Children's Hospital of Philadelphia

Annual Progress Report: 2006 Formula Grant

Reporting Period

July 1, 2009 – June 30, 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.

Anticipated 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

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

The researchers have completed a polished version of a workflow system (CNV Workshop) for analyzing high-resolution genotype data for structural variation. This system has been applied to a number of disease cohorts and has been implemented as a molecular diagnostics tool for all patients referred to CHOP for clinical genetics and cancer genetics assessments. CNV Workshop has been released to the research community as supported open-source software. CNV Workshop is fully automated and consists of multiple functionalities, as described below. To the researcher's knowledge, no other available public or commercial system provides the ability to support all steps required for CNV analysis.

<u>Data workflow</u>

- *Project management*: CNV Workshop is unique in its ability to simultaneously manage multiple project data flows at single site, with authentication and permissioning supported. In this reporting period, the researchers have processed new or additional data from cohorts for a large number of pediatric disorders, including autism, obesity, asthma, ADHD, and congenital heart defects.
- *Array platform independence*: CNV Workshop fully supports all high-throughput genotyping array, include Illumina, Affymetrix, and Nimblegen arrays.
- *Diagnostics platform*: CNV Workshop includes a data import functionality that moves and formats genotyping data into an analysis module with minimal supervision. This allows clinical and high-priority research samples to be very rapidly processed.
- *Facilitation of diagnostic workflows*: Prior versions of CNV Workshop have been utilized to some extent by CHOP for all clinical genetics and cancer genetics consults requiring genomic analysis since May of 2008, constituting over 3,000 samples. In the last year, the researchers have instituted the current version of CNV Workshop that provides all data delivery, management, and integration functions, as well as serving as the primary platform

for analysis in the clinic. This implementation completes a patient-to-genotype-to-report pipeline that provides genomic resolution, accuracy, and speed exceeding any academic or commercial operation that the researchers are aware of.

Data management and infrastructure

- *Hardware*: For use at the institution, the researchers have expanded the computational cluster for analyzing raw genotyping data (a 96-node parallel processing unit). This environment has dedicated, delineated compute nodes for clinical diagnostics, research analytics, and development/backup production. A dedicated server hosts a database that holds separate instances of clinical diagnostics, internal research, external research, and public data. Each instance is secure with an authentication layer that accommodates user, superuser, and admin roles. Users can see projects—defined as sample cohorts—for which they have authorized access. All instances share a single unified data schema.
- *Database*: A MySQL database stores all results, metadata, and genomic annotations. The database is integrated into an institutionally hosted storage area network with full backup and redundancy capabilities built in.
- *Integration*: The researchers have written an interface between CNV Workshop and the widely-used genome analysis package Galaxy, as described at a recent Galaxy Developer's workshop (http://chapmanb.posterous.com/galaxy-developer-conference-day-2-afternoon-d).

Platform

- *Software*: The researchers have completely rewritten the pilot software using a professional and uniform design process that incorporates commonly used software engineering elements. The application, named "CNV Workshop", is written in java. It uses as its major server components a MySQL database server, a web server hosting a custom instance of the GMOD Generic Genome Browser via CGI, and an Apache Tomcat server hosting the Java-based, web-enabled CNV Workshop GUI. The application is supported for use on Linux, Macintosh OS X, and (via a virtual image) Windows platforms. This system is functional for the Hospital's clinical and many research environments.
- *Distribution*: CNV Workshop has been released as an open source package at Source Forge under the GNU Affero General Public License v3: http://cnv.sourceforge.net/. Executables and source code are available. The software was originally released to the public on August 31, 2009 as v1.0. Since that time, numerous enhancements, bug fixes, and additional functionalities have been added through a series of periodic release updates. The most current version (May 13, 2010) is v1.0.10. See Figures 1-3 for examples of the user interface.
- *Awareness*: A manuscript describing CNV Workshop was published in the leading bioinformatics journal BMC Bioinformatics in February, 2010 (see below) and has been established as a "highly cited" manuscript by the journal. Among much internet mention, the software was featured online at the bioinformatics educational site OpenHelix (http://blog.openhelix.eu/?p=3666).
- *Usage*: In addition to widespread use at CHOP, CNV Workshop has been downloaded over 500 times since its public release in 2/2010.

