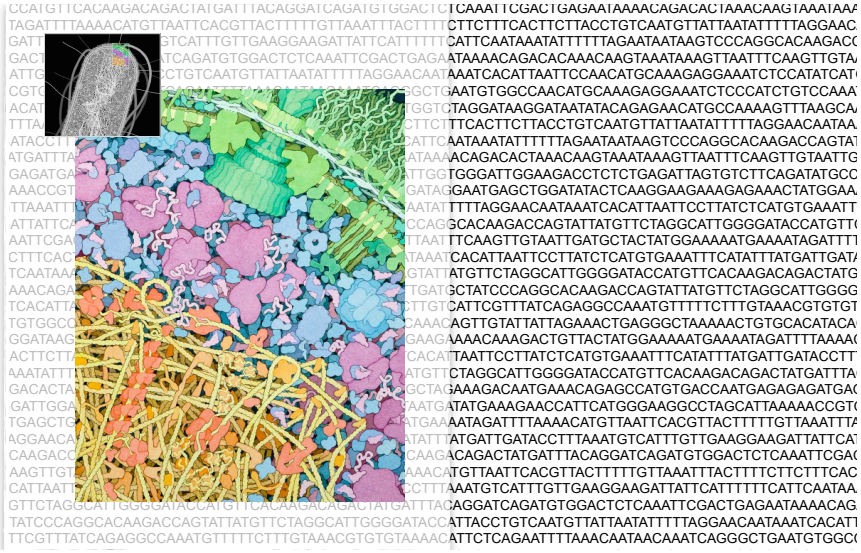
Genomics I

Biomedical Data Science: Mining and Modeling
CB&B 752 • MB&B 452
Matt Simon
Feb 2, 2021



What is genomics?

1. The **global** study of how biological **information** is encoded in genome sequence

   Genes
   Regulatory sequences
   Genetic variation

2. How this information is **read out** to produce distinct **biological outcomes**

   Gene expression and regulation
   Cellular identity, differentiation and development
   Phenotypic variation among individuals and species

In practice, many experiments that involve **deep sequencing** are considered genomics.

Overview

Genomics I: (today's lecture): Focus on sequencing technology and genomes.

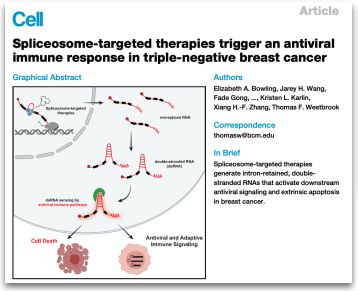
Genomics II: (Monday's lecture): Focus on applications of sequencing technology.

Overview

- Sequencing data: from wet lab to fastq.
- Applications to studying genomes and much much more.

✶ Sophisticated use of data from genomics requires an integrated understanding of the biological experiment, sample preparation and down stream computational analyses of the data.

Importance of genomics data: these data are central to most biomedical and biological


Cell

Article

Spliceosome-targeted therapies trigger an antiviral immune response in triple-negative breast cancer

Graphical Abstract

Authors

Elizabeth A. Bowling, Jarey H. Wang, Fade Gong, ..., Kristen L. Karlin, Xiang H.-F. Zhang, Thomas F. Westbrook

Correspondence

thomasw@bcm.edu

In Brief

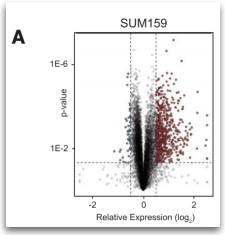
Spliceosome-targeted therapies generate intron-retained, double-stranded RNAs that activate downstream antiviral signaling and extrinsic apoptosis in breast cancer.

