Report of Phenoscape Scientific Advisory Board

San Diego, May 1st 2014

General

The Advisory Board were impressed by progress to date and congratulated the members of the project on their innovation, creativity and professionalism. The key challenge that the Advisory Board set the project for the next year was to validate the whole approach, end-to-end using a gold standard curated dataset; effectively demonstrating the success of the approach.

Our most important comments fall into five main categories;

1. Testing of the whole approach end-to-end on a gold standard dataset

- Can known fin/limb transition genes be retrieved?
- Can genes be associated with aspects of limb bud development or skeletal pattern development?

This is the highest priority for next year. A gold standard expert curated dataset is needed against which to measure success, and a metric that provides a scientific output as a main goal. This will give insights into annotation quality as well as the overall strategy.

2. Development of curation tools and metrics

The Board was impressed with the work that had been done on inter-annotator variability and benchmarking for the assessment of Charaparser. Whilst appreciating that with the goals for next year requiring the maximum annotation speed possible, we felt that the work on Charaparser was not now a priority and the work to date had shown that it still had major problems in comparison to manual curation. We recommend diversion of effort from Charaparser.

We support investigation of whether EQ writing is actually necessary for the power of the semantic similarity searches, and though this is tied into our recommendations on the pursuit and benchmarking of semantic similarity algorithms, assessing the utility of a “bag of terms’ versus a full set of curated EQ statements is something we would like to see.

We reiterate our support for the generation of a gold standard dataset against which to test semi automated methods and further development of metrics for the inter-annotator variability, which at the moment seem, following our extensive discussion on this, to be somewhat problematical,. We are however pleased that a publication is being prepared.

The proposed acquisition of further non-limb/fin characters is an important activity but this needs careful prioritisation with respect to the other consolidation activities that are going on. It is important that the potential timescale for this is realistic.

The presence/absence matrix development is promising and very interesting and we are enthusiastic about the pursuit of this, and the ability to reason over presence/absence statements. We would like to see presence/absence tied to inbred line, species, heterozygotes, homozygotes etc. and think that this could be a very useful, important and fruitful avenue to pursue.

There was suggestion that an online version of Phenex might be created; we recommend that you work closely with Monarch who may also be aiming to provide online phenotype annotation tools. Also of relevance might be the online annotation efforts currently being made by Phenodb for humans (Ada Hamosh).

3. Semantic similarity, reasoning and the knowledgebase.

We strongly support the decision to put major effort into the assessment of semantic similarity measurements, their validation and integration of the metric into the knowledgebase.

We recommend bringing GO process annotations into the knowledgebase. Partition out the processes during curation to be in a better position to analyze the genetics of development. Hone in on lifestage and subprocesses for better capture of the processes. Can organize by process based on anatomy and the classification of the papers where they are described.

4. Work on development of user interface

Put together a minimal UI quickly to relieve pressure on Jim. However, avoid spending too much time on UI requirements because it is more important to make sure that things work. For example: reproduce your command line capabilities, eg. get batch of data in, give a matrix file to download etc. Talk to John about leveraging Google cloud resources to make this easier

5. Outreach efforts to the scientific community need to continue to work to make products of the project more broadly available. Aim to have things ready for an external review, to be scheduled in a few months. Also, in addition to shifting focus for Biocurators to annotation, please also shift to bring their attention to organisms more directly relevant to the fin/limb transition (e.g. not just tetrapods).

Future Plans

- To avoid communication overhead, rebuild curation tools to facilitate communication among curators in the tool. Build on the success of the term requester.

- Take advantage of offer for free web app resources from Google. They can also offer technologies for making the tools more collaborative.

- At minimum, sustain the curation effort and survival of the tools. Tie them into science grants.

- Decide what you most want to do in the futire. What other science questions do you want to answer? If you are going to go to enhancers, consider also looking at epigenomics, microRNA expression and targets in control of gene activity, disease phenotypes. Variation of the phenotypes. Spin off tooling if you want to do that. Establish network of partnerships.

- Look at i-Corps program at NSF to help identify and develop potential markets. Consider offering services to users in other domains. Possible grant opportunities with NSF Cyberinfrastructure, NIH Collaborative projects or ORIP (NIH Office of Research Infrastructure programmes – part of the OD) . Maybe discuss possible infrastructure proposals with Franziska Grieder at ORIP.

- Within Europe Horizon 2020 will have one big huge biodiversity informatics call, and US institutions will be able to receive funding.

PNS
Thursday, 10 July 2014

Notes from meeting taken from contemporary Google Doc.

Performance to date

Apparently have not yet tested success of their approach.
Presence/Absence work looks very promising.
Creative team - lots of innovation
Very professional job. Glad everyobdy still willing to work together.

Next year

1. Highest priority is the testing of the approach with a clear metric of success.
Two approaches: Gold standard against which to measure is highest priority, and/or a metric that gives them science as the main answer.

A) Super expert curated dataset accounting for all differently curated sets (“correct”). Compare different curator outputs with that. Less risk here.

B) Go for the real data recall of the candidate genes -- but that depends on semantic similarity which itself isn’t validated. One can argue that we’re testing the system together, so separate validation less important. And in the end you want the system as a whole to work. Still risky but critical to do this soon. Set of characters they are putting in to find the genes has to be significantly overlapping with the curated characters.

Both of these are important -- end-to-end is really good to do. Intermediate testing (gold standard or semantic similarity validation) is also important.

2. Put together a **minimal** UI quickly to relieve pressure on Jim. However, avoid spending too much time on UI requirements because it is more important to make sure that things work (A). For example: reproduce your command line capabilities.Get batch of data in. Give a matrix file to download. Talk to John about leveraging Google cloud resources to make this easier.

3) Divert effort from Charaparser. This seems to be a lower priority. Appears to be too difficult a task to do it well. Go ahead and test the result of rpproposaos.l

4) Continue work on the presence/absence matrix . Tie presence/absence to inbred line, species, heterozygotes, homozygotes.

5) Outreach efforts need to continue to work to make products more broadly available. Aim to have things ready for an external review, to be scheduled in a few months. Also, in addition to shifting focus for Biocurators to annotation, please also shift to bring their attention to organisms more directly relevant to the fin/limb transition (e.g. not just tetrapods).

6) Include in the semantic similarity metric and bring into your database GO process annotations. Partition out the **processes** during curation to better set up to analyze the genetics of development [Bryan’s point about honing in on lifestage and subprocesses for better capture of the processes]. Can organize by process based on anatomy and the classification of the papers where they are described.

Future

To avoid communication overhead, rebuild curation tools to facilitate communication among curators in the tool. Build on the success of the term requester.

Take advantage of offer for free web app resources from Google. They can also offer technologies for making the tools more collaborative.

At minimum, sustain the curation effort and survival of the tool. Tie them into science grants. The second grant idea had the driving science, so that may be the most important.

Decide what you most want to do. What other science questions do you want to answer? If you are going to go to enhancers, consider also looking at epigenomics, microRNA expression and targets in control of gene activity, disease phenotypes. Variation of the phenotypes. Spin off tooling if you want to do that. Establish network of partnerships.

Look at i-Corps program at NSF to help identify and develop market. Consider offering services to users in other domains. NSF Cyberinfrastructure or NIH Collaborative something or ORIP -- what might be coming up. Francesca Grieder. Discuss a possible infrastructure.

Horizons 2020 -- there will be one big huge biodiversity call. You could be a part.