# Predicting phenotype from genotype with machine learning

Patricia Francis-Lyon, Shraddha Lanka,
Lakshmi Navin Arbatti, Gaurika Tyagi, Hung Do

University of San Francisco

*SciPy 2017*

UNIVERSITY OF SAN FRANCISCO

CHANGE THE WORLD FROM HERE

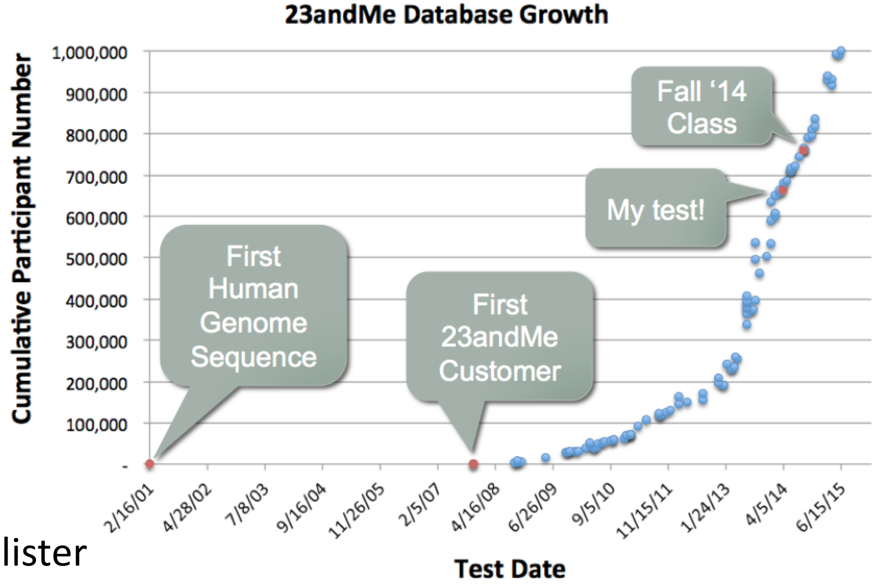
# Opportunity: availability of genomic data

- More genomic data available for predictive analytics: health risk assessment, etc.

- 23andme research:
  - 2,000,000 genotyped customers
  - 85% opt to participate in research

- OpenSNP genomic data is publicly available

- Ancestry.com, others



# Goals for genotype -> phenotype

- A general method to be applied to traits or diseases that have a significant genetic component

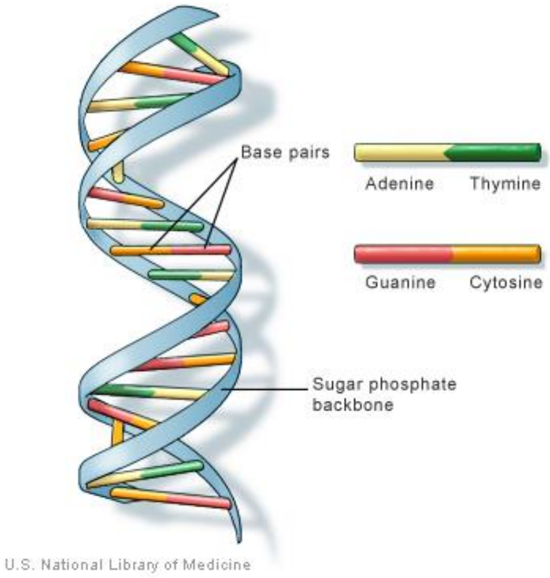
- Agnostic to application: uses no domain knowledge

- Prediction of trait

- Emphasis on **interpretability of results**: detection of novel SNPs, discovery of rules for interactions of genes

- Use publicly available tools and data sources

# Genomic information:
# some basics



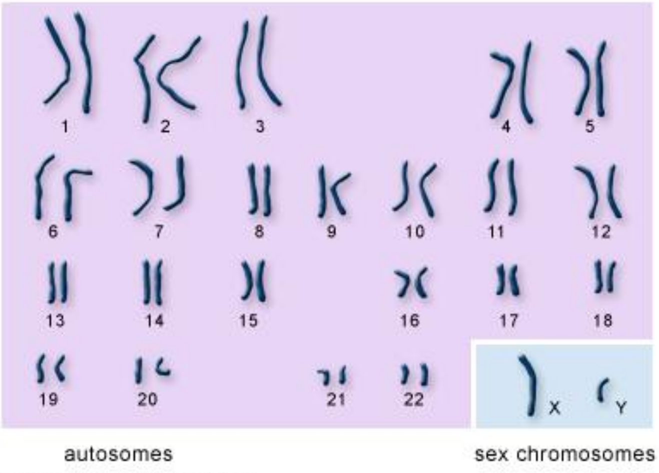
Genomic information is contained in all the cells of the body, encoded as nucleotides A,C,T,G

A person’s genome contains ~ 3 billion of these base pairs

Average difference between DNA of unrelated persons is about 0.1%

There are ~3.8 to 4 million **variants** in a genome: Single Nucleotide Polymorphisms (**SNPs**), insertions, deletions, duplications, inversions

# The human genome



U.S. National Library of Medicine

A person’s genome is organized into 46 chromosomes: 22 are paired autosomes, and 2 sex chromosomes.

Chromosomes are inherited: one of each pair from the mother and one from the father.

Typically a person has 2 copies of each gene, one on each paired chromosome. They have a combined affect on trait and functionality.

