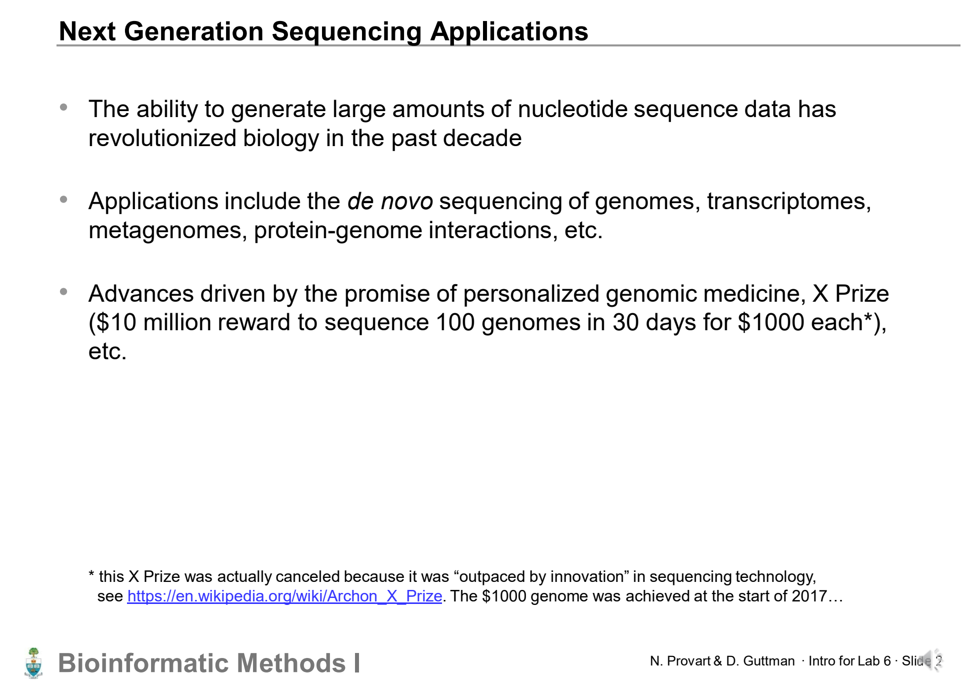
1



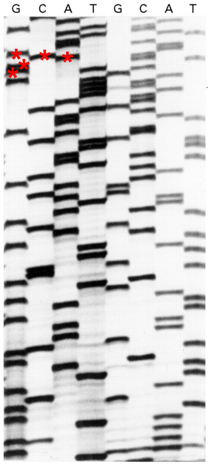
Next Generation Sequencing Applications

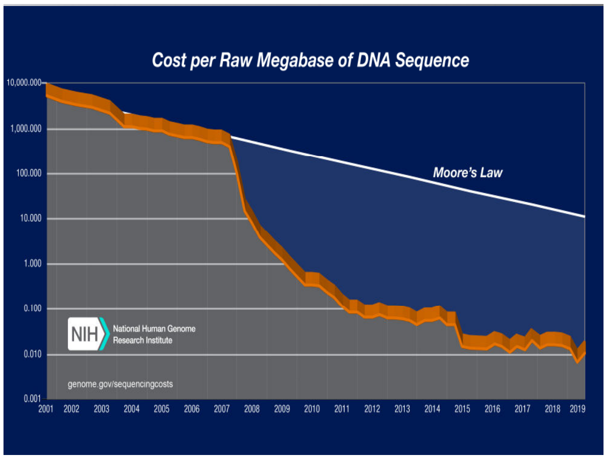
- The ability to generate large amounts of nucleotide sequence data has revolutionized biology in the past decade
- Applications include the *de novo* sequencing of genomes, transcriptomes, metagenomes, protein-genome interactions, etc.
- Advances driven by the promise of personalized genomic medicine, X Prize ($10 million reward to sequence 100 genomes in 30 days for $1000 each*), etc.

