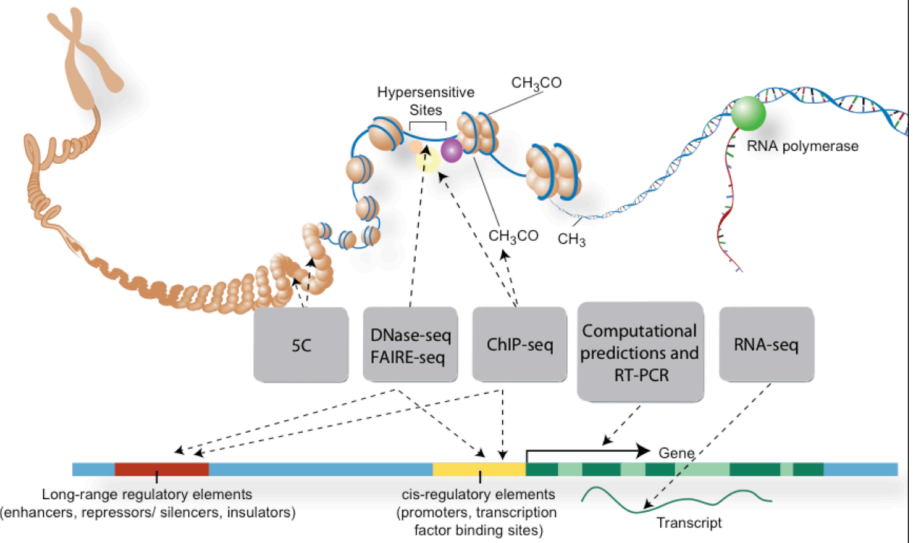


Methylation

epigenomics | The Wellcome Trust Sanger Institute | CNG

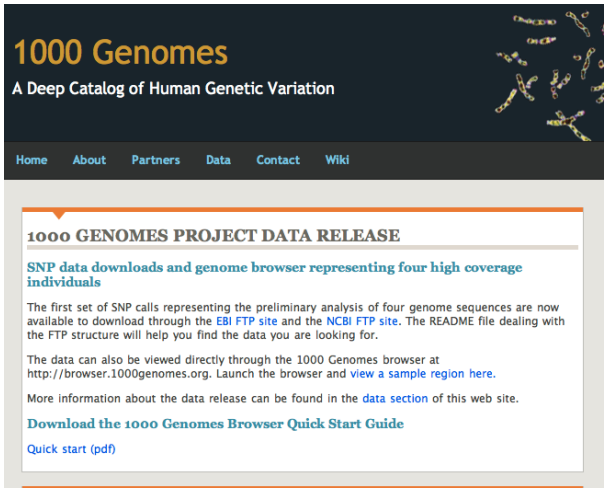
JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

ENCODE project



http://www.genome.gov/10005107

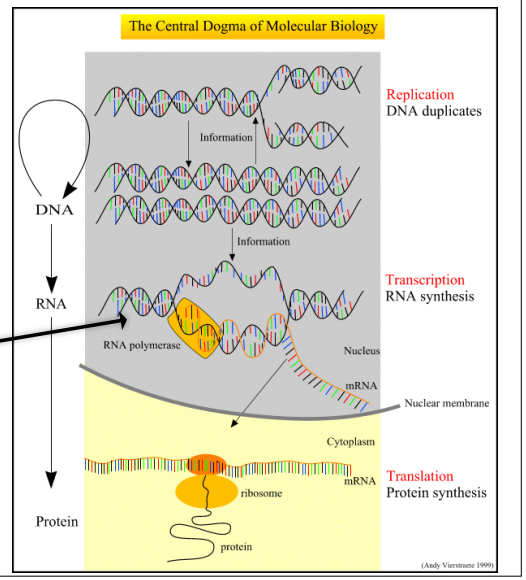
1000 Genomes Project



What Do They Measure?

