Statistical Methods for High Dimensional Biology

STAT/BIOF/GSAT 540

Course introduction + Molecular Biology Primer

Keegan Korthauer

9 January 2024

with slide contributions from Paul Pavlidis

Land acknowledgement

We respectfully recognize that the University of British Columbia Vancouver campus is located on the traditional, ancestral, and unceded territory of the xʷməθkʷəy̓əm (Musqueam) people.

You are invited to take a moment to learn about the territory you are occupying by visiting this interactive indigenous land map.

Today's topics

- What the course is about
- Course mechanics overview – full details on course website: https://stat540-ubc.github.io/
- A [fast-paced!] primer on molecular biology/genetics/genomics

Teaching team

Instructors

Keegan Korthauer

Yongjin Park

Teaching Assistants

Asfar Lathif

Ishika Luthra

Plus several guest lecturers!

Course audience

- Researchers who want to know how to analyze large data sets from biological studies
- Genomics-oriented, but covers many broadly-applicable statistical approaches
- Quick poll: your peers have varied backgrounds
  - Statistics students might find the math parts easy
  - Biology students might find the biology easy
  - We are counting on you to help make it work: help your peers!
- Formal poll due Sunday: https://canvas.ubc.ca/courses/132914/quizzes/705378

Prerequisites

Officially, none. But:

- **Statistics** – University level "Statistics 101". Prepare to get comfortable thinking about things like "probabilities" and "significance"
- **Biology** – **No requirements**, but you are expected to learn things like the difference between a DNA and RNA and a gene and a genome. We assume you are here because you are interested in biology and will pick it up as we go.
- No **R** or programming experience required but you must be prepared to do a lot of self-guided learning. *If you are completely new to R and have no other programming experience, be ready for a challenge.*
  - You'll use your own computer to run R/Git.
- Extra resources posted on Syllabus page

What you can expect to learn

- Apply tailored statistical methods to answer questions using high dimensional biological data
  - Generally applicable statistical approaches and principles
  - Specifics about some data types (esp. expression profiling)
  - Practical experience using the R/Bioconductor computing environment
- Critically evaluate analyses in the literature and avoid pitfalls in your own research
- Work with real data in a collaborative model
- Make your work reproducible, reusable, and shareable
  - Use Git/GitHub for version control
- Detailed list of topics: https://stat540-ubc.github.io/lectures
- What we don't cover:
  - Limited details on "low-level" processing
  - Limited details of underlying mathematical theory

Course mechanics

Course web site

http://stat540-ubc.github.io

- Syllabus
- Assignment submission instructions & due dates
- Rubrics
- Lecture notes
- Contact info for Instructors & TA
- Office hours information

Site is built/hosted with GitHub; all source files are stored in repositories in the STAT540-UBC GitHub Organization:

https://github.com/STAT540-UBC

Canvas site

- https://canvas.ubc.ca/courses/132914/
- Announcements
- Assignments
  - GitHub invite links
  - Gradescope submission tool
- Discussion board (Piazza)
- Grades

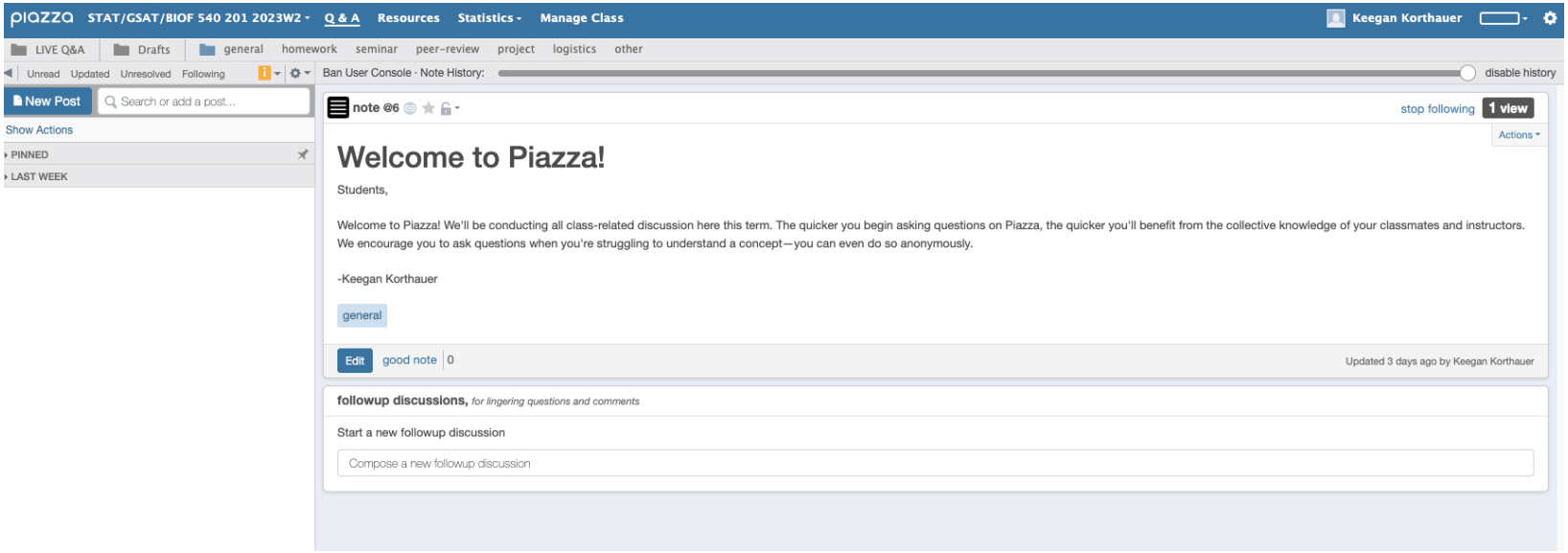
GitHub

Primary tool for coursework

- Individual (private) student repositories for each assignment
- Shared team repositories for project work

*Seminar 1 will focus on getting you up and running with **GitHub** and walks through assignment submission process with **Gradescope***

Course Communication: Piazza



Join through Canvas

- All questions about course material / assignments
- Only use email for sensitive / private matters

Lectures

- Schedule of topics/instructors: https://stat540-ubc.github.io/lectures
- Start at 9:00 AM, end at 10:20 AM
- Slides and other resources will be posted on the website as we go

Seminars (Computing Labs)

- Schedule & material (please read before attending each session): https://stat540-ubc.github.io/seminars
- Tuesdays 12:30 PM -1:30 PM
- TA-led exercises using Git, GitHub, R/RStudio
- You will follow along on your own computer
- All sessions have short 'deliverables' that you submit for a grade

First session TODAY

- Git/GitHub setup
- Demo how to submit assignments in this course
- Completion deliverable due Friday

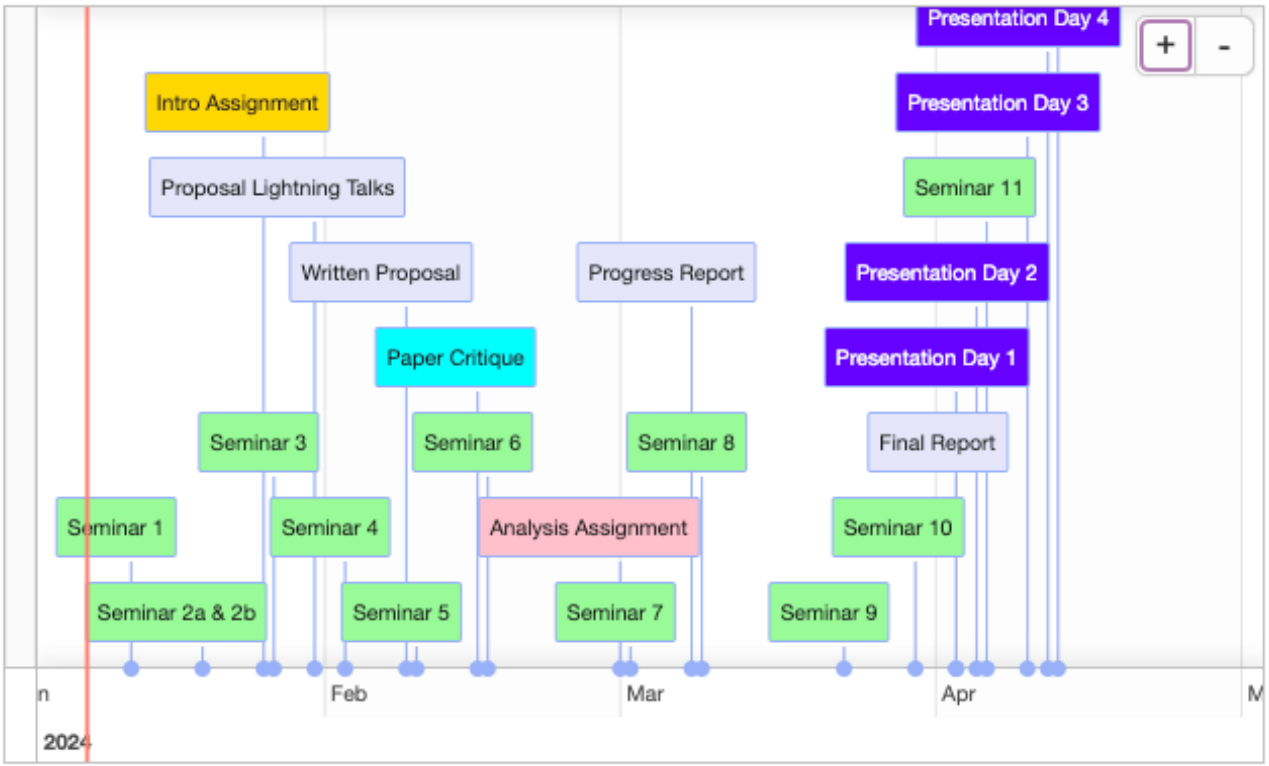
Evaluation

- **Seminar Participation (20%)**
  - 11 deliverables; keep10 highest scores each worth 2%
- **Intro Assignment (5%)**
- **Paper critique (5%)**
  - Write a short review summarizing and critiquing a paper
- **Analysis Assignment (30%)**
  - Application of techniques learned in class to a data analysis
- **Group project (40%)**
  - Proposal lightning talks (5 pts)
  - Written proposal (5 pts)
  - Progress report (5 pts)
  - Final report & repository (10 pts)
  - Oral presentation (10 pts)
  - Individual report (5 pts)

**These slides are just an overview - more detail on the website*

Key deadlines

See [course website]() or canvas for due dates



https://stat540-ubc.github.io/timeline

Late policy for assignments

- Deadlines are all by 11:59 pm (Pacific time) on the due date
- Any submission or modification after the due date **will not be graded** unless you have requested an extension
- If you anticipate having trouble meeting a deadline and need to request an extension/academic concession please reach out via email in advance

Academic Integrity

https://stat540-ubc.github.io/syllabus.html#academic-integrity

- **Do your own work**. All individual work that you submit should be completed by you and submitted by you. Do not receive or share completed coursework with students who take the course in another term.

- **Acknowledge others' ideas**. Scholars build on the work of others, and give credit accordingly. This refers to both outside sources, such as from the literature or *AI tools*, and inside sources, such as from your peers.

- **Learn to avoid unintentional plagiarism.** Visit the Learning Commons' guide to academic integrity to help you organize your writing as well as understand how to prevent unintentional plagiarism.

Privacy

https://stat540-ubc.github.io/syllabus.html#privacy

- GitHub does not have servers in Canada
- Use caution in sharing personal information

Submitting assignments

- You will use GitHub and Gradescope to submit all assignments
  - A private (only for you and the Instructors/TAs) GitHub repository will be set up for each individual assignment
  - A shared team GitHub repository (only accessible to your teammates and the Instructors/TAs) will also be set up for your project work
  - You will submit each assignment repository on Gradescope (via Canvas)
- Seminar 1 (today) will guide you through your first submission **due Friday**
- Detailed instructions also on the website: https://stat540-ubc.github.io/submission_guide

Group projects

- Begins **today** – start thinking about it! – and continues through the rest of the term
- Groups with diversity of skill sets will be formed by the teaching team following Canvas survey results (**3-4 people per group**)
- Several deliverable checkpoints throughout the term
- Overview here: https://stat540-ubc.github.io/assignments.html - final-group-project-40
- Details here: https://stat540-ubc.github.io/group_project_rubrics

**Auditors do not participate in group projects*

Group projects: where do they come from?

