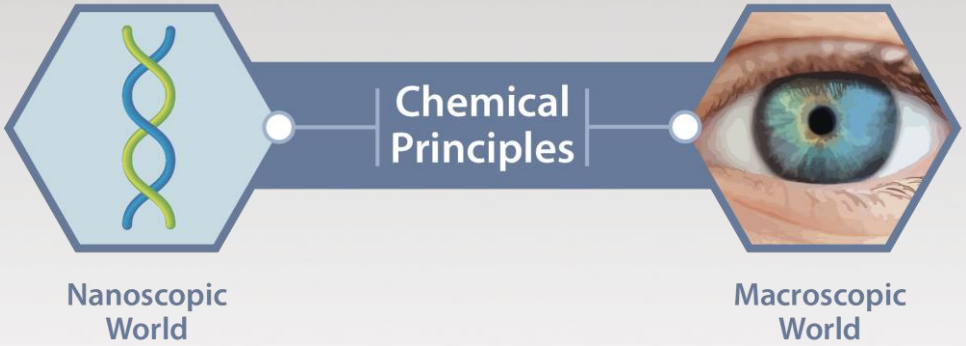
# The Molecular Sciences Made Personal

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?

# 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…

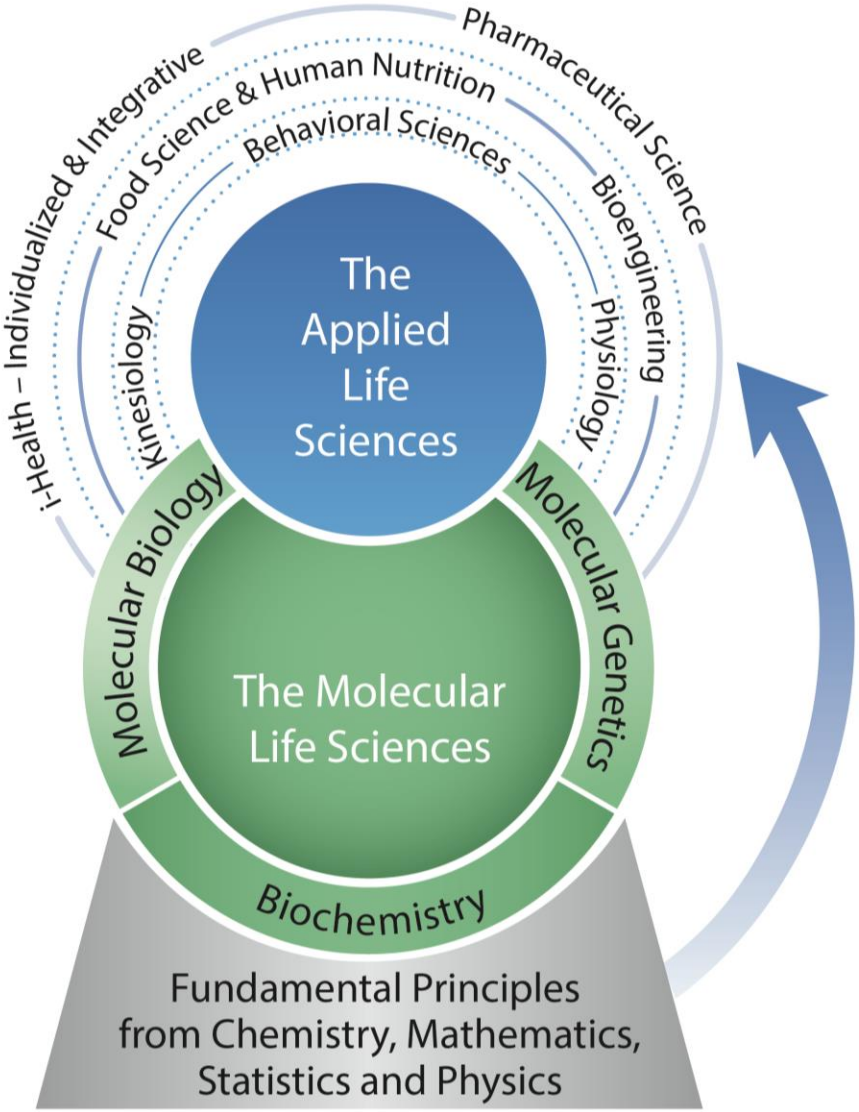
## Preparing the Next Generation of Genetic Researchers

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?



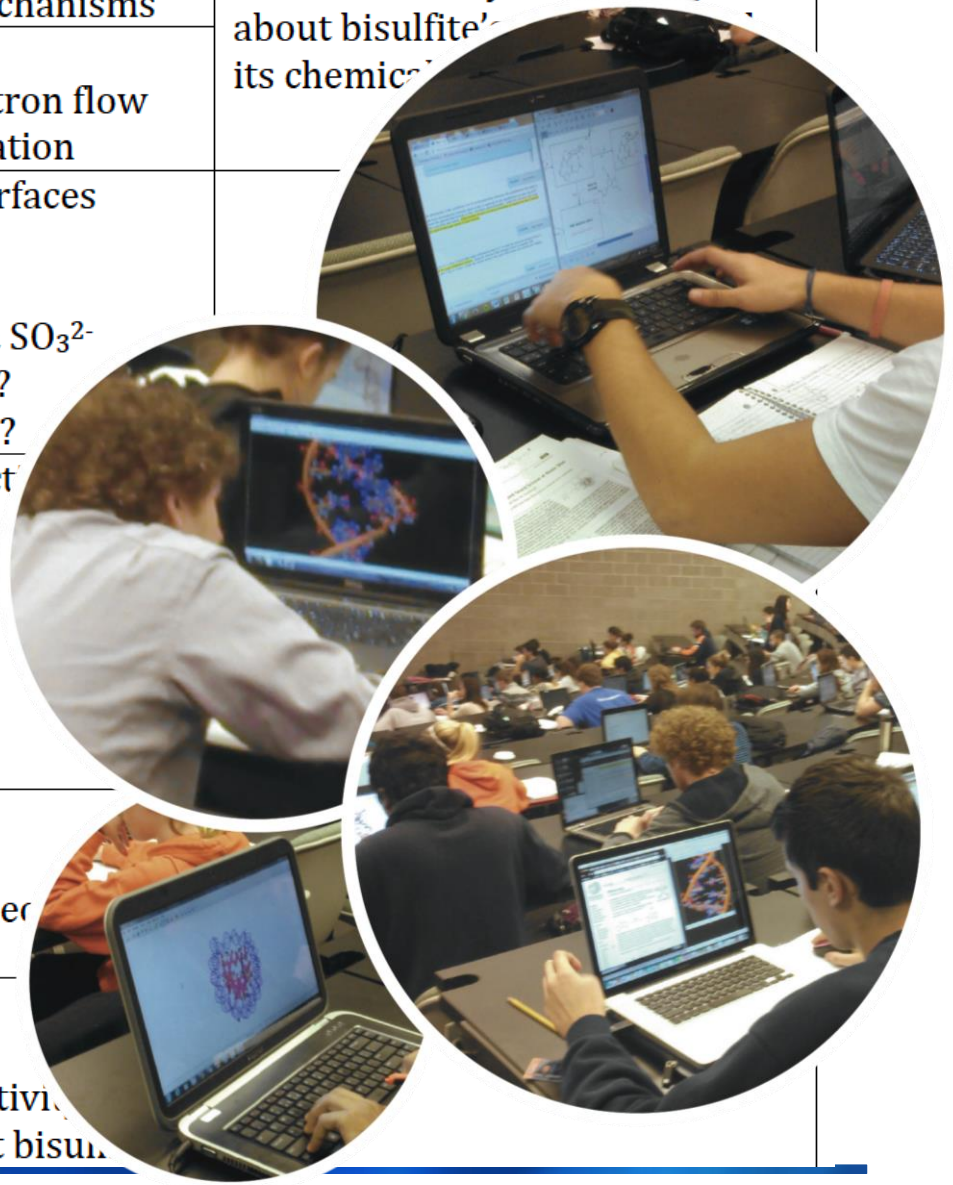

Hypothesis: Learning gains are greater when content is strongly connected to knowledge that students acquire in an aligned field.

