Complete genomics

APPLICATION NOTE

Switch from Exome to Whole Genome Studies

- Affordable: Low cost per genome for large studies
- High-quality: Better quality exomes than by selection methods
- Comprehensive: Non-coding and coding variants including SV and CNV
- Easy-to-use: Tables of research-ready annotated coding and non-coding variants
- Fast: Large studies comprising hundreds of whole genomes can be done in a few months
- Informative: Need fewer samples per study: Whole genomes contain more information (e.g., regulatory variants; allow haplotyping)

To Get the Whole Picture, Sequence the Whole Genome

Comparing whole genome sequencing and exome sequencing

Introduction

Next-generation DNA sequencing empowers scientists to identify genetic variations associated with human disease at higher resolution and greater sensitivity than previously possible. Two approaches are commonly employed -- exome sequencing and whole genome sequencing. Exome sequencing targets protein-coding regions comprising approximately 1% of the human genome, while whole genome sequencing uses an unbiased approach to investigate the majority of the human genome, including a comprehensive survey of coding and non-coding regions.

Researchers perform exome sequencing for two main reasons. First, targeted sequencing may cost less than whole genome sequencing. Secondly, at the moment, protein-coding variants are more easily interpreted than non-coding variants. However, by targeting only a specified list of protein coding sequences, DNA variations outside of these regions are missed. Moreover, exome capture by hybridization can introduce considerable coverage variability which may impact comparative analyses and therefore limit discovery efforts. In addition, some types of events such as copy number and structural variations (CNVs and SVs), as well as some insertions, deletions and block substitutions, may be difficult to detect in exome capture data. Overall, whole genome sequencing provides a more comprehensive approach for identifying the genetic causes of disease.

Disease Biology Inside and Outside the Exome

While whole genome sequencing provides a more thorough picture of the genome, there are several reasons why researchers choose to do targeted exome sequencing. Exome sequencing produces less raw sequence data than a whole human genome and therefore reduces the overall cost of the project, thus potentially allowing a larger number of samples to be studied. In addition, exome sequencing has been successful at identifying disease causing variants (1) (2).

Conversely, whole genome sequence allows researchers to consider not just the coding regions but also the non-coding functional genetic elements. Non-coding regions make up ~99% of the genome, and while a good fraction of these nucleotides may be relatively silent in terms of the impact of variation on phenotypes, thousands of specific, well-annotated and conserved elements exist outside of the protein-coding regions and have been found to be implicated in various diseases (Table 1).

In addition, it is now widely recognized that a much larger fraction of the human genome is systematically transcribed than previously thought and entirely new classes of non-protein-coding genes continue to be discovered and characterized (17). Even when some of these regions are included in exome capture kits, the list of targets tends to be incomplete as the community's catalog of these loci is continually expanding. Whole genome sequences can be re-annotated against such evolving databases, at any time, without re-sequencing. Intriguingly, a sizable fraction of loci identified by genome wide association studies (GWAS) lie within "gene deserts" or genomic regions with no known protein coding genes (18). The detailed functions of many of these non-coding regions remain to be thoroughly elucidated. Whole genome studies will be an important approach for expanding our knowledge of the role of both coding and non-coding genomic regions in human disease and basic biology.

Structural Variants Including Copy Number Changes

With whole genome sequence data, larger genomic variations can be detected using a variety of approaches; by read-depth counting (normalized read depth correlates with locus-specific copy number), discordant paired-end mappings (indicating regions of heterozygous or homozygous structural changes with respect to the reference genome), and loss of heterozygosity (LOH), where homozygous runs of smaller variants are detected within larger hemizygous deletions. These data analyses support each other and thus improve sensitivity and specificity. For example, any copy number change in a diploid region must correspond to (at least one) structural change, a deletion, insertion, or translocation, and any larger deletion should correspond to an LOH region. In addition, whole genome sequencing has the capability to detect many copy neutral events, including uniparental disomies (which cannot be detected by array CGH) as well as inversions or translocations (not detectable by array CGH or SNP arrays). Somatic structural changes in tumors have been validated as "driver mutations" in various cancers, *de novo* structural variants are well known as causes of certain developmental disorders, and inherited copy number variants have been linked to psychological disorders, HIV susceptibility, and others (19) (20) (21). Complete Genomics data have been shown to detect many such variants and the company provides

