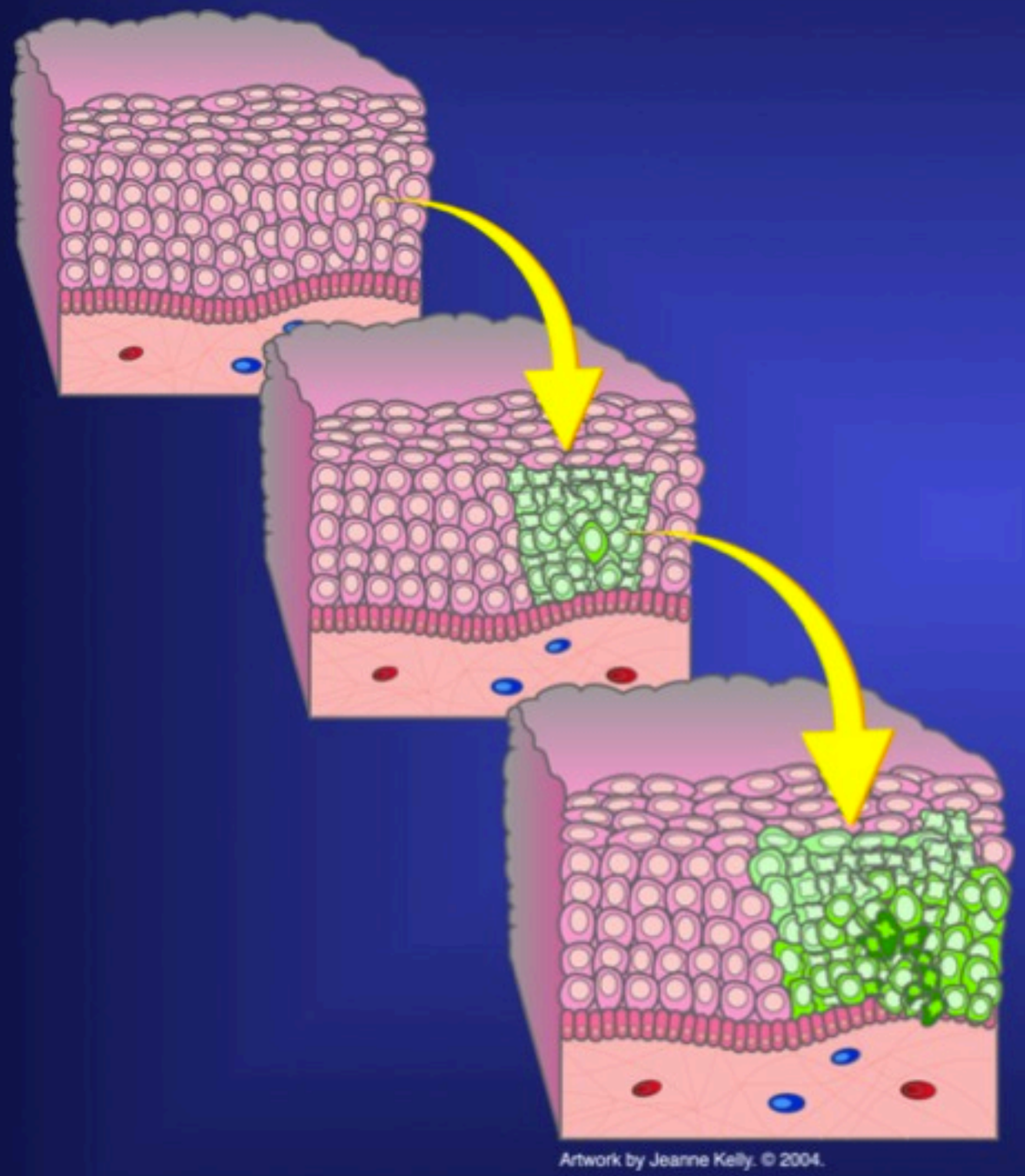
Introduction to cancer genetics and genomics

Nuria Lopez-Bigas

- Molecular bases of cancer
- Cancer alterations and technologies to detect them genome-wide
- Predictive cancer genomics
- NGS technologies in cancer research

Cancer-Associated Mutations

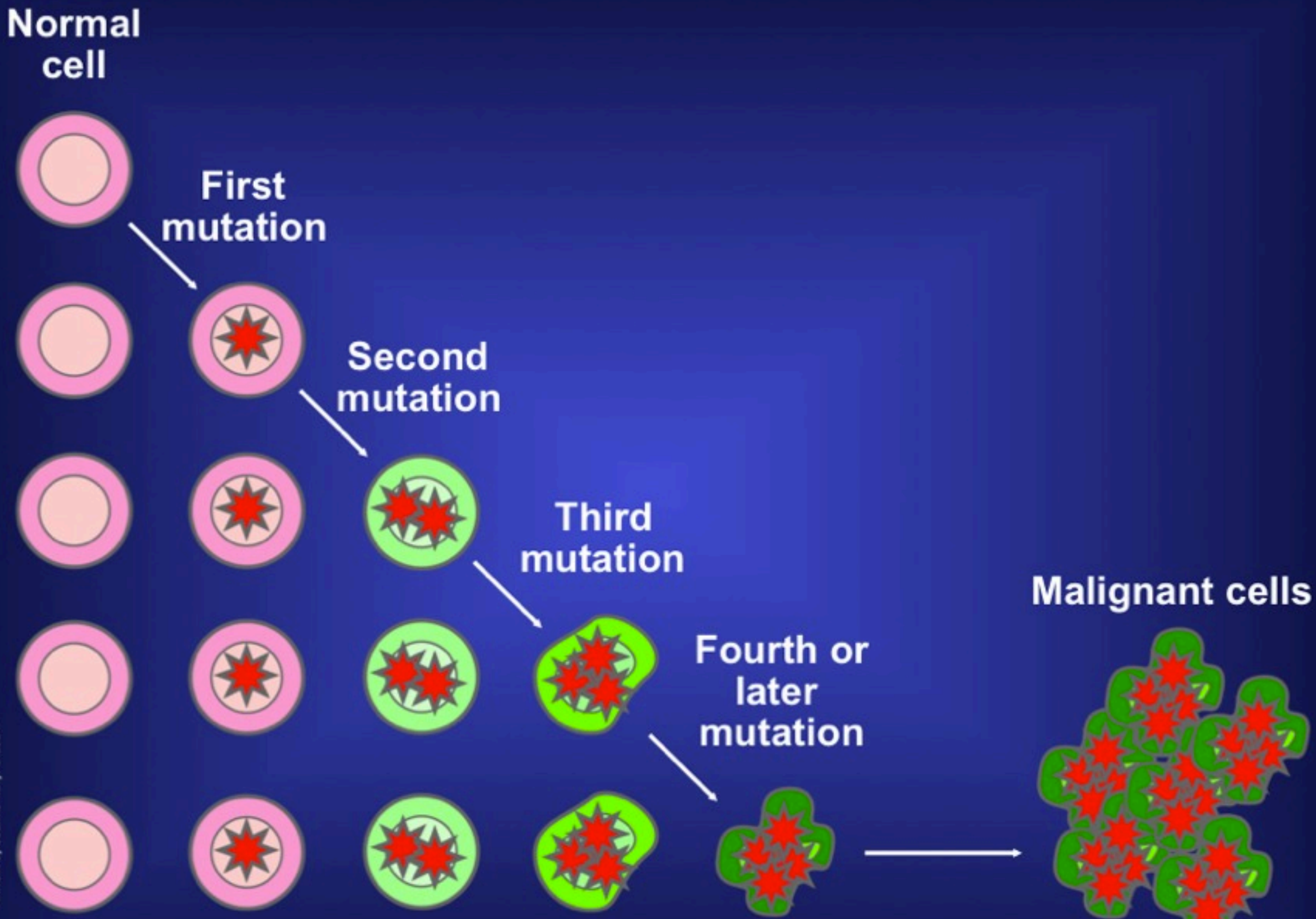


Artwork by Jeanne Kelly. © 2004.

- Oncogenes
- Tumor suppressor genes
- DNA repair genes
- Carcinogen
  - activating genes
  - deactivating genes
- Cell cycle genes
- Cell cycle checkpoint genes
- Cell death genes
- Cell signaling genes
- Cellular differentiation genes
- Cellular senescence genes
- Metastasis/invasion genes

NATIONAL CANCER INSTITUTE

Tumors Are Clonal

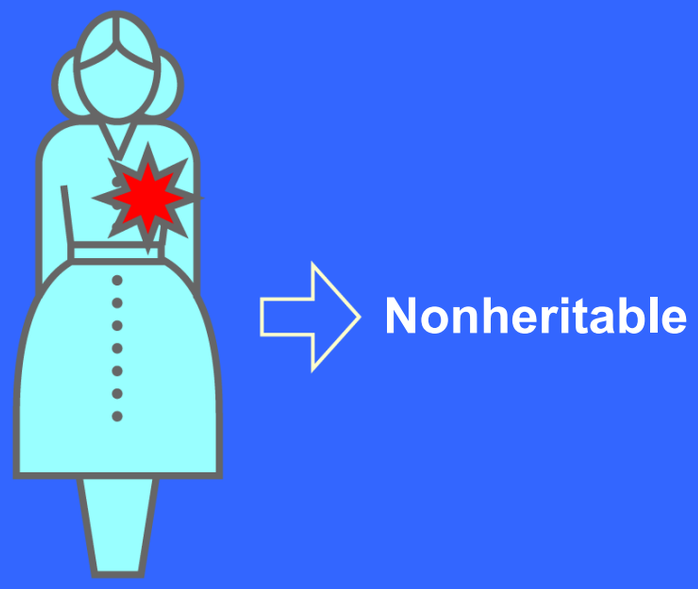


NATIONAL CANCER INSTITUTE

Mutations: Somatic and Germline

Somatic mutations

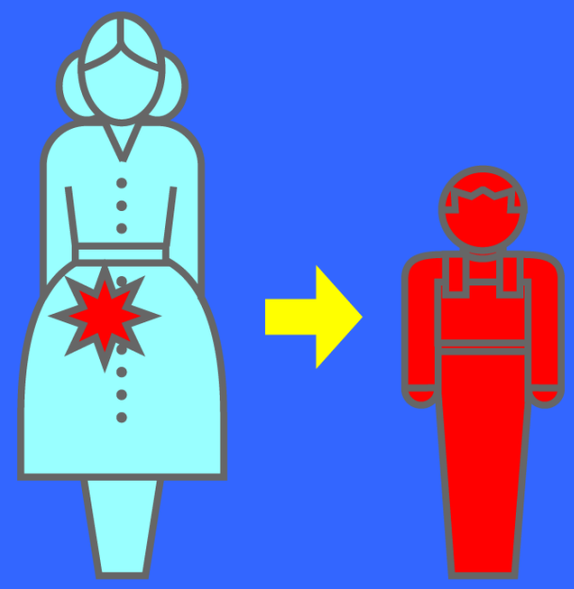
- Occur in nongermline tissues
- Are nonheritable



Somatic mutation
(e.g., breast)

Germline mutations

- Present in egg or sperm
- Are heritable
- Cause cancer family syndrome

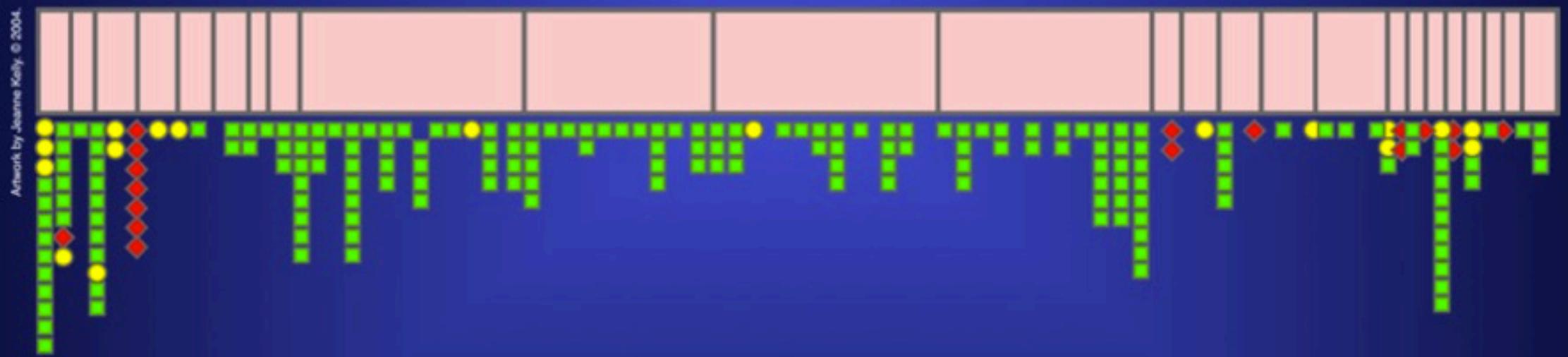


Mutation in
egg or sperm

All cells
affected in
offspring

Mutations in Cancer Susceptibility Genes: *BRCA1*

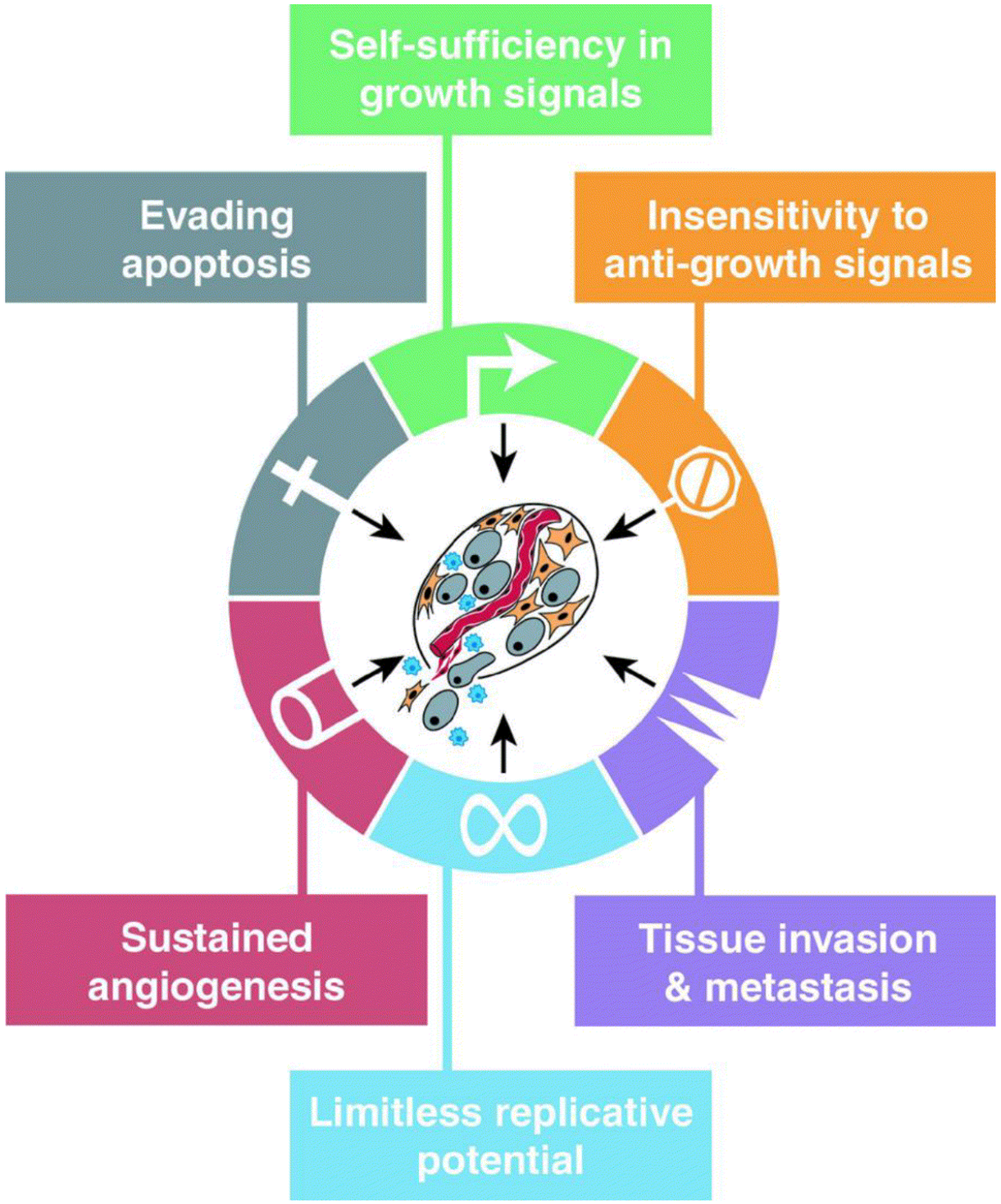
- On chromosome 17
- Autosomal dominant transmission
- Protein has role in genomic stability
- ~500 different mutations reported



- Nonsense/Frameshift
- Missense
- Splice-site

NATIONAL CANCER INSTITUTE

Hallmarks of Cancer



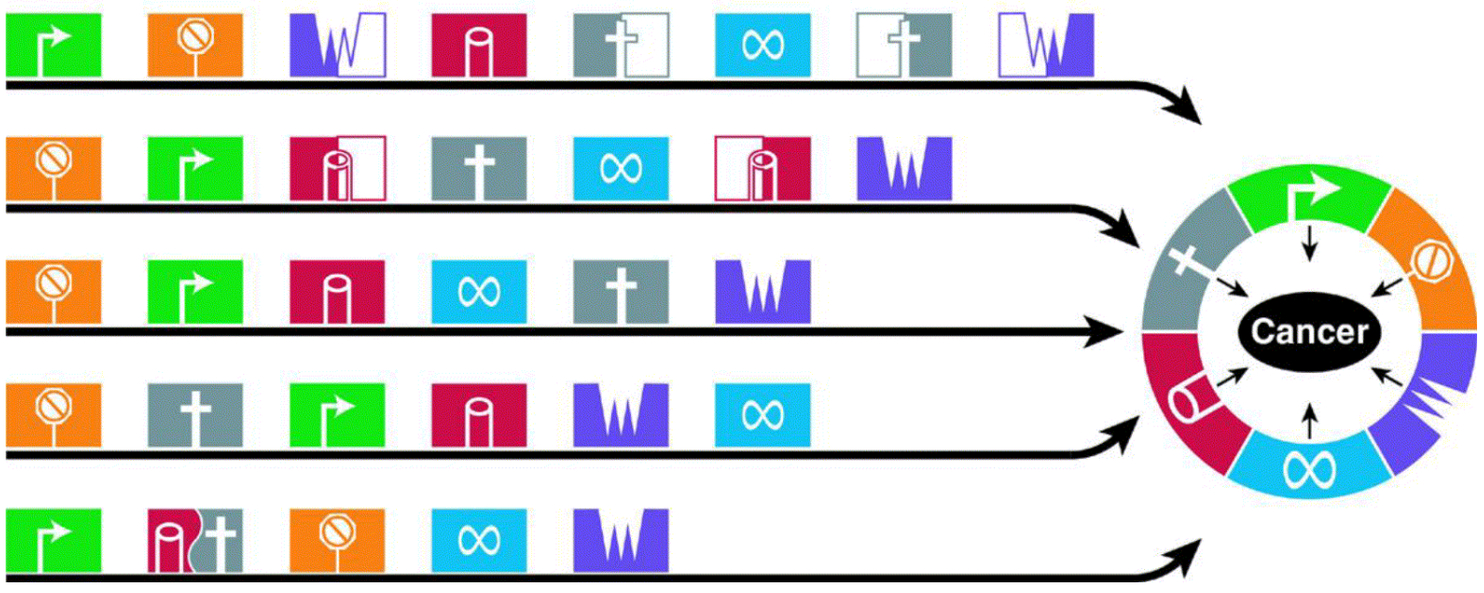
Hanahan & Weinberg. Cell. 2000 Jan 7;100(1):57-70

Hallmarks of Cancer

A

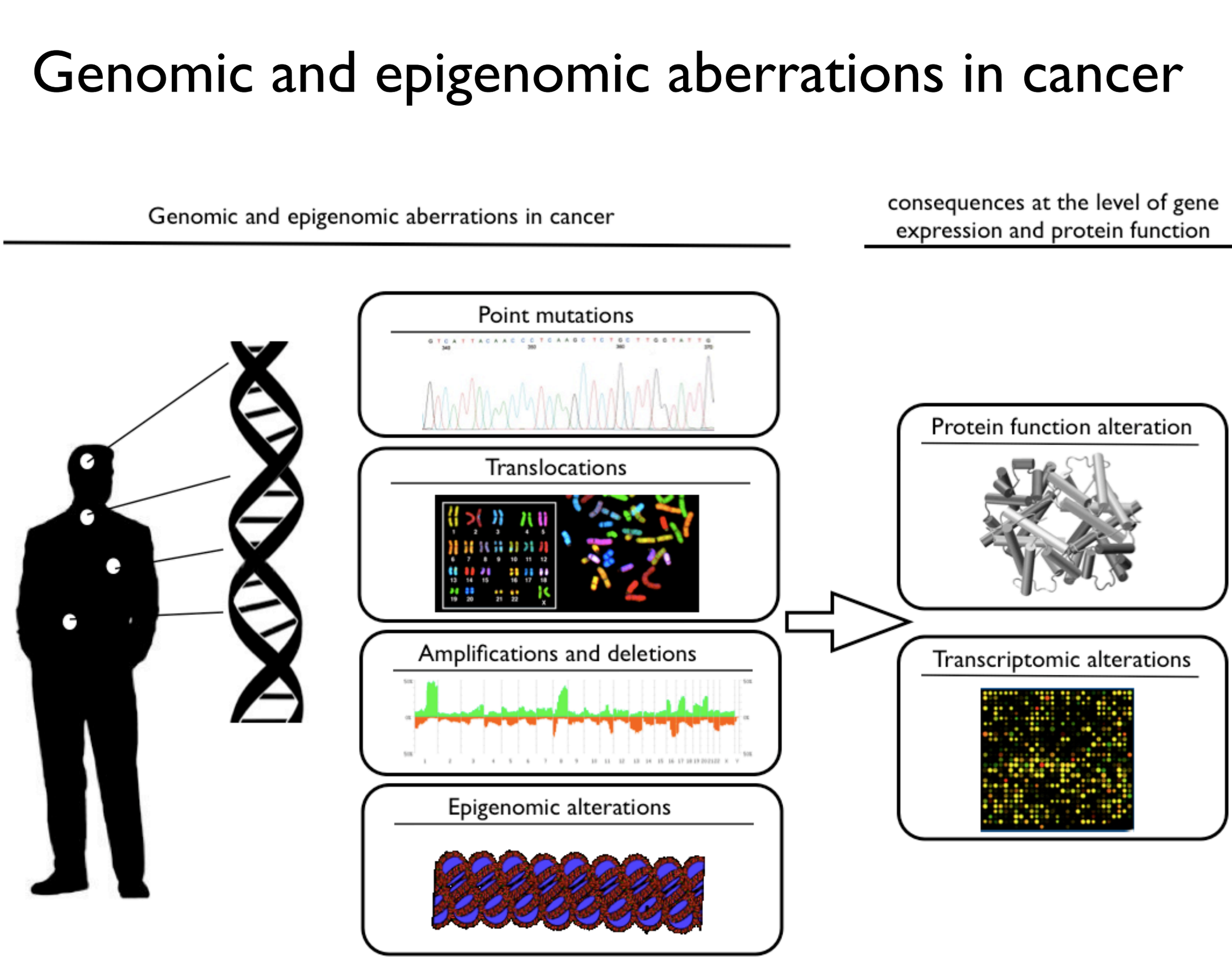
Component	Acquired Capability	Example of Mechanism
	Self-sufficiency in growth signals	Activate H-Ras oncogene
	Insensitivity to anti-growth signals	Lose retinoblastoma suppressor
	Evading apoptosis	Produce IGF survival factors
	Limitless replicative potential	Turn on telomerase
	Sustained angiogenesis	Produce VEGF inducer
	Tissue invasion & metastasis	Inactivate E-cadherin

