COM SCI C121 Week 9

Aidan Jan

May 30, 2024

Experimental Design, Robustness, and Functional Genomics

Functional Genomics

- The study of the function of many genes at once (or how genes bbehave in different contexts based on some form of a perturbation)
- "Functional genomics focuses on the **dynamic expression of gene products in a specific context**, for example, at a specific developmental stage or during a disease. In functional genomics, we try to use our current knowledge of gene function to develop a model linking genotype to phenotype."

Robustness

- Let's assume we have a *hypothesis* we want to ask:
  - If I perturb gene X, how will gene Y behave?
- There is *reality*, there is my belief of reality, and there is my tractable representation of my belief of reality
  - Which is my generative model and which is my inference model?
- Robustness refers to a model that can behave well when there are varying levels of misspecification.

Experimental design in genomics is hard because of high-dimensional sampels and so many hypotheses

- There is a lot of work in the classical statistical literature around "experimental design"
  - e.g., I'm designing a randomized controlled trial to test the efficacy of a drug. If I believe the effect to be "around" some value, how many samples do I need?
- In modern genomics, there are often many moving part in *high dimensions*.
- The fact that so many parts are moving makes this question of sample size very difficult because in some sense, every hypothesis is interrelated.

Experimental design can be used for robustness testing and for designing experiments

- Imagine you have a reasonable generative model of some experimental process
- If I want to perform an experiment in a new setting, I can simulate under that process
- Because I know what I changed in the generative model, I can evaluate how well I am doing on the inference side. If the generative model is already different than the inference model, I've done a robustness analysis for free

Functional Genomics Today: Genetic Variation

- A fair amount of work is done on the statistical genetics side for functional genomics
- Here, a perturbation is a genetic variant
- A simple example is a cis-eQTL
  - "cis" = jargon for "nearby". In practice, within 100000 bases of a gene.
  - "eQTL" = expression quantitative trait loci. Jargon for gene expression that changes as a function of the genetic variant.

Experimentally Induced Perturbations

- In the past, you pipette some lead into some cells on a dish, then look at gene expression differences
- The function transcriptional (gene expression) changes in response to a stress (lead) is an experimentally induced perturbation.

CRISPR

- CRISPR can be used to knock out genes by introducing variants
- Briefly, a short guide RNA (sgRNA) 2̃0 bases long is engineered.
- Together, with Cas9, the complex searches for that sequence, cuts the DNA, then the cell "repairs" the DNA, and totally screws up super often, thus breaking the cell (NHEJ, non-homologous end joining)

CRISPR Works to change one gene, but how about many genes?

- Remember engineered barcodes? We can engineer guides!
- There is an entire field of how to generate these efficiently
- The GeCKO library targets 19050 genes with 123411 sequences.
- It only costs $600.

How do you put CRISPR library DNA into cells

- Amplify the library DNA
- Put the library DNA in phages and have them "infect" cells.
- Each cell gets a different snippet of DNA and you have basically run 10k experiments in one.

Exercise: A Generative Model for Single-Cell Perturbations

- How would you play the oracle for generating single-cell RNA data with perturbations?

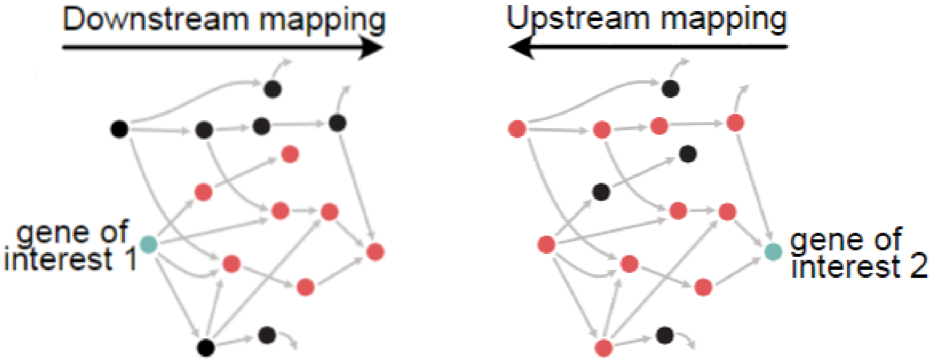
1. Choose a cell type
2. Choose a gene to knock out
3. Measure sampling (DNA)
4. Repeat for every cell in sample.

