

Statistical Methods for High Dimensional Biology

STAT/BIOF/GSAT 540

Course introduction + Molecular Biology Primer

Keegan Korthauer

10 January 2022

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 xʷməθkʷəy̕əm (Musqueam) people.

As we begin the term in a virtual learning environment, we are coming together from locations throughout Vancouver, BC, Canada, and the world.

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 – full details on course website:
<https://stat540-ubc.github.io/>
(These lecture notes are just an overview)
- A [fast-paced!] primer on molecular biology/genetics/genomics

Teaching team

Instructors



Keegan Korthauer

Yongjin Park

Teaching Assistants



Asfar Lathif



Marco Tello

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!

Prerequisites

Officially, none. But:

- **Statistics** – You should have already taken university level “Statistics 101”. You’ll get a refresher, but you should be prepared 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. If you can’t install R/Git on your computer, ask us for options.
- **Lots of extra resources posted on Syllabus page**

What you can expect to learn

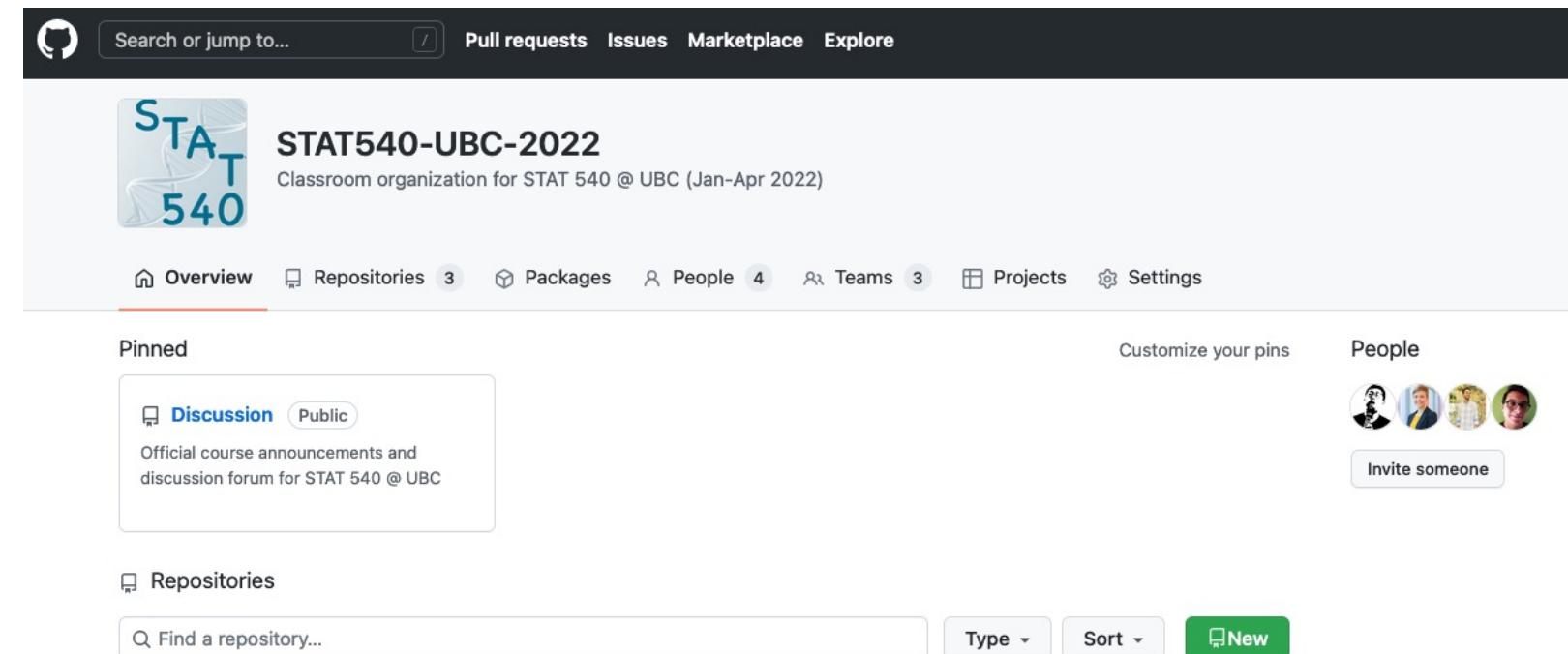
- Apply tailored statistical methods to answer questions using high dimensional biological data
 - Generally applicable statistical approaches and principles
 - Specifics about some data types (esp. expression profiling)
 - Practical experience using the R/Bioconductor computing environment
- Critically 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

STAT540-UBC-2022 Class GitHub Organization

Primary tool for administering the course

- Discussion Repository for course communication
- Individual (private) student repositories for each assignment
- Shared team repositories for project work



<https://github.com/STAT540-UBC-2022/>

Course web site

<http://stat540-ubc.github.io>

- Syllabus
- Assignment submission instructions & due dates
- Rubrics
- Lecture and seminar notes
- Contact info for Instructor & TA
- Virtual Office hour schedule

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

- Zoom links for online lecture and seminar meetings
- Zoom links for virtual office hours
- Links to livestreams for in-person lectures
- GitHub assignment invite links
- Grades
- Survey for Project Group formation (**please complete by Friday January 14 at 11:59pm**)

The screenshot shows the left sidebar of a Canvas course. The sidebar includes icons for Account, Dashboard, Courses, Calendar, Inbox, and Course Evaluation. The main content area displays a survey titled "Background & Research Interests". The survey details are: Due Jan 14 at 11:59pm, Points 0, Questions 10, and Time Limit None. Below the title is a section titled "Instructions" with a descriptive paragraph about the purpose of the survey. A blue button at the bottom right says "Take the Survey".

BIOF/GSAT/STAT 540 2021W2 > Quizzes > Background & Research Interests

2021W2

Home

Assignments

Zoom

Grades

People

Pages

Quizzes

Media Gallery

Course Evaluation

Background & Research Interests

Due Jan 14 at 11:59pm Points 0 Questions 10 Time Limit None

Instructions

The purpose of this survey is to learn about your background in the course subject areas and research interests so that we can build effective teams for the group projects. Please only complete this survey if you are taking the course for credit (i.e. don't complete the survey if you are auditing, as you will not participate in the project).

Take the Survey

Video etiquette and privacy

- Online Zoom meetings, in-person lecture streams, and their recordings are not to be made available to anyone outside the course
- Please keep your camera **off** and microphone **muted** when you aren't speaking
 - If your environment/bandwidth is suitable and you are comfortable doing so, please turn your cameras on when speaking
- Please make use of the chat feature to ask questions or make comments - the TAs will be monitoring this space

Course Communication

<https://github.com/STAT540-UBC-2022/Discussion>

Search or jump to... Pull requests Issues Marketplace Explore

STAT540-UBC-2022 / Discussion Public

Watch 4 Fork 0

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

main 1 branch 0 tags Go to file Add file Code About

kdkorthauer Update README.md ... 3a1a87e 2 days ago 11 commits

README.md Update README.md 2 days ago

README.md

Official course announcements and discussion forum for STAT 540 @ UBC

Official course announcements and discussion forum will be held in the [issues](#) of this repository.

How to use

- Subscribe to notifications for this repo (look for the 'Watch' button in the upper right, and check your email notification settings in your GitHub account).
- To start a new thread [create a new issue](#).

About

Official course announcements and discussion forum for STAT 540 @ UBC

Readme 0 stars 4 watching 0 forks

Releases

No releases published [Create a new release](#)

Packages

No packages published [Publish your first package](#)

Use “Issues” feature in the Discussion repository for

- **Official Announcements**
- All questions about course material / assignments
- Only use email for sensitive / private matters
- This repo will be made private to our class starting next week

Course Communication

<https://github.com/STAT540-UBC-2022/Discussion>

Subscribe to notifications for new 'issues'

Search or jump to... / Pull requests Issues Marketplace Explore

STAT540-UBC-2022 / Discussion Public

Watch 4 Fork 0 Star 0

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

Filters is:issue is:open Labels 6 Milestones 0 New issue

Welcome to issues!

Issues are used to track todos, bugs, feature requests, and more. As issues are created, they'll appear here in a searchable and filterable list. To get started, you should [create an issue](#).