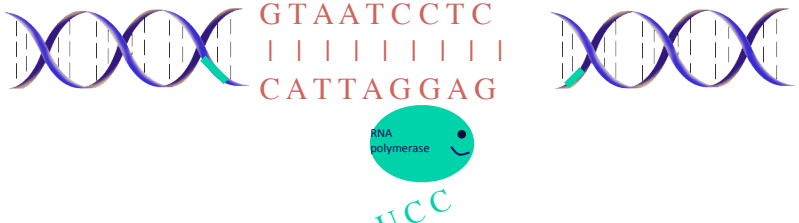
Microarrays

Gene Expression Arrays
Exon Arrays



Transcription

DNA

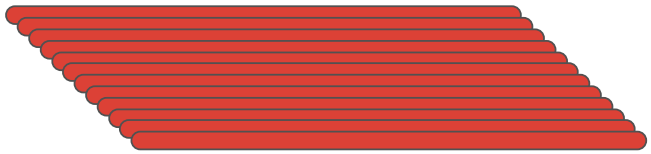


From DNA to mRNA

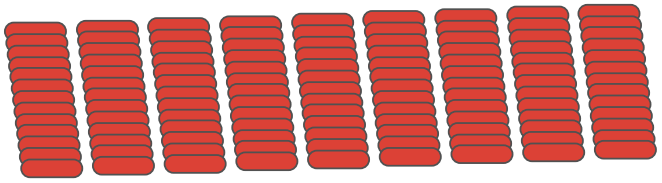
Reverse transcription

Clone cDNA strands, complementary to the mRNA

Sec-gen Sequencing

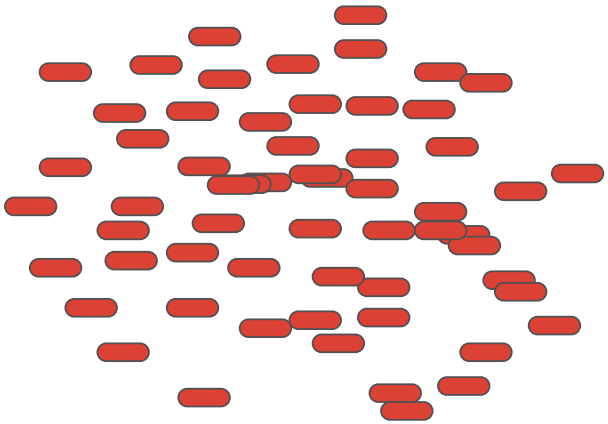


Sec-gen Sequencing



Fragmentation is random,
i.e., not equal-sized (but hard to draw)

Sec-gen Sequencing



Second-Generation Sequencing

- "Ultra high throughput" DNA sequencing
  - 3 gigabases / day vs.
  - 3 gigabases / 13 years (human genome project, more or less)

Platforms

illumina

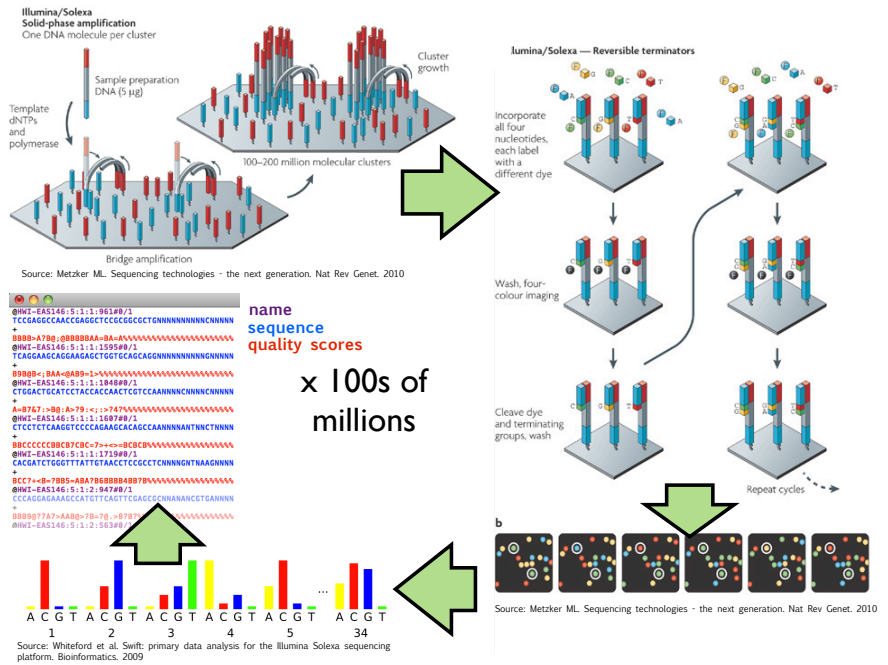
life technologies | Semiconductor Sequencing for Life™ ion torrent

454 SEQUENCING | Roche

NANOPORE

Applied Biosystems

- Millions of short DNA fragments (~100 bp) sequenced in parallel



Sequencing throughput

GA II
1.6 billion bp per day
(2008)

GA IIx
5 billion bp per day
(2009)

HiSeq 2000
25 billion bp per day
(2010)

Images: www.illumina.com/systems
Numbers: www.politigenomics.com/next-generation-sequencing-informatics
Dates: Illumina press releases

Sequencing throughput

GA II
1.6 billion bp per day
(2008)

GA IIx
5 billion bp per day
(2009)

HiSeq 2500
60 billion bp per day
(2012)

Images: www.illumina.com/systems
Numbers: www.politigenomics.com/next-generation-sequencing-informatics
Dates: Illumina press releases

Sequencing throughput

End of 2009 Mid 2010 Late 2011/Early 2012

SOLiD 3+ System
Up to 50 Gb

New SOLiD™ Systems		
Specifications	SOLiD™ 4 System	SOLiD™ 4*hq* System*
Throughput/Run**	Up to 100 Gb	Up to 300 Gb
System Accuracy	99.94%	99.99%
Cost/Genome***	As low as $6,000	As low as $3,000
Read Length	• Fragment: 50 bp • Mate-pair: 2 x 50 bp • Paired-end: 50 x 25 bp	• Fragment: 75 bp • Mate-pair: 2 x 75 bp • Paired-end: 75 x 35 bp
Multiplexing	• 48 RNA barcodes • 96 DNA barcodes	• 96 RNA barcodes • 96 DNA barcodes
Run Time	• 3 days for 35 bp • 11 days for 50 x 25 bp • 12 days for 2 x 50 bp	• 3 days for 35 bp • 12 days for 75 x 35 bp • 14 days for 2 x 75 bp

* These systems are under development and the specifications are subject to change.
** Expected throughput.
*** Prices reflect US list at optimal running efficiencies. Regional pricing may vary.

