



University of  
Zurich<sup>UZH</sup>

# BIO392 Bioinformatics of Genome Variations

Genomes: Core of "Personalized Health" & "Precision Medicine"

Michael Baudis **UZH SIB**  
Computational Oncogenomics

# BIO392: Course Schedule

<https://compbiozurich.org/UZH-BIO392/>

	Friday 2025-04-04	Tuesday 2025-04-08	Wednesday 2025-04-09	Thursday 2025-04-10	Friday 2025-04-11	Tuesday 2025-04-15	Wednesday 2025-04-16	Thursday 2025-04-17	Tuesday 2025-04-29	Wednesday 2025-04-30	Friday 2025-05-02	Tuesday 2025-05-06	Wednesday 2025-05-07
09:00 - 10:00	* Room information * Administrative - discuss times/days - exam		Jiahui: Terminal / Unix / Files	Hangjia: R environment introduction	Michael: Genomic Resources & Data Sharing		Feifei: Sequence analysis practical. FastQC, trimmomatic, BWA-MEM2, SAMtools, GangSTR, BCFtools	Michael: Genomic Data & Privacy		Feifei & Ziying: survival	Feifei: population structure		
10:00 - 11:00	Tina Siegenthaler: technical introduction (room, computer, accounts)		Jiahui: Terminal / Unix / Files	Hangjia: R exercise	Michael: Genomic Resources & Data Sharing		Feifei:: Sequence analysis practical	Michael: Genomic Data & Privacy		Feifei & Ziying: survival	Feifei: population structure		Discussion
11:00 - 12:00	* explore course site * create Github accounts and forward to bio392@compbiozurich.org *feifei&jiahui: overall schedule of the course		jiahui: SIB online introduction to Unix		Michael: Genomic Resources & Data Sharing		Feifei:: Sequence analysis practical	Michael: Genomic Data & Privacy		Feifei: survival	Feifei: population structure		
13:00 - 14:00	jiahui: Github	Michael: Introduction	Jiahui & Ziying: Python	Hangjia: CNV paper reading	Hangjia: Clinvar and Clingen	Feifei: Sequence analysis introduction. Overview of pipeline from raw reads -> variant calling & interpretation	Feifei:: STR reading up			Feifei:: analysis & interpretation. Parsing VCF (cvcf2), UCSD genome browser, ENSEMBL variant effect predictor	Feifei: population structure	Feifei & others: Presentation & Discussion	Exam revision, Q&A
14:00 - 15:00	jiahui: Github&Git exercise	Michael: Introduction	Exercise	Hangjia: Progenetix as tool for CNV frequencies etc.	Hangjia: blast	Feifei: Sequence analysis introduction	Feifei:: STR reading up			Feifei:: analysis & interpretation.	Feifei: population structure	Feifei & others: Presentation & Discussion	
15:00 - 16:30		Michael: Introduction			Hangjia: Blast exercise					Feifei:: analysis & interpretation.	Feifei: population structure	Feifei & others: Presentation & Discussion	

1992



Heidelberg

Student of medicine | doctoral thesis in molecular cytogenetics @ DKFZ (Peter Licher) | resident in clinical hematology/oncology | data, clinical studies & cancer systematics

2001



Stanford

Post-doc in hemato-pathology (Michael Cleary) | molecular mechanisms of leukemogenesis | transgenic models | expression arrays | systematic cancer genome data collection | *Progenetix* website

2003



Gainesville

Assistant professor in paediatric haematology | molecular mechanisms of leukemogenesis | focus on bioinformatics for cancer genome data analysis

2006



Aachen

Research group leader in genetics | genomic array analysis for germline alterations | descriptive analysis of copy number aberration patterns in cancer entities

2007



Zürich

Professor of bioinformatics @ DMLS (2015) | systematic assembly of oncogenomic data | databases and software tools | patterns in cancer genomes | *Progenetix* & *arrayMap* resources | GA4GH | SPHN | ELIXIR

# Our Research

## Theoretical Cytogenetics & Oncogenomics

- CNV resource
    - Data - e.g. [progenetix.org](https://progenetix.org)
    - Tools - CNV remapping, visualization, API access to resources ...
  - patterns and correlations of genomic variations in cancer
  - annotation mapping
  - API, protocols and standards contributions
- ➡ Beacon

[baudisgroup.org](http://baudisgroup.org)

baudisgroup @ UZH & SIB  
[Baudisgroup Home](#)

[Latest News & Publications](#)  
[Address](#)  
[Group Members](#)  
[Publications](#)  
[Presentations](#)  
[Projects and Open Positions](#)  
[Teaching and Seminars](#)  
[BIO390 Lectures ↗](#)  
[BIO392 Block Course ↗](#)  
[ZH Bioinfo Seminars ↗](#)  
[Progenetix ↗](#)  
[CancerCellLines ↗](#)  
[CompbioZurich ↗](#)  
[SchemaBlocks {S}\[B\] ↗](#)  
[Beacon Project ↗](#)  
[Michael Baudis @ UZH ↗](#)

## Welcome to the *baudisgroup* Pages

The *baudisgroup* website represents projects and information by the **Computational Oncogenomics Group** of the [University of Zurich \(UZH\)](#) and the [Swiss Institute of Bioinformatics \(SIB\)](#). For visitors more interested in Particle Astrophysics, we strongly recommend the website of another, although related, [Professor Baudis](#).

The Computational Oncogenomics Group's research focus lies in the exploration of structural genome variations in cancer. Our work centres around our [Progenetix](#) resource of curated molecular-cytogenetic and sequencing data. Specific projects explore computational methods, genomics of selected tumour entities and genomic variant patterns across malignancies. As members of the [Global Alliance for Genomics and Health](#), the group is developing standards in biocuration and data sharing for genomic variants and phenotypic data, for instance in driving development of the [ELIXIR Beacon](#) project. Other research is related to genome data epistemology, e.g. geographic and diagnostic sampling biases in cancer studies. Some general information can be found in [these slides](#).



Have you seen deletions in this region on chromosome 9 in Glioblastoma from a juvenile patient, in a dataset with unrestricted access?



The Beacon API v2 proposal opens the way for the design of a simple but powerful "genomics API".

## Latest News & Publications

🕒 March 23, 2025

[pgxRpi: an R/Bioconductor package for user-friendly access to the Beacon v2 API](#)

**Hangjia Zhao and Michael Baudis**

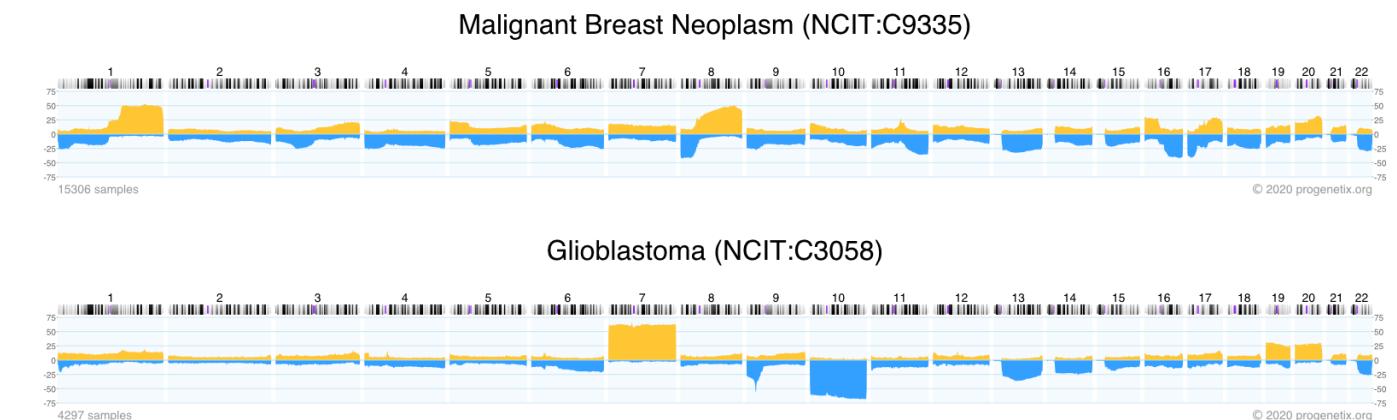
BIOARXIV PREPRINT (2025-03-23): [HTTPS://DOI.ORG/10.1101/2025.03.20.644282](https://doi.org/10.1101/2025.03.20.644282)

**Abstract:** The Beacon v2 specification, established by the Global Alliance for Genomics and Health (GA4GH), consists of a standardized framework and data models for genomic and phenotypic data discovery. By enabling secure, federated data sharing, it fosters interoperability across genomic resources. Progenetix, a reference implementation of Beacon v2, exemplifies its potential for large-scale genomic data integration, offering open access to genomic mutation data across diverse cancer types. Here we present pgxRpi, an open-source R/Bioconductor package that provides a streamlined interface to the Progenetix Beacon v2 REST API, → [Continue reading](#)

# Theoretical Cytogenetics and Oncogenomics

## ... but what does this entail @baudisgroup?

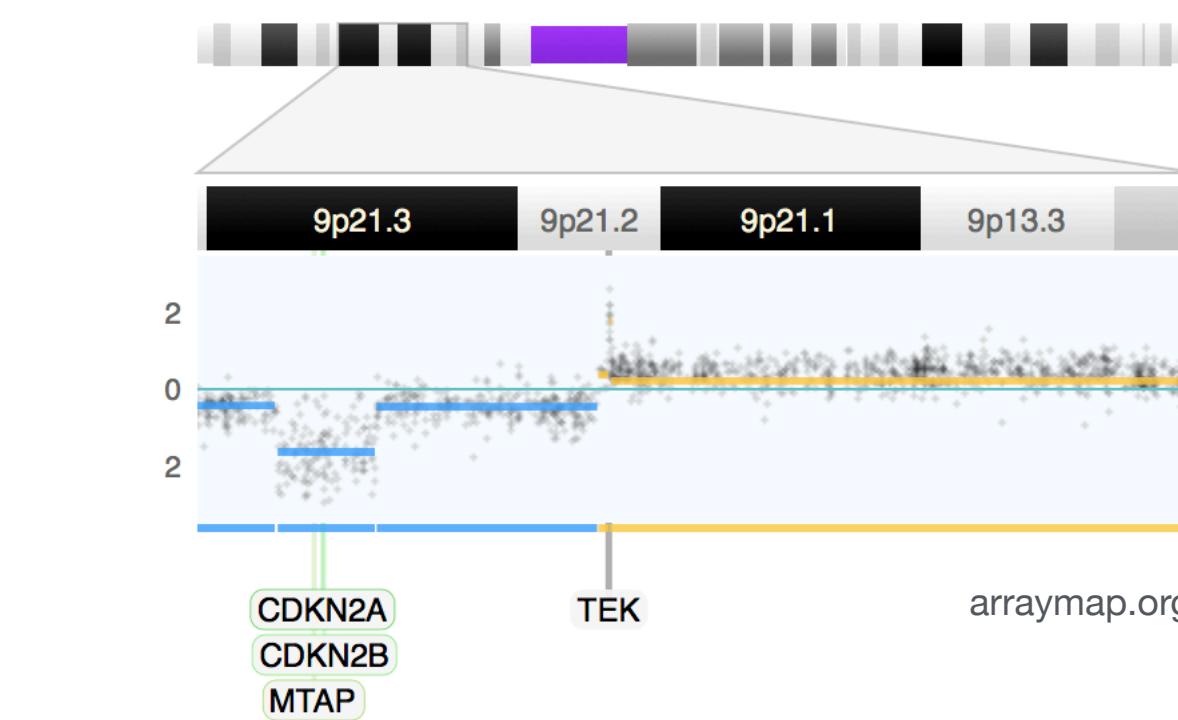
- patterns & markers in cancer genomics, especially somatic structural genome variants
- bioinformatics support in collaborative studies
- reference resources for curated cancer genome variations
- bioinformatics tools & methods
- standards and reference implementations for data sharing in genomics and personalized health
- open research data "ambassadoring"



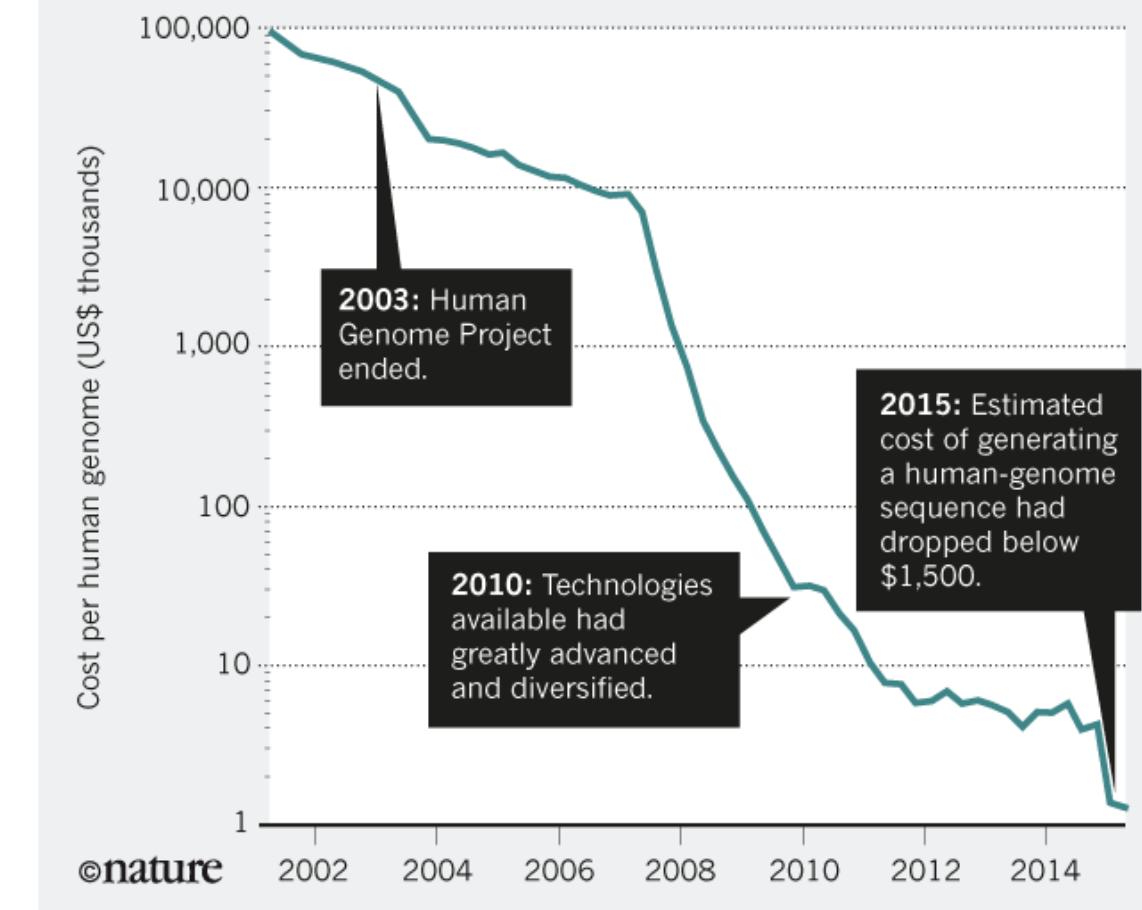


## Genome screening at the core of “Personalised Health”

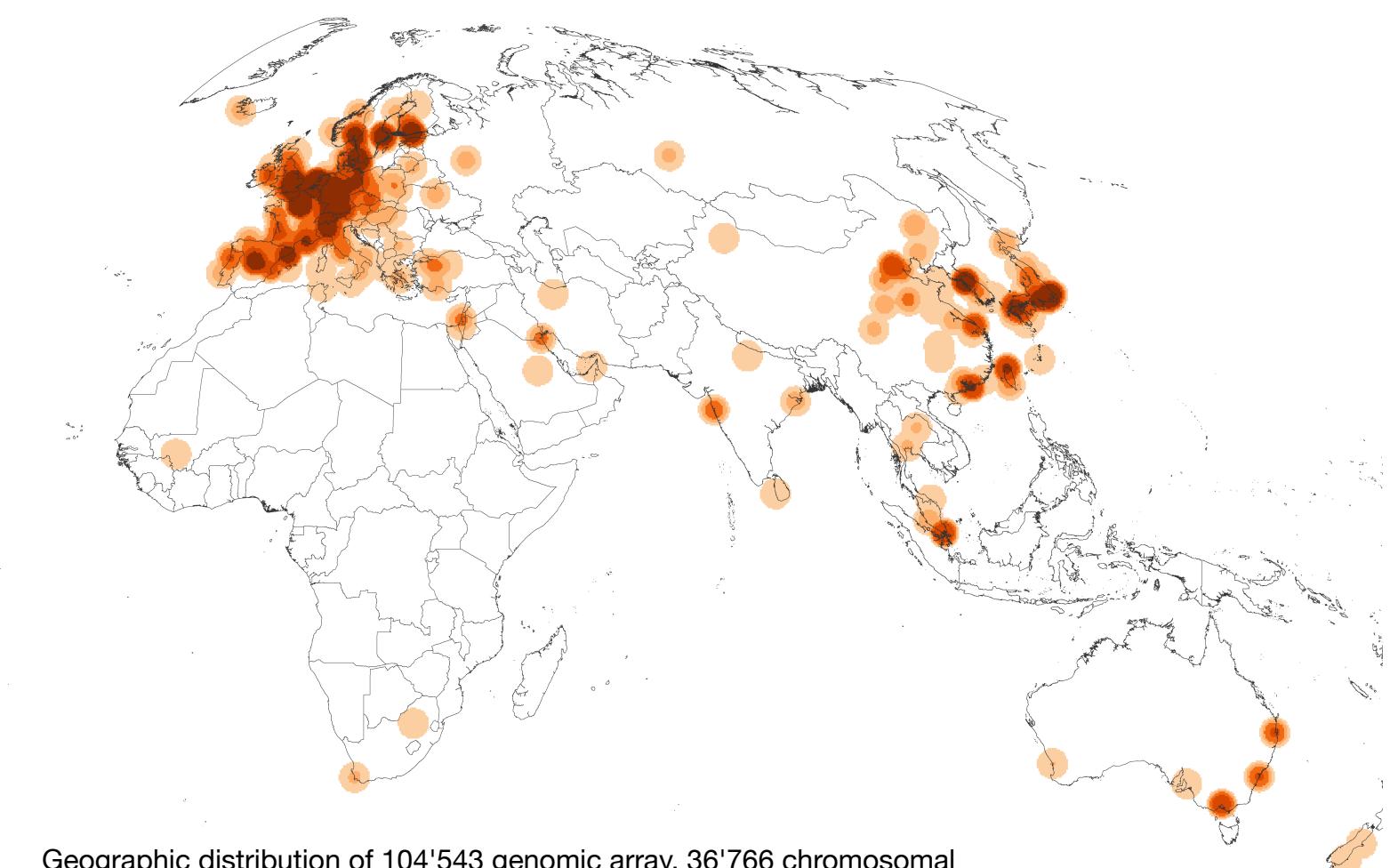
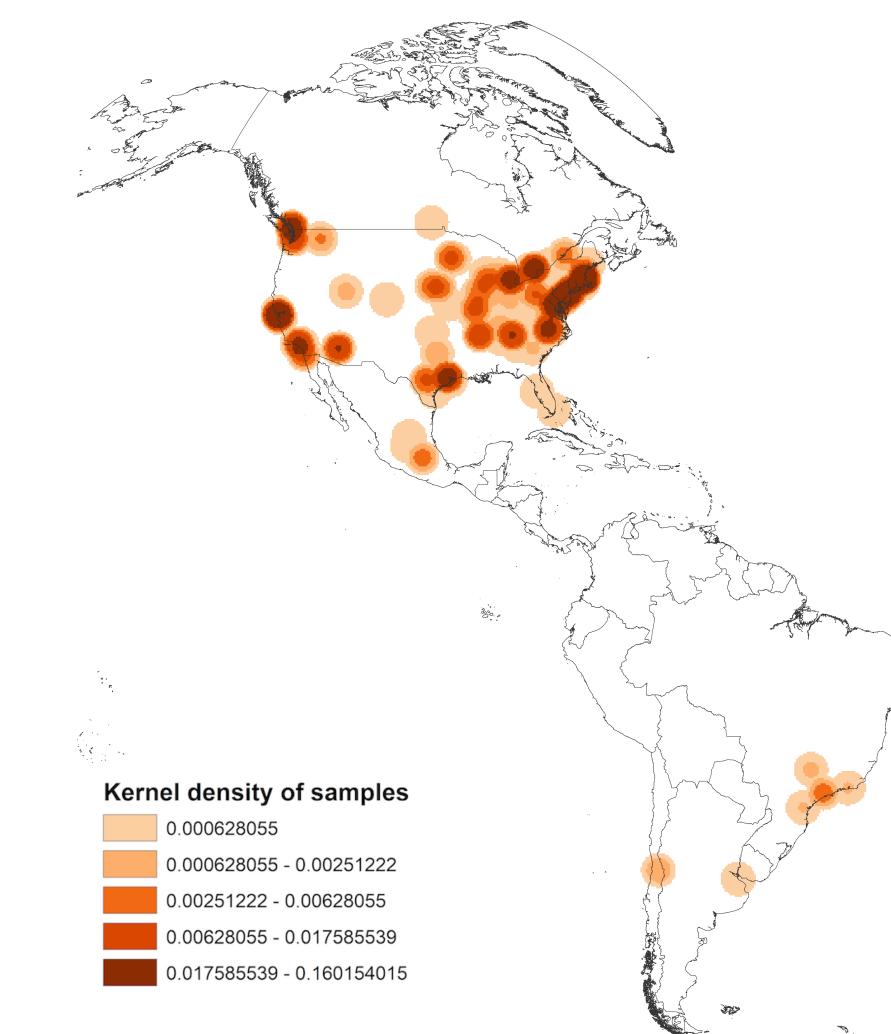
- ▶ **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications
- ▶ The unexpectedly large amount of **sequence variants** in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of **reference repositories**
- ▶ **Standardized data formats** and **exchange protocols** are needed to connect these resources throughout the world, for reciprocal, international **data sharing** and **biocuration** efforts
- ▶ Our work @ UZH:
  - ▶ **cancer genome repositories**
  - ▶ **biocuration**
  - ▶ **protocols & formats**



**BETTER, CHEAPER, FASTER**  
The cost of DNA sequencing has dropped dramatically over the past decade, enabling many more applications.



The future of DNA sequencing. Eric D. Green, Edward M. Rubin & Maynard V. Olson. Nature; 11 October 2017 (News & Views)

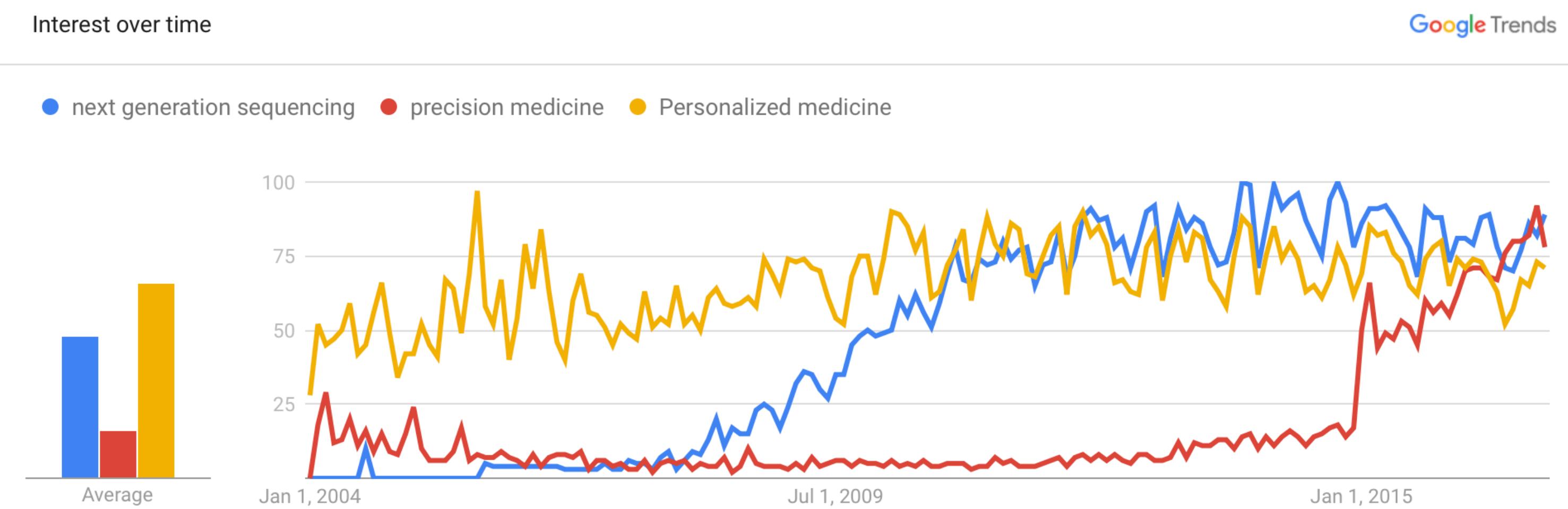


Geographic distribution of 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets

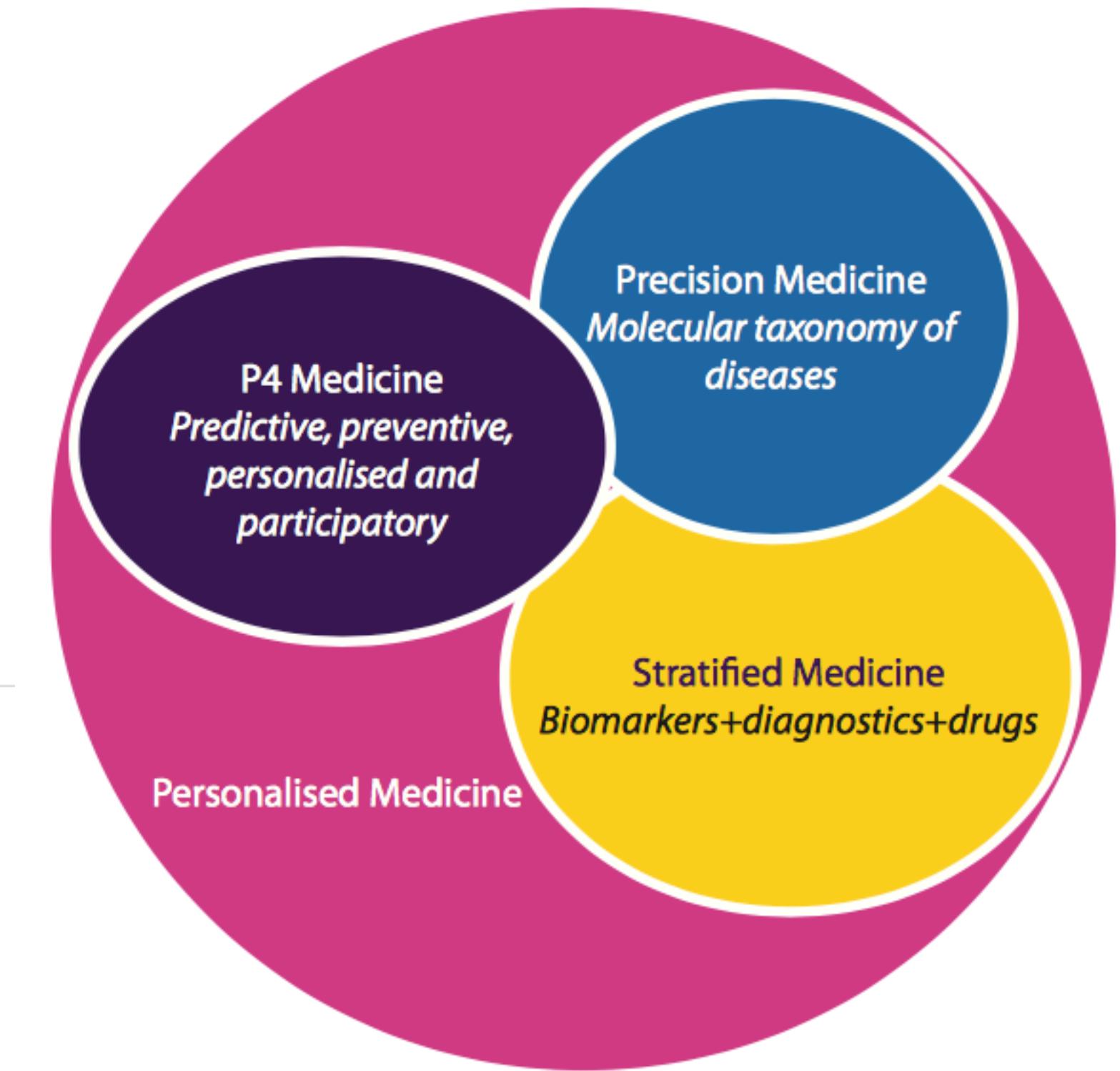
# **Begriffsbestimmung**

# Many names for one concept or many concepts in one name?

Stratified, personalised, precision, individualised, P4 medicine or personalised healthcare – all are terms in use to describe notions often referred to as the future of medicine and healthcare. But what exactly is it all about, and are we all talking about the same thing?



