SNPector: SNP inspecting tool to detect disease existence and drug response in direct naked sequencing depends on mutation index

Peter T. Habib[1], Al Samman M. Al Samman[2] and Aladdin Hamwieh[1]

[1]Department of Biodiversity and Crop Improvement, International Center for Agriculture Research in the Dry Areas (ICARDA)

[2]Department of Genome Mapping, Molecular Genetics and Genome Mapping Laboratory, Agricultural Genetic Engineering Research Institute (AGERI), Giza, Egypt

Abstract

In recent years the number of genetically originated disease cases has raisin the alarm worldwide and has sparked interest in the development of precision medicine using molecular bio-markers such as single nucleotide polymorphism (SNP) that draw the attention of researchers due to the ability of diagnosing diseases and assessing medicines effectiveness in patients with aid of advanced diagnostic computing systems to help early detection of disease and prevention. The extraction of disease-associated SNP to discover disease is bottle nick problem because of the difficulty in selecting meaningful data and discover of SNP without need of reference to guide the software. The main objective of SNPector (https://github.com/peterhabib/SNPector) is to build python-based and user-friendly software to determine the existence of SNP in given sequence based on BLAST alignment and Clinvar database of NCBI to discover associated disease with detected SNP and calculate the linkage between detected SNPs and other different SNPs with providing of SNP Chip used for SNP genotyping . Application program interface (API) used to retrieve data and Several python packages are used for the visualization of variants and gene, disease, phenotype, and drug.

1 Introduction

Single nucleotide polymorphism (SNP), the most genetic variations among the individuals experimentally approved occurs in the human genome[1]. these randomized modification in the DNA bases may create alterations in amino acid residues of protein sequences so, modify their functions that consequently result in different disease condition in the individuals. many SNPs have been reported to be a disease-linked genetic markers in the human genome, which were used to uncover genes responsible for a specific disease Distinguishing proof and portrayal of rich number of markers will be important to relate the significant

transformations in SNP and to discover their association in the advancement of malady conditions and thus, Clarification of the phenotypic-association mechanisms for these variations is vital for understanding the sub-atomic subtleties of disease beginning and creating novel therapeutic approaches[2-3]. So far, most researches has been concentrated on disease-associated SNPs (daSNPs) found in coding regions or exons, especially the non-synonymous SNPs which may change the biochemical capacity of coded proteins. Nevertheless, SNPs are also occurs in many other places of genes including: promoters, introns, 5'- and 3' UTRs. Therefore, modifications in gene expression, their consequences on disease susceptibility and medication reaction vary depending on the location of the SNPs**.** The SNPs promoter influence gene expression by altering transcription, transcription-factor binding, DNA methylation and histone modifications[4-11]**.** The exonal SNPs are vulnerable to cancer by preventing gene transcription and translation [12-14]. SNPs produce splice variants of transcripts in intronic regions, stimulate, and interrupt non-coding RNA bindings and functions[15-17]. SNPs in the 5'-3UTR affect translation, whereas SNPs in the 3'-UTR affect microRNA (miRNA) binding[16-19]. SNPs in regions far from the existing genes decrease or increase transcription of the genes through long-range cis effects [20]. In non-coding regions where ninety three percent as stated by like intron, long terminal repeats (LTRs) and intergenic regions, the majority of daSNPs found so far pose significant challenges to the researcher in the understanding of their disease participation.