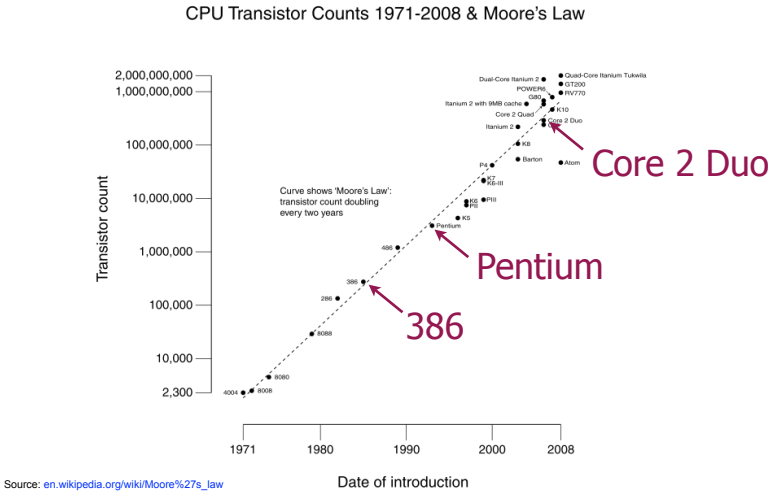
Source: www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_061241.pdf

Computational throughput

Moore's Law:

The number of transistors that can be placed inexpensively on an integrated circuit doubles approximately every two years.

Computational throughput

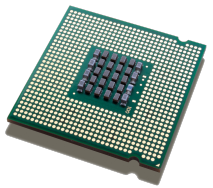


Source: en.wikipedia.org/wiki/Moore%27s_law

Throughput growth gap

4-5x per year

>



2x per 2 years

ionTorrent

Ion Semiconductor Sequencing

Speed, Scalability, Simplicity

Ion semiconductor benchtop sequencers perform real-time measurements of hydrogen ions produced during DNA replication. Ion semiconductor chips employ a massively parallel array of semiconductor sensors, to directly translate genetic information (DNA) to digital information (DNA sequence). This groundbreaking technology enables rapid and scalable sequencing across a range of applications.

Technology

Ion PGM™ Sequencer

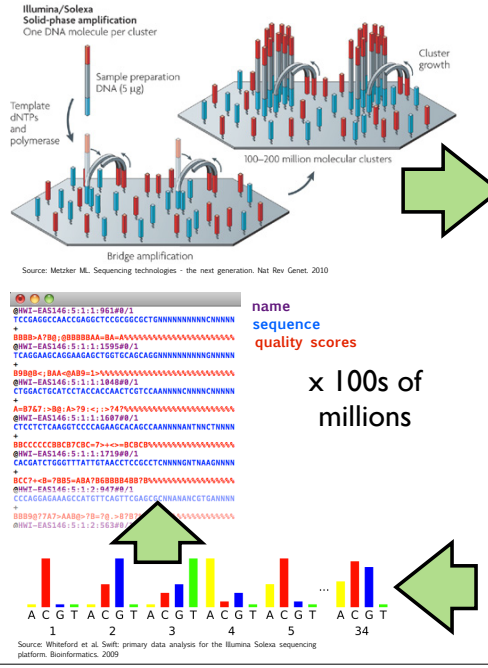
The fastest and most affordable benchtop sequencer

