|  |  |  |
| --- | --- | --- |
| **eMERGE Network: External Collaborator Manuscript Concept Sheet** | | |
| **Reference Number**  *(to be assigned by CC)* | NT298 | |
| **Submission Date** | 7/23/2018 | |
| **Project Title** | Association between variants in FBN1 and human height | |
| **Tentative Lead Investigator** *(first author)* | S. Asgari ([sasgari@broadinstitute.org](mailto:sasgari@broadinstitute.org)) | |
| **Tentative Senior Author**  *(last author)* | S. Raychaudhuri ([soumya@broadinstitute.org](mailto:soumya@broadinstitute.org)) | |
| **eMERGE Site Sponsor & Contact** | S. Weiss (restw@channing.harvard.edu) | |
| **All Other Authors** | S. Weiss, E. Karlson | |
| **Sites Participating** | We invite each site to join | |
| **Background / Significance** | Height is a highly heritable, classic polygenic trait [(Fisher 1919; Galton 1886; Visscher et al. 2006)](https://paperpile.com/c/u115be/1YPJ+vEf3+bf7C). Previous height genetic studies, done predominantly in European populations, have identified ~700 independent height-associated variants [(Wood et al. 2014; Lango Allen et al. 2010)](https://paperpile.com/c/u115be/iWDW+VxDk). However, these variants cannot explain a large portion of phenotypic variance in non-European populations [(Vilhjálmsson et al. 2015; Martin et al. 2017)](https://paperpile.com/c/u115be/G7fJ+cYbV). Our group recently performed the first large-scale genetic study of height in the Peruvian population (<https://bit.ly/2xzn2bf>). We identified a new locus on chromosome 15 that was significantly (p-value < 5e-8) associated with human height. This locus included five SNPs in FBN1. The top SNP (MAF ~0.05) was associated with a 2.2 cm reduction of height in the Peruvians population, making this variant one of the largest effect common variants ever reported. Among the five significantly associated variants in FBN1, one was a missense variant and the other four were intronic variants. These five variants were in high-LD in our data (0.94 ≤ pair-wise r2 ≤ 1). We used fine mapping strategies to pinpoint the likely causal SNP among these five SNPs. All five variants had similar posterior probabilities (0.18 < p < 0.23) making it impossible to pinpoint the variant that is deriving the observed effect on height.  *FBN1* encodes a preproprotein which is proteolytically processed to generate two proteins Fibrillin-1 and asprosin. Fibrillin-1 is an extracellular matrix glycoprotein that serves as a structural component of calcium-binding microfibrils. These microfibrils provide force-bearing structural support in elastic and non-elastic connective tissue throughout the body [(Sakai et al. 1986; Lin et al. 2002)](https://paperpile.com/c/u115be/jWq0+GQqX). Missense mutations in FBN1 are most well known to cause Marfan syndrome, but can cause a number of other diseases, most of which include skeletal problems among other symptoms. Marfan syndrome for example is characterized by tall stature, while stiff skin syndrome and acromicric dysplasia are characterized by short statute [(Ramachandra et al. 2015)](https://paperpile.com/c/u115be/GGon).  Our Aims are to:  1- To test the association between common (MAF ≥ 0.01) variants in FBN1 and height in order to pinpoint the casual variant among the five FBN1 height-associated variants and/or to discover new height associations.  2- To test the association between the burden of rare (MAF < 0.01) variants in FBN1 and height.  3- To explore the association between the five FBN1 height-associated variants with the following phenotypes:   1. Height 2. Diffuse diseases of connective tissue (ICD9 code: 710) 3. Systemic sclerosis (710.1) 4. Sicca syndrome (710.2) 5. Dermatomyositis (710.3) 6. Polymyositis (710.4) 7. Other specified diffuse diseases of connective tissue (710.8) 8. Unspecified diffuse connective tissue disease (710.9). | |