DISEASE	GENE/REGION	VARIATION	FUNCTIONAL ELEMENT	REFERENCE
Chronic Myelogenous Leukemia	BRC-Abl1	Translocation	Gene fusion	(3)
ß-thalassemia	ß-globin	CACC-box duplication	Promoter	(4)
Allergies and Aslhma	IL-10	SNP	Promoter	(5)
	TGF- ß1-1	SNP	Promoter	
	TGF- ß1-3	300bp Deletion	Promoter	
	TGF- ß2	308bp Deletion	Promoter	
Asthma	IL-4	SNP	Promoter	(6)
Hereditary thrombocythemia	TPO	Various	5' UTR	(7)
HIV susceptibility	CCL3L1	Copy number change	Gene	(8)
Alzheimer's	APP	Duplication	Gene	(9)
Breast Cancer	RAD51	SNP	5' UTR	(10)
Congenital Heart Disease	GATA4	Various	3' UTR	(11)
Hypertension	Angiotensinogen	SNP	Promoter	(12)
Breast Cancer	BRCA1 and BRCA2 (targets)	SNP	microRNA precursor gene	(13)
ADHD	PTPHD	Deletions	Various (including one intronic-only)	(14)
	GRM5	Deletion	82kb of gene	
Breast Cancer	FBXW7 and NM_018315	Deletion	46kb of 2 genes	(15)
	ERVL-MaLR and ABCA2	Translocation	LTR and 5' exon	
Coronary Artery Disease	Cdkn2a and Cdkn2b (affected)	Deletion	Intragenic cis-acting transcription regulator	(16)

Table 1. *Examples of Non-Coding and Structural Variants Implicated in Disease.*

structural variants as part of its standard service offering.

In contrast, identifying structural variations, copy number, and LOH in most exome sequence data is challenging. First, the highly uneven capture efficiency makes correlating read depth and copy number far more difficult than with whole genome data. Secondly, paired ends can be used to detect structural variants when they span both sides of a junction. Paired ends are rarely used in exome sequencing and the typical clone insert size is much larger than the average exon size. Even if paired ends are used, they only help detect structural variants when they span a junction. It is unlikely that most of the functionally relevant junctions have both ends contained in the small fraction of the genome targeted by exon capture. Finally, LOH analysis can have reduced power (depending on the size of the deletion) given the uneven spatial sampling of small variants in exome data and the complicating effects of linkage disequilibrium on SNPs within close proximity.

Technical Considerations in Exome Sequencing

Exome sequencing may be successful in finding variants in certain regions. However, it is important to understand that current commercially available exome targeting kits (for example, Agilent SureSelect or Roche/NimbleGen Sequence Capture products) typically only cover the CCDS exome set. The CCDS definition of the exome is approximately 18K of the 22K putative human genes (http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi). In addition, even within this set not all of the desired exome is typically captured by these kits as certain exomic regions have to be excluded during design of the capture probes because of hybridization thermodynamics and/or cross-hybridization potential. Segmental duplications and conserved pseudogenes create particular challenges in targeting by hybridization. Furthermore, RNA sequencing continues to identify novel exons that are not targeted in otherwise well-characterized genes (22).

Exome capture methods tend to have highly uneven yield across the targeted regions. Some regions will be greatly over-covered by mapped sequence, while others may be missed or barely covered. At the same time significant amounts of off-target DNA (including fragments flanking and partly overlapping a targeted region) will typically be captured and sequenced. To compensate, as much as possible, many expert laboratories now generate disproportionately large amounts of raw DNA sequence for these projects (23). Differences between the allele in the sample and the probe sequence based on the reference genome can result in probes which hybridize poorly. If the sample is heterozygous with one allele which is similar or identical to reference, then this reference–like allele will be preferentially captured and sequenced, potentially resulting in the site being called a false homozygote in spite of high depth of coverage. Complete Genomics uses advanced computational approaches, which, in combination with paired end reads, minimizes the impact of such "reference bias" which can also be found in alignment data. For these technical reasons, the most accurate and effective way to sequence the exome at high accuracy and completeness may be to sequence the whole human genome.

