GOLDEN HELIX
Accelerating the Quest for Significance™

A Hitchhiker's Guide to Next-Generation Sequencing

by Gabe Rudy, VP of Product Development

If you have had any experience with Golden Helix, you know we are not a company to shy away from a challenge. We helped pioneer the uncharted territory of copy number analysis with our optimal segmenting algorithm, and we recently hand crafted a version that runs on graphical processing units that you can install in your desktop. So it's probably no surprise to you that the R&D team at Golden Helix has been keeping an eye on the developments of next-generation sequencing technologies. But what may have surprised you, as it certainly did us, was the speed in which these sequencing hardware platforms and services advanced. In a matter of a few short years, the price dropped and the accuracy improved to reach today's standards where acquiring whole exome or whole genome sequence data for samples is both affordable and accurate.

In a three-part series, I'm going to cover the evolution of sequencing technologies as a research tool, the bioinformatics of getting raw sequence data into something you can use, and finally the challenges and unmet needs Golden Helix sees in the sense-making of that processed sequence data.

To start with, let's look at the story of how we got to where we are today. If you ever wondered what's the difference between an Illumina HiSeq 2000 and a SOLiD 4hq, or why it seems that every six months the purported cost of whole genome sequencing is halved, then this story is for you.

Part 1: Evolution of sequencing technologies as a research tool

To start with, let's look at the story of how we got to where we are today. If you ever wondered what's the difference between an Illumina HiSeq 2000 and a SOLiD 4hq, or why it seems that every six months the purported cost of whole genome sequencing is halved, then this story is for you.

How We Got Here

As I'm sure Frederick Sanger could tell you, DNA Sequencing is nothing new. He received the Nobel Prize (or half of one) for his technology to determine the base sequence of nucleic acids in 1980. But it seemed that it wasn't until ten years later, with the audacious pursuit of sequencing the entire human genome, that the real driver of innovation took hold: competition.

With both the Human Genome Project and Celera racing to the finish line, improvements were made to the entire pipeline: from the wet work to the sequence detection hardware to the bioinformatics. What was originally expected to, optimistically, take fifteen years was out the door in ten. Hot on the heels of these original innovators was a new round of start-ups mirroring the dot-com era with their ruthless competitiveness and speed of innovation.

First out of the gate was the venerable 454 Life Sciences, later acquired by Roche, with their large-scale parallel pyrosequencing capable of long reads of 400 to 600 bases. These read lengths allowed for the technology to sequence novel organisms without a reference genome and were able to assemble a genome *de novo* with confidence. Although using the advances of the microprocessor industry in producing highly accurate small scale parallel components, the 454 system was still fairly expensive in acquiring a Gb (billion base-pairs) of sequence data.

Over a pint of beer at the local chemist's bar near Cambridge University, a couple of Brits decided they could do better. In their informal "Beer Summit",[1] they hashed out an idea for a sequencing chemistry that had the potential to scale to a very cheap and high-throughput sequencing solution. With the biochemistry skills of Shankar Balasubramanian and the lasers of David Klenerman, a massively parallel technique of reversible terminator-based sequencing was matured and commercialized under the funding of their start-up they called Solexa. With the promise of this potential, Solexa was purchased by U.S. based Illumina. The promise held out, and the 1G Genetic Analyzer released in 2006 could sequence a personal genome for about $100,000 in three months.

Coming to market at the same time, but seeming to have just missed the wave, was the Applied Biosystems (ABI) SOLiD system of parallel sequencing by stepwise ligation. Similar to the Solexa technology of creating extremely high throughput short reads cheaply, SOLiD has the added advantage of reading two bases at a time with a florescent label. Because a single base pair change

The Panton Arms where the "Beer Summit" took place

is reflected in two consecutive di-base measurements, this two-base encoding has inherent accuracy in detecting real single nucleotide variations versus potential sequencing errors. In a seemingly otherwise head-to-head competitive spec sheet with Illumina's Solexa technology, the momentum of the market went to the company that shipped working machines out to the eager sequencing centers first. That prize was won by Illumina by a margin of nearly a year.

Drivers of the Cost Curve

With the fierce competition to stay relevant in an exploding marketplace, the three "next-generation" sequencing platforms Roche 454, Illumina and ABI SOLiD vastly improved the throughput, read length, and quality of their sequencing hardware from their initial offerings. New companies such as Ion Torrent and Pacific Biosciences began to innovate their way into the market with new sequencing technology, each with their own unique advantages–Ion Torrent with simple chemistry and inexpensive hardware and Pacific Biosciences with extremely long reads and reduced sample prep. These "third-generation" sequencing companies have the potential to completely change the cost structure of sequencing by removing entire steps of chemistry or using alternatives to complex optical instruments. But despite the allure of the novel, Illumina has set the competitive bar very high with its recent release of the HiSeq 2000 in terms of throughput and cost per Gb of sequence data produced.