Course Communication

<https://github.com/STAT540-UBC/Discussion>

Live Demo

Lectures

- Schedule of topics/instructors: <https://stat540-ubc.github.io/lectures>
- Start at 9:30am, and aim to end at 10:50am (Vancouver time); will be recorded
 - Links for online Zoom sessions and in-person streaming on Canvas
- Switching to in-person in ESB 2012 no earlier than Jan 24
- Slides will be posted on the website as we go

Seminars (Computing Labs)

Mondays 12-1 PM

- Schedule & material (please read before attending each session):
<https://stat540-ubc.github.io/seminars>
- TA-led exercises using Git, GitHub, R/RStudio
- You will follow along on your own computer
- Most sessions have ‘deliverables’ that you submit for a grade
- Meeting links posted in Canvas

First session **today**

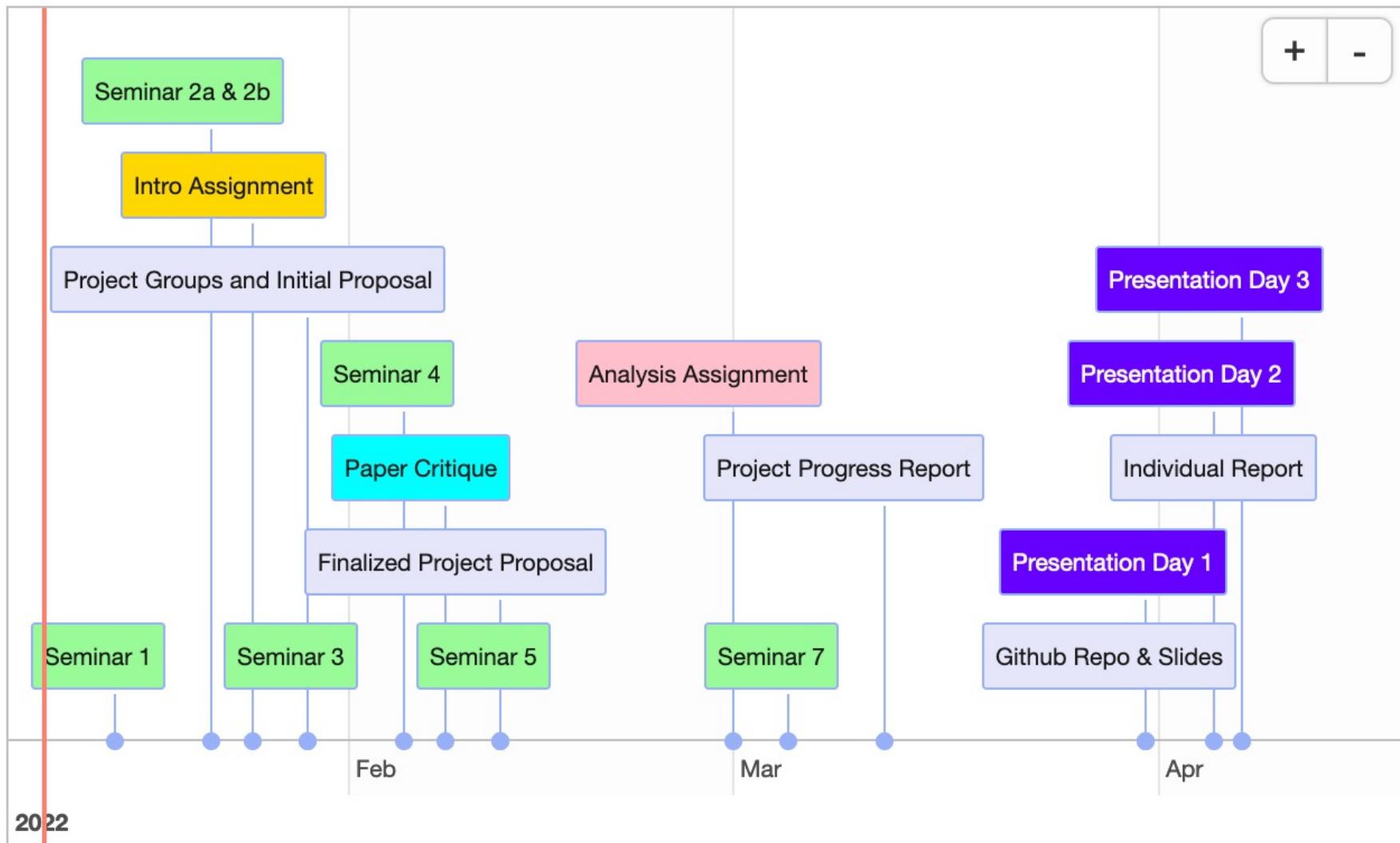
- Git/GitHub setup
- Demo how to submit assignments in this course
- Have R/RStudio installed beforehand, if possible

Evaluation

- **Intro Assignment (5%)**
- **Seminars (10%)**
 - 6 deliverables, keep 5 highest scores each worth 2%
- **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 (50%)**
 - Proposal (5 pts)
 - Progress report (10 pts)
 - Oral presentation (20 pts)
 - GitHub repository (10 pts)
 - Individual report (5 pts)

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

Key deadlines



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

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, 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.

Submitting assignments

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

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 (aimed at **4 people per group**)
- Several deliverable checkpoints throughout the term
- Overview here: <https://stat540-ubc.github.io/assignments.html#final-group-project-50>
- 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 of past group project topics

- Genomic copy number alterations for prognosis of prostate cancer
- Analysis of HPV E2 protein on host gene expression
- 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

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
- 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
- What kind of data are we taking about

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

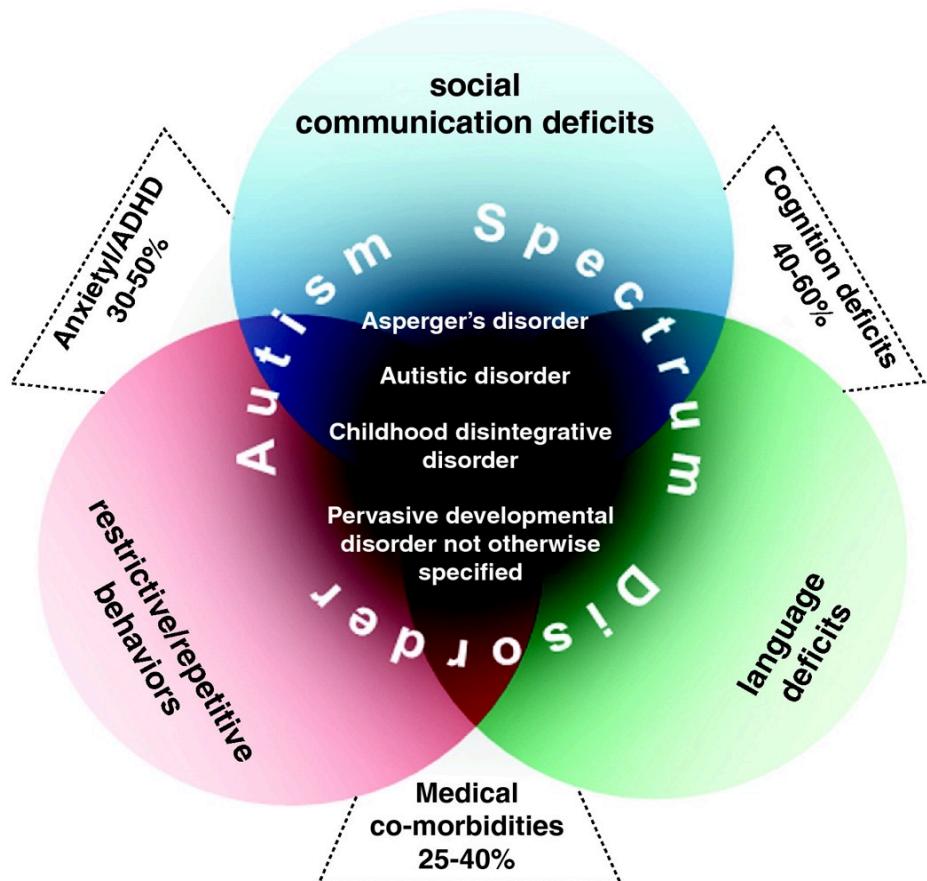
Why study genomics and genetics?

- 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)



- 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/>



Available online at www.sciencedirect.com
SciVerse ScienceDirect

Current Opinion in
Genetics & Development

Genetic architecture in autism spectrum disorder
Bernie Devlin¹ and Stephen W Scherer²

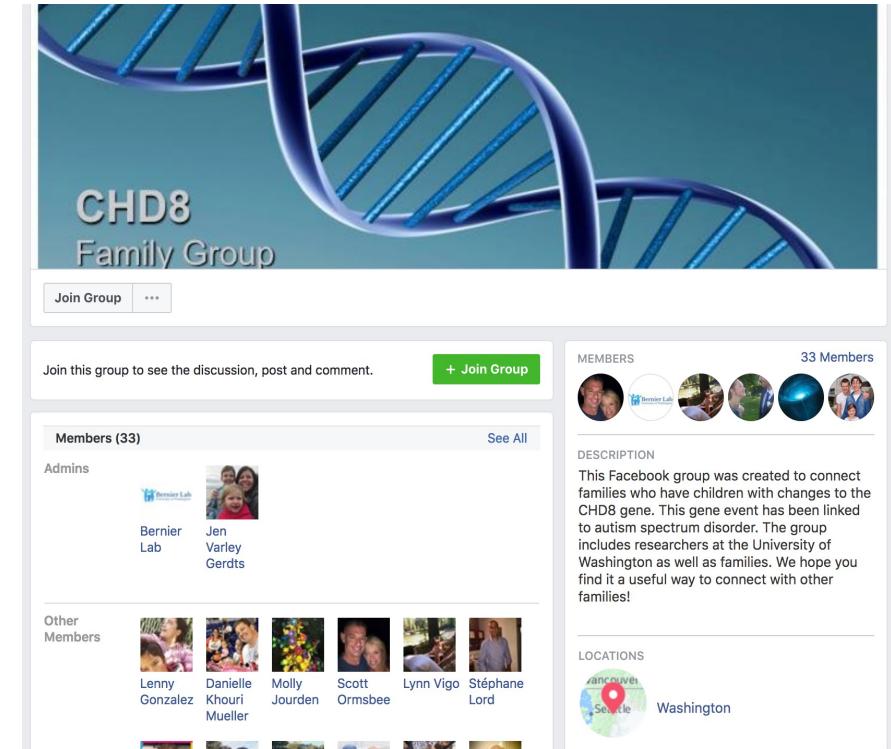
How can we study ASD biology?

- 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



The screenshot shows a Facebook group page titled "CHD8 Family Group". The cover photo features a blue-toned DNA double helix. The group has 33 members. Administrators listed are Bernier Lab and Jen Varley Gerdts. Other members shown include Lenny Gonzalez, Danielle Khouri Mueller, Molly Jourden, Scott Ormsbee, Lynn Vigo, and Stéphane Lord. The group's description states: "This Facebook group was created to connect families who have children with changes to the CHD8 gene. This gene event has been linked to autism spectrum disorder. The group includes researchers at the University of Washington as well as families. We hope you find it a useful way to connect with other families!" A location pin indicates the group is based in Seattle, Washington.