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 | • elementary steps & reaction mechanisms |
| Lesson 05<br>T Sep-10 | • proton transfer<br>• general-acid / general-base electron flow<br>• the mechanism of bisulfite formation |
| Lesson 06<br>R Sep-12 | • electrostatic potential energy surfaces (PES)<br>• frontier MOs (FMOs)<br>• FMO energies for $H_2SO_3$, $HSO_3^-$ & $SO_3^{2-}$<br>• which atom is most nucleophilic?<br>• which atom is most electrophilic? |
| Lesson 07<br>T Sep-17 | • charge– vs. FMO–controlled react<br>• filled / empty interactions<br>• generalized σ–type & π–type int |
| Lesson 08<br>R Sep-19 | • π MOs<br>• π delocalization<br>• aromaticity<br>• aromatic heterocycles |
| Lesson 09<br>T Sep-24 | • reactions of heteroaromatics<br>• the π MO as a nucleophile<br>• the nonbonding (*n*) MO as a nucle<br>• the π* MO as an electrophile |
| Lesson 10<br>R Sep-26 | • the nucleobases<br>• tautomeric forms & equilibria<br>• nucleophilic & electrophilic reactivi<br>• hydrolytic deamination (without bisu |

A chemist's way of thinking
about bisulfite'
its chemic

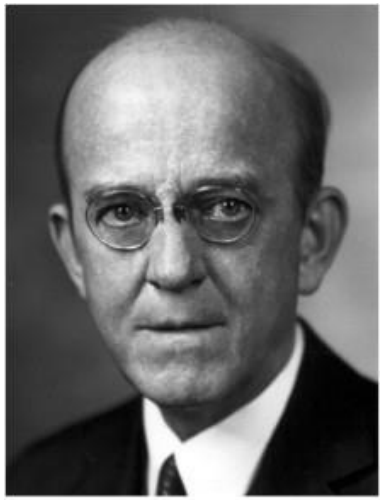


**Building a bridge between molecular genetics and organic chemistry**

# "Studies on the ***chemical*** nature of the substance inducing transformation of Pneumococcal types"

**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.*[1]—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

# The Molecular Sciences Made Personal

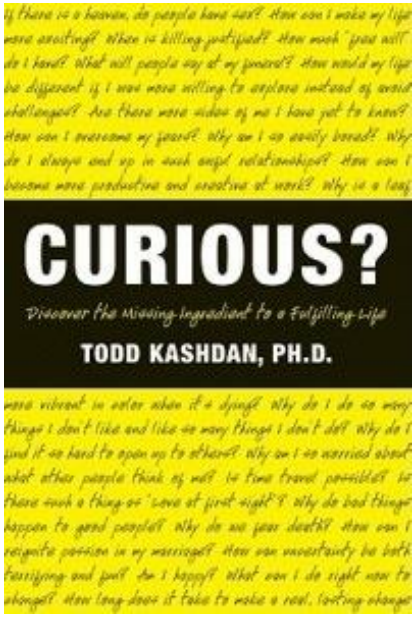
Hypothesis: Learning gains are greater when content is strongly connected to knowledge that students recently acquired in an aligned field.