Highlights

- Spliceosome-targeted therapies (STTs) induce widespread mis-spliced mRNA in cancer
- Mis-spliced, intron-retained mRNAs are an unexplored source of endogenous dsRNA
- STTs trigger antiviral signaling and extrinsic apoptosis in TNBCs via dsRNA sensors
- RNA mis-splicing in human breast cancers correlates with immune signatures

Bowling et al., 2021, Cell *184*, 384–403
January 21, 2021 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.cell.2020.12.031

Importance of genomics data: these data are central to most biomedical and biological



Bowling et al., 2021, Cell *184*, 384–403
January 21, 2021 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.cell.2020.12.031

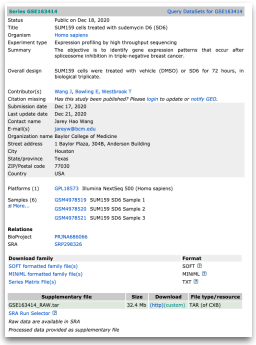
Figure 1. Spliceosome-targeted therapies stimulate antiviral signaling in MYC-driven triple-negative breast cancer

(A) Volcano plot of RNA-seq gene-expression changes due to spliceosome inhibition for two MYC-driven TNBC cell lines, SUM159 and LM2, treated with SD6 or DMSO (n = 3 biological replicates).

Data and code availability

The datasets generated during this study are available in GEO [GEO: GSE163411, GSE163414, GSE163181, GSE163188, GSE163232].

Raw data can be found in genomics databases


Series GSE163414 Query DataSets for GSE163414

Status	Public on Dec 18, 2020
Title	SUM159 cells treated with sudemycin D6 (SD6)
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	The objective is to identify gene expression patterns that occur after spliceosome inhibition in triple-negative breast cancer.
Overall design	SUM159 cells were treated with vehicle (DMSO) or SD6 for 72 hours, in biological triplicate.
Contributor(s)	Wang J, Bowling E, Westbrook T
Citation missing	Has this study been published? Please login to update or notify GEO.
Submission date	Dec 17, 2020
Last update date	Dec 21, 2020
Contact name	Jarey Hao Wang
E-mail(s)	jarey@bcm.edu
Organization name	Baylor College of Medicine
Street address	1 Baylor Plaza, 304b, Andersen Building
City	Houston
State/province	Texas
ZIP/Postal code	77030
Country	USA
Platforms (1)	GPL18573 Illumina NextSeq 500 (Homo sapiens)
Samples (6)	GSM4978519 SUM159 SD6 Sample 1 GSM4978520 SUM159 SD6 Sample 2 GSM4978521 SUM159 SD6 Sample 3
Relations	
BioProject	PRJNA686066
SRA	SRP298326

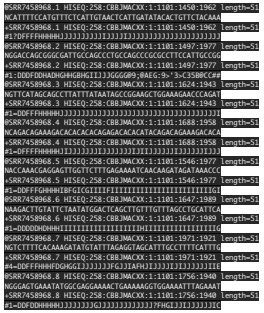
Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE163414_RAW.tar	32.4 Mb	(http)(custom)	TAR (of CXB)

SRA Run Selector

Raw data are available in SRA

Processed data provided as supplementary file



- Most journals require authors to submit their data to a database (e.g.,GEO) prior to publication.
- These databases entries contain raw data and processed data.
- These data can be use to examine the authors' claims, but also to test new hypotheses.

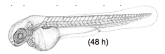
Central questions

Where do these data come from?

How does the way we collect it influence what we know?

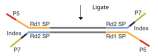
Workflow

1. Isolation of sample.



e.g., Isolate DNA and shear.

2. Library preparation



e.g., Add known sequences to the ends.

3. Sequencing

e.g., Illumina Novaseq

4. Analysis

e.g., Map to genome and interpret.

Metrics for evaluating sequencing technology

Throughput:
- Number of high quality bases per unit time
- Number of independent samples run in parallel
- Difficulty of sample preparation

Yield
- Number of useful reads per sample
- Read length

Cost
- Per run cost
- Per base cost
- Equipment
- Reagents
- Labor
- Analysis

Quality
- Accuracy per base

What is sequencing?

