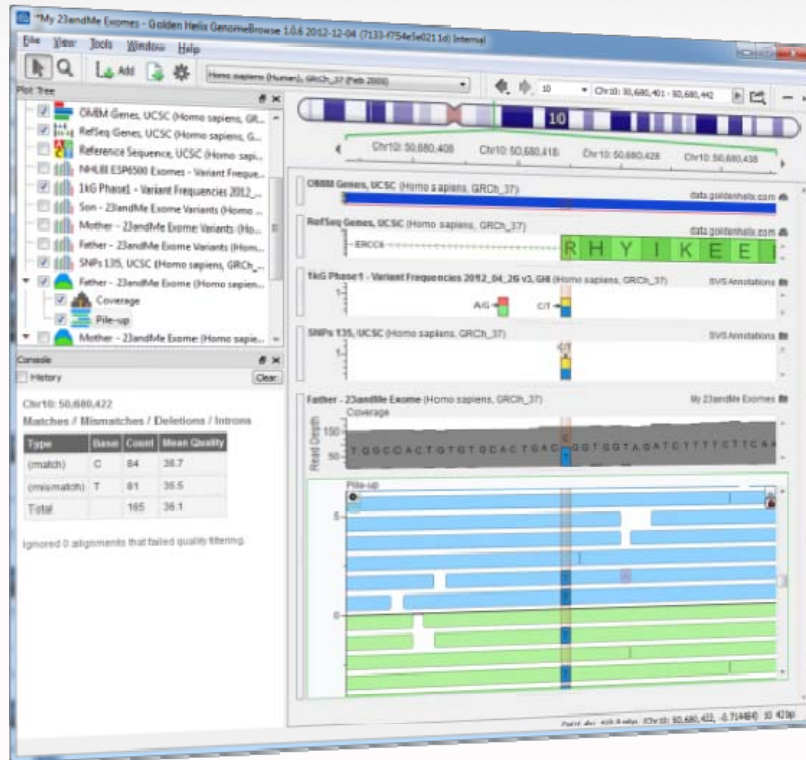


# @gabeinformatix: 23andMe Variant Analysis of My Personal Exome



Gabe Rudy

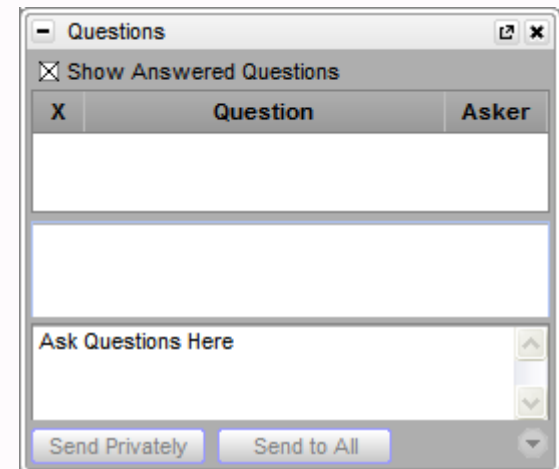
Vice President of Product Development

December 5, 2012



# Questions during the presentation

Use the Questions pane in your GoToWebinar window





- Exome Sequencing and Medical Tests
- Can Consumer Genomics Benefit from NGS?
- What I Hope to Do with My Exome
- My Exome: By The Numbers
- Summary of Results

# Exome Sequencing: Success Leads to...



Journal of  
**MEDICAL GENETICS**

- **Wave One: Exome Sequencing as Diagnosis of Last Resort by Researchers**
  - Rare, Congenital, Highly Penetrant, Presumed Monogenic
- **Wave Two: Expanded Research and Select Clinical Use**
  - Research outside above constraints
  - Clinical Exomes as end to diagnostic odyssey or even “shortcut”
- **Wave Three: Path to Personalized Medicine?**
  - Standardized clinical use
  - Consumer driven innovation?

ORIGINAL ARTICLE

## Clinical application of exome sequencing in undiagnosed genetic conditions

Anna C Need,<sup>1</sup> Vandana Shashi,<sup>2</sup> Yuki Hitomi,<sup>1</sup> Kelly Schoch,<sup>2</sup> Kevin V Shianna,<sup>1</sup> Marie T McDonald,<sup>2</sup> Miriam H Meisler,<sup>3</sup> David B Goldstein<sup>1,4</sup>

### ABSTRACT

**Background** There is considerable interest in the use of next-generation sequencing to help diagnose unidentified genetic conditions, but it is difficult to predict the success rate in a clinical setting that includes patients with a broad range of phenotypic presentations.

