FORGE : A tool to discover cell specific enrichments of GWAS associated SNPs in regulatory regions.

Ian Dunham, Eugene Kulesha, Valentina Iotchkova, Sandro Morganella and Ewan Birney

European Molecular Biology Laboratory
European Bioinformatics Institute (EMBL-EBI)
Wellcome Trust Genome Campus
Hinxton
Cambridge CB10 1SD
United Kingdom

Abstract

Genome wide association studies provide an unbiased discovery mechanism for numerous human diseases. However, a frustration in the analysis of GWAS is that the majority of variants discovered do not directly alter protein-coding genes. We have developed a simple analysis approach that detects the tissue-specific regulatory component of a set of GWAS SNPs by identifying enrichment of overlap with DNase I hotspots from diverse tissue samples. Functional element Overlap analysis of the Results of GWAS Experiments (FORGE) is available as a web tool and as standalone software and provides tabular and graphical summaries of the enrichments. Conducting FORGE analysis on SNP sets for 260 phenotypes available from the GWAS catalogue reveals numerous overlap enrichments with tissue–specific components reflecting the known aetiology of the phenotypes as well as revealing other unforeseen tissue involvements that may lead to mechanistic insights for disease.

Keywords

Genome Wide Association; GWAS; Hypersensitive sites; Regulatory Elements.

Identifying Regulatory Components of Genome Wide Association Study Hit Lists

A primary motivation for sequencing the human genome was to shed light on mechanisms involved in human disease. Since finishing the sequence there has been much activity in two areas towards that goal. In the first, extensive re-sequencing of individual genomes has provided comprehensive lists of human variations, which can in turn be examined for association with disease and other phenotypes in Genome Wide Association Studies (GWAS) [1]. In the second area, efforts have been undertaken to identify the specific sequences that enact function within the genome including, but not restricted to, regions defining genes and their controlling elements [2-4]. The aim, of course, is to understand the associations uncovered in the first approach in the context of the annotations delivered from the second.

The past few years have seen a dramatic growth in the number of variants associated with disease by GWAS [1]. An extensive catalogue of GWAS associations has been compiled containing nearly 14,000 associations of variants to phenotypes [5]. However, a crucial observation is that the majority of the variants observed do not directly affect the coding regions of protein coding genes. Notwithstanding that the reported variant for an association may be in linkage disequilibrium with a causal variant affecting a protein coding sequence, regulatory regions have been demonstrated to be linked to both specific diseases associations [6-18] (see [19] for review and further examples) and to be enriched in bulk in SNPs found across all GWAS [2, 20-22]. The ENCODE consortium reported that GWAS single nucleotide variants are substantially enriched in regulatory regions and up to 80% of GWAS variants have a potential regulatory interpretation via overlap with regulatory annotation [2, 21, 22]. Furthermore, Maurano et al [21] showed that regulatory regions revealed by the DNase-seq method show a cell specific enrichment for GWAS variants in specific phenotypes consistent with probable physiological mechanisms. Trynka et al [23] similarly found that regulatory elements identified by the histone modification H3K4me3 show a phenotypically relevant cell specific overlap with GWAS SNPs. Several tools exist to highlight the specific overlaps of individual GWAS SNPs with potential regulatory regions [24, 25]. To date however, much of the focus on this work has been on prioritising variants, rather than exploring the extensive cell type information present in the large-scale projects.

We have developed a simple but powerful approach that identifies significant cell specific enrichments in regulatory regions for sets of single nucleotide variants, typically from GWAS. We name the approach Functional element Overlap analysis of the Results of GWAS Experiments or FORGE, and have implemented it as both a rapid web tool for ENCODE[26] and Roadmap Epigenome project DNase-seq data[27] and a free-standing open source software. The web tool produces two alternative graphical outputs for exploration alongside tabulated enrichment data. FORGE analysis across all eligible phenotypes in the entire GWAS catalog [5] identifies numerous interesting patterns of enrichment by cell type and suggests tissues to focus on for future follow up studies.

