Cancer Bioinformatics

A Quickstart for Computational People

Johan Michalove*

February 2026 — Living document

cancer.silver.is

Abstract

This guide is for people who know how to code, build systems, and read papers, but who do not have a background in biology or medicine. It provides the minimum viable knowledge to begin contributing to cancer bioinformatics: what cancer is, how it is classified, how it is treated, where computation enters, what the key datasets are, what the open problems are, and where to start. It is opinionated. It assumes you learn by building. It will be updated as the field moves.

*Department of Information Science, Cornell University. jam844@cornell.edu

Contents

1 What Cancer Is **3**

1.1 The Hallmarks 3

1.2 The Central Dogma 3

2 How Cancer Is Classified **4**

2.1 Anatomical 4

2.2 Histological 4

2.3 Molecular 5

2.4 Staging 5

3 Key Mutations **5**

4 How Cancer Is Treated **6**

5 Where Computation Enters **7**

5.1 Variant Calling and Interpretation 7

5.2 Gene Expression and Subtyping 8

5.3 Digital Pathology 8

5.4 Drug Discovery and Protein Structure 8

5.5 Neoantigen Prediction 9

5.6 Clinical Trial Matching 9

5.7 Liquid Biopsy 9

6 Key Datasets and Resources **9**

7 The Infrastructure Problem **10**

8 Open Problems Worth Your Time **11**

9 Where to Start **12**

1 What Cancer Is

Cancer is evolution. Not metaphorically—literally. A cell in your body accumulates mutations that break two classes of gene: **oncogenes** (accelerators that get stuck on) and **tumor suppressors** (brakes that get knocked out). When enough of both fail, the cell divides without constraint, and natural selection at the cellular level favors the fastest-growing, hardest-to-kill variants. A tumor is a population under selection pressure.

This is why cancer is hard. You are not fighting a static enemy. You are fighting an evolutionary process that adapts to everything you throw at it.

1.1 The Hallmarks

Hanahan and Weinberg (2000, updated 2011 and 2022) defined the **hallmarks of cancer**—the capabilities a cell must acquire to become malignant:

1. Sustaining proliferative signaling
2. Evading growth suppressors
3. Resisting cell death (apoptosis)
4. Enabling replicative immortality (telomerase activation)
5. Inducing angiogenesis (growing blood vessels to feed the tumor)
6. Activating invasion and metastasis
7. Avoiding immune destruction
8. Tumor-promoting inflammation
9. Genome instability and mutation
10. Deregulating cellular energetics

The 2022 update added: unlocking phenotypic plasticity, nonmutational epigenetic reprogramming, polymorphic microbiomes, and senescent cells. Every hallmark is a potential computational target.

1.2 The Central Dogma

DNA $\rightarrow$ RNA $\rightarrow$ Protein. Cancer bioinformatics intervenes at every level:

- **Genomics** — sequencing DNA to find mutations

- **Transcriptomics** — measuring which genes are active (RNA-seq)
- **Proteomics** — measuring the proteins actually produced
- **Epigenomics** — measuring modifications that regulate gene expression without changing sequence (methylation, histone modification)
- **Metabolomics** — measuring small-molecule metabolites

When people say "multi-omics," they mean combining two or more of these layers. Integration is an open computational problem.

2 How Cancer Is Classified

Three overlapping systems. All three matter. Understanding their interactions is where STS meets oncology.

2.1 Anatomical

Where the cancer started: breast, lung, colon, prostate, pancreas, etc. This is the oldest system and still determines which clinical department treats you.

2.2 Histological

What the cells look like under a microscope:

- **Carcinoma** — epithelial tissue (skin, linings of organs). $\sim$85% of cancers.
- **Sarcoma** — connective tissue (bone, muscle, fat)
- **Lymphoma** — lymphatic system
- **Leukemia** — blood and bone marrow
- **Melanoma** — melanocytes (pigment cells)

Formalized in the ICD-O (International Classification of Diseases for Oncology) coding system. The WHO Classification of Tumours (the "Blue Books") is the reference.

