The Gabriella Miller Kids First X01 Data Analysis Collaboration Workshop Executive Summary

5635 Fishers Lane
Rockville, MD

June 16, 2017

Summary of Comments and Recommendations

Introduction

NIH program staff have been working on the Kids First program for about 2.5 years and the program is making great progress. Recently, the first whole genome dataset was released in dbGaP and by the end of FY17, the third year of our program, the Gabriella Miller Kids First Pediatric Data Resource will have been established. Ellyn Miller, Gabriella Miller's mother and founder of the Smashing Walnuts Foundation, continues to lobby for Congressional support of this 10-year effort since it is dependent upon annual appropriations.

The workshop focused on considerations for Kids First investigators as they seek to collaborate with each other and analyze data, and how the Data Resource and NIH program staff can support these endeavors. Key discussion points are summarized below, and these will be taken into consideration as the Kids First program prepares for the First Annual Meeting to be held September 6-7 and makes general programmatic decisions for future years (e.g. FOAs that would support research endeavors).

Updates from the Kids First Sequencing Centers

The Broad Institute of MIT and Harvard

- The Broad can support analysis for any X01 collaborator using the Broad infrastructure including processing and variant-calling of data through Broad best-practices pipeline (Picard alignment, GATK variant calling), integration with gnomAD dataset through the seqr platform, and rare disease analysis support through Center for Mendelian Genomics.
- Many Kids First datasets will be jointly called with the next release of the Genome Aggregation Database (gnomAD). GnomAD is a resource developed by the Broad that contains thousands of exomes and genomes from a variety of sequencing projects. The browsing tool seeks to make summary data available for the wider scientific community.
- The Broad has developed a common pipeline for *upstream* processing (raw data to BAM), which is important for generating a common input a to variant calling. All Kids First data produced at the Broad undergo that process, but they are willing to help reprocess other datasets so they can be eventually combined for variant calling.

HudsonAlpha Institute for Biotechnology – St. Jude Children's Research Hospital

- Tumor RNAseq and exome sequencing are being done at St. Jude while all WGS is performed at HudsonAlpha.

NIH Genomic Data Sharing Policy for Kids First Projects

An understanding of NIH Genomic Data Sharing (GDS) policy is extremely important when it comes to collaborating for cross cohort analyses. The language of the individual consent forms signed by study participants provides the legal foundation for how controlled-access data from enrolled participants can be shared. The institutional certification documents how the collecting PI's Institutional Review Board (IRB) interprets the language of the consent form, including any data use limitations (DULs) or data use limitation modifiers. The institutional certification is submitted to a Genomic Program Administrator (GPA) who uses this to register the study in dbGaP and generate a Data Use Certification (DUC). DULs and DUL modifiers are to be selected only if they reflect the language of the consent form, not PI or even IRB preference. "Preferences" can be written into other sections of the DUC; for example, an IRB may prefer the dataset not be used for commercial endeavors, or the PI might prefer to establish a collaboration with anyone accessing that dataset.

Secondary users (or "end-users") must request access to a dataset by submitting a Data Access Request (DAR) through dbGaP which will be reviewed by the relevant Data Access Committee (DAC). Secondary users and their supporting Institution's Signing Official must agree to the conditions of the DUC, including GDS policy as well as any DULs or DUL modifiers pertinent to the requested dataset.

Since the Data Resource is tasked with making genomic data from various pediatric cohorts broadly accessible to the research community, it is important that the Kids First program works with datasets that can be broadly shared and used. Therefore, going forward:

- Projects with broad data sharing/use ("general research use" with minimal modifiers) will be prioritized for sequencing.
- If any DULs or DUL modifiers were erroneously included, a revised institutional certification, preferably using the current NIH template, may be submitted to your GPA (after IRB approval) and the dbGaP registration will be corrected.
- In line with GDS policy, new consent forms should use language that reflects broad sharing. For guidance on developing consent language, visit https://osp.od.nih.gov/wp-content/uploads/NIH_Guidance_on_Elements_of_Consent_under_the_GDS_Policy_07-13-2015.pdf.

The Kids First team, and NIH GPAs are available to help investigators navigate the GDS policy, and to facilitate collaborative research among the pediatric research community. The Data Resource, and its coordinating center and portal, will use the dbGaP-DAC process for access to individual datasets.

