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/courses/UZH-BIO392/

	Tue Sep 19	Wed Sep 20	Thu Sep 21	Fri Sep 22	Tue Sep 26	Wed Sep 27	Thu Sep 28	Fri Sep 29	Tue Oct 3	Wed Oct 4	Thu Oct 5	Fri Oct 6	Tue Oct 10	Wed Oct 11
09:00 - 10:00		Github exercise: create user specific directories & upload/edit test files using Markdown (Ziying)	Izaskun: Terminal / Unix / Files	Izaskun: File formats for human genetic variation / file handling		Michael lecture introduction to some resources, CNVs, Progenetix	Hangjia: Clinvar and Clingen	Max & Feifei: Sequence analysis practical. FastQC, trimmomatic, BWA-MEM2, SAMtools, GangSTR, BCFtools		Max & Feifei:: analysis & interpretation. Parsing VCF (cyvcf2), UCSD genome browser, ENSEMBL variant effect predictor	Rahel: survival analysis	Michael: Genomic data risks & opportunities		Exam
10:00 - 11:00		Ziying: github desktop and terminal	Izaskun: Terminal / Unix / Files	Izaskun: File formats for human genetic variation / file handling			Hangjia: blast	Max & Feifei:: Sequence analysis practical		Max & Feifei:: analysis & interpretation.	Rahel: survival analysis	Michael: Genomic data risks & opportunities		Exam
11:00 - 12:00		Ziying: python warmup and exercise		Izaskun: File formats for human genetic variation / file handling		Hangjia: Progenetix as tool for CNV frequencies etc.	Hangija: Blast exercise	Max & Feifei:: Sequence analysis practical		Max & Feifei:: analysis & interpretation.	Rahel: survival analysis	Michael: Genomic data risks & opportunities		Exam
13:00 - 14:00	* Room information * Administrative - discuss times/days - exam	Hangjia: R enviroment introduction	Izaskun: SIB online introduction to Unix	Izaskun: short project (1000 genomes), reading, literature	Recap W1; Q&A	Task: Browse/explore genome resources and provide some notes (1-2 pages total) in a doc posted on Github (.md)	Max: Sequence analysis introduction. Overview of pipeline from raw reads -> variant calling & interpretation	Max & Feifei:: STR reading up	Michael lecture analysis principles (why surv, etc.)	Rahel: survival analysis	Rahel & Michael: presentation/discussion	graded exercise: genomic data risks ...	Exam revision, Q&A	
14:00 - 15:00	Tina Siegenthaler: technical introduction (room, computer, accounts)	Hangjia: R exercise	Izaskun: SIB online introduction to Unix	Izaskun: short project (1000 genomes), reading, literature	Literature (genome analysis techniques ...)		Max: Sequence analysis introduction	Max & Feifei:: STR reading up		Rahel: survival analysis	Rahel & Michael: presentation/discussion			
15:00 - 16:30	* explore course site * create Github accounts and forward to bio392@compbiozurich.org * Ziying&Hangjia: overall schedule of the course	Ziying: paper reading, Q & A		Izaskun: short project (1000 genomes), reading, literature	Genome technologies - brief notes about usage scenarios, pro & con					Rahel: survival analysis				

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

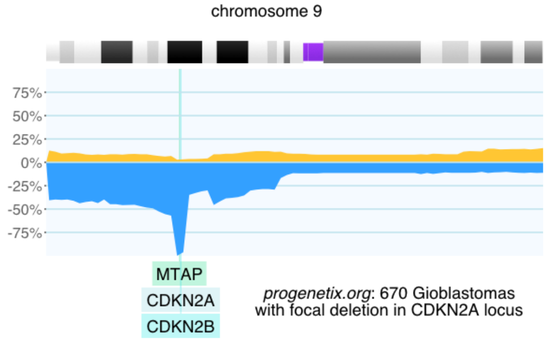
Theoretical Cytogenetics & Oncogenomics

- 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

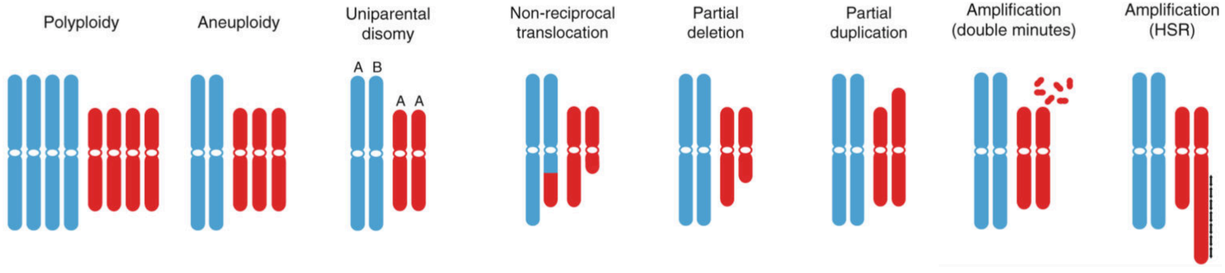
info.baudisgroup.org

Theoretical Cytogenetics and Oncogenomics Research | Methods | Standards

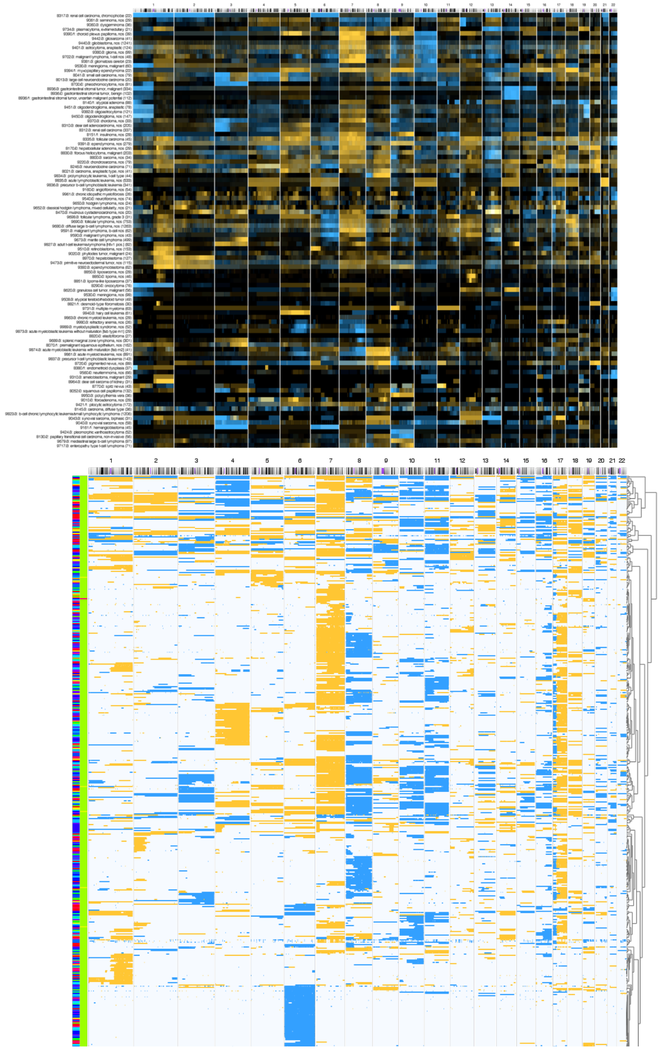


Genomic Imbalances in Cancer - Copy Number Variations (CNV)

- Point mutations (insertions, deletions, substitutions)
- Chromosomal rearrangements
- **Regional Copy Number Alterations** (losses, gains)
- Epigenetic changes (e.g. DNA methylation abnormalities)



Grade et al., 2015 Recent Results Cancer Res



Genome screening at the core of “Personalised Health”

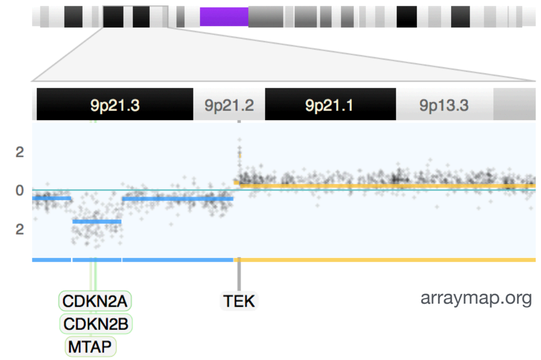
- **Genome analyses** (including transcriptome, metagenomics) are core technologies for Personalised Health™ applications
- The unexpectedly large amount of **sequence variants** in human genomes - germline and somatic/cancer - requires huge analysis efforts and creation of **reference repositories**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

