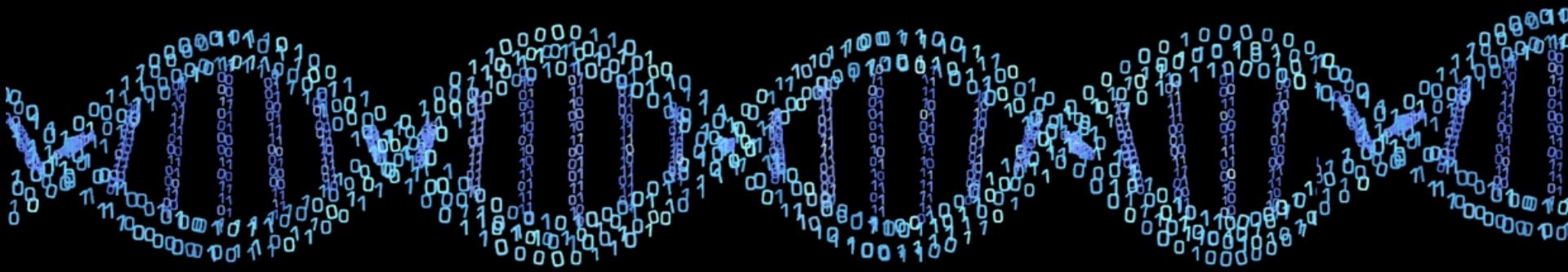
Introduction to Bioinformatics



2017

Instructors
Nicholas Navin
Ken Chen
Arvind Rao

Lecturers
Yiwen Chen
Han Liang
Wenyi Wang
Amin Momin
Rehan Akbani

Teaching Assistants
Aislyn Schalck
Visweswaran Ravi

IT Support
Alfred Valladolid
Fawad Jilani

Poll

#1
Raise your hand if you have previous programming experience

UNIX, Perl, Python, Matlab, R ?

#2
Raise your hand if you have experience analyzing next-generation sequencing data?

Course Outline

Module 1: Programming & UNIX
Module 2: Statistics & Probability
Module 3: Sequence Analysis & Next-Gen Sequencing
Module 4: Variant Detection, RNA & Phylogenetics
Module 5: Pathway Analysis & Inherited Mutations
Final Project

Format:

- Each class will consist of a 1-hour lecture and 2-hour workshop
- Workshops will be completed on the course server and/or using local computers
- PC users will need to use SSH software (ex. Putty) to log into the course server in order to complete the workshops

Grading

3 take-home projects:

HW1 – 20%
HW2 – 20%
HW3 – 20%

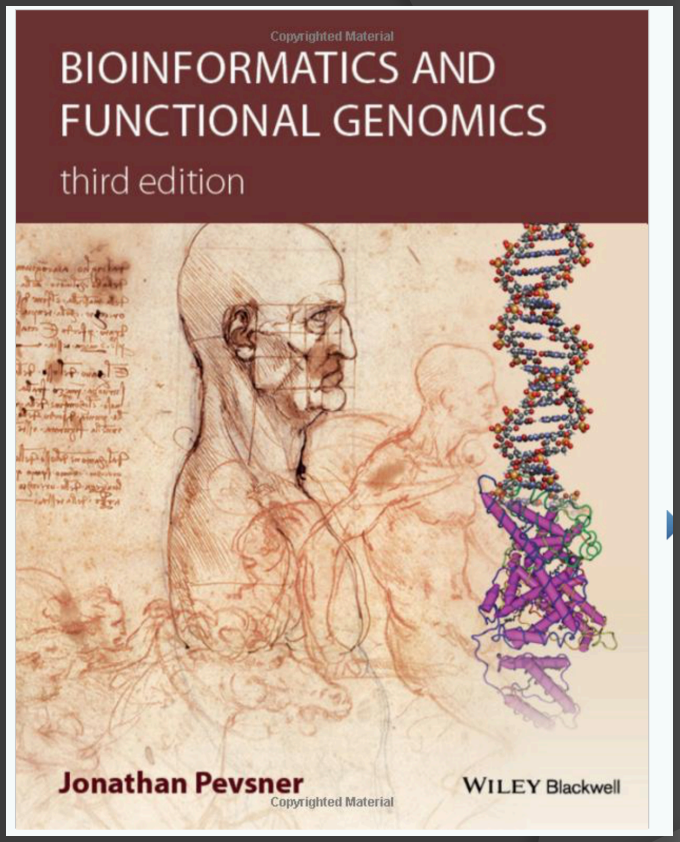
Final Project - 40% of the grade

Final Project

- Research Paper: 3 pages
- The final project will involve next-generation sequencing of a cancer cell line and analyzing the data to identify driver oncogenes and tumor suppressors that promote carcinogenesis

Homework Reading for Module 1

Book for the Course:
Bioinformatics and Functional Genomics,
3rd edition
Jonathan Pevsner



<u>Reading Assignments for First Module</u>

Introduction to Bioinformatics
Chapters 1-2

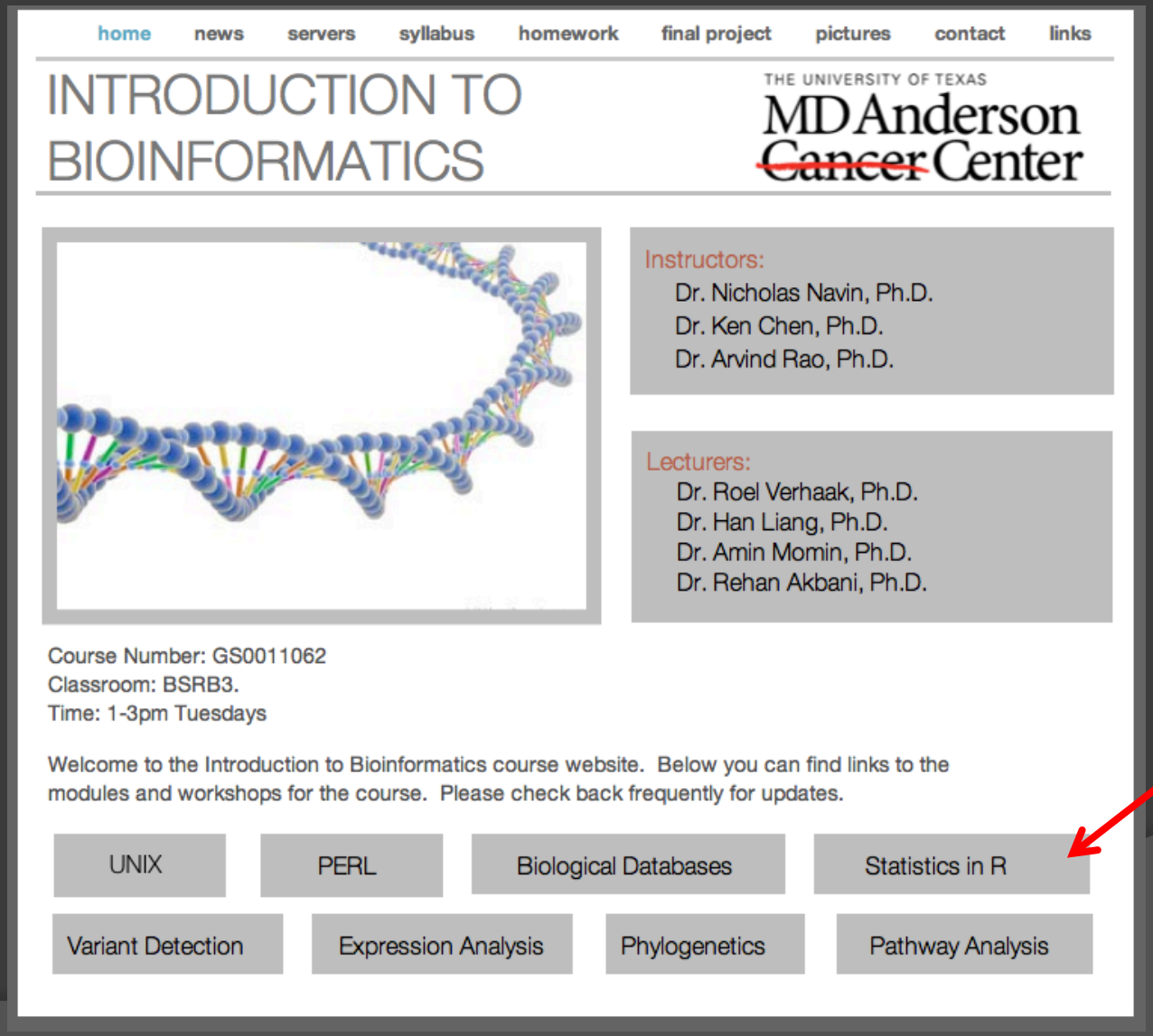
Eukaryotic Chromosomes
Chapter 8

Analysis of Next-Generation Sequencing Data
Chapter 9

The Human Genome
Chapter 20

Course Website

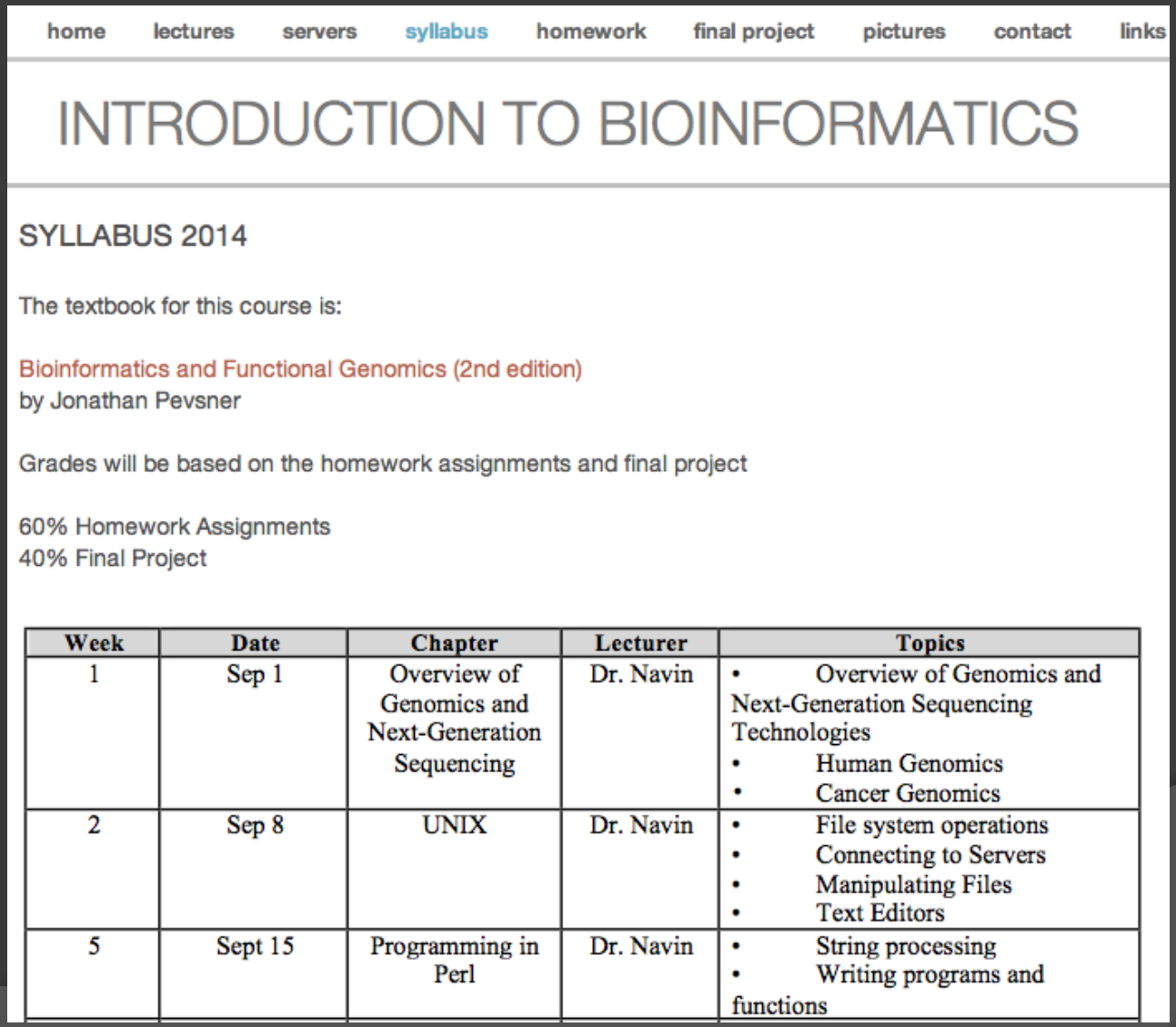
www.navinlab.com/bioinfo



Workshops from each module have links from the main webpage

Syllabus

http://www.navinlab.com/bioinfo/bioinfo/syllabus.html



home lectures servers syllabus homework final project pictures contact links

INTRODUCTION TO BIOINFORMATICS

SYLLABUS 2014

The textbook for this course is:

Bioinformatics and Functional Genomics (2nd edition)
by Jonathan Pevsner

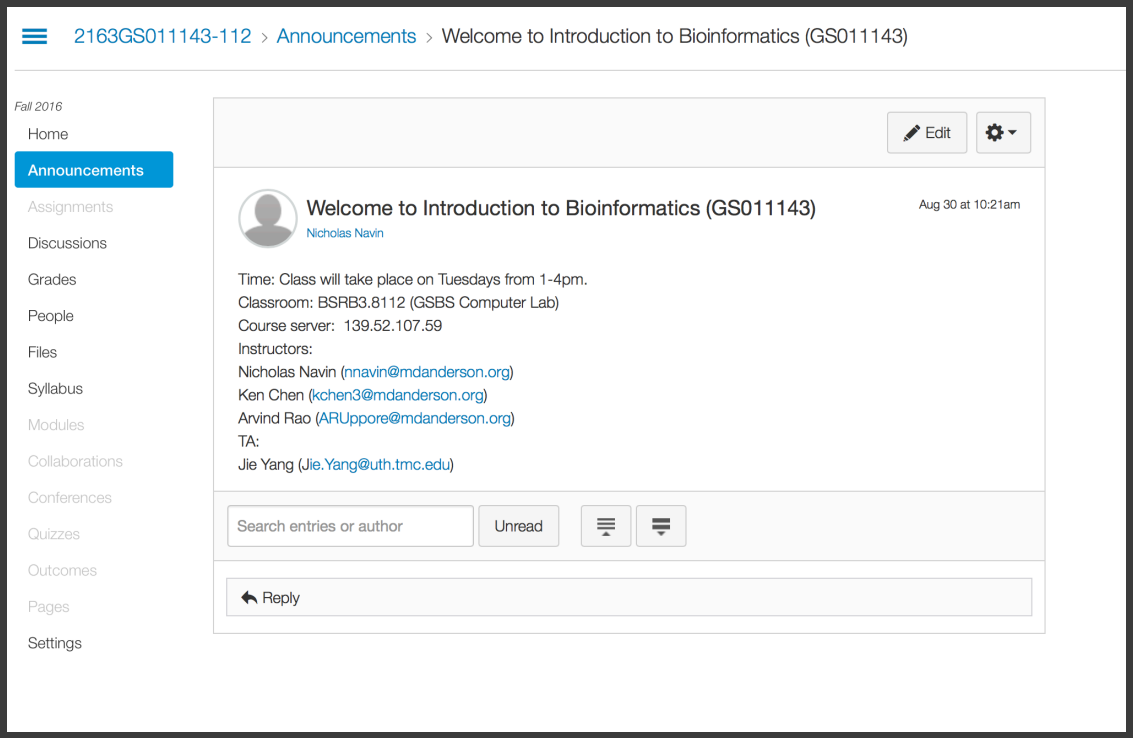
Grades will be based on the homework assignments and final project

60% Homework Assignments
40% Final Project

Week	Date	Chapter	Lecturer	Topics
1	Sep 1	Overview of Genomics and Next-Generation Sequencing	Dr. Navin	• Overview of Genomics and Next-Generation Sequencing Technologies • Human Genomics • Cancer Genomics
2	Sep 8	UNIX	Dr. Navin	• File system operations • Connecting to Servers • Manipulating Files • Text Editors
5	Sept 15	Programming in Perl	Dr. Navin	• String processing • Writing programs and functions

Bioinfo CANVAS

https://www.uth.edu/canvas



- Powerpoint Lectures
- Reading Materials (Review Papers)
- Questions and Discussions
- Weblinks to workshops
- Syllabus

Course Server

IP Address: **139.52.107.59**

note: the course server is on the **UT Health Network**

- We will pass around a contact information sheet, please check your name and email address

- USERNAME = first letter of first name + last name (lower case)
  Example: Nicholas Navin = nnavin

- PASSWORD = LAST NAME, please change after logging into the server with ‘passwd’ command

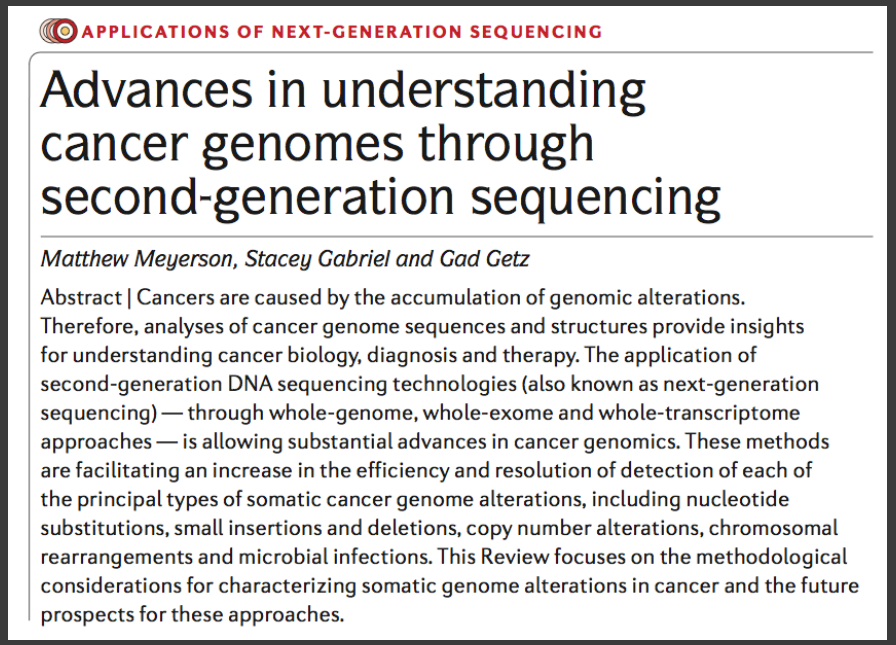
In order to connect to the server from the MD Anderson network or home, you will need to use a VPN client to connect to the UT Health network. Please follow these instructions to install Aventail Connect VPN: http://www.navinlab.com/bioinfo/bioinfo/servers.html

Introduction to Genomics

Nicholas E. Navin, Ph.D.
Associate Professor
Department of Genetics
Department of Bioinformatics

Homework Reading Assignments

- Review Papers can be downloaded from CANVAS



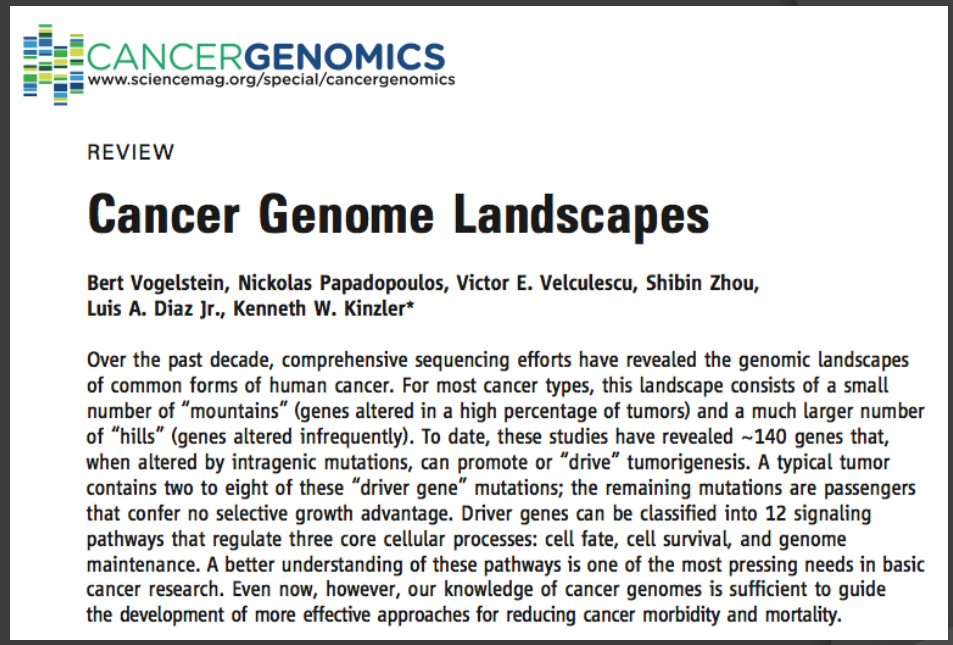
APPLICATIONS OF NEXT-GENERATION SEQUENCING

Advances in understanding cancer genomes through second-generation sequencing

Matthew Meyerson, Stacey Gabriel and Gad Getz

Abstract | Cancers are caused by the accumulation of genomic alterations. Therefore, analyses of cancer genome sequences and structures provide insights for understanding cancer biology, diagnosis and therapy. The application of second-generation DNA sequencing technologies (also known as next-generation sequencing) — through whole-genome, whole-exome and whole-transcriptome approaches — is allowing substantial advances in cancer genomics. These methods are facilitating an increase in the efficiency and resolution of detection of each of the principal types of somatic cancer genome alterations, including nucleotide substitutions, small insertions and deletions, copy number alterations, chromosomal rearrangements and microbial infections. This Review focuses on the methodological considerations for characterizing somatic genome alterations in cancer and the future prospects for these approaches.

Meyerson M, Gabriel S and Getz G
Nature Reviews Genetics 2010
PMID: 20847746



CANCERGENOMICS
www.sciencemag.org/special/cancergenomics

REVIEW

Cancer Genome Landscapes

Bert Vogelstein, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz Jr., Kenneth W. Kinzler*

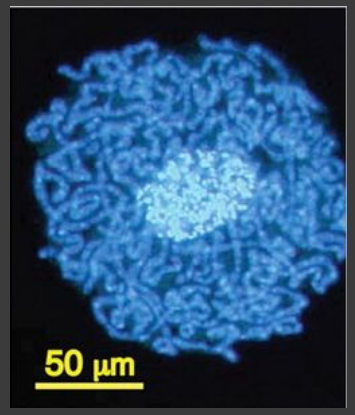
Over the past decade, comprehensive sequencing efforts have revealed the genomic landscapes of common forms of human cancer. For most cancer types, this landscape consists of a small number of "mountains" (genes altered in a high percentage of tumors) and a much larger number of "hills" (genes altered infrequently). To date, these studies have revealed ~140 genes that, when altered by intragenic mutations, can promote or "drive" tumorigenesis. A typical tumor contains two to eight of these "driver gene" mutations; the remaining mutations are passengers that confer no selective growth advantage. Driver genes can be classified into 12 signaling pathways that regulate three core cellular processes: cell fate, cell survival, and genome maintenance. A better understanding of these pathways is one of the most pressing needs in basic cancer research. Even now, however, our knowledge of cancer genomes is sufficient to guide the development of more effective approaches for reducing cancer morbidity and mortality.

