PhenoStacks: Cross-Sectional Cohort Phenotype Comparison Visualizations

Michael Glueck, Alina Gvozdik, Fanny Chevalier, Azam Khan, Michael Brudno, Daniel Wigdor

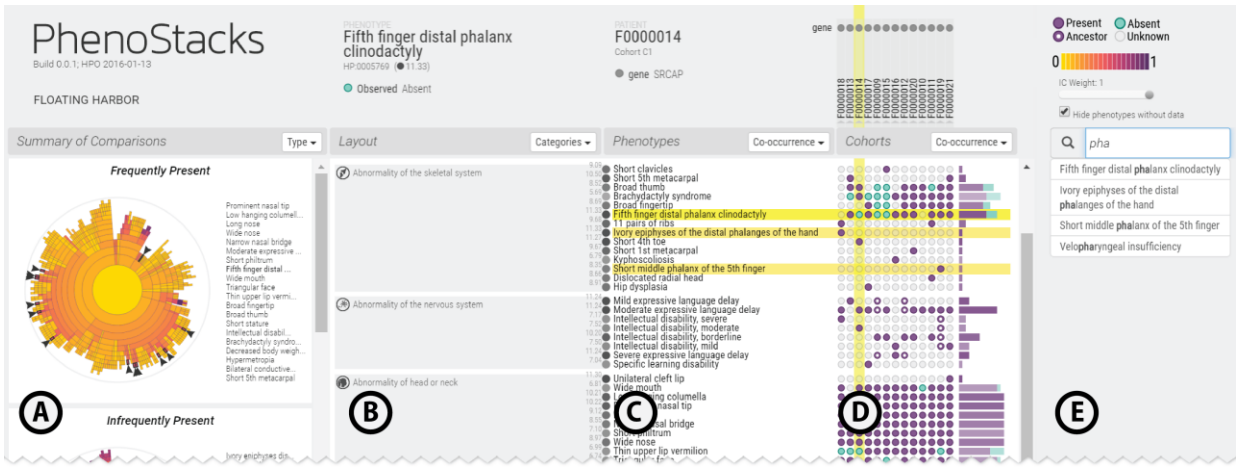

Fig. 1. PhenoStacks employs an observations plot (D) to reveal the distribution of phenotypes (rows) across patients (columns) in a cohort, which can be sorted by patient or phenotype attributes. Similar phenotypes can be grouped based on the Human Phenotype Ontology (B,C). Radial hierarchies (A) summarize global patterns. Views are linked, e.g., search results (E) are highlighted (A,C,D).

Abstract—Cross-sectional phenotype studies are used by genetics researchers to better understand how phenotypes vary across patients with genetic diseases, both within and between cohorts. Analyses within cohorts identify patterns between phenotypes and patients (e.g., co-occurrence) and isolate special cases (e.g., potential outliers). Comparing the variation of phenotypes between two cohorts can help distinguish how different factors affect disease manifestation (e.g., causal genes, age of onset, etc.). PhenoStacks is a novel visual analytics tool that supports the exploration of phenotype variation within and between cross-sectional patient cohorts. By leveraging the semantic hierarchy of the Human Phenotype Ontology, phenotypes are presented in context, can be grouped and clustered, and are summarized via overviews to support the exploration of phenotype distributions. The design of PhenoStacks was motivated by formative interviews with genetics researchers: we distil high-level tasks, present an algorithm for simplifying ontology topologies for visualization, and report the results of a deployment evaluation with four expert genetics researchers. The results suggest that PhenoStacks can help identify phenotype patterns, investigate data quality issues, and inform data collection design.

Index Terms—Cross-sectional cohort analysis, Phenotypes, Human Phenotype Ontology (HPO)

1 Introduction

The insights we garner by deepening our understanding of human genetics can improve general medical practice, from better predicting the side effects of medical interventions [14] to guiding new approaches to treat everyday diseases [1]. As studying rare genetic diseases can lead to broad insights into human genetics, genetics and genomics scientists, as well as clinical geneticists, continue to unravel the complex interactions between genes, environment, and visual manifestations (i.e., phenotypes) [14].

- *M. Glueck is with Autodesk Research and the University of Toronto (e-mail: mglueck@dgp.toronto.edu)*
- *F. Chevalier is with Inria (e-mail: fanny.chevalier@inria.fr)*
- *A. Khan is with Autodesk Research (e-mail: azam.khan@autodesk.com)*
- *A. Gvozdik and D. Wigdor are with the University of Toronto (e-mail: alina.gvozdik@mail.utoronto.ca; daniel@dgp.toronto.edu)*
- *M. Brudno is with the University of Toronto and the Hospital for Sick Children in Toronto (e-mail: brudno@cs.toronto.edu)*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org.
Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx/.

Phenotypes are observable and measureable patient traits primarily caused by genetic variation. They describe abnormalities with respect to morphology (i.e., structural features such as having a broad thumb or low-set ears), physiology (i.e., functional features such as cognitive impairment or seizures), or behavior (e.g., depression or impulsivity). Genetics and medical researchers study the spectrum of phenotypic abnormalities associated with rare genetic diseases to garner insights into the multitude of factors affecting disease characteristics and manifestations. For example, the more comprehensive patient phenotype reports, the higher the chances of discovering the gene variants responsible for a given disease [13].

Although phenotypes are powerful indicators of disease and have a long history of use in diagnostic medicine, formative interviews with genetics researchers indicated that there is a lack of standardized workflows and systematic approaches to phenotype data analysis. Individual research groups typically collect phenotype data using self-designed instruments and protocols, manage internal databases, and conduct analyses using a combination of custom scripts, spreadsheets, and statistical packages. Unlike gene-oriented analyses (e.g., [37], [40]), dedicated tools for phenotype analyses do not exist. This is in large part due to use of natural language to describe phenotypes within electronic health records (EHRs) or in published case reports. Such language results in unstructured phenotype data. While access to

centralized disease registries is growing, phenotype data collected by different investigators and institutions varies widely depending on collection protocols, diagnostic methodologies, and resource constraints, as well as measurement subjectivity and investigator diagnostic interests and priorities. Many notations, abbreviations, and synonymous terms have evolved [32] and are reported at different granularities of detail [24]. These factors can introduce uncertainty regarding the actual observations of a clinician when comparing published results [33] and complicates computational analysis.

Structured approaches to phenotype data aim to address these issues of consistency, completeness, granularity, and computability. The Human Phenotype Ontology (HPO) [18] is the first ontology for phenotypes; an on-going initiative to standardize terminology and add structure by defining relationships between phenotypes (i.e., semantic, logical, hierarchical). The HPO also facilitates interoperability with external disease databases that link genes, phenotypes, and diseases (e.g., OMIM, OrphaNet) [34]. Leveraging the HPO structure and external resource integration enables computation on phenotypes that is not possible using current nomenclatures alone (e.g., ICD-10). For example, the similarity of patients with non-overlapping phenotype observations can be calculated and concepts like diagnostic significance can be quantified [19]. Despite the popularity of this resource in the genetics community, very few tools are available to help scientists visualize, analyze, and compute on cohorts of patients described using terms from the HPO.

This work contributes PhenoStacks, a visual analysis tool that enables genetics scientists and researchers to explore phenotypic variation in cross-sectional cohorts of patients with a rare genetic disease. Interviews with domain experts identified a series of high-level cross-sectional cohort analysis tasks that guided the development and design of PhenoStacks. Within PhenoStacks, potential metrics of interest (e.g., frequency, similarities, entropy) are automatically extracted and encoded in summary charts that guide a goal-oriented visual exploration of detailed patient phenotype plots. The tool can be used to investigate the distribution of patient phenotypes within a single cohort or compare and contrast similarities and differences of phenotype distribution across patients between cohorts. In addition, we contribute a novel algorithm for simplifying ontology topology visualization and visualization concepts that can be applied to other domains where instances of taxonomies or ontologies are analyzed. PhenoStacks is available as an open source project at phenostacks.org.

2 Background and Related Work

Rare genetic diseases afflict an estimated 350 million people world-wide, with approximately 30 million patients in the US alone [10]. While there are more than 8,000 named rare genetic diseases, many more are yet to be discovered or classified [32]. Those that have been identified are difficult to diagnose, as their low occurrence means clinicians only encounter a handful of cases over their career. On average, these patients will be misdiagnosed twice, interact with four clinical specialists, and wait seven years for a correct diagnosis [39].