<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.

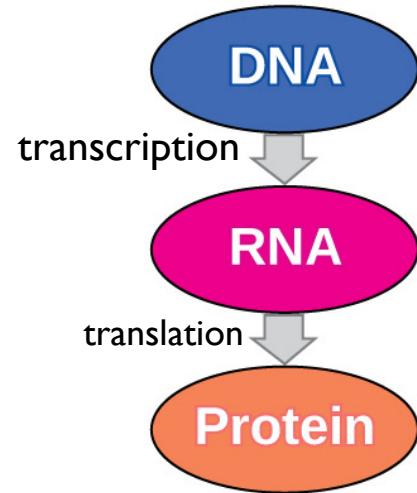
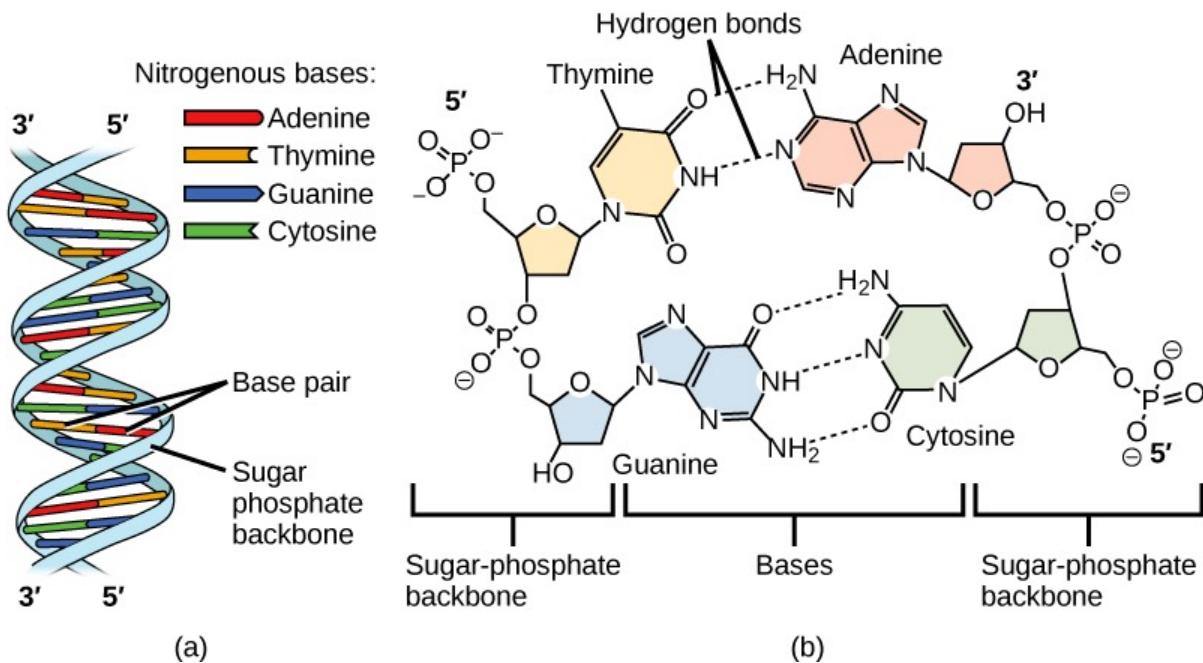


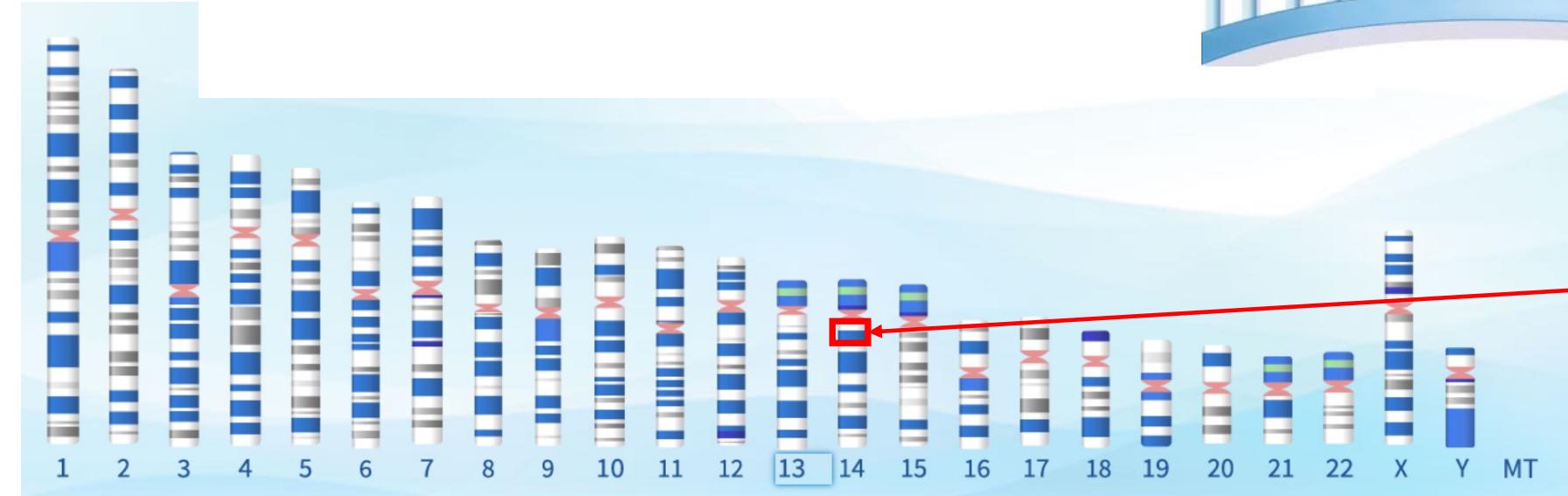
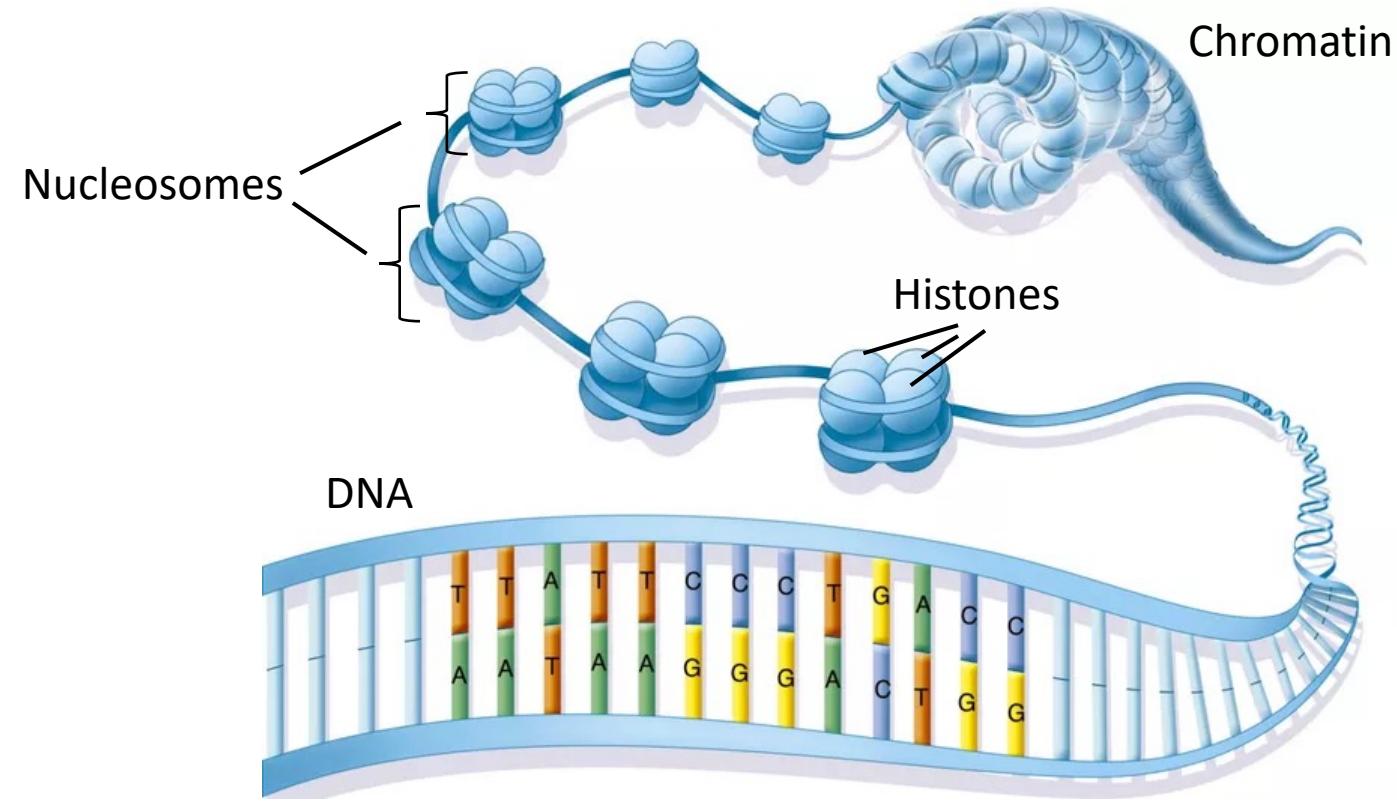
Figure 9.14, Concepts of Biology

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 we think is “functional”
- 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)

A genome (human)



CHD8 is here

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 (~15000) – broken gene copies or relics – NOT genes

How much of the human genome is functional?

