Veridical Data Science for Biomedical Research:
detecting epistatic interactions via epiTree

Bin Yu

Statistics and EECS, UC Berkeley

WNAR Webinar

Feb. 26, 2021

ve·rid·i·cal

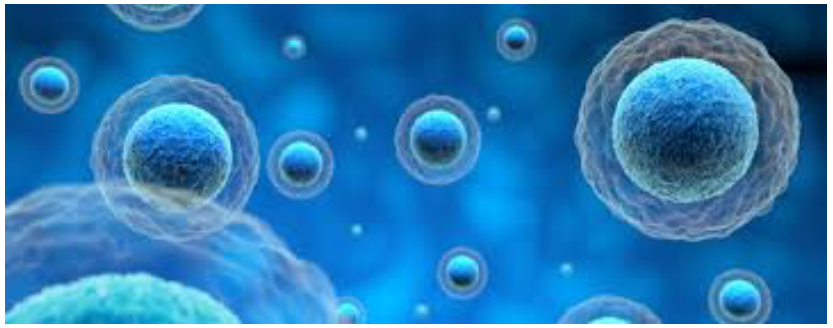
/vəˈridək(ə)l/

adjective FORMAL

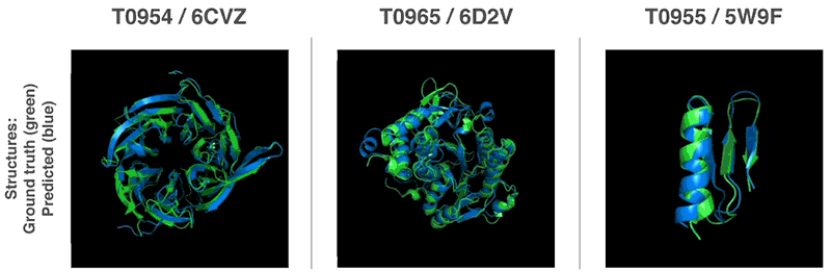
truthful.

- coinciding with reality.
  "such memories are not necessarily veridical"

Biomedical data problems are pressing



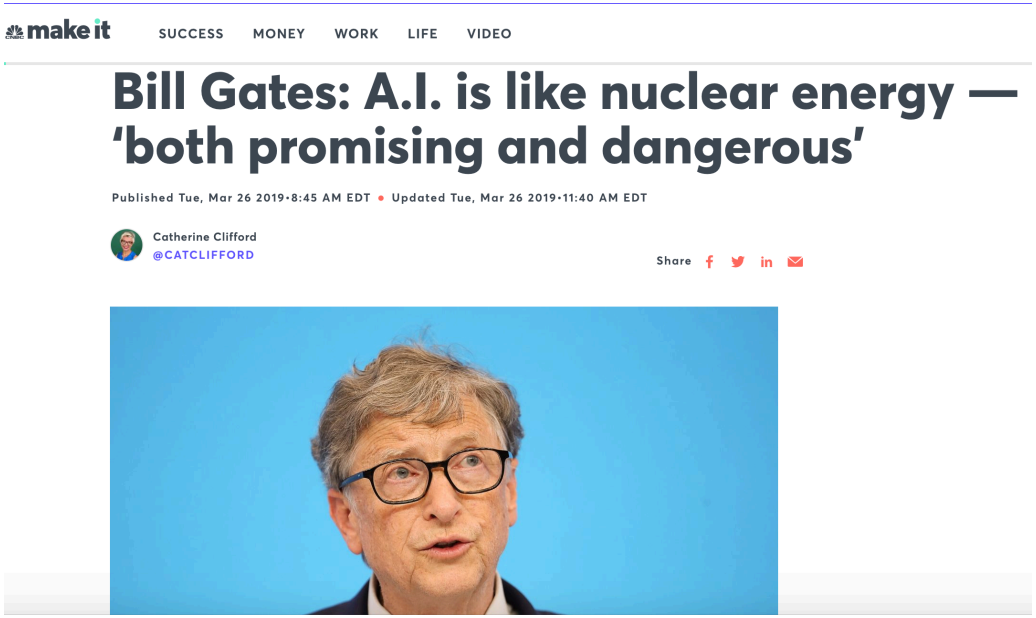
medium.com



https://deepmind.com/blog/alphafold/

website of S. Saria at JHU

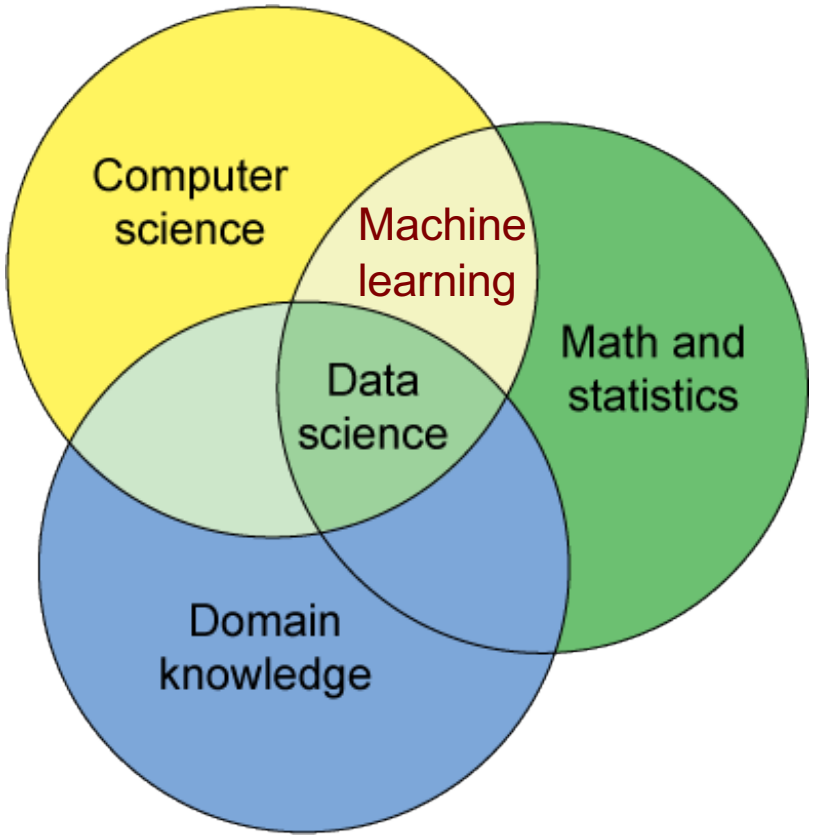
AI is part of modern life



Alexa, Siri, …
Wearable health devices
Streaming videos, on-line gaming, …
On-line news
Self-driving cars
Election campaigns
Precision medicine
Biology
Neuroscience
Cosmology
Material science
Chemistry
Law
Political science
Economics
Sociology
…

Data science is a key element of AI

Conway's Venn Diagram



Goal:

combine data with domain knowledge to make decisions and generate new knowledge

Veridical Data Science

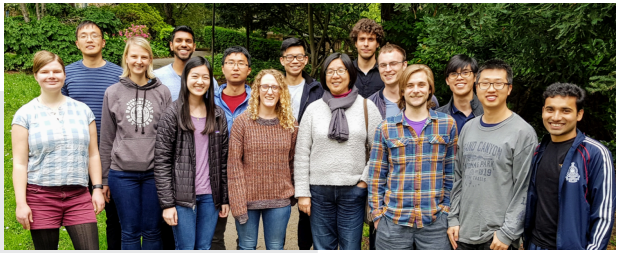
Extracts reliable and reproducible information from data,
with an enriched technical language to communicate
and evaluate empirical evidence in the context of
human decisions and domain knowledge

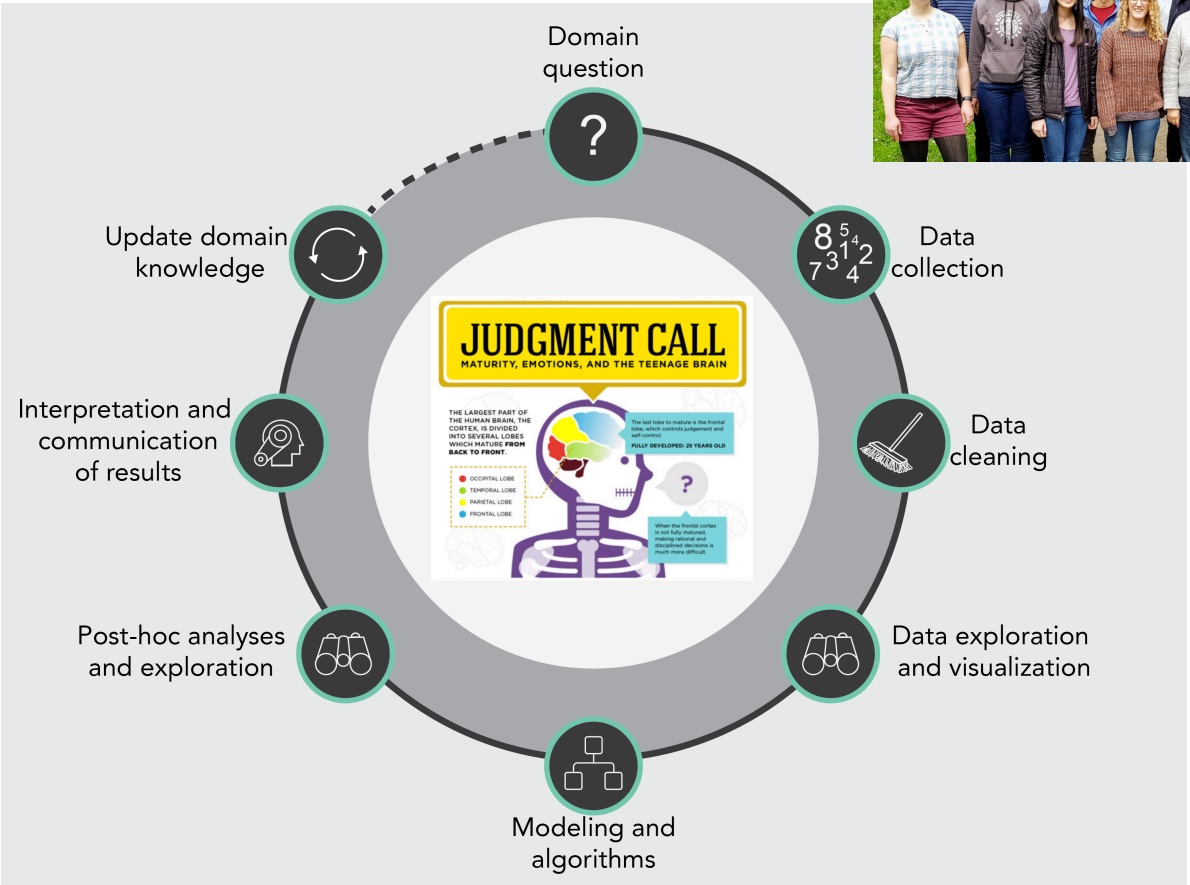
Veridical Data Science

Extracts reliable and reproducible information from data,
with an enriched technical language to communicate
and evaluate empirical evidence in the context of
human decisions and domain knowledge

It realizes promises and mitigates dangers of AI.

A holistic view of DS: a system





Missing: quality control and standardization of the process

Image credits: R. Barter and toronto4kids.com

Rest of the talk

- PCS framework for veridical data science

- PCS case study in biomedical research:
  epiTree for epistatic interactions

PCS framework for veridical data science

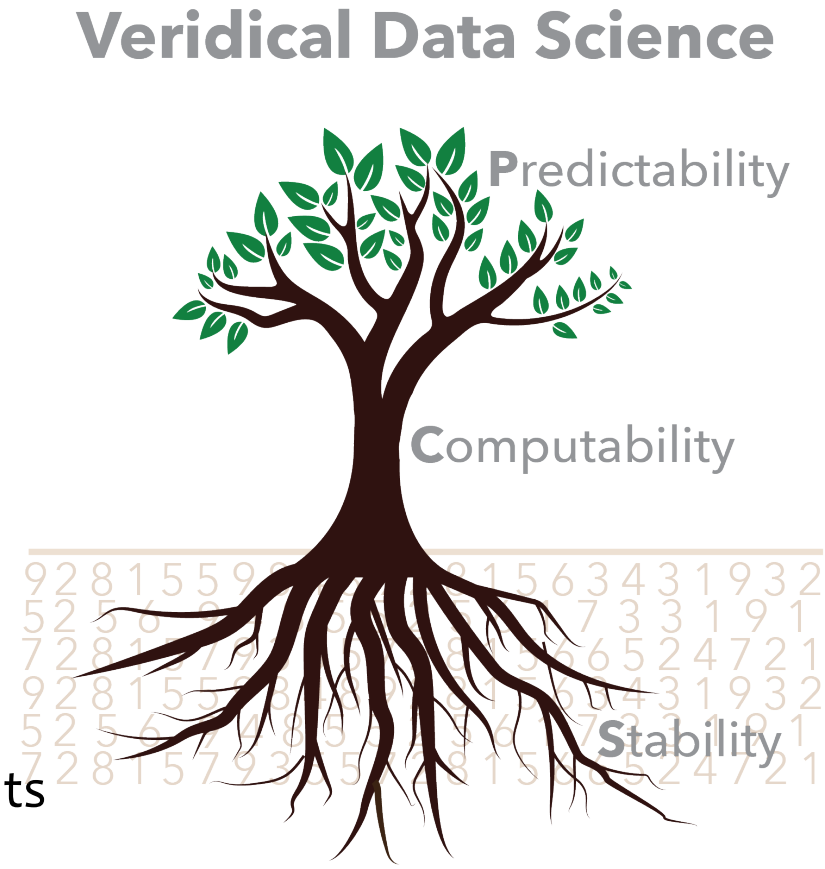
PCS framework Y. and Kumbier (PNAS, 2020)

Three principles of data science : PCS

Predictability (**P**) (ML and Stats)

Computability (**C**) (ML)

Stability (**S**) (Stats)

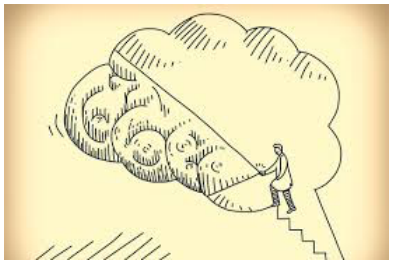


PCS bridges Breiman's two cultures.
It unifies, streamlines and expands on ideas and best practices in ML and Stats

Image credit: R. Barter

PCS connects science with engineering

- Predictability and stability embed two scientific principles: prediction and replication



- Computability is a necessity and includes data-inspired simulations

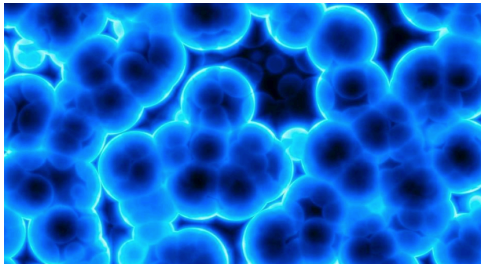


Image credits: nstat.org, hub.jhu.edu, vox.com, Andras Libal

Stability is robustness for all parts of DSLC

Bernoulli **19**(4), 2013, 1484–1500
DOI: 10.3150/13-BEJSP14

Stability

BIN YU

It unifies and extends a myriad of works on "perturbation" analysis.

