03-621 Week 3

Advanced Quantative Genetics

Aidan Jan

January 30, 2026

Mutations (Continued)

- Synonymous Substitutions Can Alter Phenotypes

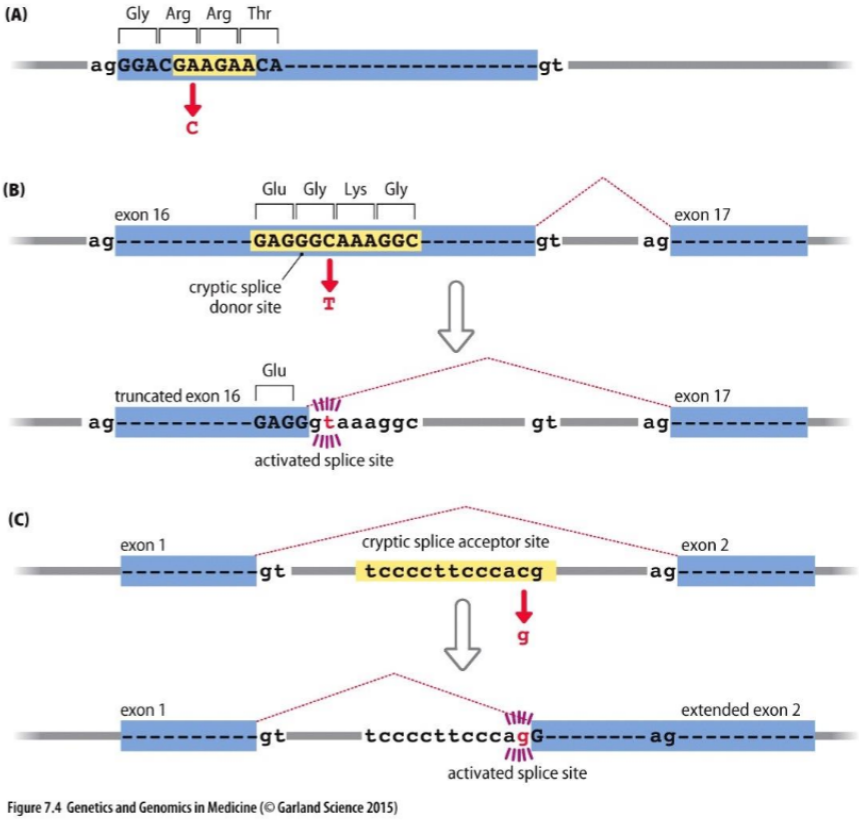


Figure 7.4 Genetics and Genomics in Medicine (© Garland Science 2015)

- Single-gene disorders can also be caused by mutations in **non-coding** RNA genes
- Large Copy Number Variations (CNVs) are common in the human genome
  - One section of DNA is duplicated many times
  - About 10-15% of the genome displays copy number variation
  - 1000 bp - 5 mbp in length
  - Most are in non-coding regions, but some contain genes
  - Too large for PCR analysis, can be detected using a microarray

- Variable nucleotide tandem repeats (VNTR) and short tandem repeats (STR) are inherited repeating stretches of DNA
  - Different individuals may have different numbers of repeats at a given locus.
  - If the number of base pairs in a repeated section is not a multiple of three, tandem repeats may cause frameshift errors.
- Some mutations that cause disease do not change the sequence of a gene product but alter the amount of gene product
  - For example, mutations in promotors or enhancers
  - Mutations in transfactors (e.g., histones)
  - Gene duplication or deletion

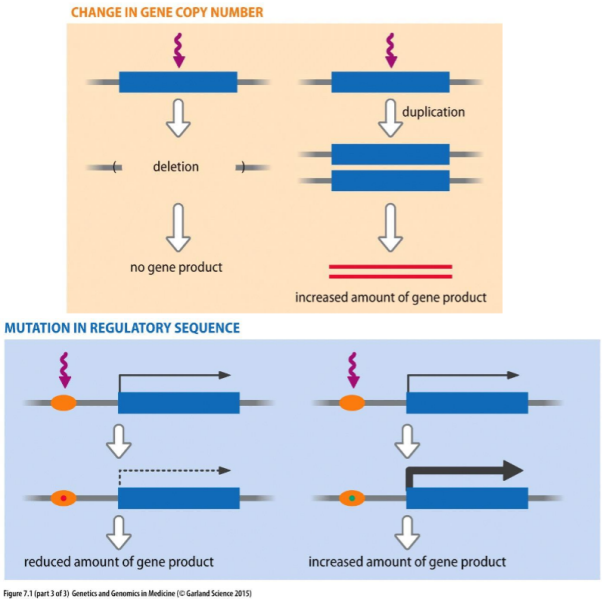


Figure 7.1 (part 3 of 3) Genetics and Genomics in Medicine (© Garland Science 2015)

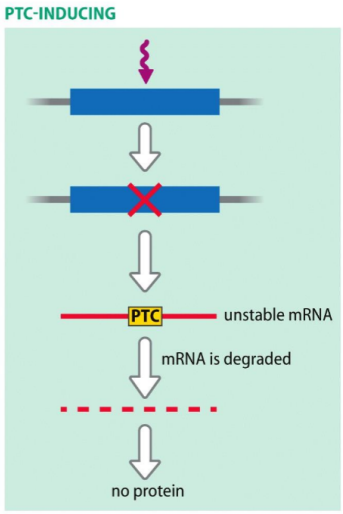
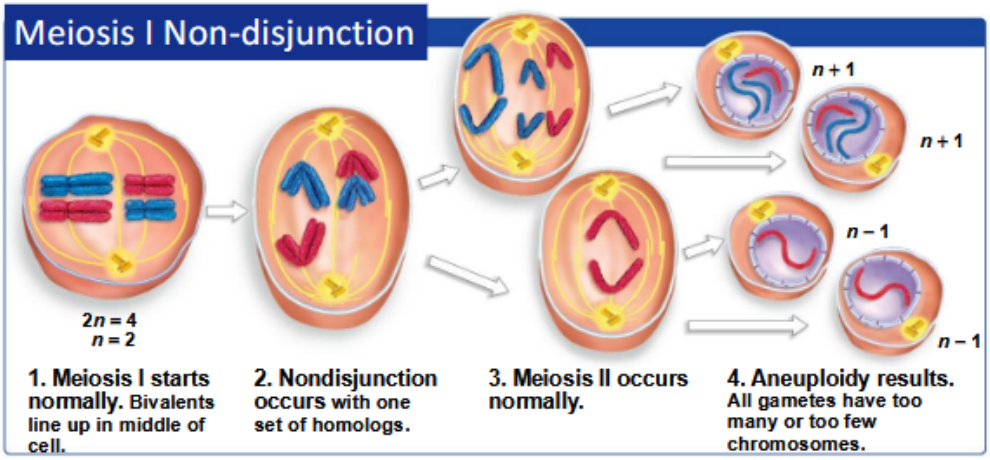


