A Framework to Analyse and Interpret Mouse Functional Genome by Prioritizing High Impact SNPs

Ahmed Arslan*

Department of Anaesthesia, Stanford University School of Medicine, 300 Pasteur Drive, Palo Alto CA 94305, The USA.
* To whom correspondence should be addressed. Email: aarslan@stanford.edu

ABSTRACT

The essential understanding of disease pathogenesis and enabling genetic findings to be used for developing new therapeutics, is missing in the identifications of genomic loci through whole genome association studies (GWAS). Here we describe a new computational method (mMap) that reduces this gap by characterizing the functional and regulatory impact of allelic variation. The method incorporates the precomputed annotations of 26 protein functional regions and eight regulatory regions and recover SNPs that fall/lie in these regions. After annotating SNPs to functional or regulatory data, method link them to biological functions and pathways, and predicts significantly disrupted biological regions, processes and pathways, by controlling false discovery through hypergeometric test. By doing so, the method limits data to human interpretation level by prioritizing SNPs that have the potential to mediate a biological phenotype. The method is applicable to procedures that rely on the understanding of the biological causal role of mouse SNPs and is available online. In two example mMap applications, including whole genomes SNPs data from 48 inbred mice strains, we identify biological mechanisms by which SNPs can regulate pathways to govern phenotypes by targeting different coding and regulatory regions, even in closely related strains.

INTRODUCTION

Genome wide association studies (GWAS) have identified many genetic factors that are associated with disease susceptibility; but we have difficulty determining how the identified SNP alleles contribute to disease susceptibility, which is critical for devising new therapies from genetic discoveries. Many SNPs identified through GWAS are located within regulatory regions, which makes it even harder to decipher the impact of allelic variation. Also, since many different SNPs can be associated with a disease phenotype in a GWAS, it can be difficult to select the true causative SNP from among the many whose allelic associations arise by chance. For all of these reasons, improved computational tools that can rapidly assess the potential impact of allelic variation on protein domain structure, post-translational modifications, protein-protein interactions, cellular signalling pathways and on promoter or enhancer function are needed. This type of computational tool could enable SNPs that are most likely to impact a trait response to be identified from among the many that are often identified in a GWAS. Moreover, this information could subsequently facilitate the translation of genetic data into actionable information that can be used to improve our understanding of disease pathogenesis and for developing new therapeutic approaches.

We previously developed an early version of this type of computational tool (yMap) for analysing the impact of allelic variation in yeast[1]. Mice are the premier model organism used for biomedical discovery; and the large number of available inbred strains has made them the ideal experimental organism for use in the genetic analysis of biomedical traits. Causative genetic factors affecting susceptibility to eye, metabolic and infectious diseases were identified when automated methods were used to filter the output of our haplotype-based computational genetic mapping (HBCGM) method[2]. The use of structured computational methods enabled true causative genetic factors to be uncovered by identifying correlated genes that: (i) were expressed within the target organ for the analysed trait; (ii) contained a codon-changing SNP; and (iii) had gene ontology (**GO**) database annotations that were related to the phenotype. Here, we further develop and adapt this approach for analysis of mouse genetic data (**mMap**). To demonstrate utility, mMAP was used to: (i) analyse genes affecting the response to drugs of abuse, and (ii) to identify SNP alleles that are unique to individual strains and have potential to impact biomedical trait responses.

Features:

The architecture for this method is outlined in Figure 1. The gene symbol and identified alleles for each SNP are analysed by this program to determine the presence of SNPs in 26 types of functional (like proteins domains and DNA-binding motifs) and 8 types of regulatory genomic regions (promoter and enhancer motifs, see below for details). The SNPs containing genes are then processed through over-representation tests to analyse the impact of SNPs-regulated biological functional, pathways and regions. This enables the impact of SNP alleles on protein functional or genome regulatory regions to be assessed. Major features of the method are lists as following:

Protein functional and regulatory annotations: For functional assessment of SNPs, we pre-compiled publicly available data from several available resources. Protein functions, protein domains, and posttranslational modification data from UniProt[3], PTMdb[4]. Additional protein domains data compiled from InterPro[5], Pfam[6] and SMART[7]. the protein localization data, as well as the transmembrane regions data and interactions data retrieved from STRING[7] db. For regulatory assessment of SNPs, data on genomic regulatory regions enhancers, promoters, promoter-flanking regions, DNA- methylation (CpG Islands), insulators, transcription start-sites (TSS) and transcription factor binding-sites (TFBS) were compiled from Encode consortium[8], Ensembl[9], VISTA[10], and UCSC genome browser[11]; structural data (Disulfide-bond, alpha-helix, Coiled-coil regions and turn) from PDB[12]; mouse organ specific bulk gene expression from Expression Atlas[13]. The data of mouse microRNAs target binding motifs was retrieved from miRBase[14] and the alternative start/stop codon mouse data from

TISdb[15]. The protein conservation is computed with Rate4Site[16], the algorithm calculates the evolutionary rate at which a residue changes, the positions which are slow to change provides a measure of being evolutionary conserved and vice-versa. The SNP positions that overlap with the coding regions are provided with conservation scores.

A user provided flat file with gene names and SNP positions processed according to the type of function choose. In case of functional analysis of SNPs, the method maps amino acid level SNPs positions to protein functional features like PTMs sites or domains, whereas in case of regulatory, the SNP with genomic coordinates is mapped to regulatory regions like promotors or enhancers. The outcome is stored in a "*functional-accessment.txt*" file (see supplementary info for all the information regarding the commends and output files content).

