Complete genomics

CASE STUDY

Whole Genome Sequencing of a Nuclear Family Reveals Candidate Genes for Rare Genetic Diseases

The Institute for Systems Biology, Seattle, Washington

Fast Facts

- Project: Whole human genome sequencing of a family
- Number of genomes: 4
- Result: Validation of candidate genes and causal variants for two Mendelian disorders

Introduction

Ever since the human genome was first sequenced, scientists have been inspired by the possibilities of using genomic information for medical research. This potential, however, has not been fully realized due to the time and expense involved in performing even just a handful of whole human genome studies. These studies can take years and require hundreds of thousands or even millions of dollars, including purchasing and managing the sequencing instrumentation and massive computing resources.

Complete Genomics is changing that, with a whole human genome sequencing service based on proprietary technology that yields high-quality results for a fraction of the cost of the leading systems on the market. With our service, researchers can directly obtain comprehensive, annotated reference and variant calls from genomic DNA samples, without spending time and money operating instruments, preparing DNA libraries, developing extensive computing infrastructure, or acquiring expertise in genome assembly methods.

The Miller Syndrome Family Study

The practical utility of the Complete Genomics approach, in collaboration with the Institute for Systems Biology (ISB), was first presented at the Personal Genomes meeting, Cold Spring Harbor, New York, in September 2009, with results later described by J. Roach, et al. in Science, 1186802, (11 March 2010). In this pilot project, Complete Genomics sequenced the individual genomes of a four-member nuclear family, including two unaffected parents and their two children who suffer from two genetic disorders: Miller Syndrome and primary ciliary dyskinesia.

The high-quality sequence data provided by Complete Genomics in combination with novel inheritance analyses enabled ISB to narrow the search for variations consistent with recessive inheritance of a rare allele to four candidate genes across the entire genome. One of these genes, DHODH, was concurrently identified as a cause for Miller Syndrome, and mutations in a second gene, DNAH5, have previously been shown to cause primary ciliary dyskinesia.

Family Genome Sequencing

Working with genome sequence data from a nuclear family offers researchers the opportunity to:

- Analyze Mendelian inheritance patterns and identify recombination sites precisely, thereby achieving even greater accuracy in variation calling than would be possible with single genomes
- Identify candidate genes consistent with various modes of disease inheritance (e.g., recessive, dominant, complex)
- Identify *de novo* mutations arising in the germline, as well as somatic mutations

In keeping with Complete Genomics' standard human genome sequencing service, the genomes in this study were all sequenced to an average depth of coverage greater than 40x. An 85-92% call rate (percentage of the bases called as reference, heterozygous or homozygous variant) was obtained for these genomes. More than 4,471,000 SNPs were identified; among these, more than 3,665,000 SNPs varied within the family. Comparison to previously published exome sequencing data for the offspring in the family yielded an estimated error rate within called bases of 8.16×10^{-6}. Inheritance-based analyses further reduced the effective genotype error rate relevant to downstream analyses to 3.3×10^{-6}.

Opportunity for Discovery

In a four-member nuclear family with two unaffected parents, such as the family in this study, genomic information is particularly rich. Informative markers are dense, and missing data can be inferred by considering inheritance patterns of the genotypes of other individuals. As a result, it is possible to precisely determine phase, haplotypes, and recombinations for entire chromosomes (Figure 1).

Identifying Disease Causing Variations

By comparing whole genome sequences of the unaffected parents and offspring affected by a rare disorder, one may identify candidate rare variations and genes responsible for the disorder. Disease models considered in this study included simple recessive and compound heterozygote inheritance. Considering rare mutations (using novelty with respect to dbSNP and other SNP databases) is a powerful constraint.

In compound heterozygote inheritance, both children separately inherit different deleterious mutations within a single gene from each parent (Figure 2). This study uncovered three compound heterozygote candidate genes, all of which were independently identified by exome sequencing (S. B. Ng et al., Nat Genet 42, 30 (Jan, 2010)). One is the likely cause of Miller Syndrome, as confirmed in unrelated affected individuals. The other explains the lung disorder.

Inheritance analysis within families can be used to identify candidate alleles that cause genetic disorders. In this study, both offspring were affected by Miller Syndrome

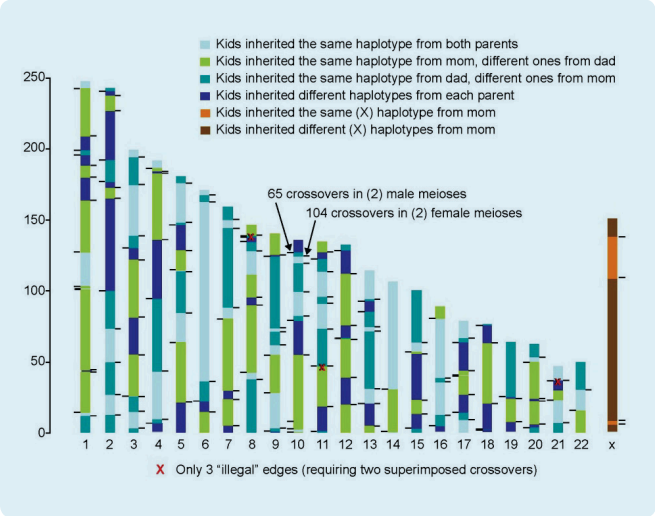

Figure 1. *Detection of crossover sites in the Miller Syndrome family study. All single recombinations that occurred in any of the four meioses generative for the children will cause a shift from one to another of these inheritance combinations. Therefore, locations of the recombinations may be identified based on the shift in inheritance patterns.*

