

CANCER GENOMICS

Lecture 1: Introduction to Cancer Genome Analysis

GENOME 541 Spring 2022

April 26, 2022



Gavin Ha, Ph.D.

Public Health Sciences Division
Human Biology Division



@GavinHa



gha@fredhutch.org



<https://github.com/GavinHaLab>

GavinHaLab.org

Overview of Cancer Genomics Module

- 1. Introduction to Cancer Genome Analysis**

- 2. Probabilistic Methods for Mutation Detection**

- 3. Probabilistic Methods for Profiling Copy Number Alteration**

- 4. Additional Topics: Tumor Heterogeneity, Mutation Detection Power, Structural Variation**

Homework Assignments and Office Hours

TA for Module: Anna-Lisa Doebley (adoebley@uw.edu)

Homework #5

Due: May 5th

Virtual Office Hours

- May 3 @ 2pm
- May 6 @ 2pm

Homework #6

Due: May 12th

Virtual Office Hours

- May 10 @ 2pm

Zoom Meeting ID: 446 356 7725

Passcode: GS541

Outline: Introduction to Cancer Genome Analysis

1. Intro to Cancer Genome Alterations

- Genomic alterations in cancer: drivers vs passengers, somatic vs germline
- Tumor evolution and heterogeneity

2. Overview of Cancer Genome Analysis

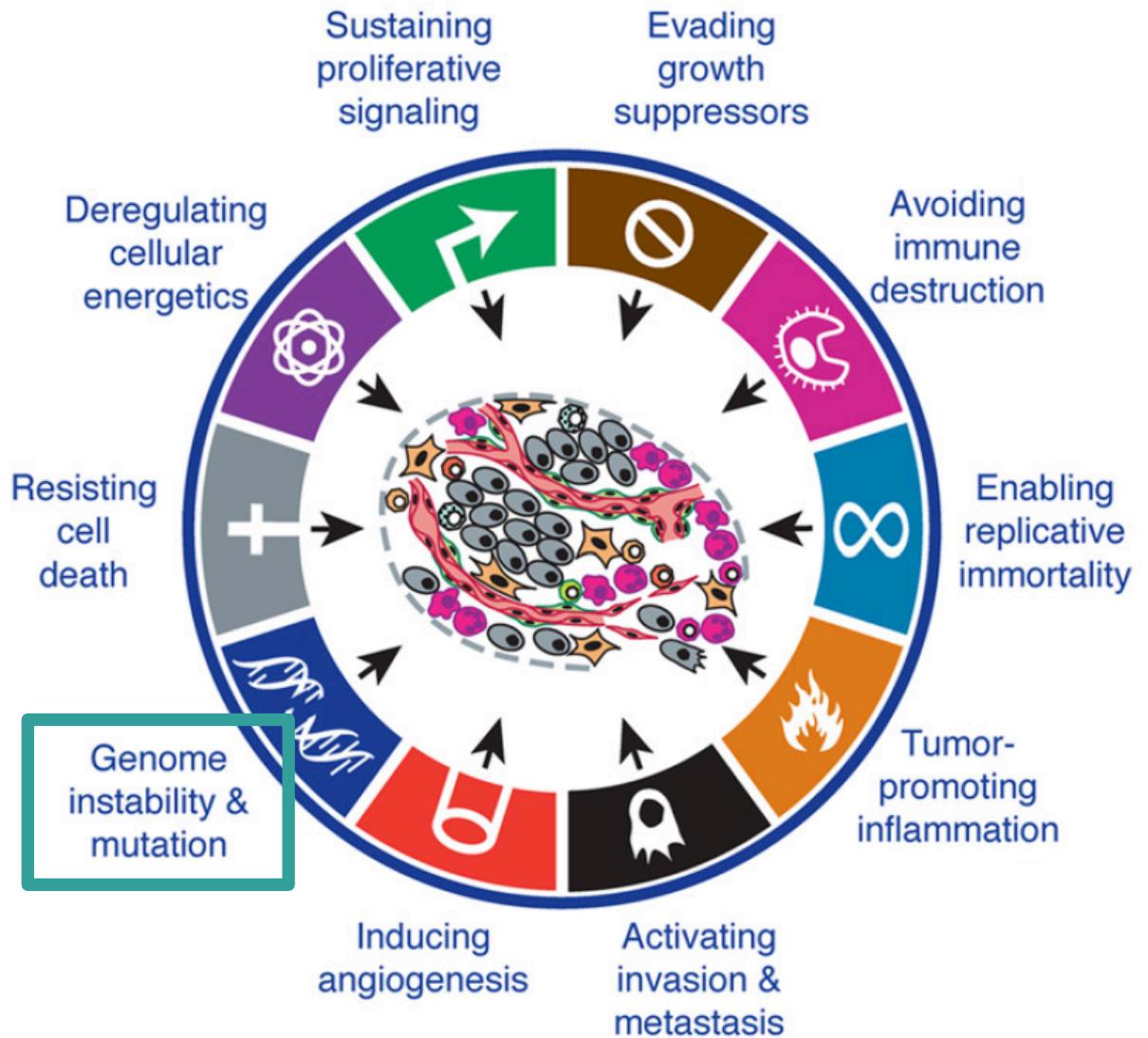
- Computational strategy and workflow
- Tumor DNA Sequencing
- Types of genomic alterations predicted from tumor sequencing
- Methods/tools/algorithms in following lectures

3. Primer on statistical modeling

- Binomial probability distribution, Bayesian statistics, parameter learning

The hallmarks of cancer

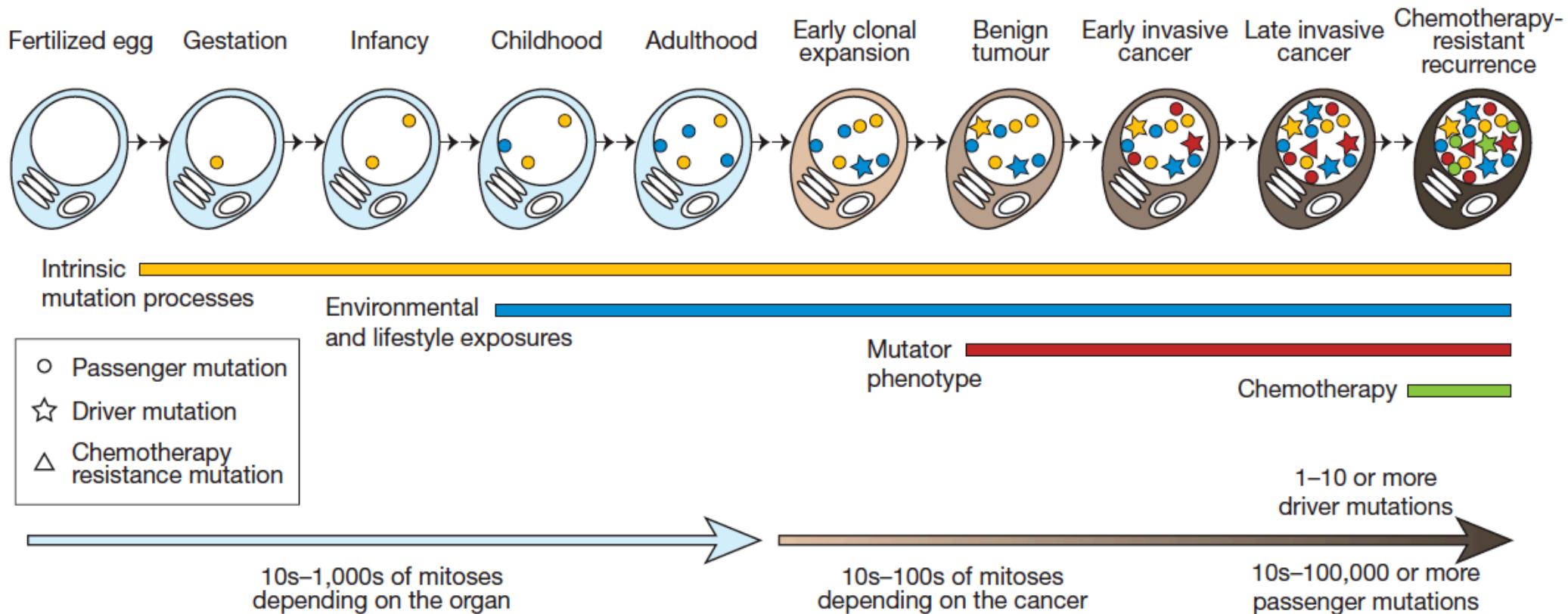
- All cancers exhibit many of these hallmarks that lead to tumor growth
- **Genome instability & mutation** is an enabling characteristic that can result in multiple hallmarks



Cancer is a disease of the genome

Cancer progression results from **mutations** acquired throughout lifetime

- Few **driver** mutations, many **passenger** mutations
- Mutational process can be intrinsic and from environmental mutagens



Genomic Variation: Somatic and Germline

Variant or Mutation or Alteration or Polymorphism

- Changes in the genome sequence of a sample compared to a reference sequence

Germline Variant

- Chromosomes: 22 autosomal pairs + 1 sex pair
 - Each set inherited from maternal and paternal germline cells
- Variant inherited from one or both parental chromosomes
- Source of genetic differences between ancestral populations and individuals
- Polymorphism: >1% frequency in a population

Somatic Variant

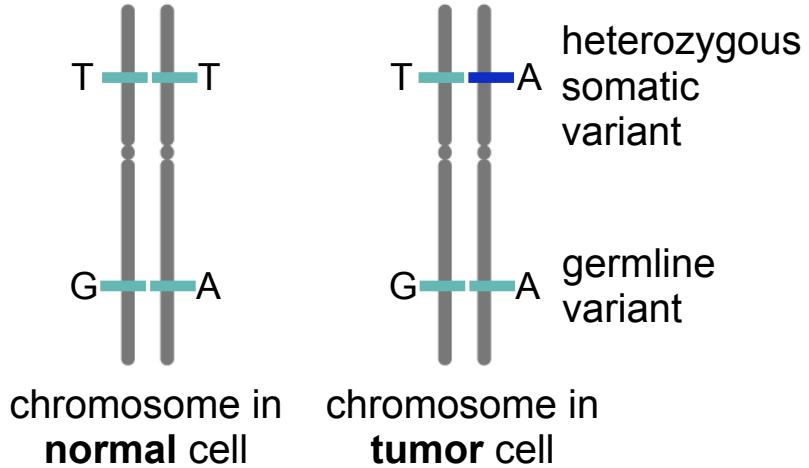
- Mutation acquired during individual's lifetime
- Important to identify in sporadic cancers and other non-familial diseases

Types of Genomic Variation: Small/Short mutations

1. Single nucleotide base substitutions

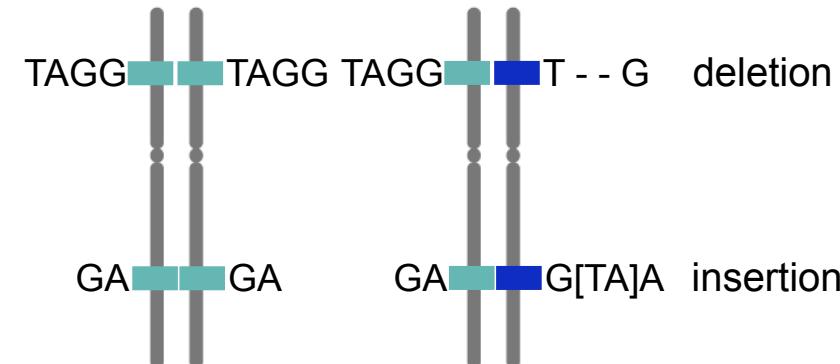
- Germline single nucleotide polymorphism (SNP)
- Somatic single nucleotide variant (SNV)

Single nucleotide variant



2. Small insertions or deletions

- Germline or somatic insertion or deletion (INDEL)
- Small indels: 1 bp - 20 bps
- Large indels: 20 - 10,000 bps



Insertion-Deletion (INDEL)

