High Throughput Sequencing and Multi-omics Bioinformatics

From Data to Knowledge

Prof. Ioannis Xenarios

CIG Center for Integrative Genomics (UNIL)
Departement of Biochemistry (UNIGE)
Health2030 Genome Center
Ludwig Institute for Cancer Research

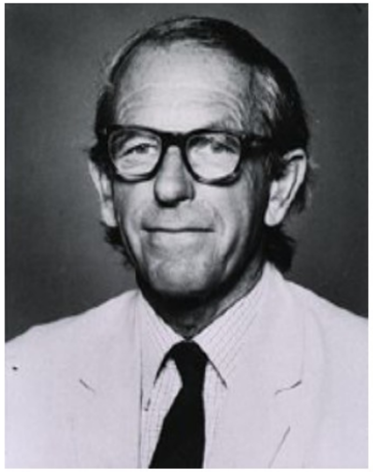
Sequencing is becoming a **commodity**

- Sequencing a whole genome allows to identify variation (SNVs)
- Sequencing and characterizing the mRNA (RNA content of a cell allows to identify – expression level, isoforms)
- Sequencing different tissues enables comparison of expression pattern amongst tissues
- Sequencing different microorganisms enables the characterization of communities of bacteria in the environment, our gut, or other human body cavities.
- Sequencing pathogenic virus allows to follow their evolution (applied nowdays to the SARS-Cov2)

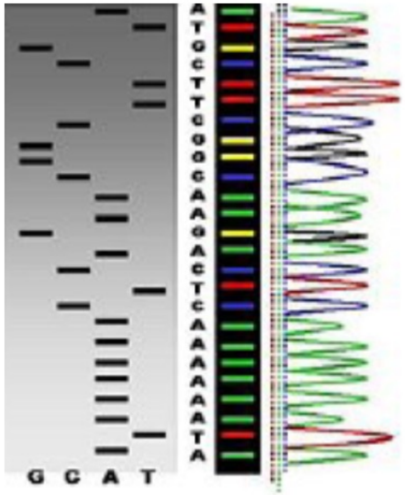
High Throughput Analysis is not yet

Last century sequencing

- Originally with a radiolabeled (P^{32}/S^{35}) nucleotides which was incorporated by DNA synthesis and subsequently used in electrophoretic gels.
- Four radiolabelled nucleotides we used and loaded separately in each « lane »
- Around 150-300bp(when you had good eyes) could be read through
- Slow painfull **(1/3 of my PhD thesis sequencing immunoglobulin and T cell clones…)**
- Improvement introduced capillary electrophoresis

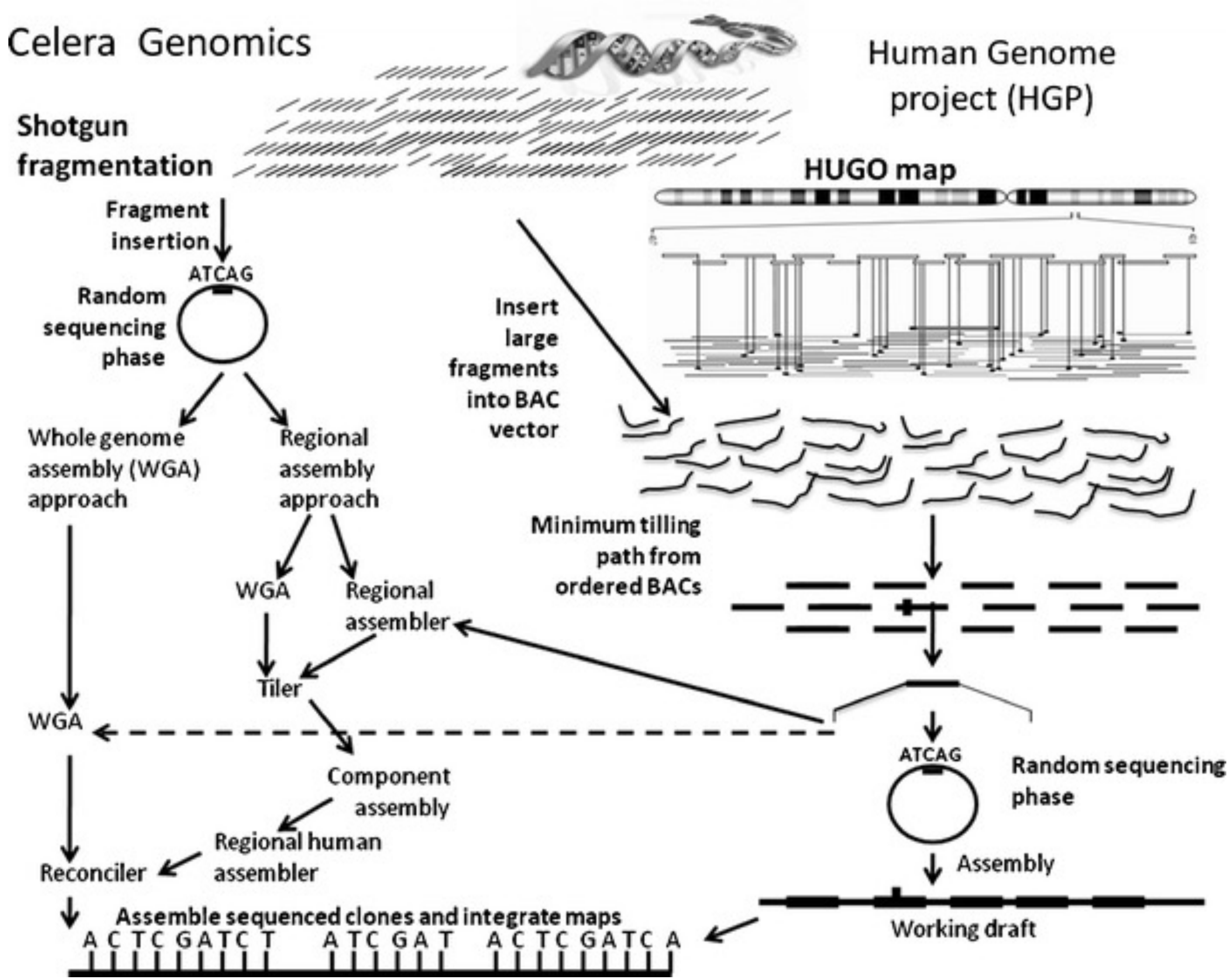


Frederick Sanger
Nobel Prize (1980)



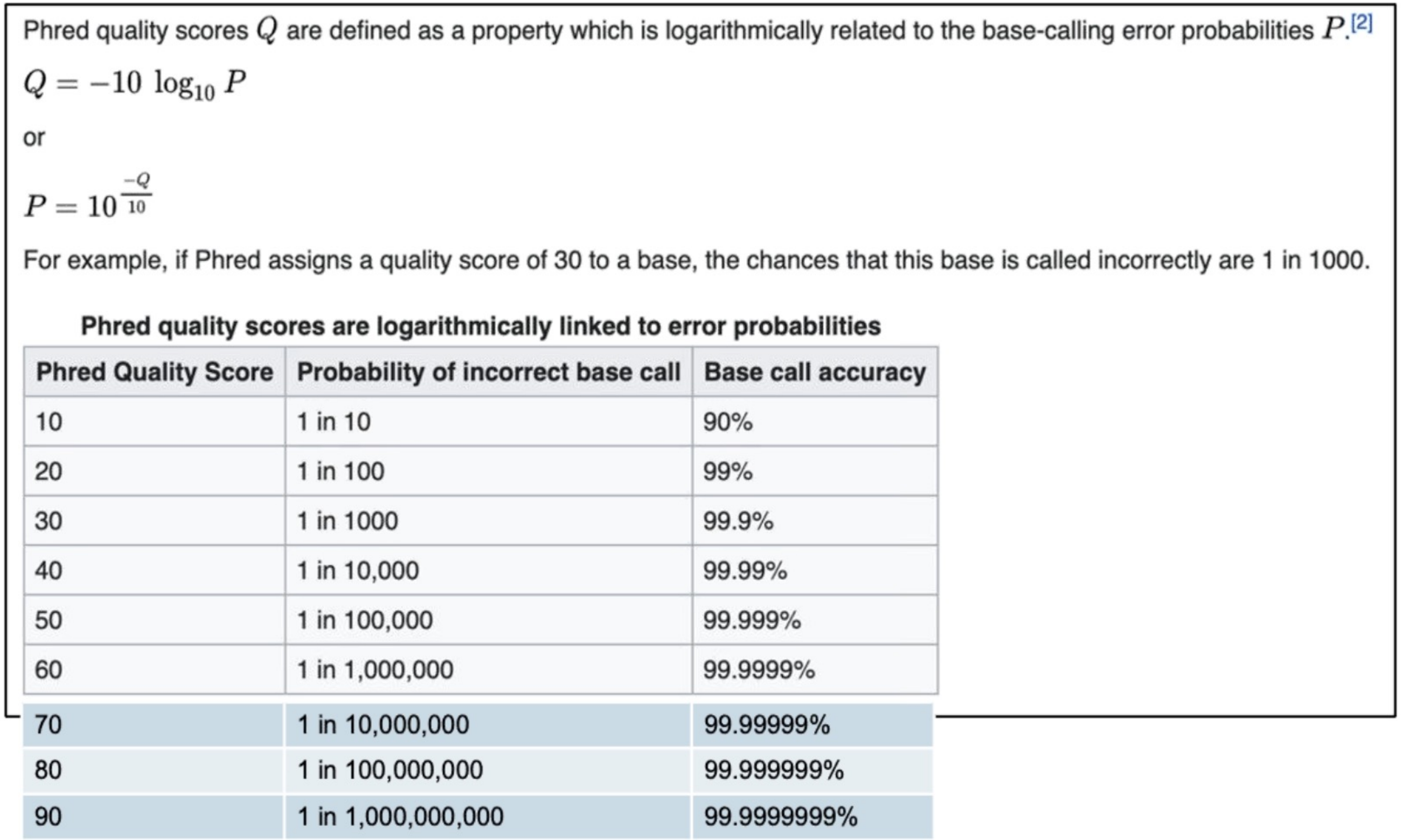
397-387
377-386
3730xl DNA Analyzer
397

How was the Human genome done


Celera Genomics
Human Genome project (HGP)
Shotgun fragmentation
HUGO map
Fragment insertion
ATCAG
Random sequencing phase
Insert large fragments into BAC vector
Whole genome assembly (WGA) approach
Regional assembly approach
Minimum tilling path from ordered BACs
WGA
Regional assembler
Tiler
WGA
Component assembly
ATCAG
Random sequencing phase
Regional human assembler
Reconciler
Assembly
Assemble sequenced clones and integrate maps
Working draft
A C T C G A T C T A T C G A T A C T C G A T C A

Technologies for sequencing DNA/RNA short/long read , single-cell and optical mapping

All these machines have an intrinsic sequencing errors



Phred quality scores Q are defined as a property which is logarithmically related to the base-calling error probabilities P.[2]

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

or

$$P = 10^{\frac{-Q}{10}}$$

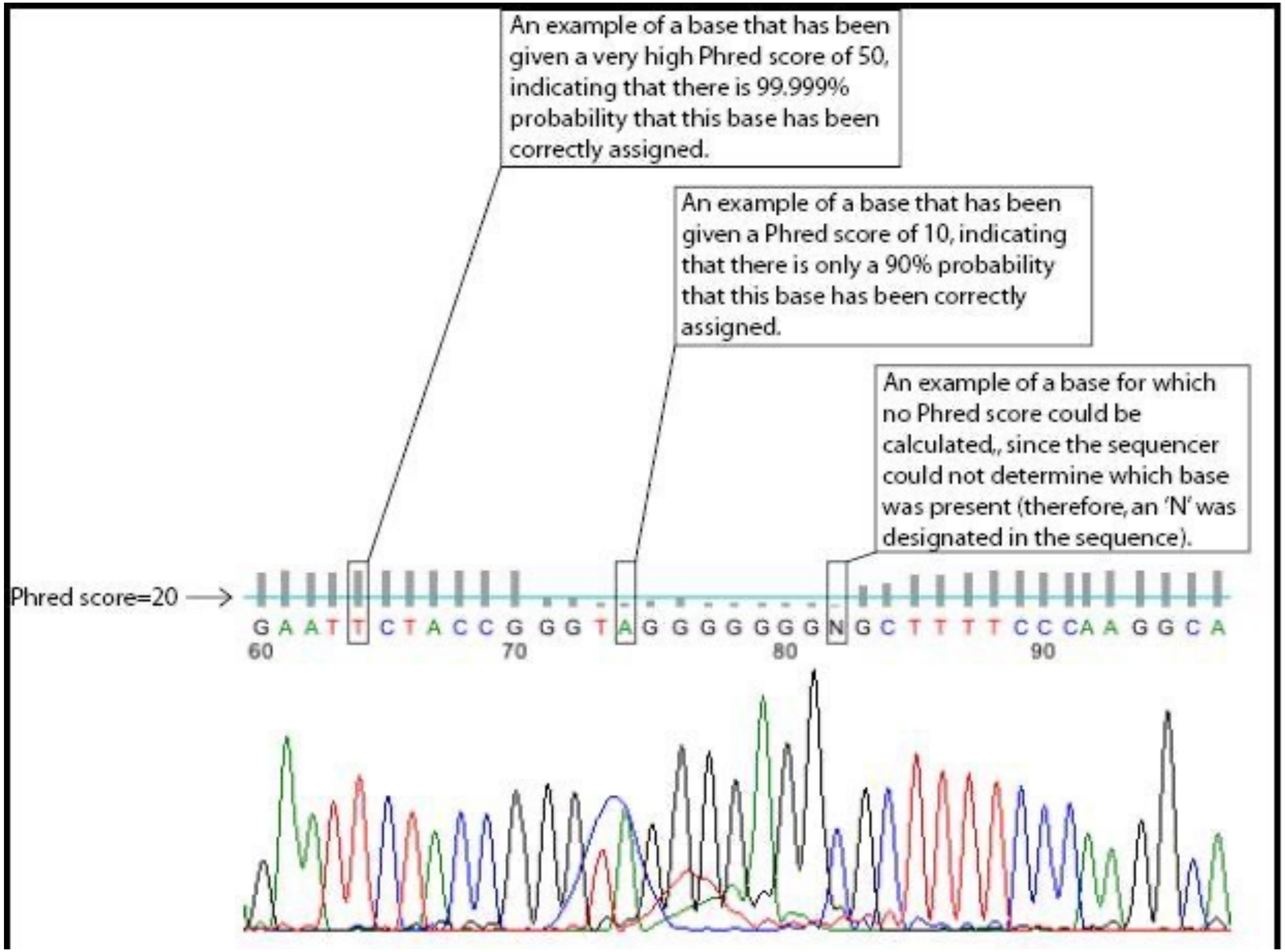
For example, if Phred assigns a quality score of 30 to a base, the chances that this base is called incorrectly are 1 in 1000.