Recommendations for the Data Resource

- Recall all variants in a common pipeline, or multiple pipelines to encompass "clinical" (identifying *the* causal variant) and "research" (unbiased calls at every locus) styles.
- Facilitate re-classification of pathogenetic variants in ClinVar as more data become available.
- Facilitate discussion and data comparisons among Kids First investigators as well as external datasets including those from genetic testing companies.
- Maintain access to raw read files (BAMs, FASTQs) because:
    - There are strengths and weakness of each pipeline
    - Access to raw data files maximizes the performance of many of tools and different methods of variant filtering
    - Rare variants may not be called or can get lost, particularly for INDELs (alignment and

mapping of INDEL boundaries are not consistent)

- Facilitate sharing analytic pipelines and protocols that each individual group is using.
- Integrate pipelines to facilitate analyses that will be most common across investigators but provide flexibility to tailor how the pipeline is used (e.g. family-based approached may use less stringent filtering than tumor-normal comparisons).
    - An integrated pipeline is particularly important for studying across childhood cancers and structural birth defects
    - Harmonizing and calling through a combined pipeline is important but it doesn't solve all problems
- Identify end-users and determine what they want to do with the data. We anticipate a spectrum of end users from "hardcore bioinformaticians" to basic scientists/ developmental biologists.
- Make aggregate and other types of data "open access" wherever possible (in line with NIH policy) to facilitate searching and identification of relevant cohorts.
    - Aggregate data that does not identify individuals, e.g., gene lists and summary statistics, can be publicly available. Other types of aggregate data (germline variants that could identify someone) need to be "controlled-access".
    - Where appropriate (e.g., for allele frequencies), the Data Resource should consider separating aggregate data between parents, any unaffected family members, or "controls" from diagnosed cases.
- Organize and coordinate a "user committee" that includes current Kids First investigators as well as external members of the pediatric research community.
    - Decisions on detailed questions like how aggregate data is organized (previous point) should be made by such a committee.
- Address how to harmonize phenotypic and clinical data (see section below).
- *See "Control Cohorts for Pediatric Analyses" section*
- *See "Phenotype Harmonization" section*

Considerations for X01 Data Analyses and Collaborations

- The NIH Office of General Council recently verified that they favor Trusted Partnerships only be established under contracts. Since the language of the 2014 Gabriella Miller Kids First Research Act specifies that this program can only use grant mechanisms, the Kids First Data Resource cannot be granted a Trusted Partnership.
- Some are concerned that the phenotyping in gnomAD is not comprehensive or consistent. 30% gnomAD genomes have some phenotype data available, a small portion are re-contactable but it is complicated to do. As published in a recent paper, the Broad recommends setting frequency thresholds calibrated for disease prevalence and tweaked based on prevalence and other factors to model the maximum count of gnomAD participants that would be consistent with those parameters.
- dbGaP was set up to control access to stored individual datasets, and it serves that purpose very well. When it comes to collaborative analyses and what you can do after you have access, individual project DULs can continue to hinder analyses even as we look forward to an NIH Data Commons environment. Legally, we must respect the DULs. The research community can help this process by making data more broadly available at the time of consent.
- Investigators can influence consent language for new projects. Broad consent language allows for flexibility in collaborations. This would require a mental shift because, while you want to protect the patient as much as possible, you also want to maximize the use of their data as

much as possible.

- Kids First investigators could collaborate to form a "synthetic consortium", so they would not have to apply for each other's datasets in dbGaP. With local IRB approvals, data could be shared among those who agree to this collaboration to compare datasets, but the group could not distribute that data to external users. The Data Coordinating Center could facilitate this data transfer among members of this "consortium".
- Some groups are starting to self-assemble into collaborations and applying for R01s. Please keep NIH staff aware of these applications when they are submitted.
- There should be continued discussion about what data should be collected in individual follow up and replication studies that will have value for the program.

Control Cohorts for Pediatric Analyses

Kids First is not able to whole genome sequence a large number of controls; however, we could broker a conversation with other NIH programs, for example ECHO and the *All of Us* program (which has delayed its plans to sequence children). Suitable control datasets may exist; however, investigators need to consider several factors for filtering, prioritizing, and validating variants associated with pediatric cancer or birth defects cohorts when using these.

Confounding Factors and How to Address Them

- **Genomic Technology:** Type of sequencing, chemistry, library preparation, coverage differences, variant calling pipelines, tools etc. Several attendees felt this is the most important confounding factor.
    - **Access to raw sequence files** (BAMS) is crucial for comparing controls with "disease". Particularly for Copy Number Variant and Structural Variant analysis, you need to reanalyze data together through the same pipeline. Aggregate data and variant calls from controls are not enough.
- **Age**: Although this is an important factor when studying rare pediatric conditions and age-related phenotypes; several attendees felt this was lowest priority and it would not be worth investing in a *pediatric* control cohort.
    - **Somatic mosaicism.** Mutations accumulate across age, which makes it difficult to identify rare variants for pediatric conditions, especially in participants >40 years old. Some argued that these are relatively easy to detect and filter out, but the concern would be that you are potentially filtering out risk variants associated with these rare pediatric conditions.
    - **Survival bias**. This factor could be estimated based on US mortality rates and factored into power calculations.
    - Age matching is important for answering questions related to telomere length and mitochondrial mutations.
    - There are certain study designs in existing datasets that could address these issues.
        - Disease prevalence in a population might be more important than matching cases to controls.
        - A broad exposure (e.g. radiation) base may be more valuable than age-matching.
    - You would need thousands of fully sequenced individuals to determine the population allele frequencies of rare alleles even if they are age matched.
- **Phenotyping**: If possible controls should be screened by qualified professionals and phenotype