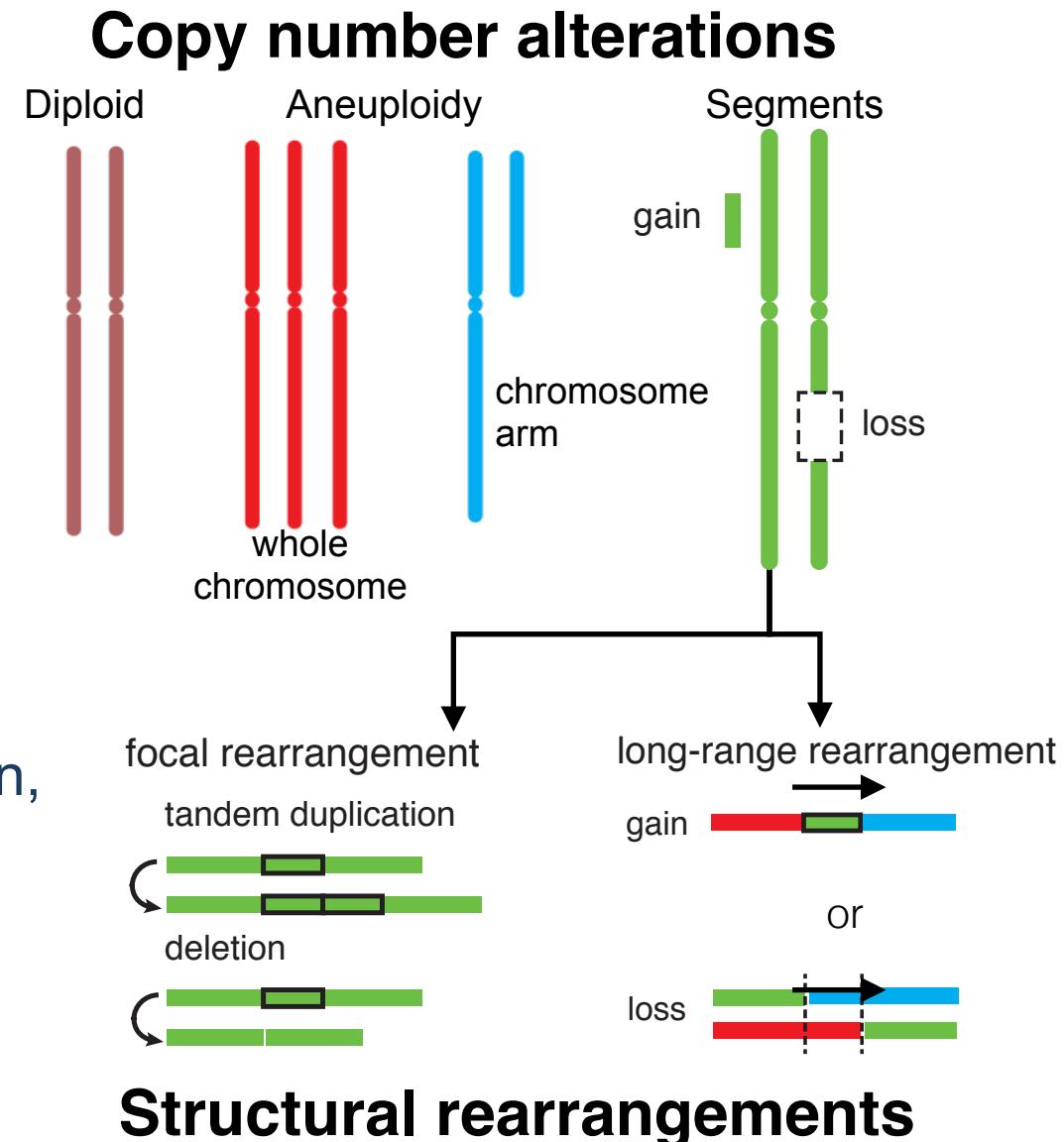
Types of Genomic Variation: Large alterations

3. Copy number changes

- Germline copy number variant (CNV) or polymorphism (CNP)
- Somatic copy number variant (CNV) or alterations (CNA)
- Size > 1 kbps, typically mega-bases (depending on resolution)

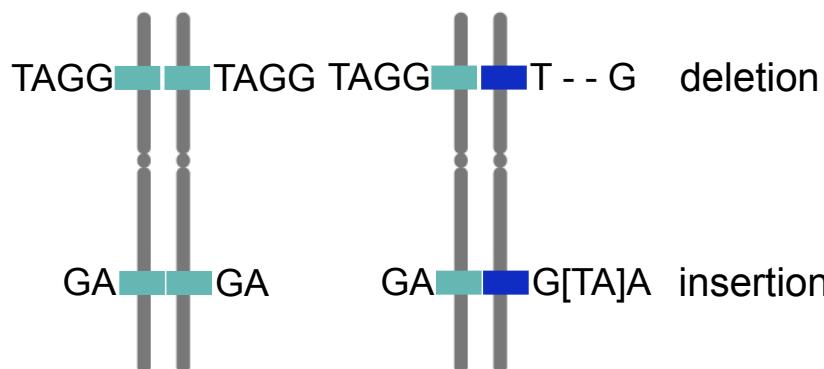
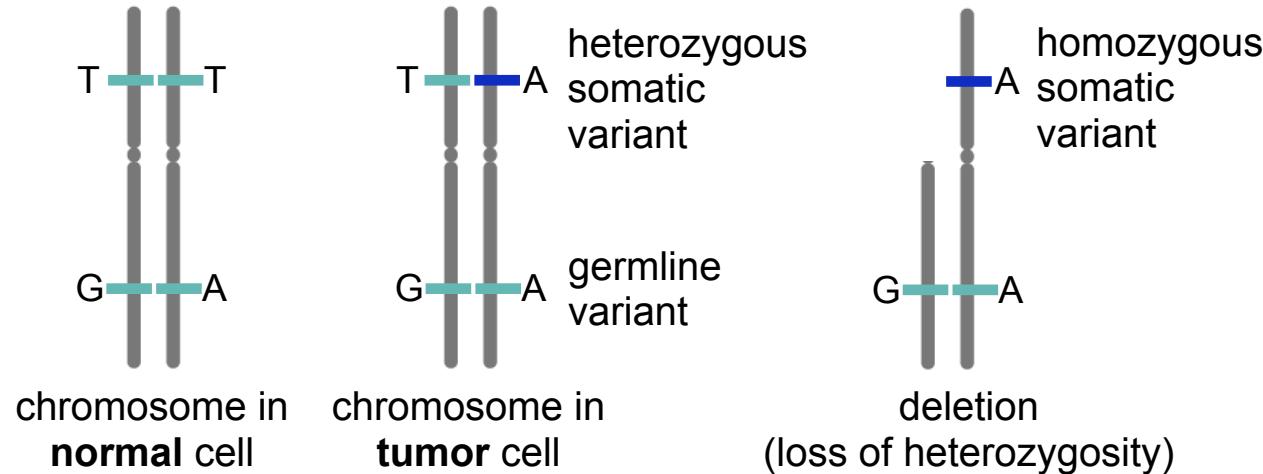
4. Structural rearrangements

- Germline or Somatic structural variant (SV)
- Simple events: deletion, duplication, inversion, translocation
- Single nucleotide resolution for breakpoints
- Size > 20 bps, typically kilo-bases to mega-bases



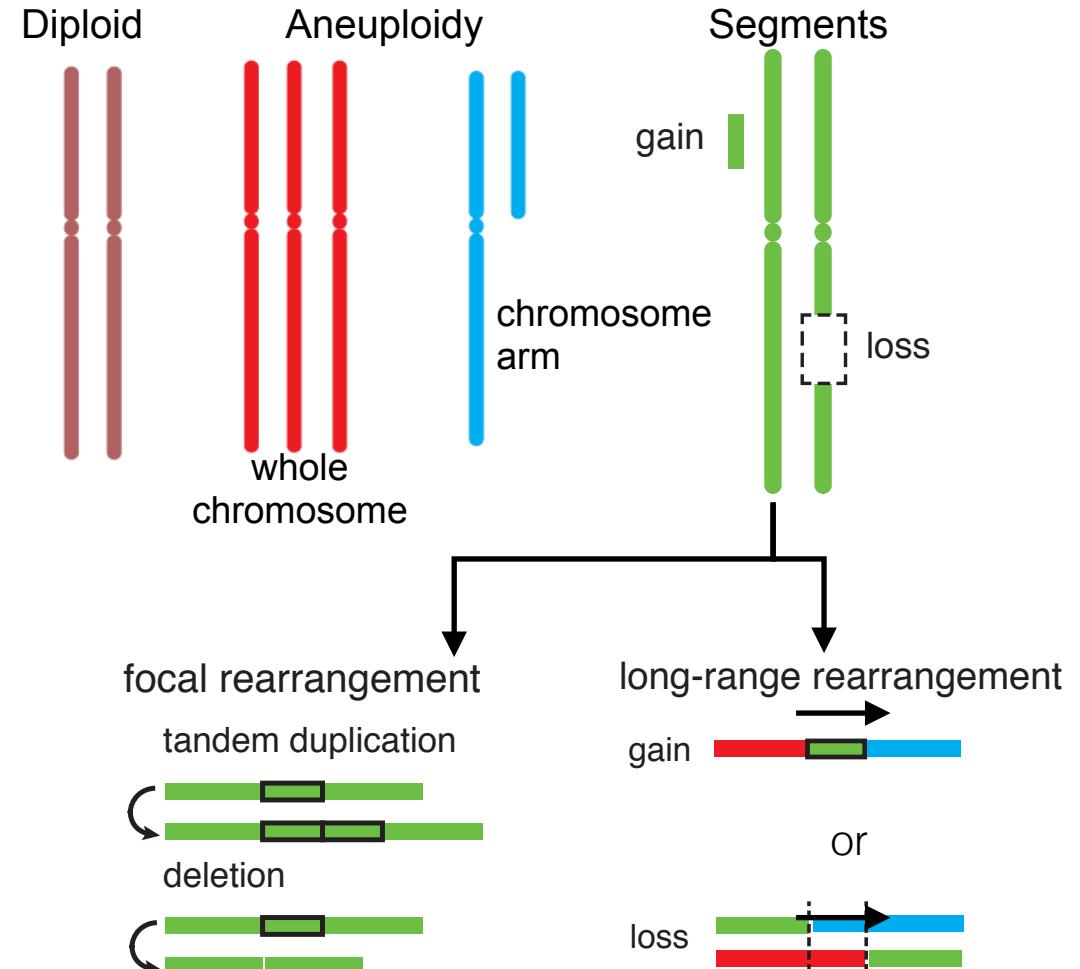
Types of Genomic Variation in Cancer

Single nucleotide variant



Insertion-Deletion (INDEL)