# Variants

Variants may occur within a protein coding region, a regulatory region, an RNA gene, or an unknown region

Some variants have no impact on function.  Ex: a gene may contain a silent mutation or loss of function may be compensated by the other copy of the gene.

Some phenotypes (traits) are Mendelian: controlled by whether the inherited gene alleles are dominant or recessive.

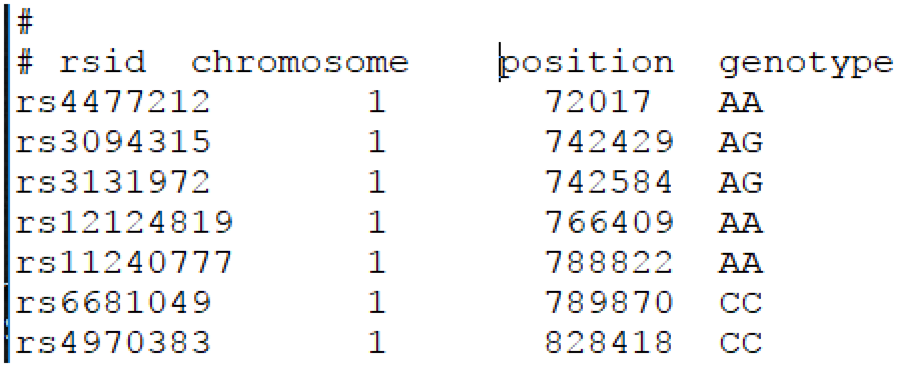
Some traits are polygenic, controlled by more than one gene in more complex patterns

# Known variants

Regions of genome containing prevalent SNPs are typically reported by a sequencing company in variant call format (vcf) or similar plain text.

Nucleotides in the SNP region are reported

From a 23andme file

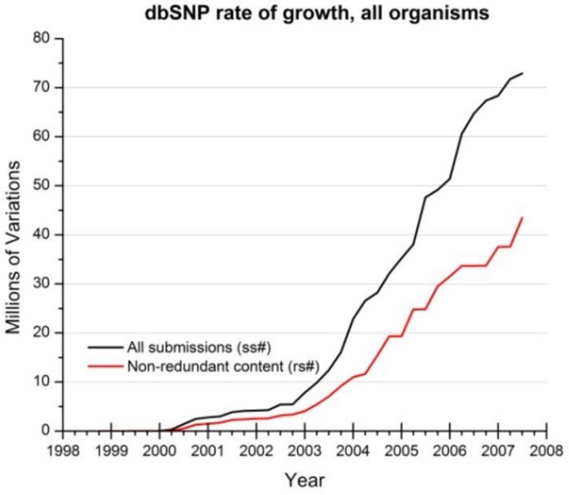


```
#
# rsid  chromosome    position   genotype
rs4477212       1       72017    AA
rs3094315       1       742429   AG
rs3131972       1       742584   AG
rs12124819      1       766409   AA
rs11240777      1       788822   AA
rs6681049       1       789870   CC
rs4970383       1       828418   CC
```

Sequencing is not perfect, called with ~ 1% error rate.

# Tools

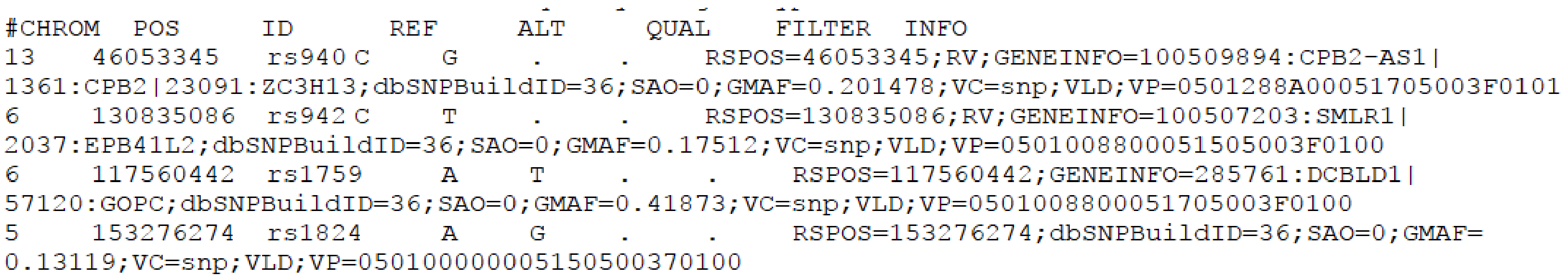
- Human genomic data from OpenSNP, mostly originating from 23andme and Ancestry.com. Full dataset ~1.5 TB. Our dataset of 830 people : 56 GB



SNP FAQ Archive [Internet]

- Information on millions of SNPs from dbSNP in .vcf format - applicable to any human trait: 206 MB
- Python : Numpy, Pandas, Scikit-learn
- R (may use from Python with rpy2)