Figure 7.1 (part 1 of 3) Genetics and Genomics in Medicine (© Garland Science 2015)

- Gene dosage changes via Meiotic Non-disjunction
  - Frequency: in all recognized pregnancies
  - 8% of *recognized* pregnancies have major chromosomal abnormalities. $> 94\%$ of this 8% undergo spontaneous abortion.
  - Aneuplodies change the amount of gene product expressed ("gene dosage"), across many genes.
  - Among the fetuses in 100000 recognized pregnancies, about 8000 have major chromosomal abnormalities, 7500 of these undergo spontaneous abortion, and 500 are born alive.

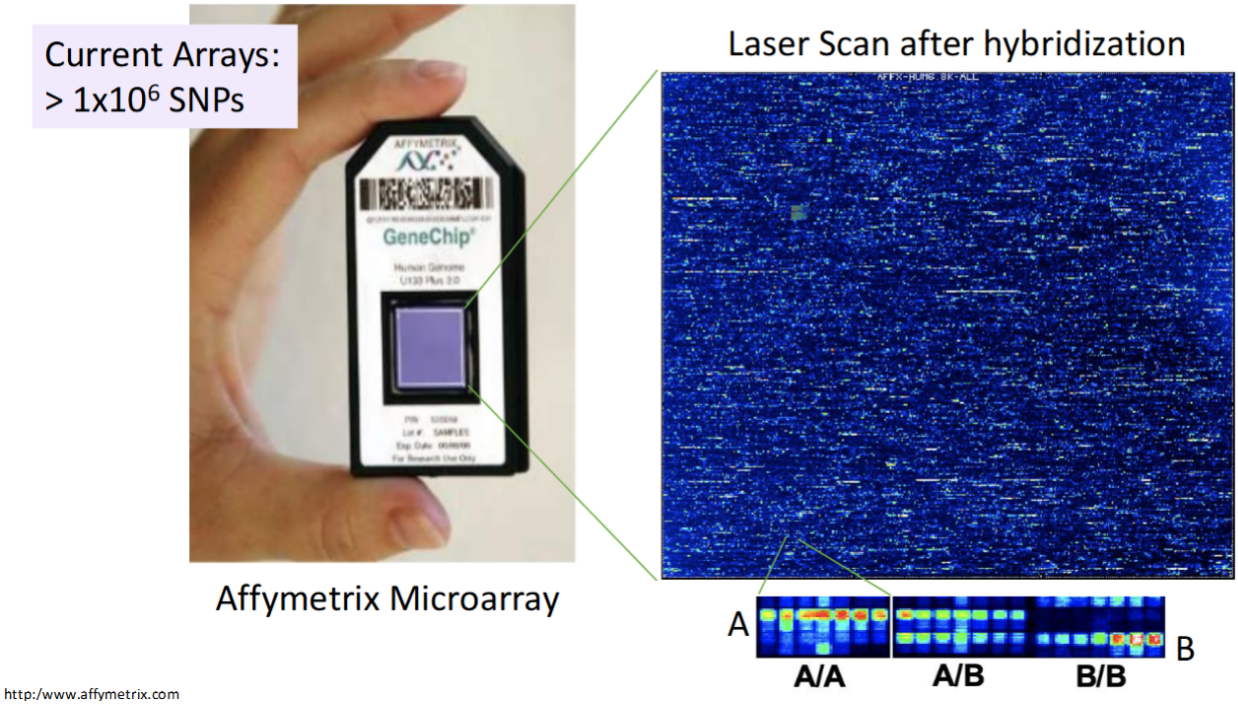


Analyzing Sequence Variation

What methods are used for analysis of sequence variation?

1. For **genotyping** large numbers of individuals:
   - DNA microarray hybridization
   - PCR (and gel electrophoresis)
   - Sequencing targeted regions
2. For variant **discovery**
   - Whole genome sequencing (WGS)

Detection of SNPs using microarray hybridization



http:/www.affymetrix.com

Microarrays can also be used to detect CNVs

The fluorescence intensity of the microarray hybridization signal increases with copy number

