# Next-Generation Sequencing (NGS)

Weigang Qiu, Ph.D.

Email: wqiu@hunter.cuny.edu

Department of Biological Sciences

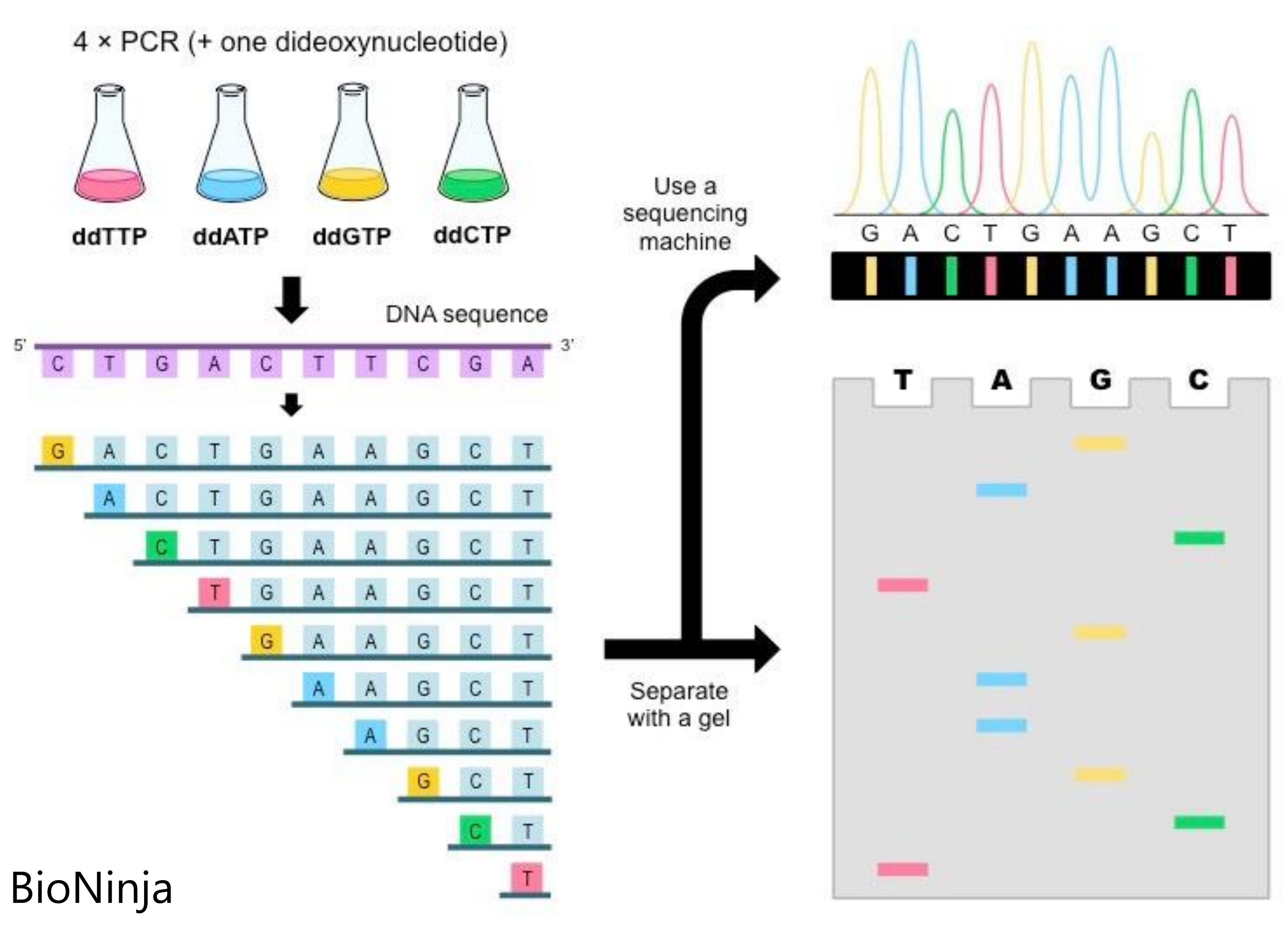
Hunter College, City University of New York (CUNY)

Spring 2026

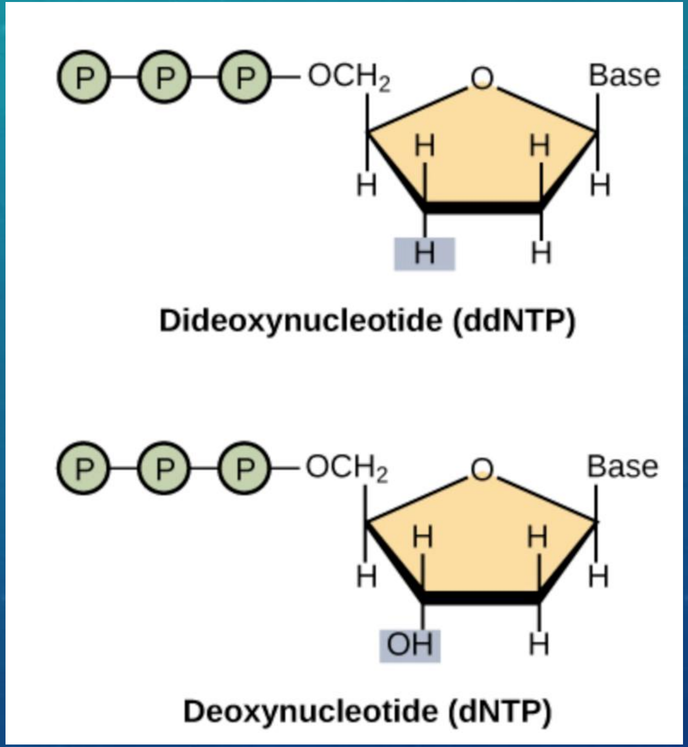
Comp Mol Bio

Research wiki: [https://wiki.genometracker.org](https://wiki.genometracker.org)

# SANGER SEQUENCING (1977)



BioNinja



Random chain termination by di-deoxynucleotides (ddNTPs)

