Relating Animal Model Phenotypes to Human Disease Genes

Project Goals:
• To develop methods and syntax for describing phenotypes using ontologies
• To compare consistency among annotators using these methods
• To query for similar phenotypes within and across species
• To provide use cases and feedback for tool development
Currently there is no easy way to connect mutant phenotypes to candidate human disease genes.

**Humans**

- Mutant Gene
  - Mutant or missing Protein
  - Mutant Phenotype (disease)

**Animal models**

- Mutant Gene
  - Mutant or missing Protein
  - Mutant Phenotype (disease model)
Sequence analysis (BLAST) can connect animal genes to human genes.

Humans

Mutant Gene

Mutant or missing Protein

Mutant Phenotype (disease)

Animal models

Mutant Gene

Mutant or missing Protein

Mutant Phenotype (disease model)
Humans and Animal models

- Mutant Gene
- Mutant or missing Protein
- Mutant Phenotype (disease)
- Mutant Phenotype (disease model)

Shared ontologies and syntax can connect mutant phenotypes to candidate human disease genes
Information retrieval is not straightforward.

<table>
<thead>
<tr>
<th>OMIM Query</th>
<th># of records</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large bones</td>
<td>251</td>
</tr>
<tr>
<td>Large bone</td>
<td>713</td>
</tr>
<tr>
<td>Enlarged bones</td>
<td>75</td>
</tr>
<tr>
<td>Enlarged bone</td>
<td>136</td>
</tr>
<tr>
<td>Big bone</td>
<td>16</td>
</tr>
<tr>
<td>Huge bone</td>
<td>4</td>
</tr>
<tr>
<td>Massive bone</td>
<td>28</td>
</tr>
<tr>
<td>Hyperplastic bones</td>
<td>8</td>
</tr>
<tr>
<td>Hyperplastic bone</td>
<td>34</td>
</tr>
<tr>
<td>Bone hyperplasia</td>
<td>122</td>
</tr>
<tr>
<td>Increased bone growth</td>
<td>543</td>
</tr>
</tbody>
</table>
Relating Animal Model Phenotypes to Human Disease Genes

Project Goals:
• To develop methods and syntax for describing phenotypes using ontologies
• To compare consistency among annotators using these methods
• To query for similar phenotypes within and across species
• To provide use cases and feedback for tool development
Annotate phenotypes of orthologous genes in human, zebrafish, and *Drosophila* using PATO and EQ syntax

Triple blind annotation of human phenotypes

Compare annotations

FlyBase

BBOP

ZFIN

OBD
Results: Number of annotations added to OBD

Human

Curate data from OMIM entries for gene and related diseases into PBD using EQ syntax

Zebrafish

Curate data from zebrafish publications for mutant and morphant phenotypes into ZFIN using EQ syntax
Results: Number of annotations added to OBD

Human
Curate data from OMIM entries for gene and related diseases into OBD using EQ syntax

10 genes
677 EQ annotations from ZFIN
314 EQ annotations from Flybase
507 EQ annotations from BBOP

Zebrafish
Curate data from zebrafish publications for mutant and morphant phenotypes into ZFIN and OBD using EQ syntax

4,355 genes and genotypes into OBD
17,782 EQ annotations into OBD
Tests of the method

1. How consistently do curators use the ontologies and EQ syntax?

2. Can the phenotype annotations for one mutation be used to retrieve annotations to another allele of the same gene?

3. Given a human phenotype, can we retrieve similar phenotypes from mutations in model organisms? Are these mutations in homologous genes?

4. Given a model organism phenotype, can we find other known (or unknown) pathway members with similar phenotypes?

5. Do zebrafish paralogs have phenotypes that are complementary to their mammalian ortholog?
~10% of the EQ statements for 1 gene
We need a quantitative method to calculate the similarity of annotations.
Similarity of phenotype annotations is calculated by reasoning across the ontologies.

Query:

Gene variant influences $E = \text{facial ganglion} + Q = \text{morphology}$

$\text{cranial ganglion} \quad \text{is_a} \quad \text{epibranhial ganglion}$

Similar annotations have the same or more general entity types

$\text{shape} \quad \text{structure} \quad \text{size}$

$\text{convex} \quad \text{folded} \quad \text{wavy}$

Similar annotations have the same or more specific PATO qualities
Similarity of phenotype annotations is calculated by reasoning across the ontologies.