**Methods** The authors present a pilot programme of whole-exome sequencing on 12 patients with unexplained and apparent genetic conditions, along with their unaffected parents. Unlike many previous studies, the authors did not seek patients with similar phenotypes, but rather enrolled any undiagnosed proband with an apparent genetic condition when predetermined criteria were met.

**Results** This undertaking resulted in a likely genetic diagnosis in 6 of the 12 probands, including the identification of apparently causal mutations in four genes known to cause Mendelian disease (*TCF4*, *EFTUD2*, *SCN2A* and *SMAD4*) and one gene related to known Mendelian disease genes (*NGLY1*). Of particular interest is that at the time of this study, *EFTUD2* was not yet known as a Mendelian disease gene but was nominated as a likely cause based on the observation of de novo mutations in two unrelated probands. In a seventh case with multiple disparate clinical features, the authors were able to identify homozygous mutations in *EFEMP1* as a likely cause for macular degeneration (though likely not for other features).

**Conclusions** This study provides evidence that next-generation sequencing can have high success rates in a clinical setting, but also highlights key challenges. It further suggests that the presentation of known Mendelian conditions may be considerably broader than currently recognised.

### INTRODUCTION

Whole-genome and whole-exome sequencing have proven remarkably successful in identifying the causes of Mendelian diseases. These analyses have generally depended on the availability of more than one unrelated affected individual and/or linkage evidence in at least one family. However, next-generation sequencing (NGS) has also succeeded in identifying causes of genetic conditions even when they are seen in only a single patient.<sup>1–3</sup>

Consequently, there is growing interest in the introduction of NGS into the clinic to aid in the diagnosis of conditions for which no genetic cause can be found with targeted testing or chromosomal arrays. However, in a clinical setting, patients with

undiagnosed genetic conditions tend to present with a wide range of clinical features, and it is often necessary to consider each patient's genome individually, rather than looking for common disrupted genes in multiple cases with a similar phenotype. It is not clear what success rate NGS approaches will achieve in providing genetic diagnoses in this more challenging setting. In this study, we have evaluated the use of NGS to provide genetic diagnoses using 12 parent-child trios in which the child had congenital anomalies and/or intellectual disabilities due to unexplained conditions presumed to be genetic. Importantly, the patients were chosen to be representative of a clinical sample of undiagnosed genetic conditions, in that they were not selected for genetic tractability or phenotypic homogeneity.

### METHODS

Exome sequencing was performed on each patient and both parents using the Illumina HiSeq2000 platform and the Agilent SureSelect Human All Exon 50Mb Kit. Detailed methods for laboratory work can be found in the online supplementary methods.

### Study population

The research protocol was approved by the Duke Institutional Review Board, and all human participants or their guardians gave written informed consent. Twelve families (child, mother and father) were recruited through the genetics clinic at Duke University Medical Center based on whether their child met two or more of the following criteria: (1) unexplained intellectual disability and/or developmental delay; (2) one major congenital anomaly; (3) 2–3 minor congenital anomalies; and (4) facial dysmorphism. In addition, the families were required to meet the following eligibility requirements: (1) both biological parents available for testing; (2) previous clinically indicated genetic testing, including a chromosomal microarray (Affymetrix 6.0, <http://www.affymetrix.com>), had been normal; and (3) no evidence of effects of teratogens, birth asphyxia or non-accidental trauma. Subjects were not eligible if the mother was pregnant at the time of enrolment. Finally, results were only returned to patients and/or patient families following confirmation of detected variants in a CLIA certified laboratory. Controls were subjects enrolled in Center for Human Genome Variation studies through Duke Institutional Review Board approved protocols (n=830).

# Exome Sequencing in Consumer Genomics



## ■ 23andMe Provides Genotyping Service

- ~1M SNPs genotyped
- 48 Diseases Carrier Status
- 57 Traits
- 20 Drug Responses
- 119 Diseases Risk Predictions

## ■ Exome done as Pilot Program

## ■ 80X coverage

## ■ Raw Data

## ■ No Interpretation

Exome 80x  
Pilot Program

Be one of the first to get your personal exome sequence  
**\$999** Enrollment Currently Closed

Sign up to be notified when ordering is available  
email  [Notify me »](#)

Announced at Health 2.0, San Francisco - September 27, 2011

**What is an exome? How is it different from a full genome sequence?**  
Your exome is the 50 million DNA bases of your genome containing the information necessary to encode all your proteins. Informally, you can think of the exome as the DNA sequence of your genes.  
Your entire genome is made up of your exome plus other DNA, consisting of three billion bases with repetitive sequences, sequences of unknown function, and DNA that does not code for proteins.