Forge Analysis approach

FORGE analysis provides a method to view the tissue specific regulatory component of a set of variants. In its current implementation, FORGE analysis takes a set of single nucleotide polymorphisms (SNPs), such as those SNPs reported above the genome wide significance threshold ($p < 5e-8$) in a GWAS study, optionally filters the SNPs to remove all bar one SNP from a region in high LD ("LD pruning") and determines whether there is enrichment for overlap with putative regulatory elements compared to a matched background of SNP sets. Initially the elements considered are DNase I hotspots generated from either the ENCODE [26] or Roadmap Epigenomics projects DNase I data by the Hotspot method [28, 29], because of both the comprehensiveness of the sites identified and the broad range of cell types for which DNase I data was available. DNase I hotspots can be regarded as regions of general DNase I sensitivity.

For each set of test SNPs, an overlap analysis is performed against the DNase I hotspots for each available cell sample separately (125 samples for ENCODE, 299 for Roadmap, described in Supplementary Table S1), and the number of overlaps is counted. Major potential confounders in this analysis are the many biases of GWAS SNP distribution on the genome. To account for this a background distribution of the expected overlap counts for this SNP set is obtained by identifying 1000 matched background SNP sets of the same number of SNPs, matching each test SNP with an equivalent SNP by decile bin for each of G+C content (GC), minor allele frequency (maf) and distance to the nearest transcription start site (TSS). The matched background SNPs sets are overlapped with the DNase I hotspots and the background distribution of overlap counts is determined. The enrichment of the test SNP set for each sample is expressed as the binomial P value of the test SNP set given the background overlap distribution. The FORGE results are presented in interactive and static graphical and tabular forms by cell type. Enrichments above the background distribution with binomial P values less than 0.01 corrected for multiple testing are considered significant and are highlighted in red in the graphical output. Enrichments with $p <= 0.05$ are also highlighted in pink (Figure 1). As the DNase I patterns are not independent between cell types, we conducted simulation experiments with randomly selected input SNPs. We chose 1000 random test SNP sets for each of a series of SNP counts ranging between 5 and 100 SNPs and conducted FORGE analysis on both ENCODE and Roadmap data. The false positive rate was determined as the number of cell type enrichments identified greater than the significance thresholds used by FORGE expressed as the proportion of the total number of sample overlap tests performed (424,000) for each SNP count. This analysis showed that the P value thresholds are reasonably well calibrated to false positive levels around 0.5-0.75 %. In practice we correct the thresholds further for multiple testing across different tissues so as to be more stringent and so expect typical false positive levels to be less. As discussed below, many of the GWAS SNP sets do not reveal any enrichment, consistent with low false positive rates.

The FORGE tool

We have implemented FORGE as a web tool available at http://browser.1000genomes.org/Homo_sapiens/UserData/Forge. The interface accepts a list of SNPs by dbSNP RefSNP identifier (RSID) or by genomic location on human genome build GRCh37 in either Variant Call Format (VCF, http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41) or Bed format (Personal Genome SNP format, http://genome.ucsc.edu/FAQ/FAQformat.html#format10), and allows specification of the background selection from two common sets of GWAS SNP typing microarrays. LD filtering is achieved at either $r^2 >= 0.8$ or $r^2 >= 0.1$ using 1000 genomes project population data. The outputs of the analysis are an interactive graphic for exploration of the analysis, a static pdf for printing or publication (Figure 1), and a table of enrichments in either an interactive or standard tab separated format.

In addition the code is available to download from https://github.com/iandunham/Forge, with the required database files available at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/browser/forge. Installed on a Macbook Pro with core i7 processor, 16Gb RAM and a solid state hard drive, a typical FORGE analysis of a set of 55 SNPs with 1000 background tests is accomplished in around 30s, and in 35s with LD filtering.

Gallery of Examples

We ran FORGE analysis on 260 phenotypes in the NCBI GWAS catalog NCBI GWAS catalog [5, 30] with a reported associated SNP count of 5 or more after LD pruning (see Supplementary Table S2 for list of phenotypes analysed and references) at genome-wide significance. Complete tables of results for all SNP sets analysed are included in the data directory of the Github release, https://github.com/iandunham/Forge. 35 and 60 out of 260 SNP sets had at least one significant enrichment at the P value thresholds of 0.05 and 0.01 after correction for multiple testing, respectively (Table 1). A set of example positive outputs from this analysis is available from http://www.1000genomes.org/forge-gwas-catalog-example-gallery11 (PDF format). Removing SNPs that directly alter a protein coding exon from the GWAS catalog sets did not substantively alter the patterns of enrichments (data not shown).