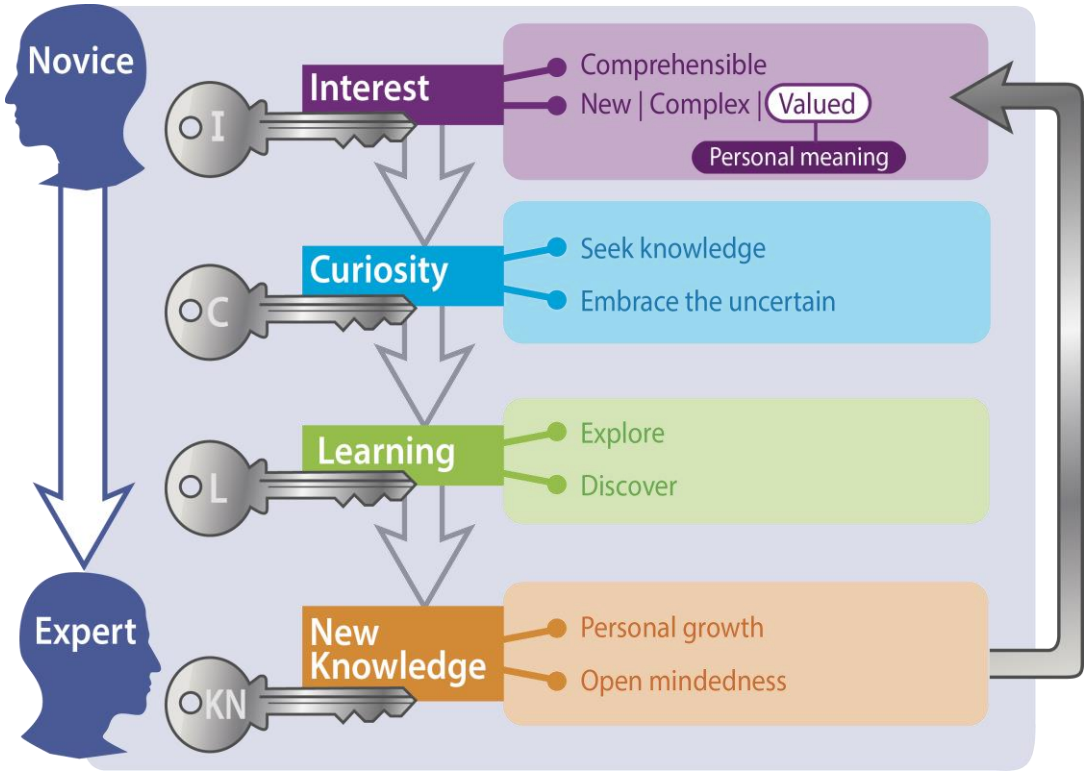
# The Molecular Sciences Made Personal

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

# The Molecular Sciences Made Personal

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.**

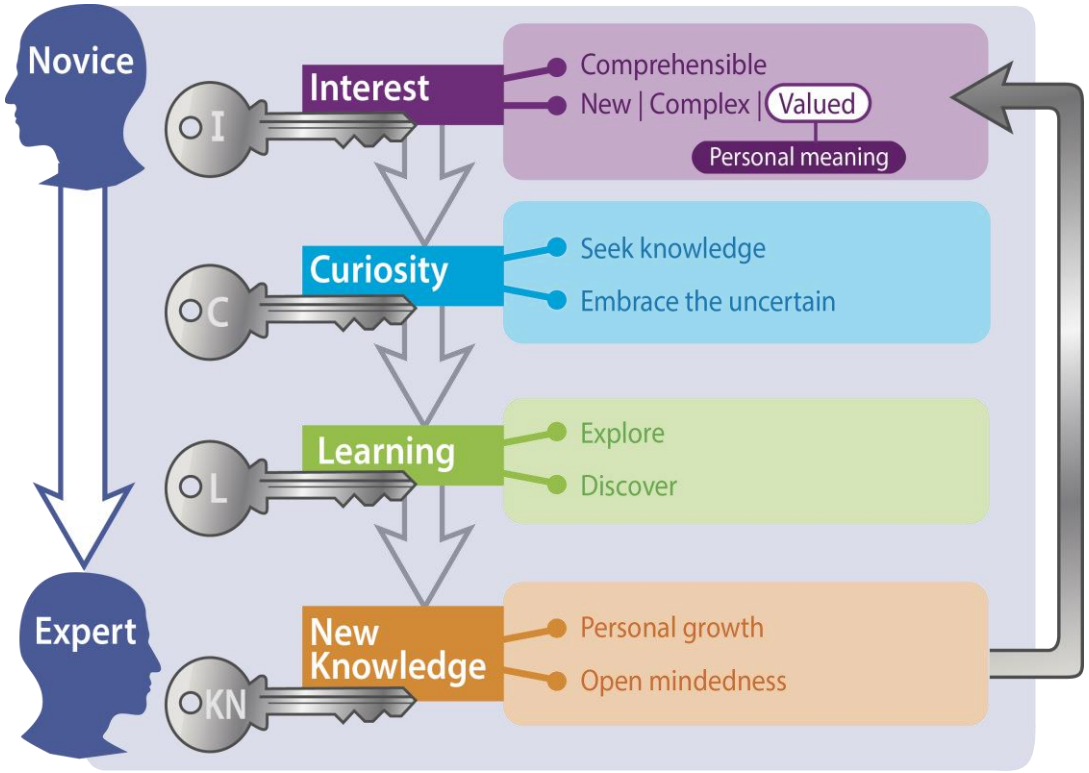


# The Molecular Sciences Made Personal

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.**

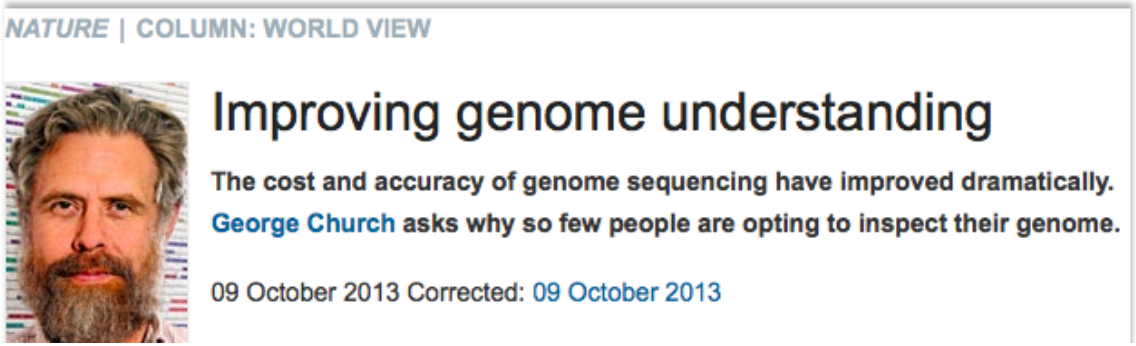
Question: How to deliver **personally meaningful content** to the individual organic chemistry student?



# The Molecular Sciences Made Personal

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?