Alongside the technological drivers, there are two other factors I see that are making sequencing a viable and affordable research tool. First is the "democratization of sequencing" effect causing more sequencing machines to show up in smaller institutes, and second is the centralization and specialization found in the "sequencing as a service" business model. Let's explore both of these and how they may influence the researcher.

Democratization of Sequencing

While larger genome institutes are snatching up the latest high throughput sequencing machines, their existing machines are often being bought by smaller institutes quite happy with the throughput of the previous generation of hardware. Roche is even building a different class of product, the 454 Junior, for people wanting easier sample prep and less expensive machines without the need for as high of throughput per run. ABI is similarly releasing the SOLiD PI at a lower price point.

The target audience for these products are not those wanting to sequence whole genome or even whole exome human samples in their own lab, but rather people who need the flexibility and turn-around time of their own sequencing machine and want to investigate targeted regions or organisms that do not require gigabases of sequence data.

This is the democratization of sequencing: the power of being able to run experiments that treat sequencers as glorified microscopes or cameras, peering into the uncharted territory of certain bacteria or capturing the result of RNA expression experiments.

Sequencing as a Service

On the other hand, if you are interested in getting full genome or exome sequences of humans, you may be interested in another emerging trend: the centralization and growing of sequencing centers concerned with reducing costs by taking advantage of their focused expertise and economies of scale.

Despite what a sequencing platform vendor may tell you, every system comes with its quirks and issues. From the sample prep, to loading and unloading the sequencer, to monitoring and processing the bioinformatics pipeline, there are a lot of places for the process to go wrong and time and money to be wasted. But, on the other hand, if you take a page out of the book of highly efficient manufacturers of the 21st century, you can see amazing consistency and accuracy with a process of continuous improvement in place.

By focusing on just human whole genome or whole exome sample processing, you gain the expertise in providing the best quality data capable of the underlying technology. Complete Genomics has taken it one step further, building their sequencing service around their own unique sequencing chemistry and hardware to have complete control over every step of the process. This would be a good place to raise the flag that just having an outsourced service provider does not eliminate the potential for batch effects and poor study design to confound your downstream analysis. I will talk more about this in Part 2 with the discussion of the analysis of sequence data.

An often undervalued part of the production of quality sequence data is the bioinformatics pipeline that takes the raw reads from the machine and does the assembly or alignment to a reference genome and finally calls the variants or differences between the consensus sequence of the sample and reference genome. Single purpose software tools used in the pipeline have been rapidly developed by the sequencing centers themselves and other bioinformatics researchers. Though built on these open source tools for the most part, the expertise and compute power required to run this pipeline benefits greatly from the economies of scale and specialization of the sequence service providers.

Though not immediately obvious, we can now see that both the democratization of sequencing and the centralization of sequencing through service providers each fulfill their own complementary market needs. If your research is focused on human samples, and you want to do large runs covering whole exomes or genomes, it makes sense to consider the benefits of sequencing services both in terms of price and quality.

Sequencing as a Service Sounds Good. Now What?

So you've determined that sequencing as a service is good way to go and may be wondering what you should be requesting from your service provider? What data should you be keeping? What do you need to ensure you will have the data in a format ready for downstream analysis?

Although you may have the option to order just the raw reads from sequencing service providers, we have discussed some reasons why it makes sense to have them run their data processing pipeline on the reads so they can provide you with data ready for downstream analysis.

In fact, the sequence alignment algorithms such as BWA have matured to the point where it doesn't even make sense to keep the raw reads in their FASTQ format once alignment has been done. To allow for the easiest use of your data in downstream analysis, you should ask for your aligned data in the now standardized and efficient BAM file format and your variant calls in the near-standardized VCF format (although variant calls in any text format is usually sufficient).