Vogelstein B, Papadopoulos N, Velculescu V, Zhou S, Diaz L, Kinzler K.
Science 2013
PMID: 23539594

Genomics

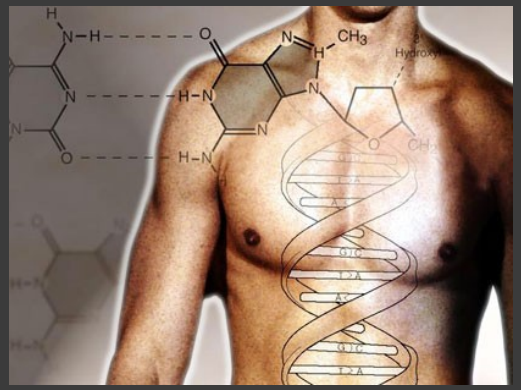
- The word genome was coined in the 1920's by Hans Winkler (Botanist)
- Genome = fusion of the words gene & chromosome
- Genomics involves the quantitative analysis of genome-wide DNA, RNA and Epigenetics using technologies such as microarrays and sequencing
- Different organisms show a large range in genome size and chromosome numbers

smallest genomes (viruses/bacteria)



Carsonella ruddii
182 genes
159,000 bases

human genome



Homo sapien
~25,000 genes
3.18 billion bases

largest genomes (plants)

Paris Japonica
not sequenced
132 billion bases

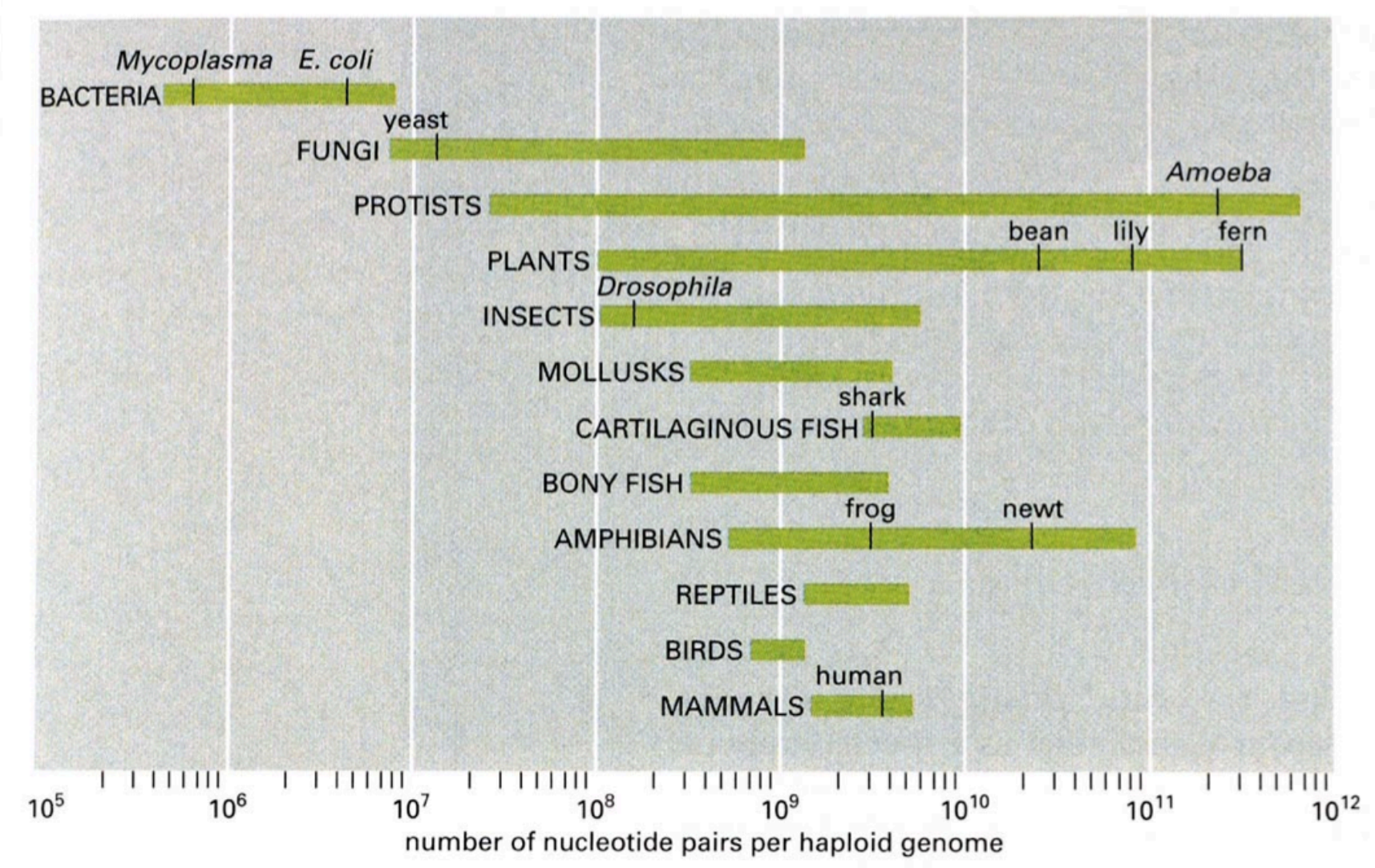
Differences in Genomes

How do genomes differ?

- GC Content
- Chromosome Numbers
- Genome Size
- Repetitive Elements
- Gene Density (% Coding Regions)
- Intron structure

Genome Sizes in Different Organisms

Genome Sizes in Different Organisms



There is no correlation between genome size and organism complexity

Molecular Biology of the Cell (MBOC)

Chromosome Numbers Differ Across Species

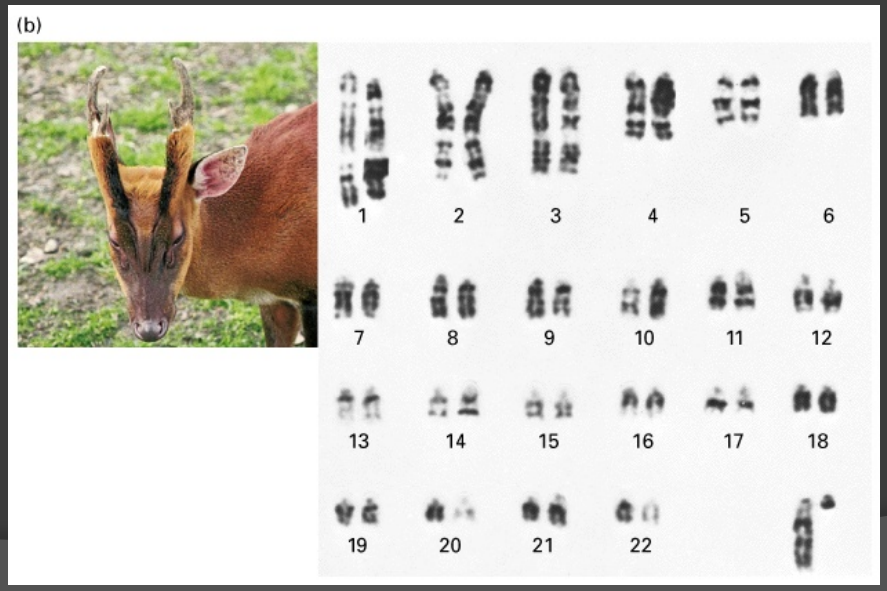
Chromosome Numbers	
Adders-tongue Fern	1440
Rattlesnake Fern	184
Carp	104
Aquatic Rat	92
Shrimp	90
African Hedgehog	90
Pigeon	80
Dog	78
Horse	64
Platypus	52
Pineapple	50
Human	46
Blue Whale	44
Mouse	40
Yeast	32
Pill millipede	30
Zebrafish	26
Rice	24
Cannabis	20
Fruit Fly	8
Mosquito	6
Jumper Ant	2

There is no clear correlation between the number of chromosomes and species complexity

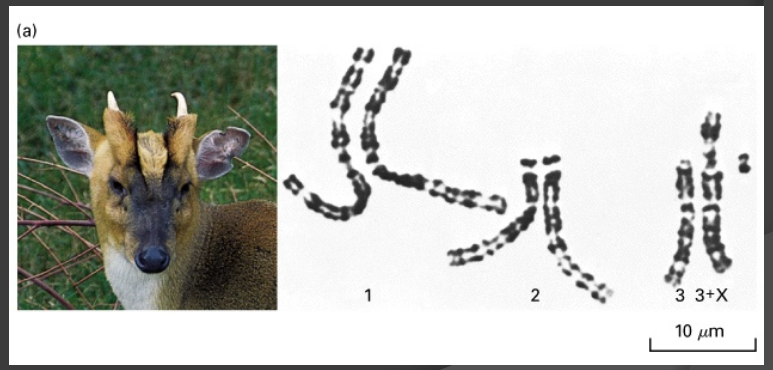
Chromosome Numbers

- Even closely related species may have vastly different chromosome numbers
- The Chinese muntjac (A) has 46 chromosomes, while the Indian Muntjac has only 7 chromosomes
- The Chinese muntjac evolved from a common ancestor of the Indian muntjac by fusing its chromosomes, which did not have a major phenotypic effect
- Both species have approximately the same number of genes

Chinese Muntjac



Indian Muntjac



Genome Characteristics of Different Organisms

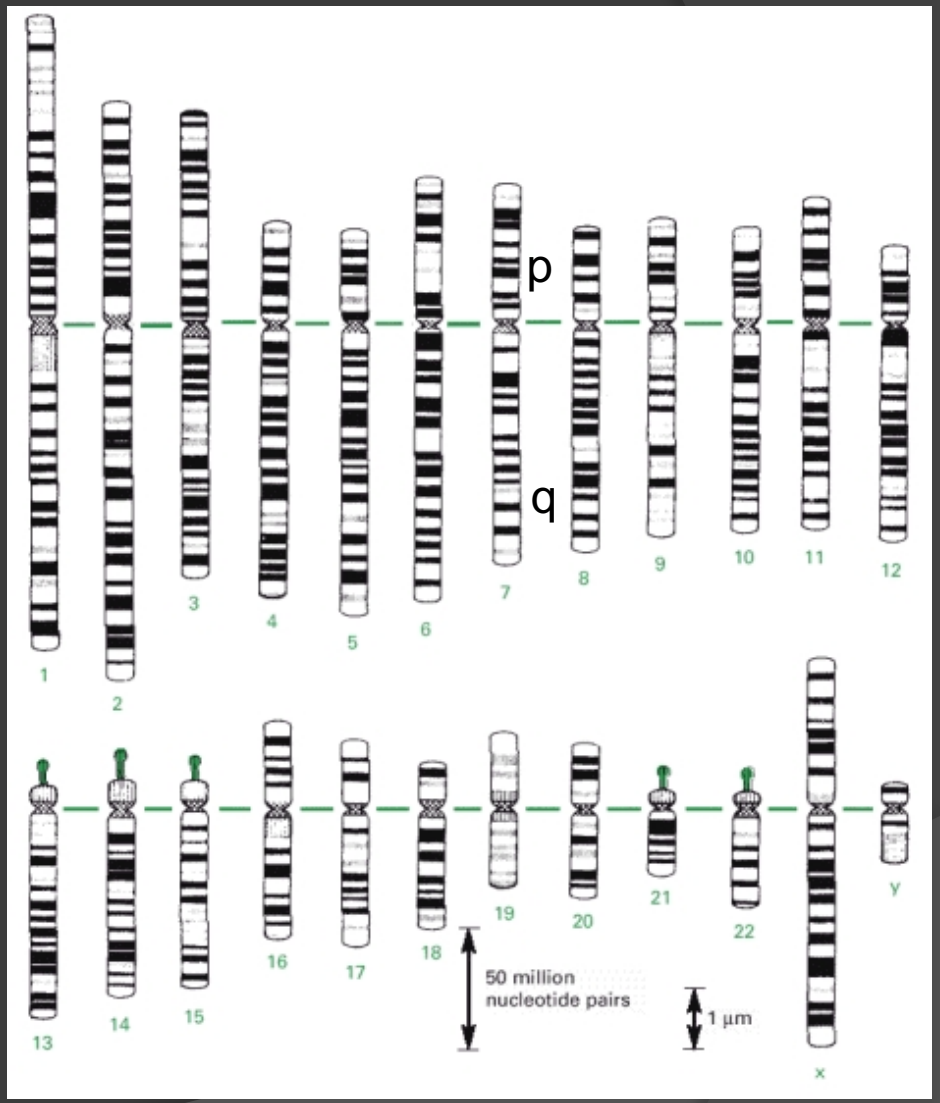
TABLE 16-1 Features of Several Sequenced Bacterial and Eukaryotic Genomes

Feature	*E. coli* K-12	Parasite[a]	Yeast[b]	Slime Mold[c]	Plant[d]	Human
Genome size, Mb	4.64	22.8	12.5	8.1	115	3289
GC content, %	50.8	19.4	38.3	22.2	34.9	41
Number of genes	4288	5268	5770	2799	25,498	20,000–25,000
Gene density, kb per gene	0.95	4.34	2.09	2.60	4.53	27
Percent coding	87.8	52.6	70.5	56.3	28.8	1.3
Number of introns	0	7406	272	3578	107,784	53,295
Repeat %	<1	<1	2.4	<1	14	46

Chromosome Maps

- First genomic maps were constructed by staining chromosomes with Giemsa (G-banding) which stains regions dark or light depending on the GC% content
- **Dark regions** have low GC content and are gene-poor, while **light regions** are gene-rich and have high GC content
- P-arm is the short arm ('petit') and Q-arm is the long arm
- Physical locations of genes on chromosomes are described using cytogenetic nomenclature:

  Ex. *KRAS* is located on 12p12.1
- Modern nomenclature is:
  Chr12: 25278036-25295130

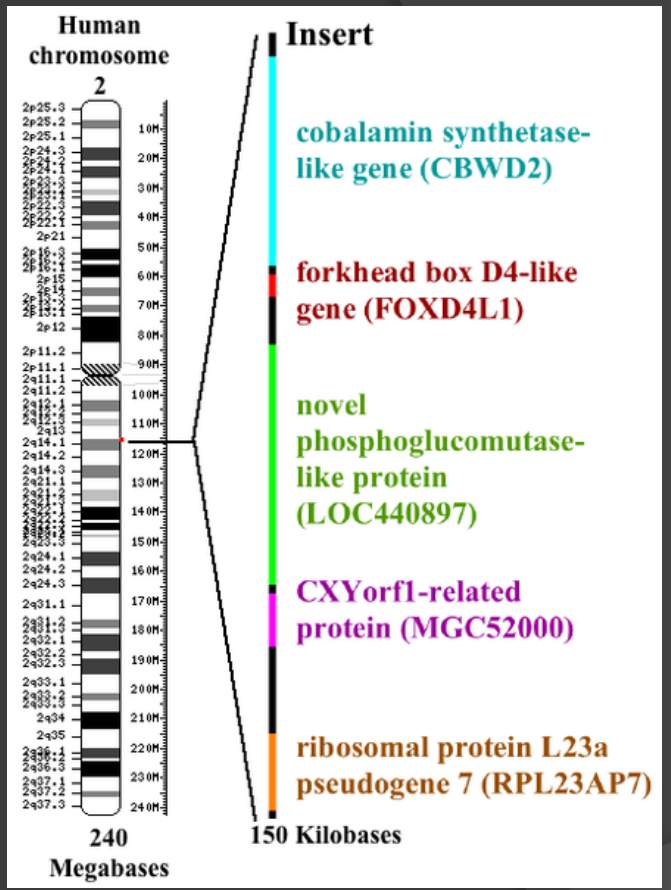


Ideogram – representation of a karyotype

Human and Primate Chromosome Evolution

- The difference between any two human genomes is about 0.1%
- The difference between a human and chimpanzee genome is 1-2%
- Most variants are SNPs, however when primates evolved into humans (via a common ancestor) there was a major chromosome rearrangement
- Human chromosome 2 evolved from the fusion of two primate chromosomes (2A and 2B)

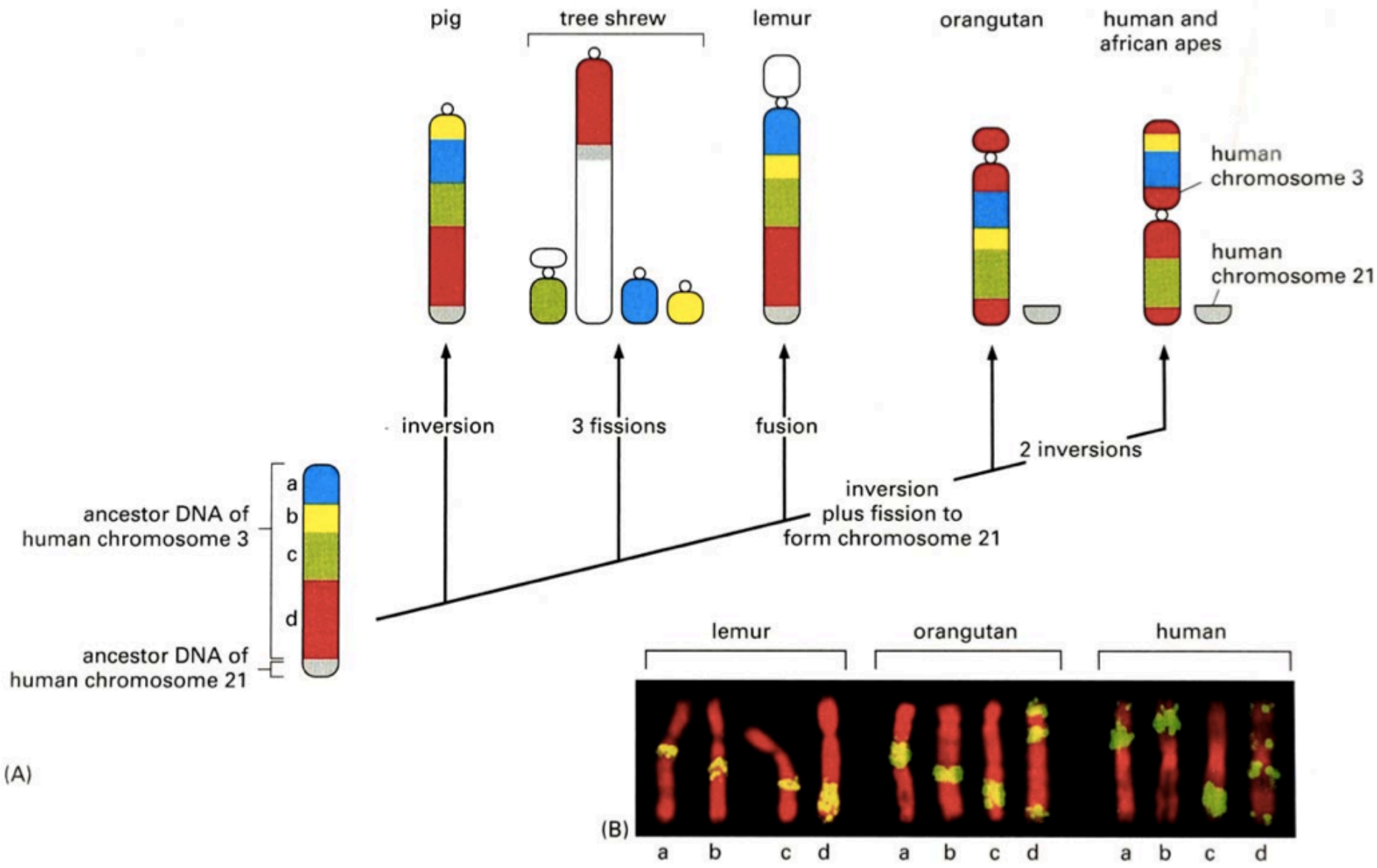
Fusion of Chromosome 2



New Genes at the 150kb fusion Site

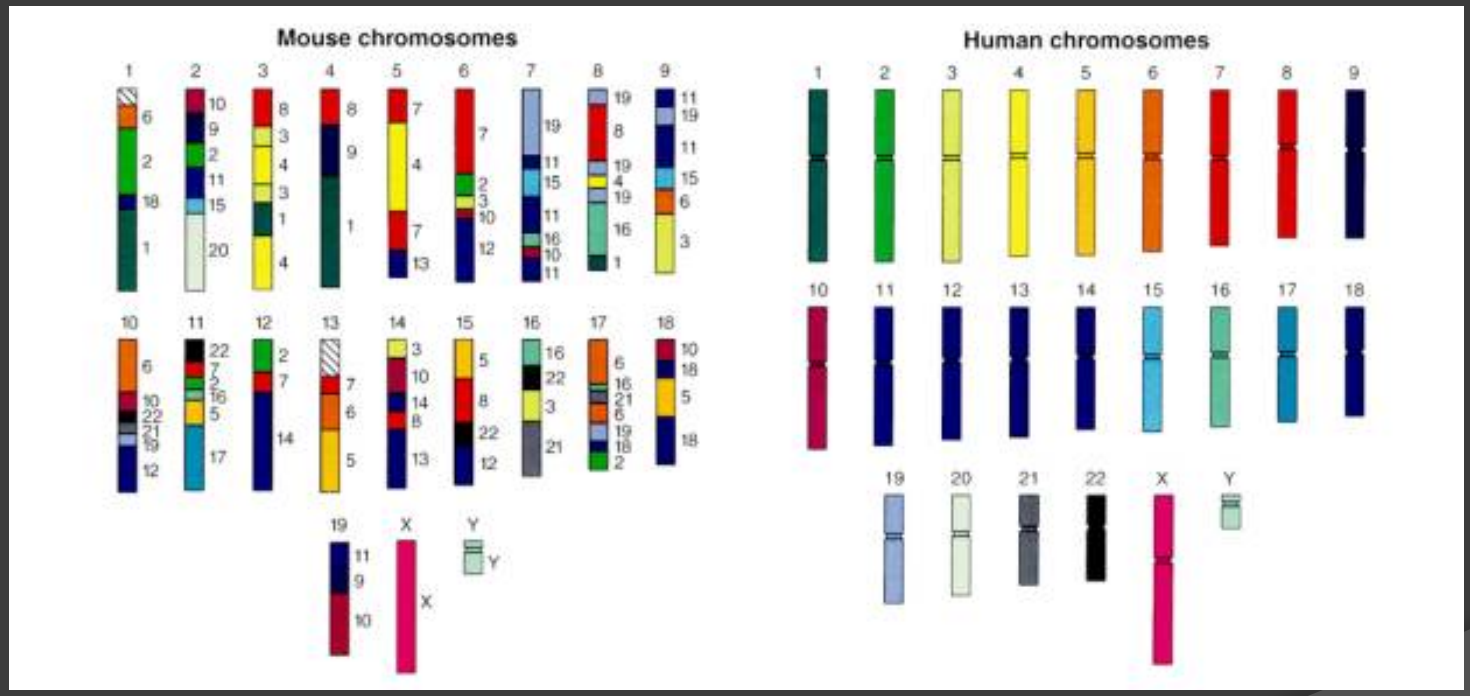
- The fusion event resulted in a 150kb stretch of new genes that may have played an important role in evolving human characteristics (ex. FOXD4L1 is a major transcription factor in human embryogenesis)