B

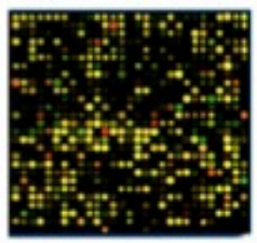
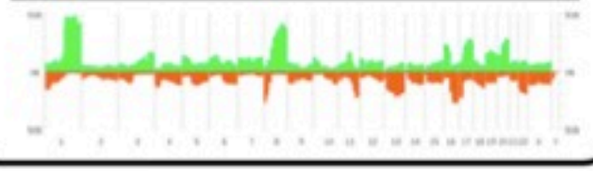
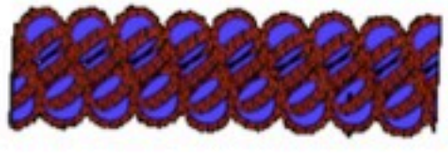


Hanahan & Weinberg. Cell. 2000 Jan 7;100(1):57-70

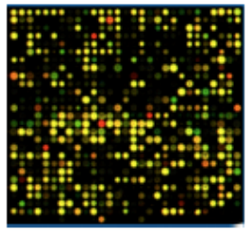
Genomic and epigenomic aberrations in cancer



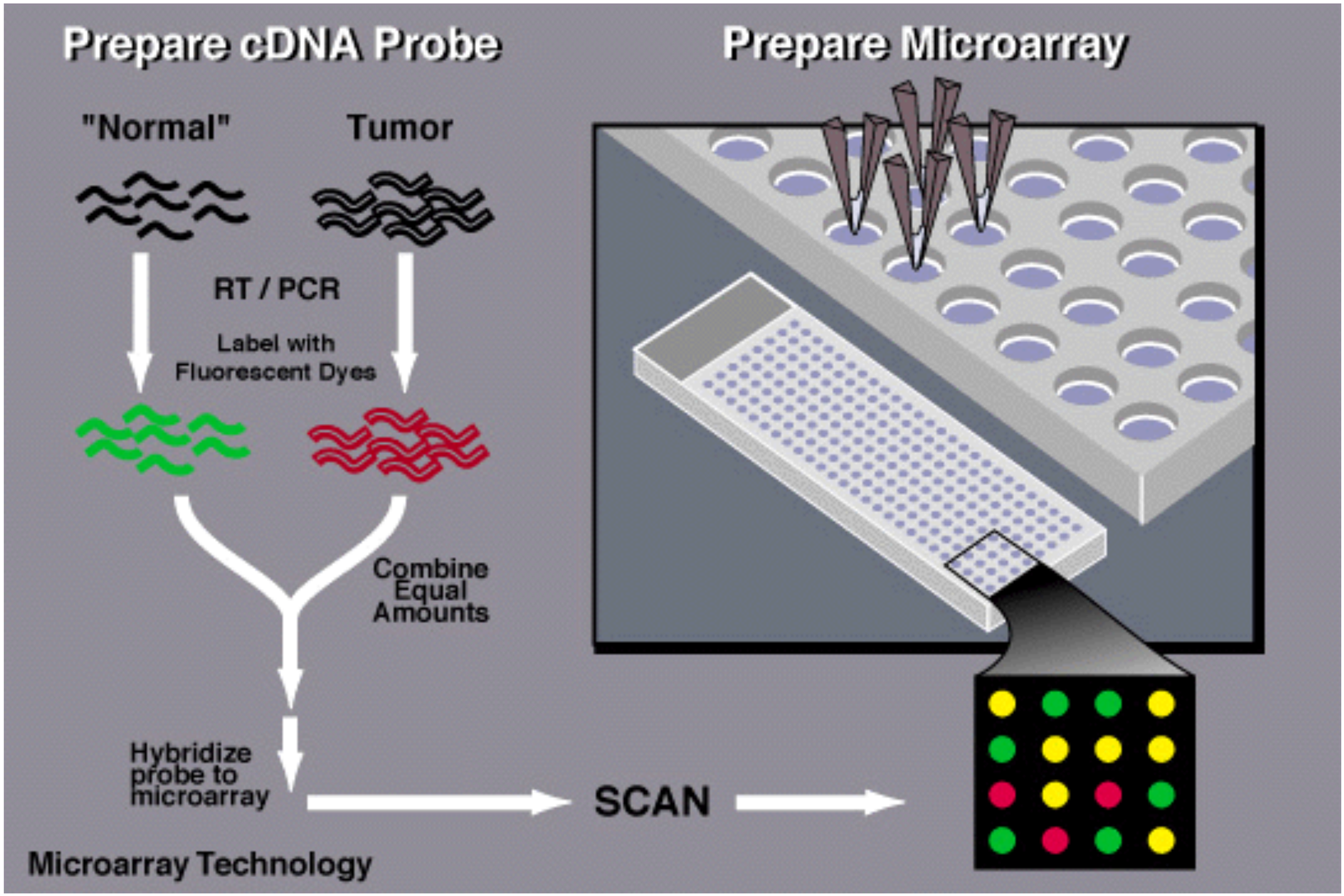
High-throughput technologies to identify cancer alterations

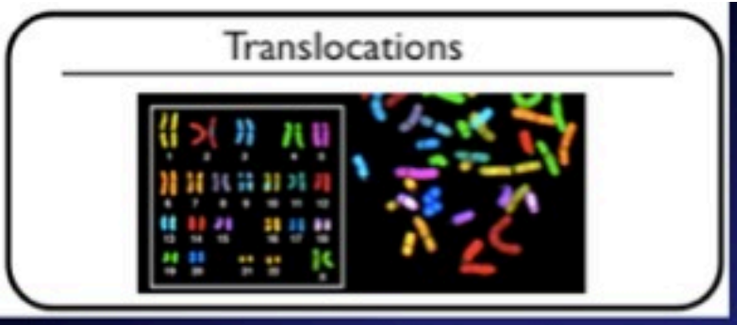
Type of alteration	High-throughput tecnology
Transcriptomic alterations 	Expression microarrays
Point mutations 	Sequencing
Translocations 	SKY, FISH
Amplifications and deletions 	Comparative Genomics Hybridization
Epigenomic alterations 	Methylation profiling

Transcriptomic alterations



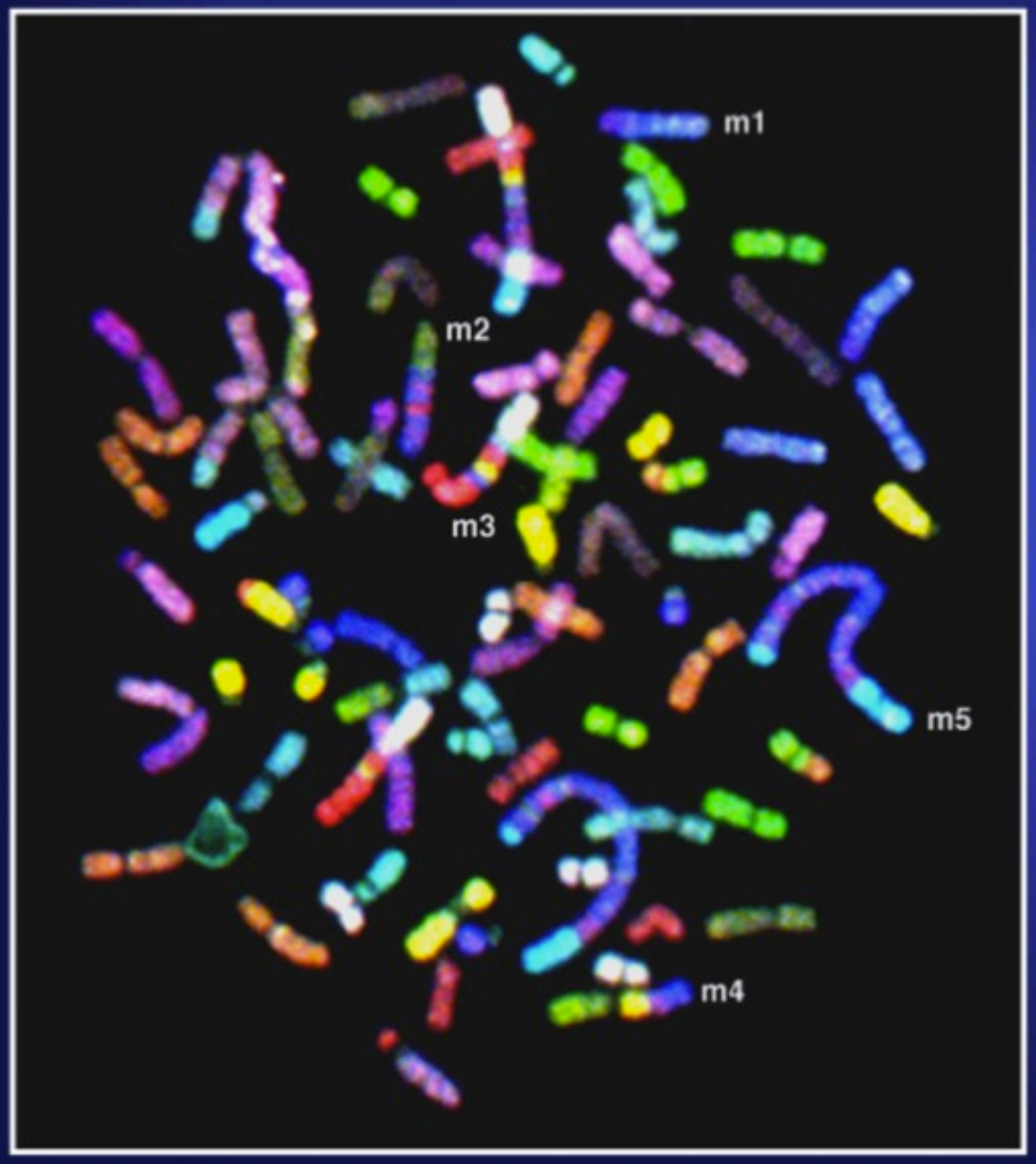
Expression microarrays





Large Deletions or Insertions

SKY chromosome painting: breast cancer



Artwork by Jeanne Kelly. © 2004.

Normal SKY chromosomes are not multicolored.

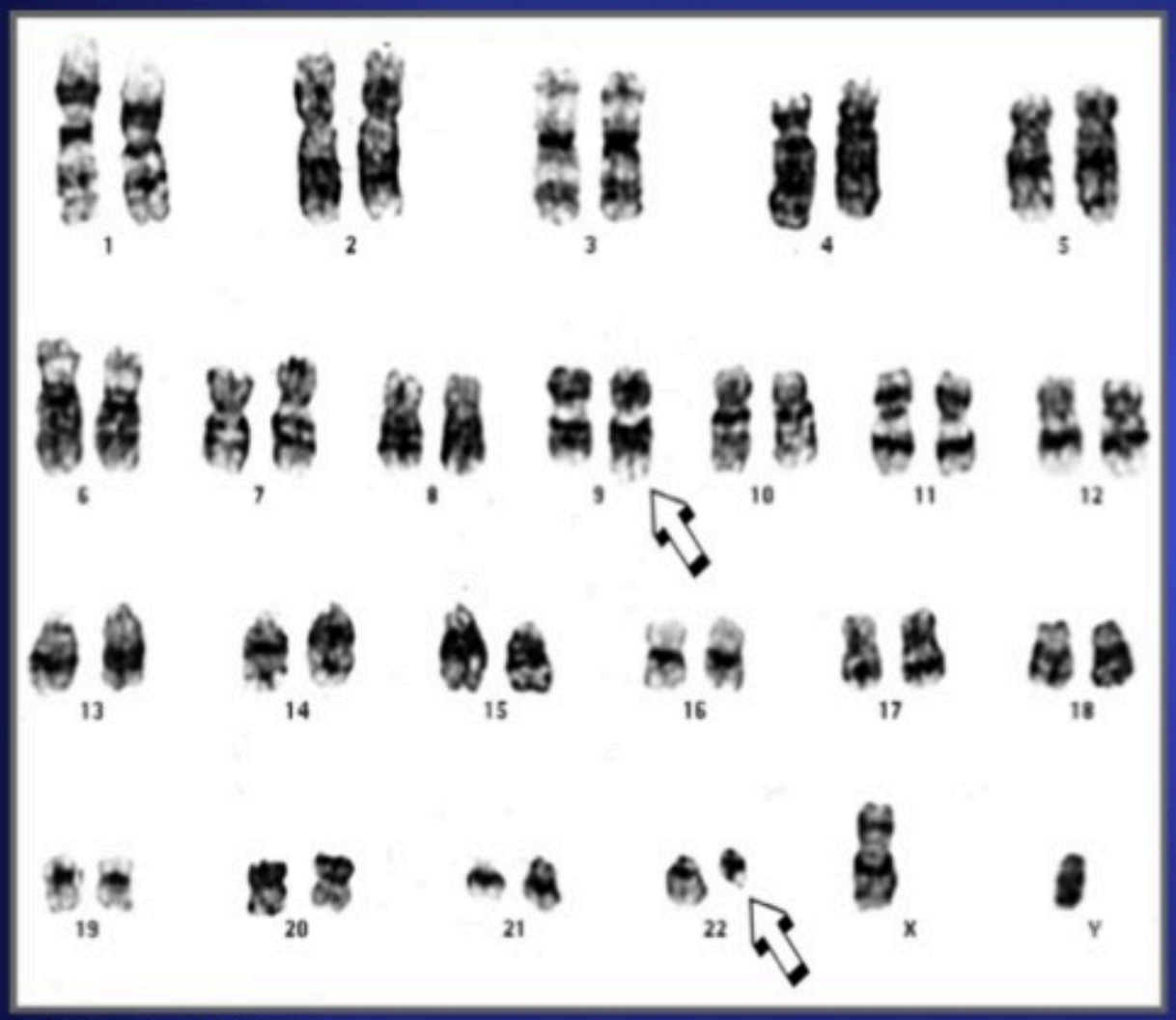
Chromosomes in breast cancer appear multicolored because they have exchanged genetic material.

NATIONAL CANCER INSTITUTE

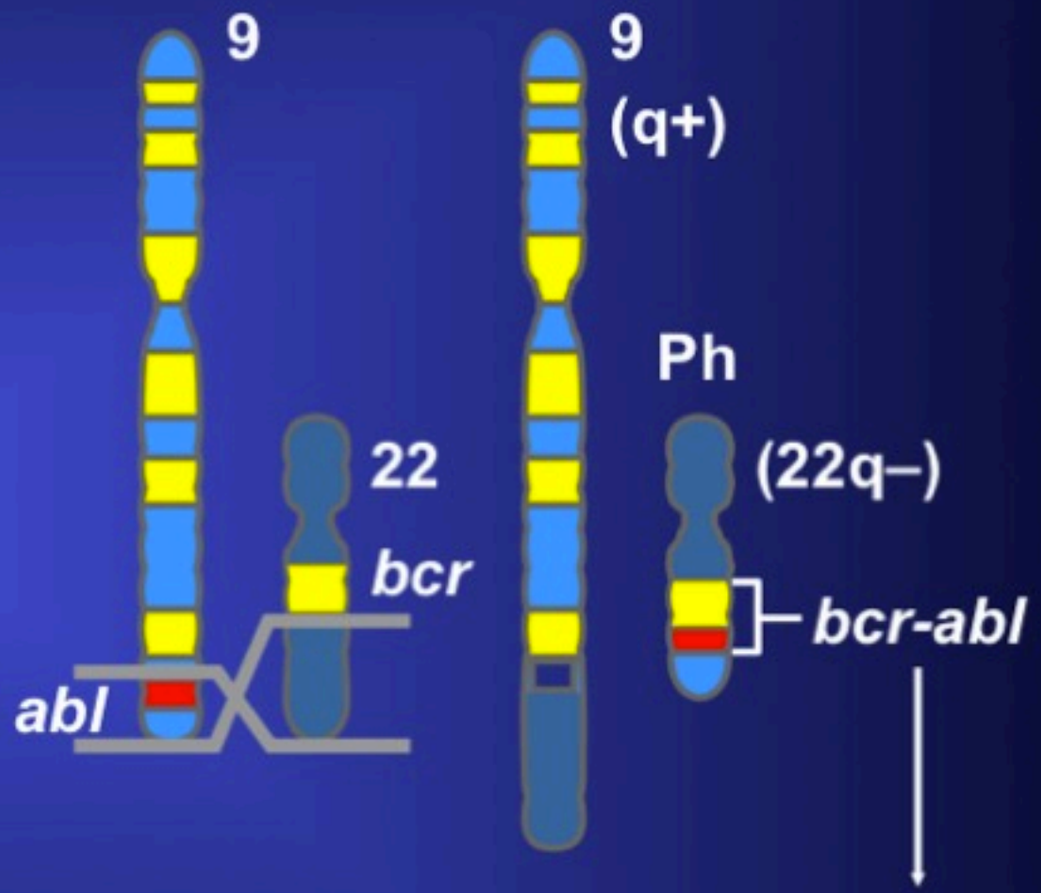
Translocations

Philadelphia Chromosome

chronic myelogenous leukemia



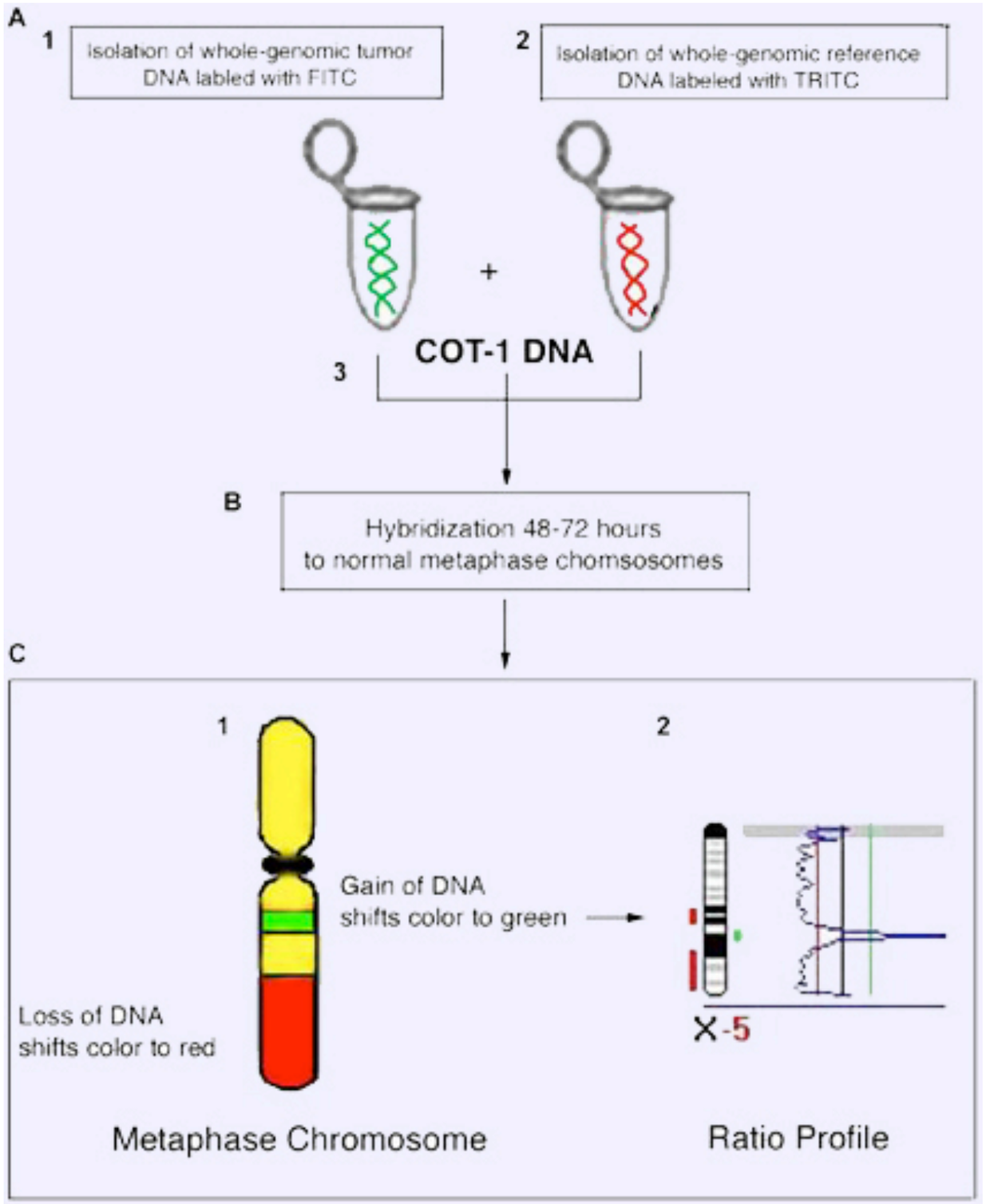
Artwork by Jeanne Kelly. © 2004.



