BIOL 469: Genomics

Exam Review

Unit 1: The History of Genomics

- 1871 – Freidrich Miescher
    - discovery of "nuclein" (now **DNA**) in the nucleus
- 1910 - Albrecht Kossel
    - discovery of five nucleotide **bases**
- 1950 - Erwin Chargaff
    - DNA base pairing (A with T, and C with G)
- 1953 – Watson and Crick (& Wilkins and Franklin)
    - double helical structure of DNA
- 1961 – Nirenberg, Khorana et al
    - "**code of life**", codons and amino acid translation
    - Translate largely universal protein-coding genes into peptide sequence → beneficial in computational genomics
- 1977 – Frederick Sanger
    - develops **DNA sequencing**; sequences **phiX174 genome** (a bacteriophage)
- 1983 – Kary Mullis → PCR
- 1990 – Human Genome Project launched
- 1995 – TIGER (The Institute for Genomic Research)
    - First **bacterial genome** sequenced (H. influenza)
- 1996 – **Yeast** sequenced and "**Dolly the Sheep**" (first cloned animal)
- 1998 – C. elegans – **first multicellular sequenced organism**
- 1999 – **Human Chromosome 22** is sequenced
- 2000 – **UCSC Genome Browser**

- 2001 – **Human Genome** sequenced (technically a draft sequence with some gaps)

    - Lander et al. is the collaborative and public one; Venter et al. is more of a private one
    - **Key findings:**
        - 20,000 – 25,000 protein-coding genes (a lot less than expected)
            - Protein-Coding Genes only make up 1.5% of the genome
        - Only 7% are vertebrate-specific (most are shared with other species)
        - Thus, the human specific components are **old and few**, and much of the genome may rather encode a "**regulatory function**"
- 2002 – **Mouse genome** sequenced
    - For the first time, we can compare genome → Pioneering study for **comparative genomics**
    - **Key findings**:
        - Large-scale **synteny** (conservation of order of genomic segment) to human genome
        - Lineage-specific **duplications**
        - 40% of DNA sequence could be aligned to human
        - 5% of mammalian DNA is under **(purifying) selection**
            - Evolution has acted to conserve the sequence for a functional reason
            - This value is MORE than the protein-coding genes in genome
                - Suggests **functional non-coding features** (~3.5%)
- 2003 – Human Genome Project completed
    - There are (only) 20,000-25,000 genes (without splicing) → so, what does the rest do?
    - **ENCODE project** is launched
        - aim is to characterize all the functional elements in the human genome (not just genes)

- 2004 – "**Metagenomics**"
    - Venter et al. (2004), **environmental shotgun sequencing** of the sargasso sea

        - Large scale sequencing of the sea water (including a lot of organisms!)
- 2005 – **HapMap** → population genomics
    - Map of Human Genetic Variation
    - "Re-sequencing"
    - **SNPs** (single nucleotide poly-morphism / change)
    - Human genetic diversity
- 2005 – First successful "**GWAS**" paper published
    - Genome-wide Association Study
    - Correlate SNP values (e.g., an "A" vs "G" in specific position of genome) against prevalence of disease
    - Applied to age-related macular degeneration (AMD)
        - SNPs now explain 65% of AMD's heritability
- 2007– **Next Generation Sequencing** (NGS)
    - You can now sequence a lot of DNA fragments together by one reaction
    - Nature "Method of the Year"
- 2008 – **1,000 Genomes Project** launched (HapMap 2.0)
    - Whole genome and exome sequencing of 1,092 individuals from 14 populations by applying next gen seq
- 2008 – **Human-accelerated regions**
    - Regions of the human genome showing excessive bp changes compared to other mammals
    - What makes the human genome unique? (ie: what sequence changes a lot in us for functional reasons? → indicates **positive selection**)
- 2009 – Coarse-grained **3D structure of human genome**
    - How genome might fold inside the nucleus → how the genome folds and unfolds efficiently for regulation
        - Forms a fractal globule like structure
- 2009 – 1st analysis of **cancer genomes** (major area of genomic today)
- 2010 – Neandertal (an old and extinctic species) Genome
    - Closest to modern Eurasians
    - Gene flow from Neandertals to non-Africans

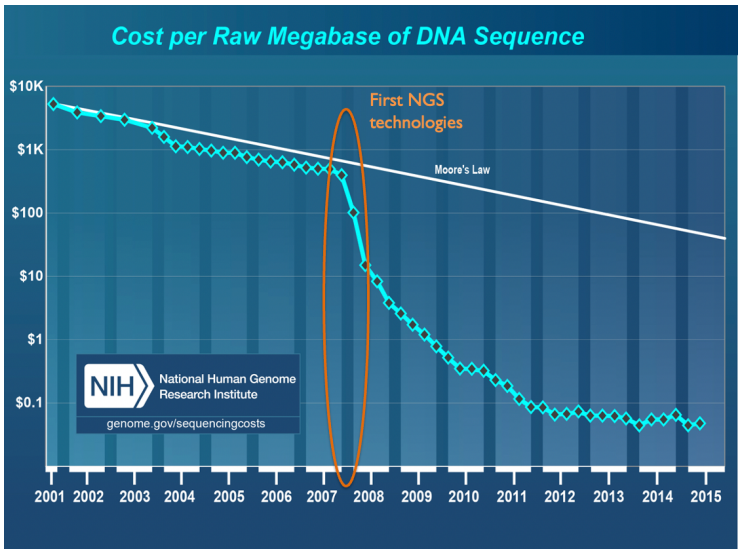
- 2010 – The first **synthetic cell** by Venter et al. 2010, Science
    - JVCI-syn1.0
    - First synthetic genome used as "software" to boot up a living cell
- 2010 – The human **gut metagenome**
    - 16S done back in 2006
    - This was the first shotgun metagenome
    - Humans are superorganisms, more bacterial than human cells
- 2012 – ENCODE study published
    - 80.4% of the genome participates in at least one biochemical RNA- and/or chromatin-associated event!
        - So, most of the genomes are in the end functional instead of junk
    - This was and still is highly controversial
- 2012 – **Personal Omics Profiling** of Health
    - Michael Snyder observed the onset of his type 2 diabetes while following a range of physiological variables.
- 2013 – genome engineering potential of **CRISPR/Cas**
    - Pioneering tool for mammalian (and beyond) genome editing
- 2014 – Several major disease genomics studies
    - Lung cancer; Schizophrenia; Ebola; Parkinson's
- 2015 – **Epigenome** Roadmap
    - focus on how DNA is regulated by epigenomic features → important for gene regulation
- 2016 – CRISPR identification of **human essential genes**
    - Without those human essential genes, you die!

- 2017 – First **CRISPR editing of human embryos**
- 2018 – First **genetically engineered babies**
    - Unregulated and done outside of scientific domain
    - Sparked major international outcry
- 2019 – 2020: Real-time **genomic epidemiology** of a viral pandemic
    - How the SARS-CoV-2 virus is mutating and spreading around the world

- In the future…
    - Billions of human genomes?
    - Synthetic genomes / genome engineering
    - Personal genomics and genome medicine
    - Real-time genomic pathogen surveillance
    - Others?