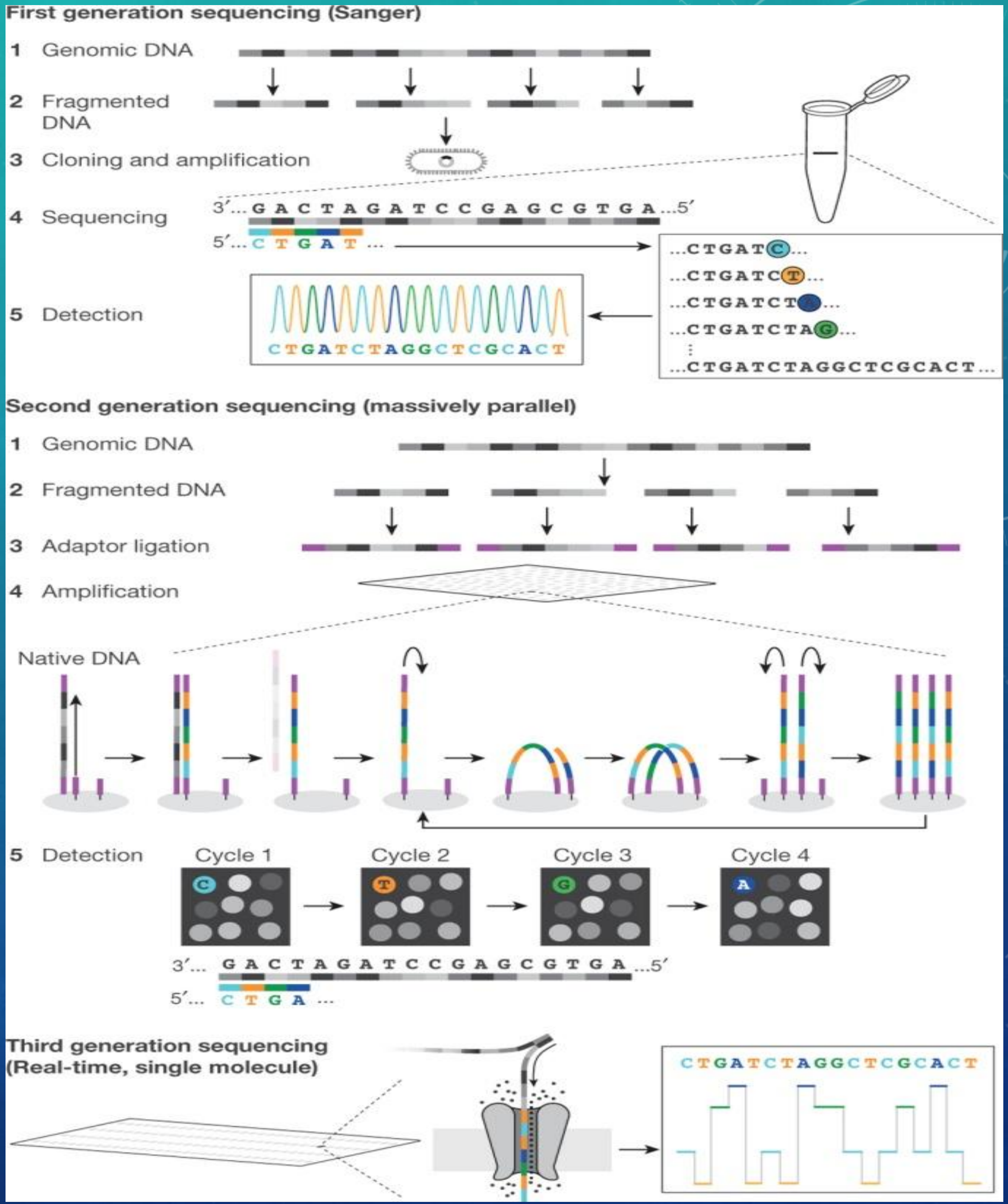
# $2^{ND}$ & $3^{RD}$ GENERATION SEQUENCING TECHNOLOGIES

Illumina (MiniSeq, MiSeq, HiSeq)

1. Library prep: fragement DNA, add adaptor (with PCR), add index (one per sample)
2. Cluster amplification: with immobilized oligos complementary to adaptor
3. Sequencing by synthesis: with florescence-labeled NTPs, imaged base-by-base

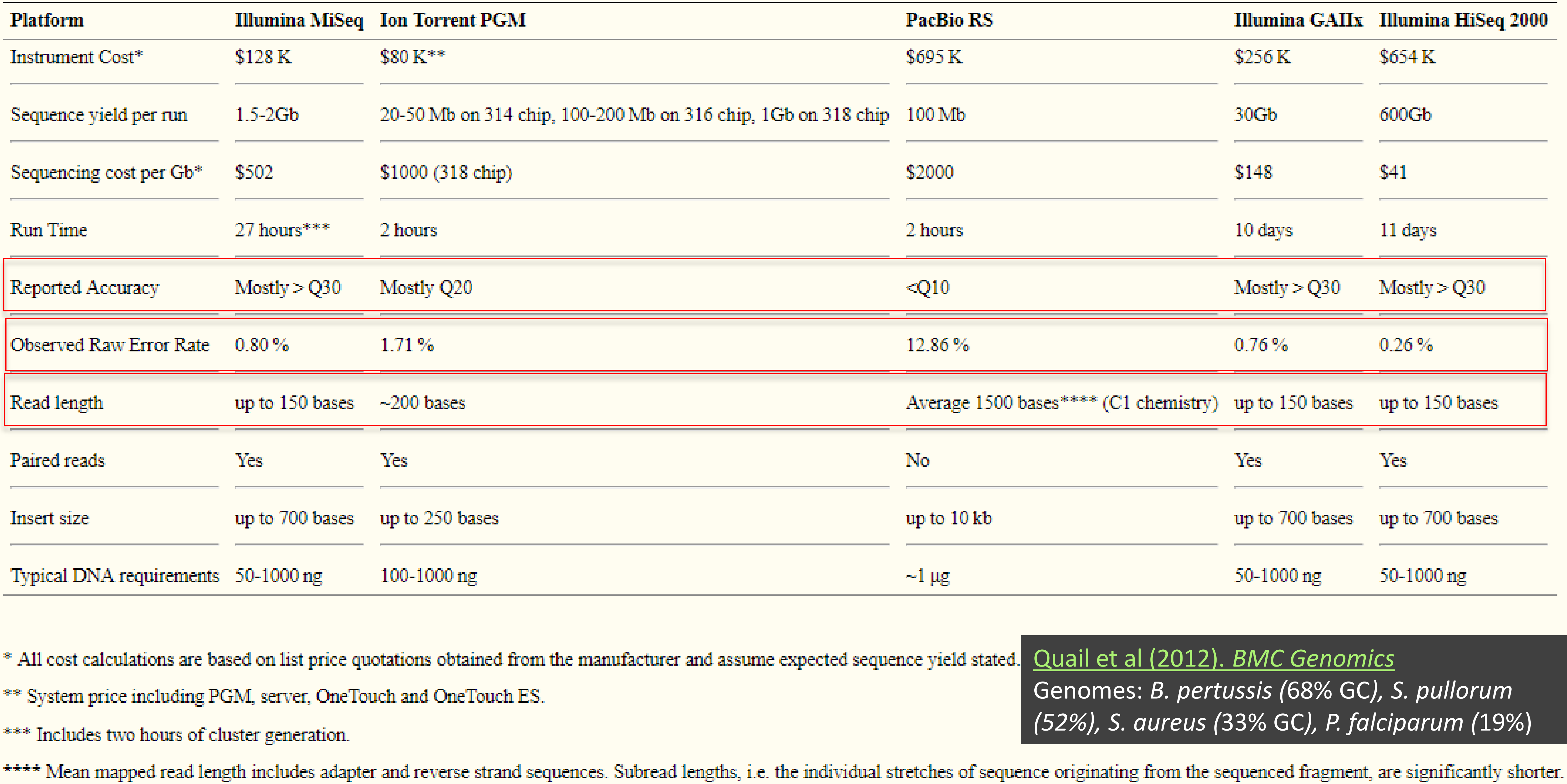
PacBio Single-Molecule Real Time (SMRT)