Similarly, annotations have the same or more general entity types.

Similar annotations have the same or more specific PATO qualities.

Similarity between two genotypes is weighted by: the total number and rarity of similar annotations, and the degree of relatedness within each ontology.
How consistently do curators use the ontologies and EQ syntax?

At the intersection of different curators’ term choices, colors merge, with white indicating consensus among all 3.
### Subsumption in similarity scoring

<table>
<thead>
<tr>
<th>subsumption</th>
<th>Gene</th>
<th>Genotype</th>
<th>Quality</th>
<th>Morphology</th>
<th>Shape</th>
<th>Thickness</th>
<th>Genotype</th>
<th>Genotype</th>
<th>Genotype</th>
<th>Genotype</th>
<th>Genotype</th>
<th>Genotype</th>
<th>Genotype</th>
<th>Genotype</th>
<th>Genotype</th>
<th>Genotype</th>
<th>Genotype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gene A</td>
<td></td>
<td></td>
<td></td>
<td>Gene A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td></td>
<td></td>
<td>Gene A</td>
<td>A-001</td>
<td>+</td>
<td>+</td>
<td>A-001</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gene A</td>
<td>A-002</td>
<td>+</td>
<td>+</td>
<td>A-002</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gene A</td>
<td>A-003</td>
<td>+</td>
<td>+</td>
<td>A-003</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene</td>
<td></td>
<td></td>
<td>Gene B</td>
<td>B-001</td>
<td>+</td>
<td>+</td>
<td>B-001</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gene B</td>
<td>B-002</td>
<td>+</td>
<td>+</td>
<td>B-002</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Gene B</td>
<td>B-003</td>
<td>+</td>
<td>+</td>
<td>B-003</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype</td>
<td></td>
<td></td>
<td>A-001</td>
<td>+</td>
<td></td>
<td></td>
<td>A-001</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype</td>
<td></td>
<td></td>
<td>A-002</td>
<td>+</td>
<td></td>
<td></td>
<td>A-002</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype</td>
<td></td>
<td></td>
<td>A-003</td>
<td>+</td>
<td></td>
<td></td>
<td>A-003</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype</td>
<td></td>
<td></td>
<td>B-001</td>
<td>+</td>
<td></td>
<td></td>
<td>B-001</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype</td>
<td></td>
<td></td>
<td>B-002</td>
<td>+</td>
<td></td>
<td></td>
<td>B-002</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genotype</td>
<td></td>
<td></td>
<td>B-003</td>
<td>+</td>
<td></td>
<td></td>
<td>B-003</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.99</td>
<td>0.6</td>
<td>0.3</td>
<td>1.1e-4</td>
<td>3.2e-5</td>
<td>6.1e-6</td>
<td>2.0e-7</td>
<td>1.9e-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Unlikely to match by chance
We can quantitate similarity and, thus, optimize consistency.
Results: best practices for phenotype curation to ensure curator consistency

1. Use the same set of ontologies
2. Use the same ID format
3. Use the same Phenote configuration
4. Constrain post-composition of entity terms to the same type and relation
5. Annotate both the anatomical entity and the process where applicable
6. Annotate to a more general term in the PATO hierarchy when the correct term is unavailable, rather than not making an annotation
7. For OMIM, annotate both the general description as well as specific alleles
Relating Animal Model Phenotypes to Human Disease Genes

Project Goals:
• To develop methods and syntax for describing phenotypes using ontologies
• To compare consistency among annotators using these methods
• To query for similar phenotypes within and across species
• To provide use cases and feedback for tool development
Can the phenotype annotations for one mutation be used to retrieve annotations to another allele of the same gene?

Example: A search for phenotypes similar to each human \textit{EYA1} allele returns other human \textit{EYA1} alleles

<table>
<thead>
<tr>
<th>\textbf{EYA1 query}</th>
<th>\textbf{target}</th>
<th>\textbf{Allele number}</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>query</td>
<td>target</td>
<td>Allele number</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

(The smaller the number, the more similar)
Given a human phenotype, can we retrieve similar phenotypes from model organisms? Are these due to mutations in orthologous genes?

