Cancer hallmarks, “omic” data, and data resources

Anthony Gitter

Cancer Bioinformatics (BMI 826/CS 838)

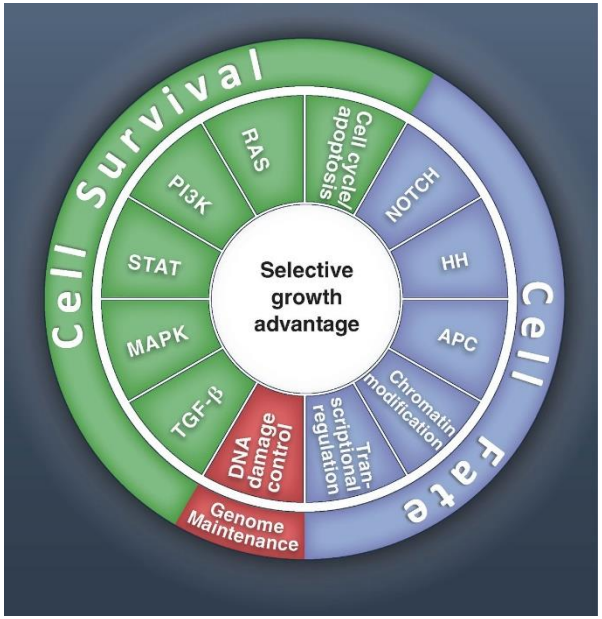
January 22, 2015

What computational analysis contributes to cancer research

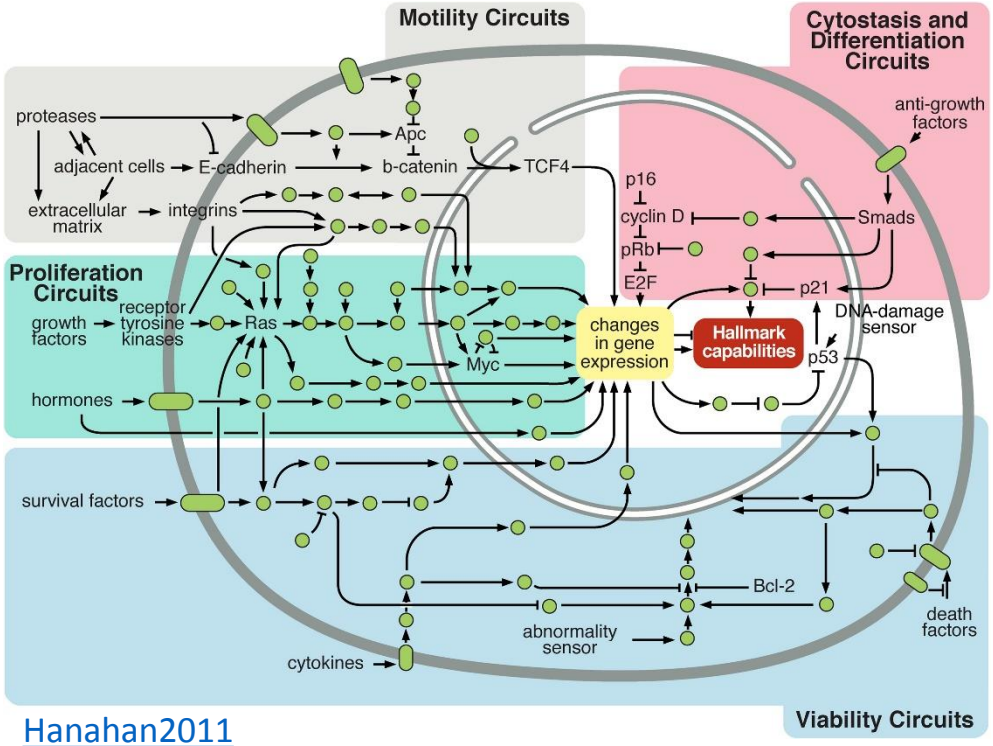
1. Predicting driver alterations
2. Defining properties of cancer (sub)types
3. Predicting prognosis and therapy
4. Integrating complementary data
5. Detecting affected pathways and processes
6. Explaining tumor heterogeneity
7. Detecting mutations and variants
8. Organizing, visualizing, and distributing data

Convergence of driver events

- Amid the complexity and heterogeneity, there is some order
- Finite number of major pathways that are affected by drivers



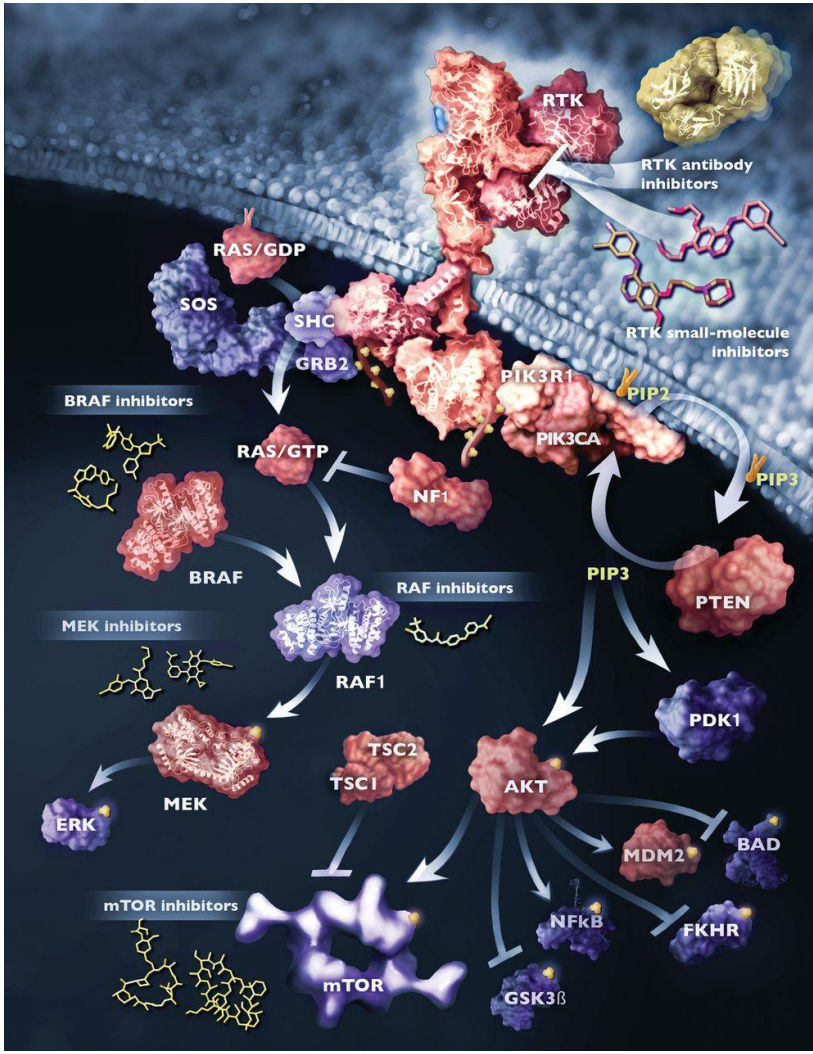
Vogelstein2013



Hanahan2011

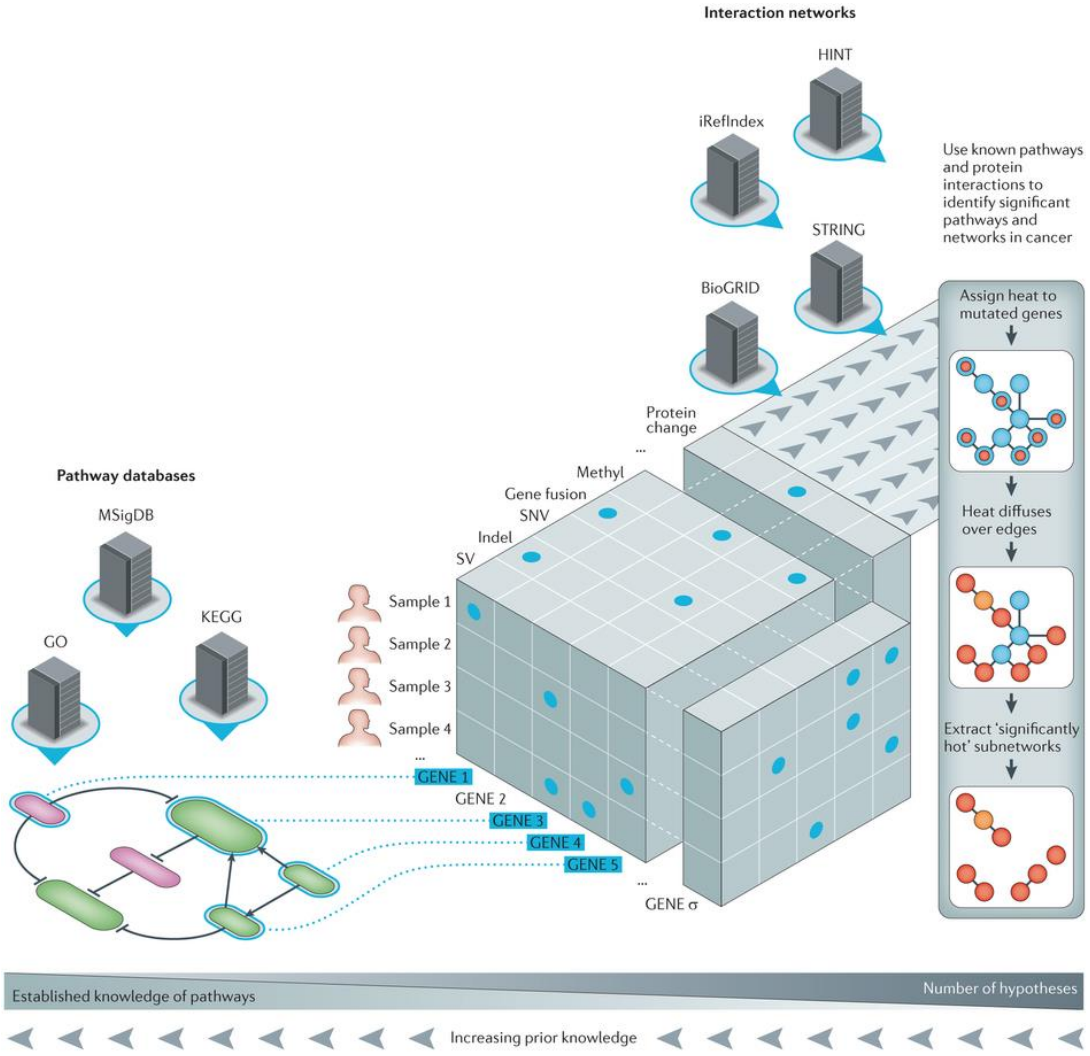
Similar pathway effects

- Tumor 1: EGFR receptor mutation makes it hypersensitive
- Tumor 2: KRAS hyperactive
- Tumor 3: NF1 inactivated and no longer modulates KRAS
- Tumor 4: BRAF over responsive to KRAS signals