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%
70	1 in 10,000,000	99.99999%
80	1 in 100,000,000	99.999999%
90	1 in 1,000,000,000	99.9999999%

https://en.wikipedia.org/wiki/Phred_quality_score

The good old Sanger … the HGP defined the Phred score



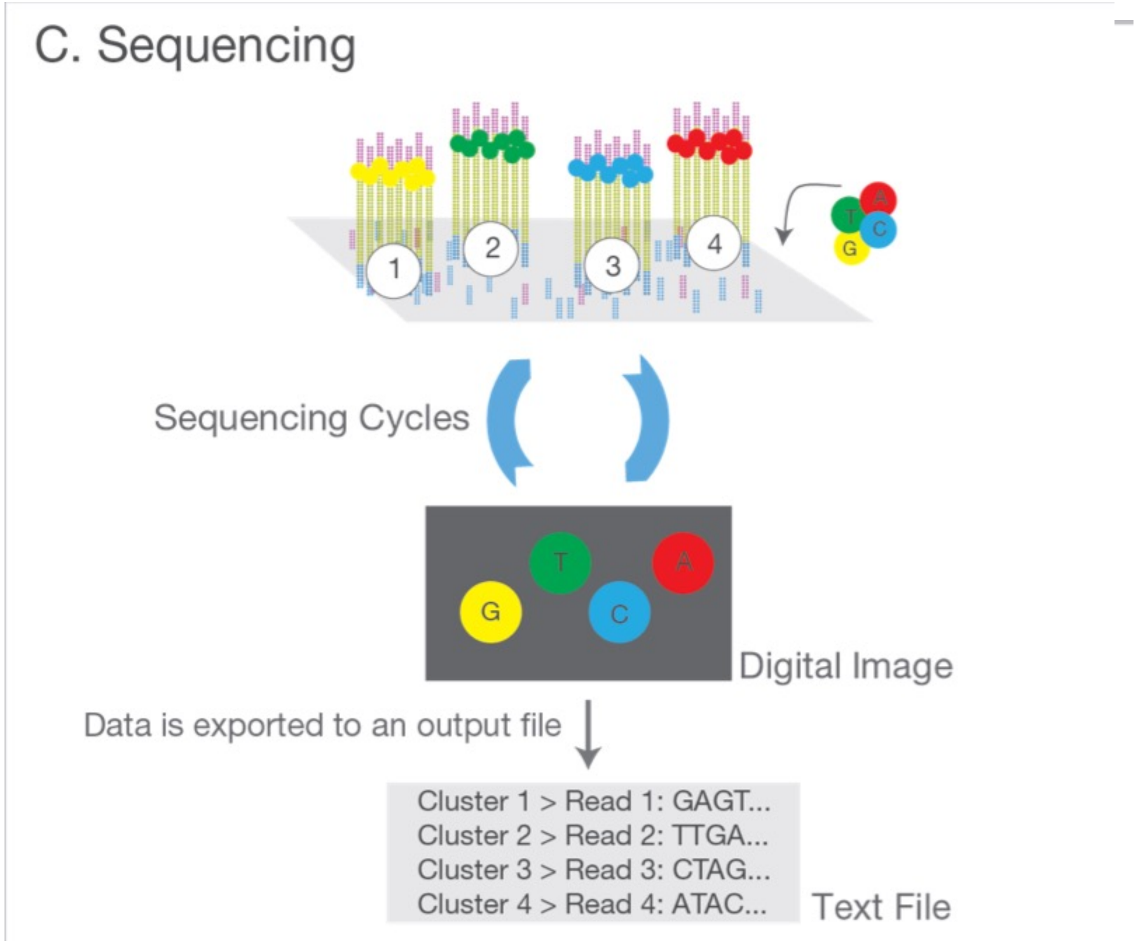
An example of a base that has been given a very high Phred score of 50, indicating that there is 99.999% probability that this base has been correctly assigned.

An example of a base that has been given a Phred score of 10, indicating that there is only a 90% probability that this base has been correctly assigned.

An example of a base for which no Phred score could be calculated,, since the sequencer could not determine which base was present (therefore, an 'N' was designated in the sequence).

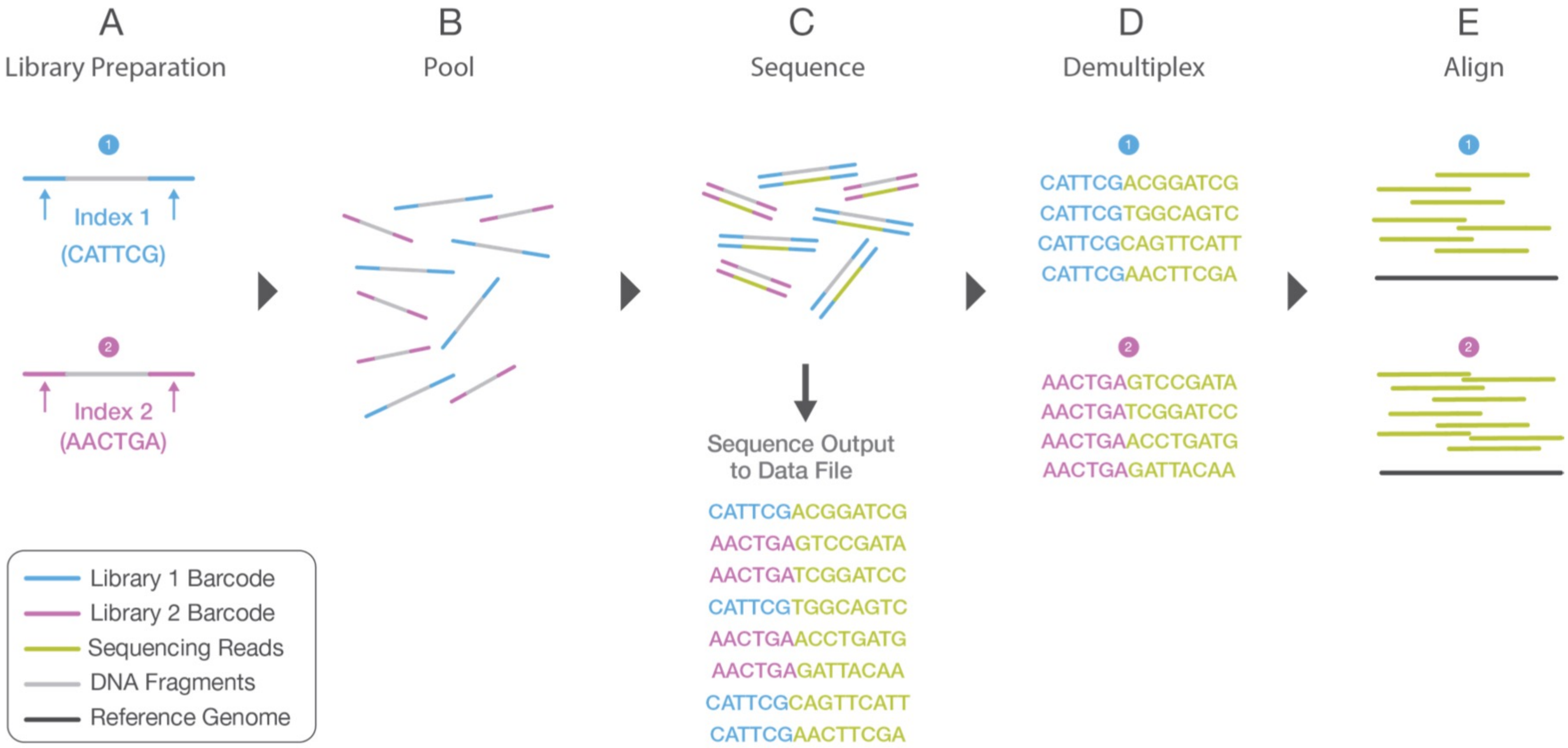
Phred score=20
G A A T T C T A C C G G G T A G G G G G G N G C T T T T C C C A A G G C A
60
70
80
90

Short read sequencing (**example Illumina**)



C. Sequencing

Cluster 1 > Read 1: GAGT...
Cluster 2 > Read 2: TTGA...
Cluster 3 > Read 3: CTAG...
Cluster 4 > Read 4: ATAC...



Sequence Output to Data File

CATTCGACGGATCG
AACTGAGTCCGATA
AACTGATCGGATCC
CATTCGTGGCAGTC
AACTGAACCTGATG
AACTGAGATTACAA
CATTCGCAGTTCATT
CATTCGAACTTCGA

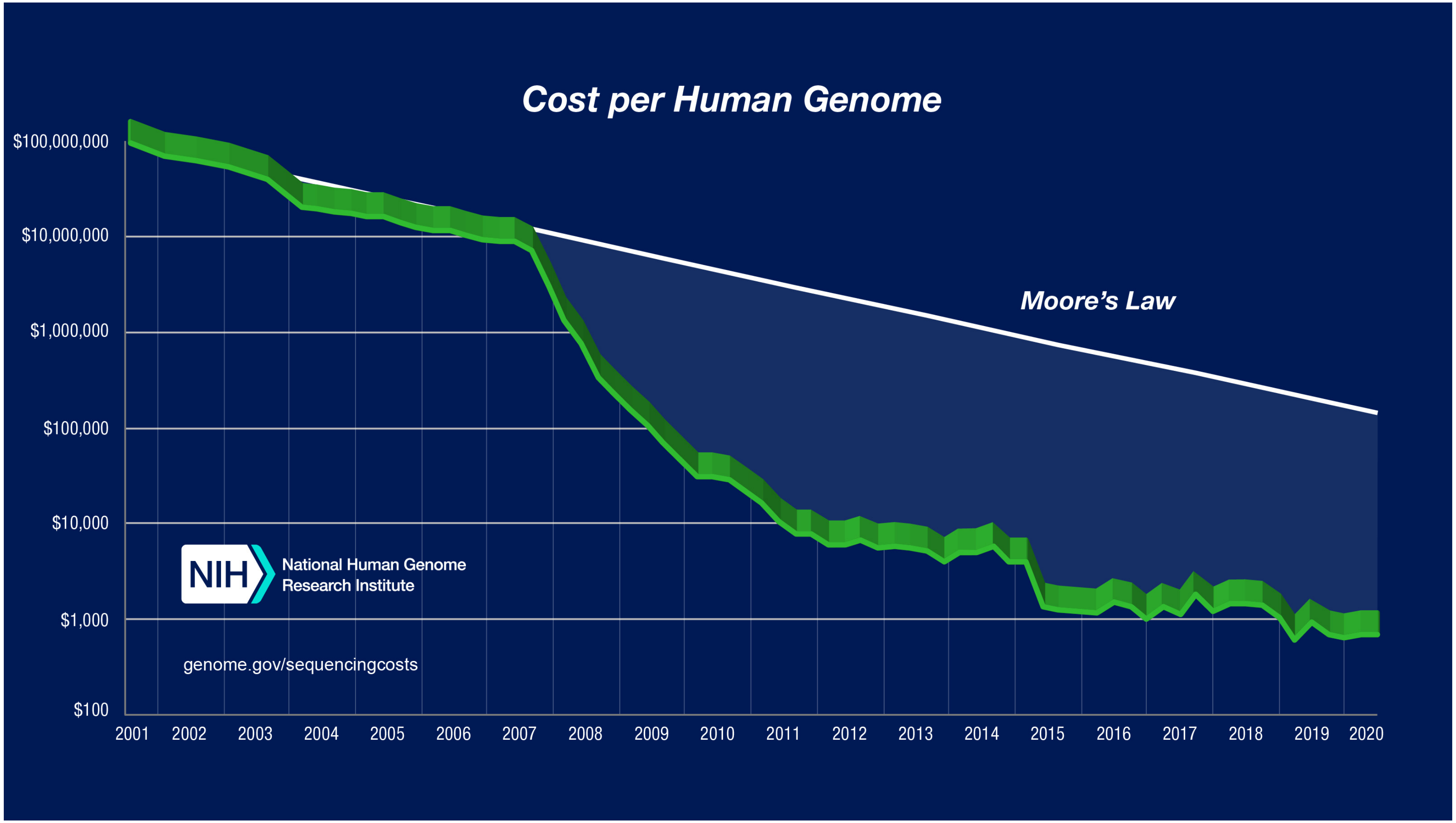
Demultiplex

1
CATTCGACGGATCG
CATTCGTGGCAGTC
CATTCGCAGTTCATT
CATTCGAACTTCGA

2
AACTGAGTCCGATA
AACTGATCGGATCC
AACTGAACCTGATG
AACTGAGATTACAA

Up to 384 unique barcode can be used per library

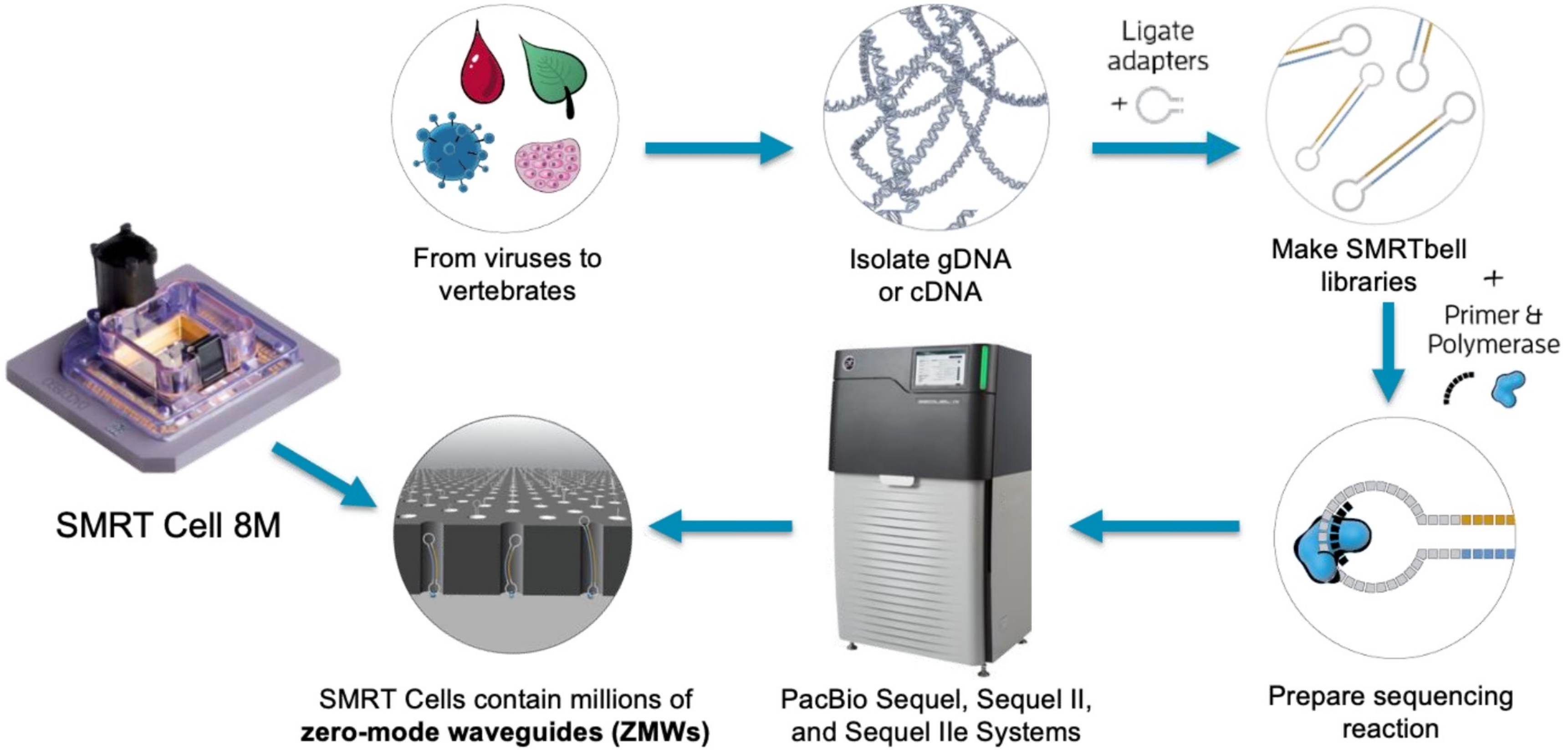
Cost per base and per genome over time



Health2030 Genome Center 850 CHF (2021)

2023 down to **150 CHF**

Long read PacBio



From viruses to vertebrates

Isolate gDNA or cDNA

Ligate adapters +

Make SMRTbell libraries

+ Primer & Polymerase

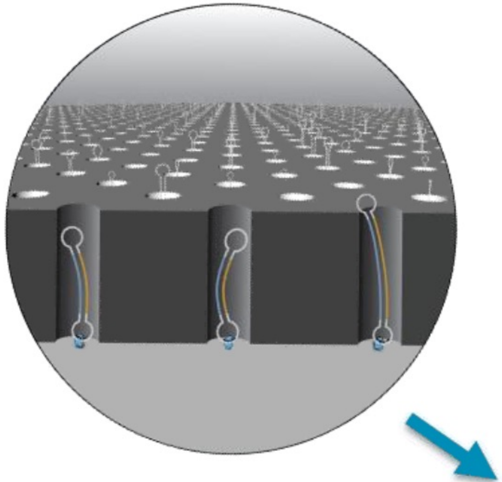
Prepare sequencing reaction

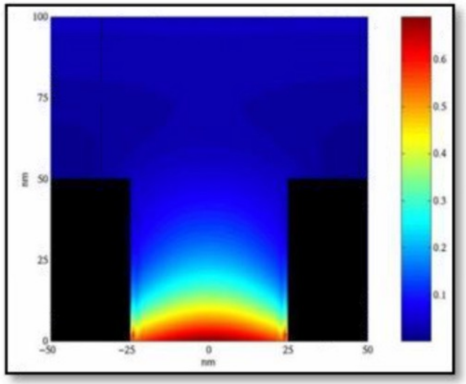
PacBio Sequel, Sequel II, and Sequel IIe Systems

