



University of
Zurich^{UZH}

BIO392

Bioinformatics of Genome Variations

Genomic Analysis Technologies...

Michael Baudis **UZH SIB**
Computational Oncogenomics



Theodor Boveri (1914)

Observations in sea urchin eggs

- **Cell-cycle checkpoints** (“Hemmungseinrichtung”)
- **Tumour-suppressor genes** (“Teilungshemmende Chromosomen”), which may be overcome by external signals, and can be eliminated during tumour progression
- **Oncogenes** (“Teilungsfoerdernde Chromosomen”) that become amplified (“im permanenten Übergewicht”)
- **Progression** (benign to malignant), w/ sequential changes of chromosomes
- Clonal origin & Genetic mosaicism
- Cancer **predisposition** through inheritance of “chromosomes” that are less able to suppress malignancy
- Inheritance of the same 'weak chromosome' from both parents leads to **homozygosity** and, consequently, to high-penetrance cancer syndromes - (e.g. xeroderma pigmentosum)
- Wounding and inflammation in tumour promotion; loss of cell adhesion in metastasis; sensitivity of malignant cells to radiation therapy (based on Hertwig *et al.*)

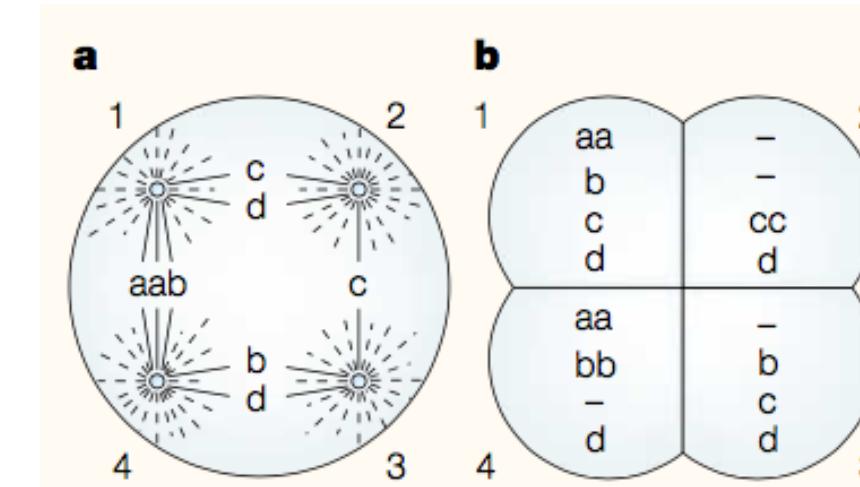
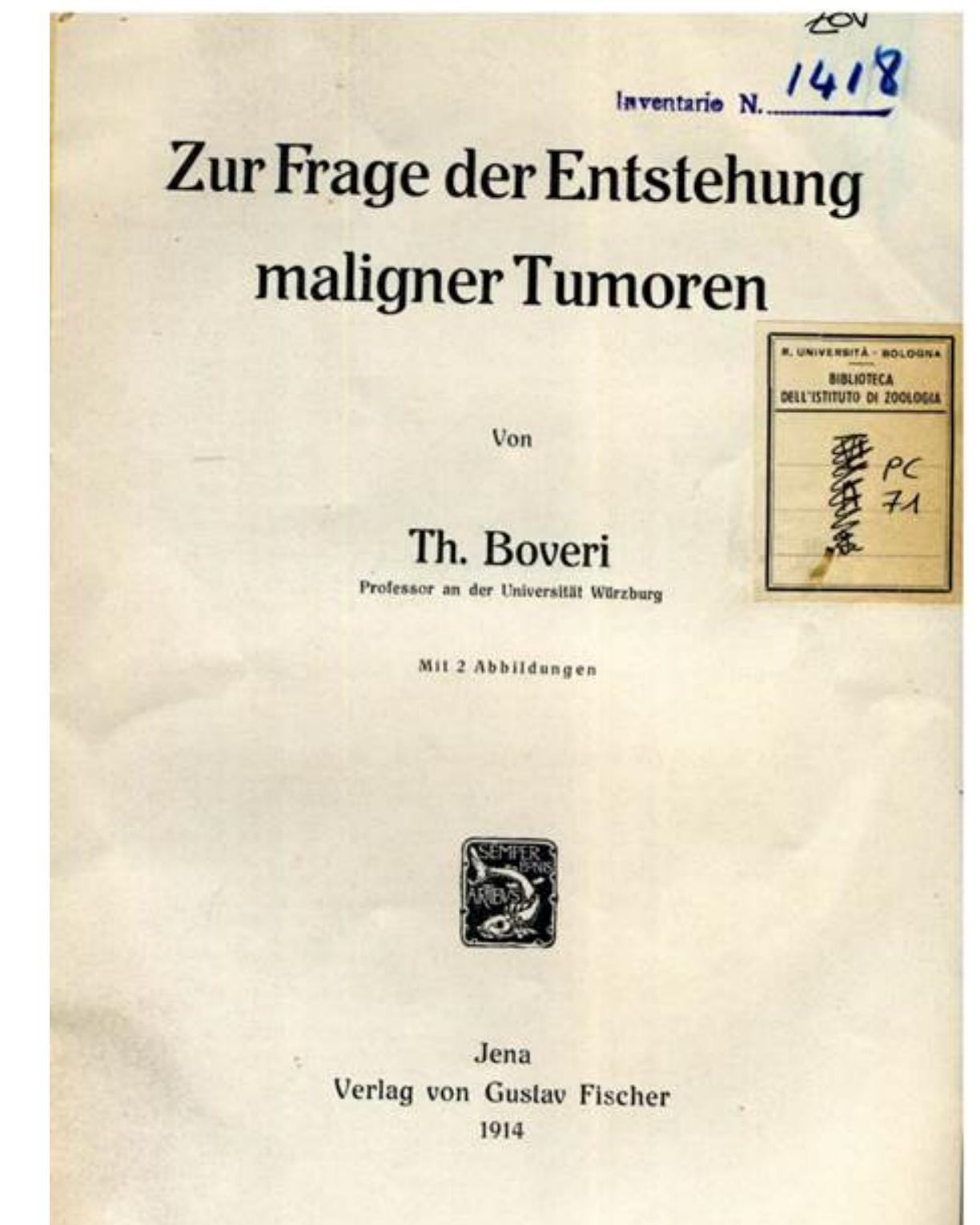
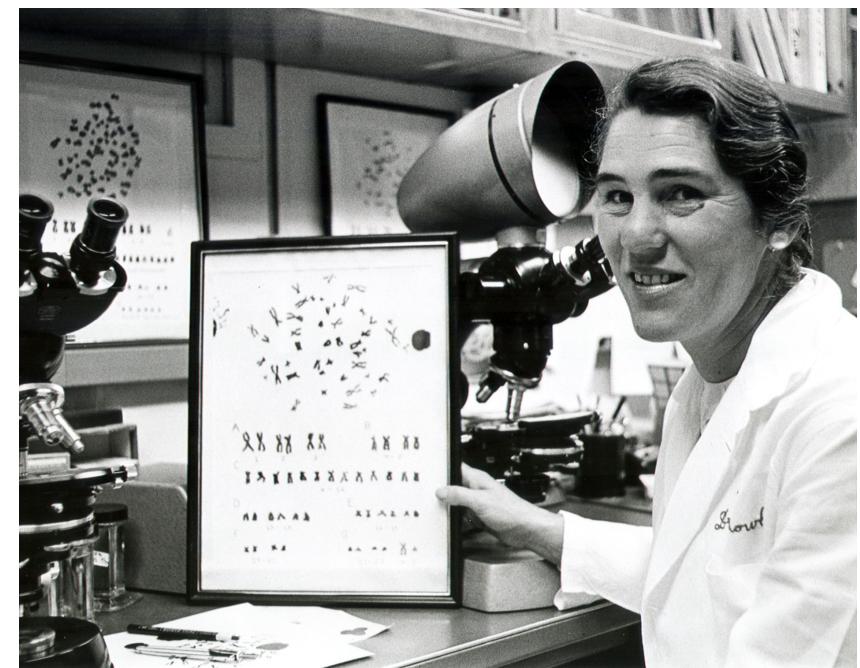


Figure 2 | **Multiple cell poles cause unequal segregation of chromosomes.** **a** | Boveri showed that fertilization of sea-urchin eggs by two sperm results in multiple cell poles. Individual chromosomes then attach to different combinations of poles — for example, one copy of chromosome c is attached to poles 1 and 2, and one copy is attached to poles 2 and 3.
b | Chromosomes are segregated to the four poles at cell division, leaving some cells with too many copies of the chromosomes and some with too few — for example, cell 2 has two copies of chromosome c and cell 4 has none.



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

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



Janet Rowley (1972/73)

Chromosomal translocations in cancer

- Recurrent chromosomal translocations in leukemias and lymphomas
- "Philadelphia chromosome" in CML (Nowell & Hungerford, 1960) represents a reciprocal translocation between chromosomes 9 and 22
- 1972: t(8;21) ALL manuscript rejected by NEJM
- 1973: t(9;22) manuscript rejected by *Nature* "with some reasonable comments and some truly wrong"
- Clinical implications: **Tyrosine Kinase inhibitors** as standard first-line therapy in CML
 - first trials in 1998 (STI-571; Imatinib/Gleevec)
 - cf. Druker BJ, Lydon NB (2000). Lessons learned from the development of an Abl tyrosine kinase inhibitor... *J Clin Invest* 2000;105:3-7