Unit 2: Genome Sequencing

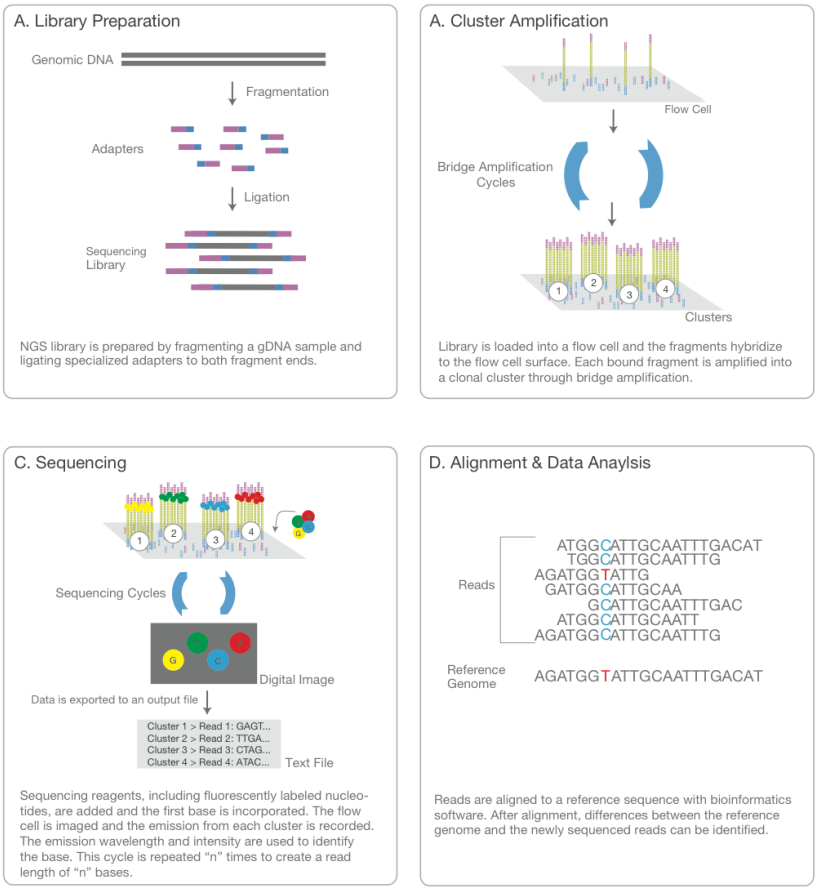
- **Next Gen Sequencing**
    - DNA sequencing costs are continuing to plummet
    - This has fueled a genomics revolution
    - Full genomes, populations, tissue samples all within one sequencing the sky is the limit...
    - Moore's Law: predicts the advances in computing technology
        - long-term trend seen in computing industry – exponential doubling of 'compute power' every two years
        - At 2007: NGS outpaces Moore's Law
            - Next-gen sequencing technology has been improving at a remarkable rate
            - This has made the many technologies both possible and widespread



    - Devices for Sequencing
        - **HiSeq**: larger scale with more data (~human genome)
        - **MiSeq**: smaller scale (~bacteria genome)
        - Nano-pore sequencing → even more portable

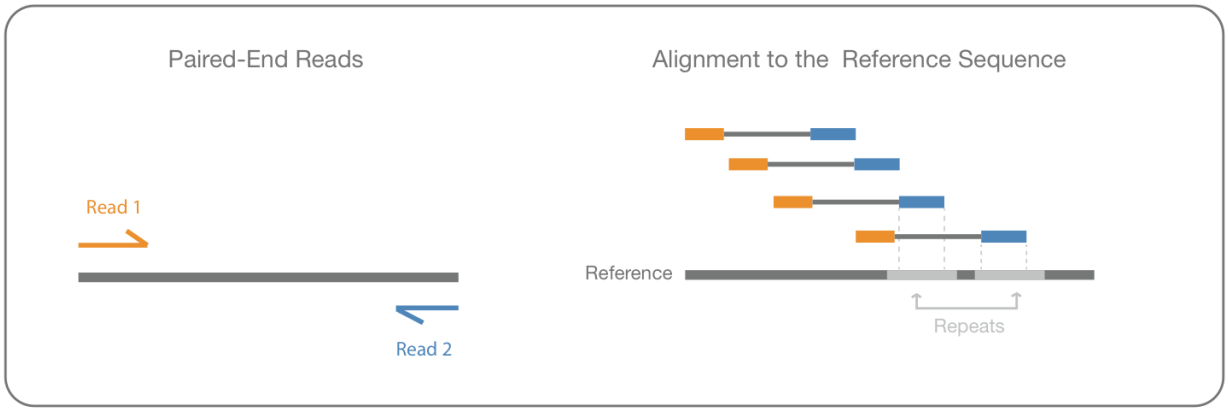
- **Illumina** next-gen sequencing
    - General idea

- DNA polymerase catalyzes incorporation of fluorescently labeled **dNTPs** (chain terminating) during DNA synthesis cycles
- During each cycle, nucleotides are identified by fluorophore excitation
- This is done in a massively parallel fashion to speed this up (million sample in one go)
    - Each cluster contains identical set of DNA sequence
    - Each cycle is creating a different length, and an enzyme cleaves of the label to allow the process to restart
    - Final file: FASTQ file

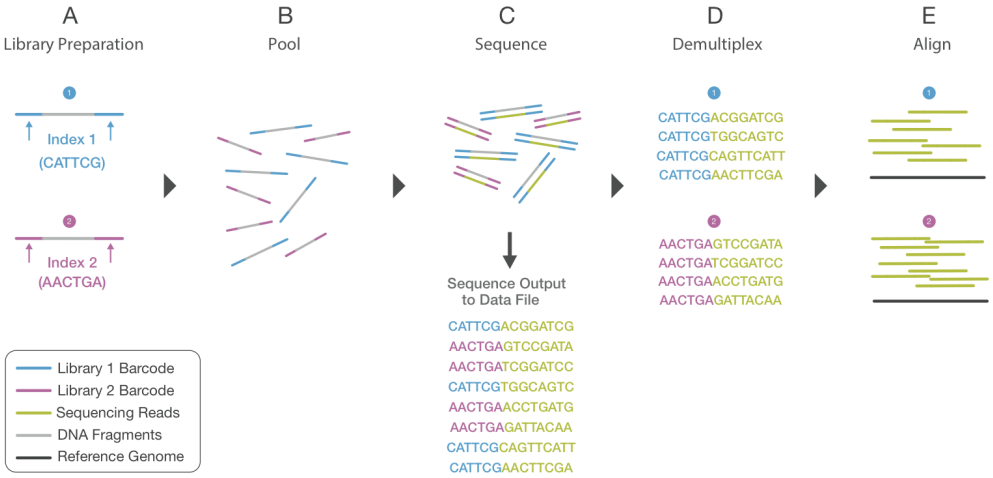


- **Paired-end Sequences**
    - Paired-end reads are sequences of both ends of a sequence fragment
        - sequence one end, then turn it around and sequence the other end

- This added information improves the accuracy to which reads can be mapped to a reference genome



- **Multiplexing**
  - Large numbers of libraries (with different DNA barcode sequences) can be pooled and sequenced simultaneously
    - Analyzing multiple samples in one sequencing run → create several libraries from different samples by attaching barcode/index to the reads (done chemically)
  - Powerful for multi-sample sequencing studies → Useful for comparative sequencing

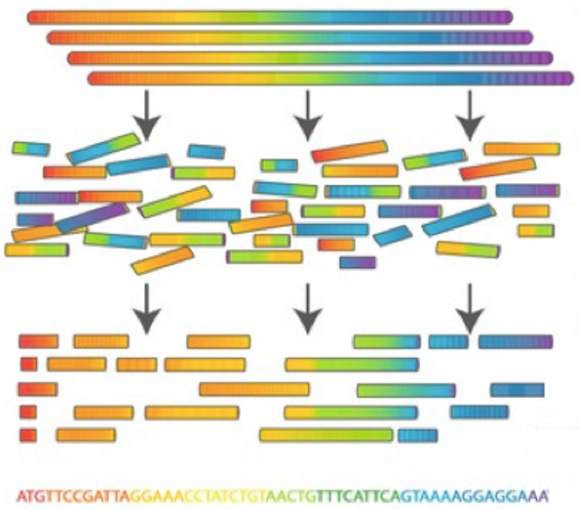


- Raw Data → Millions of Fragments / Reads
  - read lengths range approx from 50-250 bp
  - reads can be paired-end or single
  - Contained in a **.fastq file**

        - Standard for storing output of high-throughput sequencing data
        - 4 lines/sequence
            - Line 1: sequence identifier (begins with @)
            - Line 2: raw sequence calls
            - Line 3: break (begins with +)
            - Line 4: **ASCII-encoded quality score** for each call
    - Can also be in a **.fasta file**
        - Older format
        - Each sequence starts with a description (denoted by a ">" followed by the raw sequence data
    - Comparison between FASTQ and FASTA
        - FASTQ
            - generally for NGS short reads
            - essentially a FASTA with quality information
        - FASTA
            - generally for assembled sequences (contigs)
            - contain entire genomes or reference genomes
            - **.fna** vs **.faa** sometimes used for nucleotide vs aa

