University of Zurich[UZH]

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 enviroment introduction	Michael: Genoomic Resources & Data Sharing		Feifei: Sequence analysis practical. FastQC, trimmomatic, BWA-MEM2, SAMtools, GangSTR, BCFtools	Michael: Genomic Data & Privacy		Feifei & Ziying: survival	Feifei: population structure		Exam
10:00 - 11:00	Tina Siegenthaler: technical introduction (room, computer, accounts)		Jiahui: Terminal / Unix / Files	Hangjia: R exercise	Michael: Genoomic 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: Genoomic 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 (cyvcf2), 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			Hangija: Blast exercise				Feifei:: analysis & interpretation.	Feifei: population structure	Feifei & others: Presentation & Discussion		

https://drive.switch.ch/index.php/s/PB1czLjrjAKR6Q2

1992	Heidelberg	Student of medicine | doctoral thesis in molecular cytogenetics @ DKFZ (Peter Lichter) | 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

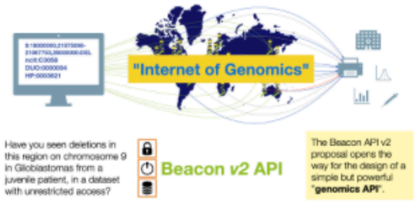
- CNV resource
  - Data - e.g. progenetix.org
  - Tools - CNV remapping, visualization, API access to resources ...
- patterns and correlations of genomic variations in cancer
- annotation mapping
- API, protocols and standards contributions

➡Beacon

baudisgroup.org

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.

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

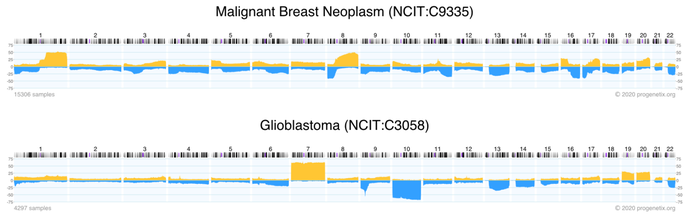
BIOARXIV PREPRINT (2025-03-23): HTTPS://DOI.ORG/10.1101/2025.03.20.644282

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"



progenetix

Genome screening at the core of “Personalised Health”

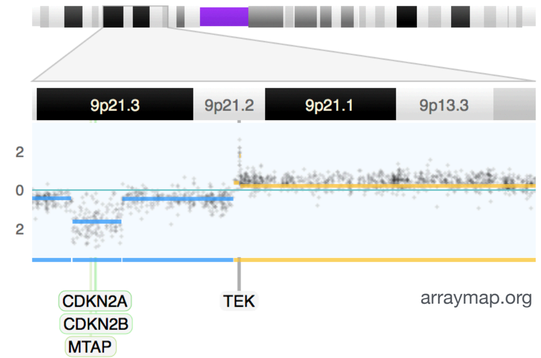
- **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications
- The unexpectedly large amount of **sequence variants** in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of **reference repositories**
- **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

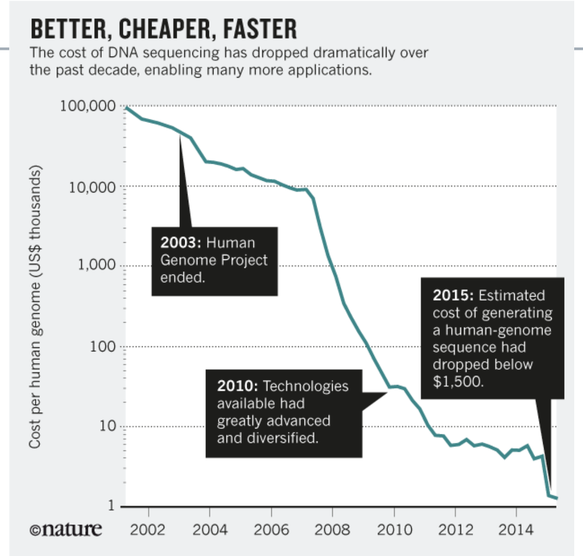
arrayMap

progenetix

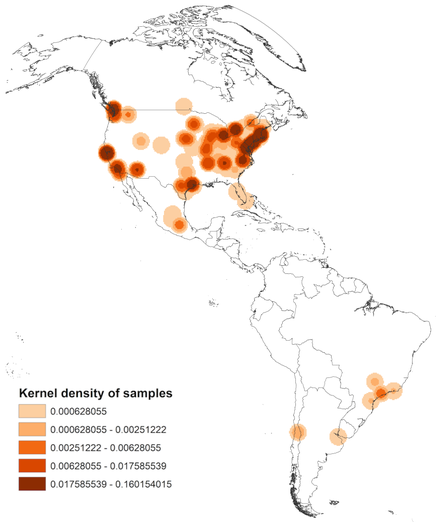
elixir

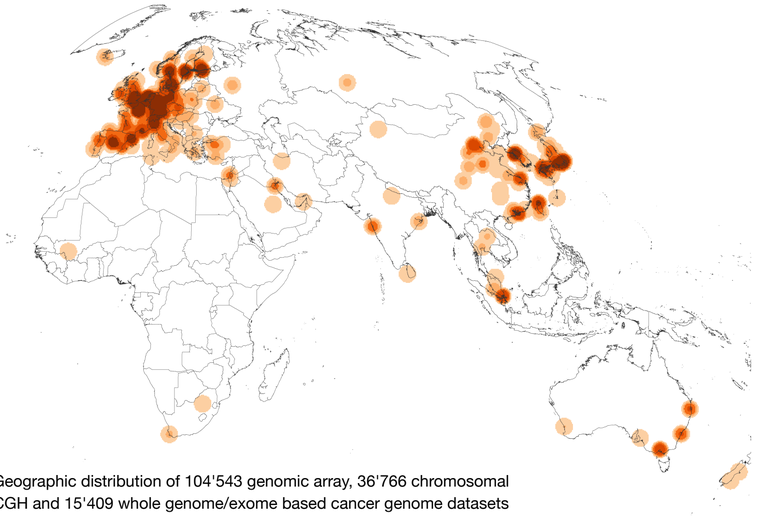
Global Alliance for Genomics & Health





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

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

Rapid whole genome sequencing and precision neonatology

Joshua E. Petrikin, MD[a,], Laurel K. Willig, MD, FAAP[b], Laurie D. Smith, MD, PhD[c], and Stephen F. Kingsmore, MB, BAO, ChB, Dsc, FRCPath[d,e]*

Genomic Classification of Cutaneous Melanoma

The Cancer Genome Atlas Network[1,*,*,*]
[1]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

Psychiatry and Clinical Neurosciences

PCN Frontier Review

Copy-number variation in the pathogenesis of autism spectrum disorder

Emiko Shishido, PhD[1,2,3], Branko Aleksic, MD, PhD[3] and Norio Ozaki, MD, PhD[3]*

...he Promotion of Science, ...pan

RESEARCH ARTICLE Open Access

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

Haoyang Cai[1,2], Nitin Kumar[1,2], Homayoun C Bagheri[3], Christian von Mering[1,2], Mark D Robinson[1,2]* and Michael Baudis[1,2]*

Barkur S. Shastry

SNP alleles in human disease and evolution

insight progress

Cancer genetics

Bruce A. J. Ponder

DISEASE MECHANISMS

Mechanisms underlying structural variant formation in genomic disorders

Claudia M. B. Carvalho[1,2] and James R. Lupski[1,3,4,5]

Abstract | With the recent burst of technological developments in genomics, and the clinical implementation of genome-wide assays, our understanding of the molecular basis of genomic disorders, specifically the contribution of structural variation to disease burden, is evolving

Consequences of genomic diversity in *Mycobacterium tuberculosis*

Mireia Coscolla[a,b], Sebastien Gagneux[a,b,*]

[a] *Department of Medical Parasitology and Infection Biology, Swiss Tropical and Public Health Institute, Socinstrasse 57, 4002 Basel, Switzerland*
[b] *University of Basel, Petersplatz 1, Basel 4003, Switzerland*

Common gene variants, mortality and extreme longevity in humans

B.T. Heijmans[a,b], R.G.J. Westendorp[b], P.E. Slagboom[a,*]

RESEARCH ARTICLE Open Access

Integrative genome-wide expression profiling identifies three distinct molecular subgroups of renal cell carcinoma with different patient ...utcome

...nfred Beleut[1,5*], Philip Zimmermann[2], Michael Baudis[3], Nicole Bruni[4], Peter Bühlmann[4], Oliver Laule[2], ...-Duc Luu[1], Wilhelm Gruissem[2], Peter Schraml[1*] and Holger Moch[1]

NEURODEVELOPMENT

Genes, circuits, and precision therapies for autism and related neurodevelopmental disorders

Mustafa Sahin* and Mriganka Sur*

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

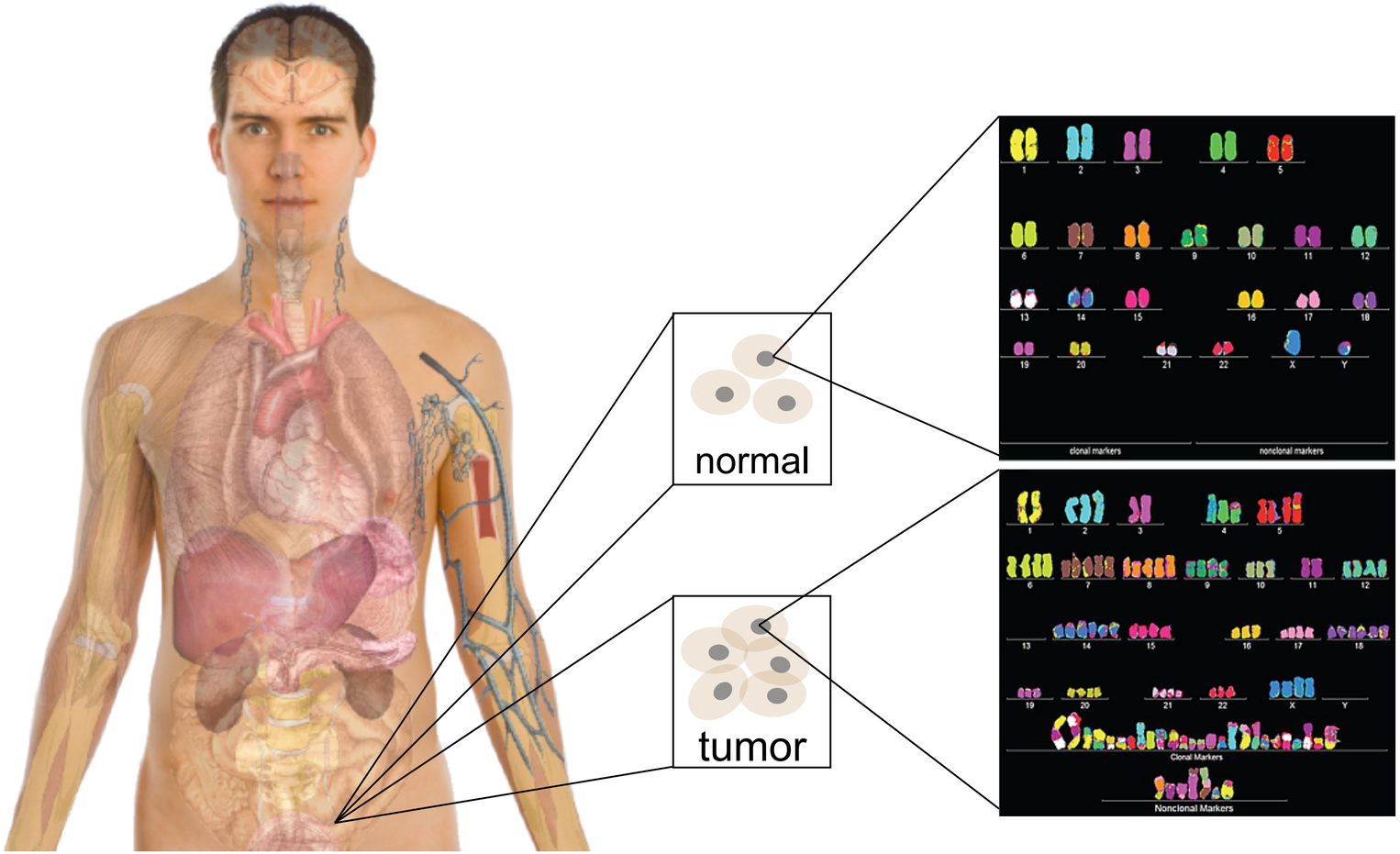
Thomas J. Lynch, M.D., Daphne W. Bell, Ph.D., Raffaella Sordella, Ph.D., Sarada Gurubhagavatula, M.D., Ross A. Okimoto, B.S., Brian W. Brannigan, B.A., Patricia L. Harris, M.S., Sara M. Haserlat, B.A., Jeffrey G. Supko, Ph.D., Frank G. Haluska, M.D., Ph.D., David N. Louis, M.D., David C. Christiani,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

The landscape of somatic copy-number alteration across human cancers

Rameen Beroukhim[1,3,4,5]*, Craig H. Mermel[1,3]*, Dale Porter[8], Guo Wei[1], Soumya Raychaudhuri[1,4], Jerry Donovan[8], Jordi Barretina[1,3], Jesse S. Boehm[1], Jennifer Dobson[1,3], Mitsuyoshi Urashima[9], Kevin T. Mc Henry[8], Reid M. Pinchback[1], Azra H. Ligon[4], Yoon-Jae Cho[6], Leila Haery[1,3], Heidi Greulich[1,3,4,5], Michael Reich[1], Wendy Winckler[1], Michael S. Lawrence[1], Barbara A. Weir[1,3], Kumiko E. Tanaka[1,3], Derek Y. Chiang[1,3,13], Adam J. Bass[1,3,4], Alice Loo[8], Carter Hoffman[1,3], John Prensner[1,3], Ted Liefeld[1], Qing Gao[1], Derek Yecies[3], Sabina Signoretti[3,4], Elizabeth Maher[10], Frederic J. Kaye[11], Hidefumi Sasaki[12], Joel E. Tepper[13], Jonathan A. Fletcher[4], Josep Tabernero[14], José Baselga[14], Ming-Sound Tsao[15], Francesca Demichelis[16], Mark A. Rubin[16], Pasi A. Janne[3,4], Mark J. Daly[1,17], Carmelo Nucera[7], Ross L. Levine[18], Benjamin L. Ebert[1,4,5], Stacey Gabriel[1], Anil K. Rustgi[19], Cristina R. Antonescu[18], Marc Ladanyi[18], Anthony Letai[3], Levi A. Garraway[1,3], Massimo Loda[3,4], David G. Beer[20], Lawrence D. True[21], Aikou Okamoto[22], Scott L. Pomeroy[6], Samuel Singer[18], Todd R. Golub[1,3,23], Eric S. Lander[1,2,5], Gad Getz[1], William R. Sellers[8] & Matthew Meyerson[1,3,5]

Personal Genomics as a Gateway into Biology

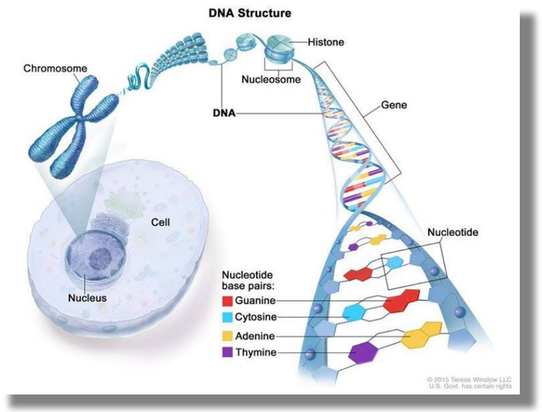
Personal genomes soon will become a commonplace part of medical research & eventually treatment (esp. for cancer). They will provide a primary connection for biological science to the general public.

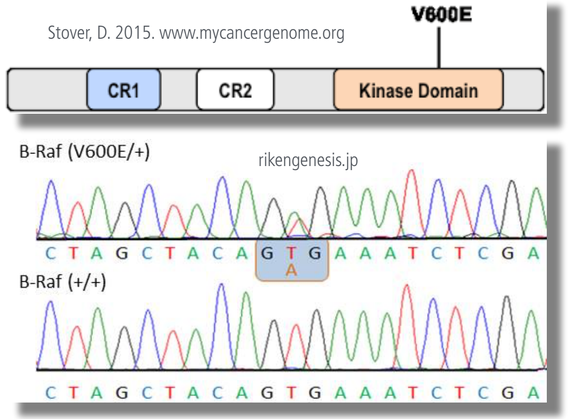


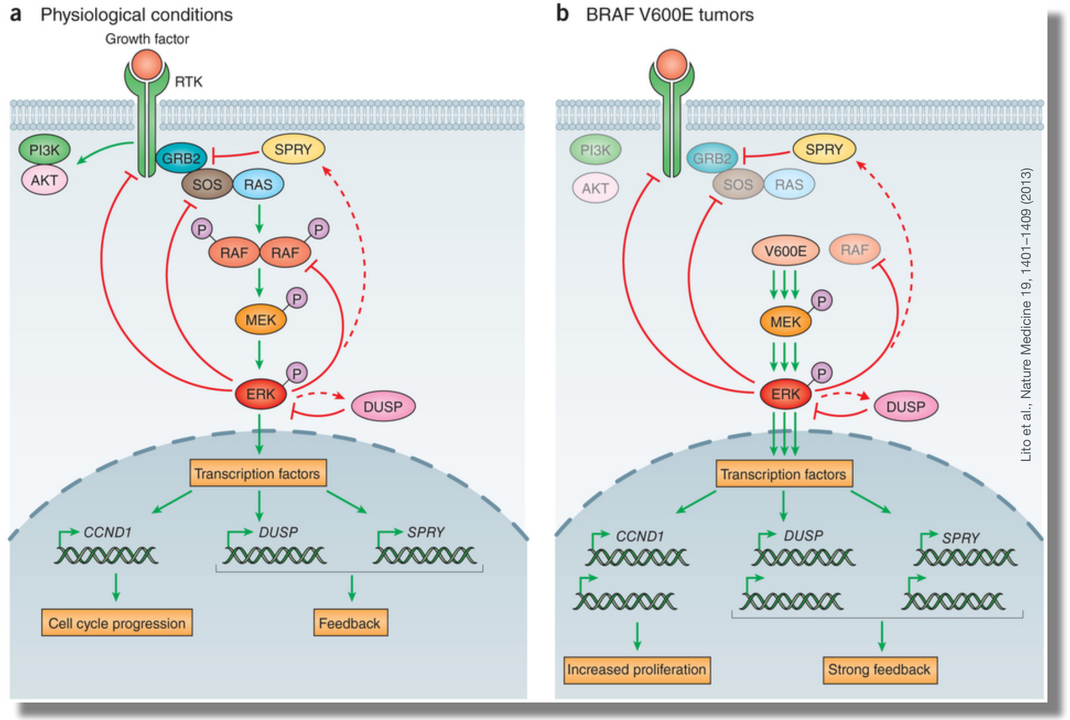
BRAF V600E (c.1799T>A) Mutation

Oncogene Activation by Single Nucleotide Alteration

- a single nucleotide exchange **T**hymidine > **A**denine 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***)





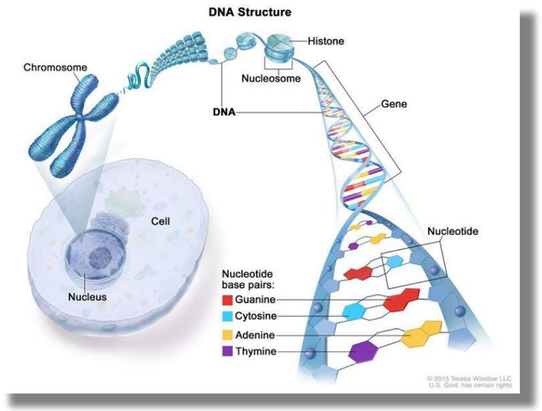


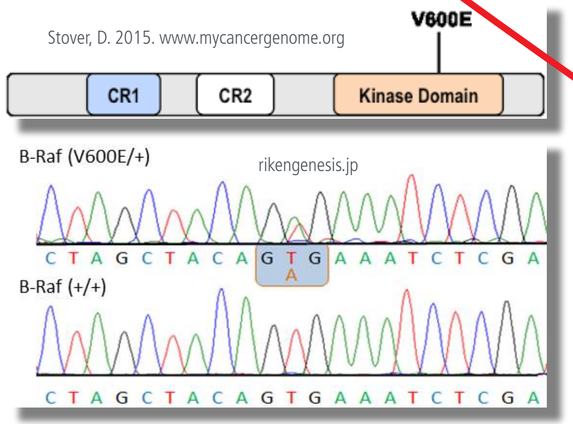
The BRAF V600E mutation leads to continuos phosphorylation of MEK, without the need for receptor based activation of the upstream pathway and loss of inhibitory feedback control.