* this X Prize was actually canceled because it was "outpaced by innovation" in sequencing technology, see https://en.wikipedia.org/wiki/Archon_X_Prize. The $1000 genome was achieved at the start of 2017…

2

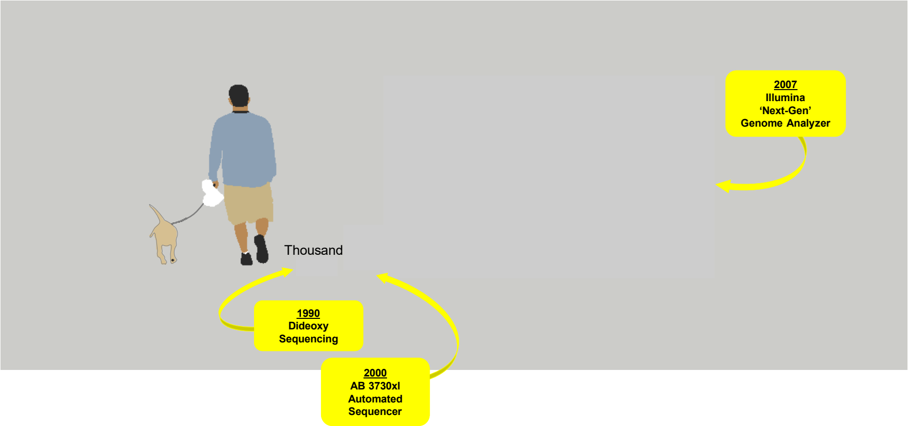
Sequencing Rates and Costs





https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Sequencing Rates and Costs

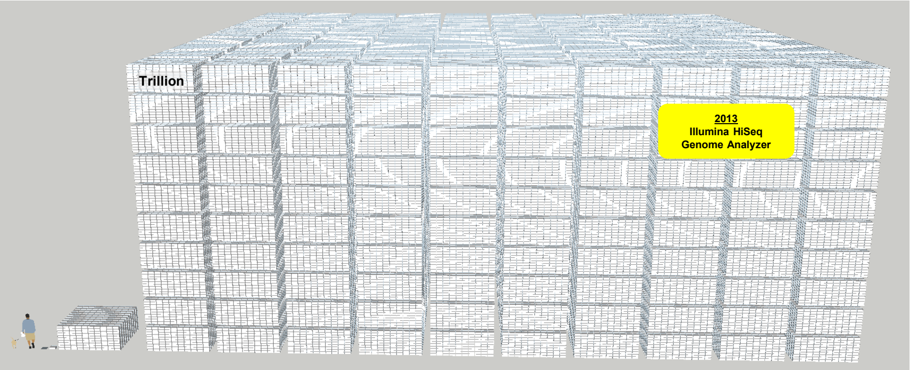


2007
Illumina
'Next-Gen'
Genome Analyzer

1990
Dideoxy
Sequencing

2000
AB 3730xl
Automated
Sequencer

Sequencing Rates and Costs

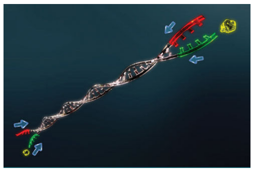
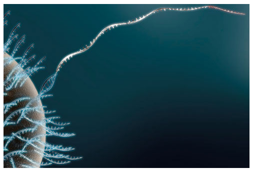


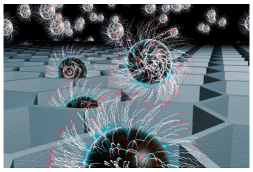
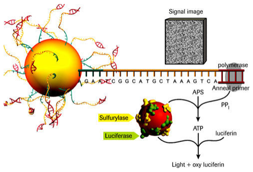
Sequencing Rates and Costs

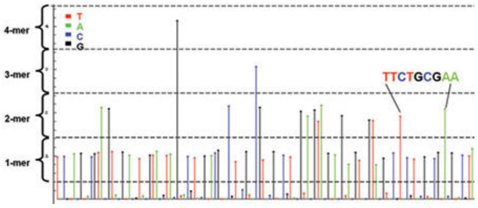
Human Gut Microbiome		Sanger
Number of Species	1000	
Average Genome Size	3 Mb	
Microbiome Size	3 Gb	
Desired Coverage	300 x	
Amount of Data Needed	~ 1 Tb	
Read Length		750bp
Number of **Runs** Needed		~14M
Cost		$667B

Platform	Reads	Read Length (bases)	Paired Ends	Run Time (days)	Yield (Gb)	Rate (days/Gb)
Sanger	96	750	No	0.5	0.00007	~7000 days

