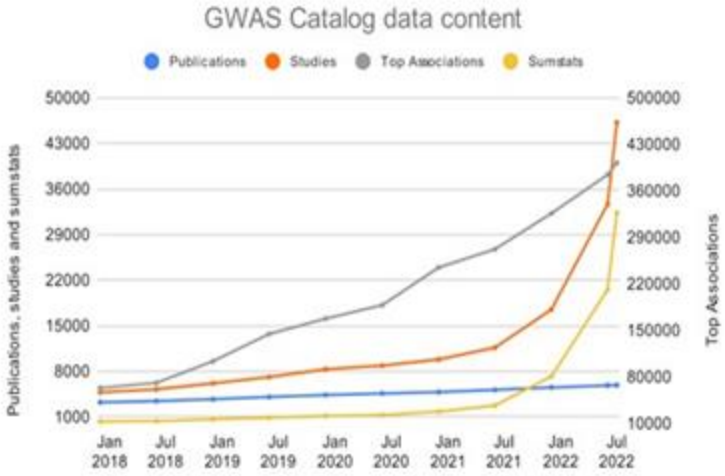
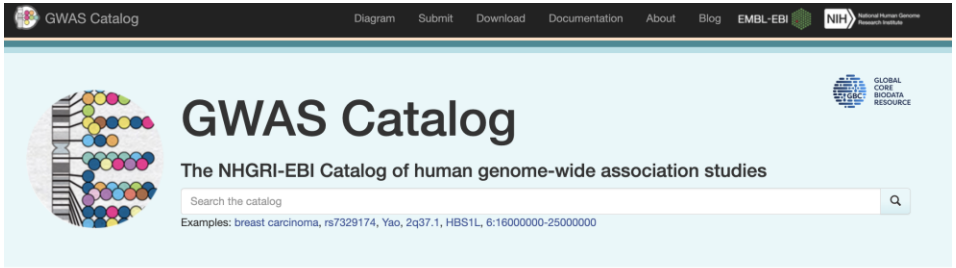
A phenome-wide association study of short tandem repeats in the UK Biobank

Celine Manigbas
NeuroPsychGen Works in Progress
PI: Dr. Andrew Sharp
December 14th, 2023

Icahn School of Medicine at Mount Sinai

Genome-wide association studies (GWAS) have identified thousands of risk loci for human disease and traits





However, GWAS signals can not fully explain estimated heritability, suggesting that other type of variants can be involved in modulating human traits

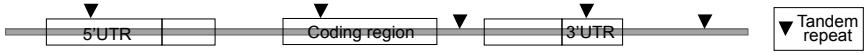
Short Tandem Repeats (STRs) represent an important source of genetic variation in the human genome

- Short tandem repeats (STRs) are segments of DNA with multiple consecutive copies of a sequence unit (motif) ranging from 1 to 6 bp.

CTACAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATACTATACTAGACTAT

Motif: GATA; 10 copies

- Over 1M STRs are annotated across the genome.



- Due to their higher mutation rate, STRs are among the most polymorphic variants in the human genome.

Genome 1	CTACAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATACTATACTAGACTAT
Genome 2	CTACAGATAGATAGATAGATAGATAGATAGATAGATAGATACTATACTAGACTAT
Genome 3	CTACAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATACTATAC
Genome 4	CTACAGATAGATAGATAGATAGATAGATAGATACTATACTAGACTAT
Genome 5	CTACAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATACTATACTAGA
Genome 6	CTACAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGATACTATAC

Motif = GATA, range = 7-12 copies

Rare expanded alleles in STRs have been shown to play a role in 60+ human diseases

Disorder	Affected gene	Repeat[a]	Repeat location	Normal repeat no.	Symptomatic repeat no.
Coding repeats					
DRPLA	*ATN1*	CAG	ORF	7–25	49–88
HD	*HTT*	CAG	ORF	6–34	36–180
SBMA	*AR*	CAG	ORF	11–24	40–62
SCA1	*ATXN1*	CAG	ORF	6–39	39–83
SCA2	*ATXN2*	CAG	ORF	15–29	34–59
SCA3	*ATXN3*	CAG	ORF	13–36	55–84
SCA6	*CACNA1A*	CAG	ORF	4–16	21–30
SCA7	*ATXN7*	CAG	ORF	4–35	34–>300
SCA17	*TBP*	CAG	ORF	25–44	45–66
Noncoding repeats					
DM1	*DMPK*	CTG	3′ UTR	5–37	>50–>2000
DM2	*CNBP*	CCTG	Intron 1	<27	75–11,000
EPM1	*CSTB*	$(C)_4G(C)_4GCG$	Promoter	2–3	30–75
FXS	*FMR1*	CGG	5′ UTR	6–52	~55–>2000
FRAXE MR	*AFF2/FMR3*	CCG	5′ end	6–25	>200
FRA12A MR	*DIP2B*	CGG	5′ UTR	6–23	?
FRDA	*FXN*	GAA	Intron 1	7–22	>66–>900
SCA10	*ATXN10*	ATTCT	Intron 9	10–29	280–4500
Coding and noncoding repeats					
SCA8	*ATXN8/ATXN8OS*	CAG and CTG	ORF and NCT	6–37	~107–250
Repeats with uncertain location					
HDL-2	*JPH3*	CAG/CTG	?	<50	>50
SCA12	*PPP2R2B*	CAG/CTG	?	<66	>66

Rare expanded alleles in STRs have been shown to play a role in 60+ human diseases

Disorder	Affected gene	Repeat[a]	Repeat location	Normal repeat no.	Symptomatic repeat no.
Coding repeats					
DRPLA	*ATN1*	CAG	ORF	7–25	49–88
HD	*HTT*	CAG	ORF	6–34	36–180

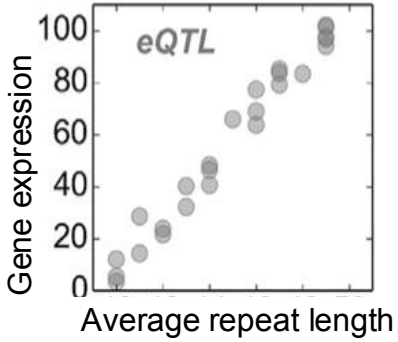
Rare expanded alleles in STRs have been shown to play a role in 60+ human diseases

Disorder	Affected gene	Repeat[a]	Repeat location	Normal repeat no.	Symptomatic repeat no.
Coding repeats					
DRPLA	*ATN1*	CAG	ORF	7–25	49–88
HD	*HTT*	CAG	ORF	6–34	36–180
SBMA	*AR*	CAG	ORF	11–24	40–62
SCA1	*ATXN1*	CAG	ORF	6–39	39–83
SCA2	*ATXN2*	CAG	ORF	15–29	34–59
SCA3	*ATXN3*	CAG	ORF	13–36	55–84
SCA6	*CACNA1A*	CAG	ORF	4–16	21–30
SCA7	*ATXN7*	CAG	ORF	4–35	34–>300
SCA17	*TBP*	CAG	ORF	25–44	45–66
Noncoding repeats					
DM1	*DMPK*	CTG	3′ UTR	5–37	>50–>2000
DM2	*CNBP*	CCTG	Intron 1	<27	75–11,000
EPM1	*CSTB*	(C),G(C),GCG	Promoter	2–3	30–75
FXS	*FMR1*	CGG	5′ UTR	6–52	~55–>2000
FRAXE MR	*AFF2/FMR3*	CCG	5′ end	6–25	>200
FRA12A MR	*DIP2B*	CGG	5′ UTR	6–23	?
FRDA	*FXN*	GAA	Intron 1	7–22	>66–>900
SCA10	*ATXN10*	ATTCT	Intron 9	10–29	280–4500
Coding and noncoding repeats					
SCA8	*ATXN8/ATXN8OS*	CAG and CTG	ORF and NCT	6–37	~107–250
Repeats with uncertain location					
HDL-2	*JPH3*	CAG/CTG	?	<50	>50
SCA12	*PPP2R2B*	CAG/CTG	?	<66	>66

Common STR length variation influences molecular phenotypes

Abundant contribution of short tandem repeats to gene expression variation in humans

Melissa Gymrek[1–4], Thomas Willems[1,4,5], Audrey Guilmatre[6,7], Haoyang Zeng[8], Barak Markus[1], Stoyan Georgiev[9], Mark J Daly[3,10], Alkes L Price[3,11,12], Jonathan K Pritchard[9,13], Andrew J Sharp[6] & Yaniv Erlich[1,4,14,15]

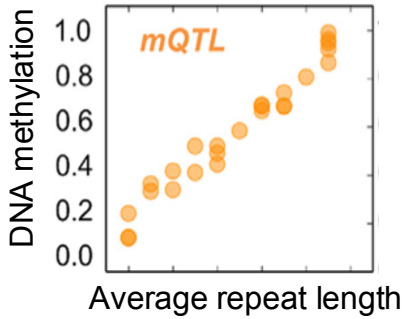


Research

Genome-wide evaluation of the effect of short tandem repeat variation on local DNA methylation

Alejandro Martin-Trujillo,[1] Paras Garg,[1] Nihir Patel,[1,2] Bharati Jadhav,[1] and Andrew J. Sharp[1]

[1]Department of Genetics and Genomic Sciences and Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, Hess Center for Science and Medicine, New York, New York 10029, USA



Why haven't the impacts of STRs on human traits been systematically investigated?