One-at-a-time methods

a. Maxam-Gilbert Sequencing
b. Sangar Sequencing

Short read deep sequencing

a. Illumina Sequencing
b. Ion Torrent

Long read deep sequencing

a. **Nanopore** based
b. **Pacific Bioscience** Sequencing

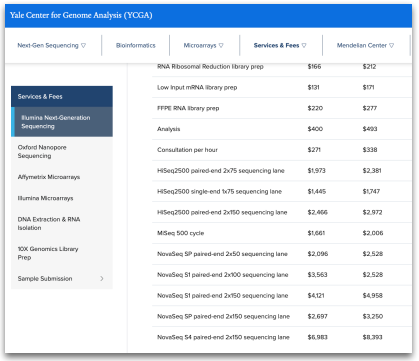
Sequencing technology	Platform	Data type	Read length (kb)		Read accuracy (%)	Throughput per flow cell (Gb)		Estimated cost per Gb (US$)	Maximum throughput per year (Gb)[a]
			N50	Maximum		Mean	Maximum		
Pacific Biosciences (PacBio)	RS II[b]	CLR	5–15	>60	87–92	0.75–1.5	2	333–933[c]	4,380
	Sequel	CLR	25–50	>100		5–10	20	98–195[d]	17,520
	Sequel II	CLR	30–60	>200		50–100	160	13–26[e]	93,440
		HiFi	10–20	>20	>99	15–30	35	43–86[e]	10,220
Oxford Nanopore Technologies (ONT)	MinION/GridION	Long	10–60	>1,000	87–98	2–20	30	50–500[f]	21,900 (MinION) 109,500 (GridION)
		Ultra-long	100–200	>1,500		0.5–2	2.5	500–2,000[f]	913 (MinION) 4,563 (GridION)
	PromethION	Long	10–60	>1,000		50–100	180	21–42[f]	3,153,600
Illumina	NextSeq 550	Single-end	0.075–0.15	0.15	>99.9	16–30	>30	50–63[g]	>47,782
		Paired-end	0.075–0.15 (×2)	0.15 (×2)		32–120	>120	40–60[g]	>70,080
	NovaSeq 6000	Single-end	0.05–0.25	0.25		65–3,000	>3,000	10–35[h]	>1,194,545
		Paired-end	0.05–0.25 (×2)	0.25 (×2)					

The technology will change, but your need to critically understand the input and output will not.

Logsdon (**2020**) *Nat Rev Genetics*

The steps of sequencing experiments

1. Sample preparation
    a. Isolation
    b. Library construction
2. Sequencing
    a. Flow cell loading
    b. Cluster generation
    c. Sequencing
    d. Processing image files
    e. De-multiplexing samples
3. Data analysis
    a. Read filtering
    b. Alignment to a genome
    c. Diverse analyses



Retrieved Feb 2, 2021:
https://medicine.yale.edu/keck/ycga/services/illuminaprices/

What is the output from an Illumina sequencing experiment?

Many reads…

```
@HWI-D00306:498:HBB89ADXX:1:1101:1180:1882 1:N:0:CGATGT
NCATCACTTTCTGCACCAGCCATGACGTCAATCTTCGTCCGAACCCCAAACTCGAGATCGGAAGAGCACACGTCTG
+
#11BBDDDFFDFBFFFIIIIIIIIIIIIIIIFEGIIIIIFIGAGIIFIII=FEEEEEFFFDDD=@9A@BBBBB=?BB<

@HWI-D00306:498:HBB89ADXX:1:1101:1167:1902 1:N:0:CGATGT
TATTGCAATATGTTAACAATCTAACAAGGAAAAAATACCCCACACAAAACAAAACACAACCCTTAGAACTGTGCTG
+
B@@FFDFFHFHHHJJIJIGIIJJJJJIJJHFIJJJJJIJJJEHHJJIJJJJIIJJJJJGHHHHFBDFFFE>CEEC
@HWI-D00306:498:HBB89ADXX:1:1101:1190:1928 1:N:0:CGATGT
ACCAAGCCACAATAAGTTAGTGTTTCCATAGTACATGCTGAGTTATTTGATCCCGTATCTATACACTGCTACTGTC
+
@<@DDDDD8CDDDGE?2<AFFBCCEEHEIEGHIIEGEIDD@CDGFFFEFIDGCFCDABFG>FBFGFGIEIFFFDDD
@HWI-D00306:498:HBB89ADXX:1:1101:1157:1931 1:N:0:CGATGT
CTGAGATTCTTTGCCATAGTCCTTAACCACTACGCAACTGCAACCAACCACCTTCCGTGGTTTGCCCTCTCGATCG
+
CCCFFFFFHHHHHHIJJIIJJJIIGHHIJGGJIGIJJJJJJJIJIIIJJJIIJJJJIIJGJJHCHFBDFFFDDECB
```

Generally ~ 2,000,000,000 reads/sequencing lane

Note: This is for an Illumina NovaSeq with current chemistry, but this number changes

What is the output from an Illumina sequencing experiment?