Clinical specialists and genetics researchers both contribute to the advanced study of rare genetic diseases in a virtuous cycle. Through direct interactions with patients, clinicians are the primary collectors of the data used by researchers to deepen our understanding of human genetics. Research results are, in turn, applied to developing more robust diagnostic tests. These advances are critical to improving patient quality of life by reducing unnecessary diagnostic testing and mitigating the uncertainty of a diagnosis. Ultimately, expanding our understanding of the underlying genetic factors affecting diseases is key to realizing the vision of personalized healthcare promised by analysis of patient genomes [13].

Given the importance of phenotypes, both specialists and genetics researchers stand to benefit from advanced tools that promote the use of consistent phenotype terminology and facilitate comprehensive phenotyping. This is precisely a scenario where visual analysis tools can excel: combining the strengths of perception and computation into an interactive process that extracts knowledge from data [17].

2.1 Role of Phenotypes

When studying rare diseases, patient phenotypes are the best method to describe symptoms and manifestations, and to identify similar cases [1][4][13][30][32]. Generally, complete and granular phenotypes are necessary to differentiate between disease subtypes and to guide diagnostic tests to confirm a specific disorder using advanced modalities (e.g., genetic testing). Due to resource constraints, current technologies render it impractical to compile a complete and granular workup of every phenotype for each patient (i.e., deep phenotyping [32]). The onus thus remains on clinicians to prioritize diagnostic tests and compile consistent and detailed phenotype reports. The reality, however, is that researchers would benefit from more detailed phenotyping than is typically necessary for diagnosis.

Although research initiatives are exploring computational approaches to infer phenotypes from EHRs (e.g., PheKB) and sharing patient data to improve the accuracy of these methods (e.g., PCORnet, NIH Collaboratory), EHR coding schemes (e.g., MeSH, ICD-10, SNOMED CT) are inconsistent and incomplete in their coverage of phenotype terms [41][44], and lack the granularity required to study rare diseases [41]. Similarly, extracting clinical information from research articles (e.g., PubMed) remains a challenging area of research [12]. The HPO complements these data mining efforts, since mapping extracted data to HPO terms provides access to existing computational approaches. The HPO is already being used successfully in novel tools that provide systematic methods of entering complete and granular patient phenotype data (e.g., PhenoTips [9]) and that facilitate matching patients with undiagnosed rare diseases based on phenotype similarity (e.g., PhenomeCentral [4], MatchMaker Exchange [30]).

2.2 Visualizations of Genetics Data

Numerous visualization tools have been developed to assist geneticists. Most of these systems visualize genotype data using three classical visualization techniques (see Schroeder et al. [35] for a review). Heatmaps are commonly employed to represent genomic values in a compact color-coded matrix (e.g., cBio [6]) or in a circular view (e.g., CircleMap [42]). Space-filling layouts based on genome coordinates have been applied to genome browsers (e.g., Savant 2 [7]) and genome comparison visualization tools (e.g., Circos [21], Mizbee [26]). Finally, network representations have been used to explore large biological networks (e.g., Cytoscape [38]) or gene pathways (e.g., VisANT [15]). Specific visualization challenges have also been addressed, such as introducing a visual language for depicting genome assemblies (e.g., AbyssExplorer [27]).

Fewer visualization tools have focused on comparing phenotypes; most existing systems are designed to better understand the causality of gene variation on phenotypes. Representations, such as Manhattan plots, have been used to reveal complex associations between genes, proteins, and phenotypes (e.g., Arena3D [36], PheWAS-View [28]). The HPO is rarely used for visualization and was only recently first employed in PhenoBlocks [11]. In this prior work, we used the HPO to visualize phenotypes in clinical diagnosis settings, supporting the pairwise comparison of patient phenotypes using explicit encoding. In the present work, we turn our focus to genetics researchers conducting cross-sectional cohort studies, where the distribution of phenotypes is compared across many patients.

2.3 Visualizations of Cohorts

Supporting analysis of patient cohorts has been addressed in the visualization literature (see Rind et al. [31] for a review), with a focus on longitudinal cohorts in clinical settings. Recent works investigated analyzing temporal constraints of cohort membership (i.e., COQUITO [20]), exploring temporal events within a cohort (i.e., CAVA [45]), and comparing temporal events between cohorts (i.e., CoCo [25]). Although our work focuses on cross-sectional cohort analyses, we share the high-level goals of comparing patients within and between cohorts to identify patterns of cohort membership in an exploratory context, however we focus on phenotypes, instead of temporal events.

2.4 Visualizations of Ontologies

Ontologies are used to capture the conceptual structure of a domain. Using graphs, the relationships between entities (e.g., causation, inheritance) can be encoded and establish a common ground for discussion and knowledge sharing. Many visualization tools have represented ontologies [16][22] and bio-ontologies [43]. Carpendale et al. [5] review bio-ontologies from the perspective of visualization, identifying challenges and research opportunities, such as annotating data and visualizing annotated content. Katifori et al. [16] surveyed a variety of visualization approaches in the literature, identifying indented lists and space-filling hierarchical layouts (e.g., icicle, radial) as popular and effective approaches. Katifori et al. note several problems remain poorly addressed in most existing systems, including clutter reduction (topologies can be complex), structure (visualizations fail to convey overall structure), scalability (large ontologies are not handled well) and inspection (querying ontologies). Clutter reduction and structure are addressed in PhenoStacks.

3 Cross-Sectional Cohort Analysis

Cohort studies are frequently used in medicine, in both longitudinal and cross-sectional designs. Longitudinal analyses are often used in clinical studies, such as evaluating the effectiveness of treatments [20][25][45]. In contrast, cross-sectional cohort studies are typically descriptive, such as quantifying features of a patient population (e.g., the prevalence of a specific symptom of a disease) or evaluating metrics to differentiate subpopulations (e.g., whether a biochemical marker correlates to disease severity). In tandem, cross-sectional and longitudinal study designs provide complementary perspectives on patient data, to first identify disease subtypes and causal factors, and then study their correlations to disease progression and treatments.

To better understand how cross-sectional cohort analysis is used in genetics research, semi-structured interviews were conducted with six domain experts: four experienced genetics researchers (MD/PhD; 6-12 years of experience), one research manager (MSc; 8 years of experience), and a research fellow (PhD; 4 years of experience). Their specializations spanned a range of rare genetic diseases, including neuromuscular, metabolic, and inflammatory bowel diseases. We first report on barriers, goals, and types of analyses and then relate these to an abstract task typology from the visualization literature that can guide the design of tools for medical cross-sectional cohort analysis.

3.1 Barriers and Pain Points

Researchers described that published results of cross-sectional cohort studies often report only descriptive statistics of phenotype occurrence (i.e., frequency, mean, variance). Although this provides a sense of the average patient in a cohort, it masks the actual distribution and co-occurrences of phenotypes for each patient. Preserving the granularity of phenotype reporting and standardized naming of phenotypes would improve the ability to reuse the results of these studies.

In any research center, there are very few patients for any given rare disease. The researchers we interviewed explained that detailed phenotypes of diagnosed cases were once disseminated within the scientific community through case reports of single patients, but this practice is in decline with increasing availability of large curated disease registries. Sharing patient phenotype data across departments, institutions, and countries is thus easier, and studies can now include larger cohorts (i.e., 10s to 100s of patients). With more data, however, data quality is an increasing concern (i.e., completeness, consistency, and granularity). Although deep phenotyping is the ideal scenario, the practice is prohibitive due to its time-consuming nature, required expertise in the disease, and lack of available resources. Reaching consensus on what phenotypes should be collected and reported is a slow process. In particular, when faced with patients with rare or undiagnosed diseases, the researchers noted it can be equally difficult to decide at what granularity to report phenotypes. Another concern was efficiently integrating data across multiple databases as the size of data continues to grow. We determined that structured phenotype data can begin to address these issues.

3.2 Analysis Goals

In genetics research, cross-sectional cohort analyses may focus *within* a cohort (i.e., patterns across patients), or *between* cohorts (i.e., patterns between cohorts). Here, we summarize the common tasks described in our interviews.

3.2.1 Within-Cohort Analysis

Within-cohort analyses are useful to discover **emergent patterns (W1)** of phenotypes between patients (e.g., frequency, distribution, co-occurrence). The results provide insight into the scope of a disease (e.g., localized vs systemic) which can differentiate symptoms from unrelated abnormalities.