1. Library prep: no sample PCR
2. Zero Mode Waveguide (ZMW) chips: DNA-Polymerase recruites tagged NTPs, producing light pulses, one base at a time



Shendure *et al. Nature* 1–9 (2017)

# A COMPARATIVE STUDY OF 3 NGS TECHNOLOGIES



| Platform | Illumina MiSeq | Ion Torrent PGM | PacBio RS | Illumina GAIIx | Illumina HiSeq 2000 |
|---|---|---|---|---|---|
| Instrument Cost* | $128 K | $80 K** | $695 K | $256 K | $654 K |
| Sequence yield per run | 1.5-2Gb | 20-50 Mb on 314 chip, 100-200 Mb on 316 chip, 1Gb on 318 chip | 100 Mb | 30Gb | 600Gb |
| Sequencing cost per Gb* | $502 | $1000 (318 chip) | $2000 | $148 | $41 |
| Run Time | 27 hours*** | 2 hours | 2 hours | 10 days | 11 days |
| Reported Accuracy | Mostly > Q30 | Mostly Q20 | <Q10 | Mostly > Q30 | Mostly > Q30 |
| Observed Raw Error Rate | 0.80 % | 1.71 % | 12.86 % | 0.76 % | 0.26 % |
| Read length | up to 150 bases | ~200 bases | Average 1500 bases**** (C1 chemistry) | up to 150 bases | up to 150 bases |
| Paired reads | Yes | Yes | No | Yes | Yes |
| Insert size | up to 700 bases | up to 250 bases | up to 10 kb | up to 700 bases | up to 700 bases |
| Typical DNA requirements | 50-1000 ng | 100-1000 ng | ~1 μg | 50-1000 ng | 50-1000 ng |

* All cost calculations are based on list price quotations obtained from the manufacturer and assume expected sequence yield stated.

** System price including PGM, server, OneTouch and OneTouch ES.

*** Includes two hours of cluster generation.

**** Mean mapped read length includes adapter and reverse strand sequences. Subread lengths, i.e. the individual stretches of sequence originating from the sequenced fragment, are significantly shorter.

Quail et al (2012). *BMC Genomics*

Genomes: *B. pertussis* (68% GC), *S. pullorum* (52%), *S. aureus* (33% GC), *P. falciparum* (19%)