SNP file from dbSNP in .vcf format



```
#CHROM  POS       ID        REF       ALT       QUAL     FILTER  INFO
13     46053345   rs940 C      G      .      .      RSPOS=46053345;RV;GENEINFO=100509894:CPB2-AS1|
1361:CPB2|23091:ZC3H13;dbSNPBuildID=36;SAO=0;GMAF=0.201478;VC=snp;VLD;VP=0501288A00051705003F0101
6      130835086  rs942 C      T      .      .      RSPOS=130835086;RV;GENEINFO=100507203:SMLR1|
2037:EPB41L2;dbSNPBuildID=36;SAO=0;GMAF=0.17512;VC=snp;VLD;VP=0501008800051505003F0100
6      117560442  rs1759      A      T      .      .      RSPOS=117560442;GENEINFO=285761:DCBLD1|
57120:GOPC;dbSNPBuildID=36;SAO=0;GMAF=0.41873;VC=snp;VLD;VP=0501008800051705003F0100
5      153276274  rs1824      A      G      .      .      RSPOS=153276274;dbSNPBuildID=36;SAO=0;GMAF=
0.13119;VC=snp;VLD;VP=050100000005150500370100
```

# Our Approach

- Build a comprehensive SNP database containing all SNP information from dbSNP joined with gene information

- Read each person file, extract those SNPs that occur in our database

- Reduce the millions of SNPs to a set of 100 or so candidates from which to select features in Supervised learning

- Employ ML tools to classify people by phenotype, rank SNPs, and develop rules for interactions of SNPs.

# Finding Candidate SNPs

- Encode each SNP as number of mutated alleles:
  ( 0, 1 or 2)

- Employ Pandas for ease of merging information on millions of SNPs per person into one data structure per class.

- Calculate aggregate statistics on each SNP on a class level: calculate the percentage of members of each class that have each SNP value.

- Select as candidate features those SNPs that exhibit within-group commonality and between-group differences.

# Machine Learning : scikit-learn

- Split into training and test sets for supervised learning
- Employ random forest feature importance to rank SNPs by importance to the prediction.
  Examine results: are there known or novel SNPs?
- Train a random forest to classify people by phenotype based on their SNPs.
- Employ cross-validation grid search to tune hyperparameters for a decision tree to classify.
- Examine the tree for known or novel relationships.
  Extract rules.

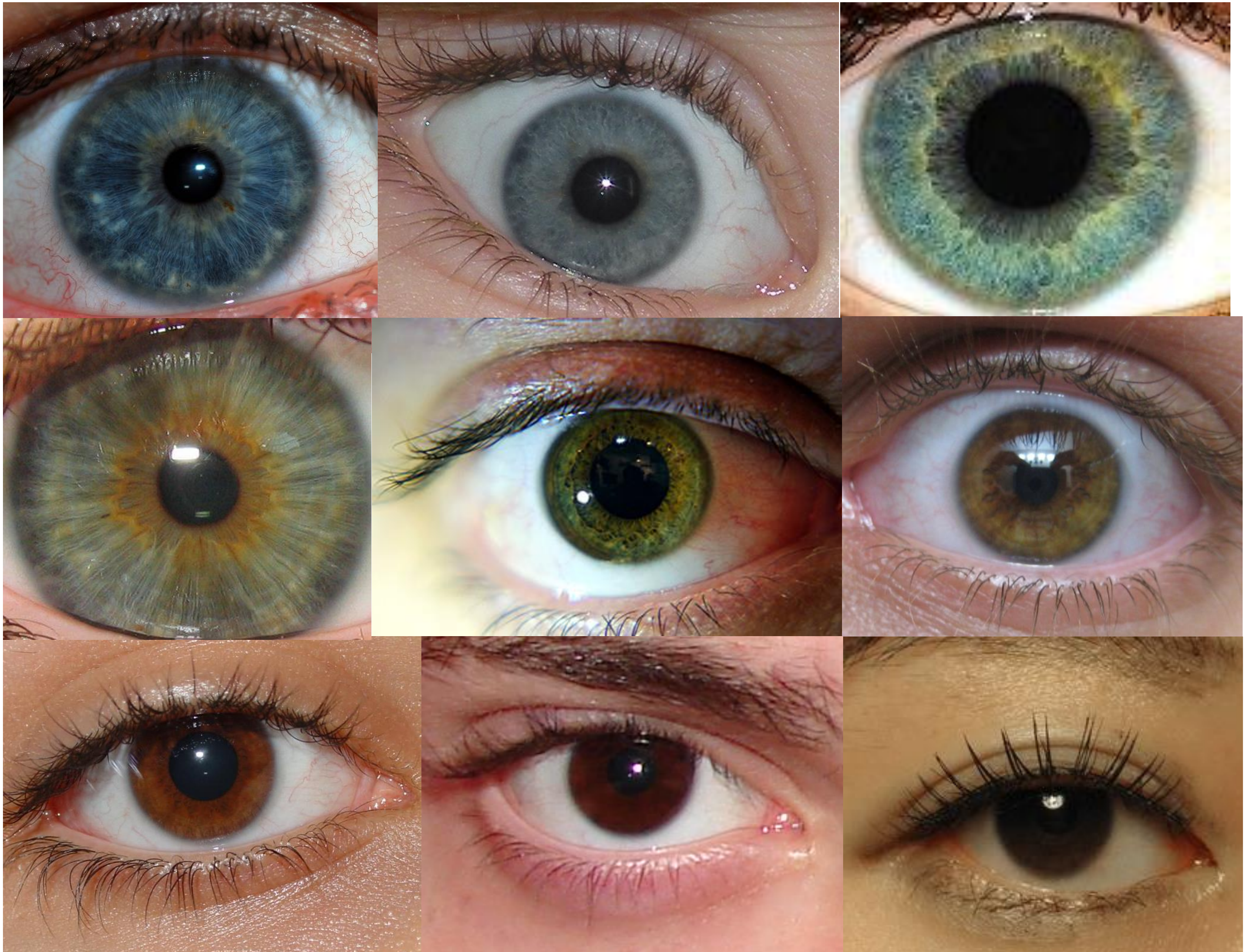
# Logistic Regression modeling

- Import  into R for all people the 15 SNPs that were ranked most important by the scikit-learn random forest:   SNPs as a) integer and  b) factor variables

- To improve regression fit, remove SNPs whose correlation is higher than 0.7. These tend to be in linkage disequilibrium, inherited together

- Detect interactions of SNPs associated with phenotype by fitting a logistic regression model to the person-SNP data.

