NGS:
technologies and challenges

Olga Vinnere Pettersson, PhD

Project Coordinator

NGI/SciLifeLab, Uppsala Node

@OlgaVPettersson

Version 8.2

Outline

INTRO

- Sequencing service at NGI-SciLifeLab

NGS general knowledge:

- History of NGS
- Current technologies

NGS Challenges:

- Sequencing artefacts
- NGS sample quality requirements

- Philosophical reflection upon NGS analysis

GANATIONALCTAC
ATCAGENOMICSGT
INFRASTRUCTURE

GANATIONALCTAC
ATCAGENOMICSGT
INFRASTRUCTURE

SciLifeLab

We are non-profit

We have technology and knowledge

We want to help you to do GREAT research

We do not want co-authorship

Let us help YOU

Operational principles of NGI

User community

- Open to all Swedish academic scientists on equal terms.
- Consultation and introduction of new protocols.
- Workshops, courses, seminars.

Cost basis

- Academic users of NGI only cover reagent cost.
- Staff salaries at NGI covered by SciLifeLab, VR, and host universities
- Premises and service contracts covered by SciLifeLab, VR, KAW and host universities.
- Capital equipment covered by KAW, VR, SciLifeLab.

Quality

- Emphasis on data quality and needs of the users.
- Illumina sequencing and genotyping processes accredited by SWEDAC, ISO/IEC 17025
- Ion and PacBio: accreditation due 2017

NATIONAL GENOMICS INFRASTRUCTURE

SciLifeLab

NGI Support

Pre-sequencing

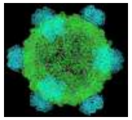
- **Project design** via discussions with users
- **Advise** in sample collection and preparation
- Case-to-case **DNA extraction service**

Post-sequencing:

- Control over produced data: making sure data meet our **high standards** in terms of quality and yield.
- Primary **analysis of human genomes** is enabled
- **Genome assembly** of PacBio data is offered as a service
- Data is delivered to **UPPMAX** (Uppsala Multidisciplinary Center for Advanced Computational Science)

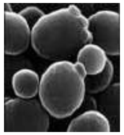
Collaborative projects for technology and method development

Education

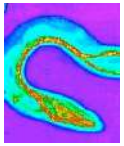


First genome: virus ϕ X 174 - 5 368 bp (1977)

First organism: *Haemophilus influenzae* - 1.5 Mb (1995)



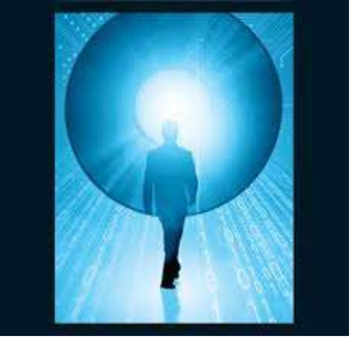
First eukaryote: *Saccharomyces cerevisiae* - 12.4 Mb (1996)



First multicellular organism: *Cenorhabditis elegans* - 100 Mb (1998-2002)

First plant: *Arabidopsis thaliana* - 157 Mb (2000)

First human genome- 3Gb (2003)



… paradigm changes

- From single genes to complete genomes
- From single transcripts to whole transcriptomes
- From single organisms to complex metagenomic pools
- From model organisms to the species you are studying
- Personal genome = personalized medicine

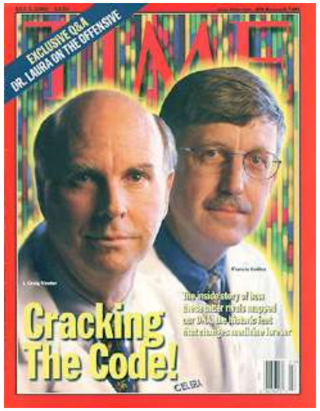
An interesting comparison…

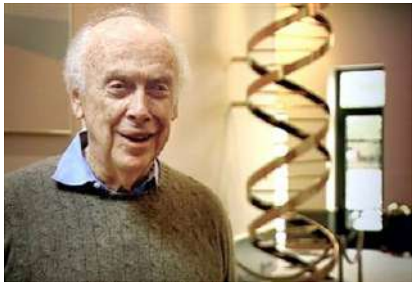
Human genome project (HUGO)
Sanger Sequencing
2.7 Billion USD

Craig Venter’s Genome
Sanger Sequencing
70 Million USD

James Watson’s Genome
454 pyro sequencing (Roche)
2 Million USD

Today’s genome
NovaSeq(Illumina)
~1 000 USD





Current Technologies

illumina®

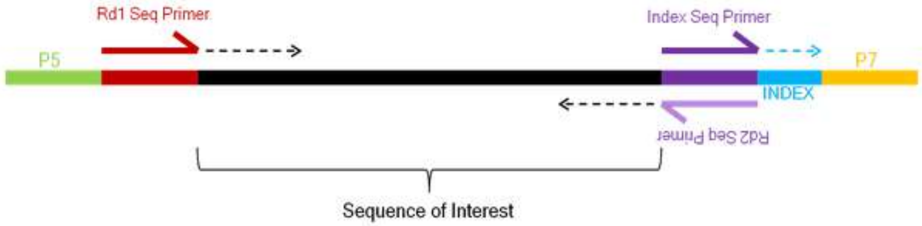
Current leader on the NGS market

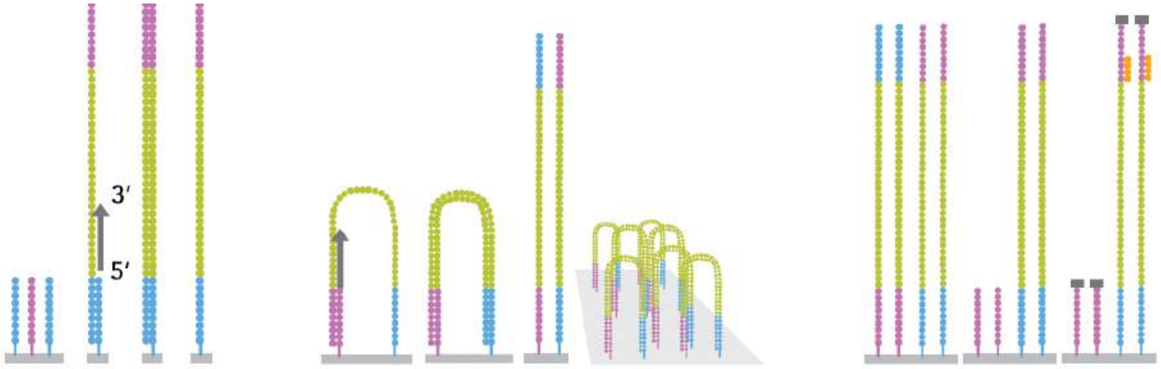
Instrument	Run time	Max output	Max reads/run	Max read length
iSeq	9.5 – 19 hrs	1.2 Gb	4 mln	PE 150
MiniSeq	4-24 hrs	7.5 Gb	25 mln	PE 150
MiSeq	4-55 hours	15 Gb	25 mln	PE 300
NextSeq series	12-48 hours	120-300 Gb	0.4 – 1 bln	PE 150
NovaSeq 6000	13-44 hours	6 Tb	20 bln	PE 250

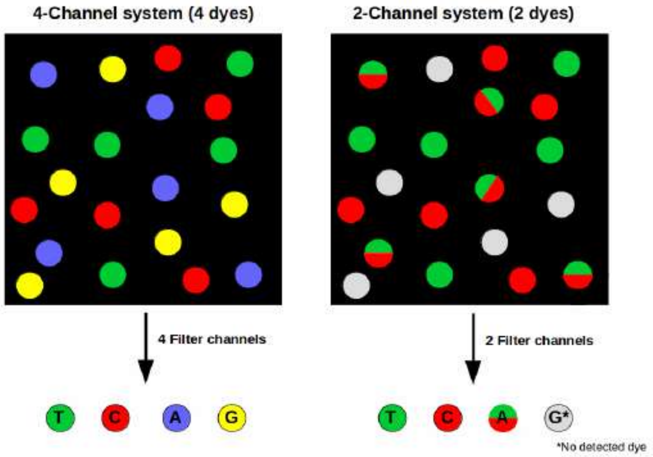
RIP: HiSeq 2500 & HiSeq X

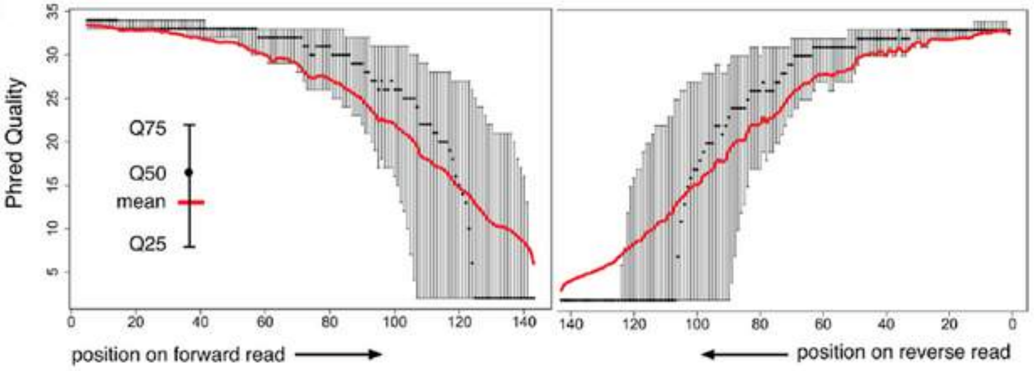
Used for everything

Illumina: bridge amplificatior



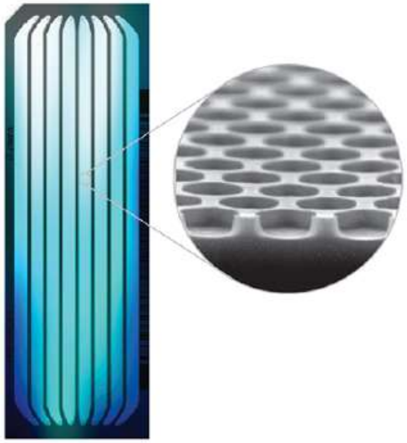


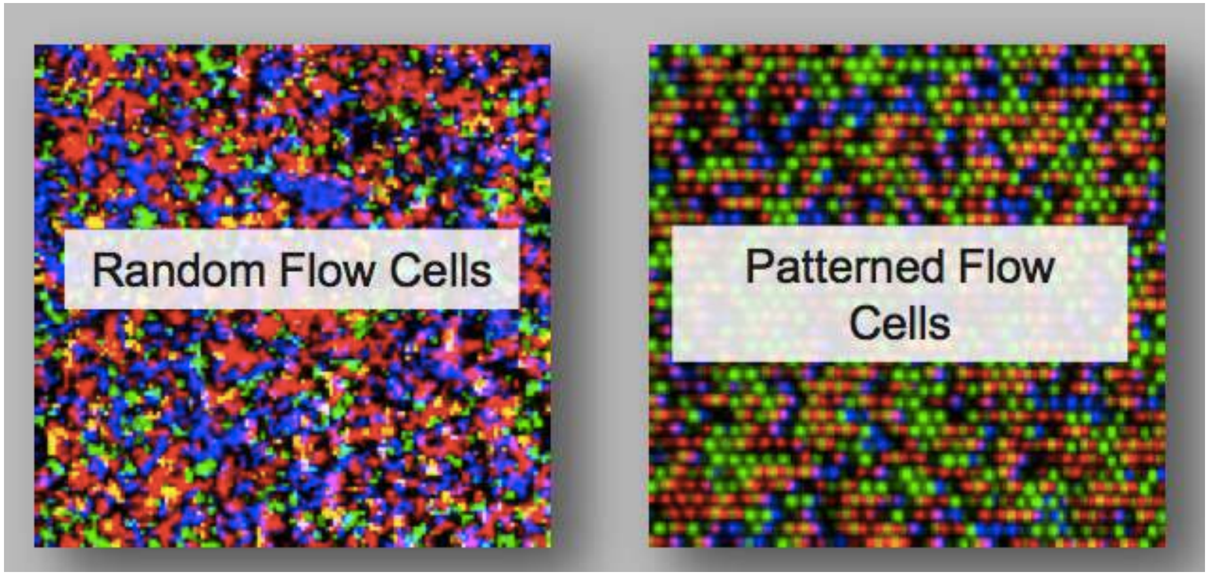




Illumina sequencing: before vs now

Patterned flow cells introduced on HiSeq X and NextSeq systems





MiSeq flow cells still do not have a patterm

ThermoFisher SCIENTIFIC

Ion S5 XL

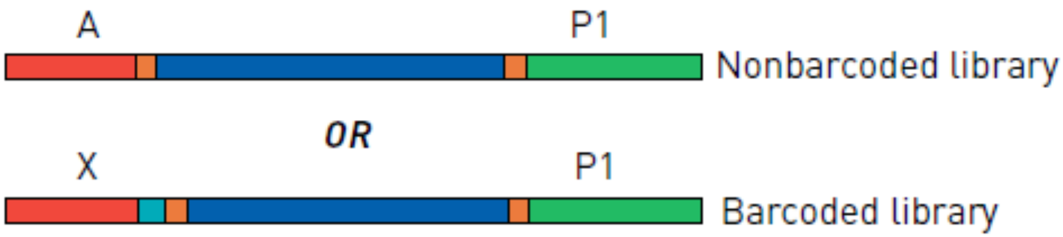
Chip:	Run time	Output	Max reads/ run	Max read length
510	2.5-4 hrs	0.3 - 0.5 Gb	2-3 mln	SE 400 bp
520	2.5-4 hrs	0.6-2 Gb	3-6 mln	SE 600 bp
530	2.5-4 hrs	3-8 Gb	15-20 mln	SE 600 bp
540	2.5-4 hrs	10-15 Gb	60-80 mln	SE 400 bp
550	2.5-4 hrs	18-20 Gb	100-130 mln	SE 200 bp