It is a minimum requirement for **interpretability, reproducibility,** and **scientific hypothesis generation or intervention design**.

Image credit: designnews.com

Stability check:
The stability principle

“

Reproducibility is imperative for any scientific discovery. More often than not, modern scientific findings rely on statistical analysis of high-dimensional data. At a minimum, reproducibility manifests itself in ***stability*** *of statistical results relative to* ***reasonable perturbations*** *to data and to the model used.*

- Y. (2013) [Stability]

Predictability for reality check
Stability tests DSLC by “shaking” every part

DSLC

Shakes come from human decisions

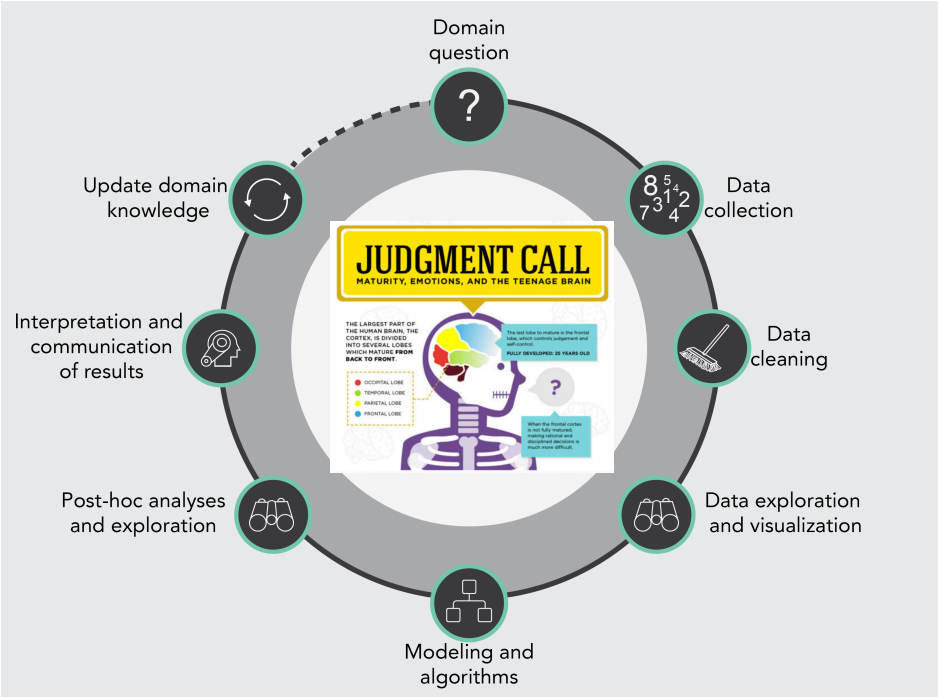
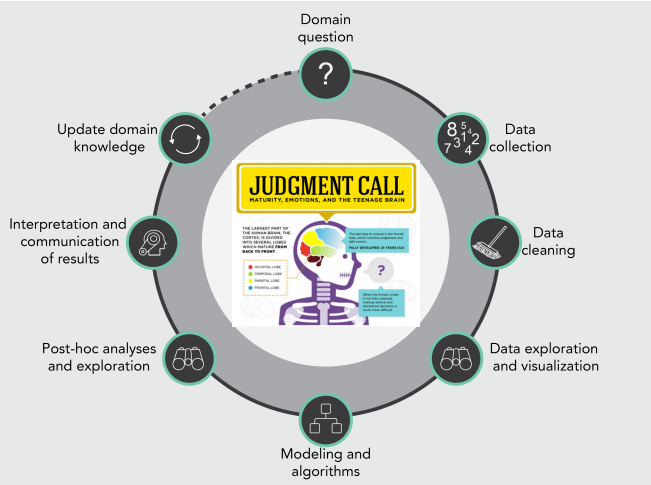


Image credits: R. Barter and toronto4kids.com

PCS workflow

- Workflow incorporates P, C, S into each step of the DSLC



- In particular, basic PCS inference applies PCS through data and model perturbations at the modeling stage (with P as a first screening step before perturbation intervals are made)

Image credits: R. Barter and toronto4kids.com

Data perturbations (existing) in the formal data analysis step

- Cross-validation
- Bootstrap
- Subsampling
- Adding small noise to data
- *Parametric bootstrapping (e.g. bootstrap in mixed effect models has several versions)
- Block-bootstrap

Data perturbations (recent)

- Data modality choices (e.g. audio vs video in PIAAC data)
- Synthetic data (mechanistic PDE models)
- Data under different environments (invariance) (e.g. different countries in PIAAC data)
- Differential Privacy (DP) (2020 US census)
- **Adversarial attacks to deep learning algorithms**

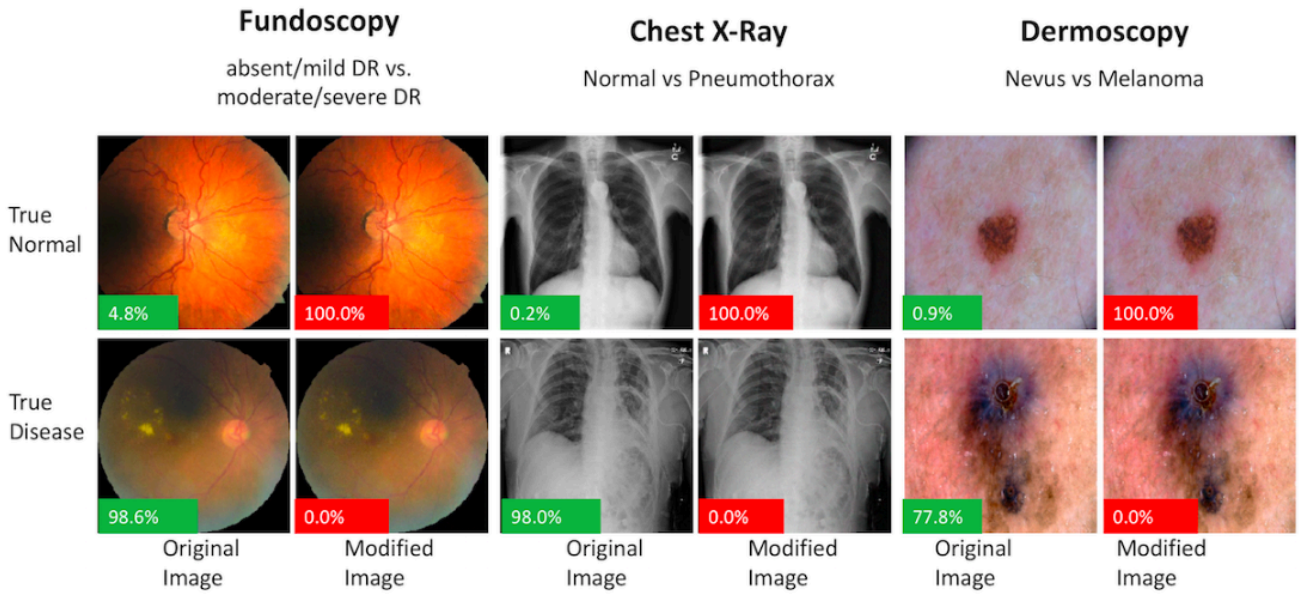


Image credits:groundai.com

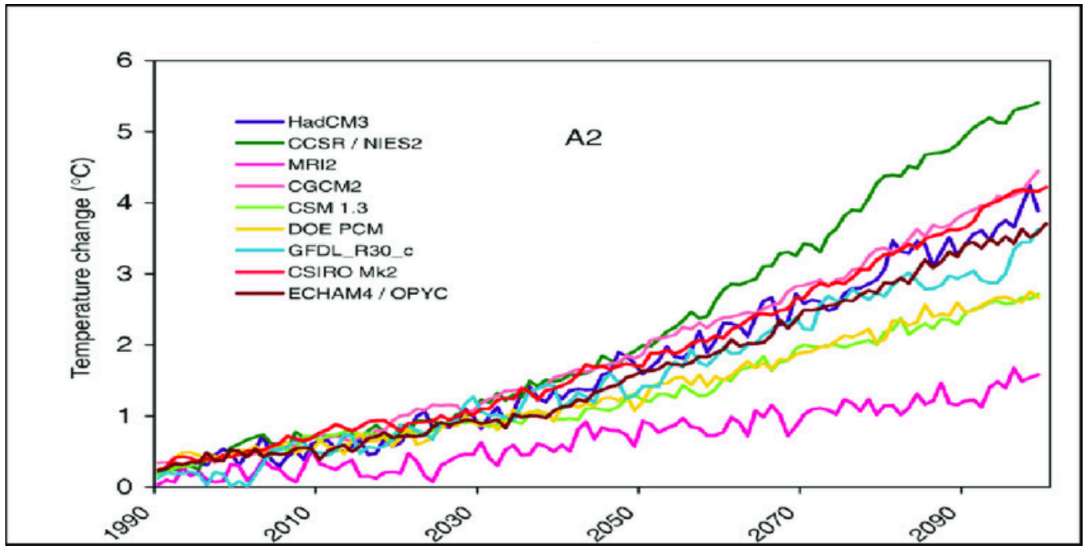
Model/algorithm perturbations (existing)

- Robust statistics
- Semi-parametric
- Lasso and Ridge
- Modes of a non-convex empirical minimization
- Kernel machines
- Sensitivity analysis in Bayesian modeling

Model/algorithm perturbations (new)

- Researcher to researcher (or team to team) perturbation

Example: 9 climate models



The change in global-mean temperature estimated by nine climate models forced by the SRES A2 emission scenario. (Source: IPCC TAR, Chapter 9)

Global mean-temp change

Human judgment calls ubiquitous in DSLC

- Which problem to work on
- Which data sets to use
- How to clean
- What plots
- What data perturbations
- What algorithm perturbations
- What post-hoc plots/results
- What interpretations
- What conclusions

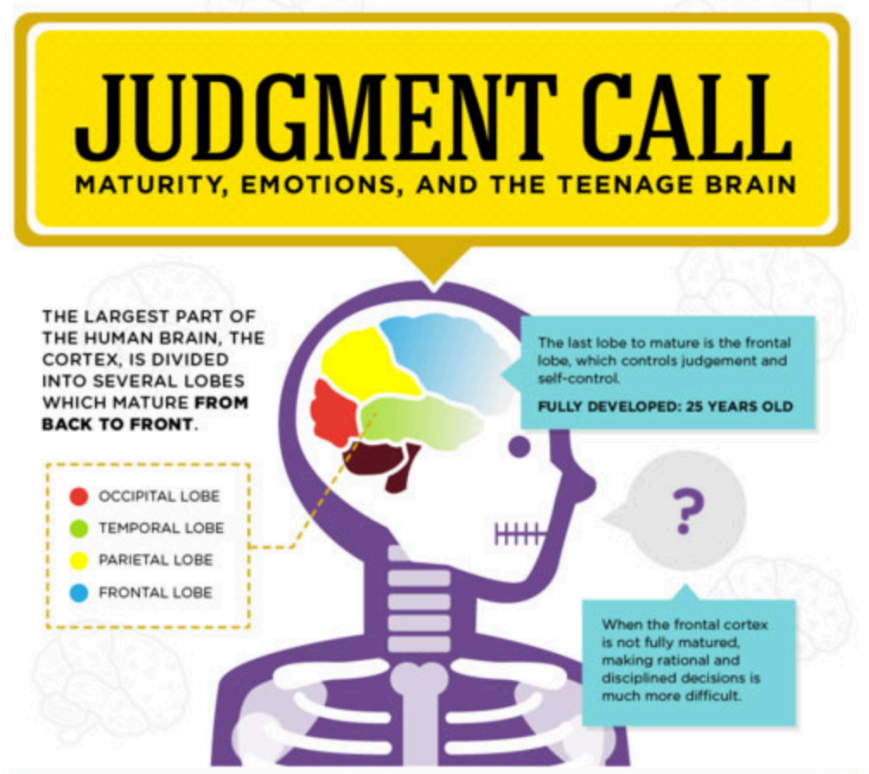
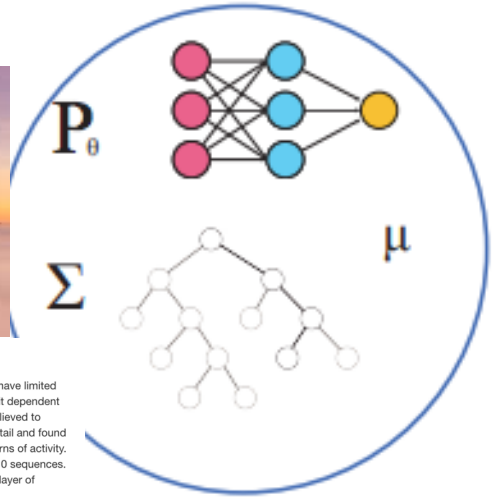


Image credits: toronto4kids.com

PCS doc. bridges reality and models on github

Reality

Models



Stability formulation

Bootstrap sampling is a widely accepted perturbation
understanding of the dependencies. However, sequen
behavior that is possible to account for. In particular, e
confer robustness to regulatory processes (Hong, Hen
that over 70% of loci they examined have anywhere fr
To account for this potential dependency along the ge
We define the stability of an interaction to be the propo
bootstrap samples using the 3 proposed perturbation

at is a useful baseline for data where we have limited
me space (i.e. nearby on the DNA) exhibit dependent
ks known as "shadow enhancers" are believed to
l. 2016) studied shadow enhancers in detail and found
et al. 2016) with highly overlapping patterns of activity.
ap perturbations using blocks of 5 and 10 sequences.
across $B = 100$ RFs trained on an outer layer of

```
# Block bootstrap for blocks of size 5 and 10
block5.tr <- makeBlocks(gene.coords, idcs=train.id, size=5)
block10.tr <- makeBlocks(gene.coords, idcs=train.id, size=10)
block5.tst <- makeBlocks(gene.coords, idcs=test.id, size=5)
block10.tst <- makeBlocks(gene.coords, idcs=test.id, size=10)
```

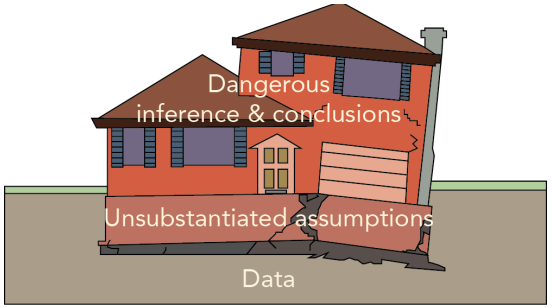


Image credit: Rebecca Barter

How to choose perturbations in PCS?