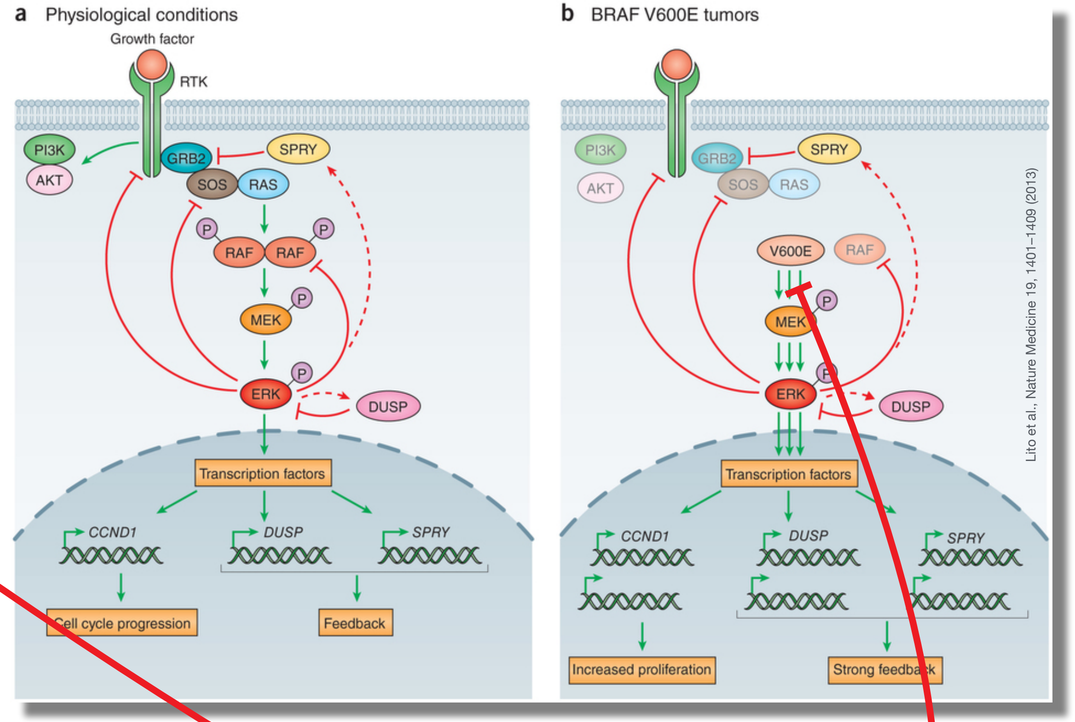
BRAF V600E (c.1799T>A) Mutation

Oncogene Activation by Single Nucleotide Alteration

- a single nucleotide exchange **T**hymidine > **A**denine 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***)





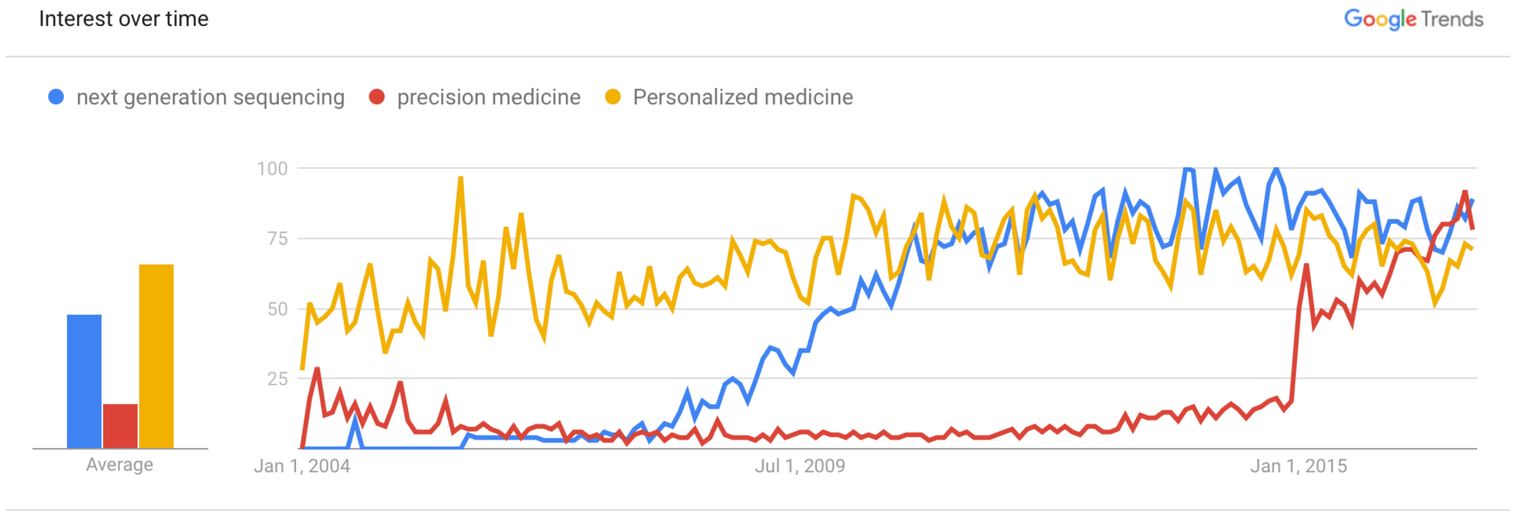


The BRAF V600E mutation leads to continuos phosphorylation of MEK, without the need for receptor based activation of the upstream pathway and loss of inhibitory feedback control.

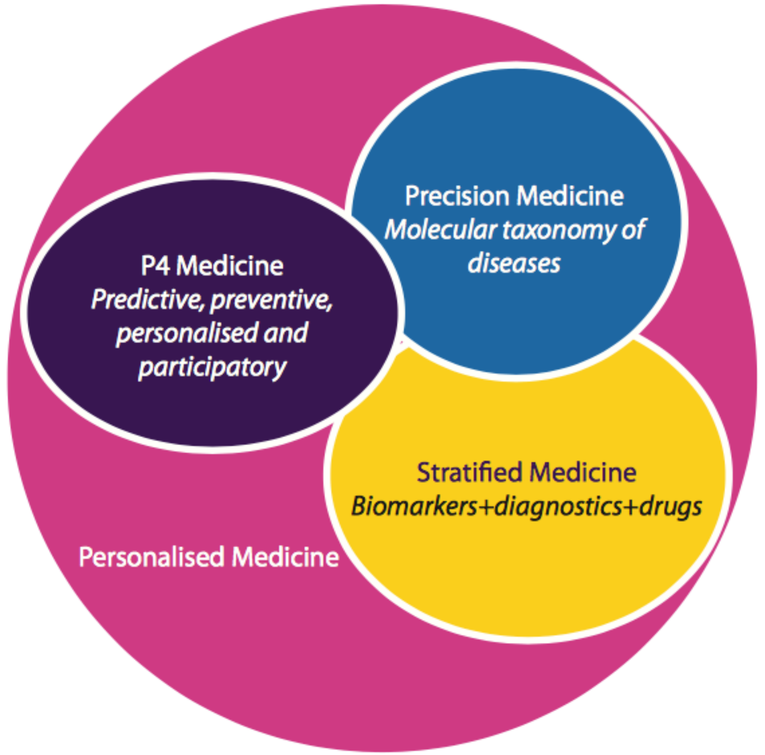
Begriffsbestimmung

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.



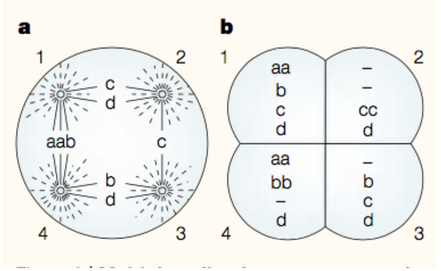
Source: PHG Foundation

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

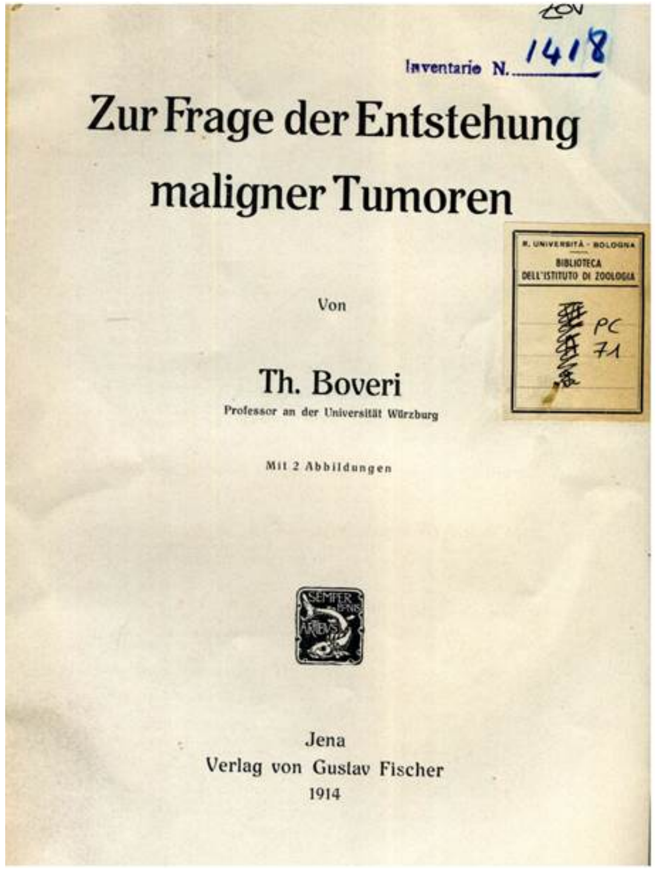
Theodor Boveri (1914)



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

Experimentelle Beobachtungen in Seeigeleiern, mit Vorhersage von:

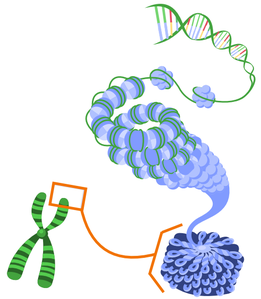
- **Tumorkontrollgene** (“Teilungshemmende Chromosomen”) zur **Zellzykluskontrolle**, in bösartigen Tumoren geschädigt
- **Tumorgene** (“Teilungsfoerdernde Chromosomen”), in Tumoren vermehrt sein
- **Progression** (gutartig zu bösartig); sequentielle Veränderungen in Chromosomen
- Klonalität & Genetische 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 Lonardo , Sergio Nasi , Simonetta Pulciani
Cancer: We Should Not Forget The Past
Journal of Cancer (2015), Vol. 6: 29-39
(for book cover & summary)

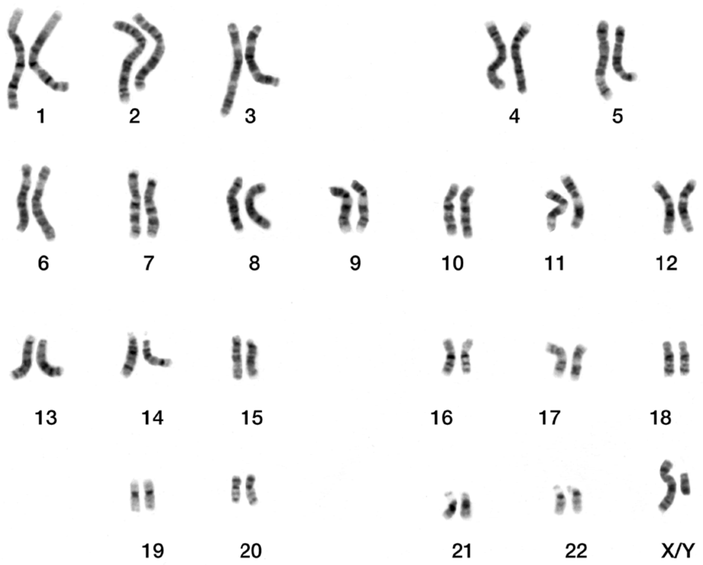
Zytogenetik

Genomik mit dem Mikroskop

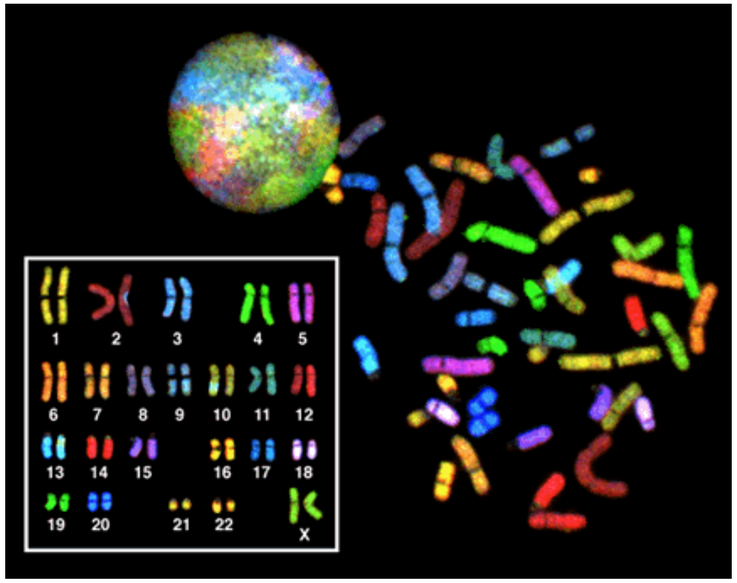


DNS/Histone cartoon modified from simpleclub.com

- 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 aktivierten Proteins in CML **die primäre Chemotherapie ersetzt**

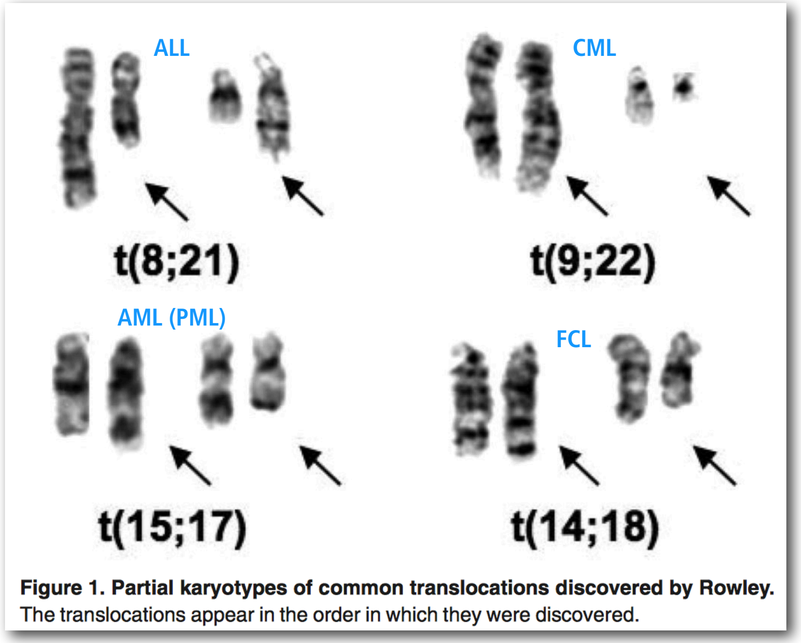
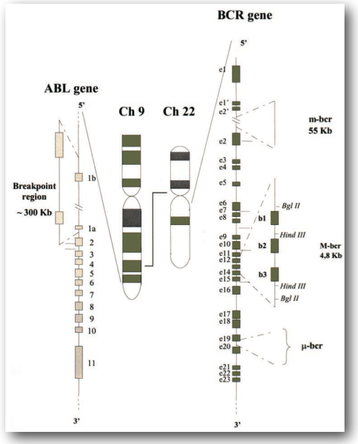
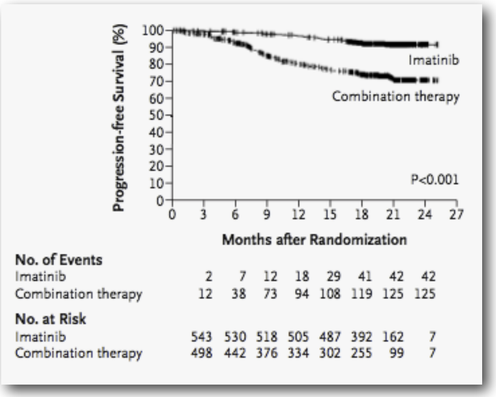


Figure 1. Partial karyotypes of common translocations discovered by Rowley. The translocations appear in the order in which they were discovered.

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



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



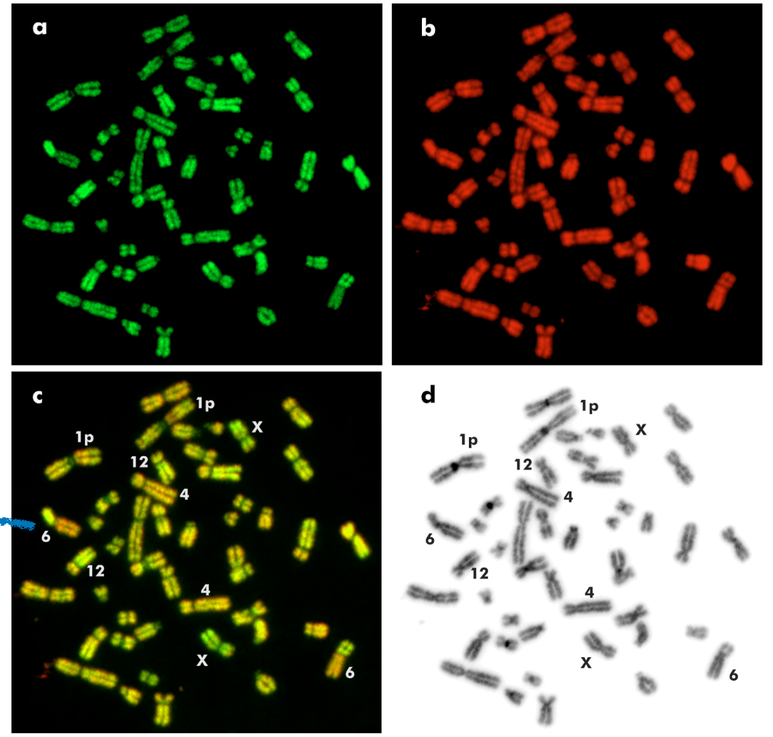
Event free Survival in first large Imatinib Trials

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

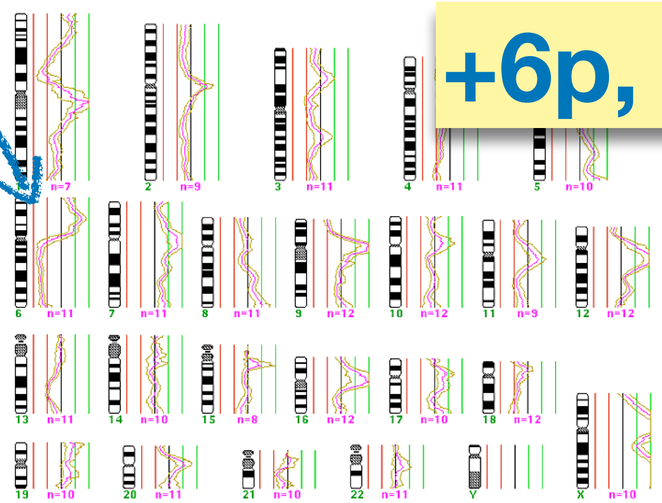
Vergleichende Genomische Hybridisierung (CGH)

Molekular-Zytogenetische Analyse Genomischer Imbalancen (CNV)

- Molekular-zytogenetische Technik zur Detektion chromosomaler Imbalancen
- **Hybridisierung** fluoreszenzmarkierter **genomischer** DNS mit normalen Metaphase Chromosomen
- Die Analyse der fluoreszierenden DNA erlaubt Aussagen über Regionen mit DNA Verlusten oder Zugewinnen
- **Indirekte** Aussagen über möglicherweise betroffene Gene anhand der Position der Signale



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



+6p, -6q...

