MINUTES
Phenoscape all-hands meeting, Chicago 2013
May 8-9, 2013

Wednesday, May 8, 2013

Morning Plenary session

Introduction (Paula)

- We are on track

- As we enter 3rd yr., need to turn attention to demonstrating utility, driving research, attracting users
- Mods finish up this year, need to coordinate how MODs and Phenoscape will move forward

Parking Lot (please put issues here that you'd like to address if not on the agenda)
Need to produce a list of deliverables desired from the project meeting

Knowledgebase GUI (Jim)

- Need for redesign due to new data, driving use cases, lessons from experience
- New data includes broader evolutionary scope, fossils, multiple MODs, gene expression
- Currently, 15,941 character states, 121,522 phenotype annotations, and 318,501 expression annotations
- Judy: Is gene expression being presented at the level of cell type? Increasingly have this data and regulation of gene expression will increasingly focus on sets of genes, co-expression sets. Note that uberon contains cell type ontology integrated, this can facilitate cross-granular query.

Jim identified interface issues based on expansion of data types in Phenoscape II (1-3), drivers from CalAcad workshop (4-5), lessons learned from Phenoscape I (6), and Paul and Nizar's ideas (7):
1. Taxonomic expansion - Jim showed default existing interface where user doesn't - know what the organisms are. Solutions: add column with taxon name;
2. Gene to taxon identification necessary; another column?
3. Gene expression (new data) needs to be incorporated. How do users want to work with these data?
- HL: how much should we bother integrating, querying, displaying, etc. data that already exist at MODs; how much replication should we have?
- Jim: unlike other db, we have an aggregate of MOD data + evolutionary data. Focus on evolutionary data and how to link to expression.
- Terry: need to indicate where data are coming from, e.g., not just 'mouse' but 'MGI mouse' vs. other mouse.
- Monte: begin interface design with use cases focused from evolutionary standpoint. Think about linking out to other data
Judy: cell expression data becoming available.
- Judy: hiding display of IDs is problematic
4. (Driven from CalAcad users): want to map phenotypes on phylogenetic tree. Have not addressed directly before; we have a taxonomy, but not a tree;
- Judy: project overlaying trees with GO annotations, could be used as a model here
-Use of Paint to propagate phenotypes was one idea for the Monarch-Phenoscape NSF proposal.
5.: Retrieve a supermatrix from Phenoscape
- Requirements: default phylogeny we hope to be provided by OpenTree, default mapping of character states to be determined

- Example: where on the tree are there changes in head shape?
- Example: genes with phenotypes similar to profile of changes between taxa A and B
6. Redesign ideas based on lessons learned from Phenoscape I
-- What sort of user is going to come to the KB and compose an EQ?
-- Issues with faceted browse: "I don't want an immaterial anatomical entity, I'm looking for portion of organism substance"
- Limitation of ontological annotation relative to original character
7. Potential for use of images in term info pages as an 'anatomy glossary'
- Nizar has annotated ~50 skeleton images for cat and hornbill
- Curation will be required to decide on preferred labels, hidden terms, etc.

Jim- Semantic similarity behind features to connect data to data via ontology machinery.

Phenoscape knowledgebase -- web user interface (Jim)

What needs to be incorporated? What is available?
Basis for discussion on priorities, ideas,
Interface development requirements

Kb.phenoscape.org
Existing web interface
Data is fish annotated
Fin model organism information

Developing the UI
Experience with using ontology driven phenotype database

New data is not just fish

Backend, new infrastructure
Semantic web technologies RDF al reasoners
Data integration and reasoning are working

Evolutionary Data , what do we need in the interface
Same types of annotations, character states

Broader array of taxa, could just add a column
Grouping we define in advance
What are the right levels of those?
Faceted browsing interface, hierarchical summary
How many data points are under each term, develop simplified taxonomy
Hits major nodes people want to know

For gene data , ZFIN phenotypes in current database
Simple answer, additional databases, add column for whatever the new thing is, label what organism it is coming from
Need to add to interface choosing (filter by organism)

What else might you want to do with gene expression data?
How do you want to deal with it in a web interface?

Integrating a lot of data from different sources
Available online, visible, query
How much should we bother in UI design in replicating browsing and querying data that can‚Äôt already be browsed and queried elsewhere

Or should we focus on those parts we can‚Äôt do elsewhere

Evolutionary data is the focus
What to do with gene expression data
Think about how to
Where does model organism come into that?

Nomenclature isn't always the same
Should be able to distinguish the organism if not the source that you‚Äôre getting the data

Should there be another column? A different color?
Wasn’t that previous problem with taxa

Use cases would be helpful

In breakout groups come up with
What the questions are you would want to be able to answer
For example, what could an evolutionary biologist get from this?
Look for that line where we provide minimal access
May not know how much data they can get and where they can get it
Linking out part

Bulk downloads of slices of dataset
Scripted analysis
Not a lot of wanting web interface to browse around
Consistent request on what Phenoscape is, want to see phenotypes on phylogeny
Where is evolutionary change in particular structure?
Haven‚Äôt really addressed directly in interface
Have taxonomy but not mapping changes to particular branches

Providing a way to view changes
Papers have identified changes
Would be useful to expose those better
Where on the tree are their changes related to head shape
Relevant nodes and branches

Redesign ideas from experience with using existing database
After using website we developed
Machinery of ontology is way too in your face on the site
What sort of user do we think will come to knowledge base and have to be able to compose EQ to do query?

