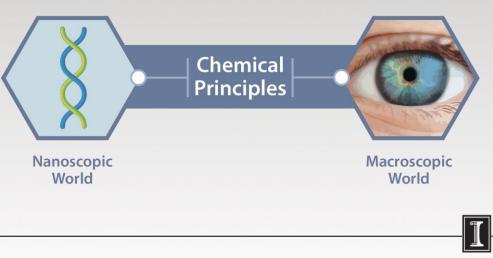
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

# The Molecular Sciences Made Personal



### Jeff Moore

Department of Chemistry University of Illinois at Urbana-Champaign Mar-25, 2015



Today's Outline •Introduction: Genetics and Chemistry? •Why Golden Helix? •Design of the educational study •What type of data to use? SNP or WES? •Maybe both types? •Is the data quality high? •What's been my experience from the pilot seat?

www.illinois.edu

## About Today's Webinar

- The webinar will outline **an educational project** aimed at developing a sequence of chemistry courses for pre-health students as a pathway to biochemistry that builds a strong foundation of molecular understanding and scientific reasoning skills.
- Aligns with the **premedical competencies** outlined in the American Association of Medical Colleges (AAMC)-HHMI report on Scientific Foundation for Future Physicians, which calls for stronger connections between course content and the underlying principles in health and medicine.
- Acquiring personal genetic data is affordable and is expected to become an important part of the healthcare industry. There is a growing need to educate prospective healthcare professionals in the interpretation of genetic data and the role of genotype-phenotype association in molecular etiology
- An investigation of "self" will not only serve to **motivate**, but also to equip students with knowledge and **hands-on experiences** to help them understand scientific and medical principles, and challenges that come with revolutionary changes, so they navigate more confidently through their future professional careers and lead change in the healthcare field.
- **Disclaimer**: With no formal education or training in genetics, I see the subject with fresh, but naïve eyes. I am, nonetheless an experienced learner. It's good for my students to an experienced learner continue to love to learn, even in the struggles of not teaching as the master.
- Constructive feedback is highly appreciated!

## Why Golden Helix? A Commitment to Education

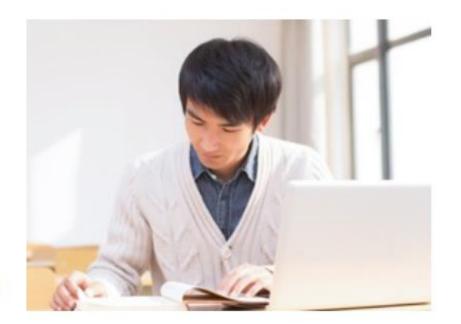
Link to Andreas' blog on...

## **Preparing the Next Generation of Genetic Researchers**

Posted on March 25, 2014 by Andreas Scherer

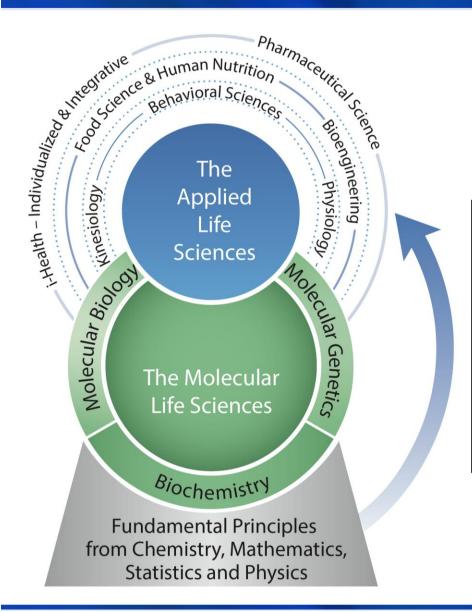
🕒 🚹 🔝

New breakthroughs are being made every day in genomics. It's a dynamic and fascinating industry, and with exceptional growth forecast in the DNA sequencing market, a new generation of people are entering the field: future researchers, clinicians, counselors and doctors. This new generation will need to learn not only the science, but also understand how to process the massive amounts of data generated with DNA sequencing (and genomics in general).



http://blog.goldenhelix.com/ascherer/preparing-the-next-generation-of-genetic-researchers/

# Do Personal Connections Lead to Greater Learning Gains?



Hypothesis: Learning gains are greater when content is strongly connected to knowledge that students acquire in an aligned field.

Question: How to teach organic chemistry using content that connects to students' knowledge from the Molecular Life Sciences?

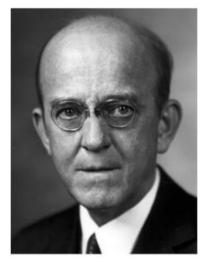
## Atoms-to-Instructions: CHEM 332 Fall 2013