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;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.

Cytogenetics

Molecular Cytogenetics

Genomics

The Human Genome Project 1990-2003


WANTED
20 Volunteers
to participate in the
Human Genome Project
a very large international scientific research effort.

The goal is to decode the human hereditary information (*human blueprint*) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

ROSWELL PARK Cancer Institute

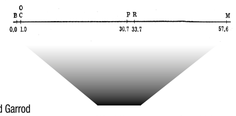
For more information please contact the Clinical Genetics Service 845-5720 (9:00 am - 3:00 pm) March 24 - 26, 1997

Landmarks in genetics and genomics

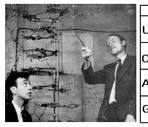
Gregor Mendel discovers laws of genetics
1865

Rediscovery of Mendel's work
1900

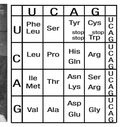
Archibald Garrod formulates the concept of human inborn errors of metabolism
1905



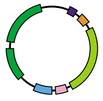
Alfred Henry Sturtevant makes the first linear map of genes
1913



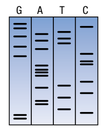
Oswald Avery, Colin MacLeod and Maclyn McCarty demonstrate that DNA is the hereditary material
1944



James Watson and Francis Crick describe the double-helical structure of DNA
1953



Marshall Nirenberg, Har Gobind Khorana and Robert Holley determine the genetic code
1966



Stanley Cohen and Herbert Boyer develop recombinant DNA technology
1972

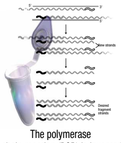
The Belmont Report on the use of human subjects in research is issued
1974

Frederick Sanger, Allan Maxam and Walter Gilbert develop DNA-sequencing methods
1977

GenBank database established
1982

First human disease gene — for Huntington's disease — is mapped with DNA markers
1983

First public discussion of sequencing the human genome
1984

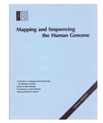


The polymerase chain reaction (PCR) is invented
1985

International Nucleotide Sequence Database Consortium formed

Muscular-dystrophy gene identified by positional cloning

First automated DNA-sequencing instrument developed
1986



First-generation human genetic map developed
1987

US National Research Council issues report on *Mapping and Sequencing the Human Genome*

Development of yeast artificial chromosome (YAC) cloning

Human Genome Organization (HUGO) formed
1988

Sequence-tagged sites (STS) mapping concept established

Cystic-fibrosis gene identified by positional cloning
1989

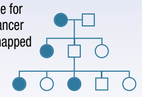
Human Genome Project
1990 2003

1990

The Human Genome Project (HGP) launched in the United States

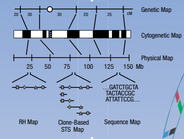
Ethical, legal and social implications (ELSI) programmes founded at the US National Institutes of Health (NIH) and Department of Energy (DOE)

First gene for breast cancer (*BRCA1*) mapped



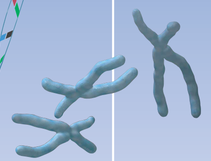
1991

First US genome centres established



1992

Second-generation human genetic map developed



Rapid-data-release guidelines established by the NIH and DOE

1993

New five-year plan for the HGP in the United States published

The Sanger Centre founded near Cambridge, UK, (later renamed the Wellcome Trust Sanger Institute)

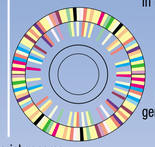
The Wellcome Trust

1994

The HGP's human genetic mapping goal achieved

1995

The HGP's human physical mapping goal achieved



First bacterial genome (*Haemophilus influenzae*) sequenced

US Equal Employment Opportunity Commission issues policy on genetic discrimination in the workplace

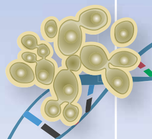
1996

First human gene map established

Pilot projects for human genome sequencing begin in the United States

First archaeal genome sequenced

Yeast (*Saccharomyces cerevisiae*) genome sequenced



The HGP's mouse genetic mapping goal achieved

Bermuda principles for rapid and open data release established

1997

The DOE forms the Joint Genome Institute

JGI
DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

US National Center for Human Genome Research becomes the National Human Genome Research Institute (NHGRI)

National Human Genome Research Institute

Escherichia coli genome sequenced

Genoscope (French National Genome Sequencing Center) founded near Paris

1998

Incorporation of 30,000 genes into human genome map

New five-year plan for the HGP in the United States published

RIKEN Genomic Sciences Center established in Japan

Roundworm (*Caenorhabditis elegans*) genome sequenced

Single-nucleotide polymorphism (SNP) initiative begins

GTGCT
GTCCT

Chinese National Human Genome Centers established in Beijing and Shanghai

1999

Full-scale human genome sequencing begins

Sequence of first human chromosome (chromosome 22) completed

2000

Draft version of human genome sequence completed

US President Bill Clinton and UK Prime Minister Tony Blair support free access to genome information

Fruitfly (*Drosophila melanogaster*) genome sequenced

Mustard cress (*Arabidopsis thaliana*) genome sequenced

Executive order bans genetic discrimination in US federal workplace

2001

Draft version of human genome sequence published

10,000 full-length human cDNAs sequenced

Mammalian Gene Collection

2002

Draft version of mouse genome sequence completed and published

Draft version of rice genome sequence completed and published

Draft version of rat genome sequence completed

2003

Finished version of human genome sequence completed

The HGP ends with all goals achieved

To be continued...

PEAS COURTESY J. BLAMIRE, CITY UNIV. NEW YORK; WATSON & CRICK COURTESY A. BARRINGTON BROWN/SPL; *SCIENCE* COVERS COURTESY AAAS

The Human Genome Project 1990-2003


WANTED
20 Volunteers
to participate in the
Human Genome Project
a very large international scientific research effort.

The goal is to decode the human hereditary information (*human blueprint*) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

No personal information will be maintained or transferred.

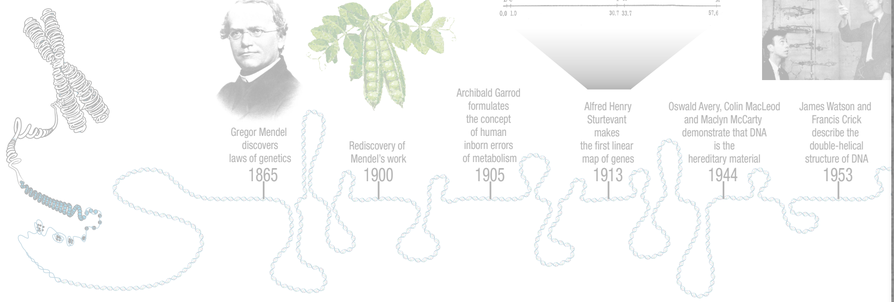
Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

ROSWELL PARK CANCER INSTITUTE

For more information please contact the Clinical Genetics Service
845-5720 (9:00 am - 3:00 pm)
March 24 - 28, 1997

Landmarks in genetics and genomics



Gregor Mendel discovers laws of genetics
1865

Rediscovery of Mendel's work
1900

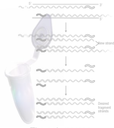
Archibald Garrod formulates the concept of human inborn errors of metabolism
1905

Alfred Henry Sturtevant makes the first linear map of genes
1913

Oswald Avery, Colin MacLeod and Maclyn McCarty demonstrate that DNA is the hereditary material
1944

James Watson and Francis Crick describe the double-helical structure of DNA
1953

First public discussion of sequencing the human genome
1984


The polymerase chain reaction (PCR) is invented
1985

International Nucleotide Sequence Database Consortium formed

Muscular-dystrophy gene identified by positional cloning

First automated DNA-sequencing instrument developed
1986


First-generation human genetic map developed
1987

US National Research Council issues report on *Mapping and Sequencing the Human Genome*

Development of yeast artificial chromosome (YAC) cloning

Human Genome Organization (HUGO) formed
1988

Sequence-tagged sites (STS) mapping concept established

Cystic-fibrosis gene identified by positional cloning
1989

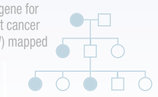
Human Genome Project
1990 – 2003

1990

The Human Genome Project (HGP) launched in the United States

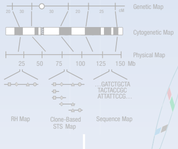
Ethical, legal and social implications (ELSI) programmes founded at the US National Institutes of Health (NIH) and Department of Energy (DOE)

First gene for breast cancer (*BRCA1*) mapped



1991

First US genome centres established



1992

Second-generation human genetic map developed

Rapid-data-release guidelines established by the NIH and DOE

1993

New five-year plan for the HGP in the United States published

The Sanger Centre founded near Cambridge, UK, (later renamed the Wellcome Trust Sanger Institute)

The Wellcome Trust

1994

The HGP's human genetic mapping goal achieved


mapping goal achieved

Bermuda principles for rapid and open data release established

National Genome Sequencing Center) founded near Paris

Single-nucleotide polymorphism (SNP) initiative begins

GTGCT
GTCCT

Chinese National Human Genome Centers established in Beijing and Shanghai

1999

Full-scale human genome sequencing begins

Sequence of first human chromosome (chromosome 22) completed

2000

Draft version of human genome sequence completed

US President Bill Clinton and UK Prime Minister Tony Blair support free access to genome information

Fruitfly (*Drosophila melanogaster*) genome sequenced

Mustard cress (*Arabidopsis thaliana*) genome sequenced

Executive order bans genetic discrimination in US federal workplace

2001

Draft version of human genome sequence published

10,000 full-length human cDNAs sequenced

Mammalian Gene Collection

2002

Draft version of mouse genome sequence completed and published

Draft version of rice genome sequence completed and published

Draft version of rat genome sequence completed

2003

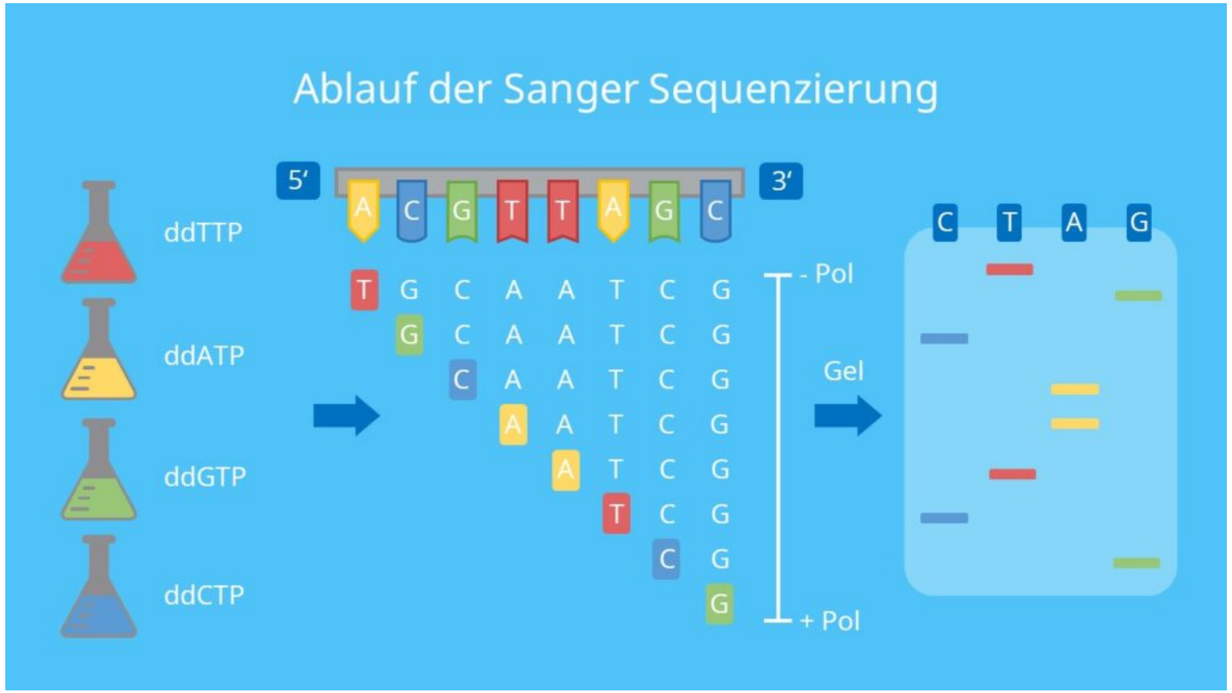
Finished version of human genome sequence completed

The HGP ends with all goals achieved

To be continued...

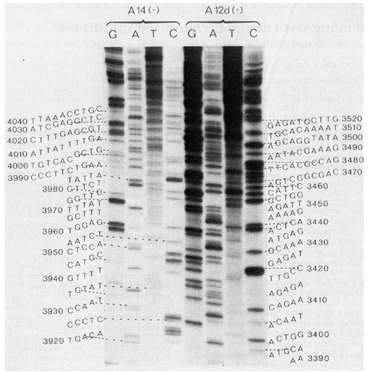
PEAS COURTESY J. BLAMIRE, CITY UNIV. NEW YORK; WATSON & CRICK COURTESY A. BARRINGTON BROWN/SPL; *SCIENCE* COVERS COURTESY AAAS

DNS Sequenzierung - Der traditionelle Weg



https://studyflix.de/biologie/dna-sequenzierung

https://wi.mit.edu/news/whitehead-human-genome-map-ushers-final-phase-us-human-genome-project



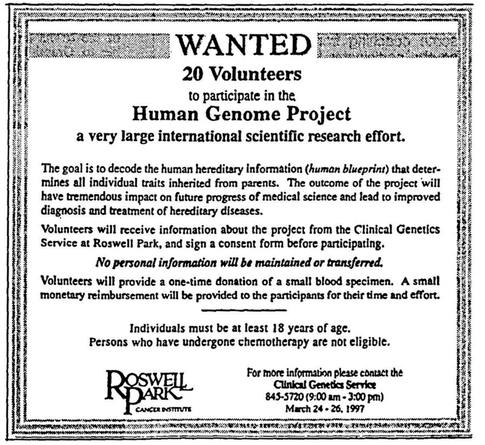
DNA sequencing with chain-terminating inhibitors. F. Sanger , S. Nicklen , A. R. Coulson. PNAS 1977

- Das humane Genomprojekt basierte auf "traditioneller" **Sanger-Sequenzierung** ("Kettenterminierung")
- die schrittweise DNA-Verlängerung auf Basis klonierter DNA Stücke, mit Sequenzlänge meist <1'000 Basen (bei einem ~3'200'000'000bp Genom)

Das Humanegenomprojekt

... lieferte einen Atlas, kein Telephonbuch

- **internationales**, mehrjähriges **Projekt** zur Entschlüsselung des menschlichen Genoms mit Hauptbeiträgen aus USA und U.K.
- weder Genom einer Person noch "repräsentative Mischung"
- "Entwurf" des kompletten Genoms in 2000; in 2003 ca. 90% fertig
- basierend auf **traditionellen Klonierungs-/ Sequenzierungstechniken**
- Kosten von ca. 3 Milliarden $
- **Meilenstein** für die verbreitetet Entwicklung und Anwendung molekularer Techniken in Biomedizinischer Forschung


WANTED

20 Volunteers

to participate in the

Human Genome Project

a very large international scientific research effort.

The goal is to decode the human hereditary information (*human blueprint*) that determines all individual traits inherited from parents. The outcome of the project will have tremendous impact on future progress of medical science and lead to improved diagnosis and treatment of hereditary diseases.

Volunteers will receive information about the project from the Clinical Genetics Service at Roswell Park, and sign a consent form before participating.

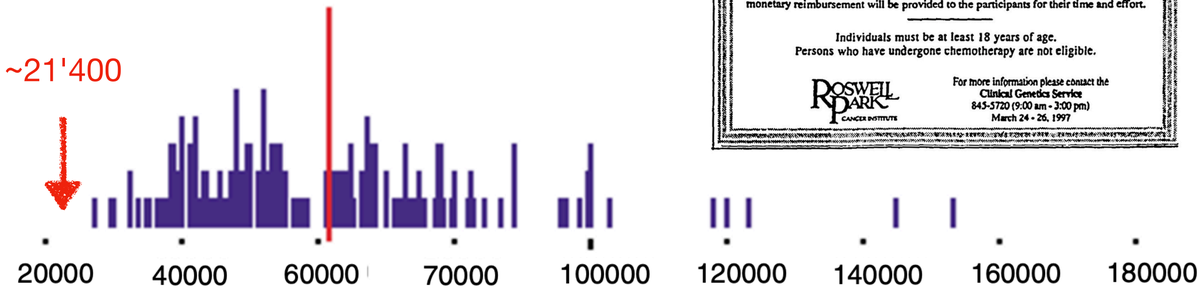
No personal information will be maintained or transferred.

Volunteers will provide a one-time donation of a small blood specimen. A small monetary reimbursement will be provided to the participants for their time and effort.

Individuals must be at least 18 years of age.
Persons who have undergone chemotherapy are not eligible.

ROSWELL PARK CANCER INSTITUTE

For more information please contact the Clinical Genetics Service 845-5720 (9:00 am - 3:00 pm) March 24 - 26, 1997



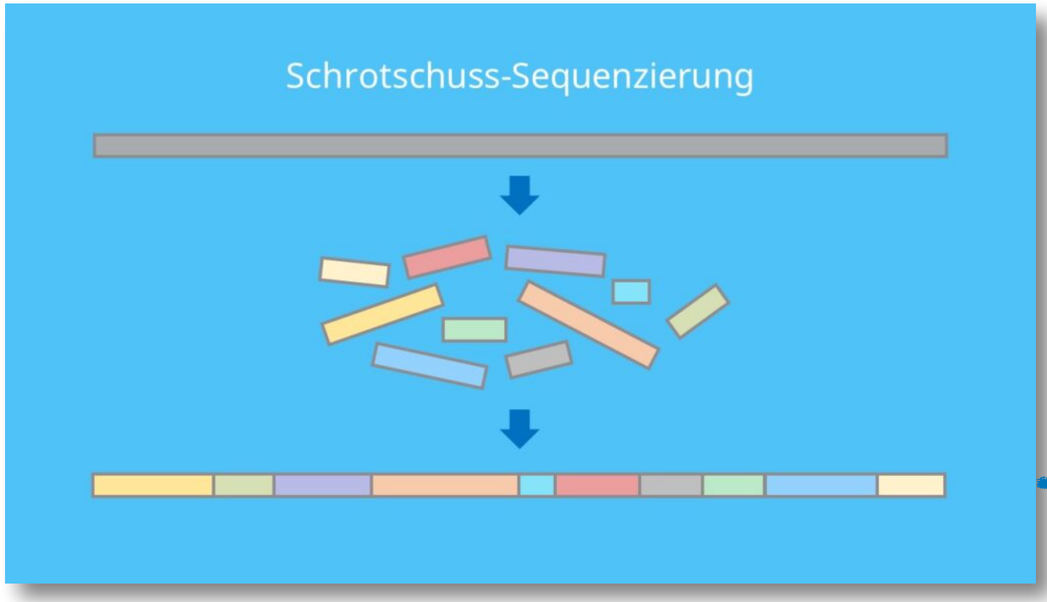
Genesweep 2000, eine Umfrage (mit Wette) von Ewan Bierney (heute Vizedirektor des European Molecular Biology Laboratory) zur **Anzahl der proteinkodierenden Gene** im menschlichen Genom. Die meisten Zahlen waren deutlich zu hoch.