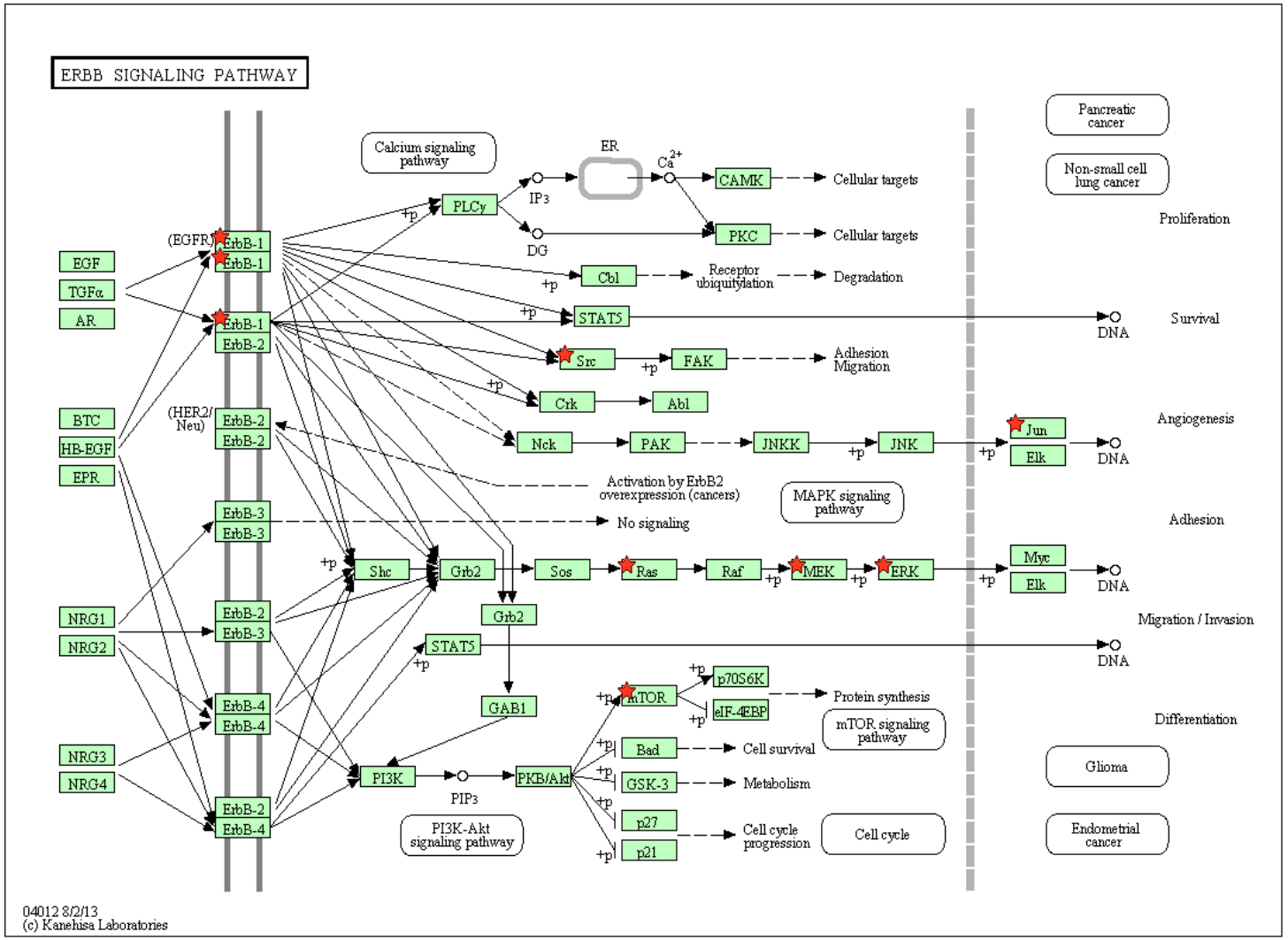
Vogelstein2013

Detecting affected pathways



Ding2014

Pathway enrichment

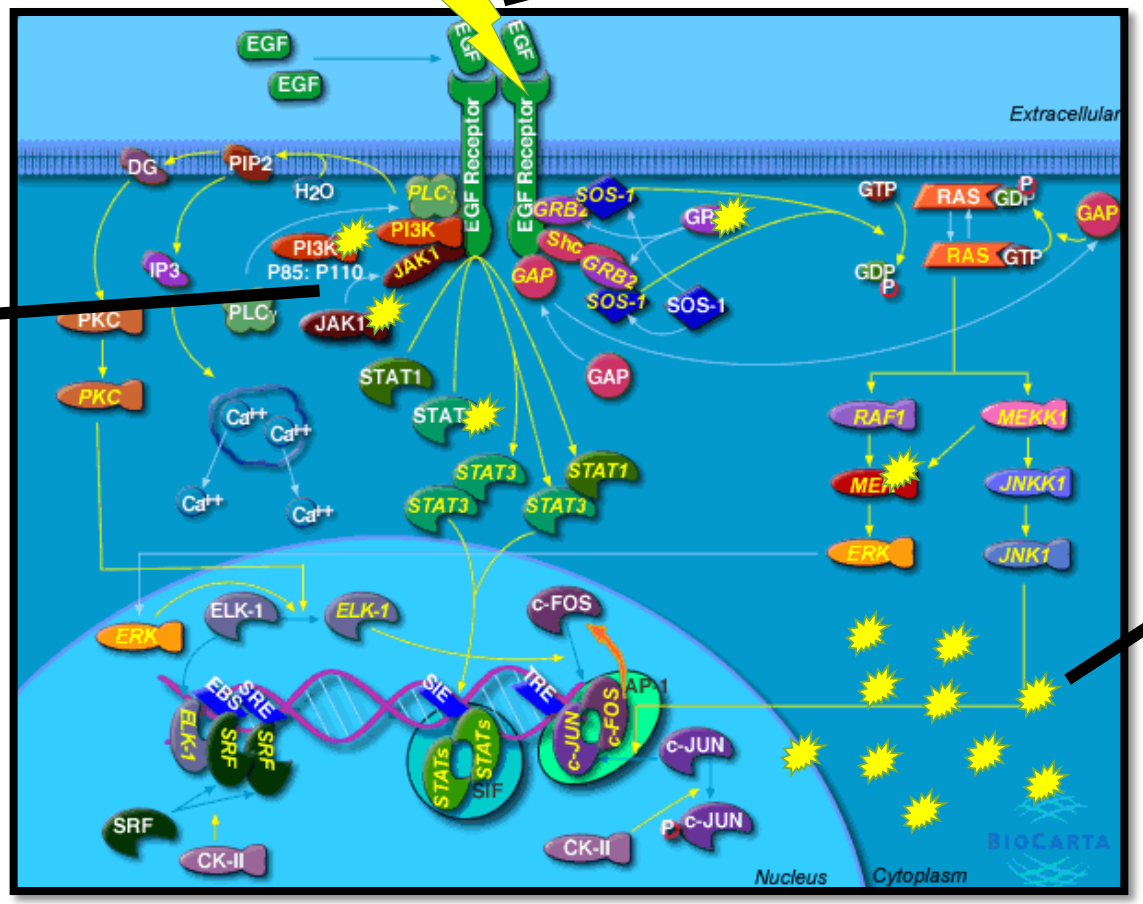


DAVID

Pathway discovery

Stimulate receptor

31% of pathway is activated

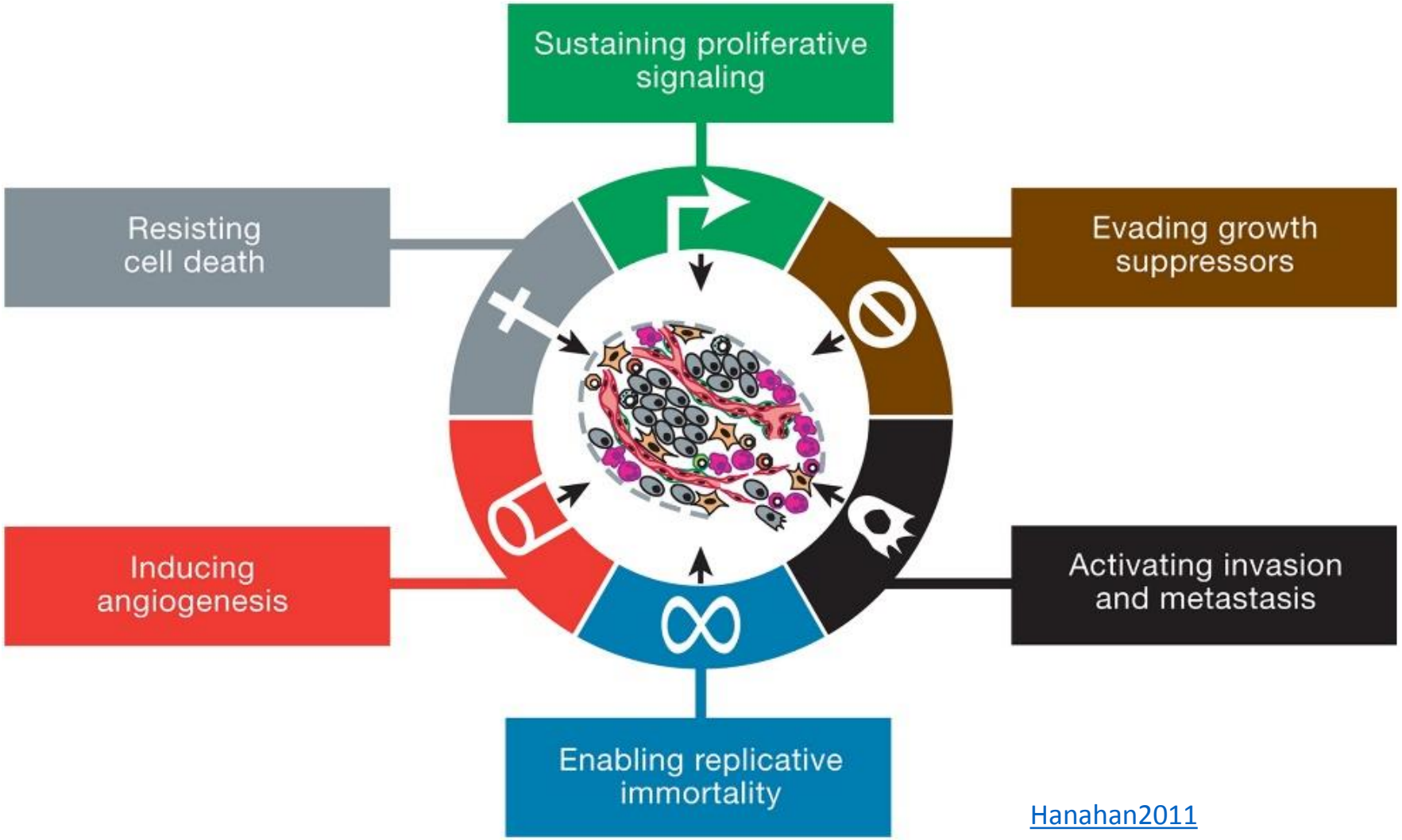


98% of activity is not covered

[BioCarta EGF Signaling Pathway]