Results

To date, the researchers have processed over 50,000 (>30,000 new) samples genotyped on various platforms, including the following:

- *>27,000 subjects accrued at CHOP through a general recruitment project:* Derived from patients and parents consented at various sick and well-child intake points throughout the CHOP clinical network, including the ER, OR, primary care facilities, and subspecialty clinics. A large number of diagnoses are associated with these individuals.
- *>20,000 samples from patients and parents with various disease diagnoses, including*:
  - ADHD
  - Autism
  - Brain malformations
  - Childhood malignancies, especially neuroblastoma, rhabdoid tumors, and medulloblastoma
  - Congenital heart defects, especially Tetrology of Fallot and left-sided lesions, as part of the Pediatric Cardiology Genetics Consortium (sponsored by a multisite grant from the National Heart, Lung, and Blood Institute)
  - Crohn's Disease
  - Hearing loss
  - Hyperinsulinism
  - Rare genomic disorders and phenotypes, including Alagille Syndrome, Cornelia de Lange Syndrome, DiGeorge Syndrome, Holoprosencephaly, Kabuki syndrome, and Multiple congenital abnormalities
- *>3,000 clinical samples for clinical diagnostics*, from the CHOP Cytogenomics and Cancer Genetics clinical laboratories

- *Public interface*: A public interface specific to CNV content identified in a set of 2,026 healthy individuals was launched in concert with a related manuscript that describes this work (http://cnv.chop.edu;see below). This site has comparable functionality to the project-specific web resource but is limited to the healthy controls. The set of controls establishes a high-resolution CNV map of over 50,000 structural variants from disease-free individuals in two populations. This resource has quickly emerged as a leading standard for human structural variation research and application since its release in July, 2009. The site has received over 700,000 web hits since launch, and the manuscript has already been cited by over 30 subsequent studies.

Functional analysis

- *Pathway-based functional analysis:* The researchers have further developed applications to identify biologically functional groups enriched for CNVs in particular disease cohorts relative to controls. This process has been applied to subjects with ADHD, autism, and heart conotruncal abnormalities. For each disease, the researchers have identified specific CNVs in the patient population and their parents that were not detected in a large number of unrelated healthy individuals. In each case, the CNV-associated gene set was significantly enriched for genes reported as candidates in studies of related disorders. For ADHD and autism, the gene

sets were also significantly enriched for genes known to be important for psychological and neurological functions, including learning, behavior, synaptic processing, central nervous system development, and glutamate-mediated neurotransmission. For conotruncal defects, enriched processes included functional pathways associated with early heart development. Together, these results suggest that rare inherited structural variations play an important role in development of these disorders and indicate sets of putative candidate genes for further study in the etiology of these diseases.

- *Phenotype analyses*: Analogous to the pathway-based enrichment methods the researchers devised, the researchers have applied a similar model to identify which traits and phenotypes observed in mouse models of genes associated with CNVs in a disease were significantly enriched. This method was initially tested for the ADHD-specific CNV gene set and then again tested for two large autism cohorts. In both cases, a number of phenotypes were significantly enriched, all of which were consistent with pathophysiology of these diseases. For autism, the 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 set of 12 genes separately implicated by three different functional methods and replicated in two independent cohorts. The researchers believe that these genes form a set of intriguing candidates for further study; these genes typically were involved in glutamate-mediated neurotransmission and synaptic function.

Clinical diagnostics

In the last year, the researchers have continued to work closely with our Clinical Diagnostics Labs to perfect SNP array-based clinical diagnostics for patients requiring genetics testing by the Hospital's Clinical Genetics and Cancer Genetics services. Samples from all patients in need of testing are now genotyped on the Illumina 660K array as a front-line diagnostic, using the latest version of our CNV Workshop platform. This system has now been in production for over two years, with over 3,000 samples processed.

Sequence-based analysis

The researchers have begun 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. Pilot studies include generating whole-exome sequence data for families with ADHD and nuclear genome-based mitochondrial disorders. In each case, the researchers are generating whole-exome sequence for a small number of nuclear families with multiple affected children. The researchers' ongoing objectives are to 1) generate a dataset for CNV and base variant analysis; 2) develop an analytical process analogous to the researchers' array-based CNV approach to identify and prioritize functional variants specific to diseased individuals; and 3) determine the utility of this process as a clinical diagnostics platform.

Academic output in FY10

Shaikh TH, Gai X, Perin JC, Glessner JT, Xie H, Murphy K, O'Hara R, Casalunovo T, Conlin LK, D'arcy M, Frackelton EC, Geiger EA, Imielinski M, Kim CE, Medne L, Wang K, Annaiah K, Bradfield JP, Dabaghyan E, Eckert A, Onyiah CC, Ostapenko S, Otieno FG, Santa E, Shaner

JL, Skraban R, Smith RM, Elia J, Goldmuntz E, Spinner NB, Zackai EH, Chiavacci RM, Grundmeier R, Rappaport EF, Grant SFA, White PS, Hakonarson H. High-resolution mapping and analysis of copy number variations in the human genome: A Data Resource for Clinical and Research Applications. Genome Research, 19:1682-90, 2009.