# Eye color for validation of method



# Eye color

- Eye color has recessive monogenic aspects:

- breaking both copies of the OCA2 gene disrupts/breaks pigment production chain, result is blue eye color

- breaking both copies HERC2 region regulating OCA2 results in OCA2 never activated, stopping pigment protein, blue eye color

- Eye color has polygenic aspects (next slide)

- Limitations of our eye color dataset:

- self-reported data, text descriptions led to judgement calls in assigning class

- no phasing information (haplotypes are unknown), so difficult to detect compound heterozygous and polygenic

# Eye color

Eye color has polygenic aspects:

If OCA2 is broken on one chromosome and HERC2 on the other, the combination can result in blue eyes while neither HERC2 nor OCA2 are fully mutated

Polygenic relationship: consider haplotype from mother (M), from father (F)

| OCA2 | Column1 | HERC2 | Column2 |
|---|---|---|---|
| M | D | M | D |
| **0** | 1 | **0** | 1 |
| 1 | **0** | 1 | **0** |
| 0 | **1** | **1** | 0 |
| **1** | 0 | 0 | **1** |

NB: zeros are unmutated, functional pigmentation genes

Bottom row:

D allele has functioning OCA2, but doesn't get turned on due to broken HERC2

M allele has functioning HERC2, but it regulates for a broken protein

Next higher row is vice-versa

# Random Forest results

Brown and blue-green eye color were predicted with 89% accuracy using no domain knowledge

Validation of method: RF top 30 SNPs and their genes:

Of the millions of human SNPs and tens of thousands of genes, all genes of the top 30 SNPs (HERC2, OCA2, SLC24A4, IRF4, SLC45A2, TYRP1, TYR) are known to be involved in eye color, and most of the top SNPs appear in the literature as having predictive value

# Random Forest

## Top 10 SNPs in order of feature importance

| Gene: SNP | Importance |
|---|---|
| HERC2:rs12913832 | most important: Han et al, Sturm et al, Eiberg et al, patent SNP, etc. |
| HERC2:rs1667394 | Rotterdam Study patent SNP, Kayser et al, Sulem et al, Sturm et al |
| HERC2:rs8039195 | Kayser et al, Candille et al, substitute in patent for rs1667394 |
| HERC2:rs11636232 | Mengel-From et al, Eiberg et al |
| OCA2:rs4778241 | Rotterdam Study patent associated SNP, Kayser et al, Eiberg et al |
| HERC2:rs16950987 | Rotterdam Study patent substitute for rs7495174, Kayser et al |
| HERC2:rs3935591 | Rotterdam Study patent substitute for rs1667394, Eiberg et al, |
| SLC24A4:rs12887171 | unknown clinical significance, 51208 bps from patent SNP rs12896399 |
| OCA2:rs7495174 | Rotterdam Study patent SNP, Kayser et al, Sulem et al, Duffy et al |
| OCA2:rs7174027 | Mengel-From et al, SNPedia, substitute in patent for rs7495174 |

Patent: Method for prediction of human iris color US 20110312534 A1

**Random Forest Confusion Matrix**

| | predicted | |
|---|---|---|
| class | blue | brown |
| blue | 119 | 18 |
| brown | 15 | 124 |

Sensitivity: 0.892
Specificity: 0.869
**Accuracy: 0.88**

# Decision Tree

To avoid overfitting, hyperparameters were tuned using **cross validation grid search** on the training set

```
param_grid={"criterion": ["gini", "entropy"],
        "min_samples_split": [.01, .015, .02, .025],
        "max_depth": [None, 4, 5],
        "min_samples_leaf": [.0025, .005, .01,.015],
        "max_features":[.3,.4,.5,]
            }
```