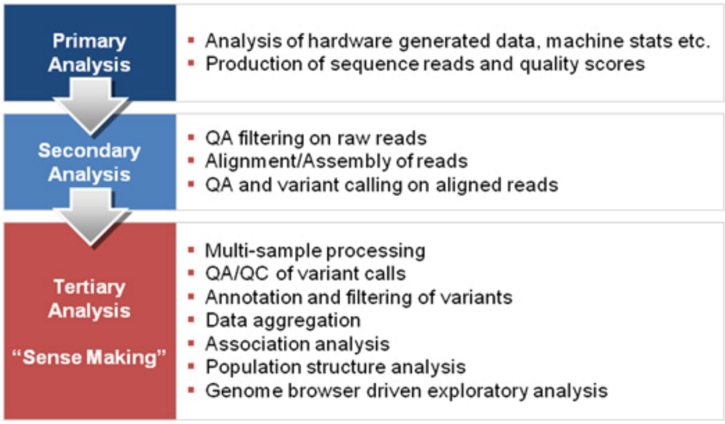
Part II: Getting raw sequence data into something you can use

When you think about the cost of doing genetic research, it's no secret that the complexity of bioinformatics has been making data analysis a larger and larger portion of the total cost of a given project or study. With next-gen sequencing data, this reality is rapidly setting in. In fact, if it hasn't already, it's been commonly suggested that the total cost of storing and analyzing sequence data will soon be greater than the cost of obtaining the raw data from sequencing machines.[2]

Next I plan to explore, in depth, what goes into the analysis of sequence data and why both the cost and complexity of the bioinformatics should not be ignored. Whether you plan to send samples to a sequencing-as-a-service center, or brave the challenges of sequencing samples yourself, this post will help distinguish the fundamental difference in analyses by their usage patterns and complexity. While some bioinformatics packages work well in a centralized, highly tuned and continuously improved pipeline, others fall into a long tail of valuable but currently isolated tools that allow you to gain insight and results from your sequence data.

Breaking Down of Sequence Analysis

The bioinformatics of sequence analysis ranges from instrument specific processing of data to the final aggregation of multiple samples into data mining and analysis tools. The software of sequence analysis can be categorized into the three stages of the data's lifecycle: primary, secondary, and tertiary analysis. I will first define these categories in more detail and then take a look at where the academic and commercial tools currently in the market fit into the categorization.



Primary Analysis

- Analysis of hardware generated data, machine stats etc.
- Production of sequence reads and quality scores

Secondary Analysis

- QA filtering on raw reads
- Alignment/Assembly of reads
- QA and variant calling on aligned reads

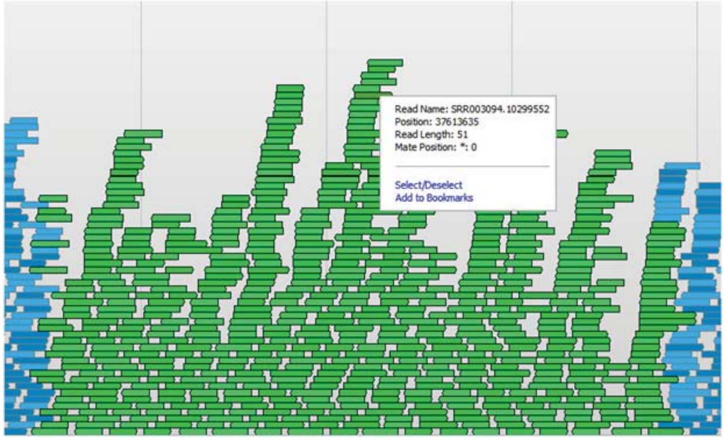
Tertiary Analysis

"Sense Making"

- Multi-sample processing
- QA/QC of variant calls
- Annotation and filtering of variants
- Data aggregation
- Association analysis
- Population structure analysis
- Genome browser driven exploratory analysis

Primary analysis can be defined as the machine specific steps needed to call base pairs and compute quality scores for those calls. This often results in a FASTQ file, which is just a combination of the sequence data as a string of A, C, G and T characters and an associated Phred quality score for each of those bases. This is the absolute bare minimum "raw" format you would ever expect to see for sequence data. While the first generation of high throughput sequencing machines, such as the Illumina G1, allowed for users to provide their own alternatives to the standard primary analysis solution, called "the Illumina pipeline", current generation machines often do this work on bundled computational hardware. The effective output of the machine is the result of the primary analysis. This output is ready for processing in a secondary analysis pipeline.

Because current sequencing technologies are generally based on the "shotgun" approach of chopping all the DNA up into smaller molecules and then generating what are referred to as "reads" of these small nucleotide sequences, it's left up to **secondary analysis** to reassemble these reads to get a representation of the underlying biology. Before this reassembly, the "raw" reads from the machine are often assessed and filtered for quality to produce the best results. Reassembly differs if the sequencing was done on an organism with a polished reference genome or if the genome is to be assembled from scratch, also referred to as *de novo* assembly. With a reference genome available, the process is much simpler, as the reads simply need to be aligned to the reference, often with some tolerance for a few base-pair errors in the sequences of the reference itself.