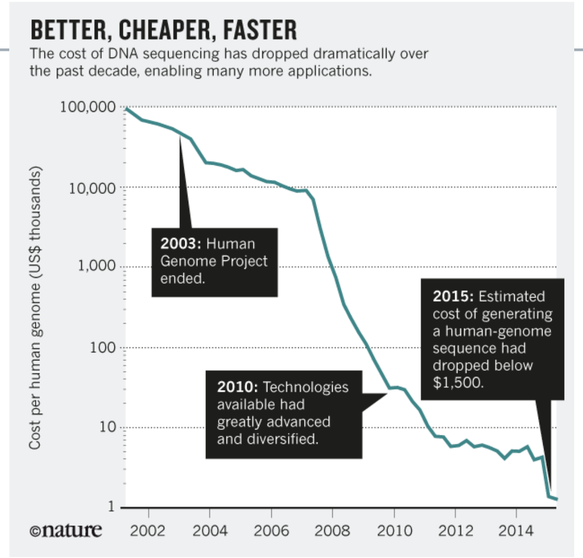
arrayMap

progenetix

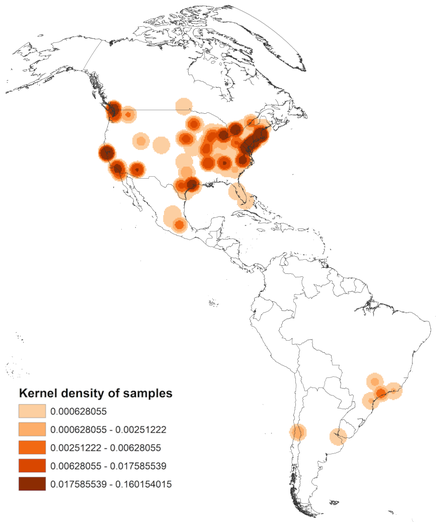
elixir

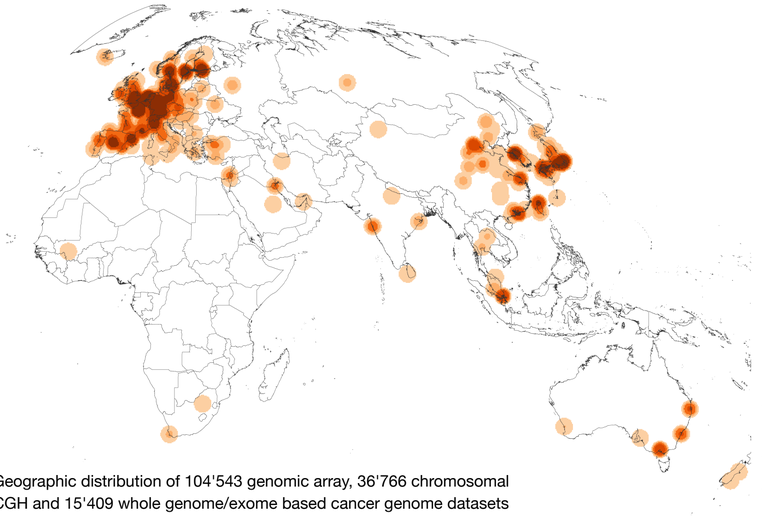
Global Alliance for Genomics & Health





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

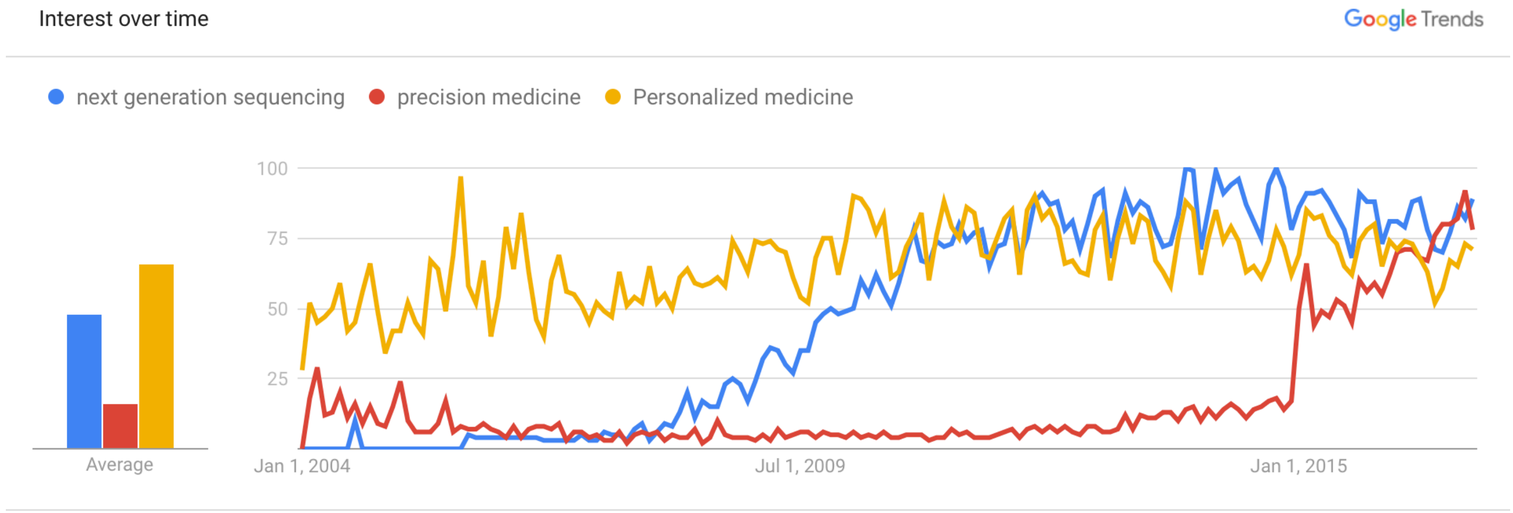




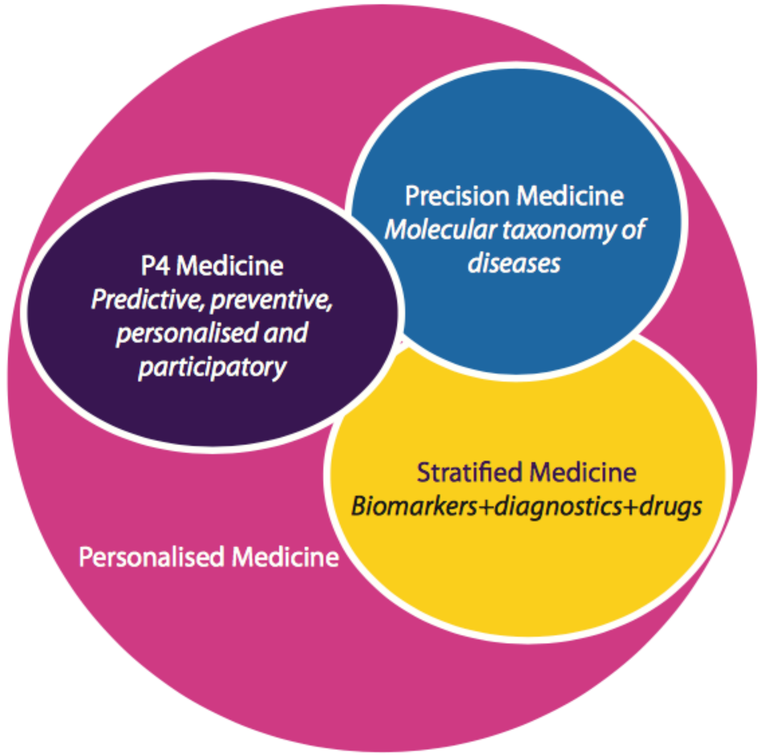
Geographic distribution of 104'543 genomic array, 36'766 chromosomal CGH and 15'409 whole genome/exome based cancer genome datasets

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.

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

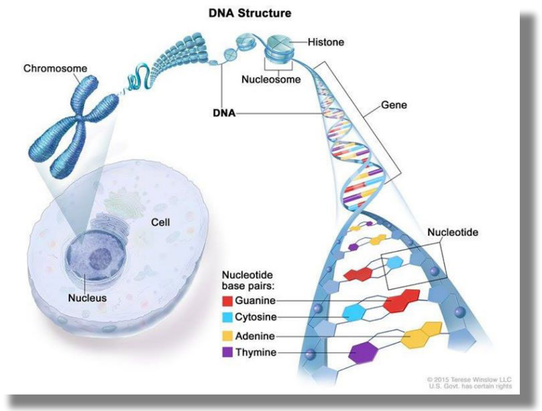
The landscape of somatic copy-number alteration across human cancers

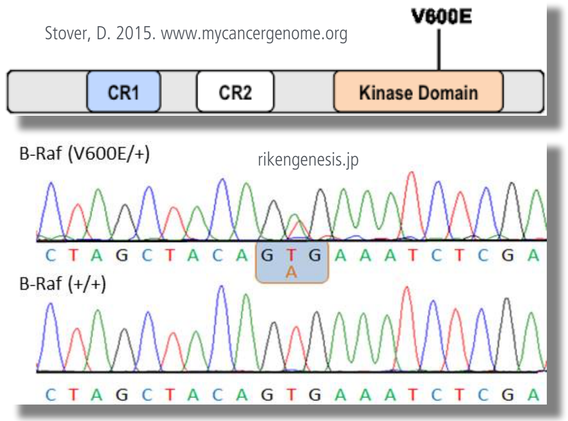
Rameen Beroukhim[1,3,4,5]*, Craig H. Mermel[1,3]*, Dale Porter[8], Guo Wei[1], Soumya Raychaudhuri[1,4], Jerry Donovan[8], Jordi Barretina[1,3], Jesse S. Boehm[1], Jennifer Dobson[1,3], Mitsuyoshi Urashima[9], Kevin T. Mc Henry[8], Reid M. Pinchback[1], Azra H. Ligon[4], Yoon-Jae Cho[6], Leila Haery[1,3], 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]

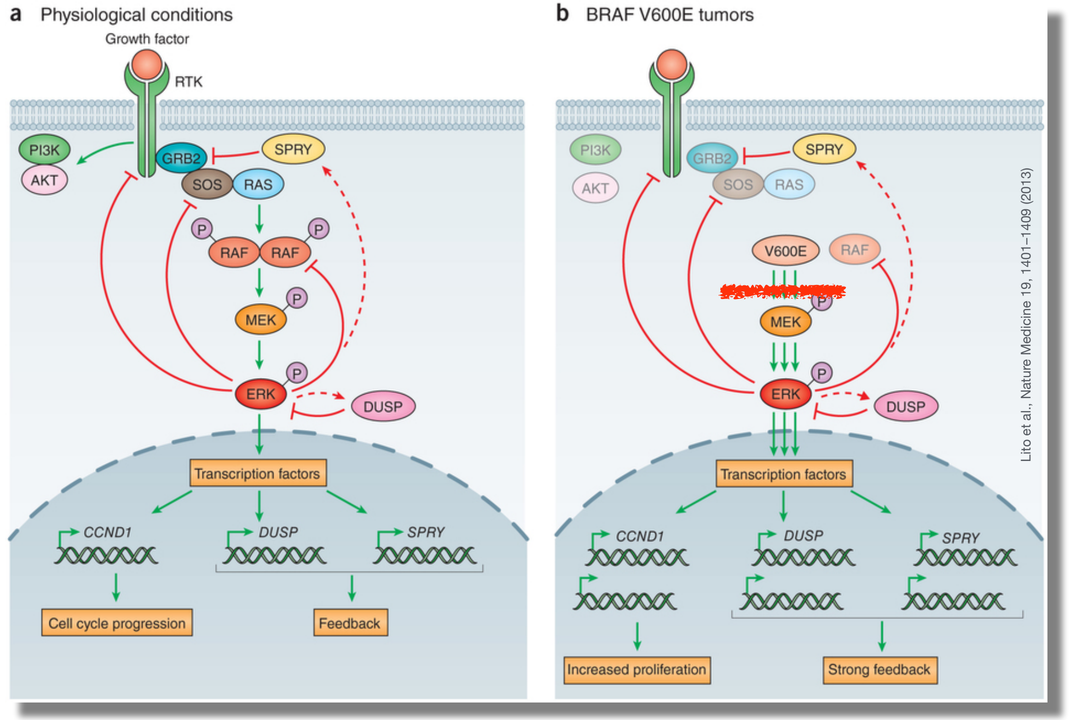
BRAF V600E (c.1799T>A) Mutation

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***)