- Many projects are based on a data set provided by a student (i.e., collected in their lab)
- Others use publicly available / published data

Examples topics of past group project topics

- Genomic copy number alterations for prognosis of prostate cancer
- Effects of mutations in histone modifying enzymes on gene expression profiles
- Modeling time-course expression of SET domain-containing genes in mouse embryos
- Gene expression in blood of humans with asthma challenged with allergen
- Characterization of placental DNA methylation profiles in preeclampsia
- Transcriptome analysis on airway epithelium exposed to traffic-related air pollution in hyper-responsive individuals
- Assessment of Differential Gene Expression in COVID-19 Infections
- MPRA Analysis of variants at the TNFAIP3 locus
- Correcting cell clustering in human embryo single cell RNAseq
- **See Canvas for example final presentation slides from last year**

Common problems with projects

- Not high-dimensional
- Research question unfocused
- Data aren't freely available
- Data requires extensive preprocessing work (not considered part of course project scope)
- Ambition mismatched to effort available (in either direction)
- Too little/too much signal
- Only reproducing an analysis from previous literature

- Want feedback? Ask us as soon as possible!

Let's get into it (today and lecture 2)

- Some biology
  - Primer for newbies (Mol. Bio. 101)
  - Hopefully some useful information for all
- Goal: understand where data is coming from

Biology self-learning resources

Free textbooks from Rice University (OpenStax)

- "Concepts of Biology": https://openstax.org/books/concepts-biology/pages/1-introduction
  - Chapters 6 and 9
- "Biology" (more in depth): https://openstax.org/books/biology-2e/pages/1-introduction
  - Chapters 14, 15, 16
- Some images in these slides come from these
- Warning: While overall high quality, they are not error-free

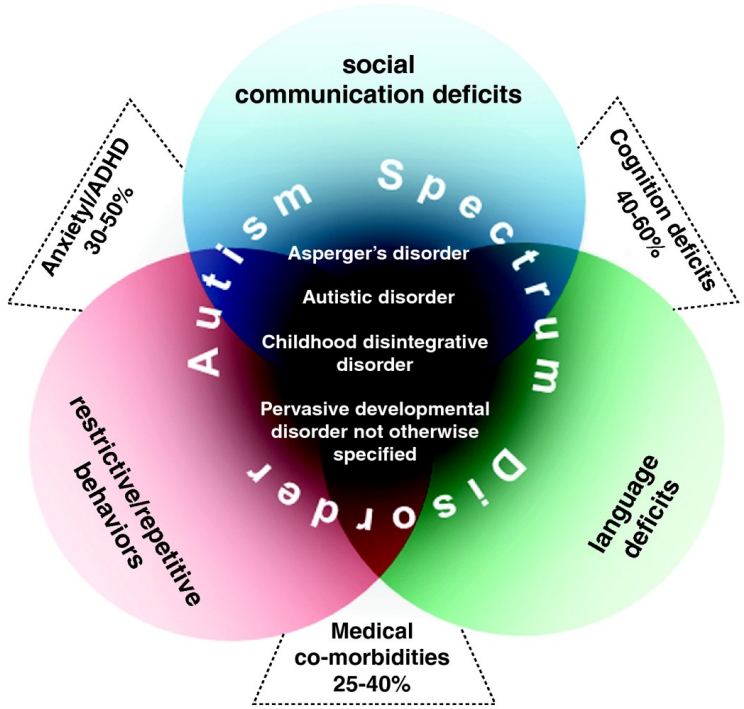
Why study genomics and genetics?

- An organism is built from many "moving parts" (e.g. proteins encoded by genes)
- Much of these have poorly understood function
- Hugely complex interactions and regulation – poorly understood
- Genetic variation and environment interact in complex ways
- Many diseases/conditions/traits have a genetic/genomic component still to be understood.

Definitions: Genetics, genomics, epigenomics

- **Genetics**: Study of DNA sequence variation
  - e.g. Association of specific DNA sequence differences with **phenotype** differences.
- **Genomics**: Study of how genomes function
  - e.g. What genes they contain, how they work, how they evolved ('comparative genomics'), how they are regulated.
- **Epigenetics/epigenomics**: Multiple definitions, but in this course we usually mean "factors that impinge on regulating genome function other than the DNA sequence itself" such as chromatin state.

Motivating example:
Biology of Autism Spectrum Disorder (ASD)



Available online at www.sciencedirect.com
SciVerse ScienceDirect
Current Opinion in Genetics & Development

Genetic architecture in autism spectrum disorder

Bernie Devlin[1] and Stephen W Scherer[2]

- Major genetic contribution
- ~1/100 children; ~¾ are boys.
- Genetically very heterogeneous
- No adequate (general) explanation nor animal models – because ASD is not just one condition

Two very broad types

- "Severe" typically nonverbal and often accompanied by intellectual disability, challenges living independently (<25%)
- "Mild" or "High-functioning" characterized by personality differences affecting social interactions (>75%)

https://autismcanada.org/
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4650984/

How can we study ASD biology?

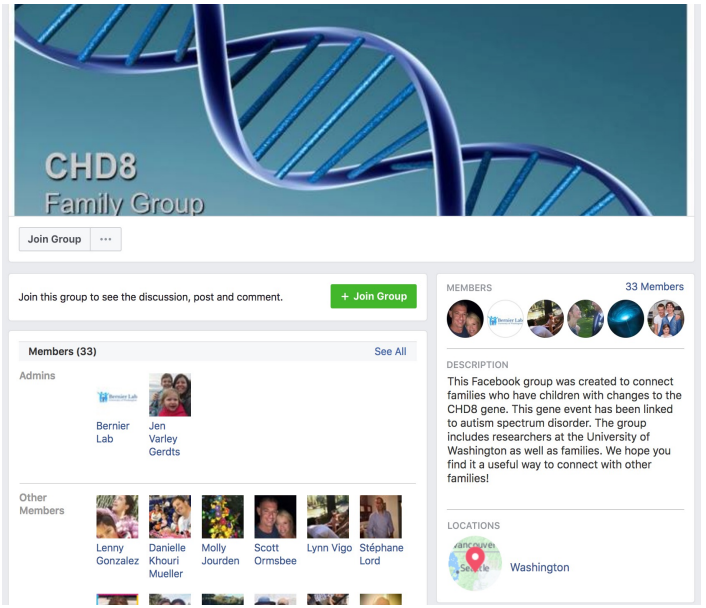
- We can look for "differences" (genetic, cellular, molecular …)
- "Reductionist" – ask questions at level of small units of biology, and build up (hopefully)
- Much more, but for our purposes this is enough

It will help to dig into a specific example ...

Our example gene: CHD8

- Individuals with certain rare mutations in CHD8 have autism
- CHD8 is the *official symbol* for the gene. Its full name is "Chromodomain-helicase-DNA-binding protein 8"
- It is a "DNA helicase that acts as a chromatin remodeling factor and regulates transcription …" http://www.uniprot.org/uniprot/Q9HCK8
- Even more info: https://www.ncbi.nlm.nih.gov/gene/57680

There is a CHD1, CHD2 … CHD9 and they have their own stories – but not necessarily for ASD



https://www.facebook.com/groups/chd8family/

Goal for rest of today: Provide some biology context, background and details that connect to the themes of STAT540

Molecular biology in one slide

Ignoring many exceptions and complications! (More details on next slides)

- **DNA**: linear arrangement (polymer) of nucleotides ('bases'), contains information to construct the organism (≈recipe, not a blueprint); provides mechanism of heritability
  - Every time a cell divides, the DNA is copied **(replicated)**
- **Gene**: a stretch of DNA that is transcribed into a functional RNA
  - Simplest organisms have ~1000 genes (e.g.: E. coli has ~4000, yeast has ~6000)
  - Most multicellular organisms have ~15-25k protein-coding genes (e.g. worms, flies, vertebrates)
  - Genes are turned on and off – not all genes are "**expressed**" in any given cell.
- **Genome**: the full complement of DNA in one cell
  - The sequence of the DNA is the **genotype**; can refer to a specific place ("locus") or overall.
  - The properties of the organism produced is the **phenotype**
- **RNA**: Immediate read-out of a gene, also made of nucleotides ("transcript")
  - Complication: Splicing. Primary transcript is of exon and intron regions, latter are removed; "**exome**" is the set of all exons. A single processed transcript is a messenger or mRNA.
  - Collection of RNAs (e.g. found in a cell; or made by a genome) is a "**transcriptome**"
- **Protein**: Major working parts of cells, encoded by genes (via RNA) and made ("translated") by the ribosome (a big molecular machine)
  - Proteins are strings of amino acids ("polypeptides"); 3 nucleotides code one AA
- DNA, RNA and protein (plus many other types of molecules used by cells) are produced using chemicals (from air and food) and energy from sunlight (directly or indirectly via food) = "**metabolism**", a process which involves the function of (at least) hundreds of genes.

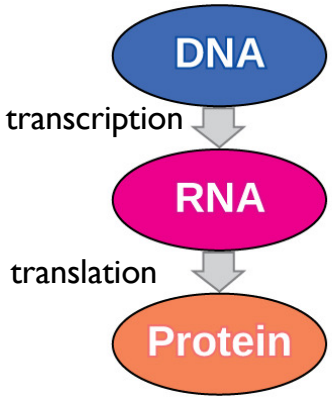
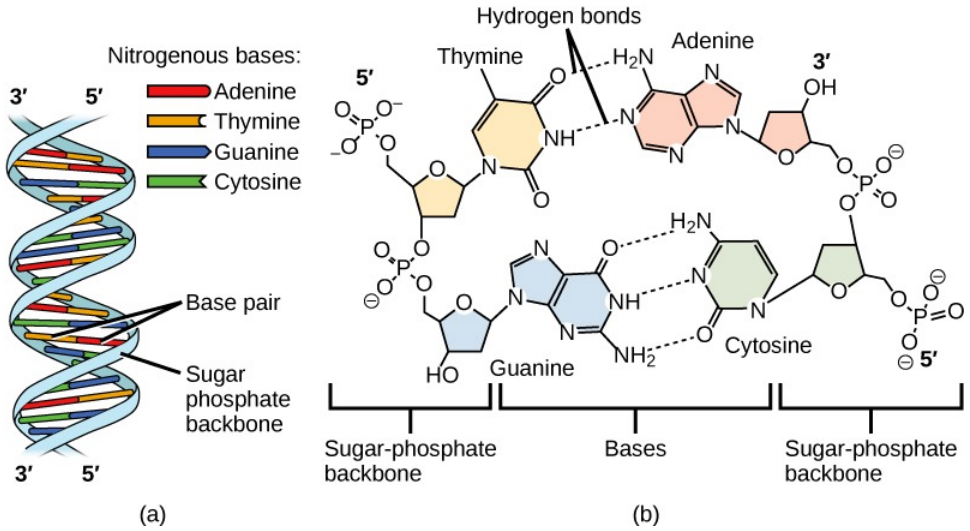