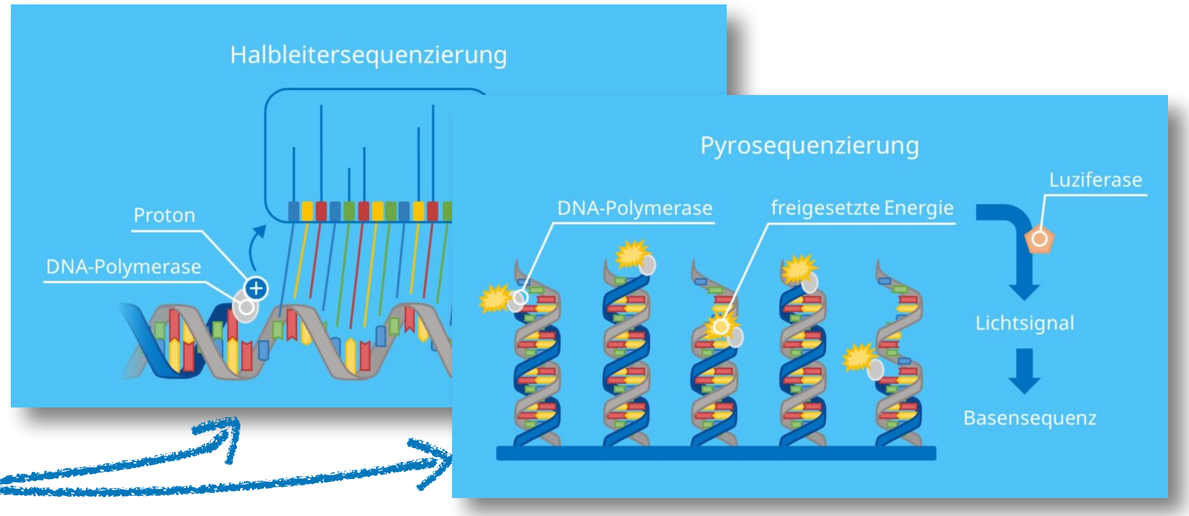
CREDIT: ILLUSTRATION BY C. FABER SMITH/SCIENCE (Graphik Science, 2007)

President Bill Clinton und Francis Collins, Leiter des HGP, bei der Vorstellung der HGP Resultate im Weissen Haus, Juni 2000 (Quelle: genome.gov)

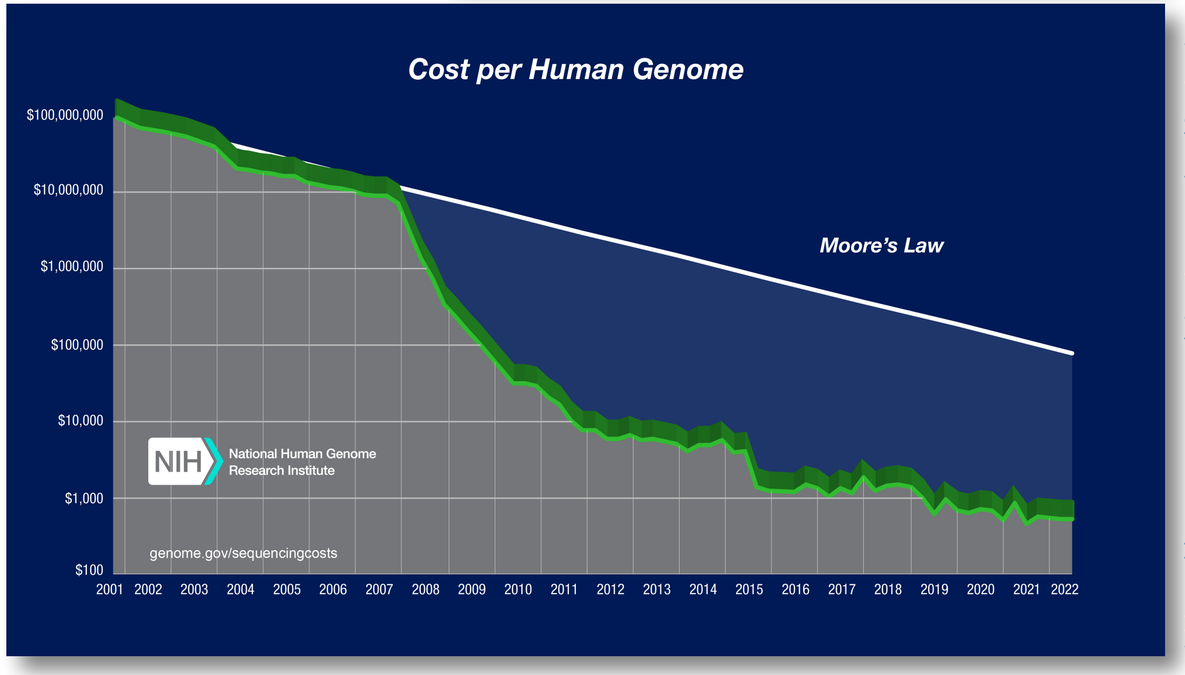
DNS Sequenzierung - "Next Generation Sequencing"

https://studyflix.de/biologie/dna-sequenzierung





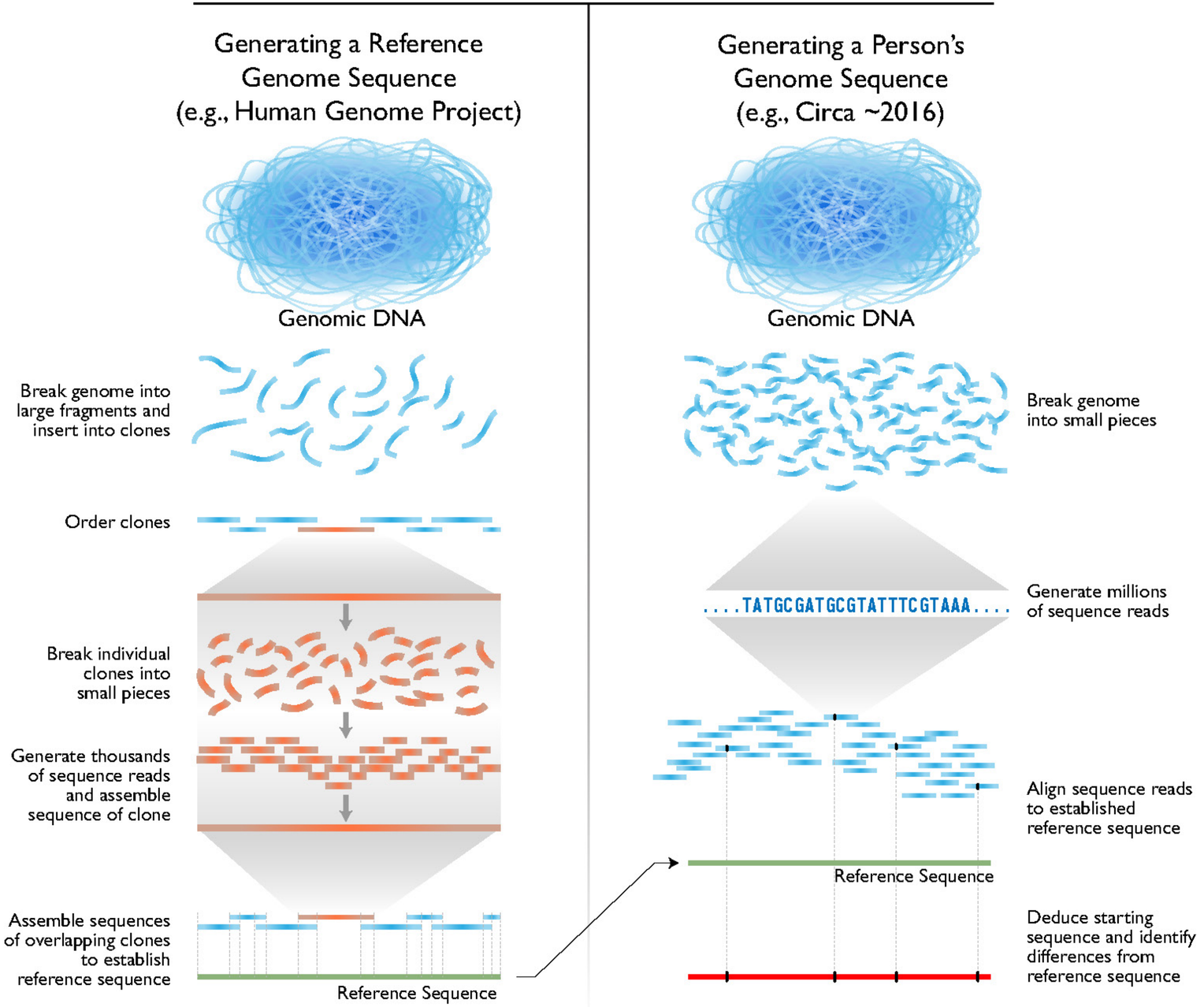
- Seit ca. 2006 radikaler Preisverfall durch "Next Generation Sequencing" Technologien
- "**Schrotschuss**-Squenzierung":
  - DNS wird zufällig fragmentiert
  - alle Fragmente werden *parallel* sequenziert
  - **Bioinformatik** zum Zusammensetzen, mit Hilfe eines ***Referenzgenoms***



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

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

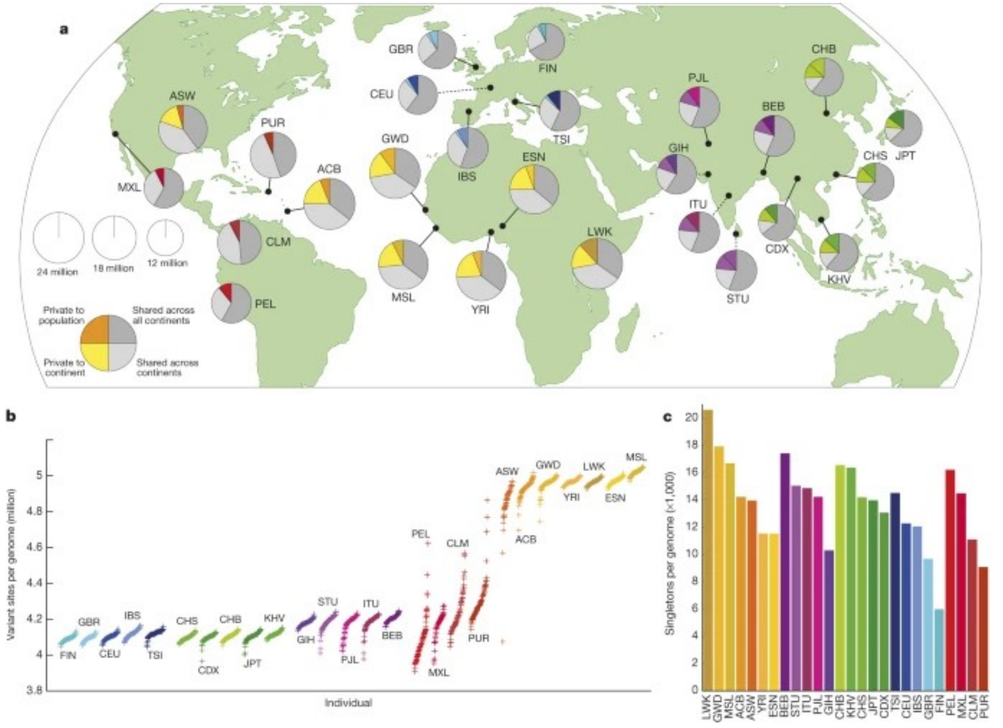
Human Genome Sequencing



Das "1000 Genome" Projekt

Genomische Varianten in Genen und nichtkodierender DNS

- **genetische Variabilität** zwischen Individuen durch Unterschieden in Millionen von einzelnen Basen (SNPs) und Tausenden von grösseren Abschnitten
- Das "1000 Genome" Projekt kartierte die Sequenzen von 2,504 Individuen aus 26 Populationen (2007-2013)
- Insgesamt ca. 80 Millionen Sequenzvarianten gefunden
- technisch ermöglicht durch "Next Generation Sequencing", **Bioinformatik** und Computertechnologie

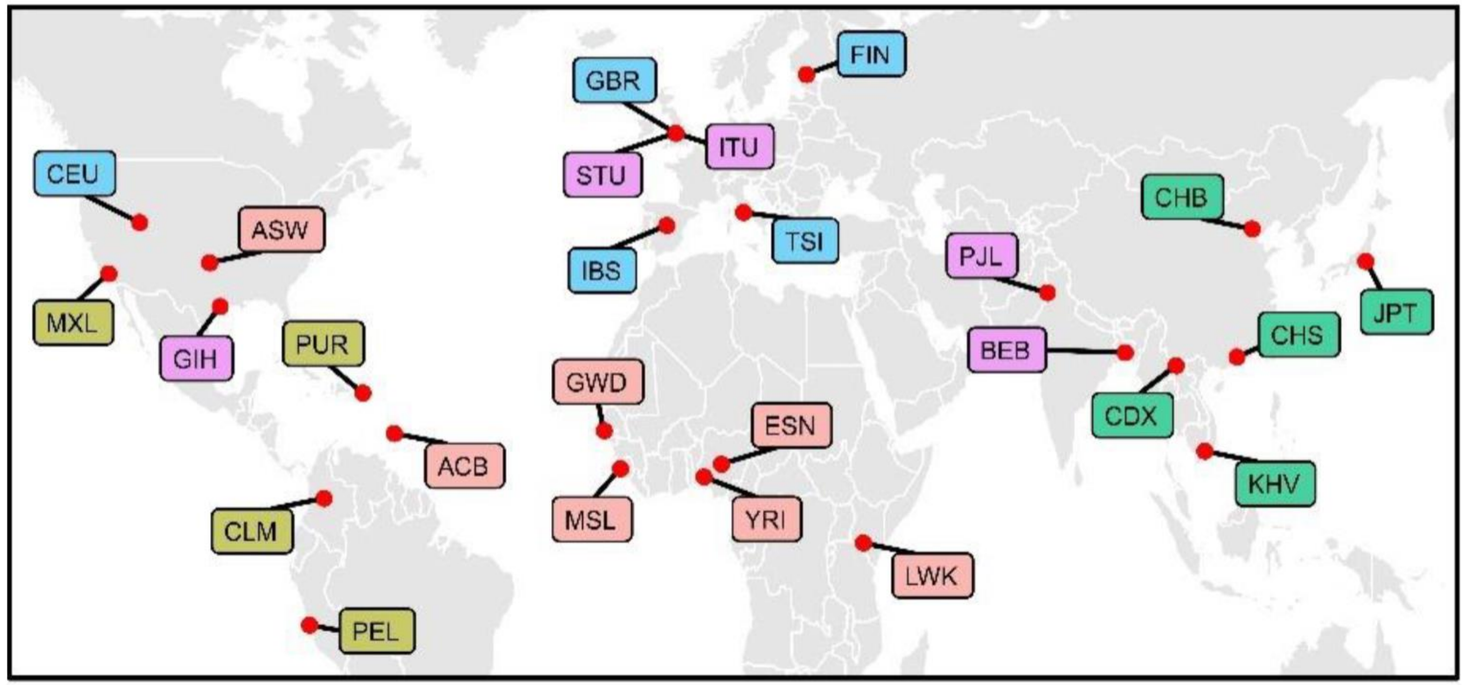


Häufigkeit variabler genomischer Loci in den verschiedenen Populationen im 1kG Projekt. In b) zeigt sich die individuelle Anzahl von polymorphen Variationen während c) die Anzahl einmaliger Sequenzvarianten per Population darstellt.

- Ein Individuum unterscheidet sich in **~5 Millionen Sequenzvarianten** und einigen Tausenden grösseren Abschnitten von einem fiktiven Referenzgenom.
- Die meisten Varianten eines Individuums sind häufig (d.h. in mehreren % der Population)
- Die Masse der Varianten ist **selten** (d.h. viele einmalige oder seltene Varianten summieren sich).
- Verschiedenen Populationen zeigen einzelne Varianten gehäuft oder in Kombination - doch die genetische Variabilität zwischen Individuen ist viel höher als zwischen Populationen.

Background

1000 Genome Projects



Superpopulation

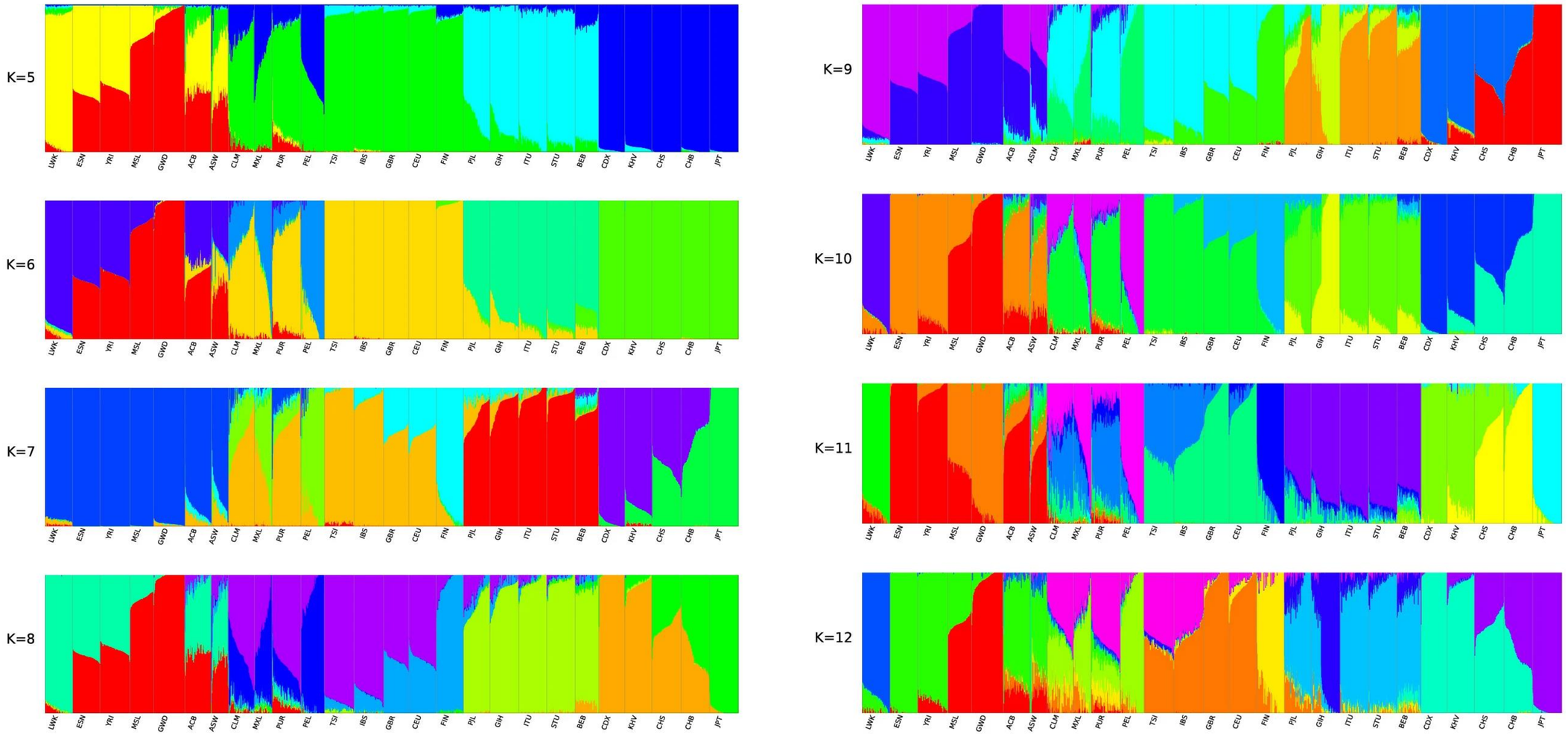
Population

Superpopulation	Population
African Ancestry	ACB African Caribbean, Barbados
	ASW African American in Southwest US, US
	ESN Esan, Nigeria
	GWD Mandinka Gambian, The Gambia
	LWK Luhya in Webuye, Kenya
	MSL Mende, Sierra Leone
	YRI Yoruba in Ibadan, Nigeria
American Ancestry	CLM Colombian in Medellin, Colombia
	MXL Mexican Ancestry in California, US
	PEL Peruvian in Lima, Peru
	PUR Puerto Rican, US
East Asian Ancestry	CDX Chinese Dai in Xishuangbanna, China
	CHB Han Chinese in Beijing, China
	CHS Han Chinese South, China
	JPT Japanese in Tokyo, Japan
	KHV Kinh in Ho Chi Minh City, Vietnam
European Ancestry	CEU Northwest European Ancestry, US
	FIN Finnish, Finland
	GBR British, England and Scotland
	IBS Iberian, Spain
	TSI Toscani, Italy
Southeast Asian Ancestry	BEB Bengali, Bangladesh
	GIH Gujarati Indians, TX, US
	ITU Indian Telugu, UK
	PJL Punjabi in Lahore, Pakistan
	STU Sri Lankan Tamil, UK

Image from https://www.internationalgenome.org/data-portal/population

Background

Population structure as inferred using the ADMIXTURE program for K = 5 to 12 in 1000 genomes project