# WHICH METHOD SHOULD I USE FOR MY STUDY?

## Use **Sanger Sequencing when:**

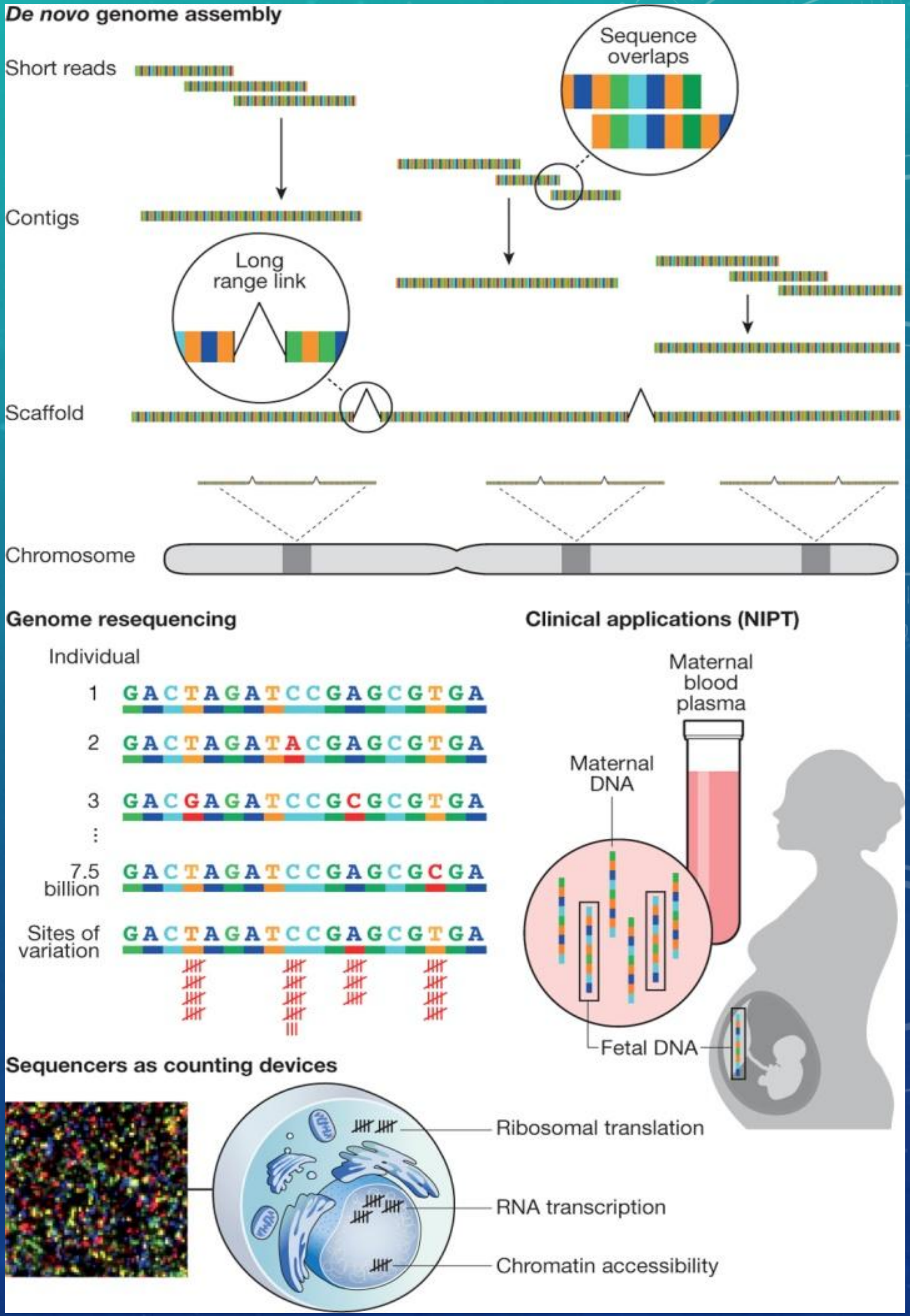
- Sequencing single genes (no mixed DNAs)
- Plenty of sample is available

## Use NGS when:

- Massively parallel & High-throughput: efficient & cost-effective
- Complex genomes: e.g., multiple plasmids
- Mixed DNA samples: e.g., cancer cells, microbiome, metagenomes
- Low input DNA: e.g., environmental DNA

# NGS APPLICATIONS ("-OMICS")

- *De novo* whole-genome sequencing (WGS)
- Exome sequencing (exons only); variant discovery (genetic polymorphisms; mutations)
- Sequencers as counting devices
  - Microbial ecology & microbiome (16S sequencing; meta-genomics)
  - Microbial genetics: sequencing of transposon-mutation library (Tn-seq)
  - Transcriptome sequencing: RNA-seq ("bulk" transcriptome)
  - Single-cell transcriptome/genome/methylation
  - Gene regulome: CHIP-seq
- Proteomics: MASS Spec; SILAC



Shendure *et al. Nature* 1–9 (2017)

# NGS FILE FORMATS & SOFTWARE TOOLS

## Sequences & Sequence Reads

### FASTA

```
>F63912
GTGGCTTACTACACACATCGTCATCTGACCCGCGAAGAGGAGCCGCAGACGCCTGACGAG
GCTTCGCTCGACCTGGCGGCTACCGATGGCATACGACTGGGCGACCGC
>X9820
GTGGCTTACTACACACATCGTCATCTGGCCCGCCTGCGCGAAGACGAGGAGCACCCGGCC
ACGCCCGGCGAAGCGACGCTGGACCTGGCCGCCACCGAGGCCATGCGCCTGGGCGACCGC
>M55212
GTGGCTTACTACACCCGCCGTCACTTGGCCCGCGAAGAAGAGGAACCGCCCACGGCCGAC
GAGGCCGTGCTCGATCTGGCCGATACCGCGGGTATGCGCCTGGGTGGTCGC
>T63266
GTGGCTTACTACACCCGCCGTCACTTGGCCCGCGAAGAAGAGGAACCGCCCACGGCCGAC
GAGGCCGTGCTCGATCTGGCCGATACCGCGGGTATGCGCCTGGGTGGTCGC
>H5708
GTGGCTTACTACACACATCGTCATCTGACCCGCGAAGAGGAGCCGCAGACGCCTGACGAG
GCTTCGCTCGACCTGGCGGCTACCGATGGCATACGACTGGGCGACCGC
>F34365
GTGGCTTACTACACACATCGTCATCTGGTCCGCCTGCGCGAAGACGAGGAGCACCCGGCC
ACGCCCGGCGAAGCGACGCTGGACCTGGCCGCCACCGAGGCCATGCGCCTGGGCGACCGC
```

bioseq:  a sequence utility developed in Qiu Lab (Hernandez et al, 2018. *BMC Bioinformatics*)

### FASTQ

```
@cluster_2:UMI_ATTCCG
TTTCCGGGGCACATAATCTTCAGCCGGGCGC
+
9C;=;=<9@4868>9:67AA<9>65<=>591
@cluster_8:UMI_CTTTGA
TATCCTTGCAATACTCTCCGAACGGGAGAGC
+
1/04.72,(003,-2-22+00-12./.-.4-
@cluster_12:UMI_GGTCAA
GCAGTTTAAGATCATTTTATTGAAGAGCAAG
+
?7?AEEC@>=1?A?EEEB9ECB?==:B.A?A
@cluster_21:UMI_AGAACA
GGCATTGCAAAATTTATTACACCCCCAGATC
+
>=2.660/?:36AD;0<14703640334-//
@cluster_29:UMI_GCAGGA
CCCCCTTAAATAGCTGTTTATTTGGCCCCAG
+
```