Cell Types:

1. $t_c \sim \text{Categorical}(p_t)$
2. $k_c \sim \text{Categorical}(p_k)$
3. for each g in numGenes:
   - $y_{cg} \sim \text{Poisson}(\mu_g \exp(\sum_a \alpha_{ga}^{\mathbb{1}\{k_i=1\}}))$

- where p_t is the proportion of cell types.
- A Categorical distribution is the same as Multinomial$(1, p_t)$
- Part 3 essentially means, what is the probability that the gene is being affected by the given knockout and not another?
- μ is the normal expression rate of a trait (some trait of some cell in standard conditions). α is the "effect size", a constant. Positive means the gene being knocked out causes expression to occur more than the mean. Negative means it expresses less than the mean. Zero means that the removal of that gene does not affect the cell (e.g., expression of the studied trait is the same as the mean).
- The goal is to use the knockout and effect data to predict which parts of genes are the most 'important' in what traits the cells express.

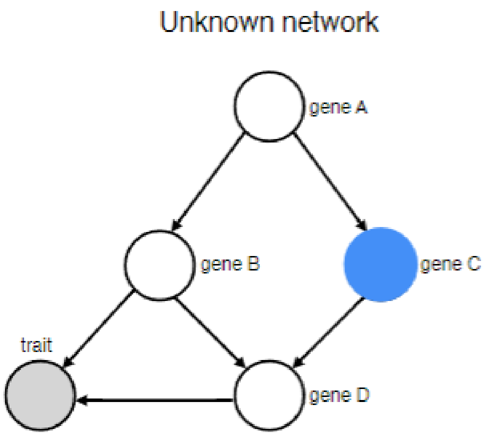
Actual Functional Genomics: FACS Screens

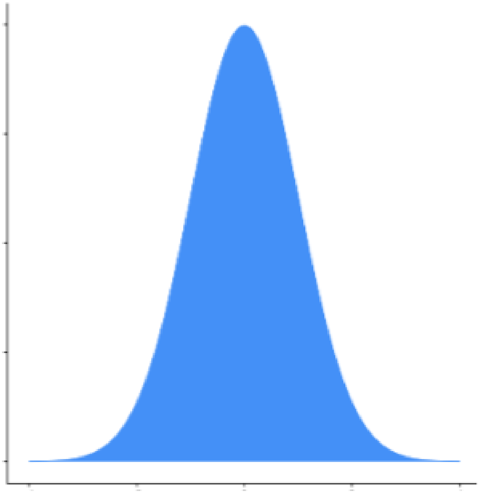


- Downstream mapping is studying what a changing a gene causes.
- Upstream mapping is studying what causes a change in a gene.

Most of the time, we want to know the changes in other genes that causes changes in the trait we are observing, but that is really hard to do since we would need to know what all the genes do.

