5-quantification

Monday, February 9, 2026 12:01 PM

Recall in our discussion a couple weeks ago about metagenomics that we sometimes care about what species are present. Sometimes, we also care about how much is present. Or, similarly, when using RNA-seq for gene expression, we want to know not just what genes are being expressed but how much.

Gene Expression Analysis

Our genome is shared by (nearly) all the cells in our body, yet different tissues do different things. What makes cells different?

One answer: which genes are being expressed

Recall: Central Dogma

$$\text{DNA} \longrightarrow \overbrace{\text{RNA} \longrightarrow \text{proteins}}^{\text{specific to tissue}}$$

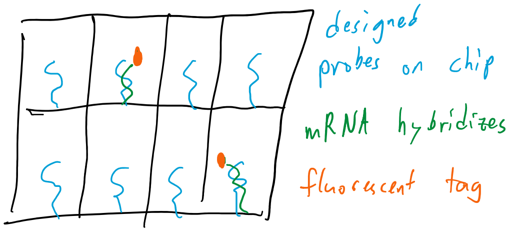
DNA: shared by all cells

RNA: mRNA can be measured instead.

proteins: we can directly measure this using lab techniques like mass spec, Western blotting, ELISA antibodies, but hard to measure entire protein landscape of cell cheaply, as proteins are very different from each other structurally.

One way: Hybridization & Scanning (Microarrays)

older tech, but still in use, since cheaper



Another way: RNA-seq ← use all the DNA analysis chemistry + algorithms for mRNA! Can look for all possible RNA at once

RNA-seq quantification

Idea: If we can align each read to a gene, then all we have to do is count up the number of reads!

Gene A:	500 reads
Gene B:	500 reads
Gene C:	1000 reads

Does this mean that genes A & B are expressed at the same rate?

NO!

What if gene A is 10 times as long?

Gene B: 500 reads

Gene C: 1000 reads

NO!! What if gene A is 10 times as long?

We need to normalize by gene length.

Solution: Normalize by gene length.

RPK (reads per kilobase)

Gene A:	500 reads	length 10 kbp	50
Gene B:	500 reads	length 1 kbp	500
Gene C:	1000 reads	length 5 kbp	200

Aside: We use kilobase here only out of historical convention.

When performing a measurement, you may want to replicate it to make sure your result isn't a flake.

e.g. Techical replicate: run the same sample again through sequencing

Clinical replicate: run samples of same phenomenon from multiple individuals.

Also, may want to compare healthy vs. diseased individuals, so need ways to compare.

How do we make sure we are comparing Apples & Apples?

	Sample 1:		Sample 2:	
Gene A	50	RPK	50	RPK
Gene B	500	RPK	1	RPK
Gene C	200	RPK	10	RPK

↔ does this imply that we have as much gene A expressed in Samples 1+2

NO! Maybe we sequenced Sample 2 with a lot less coverage, so A is actually higher?

Implicit Normalization Assumption: total amount of mRNA in cells is constant (NOT always true) — Why? Maybe cancer has put cell in overdrive.

Define: Reads Per Kilobase per Million of mapped reads (RPKM) [2008] [Mortazavi, et al] is given by

$$RPKM(G) = \frac{R}{(L/10^3)(M/10^6)}$$

R = # reads mapped to G

L = lengeth of G

M = # mapped reads in experiment

We are now normalizing by both gene length and sequencing depth. Are we done?

(100 reads)

		Length	Reads	RPK	RPKM
Ex 1	Gene A	1 kbp	500	500	500,000
	Gene B	1 kbp	500	500	500,000
	Gene C	10 kbp	0	0	0
Ex 2	Gene A	1 kbp	500	500	500,000
	Gene B	1 kbp	0	0	0
	Gene C	10 kbp	500	50	50,000

RPKM says gene A's expression is unchanged.

But how many molecules of mRNA were there in the cell?
It takes only 10% as many gene C mRNA molecules to get the same # reads.