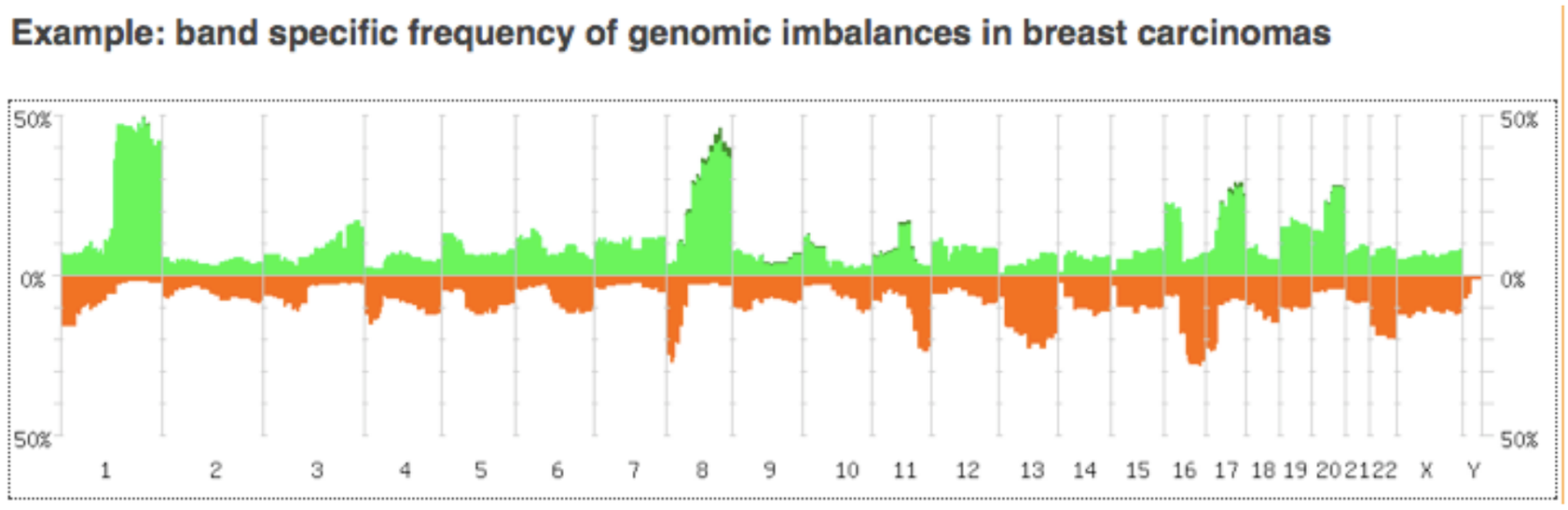
Fusion protein with tyrosine kinase activity

NATIONAL CANCER INSTITUTE

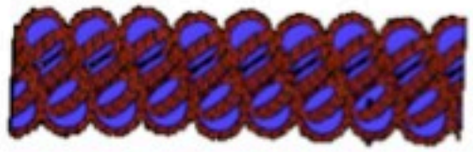
Comparative Genomics Hybridization (CGH)



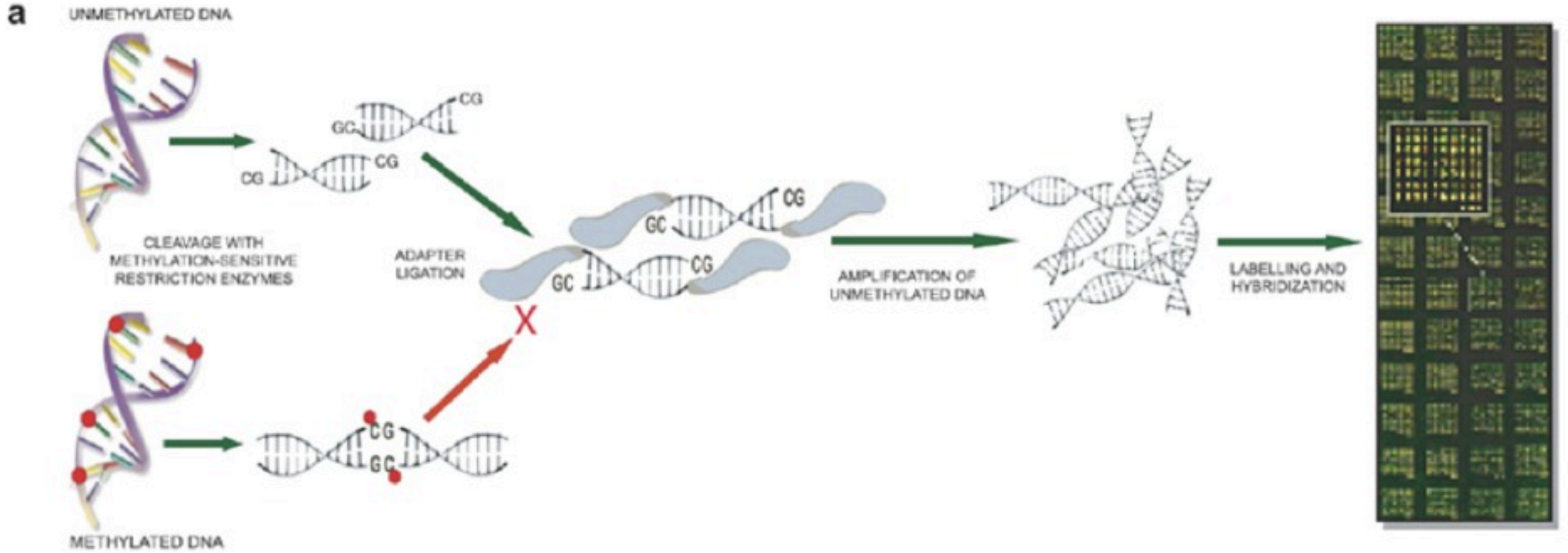
Comparative Genomics Hybridization (CGH)

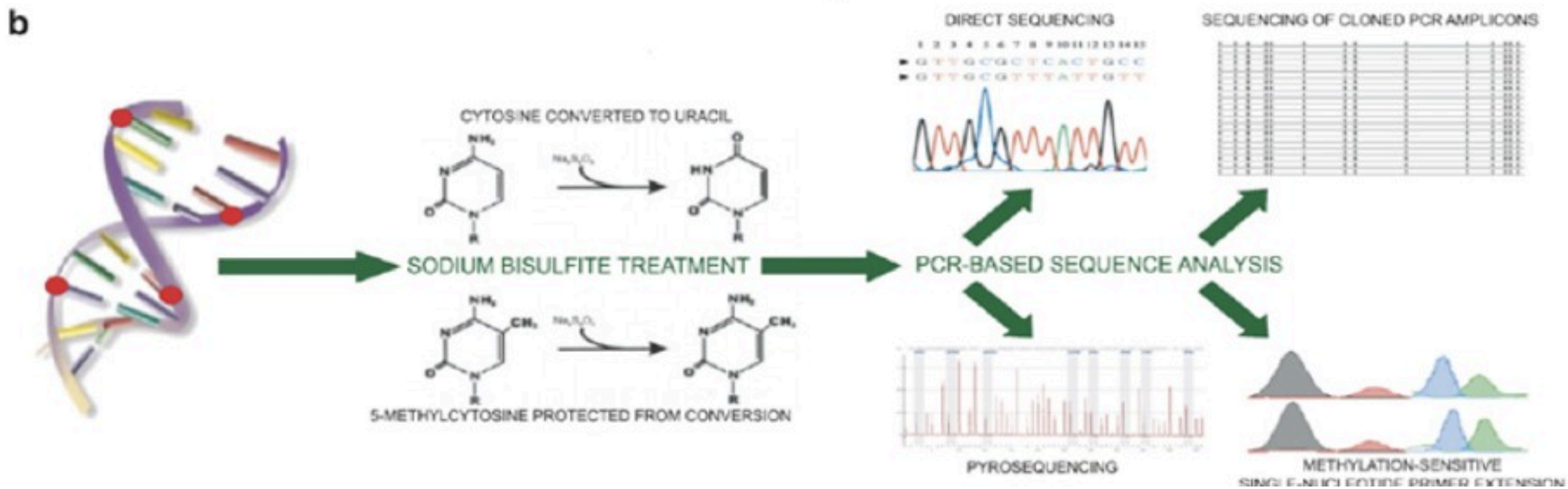


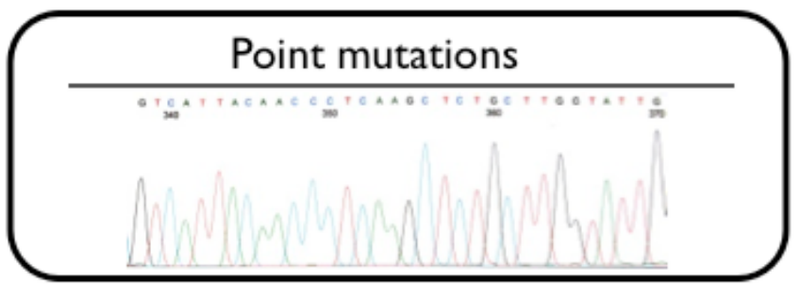
Epigenomic alterations



Methylation Profiling







DNA Sequencing

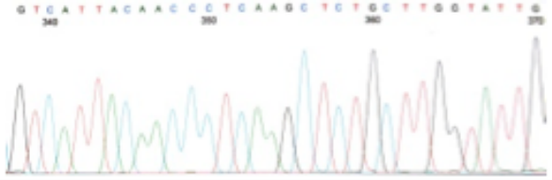
The Consensus Coding Sequences of Human Breast and Colorectal Cancers

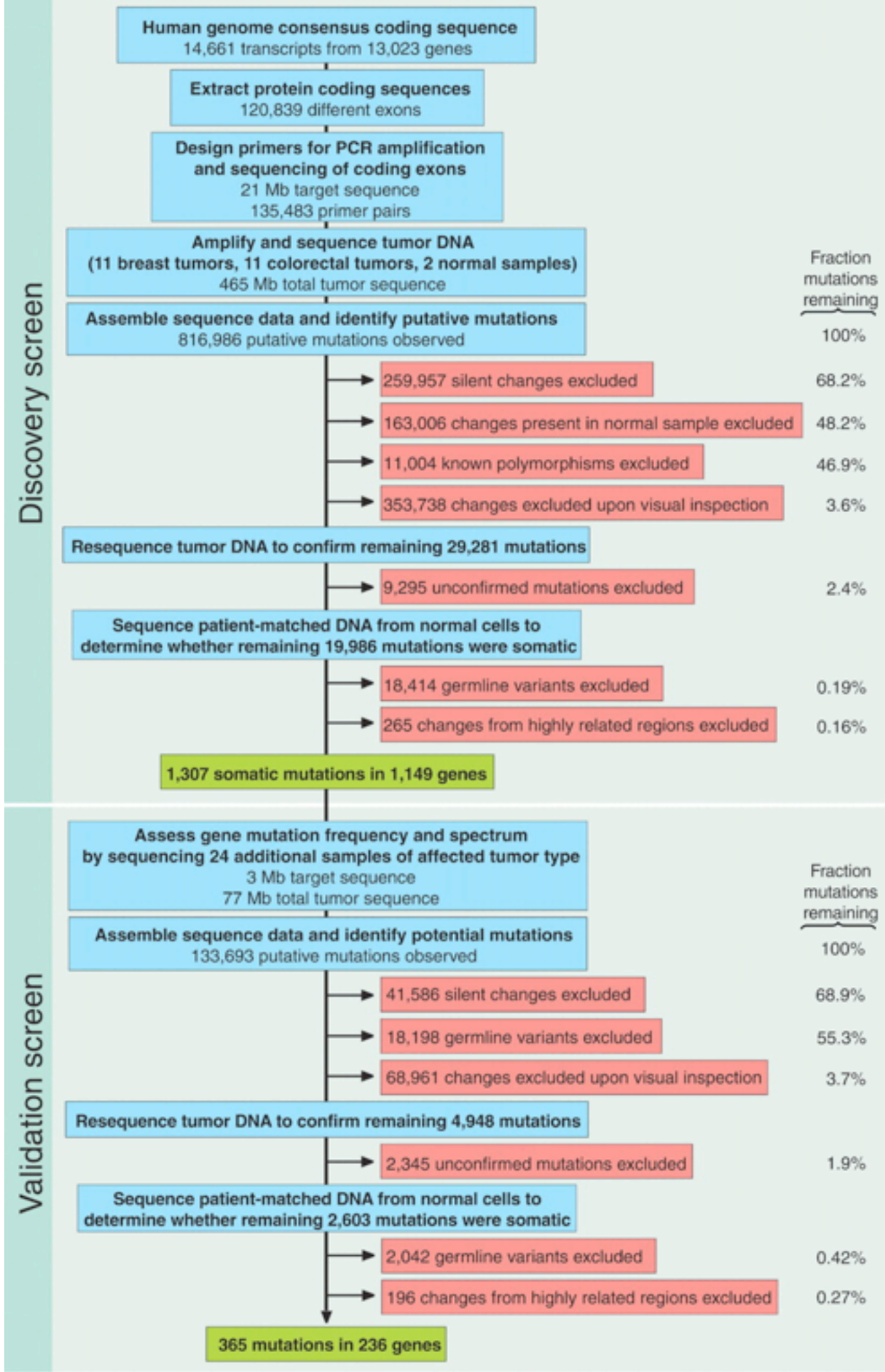
Tobias Sjöblom,[1*] Siân Jones,[1*] Laura D. Wood,[1*] D. Williams Parsons,[1*] Jimmy Lin,[1] Thomas D. Barber,[1†] Diana Mandelker,[1] Rebecca J. Leary,[1] Janine Ptak,[1] Natalie Silliman,[1] Steve Szabo,[1] Phillip Buckhaults,[2] Christopher Farrell,[2] Paul Meeh,[2] Sanford D. Markowitz,[3] Joseph Willis,[4] Dawn Dawson,[4] James K. V. Willson,[5] Adi F. Gazdar,[6] James Hartigan,[7] Leo Wu,[8] Changsheng Liu,[8] Giovanni Parmigiani,[9] Ben Ho Park,[10] Kurtis E. Bachman,[11] Nickolas Papadopoulos,[1] Bert Vogelstein,[1‡] Kenneth W. Kinzler,[1‡] Victor E. Velculescu[1‡]

The elucidation of the human genome sequence has made it possible to identify genetic alterations in cancers in unprecedented detail. To begin a systematic analysis of such alterations, we determined the sequence of well-annotated human protein-coding genes in two common tumor types. Analysis of 13,023 genes in 11 breast and 11 colorectal cancers revealed that individual tumors accumulate an average of ~90 mutant genes but that only a subset of these contribute to the neoplastic process. Using stringent criteria to delineate this subset, we identified 189 genes (average of 11 per tumor) that were mutated at significant frequency. The vast majority of these genes were not known to be genetically altered in tumors and are predicted to affect a wide range of cellular functions, including transcription, adhesion, and invasion. These data define the genetic landscape of two human cancer types, provide new targets for diagnostic and therapeutic intervention, and open fertile avenues for basic research in tumor biology.