Copy number alterations



Structural rearrangements

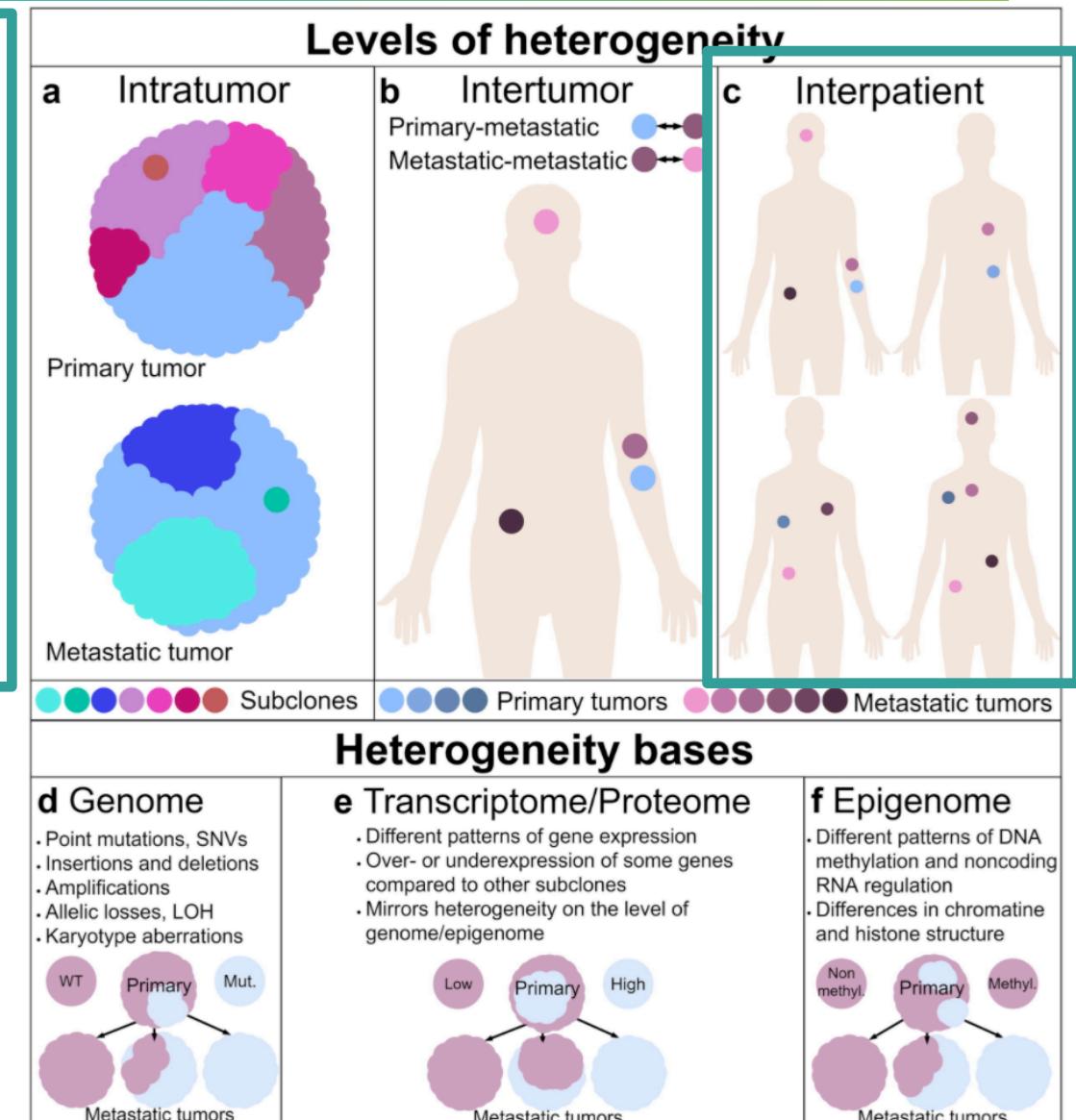
Tumors exhibit different levels of heterogeneity

Across patient populations:

1. **Cancer types:** between primary tumors of different organs or tissue-of-origin (eg. Breast and lung cancers)
2. **Tumor subtypes:** between subset of patients with tumors having similar molecular features (e.g. ER+ and ER- breast cancers)
3. **Same-subtype:** between tumors from different patients

Within an individual patient:

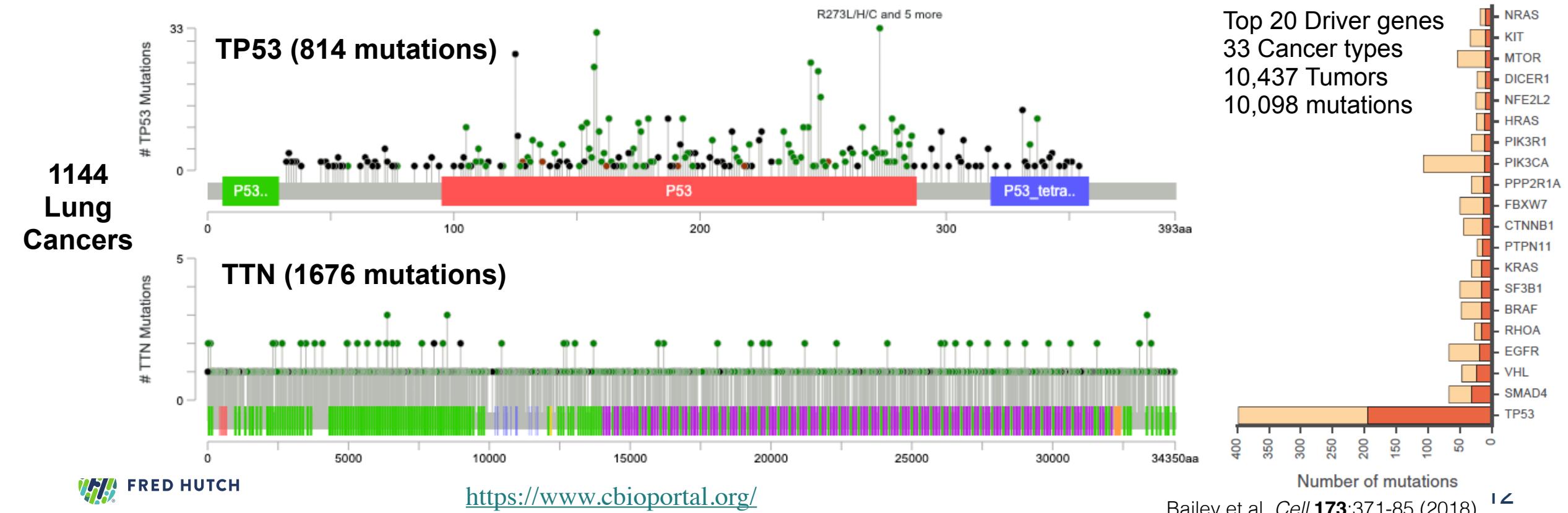
4. **Inter-tumor:** between tumors within a patient
5. **Intra-tumor heterogeneity:** between cells within a tumor lesion (e.g. tumor clones, stromal cells, infiltrating lymphocytes)



Cancer Genes: Driver vs Passenger Genomic Alterations

How do we find the mutated genes that *drive* cancer?

- **Significantly Mutated Genes:** recurrently mutated genes in patient cohorts
- Account for covariates (e.g. gene length, expression, replication timing)



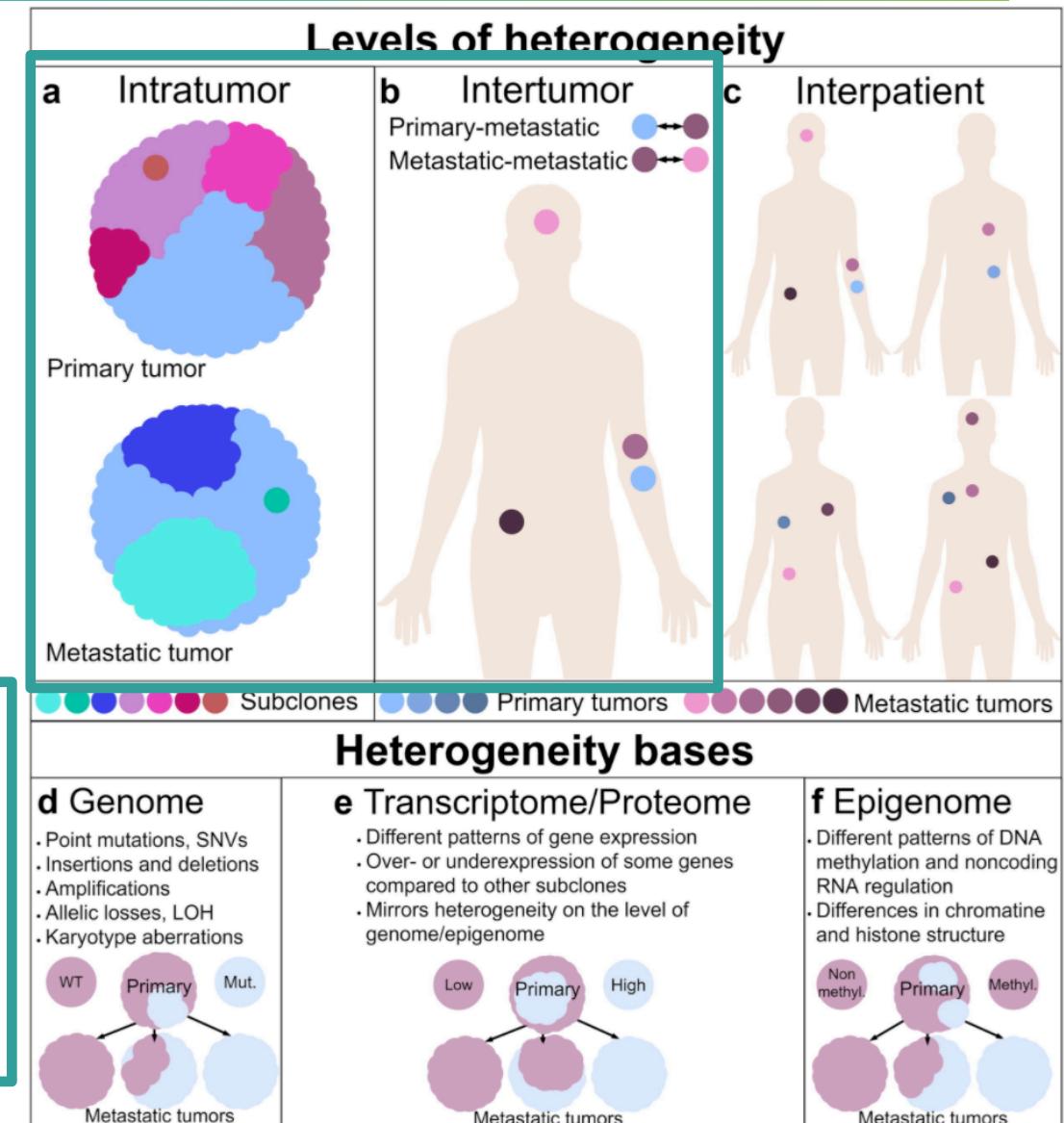
Tumors exhibit different levels of heterogeneity

Across patient populations:

1. **Cancer types:** between primary tumors of different organs or tissue-of-origin (eg. Breast and lung cancers)
2. **Tumor subtypes:** between subset of patients with tumors having similar molecular features (e.g. ER+ and ER- breast cancers)
3. **Same-subtype:** between tumors from different patients

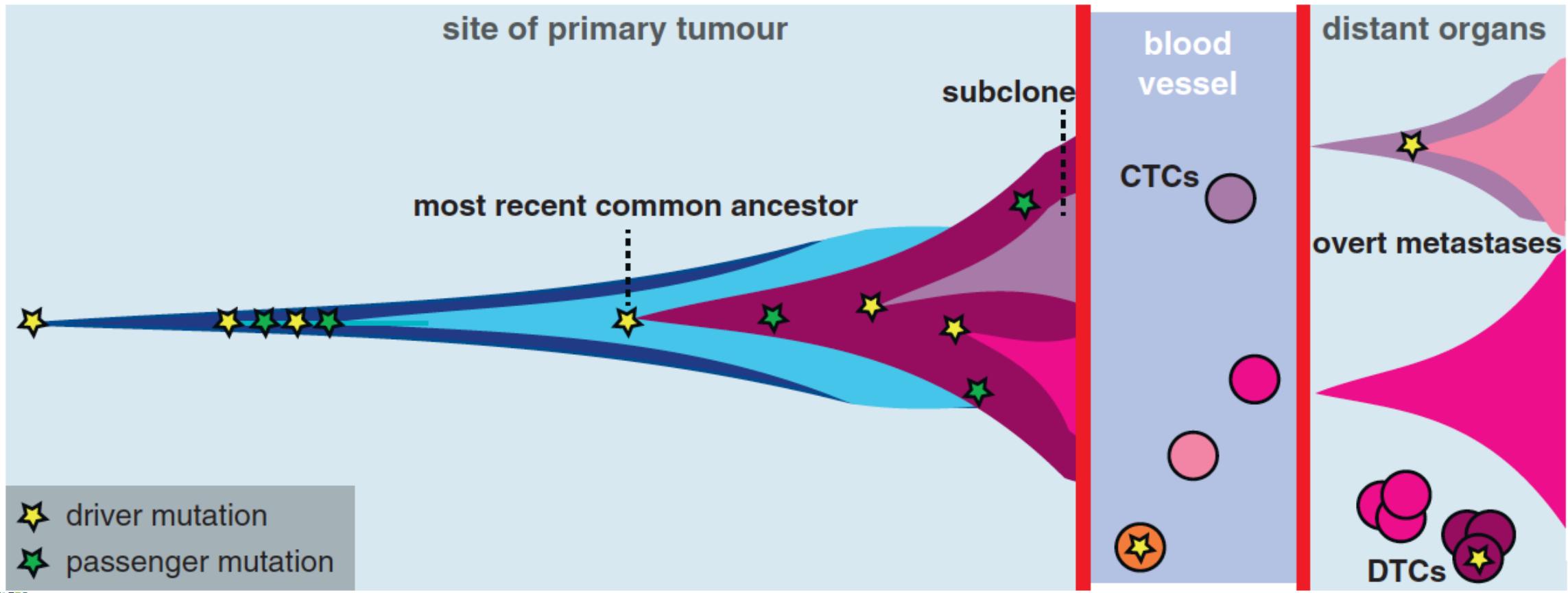
Within an individual patient:

4. **Inter-tumor:** between tumors within a patient
5. **Intra-tumor heterogeneity:** between cells within a tumor lesion (e.g. tumor clones, stromal cells, infiltrating lymphocytes)

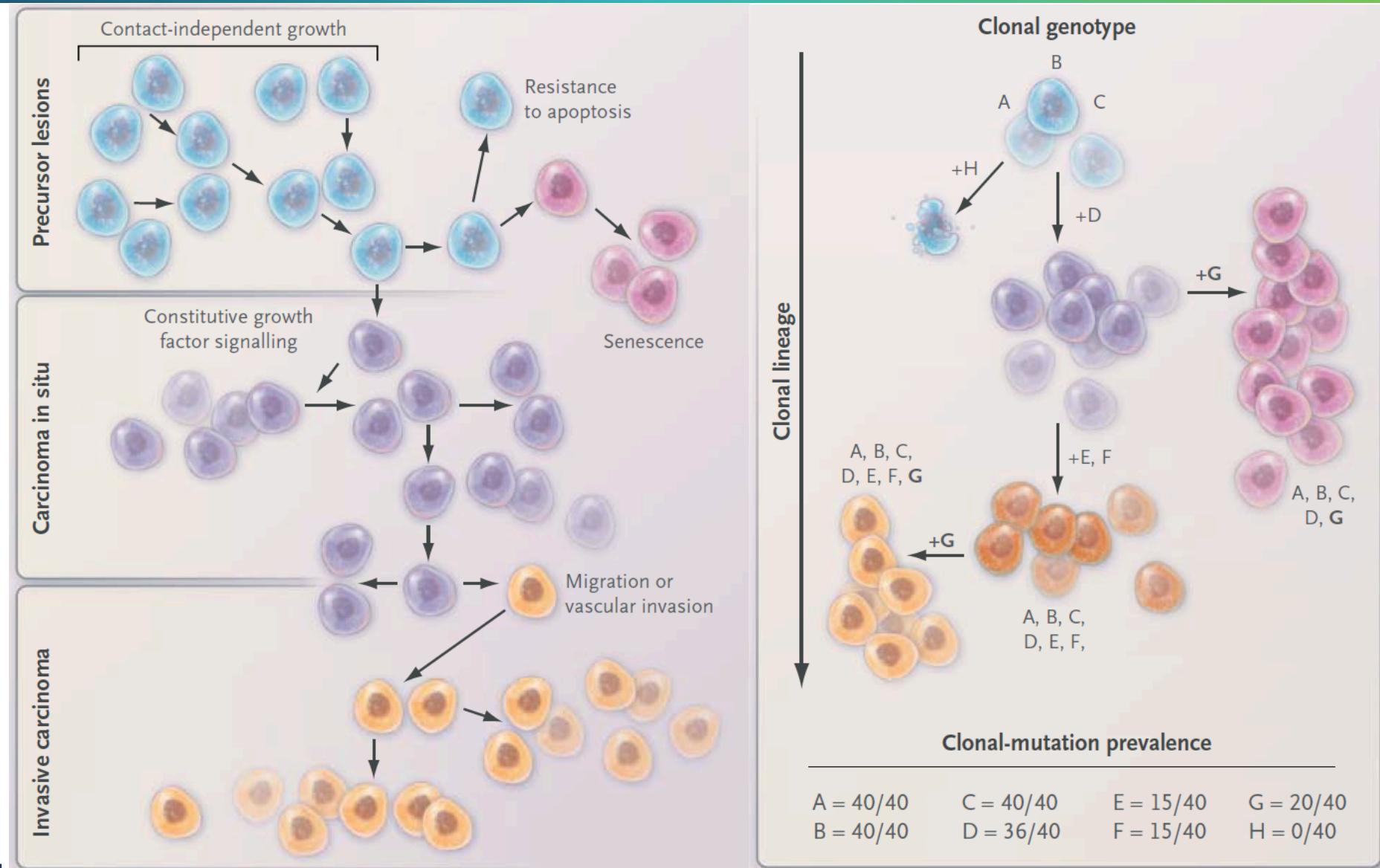


Tumors undergo genome evolution and clonal expansion

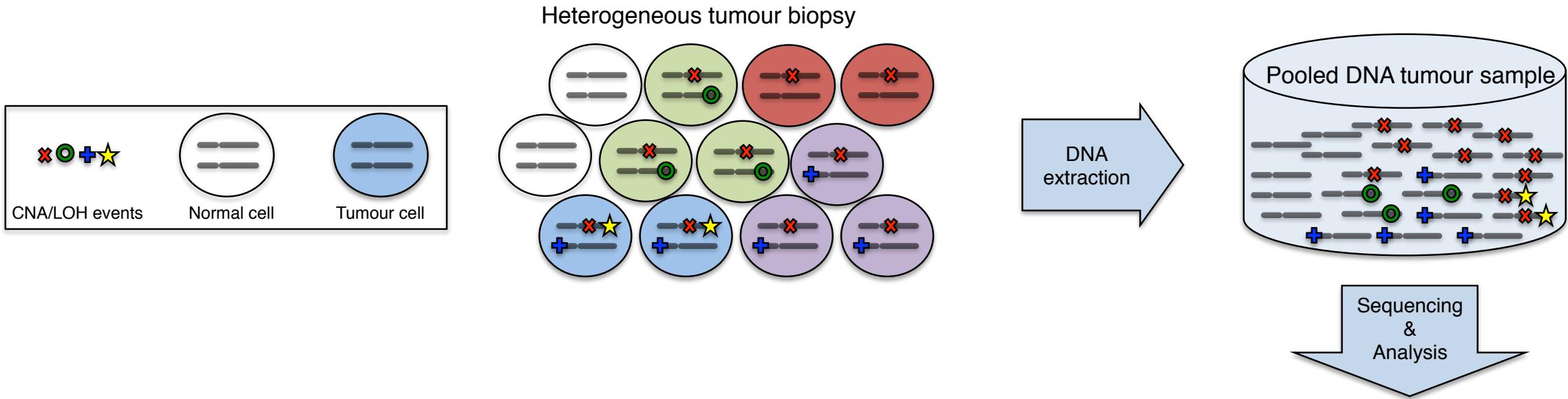
- Clonal diversity may have implications for treatment resistance
- Dynamics of clones can change in the blood and metastases



Tumor genome evolution selects for cellular phenotypes



Inferring intra-tumor genomic heterogeneity from sequencing



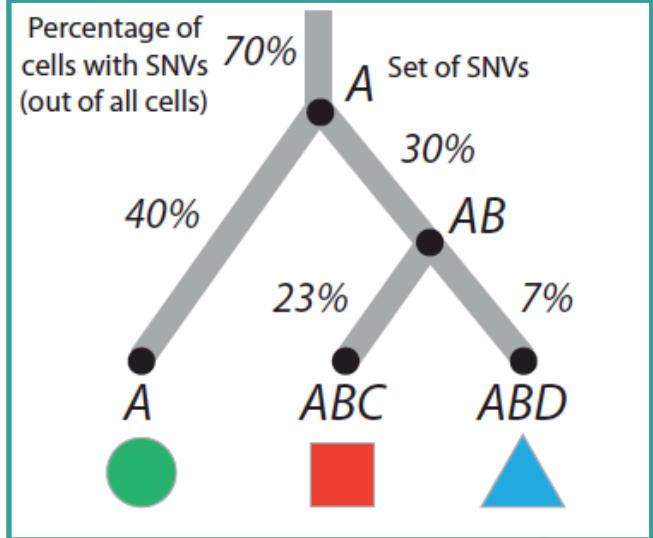
- Combined signals from normal and multiple populations of tumor cells.
- Cellular prevalence: proportion of tumor cells harboring event
- Discuss further in Lecture 4...