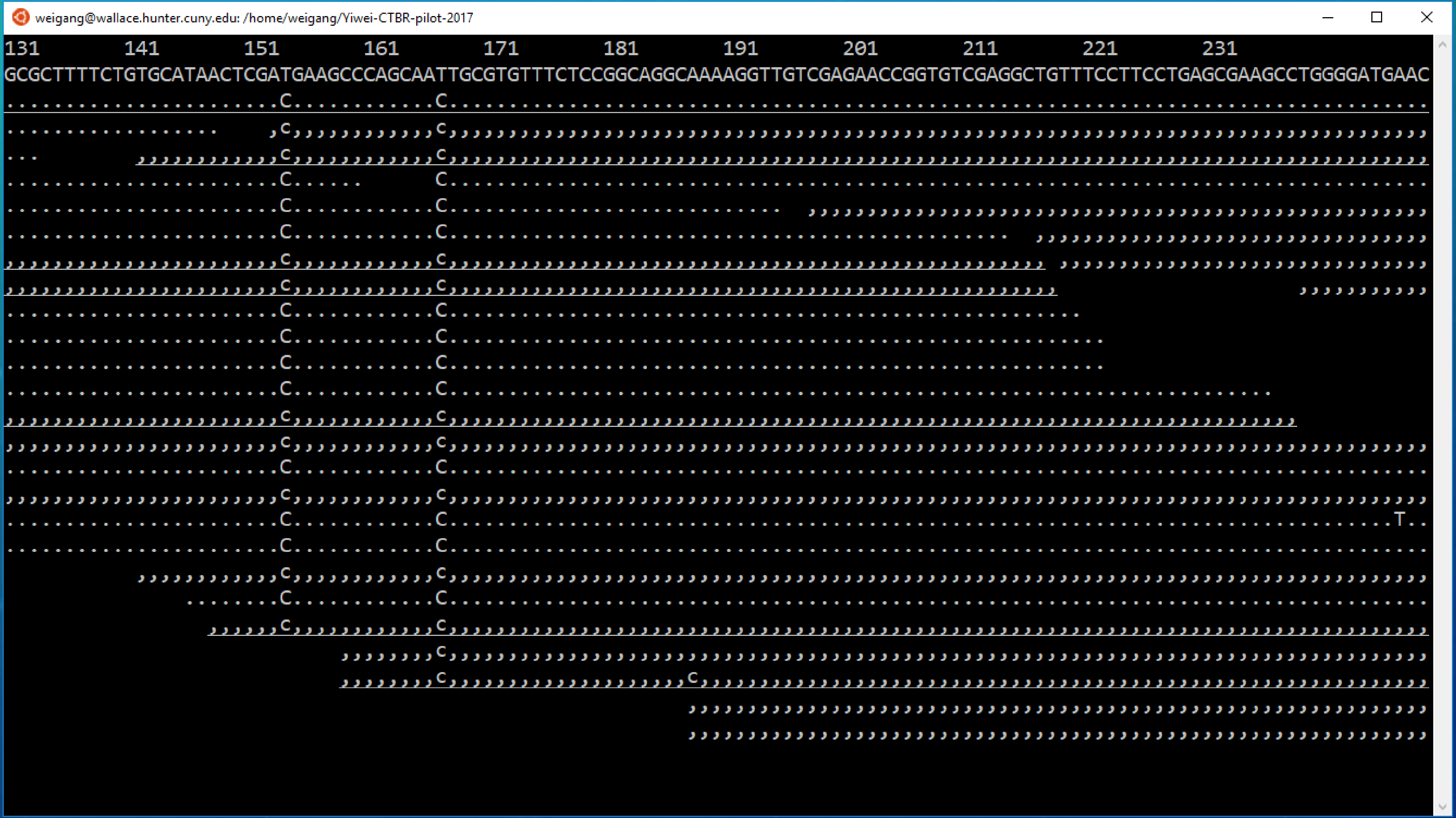
# NGS FILE FORMATS & SOFTWARE TOOLS
## Sequence Alignment/Map (SAM  & BAM)

```
SQ     SN:NC_021577    LN:6342034
@PG     ID:bwa  PN:bwa  VN:0.7.12-r1039 CL:bwa mem ref-pat5.fa 02015P1_S18_L001_R1_001.fastq.gz 02015P1_S18_L001_R2_001.fastq.gz
M04330:10:000000000-B85C4:1:1101:8493:1631      73      NC_021577       4611618 60      58M93S  =       4611618 0
CCCCGGAATAGGGCGAGGACGGGTCCTGGCGTGGCCCGTATACGTTGATGAAGCGGAATATCTCCTTTTCCATCCCATTCTTTCTTCTTTTTTAATCTATTTTTTACTCGCTTTCCAGCTTTTCCTACTC
TTTTTTCTTCATCTTCTACTT
6AAC9@7:FCA98CC@+++:+++@@@F,,6+B++669+@+,,6,:C,,,,,,,6++++,,96,96,,<966,,9,4,:,59,,,,,5,,<,,+,,95,,,,9,,8,5:5++6
M04330:10:000000000-B85C4:1:1101:8493:1631      133     NC_021577       4611618 0       *       =       4611618 0
CCTCTTCTTTCTACTTTCATTACTTCTAATTCTCTTCTTTCTACTCTTACTCCCCTTTTTCTCCTCTTACTCCCTACGCTTCCTTCTTTCTTTCCTTCTTTTTCTTCCTCTTTTTCTCTCTCCCCCTTCT
TTTTCTCTCTCCTTCCTTCTT -
6,6,,6,,;6<,6,,,6,,,,6,,6,,,,,6,6,,6,,6;,,;,,,,,;,666,,,,,;6,6,,;@,,6,;66;,;,+,4,,,,9,,,55,,,5,,5,,,5,45,4444,,
M04330:10:000000000-B85C4:1:1101:12247:1736     73      NC_021577       4067761 60      75M75S  =       4067761 0
TTCGAGGTCACCGGCTGCTCGCCGGTGTTCTGGATCATGAAGCAGCCTGTGCCGTAGGTTCTCTTCACCATGCCCTTCTTTAATCATTCCTTTCCTATCAGCTCTTCCTTCTGTTCTCCTTCCTTTCCCA
TCACCTTTCTCTCCTCTCCT
CCCCCFGGGGGGGDFG7@EFCCF7@F7FFEF,,,CC,C,,,,6,,6@,,:,,6+8,,,:,6:CEF9,:9,:,666,,9:,,,,,6,,,6<?,,69,,99,,5,5,,955,99
M04330:10:000000000-B85C4:1:1101:12247:1736     133     NC_021577       4067761 0       *       =       4067761 0
CTTCTTTTCAACATCCACAGCCTTTACTGGTTCTTTGTCTTTCTTTCTCTTTTCTTTATCCCTCTCATCCTTCTTCCTGTTTTTCTCTTCTTCTCTTCTTTTTTCTTCTTCTTCCTTCCTTCTCTTCTTT
TCTCTTCTTTCCCTTTTCT
8@,6@,6C<,,<,;;6,6,,,6,,,,;;,,,,6,,,,,,<;,;;,,6,6;,6;6,,,,66,,,6,;,,666,;@,;5,,,,,4,;55,,59,5,,,,,,,,,584,,4,49,
M04330:10:000000000-B85C4:1:1101:14009:1752     121     NC_021577       4891335 60      112S38M =       4891335 0
GAAGAAAAGAGAAAGAAAAAAGGAGAAAAAGAAGAAGAAAAAAAAAAAGTAAGAAGAGAAAGGAGAGAAAGAAAATGAACGAAAGGGAAAGAGAGGAGAAAGAAAGAACGCAGCTGAACAGCACCAGCAG
TTGCGCGCCGTTCACAGCAG
,,,3,6,,,,87,,,+7,,,,,,,4,,,,95,:,,,,+++@7+,5,,:5,,,6,,<,6,,,66,6,6,,<6,,,<,6C:,9,,@6,9,,C6,9,96,<CC<6,,C8,8+;,,
M04330:10:000000000-B85C4:1:1101:14009:1752     181     NC_021577       4891335 0       *       =       4891335 0
```