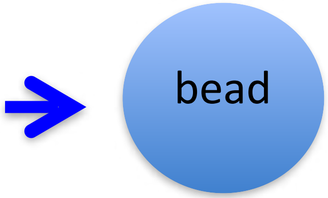
RIP: IonTorrent PGM, IonProton

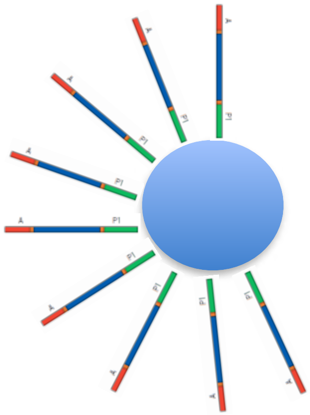
Clinical applications mainly
Standard analysis directly on the instrument

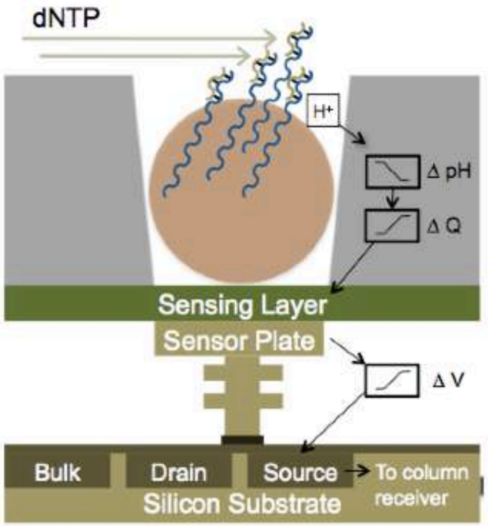
Multiplex-PCR panels

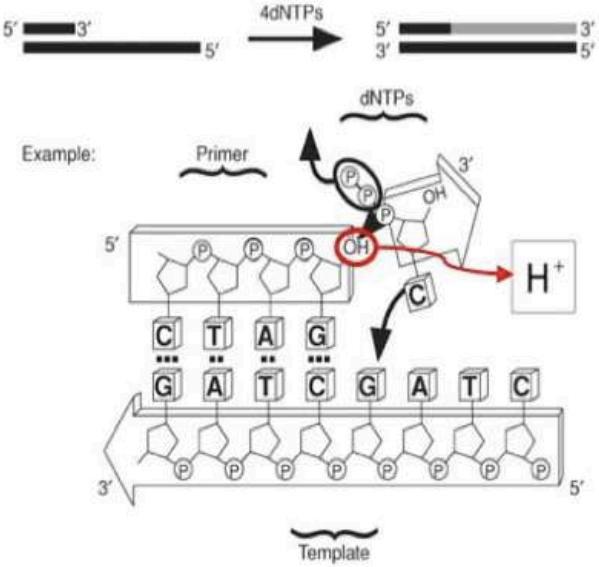
Ion Torrent: H+ ion-sensitive field effect transistors

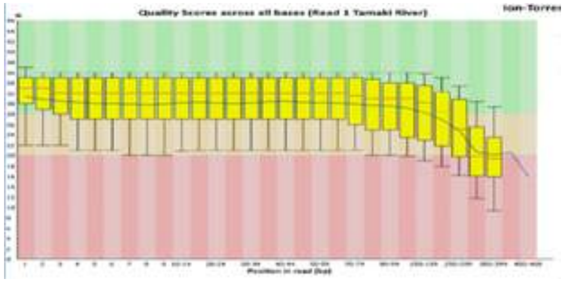












PACBIO®

Instrument	Run time /SMRT	Output /SMRT	Max reads / SMRT	Max read length*
RSII	30 min – 6 hrs	500 Mb – 2 Gb	50 000	40 kb
Sequel	30 min – 20 hrs	2 – 35 Gb	200 000	60 kb
Sequel II				
HiFi	30 hrs	*320 Gb*	*4 mln*	25 kb
CLR	15 hrs	*300 Gb*	*3 mln*	120 kb

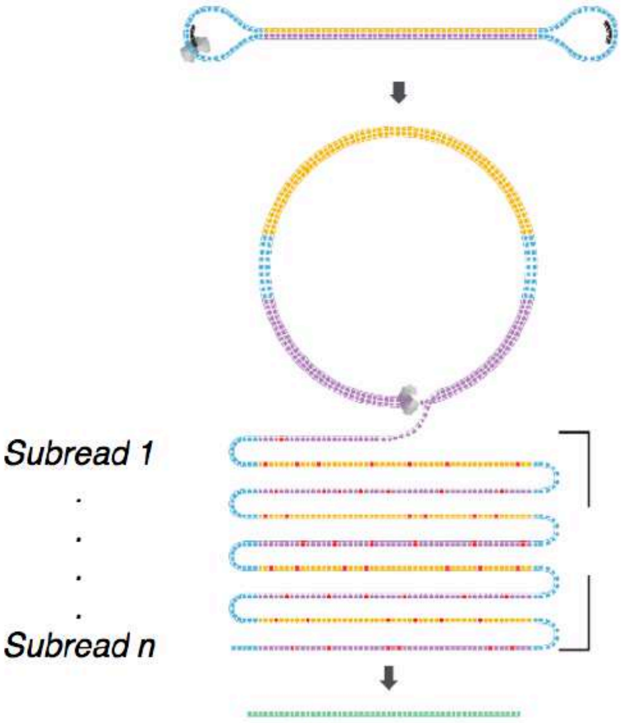
Single Molecule Real Time sequencing: SMRT

PacBio

TWO MODES OF SMRT SEQUENCING

Circular Consensus Sequencing (CCS) Mode

Inserts 10-20 kb



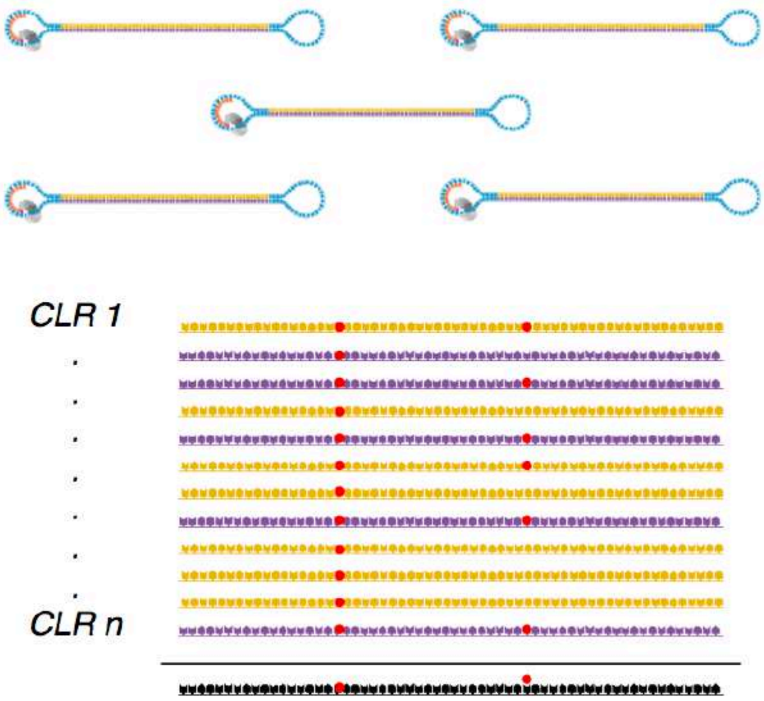
Qv 20+

HiFi READS

Single-molecule consensus sequence

Continuous Long Read (CLR) Sequencing Mode

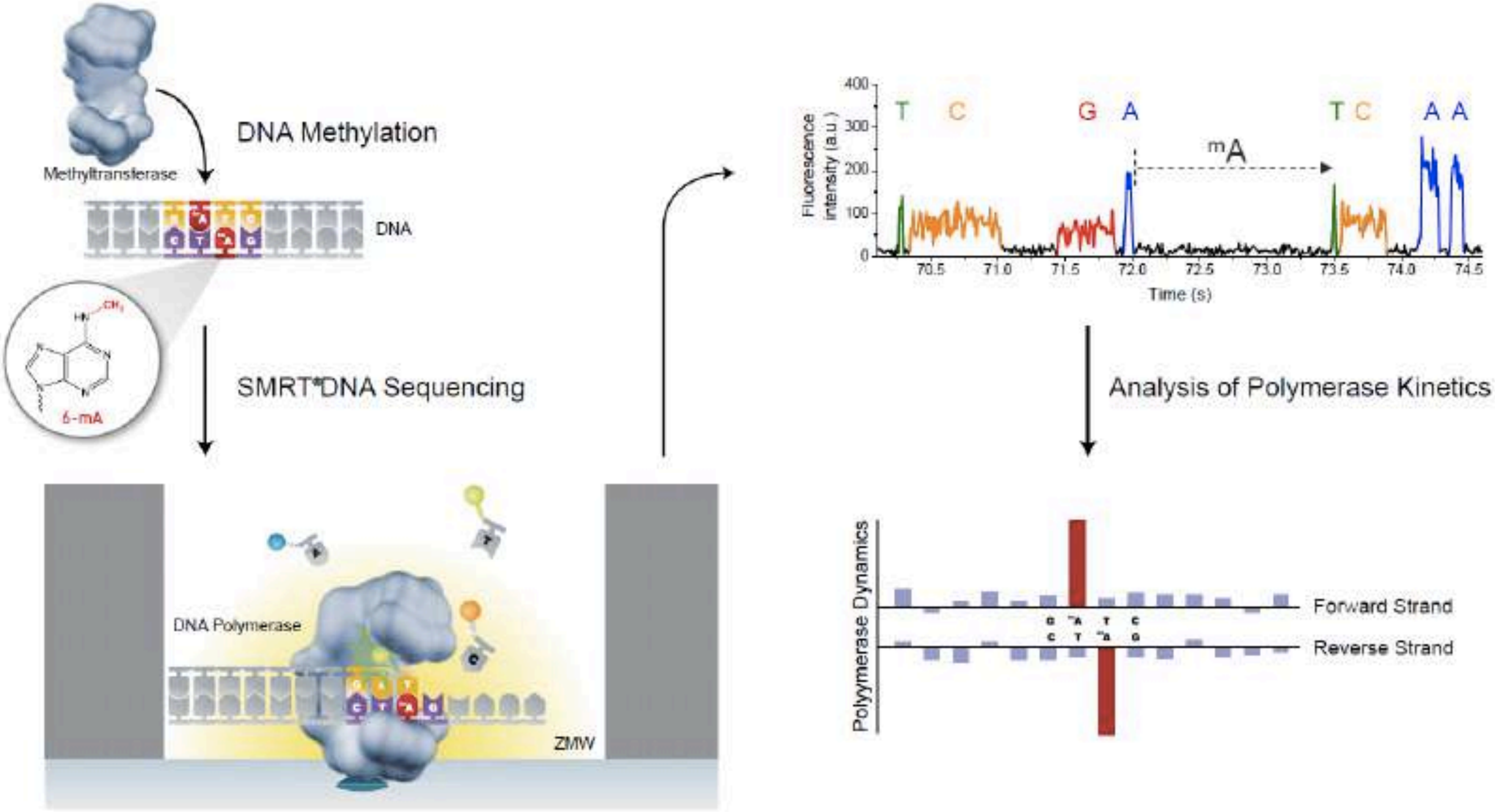
Inserts >25 kb, up to 175 kb



LONG READS

Multi-molecule consensus sequence

Base Modification: Discover the Epigenome



DNA Methylation

SMRT®DNA Sequencing

Analysis of Polymerase Kinetics

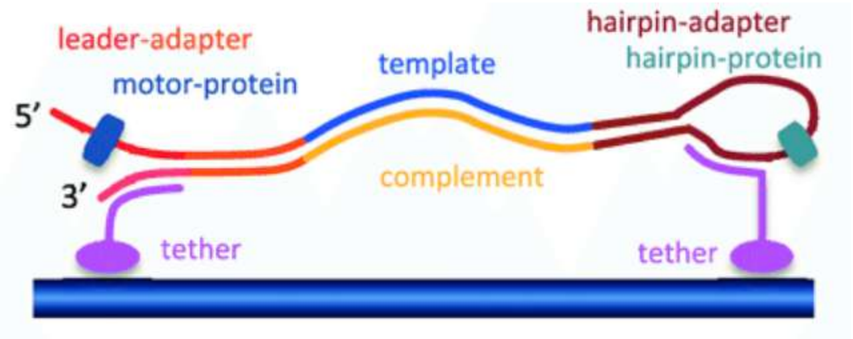
Detect base modifications using the kinetics of the polymerization reaction during normal sequencing