data should be collected in the same way data is collected for the phenotype of interest
- **Ethnicity Matching & Population Structure**. Different populations have different allele frequencies
- **DNA extraction and quality**
- **Exclude cases from population datasets**. Ensure case is not part of the controls (e.g. remove TCGA from ExAC cohort when studying cancer)
- **Overlapping variants**. When variants overlap for two different phenotypes (e.g. associated with risk of autism and cancer), be careful not to filter out causal mutations.
- **Consents**. Control datasets should have the broadest level of sharing to be used on a diversity of projects. Some datasets are restrictive to use on cancer or birth defects projects.

Small R03s or R21s could be good mechanisms for doing a side study to test whether age-matching yields interesting findings.

Existing Datasets

- 1000 Genomes: Open access data but no phenotype data available (including age and family structure for relatives). Coverage is very low which is not good for rare variant analysis.
- Austim MSSNG: Not consented for cancer research, must be used only for neurological studies.
- NDAR Autism controls: Capture problems with exomes coupled with sequencing errors. Be careful to not filter out causative genes that may overlaps between autism and cancer, for example.
- ADSP Alzheimer controls: Caucasian only, much older population (potentially confounded by somatic mosaicism)
- ExAC: Contains many cohorts/populations with a variety of conditions (not "healthy"). Remove specific cohorts related to your phenotype of interest; for example, when studying cancer remove TCGA cohort which contains breast cancer patients.
- GnomAD:
    - Example of a large datasets with multiple subpopulations (from multiple studies) so you have power to look at frequencies of rare variants
    - Individual genotype and phenotype data are not available for most individuals
    - Limited number of controls: 500 now, 2000 soon
    - Additional controls are needed for comparing de novo mutations
    - May be most useful for dominance model, but not for a recessive model
    - One challenge is that many genomes currently sequenced do not have the right consents to go into a common control dataset (e.g. Austim MSSNG). The next release of GnomAD may have up to ~20,000 European whole genomes, and ~5,000 African American whole genomes that will have either *General Research Use* or *Health/Medical/Biomedical* consent groups that could be used as individual level controls. These controls as well as some Kids First projects will go through the GATK pipeline as part of GnomAD version 3.
        - We have to strategize the best way to facilitate comparison assuming people won't want to download 25,000 genomes and run that analysis themselves.
- Internal to St. Jude:
    - Genomes 4 Kids: Clinical genomics program (WGS, WES and RNA sequencing). Developed controls with African American subjects with sickle cell disease – used to filter variants with African American cohorts.
    - St. Jude Life study: Follow every childhood cancer survivor (free of cancer for 10 years). Control group phenotyped the same as the cases.

- Other examples of potential control datasets: 23andMe, Haplotype Reference Consortium, OncoArray Consortium, dbGaP
- All of Us (formerly PMI): plans to eventually sequence children, but this has been delayed.
- CCDG (NHGRI): currently early onset heart disease, hemorrhagic stroke, autism.
- TOPMed (NHLBI): Will be 80-100,000 genomes very soon, mostly from epidemiological cohorts where they are not selected for anything. Genomes processed in a very consistent way.
- The unaffected parents and siblings of Kids First cohorts (must comply with data use limitations of the consents). Potential for confounding low-penetrance variants.

Phenotype Harmonization

With the funding of the Kids First Data Resource, the Kids First program will begin the process of phenotype harmonization for cohorts sequenced through the program. This is a very critical part of the program as consistent data will allow:

- Better prioritization of genes/variants in individual cases
- Identification of overlapping genes/gene networks across cohorts within pediatrics
- Exploration of the genetic connection between cancer and disorders of development
- Better descriptions of rare disorders
- Assessment of penetrance and identification of modifiers
- Automated, on-scale assessments to accommodate large datasets

The field has converged around Human Phenotype Ontology (HPO) terms. HPO started with incorporation of disease-associated OMIM genes and their phenotypes and constructed a structured vocabulary. More recently, HPO was updated with data from OrphaNet and DECIPHER. In addition, many other consortia and projects are using HPO with this list just being a subset:

- GWAS Central
- Phenome Central
- Extasy
- NGI
- ClinVar
- Face2Gene
- Gen2Phen
- PhenomeNet
- WebGestalt
- PhenoVar

HPO provides a starting point to choose a subset (15-25 term?) of common terms from among the ~12,000 terms. HPO is continuously evolving with new releases issued, but it is not changing very radically at this point. Tools include Phenolyzer where one can simply enter interested phenotypic terms and submit them. PhenoTips can facilitate gathering phenotype information in HPO terms. HPO applies a parent-child relation between super- and sub-classes and follows are tree model starting with five main classes:

- Phenotypic abnormality
- Mode of inheritance
- Clinical modifier
- Mortality/aging
- Frequency

Phenotypic abnormality provides the starting point. Some cohorts have overlap but others don't. Each cohort will need to decide how far down the tree to progress.
Some challenges:

- Only a small fraction (~3000) of human protein coding genes have HPO annotations.

- Differences in releases/versions among tools and organization may cause discrepancies.
- Adoption by physicians might be perceived as time-consuming; there might be difficulties in converting old records by different physicians (interpretation of older terms, e.g., "mental disorder").

A possible strawman: phenotypes to document across studies:

- Structural birth defects (including hypospadias, absent uterus)
- Hearing loss
- Neuropathies
- Neurodevelopment disorders (intellectual disability, autism, epilepsy)
- Problems with growth
- General dysmorphic features
- Cancer
- Isolated or complex
- Family history

Alternatives presented but not considered better include:

- ICD9 terms used for insurance coding – notoriously inaccurate and incomplete
- SNOMED CT used for clinical terminology to give clinical content; the US standard for electronic health information exchange (commercial product) – community has not converged on this.
- MESH (Medical Subject Headings) used to provide a hierarchically-organized terminology for indexing and cataloging of biomedical information (MEDLINE/Pubmed and other NLM databases) – least of the adopted systems.

Being able to do gene ontology across organisms: Open Biomedical Ontologies (OBO) Foundry

- Developing a family of interoperable ontologies that are both logically well-informed and scientifically accurate.

Next steps:

- Agree on classification system
- Decide which subset of phenotypes – focus on pediatrics/ level of granularity
    - Start with 25-50 of the most important elements; second pass can be more detailed
- When to complete categorization
- Where to share data
- Age at last evaluation
- Ability to recontact – make a list of which cohort cans be recontacted
- Vital status

Highlights from general discussion:

- There was consensus endorsing the use of HPO terms.
- There was agreement that the level of granularity needs discussion; it's a big issue.
    - provide all so an end user can access to level they want
    - define minimum dataset and note ability to re-contact for more details
    - define what was asked, e.g., for family history of cancer, document what was asked. An absence of a positive response is not a negative. Clearly capture was assessed and what was not assessed.

- Easiest way forward is taking information already in project databases, standardize it, and provide whatever detail is there, indicating "not evaluated" if not known. No one can afford a research assistants time to recontact (if possible) to fill in gaps.
- Standardization of dictionaries needed.
    - The Genomic Data Commons has developed Common Data Elements for cancer: https://docs.gdc.cancer.gov/Data_Dictionary/viewer/
- Need to achieve balance: healthy tension exists between end users who want everything and providers want to do minimal amount of work. Providers have the best sense of what is needed to make the best use of their data.
- PhenX has created measures and themed "domains" that capture core data elements to create public standards through a long consensus building process. This should be examined for relevant pediatric measures and domains.
- Action item – Get consensus among Kids First investigators on what core data elements to collect and input to the data resource for all cohorts.

Recommendations for the NIH and the Kids First Program

- Data storage comes with costs. While the Kids First Sequencing Centers can hold large datasets (BAMs and FASTQs) in "cold" storage at lower costs, this is at best a temporary solution as the data are not readily accessible nor sharable since the Sequencing Centers are not Trusted Partners. Until a long-term NIH solution is implemented, a short-term solution involving either dbGaP or a trusted partner is needed to make these data sharable with the research community.
- Communicate the value of sharing aggregate data to NIH policy makers, as a formal policy develops or is clarified. Releasing aggregate data can facilitate searching for a particular variant in a given cohort and determining its frequency.
- Prospectively create a centralized IRB and unified consent language that might facilitate some of the data sharing moving forward.
- Opportunities for funding follow up studies and deep phenotyping (requires re-contacting participants).
- A lot of the issues that have come up highlight that even basic cross-study collaborations and comparisons are going to be significant endeavors. New funding mechanisms to specifically jumpstart those could be useful.
- There are some discrepancies in how different DACs or even GPAs handle data sharing. Kids First works with many NIH institutes and centers and will try to communicate with GPAs and DACs as much as possible to implement a common approach to data sharing.