Read Name: SRR003094.10299552
Position: 37613635
Read Length: 51
Mate Position: *: 0

Pileup from the Savant Genome Browser

In both the case of *de novo* assembly or reference sequence alignment, you will be shooting for a desired average depth and coverage of sequence reads over the entire genome or targeted regions of interest. Depth is a measure of how many reads cover a given locus of the genome. If you were to pile up the reads to where they were assembled or mapped (pictured above), the depth would be the height of this pileup at each locus. For *de novo* assembly, a higher average depth is usually needed, so that large contigs can be formed that are then the building blocks for a draft genome. In the case of sequence alignment, higher average depth means more certainty in the "consensus" sequence of the sample and more accuracy in detecting variants from the reference.

The next step in the sequence analysis process is detection of variants. While more customizable, and sometimes considered part of tertiary analysis, variant calling lends itself to being pipelined in the same manner as secondary analysis. Variant calling is the process of accurately determining the variations (or differences) between a sample and the reference genome. These may be in the form of single nucleotide variants, smaller insertions or deletions (called indels), or larger structural variants of categorizations such as transversions, trans-locations, and copy number variants.

Tertiary analysis diverges into a spectrum of various study specific downstream investigations. Though the research or building of draft genomes for novel variants will have its own specialized tertiary analysis after *de novo* assembly, I will focus on the more mainstream practice of "resequencing" studies where sequence alignment to a reference genome was used. Out of the secondary analysis step of variant calling, you now have a more manageable set of differences between the sequenced samples and the reference, but there is still an enormous amount of data to make sense of. This is the realm of tertiary analysis.

The Resource Requirements of Primary and Secondary Analysis

As I described above, the amount of data produced by current generation, high throughput sequencing machines is enormous and continues to grow. Primary analysis solutions are largely provided by the platform providers as part of the machine's function. Whether you run their software on bundled compute resources or your own cluster, the analysis is designed to keep up with the throughput of the machine as it produces measurement information, such as images of the chemistry. In contrast, secondary analysis performs operations on the aggregated data from one or more runs of the machine. As a result, secondary analysis requires a significant amount of data storage and compute resources.

Usually quite resource intensive, secondary analysis performs a given set of algorithms and bioinformatics on a per-sample basis. This repeatable process can then be placed in an analytic pipeline that's entirely automated. Such an automated pipeline allows for even more resource utilization through scheduling. It also allows for a process of continuous improvement of the pipeline to be employed by monitoring quality metrics and tweaking or improving the algorithms and their parameters. Although these pipelines may be nothing more than an amalgamation of single purpose tools (as we will see in the next section), the economies of scale, not only in resource utilization but also in expertise and quality improvement, are realized by having secondary analysis centralized to a core lab or sequence-as-a-service provider.

Current Software Solutions

Over the past few years, the methods and algorithms designed for the primary and secondary analysis of high throughput sequence data have matured to satisfy the common needs of analysts. Most of these methods can be found in open source, single-purpose tools that are freely available to use and modify. In the case of *de novo* assembly, Velvet, written by Daniel Zerbino at EMBL-EBI uses de Bruijn graphs for genomic assembly and has set the bar for quality and accuracy. For sequence alignment to a reference genome, various Burrows-Wheeler Alignment based algorithms achieve a great balance of speed and accuracy. The original paper describing this use of the BWA algorithm[3] is implemented in the aptly named BWA package, a successor to the gold standard, but slower MAQ. Other BWA implementations include bowtie and SOAP2.

Seemingly simple in nature, a lot of work has gone into accurate variant detection for resequencing data. Single Nucleotide Variant (SNV) detection has gone through a few generations of algorithmic improvements, with implementations such as SOAPsnp and GATK representing the current lineup. Indel detection, being a newer pursuit, has not had the time for a real winner to emerge. Complete Genomics has shown that a more holistic approach may achieve better results than current best practices. Finally, copy number and other structural variant detection are still being actively developed as new methods. These types of analyses require new ways to handle the unique properties of sequence data.

Though the above methods are often implemented in single-purpose packages, they can be strung together to compose a state-of-the-art secondary analysis pipeline. This can be done manually with a considerable amount of IT customization, or through a server-oriented connector package such as Galaxy from Penn State. Commercial offerings in this area, such as CLC bio, attempt to provide an all-in-one solution for secondary analysis with their own proprietary tweaks on the standard algorithms, and a GUI for running a single sample at a time through each step.