Worldwide. 2004 - present.



Source: PHG Foundation

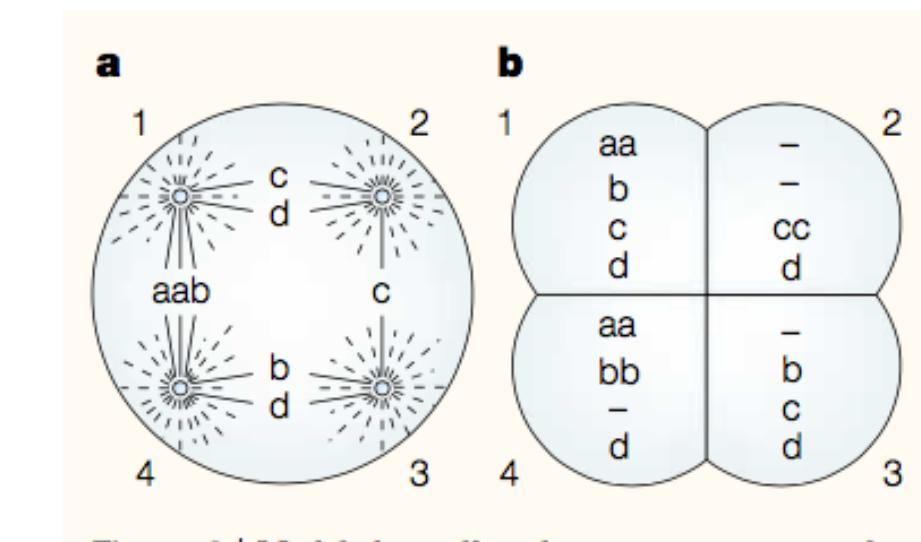
While medicine has always been "personal" and "precise" in the given context of available knowledge and technologies, the concept of "**Personalised Medicine**" describes the use of individual genome information, concept based metadata and individually targeted therapies.

**It started ... How?**



# Zelluläre Vererbung von malignen Eigenschaften

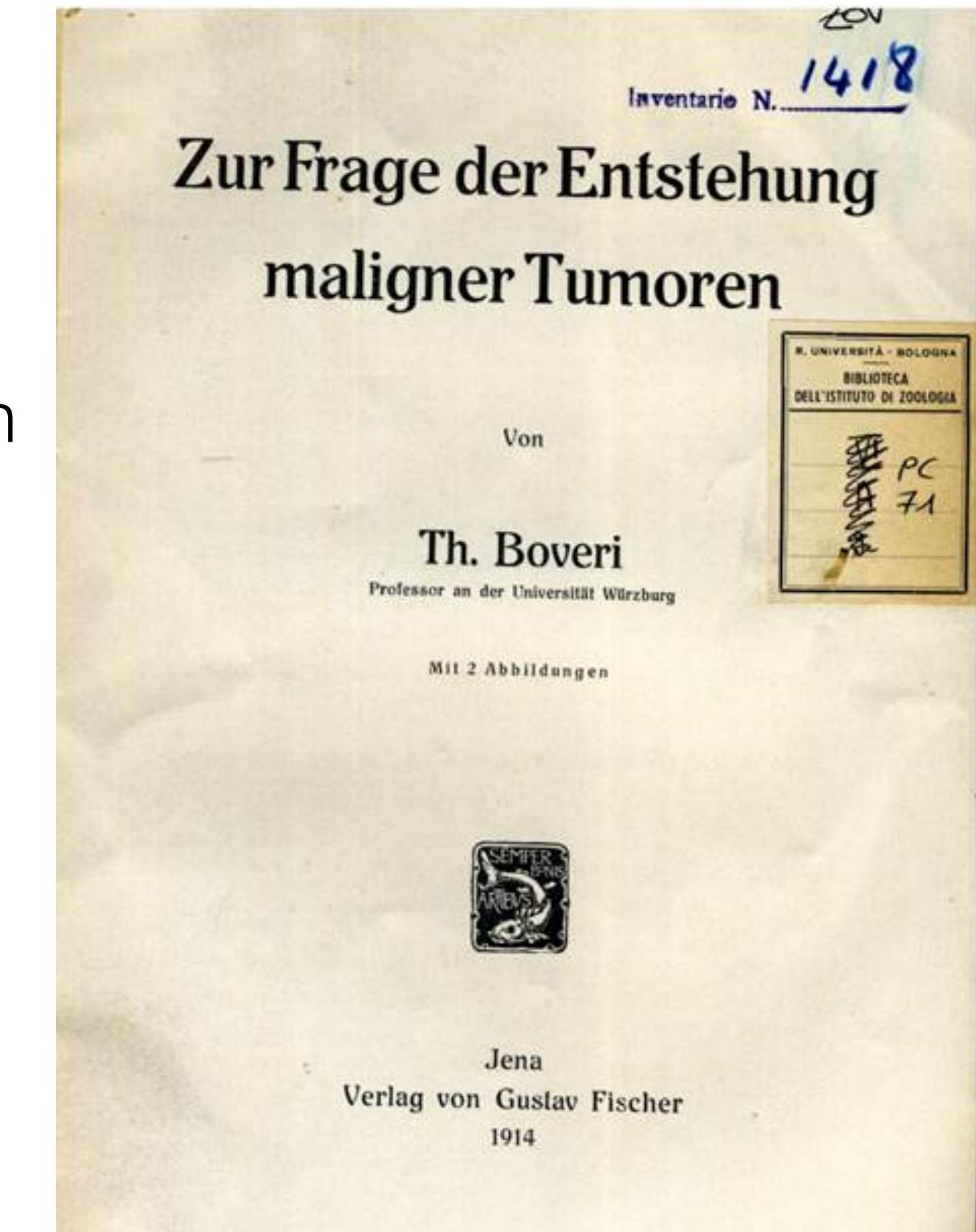
## Theodor Boveri (1914)



Allan Balmain  
Cancer genetics:  
from Boveri and  
Mendel to  
microarrays.  
NatRev Cancer  
(2001); 1: 77-82

Experimentelle Beobachtungen in Seeigeleiern, mit Vorhersage von:

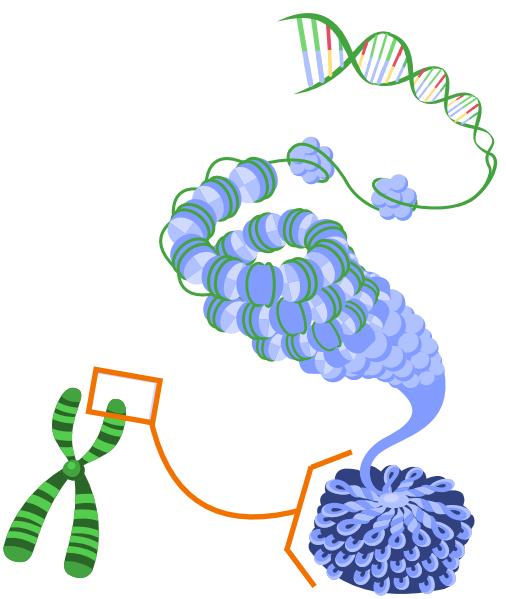
- **Tumorkontrollogene** ("Teilungshemmende Chromosomen") zur **Zellzykluskontrolle**, in bösartigen Tumoren geschädigt
- **Tumorgene** ("Teilungsfoerdernde Chromosomen"), in Tumoren vermehrt sein
- **Progression** (gutartig zu bösartig); sequentielle Veränderungen in Chromosomen
- Klonalität & Genetische Mosaike
- **Veranlagung** zu Krebs durch ererbte "Chromosomen"
- **Homozygosität** und hoher Penetranz bei Vererbung von beiden Eltern (z. Bsp. Xeroderma pigmentosum)
- Wunden und Entzündung; Metastasierung durch Adhäsionsverlust; Strahlentherapie... (basierend auf Hertwig *et al.*)



Anna Di Leonardo , Sergio Nasi , Simonetta Pulciani  
Cancer: We Should Not Forget The Past  
Journal of Cancer (2015), Vol. 6: 29-39  
(for book cover & summary)

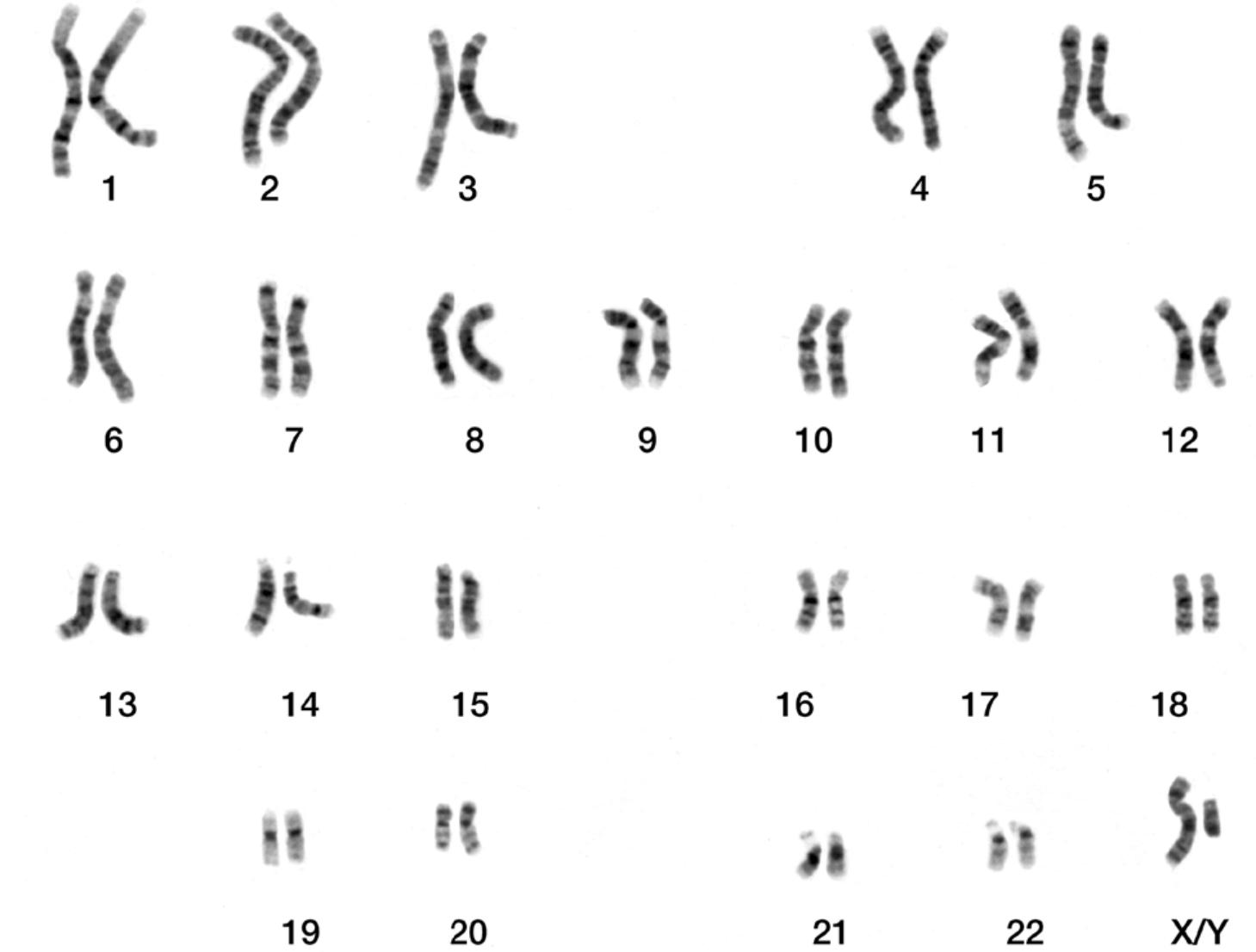
# Zytogenetik

## Genomik mit dem Mikroskop

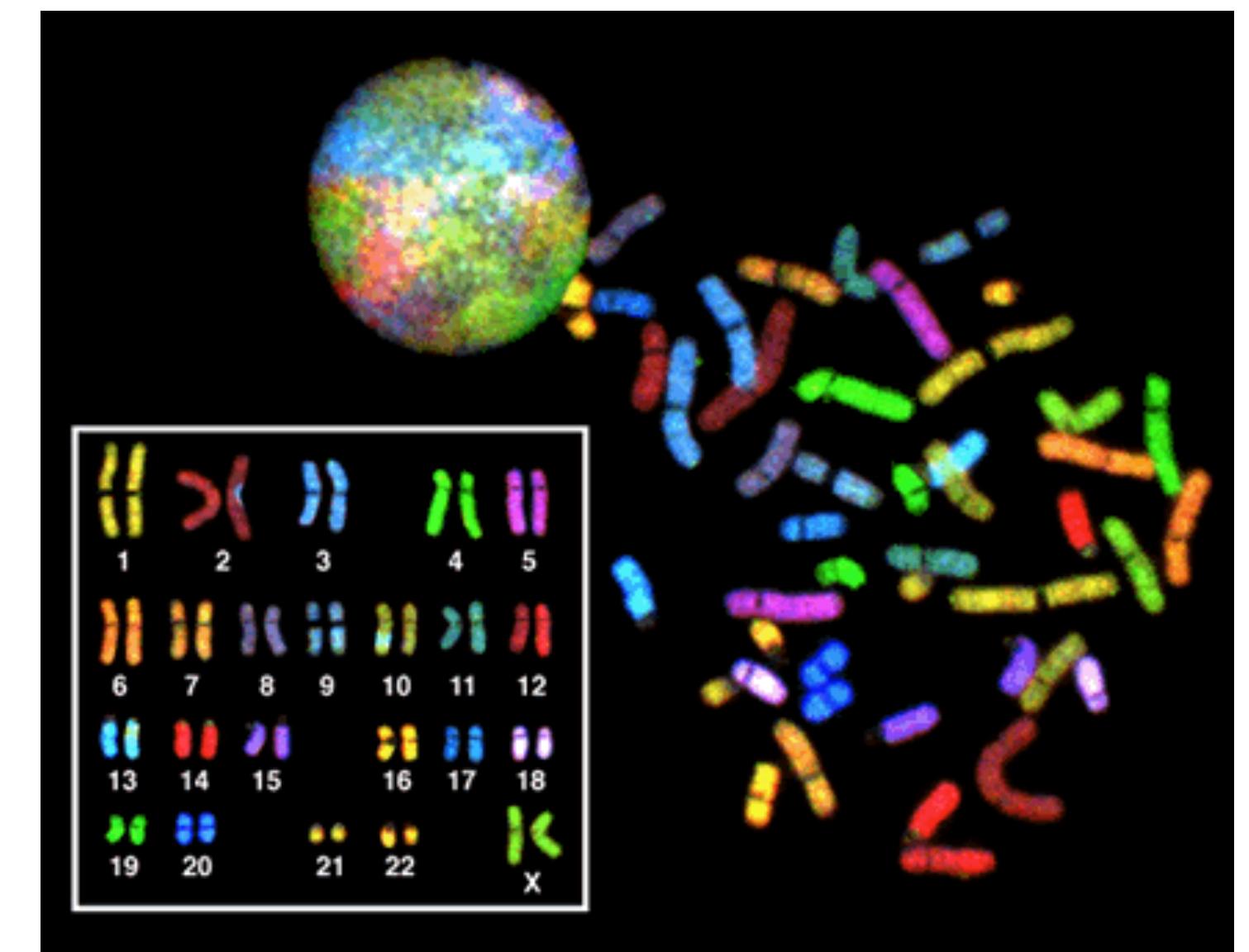


DNA/Histone cartoon modified from simpleclub.com

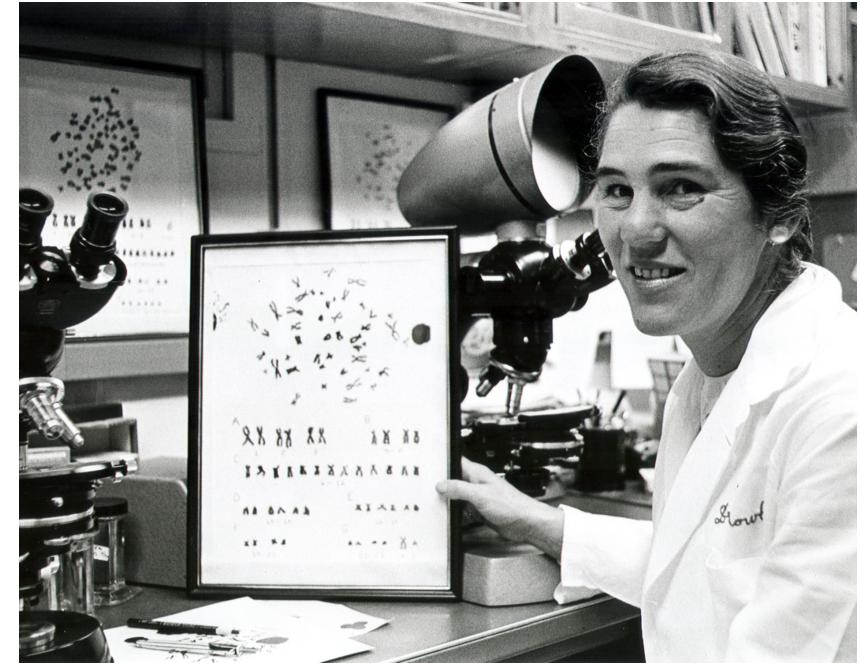
- Das menschliche Erbgut besteht aus mehr als 3 Milliarden Basenpaaren ("Di-Nucleotiden") der **DNS** (Desoxyribonukleinsäure)
- Diese können während der Zellteilung in ihrer kondensierten Form ("**Chromosomen**") sichtbar gemacht werden
- Die Analyse **chromosomaler Veränderungen** als Zeichen von Mutationen in Keimbahn oder Krebszellen ist das Feld der **Zytogenetik**
- Die molekulare Zytogenetik kombiniert Chromosomenanalyse mit **molekularen Markierungen**, zum Beispiel fluoreszierenden DNS-Sonden



Normales männliches Karyogramm. Quelle: NHGRI via Wikipedia



Normales weibliches Karyogramm mit chromosomenspezifischen "painting probes". Quelle: NHGRI via Wikipedia



# Chromosomale Translokationen in Krebs

## Janet Rowley (1972/73)

- "Philadelphia Chromosom" in Chronisch Myeloischen Leukämien (CML) von Nowell & Hungerford 1960 beschrieben
- Rowley entdeckte dass die "Marker" in einigen Leukämien durch Bruch und Verschmelzung von normaler Chromosomen entstehen
  - Bei "Philadelphia Chromosom" Aktivierung der Tyrosinkinase ABL => konstantes Wachstumssignal
  - Seit 1998 (STI-571; Imatinib/Gleevec) hat die medikamentöse Inhibition des aktivierte Proteins in CML die primäre Chemotherapie ersetzt

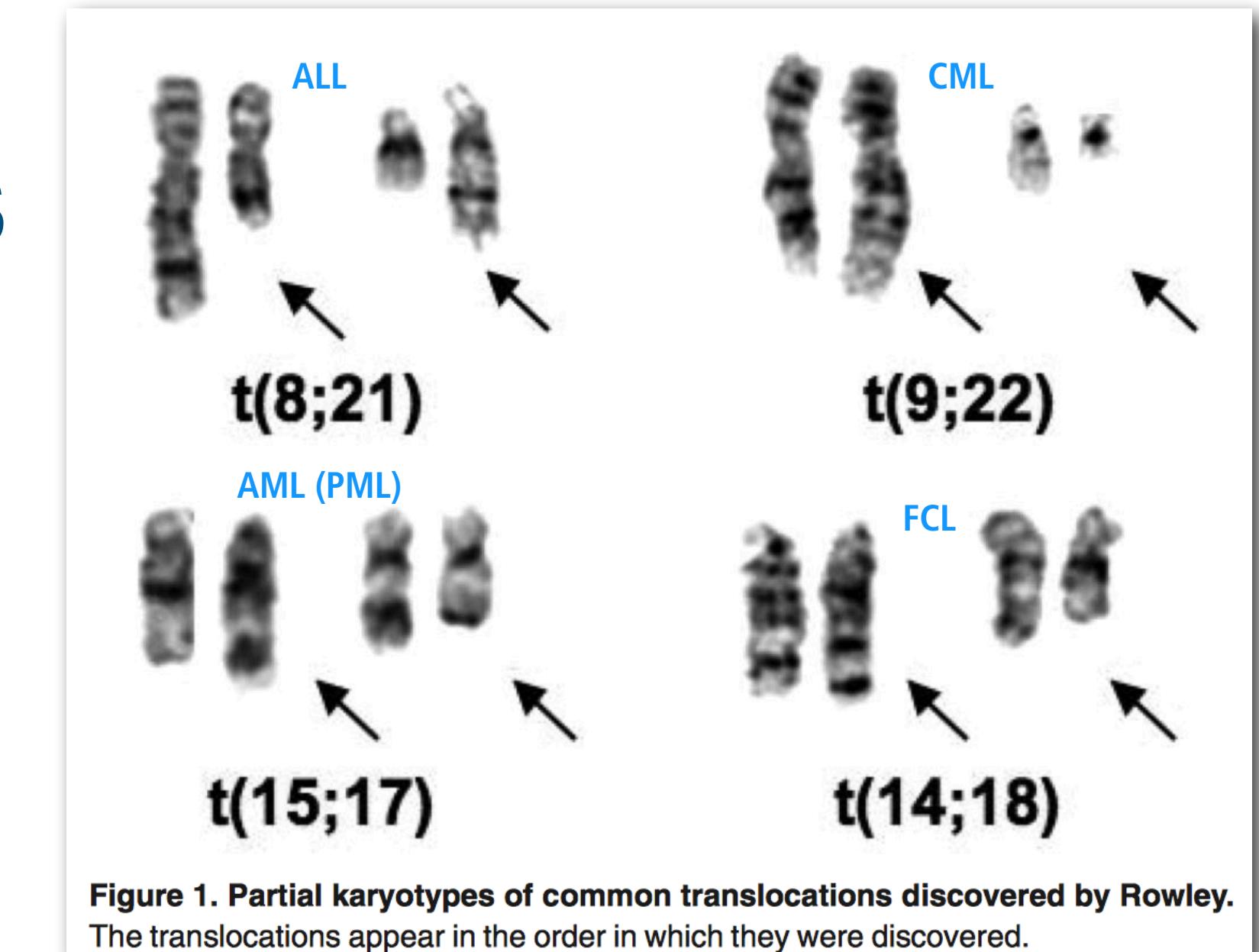
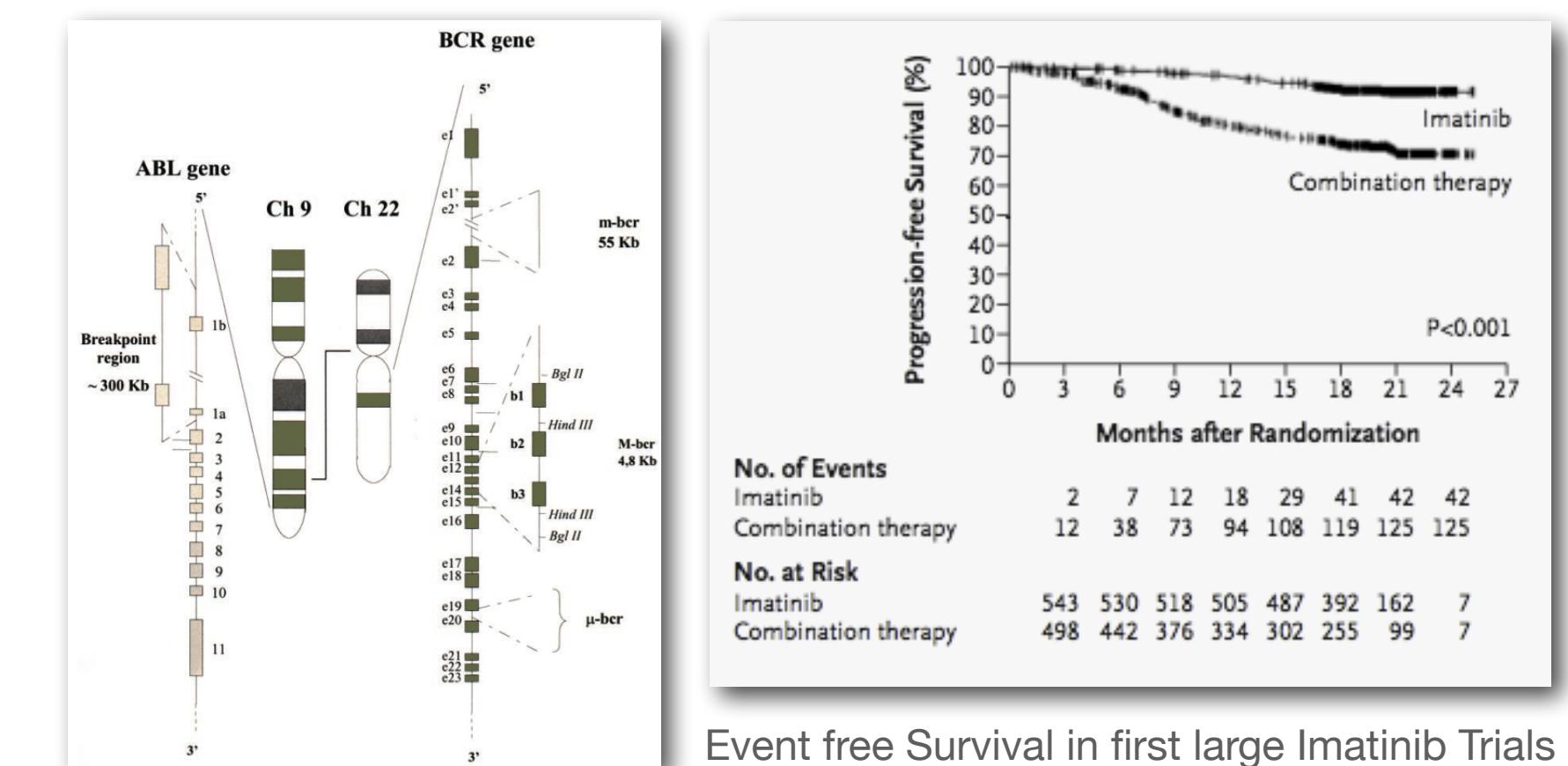


Figure 1. Partial karyotypes of common translocations discovered by Rowley.  
The translocations appear in the order in which they were discovered.