## Decision Tree Confusion Matrix Comparison

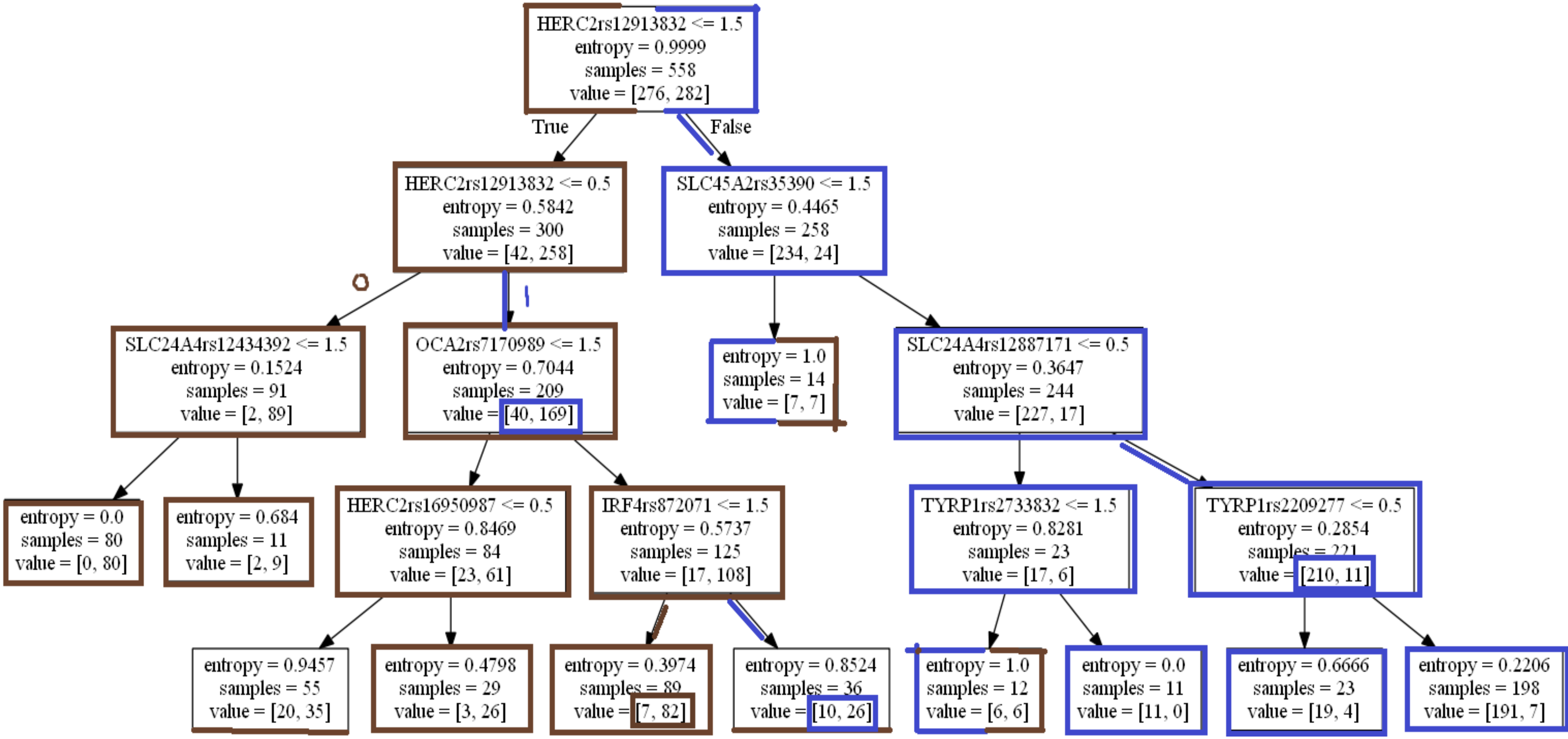
| Training set | | Test set | |
|---|---|---|---|
| 234 | 42 | 122 | 15 |
| 24 | 258 | 14 | 125 |

Sensitivity: 0.915
Specificity: 0.848
**Accuracy: 0.882**

Sensitivity: 0.899
Specificity: 0.891
**Accuracy: 0.895**

# Decision Tree

Nodes colored as predominantly blue or brown



# Decision Tree

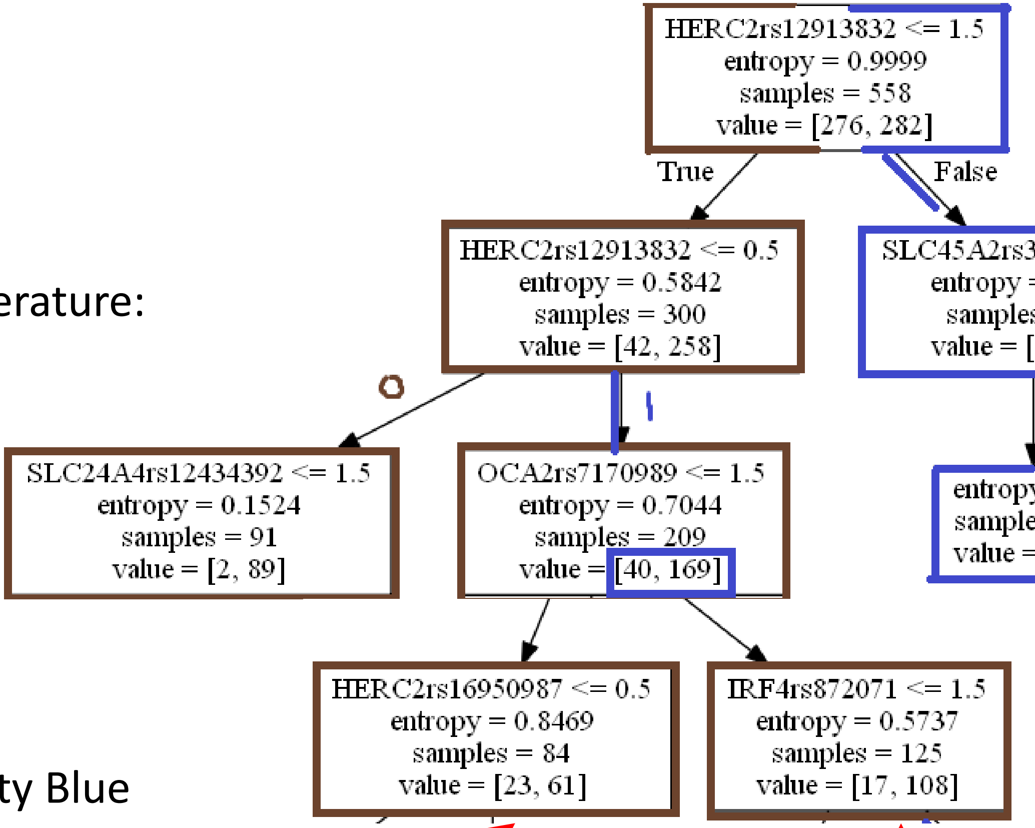
The general consensus in the literature is that **HERC2 rs12913832** alone is the most important SNP for eye color prediction. A SNP value of fully mutated (2 mutated alleles) is indicative of blue eye color