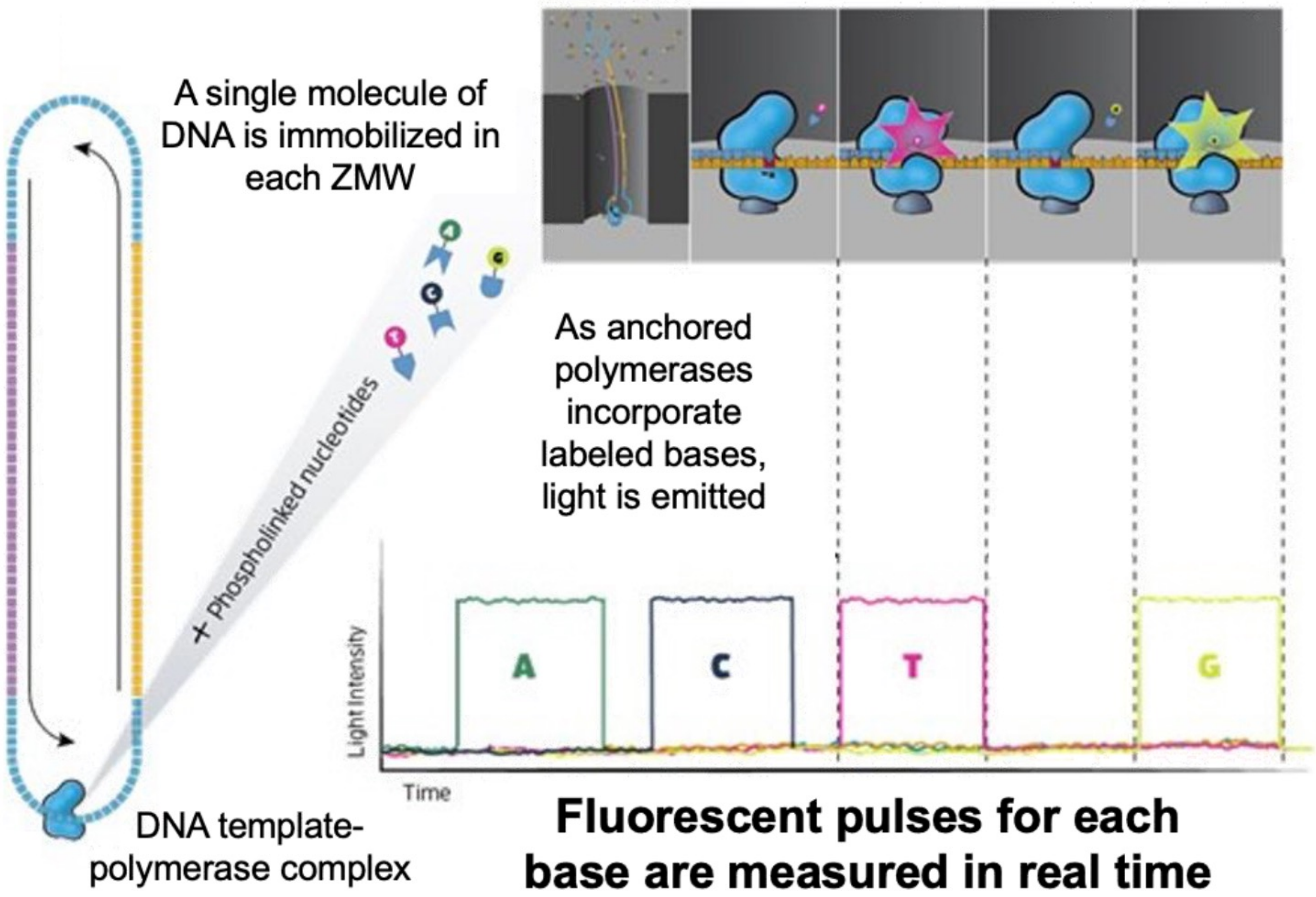
SMRT Cells contain millions of **zero-mode waveguides (ZMWs)**

SMRT Cell 8M

Watching the movie a DNA molecule





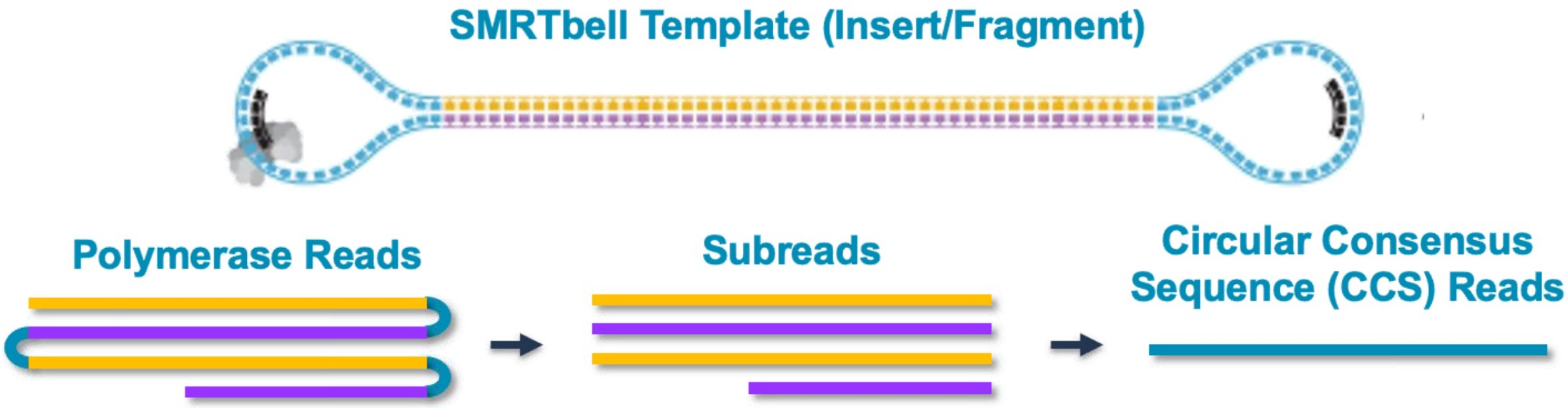


A single molecule of DNA is immobilized in each ZMW

As anchored polymerases incorporate labeled bases, light is emitted

DNA template-polymerase complex

Fluorescent pulses for each base are measured in real time



SMRTbell Template (Insert/Fragment)

Polymerase Reads

Definition:

- Linear sequence of nucleotides incorporated by polymerase while reading a SMRTbell template
- Includes adapters
- 1 molecule → 1 polymerase read

Subreads

Definition:

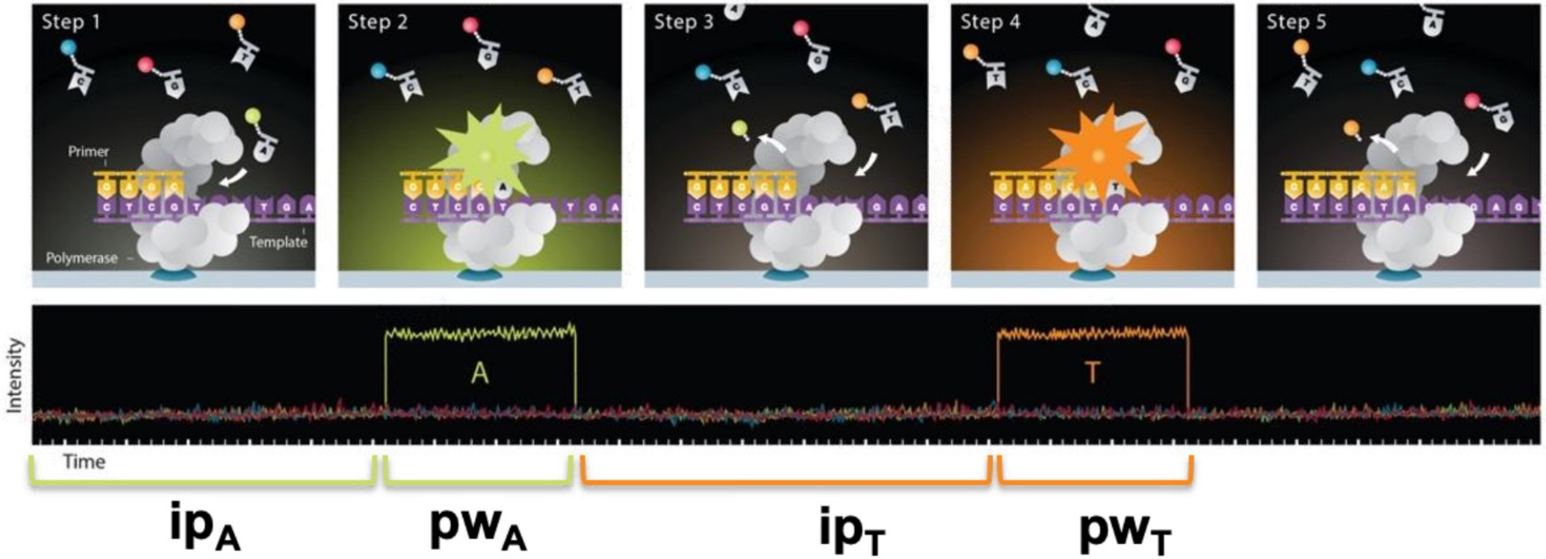
- Set of linear sequences of nucleotides in forward or reverse strand of SMRTbell template
- Adapters removed (scraps)
- 1 molecule → ≥1 subreads

Circular Consensus Sequence (CCS) Reads

Definition:

- Linear sequence of nucleotides calculated by consensus of subreads for single molecule
- **HiFi reads: ≥Q20 read quality**
- 1 molecule → 1 CCS read

PacBio capture the DNA bases modifications



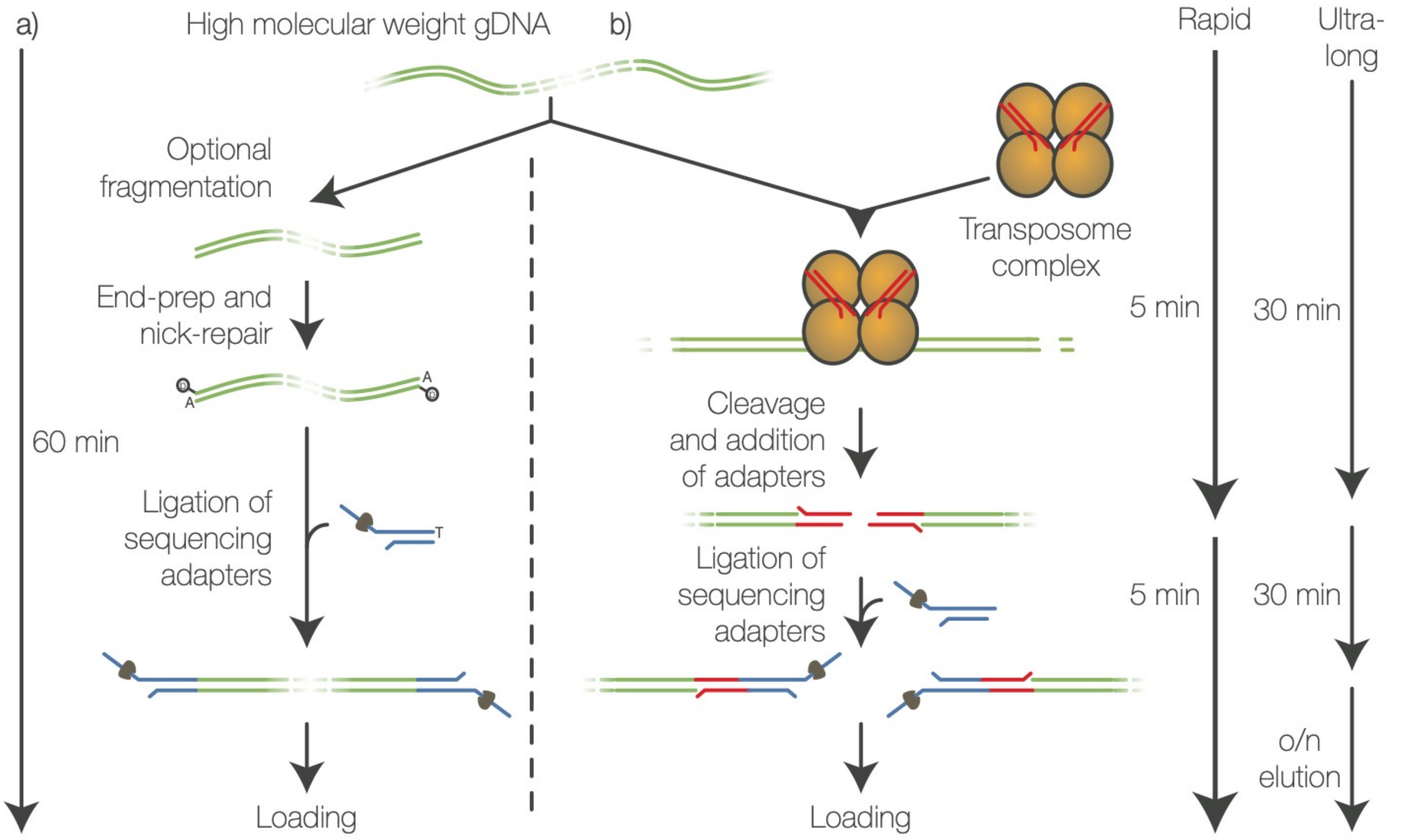
- **Per-base kinetics** encoded as frames in tags **ip & pw** (comma-separated values)
- Mean Polymerase Rate: ~2 bases per second (stochastic)
- Image Capture Rate: 100 frames per second (up to 30 hours)
  - 0.01 seconds per frame (up to 952 frames per base)