Evolution of Human Chromosomes 3 and 21



Molecular Biology of the Cell

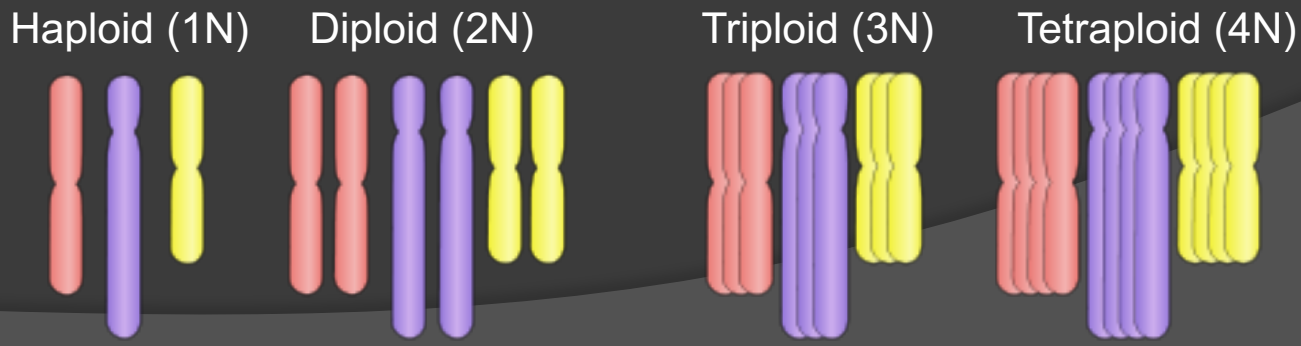
Mouse and Human Genomes



- All mouse chromosomes are Acrocentric (centromere at one end)
- Mouse and human genomes have about 90% sequence homology and roughly the same number of genes (~25,000)
- The mouse genome has many syntenic blocks of homology

DNA Content - Ploidy

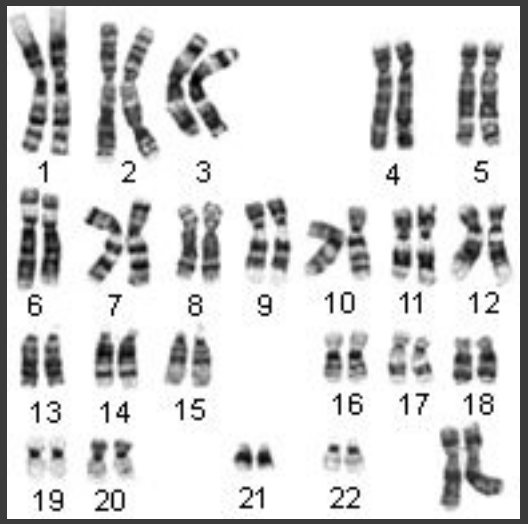
- Most mammalian genomes are diploid (2N) and contain two pairs of each chromosomes
- Gametes such as sperm and oocytes have haploid (1N) copy number
- Many plant genomes are polyploid (tetraploid 4N, hexaploid 6N or octoploid 8N)
- Euploid genomes have a normal balanced set of chromosomes (ex. 23 pairs in humans)
- Aneuploid genomes have an unbalanced number of chromosomes, and are often associated with genetic diseases and cancer
- Most human cells are diploid but there are some notable exceptions:
  1 – Liver cells fuse to become polyploid
  2 – Megakaryocytes cells in the bone fuse to become 8N – 16N
  3 – Smooth muscle cells fuse and become tetraploid



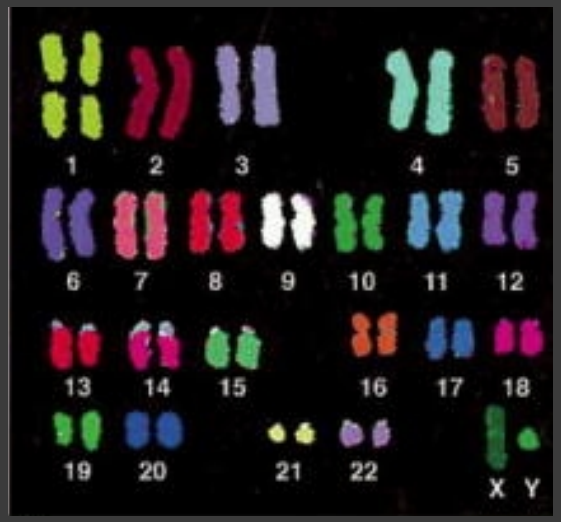
Genomic Methods & Technologies

Cytogenetic Tools

- Cytogenetic methods were among the first tools that enabled the genomic analysis of chromosomes in normal and diseased patients
- Karyotypes can be used to enumerate chromosome
- Spectral karyotyping (SKY) uses fluorescently labeled DNA probes to color chromosomes and detect chromosomal aberrations in cancer



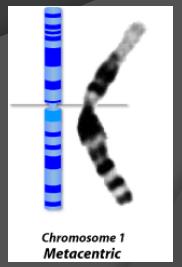
Female Karyotype

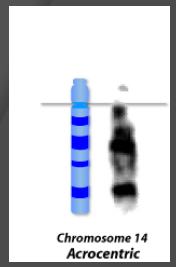


Male SKY

Chromosome Structure:

- Metacentric – centromere is in the middle
- Acrocentric – centromere is at the end of chromosome





Microarrays

- Microarrays are glass slides coated with thousands of oligonucleotide probes
- DNA or cDNA is labeled with fluorophores, hybridized to the probes and imaged with a scanner

Applications

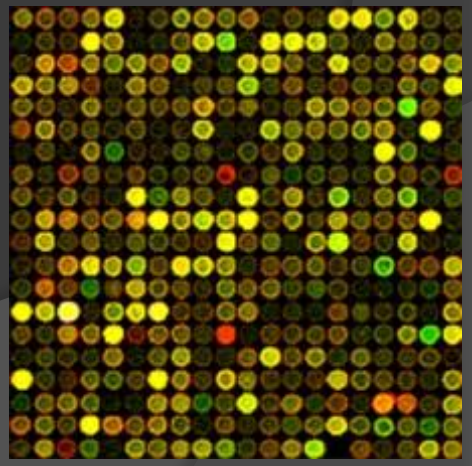
- Copy Number Profiling
- RNA Expression Analysis
- SNP Microarrays (Genome-Wide Association Studies)
- Chip on Chip (Chromatin Immunoprecipitation)

Advantages

- Cheap
- Readily available software for analysis

Disadvantages

- Data has low dynamic range
- New transcripts cannot be discovered
- Labor intensive, multiplexing samples is difficult



Proteomic Platforms

Reverse-Phase-Protein Array (RPPA)
Mass-Spectrometry (MS)
Cy-TOF (flow-sorting mass spec)

Applications

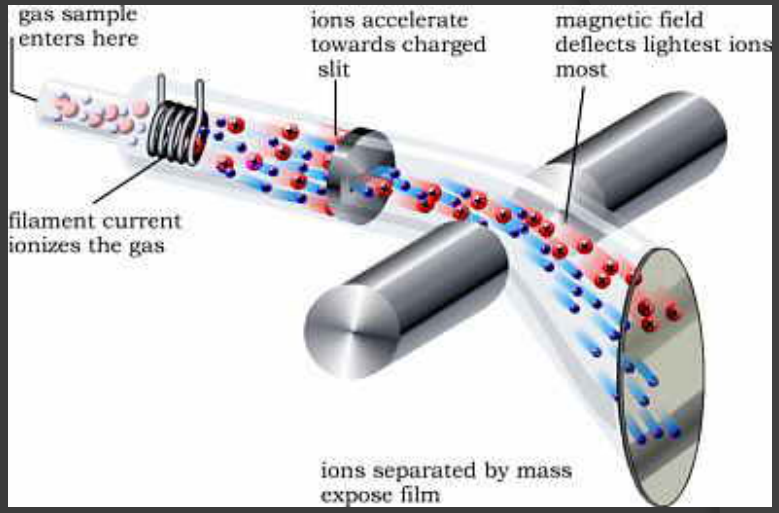
- MS can discover unknown proteins
- Measure protein levels
- Measure protein modifications (phosphorylation)

Advantages

- DNA and RNA levels may not reflect protein levels
- Protein activity from post-translational modifications can be measured

Disadvantages

- RPPA and Cy-TOF are highly dependent on antibody quality
- Cy-TOF data analysis is challenging

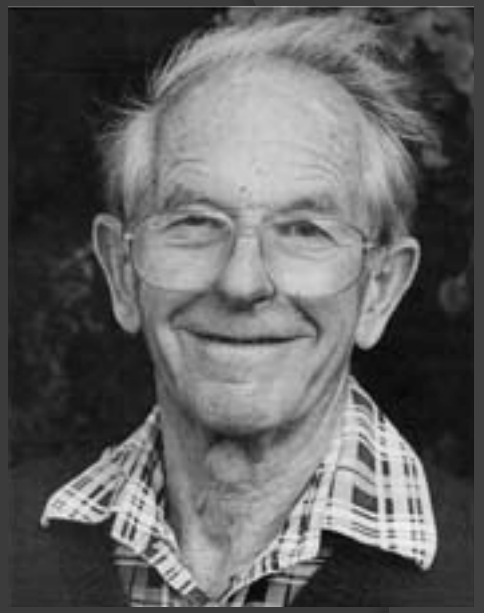


Mass-Spec

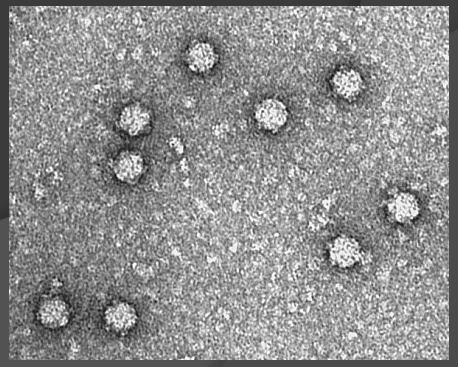
RPPA

First-Generation ‘Sanger’ DNA Sequencing

- Fred Sanger developed the first DNA sequencing method in the 1970’s using dideoxy termination chemistry
- Fred used this method to sequence the first genome: 5386 nucleotides of the PhiX (Phi174) bacteriophage, and was awarded the Nobel prize for his work
- The initial method used radioactive isotopes (P32) and polyacrylamide gels to detect nucleotide sequences
- Modern capillary sequencers use fluorophores and can sequence 700 DNA bases in each reaction
- The sanger sequencing method was the workhorse of the human genome project and is still used frequently today for targeted sequencing and validation of cloning



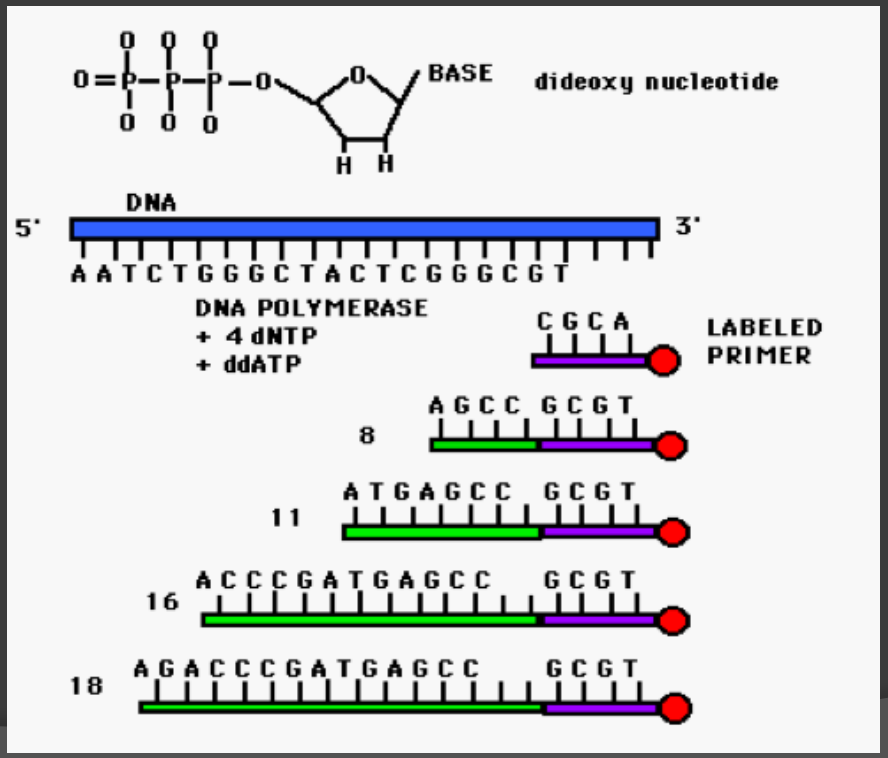
Fred Sanger is the 4th person to receive 2 nobel prizes



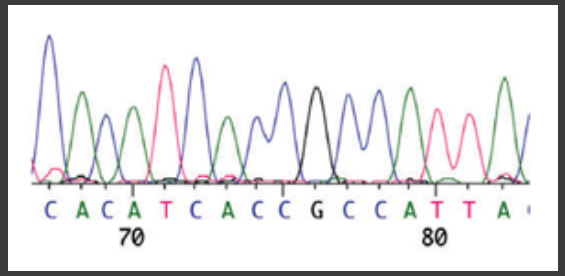
PhiX (174) Bacteriophage

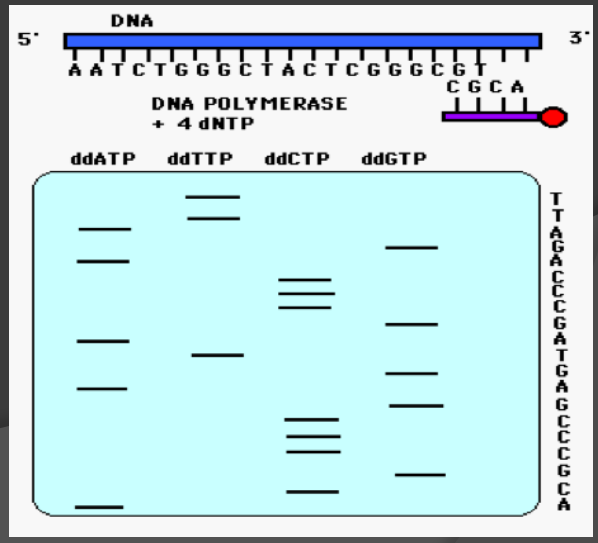
Sanger Sequencing

- Mix 1% Dideoxy nucleotide (ddNTP) with other dNTPs, which causes termination of the strand during polymerization,
- Perform 4 reactions, each using a different ddNTP
- Run the DNA products out on a polyacrylamide gel or using a fluorescence detector on a modern capillary sequencer (ABI) to produce a trace file



ddNTP lacks 3' OH group



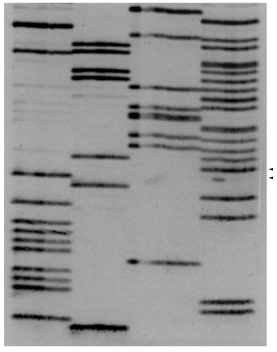


First-Generation Sequencing

1920 — Hans Winkler coins the term ‘genome’

1953 — Watson &Crick discover the structure of DNA

1975 — Fred Sanger develops dideoxy chain termination methods and sequences first genome of bacteriophage X174 which is 5368bp



1983 — Kerry Mullis invents PCR

1990 — Congress Approves Funding for Human Genome Project over next 15 years

1996 — *E.Coli* genome is sequenced

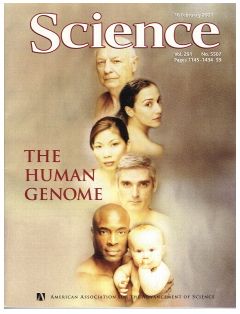
First free-living organism is sequenced *Heamophilus Influenzae*

1998 — *C. Elegans* Genome is Sequenced

ABI develops automated system for Sanger Sequencing PRISM3700

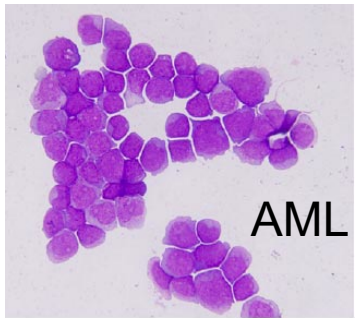
1999 — Drosophila genome is sequenced

2001 — Draft Human Genome sequence is finished



2002 — Mouse genome is sequenced

Next-Generation Sequencing



1992 — Sydney Brenner develops 'Massively Parallel Signature Sequencing' company Lynx Therapeutics

1998 — Solexa is formed and begins developing reversible terminator chemistry

2004 — Solexa and Lynx Merge

2005 — Illumina purchases Solexa; 454 develops massively parallel pyro sequencer

2006 — Illumina Releases GA2 Next-Generation Sequencer

2007 — Venter Genome is Sequenced

2008 — Watson Genome is sequenced in 4 months using 454; WashU sequences the first cancer genome

2009 — Vancouver publishes first breast cancer genome

2010 — Illumina Releases HiSeq2000 which can sequence a human genome in a week

2011-2016 — Many Large Cohorts of Cancer Patients Are Sequenced by TCGA, Broad, WashU, Sanger

Next-Generation Sequencing (NGS)

Platforms

- Illumina (Hiseq, GA2, miseq)
- LIFE technologies (Ion Torrent, Proton)

Applications

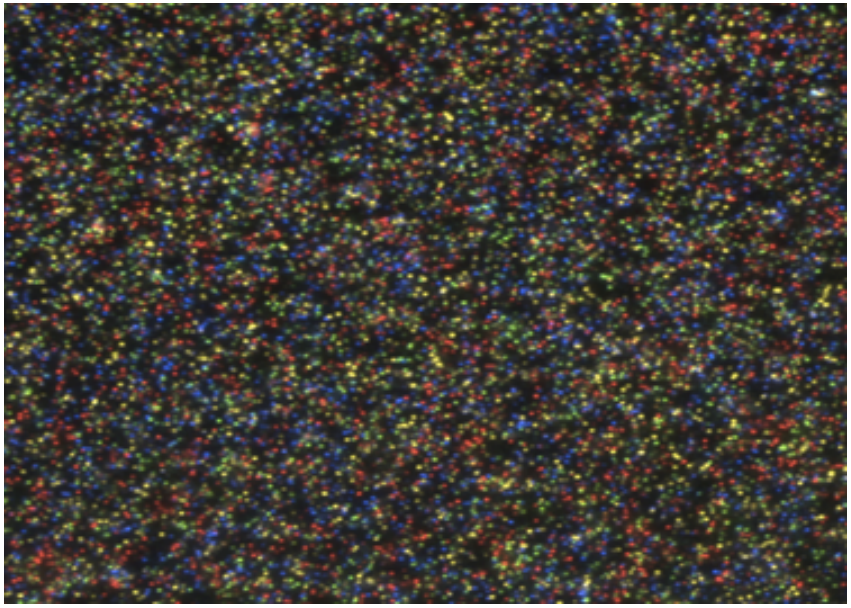
- RNA-Seq (expression, alternative splicing)
- DNA-Seq (genome, exome, panels)
- Chip-Seq
- Bisulfite-Seq

Advantages

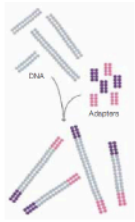
- Many different mutation types can be measured:
  - point mutations
  - Indels (small insertions and deletions)
  - Structural variants
  - Copy number aberrations
  - Transcript expression
  - Alternative splicing
  - DNA –protein interactions
- Novel transcripts can be detected

Disadvantages

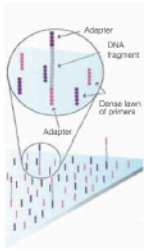
- Expensive
- Large data files need to be stored
- Data analysis is challenging



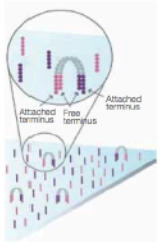
Illumina Flow-cell



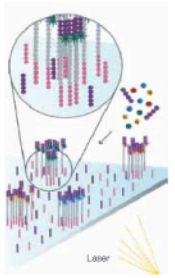
1. Prepare



2. Attach



3. Amplify



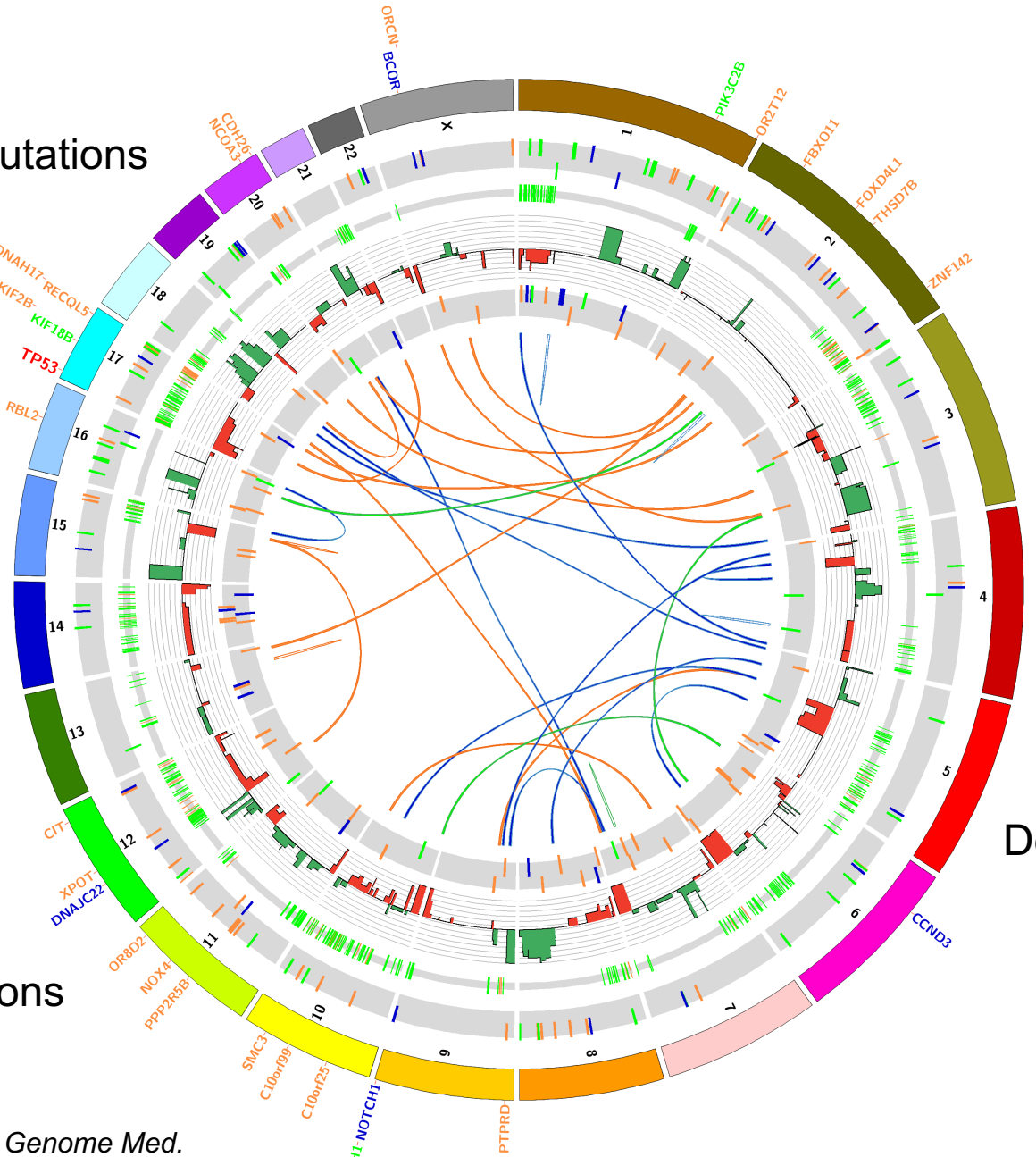
4. Image

Illumina NGS

DNA Mutations Detected in a Single Cancer Genome

Circos Plot

Triple-Negative Breast Cancer Genome

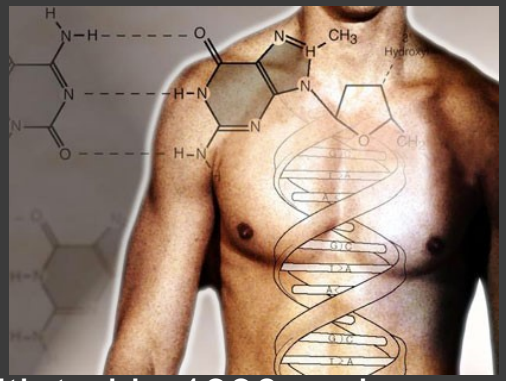


Molhatra and Wang et al. 2015, Genome Med.

