week 01: the importance of doing arithmetic

Let's start with a brief introduction to genomes, genes, mRNA transcription, and RNA-seq experiments. This is mostly intended for those of you with little biology background.

Even if you're a biologist, though, you should find this intro useful. Throughout the course, among other things, we'll emphasize *quantitative intuition*: thinking about biological systems in mathematical and physical terms. I don't mean anything sophisticated by that. I mean thinking in terms of ballpark numbers, simple quantitative relationships, and common sense. You should have a feel, even just within an order of magnitude, of when experimental results make sense. If an RNA-seq experiment tells you your favorite gene is expressed at one million RNAs per cell, you should know that a typical cell doesn't even have that many mRNAs total.

Physicists sometimes call this kind of rough and ready numerical thinking *Fermi estimation* -- a nod to Enrico Fermi's legendary capability for doing rough calculations quickly. And like everything that's important in science, there's a classic xkcd about it.

what's in the human genome

On the board, I'll go through the rough architecture of the human genome, as one example genome:

- 3Gb haploid content in 24 chromosomes (22 autosomes, X and Y)
- Protein-coding genes: 20K (probably an underestimate)
- Protein-coding gene transcription units: about 40%, counting introns
- mRNA exons: 3% (including 5' and 3' untranslated regions, UTRs)
- coding: 1%
- conserved noncoding DNA: maybe 6%-ish,

squishy in part because there's no clear threshold
- Transposable elements (TEs) in various states of decay: about 55%
- all overlapping in various ways -- for example, TEs and conserved regulatory regions in introns, UTRs.

what is a gene (an RNA perspective)

Again on the board, I'll sketch the standard canon of how we understand genes, especially from the perspective of RNA transcription (neglecting protein translation, at least for now, because the course is going to use RNA expression so much as an example of data analysis):

- structure and size of transcription units - focusing on mammalian polII
- (other txn units: rRNA -- tRNA -- snRNA, snoRNA, other small RNAs -- miRNA -- lncRNA)
- transcription start -- promoter -- 5'cap
- transcription stop -- cleavage/poly-A+ addition
- transcriptional activation: promoters, enhancers, silencers
- RNA splicing -- introns and exons
- Isoforms: alternative transcription starts, stops, and alternative splicing
- RNA transport out of nucleus, and localization in cells
- translation -- RNA quality control -- nonsense-mediated decay
- RNA stability and decay

Some people are surprised that known, annotated protein-coding gene transcription units cover so much (40%) of the genome, because they've been told that coding genes only account for 1% of the genome, and there's been so much written about how "surprising" it is that much of the human genome is transcribed, by people who apparently forgot about introns and UTRs. It's easy enough to verify just by counting in the annotation of the current human genome assembly.

RNA content of a cell

To interpret an RNA-seq experiment we want to have some intuition for what we're counting. How many RNAs are there in a cell, from each gene? Here, I'm going to use specific numbers from

Jackson, Pombo, and Iborra, *The balance sheet for transcription: an analysis of nuclear RNA metabolism in mammalian cells*, though there's many other relevant references as well.

A HeLa cell (a cultured human cancer cell line) divides every 22 hours, and it contains about 3.5M (3.5 million) ribosomes. This means the cell has to make 3.5M new ribosomes every 22h.

This presents an engineering challenge for the cell. It's hard to produce ribosomes that fast. A major component of the ribosome are the SSU (small subunit) and LSU (large subunit) ribosomal RNAs, which add up to about 7Kb of mature rRNA, processed out of a 13Kb primary transcription unit. So the cell has to be synthesizing 13Kb x 3.5M / 22h / 3600 sec/hr = 600K RNA nucleotides per second to keep up the pace, just in ribosomal RNA transcripts.

Ribosomal RNA genes are produced by RNA polymerase I, which is specialized for rRNA synthesis (mRNAs are produced by polII), but it runs at about the same speed as polII -- around 40nt/s. So we need about 15,000 active polI polymerases to polymerize 600K nucleotides/sec of rRNA.

We just can't do it with just two rRNA loci in our diploid genome. Physically, the polI machinery occupies about 50-100nt of DNA; even if we load them as fast and as tightly as possible, we could get about one polI complex every 100nt or so, which is indeed this is what's observed in HeLa: about 100-120 active polI polymerases on each 13kb rRNA transcription unit. What to do?