Tertiary Analysis Requirements

After getting sequence sample data to the stage of called variants, the real work begins in making sense of the data in the context of your study. At this point, multiple samples need to be brought together, along with phenotype and other experimental data. While primary and secondary analysis can be automated and centralized, here in the "sense making" stage, there are a plethora of potential different analysis techniques and exploratory paths. Fortunately, even whole genome sequence data processed to the level of variant calls is manageable on a researcher's modern desktop. Whole exome or targeted resequencing is even more manageable.

Unlike microarray data, where the probes are usually carefully selected to be loci of potential interest, sequence data naturally contains variants regardless of their loci or functional status. As illustrated in the categorization of operations above, the common starting point in tertiary analysis then, is to use all the annotation, functional prediction, and population frequency data available to sort through this unfiltered list. Luckily, the genetics community has really stepped up to the plate in both funding and sharing public data repositories. Population cataloging efforts such as HapMap and the 1000 Genomes Project allow for studies to be able to determine the relative frequencies of variants in their samples compared to common populations. Once variants have been filtered and ranked, new methods are gaining momentum to handle the analysis of rare variant burden as a model of disease association, as well as other methods specific to the filtering and studying of sequence data.

Common also in the exploratory analysis of sequence data, is the use of a genome browser to give historical and biological context to your data. Like these other filtering and analysis methods, a genome browser will utilize the many repositories of genome annotations and public data. You may want to see if a given locus has had GWAS results published in the OMIM database, compare variants against repositories of known SNPs, or just visualize the genes from various gene catalogs for a given region.

As we can now see, there's quite a bit to consider in terms of next-gen sequencing analysis, from producing raw reads to calling variants to actually making sense of the data, all of which can add to the complexity and cost of a sequencing study. As the field standardizes on best practices and methods for primary and secondary analysis, we can be more certain about planning for these steps and even include them in the "cost of production" of sequence data from either your own core lab or a "sequencing-as-a-service" partner.

Tertiary analysis is still in its infancy but as we experienced at IGES and GAW this year (see our report), promising new methods and workflows are beginning to emerge. Over the coming year, we are probably as excited as you are to see breakthrough discoveries being made with this technology.

Part III: "Making Sense" of all that data

The advances in DNA sequencing are another magnificent technological revolution that we're all excited to be a part of. Similar to how the technology of microprocessors enabled the personalization of computers, or how the new paradigms of web 2.0 redefined how we use the internet, high-throughput sequencing machines are defining and driving a new era of biology.

Biologists, geneticists, clinicians, and pretty much any researcher with questions about our genetic code can now more affordably and capably than ever get DNA samples sequenced and processed in their pursuit for answers. Yes, this new-found technology produces unwieldy outputs of data. But thankfully, as raw data is processed down to just the differences between genomes, we are dealing with rich and accurate information that can easily be handled by researchers on their own terms.

In this final section, I'm going to explore in more depth the workflows of tertiary analysis, focusing primarily on genotypic variants. Over the last three to four years, the scientific community has proposed a set of methods and tools for us to review as we explore the current landscape of solutions. So let's examine the current state of sense making, how the field is progressing, and the challenges that lay ahead.

Motivation Behind Next-Generation Sequencing

To properly discuss the methods and techniques of sequence analysis, we need to step back and understand the motivation behind using sequencing to search for genetic effects. We have, for the past five years, been a very productive scientific community investigating the common disease-common variant hypothesis. With significant Genome-Wide Association Study (GWAS) findings for hundreds of phenotype and disease-risk traits, we should consider this effort well spent.

Genotyping microarrays have been the workhorse of GWAS study designs as they cheaply enable the assaying of common variants on thousands, and even tens of thousands, of samples. Of course, the more we study human populations the harder time we have classifying one universal definition of what makes a common variant "common." Indeed, due to the fact that all shared variants were inherited from somewhere, even "rare" variants are not rare in some geographical locations or sub-populations. And so it follows that with better cataloging of population variants, more mileage can be had in the future with the existing GWAS study design. Even more interestingly, next-generation microarrays will be able to assay around 5 million variants, as opposed to 1-2 million with the Affy SNP 6 or the Illumina HumanOmni1. Besides trying to assay all common variants with Minor Allele Frequencies down to 1% on diverse populations, new microarrays can be based on variant catalogs for specific populations or even biologically relevant panels such as the "MetaboChip" from Illumina.