Protein pathways, network, GO-terms enrichment analysis: By mapping allelic variations to the conserved genomic regions, the impact of allelic variations on biological processes and pathways is assessed by hypergeometric tests as implemented[7]. This part of the analysis assesses the potential impact of SNPs on biological processes, pathways and functions. Additionally, a protein interaction network analysis describes the functional relationship between proteins with SNPs. A similar analysis evaluates if SNP-alleles are significantly co-localized in any of the given genomic regions, this highlights the impact of a genomic region in a given phenotype. Overall, this section aims to highlight the relevant biological regions, pathways and functions that can be disrupted due to SNP-containing genes with a high potential to mediate a phenotype (Figure 2).
In the output enrichment lists, for the clarity purposes, we integrate an approach that based on the "GO-term merging to nearest ancestral term"[17] and shorten the lists to a human readable level. Also, the visualisations are generated as a part of output, to present biological processes in a way that make interpretation easier. Overall, by doing in-depth analyses of the impact of polymorphism on protein activity and regulation, the method provides actionable information to prioritize genes from big data for further validation analyses.

Literature search: A comprehensive dataset is compiled from various text-mining resources by pooling together all the data for a given "gene-disease relation" to complement the final output of our pipeline. These gene-disease relationship text-mining data were resourced from (a) Phenolyzer[18], a tool that prioritizes genes based on information collected from several platforms; (b) OpenTarget[19] and DisGeNET[20], collections of gene-disease associations from difference sources including from literature searches; and (c) PubMed and NCBI gene db. The outcome complemented with a literature review file "*phenotype_accessment.txt*" for the genes containing SNPs of interest.

Single cell expression: Single cell expression data of mouse proteins were retrieved from publicly available datasets[21]. These precompiled datasets contain information of gene expression clusters of different cell-types. We provide the expression values of each gene with SNPs and a file "scExpression-data.txt" generated for further analyses.

The final outcome consists of files and visualizations generated for each category of analysis as well as summary files.

Example applications:

*Case study 1: Functional assessment of coding SNPs (cSNPs) affecting responses to drugs of abuse***.** Genetic factors play a crucial role in determining whether an exposed individual will become addicted to an abused drug [22]**.** Several such genetic factors are strongly associated with addiction to different drugs of abuse in GWAS, which include: ALDH2 and alcohol abuse [23]; CHRNA5, CHRNA3, and CHRNB4 with nicotine[24]; and OPRM1 with opioid abuse[25]. While the functional connection between these genes and the corresponding abused drug is very clear, there are many instances were GWAS identify genetic factors whose functional connection is less clear. We use mMap to assess the potential impact of SNP alleles present in 638 genes reported for (human or mouse) substance abuse through a literature search. A subset of these 208 genes have SNPs (from our database of 21.3M SNPs, as identified and the data was compiled previously[26]) located within the 16 different types of functional regions (Figure 3, Figure S1). With majorly cSNPs disrupted regions are protein domains (SNPs=149) and disulfide bonds (SNPs=129). One example, a gene called neuronal adhesion protein, *Nrcam*, has published genetic SNP association with vulnerability to autism, alcoholism, and substance addiction[27]. We identified 10 SNP-alleles in three different domains present in it (two SNP-alleles in Ig-like C2-type5, two in Ig-like C2-type6, six in Fibronectin type-III) with a broad range of functional classes, from protein-protein interactions (PPI) to cell adhesion, morphology, and migration. The implication of *Nrcam* protein domains in modulating functions due to SNP-alleles assign a novel role to these SNPs that they can play in addiction phenotype beyond genetic associations. Another example is, *Sema6d* protein, that plays an important role in neuronal rewiring during development, contains three allelic variations in its disulfide bonds (DiSB), can compromise its normal activities. Previously reported for its association with substance abuse, the cSNPs present in DiSB of *Sema6d* may reveal a functional insight of how structural compromise can play a role in substance addiction pathways. The cSNPs present in genes with substance addiction association reveal a previously unknown function of DiSB in the phenotype.

Together, in 208 genes, mMap identified genes with greater number of cSNPs are *Disc1* and *Syne1* (Figure 2). And pathways enrichment tests of genes with cSNPs revealed pathways like neuroactive ligand-receptor interaction (p= 3.65e-15), calcium signalling (p= 2.2e-07) and PI3K-Akt signalling (p= 6.25e-06) (Table S1). These pathways are critical for the brain development and functions and reflect a possible functional compromise that can play an important role during substance addiction. This analysis presents an additional way to prioritize phenotype relevant genes and pathways by focusing on SNPs that can disrupt protein functional regions like domains (Table S2).

Analysis of regulatory SNPs (rSNPs). It has often been difficult to determine the impact of rSNPs, but the recent production of many new data sets by the ENCODE project has increased our knowledge about the sequence and function of genomic regulatory regions[8]. This data along with other advances makes it possible assess the functional impact of many rSNP alleles. For mouse SNP analysis, the mMap framework analyses the allelic impact on eight different types of regulatory regions: promoters, enhancers, transcription start sites (TSS), CpG Islands, splice-sites, microRNA binding motifs, alternative stop/start codons and insulators. To demonstrate utility, mMap was used to analyse the impact of murine rSNP alleles present in 638 genes (Table 2) associated with addiction. We found rSNPs within promoters (SNPs=559), CpG Islands (SNPs=3955), enhancers (SNPs=3544), insulators (SNPs=72) and transcription start-sites (n=5) of these genes. As one example, Insulin receptor (*Insr)* encodes a tyrosine kinase that regulates the insulin response through activation of several intra-cellular signalling pathways. Recently shown that insulin regulate the ability of drugs that exert their role through impacting dopamine dependent neurotransmission[28]. mMap analysis identified 24 SNPs within the *Insr* promoter region (Figure S2). By affecting the binding of transcription factors to the promoters, these SNPs could have an allelic effect on *Insr* mRNA transcription. As another example, *Neurexin 3* (*Nrxn3*) encodes a protein that functions in synapse development[29], and *NRXN3* alleles have been associated with alcohol abuse [30] and cocaine dependence [31] in human GWAS. mMap analysis identified 278 murine SNPs located within or near CpG islands in *Nrxn3*. Moreover, *Nrxn3* mRNA is extensively spliced, and the different isoforms have different functional effects on the synapse[32]. In total, mMAP analysis identified 299 rSNPs in *Nrxn3*, and any of these could alter its splicing. SNPs in *Glyoxalase 1 (GLO1)* have been associated with the level of alcohol consumption, and it has been considered as a therapeutic target for treatment of alcoholism[33]. mMAP identified 34 SNPs located within or enhancer regions in *Glo1*. This suggest that rSNP alleles could impact its expression. Astrotactin-2 (*Astn2)* has been shown to effect many neurodevelopmental phenotypes[34] and null mice of its interacting partner protein *Astn1* exhibit alterations in balance and coordination[35]. mMap