Replacing a table with results that are found using the ontologies
People would be more connected to the original data
Character state, maybe not really seeing ontology terms it is annotated with
Another place where semantic similarity tools will need to play a role

Feedback from other parts of the project
Paul and Nizar talking about design changes
Enhanced term info page
Image part, how can I see these phenotypes or structures
Incorporate images in a view, takes some curation
Nizar has annotated about 50 files for cat and hornbill
Using ontology hierarchy, finding all examples of femur or replacement bone, anatomical structure - that can be done
Add into existing term info, could link to more images than just the one that comes up

Paul and Nizar, mock up, anatomy glossary
Wants to lead people through
Is that a focus for our knowledgebase? Phenotypic analysis
Could be companion site that presents anatomical terminology
Could be done, fair amount of curation required
Would want to come up with a system of annotations on terms that guide how displayed on interface

System could be reused for any type of knowledge browsing site driven by ontology
Didn't just want raw ontology, present friendlier view

Image annotations could have dual use
How best to proceed?
Femur in a mammal, half dozen things to label, pin and tag
Would be helpful to have across vertebrates, useful thing to work into outreach
Someone (anatomical expert) has to do the tagging of the anatomical entities to be displayed in

interface
Show the anatomy we're getting the genes from?
Made a list of model organisms, where interest level, where would you pick
Making place in interface for images sounds reasonable
Why can't we see phenotypes?
Mentioned getting images out of Morphbank, is there a reasonable way to add images?

Overview entry into the terms across the tree
Make it is clear where images are or are not available – e.g., see ZFIN (icon)

How much detail are you getting?
Pointing arrow with label
Scanned or photographed on uniform background
How does one know taxon connection of a particular image?
Nizar is annotating taxon and structure
Could concentrate on finishing to labeling point, skull skeleton, various images

This year will focus on making an interface, if you took what we have right now and superimposed it what kind of effort is involved in that?
Jim estimates - a few months of not absolute full time, something running in early fall
Requirements for different topics
Gene expression data, phylogenetic placement of characters

Semantic similarity (Todd)

Cluster of issues around text mining work and curation
Number of small experiments and a larger one that looked at different curators curating the same characters or phenotypes which gives some sense of how much variance there would be between them, useful in itself - guidelines for curation
Idea to use that in a more systematic way for driving the research
Gene phenotypes, comparisons between
Have built in source data for comparison
Part of what's driving the capstone
Fin limb transition - well known evolutionary development thing lots of candidate genes
If system works, what are genes from transitions?
Enrich for candidate genes we know
Path to verification for capstone
Hardest is evolutionary to evolutionary comparisons; how do we know if two profiles of phenotypes are accurate?
How to take advantage of intercurator consistency?
How much value is added by including qualities?
(MH- this is a good question. we found in earlier analyses that the answer was not much at all)
Look for evolutionary characters more similar with one or other without qualities

Would expect and hope that semantic similarity between two curators would be closer with EQ vs. E alone.
Good measure of semantic similarity - closer to same terms
Hope to maximize similarity as evaluation
Same evaluation criteria for Charaparser:
If we mine text and come up with candidate ontology terms, closer to gold standard set of annotations than another curator? How close are they?
Allows us to assess quality of evolutionary annotations without gold standard for knowing real similarities, always looking at same characters when doing comparisons

**Can use common dataset for evaluation of semantic similarity and Charaparser.
What we do have is a Washington et al. dataset (3 curators, 5 genes lots of phenotypes) and results that qualities were not useful.
What we don't have - good dataset with Phenoscape evolutionary data with different curators individually doing it
MH- having multiple curators curate the same data is a great way to determine where the curatorial guidelines are lacking. Should be done in all curation projects, imho.
Hugely important for curator activities, if qualities are not useful, if aim is to do semantic similarity
Grab entities and run
Depends on what inference we do on changes
2 characters, all we know is thorax, don't know variations
Not as important but useful - for showing the vector, directions
Stop post compositions at this point because after that it isn't adding anything

Input into how to come up with a dataset that is shared for both objectives
Largely random characters, representative of dataset as whole
With at least 2 or more curators (inter-curator comparison)
Following same guidelines
In many cases, implicit knowledge, look at paper to finish annotation, fidelity, look at characters without consulting paper
Let them use all of their outside knowledge and look at the paper
Different comparison
Use same guidelines
Might be multiple comparisons worth doing
What will affect what we will be able to say after the fact about whether it is working relative to?
Curator with external knowledge or other comparisons relative to that dataset

Prashanti's analysis of the small dataset that Wasila put together from a data jamboree (2008 with Kevin Conway, Rick Mayden, etc.):
How much variability between curators (probably overestimate, untrained)
How large of a dataset do we need in order to distinguish similarity?
Difference between similar curators and random can normalize range of difference
In methodology that have effect

Before we have curators curate a set of phenotypes don't want to invest a lot of time and curate a dataset that is too large
Estimate required sample size to
Required number of annotations
Represent various power values of experiment
At power of .95 need sample size of 100 phenotypes to detect 10% difference b/t 2 datasets
Not expert curators, guidelines not defined properly
Little difference between curator data and random data

Need feedback on how many curators we want to involve in inter-curator variability study?
What are we supposed to curate?

