Bioinformatics, 37(18), 2021, 3004–3007
doi: 10.1093/bioinformatics/btab122
Advance Access Publication Date: 24 February 2021
Applications Note

Genome analysis

PICS2: next-generation fine mapping via probabilistic identification of causal SNPs

Kimberly E. Taylor[1,*], K. Mark Ansel[2,3], Alexander Marson[1,2,4], Lindsey A. Criswell[1] and Kyle Kai-How Farh[5]

[1]Russell/Engleman Rheumatology Research Center, Department of Medicine, University of California San Francisco, CA, USA [2]Department of Microbiology and Immunology, University of California, San Francisco, CA, USA, [3]Sandler Asthma Basic Research Center, University of California, San Francisco, CA, USA, [4]Gladstone Institutes, University of California, San Francisco, CA, USA and [5]Illumina, Inc, San Diego, CA 92122 USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on February 20, 2020; revised on December 31, 2020; editorial decision on February 12, 2021; accepted on February 23, 2021

Abstract

Summary: The Probabilistic Identification of Causal SNPs (PICS) algorithm and web application was developed as a fine-mapping tool to determine the likelihood that each single nucleotide polymorphism (SNP) in LD with a reported index SNP is a true causal polymorphism. PICS is notable for its ability to identify candidate causal SNPs within a locus using only the index SNP, which are widely available from published GWAS, whereas other methods require full summary statistics or full genotype data. However, the original PICS web application operates on a single SNP at a time, with slow performance, severely limiting its usability. We have developed a next-generation PICS tool, PICS2, which enables performance of PICS analyses of large batches of index SNPs with much faster performance. Additional updates and extensions include use of LD reference data generated from 1000 Genomes phase 3; annotation of variant consequences; annotation of GTEx eQTL genes and downloadable PICS SNPs from GTEx eQTLs; the option of generating PICS probabilities from experimental summary statistics; and generation of PICS SNPs from all SNPs of the GWAS catalog, automatically updated weekly. These free and easy-to-use resources will enable efficient determination of candidate loci for biological studies to investigate the true causal variants underlying disease processes.

Availability and implementation: PICS2 is available at https://pics2.ucsf.edu.

Contact: kim.taylor@ucsf.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A fundamental problem in genome-wide association studies (GWAS) is the fine-mapping of genetic associations to causal variants, a problem made difficult by the existence of multiple variants in linkage disequilibrium (LD) on the same haplotype. The Probabilistic Identification of Causal SNPs (PICS) (Farh et al., 2015) algorithm and web application was developed as a fine mapping tool to determine the likelihood that each single nucleotide polymorphism (SNP) in LD with a reported index SNP is a true causal polymorphism. While a number of excellent fine-mapping methods have been developed (Hormozdiari et al., 2014; Huang *et al.*, 2017; Wellcome Trust Case Control Consortium *et al.*, 2012), PICS is notable for its ability to identify candidate causal SNPs within a locus using only the index SNP, which are widely available from published GWAS, whereas other methods require full summary statistics or full genotype data. The PICS algorithm utilizes the fact that for SNPs whose association is only due to LD with a true causal SNP, the strength of association scales asymptotically with r^2 to the causal SNP. However, the original PICS web application operates on a single SNP at a time, with slow performance, severely limiting its usability.

Here we describe a next-generation PICS tool, PICS2, which has been extended and updated as follows:

- The ability to request PICS analysis of batches of potentially thousands of index SNPs with much faster performance.
- Use of LD reference data generated from 1000 Genomes phase 3 (The 1000 Genomes Project Consortium *et al.*, 2015) and GRCh38 positions.

- Inclusion of the most severe consequence from the Ensembl (Cunningham *et al.*, 2019) Variant Effect Predictor (VEP) for every variant.
- Inclusion of the eQTL gene having the lowest *P*-value across all GTEx (GTEx Consortium, 2017) V8 tissues, if significant, for every variant.
- Option to generate PICS probabilities from experimental summary statistics rather than reference data.
- Support for multi-allelic markers.
- Automatic generation of PICS SNPs of all SNPs in the GWAS catalog (Buniello et al., 2019), updated weekly and available for download.
- An updated autoimmune disease PICS SNPs list available for download.
- Downloadable PICS SNPs for all GTEx V8 best eQTLs per gene, per tissue type.

2 Materials and methods

Infrastructure and performance. PICS2 is a web application in which the client html script sends HTTP requests to an Apache server, implemented in perl via CGI (Common Gateway Interface). For improved performance, the server queries a mysql database containing all LD reference data, and processes each input row independently and asynchronously.