R Sep-5	<ul> <li>elementary steps &amp; reaction mechanisms</li> </ul>	A chemist's way of thinking
Lesson 05 T Sep-10	<ul> <li>proton transfer</li> <li>general-acid / general-base electron flow</li> <li>the mechanism of bisulfite formation</li> </ul>	about bisulfite' its chemic
Lesson 06 R Sep-12	<ul> <li>electrostatic potential energy surfaces (PES)</li> <li>frontier MOs (FMOs)</li> <li>FMO energies for H<sub>2</sub>SO<sub>3</sub>, HSO<sub>3</sub><sup>-</sup> &amp; SO<sub>3</sub><sup>2-</sup></li> <li>which atom is most nucleophilic?</li> <li>which atom is most electrophilic?</li> </ul>	
Lesson 07 T Sep-17	<ul> <li>charge- vs. FMO-controlled react</li> <li>filled / empty interactions</li> <li>generalized σ-type &amp; π-type int</li> </ul>	
Lesson 08 R Sep-19	<ul> <li>π MOs</li> <li>π delocalization</li> <li>aromaticity</li> <li>aromatic heterocycles</li> </ul>	
Lesson 09 T Sep-24	<ul> <li>reactions of heteroaromatics</li> <li>the π MO as a nucleophile</li> <li>the nonbonding (n) MO as a nucleo</li> <li>the π* MO as an electrophile</li> </ul>	
Lesson 10 R Sep-26	<ul> <li>the nucleobases</li> <li>tautomeric forms &amp; equilibria</li> <li>nucleophilic &amp; electrophilic reactivit</li> <li>hydrolytic deamination (without bisur.</li> </ul>	

Building a bridge between molecular genetics and organic chemistry

**Chemical roots**: More than seventy years have passed since Avery, MacLeod and McCarty published their landmark paper revealing that DNA possessed genetic information that could transform the heritable character of cells.

Chemical analysis showed that the proportions of carbon, hydrogen, nitrogen, and phosphorus in this active portion were consistent with the chemical composition of DNA.



Oswald T. Avery

Avery, O.T., MacLeod, C.M. & McCarty, M. (**1944**) *J. Exp. Med. 79*, 137-159. *Elementary Chemical Analysis.*<sup>1</sup>—Four purified preparations were analyzed for content of nitrogen, phosphorus, carbon, and hydrogen. The results are presented in Table I. The nitrogen-phosphorus ratios vary from 1.58 to 1.75 with an average value of 1.67 which is in close agreement with that calculated

 TABLE I

 Elementary Chemical Analysis of Purified Preparations of the Transforming Substance

Preparation No.	Carbon	Hydrogen	Nitrogen	Phosphorus	N/P ratio
	per cent	per cent	per cent	per cent	
37	34.27	3.89	14.21	8.57	1.66
38B		_	15.93	9.09	1.75
42	35.50	3.76	15.36	9.04	1.69
44	—	-	13.40	8.45	1.58
Theory for sodium desoxyribonucleate	34.20	3.21	15.32	9.05	1.69

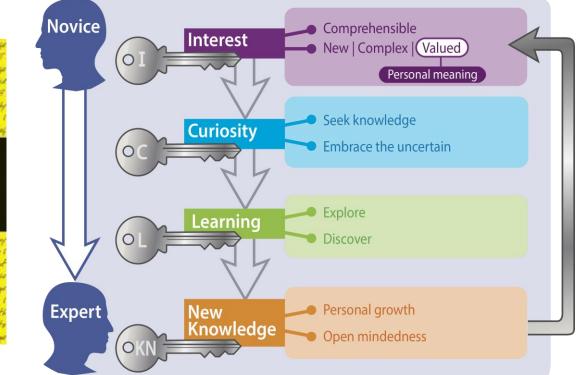
on the basis of the theoretical structure of sodium desoxyribonucleate (tetranucleotide). The analytical figures by themselves do not establish that the substance isolated is a pure chemical entity. However, on the basis of the nitrogen-phosphorus ratio, it would appear that little protein or other substances containing nitrogen or phosphorus are present as impurities since if they were this ratio would be considerably altered.

See: Current Biology (2004) 14, pp. R605–R607 doi:10.1016/j.cub.2004.07.038

Hypothesis: Learning gains are greater when content is strongly connected to knowledge that students recently acquired in an aligned field.

Hypothesis: Learning gains are greater when content is strongly connected to knowledge that students recently acquired in an aligned field

Hypothesis: Learning gains are greater when content is strongly connected to knowledge that students recently acquired in an aligned field; even greater, when the content is personal.



33 Chara is a haawan, da paopla hawa sax? Aone aan 1 waka wy lija waxa anoistingi? When is killing yostipad? Anne made graa ant? da 1 hand? What will paopla say at ng panard? Anne maded ny liga ha disparant is 1 new newa willing to anslora indeaed as evold challangud? Ana Chara mara sidaa as ma t hana yat to know? Anne van 1 mearcama ny spars? Ally an 1 sa aanty hanad? May da 1 alaoga and ny ne saak ansjet satationishipi? Anne an t barana anne and ny ne saak ansjet satationishipi? Anne an t barana anne sadadirina and anadist adationishipi? Anne ao ti barana anne and ny ne saak ansjet satationishipi? Anne ao ti barana anne and ny ne saak ansjet satationishipi? Anne ao ti barana anne sadadirina and anadist an konk?