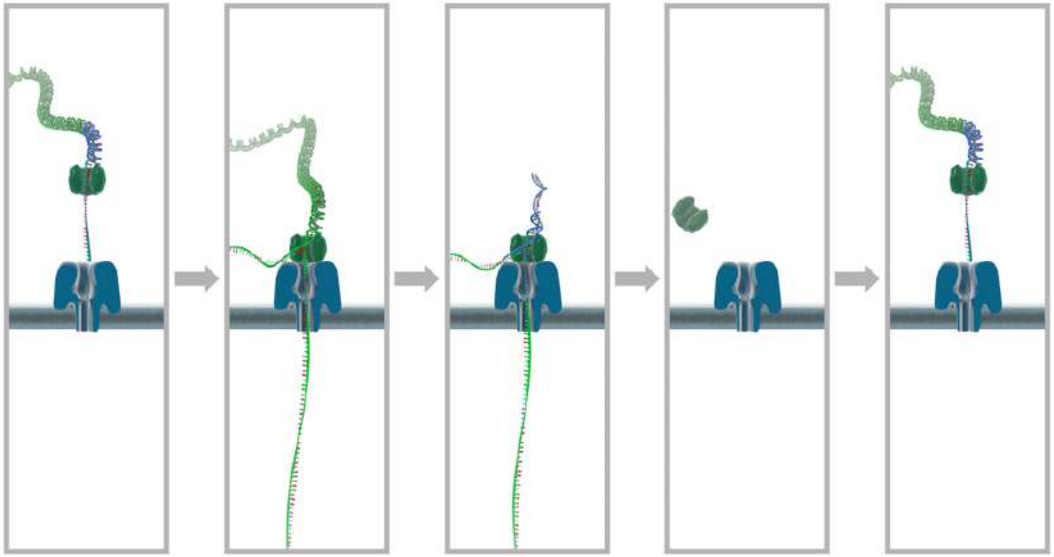
Oxford NANOPORE Technologies

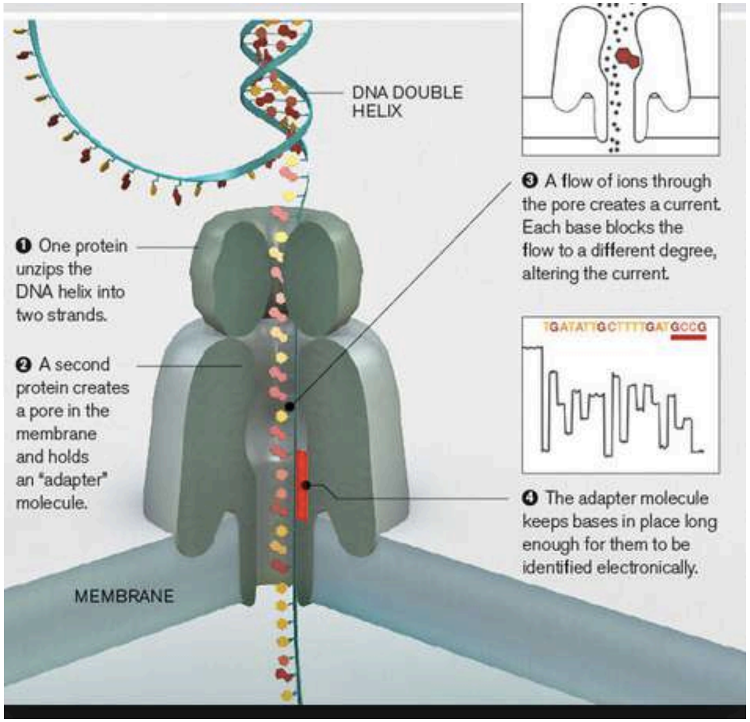
Instrument	Run time /FC	Output / FC	Nr of pores	Max read length
Flongle	16 hrs	1 Gb	126	1 Mb
MinION	24 hrs	2-15 Gb	512	1 Mb
GridION	24 hrs	2-15 Gb	512	1 Mb
PromethION	72 hrs	10 – 150 Gb	3 000	2 Mb

Q&A: *"It depends"…*

ONT: DNA + Motor + Pore







❶ One protein unzips the DNA helix into two strands.

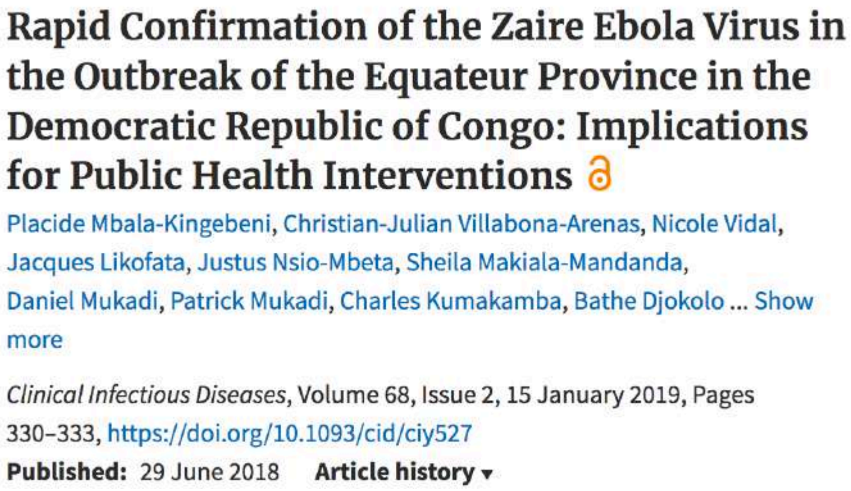
❷ A second protein creates a pore in the membrane and holds an "adapter" molecule.

❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.

❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.

Base modification info is retained

Main advantages of ONT: SPEED and PORTABILITY

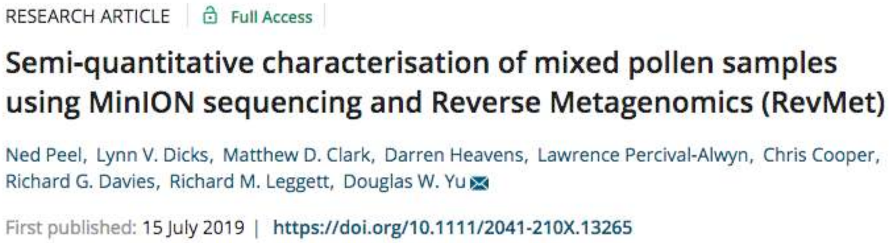


Rapid Confirmation of the Zaire Ebola Virus in the Outbreak of the Equateur Province in the Democratic Republic of Congo: Implications for Public Health Interventions

Placide Mbala-Kingebeni, Christian-Julian Villabona-Arenas, Nicole Vidal, Jacques Likofata, Justus Nsio-Mbeta, Sheila Makiala-Mandanda, Daniel Mukadi, Patrick Mukadi, Charles Kumakamba, Bathe Djokolo ... Show more

Clinical Infectious Diseases, Volume 68, Issue 2, 15 January 2019, Pages 330–333, https://doi.org/10.1093/cid/ciy527

Published: 29 June 2018 **Article history ▾**

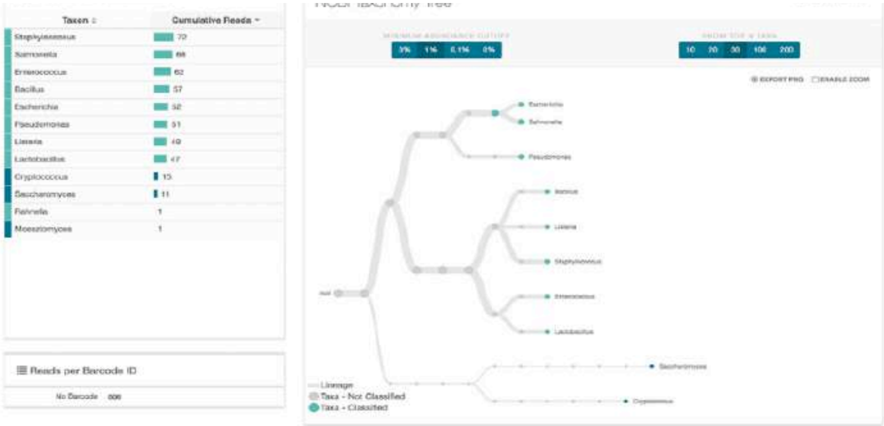


RESEARCH ARTICLE | Full Access

Semi-quantitative characterisation of mixed pollen samples using MinION sequencing and Reverse Metagenomics (RevMet)

Ned Peel, Lynn V. Dicks, Matthew D. Clark, Darren Heavens, Lawrence Percival-Alwyn, Chris Cooper, Richard G. Davies, Richard M. Leggett, Douglas W. Yu

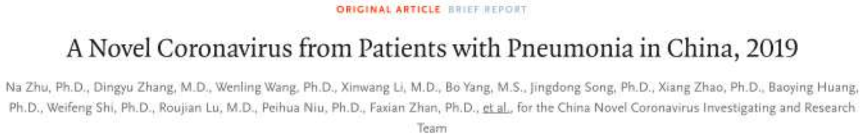
First published: 15 July 2019 | https://doi.org/10.1111/2041-210X.13265



Taxon	Cumulative Reads
Staphylococcus	72
Salmonella	68
Enterococcus	62
Bacillus	57
Escherichia	52
Pseudomonas	51
Listeria	49
Lactobacillus	47
Cryptococcus	15
Saccharomyces	11
Rahnella	1
Moesziomyces	1