**How is this different from what 23andMe already offers?**  
23andMe's current Personal Genome Service® (PGS) analyzes your DNA at approximately one million locations in the genome. The PGS® provides more than 200 detailed reports linking different genetic variants to health conditions, traits, and ancestry, as well as connecting people to other users who share DNA.  
In contrast, the exome sequencing pilot provides users with raw variant data for about 50 million bases of DNA, without reports. Over time, 23andMe will add a limited set of tools and content that will analyze exome sequencing data.

**What do I get for \$999?**  
You get access to your raw data of 50 million DNA bases at high quality (80X coverage). Over time, you will have access to new tools and content as they are developed to take advantage of your exome sequence data. Most excitingly, you'll be a trailblazer, one of the first people on the planet to know their personal exome sequence!

**When does the project start and how do I join?**  
At this time, the pilot project is full and enrollment is closed. If you are interested in joining the pilot program, you will be notified when enrollment is open.

**Why sequence my exome?**  
A person's physical structure, their body's chemical reactions and the expression of their genes are controlled by the proteins encoded in the exome. The vast majority of genetic diseases also hinge on variations in the exome. For these reasons, exome data may be useful for those exploring their personal sequence data.  
Exome data are less suitable for ancestry or genealogical research, since they will not provide mitochondrial sequence or much information on the Y chromosome.

**Who can take part in the pilot program?**  
We are offering the pilot program exclusively to current 23andMe PGS users. The exome sequencing pilot is a limited-time offer, and it is subject to change without notice. It is not available in all countries. Raw genetic data are not available for users who have opted out of sharing their data. Exome data are not available for adopters and donors.



# Still Complex and Research Oriented



- Found 8K phantom variants
- Variants Outside Exons?
- Exome Capture:

*Sequencing an exome at an average depth of 30-50x typically yields 70-75% of the target region at coverage levels of at least 20x. With clinical samples sequenced at 100x, 83-94% of the target region is typically covered at 20x or greater.*

**~EdgeBio Exome Seq Cheat Sheet**

**GATK is a Research Tool. Clinics Beware.**  
Posted on December 3, 2012 by Gabe Rudy

In preparation for a [webcast](#) I'll be giving on Wednesday on my own exome, I've been spending more time with variant callers and the myriad of false-positives one has to wade through to get to interesting, or potentially significant, variants.

So recently, I was happy to see a message in my inbox from the 23andMe exome team saying they had been continuing to work on improving their exome analysis and that a "final" analysis was now ready to download.

This meant I had both an updated "variants of interest" report as well as updated variant calls in a new VCF file. I'll get to the report in a second, which lists rare or novel variants in clinically associated genes, but first let's look at what changed in the variant calls.