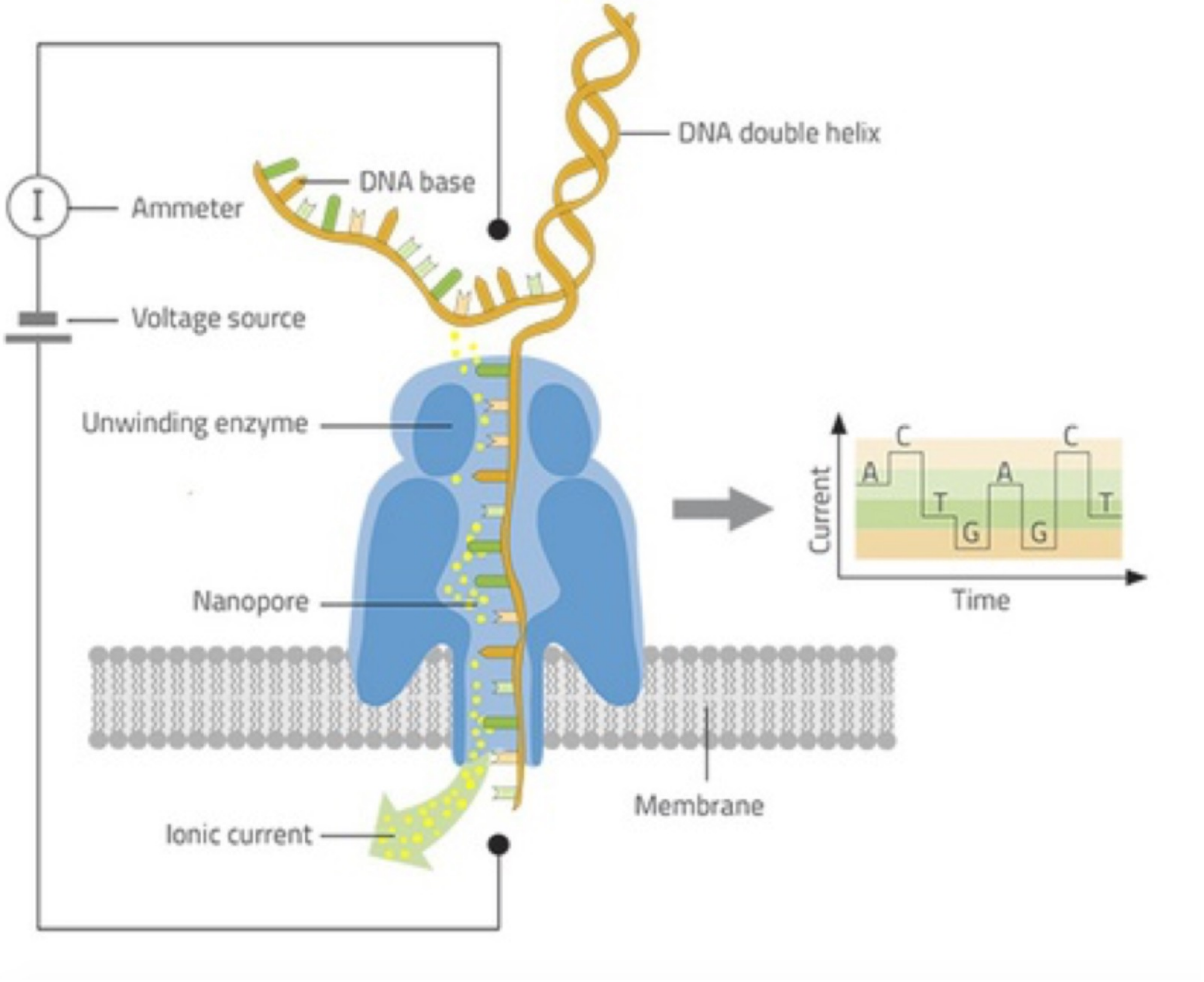
Frames	Encoding
0 .. 63	0, 1, .. 63
64, 66, .. 190	64, 65, .. 127
192, 196 .. 444	128, 129 .. 191
448, 456, .. 952	192, 193 .. 255

Long and Ultra Long read nanopore

- Small
- No imaging- just an ***ammeter***
- Low power can be used in the field (e.g Antartica)
- Can be send in space ☺…
- Ultra long read >100kbp
- Sequence Zika (Brasil), Ebola in (Africa)
- SARS-Cov2 (UK)

Long read nanopore (portable/sma



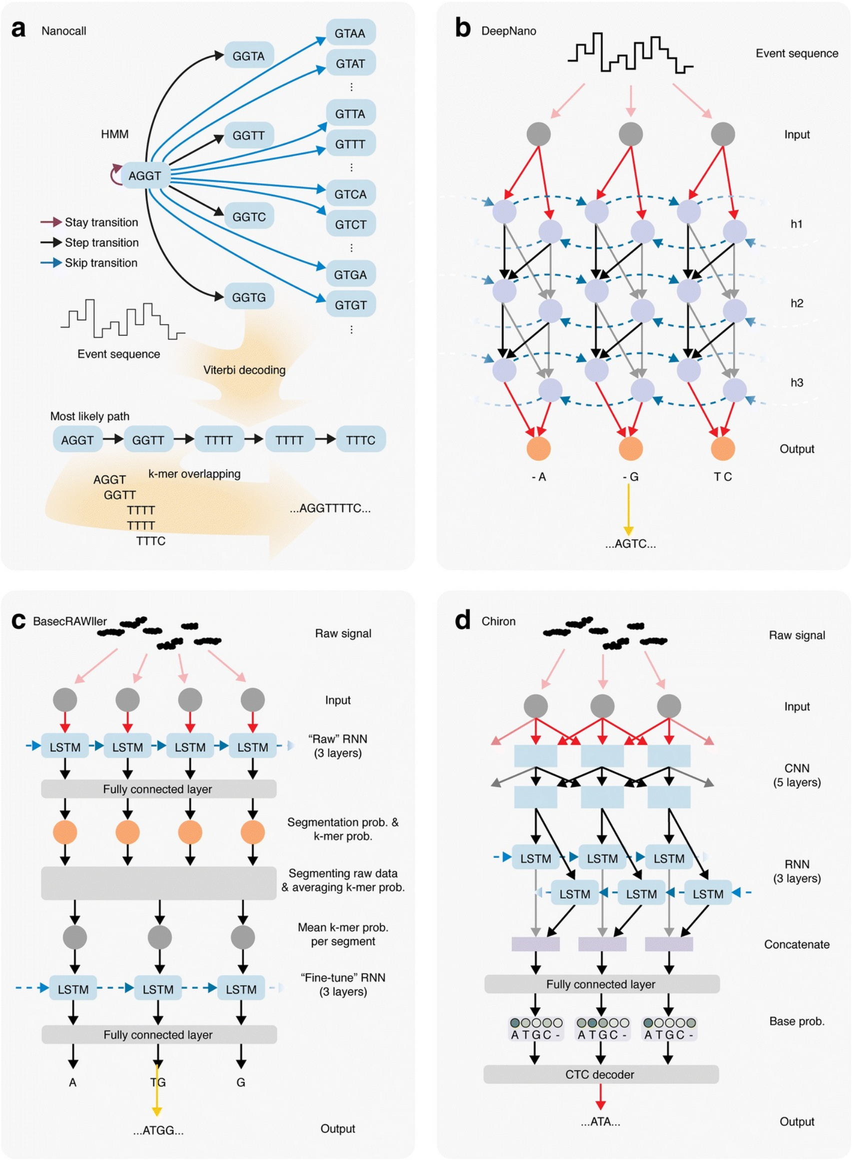


Clarke J. Nature Nanotechnology 4, 265 - 270 (2009)

As the signal is a « trace »

Computational method to call a nucleotide/modified or not is more « tricky » than fluorescence based nucleotide

Latest use of Neural Network for base calling(2018 and onward)

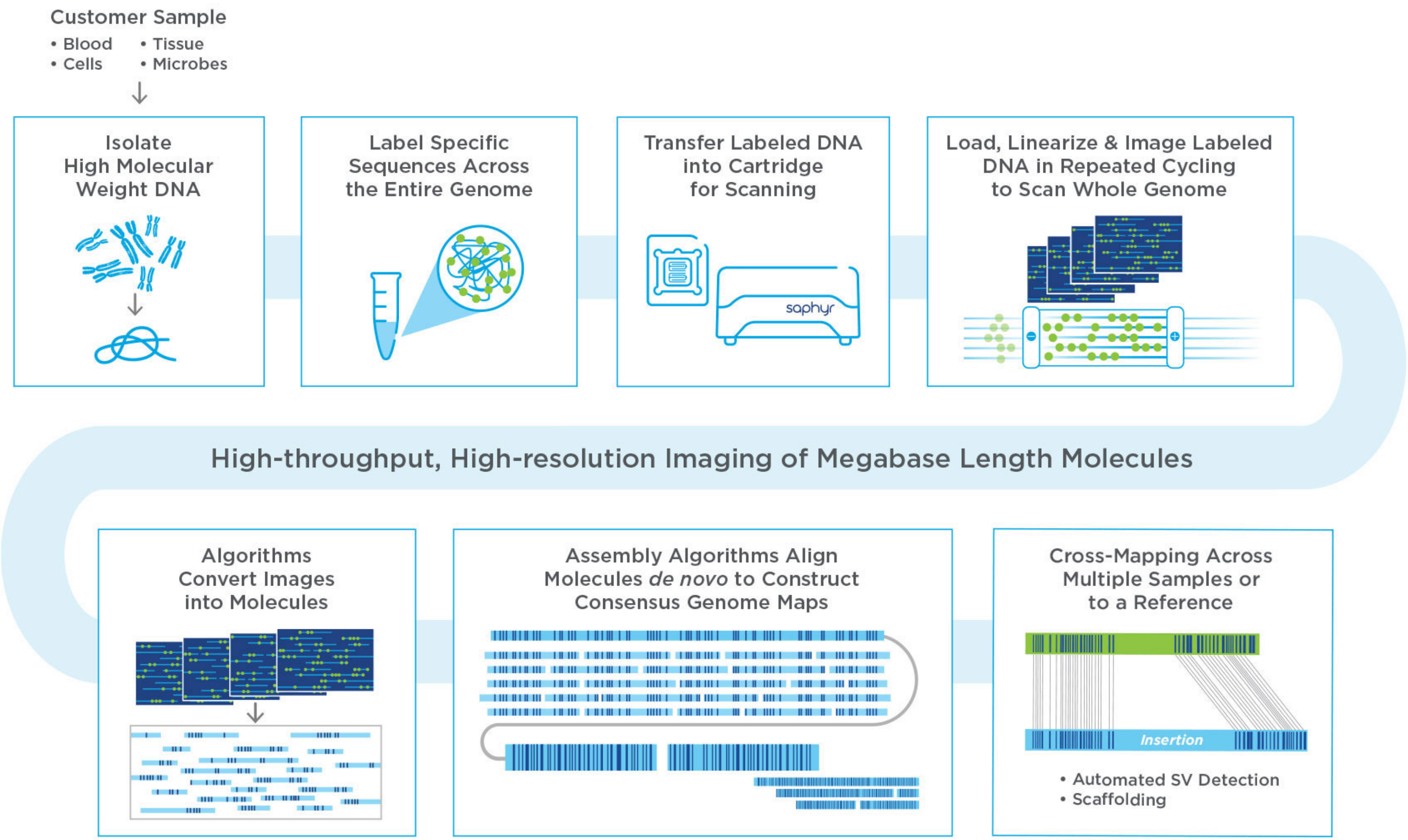


Optical genome mapping (non sequencing based metho

- Requires High molecular weight DNA
- A restriction enzyme (typically EcoRI other can be used too)
- Create a restriction map of a given from bacteria to human

Concept:

Digest *in silico* with EcoRI the human genome reference and compare the theoretical map with bionano observed map



Customer Sample

- Blood
- Cells
- Tissue
- Microbes

Isolate High Molecular Weight DNA

Label Specific Sequences Across the Entire Genome

Transfer Labeled DNA into Cartridge for Scanning

Load, Linearize & Image Labeled DNA in Repeated Cycling to Scan Whole Genome

High-throughput, High-resolution Imaging of Megabase Length Molecules

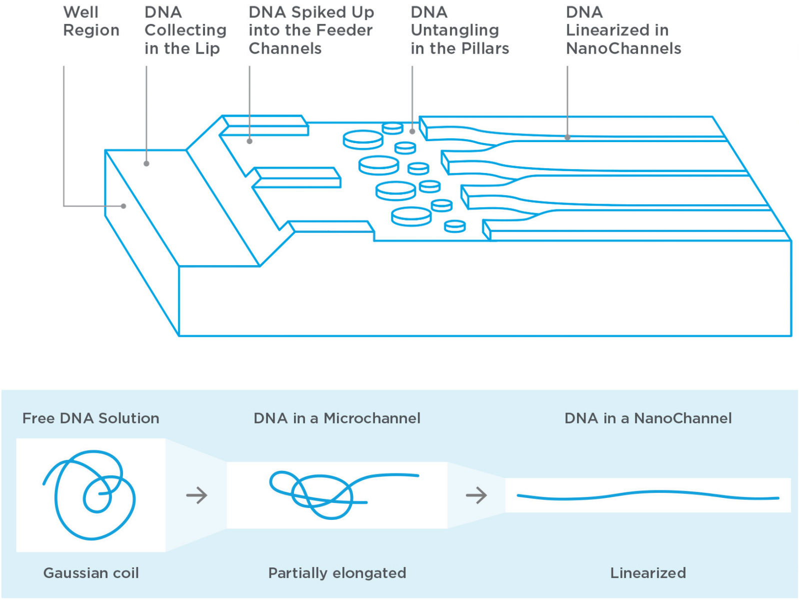
Algorithms Convert Images into Molecules

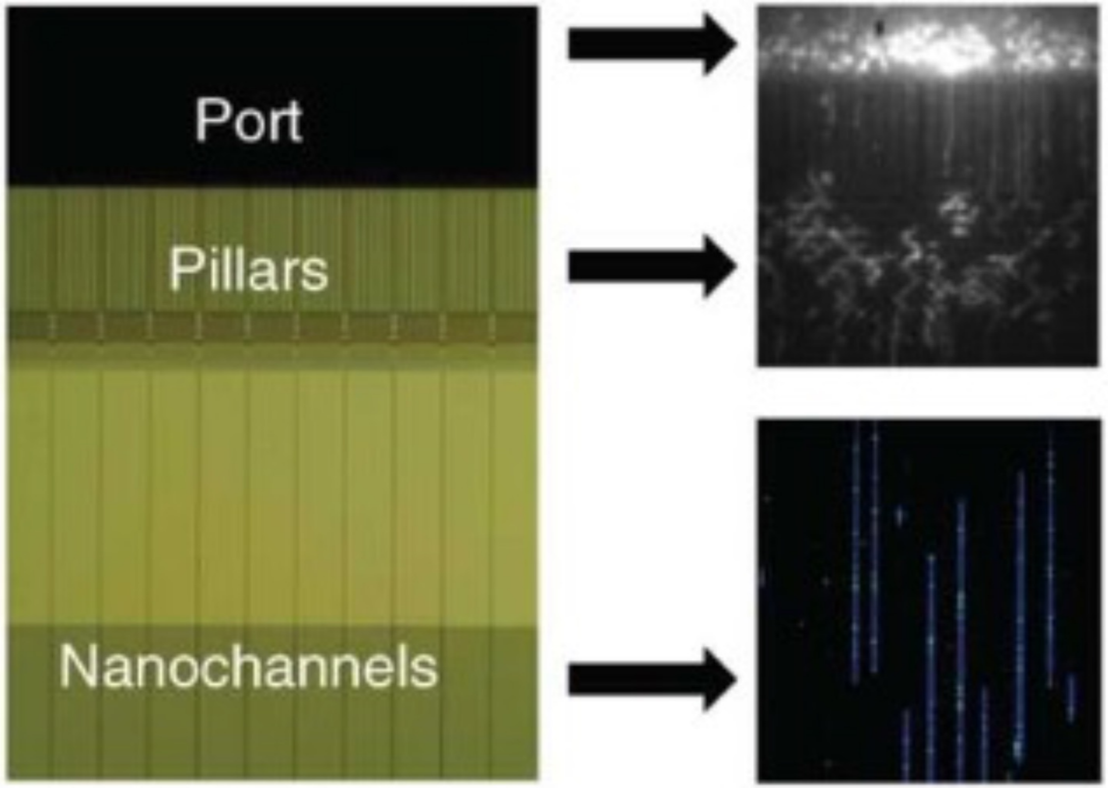
Assembly Algorithms Align Molecules *de novo* to Construct Consensus Genome Maps