The cell does it by parallelization. It amplifies the number of rRNA genes. The haploid human genome contains about 180 repeats of the ribosomal rDNA transcription unit, embedded in a 45kb repeat unit, in tandem arrays on chromosomes 13,14,15,21, and 22. About 8 Mb of our 3200 Mb genome is devoted to these tandemly arrayed rRNA units. (That's a lot, relatively. The rest of our 20,000 genes is coded by just another 35Mb total, scattered around the genome.) Since we're diploid - we have two copies of each chromosome - we have about 360 rRNA transcription units. According to Jackson et al. 2000, about 120-150 of these units in a diploid cell appear to be transcriptionally active at any one time. (I don't know what's up with the rest; that's an interesting discrepancy. I would have thought they'd all be active!) Now the factory's balance sheet works: 120-150 active units times 100-120 polI polymerases per locus times 40nt/s = 480K-720K nucleotides/sec.

Jackson et al. 2000 now switches to presenting numbers from a different mammalian cultured cell type (mouse L cells), but for Fermi-estimation-purposes, all mammalian cells are roughly the same. In total, a growing mammalian cell is synthesizing about 3000K (3M) nucleotides/sec; 39% of that active synthesis is polI on pre-rRNAs that we just talked about, 58% is polII on pre-mRNAs, and 3% is polIII on small RNA genes. Assuming that polI, polII, and polIII all move at around 40-50nt/sec, then we have something like 30K active polI polymerases, 45K polII (though Jackson says 60K), and 3K polIII.

Mouse L cells divide almost twice as fast as HeLa cells, so they're making rRNA at a total rate of about 1200Kb/sec, compared to HeLa's 600Kb/sec. Close enough for Fermi.

Happily, those deduced numbers aren't too far from other independent measurements in HeLa cells, which have estimated there's 15,000 active polI, 70,000 active polII, and 7,000 active polIII polymerases. Only about a third or so of RNA polymerases are actively engaged and transcribing genes at any given time; there's an estimated total of 300K polII polymerases in a HeLa cell. Recall there are about 3.5M ribosomes per HeLa cell -- the protein factories greatly outnumber the RNA factories, meaning that much more ATP energy equivalent is pouring through translation than through transcription.

There are about 20,000 protein-coding genes, multiplied by two because we're diploid, so if we suppose we have 60K active polII polymerases (and if we ignore other noncoding polII transcription, such as lncRNAs), we only have an average of about 1.5 RNA polymerases engaged per protein-coding gene. Only about half (10K) of genes are on in a given cell type, so call it an average of ~3 polymerases per *active* gene. Those polymerases are pretty lonely! The average polII transcription unit is around 40kb long, so polII polymerases are spaced about 10kb apart on typical active genes. A dark country road, compared to the jam-packed, hard-working polI rRNA transcription unit freeways.

The RNA population of big, intron-containing pre-mRNA transcripts is called *heterogeneous nuclear RNA* (hnRNA). Introns are spliced, typically rapidly, and splicing is happening on most transcripts even as they're still being transcribed. Only 2-3% of hnRNA (i.e. pre-mRNA) reaches the cytoplasm as mature mRNAs, after splicing and processing. Introns are big.

In terms of total RNA by mass at steady state (as opposed to synthesis rates), **75% of HeLa RNA is rRNA, 10% is tRNA, and 2.5% is mRNA** (and I think the missing fraction in these numbers is a combination of mitochondrial RNA, nuclear pre-mRNA, and small noncoding RNAs). Assuming that the 75% rRNA consists of 3.5M copies of 7Kb of SSU+LSU, and assuming that mRNAs are ~2Kb and tRNAs are ~100nt, we can estimate there are about 30M tRNAs per cell (about 10 tRNAs per ribosome), and about 400K mRNAs (about one for every ten ribosomes). That turns out to be about right!

Not all genes are expressed in every cell type; typically, about half (10K) are expressed in any given cell type. So, ballpark, with 400K mRNAs/cell and 10K active genes, we expect a mean of around 40 mRNAs/gene. (We expect a skewed distribution, with a lower median, because some genes will be highly expressed.)

When genes are "off" they're actually still being transcribed at some low rate. Turning genes "on" up-regulates them by about 100x (with a wide range, depending on the gene). A result that I think is important comes from Hebenstreit et al., *RNA sequencing reveals two major classes of gene expression levels in metazoan cells,* Molecular Systems Biology, 2011. Their key result is shown in the figure to the right. If you plot a histogram of RNA-seq expression levels in a single cell type, you tend to get a bimodal-ish distribution, as if there is an "off" peak at around 0.01-0.1 RPKM (very very roughly, 0.01-0.1 mRNA/cell), and an "on" peak around 20 mRNA/cell. (RPKM is "reads per kilobase per million reads", a normalized measure of expression level; more about that below.)

