Quantitative Genomics Notes

Nick Sun

LECTURE 2

- A genome is the collective set of DNA contained in a cell
- Cells carry a copy of an organism's genome
- DNA is a polymer of deoxyribonucleic acids (long chain of molecules)
- The order of these acids determine the biological functions that cells can execute
- These functions are encoded in subregions of the genome that are called genes
- DNA has a forward strand and a reverse strand – if we know one, we can impute the other
- DNA carries two copies of the sequence, but we only represent one copy *in silico*
- The central dogma consists of transcription and translation
- **Transcription** is when cellular machinery reads DNA at gene locations to create RNA (or transcripts, as they can be referred to). RNA is similar to DNA except that it is single stranded and contains different chemicals
- **Translation** is when cellular machiner, notably ribosome read and translate mRNA to produce proteins.
- Proteins are made up of amino acids so we have to translate nucleic acids bases into these amino acids. This is done by 3 base sequences called codons. Multiple codons can code for the same amino acid. There are 20 amino acids total.
- We can do things using genomics like comparing closely related genomes (like genomes between different people) and see how different nucleic acid compositions contribute to cholesterol levels!

LECTURE 3

- DNA sequencing is the process of discerning the specific nucleic acid composition and order of a DNA molecule (can either be a short DNA fragment or an entire genome)
- DNA synthesis is the process of creating (or writing) a new strand of DNA from a DNA template
- DNA bases are floating inside of a cell
- An enzyme called DNA polymerase reads the DNA strand that is being replicated, grabs the appropriate complementary base, adds the base to the synthesizing strand, double checks this, then continues down the line.
- We can reproduce this process in a test tube by combining a DNA template, water, DNA polymerase and nucleotides in a test tube. As this process is happening in our controlled environment, we can digitally record which bases are getting added. This is referred to as **sequencing by synthesis**
- Illumina and PacBio platforms use a sequencing by synthesis procedure.
- DNA sequencing typically cannot sequence entire genomes, though there are emergent technologies that seek to solve this problem
- We use **shotgun sequencing** to sequence large molecules of DNA. Essentially, the molecule is fragmented, allowing us to divide the genome up into smaller bits that can be exhaustively sequenced. The tradeoff is that we lose knowledge of how the smaller fragments line up with each other in the original molecules. We can think of shotgun sequencing as a procedure that randomly samples from the genome, since not all fragments will be sequenced.
- Genome assembly uses informatics procedures to reconstruct the genome from these templates.

- Each single stranded template molecule is affixed to a flowcell which is just a device that keeps the DNA molecule floating around in solution
- Special nucleotides are incorporated into the synthesizing strand of DNA, which we shine a laser on, emitting a certain light. A camera takes a picture of that flash and the machine can figure out from the light what kind of base was added.
- DNA sequencing usually happens to a cluster of identical molecules as opposed to a single molecule. This helps amplify the flashes of light so it's easier for the machine to read.
- Obviously this isn't perfect. Big problem with homopolymer runs!
- **Multiplexing** is the simultaneous sequencing of multiple DNA samples. Reduces the cost of sequencing DNA per specimen. Basically entails us affixing an artificial genetic barcode to strands corresponding with a certain organism.
- It is desireable to sequence very long templates when doing things like genome assembly, but unfortunately DNA sequencing technologies are often seriously constrained in the lengths of sequences they can read. For example, Illumina's HiSeq platform generates 150 bp of sequence data
- One strategy for getting around this limited read length is generating paired-end sequence data where both ends of a target molecule are sequenced which helps reassemble larger sequences from smaller ones. Paired end sequencing works by sequencing the templates two distinct times, once for the forward strand and another for the reverse strand.
- *Mate Pair sequencing* is another technique we can use for getting information from distant regions of the genome.
- MinION is a portable Nanopore sequencer