Comparing phenotypes between patients in a cohort can identify **outlier patients/phenotypes (W2)**. Phenotype presentations that are atypical of the cohort, or very rare phenotypes, can indicate special cases where the patient may not respond to existing therapies or there is potential to better classify disease subtypes.

Subcohort discovery (W3) can help develop more effective diagnostic tests and treatments by identifying new disease subtypes. This involves discovering groups of patients within a disease cohort that can be differentiated from each other based on measurable attributes (e.g., phenotypes, causal gene variants, biochemical markers).

Within-cohort analyses can also help **audit data quality (W4)**, ensuring that phenotype data collection is consistent and complete. For example, is the granularity sufficiently detailed to differentiate subcohorts? This is particularly important for larger studies that involve multiple investigators, institutions, and disease registries.

These analyses (W1-W4) are typically informal and exploratory undertakings. Hypotheses are later tested using statistical approaches. Such results can be used to quantify **disease prevalence (W5)** within subpopulations (e.g., geographic, demographics, ethnicity). Studying these specific incident rates can identify potential environmental influences on diseases or groups requiring special consideration.

The results of these analyses often lead to improved **disease characterization (W6)** based on identified patterns. The application of these characterizations in clinical settings can set expectations about disease severity, clinical prognosis, and the effectiveness of available treatments. This can help doctors to better plan interventions and to explain a disease to patients and families.

3.2.2 Between-Cohort Analysis

Between-cohort analyses enable **comparison of patterns across disease subtypes (B1)**. For example, age of disease onset is a factor of frequent interest. Patients who exhibit symptoms of a disease at younger ages more typically have congenital forms (i.e., inherited), while patients who develop the disorder later in life may have been exposed to environmental triggers. In addition, by comparing patients with less- and more-severe forms of a disease, more general patterns can be isolated. While outlier phenotypes and special patient cases are of significant research interest, very consistent and general patterns are most applicable to developing useful treatments.

Between-cohort analyses are important to **validate data quality (B2)**. When collecting new patient data for a cohort study, it is important to compare to existing registries to ensure the local cohort is representative. When publishing results, it is critical to describe salient similarities and differences to previously published results (e.g., case reports). When compiling patient data across registries, it is important to compare cohorts to evaluate whether it is appropriate to assimilate them. These comparisons can also be used to evaluate the observational methodologies of registries.

Results of between-cohort analyses can further be used to **inform clinical practice (B3)**. Subpopulations may be differentiated by non-clinical attributes (e.g., ethnicity) to determine whether some subpopulations require additional consideration (e.g., more frequent screening). Since the presentation of phenotypes can be influenced by treatment, analyzing phenotypes across different locales can provide insights into geographic medical practice variation.

3.3 High-Level Tasks

We synthesize the identified cross-sectional cohort analysis tasks to common visualization tasks using Brehmer & Munzner's multi-level typology [3]. All tasks fall under the *discover* branch of why tasks are performed. Geneticists engage in a variety of *search* and *query* sub-classifications, as described below:

Explore → Summarize
emergent patterns (W1), disease prevalence (W5), disease characterization (W6), comparison of patterns across disease subtypes (B1)

Locate → Identify
outlier patients/phenotypes (W2), subcohort discovery (W3), inform clinical practice (B3)

Browse → Compare
audit data quality (W4), validate data quality (B2)

4 Design Process

Having identified the common tasks of genetics researchers, we engaged two of the researchers in an iterative design process to evaluate the appropriateness of visual representations and to ensure the designs complemented their existing research workflows and mental models. We focused on analysis goals that aligned to exploratory tasks, rather than the application of research outcomes: summarizing patterns within and between cohorts (W1, B1), identifying outliers and subcohorts (W2, W3), and auditing and validating data quality (W4, B2). Each researcher provided us with a sample dataset of phenotype terms for 20-30 patients, which was representative of the size of cohorts they typically investigate.

4.1 Phenotype Distribution and Frequency

To support *Explore → Summarize*, we wanted to emphasize both the distribution of phenotype observations for each patent, as well as the frequency of each observation for each phenotype. We informed our design choices by eliciting feedback from researchers through a series of static paper-based prototypes. We considered variations using histograms, dot plots, scatterplots, and matrices to summarize and encode the distribution of phenotypes across patients and between cohorts. The researchers agreed that the histogram view provided a good summary, especially when comparing the silhouettes of phenotype observations between cohorts, but did not like that the association of observations to specific patients was obfuscated. The researchers preferred the matrix layout for looking at phenotype observations of patients, in particular a design variation that displayed phenotypes as rows, patients as columns, and encoded phenotype observations using colored cells. As matrices are often used to summarize results of genetics research, the researchers were familiar with this representation. However, this familiarity also led to confusion, since the use of colored cells resembled gene association heatmaps. One researcher commented it was awkward to see different data represented in this format, *"I know it's right, but it looks wrong."*

Based on this feedback, we refined the visual appearance of the matrix to de-emphasize cells and encoded values using row- and column-aligned glyphs. We integrated histogram summaries adjacent to each row to facilitate frequency comparisons. We call this view an *observations plot* (Fig. 1D). To compare two cohorts, two observations plots are juxtaposed and the histograms are re-aligned to a common baseline to facilitate comparisons using the silhouette. The result is visually similar to existing tools (e.g., UpSet [23]) which were also designed to facilitate visual summaries and comparisons.

4.2 Summarizing Phenotype Patterns

Locate → Identify was addressed by automatically computing metrics of phenotype observations within a cohort and between cohorts. We first asked researchers to describe the patterns they typically care about. Within cohorts, researchers focus on whether phenotypes are frequently present or absent, but they are also interested in which phenotypes do not occur often, as these may indicate outliers. They also evaluate the homogeneity of phenotype observations across patients (i.e., are all observations present, or is there a mix?). Between cohorts they identify similarities and differences, as well as outliers and homogeneity. The researchers noted that they often have difficulty generating accurate summaries of their data due to the variability in their phenotype reports. In particular, although specific terms may be used (e.g., that differentiate the severity of a phenotype), they are more interested in whether the phenotype is present in any form.

Aggregating patterns from granular to general phenotypes based on parent-child relationships of the HPO is an ideal application of the ontological structure. To support the variety of patterns the researchers requested, we automatically calculate all metrics across the cohort data and enable the user to select which ones are displayed. These *summary charts* can be used to identify starting points for deeper exploration of the details of the data (Fig. 1A). A similar approach for seeding exploration appeared in CoCo [25], although their summary metrics were not presented visually. Details of the metric calculations are described in Section 5.4.

4.3 Sorting, Filtering, and Search

To facilitate *Browse → Compare*, sorting and filtering operations are supported in the observations plot. This enables free-form exploration of patterns in data, similar to Bertin Matrices [29]. Sorting by phenotype or patient attributes reveals patterns in the observations plot, while filtering hides less important information from the view. Although none of the tasks elicited from our formative interviews explicitly addressed *Lookup*, to fill the gap in the design space, PhenoStacks also supports text-based search for specific phenotypes using natural language queries (Fig. 1E).

4.4 Phenotype Context and Relationships

PhenoBlocks [11] demonstrated benefits to representing phenotypes within the context of the semantic structure of the HPO. In terms of the task typology, representing phenotype relationships could improve *Explore → Summarize*, *Locate → Identify,* and *Browse → Compare* by adding categorical groupings to the observations plot. Thus, a key research question is whether this structure is also helpful within genetics research usage scenarios. While the researchers we engaged during the design process were not familiar with the HPO, hierarchical clustering is commonly used to add structure to results of genetic tests, using dendrograms to communicate clusters next to a matrix heatmap of test results. Inspired by these charts, and a recent evaluation suggesting that indented lists are more efficient at supporting information searches [8], we prototyped a *layout view* that displays the HPO hierarchy as a top-aligned dendrogram next to the observations plot, grouping related phenotype rows (Fig. 6A). Since the HPO is a directed acyclic graph with multiple inheritance, we converted it to a strict hierarchy by duplicating nodes with more than one parent. While this correctly mapped phenotypes to rows, feedback from researchers indicated duplication of phenotypes was confusing because it gave the false impression that there were more clusters.