Depends on how you define “function”. Reasonable definition considers genome evolution: Why is that DNA there? Does changing it affect survival?

- ~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 nearly 100% functional genomes.

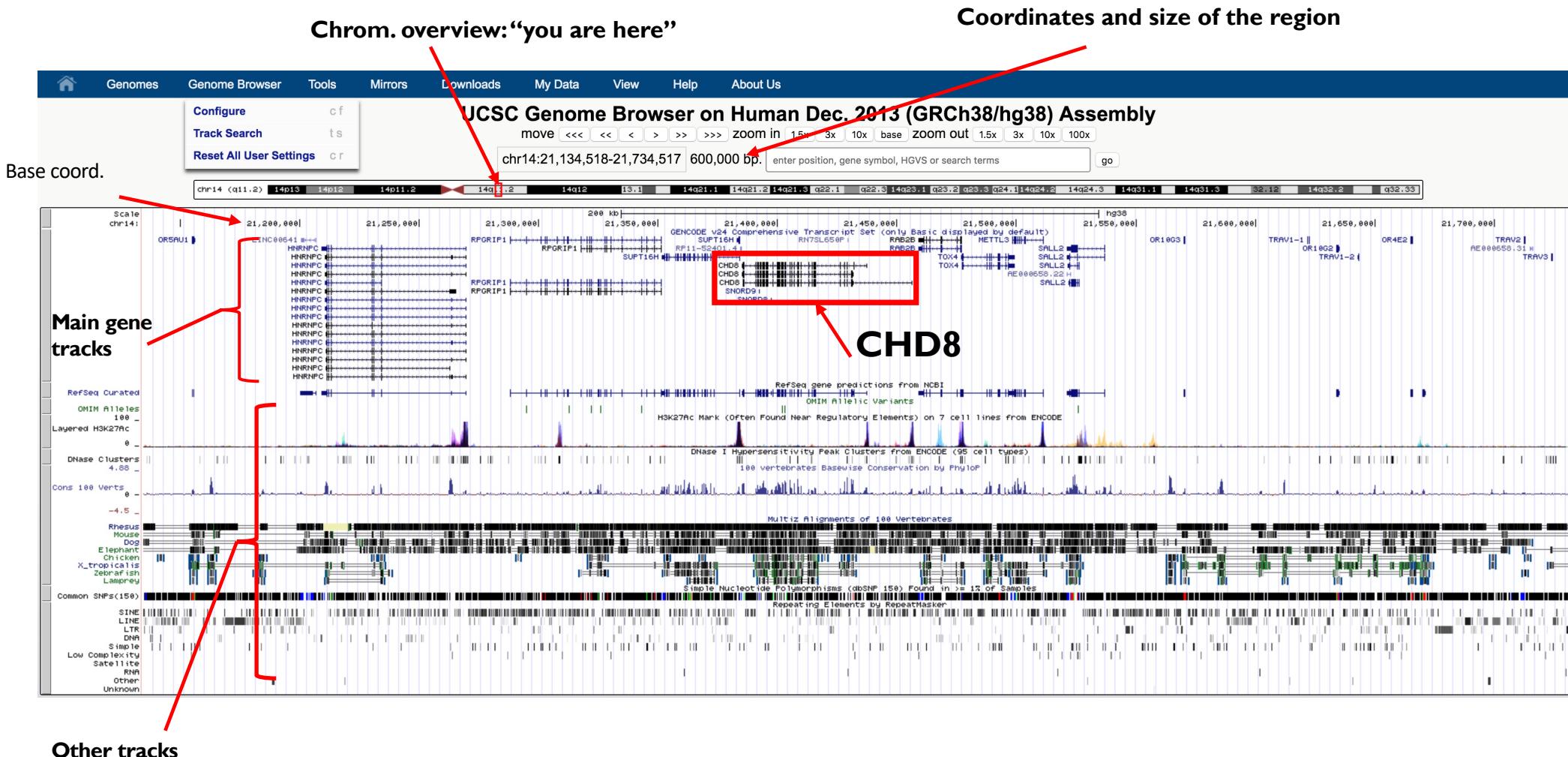
** For biological function; may have biochemical function

“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

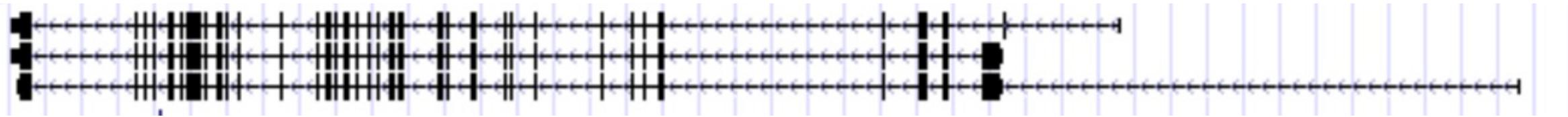
Region of chromosome 14 containing the CHD8 locus in the UCSC browser



<http://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&position=chr14%3A21134518-21734517>

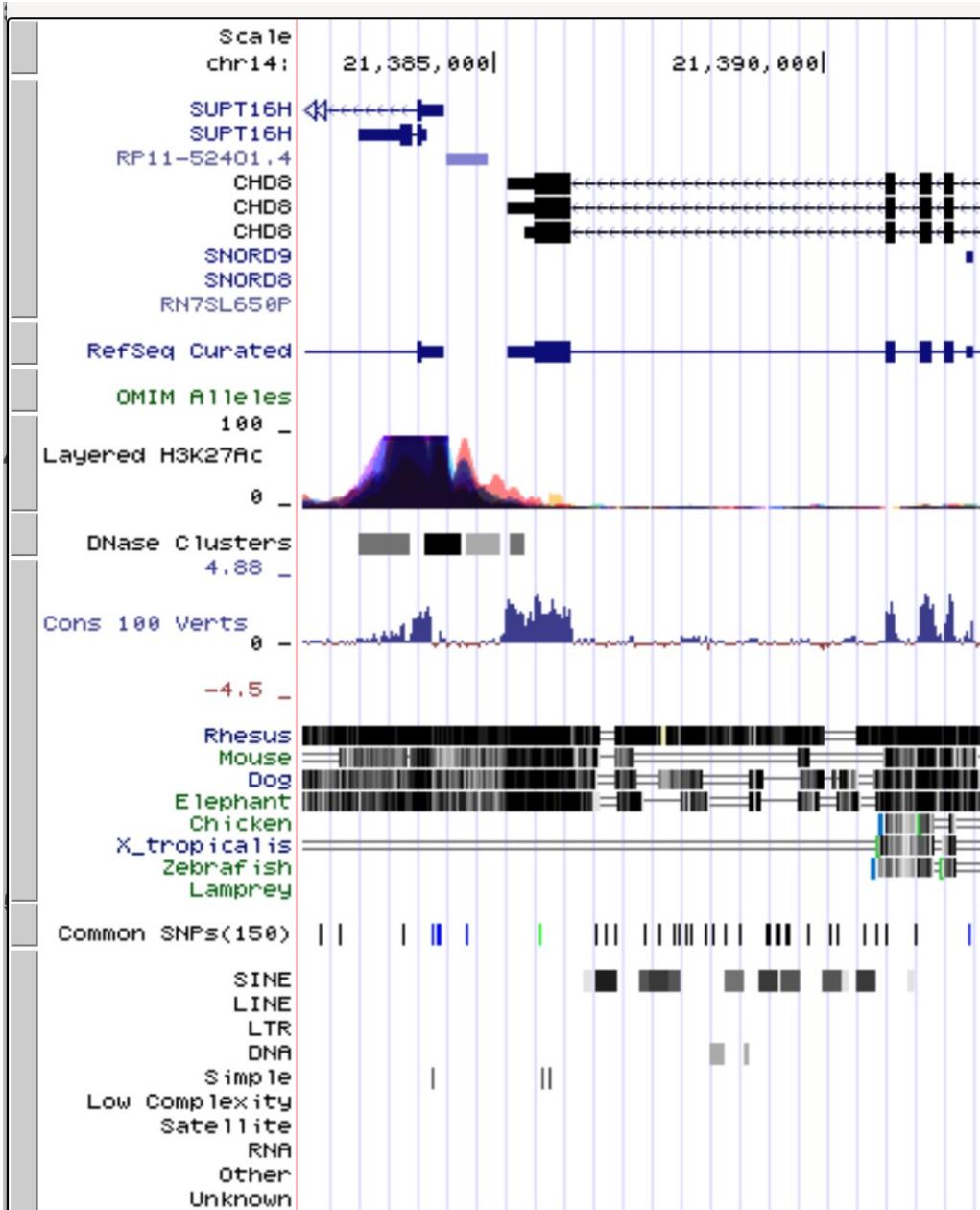
(You may see different tracks; URL is for a default view)

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)

Other genes (not necessarily in view)

Indicators of “promoter”