Regions and Variants Which May Be Missed by Exome Targeting:

- *Promoters*
- *UTR regulatory regions*
- *Intronic splicing regulators*
- *Genomic regulatory regions (for example, enhancers, insulators, silencer)*
- *Non-coding RNAs (microRNAs, snoRNAs, piRNAs, lincRNAs , endogenous siRNAs and anti-sense RNAs, and others)*
- *Copy Number and Structural Variants*

Example Studies

In a study published in Nature, 38 human multiple myeloma tumors and matched normal samples were subjected to either exome sequencing or whole genome sequencing (24). 23 patients were whole genome sequenced; 16 were exome sequenced; and 1 patient was sequenced by both methods. The results showed that the mutation frequency in coding regions was significantly less than that in intronic and intergenic regions due to negative selection pressure against mutations disrupting coding sequence. 18 statistically significant mutated non-coding regions were identified. While exome sequencing identified the majority of significantly mutated genes, half of the total protein-coding mutations occurred in chromosomal aberrations such as translocations, most of which would have been missed by sequencing only the exome. In addition, recurrent point mutations in non-coding regions would have been missed by sequencing targeted only at coding exons.

A published study in Science sequenced the individual genomes of a four-member nuclear family, including two unaffected parents and two affected children, both suffering from two apparently recessive disorders: Miller Syndrome (a developmental disorder) and primary ciliary dyskinesia (25). By comparing the whole genome sequences of the family members, and leveraging the high accuracy, high call-rate and nearly complete coverage in the whole genome sequence data, the authors constructed a complete high-resolution recombination map of the meiotic events in the family (Figure 1). This map allowed a dramatic reduction in the candidate gene regions. In addition, due to the high call rate, most sites could be directly compared between all four family members resulting in the identification of four genes that were shown to carry putatively recessive novel coding variants (heterozygous sites in the parents but diploid variants in both children). Separately, targeted exome sequencing of the same children (1) identified two of these four variants, and while in this case the two variants detected by exome sequencing may indeed prove causative for the phenotypes under study, one imagines cases where this could have resulted in a false negative result.

In the whole genome sequence data, several possibly detrimental variants also consistent with recessive inheritance were detected outside of the targeted exome. Two of the variants were in highly conserved regions, five were in known non-coding transcripts, one was in a UTR, and one was in an intronic sequence disrupting a splice site (Table 2). This latter variant is 5' of a previously unannotated exon in SP9, the ortholog of which has been implicated in a heritable developmental disorder in mice. These non-exonic variations offer further avenues for investigating the etiology of the diseases observed in this human family. Such analyses would not be possible with only the exome sequence data.

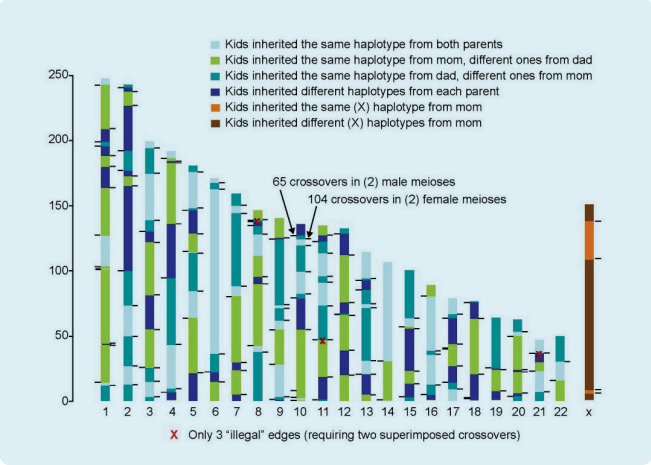

Figure 1. *Detection of Crossover Sites in the Miller Syndrome Family Study*

LIKELY RECESSIVE VARIANTS FOUND	NUMBER
Protein Coding Genes	4
Non-Protein Coding Genes	5
Splicing Regulator of a Protein-Coding Gene	1
Translational Regulatory Region of a Protein-Coding Gene	1