Figure 9.14, Concepts of Biology

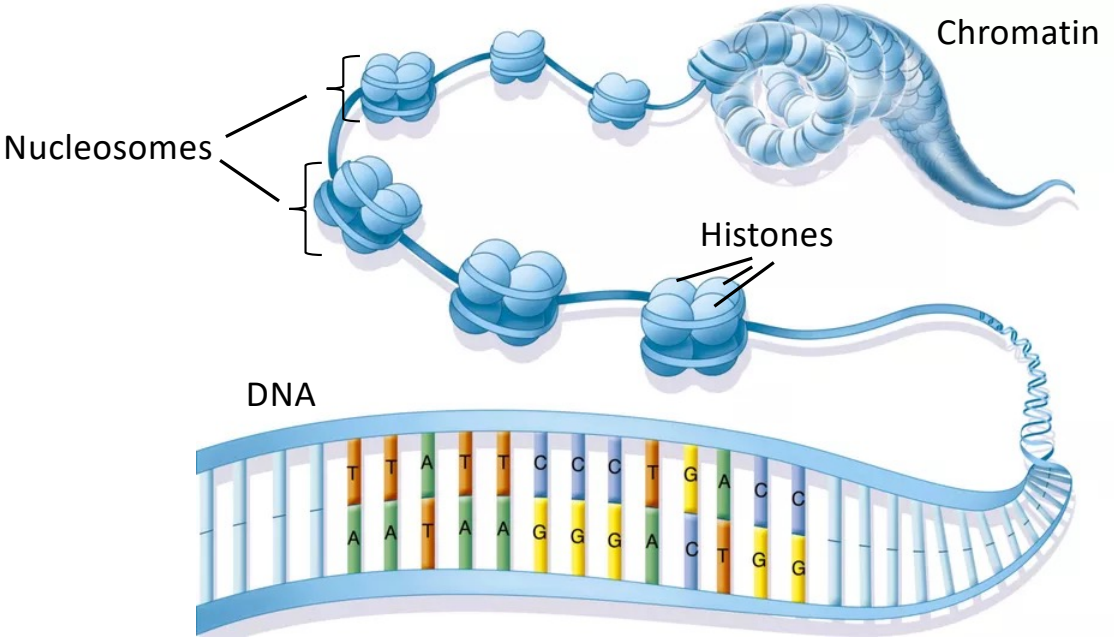
DNA: deoxyribonucleic acid

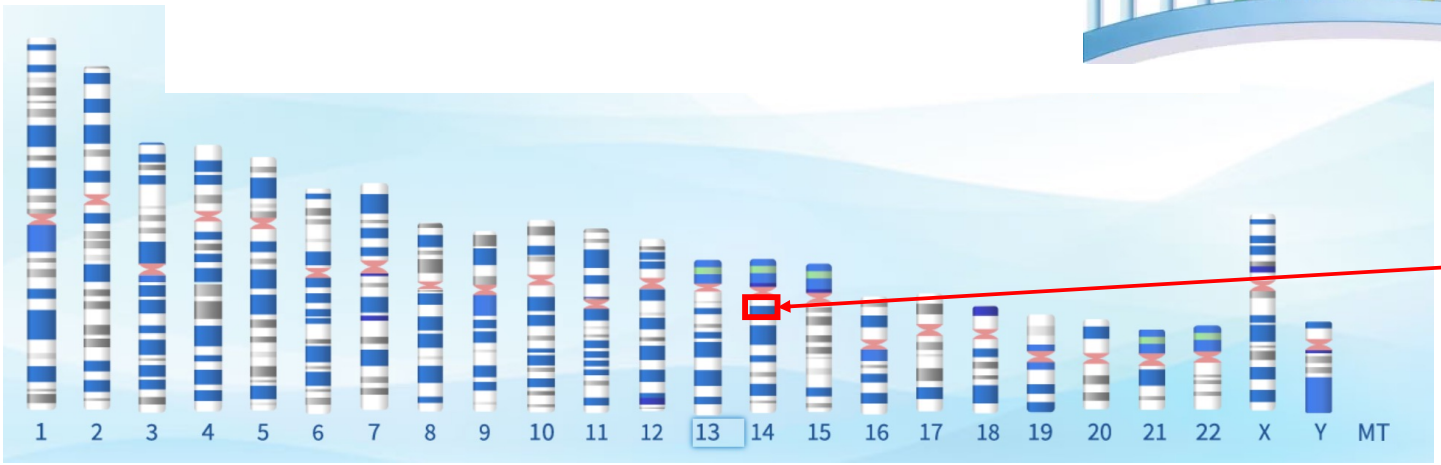


- DNA double helix consists of two strings/strands (that may separate) of bases (A, T, G, C)
  - The two strands are "complementary" and are in "opposite directions" (5' phosphate and 3' hydroxyl ends)
- A chromosome is a long piece of DNA that contains many genes
- A gene is a stretch of DNA (within a chromosome) that gets transcribed into RNA
- RNA is chemically very similar except: "U" (Uracil) instead of T; and generally we think of it as single-stranded (though it can form double helices, esp. intramolecular)

OpenStax

A genome (human)





NCBI https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml

Content of the human genome

(A typical mammalian genome)

- ~6 billion bases (Gigabases) across 46 chromosomes (3 billion "haploid")
- Of the total 46 chromosomes, 23 from mom and 23 from dad
- ~2% Exons of genes (~20,000 protein-coding genes) (~1% protein coding) = "**Exome**"; includes non-coding RNAs such as tRNAs, miRNA.
- ~20% is genes if you count introns
- 50-70% is repeats
  - Simple repeats 3% e.g. TATATATA…
  - More complex repeats e.g. LINEs, SINEs >45% (origins as transposable elements; almost all are inactive)
- Pseudogenes (~15,000) – broken copies or relics – NOT functional genes

How much of the human genome is functional?

Depends on how you define "function": ~8-15% is functional based on conservation, genetic load etc

- e.g. 8-9% is under negative selective constraint (Rands et al. 2014)
- Rest (~90%) is *non-coding or non-regulatory DNA* (historically called *junk DNA*) - basic definition: evolving "nearly neutrally"

Alternative definitions like "does something" lead to higher estimates (i.e. 100%) but can be problematic (akin to "heart's function is to make a thumping noise")

- ENCODE project: >80% human DNA has some biochemical activity (not necessarily resulting in biological *function*)
- The null hypothesis in any analysis of a sequence is that it is "non–functional"

Rands et al. : http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004525

For even more information on "Junk" see e.g. Palazzo and Gregory (2014) *The case for junk DNA* in PLOS Genetics; Doolittle et al. (2014) *Distinguishing between "Function" and "Effect" in Genome Biology* in Genome Biology and Evolution

Implications of genome content

- Only about 1% of the human genome is protein coding; another ~1% are non-protein coding genes = 60 million bases (out of 3 billion haploid)
- Assuming 10%* has biological function, another ~8% is therefore apparently "regulatory": ~5-10x the amount of protein coding! (>300 million bases)
- The other 90% is not important** (~2.7 billion bases)
- Figuring out which 10% is important is still a big challenge

* Estimates vary; these are "ballpark" values. Other mammalian genomes are similar; but other organisms can have more compact genomes – e.g. bacteria that tend to have nearly100% functional genomes.

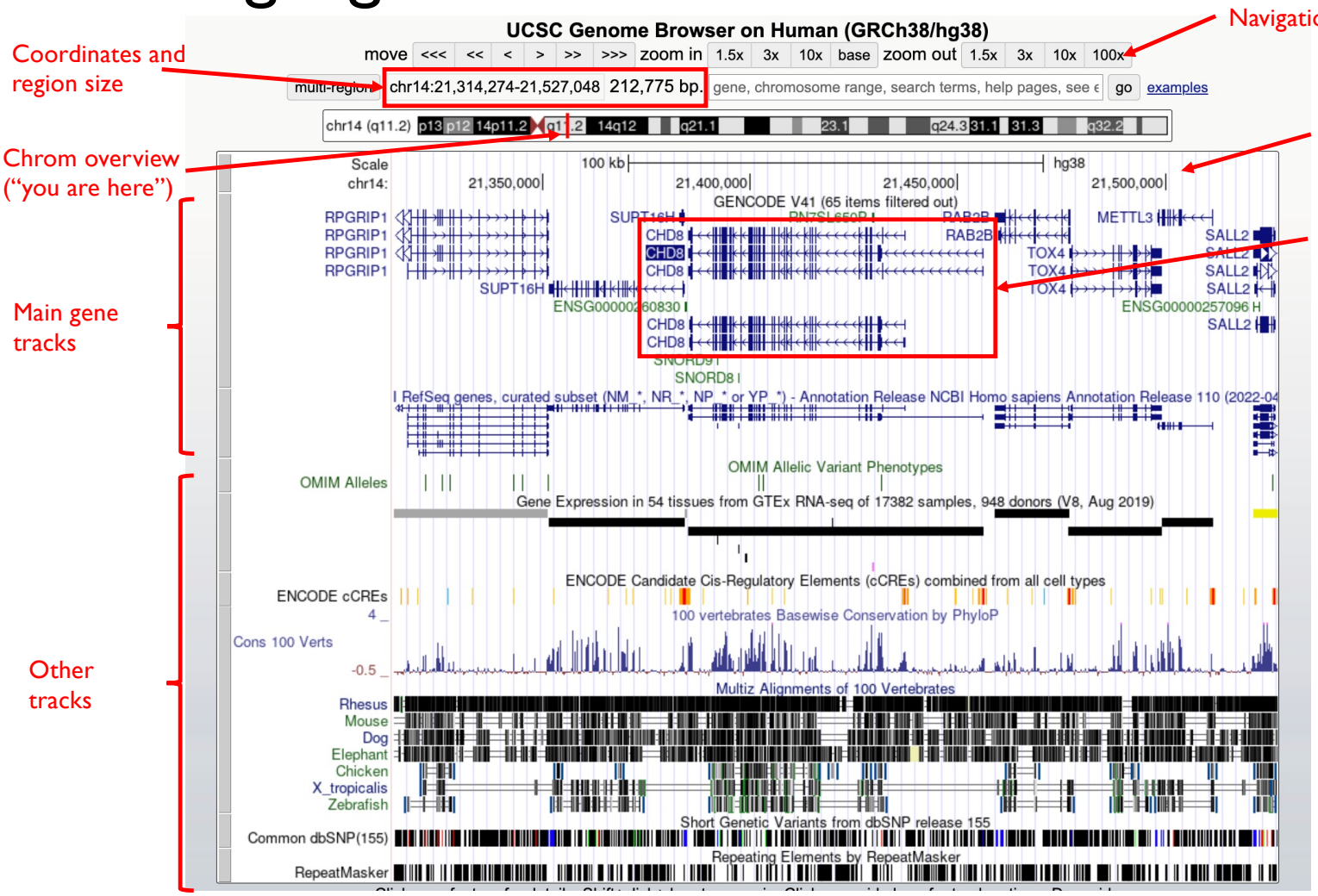
** For biological function; may have biochemical function

“Genome: bought the book; hard to read.” –Eric Lander

- This is a portion of the **reference** genome sequence of the CHD8 locus
- 54,000 bp altogether. Needless to say, we don’t usually look at it like this
- Remember there is another (complementary) strand, and there are two copies of chromosome 14 in each cell
- And because this is a fairly arbitrary reference, your genome may vary – a little

Browsing a genome

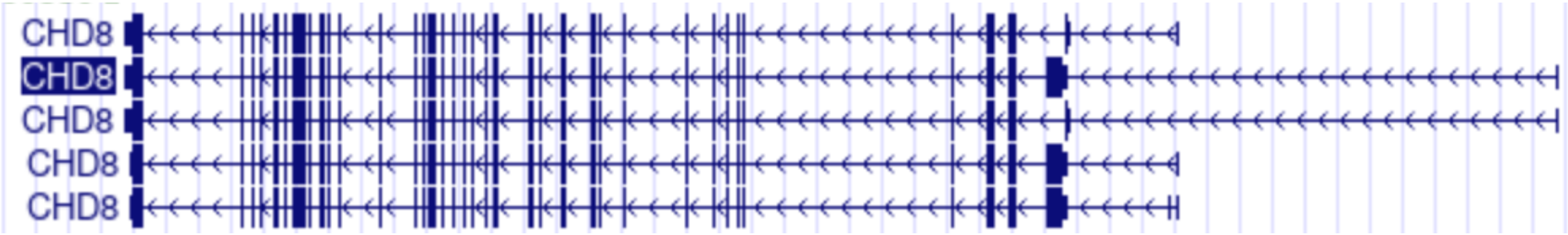
Region of chromosome 14 containing the CHD8 locus in the UCSC browser