Phosphorylation data from [Alejandro Wolf-Yadlin]

Hallmarks of cancer



Hanahan2011

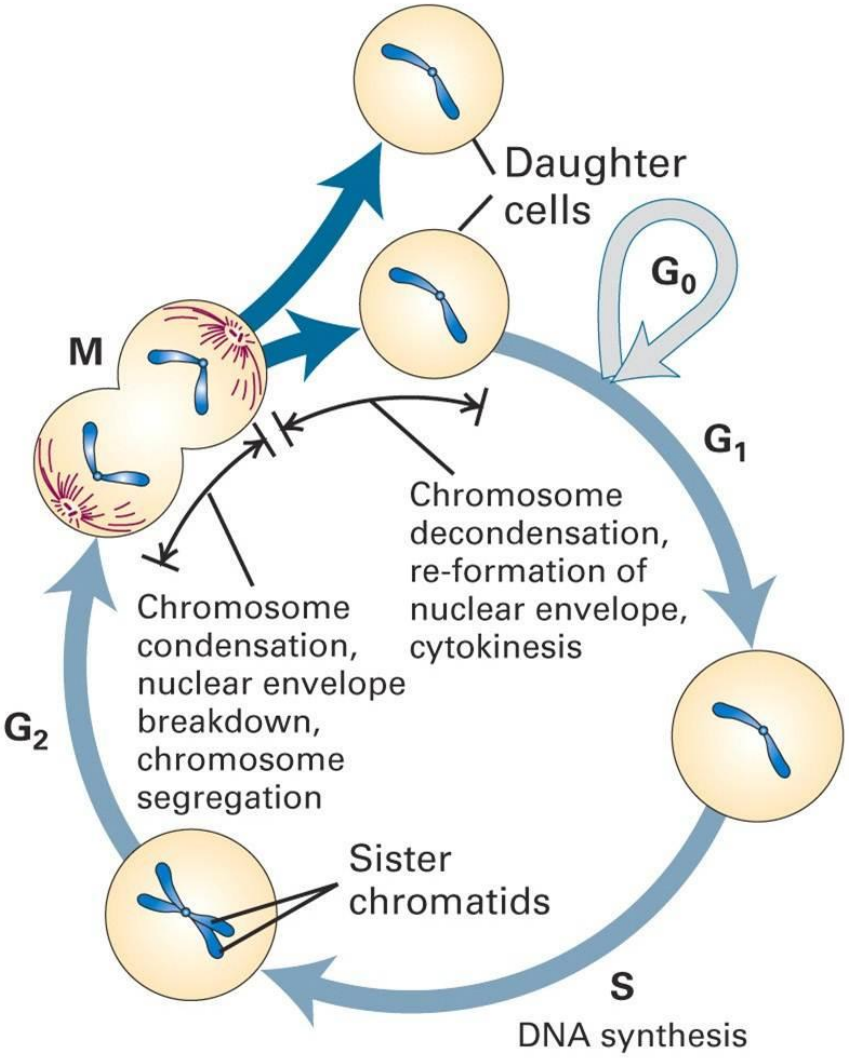
Sustaining proliferative signaling

- Cells receive signals from the local environment telling them to grow (proliferate)
- Specialized receptors detect these signals
- Feedback in pathways carefully controls the response to these signals

Evading growth suppressors

- Override tumor suppressor genes
- Some proteins control the cell's decision to grow or switch to an alternate track
  - **Apoptosis**: programmed cell death
  - **Senescence**: halt the cell cycle
- External or internal signals can affect these decisions

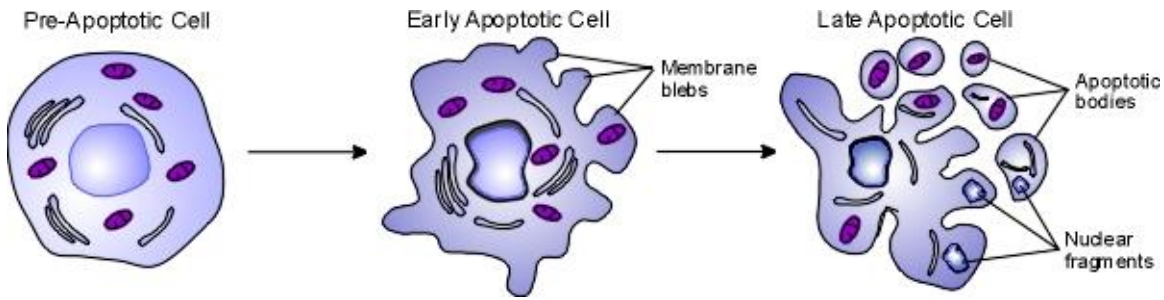
Cell cycle



Biology of Cancer

Resisting cell death

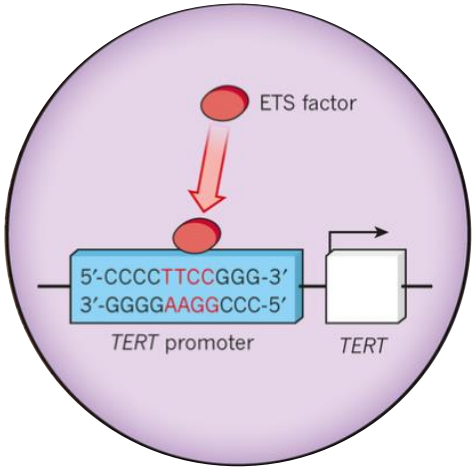
- One self-defense mechanism against cancer
- Apoptosis triggers include:
  - DNA damage sensors
  - Limited survival cues
  - Overactive signaling proteins
- Necrosis causes cells to explode
  - Destroys a (pre)cancerous cell
  - Releases chemicals that can promote growth in other cells



O'Day

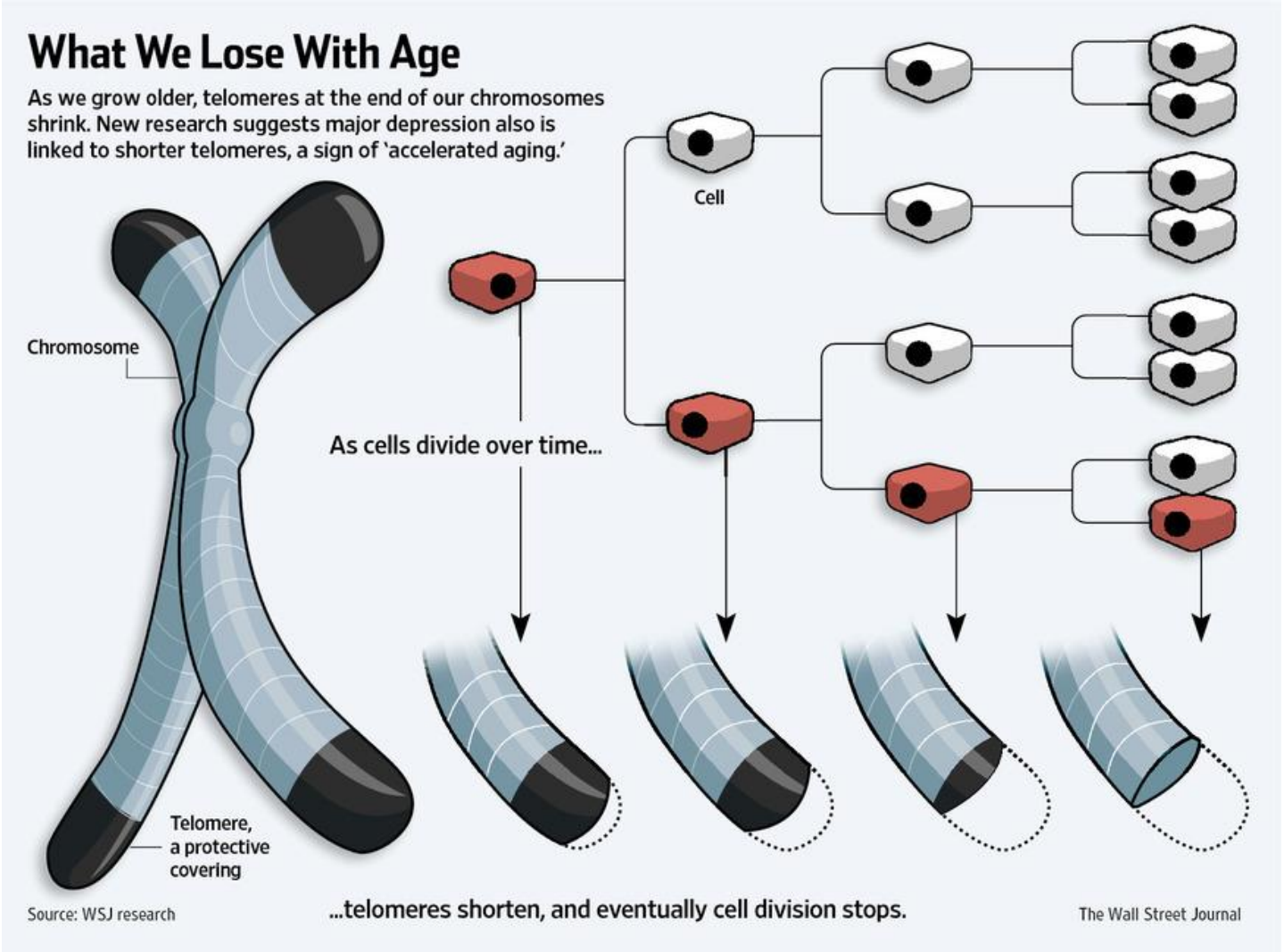
Enabling replicative immortality

- Cells typically have a limited number of divisions
- **Immortalization**: unlimited replicative potential
- Telomeres protect the ends of DNA
  - Shorten over time
  - Encode the number of cell divisions remaining
  - Can be artificially upregulated in cancer