| **Outline of Project** | ANALYSIS PLAN  1- Common variant (MAF ≥ 0.01) association testing: We will use GEMMA [(Zhou and Stephens 2014)](https://paperpile.com/c/u115be/cBKr) to performed association analysis using linear mixed model. GEMMA allows accounting for population structure and cryptic relatedness. Age, gender and the relevant PCs will be included as covariates in the model.  2- Rare variants (MAF < 0.01) association test:  A) SKAT [(Wu et al. 2011)](https://paperpile.com/c/u115be/MMKj): Null distributions are generated using SKAT\_NULL\_emmaX, which incorporates kinship structure in calculation of SKAT parameters and residuals. Age, gender and a number of relevant PCs will be included as covariates. For each gene, we test the association between height and i) all rare variants, ii) rare protein-altering variants (Loss of function (LOF) and missense) as annotated by SnpEff [(Cingolani et al. 2012)](https://paperpile.com/c/u115be/enEW).  B) Burden test: For each gene, we aggregate the information for multiple rare variants into a single burden score (*C*) as follows:  *Ci= Gij*  For subject *i*, *Gi*={ *Gi1*, *Gi2*, …, *Gim*} denotes the allele counts {0,1,2} for the *m* variants in the gene. We test the association between this score and height a linear mix model framework that allows accounting for population structure and cryptic relatedness. Age, gender and the relevant PCs will be included as covariates in the model.  3- PheWAS: We will use a linear mix model framework as implemented in lme4 package in R [(Bates et al. 2015)](https://paperpile.com/c/BVQLkv/FDSy)  that allows accounting for population structure and cryptic relatedness, to test for association between the five FBN1 height-associated variants with the phenotypes listed above*.* Age, gender and the relevant PCs will be included as covariates in the model. | |
| **Desired Data - Common Variables\***  *(Available from the CC)* | Demographics  ICD9/10 codes  CPT codes  Phecodes  BMI | Common Variable Labs  Common Variable Meds  Other: Case/Control status on Phase I and Phase II phenotypes |
| **Other Desired Data *(Available from participating sites)*** | *Please specifically list out any data elements that participating sites would collect or extract from clinical or other sources for this project (i.e. not common variables above)*  FBN1 sequencing data | |
| **Desired Genetic Data** | eMERGE I-III Merged set (HRC imputed, GWAS)  eMERGE PGx/PGRNseq data set  eMERGEseq data set (Phase III)  eMERGE Whole Genome sequencing data set  eMERGE Exome chip data set  eMERGE Whole Exome sequencing data set  Other (not listed above): | |
| **Does project pertain to an existing eMERGE Phenotype?** | Yes, if so please list  No | |
| **Planned Statistical Analyses** | Linear mixed models  PheWAS  SKAT  Gene burden test | |
| **Ethical Considerations** | This study makes use of data that is already obtained from electronic health records. The chart reviews will be performed in the Partners HealthCare Biobank subjects, a process that is IRB approved at Partners HealthCare. | |
| **Available Funding or Resources** | N/A | |
| **Target Journal** | TBD | |
| **Milestones**  *(This section should include the key dates for completion of project, including approval, project duration, draft completion, and submission.)* | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Aug | Sep | Oct | Nov | | 1 & 2. Rare and common variant association testing |  |  |  |  | | 3. PheWAS |  |  |  |  | | 4. Publication |  |  |  |  | | |

**\*Common Variables available across all datasets:**

* Demographics: sex, year of birth, decade of birth, race, ethnicity
* Codes: (repeated values & age at event): ICD, CPT, Phecodes
* BMI: (repeated value & age at event) height, weight, BMI
* Labs: (lab name, repeated lab value & age at event) Serum total cholesterol, LDL, HDL, Triglycerides, Glucose fasting/non-fasting/unknown, & White Blood Cell count
* Medications: (medication name, repeated, & age at event) Cerivastatin sodium, Rosuvastatin, Simvastatin, Fluvastatin, Pravastatin, Lovastatin, Atorvastatin, & Pitavastatin
* Other: Case/Control status on Phase I and Phase II phenotype: only on GWAS dataset participants