Janet D Rowley. Chromosomal translocations: revisited yet again  
Blood (2008), 112(6)

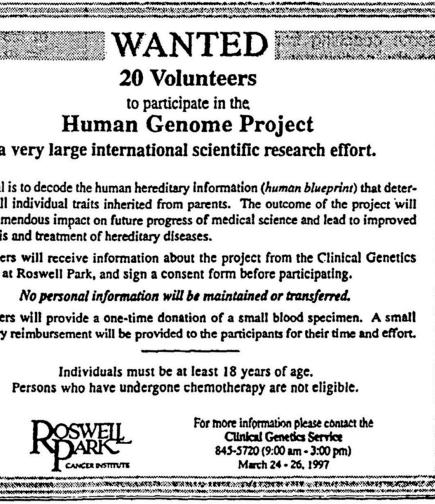


Event free Survival in first large Imatinib Trials

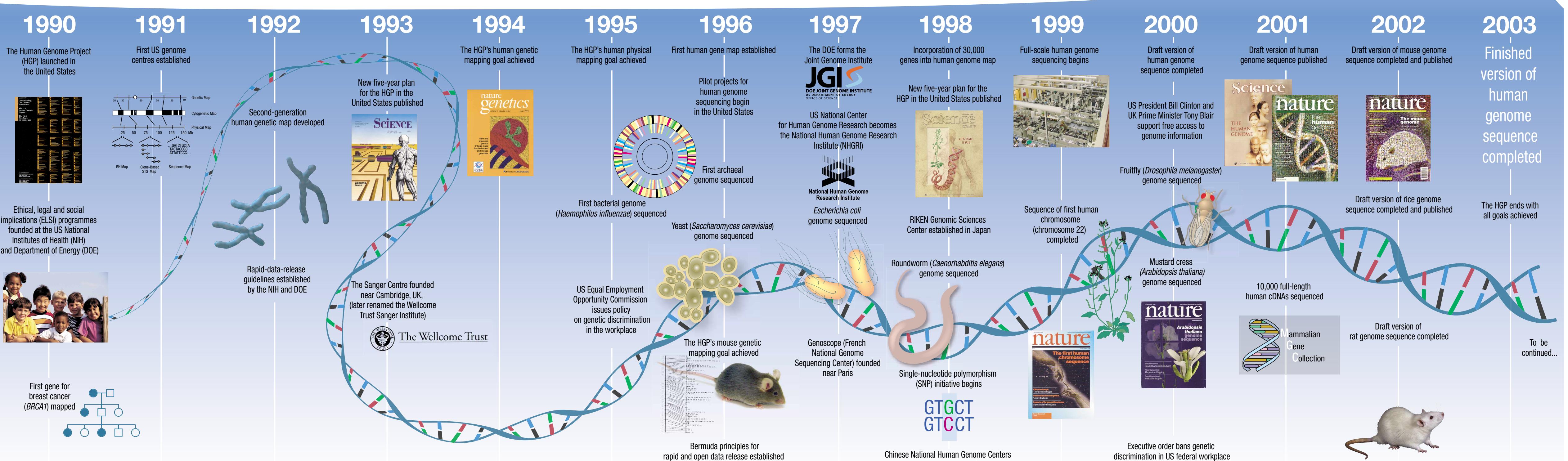
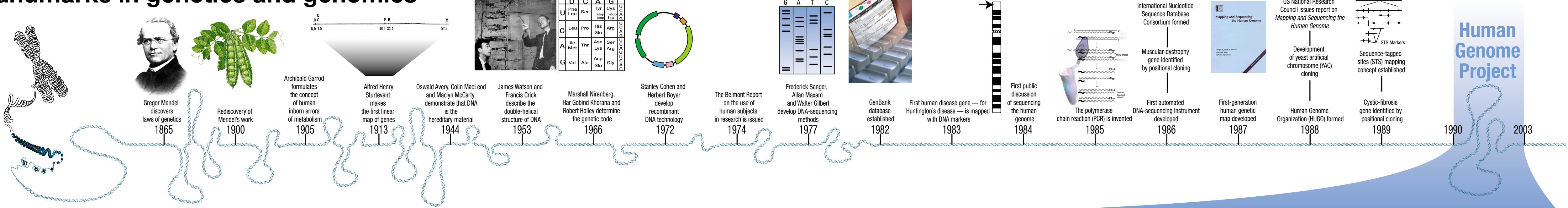
Pane et al. BCR/ABL genes ....  
Oncogene (2002), 21 (56)

O'Brien et al. Imatinib compared with interferon and low-dose cytarabine...  
NEJM (2003) vol. 348 (11)

# The Human Genome Project 1990-2003



## Landmarks in genetics and genomics



# Genome analyses at the core of Personalized Health™

Susceptibility, Pharmacogenomics, Classification, Infectious Diseases, Outcome Prediction, Lifestyle ...

doi:10.1038/nature19057

## Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek<sup>1,2,3,4</sup>, Konrad J. Karczewski<sup>1,2\*</sup>, Eric V. Minikel<sup>1,2,5\*</sup>, Kaitlin E. Samocha<sup>1,2,5,6\*</sup>, Eric Banks<sup>2</sup>, Timothy Fennell<sup>2</sup>, Anne H. O'Donnell-Luria<sup>1,2,7</sup>, James S. Ware<sup>2,8,9,10,11</sup>, Andrew J. Hill<sup>1,2,12</sup>, Beryl B. Cummings<sup>1,2,5</sup>, Taru Tukiainen<sup>1,2</sup>, Daniel P. Birnbaum<sup>2</sup>, Jack A. Kosmicki<sup>1,2,6,13</sup>, Laramie E. Duncan<sup>1,2,6</sup>, Karol Estrada<sup>1,2</sup>

Rapid whole genome sequencing and precision neonatology

CrossMark

Joshua E. Petrikirin, MD<sup>a,\*</sup>, Laurel K. Willig, MD, FAAP<sup>b</sup>, Laurie D. Smith, MD, PhD<sup>c</sup>, and Stephen F. Kingsmore, MB, BAO, ChB, Dsc, FRCPath<sup>d,e</sup>

## Genomic Classification of Cutaneous Melanoma

The Cancer Genome Atlas Network<sup>1,\*,\*\*</sup>

<sup>1</sup>Cancer Genome Atlas Program Office, National Cancer Institute at NIH, 31 Center Drive, Bldg. 31, Suite 3A20, Bethesda, MD 20892, USA

\*Correspondence: irwatson@mdanderson.org (I.R.W.), jgershen@mdanderson.org (J.E.G.), lchin@mdanderson.org (L.C.)

<http://dx.doi.org/10.1016/j.cell.2015.05.044>

doi:10.1111/pcn.12128

## PCN Frontier Review

Psychiatry and Clinical Neurosciences

### Copy-number variation in the pathogenesis of autism spectrum disorder

RESEARCH ARTICLE Open Access

Chromothripsis-like patterns are recurring but heterogeneously distributed features in a survey of 22,347 cancer genome screens

Haoyang Cai<sup>1,2</sup>, Nitin Kumar<sup>1,2</sup>, Homayoun C Bagheri<sup>3</sup>, Christian von Mering<sup>1,2</sup>, Mark D Robinson<sup>1,2\*</sup>

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani, M.D., Jeff Settleman, Ph.D., and Daniel A. Haber, M.D., Ph.D.

N Engl J Med 2004; 350:2129-2139 | May 20, 2004 | DOI: 10.1056/NEJMoa040938

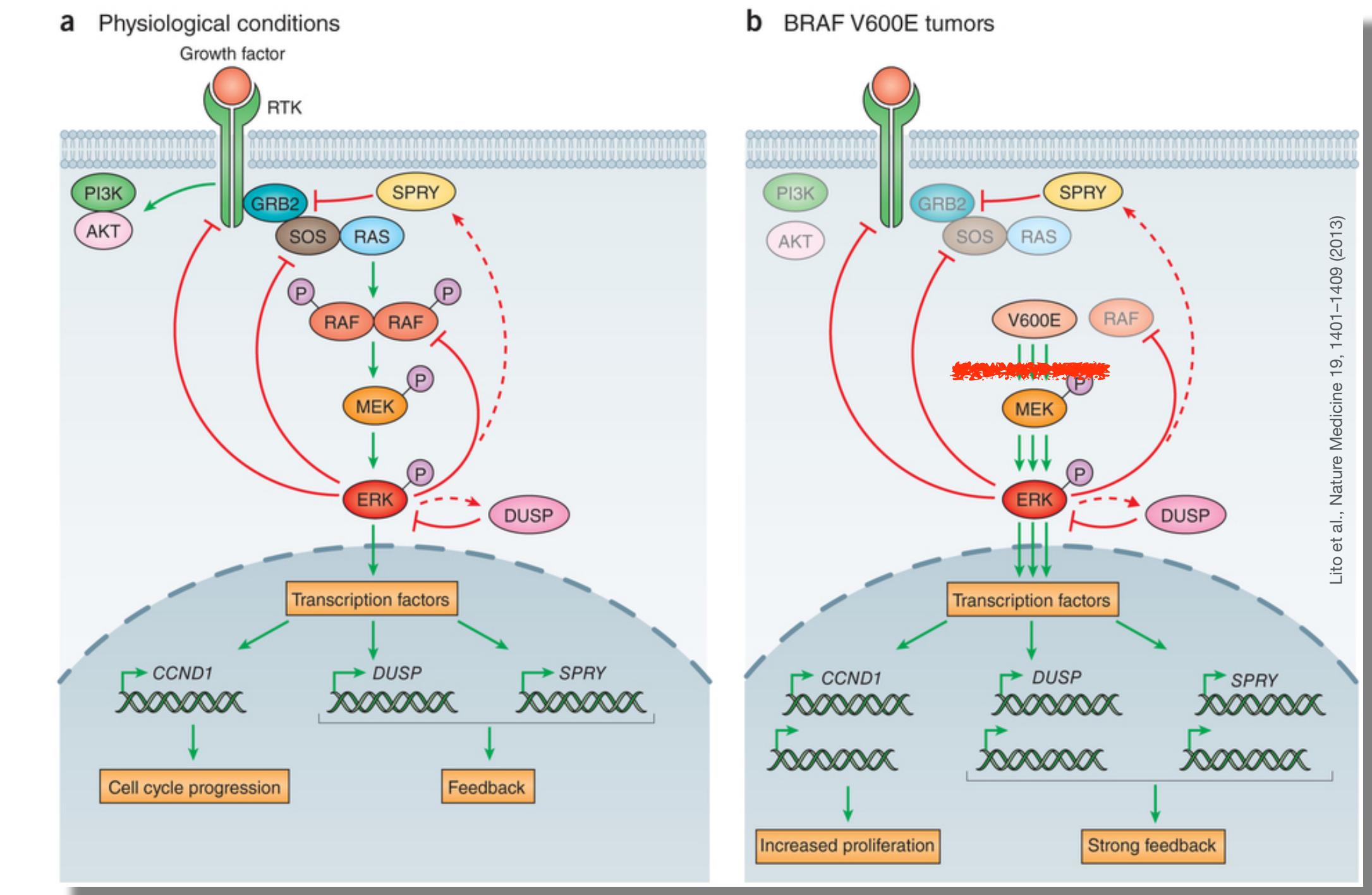
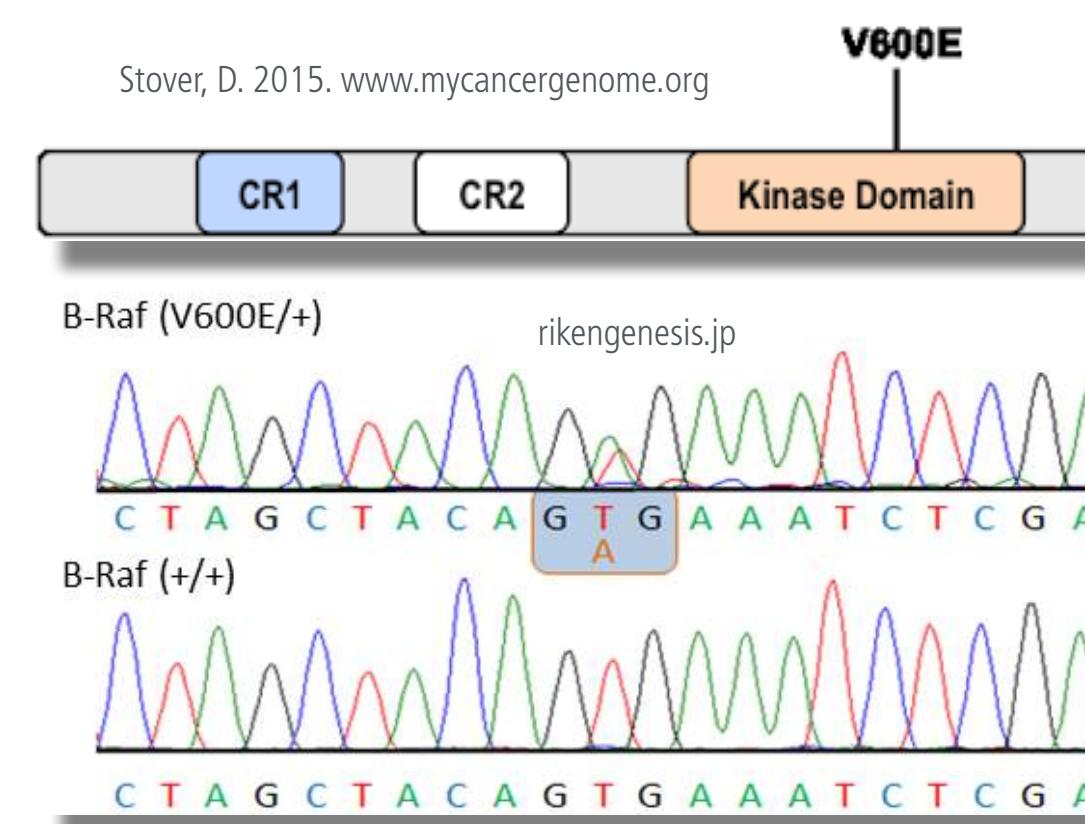
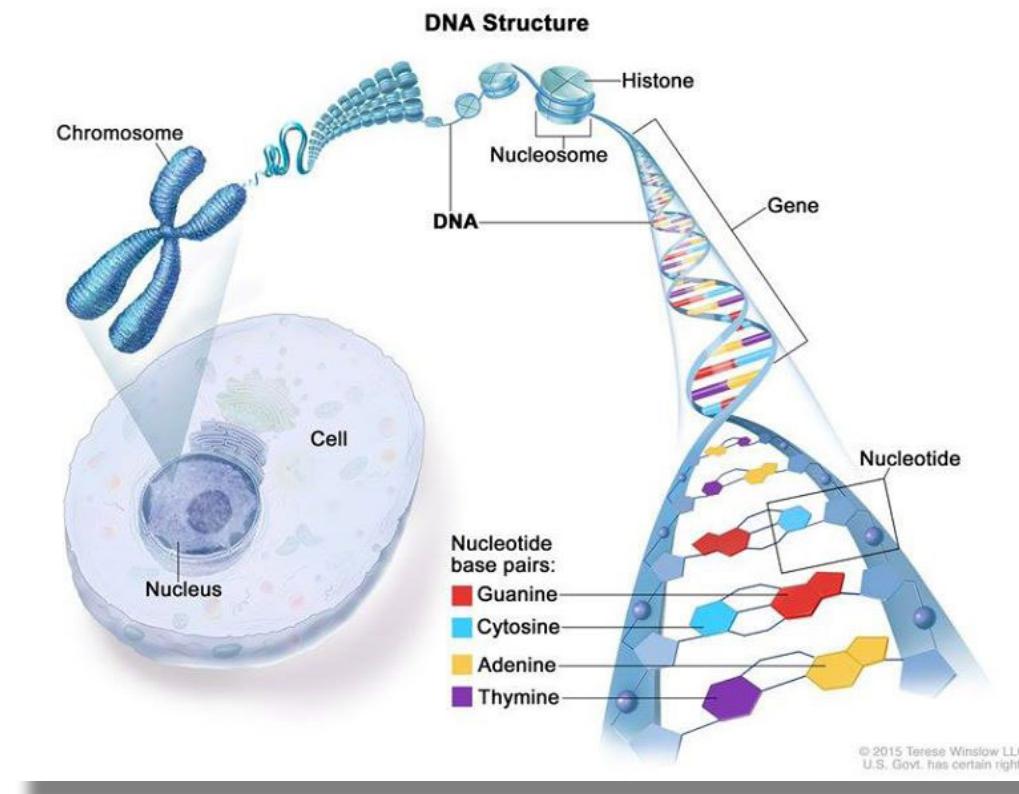
Activating Mutations in the Epidermal Growth Factor Receptor Underlying Responsiveness of Non-Small-Cell Lung Cancer to Gefitinib

Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David

# BRAF V600E (c.1799T>A) Mutation

## Oncogene Activation by Single Nucleotide Alteration

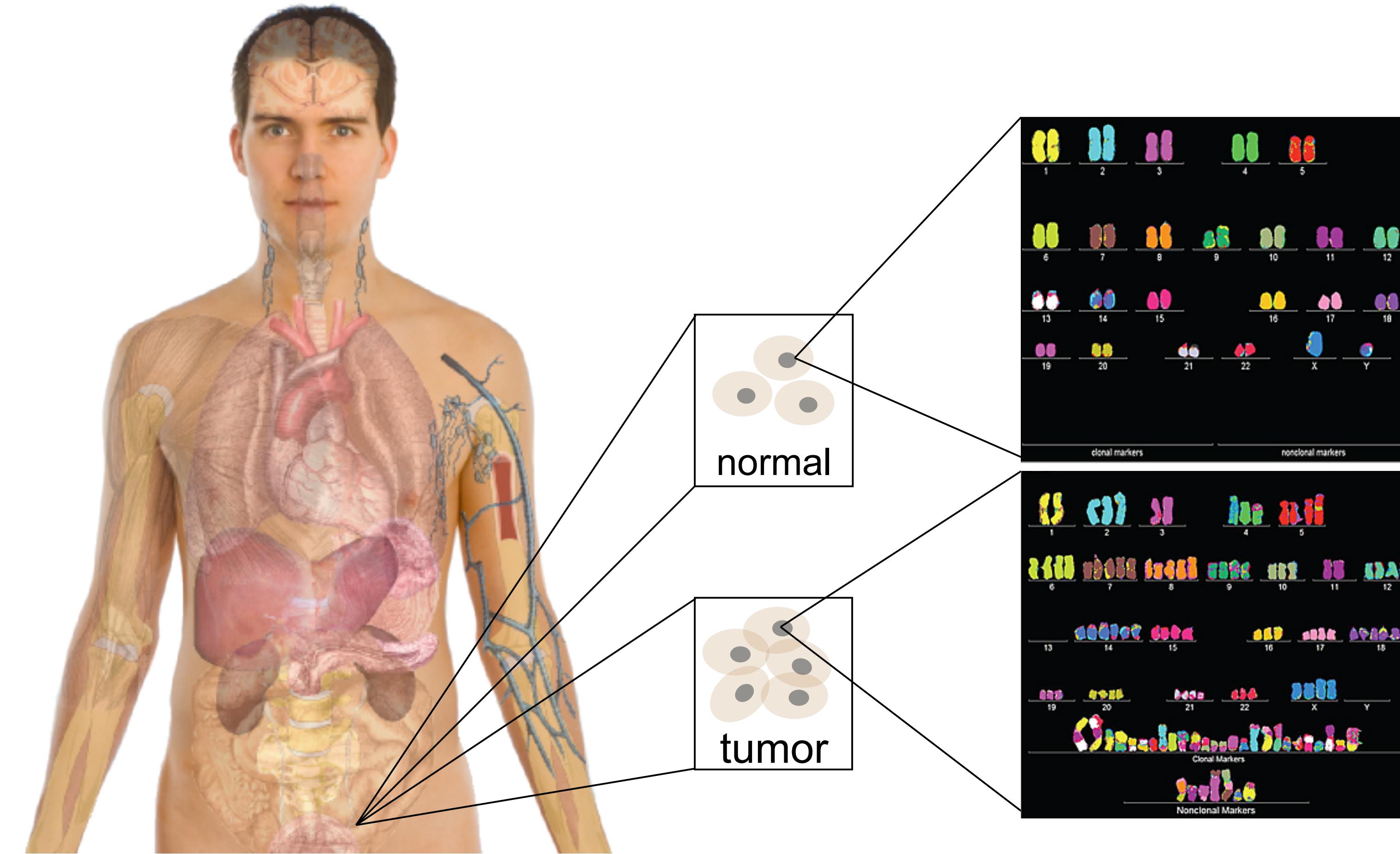
- a single nucleotide exchange Thymidine > Adenine leads to continuous RAF based activation of the MEK-ERK pathway
- BRAF V600E is a frequent mutation in >50% of malignant melanomas, but also CRC, lung ADC ...
- pharmacologic block of B-Raf (e.g. through **Vemurafenib**)



The BRAF V600E mutation leads to continuous phosphorylation of MEK, without the need for receptor based activation of the upstream pathway and loss of inhibitory feedback control.

# Personal Genomics as a Gateway into Biology

Personal genomes soon will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.



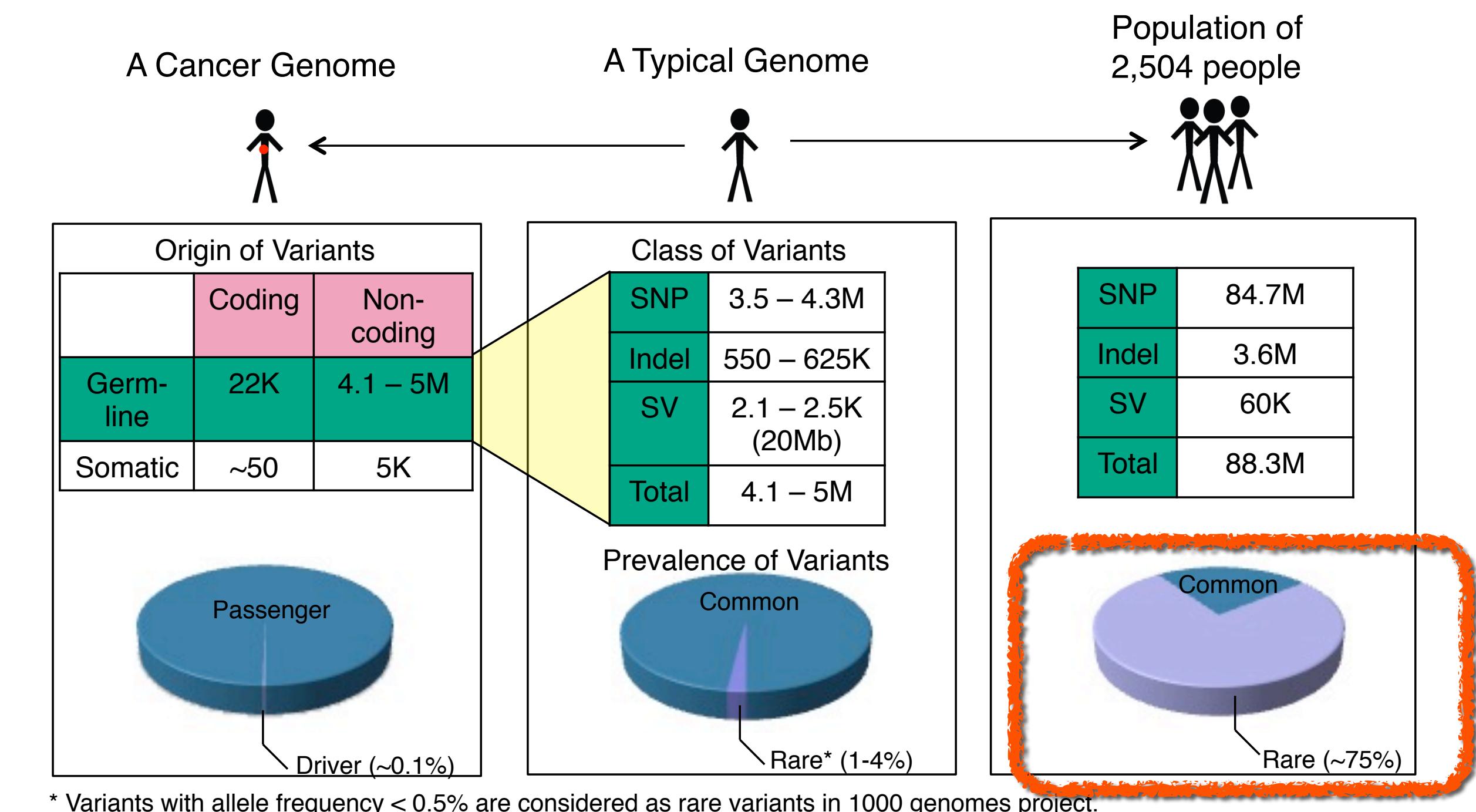
The trouble with human genome variation



# Finding Somatic Mutations In Cancer

## Many Needles in a Large Haystack

- a typical human genome (~3 billion base pairs) has ~5 million variants
- most of them are "**rare**"; i.e. can only be identified as recurring when sequencing thousands of people
- cancer cells accumulate additional variants, only **few** of which ("**drivers**") are relevant for the disease