Or other regulatory site

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

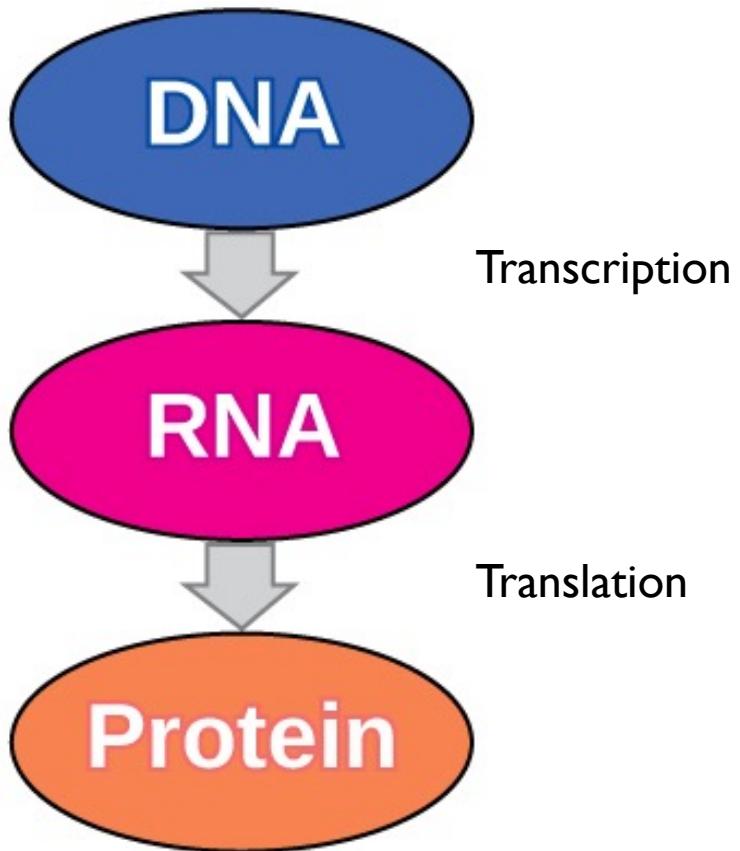
Conservation overview

More detailed conservation

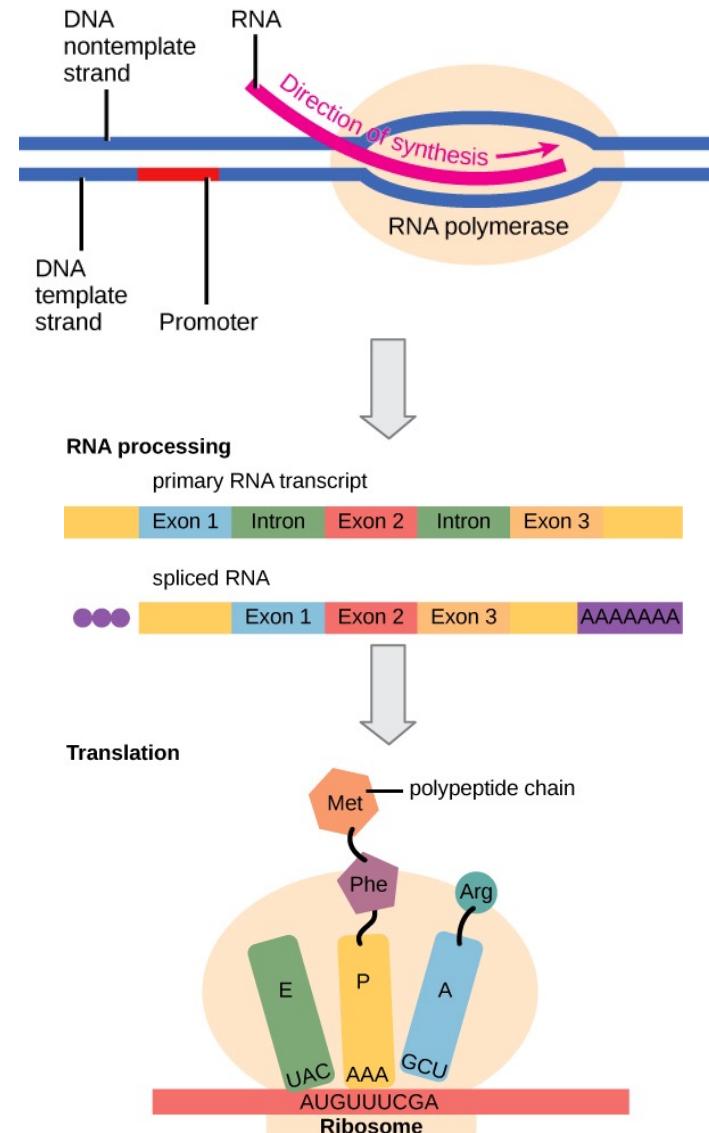
Common variants

Repeat elements

Information transformations



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”.



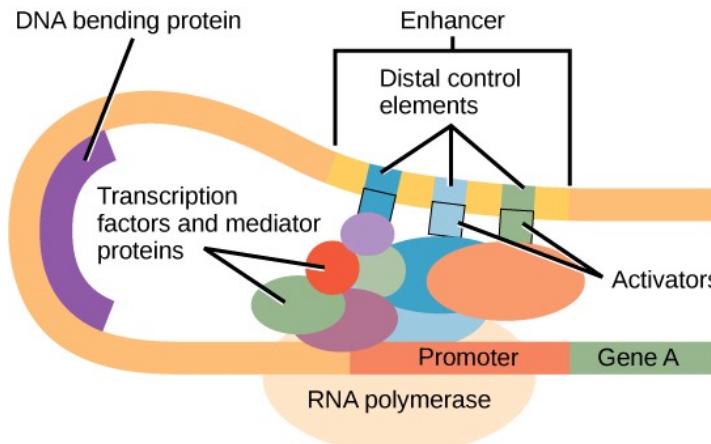
Molecular regulation

- 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:

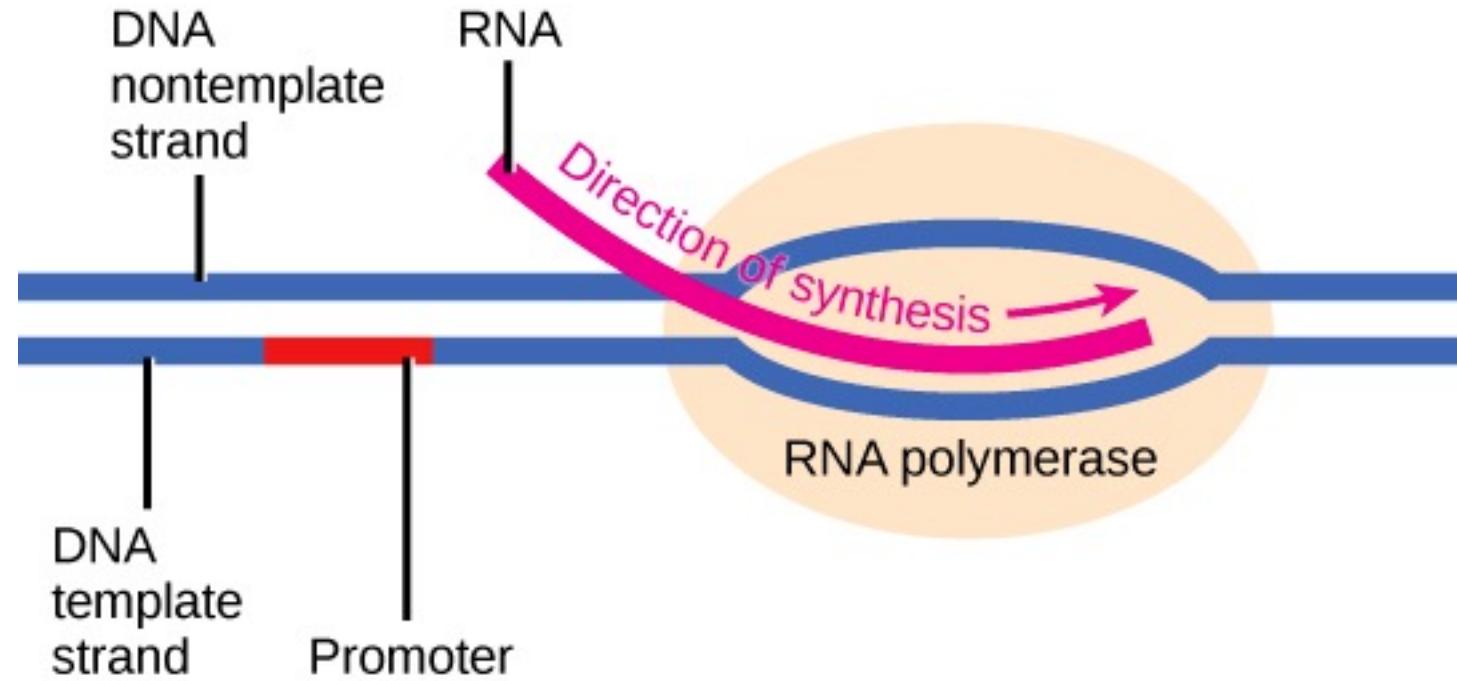
- Rate of transcription
- Rate of modification (splicing etc.), export to the cytoplasm
- Rate of degradation