- One can never consider all possible perturbations due to computation constraint
- A pledge to the stability principle in PCS would lead to null results if too many perturbations were considered
- PCS requires documentation on the appropriateness of all the perturbations
- To avoid null results, PCS encourages careful and well-founded choices of the perturbations through PCS documentation

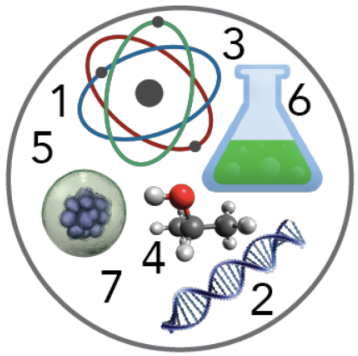
Expanding statistical inference under PCS

- Modern goal of statistical inference is to provide one source of evidence to domain experts for decision-making

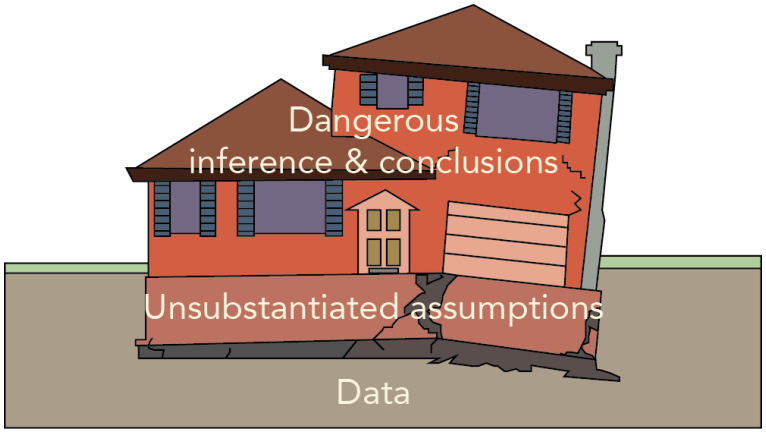
- The key is to provide data evidence in a transparent manner so that domain experts can understand as much as possible our evidence generation to evaluate the evidence strength

Traditionally, p-value has been used as evidence for decisions, but its use has been problematic that psychology journals banned it

A critical examination of probabilistic statements in statistical inference



$$? \; ? \; ? \; ? \; ? \; ? \; ? \; ? \; ? \; ? \; \mathbf{P}_\theta$$



- Viewing data as a realization of a random process is an ASSUMPTION unless randomization is explicit
- When not, using r.v. actually implicitly assumes "stability"
- If this assumption is not substantiated, all probabilistic statements are questionable
- Small p–values often measure model-bias
- The use of "true" in the "true model" is misleading – we should use other words like approximate or postulated

PCS inference (Yu and Kumbier, 2020)

P: It uses prediction error as model checking

S: It relies on data and model perturbations (with data perturbation broadly interpreted)

C: Both P and S require it

It does not rely on a probabilistic generative model assumption.

PCS inference

1. **Problem formulation:** Translate the domain question to be answered by a model/algorithm (or multiple of them and seek stability). Specify a target of interest.

   Split data into: training and test

PCS inference

1. **Problem formulation:** Translate the domain question to be answered by a model/algorithm (or multiple of them and seek stability). Specify a target of interest.

   Split data into: training and test

2. **Prediction screening for reality check or model checking:**
   Filter models/algorithms based on prediction accuracy on training set (CV)

PCS inference

1. **Problem formulation:** Translate the domain question to be answered by a model/algorithm (or multiple of them and seek stability). Specify a target of interest.

   Split data into: training and test

2. **Prediction screening for reality check or model checking:**
   Filter models/algorithms based on prediction accuracy on training set (CV)

3. **Target value perturbation distribution:** Evaluate the target of interest across "appropriate" data and model perturbations on test set

PCS inference

1. **Problem formulation:** Translate the domain question to be answered by a model/algorithm (or multiple of them and seek stability). Specify a target of interest.

   Split data into: training and test

2. **Prediction screening for reality check or model checking:**
   Filter models/algorithms based on prediction accuracy on training set (CV)

3. **Target value perturbation distribution:** Evaluate the target of interest across "appropriate" data and model perturbations on test set

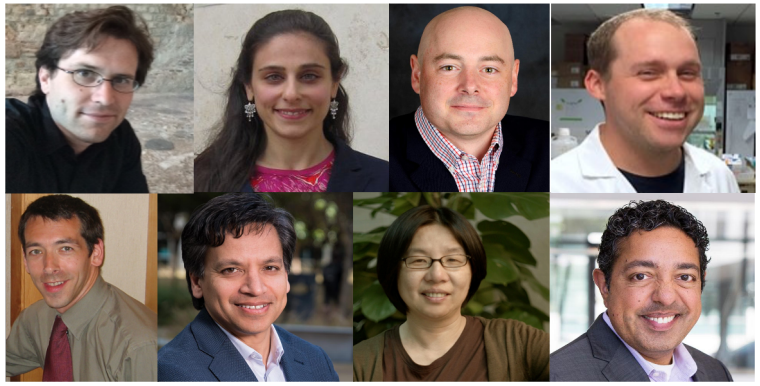
4. **Perturbation region reporting:** summarize target values over the perturbations in a distribution

PCS case study: epiTree for epistatic interactions

Multi-scale deep learning and single- cell models of cardiovascular health

PIs: Euan Ashley*, Rima Arnaout*, Ben Brown, Atul Butte, James Priest*, Bin Yu*

Collaborators: Chris Re, Deepak Srivastava



Postdocs/students:

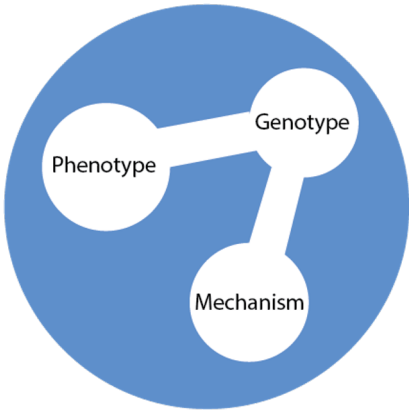
M. Behr

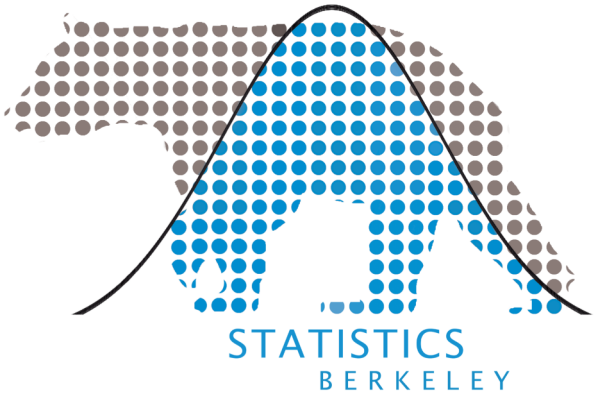
K. Kumbier

M. Aguirre

A. Cordova-Palomera

Q. Wang





Detecting epistatic interactions via epiTree

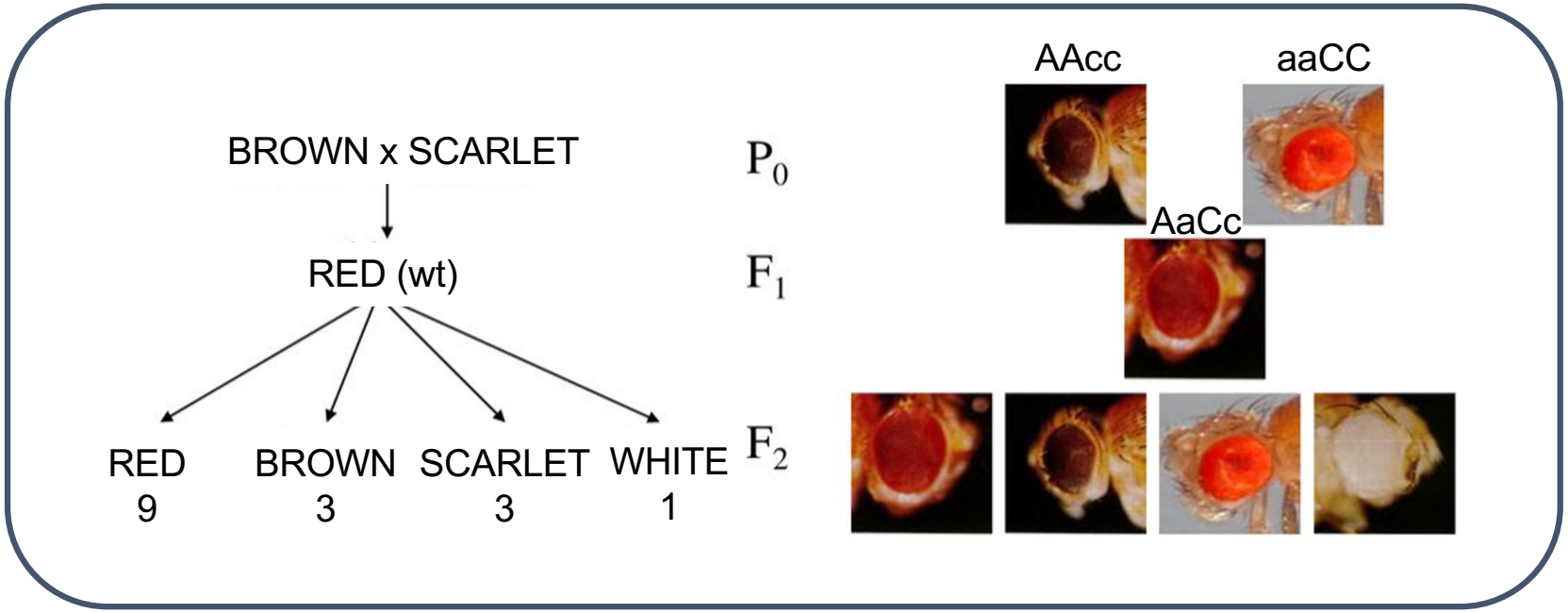
M. Behr, K. Kumbier, M. Aguirre, R. Arnaout,
E. Ashley, A. Butte, R. Arnout, B. Brown**, J. Priest, B. Yu**

https://www.biorxiv.org/content/10.1101/2020.11.24.396846v1

"Learning epistatic polygenic phenotypes with Boolean interactions"

Epistasis: a brief introduction to non-linear interactions

- A non-linear relationship between two (or more) genetic variants and a specific phenotype



Epistasis: a brief introduction to non-linear interactions

- A non-linear relationship between two (or more) genetic variants and a specific phenotype
- Traditionally formulated as a multiplicative interaction term (Fisher, 1919)

$$\text{logit}(p) = \beta_0 + \beta_1 A + \beta_2 B + \mathbf{\beta_3 A*B}$$

Major Problems

- **Definition of "non-linear" dependent on scaling of outcome**
- **Computationally intractable to test for interactions greater than order 2**

New method EpiTree for epistasis discovery

- Flexible and non-linear mathematical form
- Agnostic to scaling of outcome variable
- Suited to detect interactions greater than order 2

Positive control phenotype: red-hair

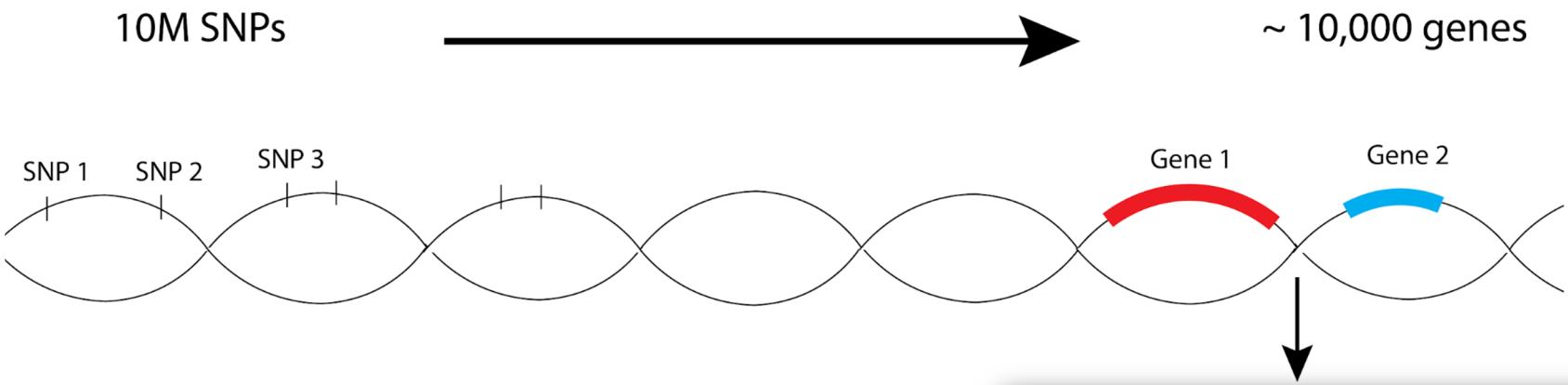
- Entirely genetic
- Governed by epistasis
- Common trait

UK Biobank

- 500,000 individuals
- Self-reported hair color
- 10,000,0000 variants from array genotype data

Biologically inspired dimension reduction:

imputed gene expression from PrediXcan (skin tissue) (Gamazon et al, '15)



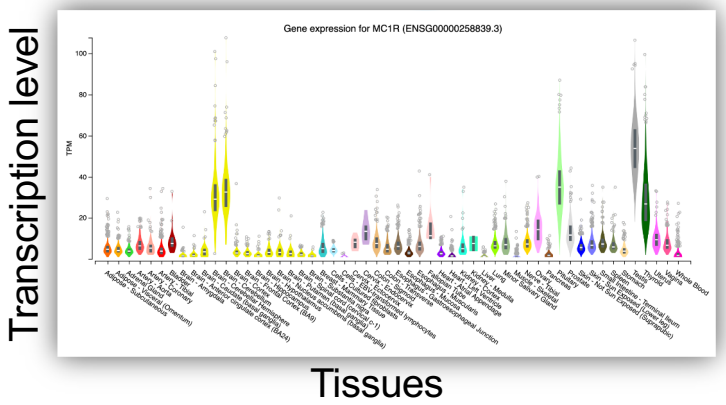
Published: 10 August 2015

A gene-based association method for mapping traits using reference transcriptome data

Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, GTEx Consortium, Dan L Nicolae, Nancy J Cox & Hae Kyung Im

