Towards reliable experimental recommendations of gene-gene interactions

Tiffany Tang
Applied and Computational Mathematics and Statistics
University of Notre Dame
ttang4@nd.edu

Our Interdisciplinary Team

PIs


Bin Yu


Ben Brown


Euan Ashley


James Priest


Victoria Parikh


Rima Arnaout


Atul Butte

Students/Postdocs


Tiffany Tang


Ana Kenney


Omer Ronen


Abhi Agarwal


Xiao Li


Karl Kumbier


Merle Behr


Qianru Wang


Nate Youlton


Chad Weldy


Weston Hughes

Outline

1. **Problem Background**
2. **Low-signal Iterative Random Forest (lo-siRF)**
   For recommending genes and gene-gene interactions
3. **Experimental Validation**

Problem Background

A Public Health Crisis: Cardiovascular Disease

Cardiovascular disease (CVD) is the leading cause of death globally and in the US

17.9 million people die each year from CVDs

32% of all deaths worldwide

2x more Americans died from CVD than COVID-19 in 2020

Risk factors for CVD: poor diet, physical inactivity, tobacco use, **genetic factors**

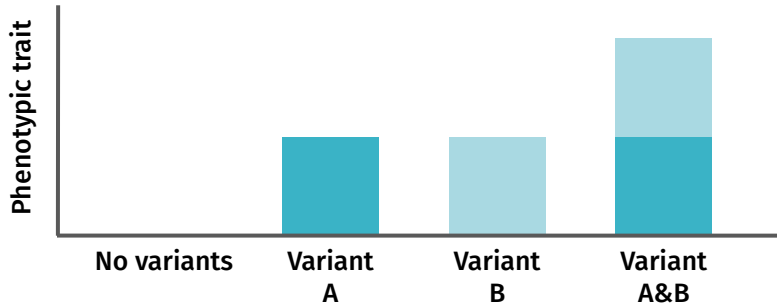
Source: World Health Organization, Ahmad and Anderson (JAMA, 2021)

Traditional tools for detecting genetic effects

Historically, research has focused on identifying genetic variants that have **marginally additive effects** on the phenotype

$$y = \beta_A x_A + \beta_B x_B + \varepsilon$$

phenotype — effect of variant A — effect of variant B — noise



Common tools: genome-wide association studies (GWAS), polygenic risk scores

[Khera et al. (2018), Bycroft et al. (2018), Shah et al. (2020), Pirruccello et al. (2020), Meyer et al. (2020), Harper et al. (2021), Khurshid et al. (2022), ...]

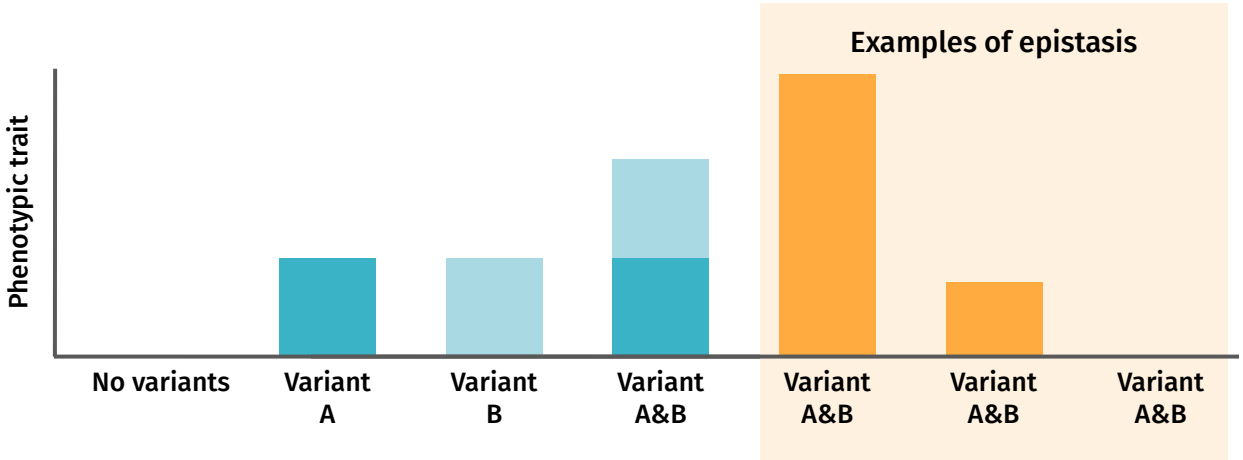
Our biological system is far more complex than this → **"missing heritability"**

Beyond the marginally additive curtain: Epistasis

Epistasis is the non-additive interaction of genetic effects*

$$y = \beta_A x_A + \beta_B x_B + \beta_{AB} x_A x_B + \varepsilon$$

- y: phenotype
- $\beta_A x_A$: effect of variant A
- $\beta_B x_B$: effect of variant B
- $\beta_{AB} x_A x_B$: interaction effect of variant A&B
- ε: noise



* Many technical definitions: see Bateson (1909), Fisher (1919), Wade et al. (2001), Cordell (2002), Ritchie and Van Steen (2018)

Barriers to detecting epistasis

Computational Challenges: exhaustive search is computationally expensive

Ex. 100,000 genetic variants → ~5 billion pairwise interactions!

Statistical Challenges: unclear what is an *appropriate* statistical model

What about higher-order interactions? (Not just pairwise)

What about nonlinear interactions? (e.g., $f(x_A)f(x_B)$, not just $x_A x_B$)

Experimental Challenges: difficult to experimentally assess small effect sizes; requires very precise, high-throughput measurements

Our aim (very broadly)

To develop an **end-to-end pipeline** for identifying **genes** and **gene-gene interactions** that affect **cardiovascular disease.**

Gene / interaction recommendation system

✓ **Computationally-tractable** interaction search engine for **higher-order, nonlinear** interactions

+

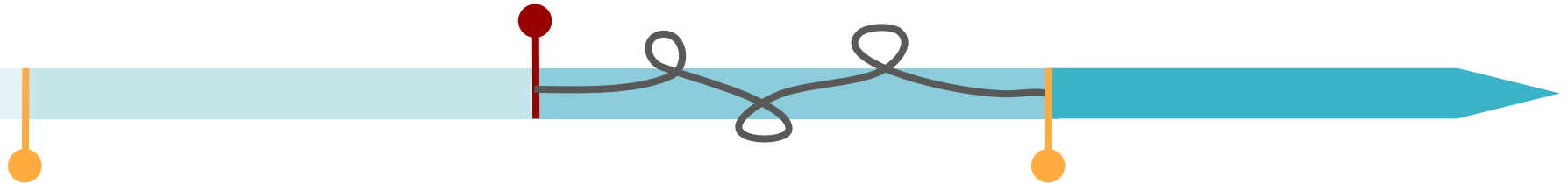
Wet-lab experimental validation

✓ **Precise, high-throughput** phenotyping via microfluidics-enabled gene silencing experiments

Which cardiovascular phenotype?

Major roadblock with **HCM:**

- **~50%** balanced classification accuracy
- Severe under-diagnosis → noisy labels



Attempt #1:

Hypertrophic Cardiomyopathy (HCM)

- High prevalence (~1 in 500)
- Team's clinical expertise
- Experimental capabilities for measuring cell size

Left Ventricular Hypertrophy

Which ca

Attempt #1:
Hypertrophic Cardio
- High prevalence
- Team's clinical e
- Experimental ca
measuring cell s

