High-throughput DNA sequencing

Robert Lyle

Department of Medical Genetics
Norwegian Sequencing Centre
Oslo University Hospital

Centre for Fertility and Health
Norwegian Institute of Public Health

Robert.Lyle@medisin.uio.no

Forskningsrådet

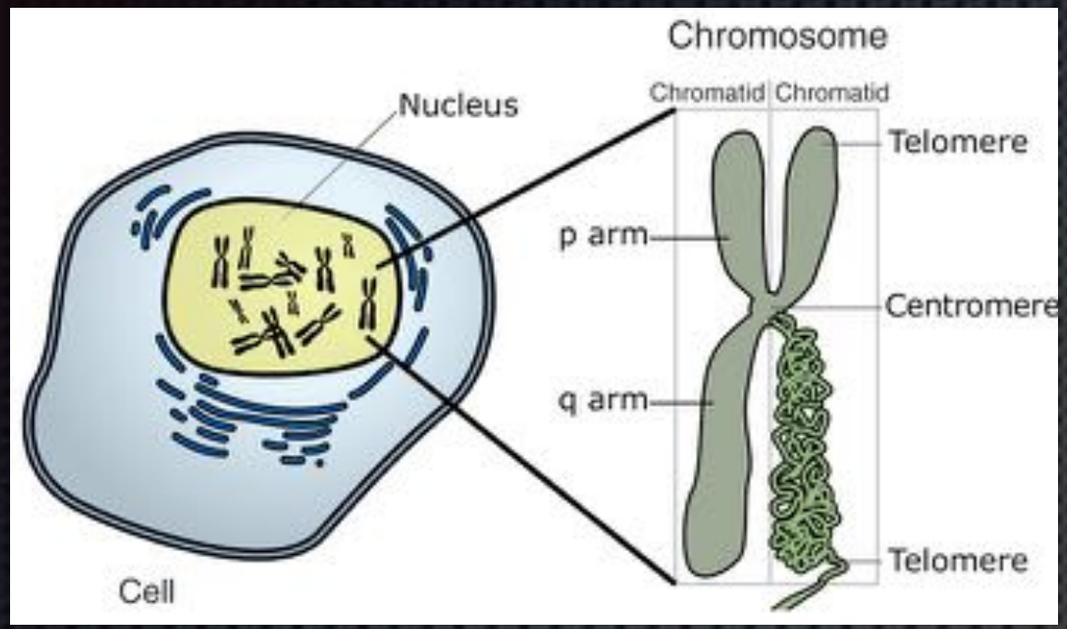
UNIVERSITY OF OSLO

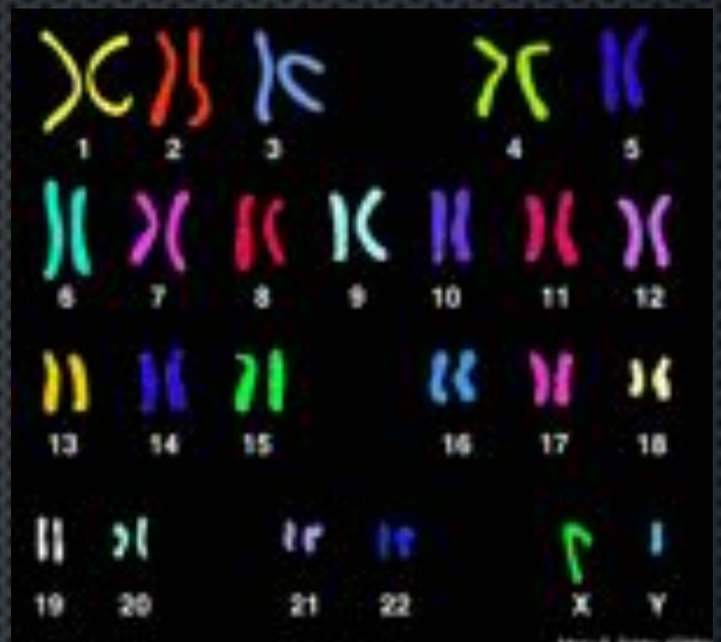
Oslo universitetssykehus

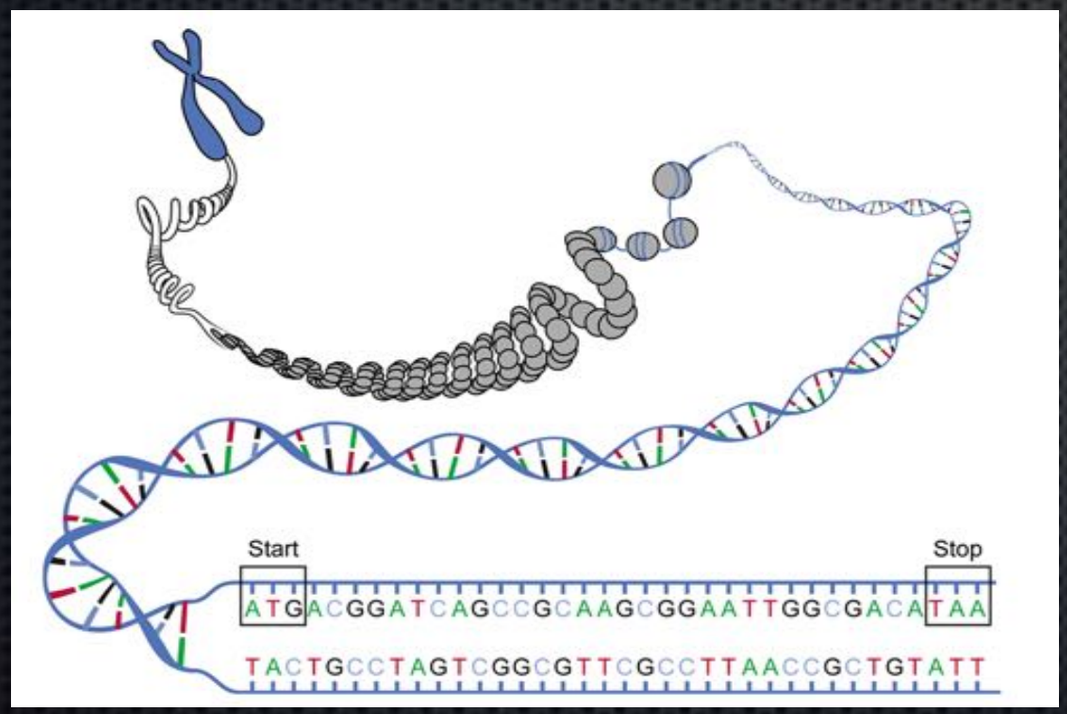
HELSE SØR-ØST

DNA sequencing

DNA







- 4 bases - A, G , C, T

G A T C
C T A G

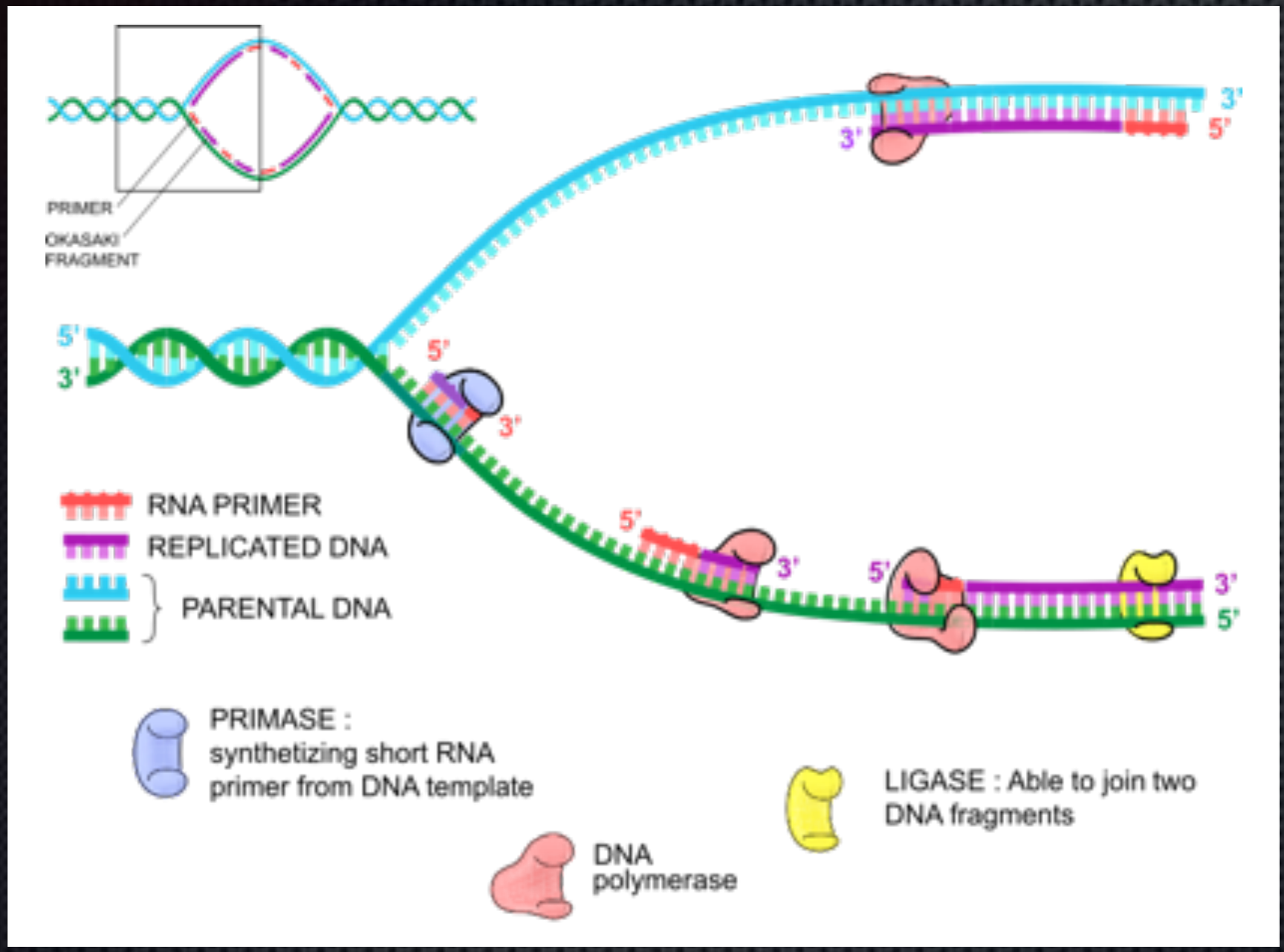
- Human ***genome*** ~3 billion bases
- How can we read the sequence of bases?

How many bases?

	bp	1	10^0
kilo	kb	1000	10^3
mega	Mb	1000000	10^6
giga	Gb	1000000000	10^9
tera	Tb	1000000000000	10^{12}

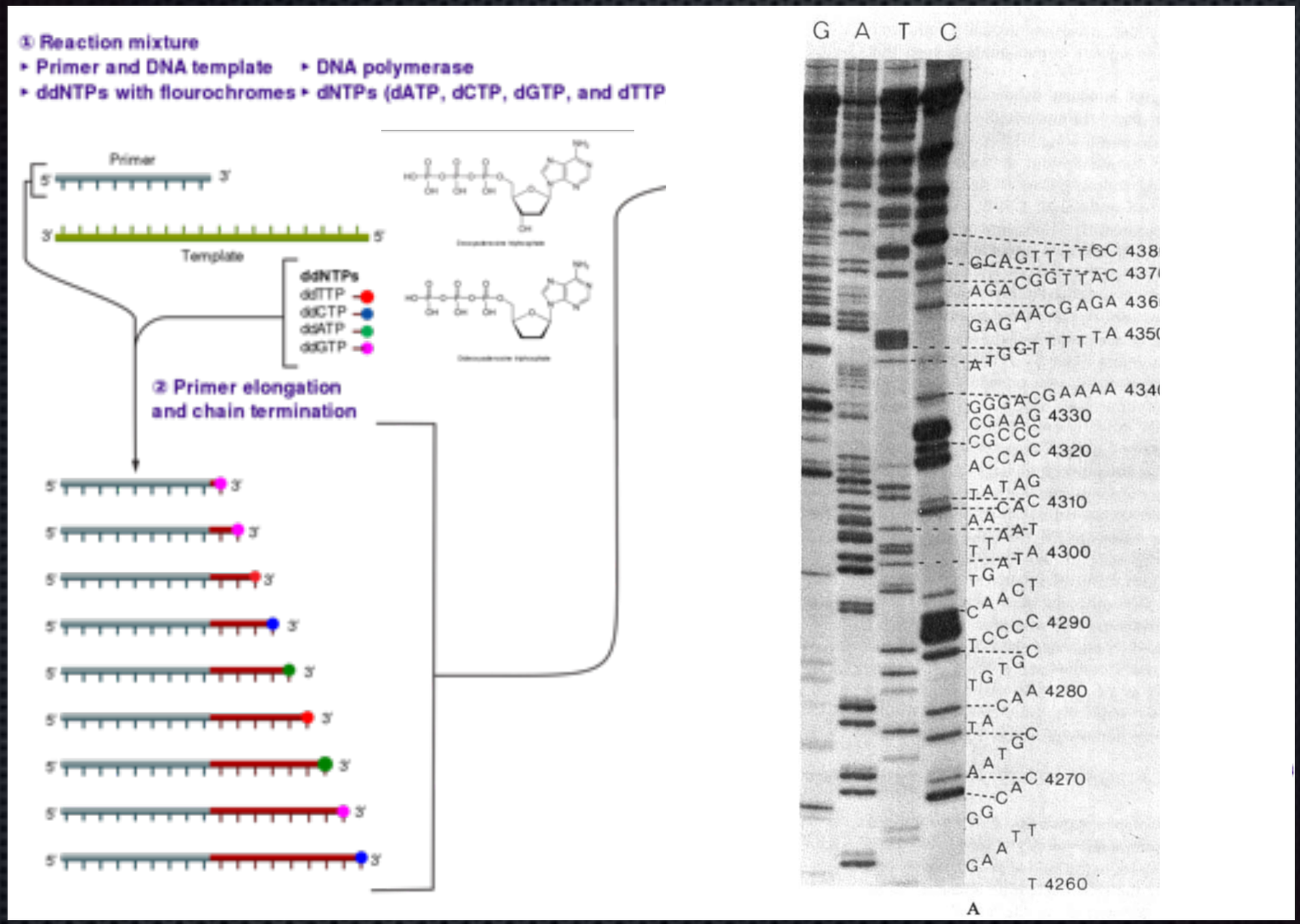
Human genome: 3 000 000 000 (3 Gb)