Figure 1 shows a series of example FORGE analyses for autoimmune disease studies on the Roadmap Epigenome samples (references for the studies are provided in Supplementary Table S2). In each case there is a clear signal for enrichment of overlap with DNase I hotspots in the blood-derived samples including cells of immune function. In more detail, for those phenotypes where there is involvement of T cell activation or invasion in the aetiology (e.g. Crohn's disease, Multiple Sclerosis) there is enrichment in the CD3, CD4 and CD8 positive samples containing T cells as well as enrichment in the CD56 positive sample

including NK cells. In addition, these disease SNPs overlap with hotspots present in the fetal thymus samples, consistent with the location of T cells maturation. Further signals specific to the individual aetiologies are also identified. Crohn's disease SNPs show enrichment of overlap with hotspots in the fetal small and large intestine samples, as well as fibroblasts and skin cells. For inflammatory bowel disease SNPs there is a much more general enrichment, in addition to the specific immune cells, which may be consistent with the more generalized inflammation. In contrast, for autoimmune diseases where the primary involvement is a B cell response (Rheumatoid arthritis (RA), Systemic lupus erythematosus (SLE)), the most pronounced overlap enrichment is in CD19 positive samples characteristic of B cell activation or circulating plasma cells. In rheumatoid arthritis there is also some overlap enrichment for samples characteristic of T cells and thymocytes, but it is relatively less, and this is much less pronounced for SLE. Thus, FORGE analysis reveals tissue specific enrichment of overlap for GWAS SNPs with regulatory regions indicative of known tissue involvement in the disease aetiology.

The tissue specific enrichment of overlap is not specific to just autoimmune disease (see results gallery at http://www.1000genomes.org/forge-gwas-catalog-example-gallery11). For instance, for QRS duration the GWAS associated SNPs are strongly enriched for overlap with fetal heart samples. GWAS SNPs associated with pulmonary function measured by spirometry are enriched for overlap with hotspots in fetal lung cells and lung cell lines. For red blood cell traits and platelet count the major overlap enrichment signal is in CD34 positive hematopoietic progenitor cells consistent with their role in both red blood cell and platelet development. In contrast for GWAS SNPs involved in height, the overlap enrichment is not tissue specific but is more general over many tissues and cell lines. There are further examples displayed in the results gallery at http://www.1000genomes.org/forge-gwas-catalog-example-gallery11, in most case consistent with expected disease aetiologies.

Discussion

FORGE (Functional element Overlap analysis of the Results of GWAS Experiments) analysis is a straightforward and fast method to examine sets of nucleotide variants, typically identified in GWAS studies, for tissue specific regulatory signals. It presents a graphical overview of overlap enrichment with DNase I hotspots that quickly provides evidence of a regulatory component to SNPs associated with a phenotype, and highlights potentially mechanistically relevant cells or tissues. A typical usage scenario would be to analyse a set of GWAS SNPs identified above genome wide significance to reveal the regulatory component of the association. Furthermore the cell or tissue enrichments might be consistent with prior expectation of the disease aetiology providing additional confidence in the SNP set identified, or might provide new insights as to potential sites of disease mechanism.

The statistical approach we used here relies on the careful matching of background behaviour of SNPs with calibrations by randomization for the per cell type enrichment. The underlying biases of GWAS SNP distributions with

respect to TSS distance, maf and GC are not easy to model parametrically. However other approaches which would make assumptions of homogeneity (such as the Poisson distribution) or of regional heterogeneity (Genome Structure Correction, [31]) would not be able to capture these known biases. It is important to note that as alternate SNP resources are utilized in GWAS, the appropriate background SNPs must be used for control. For instance a switch to genotyping by genome sequencing or extensive use of imputation requires revision of the background. New and updated background sets can be implemented as required in particular for genome sequencing GWAS approaches and higher density genotyping arrays.

Not all GWAS study SNP sets downloaded from the GWAS catalog showed overlap enrichment with the DNase I hotspots. In these cases all sample points were above the P value thresholds (blue points). This could occur because there is no regulatory component underlying the GWAS association in these phenotypes, because the associated SNPs do not contain mechanistically causal SNPs, because the relevant tissue is not present in the available DNase I datasets or because of low power in the GWAS study to detect regulatory effects. As further data sets are release by the ENCODE, Roadmap Epigenome and other projects these can be incorporated into the database to provide coverage of further cell types. In addition the approach could be readily extended to other data types including regions of specific histone modification as used by Trynka et al [23] or relevant transcription factor binding regions.

60 out of 260 sets of GWAS SNPs from the GWAS catalog for specific phenotypes had overlap enrichments detected in at least one DNase I hotspot sample (Table 1). As described above in several cases, the patterns of tissue–specific enrichment are highly evocative of the known aetiologies of the phenotypes, but can also reveal additional tissue involvements that require further investigation. We encourage interested parties to peruse the gallery of results for their own phenotypes, as well as running new SNP sets discovered in GWAS either through the web interface or with the standalone software.

Methods

DNase I Data

ENCODE consortium hotspots [26] were obtained from ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/openchrom/jan2011/combined_hotspots/. Roadmap Epigenome DNAse1 sequencing tag alignments were obtained from http://www.genboree.org/EdaccData/Current-Release/experiment-sample/Chromatin_Accessibility/. The files used correspond to that part of the Gene Expression Omnibus (GEO) accession GSE18927 beyond the data use embargo date. These alignments were processed by the Hotspot (http://www.uwencode.org/proj/hotspot/) [28, 29] method with the default parameters to give hotspot and peak files. For this analysis we choose to use the

hotspots which are regions of general DNase I sensitivity rather than peaks which are more similar to DNase I hypersensitive sites because, although the method works with peaks, hotspots reveal more tissue specific signal (data not shown). Cell and tissue assignments for each of the data sets were made using the decodings available from the ENCODE Data Coordination Center tables (https://genome.ucsc.edu/encode/cellTypes.html) or from the BioSamples database (http://www.ebi.ac.uk/biosamples/) sampleGroup SAMEG31306. A list of samples used is provided in Supplementary Table S1.

GWAS SNP data

The complete collection of SNPs discovered in GWAS studies curated in the NHGRI GWAS Catalog [5] were downloaded from http://www.genome.gov/gwastudies/ [30](Accessed 3rd September 2014). SNPs were grouped according to the annotation provided in the Disease/Trait field and only sets with 5 or more non-redundant SNPs were retained. See Supplementary Table S2 for list of SNP sets analysed. A set of files of the SNPs included in analysis for each phenotype is available in the data directory of the GitHub release, https://github.com/iandunham/Forge. For Forge analysis SNP sets were further filtered by LD pruning removing all but one SNP from a set of SNPs at $r^2 >= 0.8$ in the 1000 genomes data (see below) and were analysed for both ENCODE and Roadmap Epigenome DNase I hotspots, selecting background SNP sets from the default GWAS genotyping array SNPs.

Preparation of FORGE overlaps

The FORGE tool utilises either an SQLite (command line tool, http://www.sqlite.org) or MySQL (web tool, http://www.mysql.com) database of the overlaps of every 1000 genomes project (http://www.1000genomes.org) [32] SNP with the ENCODE and Epigenome Roadmap DNase1 hotspots. To prepare this database, SNPs from the 1000 genomes phase 1 integrated call data set were downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_call_sets and compared to indexed DNase I hotspots using tabix from the SAMtools package (http://samtools.sourceforge.net/tabix.shtml) [33] using a distributed approach on the EBI compute farm. The overlaps for each SNP were stored in a single large indexed table of SNP location and identifier with binary strings representing the presence (1) or absence (0) of overlap in each sample for each of the hotspot data sets (ENCODE or Roadmap).

Background SNP parameters

To prepare sets of background SNPs matched to the test SNP set, FORGE matches SNPs based on GC, maf and TSS distance, and repeats the overlap analysis for each of 1000 background sets. Overall population mafs were obtained from the 1000 genomes project phase 1 integrated call data set at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/integrated_c

all_sets. To control for the processes involved in selecting SNPs for genotyping, only 1000 genomes phase 1 SNPs that had been included on one of the common genotyping platforms as described in Ensembl (http://www.ensembl.org/info/genome/variation/data_description.html#variation_sets) were considered further. This left either 1875813 SNPs across various platforms (Affy GeneChip 100K Array, Affy GeneChip 500K Array Affy SNP6, HumanCNV370-Quadv3, HumanHap300v2, HumanHap550v3.0, Illumina Cardio Metabo, Illumina Human1M-duoV3, Illumina Human660W-quad) or 2231212 SNPs across the Illumina HumanOmni2.5 array. TSS distance was determined for each remaining SNP relative to the TSS defined by the Gencode project [34, 35] given in ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/gencode/jan2011/Gencodev7_CAGE_TSS_clusters_June2011.gff.gz using Bedops closest-features [36]. GC was determined for a 100 bp window centred on the SNP at base 50. The SNPs were then sorted into 1000 bins partitioned by deciles for each of GC, maf and TSS. For each SNP in a test set, the corresponding bin is identified based on its GC, maf and TSS distance, and background selections are made from that bin.

FORGE analysis

A set of SNPs can be presented to FORGE as a list of RSIDs or by genome location on human genome build GRCh37 in either VCF or Bed formats. If RSIDs are not provided in one of these formats, the genome coordinates are used to identify the RSID. SNPs not present in the 1000 genomes phase 1 integrated call data set are excluded from the analysis. With LD pruning selected a single SNP (the first in the file) is chosen from LD clusters within either $r^2 >= 0.8$ or $r^2 >= 0.1$ as specified. For each analyzable SNP in the test set, overlaps are retrieved from the FORGE database, and a count of total hotspot overlaps is recorded for each DNase I sample (cell) for the test SNP set. One hundred matching background SNP sets containing the same number of SNPs as the test SNP set are selected, matched for GC, maf and TSS distance by decile bins as described above. Overlaps for each of the SNPs in each of the background SNP sets are also retrieved from the database and an overlap count for each background set in each DNase I sample is recorded. For each test SNP set, the background probability of overlap is determined from the 1000 background set overlap counts and the probability of the observed test result under a binomial distribution is calculated. The P value thresholds of 0.05 and 0.01 are corrected for multiple testing by division by the number of tissue groupings tested, and the corrected threshold is used. The use of tissue as the unit for sample grouping is consistent with the groupings obtained by hierarchical clustering of samples by Dnase 1 data (results not shown). The corrected thresholds are therefore more stringent than established by the random trials.

FORGE outputs

FORGE generates both tabular and graphic descriptions of the enrichment of

overlap for the test SNPs for each DNase I hotspot sample. A tab-separated values (TSV) file is output including columns for the binomial P value, cell, tissue, filename of the sample hotspots, SNPs that contribute to the enrichment, and the GEO accession for each sample. This data is also provided as an interactive table produced using the Datatables (https://datatables.net/) plug-in for the jQuery Javascript library accessed through the rCharts package (http://ramnathv.github.io/rCharts/).

Each of the graphic outputs presents the $-\log_{10}$ binomial p by cell sample. A pdf graphic is generated using base R graphics (http://www.r-project.org). The interactive Javascript graphic is generated using the rCharts package (http://ramnathv.github.io/rCharts/) to interface with the dimple d3 libraries (http://dimplejs.org). In both cases cells are grouped alphabetically by tissue, and for the pdf alphabetically by cell. The interactive graphic stacks replicate samples at the same x coordinate. In each of the graphics the colouring of results by P value is consistent, blue ($p > 0.05$ equivalent after correction), pink ($0.05 => p < 0.01$), and red ($p <= 0.01$). The corrected P value threshold is given on the pdf output.

False positive rates

To estimate false positive rates, 1000 sets of SNPs at each of a series of SNP counts between 5 and 300 SNPs were randomly chosen from the 1000 genome phase 1 integrated SNP set. FORGE analysis was run for each set across the ENCODE and Roadmap Epigenome data, and the number of tests with P values less than thresholds ranging from 0.05 to 0.001 were recorded. These represent the false positives from 1000 trials at each of 424 samples i.e. 424,000 tests, and were used to calculate false positive rates at each significance threshold.

Hierarchical Clustering of DNase I samples

The hierarchical clustering solution was obtained using a multi-scale bootstrap resampling approach. We first computed a binary regulatory signature for each cell type classifying each DNase I site as active or inactive in each cell type sample. Hierarchically clustering of the binary regulatory signatures was by Euclidean distance with Ward's agglomerative method using the pvclust R/CRAN package with default values (http://cran.r-project.org/web/packages/pvclust/index.html). Finally, we identified clusters supported by the data at a bootstrap probability p value < 0.01.

Access and Source Code

FORGE is available through a web interface at http://browser.1000genomes.org/Homo_sapiens/UserData/Forge. The source code for FORGE is available on GitHub at https://github.com/iandunham/Forge with the Forge.db sqlite database and background selection hash tables available

at ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/browser/forge_11. FORGE has been successfully been installed and run on Mac OSX 10.8.4 and Red Hat Linux.

Links

Web tool

http://browser.1000genomes.org/Homo_sapiens/UserData/Forge

Web documentation

http://www.1000genomes.org/forge-analysis-11

Results Gallery

http://www.1000genomes.org/forge-gwas-catalog-example-gallery11

Source Code and Database

https://github.com/iandunham/Forge

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/browser/forge_11

Competing Interests

The authors declare that they have no competing interests.

Authors' contributions

ID and EB designed the analysis and wrote the paper. ID wrote software and ran analysis. SM conducted the clustering of cell types by DNase I regions. EK implemented the web interface. VI provided statistical advice and discussion.

Acknowledgements

Many thanks to Jan Quell for initial implementation of the pdf graphic and to Ramnath Vaidyanathan and @timelyportfolio for assistance with rCharts. We thank Prof Ajay Shah and Marc-Philip Hitz for comments on the results of cardiac phenotypes, and Graham Ritchie, and Nicole Soranzo for discussions.

Figures

Figure 1. FORGE analysis results for GWAS of several autoimmune diseases on Roadmap Epigenome DNase I hotspots. A series of FORGE analysis results are presented for autoimmune phenotype GWAS SNPs. Each point represents the Z score (y axis) of the enrichment of the test SNP set compared to matched background SNPs on a single DNase I hotspot sample, organized by tissue as indicated by the brown labels at the top of the figure, and alphabetically by cell sample (x axis). Where informative, additional labels at the bottom of the figure highlight relevant distinct cell types. Red points are at Z >= 3.39 (empirical false positive rate <= 0.005 for 25 SNPs or more), pink points at Z >= 2.58. Full lists of the cells and results for each analysis are available in the Github data directory at https://github.com/iandunham/Forge. Phenotypes are labeled beneath each result.

Tables

Table 1: A list of SNP sets with positive enrichments. GWAS SNP set gives the phenotype of the study for which these SNPs were found to be associated as recorded in the GWAS catalog. SNP Count is the number of SNPs analysed before LD pruning. High and Low columns give the number of DNas1 cell samples found to be enriched for overlap in the Forge analysis at binomial p >= 0.01 (High) and p >= 0.05 (Low) thresholds. Further details of the SNP sets analysed are given in Supplementary Table S2.

GWAS SNP Set (Phenotype)	SNP count	High	Low
Acute lymphoblastic leukemia B-cell precursor	6	6	15
Acute lymphoblastic leukemia childhood	9	0	1
Adiponectin levels	28	0	1
Allergic sensitization	10	0	1
Atrial fibrillation	14	1	13
Birth weight	7	0	1
Blood pressure	43	7	19
Breast cancer	84	13	49
Celiac disease	30	9	18
Celiac disease and Rheumatoid arthritis	12	8	12
Chronic kidney disease	27	0	8
Chronic lymphocytic leukemia	27	40	62
Colorectal cancer	26	1	4
Corneal structure	27	0	3
Crohns disease	157	39	76
Endometriosis	8	2	5
Erythrocyte sedimentation rate	5	1	3
Fasting glucose-related traits	17	0	2
Fasting glucose-related traits interaction with BMI	22	0	4
Fractional exhaled nitric oxide childhood	6	0	5
HDL cholesterol	123	1	3
Heart rate	24	0	1
Height	346	81	143
Hematological parameters	9	0	1
Hodgkins lymphoma	9	0	1
IgA nephropathy	7	1	4
Inflammatory bowel disease	117	139	207
Liver enzyme levels gamma-glutamyl transferase	26	1	6
Mean corpuscular hemoglobin	29	2	4
Mean platelet volume	51	1	2
Migraine	13	0	2

Multiple myeloma	6	2	9
Multiple sclerosis	74	32	43
Myopia pathological	37	0	2
Platelet counts	64	30	51
Primary biliary cirrhosis	26	1	2
Proinsulin levels	9	0	1
Prostate cancer	86	2	11
Pulmonary function	29	7	24
QRS duration	12	2	8
QT interval	29	0	2
Red blood cell traits	60	3	9
Renal function-related traits BUN	13	9	30
Restless legs syndrome	9	0	1
Rheumatoid arthritis	127	8	16
Schizophrenia	49	0	1
Sphingolipid levels	13	0	1
Systemic lupus erythematosus	51	7	14
Systemic lupus erythematosus and Systemic sclerosis	11	0	6
Systemic sclerosis	14	1	4
Systolic blood pressure	24	0	9
Telomere length	9	0	1
Thyroid cancer	6	0	1
Thyroid volume	6	0	2
Type 1 diabetes	53	0	1
Ulcerative colitis	88	1	1
Urate levels	49	3	13
Waist-hip ratio	8	3	25
White blood cell count	18	1	2
White blood cell types	12	1	4

Supplementary Tables

Supplementary Table S1

List of DNase I hotspot samples included in FORGE analysis. The table lists details of the 125 ENCODE and 299 Roadmap Epigenome samples in tab separated value format (tsv). The fields are

File : File name

Lab : The data-generating lab. One of UW (University of Washington, John Stamatoyannopoulos lab), Duke:UNC:UTA (Duke University, Greg Crawford lab) or combined representing a merged dataset from both labs.

Experiment type : always DNase-seq in the current implementation.

Project : Either ENCODE or Roadmap

Cell : The cell type

Tissue : Tissue name derived as described above.

Datatype : Always hotspots in the current implementation.

Short name : A short sample name used for plotting.

Individual : Either the code for the individual sample as described in Biosamples or NA if not available

GEO accession : The GEO accession where found, or "Not found" if it could not be deconvoluted.

Supplementary Table S2

List of phenotypes analysed, with non-redundant SNP counts and Pubmed identifiers for the studies involved. This table is also available in the data directory of the GitHub release, https://github.com/iandunham/Forge.

References

1. Visscher PM, Brown MA, McCarthy MI, Yang J: **Five years of GWAS discovery.** *Am J Hum Genet* 2012, **90:**7-24.
2. The ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489:**57-74.
3. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al: **The NIH Roadmap Epigenomics Mapping Consortium.** *Nat Biotechnol* 2010, **28:**1045-1048.
4. Chadwick LH: **The NIH Roadmap Epigenomics Program data resource.** *Epigenomics* 2012, **4:**317-324.
5. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci U S A* 2009, **106:**9362-9367.
6. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, Li X, Li H, Kuperwasser N, Ruda VM, et al: **From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus.** *Nature* 2010, **466:**714-719.
7. Lettice LA, Heaney SJ, Purdie LA, Li L, de Beer P, Oostra BA, Goode D, Elgar G, Hill RE, de Graaff E: **A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly.** *Hum Mol Genet* 2003, **12:**1725-1735.
8. Emison ES, McCallion AS, Kashuk CS, Bush RT, Grice E, Lin S, Portnoy ME, Cutler DJ, Green ED, Chakravarti A: **A common sex-dependent mutation in a RET enhancer underlies Hirschsprung disease risk.** *Nature* 2005, **434:**857-863.
9. Wasserman NF, Aneas I, Nobrega MA: **An 8q24 gene desert variant associated with prostate cancer risk confers differential in vivo activity to a MYC enhancer.** *Genome Res* 2010, **20:**1191-1197.
10. De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen D, Yoshinaga Y, de Jong P, et al: **A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter.** *Science* 2006, **312:**1215-1217.
11. Gaulton KJ, Nammo T, Pasquali L, Simon JM, Giresi PG, Fogarty MP, Panhuis TM, Mieczkowski P, Secchi A, Bosco D, et al: **A map of open chromatin in human pancreatic islets.** *Nat Genet* 2010, **42:**255-259.
12. Jiang Y, Shen H, Liu X, Dai J, Jin G, Qin Z, Chen J, Wang S, Wang X, Hu Z, Shen H: **Genetic variants at 1p11.2 and breast cancer risk: a two-stage study in Chinese women.** *PLoS One* 2011, **6:**e21563.
13. Lubbe SJ, Pittman AM, Olver B, Lloyd A, Vijayakrishnan J, Naranjo S, Dobbins S, Broderick P, Gomez-Skarmeta JL, Houlston RS: **The 14q22.2 colorectal cancer variant rs4444235 shows cis-acting regulation of BMP4.** *Oncogene* 2012, **31:**3777-3784.
14. Zhao JY, Yang XY, Gong XH, Gu ZY, Duan WY, Wang J, Ye ZZ, Shen HB, Shi KH, Hou J, et al: **Functional variant in methionine synthase reductase**

intron-1 significantly increases the risk of congenital heart disease in the Han Chinese population. *Circulation* 2012, **125**:482-490.

15. Harismendy O, Notani D, Song X, Rahim NG, Tanasa B, Heintzman N, Ren B, Fu XD, Topol EJ, Rosenfeld MG, Frazer KA: **9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response.** *Nature* 2011, **470**:264-268.
16. Iida A, Takahashi A, Kubo M, Saito S, Hosono N, Ohnishi Y, Kiyotani K, Mushiroda T, Nakajima M, Ozaki K, et al: **A functional variant in ZNF512B is associated with susceptibility to amyotrophic lateral sclerosis in Japanese.** *Hum Mol Genet* 2011, **20**:3684-3692.
17. Alcina A, Fedetz M, Fernandez O, Saiz A, Izquierdo G, Lucas M, Leyva L, Garcia-Leon JA, Abad-Grau Mdel M, Alloza I, et al: **Identification of a functional variant in the KIF5A-CYP27B1-METTL1-FAM119B locus associated with multiple sclerosis.** *J Med Genet* 2013, **50**:25-33.
18. Han YJ, Ma SF, Wade MS, Flores C, Garcia JG: **An intronic MYLK variant associated with inflammatory lung disease regulates promoter activity of the smooth muscle myosin light chain kinase isoform.** *J Mol Med (Berl)* 2012, **90**:299-308.
19. Lee TI, Young RA: **Transcriptional regulation and its misregulation in disease.** *Cell* 2013, **152**:1237-1251.
20. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al: **Mapping and analysis of chromatin state dynamics in nine human cell types.** *Nature* 2011, **473**:43-49.
21. Maurano MT, Wang H, Kutyavin T, Stamatoyannopoulos JA: **Widespread site-dependent buffering of human regulatory polymorphism.** *PLoS Genet* 2012, **8**:e1002599.
22. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M: **Linking disease associations with regulatory information in the human genome.** *Genome Res* 2012, **22**:1748-1759.
23. Trynka G, Sandor C, Han B, Xu H, Stranger BE, Liu XS, Raychaudhuri S: **Chromatin marks identify critical cell types for fine mapping complex trait variants.** *Nat Genet* 2013, **45**:124-130.
24. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, et al: **Annotation of functional variation in personal genomes using RegulomeDB.** *Genome Res* 2012, **22**:1790-1797.
25. Ward LD, Kellis M: **HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants.** *Nucleic Acids Res* 2012, **40**:D930-934.
26. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B, et al: **The accessible chromatin landscape of the human genome.** *Nature* 2012, **489**:75-82.
27. Stergachis AB, Neph S, Reynolds A, Humbert R, Miller B, Paige SL, Vernot B, Cheng JB, Thurman RE, Sandstrom R, et al: **Developmental fate and cellular maturity encoded in human regulatory DNA landscapes.** *Cell* 2013, **154**:888-903.

28. John S, Sabo PJ, Thurman RE, Sung MH, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA: **Chromatin accessibility pre-determines glucocorticoid receptor binding patterns.** *Nat Genet* 2011, **43:**264-268.
29. Sabo PJ, Hawrylycz M, Wallace JC, Humbert R, Yu M, Shafer A, Kawamoto J, Hall R, Mack J, Dorschner MO, et al: **Discovery of functional noncoding elements by digital analysis of chromatin structure.** *Proc Natl Acad Sci U S A* 2004, **101:**16837-16842.
30. **A Catalog of Published Genome-Wide Association Studies. Available at: www.genome.gov/gwastudies. Accessed 07 August 2013.** [www.genome.gov/gwastudies]
31. Bickel PJ, Boley N, Brown JB, Huang H, Zhang NR: **Subsampling methods for genomic inference.** . *Ann Appl Stat* 2010, **4:**1660-1697.
32. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA: **An integrated map of genetic variation from 1,092 human genomes.** *Nature* 2012, **491:**56-65.
33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
34. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al: **Landscape of transcription in human cells.** *Nature* 2012, **489:**101-108.
35. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al: **GENCODE: the reference human genome annotation for The ENCODE Project.** *Genome Res* 2012, **22:**1760-1774.
36. Neph S, Kuehn MS, Reynolds AP, Haugen E, Thurman RE, Johnson AK, Rynes E, Maurano MT, Vierstra J, Thomas S, et al: **BEDOPS: high-performance genomic feature operations.** *Bioinformatics* 2012, **28:**1919-1920.