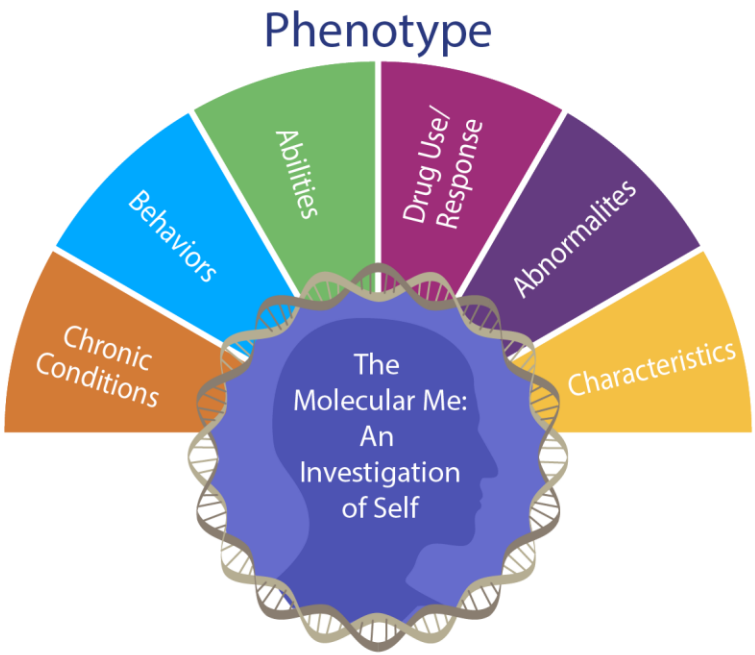

NATURE | COLUMN: WORLD VIEW

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

# Molecular Sciences Made Personal

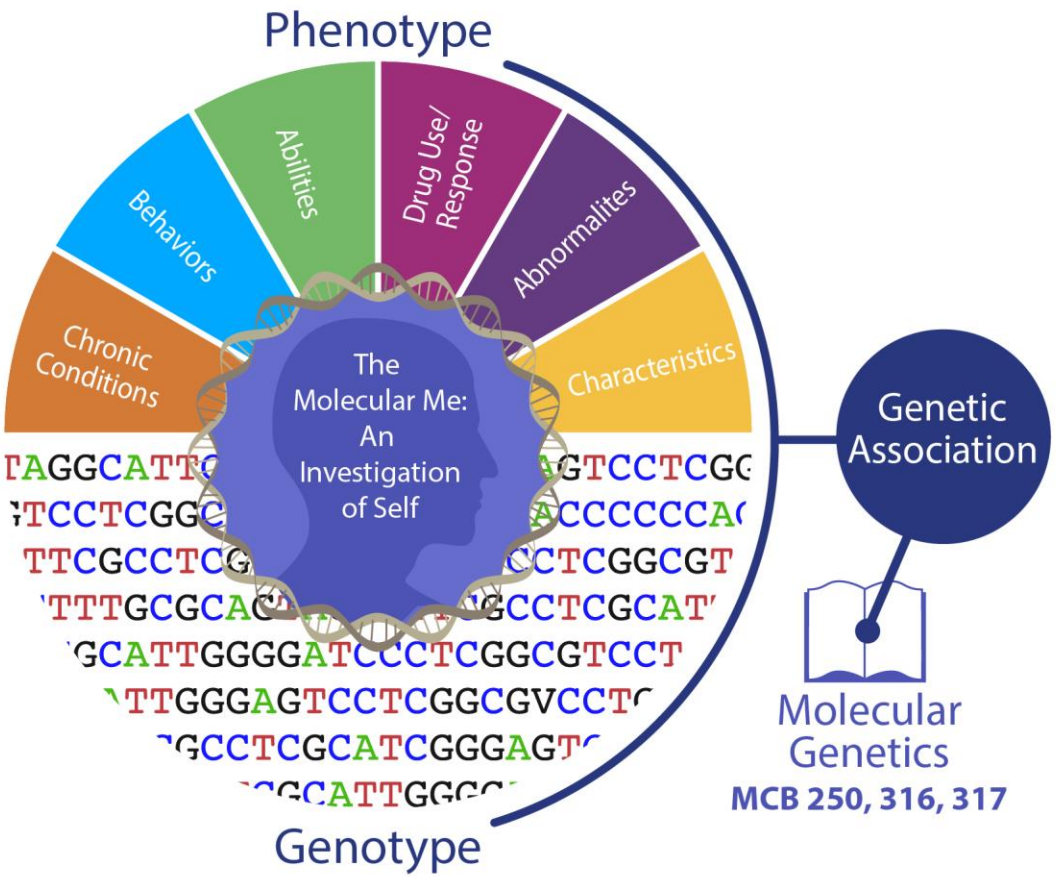


# Molecular Sciences Made Personal


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

**Stronger Connection, Greater Learning**

# Molecular Sciences Made Personal




Genetic Association

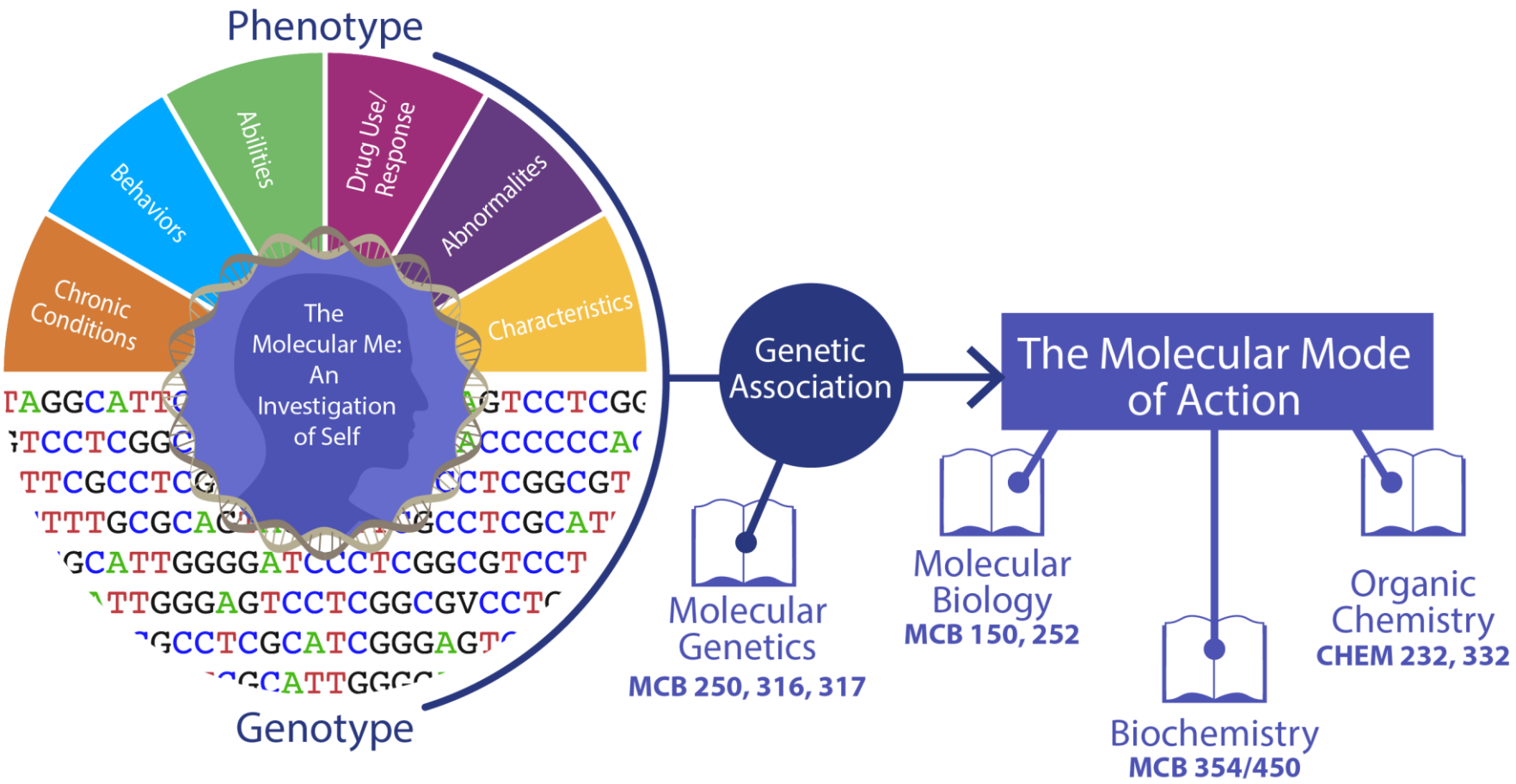
Molecular Genetics

MCB 250, 316, 317

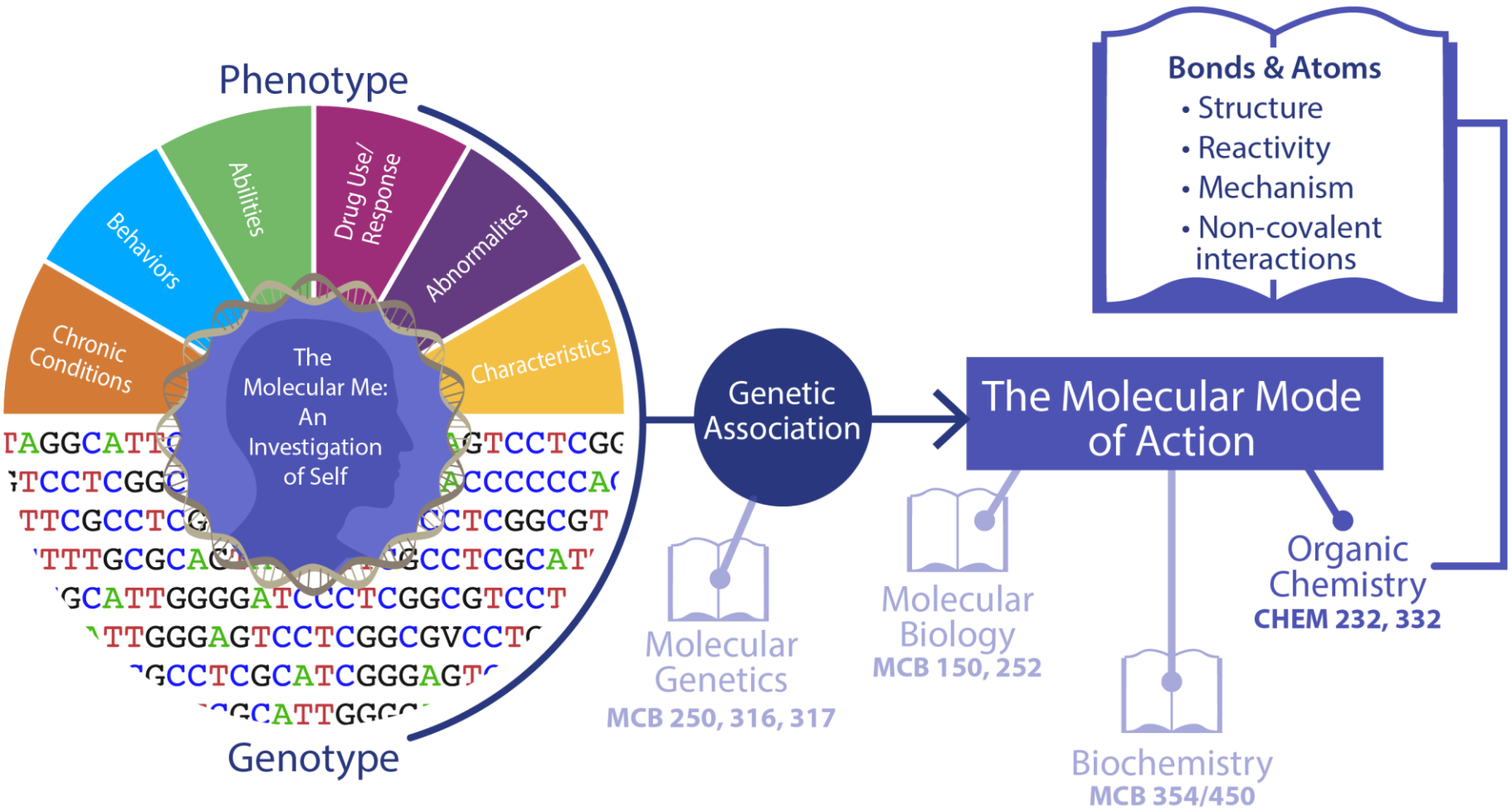
Stronger Connection, Greater Learning

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

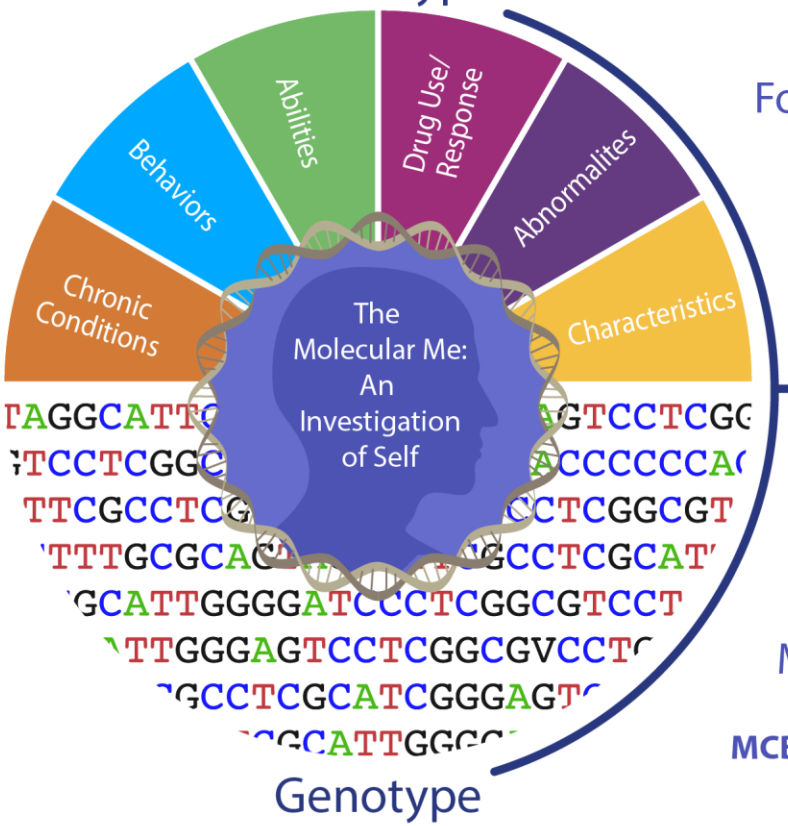
# Molecular Sciences Made Personal



# Molecular Sciences Made Personal



# Molecular Sciences Made Personal




Genetic Association

Molecular Genetics
**MCB 250, 316, 317**

The Molecular Mode of Action

Molecular Biology
**MCB 150, 252**

Biochemistry
**MCB 354/450**

Organic Chemistry