Reads per Barcode ID

No Barcode 606

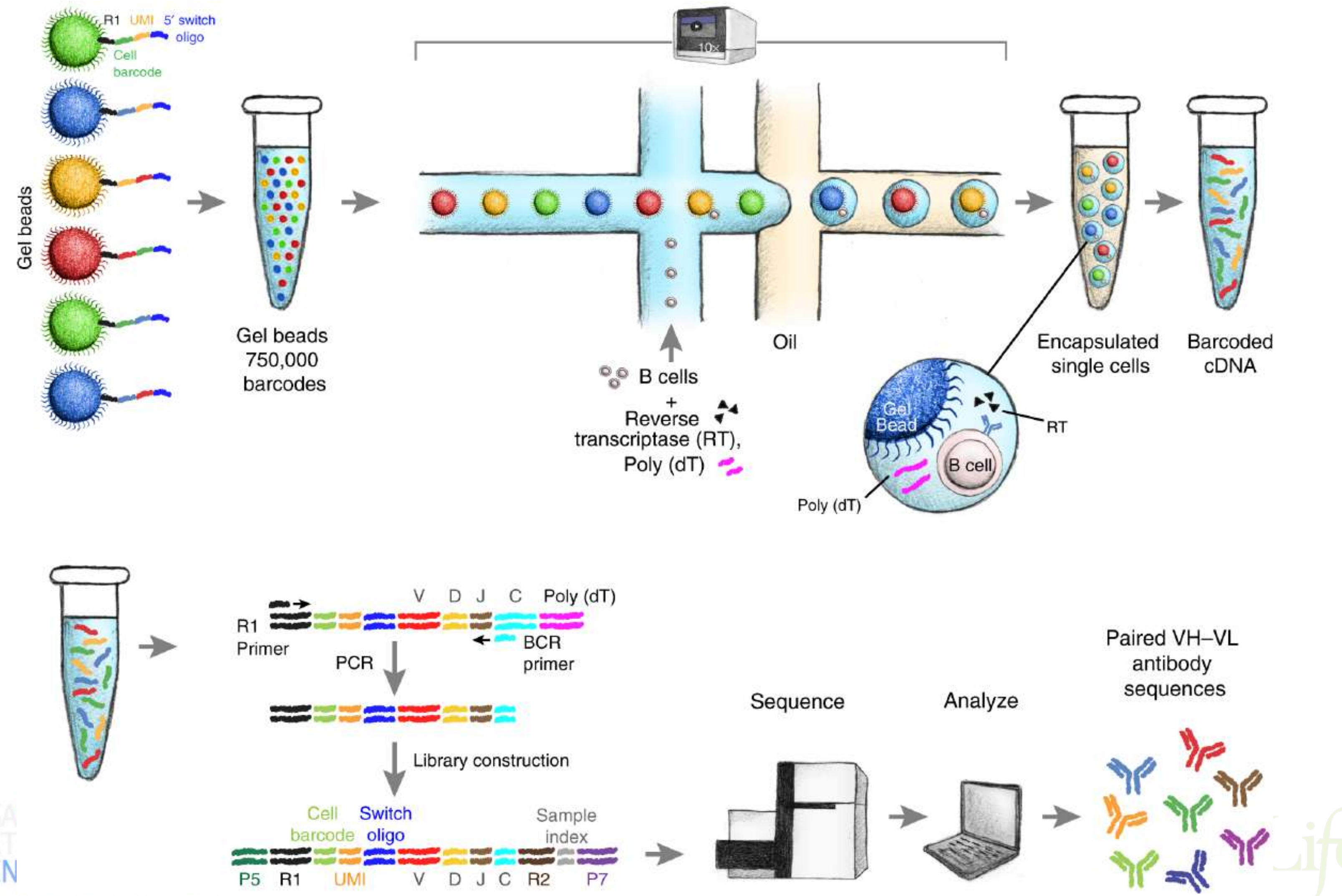


ORIGINAL ARTICLE BRIEF REPORT

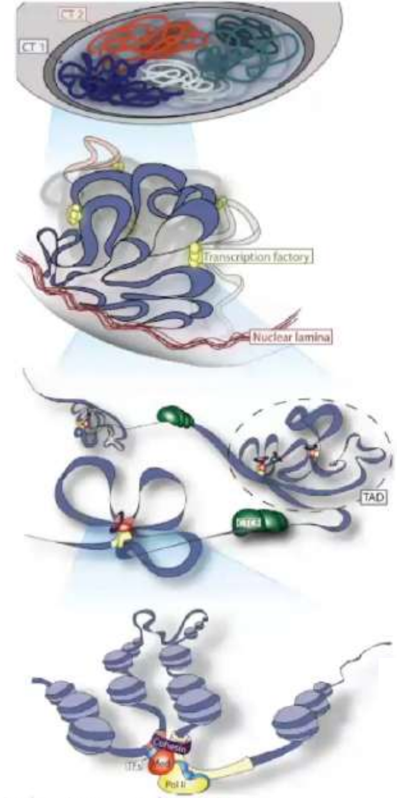
A Novel Coronavirus from Patients with Pneumonia in China, 2019

Na Zhu, Ph.D., Dingyu Zhang, M.D., Wenling Wang, Ph.D., Xinwang Li, M.D., Bo Yang, M.S., Jingdong Song, Ph.D., Xiang Zhao, Ph.D., Baoying Huang, Ph.D., Weifeng Shi, Ph.D., Roujian Lu, M.D., Peihua Niu, Ph.D., Faxian Zhan, Ph.D., et al., for the China Novel Coronavirus Investigating and Research Team

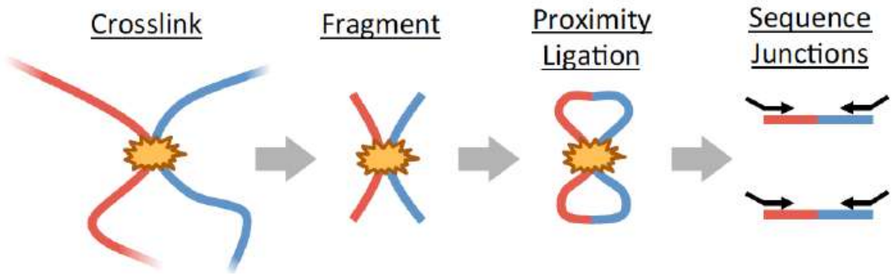
10x Genomics (Chromium)

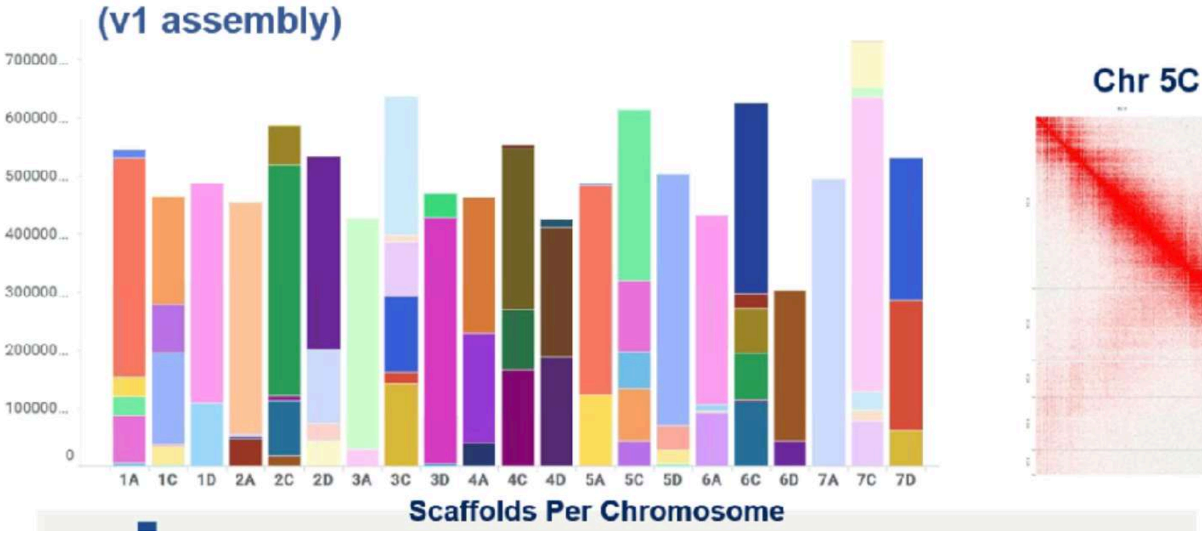


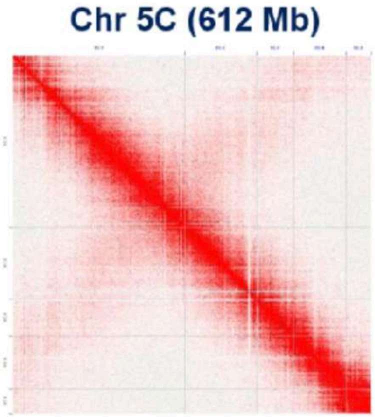
Hi-C / OmniC:
linking reads to chromosomes



Gorkin, Leung and Ren, 2014





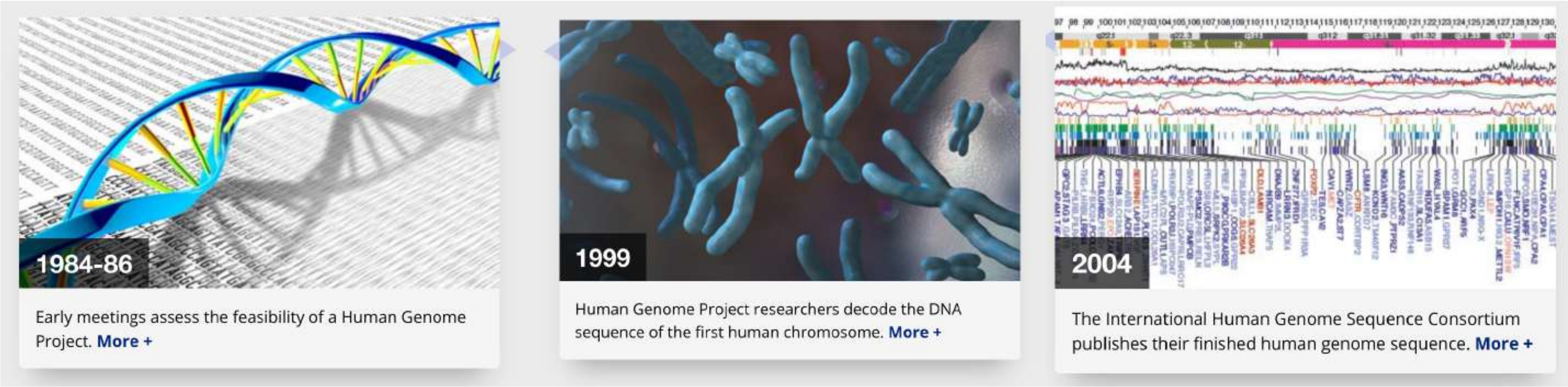


Start with a tissue!

Capture DNA bound to the same nucleosome

Make a library and sequence on Illumina NovaSeq

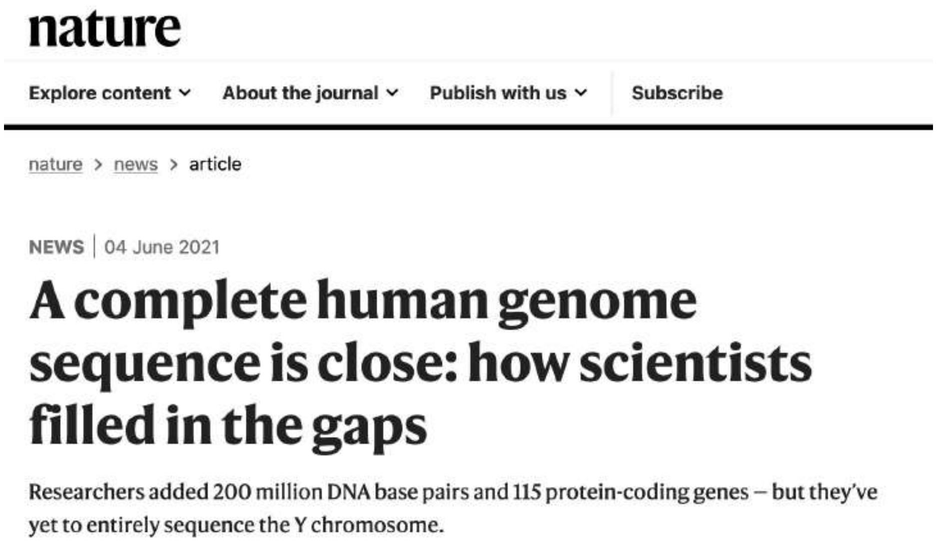
Human genome project



Early meetings assess the feasibility of a Human Genome Project. More +

Human Genome Project researchers decode the DNA sequence of the first human chromosome. More +

The International Human Genome Sequence Consortium publishes their finished human genome sequence. More +