Gai X, Perin JC, Murphy K, O'Hara R, D'arcy M, Wenocur A, Xie HM, Rappaport EF, Shaikh TH, White PS. CNV Workshop: An integrated platform for high-throughput copy number variation discovery and clinical diagnostics. BMC Bioinformatics 11:74, 2010.

Elia J, Gai X, Xie HM, Perin JC, Geiger E, Glessner JT, D'arcy M, deBeradinis R, Frackelton E, Kim C, Lantieri F, Muganga BM, Wang L, Rappaport EF, Grant SFA, Devoto M, Shaikh TH, Hakonarson H, White PS. Rare inherited structural variants found in Attention-Deficit Hyperactivity Disorder are preferentially associated with neurodevelopmental genes. Mol Psychiatry, 15:637-646, 2010.

Gai X, Elia J, Xie HM, Perin JC, Geiger E, Glessner JT, D'arcy M, deBeradinis R, Frackelton E, Kim C, Lantieri F, Muganga BM, Wang L, Rappaport EF, Grant SFA, Marcella Devoto M, Shaikh TH, Hakonarson H, White PS. Rare inherited structural variants found in Attention-Deficit Hyperactivity Disorder are preferentially associated with neurodevelopmental genes. 59th Annual Meeting of the American Society of Human Genetics, Honolulu HI, October 20-24, 2009.

Hakonarson H, Glessner JT, Wang K, Gai X, Takahashi N, Sleiman PMA, Flory JH, Zhang H, Kim C, Bradfield JP, Imielinski M, Rabin C, Shaw P, Shaikh TH, Grant SFA, Buxbaum JD, Rapoport JL, Devoto M, White PS, Elia J. Genome Wide Copy Number Variation Study Associates Metabotropic Glutamate Receptor Genes with Attention Deficit Hyperactivity Disorder. 59th Annual Meeting of the American Society of Human Genetics, Honolulu HI, October 20-24, 2009.

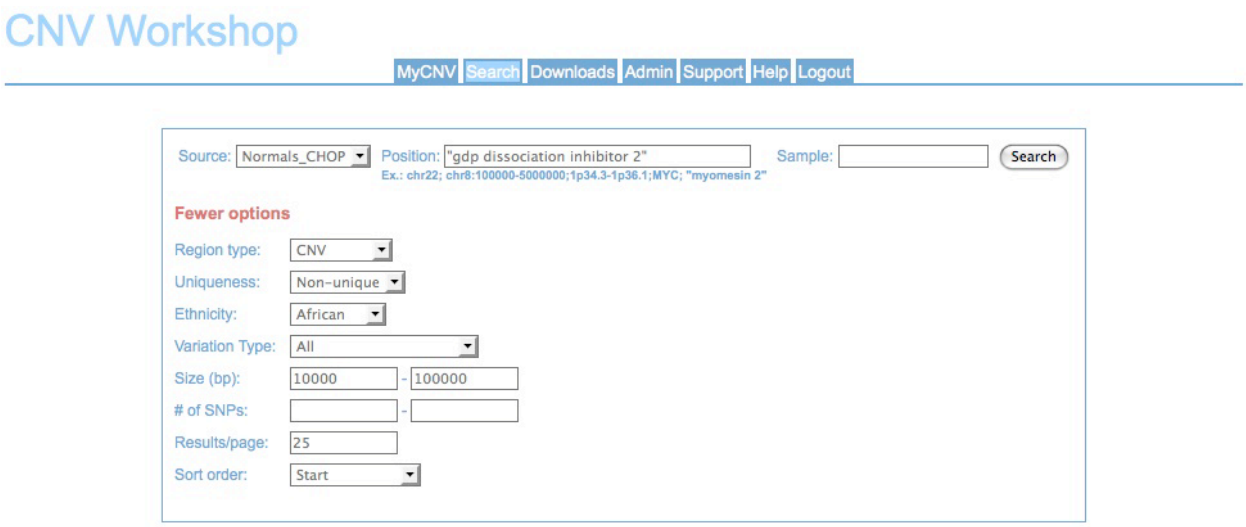

Figure 1**.** CNV Workshop query interface. Screenshot of the simple query interface in "annotated" mode for the CHOP CNV map database. Annotated mode enables a user to query by sample. Positional queries supported are chromosome(s), cytogenetic band(s), sequence position or range(s), and gene name(s). Most non-standard gene names are recognized and normalized to HGNC gene symbols. For the CHOP CNV map, additional searchable fields are CNV type (CNV, CNV region, or CNV block), ethnicity, and uniqueness (unique or non-unique/observed in multiple unrelated individuals).