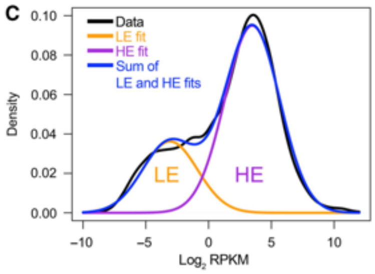

Figure 1c from Hebenstreit et al 2011 showing that mRNA expression levels per gene are bimodal-ish, sort of as if there's an "off" and an "on" distribution. Notice that the x-axis is a logarithmic scale. Fitting a normal distribution to the *logarithm* of the expression level means a so-called *lognormal* distribution. If we plotted it just as expression level, we'd we'd see a skewed distribution with a fat right tail: most genes are expressed at moderate levels, and a few at high levels. (*'open image in new tab' to enlarge.*)

This 20 mRNA/cell or so is so eerily close to what we expect from numerology that it's part of the working model that I currently have in my head when I read RNA-seq papers:

- the RNA content of a typical mammalian cell is almost all rRNA and tRNA; a tiny fraction (~2-3%) is mRNA.
- which consists of about **400K mRNAs**.
- from **10K active genes**, of 20K total.
- most active genes are expressed at around **20-40 mRNAs/cell**.
- inactive genes are **leaky, observed at 0.01-0.1 mRNAs/cell**.
- the difference between active and inactive is about **100-1000x**.

Some things about RNA-seq experiments make a surprising amount of sense in light of these ballpark numbers. For example, many people are doing cell type specific transcriptomics these days (including my lab). We expect to see 100-1000x "enrichment" for many genes in cross cell type comparisons, but instead, in many protocols, it's more typical to see 2-10x enrichment for genes that are "on" versus "off". This probably reflects the fact that a cell type specific purification protocol isn't perfect, and is carrying along substantial background contamination.

mRNA synthesis, decay, and steady state

Let's suppose, as an approximation, that mRNA levels in a cell are at steady state. Let's also assume that there are only two steps controlling mRNA levels: mRNAs are synthesized at rate k_1 (in mRNAs/hr/cell), and mRNAs decay at rate k_2 (in hr $^{-1}$). Steady state is when the number of new mRNAs synthesized equals the number that decay:

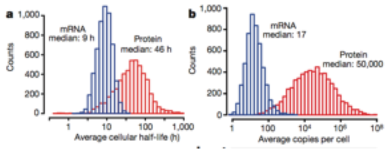

Figure 2a and 2b from Schwanhausser et al 2011 showing distribution of mammalian mRNA half-lifes (in hr), with a median around 9h, and mRNA steady-state levels (in mRNA/cell), with a median around 17. (*'open image in new tab' to enlarge.*)

$$k_1 = k_2[\text{mRNA}],$$

so at steady state,

$$[\text{mRNA}] = \frac{k_1}{k_2}.$$

So if the synthesis rate is 2 mRNAs/hr/cell, and the mRNA decay rate is 0.1 hr^-1, then there are 20 mRNAs/cell at steady state.

mRNA decay is more usually expressed by RNA biochemists in terms of the *half life* (in hours), not as the decay rate. The half life is $t_{1/2} = \frac{\ln 2}{k_2}$; and $k_2 = \frac{\ln 2}{t_{1/2}}$.

One reference for typical mammalian mRNA synthesis and decay rates is [Schwanhausser et al., 2011, *Global quantification of mammalian gene expression control*]. Some of their key figure panels are reproduced here to the right. In mammalian cells, typical synthesis rates are on the order of 2 mRNAs/hr; typical half lives are on the order of 9h; and steady-state mRNA copy numbers are on the order of 20.

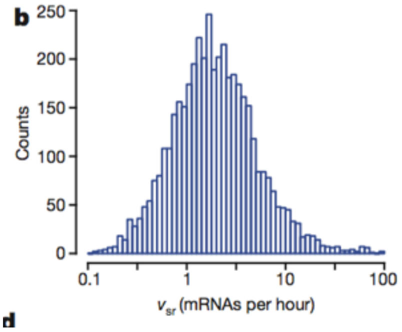

Figure 3b from Schwanhausser et al 2011 showing distribution of inferred mRNA synthesis rates, in mRNA/hr, with a median of around 2. (*'open image in new tab' to enlarge.*)

If polII initiates at 2 mRNA/hr, on two diploid copies of a ~30kb pre-mRNA transcription unit, as we saw above, the picture is very different from rRNA synthesis. At 40nt/s, it takes ~15m for polII to transcribe a typical locus, so again these numbers suggest that at any given time there's the order of zero or one polII complex on a typical gene.