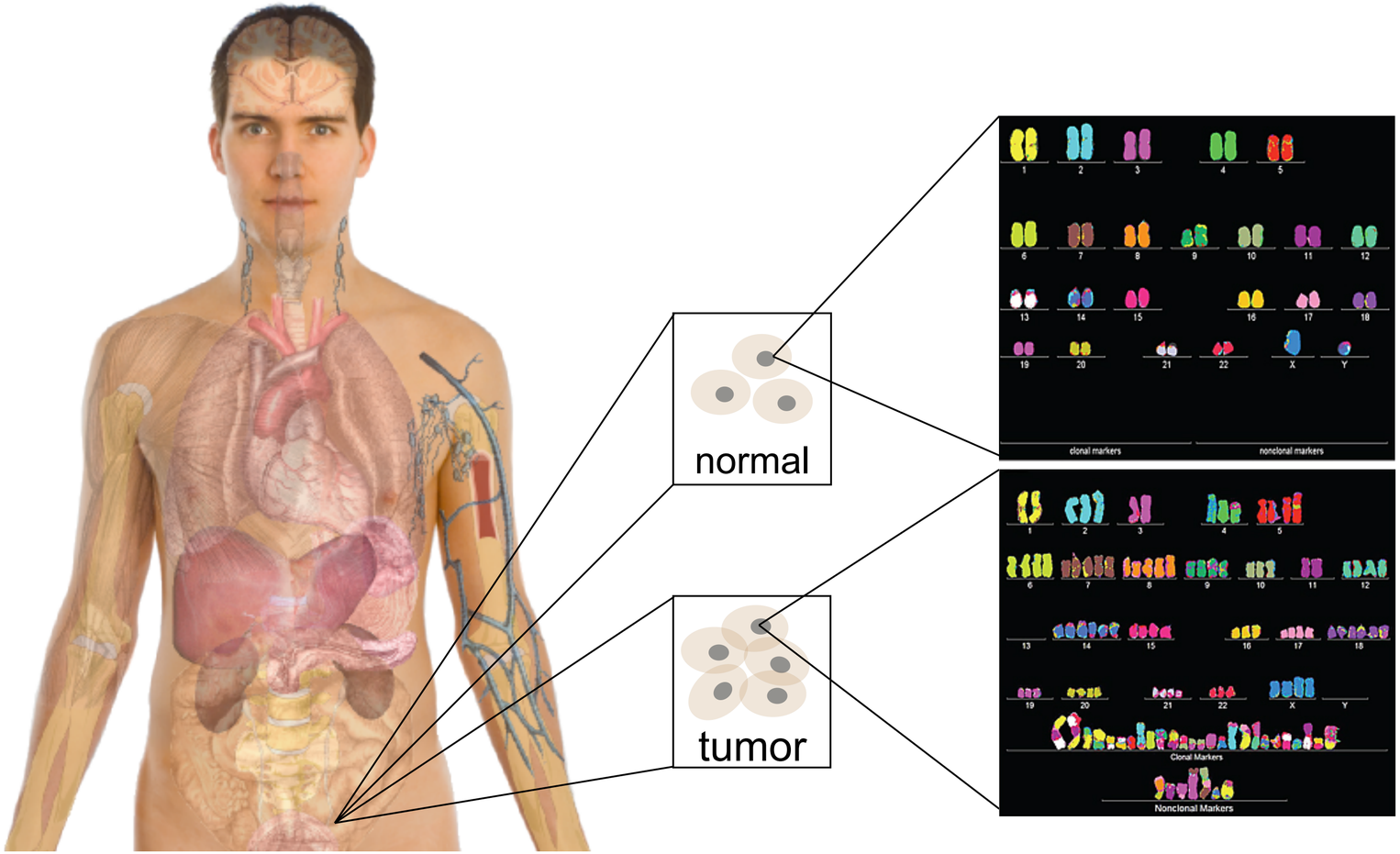




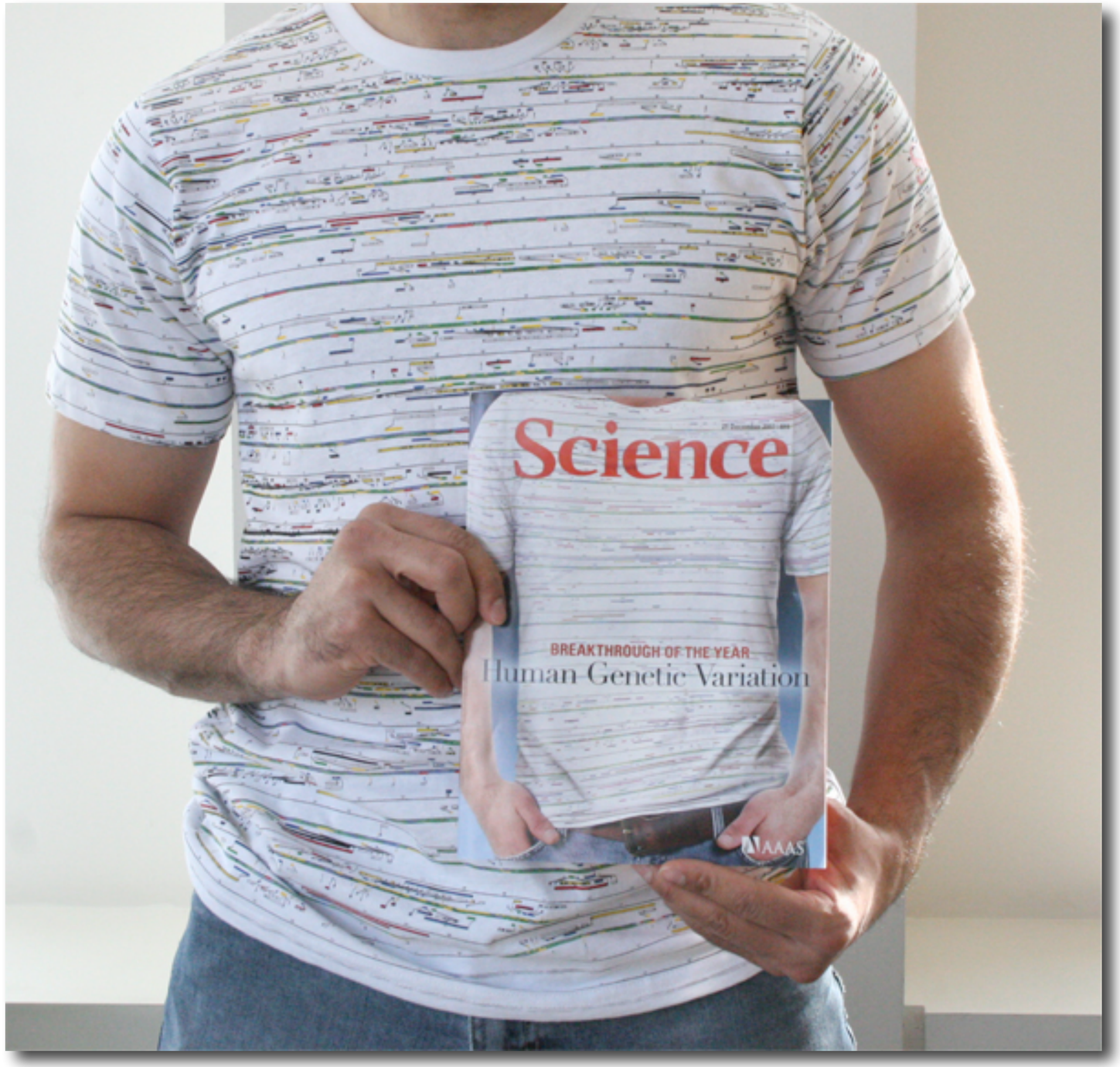
The BRAF V600E mutation leads to continuos phosphorylation of MEK, without the need for receptor based activation of the upstream pathway and loss of inhibitory feedback control.

Personal Genomics as a Gateway into Biology

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



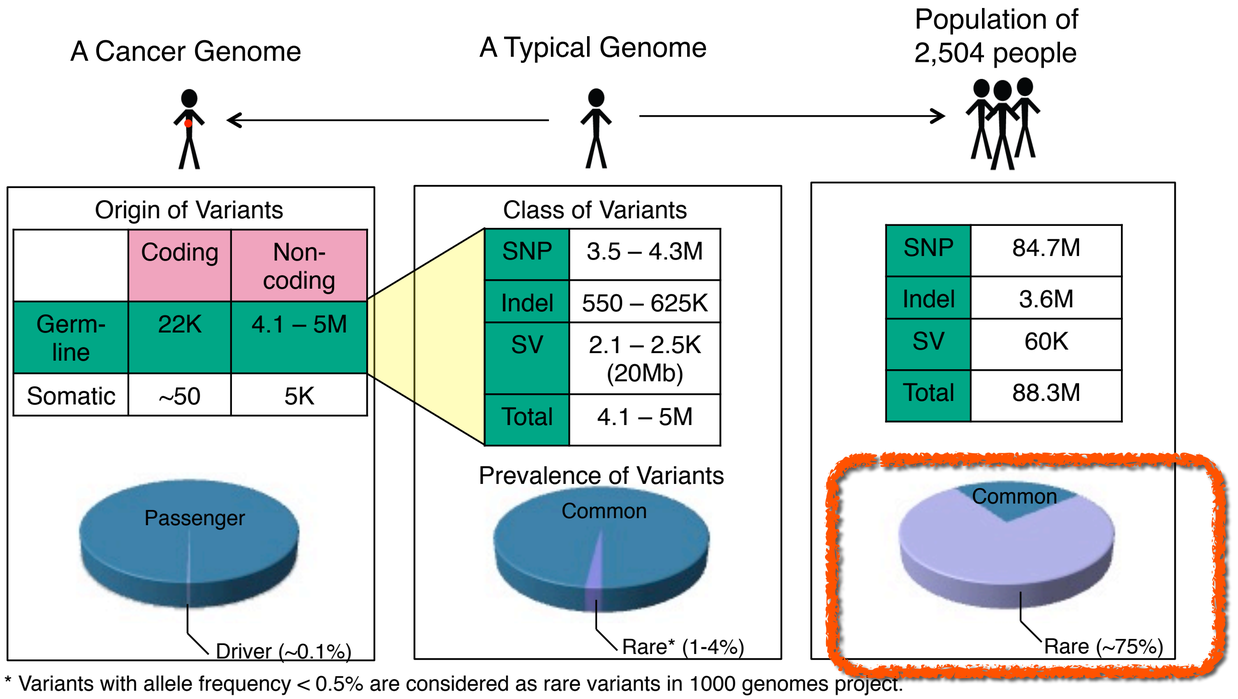
The trouble with human genome variation



Finding Somatic Mutations In Cancer

Many Needles in a Large Haystack

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



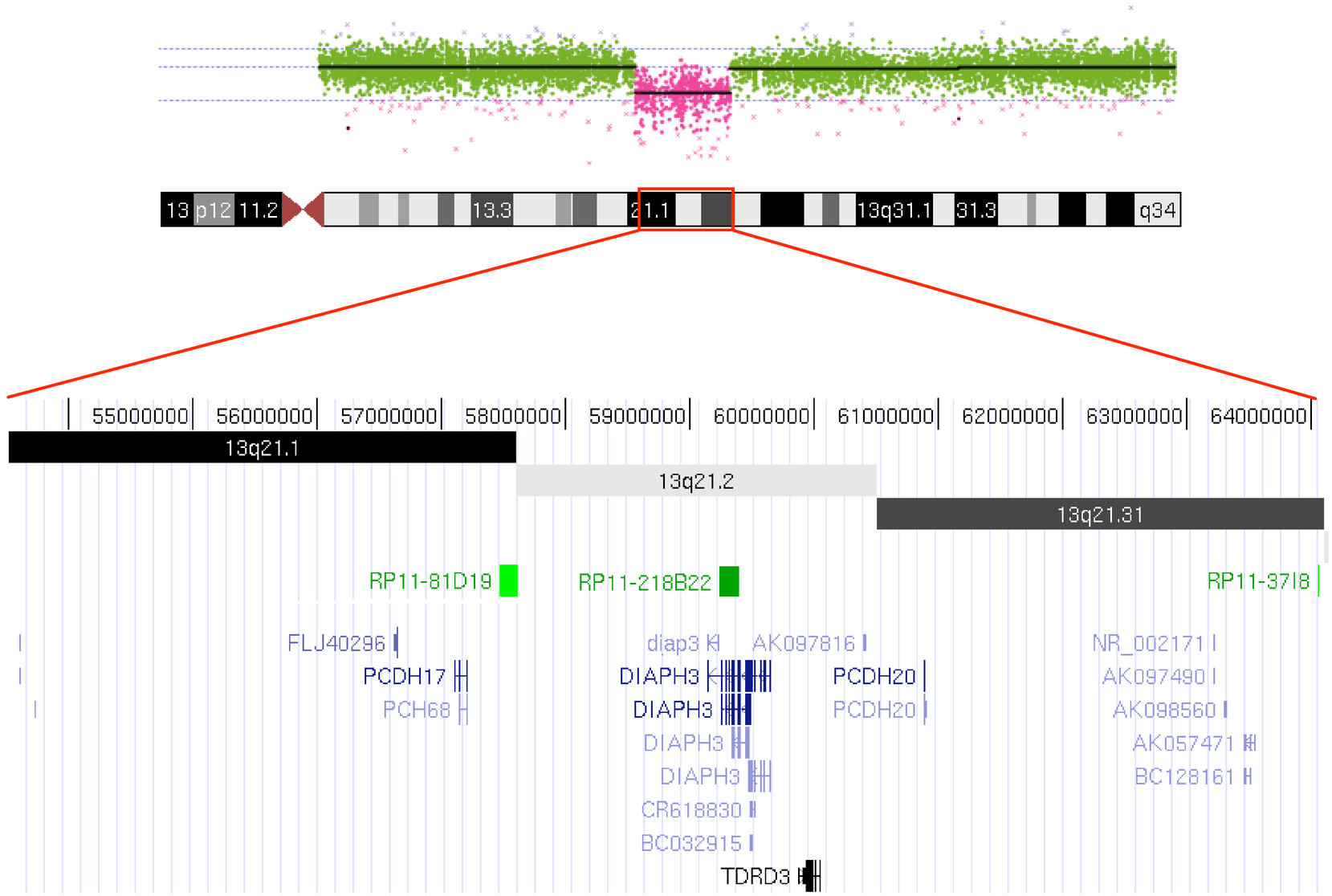
* Variants with allele frequency < 0.5% are considered as rare variants in 1000 genomes project.

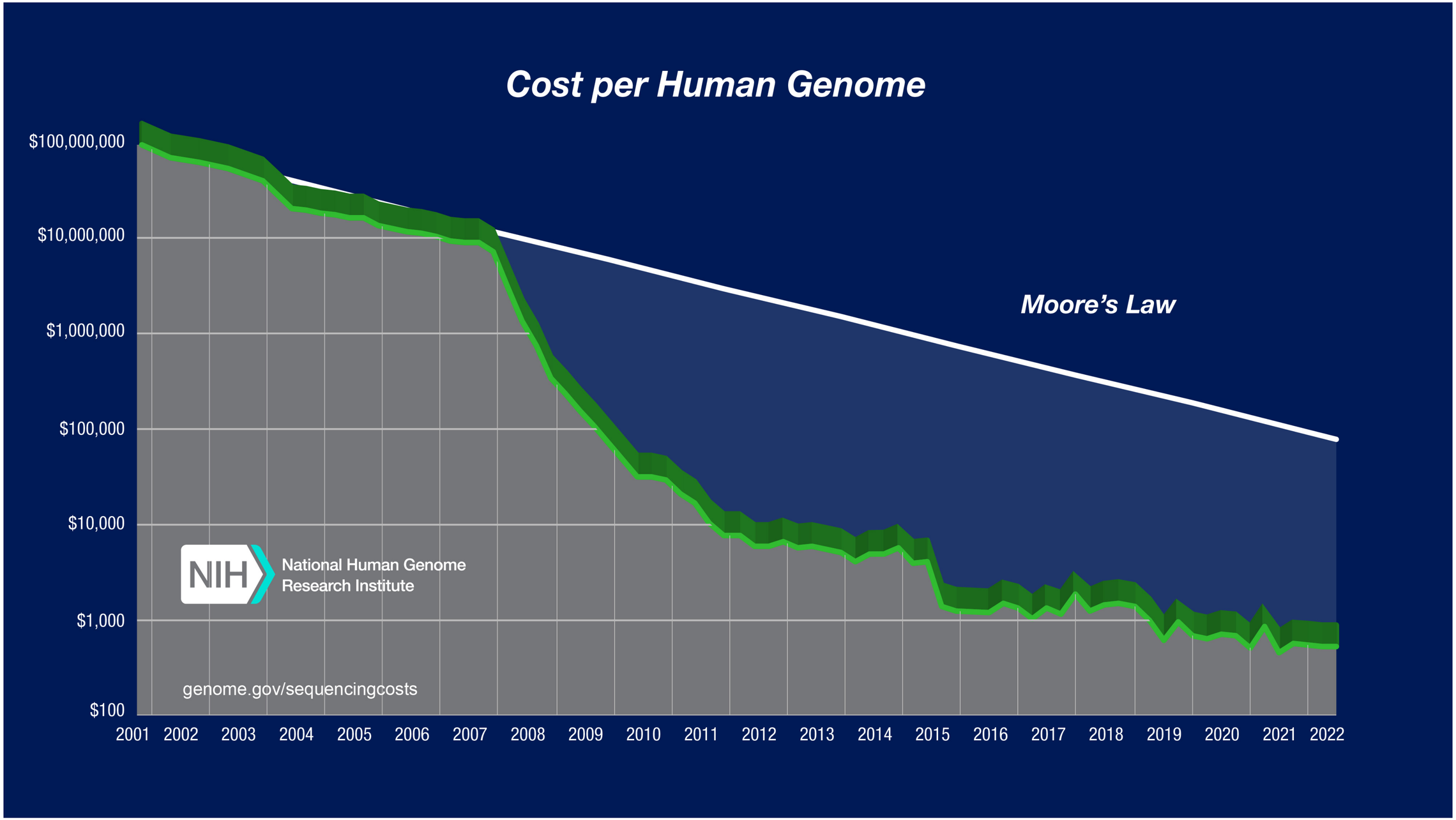
The 1000 Genomes Project Consortium, Nature. 2015. 526:68-74
Khurana E. et al. Nat. Rev. Genet. 2016. 17:93-108