Cross-Mapping Across Multiple Samples or to a Reference

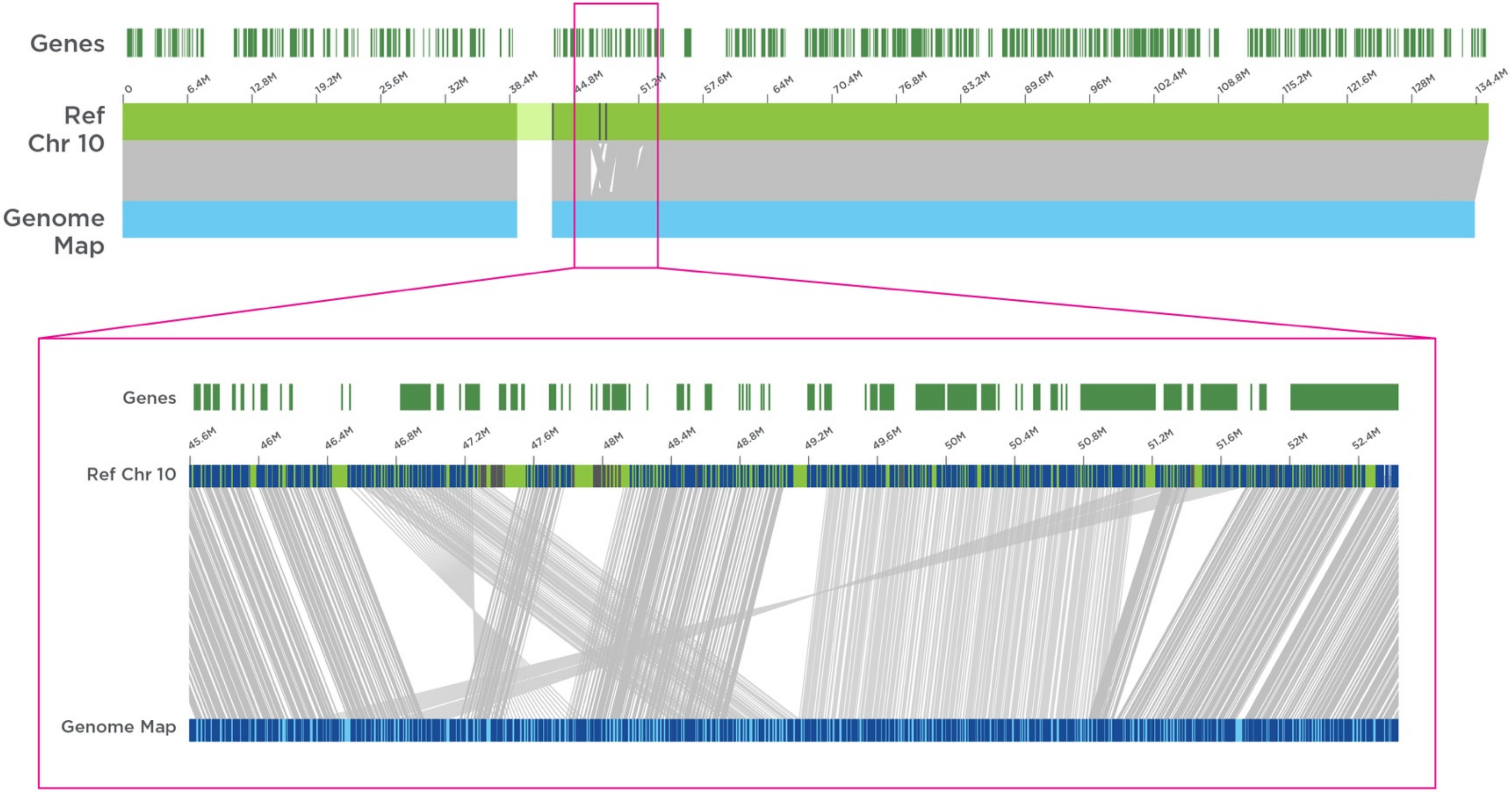
- Automated SV Detection
- Scaffolding

Gel-like to nano-channel threading of DNA –one molecule per channel

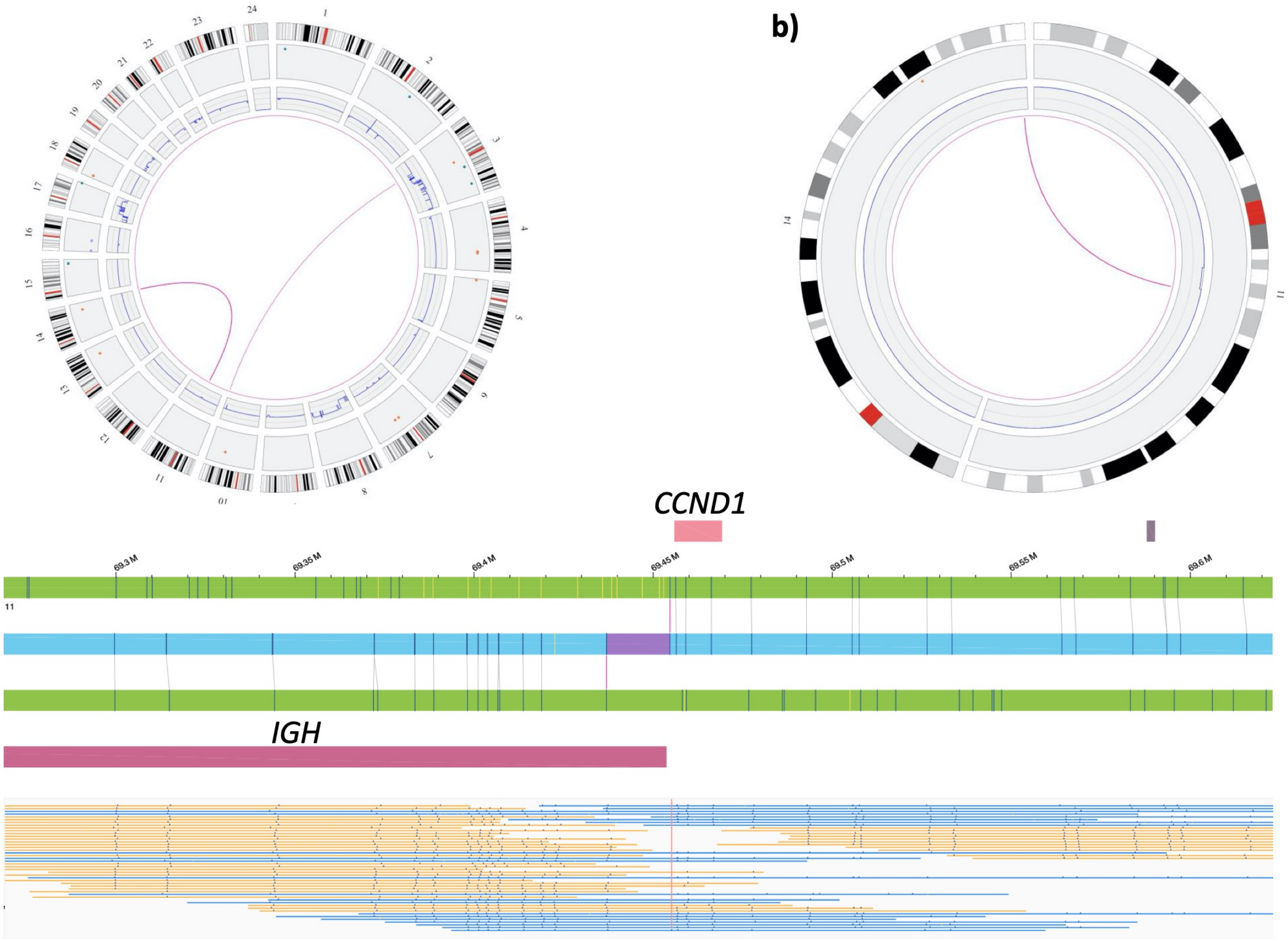

Well Region
DNA Collecting in the Lip
DNA Spiked Up into the Feeder Channels
DNA Untangling in the Pillars
DNA Linearized in NanoChannels
Free DNA Solution
DNA in a Microchannel
DNA in a NanoChannel
Gaussian coil
Partially elongated
Linearized



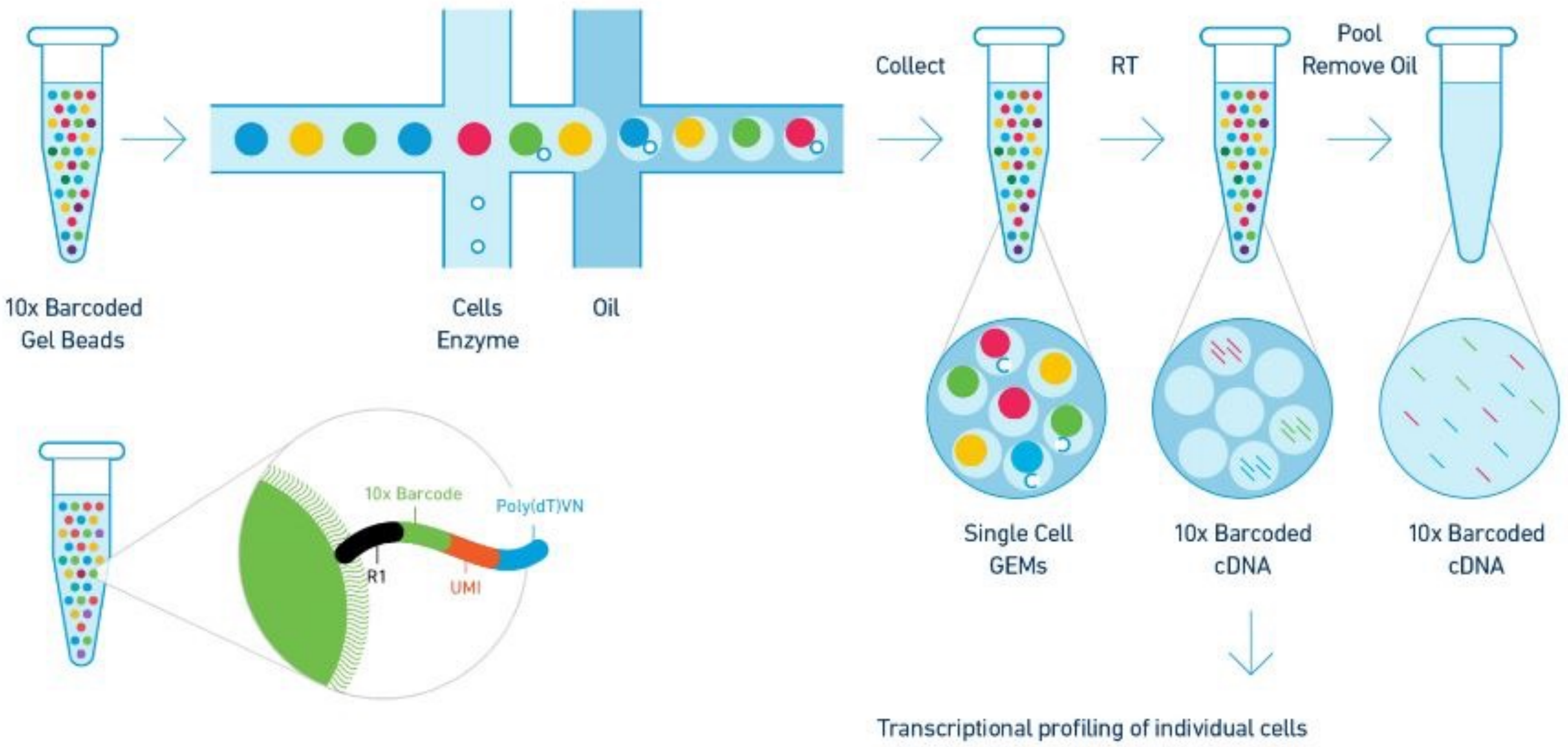
Lam et al. Nature Biotech 2012

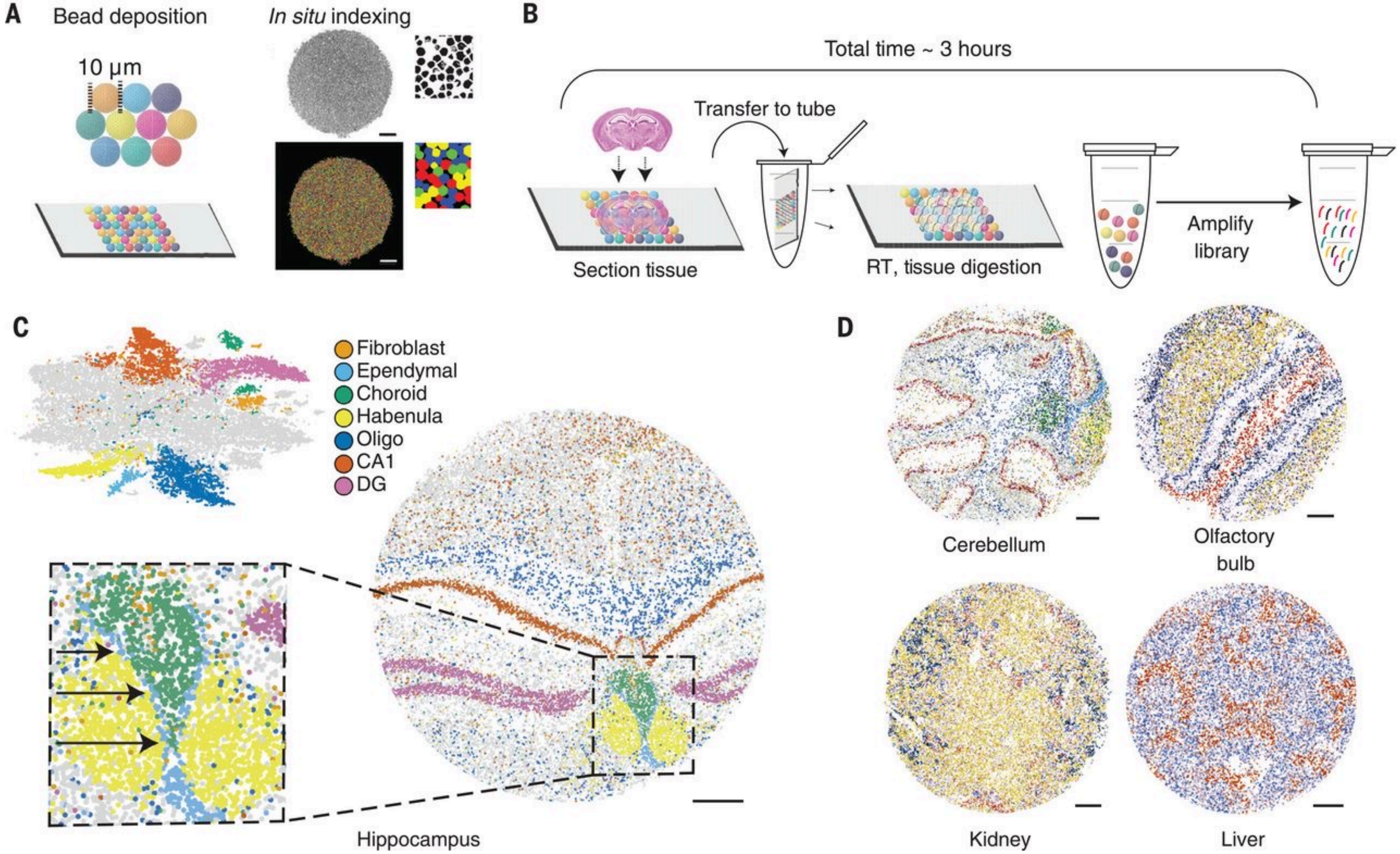

Genes
0
6.4M
12.8M
19.2M
25.6M
32M
38.4M
44.8M
51.2M
57.6M
64M
70.4M
76.8M
83.2M
89.6M
96M
102.4M
108.8M
115.2M
121.6M
128M
134.4M
Ref
Chr 10
Genome
Map
Genes
45.6M
46M
46.4M
46.8M
47.2M
47.6M
48M
48.4M
48.8M
49.2M
49.6M
50M
50.4M
50.8M
51.2M
51.6M
52M
52.4M
Ref Chr 10
Genome Map

Balanced
Chromosomal
Translocation



Single cell preparation (non adherent cells)