One read (fastq format)

```
@HWI-ST1239:178:H0KPNADXX:2:1101:3120:1979 1:N:0:TGACCA
NCTGTAGGCTGCGTAGCCTCCCTGCAGGGTAAGTGGGAGGAGAGAGAGCAGAGGGACTTAGTGGGGCTCCCCAGGG
+
#1=DDFFFHHHHHIJIJJJIJJJJJIJJJ?FHIDGIJ=GIHGIIIHGIJIHEHIHHGFFFFEEEDDDDDDDDDDD
```

1. Read identifier
2. **Sequence**
3. Quality score identifier "+"
4. Quality score

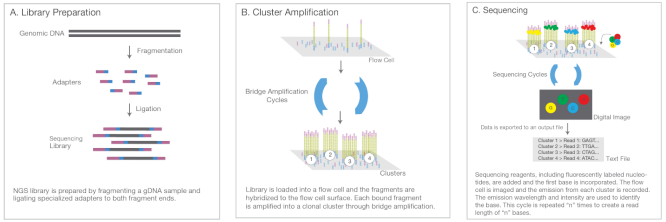
How long are the reads?

```
TATTGCAATATGTTAACAATCTAACAAGGAAAAAATACCCCACACAAAACAAAACACAACCCTTAGAACTGTGCTG
```

75 nt

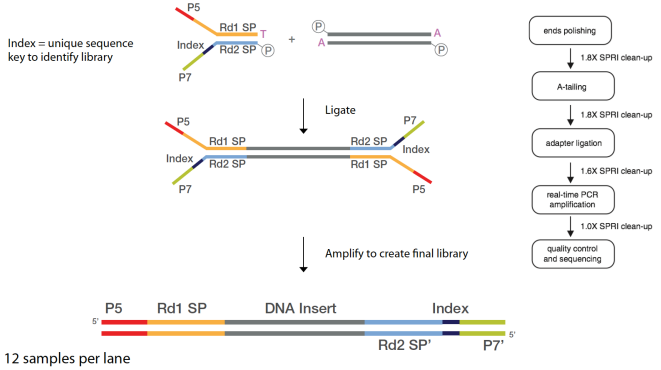
While there are other technologies that can give longer read lengths, Illumina reads are generally 50 nt - 250 nt

Where do these reads come from?



https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

Optional: How do you make a sequencing library?



12 samples per lane

Potential sources of bias:

1. Selective PCR amplification (issue of duplicates).
2. Size selection.
3. Enzyme specificities.

Challenging but possible to analyze pg quantities of DNA. (In humans, ~6 pg DNA/cell).

What is the output from an Illumina sequencing experiment?

Paired read (fastq format)

```
@HWI-ST1239:178:H0KPNADXX:2:1101:3120:1979 1:N:0:TGACCA
NCTGTAGGCTGCGTAGCCTCCCTGCAGGGTAAGTGGGAGGAGAGAGAGCAGAGGGACTTAGTGGGGCTCCCCAGGG
+
#1=DDFFFHHHHHHIJIJJJIJJJJJIJJJ?FHIDGIJ=GIHGIIIHGIJIHEHIHHGFFFFEEEDDDDDDDDDDD

@HWI-ST1239:178:H0KPNADXX:2:1101:3120:1979 2:N:0:TGACCA
NNACCTAGCCATCTGCAGTCCTCGGTCCTGTGTTAGACCAGAACTAGGTGCCCAGGCCAGGTACCACCTAATCCTT
+
##4<@@@@@@@@@?@@@?@@??????@@??@?????????????????????>???????????@>???@@@?@@??????
```

1. Read identifier
    a. **Instrument**
    b. **Flow cell**
    c. **Read ID**
    d. **Coordinates**
    e. **Which read from a paired end sample**
    f. **Which index for multiplexed read**
2. Sequence
3. Quality score identifier "+"
4. Quality score

What limits the insert size and read length?