Hypertrophy

Normal Heart

Left Ventricular Hypertrophy

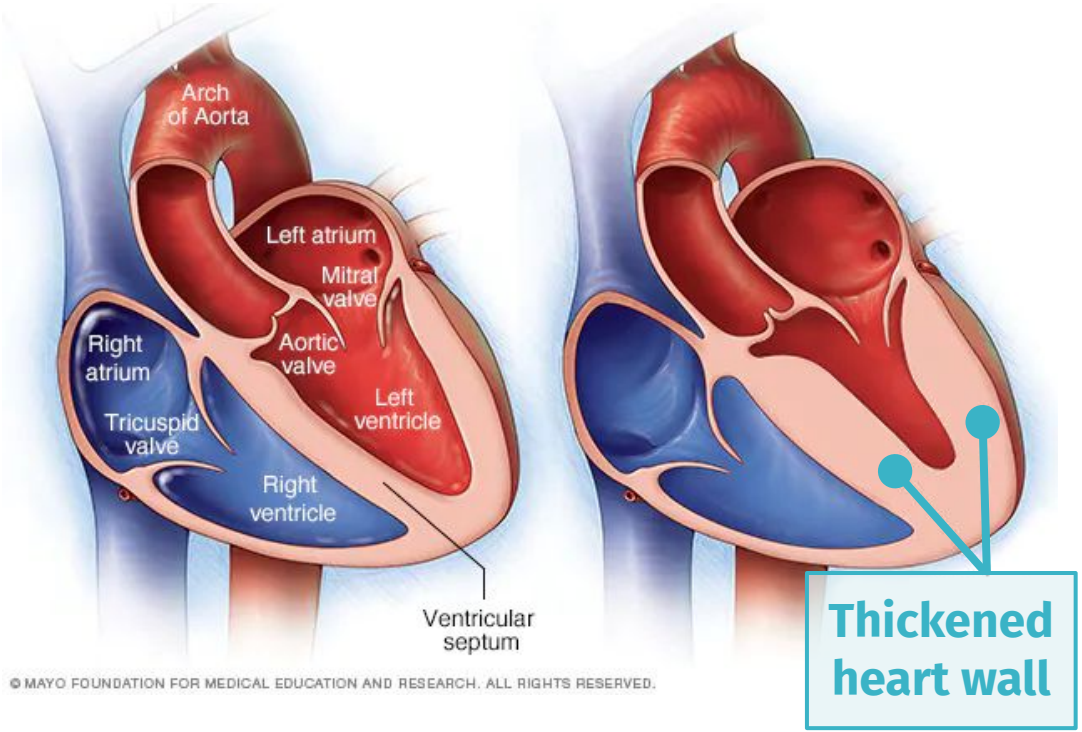
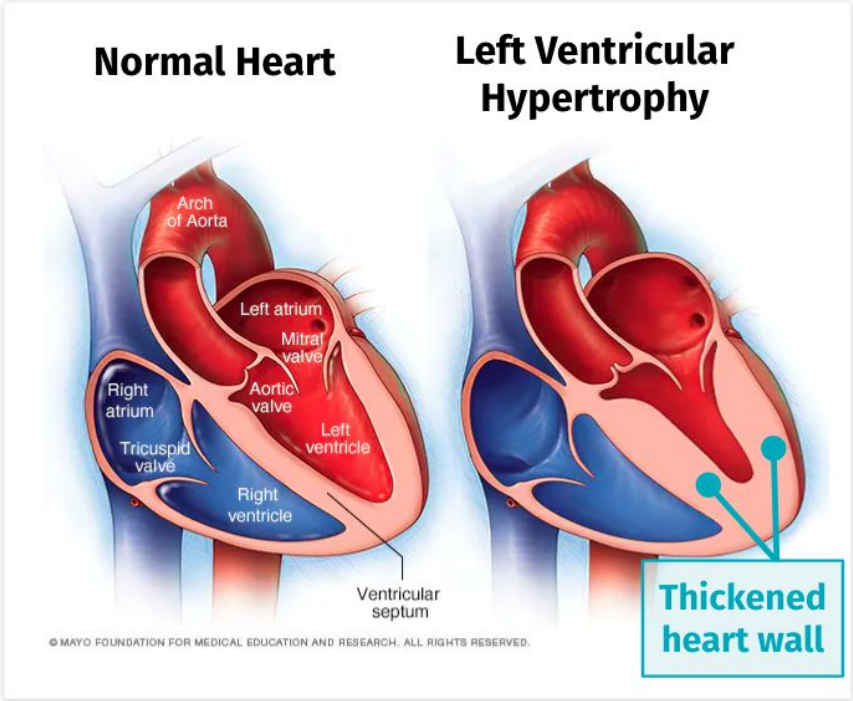


Image Source: https://www.mayoclinic.org/diseases-conditions/left-ventricular-hypertrophy/symptoms-causes/syc-20374314

Left Ventricular Hypertrophy (LVH)

Carries significant independent risk for incident heart failure, atrial arrhythmia, and sudden cardiac death

A distinguishing clinical feature of HCM



Left Ventricular Hypertrophy (LVH)

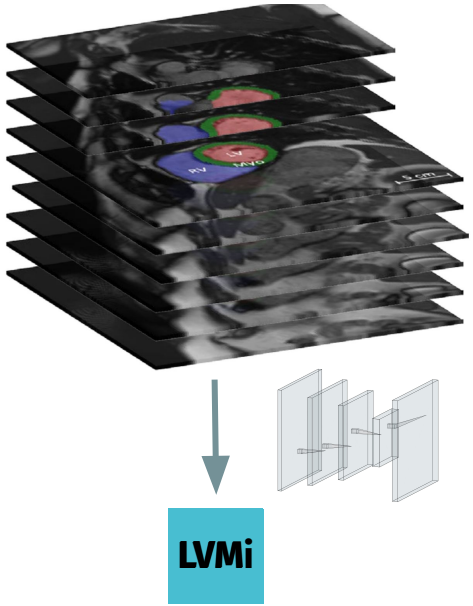
Weston Hughes

Carries significant independent risk for incident heart failure, atrial arrhythmia, and sudden cardiac death

A distinguishing clinical feature of HCM

A quantitative proxy for LVH, **left ventricular mass indexed by body surface area (LVMi)**, can be extracted from cardiac MRIs using deep learning [Bai et al. (2018)]

Cardiac MRI-derived LVMi

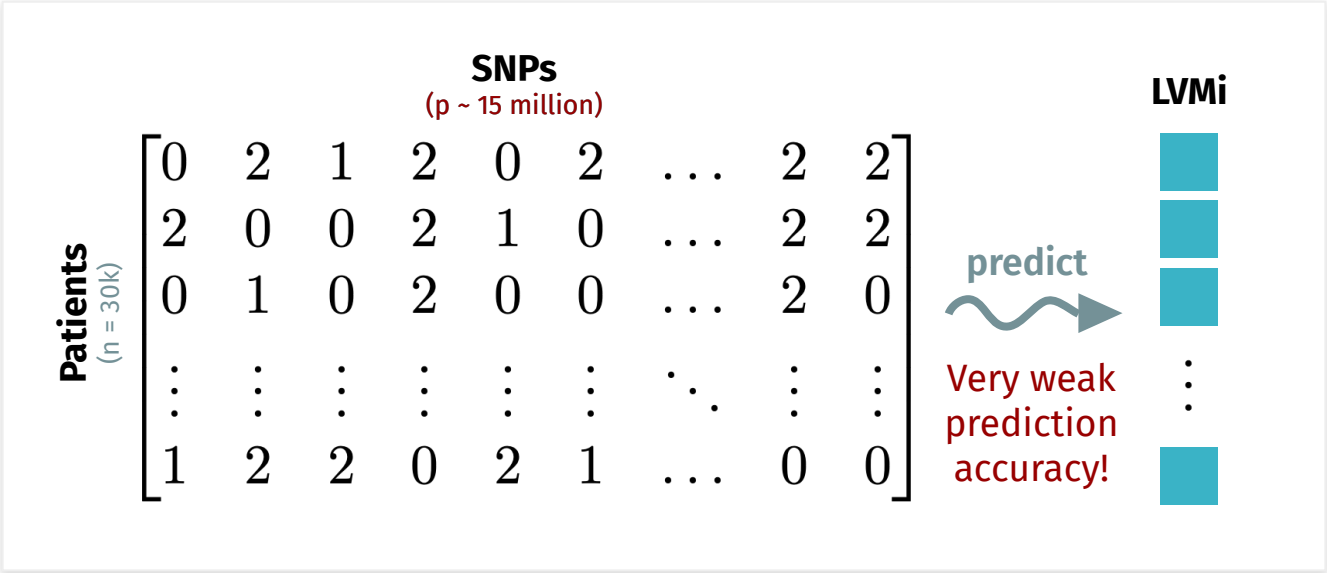


(LVM indexed by body surface area)

UK Biobank Data

n = 30K patients from white British unrelated population with cardiac MRIs

p = 15 million imputed SNPs



Epistasis...