identified 11 SNPs present within or near insulator regions of *Astn2* (Figure S2). Thus, an allelic effect within an insulator region could alter *Atn2* mRNA expression in mouse cerebellum, which could produce a neurodevelopmental effect that alters the response to drugs of abuse. Of note, murine strains show dramatically different responses to cocaine[36] and opiates. Thus, mMAP analysis indicates that there are several routes through which rSNPs in *Insr*, *Nrxn3* or *Glo1* could impact the response of inbred strains to drugs of abuse.

Of particular interest, a subgroup of addiction-related genes has SNP-alleles that either introduce (n=16) or remove (n=4) a stop codon in 14 genes across the 42 analysed strains. These 14 genes are reported previously for their phenotypically relevant genomic associations however, the role of these proteins not yet evaluated in substance addiction[37,38]. As one example, carboxylesterases are a family of enzymes that are highly expressed in liver, and are known to play a role in the metabolism of drugs of abuse[36] [39]. *Carboxylesterase 2a (Ces2a)* has two SNPs that introduce stop codons at amino acids 497 and 527, which result in the expression of a truncated form of this enzyme (Table S3). Truncated proteins are present in other animal species (including humans), and human protein truncations are shown to be associated with important medical phenotypes such as hypertension[40]. This can be of interest to put forward for the evaluation of these “knockouts” for their potential role in the development of the addiction phenotype.

mMap based biological pathways over-representation in genes with rSNPs identified pathways like neuroactive ligand-receptor interaction (p=1.08e-18), cocaine addiction (2.06e-14) and calcium signalling (p=1.0e-6). Overall, these are crucial pathways that regulate the neuroplasticity under rapid response conditions.

Conclusion of case study 1: The mMaps results of biological impact of SNPs present in genes known for substance addiction, provide novel insights at a multidimensional space of biomolecular features like protein domains and enhancers, and emphasize the importance of this approach in prioritizing the phenotypically crucial genes with regulatory and functional SNP-alleles.

Case study 2: Mouse Strains Private SNPs Analysis, To Access and Predict Their Functional Contribution(s). In a second application of mMap, we analyzed the individual genetic makeup of 42 inbred mouse strains (129P2/OlaHsd, 129S1/SvImJ,129S5/SvEvBrd, AKR/J, A/J, B10.D2-Hc<0>, BTBR + tf>/J, BUB/BnJ, BALB/cJ, C3H/HeJ, C57BL/10J, C57BL6NJ, C57BR/cdJ, C57L/J, C58/J, CBA/J, CE/J, DBA/1J, DBA/2J, FVB/NJ, I/LnJ, KK/HIL, LG/J, LP/J, MA/MyJ, MRL/MpJ, NOD/ShiLtJ, NON/ShiLtJ, NU/J, NZB/BINJ, NZO/HILtJ, NZW/LacJ, P/J, PL/J, RF/J, RHJ/LeJ, RIIIS/J, SEA/GnJ, SJL/J, SM/J, ST/bJ, SWR/J) and six wild-drive strains (CAST/EiJ, MOLF/EiJ, PWD/PhJ, PWK/PhJ, SPRET /EiJ, WSB/EiJ)

consisting of 21.3M SNPs as described previously[26], to access the contribution of various genetic (coding or non-coding) aspects in individual phenotypes. A private SNP, we described as a homozygous variable allele compared to high quality homogenous reference alleles present in the rest of the strains in our database (Table S4). We identified 3.35M (15% of total) private SNPs across 48 strains. Out of these, 11242 SNPs are coding – i.e. changing an amino acid codon – and a mMap analysis of these private cSNPs revealed 2969 SNP-alleles can be mapped to functional protein regions. Out of which, a subgroup of 1502 cSNPs overlap with annotated conserved protein domains. Functionally, these private cSNPs containing domains have a range of functions like protein interactions, stability, or signalling. Among strains with 94 cSNPs present in the annotated domains, CE/J strain has the highest number of domains disrupted. Of interest, in closely related strains, like DBA1/J and DBA/2J, the DBA/1J strain has three proteins whereas the DBA/2J has six different proteins with cSNPs present in domains. Both of these strains only have 5.6% alleles differences[41] but at functional level they do not share any of these disrupted domains. DBA/1J has altered domains, TSP type-1, Fibronectin type-III and Peptidase S8 with functions cell-to-cell communications, protein binding and Proteolytic Activity, respectively. Whereas DBA/2J has altered domains with functions including immune response (C-type lectin) and detoxification (Rhodanese). Alterations in different protein domains show the way in which private cSNPs work at a functional level to achieve a unique phenotype even for closely related strains like DBA/1J and DBA/2J. mMap identified another group of genes (n=258) with cSNPs overlapping to disulfide bonds (DiSB). Among studied strains, the CE/J has 25 cSNPs, SM/J has 21, NZO/HILtJ and RIIIS/J 20 cSNPs present in their DiSB. The cSNPs impact on DiSB may have a role to play in the phenotype of these strains (see below).