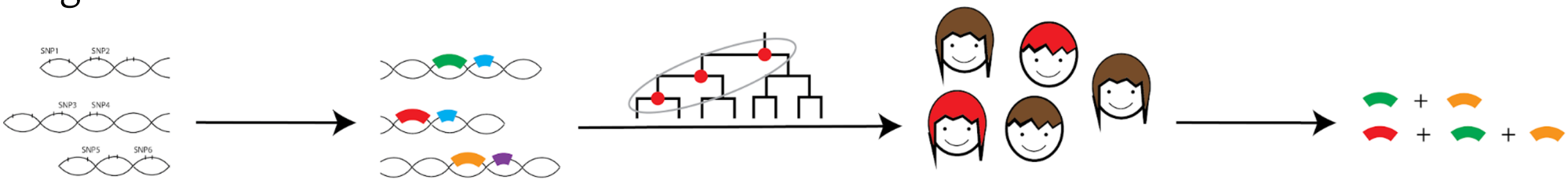
Nature Genetics **47**, 1091–1098(2015) | Cite this article

11k Accesses | **420** Citations | **79** Altmetric | Metrics



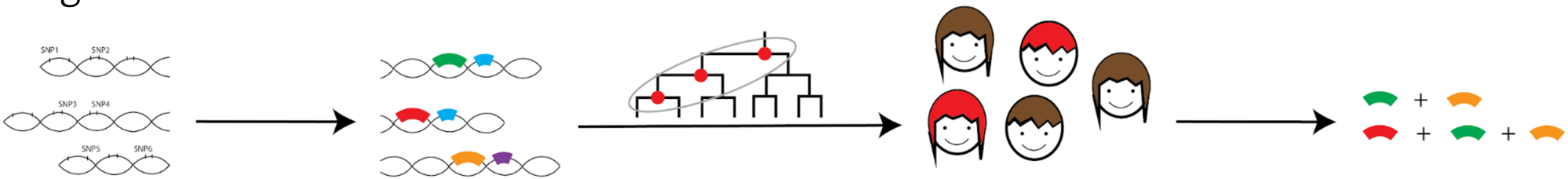
Two-step procedure:

Step 1: impute gene expression (dimension reduction) and search for interactions on gene level

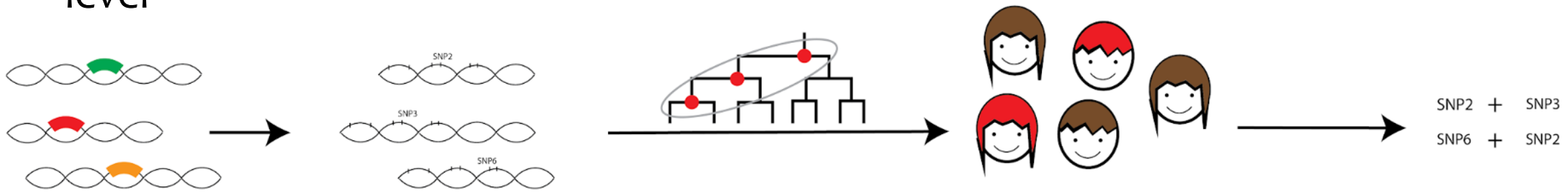


Two-step procedure:

Step 1: impute gene expression (dimension reduction) and search for interactions on gene level

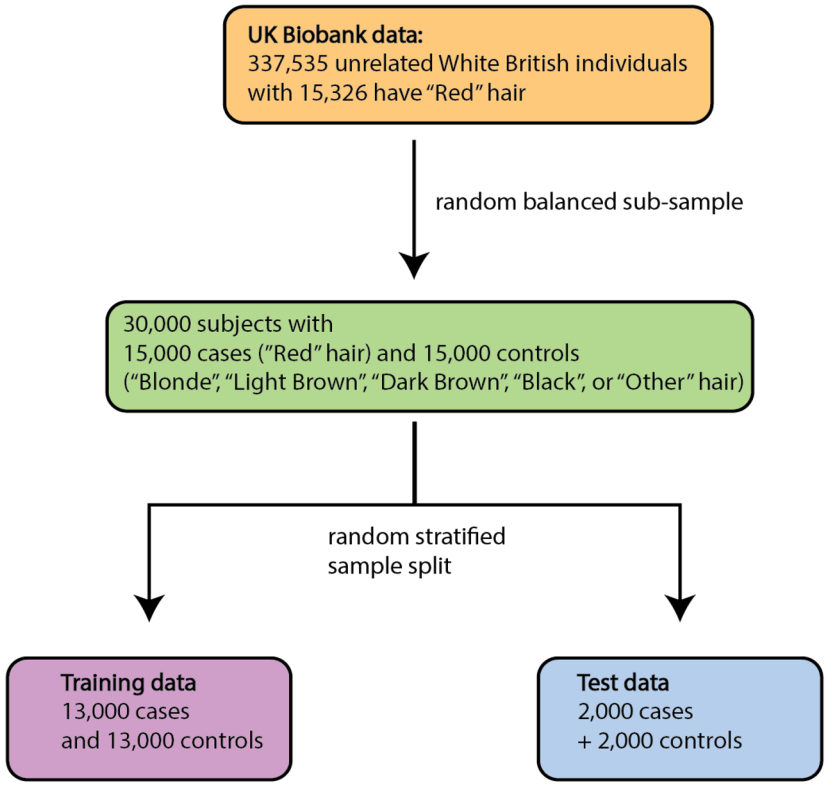


Step 2: for important genes from Step 1 extract SNPs and search for interactions on SNP level



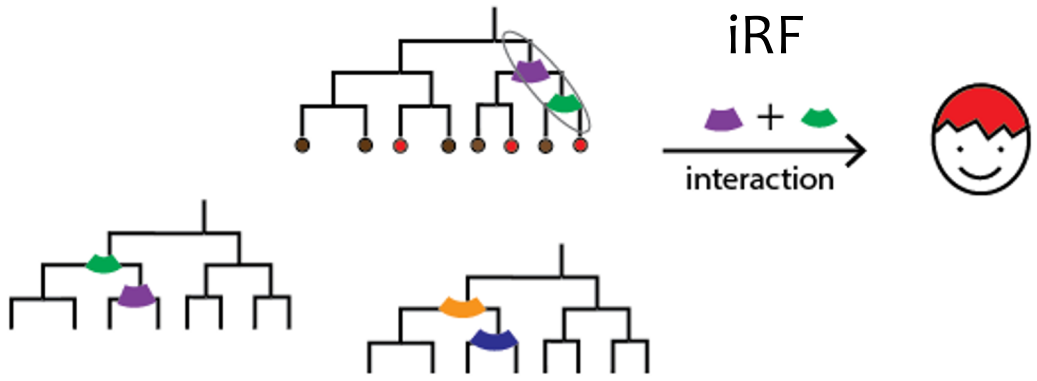
Positive control: red-hair phenotype from UKBB

- well-studied and genetically determined phenotype (Morgan et al '18)



On training data, Iterative random forests **(iRF)** selects 18 order-2 or **higher-order** interaction candidates

- Stability score in iRF is evaluated over bootstrap samples of training data



iterative Random Forests (iRFs)

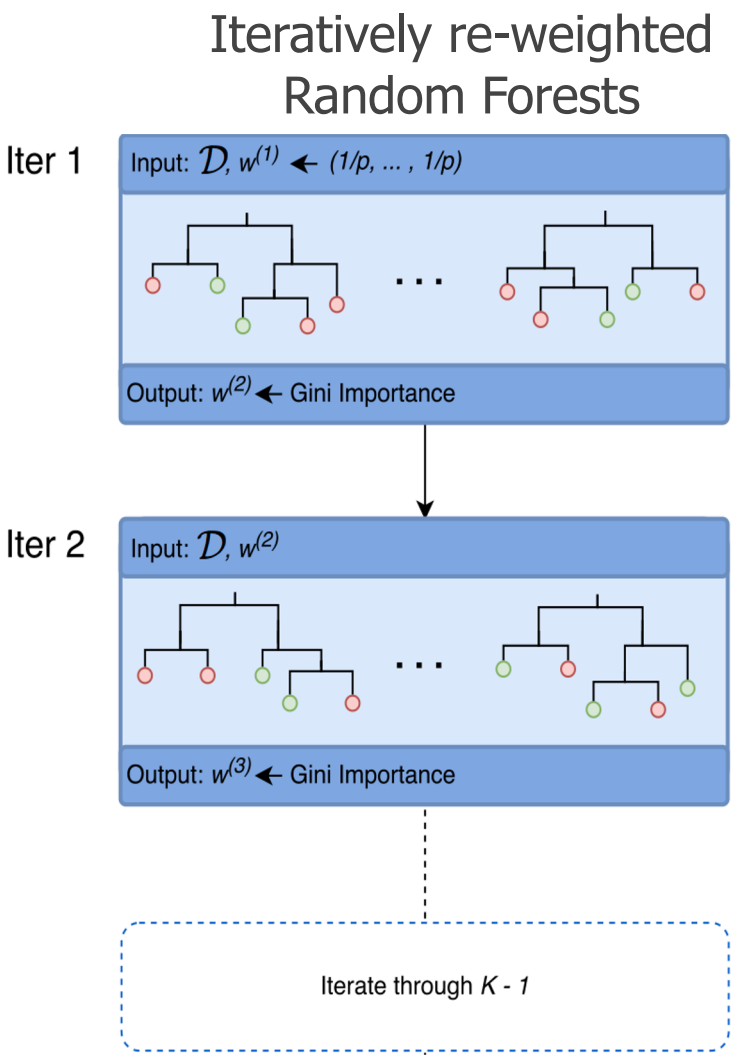
Basu, Kumbier, Brown and Yu (2018)

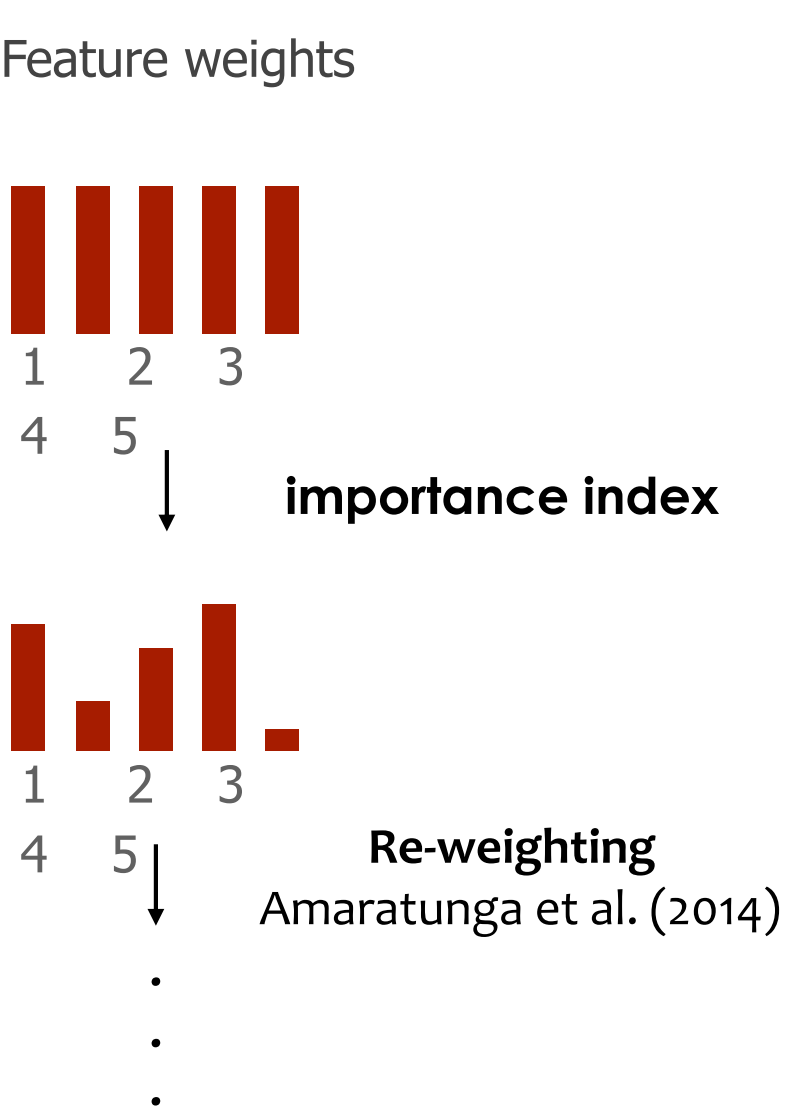
Core ideas

1. Soft dim reduction using importance index
2. Random interaction trees to find intersections of paths
3. Outer-loop bagging assesses stability

Similar computational and memory costs as RF

Iteratively re-weighted RF stabilize decision paths





Generalized RIT for Decision Trees

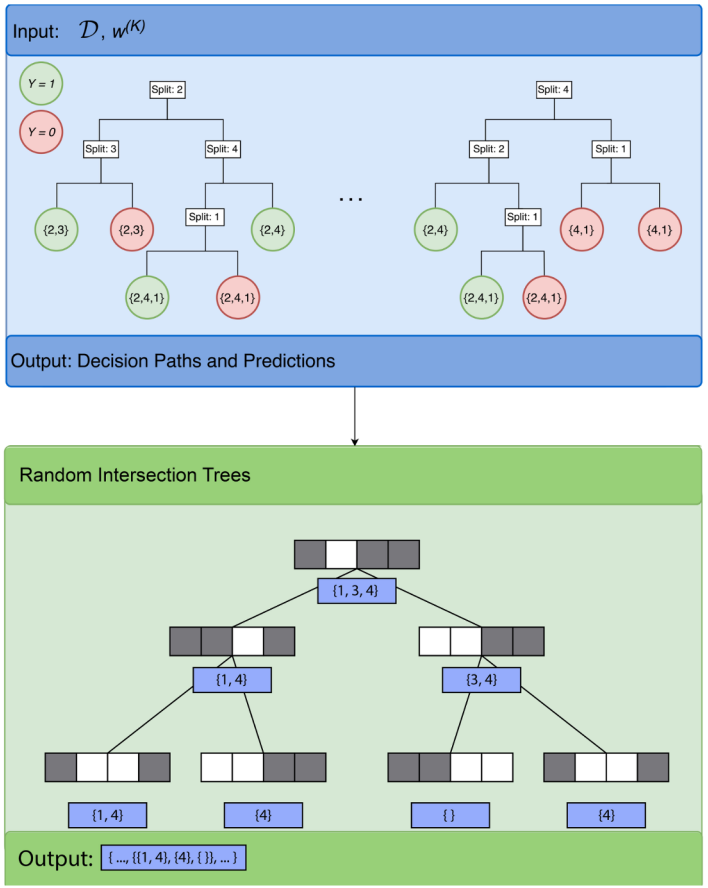
fast computation uses sparsity

(Random Intersection Trees (RIT), Shah and Meinshausen, 2014)

$\mathcal{I}_{i_t} \subseteq \{1, \ldots, p\}$ *Feature-index set* for leaf node containing observation $i = 1, \ldots, n$ in tree $t = 1, \ldots, T$