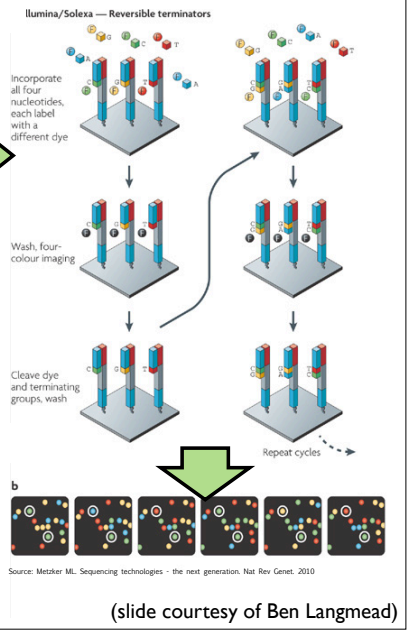
Ion Proton™ Sequencer

Genome and exome sequencing on your benchtop

Oxford Nanopore

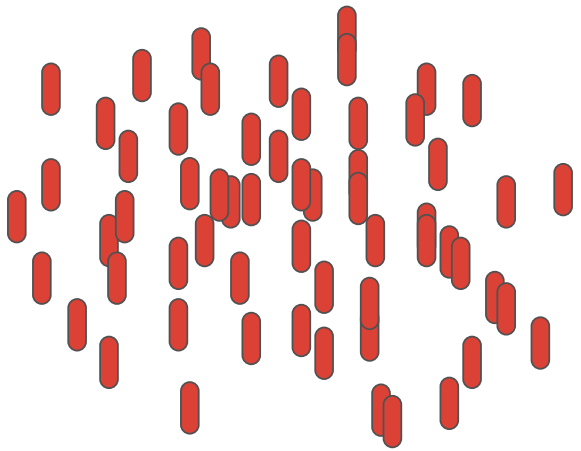
- Nanopore technology
- ultralong reads (48kb genome sequenced as one read)





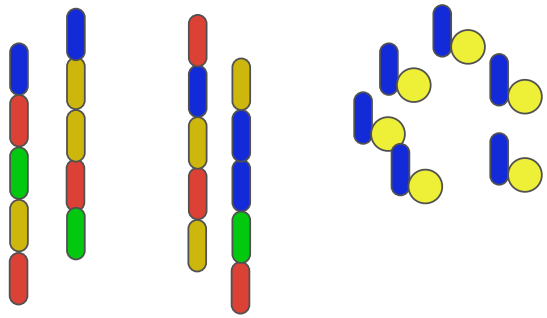
(slide courtesy of Ben Langmead)

Sec-gen Sequencing

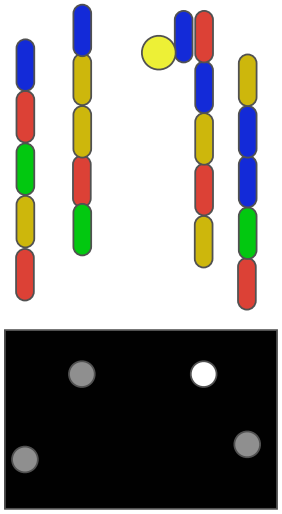


JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

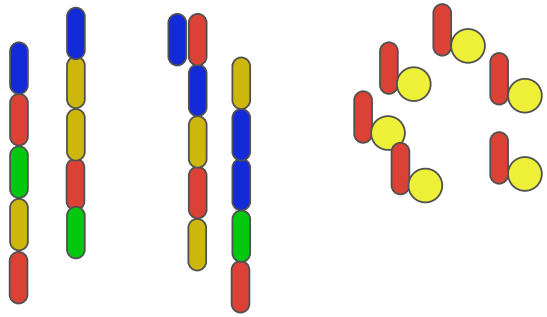
Sec-gen Sequencing



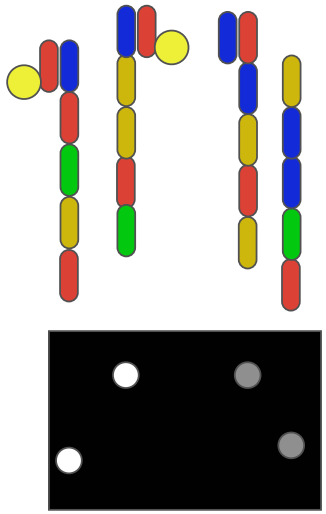
Sec-gen Sequencing



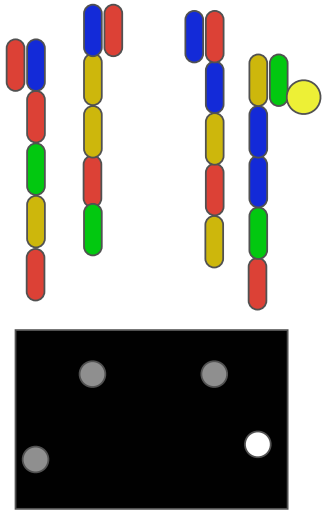
Sec-gen Sequencing



Sec-gen Sequencing



Sec-gen Sequencing



Sec-gen Sequencing

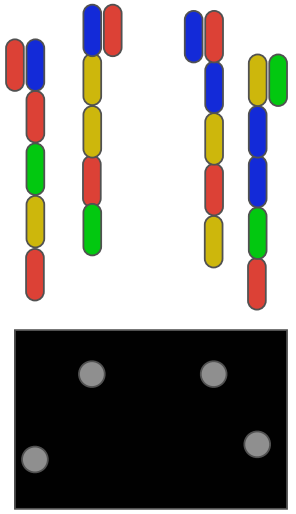
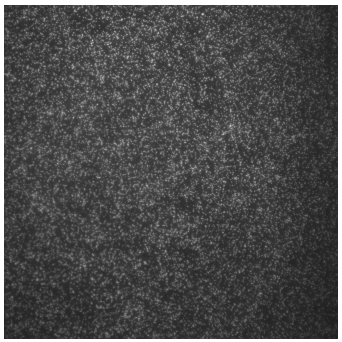
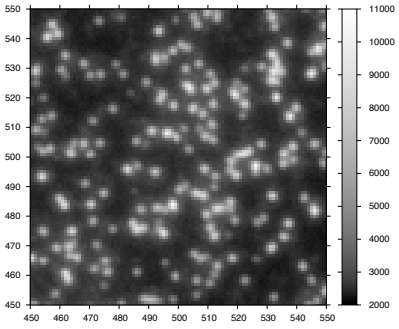


