Dec 01, 2021

Chromatin loops and expression QTL colocalization reveal novel gene targets for T1D-associated GWAS variants in immune cells

DOI
https://dx.doi.org/10.17504/protocols.io.n2bvjx5qnlk5/v1

Joaquin Reyna[1,2], Sourya Bhattacharyya[1], Nikhil Rao[1,3], Abhijit Chakraborty[1], Ferhat Ay[1,4]

[1]La Jolla Institute for Immunology, La Jolla, CA, USA.;
[2]Bioinformatics and Systems Biology Graduate Program, University of California, San Diego, La Jolla, CA.;
[3]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA.;
[4]Department of Pediatrics, University of California San Diego, La Jolla, CA, USA.

Sugar Science

Joaquin Reyna

OPEN ACCESS

DOI: https://dx.doi.org/10.17504/protocols.io.n2bvjx5qnlk5/v1

External link: https://github.com/joreynajr/dchallenge

Protocol Citation: Joaquin Reyna, Sourya Bhattacharyya, Nikhil Rao, Abhijit Chakraborty, Ferhat Ay 2021. Chromatin loops and expression QTL colocalization reveal novel gene targets for T1D-associated GWAS variants in immune cells. **protocols.io** **https://dx.doi.org/10.17504/protocols.io.n2bvjx5qnlk5/v1**

Protocol status: In development
We are still developing and optimizing this protocol

Created: November 10, 2021

Last Modified: December 01, 2021

Protocol Integer ID: 55039

Keywords: Type 1 Diabetes, HiChIP, HiC, Colocalization, eQTL, GWAS, gwas variants in immune cells type, novel gene targets for t1d, genes such asbach2, t1d gwas signal, associated gwas variant, immune cells type, immune regulation, role of immune regulation, gene expression quantitative trait loci, different immune cell population, diabetes, immune cell population, variants with chromatin loop, multiple important transcription factor, subset of these immune cell population, eqtl association for this snp, sites for multiple important transcription factor, potential target gene, novel gene target, gene expression, transcription factor, immune cell, destroying insulin, gene, chromatin loop, insulin, colocalized snp, pancrea, disease association, rna, studied gene, promoter of another gene, same snp

Disclaimer

This protocol was made as part of the DChallenge and gives a high level explanation of our steps.

Abstract

Type 1 diabetes (T1D) is a disease characterized by the destruction of β cell populations in the pancreas. Immune cells, and specifically T cells, have been implicated to play a key role in destroying insulin producing β cells by infiltrating the pancreas. To better understand the role of immune regulation in T1D, we colocalized the gene expression quantitative trait loci (eQTL) signals from 18 different immune cell populations (15 from DICE, 3 from BLUEPRINT) with T1D GWAS signals to gather non-coding variants that are likely causal for both the gene expression and the disease association. We further overlapped these variants with chromatin loops mapped in a subset of these immune cell populations to identify potential target genes of the significant non-coding SNPs. Aside from well-studied genes such as*BACH2, UBASH3A, PTPN22* and *SIRPG,* we identified *AP003774.1*, a long non-coding RNA, that is looping to a ~15kb away regulatory element overlapping a colocalized SNP (rs479777) in various T cell subsets. The looped region overlaps the promoter of another gene (promoter-promoter loop), *CCDC88B*, however, the eQTL association for this SNP is specific to *AP003774.1*and is remarkably strong for resting T cell subsets, NK cells and naïve B cells. The same SNP creates strong binding sites for multiple important transcription factors in donors with the non-reference allele leading to higher expression of *AP003774.1.* We hypothesize that the overexpression of *AP003774.1*lncRNA mediated through specific non-coding variants in different immune cell populations play a role in immune-related aspects of T1D.