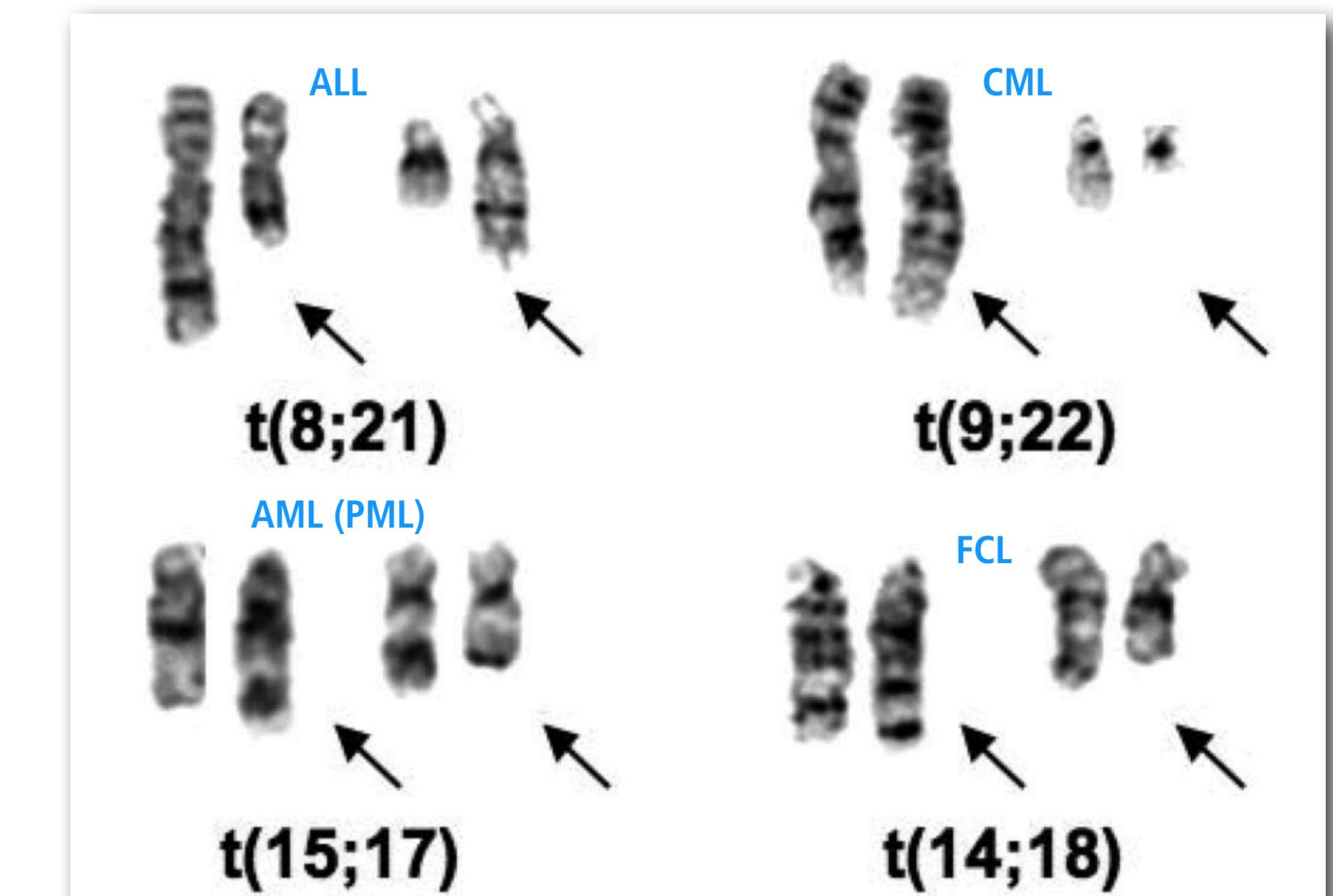
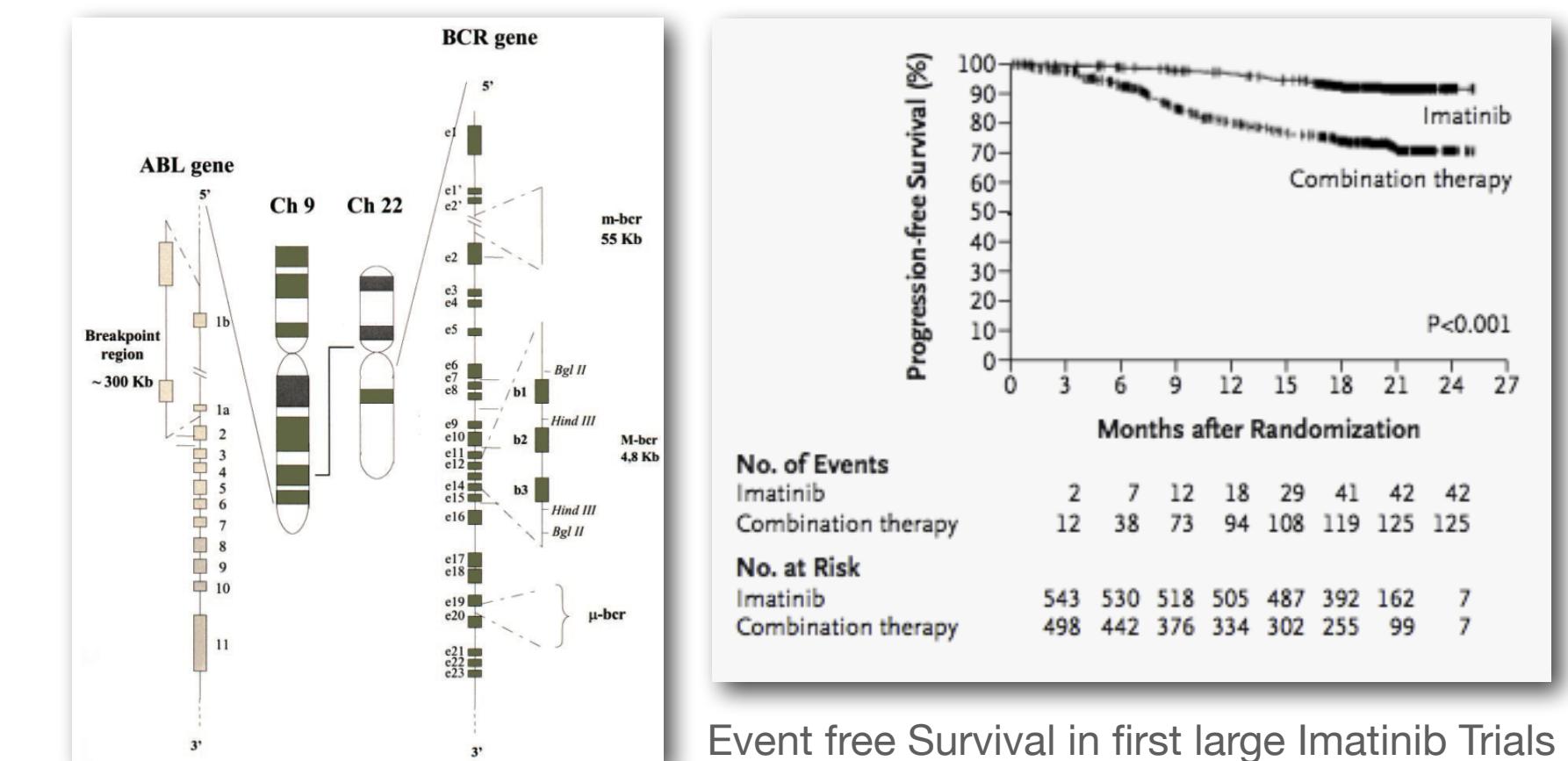


Figure 1. Partial karyotypes of common translocations discovered by Rowley.
The translocations appear in the order in which they were discovered.

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



Event free Survival in first large Imatinib Trials

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

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

Non-Sequence™ Genomics (Molecular) Cytogenetics

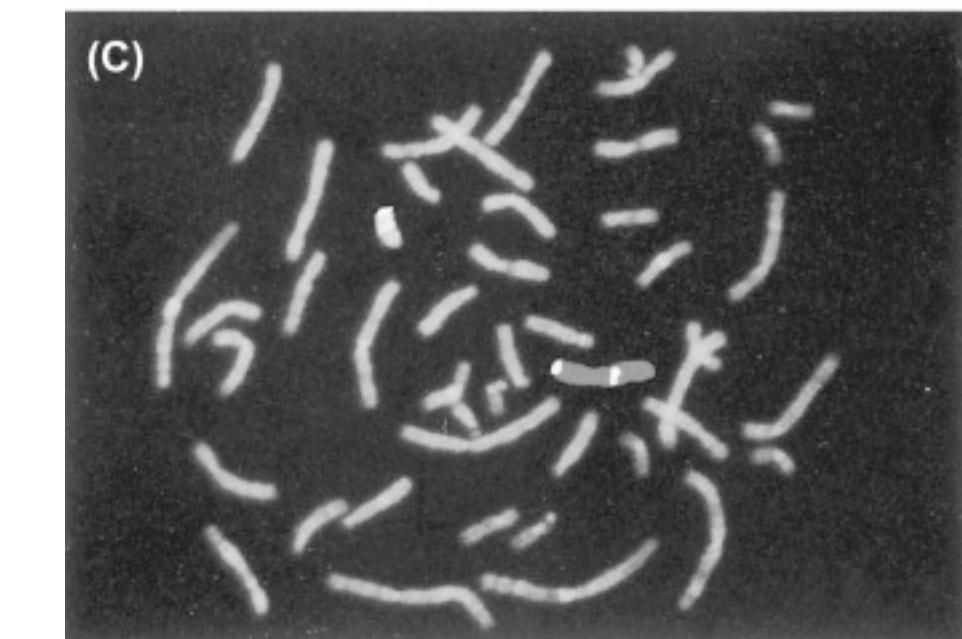
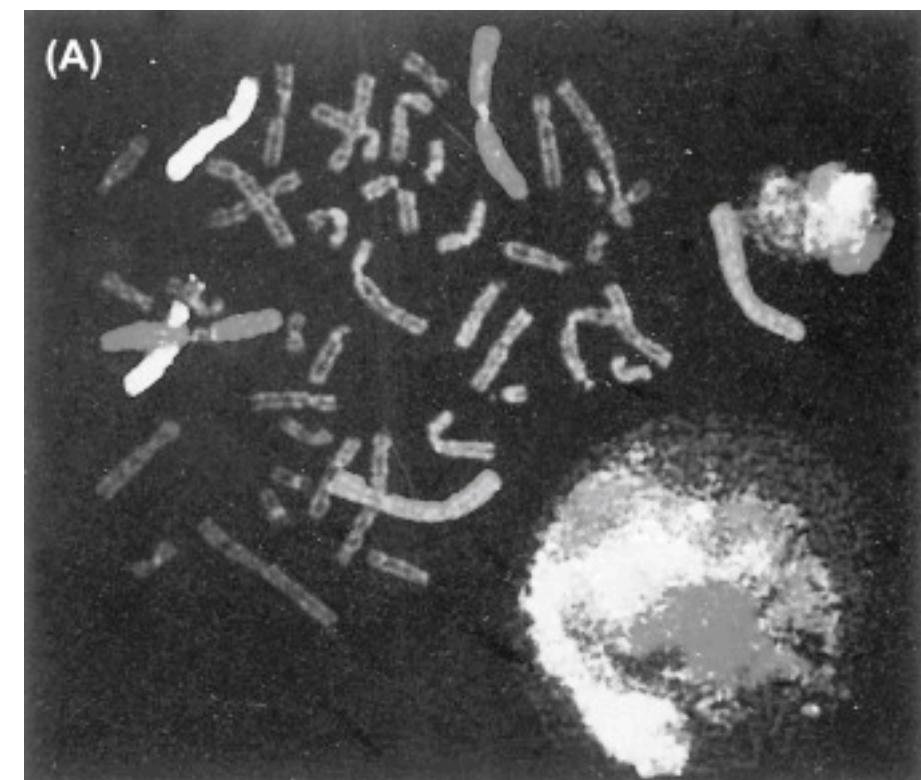
- structural changes in chromosomes can be analyzed by cytogenetics (chromosomal analysis) w/o knowledge about sequence alterations
- cytogenetics may be modified by
 - ▶ hybridization of fluorescent (or radiolabelled) probes of known DNA content or localization
 - ▶ using of fluorescently labeled "painting"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

→ **Molecular Cytogenetics**



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

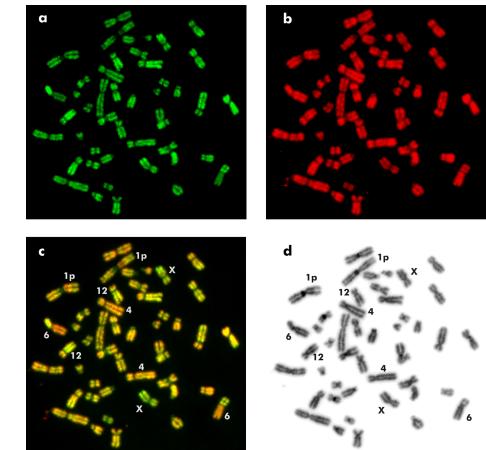
Images from "Emery and Rimoin's Principles and Practice of Medical Genetics and Genomics"
Chapter 5 - Cytogenetic Analysis -
Nancy B. Spinner & Malcolm A. Ferguson-Smith