Software: samtools (Li et al, 2009. *Bioinformatics*)

# NGS FILE FORMATS

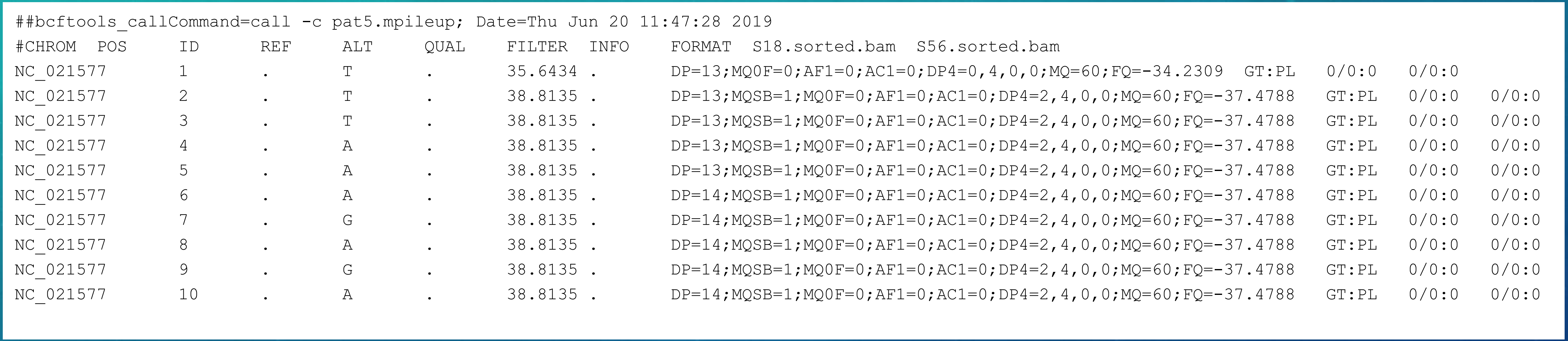
# Sequence Alignment/Map (mpileup/tiling)



weigang@wallace.hunter.cuny.edu: /home/weigang/Yiwei-CTBR-pilot-2017

```
131       141       151       161       171       181       191       201       211       221       231
GCGCTTTTCTGTGCATAACTCGATGAAGCCCAGCAATTGCGTGTTTCTCCGGCAGGCAAAAGGTTGTCGAGAACCGGTGTCGAGGCTGTTTCCTTCCTGAGCGAAGCCTGGGGATGAAC
.......................C............C..................................................................................
..................    ,c,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
...        ,,,,,,,,,,,,c,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
.......................C......      C..................................................................................
.......................C............C............................  ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
.......................C............C...............................................  ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,c,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,, ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,c,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,                    ,,,,,,,,,,,
.......................C............C.....................................................
.......................C............C.......................................................
.......................C............C.......................................................
.......................C............C.....................................................................
,,,,,,,,,,,,,,,,,,,,,,,c,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
,,,,,,,,,,,,,,,,,,,,,,,c,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
.......................C............C..................................................................................
,,,,,,,,,,,,,,,,,,,,,,,c,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
.......................C............C...............................................................................T..
.......................C............C..................................................................................
           ,,,,,,,,,,,,c,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
               ........C............C..................................................................................
                 ,,,,,,c,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
                            ,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
                            ,,,,,,,,c,,,,,,,,,,,,,,,,,,,,c,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
                                                         ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
                                                         ,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
```