Patton2013

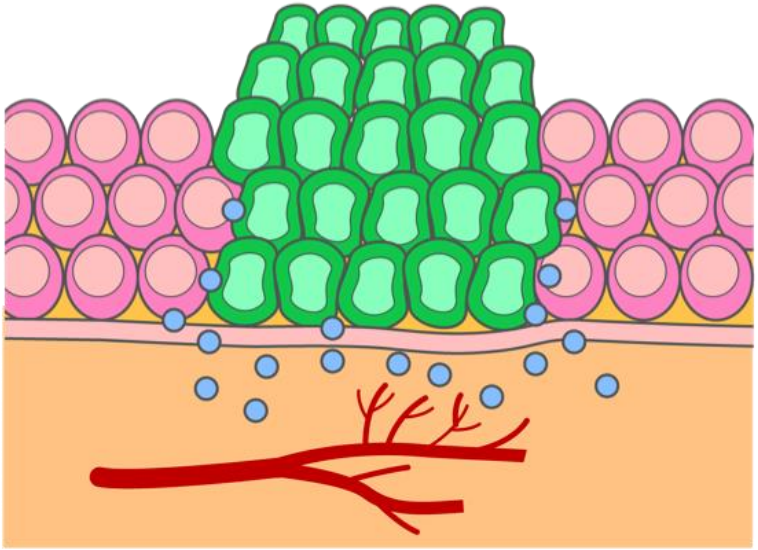
Telomere shortening

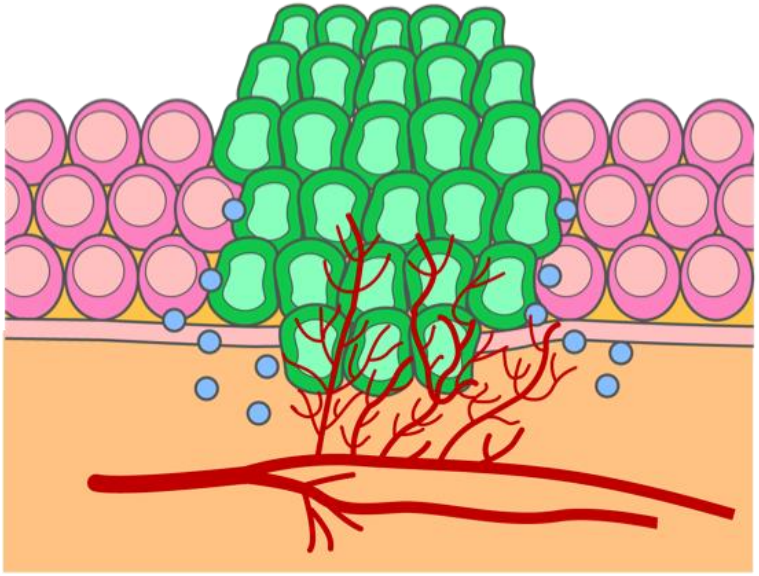


Wall Street Journal

Inducing angiogenesis

- Tumors must receive nutrients like other cells
- Certain proteins promote growth of blood vessels



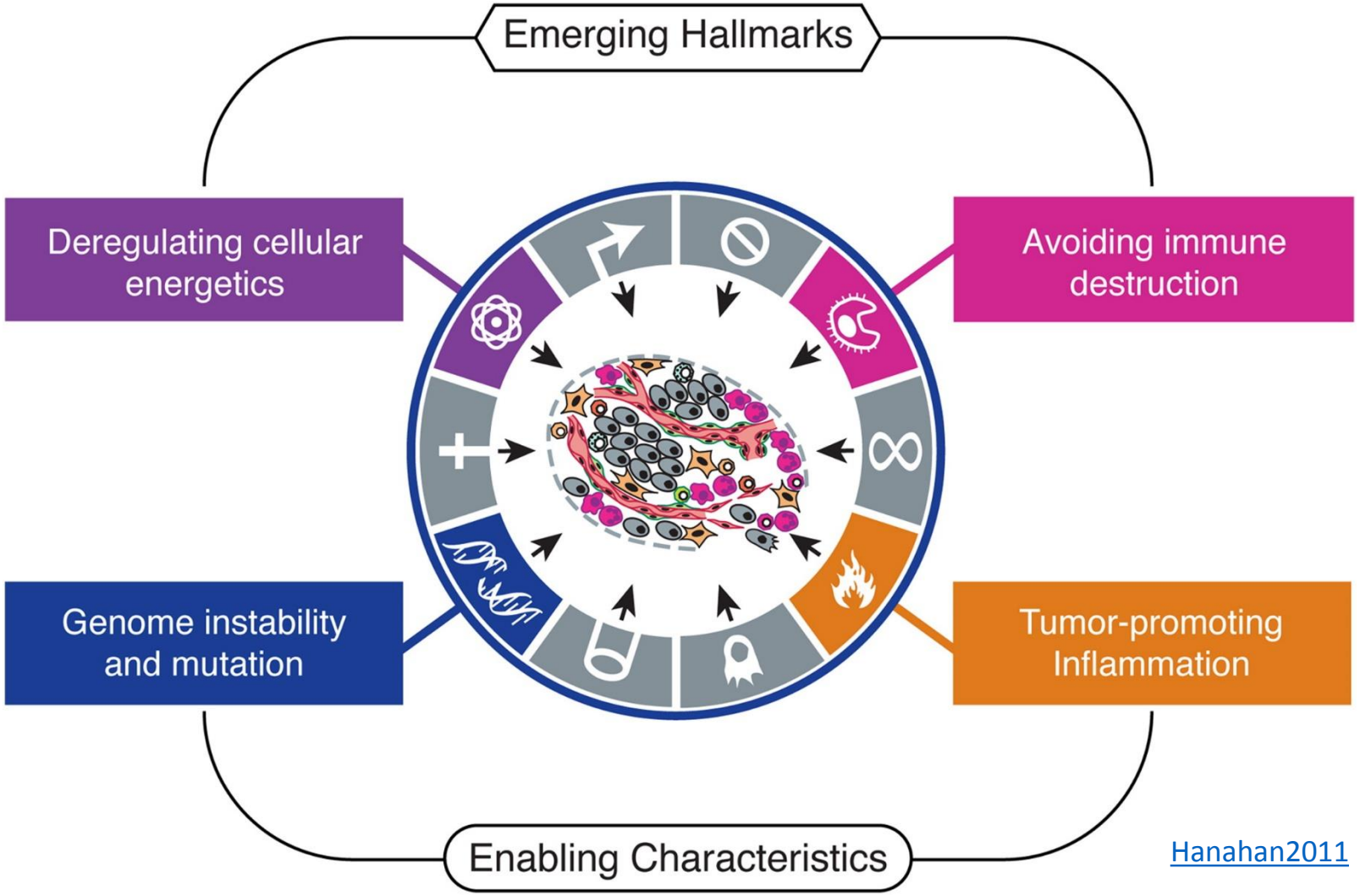


LKT Laboratories

Activating invasion and metastasis

- Cancer progresses through the aforementioned stages
- Epithelial-mesenchymal transition (EMT)

Emerging hallmarks



Hanahan2011

Genome instability and mutation

- Cancer cells mutate more frequently
- Increased sensitivity to mutagens
- Loss of telomeres increases copy number alterations

Model systems in oncology

- **Cell lines**: Cells that reproduce in a lab indefinitely (e.g. Hela cells)
- **Genetically engineered mice**: Manipulate mice to make them predisposed to cancer
- **Xenograft**: Implant human tumor cells into mice

“Omic” data types

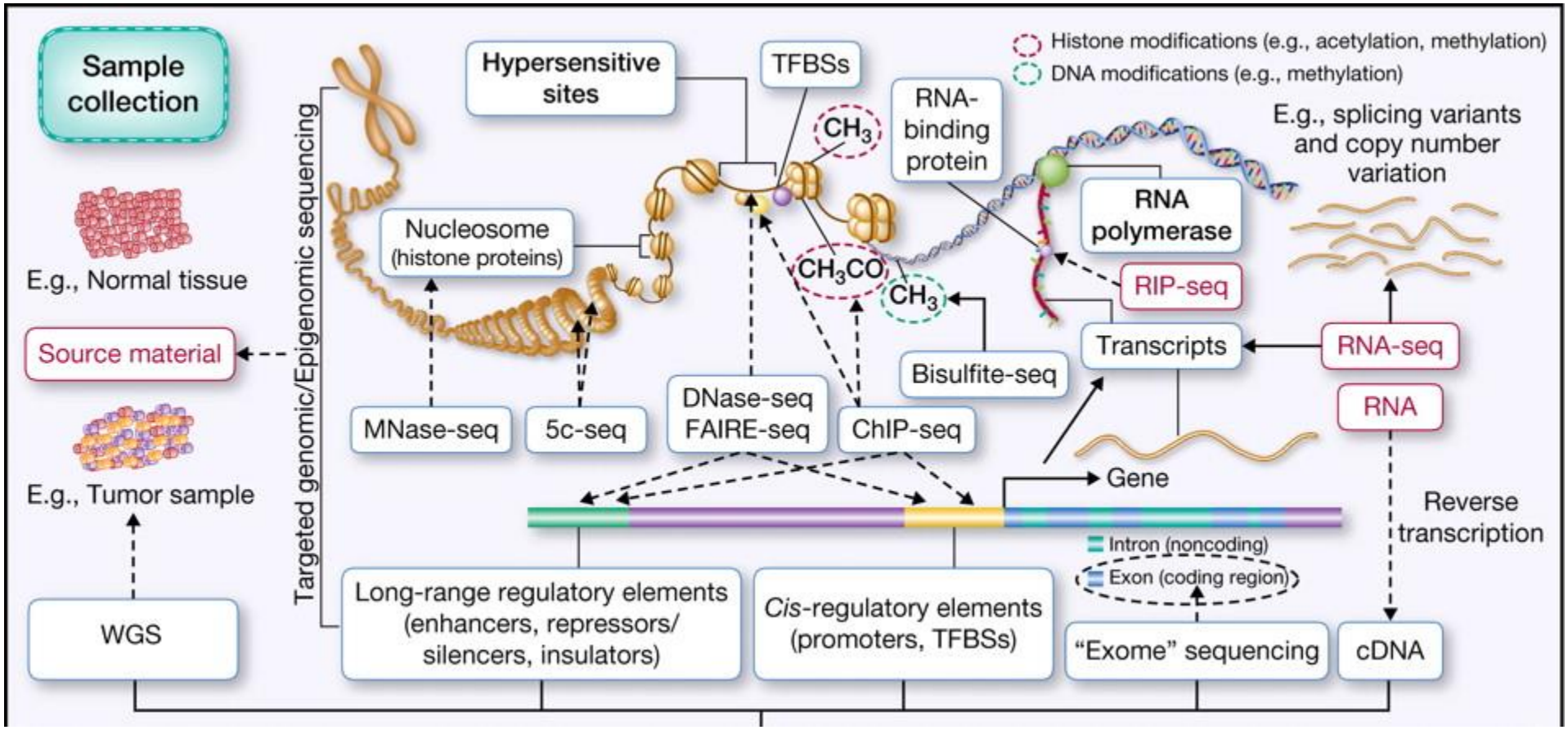
- DNA (genome)
  - Mutations
  - Copy number variation
  - Other structural variation
- RNA expression (transcriptome)
  - Gene expression (mRNA)
  - Micro RNA expression (miRNA)
- Protein (proteome)
  - Protein abundance
  - Protein state (e.g. phosphorylation)
- Protein DNA binding
- DNA state and accessibility (epigenome)
  - DNA methylation (methylome)
  - Histone modification / chromatin marks
  - DNase I hypersensitivity