**HERC2 rs12913832** appears at the top of the decision tree and divides nodes broadly into those that are primarily blue (right-hand side: **rs12913832** =2) and nodes that are primarily brown (left-hand side **rs12913832** = 0 or 1)

**HERC2 rs12913832** can be involved in polygenic recessive if both **HERC2 rs12913832** and an **OCA2** gene are partly mutated so as to prevent function of both copies of protein

**Look for evidence** of polygenic relationship on decision tree

# Decision Tree rules



Example in general agreement with literature:
Node where HERC2rs12913832 = 0 is
97.8% probability Brown

Can we find evidence of polygenic?

Caveat: we don’t know haplotype
Even so, consider:
HERC2rs12913832 = 1
OCA2rs7170989 = 2 : 13.6% probability Blue

Contrast with:
HERC2rs12913832 = 1
OCA2rs7170989 = 0 or 1: **27.4% probability Blue**
Node includes **all OCA2rs7170989 = 0** (likely brown)
plus polygenic combinations from previous slide:
**0101,1010,0110,1001** (last 2 are blue)

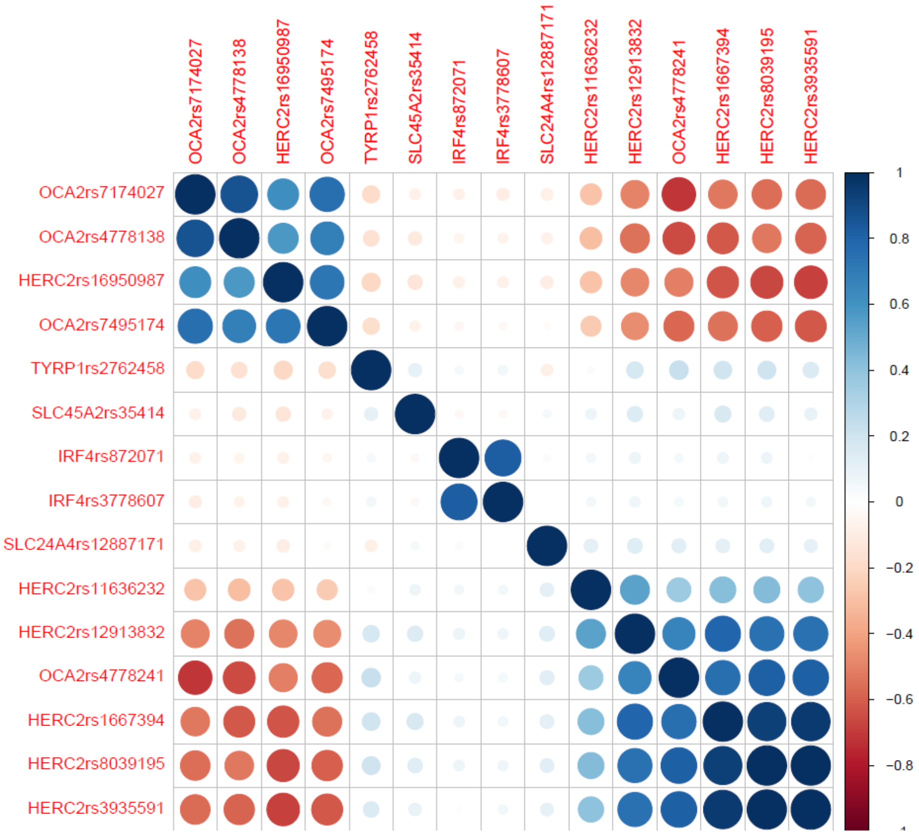
# Logistic Regression on top 15 SNPs

**Imported from all people the 15 most important SNPs (identified by RF)**

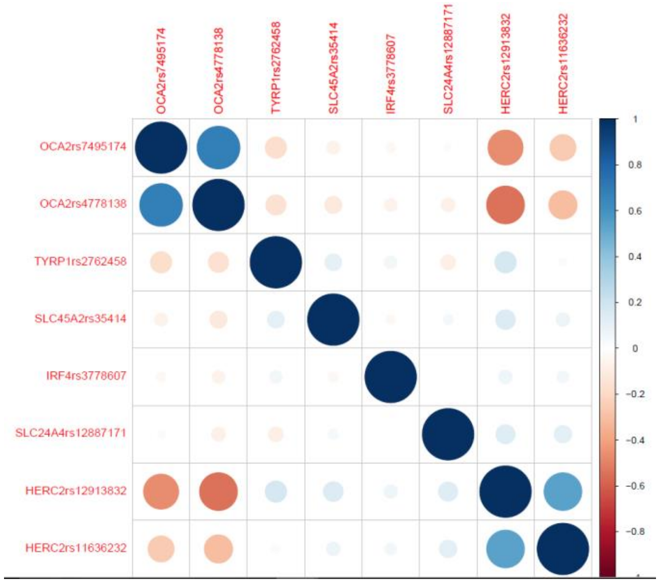
Created a dataset with int encoding of SNPs