Define Transcripts per Million (TPM)

$$TPM(G) = \frac{RPK(G)}{\sum_S RPK(S)} \cdot 10^6$$

} Normalize by number of transcript molecules, rather than total number of reads, which scale with transcript length.

[Wagner, 2012]

Most modern tools now use TPM because what we care about is # transcripts/proteins translated, rather than # reads.

Important: don't mistake the most easily measurable metric for the thing biologists care about.

Exercise: Show how to convert RPKM to TPM & vice versa. (if possible)
(in case you ever need to reanalyze old datasets)

But recall that our base assumption here is that different cells/tissues express the same total amount of protein/mRNA. What if this is not true?

Normalizing between samples

Some possible normalization assumptions: (decreasing order of assumption strength)

(1) Set of values in array are constant → Quantile normalization
(but different genes may be assigned to a value)

(1) Set of values in array are constant (but different genes may be assigned to a value) → Quantile normalization

(2) Values are normally distributed with same mean + variance → Scale Factor normalization

(3) Some (known) set of genes do not change, and should stay the same → Invariant Set normalization

Scale Factor Normalization

Let n = total number of genes

T = total number of samples

$y_{i,j}$ = expression of gene i in sample j

Let $$\underbrace{M_j = \frac{1}{n}\sum_{i=1}^{n} y_{i,j}}_{\text{sample } j \text{ mean}}, \quad \underbrace{M = \frac{1}{T}\sum_{j=1}^{T} M_j}_{\text{mean expression across samples}}.$$

Transform $\hat{y}_{i,j} = y_{i,j} - M_j + M$ (normalize just means)

Exercise: Describe how normalizing by means affects expressions given in RPKM vs. TPM.

That transforms the means. Be careful when you do this, as you might end up with non-biological values.

Assume that variances are also the same. Then

let $$\underbrace{V_j = \frac{1}{n}\sum_{i=1}^{n}\left(y_{i,j} - M_j\right)^2}_{\text{sample variance}} \qquad \underbrace{V = \frac{1}{T}\sum_{j=1}^{T} V_j}_{\text{average sample variance}}$$

And transform $$\hat{y}_{i,j} = \frac{(y_{i,j} - M_j)\sqrt{V}}{\sqrt{V_j}} + M$$

This transforms both mean + variance, assuming that your data is Gaussian to begin with.

Exercise: Think about [illegible] like [illegible] (or when) not biological gene expression data

Your data is Gaussian

Exercise: Think about whether (or when) or not biological gene expression data can be thought of as Gaussian. What goes wrong?

But what if you are in a situation where you know biologically that some genes are upregulated & total mRNA is preserved.

Exercise: Using ideas from this section, design an appropriate hybrid normalization strategy for the above.

Metagenomic profiling

Problem: How do we determine what species are present in an environmental sample, and how abundant they are?

Idea: If we know a set of candidate species' genomes, why not just map all the reads to those genomes & count them up?

Yes! But somewhat slow.

Exercise: Can you apply the same kinds of ideas from RNA-seq quantification to metagenomic profiling? What changes & what stays the same?

Pseudo-alignment, Quasi-mapping, Discriminative k-mers

Kallisto [Bray, Pimentel, Melsted, Pachter, 2016, Nature Biotech] } RNA-seq

Salmon [Patro, Duggal, Love, Irizarry, Kingsford, 2017, Nature Methods] }

Kraken [Wood, Salzberg, 2014, Genome Biology] —metagenomics

Last time we finished off with skani, which computes ANI uses an approximate alignment, which was much faster than a full alignment because it avoided seed & extend. Can we do the same thing with quantification?

Idea: To count a read for a (gene/genome) reference, we only need to know that the read maps somewhere in the reference, and not exactly the alignment.