The Human Genome

The Human Genome

- Human Genome consists of 22 pairs of autosomes and 1 pair of sex chromosomes (X,Y)
- Length of human genome is 3.18 gigabases (billion basepairs)
- Human genome consists of (approximately):
  - 25,000 genes
  - 100,000 proteins
  - 8000 lincRNAs (>200bp)
  - 1000 microRNAs



- Average size of a gene is 27,000bp
- The public human genome project was formally initiated in 1990 and headed by James Watson who received funding support from Congress
- The public human genome project involved sequencing 4 individual genomes of central European decent
- The public human genome project was the largest project in the field of biology and involved 18 Countries and cost $3 billion US tax dollars and over 20 years to complete

The Race to Sequence the Human Genome

- The public human genome project began in the 1980s and used Sanger sequencing of bacterial artificial chromosomes (BACs)
- In 1998 Craig Venter established a private company called Celera in which he announced that they would used ‘shotgun sequencing’ to beat the public project and sequence the human genome in 3 years.
- This lead to fierce race between the public and private sectors
- In the end both parties agreed to a tie and the draft human genome was announced by President Clinton in 2000 and published concurrently in *Science* and *Nature*. The final version was completed in 2003

Eric Lander

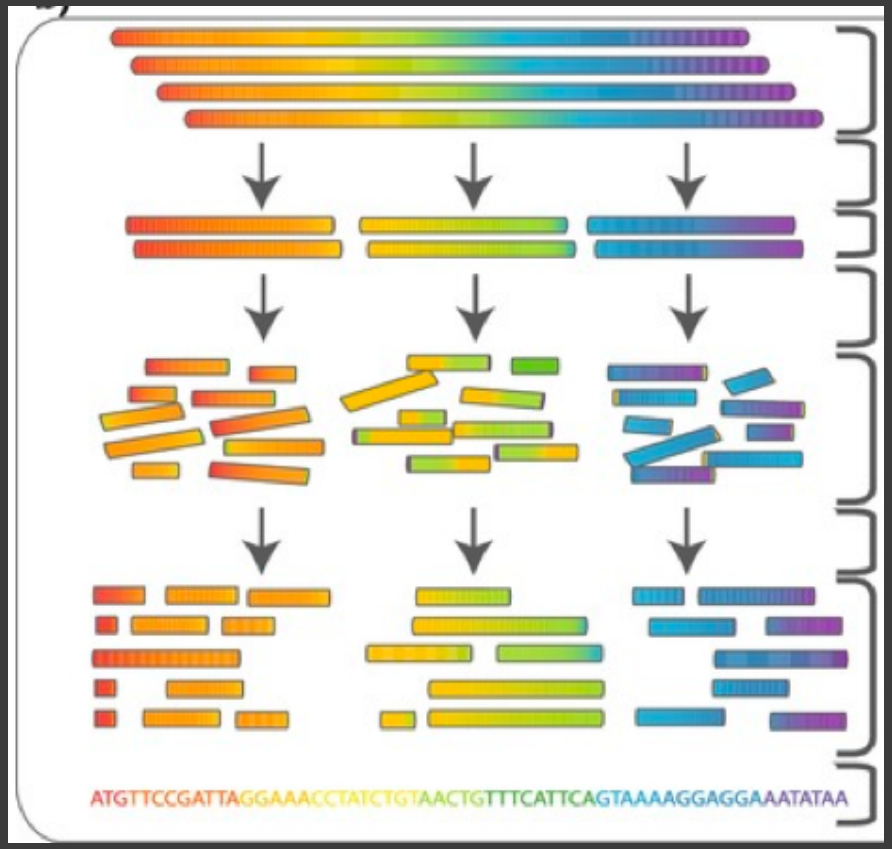
Francis Collins

Bill Clinton Announces the Completion of the Human Genome Project

Craig Venter

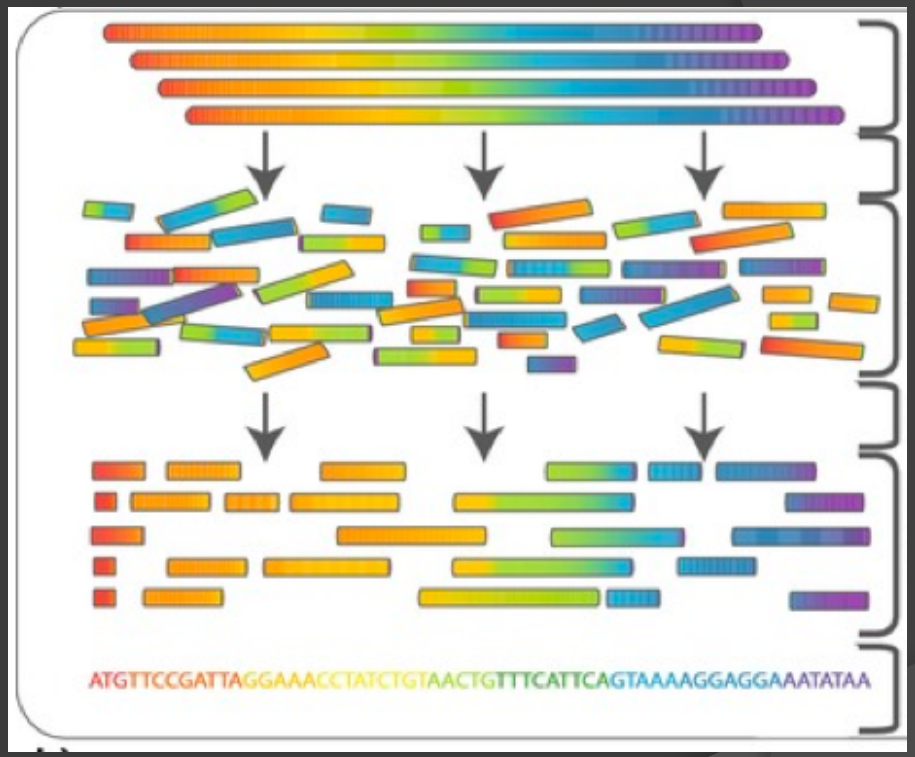
The Human Genome Project

Positional BAC Cloning



Advantages: More accurate assembly in repetitive regions of the human genome
Caveats: slow and laborious

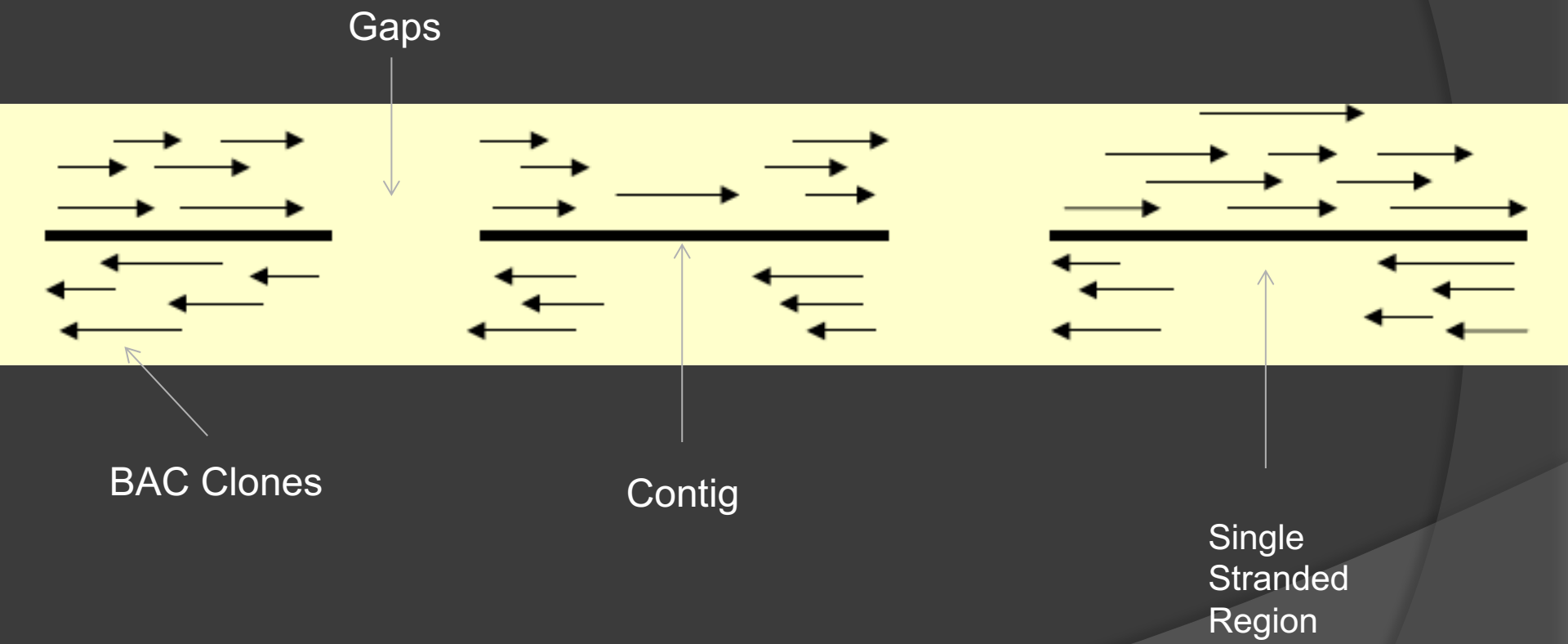
Shotgun Sequencing (Venter)



Advantages: Fast method and less labor
Caveats: Computationally very intensive to perform assembly

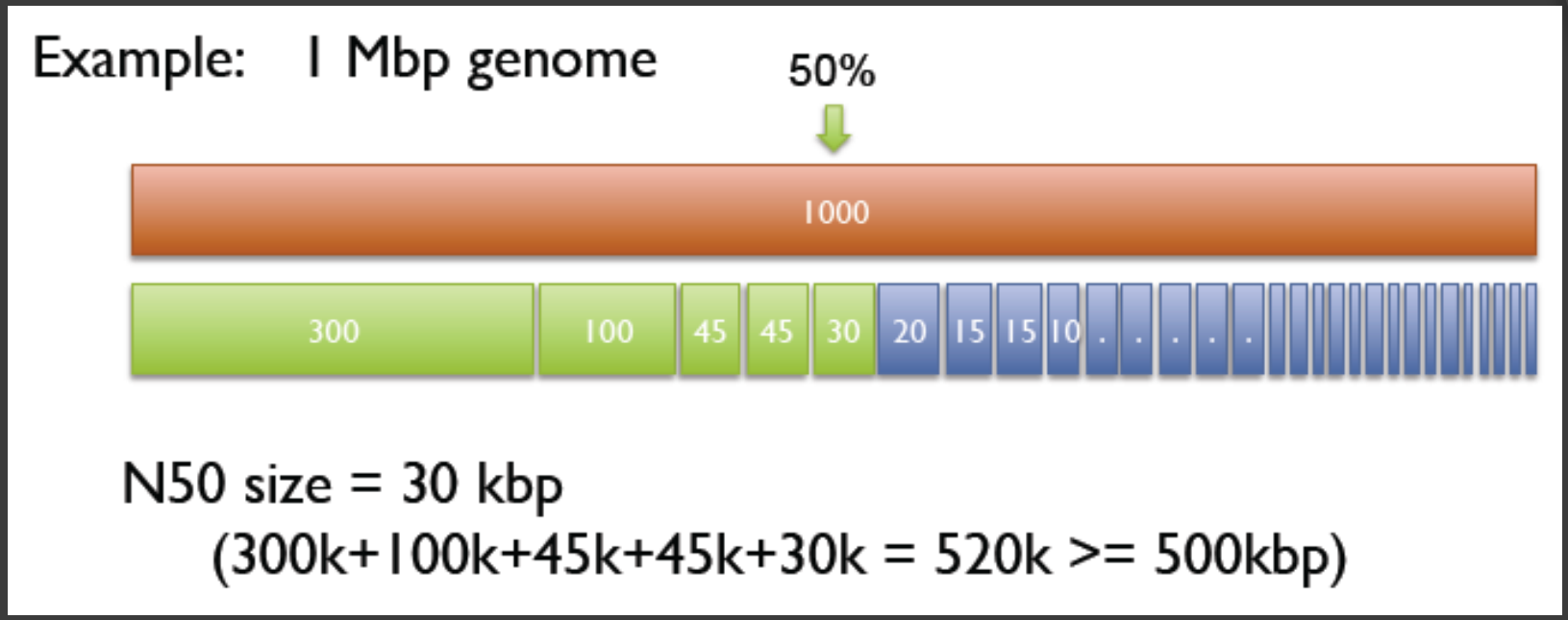
Genome Assembly

Draft Genome : 5-10X coverage depth
Final Assembly : 50 - 100X coverage depth



Genome Assembly

N50 is a metric of genome assembly performance



N50 Size – 50% of the genome is in contigs larger than this size

Accessing Human Genome Data

- Accessing the Human Genome Assembly Data:

  UCSC Human Genome Browser http://genome.ucsc.edu/
  NCBI Human Genome http://www.ncbi.nlm.nih.gov/
  Ensemble Genome Browser http://www.ensemble.org

- Several different genome assemblies are available:

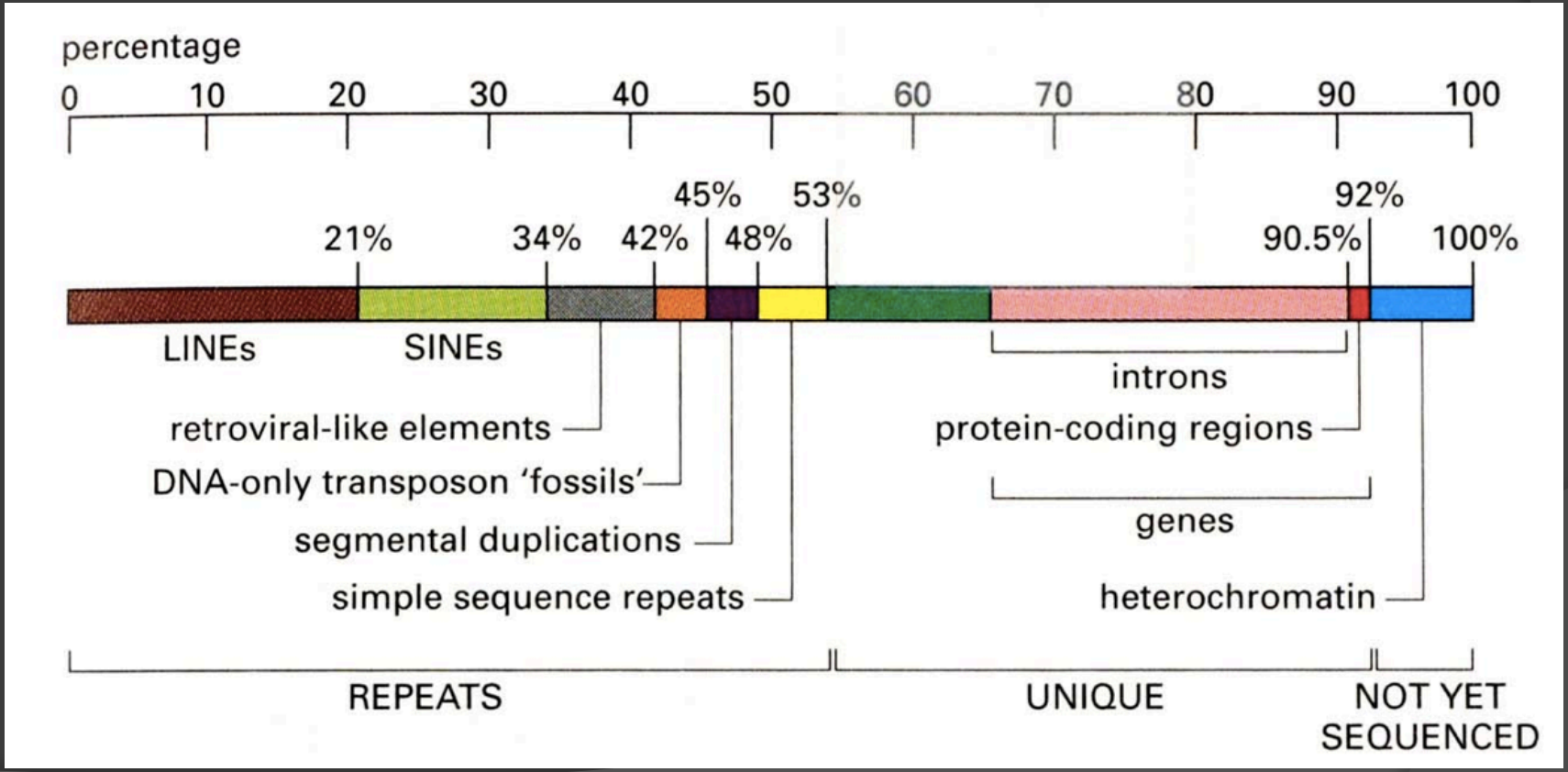
  HG18 (NCBI36) released March 2006
  HG19 (GrCH37) released Feb 2009
  HG20 (GrCh38) released 2013

- Which assembly to use?

  Currently HG19 has the most annotation tracks and is generally accepted by journals as the standard, however many scientist are now moving to HG20

Composition of the Human Genome

- Only 1% of the human genome contains coding regions (exome)
- About 24-36% of the human genome consists of introns
- The other 50% of the human genome consists of repetitive elements



Normal Genomic Variation in the Human Genome

- **Single Nucleotide Polymorphisms (SNPs)**
  -on average each human has 1 SNP every 1 kb
  -however African natives have 1 SNP every 750bp

SNP

AACTGTTGCTAGCC
TTGACAACGATCGG

AACTGTGGCTAGCC
TTGACACCGATCGG

- **Insertion-deletions (Indels)**
  - insertion or deletions of 1-100 base pairs

Indel

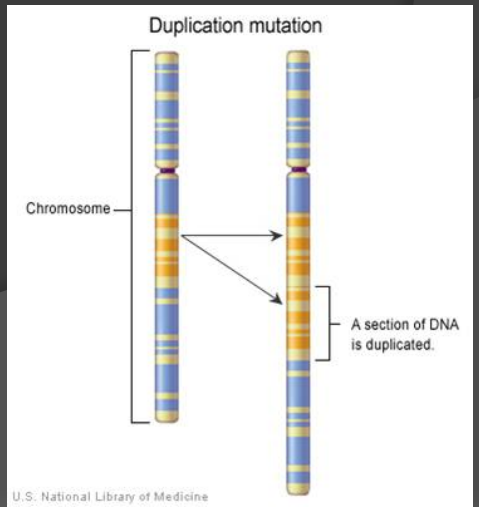
AACTGTTGCTAGCC
TTGACAACGATCGG

AACTGTCGATGCTAGCC
TTGACAGCTACGATCGG

- **Copy Number Variations (CNVs)**
  - deletions and gains
  - segmental duplications

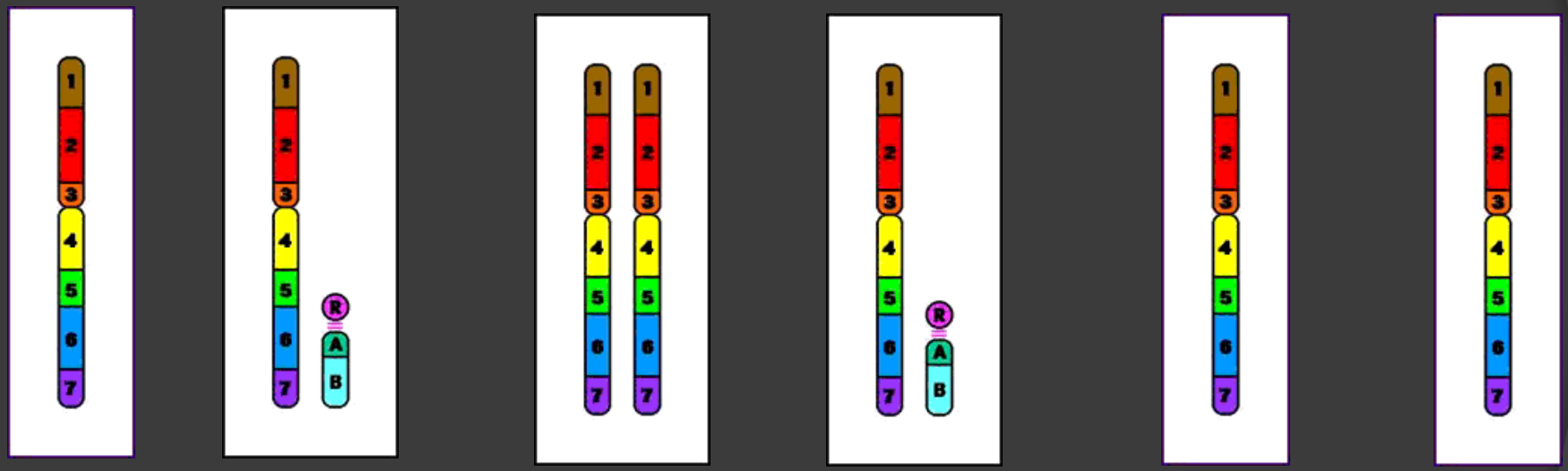
- **Structural Variations**
  - translocations
  - inversions
  - large insertions and deletions

CNV



Structural Variations in Genomes

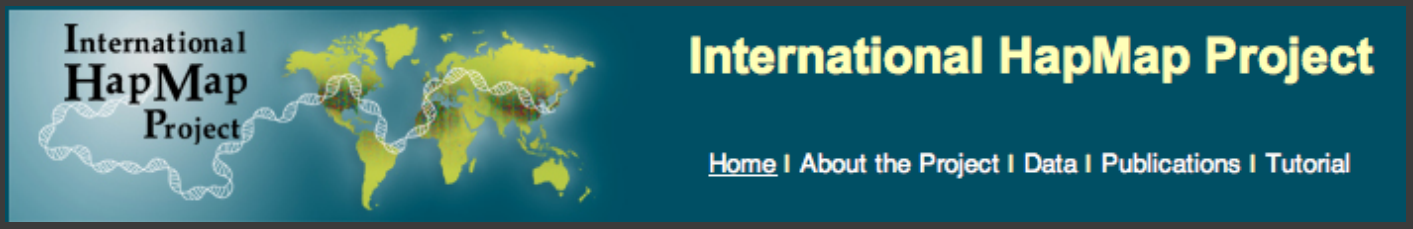
- Structural Variations (SVs) are chromosome rearrangements that occur during species evolution and remain variable in populations



Deletion Insertion Duplication Trans-location Paracentric Inversion Pericentric Inversion

- SVs also occur frequently in human cancers

Projects Investigating Normal Genome Variation



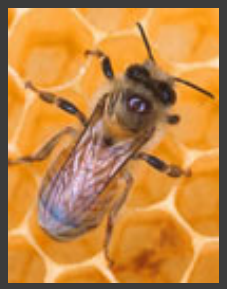
SNP microarrays of 4 populations: CEU, African, Chinese, Japanese

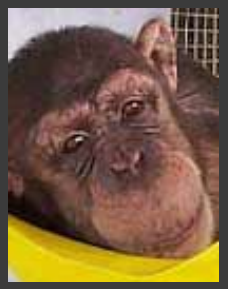
Database storing all SNPs and Indels that vary in populations