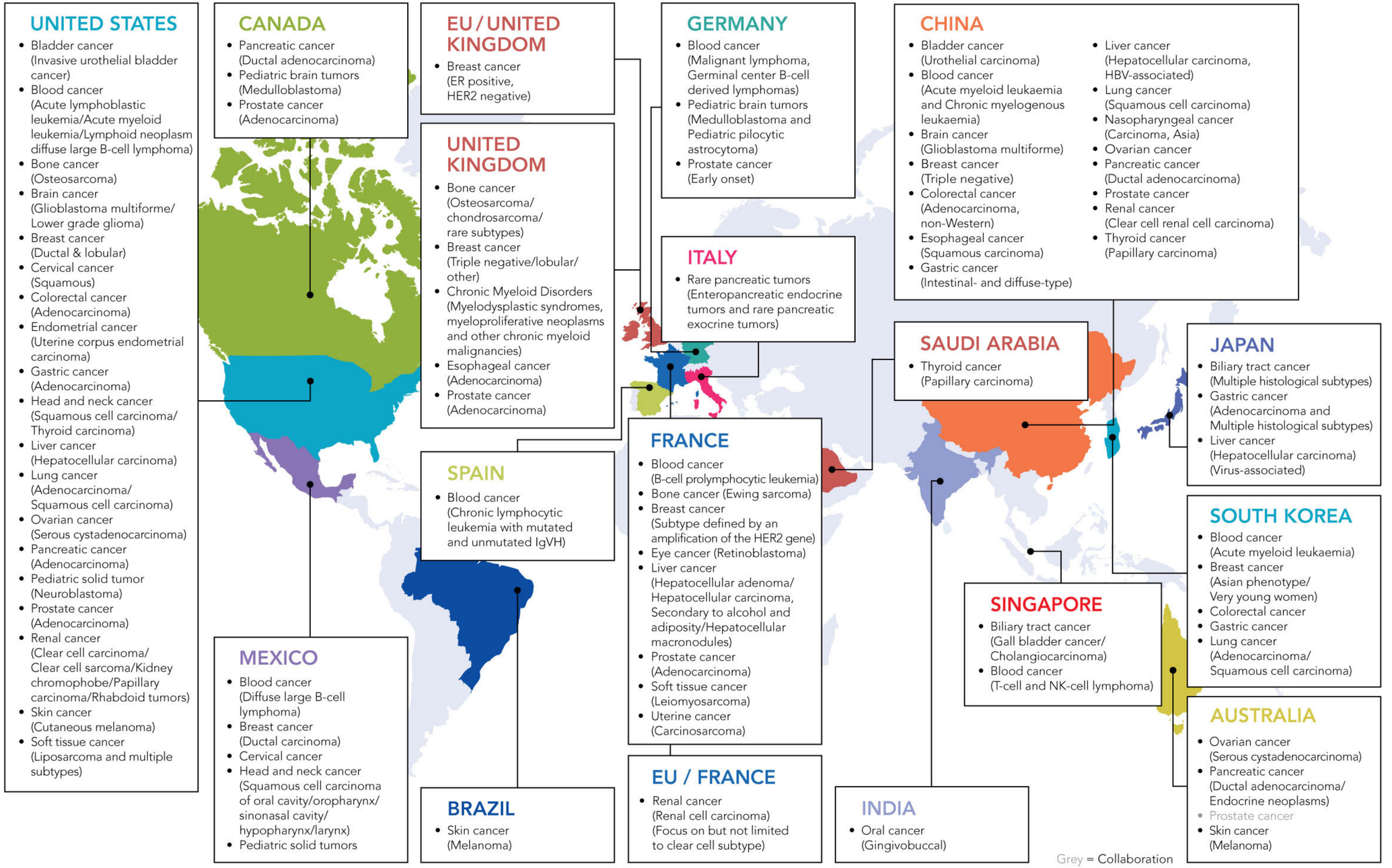


The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 526, 68–74 (2015). https://doi.org/10.1038/nature15393

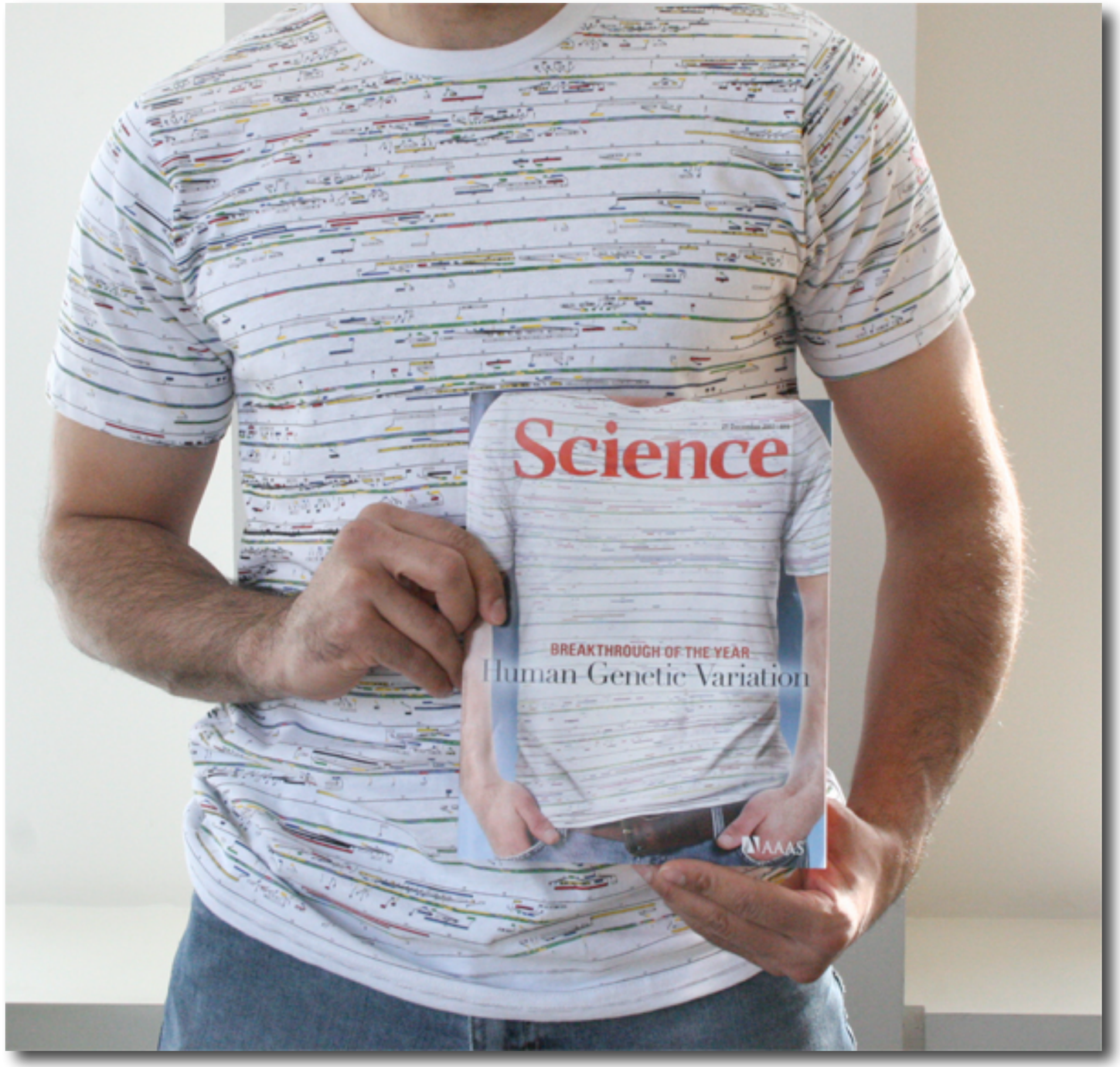
ICGC

International Cancer Genomics Consortium

Internationally coordinated research studies for cancer genome landscape analysis



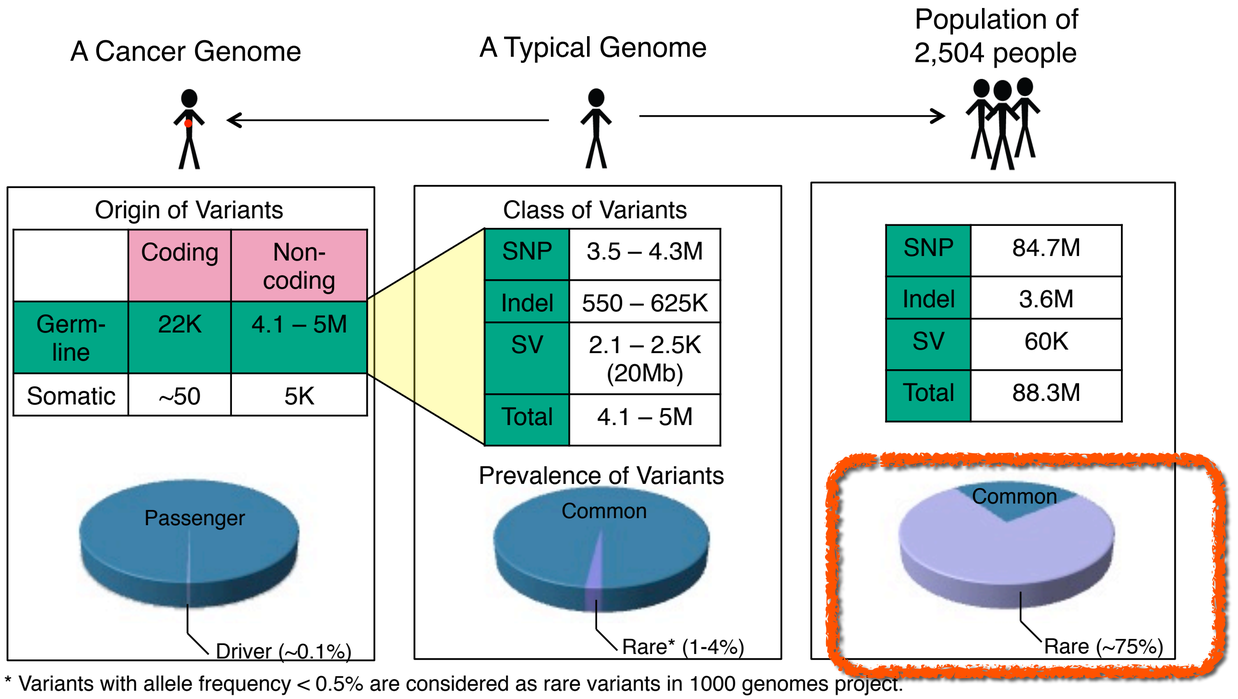
The trouble with human genome variation



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



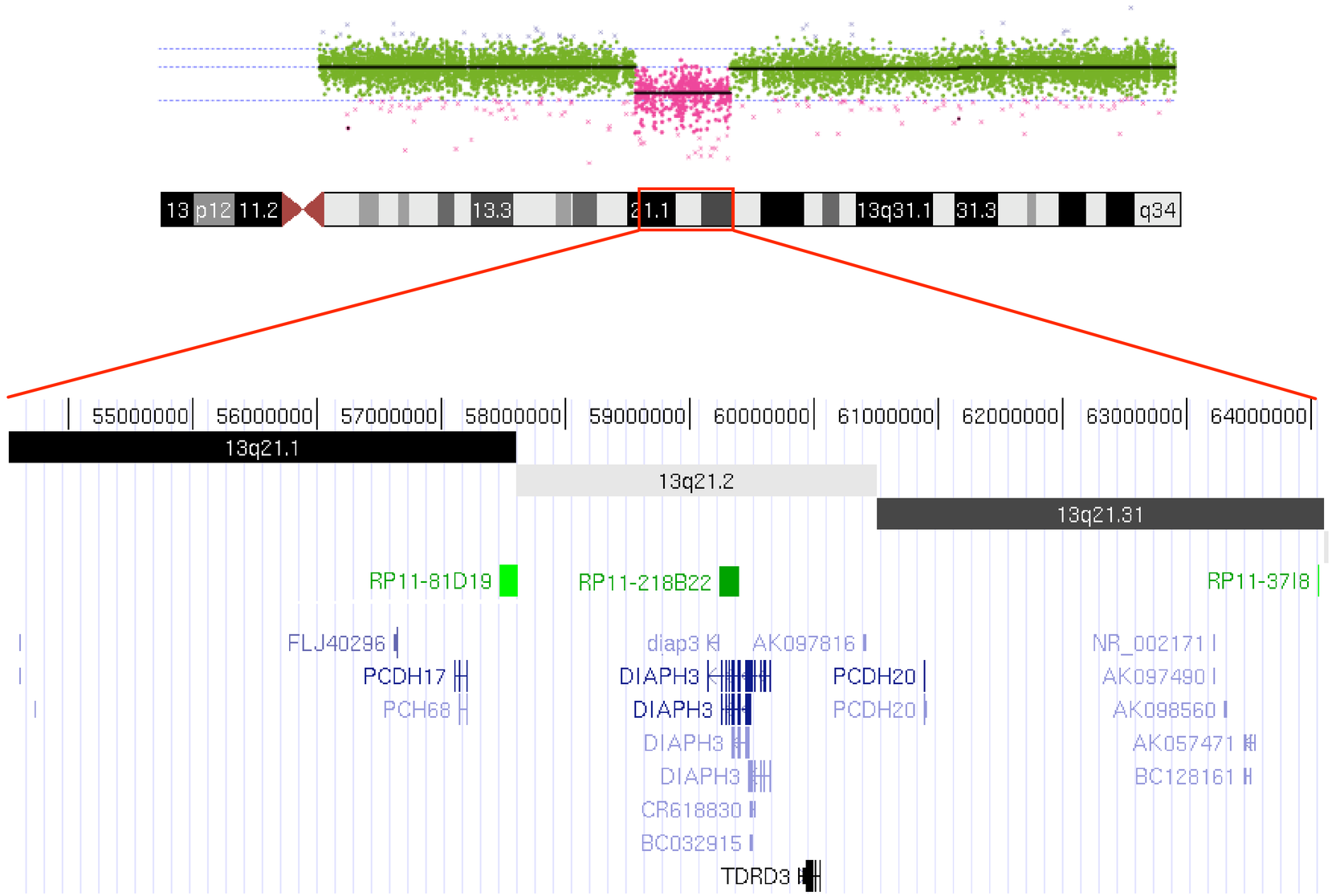
* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

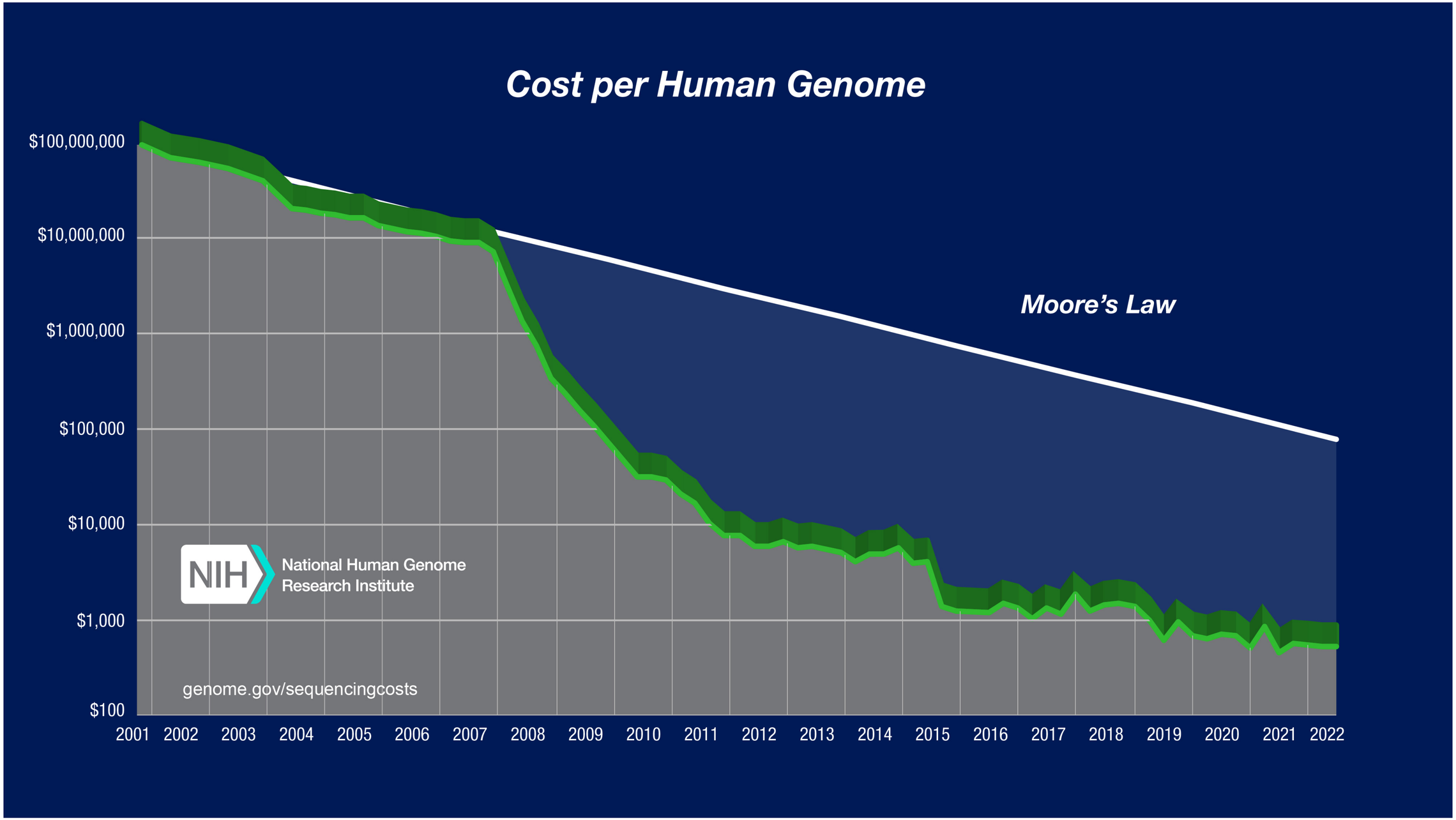
The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

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 Eggermann, and Herdit M. Schüler
American Journal of Medical Genetics Part A 146A:2417 – 2420 (2008)