Rodriguez et al. Science, 2019, Slide-Seq

Mixing technologies the only way forward

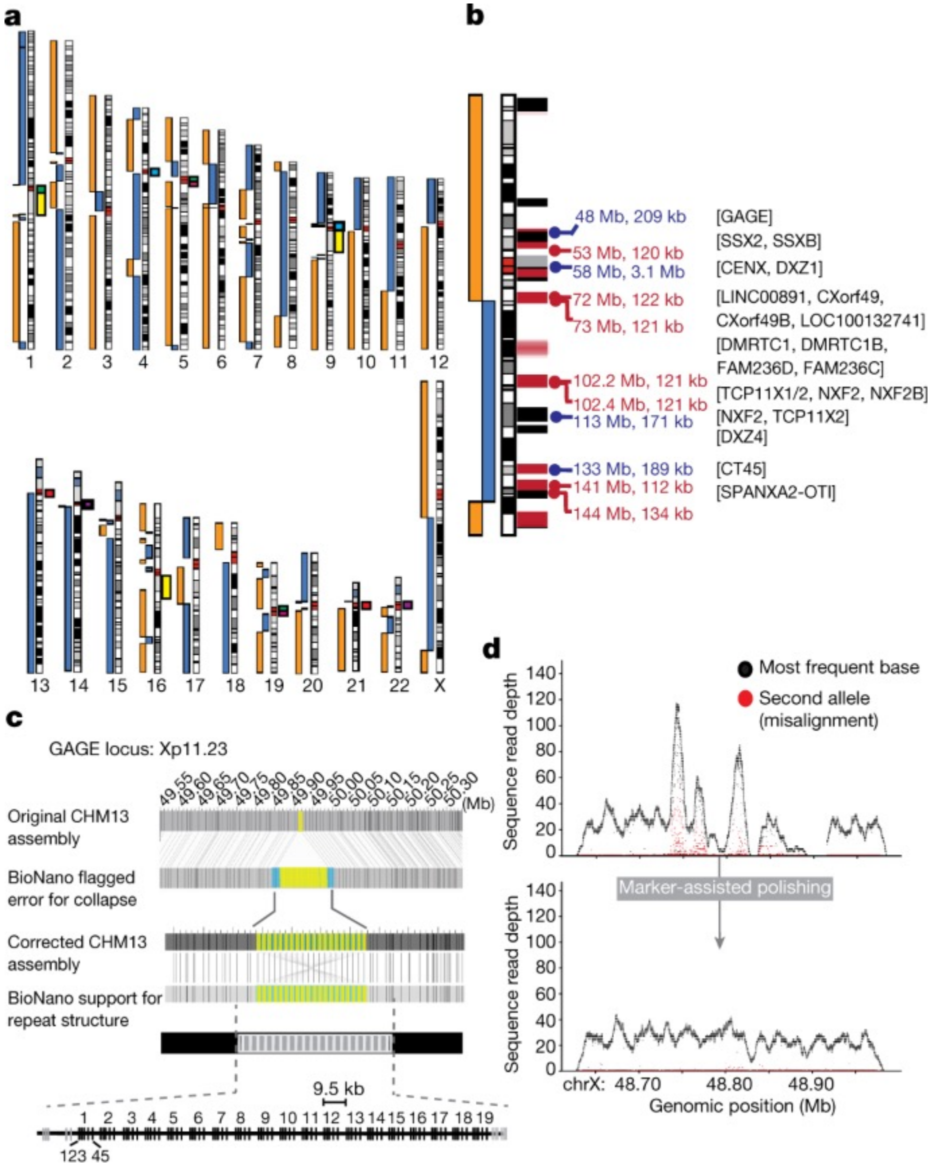
Producing a telomere

To telomere

Human genome is the next frontier

Example here the X chromosome

Miga et al. 2020 Nature : Telomere to Telomere



Sequencing technology summary

- Short – long read – optical mapping – single cell biology are ways to **understand better biology**, to map structure and capture the complexity of life

- It is use in research and more and more clinical setting and can potentially with time become part of day to day life (monitoring, alerting)

- Use to follow pathogene is one of the real life example

Non invasive

Prenatal

Diagnostic

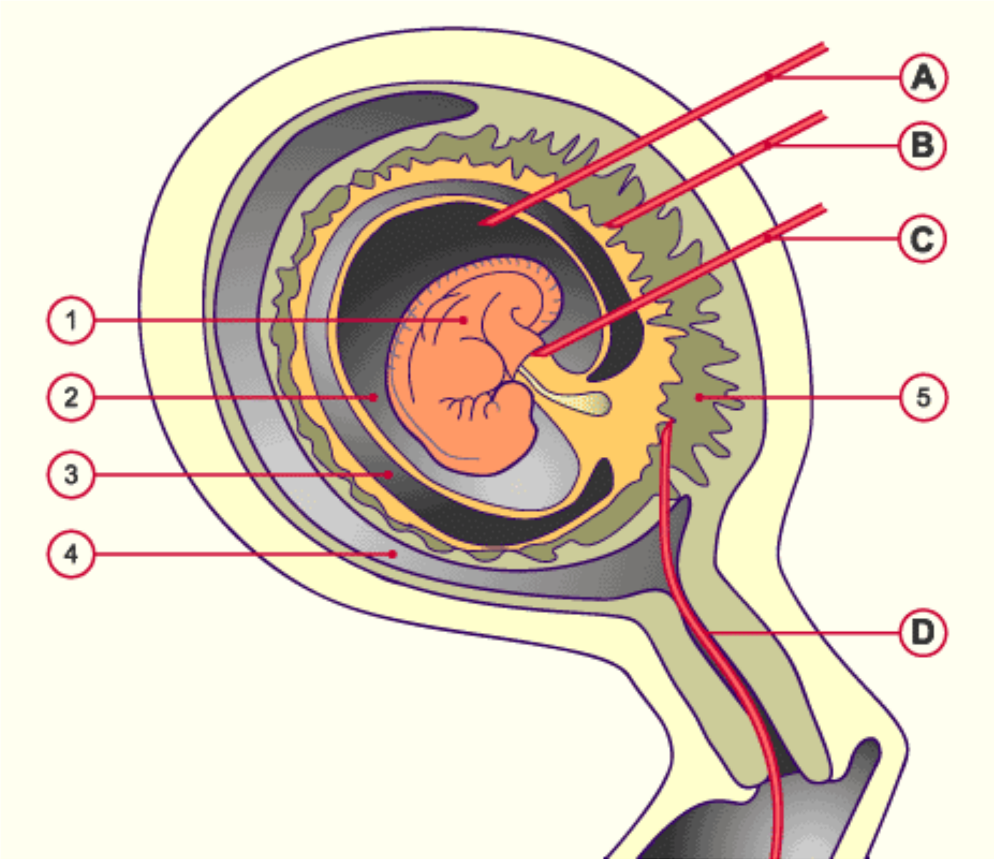
(NIPT)

Disruptive nature of sequencing

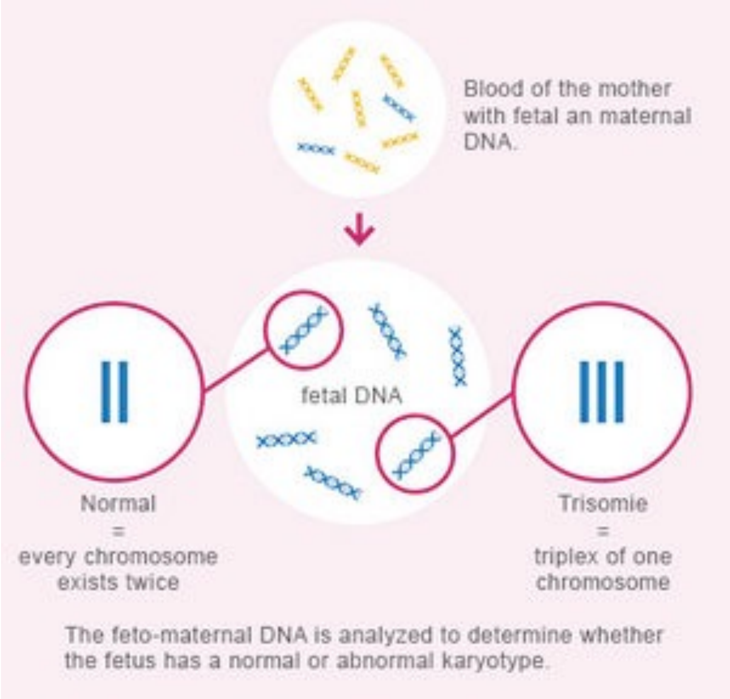
Invasive detection of T21

1. Embryo
2. Amniotic cavity
3. Chorion cavity
4. Uterine cavity
5. Chorion frondosum

A. Amniocentesis
B. Chorion biopsy
C. Umbilical blood sampling
D. Transvaginal chorion biopsy



Non-invasive detection

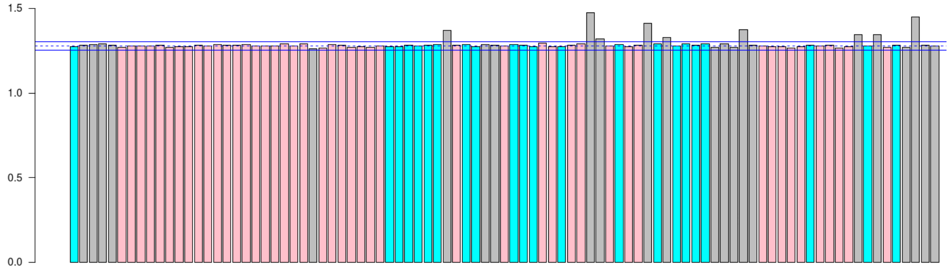

Blood of the mother with fetal an maternal DNA.
fetal DNA
Normal = every chromosome exists twice
Trisomie = triplex of one chromosome

The feto-maternal DNA is analyzed to determine whether the fetus has a normal or abnormal karyotype.

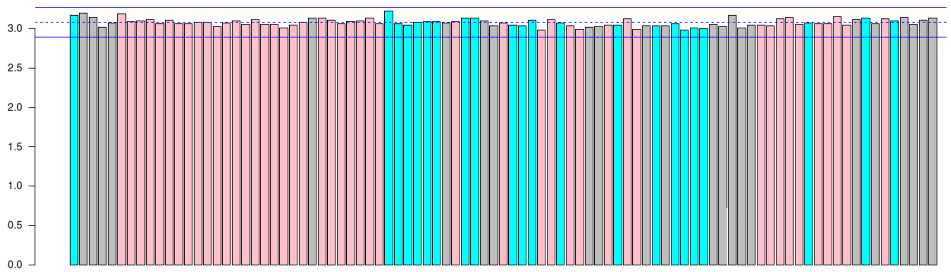
Evaluation of the state of the art

Our findings showed that although T21 is relatively easy to detect with published algorithms, other aneuploidies are more difficult to detect in a robust way

T21



T13



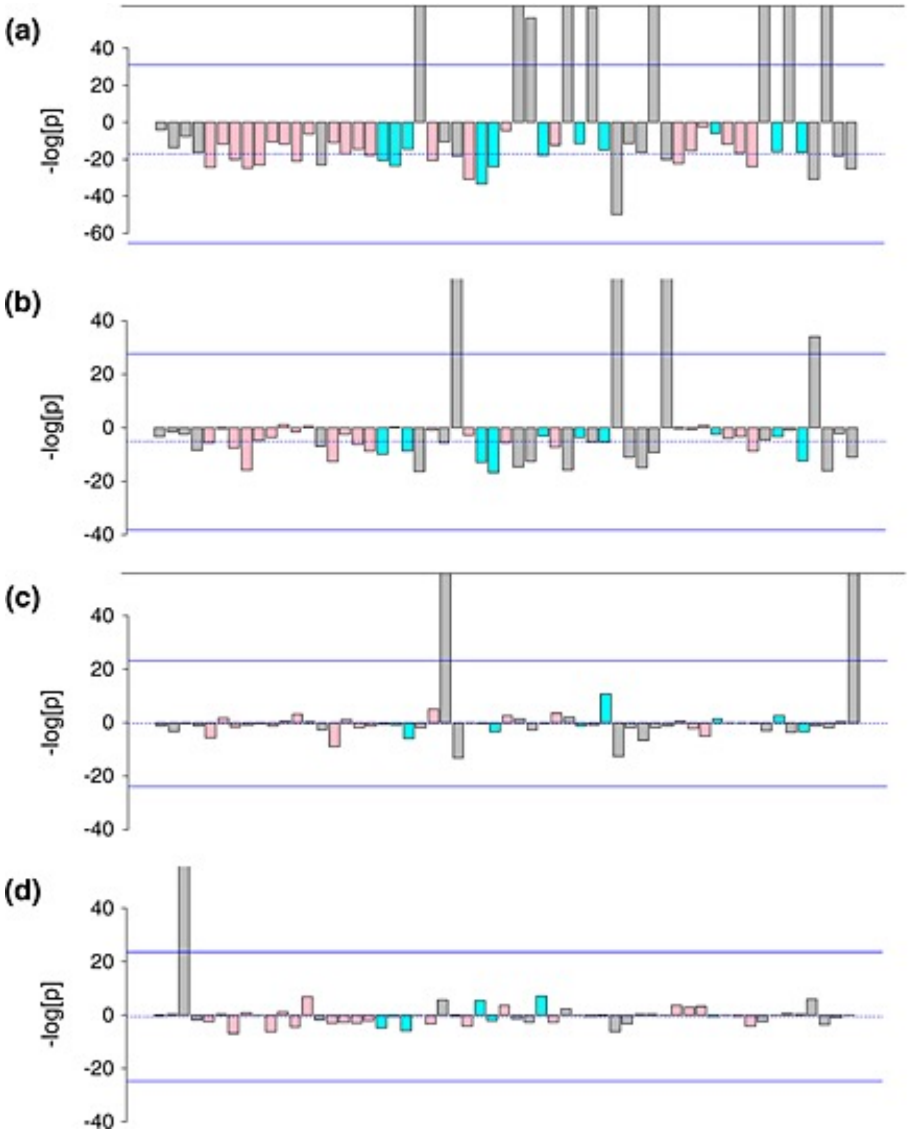
Novel algorithm development

Chr 21

Chr 18

Chr 13

Chr 22

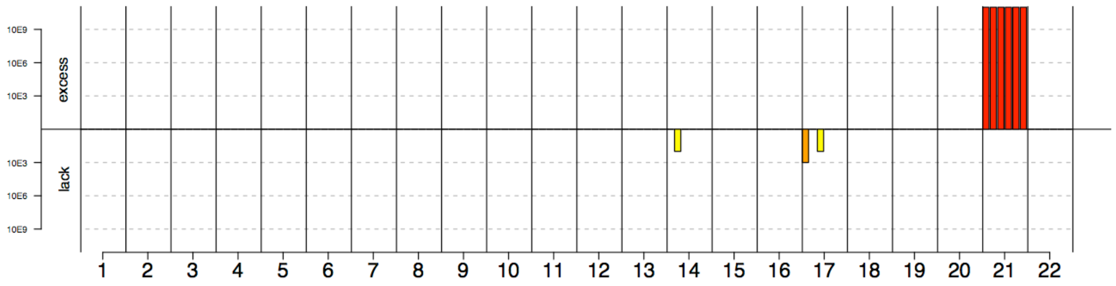


Double Blind -Clinical Trial Results

(Guex N. et al, *Prenat Diagn*. 2013 Jul;33(7):707-10)

Aneuploidy	Sensitivity	Specificity
Trisomy 21 (*n* = 39)	39 (100%, 95% CI 88.8–100)	237/237 (100%, 95%CI 98.0–100)
Trisomy 18 (*n* = 24)	23 (95.8%, 95%CI 76.8–99.7)	252/252 (100%, 95%CI 97.0–100)
Trisomy 13 (*n* = 15)	15 (100%, 95%CI 74.6–100)	261/261 (100%, 95%CI 98.1–100)
Trisomy 16 (*n* = 1)	1 (100%, 95%CI 5.4–100)	275/275 (100%, 95%CI 98.2–100)
Trisomy 22 (*n* = 2)	2 (100%, 95%CI 19.7–100)	274/274 (100%, 95%CI 98.2–100)
45,X (*n* = 15)	15 (100%, 95%CI 74.6–100)	261/261 (100%, 95%CI 98.1–100)
47,XXX (*n* = 5)	5 (100%, 95%CI 46.2–100)	271/271 (100%, 95%CI 98.2–100)

Trisomy 21 example



Very High T21 warning

When no warnings are reported, the fetus is assumed to be normal.