# NGS FILE FORMATS & SOFTWARE TOOLS
# Variant Call Format (VCF)



```
##bcftools_callCommand=call -c pat5.mpileup; Date=Thu Jun 20 11:47:28 2019
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  S18.sorted.bam  S56.sorted.bam
NC_021577       1       .       T       .       35.6434 .       DP=13;MQ0F=0;AF1=0;AC1=0;DP4=0,4,0,0;MQ=60;FQ=-34.2309  GT:PL   0/0:0   0/0:0
NC_021577       2       .       T       .       38.8135 .       DP=13;MQSB=1;MQ0F=0;AF1=0;AC1=0;DP4=2,4,0,0;MQ=60;FQ=-37.4788   GT:PL   0/0:0   0/0:0
NC_021577       3       .       T       .       38.8135 .       DP=13;MQSB=1;MQ0F=0;AF1=0;AC1=0;DP4=2,4,0,0;MQ=60;FQ=-37.4788   GT:PL   0/0:0   0/0:0
NC_021577       4       .       A       .       38.8135 .       DP=13;MQSB=1;MQ0F=0;AF1=0;AC1=0;DP4=2,4,0,0;MQ=60;FQ=-37.4788   GT:PL   0/0:0   0/0:0
NC_021577       5       .       A       .       38.8135 .       DP=13;MQSB=1;MQ0F=0;AF1=0;AC1=0;DP4=2,4,0,0;MQ=60;FQ=-37.4788   GT:PL   0/0:0   0/0:0
NC_021577       6       .       A       .       38.8135 .       DP=14;MQSB=1;MQ0F=0;AF1=0;AC1=0;DP4=2,4,0,0;MQ=60;FQ=-37.4788   GT:PL   0/0:0   0/0:0
NC_021577       7       .       G       .       38.8135 .       DP=14;MQSB=1;MQ0F=0;AF1=0;AC1=0;DP4=2,4,0,0;MQ=60;FQ=-37.4788   GT:PL   0/0:0   0/0:0
NC_021577       8       .       A       .       38.8135 .       DP=14;MQSB=1;MQ0F=0;AF1=0;AC1=0;DP4=2,4,0,0;MQ=60;FQ=-37.4788   GT:PL   0/0:0   0/0:0
NC_021577       9       .       G       .       38.8135 .       DP=14;MQSB=1;MQ0F=0;AF1=0;AC1=0;DP4=2,4,0,0;MQ=60;FQ=-37.4788   GT:PL   0/0:0   0/0:0
NC_021577       10      .       A       .       38.8135 .       DP=14;MQSB=1;MQ0F=0;AF1=0;AC1=0;DP4=2,4,0,0;MQ=60;FQ=-37.4788   GT:PL   0/0:0   0/0:0
```

Software: vcftools & bcftools

# NGS GLOSSARY

Library prep:

Sequence library: the collection of target DNA fragmented

Barcoding: Specific sequence that identifies the sample a particular read.

Multiplexing: Multiplex sequencing allows large numbers of libraries to be pooled and sequenced simultaneously in a single run

Genome assembly:

Reads: output DNA bases from both Sanger and next-generation methods

Single-end reads: reads that align to only one end of a DNA fragment.

Paired-end reads: reads that align to both ends of a DNA fragment.

Coverage: The number of sequence reads aligned to positions that cover a specific base on a target genome, or the average number of aligned reads that overlap all positions on the target genome.

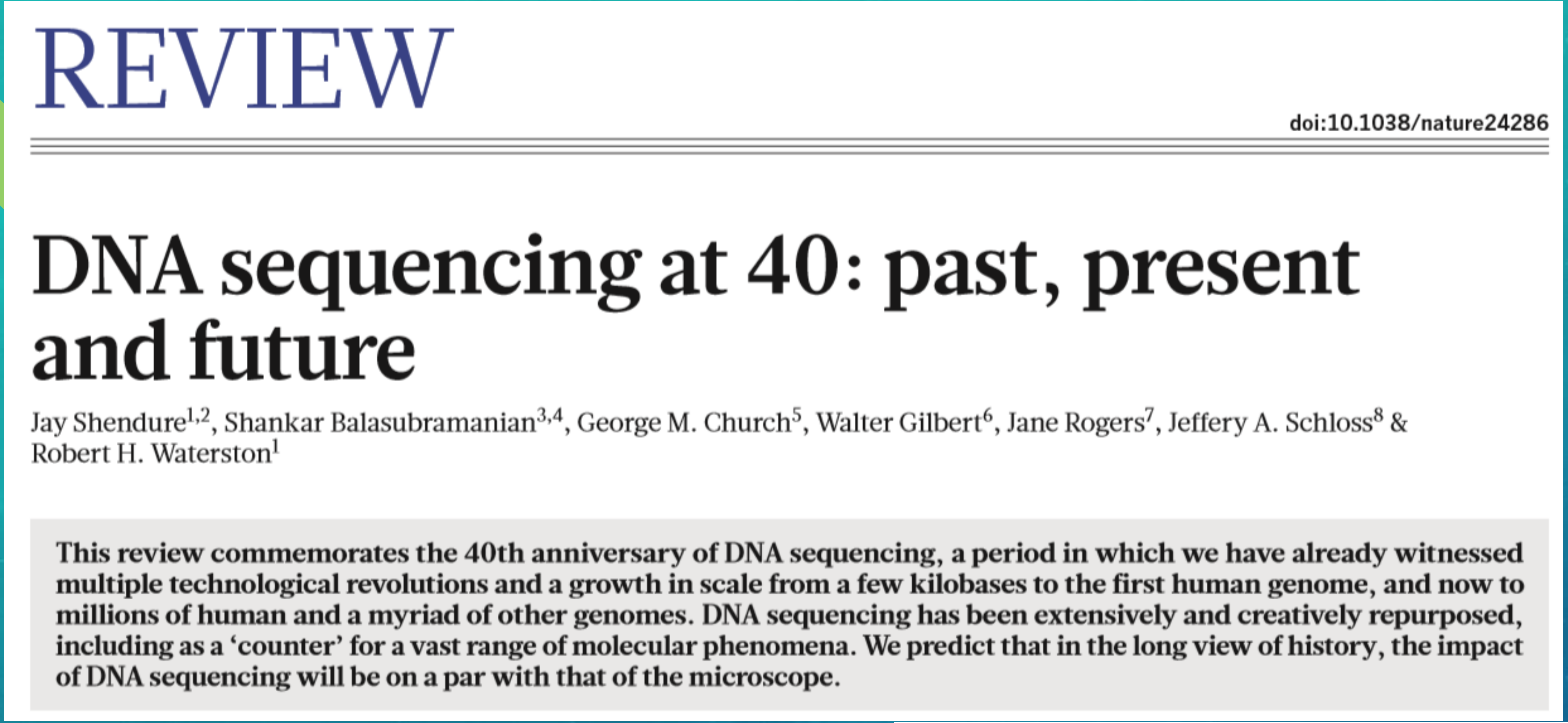
Contig: A contiguous stretch of DNA sequence that is the result of assembly of multiple overlapping sequence reads into a single consensus sequence.

Scaffold: Contigs separated by long (~10 kb) gaps, identified by mate-paired reads

*De Novo* assembly: genome assembled without the use of a known reference sequence.