[Link to interactive UCSC genome browser view of CHD8]

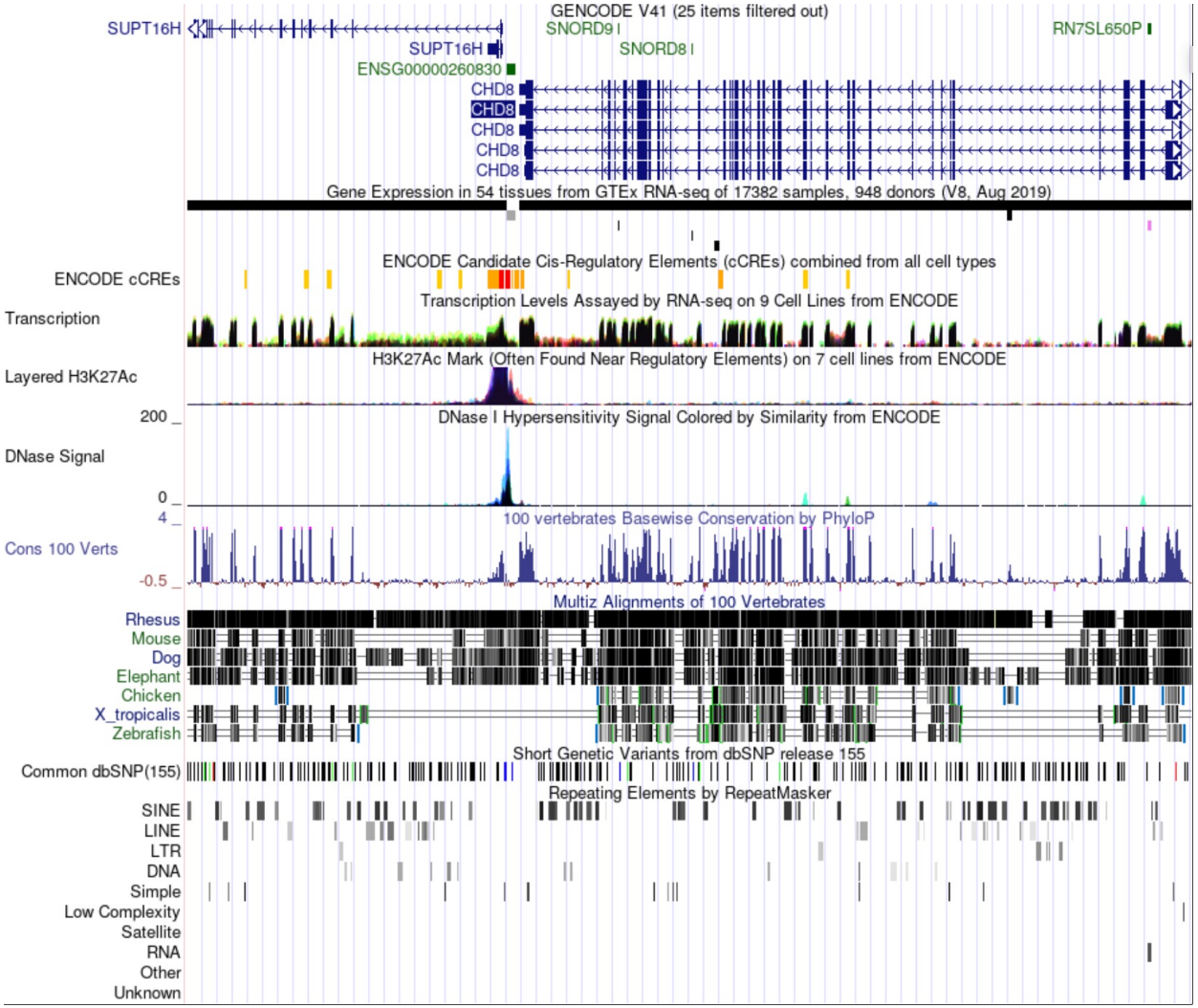
(You may see different tracks; URL is for a default view)

Reading the gene transcript models



- Thinnest lines: introns
- Thick lines: exons
- Thickest part of exons: protein coding
- <<< :direction of transcription (other choice is >>>)

What are those tracks?



Other genes (not necessarily in view)

CHD8 transcripts (3' end)

CHD8 expression and regulation in different tissues

Regulatory signals

This is likely promoter for SUPT16H
CHD8 5' end is off to the right

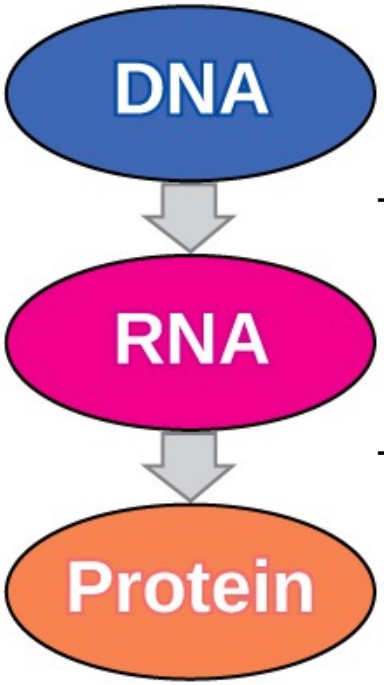
Conservation overview

More detailed conservation

Common variants

Repeat elements

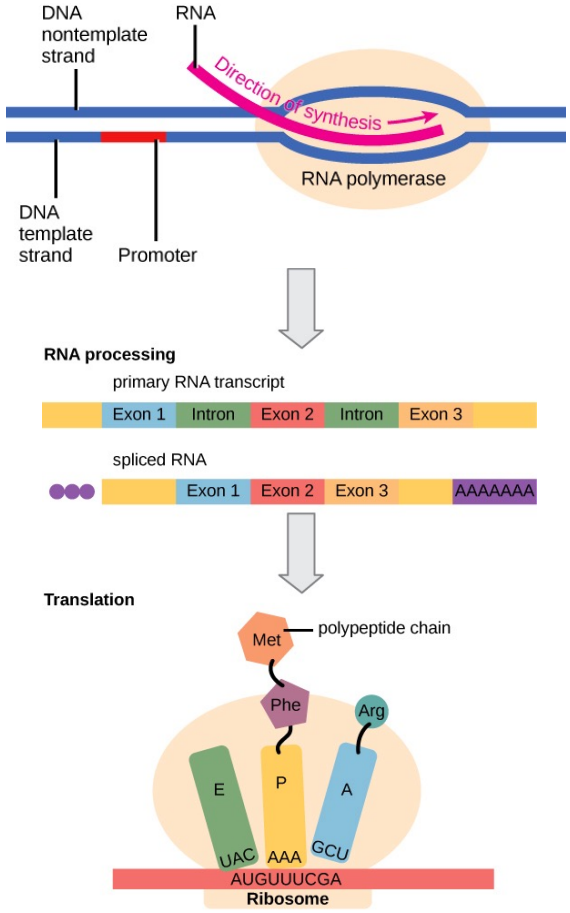
Information transformations



Transcription

Translation

Pedantic aside: OpenStax (and many others) call this the "Central Dogma of Biology" – but that refers to the claim that "Once information gets into protein, it can't get back into the DNA". This figure is just "How we make proteins".



OpenStax

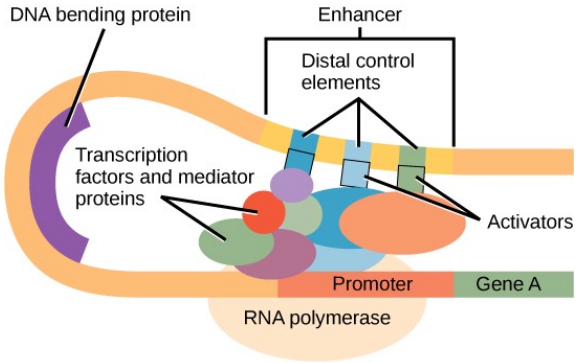
Molecular regulation

- Not all genes are active in any given cell
  - Many organisms are made of different types of cells – the differences are established by changing which genes are active
  - All organisms regulate which genes are active depending on the environment
- Regulation happens at multiple levels (transcription, translation, post-translation)
- System of signals, receptors, switches = complex "wiring" of genes with each other and with the environment – goal of "systems biology" is to understand this in full detail.
- In this course, transcriptional regulation is most relevant

At RNA level, regulatory points include:

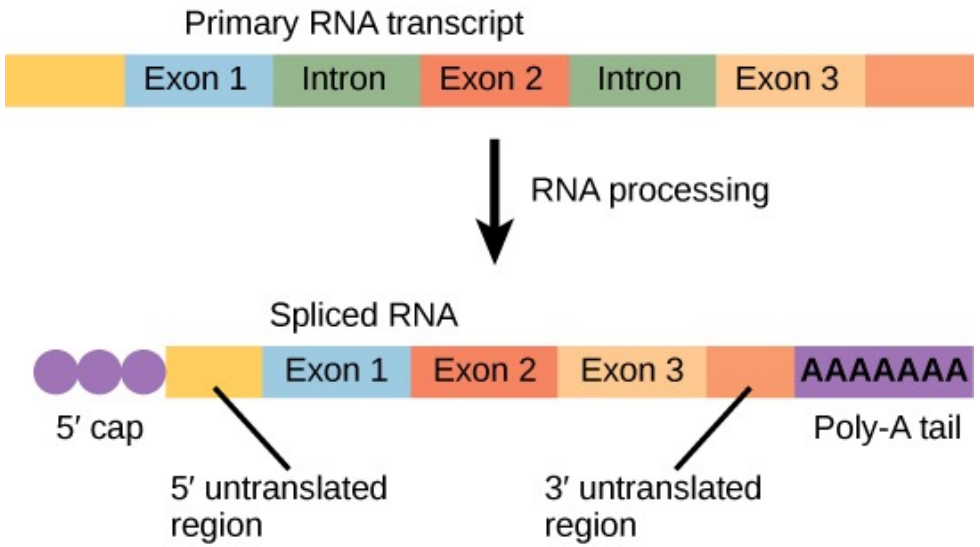
- Rate of transcription
- Rate of modification (splicing etc.), export to the cytoplasm
- Rate of degradation

Regulation of transcription initiation



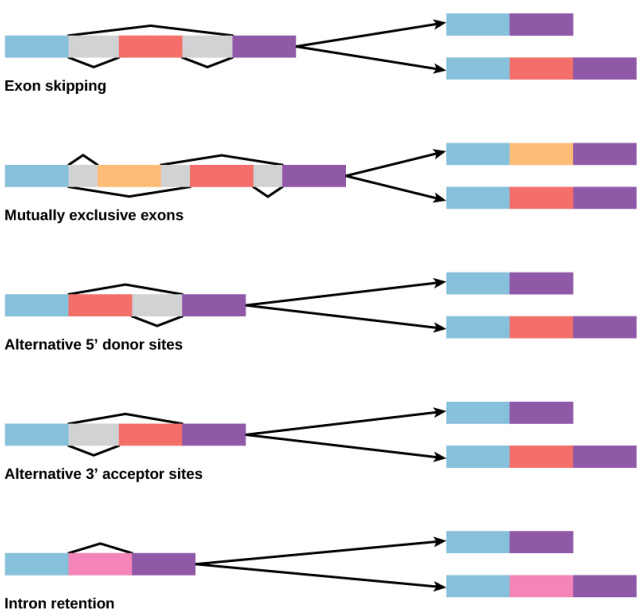
- RNA **polymerase** is the enzyme that **transcribes** the DNA into RNA
- The **promoter** is where RNA pol. gets started - has recognition sites (e.g., "TATA box") that guide binding & initiation
- All the other bits are things that can influence if the gene is "on"
  - This is very generic and over-simplified – there can be repressors and other complications
  - The exact way this works for any specific gene is rarely understood.
- **Regulation**; genes are turned on and off depending on various conditions that control whether these various bits are there (CHD8 codes for a protein that is one of these regulatory factors)

Transcript processing



Note that this figure (from OpenStax) is technically not entirely correct. The untranslated regions are actually part of the exons. "5' untranslated region" is part of Exon 1; and "3' untranslated region" is part of exon 3 (not separate regions as shown here).