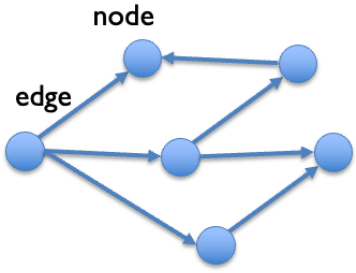
Unit 3: Genome Assembly

- **Genome Assembly**
    - We don't (yet) get entire genome sequences coming out of the sequencer
        - We get <u>fragments</u>
            - Sometimes these are <u>very short length</u> (Illumina reads are 50bp or sometimes 300 bp
    - An attempt at <u>ordering shorter sequences</u> to approximate the original sequence from which they come (by overlapping the sequences with similar fragments)



- Important Terminologies
    - **Read**
        - any sequence fragment that comes out of the sequencer
    - **k-mer**
        - A sequence of length k
        - Ex: 3-mer → sequence of 3 nucleotide in length
    - **Contig**
        - gap-less assembled sequence
        - Contig can sometimes be the entire genome (if it is really long & with no gaps)
    - **Scaffold**

        - ordering of contigs to approximate larger chromosomal sequence that may contain gaps
    - **Average Read Depth / Coverage**
        - Sum of times covered per positions / total # of positions
            - We want this value to be as high as possible
    - **% Genome Coverage**
        - % of the full genome that is covered by the assembled contigs
        - estimate the length of the complete genome to get the % genome coverage
    - ***de novo* Assembly**
        - de novo = starting from the beginning
        - assembly with no prior information / no reference genome to work with
    - **Reference-Guided Assembly**
        - Uses closely related genome to guide process
        - Align reads and contigs to reference
        - Assembly is often smaller and more fragmented than the reference genome
- **Graphs / Networks**
    - a series of **nodes** and **edges**
    - basis of most assembly algorithms
    - Represents the overlaps between reads



        - The fundamental information used to assemble genomes
    - Two types of graphs for sequencing data:
        - **overlap graphs** – used more for 454 (long read data)
        - **de Bruijn graphs** – used more for Illumina

- **The Overlap Graphs**
    - The path spells out a sequence
    - Problem – become computationally intensive when many reads are present
- **The de Bruijin Graphs**

- From an ancient math question
    - How can you go through all spots with no repeats and end up in the same place? → Solution to this is the solution to genome assembly!
- Steps
    - Find the most common k-mers from reads
    - Computer Left/Right (k-1) mer
        - Each (k-1) mer represents a node
        - Each original k-mer represents an edge
            - **Directed edge** (from L to R) represents the former k-mer, or the link between L/R k-1 mers
    - The result is a path that visits all edges exactly once → **The Eulerian Path**
        - As you move, add the last letter of the node → this way, we include the unique k-mer exactly once to reconstruct the original genome
        - Theoretically, this will reconstruct the full genome sequence (with some assumptions)
            - Assumptions:
                - we have capture all the k-mers
                - we do not have repetitive sequences less than or equal to the value of k (where there is no way to visit some paths exactly once)

- Evaluate an Assembly
    - Quality Scores to be Considered:
        - Number of contigs/scaffolds produced
            - The more the contigs, the more the genome is fragmented
        - Length of assembly
            - Is it expected based on a related genome?

        - Length of largest contig
        - % gaps (N)
        - **N50** (VERY COMMONLY USED)
            - The largest contig length at which equal-length or longer contigs cover 50% of the total assembly length
            - Cannot be used to compare the quality between different genomes (ie: e-coli vs. human) → longer genome naturally means larger N50 by chance
                - But can be a measure for quality of genome assemblers using the same reference (ie: all about e-coli assemblies)
            - If there is only one contig, N50 = length of assembly
                - We want fewer contigs with long length
        - % Coverage (when reference is available)
        - Sequencing depth
    - High quality assemblies should have…
        - Few fragments, each is very long
            - high % genome coverage (90%+)
            - high sequencing depth (10X, 50X, ...)
            - low % gaps for scaffolds
            - high N50 (# dependent on genome being assembled)

- Milestones in Genome Assembly
    - 1977: 1st complete genome – 5375 bp → Fleischmann et al., 1995: 1st free-living organism (H. influenzae) → Human genome, 2001
    - 2010, Panda Genome
        - First mammalian genome assembled using second-generation sequence
            - Illumina Genome Analyzer only
        - Assembled using SOAPdenovo

- Average read length of 52 bp (very short!)
- Generated 176-Gb usable sequence
    - 73 X coverage
- Assembled contigs cover 94% of the genome
    - remaining gaps are carnivore-specific repeats and tandem repeats
- In conclusion: we can use short reads to assemble very complex genome with great quality

Unit 4: Genome Annotation

Sequence and assembly by itself is not very useful → what does it mean & how how can we find genes and assign them functions?

- Solution: Annotation of genome to understand its function

Steps in Genome Annotation

- First, **Structural Annotation**
    - Identify genetic elements within raw genomic sequence (ex: from nb x to nb y)
    - Where are the functional elements?
- Then, **Functional Annotation**
    - Associate identified genetic elements with functions
    - What do those functional elements do?
- Sometimes we can identify the structural information, but we cannot assign it a functional information
    - label it with unknown protein or its predicted functions
- Use the UCSC Genome Browser
    - The main online portal for interaction and expliration of the human genome

Classes of Functional Elements

- **Protein-coding genes**
    - Introns, exons
- **Promoters** (usually upstream of a gene), **enhancers** (can be anywhere), and other **non-coding regulatory elements**
- **RNAs**
    - tRNAs, rRNAs, microRNAs (regulation of gene expression), siRNAs, snRNAs, exRNAs, piRNAs, long ncRNAs
- **Repetitive DNA**
    - Transposons, simple/longer repeats, etc.
    - Very prevalent in human genome

Finding Genes and Other Elements in Genomes

- Basic approach to finding genetic elements within genomes is to:

- have a pre-existing model of how these elements are supposed to look
    - **Models of Genetic Elements**
- scan raw genomic sequence with these models
- these models are stored in databases and represented as profile **Hidden Markov Models (HMMs)**

- Approaches to the prediction of location
    - ***De novo* / intrinsic approach**
        - based on a statistical model of what a gene should look like
        - Looking for the general model of the gene → more general
            - E.g., a gene-finding HMM
                - Represent basic pattern that we expect to find in a gene
                - Node → DNA characters and the codons that we expect to encounter
                - Each has a probability of what nucleotide that we are expected to see
                - Intergene model → model the sequence outside the gene → about 25% each nuecleotide
                - If we can figure out the probability of each node, we can see if our gene fits this pattern