Regulation of transcription initiation



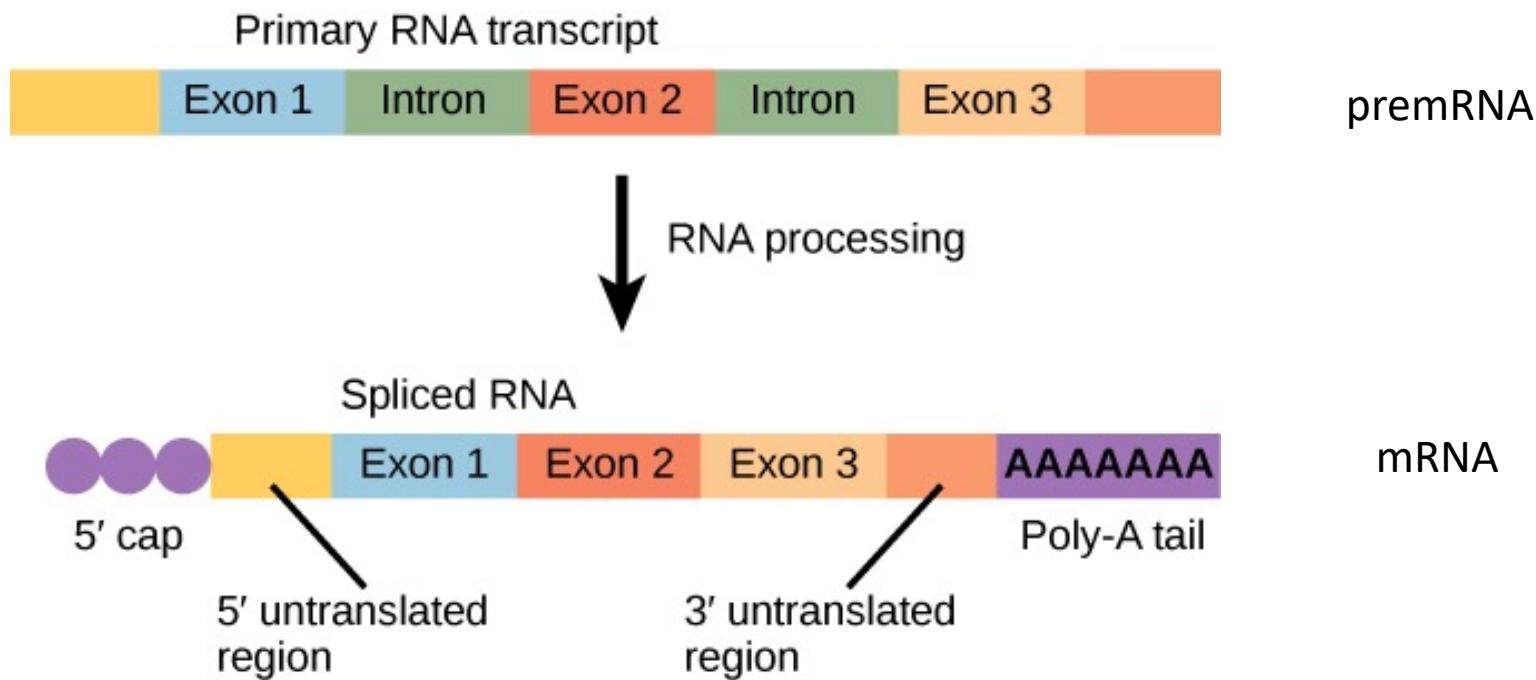
- **RNA polymerase** is the enzyme that **transcribes** the DNA into RNA
- The **promoter** is where RNA pol. gets started - has recognition sites (e.g., “TATA box”) that guide initiation
- All the other bits are things that can influence if the gene is “on”
 - This is very generic and over-simplified – 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).

Transcription



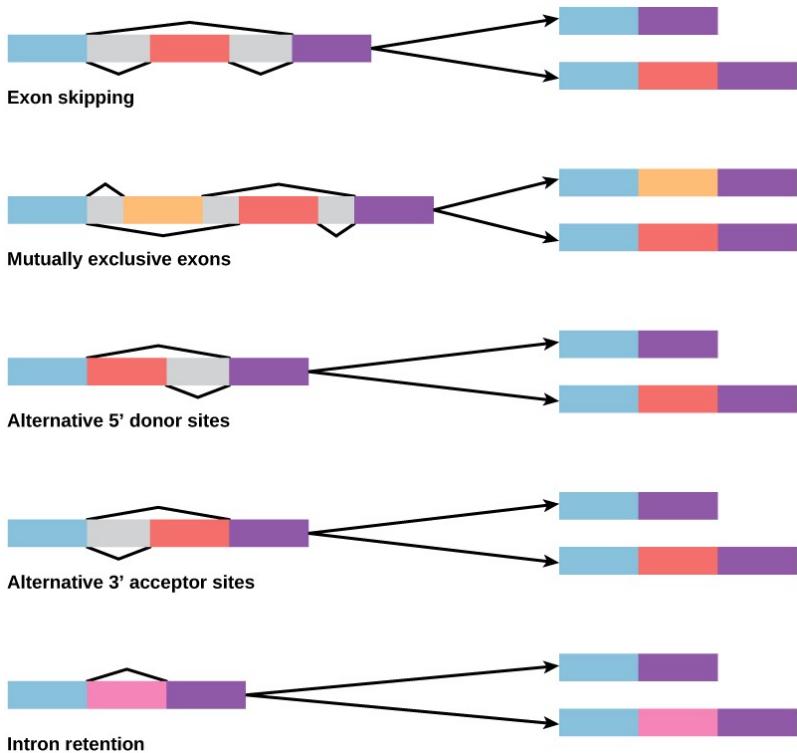
- The place where transcription begins is called the “transcription start site”
- Note that the **template** strand that is copied is the one **complementary** to the strand we read in the genome browsers, so the RNA has the sequence of the “nontemplate” strand (top strand here). The “non-template” strand has the “coding sequence”. Can be confusing!

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).

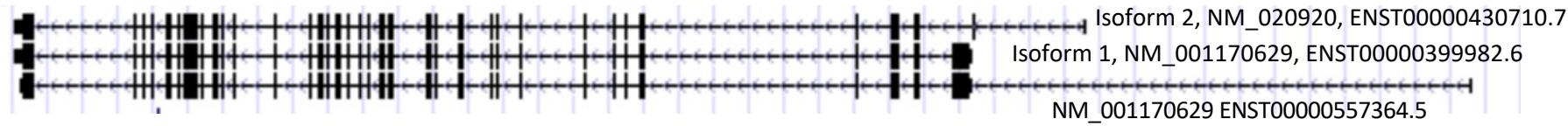
Alternative splicing



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.
- But 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 3; Ensembl lists 14
 - But 7 don't encode a protein (no open reading frame); 4 don't encode a complete protein (no stop codon); and 2 encode the exact same protein
- So perhaps **up to** two different 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**

NCBI: <https://www.ncbi.nlm.nih.gov/gene/57680>

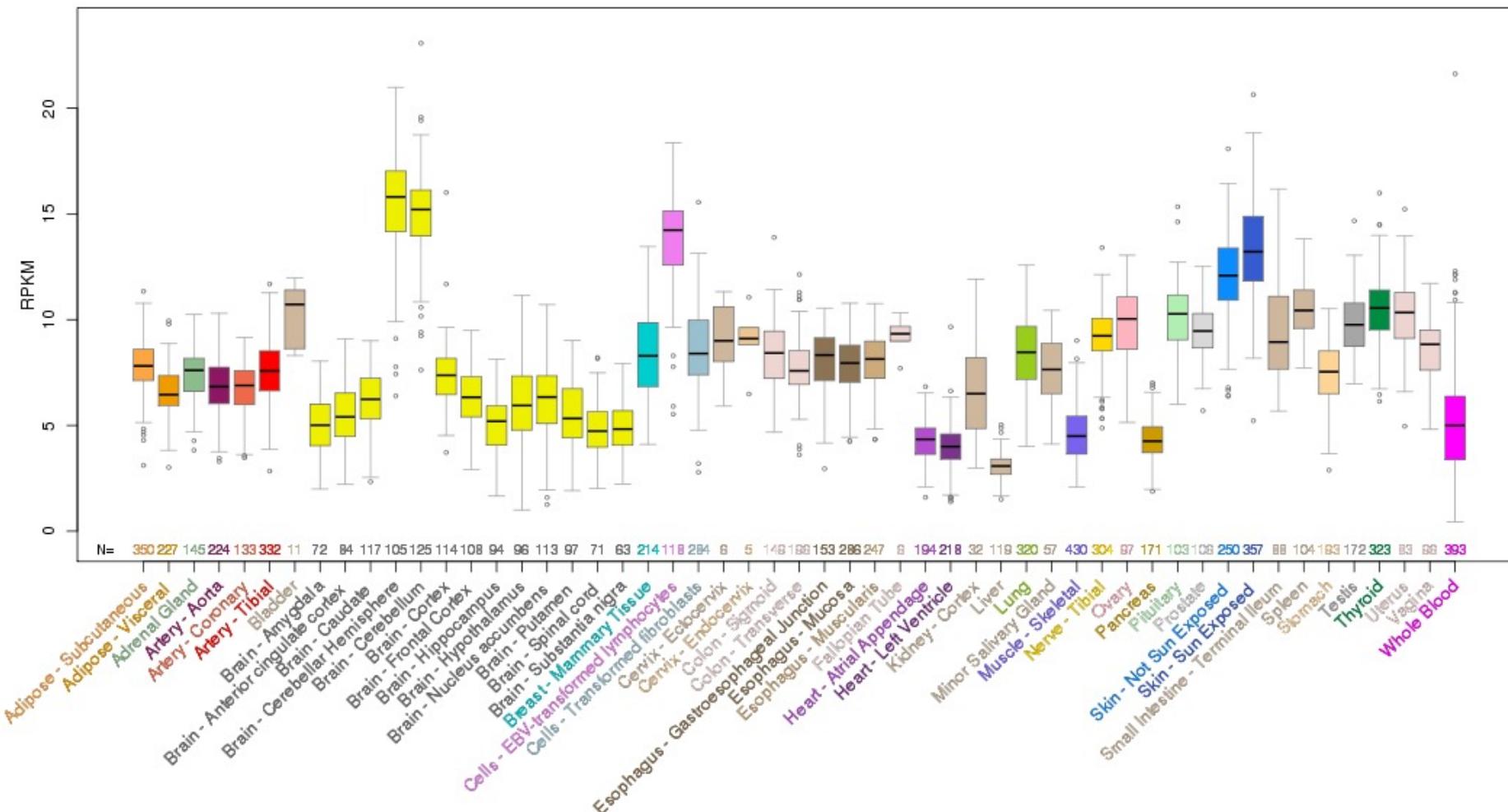
Ensembl: https://uswest.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG00000100888;r=14:21385194-21456126