$Z_{i_t} \in \{0, 1\}$ *Prediction* for the leaf node containing observation $i = 1, \ldots, n$ in tree $t = 1, \ldots, T$

$$\mathcal{S} \leftarrow \mathrm{RIT}(\{\mathcal{I}_{i_t}, Z_{i_t}\}, C)$$



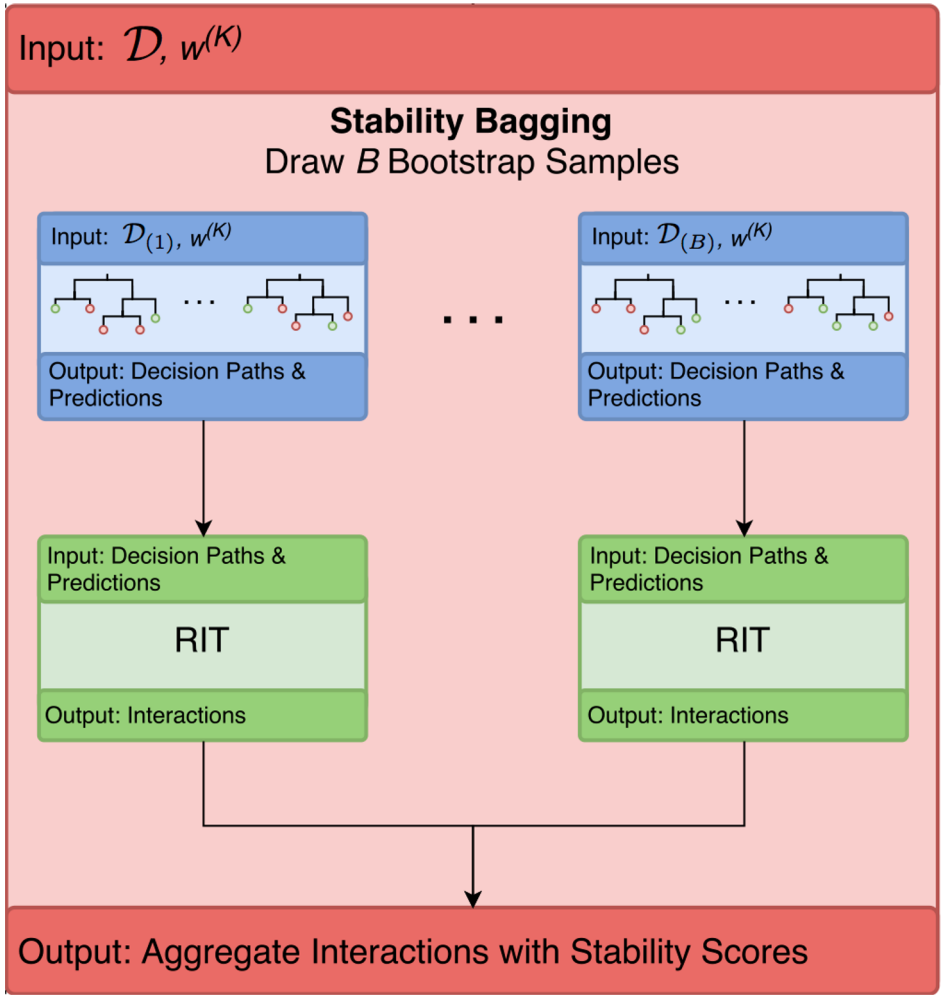
Stability bagging

Output feature interaction sets with stability scores:

$$\{S, sta(S)\}$$

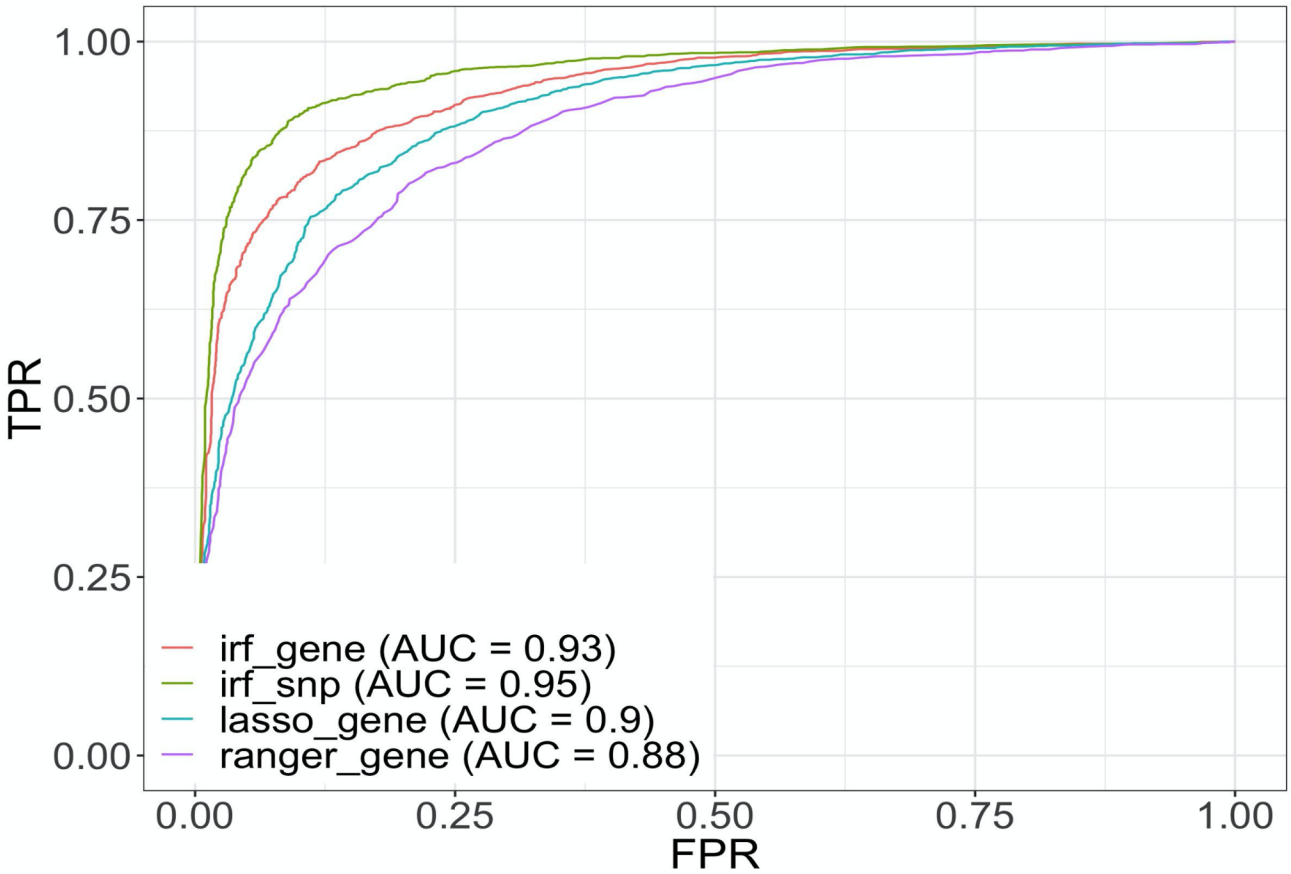
$$S \subseteq \{1, \ldots, p\}$$

$$sta(S) = \frac{1}{B} \cdot \sum_{b=1}^{B} 1(S \in \mathcal{S}_b)$$



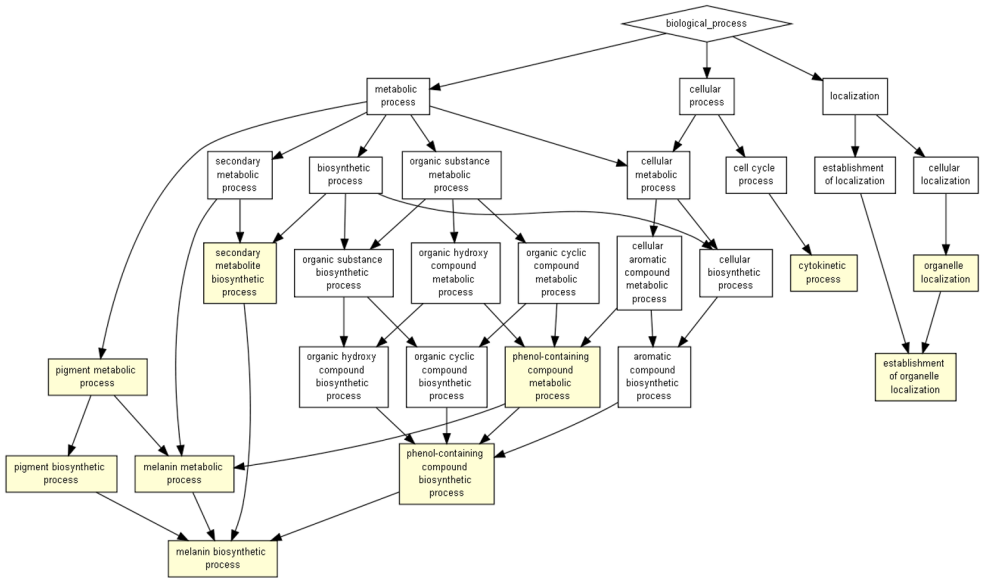
iRF uses PCS = RF (P) + RIT (C) + Stability (S)

Very high prediction accuracy on hold-out data



iRF outperforms RF (and lasso model)!

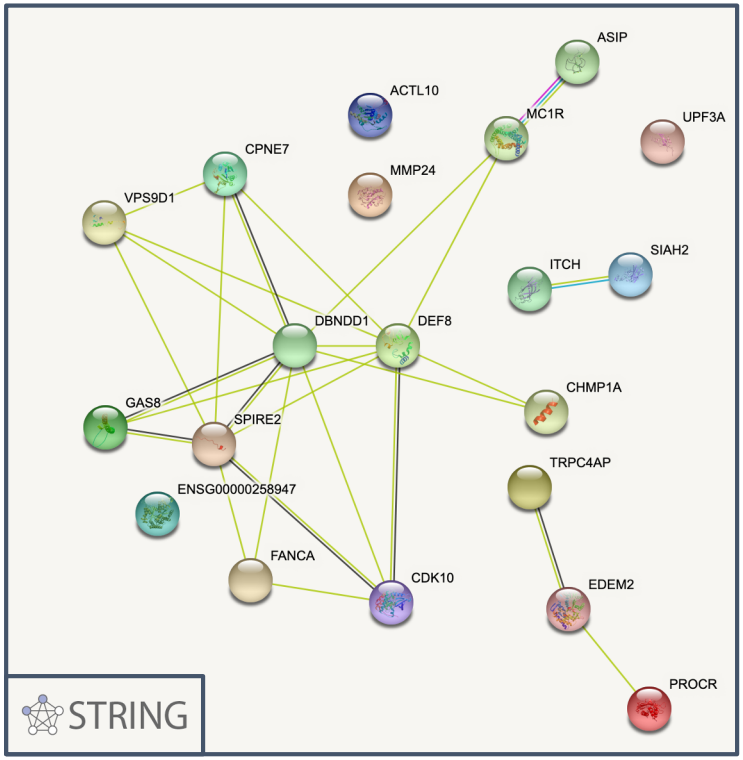
iRF recovers known genetic determinants of hair color & pigmentation:



GO Term	Description
GO:0006582	melanin metabolic process
GO:0042438	melanin biosynthetic process
GO:0044550	secondary metabolite biosynthetic process
GO:0046189	phenol-containing compound biosynthetic process
GO:0051640	organelle localization
GO:0032506	cytokinetic process
GO:0046148	pigment biosynthetic process
GO:0042440	pigment metabolic process
GO:0051656	establishment of organelle localization
GO:0018958	phenol-containing compound metabolic process

Eden et. al. BMC Bioinformatics 2009, 10:48.

iRF recovers proteins which are known to interact with each other:



Protein-protein interaction enrichment

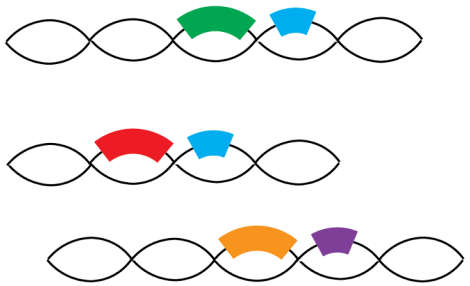
Szklarczyk et. al. Nucleic Acids Res. 2019 Jan; 47:D607-613.

Epistasis beyond multiplicative interaction

Fisher (1919): logit $(P(Y=1|A,B)) = \text{Gene A} + \text{Gene B} + \text{Int}(A, B)$

Gene level (continuous features):

1. possible forms of interaction:



- Multiplicative Int(A,B) = A*B
- Decision tree Int(A,B) = CART(A,B)
- intPredict (iRF) Int(A,B) = iRF(A,B, -)
- ...