Variability among curators accounts for one 1/3 of the variability between annotations, a lot more gain by adding annotations but not curators (need 2, may not be a lot of gain by adding 3) tripling size of annotations would add gain
100 phenotypes
Context: at least between 0 and 50 makes a lot of difference to be able to detect. What are we aiming for, what is useful?
Want to look at left end of graph, want at least 100, probably 200
10% or less
Effect size difference between random and curator

Null hypothesis, curators are giving us noise. Can we test the difference between any method and noise, comparing two subtle differences in how we do things, far from noise?
10% difference in effect size, coarse level of difference to be able to see

2 curators with same expert background? Taxonomic expertise? Want dino person curating fish stuff and fish person curating dino stuff;

Curation consistency checks, variability of persons background knowledge impacts where they annotate to a term, not so much of a concern as long as searching and indexing is against ontology structure
Outliers, annotations to outside of that sub-tree, look at the differences between curators, why are these outliers picked up?

Without paid curation effort what would be best strategy for having more consistency across?
Might be a large group of people, more inconsistent, best approach to randomly farm out and assign annotations across a group of people
Baseline level of inconsistency that wasn't taxon specific
Taxon specific things would throw us off the most, error
Even off across the board
What would be solution long term for adding data to something like this?

Any data to suggest difference between well trained expert curators and untrained non-experts?
Maybe using Washington et al paper, not on same dataset, don't currently have
Would bear on David's question. 5 untrained undergrads?

Afternoon Breakout Sessions (May 8)

Images (Paul, Monte, Nizar, Paula, and others….)

Importance of use cases in regard to images
First use case has to do with Phenoscape entry -- images would provide toehold:
Monte - all these Latin names that mean nothing. Can interface tell him what it means with pictures or images? Subdivide them in some way
In preparing for advisory board meeting knew we had human and mouse data, spent too long trying to find, easier to ask someone than find on website
- Use images to find out what the Latin names ARE.
On slide: show images of different animals, use images as search tool
Use images to solve taxonomic expansion

Images of major classes

Icons for different groups of organisms
Not intuitive to a morphologist, trying to create something intuitive for general audience and people coming into the site

Put together to show the connection between data types
EQ format, hard for people
Big interface problem
Too much data that is hard to get into

What is needed to make them possible?

First thing I'd like to see: diagrams, images
Define regions, drill down to get more detailed
We can't use images until they are annotated, if they were annotated with ontology terms, we could suck them into the database, referenced and hosted somewhere else
Retrieved and shown on the page

Beautiful, but muddied up the second the database is opened up to other people to put data in
Morphbank putting in images,

What are they looking for, what do they want to link together?
Could list key ontologies you might search

I’m looking for genes associated with mammals, how will I find them?
Looking for entities on the femur, maybe something I think I know of, is there a gene related to it?
Interface would make it take 30 minutes to determine
Entry point would be anatomy
Start with femur
Link to genes
Gene expression
Use case: Show me the genes expressed in the femur; entry anatomy, get to Paul's page:
Do with facets: get all taxa and genes annotated to femur; could use visual tags to do facets;

Not possible to degeneralize from phenotypes

Faceted approach (Paul's slide): major categories, tens of thousands of terms in each, but you drill down that way

Export function: export tree or whatever you see
Whittled down to your feature and definition, select what you need and leave with it
All data that has not been eliminated in your search into an excel spreadsheet for export

Original use case
Interested in identifying genes responsible for some evolutionary changes: Show me genes involved in transformation of the pelvic fin radials to the limb femur.
Where in phylogeny is that transition (problem solved by other group...)?
Have genes that are gained or lost in terms of expression between two states
A difference, some gained some lost, some kind of search to look at mutant phenotypes

Interface design to support this use case
I'd like to know all of the genes that when mutated change the form of one towards the other
We have positive statements of expression but no negative
Any mouse genes expressed that aren't expressed in fish?
What does that interface look like?
Take images of pelvic fin radials and a femur side by side
Hit subtract to see differences

Genes are tied to areas but no interface
Focus on fin and limb
Visual of fin - visual of limb =?

Semantic similarity test, Charaparser evaluation, curation breakout (add your name here)

Wasila (ppt)
Present where we're at with consistency and what issues came up

Curators' point of view

Broad overview of what we found so far with newly curated data
Slides refer to dataset from Alex Nizar and Wasila finished in January
Fin limb from paleontology literature

Close to 6,000 eq annotations
Updates to pato and uberon

First pass of annotation, next thing on radar is to review data with consistency review panel within phenex, what did you forget to fill in, basic EQ construction feedback

Will go back over files and review
Manual review of all of phenotypes on a spreadsheet for 6000 phenotypes
Alex and Nizar are getting started
Whether or not consistently annotating entities
Agree on understanding of what entities are and update files if inconsistencies with how they are recognizing entities
Look over use of relations and make sure consistent
Will request updates to spreadsheet on Phenoscape tracker

Review granularity of annotations
What are we aiming for with use cases and reasoning
Outcomes: annotations were done at much finer level than in Phenoscape 1
Only 5% are at attribute level
Things to size to shape to structure
25% of data in Phenoscape 1 was at coarse level
Dedicated curators
Phenoscape 1 - experts but consultants 5 hr./wk. inconsistencies popped up that way

Why didn't we annotate finer for 5%?
Paleontology literature possibility - paleo that we are annotating, descriptions are more concise, fewer words more to the point, less confusing compared to Phenoscape 1
Multiple sentences for character state, more complicated finer descriptions of anatomy
Could play into difficulties in Phenoscape 1 annotating to a finer level
What exactly was the character describing if wordy?

