Prioritization of putative target genes underpinning COVID-19 host GWAS traits based on high-resolution 3D chromosomal topology

Document version 2 based on COVID-19 host genetics GWAS release 4

Michiel J. Thiecke[1], Emma Yang[2,3], Helen Ray-Jones[2,3], Oliver S Burren[4,5], and Mikhail Spivakov[2,3]

[1]Enhanc3D Genomics Ltd, Cambridge, UK
[2]MRC London Institute of Medical Sciences, London, UK
[3]Institute of Clinical Sciences, Faculty of Medicine, Imperial College, London, UK
[4]Cambridge Institute of Therapeutic Immunology and Infectious Disease, Cambridge, UK
[5]Department of Medicine, School of Clinical Medicine, University of Cambridge, UK

Contacts: mthiecke@enhanc3dgenomics.com, mikhail.spivakov@lms.mrc.ac.uk

Summary. GWAS variants commonly map to DNA regulatory regions, many of which are located away from their target genes, coming into their proximity through 3D chromosomal interactions. We previously generated high-resolution Capture Hi-C data on the chromosomal contacts involving all annotated gene promoters in 17 human primary blood cell types (including endothelial precursors) and developed COGS, a statistical pipeline for GWAS gene prioritisation based on these data (Javierre et al., 2016). Applying COGS to COVID-19 host GWAS data using the same panel of cell types, we prioritise multiple putative associated genes such as those known to be involved in immune function (including *ETS1*, *IFNAR1/2, OAS3, CCR1* and others) and lung biology (such as *DPP9* and *FOXP4*). Full results are listed in the attached data table. These data, used in conjunction with other prioritisation approaches, will aid in the understanding of COVID-19 pathology, paving the way for novel treatments.

Methods. The COGS pipeline (Burren et al., 2017; Javierre et al., 2016) takes GWAS summary data as input, fine-maps it using Wakefield synthesis (Wakefield, 2009) and aggregates the resulting posterior probabilities of a variant being casual across all promoter-interacting regions detected using Promoter Capture Hi-C data. We have run the COGS pipeline using each of the seven COVID-19 host GWAS datasets (GRCh37-based release 4, excluding 23andMe data) and 3D Promoter Capture Hi-C data in 17 human primary blood cell types (Javierre et al., 2016). Promoter interactions exceeding the CHiCAGO interaction score of 5 (Cairns et al., 2016) in at least one cell type were supplied to COGS.

Description. We release a data table containing COGS prioritisation scores (corresponding to gene-level posterior probabilities of association) for ~42,000 genes (including both protein-coding and non-protein coding) (Table S1). We further summarised results for each of the seven GWAS in the form of gene-level Manhattan plots (Figures S1-S7). The numbers of genes showing high and medium prioritisation scores per GWAS are listed in Table S2. Example genomic profiles of SNP-level posterior probabilities alongside Promoter Capture Hi-C-detected chromosomal interactions and H3K27ac profiles in potentially causal cell types are shown in Figures S7-S10 for four prioritised genes, *OAS3, IFNAR1, ETS1,* and *CCR1*.

List of Supplementary Tables and Figures

Table S1 - COGS prioritisation scores for each of the six COVID-19 host GWAS based on Promoter Capture Hi-C data from 17 human primary blood cell types (including endothelial precursors)
Table S2 - The numbers of genes showing high (score>0.7) and medium (0.3<score<=0.75) COGS prioritisation scores per GWAS
Figures S1-S6 - Gene-level Manhattan plots summarising COGS prioritisation results for each of the seven GWAS
Figures S7-S10 - Example profiles of SNP-level posterior probabilities, promoter interactions and H3K27ac signals in potentially causal cell types for four prioritised genes *OAS3*, *IFNAR1, ETS1,* and *CCR1*.

Conflict of interest statement

M.J.T is an employee, and M.S. is a cofounder, of Enhanc3D Genomics Ltd.

References

Burren OS, Rubio García A, Javierre B-M, Rainbow DB, Cairns J, Cooper NJ, Lambourne JJ, Schofield E, Castro Dopico X, Ferreira RC, Coulson R, Burden F, Rowlston SP, Downes K, Wingett SW, Frontini M, Ouwehand WH, Fraser P, Spivakov M, Todd JA, Wicker LS, Cutler AJ, Wallace C. 2017. Chromosome contacts in activated T cells identify autoimmune disease candidate genes. *Genome Biol* **18**:165.

Cairns J, Freire-Pritchett P, Wingett SW, Várnai C, Dimond A, Plagnol V, Zerbino D, Schoenfelder S, Javierre B-M, Osborne C, Fraser P, Spivakov M. 2016. CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biol* **17**:127.

Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, Cairns J, Wingett SW, Várnai C, Thiecke MJ, Burden F, Farrow S, Cutler AJ, Rehnström K, Downes K, Grassi L, Kostadima M, Freire-Pritchett P, Wang F, BLUEPRINT Consortium, Stunnenberg HG, Todd JA, Zerbino DR, Stegle O, Ouwehand WH, Frontini M, Wallace C, Spivakov M, Fraser P. 2016. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell* **167**:1369–1384.e19.

Wakefield J. 2009. Bayes factors for genome-wide association studies: comparison with P-values. *Genet Epidemiol* **33**:79–86.