“Next-generation” sequencing (NGS)

- Revolutionized high-throughput data collection
- *-seq strategy
  - Decide what you want to measure in cells
  - Figure out how to select or synthesize the right DNA
  - Dump it into a DNA sequencer
- ~100 different *-seq applications

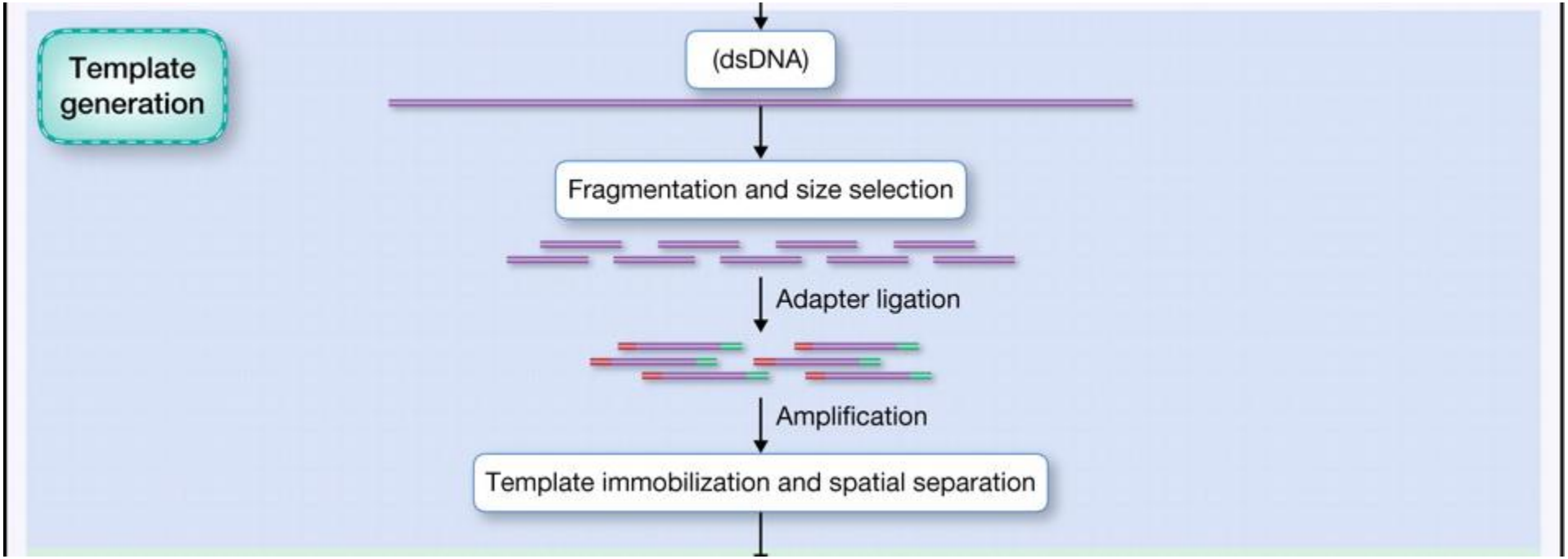
NODAI

*-seq examples



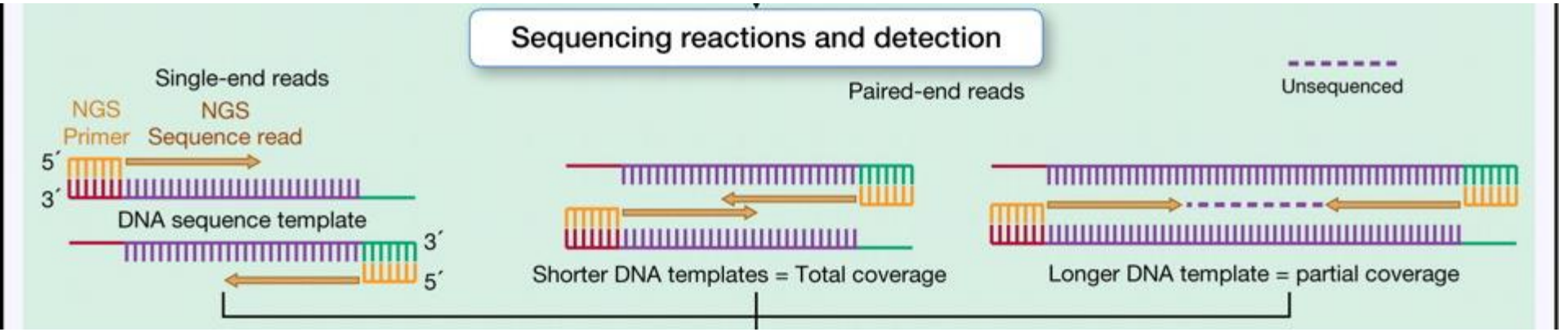
Rizzo2012

Generating DNA templates



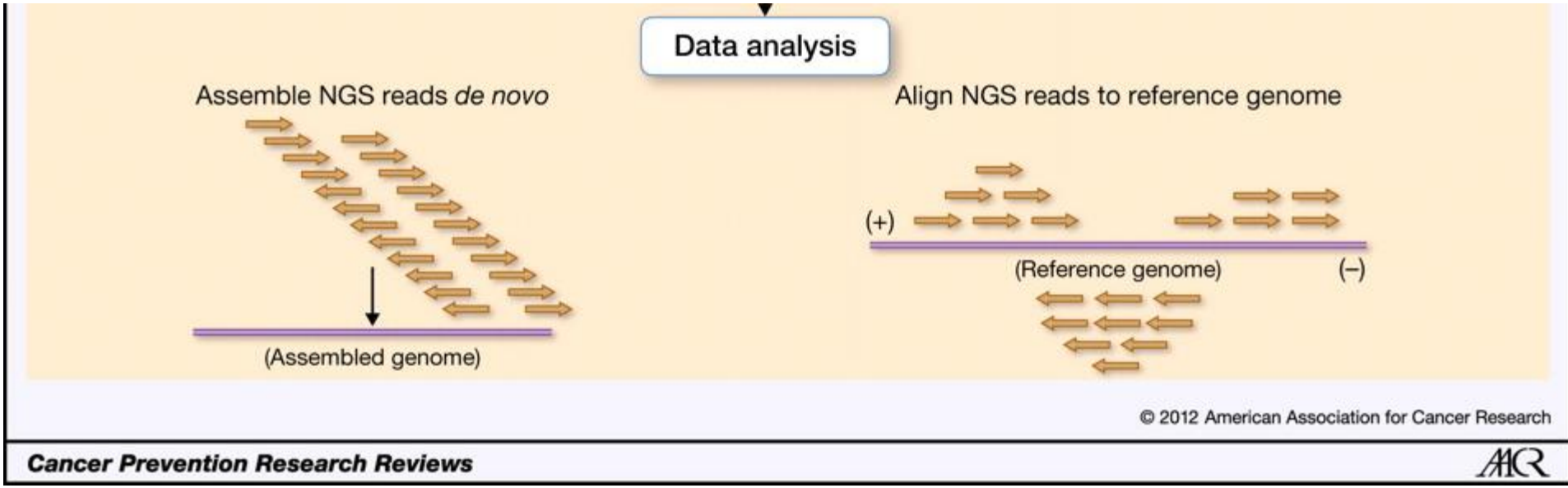
Rizzo2012

Generating reads



Rizzo2012

Assembly and alignment

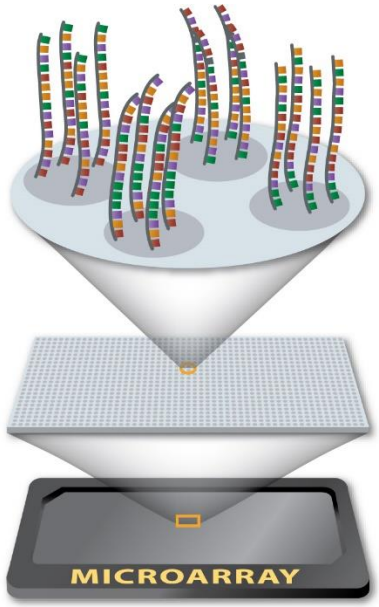


Rizzo2012

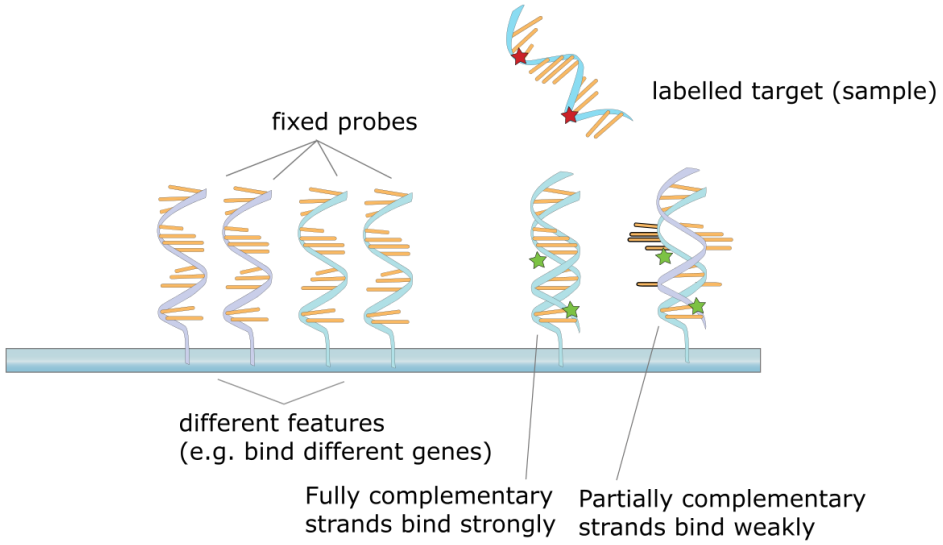
Microarrays

- High-throughput measurement of gene expression, protein DNA binding, etc.
- Mostly replaced by *-seq
- Fixed probes as opposed to DNA reads