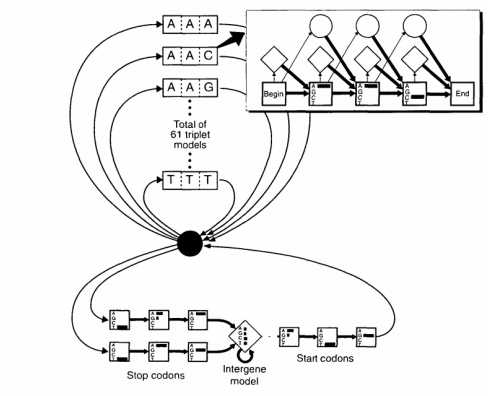
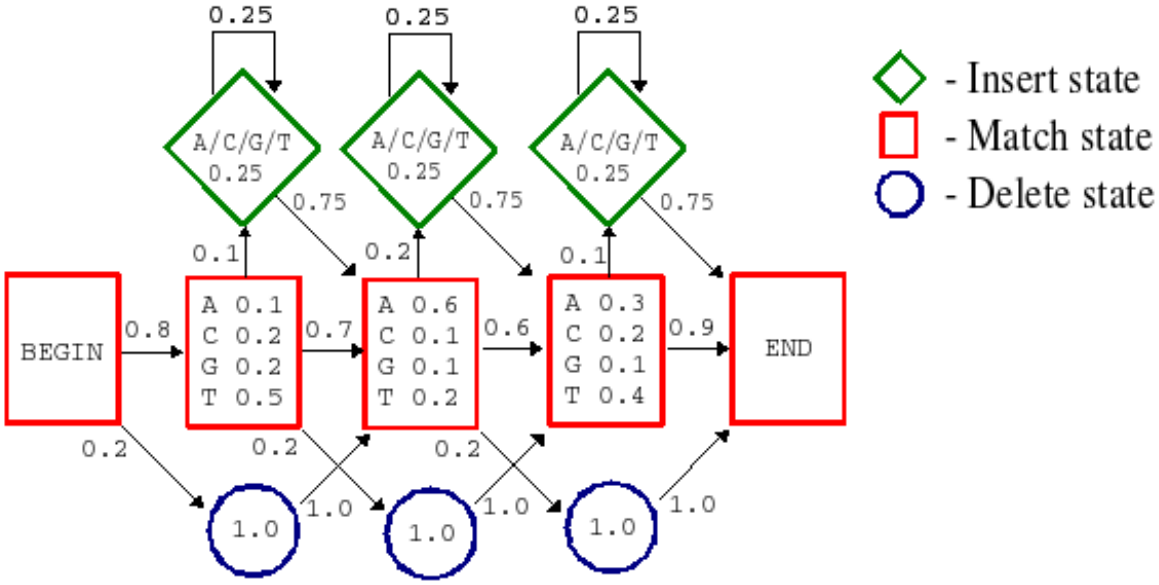


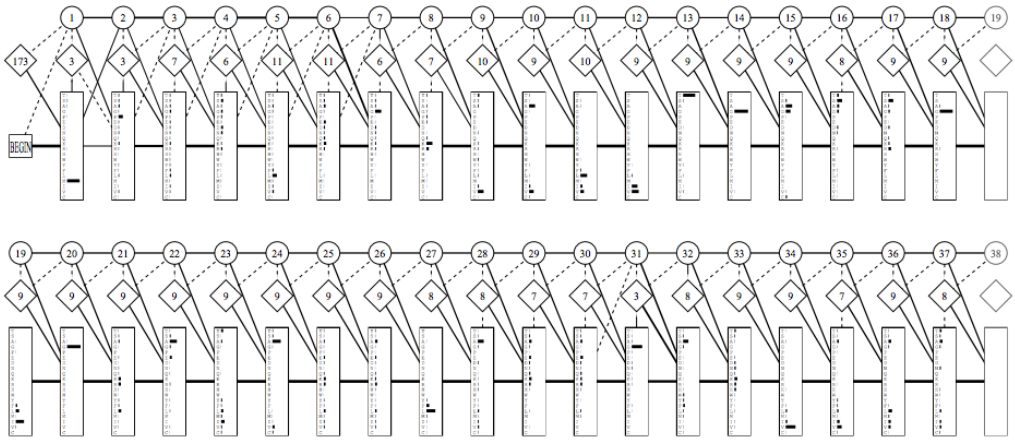
Figure 1. HMM architecture for a parser for *E.coli* DNA with a simple intergenic model. The *central state* (shaded circle), generates no nucleotides and is used to connect all the models. The 61 triplet or codon models above the central state all have identical structures, shown in detail for the codon AAC. Squares represent main states; diamonds denote a state where a nucleotide can be inserted between consecutive codon nucleotides whereas circles generate no nucleotide and can be used to delete one of the three nucleotides. The thickness of the arrows indicate the fraction of sequences making the given transition. The insert state in the middle of the intergenic model (diamond) produces random sequences from a base distribution estimated from the actual distribution of bases in the intergenic regions of the training set. The four bases have almost the same frequency.

- **Extrinsic approach**
    - external information (e.g., reference databases)
    - More specific approach

- E.g,. Detecting homology to known genes via BLAST
- E.g., a more specific HMM from sequence alignments using reference database (Profile HMM based on a DNA alignment)



- E.g., a more specific HMM from sequence alignments using reference database (Profile HMM based on a protein alignment)



- Systems for Annotating Genes / Genomes
  - Gene/protein <u>descriptions</u>
    - NCBI, UniProt

        - the description may be transferred from top BLAST match → computationally annotated genome
    - Functional terms associated with proteins
        - Gene Ontology (GO)
    - Protein and domain families
        - InterPro, TIGRFAM, EggNog, CDD, PFAM
    - RNA and DNA sequence families
        - RFAM, DFAM, RepBase
    - Metabolic pathways and organism traits
        - KEGG, BioCyc, MetaCyc, Genome Properties (predict biological traits using the known functions the gene)
        - Annotate beyond one gene

- **Xfam HMM Database**
    - Profile HMM is based on pre-computed protein (domain), RNA (non-mRNA --> functional RNA), or DNA (typical repetitive elements) family
    - PFAM, RFAM, DFAM
    - To use the HMMs
        - Individual HMMs can be searched against databases for significant matches
        - Alternatively, an entire genome/proteome can be scanned for matches to an entire database of HMMs
    - When inputting a DNA seq for PFAM → perform a six-frame translation to generation a set of protein sequence, then search using normal PFMA-A-HMMs
    - For Virus
        - Why no genes on the opposite (-) strand? → because this virus is not double stranded
        - Note the **overlapping genes** (same sequence, different protein)
            - Same DNA being used to encode different proteins
            - Clinically important because these regions may be particularly constrained and therefore good therapy targets

        - This virus condenses its genome → those overlapping regions are hard to mutate

- Match Family to Functions – GO Database
    - GO: a function vocabulary
        - Think GO as an organized list of terms that describes the biological functions of genes
    - **Gene Ontology (GO) database**
        - Collections of GO terms assigned to genes/proteins
        - Each GO term has its own identifier
    - 3 major biological aspects
        - Molecular Function, Cellular Component, Biological Process
        - Each then has its own subcategory, and so on
            - We can look specifically or at more general function → maybe there are patterns at the higher levels
    - We can now summarize predicted functions for the entire genomes
        - Look at the frequency of different high level of GO from the human and other species genome high level functional profile → gives you a more general overview
        - Here, we can compare the functional profile of different organisms
    - Other profiling methods
        - On gene families
        - On protein families (ex: using interpro)

- Assigning functions based on pathways
    - **KEGG Database**
        - Powerful for looking at metabolic pathways (reactant, product, and enzymes)
            - we can detect the presence or absence of pathways based on the

presence of proteins/genes

- Each step (reaction) in metabolism associated with protein (enzyme) sequence(s) is stored in KEGG database → **KEGG Reference Pathways**
- Given a genome, we can predict whether a reaction takes place by BLASTing all genes in that genome against KEGG
- Infer presence/absence of entire metabolic pathways

- Sometimes due to divergent evolution, part of the pathway might be present
  - it might evolve new enzymes for the rest of the functions, or just lacking half of the pathway
- Also, new (that are not in the database) metabolic pathway cannot be detected
  - Metabolic capacity is quite straightforward to predict based on detection of homology to known pathways (KEGG)
  - Novel pathways are much harder to predict
- Examples of Inferring Metabolic Potential from Genetic Information
  - **A. ferroxidans**
    - Whole-cell model for A. ferroxidans ATCC 23270
    - Known for their industrial bioleaching
    - Solubilizes copper and other metals from rocks
    - Industrial recovery of copper
    - Methods:
      - Gene modeling was performed using CRTICA and GLIMMER
      - The translated ORFs were submitted to BLAST analysis against the UNIPROT