Subclonal events
Cellular prevalence

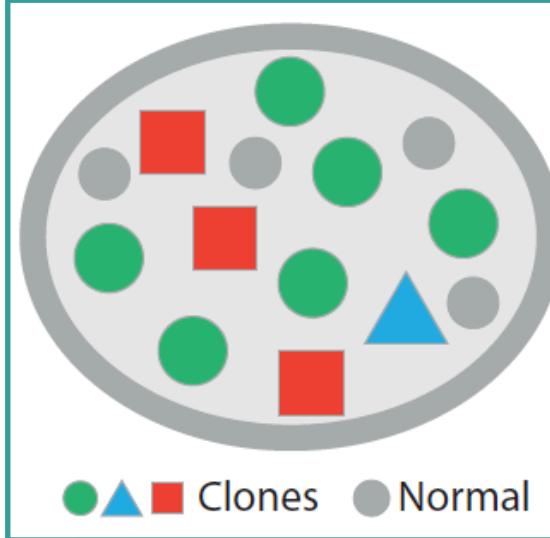
- ✖ 100%
- ✚ 50%
- 30%
- ★ 20%

Inferring evolutionary history of a tumor from sequencing

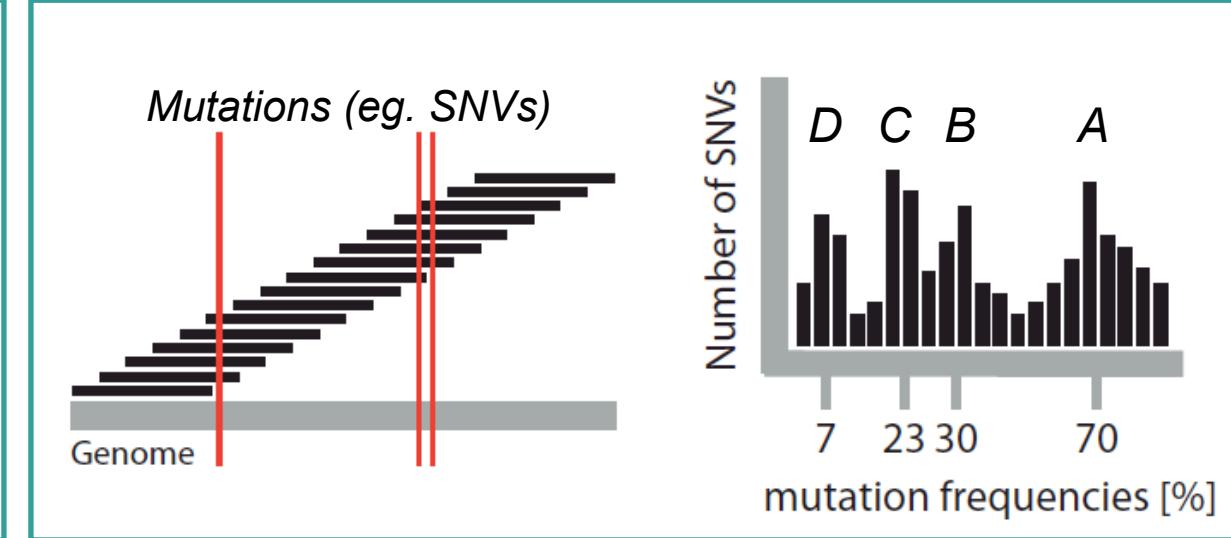
Evolutionary History



Clonal Cell Populations



Sequencing Data



3. Infer evolutionary (phylogenetic) tree

2. Infer clonal prevalence

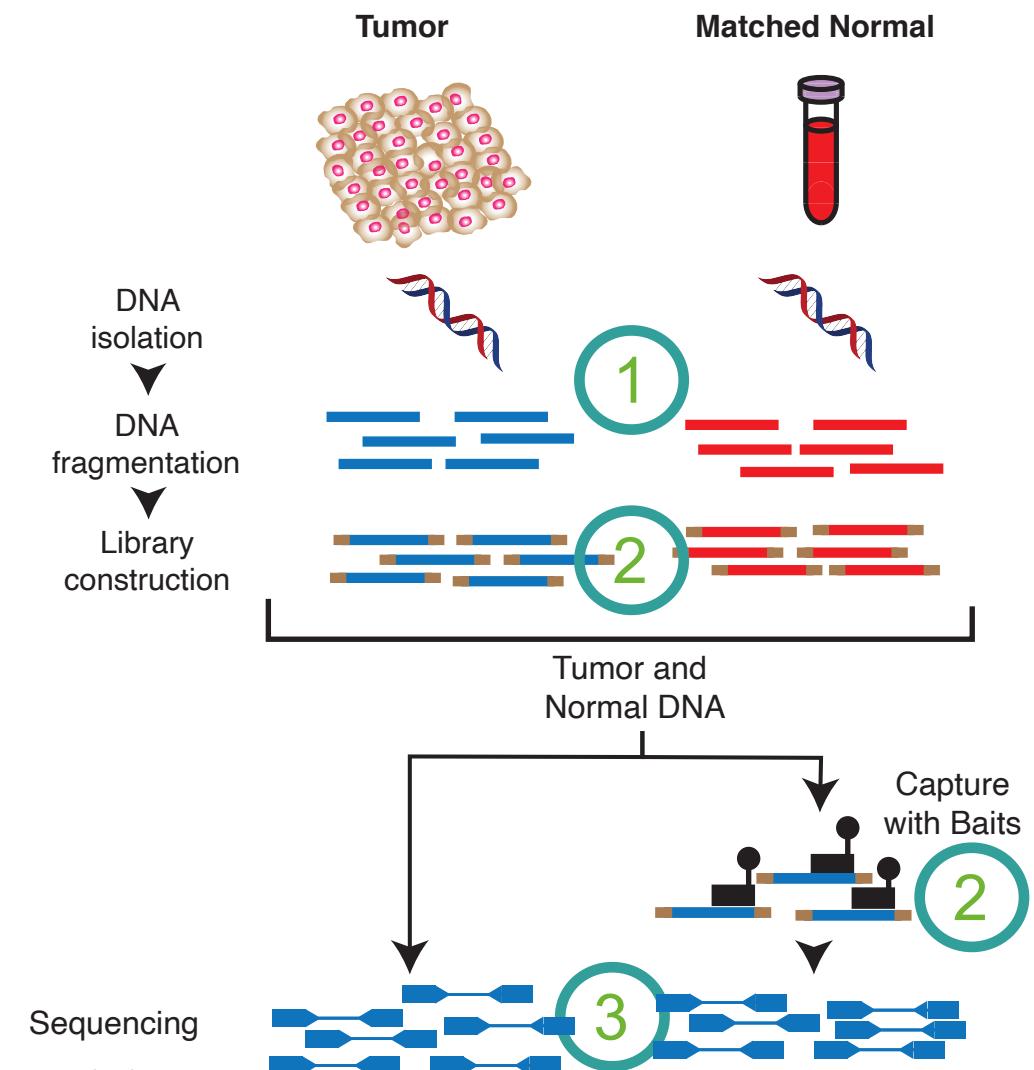
1. Mutation Calling & Analysis

2. Overview of Cancer Genome Analysis

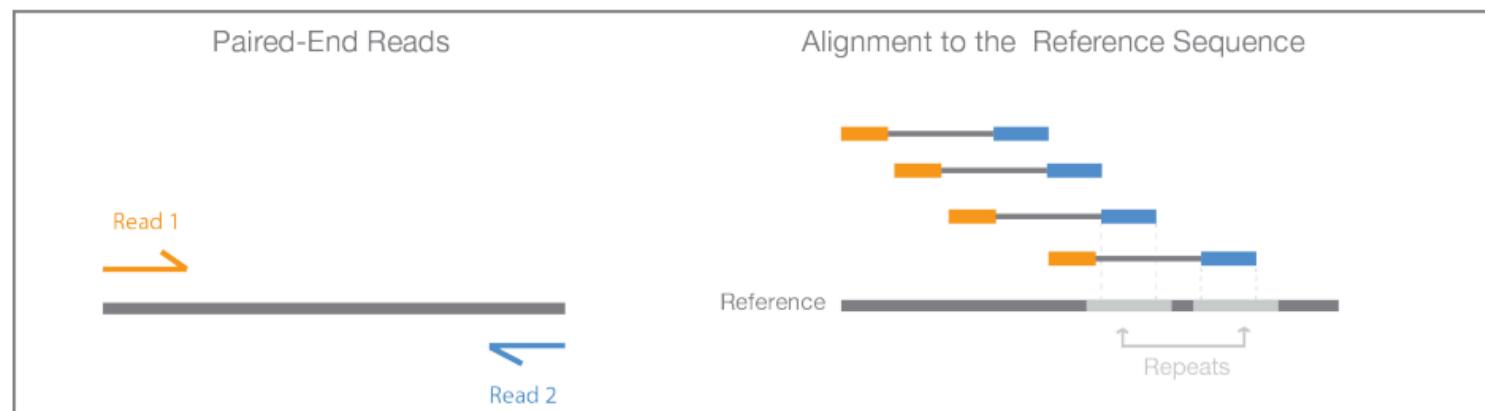
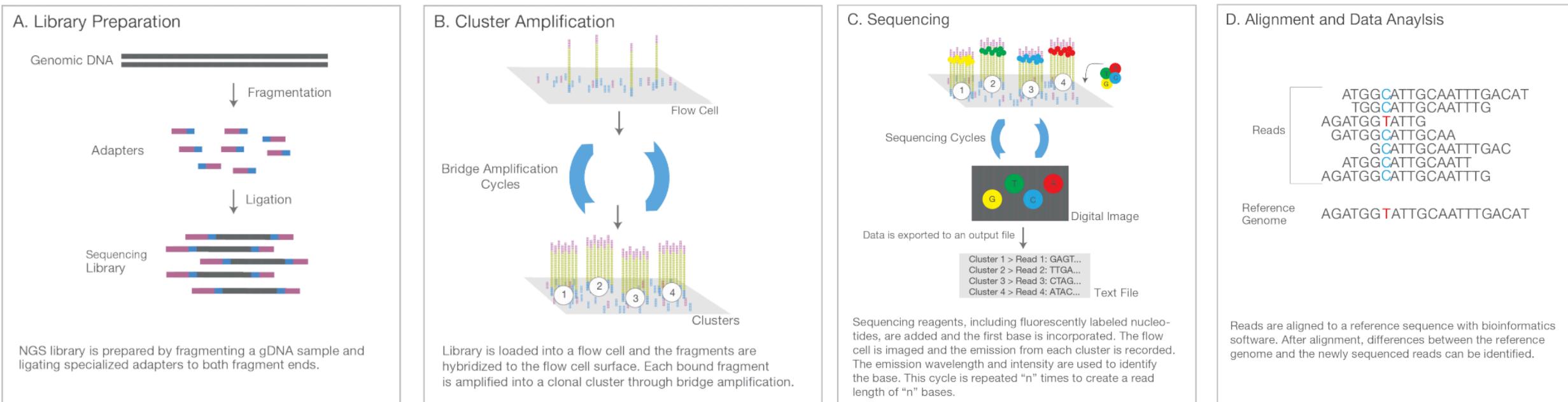
- Computational strategy and workflow
- Tumor DNA sequencing
- Whole genome vs whole exome vs targeted sequencing
- Types of genomic alterations predicted from tumor sequencing
- Methods/tools/algorithms in following lectures

General Workflow of Tumor Genome Sequencing (1)

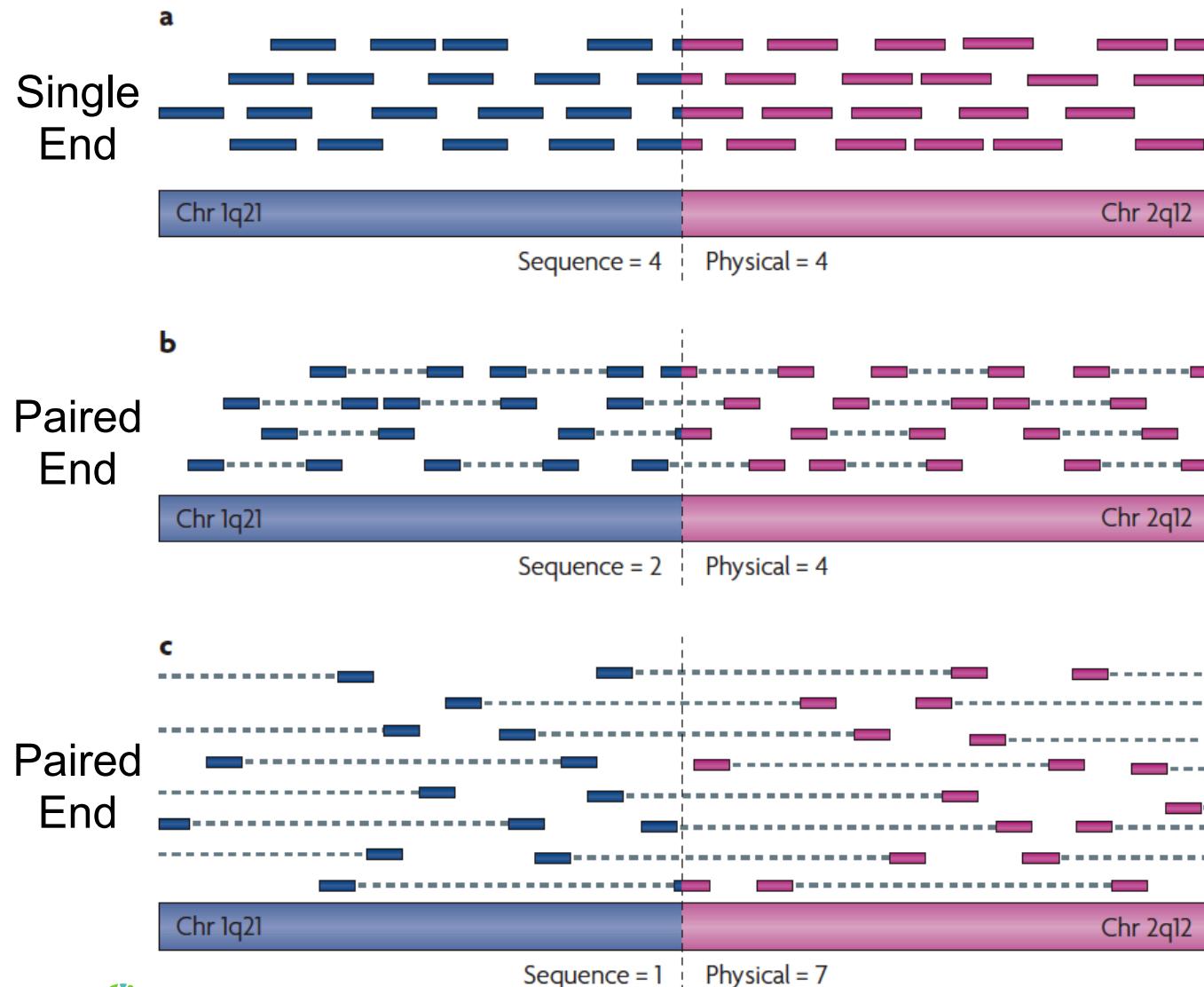
- Tumor and Normal pairing
 - Distinguish somatic and germline alterations
- Capture baits can be used to select regions
 - e.g. whole exome or targeted gene panels
- Potential sources of error can arise
 1. 8-oxoG transversions (C>A/G>T)
 2. PCR errors and GC content bias
 3. Sequencing errors