# **CURIOUS?**

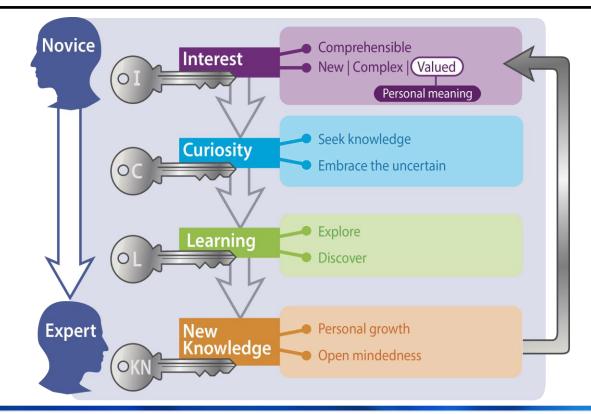
Discover the Missing Ingradiant to a folgilling size

#### TODD KASHDAN, PH.D.

nesa vibrant in antar nikan it a dyingit. Niky de 1 de sa ming Chinga'i dan't (ika and lika sa mang thinga'i dan't dat' Niky da 1 sud it ao harid ta apan ng ta atkaret Niky an i sa marriad abart nikat athar, pazzia (kinik aj mdi' in tima travnik porisibilit in Chara anuk a thing an 'sana at jirat night't Niky da bad thinga happan ta gand pazzidit. Niky da wa jaar daathit dan ann t raignita parrian in ng marriagat. Ham ann saar daathit dan ann t raignita parrian in ng marriagat. Ham an anastaity ba bat batarjang and jait' dan i bazagit athat ann i anastaity ba bat barigant man lang daar it taba ta maka a raai, ianting ahanga

Hypothesis: Learning gains are greater when content is strongly connected to knowledge that students recently acquired in an aligned field; even greater, when the content is personal.

Question: How to deliver **personally meaningful content** to the individual organic chemistry student?



Hypothesis: Learning gains are greater when content is strongly connected to knowledge that students recently acquired in an aligned field; even greater, when the content is personal.

Question: How to deliver **personally meaningful content** to the individual organic chemistry student?

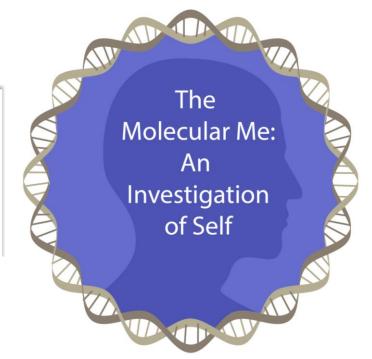
#### NATURE | COLUMN: WORLD VIEW

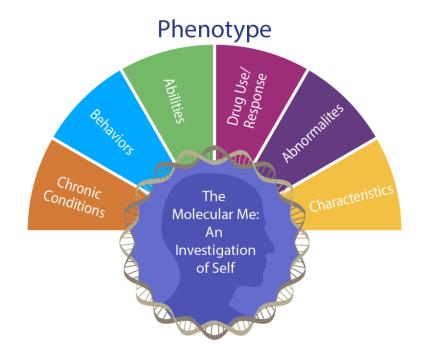


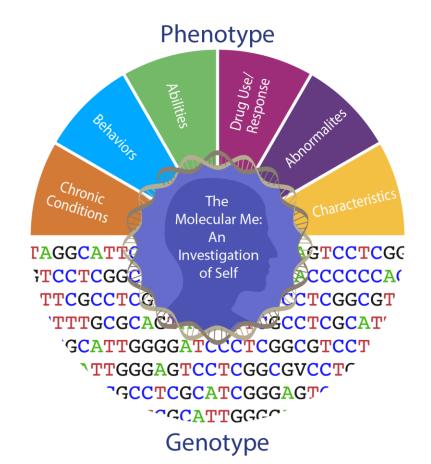
### Improving genome understanding

The cost and accuracy of genome sequencing have improved dramatically. George Church asks why so few people are opting to inspect their genome.