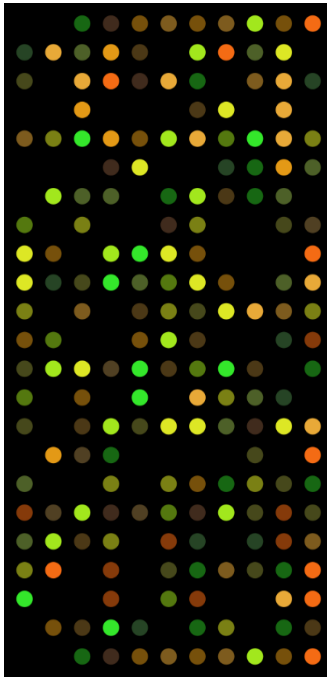
Microarray quantification



University of Utah



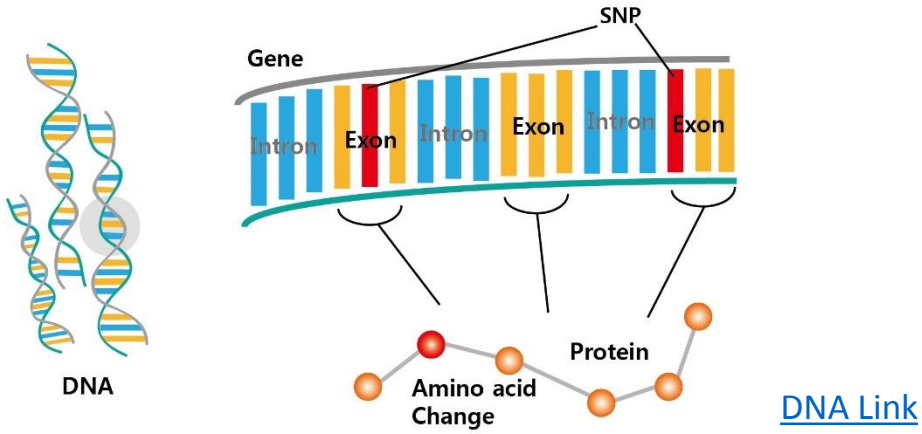
Wikipedia



Wikimedia

DNA mutations

- Whole-exome most prevalent in cancer
  - Only covers exons that form genes, less expensive

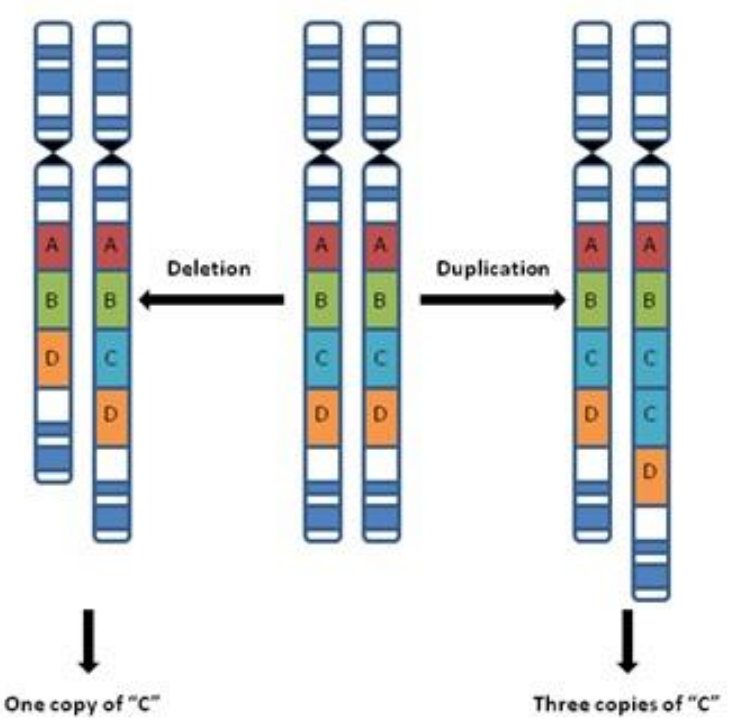


DNA Link

- Whole-genome becoming more widespread as sequencing costs continue to decrease

Copy number variation

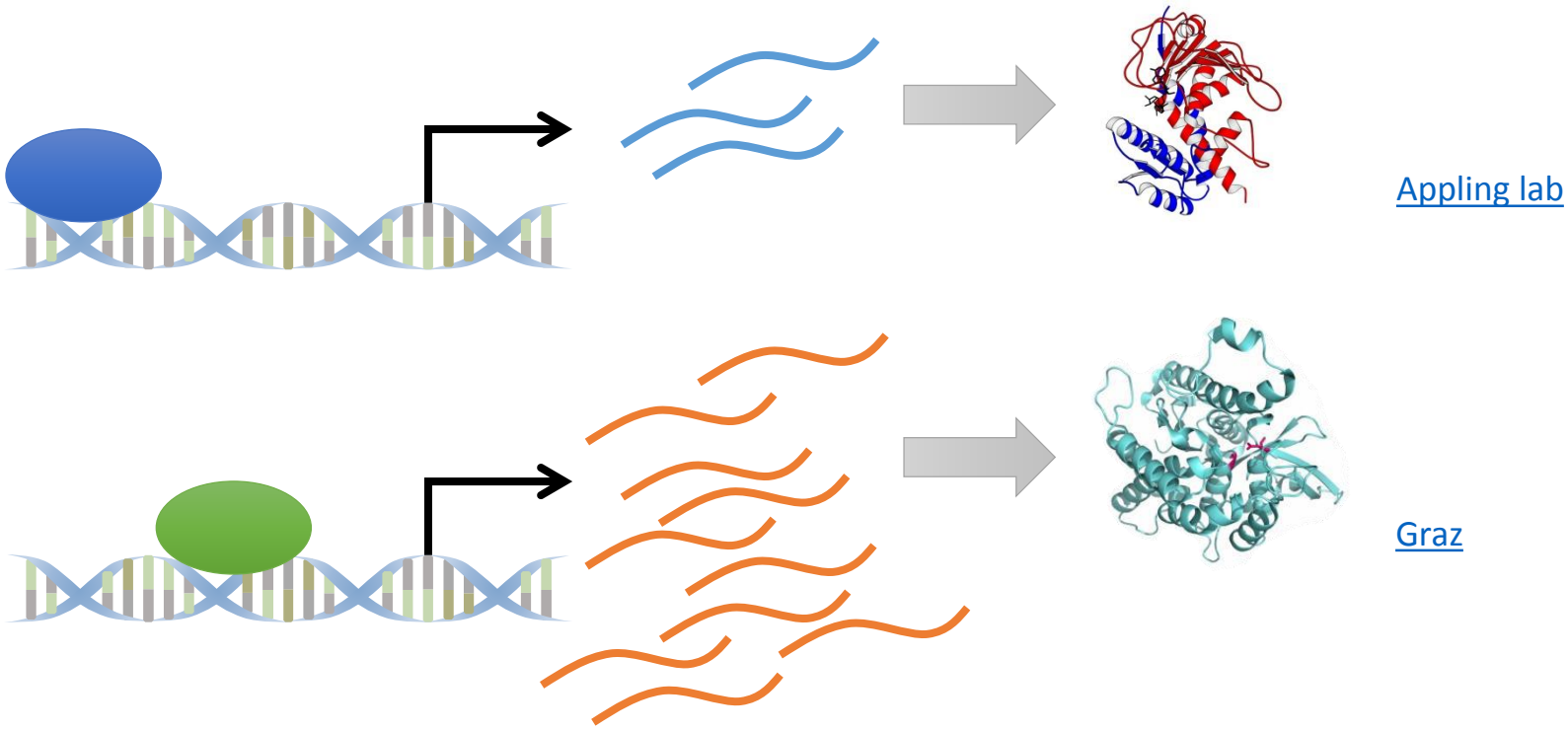
- Often represented as relative to normal 2 copies
- Ranges from a few bases to whole chromosomes
- Quantitative, not discrete, representation