Interaction was added to the HPO dendrogram, enabling individual nodes to be collapsed and expanded. One researcher liked having granular control to define a customized tree, but the other felt any navigation was cumbersome and detracted from exploring the actual data. Both agreed that general and intermediate categories were very useful, but that granular categories became redundant. This feedback led us to develop a top-level category layout (Fig. 1B) and a cluster layout, computed by a novel algorithm that simplifies the ontology topology and selects salient intermediate categories while eliminating duplications (see Section 5.7). This addresses the clutter and structure issues in ontology visualizations identified by Katifori et al. [16].

4.5 Lessons Learned

Involving researchers in the design process revealed many insights into how they conduct their analyses. The software researchers typically use to view patient records is designed primarily for data

input and queries. While basic visualizations are supported, the process requires considerable effort to select data rows and columns and checkboxes in dialog configurations. Researchers are interested in their data but want to spend minimal time tuning and configuring interfaces. To this end, we adopt an "opt-out" approach for the summary charts in PhenoStacks (i.e., show more information and then eliminate the unneeded) rather than an "opt-in" approach (i.e., start with a blank slate and build the needed views).

Our overarching goal was to support flexible free-form exploration of phenotype data, where users can smoothly transition between a variety of tasks. We quickly determined that supporting all desired views could not be easily merged into a single visual that would work in all cases. Thus, instead we opted for a range of views on the data that could be integrated in different ways to answer specific questions.

5 PhenoStacks Tool Description

PhenoStacks is a visual analysis tool that complements the existing workflows of genetics researchers, and was developed based on feedback from domain experts. The tool supports the higher-level tasks researchers engage in when analyzing cohort data: *summarize*, *identify*, and *compare* [3]. PhenoStacks can be used to analyze a single cohort or compare two cohorts and could help with data collection planning and auditing, identifying potential patterns in cohort data, and guiding the direction of in-depth analyses.

PhenoStacks was implemented using D3 [2], Python, and libraries such as SciPy to compute hierarchical clustering. In this section, we describe the data model, the user interface design, and how the design goals and user tasks are addressed through different aspects of the tool.

5.1 Data Preparation

Our data model is based on the notion of a *cohort subgraph*, a subgraph of the HPO that represents the cohort phenotype terms as leaf nodes and includes all parent nodes in a traversal to the root term (Fig. 2). The cohort subgraph contains the union of all phenotypes across all patients in the cohort, not only phenotypes thought to be associated with a given disease. Each node is labeled with *present* or *absent* for the phenotype observation of each patient in the cohort. Since the HPO describes "is-a" relationships between phenotypes, observations can be inferred for unlabeled nodes (i.e., if a phenotype is present, so are all ancestors; if absent, so are all descendants). Using these rules, unlabeled nodes are labeled with present or absent. Any remaining unlabeled nodes are then marked as *unknown*.

Since only the labels of leaf nodes in the observations plot are presented, an unknown observation may obscure an observation for a related, but more general, phenotype. This consideration is important because patient phenotype reports vary widely in granularity. Thus, PhenoStacks checks whether any unknown nodes have observations associated with ancestor phenotypes (i.e., an observation for a more general term) and labels them *present-ancestor* or *absent-ancestor*. Note that these observations come from the raw data and not the inferred labels because all unknown nodes would then be labeled as they all inherit from the root (Fig. 3).

5.2 Information Content

One calculation enabled by the HPO and its external resources is *information content* [19], a dimensionless quantification of the concept of diagnostic significance. The information content for a phenotype is higher when its occurrence is associated with fewer diseases and lower when it is more common. Thus, phenotypes with higher information content are more likely to be useful when discriminating between diseases. Information content is calculated across a given HPO instance and is independent of the specific phenotypes and patients under investigation.

5.3 User Interface

The PhenoStacks interface comprises three distinct panels that enable a user to visualize and explore the cohort subgraph from overview to detail: a Summary Panel, a Details Panel, and a Search Panel. Each panel shows a different perspective on the topology of the cohort subgraph and supports different observations.

The **Summary Panel** displays a variety of summary charts that display metrics of phenotype patterns across patients of a single cohort or between two patient cohorts (Fig. 1A). These metrics are presented within the hierarchical context of the cohort subgraph topology. Space-filling radial hierarchies were selected for the summary charts because they compactly encode values and equalize the size of both intermediate and leaf nodes. Visual emphasis on leaf nodes is critical because they are the most granular HPO terms. Similar charts were successfully used in PhenoBlocks [11]. The width of segments communicates the number of leaf nodes, with each leaf having equal weight, and all charts are sorted identically to support comparisons.

The **Details Panel** consists of a layout view (Fig. 1B), a phenotype list view (Fig. 1C), and an observations plot (Fig. 1D) that support the detailed comparison of the distribution of phenotypes across patients and between cohorts. The layout view enables the user to control HPO context by selecting different methods of collapsing, filtering, and clustering phenotypes based on the cohort subgraph topology. The phenotypes list view enables phenotype terms to be sorted. The observations plot is a matrix of the actual and inferred phenotype observations for each patient in a cohort. The frequency of each phenotype observation is summarized in an adjacent histogram. When two cohorts are compared, the histograms are arranged as a silhouette plot between the matrix of each cohort.

The **Search Panel** supports specific queries, including Boolean logic, to search for phenotypes by name or HPO ID (Fig. 1E). The search result set is persistently displayed below and matching results are highlighted in the Summary and Details Panels. Clicking on any phenotype label automatically enters the term into the search panel.

Since PhenoStacks supports fully linked and coordinated views, phenotypes, patients, and observations highlight in all charts, and display detailed information when hovered in any of the views.

5.4 Summarizing Patterns

The Summary Panel presents a variety of metrics in summary charts, explicitly encoded with color using a divergent yellow-purple scale, interpolated in Lab space. These metrics are further weighted by the information content score of each phenotype. Users can control the balance between metric and information content score, whether they are more interested in the metric ranking alone or highly ranked phenotypes with high diagnostic significance.

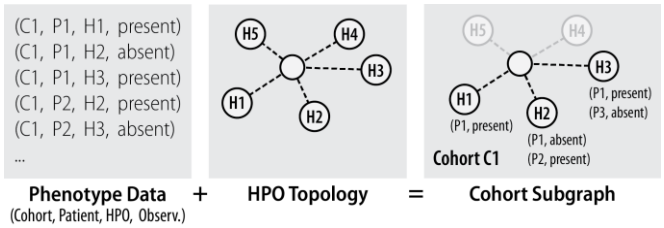

Fig. 2. Illustration of phenotype data translation to cohort subgraph representation via the topology of the HPO. Cohort subgraphs contain only HPO nodes for recorded patient phenotype observations. HPO nodes in the cohort subgraph are annotated with a list of associated patient phenotype observations.

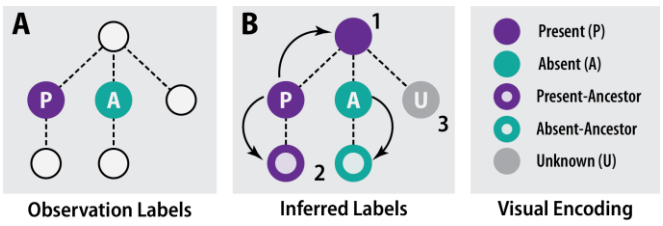

Fig. 3. (A) Cohort subgraph nodes are labeled using patient phenotype observations. (B) Labels for nodes without observations are inferred using HPO inheritance (1) and descendants reflect observations of ancestors (2). Remaining unlabeled nodes are labeled "unknown" (3).

For within-cohort scenarios, metrics of interests are: frequently present/absent, infrequently present/absent, potential outliers, and entropy to represent homogeneity. For between-cohort scenarios, these metrics are: similarly present/absent, differently present/absent, outliers, and entropy. The outlier metric is highest for phenotypes with very few present or absent observations. Entropy is calculated using the Shannon entropy equation between present, absent, and unknown observations.

These metrics are aggregated up the hierarchy and are designed to orient users towards phenotypes that align with the metrics in a general fashion; the summary charts are not designed to be carefully explored by users. A list of phenotypes ranked by the metric is displayed adjacent to each chart to highlight specifics.

These views are designed to address first contact with the dataset. "What are high level features I may be interested in?" For example, these views can quickly give a sense of the homogeneity of the cohort, or the similarity between two cohorts, at a single glance (Fig. 4).

In addition to *locate*, the summary charts also support *lookup* searches to starting points of potential interest. These use the principles of *encode*, *aggregate*, and *navigate* [3]. This addresses the issue of "Where do I start looking?" For example, phenotypes with high frequency can be identified, as well as outliers.