DNA replication



1. Unwind DNA
2. Prime with short (RNA) oligonucleotide
3. Polymerase extends base by base

Sanger sequencing

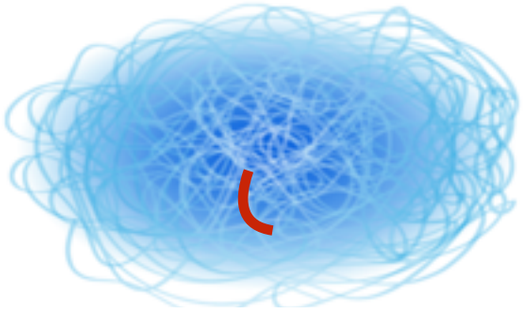


Fred Sanger (1977) - still used routinely today

High-throughput sequencing

Sanger and HTS

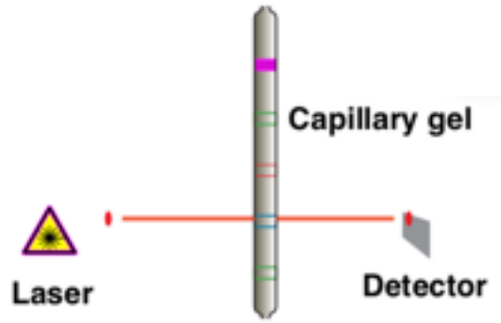
Sanger



Genomic DNA

PCR

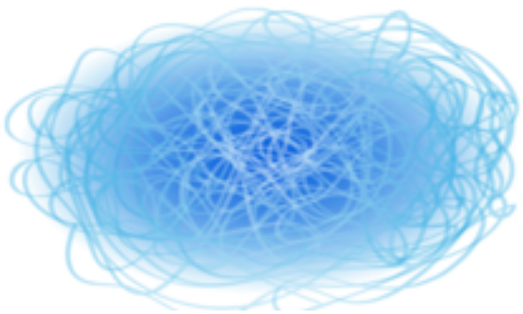
Locus of interest



16, 48, 96 reads

+ low error rate
- very low throughput

HTS

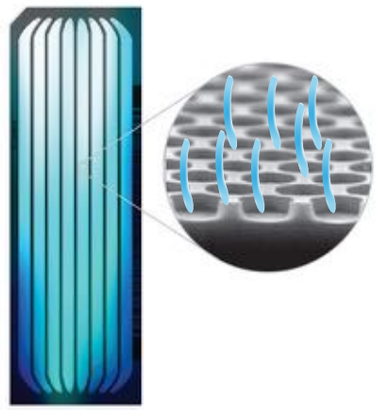


Genomic DNA

Fragment

Array

Array of DNA molecules



<20x10^9 reads

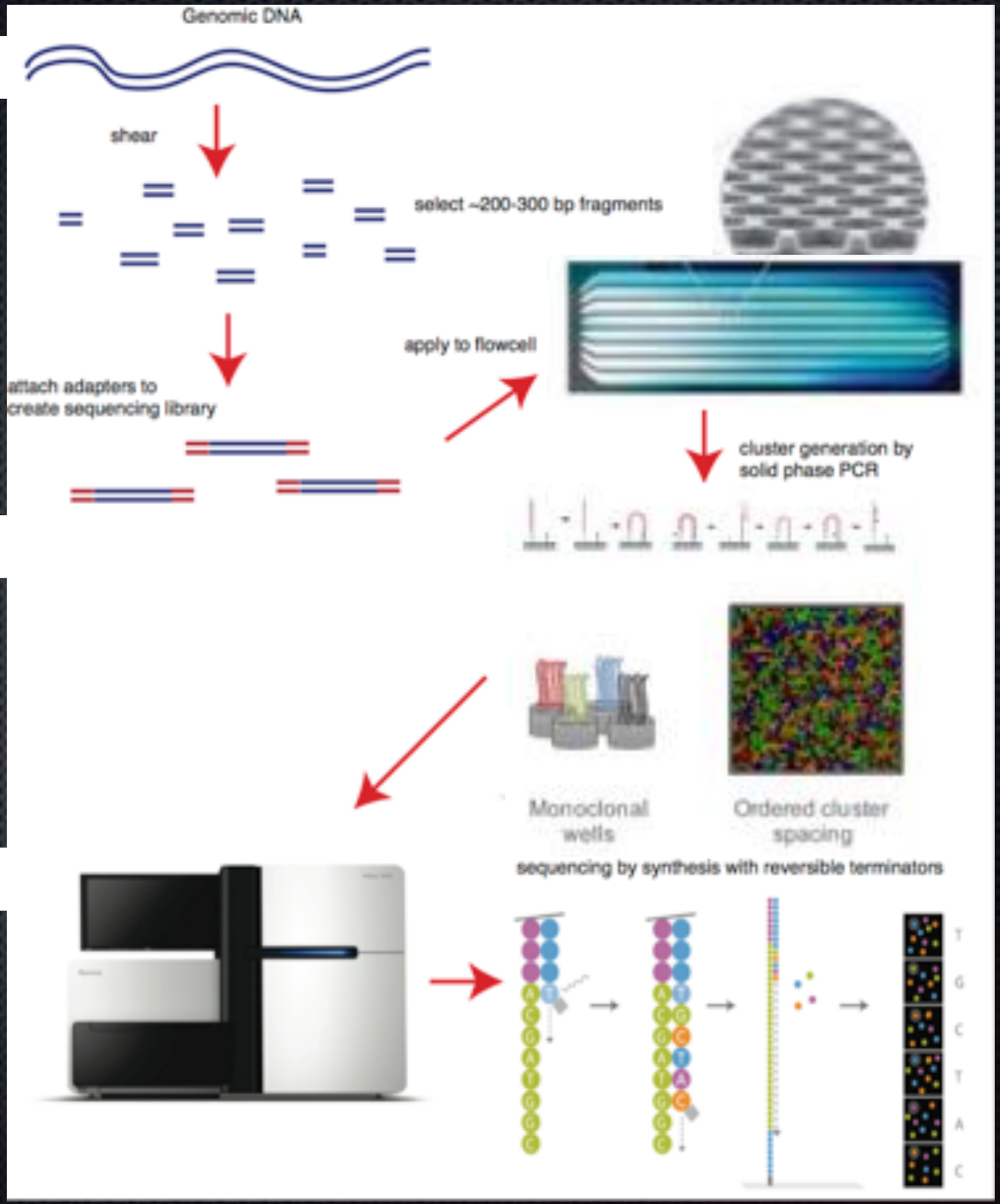
+ very high throughput
- high error rate

Illumina sequencing technology

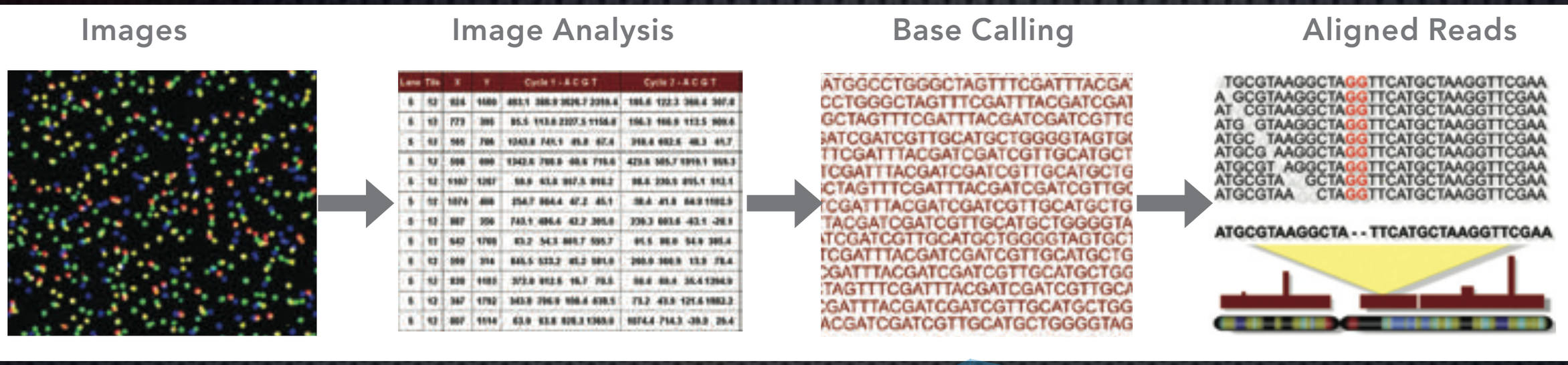
1. Library preparation

2. Cluster generation

3. Sequencing



Analysis pipeline



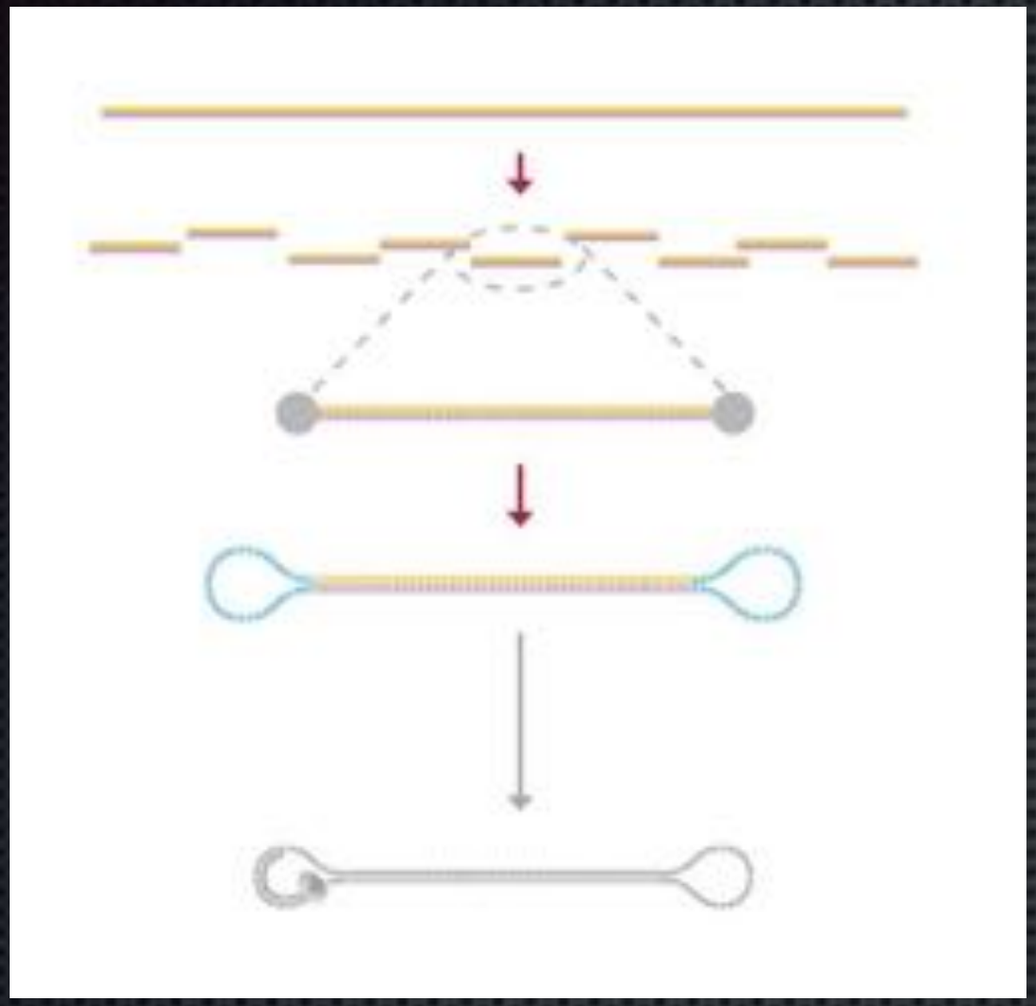
- Short fragments: 50…300 bp
- ***<20 billion DNA sequence reads***
- Single-end reads
- Paired-end reads