Could also be because of focus on fin and limb

Additional fish papers to continue experiment

Why couldn't we annotate 5% to finer level?
Stumped on how to do that
1. Constrained within phenex to simplified EQ model
Some require something more complicated
Limited
2. May need to contact expert or look at figure to make that judgment

Possibility of numerical data, ratios, count
Just can't handle those, leave that up to freetext field, recorded there

Annotating at a finer level, still considerable variation when you compare annotations among curators
One curator made very specific EQ
Vs. coarse annotation by other curator

Does overall level of curation matter within dataset
If all to fine level does it matter to use cases?
Does specificity within PATO matter?
Do complex post compositions matter?
Does consistency among level of curators matter?
What will we use as guidelines for granularity

Open question: whether we could use logic in ontology to do some consistency checking
Can we generate reports that use what’s in ontology to flag those errors and fix annotations?

Plan through the summer, opens up annotation of other non-fin limb characters

If difference is subtle between fine and less fine annotation, we may not be able to correct that

Hong Cui Charaparser (ppt)
Wasila compiled list of examples that represent different curation patterns
Looked through patterns look at all possibilities
Where could entities and characters be?
Could be anywhere
Rewrite whole algorithm
New has basic assumptions implemented, different patterns to address special cases through set derived dozen of patterns implemented in
Direct impact on how we choose for evaluation design
Answer a set of questions - presented at dry run
Concerned on how to get representative data, how we define representative?

Head and limb may have very different ways and varying complexity
Worried about over fitting
Has patterns that we identified from curators practices, those are the rules they need to follow
Algorithm, easy to add new patterns

Todd mentioned patterns Wasila identified are not the patterns Charaparser is sensitive to
Want to see example of fundamental difference in that way
What is relevance for identifying those patterns because you said there are more patterns of that source out there?
Activity for curators to identify those
Curators need patterns to guide they way they annotate
95% of patterns have been captured
With another year how many new patterns would you find? Wasila: probably not much
Relates to what Hong's interest is
Wasila
What is the set of patterns - volume of that set?
How many ways of talking about patterns?

Categorization of patterns, what specific steps to take, knowledge sources to know to arrive at reasonable EQ statement
Can be very simple ones

Looking into ontology, does not tell you that finer structure is actually sort of part of or close to larger part of the structure, Charaparser might think it is two different entities but curator knows they are related
Mistake in ontology? not a mistake. Missing link. Implicit knowledge that curator has
Should be in ontology.
Will always be a gap
Difference between curator styles, makes fork between annotations Sereno style vs. ones that aren't because syntax is so different
Ultimate EQs might be the same, but Charaparser needs to parse it completely differently
Differences in how authors describe them and how those relate to underlying abstract things
Fork in the way you process and workflow
Run over free text: no fork

Predefined style, if use style you will get high accuracy
New version - trying to get rid of fork, decision point
Not sure how well it will work
Will rely a lot more on ontology
Has to tell us it's a problem
If knowledge is there, use it without having to write into ontology without knowing if it won't
David: would like to compare performance on free-text species descriptions as well as character-states

Do you have as much from curators as you need? Most of curation patterns at this point.
Anymore in that way we can help with?
- How to choose a sample of characters? Pick from categories vs. random - go with random.

If priority is to say in future curate free text character descriptions, separate block
Include 50-100 characters that are in that way
Some way to evaluate, discussion to add

Is little circle representative?
Are patterns in Phenoscape 1 all found in little circle?
Phenoscape 1 wordier, finer descriptions
Basic patterns of related vs. monadic probably the same
Can’t be that similar, not that many things to say about limbs like that
Things like counts, whether requires 2 entities vs. 1
Few dozen categories
Raised issues around this

Example: 'pelvic, fused' but E: ischium Q: fused with RE: ilium
Requires implicit knowledge, Charaparser will never recover that

Capstone breakout (add your name here)

Presentation on capstone for advisory board?
Remind them what we are aiming for

Description of Capstone:

Generating set of genes - tricky
Requires a lot of expert input
Generate set from MODS

You need a list of important limb genes
1. Pull from lit
2. Have limb people tell you
3. Coelacanth genome project

Table: says gene, mouse homolog, zebrafish homolog
So we recognize when we ID’d in those databases
Who and how was this identified as a candidate?

If even some of it were milked out it would be nice
Single reference instead of added work of pulling out of literature

Make a test case: then scale up

Year 3: small proof of concept to scale up?
Are we properly sampling candidates?

What if we were to lump everything together, what would we return?