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

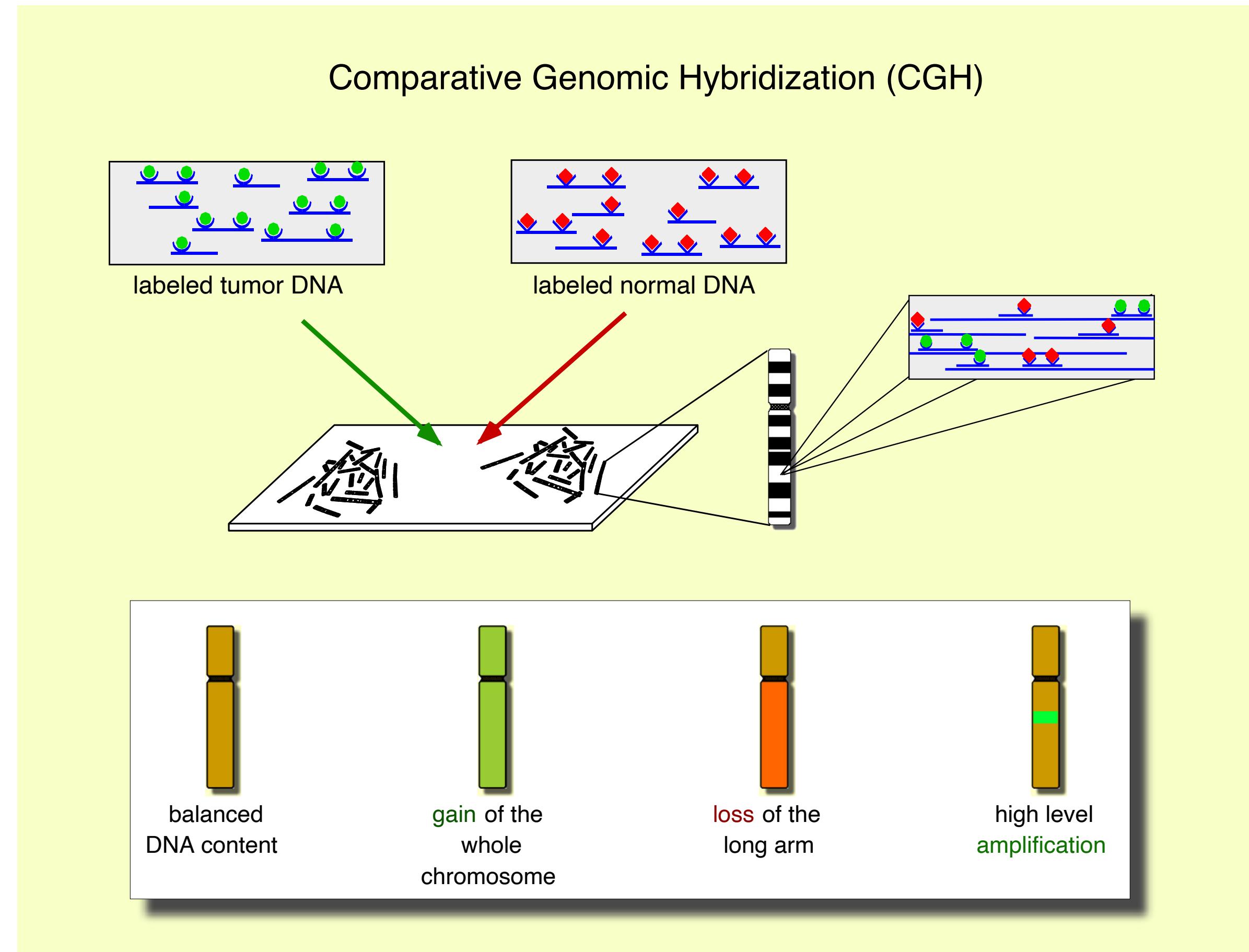
Comparative Genomic Hybridization

Molecular-Cytogenetic Technology for Genomic Imbalance Screening



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

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

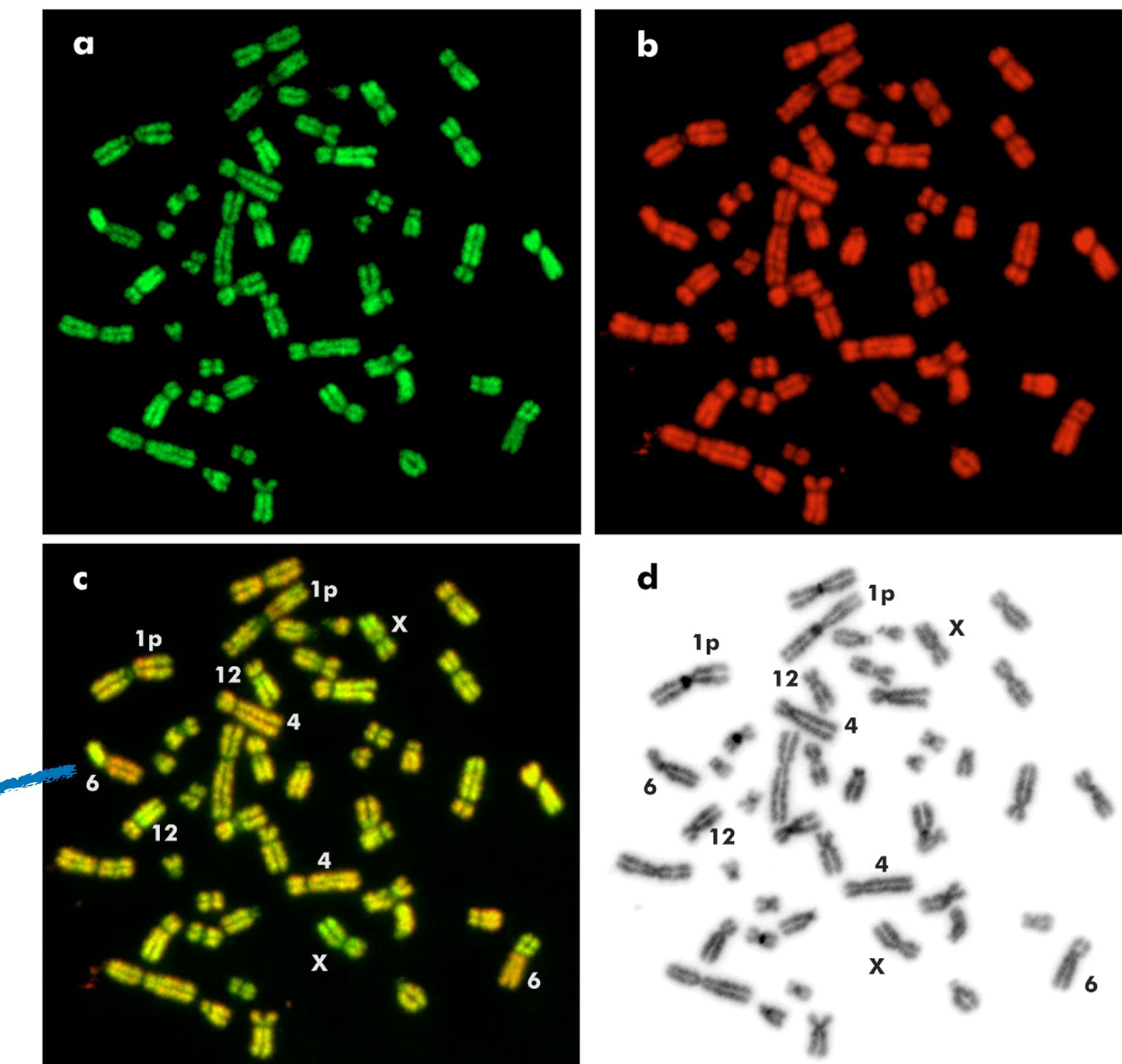


Chromosomal CGH: Normal metaphase spreads (cultured lymphocytes from healthy donors) on microscopy slides serve as the hybridization matrix for whole-genome DNA from tumor and reference tissue, labeled with different fluorophores. The regional ratio between the two colors points to (relative) changes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