**CHEM 232, 332**

Actionable Response

Drug Dosage

Drug Use & Abuse
**CHLH 243**

Counseling

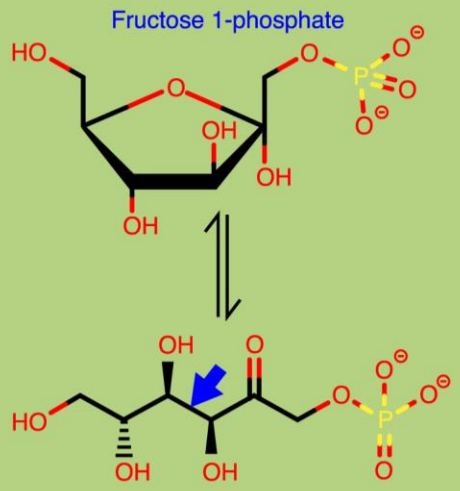
Diet

Food Sci & Nutrition
**FSHN 101**

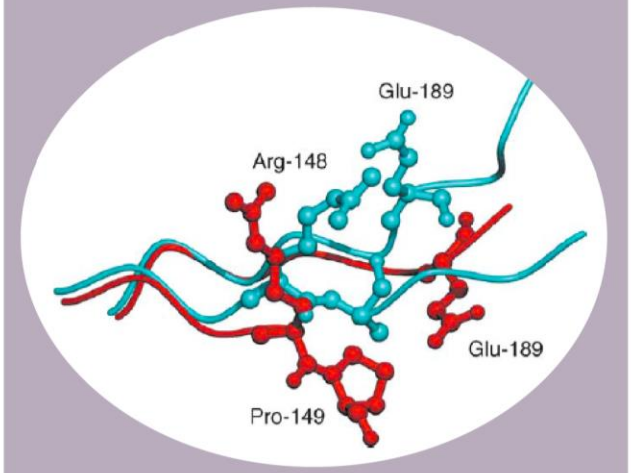
Lifestyle

# A Crosscutting Example of Molecular Sciences Made Personal

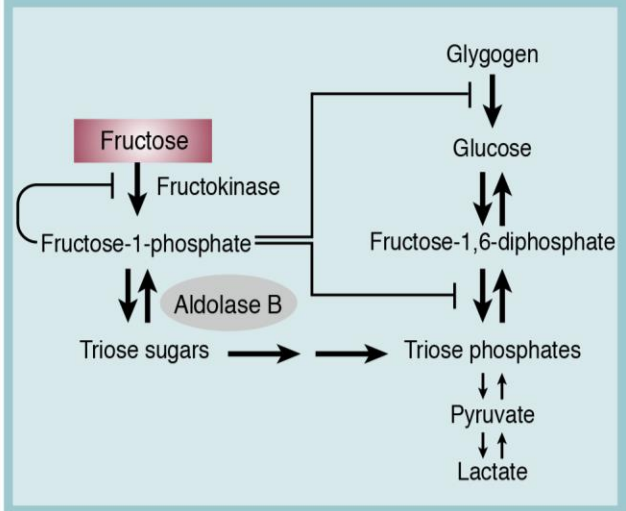
## Organic Chemistry




## Biochemistry



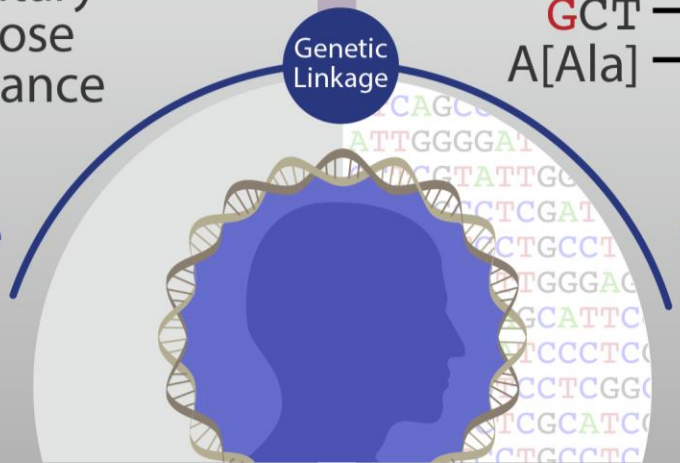

## Molecular Biology




## Investigation

HFI
Hereditary
Fructose
Intolerance

rs1800546 (C;C)
GCT → CCT
A[Ala] → P[Pro]



# Design of Study

Stronger Connection, Greater Learning

**Interventions**

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)
- Develop and teach a sequence of chemistry courses for the intervention group that 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 option 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

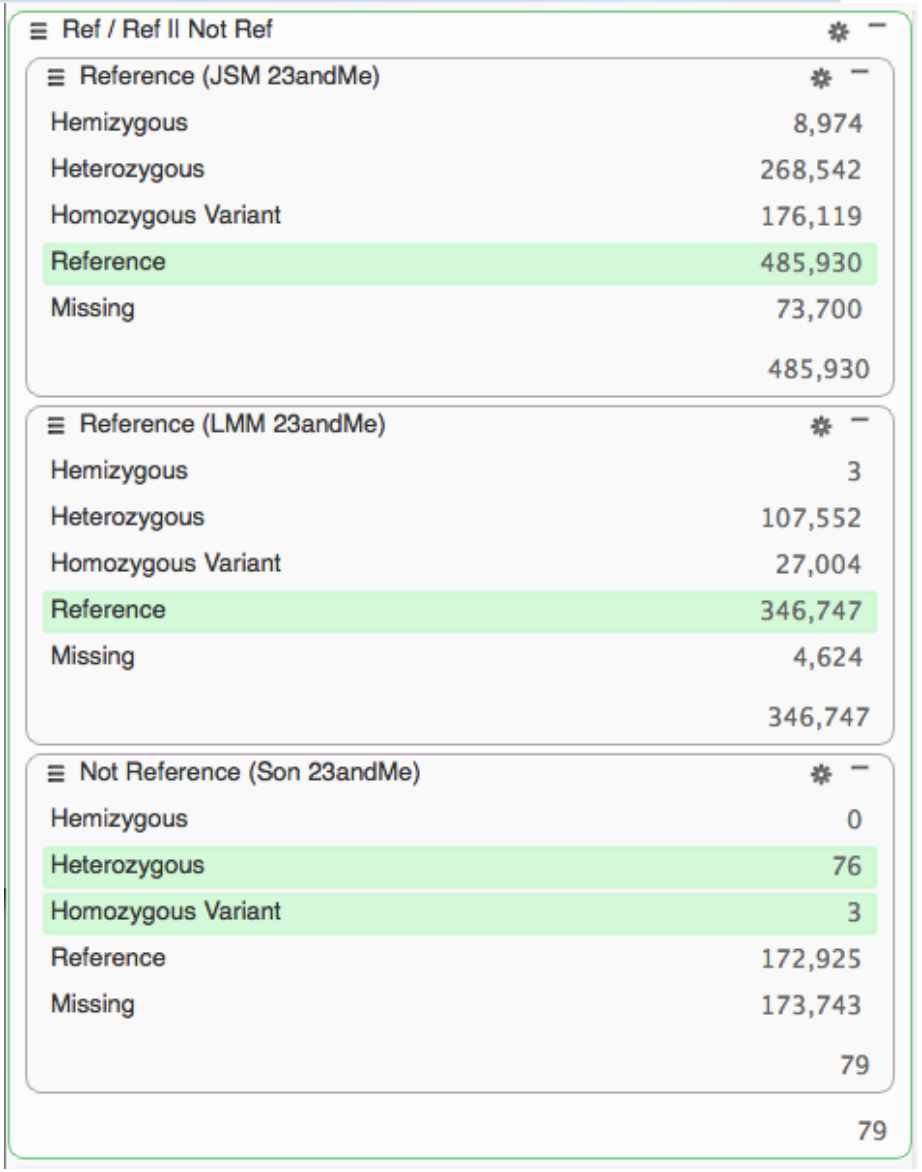
What type of data should be collected?

- DTC SNPs?
- WES?
- Both?