2 Related work

Based on the increase in accessible variants data, softwares to modify produced data can be used for the production of new knowledge in disease diagnosis and drug response studies and their SNPs involvement can be further used as a biomarker to personalize disease and drug. Many tool emerged in last few years to study the SNP effect, for example: [21] is a versatile tool for coding and non-coding regions for the study, classification, and priority of genomic variants. It provides access to a large genomic annotation array, with a range of frameworks that address various needs and easy methods for setup and study, [22] is a software tool that predicts the possible impact of amino acid replacements in physical and evolutionary comparative factors on the structure and operation of human proteins, [23] models structural changes caused by the replacement of amino acid, [24] determines whether a replacement of an amino acid influences the protein feature depending on the homology of the sequences and the amino acid physical properties. SIFT can be used for non-synonymous polymorphisms or laboratory-induced missense mutations that occur naturally[25], to rapidly categorize the results of SNPs and other forms, such as multiple nucleotide polymorphisms (MNPs) [26] A web-based suite of tools to predict and analyze protein structure, function and mutations. It has advanced remote homology detection methods to build 3D models, predict ligand binding sites and analyze the effect of amino acid variants, e.g., non-synonymous SNPs. But with all those vraiaty of software that investigate the snp effect in gene, protein and disease consequences, still we have shortage in tool that extract and check SNP existence. This is a software for extracting and verifying the presence of an SNP in a specified DNA sequence SNPector. based on various SNP clinically associated databases such as NCBI Clinvar Database, Awesome Database, and PharmGKB, SNPector is constructed. With the support of various Python modules, SNPector links the observed SNP to disease and medication, visualizing the relation between gene, SNP, drug, and phenotype.

3 Aim of work

Current software and web applications depends on given SNP information like SNP position, ID, allele, and gene in which gene located to conclude the effects and phenotype, nevertheless the only available information in some times may be the only the sequence in which SNPs are hidden without knowing any further SNP information or even the name of region from where sequence is taken. SNPector detect SNP with its position on gene, scaffold name, associated disease of phenotype, drug linked to disease, and adverse drug reaction annotation.

4 Material and Method

4.1 Data resources

SNPector use 159,184 SNPs data collected from NCBI, 1,080,551 SNP record from Awesome, and 3,932 SNP from PharmGKB, 3,932 Drug and phenotype record from PharmGKB, and to decrease storage size we used LDlink[27] application program interface (API) to download LDhap file that Calculate population specific haplotype frequencies of all haplotypes observed for a list of variants extracted from given query, LDmatrix file which is matrix of pairwise linkage disequilibrium statistics, LDproxy file that explore proxy and putatively functional variants for a query variant, and SNPchip file that is list of platforms used for detected SNP genotyping. The other data file such BLAST software[28] output file data and data for circos and network manipulated and prepared by python script.

4.2 Running SNPector

SNPector run BLAST locally to find out where given sequence is located, or query, is located on human genome. By locating the query on the genome, SNPector starts collecting all SNPs located within query range from NCBI clinvar database, then compare the allele of each collected SNP with opposite on query and save the record if SNP alternative or mutated allele matches query allele.

4.3 Results preparation

At the end of SNP scanning, the output is about three files each contain the detected SNP according to its database in addition to “BLAST_RESULT.txt” where BLAST output saved. The first file created is “FromNCBI.tsv” containing the detected SNP only, then from SNP identification number in NCBI file SNPector begins to extract the same SNP from other databases file “FromAwesom.tsv” and “FromPharmGKB.tsv”.

4.4 Result post-preparation

after generating the result files, SNPector initiate the preparation of data from which the circos and network will be build. Circos and network scripts accept specific format in order to work correctly

4.5 Visualization

SNPector use different python packages to: [1]construct circos figure that illustrate the other places in genome that have the same SNP properties according to Awesome database, [2]Build a network between gene, SNP, drug, and phenotype to illustrate how far different genes, SNP, Drugs, or disease are shared with each other, and [3]used downloaded data from LDlink to create different figure types to summarize the huge data and show how far SNPs are linked to each other and how they are distributed over the genome.

3 Results and discussion

3.1 SNPector in General

SNPector can extract variant data from given DNA sequence by investigate nucleotide by nucleotide in query and compare it with Clinvar NCBI dataset. Different databases When integrated into SNPector, it is possible to manifest the fluctuations in abundances of SNPs in query comparing with human genome variants originating from next generation sequencing projects. These input data can be then used for generate visualizations of all output data elements. Furthermore, Figure 1 provides the workflow of the tool.