But even with a new generation of microarrays around the corner, the track record for the GWAS study design has not been perfect, and the limited ability to assay only sets of known variants is being blamed as the culprit.[4] The common variants identified in association studies thus far have only accounted for a small portion of the heritability of complex traits studied. If complex diseases are driven by susceptibility alleles that are ancient common polymorphisms, as the common diseases-common variant hypothesis proposes, we should be seeing a larger portion of the heritability of these traits being accounted for by these associated polymorphisms.

The alternative hypothesis that must now be considered is that complex diseases are actually driven by heterogeneous collections of rare and more recent mutations, as with most Mendelian diseases! So how does one go about studying low-frequency or rare variants and how they contribute to genetic risk? Well, nothing beats the fidelity and comprehensiveness of whole-genome sequencing. But given that the cost of whole-genome sequencing may still be prohibitively expensive (though its rapidly dropping) for larger numbers of samples, whole exome sequencing is a good compromise.

Clearly getting the full set of variants provided by sequencing is the ideal tool, but microarrays should not be totally discounted. As our CEO, Dr. Christophe Lambert, explains in his post Missing Heritability and the Future of GWAS, there can be a productive pairing of next-generation sequencing with next-generation custom microarrays. Simply put, with the impending ability to customize microarrays, a small scale preliminary sequencing of affected individuals can be done to provide an enriched panel of rare and common variants specific to the disease population. The custom micorarrays could then be used affordably on a large set of cases and controls.

But no matter how you acquire your low frequency and rare variant data, you will quickly realize that the set of tools from the traditional GWAS study toolbox are inadequate to properly ascertain causal alleles and regions.

Sequencing and Study Design

To understand how to study sequence data, we have to better understand the nature of rare and low frequency variants. Out of the secondary analysis pipeline, you can expect an average of 20K single nucleotide variants to be detected for a single exome. That relatively small number may make the analysis of those variants seem deceptively easy. But let's try to place those 20K variants in context with some figures from the 1000 Genomes Project[5] on the cataloging of variants and an article by the CDC[6] on the expected functional implications of those variants.

- 25 million unique variant sites have been identified on a set of 629 samples
- Of those 25 million, 15 million have frequencies below 2% in the population
- 7.9 million variants are in dbSNP 129 (closest thing to a database for common variants)
- You can expect roughly 4 million variants for a single sample at the whole-genome scale
- Around 20k will be in coding regions (exome)
- 250-300 will be loss-of-function variants (biologically damaging) in annotated genes
- 50-100 will be SNPs previously implicated in diseases

With these numbers to set perspective, the scale of the problem becomes apparent. How do you use the abundance of public data, as well as good study design, to filter down to causal regions of interest? How do you distinguish the functions of variants? How do you measure the combined effect of variants and their correlation with phenotypes?

First and foremost you need to start with good study design. In my previous post I advocated taking advantage of the centralized expertise and economies of scale that sequencing-as-a-service providers deliver for the sequencing and variant calling of your samples. While this still holds, a word of caution: having dealt with the confounding effects of poor study design in the analysis of many GWAS studies, Dr. Lambert has made a plea to take batch effects seriously when planning the logistics of assaying your samples. There is no reason to believe sequencing studies would be immune to the same effects as GWAS studies. So be sure to work with your genomic center or sequencing service provider to ensure proper randomization of cases, controls, and family members across runs. Also technical replicates should be considered to measure concordance of measurements across batches.

Besides proper randomization, there are other design level techniques to ensure the highest quality data for your study. Not surprisingly, if you have the luxury of including multiple related samples in your study, you can use the family structure not only to investigate inheritance patterns but to do extra quality checks on your variant calls themselves. Another neat trick is to pool the DNA of your cases and controls into groups, which you can then deeply sequence with over 100x coverage to get a highly accurate set of variants for your various groups.

So now, with proper study design and quality control checks in place, how does one go about finding causal variants?

Analysis of Rare Variants

With the hypothesis of susceptibility alleles being a heterogeneous collection of rare and low frequency alleles, the first step in the analysis of sequence data is to categorize the relative frequency of your sequenced variants. If your sample set is enriched for the disease or trait of interest or you simply don't have enough samples to attain accurate frequencies, you will want to refer to an external catalog such as dbSNP 129 or maybe the catalog of variants from the 1000 Genomes project. Note that dbSNP 129 is often considered the last "clean" dbSNP build without many rare and unconfirmed variants from 1000 Genomes and other large scale projects being added.

