


Clinical Next-Gen Sequencing Analysis



Dr. Andreas Scherer
Golden Helix, Inc.

Copyright © 2014 Andreas Scherer

All rights reserved. No part of this book may be reproduced in any form or by any electronic or mechanical means - except in the case of brief quotations embodied in articles or reviews - without written permission from its publisher.

Copyright © 2014 Andreas Scherer

All rights reserved.

ISBN: 978-0-9908886-1-1

Table of Contents

Preface 5

Chapters

1. Adoption of Genetic Services and Technology... 7

2. Gene Panels..... 10

3. Whole Exome and Genome Tests..... 14

4. Summary..... 17

Preface

We live in interesting times. This generation of scientists, clinicians and bioinformaticians will elevate the standards for diagnosis, prediction and care, ultimately improving patient outcome for millions of people. Golden Helix will support this process by building products such as VarSeq that help to welcome in the next frontier of medical care.

In this ebook, we cover our best understanding on how this revolution in clinical diagnostics will unfold. It reviews critical success factors for the clinical adoption of next-gen sequencing technology. In addition, we discuss workflow design principles for gene panels, whole exome, and whole genome analysis. All of these are important to successfully leveraging VarSeq for clinical purposes.

A lot of people at Golden Helix have contributed to this ebook. It would have been impossible to write this without the ingenious work of our product developers. Specifically, I'd like to thank Gabe Rudy, Andrew Jesaitis, Dr. Bryce Christensen, Ashley Hintz, Greta Linse Peterson, Cheryl Rogers, Chelsey Clayton and Alyssa Burzynski for their invaluable contributions.

Andreas Scherer
September, 2014
Bozeman, Montana

Chapter: 1

Adoption of Genetic Services and Technology

The adoption of genetic services is key to our ability to provide personalized medicine in the future. The goal is to better diagnose diseases, predict their outcome, and choose the best possible care option for a patient. We still have a long way to go to achieve this goal. While there is agreement about the ultimate goal, there is still a lot of uncertainty about the timing of the adoption. Essentially, it is a widely debated topic among experts. Here is what we know. There are three phases of the adoption (see also Fig. 1):

- **Early Stage:** Strong focus on science and research; understanding underlying genetic mechanisms and pathways
- **Moderate Adoption:** Utilizing available results in the clinic on a selected basis. The science and clinical communities focus on increasing the number of therapeutic areas as well creating infrastructure to improve scale.
- **High Adoption:** At this point, genetic service is part of standard care.

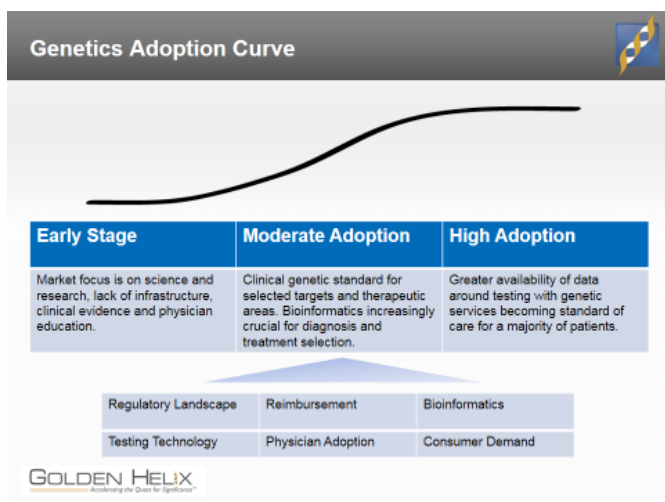


Figure 1: Genetic Services Adoption (Leslie, T. et al (2012))

There are a number of success factors that are key to drive adoption in the phase of moderate adoption (see also Leslie, T. et al (2012):