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

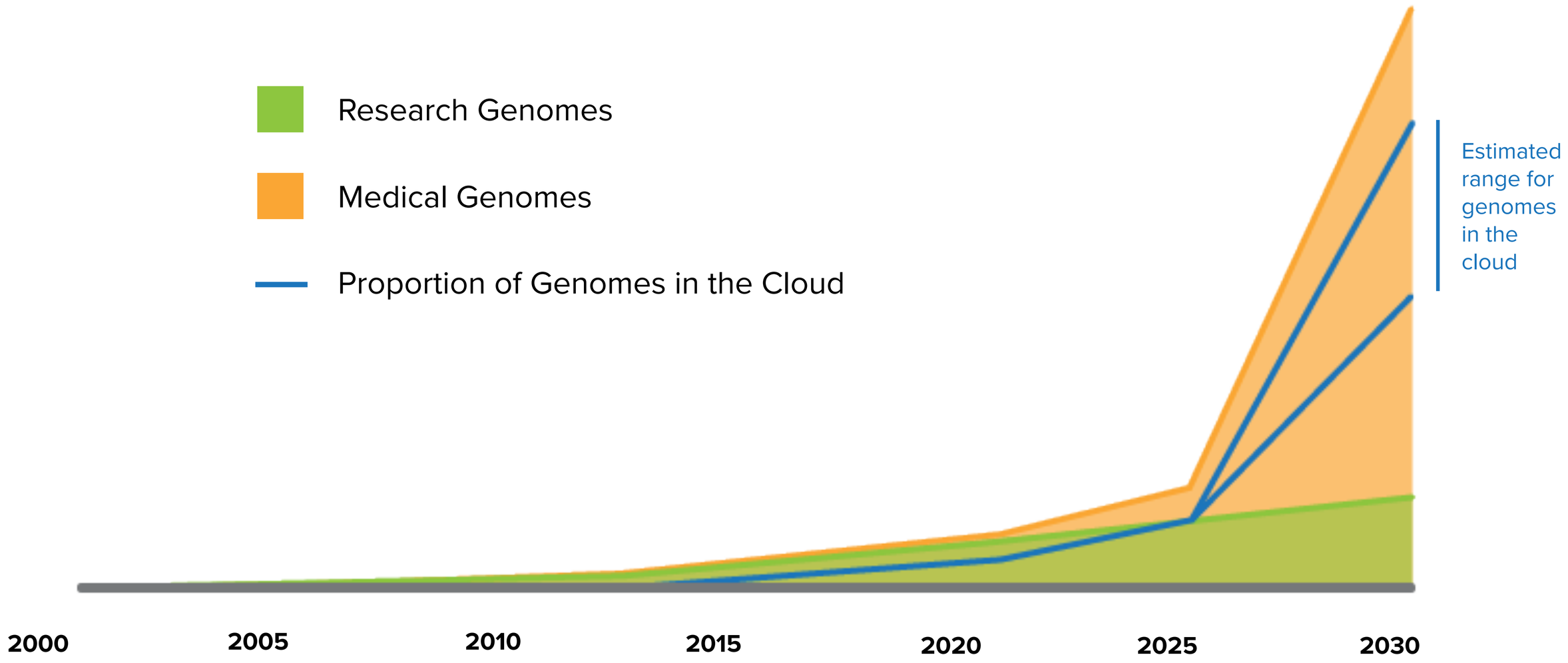
Clinical/Genomic Medicine

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

Total worldwide sequencing capability?!

Deployed sequencers of major platforms

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

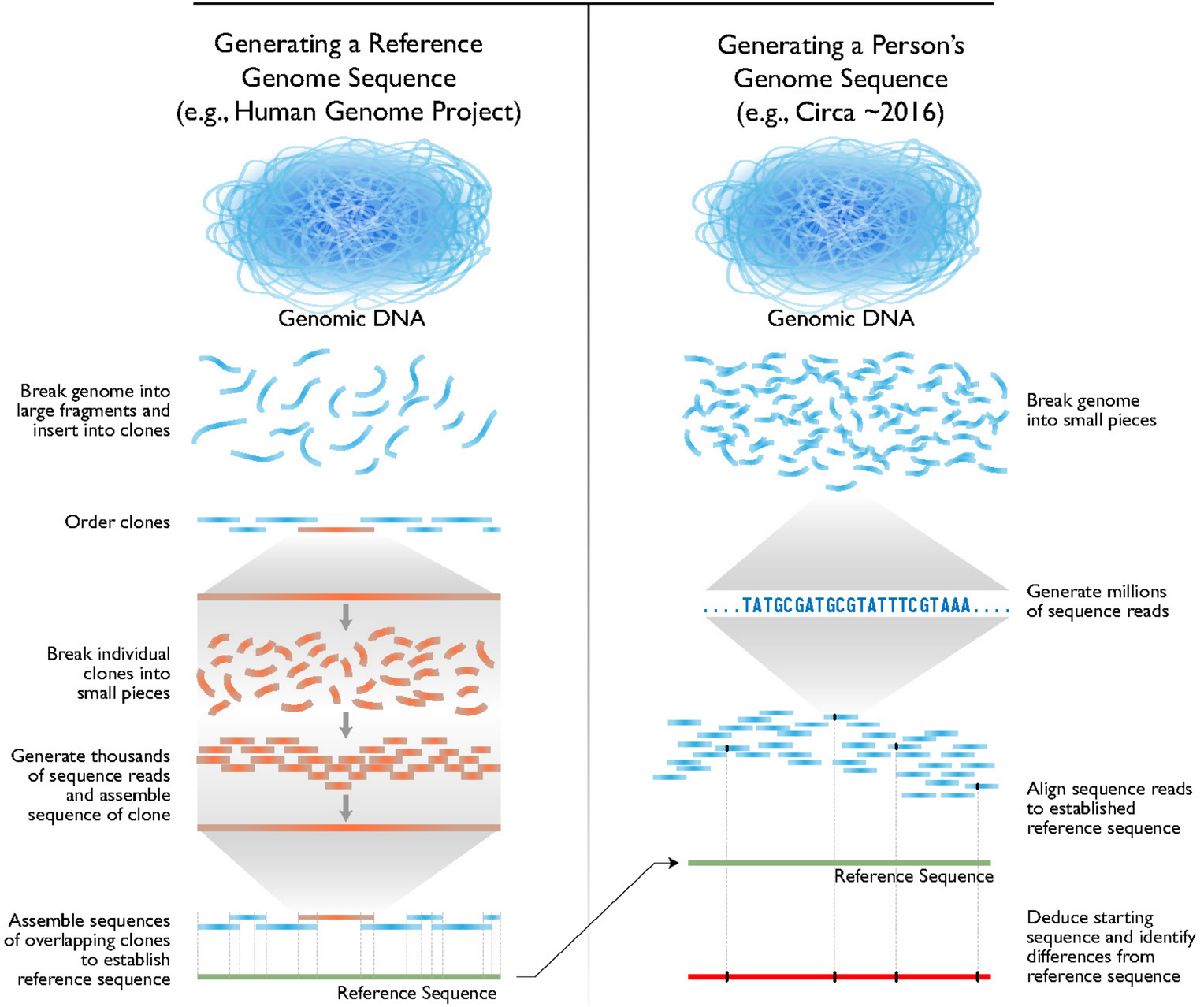
Platform	Estimated Instruments	Runs/Week	Total Runs/Year	Run Yield (Tb)	Total Sequencing Capacity Tb/year	
ONT MinION	5501	3	858156	0.05	42907.8	
ONT GridION	782	3	121992	0.25	30498	
ONT PromethION 48	67	2	6968	14	97552	
Ilm Novaseq 6000	1485	3	231660	6	1389960	
Ilm NextSeq	5430	3	847080	0.36	304948.8	
Ilm Miseq/Mini/iSeq	12340	3	1925040	0.015	28875.6	
Ion Torrent	2220	14	1616160	0.05	80808	
PacBio	577	5	150020	0.03	4500.6	
MGI - Mid/Low	2000	3	312000	0.72	224640	
MGI -T7	10	5	2600	6	15600	
Total	30412		6071676		2220290.8	Tb/year
			2331523584	Samples	70.4	Gb/s
			73.9	Samples/s		

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

Source: 41J Blog - https://41j.com/blog/2022/09/global-sequencing-capability-notes/

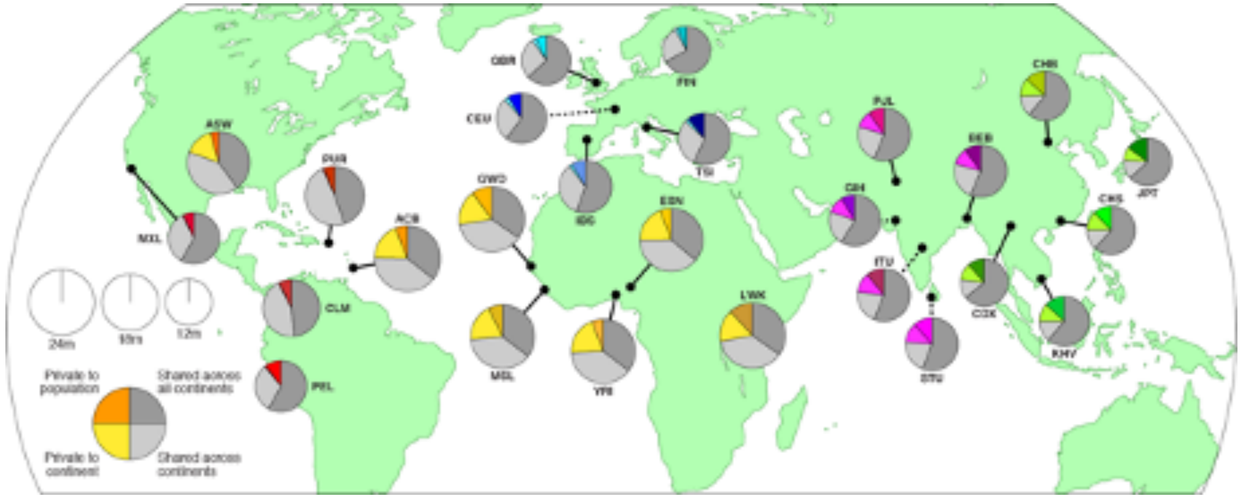
https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Human Genome Sequencing

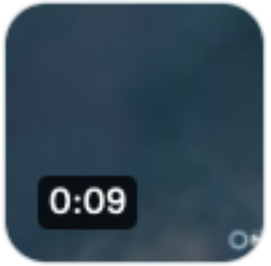


New technologies and new applications

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

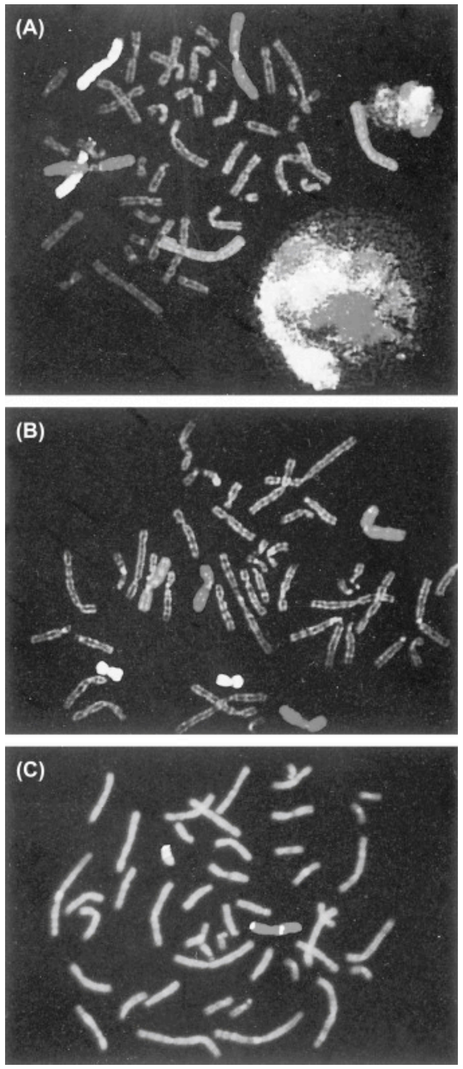
(Molecular) Cytogenetics

- 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**

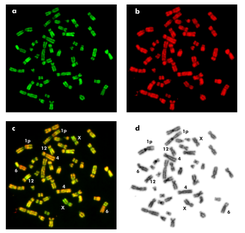
G-banded metaphase with three marker chromosomes (whose origin from chromosomes 22, 11, and 14 cannot be identified by this technique).

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).