Pacific BioSciences

- Single Molecule Real Time
- Long Readlength
- Speed

http://www.pacificbiosciences.com

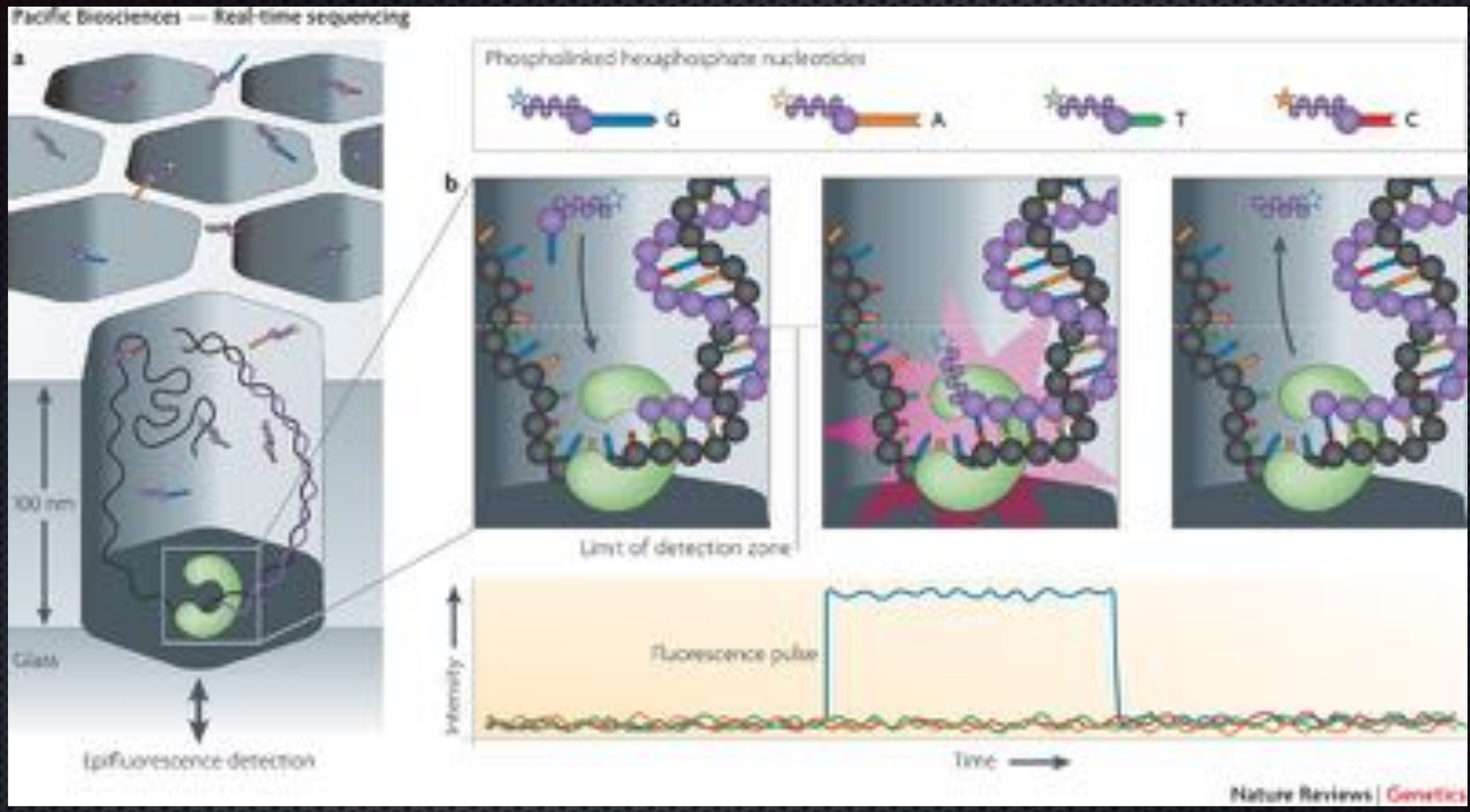
PacBio sequencing I



SMRT bell

SMRT cell

PacBio sequencing II

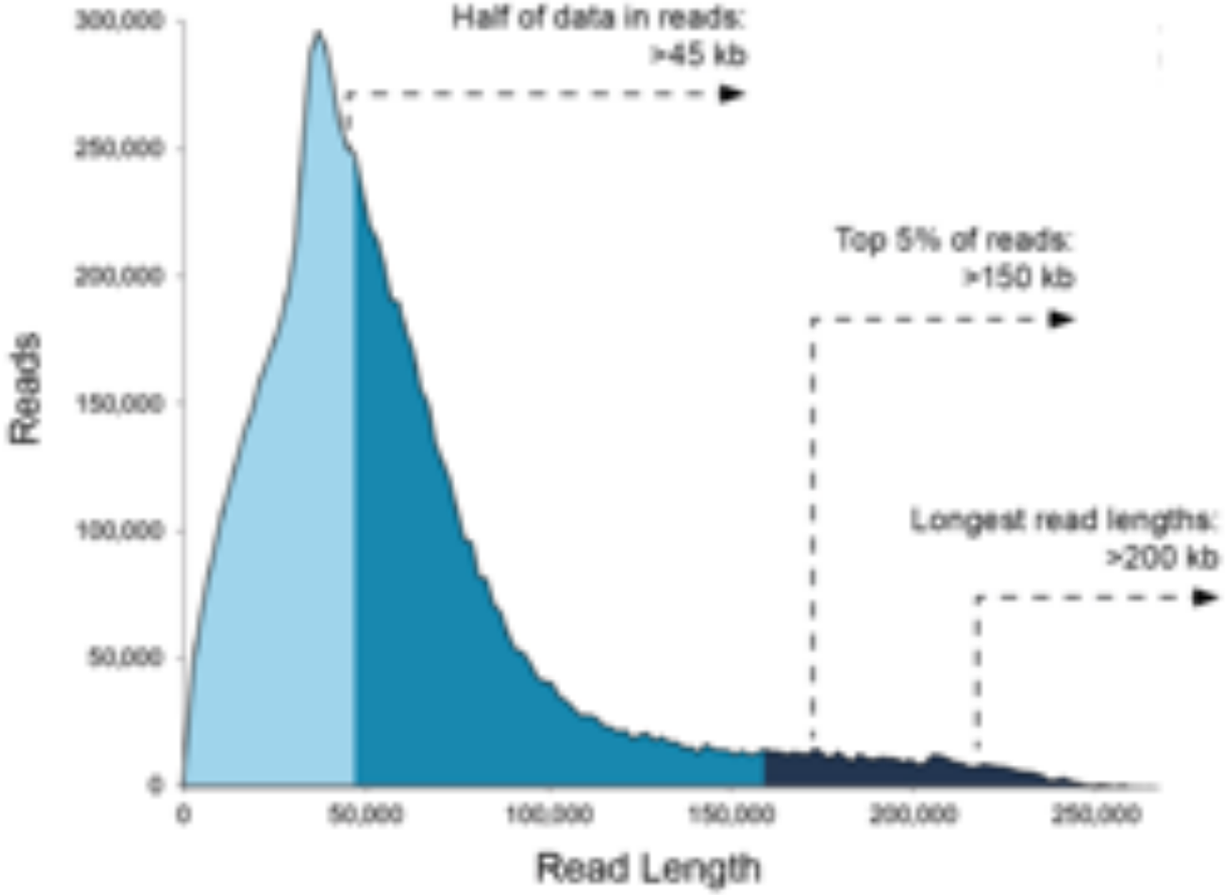


- Detection of base incorporation in single molecule

PacBio data

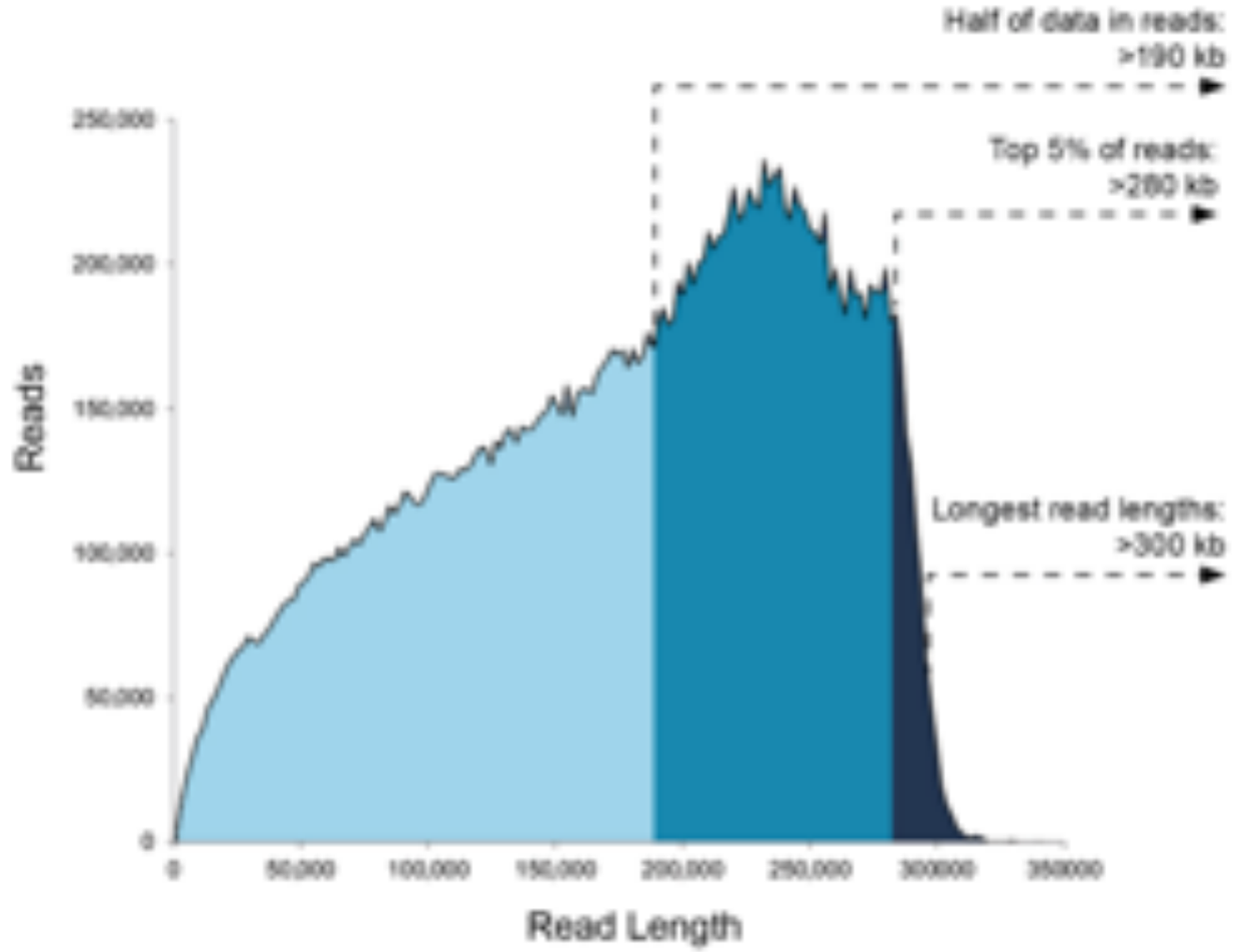
Long Read Lengths
(Libraries >20 kb)

Half of data in reads: >45 kb
Data per SMRT Cell: Up to 20 Gb



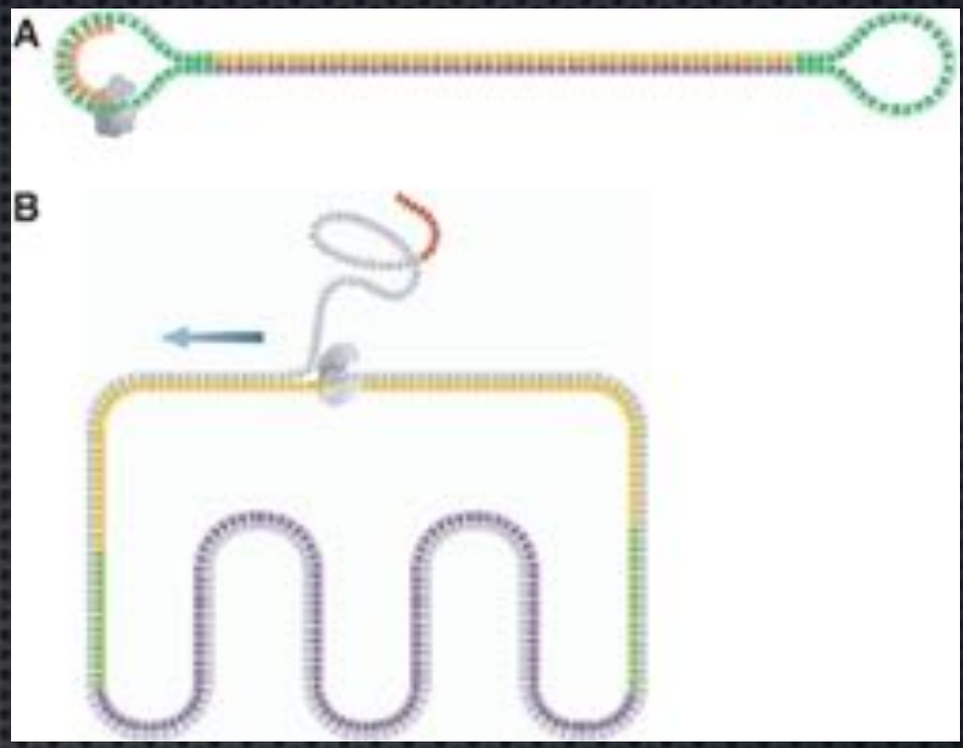
Long Read Lengths
(Libraries <20 kb)

Half of data in reads: >190 kb
Data per SMRT Cell: Up to 50 Gb

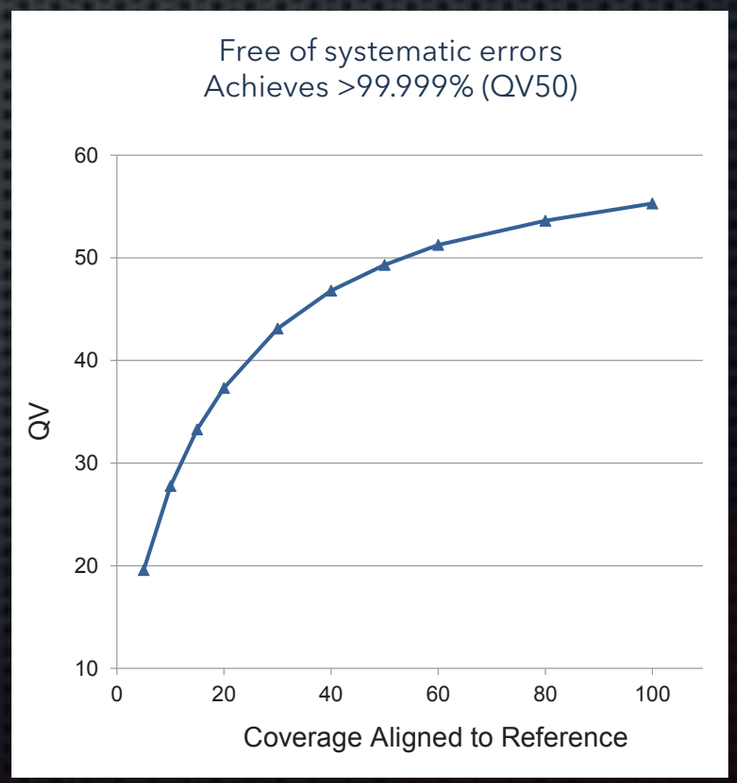


PacBio read accuracy

- Raw read accuracy is low, 87%
- Read same molecule multiple times
- Circular consensus sequencing



- Increase read depth
- Errors are randomly distributed
- Low GC bias



Sequencing instruments





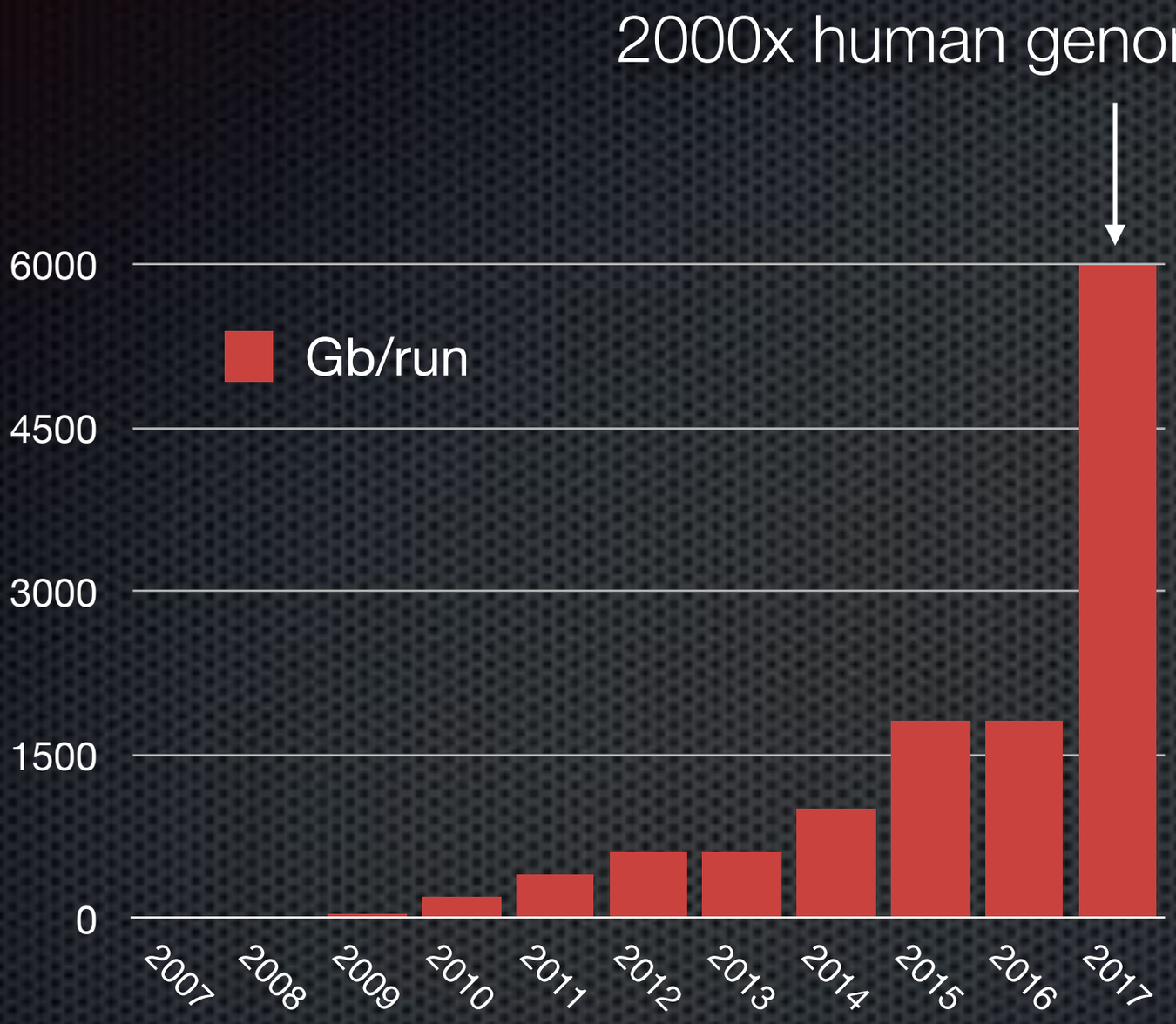

Platform	NovaSeq	HiSeq X	HiSeq 3000/4000	NextSeq	MiSeq	PacBio Sequel
@NorSeq	2	4	1/4	5	4	1/1
Run time	1-2 days	1-3 days	1-5 days	29 hours	29 hours	0.5-30 hours
Read accuracy	99%	99%	99%	99%	99%	87%
Read number	20 10e9	6 10e9	4 10e9	400000000	20000000	350000
Read length	2x150 bp	2x150 bp	2x125 bp	2x150 bp	2x300 bp	~20 kb
Output	6000 Gb	1800 Gb	1-1500 Gb	129 Gb	12 Gb	160 Gb

Types of sequencing projects

Project	Description
Resequencing	Align and compare to a reference sequence
de novo	Assemble new genome
metagenomics	Sequence DNA pool of multiple species
mRNA	Sequence cDNA for gene expression
miRNA	Sequence small RNAs for expression
ChIP	Study chromatin structure
DNA meth	Study DNA methylation

So what?

HTS throughput: data per run



Illumina NovaSeq6000

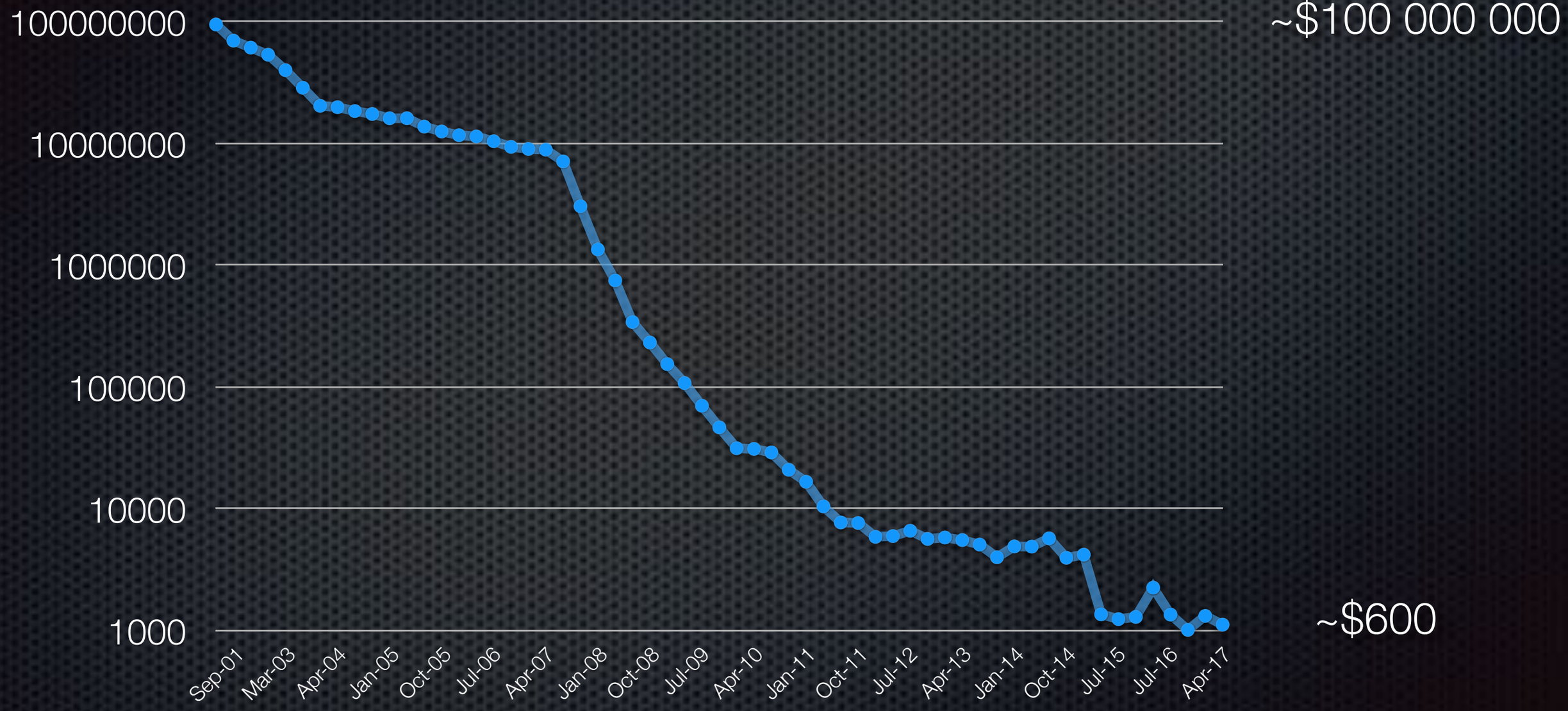
- 6 Tb (6000 Gb)
- 2000x human genome

Really, so what?

