Connecting genotype to phenotype in human genetics
(Mendelian disease)

- Many coding gene-variants, which are disease causing?
- Currently ~4500 Mendelian disease genes are known (OMIM.org).
  > Estimated that there are ~6000 – ~13,000 Mendelian diseases to be discovered.*

- Dominant conditions are common in human genetics, for autosomal disease genes (OMIM.org)
  > ~53% of disease genes show dominant genetics, ~35% disease genes show recessive genetics and
  ~12% both dominant and recessive genetics, which contribute to distinct diseases.

*Bamshad et al. 2019. American Journal of Human Genetics

Connecting genotype to phenotype in human genetics
(Mendelian disease)

From patient whole exome (exome) or genome sequencing, candidate disease genes are called
Genes of Unknown clinical Significance (GUS).
- The ClinGen database and working groups determine, from publicly available data, if a gene
  is a disease gene https://www.clinicalgenome.org/
- Similar info found at Online Mendelian Inheritance in Man (OMIM): http://omim.org/

- From patient exome or genome sequencing, often variants are found in known disease genes,
  but it is not known if the variant disrupts gene function (is pathogenic) – these are
  called Variants of Unknown clinical Significance (VUS).
- The ClinVar database and working groups determine, from publicly available data, if a
  variant in a known disease gene is pathogenic & thus likely contributing to disease or is
  benign and not disease causing.

*Bamshad et al. 2019. American Journal of Human Genetics
Approaches employed with patients

Genotype
- Exome or genome sequence of patient/proband and ideally both parents (trio).
- Inheritance data (not always available).
  - \textit{De novo} variants (in proband, but not parents), if not in normal population databases, are considered to be likely damaging to gene function.

Phenotype
- Clinical presentation from examination, biochemical tests, imaging, electrical recordings, etc.

Recurrence is a major criteria for identifying new disease genes and determining that a variant is disease causing.
MatchmakerExchange is a portal used to identify recurrence in the overall patient population.

- Disease gene: multiple unrelated patients with the same/similar phenotypic presentation that have predicted damaging variants (e.g. premature stop codons) in a gene that was not previously known to contribute to the observed phenotype [no longer a GUS].

- Disease causing (pathogenic) variant: multiple unrelated patients with the same/similar phenotypic presentation that have the predicted damaging variant at the same residue in a known disease gene [no longer a VUS].

Clinical outcome of exome sequencing of patient and parents (trio)

\begin{itemize}
  \item \textbf{New candidate disease gene (Gene of Unknown clinical Significance; GUS)}
  \item \textbf{Variant of Unknown clinical Significance (VUS) in known disease gene}
  \item \textbf{No clear candidate disease causing variant found}
\end{itemize}
What are some of the approaches to relate GUS and VUS to disease?

- Sequence databases of "normal" individuals
- Use of model organisms to test gene-variant for altered function \textit{in vivo} (addressing GUS or VUS)
- Massively parallel single gene variant analysis (addressing VUS)  
  February 13 lecture

Aggregation of "normal" human population sequencing, to approximate a population "reference" genome

Analysis of protein-coding genetic variation in 60,706 humans  
\textit{ExAC} database, from analysis of whole exome sequencing (WES)

The mutational constraint spectrum quantified from variation in 141,456 humans.  
Karczewski et al., 2020. Nature \textbf{581}:434-455  
\textit{gnomAD} database, combined analysis of WES and whole genome sequencing

gnomAD v4.0, now with ~800,000 individuals, more ethnically diverse  
(https://gnomad.broadinstitute.org/news/2023-11-gnomad-v4-0/)
Diverse populations represented (but numbers fall short for many groups)

Types of variation observed (vast majority does not contribute to disease)

Karczewski et al., 2020

ExAC/gnomAD contains exomes/genomes from single unrelated individuals

- ExAC/gnomAD has a large enough sample size that there are multiple variants per gene, including cases of independent recurrence; but not yet saturated.

- Gene-variants that have a population frequency >0.5% are defined as benign polymorphisms rather than disease causing.