Roche 454 Pyrosequencing

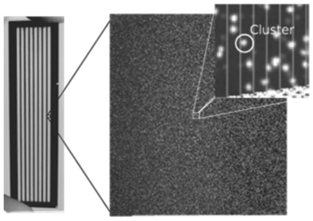



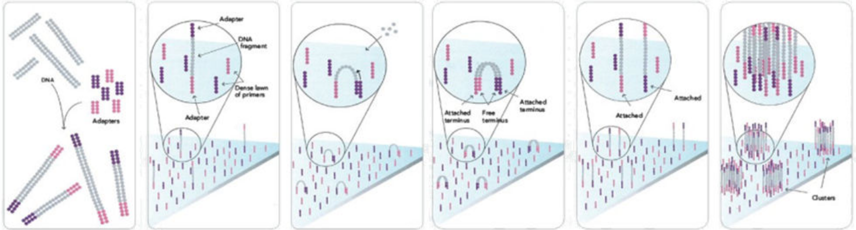





Images from 454.com website

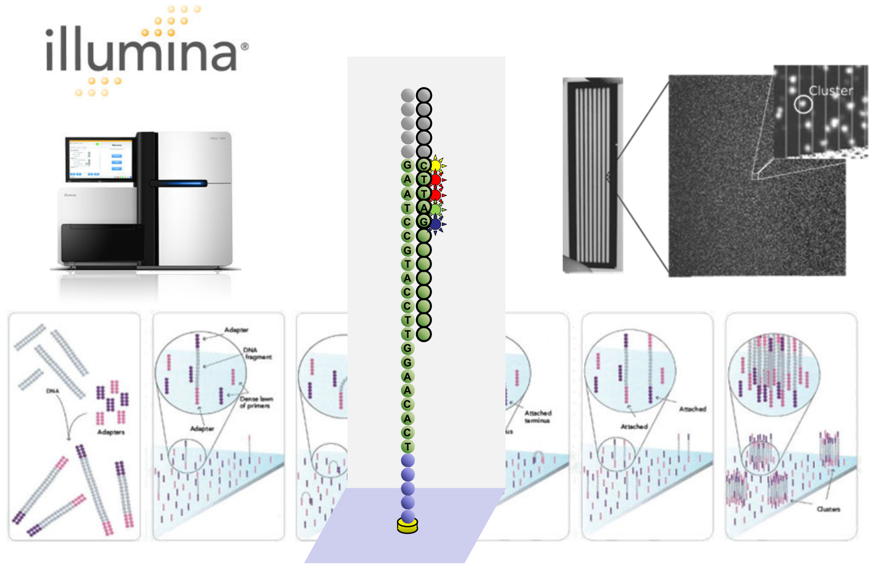
Illumina Sequencing by Synthesis





Images from Illumina.com website

Illumina Sequencing by Synthesis



Images from Illumina.com website

Next Generation Sequencing Technologies

- **Illumina GAIIx, HiSeq2000 + newer machines (short reads, sequencing by synthesis)**
- **Roche 454 FLX, GS Junior (short reads, pyrosequencing)** – *now discontinued*
- **ABI SOLiD (short reads, sequencing by ligation)** – *good for SNP calling*

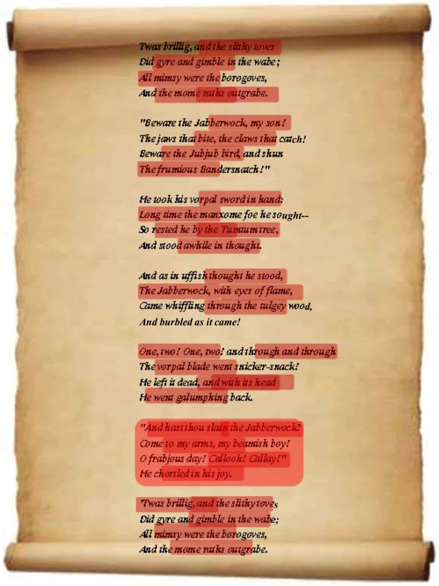
- PacBio (long reads*, real time single molecule sequencing)
- Oxford Nanopore (long reads*, nanopore sequencing)

- MGI DNBSEQ (short reads, DNA nanoball rolling circle sequencing)

* Long reads can really help with the assembly of complex eukaryotic genomes and transcriptomes, in a manner analogous to paired-end sequencing on slide 19, but in the case of these reads the intervening sequence is known!