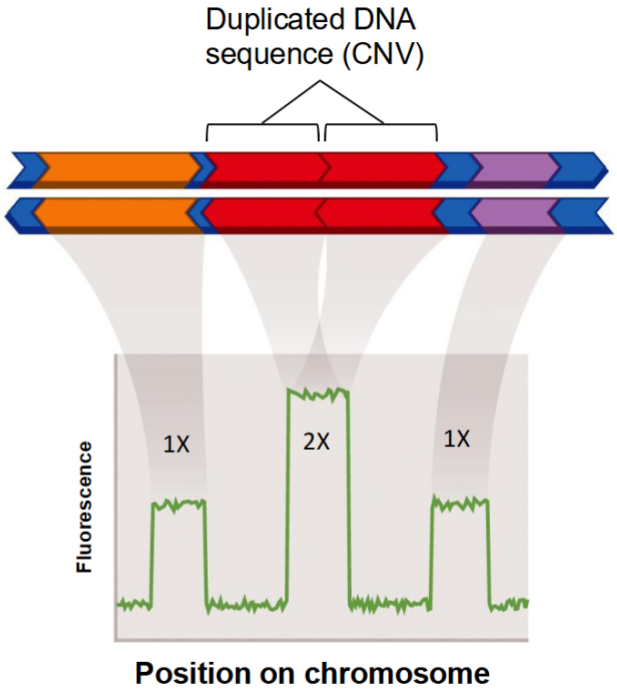


Figure 15.11
Biology: How Life Works
© 2014 W. H. Freeman and Company

VNTRs and STRs can be genotyped easily by PCR

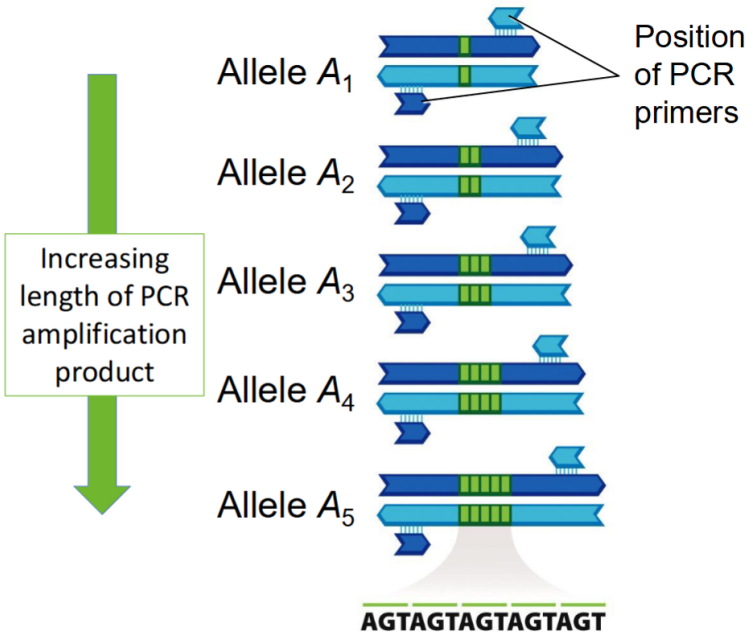


Figure 15.5a
Biology: How Life Works
© 2014 W. H. Freeman and Company

Size differences of the PCR products reveal repeat polymorphism genotypes

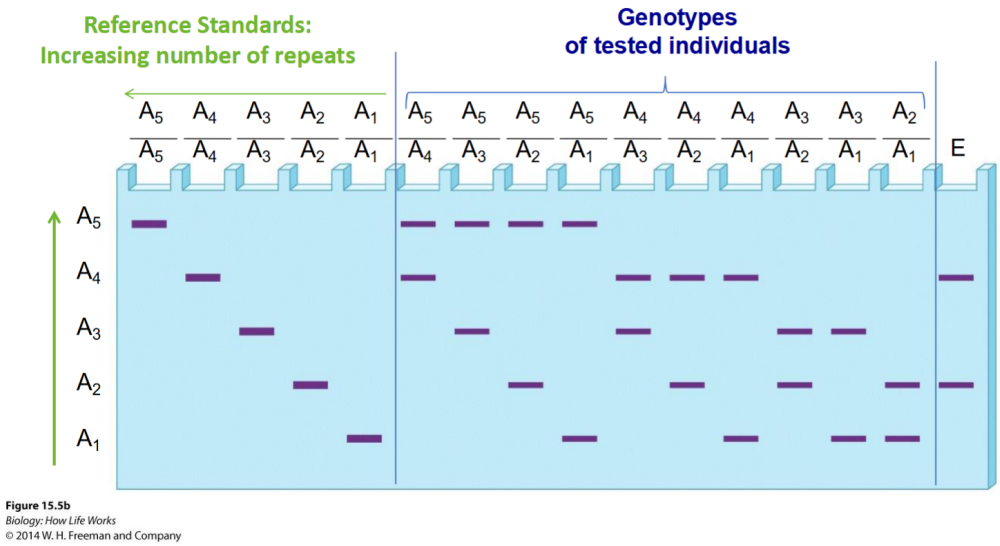


Figure 15.5b
Biology: How Life Works
© 2014 W. H. Freeman and Company

Some Current Sequencing Technologies

	Method	Read Length
dideoxy-nucleotide termination method	**Sanger**	600-1,000 bp
"2nd Generation Sequencing"	**454**	400-800 bp
	Illumina	75-600 bp
"3rd Generation Sequencing"	**PacBio**	10,000 – 25,000 bp
	Oxford Nanopore	10 Kb – 4 Mb! (many errors)

Decreasing cost per base
Increasing cost per run

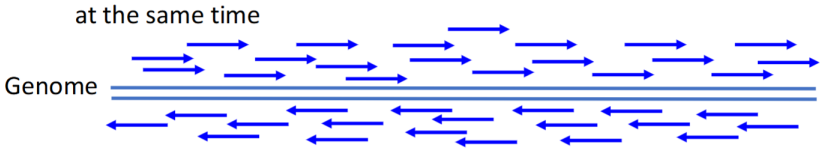
Increasing throughput (speed)

Decreasing accuracy

No method can sequence an entire genome in one reaction!

Shotgun Sequencing Approach

1. Break up the genome and sequence many small random fragments at the same time



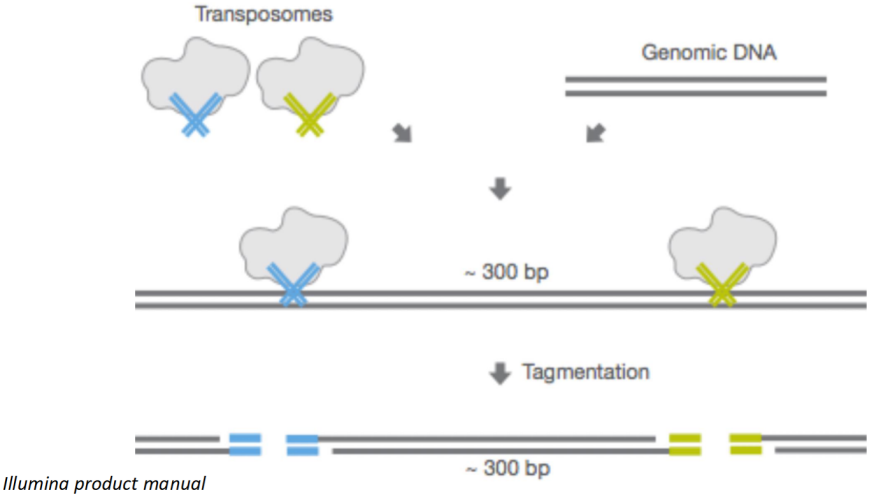
(initially we don't know where each read comes from on the genome)