3.2 Implementation

To achieve user-friendly usage, SNPector can be run from terminal with simple command line indicate the parameters to how it works. In figure(1), python3 indicate the program interpeter as it build with python programming language, scan_dna.py refers to the python file in which program is written, -blaston command blast to work and blasting the given sequence against the genome to find where sequence is located while -blastoff tells the program to switch of BLAST and use previously generated results , -modesearch to order SNPector to only find out SNPs located in range of query while -modescan deeply investigate the existence of SNP in query, -circoson to draw circos figure illustrate where SNP with same scores are located, -networkon activate script the link between SNP, disease and drug to produce network html file, -download activate the API to download data for extracted SNPs from LDlink database, -vis for running visualize and plot data to produce figures such in figure(2), and GivenSequence.fasta is the file where user paste its sequence in fasta format. Any of the previous parameter can be deactivated when replaced with -off.

```
~/Desktop/SNPector$ python3 scan_dna.py -blastoff -modesearch -circoson -networkon -download -vis -GivenSequence.fasta
```

Figure 1: command line to run SNPector

3.2 Comparability to SNP effect Tools

SNPector strength in its dependency on different database to annotate the discovered SNPs. Nevertheless, many tools provides SNPs annotation but they are still limited to given information. On the other hand, SNPector provides new tool that extract SNP from naked sequence while nothing other than the nucleotides order is known. And the following table show the option provided by each software.

Software	SNPector	Ensembl Variant Effect Predictor	PolyPhen-2	Missense3D	SIFT	SnpEff	Phyre2
SNP Detection	**Yes**	**No**	**No**	**No**	**No**	**No**	**No**
Disease and Drug annotation	**Yes**	**No**	**No**	**No**	**No**	**No**	**No**
SNP Circos Linkage	**Yes**	**No**	**No**	**No**	**No**	**No**	**No**
Gene, SNP, Drug, and Disease Network	**Yes**	**No**	**No**	**No**	**No**	**Yes**	**No**
3D SNP effect confirmation	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**
Physical and Chemical Investigation	**No**	**No**	**No**	**Yes**	**No**	**No**	**No**
Coding consequences	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
SNP annotation	**Yes**	**Yes**	**Yes**	**No**	**Yes**	**Yes**	**Yes**
SNP Effect	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

3.3 Results Deep Investigation

SNPector provides user with more deeper and visual view figure(1), of SNP as it generate Circos PDF file for each detected SNP illustrate where other SNPs that share the same properties such as: [1]prediction of protein phosphorylation sites, [2]prediction ubiquitination sites in proteins, [3]prediction of protein methylation sites, [4]prediction of sumoylation sites, and [5]predicts

substrates of N-acetyltransferase A in genome which give deeper insight. SNPector also visualize Linkage disequilibrium data using Matplotlib, Pandas, Seaborn, and Numpy python packages to create twelve type of figures that summarize the huge amount of data downloaded from LDlink such as: Dprime which is an indicator of allelic segregation for two genetic variants that ranging from 0 to 1 with higher values indicating tight linkage of alleles where value of 0 indicates no linkage of alleles and value of 1 indicates at least one expected haplotype combination is not observed, R squared (R2) that is a measure of correlation of alleles for two genetic variants, Minor allele frequency that refers to the frequency at which the second most common allele occurs in a given population which is widely used in population genetics studies because it provides information to differentiate between common and rare variants in the population. R2 values range from 0 to 1 with higher values indicating a higher degree of correlation. An R2 value of 0 indicates alleles are independent, whereas an R2 value of 1 indicates an allele of one variant perfectly predicts an allele of another variant. R2 is sensitive to allele frequency. Figure(1) shows example of output images that SNPector can generate.