Next-generation sequencing of normal individuals from different ethnic populations

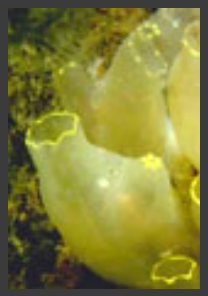
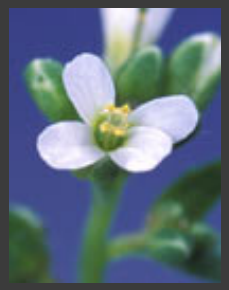
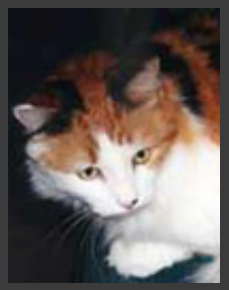
Database storing normal copy number variations from many studies

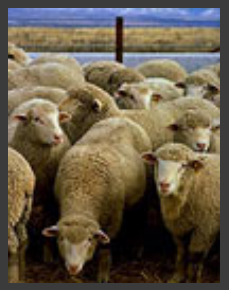
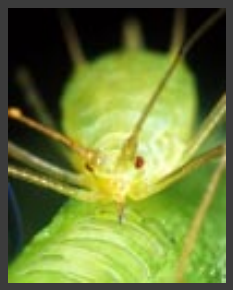
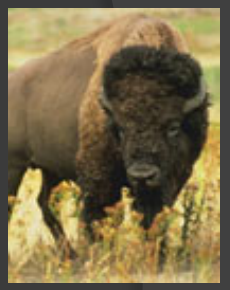
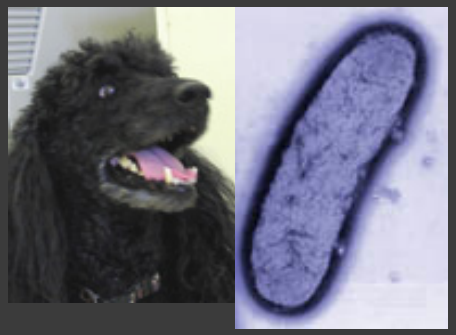
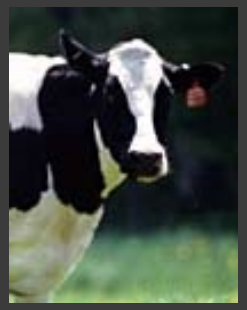
Organisms with Genomes Sequenced

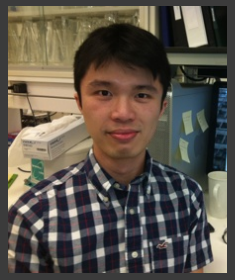
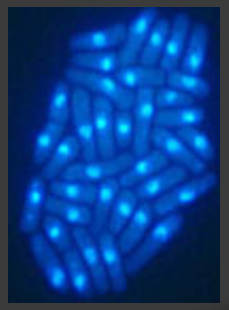








Over 180 genome assemblies have been sequenced in the last 20 years
http://www.genomenewsnetwork.org/

Archeological Genome Sequencing

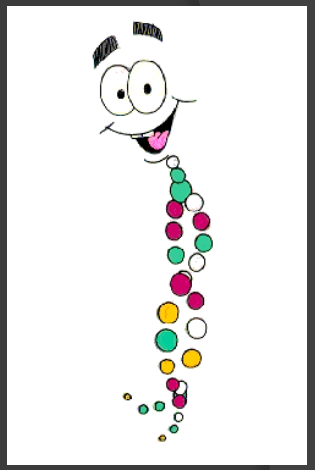
DNA is a highly stable molecule and can thus be extracted from archeological samples and extinct species for sequencing

A Draft Sequence of the Neandertal Genome

Richard E. Green,[1*†‡] Johannes Krause,[1†§] Adrian W. Briggs,[1†§] Tomislav Maricic,[1†§] Udo Stenzel,[1†§] Martin Kircher,[1†§] Nick Patterson,[2†§] Heng Li,[2†] Weiwei Zhai,[3†‖] Markus Hsi-Yang Fritz,[4†] Nancy F. Hansen,[5†] Eric Y. Durand,[3†] Anna-Sapfo Malaspinas,[3†] Jeffrey D. Jensen,[6†] Tomas Marques-Bonet,[7,13†] Can Alkan,[7†] Kay Prüfer,[1†] Matthias Meyer,[1†] Hernán A. Burbano,[1†] Jeffrey M. Good,[1,8†] Rigo Schultz,[1] Ayinuer Aximu-Petri,[1] Anne Butthof,[1] Barbara Höber,[1] Barbara Höffner,[1] Madlen Siegemund,[1] Antje Weihmann,[1] Chad Nusbaum,[2] Eric S. Lander,[2] Carsten Russ,[2] Nathaniel Novod,[2] Jason Affourtit,[9] Michael Egholm,[9] Christine Verna,[21] Pavao Rudan,[10] Dejana Brajkovic,[11] Željko Kucan,[10] Ivan Gušic,[10] Vladimir B. Doronichev,[12] Liubov V. Golovanova,[12] Carles Lalueza-Fox,[13] Marco de la Rasilla,[14] Javier Fortea,[14¶] Antonio Rosas,[15] Ralf W. Schmitz,[16,17] Philip L. F. Johnson,[18†] Evan E. Eichler,[7†] Daniel Falush,[19†] Ewan Birney,[4†] James C. Mullikin,[5†] Montgomery Slatkin,[3†] Rasmus Nielsen,[3†] Janet Kelso,[1†] Michael Lachmann,[1†] David Reich,[2,20*†] Svante Pääbo[1*†]

Sequencing the nuclear genome of the extinct woolly mammoth

Webb Miller[1], Daniela I. Drautz[1], Aakrosh Ratan[1], Barbara Pusey[1], Ji Qi[1], Arthur M. Lesk[1], Lynn P. Tomsho[1], Michael D. Packard[1], Fangqing Zhao[1], Andrei Sher[2]‡, Alexei Tikhonov[3], Brian Raney[4], Nick Patterson[5], Kerstin Lindblad-Toh[5], Eric S. Lander[5], James R. Knight[6], Gerard P. Irzyk[6], Karin M. Fredrikson[7], Timothy T. Harkins[7], Sharon Sheridan[7], Tom Pringle[8] & Stephan C. Schuster[1]



But probably not DNA from Dinosaurs – the DNA is too degraded

Genome Evolution

How do genomes evolve?

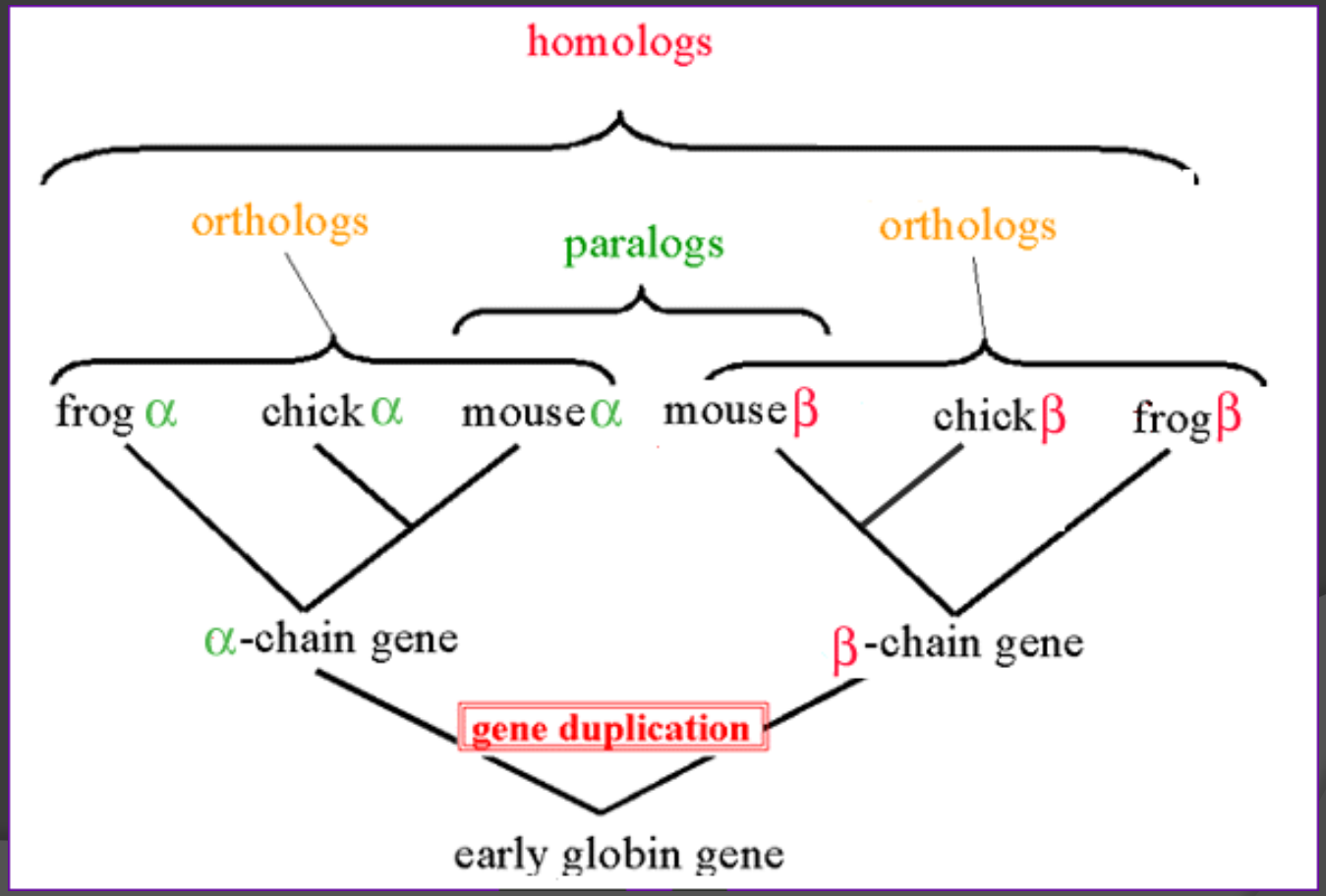
- Gene duplication and mutation
- Gene death (deletion)
- Exon Shuffling
- Insertion of novel sequences (horizontal gene transfer)

What are the mechanisms?

- Viruses and transposable elements
- Recombination and Unequal Crossing Over
- *De Novo* Mutations introduced by DNA replication errors

Homologs, Paralogs and Orthologs

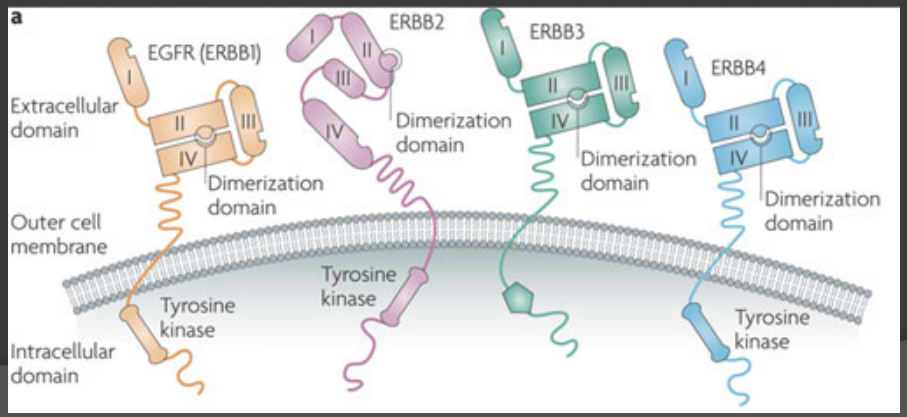
- **Homologs** – genes sharing common ancestry
- **Orthologs** – genes in different species that derive from a common ancestral gene
- **Paralogs** – genes within the same species that are related via gene duplication

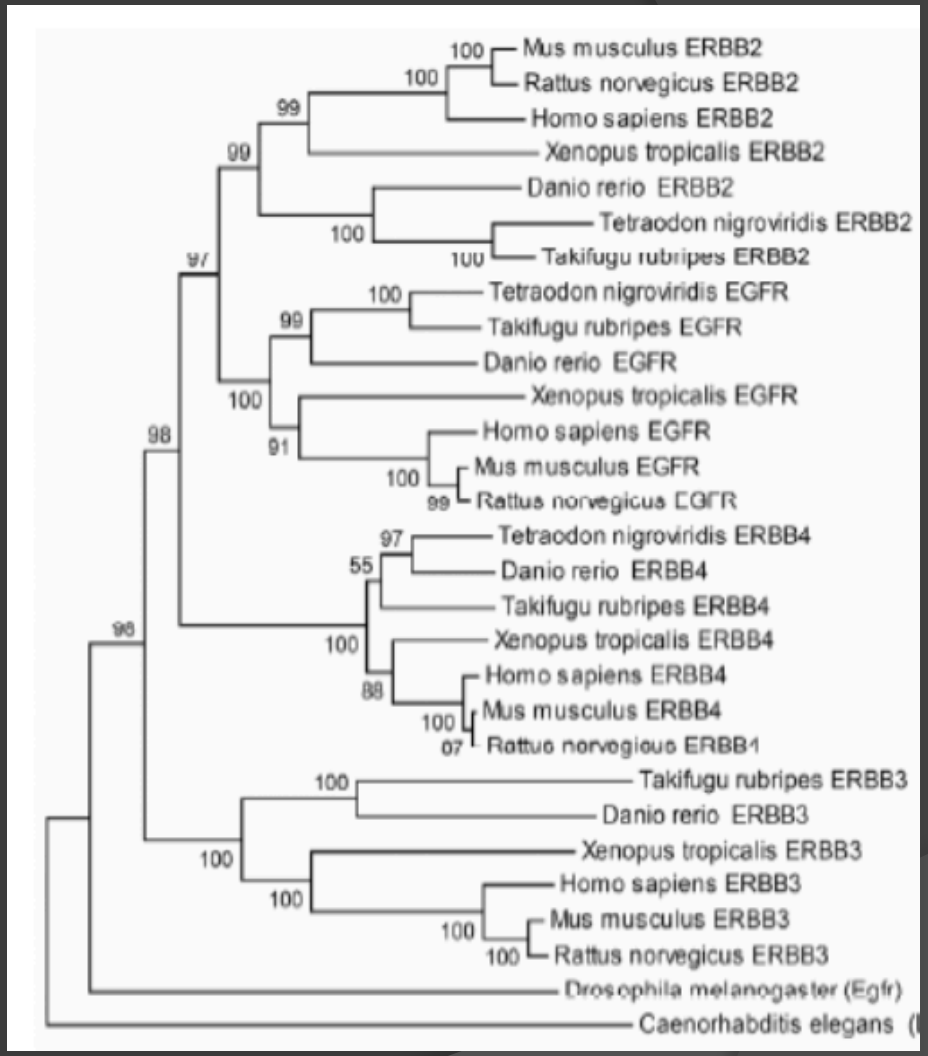


Gene Duplication and Evolution can lead to Protein Families

The EGFR family proteins are involved in cell signaling and proliferation. They have evolved through a complex series of gene duplications, deletions and mutations

Erbb1: chr1
Erbb2: chr17
Erbb3: chr3
Erbb4: chr2





Baselga & Swain *Nat Rev Cancer* 2009

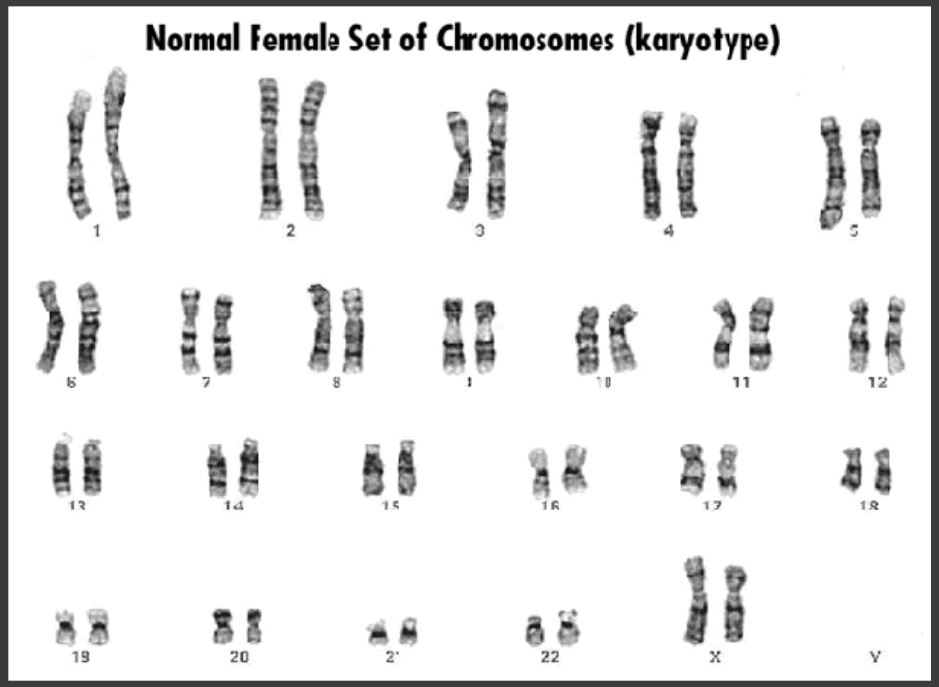
Summary I

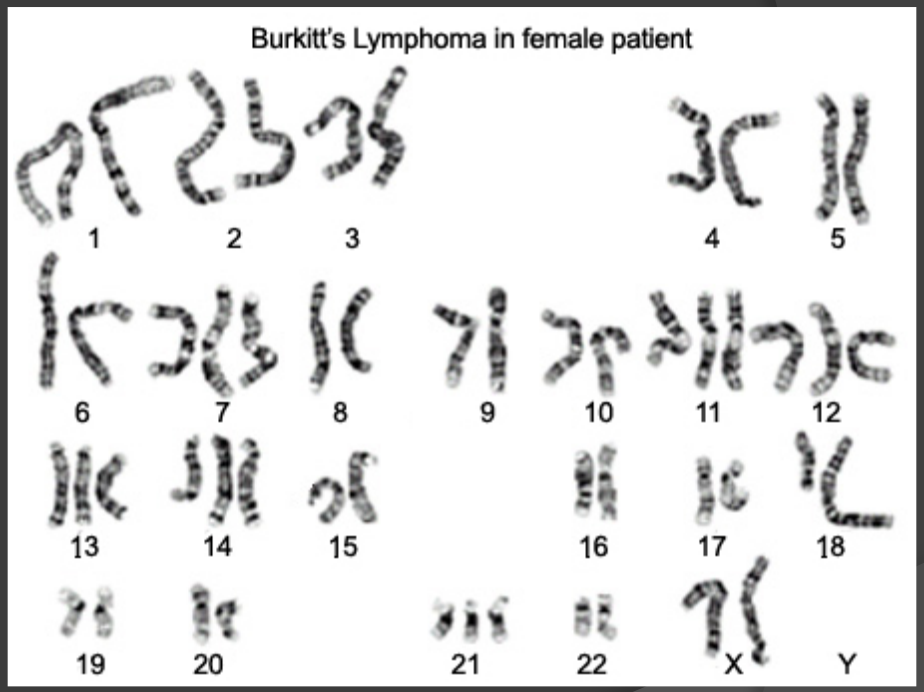
- Genomes vary in chromosome number, genome size, GC content, repetitive elements, percentage of coding regions
- Cytogenetic tools provided the first genomic information in normal and diseased genomes
- Microarrays platforms were the first quantitative tools for collecting genome-wide information on expression levels, copy number and chromatin interactions
- Next-generation sequencing (NGS) is a powerful method for analyzing genomes by providing quantitative data on : Single Nucleotide Variants, Indels, Copy Number Changes, Structural variants, RNA expression, Epigenetics, DNA-Protein binding
- The human genome project was the largest funded project in biology and the data is freely accessible for download
- Normal human genomes show significant variation in SNPs, Indels and Copy Number Variants
- Genomes evolve by gene duplication, chromosome deletion, *de novo* mutations and horizontal gene transfer (in microbes)

Cancer Genomics

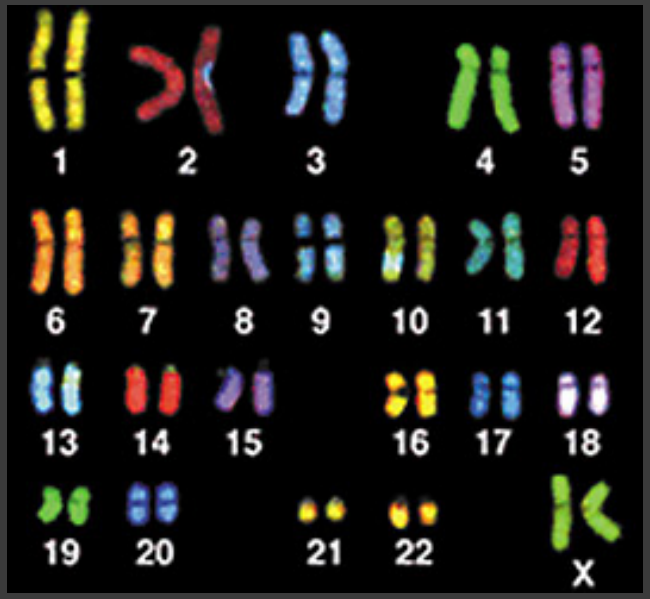
Cancer Cytogenetics

Karyotyping provided the first insight into chromosome abnormalities in cancer genomes