1. STR variation is not well tagged by SNPs, and as a result, they're ignored in standard genetic studies

2. Methods of Detection

Traditional Methods

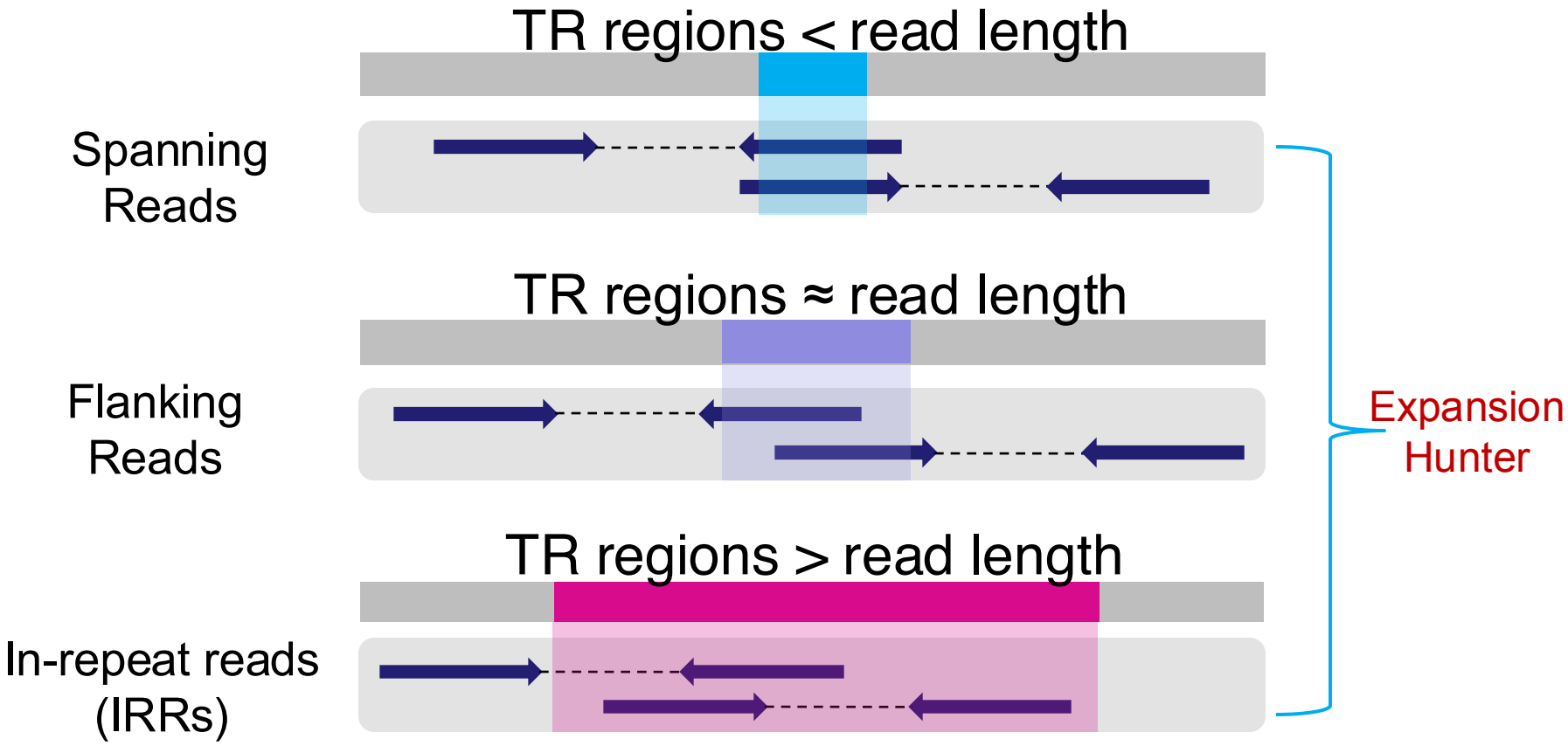
- PCR-based methods

Labor-intensive and low throughput!

Methods Using Whole Genome Sequencing

- Bioinformatic tools for genotyping

Genotyping STRs from whole genome sequencing data

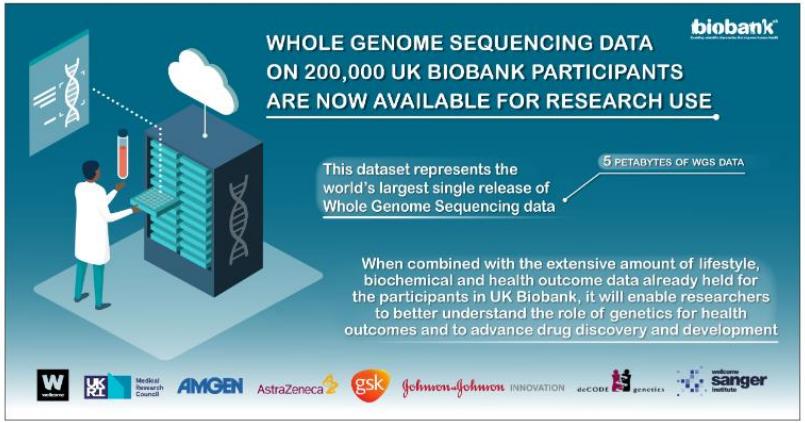


We hypothesize that common variation at STRs can modulate phenotypic variation in humans, which can be elucidated through Phenome-Wide Association Studies

Cohort
UK Biobank (UKB)

Whole Genome Sequencing data on 200,000 UK Biobank participants made available for research

November 17th 2021



biobank uk

168,544 unrelated Europeans with WGS data and tens of thousands of harmonized phenotypes

Cohort
168,544 unrelated Europeans in UKB with WGS and phenotype data

Genotype Data (STRs)
STR genotypes from WGS using ExpansionHunter

Catalog of ~36,000 STRs

- highly **polymorphic** or undergo rare expansions
- Enriched for overlap with genes and **regulatory elements**

Cohort
168,544 unrelated Europeans in UKB with WGS and phenotype data

Genotype Data (STRs)
STR genotypes from WGS using ExpansionHunter

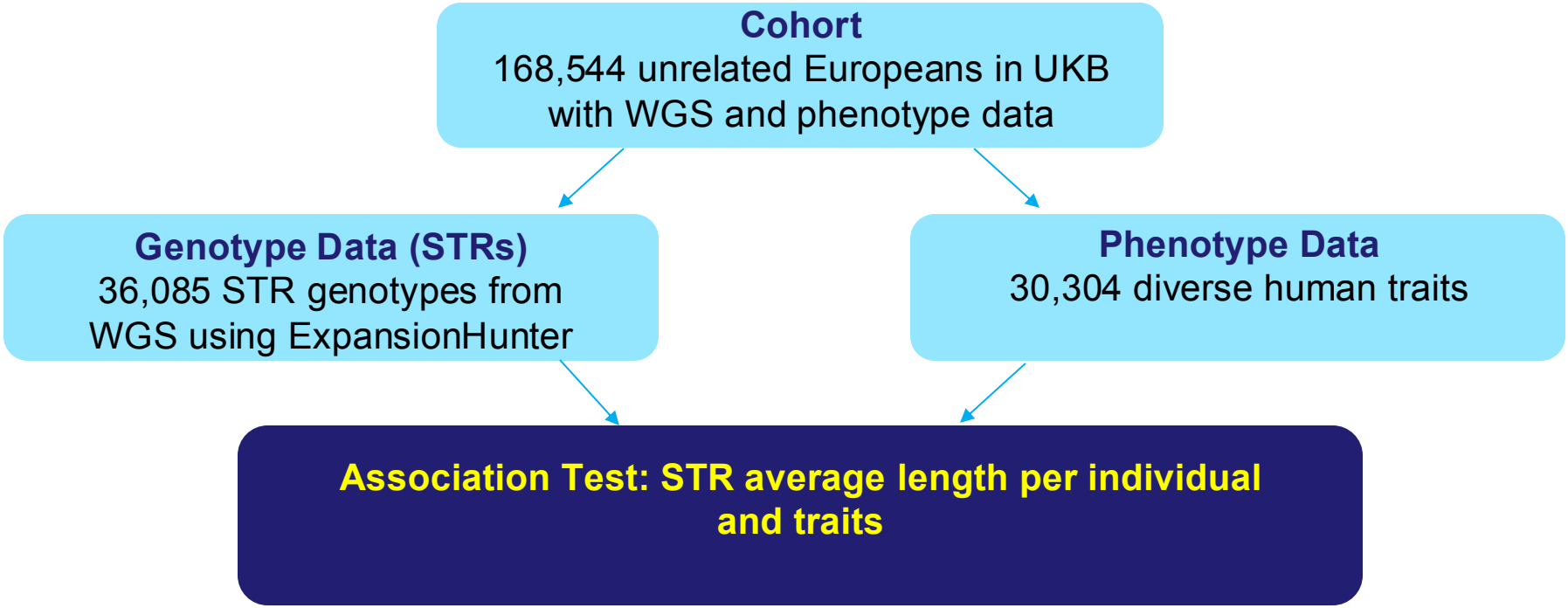
Catalog of ~36,000 STRs

- highly **polymorphic** or undergo rare expansions
- Enriched for overlap with genes and **regulatory elements**

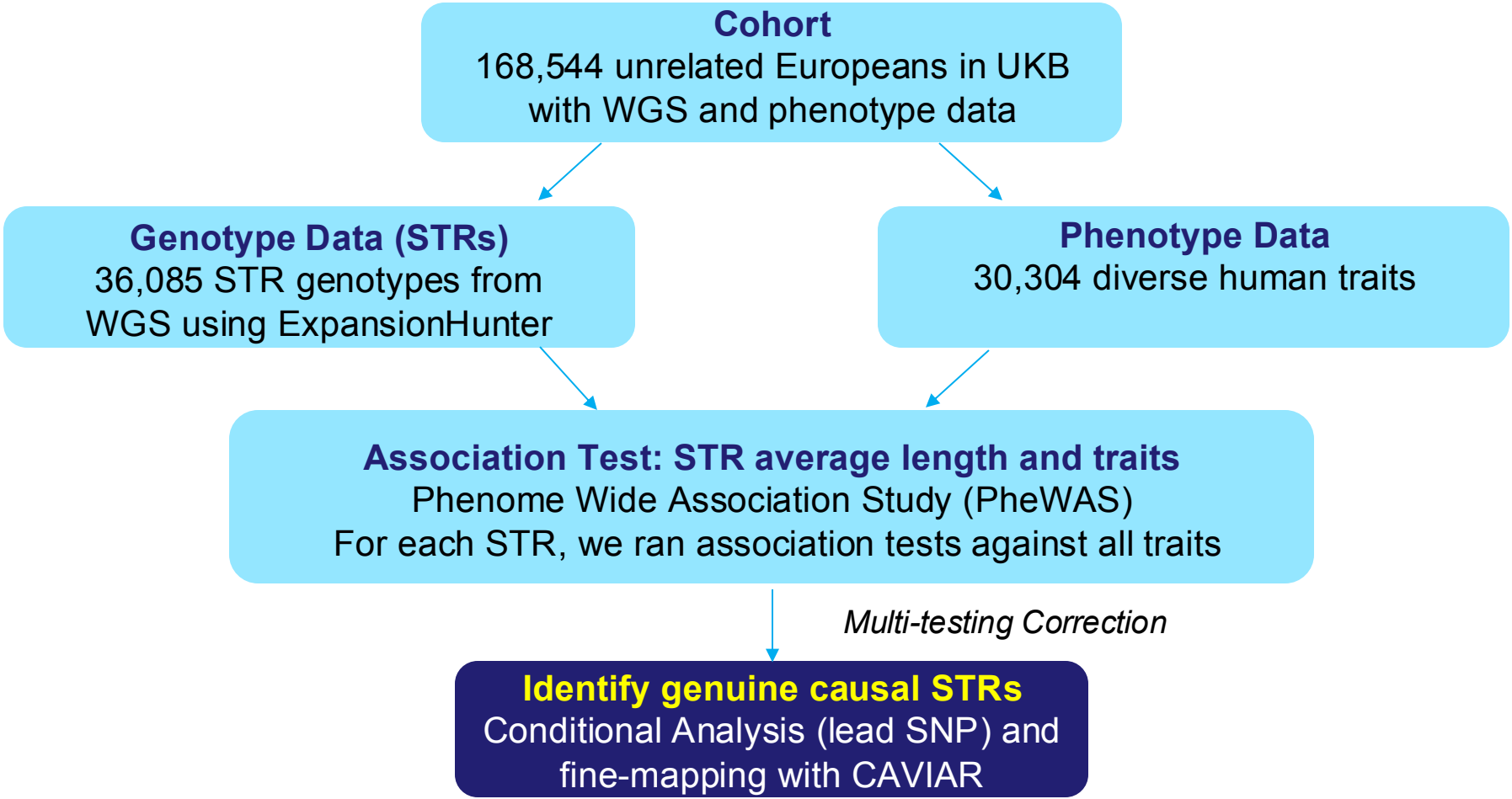
Phenotype Data
ICD codes, quantitative, categorical, and binary traits