2.3 Molecular

What mutations and expression patterns drive the cancer. This is where bioinformatics enters. Example—breast cancer has four molecular subtypes defined by receptor status and gene expression (the PAM50 assay, 50 genes):

Subtype	**Receptors**	**Prognosis**	**Treatment**
Luminal A	ER+/PR+/HER2−, low Ki-67	Best	Hormone therapy
Luminal B	ER+/PR+/HER2±, high Ki-67	Moderate	Chemo + hormone
HER2-enriched	ER−/PR−/HER2+	Aggressive, targetable	Trastuzumab
Triple-negative	ER−/PR−/HER2−	Worst	Chemo, immunotherapy

The subtype determines the treatment. The classification *is* the care. This is Bowker and Star's insight applied at the molecular level: what the taxonomy recognizes determines what the patient receives.

2.4 Staging

The TNM system: **T**umor size (T1–T4), lymph **N**ode involvement (N0–N3), distant **M**etastasis (M0/M1). Combined into stages:

- **Stage I** — small, localized
- **Stage II–III** — larger and/or regional spread
- **Stage IV** — metastatic (spread to distant organs)

Staging determines treatment aggressiveness and prognosis. It is assessed through imaging (CT, MRI, PET) and pathology.

3 Key Mutations

You will encounter these constantly. Know them.

Gene	Type	Cancers	Why It Matters
TP53	Tumor suppressor	∼50% of all	"Guardian of the genome." Most frequently mutated gene in cancer.
BRCA1/2	DNA repair	Breast, ovarian	Hereditary risk. Angelina Jolie. PARP inhibitors exploit the broken repair.
KRAS	Oncogene	Pancreatic, colorectal, lung	"Undruggable" for 40 years. Sotorasib (2021) was first KRAS inhibitor.
EGFR	Growth receptor	Lung	Targetable with erlotinib, osimertinib. Common in nonsmokers.
HER2 (ERBB2)	Growth receptor	Breast, gastric	Trastuzumab (Herceptin) transformed HER2+ breast cancer.
PIK3CA	Signaling	Breast, colorectal	Common activating mutation. Alpelisib approved 2019.
BRAF	Signaling	Melanoma, colorectal	V600E mutation. Vemurafenib.
BCR-ABL	Fusion	CML (leukemia)	Imatinib (Gleevec). The original targeted therapy success story.

Drivers vs. passengers. A tumor may carry thousands of mutations. Most are passengers—they happened during replication but don't contribute to cancer. A handful are drivers—they confer selective advantage. Distinguishing drivers from passengers is a core bioinformatics problem.

4 How Cancer Is Treated

1. **Surgery** — physical removal. Still the most effective for solid tumors when caught early.
2. **Chemotherapy** — cytotoxic drugs that kill fast-dividing cells (including healthy

ones—hence hair loss, nausea, immunosuppression).

3. **Radiation** — targeted DNA damage via ionizing radiation.
4. **Targeted therapy** — drugs designed to hit specific molecular targets. Imatinib for BCR-ABL. Trastuzumab for HER2. This is where genomics directly informs treatment.
5. **Immunotherapy** — unleashing the patient's immune system. Checkpoint inhibitors block PD-1/PD-L1 (the "don't eat me" signal tumors use to hide). 2018 Nobel Prize (Allison and Honjo). Game-changer for melanoma, lung, some breast.
6. **CAR-T cell therapy** — engineer the patient's own T-cells to recognize tumor antigens. Revolutionary for blood cancers (certain leukemias and lymphomas). Solid tumors remain a frontier.
7. **Hormone therapy** — block hormones that fuel growth (tamoxifen, aromatase inhibitors for ER+ breast cancer).
8. **PARP inhibitors** — exploit BRCA mutations. If the tumor can't repair DNA and you block the backup repair pathway (PARP), the cell dies. Synthetic lethality.

Drug resistance is the central clinical problem. Tumors evolve under treatment pressure. A drug kills 99.9% of cells; the 0.1% that survive carry resistance mutations and repopulate. Tracking this clonal evolution through longitudinal sequencing is a bioinformatics problem.

5 Where Computation Enters

5.1 Variant Calling and Interpretation

You sequence a tumor genome. You get millions of variants relative to the reference genome. Which ones matter?