2. Assemble the random reads into the complete, continuous sequence by searching for overlaps computationally, and generate a consensus

```
                                ATTGTTCCCACAGACCG
                        CGGCGAAGCATTGTTCC      ACCGTGTTTTCCGACCG
                ACCTCGATGCCGGCGAAG  TTGTTCCCACAGACCGTG TTTCCGACCGAAATGGC
                        ATGCCGGCGAAGCATTGC     ACAGACCGTGTTTCCCGA
TAATGCGACCTCGATGCC     AAGCATTGTTCCCACAG     TGTTTTCCGACCGAAAT
                TGCCGGCGAAGCGTTGT                      CCGACCGAAATGGCTCC
```

Consensus:
TAATGCGACCTCGATGCCGGCGAAGCATTGTTCCCACAGACCGTGTTTTCCGACCGAAATGGCTCC

Adding *known* sequence adapters for high-throughput sequencing by transposon insertion



Illumina product manual

Sequencing the Genomes of Trios

We can detect new mutations by sequencing a child plus both parents, and comparing their genomes.

- How do we tell whether a mutation arose in the germline of the father or mother?
- We can look at linked genes. We try to find another SNP nearby, and see if we can match one of the strands to one of the parents.
- There tend to be a lot more mutations (10x) in the open reading frame (ORF), and even more in mitochondria.
- Mutation rates differ among genes. Some genes mutate orders of magnitude more than others.

The Human pathogenic allelic load is not well known yet

Estimates from the 1000 genomes project:

- 400 damaging DNA variants per individual
- 100 severe loss-of-function mutations
  - 20 genes homozygous for inactivated allele (null)
  - 60 missense variants that damage protein structure

Databases

Genomewide databases:

- Human Gene Mutation Database: comprehensive data on germline mutations in nuclear genes associated with human inherited disease. (http://www.hgmd.org)
- COSMIC: comprehensive catalog of somatic mutations in cancer. (http://www.sanger.ac.uk/genetics/CGP/cosmic/)
- MITOMAP: mitochondrial genome database with prominent sections devoted to disease-associated mutations in mt-tRNA, mt-rRNA, and mitochondrial coding and regulatory sequences. (http://www.mitomap.org/)

Locus-specific databases:

- Phenylalanine Hydroxylase Locus Knowledgebase: a list of mutations at the *PAH* locus, mostly centered on mutations causing phenylketonuria. (http://www.pahdb.mcgill.ca)

Databases by mutation category:

- SpliceDisease Database: disease-associated splicing mutations. (http://cmbi.bjmu.edu.cn/sdisease)

Databases of human genetic variation

- dbSNP: SNPs and other short genetic variations. (http://www.ncbi.nlm.nih.gov/SNP/index.html)
- dbVar: genomic structural variation. (http://www.ncbi.nlm.nih.gov/dbvar/)
- DGV: genome structural variation. (http://dgv.tcag.ca/)
- ALFRED: allele frequencies in human populations. (http://alfred.med.yale.edu/alfred/index.asp)

Summary

- Whole genome sequencing is becoming cheaper and faster: practical for large numbers of individuals
- Targeted sequencing of specific regions
- Other high throughput methods for detection of molecular markers

Methods besides sequencing:

- Single nucleotide polymorphisms (SNPs) ←Microarray hybridization
- Variable number of tandem repeats (VNTPs) ←PCR + gel electrophoresis
- Copy number variation (CNV) ←Microarray hybridization

Making Sense of the Human Genome

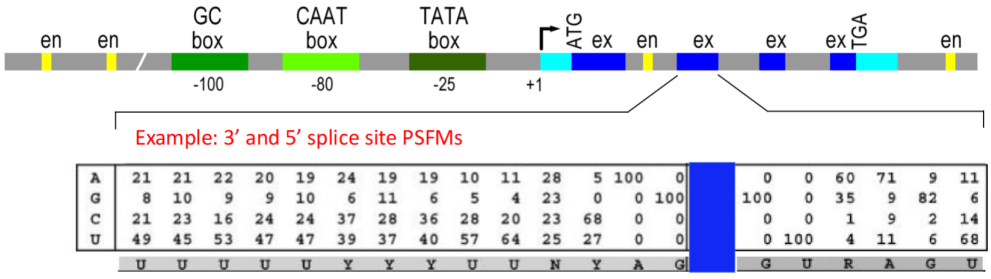
"Annotation" to describe the genome contents

1. Computational gene prediction
2. Functional assays:
   - RNA sequencing to identify transcribed regions
   - Many other experimental approaches (details later)
3. Comparative Genomics:
   - Evolutionary conservation of function regions

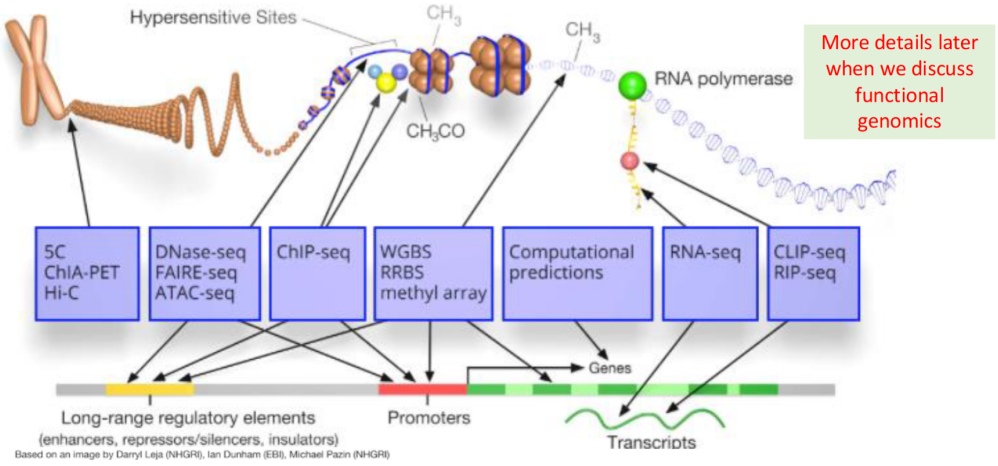
1. Computational Gene Prediction

Goal: find protein coding and RNA genes in the genome. What do we look for?