Notice that I'm sneakily slipping back and forth between "mean" and "typical" (i.e. median). If we were talking about normal (Gaussian) distributions of things, mean and median are the same thing. But the distributions we're talking about are skewed: notice that the plots from the Schwanhausser paper are roughly lognormal. So I'm not too bothered by the discrepancy between less than one polII complex per "typical" gene, versus three or so polII's per gene *on average*; the lognormal distribution shifts the mean upwards from the median, the same way that average household income in the US looks pretty good until you look at the median income.

With small numbers like 20 mRNA/cell, you can run into problems with continuous approximations like the steady-state equations above. This is why there's a lot of interest in stochastic dynamics of small numbers of things per cell. For example, many promoters don't fire at a steady rate; single cell experiments have shown that they "burst", with bursts of initiation followed by more quiescent states. We're neglecting that in these ballpark numbers.

This steady state approximation neglects that dilution effect of cell growth and division: if there's 20 mRNAs/cell, then the cell has to make 20 new mRNAs per cell division to keep things constant. In a HeLa cell doubling every 22hr, that's not an insignificant concern.

an RNA-seq experiment

Conceptually, an RNA-seq experiment is straightforward, but the details can matter quite a bit in the interpretation.

In principle, "all" we're going to do is purify RNA transcripts from a biological sample, and measure the relative level of each transcript i in that sample -- τ_i , in units of transcripts per million transcripts (TPM) -- essentially by counting a random sample of short RNA fragments.

In practice, we need to think about what's going on at each step:

- **the biological sample** might be a cell culture, a whole animal that we're grinding up, a dissected tissue, or some sort of purified subset of cells. The RNA-seq experiment is going to give us **population averaged measurements** over all the cells in this sample, so we need to think about how heterogeneous it is, especially if we're going to compare different samples. Even within a cell type, mRNA expression levels may change over time (age, circadian rhythms...), with environmental conditions (light, food, behavior...), or with individual differences (genotype, sex). We may need to think about how to hold relevant confounding variables constant in our experimental design. If the sample is composed of a mixture of different cell types, population averages can change just because of changes in their proportion. For example, if you kick off an inflammatory response in your tissue in your experiment, it may be infiltrated with immune cells; you'll see an enrichment of genes expressed by those cells not necessarily because the genes turned on, but because the proportion of cells changed.

- **the RNA prep**, where we purify the RNA from the sample, may specifically favor some subsets of RNAs over others. Small RNAs (<100nt or so) may tend to get lost in standard RNA purification steps. mRNAs are often purified by poly-A+ selection, which is designed to select against noncoding RNAs (rRNA and tRNA).

- **the library prep**, where we make a set of (probably fragmented) double-stranded DNAs from the RNA, may also favor certain RNAs over others. Some libraries are dT-primed, so they will favor the 3'end of polyA+ mRNAs (and the occasional genomically-encoded poly-A+ stretch); random-primed libraries shouldn't, but will tend to prime on non-mRNAs (like rRNA) which may not be what you want.

- even **the sequencing protocol**, where we obtain short DNA sequences from one or both ends of our dsDNA library fragments, may have some biases, for example with respect to sequence (GC%) composition.

Nobody should be calling an RNA-seq experiment "unbiased". A good description of an RNA-seq experiment should describe all the steps that may have biased the sequences that ended up being collected.

from mapped read counts to TPM

Each gene i expresses some number of mRNA transcripts per cell; let's call that number t_i. (We're already making a strong simplifying assumption -- that there's only one mRNA splicing isoform per gene -- but let's go with that for now.) In an RNA-seq experiment, we've usually taken our RNA sample from an unknown number of cells, so we're generally going to have to think in terms of relative proportions, not absolute numbers. In a large sample, the proportion of mRNA transcripts from gene i is $\tau_i = \frac{t_i}{\sum_j t_j}$. These unknown τ_i expression levels are what we want to infer from our observed counts of mapped reads.

In RNA-seq data processing, we map short sequence reads to annotated transcript sequences, which gives us a number of mapped reads per transcript: c_i. These counts c_i are our observed data. The proportion of mapped reads that mapped to transcript i is $\nu_i = \frac{c_i}{\sum_j c_j}$. We can also consider the normalized ν_i to be observed data.