Pipeline: Raw reads (FASTQ) $\rightarrow$ alignment (BWA-MEM2) $\rightarrow$ variant calling (GATK, DeepVariant, Mutect2 for somatic) $\rightarrow$ annotation (VEP, ANNOVAR) $\rightarrow$ interpretation (ClinVar, OncoKB, COSMIC).

The hard part: Somatic variant calling (tumor vs. normal) is harder than germline because tumors are heterogeneous—they contain subclones at different frequencies. Low variant allele frequency (VAF) variants are hard to distinguish from sequencing errors.

5.2 Gene Expression and Subtyping

RNA-seq measures which genes are active and at what level. Differential expression analysis (DESeq2, edgeR) identifies genes that are up- or down-regulated in tumor vs. normal. Clustering on expression profiles yields molecular subtypes (this is how PAM50 works).

Single-cell RNA-seq (scRNA-seq) resolves heterogeneity within a tumor. Tools: Cell Ranger (10x Genomics), Seurat (R), Scanpy (Python). A single tumor biopsy can yield 10,000+ cells with distinct expression profiles. Spatial transcriptomics (Visium, MERFISH) adds location: which cells are where within the tissue architecture.

5.3 Digital Pathology

Whole slide images (WSIs) of stained tissue at 40x magnification produce images of $\sim$100,000 $\times$ 100,000 pixels. Pathologists have read these manually for a century. Foundation models now match or exceed pathologist performance on specific tasks: tumor detection, grading, biomarker prediction from H&E stains alone (predicting molecular status from morphology without sequencing).

Key models: PathAI, Paige, CLAM (attention-based multiple instance learning), UNI, CONCH, Virchow (foundation models for histopathology, 2023–2024).

5.4 Drug Discovery and Protein Structure

AlphaFold (DeepMind, 2020) solved protein structure prediction from sequence. AlphaFold2 predicts structures to experimental accuracy. This accelerates target identification and drug design: if you know the 3D structure of a protein, you can computationally screen millions of compounds for binding affinity (virtual screening, molecular docking).

RoseTTAFold, ESMFold, and protein language models (ESM-2, ProtTrans) extend this to function prediction, interaction prediction, and design of novel proteins.

5.5 Neoantigen Prediction

Somatic mutations produce novel peptides (neoantigens) that can be presented on the cell surface by MHC molecules and recognized by T-cells. Predicting which mutations will produce immunogenic neoantigens involves: peptide-MHC binding prediction (NetMHCpan), T-cell receptor recognition modeling, and clonality assessment. This is the computational backbone of personalized cancer vaccines (BioNTech, Moderna's mRNA-4157).

5.6 Clinical Trial Matching

$>$10,000 active cancer clinical trials, each with complex eligibility criteria written in clinical language. Matching a patient's genomic profile, medical history, and demographics to eligible trials is an NLP and knowledge graph problem. Current systems: TrialGPT, Tempus, Foundation Medicine's FoundationOne CDx.

5.7 Liquid Biopsy

Detecting circulating tumor DNA (ctDNA) in blood. The promise: early detection (find cancer before symptoms), monitoring (track treatment response without biopsies), minimal residual disease detection (is any cancer left after treatment?). The computational challenge: ctDNA is a tiny fraction of total cell-free DNA ($<$0.1% in early-stage cancer). Signal-in-noise at the extreme.

Key companies: Grail (Galleri multi-cancer early detection test), Guardant Health, Foundation Medicine.

6 Key Datasets and Resources

Resource	What It Is
TCGA	The Cancer Genome Atlas. 33 cancer types, 20,000+ samples, multi-omic (genomic, transcriptomic, epigenomic, clinical). The foundation.

ICGC/ARGO	International Cancer Genome Consortium. Global equivalent of TCGA.
GENIE (AACR)	Real-world clinical genomic data from major cancer centers. 200,000+ samples.
SEER	NCI Surveillance, Epidemiology, and End Results. Cancer registry. Epidemiological data.
COSMIC	Catalogue of Somatic Mutations in Cancer. Curated database of known cancer mutations.
ClinVar	NCBI database of variant-disease relationships.
OncoKB	Memorial Sloan Kettering's precision oncology knowledge base. Annotates variants with clinical actionability.
cBioPortal	Interactive exploration of TCGA and other cancer genomic datasets. Start here.
GDC	Genomic Data Commons. NCI's unified data portal for TCGA, TARGET, etc.
ClinicalTrials.gov	Registry of clinical trials. API available.
UniProt	Protein sequences and annotations.
PDB	Protein Data Bank. 3D structures.
AlphaFold DB	Predicted structures for ∼200M proteins.