Additionally, mMap analysis suggested a total of 313,384 (8.95% of private) SNPs potentially impacting genomic regulatory regions of the studied strains. Major regulatory regions that are disrupted include CpG islands (CGI), genomic enhancers, and promoters (Figure 4). Out of these rSNPs, 8104 rSNPs (2.7%) fall in the CGI, 22543 (7.7%) in the enhancers, and 49438 (17 %) in the promoter regions of nearby genes (Figure 4). These results emphasize the potential variations in epigenetic and transcription regulatory mechanisms behind the phenotypic differences among mouse strains, in addition to protein codon changing SNPs. Promoters play crucial role in genomic and protein functions by transcription regulation. A set of genes (n=11622) identified by mMap has disrupted nearby promoters by private rSNPs. Among most affected strains, the C57L/J strains have 2842, SW/R has 2276, and 129S5/SvEvBrd has 1945 gene. Enhancers promote gene transcription by facilitating the binding of activator proteins. mMap identified 3948 genes have rSNPs present in their nearby enhancers, that can mediate the transcription of these genes. Among strains,

SPRET/EiJ (n =1806), CAST (n=675), CE/J (n = 579), SM/J (n=483), and 129S5/SvEvBrd (n=448) have the highest number of genes with potentially disrupted nearby enhancers by their private rSNPs.

Another region of particular interest uncovered by the application of mMap was CGI (n=3585 genes). Different strains show different number of disrupted CGI disrupted by private rSNPs. 129S5/SvEvBrd (n=560), RIIIS/J (n=301), and B10 (n=287) strains have the highest number of genes with rSNPs in nearby CGIs. One example is of a phenotypically important gene, G1/S-specific cyclin-D1 (*Ccnd1*), which has rSNPs in the RIIIS/J strain, the gene has been reported for eye- and lens-related biomedical traits[42]. This strain is known for the spontaneous development of lens and cataract phenotypes. mMap analysis suggested a potential novel transcription mechanism via rSNPs mediated methylation of CGI region nearby to *Ccnd1* gene, which can play an epigenetic regulation of vision phenotype of this strain. In KK/HiJ strain, the pathways enrichment among genes with disrupted nearby CGI include macromolecule biosynthesis (p=5.05 e-05) and metabolism regulation (p=0.00023). The KK/HiJ which serves as a model to study type-2 diabetes mellitus (DM-2) and the genes with rSNPs containing CGI could be prioritized for further assessment of their role in KK/HiJ metabolism and in DM-2 disease phenotype.

Together, the private SNPs show an individual trend in functional and regulatory genomic regions to disrupt these features in a particular individual way that is unique and had not been explored before. *In conclusion of case study2*: Both, private cSNPs and rSNPs, based mMap results are an important step to demonstrate how individual SNPs can regulate different regulatory and functional circuits at individual phenotype level.

DISCUSSION

We provided a computational pipeline with two data analyses examples to highlight the importance of linking big SNPs data to biological features in order to interpret the potential consequences of SNPs onto the phenotypes. A method like this was overdue to fill the knowledge gap of number of SNPs present in the genome of a premier biomedical model organism like mouse and their biological importance. The mMap pipeline can greatly improve the data interpretation by demonstrating the biological role that SNP-alleles play in a disease phenotype. Importantly, the biological understanding of genetically associated loci is crucial for the formulation of therapeutics against genetic targets. In both mMap applications, the SNPs present in regulatory regions outnumber the SNPs detected in the protein coding regions, which is quite important finding. Consistent with previous results[43], our findings show that the transcription regulation plays a potentially crucial role in mediating a

biomedical phenotype susceptibility. This highlights the need to revisit the conventional paradigm that focuses on targeting coding SNPs as important regulators of phenotypes.

In general, the common mechanisms of substance response are either not known or difficult to define. However, our approach fills this knowledge gap by showing how genetically associated loci assert their pathway-level influence by modulating the protein regulatory and functional features to impact a phenotype. we examined the biological consequences of SNP-alleles from 638 previously reported genes for their genetic association with the phenotype. Previously GWAS reported loci for human alcoholism and alcohol consumption in mice, including autism susceptibility candidate 2 (*AUTS2*) gene[44], mu-opioid receptor[45] (OPRM1), Ankyrin Repeat and Kinase Domain Containing-1[46] (ANKK1), Nesprin-1 (SYNE1), and GABA receptor alpha2 subunit [47]. We identified the SNPs present in these – and other – genes that can mediate response to alcohol by affecting both genetic and epigenetic elements. *Auts2* has four SNP-alleles in regulatory enhancer regions which can impact its expression, *Oprm1* has 1 allelic variation in its structural alpha helix, *Ankk1* has four SNP-alleles in functionally crucial protein kinase domain which can mediate phosphorylation events of *Ankk1* kinase, whereas *Syne2* has 6 SNP-alleles in promoter regions that can affect its transcription and 25 alleles in structurally important Spectrin regions. This investigation is critical to understand a biological role that SNPs can play in alcoholism. Likewise, genes reported for opioid and cocaine responses have several SNPs present that can change the epigenetic and regulatory architecture. The central opioid addiction regulator, *Oprm1*, has hypermethylated promoter and histone deacetylation associated with heroin addiction[48]. We detected SNPs (n=6) present in nearby CGI of *Oprm1* gene that can change the methylation landscape of the gene and addiction regulation phenotype. Another interesting fact, the pathways enrichment analyses revealed neuroplasticity pathways that are common to these and other SNPs carrying addiction genes. This shows that the underlying common routes that these genes take in different types substance addiction. It is also noteworthy that our method accurately predicted the pathways known to play crucial role in addiction.