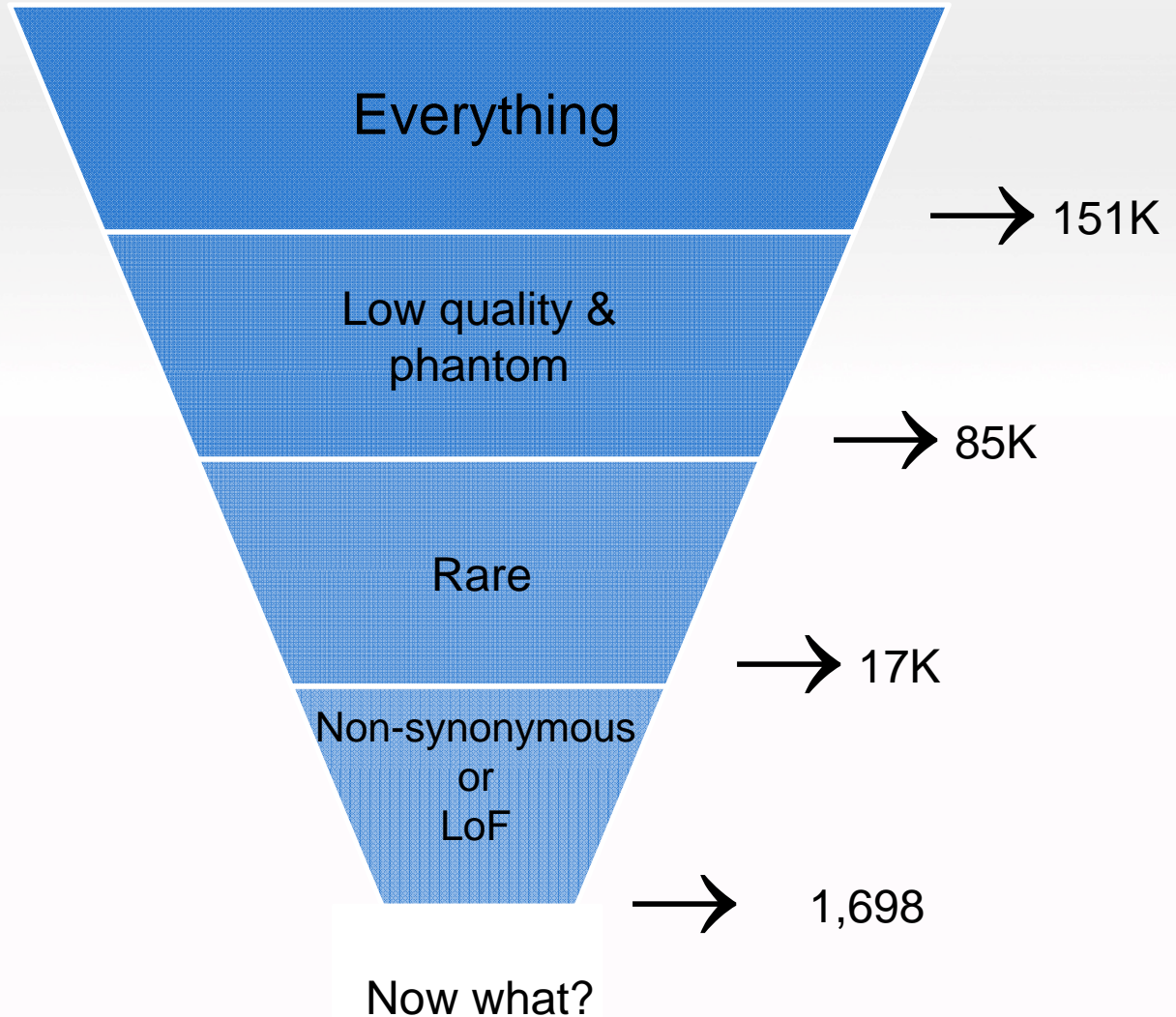
**New... and improved?**  
At first blush, the variant files look quite different as you can see in the first Venn diagram below comparing variants (both unique and common) between the original and new files. But after I applied a conservative filter (Genotype Quality (GQ) > 10 and Read Depth (RD) > 10) on the variants, things start to look less dramatic. So what is with all the new variants? It looks like many are just more aggressive variant calls. In fact there were ten thousand variants with a read depth of just one (a single read) in the new file!

Category	Count
Original VCF File (Total)	106,760
"Final" VCF File (Total)	152,205
Unique to Original	974
Unique to Final	46,419
Shared	105,786

# Filtering Strategy



- Follow best practice for high-impact variants
- Transparent and reproducible in software
- Interpretation more open ended



Quality Histograms [145]

File View Help

Quality Histograms [145]

File View Help

X-Axis Zoom 0 - 30.3653

Graph Control Interface

- User Graphs
  - Graph 1
    - Read Depth
  - Graph 2
    - LK8250\_GQ

Graph Axes Add Item

Title  Legend

Bin Count 250

X-Min: 0 X-Max: 30.3653

Y-Min: # Y-Max: #

Data Console

Current History User Annotations

**Read Depth**

Read Depth Bin	Count
0	10000
1	9000
2	7500
3	6500
4	5500
5	4800
6	4200
7	3800
8	3500
9	3200
10	3000
11	2800
12	2600
13	2400
14	2200
15	2100
16	2000
17	1900
18	1800
19	1700
20	1600
21	1500
22	1400
23	1300
24	1200
25	1100
26	1000
27	900
28	800
29	700
30	600

**LK8250\_GQ**

Genotype Quality Bin	Count
0	2000
1	2000
2	3000
3	12000
4	6000
5	2000
6	7000
7	2000
8	1000
9	4000
10	1000
11	1000
12	3000
13	1000
14	1000
15	3000
16	1000
17	1000
18	2000
19	1000
20	1000
21	2000
22	1000
23	1000
24	2000
25	1000
26	1000
27	2000
28	1000
29	1000
30	2000