NGS Assembly

- Cut poem into short “reads”
- Each part of poem is read multiple times
- Assemble poem by assembling overlapping reads



Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

"Beware the Jabberwock, my son!
The jaws that bite, the claws that catch!
Beware the Jubjub bird, and shun
The frumious Bandersnatch!"

He took his vorpal sword in hand:
Long time the manxome foe he sought--
So rested he by the Tumtum tree,
And stood awhile in thought.

And as in uffish thought he stood,
The Jabberwock, with eyes of flame,
Came whiffling through the tulgey wood,
And burbled as it came!

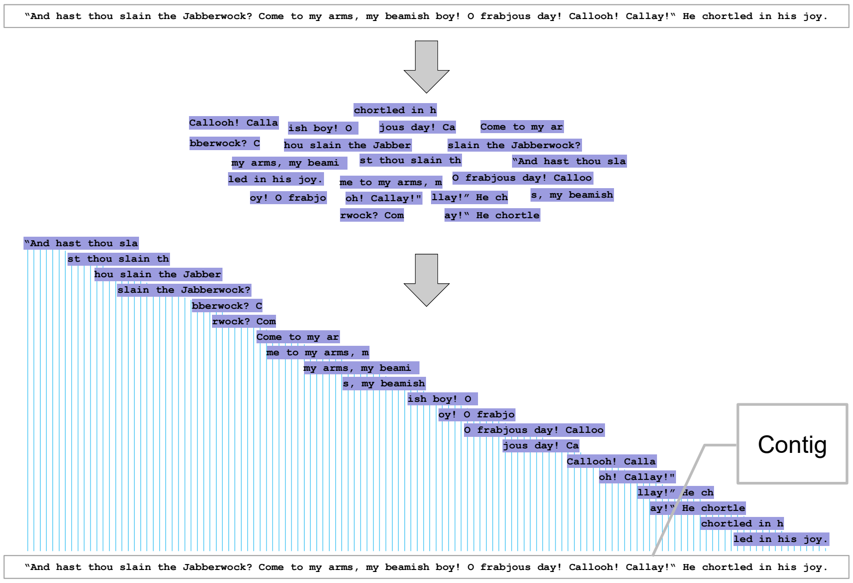
One, two! One, two! and through and through
The vorpal blade went snicker-snack!
He left it dead, and with its head
He went galumphing back.

"And hast thou slain the Jabberwock?
Come to my arms, my beamish boy!
O frabjous day! Callooh! Callay!"
He chortled in his joy.

'Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

“Jabberwocky” by Lewis Carroll, 1872

Assembly



“And hast thou slain the Jabberwock? Come to my arms, my beamish boy! O frabjous day! Callooh! Callay!“ He chortled in his joy.

chortled in h
Callooh! Calla
ish boy! O
jous day! Ca
Come to my ar
bberwock? C
hou slain the Jabber
slain the Jabberwock?
my arms, my beami
st thou slain th
“And hast thou sla
led in his joy.
me to my arms, m
O frabjous day! Calloo
oy! O frabjo
oh! Callay!"
llay!” He ch
s, my beamish
rwock? Com
ay!“ He chortle

“And hast thou sla
st thou slain th
hou slain the Jabber
slain the Jabberwock?
bberwock? C
rwock? Com
Come to my ar
me to my arms, m
my arms, my beami
s, my beamish
ish boy! O
oy! O frabjo
O frabjous day! Calloo
jous day! Ca
Callooh! Calla
oh! Callay!"
llay!” He ch
ay!“ He chortle
chortled in h
led in his joy.

“And hast thou slain the Jabberwock? Come to my arms, my beamish boy! O frabjous day! Callooh! Callay!“ He chortled in his joy.

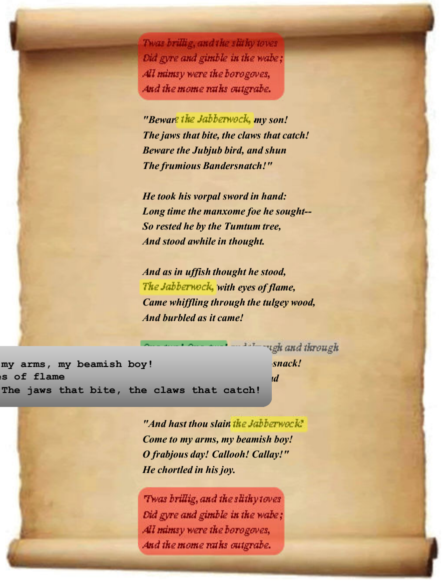
Assemby – Strategy

- Cut poem into short “reads”
- Each part of poem is read multiple times
- Assemble poem by assembling overlapping reads