Genome Sequencing: Massively Parallel Sequencing



Genome Sequencing: Sequence vs Physical Coverage

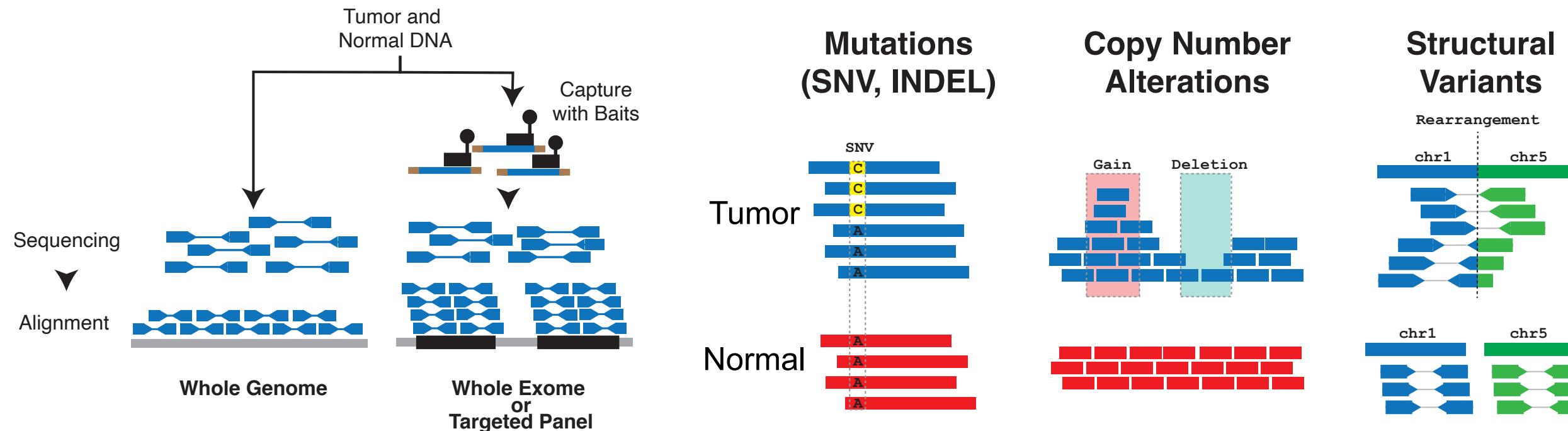


Sequence Coverage = number of sequenced reads spanning locus

Physical Coverage = number of DNA fragments spanning locus

- Mutation detection rely on sequence coverage
- Rearrangement detection rely on both

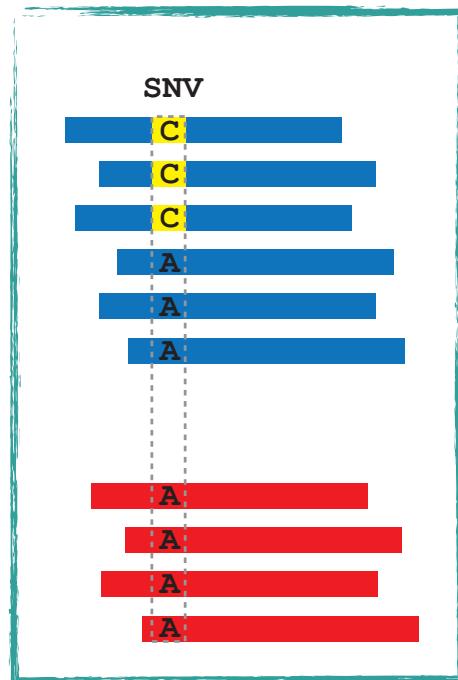
General Workflow of Tumor Genome Sequencing (2)



Whole Genome Sequencing	Whole Exome Sequencing	Targeted Gene Sequencing
<ul style="list-style-type: none"> • Genome-wide (unbiased) • 0.1-100x genome coverage 	<ul style="list-style-type: none"> • Exons (2% of genome) • 50-500x target coverage 	<ul style="list-style-type: none"> • Target regions (1-5Mb) • 100-25000x target coverage
<ul style="list-style-type: none"> • More sequencing required • Expensive 	<ul style="list-style-type: none"> • Less sequencing required • Cost-effective 	<ul style="list-style-type: none"> • Least sequencing required • Panel design costs
<ul style="list-style-type: none"> • Coding/Non-coding mutations • Copy number alterations • Structural variation 	<ul style="list-style-type: none"> • Coding mutations (all genes) • Copy number alterations • Gene fusions rearrangements 	<ul style="list-style-type: none"> • Coding mutations (selected) • Targeted rearrangements

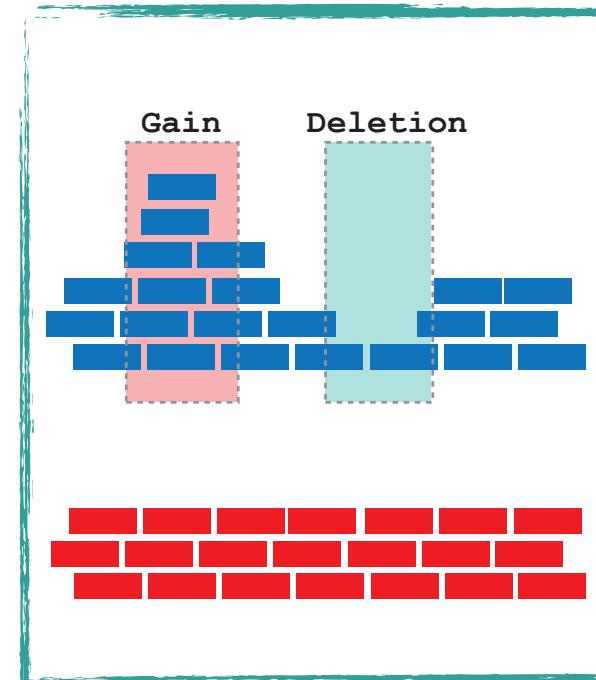
Types of Genomic Alterations Predicted from Sequencing

Mutations (SNV, INDEL)



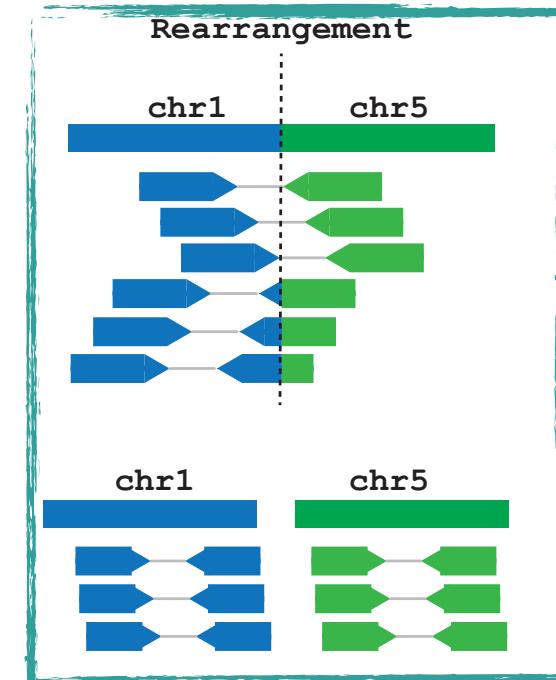
Lecture 2

Copy Number Alterations



Lecture 3

Structural Variants



Lecture 4?

Genome Sequencing: International Consortia & Projects

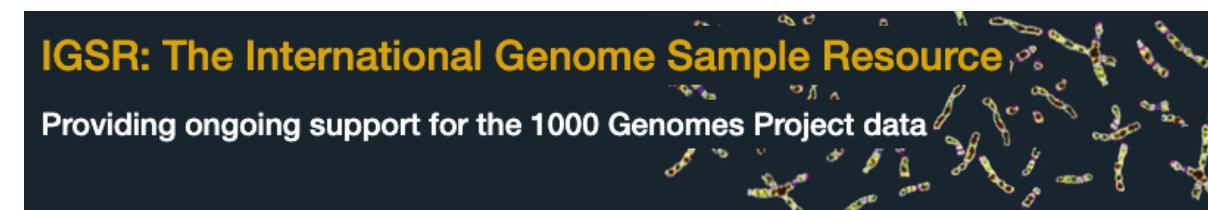
1000 Genomes Project (<https://www.internationalgenome.org/>)

UK10K (<https://www.uk10k.org/>)

The 100,000 Genomes Project

(<https://www.genomicsengland.co.uk/>)

- Rare disease, cancer, infectious disease



UK10K

Rare Genetic Variants in Health and Disease



#100kThankYous

Genome 10K Project (<https://genome10k.soe.ucsc.edu/>)

- Genomic “zoo” of 16,000 vertebrate species



Exome Aggregation Consortium (ExAC) (<http://exac.broadinstitute.org/>)

Genome Aggregation Database (gnomAD) (<https://gnomad.broadinstitute.org/>)

The Cancer Genome Atlas (TCGA) (<https://portal.gdc.cancer.gov/>)

International Cancer Genome Consortium (ICGC) (<https://icgc.org/>)



Cancer Genome Sequence Data: Databases & Online Resources

NATIONAL CANCER INSTITUTE
GDC Data Portal

Home Projects Exploration Analysis Repository

Quick Search Manage Sets Login Cart 0 GDC Apps

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

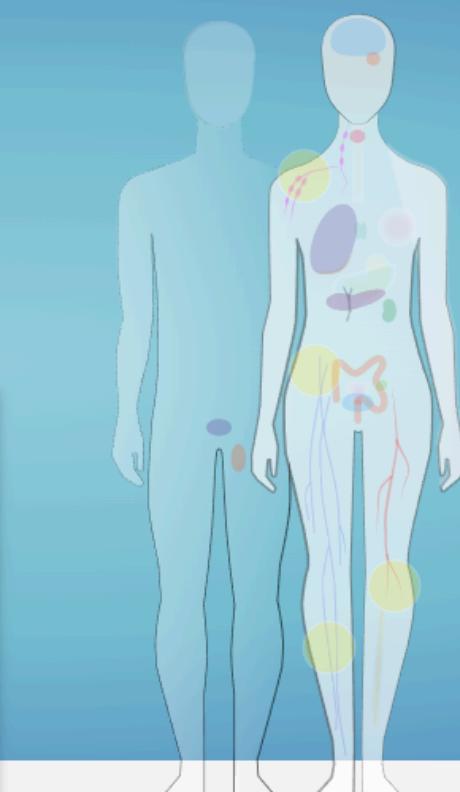
Projects Exploration Analysis Repository

e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Data Portal Summary

Data Release 22.0 - January 16, 2020

PROJECTS	PRIMARY SITES	CASES
64	67	83,709
FILES	GENES	MUTATIONS
526,931	22,872	3,142,246



Cases by Major Primary Site

Primary Site	Cases (in 1000s)
Adrenal Gland	0.1
Bile Duct	0.1
Bladder	1.5
Blood	0.5
Bone	0.5
Bone Marrow	9.0
Brain	1.2
Breast	8.5
Cervix	0.5
Colorectal	7.8
Esophagus	1.2
Eye	0.2
Head and Neck	1.8
Kidney	3.2
Liver	1.2
Lung	11.2
Lymph Nodes	0.5
Nervous System	3.5
Ovary	3.2
Pancreas	2.5
Pleura	0.5
Prostate	2.5
Skin	2.5
Soft Tissue	1.2
Stomach	1.2
Testis	0.5
Thymus	0.5
Thyroid	1.2
Uterus	2.5

Cancer Genome Sequence Data: Databases & Online Resources

cBioPortal FOR CANCER GENOMICS

Data Sets Web API R/MATLAB Tutorials FAQ News Visualize Your Data About Login

Query Quick Search Beta! Download Please cite: Cerami et al., 2012 & Gao et al., 2013

Select Studies for Visualization & Analysis: 0 studies selected (0 samples) Search...

