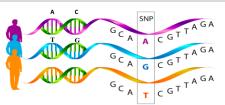# Discovering SNPs and Predicting Phenotype from Genotype

Gaurika Tyagi, Poojan Suri, Shraddha Lanka and Siby Sabu

University of San Francisco

Background

Single Nucleotide Polymorphism (SNPs) are variations at a single point in a specific position in the genetic sequence. Variations in gene that arise are called **Alleles**.

Wild Type (WT): It is the "norm", SNP is a deviation (mutation) from the norm

Mendelian Genetics v/s polygenic nature of eye color.

Impact

Forensic analysis to identify suspects from DNA

Predict risk for human disease phenotypes.

Current Research Status

- Research is limited; Based on specific populations
- SNPs reported in various formats: Allele and Strands
- Classification of Eye Color: 2 vs 3 categories
- SNPedia, dbSNP databases for SNP WT information
- Identified 15+ SNPs shown to affect eye color
- 23andMe and Ancestry.com – Commercial platforms for whole genome sequencing

Data

Data Source (sequence data): openSNP.org - 913 individuals

	SNP1	SNP2	SNPn	Eye Color
human_1	0	1	0	Brown
human_2	1	2	1	Blue

Analysis

Genes Responsible: HERC2 and OCA2
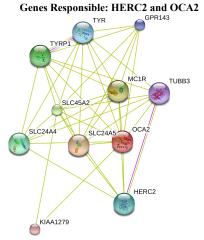

Figure 1: OCA2 Protein Associations

Figure 2: Novel Protein Family Associations with OCA2

Known Interactions	Predicted Interactions	Others
from curated databases	gene neighborhood	textmining
experimentally determined	gene fusions	co-expression
	gene co-occurrence	protein homology

Pre-Processing

Common SNPs + Eye Color dataset: 476 SNPs for 760 users

rsid	user4124	REF	ALT	
rs17399569	GG	G	A	0
rs2487670	TC	C	T	
rs9792946	AG	G	A	
rs10915421	GG	A	G	2
rs4233262	TC	C	T	
rs428001	TC	T	C	1
rs7418357	AC	C	A	
rs11121600	CC	A	G	3
rs7532024	CC	A	G	

476 Features → 124 Features → 53 Features

Genetic Algorithm → Gradient Boosting Machine → Predict Eye color

Analysis

If rs560563 > 0 and rs360490 < 0 then class = "brown"
(CSMD2)   (KIA1522)

Prediction Accuracy: 87%

SNP ID	GENE	FUNCTION/PATHWAY OF THE GENE	ANNOTATION
rs10753239	SDC3	Thought to act as coreceptors, especially for G protein-coupled receptors.	NOG 112892
rs66904931	ARHGEF12	Associated with Leukemia. Related to GTPase activator activity and G-protein coupled receptor binding.	COG 5422
rs3121537	AGBL4	Mutations in this gene are associated with **retinitis pigmentosa.**	COG 2866
rs2803273	AGBL4	Mutations in this gene are associated with **retinitis pigmentosa.**	COG 2866
rs2246105	AGBL4	Mutations in this gene are associated with **retinitis pigmentosa.**	COG 2866
rs5007951	CSMD2	Its expression is elevated in some head and neck cancer cell lines.	NOG 01722
rs10799006	CSMD2	Its expression is elevated in some head and neck cancer cell lines.	NOG 01722
rs560563	CSMD2		NOG 01722
rs360490	KIAA1522		NOG 135987
rs12406072	SLC2A1	Major glucose transporter in the mammalian blood-brain barrier. Functions as a receptor for human T-cell leukemia virus (HTLV) I and II.	COG 0477
rs7546822	LDLRAP1	Lipoprotein metabolism	NOG 17966
rs3790688	DIRAS3	Encodes a member of the ras superfamily associated with growth suppression.	COG 1100

Acknowledgements