7 The Infrastructure Problem

TCGA—the single most important dataset in cancer genomics—is approximately 80% white. Polygenic risk scores developed on European-ancestry populations perform significantly worse on African, Asian, and Latino populations. The molecular subtypes defined on these datasets may not capture the full spectrum of disease biology across human populations.

Triple-negative breast cancer—the subtype with the worst prognosis, the fewest targeted therapies, and the greatest need for computational research—disproportionately affects Black women. The populations with the greatest burden of disease are the least represented in the data.

This is the same problem described in the companion paper on voice AI infrastructure: whose needs does the taxonomy recognize? Whose variants are in the database? Whose cancers get the precision medicine?

Open problems:

- Population-specific variant interpretation (a variant "of uncertain significance" may only be uncertain because the reference population is too narrow)
- Federated learning across institutions without centralizing patient data
- FAIR data principles (Findable, Accessible, Interoperable, Reusable) applied to cancer genomic data
- Data sovereignty for indigenous and underrepresented populations
- Algorithmic fairness in diagnostic AI across demographics

8 Open Problems Worth Your Time

If you are a computational person looking for where to contribute, these are unsolved and consequential:

1. **Early detection from liquid biopsy.** Detecting cancer from a blood draw before symptoms appear. Signal processing at the limit.
2. **Predicting drug resistance.** Modeling clonal evolution under treatment pressure. Can we anticipate resistance before it emerges?
3. **Multi-omics integration.** Combining genomic, transcriptomic, proteomic, and clinical data into a unified patient model. No consensus method exists.
4. **Foundation models for pathology.** Training general-purpose vision models on millions of whole slide images. The ImageNet moment for histopathology is happening now.
5. **Neoantigen prediction for solid tumors.** Current vaccines work for some cancers. Predicting which neoantigens will trigger immune response in solid tumors remains unreliable.
6. **Automated wet labs.** Robotic systems that can run high-throughput experiments autonomously—cell culture, drug screening, sequencing library prep. Self-driving labs are emerging but not yet standard.

7. **Clinical trial matching at scale.** NLP over eligibility criteria + patient records + genomic profiles. The data exists. The matching doesn't.

8. **Cancer in the Global South.** Computational infrastructure for cancer genomics in low- and middle-income countries. The data gap is a death gap.

9. **Federated cancer data networks.** Privacy-preserving analysis across institutions without centralizing data. Technically feasible. Institutionally hard.

10. **Closing the loop: genotype to phenotype.** We can sequence a tumor. We can catalog the mutations. We still cannot reliably predict, from sequence alone, how the cancer will behave. This is the hardest problem in the field.

9 Where to Start

1. **Read Hanahan & Weinberg** (2000, 2011, 2022). The hallmarks papers. Non-negotiable.

2. **Explore cBioPortal.** Load a TCGA dataset (start with breast, BRCA). Click through the mutations, the expression data, the survival curves. Get your hands in the data.

3. **Run a variant calling pipeline.** Download a small TCGA BAM file from GDC. Align, call variants, annotate. Feel the pipeline.

4. **Do a differential expression analysis.** TCGA RNA-seq data, DESeq2 or edgeR. Tumor vs. normal. See which genes light up.

5. **Train a classifier.** Take PAM50 labels and raw expression data. Build a simple classifier (random forest, then try a neural net). See how molecular subtyping works from the inside.

6. **Read one clinical trial protocol.** ClinicalTrials.gov, pick a breast cancer trial. Read the eligibility criteria. Now imagine writing code to match patients to it.

7. **Look at a whole slide image.** OpenSlide (Python library). Download a sample from TCGA. Zoom in. Understand what pathologists see.

This is a living document. It will be updated as the field moves and as the author learns. Corrections,

additions, and arguments welcome at jam844@cornell.edu. For Ariel, Bea, and those we lost too soon.