Nature of data: no fine-scale data

Questions out of that
Phenotypic profile between this point and this point, based on that what are the set of candidate genes relevant to that phenotypic profile?
Internode profile
And set of genes related to those phenotypes
How many of those, how much redundancy is there across that phenotypic profile?
Say 10 phenotypes in 8 are 3 of the same genes involved in?
That would be interesting outcome

Searching on profile basis

Post fact decompose it

What is the next step? Recovery step. Science Step.
Is there anything fundamentally interesting and new?

Interesting to know if things in which interesting phenotypes with genetic basis that seem fundamentally different -- either noise or interesting

Can we do anything after the fact, have we discovered new knowledge or is that an error in the process?

How is it that we would test that it was new knowledge vs. noise?
Elevates the rank of this capstone in the world of literature

Look at genetic pathways to see if there is indication they are connected in similar pathway

Expectation might be that it would be more frequent to come up in pathways we already knew about

Independent information we could hold out to bring afterwards, now look at this independent data
From another taxon, salamander or some non MOD taxon

Ideal - study underway, as we discovered this, these data were being collected

How do we get to phenotypic profile stage?
Wait 2 weeks.

Decision to be made: using EQ or going back to character states
A lot of EQs that don't differ between different character states

Need EQs at end, as long as we can say these character states are variable here and then we go to the mapping that we have

Force PATO to change things to make it more granular

Plan for coming up with dataset of candidates
(Search of mouse, zebrafish, etc. gene annotations to recover candidates for these transitions)
Semantic similarity search with ranked results

Method to bring that back (technique to result in list is done)

Actual list of candidates is done outside of system
Decisions to be made as to how to collect it, how big it should be

How do we make that list?
Hall's book fin to limb transition
Put a gene in each of those categories

Zfin expression data and phenotype data to line up against it
A lot of interaction
Other key papers get them in there
Small groups of experts
Other papers to curate
Find what there is in the literature to find what they should have in their databases

If we recover genes that aren't in literature, some things need to include in gold standard knowledge - make sense of what worked and didn't, not cherry picking, it was or it wasn't so it is a fair evaluation

Pick the ones the paper is focused on, the rest of the literature, part without being focus would be test, genes involved in processes but in transition, what's different

Regulators
Enhancer elements that are identified

Because mutants wouldn't pick up changes
Same genes in both but regulated differently
No difference in terms of key genes
Differently regulated

Might see it in spatial gene expression pattern
Transition may come down to regulatory
Spatial patterns of expression
Both present but present differently

All we can search with EQs are, could take entities and search expression patterns but not a lot of discrimination, lots of genes expressed in any given tissue

To an extent can't do anything about gene expression?

Easiest, gene loss patterns
Relevant to origin of tetrapod's, specifically focus on
Small subset of genes, picked them ahead of time

Could make it a priority to work on that list, we need it a head of time
We need to make sure the list is classified the way we want to study ahead of time

Classification Paula made is great
If encounter genes that are candidates, but we don't pick up, after the fact what we don't pick up are the ones we know to be differentially expressed but knock out doesn't show phenotype

3 categories of things; stories for all of them. As little fishing as possible
Know ahead of time how we're going to test if we believe them to be false positives or false negatives

Regulatory differences and not so much presence absence of genes being expressed

Very important component
Concentration of particular expression

Paper David is talking about is all enhancer-based

If we're trying to capture through Phenoscape genes important to fin to limb

Genes don't change, regulation does
Capturing that through phenotypes we observe

Phenotype is changing; genes are the same.

Infer that regulatory elements are somehow
Not going to be able to find regulatory elements

Propose set of genes that may have differential regulation that leads to that phenotype

Ruling out classes of those types of genes
Patterning, positioning initiation outgrowth
Genes inferred by that didn't involve outgrowth
When we have phenotypic profile, only infers subset of those
That would potentially be interesting
May have ruled out
Seem genes involved in outgrowth not as fundamentally important in transition
More of a synthesis type of thing
Maybe there are certain pathways worth investigating further
Never going to be able to give you specifics on how pathway itself has changed

Going to have to test can't just infer that
Other is gene loss, phenotype differences are interested and do seem to involve gene gain or loss

Hard to infer, would have to be absence of something in one of the MOD datasets

All genes in the box. Set of Phenoscape candidates returned, some are misfired noise, some are real genes that candidate list doesn't have
new discovery for example, ones that are candidates from literature, genes someone thought would be good, in fact not involved in actual transition

Issue with how long list is
Could make it encompass almost the whole box, of course you would find stuff in it, but also so much noise to not know how much of a difference was made

What would make a top candidate, in the absence of having done experiments, what would be good candidate for belonging to transition?
For genes at top of the list WHY at top of list?

Same gene expressed at different times

How do we decide to use 1,000, top 100 top vs. low?
Paula has a list of fin positioning, initiation, outgrowth, and patterning genes for fin/limb based on Hall’s book.

Candidate genes: how do you rank them, what to base it on?
How do we rank them by strength of evidence?

Should we include these because in chicken we know there is a regulatory difference?

Ones directly proposed in literature to be fin to limb
I.e. this causes digit formation
Ones specifically chosen by authors, these are candidates, and reasons for doing it and citations; need to keep track of species for below.