Science 13 October 2006

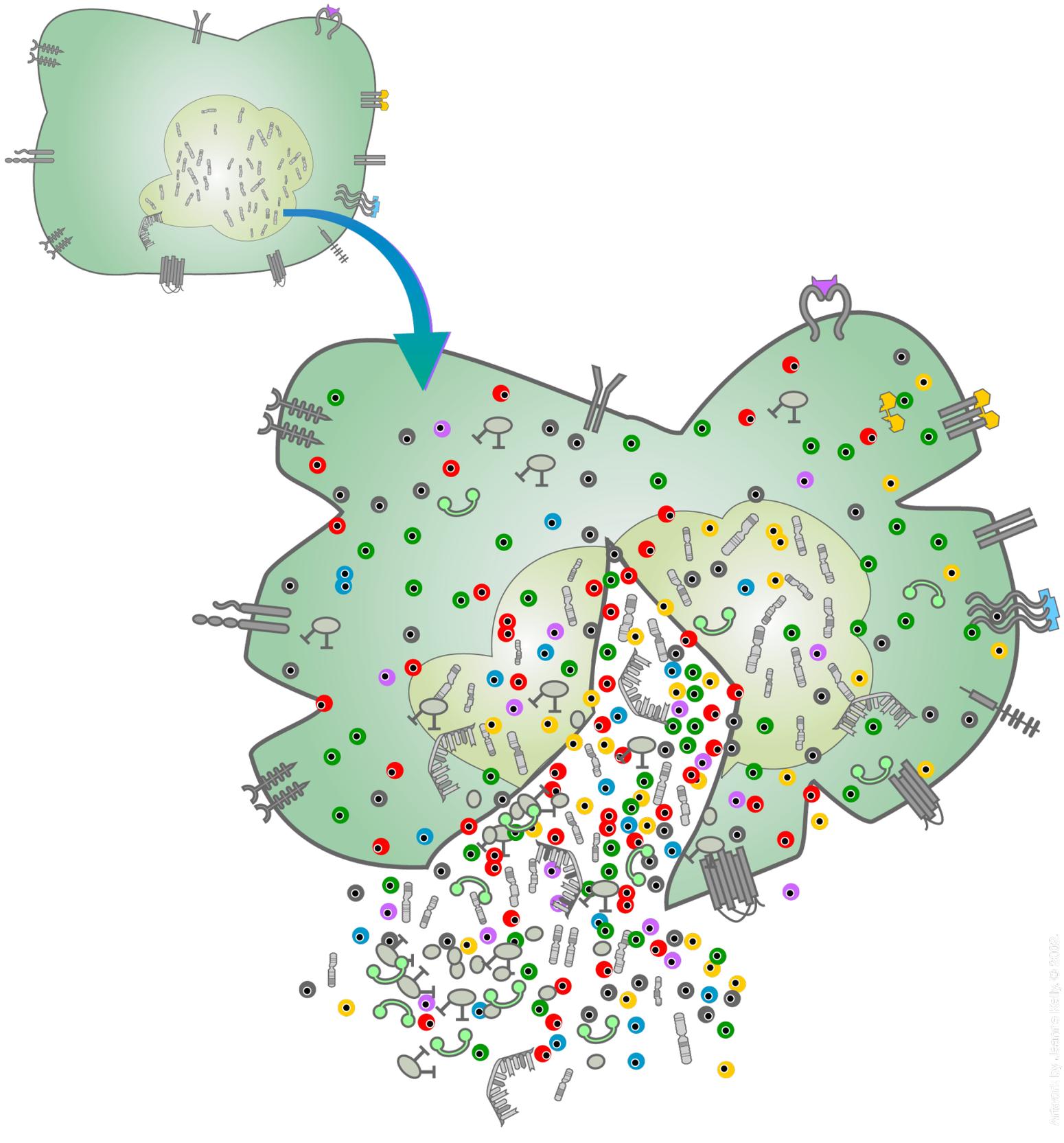
Point mutations



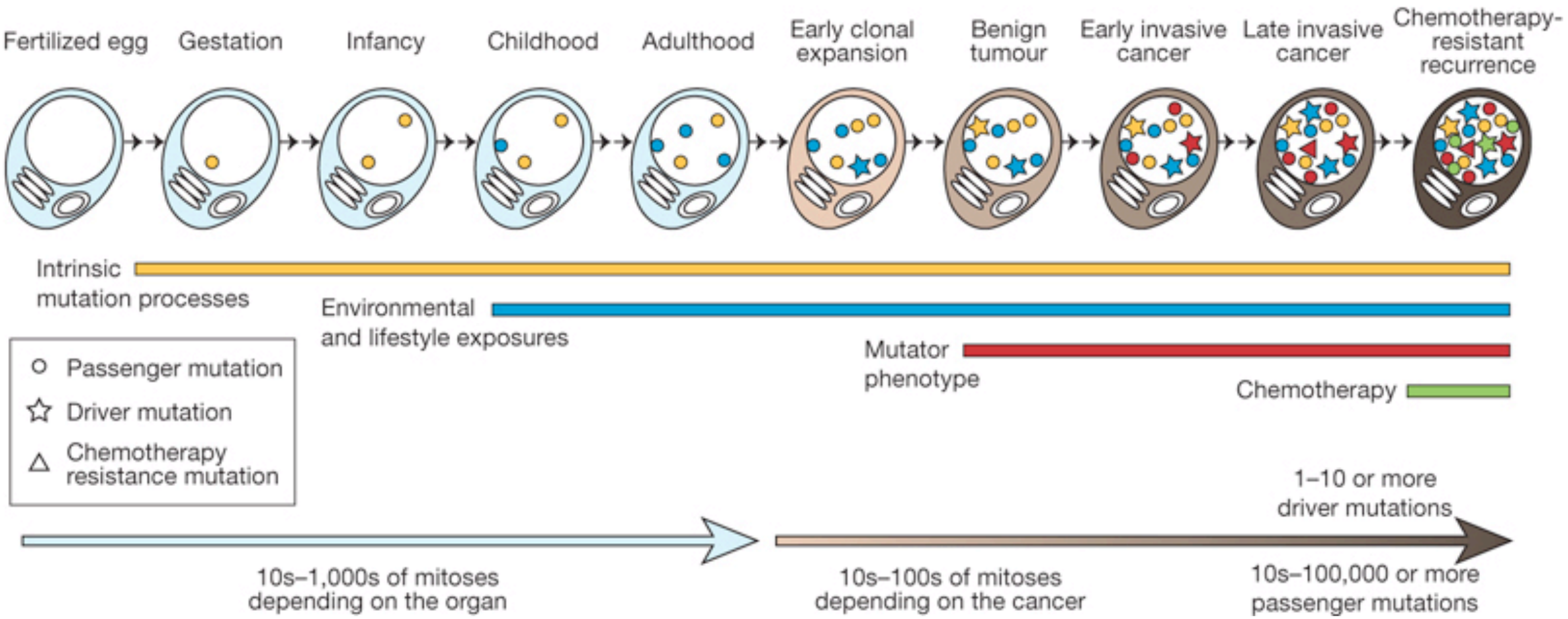


Sjoblom et al., Science 2006

Detecting the Troublemakers

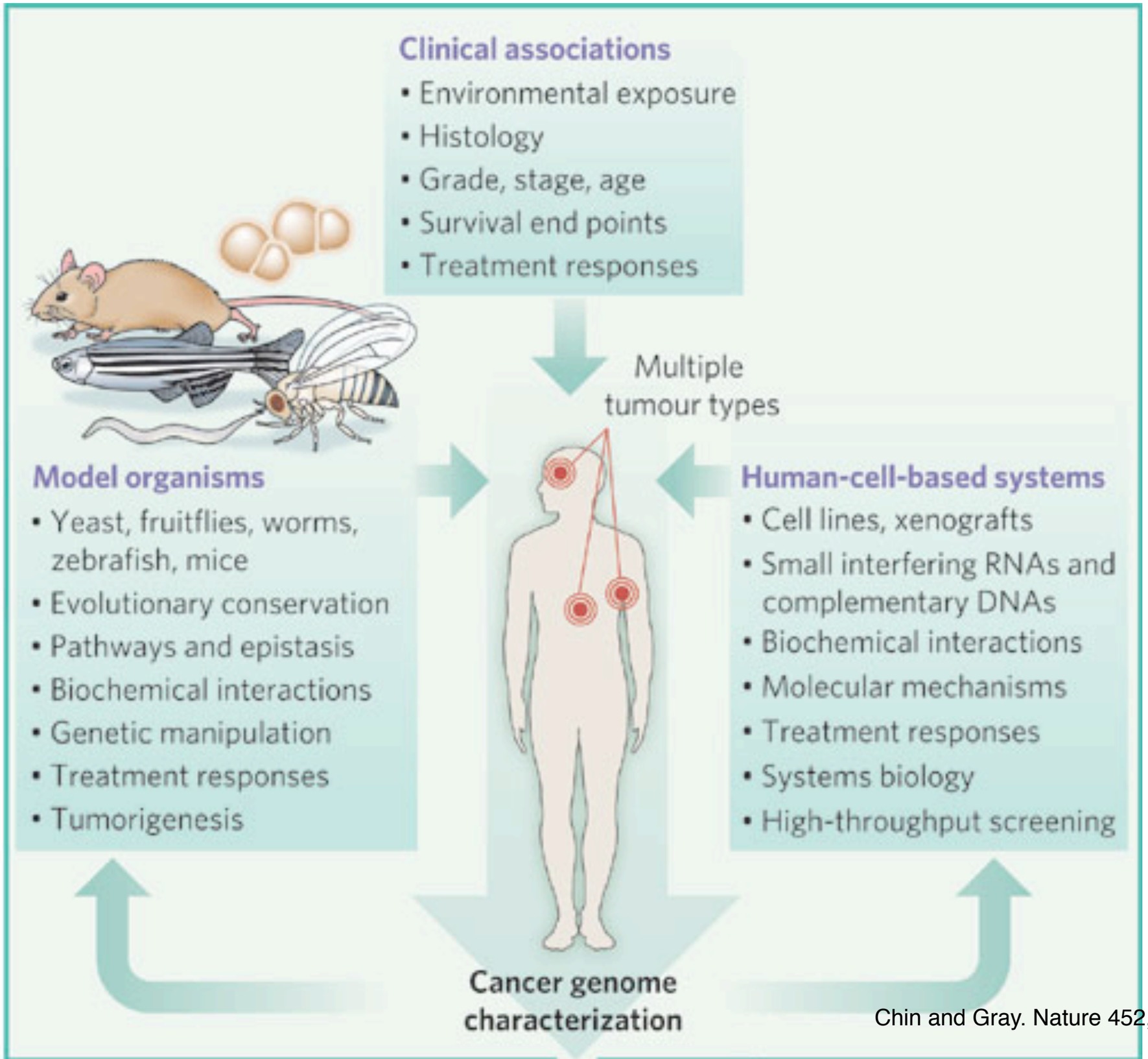


Drivers vs Passengers

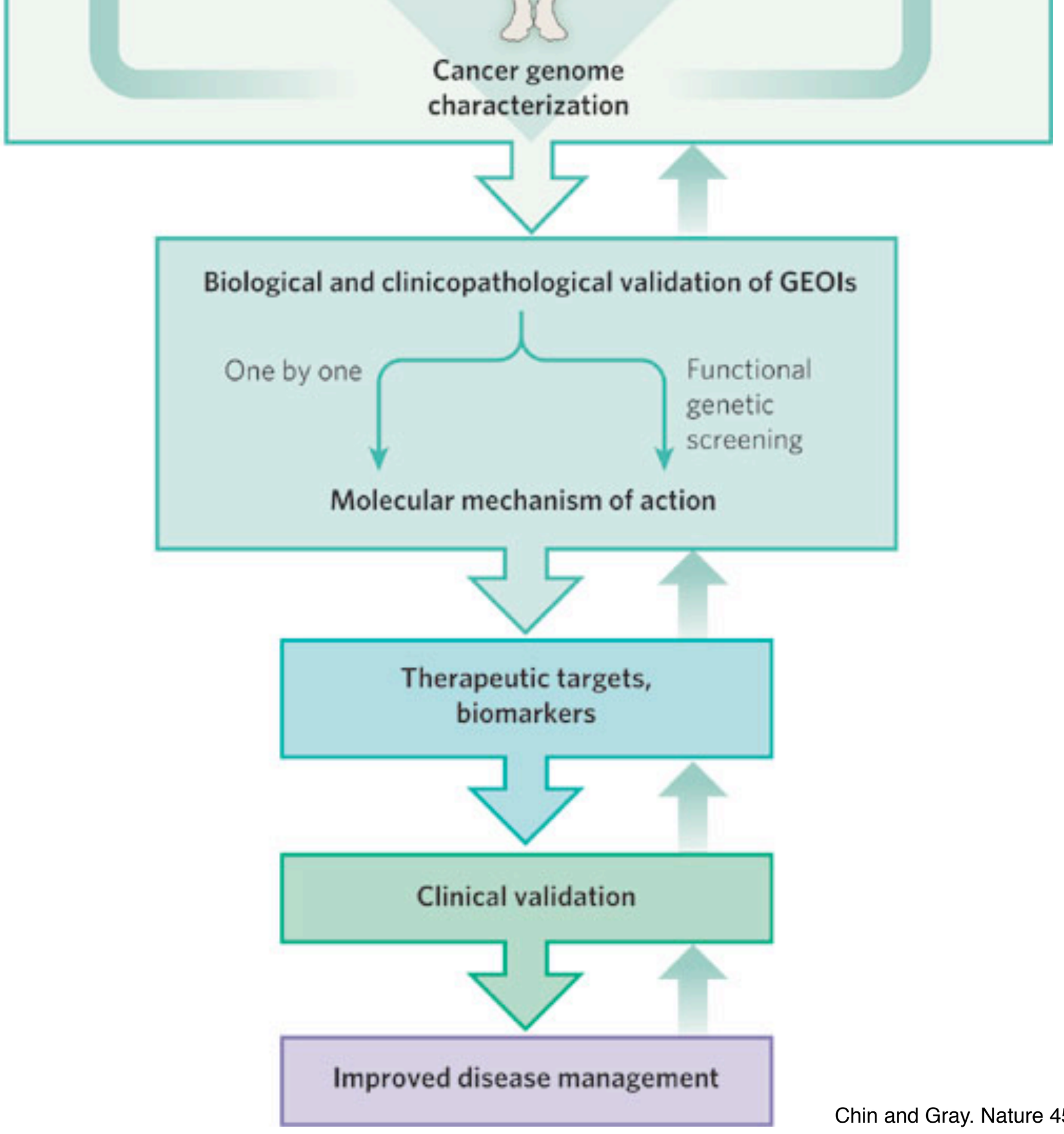


Straton et al., Nature. 2009 Apr 9;458(7239):719-24.

Cancer genome characterization

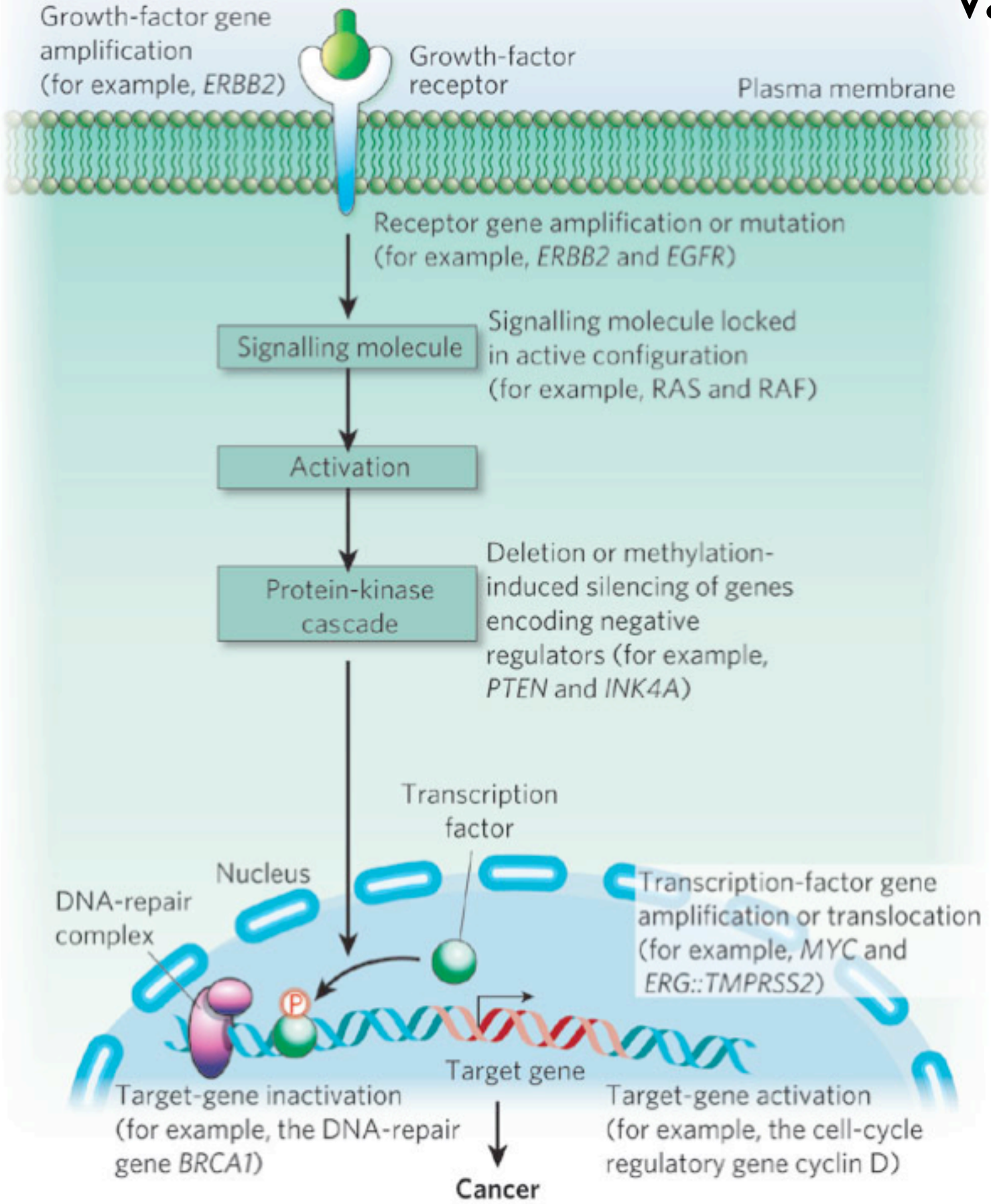


Chin and Gray. Nature 452, 553-563(3 April 2008)



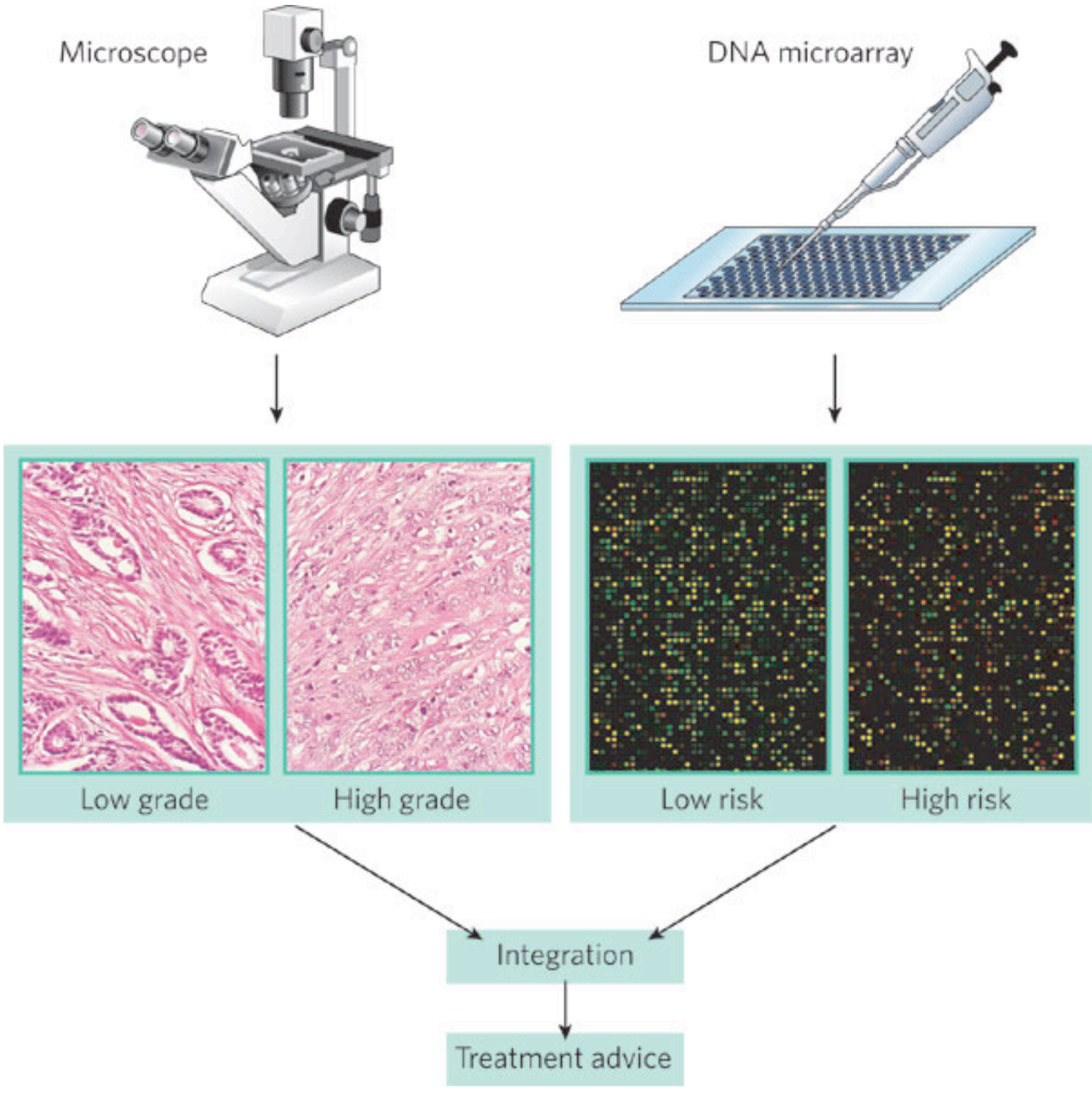
Chin and Gray. Nature 452, 553-563(3 April 2008)

Various alterations affecting the same pathway



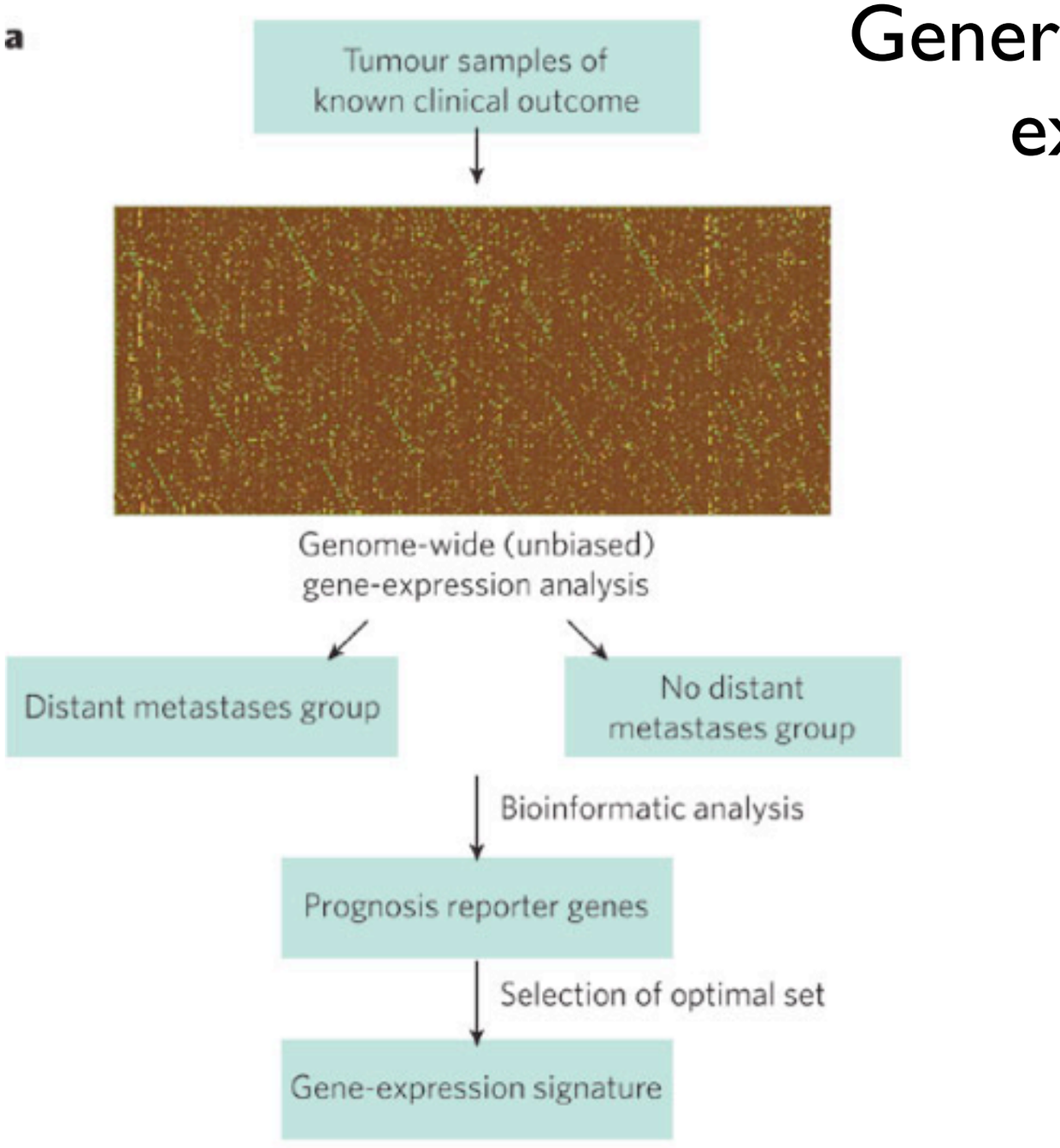
Chin and Gray. Nature 452, 553-563(3 April 2008)

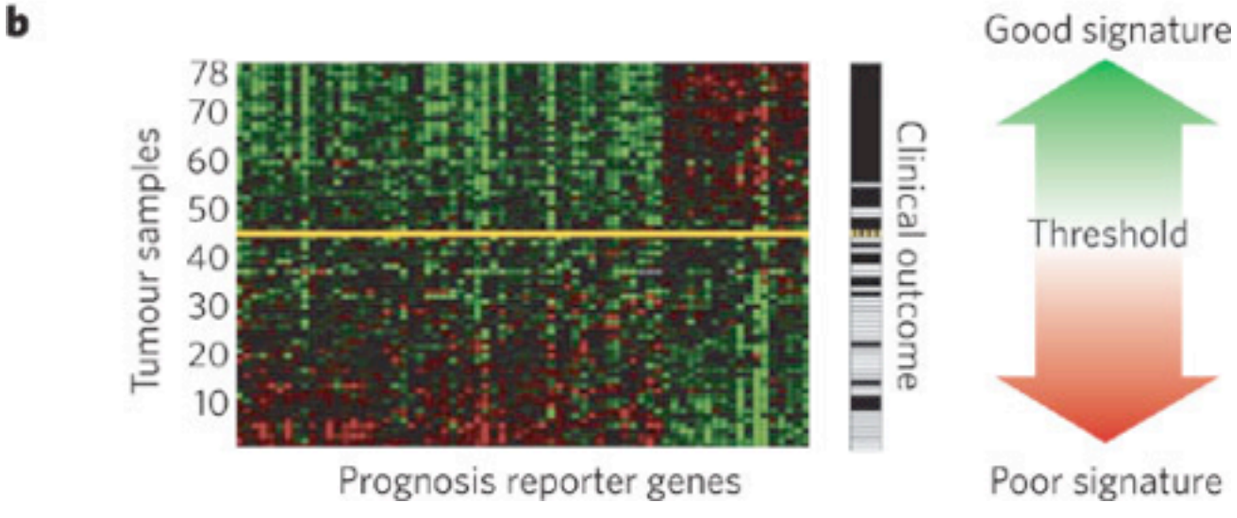
Molecular Diagnostics of Cancer



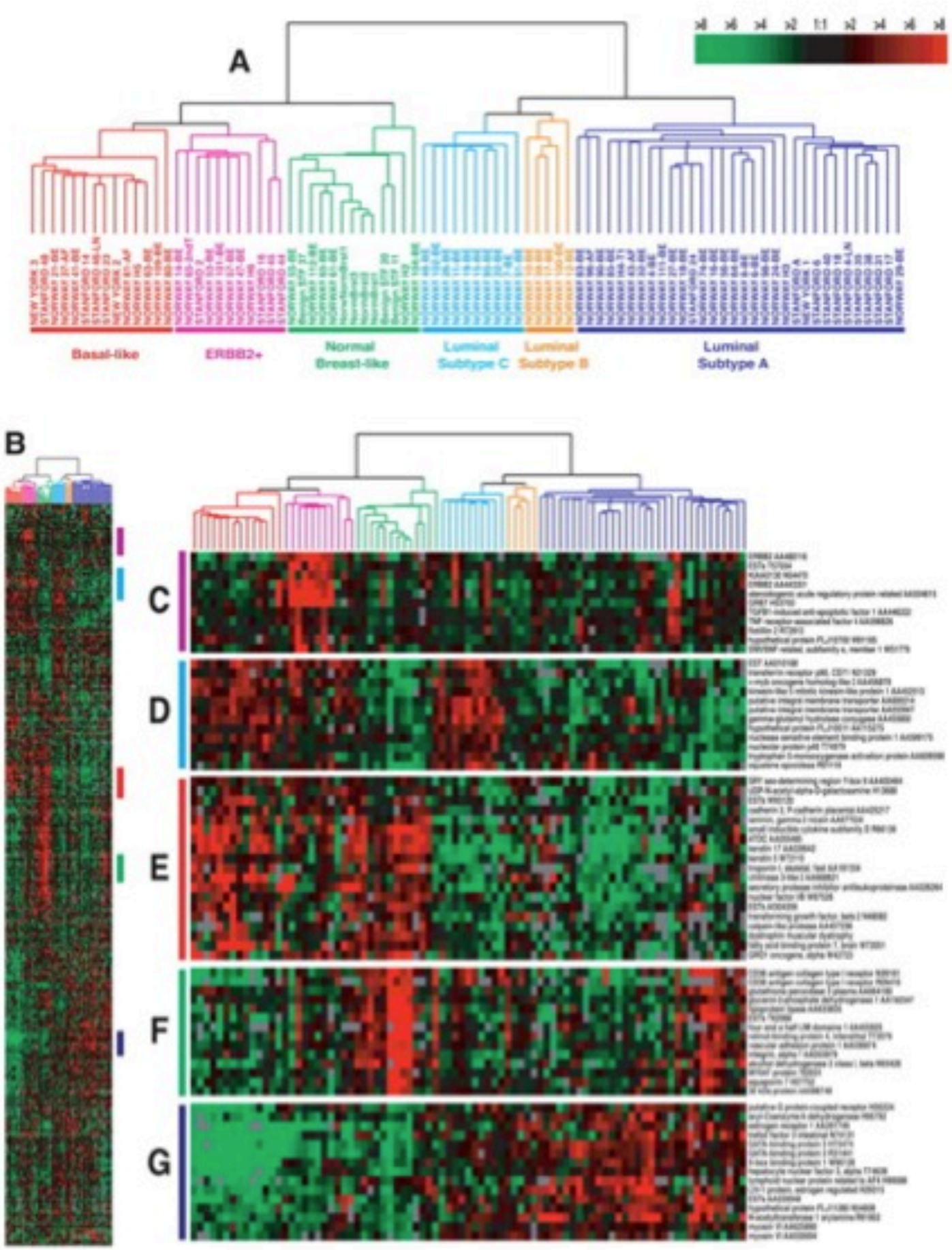
Laura J. van 't Veer & René Bernards. Nature 452, 564-570(3 April 2008)