        - these amino acid sequences were then used to query the following databases
    - **C. acetobutylicum**
        - Commercially valuable bacterium
        - Acetone-butanol-ethanol (ABE) fermentation
        - C. acetobutylicum most widely used organism
        - Renewed interest as a biofuel
        - Also active research in its use to produce solvents from diverse substrates
        - It has solventogenesis enzymes and novel cellulosome enzymes
        - By understanding how the organisms are beneficial for industry, we can modify the enzyme to make it even better
    - **Algae**
        - Lots of potential for algae in biofuel production → bioenergy potential
        - Few model strains are viable
        - Genome sequence of Nannochloropsis gaditana suggests it may be commercially useful
            - "N. gaditana has an expanded repertoire of genes involved in both TAG assembly and lipid degradation"
- Software for Genome Annotation
    - Automated Genome Annotation
        - NCBI
        - Ensembl
        - MAKER (for Eukaryotes)
        - Prokaryotes only
            - **Prokka (standalone tool) (Command-line Tool)**
            - RAST
            - JGI/DOE IMG

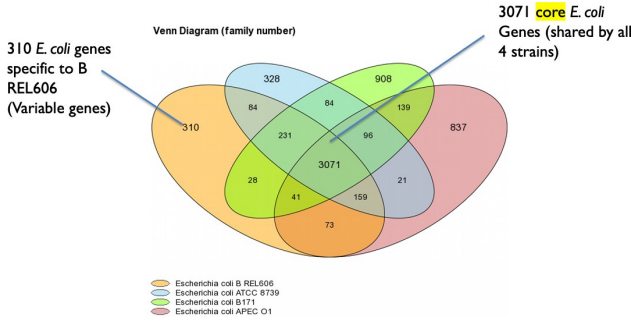
- Genome Annotation File Formats
  - .fna or .fa – fasta file
    - raw DNA sequence
  - .gbk – Genbank file
    - Genbank file containing meta-data, sequence, and annotations
  - .gff – GFF3 file containing only the annotations (coordinates relatively to .fna file)
    - No sequence
- **Prokka**
  - an automated bacterial genome annotation pipeline
    - A tool that chains multiple tools to complete a big task
  - Designed for prokaryotes
  - Starts with raw fasta file (DNA sequence)
  - Finds genes, tRNA, rRNA, and other genomic elements
  - **Fast** – annotates a 4 Mbp bacterial genome in 10 minutes on a typical quad-core computer
  - Then annotates CDSs (coding regions) by:
    - BLASTing against RefSeq and Uniprot
    - HMMscan against PFAM
  - Run it using: prokka contigs.fa
- **Maker**
  - For prokaryotic AND eukaryotic genomes
  - Identifies and masks out repeat elements
  - Aligns known **expressed sequence tags (ESTs)** and proteins to the genome → Gives you predictions of the function
  - Synthesizes these data into final annotations
  - Produces evidence-based quality values for downstream annotation management
- **NCBI Eukaryotic Genome Annotation** → there are unique challenges to euk.

Unit 5: Comparative Genomics

- Three main ways of comparing genomes
    - Comparing **gene sets**
        - E.g., Predict and compare genes, functions, pathways
    - Comparing **genome structure**
        - E.g., Identify large-scale chromosomal rearrangements (ie: between different regions)
    - Comparing **genome sequence**
        - E.g., Align entire genomes and inspect alignments for interesting patterns (sequence conservation, etc.)

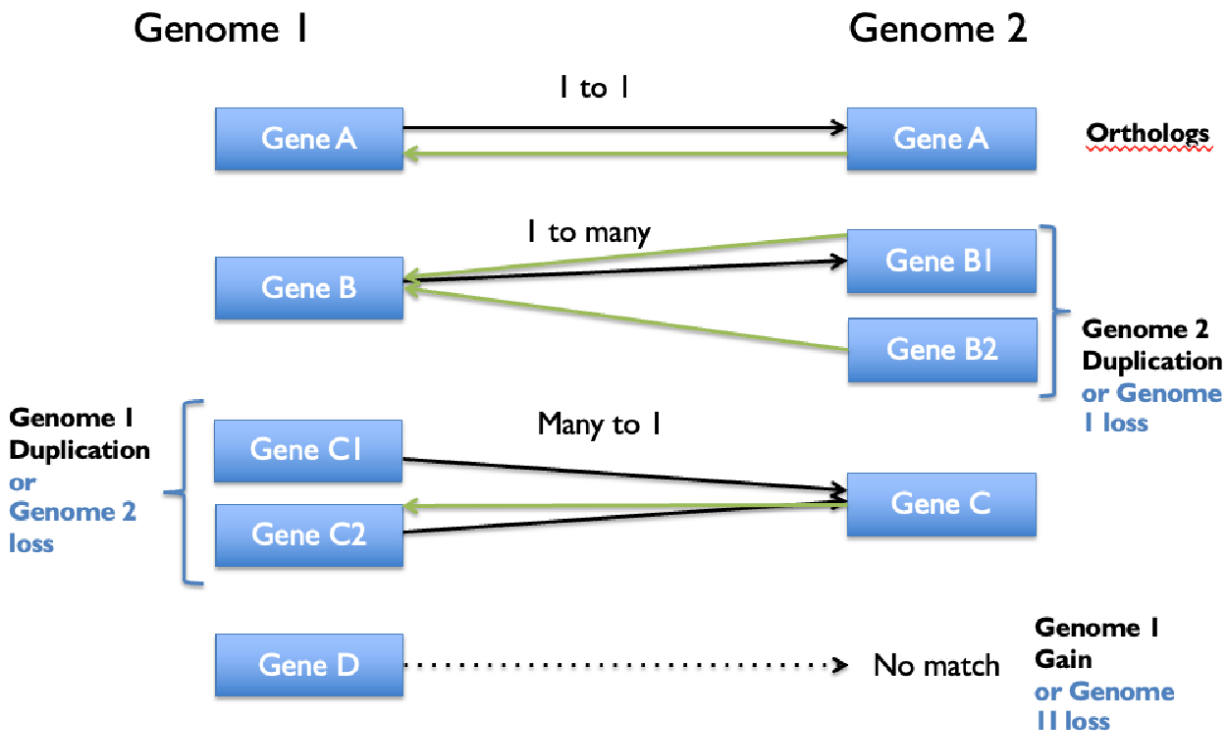
Comparing Genes Sets

- Definitions
    - Terminology more commonly applied to bacterial and archaeal genomes
        - Even closely related strains can differ widely in gene content
    - **Pan genome**
        - Full complement of genes in a group of organisms; relevant to the metagenome and ecological function
    - **Core genome**
        - Genes shared by the whole group
    - **Variable genome**
        - Genes specific to one or a subset of organisms; may encode lineage-specific (species specific) biological traits

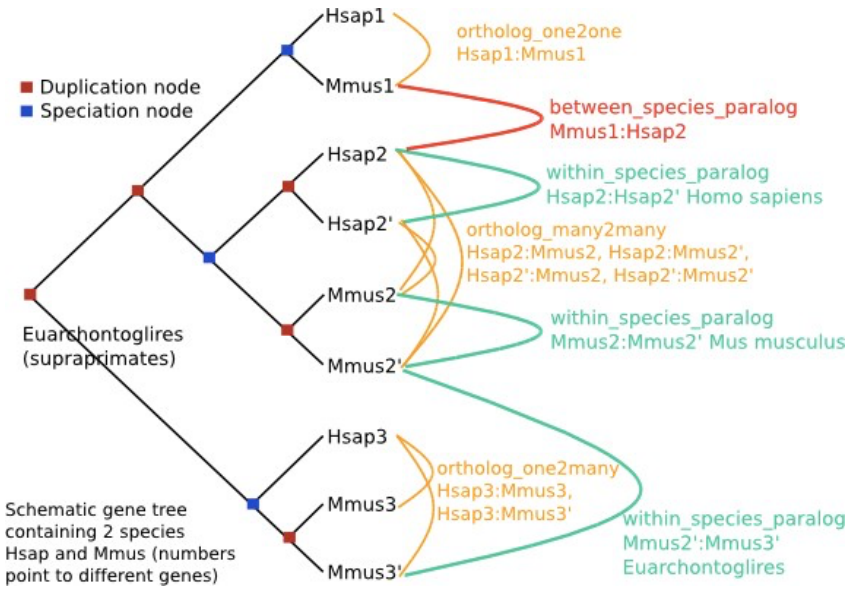


- Compare Gene Sets
    - We need to align all predicted genes from genome 1 to those from genome 2
        - Can also be used for **annotation** if one genome is considered a reference
    - This will help define:
- **Orthologs**
        - same gene in different genomes
- **Paralogs**
        - gene duplicates within a genome
- **Novel genes**
        - completely new genes with no homologs
        - May be due to **horizontal transfer** or **de novo evolution**
- Automated orthology and paralogy peplines
    - Some good ones are:
        - Ensembl compara
        - Eggnog
        - OrthoMCL
        - OrthoDB
    - A basic method: **All-by-all BLAST**
        - Ex: for 2 genomes → Genome 1 vs. Genome 2 + genome 2 vs. genome 1
        - "one to one" (reciprocal) matches are used to predict **orthologs**
        - But what about "one to many" and "many to one"?
    - Whole genome/proteome BLAST
        - One of the most commonly used commands in all of bioinformatics
            - This is the **-outfmt 6** option (to present in a table) in blast+
        - Ex: BLAST all proteins in proteome1.fa against proteome2.fa
            - blastp -query proteome1.fa -db proteome2.fa -outfmt 6