LECTURE 4

- There are three basic kinds of DNA sequencing error: substitutions, insertions, and deletions
- Homopolymer runs are often affected by insertions and deletions
- Chromatograms can vary between high quality and low quality reads (depends on the smoothness and distinctness of the peaks)
- We often represent the probability that a called base is an error as p. Under the hood of the machine, it is not only calling bases as cluster emitted signals, it is also using patterns of signal emission to discern whether a called base is likely an error or not.
- The Phred quality score is taken as -10log(p) where log is log base 10. Higher score is better, with a really good score being 40, corresponding to a base accuracy call of 99.99%
- Sequence data files (FASTQ) files include both base calls and corresponding quality scores.
- FASTA and SAM files typically do not have the associated quality scores in them
- Sequence quality tapers off as read length increases
- Almost all quality control procedures require you the user to apply some sort of threshold that stratifies acceptable quality scores from unacceptable quality sequences
- Since you are generating a lot of data, it is usually alright to be conservative.

LECTURE 5

- Mutations or changes to an organism's DNA can affect an organism's phenotype.
- Evolution results in sequence dissimilarity among related genomes
- Homologs are genes that share a common ancestor: homologs can be used to infer the function of a new gene, determine which nucleic acids are critical to a gene product's function, or quantify how rapidly a species has evolved.
- Suppose that we are sequencing a genome from a new mammalian species for the first time and we find a gene that looks a lot like a gene in other mammals. We might reasonably impute that this new gene does a similar function in the cell.

- More similar sequences will tend to be more closely related. We can use our knowledge of sequence similarity to construct phylogenetic trees.

LECTURE 6

- Biological sequence alignment centers on orienting sequences relative to one another such that the residues which share a common ancestor are placed in the same column.
- A **pairwise alignment** is an alignment between two sequences and are useful for determining where similarities and differences lie between the aligned sequences.
- Insertions and deletions are commonly abbreviated as indels or gaps.
- There are lots of possible alignments between two sequences, so how do we decide which alignment is the best? Let's try using a scoring system to see which alignment is the best.
- k-mer frequencies provide us with a framework for assessing alignment probabilities.
- We will typically fit an alignment null model, which is under the assumption that each residue occurs independently of the rest of the alignment with some frequency q. We then fit an alignment match model which assumes that aligned residues a and b occur with a joint probability p.
- We then look at the odds ratio between the match model and the null model. If this odds ratio is greater than 1, that might suggest that the match model performs better than the null model for that particular alignment.
- We often use a log-transformation of the odds ratio to get an additive scoring system.
- We get our p and q values from a substitution matrices, which is basically a product of years of research from experts in molecular sequence alignment who measured the frequencies of each base to produce a background residue distribution for the q values as well as a residue match probability to get the p probability values.
- The BLOSUM50 Matrix is such an example
- Aligning indels is pretty tough, there are basically two schools of thought: using an indel independence model (assumes the indels are independent of each other) and an affine gap model which assumes that consecutive indels might represent a single mutation event.
- Indel independence applies a fixed penalty to each column that contains a gap, whereas the affine gap model uses a gap penalty which is a function of the length of the gap.
- Be careful: the optimal alignment does not necessarily mean that the alignment is meaningful!
- Expectation values (e-values) convert scores into useful statistics! They tell us the probability that a random sequence could produce a score at least as large as what we have observed. They are essentially a normalized score.

LECTURE 7

- There are three general types of pairwise alignments: global alignment where we try to align a whole sequence (Needleman-Wunsch, GGSEARCH), local alignment (BLAST) where we only try to align part, a Glocal (Bowtie, BWA) where we try to align one sequence to part of another sequence like a genome.
- This lecture basically goes over how to do Needleman Wunsch.

LECTURE 8

- Genome assembly is the process of ordering and merging DNA sequencing reads together into a larger contiguous sequence that represents the underlying genome from which the reads were derived.
- Why do we assemble sequences? Informatics on short sequences is hard, and for now it's the only way of getting a genome sequence. This also provides a way of assessing the thoroughness of sequencing.