2. possible forms of scale: logit, penetration, …

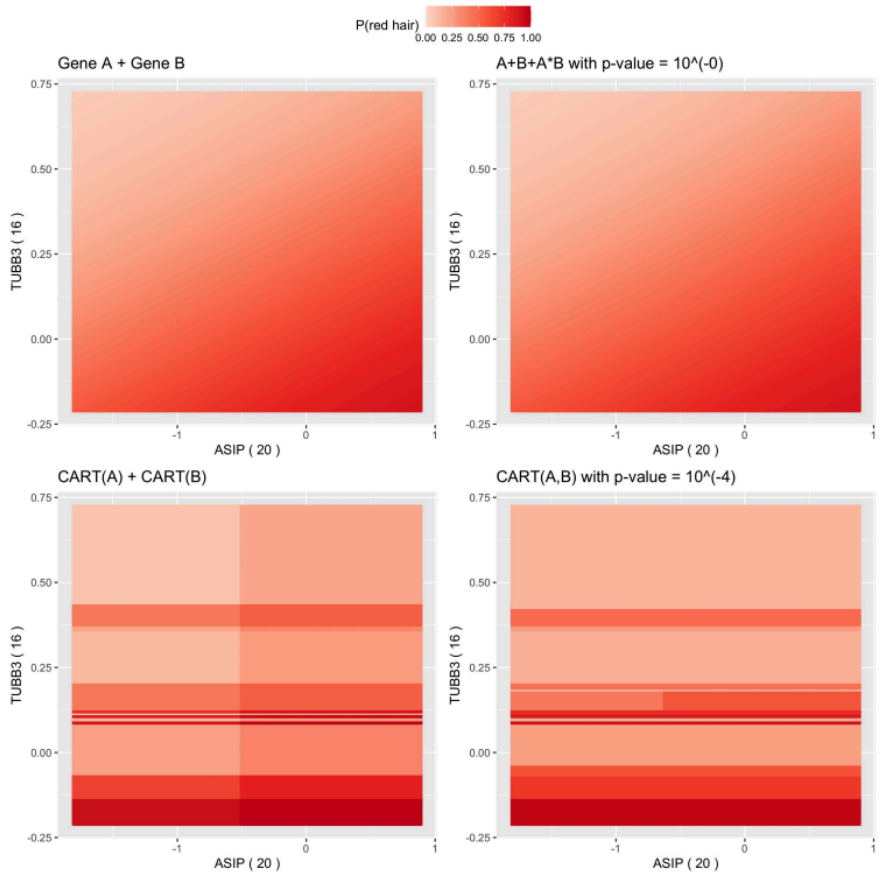
We use penetration $P(y = 1 \mid A, B)$ **for scale and biologically meaningful CART interaction**

CART models fit better than linear

Response surface for ASIP - TUBB3

Linear

CART



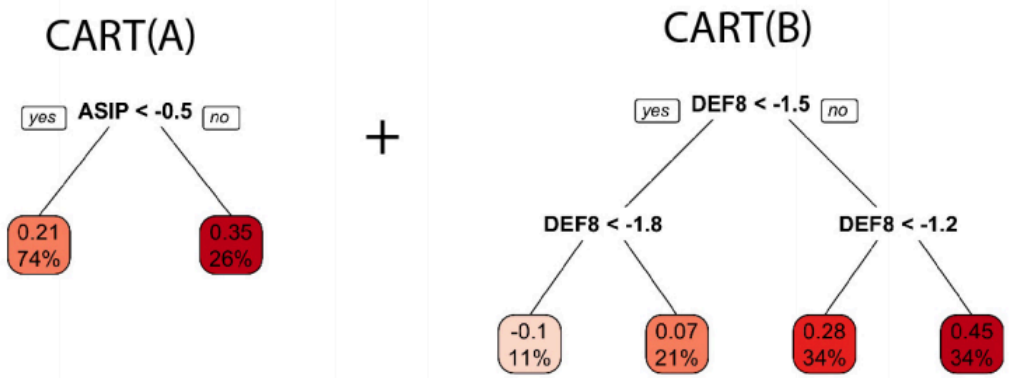
No-epistasis

Epistasis

CART fitted models for an interaction by iRF

A = ASIP, B = DEF8

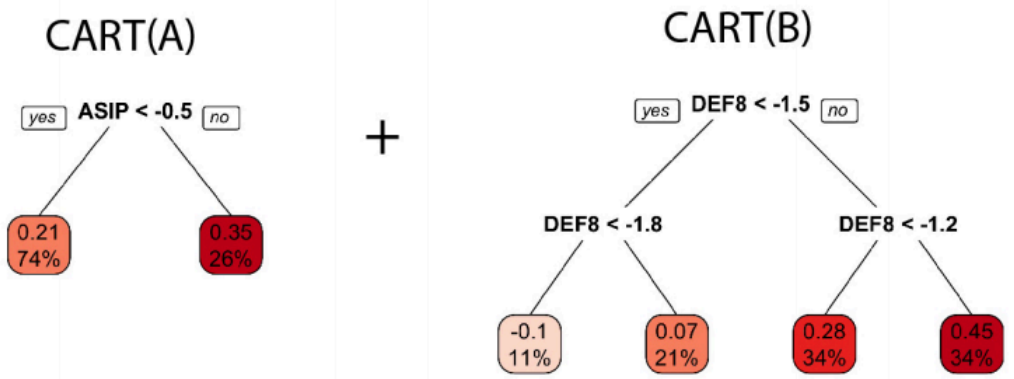
No-epistasis model (Null):



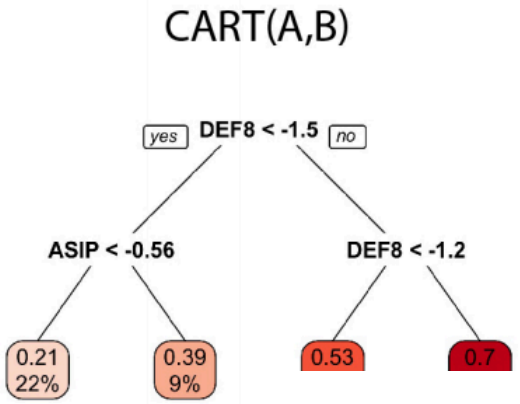
CART fitted models for an interaction by iRF

A = ASIP, B = DEF8

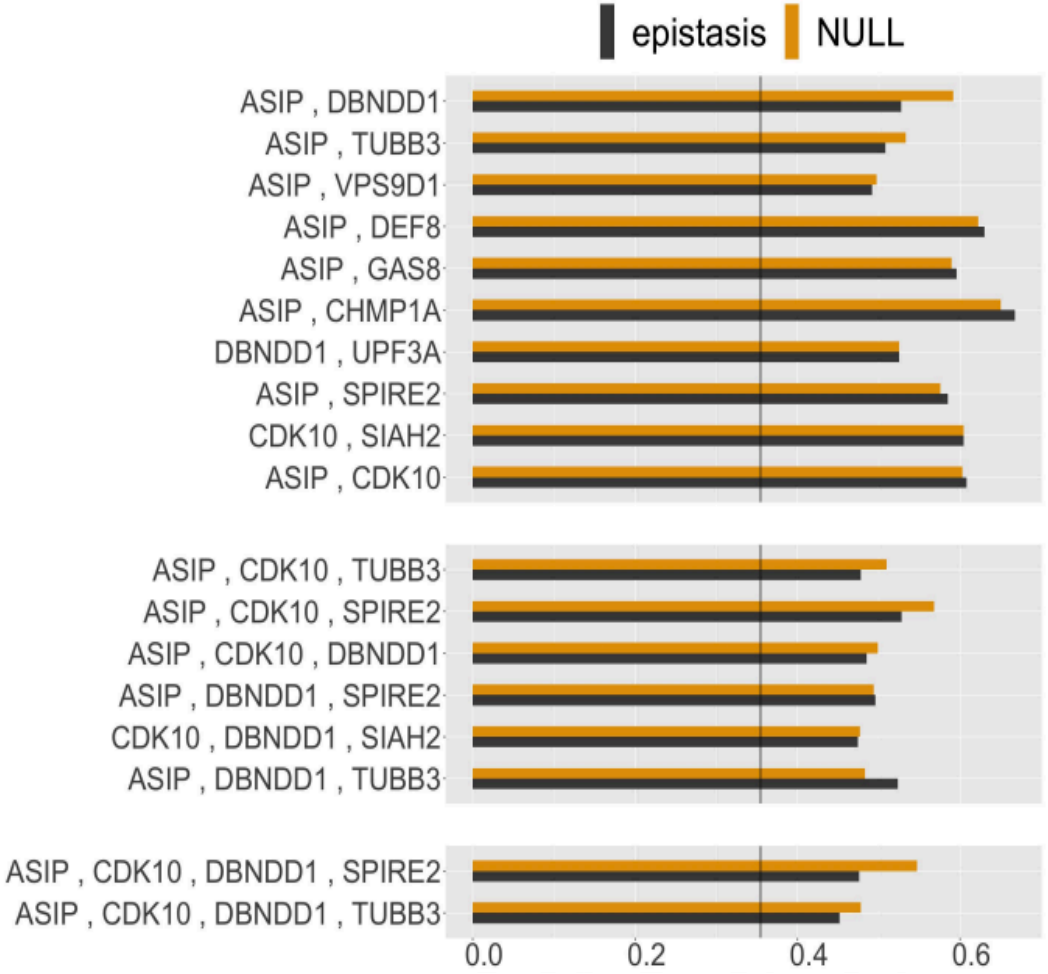
No-epistasis model (Null):



Epistasis model (Alternative):

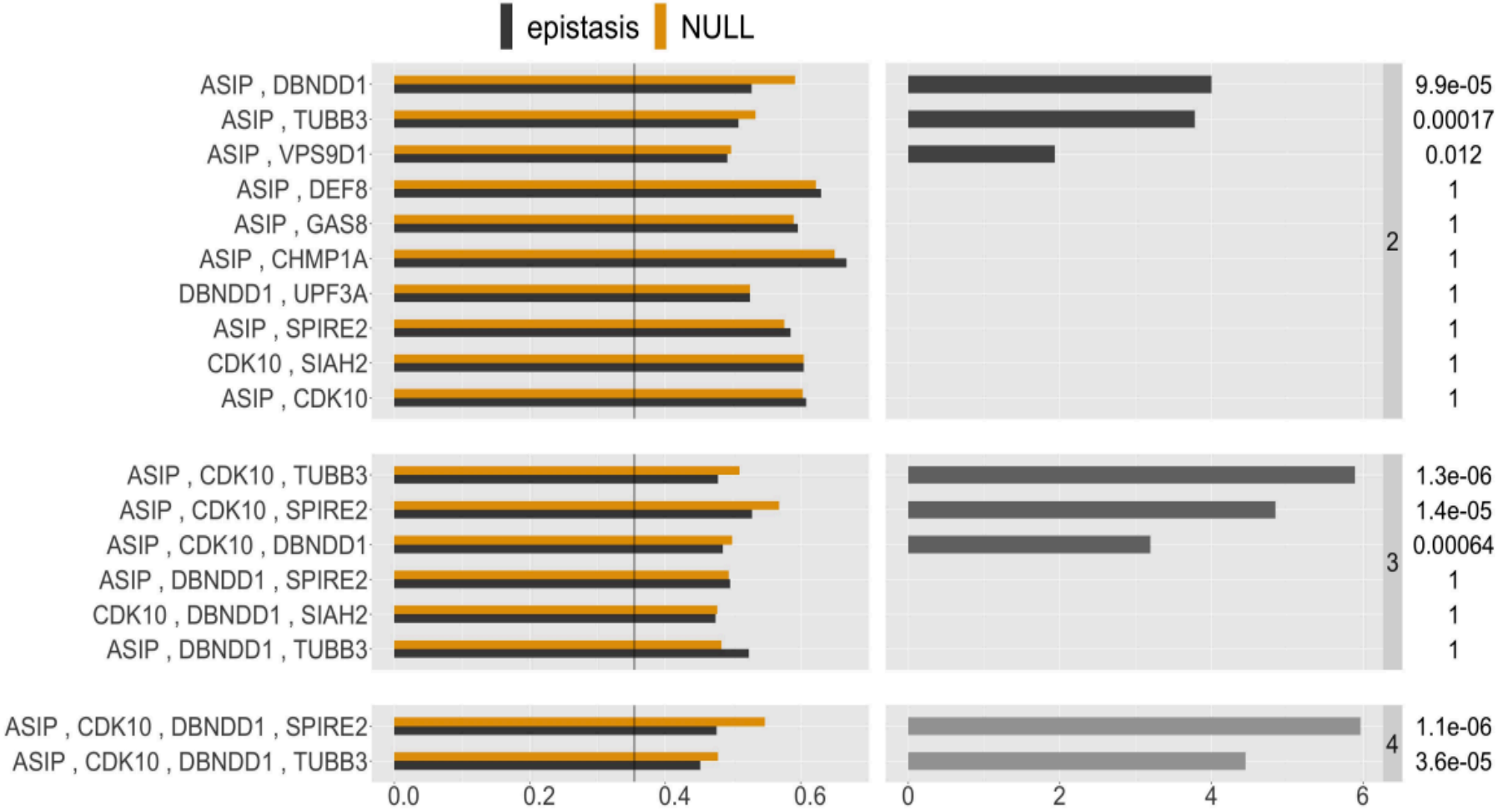


PCS inference on interactions found by iRF (stability score > 0.5)



Prediction Error

PCS inference on interactions found by iRF (stability score > 0.5)



Prediction Error

-log10 (PCS p-value)

epiTest PCS p-value calculation on test data

- If epistasis model gives worse prediction, set p-value=1
- Otherwise, p-value is calculated based on a refined comparison of the two models while taking into account *test data variability

  As a result, the PCS p-values are realistic, e.g. 10^-5 or 10^-2

  unlike 10^-11 etc from logistic Chi-sq tests (using CART terms)

  epiTree test works for more than **higher-order interactions**

PCS p-value calculation details: genes A and B

After passing the prediction screening, we bootstrap over the test data to evaluate PCS p-value.

- For each bootstrap sample m (out of M), we obtain $p_{0|m}$, an n- vector of estimated penetrance using gene expression levels under the null. we simulate null responses $Y0|m \sim$ Bernoulli($p_{0|m}$).
- For each m, we obtain $Y|m \in \{0, 1\}^n$, an n-vector consisting of the observed responses

$$\text{PCS p-value} = \text{average over m of } I_{\{T(Y|m) > T(Y0|m)\}}$$

where T(Y) is the Bernoulli log-likelihood ratio of null over alternative

PCS p-value is conservative

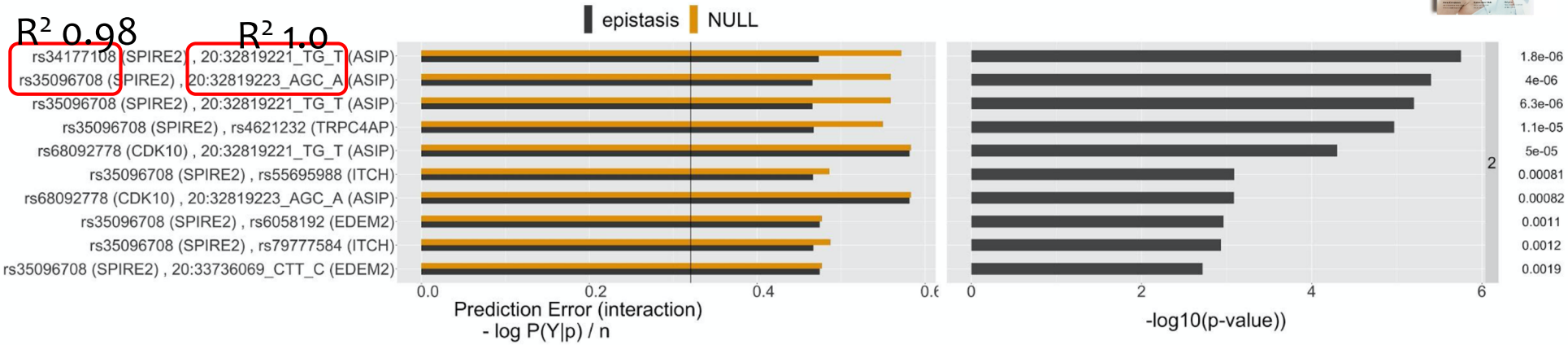
- PCS p-value =1 if null model does better prediction than alternative on test data

- On the test set, PCS null perturbation seems to correspond to a ”fattened” version of the sharp null distribution

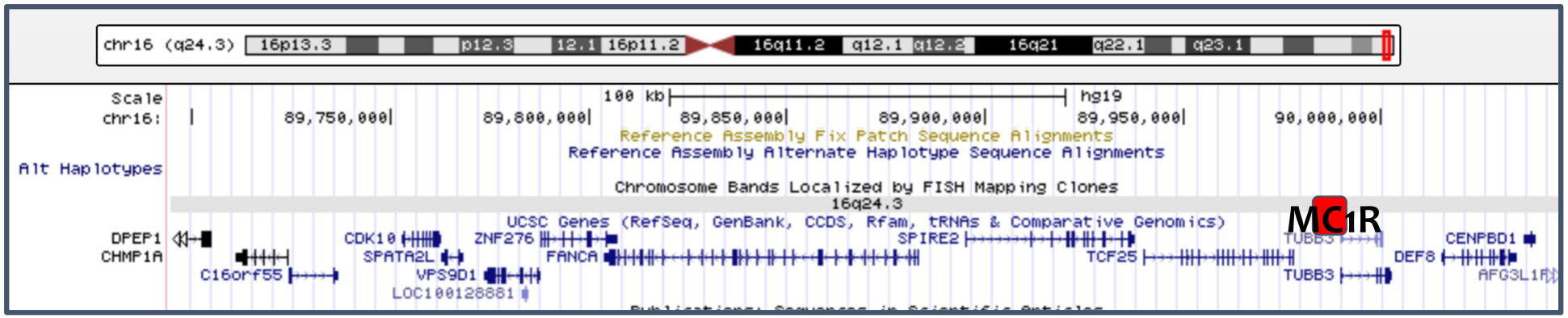
- PCS p-value seems smaller than the corresponding traditional p-value

We can make things precise in a simple linear regression model and more work is on-going.