Think about the situation where the fraction of circulating fetal DNA is very low…

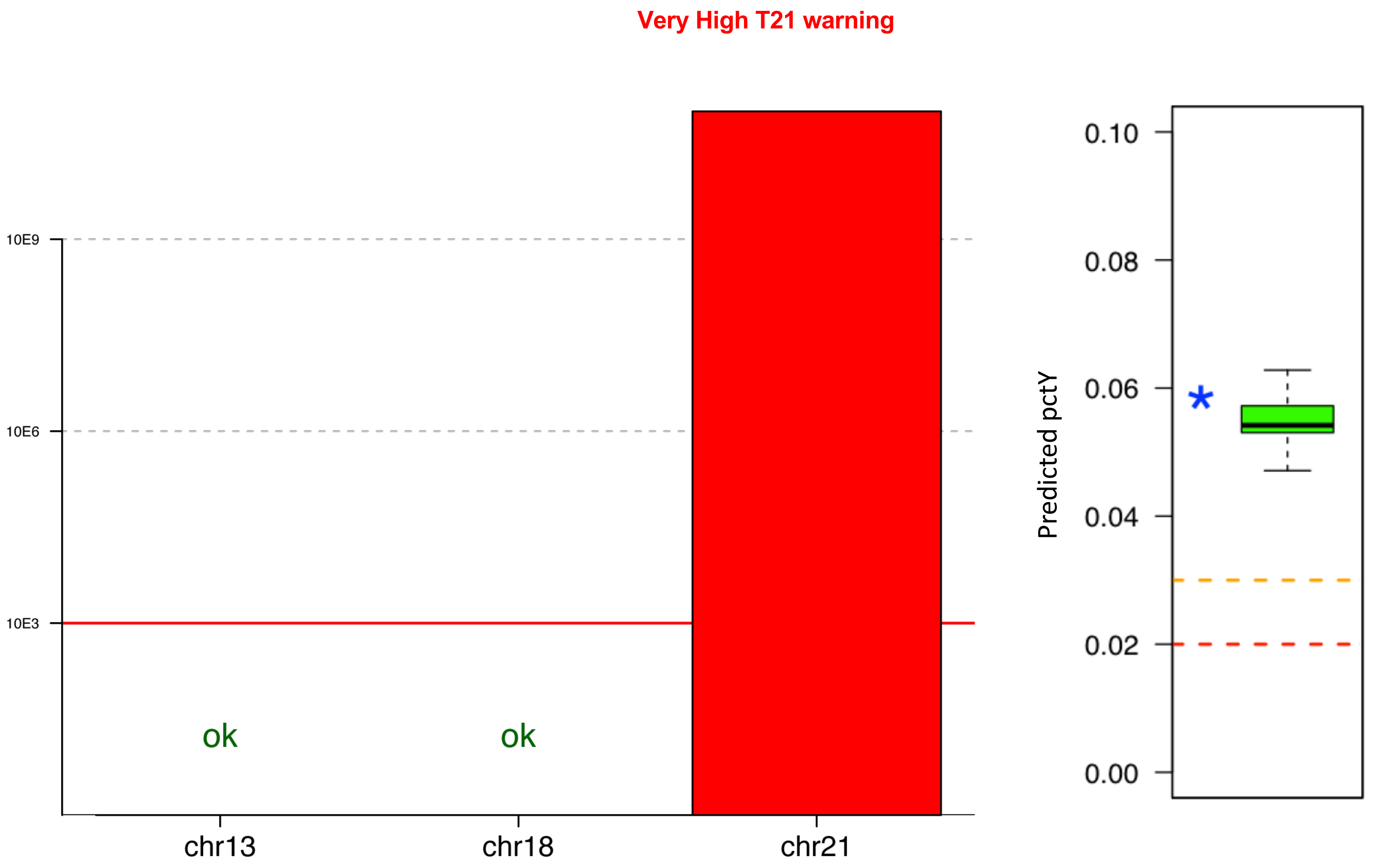
Evaluation of Fetal Fraction

- Percentage of total tags mapping on chrY is theoretically proportional to the fetal fraction (for males)
- Equivalent measure does not exist for females

Goal:

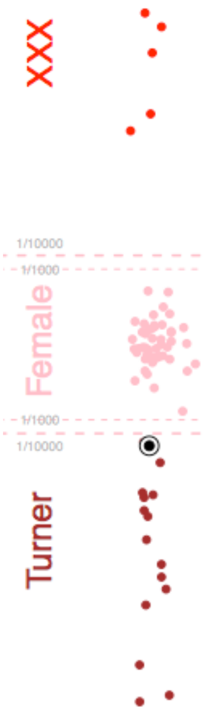
- Identify a “universal” measure derived from autosomes that reflects the fetal fraction

known T21 male with high Fetal Fraction

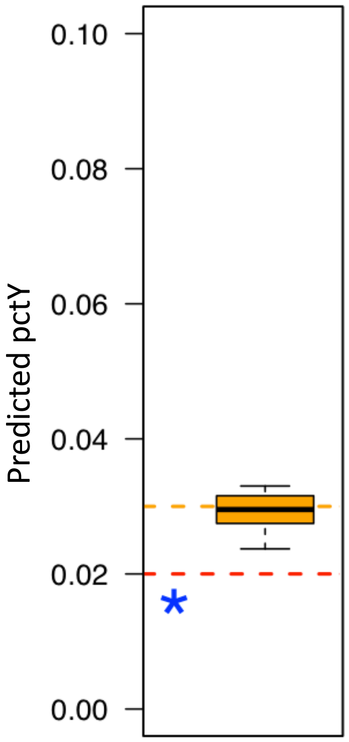


Example: female, borderline 46,XX – 45,X

blood sample taken at **11 week** of pregnancy

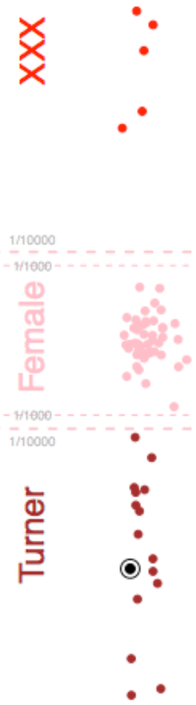


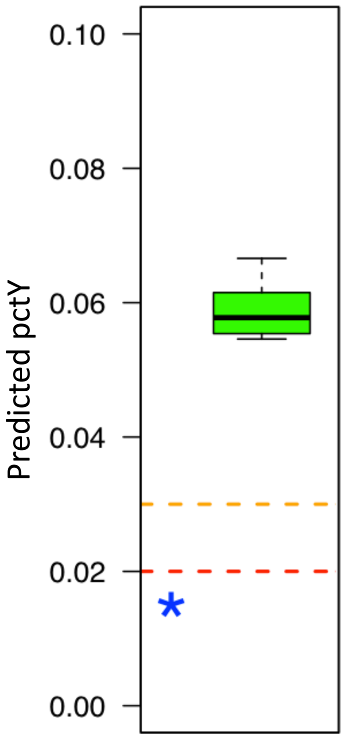
Low Monosomy X warning



Same pregnancy

blood sample taken at **13 weeks** of pregnancy





High Monosomy X warning

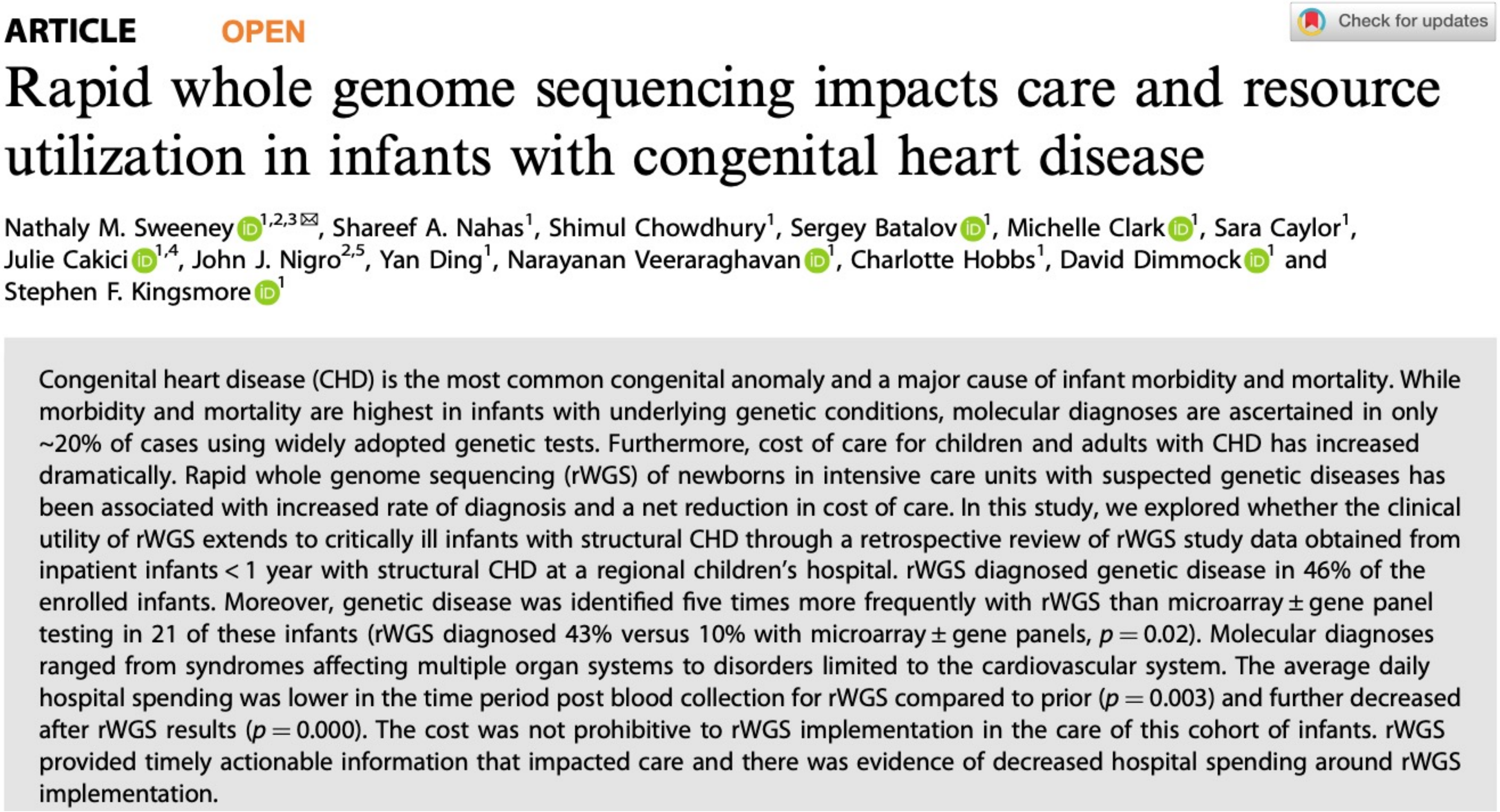
PrenDia – Prenatal Trisomy Detection: development timelines

- June 2012 – approached by medisupport
- October 2012 – calibrate new method on known samples
- December 2012 – validate by double blind experiment
- March 2013 – announce new test
- April 2013 – CE marking
- May 2013 – on the market (PrenDia by GeneSupport)
- June 2014 – fetal Fraction detection in production
- July 2014 – Reimbursed by the LaMal

Rapid
Whole Genome Sequence in Intensive Care Unit

(Genetic Disease/Rare Variants)

And
Pediatric Oncology



ARTICLE OPEN

Rapid whole genome sequencing impacts care and resource utilization in infants with congenital heart disease

Nathaly M. Sweeney[1,2,3]✉, Shareef A. Nahas[1], Shimul Chowdhury[1], Sergey Batalov[1], Michelle Clark[1], Sara Caylor[1], Julie Cakici[1,4], John J. Nigro[2,5], Yan Ding[1], Narayanan Veeraraghavan[1], Charlotte Hobbs[1], David Dimmock[1] and Stephen F. Kingsmore[1]

Congenital heart disease (CHD) is the most common congenital anomaly and a major cause of infant morbidity and mortality. While morbidity and mortality are highest in infants with underlying genetic conditions, molecular diagnoses are ascertained in only ~20% of cases using widely adopted genetic tests. Furthermore, cost of care for children and adults with CHD has increased dramatically. Rapid whole genome sequencing (rWGS) of newborns in intensive care units with suspected genetic diseases has been associated with increased rate of diagnosis and a net reduction in cost of care. In this study, we explored whether the clinical utility of rWGS extends to critically ill infants with structural CHD through a retrospective review of rWGS study data obtained from inpatient infants < 1 year with structural CHD at a regional children's hospital. rWGS diagnosed genetic disease in 46% of the enrolled infants. Moreover, genetic disease was identified five times more frequently with rWGS than microarray ± gene panel testing in 21 of these infants (rWGS diagnosed 43% versus 10% with microarray ± gene panels, $p = 0.02$). Molecular diagnoses ranged from syndromes affecting multiple organ systems to disorders limited to the cardiovascular system. The average daily hospital spending was lower in the time period post blood collection for rWGS compared to prior ($p = 0.003$) and further decreased after rWGS results ($p = 0.000$). The cost was not prohibitive to rWGS implementation in the care of this cohort of infants. rWGS provided timely actionable information that impacted care and there was evidence of decreased hospital spending around rWGS implementation.

https://www.nature.com/articles/s41525-021-00192-x.pdf

Service development: Rapid WGS in critically ill infants

"I have never seen a diagnostic tool that's made such huge impact in intensive care medicine in all my years of practice"

Mario Rojas, MD
NICU Medical Director Valley Children's Hospital

project baby bear | Rady Children's Hospital San Diego

Nicklaus Children's Hospital | Project Baby Manatee

project baby deer

Australian Genomics

Karolinska Institutet

EXETER CLINICAL LABORATORY INTERNATIONAL

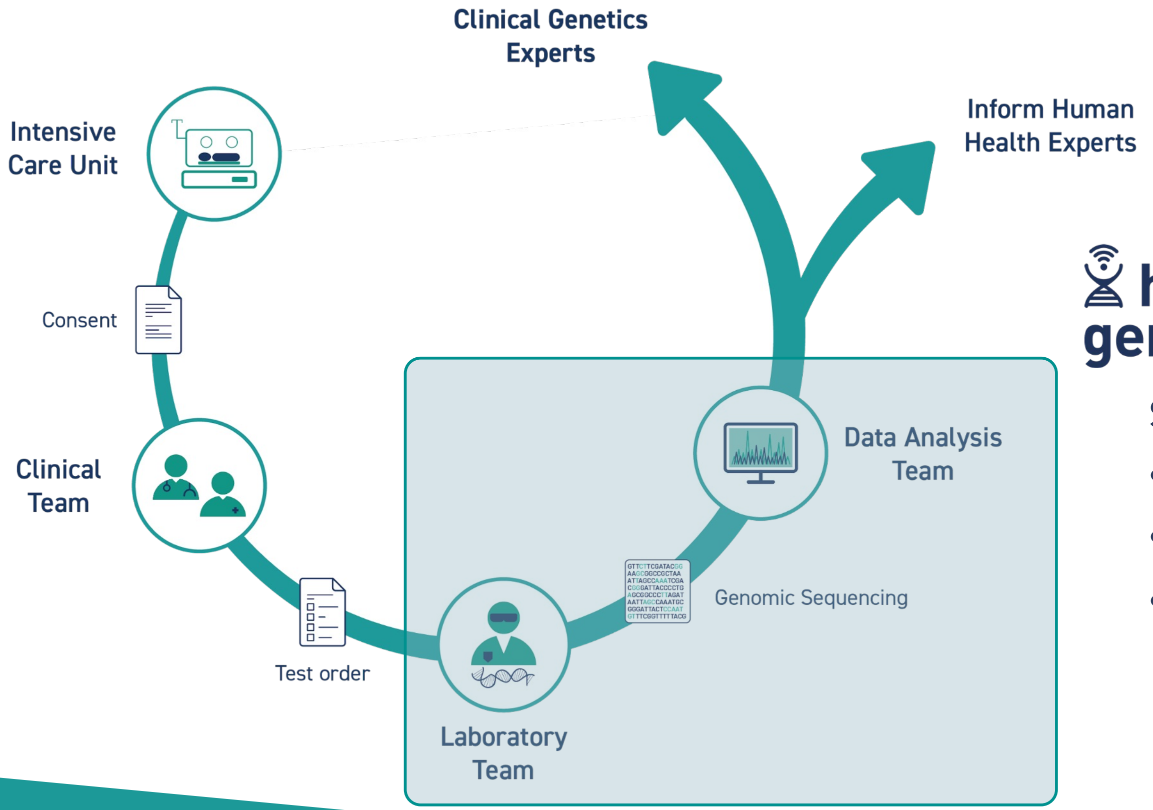
Figures and data from The Project Baby Bear Final Report;
Rady Children's Hospital (CA, USA) 2020

