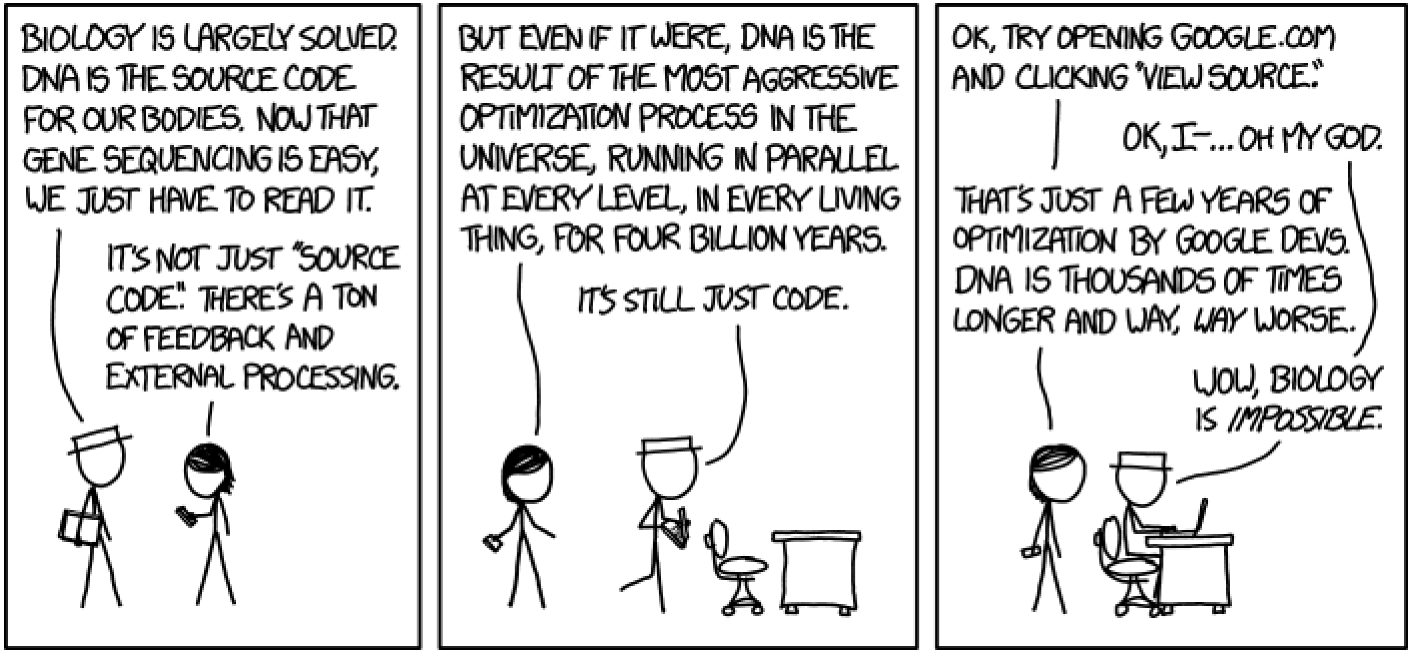
MIMM4750G
Next-generation sequence analysis


BIOLOGY IS LARGELY SOLVED. DNA IS THE SOURCE CODE FOR OUR BODIES. NOW THAT GENE SEQUENCING IS EASY, WE JUST HAVE TO READ IT.
IT'S NOT JUST "SOURCE CODE." THERE'S A TON OF FEEDBACK AND EXTERNAL PROCESSING.
BUT EVEN IF IT WERE, DNA IS THE RESULT OF THE MOST AGGRESSIVE OPTIMIZATION PROCESS IN THE UNIVERSE, RUNNING IN PARALLEL AT EVERY LEVEL, IN EVERY LIVING THING, FOR FOUR BILLION YEARS.
IT'S STILL JUST CODE.
OK, TRY OPENING GOOGLE.COM AND CLICKING "VIEW SOURCE."
OK, I—... OH MY GOD.
THAT'S JUST A FEW YEARS OF OPTIMIZATION BY GOOGLE DEVS. DNA IS THOUSANDS OF TIMES LONGER AND WAY, WAY WORSE.
WOW, BIOLOGY IS *IMPOSSIBLE.*

What is NGS?