Parameter	ABI 3100	ABI 3730	NovaSeq6000
Read length	~700	~700	150 (x2)
Reads per run	16	96	20000000000
Run time	2 hours	30 minutes	2 days
Time for 1x human genome (3 Gb)	120 years	15 years	~90 seconds

Sequencing costs

First human genome ~$3 000 000 000



www.genome.gov/sequencingcosts/

Earth BioGenome Project

- “Sequencing life for the future of life”
- “…a moonshot for biology that aims to sequence, catalog, and characterize the genomes of all of Earth’s eukaryotic biodiversity over a period of 10 years.”

- Sequencing and annotate ~1.5 million known eukaryotic species
- Cost estimate: $4.7 billion USD
- Less than the cost of creating the first draft human genome

https://www.earthbiogenome.org

NorSeq

National Consortium for Sequencing and Personalized Medicine

NorSeq-Tromsø

NorSeq-Trondheim

NTNU
Norwegian University of Science and Technology

nor seq

NorSeq-Bergen

UNIVERSITY OF BERGEN
Haukeland University Hospital

NorSeq-Oslo

NorSeq-Cancer

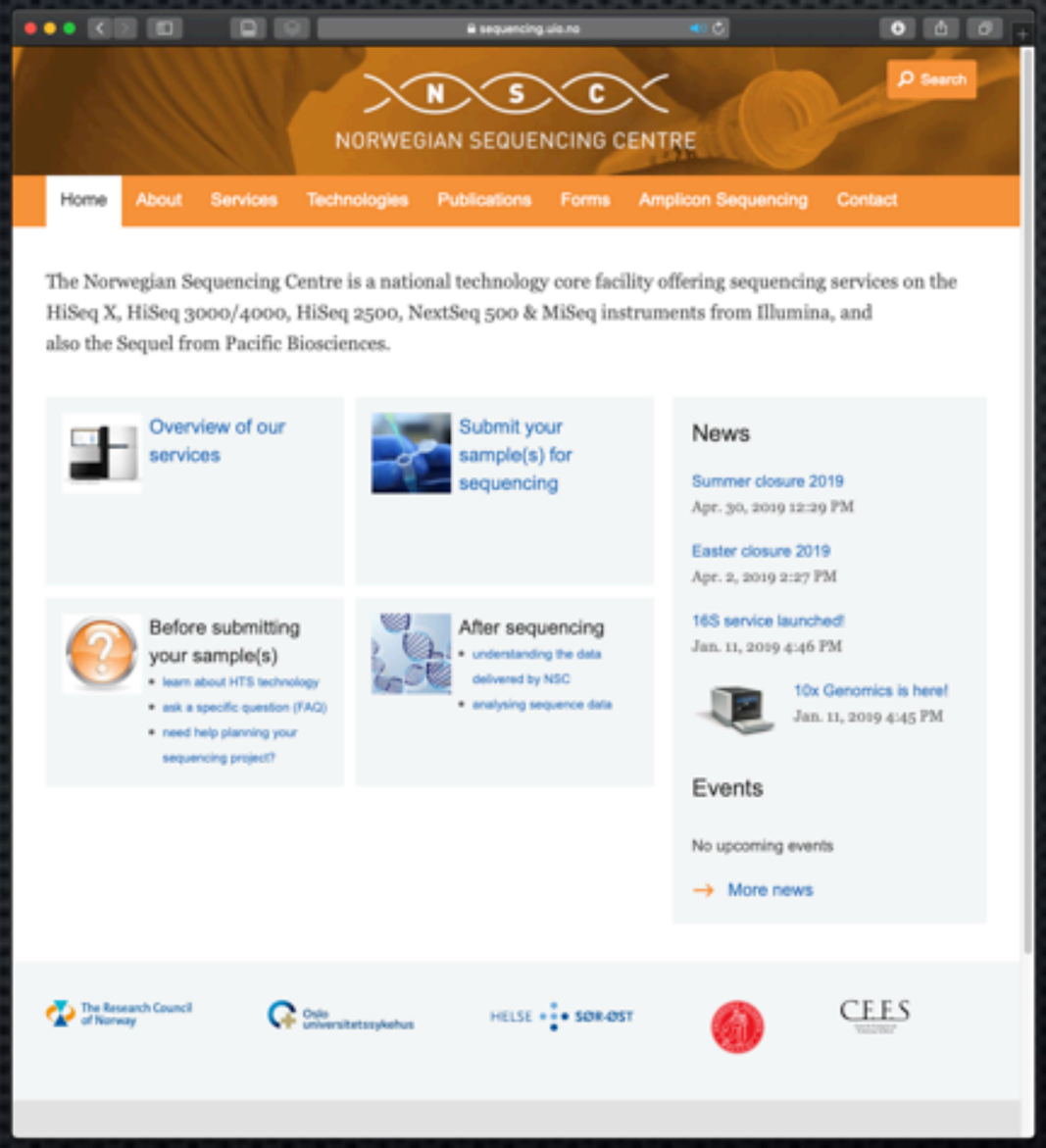
NSC
NORWEGIAN SEQUENCING CENTRE
Oslo University Hospital

post@sequencing.uio.no

https://www.norseq.org/kontakt/

Norwegian Sequencing Centre (NSC)

Your local, friendly HTS core facility

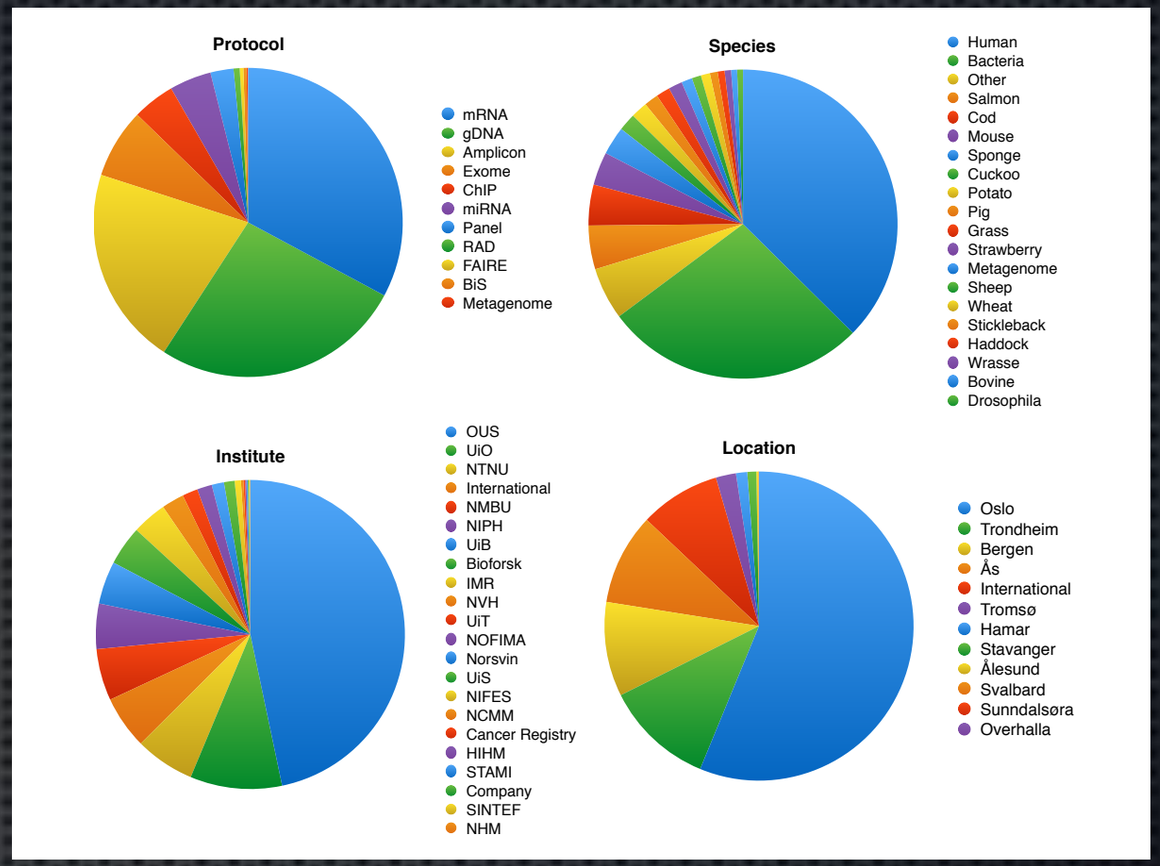


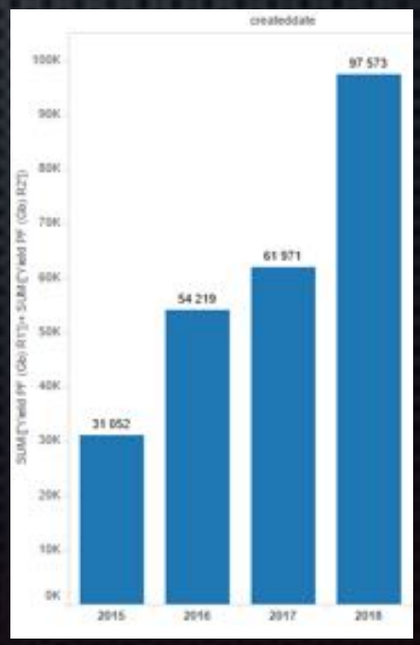
post@sequencing.uio.no

User statistics

2020

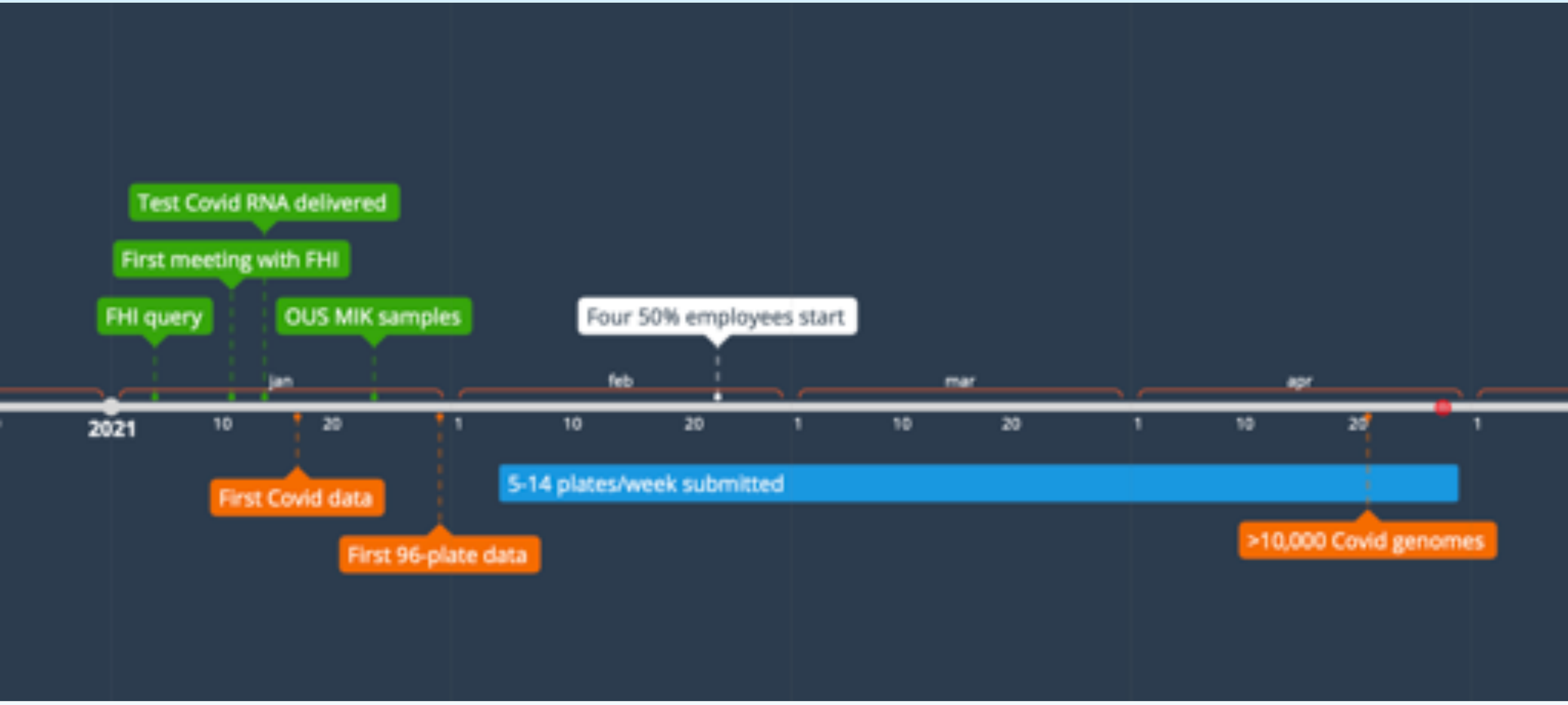
- NorSeq
  - 319 users
  - 464 projects
  - 164 publications