- Basically we compare the ends of our reads and see where they match up. This might indicate that the reads were sampled from the same location on the genome.
- We do this for all read pairs and build long chains of reads that overlap with one another. These chains are merged together into a single long sequence that we call a **contig**.
- It is unusual to get a single contig that is the length of the whole genome. Often we have to use mate paired sequencing to order contigs into **scaffolds**
- Sequencing gaps are assembly gaps for which we know that there should be something between two contigs but we were unable to capture this sequence.
- Couple different ways to assess assembly quality: total size, number of contigs, number of scaffolds, mean contig/scaffold size, and N50. We can compare different approaches using these metrics and decide which produces the best assembly.
- Assemblies with higher N50 values are those that tend to have longer contigs and are thus presumed to be higher quality.
- Many aspects of our data might impact the quality of our assembly: read length, mate pair distance, read depth, sequencing error rates, repeat frequency and repeat size of the genome
- Several types of assembly errors: small nucleotide changes (replacing the correct base with a different ones), tandem repeat copy number changes (assembly contains more of fewer number of copies than exist in the genome because it collapsed sequences together), large scale rearrangements of genome (the assembly merges physically distinct genomic location)
- Quantifying coverage is helpful since it might tell us if we need more data to improve our ability to grown contigs during assembly.
- **Expected coverage** is commonly modeled using the poisson distribution where lambda is equal to the average number of times a base is sequenced (Usually referred to as the Lander-Waterman statistic)
- Deviations from our expected coverage might indicate biases, incorrect estimates of genome length, or assembly error

LECTURE 9

- There are three major classes of assembly algorithms: greedy assemblers, overlap-layour-consensus assemblers, and lastly de Bruijn graph assemblers.
- Greedy assemblers grow the contigs by joining the read with the best match to the end
- Pretty naïve but may be good when there are only a few sequences. Reads from repeats might throw off the assembler and we end up with contigs that do not represent the underlying genome well.
- Overlap-layout-consensus is a graph based approach by looking at sequence similarity between all reads, identifying all pairs of reads that overlap, and using this information to build an assembly graph. This is an effective approach, but it is not computationally efficient since it requires us to compare the sequence similarity of all pairs of reads.
- De Bruijn graphs use prefixes and suffixes as nodes and k-mers are the edges, so that's kind of a cool innovation. This approach is efficient and performs well! It is the most widely used assembly algorithm

LECTURE 10

- **Genome annotation** is identifying the location with a genome sequence for various features: genes, coding sequences, transcriptional regulatory elements, etc. Basically "ok now that we have an assembled genome... where is stuff located?"
- Open reading frames are regions of a genome which might encode a gene. They begin with a start codon and end with a stop codon
- This is only part of the process though. Not all ORFs might actually be genes.
- There are 6 possible ways to read a DNA sequence! Shifting your reading frame gives you three distinct ways to read, and then factoring in that there are two strands of DNAs and each fragment has a complementary but distinct sequence, we have to look for ORFs in both of those sequences

LECTURE 11

- Ab initio gene prediction (ab initio means "from the beginning") is the idea of training probabilistic models based on prior observations to predict genes in the new sequence
- Common ways to train models: identifying sets of similar types of regions as your training set (ex: regions are known start codons), align these regions, identify positions in the alignment that encode disproportionately higher frequencies across the alignment (common characters are probably important) and use this data to calculate the probability that another sequence also contains the genomic characteristics that we are interested in.
- This ab initio approach can help us find promoters which are often located upstream from a gene's start codon and affects a gene's expression
- Hidden Markov Models (HMM) are an example of one of the probabilistic models we can train.
- HMMs consider a hidden collection of states that cannot be directly observed but can be imputed from observations if we know certain probabilities.
- HMMs can be used to predict genes where the states are typically "intron" vs "exon". Exons code for things, introns do not.
- Regardless of your choice of model, it's important to carefully consider the training data you are selecting to build your model. Different genomes manifest different genomic characteristics.
- Cross-validation is a common strategy for assessing annotation accuracy. Essentially, we save parts of our data as testing data against which we evaluate the performance of our classifier, in this case, our ab initio classifier which determines if a location is a gene or not.
- Different methods will give you different results, but there are tools called combiners which merge different results into a consensus annotation.