- e.g., for protein-coding genes:
  - search for open reading frames ("ORFs")
  - search for matches to models of regulatory elements
- Shortcomings: high error rates, missed genes or exons, incorrect gene models

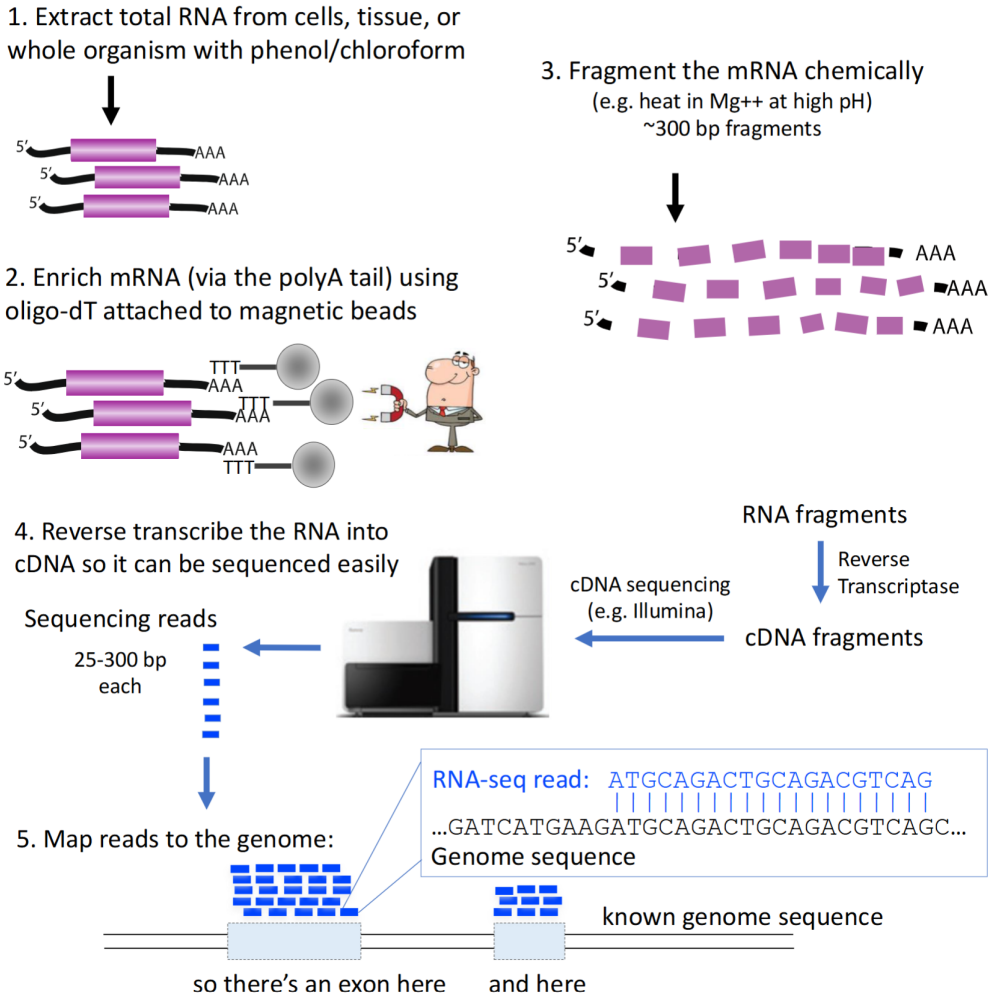


2. Functional assays: ENCODE project



Still incomplete - ongoing, but lots of data already.

Example: mRNA-sequencing (mRNA-seq) to map exons



ENCODE project current outcomes

Gene Class	Number
Protein-coding genes	19379
RNA genes	27599
Long non-coding RNA genes	19993
Small non-coding RNA genes	7566
Pseudogenes	14737
Processed pseudogenes	10633
Unprocessed pseudogenes	3571
Other pseudogenes	367

What have we learned?

- An average protein coding gene produces on average:
  - 6 different RNA transcripts
  - Out of the 6, 4 contain protein coding sequences
  - The other 2 are non-coding transcripts
- How are different transcripts produced?
  - Alternative splicing
  - Alternative transcription start sites
  - Alternative polyadenylation sites
- 75% of genome is transcribed in at least one cell type
  - But only 1.2% of genome is protein-coding!
  - Not all 75% transcribed in a single cell
- RNA transcripts from **both** strands are common
  - Within genes
  - Within intergenic regions
- Caveat: Much of this RNA is produced at very low levels that may just be background noise or a result of sloppy transcription regulation.
- Functional Biochemical signals analyzed so far:
  - Binding sites for 119 of 1800 known transcription factors
  - 20% of chemical modifications of DNA or histones (epigenetic markers)
  - Positions occupied by histone variants
  - Patterns of chromatin 3D structure
  - 95% of genome lies within 8kb of a DNA-protein interaction!

3. Comparative genomics (Evolutionary Conservation)

Identifying genes and other functionally important DNA elements through evolutionary conservation.