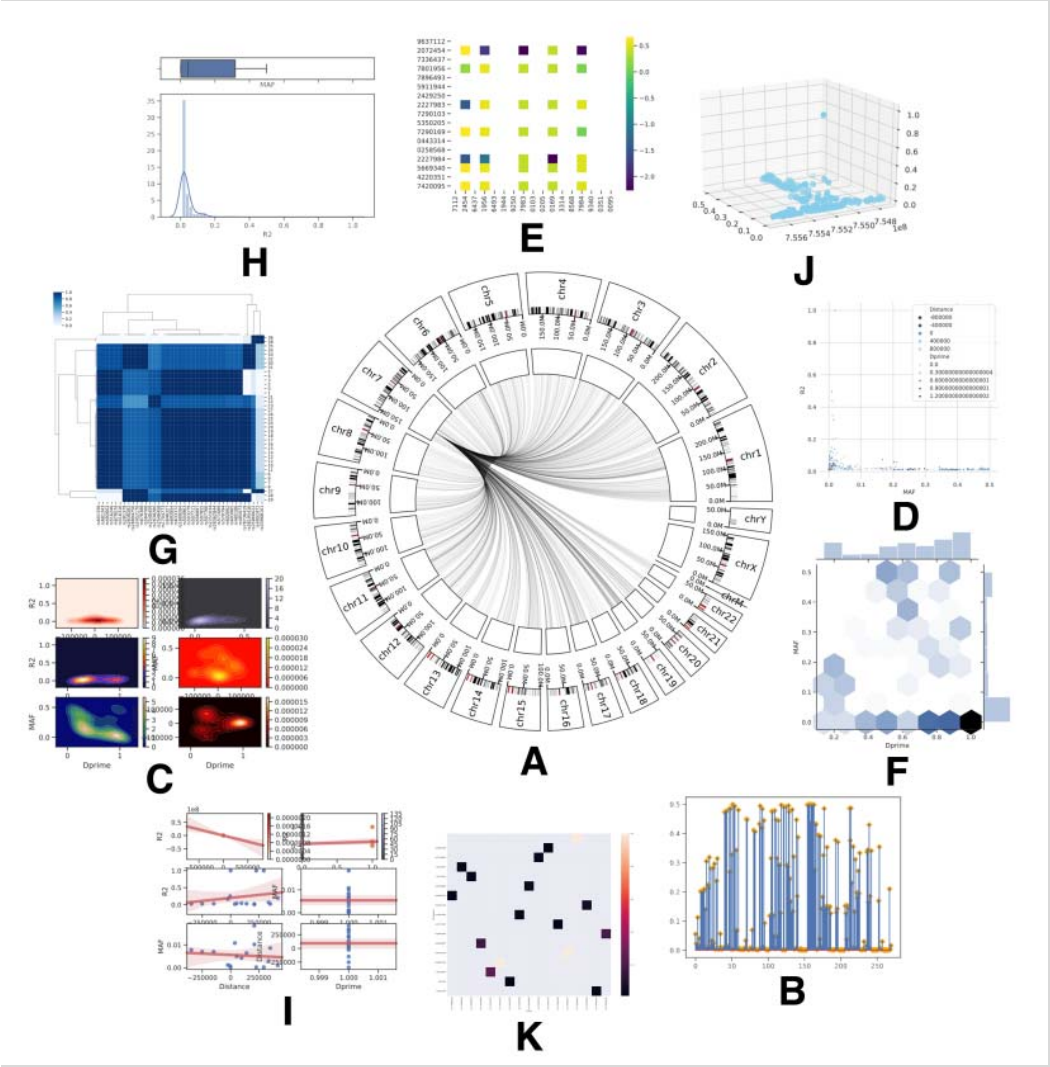


Figure 2: (A) is circos figure and it is in our case used to illustrate where other SNPs that have same proprieties are located. (B) is Lollipop figure that show values each with head to be more distinguishable specially in lower values. (C) Counter Plot between two value creating this colored shade and its mean more contrast means higher value. (D) is Numerical Schematic figure show the distribution between four values by plotting and scaling color contrast according to other to values. (E) is heat map between SNP linkage disequilibrium matrix to show how each two SNPs are linked. (F) is marginal plot combine between column graph and plot both show the relationship between two values. (G) is dendogram with heatmap which show valuable information of how far all SNP are linked to each others. (H) Histogram with box plot to compare visually between two values. (I) plotting illustrate the regression fit of Two plotted value. (J) 3D plot of three values. (K) Annotated heatmap show the plotted value with its number on it with changing in color contrast according to the number.

4 Summary and Conclusion

One of our currently virulence growing catastrophic health problem is the accumulation of genetic disease in our genome, and that make it worse is not detecting it before it occurs ti reduce its severity or even cure it and this hard part of diagnosis and treatment can not be done perfectly without omics sciences and and computer sciences. SNPector combine between computer and omics sciences to provide the user with: in detail explanation of Disease associated SNPs in given query, network of disease and treatment, illustrations of linked disequilibrium detected SNPs on the same query, SNP list of minor allele frequency and R square linked with discovered SNPs in query, and list of SNP genotyping Chips for detected SNPs.

5 References

[1] Chaudhary, Renu, et al. "Role of single nucleotide polymorphisms in pharmacogenomics and their association with humandiseases." Drug metabolism reviews 47.3 (2015): 281-290.

[2] Welter, Danielle, et al. "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations." Nucleic acids research 42.D1 (2013): D1001-D1006.

[3] Stranger, Barbara E., Eli A. Stahl, and Towfique Raj. "Progress and promise of genome-wide association studies for human complex trait genetics." Genetics 187.2 (2011): 367-383.

[4] Schirmer, Markus A., et al. "Relevance of Sp binding site polymorphism in WWOX for treatment outcome in pancreatic cancer." JNCI: Journal of the National Cancer Institute 108.5 (2016).

[5] Wu, Huazhang, et al. "A novel functional TagSNP Rs7560488 in the DNMT3A1 promoter is associated with susceptibility to gastric cancer by modulating promoter activity." PLoS One 9.3 (2014): e92911.

[6] Fan, Hong, et al. "A functional polymorphism in the DNA methyltransferase-3A promoter modifies the susceptibility in gastric cancer but not in esophageal carcinoma." BMC medicine 8.1 (2010): 12.

[7] Zhu, Xiao, et al. "The rs391957 variant cis-regulating oncogene GRP78 expression contributes to the risk of hepatocellular carcinoma." Carcinogenesis 34.6 (2013): 1273-1280.

[8] Xu, Ming, et al. "Functional promoter rs2295080 T> G variant in MTOR gene is associated with risk of colorectal cancer in a Chinese population." Biomedicine & pharmacotherapy 70 (2015): 28-32.

[9] Shao N, Li J, Xu B, Wang Y, Lu X, Feng N. Role of the functional variant (-652T>G) in the XRCC4 promoter in prostate cancer. Mol Biol Rep. 2014; 41:7463–7470.

[10] Rintisch, Carola, et al. "Natural variation of histone modification and its impact on gene expression in the rat genome." Genome research 24.6 (2014): 942-953.

[11] He C, Xu Q, Tu H, Sun L, Gong Y, Liu J, Yuan Y.Polymorphic rs9471643 and rs6458238 upregulate PGC transcription and protein expression in overdominant or dominant models. Mol Carcinog. 2016; 55:586–99.

[12] Yang, Lei, et al. "Effects of a functional variant c. 353T> C in snai1 on risk of two contextual diseases. Chronic obstructive pulmonary disease and lung cancer." American journal of respiratory and critical care medicine 189.2 (2014): 139-148.