MindSpec

Gene expression

- Transcript (messenger RNA) abundance

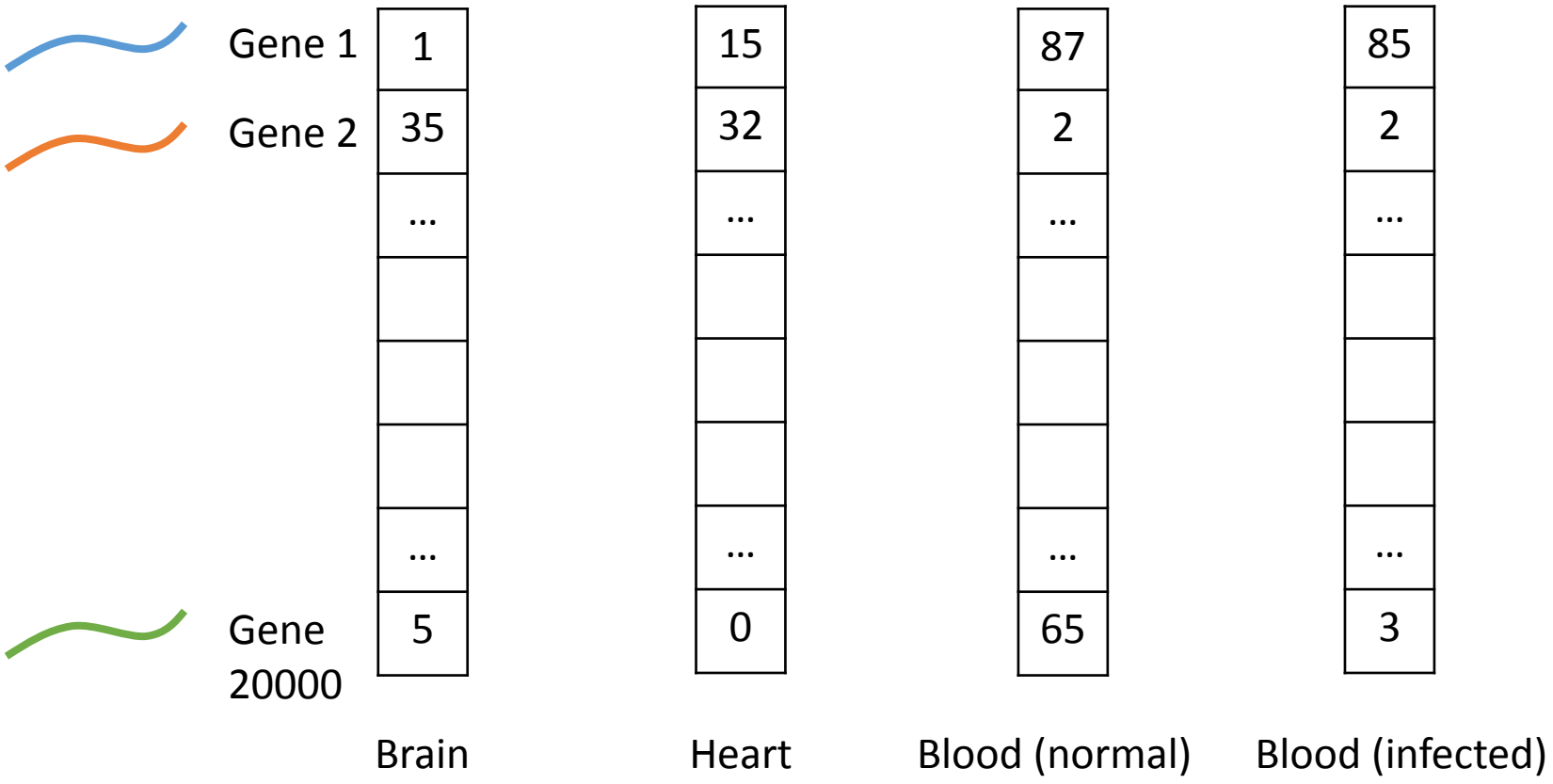


Appling lab

Graz

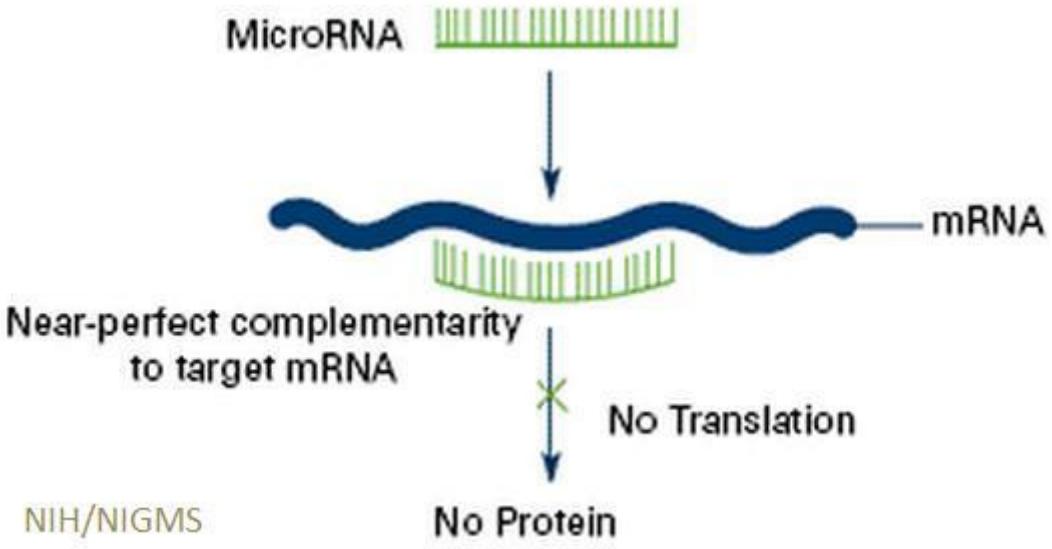
Genome-wide gene expression

- Quantitative state of the cell



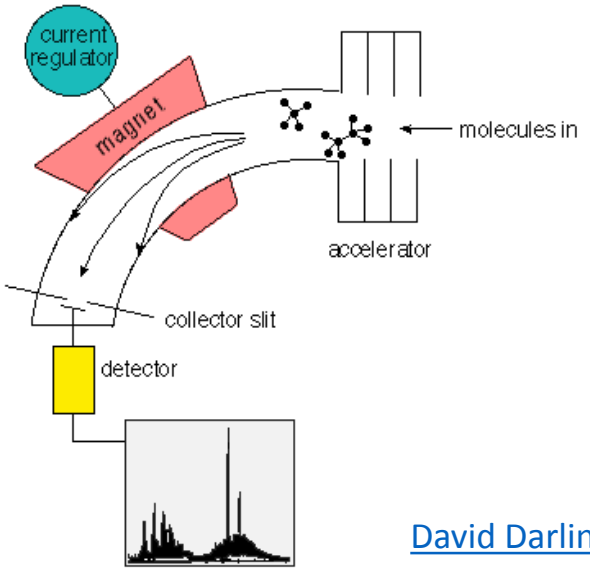
miRNA expression

- microRNA (miRNA)
  - ~22 nucleotides
  - Does not code for a protein
  - Regulates gene expression levels by binding mRNA



Protein abundance

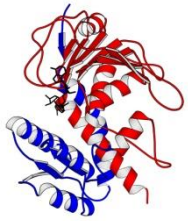
- Protein abundance is analogous to gene expression
- Not perfectly correlated with gene expression
- Harder to measure
- Mass spectrometry is almost proteome-wide
  - Vaporize molecules
  - Determine what was vaporized based on mass/charge

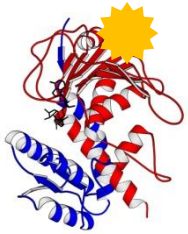


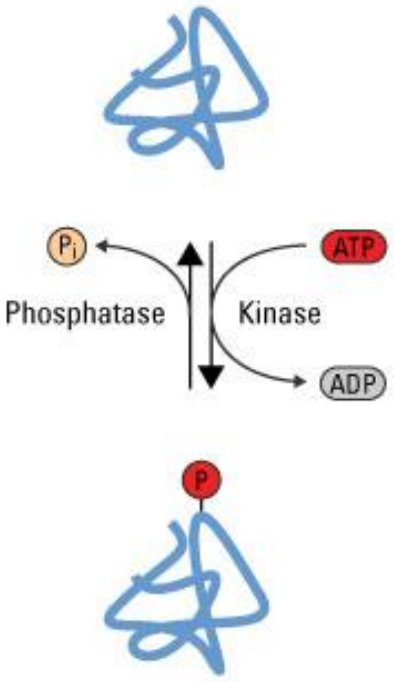
David Darling

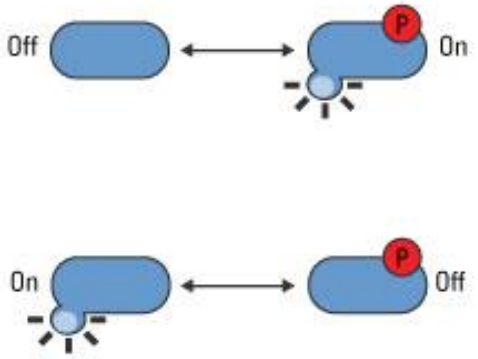
Protein state

- Chemical groups added to mature protein
- Phosphorylation is the most-studied
- Analogous to Boolean state





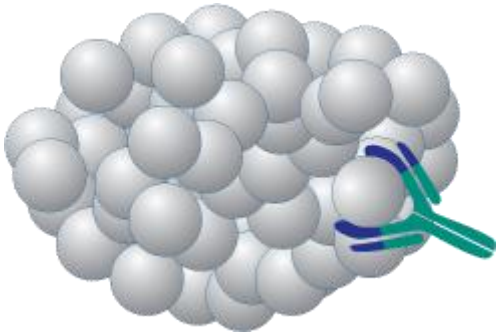




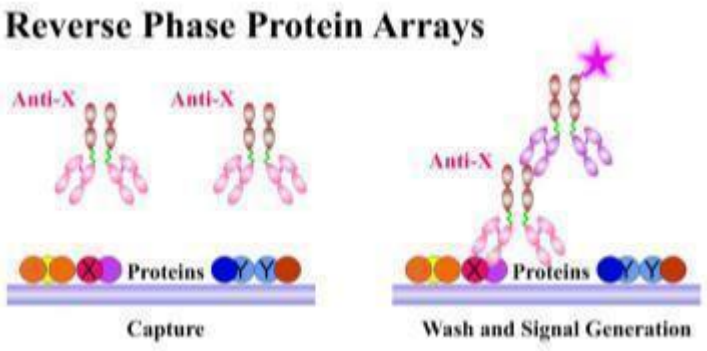
Pierce

Protein arrays

- Currently more common in cancer datasets
- Measure a limited number of specific proteins using antibodies
- Protein abundance or state