30,304 diverse human traits

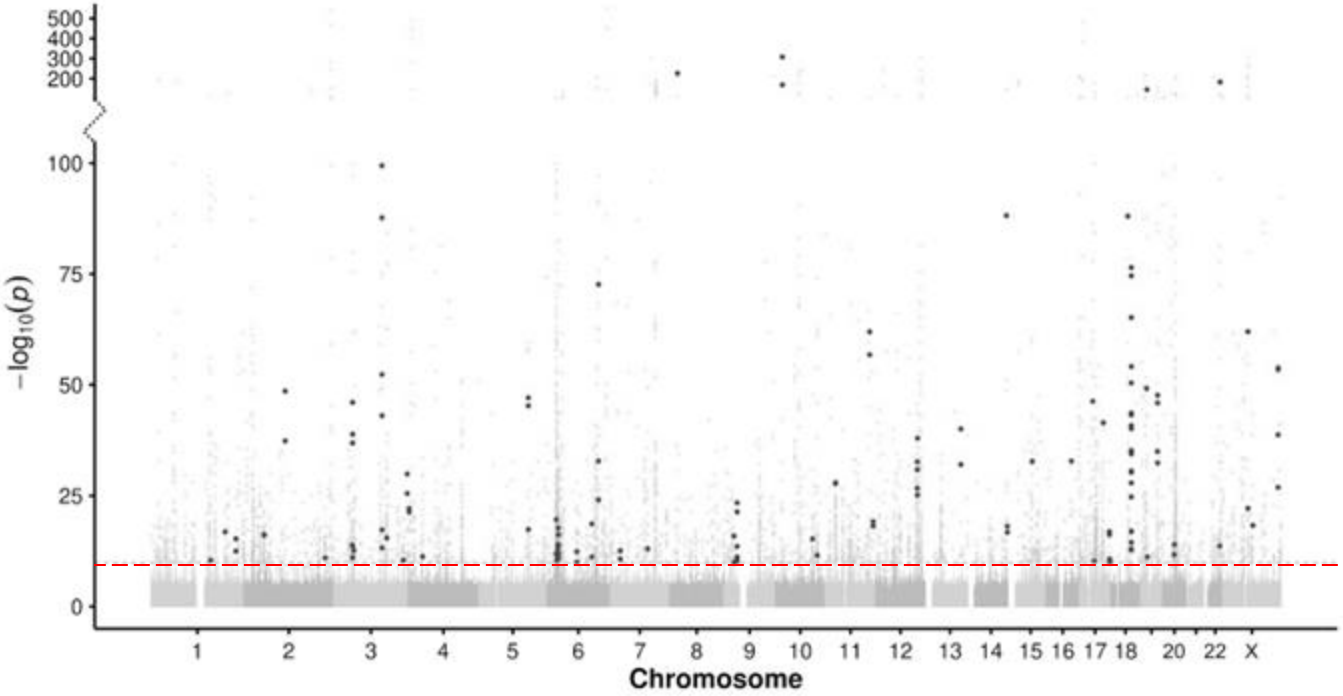
- 1,353 Quantitative Traits
- 28,951 Binary Traits



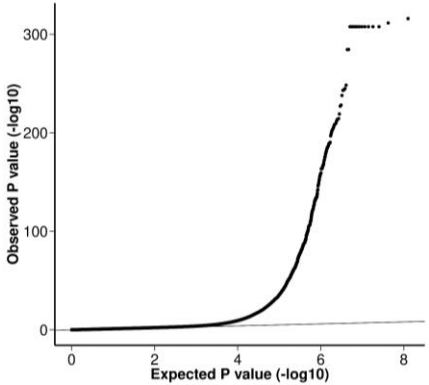
Phenome Wide Association Study (PheWAS)
For each STR, we ran association tests against all traits using REGENIE


Cohort
168,544 unrelated Europeans in UKB
with WGS and phenotype data
Genotype Data (STRs)
36,085 STR genotypes from
WGS using ExpansionHunter
Phenotype Data
30,304 diverse human traits
Association Test: STR average length and traits
Phenome Wide Association Study (PheWAS)
For each STR, we ran association tests against all traits
Multi-testing Correction
Identify genuine causal STRs
Conditional Analysis (lead SNP) and
fine-mapping with CAVIAR

Our PheWAS identified a total 5,378 associations involving 1,635 STRs and 461 traits



λ Genomic inflation = 1.061



Bonferroni $p < 1.45 \times 10^{-10}$

Pathogenic intronic CAG repeat within *TCF4* has been previously associated with corneal dystrophy

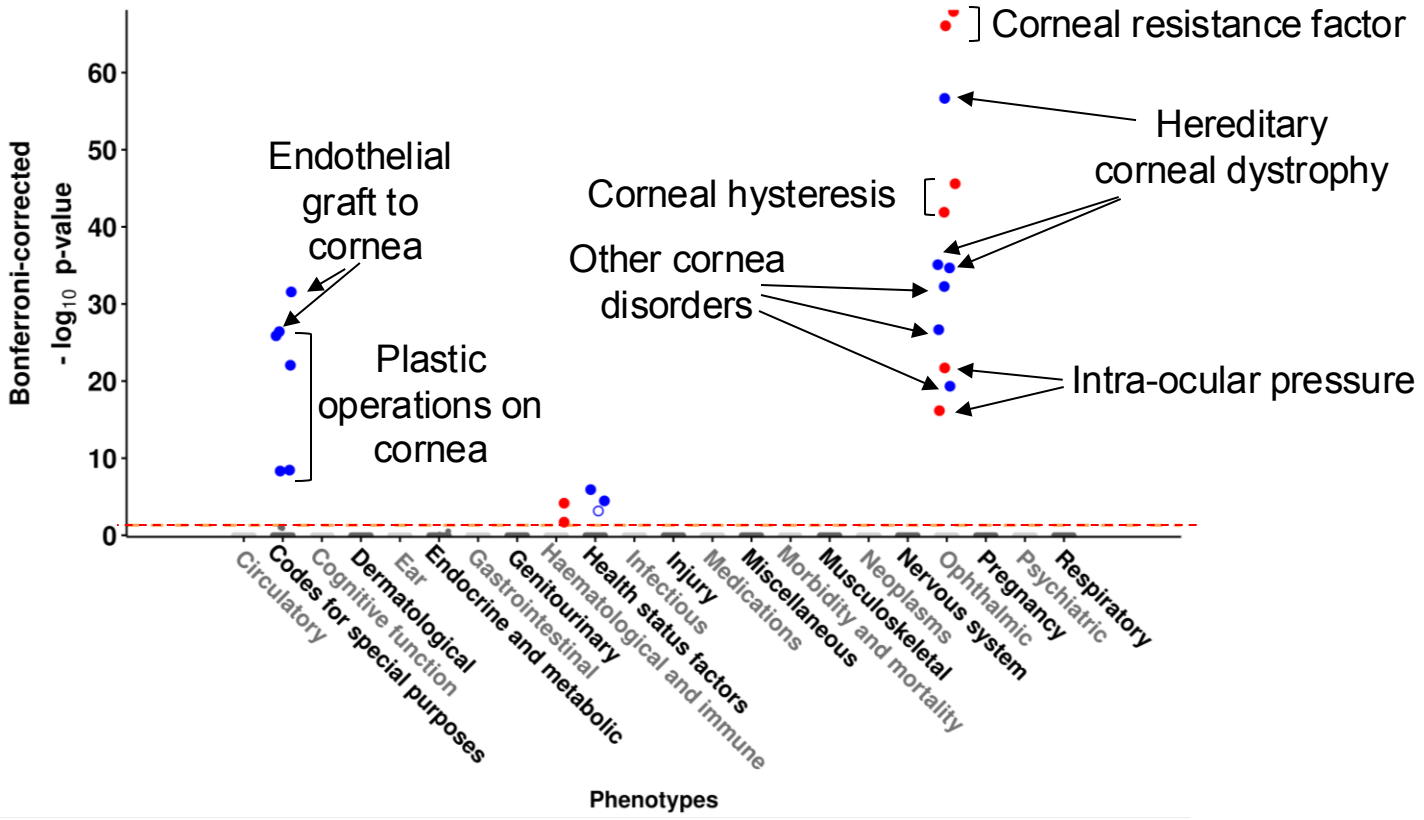
TCF4-mediated Fuchs endothelial corneal dystrophy: Insights into a common trinucleotide repeat-associated disease

Michael P. Fautsch [a,*,1], Eric D. Wieben [b,1], Keith H. Baratz [a,1], Nihar Bhattacharyya [c,1], Amanda N. Sadan [c,1], Nathaniel J. Hafford-Tear [c,1], Stephen J. Tuft [c,d,1], Alice E. Davidson [c,1]

[a] *Department of Ophthalmology, 200 1st St SW, Mayo Clinic, Rochester, MN, 55905, USA*
[b] *Department of Biochemistry and Molecular Biology, 200 1st St SW, Mayo Clinic, Rochester, MN, USA*
[c] *University College London Institute of Ophthalmology, London, ECIV 9EL, UK*
[d] *Moorfields Eye Hospital, London, EC1V 2PD, UK*