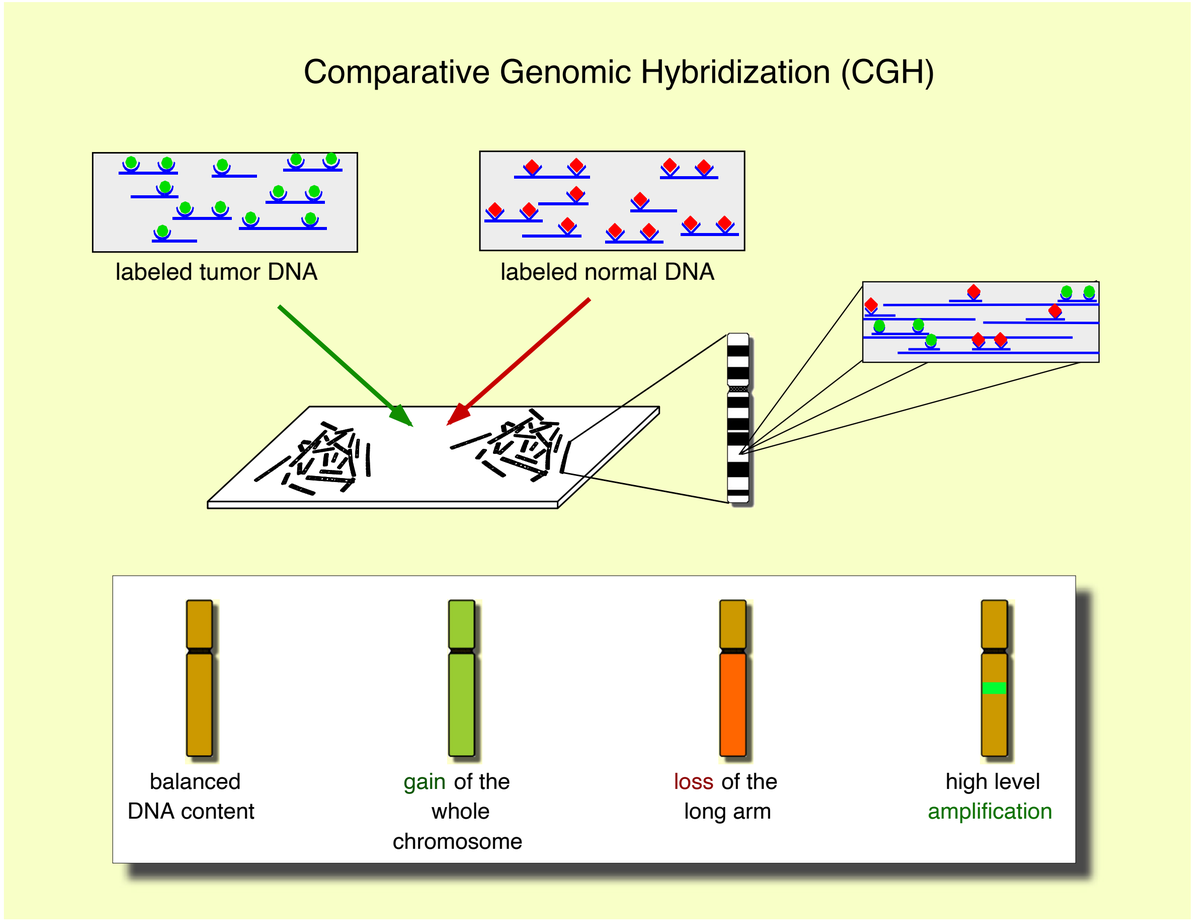
Comparative Genomic Hybridization



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.



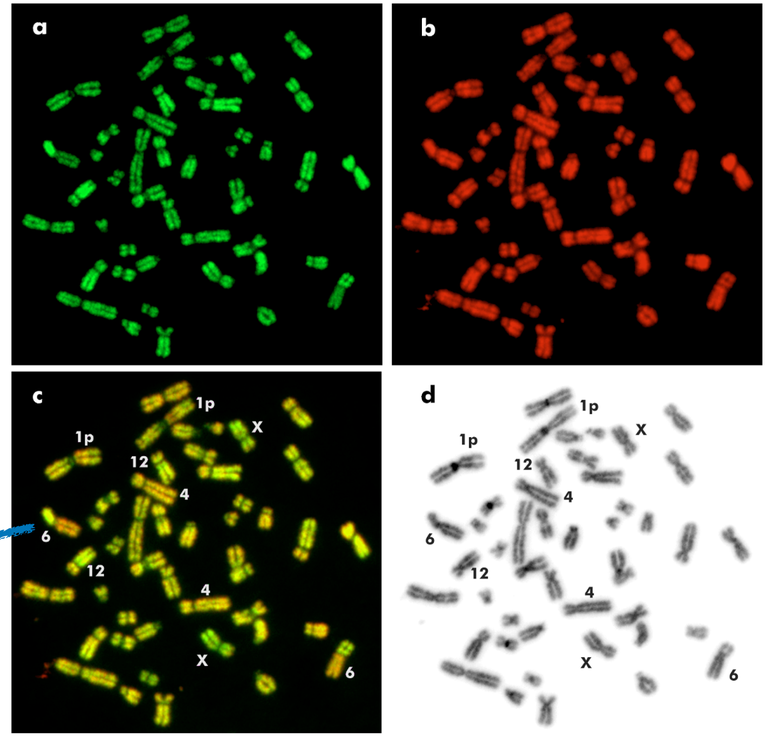
Chromosomal CGH: Normal metaphase spreads (cultured lymphocytes from healthy donors) on microscopy slides serve as the hybridization matrix for whole-genome DNA from tumor and reference tissue, labeled with different fluorophores. The regional ratio between the two colors points to (relative) changes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

Comparative Genomic Hybridization

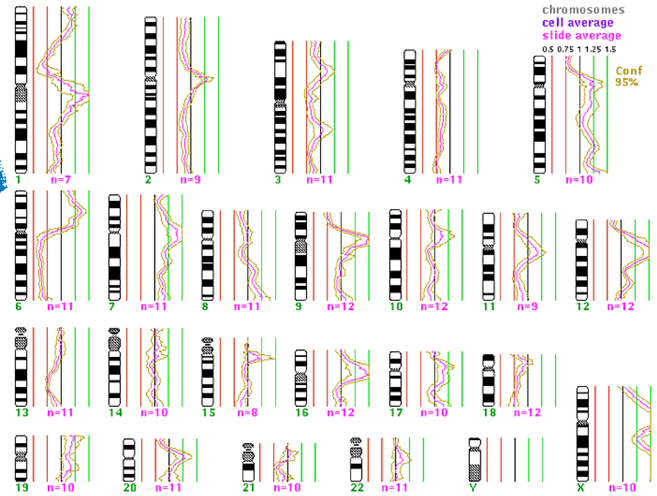
Molecular-Cytogenetic Technology for Genomic Imbalance Screening

- 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.



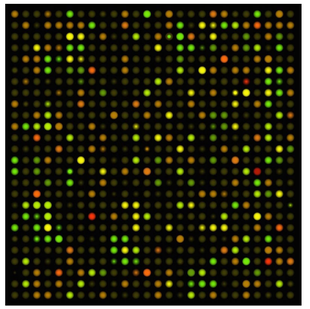
CGH-Experiment: **a** Hybridisierung mit Tumor-DNA; **b** Hybridisierung mit normaler menschlicher DNA als Kontrolle; **c** Überlagerung der Signale; **d** Bänderungsfärbung zur Identifizierung der Chromosomen



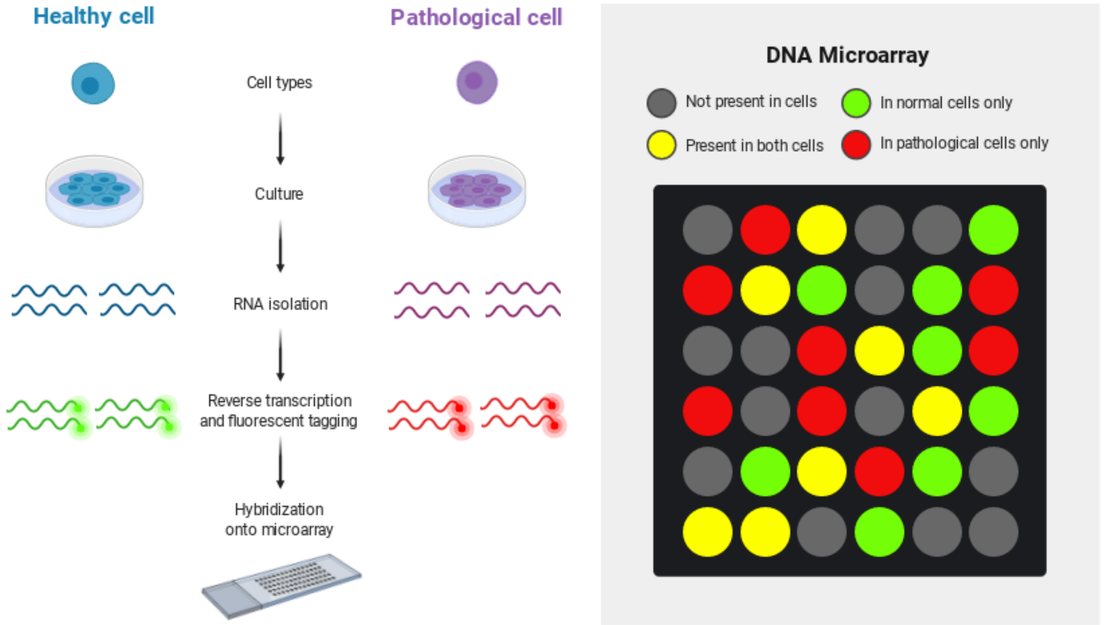
Auswertung: Summationsprofil der computergestützten Analyse mehrerer Metaphasen des dargestellten Falles; die Profilausschläge stehen für Zugewinne bzw. Verluste von chromosomalen Anteilen im Tumorgenom

Array Comparative Genomic Hybridization

Molecular Hybridization Technology for Genomic Imbalance Screening



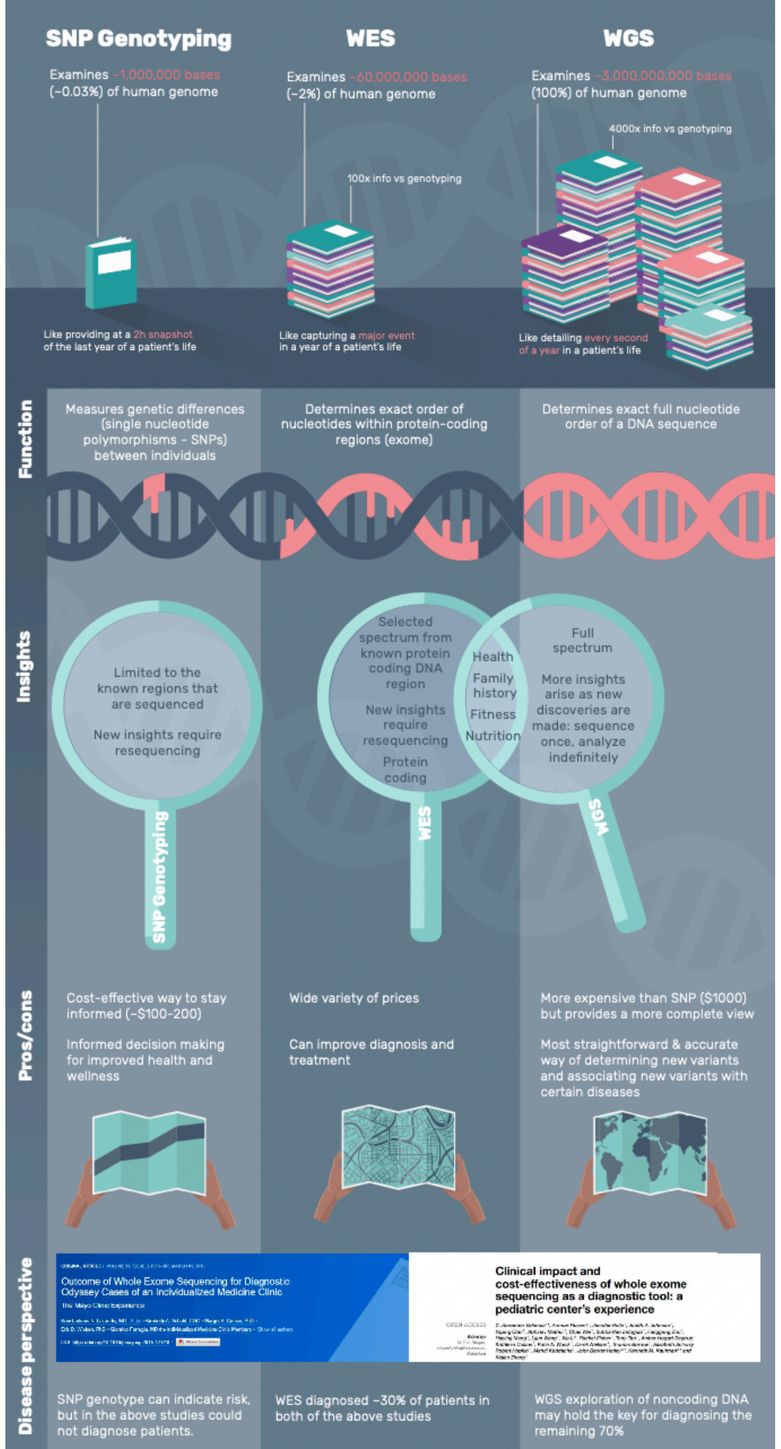
- Molecular-cytogenetic technique to identify genomic copy number variations (CNV/ CNA) for given sequences (multi-kb to Mb)
- based on ***in situ*** suppression **hybridization** of labeled **genomic** tumor and reference DNA against spotted or synthesized DNA clones or oligonucleotides
- analysis of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

