Children's Hospital of Philadelphia

Annual Progress Report: 2006 Formula Grant

Reporting Period

July 1, 2010 – December 31, 2010

Formula Grant Overview

The Children's Hospital of Philadelphia received $3,315,523 in formula funds for the grant award period January 1, 2007 through December 31, 2010. Accomplishments for the reporting period are described below.

Research Project 1: Project Title and Purpose

Genome Copy Number Variation Analytics - This project seeks to discover and determine novel underlying genetic causes of and contributors to a variety of childhood diseases, by leveraging our institutional strengths in human genome surveillance, the use of electronic medical records, and computational techniques for processing and analyzing large and complex data sets. We propose to: 1) establish a computationally-based analysis pipeline for identifying variations in the genomes of individuals with or at risk for diseases with a known genetic contribution; 2) create tools and data visualization interfaces to assist with determining whether an individual has a genome copy number alteration that is likely associated with that individual's diagnosed or predisposed disease; and 3) to refine the analytical process for suitability as a novel and superior platform for clinical diagnostics of genetic abnormalities.

Duration of Project

1/1/2007 - 12/31/2010

Project Overview

Much human genetic variation arises from variations in genome copy number (CNVs). Certain CNVs disrupt one or more genes to the extent that they contribute to a disease etiology; such diseases are referred to as "genomic disorders". Recent evidence suggests that CNVs comprise up to 12% of individual genomes, and that certain of these differences likely contribute significantly to human disease. Despite known correlations between certain CNVs and particular genomic disorders (e.g., Down Syndrome, epilepsy, mental retardation), CNVs likely play an as-yet-undiscovered role in many common diseases. The Children's Hospital of Philadelphia (CHOP) has heavily invested in high-volume, high-resolution genome scanning technologies as well as creating an all-encompassing electronic health record (EHR) system. These strategies uniquely position CHOP to be able to identify CNVs on a large scale, and to subsequently discover genetic etiologies for a number of childhood diseases for which a genomic component exists, including birth defects, autism, ADHD, and cancer. This project will create analytical

processes to maximize the value of CHOP's wealth of clinical data and experimental expertise. In turn, the resulting information will accelerate translation of these research findings into successes in diagnostic, therapeutic, and personalized medicine. The Specific Aims are to:

- *Develop methods for high-throughput analysis of genome surveillance data for identifying CNVs.* We will develop ways to automatically process and manage high-resolution genome surveillance data. This data will be analyzed to identify possible copy-number variations, calculate the probability that these events are actual CNVs, and determine whether these CNVs are likely to be normal variations in the population.
- *Create tools for assessing CNVs in a wider biomedical context.* We will integrate known information regarding genomics and diseases to annotate the identified CNV regions, such as gene annotations and the biomedical literature. This information will be used as a visualization platform to enable investigators to assess the likelihood that certain CNV events observed in children play a role in a disease process.
- Develop a process for using the analytical results for improved clinical diagnostics and delivery to the EHR. We will determine the parameters by which CNVs identified by our analyses can be labeled as CNVs causative for known and established genomic diseases. These diagnostic labels will be assessed for accuracy and for their suitability for clinical diagnostics and EHR-based clinical decision support.

Principal Investigator

Peter S. White, PhD
Research Associate Professor of Pediatric Oncology
The Children's Hospital of Philadelphia
Abramson Research Center
34th Street and Civic Center Boulevard
Philadelphia PA 19104-4318

Other Participating Researchers

Ge Zhang, PhD, Xiaowu Gai, PhD, Jonathan Crossette, MPH, Svetlana Ostapenko, MA, Mark Ramos, BS, Mark Porter, MS, Ryan O'Hara, BS, Adam Wenocur, BS, Kevin Murphy, BS, Juan Perin, MS, Monica D'arcy, MS - employed by Children's Hospital of Philadelphia (CHOP)

Expected Research Outcomes and Benefits

As the nation's leading pediatric hospital, CHOP is making considerable investment in data generation and data management infrastructures, including an electronic health record that records over 1 million visits per year and a genomics initiative to generate ultra-high resolution snapshots of the genomes of 100,000 children in the next two years. This project will provide means for realizing the potential of these unprecedented information stores by identifying signature events in individual genomes that drive or contribute to a variety of childhood diseases.