- **Regulatory Environment:** This is a wide open field. The FDA is paying increasing attention to how genetic information about individuals is handled. Very publicly, it engaged with 23andMe to review their processes and procedures. While this gives the adoption process a pause in the short-term, it might mean that in the mid- and long-run, we can hope for a consolidated governance body protecting patients while enabling clinicians to make informed decisions.
- **Testing Technology:** There is already clinical experience leveraging genetic markers via gene panels. Many experts believe that the usage of Next-Gen sequencing data, including whole exomes and genomes, will ultimately give us the insights needed to diagnose or predict diseases and select treatment options at the highest level possible. The two remaining chapters of this e-book will discuss this technical topic in further detail.
- **Reimbursement:** The adoption of these tests in the clinic hinges on the ability for doctors to recover the expense associated with genetic tests. For this to occur, the quality of the results of these tests has to be very high and their clinical utility proven. Also, price points have to become reasonable for payers and insurance companies to accept these tests as part of standard care.
- **Physician Education and Acceptance:** We are facing the need to educate a wide range of health care specialists involved in designing, conducting, interpreting and utilizing genetic tests: translational researchers, pathologists, geneticists, genetic counselors, biostatisticians, etc. This also includes the personnel supporting these professions, such as nurses and medical assistants.
- **Bioinformatics Capability:** There is a plethora of new tests in the development stage that require a massive computing resource to deal with sheer amount of data. The storage requirements for a whole genome, depending on the coverage, ranges between 100 and 200 GB for a single person. Added to that are the results of the variant analysis and other datasets generated as part of analysis. For a wider scale usage, these datasets and their interpretation have to be part of the patient record. This requires investment in an infrastructure to provide, gather, store, and research and clinically interpret this data on a large scale that is not currently in place.

- **Patient Demand:** Any test requires patient consent. Because of this, a more global adoption of genetic tests is also dependent on patients agreeing to use their DNA for this purpose. So far, the early successes of direct consumer companies such as 23andMe seems indicative of a general positive attitude of patients towards utilizing their DNA information for clinical purposes. Ease of use and affordability will be crucial. Also, as with any new technology being rolled out on a global basis, there are concerns that need to be publicly discussed. Data security or the potential use of this data for purposes outside of the specific clinical use by insurers or other third parties must be considered.

The genetic testing technology and infrastructure is evolving quickly. There are a few areas where adoption is expected first with significant testing volume:

- Oncology
- Inflammation
- Rare diseases
- Pediatrics

Beyond these areas, there is strong potential in areas such as obesity, diabetes, and cardiac disorders. In addition to this, there is a huge uptick in adoption expected in the area of pharmacogenomics to determine safety, efficacy and cost of care. New applications of genetic testing will result in changes to current care teams and processes. It will reshape how pathologists, oncologists, geneticists, genetic counselors, biostatisticians and bioinformaticians work together.

Chapter: 2

Gene Panels

Gene Panels are well established methods diagnosing a large variety of disorders based on protocols and library preparation kits from a number of manufacturers such as Illumina, Life Technologies and Qiagen. Gene Panels based on modern sequencing platforms such as Ion Torrent PGM or Illumina MiSeq have some distinct advantages over whole genome or exome sequencing workflows.

1. **Coverage:** They allow a much higher coverage of the targeted region (>1000x). Analyzing whole genomes or exomes usually only allows coverages in a much lower range (<100x).
2. **Frequency of variants:** They are much more effective at detecting somatic variants such as BRAF V600E which occur at much lower allelic percentages due to the sample being only partially made up of tumor cells. Whole exome or genome analysis is more suitable for germline mutations that are present in virtually every cell.
3. **Complexity of bioinformatics:** Gene panels are of moderate complexity. The data volume is very manageable. Even if we look at multi-gene panels, we are typically dealing with thousands of variants across a manageable number of genes (e.g. up to 100). The bioinformatics can be done on a modest-sized server. Whole exome and genome workflows produce a much larger amount of data in comparison.
4. **Clinical relevance:** Due to their targeted nature, gene panels are designed to discover variants of clinical relevance. The practitioner is more likely to be presented with actionable data compared to whole exome and genome workflows that are more likely to produce a higher number of variants of uncertain clinical significance.

Carrier screening	Counsyl is now using gene panels rather than genotyping for carrier screens; MiSeq DX is FDA approved as a device for carrier screening for Cystic Fibrosis.
Targeted Cancer Therapy	Ion Torrent Cancer Hotspot Panel v2
Cancer Risk Assessment	BRCA 1 and 2 cancer panels (and others) (Myriad, Invitae, GeneDx)
Cancer Risk Assessment Neonatal respiratory distress	Ambrygen
Multi-gene panels	Invitae's multi-gene panel includes hundreds of genes and implicated in a multitude of diseases; Illumina markets a 46-gene panel for heritable cardiomyopathy; see also Kuorian, A. W. et al (2014)

Figure 2: Examples of Gene Panels

Building a state of the art gene panel

We see an explosion of gene panels coming out of research organizations. The advances in bioinformatics allow core labs to produce targeted panels that are specific to the needs of their clients. How do we build these gene panels? Here are some key design factors:

1. **Simplicity:** The complexity of designing and utilizing gene panels in the clinic is not due to the size of data for an individual sample. It's all about being able to handle a high level of throughput of samples. Labs receive an ever increasing number of samples. Hence, the processing of those datasets must be
 - a. Efficient
 - b. Accurate
 - c. Reproducible
2. **Verification is key:** In the clinic it's not about the discovery of a new variant which is key on the research side. For the clinician, it is just as important to be able to determine that a genomic region doesn't have a mutation as it is to determine that such a variant exists. Clinicians need certainty on both sides of the equation.