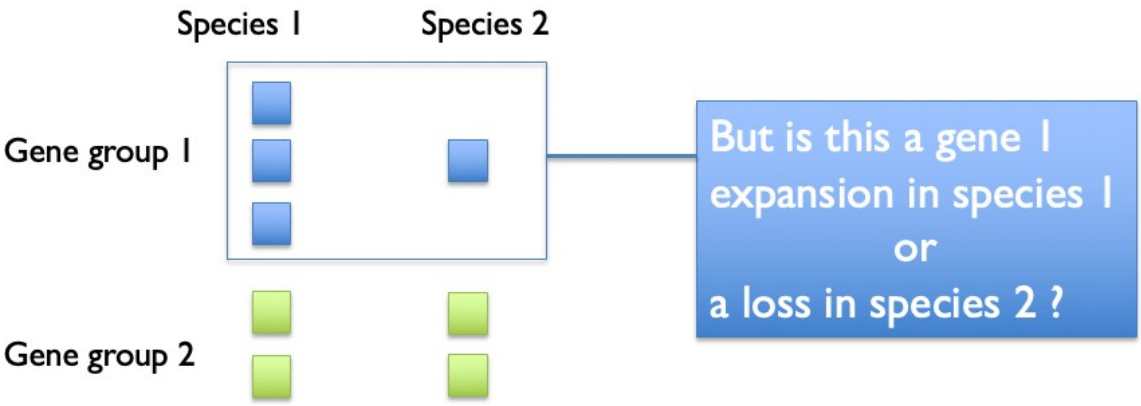
- Ortholog Mapping
    - Top reciprocal match → **ortholog**
        - Due to shared common ancestor
    - Other cases
        - B1 and B2 are **paralogs**
            - 2 mapped to 1
            - due to duplication or genome 1 loss
        - D has no detected homology
            - novel gene due to Genome 1 gain or genome 2 loss



- Tree Based Methods
    - <u>BLAST may incorrectly identify relationships</u>
        - Top BLAST hit is often the closest homolog but not always
        - Tree-based methods are more accurate
    - Orthology analysis of a gene family tree using in the **Ensembl Compara Database**
        - Pre-compute the tree to allow predictions
        - Human's Hsao1 closest phylogenetic relative is mouse Mmus1
            - **ortholog** of 1 to 1
        - 2 human genes (Hsap2 and Hsap2') next to each other (but not ortholog)
            - gene duplication (**paralog within species**)
        - 2 human genes are mostly closed to 2 mouse genes
            - **ortholog of many to many**
        - 1 human gene next to 2 mouse genes
            - ortholog of one to many (where many are within species paralog)
        - Blue node → speciation node; red nodes → duplication nodes



- Ensembl Compara Prediction types
    - **Orthologs**
        - 1-to-1
        - 1-to-many
        - many-to-many
    - **Paralogs**
        - Within-species paralogs (1-to-1)
        - Across-tree/between spicies paralogs (1:many and many:many)
        - Fragments of the same predicted gene (gene splitting → one to many parts)
- >= 3 genomes
    - Using BLAST or a tree-based method, we can group all genes from one or more genomes into orthology groups like this:



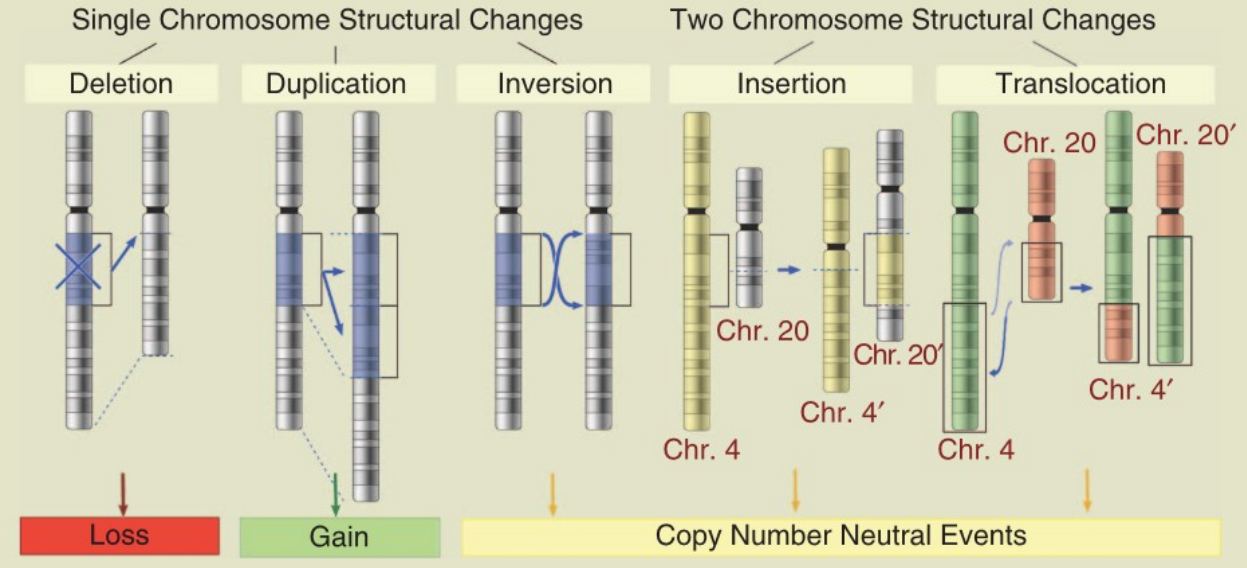
    - if this is all we have, we do not know the answer
        - yet if we bring in more genomes → we can try to figure out by looking through the homology amongst different genes
    - Example: Expansion of cytochrome P450 protein families in mouse, human and pufferfish

Functional Comparison of Gene Sets

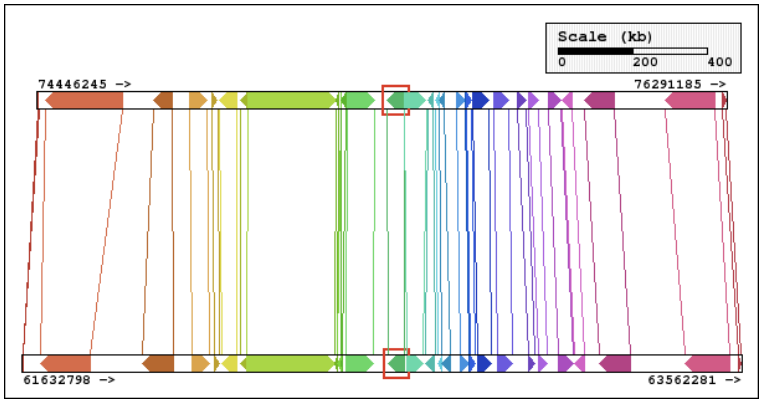
- After identifying the shared vs. linkage specific genes
    - Biologically interpret these differences. How?
        - Do the amplified/duplicated genes associate with particular functions?
        - Do the lost genes associate with particular functions?
        - Have some genes or functional categories undergone accelerated evolution?
- Functional Summaries
    - We can now tally up the GO terms, KEGG pathways, etc. for the lineage-specific gene duplications, deletions, etc.
    - Alternatively, we can compare the **total function frequencies** between organisms → **functional profiling**
        - Use Table
        - Use Circular Plot
        - Use Heat Map → Green: underrepresentation; Red: overrepresentation