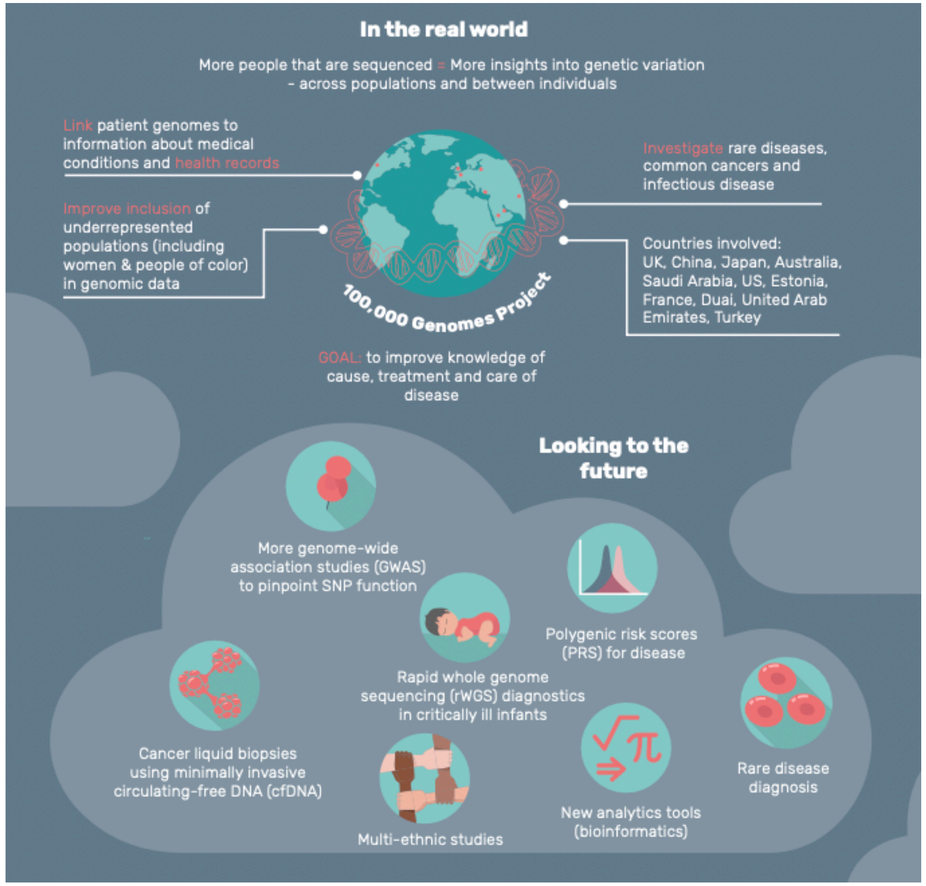


- Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes Cancer. 1997; 4 (20):399-407.
- Albertson DG, Pinkel D. Genomic microarrays in human genetic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