Gb/year

Covid sequencing at NSC

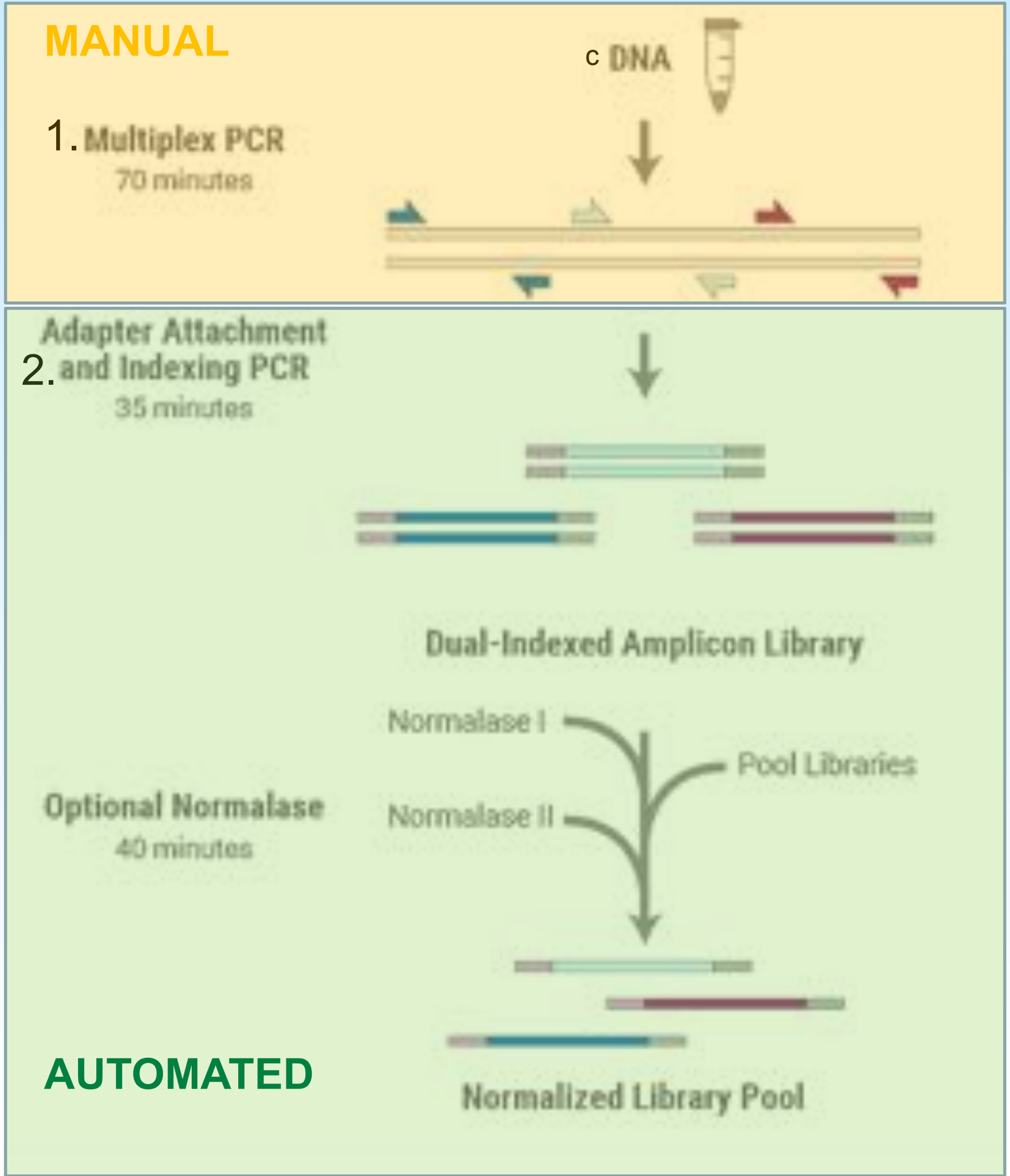


Slide from Gregor Gilfillan

Swift-Covid SNAP*-seq Overview

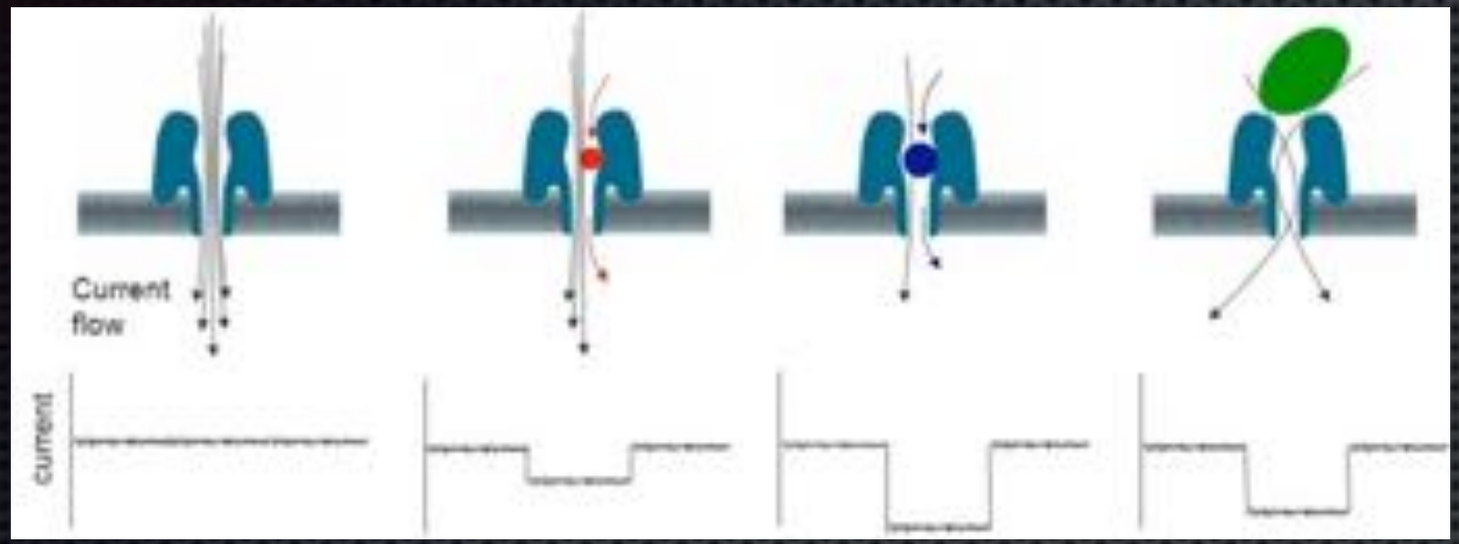
* Swift Normalase Amplicon Panel

Sequencing:
Want ca. 1 M reads per sample =
Up to 384 samples per 1/2 NovaSeq SP flowcell (150 bp PE).



New(er) technologies

Oxford Nanopore



- Read lengths 100s kb
- longest read 3 Mb
- G A T C mC hmC…
- Directly sequence RNA

http://www.nanoporetech.com/

Oxford Nanopore sequencers

MinION
Small-scale

PromethION
Large genomes

SmidgION
?

LETTER

doi:10.1038/nature16996

Real-time, portable genome sequencing for Ebola surveillance

Joshua Quick[1]*, Nicholas J. Loman[1]*, Sophie Duraffour[2,3]*, Jared T. Simpson[4,5]*, Ettore Severi[6]*, Lauren Cowley[7]*, Joseph Akoi Bore[2], Raymond Koundouno[2], Gytis Dudas[8], Amy Mikhail[7], Nobila Ouédraogo[9], Babak Afrough[2,10], Amadou Bah[2,11], Jonathan H. J. Baum[2,3], Beate Becker-Ziaja[2,3], Jan Peter Boettcher[2,12], Mar Cabeza-Cabrerizo[2,3], Álvaro Camino-Sánchez[2], Lisa L. Carter[2,13], Juliane Doerrbecker[2,3], Theresa Enkirch[2,14], Isabel García-Dorival[2,15], Nicole Hetzelt[2,12], Julia Hinzmann[2,12], Tobias Holm[2,3], Liana Eleni Kafetzopoulou[2,16], Michel Koropogui[2,17], Abigael Kosgey[2,18], Eeva Kuisma[2,10], Christopher H. Logue[2,10], Antonio Mazzarelli[2,19], Sarah Meisel[2,3], Marc Mertens[2,20], Janine Michel[2,12], Didier Ngabo[2,10], Katja Nitzsche[2,3], Elisa Pallasch[2,3], Livia Victoria Patrono[2,3], Jasmine Portmann[2,21], Johanna Gabriella Repits[2,22], Natasha Y. Rickett[2,15,23], Andreas Sachse[2,12], Katrin Singethan[2,24], Inês Vitoriano[2,10], Rahel L. Yemanaberhan[2,3], Elsa G. Zekeng[2,15,23], Trina Racine[25], Alexander Bello[25], Amadou Alpha Sall[26], Ousmane Faye[26], Oumar Faye[26], N'Faly Magassouba[27], Cecelia V. Williams[28,29], Victoria Amburgey[28,29], Linda Winona[28,29], Emily Davis[29,30], Jon Gerlach[29,30], Frank Washington[29,30], Vanessa Monteil[31], Marine Jourdain[31], Marion Bererd[31], Alimou Camara[31], Hermann Somlare[31], Abdoulaye Camara[31], Marianne Gerard[31], Guillaume Bado[31], Bernard Baillet[31], Déborah Delaune[32,33], Koumpingnin Yacouba Nebie[34], Abdoulaye Diarra[34], Yacouba Savane[34], Raymond Bernard Pallawo[34], Giovanna Jaramillo Gutierrez[35], Natacha Milhano[6,36], Isabelle Roger[34], Christopher J. Williams[6,37], Facinet Yattara[17], Kuiama Lewandowski[10], James Taylor[38], Phillip Rachwal[38], Daniel J. Turner[39], Georgios Pollakis[15,23], Julian A. Hiscox[15,23], David A. Matthews[40], Matthew K. O'Shea[41], Andrew McD. Johnston[41], Duncan Wilson[41], Emma Hutley[42], Erasmus Smit[43], Antonino Di Caro[2,19], Roman Wölfel[2,44], Kilian Stoecker[2,44], Erna Fleischmann[2,44], Martin Gabriel[2,3], Simon A. Weller[38], Lamine Koivogui[45], Boubacar Diallo[34], Sakoba Keïta[17], Andrew Rambaut[8,46,47], Pierre Formenty[34], Stephan Günther[2,3] & Miles W. Carroll[2,10,48,49]

**The Ebola virus disease epidemic in West Africa is the largest on record, responsible for over 28,599 cases and more than 11,299 deaths[1]. Genome sequencing in viral outbreaks is desirable to characterize the infectious agent and determine its evolutionary rate. Genome sequencing also allows the identification of signatures of host adaptation, identification and monitoring of diagnostic targets, and characterization of responses to vaccines and treatments. The Ebola virus (EBOV) genome substitution rate in the Makona strain has been estimated at between 0.87×10^{-3} and 1.42×10^{-3} mutations pe[illegible] in each geno[illegible] to identify di[illegible] Genome sequ[illegible] evolution and[illegible] Sequence dat[illegible] the results ar[illegible] Genomic su[illegible] owing to a lack of local sequencing capacity coupled with practical difficulties transporting samples to remote sequencing facilities[9]. To address this problem, here we devise a genomic surveillance system that utilizes a novel nanopore DNA sequencing instrument. In April 2015 this system was transported in standard airline luggage to Guinea and used for real-time genomic surveillance of the ongoing epidemic. We present sequence data and analysis of 142 EBOV samples collected during the period March to October 2015. We were able to generate results less than 24 h after receiving [illegible]

Figure 1 | Deployment of the portable genome surveillance system in Guinea. a, We were able to pack all instruments, reagents and disposable consumables within aircraft baggage. **b**, We initially established the genomic surveillance laboratory in Donka Hospital, Conakry, Guinea. **c**, Later we moved the laboratory to a dedicated sequencing laboratory in

ARTICLE

doi:10.1038/nature22040

Virus genomes reveal factors that spread and sustained the Ebola epidemic

A list of authors and their affiliations appears at the end of the paper

The 2013–2016 West African epidemic caused by the Ebola virus was of unprecedented magnitude, duration and impact. Here we reconstruct the dispersal, proliferation and decline of Ebola virus throughout the region by analysing 1,610 Ebola virus genomes, which represent over 5% of the known cases. We test the association of geography, climate and demography with viral movement among administrative regions, inferring a classic 'gravity' model, with intense dispersal between larger and closer populations. Despite attenuation of international dispersal after border closures, cross-border transmission had already sown the seeds for an international epidemic, rendering these measures ineffective at curbing the epidemic. We address why the epidemic did not spread into neighbouring countries, showing that these countries were susceptible to substantial outbreaks but at lower risk of introductions. Finally, we reveal that this large epidemic was a heterogeneous and spatially dissociated collection of transmission clusters of varying size, duration and connectivity. These insights will help to inform interventions in future epidemics.

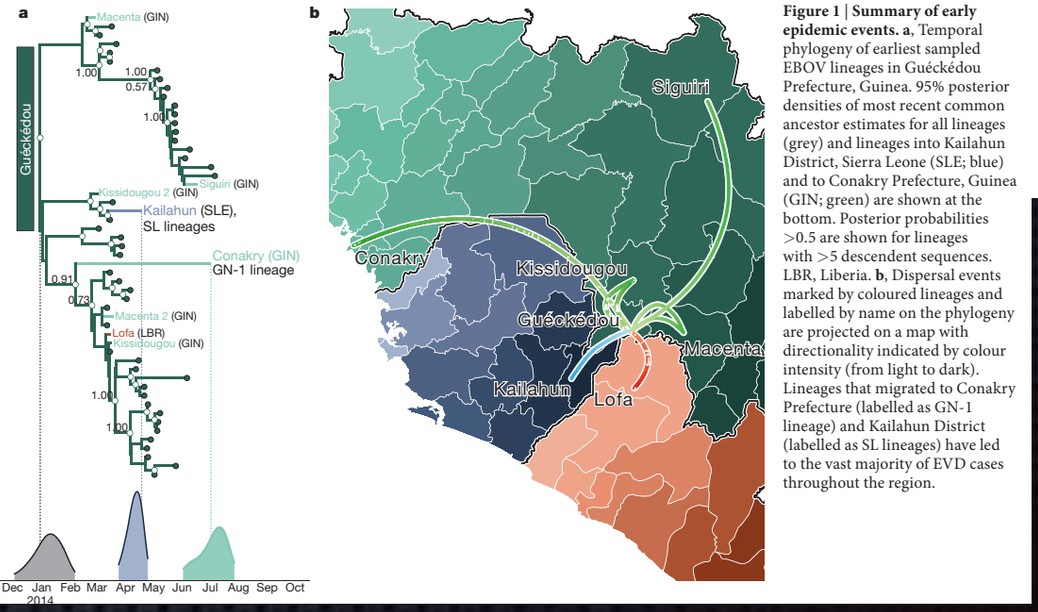


Figure 1 | Summary of early epidemic events. a, Temporal phylogeny of earliest sampled EBOV lineages in Guéckédou Prefecture, Guinea. 95% posterior densities of most recent common ancestor estimates for all lineages (grey) and lineages into Kailahun District, Sierra Leone (SLE; blue) and to Conakry Prefecture, Guinea (GIN; green) are shown at the bottom. Posterior probabilities >0.5 are shown for lineages with >5 descendent sequences. LBR, Liberia. **b**, Dispersal events marked by coloured lineages and labelled by name on the phylogeny are projected on a map with directionality indicated by colour intensity (from light to dark). Lineages that migrated to Conakry Prefecture (labelled as GN-1 lineage) and Kailahun District (labelled as SL lineages) have led to the vast majority of EVD cases throughout the region.

Nature 530, 228–232 (11 February 2016)

News

Mobile laboratories use LamPORE COVID-19 test, UK Government LamPORE evaluation report shows high accuracy

Fri 29th January 2021

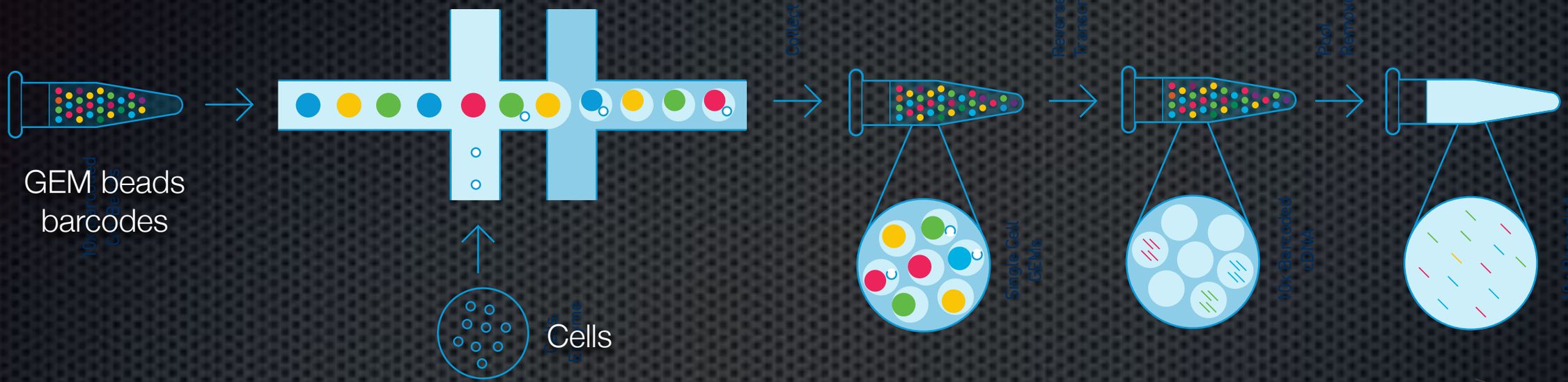
Mobile COVID-19 testing laboratories containing Oxford Nanopore's LamPORE test are now being deployed in a pilot programme, to support testing efforts, including in remote locations.