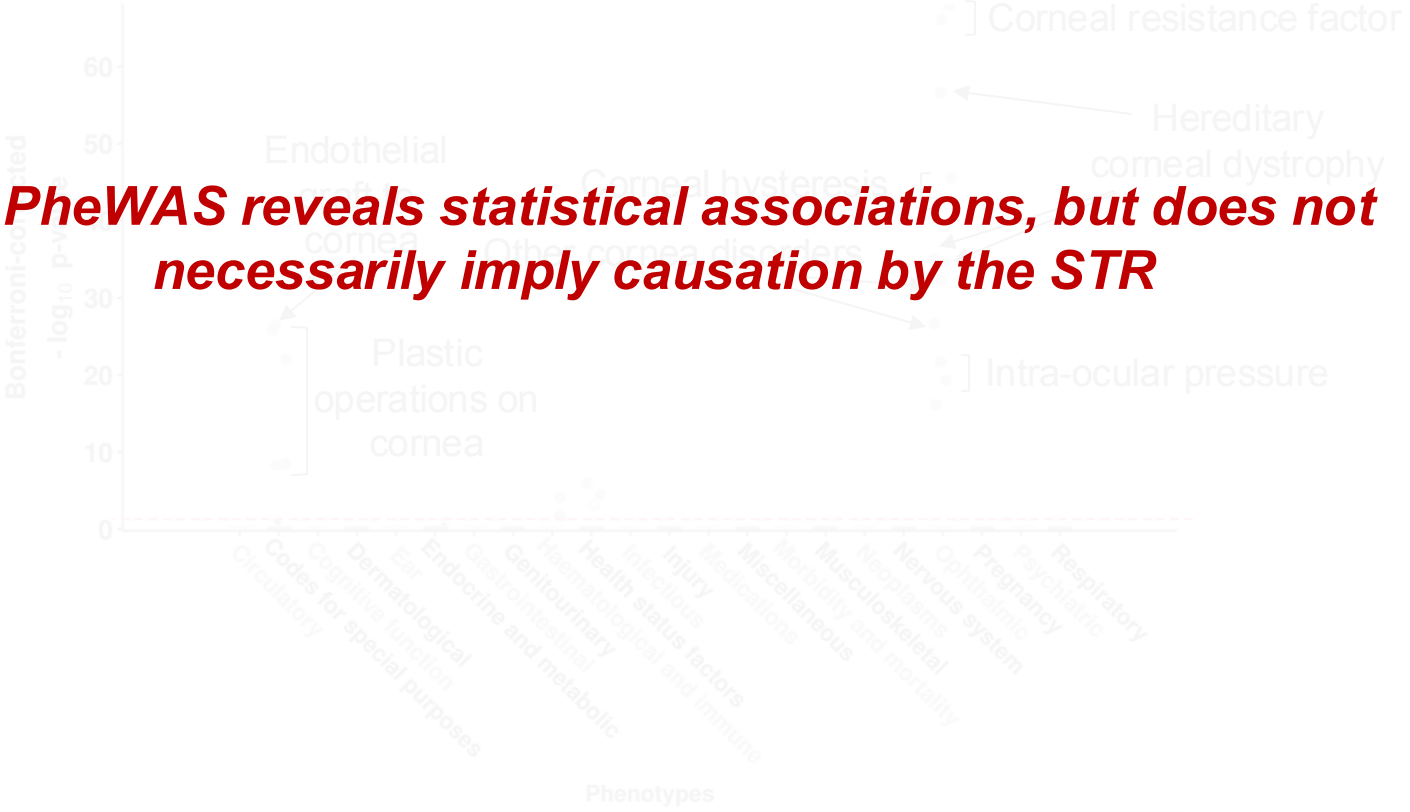
Association of intronic CAG repeat within *TCF4* with ocular-related phenotypes

We grouped traits into ~20 descriptive categories



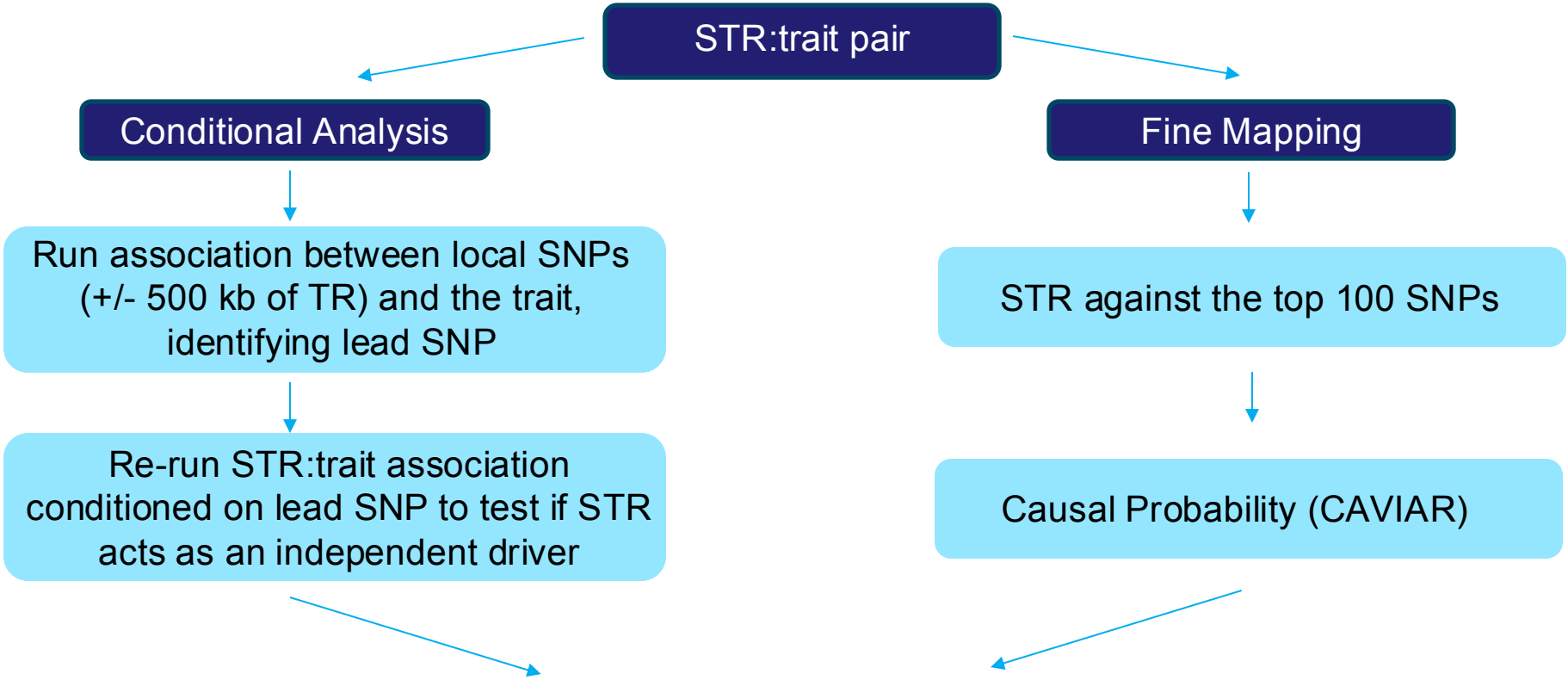
Association of intronic CAG repeat within *TCF4* with ocular-related phenotypes

We grouped traits into ~20 descriptive categories



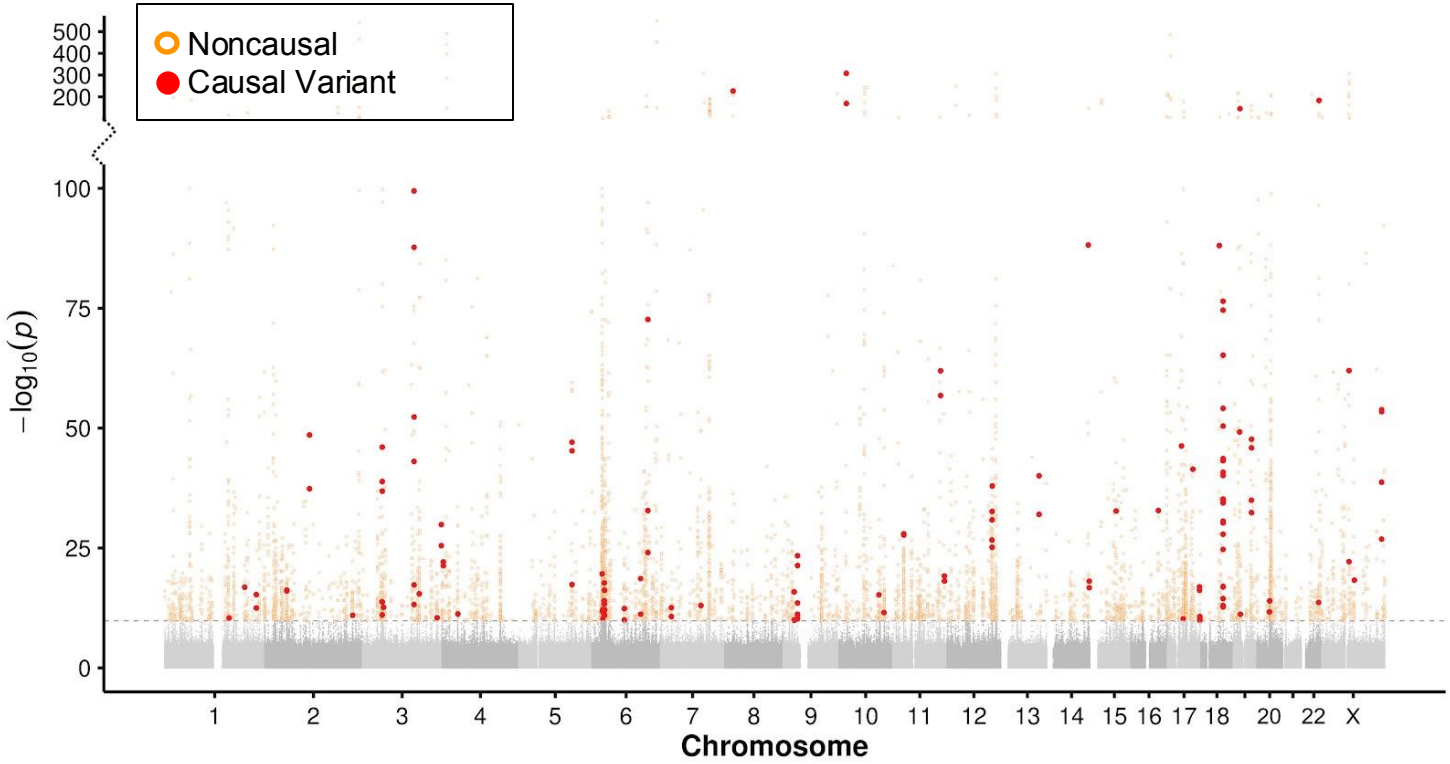
… PheWAS reveals statistical associations, but does not necessarily imply causation by the STR