Image Attribution

https://thenounproject.com/term/weaving/

Main Colocalization Pipeline

1 **Download the GWAS summary statistics (uses GRCh38 coordinates)**
Download GWAS summary statistics from the GWAS catalogue.
Download link:
http://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/GCST90014001-GCST90015000/GCST90014023/GCST90014023_buildGRCh38.tsv

Citation

Chiou, J., Geusz, R.J., Okino, ML. et al (2021)
. Interpreting type 1 diabetes risk with genetics and single-cell epigenomics. Nature.

https://doi.org/10.1038/s41586-021-03552-w LINK

1.1 **Remap the GWAS SNP coordinate to GRCh37**

Command

liftover

```
liftOver -bedPlus=3 -tab <in bed> <chain file> <out lift file>
<unmapped>
```

1.2 **Filter for SNPs with genome-wide significance (5e-8)**

1.3 **Reformat the file into the required format for colocalization**
Estimate standard error of the SNP using beta and MAF values. Standard error is beta/z-score

2 **Download the eQTL summary statistics (uses GRCh37 coordinates)**

Many eQTL studies have already been completed with summary statistics. For our project we downloaded preprocessed data from the Mu et al., 2021 which contains Blueprint and DICE eQTL results.
Download link: https://zenodo.org/record/4480206#.YXcophrMJyw

Citation

Mu, Z., Wei, W., Fair, B. et al (2021)
. The impact of cell type and context-dependent regulatory variants on human immune traits.
Genome Biol.

https://doi.org/10.1186/s13059-021-02334-x LINK

3 **Run colocalization between the T1D GWAS and all eQTL studies**
The coloc package that we used can be referenced here: https://cran.r-project.org/web/packages/coloc/index.html

Command

Colocalization scripts generated by Sourya Bhattacharyya.

Colocalization

```
Rscript Colocalization_Analysis_GWAS.R <in GWAS> <output directory>
<in eQTL>
```

4 **Intersect colocalized SNPs with HiChIP data**
Obtain HiChIP-seq data for several cell types. In our case, we obtained data from CD4+ T-cells, CD8+ T-cells, classical monocytes, non-classical monocytes, naive B cells, natural killer, T follicular helper, Th1, Th17, Th2 and TH1/17, TREGMEM, and TREGNAIVE cells.

5 **Generate a Master table of SNP-gene pairs with loops**

For this master table we focused on colocalized SNPs, their colocalized eGenes and other genes as well.

5.1 **Extract all SNP-gene pairs that are +/- 500kb from a colocalized SNP**

Command

Intersect two lists of genetic loci.

bedtools intersect

```
bedtools intersect <bed1> <bed2> > <output bed>
```

5.2 **Intersect loops with SNP-gene pairs**

Command

Intersect two lists of genetic loci pairs.

bedtools pairtopair

```
bedtools pairtopair <bedpe1> <bedpe2> > <output>
```

5.3 **Label each SNP-Gene pair**

For each SNP-Gene pair label whether it is an eQTL, colocalized pair, contains a loop and all metadata.

Gene Candidate Prioritization

6 **Investigate Candidate Genes using the WashU Epigenome Browser**

6.1 **Generate BED + Index files for 1D tracks**

- single-cell ATAC-seq data
- SNP and gene locations

Command

Compress bed file with bgzip.

bgzip

```
bgzip <input> <output>
```

Command

Create an index for compressed bed file.

tabix

```
tabix <bed.gz>
```

6.2 **Generate BEDPE + Index files for 2D tracks**

- loop data for all cell lines

Command

bgzip

```
bgzip <input> <output>
```

Command

tabix

```
tabix <bed.gz>
```

6.3 **Add ChromHMM tracks from the Public Hubs**

6.4 **Look for genes with several SNPs which overlap important loops**

7 **Investigate TF binding sites using the Genome Browser + others**

7.1 **Load JASPAR tracks, query SNP locations, investigate motifs**
- JASPAR tracks can be added from the Public Hub
- **https://jaspar.genereg.net/**

7.2 **Query SNPs using ADASTRA website**
https://adastra.autosome.ru/zanthar

8 **Confirm expression of candidate genes using the DICE database**
https://dice-database.org/

Citations

Step 1

Chiou, J., Geusz, R.J., Okino, ML. et al. Interpreting type 1 diabetes risk with genetics and single-cell epigenomics
https://doi.org/10.1038/s41586-021-03552-w

Step 2

Mu, Z., Wei, W., Fair, B. et al. The impact of cell type and context-dependent regulatory variants on human immune traits
https://doi.org/10.1186/s13059-021-02334-x