Example: Reference A: ···ACCGATTACACC·····

Reference B: ·····TCTACCACGTA····

Read: ACTGATTACAGC

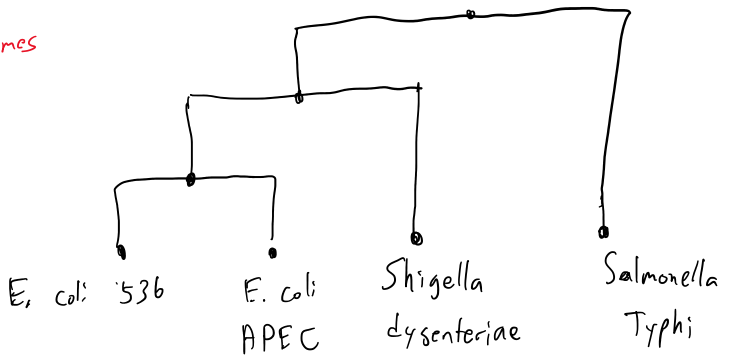
Can infer immediately that IF it maps to one of A or B, it likely map to A.

Discriminative k-mers are k-mers that are unique to a reference. If we spot it in a read, we can immediately count that read without alignment.

What do we do about reads without discriminative k-mers?

Kraken uses a phylogenetic tree to assign it higher up the tree.

which bins reads by bacterial genomes



This only works if we have enough discriminative k-mers.

Idea 2: Look at the k-mer spectrum of a read and see if it is compatible with a gene.

} introduced term "pseudoalignment"

More precisely, for any read, look at all its k-mers that are present in any reference. A read is compatible with a reference X if all its k-mers (that are in any reference) are in reference X.

Instead of using perfectly unique k-mers, we can use k-mers that are in multiple genes, & just look for compatible genes.

Example: Gene A: ACCTAACCGATTACACC

Gene B: GCTAACCGATTTCATTC

read: AGATTCCACC → only compatible with the gene A

read: GATTGTAACC → compatible with both

Idea 3: Chain together maximal exact matches (MEMs) } sometimes called "quasi-mapping"

Idea 3: Chain together maximal exact matches (MEMs) } sometimes called "quasi-mapping"

Define: A maximal exact match cannot be extended either left or right without mismatch.
Can use bi-directional FM-index to find.

Exercise: Provide an efficient algorithm to find all MEMs.

Gene A: ACCTAACCGATTACACC

Gene B: GCTAACCGATTTCATTC

read: AGATTCCACC → still compatible with A

read: GATTGTAACC → not compatible because of chaining

Exercise: Think about the trade-offs to chaining with MEMs vs. chaining with k-mers.

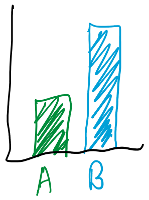
Differential Expression Analysis – Statistical Caveats

We now can quantify expression of genes or abundance of bacteria.
This is a starting point for understanding clinical relevance.

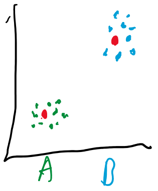
Example: Suppose expression of gene X doubles in condition B compared to condition A.
How reliable is this observation?

(In early days, this threshold of 2-fold increase was used to select genes)

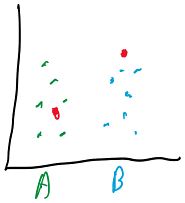
Problem:



Could be from



or



Comes down to variation estimation.

What is the chance of observing the results we did by random chance.

If we have a ton of data, can learn distributions directly.
Otherwise, may have to assume appropriate strong statistical model

e.g. Gaussian, binomial, Poisson, negative binomial, etc.
(bell curve) (successes in coin flips) (rare event counter) (number of coin flips needed to obtain # successes)

(bell curve) (successes in coin flips) (rare event counter) (number of coin flips needed to obtain # successes)

often useful for biological replicates because maybe you did the experiment multiple times until you got enough successes to please your PI.

Recall: Hypothesis testing from statistics! p-values, Bonferonni correction, etc (False Discovery rate with multiple hypotheses)

Exercise: How might you apply these kinds of statistical ideas when designing an algorithm that relies on chaining MEMs like Salmon?