1. Experimental evidence at top of the list; Experiments actually in the species where there is a polymorphism

Positive as highest level (breaking something is easier than making something)

2. Gene expression from in situ data; Expressed in the right tissues at the right times (right place right time);

3. Gene expression from microarray; from a giant expression profile analysis for microarray (We are not getting microarray data)

Return list of candidates
Some of the lit candidates or external candidates we get some we don't and we can make sense after the fact, the ones due to (weakest evidence, biggest noise) better way to come up with candidate genes if we use Phenoscape as gold standard for finding them

Mouse has so many different and very sophisticated ways of gathering experimental

Semantic search method developing to bring back set of phenotypes

Lit of mutation analysis

Will be missing a lot of power potentially by not having other phenotypes on branches may get very different set of genes

Which ones of those do we want to make sure we annotate, subset of 9 thousand

You mean cherry pick annotation?
Run character states across the whole tree
By hand?
We have all limb stuff, run limb stuff and EQs, run characters independently
Same thing with rest of body

1500 characters, 6000 phenotypes
More likely to be important than random one
I think we're ok

Other things that are connected to those transitions, changes in neck for example
Might pull up a different profile of genes
Gene profile altered by pleiotropy

Go into knowledgebase
What you are querying is in model system data if there are those genes that happen to be involved

Can't discover interesting features about neck phenotypes if we don't have them

Reduce the number of terms we return that have nothing to do with fins and limbs

Would only come up in the event post hoc you came up with one of those genes

Hardest part would be ranking
Classifying
If we had 3 priorities
Frequency with which they are cited in papers in response to limb
1 analytical way of doing it, generate a list probably do in a variety of ways, say we have expert curated set, might also say how many times do they appear in pub med associated with that gene name, way of getting some sense of the frequency it is talked about in the context of that phenotype

If we are missing a gene that is talking about 10,000 times and it doesn't' turn up in this it would seem of note
Guide to getting list together as well
Making sure we're comprehensive

Don't just want genes involved in transition
Genes believed to be involved in transition

Hand developing a list

Not just pulling together candidates have proposed into one pool
Manipulating that pool
Pull together things in literature and say what evidence is for each of them

Mix of total pool and make a two-tiered priority list
Everything in the pool and everything that people talking about evo devo in fin limb have talked about

Not just every gene in the limb

Would one starting point be all of the genes involved in the skeleton from the MODs easy pool
But we don’t want the whole pool
Total pool
Of that we have sub class that are those things we think are important
Also ongoing subproject for MODs to look at data Venn diagram
Most surprising thing would be to turn up genes that weren't even on the bigger list, not on MOD list

Some phenotype somewhere else is somehow similar

Paula has gotten some going through Hall list, in mock databases? Looked in lit, some cases they missed something,

If we find a limb gene not in any of MODs but in MOD literature SHOULD be in database, otherwise we can't find it.

In theory we should be able to find it, if it was in there

Who will take leadership on the list?
How big is your current list
Couple hundred
50 of that molecular systematics lab

Paula can take leadership
Including: David, Alex

Do what extent do we want genes that are not in the MODs
We REALLY want them
If they are in the MODs
Find them on the basis of other features

What is timeline in relation to when it would even be useful
Semantic similarity
Thinking Fall

Good time frame
Should have first pass of that search and list in October

Nizar and Paul and Alex could work on mapping transitions
What are the phenotypes that we include in those profiles?

Strength of list in relation to test case with and without ancillary characters

Could be done is to say we annotated fin limb throughout tree, what if we took random internodes, would we pull out same candidates

Worth bringing up to advisory board? Skeleton of how it would be done
Idea of how to make it more objective
Get text and number of limb gene associated genes across abstracts in last 10 years

If it's in abstracts it can be done

Could be a set of terms you query against abstracts with gene names

MOD breakout group:

MGI:
* Mapping of MP to EQ is incomplete - there is no funding to make complete - but Monte says mapping is continuing
** Chris says the morphological parts are reasonably complete and high quality
* There are some expression annotations that reference MGI anatomy terms that have not yet been linked to MA or EMAPA anatomy ontology terms
** But they continue to do this so data reports will automatically get better
* Terry and Judy advocate creation of any MGI reports needed by Phenoscape, on the MGI download site
*Document what the mods are doing (filtering, etc.) to generate the download file that Phenoscape is picking up
Summary:
   * MGI data problems - require specialized reports from MGI (this can probably be accomplished by the end of year 2:
            * Complex genotypes - human transgenes
            * Terry will look at ZFIN report and work with Jim to generate a new report format
            * Not all MA terms have IDs
            * New terms have IDs
            * MGI is adding IDs retrospectively as needed
            * MP>EQ mapping is incomplete
            * May not be a serious problem because majority are mapped
            * PATO project is working on this, but winding down
   * To do:
            * Write out requirements for reports and what is filtered out by ZFIN & MGI for those reports - this becomes the Phenoscape standard data format- create Google doc so that all can contribute to central place
            * Write manuscript that includes standards for appropriate use of MOD data

Xenbase:

* 100 papers on limb phenotypes - mainly limb regeneration
* Have completed ~6
* Will require 50 FTE days to complete
* VG and Yvonne have been using Skype to achieve curator consistency