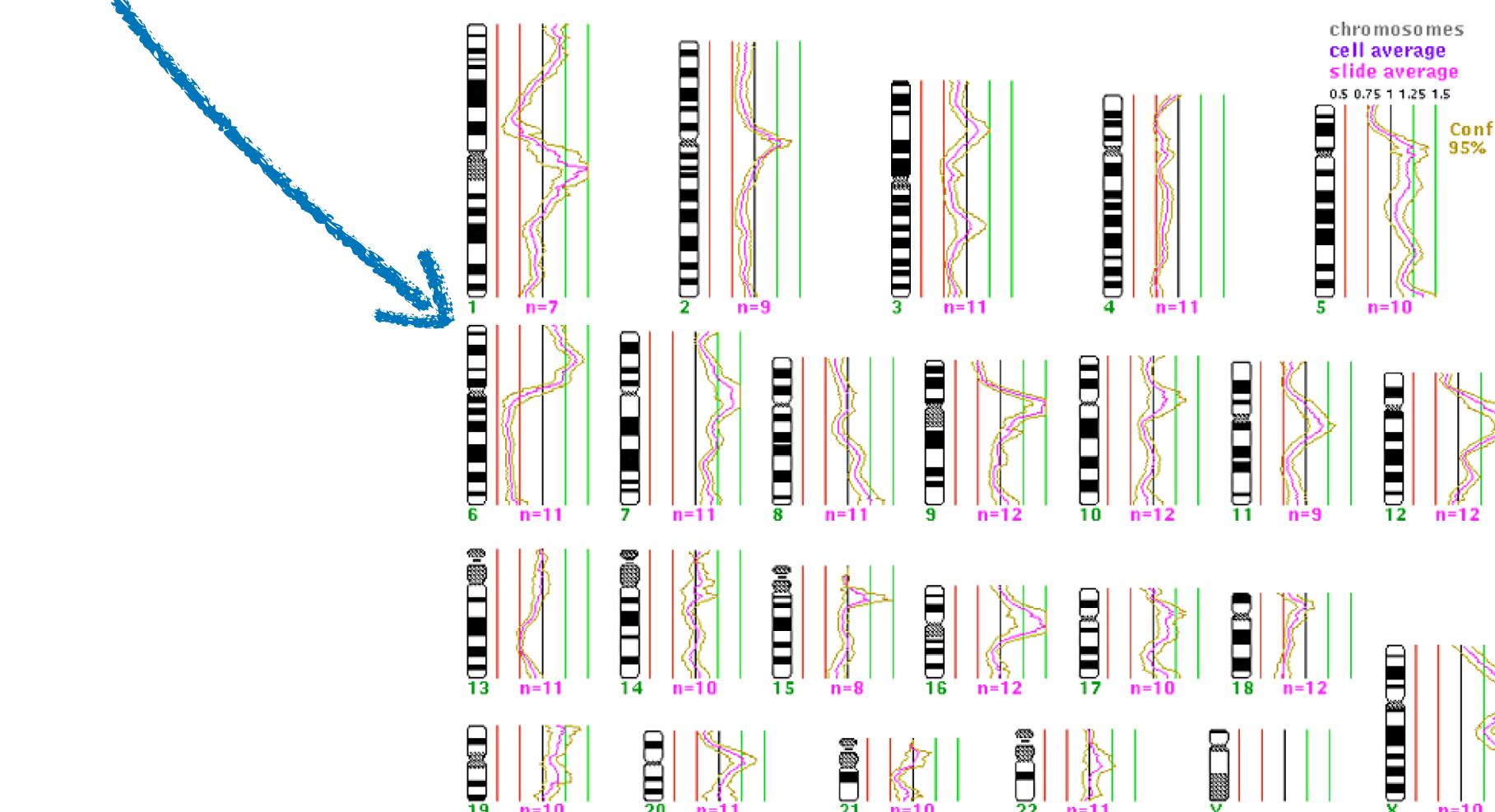
Comparative Genomic Hybridization

Molecular-Cytogenetic Technology for Genomic Imbalance Screening

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



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



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

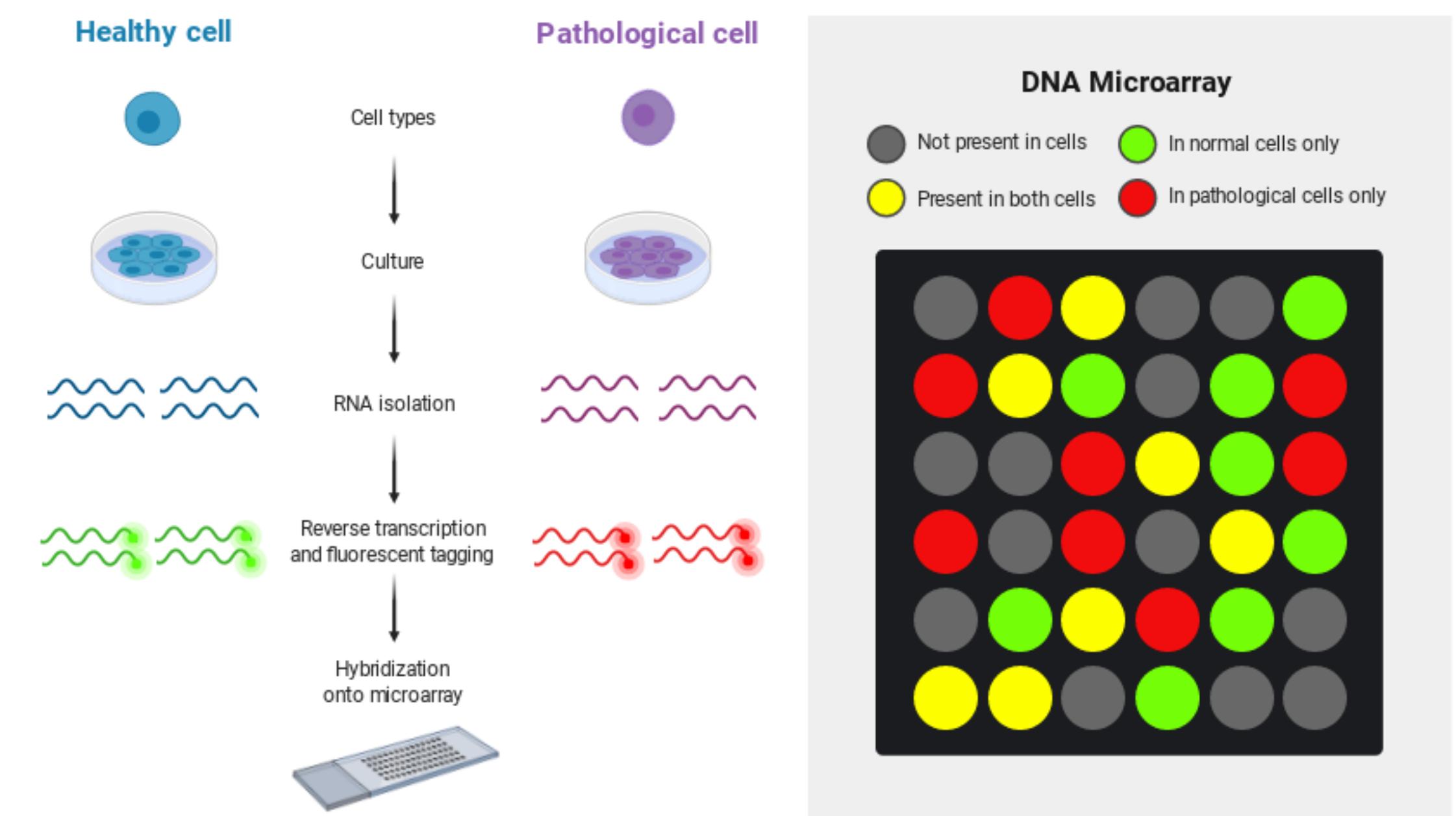
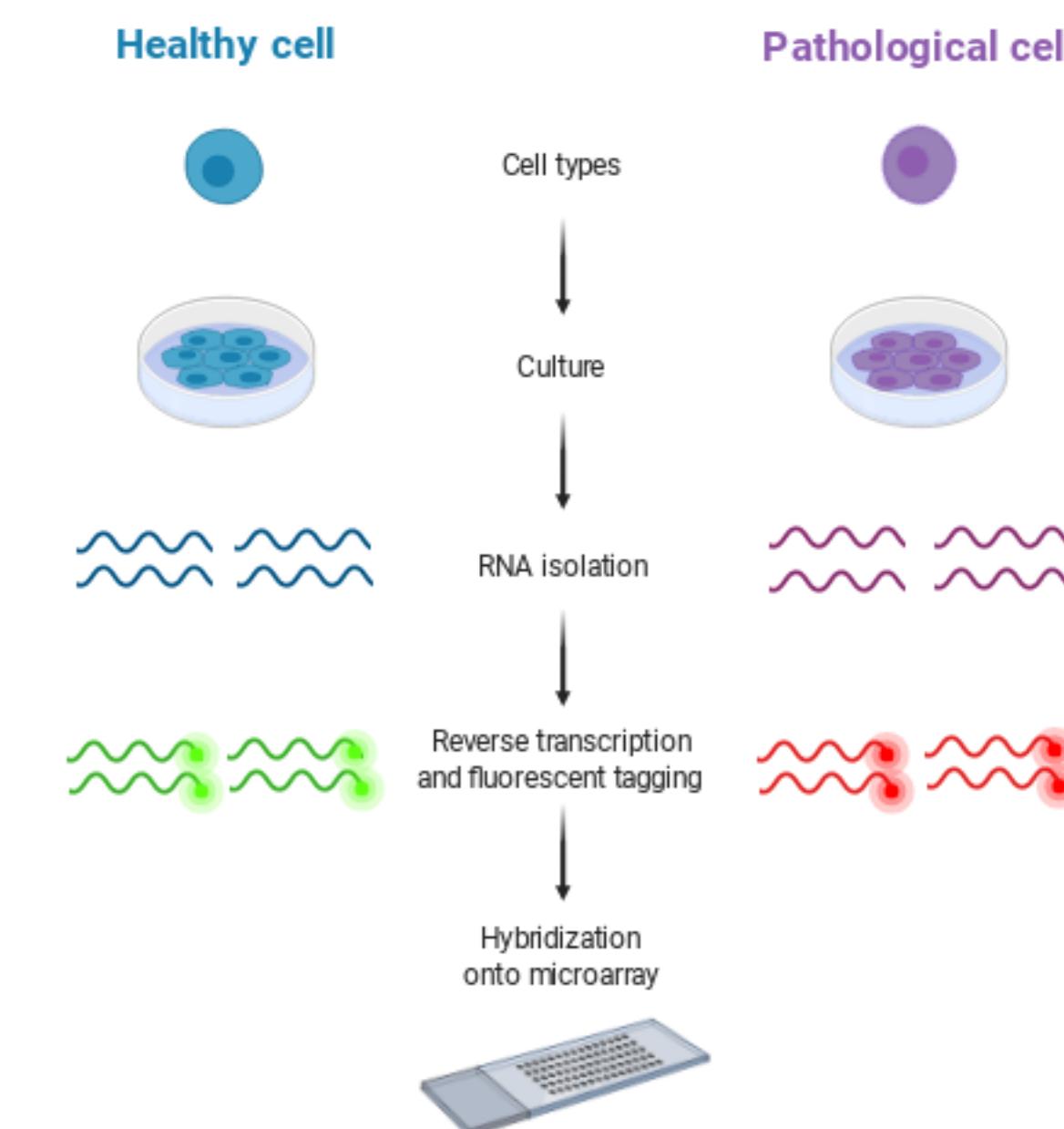
- Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science. 1992;258:818-821.
- Joos S, Scherthan H, Speicher MR, Schlegel J, Cremer T, Lichter P. Detection of amplified DNA sequences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