With a classification of your variants, the next step is to search for regions where the heterogenous burden of rare, low frequency and common variants is strongly correlated with the trait under study. Traditional association techniques used in GWAS studies do not have the power to detect associations with these rare variants individually or provide tools for measuring their compound effect. To this end, there has been a field-wide development of analytic approaches for testing disease association with sequence data.

The first round of methods focused on finding association regionally, based on accounting for the presence of at least one rare variant or a counting of the number of rare variants. Cohort Allelic Sum Test (CAST; Cohen et al.)[7] was developed in 2004 to simply count and compare the number of individuals with one or more mutations in a region or gene between affected and unaffected groups. In 2008, Li and Leal published the Combined Multivariate and Collapsing (CMC)[8] method which similarly collapsed the rare variants into a single covariate that is analyzed along with all the common variants in the region using a multivariate analysis. Finally, Madsen and Browning published a method[9] in 2008 that described a weighted-sum approach that tests group-wise association with disease while allowing for rare variants to have more weight than common variants.

The second round of rare variant analysis methods attempt to address a couple confounding issues. First is that not all variants, regardless of their rarity, are created equal. In fact, it's possible for there to be both protective, neutral, and damaging variants all in the same gene or region being tested. While neutral variants generally just reduce power, they are fairly easy to filter out (see next section). The presence of both protective and damaging variants in the same region without knowing which is which is especially confounding. Using a very neat set of mathematical functions, called c(α), you can test for the mixture of neutral, risk, and protective variants. Benjamin Neale has presented on this approach at ASHG 2010 and other venues, but details of the method have not yet been published.

Suzanne Leal, the author behind the CMC method, has proposed an enhanced method with Dajiang Liu, which they call Kernel Based Adaptive Cluster (KBAC).[10] This new method can account for other more general confounders such as age, sex, and population substructure in its association testing.

Almost all of these methods benefit from having another dimension of variant classification other than assessing its population frequency, and that's inferring or predicting its functionality.

Getting More Clarity with Functional Prediction

Although we expect rare variants to be enriched for functional alleles (given the view that functional allelic variants are subject to purifying selection pressure), properly understanding and accounting for the functional significance or insignificance of variants improves the understanding of the contribution of those variants. Variants come in the form of single nucleotide variants (SNV), and insertions and deletions (often abbreviated as "indels"). When SNVs occur in coding regions, they are either synonymous (in that they code the same amino acid) or non-synonymous. Non-synonymous single base pair substitutions can then be classified into missense mutations where one amino acid is replaced with another, nonsense mutations where an acid codon is replaced with a stop codon, and splice site mutations where signals for exon-intron splicing are created or destroyed. Small insertions or deletions may introduce frameshift errors by adding or removing nucleotides that are not a multiple of three, hence entirely changing the downstream amino acid coding.

Although it seems these types of mutations are easy to classify (besides a mutation as being missense), what we really care about is whether the amino acid substitution affects the protein function. Protein function prediction is a world of complexity itself and quickly leads to understanding entire organisms and their networks. Programs such as SIFT and PolyPhen2 are able make a decent prediction of how likely a mutation is damaged by looking at things like how a protein sequence has been conserved through evolution as a proxy to its functional importance.

If you're trying to get a perfect picture of the functional importance of all your exonic variants, you'll see it's already tough. When you throw into the mix splicing mutations that alter splice sites, affect splicing enhancers, or activate cryptic splice sites, things get a bit tougher. When you try to account for all potential gene regulation sites upstream and downstream, it gets near impossible. Although many of these things can be worked out through carefully run lab experiments on a gene-by-gene basis, possibly the best we can do at a genome-wide scan level is to keep variants within reasonable distances of genes and close to splice sites in our analysis as we filter.

Along with the variant calls themselves, a secondary analysis pipeline usually provides quality scores and the depth of the pileup where a variant was called. You can use this information to either pre-filter variants to a higher quality set or investigate your variants in your region of interest to ensure its biological validity. At that level of individually confirming variants, you may want to visualize the aligned sequences themselves around a variant. If things seem a bit wonky, doing a *de novo* assembly of that local set of reads may clean up assembly.

Other Analyses on Sequence Data

Besides trying to nail down the elusive genetic components of complex diseases through variant analysis, DNA sequencing is an excellent tool for other study types. In the case of rare penetrant Mendelian diseases, sequencing whole genomes of affected individuals and their families can lead to near instant success in discovering the causal mutations.[11] The analysis is not one of association, but really of data-mining. Just doing set-filters between the variants of different samples helps illuminate novel or compound heterozygous mutations in just a few, easy to follow-up on regions.