Generating a prognostic gene-expression signature





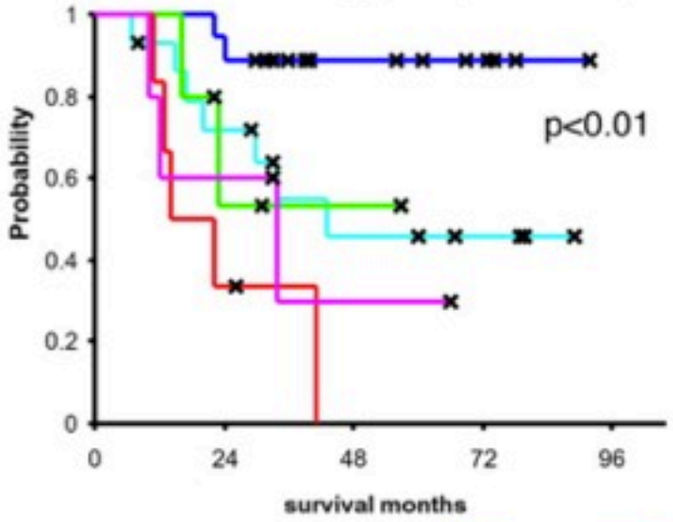
Laura J. van 't Veer & René Bernards. Nature 452, 564-570(3 April 2008)



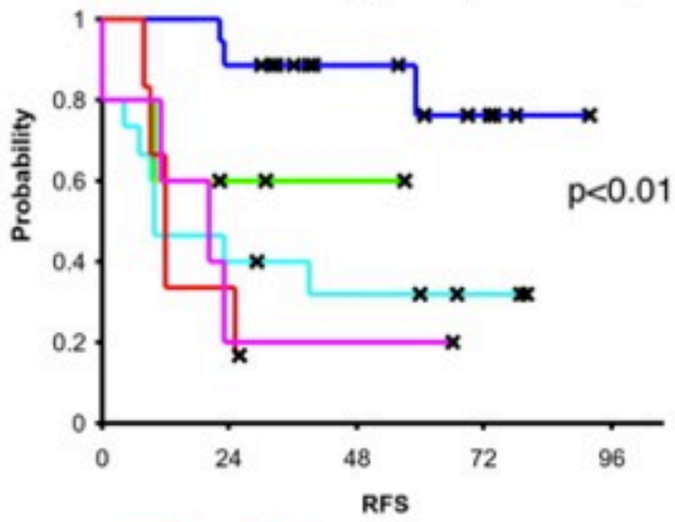
Sørlie T et al. PNAS 2001;98:10869-10874

Overall and relapse-free survival analysis of the 49 breast cancer patients, uniformly treated in a prospective study, based on different gene expression classification.

A 5 tumor subtypes (based upon Fig 1)

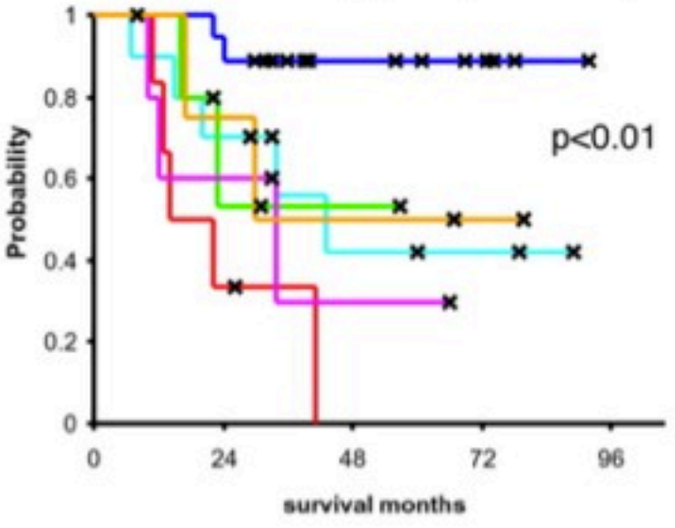


B 5 tumor subtypes (based upon Fig 1)



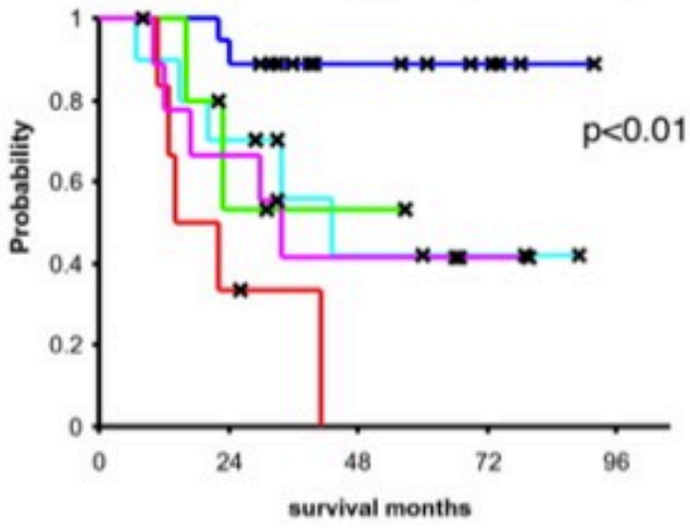
X Censored, Lum A, Lum B+C, NorB-like, Basal, ERBB2+

C 6 tumor subtypes (based upon Fig 1)



X Censored, Lum A, Lum C, NorB-like, Basal, ERBB2+, Lum B

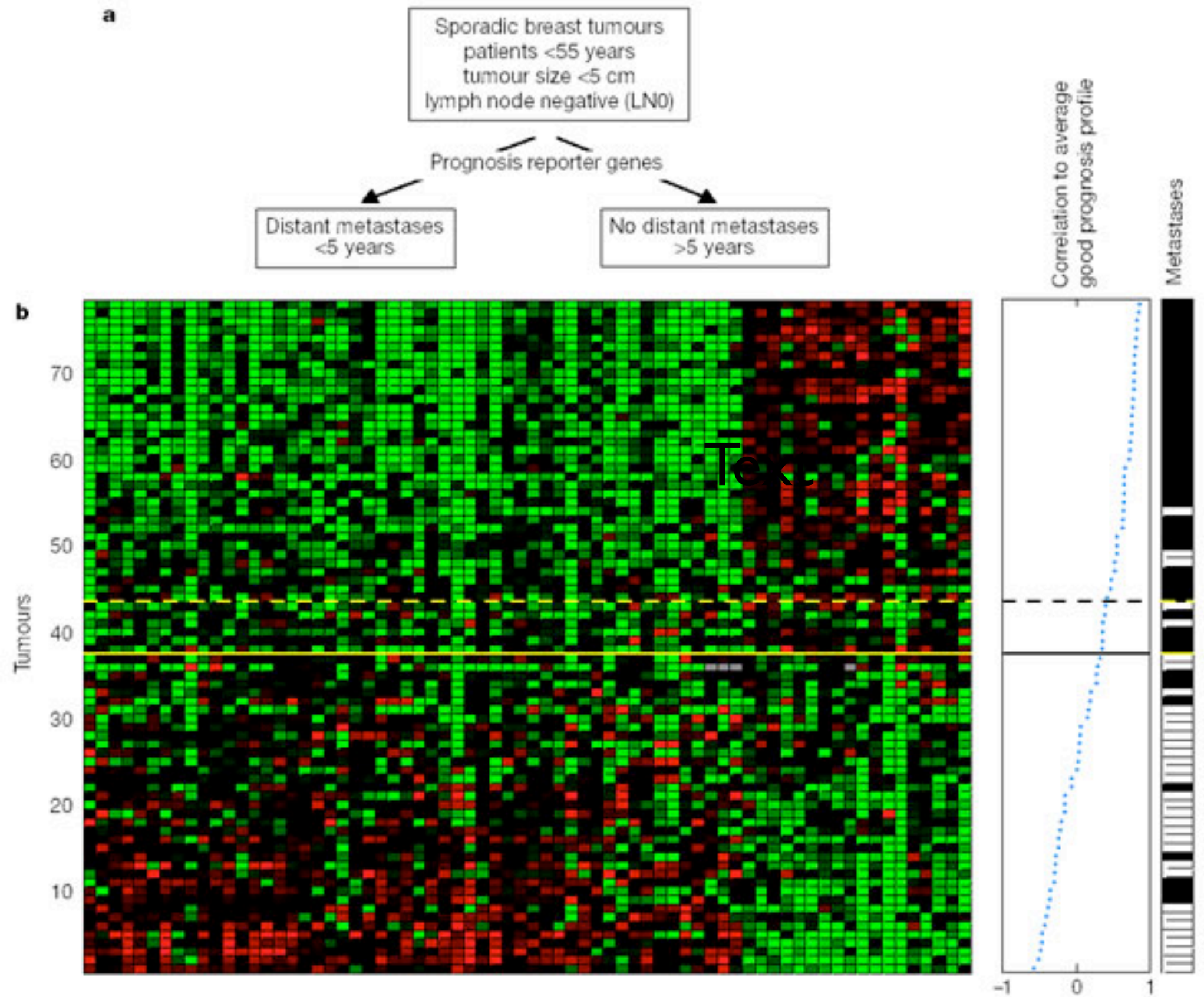
D 5 tumor subtypes (based upon Fig 5)



X Censored, Lum A, Lum C, NorB-like, Basal, ERBB2+ and Lum B

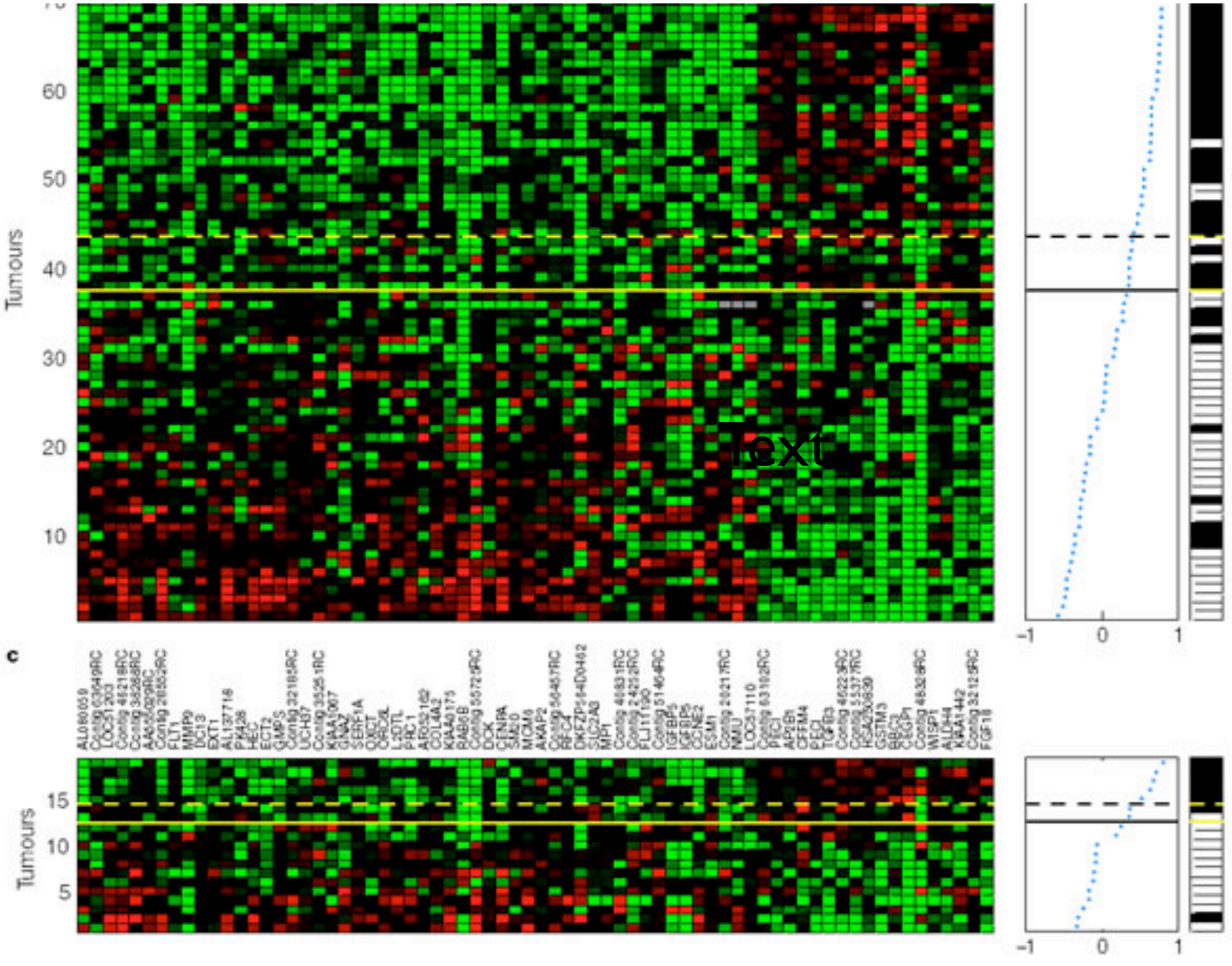
Sørlie T et al. PNAS 2001;98:10869-10874

70 gene prognostic signature in breast cancer



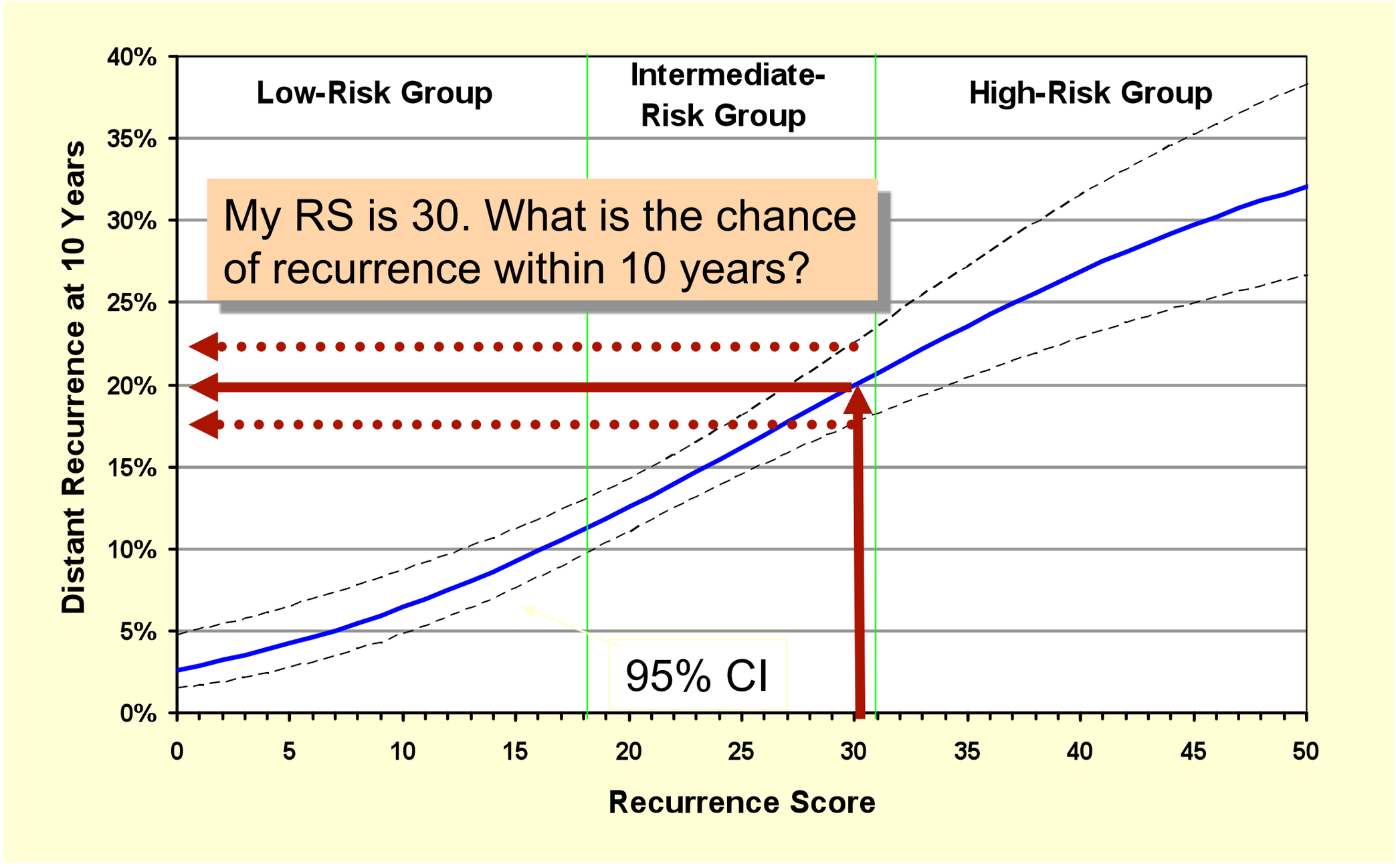
Laura J. van 't Veer, et al. Nature 415, 530-536(31 January 2002)

70 gene prognostic signature in breast cancer

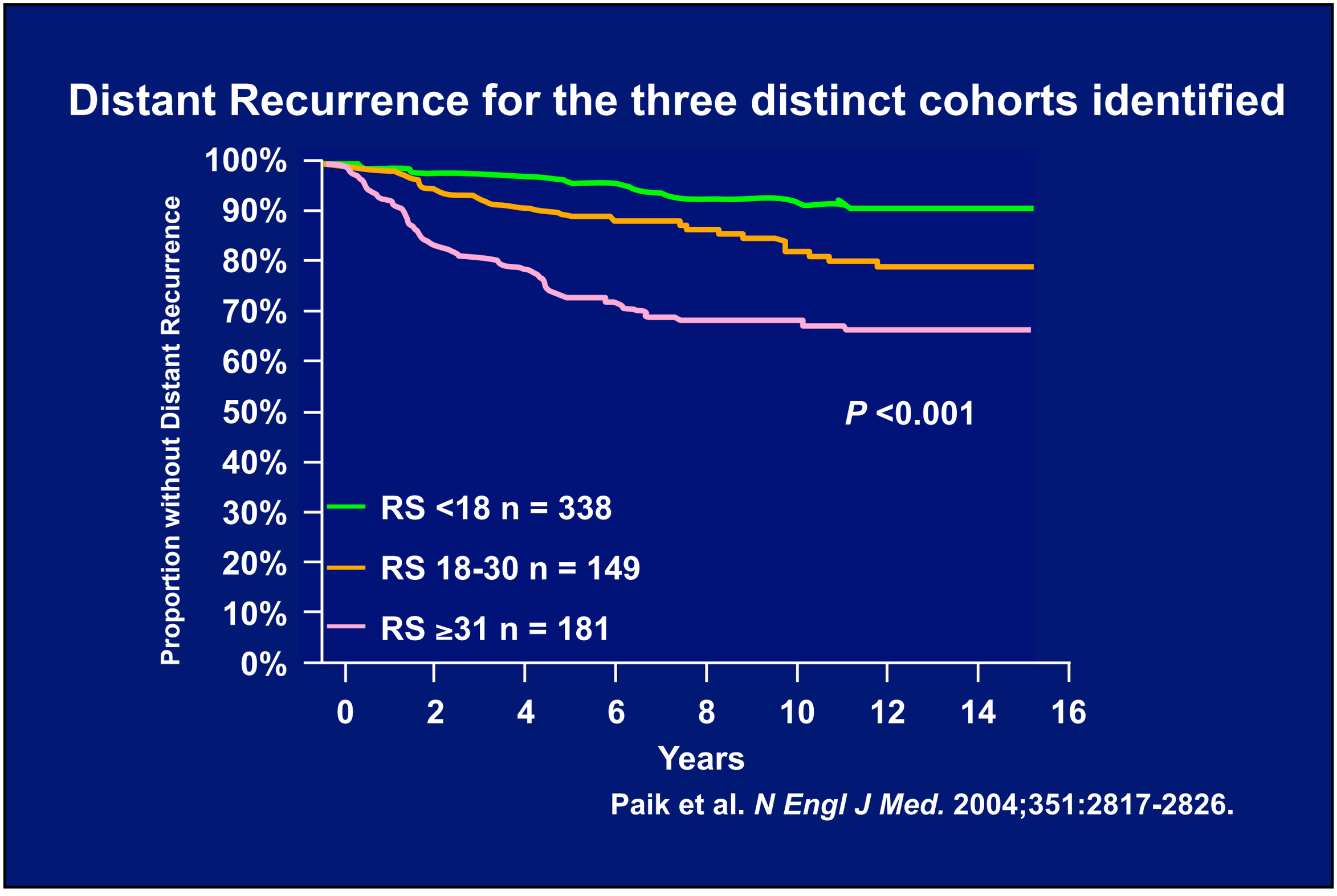


Laura J. van 't Veer, et al. Nature 415, 530-536(31 January 2002)

Oncotype DX: 21 gene signature in breast cancer for predicting distance recurrence