5.5 Exploring Patterns

The Detail Panel can be sorted across phenotypes and across patients to identify patterns in the observations plot. Phenotypes can be sorted alphabetically, by information content, by frequency, and by co-occurrence. Patients can be sorted by ID, dataset-specific patient attributes (e.g. age, severity, diagnostic test scores, treatments), frequency, and by co-occurrence. This supports the *explore* task using the *encode* and *navigate* principles [3].

Frequency is calculated by counting the number of present and absent glyphs in each column or row, for patients and phenotypes, respectively. The sort order is determined first by present count, then by absent count, then by information content. Several methods were explored, but based on researchers' feedback, this was the most intuitive sorting that accounted for both presence and absence.

Co-occurrence sort order is calculated based on the results of hierarchical clustering. To accomplish this, observations are labeled (present: 1, unknown: 0, absent: -1) and a Euclidean distance metric was used. Determining precise clusters would require user input and familiarity with the clustering technique to verify and refine the results. To simplify the user experience, the ordering of phenotypes or patients from the resulting clustering output is extracted. Functionally, this achieves a sort order that groups similar phenotypes or patients next to each other, while mitigating the need for precise refinement of the clustering results (Fig. 5). For the purposes of an exploratory tool, this provides a quick and flexible estimate of potential clusters which can be more robustly evaluated in later analyses.

Phenotype observations are encoded using circular glyphs. Color encodes the type of observation: present as purple, absent as blue, and unknown as light grey. Present-ancestor and absent-ancestor observations are represented as outlined glyphs, using the same color as present or absent to indicate the same class of observation (Fig. 8). Present and absent observations are important to differentiate, since present phenotypes gain diagnostic significance as they become more granular, while absent phenotypes gain diagnostic significance as they become more general.

Phenotype information content and patient attributes are encoded using a monochrome scale, so as not to overpower the visualization with additional colors. Patient attribute values can be categorical or numeric. The logical ordering of categorical values along the scale can be dynamically defined and customized for a given dataset. Ratio values are binned into quantiles prior to display and the number of quantiles can also be dynamically selected. In both cases, attribute values are mapped to perceptually equidistant shades of grey.

To reduce the number of text labels, and to support quicker identification, the layout views adopt the set of HPO category icons developed in PhenoBlocks [11].

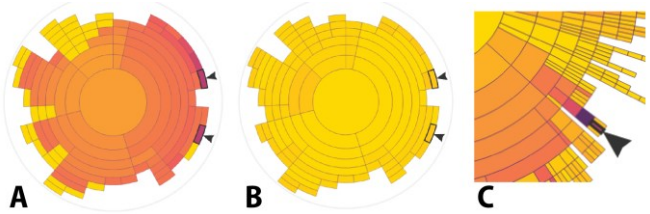

Fig. 4. Each summary chart encodes a specific metric (e.g., frequency, similarity, entropy, outliers), weighted by the diagnostic significance score. (A) Darker colors indicate stronger metrics, e.g., two cohorts share important phenotypes, while (B) lighter colors show weaker scores, e.g., two cohorts lack diagnostically significant differences. (C) Ancestor nodes aggregate the metrics of children, e.g., a darker purple parent indicates this general term appears in all patients.

Fig. 5. Observations can be sorted by phenotypes and/or patients (e.g., co-occurrence, frequency, patient attributes) to reveal patterns.

5.6 Finding Specifics

A global search enables specific phenotypes to be located by name or HPO ID. Search results are persistently listed below the search input to encourage broader search queries to find groups of phenotypes (e.g., searching for "musc" will return a range of muscle/muscular terms). Matches are highlighted in the phenotype list and summary views.

5.7 Ontology Topology Simplification

The layout view enables the selection of different levels of context based on the topology of the HPO. The initial view shows no structure, displaying a flat list of all phenotypes. The user can choose to add grouping to the phenotype terms through a category layout, which uses the first level HPO terms to group phenotypes at the systems-level (e.g., skeletal, muscular, nervous). Alternately, a tree layout supports free-form collapsing/expanding control to explore the entire HPO hierarchy. Adding grouping through the layout view engages *browse* through *encode* and *arrange* principles [3].

Due to multiple inheritance in the HPO, the more the hierarchy is expanded, the greater the complexity and visual redundancy displayed in the observations plot, since by default we duplicate phenotypes that fall under multiple categories (Fig. 6A). To balance the goals of increasing detail while minimizing complexity, a novel algorithm was developed to simplify the topology of the HPO. At a high-level, the algorithm eliminates multiple inheritance by extracting a single path from the root to each leaf phenotype, based on an evaluation of the

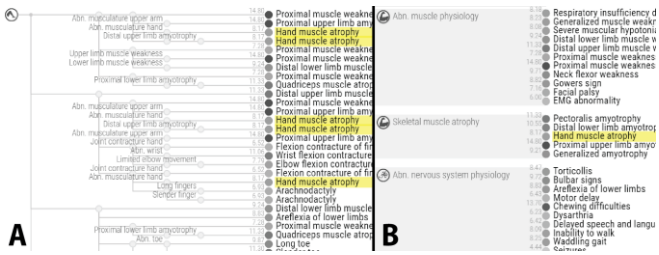

Fig. 6. Example representations of the HPO topology in the layout view. (A) The tree layout can be manually expanded and collapsed, but results in phenotype duplication due to multiple inheritance. (B) The cluster layout categorizes phenotypes using our topology simplification algorithm eliminating duplications. A phenotype is highlighted in yellow.

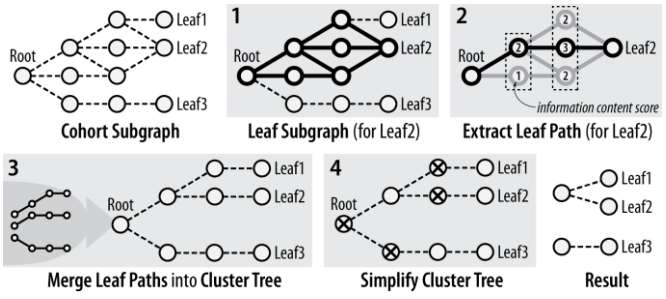

Fig. 7. Illustration of our ontology simplification algorithm.

information content of intermediate phenotypes. Phenotypes that do not contribute to the grouping are then removed (i.e., parents with only a single child). Conceptually, the result is a detailed and diagnostically significant categorization for each phenotype that minimizes complexity (Fig. 6B). The algorithm is as follows (Fig. 7):

1. For each leaf in the cohort subgraph, create a *leaf subgraph* containing all paths back to the root.
2. For each leaf subgraph, use a greedy depth-first approach to extract the path from root to leaf that passes through the child with the *highest* information content at each depth. If more than one path is found, select the path with the highest cumulative information content.
3. Merge all extracted leaf paths into a new *cluster tree.*
4. Traverse the cluster tree depth-first and remove nodes with only one child, stopping when a node has more than one child, and ensuring leaf nodes have at least one parent. Remove the root node.

The fourth step of the algorithm results in a cluster tree that contains all leaf nodes, without duplications, grouped using diagnostically significant intermediate phenotype categories. Since this method is based on the topology of the HPO and information content score, it is independent of the data being displayed. However, depending on the phenotypes exhibited by patients in a cohort, categories may be more or less specific; disparate phenotypes will be categorized more granularly, while related phenotypes will be clustered together under more general categorizations.

This method requires more robust evaluation, and does not currently support user intervention. For example, a parameter could be introduced to control the aggressiveness of intermediate phenotype removal (Step 4), to control the granularity of computed categories. In terms of our goals in this work, the implementation was sufficient to evaluate whether we could compute an intuitive simplified topology using the concept of diagnostic significance.

6 Expert Evaluations

To evaluate PhenoStacks, we deployed the tool in the labs of four experienced genetics researchers. All four participated in the formative interviews, but only one participated in the iterative design process. We asked each researcher to provide a cross-sectional cohort dataset they were studying, or had previously analyzed, as well as patient attributes they were interested in evaluating. Semi-structured interviews were conducted to learn more about each disease and the particular aspects of the cohorts that they were interested in exploring.