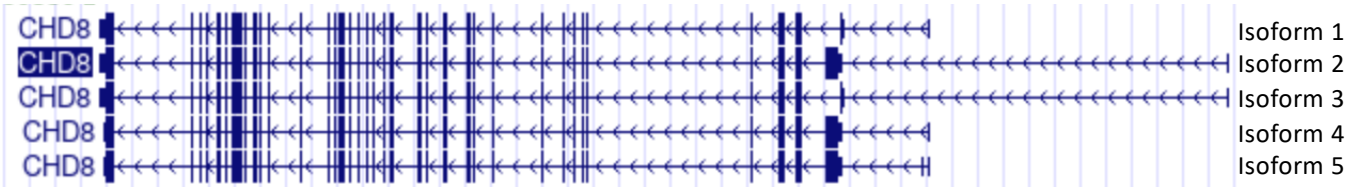
Alternative splicing



Introns: Grey
Black lines show places that will be joined together by splicing

- One gene can have more than one transcript ("isoform")
  - This can happen due to "alternative splicing"
  - Also: genes can have more than one transcript start site/end site
- So one gene can encode multiple transcripts and *may* encode more than one protein
- It's common to designate one isoform as "primary" and the others as "alternative", but it's arbitrary - they are all alternative
- Usually the differential function (if any) of the various transcripts are not known

CHD8 has multiple annotated isoforms in human



- NCBI lists 2 isoforms
- UCSC shows 5
- Ensembl lists 24
  - But only 12 encode a protein, and only 9 unique proteins, and only 2 of those are 'complete'
- So perhaps ***up to*** two different functional proteins (these are the two NCBI recognizes)
- I'm not aware of any evidence they are both important
  - It's important not to assume they are important - transcription and splicing are **noisy**

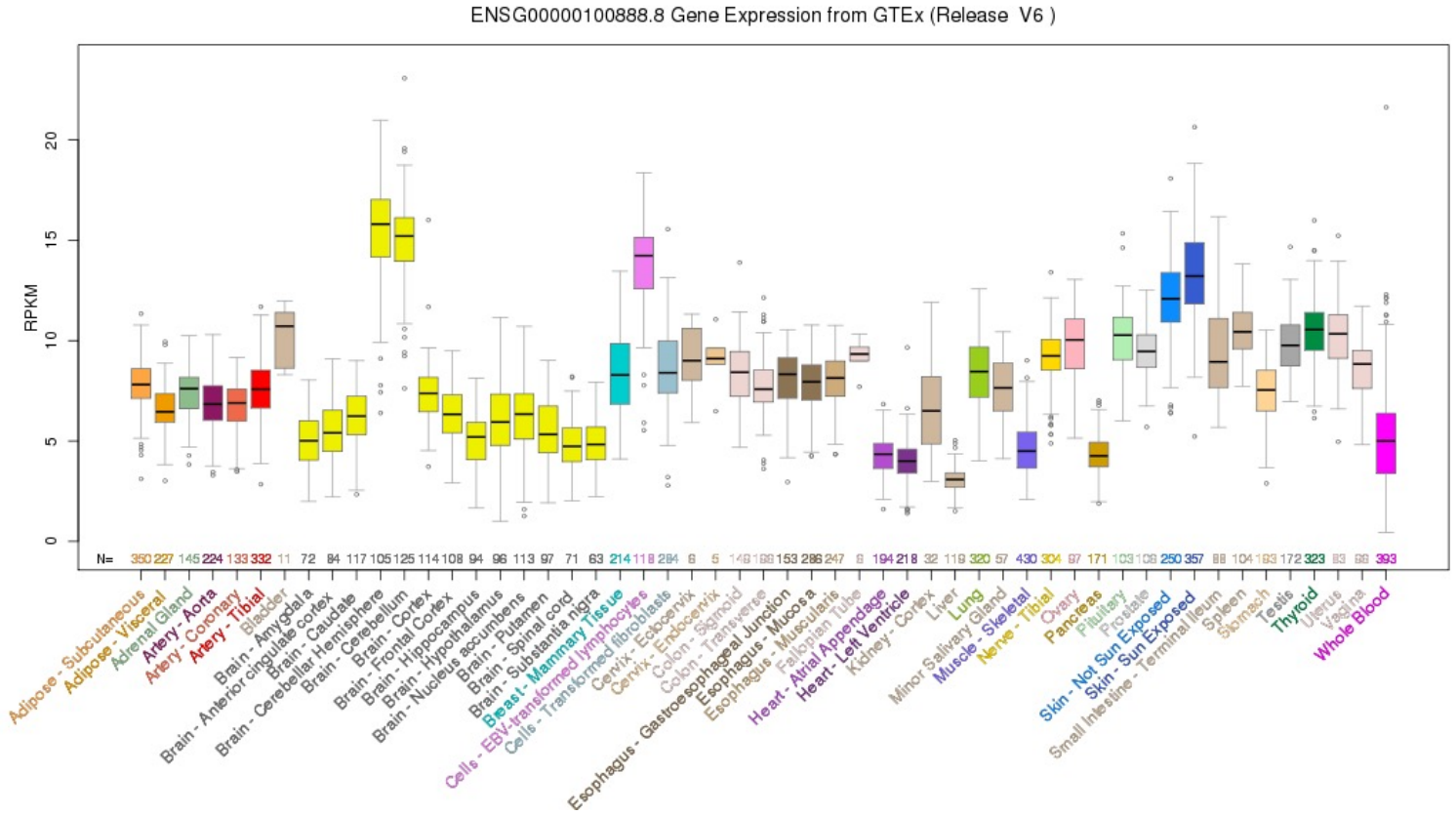
Some context: https://sandwalk.blogspot.com/2018/12/the-persistent-myth-of-alternative.html

A CHD8 mRNA

>NM_001170629.1 Homo sapiens chromodomain helicase DNA binding protein 8 (CHD8), transcript variant 1, mRNA