LD reference data. We generated LD data for all 5 populations of the 1000 Genomes phase 3 GRCh38 data (European EUR, African AFR, Amerindian AMR, East Asian EAS, South Asian SAS) and all subjects (ALL). [We also recovered an additional 1.9 M SNPs which were missing in the GRCh38 dataset with error code ERROR_RSIDS_NOT_IN_DBSNP149. We lifted over the SNPs from 1000 Genomes GRCh37 data to GRCh38 using the UCSC liftover tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver) and mapped SNP IDs to dbSNP 151 using the NCBI mapping file.] For each population, LD was computed using PLINK (Purcell et al., 2007) for all variant pairs with <1 M base-pair distance in between or <100 000 SNP variants in between, and stored in the PICS2 database if $r^2 \geq 0.5$. The LD database also contains all SNPs in the original reference data, so that PICS2 results can distinguish between independent SNPs not in LD ($r^2 < 0.5$) with any other SNP and those that are not present in the reference data.

Multi-allelic support. For multi-allelic markers, prior to LD calculation reference VCF rows were split into one row per alternate allele using bcftools (https://samtools.github.io/bcftools/bcftools.html) and SNPs renamed to chr:pos:ref:alt format (e.g. '22:41400874: A: AC'). For input IDs—RSID or chromosome:position—without allele designations but which correspond to a multi-allelic marker and have therefore been renamed, we use the most common alternate allele; the chr:pos:ref:alt format can be used as input to request other alternate alleles.

Summary statistics. A new summary statistics feature is provided in a separate user interface. Given a user-specified significance threshold and window size, we compute PICS probabilities for all windows around significant SNPs in the provided data, merging windows when there are multiple significant SNPs within a window. The SNP with the most significant *P*-value in each window is treated as the index SNP for PICS, and r^2 with other SNPs in the window is estimated as the ratio of –log10(*P*-value) of the index SNP to –log10(*P*-value) of the SNP in LD. Assuming that the second SNP association is completely due to LD, by the formula for the Armitage Trend Test it will scale linearly with r^2 to the causal SNP. In addition, for reference we provide the upper bound Bayes factor (Stephens and Balding, 2009) for each association.

GWAS catalog. To automatically update the GWAS catalog PICS SNPs, we query the latest download from the NHGRI-EBI site (www.ebi.ac.uk) via a weekly cron task and determine if there have been changes since our previous release. Corresponding PICS2 requests for each entry are sent directly to the server using the linux curl command, and new results appended to the previous release.

Autoimmune disease SNP list. Autoimmune and inflammatory diseases listed in Supplementary Table S1 were extracted from the GWAS catalog as of January 11, 2019. Studies were kept if there were at least 5 entries with $P \leq$ 5e-8 (Supplementary Table S2); all SNPs in those studies with $P \leq$ 1e-6 were retained. For each disease, the SNP list was LD-pruned using the clump feature of PLINK (beginning with highest associated SNP, surrounding SNPs are pruned if $r^2 > 0.5$ within 250 kb), then run through PICS2. All SNPs with PICS probability > 0.025 are reported. Annotation was done using VEP for GENCODE primary transcripts; if there was no primary transcript than one transcript of the highest GENCODE level was selected at random.

Application to GTEx SNPs. We ran PICS2 on all GTEx (GTEx Consortium, 2017) version 7 eQTLs (best eQTL per gene if q-value ≤ 0.05; with permutation *P*-value as input) for 12 tissue types (Supplementary Table S3). Using VEP we obtained the distributions of variant types in these sets as well as for the autoimmune and full GWAS catalog SNPs, for all unique SNPs having a pics probability ≥ 0.025. We determined the overlap of the GTEx PICS SNPs and autoimmune PICS SNPs by disease and tissue type. Finally, in order to determine those SNPs with the strongest evidence of being causal for autoimmunity via the regulation of gene expression, we selected those SNPs in the PICS autoimmune-eQTL overlap with both PICS probabilities (as an autoimmune causal SNP and as an eQTL) greater than 85%.

GTEx annotation and downloads. LD-based PICS output includes the gene for which a variant has the strongest evidence for being a GTEx version 8 eQTL; namely having qval > 0.05 and having the lowest gene *P*-value among all tissue types. Also, downloadable resources include PICS variants (PICS probability > 0.025) for all GTEx version 8 best eQTL per gene (qval > 0.05) for all tissue types.