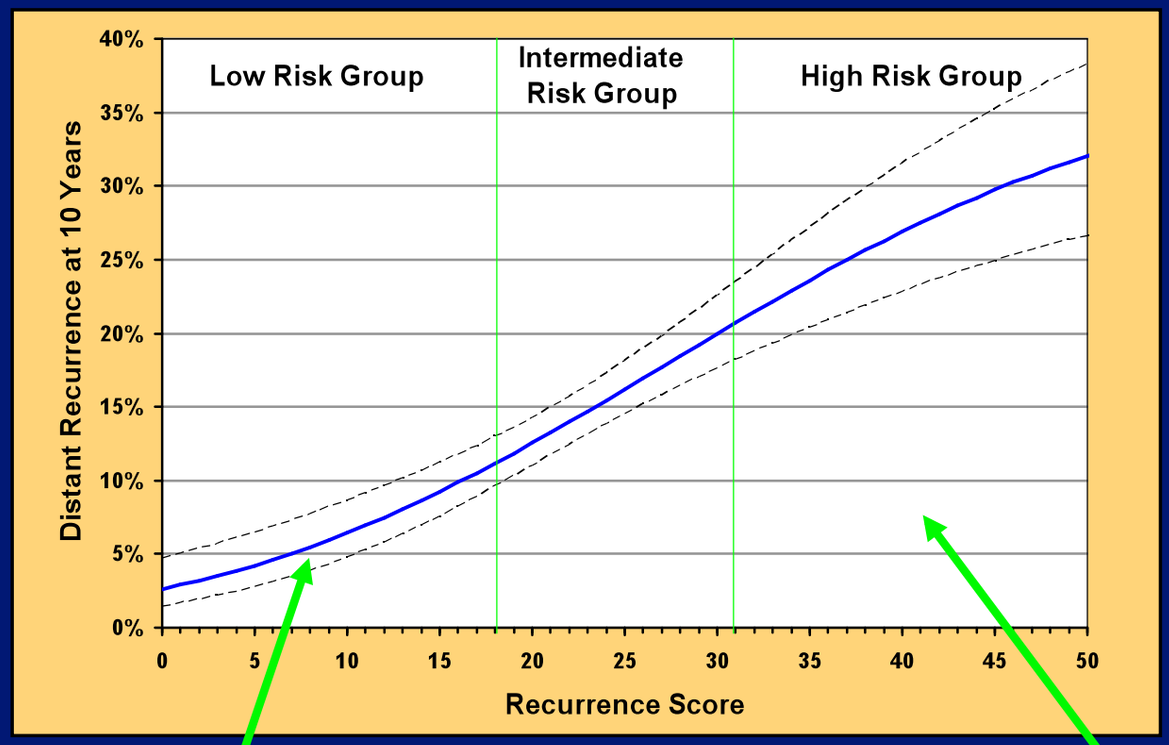


Oncotype DX: 21 gene signature in breast cancer for predicting distance recurrence



Oncotype DX: 21 gene signature in breast cancer for predicting distance recurrence

Recurrence Score in N-, ER+ patients



Lower RS's

- **Lower likelihood of recurrence**
- **Greater magnitude of TAM benefit**
- **Minimal, if any, chemotherapy benefit**

Higher RS's

- **Greater likelihood of recurrence**
- **Lower magnitude of TAM benefit**
- **Clear chemotherapy benefit**

1) Paik et al NEJM 2004, 2) Habel et al Breast Cancer Research 2006
3) Paik et al JCO 2006, 4) Gianni et al JCO 2005

oncotype DX® Breast Cancer Assay

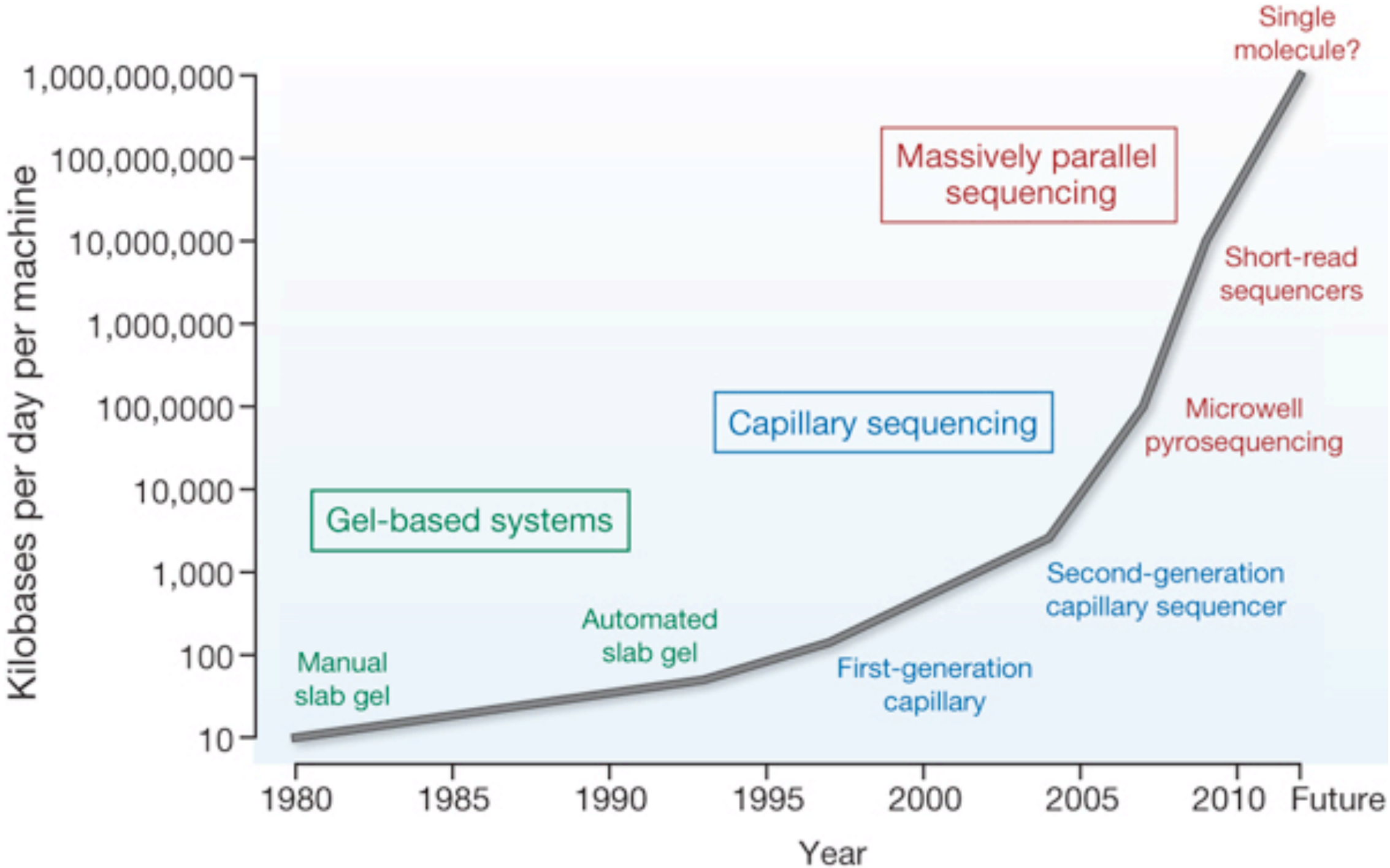
Login | Press Room | Register for Updates | Contact Us | Search

PATIENTS & CAREGIVERS | HEALTHCARE PROFESSIONALS | MANAGED CARE ORGANIZATIONS | CUSTOMER SUPPORT

The Oncotype DX® Breast Cancer Assay helps you find an answer

oncotype DX® Breast Cancer Assay

Next Generation Sequencing Revolution



Straton et al., Nature. 2009 Apr 9;458(7239):719-24.

Basics of the “old” technology

- Clone the DNA.
- Generate a ladder of labeled (colored) molecules that are different by 1 nucleotide.
- Separate mixture on some matrix.
- Detect fluorochrome by laser.
- Interpret peaks as string of DNA.
- Strings are 500 to 1,000 letters long
- 1 machine generates 57,000 nucleotides/run
- Assemble all strings into a genome.

Basics of the “new” technology

- Get DNA.
- Attach it to something.
- Extend and amplify signal with some color scheme.
- Detect fluorochrome by microscopy.
- Interpret series of spots as short strings of DNA.
- Strings are 30-300 letters long
- Multiple images are interpreted as 0.4 to 1.2 GB/run (1,200,000,000 letters/day).
- Map or align strings to one or many genome.

Template preparation

- Clonally amplified templates from a single DNA molecule
  - Emulsion PCR
  - Solid Phase Amplification
- Single DNA molecule
  - Primer immobilized
  - Template immobilized
  - Polymerase immobilized

Sequencing

- Imaging

- Data processing

Template preparation

Sequencing

Imaging

Data processing

Template preparation

Clonally amplified templates from a single DNA molecule

- Emulsion PCR
- Solid Phase Amplification

Single DNA molecule

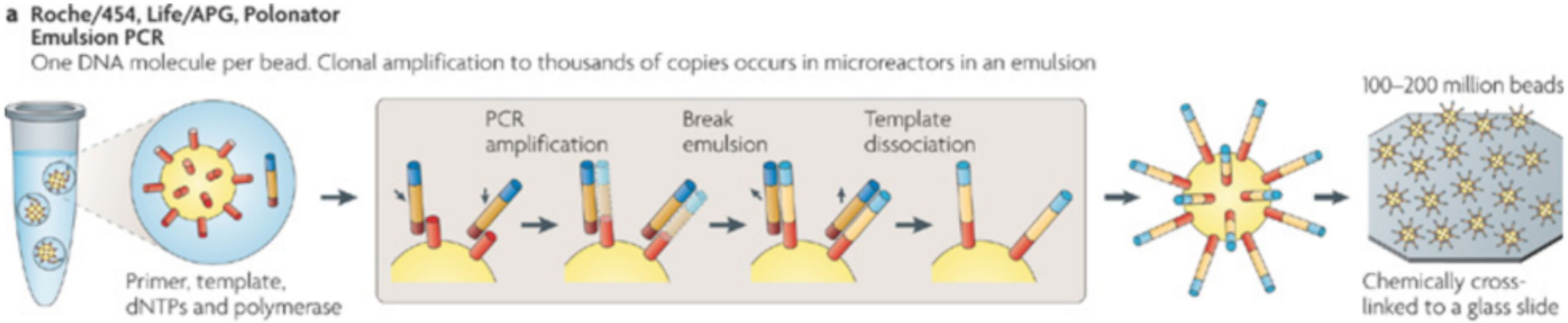
- Primer immobilized
- Template immobilized
- Polymerase immobilized

Template preparation

Clonally amplified templates from a single DNA molecule

Emulsion PCR

Solid Phase Amplification



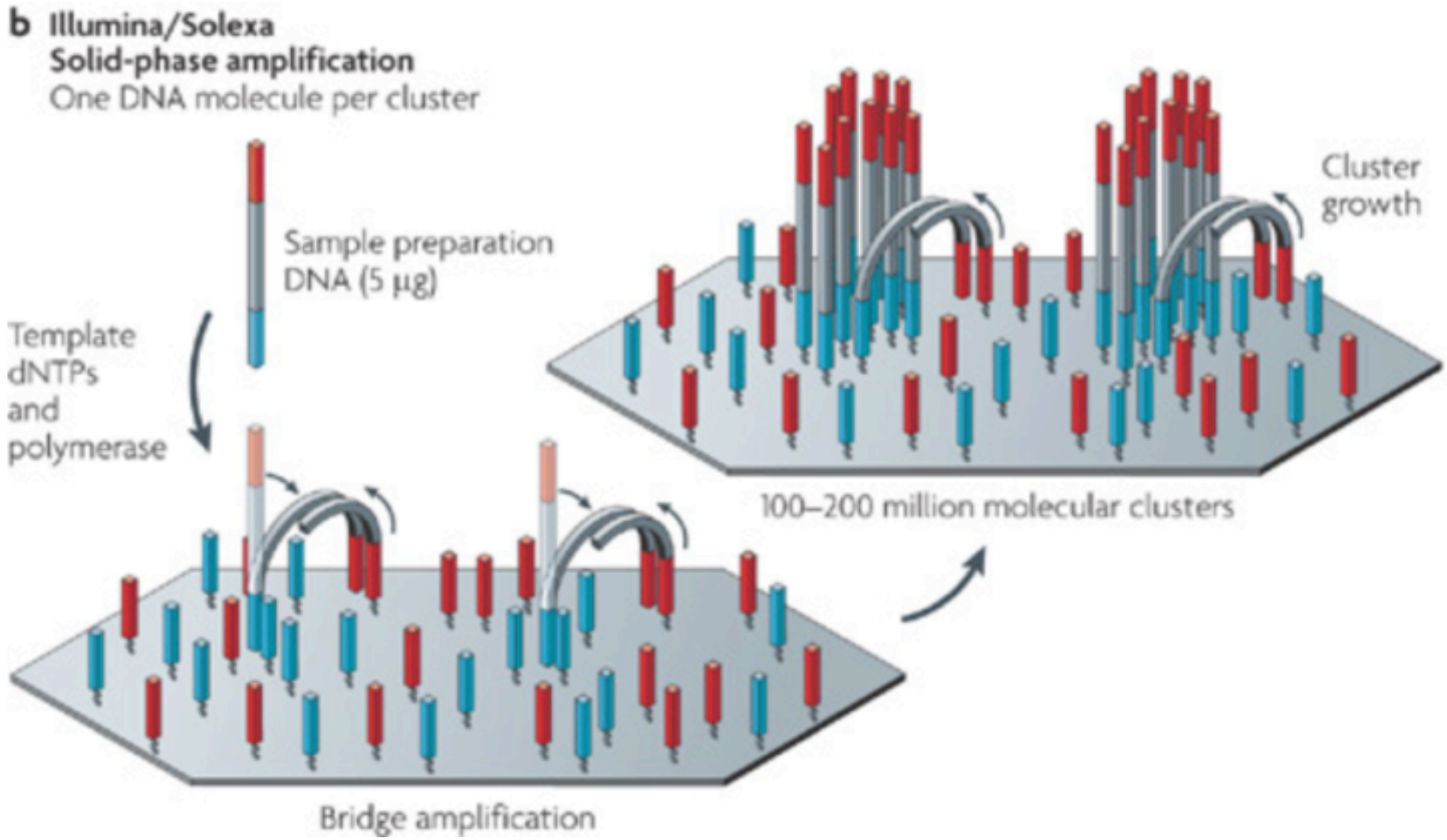
Michael L. Metzker. Nature Reviews Genetics 11, 31-46 (January 2010)

Template preparation

Clonally amplified templates from a single DNA molecule

Emulsion PCR

Solid Phase Amplification



Michael L. Metzker. Nature Reviews Genetics 11, 31-46 (January 2010)

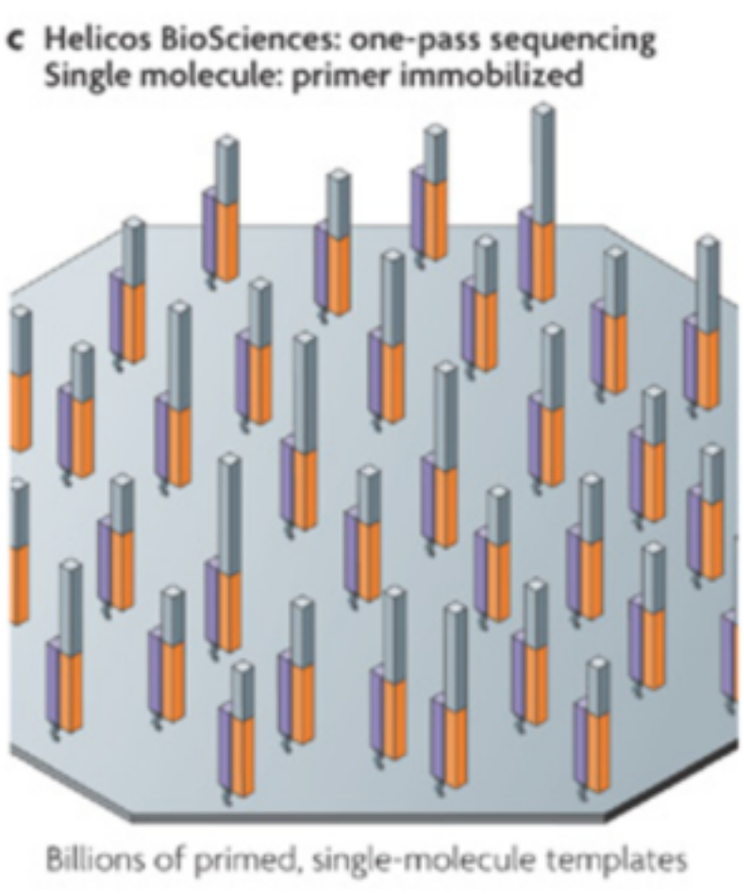
Template preparation

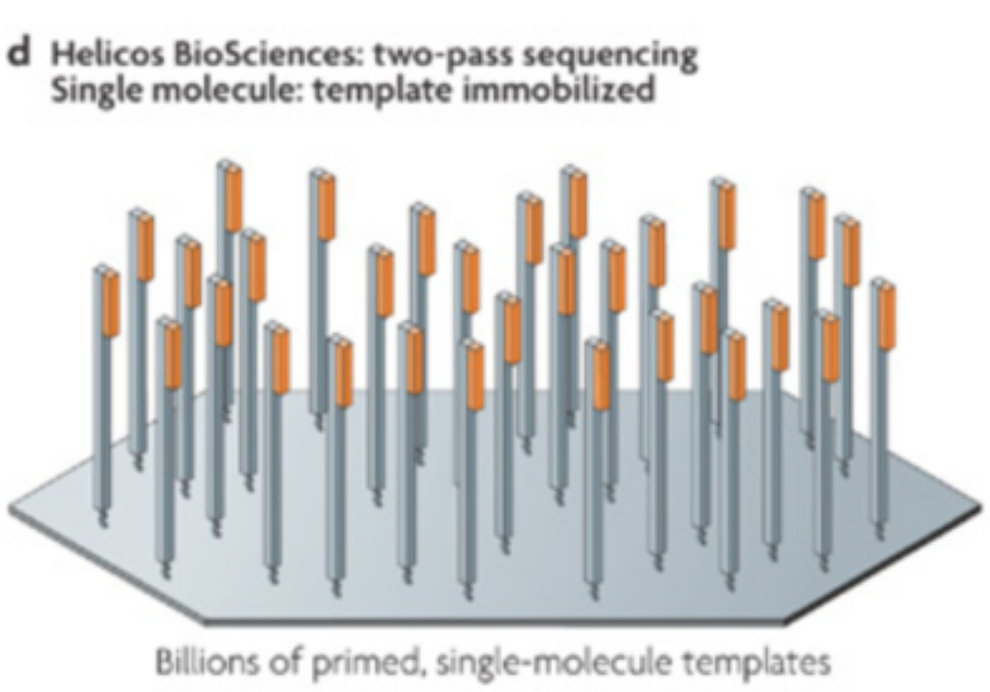
Clonally amplified templates from a single DNA molecule

- Emulsion PCR
- Solid Phase Amplification

Single DNA molecule

- Primer immobilized
- Template immobilized
- Polymerase immobilized





Michael L. Metzker. Nature Reviews Genetics 11, 31-46 (January 2010)