LECTURE 12

- **Functional annotation** is the process of imputing the biological functions that a gene encodes (still a big challenge today!)
- Recall that one of the main reasons we sequence a genome is to understand which biological functions an organism can execute. Once we have identified the genes in a genome, how do we determine its function. We usually impute them using bioinformatics techniques.
- We usually use homologs to annotate gene functions. If a new gene sequence is homologous to a gene family for which we have some functional information, then it might be reasonable to infer the families functional annotation to the new homolog
- By comparing a sequence to a database of known genes using tools like BLAST, we can identify potential homologs, obtain evidence that the sequence codes a protein, obtain insight into what the protein does if that database knows the function of the homologs
- **Orthologs** are homologs which share a common ancestor through a speciation event. When genomes duplicate upon cell division, each sister gets an ortholog
- **Paralogs** are homologs that share a common ancestor through a gene duplication event
- Paralogs may not retain their ancestral function, since having paralogs which execute redundant functions in the same genome might not be helpful to the organism. Also, since there are multiple paralogs in a gene, that might give the gene more leeway for mutations since the original function will be preserved in the other copies of the gene.
- Functional annotation is a type of classification problem where we are given gene sequences which are then grouped into protein families which *are then grouped* into functional gene families.
- Sequence alignment is most frequently used to assess family membership – sequences that share significant alignment similarity will tend to share a common ancestor. Thus, the quality of our alignment plays a big role in or ability to impute a gene's biological function.
- A large number of mutations can challenge the identification of similar sequences, for example, consider distant homologs or rapid evolution. Convergent evolution can result in unrelated sequences looking similar!

- A variety of gene family databases exist and can aid in functional inference: COG, Pfma, KEGG, SiftingFamilies
- Not all proteins can be classified into a reference database family though, they might have not detectable homologs or they might be completely spurious!
- Other limits of functional annotation: protein classification can just be wrong, reference database might have errors, and functions within a family can change – other members of the family might not be consistent with your gene!
- De Novo clustering can help reveal new families which might be helpful down the road if you have a bunch of ambiguous proteins, in the hope that future work finds the function of at least one member
- Incorporating phylogenetic information into our analysis might help us ensure that we have meaningful annotations

LECTURE 13

- **Comparative genomics** involves comparing genomes with the objective of identifying similarities or differences between the genomes that might help us figure out the underlying biology or evolution of the organism under assessment.
- We might be able to discern which genes were encoded in the genome of the evolutionary ancestor of the organism that we are studying, and even impute the particular genes that were gained and lost over the evolutionary history of these organisms
- Closely related organisms can vary wildly in shared gene content – some organisms might have gained genes since splitting from a common ancestor
- Conserved sequences might be functionally important! When thinking about sequence conservation, the have to consider the underlying phylogeny relating to the organism in question. Conservation between humans and chimps might be less informative than a comparison between humans and fish
- Neutral evolution is evolution with no selection. Organism die and/or pass on genes due to random chance. It should yield a relationship between the frequency of amino acid mutations and divergence time.
- Adaptive evolution is evolution in which the organisms had to adapt to a selection factor (predation, climate change, etc.)
- The *relative rates test* allows us to determine if our data matches neutral or adaptive evolution more. But there are limitations: it requires an outgroup and assumes that molecules evolve consistently over the course of their evolution
- **synonymous mutations** are mutations that have no impact on the corresponding amino acid.
- **nonsynonymous mutations** are mutations that do impact the amino acid.
- The dN/dS test is useful here then since it is the ratio between the nonsynonymous and synonymous mutation rates. If dN/dS is high, that means there are more nonsynonymous mutations on average than synonymous ones, so there is accelerated evolution. If there is a low value of dN/dS, then there is conservation. A value of 1 means neutral evolution
- Comparative genomics often produces a list of genes: genes under selection, lineage specific gene duplications, etc.
- A question that might emerge after you get a list of interesting features for your organism is "Is a given function important to my organism?" One way to assess this is by asking: "Is a particular functional category more likely to be in my set of interesting features than not?"
- **Functional enrichment tests** group genes based on their function and quantifies whether one list contains a disproportionate abundance of a function relative to another.
- We do an enrichment test by: 1) categorizing the genes into functional groups, 2) calculating the background frequency of the functional group in the genome 3) calculating the frequency in the list of interesting genes and 4) determining if the target frequency is likely to have been randomly sampled from the background frequency using a hypergeometric distribution