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

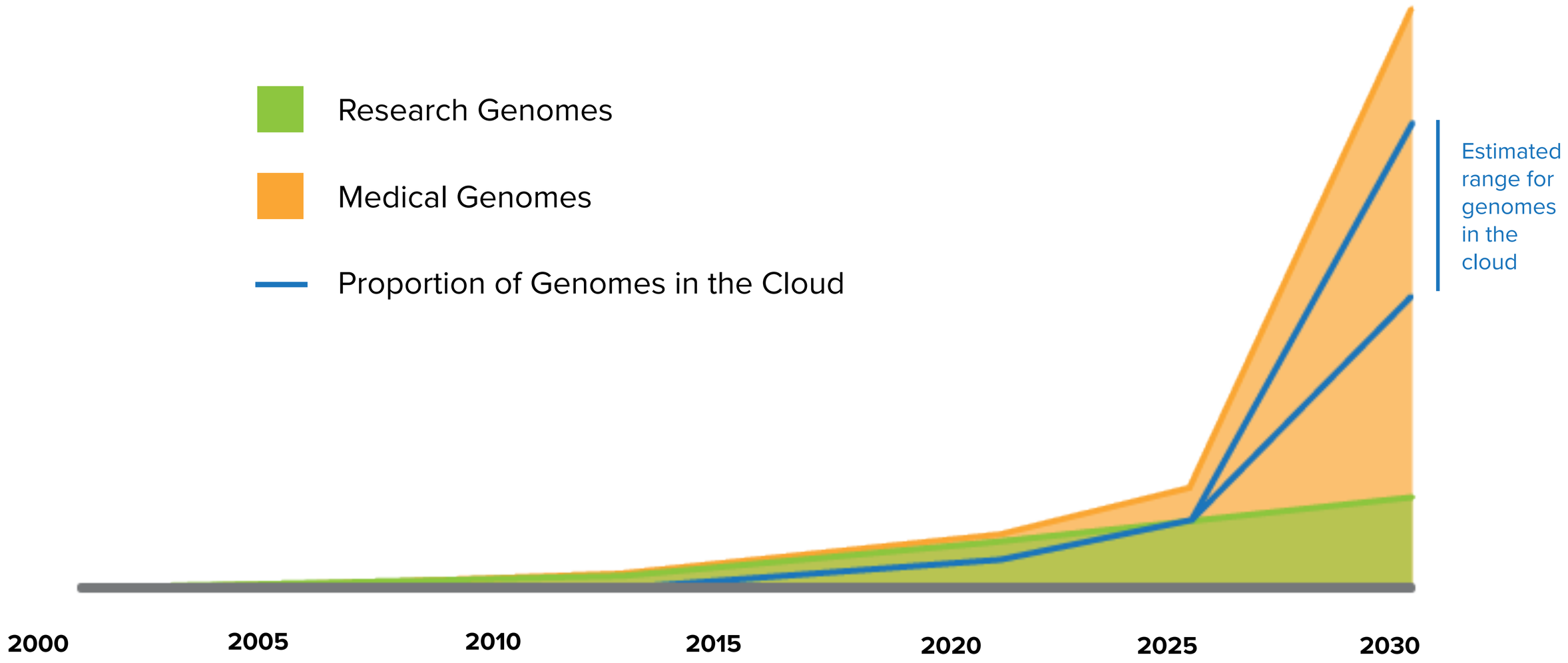
Clinical/Genomic Medicine

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

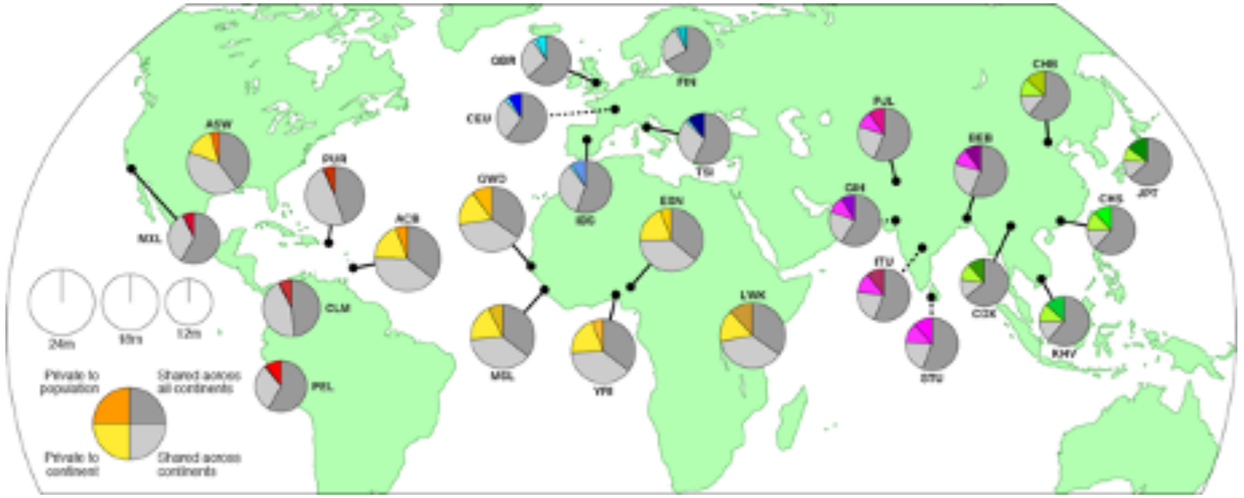
National cohorts > 100k genotyped or sequenced at least 25k

National cohorts > 100k people active collection now

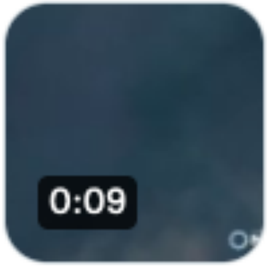
Planning national cohorts > 100k

New technologies and new applications

Global Alliance
for Genomics & Health



Oxford Nanopore @nanopore · 5 Oct



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[UZH]

BIO392
Bioinformatics of Genome Variations

Genomic Analysis Technologies...

Michael Baudis **UZH SIB**
Computational Oncogenomics

Non-Sequence™ Genomics

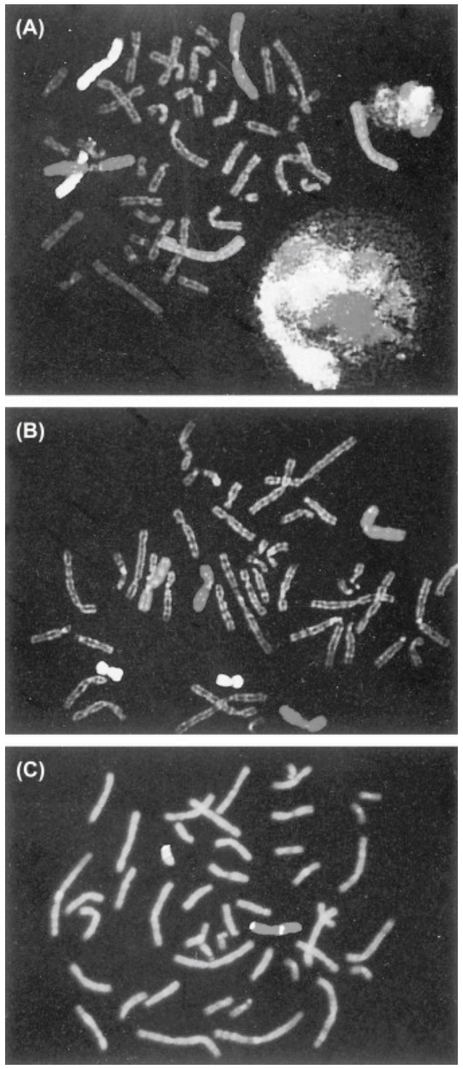
(Molecular) Cytogenetics

- structural changes in chromosomes can be analyzed by cytogenetics (chromosomal analysis) w/o knowledge about sequence alterations
- cytogenetics may me 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**

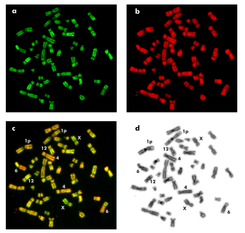
G-banded metaphase with three marker chromosomes (whose origin from chromosomes 22, 11, and 14 cannot be identified by this technique).

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).

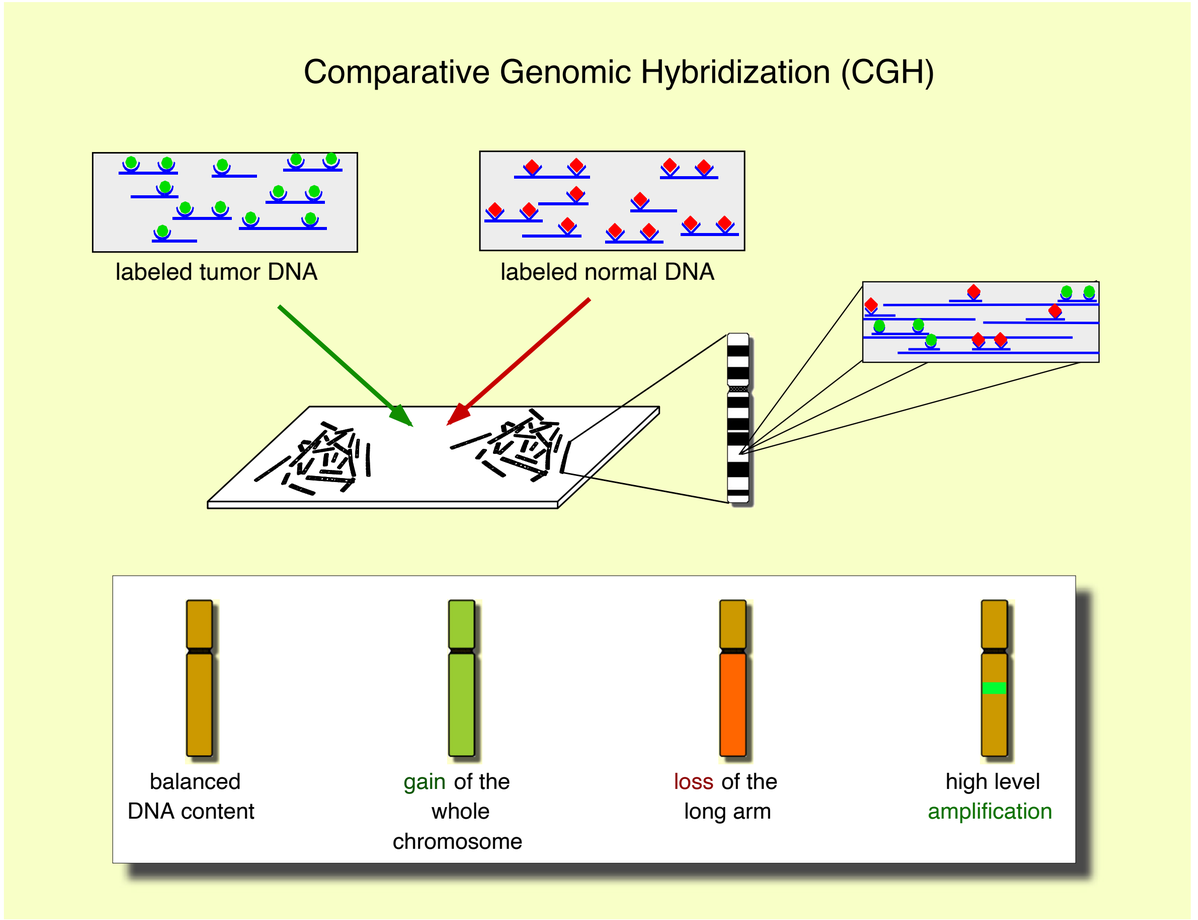
Comparative Genomic Hybridization



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;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.



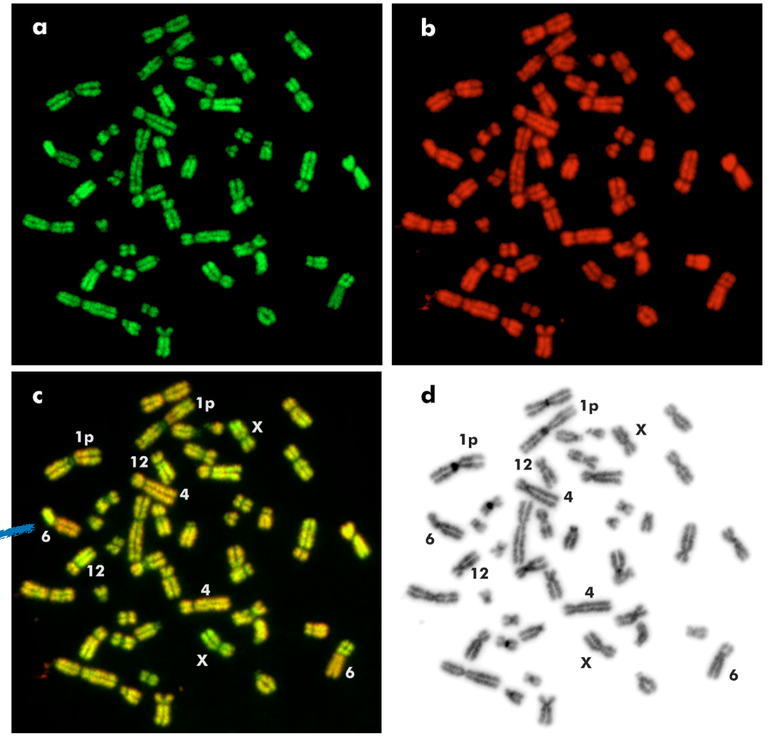
Chromosomal CGH: Normal metaphase spreads (cultured lymphocytes from healthy donors) on microscopy slides serve as the hybridization matrix for whole-genome DNA from tumor and reference tissue, labeled with different fluorophores. The regional ratio between the two colors points to (relative) changes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

Comparative Genomic Hybridization

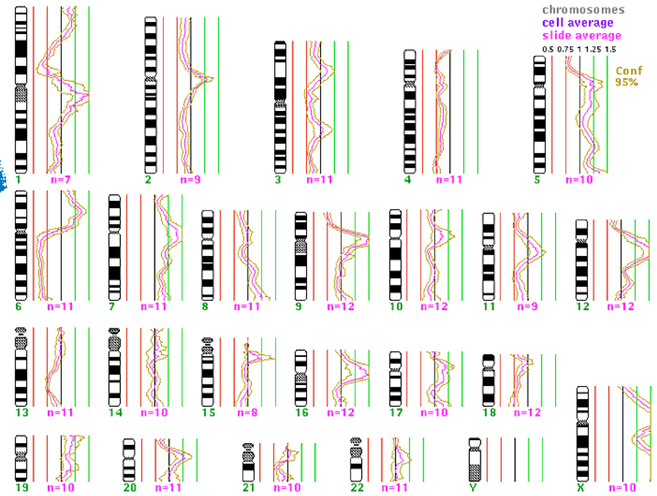
Molecular-Cytogenetic Technology for Genomic Imbalance Screening

- 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)

- 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;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.



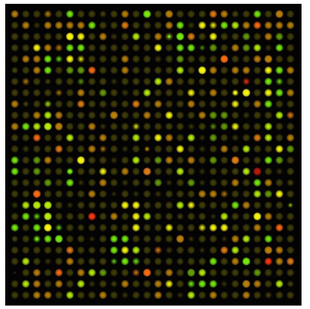
CGH-Experiment: **a** Hybridisierung mit Tumor-DNA; **b** Hybridisierung mit normaler menschlicher DNA als Kontrolle; **c** Überlagerung der Signale; **d** Bänderungsfärbung zur Identifizierung der Chromosomen



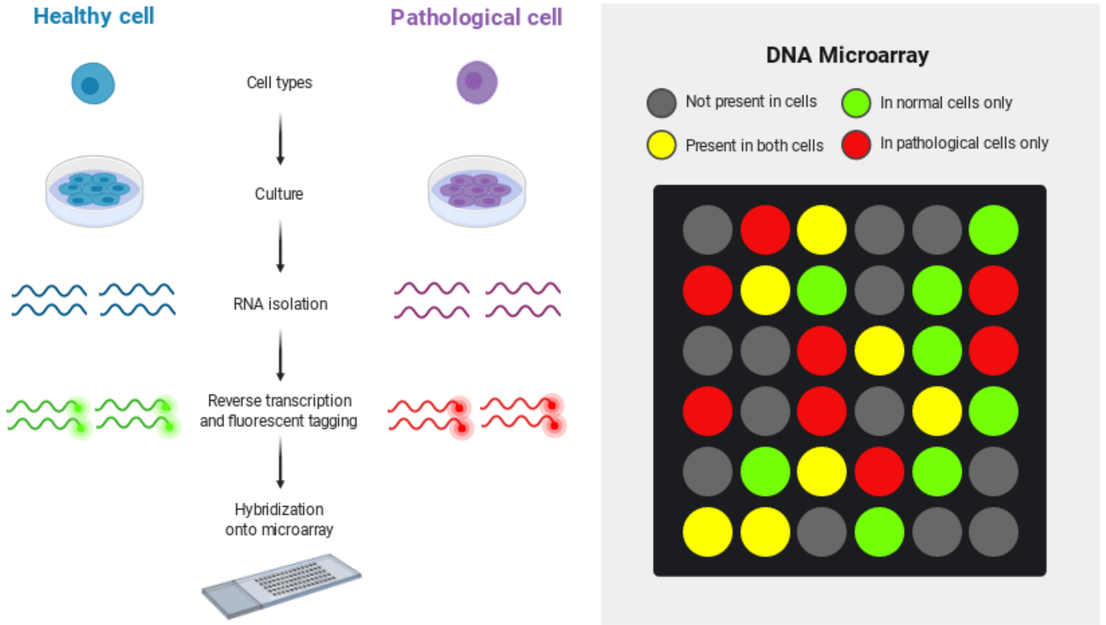
Auswertung: Summationsprofil der computergestützten Analyse mehrerer Metaphasen des dargestellten Falles; die Profilausschläge stehen für Zugewinne bzw. Verluste von chromosomalen Anteilen im Tumorgenom

Array Comparative Genomic Hybridization

Molecular Hybridization Technology for Genomic Imbalance Screening



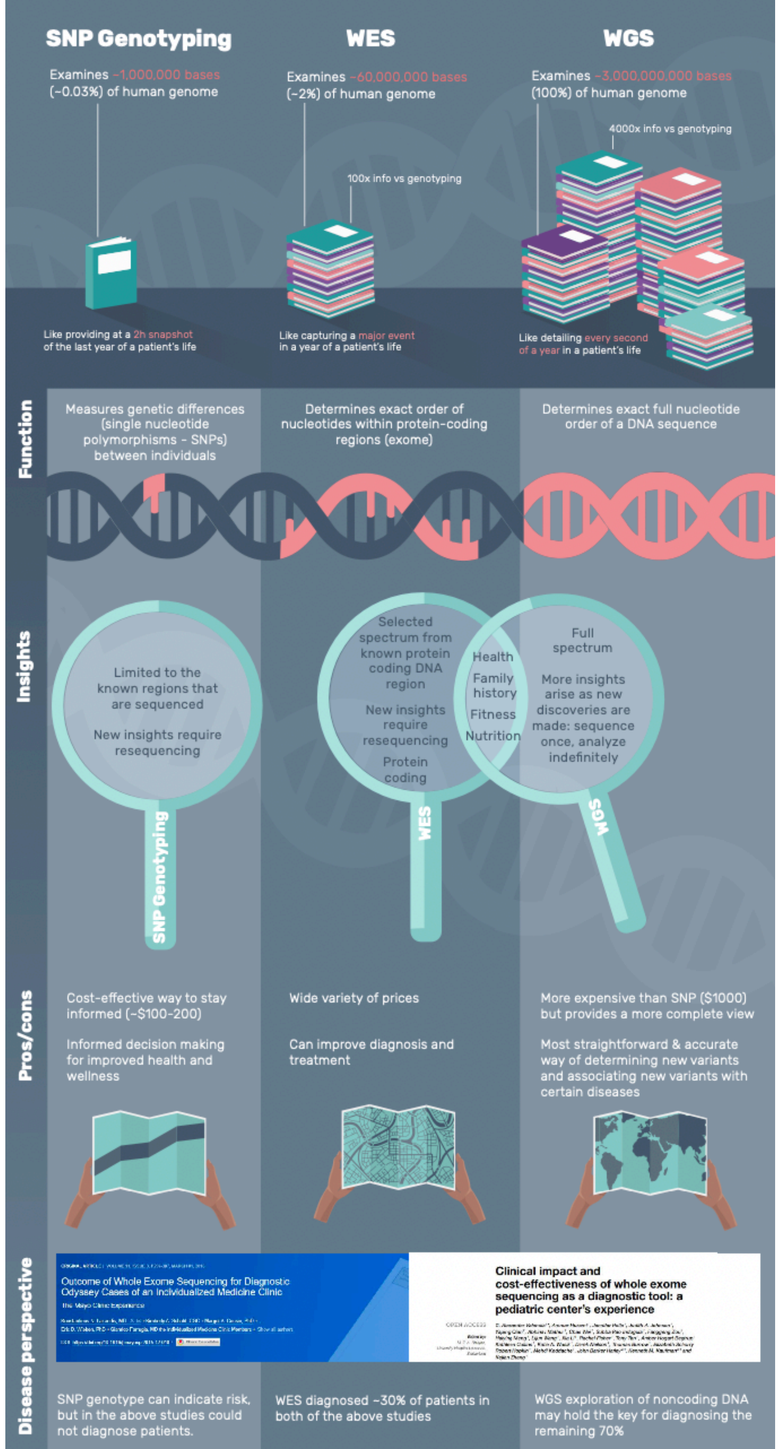
- Molecular-cytogenetic technique to identify genomic copy number variations (CNV/ CNA) for given sequences (multi-kb to Mb)
- based on ***in situ*** suppression **hybridization** of labeled **genomic** tumor and reference DNA against spotted or synthesized DNA clones or oligonucleotides
- analysis of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

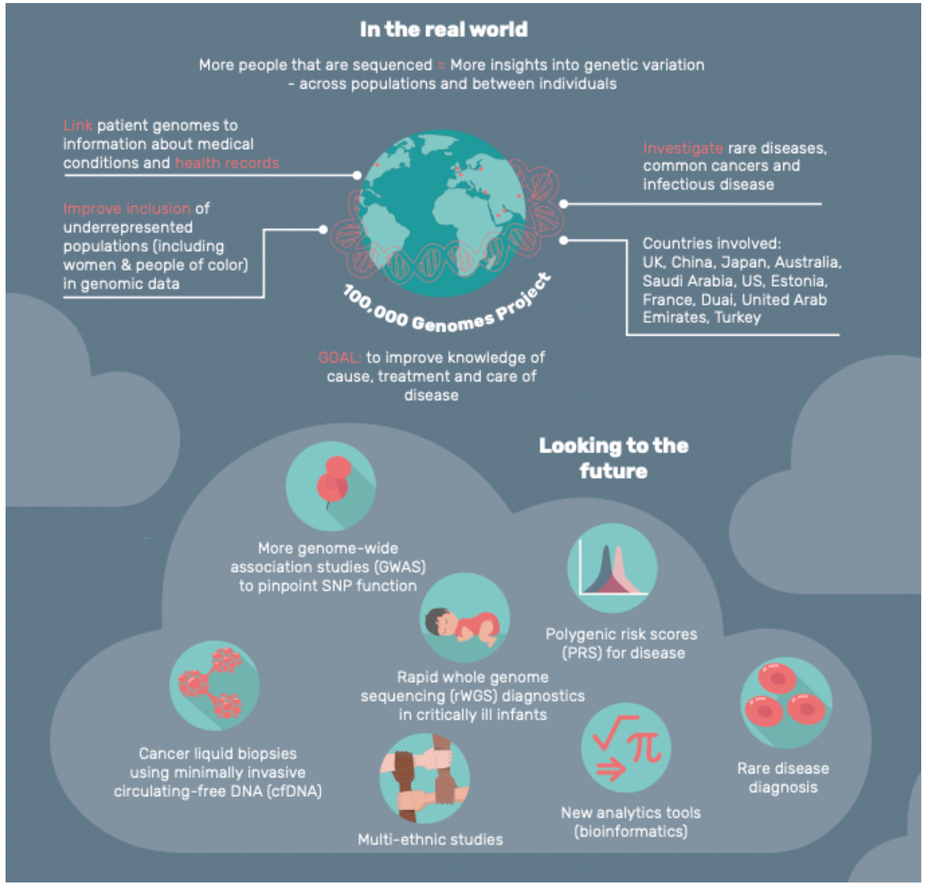


- Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes Cancer. 1997; 4 (20):399-407.
- Albertson DG, Pinkel D. Genomic microarrays in human genetic 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.