Identification of causal STRs using two complementary approaches

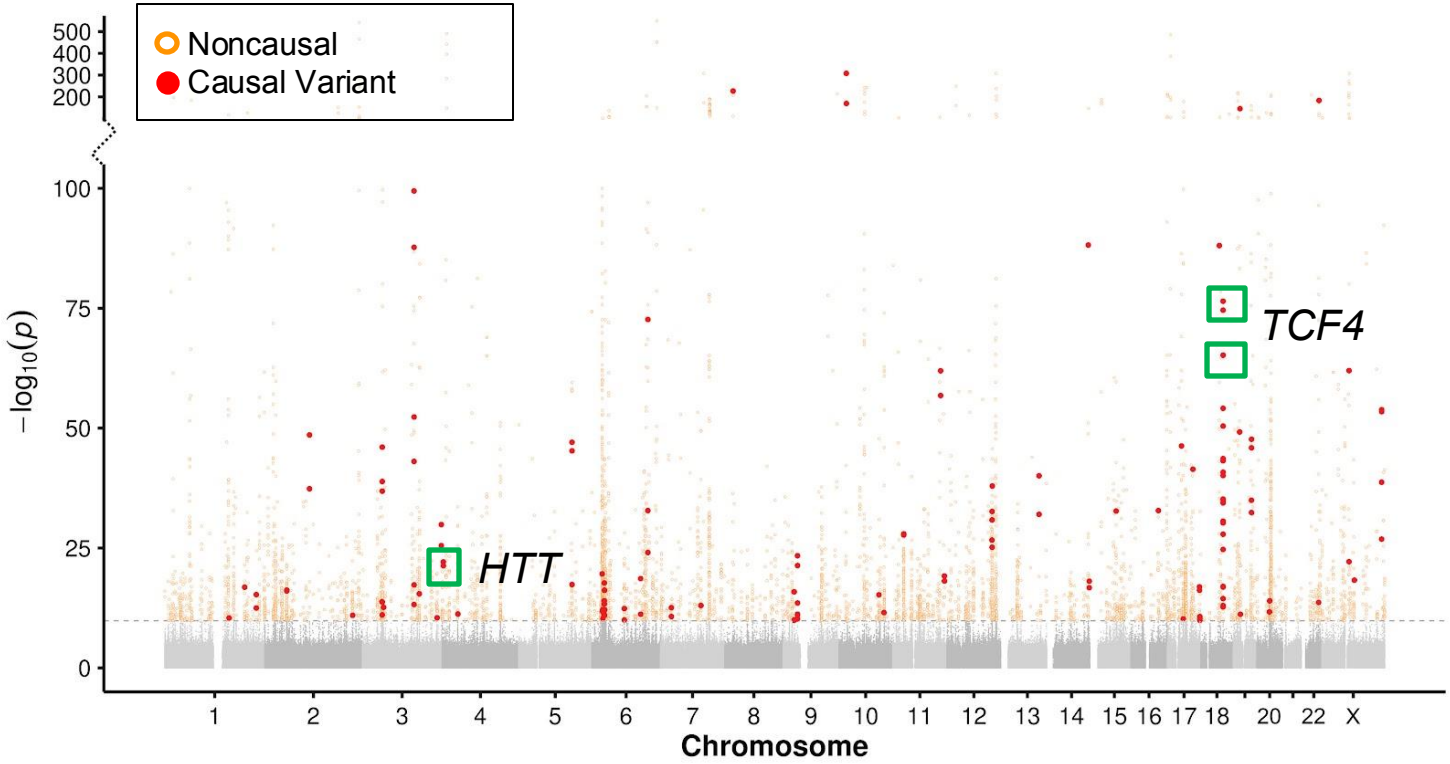


Casual STR: Conditional Analysis retains significance & show highest causal probability from CAVIAR

Our causal analysis identified a total of 47 causal STRs involving 101 associations

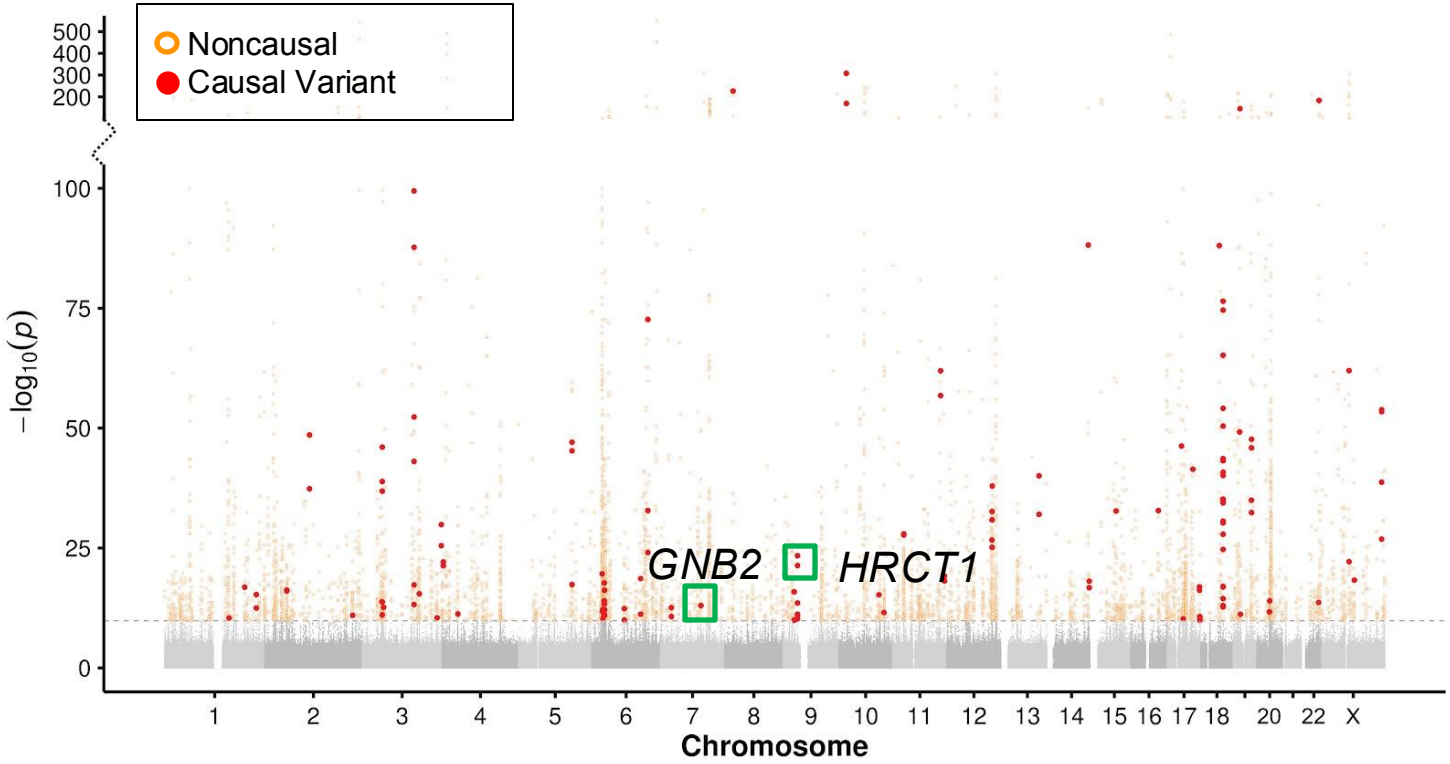


Our causal analysis identified a total of 47 causal STRs involving 101 associations



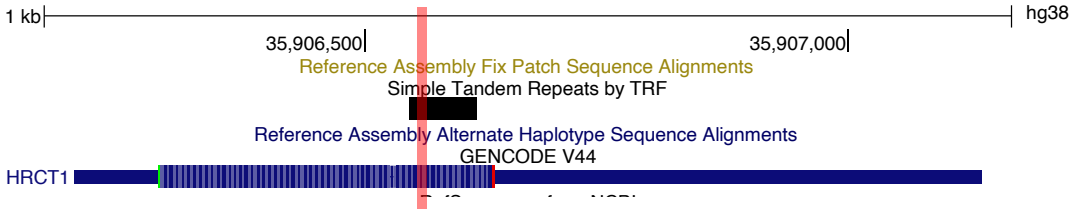
Known pathogenic repeats confirm that causal analysis is correctly identifying causal STRs

Our causal analysis identified a total of 47 causal STRs involving 101 associations