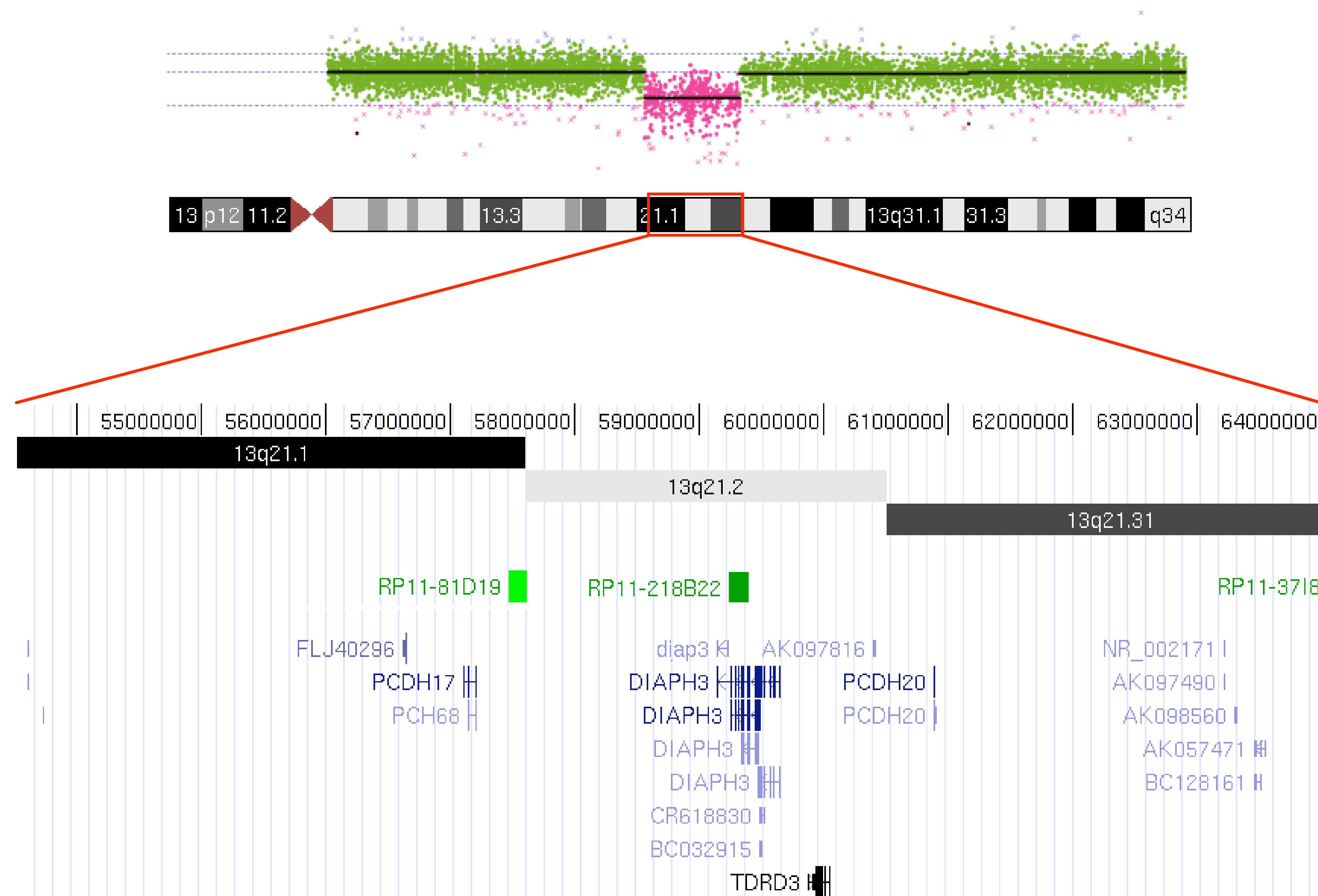


The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74  
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

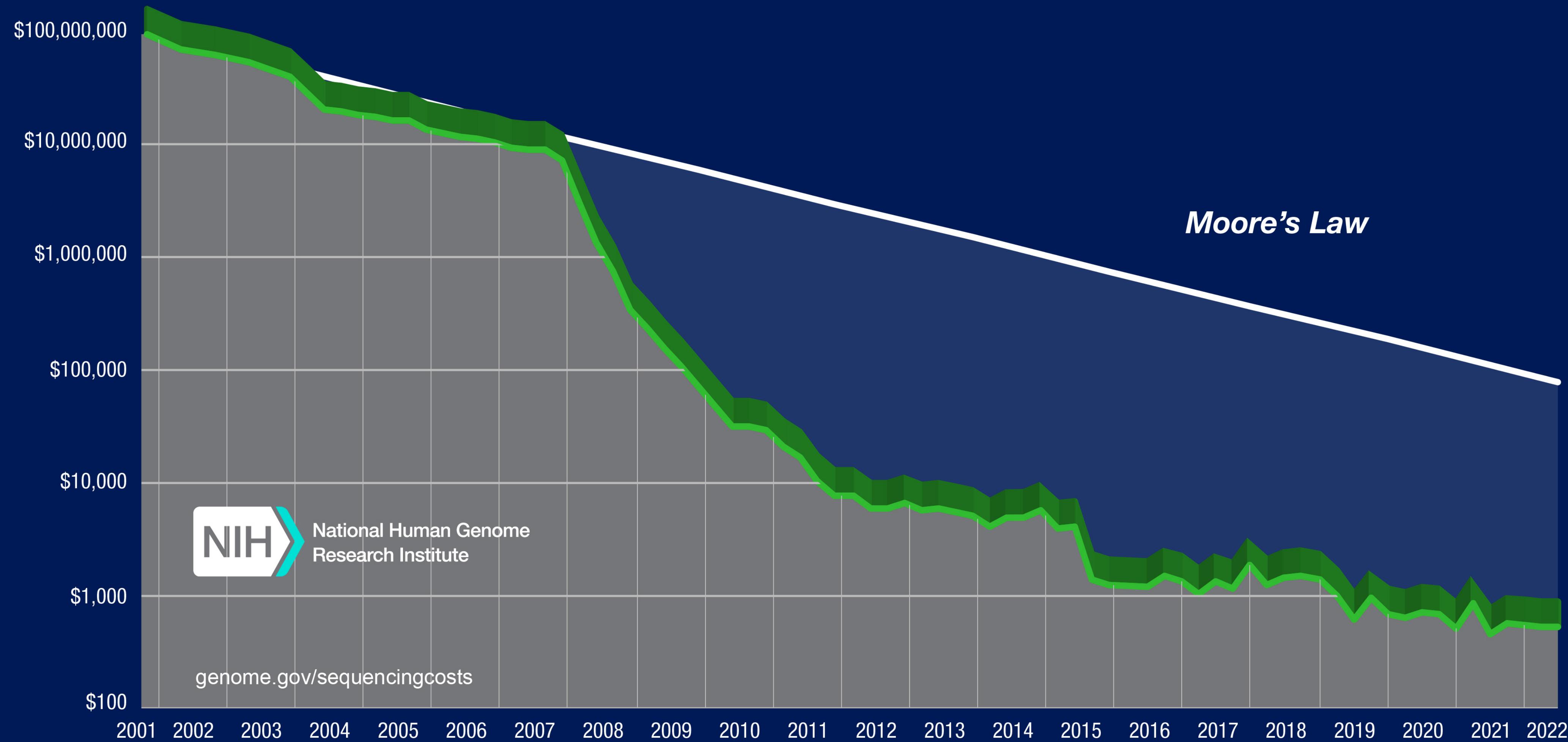
Graphic adapted from Mark Gerstein (GersteinLab.org; @markgerstein)

# Nobody is perfect (?)

A 10.7 Mb Interstitial Deletion of 13q21 Without Phenotypic Effect Defines a Further Non-Pathogenic Euchromatic Variant  
Andreas Roos, Miriam Elbracht, Michael Baudis, Jan Senderek, Nadine Schönherr, Thomas Eggemann, and Herdit M. Schüler  
*American Journal of Medical Genetics Part A* 146A:2417 – 2420 (2008)



## *Cost per Human Genome*



<https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>



nature

<https://doi.org/10.1038/s41586-021-04103-z>

## Accelerated Article Preview

# Exome sequencing and analysis of 454,787 UK Biobank participants

Received: 9 July 2021

Accepted: 6 October 2021

Accelerated Article Preview Published  
online 18 October 2021

Cite this article as: Backman, J. D. et al.  
Exome sequencing and analysis of 454,787  
UK Biobank participants. *Nature*  
<https://doi.org/10.1038/s41586-021-04103-z>  
(2021).

Joshua D. Backman, Alexander H. Li, Anthony Marcketta, Dylan Sun, Joelle Mbatchou, Michael D. Kessler, Christian Benner, Daren Liu, Adam E. Locke, Suganthi Balasubramanian, Ashish Yadav, Nilanjana Banerjee, Christopher Gillies, Amy Damask, Simon Liu, Xiaodong Bai, Alicia Hawes, Evan Maxwell, Lauren Gurski, Kyoko Watanabe, Jack A. Kosmicki, Veera Rajagopal, Jason Mighty, Regeneron Genetics Center, DiscovEHR, Marcus Jones, Lyndon Mitnaul, Eli Stahl, Giovanni Coppola, Eric Jorgenson, Lukas Habegger, William J. Salerno, Alan R. Shuldiner, Luca A. Lotta, John D. Overton, Michael N. Cantor, Jeffrey G. Reid, George Yancopoulos, Hyun M. Kang, Jonathan Marchini, Aris Baras, Gonçalo R. Abecasis, Manuel A. Ferreira

# 200+ Genomic Data Initiatives Globally

Clinical/Genomic  
Medicine



Research



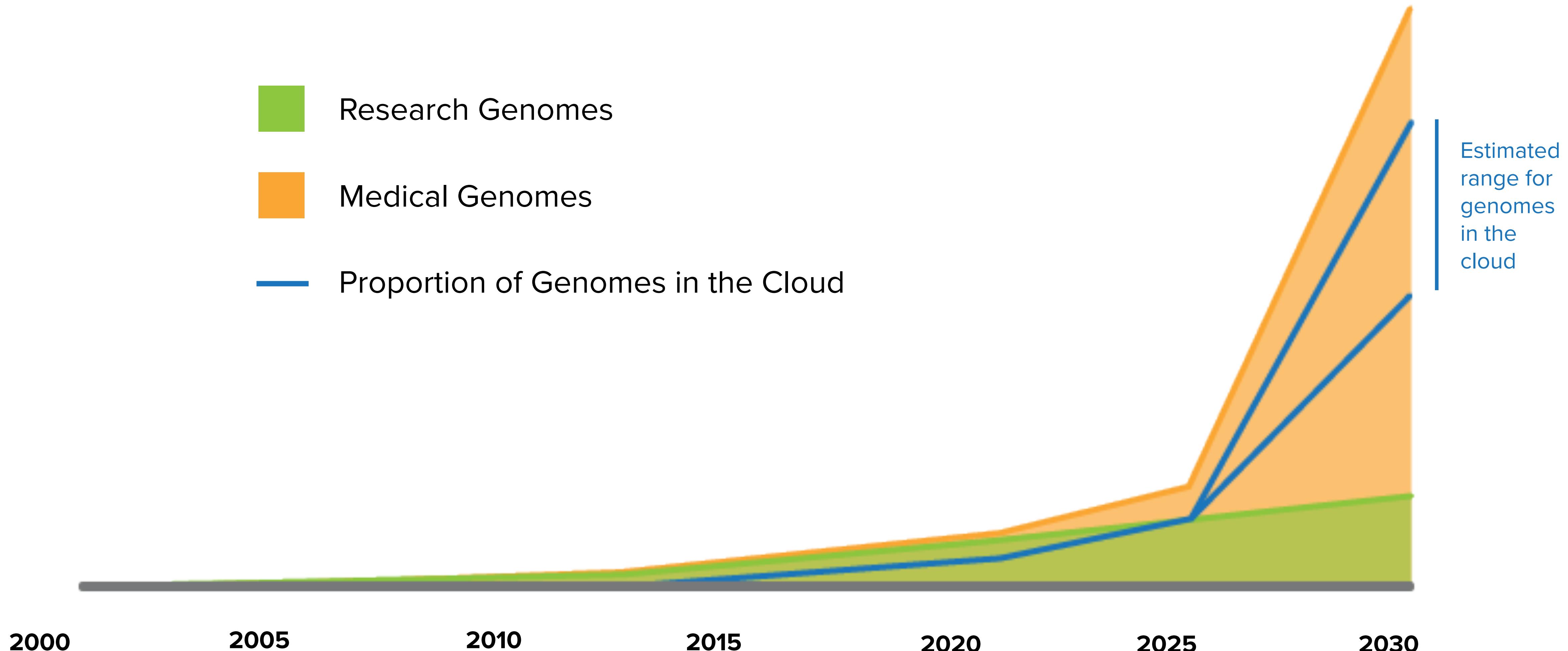
National



Cohorts



# How Many Genomes?



# How Many Genomes?



**RESEARCH**



**HEALTHCARE**

**60M** individuals  
**132.5M** sequences



**CLINICAL TRIALS**

**2.7-3M** individuals



**COHORTS**

**140M** individuals

# National Medical Genome Projects and Cohorts (2018)



## Medical Genomes

- Countries with active national medical genome projects
- Countries with some activity of medical genomics
- Countries planning medical genome projects

## Cohorts

- National cohorts > 100k genotyped or sequenced at least 25k
- National cohorts > 100k people active collection now
- Planning national cohorts > 100k

# Total worldwide sequencing capability?!

## Deployed sequencers of major platforms

- deployed sequencers of major platforms (estimate end of 2021)
- "While 3 Exabases seems like a awful lot, it's worth noting that this still isn't enough to sequence every human born.  
Somewhere in the region of 4 children are born a second, our 100Gb/s wouldn't even let us fully sequence one of them (at 30x)" - Nava Whiteford

Platform	Estimated Instruments	Runs/Week	Total Runs/Year	Run Yield (Tb)	Total Sequencing Capacity Tb/year
ONT MinION	5501	3	858156	0.05	42907.8
ONT GridION	782	3	121992	0.25	30498
ONT PromethION 48	67	2	6968	14	97552
Illum Novaseq 6000	1485	3	231660	6	1389960
Illum NextSeq	5430	3	847080	0.36	304948.8
Illum Miseq/Mini/iSeq	12340	3	1925040	0.015	28875.6
Ion Torrent	2220	14	1616160	0.05	80808
PacBio	577	5	150020	0.03	4500.6
MGI - Mid/Low	2000	3	312000	0.72	224640
MGI -T7	10	5	2600	6	15600
Total	30412		6071676		2220290.8

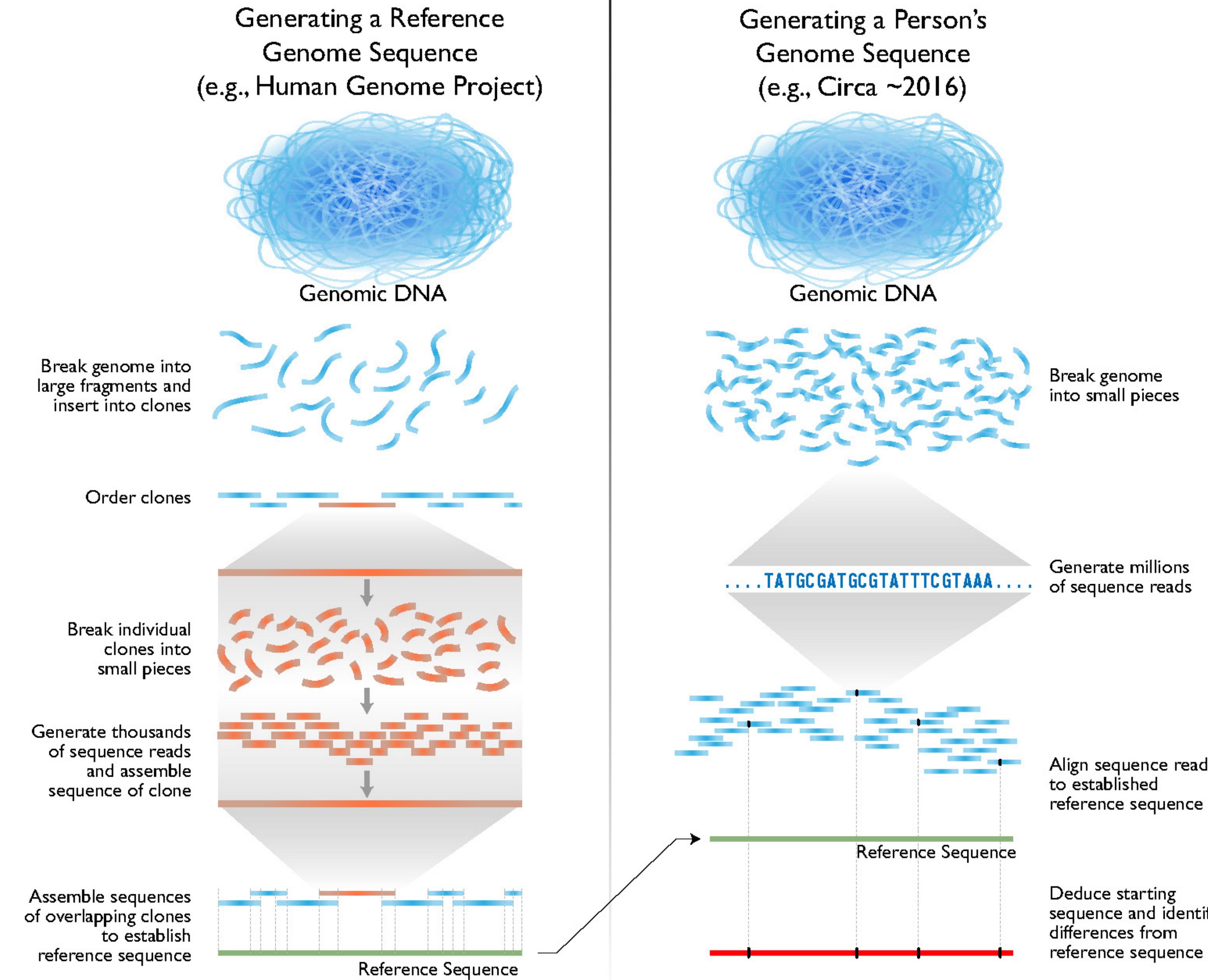
2331523584 Samples  
73.9 Samples/s

70.4 Gb/s

Tb/year

Estimated sequencing capability if everything would run continuously at full speed...

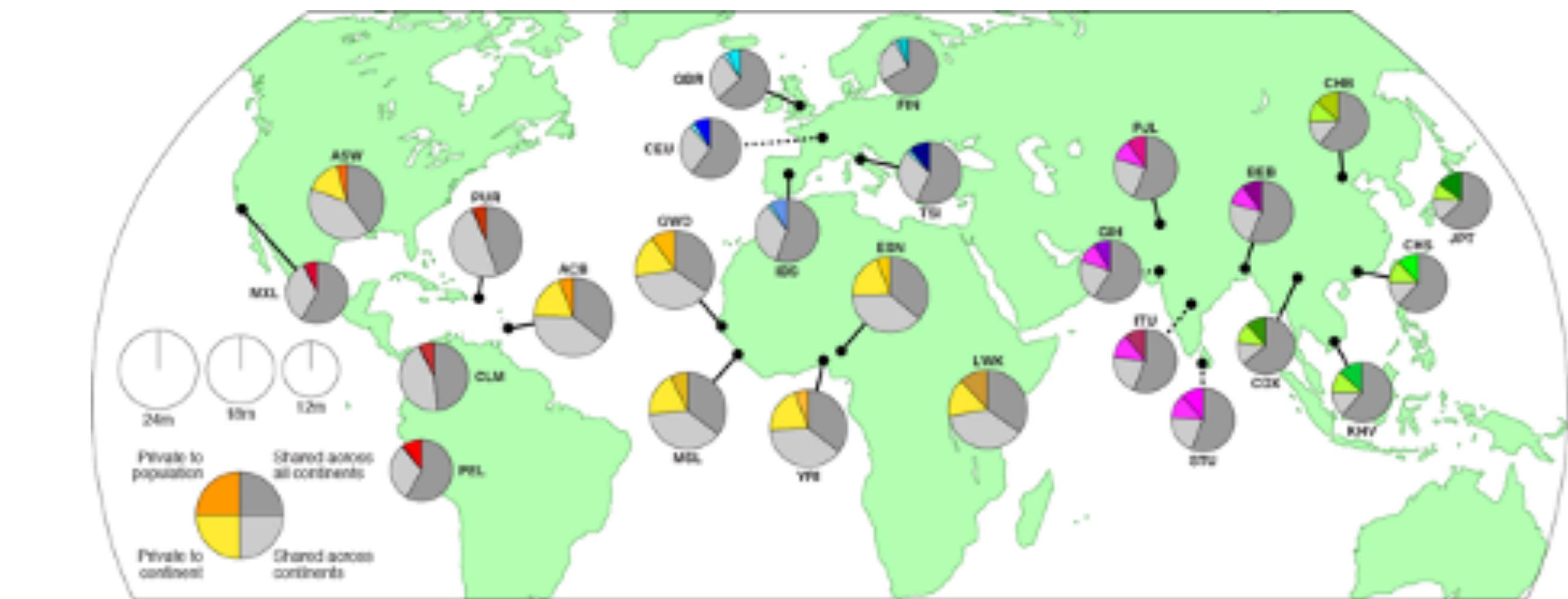
## Human Genome Sequencing



## New technologies and new applications



Global Alliance  
for Genomics & Health



Oxford Nanopore @nanopore · 5 Oct

0:09

91 minutes doesn't sound like long does it? Well, it's enough time for Luna Dijrackor and team to characterise a brain tumour during neurosurgery using #MinION — see for yourself #anythinganyoneanywhere #realrealtime ...



University of  
Zurich<sup>UZH</sup>

# BIO392

# Bioinformatics of Genome Variations

## Genomic Analysis Technologies...

Michael Baudis **UZH SIB**  
Computational Oncogenomics

# Non-Sequence™ Genomics (Molecular) Cytogenetics

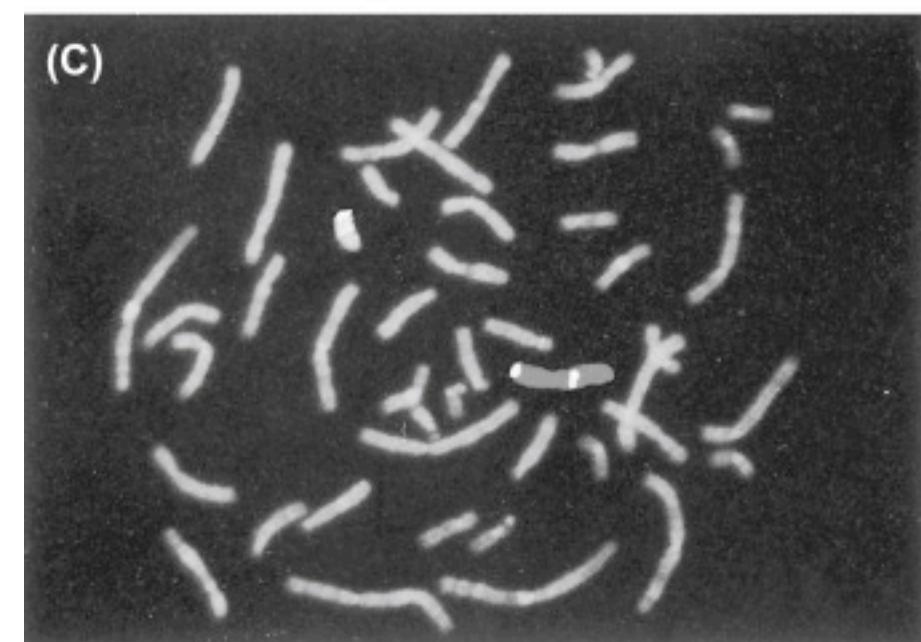
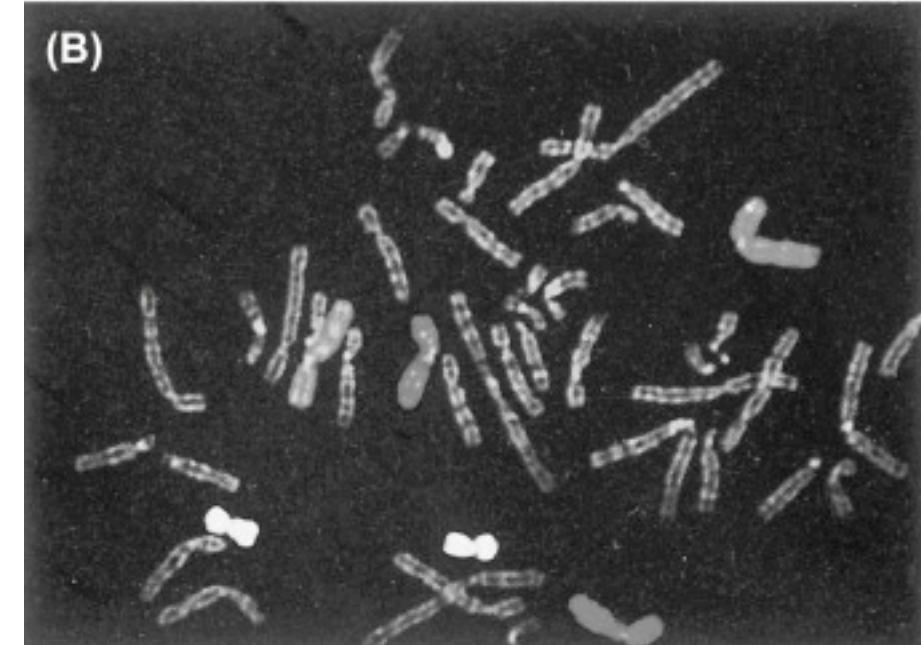
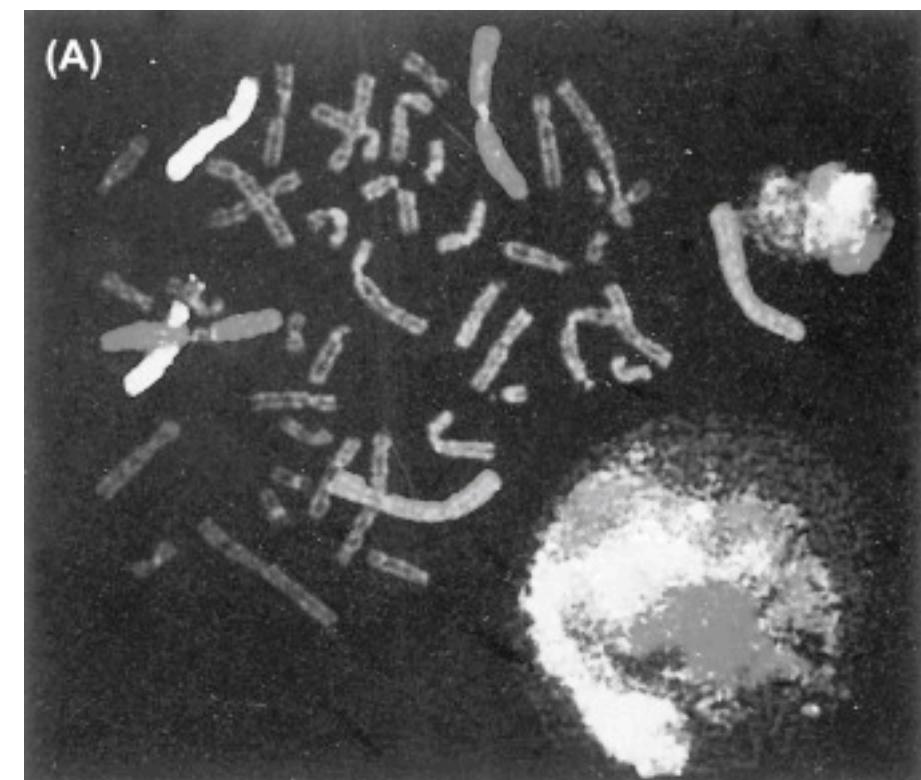
- structural changes in chromosomes can be analyzed by cytogenetics (chromosomal analysis) w/o knowledge about sequence alterations
- cytogenetics may be modified by
  - hybridization of fluorescent (or radiolabelled) probes of known DNA content or localization
  - using of fluorescently labeled "painting" libraries
  - reversing the hybridization - hybridizing DNA of interest to known metaphases

→ **Molecular Cytogenetics**



G-banded metaphase with three marker chromosomes (whose origin from chromosomes 22, 11, and 14 cannot be identified by this technique).