SNP interactions recovered by epiTree

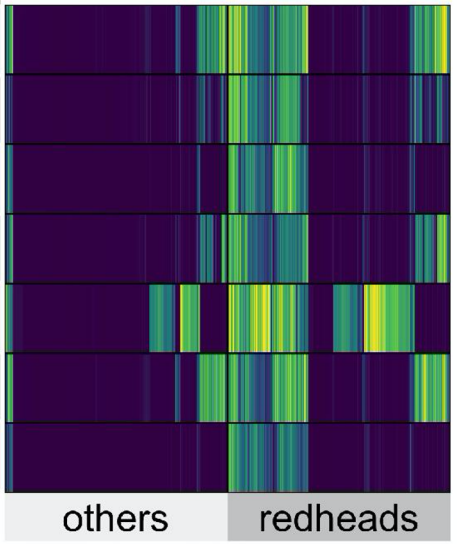


epiTree recovers ***MC1R - ASIP*** interaction also on the SNP level!



Interactions are “active” for subset of subjects

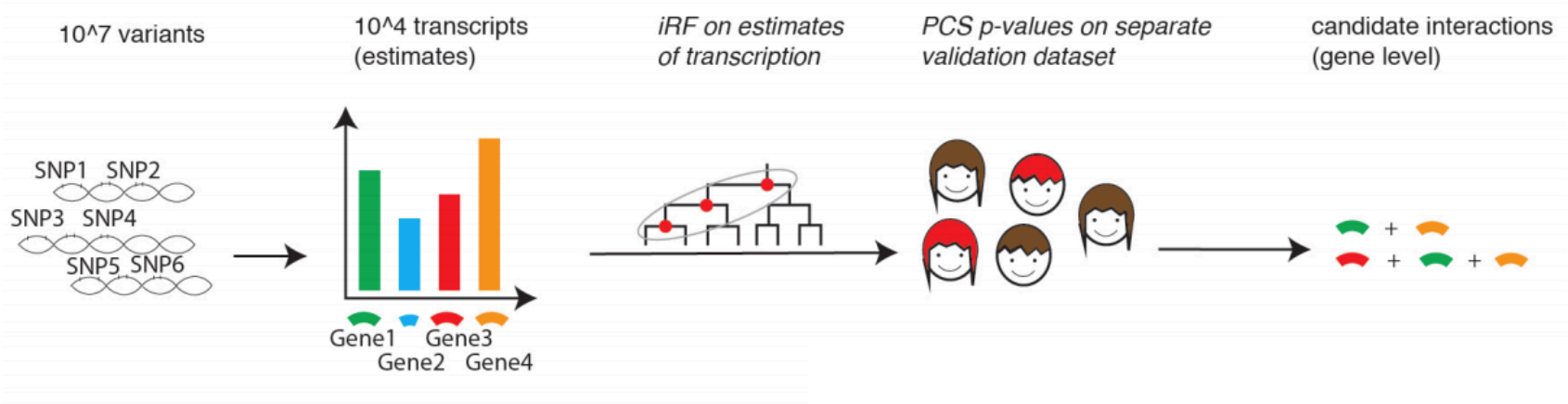
genes	chr
ASIP , CDK10	20 , 16
ASIP , CDK10 , TUBB3	20 , 16 , 16
DBNDD1 , SIAH2 , CDK10	16 , 3 , 16
ASIP , CDK10 , SPIRE2	20 , 16 , 16
DBNDD1 , UPF3A	16 , 13
SIAH2 , CDK10	3 , 16
DBNDD1 , ASIP , CDK10 , SPIRE2	16 , 20 , 16 , 16



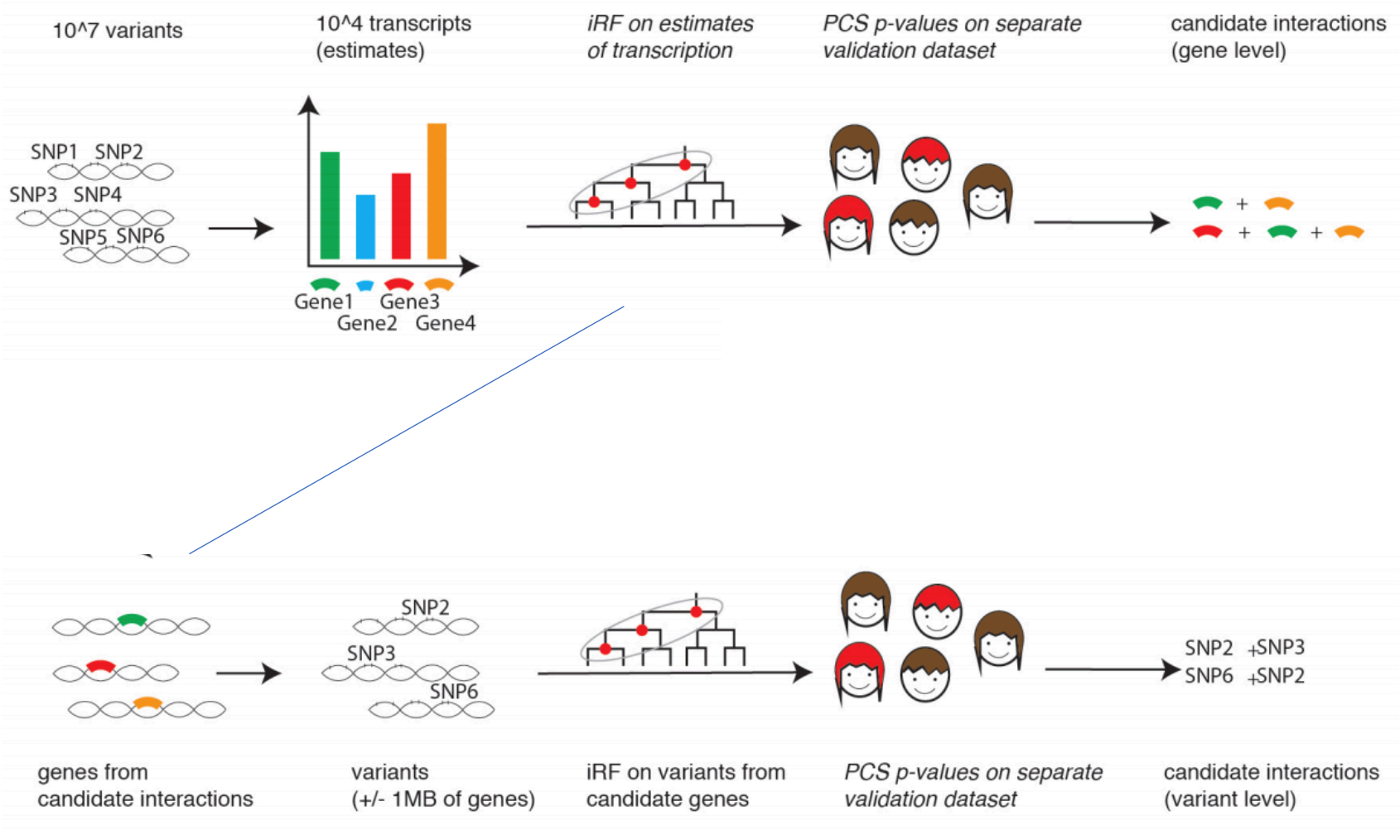
Using a **superheat** plot, Barter and Y. (2018) and **R package**

Novel higher order interactions: ASIP, CDK10, TUBB3
Possible new red-hair genes: UPF3A and SIAH2 (with MC1R related genes)

epiTRee pipeline in one figure



epiTRee pipeline in one figure



Summary

Veridical data science (trustworthy AI) through PCS

- PCS framework (workflow and documentation on github)
- Eight PCS case studies: **iRF, epiTRee**, ESCV, DeepTune, statNMF, staDISC, staDRIP
- PCS is useful for evaluation as well: PCS can stress-test clinical decision rules
- Domain knowledge is important and **PCS** generates testable results for external validation (experiments or other studies)

On-going:
Biohub project for cardiovascular health

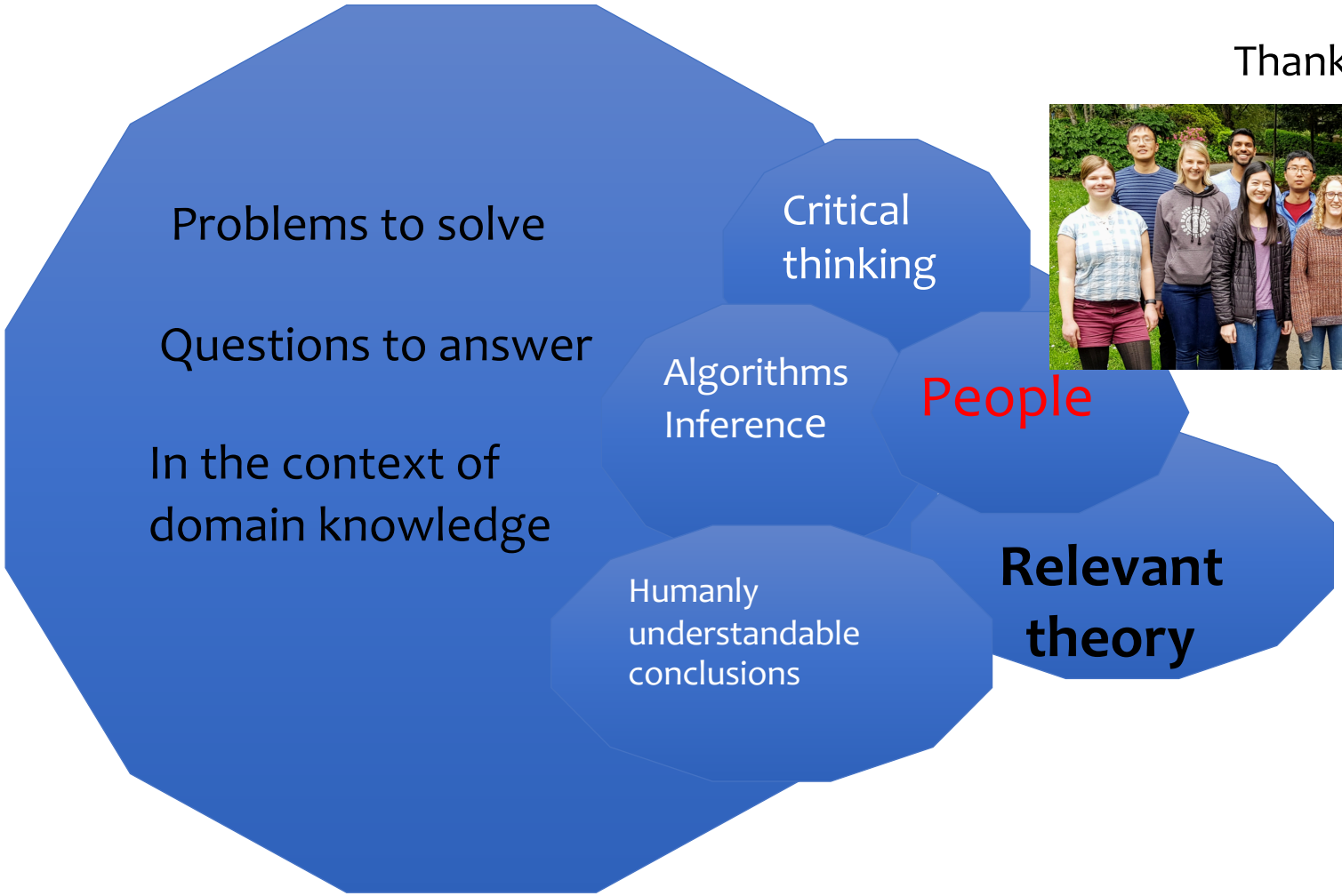
Cardiovascular phenotypes from MRI: (n = 30,000 subjects, continuous trait)

1. So far: aorta size, BAV/TAV, LVM, etc
2. Much less data available (rare variants)
3. Genetic association more complex:

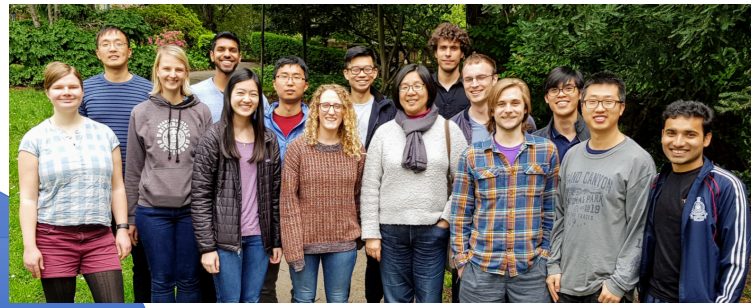
   predictability and stability are both low

1. Need for finding appropriate possibly new phenotype(s)

People make "veridical" happen



Thanks to my group



Upcoming book on veridical data science (2021)

Veridical Data Science: A Book

Bin Yu[1,2] and Rebecca Barter[1]
[1]Department of Statistics, UC Berkeley
[2]Department of Electrical Engineering and Computer Science, UC Berkeley


Berkeley
UNIVERSITY OF CALIFORNIA

What skills does the book teach?