Our contribution

We develop an **end-to-end pipeline** for identifying **genes** and **gene-gene interactions** that affect **left ventricular mass**.

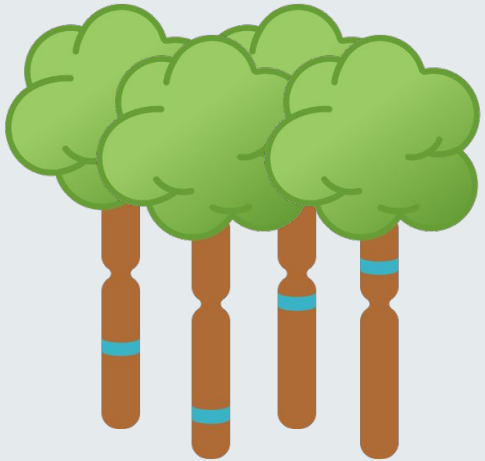
Gene / interaction recommendation system

- ✓ **Computationally-tractable** interaction search engine for **higher-order, nonlinear** interactions
- ✓ Tailored for **low-signal** phenotypes
- ✓ Suitable for **high-dimensional** data

Wet-lab experimental validation

- ✓ **Precise, high-throughput** phenotyping via microfluidics-enabled knockdown experiments

Low-signal iterative random forest (lo-siRF)



Challenges

- **High-dimensionality**
- **Finding interactions**
- **Very low signal**

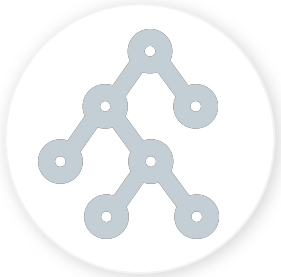
Low Signal iRF (lo-siRF)

- **Domain-inspired dimension reduction**
  via GWAS
- **Iterative Random Forest (iRF)**
  A **computationally-efficient** search engine to find **stable, higher-order, nonlinear** interactions
  [Basu, Kumbier, Brown, Yu (2018)]
- **Binarize LVMi phenotype**
  “Simplifying” the problem
- **A new RF feature importance score**
  Leverages SNP correlations to aggregate weak SNP-level importances into more stable, stronger gene-level importances

lo-siRF for gene (interaction) recommendations

Dimension reduction | **Fit iRF on binarized LVMi** | **Rank genes / interactions**

lo-siRF for gene (interaction) recommendations

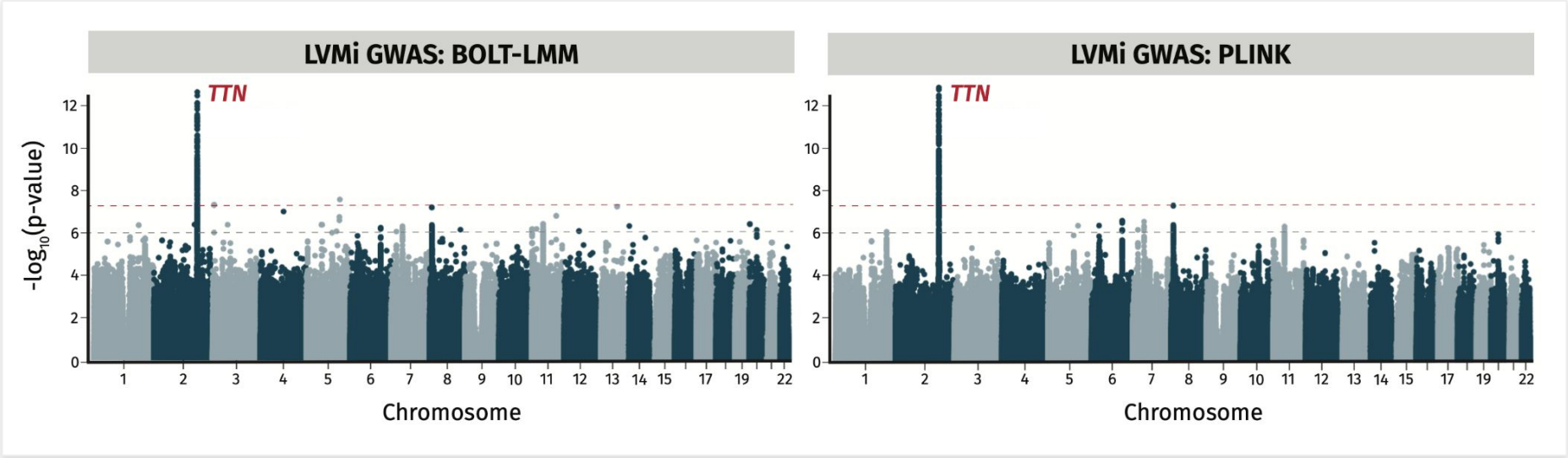


Dimension reduction | **Fit iRF on binarized LVMi** | **Rank genes / interactions**

Dimension Reduction via GWAS

Run GWAS using BOLT-LMM [Loh et al. (2015)] and PLINK [Purcell et al. (2015)]

Select union of top 1000 SNPs from each GWAS run → ~1400 SNPs



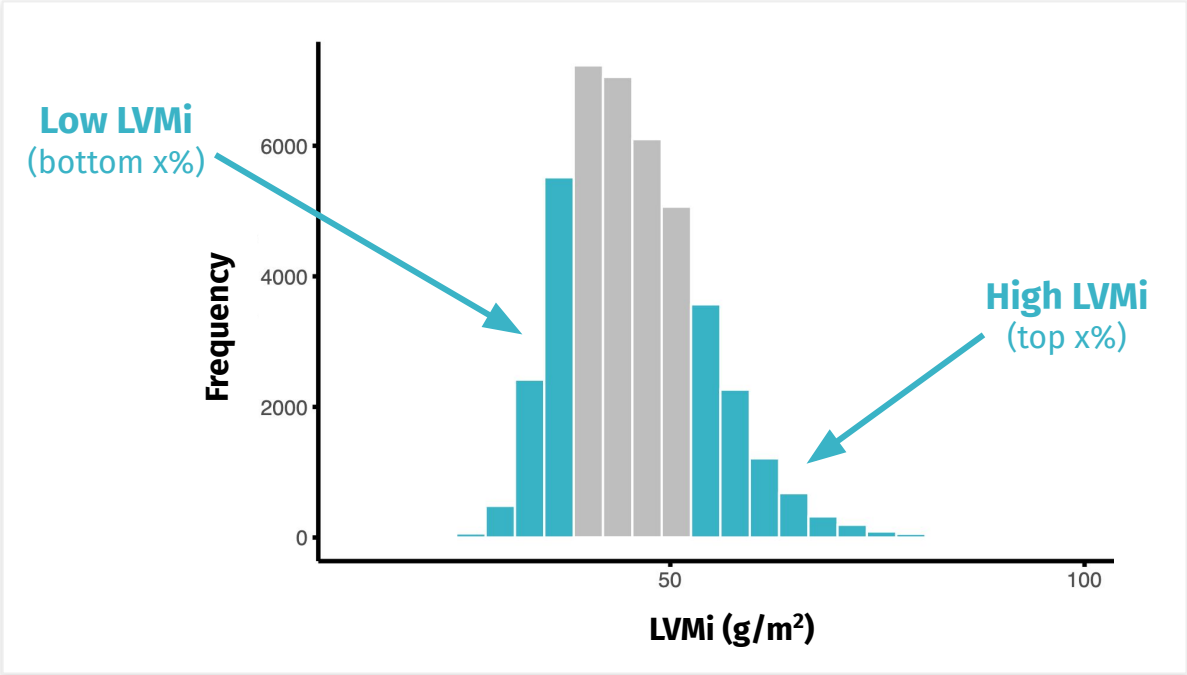
lo-siRF for gene (interaction) recommendations

Dimension reduction | **Fit iRF on binarized LVMi** | **Rank genes / interactions**

- Run **GWAS** using two methods: BOLT-LMM and PLINK
- Select union of top 1000 SNPs from each GWAS method

