Distribute: April 6, 2026 Due: April 16, 2026, 3:00 PM via email to David McDonagh

Problem Set 1 – Cancer Biology – Spring 2026

1. From Pablo Sanchez Vela's Lecture:

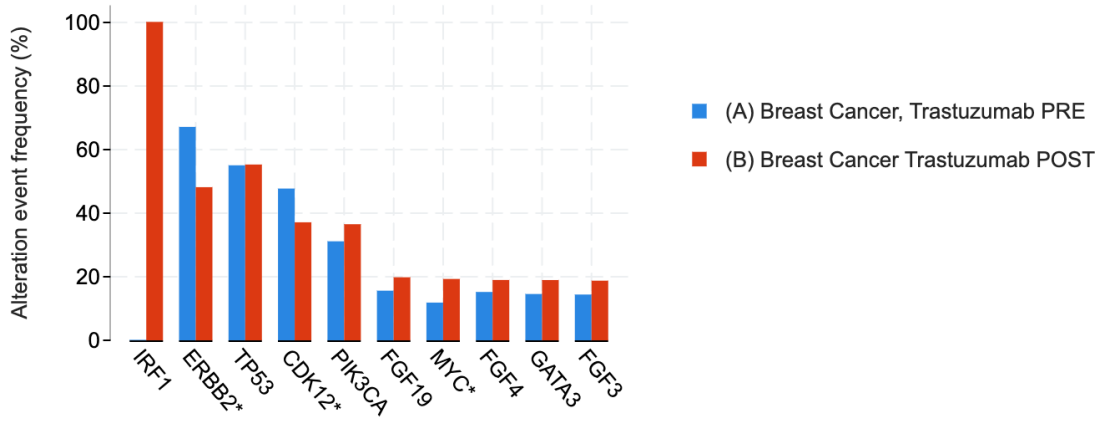
As part of your PhD project you are interested in understanding why treatment doesn't always eradicate cancer.

To generate some preliminary hypothesis, you first decide to evaluate the MSK-IMPACT clinical sequencing data using the cBioPortal platform*.

One the most commonly genes found to be amplified in the most common type of cancer in women is HER2. Anti-HER2 therapies like Trastuzumab have revolutionized the treatment landscape of breast cancer. Nonetheless, this disease is not yet curable once it has metastasized.

To understand how the mutational landscape of breast cancers changes upon anti-HER2 treatment exposure, you decide to depict the most frequently mutated genes from patients before and after treatment with Trastuzumab. This identifies *IRF1* as uniquely mutated in post-trastuzumab samples**.

Please, design a project with observational and experimental data to study the role of *IRF1* mutations in mediating resistance, if any, to this anti-HER2 therapy using the "hallmarks of cancer" as a guide. Provide at least one of observational and one experimental assay for each hallmark and be specific about the proposed tests and controls as discussed in class (1-2 pages) ***.



*To explore sequencing and clinical information from patients that have undergone MSK-IMPACT testing, go to cBioPortal (Clinical Sequencing Cohort): https://cbioportal.mskcc.org/study/summary?id=mskimpact

** While interesting as an example for this exam, the results from these fast exploratory analyses should be interpreted with caution in the appropriate context. In your real project consider include clinical, diagnostic, and biological experts from the relevant disease area(s) in the planning and execution of your study. The disease management team (DMT) representative on the MH Research Data Governance Committee can serve as an initial contact to assist in finding an appropriate collaborator. Most studies, especially those analyzing clinical outcomes or large cohorts, will benefit from inclusion of a statistician.

***The objective of this exam is to evaluate your use of your agility to leverage the cancer hallmarks as tools to uncover novel biology and not your knowledge about the actual known functions of *IRF1*.

Problem Set 1 – Cancer Biology – Spring 2026

2. From Dr Adam Widman's lecture:

Liquid biopsy assays aim to detect tumor-derived genomic signals in plasma under conditions of extreme biological dilution and background noise. The choice of genomic features, sequencing strategy, and analytical framework must be guided by both biological constraints and statistical limits of detection.