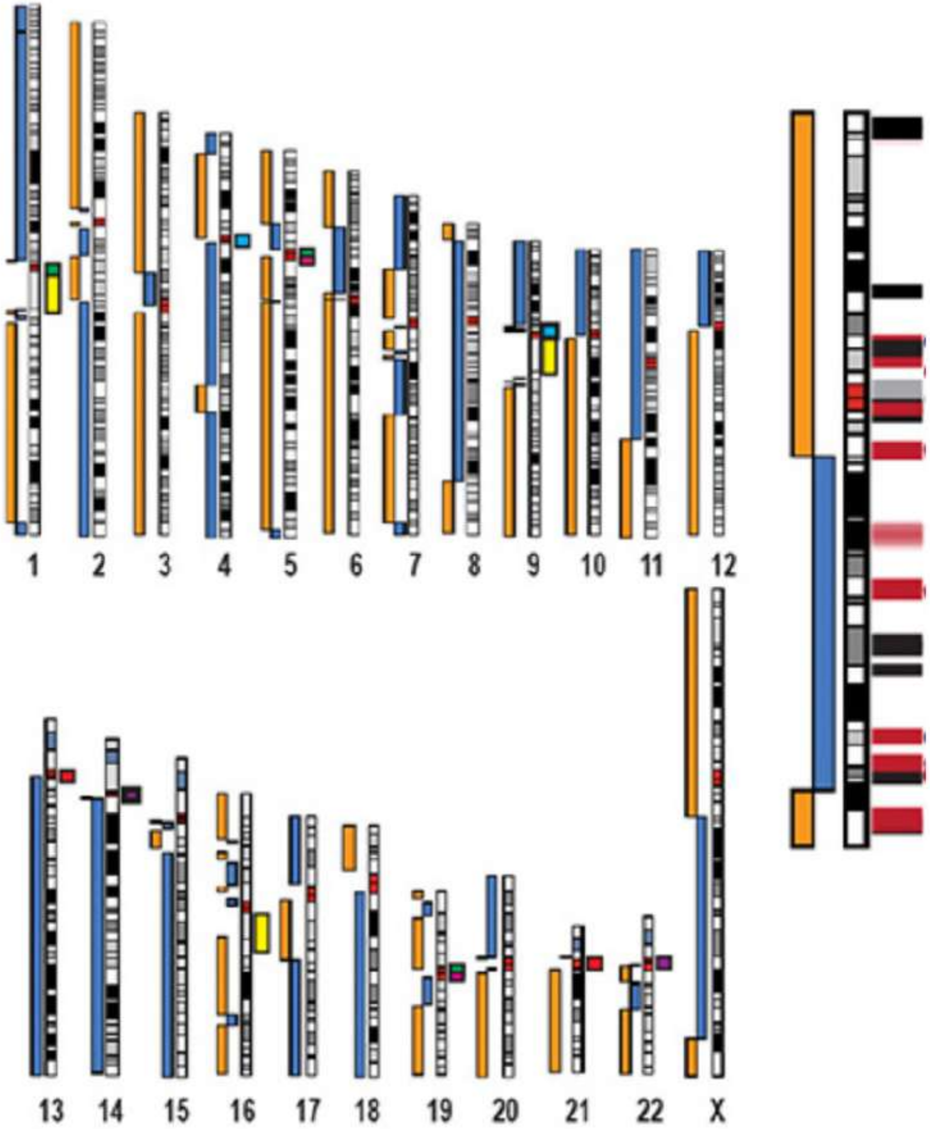


NEWS | 04 June 2021

A complete human genome sequence is close: how scientists filled in the gaps

Researchers added 200 million DNA base pairs and 115 protein-coding genes – but they've yet to entirely sequence the Y chromosome.

Zooming into the dark matter:



Telomere-to-telomere assemblies are now achieved with long reads

Technologies and Applications at NGI

NGS technologies

Short read NGS

Whole genome re-sequencing
RNA-seq
Exome
Targeted re-seq
Panels
Amplicons up to 600 bp

Long-read NGS

De novo genome sequencing
Whole-transcript sequencing
Structural variant resolving
Allele phasing
Targeted re-seq
Amplicons up to 13 kb

Research and development

SWEDAC · ACKREDITERING ·
Ackred. nr 1239
Provning
ISO/IEC 17025

NGS Technologies: SUMMARY

- Development goes VERY FAST
- All technologies have their PROs and CONs
- One technology does not suit all the applications
- In some projects, several technologies should be combined

BREAK

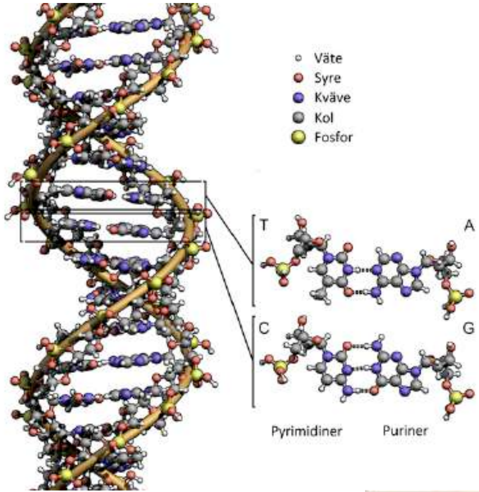
Making sense of genomics data:

Understanding sequencing bias

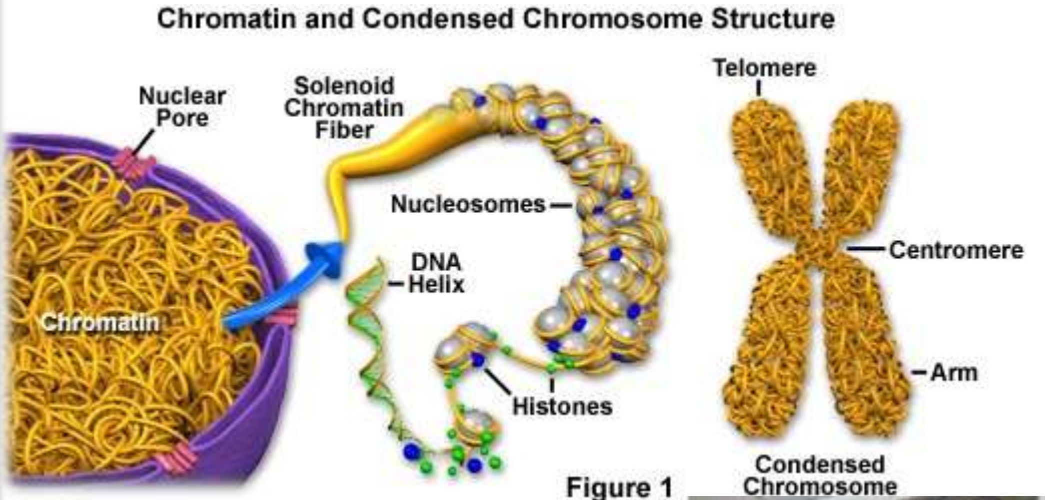
You do not see them before it is too late

Sequencing artefacts: what are they?

Sequencing a representative, completely randomized subsample:
it starts with input material



What textbook tells you

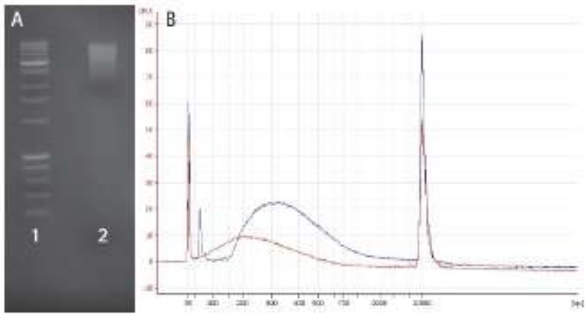


Brutal reality

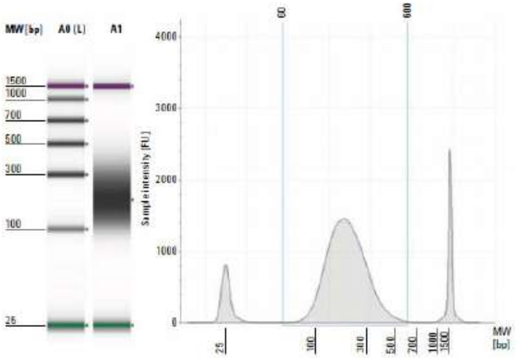
Do not forget:
DNA in solution behaves differently

Sequencing artefacts: what are they?

Sequencing a representative, completely randomized subsample:
continues with library preparation

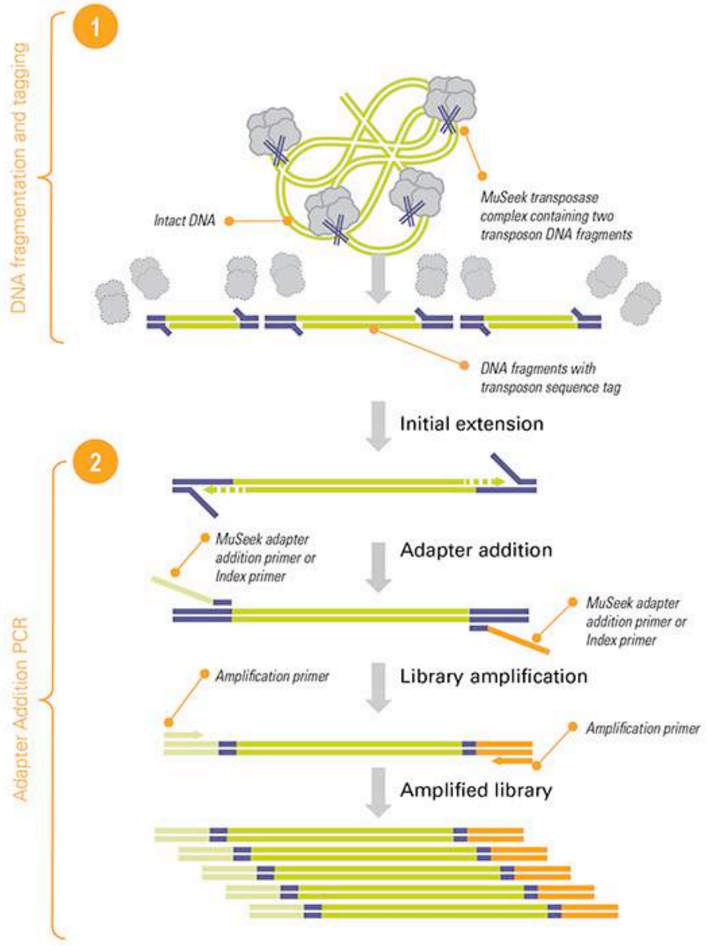


Input sample



Shearing and size-selection

Loosing molecules all the way


1
DNA fragmentation and tagging
Intact DNA
MuSeek transposase complex containing two transposon DNA fragments
DNA fragments with transposon sequence tag
Initial extension
2
Adapter Addition PCR
MuSeek adapter addition primer or Index primer
Adapter addition
MuSeek adapter addition primer or Index primer
Amplification primer
Library amplification
Amplification primer
Amplified library

Less material -> more amplification cycles

Sequencing artefacts: what are they?

PCR bias – important source of sequencing artefacts

PCR steps involved in any NGS but PacBio and Oxford Nanopore:

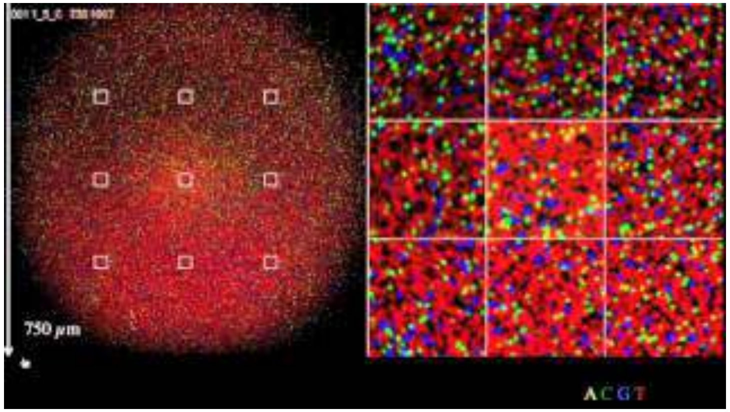
1. Library amplification
2. Amplification during templating (Illumina – on glass; Ion – emPCR)

Main PCR bias:

1. Size: shorter fragments amplify faster -> higher sequencing signal and coverage
2. Polymerase errors
   slippage in low complexity regions
   incorporation of erroneous bases & indels
3. GC-bias (fragments with high GC diminish to 1/10th from initial amount)

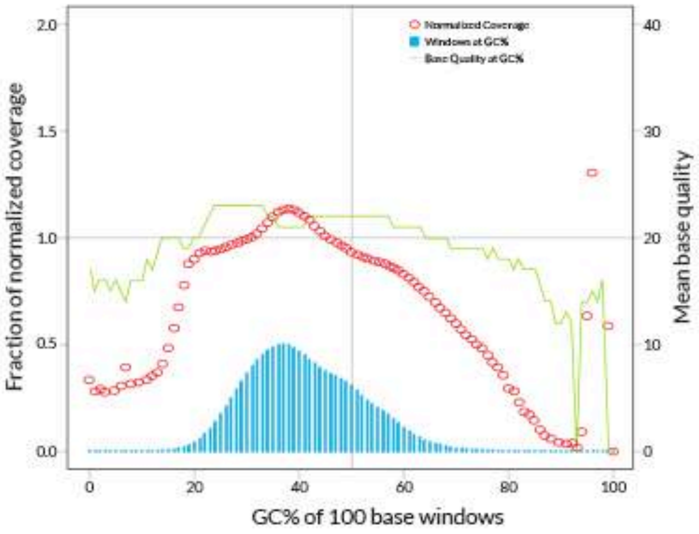
Sequencing artefacts: what are they?