Upstream Regulators can be Inferred by Perturbations

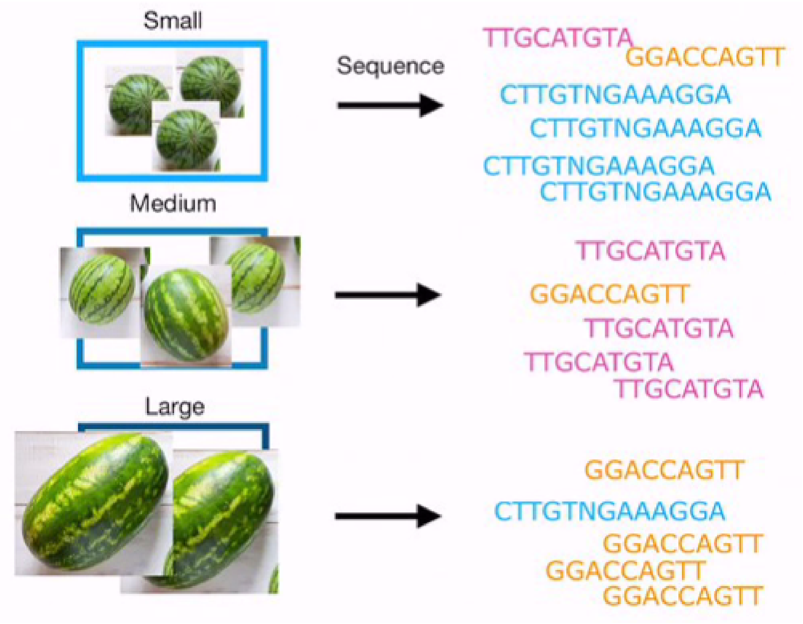




Suppose we care about the trait. We do a "do() operation" on gene B. Basically, we set the expression of Gene B to 0 by knocking out the gene. (`do(Gene B = 0)`). Now, the only genes affecting the trait are genes A, C, and D, in a straight line. Now, if we modify Gene C, we can see its effect on the trait.

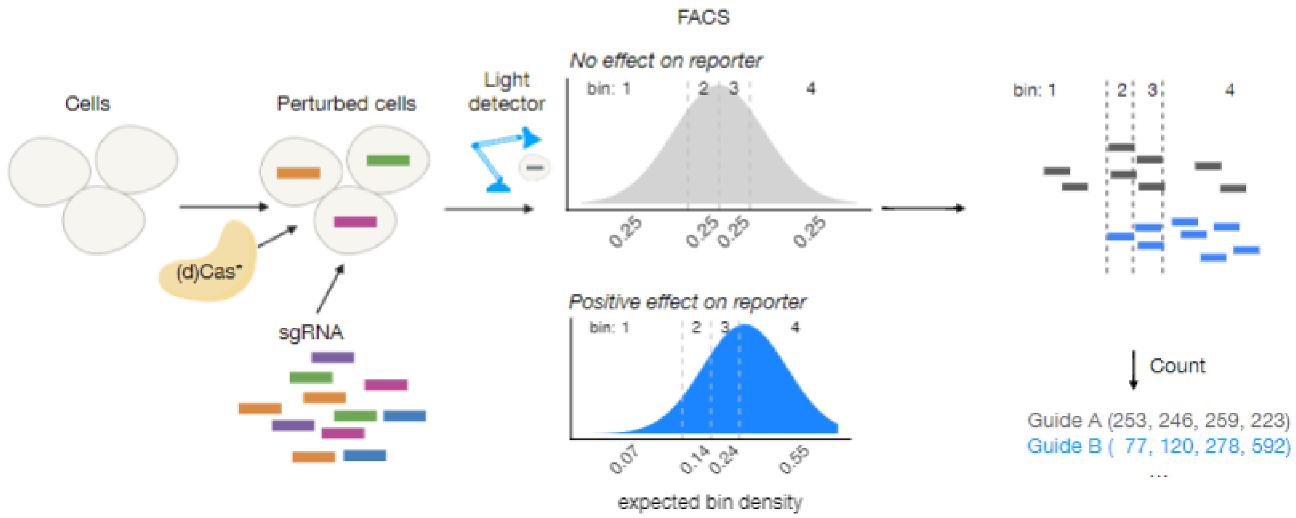
We can use this technique, for say, finding genetic variants associated with size of watermelons.

1. Take the seeds from one watermelon and mutate the seeds
2. Grow watermelons and sort them by size.
3. Analyze DNA of small melons vs big melons.