LVMi Binarization

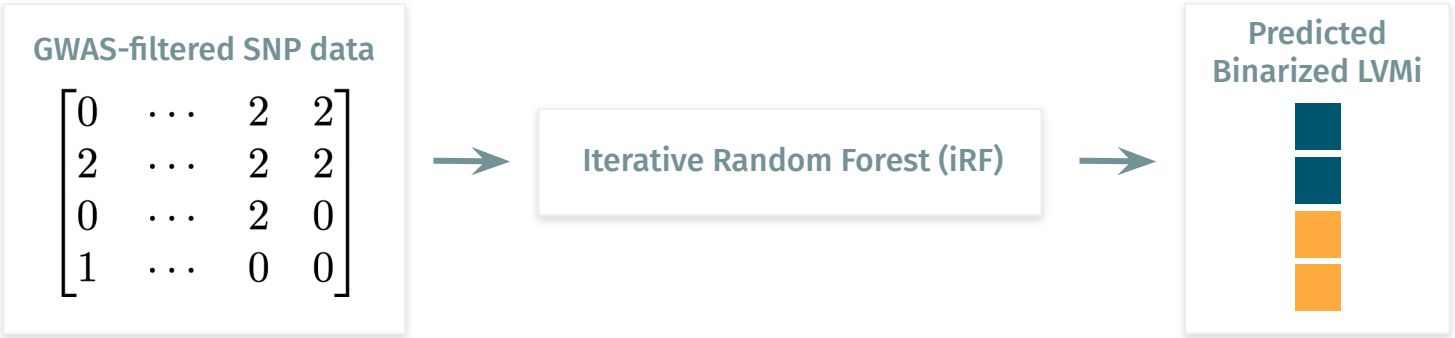
Binarize LVMi phenotype into high and low groups to **“simplify”** the problem (using multiple thresholds: 15%, 20%, 25%)



Fit iRF on the binarized iLVM

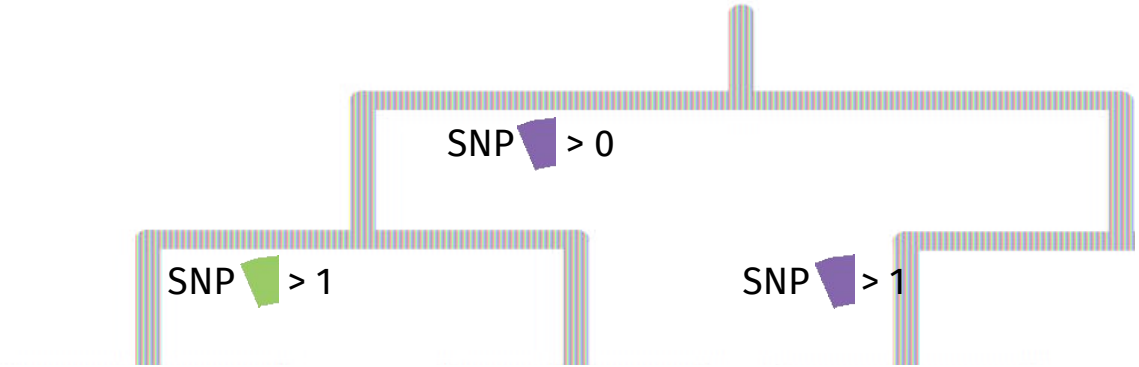
For each binarization threshold (15%, 20%, 25%):

Fit iRF using GWAS-filtered SNP data to predict binarized LVMi



Basu, Kumbier, Brown, Yu. "Iterative random forests to discover predictive and stable high-order interactions." PNAS (2018)

Decision Tree



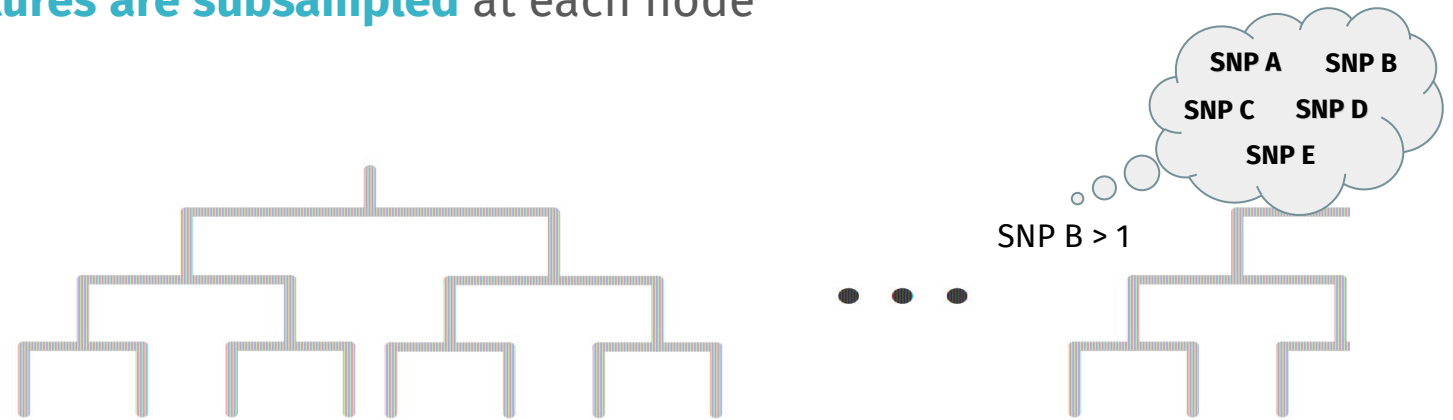
Predicted probability of high LVMi:

0.1 0.8 0.5 0.9

Random Forest

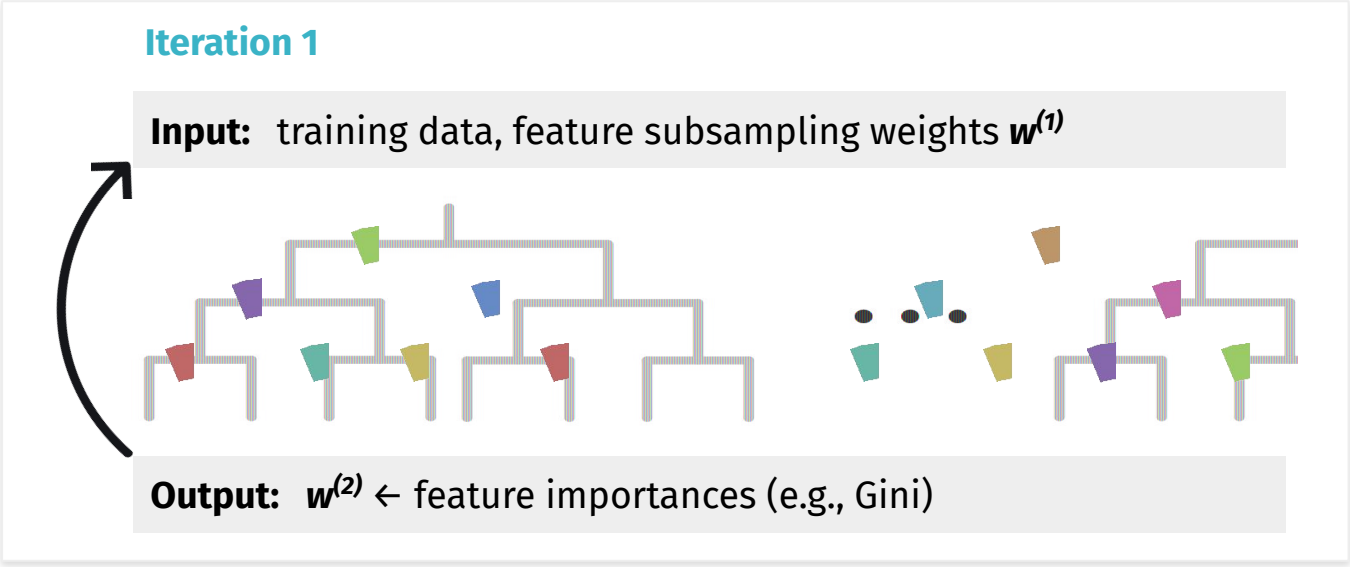
A **collection of decision trees**, where

- each tree is fitted on a different **bootstrap** version of the data
- **features are subsampled** at each node



Iterative Random Forest (iRF)