Today we'll conduct data validation analyses using VarSeq

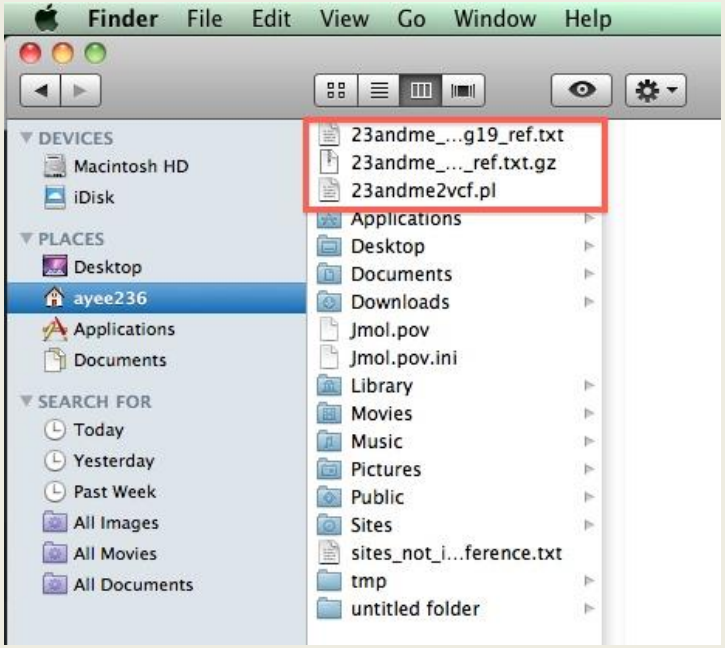
# Data validation by a systematic hereditary analysis in VarSeq

- 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 v4 microarray chip
- Data sheet 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 23andme2vcf converter (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



# Instructions for .txt to .vcf conversion

- **Windows** users: Detailed instructions and an instructional video for the 23andme2vcf converter
- **Mac** users: Detailed instructions and an instructional video 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 to validate genotype data

| | | Mother | | |
|---|---|---|---|---|
| | | Ref (0/0) | ht (0/1) | Hvar (1/1) |
| Father | Ref (0/0) | Ref | Ref or ht | ht |
| | ht (0/1) | Ref or ht | Ref, ht or Hvar | ht or Hvar |
| | Hvar (1/1) | ht | ht or Hvar | Hvar |

Inheritance **inconsistencies** used to identify genotyping errors

| | | Mother | | |
|---|---|---|---|---|
| | | Ref (0/0) | ht (0/1) | Hvar (1/1) |
| Father | Ref (0/0) | Not Ref | Hvar | Not ht |
| | Ht (0/1) | Hvar | | Ref |
| | Hvar (1/1) | Not ht | Ref | Not Hvar |

# Findings from systematic heritability analysis

Inheritance **consistencies** used to validate genotype data

| | | Mother | | |
|---|---|---|---|---|
| | | Ref (0/0) | ht (0/1) | Hvar (1/1) |
| Father | Ref (0/0) | 172,925 | 61,941 | 14,991 |
| | ht (0/1) | 61,583 | | 33,190 |
| | Hvar (1/1) | 15,068 | 34,308 | 48,202 |

Inheritance **inconsistencies** used to identify genotyping errors

| | | Mother | | |
|---|---|---|---|---|
| | | Ref (0/0) | ht (0/1) | Hvar (1/1) |
| Father | Ref (0/0) | 79 | 5 | 17 |
| | Ht (0/1) | 6 | | 7 |
| | Hvar (1/1) | 7 | 1 | 31 |

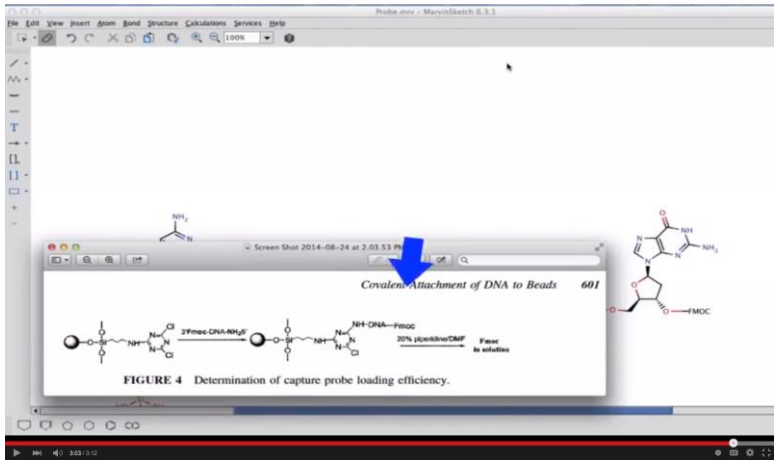
# Summary of results for systematic heritability analysis

Percentage of **inconsistent SNPs** in the trio data set.

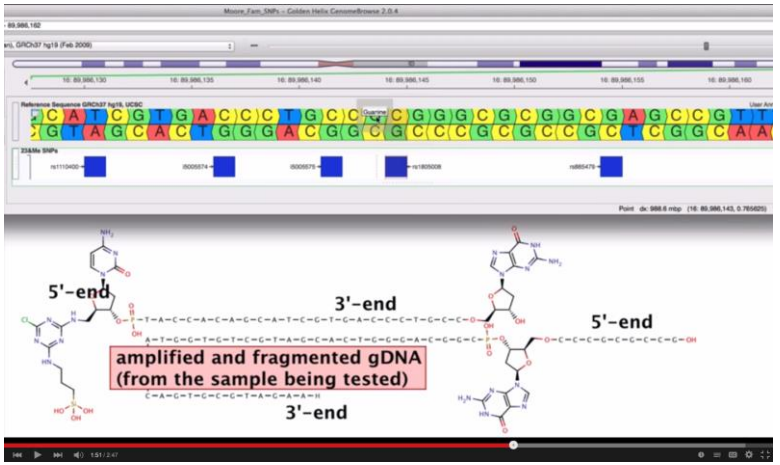
| | | Mother | | |
|---|---|---|---|---|
| | | Ref (0/0) | ht (0/1) | Hvar (1/1) |
| Father | Ref (0/0) | 0.05% | 0.008% | 0.11% |
| | ht (0/1) | 0.01% | | 0.02% |
| | Hvar (1/1) | 0.05% | 0.003% | 0.06% |

- The quality of the SNP data from 23andMe is **high** 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 GenomeBrowse to MarvinSketch (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)


1. Frank J Steemers, Weihua Chang, Grace Lee, David L Barker, Richard Shen & Kevin L Gunderson **Whole-genome 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™ 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

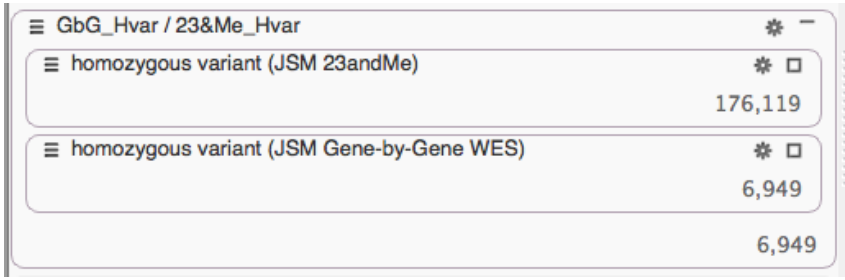
# Comparison of Variants from WES Data with SNP Chip Microarray Data Using VarSeq

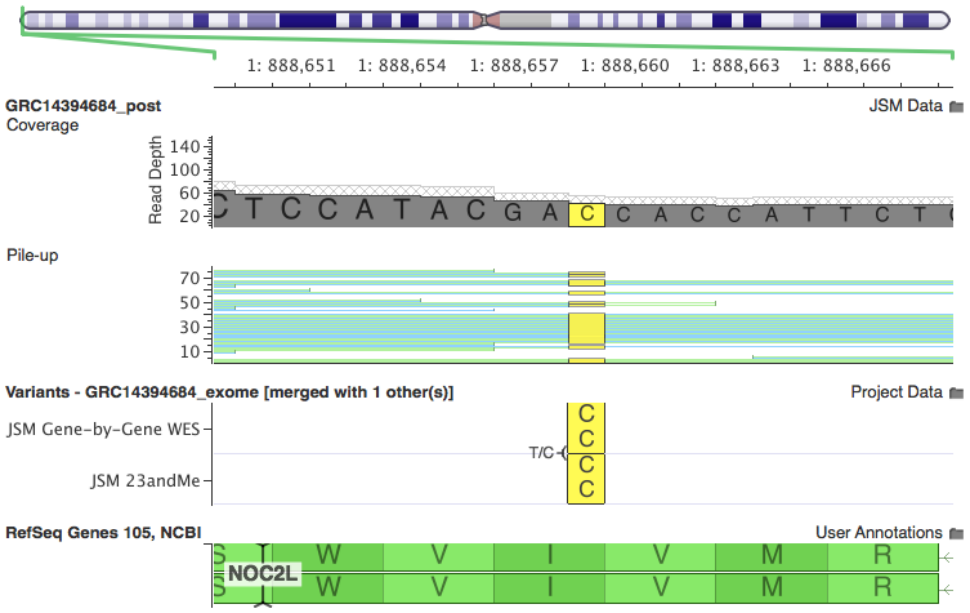
## Method

- 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 **no** 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.