Previous studies on the examination of SNPs present in inbred mouse strains were largely focusing on the analysis of allele frequencies[49]. At one instance, authors identified SNPs from 36 inbred strains and predicted underlying functional effects of private SNPs on individual strain phenotypes, but the overall prediction did not include annotations of genome regulatory or functional regions[50]. In general, these and other such studies have greatly increased the identification and knowledge of genetic variations present in mouse populations. However, the lack of detailed functional understanding has left unexplored gaps

between SNP data and known biological features of different mouse strains. We explored 3.35M private SNPs from a deserve group of 48 mouse strains to detect their influence on functional and regulatory genetic horizon at an individual strain level. Among major private SNPs containing regulatory elements, the CGI and enhancers appear as most polymorphic regions. In comparison to functional regions, the regulatory regions may have much more influence in the phenotypic differences, even in closely related strains. By doing so, we show that our approach provides an additional way to observe the regulatory mechanisms by which private SNPs can mediate important functions at an individual phenotype level.

Future prospects: Our method can provide an early glimpse of the role played by SNPs at an individual strain level. It thereby provides a foundation of future method development to examine and to interpret the functional contribution of individual genetic makeup at whole genome. However, it is also important to consider that the predictive efficiency of such a method can greatly improve through the availability of 'complete' data sets from new and improved experimental methods, like single cell multiple-omics analyses with additional information of regulatory and functional genome regions. In future, we therefore plan to include additional datatypes to make mMap even more useful. Also, both applications of mMap revealed that the majority of SNPs are either intergenic or intronic which cannot be mapped to present mouse data of functional and regulatory regions. In this case, the inclusion of conservation-based computational epistasis methods will certainly help in defining a role for intergenic SNPs and increase method's evaluation strength.

Conclusion: We developed a computational tool that can analyse SNP data through characterizing the impact of allelic variation on genomics functional features like protein domain structure, post-translational modifications, protein-protein interactions, and on regulatory features like promoter or enhancer regions. The hallmark of this approach is linking the conserved genomic regions impacted by SNP-alleles to biological pathways and functions that can potentially disrupt phenotypes. We first applied mMap to show the impact of SNP-alleles on the known genetic factors of substance addiction. A second application of this approach on the individual SNP profiles of mouse strains highlighted the contribution of private allelic variations in individual phenotypes of each strain.

AVAILABILITY

source-code is available via a GitHub page.

Source code : https://github.com/AhmedArslan/mMap

Operatizing system(s): Mac OS X and Windows
Programming language: Python

Reference:

1. Arslan, A. & Noort, V. Van. Sequence analysis yMap : an automated method to map yeast variants to protein modifications and functional regions. **33**, 571–573 (2017).
2. Zheng, M., Dill, D. & Peltz, G. A better prognosis for genetic association studies in mice. *Trends in Genetics* (2012). doi:10.1016/j.tig.2011.10.006
3. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky1049
4. Huang, K. Y. *et al.* DbPTM in 2019: Exploring disease association and cross-Talk of post-Translational modifications. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky1074
5. Mitchell, A. L. *et al.* InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky1100
6. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky995
7. Szklarczyk, D. *et al.* STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* (2015). doi:10.1093/nar/gku1003
8. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
9. Aken, B. L. *et al.* The Ensembl gene annotation system. *Database (Oxford).* (2016). doi:10.1093/database/baw093
10. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser - A database of tissue-specific human enhancers. *Nucleic Acids Res.* (2007). doi:10.1093/nar/gkl822
11. James Kent, W. *et al.* The human genome browser at UCSC. *Genome Res.* (2002). doi:10.1101/gr.229102. Article published online before print in May 2002
12. Berman, H. M. *et al.* The protein data bank. *Acta Crystallogr. Sect. D Biol. Crystallogr.* (2002). doi:10.1107/S0907444902003451
13. Papatheodorou, I. *et al.* Expression Atlas: Gene and protein expression across multiple studies and organisms. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gkx1158
14. Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* (2015). doi:10.7554/eLife.05005
15. Lee, S. *et al.* Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution. *Proc. Natl. Acad. Sci. U. S. A.* (2012). doi:10.1073/pnas.1207846109
16. Pupko, T., Bell, R. E., Mayrose, I., Glaser, F. & Ben-Tal, N. Rate4Site: An algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. in *Bioinformatics* (2002). doi:10.1093/bioinformatics/18.suppl_1.S71
17. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS One* (2011). doi:10.1371/journal.pone.0021800
18. Yang, H., Robinson, P. N. & Wang, K. Phenolyzer: Phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods* (2015). doi:10.1038/nmeth.3484
19. Koscielny, G. *et al.* Open Targets: A platform for therapeutic target identification and Validation. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkw1055
20. Piñero, J. *et al.* DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkw943
21. Franzén, O., Gan, L. M. & Björkegren, J. L. M. PanglaoDB: A web server for exploration of mouse and human single-cell RNA sequencing data. *Database* (2019). doi:10.1093/database/baz046
22. Bierut, L. J. Genetic Vulnerability and Susceptibility to Substance Dependence. *Neuron* (2011). doi:10.1016/j.neuron.2011.02.015
23. Edenberg, H. J. The genetics of alcohol metabolism: Role of alcohol dehydrogenase and aldehyde dehydrogenase variants. *Alcohol Res. Heal.* **30**, 5–13 (2007).
24. Saccone, N. L. *et al.* The CHRNA5-CHRNA3-CHRNB4 nicotinic receptor subunit gene cluster affects risk for nicotine dependence in African-Americans and in European-Americans. *Cancer Res.* (2009). doi:10.1158/0008-5472.CAN-09-0786
25. Bond, C. *et al.* Single-nucleotide polymorphism in the human mu opioid receptor gene alters β-endorphin binding and activity: Possible implications for opiate addiction. *Proc. Natl. Acad. Sci. U. S. A.* (1998). doi:10.1073/pnas.95.16.9608