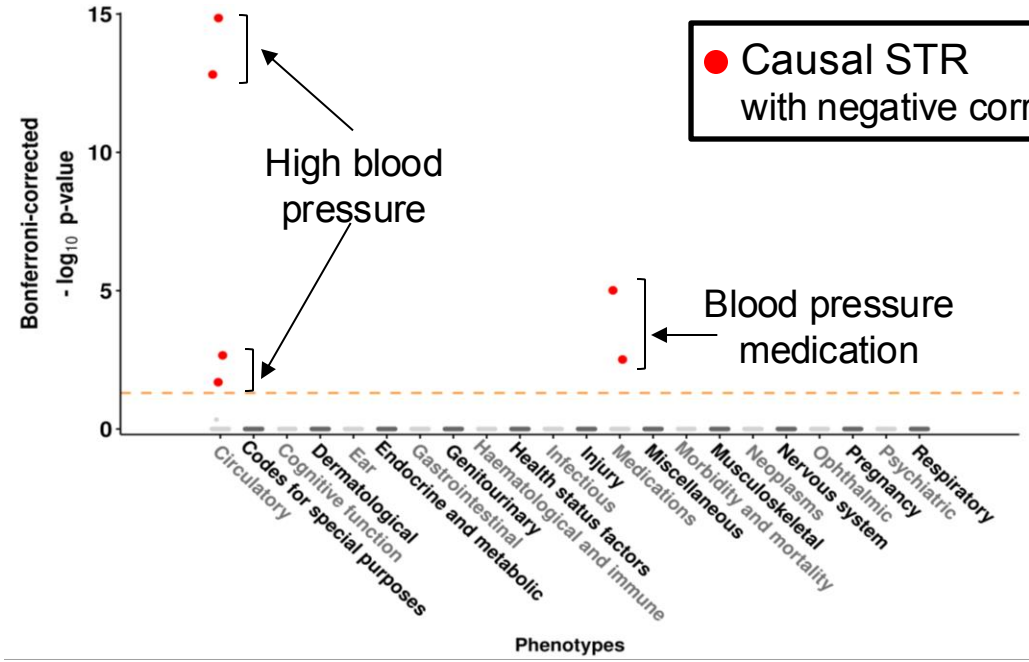


Novel Associations of Interest

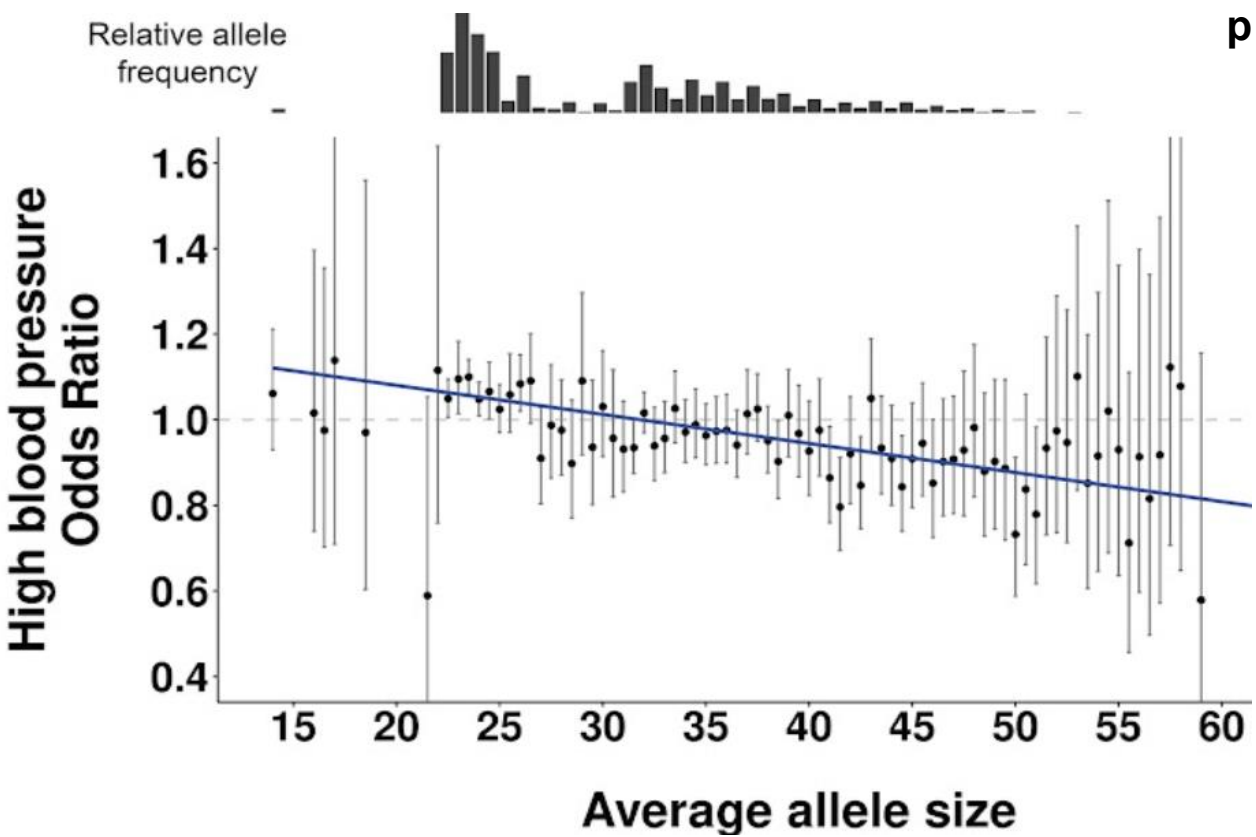
Novel association between exonic CCA repeat within *HRCT1* and high blood pressure



Gene function unknown

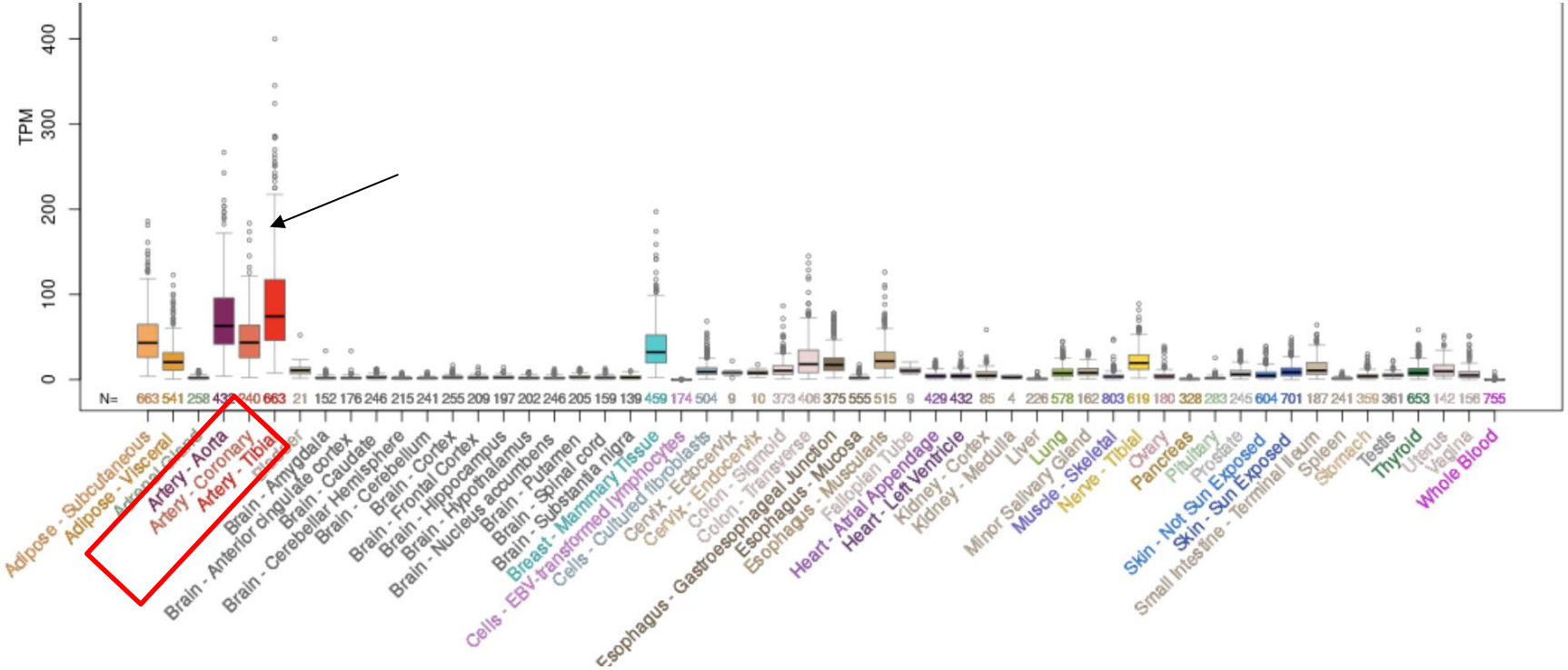


Novel association between exonic CCA repeat within *HRCT1* and high blood pressure

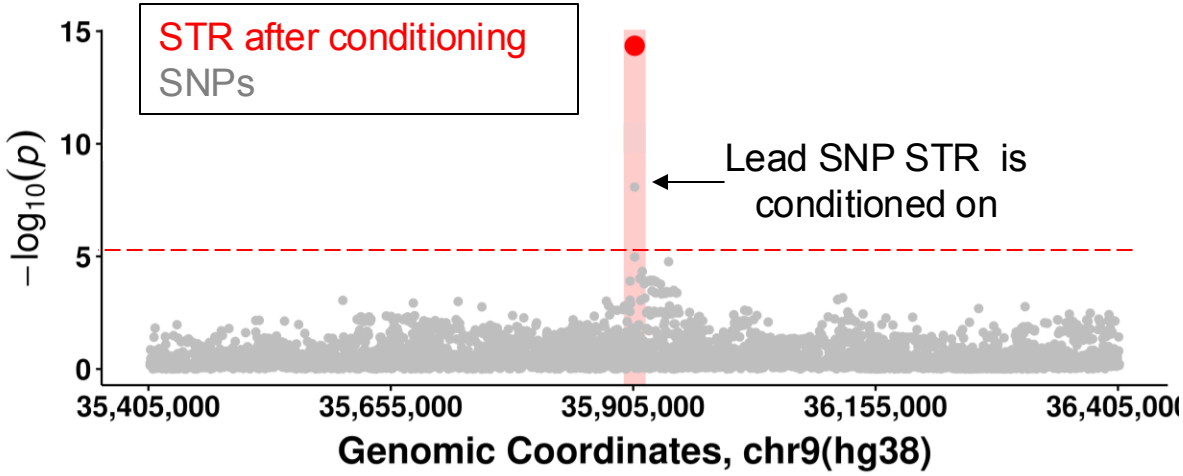


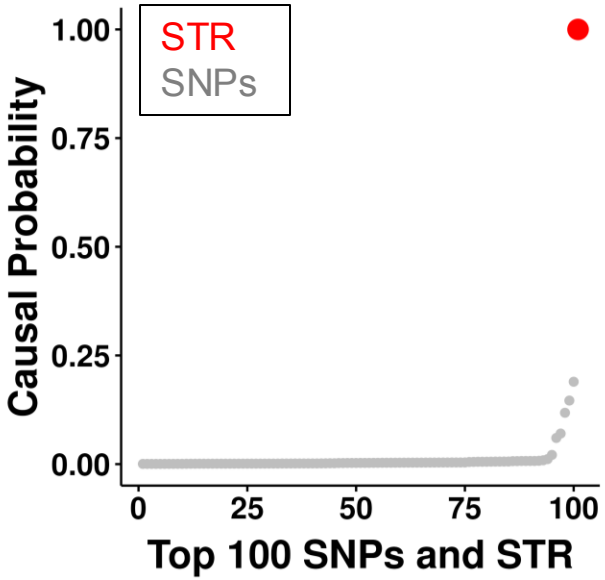
~20% decrease for the risk of high blood pressure in the longest average alleles

HRCT1 shows highest expression in artery tissues



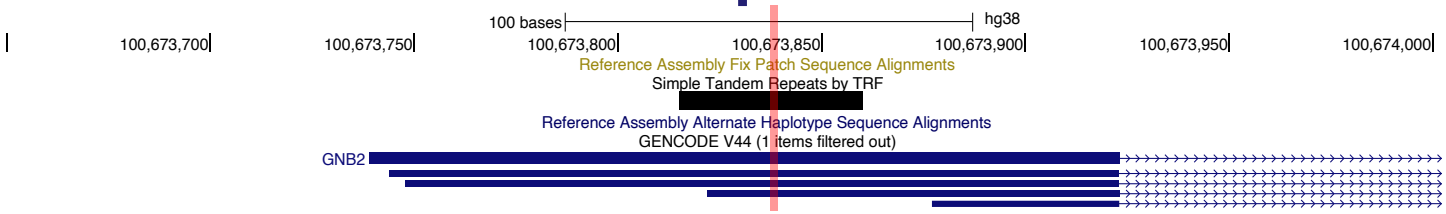
Conditional analysis and fine mapping results for exonic CCA repeat within *HRCT1* and high blood pressure

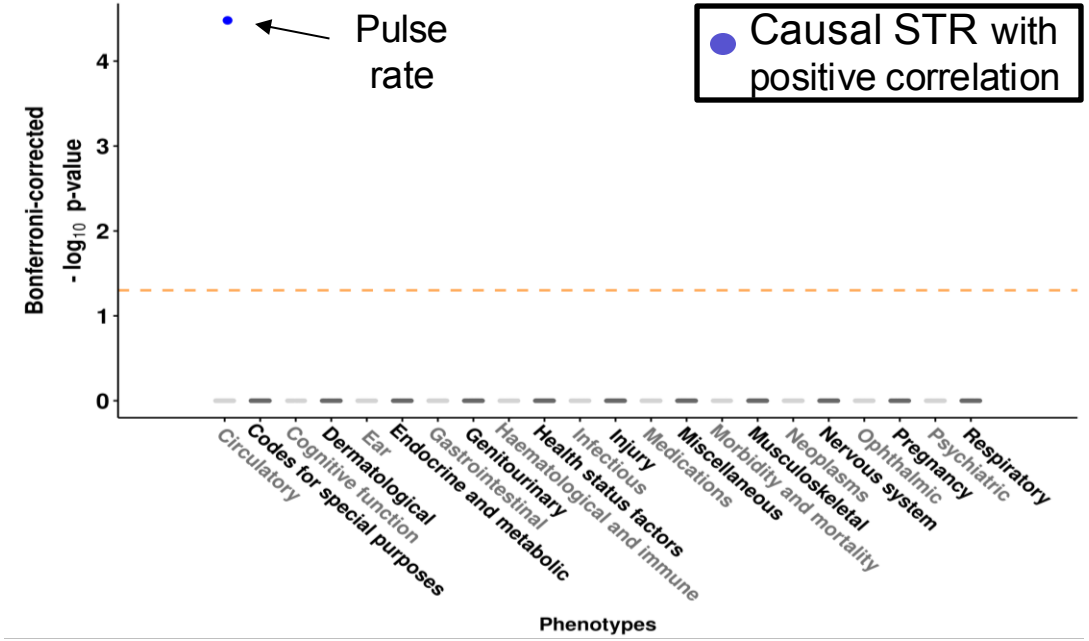




- ✓ HRCT1 repeat retains significance, indicating that the STR independently drives the association with high blood pressure
- ✓ HRCT1 repeat has the highest causal probability reported by CAVIAR