PromethION

- Integrated compute allows real-time basecalling and onward analysis
- 48 Flow Cells, up to 3,000 nanopores
- Up to 12 Tb in 48 hours for the whole device
- Flow Cells can be run individually or concurrently
- Read modified bases (5mC…)
- Direct RNA sequencing
- Longest read >1 000 000

Single-cell sequencing

10x Genomics Chromium system



Single Cell Genomics
Copy Number Variation

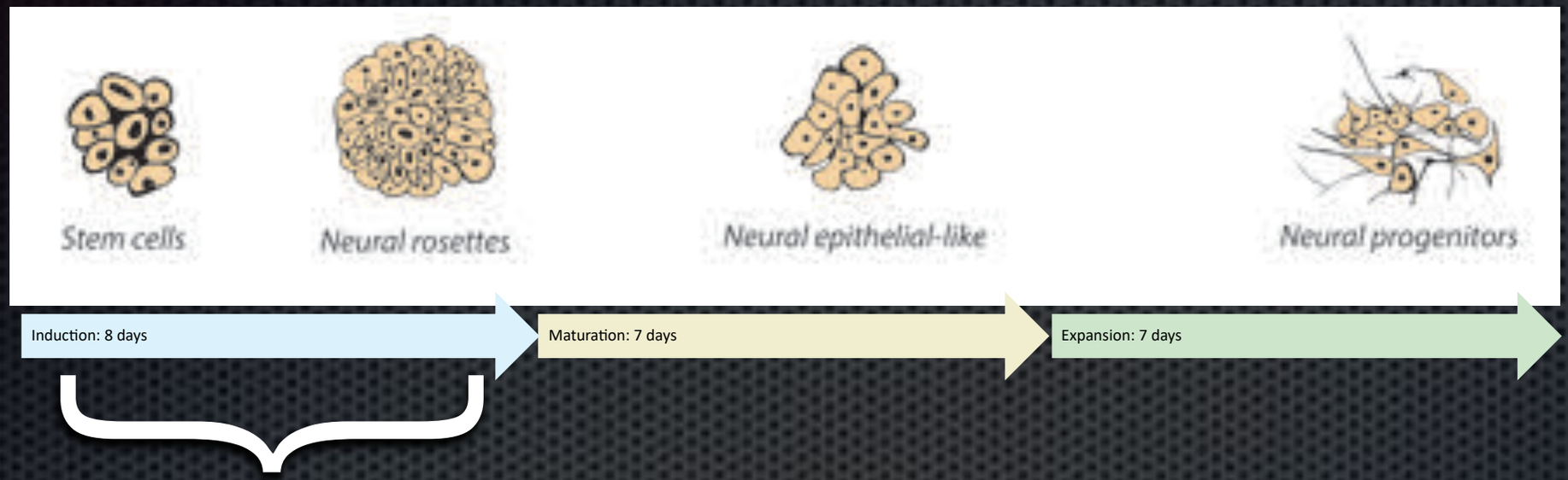
Single Cell Epigenomics
Chromatin Accessibility *NEW!*

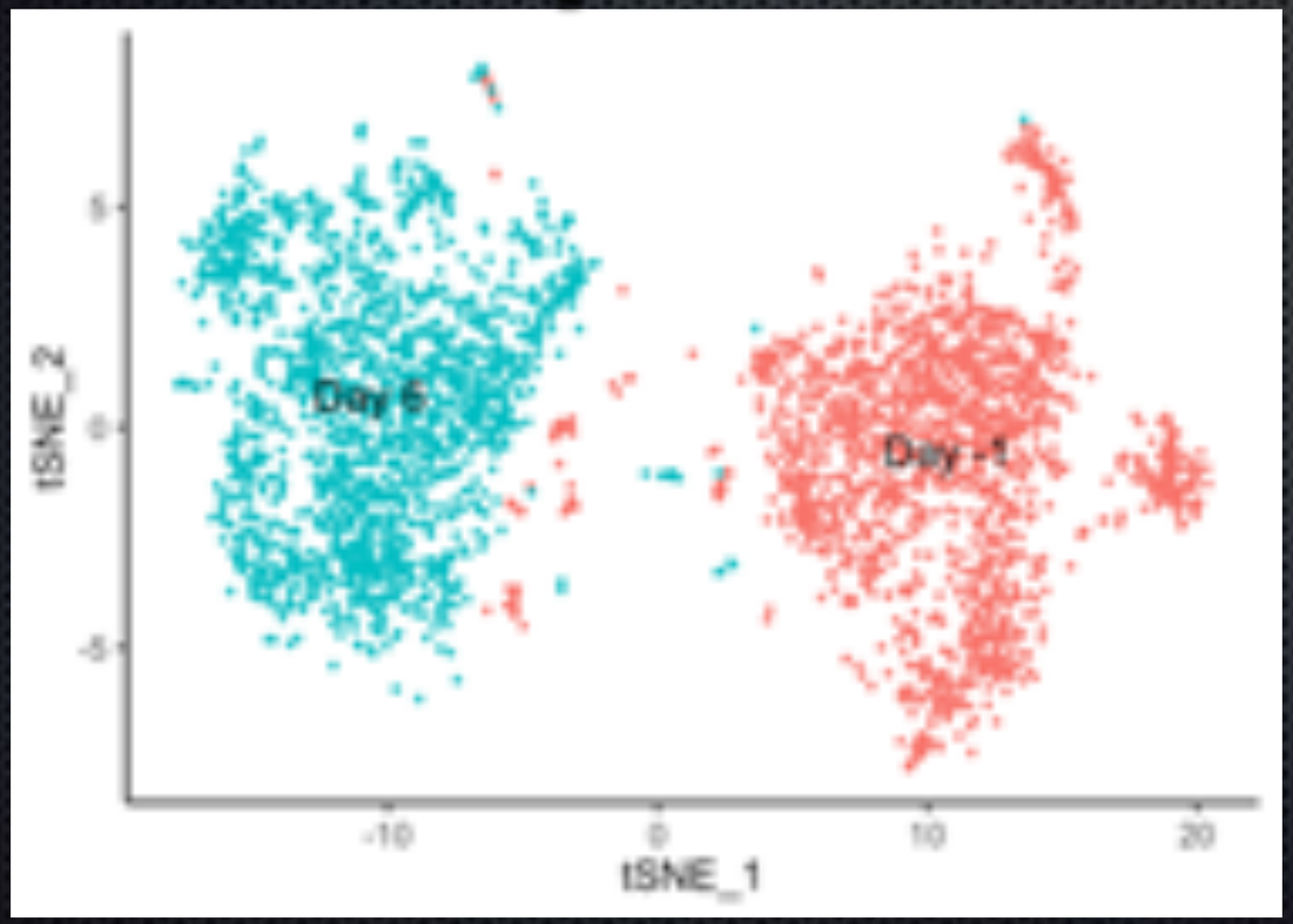
Single Cell Transcriptomics
Gene Expression Profiling *IMPROVED!*
Gene Expression CRISPR Screening *NEW!*
Gene & Cell Surface Protein *NEW!*
Immune Profiling
Immune Profiling & Cell Surface Protein *NEW!*
Immune Profiling & Antigen Specificity *NEW!*

Linked-Reads Genomics
Whole Genome Sequencing
Exome Sequencing
de novo Assembly

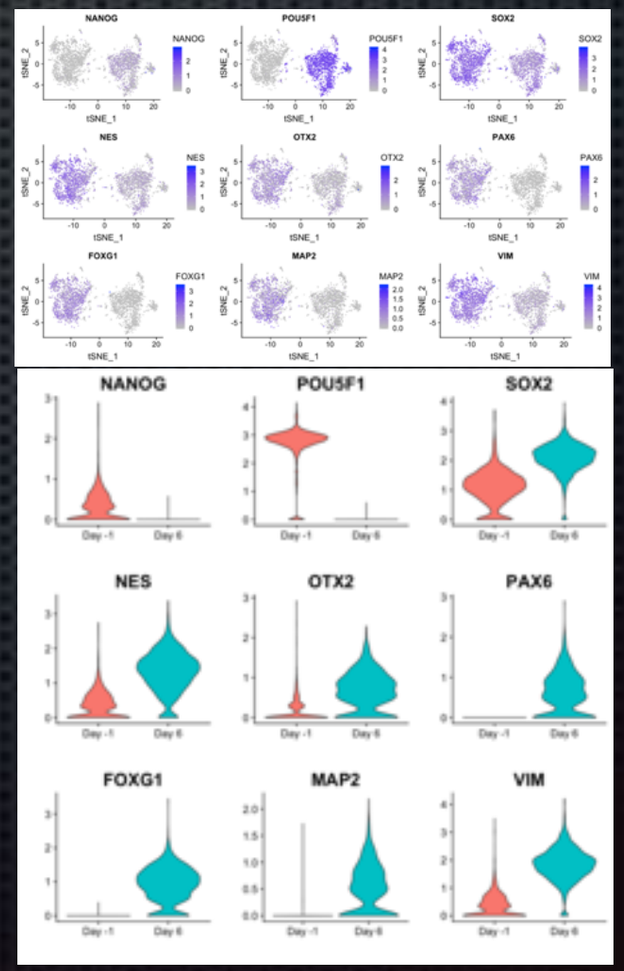
ous-seq@sequencing.uio.no

Single-cell RNAseq





Cells

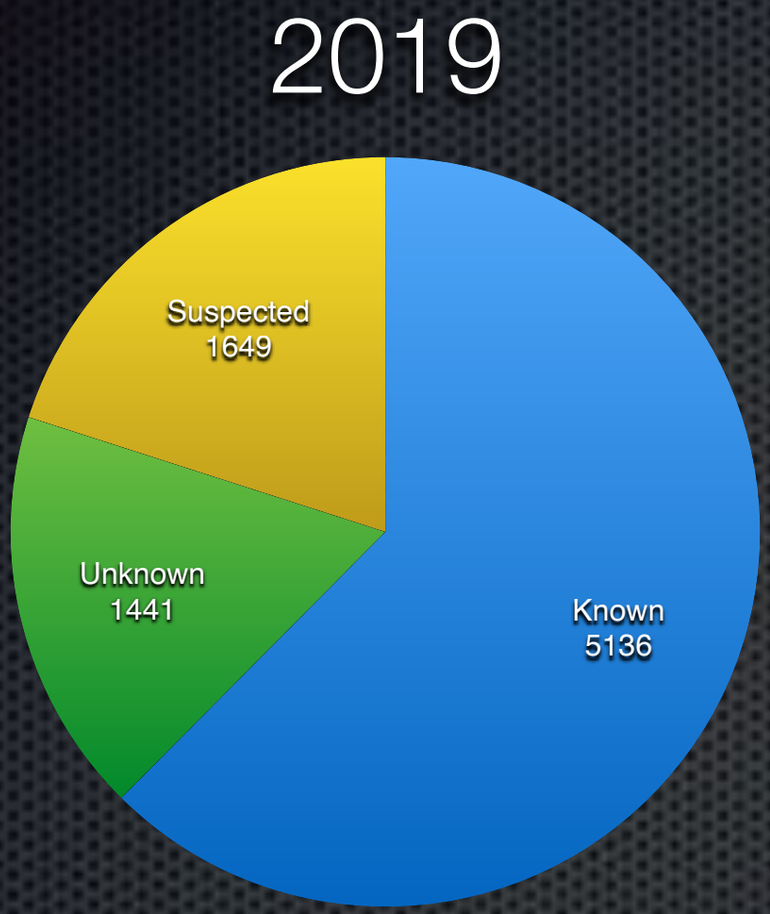


Genes

HTS and medical genetics

Exome sequencing

Mendelian disease in man



- 5136 known molecular basis
- 1649 suspected molecular basis
- 1441 unknown

OMIM
online inheritance in man

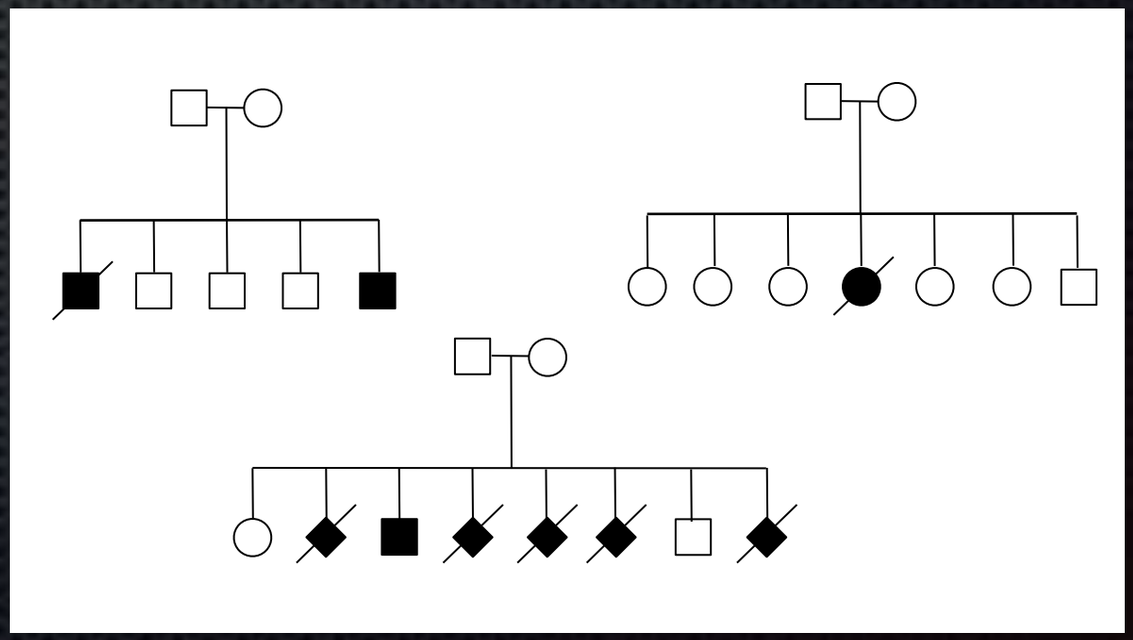
www.omim.org

Genetic disease: the challenge

- Single gene (monogenic/Mendelian) disorders
- ~50% children with congenital syndromes do not receive a firm diagnosis
- Many of these children have disorders that are primarily genetic in origin
- Clinical phenotypes can be very rare, non-specific and variable
- Many phenotypes are **genetically heterogeneous** - many genes/one disease

- Human genome is (quite) big
  - ~3 billion bases pairs
  - ~20 000 genes

- How can we (rapidly) identify a mutation causing disease?

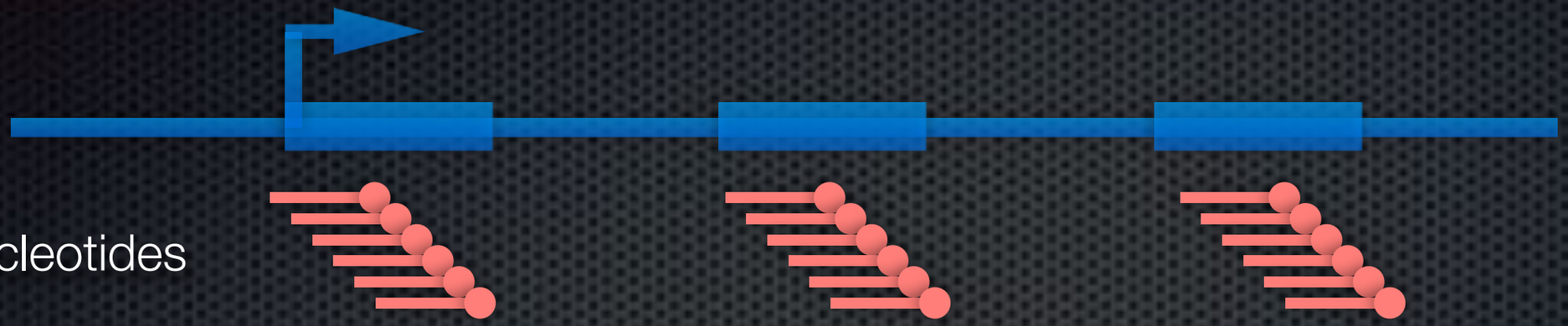


Where are disease causing variants?