- What's the rationale?
- Functional sequences should be constrained and thus conserved over evolution due to **purifying selection** (aka. negative selection)

This allows us to:

- Confirm/refine gene models
- Discover missed genes and elements
- Evaluate functional significance

Comparative Genomics Tools

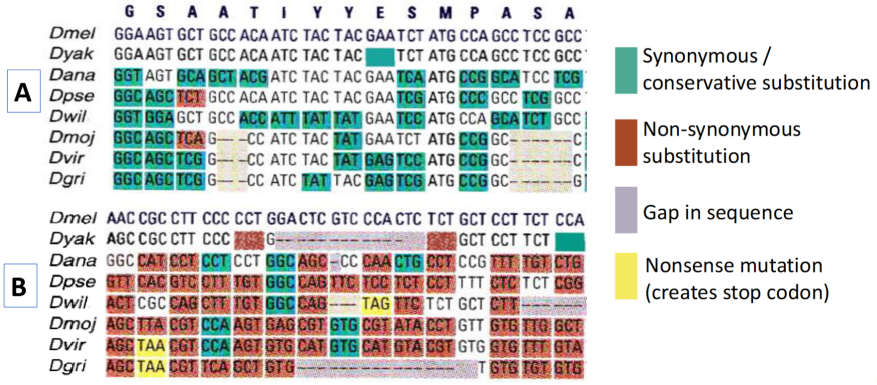
Program / Database	Internet Address	Description
BLAST programs	http://blast.ncbi.nlm.nih.gov	suite of programs for comparing query nucleotide sequences or amino acid sequences against each other or against databases of recorded sequences
BLAT	http://genome.ucsc.edu	for rapid matching of a query sequence to a genome; available through the University of Santa Cruz genome server
HomoloGene	http://www.ncbi.nlm.nih.gov/homologene	for a gene of interest, lists homologs found in other species

Differences in Sequence Conservation can Predict Type of Sequence Function

What is the consequence of these four different mutations in a coding sequence?

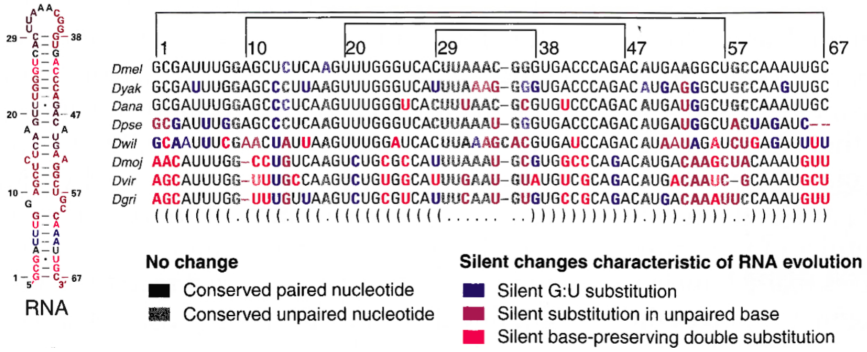
```
GCTTGCGGTGGGAGGAGGCGGCTGAGGCGGAAGGACACACGAGGC
|||  |||||||  |   ||||||||||||||||||  ||||||||||
GCTCGCGGTGGAA---GGCGGCTGAGGCGGAAG--CACACGAGGC
   ^^^    ^^^
```

From here, we can look at the following aligned sequences from different organisms. Which one is likely to have come from a protein coding sequence? Why?



In this case, A is more likely to have come from a protein coding sequence since most of the mutations are synonymous or conservative. Changing amino acids would cause damage to the organism, and choice B has too many non-synonymous substitutions that it is unlikely to come from a protein-coding sequence. If B was a protein coding sequence, organisms with that many mutations would probably die.

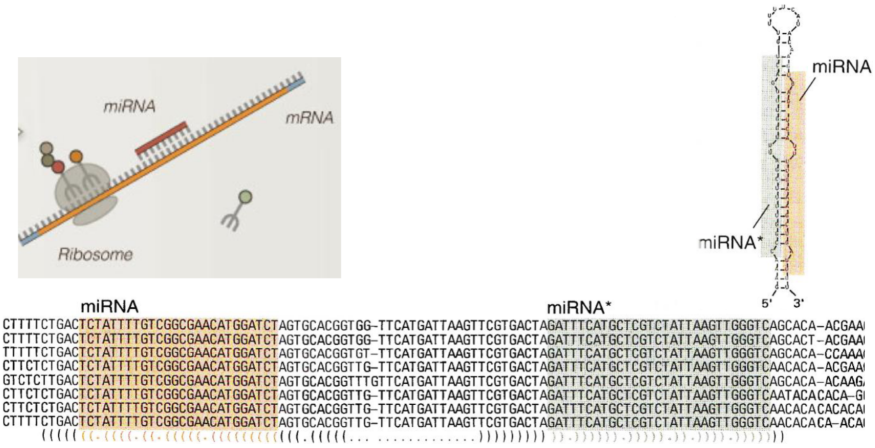
Compensatory Variation in Structural RNAs



Base-pairing is conserved, but specific sequence is not.

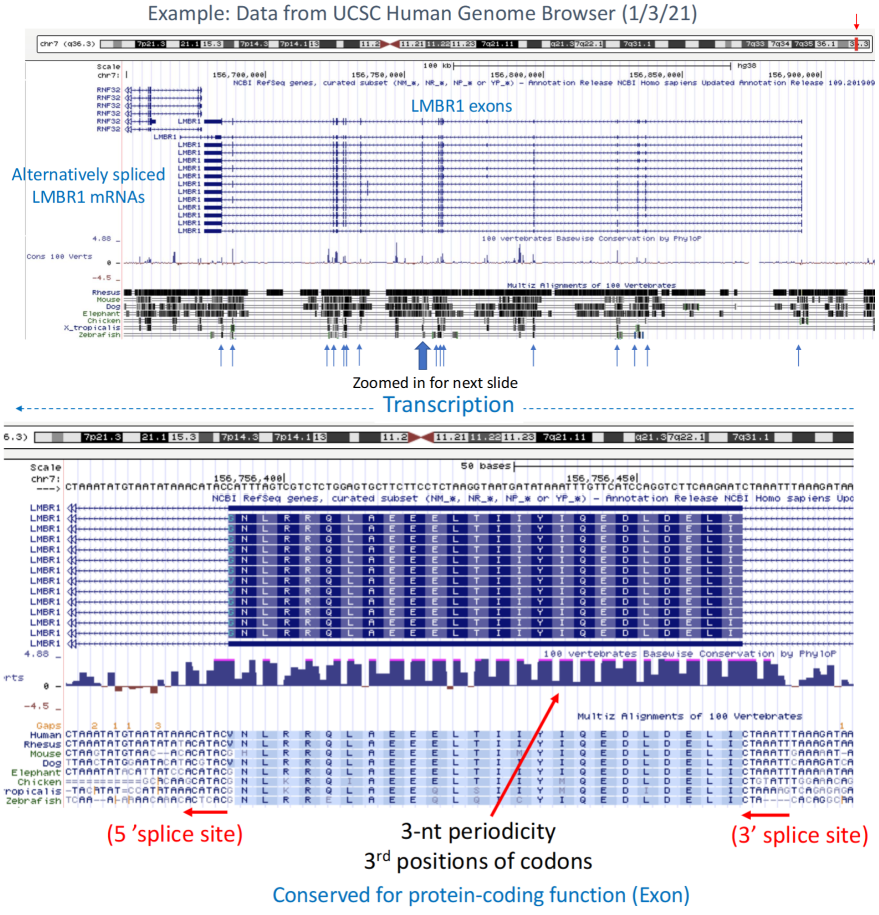
Structure and Sequence Conservation in Regulatory RNAs

