Idea Nomination: Common Fund Kids First Research Act

Title of proposed program: Gabriela Miller Kids First Data Commons

What is the major obstacle/challenge/opportunity that the Common Fund should address?
There are many unanswered questions in pediatric research that require well-curated medical, genomic, and physiological data from large numbers of children. What are the origins of diseases? Which diseases and conditions are associated with each other, and why? What treatments work best for a given condition?

The existence of many pediatric cohorts across the NIH creates an opportunity for integrated data sets that would enable new studies that require data from larger sets of patients. This would expand the utility of individual data sets, allowing entirely new types of questions to be addressed. The challenges are computational and logistical. Data integration of several types of data would be needed, and existing cohorts would need to adapt to acquire data that meet necessary data standards for the Data Commons. New/additional data would need to be acquired for participating cohorts.

What would the goals of the program be?
1. Enhance IC-funded cohort studies by supporting acquisition of additional data, with the expectation that certain core data types, including genotyping or sequencing, would be collected for all participating cohorts. Cohorts would apply to be part of the Data Commons and would be selected based on the value of data that have been gathered, value of data to be acquired, and vision for compelling use of integrated datasets.
2. Develop computational infrastructure, tools, and methods that would support data sharing, interoperability, mining, and collaboration across sites
3. Demonstrate utility of integrated data through support of projects that use the data to address clinically meaningful questions

What planning activities led to the identification and development of this topic?
1. IC-specific or multi-IC planning activities led to plans for individual cohorts. The concept of integrating data across these cohorts emerged from the Kids First Working Group.
2. BD2K planning activities also contributed to the concept of Data Commons generally, and this program would be developed in coordination with BD2K.

Why is a trans-NIH strategy needed to achieve these goals?
Since individual cohorts are managed by different ICs, the integration of data across cohorts will require trans-NIH cooperation.

What activities might form the strategic plan for this topic?
1. Identification of high-priority "use cases" – Additional community input is needed to articulate the most compelling types of analyses which a Data Commons could support. Workshops or other means of acquiring community input would be conducted in early FY 2015.
2. Cohort Enrichment – This initiative would make funds available to cohorts to expand their data acquisition. Use cases of the integrated data would also be described through these applications; cohorts would be selected based on the planned use of the Commons data and on broad utility of data that the cohort would provide to the Data Commons. This initiative could begin in FY15 through competitive revisions to existing awards or through provision of supplemental funds to competing awards for new cohorts being launched by ICs this year.

3. Pediatric Data Commons - This center will develop the computational requirements for data integration and would establish a "single access point" for diverse types of data.
4. Data Coordination Center – Potentially part of the Data Commons award, the DCC would work with individual research award sites to acquire data and with the Data Commons to ensure that data meet the necessary standards.
5. Demonstration Projects – This initiative would support projects that mine the integrated data sets to address clinically relevant problems and to illustrate the utility of the resource for the broader community.

If a Common Fund program on this topic achieved its objectives, what would be the impact?
1. Establishment of a single access point (the Commons) for users to access and query several different types of data from thousands of children would enable many clinically important questions to be addressed. It would expand utility of the data beyond the immediate purposes for which each cohort was initiated and provide a resource for the greater community.
2. Once established, this data resource would be catalytic for future IC-supported research. It is hypothesis generating. Demonstration projects conducted through this program will illustrate the utility of the approach to the broader community.