Genome Analysis

A “progressing technologies” view



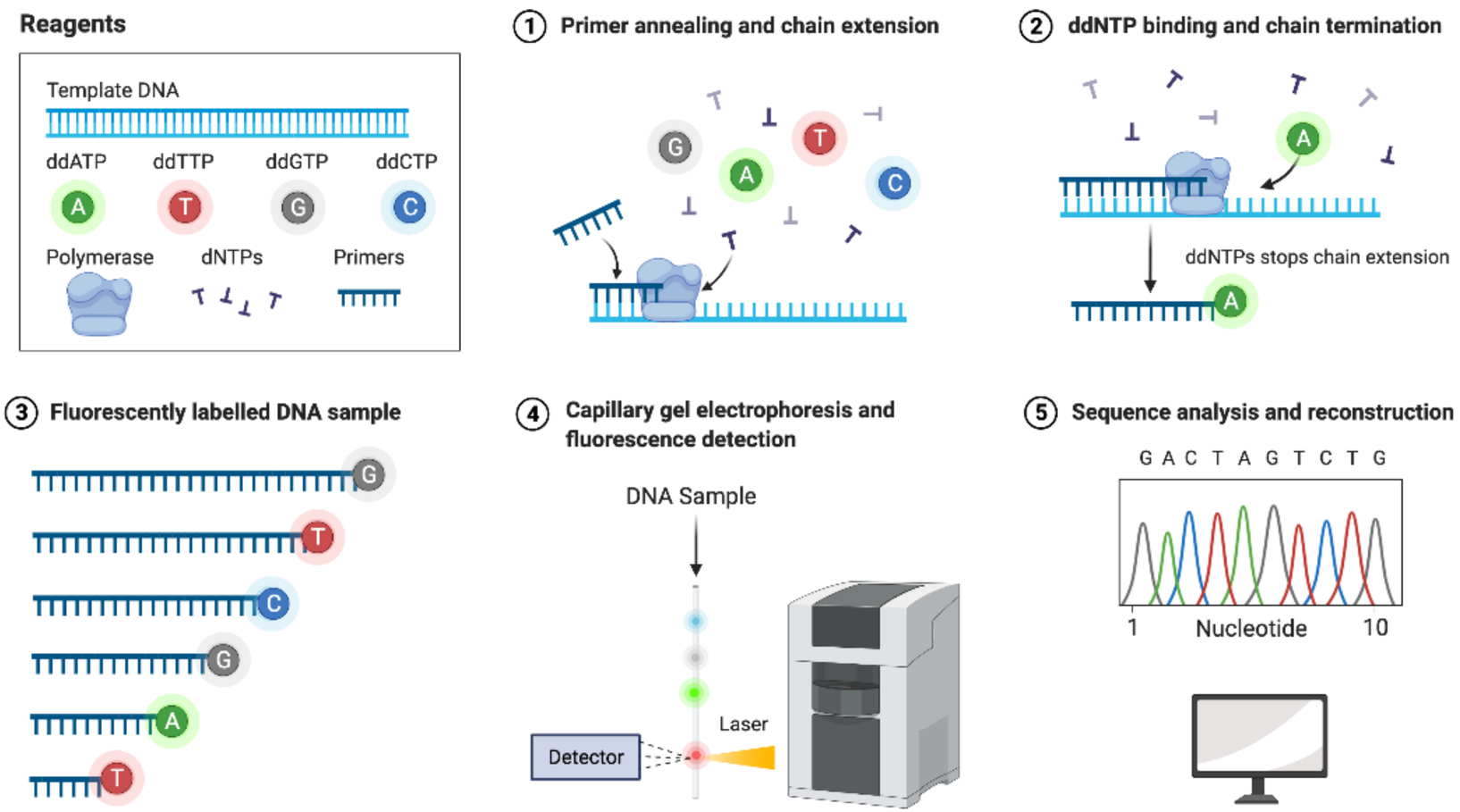


BioTechniques

psomagen

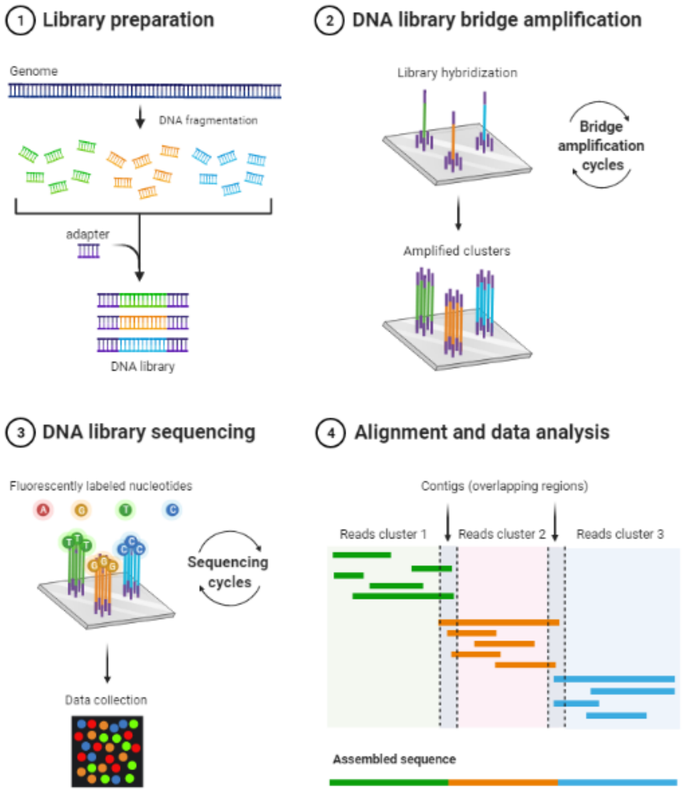
Genome Analysis

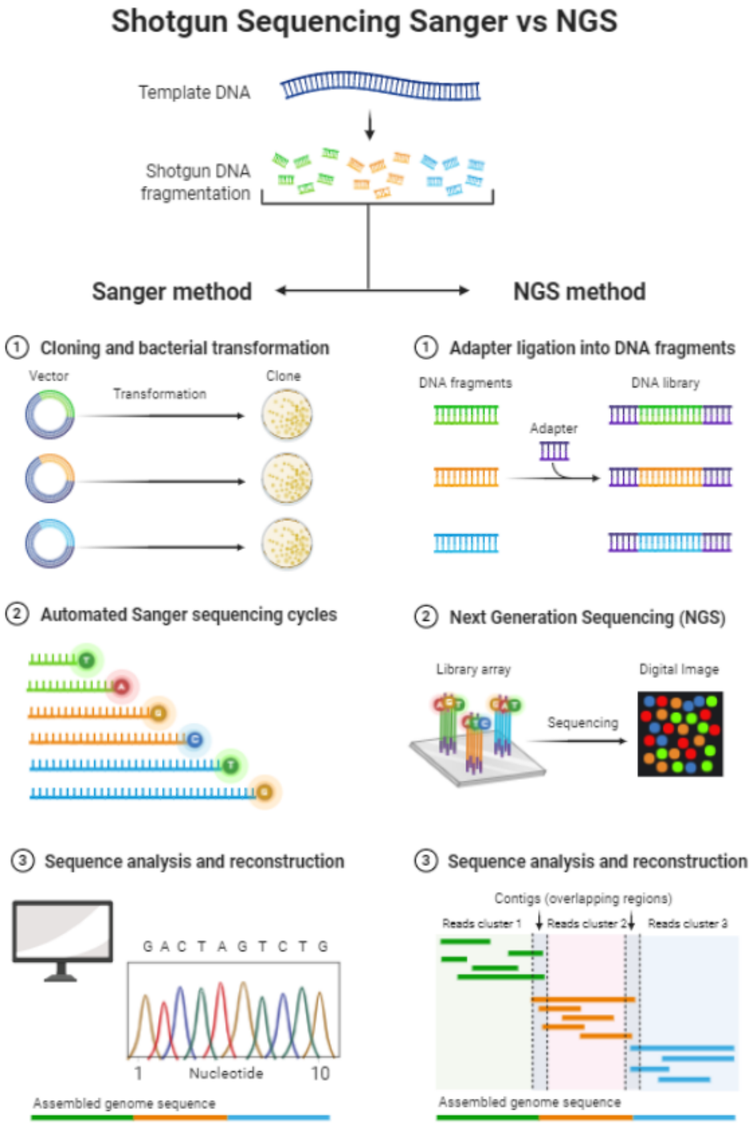
Sanger Sequencing



Genome Analysis

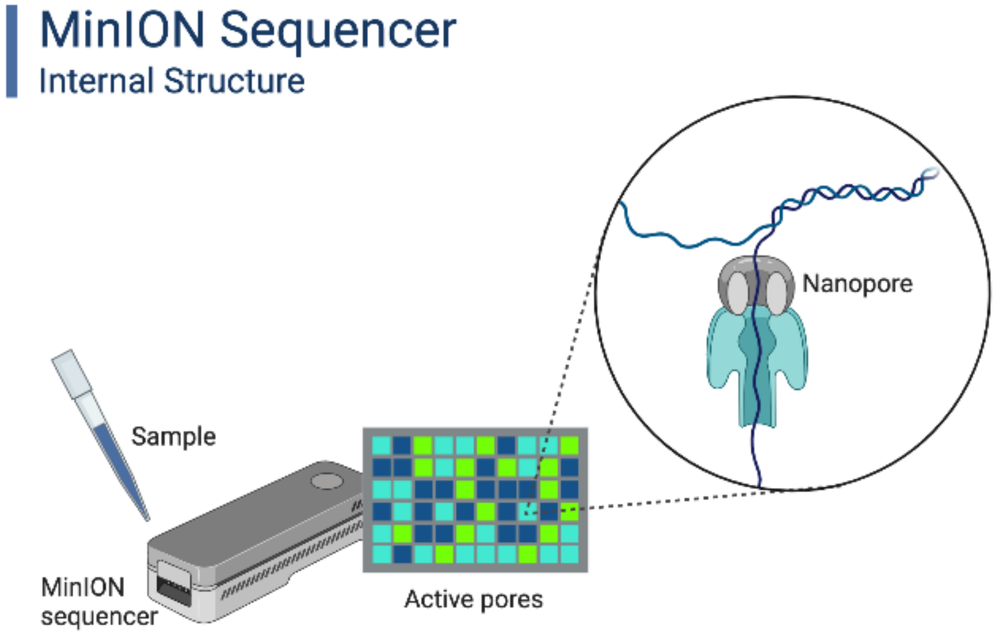
NGS vs. Sanger Sequencing

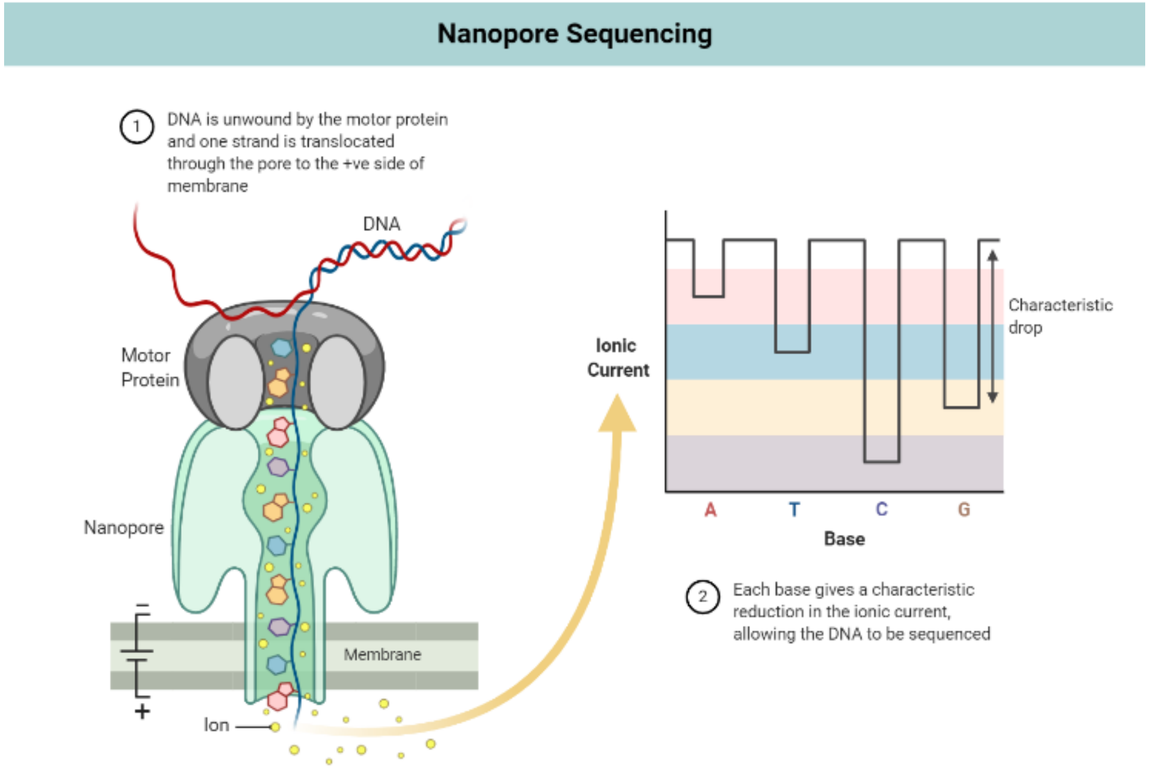




Genome Analysis

Nanopore Sequencing

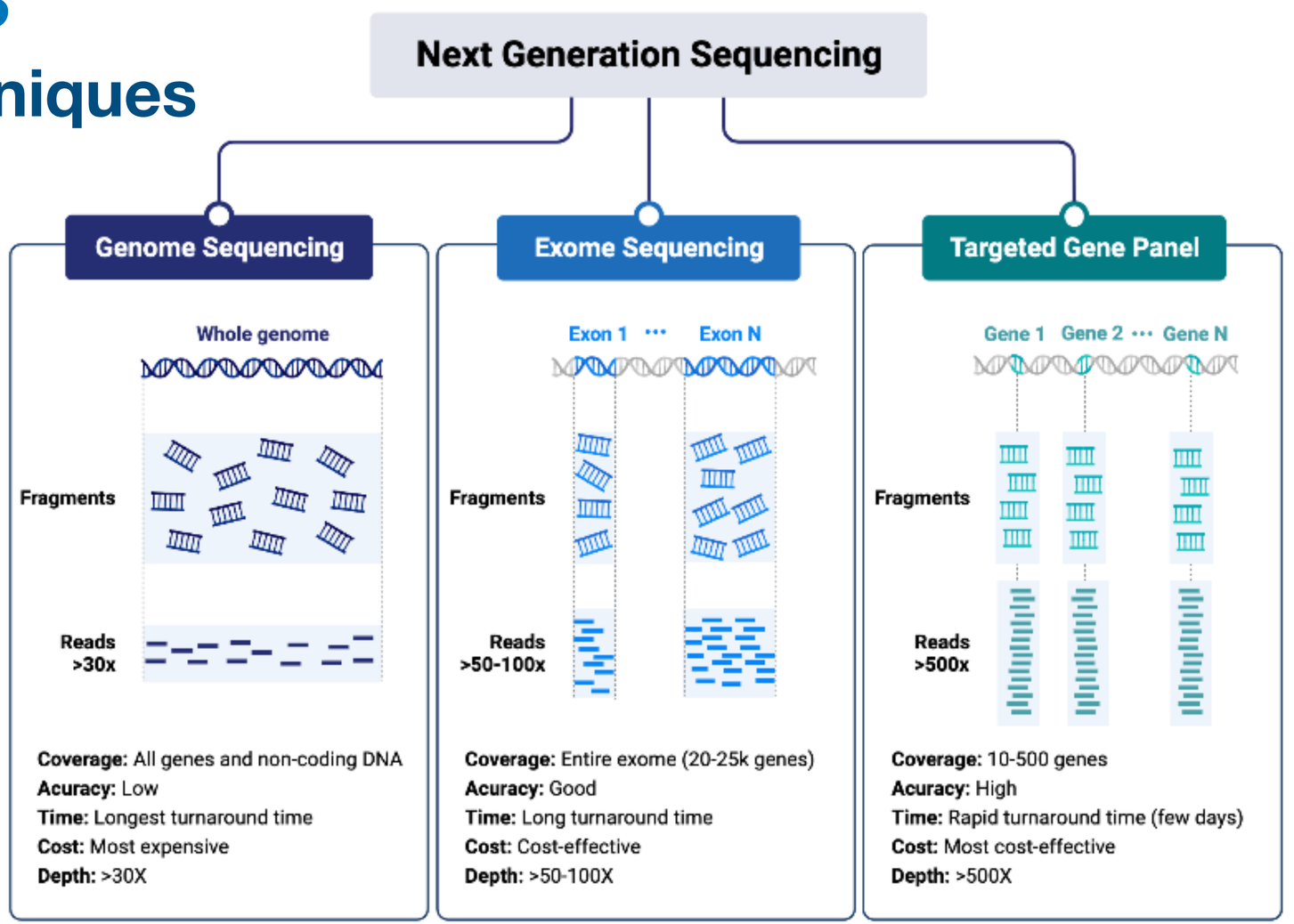




Genome Analysis

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

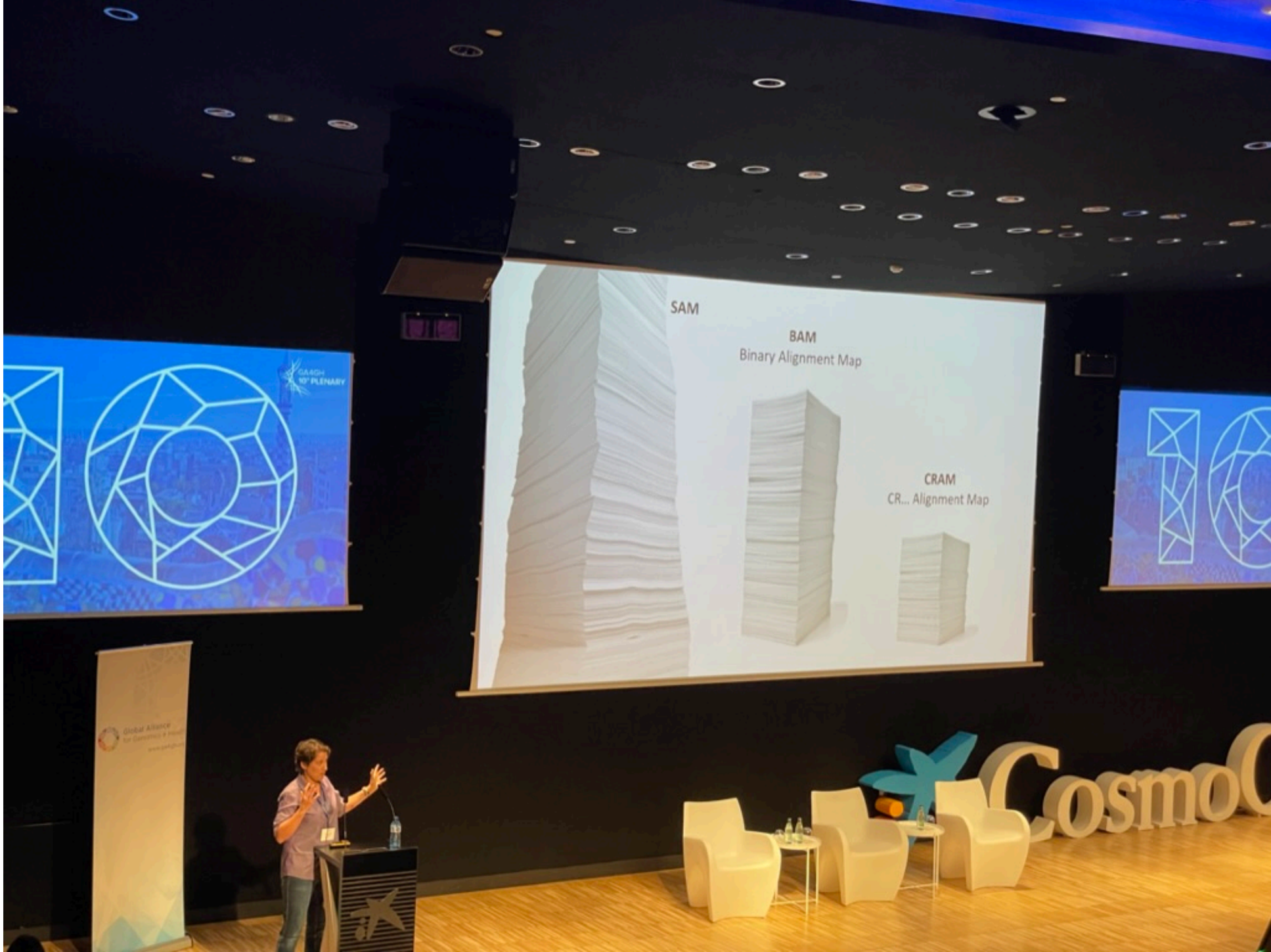
Types | Sizes | Use Cases

What is a PB, for human genomes?