09 October 2013 Corrected: 09 October 2013

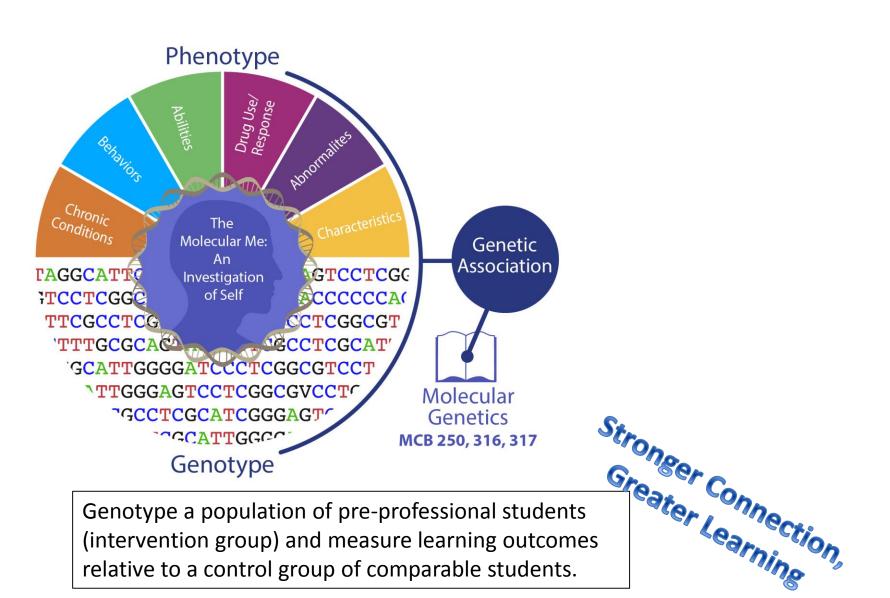


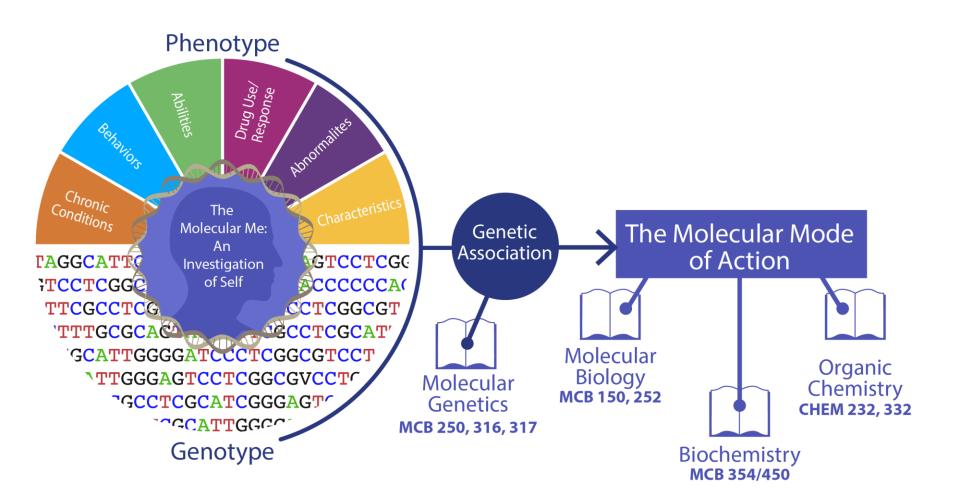


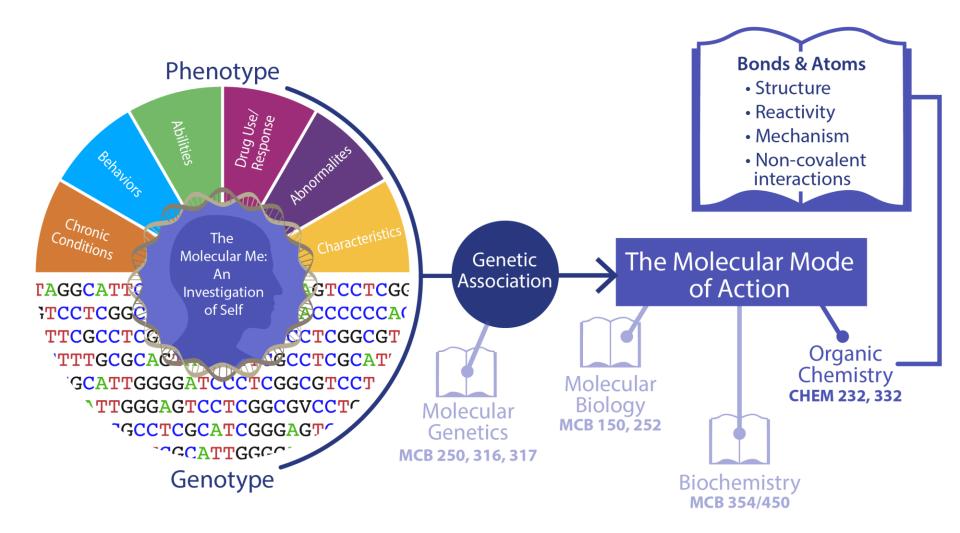


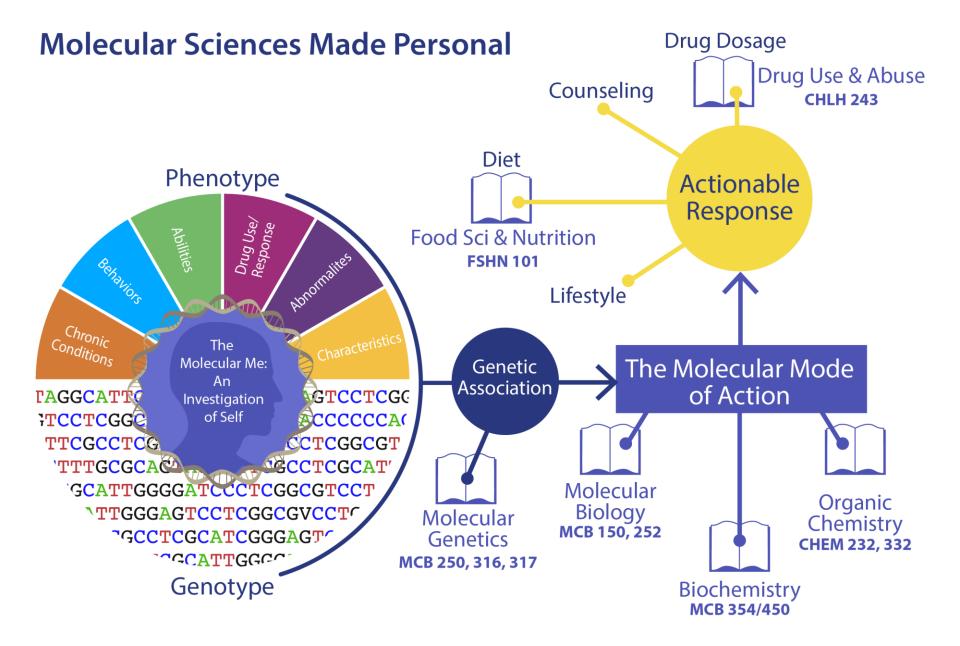
Genotype a population of pre-professional students (intervention group) and measure learning outcomes relative to a control group of comparable students.

Stronger Connections ts nes .s.