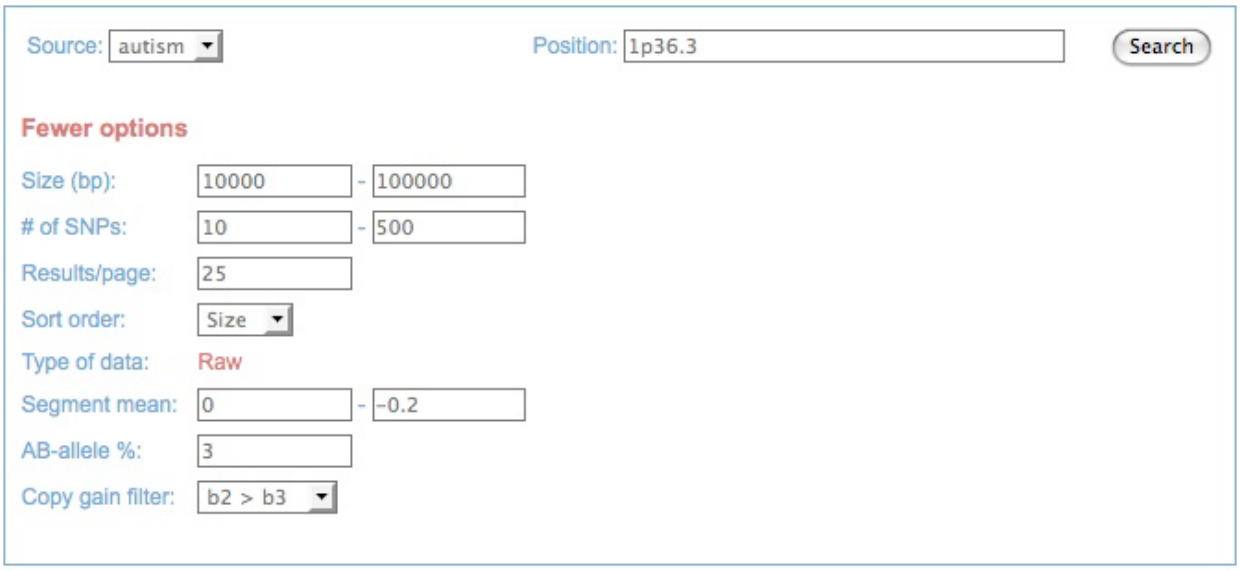

Figure 2**.** Screenshot of the advanced query interface in "raw" mode for a database of autism samples. Raw mode enables a user to query by segment mean, minor (AB) allele frequency, and to filter results based upon allelic ratio statistics.

#		Type	Sample	Band	Start	End	SNPs	segSize	segMean	GAD	Custom Track	DGV CNVs	UCSC	Content
1	☐	hetero	AU028201_98HI0113B	16p12.2	20,559,524	20,580,129	4	20,605	-0.828	METABOLIC		4	Map	ACSM1 *Fable Entrez Gene*
2	☐	hetero	AU1282302_04HI3189A	16p12.3	20,391,462	20,408,047	4	16,585	-0.543			5	Map	ACSM2A *Fable Entrez Gene*
3	☐	hetero	AU1282202_04HI3191A	16p12.3	20,391,462	20,408,047	4	16,585	-0.606			5	Map	ACSM2A *Fable Entrez Gene*

Figure 3. Presentation of query results in CNV Workshop. Depicted are results for a chromosome 16 query of sequence position range 20,300,000-20,610,000 in an autism cohort. Top panel: graphical display. Layers (top to bottom) represent the sequence position (top), cytogenetic bands, CNVs observed in the autism cohort, CNVs in the CHOP CNV map, CNVs in the Database of Genomic Variants (labels indicate the study), phenotypes of Genetic Association Database studies, and UCSC Known Genes. All glyphs hyperlink to corresponding database records. Bottom panel: tabular display for a subset of CNVs. Sortable column headers are colored red or (for current sort order) green. Each row value colored red denotes a hyperlink to a corresponding external database record. Checkboxes at far left allow a user to save selected CNV sets in a MyCNV clipboard.