Referenced assembly: genome assembled with the use of a known reference sequence.

# KEY REFERENCES



REVIEW

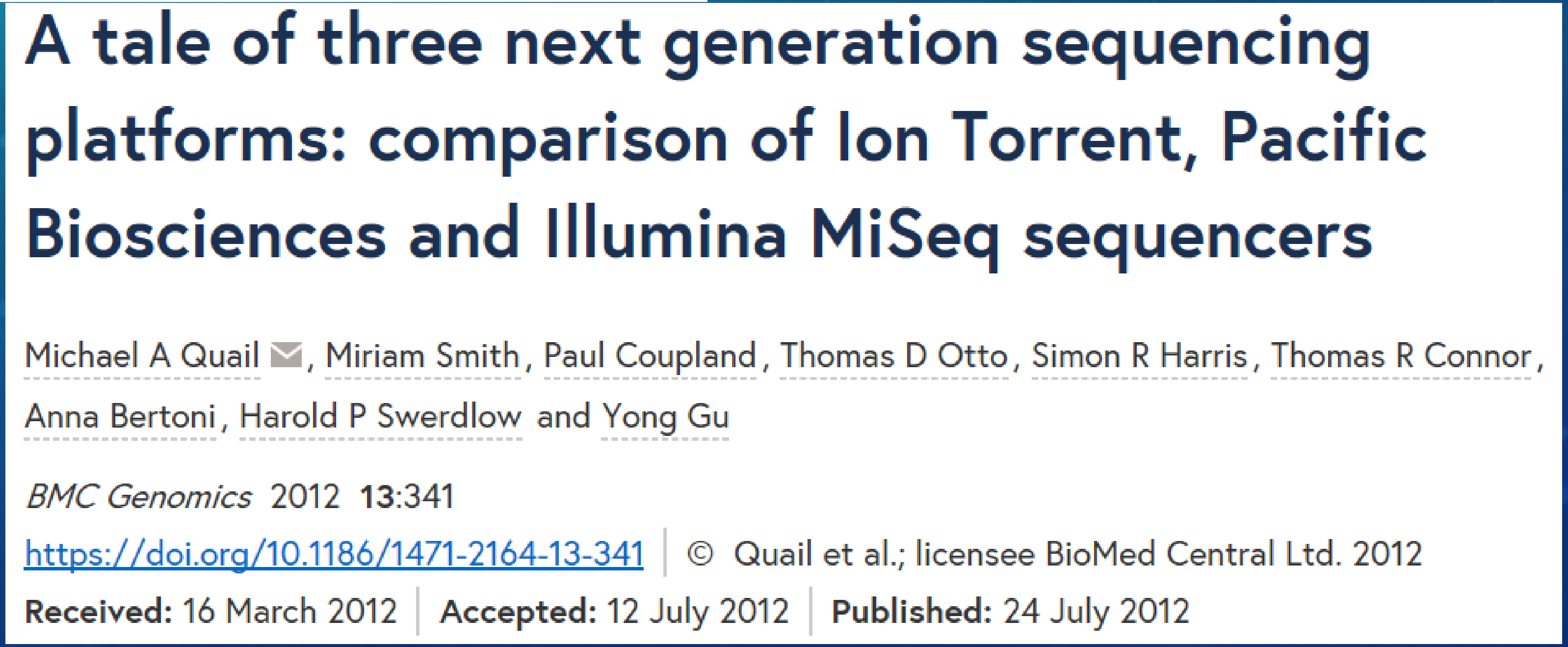
doi:10.1038/nature24286

## DNA sequencing at 40: past, present and future

Jay Shendure[1,2], Shankar Balasubramanian[3,4], George M. Church[5], Walter Gilbert[6], Jane Rogers[7], Jeffery A. Schloss[8] & Robert H. Waterston[1]

**This review commemorates the 40th anniversary of DNA sequencing, a period in which we have already witnessed multiple technological revolutions and a growth in scale from a few kilobases to the first human genome, and now to millions of human and a myriad of other genomes. DNA sequencing has been extensively and creatively repurposed, including as a ‘counter’ for a vast range of molecular phenomena. We predict that in the long view of history, the impact of DNA sequencing will be on a par with that of the microscope.**

Shendure *et al. Nature* 1–9 (2017)



## A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers

Michael A Quail ✉, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow and Yong Gu

*BMC Genomics* 2012 **13**:341

https://doi.org/10.1186/1471-2164-13-341 | © Quail et al.; licensee BioMed Central Ltd. 2012

**Received:** 16 March 2012 | **Accepted:** 12 July 2012 | **Published:** 24 July 2012

Quail et al (2012). *BMC Genomics*

# DISSECTING AN NGS-BASED RESEARCH PAPER

- **Search engine: PubMed (https://pubmed.ncbi.nlm.nih.gov/). Restrict on: year; associated data; open access**
- **Background**: Research question, Significance & Hypothesis
- Materials & Samples: sample sizes, positive and negative controls
- NGS platform: Illumina, PacBio, Nanopore
- Nature of NGS data
  - Whole genome/Exome
  - Transcriptome (Bulk RNA-Seq)
  - Single-cell transcriptomes
  - Microbiome (16S rRNA); Meta-genomes (community genomics, for uncultivable species)
  - CHIP-Seq (transcription binding sites; gene regulation)
  - Proteomics (SiLAC)
- Methods/Computation
  - Programs were used to generate NGS data
  - Programming languages: Linux/BASH, R/Rstudio/R packages, Python/Python module
- Methods/Statistical analysis: t-test, clustering, principle-component analysis (PCA), linear regression? ANOVA?
- Data & Code availability (Do NOT pick paper without Full dataset)
  - Raw data & sequencing reads: GenBank accession numbers (BioProject; SRA; GEO)
  - Processed data sets: e.g., genome assemblies; mRNA or species counts (in Excel format)
  - Code: iPython Notebook; R markdowns (Github)
- Results/Visualization: Data visualization: heatmap, barplot, boxplot, or scatterplot?
- Results/Statistical analysis: Data visualization: heatmap, barplot, boxplot, or scatterplot?
- Conclusion: Biological conclusions
- Full citation: First author (last name only) *et al* (2020). *Journal name* & page, URL