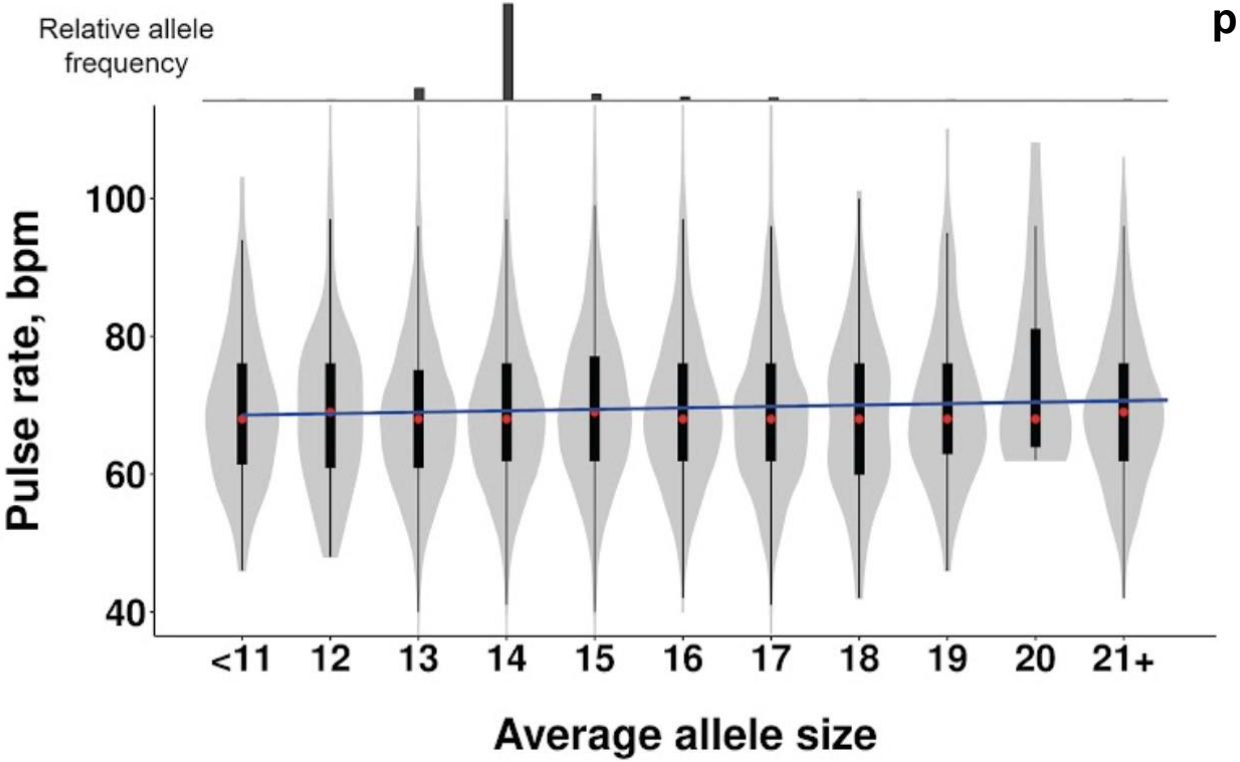
Novel association between CGC repeat in the promoter of *GNB2* and pulse rate





Novel association between CGC repeat in the promoter of *GNB2* and pulse rate

$p = 9.59 \times 10^{-14}$



Larger alleles have ~1bpm increase in pulse rate

Novel association between CGC repeat in the promoter of *GNB2* and pulse rate

Check

Molecular Medicine

A Mutation in the G-Protein Gene *GNB2* Causes Familial Sinus Node and Atrioventricular Conduction Dysfunction

Birgit Stallmeyer,* Johanna Kuß,* Stefan Kotthoff, Sven Zumhagen, Kirsty Vowinkel, Susanne Rinné, Lina A. Matschke, Corinna Friedrich, Ellen Schulze-Bahr, Stephan Rust, Guiscard Seebohm, Niels Decher, Eric Schulze-Bahr

Relative allele frequency

40

<11 12 13 14 15 16 17 18 19 20 21+

Average allele size

Larger alleles have ~1bpm increase in pulse rate

Replication in All of Us

Replication Cohort

- In All of Us, we used ~89,000 individuals from diverse backgrounds with WGS and trait information

Association Testing

- We matched causal traits identified in UK Biobank (UKB) with traits in All of Us, followed by targeted genotyping for causal STRs.
- We then conducted association tests between the average allele length of causal STRs per individual and the respective trait in All of Us.

After matching traits and filtering were able to replicate 74% of causal associations

Exonic CCA *HRCT1 repeat* and essential hypertension

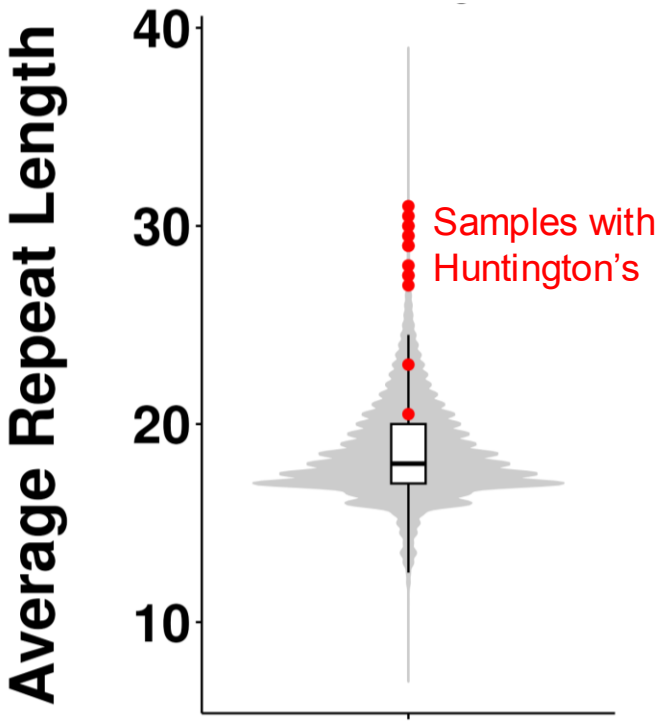
$p = 2.30 \times 10^{-3}$

GCC *GNB2* repeat and pulse rate

- $p = 3.88 \times 10^{-8}$

Novel trait associations are driven by common STR allele variation

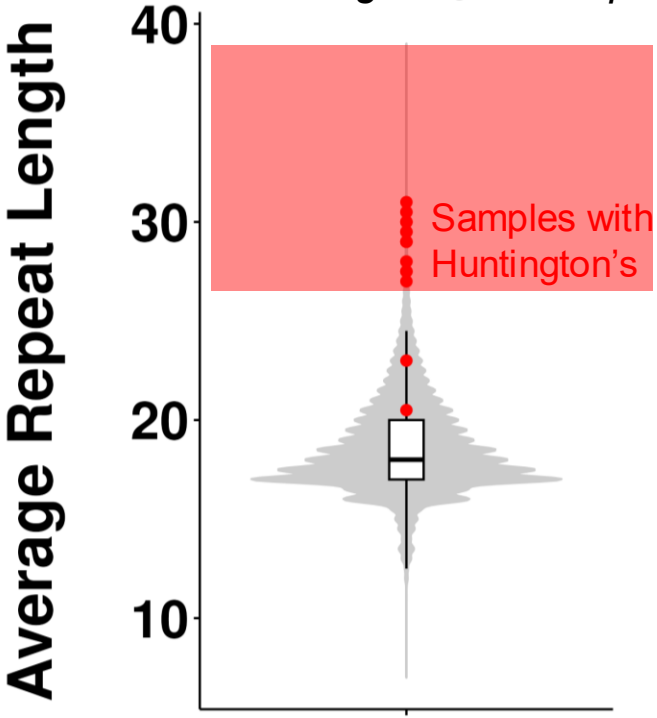
Rare expansions for the CAG repeat within HTT



Pathogenic STRs are driven by rare repeat expansions

Novel trait associations are driven by common STR allele variation

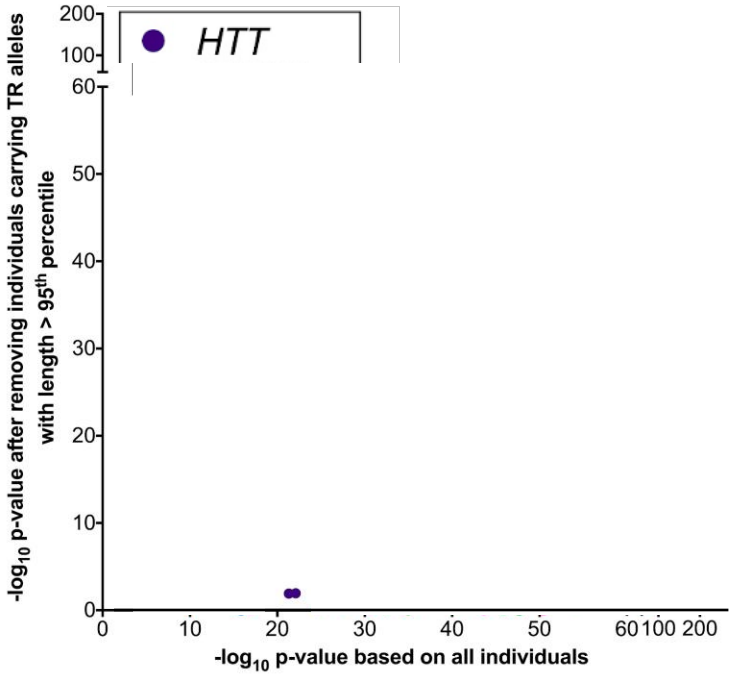
For each causal STR:trait pair, we repeated association analysis excluding samples with alleles longer than 95th percentile