==============================

- Discussion of knowledge base expression functionality
- Cal Academy follow-up (Nizar Wasila and Alex)

Knowledgebase MOD working group

CalAcademy follow up group

Phenoscape API for MOD linkout
* Monte wants to know what is known about a gene in other taxa
** Could go to Phenoscape and see view of phenotypes from all MOD taxa
** All phenotypes associated with a gene through Phenoscape
** Link to that from ZFIN, or use API to show result in ZFIN
* Or, structure-centric query, see what genes affect
* Need orthology data for some of these use cases
* Where could you link out from a gene page from a MOD?
** Phenoscape gene page is boring
** But would be useful if page synthesized data across MODs
* Perhaps structure-based linkout is most useful, since could show phenotypes across tree

BREAKOUT: Homology I presentation and breakout group (Hilmar, Paula, David, Todd, Chris)

- Need to add the thousands of default homology assertions to first slide for AB
- Infraorbital series example of serial homology built into the ontology
- Expressiveness required for negative homology assertions:

o Negative assertions that accompany a positive one can be reduced to opposing evidence tagging the positive homology, marking the evidence as controversial

o Negative assertions that stand in opposition to the ontology to be resolved by modifying the ontology (e.g., by splitting classes), or altering how the ontology is being used for annotation (e.g., by using a taxon-specific class if one already exists).

Thursday, May 9, 2013

Melissa, Monarch presentation

--Standardizing MOD data, e.g., zfish and mouse, genotype & phenotype, but worm db, allele & phenotype
--- Make available services and tools for other programs to use
- Education & outreach to make phenotype data available
- Populating triplestore with genotype-phenotype data

BREAKOUT: Homology II (Hilmar, Paula, David, Todd, Chris, Hong)

- Can reasoning over 'built in' serial homology (e.g., digit class 'digit homologous_to some digit') yield too “promiscuous” results?
Query: left manual digit 1 and all its homologues.
Returns phenotypes for all digits have the left, right, fore, and hind limbs. - This is OK because not many phenotypes actually;

David & Paula think it's ok to get all results above; can narrow down by, e.g., suggesting that we could turn off serial homology assertions

Simply returning results versus also explaining them and letting the user control them on the fly; or: how far can we limit on-the-fly user choice:
- E.g., 'including parts' -- to naive user the results from w/ and w/o including parts look very different, though overlapping
**UI note from Chris: put substructures later in list so user understands overlap more readily.
--Can we do something analogous,'include homologues’ checkbox?
--Include homologues and parts: what should this mean? Homologues plus their parts? Parts plus homologues plus their parts? Parts plus homologues plus their parts plus homologues of parts plus their parts?
Chris: suggests that we do all the possible searches and show biologists to judge what would be valuable/reasonable to return.
Todd: suggested that we not make subjective judgment but use semantic similarity to assess which set of results is most valuable
Chris: UI influences user reaction
Use the broadest possible definition? Would it be helpful to find ways to rank results in interface - to help with very broad results

Paula suggests that controversial homologies be relegated to annotations that can be displayed as info about a class
* Up to the user to view and potentially include additional classes in their search
* Alex to reconcile each homology statement with respect to Uberon

Two models in OWL, one of which returns instance data, but is more complex; do we need to return instances? Museum database connection to individual specimens would be such a case.

BREAKOUT: Interoperability and API (Todd, Paula, names to be filled in)

Possible data from Phenoscape: taxa, specimens, phenotypes, character descriptions, anatomy, pubs, genes, gene expression,

Users:
1. Evolutionary functional genomics/molecular evolution (genome-centric information cloud) - need phenotype to navigate to taxon
- Would want all data; likely to integrate with expression data, loss, etc. systems biology graphs, Gill's stuff
Monte: would want basic ape so all data can be provided;

2. Biodiversity informatics (taxon based) - need phenotype to go to gene....
--Museum folks e.g., eol, museums, morph, treebase, open tree, iPlant,
Want: all but genes,

3. Non-MODs -- need more use cases;

Services provided by Phenoscape:
1. PhenoBlast: similarity analysis of entities (anatomy) for gene expression (in anatomical entities) or anatomical entities
2. precalculated genesets across MODS associated with anatomy,
3. Have a PhenoMine (Judy), a way of serving up data, would be familiar to users coming from mod background;
4. Taxon phenotypes, pheno-taxon matrix;

Funders and users need to come to MOD sites from human biology entry point to solve human disease standpoint;

Summary: id external groups and help them; work with internal MOD groups w/in Phenoscape

Discussion of Phenoscape Year 3 Timeline

Livezey dataset? Not curated yet - huge number of characters would take Nizar months to annotate.

Melissa: Could do definition check and interface development at the same time.

Start with interface work this summer - mockups, discussions.

Work on Matrix/Tree tool should start now - have something coarse/in progress ready for SVP meeting (LA).

Similarity Tool - work with Monarch, lots of similarities, overlap.

Capstone - candidate gene list by August

Conferences - SVP, Barcelona.

Report out from breakout groups (Thursday afternoon, May 9)

1. Monte breakout (Mod closeout)

1. complex genotypes – terry will work with Jim to generate a report format (this can be accomplished by the end of year 2)
2. Not all MA terms have IDs
a. May not be a serious problem
b. New terms get IDs
c. MGI is adding IDs retrospectively as needed

Manuscript that includes challenges with integrating data from different MODs.