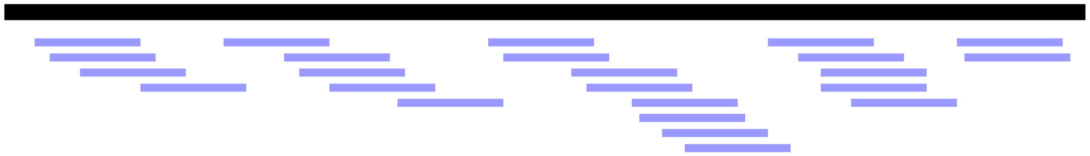
- A catch-all term for specialized technology for performing genetic sequencing reactions on a very large scale.
- NGS platforms generate gigabytes or *terabytes* of sequence data in a day.
- The field of bioinformatics grew largely from the need to make sense of these data.

NGS applications

- Whole-genome sequencing
- Exome sequencing
- CHiP-seq
- **Deep sequencing**
- RNA sequencing (RNA-seq)
- **Metagenomics**

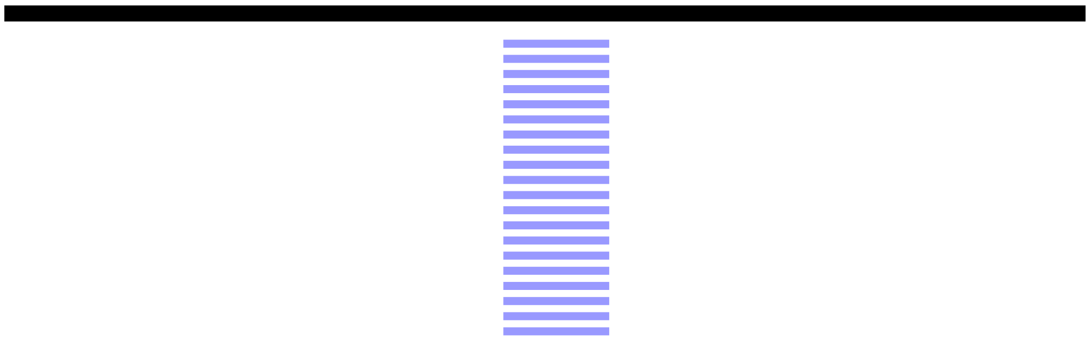
Whole-genome sequencing

- NGS platforms are converging towards \$10 per Gbp.
- Bacterial genomes range from about 100 kbp to nearly 20 Mbp
- May be cheaper to randomly shear template and use NGS for "shotgun" sequencing.
- More information than targeted gene marker sequencing (*e.g.*, variable number tandem repeat sequencing in *M. tuberculosis*).



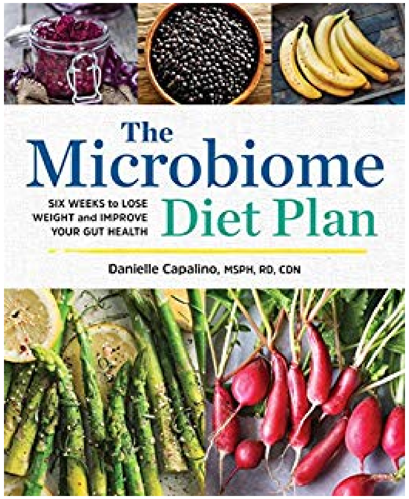
Deep sequencing

- Sequence a specific region of the genome from thousands of copies in the pathogen population.
- Useful to measure the frequency of some variant in the infection.
- Can yield a sequence alignment for reconstructing within-host evolution.



Metagenomics

- Reads cover many genomes from different organisms in the same sample.
- Often used to characterize the microbial composition of a sample.
- Useful to discover novel pathogens or to detect pathogens that cannot be cultivated.