Array Comparative Genomic Hybridization

Molecular Hybridization Technology for Genomic Imbalance Screening

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

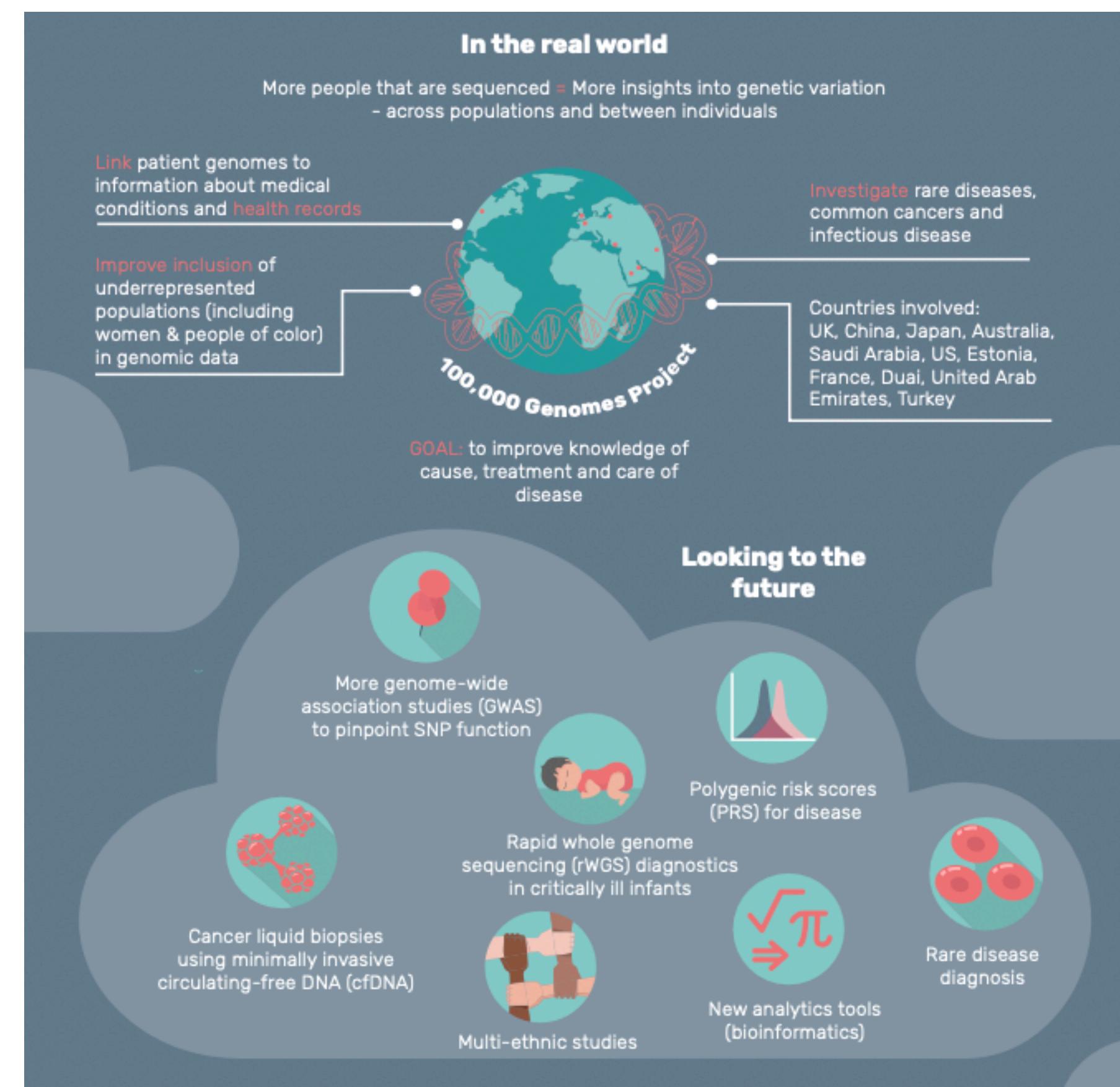
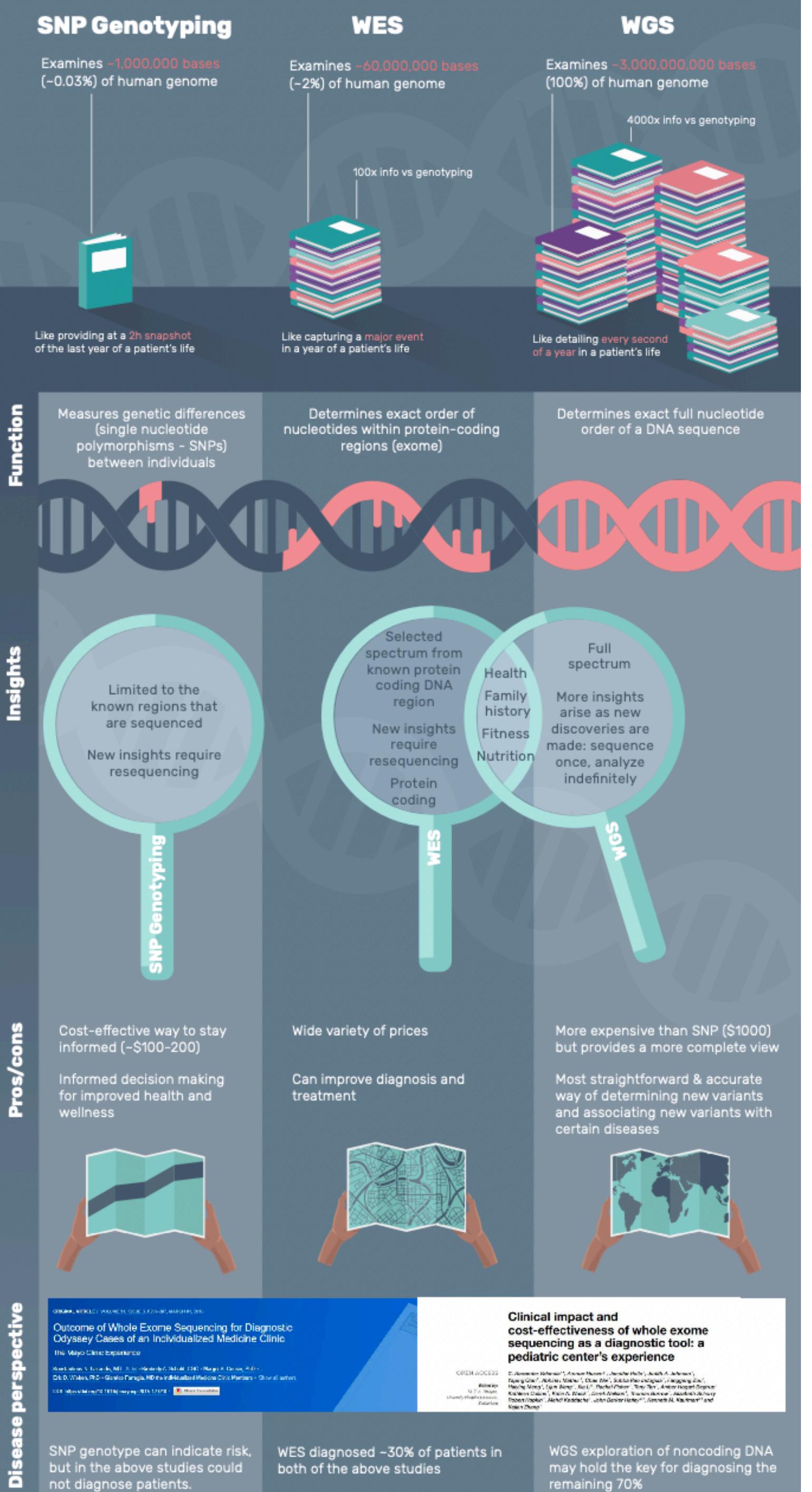
- Solinas-Toldo S, Lampel S, Stilgenbauer S, Nickolenko J, Benner A, Dohner H, Cremer T, Lichter P. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. *Genes Chromosomes Cancer.* 1997; 4 (20):399-407.
- Albertson DG, Pinkel D. Genomic microarrays in human genetic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