GCCAGTACCCCCCCTCCCCTTCCCTCCCTAGACCTTGGAGAAGTACCCTCCATTACTTTCCCAAGATGGCAGACCCCATCATGGATCTGTTCGATGACCCAAATTTATTTGGCCTGGACTCTCTGACTGATGACAGCTTTAACCAGGTCACACAAGACCCCATTGAGGAAGCCCTTGGACTGCCAAGCT
CTCTGGACTCCTTGGATCAGATGAACCAGGATGGTGGAGGTGGTGATGTGGGGAATTCATCAGCAAGTGAACTGGTCCCTCCACCAGAGGAAACAGCTCCCACAGAACTTTCCAAAGAATCCACAGCTCCAGCTCCAGAATCCATAACCTTGCATGATTATACCACTCAGCCTGCCAGCCAGGAGCAGC
CAGCCCAACCTGTCTTACAGACATCGACGCCAACATCAGGACTTTTGCAAGTCTCCAAGAGCCAGGAGATCCTGAGCCAAGGGAATCCTTTCATGGGTGTCTCTGCCACAGCTGTCTCCTCCAGTAGTGCTGGAGGGCAGCCACCTCAGTCAGCCCCTAAGATTGTTATCCTTAAGGCCCCACCAAGCT
CCTCAGTCACTGGTGCCCATGTGGCACAAATTCAGGCCCAAGGTATCACCAGCACAGCTCAGCCCCTGGTGGCAGGCACAGCCAATGGTGGAAAAGTCACTTTTACCAAAGTGCTAACCGGCACACCCCTTCGACCAGGTGTTTCCATTGTCTCTGGTAATACAGTGTTGGCCGCCAAGGTCCCTGGGA
ACCAGGCTGCTGTTCAGCGCATTGTCCAGCCCAGCCGACCAGTAAAGCAGCTGGTCCTCCAGCCAGTTAAGGGTTCAGCTCCTGCTGGAAACCCTGGGGCCACAGGGCCCCCACTGAAGCCTGCAGTTACACTGACCTCTACACCTACCCAGGGTGAATCGAAACGCATCACCCTGGTCCTCCAGCAGC
CACAGTCTGGAGGTCCCCAAGGACATCGGCATGTTGTGCTAGGGAGTCTACCAGGCAAGATAGTGTTACAGGGCAACCAGCTGGCAGCCCTGACTCAAGCCAAGAATGCCCAAGGGCAGCCTGCCAAGGTAGTAACTATCCAGCTGCAGGTGCAGCAGCCACAGCAAAAAATCCAGATTGTACCACAAC
CACCATCATCGCAGCCACAGCCCCAGCAGCCACCCTCCACCCAGCCAGTGACTCTGTCCTCTGTACAGCAGGCTCAGATAATGGGACCAGGACAAAGCCCAGGACAAAGACTTTCAGTACCAGTCAAGGTGGTACTGCAGCCACAGGCTGGCTCTTCCCAAGGGGCCTCTTCTGGTAGTTA
AAGTTCTGAGTGCCAGTGAAGTGGCAGCTTTGTCATCACCAGCAAGCTCTGCTCCTCATTCGGGGGGAAAGACAGGAATGGAGGAAAACCGCAGATTGGAACACCAGAAGAAGCAAGAGAAAGCAAATCGGATTGTAGCAGAGGCCATTGCGAGAGCCCGTGCCCGCGGTGAGCAGAACATACCTCGAG
TCTTAAATGAGGACGAGTTGCCCAGCGTTCGGCCAGAGGAGGAAGGCGAGAAGAAACGCAGGAAGAAGAGTGCTGGGGAGAGGCTGAAAGAGGAGAAGCCAAAGAAGAGTAAAACATCTGGTGCCTCCAAAACAAAGGGCAAGAGCAAGCTCAACACCATCACTCCTGTAGTGGGTAAGAAGAGAAAAC
GTAATACCTCATCTGATAATTCAGATGTGGAAGTCATGCCTGCACAGTCACCTCGAGAAGATGAAGAAAGCAGCATTCAGAAGAGACGCTCAAACCGCCAAGTTAAGCGAAAAAAATATACAGAGGACCTGGATATAAAGATCACAGATGATGAAGAAGAAGAAGAGGTGGATGTAACTGGTCCAATAA
AACCTGAGCCTATCCTCCCTGAACCAGTGCAAGAACCAGATGGCGAGACTTTGCCTTCCATGCAGTTCTTTGTGGAGAATCCCAGTGAAGAAGATGCAGCCATTGTAGACAAAGTGCTTTCTATGCGGATTGTGAAGAAGGAGCTCCCTTCTGGACAATATACTGAAGCAGAAGAATTCTTTGTCAAGT
ACAAGAACTACTCCTATCTGCATTGTGAATGGGCTACTATCTCCCAACTAGAGAAGGATAAGAGGATACATCAAAAATTAAAGCGCTTCAAAACCAAAATGGCTCAGATGAGACACTTCTTCCATGAGGATGAAGAGCCCTTTAATCCAGACTACGTAGAGGTGGATAGGATATTGGATGAGTCTCACA
GTATTGACAAGGACAATGGGGAGCCCGTTATTTACTACCTGGTGAAATGGTGCTCTCTGCCCTATGAGGATAGCACATGGGAGCTAAAAGAAGATGTTGATGAGGGCAAGATTCGAGAATTTAAACGGATTCAGTCAAGGCACCCAGAACTCAAAAGGGTGAATCGTCCGCAGGCAAGTGCCTGGAAGA
AATTGGAGCTATCACATGAATATAAAAACAGAAACCAGCTACGGGAATATCAGTTGGAAGGGGTTAATTGGCTGCTCTTTAATTGGTATAACAGGCAGAACTGCATCCTGGCTGATGAGATGGGATTGGGCAAAACTATTCAGTCCATTGCCTTCTTGCAGGAAGTATATAATGTGGGCATCCATGGTC
CCTTCTTGGTCATTGCCCCACTGTCCACAATTACTAACTGGGAGCGAGAATTTAATACATGGACAGAAATGAACACTATTGTGTACCATGGCAGTCTGGCCAGCAGGCAGATGATTCAACAGTATGAAATGTACTGCAAAGATTCACGGGGACGCCTCATCCCAGGCGCATACAAGTTTGACGCTCTGA
TCACCACTTTTGAGATGATTTTGTCAGATTGTCCTGAGCTTCGTGAAATTGAATGGCGTTGTGTTATCATTGATGAAGCCCATCGACTGAAAAACCGTAATTGCAAGCTGCTTGATAGTCTCAAGCACATGGACCTGGAACACAAGGTGCTACTCACAGGAACACCATTGCAAAATACTGTAGAAGAAC
TGTTTAGCTTGCTTCATTTCTTGGAACCGTCACAATTTCCCTCAGAATCAGAGTTTCTCAAGGACTTTGGGGATCTCAAGACAGAGGAACAGGTTCAAAAGCTACAGGCCATTCTTAAGCCAATGATGCTGAGAAGACTCAAAGAGGATGTTGAAAAAAACTTGGCACCCAAACAGGAAACAATTATTG
AAGTAGAGCTGACTAATATCCAGAAGAAATACTATCGGGCTATTTTGGAGAAGAATTTCTCCTTCCTTTCCAAAGGGGCAGGTCATACCAACATGCCTAATCTACTTAACACAATGATGGAGTTGCGCAAGTGCTGCAACCACCCATATCTCATCAATGGTGCTGAAGAAAAAATCCTAACAGAATTCC
GTGAAGCTTGCCATATTATACCTCATGACTTTCACCTGCAGGCCATGGTTCGTTCAGCCGGCAAACTGGTTCTTATTGACAAGTTGCTTCCAAAGCTTAAAGCTGGTGGCCATAAAGTTCTGATCTTCTCTCAGATGGTGCGCTGCCTAGACATCCTAGAGGATTATTTAATCCAGAGGAGGTACTTAT
ATGAACGTATTGATGGGCGAGTTAGAGGCAACCTTCGACAGGCTGCCATTGACCGCTTCAGCAAGCCTGACTCAGACCGCTTTGTCTTCTTACTGTGTACCCGGGCTGGTGGACTTGGTATTAATCTTACAGCTGCTGATACCTGCATCATCTTTGATTCAGACTGGAATCCACAAAATGACCTGCAGG
CCCAGGCACGATGTCATCGAATTGGGCAGAGCAAAGCTGTGAAGGTGTACCGCCTCATCACTCGTAATTCCTACGAGAGAGAGATGTTTGATAAGGCCAGCCTCAAGTTGGGGTTGGATAAGGCTGTGCTTCAATCCATGAGTGGTCGGGATGGCAACATTACTGGAATCCAACAGTTCTCTAAGAAGG
AGATTGAAGATCTTTTAAGAAAAGGAGCATATGCAGCCATCATGGAGGAAGATGATGAAGGCTCCAAGTTTTGTGAAGAGGACATTGACCAGATCTTGTTAAGACGAACTACAACCATCACCATTGAATCTGAAGGAAAAGGTTCCACCTTTGCTAAGGCAAGCTTTGTTGCTTCTGAAAACAGGACAG
ATATTTCTTTGGATGACCCCAACTTTTGGCAAAAGTGGGCCAAAAAGGCTGACCTAGACATGGATCTGCTCAACAGCAAGAATAATTTGGTAATTGACACACCTAGAGTACGAAAACAAACGCGCCACTTTAGCACTCTGAAAGATGATGACCTGGTGGAATTCTCTGATTTGGAAAGTGAGGATGATG
AGCGGCCACGCTCCCGCAGACATGACCGTCATCATGCCTATGGGCGCACTGACTGCTTTCGGGTGGAAAAGCATCTCCTGGTATATGGTTGGGGACGATGGCGAGATATTTTATCTCATGGACGCTTCAAGCGACGTATGACTGAACGAGATGTGGAGACCATTTGTCGGGCCATTCTCGTGTACTGTC
TTCTACACTACCGTGGGGATGAAAATATTAAAGGCTTCATCTGGGACTTGATTAGCCCAGCTGAAAATGGCAAGACAAAAGAATTGCAGAATCATTCAGGTCTATCTATCCCTGTGCCTCGTGGACGCAAAGGAAAAAAAGTAAAGTCACAAAGCACTTTTGATATCCATAAGGCAGATTGGATCCGGA
AATATAACCCTGACACTTTGTTCCAAGATGAAAGTTATAAGAAGCACTTGAAACATCAGTGTAACAAGGTACTGTTGCGGGTACGAATGCTATACTACCTGAGGCAGGAGGTTATTGGAGACCAAGCAGAAAAGGTGTTAGGGGGTGCGATTGCCAGTGAGATTGACATATGGTTCCCAGTAGTGGATC
AACTGGAGGTTCCAACAACTTGGTGGGACAGTGAGGCTGACAAGTCGCTGCTCATTGGAGTCTTTAAACATGGCTATGAGAAATATAATACCATGAGGGCAGACCCAGCCTTATGTTTCCTAGAAAAGGCTGGCCGACCAGATGACAAAGCAATTGCAGCAGAACATCGAGTGTTGGATAACTTCTCTG
ACATAGTAGAAGGGGTTGACTTTGATAAAGATTGTGAAGATCCTGAATATAAACCACTCCAAGGTCCCCCAAAGGACCAAGATGATGAGGGTGATCCCTTGATGATGATGGATGAGGAGATCTCAGTGATTGATGGAGATGAAGCCCAGGTGACCCAACAGCCAGGCCATTTATTCTGGCCTCCGGGCT
CTGCCCTAACAGCTAGGCTTCGGCGTCTAGTAACAGCGTATCAGCGCAGCTACAAGAGAGAACAAATGAAGATAGAGGCTGCAGAACGTGGGGACCGGCGAAGGCGGCGTTGTGAAGCAGCCTTCAAGCTGAAAGAAATTGCACGGCGGGAGAAACAACAACGATGGACAAGGCGTGAACAAACTGATT
TTTATCGAGTGGTGTCTACGTTTGGTGTGGAATATGACCCTGACACCATGCAGTTCCATTGGGATCGCTTCCGCACTTTTGCTCGACTAGACAAAAAGACAGATGAAAGCCTTACCAAGTACTTCCATGGCTTTGTGGCCATGTGCCGCCAAGTATGCCGCCTTCCCCCAGCAGCTGGAGATGAACCCC
CCGACCCTAACCTGTTCATTGAGCCCATCACTGAGGAGAGAGCCTCACGGACTCTCTACCGTATAGAATTGCTTCGGCGCTTACGGGAACAAGTTTTATGCCACCCCCTTTTGGAAGATCGGCTGGCATTGTGTCAGCCTCCTGGTCCTGAATTGCCCAAATGGTGGGAGCCTGTTCGGCATGATGGGG
AGCTTCTAAGAGGGGCAGCCCGCCATGGGGTGAGCCAAACAGACTGCAACATCATGCAGGACCCAGACTTCTCTTTTCTGGCTGCCCGTATGAATTATATGCAGAACCATCAAGCAGGAGCACCAGCTCCATCCTTGTCACGCTGCTCTACTCCACTGCTGCACCAGCAGTATACCTCACGCACTGCCT
CACCACTGCCCCTGCGCCCAGATGCTCCTGTTGAAAAGTCACCCGAGGAGACAGCTACCCAGGTCCCCAGTCTGGAGAGTCTGACTTTAAAGCTAGAGCACGAGGTGGTGGCCAGGAGCCGACCAACCCCACAAGACTATGAGATGCGACGATATCCCCCTCTGATACTACCCCTCTGGTTTCCCGGAGTG
TTCCACCAGTCAAACTGGAGGATGAGGATGATTCGGACTCTGAGCTGGACTTGAGCAAGCTGTCACCATCTTCTTCTTCTTCCTCATCCTCATCCAGCTCCAGCTCCAGCACTGATGAGAGTGAGGATGAGAAGGAAGAGAAGCTAACTGACCAGTCCCGCTCAAAGCTCTATGATGAAGAGAGTCTCC
TGTCCCTCACTATGTCCCAAGATGGATTCCCAAATGAAGATGGAGAACAAATGACCCCTGAGCTTCTGCTACTGCAGGAAAGACAAAGAGCCTCTGAGTGGCCCAAGGATCGTGTCCTGATAAACCGTATTGACCTCGTCTGCCAGGCTGTACTCTCAGGGAAGTGGCCTTCTAGCCGTAGGAGCCAGG
AAATGGTAACAGGAGGAATTTTGGGGCCAGGCAACCACTTGCTAGACAGTCCCTCATTGACTCCTGGAGAATATGGTGACTCTCCAGTCCCCACACCACGAAGTAGTAGTGCAGCTTCCATGGCAGAGGAGGAAGCATCTGCAGTCAGCAGCAGCGGCAGCCCAGTTCACCAAACTTCGCCGAGGCATGG
ATGAAAAGGAGTTTACAGTTCAAATCAAAGATGAGGAAGGATTGAAGTTAACATTCCAGAAGCACAAGTTGATGGCGAATGGAGTAATGGGAGATGGACATCCACTGTTTCATAAGAAGAAGGGGAACAGAAAGAAGCTAGTAGAGCTGGAGGTGGAGTGCATGGAAGAGCCTAATCACCTTGATGTGG
ACCTGGAGACCCGGATCCCTGTCATCAATAAGGTGGATGGTACTTTGCTGGTGGGTGAGGATGCCCCTCGCCGGGCTGAACTGGAGATGTGGTTACAGGGTCATCCAGAGTTTGCTGTTGATCCCCGATTTCTAGCGTATATGGAGGATCGCAGAAAACAGAAGTGGCAAAGATGTAAAAAAAATAATA
AGGCAGAATTGAACTGTTTGGGAATGGAACCAGTACAGACAGCTAACTCTAGAAATGGGAAAAAGGGTCATCACACTGAAACGGTGTTCAACCGGGTTTTGCCAGGGCCTATTGCACCAGAGAGCAGCAAGAAGCGGGCCCGTAGGATGCGACCAGACCTTTCTAAGATGATGGCCCTCATGCAGGGTG
GAAGCACTGGGTCTCTATCTCTGCATAACACGTTCCAACACAGCAGTAGTGGCCTACAGTCTGTGTCATCTTTGGGTCACAGCAGTGCCACTTCTGCATCTTTGCCTTTTATGCCATTTGTGATGGGTGGTGCACCATCATCCCCTCATGTAGACTCCAGCACCATGCTTCATCACCACCACCACCACC
CCCACCCCACCATCACCACCATCACCATCCAGGCTTGAGAGCCCCTGGCTACCCCTCTTCACCAGTGACTACCGCCTCTGGTACTACCTTGCGGTTGCCACCACTGCAACCTGAGGAGGATGACGATGAGGATGAAGAAGATGATGATGACTTATCTCAGGGCTATGATAGCTCAGAAAGGGACTTCT
CACTCATTGATGATCCTATGATGCCAGCTAACTCAGACTCCAGTGAAGATGCTGATGACTGAAGCCCCAGCATGGGCCCCATTGCTTGGGCGGCTGCTGTATTTTCATTTACTCTGGCCCTTGGACTATGGAAACGTGGGAGGGGCAGGGGAGATGTGGGGAAGTCCAGGACTCCAGGAGGTGAAAAGG
AAAAAAAAAAAAAAATGTACCTGATTGCTCCCAATTATGAGAGGATTGGGTGGGCAGGGGAACTCCTAAAATAATACATGACCACTTCCTCATTTCTGGGGAAGGAAAGGAGACTAGAGCAGCTGGTGTGCTCACCCCTCCCTAGTCACCTCCATTAACCACAGACTATGTAGCGCTGGCCCTAGCCTC
TGGCAGAGCCTGTTCCTGGCCGAACTGTGGATACAGCTGGAGGGTCAGGAACTGTTACCTTCTTTCCCCTTGGCATTAATAAATTTAAGTTAATCCTTTTCCAAAAAAAAAAAAAAAAAAAAAAAA