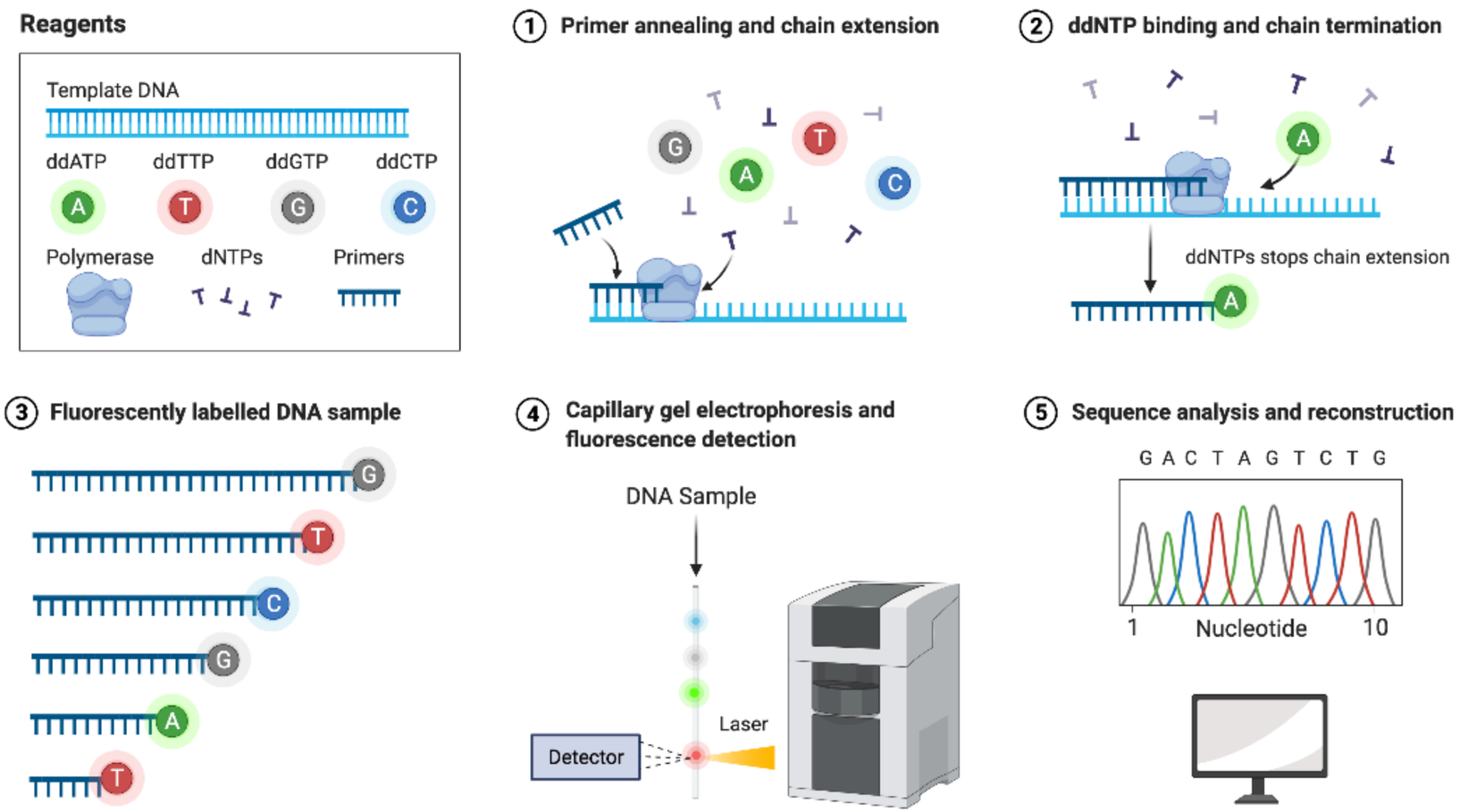


BioTechniques

psomagen

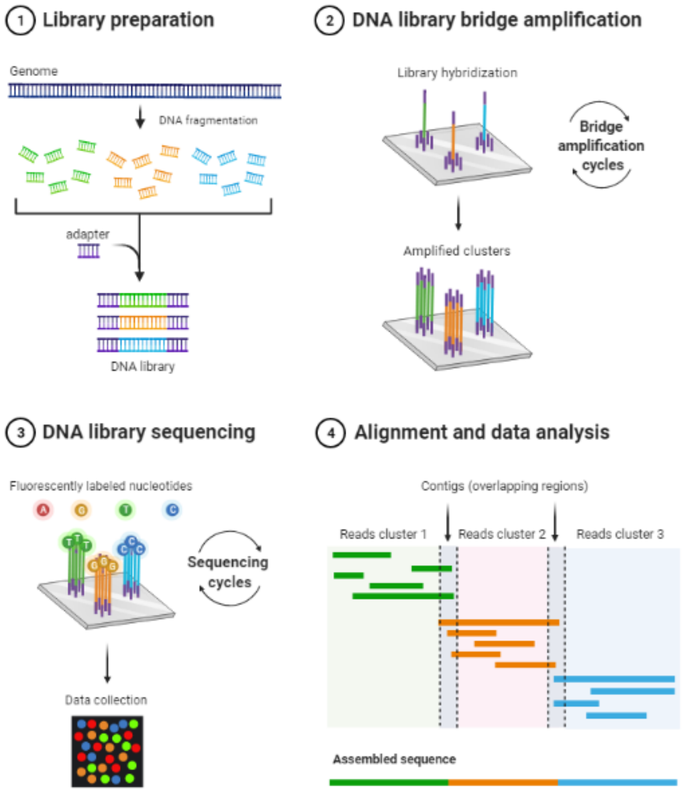
Genome Analysis

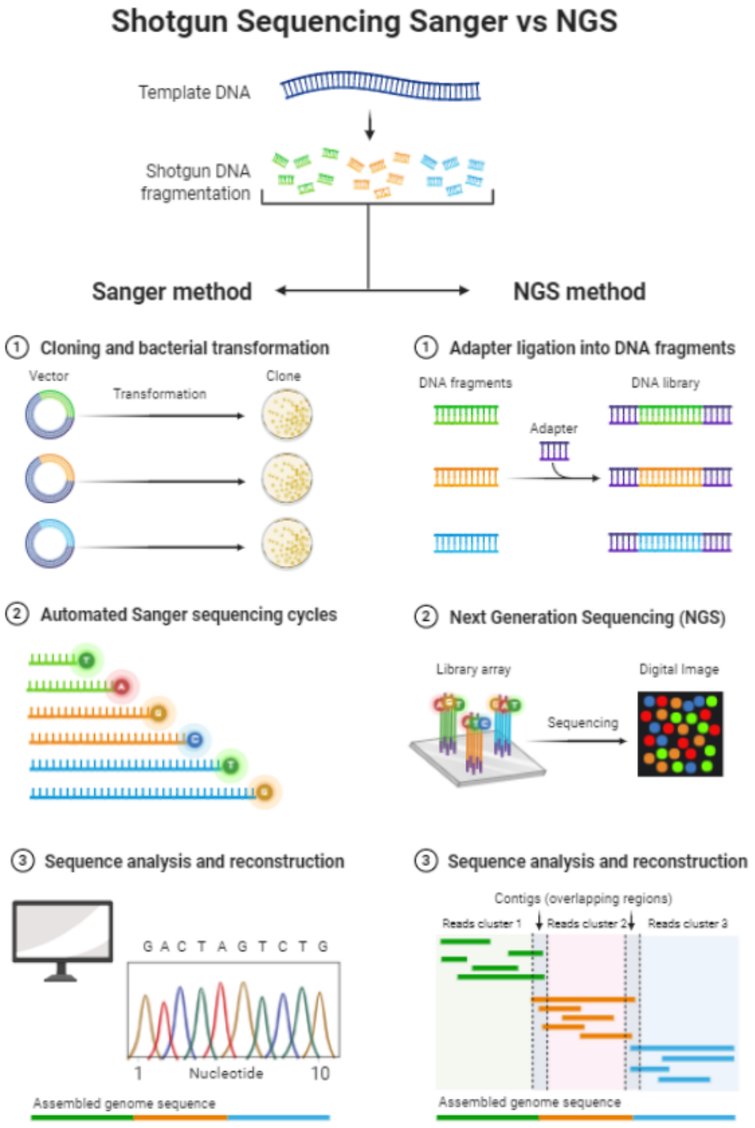
Sanger Sequencing



Genome Analysis

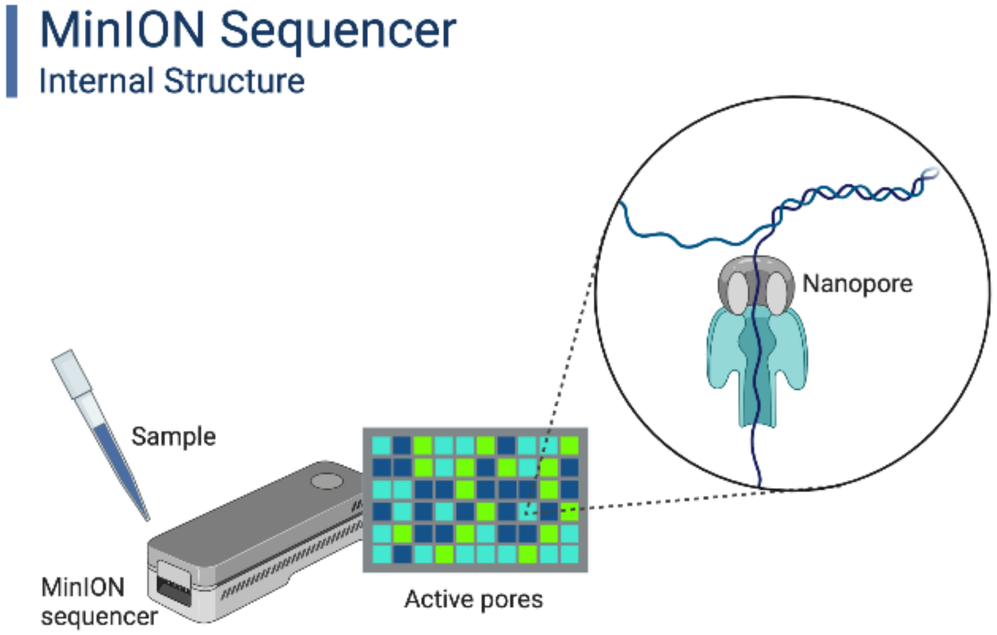
NGS vs. Sanger Sequencing

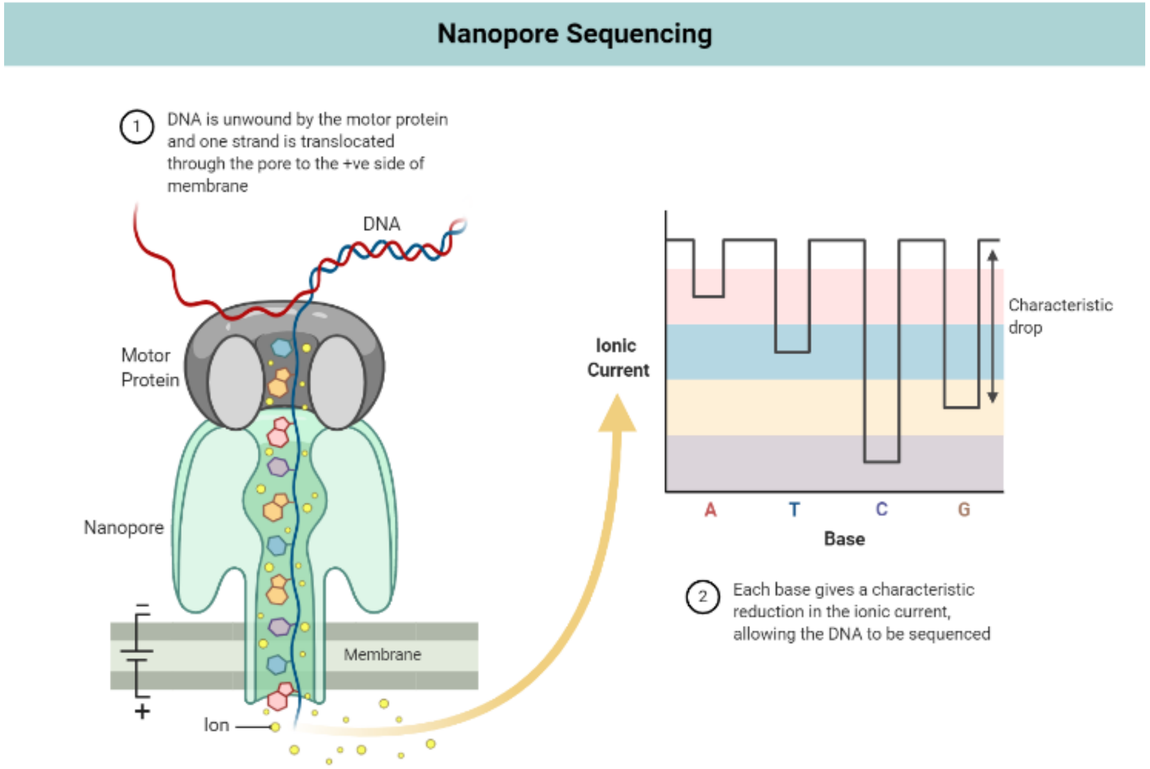




Genome Analysis

Nanopore Sequencing

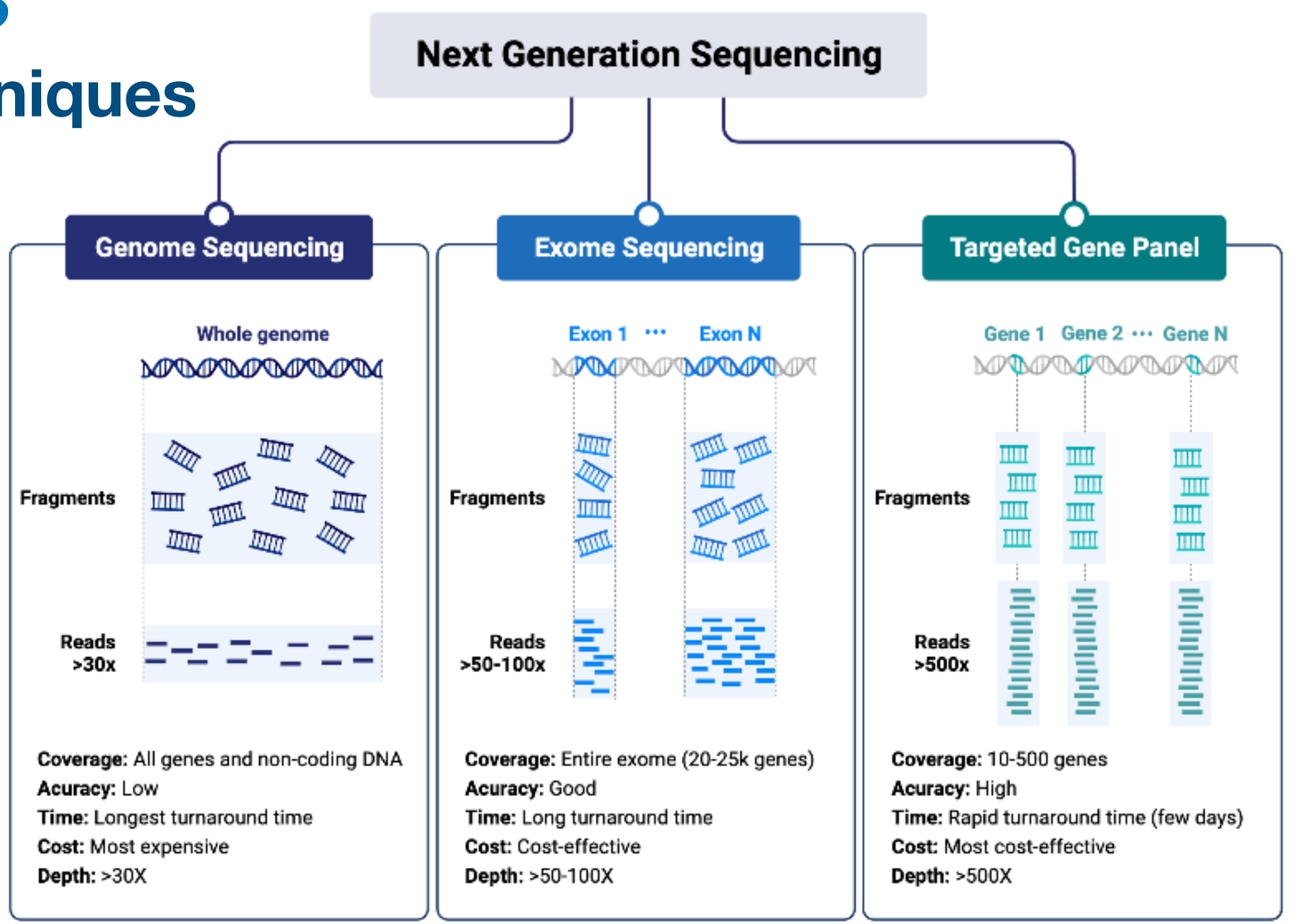




Genome Analysis

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

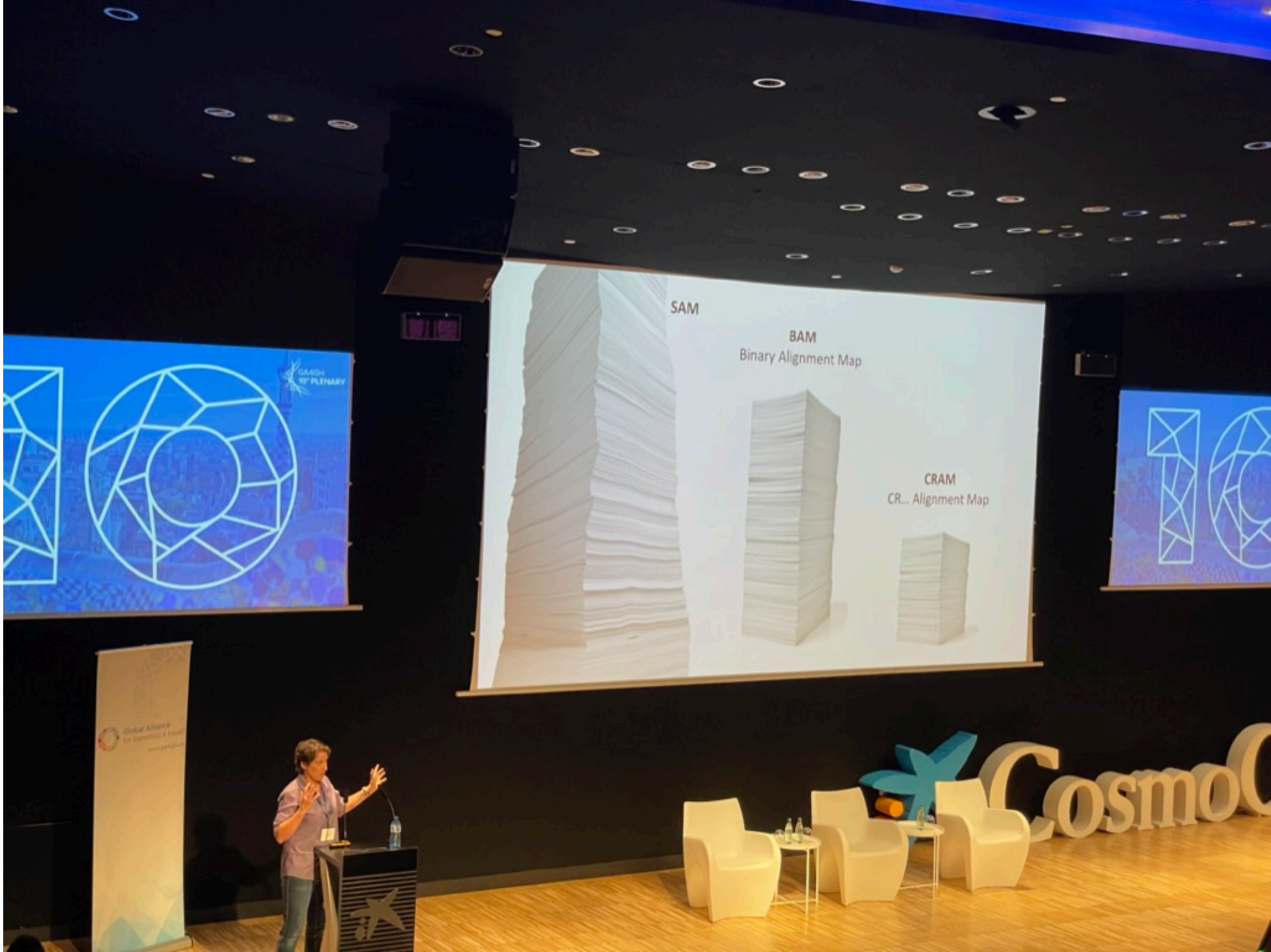
Types | Sizes | Use Cases

What is a PB, for human genomes?

It depends...

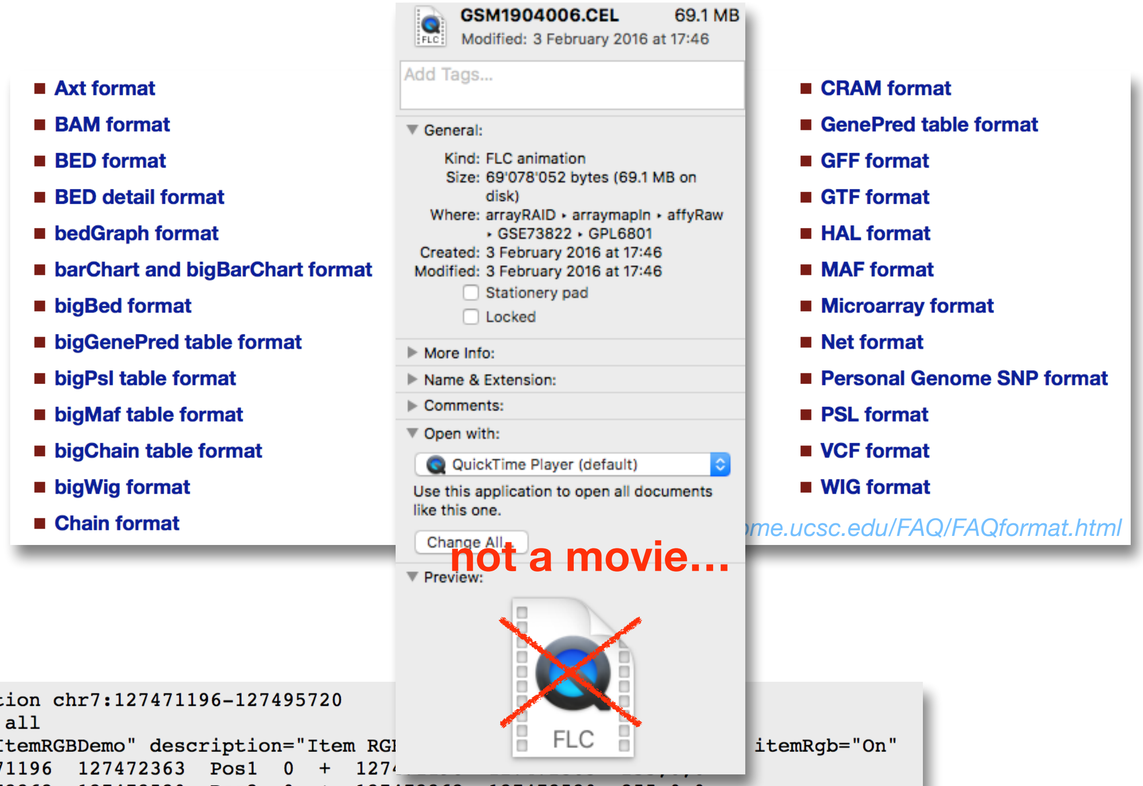
- 2 bits per base are sufficient to encode TCGA
  - using 00, 01, 10, 11
  - [TCGA]{3’000’000’000}
  - 2 * 3 * 10^9b = 6, 000, 000, 000 b
  - perfect genome (no overhead): ~715 MB
  - 1PB => ~1’400’000 genomes
- according to Swiss online store (Sep 2021) ~35'000CHF (65x16TB á CHF550)
- this is less than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