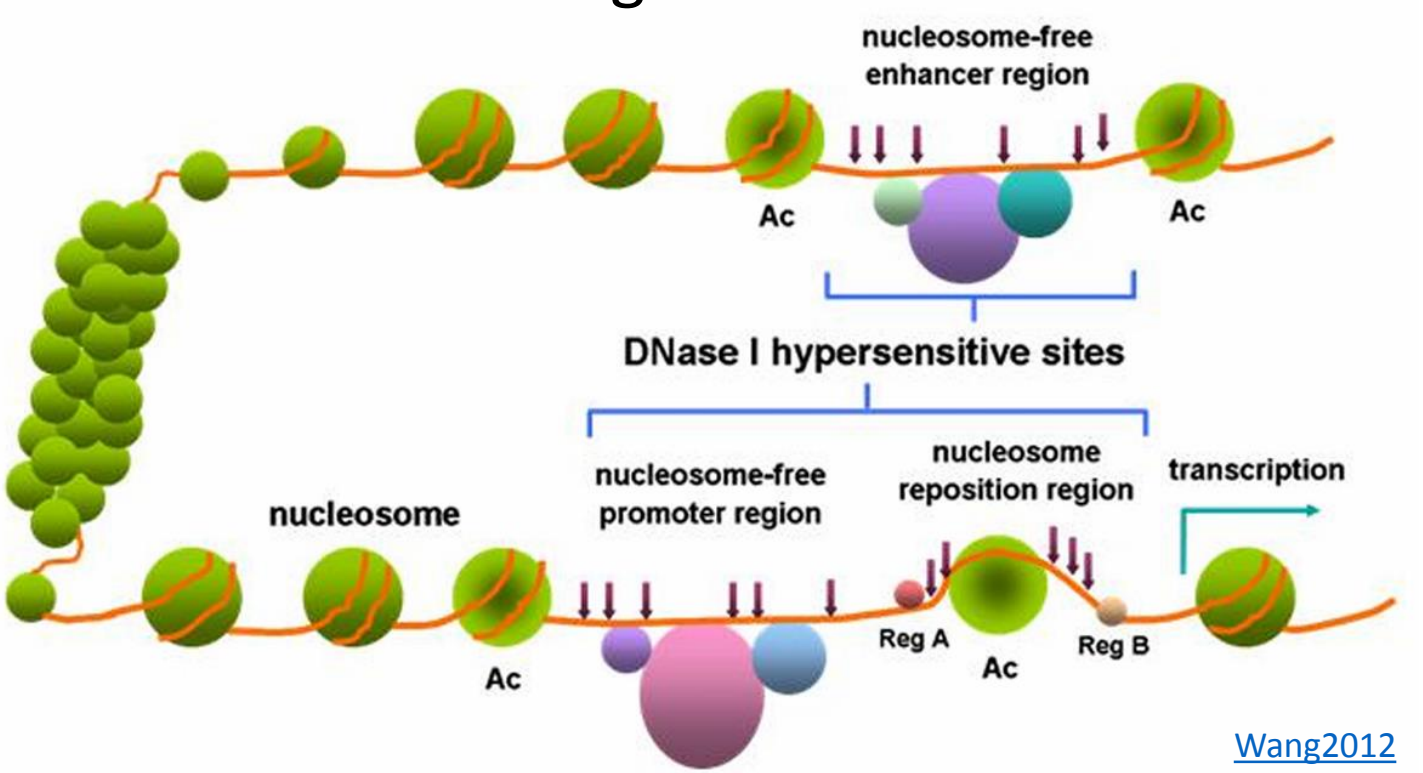
R&D



MD Anderson

Transcriptional regulation

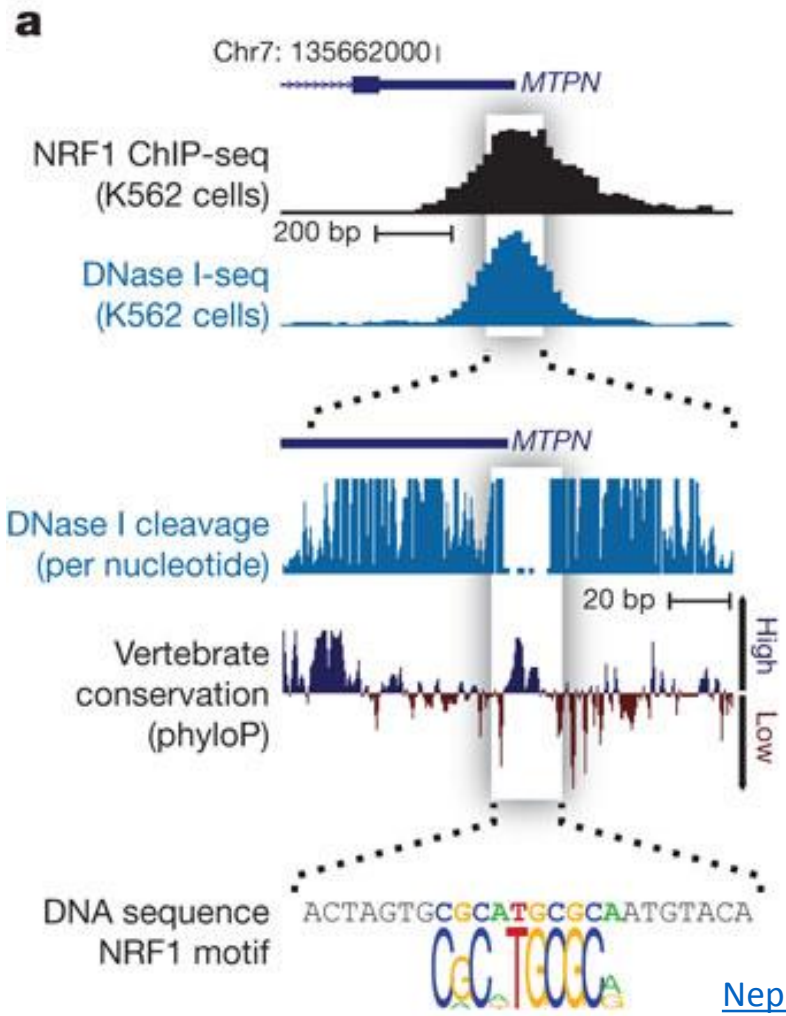
- ChIP-seq directly measures transcription factor (TF) binding but requires a matching antibody
- Various indirect strategies



Wang2012

Predicting regulator binding sites

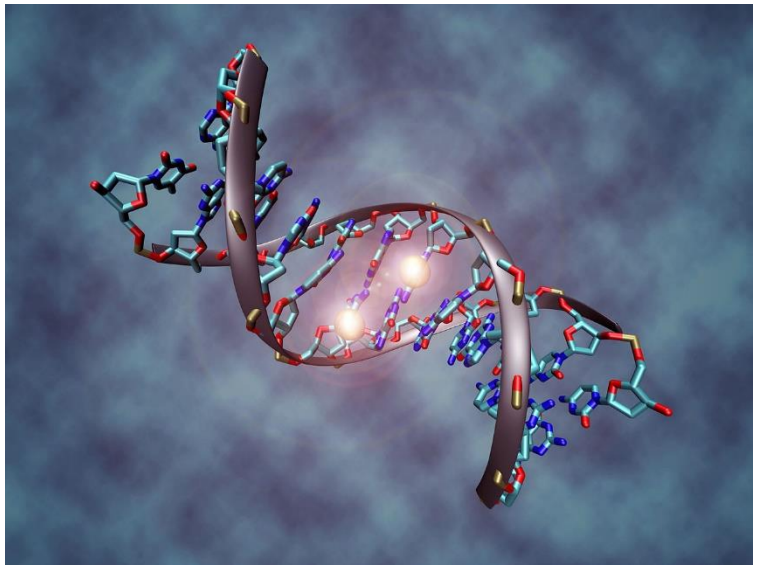
- Motifs are signatures of the DNA sequence recognized by a TF
- TFs block DNA cleavage
- Combining accessible DNA and DNA motifs produces binding predictions for hundreds of TFs



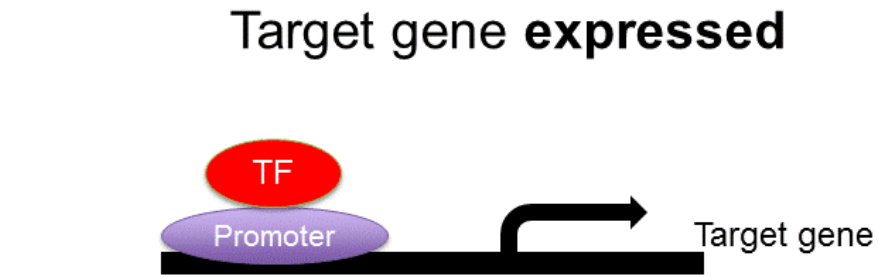
Neph2012

DNA methylation

- Methylation is a DNA modification (state change)
- Hyper-methylation suppresses transcription
- Methylation almost always at C



Wikimedia

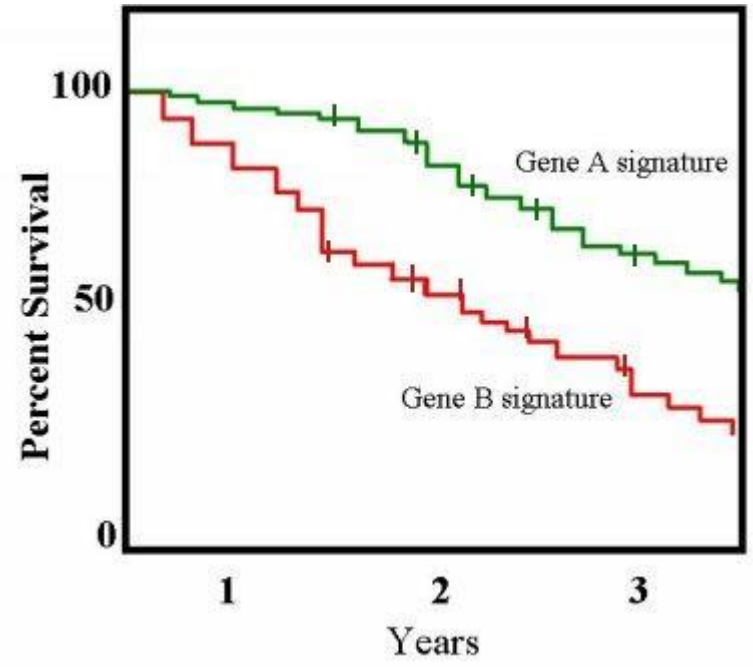


Learn NC

(Fry, 2011)

Clinical data

- Age, sex, cancer stage, survival
- Kaplan–Meier plot



Wikipedia

Large cancer datasets

- Tumors
  - The Cancer Genome Atlas (TCGA)
  - Broad Firehose and FireBrowse access to TCGA data
  - International Cancer Genome Consortium (ICGC)
- Cell lines
  - Cancer Cell Line Encyclopedia (CCLE)
  - Catalogue of Somatic Mutations in Cancer (COSMIC)
- Cancer gene lists
  - COSMIC Gene Census
  - Vogelstein2013 drivers

Interactive tools for cancer data

- cBioPortal
- TumorPortal
- Cancer Regulome
- Cancer Genomics Browser
- StratomeX

Gene and protein information

- TP53 example
  - GeneCards
  - UniProt
  - Entrez Gene

Pathway and function enrichment

- Database for Annotation, Visualization and Integrated Discovery (DAVID)
- Molecular Signatures Database (MSigDB)

Gene expression data

- Gene Expression Omnibus (GEO)
- ArrayExpress

Protein interaction networks

- iRefIndex and iRefWeb
- Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)
- High-quality INTeractomes (HINT)

Transcriptional regulation

- Encyclopedia of DNA Elements (ENCODE)
- DNA binding motifs
  - TRANSFAC
  - JASPAR
  - UniPROBE

miRNA binding

- miRBase
- TargetScan