Pathogenic STRs are driven by rare repeat expansions

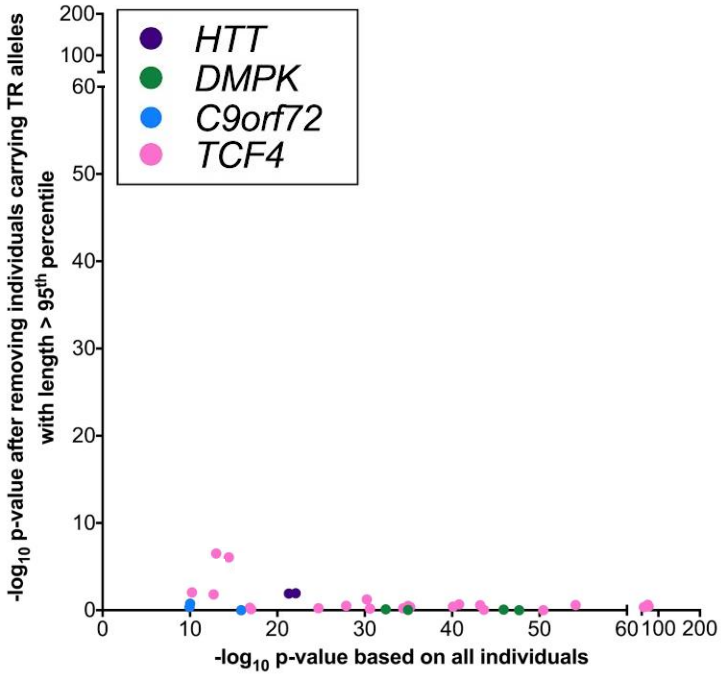
Novel trait associations are driven by common STR allele variation

For each causal STR:trait pair, we repeated association analysis excluding samples with alleles longer than 95th percentile



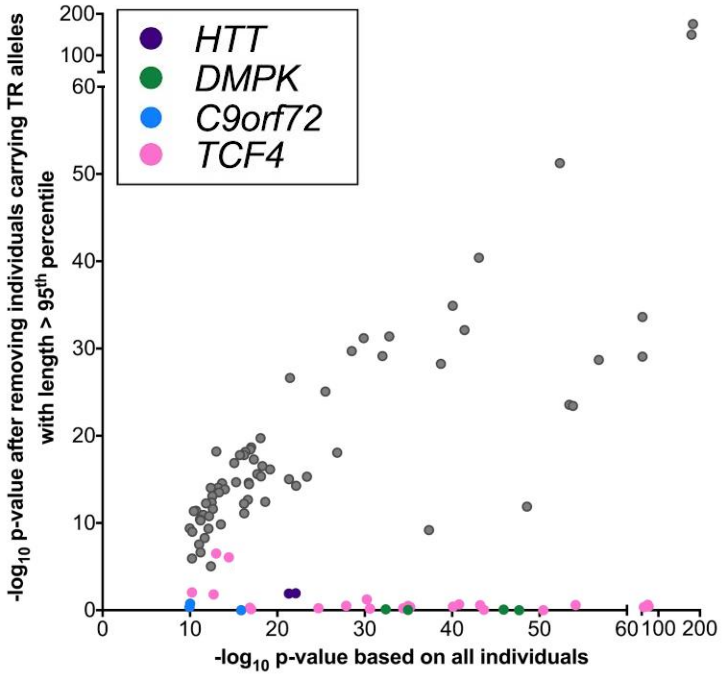
Novel trait associations are driven by common STR allele variation

For each causal STR:trait pair, we repeated association analysis excluding samples with alleles longer than 95th percentile



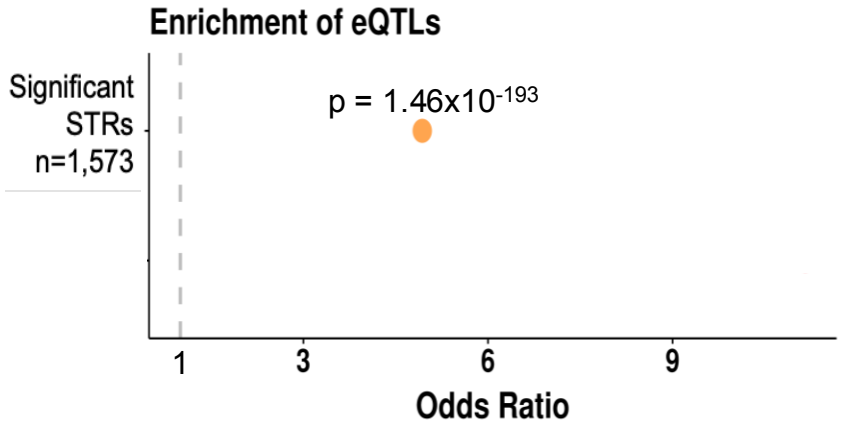
Novel trait associations are driven by common STR allele variation

For each causal STR:trait pair, we repeated association analysis excluding samples with alleles longer than 95th percentile



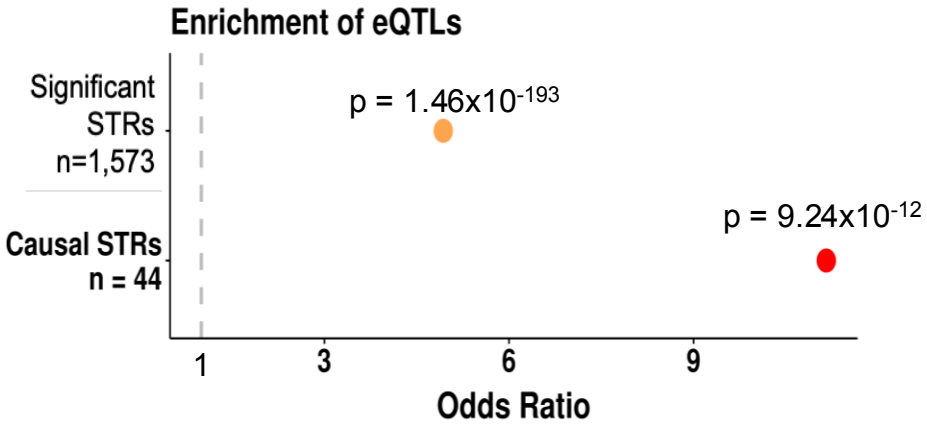
Causal TRs are enriched for STRs that impact expression levels of nearby genes

Using GTEx data, we performed eQTL analysis using RNA-seq data and STR genotypes to identify functional STRs



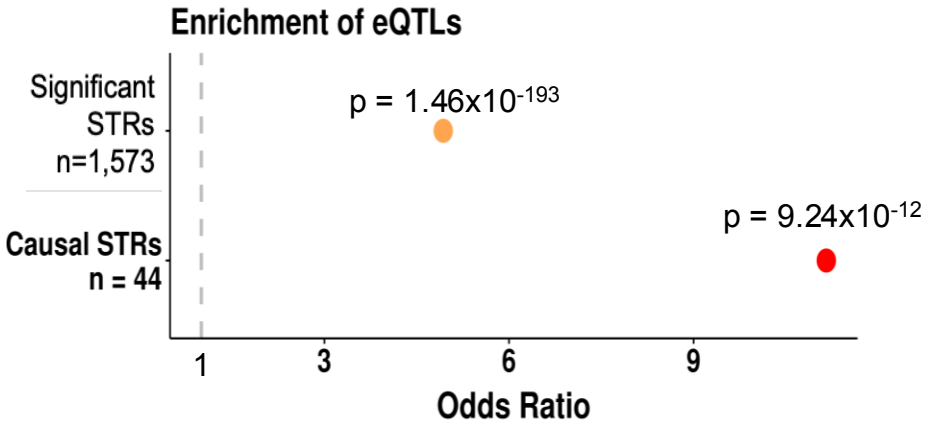
Causal TRs are enriched for STRs that impact expression levels of nearby genes

Using GTEx data, we performed eQTL analysis using RNA-seq data and STR genotypes to identify functional STRs

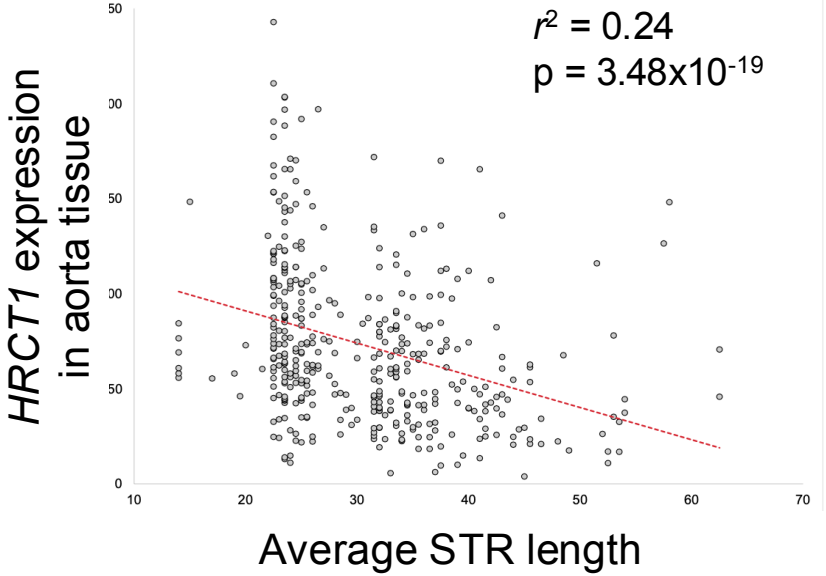


Causal TRs are enriched for STRs that impact expression levels of nearby genes

Using GTEx data, we performed eQTL analysis using RNA-seq data and STR genotypes to identify functional STRs



Exonic STR in *HRCT1* has strong negative correlation with expression level of *HRCT1* in aorta



Conclusions

1. Our PheWAS provided a comprehensive evaluation of the impact of STR variation on human traits

Conclusions

1. Our PheWAS provided a comprehensive evaluation of the impact of STR variation on human traits.

2. Causal analysis showed that a subset of STRs represent the causal variant responsible for the phenotypic variation.

Conclusions

1. Our PheWAS provided a comprehensive evaluation of the impact of STR variation on human traits

2. Causal analysis showed that a subset of STRs represent the causal variant responsible for the phenotypic variation.

3. Causal STRs are strongly enriched for loci involved in the expression levels of nearby genes, providing insights about the molecular mechanism by which STRs regulate the associated trait.

Conclusions

1. Our PheWAS provided a comprehensive evaluation of the impact of STR variation on human traits.

2. Causal analysis showed that a subset of STRs represent the causal variant responsible for the phenotypic variation.

3. Causal STRs are strongly enriched for loci involved in the expression levels of nearby genes, providing insights about the molecular mechanism by which STRs regulate the associated trait.

4. Our results highlight the role of multi-allelic variants as contributors to the “missing heritability” of the genome and the importance of incorporating variation at STRs in future genetic studies.