Expected outcomes and benefits will include the following:

- A dataflow, policies, and procedures that allow the high-resolution genotypes of all eligible individuals to be analyzed for the presence of copy number variations.
- An automated data flow and data management system where genotypic data, de-identified phenotypic annotations, analysis results, and integrated external data are aggregated in a queryable relational database.
- Analysis tools for precise, efficient, and facile batch visualization and querying of predicted CNVs, their underlying genotype data points, and relevant integrated genomic and disease-specific data sets.
- A method for identifying putative CNVs from the genotype data and for prioritization ranking of the CNV predictions, based upon quality and quantity of the implicating genotype data and associated information.
- Methods for determining the likelihood that an identified CNV is causative or associated for a disease, and determination of the associated disease for those diseases with clear and established genetic etiologies.
- Development of methods for assuring that assigned disease labels are sufficiently accurate for use in clinical diagnostics, and for determining how the analysis pipeline can be modified to be suitable for diagnostics.
- Development of methods for delivering diagnostic predictions and supportive evidence to an EHR environment for electronically-enabled clinical decision support.

Summary of Research Completed

In the current reporting period, researchers have made additional progress for all Aims of the project. The researchers have continued to make enhancements to the automated workflow for CNV detection in both the research and clinical diagnostics arenas using high-resolution SNP arrays. The analysis software platform has undergone additional iterations and has been used by a substantial number of research groups worldwide. The researchers have utilized the workflow to identify CNVs in a variety of pediatric disorders; these discoveries have led to additional insights into the genetic basis of the targeted diseases, and data generated from these discoveries has been used by many biomedical researchers. Researchers have continued to utilize the workflow for clinical diagnostic purposes at Children's Hospital of Philadelphia (CHOP). Researchers have also applied knowledge of the CNV detection process to more effectively analyze next-generation sequence data for CNV content.

<u>Data workflow</u>

- *CNV Detection and Analysis*: Subsequent to the release and publication of the CNV detection and analysis platform CNV Workshop in February, 2010, researchers have added numerous additional features and improvements to the software. The most recent version of the software is CNV Workshop 2.011, which was released to the open-source community in late August, 2010. Enhancements include:
    - Enables labeling hemizygous variations for males on chromosome X
    - GBrowse graphical displays are enabled for searches
    - Simultaneous searching of all non-public data sets to which the user has access
    - Allows separate email address to which exception (error) email messages are sent.

- *CNV Workshop utilization*: A description of CNV Workshop was published in BMC Bioinformatics in the prior reporting period, and the software has been made available without restrictions through the SourceForge software repository. Since that time, the manuscript has been accessed over 4,000 times, yielding a "Highly Accessed" tag from the journal. The latest version of the software (8/2010) has been downloaded over 350 times (cumulative >1,000) and is in use by many investigators in the genome variation research community.
- *Facilitation of diagnostic workflows*: Researchers have implemented the current version of CNV Workshop for use in all clinical genetics and cancer genetics consults requiring genomic analysis at our institution; prior version of the software have been in place since May of 2008, and it has collectively analyzed over 4,000 subjects. This implementation completes a patient-to-genotype-to-electronic health record pipeline that provides genomic resolution, accuracy, and speed exceeding any academic or commercial operation for which the researchers have knowledge.

Results

To date, the researchers have processed nearly 60,000 (>10,000 new) samples genotyped on various platforms. Recent projects in which researchers have directed CNV analysis using the approaches developed in prior years include:

- *ADHD* (>500 patients and family members). A characterization of CNV content in a large set of trios with attention-deficit hyperactivity disorder. This work has led to the first genome-wide study of structural variation in patients with ADHD
- *Autism* (4,000 patients and family members). Analysis of CNV content for a large, uniformly genotyped cohort of individuals with ideopathic autism. This work has led to a landmark study that demonstrates statistical support for an association between inherited CNVs found in autism patients and specific neurological functions, including synaptic development and glutamatergic signaling.
- *Congenital heart defects* (>2,000 patients and family members). Analysis of patient populations with left-sided heart defects and conotruncal abnormalities for CNV content. Researchers have incorporated this work as CHOP's contribution to the NHLBI/NIH-sponsored Bench-to-Bassinet program (http://www.benchtobassinet.com/).
- *Clinical diagnostics* (>4,000 samples; all subjects from the CHOP Cytogenomics and Cancer Genetics clinical laboratories receiving a genetic consult). The work has led to the first fully integrated genetics diagnostic workflow utilizing SNP arrays as a general screening tool in an academic research environment.
- *CHOP CNV map* (>2,000 subjects). Analysis of CNV content to support the CHOP CNV resource, a public interface specific to CNV content identified in a set of 2,026 healthy individuals (http://cnv.chop.edu). Since its release in July, 2009, this resource has emerged as a leading standard for human structural variation research and application, having received over 1.4 million web hits since launch, and the manuscript has now been cited by over 50 subsequent studies.

Functional analysis

- *Pathway-based functional analysis:* The researchers have continued to refine applications to identify biologically functional groups enriched for CNVs in particular disease cohorts relative to controls. This process has now been perfected for three new large cohorts: autism, conotruncal defects, and left-sided heart lesions. For each, the researchers have identified inherited CNVs in the patient population that were not detected in healthy individuals. In each case, the CNV-associated gene set was significantly enriched for genes reported as candidates in studies of related disorders. For autism, the gene set was significantly enriched for genes known to be important for psychological and neurological functions, including synaptic processing, central nervous system development, and glutamate-mediated neurotransmission. For conotruncal defects, enriched processes included functional pathways associated with early heart development. These results suggest that rare inherited structural variations play an important role in development of these disorders, indicate sets of putative candidate genes for further study in the etiology of these diseases.
- *Phenotype analyses*: Researchers have perfected a model similar to that used for functional attribute-based CNV enrichment to identify traits and phenotypes observed in mouse models of genes associated with CNVs in a disease. This has been most recently applied successfully for autism, where researchers selected a set of genes preferentially overlapping CNVs exclusive to disease subjects that were implicated both by pathway and mouse phenotype enrichment studies. This identified a small set of genes implicated by three different functional methods and replicated in two independent cohorts. This approach suggests a possible way forward toward coalescing risk-associated genome variation data that is distributed across many variants and pathogenic loci for complex disorders.

Sequence-based analysis

The researchers have continued to explore the use of next-generation sequencing as a higher resolution platform for identifying structural variation and base changes, both for research and diagnostic purposes. The pilot study, seeded by the current project and currently continued with support from other sources, includes generating whole-exome sequence data for families with ADHD and nuclear genome-based mitochondrial disorders. Researchers have generated whole-exome sequence for a small number of nuclear families with multiple affected children, and researchers have applied aspects of the variant detection and management pipeline to assess variants. Sequence has been generated for approximately 30 individuals. The researchers have established a semi-automated workflow for sequence trimming, quality control, assembly, alignment, and annotation with genomic features. The researchers' ongoing objectives are to further develop an analytical process analogous to the array-based CNV approach to identify and prioritize functional variants specific to diseased individuals, and to determine the utility of this process as a clinical diagnostics platform.

Academic output

White PS, Xie HM, Perin JC, Takahashi N, Murphy K, Wenocur AS, D'arcy M, O'Hara RJ, Goldmuntz E, Grice DE, Shaikh TH, Hakonarson H, Buxbaum JD, Elia J, Gai X. Rare Structural

Variation of Synapse and Neurotransmission Genes in Autism. 60th Annual Meeting of the American Society of Human Genetics, Washington, DC, November 2-6, 2010.

Hakonarson H, Glessner JT, Wang K, Takahashi N, Shtir CJ, Sleiman PMA, Zhang H, Kim CE, Lyon GJ, Flory JH, Bradfield JP, Imielinski M, Hou C, Fackelton EC, Middleton F, Todorov AA, Reif A, Franke B, Lesch KP, Anney R, Shaw P, Devoto M, Grant SFA, White P, Buxbaum JD, Rapoport JL, Williams NM, Nelson SF, Faraone SV, Elia J. Genome Wide Copy Number Variation Study Associates Metabotropic Glutamate Receptor Genes with Attention Deficit Hyperactivity Disorder. 60th Annual Meeting of the American Society of Human Genetics, Washington, DC, November 2-6, 2010.