Problem

- Repetitive regions

```
          “And hast thou slain             ? Come to my arms, my beamish boy!
And as in uffish thought he stood,  >  the Jabberwock  <  , with eyes of flame
                          "Beware             , my son! The jaws that bite, the claws that catch!
```



Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

"Beware the Jabberwock, my son!
The jaws that bite, the claws that catch!
Beware the Jubjub bird, and shun
The frumious Bandersnatch!"

He took his vorpal sword in hand:
Long time the manxome foe he sought--
So rested he by the Tumtum tree,
And stood awhile in thought.

And as in uffish thought he stood,
The Jabberwock, with eyes of flame,
Came whiffling through the tulgey wood,
And burbled as it came!

ugh and through
snack!
d

"And hast thou slain the Jabberwock?
Come to my arms, my beamish boy!
O frabjous day! Callooh! Callay!"
He chortled in his joy.

Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

Assembly – Challenges

- 50 - 1000 Epic Poems (for human microbiome)
- Each poem is 500,000 – 10,000,000 characters long
- Each poem is written in an alphabet of only four letters
- Some poems are present in billions of copies, while others only present in only tens of copies
- Each poem is randomly cut up into “reads” ranging in length from 75 – 500 characters long
- There are as many as a billion such reads present in no particular order

Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

"Beware the Jabberwock, my son!
The jaws that bite, the claws that catch!
Beware the Jubjub bird, and shun
The frumious Bandersnatch!"

He took his vorpal sword in hand:
Long time the manxome foe he sought--
So rested he by the Tumtum tree,
And stood awhile in thought.

And as in uffish thought he stood,

Twas brillig, and the slithy toves
Did gyre and gimble in the wabe;
All mimsy were the borogoves,
And the mome raths outgrabe.

Mapping Reads to Genomes - Overview

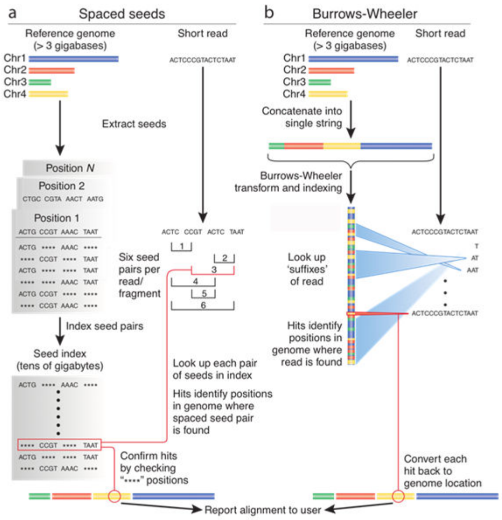


Image from Trapnell & Salzberg (2009) Nature Biotech. 27:455-457.

De novo Genome Assembly

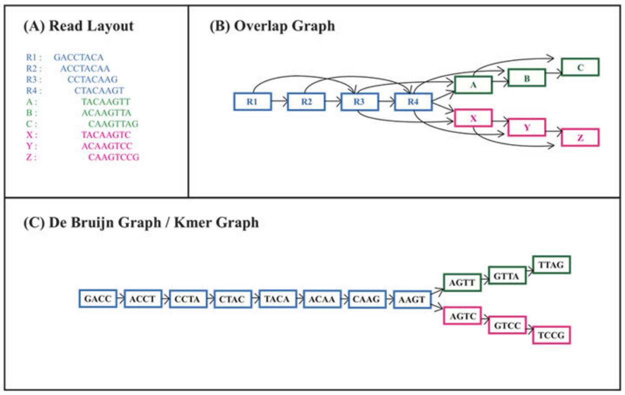


Image courtesy of Jessica Yang, M.Sc. Thesis (2011), adapted from Schatz et al. (2010) Genome Res. 20:1165-73.