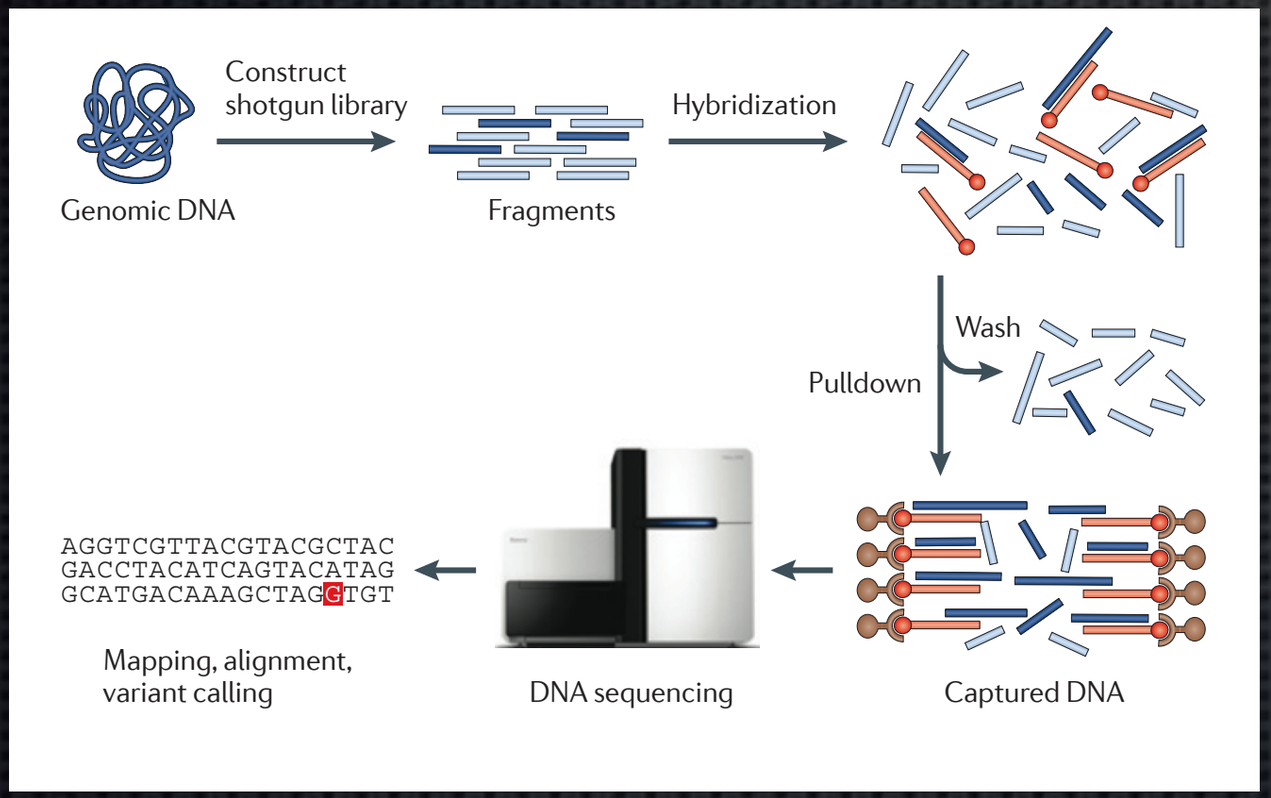
- Could be anywhere in the genome
- Genome sequencing still expensive

- Where are the majority of disease-causing variants?
  - The **exome** - protein coding regions of the genome
    - ~1.5% of the genome
    - ~80% of mutations with large effects on disease-related traits
  - ***Exome sequencing***

Exome sequencing

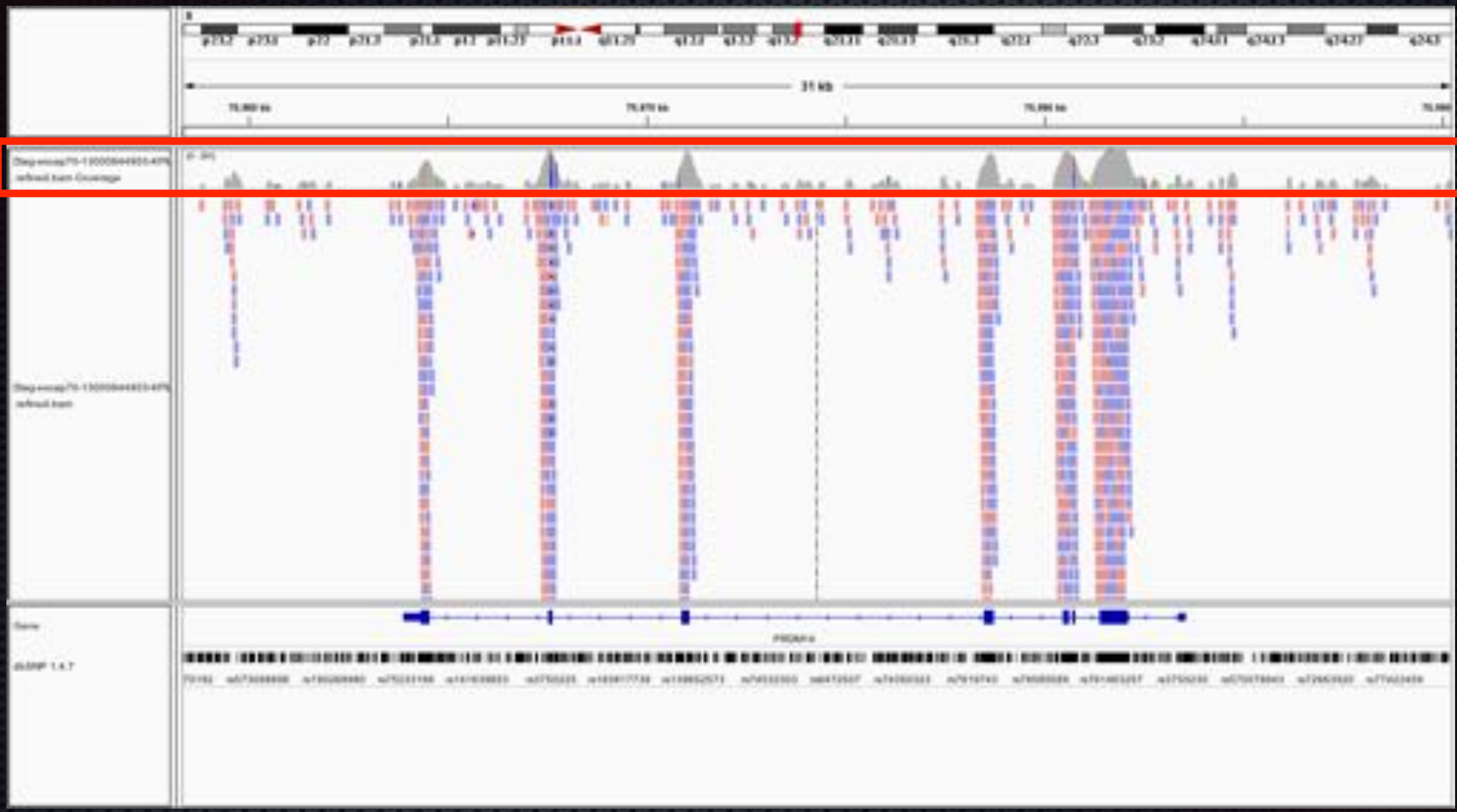


oligonucleotides



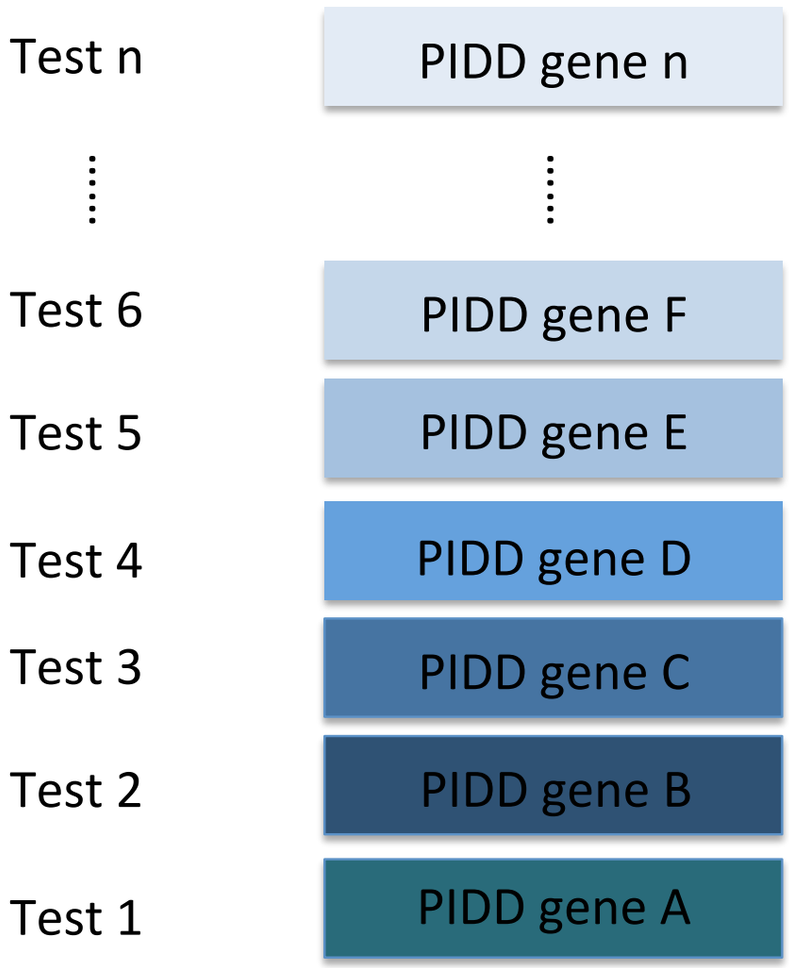
- Agilent SureSelect kit
- Sequence 4 samples/lane HiSeq
- ~100x coverage per exome
- bwa for alignment
- GATK to call variants

Exome sequencing data



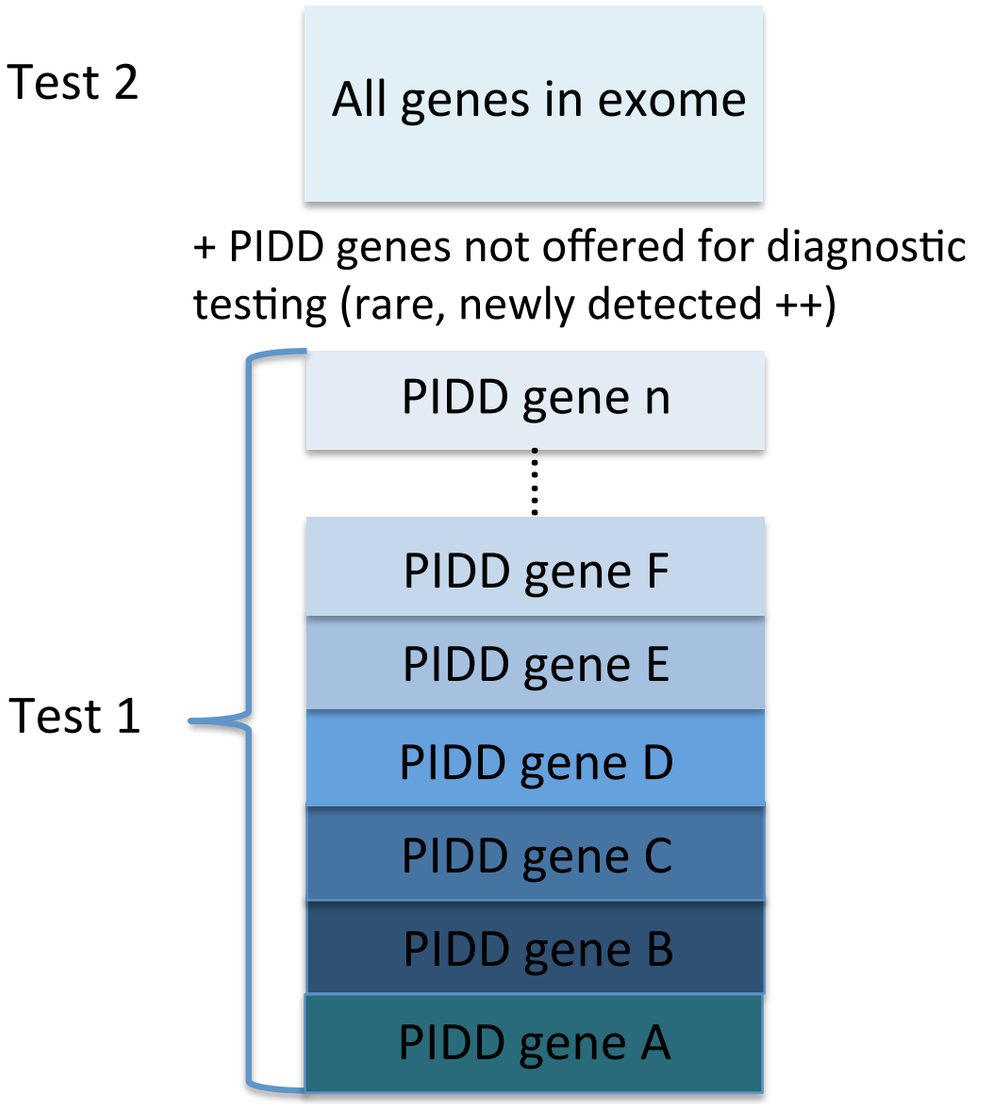
Conventional genetic analysis

1. Molecular genetic diagnosis in known PIDD gene offered for diagnostic testing
2. No disease-causing variant found in known PIDD genes



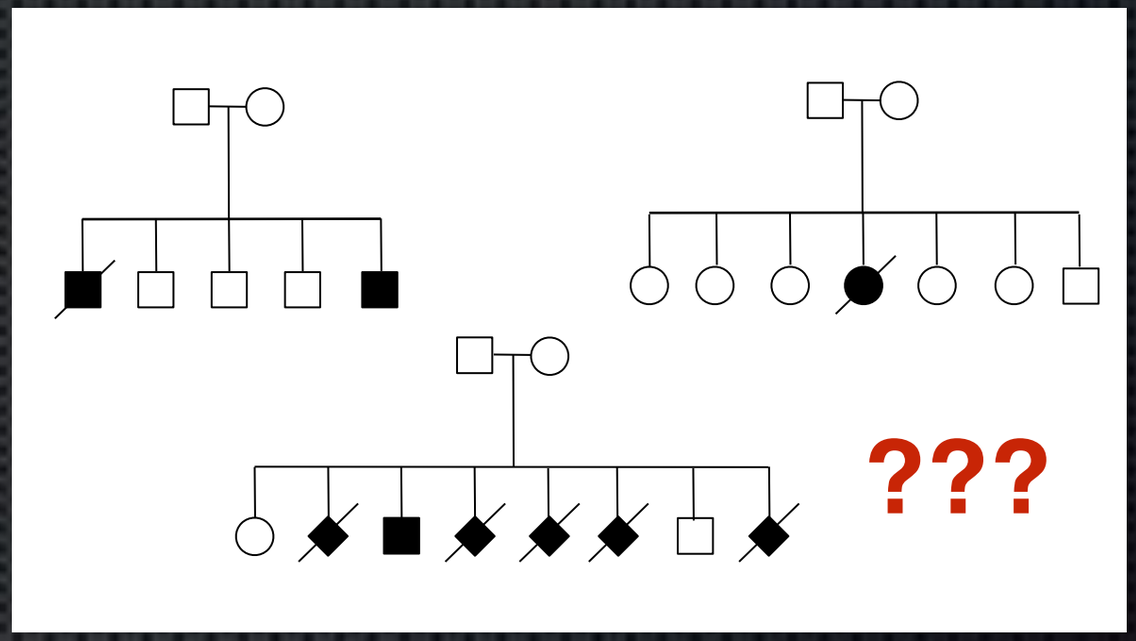
Exome sequencing

1. Molecular genetic diagnosis in known PIDD gene
2. No disease-causing variant found in known PIDD genes



What's in an exome?