Image Analysis



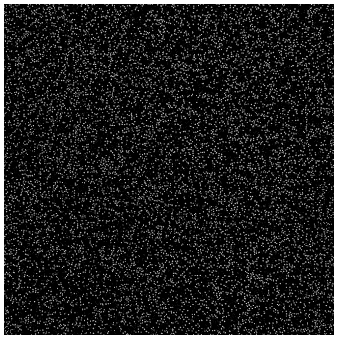


An input image and zoomed in section

Image Analysis

- 4 images per cycle
- ~100 tiles
- Analysis:
  - Filtering
  - Background subtraction
  - Thresholding
- Each image analysis independent (so can parallelize)

Image Analysis



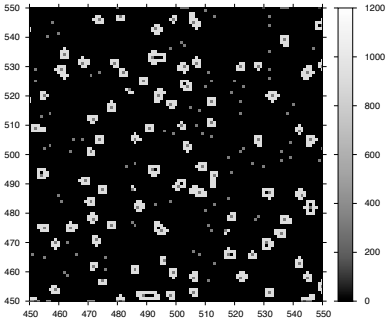
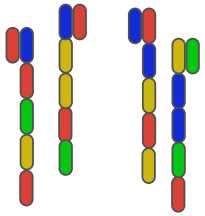
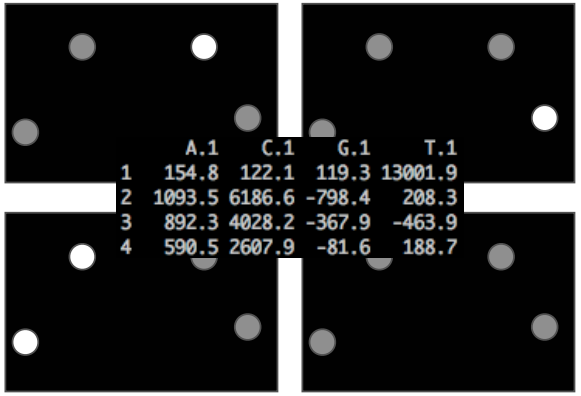