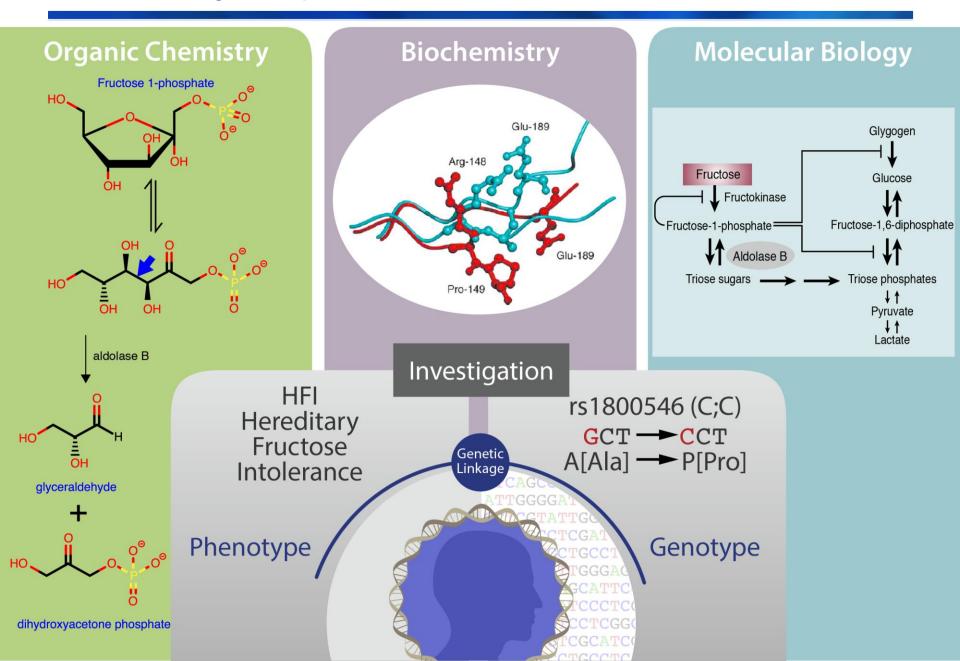








### A Crosscutting Example of Molecular Sciences Made Personal



#### Interventions

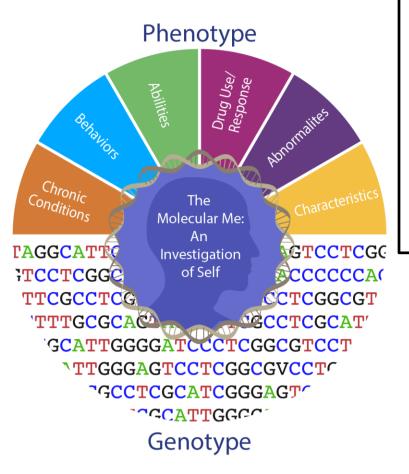
Stronger Connection, Greater Learning To test the "stronger connection, greater learning" hypothesis.

- 1. Course content approach
- 2. Personal genotyping approach

### **Specific Aims**

- Establish logistics and protocols for genotyping a group of students (intervention group).
- Pilot genotype activities:
  - Teaching *The Molecular Me* Freshman Discovery Course to non-science majors (Fall 2014)
  - Teaching *The Molecular Me* to Illinois OLLI students (Spring 2015).
- Genotype intervention students as they finish MCB 150 (i.e., Fall 2015) ٠
- Develop and teach a sequence of chemistry courses for the intervention group that ۰ integrates use of personal genetic data (Fall 2015).
- Measure learning outcomes in Biochemistry (e.g., Fall 2017) for intervention group relative to control group.

## **Genotyping Protocols and Logistics**



- Each student will be given the <u>option</u> to undergo genome-wide genotyping
- Student-managed data
- IRB approval
- To qualify, participants must demonstrate an understanding of risks and protocols
- SNP Genotyping vs. WES or both?
- Oragene saliva collection kit
  - 6-8 weeks to receive genotype data
  - Learn from others Personalized genomic services already used in the classroom

#### What type of data should be collected?

- DTC SNPs?
- WES?
- Both?

Today we'll conduct data validation analyses using VarSeq

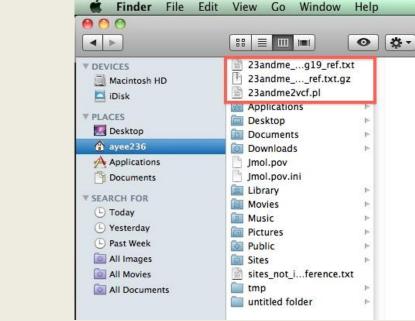
## Data validation by a systematic hereditary analysis in VarSeq

- Genotype data for mother and father obtained by 23andMe in Nov-2013 using their v3 microarray chip
- Son's genotype data obtained by 23andMe in Jun-2014 using the <u>v4 microarray chip</u>
- <u>Data sheet</u> for Illumina's HumanOmniExpress-24 format chip reports reproducibility values between 99.9 to 99.99%
- Genotype data from 23andMe were downloaded in their standard .txt format and converted to .vcf files using the open source program <u>23andme2vcf converter</u> (see next slide for a detailed set of instructions including Mac and PC platforms)
- Trio data loaded as .vcf files into VarSeq using Hereditary Gene Panel Starter Template
- Computed genotype zygosity for all three samples using VarSeq algorithm
- For each possible combination of parent's zygosity, examined son's SNPs for inconsistencies (Note: due to the change in microarray chip, a significant fraction of SNPs were not aligned)
- Only considered autosomal DNA