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

| Number of SNPs in each category | | 23andMe Genotype Assignment | | |
|---|---|---|---|---|
| | | Ref (0/0) | ht (0/1) | Hvar (1/1) |
| WES Variant Calls | ht (0/1) | 12 | 9,662 | 6 |
| | 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% |

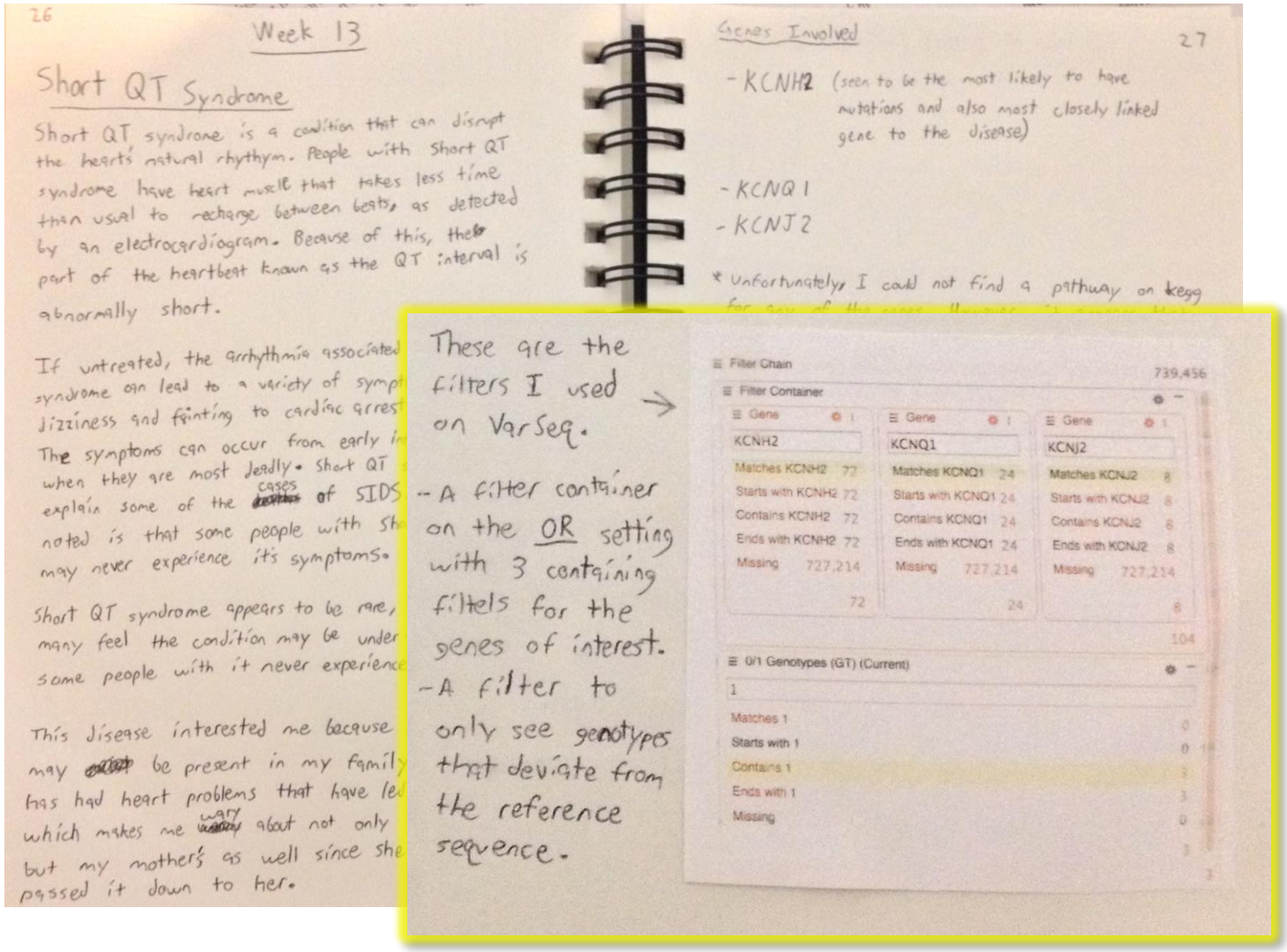
# Findings: WES Variant Quality for Consistent vs. Inconsistent SNPs

| WES **Read Depths** | | 23andMe Genotype Assignment | | |
|---|---|---|---|---|
| | | Ref (0/0) | ht (0/1) | Hvar (1/1) |
| WES Variant Calls | ht (0/1) | 122 | 79 | 45 |
| | Hvar (1/1) | 33 | 14 | 73 |

| WES **Genotype Qualities** | | 23andMe Genotype Assignment | | |
|---|---|---|---|---|
| | | 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



26

Week 13

Short QT Syndrome

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

If untreated, the arrhythmia associated syndrome can lead to a variety of sympt dizziness and fainting to cardiac arrest The symptoms can occur from early in when they are most deadly. Short QT explain some of the ~~deaths~~ cases of SIDS noted is that some people with Sh may never experience it's symptoms.

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

This disease interested me because may ~~exist~~ be present in my family has had heart problems that have le which makes me ~~wary~~ wary about not only but my mother's as well since she passed it down to her.

Genes Involved

27

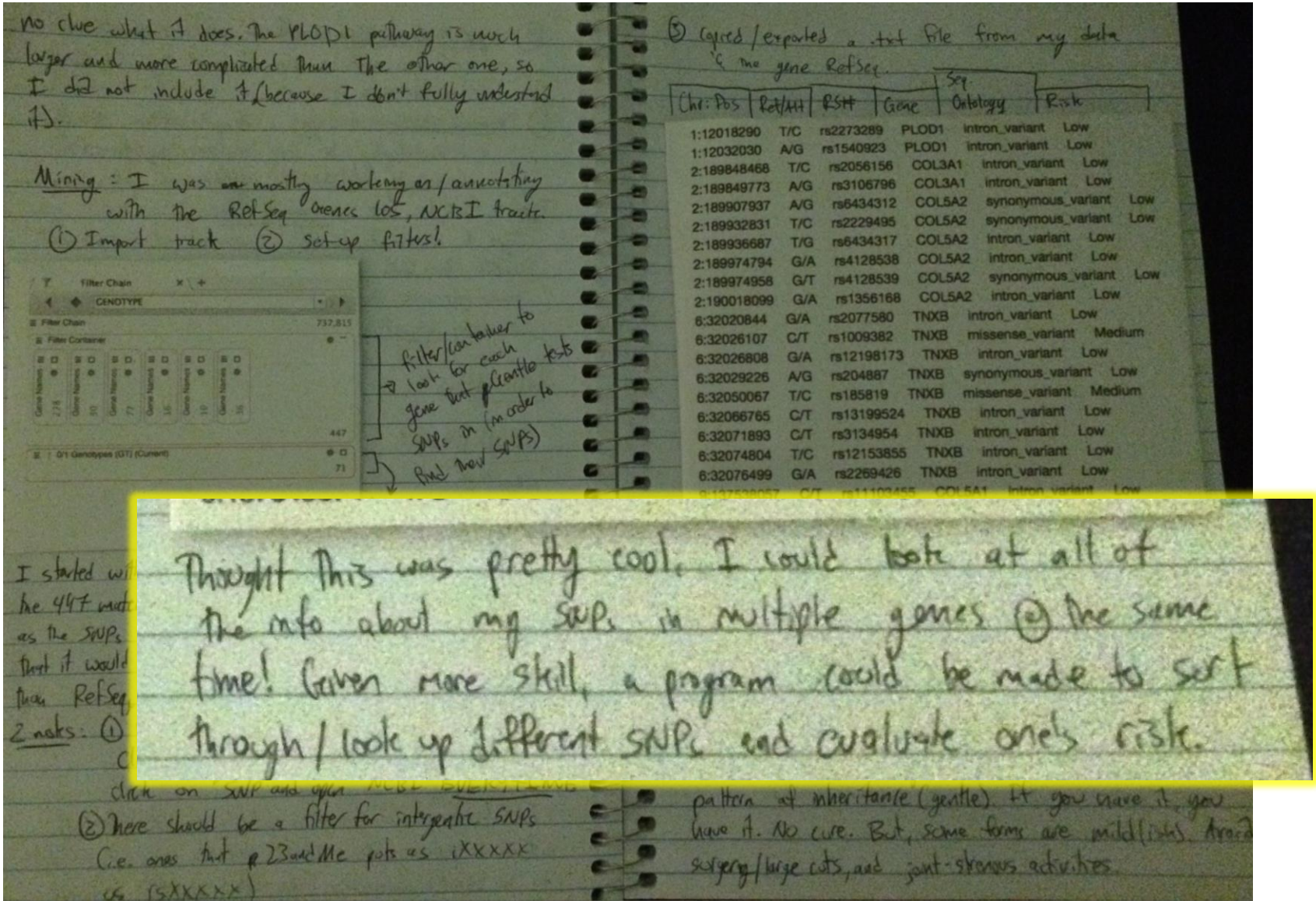
- KCNH2 (seen to be the most likely to have mutations and also most closely linked gene to the disease)
- KCNQ1
- KCNJ2