LECTURE 14

- **Population genomics** is a branch of comparative genomics that involves comparing genomic features among closely related individuals, for ex: comparing the genomes of many different humans
- We usually do population genomics to find the genomic features that characterize differences between individuals in a population
- There are lots of different ways to conduct a population genomic analysis: we can try assembling each genome and then aligning them to get the orthologous positions of each genome, but this is a lot of work. We could try just using a high quality reference genome and then take sequencing reads for individuals and map them onto this reference genome.
- Individuals in a population can exhibit distinct differences in their genomes (polymorphisms). There are two main types of polymorphisms: single nucleotide polymorphisms (SNPs) which are different nucleotides that occupy the same location of the genome across individuals and Copy Number variances (CNVs) which are repetitive sequences that differ in their copy number across individuals.
- Polymorphisms stratify individuals, nucleotide differences stratify species. There might be shared polymorphisms though, locations that vary between species and individuals in that species so the difference might come from a common ancestor
- Alleles are specific variants that occupy each genomic locus. Each individual may carry up to two different alleles for any given genetic locus
- **Gene flow** is the migration of individuals into a population and can introduce new alleles or later existing allele frequencies
- **Genetic drift** are random demographic processes where some people just do not get to pass on their genes and can also impact allele frequencies
- **Natural selection** is the process by which nature biases the reproductive success of individuals based on their phenotype
- **Negative selection** is where mutations are unfavorable.
- **Positive selection** is where mutations are favorable and improve fitness

LECTURE 15

- Genetic drift is when allelic variation occurs through a stochastic process; there is in other words neutral evolution
- We can use the Wright Fisher model of populations to predict changes in allele frequencies based on random chance. This model makes heavy use of the binomial distribution.
- Drift is often used as a null model to compare against selection
- Most alleles are subject to both selection and drift; when populations are small, random sampling has a larger effect than when populations are large. When selection coefficients are small, selection has a small impact.
- **McDonald-Kreitman Test** is a test of selection that compares polymorphisms within and between species – if evolution is protein neutral, then the percentage of mutations that alter amino acids should be the same across evolution. If this percentage differs, then selection has likely occurred. We basically employ Fisher's Exact test to calculate the probability that the percentage of mutations that alter amino acids are the same between and within species.

LECTURE 16

- **Functional Genomics** is the process of determining, usually through inference, which regions of the genome encode functional elements. Through this process, we are able to learn how the genome functions and also the specific locations within the genome at which various functions are located.
- Functional elements include: exons, protein domains, and catalytic protein residues, promoters and enhancers, telomeres, centromeres, etc.

- How do we go from a bunch of As, Ts, Cs, and Gs and get a human being (or any other organism)?
- Evolutionary signatures of sequence conservation can reveal functional elements. Highly conserved sequences over evolutionary timescales may indicate function (remember, not just exons but possibly regulatory elements that affect expressions of genes)
- A **Position Probability Matrix** is a useful tool that encodes frequencies and enables location of conserved elements
- There are limits to using residue frequency to impute conservation: we assume a lot of things in using this approach (independence of positions, equiprobability of residues, no consideration of phylogenetic context, etc.)

LECTURE 17

- **Transcriptomics** is the study of the transcriptomes i.e. the total set of gene transcripts (RNA) to clarify how the genome links to the phenotype

LECTURE 18

- There are several techniques for measuring gene expression, but by far the most common in use is RNA-Seq which actually involves DNA sequencing despite what the name may suggest
- The process involves 1) Obtaining RNA samples and preparing sequencing libraries of cDNA using reverse transcriptase 2) generating RNA-Seq reads 3) Using a read mapping algorithm like Bowtie or BWQ to determine the genomic locus that expressed the RNA that generated the read 4) Quantify the gene's expression based on mapped data 5) Use statistical tools to determine if gene expression varies across study condition
- We have to use a read mapping algorithm because RNA-seq studies produces millions of reads per sample. That's a lot of data to manually align. Bowtie and BWA are very fast – they essentially compare the sequences of a read against an entire genome but they are not robust and do not handle mutations well. Additionally, they are inefficient for long reads.
- Recall that we are mapping reads to the genome for two purposes: find where a gene is located and also assess its transcript abundance
- RPKM is a normalized metric that we use to correct for confounding that might impact our abundance frequency calculations. You can compare RPKM values to other genes in your sample or across different samples and get an idea of how abundant your protein is compared to others.
- Our ability to predict a gene's expression for lowly expressed genes is relatively inaccurate because of a mean variance relationship in transcriptomic data. We may need to do more RNA-Seq reads if we are interested in studying these genes.
- The negative binomial distribution performs well with this overdispersed RNA-Seq data for calculating probabilities of different gene expressions between groups (usually we pick groups that display different phenotypes).