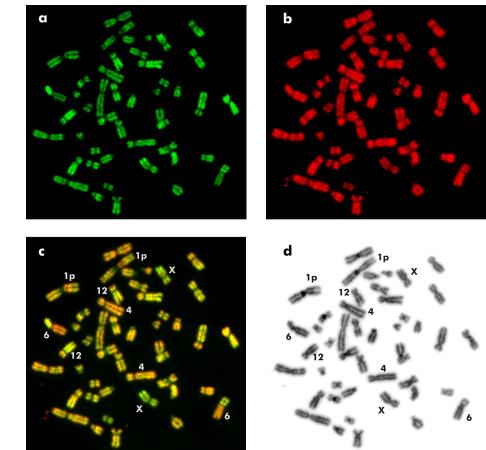
Images from "Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics"  
Chapter 5 - Cytogenetic Analysis -  
Nancy B. Spinner & Malcolm A. Ferguson-Smith



Chromosome painting using chromosome-specific probes (CAMBIO Ltd., Cambridge, England) from flow-sorted chromosomes (DAPI counterstain). (A) Chromosome 1 (red), chromosome 2 (green), and chromosome 6 (yellow). Interphase nucleus reveals chromosome domains within nucleus. (B) Chromosome 7 (green), chromosome 11 (red), and chromosome 20 (yellow). (C) X chromosome (red), Y chromosome (green). Note Y signal (yellow) on XY homologous regions of Xp (tip) and Xq (proximal third) and X signal (yellow) on XY homologous region of Yp (tip).

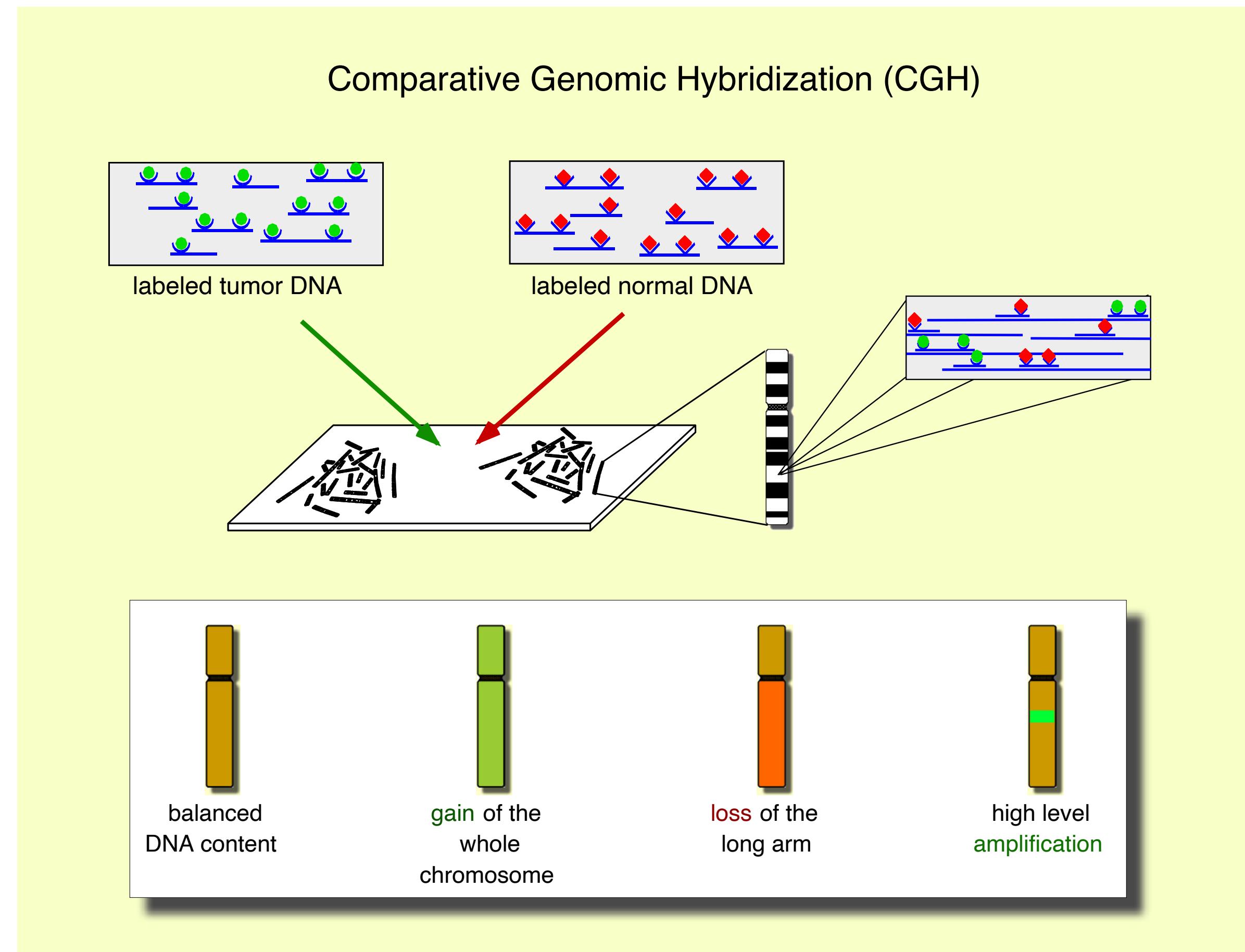
# Comparative Genomic Hybridization

## Molecular-Cytogenetic Technology for Genomic Imbalance Screening



- Molecular-cytogenetic technique to identify regional genomic copy number variations (CNV/CNA)
- based on ***in situ*** suppression **hybridization** of labeled **genomic** tumor and reference DNA against a karyotypically normal metaphase chromosomes
- analysis of relative fluorescence ratio allows **semi-quantitative copy number** read-out
- **indirect** attribution of involved target genes through cytogenetic bands (megabase resolution)

- Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*. 1992;258(5083):818-821.
- Joos S, Scherthan H, Speicher MR, Schlegel J, Cremer T, Lichter P. Detection of amplified DNA sequences by reverse chromosome painting using genomic tumor DNA as probe. *Hum Genet*. 1993;90:584-589.

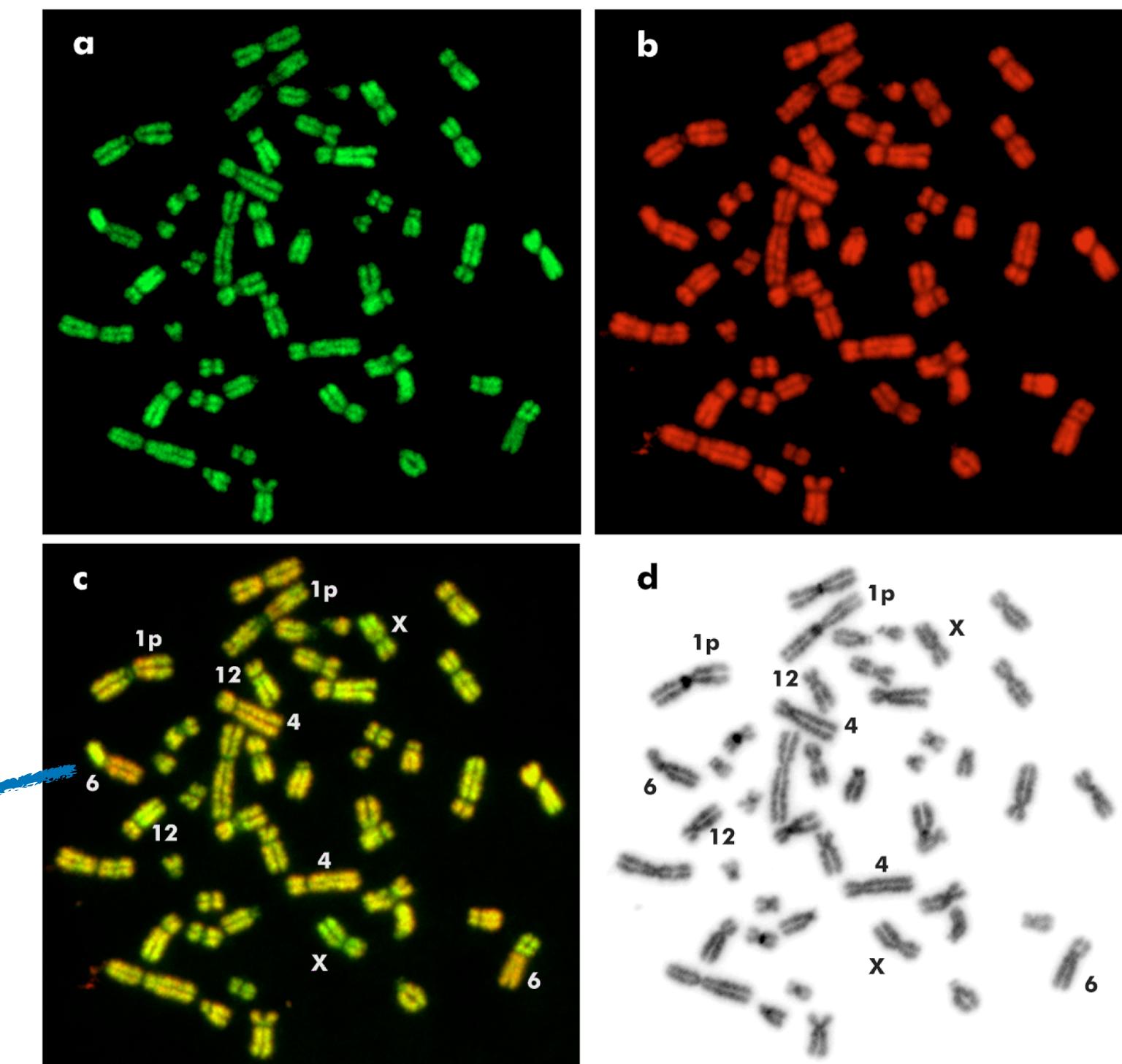


Chromosomal CGH: Normal metaphase spreads (cultured lymphocytes from healthy donors) on microscopy slides serve as the hybridization matrix for whole-genome DNA from tumor and reference tissue, labeled with different fluorophores. The regional ratio between the two colors points to (relative) changes in the copy number in the tumor DNA. Michael Baudis, 1998

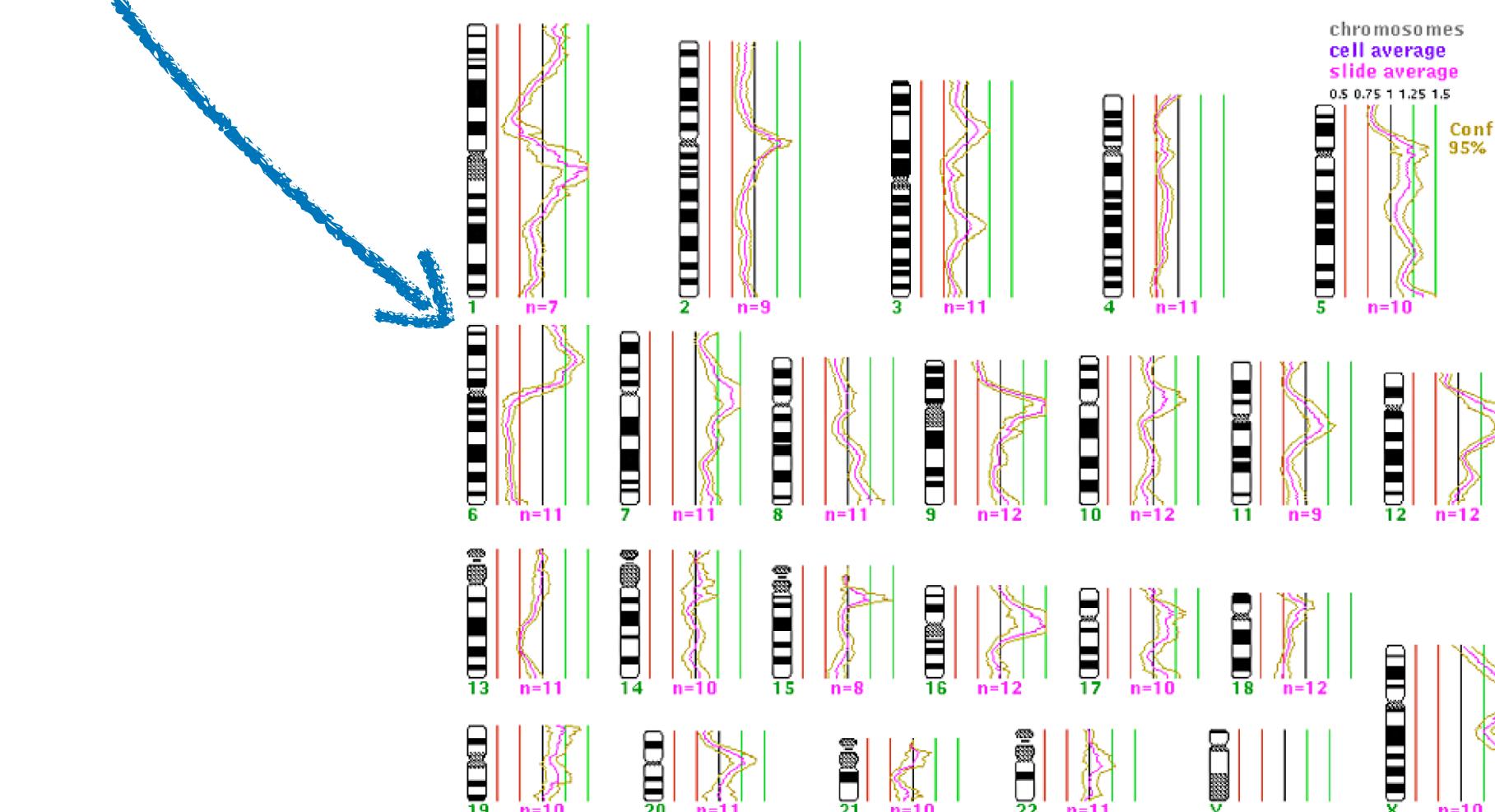
# Comparative Genomic Hybridization

## Molecular-Cytogenetic Technology for Genomic Imbalance Screening

- Molecular-cytogenetic technique to identify regional genomic copy number variations (CNV/CNA)
- based on ***in situ*** suppression **hybridization** of labeled **genomic** tumor and reference DNA against a karyotypically normal metaphase chromosomes
- analysis of relative fluorescence ratio allows **semi-quantitative copy number** read-out
- **indirect** attribution of involved target genes through cytogenetic bands (megabase resolution)



CGH-Experiment: **a** Hybridisierung mit Tumor-DNA; **b** Hybridisierung mit normaler menschlicher DNA als Kontrolle; **c** Überlagerung der Signale; **d** Bänderungsfärbung zur Identifizierung der Chromosomen



Auswertung: Summationsprofil der computergestützten Analyse mehrerer Metaphasen des dargestellten Falles; die Profilausschläge stehen für Zugewinne bzw. Verluste von chromosomalen Anteilen im Tumorgenom

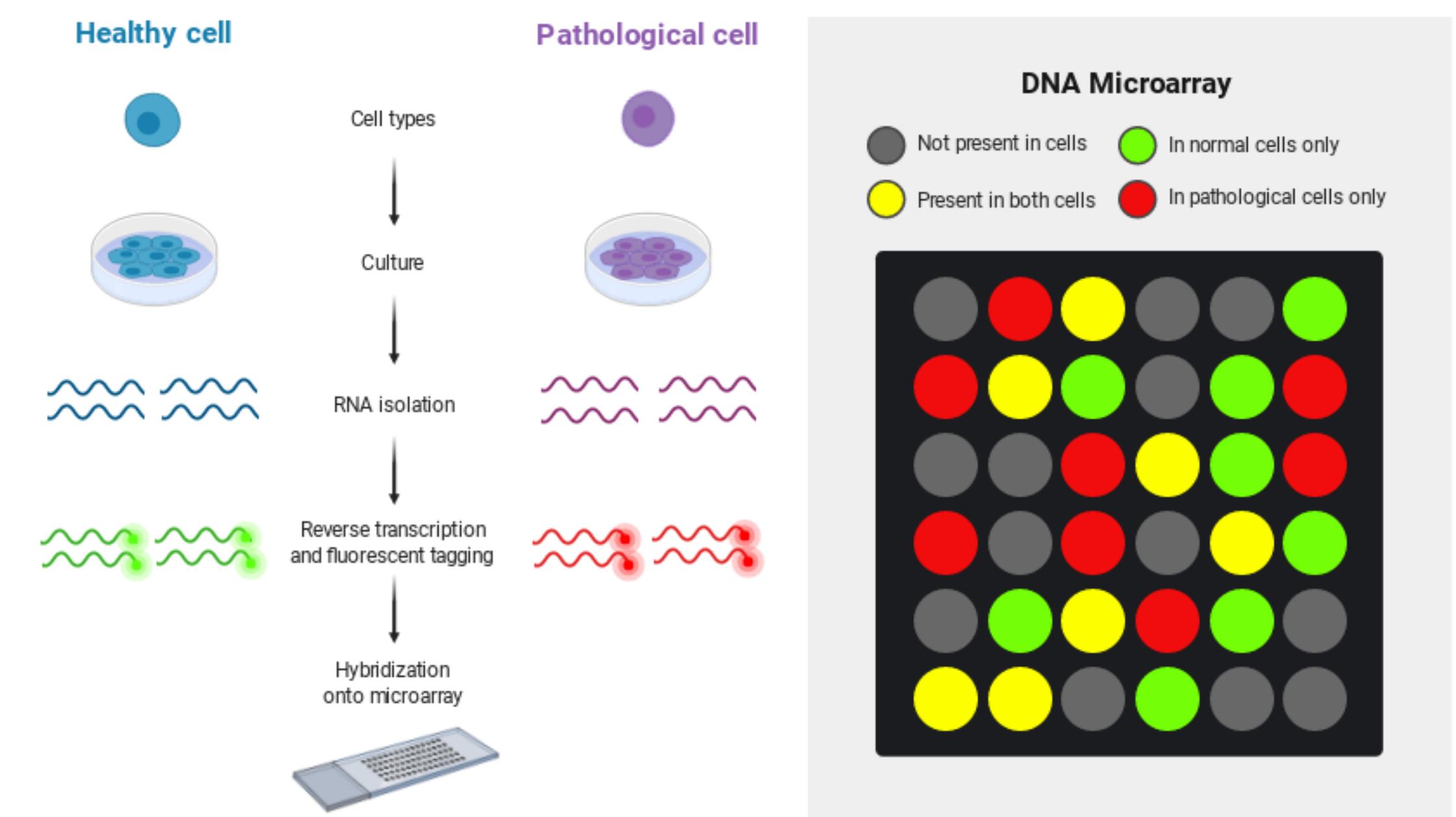
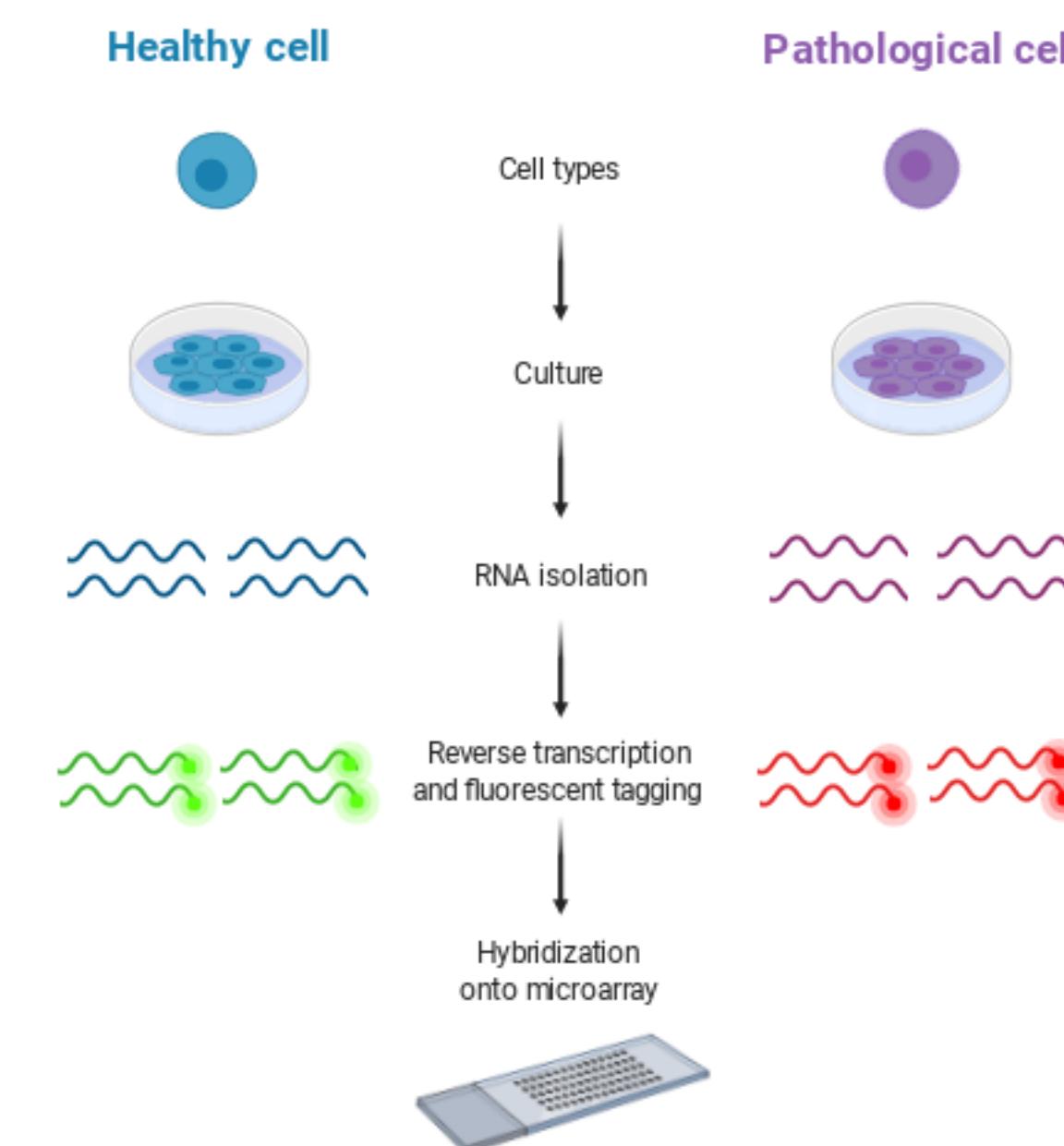
- Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 1992;258:818-821.
- Joos S, Scherthan H, Speicher MR, Schlegel J, Cremer T, Lichter P. Detection of amplified DNA sequences by reverse chromosome painting using genomic tumor DNA as probe. Hum Genet. 1993;90:584-589.