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

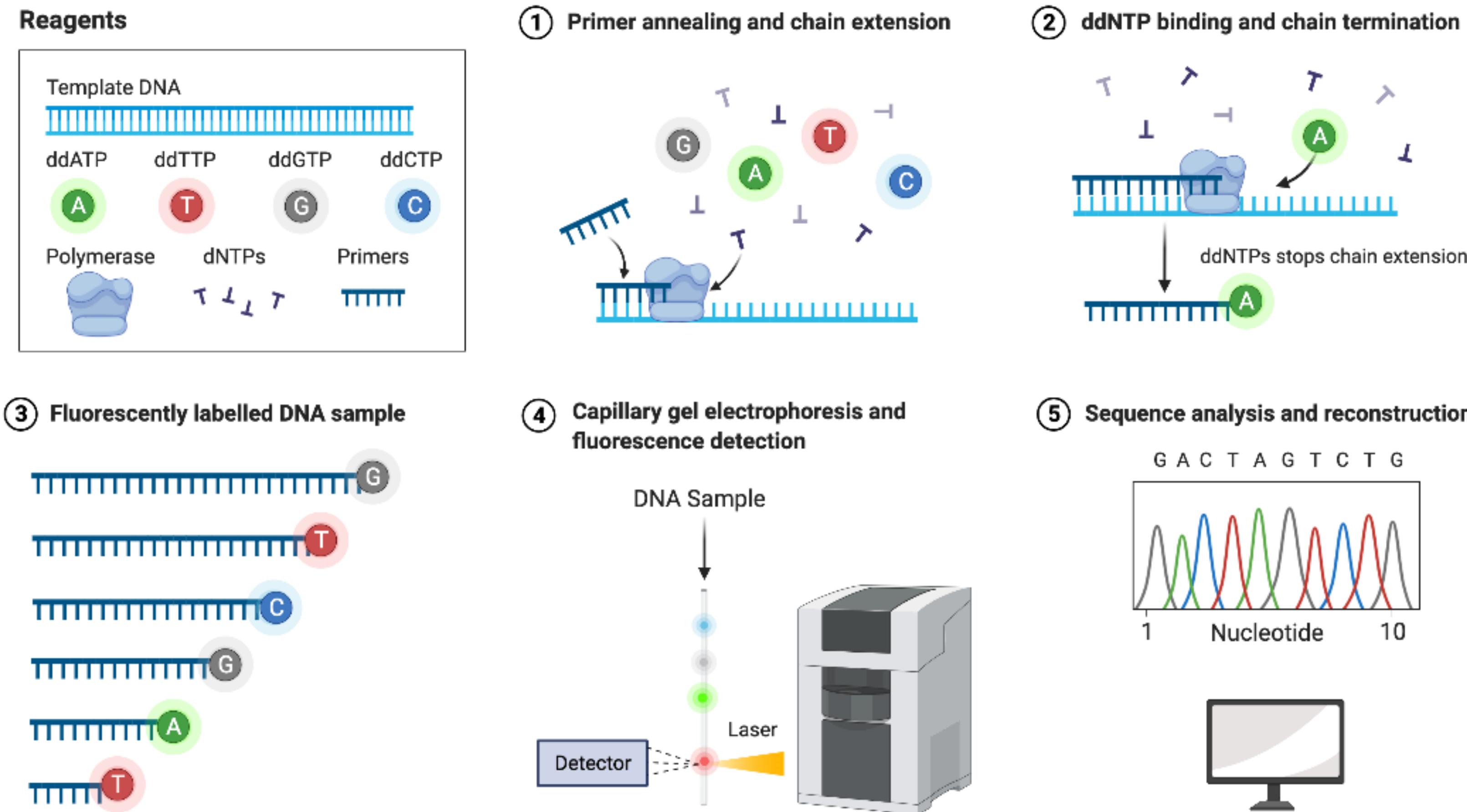
Genome Analysis

A “progressing technologies” view



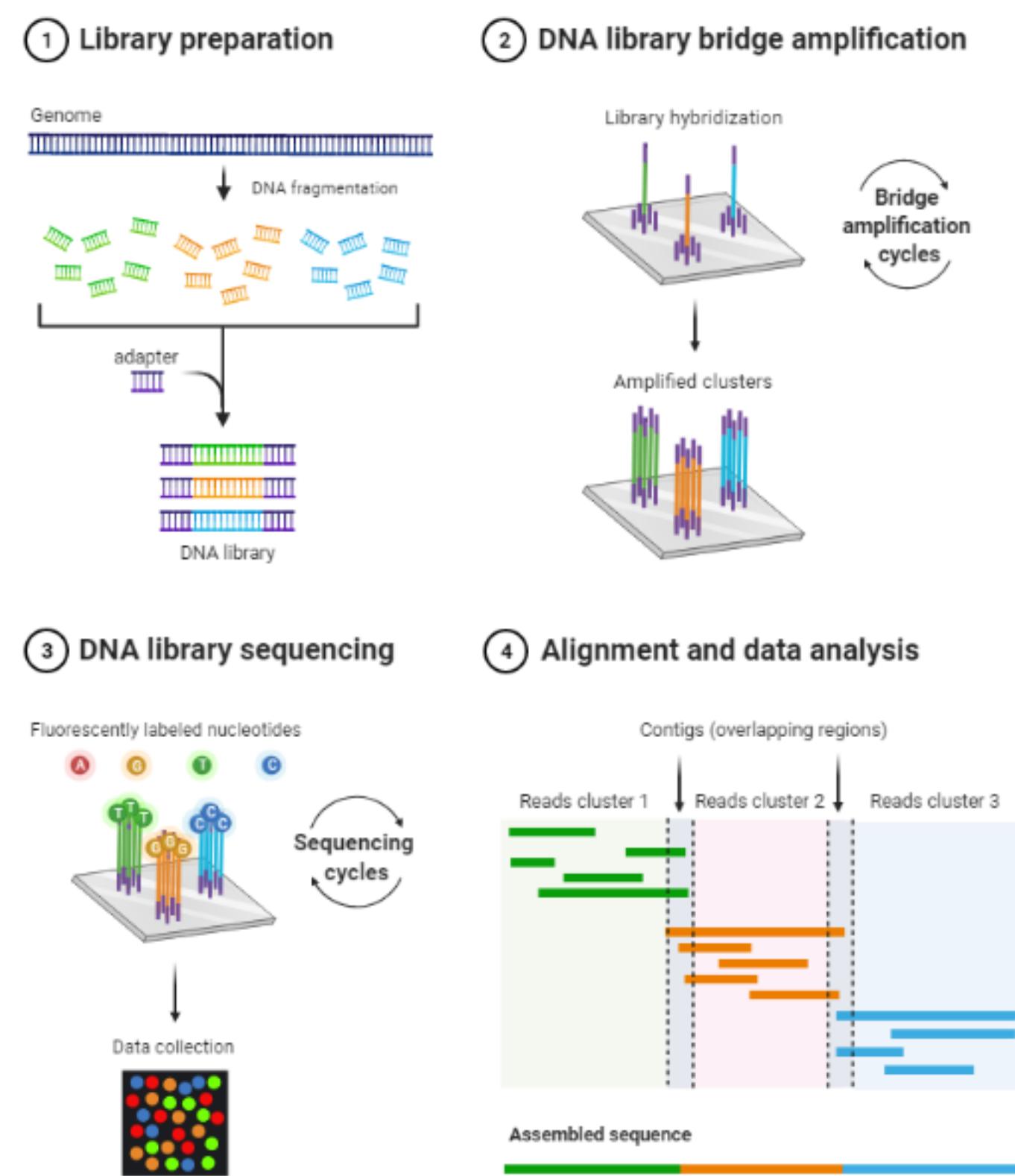
Genome Analysis

Sanger Sequencing

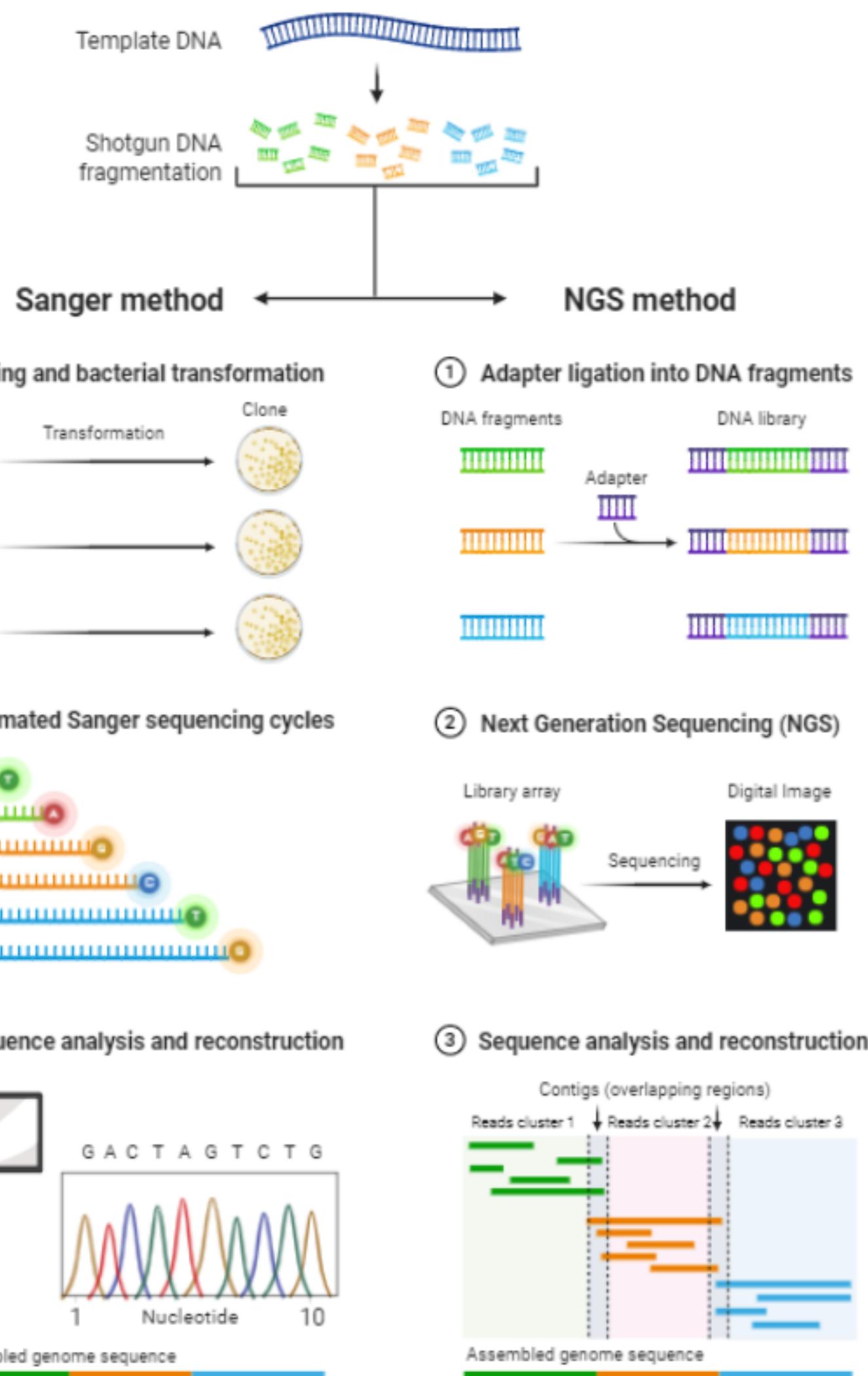


Genome Analysis

NGS vs. Sanger Sequencing



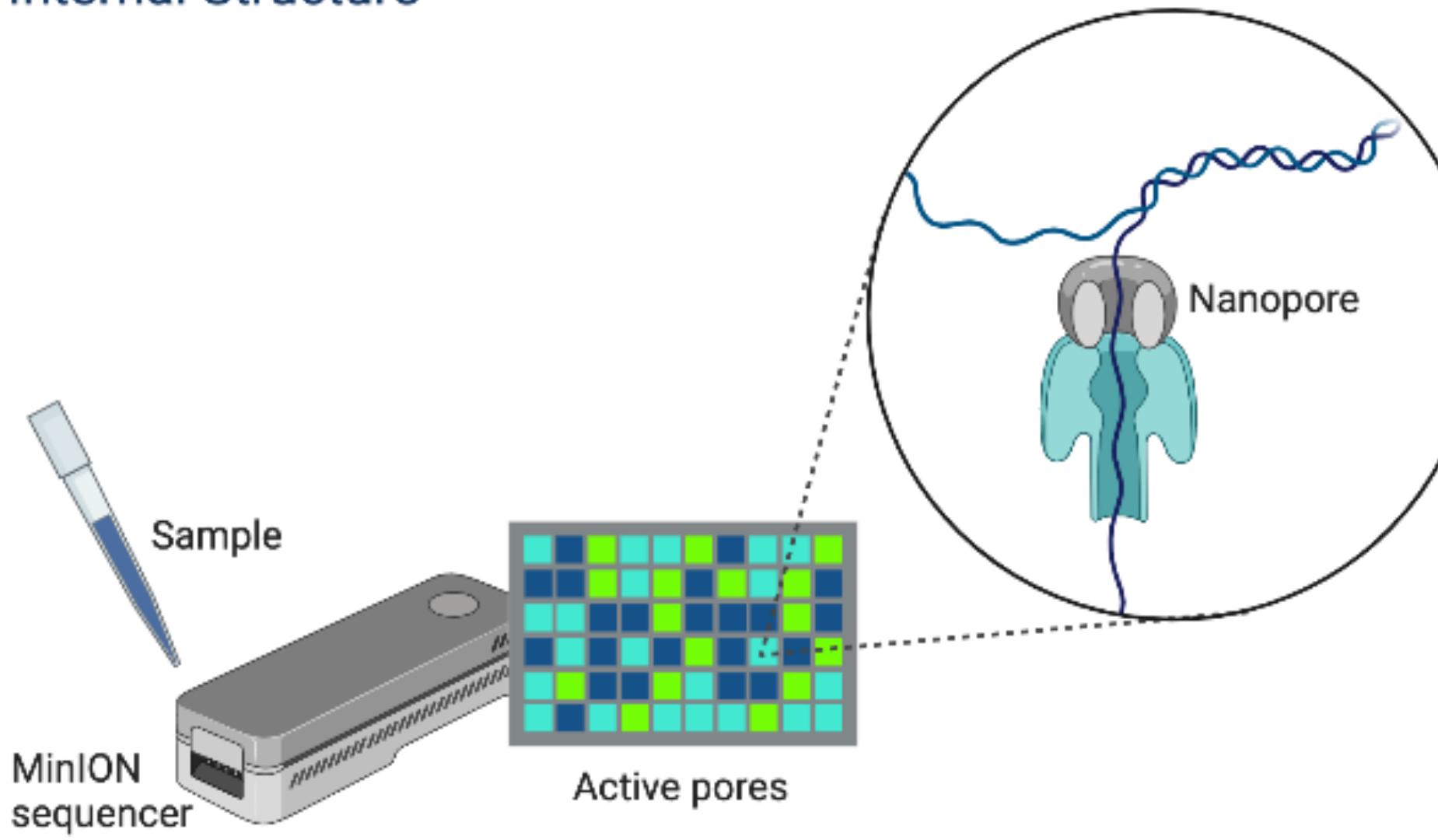
Shotgun Sequencing Sanger vs NGS



Genome Analysis

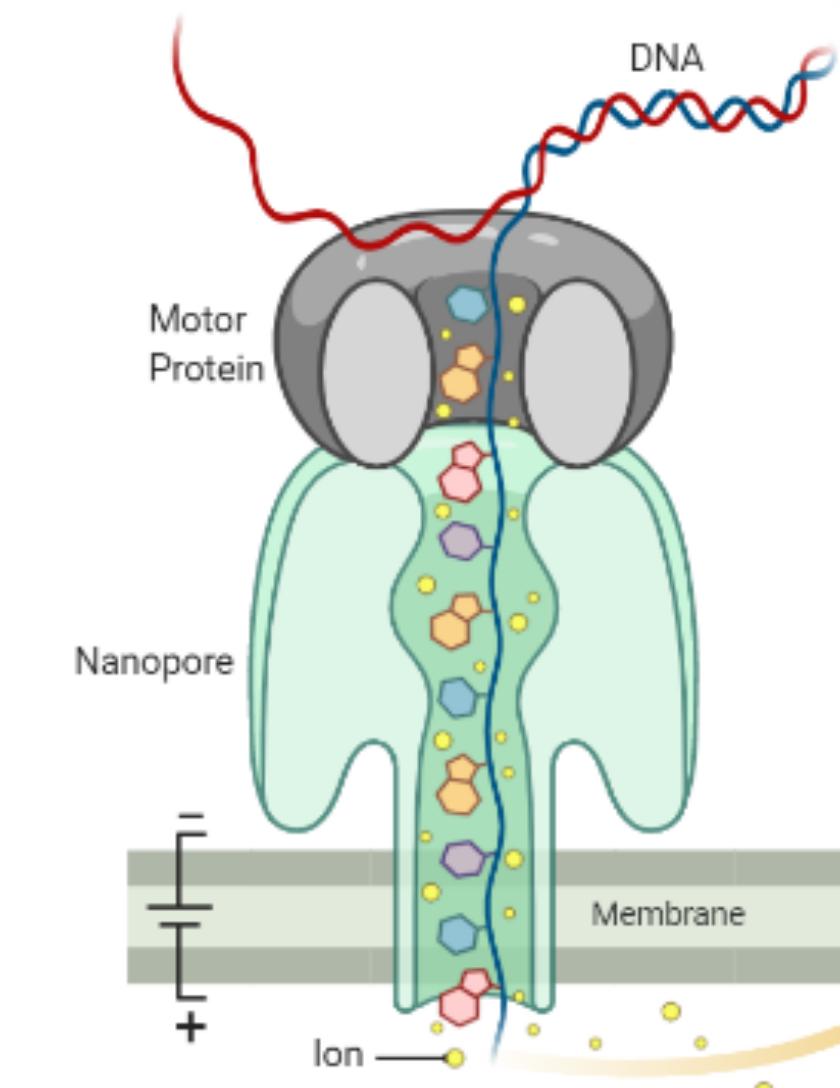
Nanopore Sequencing

MinION Sequencer Internal Structure

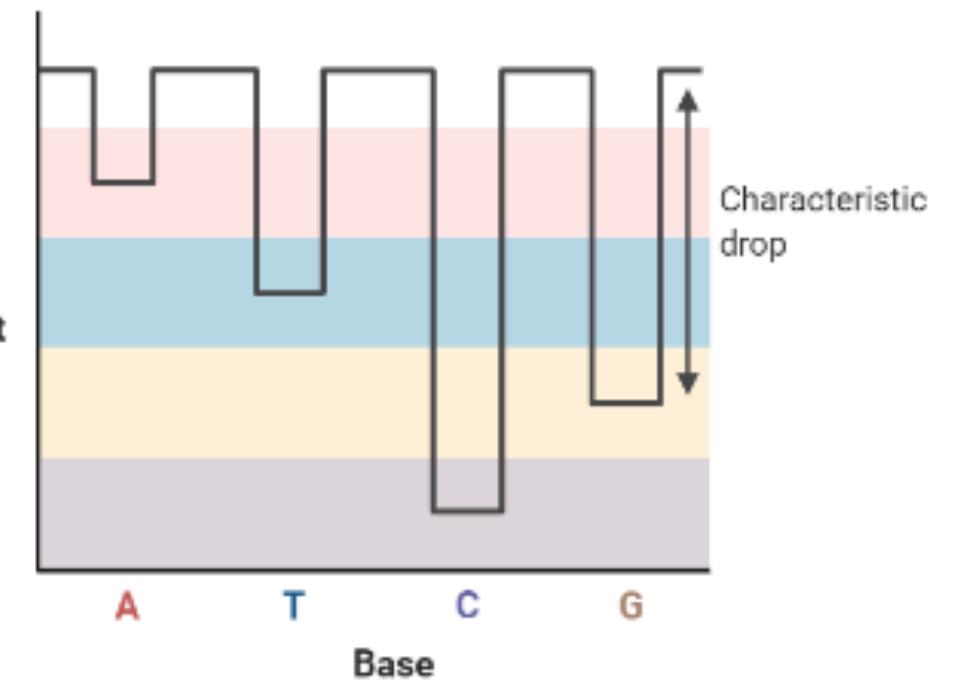


Nanopore Sequencing

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



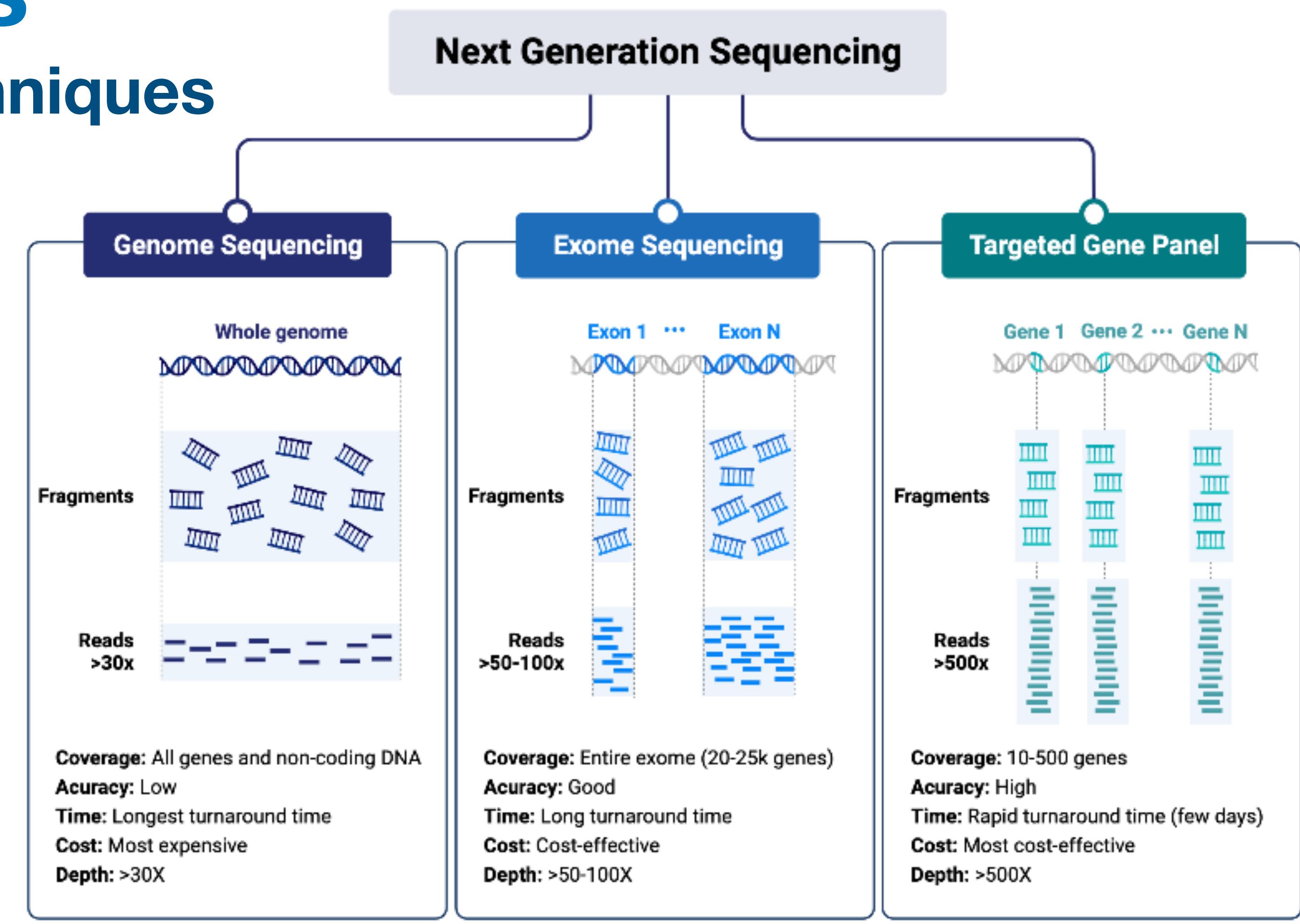
- 2 Each base gives a characteristic reduction in the ionic current, allowing the DNA to be sequenced



Genome Analysis

Comparison of NGS Techniques

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



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

Genomic File Formats

Types | Sizes | Use Cases

What is a PB, for human genomes?

It depends...

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



Genomic File Formats

- **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)

GSM1904006.CEL 69.1 MB
Modified: 3 February 2016 at 17:46