Comparing Genome Structures

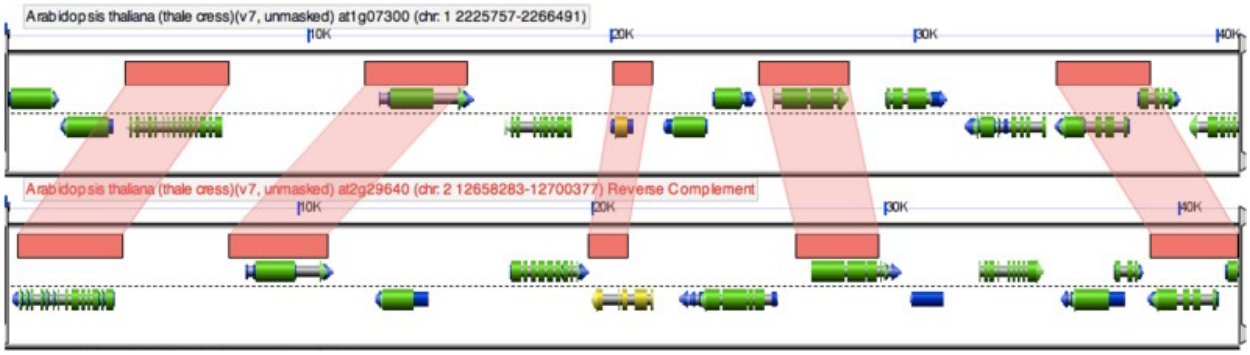
- Big Q:
    - How does one genome relate to another in terms of broader chromosomal homology?
        - Solving this problem also relates to whole-genome alignment
        - This is key to modeling genome evolution
        - Chromosomal re-arrangements (e.g., duplications and deletions) can also have adaptive/functional consequences
- Types of Chromosomal Rearrangements
    - When doing gene set comparison, those differences might not be detected
        - since in gene set, we are not looking at location



- **Synteny Analysis**
    - **Synteny**
        - conserved chromosomal blocks between species
        - Gene order largely consistent (conserved)
        - Reflects shared ancestral genome characteristics
    - Finding syntenic regions between genomes is a critical step in further whole genome alignment
    - Lining up the gene
        - same gene is coloured the same way (ie: orthologs)
        - colinear pattern → conserved 1 to 1 in gene order



- **Whole genome duplication**
  - Synteny after whole genome duplication in Arabidopsis
  - This is the alignment of chromosome 1 to chromosome 2 → line up in a highly syntenic way → 2 regions are homologs → duplication of ancestor regions



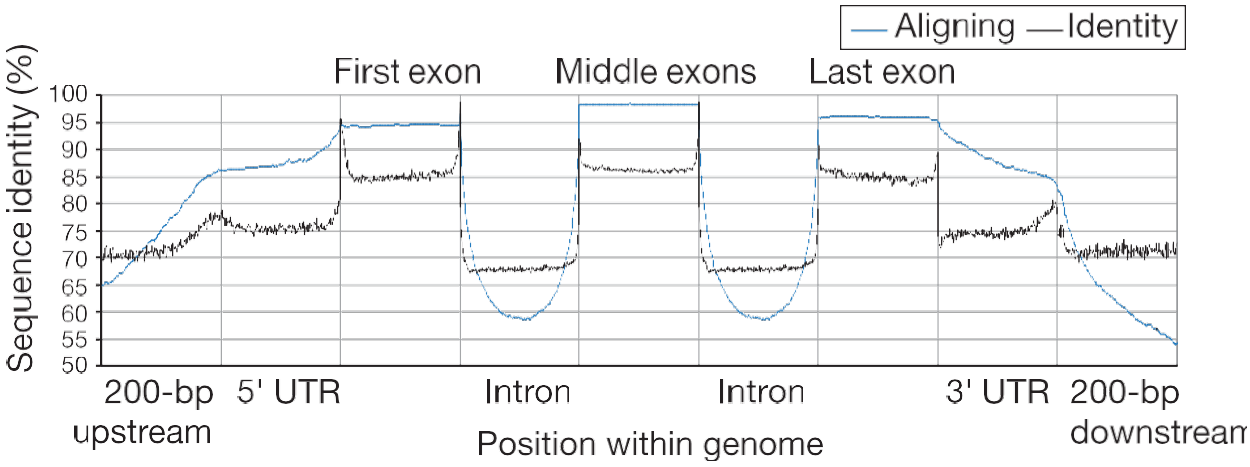
  - By finding the syntenic blocks, we also find the differences…
    - Synteny analysis also helps identify **chromosomal rearrangements**
      - Deletions, duplications, inversions, translocations
      - Greater evolution differences → more shuffle → less synteny
  - Basic synteny mapping
    - **BLAST** genome 1 against genome 2 → BLAST -outfmt 6
    - Visualize high-scoring pair matches in tools such as **Artemis / ACT**
      - Make sure that you are looking at the orientation correctly & start at the same positions
  - Advanced Synteny Mapping
    - **MAUVE**
      - Based on alignment of "**anchors**" (long ungapped matches between genomes)
      - Seed and extend
    - **Mercator**
      - Used by Ensembl

- **BRIG**
    - BLAST Ring Image Generator (BRIG)
    - Images show central reference genome with other genomes as concentric rings
        - Thus, relevant for circular (prokaryotic) genome comparisons
        - Drawback: does not show the unique insertion in the other genomes (since it is only comparing to the central ref genome)
    - Presence, absence, truncation, or sequence variation can be highlighted

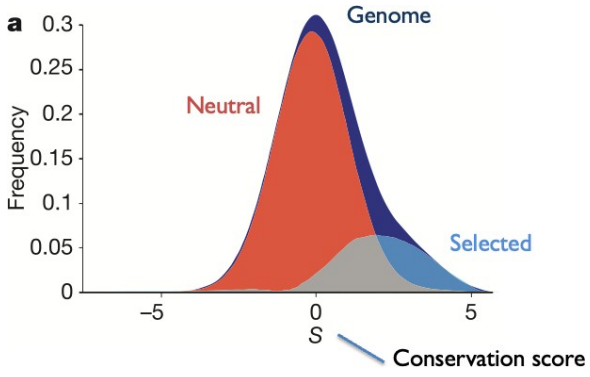
Analysis of Conservation Patterns from Mammalian Genome Alignments

- Mouse Genome
    - Mouse genome sequenced in 2002
    - Enormously important
        - Model organism for biomedical research
        - Second mammalian genome (after human), enabled sequence comparison!
        - First chance to measure genome-wide patterns of conservation patterns between two mammals
        - First chance to see which regions of the human genome are conserved and unconserved

- Conservation Levels for Different Genomic Elements
    - One of the most amazing figures in all of comparative genomics!
    - All genes compared between human and mouse summarized by <u>region-specific conservation patterns</u>
        - Conservation amongst gene structure
    - Provides a picture of the average gene and its functional constraint
        - Most conserved part: at the end and beginning of the exons → splice sites



- Proportion of the Human Genome Under Varying Levels of Sequence Conservation
    - Genome: more distribution towards **the right side** (higher conservation score)
        - since it contains a lot of sequences that are selected for