E Ref / Ref II Not Ref	* -
■ Reference (JSM 23andMe)	* -
Hemizygous	8,974
Heterozygous	268,542
Homozygous Variant	176,119
Reference	485,930
Missing	73,700
	485,930
■ Reference (LMM 23andMe)	* -
Hemizygous	3
Heterozygous	107,552
Homozygous Variant	27,004
Reference	346,747
Missing	4,624
	346,747
■ Not Reference (Son 23andMe)	÷ -
Hemizygous	0
Heterozygous	76
Homozygous Variant	3
Reference	172,925
Missing	173,743
	79
	79

## Instructions for .txt to .vcf conversion

- Windows users: Detailed <u>instructions</u> and an instructional <u>video</u> for the 23andme2vcf converter
- Mac users: Detailed <u>instructions</u> and an instructional <u>video</u> for the 23andme2vcf converter



Special thanks to Mr. Albert and Ms. Anna Yee for writing and testing the instructions and for preparing the instructional videos

## Data validation via systematic heritability analysis

	Inheritance consistencies used			Mother	
to validate genotype data		Ref (0/0)	ht (0/1)	Hvar (1/1)	
		Ref (0/0)	Ref	Ref or ht	ht
	Father	ht (0/1)	Ref or ht	Ref, ht or Hvar	ht or Hvar
		Hvar (1/1)	ht	ht or Hvar	Hvar

	Inheritance inconsistencies used			Mother	
to identify genotyping errors		errors	Ref (0/0)	ht (0/1)	Hvar (1/1)
		Ref (0/0)	Not Ref	Hvar	Not ht
	Father	Ht (0/1)	Hvar		Ref
	Ľ.	Hvar (1/1)	Not ht	Ref	Not Hvar

## Findings from systematic heritability analysis

Inheritance <b>co</b>				Mother	
to validate genotype data		Ref (0/0)	ht (0/1)	Hvar (1/1)	
	ir	Ref (0/0)	172,925	61,941	14,991
	Father	ht (0/1)	61,583		33,190
	Ľ.	Hvar (1/1)	15,068	34,308	48,202

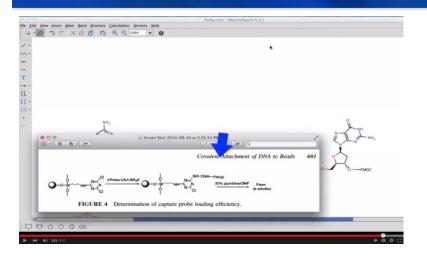
	Inheritance <b>inconsistencies</b> used to identify genotyping errors			Mother	
to identify gen			Ref (0/0)	ht (0/1)	Hvar (1/1)
	j.	Ref (0/0)	79	5	17
	Father	Ht (0/1)	6		7
	Ľ	Hvar (1/1)	7	1	31

## Summary of results for systematic heritability analysis

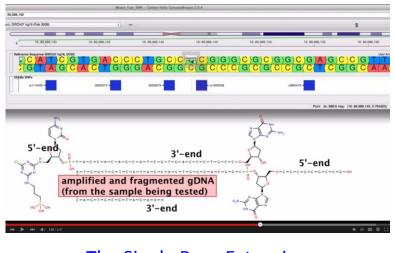
ir		tent SNPs		Mother	
	in the trio data set.		Ref (0/0)	ht (0/1)	Hvar (1/1)
		Ref (0/0)	0.05%	0.008%	0.11%
	Father	ht (0/1)	0.01%		0.02%
	Ľ.	Hvar (1/1)	0.05%	0.003%	0.06%

- The quality of the SNP data from 23andMe is <u>high</u> consistent with specifications reported in Illumina's technical data sheets.
- Note that the **inconsistent SNP** data (table above) required 3 measurements for each entry (which of the three is in error cannot be determined).
- A reproducibility test on independent samples from the same individual is a worthwhile alternative for educational purposes (not yet performed).
- The inconsistencies are currently being analyzed for systematic genotyping errors.
- VarSeq provides a simple yet powerful way to validate SNP data!

# A Look Under the Hood: SNP Chip Genotyping Chemistry



BeadArray and Oligonucleotide Chemistry for SNP Genotyping Technology



The Single Base Extension Method of SNP Detection

**Quick demo**: From <u>GenomeBrowse</u> to <u>MarvinSketch</u> (i.e., Letters to Bonds & Atoms) Let's look into GRCh37 at 8: 18,257,451 - 18,258,779 (an exon of the *NAT2* gene)

- Frank J Steemers, Weihua Chang, Grace Lee, David L Barker, Richard Shen & Kevin L Gunderson Wholegenome genotyping with the single-base extension assay. *Nature Methods* 2006, 3, 31 -33 doi:10.1038/nmeth842
- 2. Frank J. Steemers, Kevin L. Gunderson Whole genome genotyping technologies on the BeadArray<sup>™</sup> platform. *Biotech J* 2007, *2*, 41-49 DOI: 10.1002/biot.200600213
- 3. Illumina Technical Note on Infinium® II Assay Workflow. Pub. No. 370-2006-027 07Dec06