- Based on frequency of observed versus expected variants in a given gene, can assess gene specific tolerance to variation.
  - Valid for dominant phenotypic consequence – haplo-insufficiency
  - Not valid for recessive phenotypic consequence (phase relevant)
ExAC/gnomAD contains exomes/genomes from single unrelated individuals

- ExAC/gnomAD has a large enough sample size that there are multiple variants per gene, including cases of independent recurrence; but not yet saturated.

- Gene-variants that have a population frequency >0.5% are considered as benign polymorphisms rather than disease causing.

- Based on frequency of observed versus expected variants in a given gene, can assess gene specific tolerance to variation.
  - Valid for dominant phenotypic consequence - haplo-insufficiency
  - Not valid for recessive phenotypic consequence (phase relevant)

*** In human genetics, "loss of function" means protein truncating variant (PTV) that arise from premature stop or splice site disrupting variants. These are largely assumed by the community to be null (but see Coban-Akdemir et al. 2018 Am J Hum Genet. 103:171-187 for rare exceptions).

pLI (Lek et al., 2016)

- pLI is a metric developed by ExAC.
- pLI estimates the probability of being loss-of-function intolerant for each gene (in the heterozygous state - thus likely to be haploinsufficient).
- Calculate expected number of variants in each functional class (synonymous, missense, loss-of-function) for each gene based on a local sequence context model.
- Observed number of variants in each class are counted in ExAC & gnomAD data.
- pLI is calculated using the deviation of the observed loss-of-function counts from the expected counts.
- In gnomAD introduced observed/expected (o/e), viewed as a continuous function.
Gene A - Frequency of expected versus observed loss of function variants are equal

- 11,400 (58%) of genes are confidently considered as loss of function tolerant (pLI < 0.9);
- individuals heterozygous for loss of function variants are not selected against in the population
Gene B - Frequency of expected versus observed loss of function variants are different.

- ~2,276 (~16%) of genes are confidently considered extremely loss of function intolerant (pLI > 0.9)
- >90% of known haploinsufficient genes have pLI > 0.9
- ~30% of genes that are loss of function intolerant have been linked to disease, the remaining ~70% of loss of function intolerant genes have yet to be linked to disease.
Cautions with ExAC/gnomAD data

- Lack of constraint/tolerance does not necessarily reflect absence of a lethal disease or status of the disease gene in the heterozygous state: 
  \( BRCA1 \) has a pLI = 0 (o/e = 0.73)

- There is no phenotype data available for individual in this database. While individuals with penetrant early onset disease have been excluded, the dataset certainly contains individuals with Mendelian disease (not all "normals" are normal).

- The database is underrepresented for individuals of Middle Eastern, African & Asian populations

Using model organisms to investigate human gene-variants

- For a human gene-variant GUS or VUS, model organisms can provide functional (experimental) data about the consequences of the gene-variant relative to the biology of the model organism.

- This requires that the model organism has an orthologous gene, and that the variant residue is conserved, so that it can be modeled by CRISPR/Cas9 genome editing; alternate approach employs rescue with the human ortholog (if rescues).

- Can assess if a gene-variant found in a patient is damaging in C. elegans by knocking the variant (var) into the C. elegans ortholog at the corresponding residue and determine if there is a phenotype in the homozygous state. 
  > If there is a phenotype (damaging), this suggests that the variant is likely damaging in the patient, contributing to one or more of the patient's phenotypes.

- Determine genetic mechanism of the variant, using the genetic principles I described. This is important given the high prevalence of dominant presentation that is not haploinsufficiency.
Investigating gene function - mechanism in C. elegans

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Phenotype</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>++/++</td>
<td>wildtype</td>
<td>normal function</td>
</tr>
<tr>
<td>ko/ko</td>
<td>present</td>
<td>gene is necessary for functions that are defective</td>
</tr>
<tr>
<td>ko/ko</td>
<td>absent</td>
<td>gene is not necessary for tested phenotype</td>
</tr>
<tr>
<td>ko/++</td>
<td>wildtype</td>
<td>gene is recessive (not haploinsufficient)</td>
</tr>
<tr>
<td>ko/ko</td>
<td>present</td>
<td>gene is haploinsufficient</td>
</tr>
</tbody>
</table>