Add Tags...

General:

- Kind: FLC animation
- Size: 69'078'052 bytes (69.1 MB on disk)
- Where: arrayRAID • arraymapln • affyRaw • GSE73822 • GPL6801
- Created: 3 February 2016 at 17:46
- Modified: 3 February 2016 at 17:46
- Stationery pad
- Locked

More Info...
Name & Extension:
Comments:
Open with:
QuickTime Player (default)
Use this application to open all documents like this one.
Change All...
Preview:

not a movie...

itemRgb="On"

genome.ucsc.edu/FAQ/FAQformat.html

Axt format
BAM format
BED format
BED detail format
bedGraph format
barChart and bigBarChart format
bigBed format
bigGenePred table format
bigPsl table format
bigMaf table format
bigChain table format
bigWig format
Chain format

CRAM format
GenePred table format
GFF format
GTF format
HAL format
MAF format
Microarray format
Net format
Personal Genome SNP format
PSL format
VCF format
WIG format

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

```

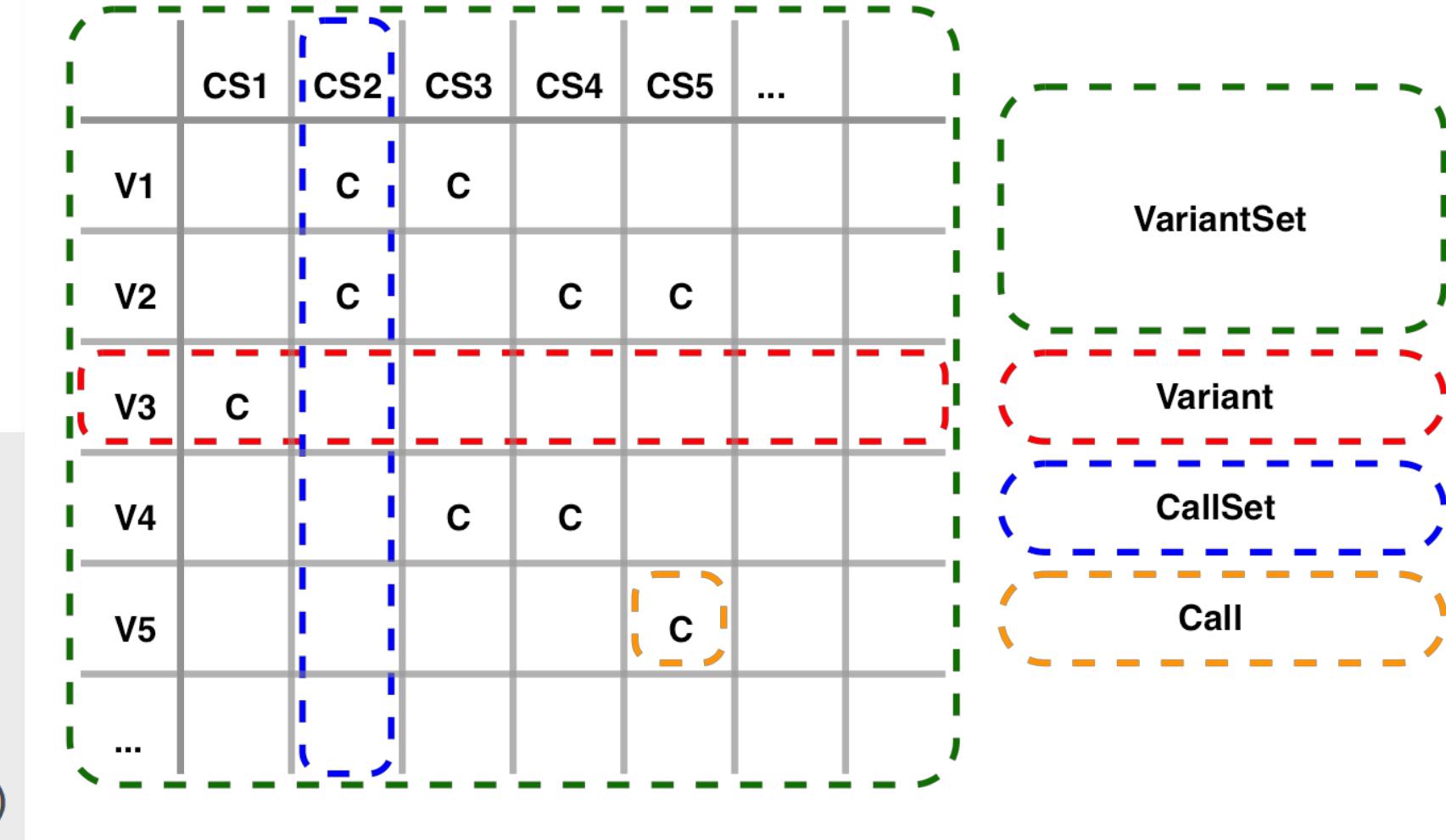
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype 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">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1 SAMPLE2
1 1 . ACG A,AT .
1 2 rs1 C T,CT .
1 5 . G <DEL> .
1 100 . T .