* Unfortunately, I could not find a pathway on kegg

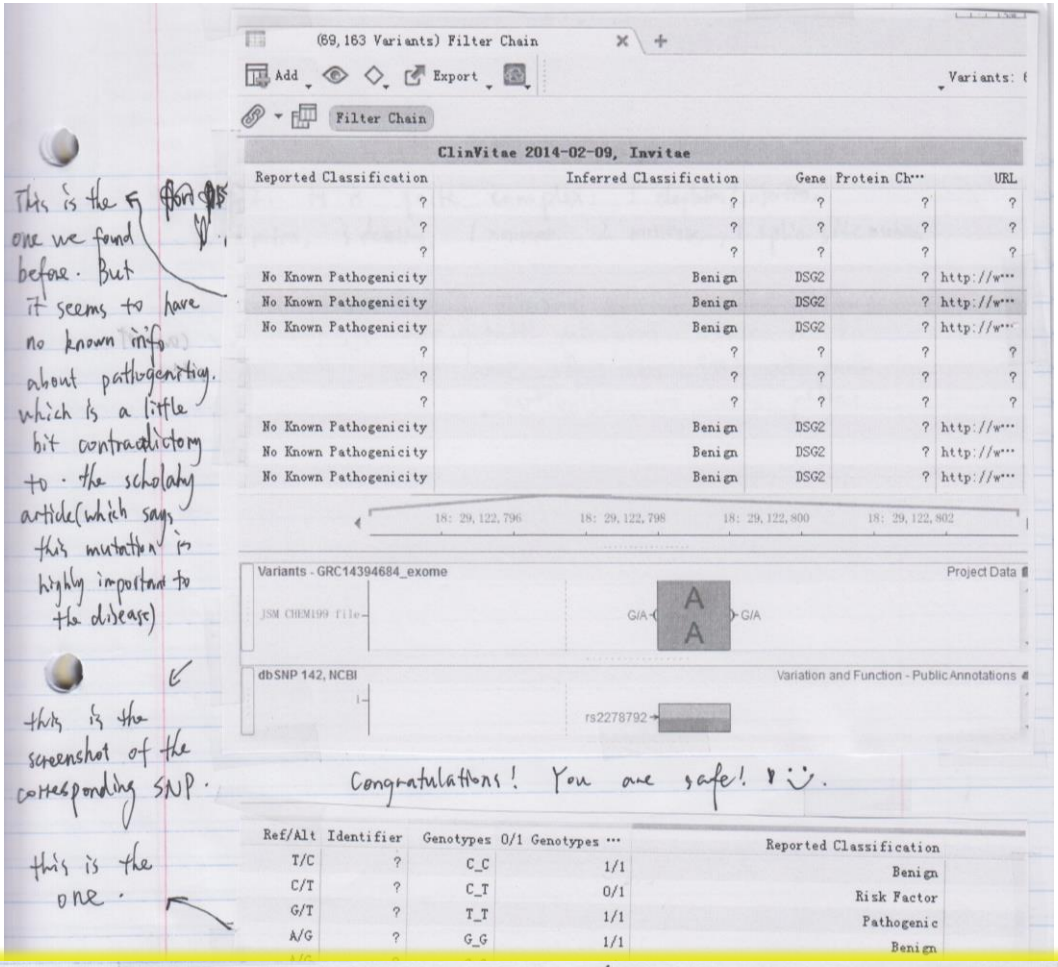
These are the filters I used on VarSeq. →

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

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



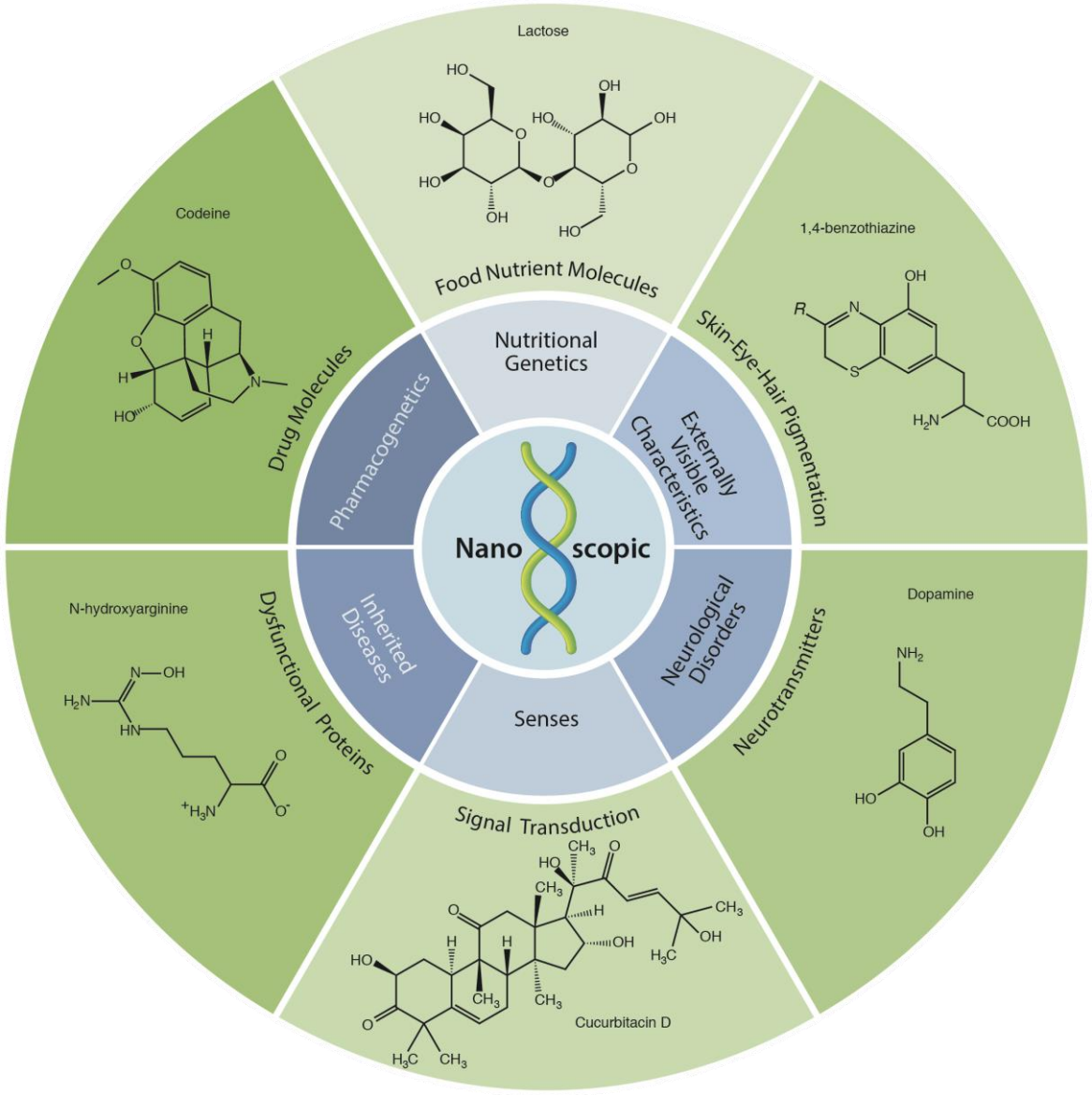
# VarSeq in the Wild
## Student Journal Entries a Few Weeks After VarSeq's Initial Release




It saves me a lot of work to understand what the genotypes mean by using this tool!

# 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)