- Compare two exomes
- ~20 000 variants
- ~10 000 silent
- ~10 000 missense
- ~100 nonsense
  - loss-of-function variant

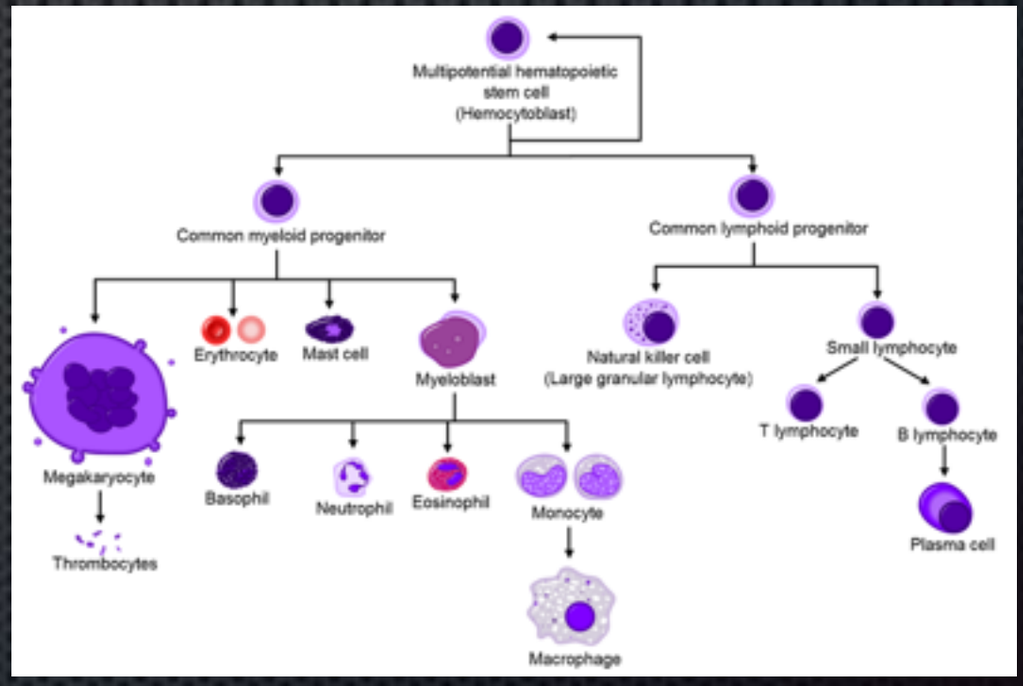


Challenge to identify disease-causing variant

Primary immunodeficiency diseases (PIDD)

- Part of the body's immune system is missing or does not function normally
- Clinically and genetically heterogenous
- Symptoms range from severe/life-threatening to milder phenotype: infections, autoimmune or lymphoproliferative phenomena, skin affection or various dysmorphic features
- Occurrence of 1:500

- Primary -> genetic origin
- > 300 genes known
- Many unknown genes
- ~50% cases lack genetic diagnosis



- Specific diagnosis can direct targeted and curative therapy

Project organisation

Strategy: **All** PIDD phenotypes -WES will yield unbiased and novel insights

ExomeSeq Oslo, Norway
Families: 123
Participants: 244

ExomeSeq Houston, TX
Families: >200
Participants: >400

ExomeSeq + CNV screen
Families: 78
Participants: 83

278 families

ExomeSeq + CNV screen
Families: 200
Participants: 400

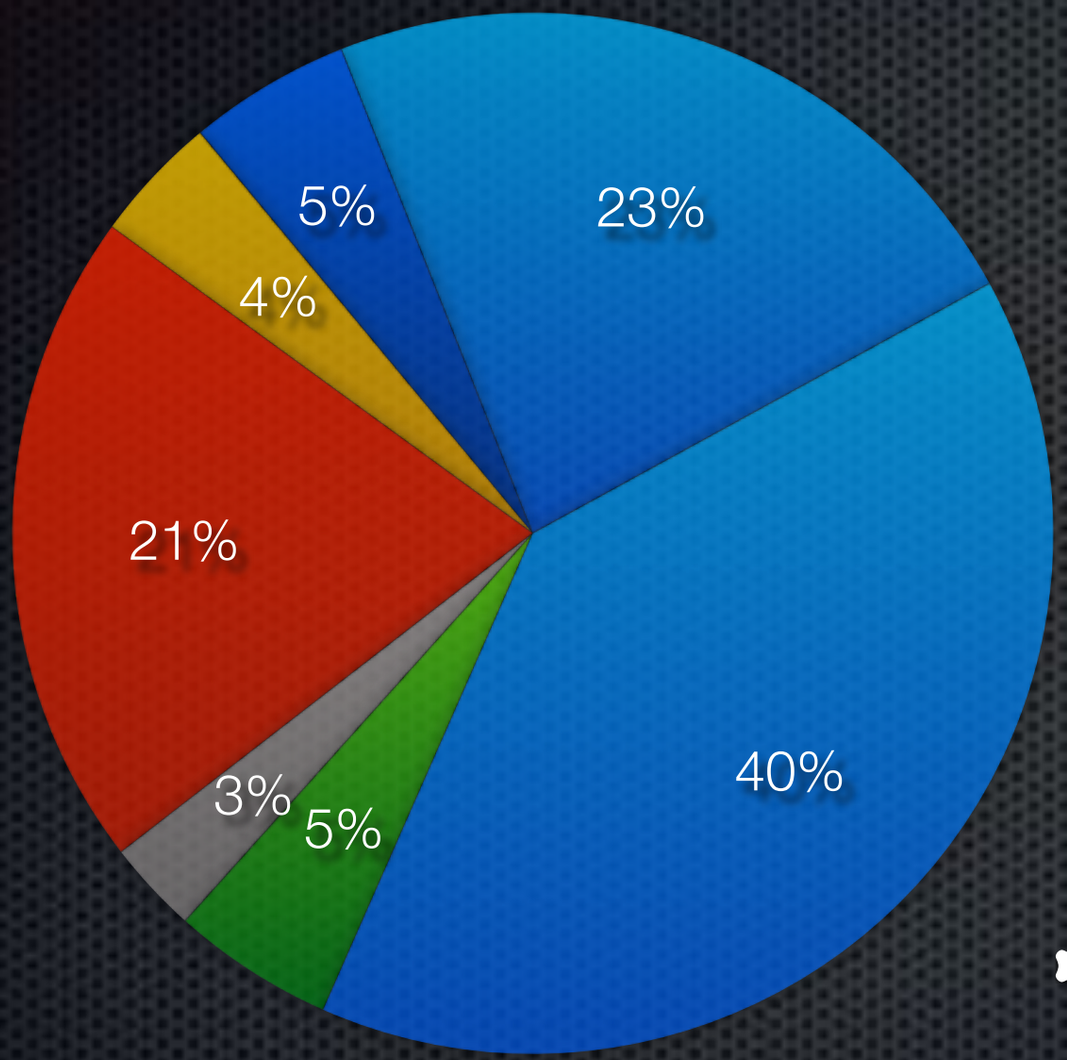
Evaluation of genetic findings

PIDD gene list

Exome-wide

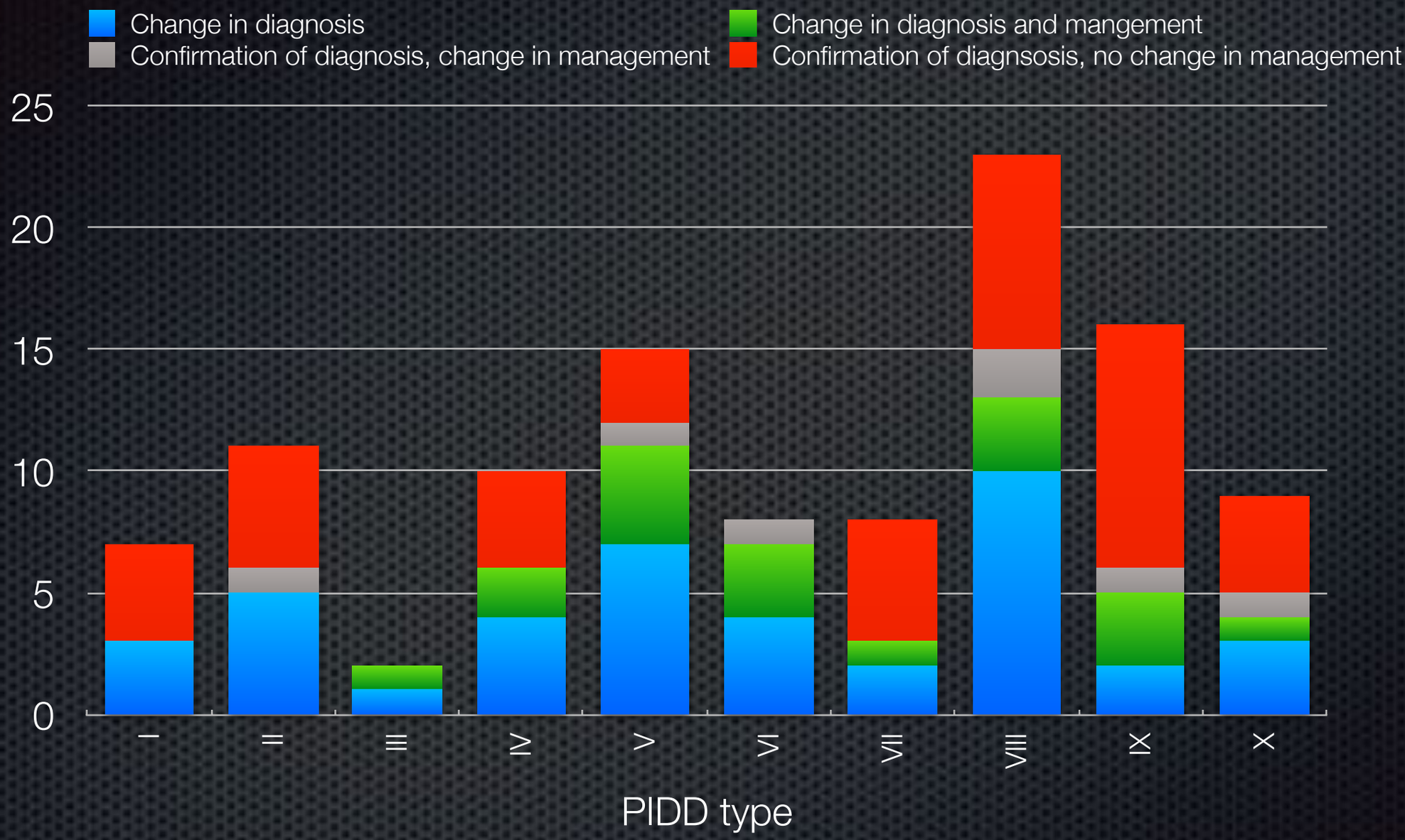
Diagnosis in 278 families

- SNPs and CNVs



- 40-45% diagnostic rate
- 110 families: 88 different genes

Why does a genetic diagnosis matter?



- 55% change in diagnosis
- 25% change in management

REVIEW

Case for genome sequencing in infants and children with rare, undiagnosed or genetic diseases

David Bick,[1] Marilyn Jones,[2] Stacie L Taylor,[3] Ryan J Taft,[3] John Belmont[3]

Table 1 Select studies illustrating the diagnostic variability of genetic and genomic testing.

Study		Publication date	Number of subjects	Age (mean or median)	Clinical indication	Technology	Diagnosis rate (%)
Soden *et al*[25]	*Sci Trans Med*	Oct 2014	100	7 years	NDD	GS R-GS ES	47 73 40
Lee *et al*[97]	*JAMA*	Nov 2014	814	>18 years	Any	ES	26
Yang *et al*[98]	*JAMA*	Nov 2014	2000	6 years	DD	ES	25.2
Wright *et al*[99]	*Lancet*	Dec 2014	1133	6 years	NDD	ES	27
Gilissen *et al*[44]	*Nature*	Jul 2014	50	>18 years	ID	GS	42
Willig *et al*[47]	*Lancet Respir Med*	May 2015	35	>4 months	Any	R-GS SCP	57 9
Petrikin *et al*[48]	*SEM Perinatol*	Dec 2015	35	26 days	Any	GS	57
Stavropoulos *et al*[43]	*NPJ Genom Med*	Jan 2016	110	>18 years	NDD	GS CMA CMA+TGS	34 8 13
Rump *et al*[96]	*BMC Med Genom*	Feb 2016	38	10 years	ID	ES	29
Vissers *et al*[74]	*Genet Med*	Mar 2017	150	>18 years	NDD	ES SCP	29.3 7.3
Lionel *et al*[49]	*Genet Med*	Aug 2017	103	>18 years	Any	GS	41
Petrikin *et al*[50]	*NPJ Genom Med*	Feb 2018	65	>4 months	Any	R-GS Standard tests	31 3
van Diemen *et al*[100]	*Paediatrics*	Oct 2017	23	>12 months	Any	R-GS	30
Farnaes *et al*[71]	*NPJ Genome Med*	Apr 2018	42	>4 months	Any	R-GS Standard tests	48 10

CMA, chromosomal microarray; DD, developmental delay;ES, exome sequencing; GS, genome sequencing; ID, intellectual disability;NDD, neurodevelopmental disorder; R-GS, rapid genome sequencing; SCP, standard care pathway; TGS, targeted gene sequencing.

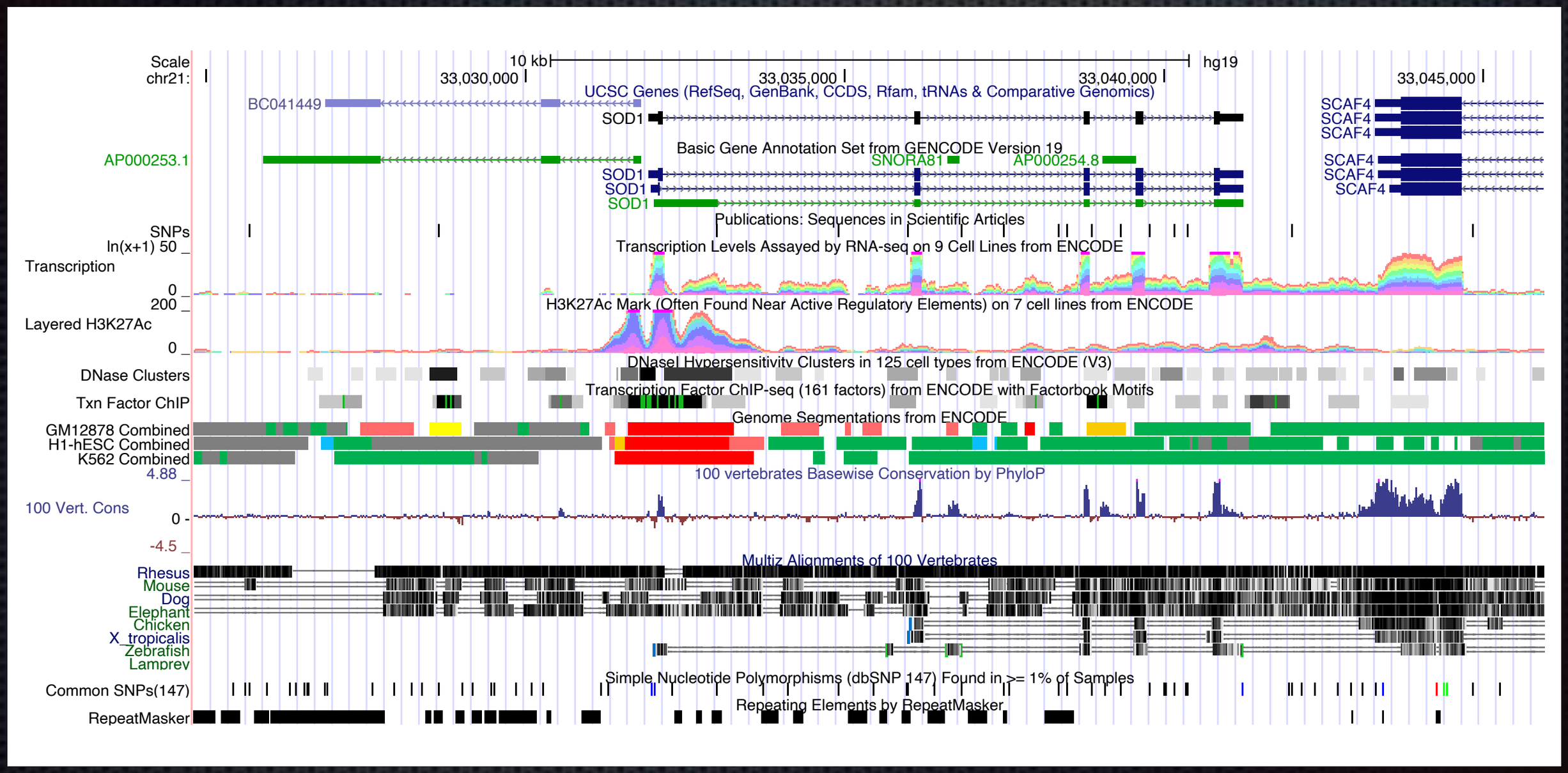
Variants in genome sequencing

- SNVs
  - Exome - ~20 000
  - **Genome - ~3 500 000**
- CNVs: Mbs

- How do we deal with so many variants?
- Genome contains many non-coding functional regions

Encode: ENCyclopedia of DNA Elements

Systematic effort to identify functional regions of the genome



- Evidence for function

Summary

- High-throughput sequencing
  - Dramatic increase in sequence production
  - Many applications on one platform
  - Field still moving very quickly
  - Huge impact on biology