## Example in GenomeBrowse of Quality Filter

# Population Catalog and Variant Classification



Non-coding		Variant	Rare (Novel)
	Intergenic	8,462	3,609 (1,130)
	Intronic	48,826	7,516 (4,418)
	UTR 3/5	4,128	669 (303)
	Non-coding	1,643	648 (183)

Coding		Variant	Rare (Novel)
	Splicing	79	28 (17)
	Frameshift Ins/Del	196	138 (118)
	Stop gain/loss	113	31 (9)
	Non-synonymous	10,252	1,501 (447)
	Ins/Del	256	162 (89)
	Synonymous	11,080	885 (215)
Unknown	589	176 (36)	



- Regions of Chromosomal Duplication (SuperDups)
- Look at genes in OMIM (most)
- Use predictions of genes as recessive/haploinsufficient to weed out low-priority genes
- For NonSynonous missense variants can use functional prediction to annotate



## Review Homozygous Variants

# Genes of Interest and Homozygous Variants



		Loss of Function			
		Rare	!Dups	OMIM	Rec Genes
Loss of Function	Splicing	28	17	12 (2)	0 (0)
	Frameshift Del	60	44	32 (4)	1 (0)
	Frameshift Ins	78	66	46 (5)	3 (1)
	Stop gain	31	8	7 (0)	0 (0)

		Non-Synonymous			
		Rare	!Dups	OMIM	Rec Genes
Non-Synonymous	Damaging (3/3)	337	108	85 (0)	9 (0)
	Damaging (2/3)	205	139	97 (0)	7 (0)
	Damaging (1/3)	136	194	129 (1)	12 (0)
	Tolerated/Unk	781	339	204 (4)	10 (1)

# Homozygous Variants



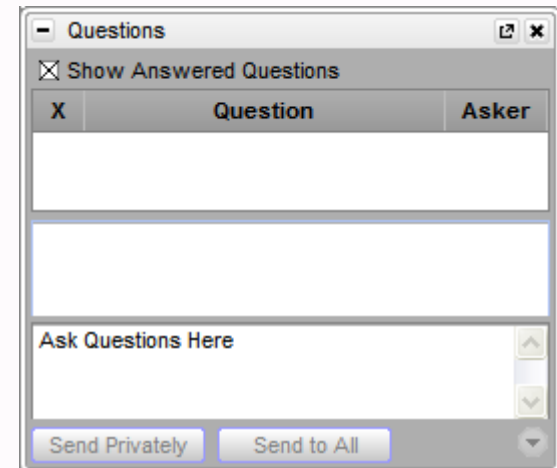
Note	Variant	LK8250_A D	LK8250_ DP	LK8250_ GQ	Gene(s)	Classification	HGVS Coding 1
common	1:54605319-Ins	50,26	76	99	CDCP2	Frameshift Ins	c.1224_1225insC
reference	2:71062833-Ins	106,1	107	99	CD207	Splicing	
in-wife	5:156721864-Ins	6,91	97	99	CYFIP2	Frameshift Ins	c.279_280insC
bad-call	6:44269193-Del	120,1	121	99	AARS2	Frameshift Del	c.2607delG
bad-call?	10:46999604-SNV	21,140	161	99	GPRIN2	Nonsyn SNV	c.724A>G
common	12:26834806-Ins	95,1	96	99	ITPR2	Splicing	
in-wife	14:63784408-Ins	3,141	144	99	GPHB5	Frameshift Ins	c.156_157insC
bad-call	17:7606722-Del	161,6	167	99	WRAP53	Frameshift Del	c.1565delC
bad-call	19:54649671-Del	142,1	143	99	CNOT3	Frameshift Del	c.729delT
in-wife	22:19189004-Ins	6,183	189	99	CLTCL1	Frameshift Ins	c.3601_3602insG
VUS	X:16657321-SNV	0,54	54	99	CTPS2	Nonsyn SNV	c.1342A>C
pathogenic	X:38226614-SNV	0,29	29	84.27	OTC	Nonsyn SNV	c.148G>A
VUS	X:100496711-SNV	0,65	66	99	DRP2	Nonsyn SNV	c.380C>T
VUS/in-5-M	X:105167411-SNV	0,16	16	48.13	NRK	Nonsyn SNV	c.2912A>G
wrong-geno	X:112022302-Ins	61,1	62	99	AMOT	Frameshift Ins	c.3080_3081insCC
VUS/common	X:150349559-Del	106,4	110	96.99	GPR50	Frameshift Del	c.1504_1514delACCACTG GCCA





# Do You Have Any Questions?

Use the Questions pane in your GoToWebinar window



406-585-8137  
info@goldenhelix.com