Core idea: iRF induces **stability** in the RF to improve **interpretability**



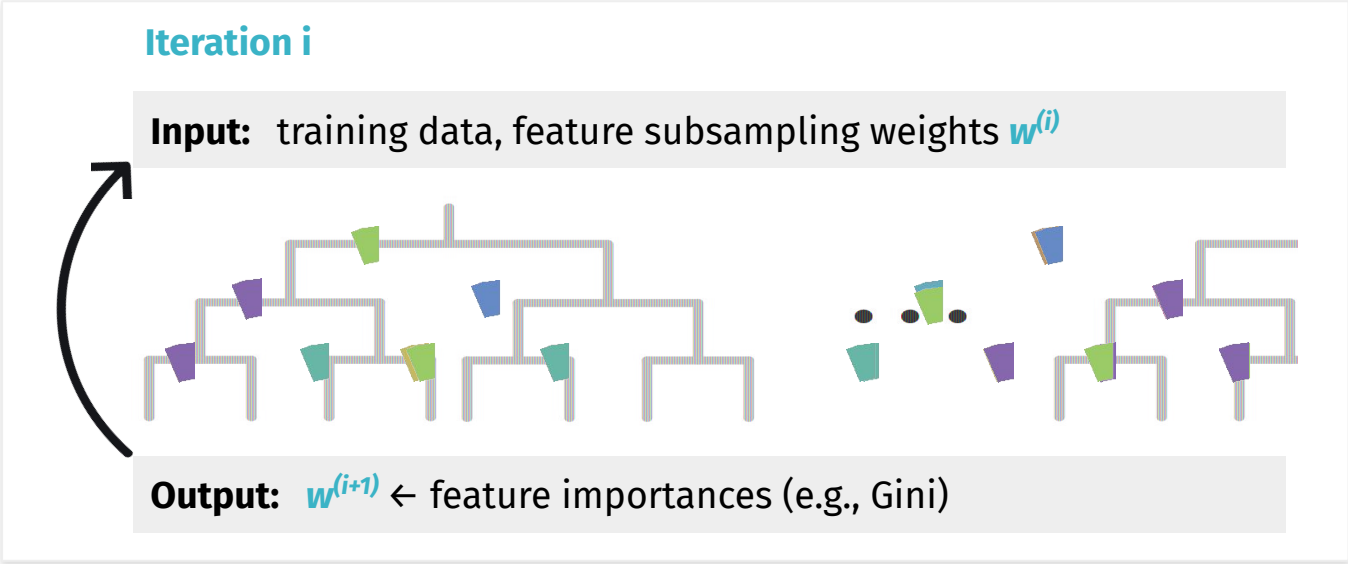
Basu, Kumbier, Brown, Yu. "Iterative random forests to discover predictive and stable high-order interactions." PNAS (2018)

Iterative Random Forest (iRF)

Core idea: iRF induces **stability** in the RF to improve **interpretability**

**without sacrificing prediction accuracy*

iRF → **~55%** classification accuracy for LVMi and better than other ML methods



Basu, Kumbier, Brown, Yu. "Iterative random forests to discover predictive and stable high-order interactions." PNAS (2018)

Iterative Random Forest (iRF)

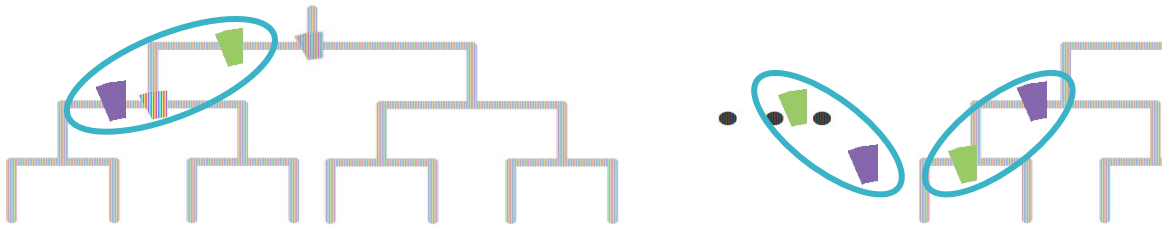
Core idea: iRF induces **stability** in the RF to improve **interpretability**

**without sacrificing prediction accuracy*

iRF → **~55%** classification accuracy for LVMi and better than other ML methods

Last Iteration

Do: Find *interactions* (commonly co-occurring features)



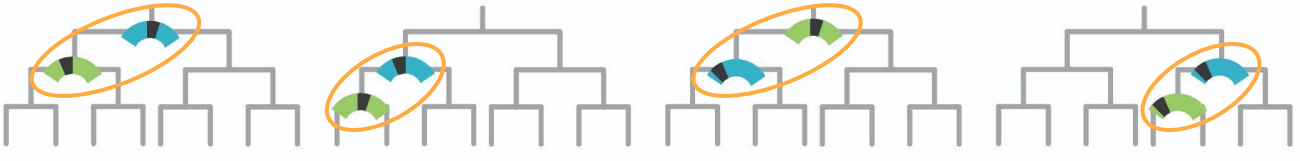
Basu, Kumbier, Brown, Yu. "Iterative random forests to discover predictive and stable high-order interactions." PNAS (2018)

Limitations of iRF

1. iRF identifies candidate interactions based upon their stability within the RF

   Problem: Low-signal phenotype + highly-correlated features → SNP-SNP interactions are highly unstable

   Solution: Aggregate SNPs at the gene level → gene-gene interactions



Substitute SNPs in RF by their **corresponding genes** to take advantage of **correlation** structure between nearby SNPs

Limitations of iRF

1. iRF identifies candidate interactions based upon their stability within the RF

   Problem: Low-signal phenotype + highly-correlated features → SNP-SNP interactions are highly unstable

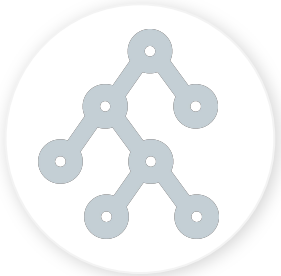
   Solution: Aggregate SNPs at the gene level → gene-gene interactions

2. iRF ranks candidate interactions based upon their frequency within the RF

   Problem: Longer genes will naturally be more frequent in the RF

   Solution: A new gene-level (or group) feature importance score

lo-siRF for gene (interaction) recommendations



Dimension reduction

- Run **GWAS** using two methods: BOLT-LMM and PLINK
- Select union of top 1000 SNPs from each GWAS method

Fit iRF on binarized LVMi

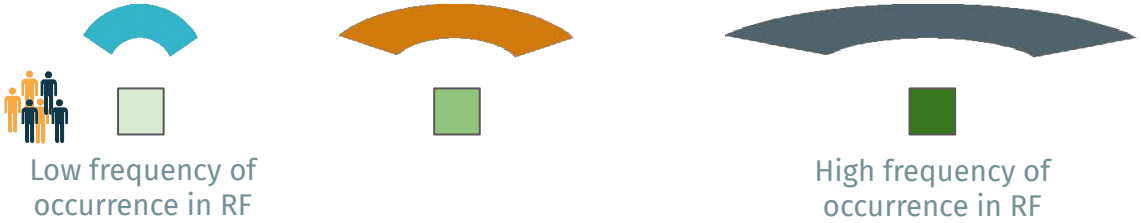
- **Binarize** iLVM phenotype into high and low groups to *"denoise"* (using multiple thresholds: 15%, 20%, 25%)
- Fit **iRF** on SNP data to extract candidate gene interactions