Expression profile for CHD8 across adult tissues

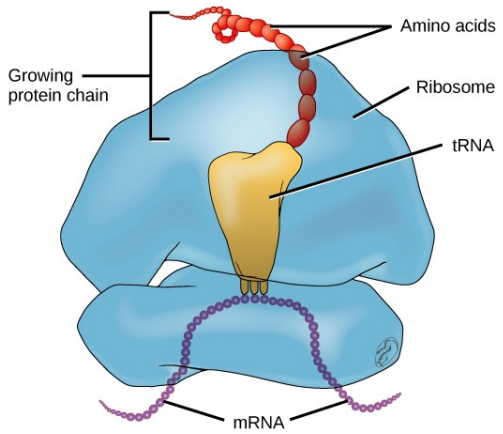


- Measured by RNA-seq (more on this later in the course!) - RPKM is a measure of expression level. Image via UCSC genome browser.
- Boxes are summaries of multiple samples from the named tissue
- CHD8 is expressed broadly; somewhat higher in some places like cerebellum and skin

http://genome.ucsc.edu/cgi-bin/hgGene?hgg_gene=uc001was.3&hgg_prot=ENST00000430710.7&hgg_chrom=chr14&hgg_start=21385193&hgg_end=21437243&hgg_type=knownGene&db=hg38

Or see at https://www.gtexportal.org/home/gene/ENSG00000100888.8

Translation: making proteins



First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC: Phe; UUA, UUG: Leu	UCU, UCC, UCA, UCG: Ser	UAU, UAC: Tyr; UAA: Stop; UAG: Stop	UGU, UGC: Cys; UGA: Stop; UGG: Trp	U C A G
C	CUU, CUC, CUA, CUG: Leu	CCU, CCC, CCA, CCG: Pro	CAU, CAC: His; CAA, CAG: Gln	CGU, CGC, CGA, CGG: Arg	U C A G
A	AUU, AUC, AUA: Ile; AUG: Met	ACU, ACC, ACA, ACG: Thr	AAU, AAC: Asn; AAA, AAG: Lys	AGU, AGC: Ser; AGA, AGG: Arg	U C A G
G	GUU, GUC, GUA, GUG: Val	GCU, GCC, GCA, GCG: Ala	GAU, GAC: Asp; GAA, GAG: Glu	GGU, GGC, GGA, GGG: Gly	U C A G

Recall: a protein is a sequence of amino acids; there are 20 different amino acids

- Input is the mRNA transcript, output is a protein (polypeptide)
- Ribosome is a "molecular machine" that performs this task
- Translation is initiated at a start triplet (codon) (AUG)
- Translation ends at a stop codon (UAA, UAG or UGA)
- The translated part of the mRNA is an "**open reading frame**" (ORF)
  - Changes to the DNA sequence can mess up the open reading frame: **frameshifts** and **nonsense** mutations
  - Or change a codon to one for a different amino acid

OpenStax

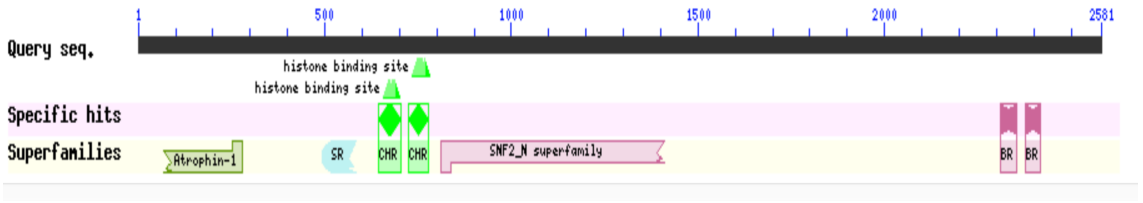
Note: RNA can be functional by itself: not every gene encodes a protein – tRNA is one example, and ribosome itself is partly made of RNA

CHD8 protein

>Q9HCK8
MADPIMDLFDDPNLFGLDSLTDDSFNQVTQDPIEEALGLPSSLDSLDQMNQDGGGGDVGN
SSASELVPPPEETAPTELSKESTAPAPESITLHDYTTQPASQEQPAQPVLQTSTPTSGLL
QVSKSQEILSQGNPFMGVSATAVSSSSAGGQPPQSAPKIVILKAPPSSSVTGAHVAQIQA
QGITSTAQPLVAGTANGGKVTFTKVLTGTPLRPGVSIVSGNTVLAAKVPGNQAAVQRIVQ
PSRPVKQLVLQPVKGSAPAGNPGATGPPLKPAVTLTSTPTQGESKRITLVLQQPQSGGPQ
GHRHVVLGSLPGKIVLQGNQLAALTQAKNAQGQPAKVVTIQLQVQQPQQKIQIVPQPPSS
QPQPQQPPSTQPVTLSSVQQAQIMGPGQSPGQRLSVPVKVVLQPQAGSSQGASSGLSVVK
VLSASEVAALSSPASSAPHSGGKTGMEENRRLEHQKKQEKANRIVAEAIARARARGEQNI
PRVLNEDELPSVRPEEEGEKKRRKKSAGERLKEEKPKKSKTSGASKTKGKSKLNTITPVV
GKKRKRNTSSDNSDVEVMPAQSPREDEESSIQKRRSNRQVKRKKYTEDLDIKITDDEEEE
EVDVTGPIKPEPILPEPVQEPDGETLPSMQFFVENPSEEDAAIVDKVLSMRIVKKELPSG
QYTEAEEFFVKYKNYSYLHCEWATISQLEKDKRIHQKLKRFKTKMAQMRHFFHEDEEPFN
PDYVEVDRILDESHSIDKDNGEPVIYYLVKWCSLPYEDSTWELKEDVDEGKIREFKRIQS
RHPELKRVNRPQASAWKKLELSHEYKNRNQLREYQLEGVNWLLFNWYNRQNCILADEMGL
GKTIQSIAFLQEVYNVGIHGPFLVIAPLSTITNWEREFNTWTEMNTIVYHGSLASRQMIQ
QYEMYCKDSRGRLIPGAYKFDALITTFEMILSDCPELREIEWRCVIIDEAHRLKNRNCKL
LDSLKHMDLEHKVLLTGTPLQNTVEELFSLLHFLEPSQFPSESEFLKDFGDLKTEEQVQK
LQAILKPMMLRRLKEDVEKNLAPKQETIIEVELTNIQKKYYRAILEKNFSFLSKGAGHTN
MPNLLNTMMELRKCCNHPYLINGAEEKILTEFREACHIIPHDFHLQAMVRSAGKLVLIDK
LLPKLKAGGHKVLIFSQMVRCLDILEDYLIQRRYLYERIDGRVRGNLRQAAIDRFSKPDS
DRFVFLLCTRAGGLGINLTAADTCIIFDSDWNPQNDLQAQARCHRIGQSKAVKVYRLITR
NSYEREMFDKASLKLGLDKAVLQSMSGRDGNITGIQQFSKKEIEDLLRKGAYAAIMEEDD
EGSKFCEEDIDQILLRRTTTITIESEGKGSTFAKASFVASENRTDISLDDPNFWQKWAKK
ADLDMDLLNSKNNLVIDTPRVRKQTRHFSTLKDDDLVEFSDLESEDDERPRSRRHDRHHA
YGRTDCFRVEKHLLVYGWGRWRDILSHGRFKRRMTERDVETICRAILVYCLLHYRGDENI
KGFIWDLISPAENGKTKELQNHSGLSIPVPRGRKGKKVKSQSTFDIHKADWIRKYNPDTL
FQDESYKKHLKHQCNKVLLRVRMLYYLRQEVIGDQAEKVLGGAIASEIDIWFPVVDQLEV
PTTWWDSEADKSLLIGVFKHGYEKYNTMRADPALCFLEKAGRPDDKAIAAEHRVLDNFSD
IVEGVDFDKDCEDPEYKPLQGPPKDQDDEGDPLMMMDEEISVIDGDEAQVTQQPGHLFWP
PGSALTARLRRLVTAYQRSYKREQMKIEAAERGDRRRRRCEAAFKLKEIARREKQQRWTR
REQTDFYRVVSTFGVEYDPDTMQFHWDRFRTFARLDKKTDESLTKYFHGFVAMCRQVCRL
PPAAGDEPPDPNLFIEPITEERASRTLYRIELLRRLREQVLCHPLLEDRLALCQPPGPEL
PKWWEPVRHDGELLRGAARHGVSQTDCNIMQDPDFSFLAARMNYMQNHQAGAPAPSLSRC
STPLLHQQYTSRTASPLPLRPDAPVEKSPEETATQVPSLESLTLKLEHEVVARSRPTPQD
YEMRVSPSDTTPLVSRSVPPVKLEDEDDSDSELDLSKLSPSSSSSSSSSSSSSSTDESED
EKEEKLTDQSRSKLYDEESLLSLTMSQDGFPNEDGEQMTPELLLLQERQRASEWPKDRVL
INRIDLVCQAVLSGKWPSSRRSQEMVTGGILGPGNHLLDSPSLTPGEYGDSPVPTPRSSS
AASMAEEEASAVSTAAAQFTKLRRGMDEKEFTVQIKDEEGLKLTFQKHKLMANGVMGDGH
PLFHKKKGNRKKLVELEVECMEEPNHLDVDLETRIPVINKVDGTLLVGEDAPRRAELEMW
LQGHPEFAVDPRFLAYMEDRRKQKWQRCKKNNKAELNCLGMEPVQTANSRNGKKGHHTET
VFNRVLPGPIAPESSKKRARRMRPDLSKMMALMQGGSTGSLSLHNTFQHSSSGLQSVSSL
GHSSATSASLPFMPFVMGGAPSSPHVDSSTMLHHHHHHPHPHHHHHHHPGLRAPGYPSSP
VTTASGTTLRLPPLQPEEDDDEDEEDDDDLSQGYDSSERDFSLIDDPMMPANSDSSEDAD
D

These are one-letter codes for amino acids (2581 of them). Like the DNA sequence, can't do much with this by eye. But comparing this sequence to other proteins tells us more.

Many proteins can be described as a set of "domains" that are relatively modular



This shows that CHD8 has a "SNF2_N" = "SNF2 family N-terminal domain"; Found in proteins involved in a variety of processes. It is found in some proteins that bind to DNA and are involved in gene regulation by changing the bending or packing of DNA.

https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi, search for Q9HCK8

Genetic variation

- Differences in DNA sequence between two individuals of the same species (different genotypes)
- Most variation has no effect on phenotype
  - In non-functional part of genome or is otherwise silent (degeneracy of amino acid code)
- Some variation has effects, but of little consequence
  - Example: fingerprint patterns (arches, whorls) are highly heritable
- Some variation is "deleterious"
  - Slightly deleterious: increases your risk of disease; can be hard to detect
  - Highly deleterious: mutations cause disease; e.g. Cystic fibrosis
- Even more rarely variation can be "beneficial"
- In population genetics "deleterious" and "beneficial" are defined by effects on **reproductive success** (how many descendants you have), but we can also talk about it in terms of the effect on the gene's biochemical function.
- Phrases in the media like "having the gene for___" are usually not correct – often we are talking about genetic variation *within* a gene, not the presence or absence of it

Types of variants

- Localized variants
  - Single base changes (next slide)
  - Small insertions and deletions (indels)
- Structural variants
  - Copy number variants (larger insertions, deletions, duplications)
  - Translocations, inversions

In general, these can have a variety of impacts
- "**Likely gene damaging**" (LGD) – e.g. deletion or nonsense and frameshift SNV
- Changes a **conserved** amino acid – more likely to have impact
- Grab bag: "Variant of unknown significance" (VUS)
- Or no obvious impact ("silent" or "synonymous")

Terminology: The collection of existing genotypes at a locus are the "alleles"

Single nucleotide variants (SNV)

- A base that varies across individuals
- By convention, the term "**single nucleotide polymorphism**" (SNP) is used for **common** variants - the less common ("minor") allele is in >1% of the population (e.g., 15% A, 85% C are the allele frequencies)
  - A source of differences among people, but usually don't break anything important (more on that later)
  - SNPs are easily assayed genome-wide using **microarray-based methods** we'll discuss in the next lecture
  - There are databases of these known variable sites (i.e. dbSNP)
  - This is what services like 23andme look at.
- We use the term "**single nucleotide variant**" to refer to the more general class – disregarding how rare they are.
  - Rare variants are typically found using **direct DNA sequencing methods**. Also a topic for the next lecture.
  - In general, SNVs can be highly deleterious, or not.