De novo Genome Assembly, e.g. Velvet

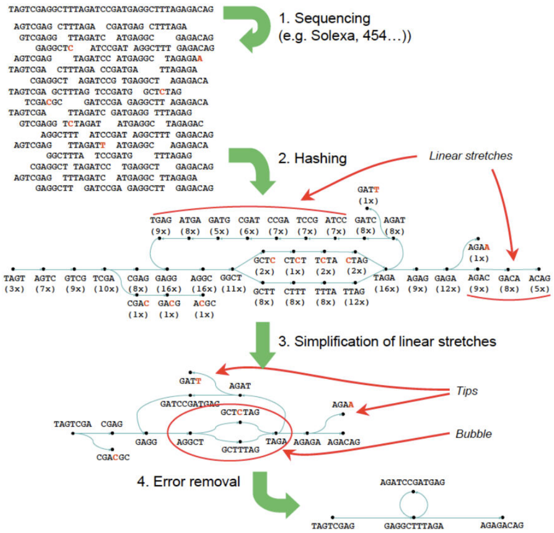

TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG
AGTCGAG CTTTAGA CGATGAG CTTTAGA
GTCGAGG TTAGATC ATGAGGC GAGACAG
GAGGCTC ATCCGAT AGGCTTT GAGACAG
AGTCGAG TAGATCC ATGAGGC TAGAGAA
TAGTCGA CTTTAGA CCGATGA TTAGAGA
CGAGGCT AGATCCG TGAGGCT AGAGACA
TAGTCGA GCTTTAG TCCGATG GCTCTAG
TCGACGC GATCCGA GAGGCTT AGAGACA
TAGTCGA TTAGATC GATGAGG TTTAGAG
GTCGAGG TCTAGAT ATGAGGC TAGAGAC
AGGCTTT ATCCGAT AGGCTTT GAGACAG
AGTCGAG TTAGATT ATGAGGC AGAGACA
GGCTTTA TCCGATG TTTAGAG
CGAGGCT TAGATCC TGAGGCT GAGACAG
AGTCGAG TTTAGATC ATGAGGC TTAGAGA
GAGGCTT GATCCGA GAGGCTT GAGACAG
1. Sequencing
(e.g. Solexa, 454…))
2. Hashing
Linear stretches
GATT (1x)
TGAG (9x) ATGA (8x) GATG (5x) CGAT (6x) CCGA (7x) TCCG (7x) ATCC (7x) GATC (8x) AGAT (8x)
AGAA (1x)
GCTC (2x) CTCT (1x) TCTA (2x) CTAG (2x)
TAGT (3x) AGTC (7x) GTCG (9x) TCGA (10x) CGAG (8x) GAGG (16x) AGGC (16x) GGCT (11x)
GCTT (8x) CTTT (8x) TTTA (8x) TTAG (12x)
TAGA (16x) AGAG (9x) GAGA (12x) AGAC (9x) GACA (8x) ACAG (5x)
CGAC (1x) GACG (1x) ACGC (1x)
3. Simplification of linear stretches
GATT
AGAT
GATCCGATGAG
GCTCTAG
TAGTCGA CGAG
GAGG AGGCT
GCTTTAG
TAGA AGAGA AGACAG
AGAA
CGACGC
Tips
Bubble
4. Error removal
AGATCCGATGAG
TAGTCGAG GAGGCTTTAGA AGAGACAG

Image courtesy of Jim Noonan, Yale University.

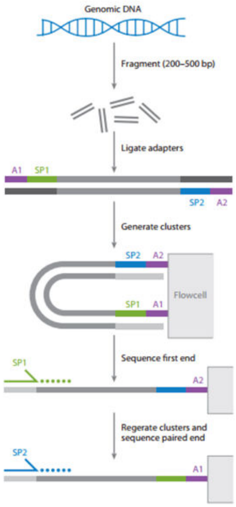
Assembly Quality Assessment Scores

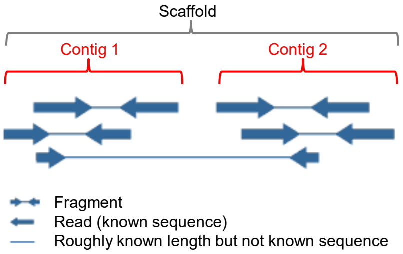
- Typically, several assemblers (such as ABySS, SOAPdenovo, Velvet etc.) are used and then a final assessment of quality is made using parameters such as:
- Number of Contigs…generally, the fewer the better
- N50: the maximum length *L* such that 50% of all bases lie in contigs at least *L* bases long…generally, the longer the better
- Coverage: the number of reads covering each base in a contig

Note that *de novo* genome (and transcriptome) assembly, especially from short reads, is a very active area of research. The first “assemblathon” was recently held to assess many different methods.