A query for phenotypes similar to:

Human *EYA1* variant OMIM:601653
MP:deafness = E = Sensory perception of sound  Q = absent

returns:

Mouse *Eya1* ^{bor/bor} and *Eya1* ^{tm1Rilm/tm1Rilm}
E = Sensory perception of sound  Q = decreased
Given a human phenotype, can we retrieve similar phenotypes from model organisms? Are these due to mutations in orthologous genes?

A query for phenotypes similar to:

**Human EYA1 variant OMIM:601653**
MP:deafness = E = Sensory perception of sound Q = absent

returns:

**Mouse Eya1^bor/bor** and **Eya1^tm1Rilm/tm1Rilm**
E = Sensory perception of sound Q = decreased

These similarities are based on the same GO term entity. Anatomical cross-species queries require classification of anatomical structures, in the different species, based on function and/or homology.
Currently, different terms are often used to describe anatomy in different species.
Future queries across species will utilize homology annotations.
Human and zebrafish SOX9 annotations are similar

Human, SOX9 (Campomelic dysplasia)
- Scapula: hypoplastic
- Lower jaw: decreased size
- Heart: malformed or edematous
- Phalanges: decreased length
- Long bones: bowed

Zebrafish, sox9a (jellyfish)
- Scapulocorocoid: aplastic
- Cranial cartilage: hypoplastic
- Heart: edematous
- Pectoral fin: decreased length
- Cartilage development: disrupted
Curation of mutant phenotypes and human diseases using common ontologies & syntax can provide candidate genes and animal models of disease.
Given a model organism phenotype, can we find other known (or unknown) pathway members based on similar phenotypes?

Similarity search for zebrafish \( shha^{t4/t4} \) identifies pathway members

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Congruence</th>
<th># of alleles</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>( disp1^{ty60} )</td>
<td>9.6E-19</td>
<td>6</td>
<td>regulates long range Shh signaling</td>
</tr>
<tr>
<td>( gli1^{ts269} )</td>
<td>6.6E-13</td>
<td>2</td>
<td>downstream transcriptional repressor</td>
</tr>
<tr>
<td>( wnt5b^{le1c} )</td>
<td>4.3E-11</td>
<td>5</td>
<td>downstream target gene</td>
</tr>
<tr>
<td>( smo^{hi1640Tg} )</td>
<td>4.1E-11</td>
<td>4</td>
<td>membrane protein mediates Shh intracellular signaling pathway</td>
</tr>
<tr>
<td>( hhip^{hu540a} )</td>
<td>3.1E-5</td>
<td>2</td>
<td>binds Shh in membrane</td>
</tr>
</tbody>
</table>
Do zebrafish paralogs have phenotypes that are complementary to their mammalian ortholog?

<table>
<thead>
<tr>
<th>Mouse $Notch1^{-/-}$</th>
<th>Zebrafish $notch1a^{-/-}$</th>
<th>Zebrafish $notch1b$ MO</th>
</tr>
</thead>
<tbody>
<tr>
<td>somite formation</td>
<td>somite</td>
<td></td>
</tr>
<tr>
<td>neural tube</td>
<td>neural plate</td>
<td></td>
</tr>
<tr>
<td></td>
<td>nervous system</td>
<td></td>
</tr>
<tr>
<td></td>
<td>motor axons peripheral</td>
<td></td>
</tr>
<tr>
<td>ectoderm development</td>
<td></td>
<td>ectoderm development</td>
</tr>
<tr>
<td>myocardium</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pericardial sacs</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cardiac muscle</td>
<td></td>
<td></td>
</tr>
<tr>
<td>angiogenesis</td>
<td></td>
<td>blood vessel morphogenesis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>intersegmental vessel</td>
</tr>
<tr>
<td></td>
<td></td>
<td>endothelial cell migration</td>
</tr>
<tr>
<td>notochord</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Relating Animal Model Phenotypes to Human Disease Genes

Project Goals:
• To develop methods and syntax for describing phenotypes using ontologies
• To compare consistency among annotators using these methods
• To query for similar phenotypes within and across species
• To provide use cases and feedback for tool development
Results of tool development

Unanticipated outcomes for ZFIN:

- Phenote is now used by ZFIN collaborators to submit data
- Phenote checks versioning of the ontologies, eliminating the need to update templates for researchers
- Phenote has a familiar spreadsheet feel
- Phenote supports reading and writing in multiple file formats
- Phenote provided template for ZFIN curator interface development