Rank genes / interactions

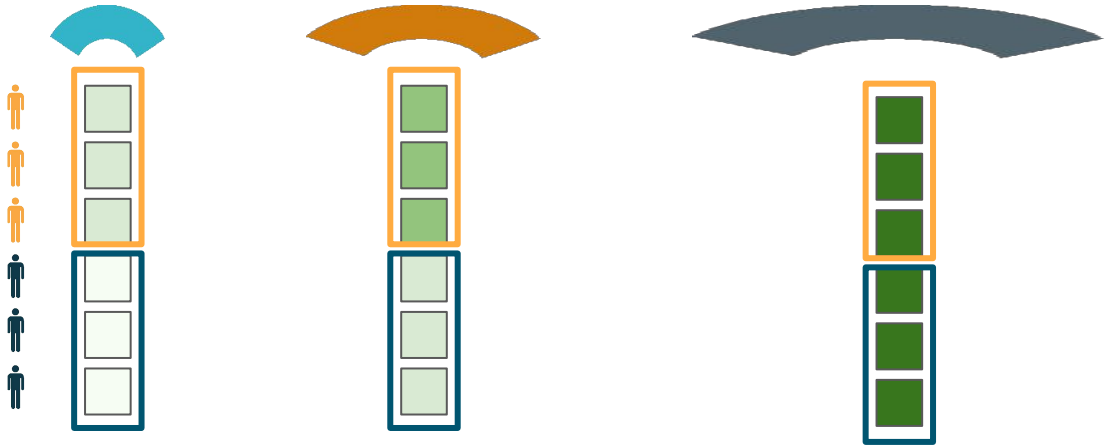
- Using a **new stability-based importance score** to aggregate SNP-level importances from iRF into a **gene-level** score

A paradigm shift for feature importances

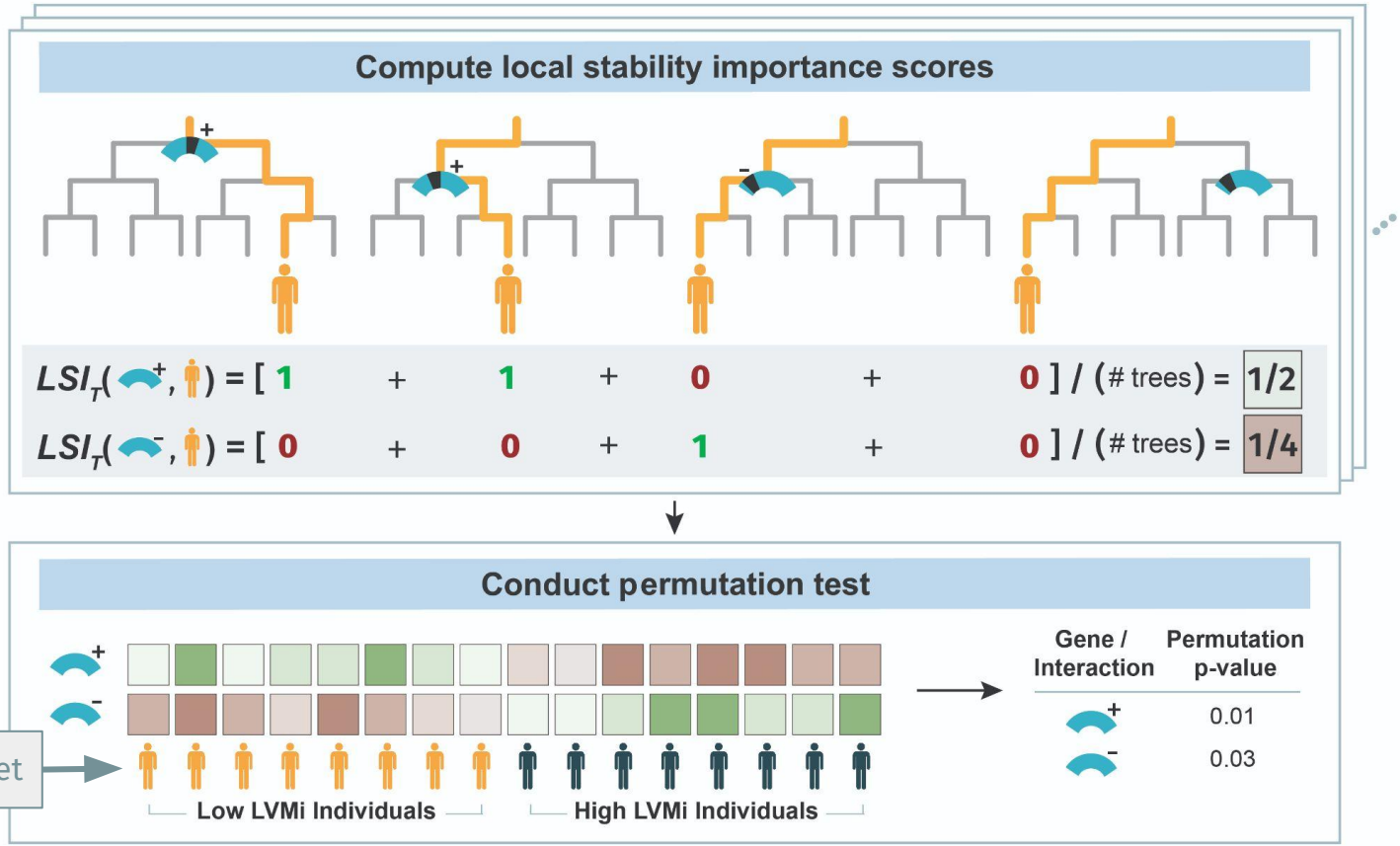
iRF
"Global importance"



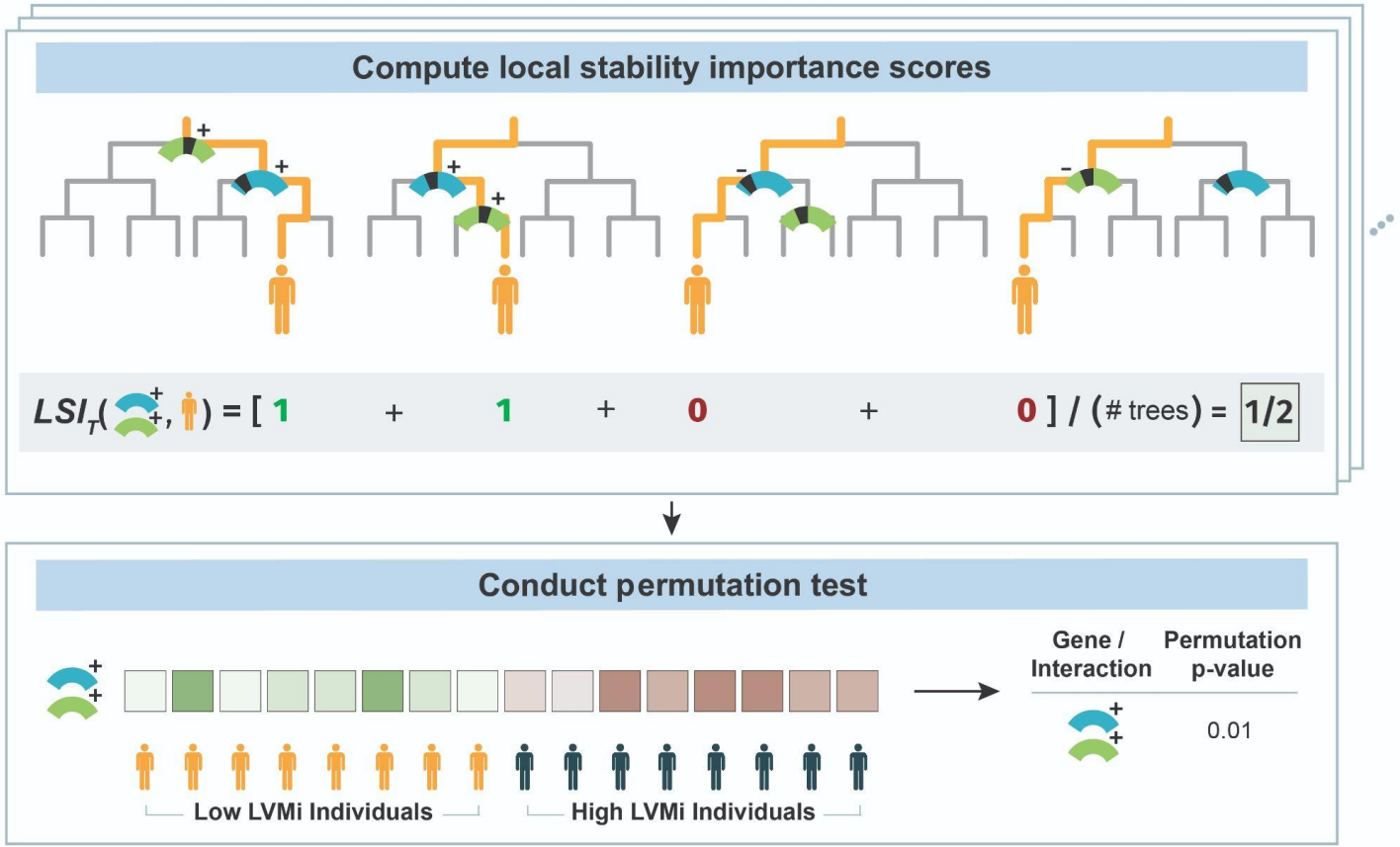
lo-siRF
"Local importance"
(per-individual)