3 Results

PICS2. The PICS2 web application is currently available at https://pics2.ucsf.edu. We have successfully submitted batches of 50 000 and 100 000 input rows to the LD-based and summary statistics (SS) interfaces, respectively. LD-based PICS typically executes over 10 request rows/second; summary statistics processes several hundred rows per second. Supplementary Tables S4A–B and S5A–C show examples of LD versus SS PICS2 output, where SS input *P*-values are estimated from PICS probabilities.

Autoimmune and GWAS catalog PICS SNPs. These SNP sets are available on the PICS2 Data Portal. The autoimmune disease SNPs dataset contains 14 719 disease-SNP combinations across 24 autoimmune and inflammatory disease categories, with annotations as described above. As of this writing, the latest version of the GWAS SNPs were produced from the January 21, 2021 GWAS catalog and contains 9.8 million variants (2.0 million unique) across 4717 disease traits. Supplementary Figure S1 shows the distribution of variant types from VEP (all consequences, exon consequences and regulatory biotypes) for the autoimmune diseases, full GWAS catalog (January 31, 2020), and eQTL PICS SNPs having pics probability ≥ 0.025.

Overlap between autoimmune and GTEx PICS SNPs. Supplementary Table S3 shows the overlapping PICS SNPs (probability > 0.025) between the autoimmune disease and eQTL SNP sets. As previously reported, overlap is relatively modest, with average overlap per tissue type ranging from 3-8% and average overlap per disease ranging from 1% to 38% (Kawasaki disease is an outlier with 41% (12/29) of disease SNPs overlapping with eQTLs in most tissue types, but with a small number of SNPs). Table 1 shows all SNPs with PICS probabilities >85% for both an autoimmune disease category and an eQTL tissue type, thus these SNPs are those most likely to be causal and functioning as eQTLs. SNP rs1893592, found in an extended splice motif variant in the UBASH3A gene is mostly strongly implicated as an autoimmunity eQTL with high probabilities in 4 autoimmune disease categories (celiac, primary

Table 1. Variants with >85% AI and eQTL PICS probabilities

Gene symbol	GWAS autoimmune disease category	GTEx tissue type	RSID	AI disease PICS probability	eQTL PICS probability
CCL20	Inflammatory bowel disease	Thyroid	rs1811711	1	1
CCL20	Ulcerative colitis	Thyroid	rs1811711	1	1
TRAF3IP2-AS1	Psoriasis	Transformed fibroblasts	rs33980500	0.9999	0.9797
TRAF3IP2-AS1	Psoriasis	Unexposed skin	rs33980500	0.9999	0.9797
TRAF3IP2-AS1	Psoriasis	Thyroid	rs33980500	0.9999	0.9797
TRAF3IP2-AS1	Psoriasis	Whole blood	rs33980500	0.9999	0.9797
UBASH3A	Primary sclerosing cholangitis	Transverse colon	rs1893592	0.9992	0.8706
UBASH3A	Primary sclerosing cholangitis	Lung	rs1893592	0.9992	0.9494
UBASH3A	Primary sclerosing cholangitis	Thyroid	rs1893592	0.9992	0.9494
UBASH3A	Primary sclerosing cholangitis	Whole blood	rs1893592	0.9992	0.9494
UBASH3A	Rheumatoid arthritis	Transverse colon	rs1893592	0.9989	0.8706
UBASH3A	Rheumatoid arthritis	Lung	rs1893592	0.9989	0.9494
UBASH3A	Rheumatoid arthritis	Thyroid	rs1893592	0.9989	0.9494
UBASH3A	Rheumatoid arthritis	Whole blood	rs1893592	0.9989	0.9494
UBASH3A	Celiac disease	Transverse colon	rs1893592	0.9958	0.8706
UBASH3A	Celiac disease	Lung	rs1893592	0.9958	0.9494
UBASH3A	Celiac disease	Thyroid	rs1893592	0.9958	0.9494
UBASH3A	Celiac disease	Whole blood	rs1893592	0.9958	0.9494
UBASH3A	Autoimmunity[a]	Transverse colon	rs1893592	0.995	0.8706
UBASH3A	Autoimmunity[a]	Lung	rs1893592	0.995	0.9494
UBASH3A	Autoimmunity[a]	Thyroid	rs1893592	0.995	0.9494
UBASH3A	Autoimmunity[a]	Whole blood	rs1893592	0.995	0.9494
OVOL1	Allergy	EBV-transformed lymphocytes	rs10791824	0.9987	0.8908
JAZF1	Systemic lupus erythematosus	Pancreas	rs10245867	0.9967	0.9765
IRF4	Vitiligo	Lung	rs12203592	0.9957	0.9479
IRF4	Vitiligo	Whole blood	rs12203592	0.9957	0.9401
IL27	Type 1 diabetes	Whole blood	rs151234	0.9823	0.9039
EVI5	Systemic lupus erythematosus	Thyroid	rs6662618	0.9676	0.8856
EIF3C	Juvenile idiopathic arthritis	Pancreas	rs12928404	0.9648	0.8839
TUFM	Juvenile idiopathic arthritis	Transformed fibroblasts	rs12928404	0.9648	0.9104
RTEL1	Juvenile idiopathic arthritis	Whole blood	rs2738774	0.9509	0.8931