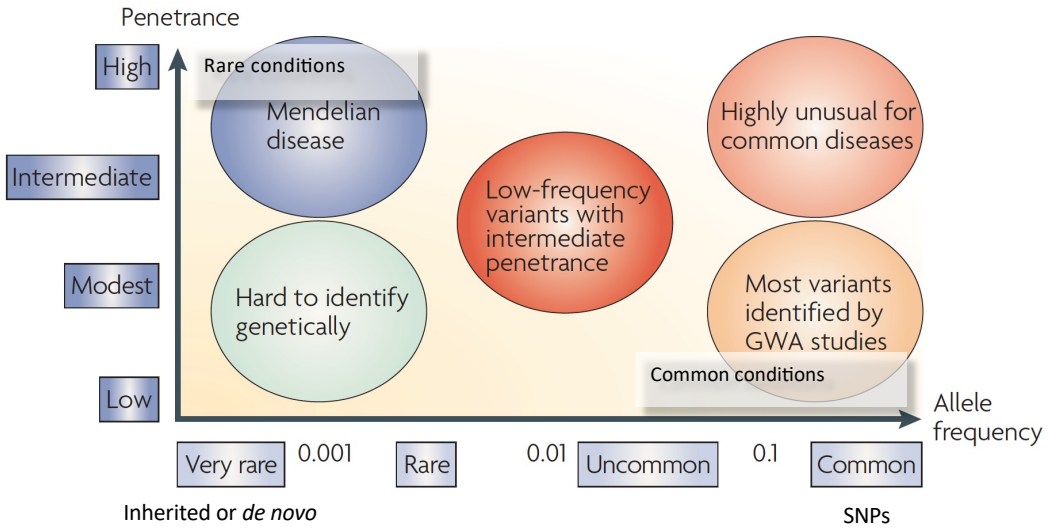
Variation in the human genome

- Compare any two random people: millions of differences - Still, most of genome is identical between any two people (>99%)
- Most differences are inherited SNPs - @ 1% MAF there is one every ~1000 bases (in human; on average)
- Also ~100 new ("*de novo*") mutations per generation; most are not deleterious
  (they're in the 90% of genome that is non-functional, or don't break anything, etc)

MAF = minor allele frequency

ExAC: http://exac.broadinstitute.org/, https://www.nature.com/articles/nature19057)

Rule of thumb: Rare diseases aren't caused by common variants



Related: "Common disease-common variant hypothesis"

Situation for ASD (very roughly)

- "Severe" form (nonverbal, intellectual disability) more explained by rare *de novo* gene-damaging variants in any of a number of genes (at least dozens), each accounting for <1% of such cases
- "Mild" more common form is most influenced by common variants (SNPs). Probably hundreds if not thousands of variants of small effect play a role

www.pnas.org/cgi/doi/10.1073/pnas.1409204111

https://www.biorxiv.org/content/early/2017/11/25/224774

Tammimies et al., 2016 JAMA. 2015;314(9):895-903

Implication of CHD8 in ASD

Disruptive *CHD8* Mutations Define a Subtype of Autism Early in Development

Raphael Bernier,[1,19] Christelle Golzio,[2,19] Bo Xiong,[3,19] Holly A. Stessman,[3,19] E
Kali Witherspoon,[3] Jennifer Gerdts,[1] Carl Baker,[3] Anneke T. Vulto-van Silfhout
Marco Fichera,[5,6] Paolo Bosco,[5] Serafino Buono,[5] Antonino Alberti,[5] Pinella Fa
Lisenka E.L.M. Vissers,[4] Ludmila Francescatto,[2] Heather C. Mefford,[11] Jill A. R
Brian J. O'Roak,[14] Matthew Pawlus,[15] Randall Moon,[15,18] Jay Shendure,[3] Davi
Corrado Romano,[5] Bert B.A. de Vries,[4] Nicholas Katsanis,[2] and Evan E. Eichle

Cell *158*, 263–276, July 17, 2014

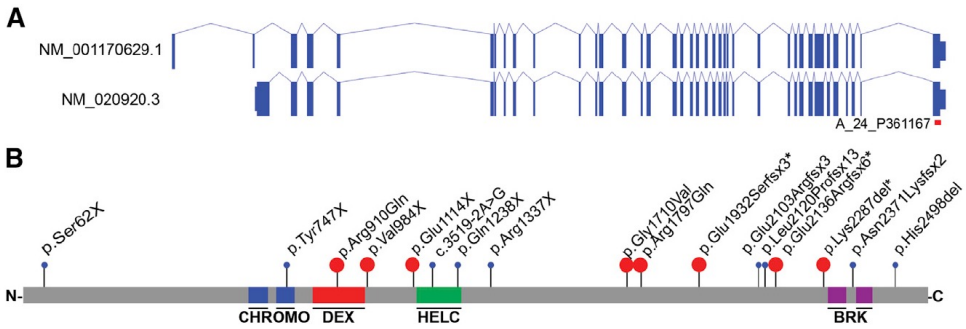


Figure 1. Spectrum of *CHD8* Mutations in Autism Spectrum Disorder

(A and B) (A) Gene isoforms 1 and 2 and (B) protein models of *CHD8* with proband putative disruptive mutations indicated. The location of the gene expression array probe used for Figure 3 (A_24_P361167) is shown in (A) in red. Events in blue were reported previously (Neale et al., 2012; O'Roak et al., 2012a). Events in red are novel. (*) Diagnosis of intellectual disability (Table 1). See also Tables S1 and S2 and Figure S1.

This paper summarizes **very rare *de novo*** single nucleotide variants (SNVs) causing a **likely loss of function allele** of CHD8 and ASD w/ID, macrocephaly and dysmorphic facial features.

- Damaging variants never found in individuals who do not have ASD/NDD.
- Now one of the best-established ASD-assoc. genes; though <0.5% of people with an ASD have this
- These individuals have one copy of the gene that is ok (**heterozygous**); if both copies are broken: probably lethal (never observed)
- For more information on this syndrome see also https://omim.org/entry/615032

Why does having a bad copy of CHD8 cause ASD?

- Nobody knows exactly
- But the reductionist approach has given a lot of information about what CHD8 does in cells

Short answer: CHD8 regulates expression of many genes. Some of the genes it regulates are important for brain development (and other things). Losing one copy isn't fatal but it results in an obvious phenotype

CHD8 regulates transcription

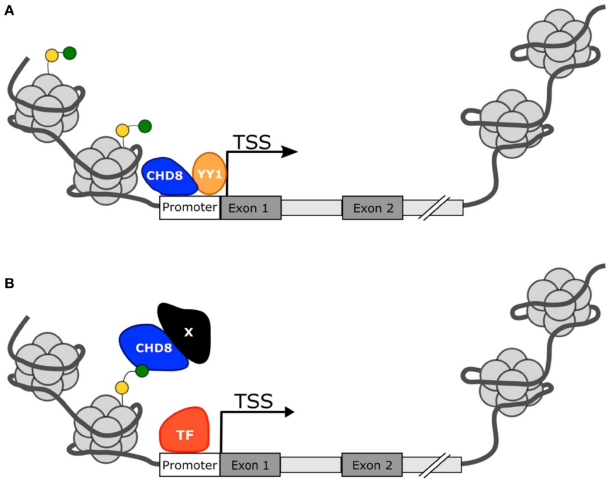
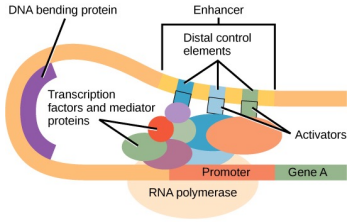


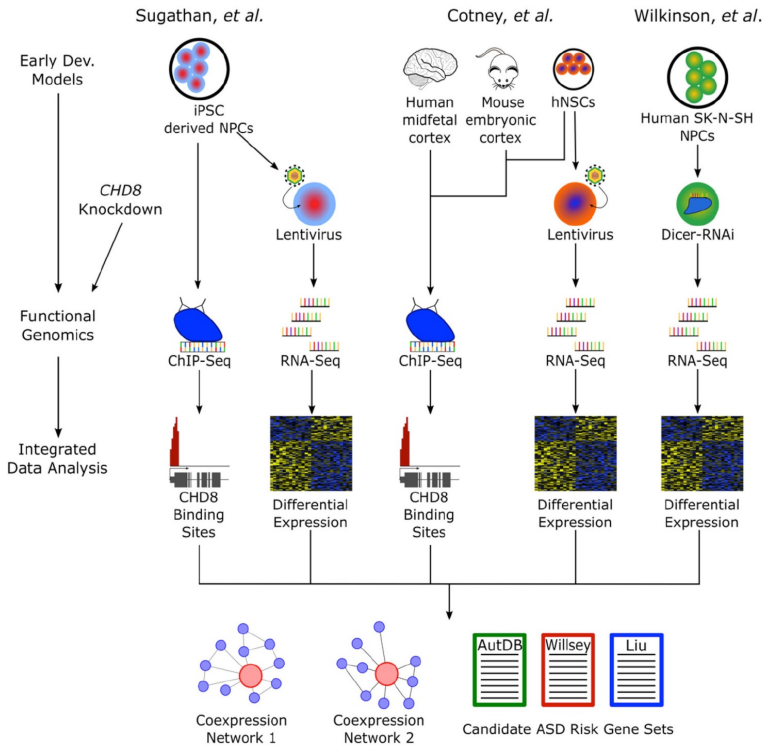
FIGURE 3 | Proposed mechanisms for CHD8 transcriptional activation. (A) CHD8 is most commonly found near active transcription start sites with histone modifications H3K4me3 (green circle) or H3K27ac (yellow circle). CHD8 may directly activate genes by directly binding near the transcriptional start site and promote transcription factor activity or recruitment. **(B)** CHD8 may indirectly activate genes through interactions between modified histone sites and other co-regulators to make chromatin more assessable.

https://www.frontiersin.org/articles/10.3389/fnins.2015.00477/full



- It can bind to DNA, increasing or decreasing transcription of a nearby gene
- And/or it can bind to other proteins which are bound to the DNA, also increasing or decreasing transcription of nearby genes
- These kinds of intermolecular interactions can be at least partly specific – there are only some places in the genome where CHD8 can bind, and only some of those do anything

What genes does CHD8 regulate, and what happens when it is broken?



https://www.frontiersin.org/articles/10.3389/fnins.2015.00477/full

Figure is from a 2015 review, summarizing three studies:

- Look for where CHD8 binds to DNA (ChIP-seq)
  - Thousands of binding sites, much in/near promoters
- Decrease CHD8 in cells/animals and see what goes wrong (RNA-seq, et al.)
  - Thousands of RNAs go up or down in levels
- Some of the "regulatees" are other ASD-associated genes

Some of the topics we'll get back to

- Measuring transcription – Microarrays/RNA-seq, and how we analyze such data e.g. differential expression (much of the course)
- Sequencing genomes, genotyping arrays and SNPs
- Coexpression and gene sets (later in course)
- And more …

Reminders

- Survey for Project Group formation on canvas (submit by Sunday January 14 at 11:59pm)
- Seminar 1 submission due Friday January 12 at 11:59pm