One read (fastq format)

```
@HWI-D00306:498:HBB89ADXX:1:1101:1180:1882 1:N:0:CGATGT
NCATCACTTTCTGCACCAGCCATGACGTCAATCTTCGTCCGAACCCCAAACTCGAGATCGGAAGAGCACACGTCTG
+
#11BBDDDFFDFBFFFIIIIIIIIIIIIIIIFEGIIIIIFIGAGIIFIII=FEEEEEFFFDDD=@9A@BBBBB=?BB<
```

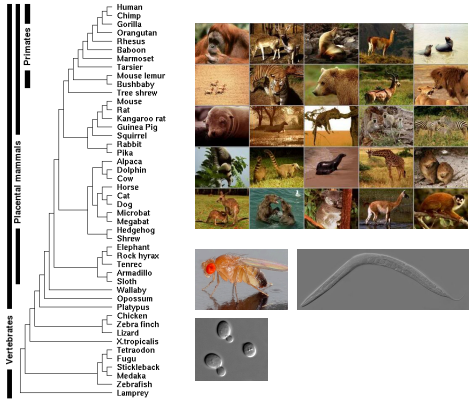
- For each single end read: Incomplete incorporation of bases.

- For the size of the insert (especially for paired end analysis): Ability to get consistent clusters.

What do I do with my sequencing reads?

Source: Slate via Noonan

Many reference genomes are available

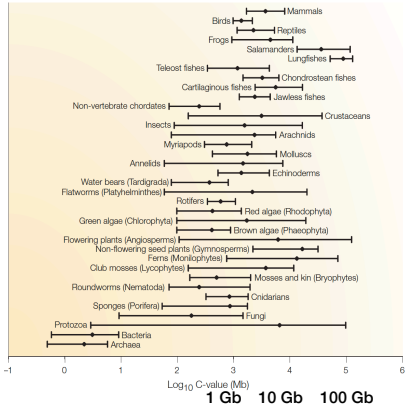


There is a wide range of genome sizes.

kb = 1000 bp
Mb = 1x10^6 bp
Gb = 1x10^9 bp
Tb = 1x10^{12} bp

Human haploid genome ~ 3 Gb

75 nt x 3x10^8 reads/lane is about the right scale, but the amount of **coverage** necessary depends on application.



Sequencing of the human genome

Victory declared **2003**

•Industrialization of Sanger sequencing, library construction, sample preparation, analysis, etc.

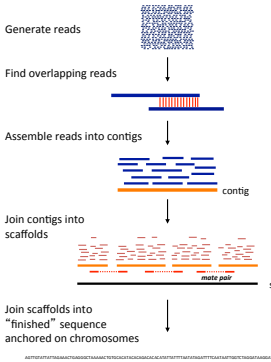
•$3 billion total cost

•1 Gb/month at largest centers (2005)

HGSC
National Human Genome Research Institute
JGI DOE JOINT GENOME INSTITUTE
wellcome trust sanger institute

Novaseq 20 billion reads 2x150 bp. $1000 -> $100/genome.

How to assemble a genome



Assembly quality criteria:

Accuracy: number of errors (Human << 1/100,000 bp)

Contiguity: number of gaps (Human: est. 357)

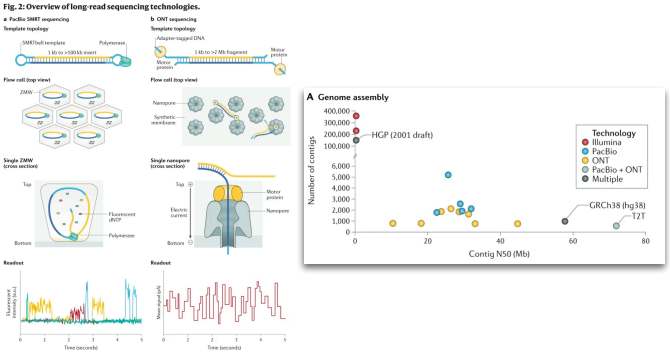
Coverage:
Average number of reads representing a particular position in the assembly

Human, Mouse, Rat: > 20x
Chimpanzee: ~6x
Squirrel: ~2x

Scaffold_0: 12,865,123 – 12,965-110

Chr5: 133,876,119 – 134,876,119

The importance of long read sequencing



Logsdon (**2020**) *Nat Rev Genetics*

What types of annotation do we have/want?