It depends...

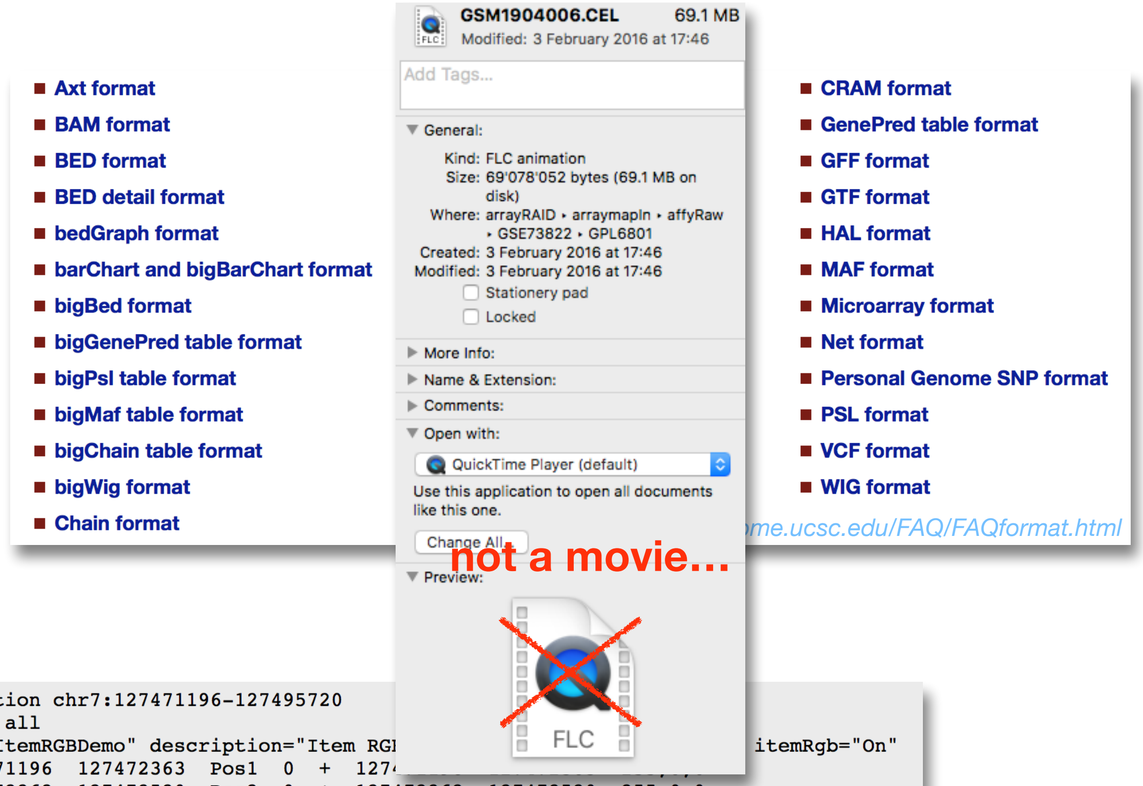
- 2 bits per base are sufficient to encode TCGA
  - using 00, 01, 10, 11
  - [TCGA]{3’000’000’000}
  - 2 * 3 * 10^9b = 6, 000, 000, 000 b
  - perfect genome (no overhead): ~715 MB
  - 1PB => ~1’400’000 genomes
- according to Swiss online store (Sep 2021) ~35'000CHF (65x16TB á CHF550)
- this is less than 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)




```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB                                    itemRgb="On"
chr7    127471196  127472363  Pos1  0  +  127                                    ,
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

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

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.

https://samtools.github.io/hts-specs/

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.

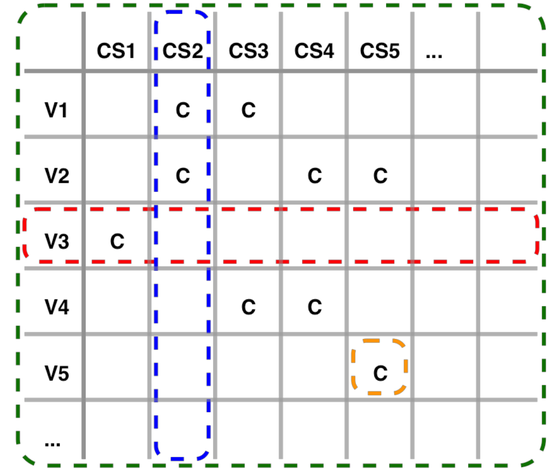
Transfer protocols

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">
```

Body

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0|1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1|0:77	1/1:95
1	100		T	<DEL>	.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

Mandatory header lines

Optional header lines (meta-data about the annotations in the VCF body)

Reference alleles (GT=0)

Alternate alleles (GT>0 is an index to the ALT column)

Phased data (G and C above are on the same chromosome)

Deletion

SNP

Large SV

Insertion

Other event

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

Global Alliance for Genomics & Health

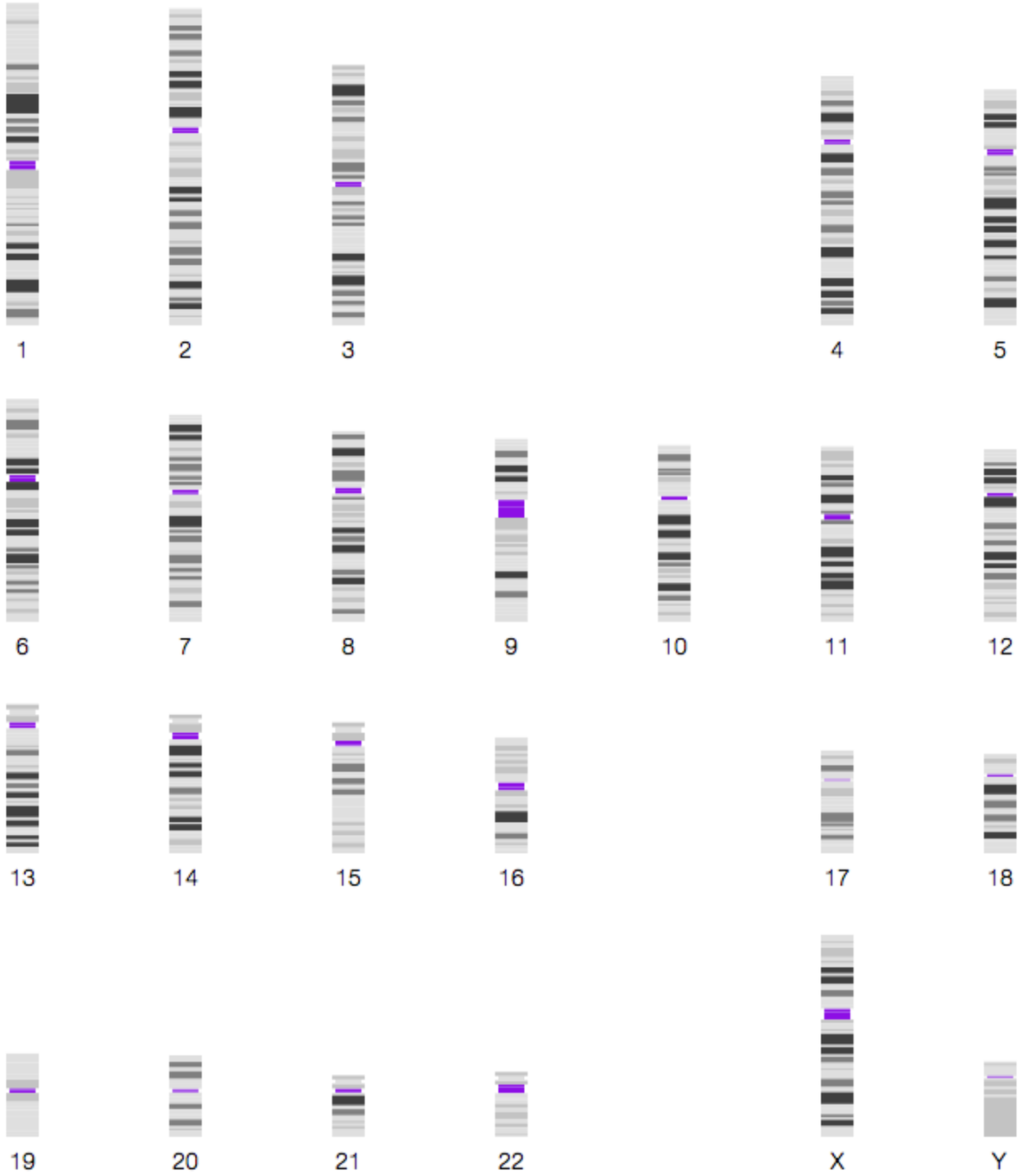
Task: Estimate Storage Requirements for 1000 Genomes

- 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
	3080419480

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
	3070128059	**3080419480**	**3095677412**	**3088781199**	**18653140**

Genome Liftover

Moving between genome editions

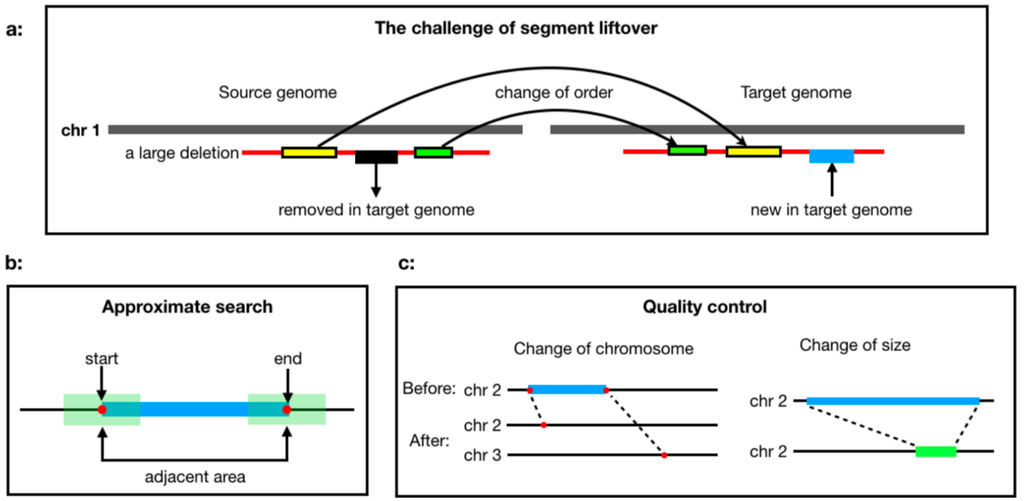
- different genome editions lead to shifting positions of defined elements such as genes
- local regions are frequently stable between editions
- shifts from change in regional lengths are defined in "chain files"
- chain files serve as guides for positional remapping using liftover methods
- Task: Read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

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]

[1]Institute of molecular Life Sciences, University of Zürich, Zürich, CH-8057, Switzerland
[2]Swiss Institute of Bioinformatics, University of Zürich, Zürich, CH-8057, Switzerland



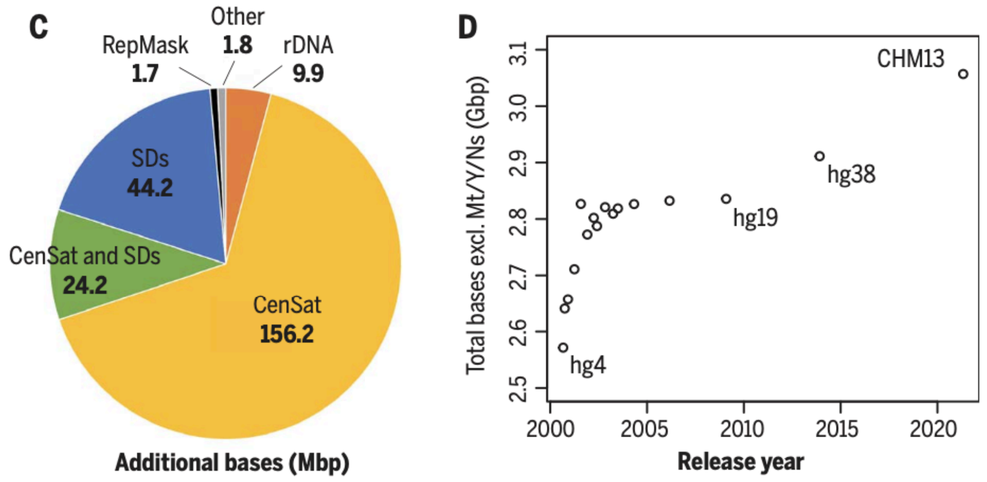
National Human Genome Research Institute

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

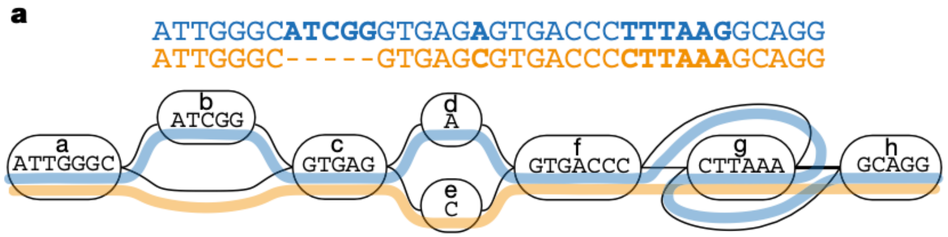
The complete sequence of a human genome

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

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[1,2,3,60], Mobin Asri[4,60], Jana Ebler[5,6,60], Daniel Doerr[5,6], Marina Haukness[4], Glenn Hickey[4], Shuangjia Lu[1,2], Julian K. Lucas[4], Jean Monlong[4], Haley J. Abel[7], Silvia Buonaiuto[8], Xian H. Chang[4], Haoyu Cheng[9,10], Justin Chu[9], Vincenza Colonna[8,11], Jordan M. Eizenga[4], Xiaowen Feng[9,10], Christian Fischer[11], Robert S. Fulton[12,13], Shilpa Garg[14], Cristian Groza[15], Andrea Guarracino[11,16], William T. Harvey[17], Simon Heumos[18,19], Kerstin Howe[20], Miten Jain[21], Tsung-Yu Lu[22], Charles Markello[4], Fergal J. Martin[23], Matthew W. Mitchell[24], Katherine M. Munson[17], Moses Njagi Mwaniki[25], Adam M. Novak[4], Hugh E. Olsen[4], Trevor Pesout[4], David Porubsky[17], Pjotr Prins[11], Jonas A. Sibbesen[26], Jouni Sirén[4], Chad Tomlinson[12], Flavia Villani[11], Mitchell R. Vollger[17,27], Lucinda L. Antonacci-Fulton[12], Gunjan Baid[28], Carl A. Baker[17], Anastasiya Belyaeva[28], Konstantinos Billis[23], Andrew Carroll[28], Pi-Chuan Chang[28], Sarah Cody[12], Daniel E. Cook[28], Robert M. Cook-Deegan[29], Omar E. Cornejo[30], Mark Diekhans[4], Peter Ebert[5,6,31], Susan Fairley[23], Olivier Fedrigo[32], Adam L. Felsenfeld[33], Giulio Formenti[32], Adam Frankish[23], Yan Gao[34], Nanibaa' A. Garrison[35,36,37], Carlos Garcia Giron[23], Richard E. Green[38,39], Leanne Haggerty[23], Kendra Hoekzema[17], Thibaut Hourlier[23], Hanlee P. Ji[40], Eimear E. Kenny[41], Barbara A. Koenig[42], Alexey Kolesnikov[28], Jan O. Korbel[23,43], Jennifer Kordosky[17], Sergey Koren[44], HoJoon Lee[40], Alexandra P. Lewis[17], Hugo Magalhães[5,6], Santiago Marco-Sola[45,46], Pierre Marijon[5,6], Ann McCartney[44], Jennifer McDaniel[47], Jacquelyn Mountcastle[32], Maria Nattestad[28], Sergey Nurk[44], Nathan D. Olson[47], Alice B. Popejoy[48], Daniela Puiu[49], Mikko Rautiainen[44], Allison A. Regier[12], Arang Rhie[44], Samuel Sacco[30], Ashley D. Sanders[50], Valerie A. Schneider[51], Baergen I. Schultz[33], Kishwar Shafin[28], Michael W. Smith[33], Heidi J. Sofia[33], Ahmad N. Abou Tayoun[52,53], Françoise Thibaud-Nissen[51], Francesca Floriana Tricomi[23], Justin Wagner[47], Brian Walenz[44], Jonathan M. D. Wood[20], Aleksey V. Zimin[49,54], Guillaume Bourque[55,56,57], Mark J. P. Chaisson[22], Paul Flicek[23], Adam M. Phillippy[44], Justin M. Zook[47], Evan E. Eichler[17,58], David Haussler[4,58], Ting Wang[12,13], Erich D. Jarvis[32,58,59], Karen H. Miga[4], Erik Garrison[11]✉, Tobias Marschall[5,6]✉, Ira M. Hall[1,2]✉, Heng Li[9,10]✉ & Benedict Paten[4]✉

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[1]. 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 enviroment introduction	Michael: Genoomic Resources & Data Sharing		Feifei: Sequence analysis practical. FastQC, trimmomatic, BWA-MEM2, SAMtools, GangSTR, BCFtools	Michael: Genomic Data & Privacy		Feifei & Ziying: survival	Feifei: population structure		Exam
10:00 - 11:00	Tina Siegenthaler: technical introduction (room, computer, accounts)		Jiahui: Terminal / Unix / Files	Hangjia: R exercise	Michael: Genoomic 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: Genoomic 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 (cyvcf2), 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			Hangija: 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