The phenotype data was provided in text format and translated to HPO terms using a dictionary mapping approach; the mapping schema was verified by the researchers. We introduced PhenoStacks with their data, explained the visual encoding, and demonstrated the interface to each researcher and their associates. The deployment period lasted 3-7 days, depending on availability. Researchers were instructed to use the tool for at least 1 hour over this period and encouraged to engage research associates in these explorations. Deployment periods ended with semi-structured interviews, asking participants to demonstrate insights garnered using PhenoStacks and to reflect on its utility as part of their workflow. HPO layouts were ranked to evaluate the output of our simplification algorithm, and we elicited general feedback and suggestions for improvements. All participants confirmed they had used the tool independently, as well as with members of their research teams, for a total of 1-2 hours each over the deployment period.

6.1 Usage Scenario Case Studies

First, we summarize descriptions of each disease and how the tool was used by the researchers for their specific tasks.

6.1.1 Inflammatory Bowel Disease

Inflammatory Bowel Disease (IBD) describes a range of genetic disorders that affect the gastrointestinal (GI) system. Patients display widely heterogeneous symptoms due to multi-factorial causes (e.g., genetics, lifestyle, environment) which makes investigating the disease challenging. The IBD researcher was interested in comparing patients with early onset IBD against patients who develop the disease later in life. This task aligns to finding patterns between cohorts (B1), with the goal of differentiating congenital and adult onset forms of the disease. He provided two between-cohort datasets that focused on patients with severe presentations of the disease. The first compared 17 and 6 patients, while the second compared 28 and 19 patients.

We conducted evaluation interviews with two research associates because the researcher was unavailable. While transcribing the data to HPO terms, it became clear that the granularity of terms involved in IBD were not yet sufficiently defined in the HPO. This presents an opportunity to contribute modifications to the HPO to expand their coverage for this disease. However, the potential of using standardized and structured phenotype data was immediately clear. *"I can see a tremendous amount of utility for the tool… it really makes patterns clear,"* one commented. *"Being able to show this is where you have commonality and this is where you start seeing differences. That would be powerful in and of itself; even before you reach the functional and translational implications."* They were also excited about the potential to compare cohort data along other vectors, such as geographic location, to assess potential environmental factors.

This discussion echoed a common pain point for researchers regarding data collection, which often aligns with clinician-patient interactions, but is less suited to research goals. The phenotyping is very broad, less systematic, and detail is not consistent. Investing in standardized data collection methods and developing guidelines to improve consistency would facilitate researcher collaborations, within and between institutions. The difficulty lies in reaching consensus regarding data collection granularity with the clinicians who actually record the data. The researchers felt PhenoStacks would be a compelling artifact to communicate the benefits and to ground discussion.

6.1.2 Myotonic Dystrophy

Myotonic dystrophy is a hereditary multi-systemic disease characterized by progressive degeneration of muscle that weakens the musculoskeletal system, affects cardiovascular function, and results in endocrine changes. The researcher we interviewed wanted to evaluate patterns within a single cohort of 40 patients. In particular, he wanted to identify similarities in the cohort (W1) and look for subcohorts (W3) based on attributes of the patients, such as age and a potential measure of disease severity. He indicated that finding biochemical metrics that consistently correlate with aspects of the disease, such as severity, are critical to selecting appropriate treatments for patients.

The researcher appreciated sorting patients and phenotypes by co-occurrence. *"I can see very nice clusters of [phenotypes] coming together, which really make sense knowing the disease well."* He commented the summary charts complemented the observations plot. By seeing the overall patterns, he was guided to confirm where the numbers were coming from. Sorting the patients by the potential severity metric clearly revealed a progression to more severe patient cases, noted by an increased presentation of severe phenotypes, suggesting the metric may indeed be a valid measure of disease severity (Fig. 8). Moreover, sorting by age could help to investigate a suspected inflection point, where the congenital version of the disease transitions into the more typical, adult degenerative version.

While he did not state it as one of his primary tasks, the researcher commented that outliers were very easy to spot (W2). He identified two examples of severe cases that stood out, due to the severity of the phenotypes, guided by hotspots in the outlier summary chart.

Exposure to the HPO raised questions about the correctness and completeness of his data. While common phenotypes were clear, he had hoped to see more differentiation along severe symptoms. He identified several phenotypes that he felt could be recorded at higher granularity to better capture subtle variations between patients (W4). This suggests that PhenoStacks could be used to audit data collection and help in data collection planning by identifying where deeper phenotyping could better differentiate subcohorts of patients.

6.1.3 Myotubular Myopathy

Myotubular myopathy is a congenital neuromuscular disease marked by abnormal skeletal muscle fiber cells with diminished function, resulting in muscle weakness in patients. Patients present with a variety of phenotypes and researchers are still trying to better understand the genetic causes and differentiate subclasses of the disease. The researcher we interviewed was conducting a study to compare phenotypes of a new patient cohort to those of existing published case reports in the literature (B2). The cohorts contained 12 and 10 patients. He was particularly interested in the similarities and specific differences between the two cohorts.

He commented that overall the tool was very useful for his research goals. Visually comparing the similarities and differences using the observations plot was far easier than the spreadsheet-based approach he was using before. The grouping of phenotypes based on the cluster layout helped to merge cases where different terms were used to describe similar clinical features. *"In the [case study] cohort, we saw a lot of 'Respiratory insufficiency due to muscle weakness' and we thought: no one in our cohort has that, but then you go down to 'Respiratory failure requiring ventilation'—these are basically the same thing—and then the data looks more comparable."*

The researcher also stated that in general, the tool would be especially useful when collecting new data. The new data could be quickly compared to the existing cohort and evaluated to see if the new data was robust enough (i.e., complete, granular, and consistent with other patients). This maps to the data auditing task (W4) we identified in our formative interviews, but did not originate from discussions with this researcher.

6.1.4 Phenylketonuria

Phenylketonuria (PKU) is a congenital disease resulting in impaired metabolism of the amino acid phenylalanine. Untreated, the disease affects the neurological system, leading to intellectual disabilities and other severe medical problems. The researcher we interviewed is also a clinician and was interested in the variation of phenotypes within a cohort of his clinical patients to identify cases where the treatment regimen has not been effective. Thus, this use case was more oriented towards clinical diagnostics. He provided us with phenotype data for 12 random patients. Several attributes were provided for each patient, including metrics of disease management and severity.

During the interview, he first demonstrated how the observations plot revealed patterns he expected to see. For example, when sorting by increasing severity: *"It's nice to see it work, with neurological features appearing [in more severe cases]."* He commented that the "frequently present" summary chart was very useful for diagnostic purposes, since darker colors immediately indicated something he should investigate. He also commented that the visualization could be used to identify groups of patients within the cohort (W3), for example patients who are not managing their disease effectively.

He also used the tool to explain hypotheses about why the treatment regimen was not meeting the needs of certain patients. For example, if all patients had the disease under control, the measure of management should be roughly equivalent, but sorting by severity indicated that the management score was worse for certain severe cases. These are cases where the current treatment is not working and the treatment should be reassessed. *"[PhenoStacks] provides me with a nice visualization for management... to see where we are in meeting our expectations. I know what I should be seeing, and when I'm not seeing it, it's informative."*

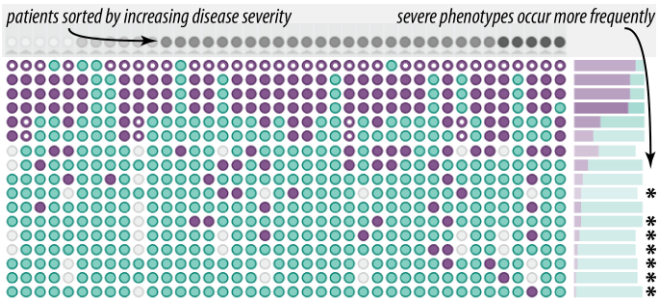

Fig. 8. An example observations plot from the myotonic dystrophy dataset. Sorting patients by severity reveals a pattern of increasing present observations (purple) of more severe phenotypes (lower in the chart) as we move from less to more severe patient cases (to the right).

6.2 General Feedback

Having described detailed usage scenarios and comments from the researchers, we now address higher-level shared feedback.

6.2.1 HPO Layout Methods

The cluster layout method (using our simplification algorithm), was the preferred layout (2/3 researchers and both research associates), followed by the flat view. These participants were all focused on research-oriented use cases, where they were either interested in seeing global co-occurrences, or investigating particular sub-systems, which was easier to do with the clusters layout; the category layout was too broad, and the tree layout had too many duplications. One researcher commented that for his data, the clusters were sometimes too specific, and he would like some control to "relax" the output so that more general categories would be selected. While category granularity preference differed by researcher, these results indicate that computational approaches to simplifying the HPO topology is a promising direction to aid contextualization of phenotype data.