# Array Comparative Genomic Hybridization

## Molecular Hybridization Technology for Genomic Imbalance Screening

- Molecular-cytogenetic technique to identify genomic copy number variations (CNV/CNA) for given sequences (multi-kb to Mb)
- based on ***in situ*** suppression **hybridization** of labeled **genomic** tumor and reference DNA against spotted or synthesized DNA clones or oligonucleotides
- analysis of relative fluorescence ratio allows **semi-quantitative copy number** read-out
- **direct** attribution of involved target genes through known sequence content

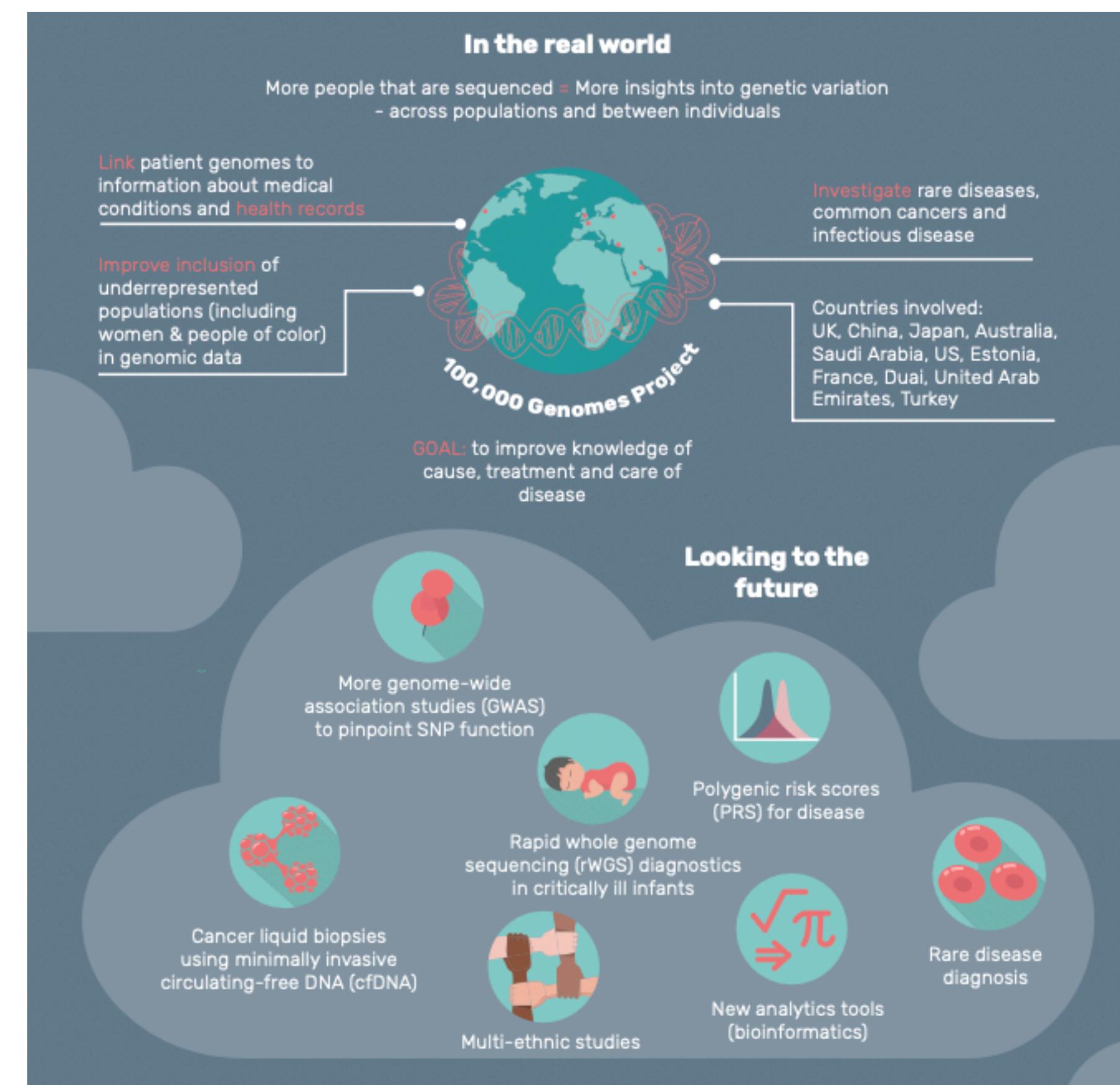
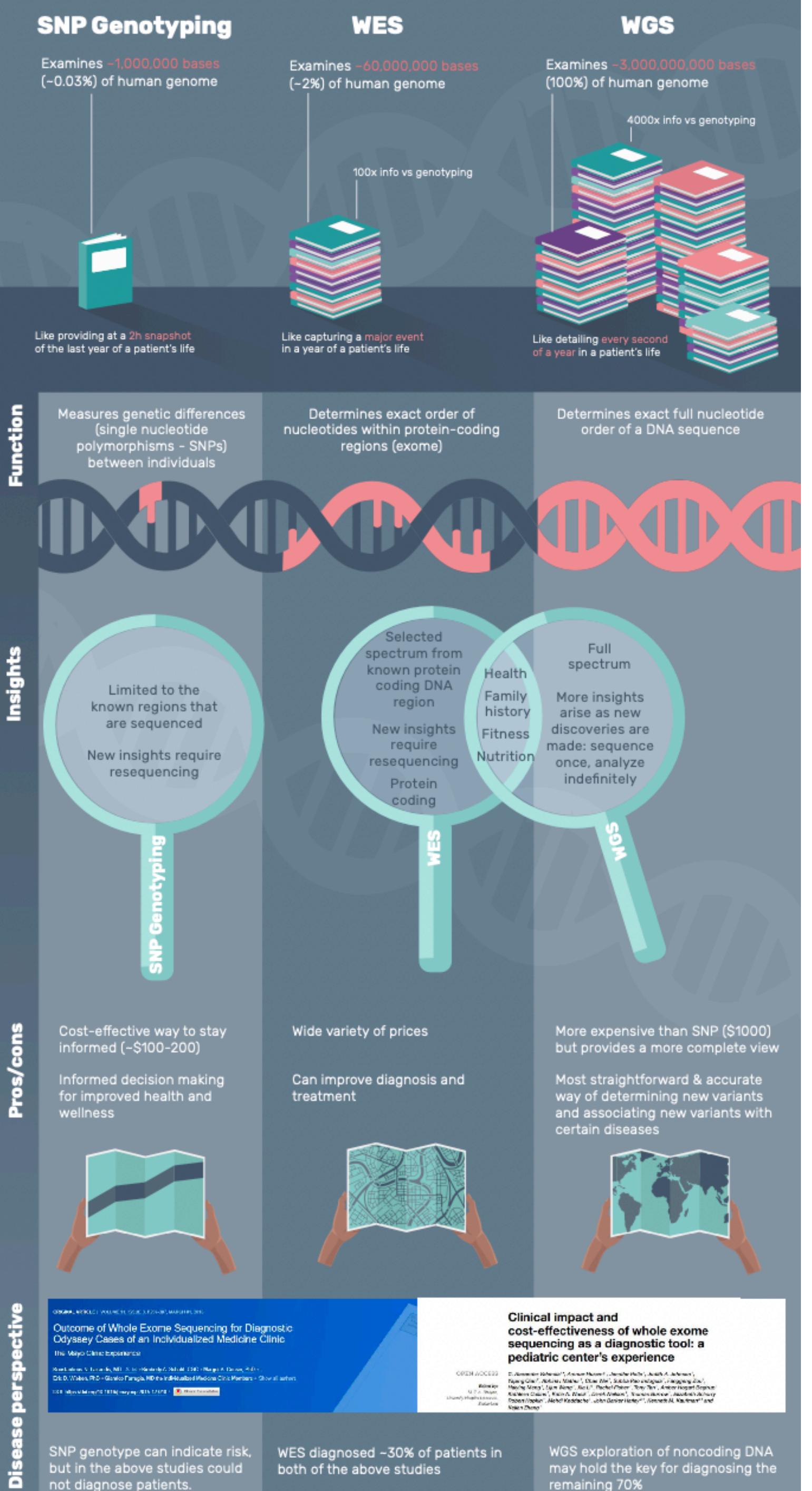
- Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer.* 1997; 4 (20):399-407.
- Albertson DG, Pinkel D. Genomic microarrays in human genetic disease and cancer. *Hum Mol Genet.* 2003; 12 Spec No 2 R145-52.
- Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR, Futreal PA, Weber B, Shapero MH, Wooster R. High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.* 2004; 14 (2):287-295.



Reprinted from "DNA Microarray", August 2019, retrieved from <https://app.biorender.com/biorender-templates/figures/all/t-5e41b61b0dd2690088b72481-dna-microarray> © 2022 by BioRender.

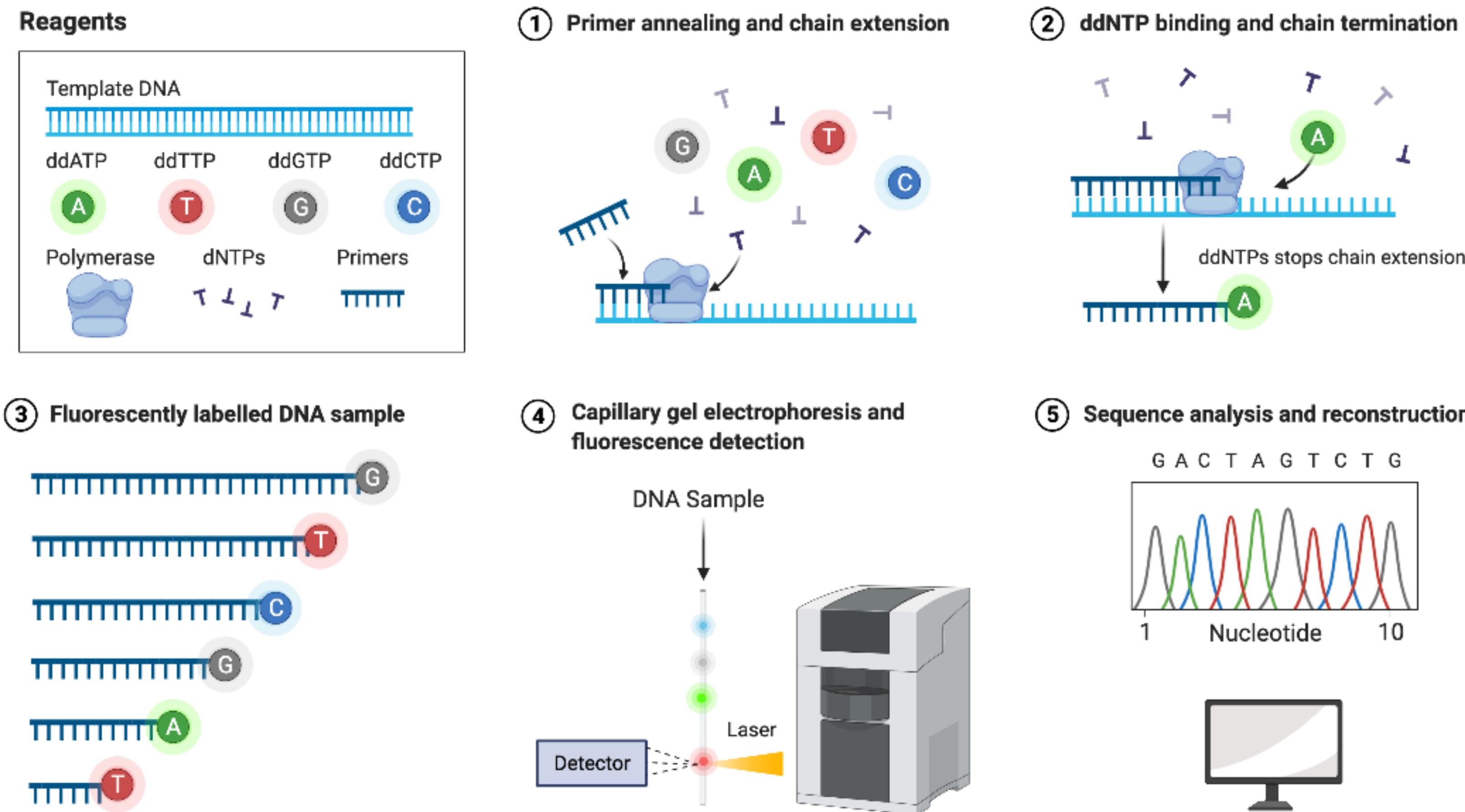
# Genome Analysis

## A “progressing technologies” view



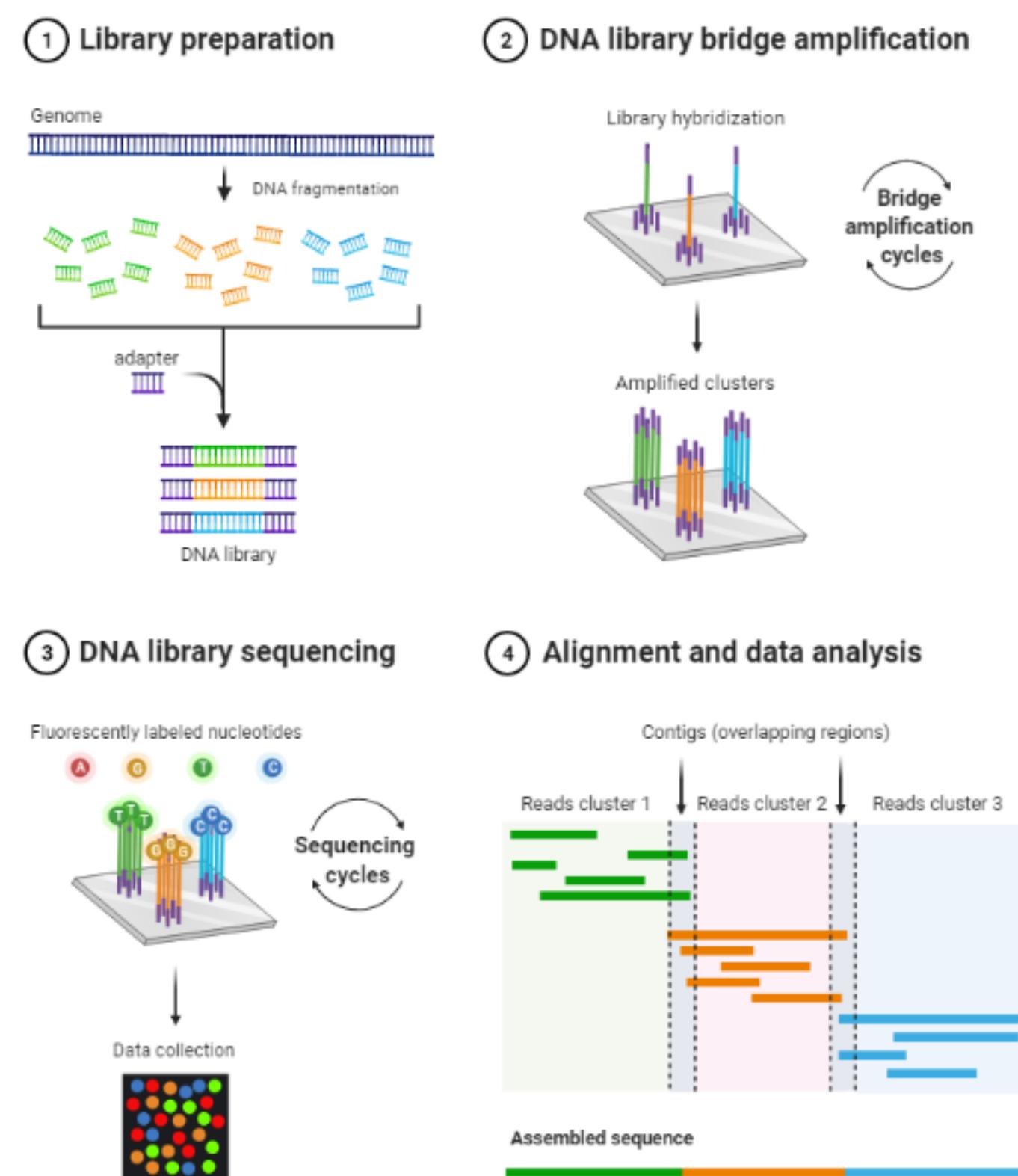
# Genome Analysis

## Sanger Sequencing

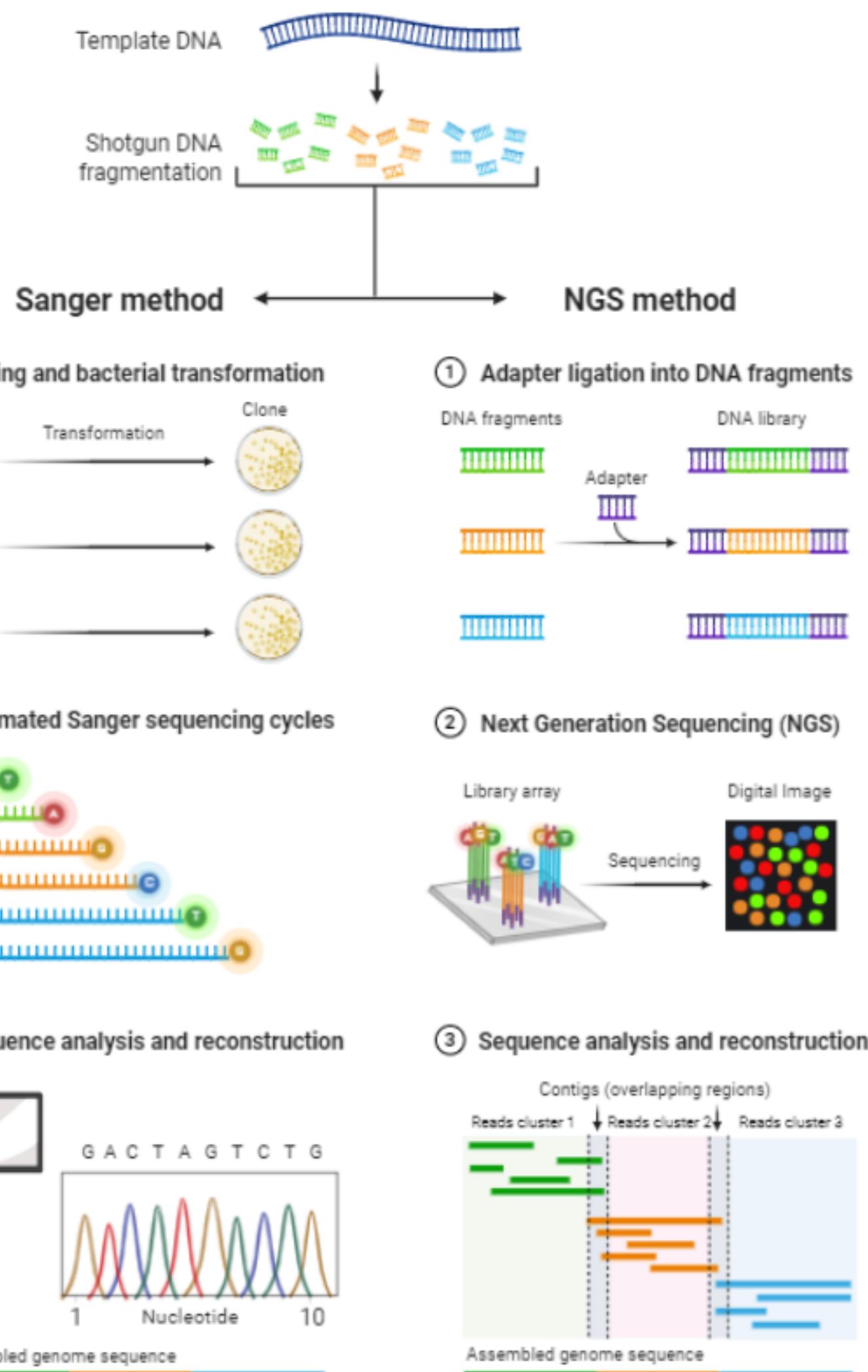


# Genome Analysis

## NGS vs. Sanger Sequencing



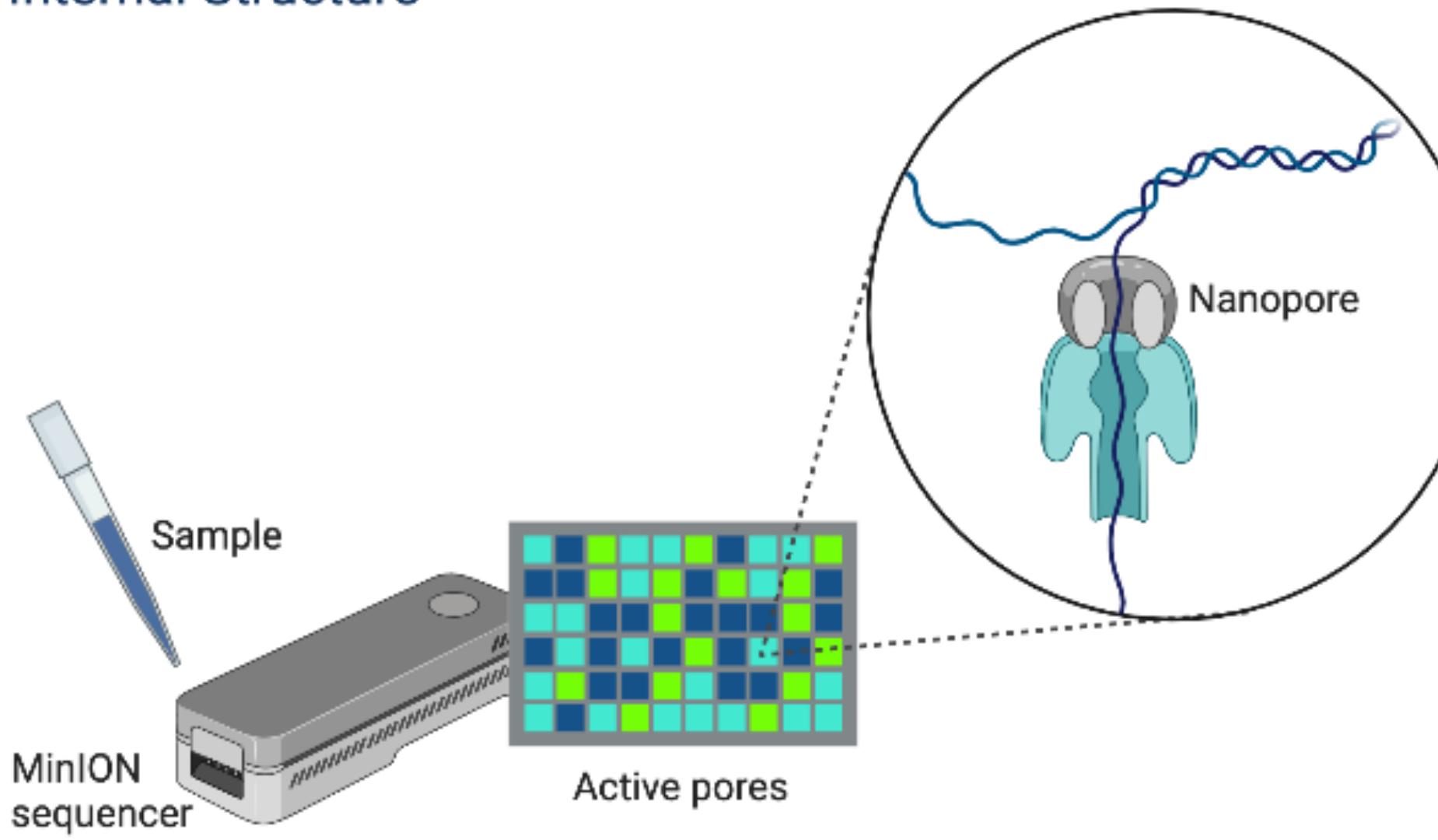
### Shotgun Sequencing Sanger vs NGS



# Genome Analysis

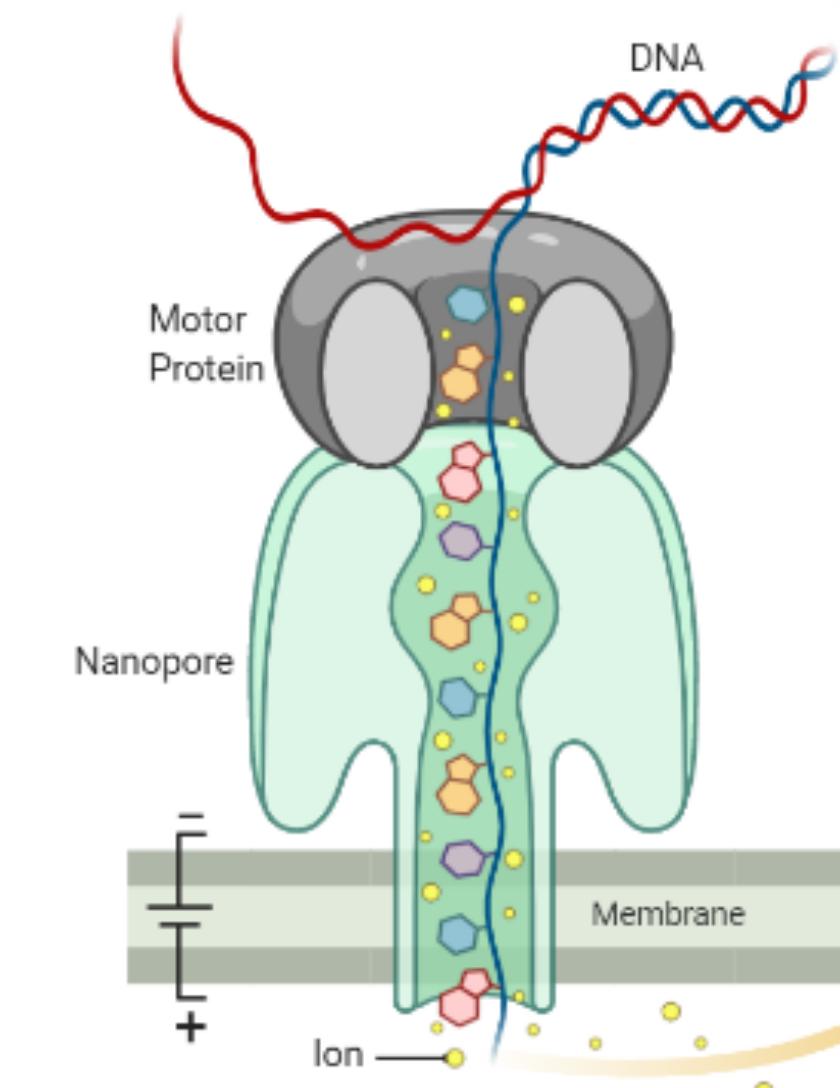
## Nanopore Sequencing

### MinION Sequencer Internal Structure

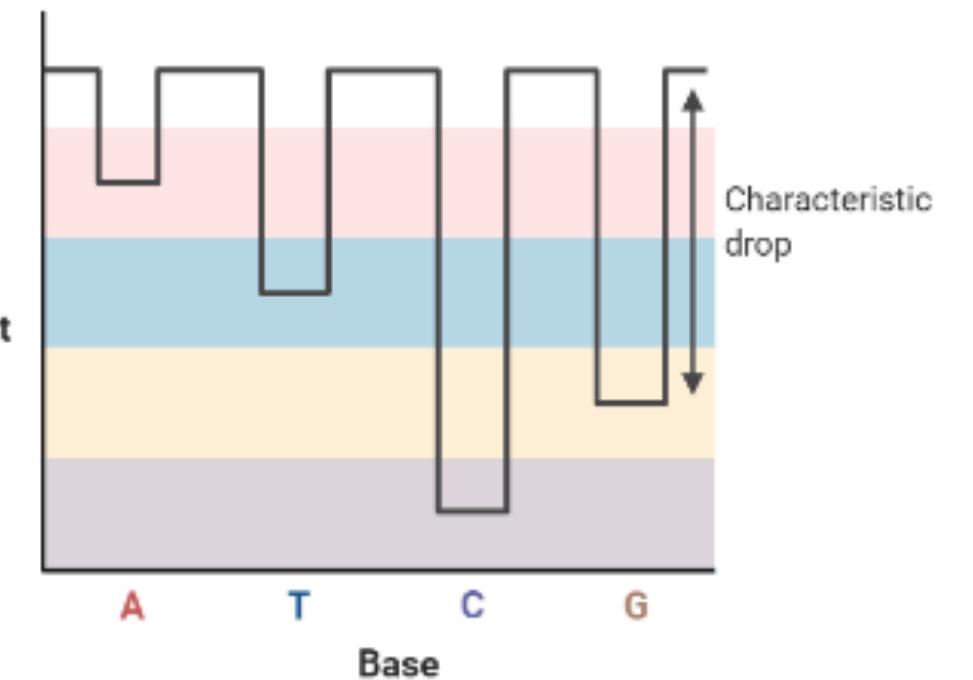


### Nanopore Sequencing

- 1 DNA is unwound by the motor protein and one strand is translocated through the pore to the +ve side of membrane



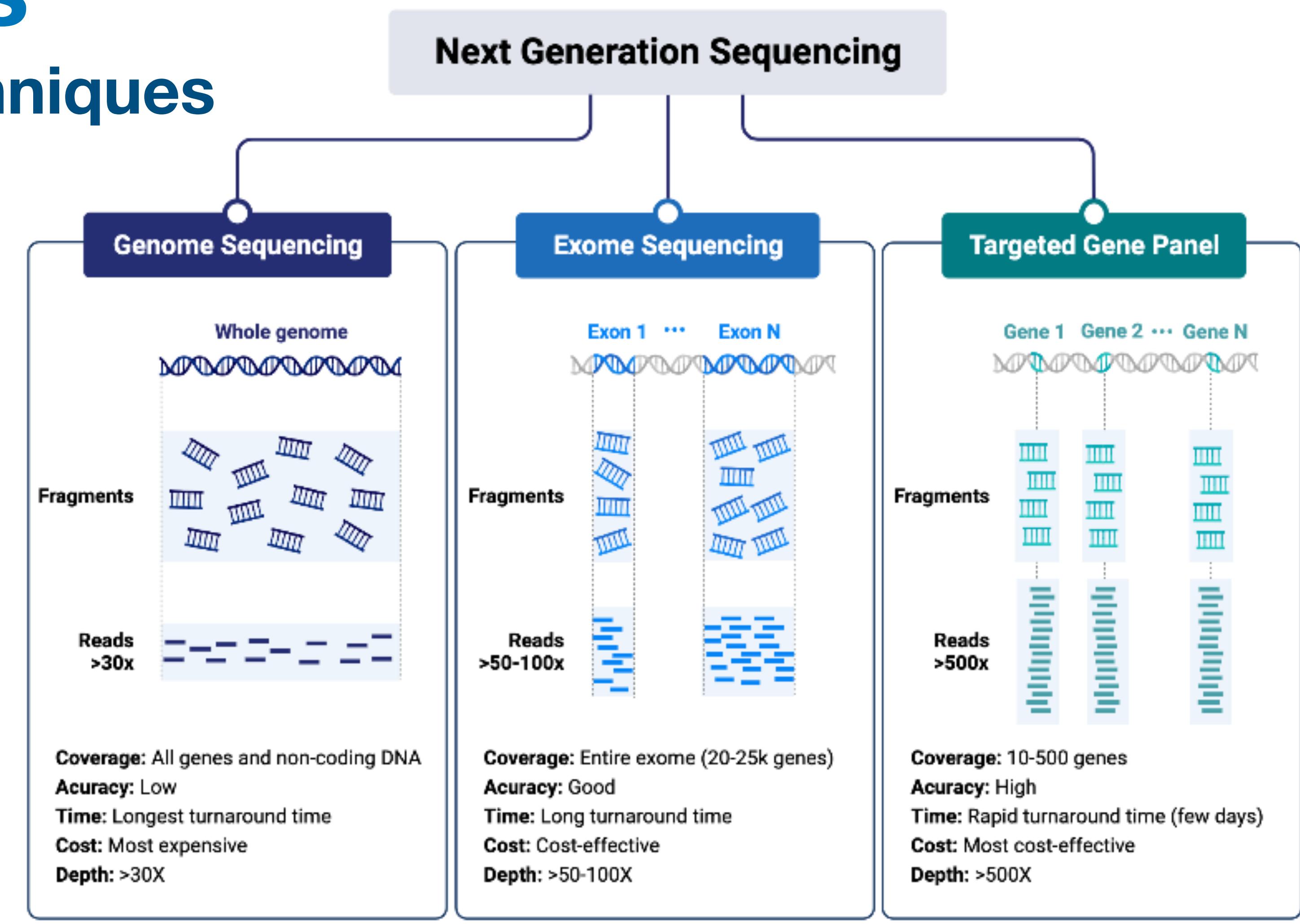
- 2 Each base gives a characteristic reduction in the ionic current, allowing the DNA to be sequenced



# Genome Analysis

## Comparison of NGS Techniques

- current NGS technologies present various compromises between coverage, precision, speed & accuracy
- the bottleneck in WGS is not in read production costs but in interpretation and storage
- widespread use of incomplete genome profiling technologies avoids problems of hard to interpret or not "actionable" results but limits the expansion of our "knowledge horizon"



*Template adapted from: Dr. Roshini Abraham  
Clinical Immunologist at Nationwide Children's Hospital*

# Genomic File Formats

Types | Sizes | Use Cases

# What is a PB, for human genomes?

It depends...