- Human Genome Conservation
    - Only 5% of the human genome appears “conserved”
        - Protein-coding genes only make up 1.5% of bases
            - Genes are not everything → there are a lot more non-coding sequences that are conserved too
    - This leaves 3.5%: functional non-coding elements that are under negative or purifying selection
- Updated and More Detailed Analysis
    - Compared between many more mammalian species
    - 4.2% of genome is putatively constrained & with ~1 million putative regulatory elements → more regulatory elements thatn there are genes
    - The functional distribution of conserved sequences depends on species
- Functions of conserved non-coding sequences
    - **Enhancers**
        - Recruit transcription factors and control the spatial and temporal expression of genes by binding cell- and tissue-specific transcription factors
        - can be very far from the gene
    - Looking for Enhancers
        - First, enhancer candidates identified as non-coding elements with extreme evolutionary conservation located far from genes
        - Then, tested through transgenic mouse assays
            - Sequence is fused to a Hsp68 reporter and LacZ reporter gene
            - Microinjected into fertilized eggs
            - Embryos harvested and stained
    - Many highly conserved non-coding elements are tissue-specific enhancers
        - Enhancer can be diverse and specific for the expression pattern

- **Ultra-conserved Elements (UCEs)**
    - Long non-coding sequences with perfect conservation in distantly relatedmammals and even other vertebrates
    - Even more conserved than protein-coding genes!
    - Enriched functions associated with Genes near UCEs → prob in gene regulation
    - Why are some enhancers so conserved?
        - Some enhancers composed of densely packed, evolutionarily conserved, **transcription factor binding sites**
            - Mutation will lead to disruption of binding
- Conservation Analysis of the Human Genome
    - 3%-8% of the human genome is conserved in vertebrates and/or other mammals
    - In all species groups, the most **highly conserved elements (HCEs)** are quite long(hundreds or thousands of bp). Less than half of these are exons.
    - Many of the HCEs are in 3'UTRs
    - Many HCEs in UTRs show evidence of local RNA structure
    - Many HCEs found in gene deserts -> these may be **distal enhancers**
- **Reconstructing Ancestral Genome**
    - Highly conserved sequences show us what DNA has been conserved sinceancient common ancestors
    - Is it possible to reconstruct ancestral sequences from extant (modern) ones?
        - Ensembl uses ORPHEUS
            - Probabilistic method to reconstruct value of each base in ancestor
            - Also handles insertions and deletions
        - From this it is possible to infer the each of each base
    - Reconstruction can also help us to annotate genomes
        - Some patterns become clearer after reconstruction (due to the removalof neutral mutations)

Comparative mammalian genomics and the genomic basis of human-specific traits

- What happened along the branch to human?
    - By only comparing to chimpanzee, we cannot know for sure
        - if we see differences, we still don't know which lineage they happened in → Is this a change in human or in chimp?
    - We need an **outgroup** to do that → we can then define the lineage in which the change occurred
- How similar are we to chimps
    - Often cited that we are 99% identical genomically
        - Not 100% accurate if you count the insertion and deletion
    - Actually 4% difference (96% identical)
        - ~35 million SNPs (substitution, 1%)
        - 90 Mb of insertions and deletions (>3%, structural variation)
        - A few chromosomal rearrangements
        - And chimps have 24 pairs of chromosomes (not 23)
            - Duplicated chromosome 2 → 2A and 2B
    - There is a huge gap between genomic and phenotypic differences
        - Each change could correspond to some phenotypic differences (but not all, since some are probably neutral)
            - But each of the change in human phenotypic trait must be somehow linked to genomes
            - ex: structural difference between human and champ brain
                - Evolution of structure of the brain → human has some unique anatomical brain structure
            - Ex: structural difference in skeletal structures for upright walking
            - Ex: morphology of the foot → human is adapted for upright walking, yet champ is adapted for gripping and climbing trees
        - How do we filter these millions of genomic differences to identify changes explaining human-specific traits?

- Look for two things
    - Gains and losses of genomic elements (genes, non-coding regulatory regions)
    - Genomic elements that have undergone **accelerated evolution**
        - Significant change compared to other regions
- Human Specific Gain and Losses
    - Several studies have examined complete loss or "**pseudogenization**" in the human genome
        - rapid degeneration and excessive function-altering mutations
        - **pseudogenes**: genes that are functional in ancestor but have decayed / acquired significant function-altering mutations in descendent
    - Deleted or pseudogenized genes in human include…
        - Olfactory receptor (OR) genes
            - Humans ORs have pseudogenized 4X faster than other lineages
                - There is a sudden increase in pseudogene accumulation in human
            - Likely due to reduced chemosensory dependence due to change in life style → you do not depend so much on chemosensory to interact with the environment
        - Keratin (hair) Genes
            - The ortholog of the gene is functional in chimp but inactivated in human → loss of body hair?
- A genome-wide screen for human-specific loss by McClean et al., 2011
    - Looked for sequences conserved in chimp and other mammals but deleted only in humans (and fixed in human population so that it is not a variant)
        - **hCONDELs** → human conserved deletions
        - Conservation -> functional importance
        - Deletion -> possible human phenotype change
    - Result: 583 human-specific deletions of conserved non-coding elements found in almost all human chromosomes

        - 510 conserved deletions were independently validated
        - Only 1 is protein-coding and 509 are non-coding (regulatory?)
            - To verify, look at functional annotation of genes near the deleted region
                - potential functions that the regulatory seq is affecting
                - result: there are some very specific stuff --> shows how the deletion might affect the structure
        - Example: human-specific enhancer loss
            - Look for functions of the deleted enhancer by using enhancer assay
            - Loss of a forebrain enhancer of the tumor suppressor gene GADD45g
            - May coincide with expansion of specific brain tissues in human
                - Increased cell growth → deletion could lead to expansion of the cortical region of the brain

- Gene Gain
    - Note: Difficulty measuring this
        - Duplication-rich regions hard to distinguish by regular short read sequencing → natural of how we assemble the reads using trees
    - **Copy number variants (CNVs)** can therefore go unnoticed
        - One strategy → look for sudden increase in coverage
    - Several studies have now looked at human-specific copy number variants
        - Compared human, chimpanzee, orangutan and gorilla
        - 53 families with increased copy number variation in the human lineage
        - Numerous gene expansions tied to brain development genes!
            - Ex: A human-specific duplicated gene plays a role in neocortical proliferation and folding!

- **Accelerated Evolution**
    - Instead of gene duplications/deletions, adaptation may also occur through modification of existing coding or non-coding sequences
    - E.g., a beneficial sequence variant may arise in an existing gene, and increase in frequency due to **positive selection**
- Measuring positive selection
    - **Ka/Ks**
        - compare **non-synonymous** (altering aa) to **synonymous** (non-altering aa, neutral change) changes (protein coding genes only)
        - Ex: Spermatogenesis protein PRM1 has high Ka/Ks (>1)
            - High Ka/Ks ratio = undergo accelerated change from to change of aa → This gene might have gone under positive selection
        - Ex: Highest Ka/Ks ratio (potentially greatest positive selection) is at epidermal and olfactory
            - This method alone cannot differentiate between beneficial selection and loss of function (since loss of function is also associated with change in aa) → it just indicates interesting functional changes
    - **Population methods**
        - nature and frequency of allele diversity within a population – will touch on this in population genomics lecture
- Ex: FOXP2 and Language
    - FOXP2 → **Transcription factor**
    - Extremely conserved in mammals
        - Yet acquired 2 substitutions along the human lineage
        - Implicated in origin of human language → might explain the unique capability of human in speech

Accelerated evolution in non-coding elements

- Find non-coding sequences that are extremely conserved in mammals but have mutated dramatically in human
- Strategy
    - Looking for very conserved non-coding regions in other mammals but not in human
    - Requiring mammalian conservation narrows search to functionally constrained regions
    - Accelerated mutation in human predicts human-specific adaptations
- These have been called **human-accelerated regions (HARs)** or **human-accelerated conserved non-coding elements (HACNS)**

Human-accelerated regions

- 49 HARs
    - Mostly non-coding
        - 66% intergenic (b/w genes); 32% intronic; 1.5% protein-coding; 0.5% UTR
- Example: HAR 1
    - The sequence that has undergone the most human-specific mutation
    - Part of a **long non-coding RNA**
    - Expressed in Cajal-Rezius **neurons** in **the developing brain**
- Example: HAR 2
    - **Intronic regulatory element**
    - Hypothesized to have "contributed to the evolution of the uniquely opposable human thumb, and possibly also modifications in the ankle or foot that allow humans to walk on two legs"