PCR bias – important source of sequencing artefacts

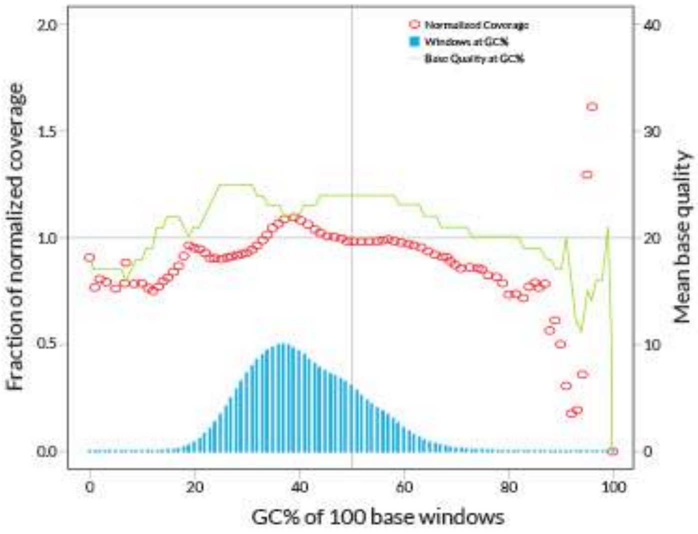


Clusters with shorter fragments grow faster -> quality signal from smaller clusters worsens

GC bias & genome coverage



Heavily amplified library

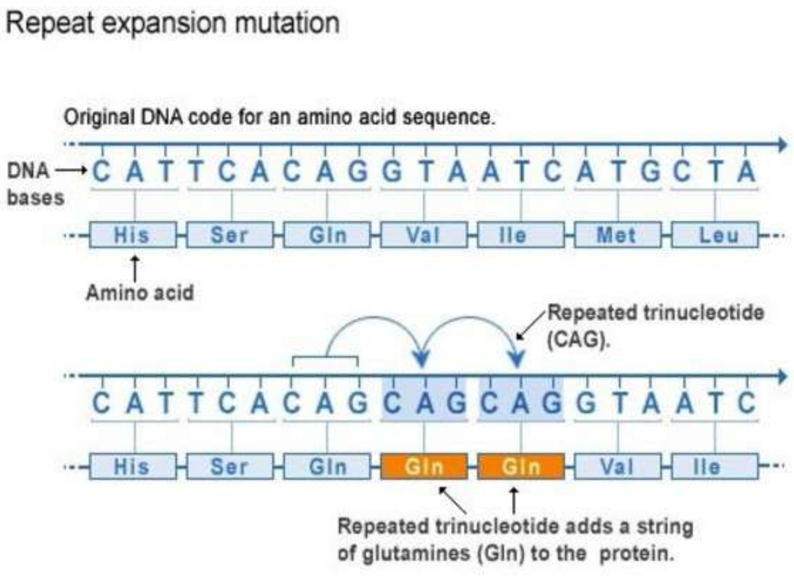


PCR-free library

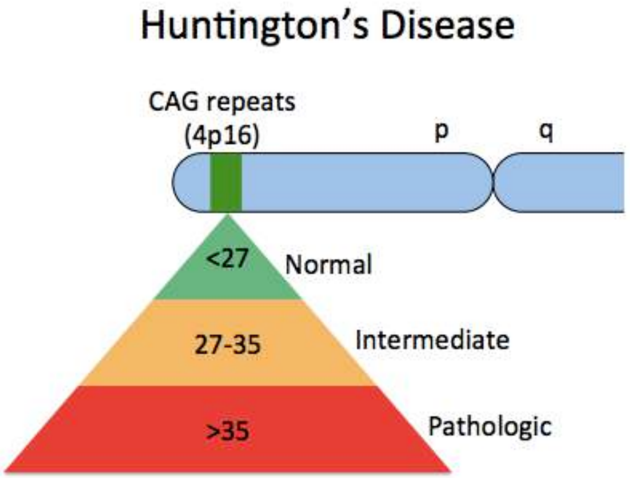
Sequencing artefacts: what are they?

PCR bias – important source of sequencing artefacts

Polymerase slippage – low complexity regions


Repeat expansion mutation
Original DNA code for an amino acid sequence.
DNA bases
C A T T C A C A G G T A A T C A T G C T A
His Ser Gln Val Ile Met Leu
Amino acid
Repeated trinucleotide (CAG).
C A T T C A C A G C A G C A G G T A A T C
His Ser Gln Gln Gln Val Ile
Repeated trinucleotide adds a string of glutamines (Gln) to the protein.

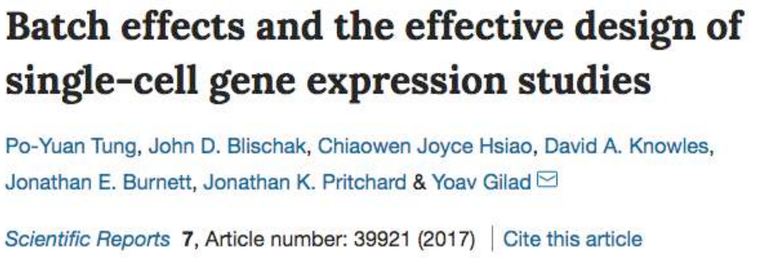
U.S. National Library of Medicine



Huntington's disease:

- Inherited disorder resulting in brain cell death
- Decline of motoric and cognitive functions
- Common onset: 30-50 years of age
- No cure
- Causative genetic variant: CAG-repeat expansion in *HTT* gene

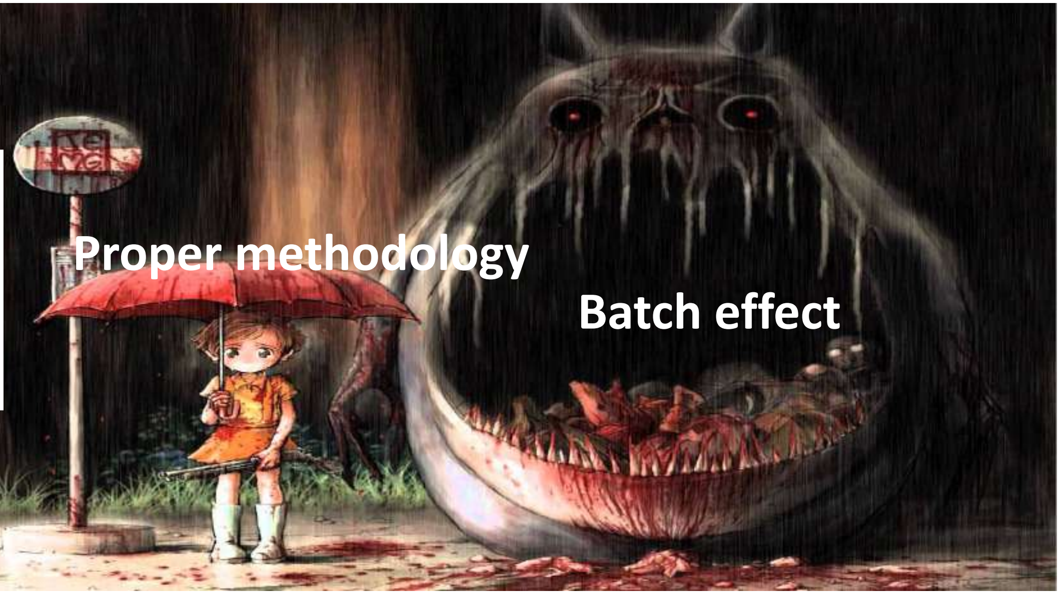
Batch Effects



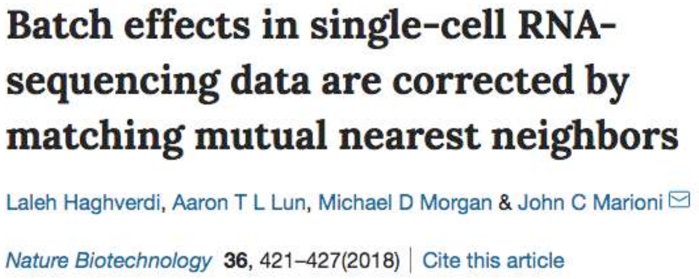
Batch effects and the effective design of single-cell gene expression studies

Po-Yuan Tung, John D. Blischak, Chiaowen Joyce Hsiao, David A. Knowles, Jonathan E. Burnett, Jonathan K. Pritchard & Yoav Gilad ✉

Scientific Reports **7**, Article number: 39921 (2017) | Cite this article



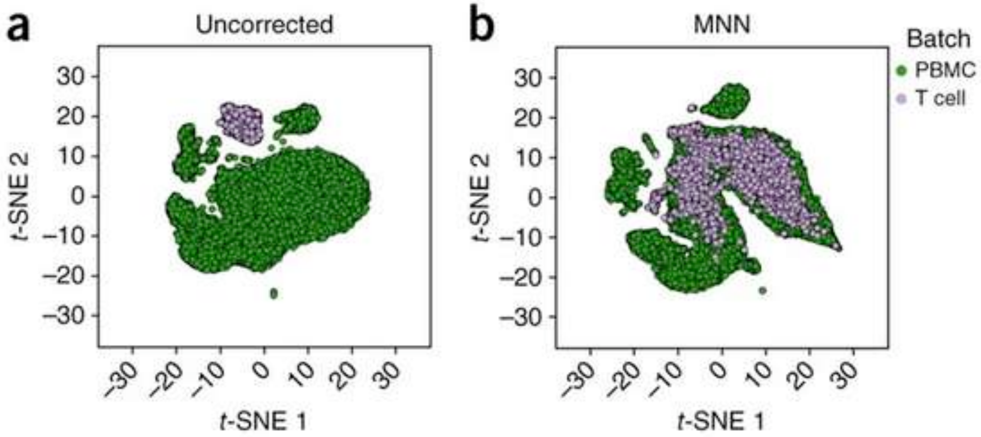
Single-cell & RNA sequencing



Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors

Laleh Haghverdi, Aaron T L Lun, Michael D Morgan & John C Marioni ✉

Nature Biotechnology **36**, 421–427(2018) | Cite this article



Sequencing bias: SUMMARY

- Keep in mind that they are there
- Coverage varies across the genome
- One technology does not suit all the applications
- Beware of batch effects

SAMPLE QUALITY REQUIREMENTS

Garbage in – garbage out:

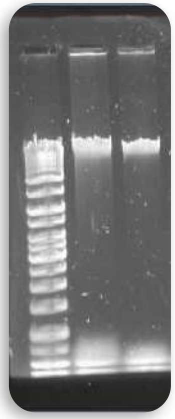
Sequencing success always depends on the **sample quality**.

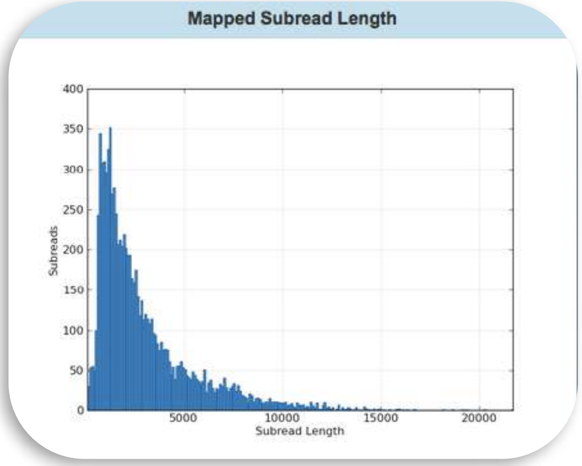
NGS-quality DNA and

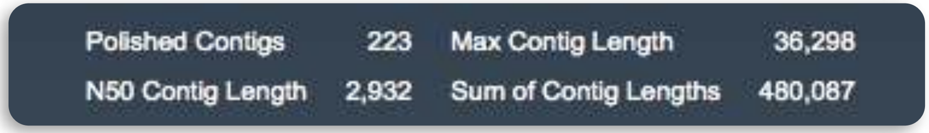
PCR-quality DNA

are two completely different things.