The Microbiome Diet Plan
SIX WEEKS to LOSE WEIGHT and IMPROVE YOUR GUT HEALTH
Danielle Capalino, MSPH, RD, CDN

NGS databases

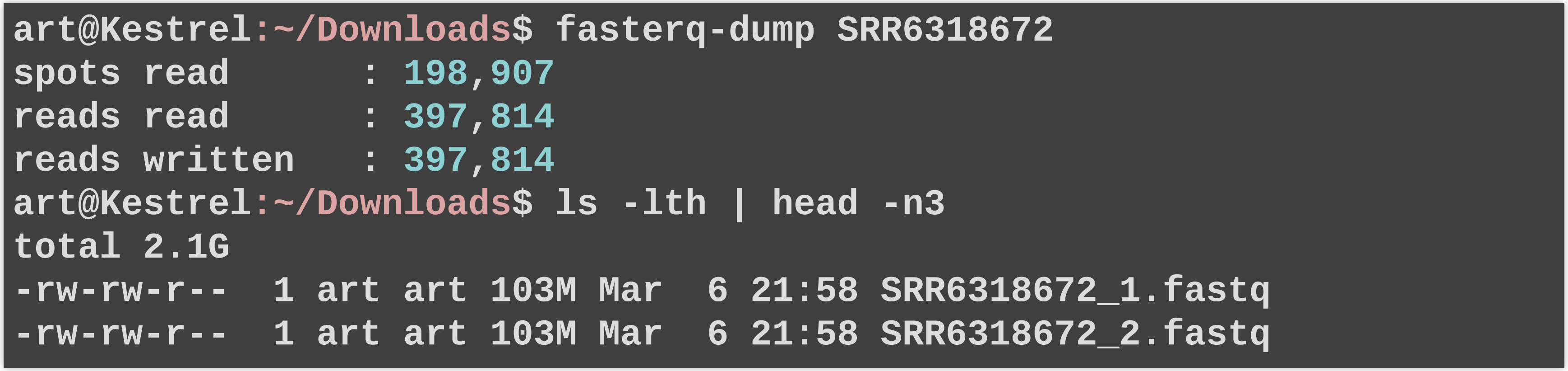
- Storing and distributing NGS data created a unique problem for those maintaining pubilc databases of conventional sequences.
- NCBI created the Short Read Archive (now the *Sequence* Read Archive).
- Partnership with EMBL-EBI European Nucleotide Archive and the National Institute of Genetics DNA Data Bank of Japan.

SRA Toolkit

- NCBI requires users to use its own open-source software to download data
- https://github.com/ncbi/sra-tools
- `fasterq-dump` uses multi-threading and file caching to make downloads faster
- current release for Linux only!

fasterq-dump

Using `fasterq-dump` to retrieve a WGS data set of *Helicobacter pylori*.



```
art@Kestrel:~/Downloads$ fasterq-dump SRR6318672
spots read      : 198,907
reads read      : 397,814
reads written   : 397,814
art@Kestrel:~/Downloads$ ls -lth | head -n3
total 2.1G
-rw-rw-r--  1 art art 103M Mar  6 21:58 SRR6318672_1.fastq
-rw-rw-r--  1 art art 103M Mar  6 21:58 SRR6318672_2.fastq
```

NGS data formats

- Recall that FASTQ is like an expanded version of FASTA

```
@SRR6318672.2 2 length=251
GGATAAAATGATACCCGCTTTTTTGATCACGCCCATTTCTAGCCAGATCGCTGGTAAAGTCATCGCGCAAGT
+SRR6318672.2 2 length=251
BCCCCFFFFFFFGGGGGGGGGGHGGGHHHHGHGGGHHHHHHHGHHHHHHHGGGGHHHHGHHHHHHGGGGGGGH
```

- Row 1 has @ prefix contains sequence label.
- Row 2 contains nucleotide sequence.
- Row 3 has + prefix and sometimes repeats label.
- Row 4 contains quality scores.

Quality scores

- A quality score is a log-transform of the estimated probability (P) of an incorrect base call.

$$Q = -10 \times \log_{10} P$$

- So if the error probability is 0.01 (1%), then $\log_{10}(0.01) = -2$, and $Q = 20$.

INCA7, Q1

Encoding quality scores

- FASTQ uses ASCII encoding to convert quality scores from numbers to single characters.