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

A CHD8 mRNA

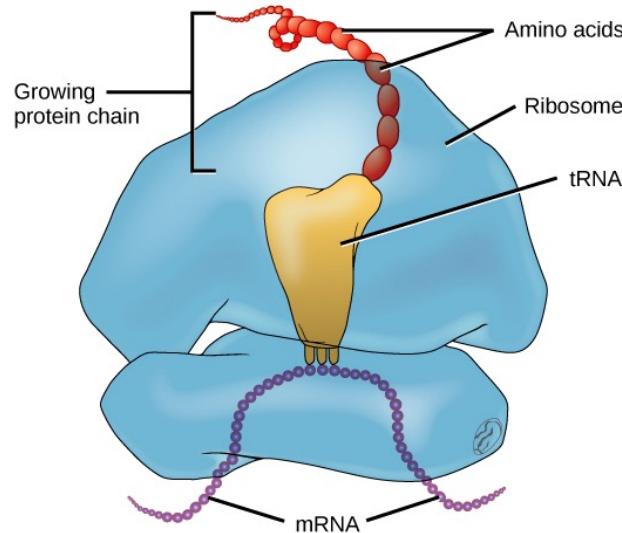
Expression profile for CHD8 across adult tissues

ENSG00000100888.8 Gene Expression from GTEx (Release V6)



- 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

Translation: making proteins



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

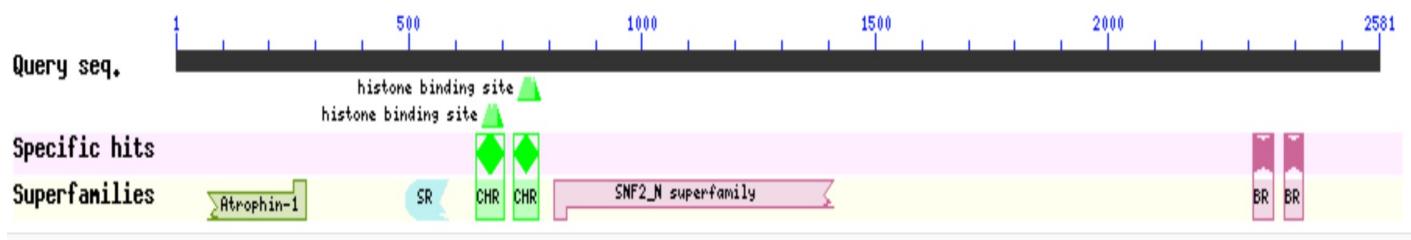
Note: RNA can be functional by itself: not every gene encodes a protein – tRNA is one example, and ribosome itself is partly made of RNA

CHD8 protein

>Q9HCK8
MADPIMDLFDDPNLFGLDSLTDSDFNQVTQDPIEEALGLPSSLSDLQMNQDGGGGDVGN
SSASELVPPEETAPTELSKESTAPAPESITLHDYTTQPASQECPAQPVLTQSTPTSGLL
QVSKSQEILSQQNPFGMSATAVSSSSAGGQQPQSAPKIVILKAPPSSVTGAHVAQIQA
QGITSTAQPLVAGTANGGKVFTKVLGTPLRGPSIVSGNTVLAAKVPGNQAAVQRIVQ
PSRPPVKQLVLQPVKGSAPAGNPAGTGPPLKPAVTLTSTPTQGESKRITLVLQPOQSGGPQ
GHRHVVVLGSLPGKIVLQGNQLAALTQAKNAQGQPAKVVTIQLQVQOPQQKIQIVQPQPPSS
QPQPQQPPSTQPVTLLSVQQAQIMGPQSPGQRLSVPVKVVLQPQAGSSQGASSGLSVVK
VLSASEVAALSSPASSAPHSGGKTGMEEENRRLEHQKKQEAKANRIVAEAIARARARGEQNI
PRVLNEDELPSPVRPEEEGEKKRKKSAGERLKEEKPKSKTSGASKTKGSKSLNTITPVV
GKRRKRNTSSNDSDVEVMPAQSPREDEESSIQKRRSRNRQVKRKYTEDLDIKITDDEEEE
EVDTVTPGIKPEPILPEPVQEPDGETLPSMQFFVENPSEEDAAIVDKVLSMRIVKKEPLSG
QYTEAEFFVVKYKNYSYLHCEWATISQLEKDKRIHQKLKRFKTKMAQMRRHFFHEDDEPFN
PDYVEVDRILDESHSIDKDNGEPIYYLVWKWCSLPYEDSTWELKEDVDEGKIREFKRIQS
RHPELKRVNRQASAWKKELESHEYKNRNLREYQLEGVNWLFWYNRNQNCILADEMGL
GKTIQSIAFLQEVNVGIGHGPFLVIALSTITNWEREFNTWTMNTIVYHGSLASRQMIQ
QYEMYCKDSRGRLLIPGAYKFDALITTEFILSDCPELREWRCVIIDEAHRLKRNRCNL
LDSLKHMDLEHKVLLTGTLQNTVEELFSLLHFLEPSQFSESEFLKDFGDLKTEEQVQK
LQAILKPMMLRRLKEDVEKNLAPKQETIIEVELTNIQKKYRAILEKNFSFLSKGAGHTN
MPNLLNTMMELRKCCNHPYLINGAEEKILTEFREACHIIPHDFHLQAMVRSAGKLVLIDK
LLPKLKAGGHVKVLIFSQMVRCLDILEDYLIQRRRYLERIDGRVGRGNLRQAAIDRFSKPDS
DRFVFVLLCTRAGGLGINLTAADTCIIFDSDWNPQNQLQAQARCHRIGQSKAVKVYRLITR
NSYEREMFDKASLKLGLDKAVLQSMSGRDGNITGIQQFSKKEIEDLLRKAYAAIMEEDD
EGSKFCEEDIDQILLRRTTTITIESEGKGSTFAKASFVASENRTDISLDDPNFWQKWA
ADLDM DLLNSKNNLVIDTPVRKQTRHFSTLKD LVEFSDLESEDDEPRSRHRRHHAA
YGR TDCFRV EKHLLV YGWGRWRDILSHGRFKRRMTERDVETICRAILVYCLLHYRGDENI
KGFIWDLISPAENGKTKELQNHSGL SIPVPRGRKGKKVKSQSTFDIHKADWIRKYNPD
FQDESYKKHLKHQCNCNKVLLRVRMLYYLQRQE VIGDQAEKVLGGAIASEIDIWF
PTTWWDSEADKSLLIGVFKHGYEKYNTMRADPALCFLEKAGR PDDKAIAAEH
RVLDNFSD IVEGVDFDKCDEDPEYKPLQGPPKDQDEGDPLMMDEEISV
IDGDEAQVTQQP GHLFWP
PGSALTARLRLVTAYQRSYKREQM KIEAAERGDRRRRCEAFKLKEIARREK
QQRWTR REQTDFYRVVSTFGVEYDPDTMQFHWD RFRTFARLKKTDES LTKYFHGF
VAMCRQVCRL PPAAGDEPPDPNL FIEPITEERASRTLYIELLRLRREQVLCHPLLED
R LACQPPGP ELPKWWEPV RHDGELL RGAARHGV SQTDCNIMQDPDF SFLA
ARMNYM QN H QAGA PAPS LSRC STPL LHQQY TSRT ASPLPLR DPAP
VEK SPEET ATQVPSLES LT KLEHEV VARS RPT P QD
YEMRVSPSDTPLV SRVPPV KLE DEDDS DSEL DL SKL SP
SSSSSSSSSSSSSSST DESED EKEEKL TDQSR SKL YDEE SLLS
LTMS QDGFP NEDGE QMTP ELLL L QER QRA SEWP
KDR VL INR IDL V CQAVL SGS KWP SRRS QEMV
TGG I L GP GHN LLDSPS LTP GEY GDSP VPT PRSS
AASMAEEF AAS AVST AAAQFT KLR GMDE KEFTV QIK
DEE GLK LTF QK HKLM VGMD GH
PLF HKKKGNRKL V LEV E CME EP NHLD V DLE TRIP VINK
V DGT LLV GED A PRA ELEM W
LQGHPEFAV D P RFLAY M ED RR KQK W QR CK NNKA ELN
CL GM EP V QT AN S RN GKK GH HTET
VFNRVLP GP IA PESS KKR ARR MR PDL SK MM AL MQGG
ST GS LS LH NT FQH SSS GL QSV SSL
GH SSAT S AS LPF MPF FVM GGAP S P HVD S TSML
HHHHHHH PHPHHHHHH PG L RAP GYP PSS P
VTTASGTT LRLPPL QPE E DD DE E D D D L S QG YD
SER DF SL I DD P M P M PAN SD S E AD 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
 - Occurs in a non-functional part of the 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)