Image after processing. *This is old, cluster density is much higher now*

Sec-gen Sequencing

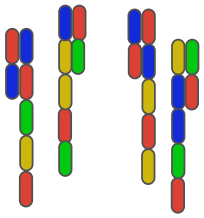


First Cycle

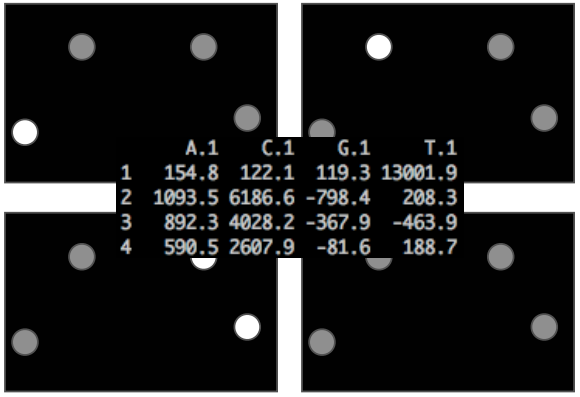


JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Sec-gen Sequencing

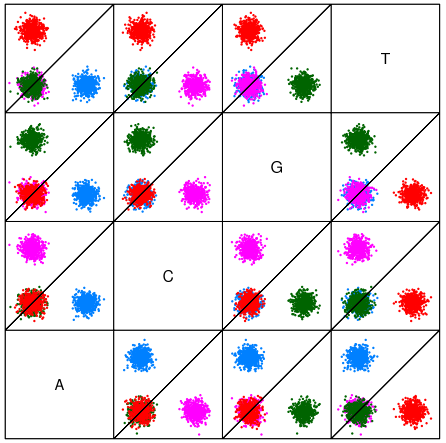


Second Cycle



JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

A Thought Experiment

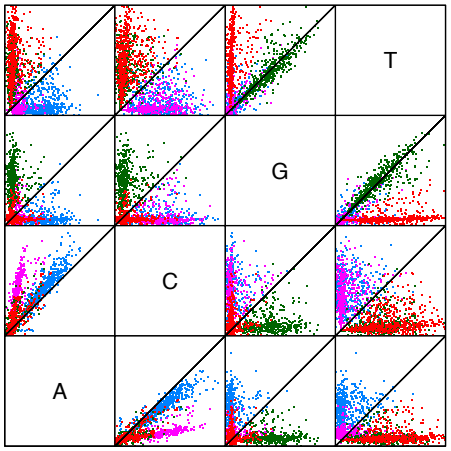


Four-channel fluorescence intensity, cycle 1

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

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Fluorescence Intensity



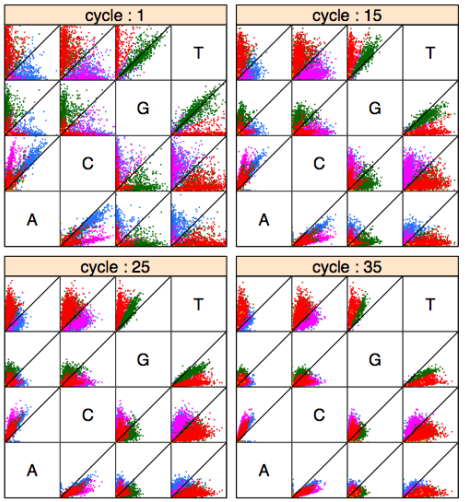
Four–channel fluorescence intensity, cycle 1

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

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

Fluorescence Intensity

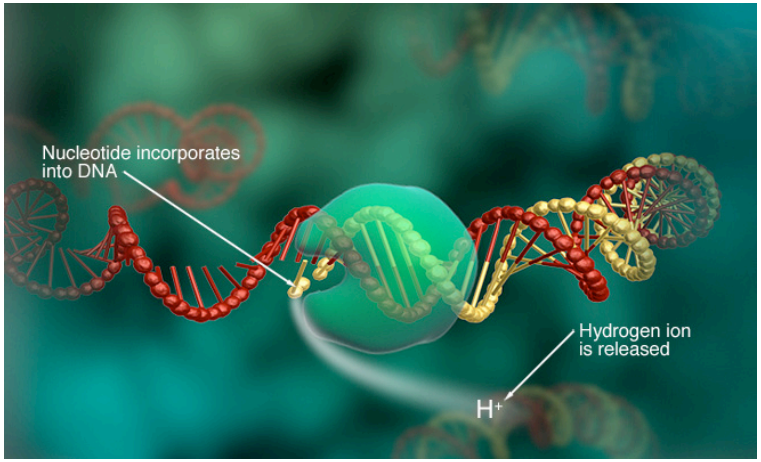
MORE ON THIS LATER IN THE COURSE!