Created a dataset with SNPs as  factor levels (like one hot encoding)

Removed 7 highly correlated (cor > .7)  SNPs



Correlation of top 15 SNPs



Correlation of remaining 8 SNPs

# Logistic regression on uncorrelated SNPs as int

**Ran step function to find best model of all combinations of main effects and 2-way interactions: then refit with only significant variables**

**Found 2 significant main effects and 3 significant interactions:**

```
Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                         5.8174     0.4920  11.825  < 2e-16 ***
HERC2rs12913832                    -2.5663     0.3404  -7.540 4.72e-14 ***
TYRP1rs2762458                     -1.1576     0.3727  -3.106 0.001895 **
TYRP1rs2762458:IRF4rs3778607       -0.3548     0.1339  -2.649 0.008077 **
HERC2rs12913832:SLC24A4rs12887171 -0.8384     0.2209  -3.796 0.000147 ***
TYRP1rs2762458:SLC24A4rs12887171    0.7311     0.2341   3.123 0.001789 **
```

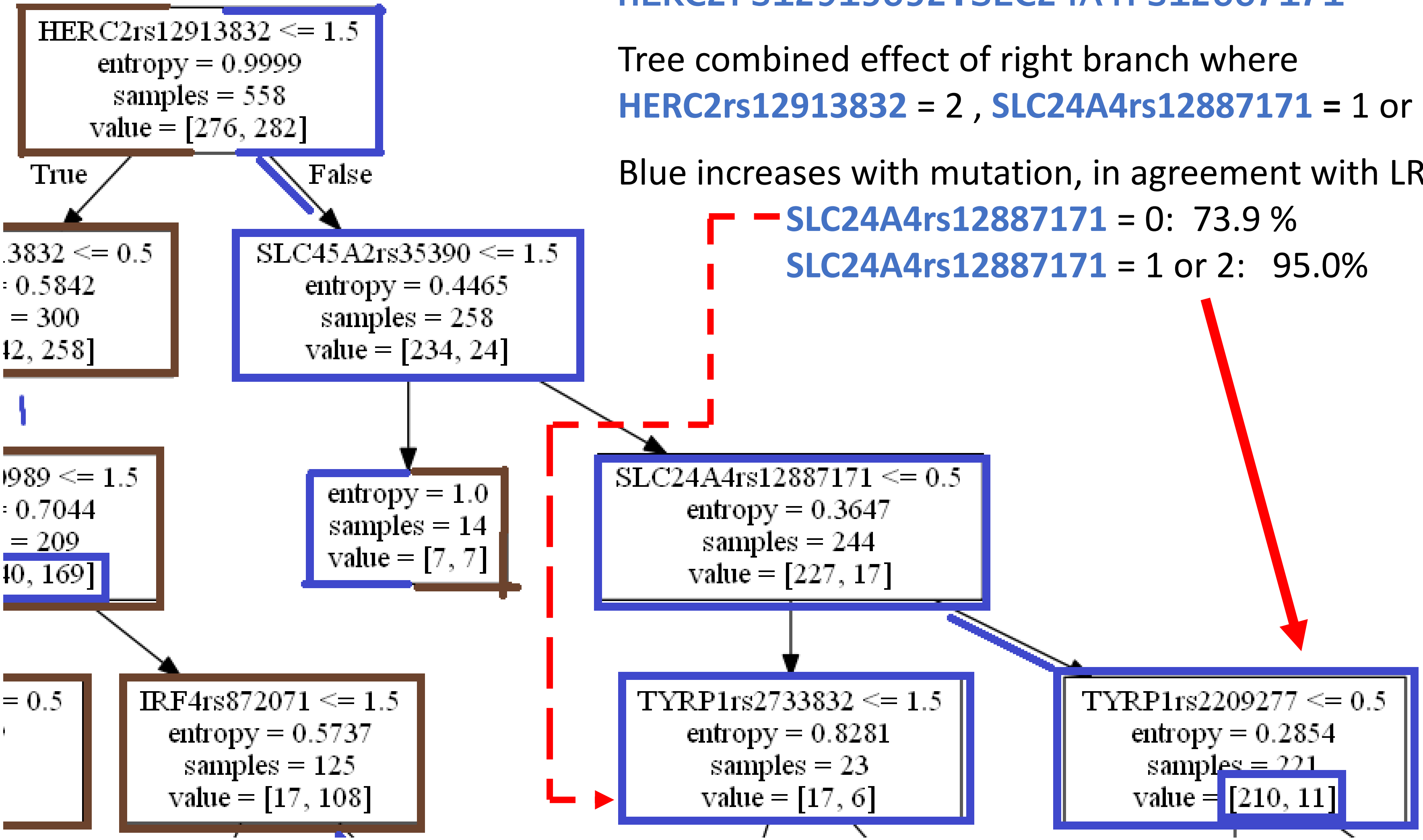
HERC2rs12913832:SLC24A4rs12887171 a unit increase in this value results in a decrease in the log odds of having brown eyes
exp(-0.8384)= 0.43: reduction in odds of having brown eyes by factor of .43