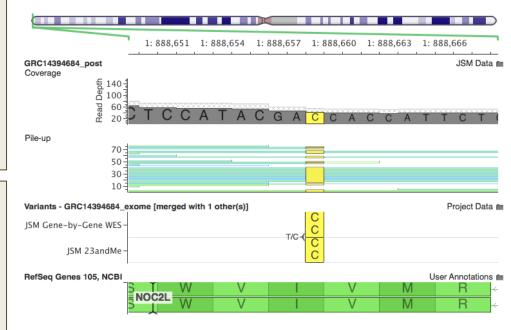
#### Method

- WES data obtained Sep-2014 by Gene-by-Gene, Average 70X Coverage.
- .vcf files loaded into VarSeq using Hereditary Gene Panel Starter Template
- Matched WES variant calls to available SNPs
- Systematically compared **consistencies** and **inconsistencies** for all matched variants
- Applied <u>no</u> quality controls to WES .vcf file
- Compared Read Depths and Genotype Qualities of WES data at consistent vs. inconsistent variant data
- Found one suspect region in WES data (EPPK1)
- Only considered autosomal DNA

#### WES Methodology

- Enrichment: Nextera Rapid Capture Expanded Exome Kit FC-140-1006
- Platform: Illumina HiSeq
- Analysis: Gene By Gene uses the Arpeggi Engine for NGS analytics. They claim the pipeline has been vetted and shown to be more accurate than traditional tools for alignment, variant calling, and variant annotation.

≡ GbG_Hvar / 23&Me_Hvar	* -
≡ homozygous variant (JSM 23andMe)	÷ 🗆
	176,119
	* □
	6,949
	6,949



## Findings: Variant Comparison for WES vs. SNP Chip Data Sets

	r of SNPs	23andMe Genotype Assignment			
in each category		Ref (0/0)	ht (0/1)	Hvar (1/1)	
WES /ariant Calls	ht (0/1)	12	9,662	6	
W Vari Ca	Hvar (1/1)	2	123	6948	

Data type	Value
Total no. aligned variants	16,753
Number of inconsistencies	143
Percent of inconsistent calls	0.85%

## Findings: WES Variant Quality for Consistent vs. Inconsistent SNPs

WES Read Depths		23andMe Genotype Assignment			
		Ref (0/0)	ht (0/1)	Hvar (1/1)	
ES ant IIs	ht (0/1)	122	79	45	
WES Varian Calls	Hvar (1/1)	33	14	73	

WES Genotype		23andMe Genotype Assignment			
Qualitie	25	Ref (0/0) ht (0/1) Hvar (1/1)			
WES Variant Calls	ht (0/1)	2551	1771	71	
	Hvar (1/1)	58	48	1586	

- For most of the inconsistent categories, WES Read Depths and Genotype Qualities reflect poorer data than the consistent categories.
- This is not the case for the WES (0/1) genotype | 23&Me (0/0) genotype
- Could I possibly lead 100+ students through an analysis of their WES data?

### VarSeq in the Wild Student Journal Entries a Few Weeks After VarSeq's Initial Release

## Short QT Syndrome

Short QT syndrone is a condition that can disrupt the hearts natural rhythym. People with short QT syndrome have heart muscle that takes less time than usual to recharge between bests, as detected by an electrocardiogram. Because of this, the part of the heartbest known as the QT interval is abnormally short.

Week 13

If untreated, the arrhythmia associated syndrome on lead to a variety of sympt Jizziness and fainting to cardiac arrest The symptoms can occur from early in when they are most deadly . short QT explain some of the cases of SIDS - A filter container noted is that some people with sh may never experience it's symptoms.

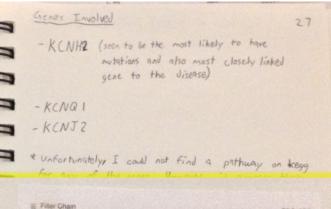
short QT syndrome appears to be me, many feel the condition may be under some people with it never experience

This Jiseque interested me becque may and be present in my famil has had heart problems that have le which makes me water about not only but my mother's as well since she passed it down to her.



These are the filters I used > on Varseq.

on the OR setting with 3 containing filtels for the penes of interest. -A filter to only see genotypes that deviate from the reference sequence.



Filter Container		0-0
E Gene 🏟 1	E Gene 👩 (	E Gene 🌼 1
KCNH2	KCNQ1	KCNJ2
Matches KCNH2 72	Matches KCNQ1 24	Matches KCNJ2 8
Starts with KCNH2 72	C. C	Starts with KCNJ2 8
	Contains KCNQ1 24	Contains KCNJ2 8
	Ends with KCNQ1 24	Ends with KONJ2 8
Missing 727,214	Missing 727,214	Missing 727,214
72	24	
0/1 Genotypes (GT) (Cu	urrent)	104
0/1 Genotypes (GT) (Co	urrent)	104 • -
	urrent)	104 • -
Aatches 1	urrent)	104 • -
Aatches 1 Starts with 1	urrent)	0-
Aatches 1 Starts with 1 Contains 1	urrent)	• -
E 0/1 Genotypes (GT) (Co Matches 1 Starts with 1 Contains 1 Ends with 1	urrent)	• - 0 0 •
Aatches 1 Starts with 1 Contains 1	urrent)	• - 0 0
Aatches 1 Starts with 1 Contains 1 Enda with 1	urrent)	• - 0 0 3 3