Assemblathon 1: Earl et al. (2011) Genome Research. DOI: 10.1101/gr.126599.111.

How Paired-End Reads Can Facilitate Assembly





From Illumina.com website and http://genome.jgi.doe.gov/help/scaffolds.html

RNA-Seq

RNA-seq is a powerful method for quantifying steady-state mRNA expression levels and detecting alternative splicing events in transcriptomes.

An RNA-seq pipeline involves creating cDNA, shearing, adding adapters, NGS, mapping reads, and summarizing the read counts (e.g. FPKM – fragments per kilobase per million fragments mapped). Evaluation of alternative splicing events may be visualized in a genome browser.

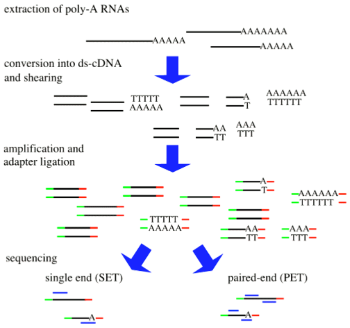
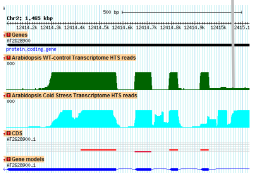
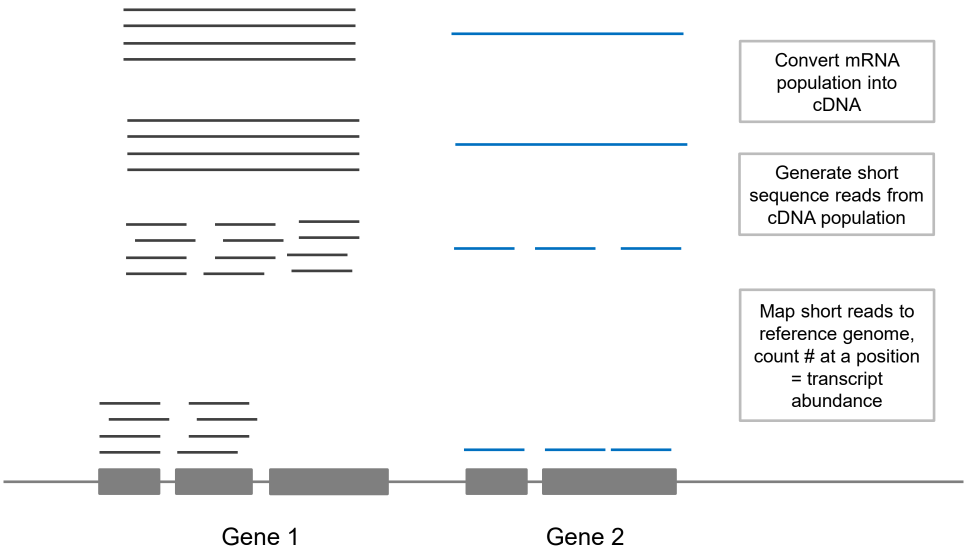


Image courtesy of http://cmb.molgen.mpg.de

Short reads mapped to reference genome in Gbrowse/JBrowse instance (height = # of reads at a given position)



RNA-Seq can be used to identify alternative splicing events



Convert mRNA population into cDNA

Generate short sequence reads from cDNA population

Map short reads to reference genome, count # at a position = transcript abundance

Alternative splicing is a common occurrence

“Of the 25,800 genes expressed along the [maize] leaf gradient we detected evidence of alternative splicing at 9,492 genes…Only 20,999 of [these] contain introns, so 56.4% of all possible targets showed evidence of alternative splicing.”