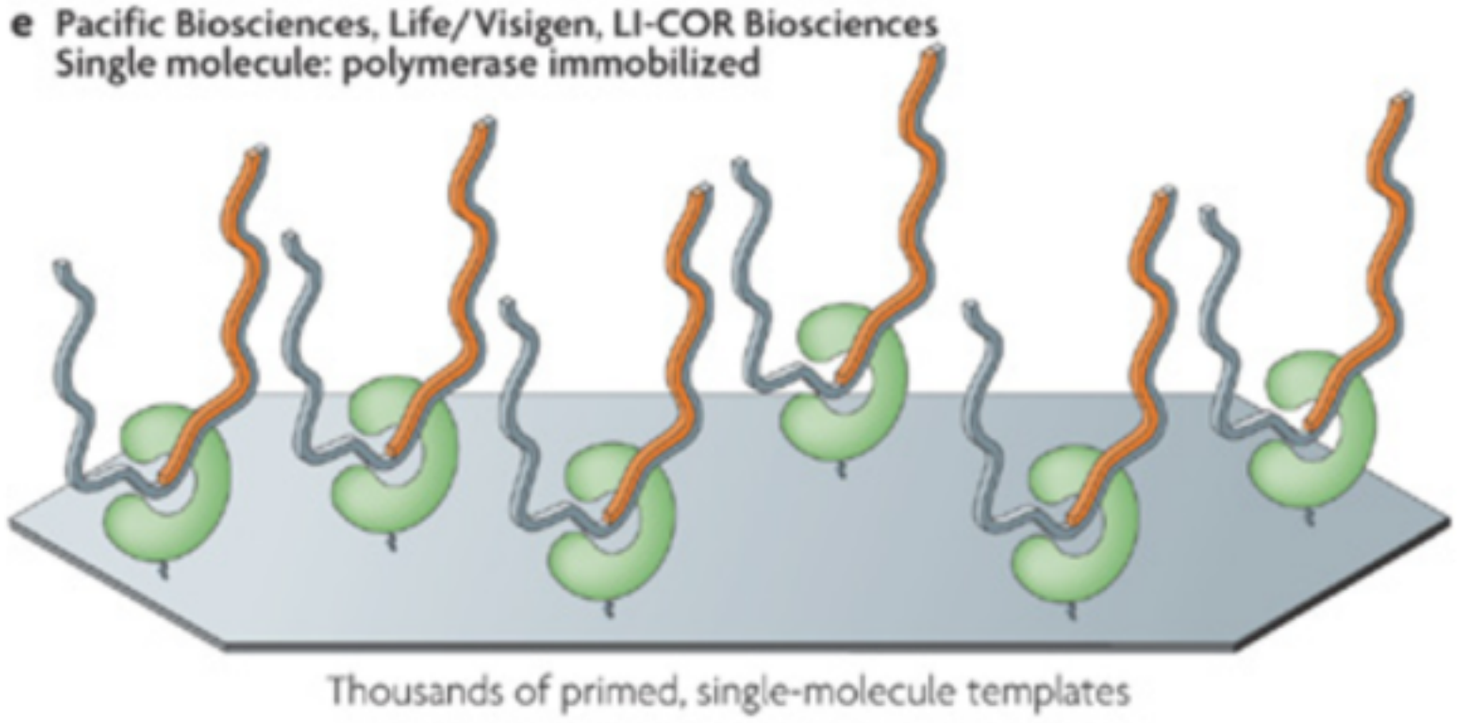
Template preparation

Clonally amplified templates from a single DNA molecule

- Emulsion PCR
- Solid Phase Amplification

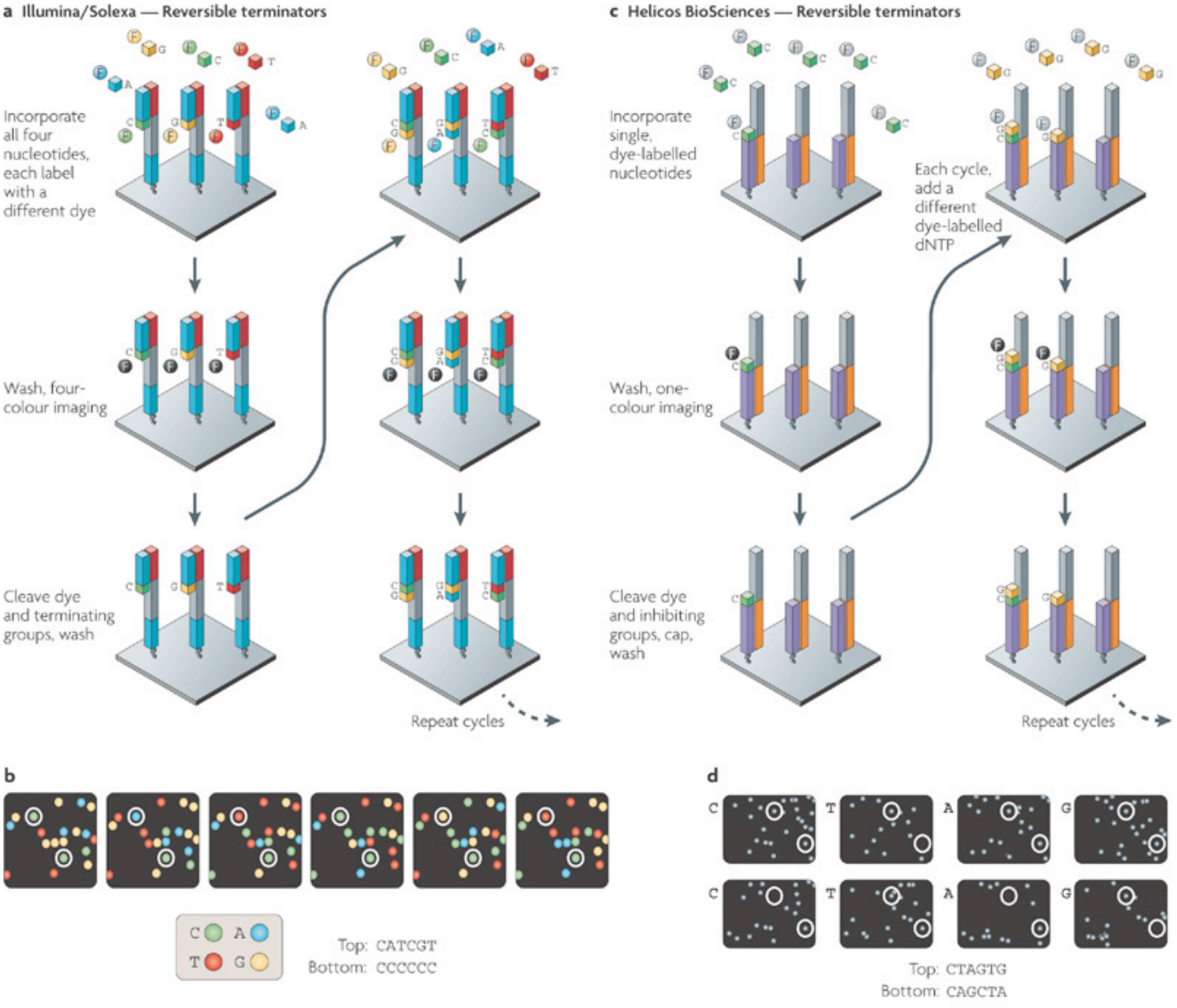
Single DNA molecule

- Primer immobilized
- Template immobilized
- Polymerase immobilized



Michael L. Metzker. Nature Reviews Genetics 11, 31-46 (January 2010)

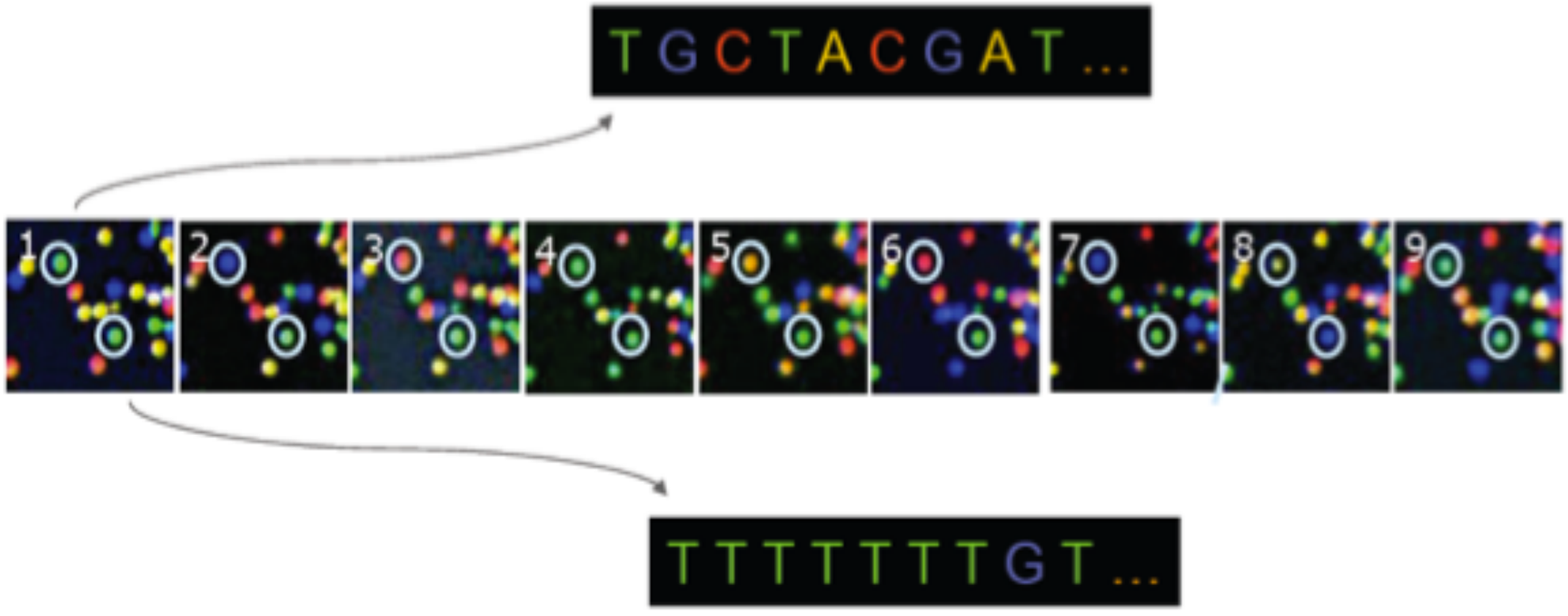
Sequencing/Imaging



Michael L. Metzker. Nature Reviews Genetics 11, 31-46 (January 2010)

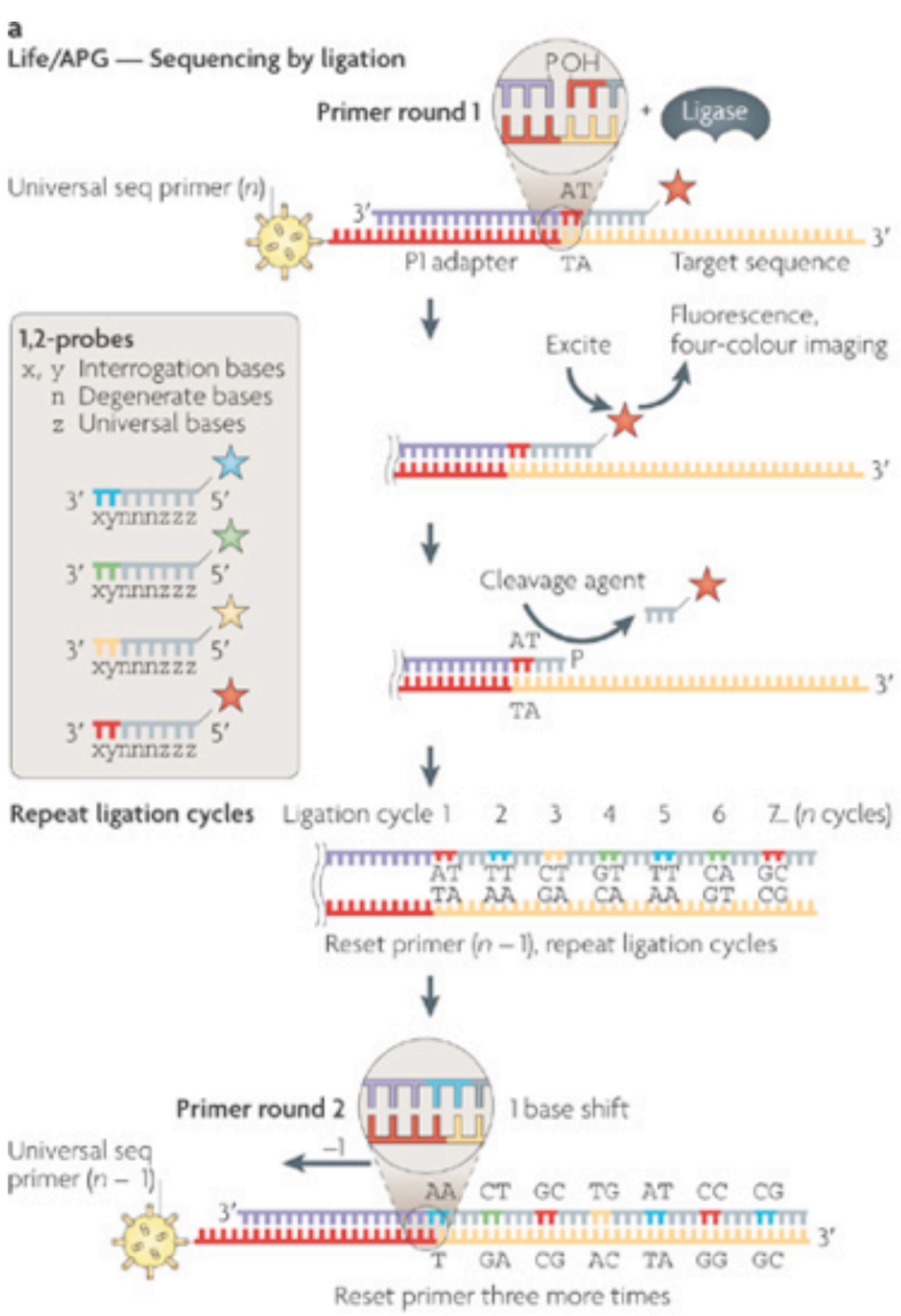
Sequencing/Imaging

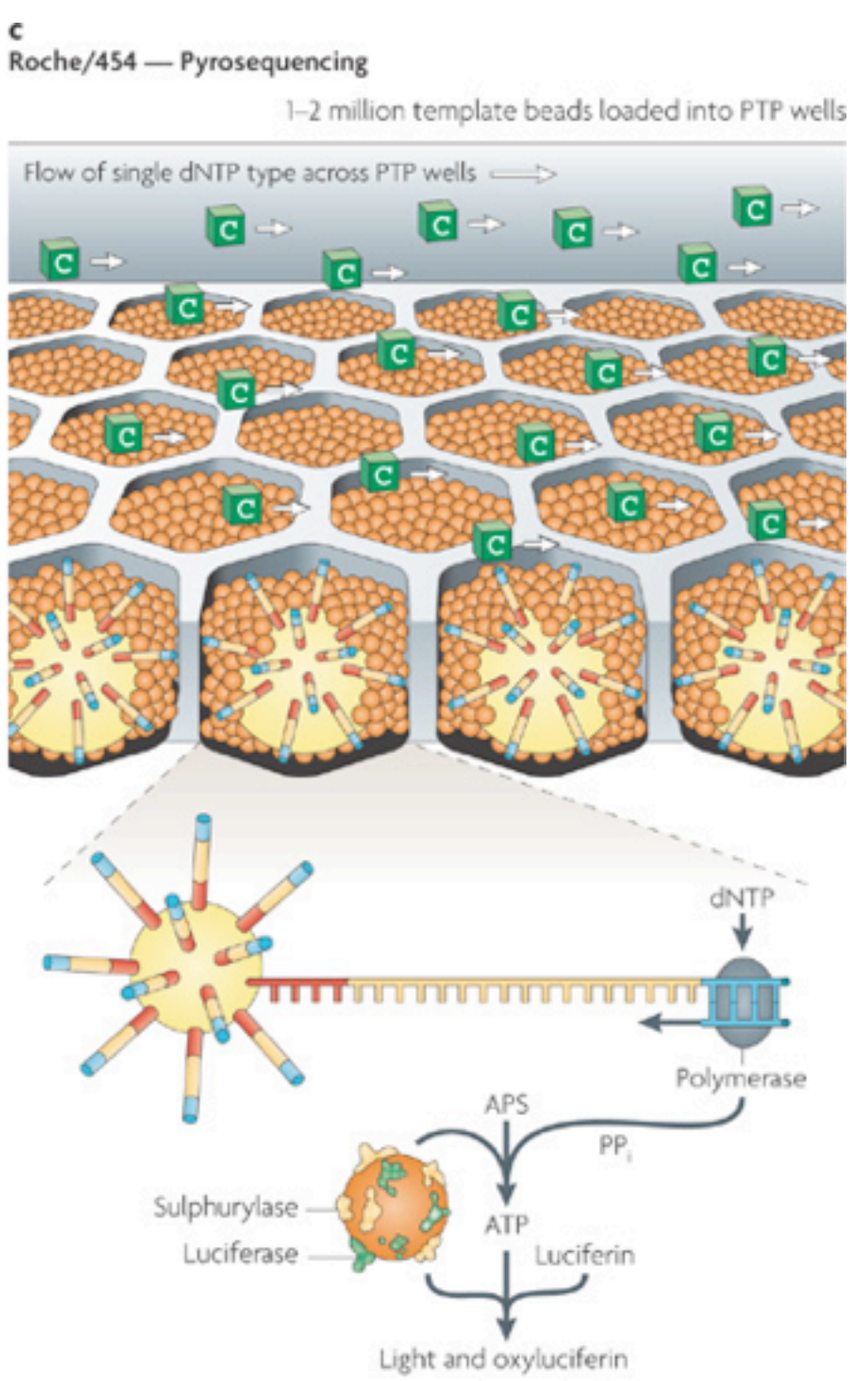
Base calling from raw data

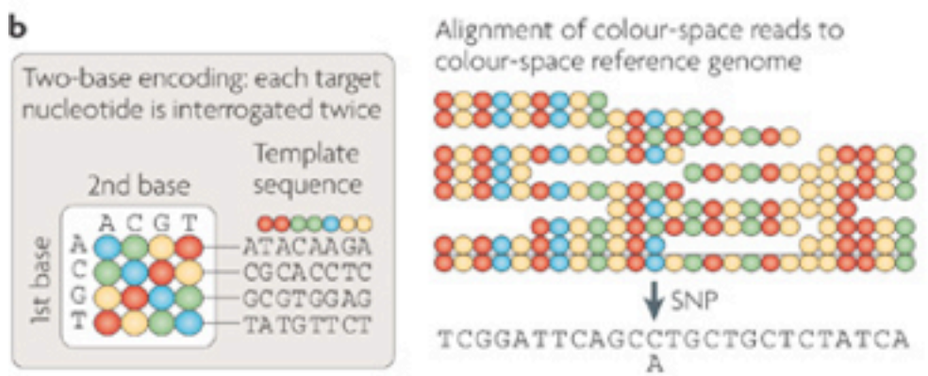


The identity of each base of a cluster is read off from sequential images

Sequencing/Imaging

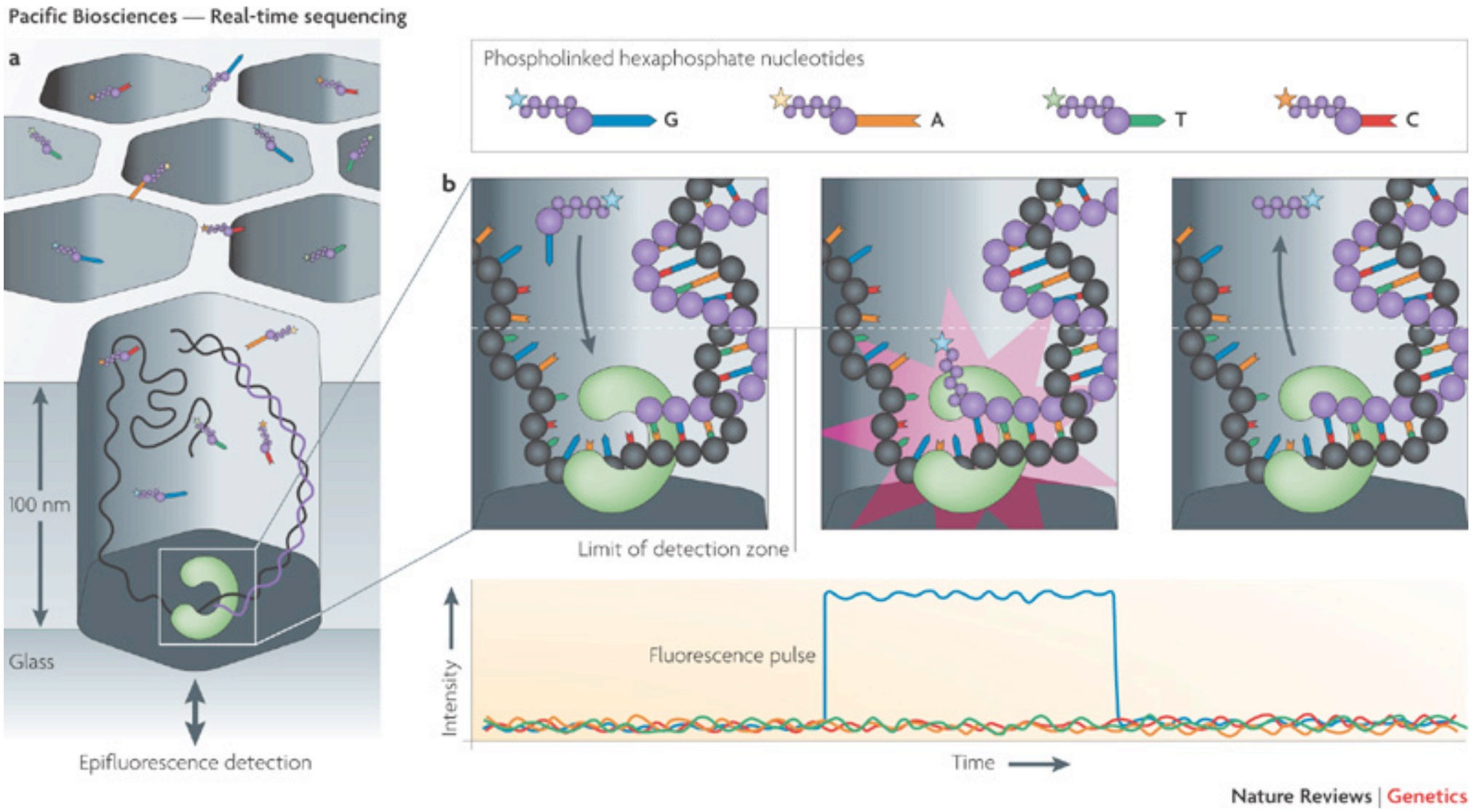






Michael L. Metzker. Nature Reviews Genetics 11, 31-46 (January 2010)

Sequencing/Imaging



Michael L. Metzker. Nature Reviews Genetics 11, 31-46 (January 2010)

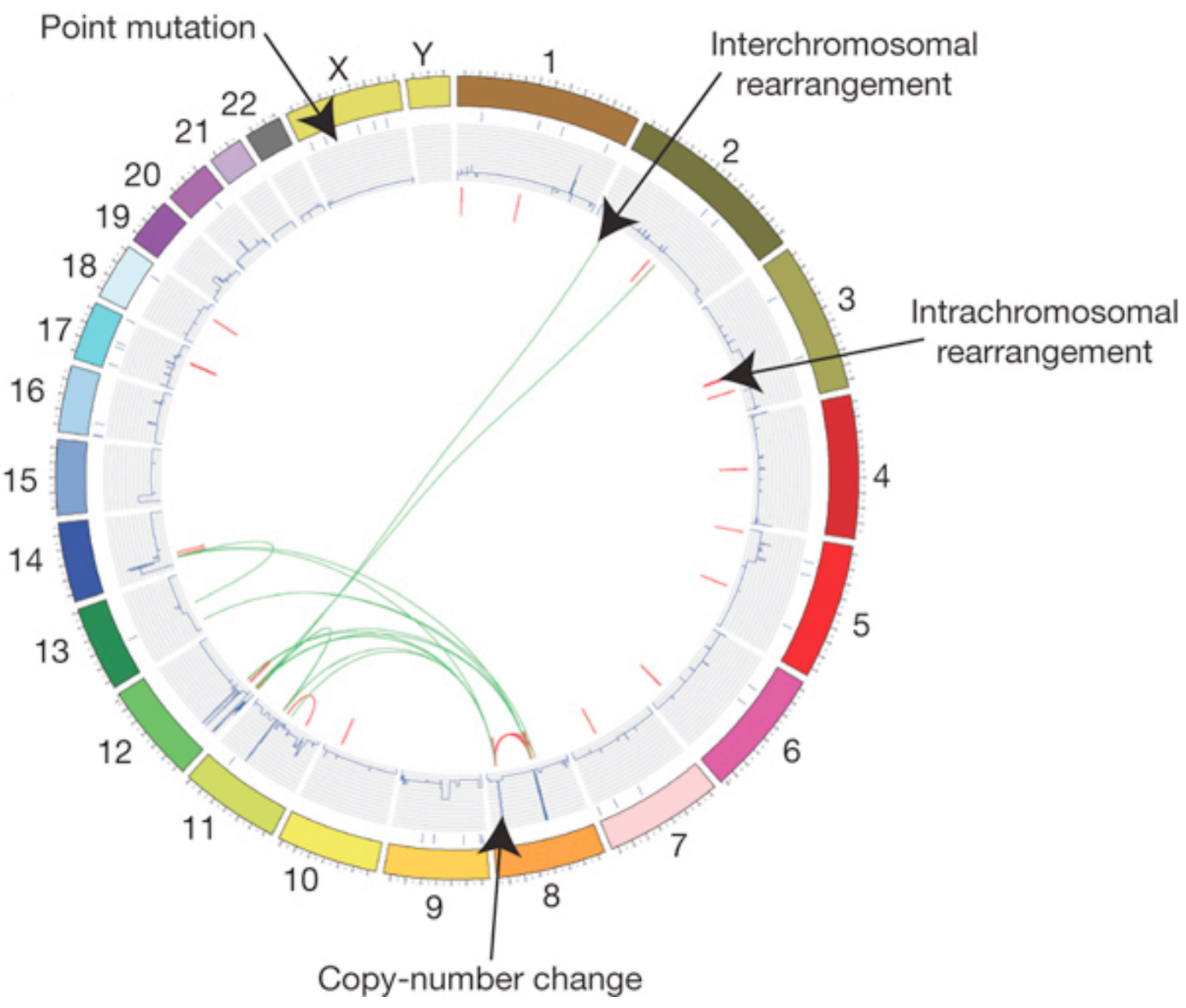
Platform	Library/ template preparation	NGS chemistry	Read length (bases)	Run time (days)	Gb per run	Machine cost (US$)	Pros	Cons	Biological applications	Refs
Roche/454's GS FLX Titanium	Frag, MP/ emPCR	PS	330*	0.35	0.45	500,000	Longer reads improve mapping in repetitive regions; fast run times	High reagent cost; high error rates in homo-polymer repeats	Bacterial and insect genome *de novo* assemblies; medium scale (<3 Mb) exome capture; 16S in metagenomics	D. Muzny, pers. comm.
Illumina/ Solexa's GA_{II}	Frag, MP/ solid-phase	RTs	75 or 100	4‡, 9§	18‡, 35§	540,000	Currently the most widely used platform in the field	Low multiplexing capability of samples	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Life/APG's SOLiD 3	Frag, MP/ emPCR	Cleavable probe SBL	50	7‡, 14§	30‡, 50§	595,000	Two-base encoding provides inherent error correction	Long run times	Variant discovery by whole-genome resequencing or whole-exome capture; gene discovery in metagenomics	D. Muzny, pers. comm.
Polonator G.007	MP only/ emPCR	Non-cleavable probe SBL	26	5§	12§	170,000	Least expensive platform; open source to adapt alternative NGS chemistries	Users are required to maintain and quality control reagents; shortest NGS read lengths	Bacterial genome resequencing for variant discovery	J. Edwards, pers. comm.
Helicos BioSciences HeliScope	Frag, MP/ single molecule	RTs	32*	8‡	37‡	999,000	Non-bias representation of templates for genome and seq-based applications	High error rates compared with other reversible terminator chemistries	Seq-based methods	91
Pacific Biosciences (target release: 2010)	Frag only/ single molecule	Real-time	964*	N/A	N/A	N/A	Has the greatest potential for reads exceeding 1 kb	Highest error rates compared with other NGS chemistries	Full-length transcriptome sequencing; complements other resequencing efforts in discovering large structural variants and haplotype blocks	S. Turner, pers. comm.