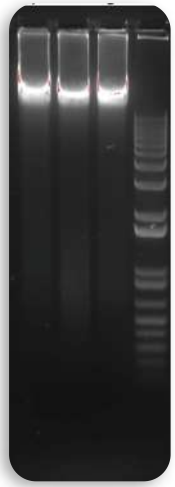
2013: a wake-up call

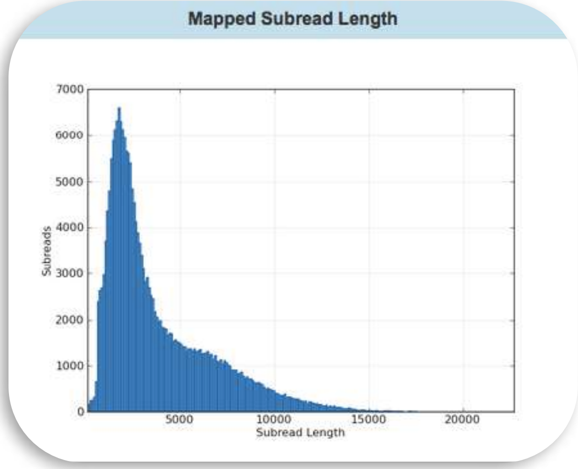


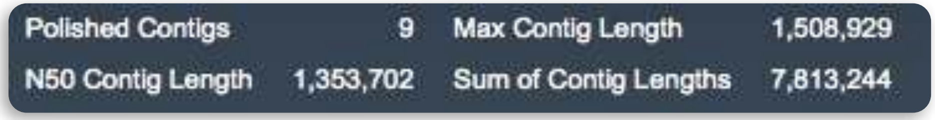




Polished Contigs	223	Max Contig Length	36,298
N50 Contig Length	2,932	Sum of Contig Lengths	480,087







Polished Contigs	9	Max Contig Length	1,508,929
N50 Contig Length	1,353,702	Sum of Contig Lengths	7,813,244

For Long Reads one needs to have *long and pure* DNA

SciLifeLab

DNA quality and inhibition of sequencing

Short-read technologies: PCR inhibition
Long-read technologies are PCR-free, but one sequences native DNA “as is”.

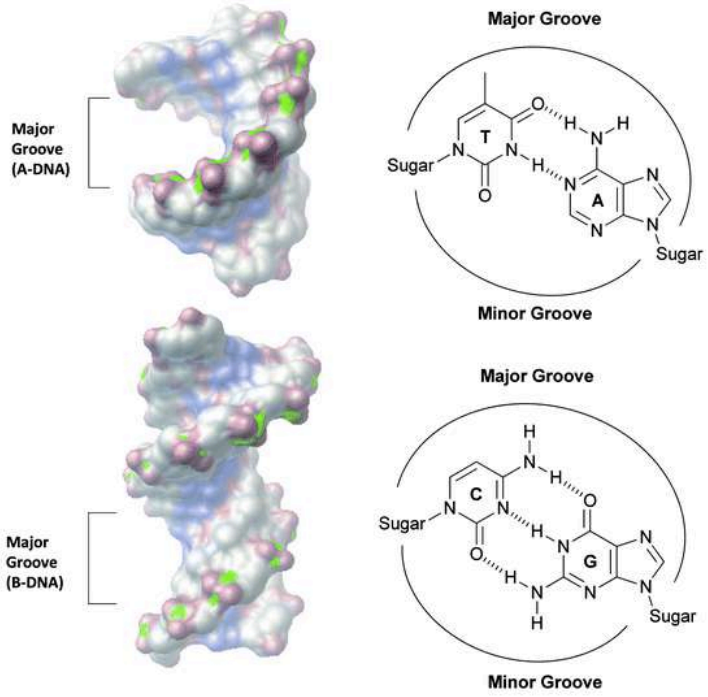
DNA-binders:

- Proteins
- Polyphenols
- Secondary metabolites (e.g. toxins)
- Pigments
- Polysaccharides

Polymerase inhibitors:

- Salts
- Phenol
- Alcohols

Physical inhibiting factors – debris



*Hamilton & Arya, Nat. Prod. Rep., 2012, **29**, 134-143*

What do absorption ratios tell us?

Pure DNA 260/280: 1.8 – 2.0

< 1.8:

Too little DNA compared to other components of the solution; presence of organic contaminants: proteins and phenol; glycogen - **absorb at 280 nm**.

> 2.0:

High share of RNA.

Pure DNA 260/230: 2.0 – 2.2

<2.0:

Salt contamination, humic acids, peptides, aromatic compounds, polyphenols, urea, guanidine, thiocyanates (latter three are common kit components) – **absorb at 230 nm**.

>2.2:

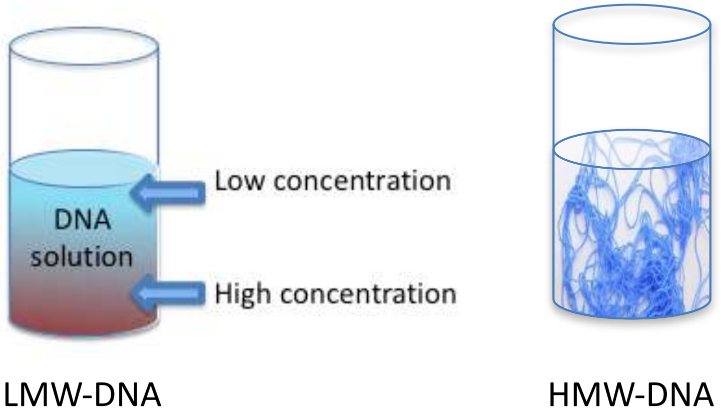
High share of RNA, very high share of phenol, **high turbidity,** dirty instrument, wrong blank.

Photometrically active contaminants:

phenol, polyphenols, EDTA, thiocyanate, protein,

RNA, nucleotides (fragments below 5 bp)

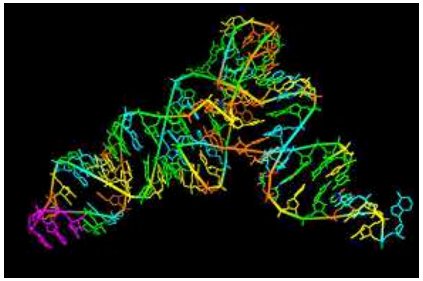
How to make a correct measurement



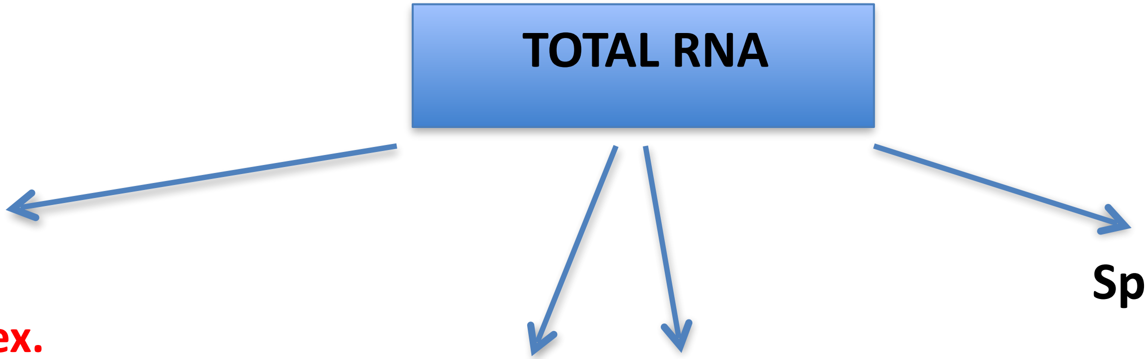
- Thaw DNA completely
- Mix gently (**never vortex!**)
- Put the sample on a thermoblock: 37°C, 15-30 min
- Mix gently
- **Dilute 1:100** (if HMW)
- Mix gently
- Make a measurement with an appropriate blank

- **NANODROP is Bad**. Point.
- Use Qubit, or PicoGreen.

What about RNA?



Transcriptome sequencing (RNA-seq)



TOTAL RNA

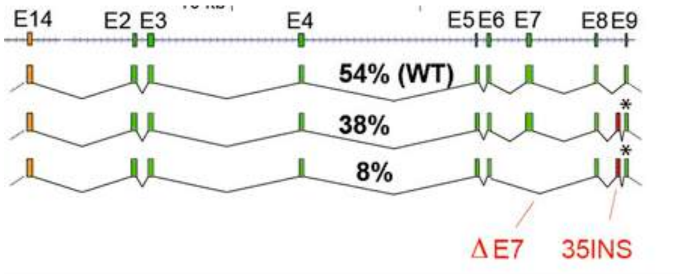
mRNA

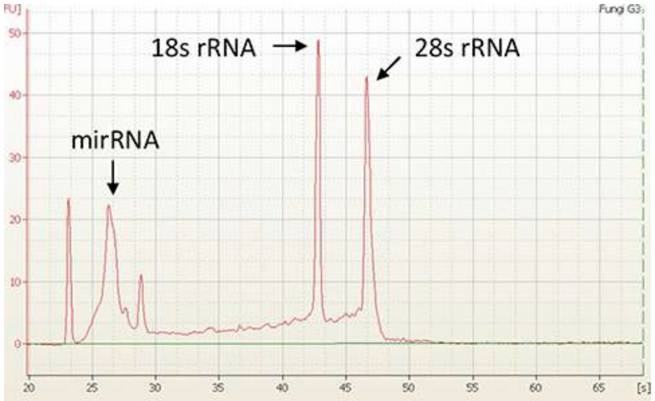
- **Dif.ex.**
- Annotation

Non-codingRNA **miRNA**

- Transcriptional regulation

Splice isoforms





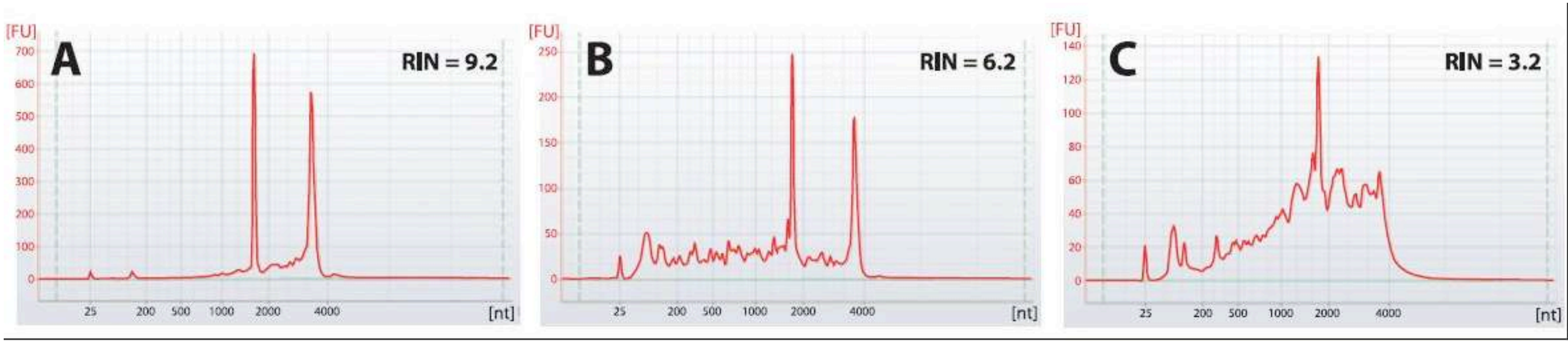
Sample prep: RNA

mRNA degrades FAST

Freeze sample or place it in RNA-later within 30 sec *(if possible)*

Chose a correct kit for your particular application!
Always treat samples with DNase

Differential expression, miRNA – **RIN value over 8.0**
Aim for 4 biological replicates

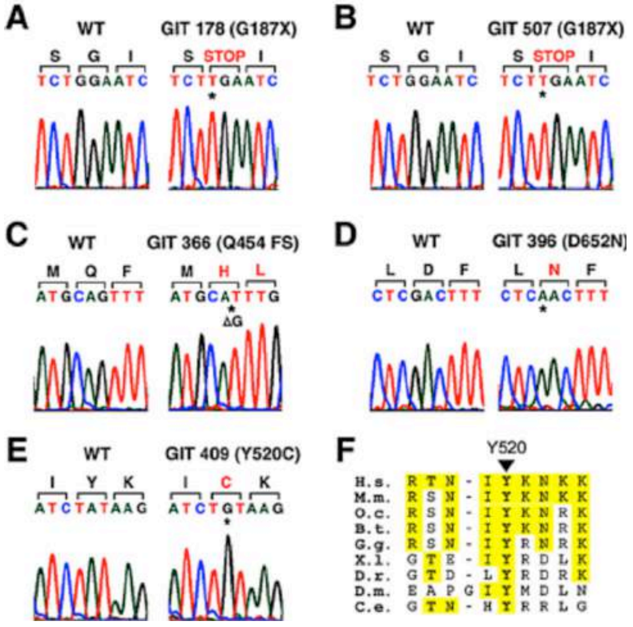


Sample prep: SUMMARY

- Sequencing success depends on the sample quality
- DNA quality is **essential** for PacBio and ONT sequencing
  … as well as PCR-free Illumina libraries & linked reads!
- Basic understanding of biochemistry is needed
- **NGS-grade sample ≠ PCR-grade sample**
- **Be cautious with data interpretation**

1 minute of phylosophy

Genome is not a linear string of bases!!