Investigating gene-variant genetic mechanism

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Phenotype</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>var/var</td>
<td>present</td>
<td>var is damaging</td>
</tr>
<tr>
<td>var/var</td>
<td>= ko/ko</td>
<td>var is null</td>
</tr>
<tr>
<td>var/var</td>
<td>&lt; ko/ko</td>
<td>var is hypomorph</td>
</tr>
<tr>
<td>var/var</td>
<td>&gt; ko/ko</td>
<td>var is likely change of function</td>
</tr>
<tr>
<td>var/var</td>
<td>opposite of ko/ko</td>
<td>var is likely hypermorph</td>
</tr>
<tr>
<td>Gene is not haploinsufficient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>var/ko</td>
<td>var is a change of function</td>
<td></td>
</tr>
<tr>
<td>=&gt; ko/ko</td>
<td>var is likely dominant negative/antimorph</td>
<td></td>
</tr>
<tr>
<td>var/ko</td>
<td>= or &lt; ko/ko</td>
<td>var is likely dominant negative/antimorph</td>
</tr>
<tr>
<td>Gene is haploinsufficient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ko/var</td>
<td>var is likely end of haploinsufficient gene</td>
<td></td>
</tr>
<tr>
<td>&gt; ko/var</td>
<td>var is likely dominant negative/antimorph</td>
<td></td>
</tr>
<tr>
<td>= or &lt; ko/var</td>
<td>var is likely dominant negative/antimorph</td>
<td></td>
</tr>
<tr>
<td>Effect of adding wildtype gene copies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wt/wt</td>
<td>var is not suppressed var/+ phenotype</td>
<td></td>
</tr>
<tr>
<td>var/wt</td>
<td>var is a dominant negative/antimorph</td>
<td></td>
</tr>
<tr>
<td>= or &lt; ko/wt</td>
<td>var is not enhanced var/+ phenotype</td>
<td></td>
</tr>
<tr>
<td>ko/wt</td>
<td>var is a hypermorph</td>
<td></td>
</tr>
<tr>
<td>ko/wt</td>
<td>var is a neomorph or dominant negative/antimorph</td>
<td></td>
</tr>
</tbody>
</table>

Key
- + - wildtype (wt) allele
- ko – knockout, or null, elimination of gene activity (called loss of function in human genetics)
- var - variant; missense, stop, frameshift, insert
- wt-sct - wt single-copy transgene containing the genomic locus
- = - phenotypes equivalent
- > - phenotype is stronger than
- < - phenotype is weaker than
- >> - phenotype is much weaker than
- > - phenotype is much stronger than

General findings in model organisms, related to humans

- For a dominant mode of variant action - hypermorphic and antimorphic mutations, to the extent to which it has been analyzed, show conservation between orthologous gene-variants in different species (e.g., activated Ras, antimorphic myosin).

- Haploinsufficiency for orthologous genes is significantly less conserved. Different species titrate gene activity differently. As expected, mouse shows more similarities with humans for haploinsufficiency than more distantly related species, yet there are differences in which genes are haploinsufficient.
Using model organisms to investigate human gene-variants

- Modeling can determine the cellular, subcellular or molecular mechanism that variant causes disease.
- Devise potential treatment strategies through suppressor genetics & drug screening.
- Note that gross/organismal phenotype of the patient disease gene-variant in the model organism often does not resemble disease presentation in humans (called phenologs), although at the molecular level phenotype is usually the same.  
  (Golden, 2017, Mol Reprod Dev. 84(11):1118-1132)

Importance of determining the genetic mechanism of gene-variant defect
Genetic mechanism of gene-variant defect

Null → Gene-variant → Hypermorph

Hypomorph → Gene-variant → Dominant negative/ Antimorph, Neomorph

Provides conceptual route for treatment of the gene-variant defect

"Compensate" - Activate downstream or parallel pathway, restore wild type product

Null → Gene-variant → Hypermorph

Hypomorph or haploinsufficiency → "Activate" - Increase gene product function, inhibit negative regulator

"Inhibit" - Decrease level or activity of hypermorphic product, inhibit positive regulator

"Eliminate" - Decrease level or activity of toxic product

Dominant negative/ Antimorph, Neomorph