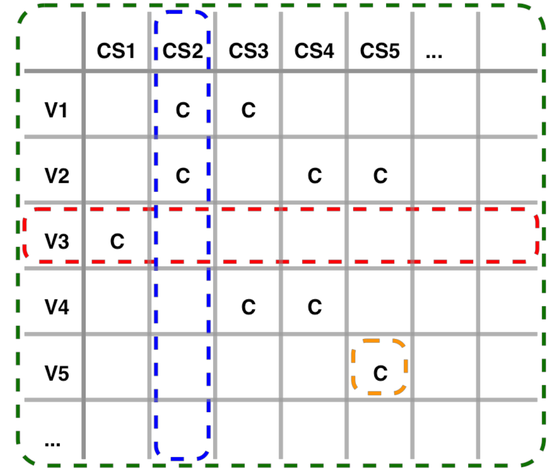



```
browser position chr7:127471196-127495720
browser hide all
track name="ItemRGBDemo" description="Item RGB                                    itemRgb="On"
chr7    127471196  127472363  Pos1  0  +  127                                    ,
chr7    127472363  127473530  Pos2  0  +  127472363  127473530  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

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

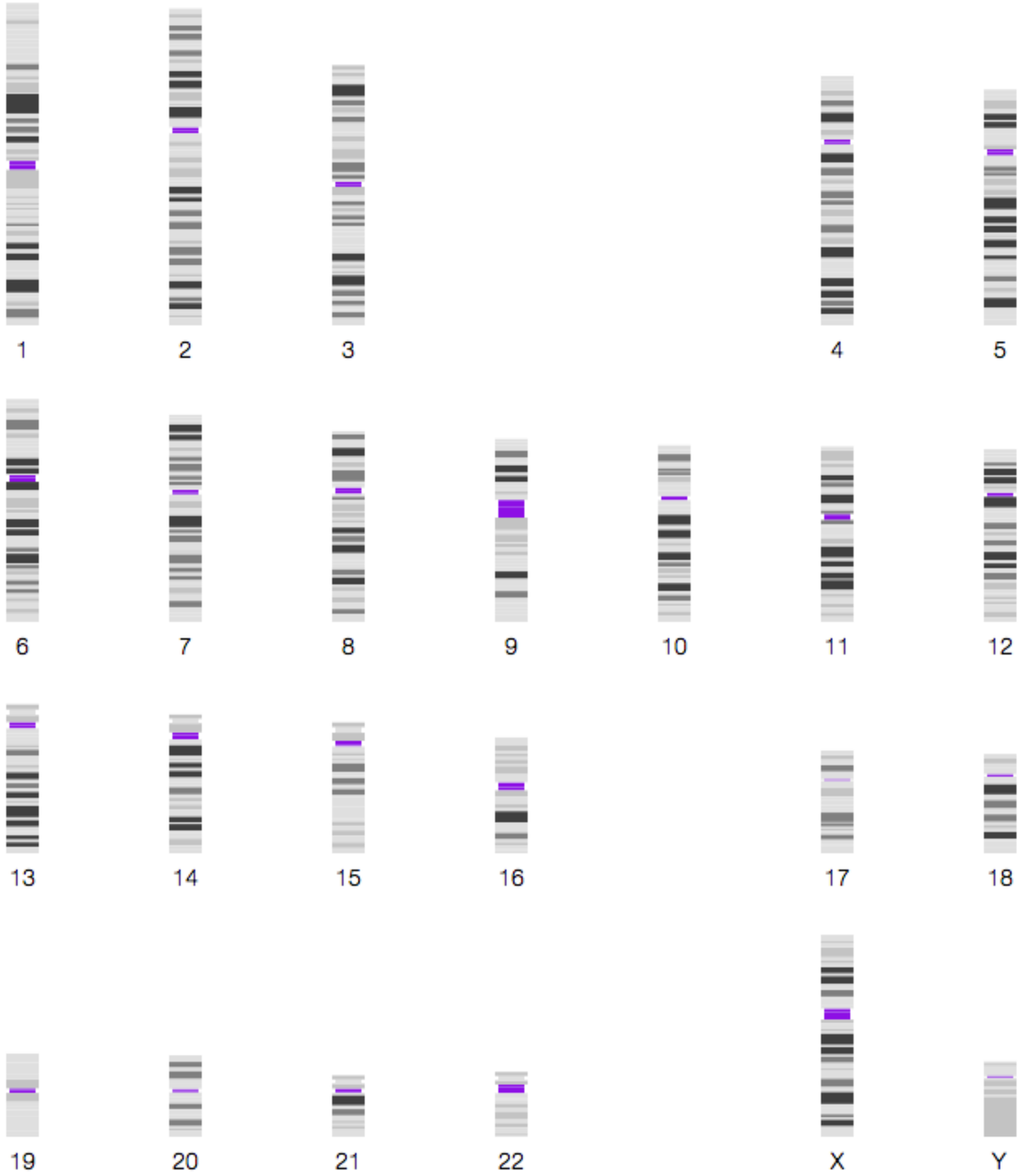
Task: Estimate Storage Requirements for 1000 Genomes

- 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	248'956'422
2	242'193'529
3	198'295'559
4	190'214'555
5	181'538'259
6	170'805'979
7	159'345'973
8	145'138'636
9	138'394'717
10	133'797'422
11	135'086'622
12	133'275'309
13	114'364'328
14	107'043'718
15	101'991'189
16	90'338'345
17	83'257'441
18	80'373'285
19	59'128'983
20	64'444'167
21	46'709'983
22	50'818'468
X	156'040'895
Y	57'227'415
	3'080'419'480

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	246'127'941	247'249'719	249'250'621	248'956'422	2'828'481
2	243'615'958	242'951'149	243'199'373	242'193'529	-1'422'429
3	199'344'050	199'501'827	198'022'430	198'295'559	-1'048'491
4	191'731'959	191'273'063	191'154'276	190'214'555	-1'517'404
5	181'034'922	180'857'866	180'915'260	181'538'259	503'337
6	170'914'576	170'899'992	171'115'067	170'805'979	-108'597
7	158'545'518	158'821'424	159'138'663	159'345'973	800'455
8	146'308'819	146'274'826	146'364'022	145'138'636	-1'170'183
9	136'372'045	140'273'252	141'213'431	138'394'717	2'022'672
10	135'037'215	135'374'737	135'534'747	133'797'422	-1'239'793
11	134'482'954	134'452'384	135'006'516	135'086'622	603'668
12	132'078'379	132'349'534	133'851'895	133'275'309	1'196'930
13	113'042'980	114'142'980	115'169'878	114'364'328	1'321'348
14	105'311'216	106'368'585	107'349'540	107'043'718	1'732'502
15	100'256'656	100'338'915	102'531'392	101'991'189	1'734'533
16	90'041'932	88'827'254	90'354'753	90'338'345	296'413
17	81'860'266	78'774'742	81'195'210	83'257'441	1'397'175
18	76'115'139	76'117'153	78'077'248	80'373'285	4'258'146
19	63'811'651	63'811'651	59'128'983	59'128'983	-4'682'668
20	63'741'868	62'435'964	63'025'520	64'444'167	702'299
21	46'976'097	46'944'323	48'129'895	46'709'983	-266'114
22	49'396'972	49'691'432	51'304'566	50'818'468	1'421'496
X	153'692'391	154'913'754	155'270'560	156'040'895	2'348'504
Y	50'286'555	57'772'954	59'373'566	57'227'415	6'940'860
	3'070'128'059	**3'080'419'480**	**3'095'677'412**	**3'088'781'199**	**18'653'140**

Genome Liftover

Moving between genome editions

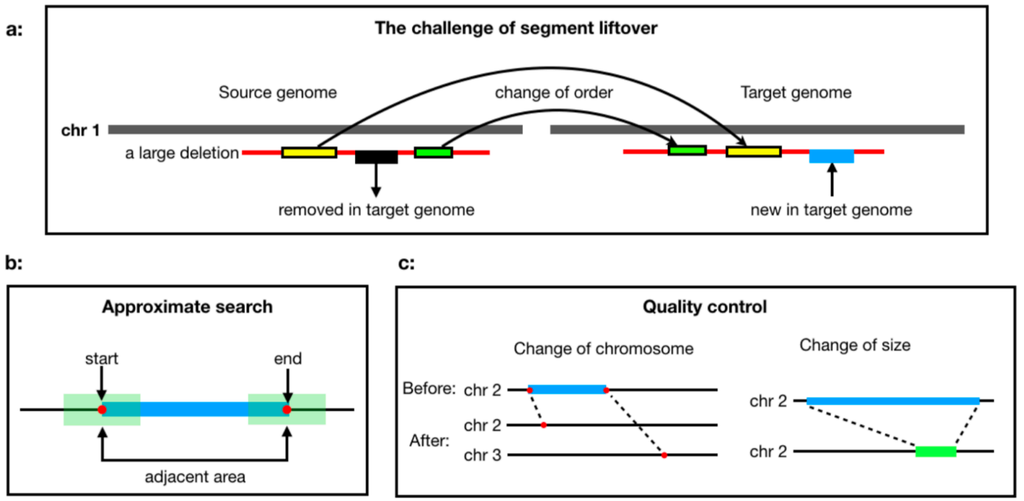
- different genome editions lead to shifting positions of defined elements such as genes
- local regions are frequently stable between editions
- shifts from change in regional lengths are defined in "chain files"
- chain files serve as guides for positional remapping using liftover methods
- Task: Read 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.

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
  - SNP, aCGH arrays
  - SKY, M-FISH
  - chromosomal CGH

➡ notes about usage (research, clinical, historical vs. current)

- "T2T genome"
  - What technologies enabled this?

BIO392: Course Schedule

https://compbiozurich.org/courses/UZH-BIO392/

	Tue Sep 19	Wed Sep 20	Thu Sep 21	Fri Sep 22	Tue Sep 26	Wed Sep 27	Thu Sep 28	Fri Sep 29	Tue Oct 3	Wed Oct 4	Thu Oct 5	Fri Oct 6	Tue Oct 10	Wed Oct 11
09:00 - 10:00		Github exercise: create user specific directories & upload/edit test files using Markdown (Ziying)	Izaskun: Terminal / Unix / Files	Izaskun: File formats for human genetic variation / file handling		Michael lecture introduction to some resources, CNVs, Progenetix	Hangjia: Clinvar and Clingen	Max & Feifei: Sequence analysis practical. FastQC, trimmomatic, BWA-MEM2, SAMtools, GangSTR, BCFtools		Max & Feifei:: analysis & interpretation. Parsing VCF (cyvcf2), UCSD genome browser, ENSEMBL variant effect predictor	Rahel: survival analysis	Michael: Genomic data risks & opportunities		Exam
10:00 - 11:00		Ziying: github desktop and terminal	Izaskun: Terminal / Unix / Files	Izaskun: File formats for human genetic variation / file handling			Hangjia: blast	Max & Feifei:: Sequence analysis practical		Max & Feifei:: analysis & interpretation.	Rahel: survival analysis	Michael: Genomic data risks & opportunities		Exam
11:00 - 12:00		Ziying: python warmup and exercise		Izaskun: File formats for human genetic variation / file handling		Hangjia: Progenetix as tool for CNV frequencies etc.	Hangija: Blast exercise	Max & Feifei:: Sequence analysis practical		Max & Feifei:: analysis & interpretation.	Rahel: survival analysis	Michael: Genomic data risks & opportunities		Exam
13:00 - 14:00	* Room information * Administrative - discuss times/days - exam	Hangjia: R enviroment introduction	Izaskun: SIB online introduction to Unix	Izaskun: short project (1000 genomes), reading, literature	Recap W1; Q&A	Task: Browse/explore genome resources and provide some notes (1-2 pages total) in a doc posted on Github (.md)	Max: Sequence analysis introduction. Overview of pipeline from raw reads -> variant calling & interpretation	Max & Feifei:: STR reading up	Michael lecture analysis principles (why surv, etc.)	Rahel: survival analysis	Rahel & Michael: presentation/discussion	graded exercise: genomic data risks ...	Exam revision, Q&A	
14:00 - 15:00	Tina Siegenthaler: technical introduction (room, computer, accounts)	Hangjia: R exercise	Izaskun: SIB online introduction to Unix	Izaskun: short project (1000 genomes), reading, literature	Literature (genome analysis techniques ...)		Max: Sequence analysis introduction	Max & Feifei:: STR reading up		Rahel: survival analysis	Rahel & Michael: presentation/discussion			
15:00 - 16:30	* explore course site * create Github accounts and forward to bio392@compbiozurich.org * Ziying&Hangjia: overall schedule of the course	Ziying: paper reading, Q & A		Izaskun: short project (1000 genomes), reading, literature	Genome technologies - brief notes about usage scenarios, pro & con					Rahel: survival analysis				

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