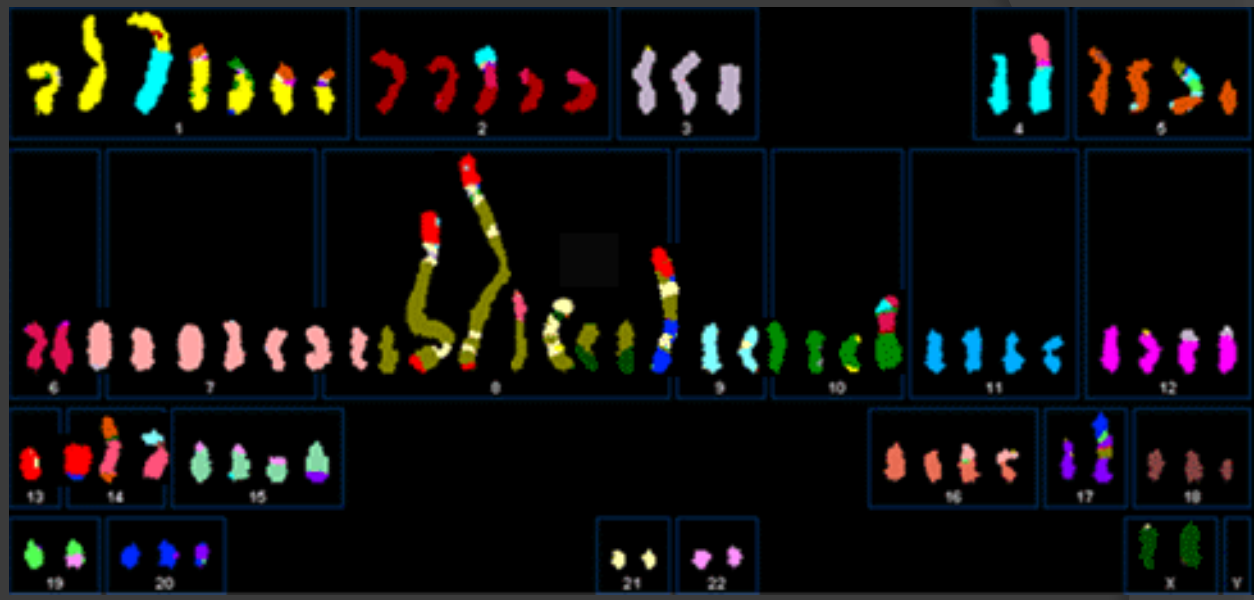




Spectral Karyotyping



Normal Female Genome

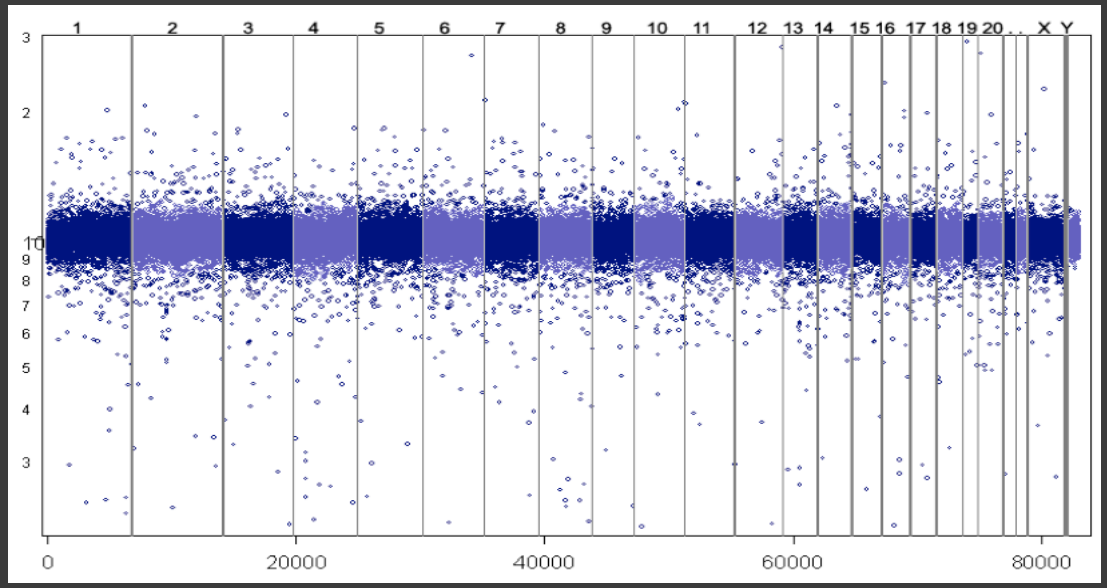


Breast Cancer Genome

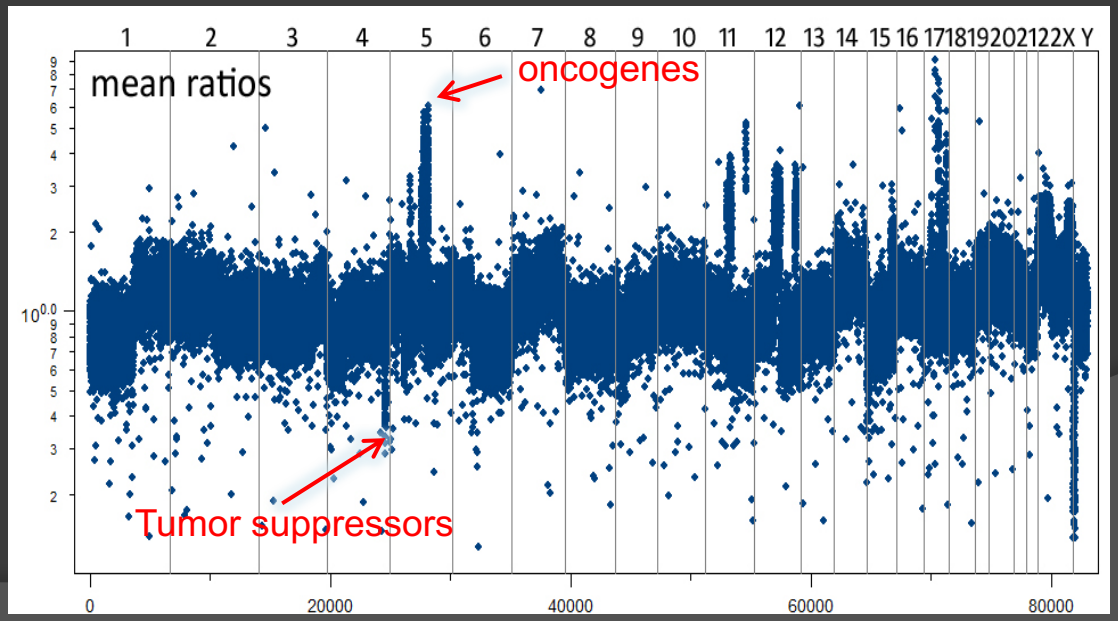
- SKY (Spectral Karyotyping) is a cytogenetic technique in which fluorescent probes are designed to hybridize to each chromosomes with a different color, allowing translocations, deletions and amplifications to be identified
- SKY has limited resolution: only rearrangements > 1mb can be detected

CGH Microarrays

Comparative Genomic Hybridization (CGH) Microarrays can measure genome-wide copy number aberrations at high-resolution to detect amplifications and deletions in cancer genomes

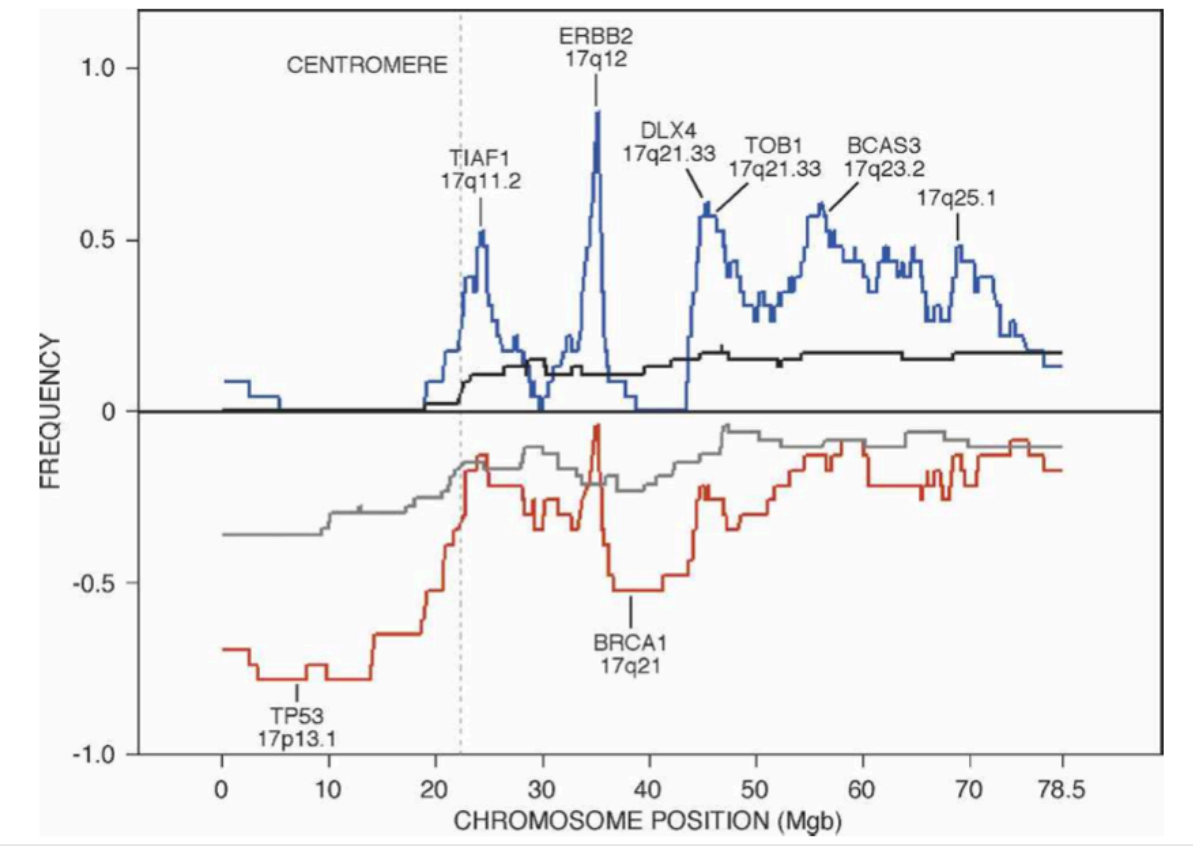


Copy Number Profile of a Normal Female



Copy Number Profile of a Breast Tumor Genome

CGH Microarrays

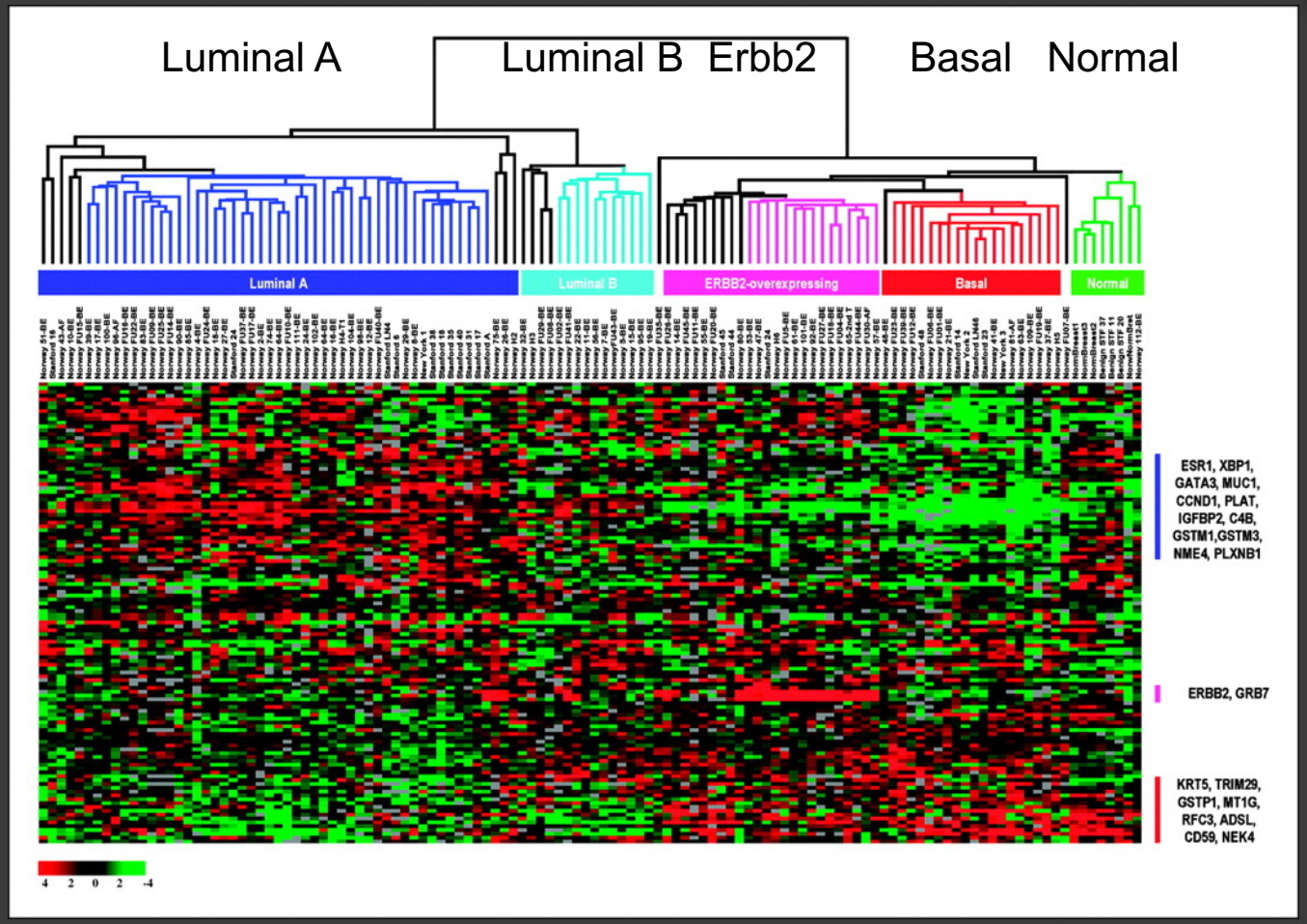


Chr17 Frequency Plot of 200 breast cancers

- Oncogenes (ERBB2, BCAS3) are often amplified in cancer genomes
- Tumor suppressors are often deleted (TP53, BRCA1) or show LOH

Hicks et al. 2006

Gene Expression Microarrays



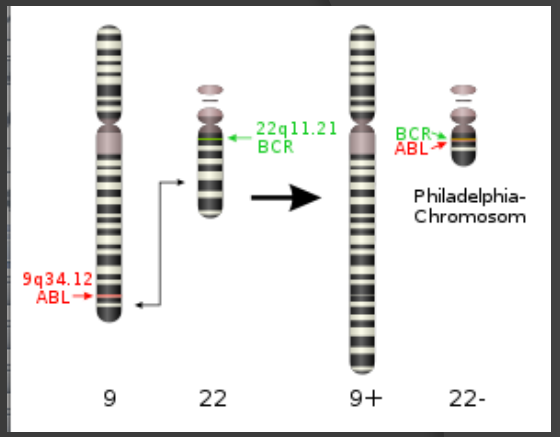
- Gene expression microarrays can measure the mRNA levels of the 25,000 genes in a cancer transcriptome
- Oncogenes are often overexpressed, while tumor suppressors often downregulated
- In some cancers (such as breast) it has been possible to identify ‘gene signatures’ that can stratify patient populations and predict survival

Sorlie et al. 2003

Gene Fusions in Cancer

BCR-ABL

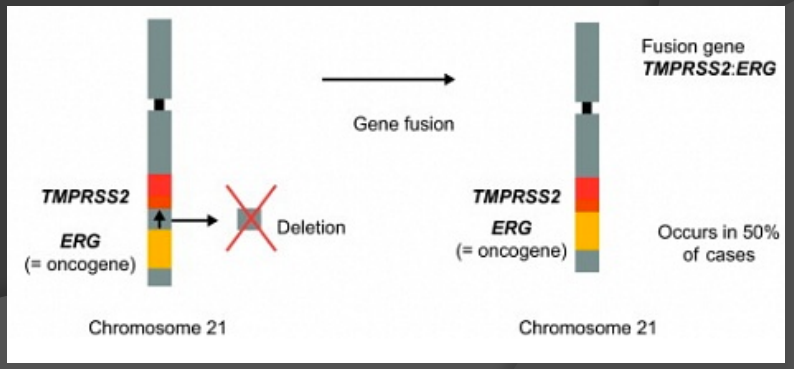
- Translocation between chrom 9 and 22 creates a fusion gene in 95% of cancer patients with Chronic Myelogenous Leukemia (CML)
- BCR is a breakpoint cluster gene and *Abl1* is an oncogene involved in cell proliferation and differentiation
- The fusion creates a constitutively activated tyrosine kinase oncogene that drives cancer



Gleevec (imatinib) is an oral drug that targets the tyrosine kinase activity of the BCR-ABL fusion gene and can put CLL patients in complete remission for many years

TMPRSS2-ERG

- The TMPRSS2-ERG fusion gene was Identified by paired-end next-generation sequencing
- Fusion on chromosome 21q22
- This translocation occurs in 70% of prostate cancers which occurs in 70% of patients (Maher et al., *Nature* 2010)



A Classical View of Cancer Progression

The text-book view of cancer is that 2-5 'hits' in cancer genes are required for tumor growth and invasion

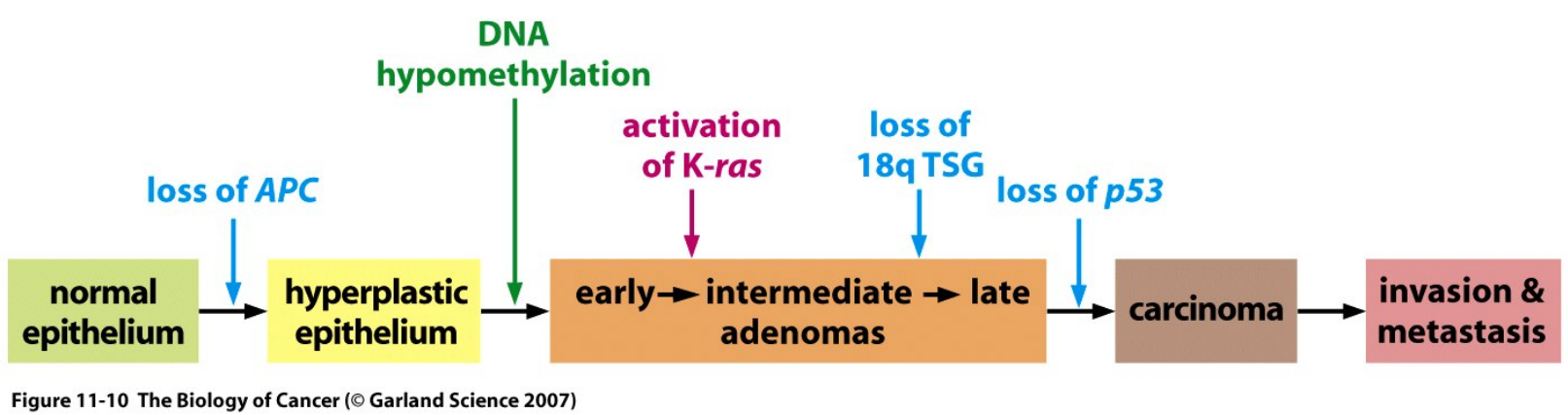
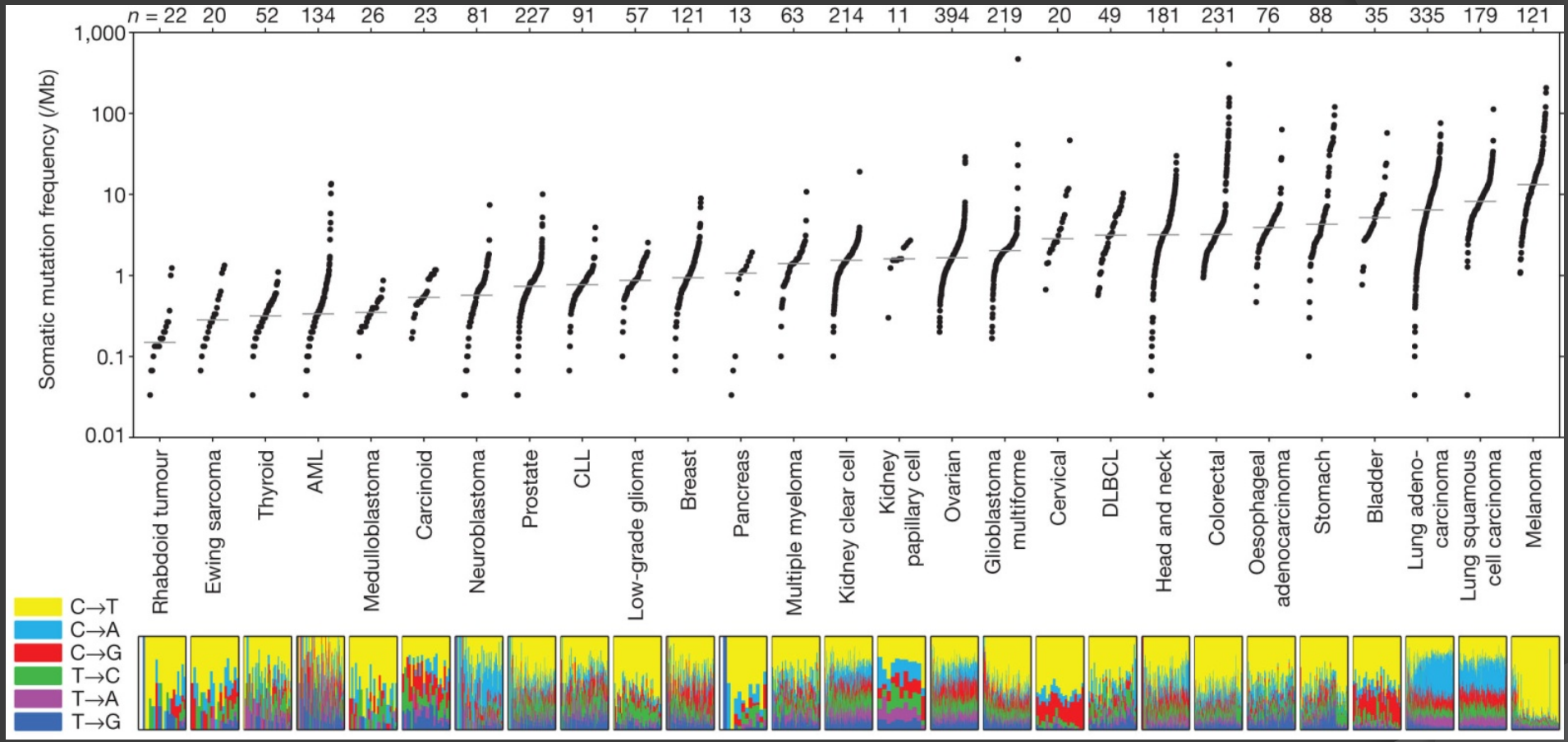


Figure 11-10 The Biology of Cancer (© Garland Science 2007)

However, next-generation sequencing studies are showing that cancer genomes can harbor hundreds of nonsynonymous mutations that may affect protein function

Cancer Genomes Harbor Thousands of Somatic Mutations



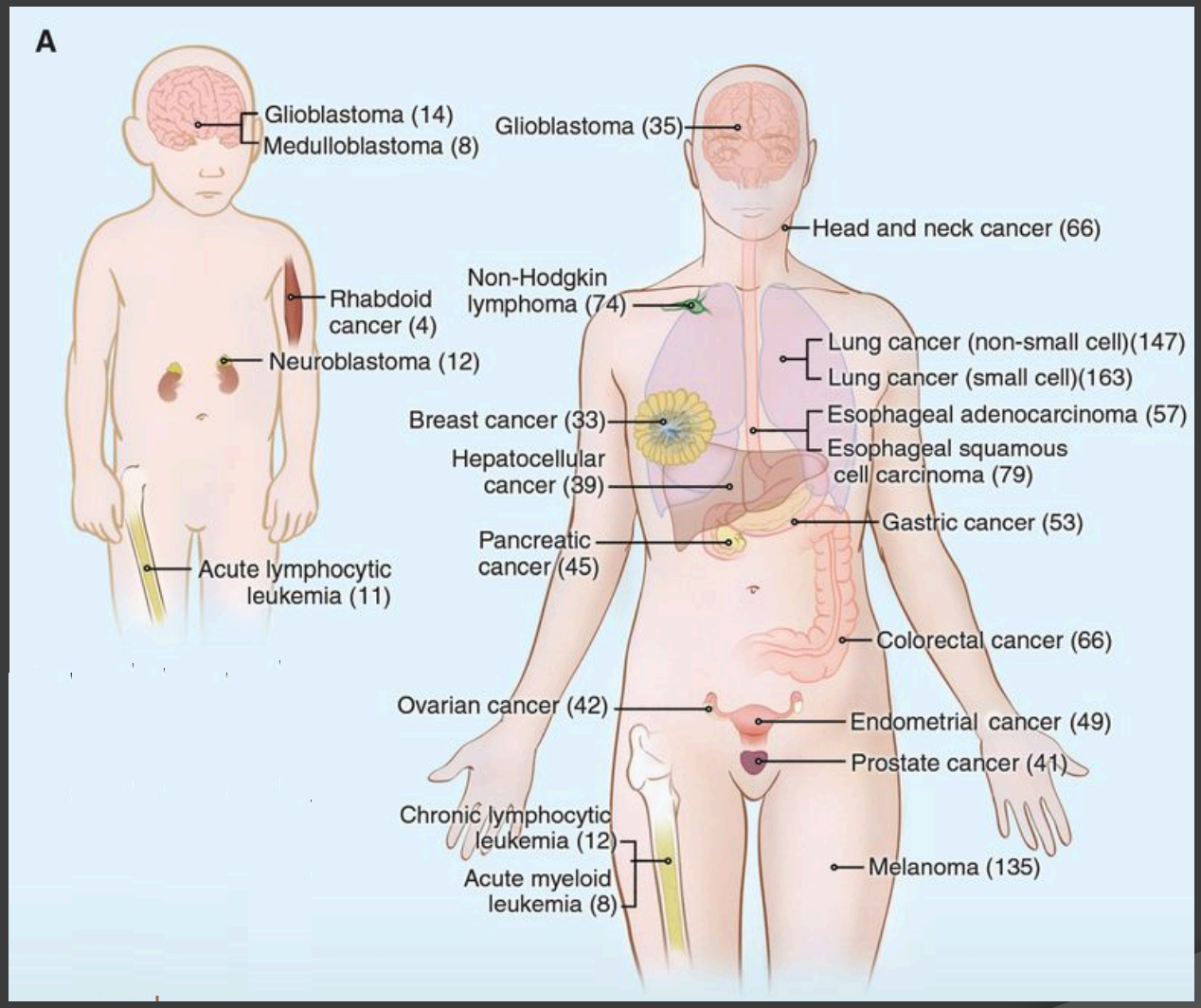
Trends:

- Hematopoetic tumors have lower mutation frequencies, while solid tumors generally have higher frequencies
- Mutation spectrum changes in different cancers, particularly lung carcinomas (C>A transversions) and Melanomas (C>T transitions)

MS Lawrence *et al. Nature* **000**, 1-5 (2013)

Nonsynonymous Mutations in Different Cancer Types

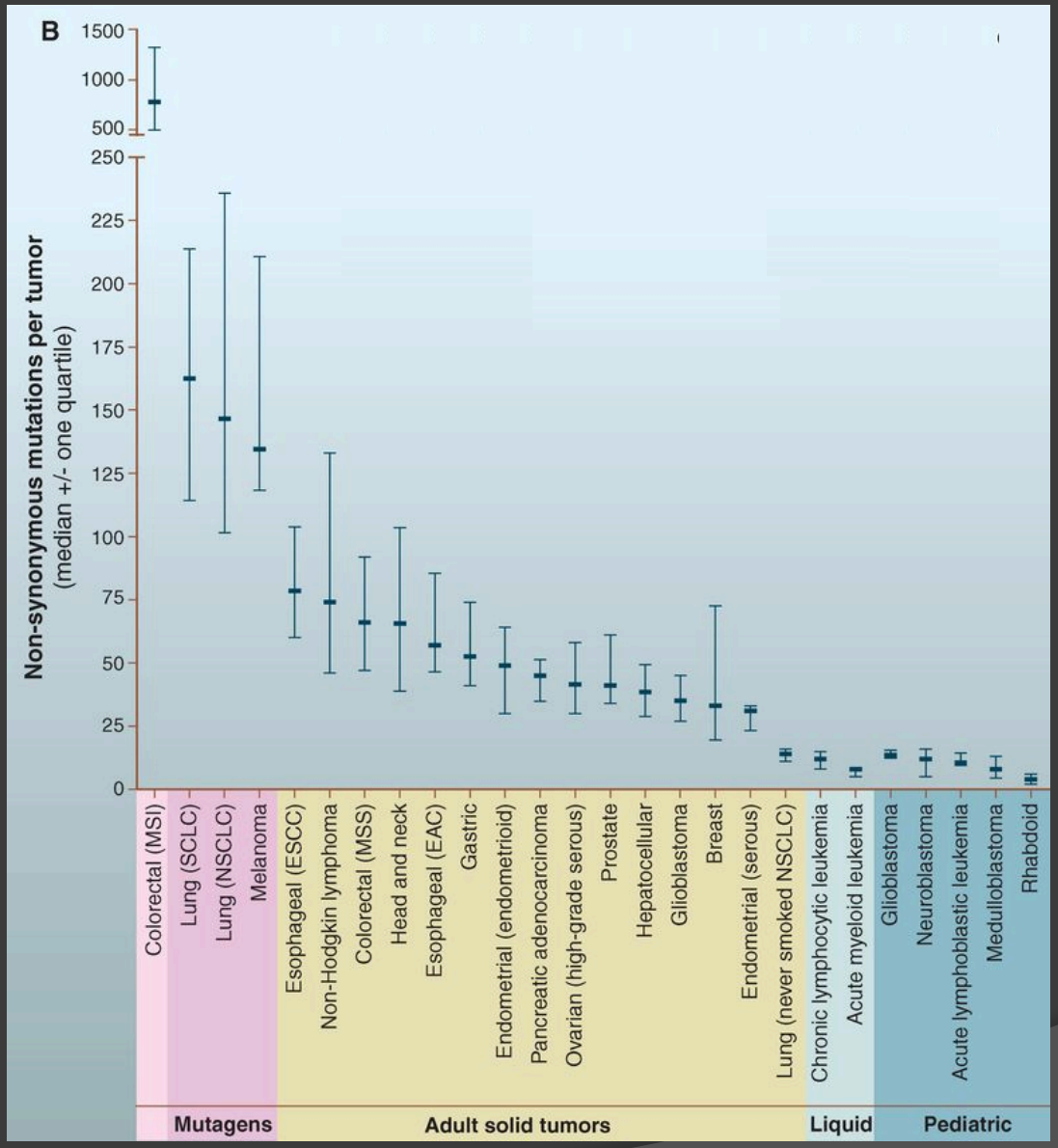
Vogelstein et al. (2013) *Science*



- Median number of nonsynonymous mutations varies greatly across different cancer types
- Pediatric cancers have the lowest numbers of somatic mutations

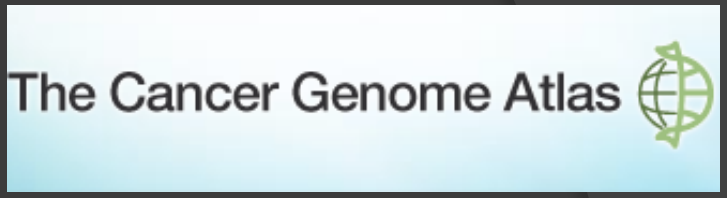
Nonsynonymous Mutations in Different Cancer Types

Vogelstein et al. (2013) *Science*



Mutagenic cancers have the highest number of somatic mutations

Large-Scale Cancer Genome Sequencing Studies



The Cancer Genome Atlas (TCGA)

- http://tcga-data.nci.nih.gov/tcga/
- Contains copy number data, expression data, protein data and next-generation sequencing mutational data on 30 different cancer types
- All data is free to download, however the sequencing data requires special permission for studying, due to patient privacy concerns
- MD Anderson is a Genome Data Analysis Center for TCGA

The International Cancer Genome Consortium

- http://icgc.org
- Copy number, expression, methylation and DNA sequencing data on 50 different cancer types
- All data can be downloaded, but sequencing data is also restricted

Cancer Patients Share Few Common Mutations

ARTICLE

doi:10.1038/nature10166

Integrated genomic analyses of ovarian carcinoma

The Cancer Genome Atlas Research Network*

316 ovarian cancers were analyzed by exome sequencing and identified 19,396 somatic mutations in genes, but only 1 gene occurred at a frequency > 5%

Gene	No. of mutations
TP53	302
BRCA1	11
CSMD3	19
NF1	13
CDK12	9
FAT3	19
GABRA6	6
BRCA2	10
RB1	6

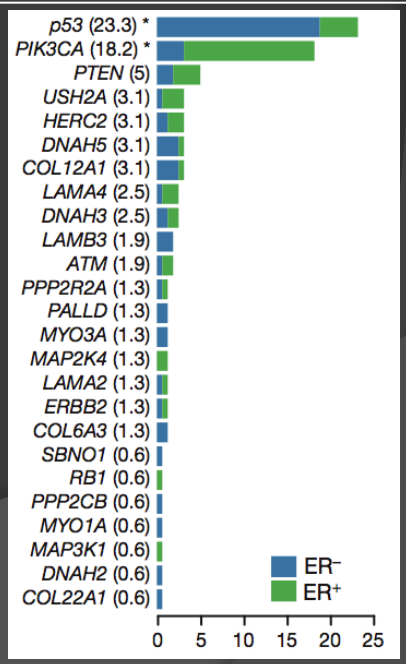
LETTER

doi:10.1038/nature10933

The clonal and mutational evolution spectrum of primary triple-negative breast cancers

Sohrab P. Shah[1,2], Andrew Roth[1,2]*, Rodrigo Goya[3]*, Arusha Oloumi[1,2]*, Gavin Ha[1,2]*, Yongjun Zhao[3]*, Gulisa Turashvili[1,2]*, Jiarui Ding[1,2]*, Kane Tse[3]*, Gholamreza Haffari[1,2]*, Ali Bashashati[1,2]*, Leah M. Prentice[1,2], Jaswinder Khattra[1,2], Angela Burleigh[1,2], Damian Yap[1,2], Virginie Bernard[4], Andrew McPherson[1,2], Karey Shumansky[1,2], Anamaria Crisan[1,2], Ryan Giuliany[1,2], Alireza Heravi-Moussavi[1,2], Jamie Rosner[1,2], Daniel Lai[1,2], Inanc Birol[3], Richard Varhol[3], Angela Tam[3], Noreen Dhalla[3], Thomas Zeng[3], Kevin Ma[3], Simon K. Chan[3], Malachi Griffith[3], Annie Moradian[3], S.-W. Grace Cheng[3], Gregg B. Morin[3,5], Peter Watson[1,6], Karen Gelmon[6], Stephen Chia[6], Suet-Feung Chin[7,8], Christina Curtis[7,8,9], Oscar M. Rueda[7,8], Paul D. Pharoah[7], Sambasivarao Damaraju[10], John Mackey[10], Kelly Hoon[11], Timothy Harkins[11], Vasisht Tadigotla[11], Mahvash Sigaroudinia[12], Philippe Gascard[12], Thea Tlsty[12], Joseph F. Costello[13], Irmtraud M. Meyer[5,14,15], Connie J. Eaves[16], Wyeth W. Wasserman[4,5], Steven Jones[3,5,17], David Huntsman[1,2,18], Martin Hirst[3,15,19], Carlos Caldas[7,8,20,21], Marco A. Marra[3,5] & Samuel Aparicio[1,2]

Identified 2414 somatic mutations in genes, but only 2 genes occurred above frequencies of 5%



Novel Cancer Genes Discovered by Sequencing

Energy metabolism genes:

- **IDH1** mutations in 16% AML (Mardis et al. , NEJM 2009)
- **IDH1** mutations in 80% of glioblastomas (Yan et al. 2009, NEJM)

DNA Methyltransferases:

- **DNMT3A** mutations in 22% of AML patients (Ley et al. 2010)

Histone modifying gene mutations:

- **SETD2**, **JARID1C** and **UTX**) in Clear Cell Renal Carcinoma (Dalgliesh et al. 2010)

Developmental gene mutations:

- **NOTCH1** inactivating mutations in 15% of Head and Neck Squamous Cell Carcinoma (Agrawal et al. 2011)

Integrin and ECM gene mutations:

- ITG, COL, MYH, MYO in Triple-Negative Breast Cancers (Shah et al. 2012)

Driver vs. Passenger Mutations

Driver Mutation - a mutation that provides a selective advantage to a clone and increases its fitness (survival or reproduction). Driver mutations lead to clonal expansions and tumor progression.

Two classes

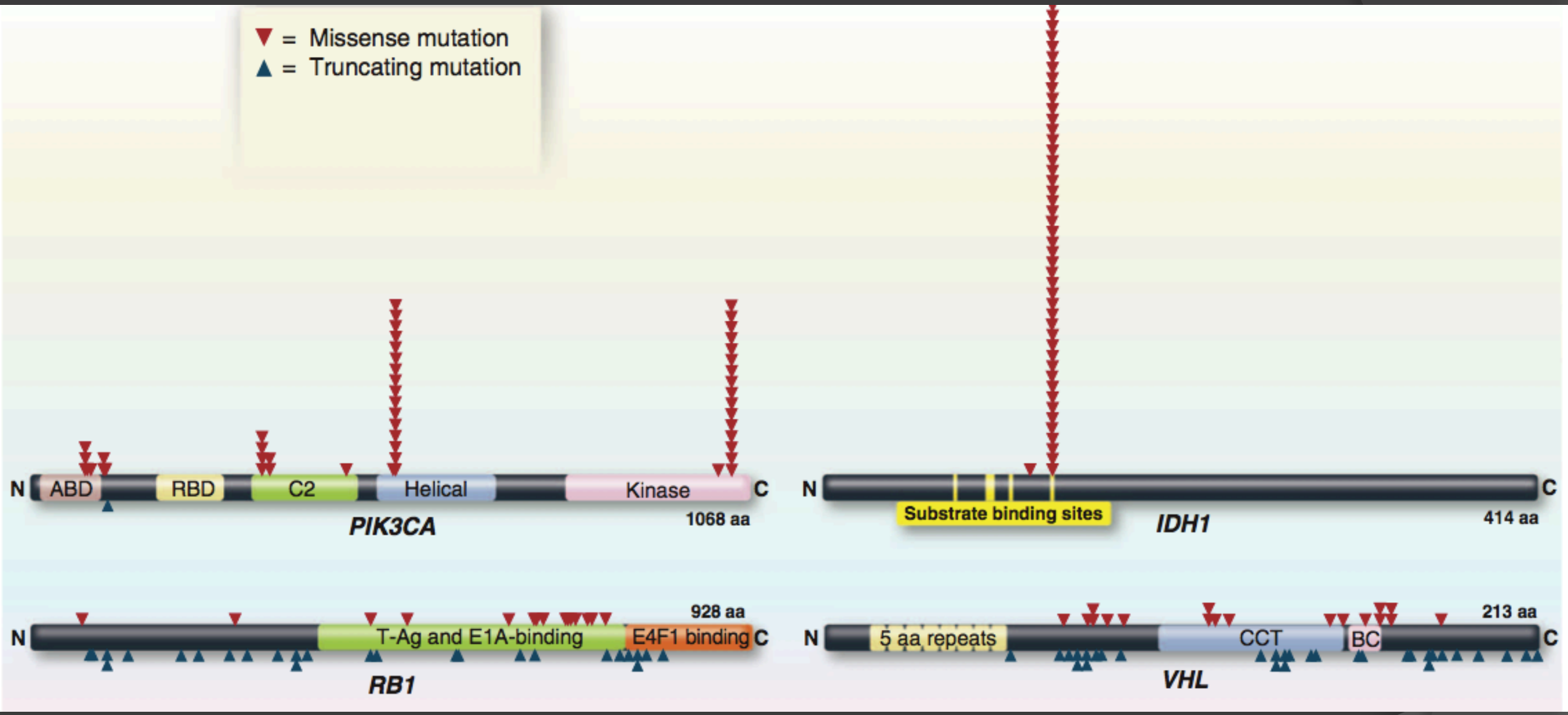
- **Activating Mutations –** increase protein levels/activity (occur in oncogenes)
- **Inactivating Mutations** – decrease protein levels/function (occur in tumor suppressors)

Passenger Mutation - a mutation that has no effect on the fitness of a clone but may be associated with a clonal expansion because it occurs in the same genomic region as a driver mutation.

How to distinguish between driver and passenger mutations?

Activating Driver Mutations are Recurrent in Patients

Data from COSMIC



- **Activating missense** driver mutations (ex. Oncogenes) often occur at the same nucleotide position in many different patients and cancer types (not true for **nonsense** truncating mutations)
- Passenger mutations are randomly distributed across genes
- Inactivating nonsense mutations (ex. Tumor suppressors) cluster in protein domains

Vogelstein et al. (2013) *Science*

Prediction Algorithms for Identifying Driver Mutations

SIFT

- http://sift.jcvi.org
- Predicts whether an amino acid substitution is likely to have a damaging effect on protein function
- Prediction scores range from 0-1, where < 0.05 is considered damaging
- SIFT uses sequence conservation across species as predicted by PSI-BLAST

POLYPHEN2

- http://genetics.bwh.harvard.edu/pph2/
- 'Polymorphism Phenotyping 2'
- Predicts whether an amino acid substitution is likely to have a damaging effect on protein function
- Prediction scores range from 0-1, where > 0.5 is considered damaging (note: the inverse from SIFT)
- Polyphen uses protein structure information and properties of amino acids to predict the functional impact of mutations

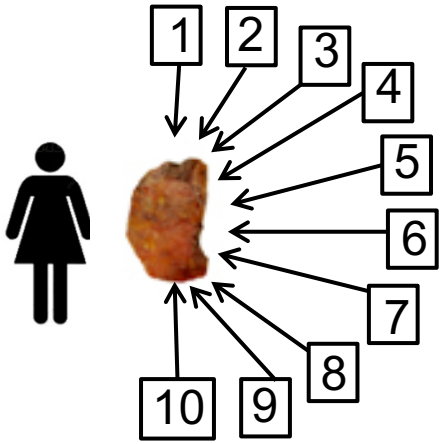
Personalized Medicine

- The objective of personalized (precision) medicine is to sequence the genome of a patient's tumor and identify *actionable* mutations for targeted therapies, instead of using broad chemotherapies
- At MD Anderson the Institute for Personalized Cancer Therapy (IPCT) and many clinical trials use NGS and targeted therapy to treat cancer patients
- Currently over 100 targeted therapies are commercially available and hundreds more are in phase II and phase III clinical trials

Protein Target	Targeted-Therapy	Cancer
ERBB2	Trastuzumab, lapatinib	breast
EGFR	Iressa, Tarceva, Cetuximab	Lung, colon
mTOR	Torisel	renal
VEGF	Avastin	Lung, breast, colon
BRAF, VEGF2, PDGF	Sorafenib	melanoma
ABL, C-KIT, PDGFR	Gleevec	CLL
PARP	Olaparib	Breast, ovarian
ER	Tamoxifin, toremifene	Breast
Aromatase	Arimidex, aromasin	Breast
CD20	Rituximab, Ofatumumab	CLL

Inter-tumor vs. intra-tumor heterogeneity

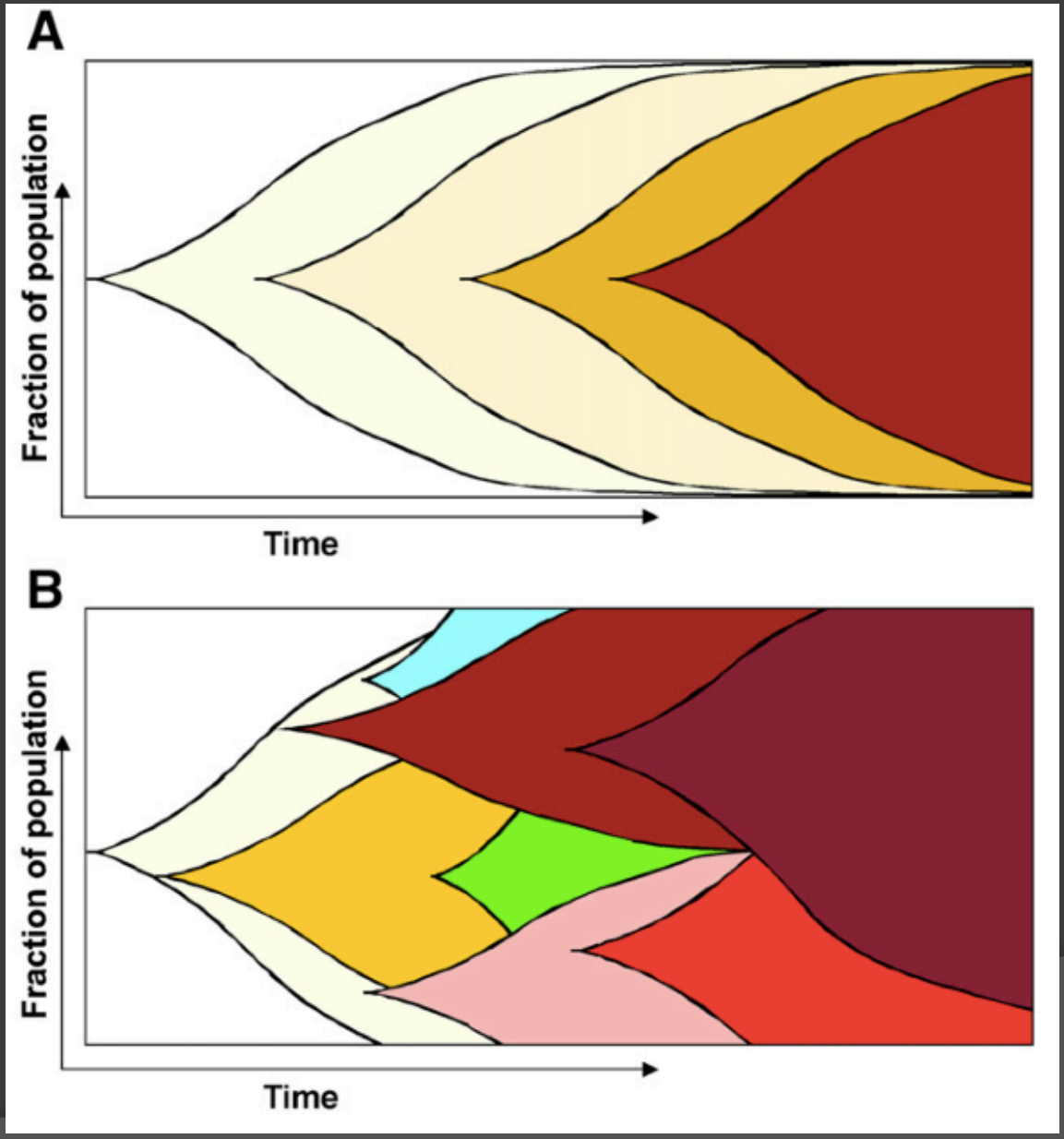
Intratumor heterogeneity - Mutational differences between cells that occur *within* individual tumors



Intertumor heterogeneity - Mutational differences that occur between different cancer patients