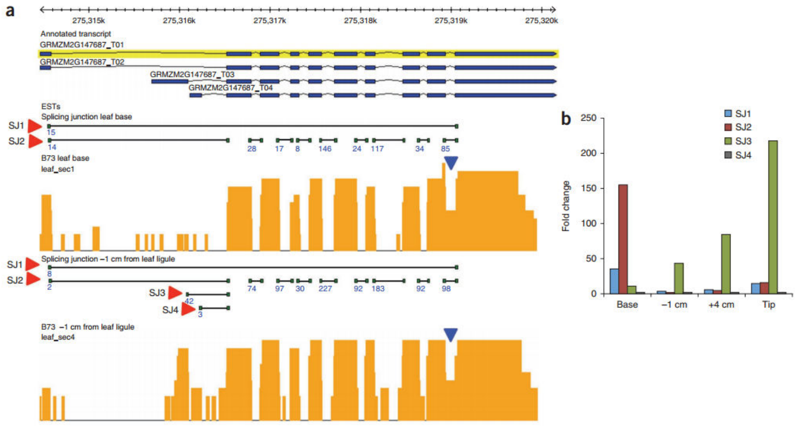
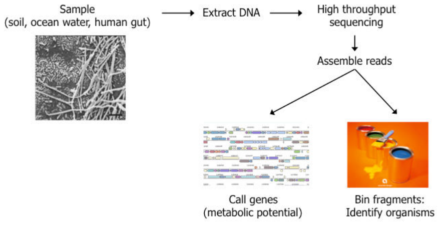


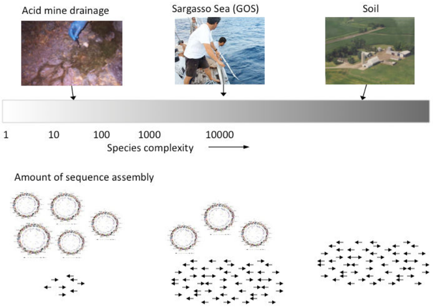
Figure 3 Alternative splicing of GRMZM2G147687. **(a)** Genome browser shows the alignment of reads to splice junctions (green) and exons (yellow) of GRMZM2G147687. Red arrows indicate the alternative splice junctions within the first exon, and blue arrows indicate putative intron retention events (see also **Supplementary Fig. 3**). **(b)** Results of qRT-PCR showing accumulation of four isoforms along the developmental gradient.

Li~Brutnell (2010). Nature Genetics 42: 1060-1067

Metagenomics

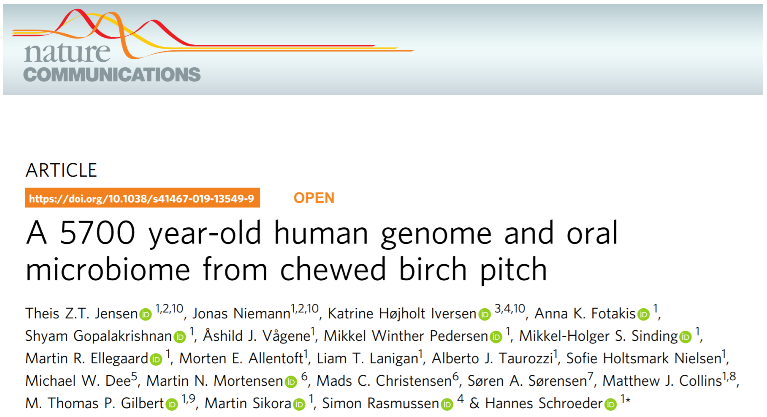
"The application of modern genomics techniques to the study of communities of microbial organisms directly in their natural environments, *bypassing the need for isolation and lab cultivation of individual species*"





Definition: Chen & Pacter (2005) PLoS CompBio 1:e24. Images courtesy of JGI.

Some amazing things are possible with NGS!


nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-13549-9 OPEN

A 5700 year-old human genome and oral microbiome from chewed birch pitch

Theis Z.T. Jensen[1,2,10], Jonas Niemann[1,2,10], Katrine Højholt Iversen[3,4,10], Anna K. Fotakis[1], Shyam Gopalakrishnan[1], Åshild J. Vågene[1], Mikkel Winther Pedersen[1], Mikkel-Holger S. Sinding[1], Martin R. Ellegaard[1], Morten E. Allentoft[1], Liam T. Lanigan[1], Alberto J. Taurozzi[1], Sofie Holtsmark Nielsen[1], Michael W. Dee[5], Martin N. Mortensen[6], Mads C. Christensen[6], Søren A. Sørensen[7], Matthew J. Collins[1,8], M. Thomas P. Gilbert[1,9], Martin Sikora[1], Simon Rasmussen[4] & Hannes Schroeder[1*]