(a) Genomic signal across clinical contexts (10)
For each of the following settings, describe a strategy that can help boost ctDNA signal above background noise:

Settings:

- Minimal residual disease (MRD) after surgery in early-stage cancer
- Treatment monitoring
- Resistance mutation detection
- Early detection / screening
- Therapy selection

(b) Statistical limits and experimental consequences (10)
Early detection and MRD both target extremely low tumor fractions but differ fundamentally in prior probability and acceptable error rates. Explain how these differences affect:

- Required specificity and tolerance for false positives
- Assay design (e.g., breadth vs depth, longitudinal sampling)
- Experimental design choices, including controls and validation cohorts

---

Problem Set 1 – Cancer Biology – Spring 2026

3. From Dr Ivan Maillard's lecture:

Please imagine that you are tasked to create a ligand-receptor system inspired by the biology of Notch signaling. The system should be generally based on the principles of ligand-mediated proteolytic receptor activation, but differ from Notch signaling itself. Can you imagine a biological process, or a set of biological processes, that could be regulated efficiently by such a system? can you describe possible designs for components of this system? and possible therapeutic applications? Please use your imagination and/or inspiration from synthetic biology

Problem Set 1 – Cancer Biology – Spring 2026

4. Based on Dr Jian Carrot-Zhang's lecture (submitted by Tejiri Agbamu):

Each section is worth 5 points.
Entire answer should be 1.25-2 pages.

You've been approved for an IRB, and a pathologist has sent you tumor samples from 30 lung adenocarcinoma (LUAD) patients. You perform MSK-IMPACT on the tumors and matched blood normals. You come across some of the following challenges/questions during your genomic analysis workflow.

a) In analysis, you observe that a variant is present at ~15% of tumor reads, and your tumor purity is estimated at 60%. Explain how tumor purity and subclonal architecture each affect the observed variant allele fraction (VAF). What are some of the challenges to calling somatic mutations in low purity samples and what are some computational approaches that exist to account for them?
b) You have a reason to suspect that there are structural variants (SVs) present in one of your samples. After loading your BAM into IGV you notice three sequence features that are used to detect SVs using short-read targeted sequencing. **1)** What are they? **2)** What are the most common structural variants in LUAD and **3)** at what rate do they appear in the MSK-IMPACT LUAD cohort (https://cbioportal.mskcc.org/study/summary?id=mskimpact)?
c) Chromothripsis (CTH) occurs at a rate of 40% in LUADs. CTH and breakage-fusion-bridge (BFB) cycles both generate complex rearrangements. **1)** What is one hallmark that distinguishes these two? **2)** How does extrachromosomal DNA (ecDNA) arise from these events and how does it provide a proliferative advantage?
d) MSK-IMPACT offers several feasible advantages over whole genome sequencing (WGS) such as cost and time, however there are several limitations. **1)** Describe three limitations that targeted sequencing panels have in comparison to WGS**. 2)** Then explain how the GLIMPSE imputation framework leverages off-target reads to partially overcome some of these limitations.

Problem Set 1 – Cancer Biology – Spring 2026

5. Based on Dr Scott Lowe's lecture (submitted by Isabella Del Priore):

TP53 is the most commonly mutated gene across cancers and is frequently mutated in pancreatic cancer. Most p53 mutations observed across cancers are missense mutations, with 25% of those falling into 5 'hotspot' mutations. One quarter of *TP53* mutations are nonsense or frameshift mutations predicted to encode truncated proteins, whereas the remainder consists of splice site single nucleotide variants and in-frame indels of unclear biological significance. These mutations typically occur on a single allele, while the second *TP53* allele is disabled by loss of heterozygosity (LOH) by segmental deletion.

While analyzing pancreatic cancer sequencing data, you identify a p53 mutation of unknown significance, R277G, and decide to characterize it. How would you test the functional contributions of this mutation to cancer progression using mouse models of pancreatic cancer? Please discuss the following:

How would you test if the R277G mutation confers a loss or gain of function?

How would you test the downstream functional output of the R277G mutation (ie, if it controls any specific gene programs)?

How would you test if the R277G mutation confers a fitness advantage?