The RNA-seq experiment doesn't measure τ_i directly. We made a fragment library and obtained sequence reads from the ends of each fragment, so long mRNAs are more likely to be sampled than short ones. A basic assumption of short-read RNA-seq analysis is that we sample fragments, and therefore reads, from mRNA transcripts with a probability proportional to $\tau_i \ell_i$, where ℓ_i is the transcript length. (Imagine simulating the experimental process with a Python script.) So:

$$\nu_i = \frac{\tau_i \ell_i}{\sum_j \tau_j \ell_j}$$

and therefore:

$$\tau_i = \frac{\nu_i}{\ell_i} \left(\sum_j \frac{\nu_j}{\ell_j} \right)^{-1}$$

That is: first we normalize each ν_i individually by the mRNA length ℓ_i; then we normalize by the total sum of $\frac{\nu_i}{\ell_i}$ over all genes; and this gives us our estimates of τ_i.

It's convenient to scale τ_i by a constant that's on the same order as the number of mRNAs per cell: one million. (Fermi estimation again.) Thus τ_i values are reported in units of "transcripts per million transcripts" (TPM), which you can think of as being on the order of mRNAs per cell (very roughly, because different cells have more or fewer than 10^6 mRNAs in them).

Look back over that and you'll see several assumptions that we could try to improve on. We don't have to assume that each gene gives rise to only a single mRNA isoform, but because isoforms overlap (share exon sequences) we wouldn't be able to assume that there was an observable 1:1

relationship between a mapped read and a particular mRNA i; we'd have to make a statistical model that treats the mapping of a read to a specific isoform i as something we'd need infer. We also don't have to assume that all nucleotide positions are uniformly sampled by short reads, because of biases in the RNA-seq library generation and sequencing procedure; we could model that bias. We'll learn some techniques for such statistical modeling as we go forward in the course.

TPM vs. RPKM

The mathematical model above was explained in a seminal paper from Colin Dewey's group [Li et al., Bioinformatics 2010]. I've followed their notation. But it isn't the procedure that was first used in RNA-seq, when RNA-seq was first introduced by Barbara Wold's group [Mortazavi et al., Nature Methods 2008]. In Mortazavi et al., a different procedure was used, and gene expression estimates were obtained in units of "reads per kilobase per million mapped reads" (RPKM). (When people started using paired-end sequencing more often, it became more reasonable to talk in terms of mapping library *fragments* instead of independent *reads*, thus fragments per kilobase per million mapped reads (FPKM), but RPKM and FPKM are the same thing for our purposes.)

Converting mapped read counts ν_i to RPKM gives us $\frac{\nu_i}{\ell_i} \cdot 10^9$. ($10^9$ because *per kilobase per million*.) This is *proportional* to τ_i (within a given sample) but it's unnormalized. We would need to normalize by $10^{-3} \sum_j \tau_j \ell_j$ if we want to convert RPKM to TPM: that is, **by the abundance-weighted mean mRNA transcript length**. If the abundance-weighted mean mRNA transcript length is 1kb, TPM and RPKM are the same thing.

So do we care? A problem with RPKM arises when we start trying to compare across samples, because different samples don't necessarily have the same abundance-weighted mean mRNA transcript length. Maybe in one sample some long mRNAs went up, and some short mRNAs went down. Now we could see that an mRNA transcript i, present at exactly the same proportion τ_i in the mRNA populations of the two samples, would show *different* RPKM measures just because some *other* genes shifted their expression and altered the abundance-weighted mean mRNA transcript length.

Don't dis RPKMs too much though, even if it's amusing to see Lior Pachter flagellate himself over it, because you can get a related normalization artifact from TPMs too, and this artifact is probably even more common. The *relative* abundance of τ_i of transcript i is obviously going to be affected by the absolute expression level of every *other* mRNA in the cell, so it is a mistake to assume that a change in τ_i necessarily means a change in the expression level of gene i (in RNAs/cell). If you turn a bunch of some genes up, the relative proportion of other genes must go down even if their absolute concentration remains unchanged. One common way this can happen is if you alter the growth rate, which changes the expression level of a large battery of genes having to do with making ribosomes (among other things).

There's been an amazing amount of wailing and gnashing of teeth over RPKM vs. TPM. Much of it seems confusing and/or wrong to me. I think it's just an example of people passing along lore without having a quantitative understanding of what they're talking about. The Li et al. 2010 paper is clear and correct.