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
|                         |    |         |
0........................26...31........41
```

- This saves a lot of space and makes it easier to see how scores relate to different bases:

```
GGATAAA
+
BCCCCFF
34 35 35 35 35 37 37
```

What is ASCII?

- American Standard Code for Information Interchange
- Data are stored and processed as binary `0`s and `1`s.
- ASCII is a map of binary numbers to human-readable characters.

Binary	Decimal	ASCII
010 0110	38	&
011 0101	53	5
100 1001	73	I
100 1010	74	J

ASCII for FASTQ

- The current version of Illumina systems subtracts 33 from the decimal value of each ASCII character.
- $\mathtt{I}$ becomes $73 - 33 = 40$, which gives us the Q score.

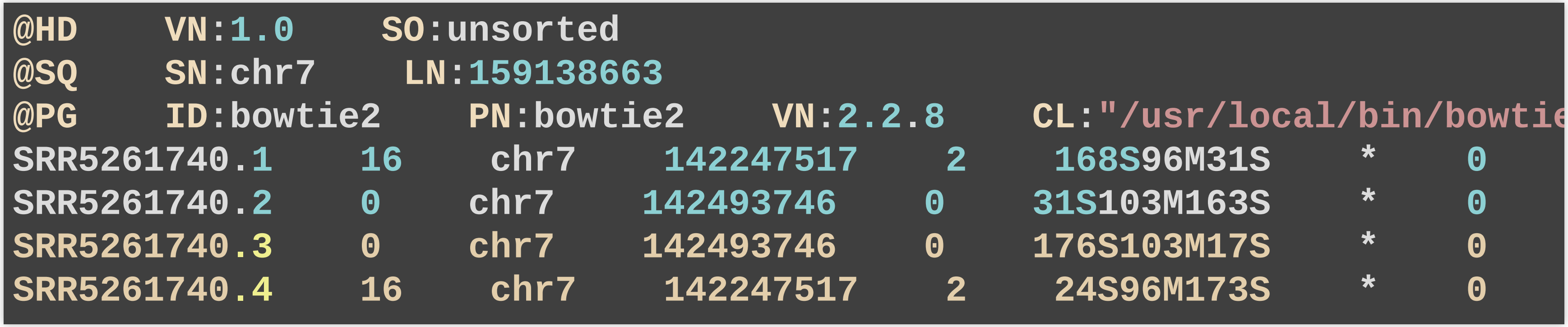
INCA7, Q2

Data compression

- FASTQ files are often stored in a `gzip` format.
- `gzip` is a UNIX (GNU) compression/decompression program.
- This program essentially replaces repeating sequences in the data with an instruction to copy forward the first instance.
- Generally reduces a FASTQ file down about 3-fold.
- Some programs can process the gzipped FASTQs!

SAM format

- The SAM (Sequence Alignment/Map) format has become a standard output format for programs that align NGS reads to reference genomes.
- It is a tabular, tab-separated data format.
- Comments at top of file prefixed with `@`



```
@HD     VN:1.0     SO:unsorted
@SQ     SN:chr7     LN:159138663
@PG     ID:bowtie2     PN:bowtie2     VN:2.2.8     CL:"/usr/local/bin/bowtie
SRR5261740.1     16     chr7     142247517     2     168S96M31S     *     0
SRR5261740.2     0     chr7     142493746     0     31S103M163S     *     0
SRR5261740.3     0     chr7     142493746     0     176S103M17S     *     0
SRR5261740.4     16     chr7     142247517     2     24S96M173S     *     0
```

SAM format

- Each line in a SAM corresponds to a read and contains the following information:

#	Name	Description
1	QNAME	Read label
2	FLAG	Bitwise flags
3	RNAME	Reference seq.
4	POS	Map location of 1st base in read
5	MAPQ	Mapping quality
6	CIGAR	Compact idiosyncratic gapped alignment report
7	RNEXT	Ref. seq. of mate
8	PNEXT	Map location of 1st base in mate
9	TLEN	Insertion length
10	SEQ	Read sequence
11	QUAL	Read quality string

Bitwise flags

- A decimal number is a compact way to store a series of bits.
- The decimal number 99 maps to the binary number `00001100011`.

Bit	Description	Bit	Description
1	read is paired	64	first in pair
2	read is mapped in a proper pair	128	second in pair
4	read is not mapped	256	not primary alignment
8	mate is not mapped	512	read fails platform quality checks
16	read is reverse strand	1024	read is PCR/optical duplicate
32	mate is reverse strand	2048	supplementary alignment

INCA7, Q3

CIGAR

- Compact Idiosyncratic Gapped Alignment Report
- A string representation of how the read aligns to the reference

Token	Description
M	Matched
I	Insertion
D	Deletion
S	Soft clip

- For example, `5S45M3I89M1S` means a 5nt soft clip, 45nt match, 3nt insertion, 89nt match, and 1nt soft clip.