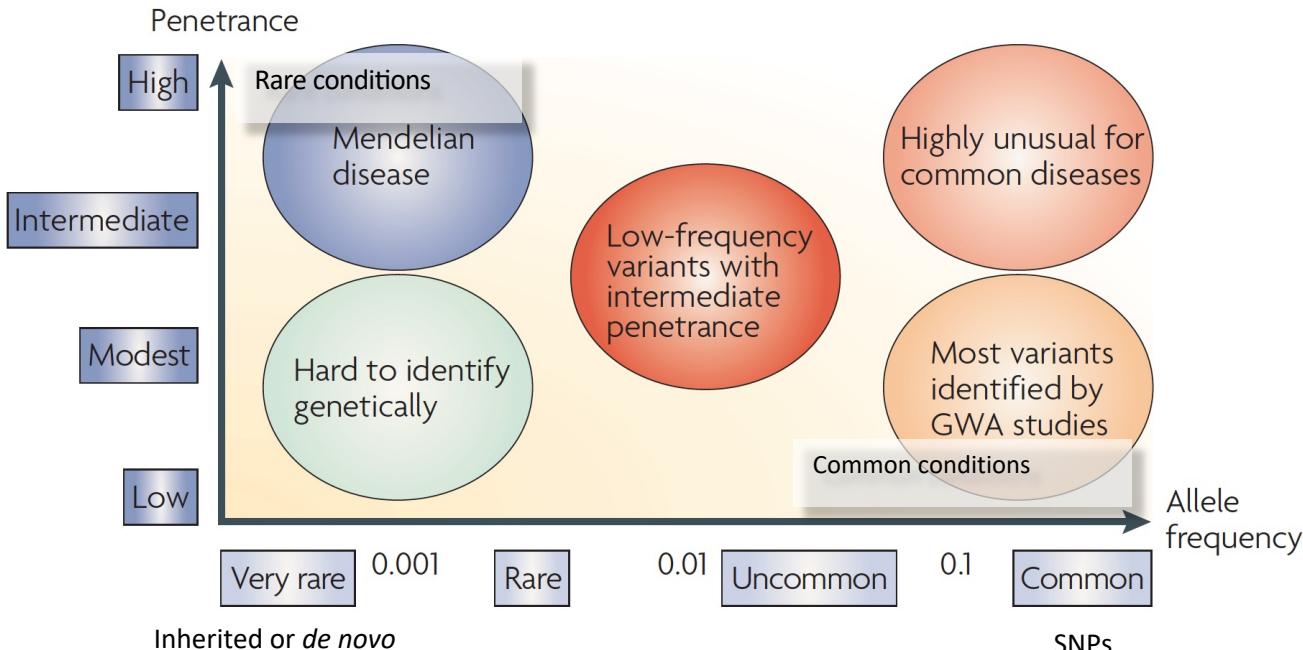
- A base that varies across individuals
- By convention, the term “**single nucleotide polymorphism**” (SNP) is used for **common** variants - the less common (“minor”) allele is in >1% of the population (e.g., 15% A, 85% C are the allele frequencies)
 - A source of differences among people, but usually 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

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 Francescato,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

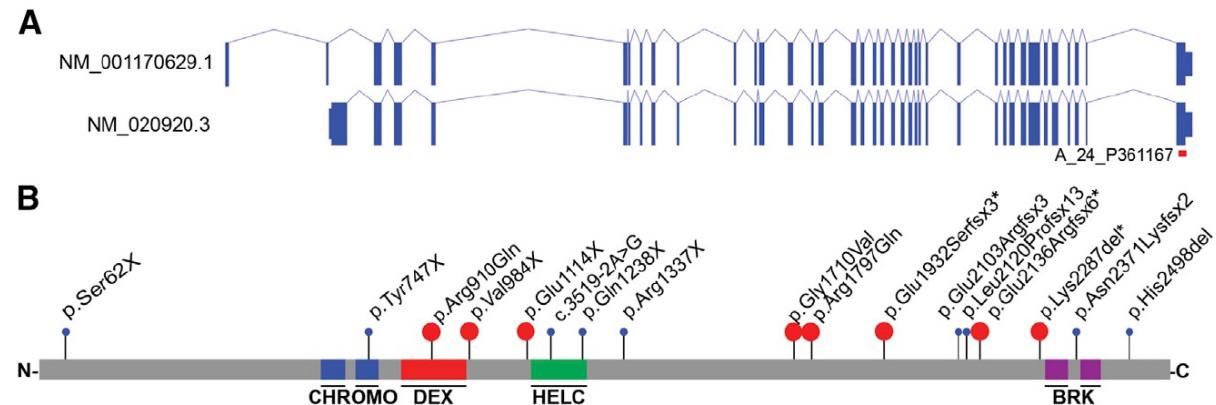


Figure 1. Spectrum of *CHD8* Mutations in Autism Spectrum Disorder

(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

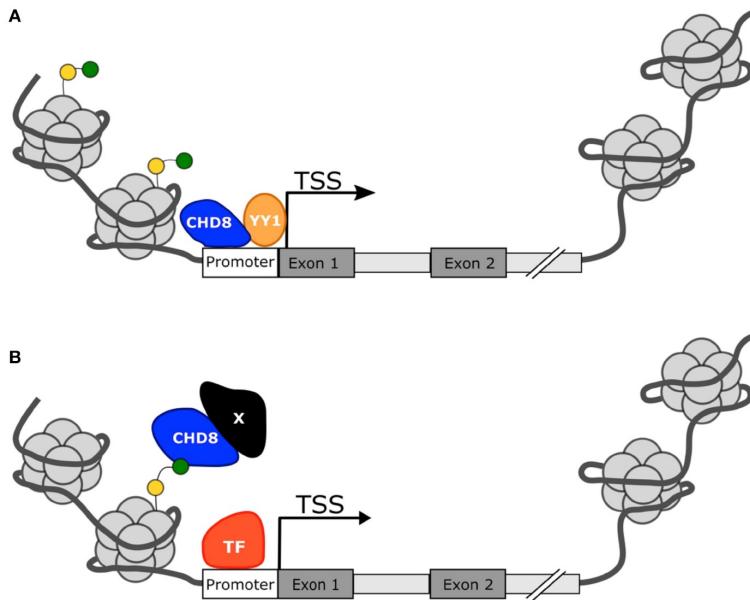
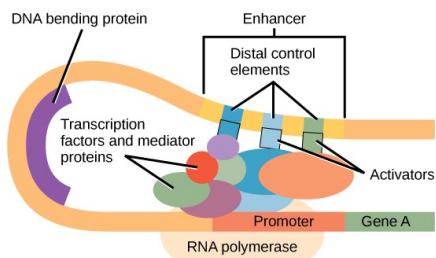


FIGURE 3 | Proposed mechanisms for CHD8 transcriptional activation. **(A)** CHD8 is most commonly found near active transcription start sites with histone modifications H3K4me3 (green circle) or H3K27ac (yellow circle). CHD8 may directly activate genes by directly binding near the transcriptional start 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 accessible.

<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?

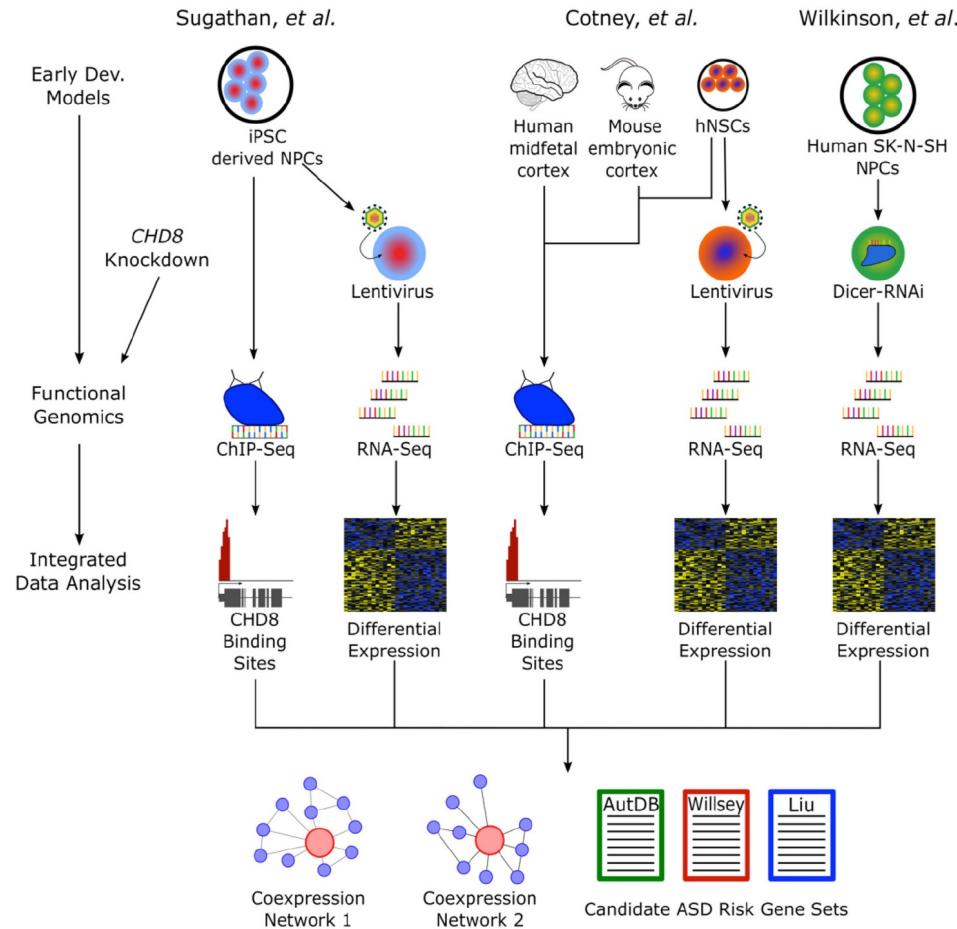


Figure is from a 2015 review, summarizing three studies:

- 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 ...