Table 2. *List of Variants Identified in Roach et al.*

Summary

For many types of disease research, one should consider sequencing approaches that interrogate the entire human genome. Whole genome sequencing provides the most comprehensive analysis of coding, non-coding and other functional elements. Whole genome sequencing further enables investigation of novel and recently discovered elements, and provides the ability to see types of variation (CNV, SV, and LOH), which may be technically difficult to observe accurately in exome sequence data.

Complete Genomics offers whole human genome sequencing to at least 40X (~120 GB) mapped coverage. Accurate calls are made for ≥90% (typically >95% for both entire genome and exome) of the entire genome.

Complete Genomics' Sequencing Service

Complete Genomics is the world's first company dedicated to providing large-scale human genome sequencing and analysis as an outsourced service. We make accurate whole genome sequencing affordable, easy, and reproducible. Our unique patterned DNA nanoarray and unchained base ready DNA sequencing technology is highly accurate and efficient. We are the only company to provide rich research-ready genomic variant data, as we provide researchers with finished sequences and annotated variant reports. Our variant files include highly informative confidence scores to balance sensitivity and specificity with explicit differentiation of "no-variant" from a "no-call" position.

We are leading the way in reducing the price of whole genome sequencing, providing researchers with rich, full genome dataset at 40X mappable coverage. For $5,000 per sample in small studies and even less for studies over 50 samples, customers of Complete Genomics can

Complete Genomics provides high quality, whole human genome sequencing and analysis at a price that enables researchers to carry out large-scale disease studies. The company offers its human genome sequencing as a service through its own commercial-scale genome center. This development provides pharmaceutical, biotechnology and other medical research customers with easy access to affordable, high quality, research-ready data for accelerated scientific discovery.

get whole genome data which includes sequence variants (SNPs, small indels, CNVs, and SVs), data summary reports, and full set of supporting data for these results. And as a long-term investment in future genetic studies, Complete Genomics offers a cost-effective solution because it enables researchers to get more comprehensive genetic information from a single dataset.

Whole genome sequencing enables:

- ***Total discovery**: The detection of variations in >90% of the human genome including non-coding regions, as well as the identification of copy number and structural variations.*
- ***Comprehensive study design**: This includes the ability to more reliably compare variants across genomes in disease studies.*
- ***Greater value**: For almost the same cost of sequencing one exome, Complete Genomics offers a whole human genome dataset with a minimum 40X (~120 GB) mapped coverage and ≥90% of the calls on both alleles within the coding and non-coding regions of the genome. For the Higher Coverage Product, double the data is delivered, resulting in a minimum of 80X average coverage or approximately 240 Gb per sample.*

For more information about Complete Genomics, Inc., our technology, or sequencing services, please visit our website: www.completegenomics.com.