[a] Multiple autoimmune diseases in a single study.

sclerosing cholangitis, rheumatoid arthritis, and multiple autoimmune diseases) and with eQTL evidence in 4 tissues (transverse colon, lung, thyroid, and whole blood). UBASH3A negatively regulates T-cell signaling, facilitates T-cell apoptosis and variants in the gene have been associated with additional autoimmune diseases including systemic lupus erythematosus, atopic dermatitis, and type 1 diabetes (T1D) (Diaz-Gallo et al., 2013; Ge and Concannon, 2018; Li et al., 2017).

4 Conclusion

In summary, we have provided a next-generation fine-mapping tool, PICS2, to identify candidate causal SNPs for thousands of loci implicated in GWAS. In addition, we regularly update PICS2 SNPs from the GWAS catalog available for download from our web site. These free and easy-to-use resources will enable the genetics community to more efficiently determine and annotate candidate loci for biological studies to help determine the true causal variants underlying disease processes.

Funding

Our study was funded by PREMIER, a National Institutes of Health / National Institute of Arthritis and Musculoskeletal and Skin Diseases P30 Center for the Advancement of Precision Medicine in Rheumatology at UCSF (P30AR070155, K.T., L.C.), NIH/National Institute of Allergy and Infectious Diseases R01HG008140 (K.T., A.M., L.C.) and by the Sandler Asthma Basic Research Center, NIH R21AI1280471 (K.M.A.). A.M. holds a Career Award for Medical Scientists from the Burroughs Wellcome Fund, the Lloyd J. Old STAR award from the Cancer Research Institute (CRI), is an investigator at the Chan Zuckerberg Biohub and has received funding from the Innovative Genomics Institute (IGI) and the Parker Institute for Cancer Immunotherapy (PICI).

Conflicts of Interest: K.F. is employed by Illumina, Inc. A.M. is cofounder, member of the Boards of Directors and member of Scientific Advisory Boards of Spotlight Therapeutics and Arsenal Biosciences. A.M. has served as an advisor to Juno Therapeutics, was a member of the scientific advisory board at PACT Pharma and was an advisor to Trizell. A.M. has received honoraria from Merck and Vertex, a consulting fee from AlphaSights, and is an investor in and informal advisor to Offline Ventures. A.M. owns stock in Arsenal Biosciences, Spotlight Therapeutics and PACT Pharma. The Marson lab has received research support from Juno Therapeutics, Epinomics, Sanofi, GlaxoSmithKline, Gilead and Anthem.

References

Buniello,A. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.

Cunningham,F. *et al.* (2019) Ensembl 2019. *Nucleic Acids Res.*, **47**, D745–D751.

Diaz-Gallo,L.M. *et al.* (2013) Evidence of new risk genetic factor to systemic lupus erythematosus: the UBASH3A gene. *PLoS One*, **8**, e60646.

Farh,K.K. *et al.* (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, **518**, 337–343.

Ge,Y., and Concannon,P. (2018) Molecular-genetic characterization of common, noncoding UBASH3A variants associated with type 1 diabetes. *Eur. J. Hum. Genet.*, **26**, 1060–1064.

GTEx Consortium (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.

Hormozdiari,F. *et al.* (2014) Identifying causal variants at loci with multiple signals of association. *Genetics*, **198**, 497–508.

Huang,H. *et al.*; International Inflammatory Bowel Disease Genetics Consortium. (2017) Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature*, **547**, 173–178.

Li,Y. *et al.* (2017) Association of UBASH3A gene polymorphism and atopic dermatitis in the Chinese Han population. *Genes Immun.*, **18**, 158–162.

Purcell,S. *et al.* (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.*, **81**, 559–575.

Stephens,M., and Balding,D. (2009) Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.*, **10**, 681–690.

The 1000 Genomes Project Consortium. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Wellcome Trust Case Control Consortium. *et al.* (2012) Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.*, **44**, 1294–1301