The PKU researcher preferred the tree layout, and then the cluster view. He explained that for his purposes (i.e., clinical diagnostics), he was more interested in specific phenotypes, and less about how they relate to each other. In fact, while the structure of the HPO made sense to him, it was contrary to his mental model of diagnosis, which started with detailed presentations, not categories. His preferences thus aligned with the layouts that showed the most detailed view. This suggests that representing phenotypes in the context of the HPO may be less critical when designing visualizations for clinical diagnostics.

Overall, we received positive feedback on the algorithm to cluster phenotypes to eliminate duplications. The IBD research associates especially liked that it not only removed duplicates, but also presented phenotypes in categories that made the most sense. *"Cholangitis is a co-presenting liver abnormality, while it is also related to immune system, this isn't how we think about it. The cluster view groups the two liver phenotypes together, separating it from the GI phenotypes. This makes sense and it's helpful to see that separation."* (Fig. 9)

6.2.2 Data Quality Concerns

We worked with the researchers to map the phenotype terminology and clinical results of their data to the HPO. For the most part this was straight-forward, but as we described for IBD, sometimes HPO terms were missing. Since the HPO is an ongoing project, engaging researchers across a wider variety of diseases is an opportunity to benefit from their expertise to help improve the HPO.

During the process, the benefits of the structured phenotype data became clear to all participants. Ambiguities in the description of phenotypes in the researchers' datasets were uncovered, making it unclear what HPO term was appropriate, echoing issues reported in the literature. In these cases, a more general HPO term was used, but for the researchers this underscored the benefits of integrating the HPO into the data collection process at the beginning. By visualizing phenotypes using the HPO, deeper phenotyping benefits also became clear, because the researchers could see areas where they were not

seeing the differentiation they expected because the phenotypes in their data were too general. The researchers also noted that if their collaborators were also using the HPO, it would make it easier to merge datasets, even if they are recorded at different granularities, because the topology of the HPO accommodates these differences. Thus, these observations suggest that visualizations can help communicate the benefits of structured phenotype data in supporting consistency, completeness, and granularity.

6.2.3 Sharing and Collaboration

All researchers commented that the visualization would be good for communicating with collaborators and presenting and sharing results. The different layout views were seen as useful for projects involving collaborators with different levels of expertise, for example geneticists who are less familiar with specific phenotypes and more focused on the systems-level. Seeing the categories would help bridge knowledge gaps with unfamiliar terms by showing how they relate to each other.

The observations plot was also seen as a valuable aid when discussing and communicating insights. Being able to see the patterns and show them to other researchers was seen as far better than simply sharing the information using text descriptions. For example, the categorizations using the HPO topology clearly delineate related phenotypes and also indicate specific subsystems that may play a role in the disease, helping to differentiate localized and systemic symptoms. To compare results, researchers could ask collaborators to describe the pattern of symptoms they are seeing within a certain category, rather than simply sharing specific phenotypes. This can also support collaboration between experts with different specializations.

6.2.4 Suggested Improvements

The researchers were eager to share ideas to improve the tool. The PKU and myotonic dystrophy researchers, who considered the task of subcohort discovery (W3), suggested additional features to organize patients in the observations plot. For example, one suggested visually separating patients into distinct subgroups within each cohort. Both researchers wanted to collapse subgroups into a single column to hide patients that were less important to the analysis at hand. They also both wanted to add *ad hoc* attributes to the patients dynamically during analysis, to impose custom sort orders on the patients. This desire for greater interactive control over patient organization suggests that participants were embracing interactive visual analysis as part of their workflows, and could envision more advanced features.

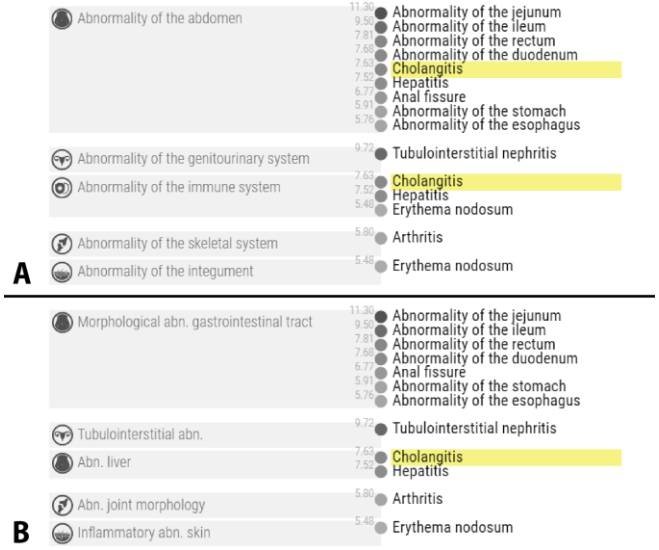

Fig. 9. Comparison between category and cluster layouts. Cholangitis, an inflammation of the liver, (A) appears under both Immune System and Abdomen in the category view, where it is grouped with mostly gastrointestinal abnormalities. (B) The cluster layout selects a more meaningful organization that separates Gastrointestinal abnormalities and Liver abnormalities using our topology simplification algorithm.

The IBD research associates suggested labeling certain phenotypes "very important", so that the frequency scores of these would be highlighted, regardless of their diagnostic significance, *"If you know you're interested in certain phenotypes, it would be nice to see them immediately."* While the information content does map to diagnostic significance, in some cases they were interested in the presentation of phenotypes that were more common, but very significant when they co-present with other phenotypes in IBDs.

In a similar vein, the myotonic dystrophy researcher wanted to add attributes to the phenotypes, similar to the patient attributes, that could be used to sort the phenotypes list differently. In particular, a measure of frequency within the disease or a correlation with severity was desired, as this would impose unique clustering when paired with co-occurrence in patients.

Finally, all the researchers and research associates wanted a longitudinal version of this tool, since there were aspects of their patients that they wanted to visually track over time.

6.3 Study Limitations

This study is not without its limitations. The deployment period was short and variability of phenotype coverage in the HPO impacted the depth of some data explorations. However, we engaged researchers across a variety of specializations and garnered insightful feedback. The researchers were exposed to structured phenotype data and gained increased sensitivity to the quality of their own data.

A long-term deployment study of PhenoStacks is being planned. Working with a researcher and his team, we are submitting updates to the HPO so that the phenotypes are well represented. PhenoStacks will be deployed among a large group of researchers participating in a data collection workshop to support data auditing and analyses thereafter.

7 Conclusion and Future Work

This work introduced PhenoStacks, a visual analysis tool to support cross-sectional cohort phenotype analyses. A novel algorithm was developed to simplify the visual display of the HPO topology and eliminate duplicated terms. Based on feedback from expert genetics researchers, this algorithm and the layouts that were produced were preferred over category and tree layouts. Deployment study results indicated many potential applications within the researchers' existing workflows. These visualization concepts can be applied in other domains where instances of taxonomies or ontologies are analyzed.

The deployment study identified several improvements that we plan to address in subsequent open-source releases. Agglomerating patient columns would benefit analyses, but also address scalability to larger cohorts. Methods of collapsing the HPO topology (e.g., fish-eye distortion) could further mitigate clutter when cohorts contain a wide variety of phenotypes. Both approaches could integrate user-authored tags (e.g., important phenotypes) and *ad hoc* attribute definition. Given a richer range of attributes, alternate visualizations of attributes are worth investigating (e.g., scatterplots to evaluate correlations).

Based on feedback, we plan to extend PhenoStacks to longitudinal cohort analyses. While clinical studies consider sequences of temporal events (e.g., hospital visits, treatments), genetics researchers focus on how phenotypes of a patient change over time. These *natural history studies* track disease progression in cohorts, e.g., investigating the effect of environmental exposures or lifestyle on the manifestation of phenotypes. We see promise in applying approaches, such as temporal metrics (as in CoCo [25]), to longitudinal phenotype analyses.

Acknowledgments

This work was partially funded by Genome Canada and Ontario Genomics through a Bioinformatics/Computational Biology grant to Dr. Brudno and by a CPER Nord-Pas de Calais / FEDER DATA Advanced data science and technologies grant to Inria Lille Nord-Europe. A special thanks to Michelle Annett, Hali Larsen, Peter Hamilton, John Hancock, the researchers who participated in our studies, and the anonymous reviewers for their thoughtful feedback.