Although small SNV and indel variants are the current focus for association studies with sequence data, larger structural variants such as copy number variants (CNV) and other inherited genomic features such as runs of homozygosity (ROH) can also be studied with whole-genome and sometimes even whole exome sequence data. Many methods have been proposed to tackle these tasks, and as sequencing becomes more affordable, we expect further focus on maturing the tools to support these studies.

Finally, outside common and rare Mendelian diseases, the quest for understanding the mutations driving cancer are always in need of more precise and comprehensive measurements of the genome. In the variant-abundant world of cancers, nothing beats having a whole-genome scan of your cancer and normal samples for comparative analysis.

Current and Future Solutions

As we have seen in our exploration, tertiary analysis is where things really open up to the inquisitive process researchers apply in discovering regions of importance in the vast genomes of their samples. For a lot of the filtering steps I've described here, programs such as ANNOVAR can take large lists of variants and help narrow them down. Although many of the collapsing and testing methods designed for this field have been around for years, it's hard to point to any package that allows easy use of them. Similarly, functional prediction and investigating variants in the context of a genome browser all require variant-by-variant queries to various websites.

Golden Helix hopes to open up this realm of analysis to more researchers with integrated and powerfully extensible tools that grow to reflect the best of breed methods and workflows for tertiary analysis of sequence data. In the release of SNP and Variation Suite version 7.4, we kick off that focus with support of many of the steps described here, and more to come. We hope you will be one of our collaborators in this exploration!

References

1 Davies, K. (September 28, 2010) The Solexa Story. *Bio-IT World*. http://www.bio-itworld.com/2010/issues/sept-oct/solexa.html

2 What to Do with All That Data? (October 7 , 2010) *Genome Web Daily Scan*. http://www.genomeweb.com/blog/what-do-all-data

3 Li, H and Durbin, R (May 18, 2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 2009 Jul 15;25(14):1754-60. http://www.ncbi.nlm.nih.gov/pubmed/19451168

4 Hayden, E (September 22, 2009) Genomics shifts focus to rare diseases. *Nature News*, 461, 459, doi:10.1038/461458a. http://www.nature.com/news/2009/090922/full/461458a.html

5 1000 Genomes Project. Retrieved January 18, 2011. http://www.1000genomes.org

6 Khoury, M (June 2010) Dealing With the Evidence Dilemma in Genomics and Personalized Medicine. *Clinical Pharmacology & Therapeutics*, 87, 635-638, doi:10.1038/clpt.2010.4. http://www.nature.com/clpt/journal/v87/n6/full/clpt20104a.html

7 Cohen, J et al (August 6, 2004) Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, 305(5685):869-72. http://www.ncbi.nlm.nih.gov/pubmed/15297675

8 Li, B and Leal, S (August 7, 2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal of Human Genetics*, 83(3):311-21. http://www.ncbi.nlm.nih.gov/pubmed/18691683

9 Madsen, B and Browning, S (February 13, 2009) A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genetics*, 5(2): e1000384. doi:10.1371/journal.pgen.1000384. http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.1000384

10 Liu, D and Leal, S (October 14, 2010) A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. *PLoS Genetics*, 6(10): e1001156. doi:10.1371/journal.pgen.1001156. http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.1001156

11 Roach, J et al (April 30, 2010) Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science*, 328(5978):636-639, doi:10.1126/science.1186802. http://www.sciencemag.org/content/328/5978/636.abstract

About Gabe Rudy

Gabe Rudy is GHI's Vice President of Product Development. Gabe is continually scouting the fast changing fields of both genetics analysis and software development. This in turn leads to curating features and infrastructure improvements so that the GHI development team can regularly provide our valued customers with top-of-the-line software releases to further accelerate their research. Gabe joined Golden Helix while still an undergrad student at MSU-Bozeman and then ventured to the University of Utah where he received his masters degree in Computer Science.

About Golden Helix

Founded in 1998, Golden Helix is known for helping genetic research groups working with large-scale DNA-sequencing or microarray data overcome the frustration and challenges of bioinformatic roadblocks: delayed projects, lack of quality findings, and low productivity. By empowering researchers with highly effective software tools, world-class support, and an array of complementary analytic services, we refute the notion that analysis has to be difficult or time consuming. Golden Helix's flagship software product, SNP & Variation Suite (SVS), is an integrated collection of powerful data management, quality assurance, visualization, and tertiary analysis tools for genetic data. SVS is delivered in a user-friendly, scalable platform and is supported by a team of highly trained bioinformaticians, statistical geneticists, and computer scientists that together make advanced statistical and bioinformatic methods accessible to scientists of all levels.