26. Peltz, G. *et al.* Next-generation computational genetic analysis: Multiple complement alleles control survival after Candida albicans infection. *Infect. Immun.* (2011). doi:10.1128/IAI.05666-11
27. Marui, T. *et al.* Association of the neuronal cell adhesion molecule (NRCAM) gene variants with autism. *Int. J. Neuropsychopharmacol.* (2009). doi:10.1017/S1461145708009127
28. Daws, L. C. *et al.* Insulin signaling and addiction. *Neuropharmacology* (2011). doi:10.1016/j.neuropharm.2011.02.028
29. Harkin, L. F. *et al.* Neurexins 1-3 each have a distinct pattern of expression in the early developing human cerebral cortex. *Cereb. Cortex* (2017). doi:10.1093/cercor/bhw394
30. Hishimoto, A. *et al.* Neurexin 3 polymorphisms are associated with alcohol dependence and altered expression of specific isoforms. *Hum. Mol. Genet.* (2007). doi:10.1093/hmg/ddm247
31. Kelai, S. *et al.* Nrxn3 upregulation in the globus pallidus of mice developing cocaine addiction. *Neuroreport* (2008). doi:10.1097/WNR.0b013e3282fda231
32. Aoto, J., Földy, C., Ilcus, S. M. C., Tabuchi, K. & Südhof, T. C. Distinct circuit-dependent functions of presynaptic neurexin-3 at GABAergic and glutamatergic synapses. *Nat. Neurosci.* (2015). doi:10.1038/nn.4037
33. de Guglielmo, G., Conlisk, D. E., Barkley-Levenson, A. M., Palmer, A. A. & George, O. Inhibition of Glyoxalase 1 reduces alcohol self-administration in dependent and nondependent rats. *Pharmacol. Biochem. Behav.* (2018). doi:10.1016/j.pbb.2018.03.001
34. Lionel, A. C. *et al.* Disruption of the ASTN2/TRIM32 locus at 9q33.1 is a risk factor in males for autism spectrum disorders, ADHD and other neurodevelopmental phenotypes. *Hum. Mol. Genet.* **23**, 2752–2768 (2014).
35. Adams, N. C., Tomoda, T., Cooper, M., Dietz, G. & Hatten, M. E. Mice that lack astrotactin have slowed neuronal migration. *Development* (2002).
36. Cervantes, M. C., Laughlin, R. E. & Jentsch, J. D. Cocaine self-administration behavior in inbred mouse lines segregating different capacities for inhibitory control. *Psychopharmacology (Berl).* (2013). doi:10.1007/s00213-013-3135-4
37. Stringer, S. *et al.* Genome-wide association study of lifetime cannabis use based on a large meta-analytic sample of 32330 subjects from the international cannabis consortium. *Transl. Psychiatry* (2016). doi:10.1038/tp.2016.36
38. Frank, J. *et al.* Genome-wide significant association between alcohol dependence and a variant in the ADH gene cluster. *Addict. Biol.* (2012). doi:10.1111/j.1369-1600.2011.00395.x
39. Jones, R. D., Taylor, A. M., Tong, E. Y. & Repa, J. J. Carboxylesterases are uniquely expressed among tissues and regulated by nuclear hormone receptors in the mouse. *Drug Metab. Dispos.* (2013). doi:10.1124/dmd.112.048397
40. Deboever, C. *et al.* Medical relevance of protein-truncating variants across 337,205 individuals in the UK Biobank study. *Nat. Commun.* (2018). doi:10.1038/s41467-018-03910-9
41. Stylianou, I. M. *et al.* Differences in DBA/1J and DBA/2J reveal lipid QTL genes. *J. Lipid Res.* (2008). doi:10.1194/jlr.M800244-JLR200
42. Das, G., Clark, A. M. & Levine, E. M. Cyclin D1 inactivation extends proliferation and alters histogenesis in the postnatal mouse retina. *Dev. Dyn.* (2012). doi:10.1002/dvdy.23782
43. Li, C. Y. *et al.* Meta-analysis and genome-wide interpretation of genetic susceptibility to drug addiction. *BMC Genomics* (2011). doi:10.1186/1471-2164-12-508
44. Schumann, G. *et al.* Genome-wide association and genetic functional studies identify autism susceptibility candidate 2 gene (AUTS2) in the regulation of alcohol consumption. *Proc. Natl. Acad. Sci. U. S. A.* (2011). doi:10.1073/pnas.1017288108
45. Ramchandani, V. A. *et al.* A genetic determinant of the striatal dopamine response to alcohol in men. *Mol. Psychiatry* (2011). doi:10.1038/mp.2010.56
46. Grzywacz, A. *et al.* Influence of DRD2 and ANKK1 polymorphisms on the manifestation of withdrawal syndrome symptoms in alcohol addiction. *Pharmacol. Reports* (2012). doi:10.1016/S1734-1140(12)70909-X
47. Bierut, L. J. *et al.* A genome-wide association study of alcohol dependence. *Proc. Natl. Acad. Sci. U. S. A.* (2010). doi:10.1073/pnas.0911109107
48. Meyer, W. H., Houghton, J. A., Lutz, P. J. & Houghton, P. J. Hypoxanthine:Guanine Phosphoribosyltransferase Activity in Xenografts of Human Osteosarcoma. *Cancer Res.* (1986).
49. Doran, A. G. *et al.* Deep genome sequencing and variation analysis of 13 inbred mouse strains defines candidate phenotypic alleles, private variation and homozygous truncating mutations. *Genome Biol.* (2016). doi:10.1186/s13059-