LECTURE 19

- **Epigenetics** are extragenomic modifications that impact genome structure and modification which subsequently alter how many genes get expressed.
- Environmental stimuli and stress can trigger epigenetic modifications and changes
- Epigenome is composed of chromatin, a complex of DNA and proteins (histones). The location of where these histones bind can have a big impact on how certain genes are expressed because if a gene is coiled around a histone, it cannot be transcribed.
- This processed is called silencing and is thought to be one of the drivers between differentiation of an eye cell and a skin cell for instance

- **ChIP-Seq** stands for chromatin immunoprecipitation sequencing and is a tool to find the loci where proteins of interest bind to
- The process basically involves 1) extracting DNA from cells 2) crosslinking proteins to DNA so they cannot come undone 3) shotgun sequence the genome 4) use immunoprecipitation to grab all the bits of DNA which are bound to our protein of interest and remove everything else
- Once we have our ChIP-Seq reads, we can analyze where the reads map to on the genome and discern the protein binding location using a read mapping algorithm
- Finding these binding locations involves using analytical techniques to find peaks in the output. This involves *tag shifting* for paired end sequencing data to combine two peaks into one, and removing spurious peaks using a background distribution
- We can quantify peak abundance using RPM which we typically want to use over a small genomic region
- We often use a control (usually running ChIP-Seq with a nonspecific antibody) to help with peak calling. The peak calling itself is often used with a Poisson distribution where lambda is calculated as the max of a variety of different calculations (background, sliding windows, etc.)

LECTURE 20

- **Microbiomes** are communities of microscopic organisms. Studying these communities has been historically difficult because not everything grows well on a petri dish.
- We can use DNA sequences as a way of finding all the organisms present in a particular community and in what abundance
- *Environmental DNA sequencing* has two main strategies: amplicon sequencing and shotgun metagenomics. Amplicon sequencing is more commonly used.
- Shotgun metagenomics essentially throws everyone together in shotgun sequencing. Very expensive and hard to analyze.
- Amplicon sequencing targets and sequences a specific genetic locus (usually a gene) from each organism's genome since this is relatively cheap.
- Amplicon sequencing makes use of PCR (polymerase chain reaction) which is essentially taking a DNA molecule, putting it in a test tube, and amplifying a specific subregion of the molecule using primers. PCR outputs a huge number of amplicon DNA molecules that are nearly identical to the region we targeted with PCR.
- We usually target the 16SSU-rRNA gene which is taxonomically distinct and has pieces which are conserved and other pieces which are hypervariable. We can use primers for the conserved bits and use the hypervariable bits to distinguish who is in our sample. We often use 16S reads to find our OTUs.
- **OTU** stands for Operational Taxonomic Unit, which essentially the equivalent of species for bacteria and archaea.
- Identifying OTUs has two different strategies: de Novo clustering and classification against a reference database.
- Clustering can idenfiying novel OTUs and is not vulnerable to errors in the reference data, but classification has the advantage of speed and avoiding spurious OTUs.
- **Alpha diversity** is the diversity within a community: often measured with richness, Shannon Entropy, or similar metrics
- **Beta diversity** is the diversity between communities: often measured with Bray-Curtis, UniFrac, or a similar metrics
- Caveats to metagenomics: PCR will fail to resolve some organisms and PCR can also introduce errors into your DNA . 16S analyses provide no direct insight into the functions encoded in the genomes of the organisms that comprise a microbiome. OUT clustering is imperfect and can create errors that over or underinflate diversity estimates.