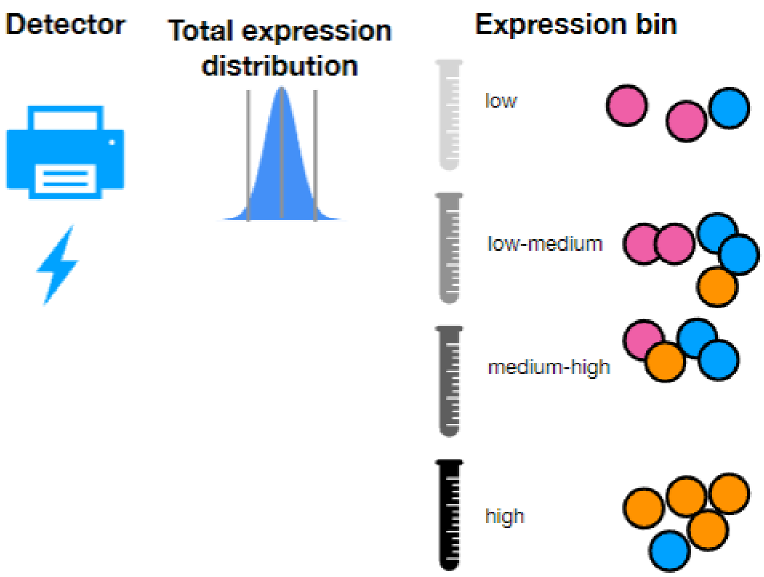


In this example, watermelon size was determined by which segments appeared more often (likely correlates to the amount of certain proteins synthesized.)

Forward Genetics Using FACS-Based Screens Finds Upstream Regulators of a Gene



1. Apply sgRNA and viruses to create perturbed cells. In FACS, genes are tagged with fluorescent molecules, the color of which is a "barcode" for the gene received.
2. Let the cells grow and observe the expression of the trait of interest.
3. Group all the traits into "bins"
4. Use a light detector on each bin to determine a correlation between the sets of genes (color) with the amount of expression.



How Do We Design Better Screens?

A big problem with flow cytometry is that you need a lot of cells. There will be read errors, dead cells, and a lot of variation, so to get a good dataset, millions of cells are needed.

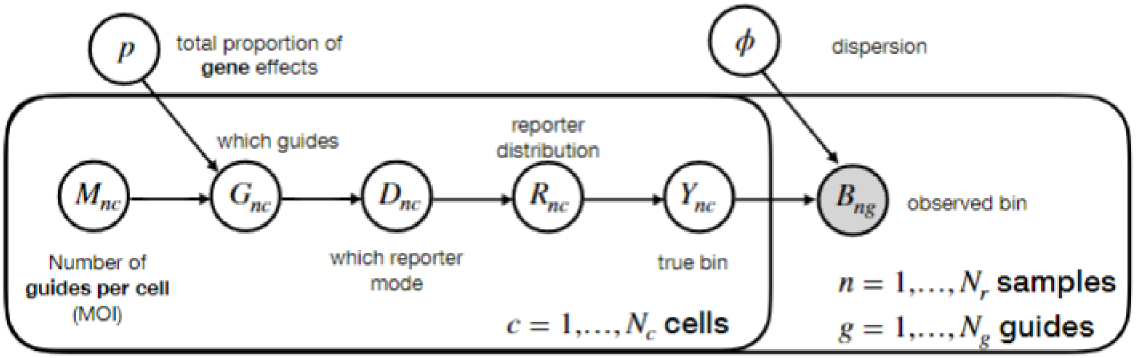
The **goal** is to design a screen such that we have high confidence in calls and probe as many targets as possible

Approach:

1. Develop a model
2. Learn parameters from existing data
3. Simulat unobserved conditions
4. Evaluate performance
5. Repeat

Pre-data Modeling Enables Efficient Screen Design...

... by using prior data for informative predictions


p
total proportion of gene effects
ϕ
dispersion
which guides
reporter distribution
M_{nc}
G_{nc}
D_{nc}
R_{nc}
Y_{nc}
B_{ng}
observed bin
Number of guides per cell (MOI)
which reporter mode
true bin
$c = 1, \ldots, N_c$ cells
$n = 1, \ldots, N_r$ samples
$g = 1, \ldots, N_g$ guides