Quick select: TCGA PanCancer Atlas Studies Curated set of non-redundant studies

PanCancer Studies

- PanCancer Studies 3
- Cell lines 3
- Adrenal Gland 3
- Ampulla of Vater 1
- Biliary Tract 9
- Bladder/Urinary Tract 15
- Bone 2
- Bowel 10
- Breast 16
- CNS/Brain 19
- Cervix 2
- Esophagus/Stomach 14
- Eye 3
- Head and Neck 13
- Kidney 17
- Liver 8
- Lung 21
- Lymphoid 20
- Myeloid 9
- Other 15
- Ovary/Fallopian Tube 4

PanCancer Studies

- MSK-IMPACT Clinical Sequencing Cohort (MSKCC, Nat Med 2017) 10945 samples   
- Pan-Lung Cancer (TCGA, Nat Genet 2016) 1144 samples   
- Pediatric Pan-cancer (Columbia U, Genome Med 2016) 103 samples   

Cell lines

- Cancer Cell Line Encyclopedia (Broad, 2019) 1739 samples   
- Cancer Cell Line Encyclopedia (Novartis/Broad, Nature 2012) 1020 samples   
- NCI-60 Cell Lines (NCI, Cancer Res 2012) 67 samples   

Adrenal Gland

Adrenocortical Carcinoma

- Adenoid Cystic Carcinoma Project (2019) 1049 samples   
- Adrenocortical Carcinoma (TCGA, Firehose Legacy) 92 samples   
- Adrenocortical Carcinoma (TCGA, PanCancer Atlas) 92 samples   

Ampulla of Vater

Ampullary Carcinoma

- Ampullary Carcinoma (Baylor College of Medicine, Cell Reports 2016) 160 samples   

Biliary Tract

Cholangiocarcinoma

- Cholangiocarcinoma (MSK, Clin Cancer Res 2018) 195 samples   
- Cholangiocarcinoma (National Cancer Centre of Singapore, Nat Genet 2012) 15 samples   
- Cholangiocarcinoma (National University of Singapore, Nat Genet 2012) 8 samples   
- Cholangiocarcinoma (TCGA, Firehose Legacy) 51 samples   
- Cholangiocarcinoma (TCGA, PanCancer Atlas) 36 samples   
- Intrahepatic Cholangiocarcinoma (JHU, Nat Genet 2013) 40 samples   

→ INTRAHEPATIC CHOLANGIOCARCINOMA

- Intrahepatic Cholangiocarcinoma (Shanghai, Nat Commun 2014) 103 samples   

Gallbladder Cancer https://www.cbiportal.org/

What's New @cbiportal 

cBioPortal 

We are hosting a webinar series to teach cBioPortal features to beginner and advanced users. Sessions will be held on five consecutive Thursdays at 11 AM EDT, starting on April 30th. Please register here: bit.ly/cbiportal-web...

Subscribe

Cancer Studies

The portal contains 283 cancer studies ([details](#))

Cases by Top 20 Primary Sites



Primary Site	Number of Cases
Breast	~9,500
Prostate	~7,500
CNS/Brain	~6,500
Lung	~6,000
Lymphoid	~5,500
Bowel	~4,500
Kidney	~4,000
Stomach	~3,500
Myeloid	~3,000
Bladder	~2,500
Skin	~2,000
Uterus	~1,500
Head/Neck	~1,500
Ovary	~1,500
Thyroid	~1,500
Liver	~1,500
PNS	~1,000
Adrenal Gland	~1,000
Pancreas	~1,000
Soft Tissue	~1,000

Cancer Genome Sequence Data: Databases & Online Resources



Cancer genomics data sets visualization, analysis and download.

Quick Search

Search

e.g. BRAF, KRAS G12D, DO35100, MU7870, Fl998, apoptosis, Cancer Gene Census, imatinib, GO:0016049

Advanced Search

By donors

By genes

By mutations

Data Release 28

March 27th, 2019

Cancer projects 86

Cancer primary sites 22

Donor with molecular data in DCC 22,330

Total Donors 24,289

Simple somatic mutations 81,782,588

Download Release

3. Primer on statistical modeling

- Probability
 - Unsupervised learning, probability rules & Bayes' theorem
 - Binomial distribution, Bayesian statistics
 - Beta-binomial model example
- Mixture models, EM inference
- References:
 - Murphy, K. (2012). Machine Learning: A Probabilistic Perspective. MIT Press. ISBN: 9780262018029
 - Bishop, C. M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics). Springer. ISBN: 0387310738
 - <https://www.cs.ubc.ca/~murphyk/Teaching/CS340-Fall06/reading/bernoulli.pdf>

Sequencing Data Analysis Requires Probabilistic Models

- Sequencing data contain uncertainty due to
 - Technical noise from imperfect measurements & errors
 - Biological features in the signal measurements
- How do we predict genomic alterations accounting for these features and noise?
 - Need approaches to learn the patterns of these features from the data...

Types of machine learning:

- Supervised: output data y , input data x , and *training set* $D = \{(x, y)\}$
 - Classification (y are labels), Regression (y is continuous)
- Unsupervised: Only given input data $D = \{x\}$, *learn the patterns of the data*
 - E.g. clustering input data x into K clusters by estimating their assignments z

Primer: Probability Theory

Let X be a random variable. The probability for the event $X = x$ for some value x is $p(X = x)$ or $p(x)$ for short. Let Y be another random variable.

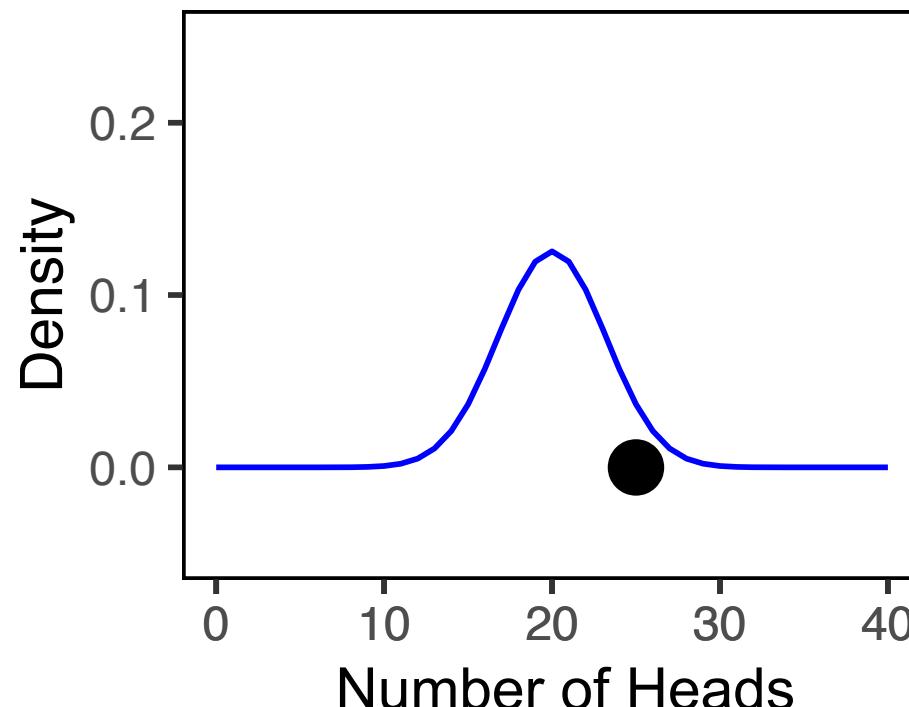
Probability Rules

- **Sum rule:** $p(X) = \sum_Y p(X, Y)$
- **Product rule:** $p(X, Y) = p(Y|X)p(X)$ and $p(Y, X) = p(X|Y)p(Y)$
- Conditional Probabilities: $p(Y|X) = \frac{p(X, Y)}{p(X)}$
- Marginal Probabilities: $p(X) = \sum_Y p(Y, X) = \sum_Y p(X|Y)p(Y)$
- **Bayes' Theorem (rule):** $p(Y|X) = \frac{p(X, Y)}{p(X)} =$

Probability distribution: Binomial

Binomial Distribution: Referee Coin Toss Example

- A referee has a coin that he uses to decide which team gets first possession. She tossed the coin N times last season, once per game. We assume this coin was fair and had a probability $\mu = 0.5$ for showing a head. We kept track of the number of heads x that appeared.
- What is the probability of seeing a specific number of heads? e.g. $x = 25$ out of $N = 40$ tosses



Probability distribution: Binomial

Binomial Distribution: Referee Coin Toss Example

- A referee has a coin that he uses to decide which team gets first possession. She tossed the coin N times last season, once per game. We assume this coin was fair and had a probability $\mu = 0.5$ for showing a head. We kept track of the number of heads x that appeared.
- What is the probability of seeing a specific number of heads? e.g. $x = 25$ out of $N = 40$ tosses

Probability mass function

- Let X be the random variable representing the number of heads. If the probability of heads is μ , then X has a binomial distribution, $X \sim \text{Bin}(N, \mu)$ or $p(X = x | N, \mu) = \text{Bin}(x | N, \mu)$ where

$$\text{Bin}(x | N, \mu) = \binom{N}{x} \mu^x (1 - \mu)^{N-x}$$

$$\binom{N}{k}$$

number of ways the 25 heads
is observed among the sequence of
40 tosses.

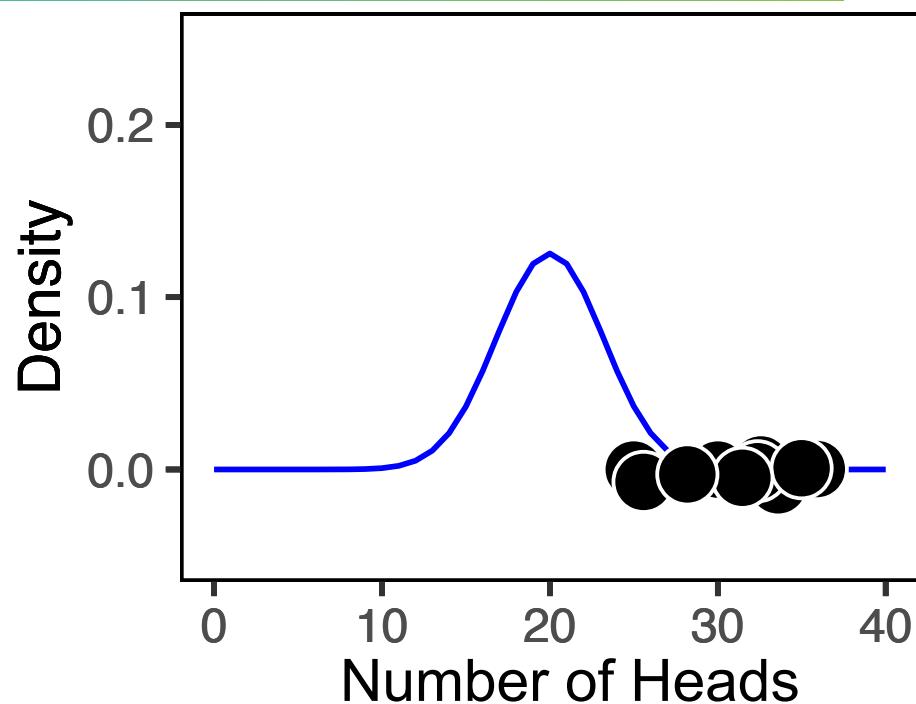
- Our coin-toss example: for $x = 25$ out of $N = 40$ and a fair coin $\mu = 0.5$

Binomial likelihood model

- Suppose there are T different referees who toss the same coin $N = \{1, \dots, N_T\}$ times and come up with head counts $x = \{1, \dots, x_T\}$.
- Assuming the referees' tosses are *independent* and *identically distributed (iid)*, what is the probability of observing the head counts given the coin (e.g. $\mu = 0.5$)?

$$p(x_{1:T} | N_{1:T}, \mu) = \prod_{i=1}^T \text{Bin}(x_i | N_i, \mu) \quad \text{Likelihood}$$

- What if the coin wasn't fair and the probability of heads, μ , might not be 0.5?