Intratumor Heterogeneity

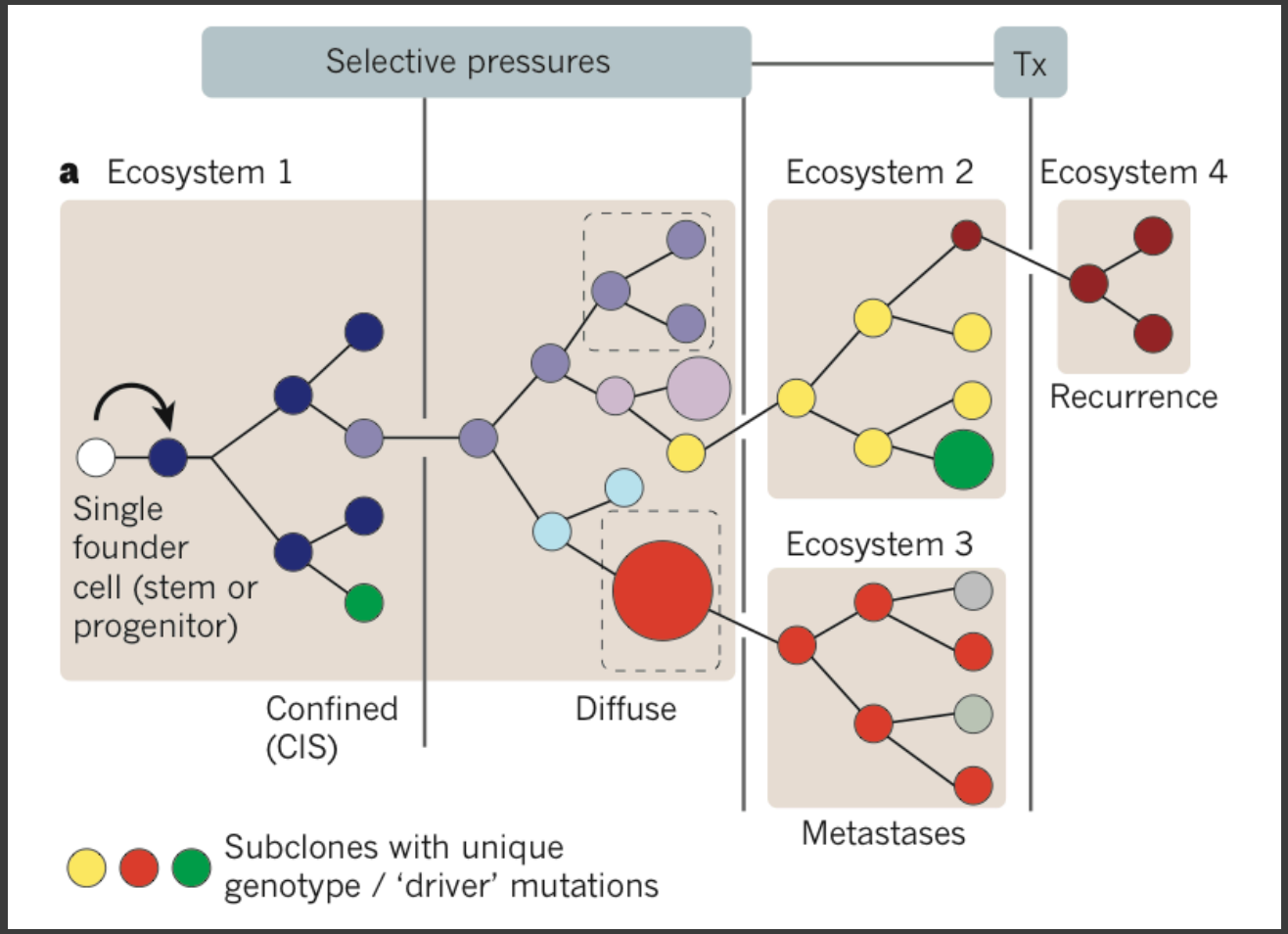
Monoclonal Tumor



Traditional View of Cancer

Polyclonal Tumor

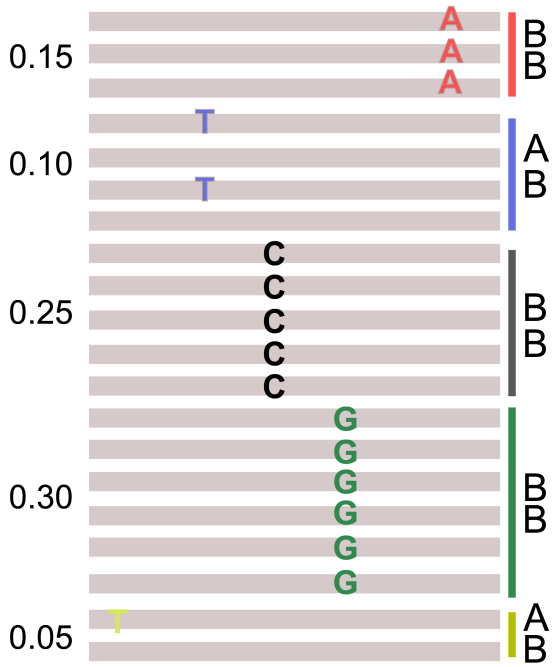
Tumors Are Evolving Ecosystems



Tumors are diverse populations of cells with different fitnesses, that encounter selective pressures such as nutrient deprivation, hypoxia, physical barriers, and chemotherapy.

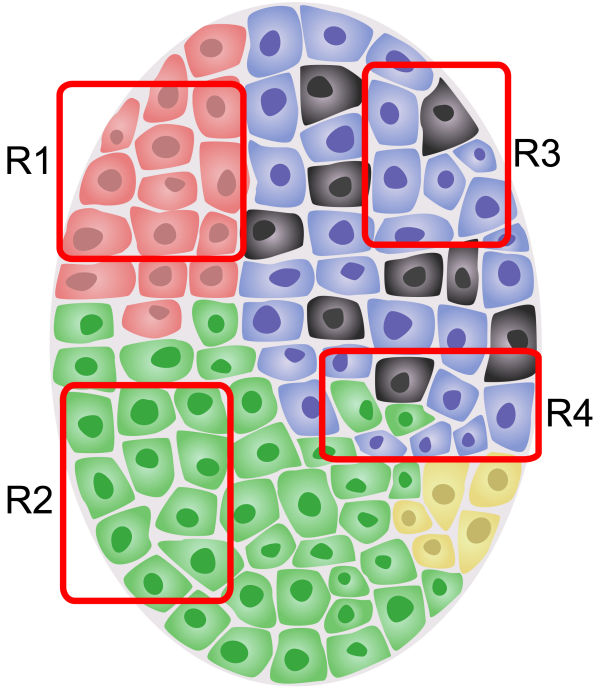
Greaves and Maley, *Nature* 2012

Sequencing Methods for Resolving Intratumor Heterogeneity



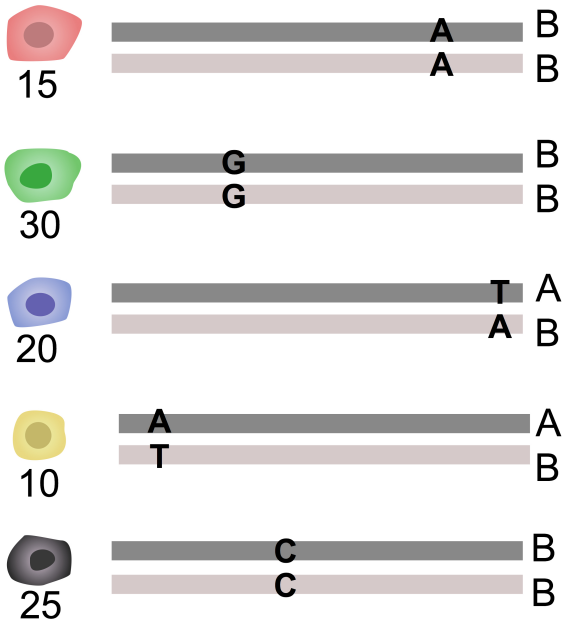
deep-sequencing

Nik-Zainal et al. (2012) *Cell*
Shah et al. (2012) *Nature*
Carter et al. (2012) *Nat. Biotech*
Campbell et al. (2008) *PNAS*
Miller et al. (2014) *PLOS Comp Bio*
Ha et al. (2014) *Genome Res*
Roth et al. (2014) *Nature Methods*
Eirew et al. (2014) *Nature*



multi-region sequencing

Gerlinger et al. (2012) *NEJM*
Gerlinger et al. (2012) *Nat. Gen.*
Zhang et al. (2014) *Science*
De Bruin et al. (2014) *Science*
Yates et al. (2015) *Nature Med.*
Sottoriva et al. (2012) *PNAS*
Molhatra et al. (2015) Gen Med

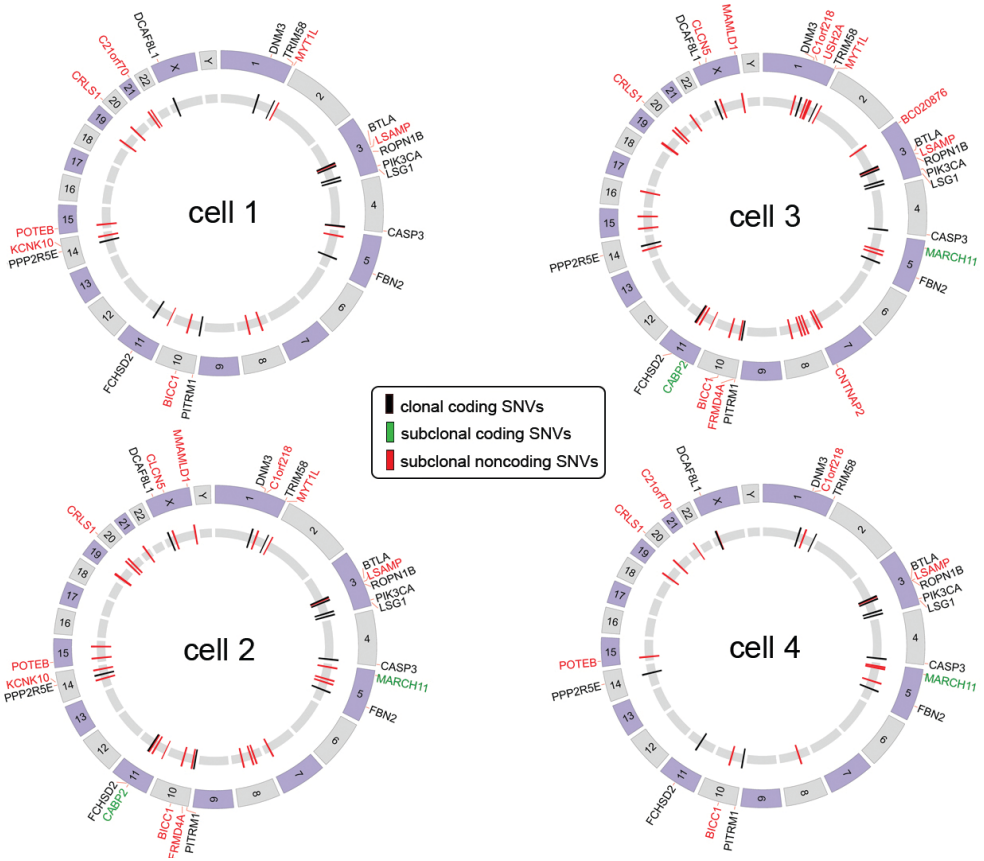


single cell sequencing

Navin et al. (2011) *Nature*
Hou et al. (2012) *Cell*
Xu et al. (2012) *Cell*
Zong et al. (2012) *Science*
Wang et al. (2014) *Nature*
Patel et al. (2014) *Science*
Leung et al. (2015) *Genome Bio.*
Zafar et al. (2016) *Nature Meth.*
Gao et al. (2016) *Nature Gen.*

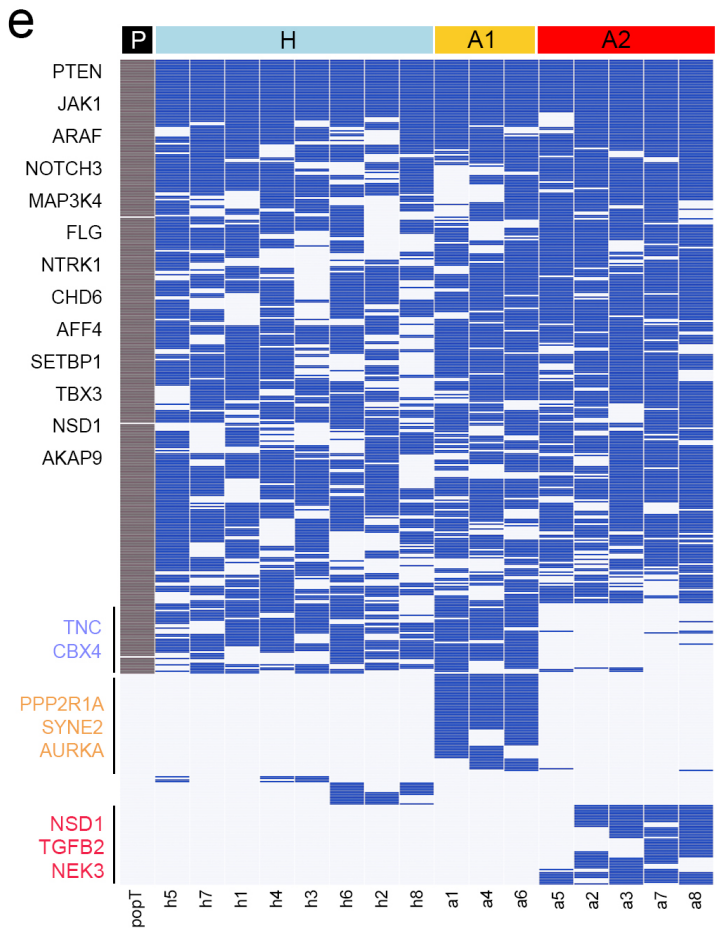
Single Cell Genome Sequencing in Breast Cancer Patients

Single Cell Genome Sequencing of an Estrogen-Receptor Positive Breast Cancer Patient



No two tumor cells are genetically identical, but all cells Share a common evolutionary origin, from a single normal cell

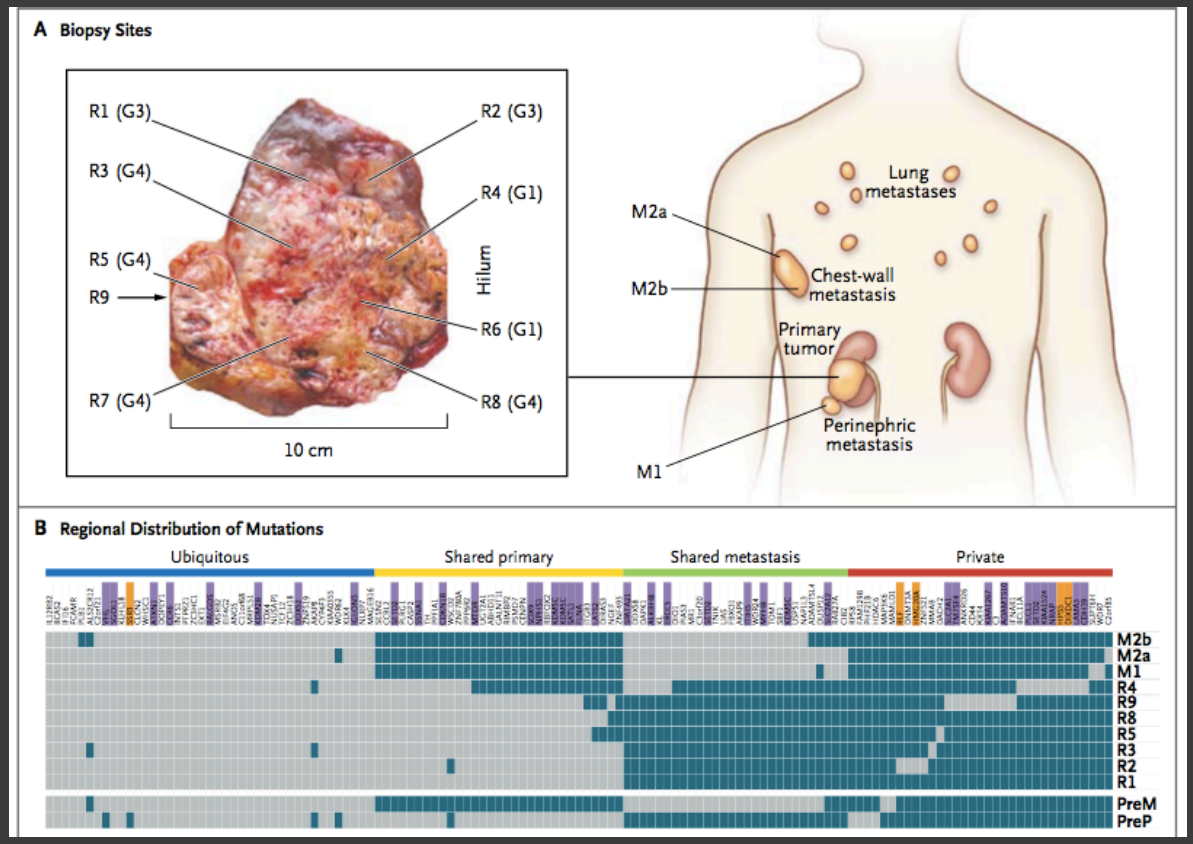
Single Cell Exome Sequencing of an Triple-Negative Breast Cancer Patient

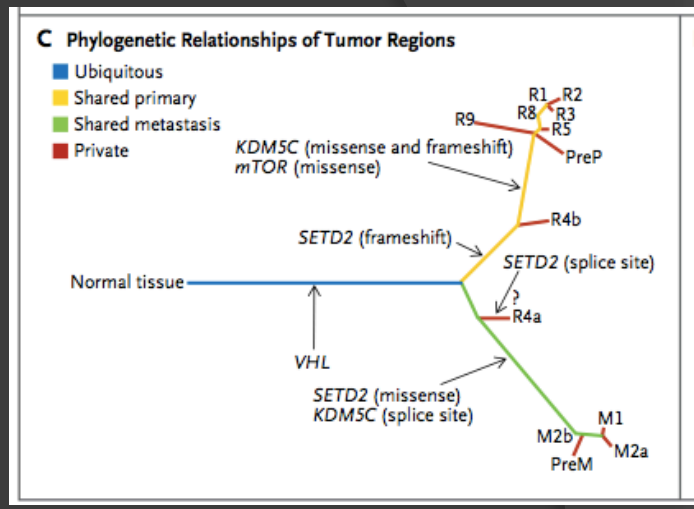


Three major tumor subpopulations were identified with different driver mutations

Wang et al. 2014, *Nature*

Spatial Genetic Heterogeneity inside Tumors



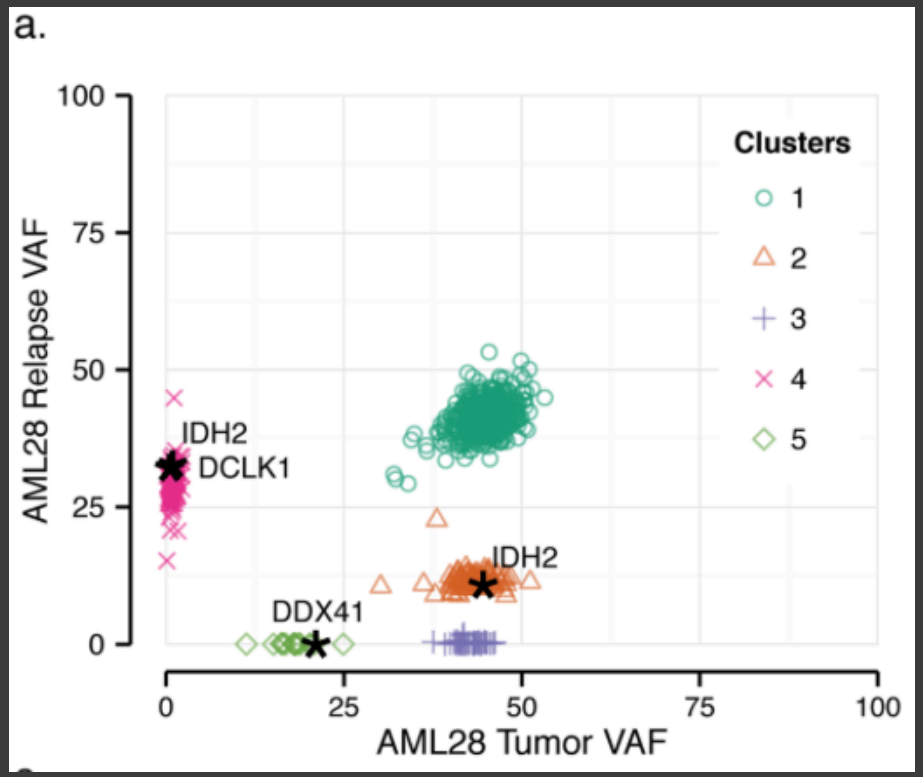


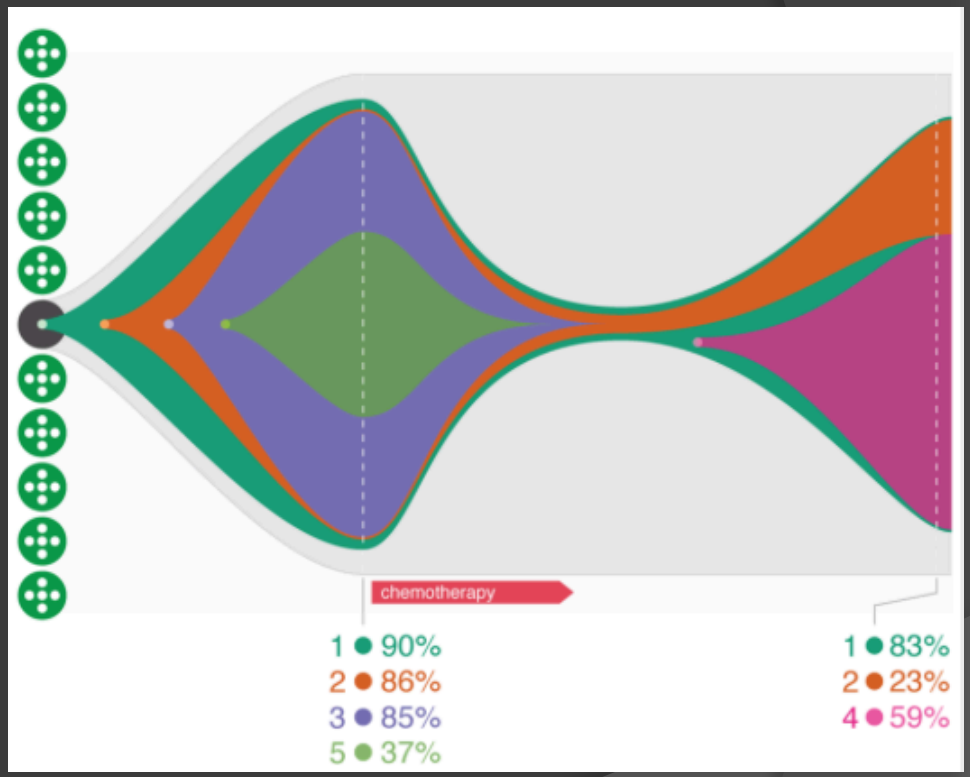
- Multiple spatially distinct regions were sampled in kidney tumors and used for exome sequencing to identify somatic mutations
- Only 63-69% of the somatic mutations were detected in all regions; many oncogenes and tumor suppressors (MTOR, SETD2 and PTEN) occurred only in distinct regions within the tumor
- The phylogenetic lineage and evolutionary history was inferred by comparing mutations from different geographical sites within the tumor

Swanton et al 2012, NEJM

Resolving Intratumor Heterogeneity by Deep-Sequencing

- Tumor subpopulations can also be estimated by deep-sequencing bulk tumor tissue at high coverage depth (ex. 100 – 1000X)
- Mutation frequencies can be calculated from the number of sequencing reads obtained from the tumor cells





- Clustering of mutation frequencies can identify groups of cells that share similar mutation frequencies and are assumed to be clonal subpopulations
- In this example SciClone was applied to an AML patient with pre and post chemotherapy samples, which identified 5 subpopulations

Miller et al. 2013, *PLOS Comp Biology*

Summary II

- Next-generation sequencing provides a powerful new approach to detect mutations, transcriptional changes and epigenomic modifications in cancer genomes
- Personalized medicine involves treating patients based on the unique mutations present in their tumors
- Distinguishing between driver and passenger mutations is a major challenge in bioinformatics
- Inter-patient heterogeneity refers to the many differences in mutations between patients with the same cancers
- Intra-tumor heterogeneity refers to the many genetic differences from cell to cell within an individual tumor
- Sequencing Methods for resolving intratumor heterogeneity include: multi-region sequencing, deep-sequencing and single cell sequencing