2. Todd: Capstone

-

3. Hilmar: Homology

- incorporate uncontroversial homology statements
- timeline

4. Jim: KB interface:

- Reduce ontology machinery displayed – made Jim think in terms of semantic similarity
- View data on trees (support for CalAcad projects)
- Tool to do similarity comparisons between lists of entities;
- Display of image data/glossary (Paul and Nizar)
- Display of gene expression data in relation to evo data
- Display of taxonomic data

5. **Wasila: papers**

- Annotation of characters across anatomy will involve ontology development

ACTION ITEMS THAT CAME OUT OF THE PROJECT MEETING BREAKOUTS AND SUBSEQUENT DISCUSSION ON MAY 9, 2013

Also see timeline here:
https://docs.google.com/spreadsheet/ccc?key=0Apgi__7Z2km5dG5qdEdmbXhMcWdLbl9fVjJQU21ERVE&usp=sharing

Jim:

- work with Terry on MOD data import
- integrate Xenbase phenotypes
- initiate standard data format report within Google docs for MOD paper
- integrate homology into KB (not too much engineering work)
- UI development

o integrate homology

- Integrate matrix download/tree mapping tool (support for CalAcad)
- Develop phenotypic profile comparison/similarity tool (Monte, Jim)
- Help Paul and Nizar get a start on anatomy/glossary interfaces
- Update Phenex/Charaparser for evaluation before 1 July 2013
- Other: software documentation, manuscripts, iPlant

Alex:

- to develop a list of candidate genes from standard sources (check with Monte) and develop a table using Go evidence codes, species, and type of action (position, etc. (timeline – by fall, same time as semantic similarity to be done)

- follow up on Uberon homology assertions, reconciling with our list
- curation of 200 characters for test data set (1 month?) plus new term additions (1 month)

Paula:

- confirm timeline of temporary hiring, 1.5 years at USD?
- Advertise for developer
- Help Paul and Nizar get a start on their interface work
- Work with Wasilla to collect papers for semantic sim/Charaparser eval, even number of characters per paper
- Candidate list of genes for capstone eval -

- Workshop for interoperability in Jan/Feb?

Nizar:

- curation of 200 characters for test data set
- work on interface with Paul

Wasila:

- size annotation guidelines with Hong, others
- selection of 200 characters across a set of papers (or curation?)
-

May 10, 2013 (Chicago) Advisory Board meeting with Phenoscape

Paula's intro:
Cyndy: How are you funding your 8 collaborative projects?

Hong's presentation (Charaparser)
AR: what are you trying to scale?
John: Highlighting context in document might be helpful

Schofeld: Quick access to information from article; 2 min from expert user to assign best model; iterative process, y,y,n,n,n pooled, cycle back…consider annotating 70% of an article and move on;
Key piece is an interface such that expert doesn't have to dig for ontology terms; here is the complexity of eq syntax;

- Concentration on developing an efficient user interface
- Will be published later this year

Alan: choosing faster than dragging;
Enable re-running such that users do not have to go through it all again

Cyndy: what are your metrics for determining success? (Hong, matching eq statements).
Goals…. Precision recall – 60-70%

Alan: How to help (or have Google help) you put studies together.

4:30 pm (Friday) notes from Advisory Board conversation with Phenoscape after their closed door session; also will be in their advisory board report

Phenex/Charaparser: show other EQs

Put into place a curation review process where one person immediately reviews the comments by someone else; useful chain of information of how annotation decisions are made; spirit of the comment is to have a record of decision making regarding annotations;

Independent audits of the annotations; optimize curation process – need clear metrics of optimality; bring in hci or industry partner to help with….

Document what our expectations are about user community, possibly not realistic, Phenome web infrastructure, trade data and queries, see how they compare (!), and see if you can work collaboratively.

Please provide more details on collaborative projects next year.

Ontologies – attribution, nice to see that we are thinking about it, but more details on models of attribution; historical tracking of attribution;

Impact of disagreement on upper level terms – look for implications

Taxonomic ontologies- should look together to eol, etc. for source classifications,

Bring manual curation audit process (manual) into community tools area

Homology clarification

Semantic similarity: did not understand ss calculation for the lineages. Do not dilute effort to go to other domains.

Broader impact piece: copyright and licensing so they can be shared; junior biocurators should express techniques etc.

Future, usability of tools; streamline annotation and curation, make tools more approachable;

Improving ui will have a dramatic impact;

Generate results; work on learning something about fin/limb;

Generalizing phenex for future funding, with other communities, in other domains;

Like to see us doing experiments – end result is testing of hypotheses, someone has to go out and sequence genes from specific species, tends to be a bit lacking; drive experimentation informatically;

Interface:

- John: woman who will how to design user experience studies; can do it over Google hangout;
- University hci groups might take it on….work out a set of user studies
- People who have excellent design sense,

Todd asked about API, phenotype data sucked in by other projects:

Phenoscape I interface was driven by our API; Phenoscape II,
John: use API in place; if easy to address use case, develop for it. Endless process and doing

work that no one needs