~3 billion bp

Genes:
- Coding, noncoding, miRNA, etc.
- Isoforms
- Expression

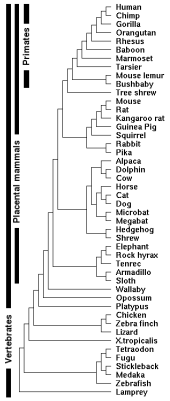
Genetic variation:
- SNPs and CNVs

Sequence conservation

Regulatory sequences:
- Promoters
- Enhancers
- Insulators

Epigenetics:
- DNA methylation
- Chromatin

Degrees of genomic annotation vary widely



ENCODE and modENCODE

Human, Mouse (Fly, Worm, Yeast):
- Chromosome assemblies
- Dense gene and regulatory maps, variation, etc.

Other models (Dog, Chicken, Zebrafish):
- Chromosome assemblies
- Partial gene maps; variation; little regulatory data

Low coverage vertebrate genomes:
- Scaffold assemblies
- Few annotated genes
- Used for comparative purposes

Where do you look for existing annotations?

UCSC Genome Browser (genome.ucsc.edu):
Visualization, data recovery, simple analysis
(also http://genome-preview.ucsc.edu/)

ENSEMBL (ensembl.org):
Visualization, data recovery, simple analysis

Integrative Genomics Viewer
(broadinstitute.orgsoftware/igv/):
Local genome viewer (visualize local and remote data)

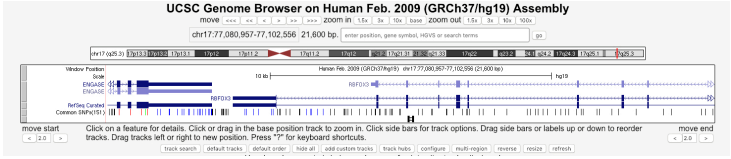
Galaxy (main.g2.bx.psu.edu):
Complex data analysis and workflows

Example of a genome browser track (UCSC)

Chr5: 133,876,119 – 134,876,119

Our specific example:

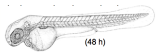
```
@HWI-ST1239:178:H0KPNADXX:2:1101:3120:1979 1:N:0:TGACCA
NCTGTAGGCTGCGTAGCCTCCCTGCAGGGTAAGTGGGAGGAGAGAGAGCAGAGGGACTTAGTGGGGCTCCCCAGGG
+
#1=DDFFFHHHHHHIJIJJJIJJJJJIJJJ?FHIDGIJ=GIHGIIIHGIJIHEHIHHGFFFFEEEDDDDDDDDDDD

@HWI-ST1239:178:H0KPNADXX:2:1101:3120:1979 2:N:0:TGACCA
NNACCTAGCCATCTGCAGTCCTCGGTCCTGTGTTAGACCAGAACTAGGTGCCCAGGCCAGGTACCACCTAATCCTT
+
##4<@@@@@@@@@?@@@?@@??????@@??@?????????????????>??????????@>???@@@?@@??????
```



Workflow

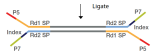
1. Isolation of sample.

e.g., Isolate DNA and shear.



2. Library preparation

e.g., Add known sequences to the ends.



3. Sequencing

e.g., Illumina Novaseq

4. Analysis

e.g., Map to genome and interpret.

Using sequencing to annotate the genome

1. Where are the cis-acting regulatory elements in DNA?
   - A. DNase I hyper-sensitivity mapping (DNase-Seq).
   - B. FAIRE to map regulatory elements.
2. Where do transcription factors bind?
   - C. ChIP-seq of transcription factors (or in high res, ChIP-exo)
   - D. Nucleosome mapping (MNase-Seq).
3. Where are different histone modifications found?
   - E. ChIP-Seq of histone modifications.
   - F. ChIP-Seq of chromatin writers, readers and erasers.
4. Where is RNA polymerase transcribing?
   - G. ChIP-Seq of polymerase.
   - H. GRO-Seq, NET-Seq and TT-Seq to measure RNA in the polymerase active site..
5. How is the genome organized in 3D?
   - I. 4C/5C/Hi-C to measure chromatin conformation.

Applications of sequencing technology next week.

Conclusions

- Sequencing technology is central to our understanding of biology.
- The decrease in cost and increase in throughput make sequencing data increasingly ubiquitous.

✳ Sophisticated use of data from genomics requires an integrated understanding of the biological experiment, sample preparation and down stream computational analyses of the data.