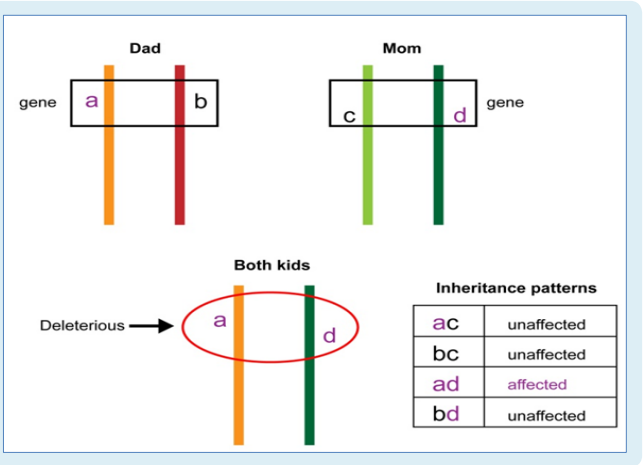

Figure 2. *Potential disease-causing genes discovered by compound heterozygote inheritance. The detrimental variants were inherited by both kids from the heterozygous, unaffected parents.*

and primary ciliary dyskinesia, rare disorders for which prior data are consistent with a simple recessive mode of inheritance.

Since both offspring are affected, genes consistent with recessive inheritance must lie in regions of the genome where they share both parental haplotypes, thereby limiting the search to about a quarter of the genome (22% in this family, based on the high-resolution recombination map obtained from a comparative analysis of the familial genomes as part of this study). Furthermore, both diseases are rare and are also likely to be caused by very rare variants not present in dbSNP or any other database. Under these constraints, only two non-synonymous SNPs, both missense variations in the CES1 gene, matched the simple recessive mode, and three genes (DHODH, DNAH5, and KIAA0556) were consistent with the compound heterozygote mode.

DHODH was identified as the primary gene for Miller Syndrome based on sequencing the exomes of these two affected offspring along with the exomes of two unrelated Miller Syndrome patients (S. B. Ng et al., Nat Genet 42, 30 (Jan. 2010)). Mutations within DNAH5 have been previously identified as a cause of primary ciliary dyskinesia (H. Olbrich et al., Nat Genet 30, 143 (Feb. 2002)), and so were likely the cause in this family as well. Thus, an analysis of whole-genome data within a single family yielded two disease genes out of a total of four disease-associated genes consistent with recessive inheritance. The roles of the other two genes, if any, are unknown.

Outside exons, several possibly detrimental variants consistent with simple recessive inheritance were detected:

- Two in highly conserved regions,
- One in an intronic sequence near a splice site,
- Five in non-protein coding transcripts, and
- One in a UTR

In particular, one non-exonic candidate SNP disrupts a putative acceptor splice site 5' of a previously unannotated exon in SP9, the mouse ortholog of which has been implicated in embryonic skeletal malformation. These non-exonic variations offer further avenues for investigating the cause of the two diseases observed in this family.

A Proven Approach

The results of this study underscore the accuracy and simplicity of Complete Genomics' human genome sequencing service. As the world's first company dedicated to large-scale whole human genome sequencing and analysis as a service, Complete Genomics enables scientists to conduct human disease research on up to thousands of genomes. Complete Genomics offers unparalleled genomic sequencing expertise via a simple outsourced service solution. There are no instruments to purchase, maintain, and upgrade; no laboratories to build and staff; and no issues with scaling up or scaling down as your capacity requirements change. And by focusing exclusively on the human genome, Complete Genomics has optimized its sequencing technology for the study of human DNA, allowing it to provide the assembled sequences and variants reproducibly, quickly and with the highest accuracy. Access to affordable human genome sequencing on a large scale allows researchers to gain a deeper understanding of the genetic mechanisms underlying drug response and complex disease.

Complete Genomics Sequencing Service

Complete Genomics' Sequencing Service provides researchers:

- Most accurate, most comprehensive data results – Leading to faster biological interpretation
- Easy-to-use service – End-to-end project management; trusted informatics support
- Highly scalable and reliable service – Sequence very large projects with shorter time-to-results
- Outsourced solution – No capital expenditure required

Researchers ship high-molecular weight, genomic DNA directly to Complete Genomics. In return, they receive whole genome sequence data ready for biological interpretation.

References

1. J. Roach, et al., Science, 1186802 (March 11, 2010).
2. R. Drmanac, et al., Science 327, 78 (January 2010).
3. S.B. Ng, et al., Nat Genet 42, 30 (January 2010).
4. H. Olbrich, et al., Nat Genet 30, 143 (February 2002)

www.completegenomics.com info@completegenomics.com
2071 Stierlin Court, Mountain View, CA 94043 USA Tel 650.943.2800

Complete genomics

Published in U.S.A., March 2012, CS_MS-02