016-1024-y

50. Timmermans, S., Van Montagu, M. & Libert, C. Complete overview of protein-inactivating sequence variations in 36 sequenced mouse inbred strains. *Proc. Natl. Acad. Sci. U. S. A.* (2017). doi:10.1073/pnas.1706168114

TABLE AND FIGURES LEGENDS

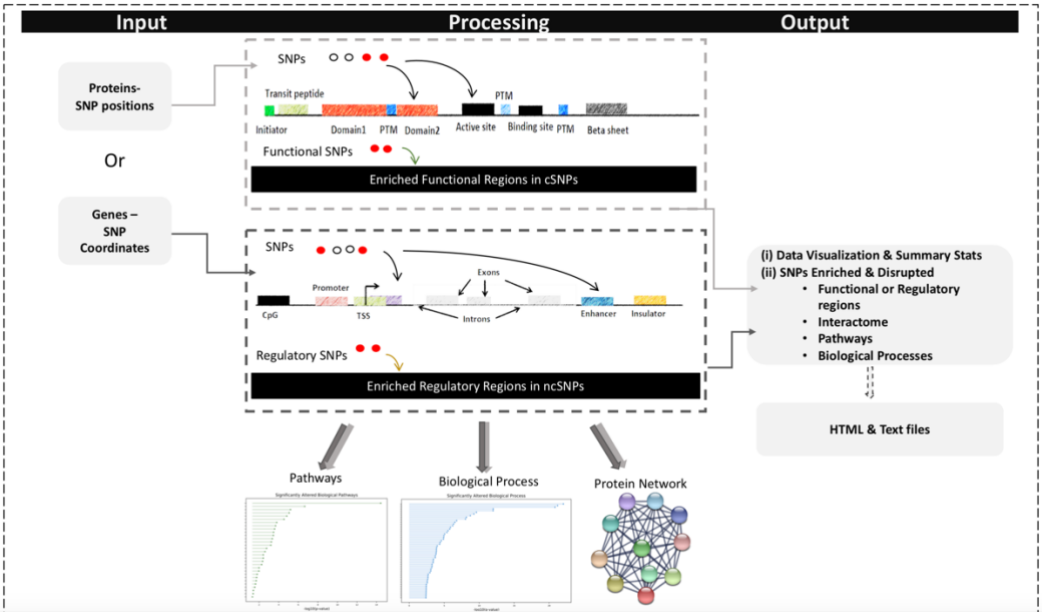

Figure 1. mMap framework**.** (Input) user provided data is processed and (in "Processing" step) mapped to the regulatory or functional regions, depending on the type of approach choose; to prioritize genes. The prioritized SNP containing genes are then processed through different additional tools/analyses including (i) biological processes enrichment (ii) KEGG pathways enrichment (iii) functional or regulatory region enrichment (iv) biological network analysis. (Output) a comprehensive report and data visualisations are generated as the part of final outcome (see methods).

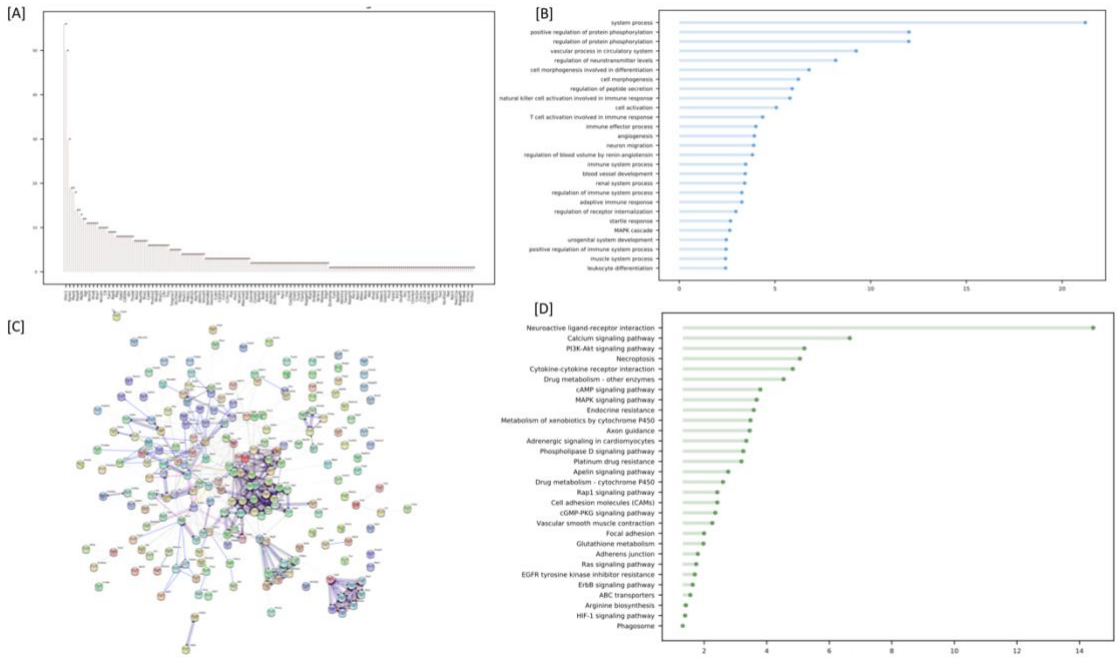

Figure 2: Example of data visualizations of results output. [a] number of SNPs on interest present in each gene [b] the biological processes and [d] pathways over-presentation of genes with SNPs of interests. All these barcharts are implemented in python's matlibplot package. [c] the interactomes of genes with SNPs shows the protein-protein interactions implemented from String-db.