A new stability-based feature importance score for RF



Extending to gene-gene interactions



lo-siRF for gene (interaction) recommendations

Dimension reduction

- Run **GWAS** using two methods: BOLT-LMM and PLINK
- Select union of top 1000 SNPs from each GWAS method

Fit iRF on binarized LVMi

- **Binarize** iLVM phenotype into high and low groups to *"denoise"* (using multiple thresholds: 15%, 20%, 25%)
- Fit **iRF** on SNP data to extract candidate gene interactions

Rank genes / interactions

- Using a **new stability-based importance score** to aggregate SNP-level importances from iRF into a **gene-level** score

A red-herring test for our collaborators


Chad Weldy

“Domain expert opinion solicitation with negative controls”

We presented three lists to our cardiology experts:

1. Top-ranked genes/interactions
2. Mid-ranked genes/interactions
3. Random genes/interactions

Collaborators (Chad and Euan) chose list #1 :)

LVMi gene & gene-gene interaction recommendations

Considering only those genes and gene-gene interactions that were stably important across all three binarization thresholds, **lo-siRF identifies**

- **genes** that are **well-known** to impact cell size

  TTN *IGF1R*

- **plausible candidate genes** that are known to be **associated with the heart**

  CCDC141 *RSPO3* *LSP1*

- **stable gene-gene interactions**

  CCDC141–IGF1R *CCDC141–TTN* *CCDC141–TNKS*

Experimental Validation

Overview: Experimental Workflow


Qianru Wang


Nate Youlton

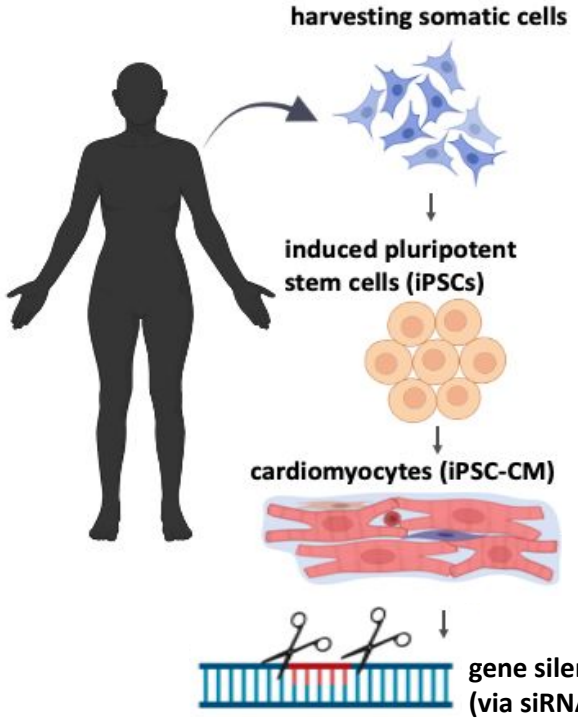
How do the size of heart cells change when we silence a gene or pair of genes?

Silenced genes/gene pairs:

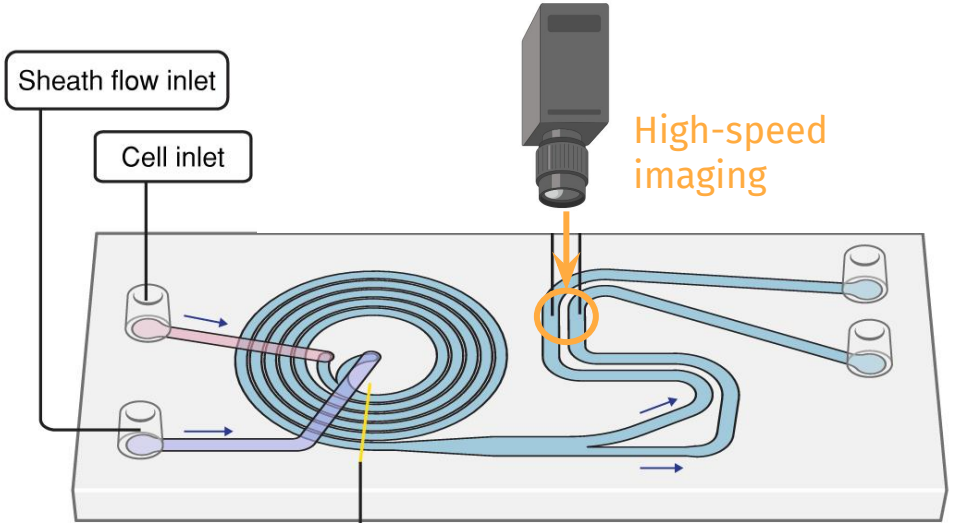
1. *CCDC141*
2. *IGF1R*
3. *TTN*
4. *CCDC141* and *IGF1R* (interaction)
5. *CCDC141* and *TTN* (interaction)

Across two cell lines:

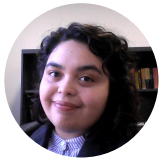
1. Healthy cell line
2. HCM cell line



High-throughput microfluidics + image processing

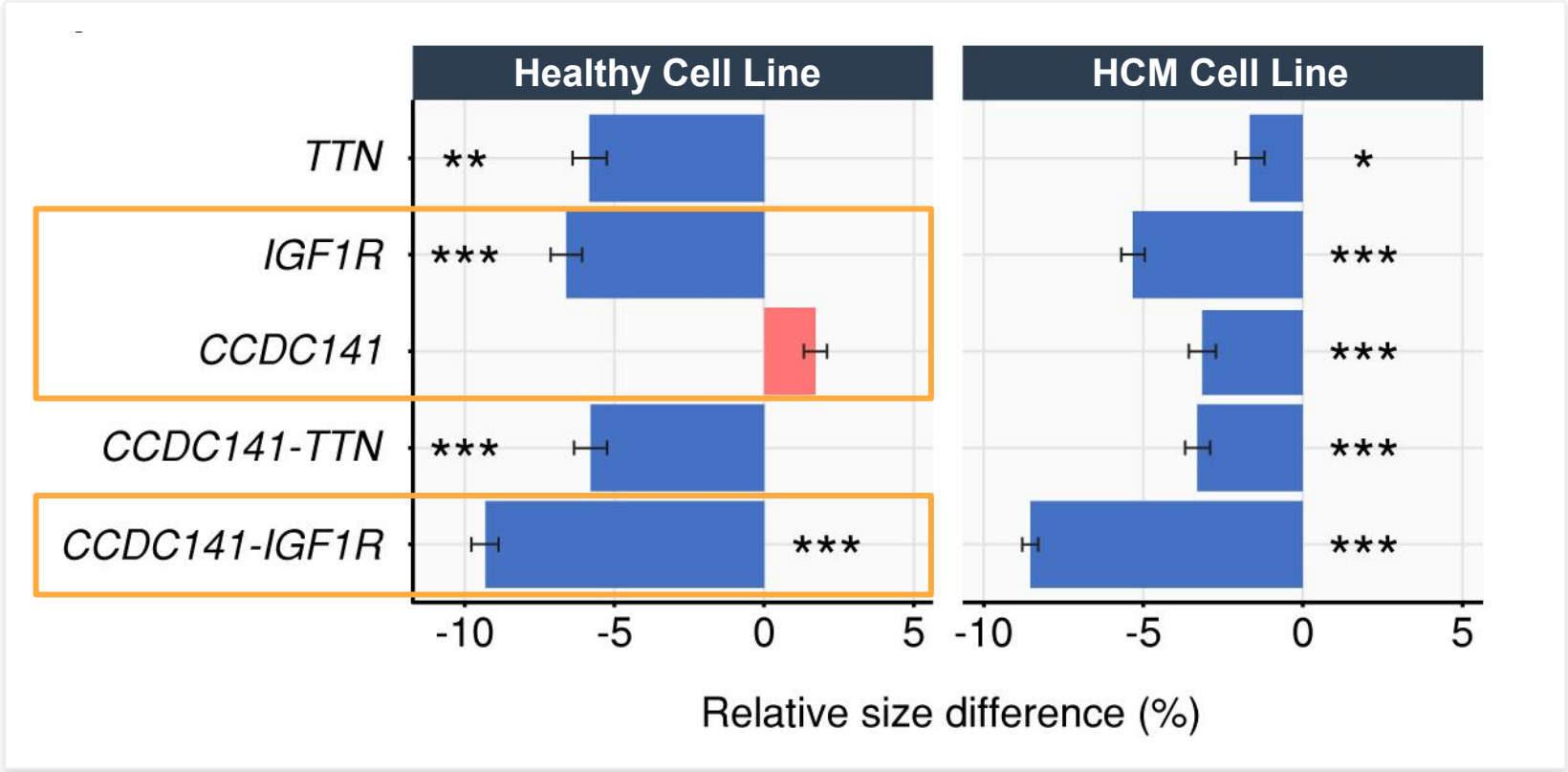



Qianru Wang

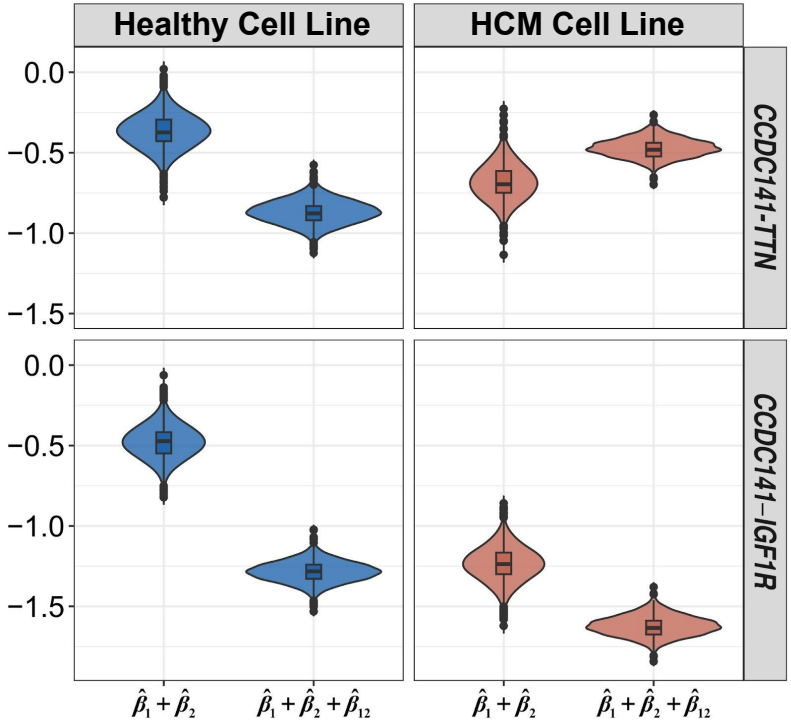

Ana Kenney


Omer Ronen

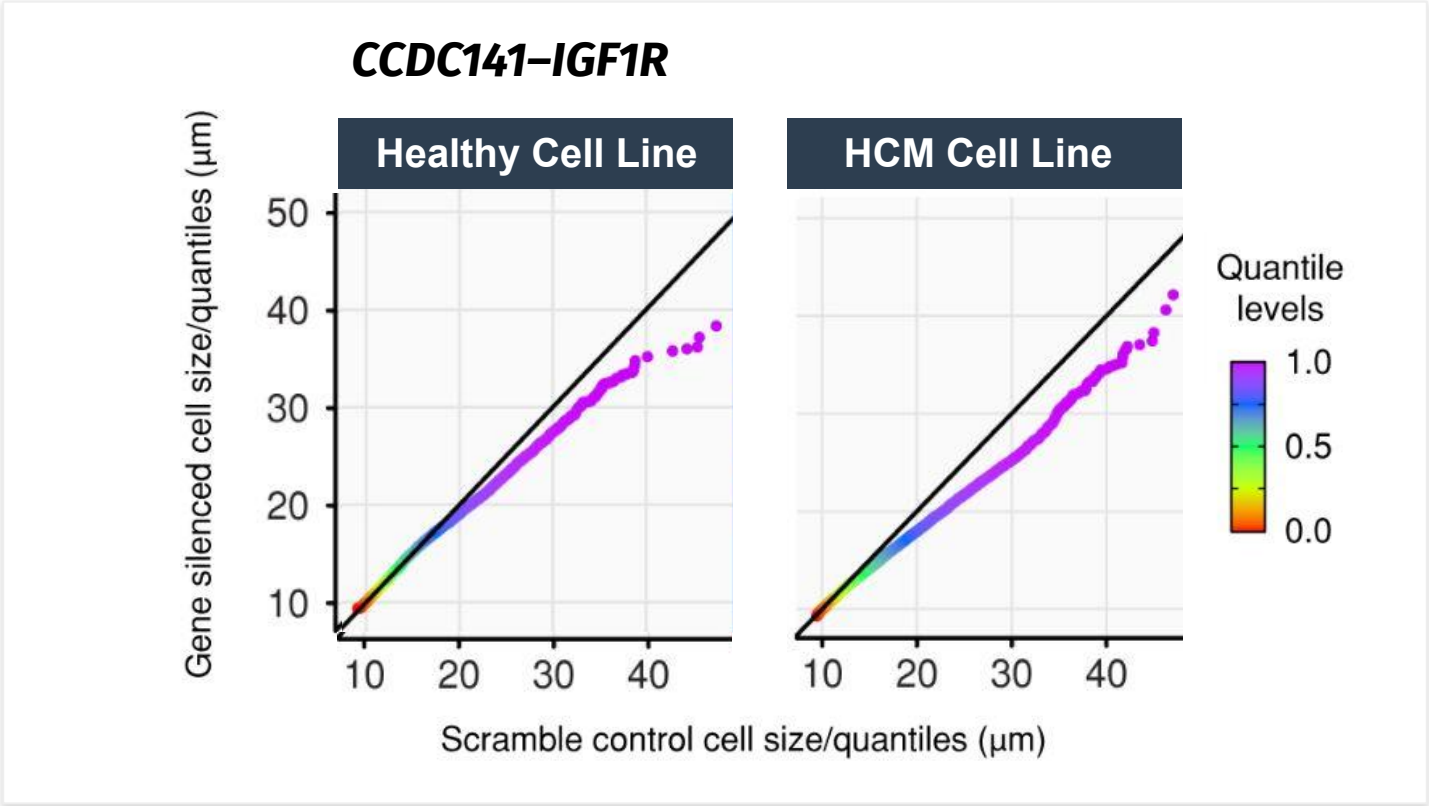
Knocking down genes led to decrease in heart cell sizes



Epistatic effect sizes



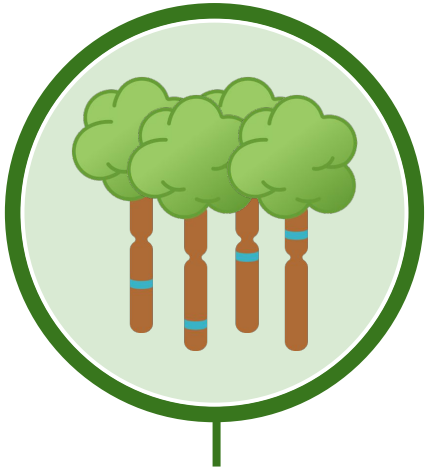
Size differences are most pronounced for large heart cells



Key Takeaways

Scientific Discovery

Experimentally validated epistatic regulation of cardiac hypertrophy



lo-siRF

A gene and gene-gene interaction search engine, tailored for low-signal data

Only possible through collaboration

Iterative cycle of feedback is crucial for a successful and impactful interdisciplinary collaboration

Thank you!

Q. Wang*, TT*, ..., B. Yu, E. Ashley. “Epistasis regulates genetic control of cardiac hypertrophy.” *Nature Cardiovascular Research* (2025).