A five step process toward sound gene panel analysis

Let's take a look how a generic workflow functions. Let's assume we have a VCF file from our variant caller such as GATK UnifiedGenotyper, FreeBayes, or SomaticSniper containing the variants related to a list of genes.

Step 1: Verify variants are of "high quality": For example there is the QUAL field in the VCF field expressing the assessment of the data quality as a direct output of the sequencing process. This number is based on the Phred quality score. It essentially tells us the probability that a REF/ALT polymorphism exists at a particular site. The Phred quality scores are logarithmically linked to error probabilities (see Fig. 3). In general a quality score of +20 is deemed to be proficient.

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

Figure 3: Phred scores

Step 2: Verify sufficient coverage over amplicon: At this point, we want to know the average read depth over an amplicon which will vary based on the library prep kit. In practice this can vary from 20x to +1500x. If the region of interest falls in an amplicon with insufficient coverage, then it's best to resequence the sample. While this adds costs and takes time, it's a crucial quality control step ensuring that the clinician is basing her best judgment on reliable data.

Step 3: Remove variants that are present outside of regions sequenced: Without trying to state the obvious, there shouldn't be any variants outside of the sequenced regions in our VCF file to begin with. If they exist, this is due to errors of the sequencing or bioinformatics process. In any case, they are not useful for the interpretation of a test and should be removed.

Step 4: Verify presence or absence of variants at all "hotspots": Based on the target of the test there is a list of hotspots. Those are locations of known variants of interest and must be called as being present or absent in a sample. For example p53 is the most frequently altered tumor suppressor gene in a wide spectrum of human tumors; see Monti, P. et al (2003).

Step 5: Annotate Variants against databases of interest: After Step 4, we have essentially a short list of variants that are “suspicious” within the regions of interest. At this point we need to enrich the variant data set with clinical relevant content. The research community is maintaining and updating major public database such as:

- COSMIC
- OMIM
- dbSNP

In this step we annotate against those or similar databases to enrich our variant list with clinical relevant actionable information. On a related note, beyond the publicly available annotation sources, major research institutes and hospitals also maintain their own databases. So, clinicians not only annotate against public datasets, but also against their own cases observed within their institutions as well as commercially available datasets.

Cancer Hotspot Panel

Cancer Hotspot Panel

Sample 1

Cancer Hotspot Panel

Passes Variant Caller Quality

Alternate Allele Frequency

Read Depth at Variant Site

In Cosmic

CHPv2 Amplicon Match

CHPv2 Hotspot Match

1,649

1,626

18

12

2

2

(18 Variants) Alternate Allele Frequency

Alt Allele Freq

Vars: 18

Variant Site	Chr Pos	Genotypes	Zygosity	RefSeq	Gene Name	Cosmic	Amplicon	Hotspot
						Match?	Match?	Match?
2:209113192	A	G	Heterozygous	IDH1	False	True	True	True
2:212812097	C	T	Heterozygous	ERBB4	False	True	False	True
3:178917005	A	G	Heterozygous	PIK3CA	False	True	False	False
3:178927410	A	G	Heterozygous	PIK3CA	True	True	False	False
4:1807894	A	A	Homozygous Variant	FGFR3	False	True	False	False
4:55141055	G	G	Homozygous Variant	PDCRFA	True	True	True	True
4:55152040	C	T	Heterozygous	PDCRFA	True	True	True	True
4:55946354	G	T	Heterozygous	KDR	False	True	False	True
5:112175770	A	G	Heterozygous	APC	False	True	True	True
5:149433596	G	T	Heterozygous	CSF1R	False	True	False	True
5:149433597	A	G	Heterozygous	CSF1R	False	True	False	True
7:55249063	A	G	Heterozygous	EGFR,EGFR...	True	True	True	False
7:116340269	C	T	Heterozygous	MET	False	True	False	True
9:139399409	CAC,CAC		Reference	NOTCH1	True	True	True	True
10:43613843	T	T	Homozygous Variant	RET	False	True	True	False