For example, micro-RNAs (miRNAs)

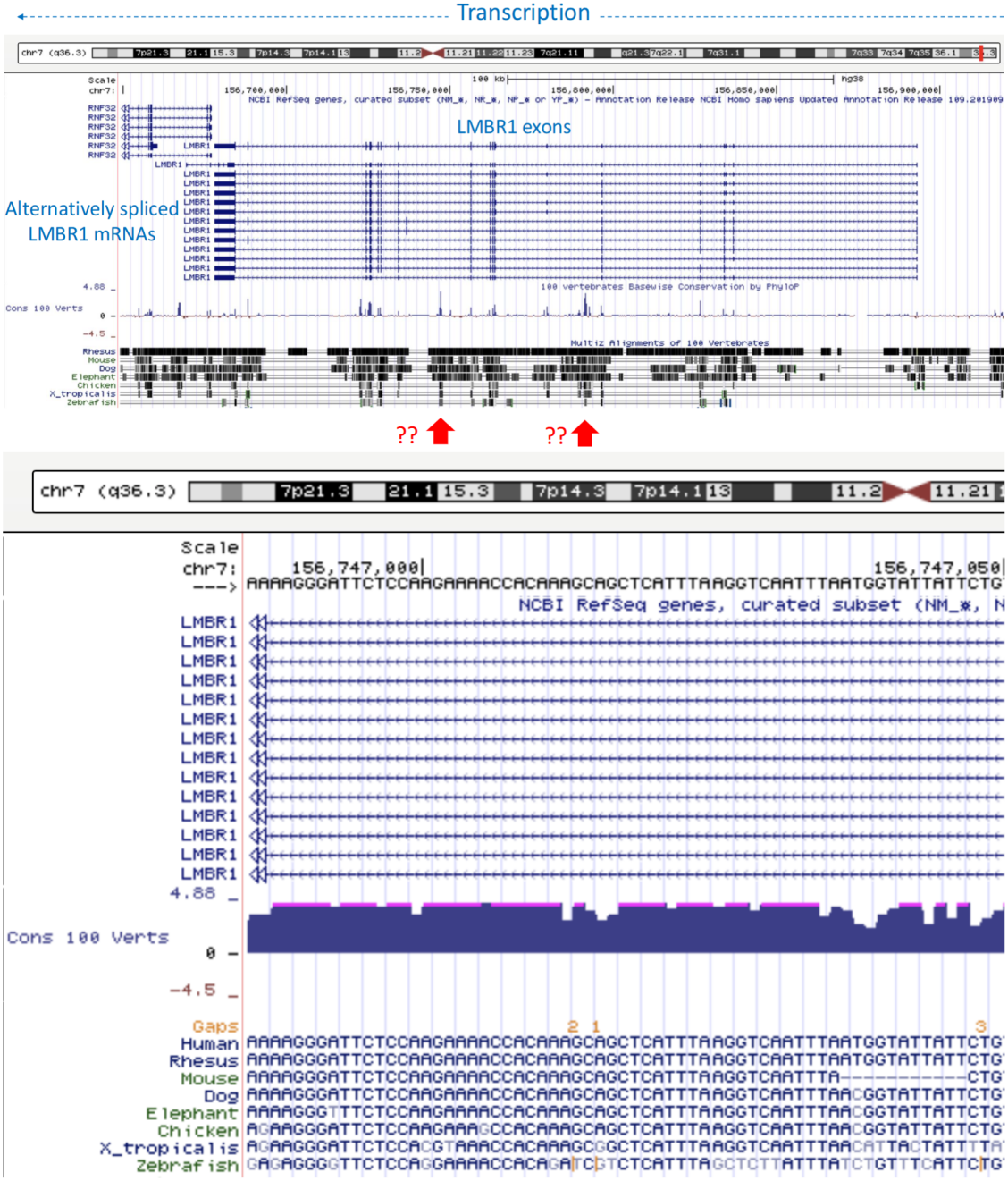


Predicting Sequence Function using Differences

Differences in sequence conservation can help predict sequence function in a **coding element** in LMBR1. For example, data from UCSC Human Genome Browser (1/3/21)

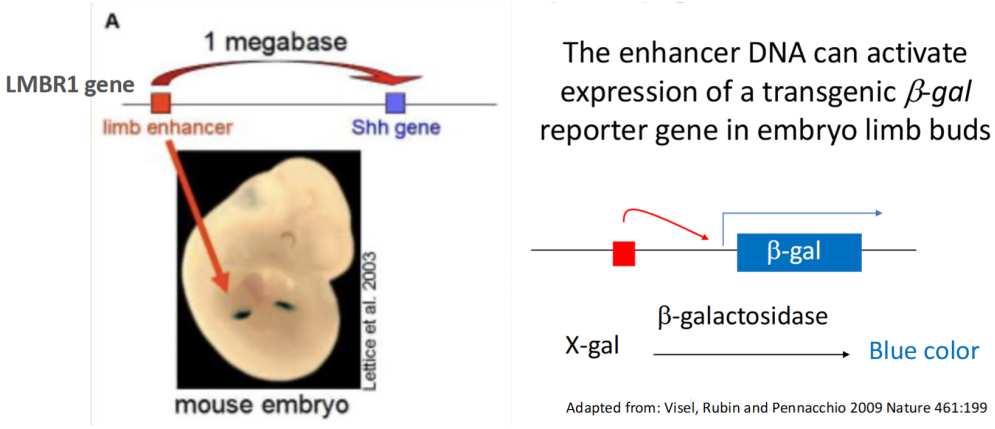


Differences in sequence conservation can also help predict sequence function in a **non-coding element** in LMBR1.

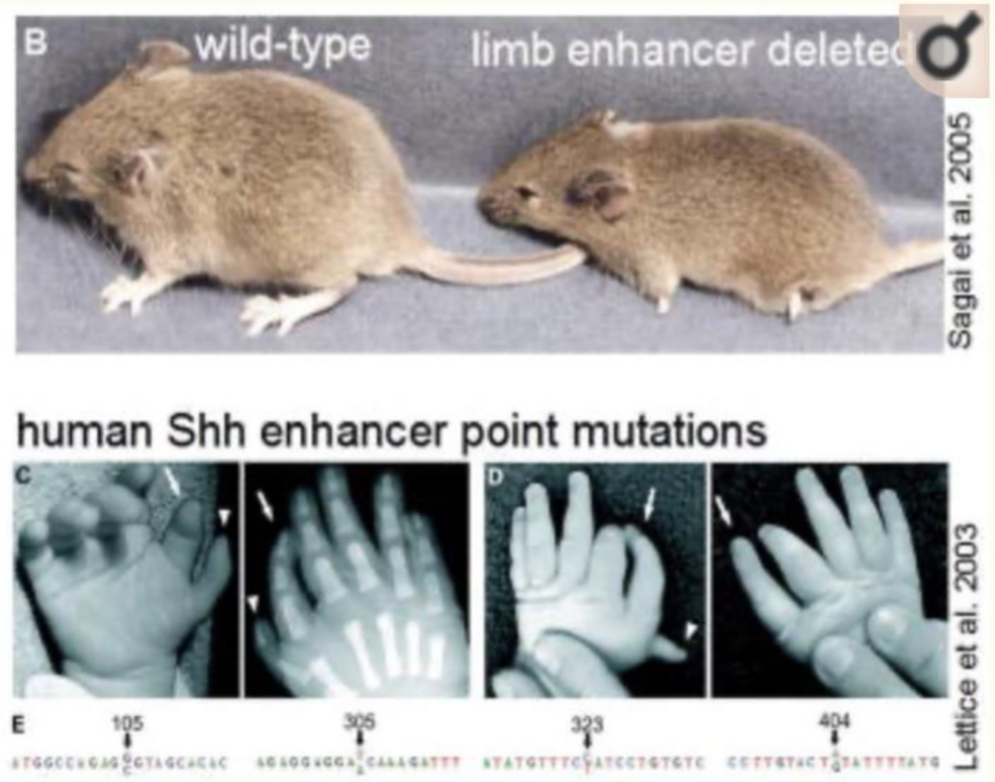


Unlike the coding region, this set has no exon detected by RNA sequencing. Additionally, there is no periodicity in the substitutions (implying mutations happen everywhere). Also, reading frames are not preserved; deletions or insertions are not necessarily in multiples of 3.

- The conserved element within an intron of LMBR1 gene is a limb-specific enhancer for transcription of the *Sonic hedgehog* gene.



Deleting or mutation of the enhancer alters *Shh* expression and affects limb development.



Overview of Comparative Genomic Results

Only 5% of genome shows strong conservation with other mammals! **Does this mean most of the genome is NOT functionally important?**

What else is in there?

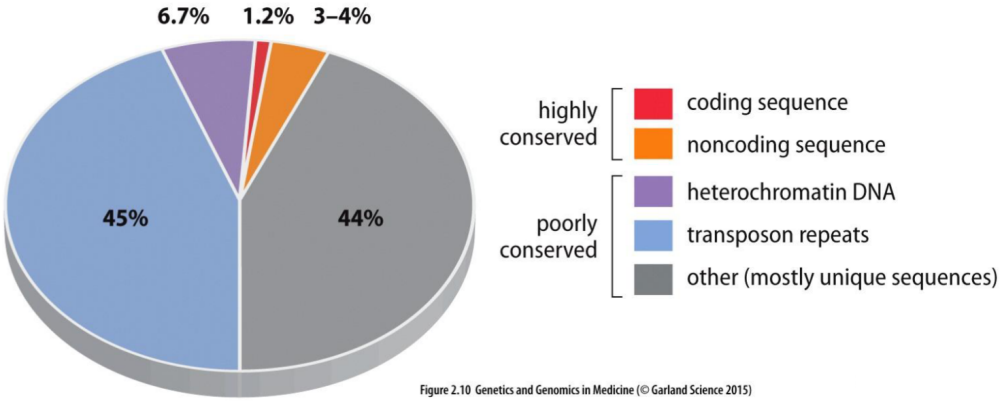


Figure 2.10 Genetics and Genomics in Medicine (© Garland Science 2015)

Barbara McClintock won the Nobel Prize in Physiology or Medicine in 1983: "for her discovery of mobile genetic elements."

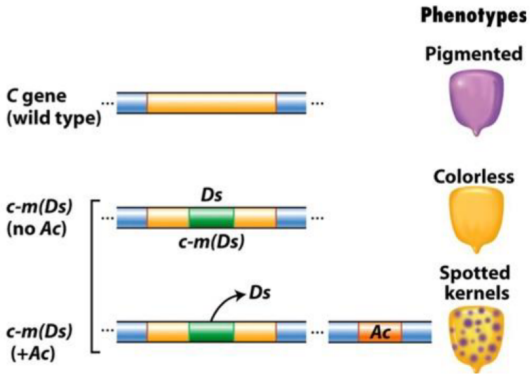


Figure 14-4
Introduction to Genetic Analysis, Ninth Edition
© 2008 W. H. Freeman and Company