[13] Fang, Wenxiang, et al. "The functional polymorphism of NBS1 p. Glu185Gln is associated with an increased risk of lung cancer in Chinese populations: Case–control and a meta-analysis." Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 770 (2014): 61-68.

[14] Griseri, Paola, et al. "A synonymous polymorphism of the Tristetraprolin (TTP) gene, an AU-rich mRNA-binding protein, affects translation efficiency and response to Herceptin treatment in breast cancer patients." Human molecular genetics 20.23 (2011): 4556-4568.

[15] Pagenstecher, Constanze, et al. "Aberrant splicing in MLH1 and MSH2 due to exonic and intronic variants." Human genetics 119.1-2 (2006): 9-22.

[16] Li, Lijuan, et al. "Association between polymorphisms in long non-coding RNA PRNCR1 in 8q24 and risk of colorectal cancer." Journal of Experimental & Clinical Cancer Research 32.1 (2013): 104.

[17] Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RK, Hua Y, Gueroussov S, Najafabadi HS, Hughes TR, Morris Q, Barash Y, Krainer AR, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015; 347:1254806.

[18] Xu, Qian, et al. "A new polymorphism biomarker rs629367 associated with increased risk and poor survival of gastric cancer in chinese by up-regulated miRNA-let-7a expression." PloS one 9.4 (2014): e95249.

[19] Xu, Qian, et al. "The interaction effects of pri-let-7a-1 rs10739971 with PGC and ERCC6 gene polymorphisms in gastric cancer and atrophic gastritis." PloS one 9.2 (2014): e89203.

[20] Xu, Qian, et al. "Promoter polymorphisms in trefoil factor 2 and trefoil factor 3 genes and susceptibility to gastric cancer and atrophic gastritis among Chinese population." Gene 529.1 (2013): 104-112.

[21] Dunna, Nageswara Rao, et al. "Association of thymidylate synthase 5'-UTR 28bp tandem repeat and serine hydroxymethyltransfarase C1420T polymorphisms with susceptibility to acute leukemia." Asian Pac J Cancer Prev 15.4 (2014): 1719-23.

[22] He, Huiling, et al. "Multiple functional variants in long-range enhancer elements contribute to the risk of SNP rs965513 in thyroid cancer." Proceedings of the National Academy of Sciences 112.19 (2015): 6128-6133.

[23] McLaren, William, et al. "The ensembl variant effect predictor." Genome biology 17.1 (2016): 122.

[24] Adzhubei, Ivan, Daniel M. Jordan, and Shamil R. Sunyaev. "Predicting functional effect of human missense mutations using PolyPhen 2." Current protocols in human genetics 76.1 (2013): 7-20.

[25] Ittisoponpisan, Sirawit, et al. "Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated?." Journal of molecular biology 431.11 (2019): 2197-2212.