```

Body

Annotations:

- Deletion**: Variant at position 5 is a deletion ().
- SNP**: Variant at position 2 is a SNP (rs1).
- Large SV**: Variant at position 5 is a large structural variant (SVTYPE=DEL; END=300).
- Insertion**: Variant at position 2 has an insertion (T).
- Other event**: Variant at position 5 is an other event ().
- Mandatory header lines**: Lines starting with ##.
- Optional header lines (meta-data)**: Lines starting with ##INFO or ##FORMAT.
- Reference alleles (GT=0)**: Reference alleles for the variants.
- Alternate alleles (GT>0 is an index to the ALT column)**: Alternate alleles for the variants.
- Phased data**: Phased data for the variants across samples (SAMPLE1 and SAMPLE2).



Variant
Call
Format

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

Task: Estimate Storage Requirements for 1000 Genomes

How much computer storage is required for 1000 Genomes

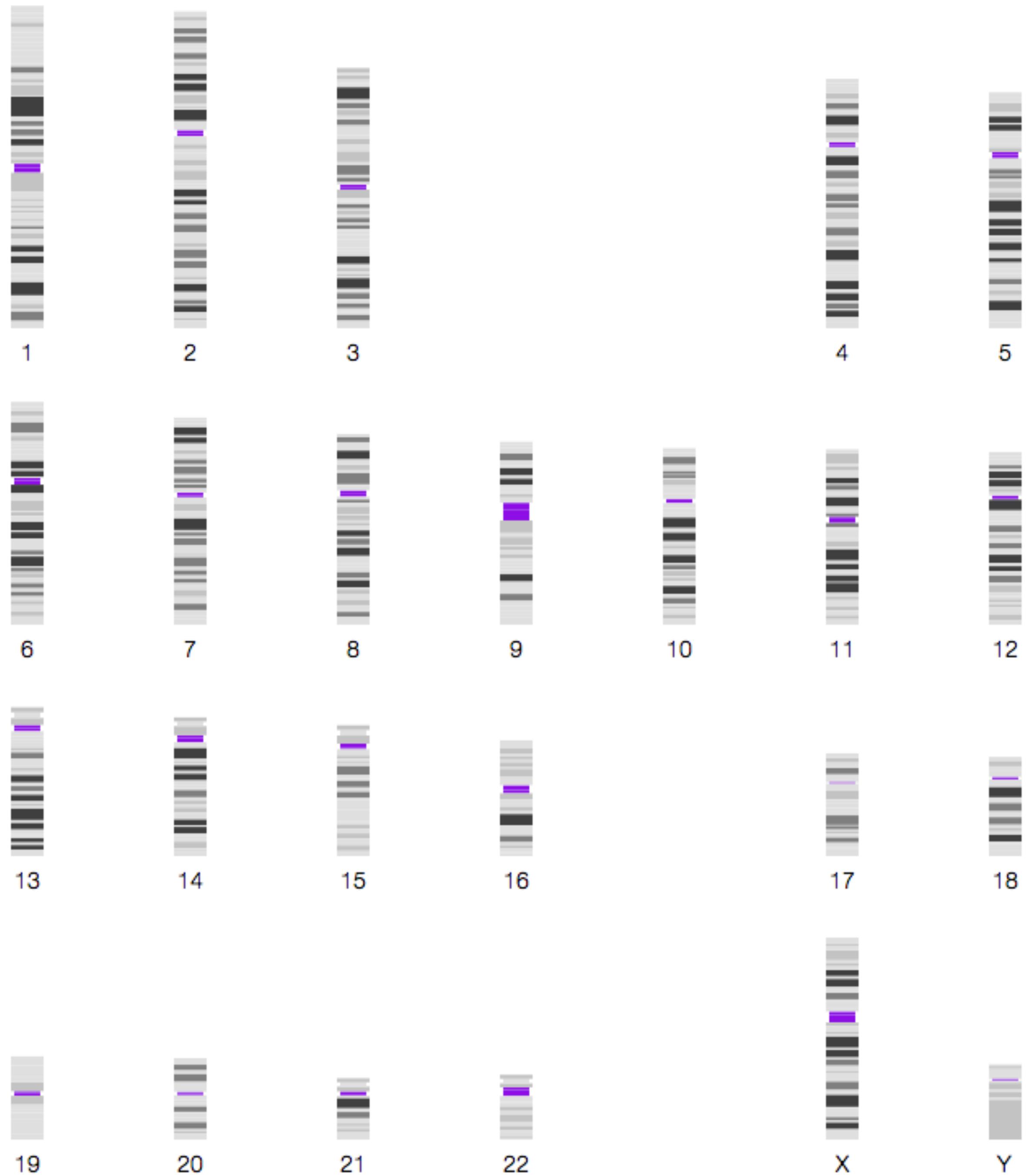
- WES & WGS
- Different file formats
 - SAM
 - BAM
 - VCF
 - FASTA
- Associated costs
 - Cost factors
 - Raw Storage costs
- Familiarize with VCF format
 - ➡ specification in article collection



IBM-storage-unit-3500-Schiphol-1957

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

SOFTWARE TOOL ARTICLE

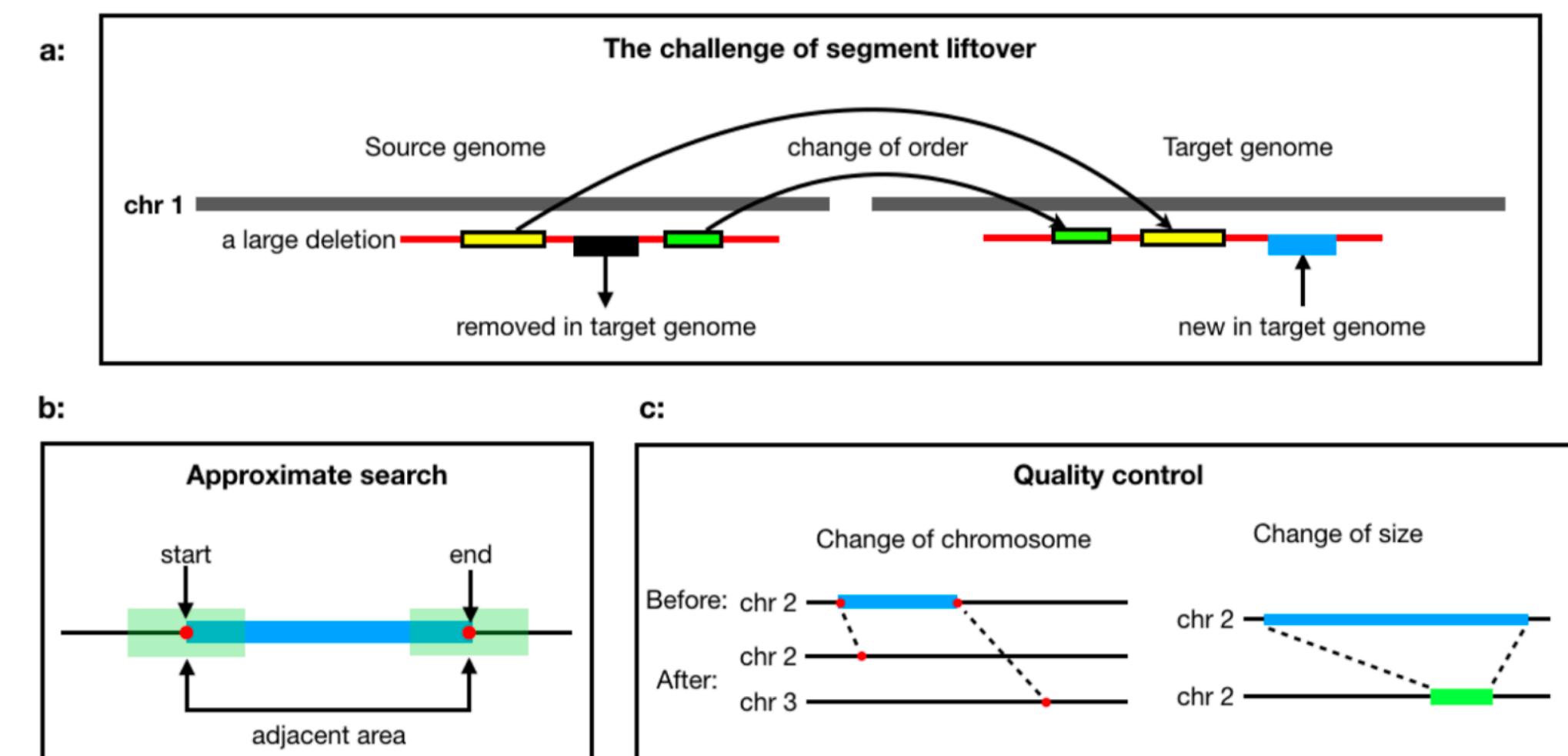
REVISED **segment_liftover** : a Python tool to convert segments between genome assemblies [version 2; referees: 2 approved]

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

¹Institute of molecular Life Sciences, University of Zürich, Zürich, CH-8057, Switzerland

²Swiss Institute of Bioinformatics, University of Zürich, Zürich, CH-8057, Switzerland

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



Task: Reading up on Genome Technologies

- General NGS technologies
- count based vs. intensity based as principle
- bonus: dig deeper for some molecular-cytogenetic techniques:
 - SNP, aCGH arrays
 - SKY, M-FISH
 - chromosomal CGH