References

[1] Baynam, G., Walters, M., Claes, P., ... & Goldblatt, J. (2015). Phenotyping: Targeting genotype's rich cousin for diagnosis. *J Paediatr Child Health*, 51(4), 381-386.

[2] Bostock, M., Ogievetsky, V., & Heer, J. (2011). D^3 data-driven documents. *IEEE TVCG*, 17(12), 2301-2309.

[3] Brehmer, M., & Munzner, T. (2013). A multi-level typology of abstract visualization tasks. *IEEE TVCG*, 19(12), 2376-2385.

[4] Buske, O. J., Girdea, M., Dumitriu, S., ... & Links, A. E. (2015). PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. *Hum Mutat*, 36(10), 931-940.

[5] Carpendale, S., Chen, M., Evanko, D., ... & Strobelt, H. (2014). Ontologies in biological data visualization. *IEEE Comp. Graph. Appl. Magazine*, 34(2), 8-15.

[6] Cerami, E., Gao, J., Dogrusoz, U., ... & Antipin, Y. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov*, 2(5), 401-404.

[7] Fiume, M., Smith, E. J., Brook, A., … & Brudno, M. (2012). Savant Genome Browser 2: visualization and analysis for population-scale genomics. *Nucleic Acids Res*, 40(1), 615-621.

[8] Fu, B., Noy, N. F., & Storey, M. A. (2015). Eye tracking the user experience–An evaluation of ontology visualization techniques. *Semantic Web Journal*, 1-19.

[9] Girdea, M., Dumitriu, S., Fiume, M., ... & So, J. (2013). PhenoTips: patient phenotyping software for clinical and research use. *Hum Mutat*, 34(8), 1057-1065.

[10] Global Genes. Accessed October 21, 2015. https://globalgenes.org/rare-diseases-facts-statistics

[11] Glueck, M., Hamilton, P., Chevalier, F., Breslav, S., Khan, A., Wigdor, D., & Brudno, M. (2016). PhenoBlocks: Phenotype Comparison Visualizations. *IEEE TVCG*, 22(1), 101-110.

[12] Groza, T., Hunter, J., & Zankl, A. (2013). Mining skeletal phenotype descriptions from scientific literature. *PloS one*, 8(2), e55656.

[13] Hennekam, R., & Biesecker, L. G. (2012). Next-generation sequencing demands next-generation phenotyping. *Hum Mutat*, 33(5), 884-886.

[14] Houle, D., Govindaraju, D. R., & Omholt, S. (2010). Phenomics: the next challenge. *Nat Rev Genet*, 11(12), 855-866.

[15] Hu, Z., Hung, J. H., Wang, Y., Chang, Y. C., Huang, C. L., Huyck, M., & DeLisi, C. (2009). VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res*, gkp406.

[16] Katifori, A., Halatsis, C., Lepouras, G., Vassilakis, C., & Giannopoulou, E. (2007). Ontology visualization methods—a survey. *CSUR*, 39(4), 10.

[17] Keim, D., Andrienko, G., Fekete, J. D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual analytics: Definition, process, and challenges, *Information Visualization*, 154-175.

[18] Köhler, S., Doelken, S. C., Mungall, C. J., ... & Robinson, P. N. (2013). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Res*, gkt1026.

[19] Köhler, S., Schulz, M. H., Krawitz, P., ... & Robinson, P. N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet*, 85(4), 457-464.

[20] Krause, J., Perer, A., & Stavropoulos, H. (2016). Supporting iterative cohort construction with visual temporal queries. *IEEE TVCG*, 22(1), 91-100.

[21] Krzywinski, M., Schein, J., Birol, I., ... & Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res*, 19(9), 1639-1645.

[22] Lanzenberger M, Sampson J, Rester M. (2010) Ontology Visualization: Tools and Techniques for Visual Representation of Semi-Structured Meta-Data. *J Univer Comput Sci*, 16(7):1036–1054.

[23] Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014). UpSet: visualization of intersecting sets. *IEEE TVCG*, 20(12), 1983-1992.

[24] Lussier, Y. A., & Liu, Y. (2007). Computational approaches to phenotyping: high-throughput phenomics. *J Am Thorac Soc*, 4(1), 18-25.

[25] Malik, S., Du, F., Monroe, M., … & Shneiderman, B. (2015). Cohort comparison of event sequences with balanced integration of visual analytics and statistics. *ACM SIGCHI*. 38-49.

[26] Meyer, M., Munzner, T., & Pfister, H. (2009). MizBee: a multiscale synteny browser. *IEEE TVCG*, 15(6), 897-904.

[27] Nielsen, C. B., Jackman, S. D., Birol, I., & Jones, S. J. (2009). ABySS-Explorer: visualizing genome sequence assemblies. *IEEE TVCG*, 15(6), 881-888.

[28] Pendergrass, S. A., Dudek, S. M., Crawford, D. C., & Ritchie, M. D. (2012). Visually integrating and exploring high throughput pwas results using PheWAS-View. *BioData Min*, 5(5).

[29] Perin, C., Dragicevic, P., & Fekete, J. D. (2014). Revisiting Bertin matrices: New interactions for crafting tabular visualizations. *IEEE TVCG*, 20(12), 2082-2091.

[30] Philippakis, A. A., Azzariti, D. R., Beltran, S., ... & Dumitriu, S. (2015). The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum Mutat*, 36(10), 915-921.

[31] Rind, A., Wang, T. D., Aigner, W., … & Shneiderman, B. (2011). Interactive information visualization to explore and query electronic health records. *Found Trends Hum-Comput Interact*, 5(3), 207-298.

[32] Robinson, Peter N. (2012). Deep phenotyping for precision medicine. *Hum Mutat*, 33(5), 777-780.

[33] Robinson, P. N., Köhler, S., Bauer, S., … & Mundlos, S. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet*, 83(5), 610-615.

[34] Robinson, P. N., Mungall, C. J., & Haendel, M. (2015). Capturing phenotypes for precision medicine. *Mol Case Stud*, 1(1), a000372.

[35] Schroeder, M. P., Gonzalez-Perez, A., & Lopez-Bigas, N. (2013). Visualizing multidimensional cancer genomics data. *Genome Med*, 5(1), 9.

[36] Secrier, M., Pavlopoulos, G. A., Aerts, J. & Schneider, R. (2012). Arena3D: visualizing time-driven phenotypic differences in biological systems. *Bioinformatics,* 13(45).

[37] Shachak, A., & Fine, S. (2008). The effect of training on biologists acceptance of bioinformatics tools: A field experiment. *J Am Soc Info Sci Tech*, 59(5), 719-730.

[38] Shannon, P., Markiel, A., Ozier, O., … & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13, 2498–2504.

[39] Rare Disease Impact Report: Insights from patients and the medical community. (2013). Shire HTG Technical Report. Accessed October 21, 2015. http://www.geneticalliance.org.uk/docs/e-update/rare-disease-impact-report.pdf

[40] Shyr, C., Kushniruk, A., van Karnebeek, C. D., & Wasserman, W. W. (2015). Dynamic software design for clinical exome and genome analyses: insights from bioinformaticians, clinical geneticists, and genetic counselors. *J Am Med Inform Assoc*, ocv053.

[41] Sollie, A., Sijmons, R. H., Lindhout, D., ... & Wijburg, R. (2013). A New Coding System for Metabolic Disorders Demonstrates Gaps in the International Disease Classifications ICD-10 and SNOMED-CT, Which Can Be Barriers to Genotype–Phenotype Data Sharing. *Hum Mutat*, 34(7), 967-973.

[42] Vaske, C. J., Benz, S. C., Sanborn, J. Z., ... & Stuart, J. M. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12), i237.

[43] Vercruysse, S., Venkatesan, A., & Kuiper, M. (2012). OLSVis: an animated, interactive visual browser for bio-ontologies. *Bioinformatics*, 13(1), 116.

[44] Winnenburg, R., & Bodenreider, O. (2014). Coverage of phenotypes in standard terminologies. *Joint Bio-Ont and BioLINK ISMB*, 41-44.

[45] Zhang, Z., Gotz, D., & Perer, A. (2014). Iterative cohort analysis and exploration. *Information Visualization*, 1473871614526077.