Zoom (2: 189,519,822, 0) 1 - Y 3.1 Cbp

Figure 4: Filtering data from an Ion Torrent Cancer Hotspot gene panel using VarSeq

Chapter: 3

Whole Exome and Genome Tests

As discussed, Gene Panels are testing the “usual suspects”. They are helpful to diagnose well understood disorders and their genetic markers. What do we do when all known options have been exhausted? Here, the use of whole exome or whole genome sequencing comes into play. This type of test lets us find out what we don’t know yet. It’s currently designed to support the discovery in a clinical setting. It is openly debated among experts, whether or not whole exome and genome tests will eventually become the standard form of genomic testing. The discussion is shaped by the fact that sequencing costs have dropped and efficiency has improved significantly over the last decade. This trend is expected to continue in the foreseeable future. At some point the economics for price efficient whole exome and genome tests seem unbeatable. Let’s assume a patient is tested for various cancers. A clinician might order the standard tests for EGFR, KRAS, BRAF, and ALK just to name a few. In the future, a whole exome or genome tests might give him all the results she needs with one test, assuming that the required coverage can be achieved at a competitive price point. However, as it stands today, whole exome and genome sequencing are most effective for the discovery of rare variants underlying Mendelian disorders (see Biesecker, L. G. (2010) and Yang, Y. et al (2013)).

Guiding design principles

1. **Filter Design:** When setting up filters, some aspects have to be taken into account. First, as filters are made more specific, the chances increase of eliminating high-impact variants. On the flip side, without sufficient filtering power, important variants are lost in a sea of noise.
2. **Speed:** The fine tuning of filters is an iterative process. It’s important that the underlying tool is highly responsive to user input.
3. **High quality annotations:** Faulty or incomplete annotation sources compromise the filtering process. We at Golden Helix spend significant resources to prevent this from happening.
4. **Handling of inheritance patterns:** Specifically, Mendelian disorders follow a particular inheritance patterns. The underlying tool needs to be able to segregate data into easily understood categories.

The workflow for whole exome and genome tests are essentially conducted in three major steps.

Step 1: Apply filters common to all filtering strategies

- a) Quality: Like with gene panels, we need to test for quality of the input data.
- b) Frequency: We expect that disease-causing variants will be rare; at least they will be no more common than the disease of interest. So we need to confirm that we are only considering rare variants in most cases. For that, a cross reference with the 6500 Exomes database (<http://evs.gs.washington.edu/EVS/>) comes in handy.
- c) Variant Classification: Disease causing variants usually have a coding (including splicing) ontology (e.g. nonsynonymous SNPs, frameshift variants, splicing variants).
- d) Phenotype Association: Finally we want to make sure that the variant is located in genes associated with a plausible phenotype.

Step 2: Filters for inheritance pattern: In this step we are categorizing variants according to all relevant inheritance patterns.

- a) De Novo: Child is heterozygous, parents are reference homozygous
- b) Maternal heterozygous: Child and mother are heterozygous
- c) Paternal heterozygous: Child and father are heterozygous
- d) Compound heterozygous: This is a bit more complex. In this case we are looking for the “two-hit” model of recessive alleles in a single gene. That is, the child has a compound heterozygous mutation if it has two or more recessive mutations for the same gene, where at least one mutation was inherited from each parent.
- e) Homozygous: Child is homozygous

Step 3: Additional Annotations: In the final step we will enrich the identified variants with datasets that help the clinician to look for clinically relevant data. Here are a few examples:

- a) 1000 Genomes Variant Frequencies
- b) dbSNP
- c) OMIM
- d) dbNSFP
- e) Clinvar

As was the case for gene panels, clinicians want to have the ability to annotate against custom databases that, for example, capture the cases in their own institution.

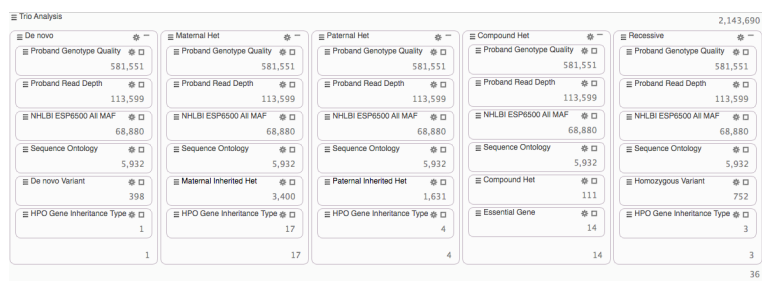


Figure 5: Trio analysis of whole exome data using VarSeq

Chapter: 4

Summary

The field of Clinical Next-Gen Sequencing Analysis is highly dynamic. We are essentially building the MRI for the genome. In this process the latest research on disorders is combined with our understanding of the best treatment option at any given time. We have collectively entered probably the most exciting part of the adoption curve. For over fifteen years, we at Golden Helix have provided tools for the research community to better understand the underlying mechanisms that cause disease.