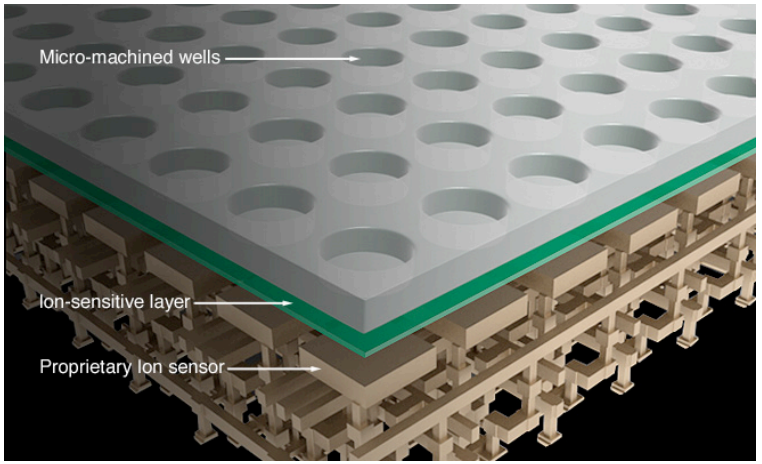
Four–channel fluorescence intensity

JOHNS HOPKINS BLOOMBERG SCHOOL of PUBLIC HEALTH

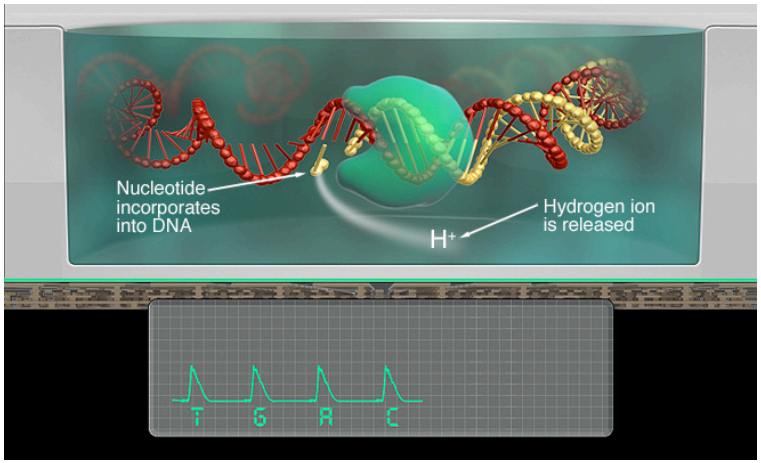
Ion Torrent



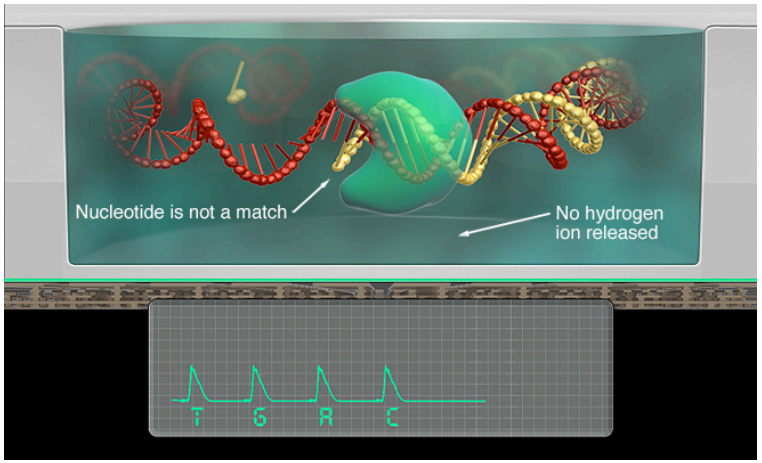
Ion Torrent



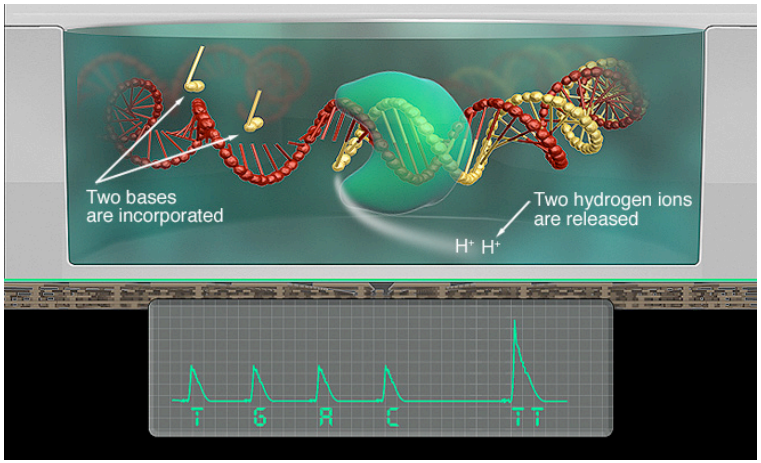
Ion Torrent

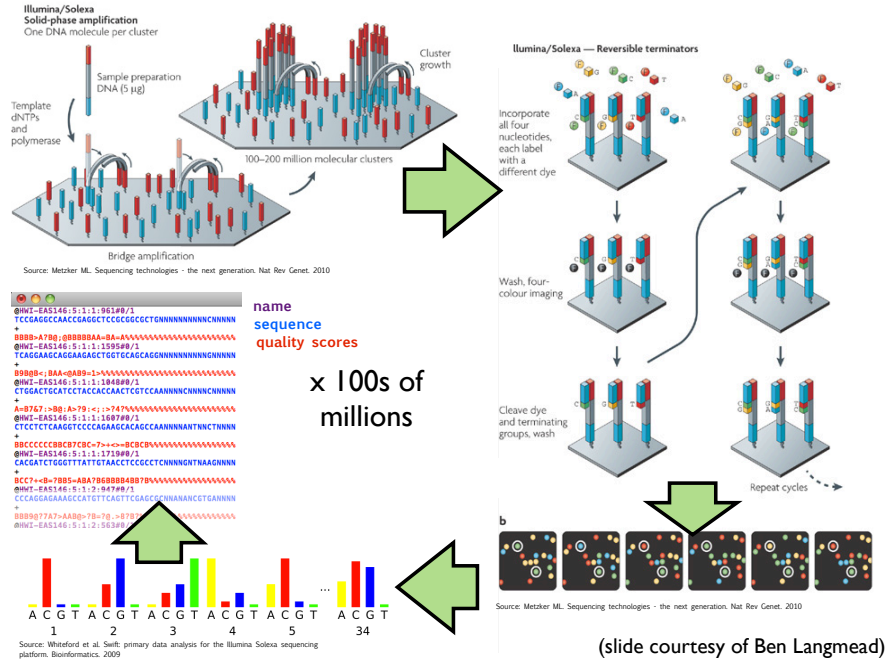


Ion Torrent



Ion Torrent

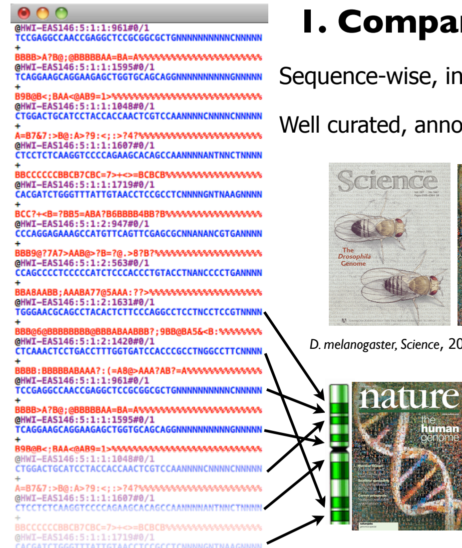




(slide courtesy of Ben Langmead)

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

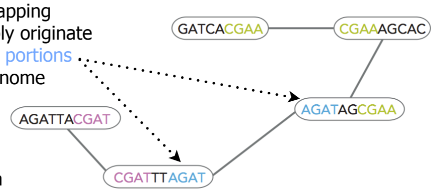
From reads to evidence

2. *de novo*

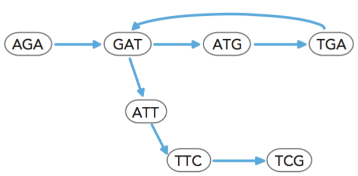
Assume nothing! - let reads tell us everything

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

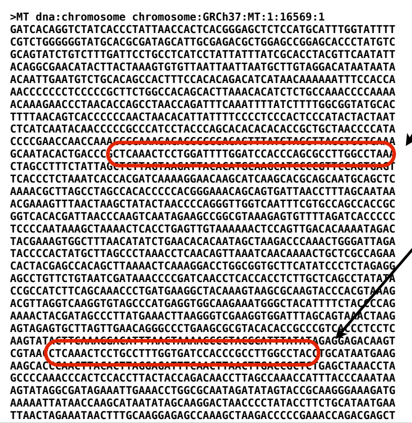
Source: De Novo Assembly Using Illumina Reads. Illumina. 2010
http://www.illumina.com/Documents/products/technotes/technote_denovo_assembly.pdf

Mapping

Take a read:

```
CTCAAACTCCTGACCTTTGGTGATCCACCCGCCTNGGCCTTC
```

And a reference sequence:



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
TACCCCACTATGCTTAGCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAA
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
TACCCCACTATGCTTAGCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAA
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
TACCCCACTATGCTTAGCCTAAACCTCAACAGTTAAATCAACAAAACTGCTCGCCAGAA
CACTACGAGCCACAGCTTAAAACTCAAAGGACCTGGCGGTGCTTCATATCCCTCTAGAGG
AGCCTGTTCTGTAATCGATAAACCCCGATCAACCTCACCACCTCTTGCTCAGCCTATATA
CCGCCATCTTCAGCAAACCCTGATGAAGGCTACAAAGTAAGCGCAAGTACCCACGTAAAG
ACGTTAGGTCAAGGTGTAGCCCATGAGGTGGCAAGAAATGGGCTACATTTTCTACCCCAG
AAAACTACGATAGCCCTTATGAAACTTAAGGGTCGAAGGTGGATTTAGCAGTAAACTAAG
AGTAGAGTGCTTAGTTGAACAGGGCCCTGAAGCGCGTACACACCGCCCGTCACCCTCCTC
AAGTATACTTCAAAGGACATTTAACTAAAACCCCTACGCATTTATATAGAGGAGACAAGT
CGTAACCTCAAACTCCTGGCCTTTGGTGATCCACCCGCCTTGGCCTACCTGCATAATGAA
AAGCACCCAACTTACACTTAGGAGATTTCAACTTAACTTGACCGCTCTGAGCTAAACCTA
GCCCCAAACCCACTCCACCTTACTACCAGACAACCTTAGCCAAACCATTTACCCAAATAA
AGTATAGGCGATAGAAATTGAAACCTGGCGCAATAGATATAGTACCGCAAGGGAAAGATG
AAAAATTATAACCAAGCATAATATAGCAAGGACTAACCCCTATACCTTCTGCATAATGAA
TTAACTAGAAATAACTTTGCAAGGAGAGCCAAAGCTAAGACCCCCGAAACCAGACGAGCT

Which hypothesis is better?

Hypothesis 1:

Q=30

Read

CTCAAACTCCTGACCTTTGGTGATCCA
CTCAAACTCCTGCCCTTTGGTGATCCA

Reference

Hypothesis 2:

Q=10

Read

CTCAAACTCCTGACCTTTGGTGATCCA
CTCAAACTCCTGACCTTTCGTGATCCA

Reference

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

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

Read

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

Reference

Penalty = 45

Q=10

Read

```
CTCAAACTCCTGACCTTTGGTGATCCA
||||| |||||||||||||| ||||||
CTCAA-CTCCTGACCTTTGGCGATCCA
```

Reference

Penalty = 55

Q=40

Read

```
CTCAAACTCCTGACCTTTGGTGATCCA
|||||||||||||||||||||| ||||
CTCAAACTCCTGACCTTTGGTGCTCCA
```

Reference

Penalty = 30

Resources

- Bowtie: ultra-fast mapping of short reads to reference genome

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 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: t GB for the human genome (2.9 GB for paired-end).

- http://bowtie-bio.sourceforge.net

RNA-seq differential expression