- 2 bits per base are sufficient to encode TCGA
  - using 00, 01, 10, 11
  - [TCGA]{3'000'000'000}
  - $2 * 3 * 10^9 b = 6,000,000,000 b$
  - perfect genome (no overhead): ~715 MB
  - 1PB => ~1'400'000 genomes
- according to Swiss online store (Sep 2021) ~35'000CHF (65x16TB á CHF550)
- this is less than a PhD position per year in Switzerland ...
- (real costs are 2x that, + duplication, facilities, service ... => ~500'000CHF)
- **However: A single 30x BAM file => 100GB**
- Still: 500'000CHF => 1PB => 10'000 genomes => 50CHF/genome (BAM format)



# Genomic File Formats



# Genomic File Formats

- **text** or **binary** file formats, optimised for specific types of biological data
- examples from genomics:
  - **BAM** - binary version of Sequence Alignment/Map (SAM)
  - CRAM - compressed version of BAM with multiple optimization and differential access options
  - **BED** (Browser Extensible Data) -flexible way to define the data lines in an genome browser annotation tracks
  - **VCF** (Variant Call Format)

The file info window displays the following details:

- File name: GSM1904006.CEL
- Size: 69.1 MB
- Modified: 3 February 2016 at 17:46
- Kind: FLC animation
- Where: arrayRAID → arraymapln → affyRaw → GSE73822 → GPL6801
- Created: 3 February 2016 at 17:46
- Modified: 3 February 2016 at 17:46
- Stationery pad
- Locked

Below the file info window, a list of genomic file formats is shown:

- Axt format
- BAM format
- BED format
- BED detail format
- bedGraph format
- barChart and bigBarChart format
- bigBed format
- bigGenePred table format
- bigPsl table format
- bigMaf table format
- bigChain table format
- bigWig format
- Chain format
- CRAM format
- GenePred table format
- GFF format
- GTF format
- HAL format
- MAF format
- Microarray format
- Net format
- Personal Genome SNP format
- PSL format
- VCF format
- WIG format

At the bottom, a BED file example is provided:

```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB Demo"
chr7 127471196 127472363 Pos1 0 + 127472363
chr7 127472363 127473530 Pos2 0 + 127472363 127473530 255,0,0
chr7 127473530 127474697 Pos3 0 + 127473530 127474697 255,0,0
chr7 127474697 127475864 Pos4 0 + 127474697 127475864 255,0,0
chr7 127475864 127477031 Neg1 0 - 127475864 127477031 0,0,255
chr7 127477031 127478198 Neg2 0 - 127477031 127478198 0,0,255
chr7 127478198 127479365 Neg3 0 - 127478198 127479365 0,0,255
chr7 127479365 127480532 Pos5 0 + 127479365 127480532 255,0,0
chr7 127480532 127481699 Neg4 0 - 127480532 127481699 0,0,255
```

BED file example

# SAM/BAM and related specifications

These documents are maintained by the Large Scale Genomics work stream of the Global Alliance for Genomics & Health ([GA4GH](#)). Information on GA4GH procedures and how to get involved is [available here](#). Lists of contributors and acknowledgements can generally be found in each individual specification document.

## Specifications:

- [SAM v1](#)
- [SAM tags](#)
- [CRAM v2.1](#)
- [CRAM v3.x](#)
- [CRAM codecs](#)
- [BCF v1](#)
- [BCF v2.1](#)
- [CSI v1](#)
- [Tabix](#)
- [VCF v4.1](#)
- [VCF v4.2](#)
- [VCF v4.3](#)
- [VCF v4.4](#)
- [BED v1](#)
- [crypt4gh](#)
- [Htsget](#)
- [Refget](#)

## Alignment data files

[SAMv1.tex](#) is the canonical specification for the SAM (Sequence Alignment/Map) format, BAM (its binary equivalent), and the BAI format for indexing BAM files. [SAMtags.tex](#) is a companion specification describing the predefined standard optional fields and tags found in SAM, BAM, and CRAM files. These formats are discussed on the [samtools-devel mailing list](#).

[CRAMv3.tex](#) is the canonical specification for the CRAM format, while [CRAMv2.1.tex](#) describes its now-obsolete predecessor. [CRAMcodecs.tex](#) contains details of the CRAM custom compression codecs. Further details can be found at [ENA's CRAM toolkit page](#) and [GA4GH's CRAM page](#). CRAM discussions can also be found on the [samtools-devel mailing list](#).

The [tabix.tex](#) and [CSIV1.tex](#) quick references summarize more recent index formats: the tabix tool indexes generic textual genome position-sorted files, while CSI is [htslib](#)'s successor to the BAI index format.

## Unaligned sequence data files

We do not define or endorse any dedicated unaligned sequence data format. Instead we recommend storing such data in one of the alignment formats (SAM, BAM, or CRAM) with the unmapped flag set. However for completeness, we list the commonest formats below with external links.

[FASTA](#) is an early sequence-only format originally defined by William Pearson's tool of the same name.

[FASTQ](#) was designed as a replacement for FASTA, combining the sequence and quality information in the same file. It has no formal definition and several incompatible variants, but is described in a paper by Cock et al.

## Variant calling data files

[VCFv4.4.tex](#) is the canonical specification for the Variant Call Format and its textual (VCF) and binary (BCF) encodings, while [VCFv4.1.tex](#), [VCFv4.2.tex](#) and [VCFv4.3.tex](#) describe their predecessors. These formats are discussed on the [vcftools-spec mailing list](#).

[BCFv1\\_qref.tex](#) summarizes the obsolete BCF1 format historically produced by [samtools](#). This format is no longer recommended for use, as it has been superseded by the more widely-implemented BCF2.

[BCFv2\\_qref.tex](#) is a quick reference describing just the layout of data within BCF2 files.

## Discrete genomic feature data files

[BEDv1.tex](#) is the canonical specification for the GA4GH Browser Extensible Data (BED) format.

## File encryption

[crypt4gh.tex](#) is the canonical specification of the crypt4gh format which can be used to wrap existing file formats in an encryption layer.

## Transfer protocols

[Htsget.md](#) describes the *hts-get* retrieval protocol, which enables parallel streaming access to data sharded across multiple URLs or files.

[Refget.md](#) enables access to reference sequences using an identifier derived from the sequence itself.

# The VCF file format

## Standard for genomic variant representation

### Example

VCF header

```

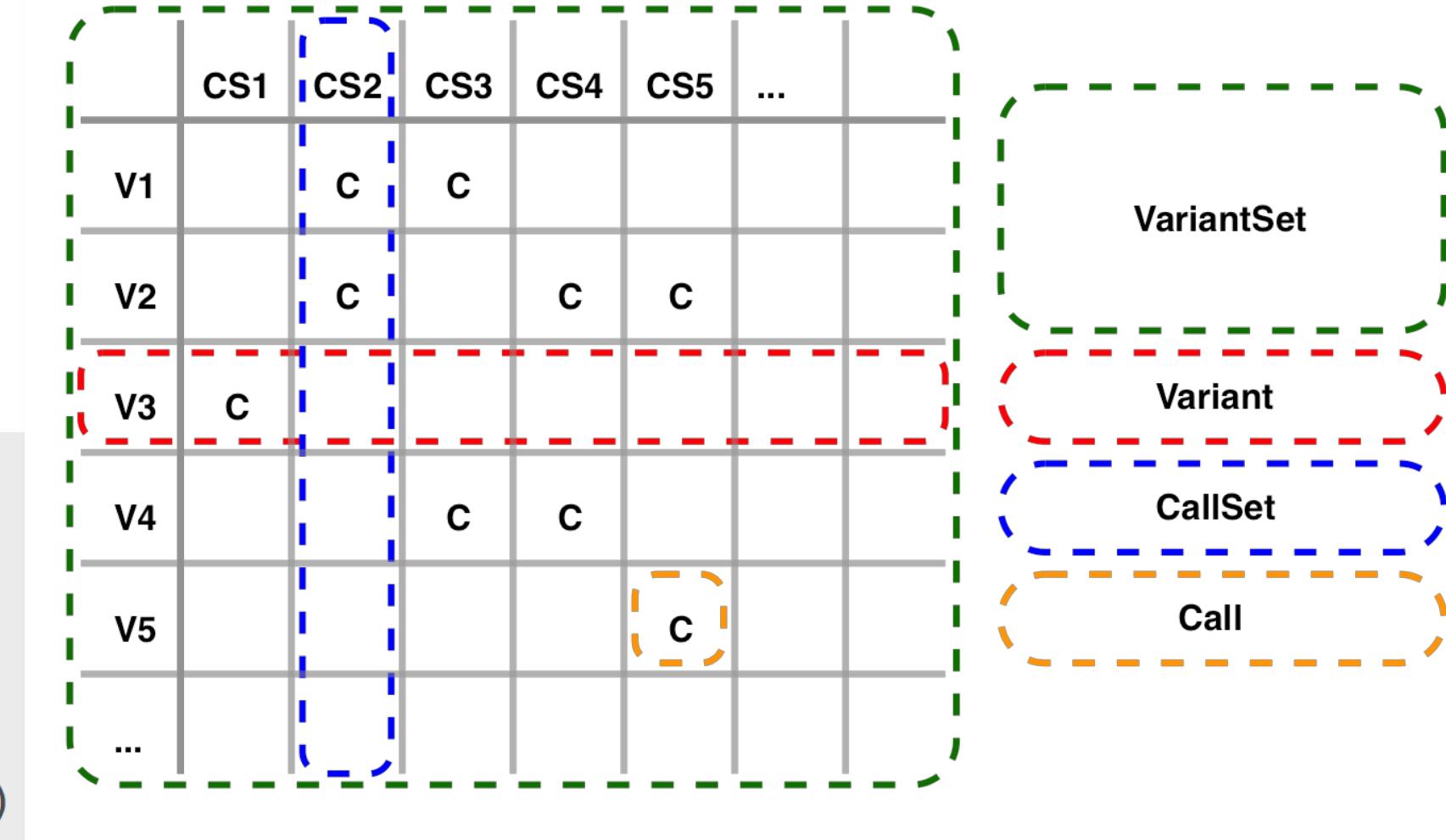
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT .
1 2 rs1 C T,CT .
1 5 . G <DEL> .
1 100 . T .

```

Body

Annotations:

- Deletion**: Variant at position 5 is a deletion (<DEL>).
- SNP**: Variant at position 2 is a SNP (rs1).
- Large SV**: Variant at position 5 is a large structural variant (SVTYPE=DEL; END=300).
- Insertion**: Variant at position 2 has an insertion (T).
- Other event**: Variant at position 5 is an other event (<DEL>).
- Mandatory header lines**: Lines starting with ##.
- Optional header lines (meta-data)**: Lines starting with ##INFO or ##FORMAT.
- Reference alleles (GT=0)**: Reference alleles for samples 1 and 2.
- Alternate alleles (GT>0 is an index to the ALT column)**: Alternate alleles for samples 1 and 2.
- Phased data**: Phased data for samples 1 and 2 across the variants.



Variant  
Call  
Format

- stores the results of a single or multiple interpretations of genome sequencing datasets, in comparison to a reference genome
- standard format for file-based storage of human genome variants

# Task: Estimate Storage Requirements for 1000 Genomes

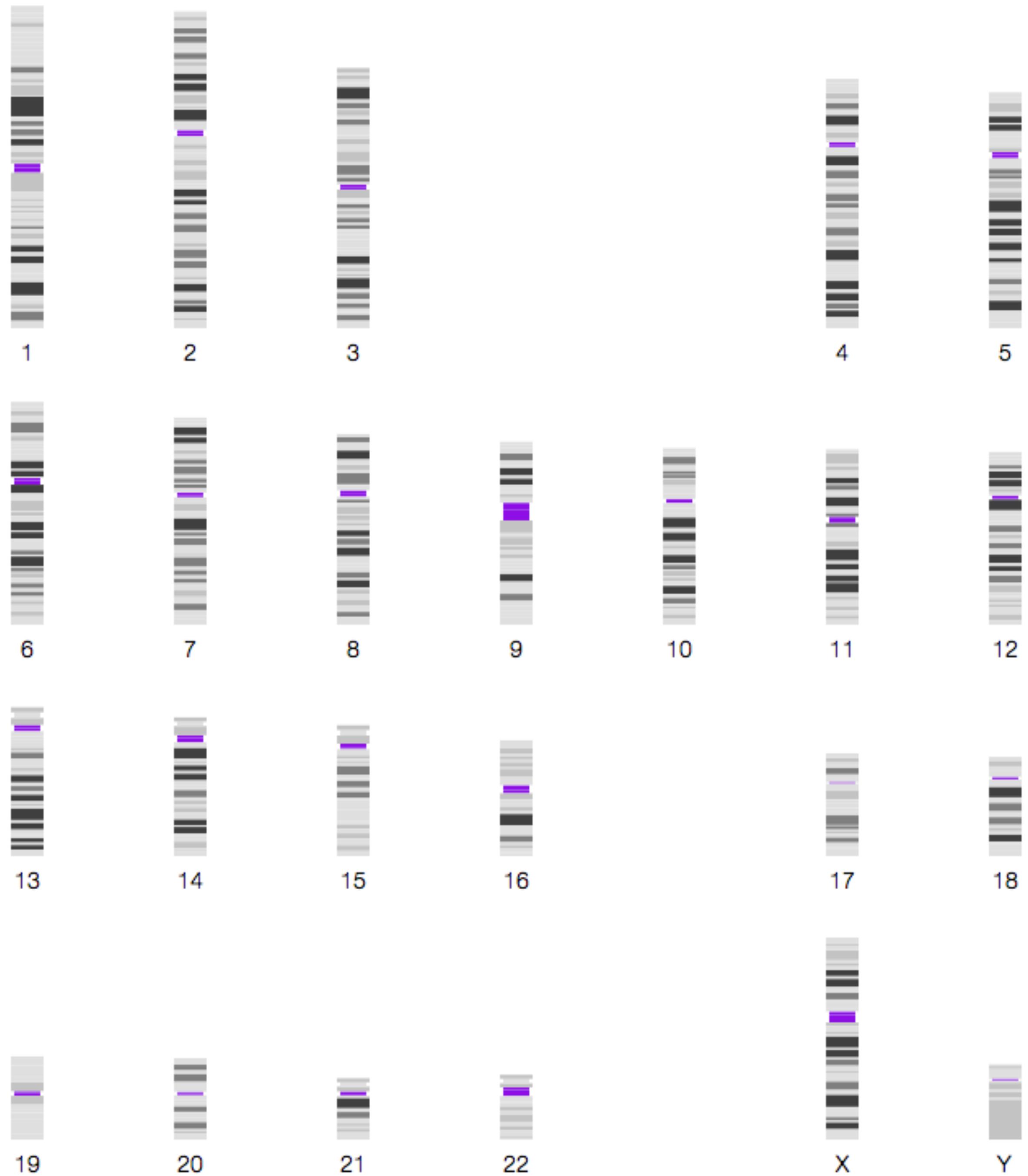
- How much computer storage is required for 1000 Genomes
  - WES & WGS
  - Different file formats
    - SAM
    - BAM
    - CRAM
    - VCF
    - FASTA
  - Associated costs
    - Cost factors
    - Raw Storage costs

Please provide 1-page size estimates and reasoning for the use of the different file types (i.e. which would you use for storing called variants, which for full archival purposes, browser visualisation), for 3-5 formats.

Submit your files (.md) per pull request to your Github directory.

# **Genome Editions**

## **Sizes | positions | mappings**



Chromosome	Basepair length (GRCh38)
1	248956422
2	242193529
3	198295559
4	190214555
5	181538259
6	170805979
7	159345973
8	145138636
9	138394717
10	133797422
11	135086622
12	133275309
13	114364328
14	107043718
15	101991189
16	90338345
17	83257441
18	80373285
19	59128983
20	64444167
21	46709983
22	50818468
X	156040895
Y	57227415
	<b>3080419480</b>



genome.ucsc.edu  
cytoBand\_UCSC\_hg18.txt

chr1	0	2300000	p36.33	gneg
chr1	2300000	5300000	p36.32	gpos25
chr1	5300000	7100000	p36.31	gneg
chr1	7100000	9200000	p36.23	gpos25
chr1	9200000	12600000	p36.22	gneg
chr1	12600000	16100000	p36.21	gpos50
chr1	16100000	20300000	p36.13	gneg
chr1	20300000	23800000	p36.12	gpos25
chr1	23800000	27800000	p36.11	gneg
chr1	27800000	30000000	p35.3	gpos25
chr1	30000000	32200000	p35.2	gneg
chr1	32200000	34400000	p35.1	gpos25
chr1	34400000	39600000	p34.3	gneg
chr1	39600000	43900000	p34.2	gpos25
chr1	43900000	46500000	p34.1	gneg
chr1	46500000	51300000	p33	gpos75
chr1	51300000	56200000	p32.3	gneg
chr1	56200000	58700000	p32.2	gpos50
chr1	58700000	60900000	p32.1	gneg
...	...	...	...	...
chrX	130300000	133500000	q26.2	gpos25
chrX	133500000	137800000	q26.3	gneg
chrX	137800000	140100000	q27.1	gpos75
chrX	140100000	141900000	q27.2	gneg
chrX	141900000	146900000	q27.3	gpos100
chrX	146900000	154913754	q28	gneg
chrY	0	1700000	p11.32	gneg
chrY	1700000	3300000	p11.31	gpos50
chrY	3300000	11200000	p11.2	gneg
chrY	11200000	11300000	p11.1	acen
chrY	11300000	12500000	q11.1	acen
chrY	12500000	14300000	q11.21	gneg
chrY	14300000	19000000	q11.221	gpos50
chrY	19000000	21300000	q11.222	gneg
chrY	21300000	25400000	q11.223	gpos50
chrY	25400000	27200000	q11.23	gneg
chrY	27200000	57772954	q12	gvar

# Cytogenetic band Sizes

chromosome	band start position	band stop position	cytogenetic band	staining intensity	band size
chr6	63400000	63500000	q11.2	gneg	100000
chr15	64900000	65000000	q22.32	gpos25	100000
chr17	22100000	22200000	p11.1	acen	100000
chrX	65000000	65100000	q11.2	gneg	100000
chrY	11200000	11300000	p11.1	acen	100000
chr17	35400000	35600000	q21.1	gneg	200000
chr3	44400000	44700000	p21.32	gpos50	300000
chr3	51400000	51700000	p21.2	gpos25	300000
chr9	132500000	132800000	q34.12	gpos25	300000
chr13	45900000	46200000	q14.13	gneg	300000
chr15	65000000	65300000	q22.33	gneg	300000
chr1	120700000	121100000	p11.2	gneg	400000
chr8	39500000	39900000	p11.22	gpos25	400000
chr9	72700000	73100000	q21.12	gneg	400000
chr16	69400000	69800000	q22.2	gpos50	400000
chr19	43000000	43400000	q13.13	gneg	400000
chr9	70000000	70500000	q13	gneg	500000
chr20	41100000	41600000	q13.11	gneg	500000
...	...	...	...	...	...
chr9	51800000	60300000	q11	acen	8500000
chrX	76000000	84500000	q21.1	gpos100	8500000
chr11	76700000	85300000	q14.1	gpos100	8600000
chr13	77800000	86500000	q31.1	gpos100	8700000
chr7	77400000	86200000	q21.11	gpos100	8800000
chr8	29700000	38500000	p12	gpos75	8800000
chr3	14700000	23800000	p24.3	gpos100	9100000
chr5	82800000	91900000	q14.3	gpos100	9100000
chr6	104800000	113900000	q21	gneg	9100000
chrX	120700000	129800000	q25	gpos100	9100000
chr9	60300000	70000000	q12	gvar	9700000
chr1	212100000	222100000	q41	gpos100	10000000
chr1	128000000	142400000	q12	gvar	14400000
chr1	69500000	84700000	p31.1	gpos100	15200000
chrY	27200000	57772954	q12	gvar	30572954

Positional genomic data has to be evaluated  
in the context of the correct edition

Chromosome	Basepairs 2003 (HG16)	Basepairs 2006 (HG18)	Basepairs 2009 (HG19)	Basepairs 2013 (GRCh38)	HG16 => HG19
1	246127941	247249719	249250621	248956422	2828481
2	243615958	242951149	243199373	242193529	-1422429
3	199344050	199501827	198022430	198295559	-1048491
4	191731959	191273063	191154276	190214555	-1517404
5	181034922	180857866	180915260	181538259	503337
6	170914576	170899992	171115067	170805979	-108597
7	158545518	158821424	159138663	159345973	800455
8	146308819	146274826	146364022	145138636	-1170183
9	136372045	140273252	141213431	138394717	2022672
10	135037215	135374737	135534747	133797422	-1239793
11	134482954	134452384	135006516	135086622	603668
12	132078379	132349534	133851895	133275309	1196930
13	113042980	114142980	115169878	114364328	1321348
14	105311216	106368585	107349540	107043718	1732502
15	100256656	100338915	102531392	101991189	1734533
16	90041932	88827254	90354753	90338345	296413
17	81860266	78774742	81195210	83257441	1397175
18	76115139	76117153	78077248	80373285	4258146
19	63811651	63811651	59128983	59128983	-4682668
20	63741868	62435964	63025520	64444167	702299
21	46976097	46944323	48129895	46709983	-266114
22	49396972	49691432	51304566	50818468	1421496
X	153692391	154913754	155270560	156040895	2348504
Y	50286555	57772954	59373566	57227415	6940860
	<b>3070128059</b>	<b>3080419480</b>	<b>3095677412</b>	<b>3088781199</b>	<b>18653140</b>