Veridical Data Science (VDS) will teach the critical thinking, analytic, human-interaction and communication skills required to effectively formulate problems and find reliable and trustworthy solutions.
VDS explains concepts using visuals and plain English, rather than math and code.
The primary skills taught are:

Critical thinking

Readers will learn to:

Formulate answerable questions using the data available
Scrutinize all analytic decisions and results
Document all analytic decisions
Appropriate common techniques to unfamiliar situations
Deal with real, messy data

Technical skills

Data processing
Data cleaning
Exploratory Data Analysis
Data merging

Algorithmic
Dimension reduction
Clustering
Least Squares & ML
Regularization

Stability-based inference
Inference
Causal Inference
Perturbation Intervals
Trustworthiness Statements

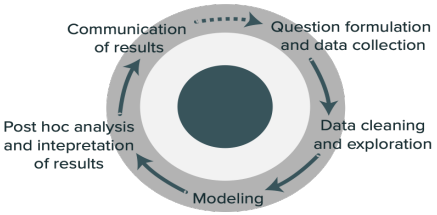
Communication

Exploratory Visual Summaries
Preparing explanatory visual and numeric summaries for explaining data and findings to an external audience

Written reports
Preparing written analytic reports for case studies based on real, messy data

Core guiding principles for the book

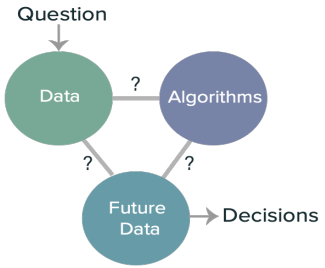
The DS Lifecycle



The Data Science Lifecycle is an iterative process that takes the analyst from problem formulation, data cleaning, exploration, algorithmic analysis, and finally to obtaining a verifiable solution that can be used for future decision-making.

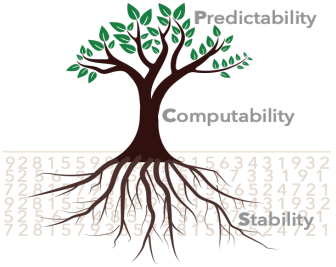
Blending together concepts from statistics, computer science and domain knowledge, the data science life cycle is an iterative process that involves human analysts learning from data and refining their project-specific questions and analytic approach as they learn.

Three realms



Readers will learn to view every data problem through the lens of connecting the three realms:
(1) the question being asked and the data collected (and the reality the data represents)
(2) the algorithms used to represent the data
(3) future data on which these algorithms will be used to guide decision-making.
Guiding the reader to connect the three realms is a means of guiding the reader through the data science lifecycle.

PCS framework



The PCS framework provides concrete techniques for finding evidence for the connections between the three realms.
Predictability: if the patterns found in the original data also appear in withheld or new data, they are said to be predictable. If an analysis or algorithm finds predictable patterns, then these patterns are likely to be capturing real phenomena.
Computability: algorithmic and data efficiency and scalability is essential to ensuring that the results and solutions (e.g. a predictive algorithm) can be efficiently applied to new data.
Stability: minimum requirement for reproducibility. If results change in the presence of minor modifications of the data (e.g. via perturbations) or human analytic decisions, then there might not be a strong connection between the analysis/algorithms and the reality that underlies the data.

Intended Reader/Audience

Anyone who wants to learn the intuition and critical thinking skills to become a data scientist or work with data scientists.
Neither a mathematical nor a coding background is required.
VDS could form the basis of a semester- or multi-semester-long introductory data science university course, either as an upper-division undergraduate or early graduate-level course.

Interested? Get in touch!

Bin Yu
Email: binyu@stat.berkeley.edu
Website: https://www.stat.berkeley.edu/~binyu/Site/Welcome.html

Rebecca Barter
Email: rebeccabarter@berkeley.edu
Website: www.rebeccabarter.com
Twitter: @rlbarter

Thank you!

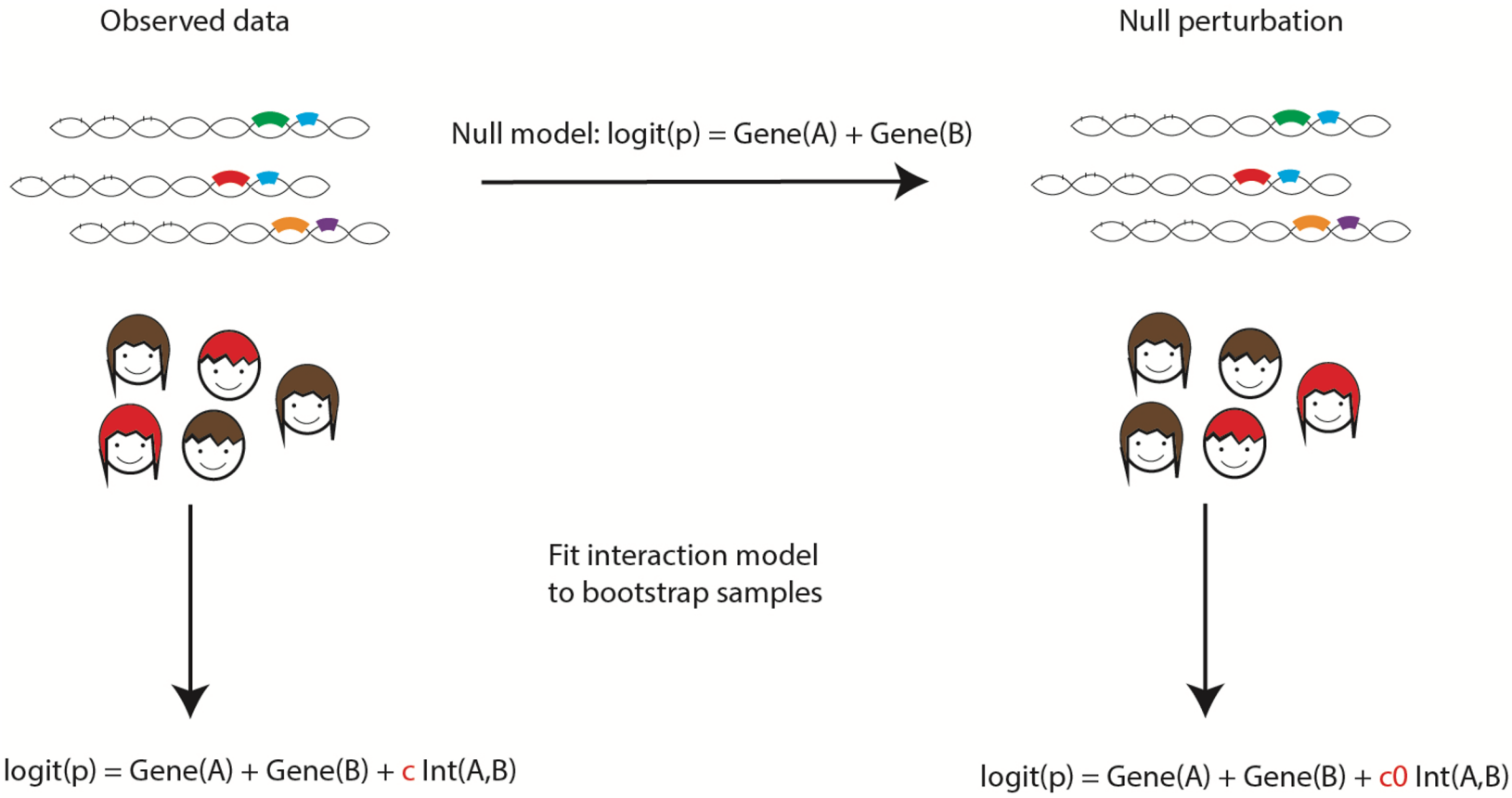
Check out PCS or iRF or epiTree for your projects

1. B. Yu and K. Kumbier (2020), **"Veridical data science",** PNAS. --- PCS framework

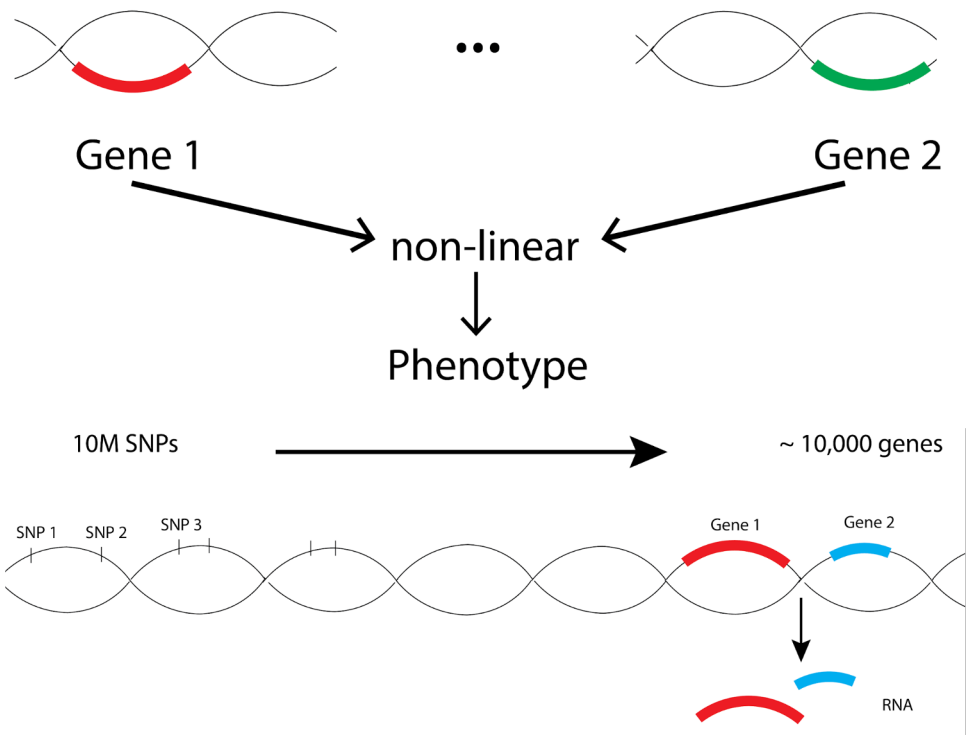
2. S. Basu, K. Kumbier, B. Brown and B. Yu (2018). **"Iterative random forests to discover predictive and stable high-order interactions",** PNAS (code available)

3. **M. Behr, K. Kumbier**, M. Aguirre, R. Arnaout, E. Ashley, A. Butte, R. Arnout, B. Brown**, J. Priest, B. Yu** (2020). **"Learning epistatic polygenic phenotypes with Boolean interactions"** https://www.biorxiv.org/content/10.1101/2020.11.24.396846v1 (code available)

Beyond distributional p-values: PCS p-values



Epistasis: non-linear (higher order) interactions



Biologically inspired dimension reduction: imputed gene expression (Gamazon et al, '15)

Data: UK Biobank

- **Positive control:** well-studied and largely genotype determined redhead phenotype (Morgan et al '18)
- Balanced sample with ~15,000 redheads
- ~ 10M imputed SNPs (~ 800K measured directly) and ~ 10K imputed gene expression

Recovered epistasis (gene level):

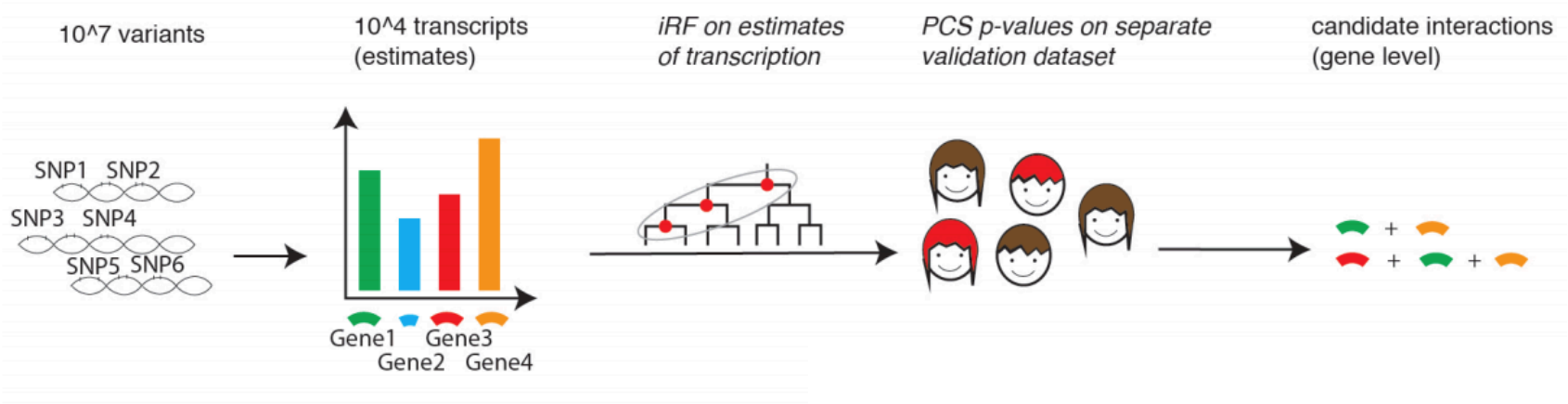
	genes	chr	stability	p_cart	p_cart_PCS	p_multi
1	ASIP , CDK10	20 , 16	1.00	10^-11	10^-5	10^0
2	ASIP , CDK10 , TUBB3	20 , 16 , 16	0.78	10^-2	10^-1	10^0
3	DBNDD1 , SIAH2 , CDK10	16 , 3 , 16	0.72	10^-1	10^-1	10^-1
4	ASIP , CDK10 , SPIRE2	20 , 16 , 16	0.70	10^0	10^0	10^-1
5	DBNDD1 , UPF3A	16 , 13	0.62	10^-4	10^-2	10^0
6	SIAH2 , CDK10	3 , 16	0.54	10^0	10^0	10^0
7	DBNDD1 , ASIP , CDK10 , SPIRE2	16 , 20 , 16 , 16	0.52	10^-2	10^-1	10^0

- Results suggest **higher order (beyond pairwise)** epistasis involving these genes, which were not tested for previously.
- **Previously reported eQTL epistasis** between ASIP (chromosome 20) and MC1R (chromosome 16) related genes, e.g., CDK10, DBNDD1, recovered (Morgan et al '18)
- Results suggest **interactions resembles decision tree structure** as opposed to classical multiplicative

iRF recovers known genetic determinants of hair color & pigmentation:

Gene	**Studies on hair color**
DEF8	Morgan et al(‘18), Lin et al (‘15), Kichaev (‘18)
SPATA2L	Morgan et al(‘18)
SPIRE2	Morgan et al(‘18), Kichaev (‘18)
FANCA	Morgan et al(‘18), Galván-Femenía (‘18), Kichaev et al (‘18), Eriksson et al (‘10),
CDK10	Morgan et al(‘18), Han et al (‘08), Song et al (‘14)--Melanoma risk
CHMP1A	No hair color findings so far! Related traits: Visconti et al (‘18) (low tan response), Kichaev et al (‘18) (sunburns)
CPNE7	Morgan et al(‘18)

epiTRee pipeline in one figure



epiTRee pipeline in one figure