### VarSeq in the Wild Student Journal Entries a Few Weeks After VarSeq's Initial Release

no clue what it does . The PLODI putheray is noch D layred / expanded a tot file from my data larger and more complicated thun The other one, so I do not include A (because I don't fully indicated. me yene Refsey 1) 52273289 PLOD1 1:12018290 1:12032030 Mining : I was an mostly working or / anuctshay with the Ref Seq oreners los, NCISI tracter COL3A1 COL5A2 rs6434312 152229495 COL5A2 2:189932831 () Import track (2) set up fittes! COL5A2 rs6434317 1.02 intron variant G/A rs4128538 2:189974794 synonymous variant LOW COL5A2 rs4128539 2:190018099 G/A rs1356168 COL5A2 \*13 P TNXB 6:32020844 Medium 6:32026107 C/T intron variant rs12198173 6:32026808 G/A TNXB 6:32029226 A/G rs185819 6:32050067 T/C 6:32066765 rs13199524 rs3134954 TNXB C/T 6:32071893 1.0% 1512153855 6:32074804 TNXB re2260426 6:32076499 G/A Thought This was pretty cool. I could at all of pote I started w he 447 mas my sup, in multiple gone as the SAUPS that it would more skill, a program could than Re-Se through / look up different SNPL and cualwate ones risk. 2 nots: () pattern at inheritance (gentle). It you have it now Dhere should be a filter for interprete SNPS have A. No cure. But, some forms are mildlishs thread Cie ones that & 23 and the pote as iXXXXX surgery large cuts, and joint strenous activities

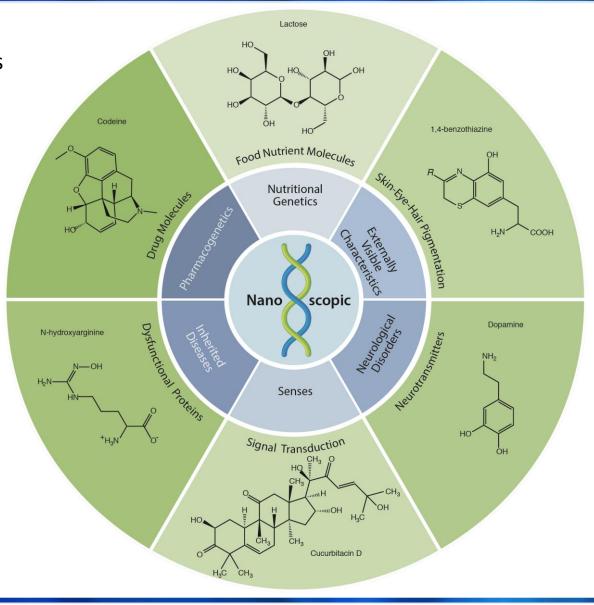
### VarSeq in the Wild Student Journal Entries a Few Weeks After VarSeq's Initial Release

		(69,163 Variants) Filter Chain X +						
		Add . I C Export .						
		Ø - E Filter Chain						
		ClinVitae 2014-02-09, Invitae						
-1. 1	I - ADAK	Reported Classification	Inferred Classification	Gene Prote	in Ch URL			
1Hs	is the 5 \$67 \$15	?	?	?	2 2			
ONC W	re found V.	?	?	?	? ?			
helos	2. But	No Known Pathogenicity	Benign	DSG2	? http://w***			
	eems to have	No Known Pathogenicity	Benign	DSG2	? http://w***			
		No Known Pathogenicity	Benign	DSG2	? http://w***			
no	known mfo	?	?	?	? ?			
abou	nt pathogenition.	?	?	?	? ?			
	h is a little	?	?	?	? ?			
		No Known Pathogenicity	Benign	DSG2	? http://w***			
	contradictory	No Known Pathogenicity	Beni gn	DSG2	? http://w***			
	the scholary	No Known Pathogenicity	Benign	DSG2	? http://w***			
artid	le(which sources	18: 29, 122, 796 18: 29, 122, 798 18: 29, 122, 800 18: 29, 122, 802						
this	i mutation is							
		Variants - GRC14394684_exome			Project Data 🛙			
his	hly important to		A		1			
+	to disease)	JSM CHEM190 file-	G/A C G/A					
			A					
		dbSNP 142, NCBI		Variation and Fun	action - Public Annotations a			
-46.00	is the	1-	rs2278792 +					
-121	nshot of the				and the second s			
scree	nshot of the	C. Hickey I	You are safe!	v · ·.				
Cotted	sponding SNP.	Congratulations!	for one gard.					
		Ref/Alt Identifier Genotypes 0/1 Ge	notymer III	2	No. of Concession, Name of			
14	s is the	T/C ? C_C	1/1	Reported Class				
		C/T ? C_T	0/1	P.	Benign sk Factor			
C	ne.	G/T ? T_T	1/1		athogenic			
		A/G ? G_G	1/1		Benign			
		- Malana al anti-	1					
It saves	me	n bot of n ny using th	ork to	unde	stand u	what the		
	,	, , , , , , , , , , , , , , , , , , , ,	. 1 .	1				
aenotyper h	nean b	m using th	is trol					
your jpes								
V V1		. 1						

University of Illinois at Urbana-Champaign

# **Ongoing and Future Plans**

Prepare teaching modules that connect phenotype with genotype and illustrate concepts in chemistry.



## Acknowledgments

#### **Golden Helix**

- Andreas Scherer
- Andrew Jesaitis
- Bryce Christensen
- Ashley Hintz

#### University of Illinois at Urbana-Champaign

- Milind Basole
- Yury Borukhovich
- Rajat Chadha
- Marilou Landes
- Nicholas Llewellyn
- Doug Mills
- Christian Ray
- Diana Steele

#### University of Kentucky

- Prof. Robert Grossman (U Kentucky)
- Prof. Raphael Finkel (U Kentucky)