Predicted number of PICU patients for rWGS

	Estonia	Switzerland
Population	1.329 Mio	8.698 Mio
Number of live births (2021)	13,272	89,644
Number of patients in Level III PICUs (per year)	520-550	3,406-3,602
Newborns among the PICU patients	Approx. 210-230 (40-45%)	Approx. 1,375-1,507
Newborns with suspected genetic disease in PICU	Approx. 88 (38-41%)	Approx. 576
Number of newborn PICU patients considered for rWGS (per year)	50-75	328-491

Data kindly shared by Prof. Tuuli Metsvaht, Pediatric Intensive Care Unit, Tartu University Hospital, Estonia

Automated and standardized rapid WGS workflow



health2030 genome center

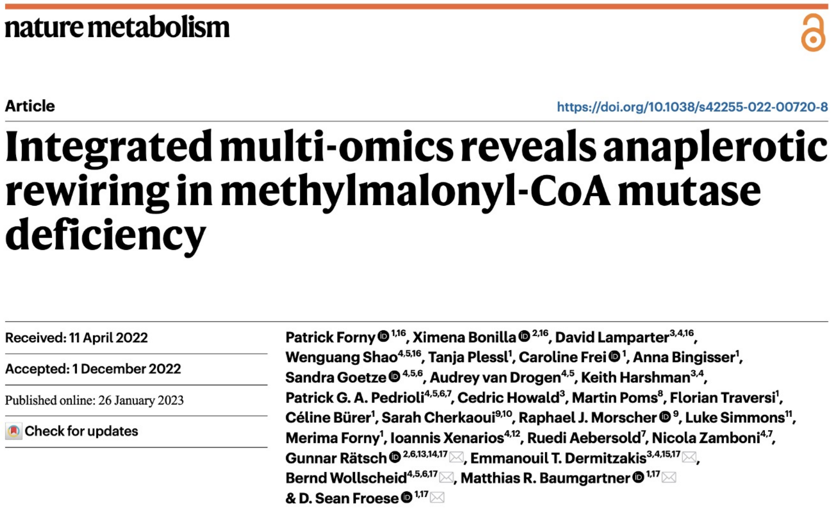
Streamlining:

- Sequencing
- First line analytics
- Initial variant filtering

Workflow adapted from Zornitza Stark at the GA4GH Plenary 2021

Development and validation of the rWGS workflow

UNIVERSITÄTS-
KINDERSPITAL
ZÜRICH



nature metabolism

Article

https://doi.org/10.1038/s42255-022-00720-8

Integrated multi-omics reveals anaplerotic rewiring in methylmalonyl-CoA mutase deficiency

Received: 11 April 2022

Accepted: 1 December 2022

Published online: 26 January 2023

Check for updates

Patrick Forny[1,16], Ximena Bonilla[2,16], David Lamparter[3,4,16], Wenguang Shao[4,5,16], Tanja Plessl[1], Caroline Frei[1], Anna Bingisser[1], Sandra Goetze[4,5,6], Audrey van Drogen[4,5], Keith Harshman[3,4], Patrick G. A. Pedrioli[4,5,6,7], Cedric Howald[3], Martin Poms[8], Florian Traversi[1], Céline Bürer[1], Sarah Cherkaoui[9,10], Raphael J. Morscher[9], Luke Simmons[11], Merima Forny[1], Ioannis Xenarios[4,12], Ruedi Aebersold[7], Nicola Zamboni[4,7], Gunnar Rätsch[2,6,13,14,17] ✉, Emmanouil T. Dermitzakis[3,4,15,17] ✉, Bernd Wollscheid[4,5,6,17] ✉, Matthias R. Baumgartner[1,17] ✉ & D. Sean Froese[1,17] ✉

Singleton analysis of three unrelated patients with known MMA variants

- Filters: MAF ≤ 0.1; Consequences
- HP:0012120 Methylmalonic aciduria
- Genomics England PanelApp:
  - Inborn errors of metabolism
  - Sever Paediatric Disorders (only Green genes, high evidence)

Results:

All expected genes found in the 3 samples by Congenica AI

- MMA017: *MUT* (ranked #1)
- MMA185: *SUCLA2* (ranked #1)
- MMA196: *ACSF3* (ranked #3)

Congenica AI prioritizes expected variants and drastically reduces the number of variants to interpret

Development and validation of the rWGS workflow

UNIVERSITÄTS-KINDERSPITAL ZÜRICH

Trio analysis of three PICU patients with unknown or blinded variants

Prof. Dr. Matthias Baumgartner

Prof. Dr. Johannes Häberle

- Filters: MAF ≤ 0.1; Consequences
- Genomics England PanelApp:
  - Sever Paediatric Disorders (only Green genes, high evidence)

	Congenica AI filtered SNVs/genes	WGS at the Genome Center	Previously performed WES
Patient 1	16/14	**Clinically relevant variant**	No variants to report
Patient 2	17/15	**Clinically relevant variant**	Clinically relevant variant
Patient 3	10/10	No variants to report	No variants to report

Swiss Ped Health

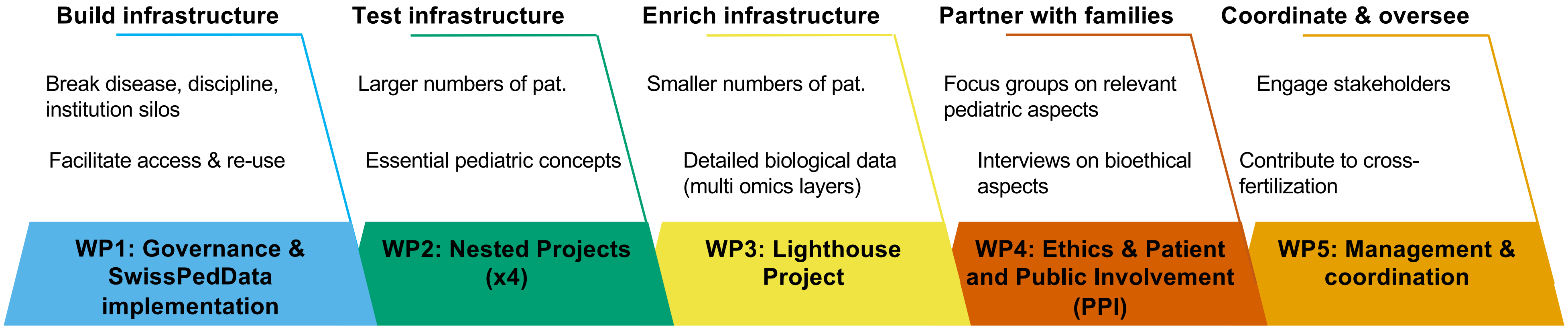
A Pediatric National Data Stream

SwissPedHealth

UNIVERSITÄTS-KINDERSPITAL ZÜRICH

Strategic Focus Area
Personalized Health
and Related Technologies

SPHN



Build infrastructure

Break disease, discipline, institution silos

Facilitate access & re-use

WP1: Governance & SwissPedData implementation

Test infrastructure

Larger numbers of pat.

Essential pediatric concepts

WP2: Nested Projects (x4)

Enrich infrastructure

Smaller numbers of pat.

Detailed biological data (multi omics layers)

WP3: Lighthouse Project

Partner with families

Focus groups on relevant pediatric aspects

Interviews on bioethical aspects

WP4: Ethics & Patient and Public Involvement (PPI)

Coordinate & oversee

Engage stakeholders

Contribute to cross-fertilization

WP5: Management & coordination

WP Leads:

Julia Bielicki,
Claudia Kuehni

Claudia Kuehni,
Julia Bielicki

Matthias Baumgartner,
Jacques Fellay

Klara Posfay-Barbe,
Effy Vayena

Luregn Schlapbach,
Julia Vogt

We aim to make routine data from children’s hospitals interoperable, harmonized and quality-controlled using a **modular and scalable approach**

Slide credit: Sean Froese & Rebeca Mozun

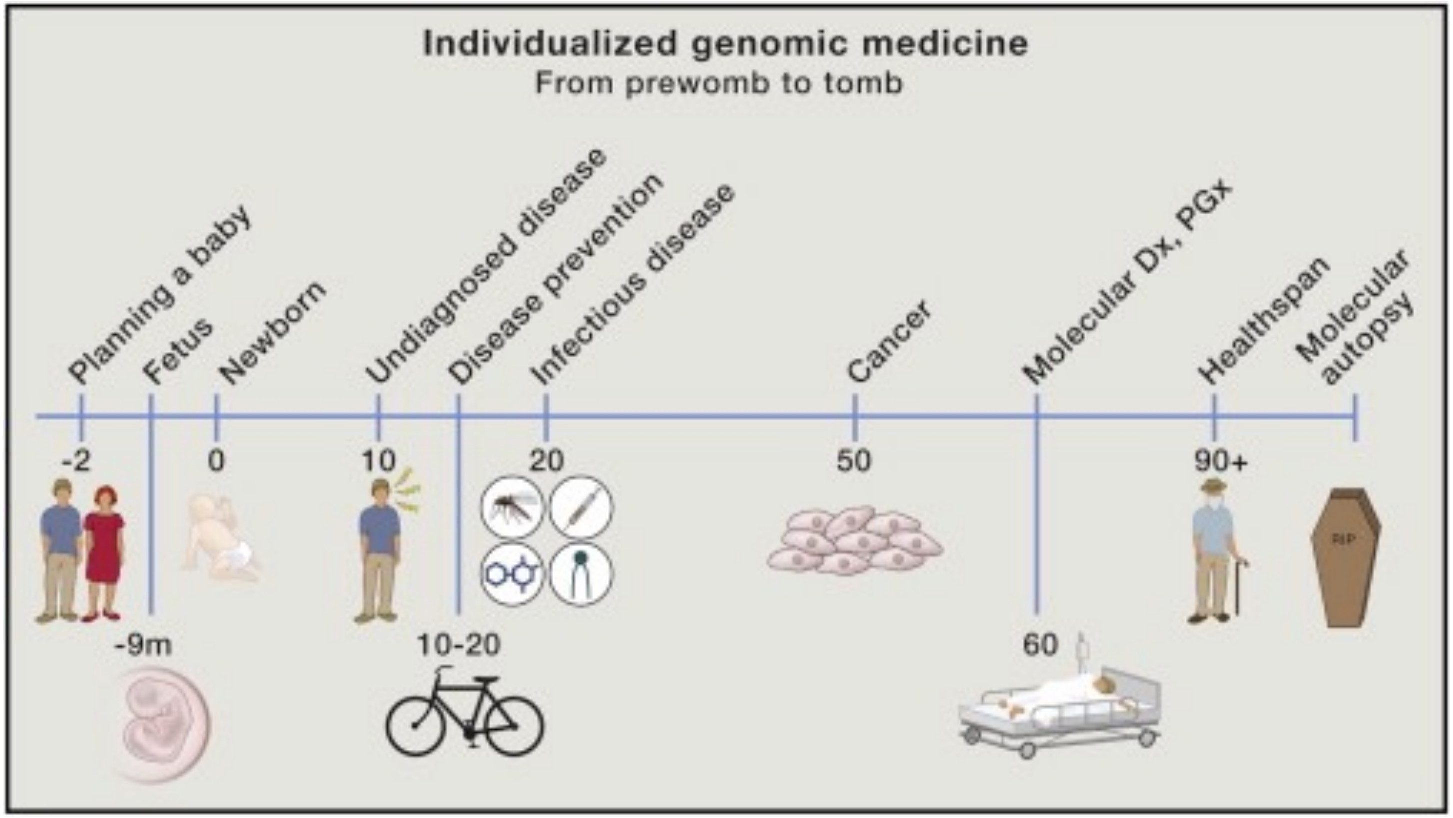
What is the future

# of high throughput

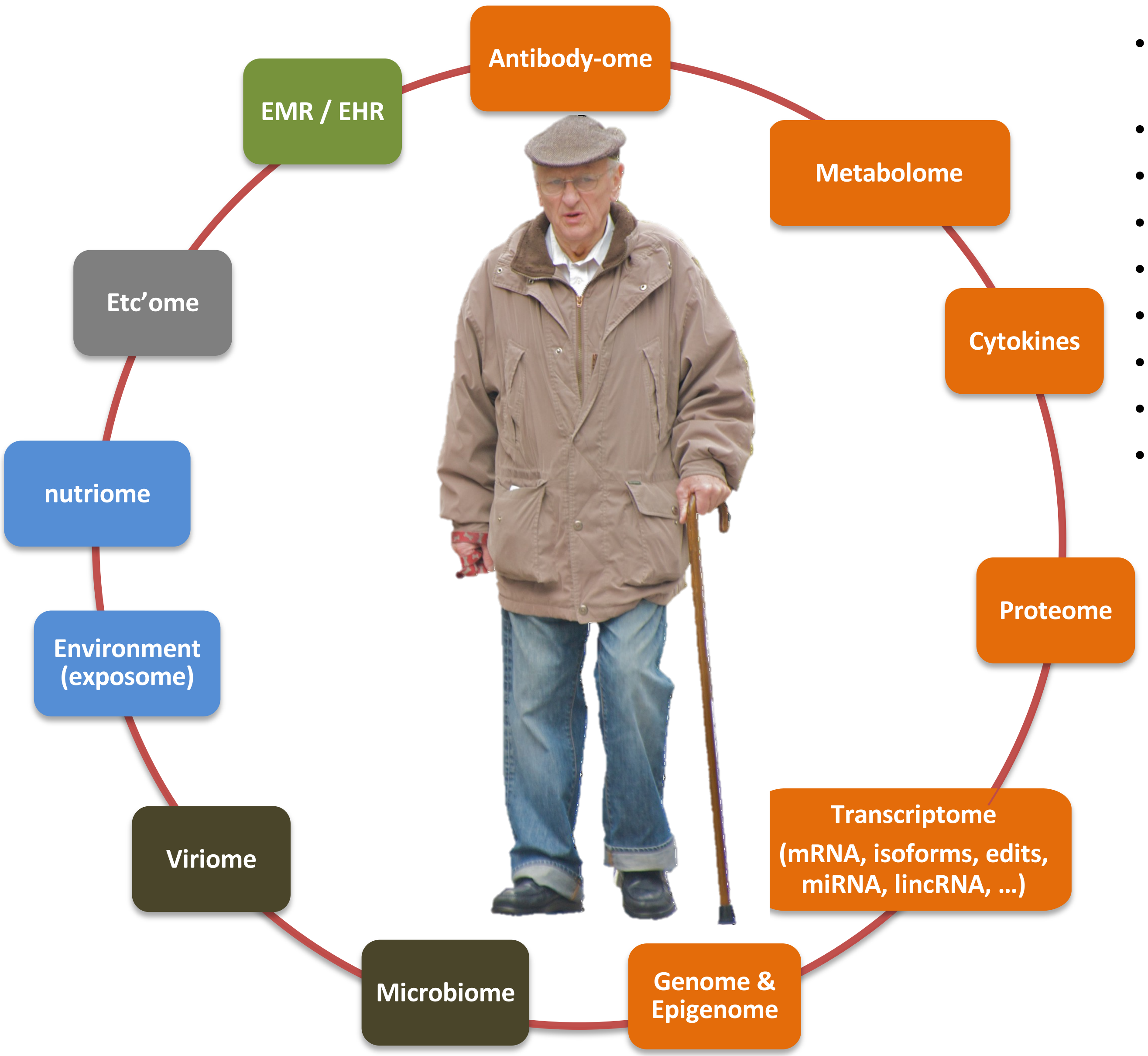
Bioinformatics and medecine

Sequencing over the entire life

- Cell free DNA can be used as indicator of what type of treatment or disease the person will have



Repertoire of (*longitudinal*) ‘omics data available



Antibody-ome

Metabolome

Cytokines

Proteome

Transcriptome (mRNA, isoforms, edits, miRNA, lincRNA, …)

Genome & Epigenome

Microbiome

Viriome

Environment (exposome)

nutriome

Etc’ome

EMR / EHR

- Improve biological and medical knowledge
- Improve disease definition
- Discovery diagnostic markers
- Discovery prognostic markers
- Understanding early pathophysiology
- Disease stratification
- Patient stratification
- New therapeutic leads
- Adapt therapies to the above

PERSONAL DATA

« PRECISION » MEDICINE