As our field is preparing to take the next step, we will do the same for clinicians, genetic counselors, and any other medical specialist in need for the best possible tool support to come up with reliable clinical diagnosis. By doing so, we have the opportunity to improve patient outcome, helping to fight the diseases of our time such as cancer, Alzheimer's, and diabetes.

Try VarSeq for Free

Discover VarSeq

For more resources visit: www.goldenhelix.com

Bibliography

Biesecker, L. G. (2010) Leslie G. Biesecker, "Exome sequencing makes medical genomics a reality", pp. 13-14, *Nature Genetics* Vol 42, 2010.

Kurian, A. W. et al (2014) Allison W. Kurian, Emily E. Hare, Meredith A. Mills, Kerry E. Kingham, Lisa McPerhson, Alice S. Whittermore, Valerie McGuire, Uri Ladabaum, Yuya Kobayashi, Stephen E. Lincoln, Michele Cargill and James M. Ford, Clinical Evaluation of Multiple-Gene Sequencing Panel for Hereditary Cancer Risk Assessment, *Journal of Clinical Oncology*, April 2014.

Leslie, T. et al (2012) T. Leslie, D. Agar, S. Fielding, S. Miller, "Market Trends in Genetic Services, Booz Allen Hamilton, 2012.

McKinsey (2013) *Personalized Medicine*, McKinsey, 2013.

Monti, P. et al (2003) Paola Monti, Paola Campomensosi, Yari Ciribilli, Raffaella Iannone, Anna Aprile, Alberto Inga, Mitsuhiro Tada, Paola Menichini, Angelo Abbondandolo and Gilberto Fronaz, "Characterization of the p53 mutants ability to inhibit p73 beta transactivation using a yeast-based functional assay, pp. 5252-5260, *Oncogene* (2003)

Yang Y. et al (2013) Yang Y., Muzny D. M., Reid, J. G., Bainbridge, M. N., Willis, A., Ward, P. A., Braxton A., Beuten, J., Xia, F., Niu, Z., Hardison, M., Person, R., Bekheima, M. R., Leduc, M. S., Kirby, A., Pham, P., Scull, J., Wang, M., Ding, Y., Plon, S. E., Lupski, J. R., Beaudet, A. L., Gibbs, R. A., Eng, C. M., "Clinical whole-exome sequencing for the diagnosis of mendelian disorders", pp. 1502-1511, Vol 369(16), *N Engl J Med*, Oct 17, 2013.

About the Author

Dr. Andreas Scherer is CEO of Golden Helix. The company has been delivering industry leading bioinformatics solutions for the advancement of life science research and translational medicine for over a decade. Its innovative technologies and analytic services empower scientists and healthcare professionals at all levels to derive meaning from the rapidly increasing volumes of genomic data produced from microarrays and next-generation sequencing. With its solutions, hundreds of the world's top pharmaceutical, biotech, and academic research organizations are able to harness the full potential of genomics to identify the cause of disease, improve the efficacy and safety of drugs, develop genomic diagnostics, and advance the quest for personalized medicine. Golden Helix products and services have been cited in over 850 peer-reviewed publications.

He is also Managing Partner of Salto Partners, a management consulting firm headquartered in the DC metro area. He has extensive experience successfully managing growth as well as orchestrating complex turnaround situations. His company, Salto Partners, advises on business strategy, financing, sales and operations. Clients are operating in the high tech and life sciences space.

Dr. Scherer holds a PhD in computer science from the University of Hagen, Germany, and a Master of Computer Science from the University of Dortmund, Germany. He is author and co- author of over 20 international publications and has written books on project management, the Internet, and artificial intelligence. His latest book, "Be Fast Or Be Gone", is a prizewinner in the 2012 Eric Hoffer Book Awards competition, and has been named a finalist in the 2012 Next Generation Indie Book Awards!

Connect with Dr. Scherer:

