GREEN-DB: a framework for the annotation and prioritization of non-coding regulatory variants in whole-genome sequencing

Giacopuzzi E

1. Wellcome Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK.
2. National Institute for Health Research Oxford Biomedical Research Centre, Oxford, OX4 2PG, UK

edoardo.giacopuzzi@well.ox.ac.uk

Popitsch N

1. Wellcome Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK.
2. Institute of Molecular Biotechnology of the Austrian Academy of Sciences (IMBA), Dr. Bohr-Gasse 3, VBC, 1030, Vienna, Austria

niko.popitsch@well.ox.ac.uk

Taylor JC

1. Wellcome Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK.
2. National Institute for Health Research Oxford Biomedical Research Centre, Oxford, OX4 2PG, UK

jenny.taylor@well.ox.ac.uk

Corresponding author: Taylor JC, jenny.taylor@well.ox.ac.uk

Abstract

Background: Non-coding variants have emerged as important contributors to the pathogenesis of human diseases, not only as common susceptibility alleles but also as rare high-impact variants. Despite recent advances in the study of regulatory elements and the availability of specialized data collections, the systematic annotation of non-coding variants from genome sequencing remains challenging.

Results: We integrated 24 data sources to develop a standardized collection of 2.4 million regulatory elements in the human genome, transcription factor binding sites, DNase peaks, ultra-conserved non-coding elements, and super-enhancers. Information on controlled gene(s), tissue(s) and associated phenotype(s) are provided for regulatory elements when possible. We also calculated a variation constraint metric for regulatory regions and showed that genes controlled by constrained regions are more likely to be disease-associated genes and essential genes from mouse knock-out screenings. Finally, we evaluated 16 non-coding impact prediction scores providing suggestions for variant prioritization. The companion tool allows for annotation of VCF files with information about the regulatory regions as well as non-coding prediction scores to inform variant prioritization. The proposed annotation framework was able to capture previously published disease-associated non-coding variants and its integration in a routine prioritization pipeline increased the number of candidate genes, including genes potentially correlated with patient phenotype, and established clinically relevant genes.

Conclusion: We have developed a resource for the annotation and prioritization of regulatory variants in WGS analysis to support the discovery of candidate disease-associated variants in the non-coding genome.

Keywords (3 to 10)

Non-coding variants, whole-genome sequencing analysis, regulatory elements, variant prioritization, GREEN-DB, GREEN-VARAN, human genomics

Background

The precise spatiotemporal control of gene expression plays a fundamental role in developmental processes and cellular functions and consequently, is essential in determining human phenotypes [1–3]. Gene expression is controlled by the interaction of distal regulatory elements, such as enhancers and silencers, with gene promoters mediated by complex networks of transcription factors (TF) binding to these genomic regions [4–7]. Thus sequence variants within these regulatory regions can alter TF binding and/or enhancer-promoter interactions, resulting in gene expression dysregulation and eventually disease [8–13]. The contribution of regulatory regions in human diseases is also supported by a myriad of genome-wide association studies (GWAS), showing that most disease-risk variants lie in non-coding regions [14–16]. In recent years, our knowledge about regulatory elements across the human genome, their tissue-specific activities, and the set of genes they control has substantially improved due to a large number of conducted genomic, epigenomic, and transcriptomics studies. Main functional elements in the human genome, such as enhancers, promoters, and TF binding sites, have been extensively mapped by large international collaborations like ENCODE [17,18] and FANTOM5 [19,20]. Several dedicated resources have subsequently been developed, integrating and extending these datasets to generate a more detailed picture of regulatory elements [21–26]. Meanwhile, the application of novel computational [26–29] and high-throughput screening methods [30–33] has substantially improved our understanding of how regulatory elements control their respective target genes while several *in-silico* methods have been developed to better predict the impact of non-coding regulatory variants [34–40].

The increasing adoption of whole-genome (WGS) over whole-exome (WES) sequencing now allows for the comprehensive investigation of human variants in disease studies, including variants affecting these regulatory regions. The accurate identification, interpretation, and prioritization of disease-relevant variants from WGS studies requires standardized resources for their annotation in routine bioinformatics pipelines. Whilst there is a large variety of annotation methods and databases available for coding variants [41,42], resources for programmatic annotation of regulatory variants and their respective target gene(s) are still lacking. Ideally, such resources would include a catalog of regulatory regions and functional elements together with a set of impact prediction scores [40,43]. However, the resources and databases currently available in this field are often presented in a format not suitable to this task, and information about controlled gene(s) and tissue(s) of activity is difficult to access programmatically.

Here, we present a unified framework that can be used to process standard variant call format (VCF) files to generate a comprehensive annotation of non-coding variants. For this aim, we have created a comprehensive resource, entitled GREEN-DB (Genomic Regulatory Elements ENcyclopedia Database), integrating a collection of ~2.4M regulatory elements, additional functional elements (TFBS, DNase peaks, ultra-conserved non-coding elements (UCNE), and super-enhancers), and 7 non-coding impact prediction scores. Information on the controlled gene(s), tissue(s), and associated phenotype(s) are provided in GREEN-DB when possible and information is compiled in standard BED and SQLite (https://www.sqlite.org/) file formats.

Results

The GREEN-DB *database*

We have created a comprehensive collection of potential regulatory regions in the human genome including ~2.4M regions from 16 data sources covering ~1.5Gb evenly distributed across chromosomes (Figure 1A, B, and Supplementary Figure 1). A summary of the information represented in GREEN-DB is given in Table 1 while detailed region counts are summarized in Supplementary Table 1. As expected, these regions are mostly constituted by intronic and intergenic bases (Supplementary Figure 2) and overall they cover ~60% of introns and ~40% of intergenic space. However, a smaller but significant overlap was observed also with UTR and other exonic regions (Figure 1B, detailed in Supplementary Tables 2, 3). We have grouped regulatory regions into the following five categories 5 categories: bivalent (regions showing both activation and repression activity), enhancer, insulator, promoter, silencer; with enhancer and promoters representing the majority of regions (Figure 1C). Each region is described by its genomic location, region type, method(s) of detection, data source and closest gene; ~35% of regions are annotated with controlled gene(s), ~40% with tissue(s) of activity, and ~14% have associated phenotype(s) (Figure 1E). These data are organized in 6 distinct tables in an SQLite database allowing for rapid querying based on genomic interval(s) and/or gene(s) of interest (the database structure is described in Supplementary Results and depicted in Supplementary Figure 3). GREEN-DB regions are also provided as an extended bed file for easy integration into existing analysis pipelines.

We tested these regions for enrichment with genomic features associated with transcriptional activity using Fisher's exact test and found that several of them were significantly enriched, including transcription factors binding sites (TFBS, OR 9.67) and DNase hypersensitivity peaks (OR 13.13) from ENCODE, GTeX significant eQTLs (OR 2.93) and ultra-conserved non-coding elements (UCNE, OR 8.35); while they are depleted for difficult-to-address regions such as segmental duplication (SegDup, OR 0.45) and low-complexity regions (LCR, OR 0.22). Finally, GREEN-DB regions are also enriched for a curated set of non-coding disease-causing mutations (OR 2.05) (Figure 2A and Supplementary Table 4). We furthermore examined the distribution of PhyloP100 conservation values and ReMM prediction scores across GREEN-DB regions, compared to random regions with comparable size and distribution across the genome. GREEN-DB regions appeared more conserved than random regions, showing a larger proportion of bases with PhyloP100 scores above 1, 1.5, and 2 (p-value < 2.2E-16 for all comparisons, Mann–Whitney U test) (Figure 2B and Supplementary Figure 4, 5). Similarly, both per-region median and maximum values of the ReMM score are significantly higher for GREEN-DB regions compared to random regions (Figure 2C, D). The median value of per-region median ReMM score is 0.568 in GREEN-DB regions and 0.363 in random regions (p-value < 2.2E-16, Mann–Whitney U test), while the median value of per-region maximum ReMM score is 0.919 in GREEN-DB regions and 0.803 in random regions (p-value < 2.2E-16).

Gene regulatory space

Overall, ~58% of GREEN-DB regions have a putative association to one or multiple genes, either because these associations were determined experimentally (~35%), or because of a gene in close proximity (distance ≤ 10kb, ~23%), that can be suggested as a controlled gene (Figure 3A). Considering the 839,807 regions with a validated region-gene association, they interact with a total of 48,246 different genes, covering 67% of all genes and 97% of protein-coding genes from ENCODE v33 basic set. Controlled genes also cover 97, 98, and

100% of clinically relevant genes from PanelApp [44], ClinVar (pathogenic genes only), and ACMG actionable genes list, respectively (Supplementary Table 5). Distal control elements (silencers, enhancers, bivalent) are mostly located outside the controlled gene, balanced between up- or downstream locations, while most of the promoter regions are located inside genes or upstream of them (Figure 3B). As expected, the distance between a region and its controlled gene(s) is larger for enhancers and silencers, which appear to be mostly located from about 10kb up to several Mb away from their controlled gene (Figure 3C). When we analyzed the relationship between GREEN-DB regions and their controlled genes (taking only experimentally associated genes into account), we saw that the closest gene is among annotated controlled genes only for ~40% of enhancers and ~12% of silencers, while this proportion is much higher (~70%) for promoters. Even when the closest gene is controlled, it is the only associated gene in just 24% and 5% of cases for enhancers and silencers, respectively. Interestingly, even when considering only GREEN-DB regions located within a gene, this gene is among the controlled ones in less than 50% of cases for enhancers, silencer, and bivalent regions (Figure 3D).

The region-to-gene relationship showed a high degree of specificity, with most regions controlling less than 5 genes, while several genes are controlled by multiple regions (Supplementary Figure 6). Regions active in multiple tissues usually control more genes, suggesting a tissue-specific region-to-gene relationship (Supplementary Figure 7).

Finally, gene-set enrichment analysis performed on the 491 genes with an extremely large regulatory-space showed that these genes are strongly enriched for essential genes derived from mouse studies (p-value 7.88E-97, FDR 1.11E-91) as well as genes involved in developmental processes, cell differentiation and other essential biological functions (Supplementary Table 6).

Regions constrained against sequence variation

Based on data from gnomAD v3, we calculated a constraint metric for GREEN-DB regions ranging from 0 to 1, so that regions with higher values have lower than expected numbers of variants. We ranked GREEN-DB regions based on this metric and defined as constrained the 23,102 regions above the 99th percentile of the distribution (mostly enhancers and promoters, Supplementary Figure 8). Comparison with other regions in GREEN-DB showed that constrained regions are more conserved (Supplementary Figure 8C) and are significantly enriched for tissue- and gene-specific regions, namely regions active in a single tissue or controlling a single gene ($p < 2.2E-16$, Supplementary Figure 9). When comparing the maximum constraint value of associated GREEN-DB regions, genes in the ClinVar pathogenic and essential groups are controlled by regions with higher constraint value compared to other genes ($p < 2.2E-16$, Matt-Withney U test, Figure 4). The 89.2% of ClinVar pathogenic genes and 93.6% of essential genes from knock-out screenings are associated with a region above the 90th percentile of constraint (Figure 4C, D). Overall, constrained regions control 5,154 genes based on GREEN-DB annotations and these genes are strongly enriched for essential genes and genes bearing pathogenic variants in ClinVar (FDR 1.89E-222 and 1.22E-153, respectively). The complete results of our enrichment analysis are reported in Supplementary Table 7.

Evaluation of non-coding impact prediction scores

We considered 29 previously published prediction scores that can be applied to evaluate the impact of non-coding variants. Of these, 13 do not provide pre-computed values or were developed for somatic variants only and were thus removed from further analyses. Using the curated set of disease-causing, non-coding variants from [36], we evaluated the performance of the remaining 16 scores in classifying disease-causing variants. The GWAVA algorithm obtained the best results according to the OPM metric with values of 0.58 and 0.57 for 2 of the 3 GWAVA scores (Supplementary Table 8). However, available GWAVA pre-computed scores only cover 1.6 % of the genome, limiting its application in WGS annotation. NCBoost, FATHMM-MKL / -XF and ReMM also showed good classification performances (OPM values 0.449, 0.434, 0.427, 0.422 respectively). Finally, to maximize classification performance, genomic coverage of annotations and the diversity of computational approaches, we selected NCBoost, FATHMM-MKL and ReMM as the best scores combination. However, no single scores seemed able to robustly remove false-positive calls while maintaining high sensitivity (Supplementary Figure 10 and Supplementary Table 8). Indeed, when TPR is set to 0.9, the FDR is above 0.8 for all scores, while controlling the FDR ≤ 0.5 results in TPR values below 0.5 for all scores except NCBoost (0.53) and GWAVA version 1 (0.56). To assist the use of these scores in variant analyses, we also computed the score thresholds corresponding to TPR ≥ 0.9, FDR ≤ 0.5, and maximum ACC (detailed metrics for each threshold are shown in Supplementary Figure 11 and Supplementary Table 9).

Validation using non-coding, disease-associated variants

We applied GREEN-DB annotations to a set of 61 variants with previously demonstrated regulatory effects on disease genes (40 promoter and 21 enhancer variants associated with 17 different genes). Our annotations were able to capture all tested variants, linking all of them to the expected gene. When considering the 7 non-coding impact prediction scores evaluated in this paper, 97% (59/61) of tested variants were classified as “deleterious” by at least 1 score, and 51% (31/61) by at least 3 different scores when applying the calculated FDR50 thresholds. When applying the less stringent TPR90 threshold, almost all variants (97%) were classified as “deleterious” by at least 3 scores (Table 2). Details on each variant are reported in Supplementary Table 10.

A framework for annotation and prioritization of non-coding variants from WGS

We created a tool (GREEN-VARAN: Genomic Regulatory Elements ENcyclopedia VARiant ANnotation) for the annotation and prioritization of non-coding variants which integrates all the collected information: regulatory elements from GREEN-DB, non-coding impact prediction scores, and additional genomic features relevant to gene regulation (TFBS, UCNE, DNase peaks, super-enhancers and enhancer loss of function (LoF) predictions). The tool is written in Python 3 and processes the output of vcfanno [45] to produce variant annotations containing: regulatory region type and IDs, controlled gene(s), closest gene(s) with their distance, overlap with any of the additional genomic features. Pre-computed values from the tested non-coding prediction scores are distributed together with GREEN-DB and can be annotated using the tool. Finally, the tool allows the user to tag/filter variants based on genes of interest as well as select/tag non-coding variants only if they are associated with a gene already affected by a coding variant of a given impact (based on snpEff impact ranking). Given an annotated VCF or a list of region IDs, GREEN-VARAN allows querying GREEN-DB to retrieve additional details such as tissue of activity and data source. More details on the annotations generated by the tool are given in Supplementary Results.

Impact on WGS variant prioritization

We added our annotation framework to a standard pipeline applied to prioritize small variants from 90 WGS pedigrees to evaluate the impact of adding non-coding annotations to the number of possible candidate variants and genes. Considering variants identified in each individual, we found a median of ~2.29M variants in GREEN-DB regions, including a median of 44,837 rare (population AF < 0.01) and 8,472 rare deleterious variants (based on impact prediction scores) (Supplementary Figure 12). When looking at rare variants that segregate with the phenotype in each pedigree, adding GREEN-DB annotations increases the median number of candidate variants from 1,764 (exonic variants only) to 77,725 (filter step1 in Figure 5A). Filtering based on prediction scores reduces the median number of variants to 4,941 (4,792 considering only non-coding variants, filter step2 in Figure 5A). A significant proportion of prioritized non-coding variants affected genes with a potential role for the family phenotype, based on HPO-profile gene ranking (filter step3 in Figure 5A). The newly annotated non-coding variants have a particular impact on identifying compound heterozygote candidate variants and interestingly they also create new combinations with prioritized coding variants. Indeed, the median number of compound heterozygotes involving one protein-altering and one non-coding variant is 835, 27, and 2 in filtering steps 1, 2, and 3 respectively (Figure 5B). Adding non-coding annotations resulted in a larger number of candidate genes including genes likely relevant for the disease phenotype based on HPO profiles. The median number of candidate genes selected was 16 when considering coding variants only and 302 when including the new annotations (step3, Figure 5C). Similarly, when considering only a subset of clinically relevant genes from PanelApp or Clinvar, the median number of selected candidates increased from 7 and 7 (exonic variants only) to 107 and 115 (including GREEN-DB variants), respectively (Supplementary Figure 13). Ranges of selected candidate variants for each filtering step are reported in Supplementary Table 11.

Discussion

Non-coding regions of the genome have clearly been implicated in disease risk from a plethora of GWAS studies [14–16] and, more recently, various WGS studies have also highlighted the role of pathogenic rare variants in the non-coding space [8,10–12]. Whilst information about the types and locations of regulatory regions has been described previously in the literature [17,19–25], the systematic interrogation of these in clinical whole-genome sequencing data from patients with rare diseases remain challenging and limited by the lack of systematic resources easy to access programmatically [46–48]. To fill this gap, we have developed a framework for the systematic annotation of non-coding variants including an extensive catalog of regulatory regions and a set of tools and resources that can be integrated into routine bioinformatics pipelines to annotate non-coding variants and improve their interpretation and prioritization in disease studies.

We have collected and curated data from published, experimental, and computational sources to create a catalog providing a standardized representation for ~2.4 million regulatory elements in the human genome (GREEN-DB). To support the interpretation of the impact of genetic variants, each regulatory region is annotated with a rich set of information: 1) genomic location; 2) a standardized definition of its role (promoter, enhancer, silencer, bivalent, insulator); 3) its known controlled gene(s) and tissue(s) of activity; 4) its closest gene; 5) its potential phenotype association(s) based on GWAS datasets and Human Phenotype Ontology; 6) a constraint metric representing the tolerance of the region to genetic variation. An actual role of these regions as regulatory elements is supported by their significant overlap with motifs of recognized regulatory importance (like TFBS and DNAse hypersensitivity sites), as well as variants involved in gene expression regulation (GTeX eQTLs) and non-coding variants involved in human diseases. Moreover, they also score highly when compared to random regions of similar length considering ReMM prediction scores and PhyloP100 conservation values, suggesting that they are capturing a functionally relevant portion of the genome.

To interpret the biological role of a regulatory region, it is essential to know the genes it controls and in which tissues it is active. In GREEN-DB we collected and curated experimentally validated region-gene links and tissue information for ~35% and ~40% of the regions, respectively. Overall, GREEN-DB provides regulatory information for 48,246 genes, including most of the clinically relevant genes from PanelApp, ClinVar and ACMG, supporting its usefulness in human disease research. Although it has long been recognized that there is some degree of spatial relationship between regulatory regions such as promoters and enhancers and the genes they control, with promoter elements being closer and silencers more distal to their dependent genes [18,19,49], our analysis confirms the complexity of the relationship between regulatory regions and controlled genes that can not easily be explained by spatial proximity in the (linear) genome as previously demonstrated, e.g., by high-throughput studies of chromatin interactions [32,50–52].

Indeed for silencer and enhancer elements, the controlled gene was the closest gene in only 5% and 24% of cases respectively, whilst regulatory regions within a gene only exert regulatory control on that specific gene in less than half of the cases. Even if we cannot exclude that these observations may be influenced by incomplete annotation of controlled genes, this has considerable implications especially for GWAS studies, where the search for disease-associated genes often starts with proximity to the most significantly associated SNPs [53,54]. Overall, we observed a high degree of specificity in the region-gene relationship, with a large fraction of regions controlling less than 5 genes, even if this result may be affected by incomplete annotations of the controlled genes. On the other hand, most genes are controlled by multiple regulatory regions which in consequence means that very different, spatially distant genomic regions may have a similar phenotypic impact. This makes the comprehensive annotation of all regions that influence/regulate the normal activity of a gene so important for understanding the consequences of genomic variants of the respective gene function. A correlation emerged between the number of controlled genes and the number of active tissues for each region, confirming the tissue-specific nature of gene regulation and supporting the idea that alterations in a regulatory region can have different impacts in different tissues [12]. Consistent with other studies [55], we found that essential, ubiquitously expressed housekeeping genes and genes involved in human diseases had a larger regulatory space, namely a higher number of associated regulatory regions, that can contribute to fine-tune their expression and increase their tolerance to single disruptive mutations in one of the associated regions.

We also integrated information from GWAS studies and HPO databases to provide a possible associated phenotype for ~15% of the regions. This resource will be useful for the interpretation of new variants found in the regulatory regions, providing hypotheses on their potential biological impact. The fact that only a limited number of regions has an associated phenotype, despite a large number of GWAS hits available [56–58], can be explained by several reasons. In some cases, the phenotypic effect of alterations in a single regulatory region may be small due to the redundancy of these control regions and their tissue-specific effect, resulting in weak associations and thus reduced the significance of SNPs from GWAS studies. On the other hand, this also underlines how the impact of rare disrupting variants in the non-coding space is largely unexplored and how resources like GREEN-DB can inform our understanding of human diseases.

Using data from gnomAD v3 [59] we calculated a constraint metric that reflects the tolerance of each region to sequence variations. The maximum constraint value for regions controlling essential genes and genes involved in human diseases is significantly higher compared to other genes, suggesting that this metric can be used effectively to prioritize regions more relevant in disease studies. This idea is further supported by the analysis of regions under strong constraint (constraint value ≥ 0.99) that were more conserved than other regions in the database and associated with genes strongly enriched for essential genes and genes involved in human diseases.

To further assist the interpretation of variants located in regulatory regions, we collected pre-computed values from 16 different impact prediction algorithms and compared their ability to classify a curated set of established disease-causing non-coding variants. Overall, we must take into account that such comparisons are (i) limited by the nature of the known variants collected so far, which are mostly variants near to the affected gene and poorly captured distant regulatory elements [48]; and (ii) by the potential overlap of the test variants with the training sets used by each algorithm, which are often unknown. Based on OPM value (a metric developed to better summarize classification performances [60]), GWAVA [61], NCBoost [36], FATHMM [62,63] and ReMM [38] algorithms emerged as the best performing scores (OPM values: GWAVA_1 0.584, GWAVA_2 0.576, NCBoost 0.449, FATHMM-MKL 0.434, FATHMM-XF 0.427, ReMM 0.422; Supplementary Table 8), probably reflecting their specific training on disease-associated variants and the integration of functional region annotations. On the other hand, the poor performance of FIRE (OPM: 0.175) can be explained considering that this prediction model has been trained on eQTLs [37], which represent a completely different type of regulatory variants. Based on ROC curve analysis, we also provided suggested thresholds for variant classification, based on the desired level of sensitivity and FDR, which can be useful in prioritizing high-impact non-coding variants.

The combination of the information present in GREEN-DB with these prediction scores can effectively capture variants involved in human diseases, as shown by our ability to recapitulate known disease-associated variants from the literature. Indeed, considering a collection of 61 variants from 3 different publications [64–66], representing both close and distant regulatory variants, our annotations allowed us to associate them with the correct controlled gene and classify them as “deleterious” considering a stringent threshold for one (52/61) or multiple (46/61) of the 10 best-performing prediction scores. Specifically, the proposed combination of NCBoost, FATHMM-MKL and ReMM was able to correctly classify as “deleterious” 42 (69%) and 60 (98%) variants using the stringent FDR50 and the more relaxed TPR90 thresholds, respectively.

When it comes to the analysis of non-coding variants, the large number of such variants present in each person’s genome and the infancy of any robust clinical annotation means the application of WGS for rare disease patients diagnosis still presents a considerable challenge [40,67,68]. Therefore, whilst most clinical diagnostic labs now utilise whole exome sequencing, few have yet transitioned to whole genome sequencing [69,70]. Nonetheless, the diagnostic yield for such WES tests still rarely attains the 50% mark [71–73] indicating that the non-coding genome is likely to harbor many variants of clinical diagnostic significance. Even dedicated clinical WGS programmes such as the UK's 100,000 Genomes Project [74] do not routinely interrogate non-coding regulatory regions in their patient genomes, while others only take into account large variants (i.e. deletions) associated to a limited list of diagnostic-grade genes [75].

Our companion tool (GREEN-VARAN) brings together in a single annotation framework information from GREEN-DB, non-coding impact prediction scores and population AF annotations, creating a system suitable for systematic WGS variants annotation. The potential of the tool in prioritizing relevant variants in rare-diseases is shown by the results obtained when applied to an internal dataset of 90 WGS family cases. The median number of rare variants segregating with the phenotype per case increased from 1,764 when considering exome variants alone to 77,725 when including genome-wide variants overlapping GREEN-DB regions. Using prediction scores as a filter reduced this number to ~5000 and applying an HPO-based prioritization strategy based on the family phenotype further reduced the candidate genes to a median of 302 per case (compared to 16 when considering exonic variants only) and this number can be further reduced by selecting only compound heterozygous involving at least one exonic variant. The same trend applies when this pipeline is applied considering only clinical relevant genes from PanelApp or ClinVar suggesting that inclusion of our non-coding annotation can reveal previously ignored candidate genes likely to have an impact on patient phenotype. Whilst this number of candidate genes is still too many for a diagnostic lab to consider, this is certainly in the realms of the possible for research-based inspection, especially in otherwise difficult to solve cases. The application of GREEN-VARAN annotations can have a particular impact on the analysis of compound heterozygous variants and be useful in identifying second non-coding hits in biallelic candidate genes where only a single coding variant has been identified, an approach that already resulted in increased diagnosis in a recent large clinical WGS study [75]. In summary therefore, we have compiled an extensive and highly curated dataset of regulatory regions (GREEN-DB) and a tool for readily annotating regulatory variants from whole genome sequencing data (GREEN-VARAN). We expect that these resources will be of particular value for clinical scientists and researchers working with WGS data, supporting the identification of pathogenic rare disease variants in the non-coding space. Information collected in GREEN-DB provides a valuable resource to understand the complex spatial relationships between regulatory elements and their controlled genes, which will be useful for rare disease diagnosis as well as for the interpretation of common disease variants from GWAS.

Conclusion

We have developed a framework for the annotation and prioritization of regulatory variants in WGS analysis supporting the discovery of candidate disease-associated variants in the non-coding regions. This includes an extensive collection of regulatory regions, information on the controlled genes and a companion tool that easily integrates into existing bioinformatic pipelines to readily annotate non-coding variants from WGS data. The resources presented here therefore represents a significant advance for clinical diagnostics labs and researchers engaged in analyzing patient genomes.

Methods

Data collection

To compile an up-to-date, standardized collection of regulatory elements in the human genome (GREEN-DB) we collected and aggregated information from 16 different sources, including 7 previously published curated databases, 6 experimental datasets from recently published articles, and predicted regulatory regions from 3 different algorithms. Four additional datasets were included to integrate region to gene/phenotype relationships. The full list of data sources and references is reported in Supplementary Table 12. We also collected additional data useful in evaluating the regulatory role of genomic regions, including TFBS and DNase peaks, ultraconserved non-coding elements (UCNE), super-enhancer definitions, and enhancer LoF tolerance (Supplementary Table 13) as well as 9 scores developed to predict the regulatory impact of non-coding variants (Supplementary Table 14).

Data processing

Data collected from the various data sources were processed to generate a standardized collection of regulatory regions, their controlled gene(s), method(s) of detection, tissue(s) of activity, and associated phenotype(s). In the standardized tables, each region is represented by its genomic coordinates and annotated with a standard region type (bivalent, insulator, promoter, enhancer, silencer), closest gene(s) information, and a unique ID, used to link the region with additional annotations. Standardized regions and annotations from each data source were integrated into an SQLite database with 6 global tables (GRCh37 / GRCh38 regions, genes, tissues, methods, phenotypes), with the region table converted to GRCh38 coordinates using UCSC LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Regions provided in the additional datasets of TFBS, DNase clusters, super-enhancer, UCNE, and enhancer LoF tolerance were processed similarly and the overlaps between these additional regions and GREEN-DB regions were pre-computed and stored in ID-to-ID link tables. The detailed description of data processing steps is provided in the Supplementary Methods.

Evaluation of collected regulatory regions

To evaluate the ability of GREEN-DB to capture regions relevant for expression regulation, we evaluated the overlap between the collected regions and 5 established sets of regions associated with functional genome elements: ENCODE TFBS, ENCODE DNase hyper-sensitivity clusters, UCNE regions, and a curated set of non-coding disease-associated variants (from [36]). We used Fisher's exact test to assess the enrichment/depletion of each of these classes among the GREEN-DB regions. Since most of the GREEN-DB regions are located in the intergenic space, we verified the overlap with potentially uninformative repetitive regions performing the same test also for genome low-complexity regions (as defined in [76]) and segmental duplications. Finally, we evaluated the degree of conservation and the ReMM score distribution for GREEN-DB regions compared to random regions across the genome. We used the ReMM score, which has emerged as the best performing non-coding impact prediction score with genome-wide coverage from our analysis, to assess the actual ability of GREEN-DB to capture disease-relevant genomic regions. First, we generated a set of control regions by randomly picking from each chromosome (excluding centromeric and telomeric regions) the same number of regions seen in GREEN-DB, with comparable size distribution (Supplementary Figure 14). For each region in the random and GREEN-DB sets, we calculated the fraction of bases having a PhyloP100 score above 1, 1.5, and 2 (higher values indicate more conservation) and the median and maximum ReMM values. We compared the distributions of these values between control and GREEN-DB regions using the Mann–Whitney U test.

Analysis of gene regulatory space

We evaluated the relationship between regions and associated genes for the 839,807 regions for which we have collected validated associations. Based on gene definitions from the GENCODE v33 basic set, we investigated where each region was located with respect to each of its associated genes (upstream, downstream, or inside the gene) and the region-gene distances for each of 4 main region types having associated genes (bivalent, enhancer, promoter, silencer). Finally, we evaluated the proportion of regions for which the closest gene was among the associated genes or was the only controlled gene, and the proportion of regions located within a gene, but controlling other distant ones. For the 48,246 genes captured in the GREEN-DB, we calculated the number of associated genes per region (GxR) and the number of associated regions per gene (RxG). We correlated GxR value with the number of tissues per region to assess if regions controlling multiple genes are more likely to do so in a tissue-specific manner. Correlation significance was tested using Spearman's correlation test. We then selected 491 genes with an extremely large regulatory-space, defined as those in the 99th percentile of RxG distribution (genes with at least 182 associated regions), and used the hypergeometric test to assess their enrichment across Gene Ontology groups and canonical pathways from MSigDB v7.1 as well as essential genes derived from cell-culture (283) or mouse knock-outs (2,454) and genes bearing any pathogenic/likely pathogenic mutation in ClinVar (4,588). Essential genes lists were obtained from https://github.com/macarthur-lab/gene_lists (core_essential and mgi_essential lists). FDR of the 13,960 performed tests was controlled using the Benjamini–Yekutieli method [77].

Identification of regions under variation constraint

Variants from gnomAD v3 WGS dataset were used to assess regions under variation constraint. For each region in GREEN-DB, we first calculated the number of gnomAD PASS variants. To prevent the detection of false-positive constrained regions due to partial inaccessibility, we filtered regulatory regions that overlap more than 50% with known segmental duplications (segdup) or low-complexity regions (LCR). We furthermore removed regions on chrY and chrM leaving us with 100,768 annotated regulatory regions. For all remaining regions, we computed the region's GC density (GC) as a proxy for the region's mutability owing to the spontaneous deamination of methylated cytosines. We then created a linear regression model with the number of variants as the dependent variable:

$$N_{var} = length + GC + segdup + LCR$$

N_{var}, GC density, and sequence length variables were transformed to approximate normality using Blom's transformation [78]. In the case of LCR and segdup, the majority of values were equal to 0 making the above transformation ineffective. Instead, we treated these two variables as binary by setting all non-zero values equal to 1. Each region's degree of constraint was measured on the basis of its distance from the resulting regression line. The residuals from the model were ranked from lowest to higher, and assigned a percentile such that regions with the lowest residual value are assigned the highest percentile, reflecting the highest predicted constraint (regions with fewer than expected variants). Regions above the 99th percentiles were considered as constrained regions. We used Fisher's exact test to assess if these regions were enriched for regions controlling a single gene or active in a single tissue (tissue- and gene-specific regions) compared to all regions present in GREEN-DB. For each gene present in GREEN-DB, we selected the highest constraint value across associated regions and used the Mann–Whitney U test to compare value distributions between genes belonging to the ClinVar pathogenic or essential genes groups and all the other genes reported in GREEN-DB. For genes controlled by constrained regions, we performed gene-set enrichment analysis (GSEA) across Gene Ontology groups, canonical pathways, essential genes, and ClinVar pathogenic genes as described above.

Evaluation of non-coding impact prediction scores

With the aim of providing a framework useful for variant prioritization, we evaluated the usability of 26 non-coding variant impact prediction scores when applied to WGS data analysis for rare diseases. Among these, we excluded: 10 scores because they do not provide pre-computed values, making them difficult to apply programmatically; 2 scores that were developed specifically for somatic variants; 1 score that provide only disease-specific predictions for a limited set of phenotypes (see Supplementary Table 14). Of the remaining 13 scores, GWAVA and EIGEN provide 3 and 2 different prediction values respectively, for a total of 16 predictors. We compared the performances of these 16 scores when applied to a set of known disease-causing non-coding variants. For this purpose, we used a set of curated disease-associated and neutral variants from [36], including 725 true positive examples and 7,250 negative examples.

For each score the evaluation was limited to the subset of scored variants (see Supplementary Table 8). Classification performances were evaluated in R using the ROCR package [79] and 3 suggested thresholds for classification were computed: (i) max_ACC: score value achieving maximum accuracy; (ii) TPR90: filtering value corresponding to TPR ≥ 90%; (ii) FDR50: filtering value able to control FDR ≤ 50% with the maximum TPR. For a better representation of the overall classification performances of each score we also computed the overall performance measure (OPM) as described in [60], that better captures the performance of a score when used for filtering purposes.

Application to known examples of disease-causing regulatory variants

To test the potential of GREEN-DB and the proposed annotations to capture non-coding variants relevant in human diseases, we evaluated a set of 61 variants that have been described to have a regulatory effect on disease genes [64–66]. This set includes 40 promoter and 21 enhancer variants associated with 17 different genes. For each variant, we evaluated the overlap with GREEN-DB regions and other functional regions (TFBS, DNase, UCNE, dbSuper) and checked whether the affected gene from the original publication is among the ones reported in our database as controlled by the regions overlapping with the variant. Additionally, we assessed if these variants can be classified as "deleterious" based on the FDR50 thresholds we computed for the non-coding impact prediction scores.

Preparation of the WGS test dataset

To test the impact of our new annotations on the variant prioritization for rare diseases, we applied them to a set of 90 family cases from an internal WGS cohort (8 duos, 66 trios, 12 quads, 4 quintets). For each case, a ranked list of genes potentially relevant for the family phenotype was calculated based on the respective HPO profile using GADO [80]. WGS was performed at a minimum 30X mean coverage, reads aligned to GRCh38 using bwa v0.7.15 [81], and duplicated reads marked using samblaster v0.1.24 [82]. Small variants were identified from single individual BAM files using deepvariant v0.9.0 [83] and single individual gVCF were merged in a single cohort VCF using GLnexus v1.2.6 [84] with deepvariantWGS optimized settings. Variants were filtered retaining only variants with quality above 20 and at least 1 individual with GQ ≥ 20. The filtered VCF was annotated using SnpEFF v4.3 [85] and then our tool was used to integrate non-coding annotations. First, we computed the number of variants located in GREEN-DB regions in each individual considering all variants, only rare variants (gnomAD / 1000G global population AF < 0.01) or rare variants classified as deleterious applying the FDR50 thresholds by at least one of ReMM, NCBoost and FATHMM-MKL. Then, we evaluated the number of candidate disease-related variants in each pedigree applying a 3 steps prioritization: (i) rare variants (population AF < 0.01 in 1000G / gnomAD populations and AF < 0.1 in the cohort) segregating with the disease phenotype, located within a exon/splice-site or a GREEN-DB region with associated gene; (ii) potentially deleterious variants based on prediction scores: LoF variants, missense variants with CADD ≥ 20, non-coding variants in GREEN-DB regions with ReMM, NCBoost or FATHMM-MKL ≥ FDR50 thresholds (see Supplementary Table 10); (iii) affecting genes in the 90th percentile of GADO Z-score (representing genes more relevant to the disease based on the respective HPO profile). At each step, we evaluated the total number of variants, the number of coding variants, and the number of variants in GREEN-DB regions. Additionally, we computed the number of compound heterozygotes combinations involving variants in GREEN-DB regions and a combination of a protein altering variant (LoF or missense with CADD ≥ 20) variant plus a GREEN-DB region variant. The same procedure was repeated considering only genes with pathogenic/likely pathogenic annotation in ClinVar and genes reported in PanelApp disease panels.

Abbreviations

OR: Odds-ratio

TPR: True positive rate (sensitivity)

TNR: True negative rate (specificity)

FDR: False discovery rate

ACC: Accuracy

HPO: Human Phenotype Ontology

TFBS: Transcription factor binding site

UCNE: Ultra-conserved non-coding element

AUC: Area under the curve

OPM: Overall Performance Measure

Declarations

Ethics approval and consent to participate

Written informed consent for all participants was obtained under the Molecular Genetic and Analysis and Clinical Studies of Individuals and Families at Risk of Genetic Disease (MGAC) study protocol, approved by the West Midlands Research Ethics Committee, reference 13/WM/0466.

Consent for publication

Not applicable

Availability of data and materials

GREEN-VARAN toolset is available in the GitHub repository

https://github.com/edg1983/GREEN-VARAN

GREEN-DB is available in the Zenodo repository https://zenodo.org/record/3981033

Competing interests

The authors declare that they have no competing interests

Funding

This research was funded and supported by the Wellcome Trust and Department of Health as part of the Health Innovation Challenge Fund scheme (Wellcome Core Award Grant Number 203141/Z/16/Z). This work was also funded and supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC) and by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health or Wellcome Trust.

Author's contributions

EG: initial idea; conceived the project; collected and curated the datasets; compiled the database; developed the annotation tools; authored the paper.

NP: conceived the project; contributed ideas; tested the database; co-authored the paper.

JT: conceived the project; contributed ideas; tested the database; provided funding; co-authored the paper.

Acknowledgments

Authors acknowledge Dr. Dimitris Vavoulis for critical discussion on the constraint model analysis.

References

1. Gamazon ER, Segrè AV, van de Bunt M, Wen X, Xi HS, Hormozdiari F, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. Nat Genet. 2018;50:956–67.

2. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. Human genomics. The human transcriptome across tissues and individuals. Science. 2015;348:660–5.

3. Furlong EEM, Levine M. Developmental enhancers and chromosome topology. Science. 2018;361:1341–5.

4. Danino YM, Even D, Ideses D, Juven-Gershon T. The core promoter: At the heart of gene expression. Biochim Biophys Acta. 2015;1849:1116–31.

5. Schoenfelder S, Fraser P. Long-range enhancer-promoter contacts in gene expression control. Nat Rev Genet. 2019;20:437–55.

6. Zhang Y, Wong C-H, Birnbaum RY, Li G, Favaro R, Ngan CY, et al. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature. 2013;504:306–10.

7. Catarino RR, Stark A. Assessing sufficiency and necessity of enhancer activities for gene expression and the mechanisms of transcription activation. Genes Dev. 2018;32:202–23.

8. Zhang G, Shi J, Zhu S, Lan Y, Xu L, Yuan H, et al. DiseaseEnhancer: a resource of human disease-associated enhancer catalog. Nucleic Acids Res. 2018;46:D78–84.

9. GTEx Consortium, Laboratory DA &coordinating C (ldacc)—analysis WG, Statistical Methods groups—Analysis Working Group, Enhancing GTEx (eGTEx) groups, NIH Common Fund, Nih/nci, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.

10. Li X, Kim Y, Tsang EK, Davis JR, Damani FN, Chiang C, et al. The impact of rare variation on gene expression across tissues. Nature. 2017;550:239–43.

11. Zhang F, Lupski JR. Non-coding genetic variants in human disease. Hum Mol Genet. 2015;24:R102–10.

12. Spielmann M, Mundlos S. Looking beyond the genes: the role of non-coding variants in human disease. Hum Mol Genet. 2016;25:R157–65.

13. Castel SE, Cervera A, Mohammadi P, Aguet F, Reverter F, Wolman A, et al. Modified penetrance of coding variants by cis-regulatory variation contributes to disease risk. Nat Genet. 2018;50:1327–34.

14. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106:9362–7.

15. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90:7–24.

16. Corradin O, Scacheri PC. Enhancer variants: evaluating functions in common disease. Genome Med. 2014;6:85.

17. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

18. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583:699–710.

19. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507:455–61.

20. FANTOM Consortium and the RIKEN PMI and CLST (DGT), Forrest ARR, Kawaji H, Rehli M, Baillie JK, de Hoon MJL, et al. A promoter-level mammalian expression atlas. Nature. 2014;507:462–70.

21. Hait TA, Amar D, Shamir R, Elkon R. FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. Genome Biol. Genome Biology; 2018;19:56.

22. Wang J, Dai X, Berry LD, Cogan JD, Liu Q, Shyr Y. HACER: an atlas of human active enhancers to interpret regulatory variants. Nucleic Acids Res. Oxford University Press; 2019;47:D106–12.

23. Dreos R, Ambrosini G, Groux R, Cavin Périer R, Bucher P. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. Nucleic Acids Res. 2017;45:D51–5.

24. Visel A, Minovitsky S, Dubchak I, Pennacchio LA. VISTA Enhancer Browser--a database of tissue-specific human enhancers. Nucleic Acids Res. 2007;35:D88–92.

25. Moore JE, Pratt HE, Purcaro MJ, Weng Z. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. Genome Biol. 2020;21:17.

26. Wu Z, Ioannidis NM, Zou J. Predicting target genes of noncoding regulatory variants with ICE. Bioinformatics [Internet]. 2020; Available from: http://dx.doi.org/10.1093/bioinformatics/btaa254

27. Libbrecht MW, Rodriguez OL, Weng Z, Bilmes JA, Hoffman MM, Noble WS. A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types. Genome Biol. Genome Biology; 2019;20:180.

28. Li Y, Shi W, Wasserman WW. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinformatics. 2018;19:202.

29. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping

and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473:43–9.

30. Pang B, Snyder MP. Systematic identification of silencers in human cells. Nat Genet. 2020;52:254–63.

31. Gasperini M, Hill AJ, McFaline-Figueroa JL, Martin B, Kim S, Zhang MD, et al. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell. 2019;176:377–90.e19.

32. Jung I, Schmitt A, Diao Y, Lee AJ, Liu T, Yang D, et al. A compendium of promoter-centered long-range chromatin interactions in the human genome. Nat Genet. Springer US; 2019;51:1442–9.

33. Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-Contact model of enhancer specificity from thousands of CRISPR perturbations [Internet]. 2019 [cited 2020 Aug 5]. p. 529990. Available from: https://www.biorxiv.org/content/10.1101/529990v1

34. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet. 2018;50:1171–9.

35. Huang Y-F, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat Genet. Nature Publishing Group; 2017;49:618–24.

36. Caron B, Luo Y, Rausell A. NCBoost classifies pathogenic non-coding variants in Mendelian diseases through supervised learning on purifying selection signals in humans. Genome Biol. 2019;20:32.

37. Ioannidis NM, Davis JR, DeGorter MK, Larson NB, McDonnell SK, French AJ, et al. FIRE: functional inference of genetic variants that regulate gene expression. Bioinformatics. 2017;33:3895–901.

38. Smedley D, Schubach M, Jacobsen JOB, Köhler S, Zemojtel T, Spielmann M, et al. A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. Am J Hum Genet. 2016;99:595–606.

39. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12:931–4.

40. Lee PH, Lee C, Li X, Wee B, Dwivedi T, Daly M. Principles and methods of in-silico prioritization of non-coding regulatory variants. Hum Genet. 2018;137:15–30.

41. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nat Rev Genet. 2017;18:599–612.

42. Butkiewicz M, Bush WS. In Silico Functional Annotation of Genomic Variation. Curr Protoc Hum Genet. 2016;88:6.15.1–6.15.17.

43. Worthey EA. Analysis and annotation of whole-genome or whole-exome

sequencing-derived variants for clinical diagnosis. Curr Protoc Hum Genet. 2013;79:Unit 9.24.

44. Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nat Genet. 2019;51:1560–5.

45. Pedersen BS, Layer RM, Quinlan AR. Vcfanno: fast, flexible annotation of genetic variants. Genome Biol. 2016;17:118.

46. Gloss BS, Dinger ME. Realizing the significance of noncoding functionality in clinical genomics. Exp Mol Med. 2018;50:97.

47. Zhu Y, Tazearslan C, Suh Y. Challenges and progress in interpretation of non-coding genetic variants associated with human disease. Exp Biol Med . 2017;242:1325–34.

48. French JD, Edwards SL. The Role of Noncoding Variants in Heritable Disease. Trends Genet [Internet]. 2020; Available from: http://dx.doi.org/10.1016/j.tig.2020.07.004

49. Narlikar L, Ovcharenko I. Identifying regulatory elements in eukaryotic genomes. Brief Funct Genomic Proteomic. 2009;8:215–30.

50. Ulianov SV, Gavrilov AA, Razin SV. Nuclear compartments, genome folding, and enhancer-promoter communication. Int Rev Cell Mol Biol. 2015;315:183–244.

51. Mishra A, Hawkins RD. Three-dimensional genome architecture and emerging technologies: looping in disease. Genome Med. 2017;9:87.

52. Jerković I, Szabo Q, Bantignies F, Cavalli G. Higher-Order Chromosomal Structures Mediate Genome Function. J Mol Biol. 2020;432:676–81.

53. Brodie A, Azaria JR, Ofran Y. How far from the SNP may the causative genes be? Nucleic Acids Res. 2016;44:6046–54.

54. Cano-Gamez E, Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Front Genet. 2020;11:424.

55. Wang X, Goldstein DB. Enhancer Domains Predict Gene Pathogenicity and Inform Gene Discovery in Complex Disease. Am J Hum Genet. 2020;106:215–33.

56. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101:5–22.

57. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47:D1005–12.

58. Leslie R, O’Donnell CJ, Johnson AD. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics. 2014;30:i185–94.

59. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581:434–43.

60. Niroula A, Urolagin S, Vihinen M. PON-P2: prediction method for fast and reliable identification of harmful variants. PLoS One. 2015;10:e0117380.

61. Ritchie GRS, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nat Methods. 2014;11:294–6.

62. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31:1536–43.

63. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. 2018;34:511–3.

64. Brandt M, Kim-Hellmuth S, Ziosi M, Gokden A, Wolman A, Lam N, et al. An autoimmune disease risk variant has a trans master regulatory effect mediated by IRF1 under immune stimulation [Internet]. 2020 [cited 2020 Aug 6]. p. 2020.02.21.959734. Available from: https://www.biorxiv.org/content/10.1101/2020.02.21.959734v1

65. Kircher M, Xiong C, Martin B, Schubach M, Inoue F, Bell RJA, et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. Nat Commun. 2019;10:3583.

66. Lecerf L, Kavo A, Ruiz-Ferrer M, Baral V, Watanabe Y, Chaoui A, et al. An impairment of long distance SOX10 regulatory elements underlies isolated Hirschsprung disease. Hum Mutat. 2014;35:303–7.

67. Lappalainen T, Scott AJ, Brandt M, Hall IM. Genomic Analysis in the Age of Human Genome Sequencing. Cell. 2019;177:70–84.

68. Posey JE. Genome sequencing and implications for rare disorders. Orphanet J Rare Dis. 2019;14:153.

69. Soden SE, Saunders CJ, Willig LK, Farrow EG, Smith LD, Petrikin JE, et al. Effectiveness of exome and genome sequencing guided by acuity of illness for diagnosis of neurodevelopmental disorders. Sci Transl Med. 2014;6:265ra168.

70. Bick D, Fraser PC, Gutzeit MF, Harris JM, Hambuch TM, Helbling DC, et al. Successful Application of Whole Genome Sequencing in a Medical Genetics Clinic. J Pediatr Genet. 2017;6:61–76.

71. Ferretti L, Mellis R, Chitty LS. Update on the use of exome sequencing in the diagnosis of fetal abnormalities. Eur J Med Genet. 2019;62:103663.

72. Mone F, Eberhardt RY, Morris RK, Hurles ME, Mcmullan DJ, Maher ER, et al. COngenital heart disease and the Diagnostic yield with Exome sequencing (CODE Study): prospective cohort study and systematic review. Ultrasound Obstet Gynecol [Internet]. 2020;

Available from: http://dx.doi.org/10.1002/uog.22072

73. Smith HS, Swint JM, Lalani SR, Yamal J-M, de Oliveira Otto MC, Castellanos S, et al. Clinical Application of Genome and Exome Sequencing as a Diagnostic Tool for Pediatric Patients: a Scoping Review of the Literature. Genet Med. 2019;21:3–16.

74. Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ. 2018;361:k1687.

75. Turro E, Astle WJ, Megy K, Gräf S, Greene D, Shamardina O, et al. Whole-genome sequencing of patients with rare diseases in a national health system. Nature. 2020;583:96–102.

76. Li H. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30:2843–51.

77. Benjamini Y, Yekutieli D. The Control of the False Discovery Rate in Multiple Testing under Dependency. Ann Stat. Institute of Mathematical Statistics; 2001;29:1165–88.

78. Blom G. Statistical estimates and transformed beta-variables [Internet]. Almqvist & Wiksell; 1958 [cited 2020 Aug 6]. Available from: https://www.diva-portal.org/smash/record.jsf?pid=diva2:516729

79. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21:3940–1.

80. Deelen P, van Dam S, Herkert JC, Karjalainen JM, Brugge H, Abbott KM, et al. Improving the diagnostic yield of exome- sequencing by predicting gene-phenotype associations using large-scale gene expression analysis. Nat Commun. 2019;10:2837.

81. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

82. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 2014;30:2503–5.

83. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36:983–7.

84. Yun T, Li H, Chang P-C, Lin MF, Carroll A, McLean CY. Accurate, scalable cohort variant calls using DeepVariant and GLnexus [Internet]. 2020 [cited 2020 Aug 6]. p. 2020.02.10.942086. Available from: https://www.biorxiv.org/content/10.1101/2020.02.10.942086v2

85. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly . 2012;6:80–92.

Tables

Table 1. Summary of GREEN-DB information

GREEN-DB	No. of Elements	Mean size (bp)	Bases covered
Enhancer	1,832,830	1,111	1,449,153,178
Promoter	565,323	580	234,315,553
Silencer	4,302	208	11,210,309
Bivalent	8,409	1,348	894,792
Insulator	23	741	17,504
All regions	2,410,887	988	1,502,180,018
With controlled gene(s)	839,511		
With tissue information	941,874		
With phenotype information	349,008		

The table summarizes the main statistic about regions in GREEN-DB, reporting counts and number of genomic basis covered

Table 2. Validation of GREEN-DB annotations using published non-coding disease-associated variants

				Above FDR50 threshold			Above TPR90 threshold		
Variant	**N**	**In GREEN-DB**	**Expected gene**	**≥ 1 score**	**≥ 2 scores**	**≥ 3 scores**	**≥ 1 score**	**≥ 2 scores**	**≥ 3 scores**
All	61	61	61	52 (42)	46 (30)	24 (14)	61 (60)	61 (55)	60 (32)
Enhancer	21	21	21	14 (13)	14 (11)	10 (6)	21 (20)	21 (19)	20 (9)
Promoter	40	40	40	38 (29)	32 (19)	24 (8)	40 (40)	40 (36)	40 (23)

The table reports results obtained applying GREEN-DB annotations and prediction scores to a set of known non-coding disease-associated variants collected from the literature. The number of variants captured by GREEN-DB regions (In GREEN-DB) and with the expected genes among controlled genes from the database (Expected gene) is reported. Additionally, the table reports the number of variants above the FDR50/TPR90 threshold considering at least 1, 2, or 3 scores among the top 10 non-coding impact prediction scores or, in brackets, the 3 scores selected as the best combination (NCBoost, FATHMM-MKL, ReMM).

Figure Legends

Figure 1. Summary statistic of regions collected in the GREEN-DB

(A) GREEN-DB collects human regulatory regions from 16 different sources including curated databases, experimental assays, and computational predictions. (B) Number of bases captured by these regions across different genomic locations and covered fraction of each genomic location (label on top of bars). (C) GREEN-DB contains bivalent, enhancer, insulator, promoter, and silencer regions with sizes mostly between 100 and 1000 bp (D). (E) Fraction of regions with associated gene, phenotype and tissue information. Phenotype information was derived from GWAS studies (via overlap of significant SNPs with GREEN-DB regions), HPO (via controlled genes), and DiseaseEnhancer dataset.

Figure 2. Evaluation of the GREEN-DB regions

(A) Using Fisher's exact test, we assessed the presence of an enriched overlap between regions collected in the GREEN-DB and several genomic features involved in transcriptional activity: DNase HS peaks (Dnase) and transcription factors binding sites (TFBS) from ENCODE, ultraconserved non-coding elements (UCNE), and significant eQTLs from GTeX v8 (GteX eQTLs). We also tested if GREEN-DB regions were enriched for a curated set of disease-causing non-coding variants (TrueSet Vars) and a set of regions difficult to sequence such as segmental duplications (SegDup) and low-complexity regions (LCR). (B) Considering the PhyloP100 conservation value distribution for each region, GREEN-DB regions have a higher proportion of highly conserved bases per region at either 1, 1.5, or 2 thresholds compared to random regions. GREEN-DB regions also showed higher per region median (C) and maximum (D) ReMM score compared to a set of random regions with comparable size and distribution across the genome. Triple stars indicates p-value < 0.001

Figure 3. Gene regulatory space

(A) Overall 839,807 regions (~35%) in GREEN-DB are experimentally associated with a controlled gene, while another 23% have a gene in close proximity. Considering only experimental associations, distant control elements are mostly located up- or downstream of a gene, with a smaller proportion observed within genes (B). The distance between a region and its controlled gene(s) is larger for enhancers, silencers, and bivalent, with most regions located between 10kb and several Mb away from the controlled gene (C). Interestingly, a large proportion of these regions may not control their closest gene(s) even when they are located within a specific transcript (D).

Figure 4. Constraint regions control diseases-associated and essential genes

For each gene reported in GREEN-DB, we considered the maximum constraint value across the associated regions and compared the distribution of these values between general genes and genes in the ClinVar pathogenic (A) or essential genes groups (B). Both groups appear to be controlled by regions with higher constraint. For various constraint value tranches, we calculated the fraction of ClinVar (C) or essential (D) genes controlled by at least one region in the corresponding tranche. Both groups show a large fraction of genes controlled by regions with constraint value ≥ 0.9. Triple stars indicates p-values < 0.001

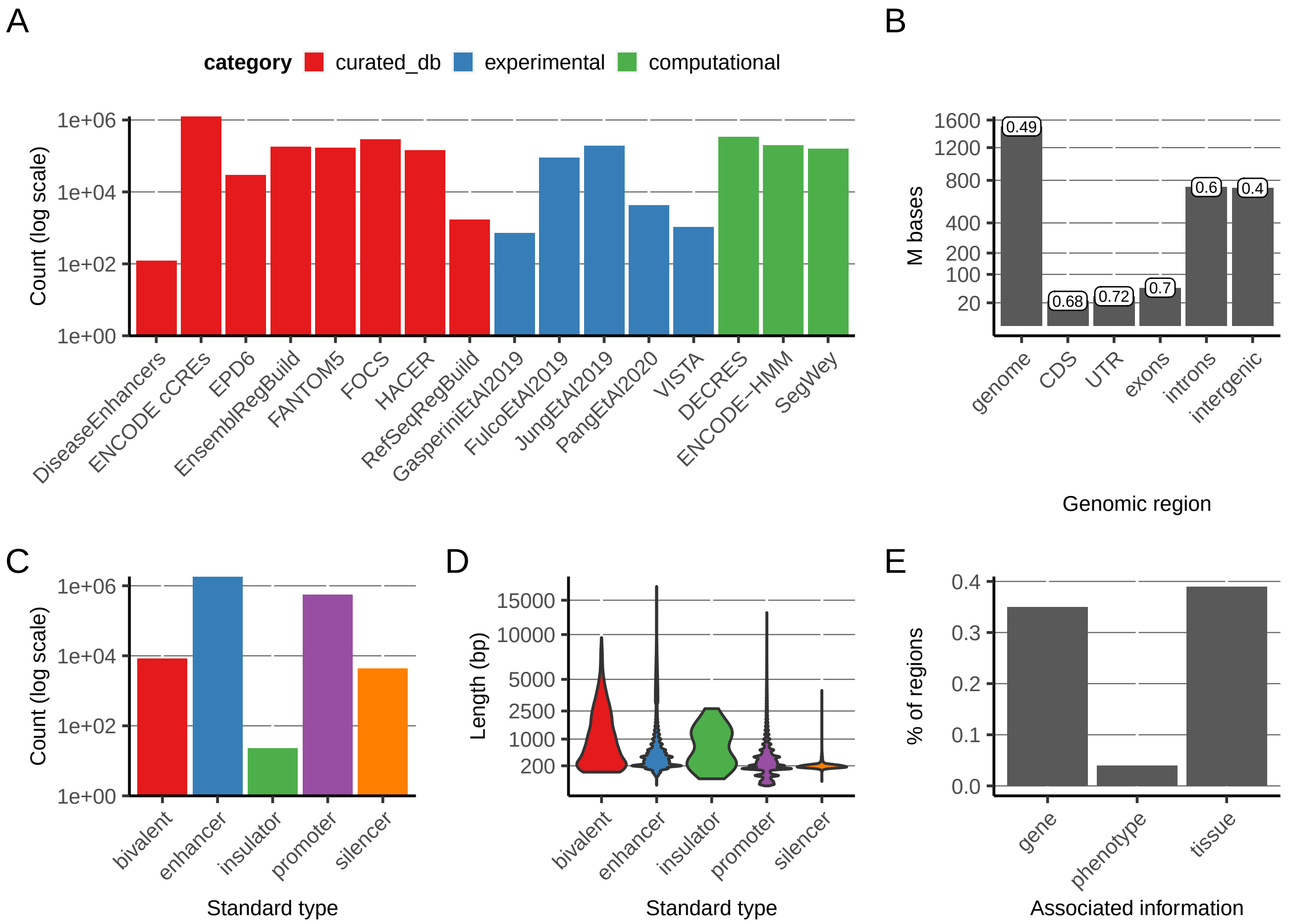
Figure 5. Impact of non-coding annotations on WGS variant prioritization

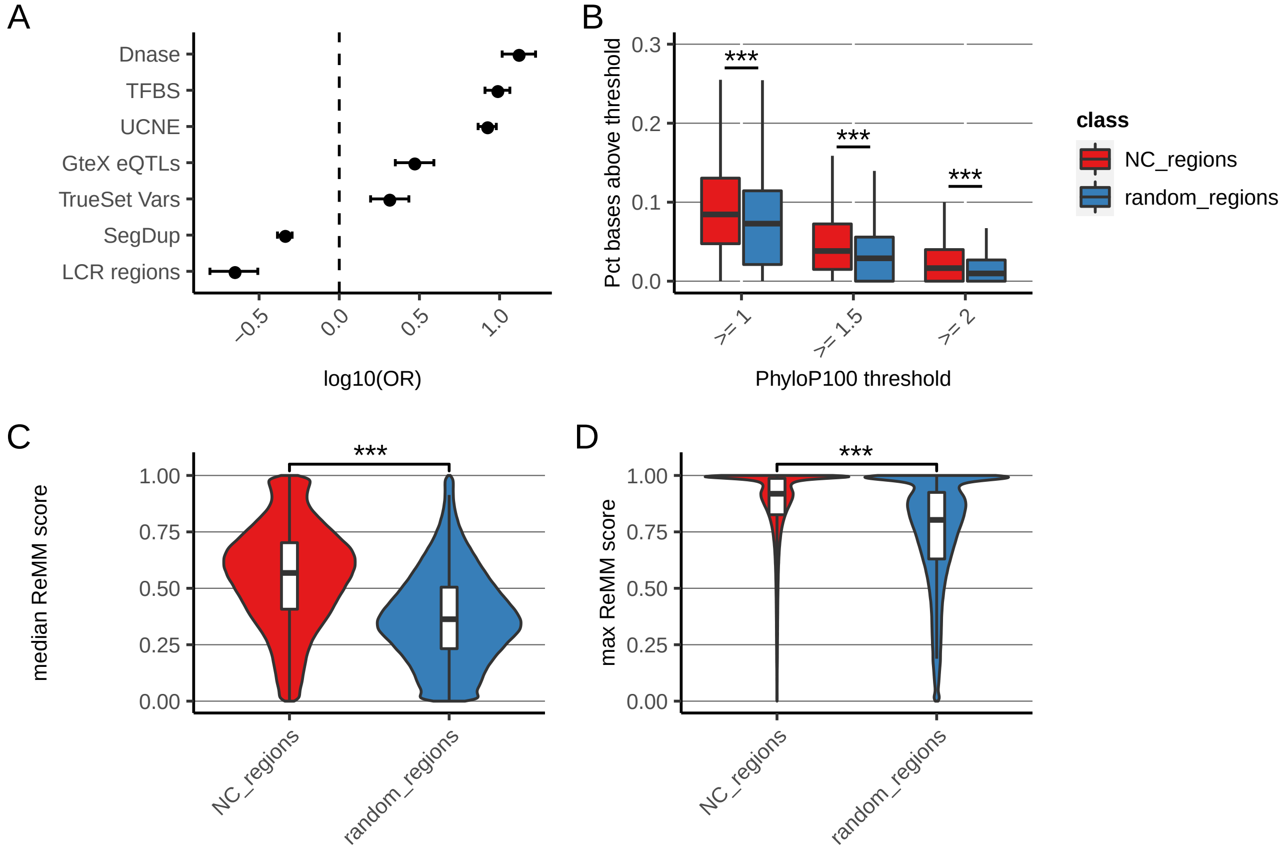
The violin plots represent the number of candidate variants (A), compound heterozygotes (B), and genes (C) found in 90 WGS pedigrees following the 3 step variant prioritization pipeline described in the main text. (A) and (C) report counts considering all, all exonic and all non-coding GREEN-DB variants (NC regions vars). In (B) the number of possible compound heterozygotes is reported considering any combination (All comphet), combinations including a non-coding variant annotated with GREEN-DB (Include NC var) and combinations including one non-coding variant and one protein-altering variant (LoF or missense with CADD ≥ 20) (Coding var+NC var).

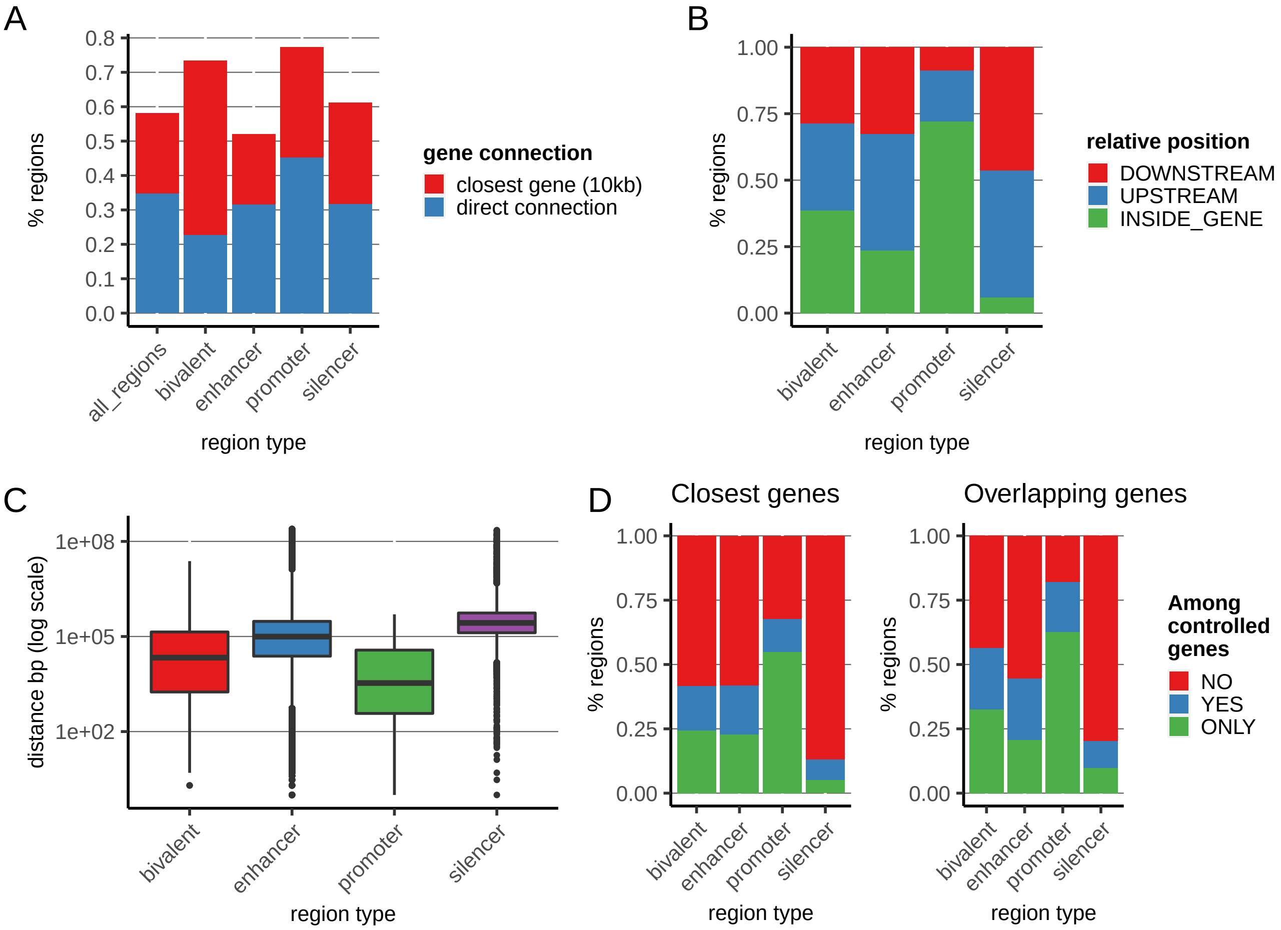
Additional files

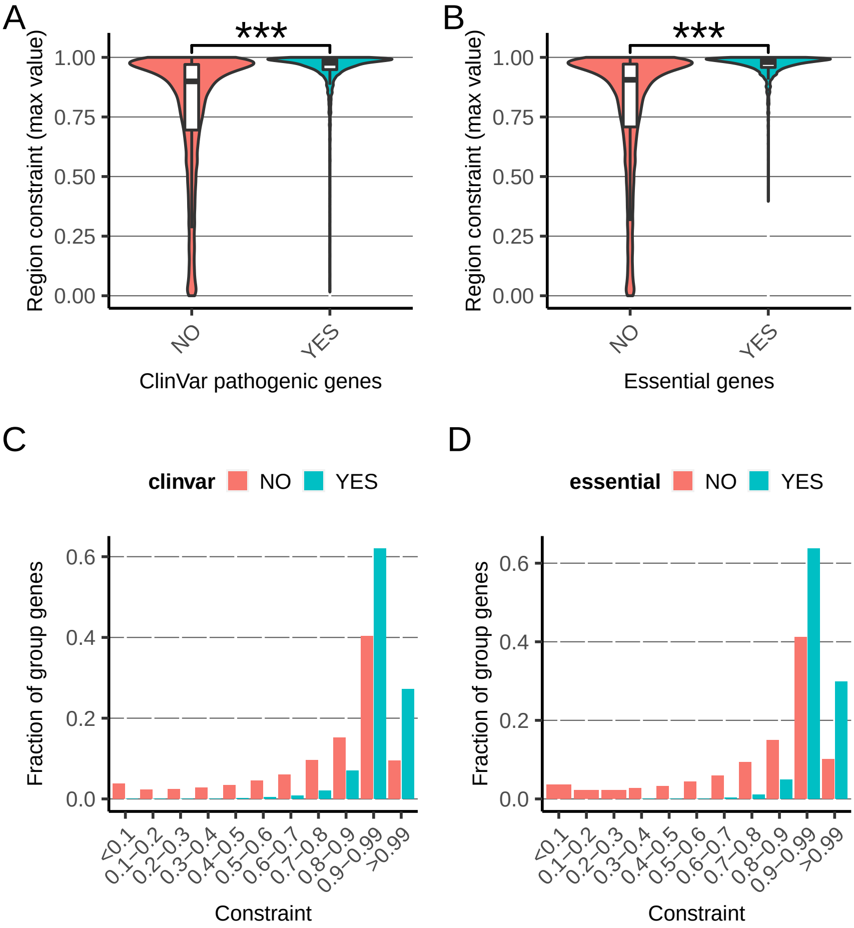
Additional File 1. Supplementary Tables 1-14 (.xls)

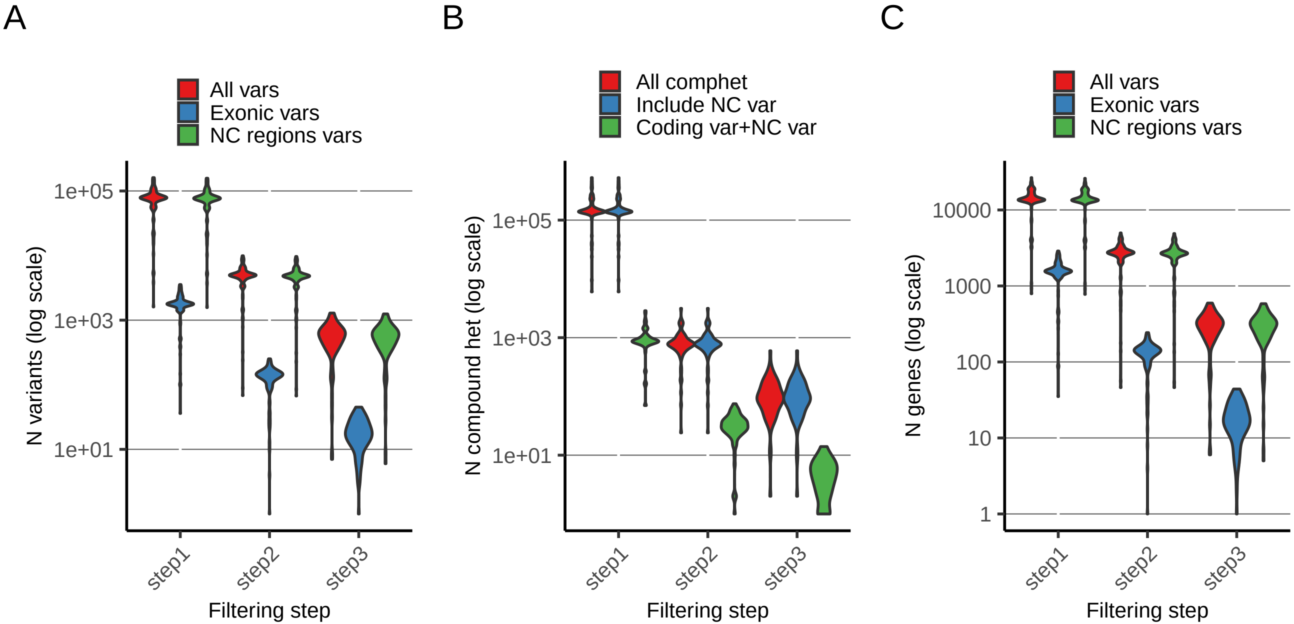
Additional File 2. Supplementary Methods and Results. Supplementary Figures 1-14. (.pdf)


A
category
curated_db
experimental
computational
Count (log scale)
1e+06
1e+04
1e+02
1e+00
DiseaseEnhancers
ENCODE cCREs
EPD6
EnsemblRegBuild
FANTOM5
FOCS
HACER
RefSeqRegBuild
GasperiniEtAl2019
FulcoEtAl2019
JungEtAl2019
PangEtAl2020
VISTA
DECRES
ENCODE−HMM
SegWey
B
M bases
1600
1200
800
400
200
100
20
0.49
0.68
0.72
0.7
0.6
0.4
genome
CDS
UTR
exons
introns
intergenic
Genomic region
C
Count (log scale)
1e+06
1e+04
1e+02
1e+00
bivalent
enhancer
insulator
promoter
silencer
Standard type
D
Length (bp)
15000
10000
5000
2500
1000
200
bivalent
enhancer
insulator
promoter
silencer
Standard type
E
% of regions
0.4
0.3
0.2
0.1
0.0
gene
phenotype
tissue
Associated information


A
Dnase
TFBS
UCNE
GteX eQTLs
TrueSet Vars
SegDup
LCR regions
−0.5
0.0
0.5
1.0
log10(OR)
B
Pct bases above threshold
0.3
0.2
0.1
0.0
***
***
***
>= 1
>= 1.5
>= 2
PhyloP100 threshold
class
NC_regions
random_regions
C
median ReMM score
1.00
0.75
0.50
0.25
0.00
***
NC_regions
random_regions
D
max ReMM score
1.00
0.75
0.50
0.25
0.00
***
NC_regions
random_regions


A
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
% regions
all_regions
bivalent
enhancer
promoter
silencer
region type
gene connection
closest gene (10kb)
direct connection
B
1.00
0.75
0.50
0.25
0.00
% regions
bivalent
enhancer
promoter
silencer
region type
relative position
DOWNSTREAM
UPSTREAM
INSIDE_GENE
C
1e+08
1e+05
1e+02
distance bp (log scale)
bivalent
enhancer
promoter
silencer
region type
D
Closest genes
Overlapping genes
1.00
0.75
0.50
0.25
0.00
% regions
bivalent
enhancer
promoter
silencer
region type
Among controlled genes
NO
YES
ONLY


A
***
Region constraint (max value)
1.00
0.75
0.50
0.25
0.00
NO
YES
ClinVar pathogenic genes
B
***
Region constraint (max value)
1.00
0.75
0.50
0.25
0.00
NO
YES
Essential genes
C
clinvar
NO
YES
Fraction of group genes
0.6
0.4
0.2
0.0
<0.1
0.1–0.2
0.2–0.3
0.3–0.4
0.4–0.5
0.5–0.6
0.6–0.7
0.7–0.8
0.8–0.9
0.9–0.99
>0.99
Constraint
D
essential
NO
YES
Fraction of group genes
0.6
0.4
0.2
0.0
<0.1
0.1–0.2
0.2–0.3
0.3–0.4
0.4–0.5
0.5–0.6
0.6–0.7
0.7–0.8
0.8–0.9
0.9–0.99
>0.99
Constraint


A
All vars
Exonic vars
NC regions vars
1e+05
1e+03
1e+01
N variants (log scale)
step1
step2
step3
Filtering step
B
All comphet
Include NC var
Coding var+NC var
1e+05
1e+03
1e+01
N compound het (log scale)
step1
step2
step3
Filtering step
C
All vars
Exonic vars
NC regions vars
10000
1000
100
10
1
N genes (log scale)
step1
step2
step3
Filtering step