*Average read-lengths. ‡Fragment run. §Mate-pair run. Frag, fragment; GA, Genome Analyzer; GS, Genome Sequencer; MP, mate-pair; N/A, not available; NGS, next-generation sequencing; PS, pyrosequencing; RT, reversible terminator; SBL, sequencing by ligation; SOLiD, support oligonucleotide ligation detection.

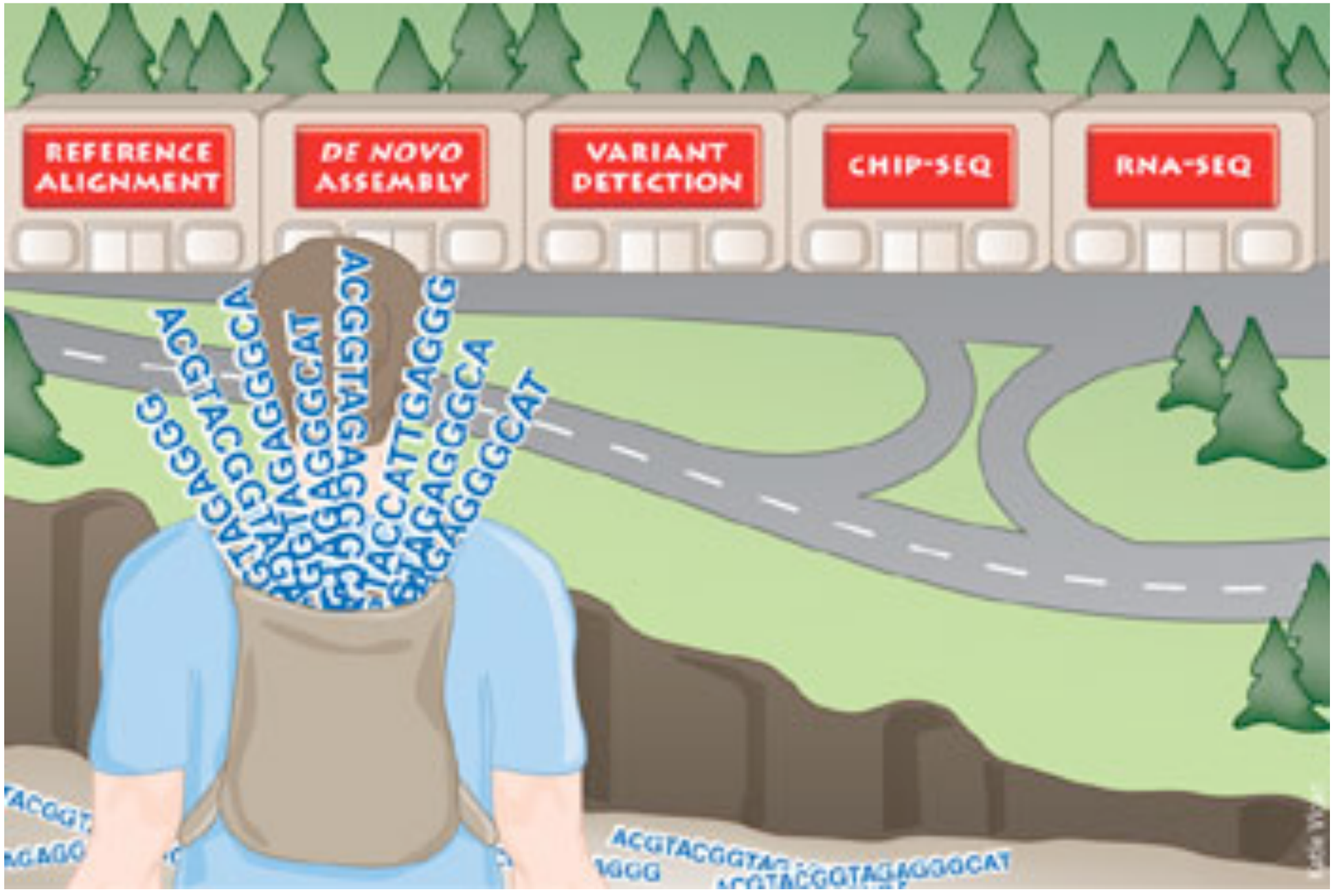
Personal Genome	Platform	Genomic template libraries	No. of reads (millions)	Read length (bases)	Base coverage (fold)	Assembly	Genome coverage (%)*	SNVs in millions (alignment tool)	No. of runs	Estimated cost (US$)
J. Craig Venter	Automated Sanger	MP from BACs, fosmids & plasmids	31.9	800	7.5	*De novo*	N/A	3.21	>340,000	70,000,000
James D. Watson	Roche/454	Frag: 500 bp	93.2‡	250§	7.4	Aligned*	95‖	3.32 (BLAT)	234	1,000,000¶
Yoruban male (NA18507)	Illumina/ Solexa	93% MP: 200 bp	3,410‡	35	40.6	Aligned*	99.9	3.83 (MAQ)	40	250,000¶
		7% MP: 1.8 kb	271	35				4.14 (ELAND)		
Han Chinese male	Illumina/ Solexa	66% Frag: 150–250 bp	1,921‡	35	36	Aligned*	99.9	3.07 (SOAP)	35	500,000¶
		34% MP: 135 bp & 440 bp	1,029	35						
Korean male (AK1)	Illumina/ Solexa	21% Frag: 130 bp & 440 bp	393‡	36	27.8	Aligned*	99.8	3.45 (GSNAP)	30	200,000¶
		79% MP: 130 bp, 390 bp & 2.7 kb	1,156	36, 88, 106						
Korean male (SJK)	Illumina/ Solexa	MP: 100 bp, 200 bp & 300 bp	1,647‡	35, 74	29.0	Aligned*	99.9	3.44 (MAQ)	15	250,000¶,#
Yoruban male (NA18507)	Life/APG	9% Frag: 100–500 bp	211‡	50	17.9	Aligned*	98.6	3.87 (Corona-lite)	9.5	60,000¶,**
		91% MP: 600–3,500 bp	2,075‡	25, 50						
Stephen R. Quake	Helicos BioSciences	Frag: 100–500 bp	2,725‡	32§	28	Aligned*	90	2.81 (IndexDP)	4	48,000¶
AML female	Illumina/ Solexa	Frag: 150–200 bp‡‡	2,730‡,‡‡	32	32.7	Aligned*	91	3.81‡‡ (MAQ)	98	1,600,000‖
		Frag: 150–200 bp§§	1,081‡,§§	35	13.9		83	2.92§§ (MAQ)	34	
AML male	Illumina/ Solexa	MP: 200–250 bp‡‡	1,620‡,‡‡	35	23.3	Aligned*	98.5	3.46‡‡ (MAQ)	16.5	500,000‖
		MP: 200–250 bp§§	1,351‡,§§	50	21.3		97.4	3.45§§ (MAQ)	13.1	
James R. Lupski CMT male	Life/APG	16% Frag: 100–500 bp	238‡	35	29.6	Aligned*	99.8	3.42 (Corona-lite)	3	75,000¶,¶¶
		84% MP: 600–3,500 bp	1,211‡	25, 50						

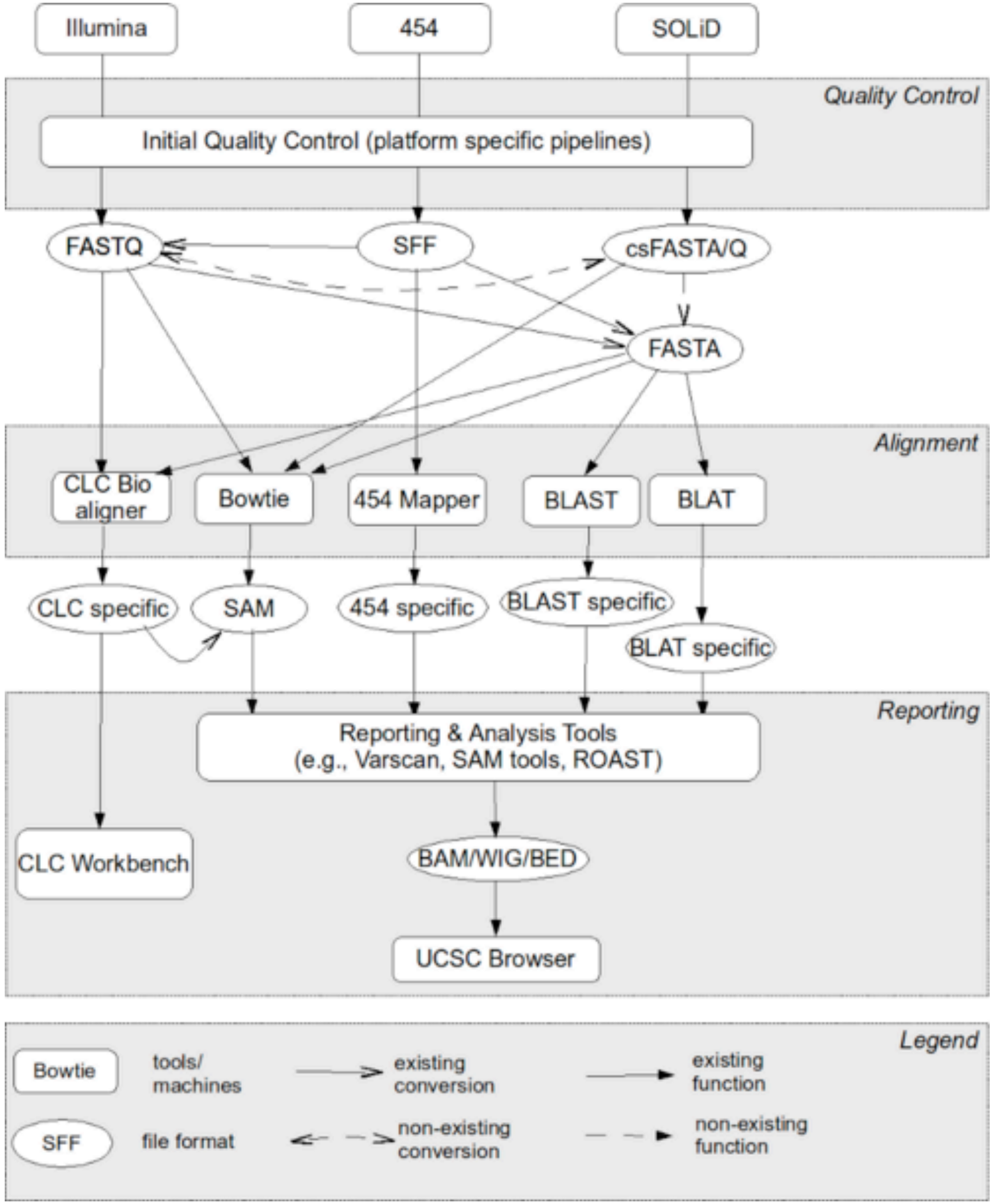
*A minimum of one read aligning to the National Center for Biotechnology Information build 36 reference genome. ‡Mappable reads for aligned assemblies. §Average read-length. ‖D. Wheeler, personal communication. ¶Reagent cost only. #S.-M. Ahn, personal communication. **K. McKernan, personal communication. ‡‡Tumour sample. §§Normal sample. ‖‖Tumour & normal samples: reagent, instrument, labour, bioinformatics and data storage cost, E. Mardis,

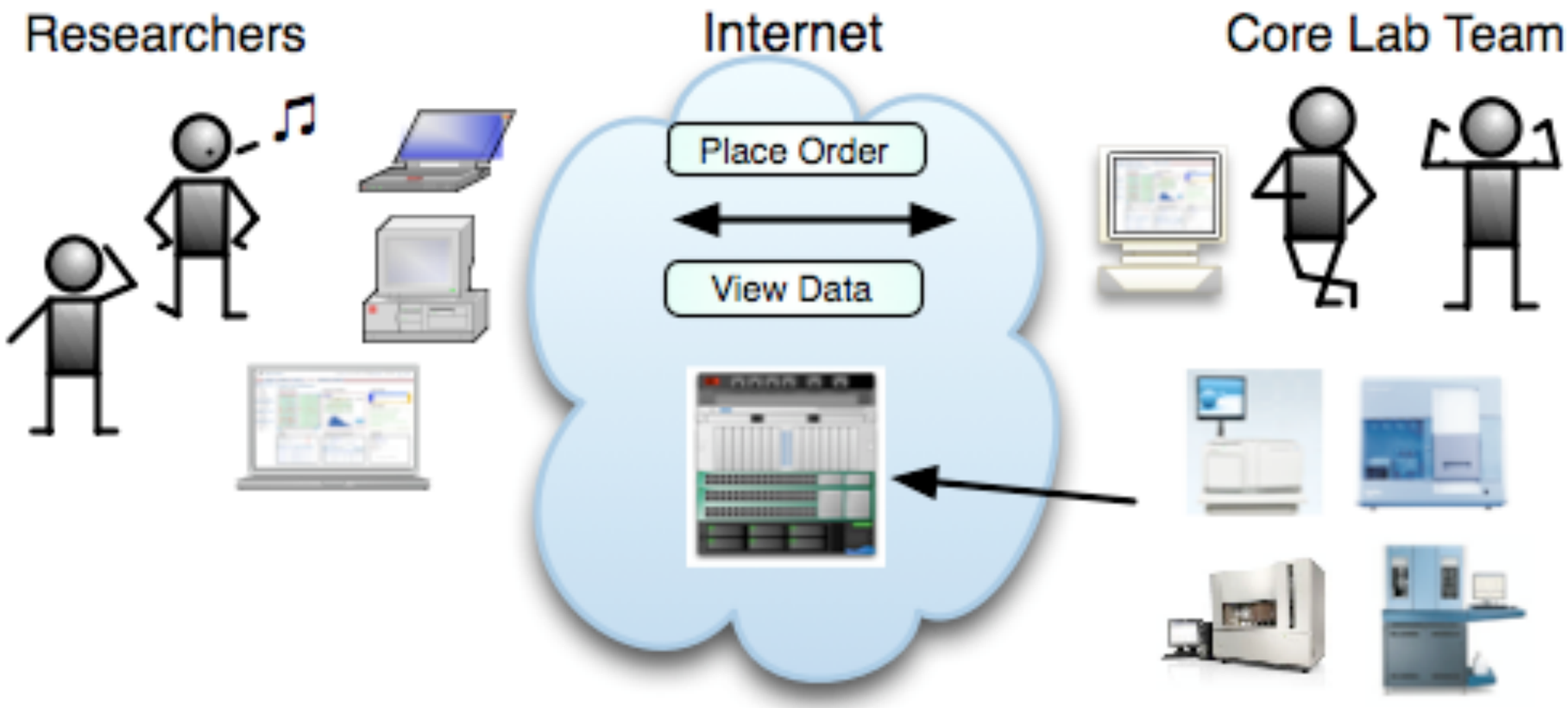
Sequencing tumour genomes



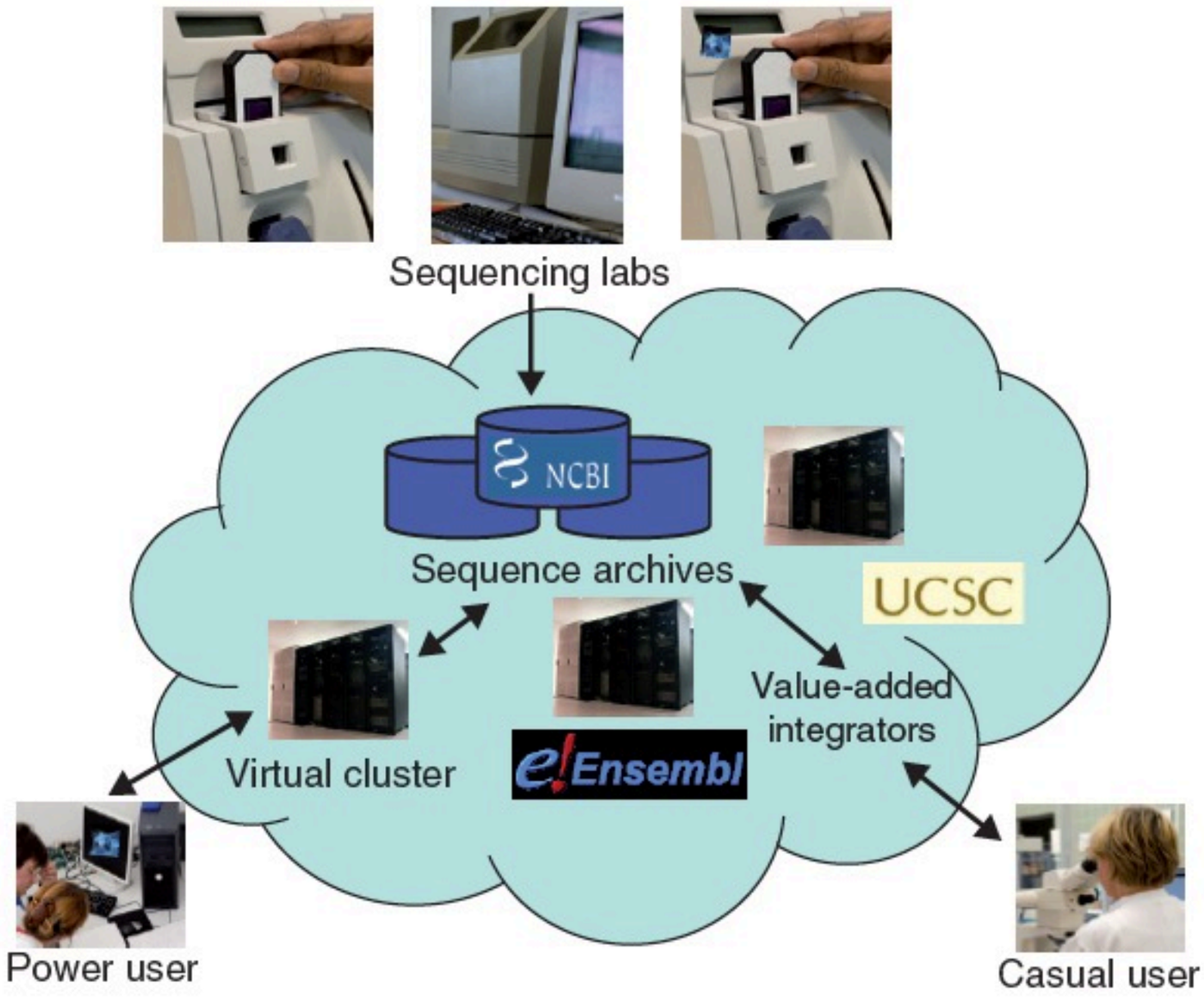
Straton et al., Nature. 2009 Apr 9;458(7239):719-24.


REFERENCE ALIGNMENT
DE NOVO ASSEMBLY
VARIANT DETECTION
CHIP-SEQ
RNA-SEQ


Illumina
454
SOLiD
Quality Control
Initial Quality Control (platform specific pipelines)
FASTQ
SFF
csFASTA/Q
FASTA
Alignment
CLC Bio aligner
Bowtie
454 Mapper
BLAST
BLAT
CLC specific
SAM
454 specific
BLAST specific
BLAT specific
Reporting
Reporting & Analysis Tools
(e.g., Varscan, SAM tools, ROAST)
CLC Workbench
BAM/WIG/BED
UCSC Browser
Legend
Bowtie
tools/ machines
existing conversion
existing function
SFF
file format
non-existing conversion
non-existing function



An additional challenge for research groups is dealing with high end computing infrastructures that are out of reach in terms of both cost and experience for most groups. Unlike the situation genome centers faced 10 years ago where such resources were best managed locally, NGS infrastructures, today are readily available through Internet-based resources. So called "Cloud Computing" resources provide storage, backup and computational resources using "pay-as-you-go" pricing models. These on-demand services, eliminate both IT management issues and costly mistakes that occur when local systems are under or over planned.



Stein LD. The case for cloud computing in genome informatics. Genome Biology 2010, 11:207 (5 May 2010)