Works Cited

1. Exome sequencing identifies the cause of a mendelian disorder. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. 2009, Nature Genetics, Vol. 42, pp. 30-35.
2. Targeted capture and massively parallel sequencing of 12 human exomes. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. 2009, Nature, Vol. 461, pp. 272-6.
3. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Rowley, JD. 1973, Nature, Vol. 243, pp. 290–3.
4. Thalassemia intermedia: Moderate reduction of beta globin gene transcirptional activity by a novel mutation of the prximal CACC promoter element. Kulozik, A. 1991, Blood, pp. 2054-2058.
5. Interleukin-10 and transforming growth factor-beta promoter polymorphisms in allergies and asthma. Hobbs, K. 1998, American Journal of Respiratory and Critical Care Medicine, pp. 1958-1962.
6. Association between a sequence variant in the IL-4 gene promoter and FEV(1) in asthma. Burchard, E. 1999, American Journal of Respiratory and Critical Care Medicine, pp. 919-922.
7. Translational pathophysiology: a novel molecular mechanism of human disease. RC, Cazzola M and Skoda. 2000, Blood, Vol. 95, pp. 3280-8.
8. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Gonzalez, E. 2005, Science, pp. 1434-1440.
9. APP duplication is sufficient to cause early onset Alzheimer's dementia with cerebral amyloid angiopathy. Sleegers, K. 2006, Brain, pp. 2977-2983.
10. RAD51 135G-->C modifies breast cancer risk among BRCA2 mutation carriers: results from a combined analysis of 19 studies. Antoniou, A C, et al. 2007, American Journal of Human Genetics, Vol. 81, pp. 1186-200.
11. Mutations in the 3'-untranslated region of GATA4 as molecular hotspots for congenital heart disease (CHD). Reamon-Buettner SM, Cho SH, Borlak J. 2007, BMC Med Genet, Vol. 8, p. 38.
12. Meta-Analysis of the association of 4 angiotensinogen polymorphisms with essential hypertension: A

role beyond M235T? Pereira, T.V. 2008, Hypertension, pp. 1-6.
13. A functional polymorphism in the miR-146a gene and age of familial breast/ovarian cancer diagnosis. Shen J, Ambrosone CB, DiCioccio RA, Odunsi K, Lele SB, Zhao H. 2008, Carcinogenesis, Vol. 29, pp. 1963-6.
14. Rare structural variants found in attention-deficit hyperactivity disorder are preferentially associated with neurodevelopmental genes. Elia, J, et al. 2009, Molecular Psychiatry , Vol. doi: 10.1038/mp.2009.57.
15. Genome remodelling in a basal-like breast cancer metastasis and xenograft. L, Ding and Ellis MJ, Li S, Larson DE, Chen K, Wallis JW, Harris CC, et al. 2009, Nature, Vol. 464, pp. 999-1005.
16. Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. Visel A, Zhu Y, May D, Afzal V, Gong E, Attanasio C, Blow MJ, Cohen JC, Rubin EM, Pennacchio LA. 2010, Nature, Vol. 464, pp. 409-412.
17. The genetic signatures of noncoding RNAs. Mattick, JS. 2009, PLoS Genet., Vol. 5, p. e1000459. Epub 2009 Apr 24.
18. The HapMap and genome-wide association studies in diagnosis and therapy. Manolio TA, FS Collins. 2009, Annual Review of Medicine, Vol. 60, pp. 443-456.
19. Detection and mapping of amplified DNA sequences in breast cancer by comparative genomic hybridization. Kallioniemi, A, et al. 1994, Proc. Natl. Acad. Sci., Vol. 91, pp. 2156-2160.
20. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Gonzalez, E, et al. 2005, Science, Vol. 307, pp. 1434-1440.
21. APP duplication is sufficient to cause early onset Alzheimer's dementia with cerebral amyloid angiopathy. Sleegers K, Brouwers N, Gijselinck I, Theuns J, Goossens D, Wauters J, Del-Favero J, Cruts M, van Duijn CM, Van Broeckhoven C. 2006, Brain, Vol. 129, pp. 2977-2983.
22. Novel Exon of Mammalian ADAR2 Extends Open Reading Frame. Maas S, Gommans WM. 2009, PLOS One, Vol. 4, p. e4225.
23. Exome sequencing: A flash in the pan? Perkel, JM. 2010, Science Business Office Feature, Vol. DOI: 10.1126/science.opms.p1000042, pp. 149-251.
24. Initial Genome Sequencing and Analysis of Multiple Myeloma. Chapman, et al. 2011, Nature, Vol. 471 pp. 467-472.
25. Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, Shendure J, Drmanac R, Jorde LB, Hood L, Galas DJ. 2010, Science, Vol. DOI: 10.1126/science.1186802.

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, AN_EX-02