	# of tosses (N)	# of heads (x)
Referee 1	40	25
Referee 2	42	35
Referee 3	39	27
Referee T	x_T	N_T

Maximum likelihood estimation (MLE)

- What is the probability of heads, μ , of this coin given the evidence?
- We can estimate this model parameter using
maximum likelihood estimation

$$p(x_{1:T} | N_{1:T}, \mu) = \prod_{i=1}^T \text{Bin}(x_i | N_i, \mu)$$

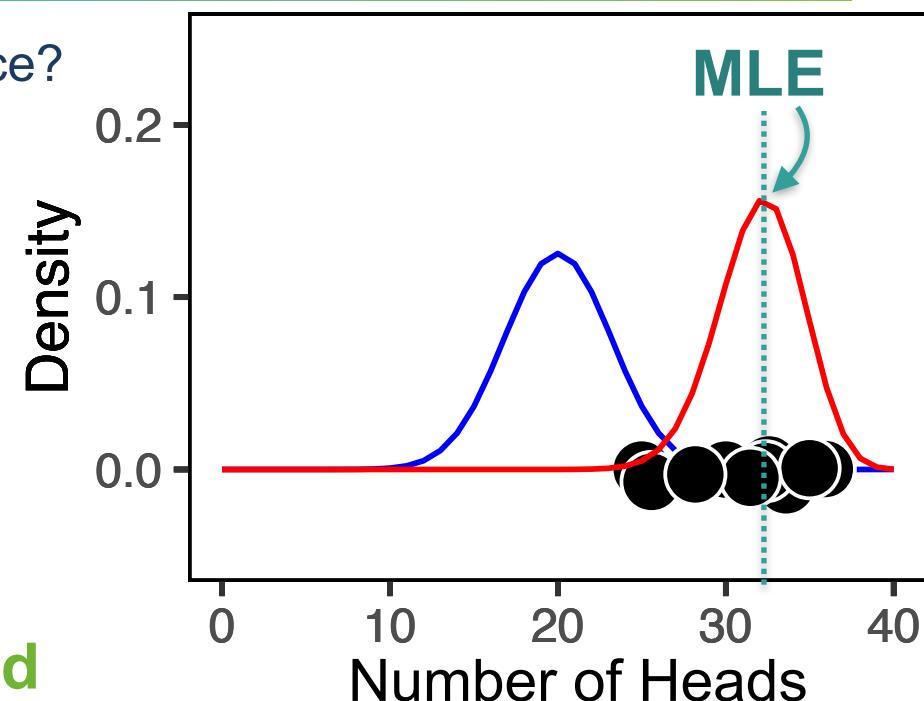
Likelihood

$$\log p(x_{1:T} | N_{1:T}, \mu) = \sum_{i=1}^T \log \text{Bin}(x_i | N_i, \mu)$$

Log-likelihood

$$\hat{\mu} = \frac{\sum_{i=1}^T x_i}{\sum_{i=1}^T N_i}$$

MLE



Bayesian Statistics: Prior distribution for model parameters

Likelihood for Binomial Model

$$p(x_{1:T} | N_{1:T}, \mu) = \prod_{i=1}^T \text{Bin}(x_i | N_i, \mu)$$

Likelihood

- MLE uses the evidence to estimate parameter $\hat{\mu}$ but our sample size is small and MLE may **overfit**
- **Zero count or sparse data problem:** If you have a bad record keeper who only tallies coin tosses from referees who never tosses a tail, then does that mean the concept of tails on a coin does not exist at all?
- Can we capture a more natural expectation of how a coin might behave? Also, what if we have some knowledge that the coin might be biased?

Prior Distribution for binomial parameter, μ

- The proportion of heads is between 0 and 1 ($\mu \in [0,1]$) and can be sampled from a distribution itself
- μ can be drawn from a Beta distribution, which is in the interval $[0,1]$, with **hyper-parameters** α and β

$$\mu \sim \text{Beta}(\alpha, \beta)$$

$$p(\mu) = \text{Beta}(\mu | \alpha, \beta)$$

Prior

	# of tosses (N)	# of heads (x)	Prop. of heads
Referee 1	40	25	0.63
Referee 2	42	35	0.83
Referee 3	39	27	0.69
Referee T	x_T	N_T	x_T/N_T

Bayesian statistics: Posterior for Beta-Binomial Model (1)

Binomial likelihood and Beta prior

- T different head counts $x = \{1, \dots, x_T\}$ for $N = \{1, \dots, N_T\}$ sets of tosses and a **prior** distribution on μ (prob. of heads)

$$p(x_{1:T} | N_{1:T}, \mu) = \prod_{i=1}^T \text{Bin}(x_i | N_i, \mu) \quad \text{Likelihood}$$
$$p(\mu) = \text{Beta}(\mu | \alpha, \beta) \quad \text{Prior}$$

- To estimate parameter μ in a Bayesian framework
 - We need the **posterior**, $p(\mu | x)$, but only have $p(x | \mu)$ and $p(\mu)$

- Recall Bayes' Theorem:

$$\frac{\text{Likelihood} \quad \text{Prior}}{\text{Posterior}} = \frac{p(X | Y)p(Y)}{\sum_{Y'} p(X | Y')p(Y')} \propto$$

- The **posterior** is our **belief state** by combining evidence from observations and our prior beliefs.

Bayesian statistics: Posterior for Beta-Binomial Model (2)

Beta-Binomial Model: Posterior distribution

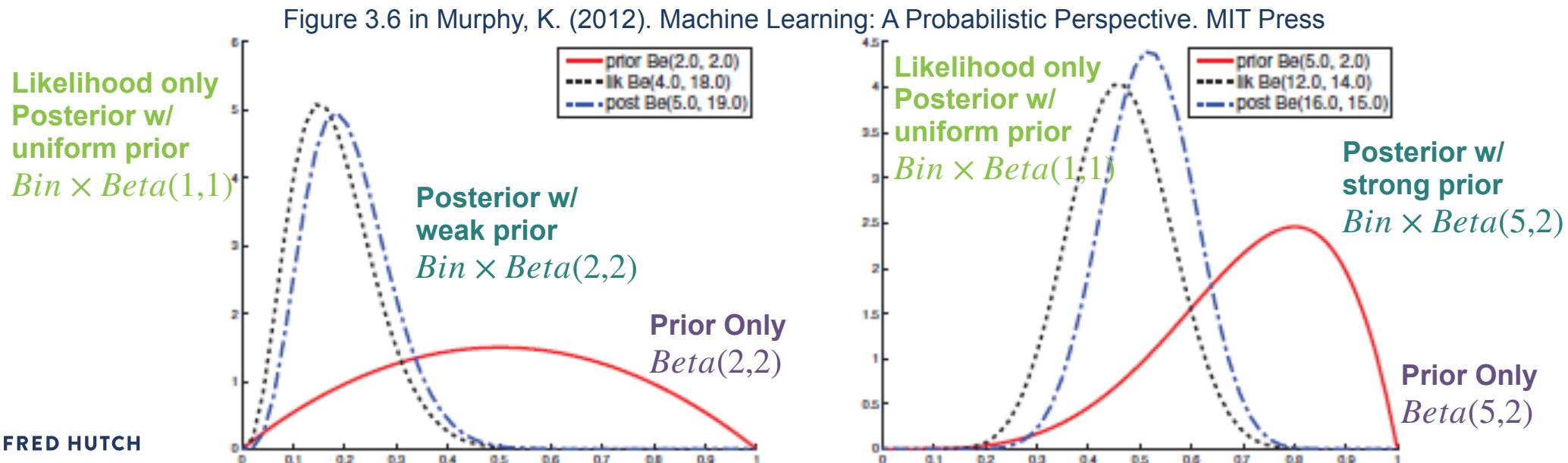
- To estimate the model parameter μ in a Bayesian framework, we compute the *posterior*, $p(\mu | x)$

$$p(\mu | x_i) \propto Bin(x_i | N_i, \mu) \times Beta(\mu | \alpha, \beta)$$

- Beta is a *conjugate prior* for the binomial; *the product of binomial and Beta has the form of a Beta*

$$p(\mu | x_i) \propto Bin(x_i | N_i, \mu) \times Beta(\mu | \alpha, \beta) = Beta(\mu | x_i + \alpha, N_i - x_i + \beta)$$

Likelihood Prior Posterior



Bayesian statistics: MAP estimate

Beta-Binomial Model: Posterior distribution

$$p(\mu | x_i) \propto Bin(x_i | N_i, \mu) \times Beta(\mu | \alpha, \beta) = Beta(\mu | x_i + \alpha, N_i - x_i + \beta)$$

Posterior

- Then, what is the probability of heads, μ , of this coin given the **evidence** and the **prior**?

Maximum a posteriori (MAP) estimate

- From the posterior, we can estimate the parameter using the **maximum a posteriori (MAP)**, $\hat{\mu}_{MAP}$
- MAP refers to the mode of the posterior distribution and the mode of a Beta is $\frac{\alpha - 1}{\alpha + \beta - 2}$
- Since the posterior has the form of a Beta distribution, then the MAP is $\frac{\alpha' - 1}{\alpha' + \beta' - 2}$
- .

$$\alpha' = x_i + \alpha$$

$$\beta' = (N_i - x_i) + \beta$$

$$\hat{\mu}_{MAP} = \frac{x_i + \alpha - 1}{N_i + \alpha + \beta - 2}$$

MAP

Section 3.3 in Murphy (2012).
Machine Learning: A Probabilistic
Perspective. MIT Press

Mapping the Referee Example to Mutation Calling

Referee Coin Toss Example

Data

Referees $1, \dots, T$

For each Referee i

- Coin Tosses: N_i
- Count of heads: x_i
- Count of tails: $N_i - x_i$

Parameters

Probability to draw coins: π_{fair} , π_{heads} , π_{tails}

Probability of heads for 3 types of coins

$$\mu_{\text{fair}}, \mu_{\text{heads}}, \mu_{\text{tails}}$$

Responsibilities

Probability that Referee i used coin k : $\gamma(Z_i = k)$

Mutation Calling from Sequencing Data

Data

Genomic loci $1, \dots, T$

For each locus i

- Depth (total reads): N_i
- Count of reference reads: x_i
- Count of variant reads: $N_i - x_i$

Parameters

Probability of genotypes: π_{AA} , π_{AB} , π_{BB}

Probability of reference base for 3 genotypes:

$$\mu_{AA}, \mu_{AB}, \mu_{BB}$$

Responsibilities

Probability that locus i has genotype k : $\gamma(Z_i = k)$

Mixture Models: Online Tutorial and Resource

fiveMinuteStats (<https://stephens999.github.io/fiveMinuteStats/>)

by **Dr. Matthew Stephens**, Professor in Statistics & Human Genetics at University of Chicago

1. Introduction to mixture models with probabilistic derivations and R code

- Examples with Bernoulli and Gaussian models
- https://stephens999.github.io/fiveMinuteStats/intro_to_mixture_models.html

2. Introduction to EM with Gaussian Mixture Model example and R code

- https://stephens999.github.io/fiveMinuteStats/intro_to_em.html

Homework #5: Single-nucleotide Genotype Caller

Implement a standard binomial mixture model described in Lecture 2.

- Learn the parameters and infer the genotypes
- Annotate the mutation status for a set of genomic loci.
- Expected outputs for each question will be provided so that you can check your code.
- RStudio Markdown and Python Jupyter Notebook templates provided.

Due: May 5th, 2022

Virtual Office Hours with Anna-Lisa Doebley (adoebley@uw.edu)

- May 3 @ 2pm, May 6 @ 2pm (Zoom Meeting ID: 446 356 7725, Passcode: GS541)