In general agreement with decision tree (next slide)

**Logistic Regression with HERC2 rs1291 as only predictor shows good fit:**
In general agreement with literature and decision tree (previous slide)

Pseudo $R^2$ for logistic regression:
Hosmer and Lemeshow $R^2$: 0.511
Cox and Snell $R^2$: 0.507
Nagelkerke $R^2$: 0.676

# Decision tree comparison with logistic regression



Interaction found in logistic regression (LR):
**HERC2rs12913832:SLC24A4rs12887171**

Tree combined effect of right branch where
**HERC2rs12913832** = 2 , **SLC24A4rs12887171** = 1 or 2

Blue increases with mutation, in agreement with LR:

**SLC24A4rs12887171** = 0: 73.9 %
**SLC24A4rs12887171** = 1 or 2: 95.0%

# Future

- Apply this method to a human disease with a significant genetic component to create a risk assessment tool

- Extend the method to employ elastic net logistic regression for logistic regression with feature selection, XGBoost for decision trees with pruning

# Conclusion

- This is a general method for using supervised learning to predict phenotype from human genomes
- We focus on gaining understanding:  detecting SNPs important to prediction, elucidating interactions and relationships between SNPs and genes
- Testing with a well-studied problem achieved good prediction. We also detected all genes known be implicated in eye color and the SNPs reported to be most influential
- We employ publicly available tools and data in an approach that may be used for the different organisms in dbSNP databases
- Our code will be publicly available

# Sources: Eye color studies for validation of detected SNPs (slide 17)

Manfred Heinz Kayser, Fan Liu, Albert Hofman. Patent: “Method for prediction of human iris color” US 20110312534 A1 (2011), data source is Rotterdam Study, Hofman et al (2007)

Kayser, Manfred, Fan Liu, A. Cecile JW Janssens, Fernando Rivadeneira, Oscar Lao, Kate van Duijn, Mark Vermeulen et al. "Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene." *The American Journal of Human Genetics* 82, no. 2 (2008): 411-423.

Sulem, Patrick, Daniel F. Gudbjartsson, Simon N. Stacey, Agnar Helgason, Thorunn Rafnar, Kristinn P. Magnusson, Andrei Manolescu et al. "Genetic determinants of hair, eye and skin pigmentation in Europeans." *Nature genetics* 39, no. 12 (2007): 1443-1452.

Sturm, Richard A., David L. Duffy, Zhen Zhen Zhao, Fabio PN Leite, Mitchell S. Stark, Nicholas K. Hayward, Nicholas G. Martin, and Grant W. Montgomery. "A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color." *The American Journal of Human Genetics* 82, no. 2 (2008): 424-431.

Eiberg, Hans, Jesper Troelsen, Mette Nielsen, Annemette Mikkelsen, Jonas Mengel-From, Klaus W. Kjaer, and Lars Hansen. "Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression." *Human genetics* 123, no. 2 (2008): 177-187.

Han, Jiali, Peter Kraft, Hongmei Nan, Qun Guo, Constance Chen, Abrar Qureshi, Susan E. Hankinson et al. "A genome-wide association study identifies novel alleles associated with hair color and skin pigmentation." *PLoS genetics* 4, no. 5 (2008): e1000074.

Mengel-From, Jonas, Claus Børsting, Juan J. Sanchez, Hans Eiberg, and Niels Morling. "Human eye colour and HERC2, OCA2 and MATP." *Forensic Science International: Genetics* 4, no. 5 (2010): 323-328.

Mengel-From, Jonas, Terence H. Wong, Niels Morling, Jonathan L. Rees, and Ian J. Jackson. "Genetic determinants of hair and eye colours in the Scottish and Danish populations." *BMC genetics* 10, no. 1 (2009): 88.

Duffy, David L., Grant W. Montgomery, Wei Chen, Zhen Zhen Zhao, Lien Le, Michael R. James, Nicholas K. Hayward, Nicholas G. Martin, and Richard A. Sturm. "A three–single-nucleotide polymorphism haplotype in intron 1 of OCA2 explains most human eye-color variation." The American Journal of Human Genetics 80, no. 2 (2007): 241-252.

Candille, Sophie I., Devin M. Absher, Sandra Beleza, Marc Bauchet, Brian McEvoy, Garrison Nanibaa’A, Jun Z. Li et al. "Genome-wide association studies of quantitatively measured skin, hair, and eye pigmentation in four European populations." *PLoS One* 7, no. 10 (2012): e48294.

# Sources: Eye color images from Wikipedia commons (slide 13):

Top row:
By JDrewes - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=3117810
By Larali21 - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=22281525
By Grodrig1 - Taken with a cell phone camera, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=18953885

Middle row:
By Picture taken by Matthew Goldthwaite. - 영국말 위키백과에 있는 en:Image:Humaniris.jpg, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=1622680
By Rick - Owner, CC0, https://commons.wikimedia.org/w/index.php?curid=13400572
By Manicjedi - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=16812864

Bottom row:
CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=68654
By Arctice at the English language Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=6910071
By Cecikierk - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=12347264