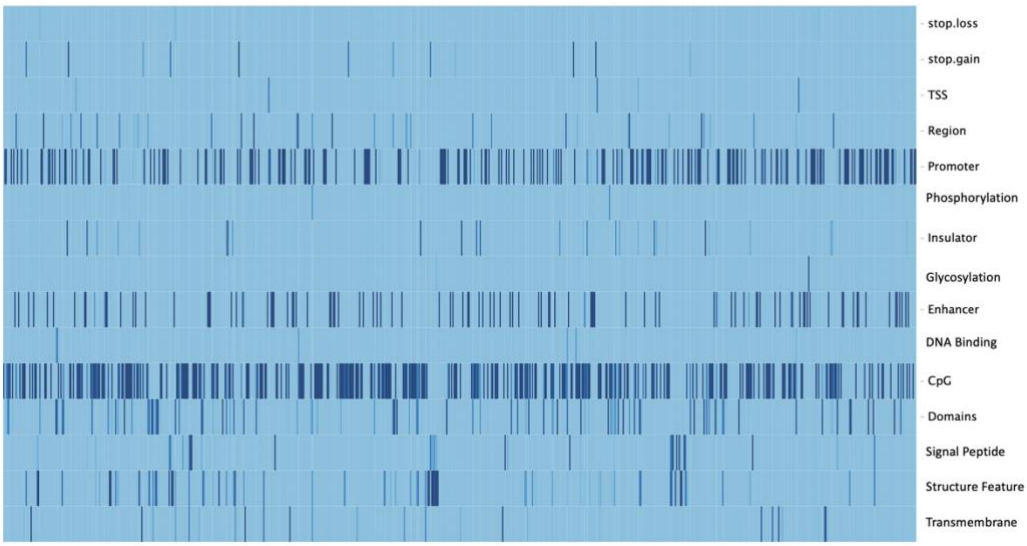

Figure 3: The genes of substance addiction show variable number of SNPs overlap with different genomic regions, with regulatory regions like promoters and CpG Islands are most disrupted as SNPs data largely disrupt these areas.

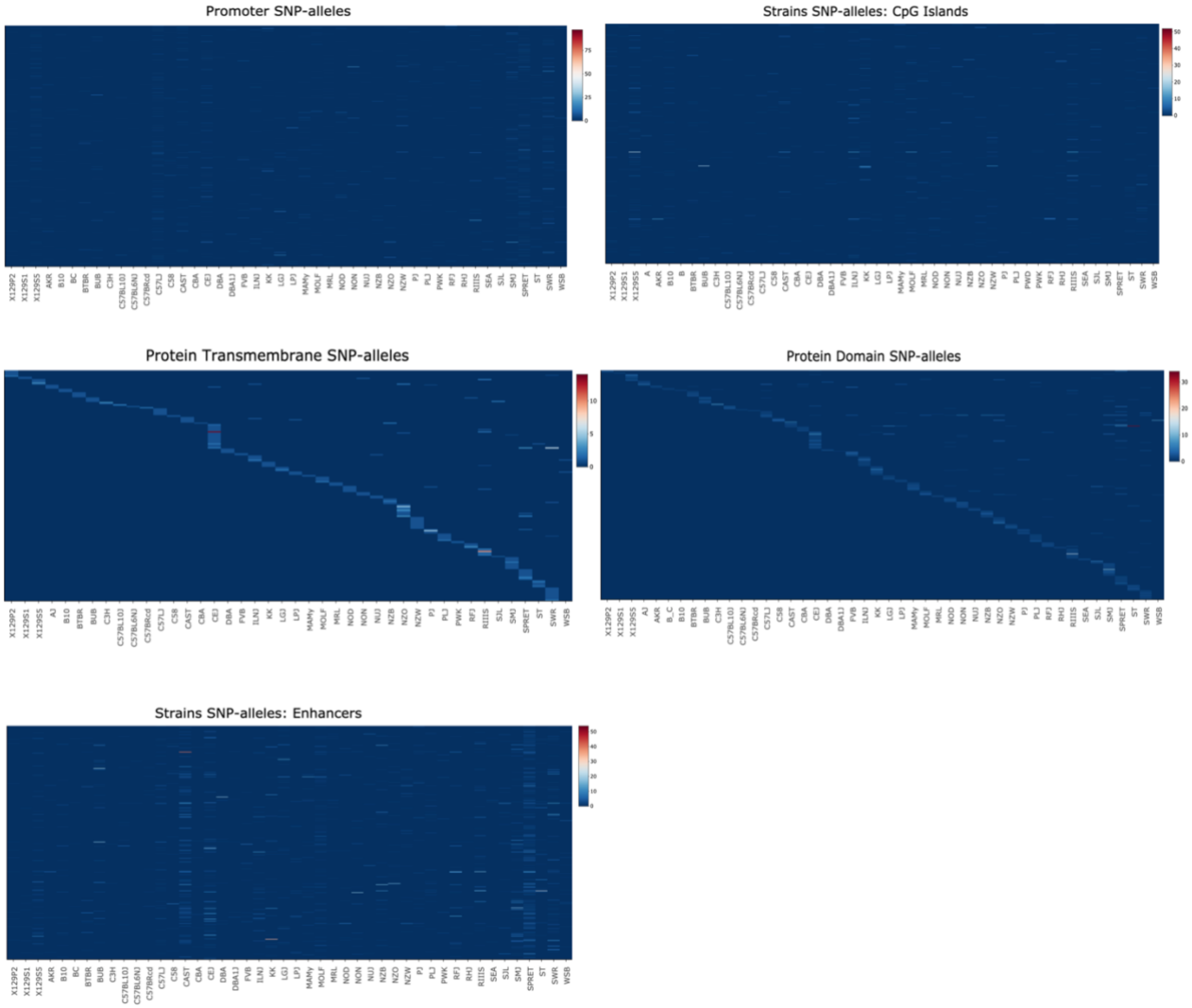

Figure 4. Different trends of stains private SNPs overlapping with various genomic regions. The horizontal axis represents different mice strains and vertical axis contains number of SNPs in each gene.