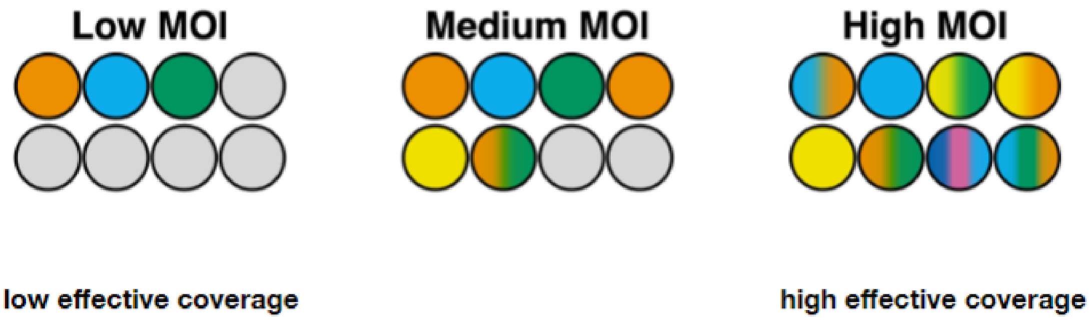
This is a generative model.

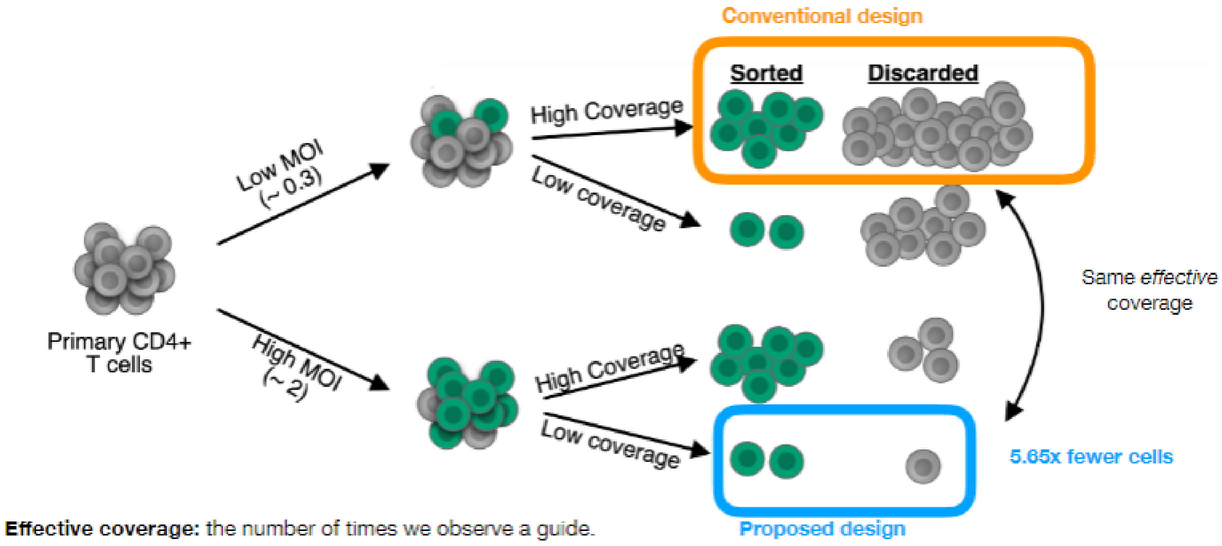
- $M_{nc} \sim \text{Poisson}(\lambda)$
- We pick a (single) cell with some number of guides in it.
- The G_{nc} node picks which guides they are, where how much each guide affects the observed trait is drawn from p, the total proportion of gene effects. $G_{nc} \sim \text{Mult}(M_{nc}, p)$
  - $p \sim \text{Dirichlet}(\frac{1}{N_g}\mathbb{1}_{N_g}\delta)$
- Skip D_{nc} for now, since it is somewhat complicated.
- At this point, if we draw many cells, as long as it is reasonably many, we will end up with a similar distribution for genes, since all the numbers so far are drawn from distributions.
- The reporter distribution essentially shifts the distribution to the "true distribution" (this is a linear transformation on the mean of the distribution)
- This is repeated over c cells to have a observed bin for a sample. This is then repeated over all samples and all guides.

[FILL]

Multiplicity of Infection (MOI)

- Multiplicity of Infection (MOI) = the rate of virions to cells
- Effective coverage = the number of times we observe a guide
- The higher the MOI, the better. A low MOI leads to low effective coverage - most cells are wasted because they were never genetically altered.





Effective coverage: the number of times we observe a guide.