# Genome Liftover

## Moving between genome editions

SOFTWARE TOOL ARTICLE

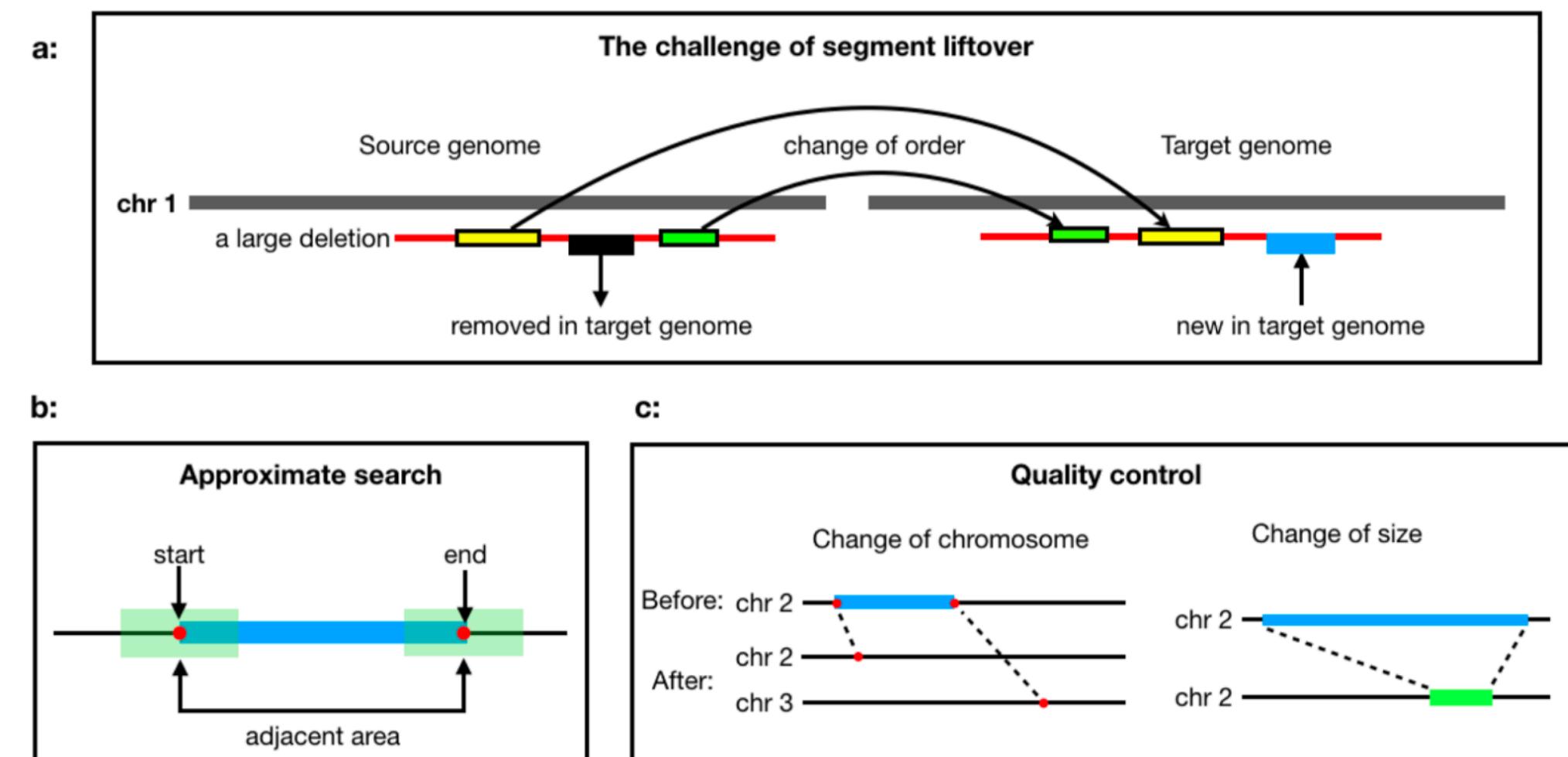
REVISED **segment\_liftover** : a Python tool to convert segments between genome assemblies [version 2; referees: 2 approved]

Bo Gao  1,2, Qingyao Huang  1,2, Michael Baudis  1,2

<sup>1</sup>Institute of molecular Life Sciences, University of Zürich, Zürich, CH-8057, Switzerland

<sup>2</sup>Swiss Institute of Bioinformatics, University of Zürich, Zürich, CH-8057, Switzerland

- different genome editions lead to shifting positions of defined elements such as genes
- local regions are frequently stable between editions
- shifts from change in regional lengths are defined in "chain files"
- chain files serve as guides for positional remapping using liftover methods
- Task: Read up on liftover techniques (starting w/ our article) & explore resources and other applications



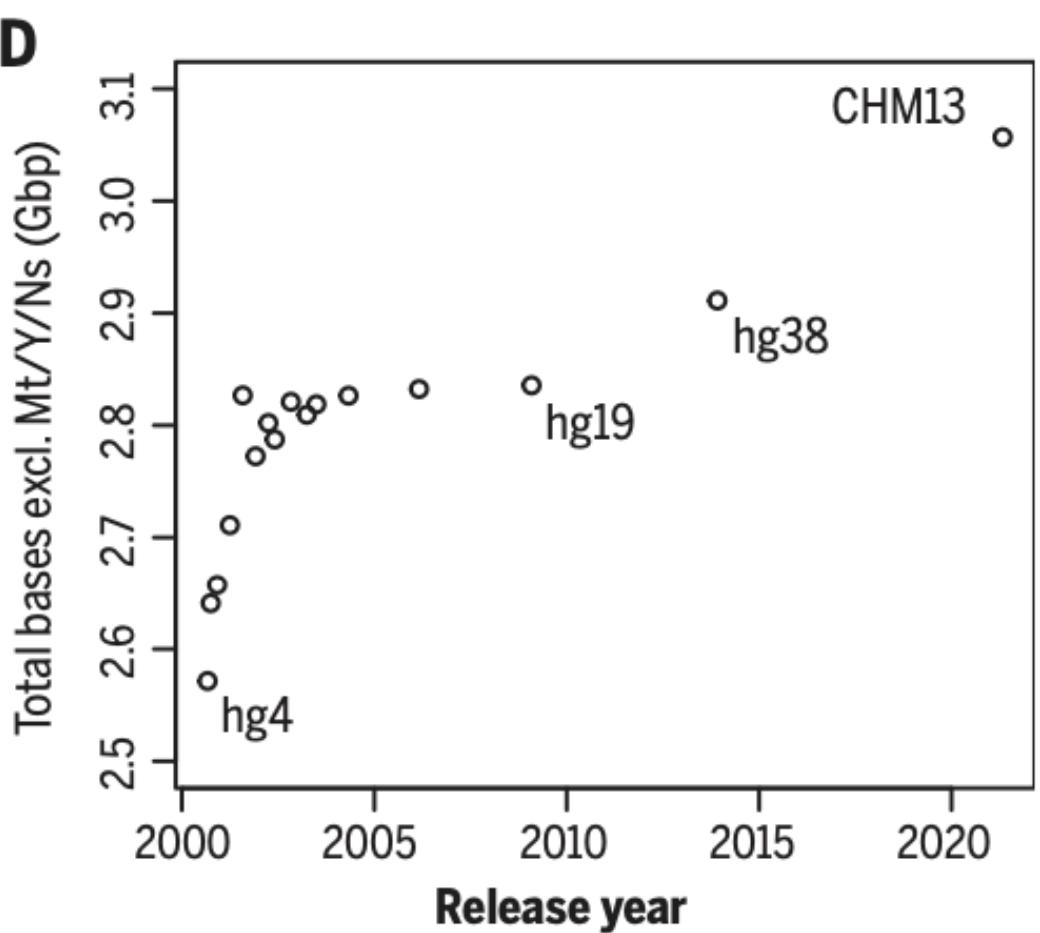
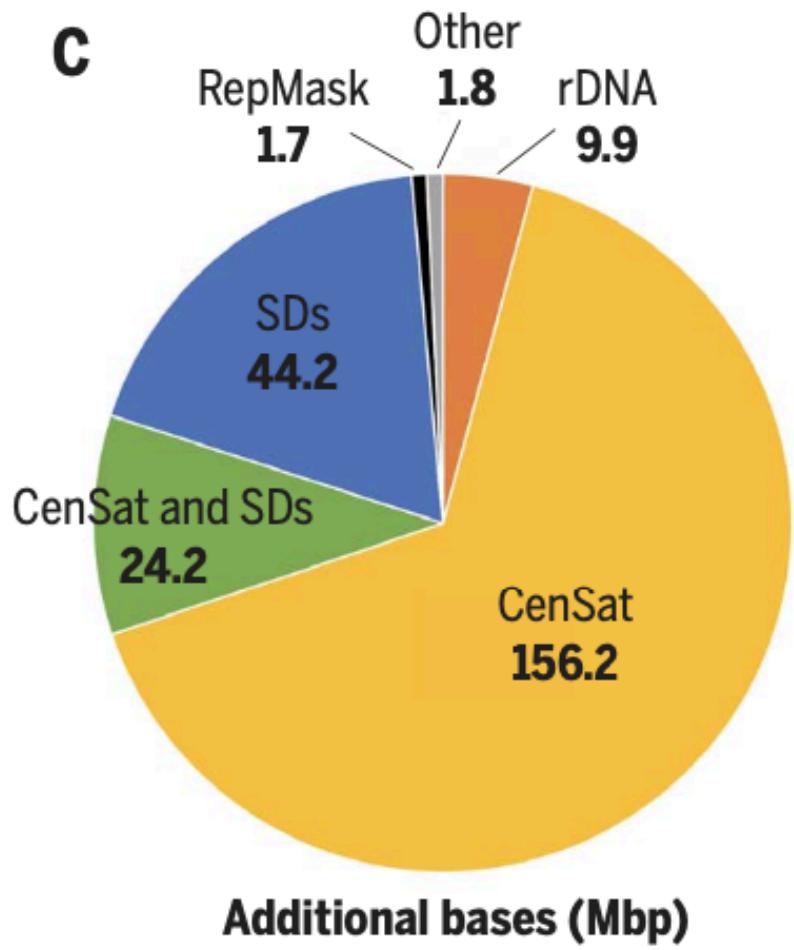
# Telomere-to-Telomere (T2T)

The first complete, gapless sequence of a human genome.



# A Complete (Human) Genome

## The T2T consortium resolves ambiguous genomic regions using a mix of sequencing technologies



**C**) Additional (nonsyntenic) bases in the CHM13 assembly relative to GRCh38 per chromosome by sequence type. (Note that the CenSat and SD annotations overlap.) RepMask, RepeatMasker.

**D**) Total nongap bases in UCSC reference genome releases dating back to September 2000 (hg4) and ending with T2T-CHM13 in 2021. Mt/Y/Ns, mitochondria, chrY, and gaps.

### RESEARCH ARTICLE

#### HUMAN GENOMICS

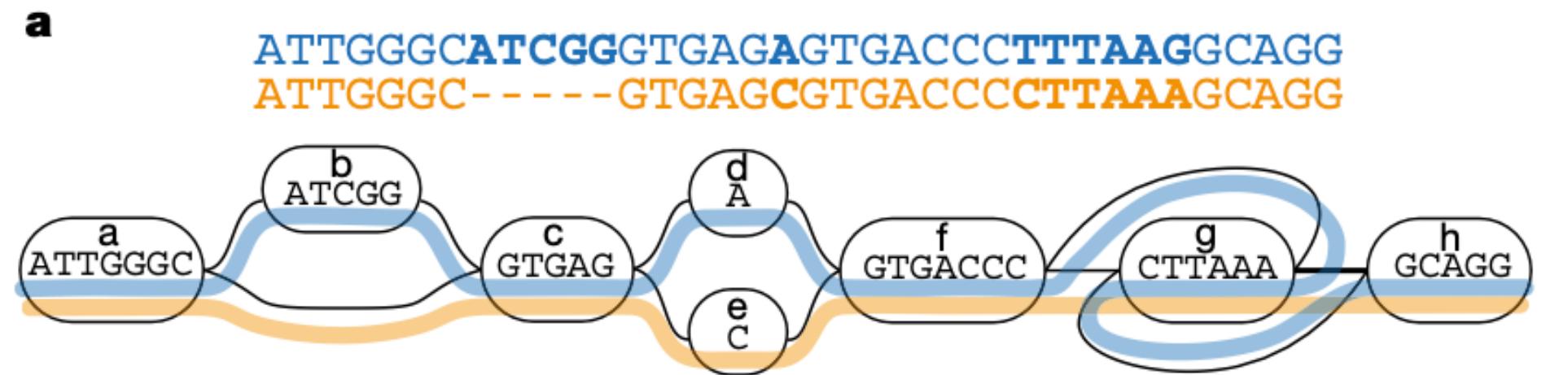
## The complete sequence of a human genome

Sergey Nurk<sup>1†</sup>, Sergey Koren<sup>1†</sup>, Arang Rhee<sup>1†</sup>, Mikko Rautiainen<sup>1†</sup>, Andrey V. Bzikadze<sup>2</sup>, Alla Mikheenko<sup>3</sup>, Mitchell R. Vollger<sup>4</sup>, Nicolas Altemose<sup>5</sup>, Lev Uralsky<sup>6,7</sup>, Ariel Gershman<sup>8</sup>, Sergey Aganezov<sup>9†</sup>, Savannah J. Hoyt<sup>10</sup>, Mark Diekhans<sup>11</sup>, Glennis A. Logsdon<sup>4</sup>, Michael Alonge<sup>9</sup>, Stylianos E. Antonarakis<sup>12</sup>, Matthew Borchers<sup>13</sup>, Gerard G. Bouffard<sup>14</sup>, Shelise Y. Brooks<sup>14</sup>, Gina V. Caldas<sup>15</sup>, Nae-Chyun Chen<sup>9</sup>, Haoyu Cheng<sup>16,17</sup>, Chen-Shan Chin<sup>18</sup>, William Chow<sup>19</sup>, Leonardo G. de Lima<sup>13</sup>, Philip C. Dishuck<sup>4</sup>, Richard Durbin<sup>19,20</sup>, Tatiana Dvorkina<sup>3</sup>, Ian T. Fiddes<sup>21</sup>, Giulio Formenti<sup>22,23</sup>, Robert S. Fulton<sup>24</sup>, Arkarachai Fungtammasan<sup>18</sup>, Erik Garrison<sup>11,25</sup>, Patrick G. S. Grady<sup>10</sup>, Tina A. Graves-Lindsay<sup>26</sup>, Ira M. Hall<sup>27</sup>, Nancy F. Hansen<sup>28</sup>, Gabrielle A. Hartley<sup>10</sup>, Marina Haukness<sup>11</sup>, Kerstin Howe<sup>19</sup>, Michael W. Hunkapiller<sup>29</sup>, Chirag Jain<sup>1,30</sup>, Miten Jain<sup>11</sup>, Erich D. Jarvis<sup>22,23</sup>, Peter Kerpeljiev<sup>31</sup>, Melanie Kirsche<sup>9</sup>, Mikhail Kolmogorov<sup>32</sup>, Jonas Korlach<sup>29</sup>, Milinn Kremitzki<sup>26</sup>, Heng Li<sup>16,17</sup>, Valerie V. Maduro<sup>33</sup>, Tobias Marschall<sup>34</sup>, Ann M. McCartney<sup>1</sup>, Jennifer McDaniel<sup>35</sup>, Danny E. Miller<sup>4,36</sup>, James C. Mullikin<sup>14,28</sup>, Eugene W. Myers<sup>37</sup>, Nathan D. Olson<sup>35</sup>, Benedict Paten<sup>11</sup>, Paul Peluso<sup>29</sup>, Pavel A. Pevzner<sup>32</sup>, David Porubsky<sup>4</sup>, Tamara Potapova<sup>13</sup>, Evgeny I. Rogaev<sup>6,7,38,39</sup>, Jeffrey A. Rosenfeld<sup>40</sup>, Steven L. Salzberg<sup>9,41</sup>, Valerie A. Schneider<sup>42</sup>, Fritz J. Sedlazeck<sup>43</sup>, Kishwar Shafin<sup>11</sup>, Colin J. Shew<sup>44</sup>, Alaina Shumate<sup>41</sup>, Ying Sims<sup>19</sup>, Arian F. A. Smit<sup>45</sup>, Daniela C. Soto<sup>44</sup>, Ivan Sovic<sup>29,46</sup>, Jessica M. Storer<sup>45</sup>, Aaron Streets<sup>5,47</sup>, Beth A. Sullivan<sup>48</sup>, Françoise Thibaud-Nissen<sup>42</sup>, James Torrance<sup>19</sup>, Justin Wagner<sup>35</sup>, Brian P. Walenz<sup>1</sup>, Aaron Wenger<sup>29</sup>, Jonathan M. D. Wood<sup>19</sup>, Chunlin Xiao<sup>42</sup>, Stephanie M. Yan<sup>49</sup>, Alice C. Young<sup>14</sup>, Samantha Zarate<sup>9</sup>, Urvashi Surti<sup>50</sup>, Rajiv C. McCoy<sup>49</sup>, Megan Y. Dennis<sup>44</sup>, Ivan A. Alexandrov<sup>3,7,51</sup>, Jennifer L. Gerton<sup>13,52</sup>, Rachel J. O'Neill<sup>10</sup>, Winston Timp<sup>8,41</sup>, Justin M. Zook<sup>35</sup>, Michael C. Schatz<sup>9,49</sup>, Evan E. Eichler<sup>4,53\*</sup>, Karen H. Miga<sup>11,54\*</sup>, Adam M. Phillippy<sup>1\*</sup>

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion-base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

# A (Human) Pangenome

## Leveraging recent sequencing technologies and advances in genome informatics



A pangenome variation graph comprising two elements: a sequence graph, the nodes of which represent oriented DNA strings and bidirected edges represent the connectivity relationships; and embedded haplotype paths (coloured lines) that represent the individual assemblies.

## Article

### A draft human pangenome reference

<https://doi.org/10.1038/s41586-023-05896-x>

Received: 9 July 2022

Accepted: 28 February 2023

Published online: 10 May 2023

Open access

Check for updates

Wen-Wei Liao<sup>1,2,3,60</sup>, Mobin Asri<sup>4,60</sup>, Jana Ebler<sup>5,6,60</sup>, Daniel Doerr<sup>5,6</sup>, Marina Haukness<sup>4</sup>, Glenn Hickey<sup>4</sup>, Shuangjia Lu<sup>1,2</sup>, Julian K. Lucas<sup>4</sup>, Jean Monlong<sup>4</sup>, Haley J. Abel<sup>7</sup>, Silvia Buonaiuto<sup>8</sup>, Xian H. Chang<sup>4</sup>, Haoyu Cheng<sup>9,10</sup>, Justin Chu<sup>9</sup>, Vincenza Colonna<sup>9,11</sup>, Jordan M. Eizenga<sup>4</sup>, Xiaowen Feng<sup>9,10</sup>, Christian Fischer<sup>11</sup>, Robert S. Fulton<sup>12,13</sup>, Shilpa Garg<sup>14</sup>, Cristian Groza<sup>15</sup>, Andrea Guerracino<sup>11,16</sup>, William T. Harvey<sup>17</sup>, Simon Heumos<sup>18,19</sup>, Kerstin Howe<sup>20</sup>, Miten Jain<sup>21</sup>, Tsung-Yu Lu<sup>22</sup>, Charles Markello<sup>4</sup>, Fergal J. Martin<sup>23</sup>, Matthew W. Mitchell<sup>24</sup>, Katherine M. Munson<sup>17</sup>, Moses Njagi Mwaniki<sup>25</sup>, Adam M. Novak<sup>4</sup>, Hugh E. Olsen<sup>4</sup>, Trevor Pesout<sup>4</sup>, David Porubsky<sup>17</sup>, Pjotr Prins<sup>11</sup>, Jonas A. Sibbesen<sup>26</sup>, Jouni Sirén<sup>4</sup>, Chad Tomlinson<sup>12</sup>, Flavia Villani<sup>11</sup>, Mitchell R. Vollger<sup>17,27</sup>, Lucinda L. Antonacci-Fulton<sup>12</sup>, Gunjan Baid<sup>28</sup>, Carl A. Baker<sup>17</sup>, Anastasiya Belyaeva<sup>28</sup>, Konstantinos Billis<sup>23</sup>, Andrew Carroll<sup>28</sup>, Pi-Chuan Chang<sup>28</sup>, Sarah Cody<sup>12</sup>, Daniel E. Cook<sup>28</sup>, Robert M. Cook-Deegan<sup>28</sup>, Omar E. Cornejo<sup>30</sup>, Mark Diekhans<sup>4</sup>, Peter Ebert<sup>5,6,31</sup>, Susan Fairley<sup>23</sup>, Olivier Fedrigo<sup>32</sup>, Adam L. Felsenfeld<sup>33</sup>, Giulio Formenti<sup>32</sup>, Adam Frankish<sup>23</sup>, Yan Gao<sup>34</sup>, Nanibaa' A. Garrison<sup>35,36,37</sup>, Carlos Garcia Giron<sup>23</sup>, Richard E. Green<sup>38,39</sup>, Leanne Haggerty<sup>23</sup>, Kendra Hoekzema<sup>17</sup>, Thibaut Hourlier<sup>23</sup>, Hanlee P. Ji<sup>40</sup>, Eimear E. Kenny<sup>41</sup>, Barbara A. Koenig<sup>42</sup>, Alexey Kolesnikov<sup>28</sup>, Jan O. Korbel<sup>23,43</sup>, Jennifer Kordosky<sup>17</sup>, Sergey Koren<sup>44</sup>, HoJoon Lee<sup>40</sup>, Alexandra P. Lewis<sup>17</sup>, Hugo Magalhães<sup>5,6</sup>, Santiago Marco-Sola<sup>45,46</sup>, Pierre Marijon<sup>5,6</sup>, Ann McCartney<sup>44</sup>, Jennifer McDaniel<sup>47</sup>, Jacquelyn Mountcastle<sup>32</sup>, Maria Nattestad<sup>28</sup>, Sergey Nurk<sup>44</sup>, Nathan D. Olson<sup>47</sup>, Alice B. Popejoy<sup>48</sup>, Daniela Puiu<sup>49</sup>, Mikko Rautiainen<sup>44</sup>, Allison A. Regier<sup>12</sup>, Arang Rhee<sup>44</sup>, Samuel Sacco<sup>30</sup>, Ashley D. Sanders<sup>50</sup>, Valerie A. Schneider<sup>51</sup>, Baergen I. Schultz<sup>33</sup>, Kishwar Shafin<sup>28</sup>, Michael W. Smith<sup>33</sup>, Heidi J. Sofia<sup>33</sup>, Ahmad N. Abou Tayoun<sup>52,53</sup>, Françoise Thibaud-Nissen<sup>51</sup>, Francesca Floriana Tricomi<sup>23</sup>, Justin Wagner<sup>47</sup>, Brian Walenz<sup>44</sup>, Jonathan M. D. Wood<sup>20</sup>, Aleksey V. Zimin<sup>49,54</sup>, Guillaume Bourque<sup>55,56,57</sup>, Mark J. P. Chaisson<sup>22</sup>, Paul Flicek<sup>23</sup>, Adam M. Phillippy<sup>44</sup>, Justin M. Zook<sup>47</sup>, Evan E. Eichler<sup>17,58</sup>, David Haussler<sup>4,58</sup>, Ting Wang<sup>12,13</sup>, Erich D. Jarvis<sup>32,58,69</sup>, Karen H. Miga<sup>4</sup>, Erik Garrison<sup>11</sup>, Tobias Marschall<sup>5,6</sup>, Ira M. Hall<sup>1,2</sup>, Heng Li<sup>9,10</sup> & Benedict Paten<sup>4</sup>

Here the Human Pangenome Reference Consortium presents a first draft of the human pangenome reference. The pangenome contains 47 phased, diploid assemblies from a cohort of genetically diverse individuals<sup>1</sup>. These assemblies cover more than 99% of the expected sequence in each genome and are more than 99% accurate at the structural and base pair levels. Based on alignments of the assemblies, we generate a draft pangenome that captures known variants and haplotypes and reveals new alleles at structurally complex loci. We also add 119 million base pairs of euchromatic polymorphic sequences and 1,115 gene duplications relative to the existing reference GRCh38. Roughly 90 million of the additional base pairs are derived from structural variation. Using our draft pangenome to analyse short-read data reduced small variant discovery errors by 34% and increased the number of structural variants detected per haplotype by 104% compared with GRCh38-based workflows, which enabled the typing of the vast majority of structural variant alleles per sample.

# Task: Estimate Storage Requirements for 1000 Genomes

## How much computer storage is required for 1000 Genomes

- WES & WGS
- Different file formats
  - SAM
  - BAM
  - VCF
  - FASTA
- Associated costs
  - Cost factors
  - Raw Storage costs
- Familiarize with VCF format  
→specification in article collection



IBM-storage-unit-3500-Schiphol-1957

Please provide 1-page size estimates and reasoning for the use of the different file types (i.e. which would you use for storing called variants, which for full archival purposes, browser visualisation), for 3-5 formats. Submit your files (.md) per pull request to your Github directory.

# Task: Reading up on Genome Technologies

- General NGS technologies
- count based vs. intensity based as principle
- long and short read technologies
  - ▶ advantages/applications for either
- dig deeper for some (molecular)-cytogenetic techniques:
  - ▶ banding analysis, SKY, M-FISH
  - ▶ SNP, aCGH arrays
  - ▶ chromosomal CGH
- ➔ notes about usage (research, clinical, historical vs. current)
- "T2T genome"
  - ▶ What technologies enabled this?
- Graph Genomes
  - ▶ Principles?

# BIO392: Course Schedule

<https://compbiozurich.org/UZH-BIO392/>

	Friday 2025-04-04	Tuesday 2025-04-08	Wednesday 2025-04-09	Thursday 2025-04-10	Friday 2025-04-11	Tuesday 2025-04-15	Wednesday 2025-04-16	Thursday 2025-04-17	Tuesday 2025-04-29	Wednesday 2025-04-30	Friday 2025-05-02	Tuesday 2025-05-06	Wednesday 2025-05-07
09:00 - 10:00	* Room information * Administrative - discuss times/days - exam		Jiahui: Terminal / Unix / Files	Hangjia: R environment introduction	Michael: Genomic Resources & Data Sharing		Feifei: Sequence analysis practical. FastQC, trimmomatic, BWA-MEM2, SAMtools, GangSTR, BCFtools	Michael: Genomic Data & Privacy		Feifei & Ziying: survival	Feifei: population structure		
10:00 - 11:00	Tina Siegenthaler: technical introduction (room, computer, accounts)		Jiahui: Terminal / Unix / Files	Hangjia: R exercise	Michael: Genomic Resources & Data Sharing		Feifei:: Sequence analysis practical	Michael: Genomic Data & Privacy		Feifei & Ziying: survival	Feifei: population structure		Discussion
11:00 - 12:00	* explore course site * create Github accounts and forward to bio392@compbiozurich.org *feifei&jiahui: overall schedule of the course		jiahui: SIB online introduction to Unix		Michael: Genomic Resources & Data Sharing		Feifei:: Sequence analysis practical	Michael: Genomic Data & Privacy		Feifei: survival	Feifei: population structure		
13:00 - 14:00	jiahui: Github	Michael: Introduction	Jiahui & Ziying: Python	Hangjia: CNV paper reading	Hangjia: Clinvar and Clingen	Feifei: Sequence analysis introduction. Overview of pipeline from raw reads -> variant calling & interpretation	Feifei:: STR reading up			Feifei:: analysis & interpretation. Parsing VCF (cvcf2), UCSD genome browser, ENSEMBL variant effect predictor	Feifei: population structure	Feifei & others: Presentation & Discussion	Exam revision, Q&A
14:00 - 15:00	jiahui: Github&Git exercise	Michael: Introduction	Exercise	Hangjia: Progenetix as tool for CNV frequencies etc.	Hangjia: blast	Feifei: Sequence analysis introduction	Feifei:: STR reading up			Feifei:: analysis & interpretation.	Feifei: population structure	Feifei & others: Presentation & Discussion	
15:00 - 16:30		Michael: Introduction			Hangjia: Blast exercise					Feifei:: analysis & interpretation.	Feifei: population structure	Feifei & others: Presentation & Discussion	

# Genome analyses at the core of Personalized Health™

## There'll be Sequencing Everywhere...

- Genome analyses (including transcriptome, metagenomics) are the **core technologies** for Personalized Health™ applications
- In the context of **academic medicine**, this requires
  - standard sample acquisition procedures & central **biobanking**
  - **core sequencing facility** (large throughput, cost efficiency, uniform sample and data handling procedures)
- secure **computing/analysis** platform
- Standardized **data formats** and **sample identification** procedures
- Metadata rich, reference **variant resource(s)** & expertise
- participation in reciprocal, international **data sharing** and **biocuration** efforts