Mutations in coding regions only

↓

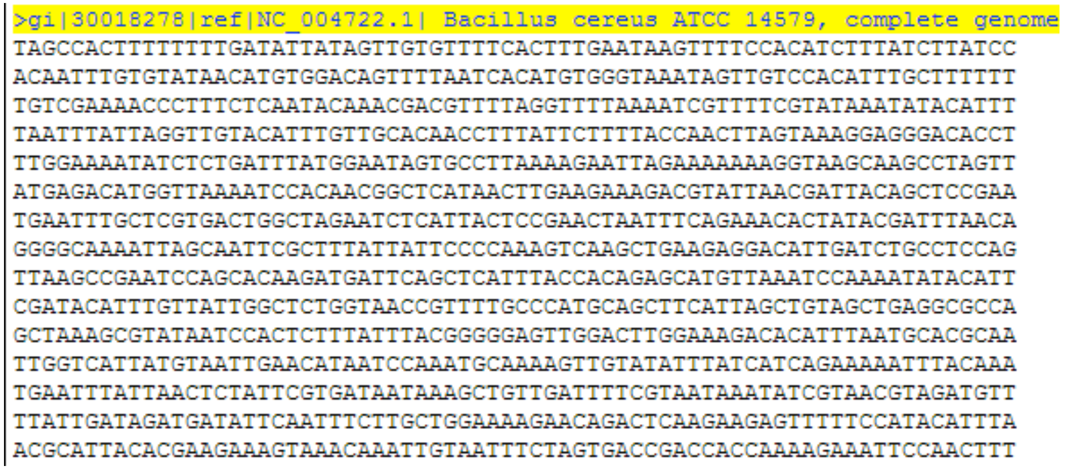
Transcriptional & post-transcriptional regulation

↓

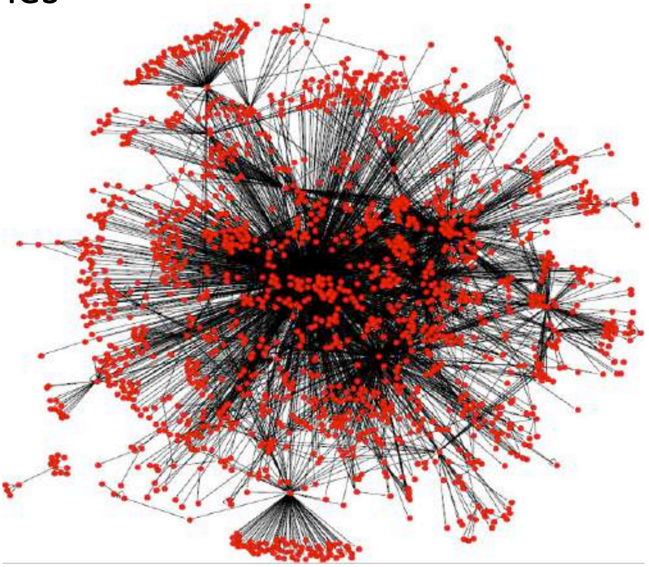
Epigenetics

↓

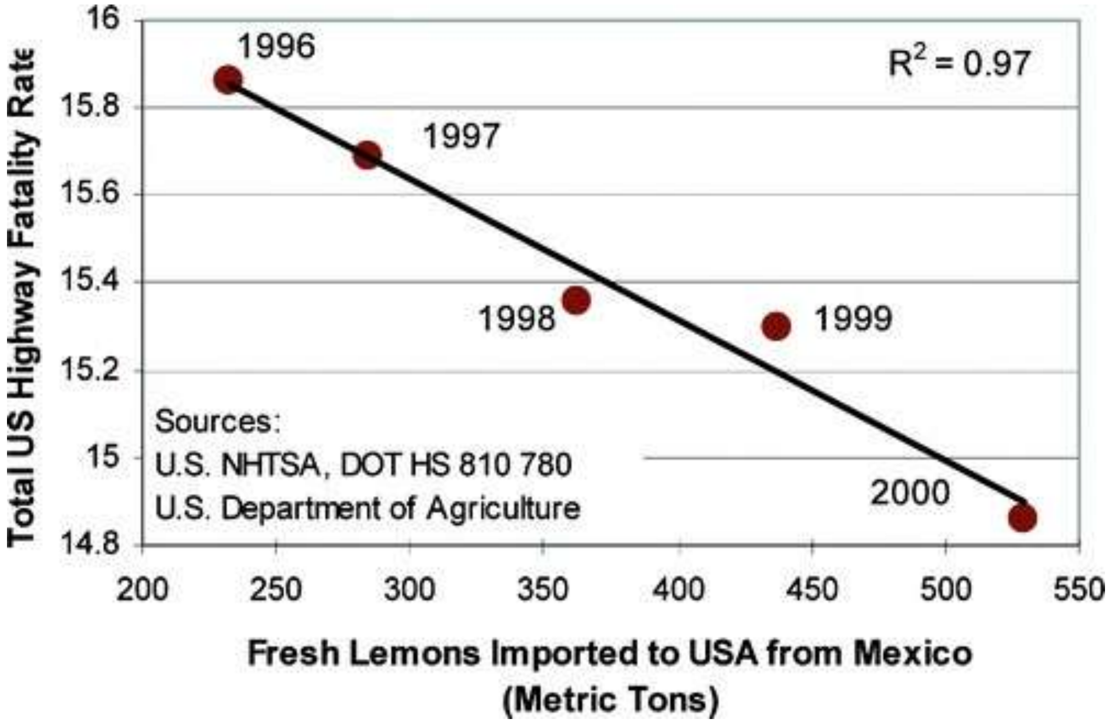
Proximity in chromosomes



```
>gi|30018278|ref|NC_004722.1| Bacillus cereus ATCC 14579, complete genome
TAGCCACTTTTTTTTGATATTATAGTTGTGTTTTCACTTTGAATAAGTTTTCCACATCTTTATCTTATCC
ACAATTTGTGTATAACATGTGGACAGTTTTAATCACATGTGGGTAAATAGTTGTCCACATTTGCTTTTTT
TGTCGAAAACCCTTTCTCAATACAAACGACGTTTTAGGTTTTAAAATCGTTTTCGTATAAATATACATTT
TAATTTATTAGGTTGTACATTTGTTGCACAACCTTTATTCTTTTACCAACTTAGTAAAGGAGGGACACCT
TTGGAAAATATCTCTGATTTATGGAATAGTGCCTTAAAAGAATTAGAAAAAAAGGTAAGCAAGCCTAGTT
ATGAGACATGGTTAAAATCCACAACGGCTCATAACTTGAAGAAAGACGTATTAACGATTACAGCTCCGAA
TGAATTTGCTCGTGACTGGCTAGAATCTCATTACTCCGAACTAATTTCAGAAACACTATACGATTTAACA
GGGGCAAAATTAGCAATTCGCTTTATTATTCCCCAAAGTCAAGCTGAAGAGGACATTGATCTGCCTCCAG
TTAAGCCGAATCCAGCACAAGATGATTCAGCTCATTTACCACAGAGCATGTTAAATCCAAAATATACATT
CGATACATTTGTTATTGGCTCTGGTAACCGTTTTGCCCATGCAGCTTCATTAGCTGTAGCTGAGGCGCCA
GCTAAAGCGTATAATCCACTCTTTATTTACGGGGGAGTTGGACTTGGAAAGACACATTTAATGCACGCAA
TTGGTCATTATGTAATTGAACATAATCCAAATGCAAAAGTTGTATATTTATCATCAGAAAAATTTACAAA
TGAATTTATTAACTCTATTCGTGATAATAAAGCTGTTGATTTTCGTAATAAATATCGTAACGTAGATGTT
TTATTGATAGATGATATTCAATTTCTTGCTGGAAAAGAACAGACTCAAGAAGAGTTTTTCCATACATTTA
ACGCATTACACGAAGAAAGTAAACAAATTGTAATTTCTAGTGACCGACCACCAAAAGAAATTCCAACTTT
```

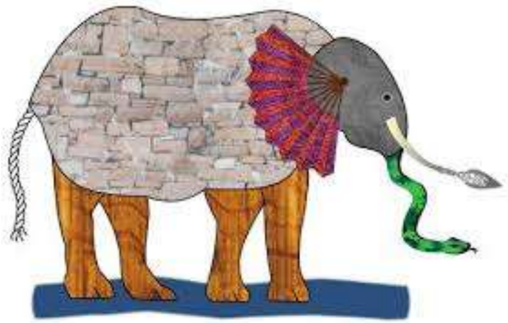


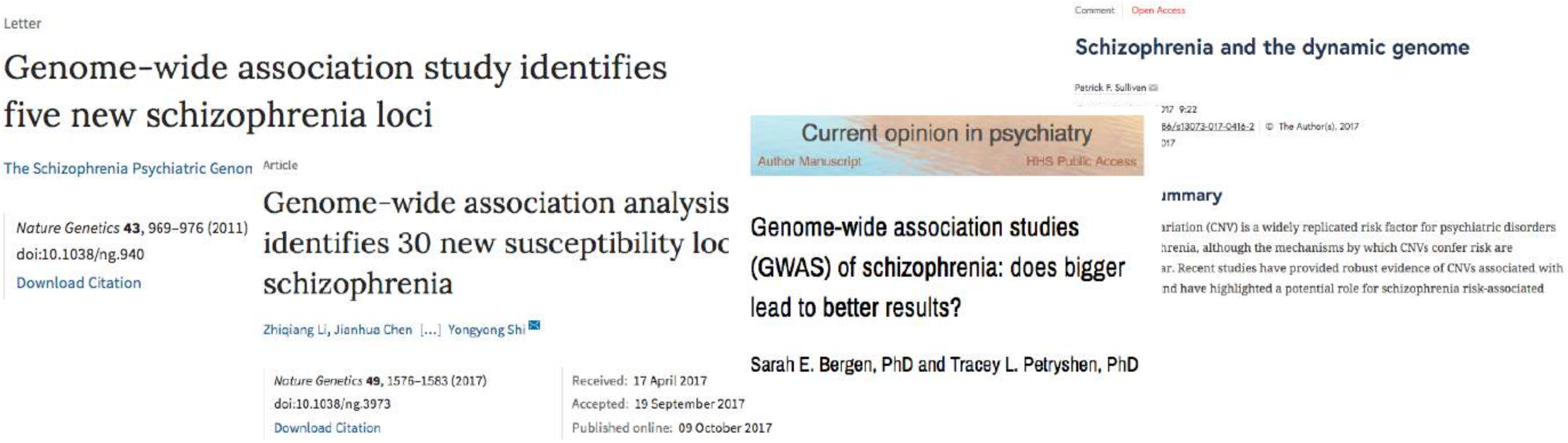
Never forget: Correlation vs Causation



Reduction in export of fresh lemons from Mexico causes significant reduction of highway traffic fatality rates in the US!

Blind men & an elephant





Letter

Genome-wide association study identifies five new schizophrenia loci

The Schizophrenia Psychiatric Genon

Nature Genetics **43**, 969–976 (2011)
doi:10.1038/ng.940
Download Citation

Article

Genome-wide association analysis identifies 30 new susceptibility loc schizophrenia

Zhiqiang Li, Jianhua Chen [...] Yongyong Shi

Nature Genetics **49**, 1576–1583 (2017)
doi:10.1038/ng.3973
Download Citation

Received: 17 April 2017
Accepted: 19 September 2017
Published online: 09 October 2017

Current opinion in psychiatry
Author Manuscript HHS Public Access

Genome-wide association studies (GWAS) of schizophrenia: does bigger lead to better results?

Sarah E. Bergen, PhD and Tracey L. Petryshen, PhD

Comment | Open Access

Schizophrenia and the dynamic genome

Patrick F. Sullivan

ɔ17 9:22
86/s13073-017-0416-2 | © The Author(s). 2017
ɔ17

ımmary

ariation (CNV) is a widely replicated risk factor for psychiatric disorders
hrenia, although the mechanisms by which CNVs confer risk are
ar. Recent studies have provided robust evidence of CNVs associated with
nd have highlighted a potential role for schizophrenia risk-associated

NGS and its challenges: SUMMARY

- Technologies develop VERY FAST.
- Beware of sequencing bias.
- Sequencing result depends on sample quality.
- Consult experts when it comes to experimental design and technology choice.

- ***Do nor forget the elephant…..***

THANK YOU!