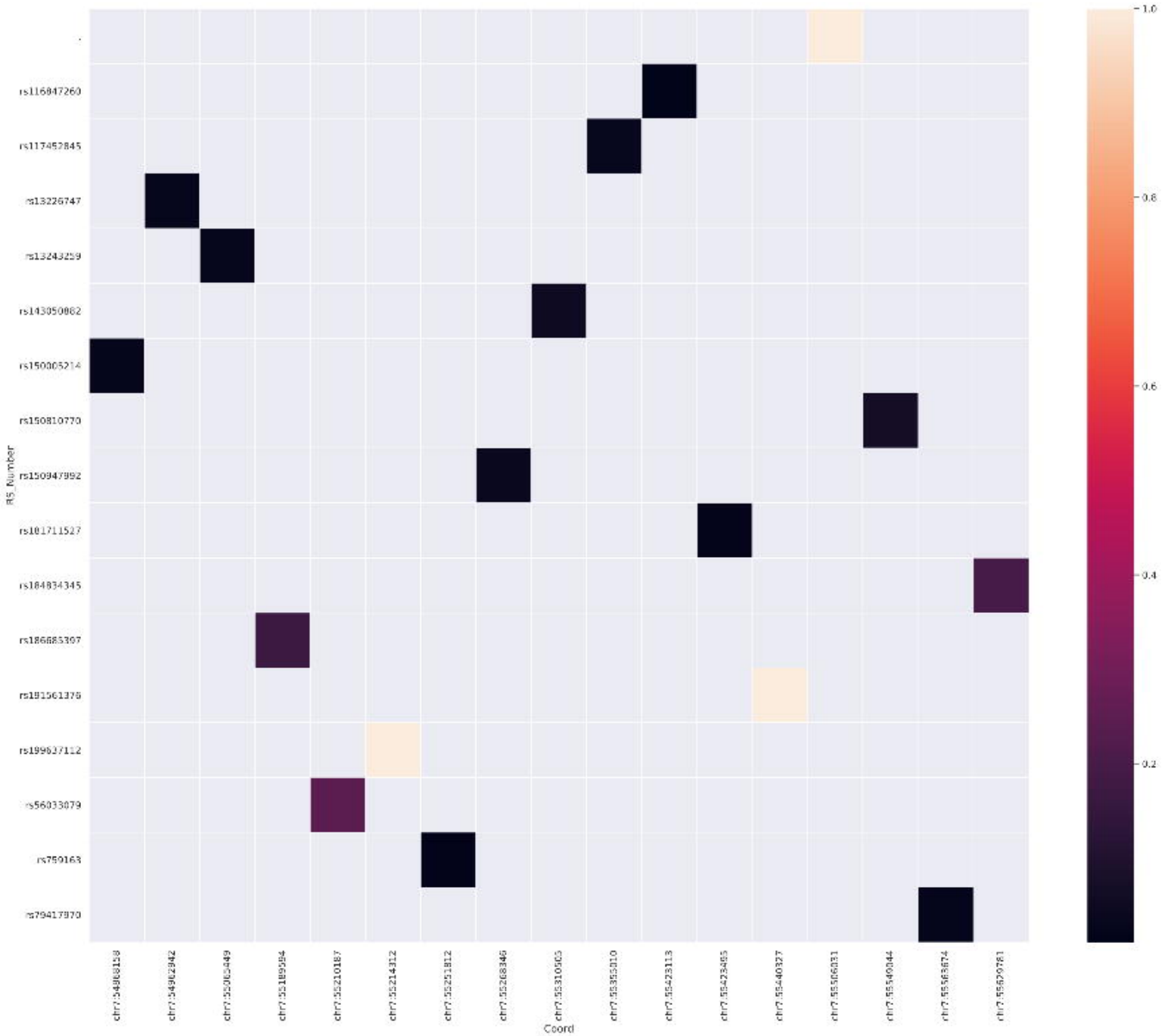
[26] Ng, Pauline C., and Steven Henikoff. "SIFT: Predicting amino acid changes that affect protein function." Nucleic acids research 31.13 (2003): 3812-3814.

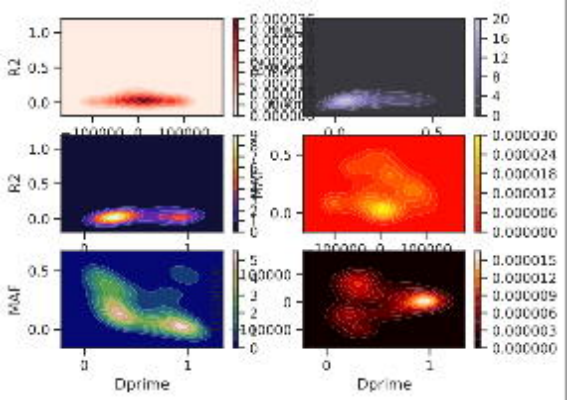
[27] Cingolani, Pablo, et al. "A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3." Fly 6.2 (2012): 80-92.

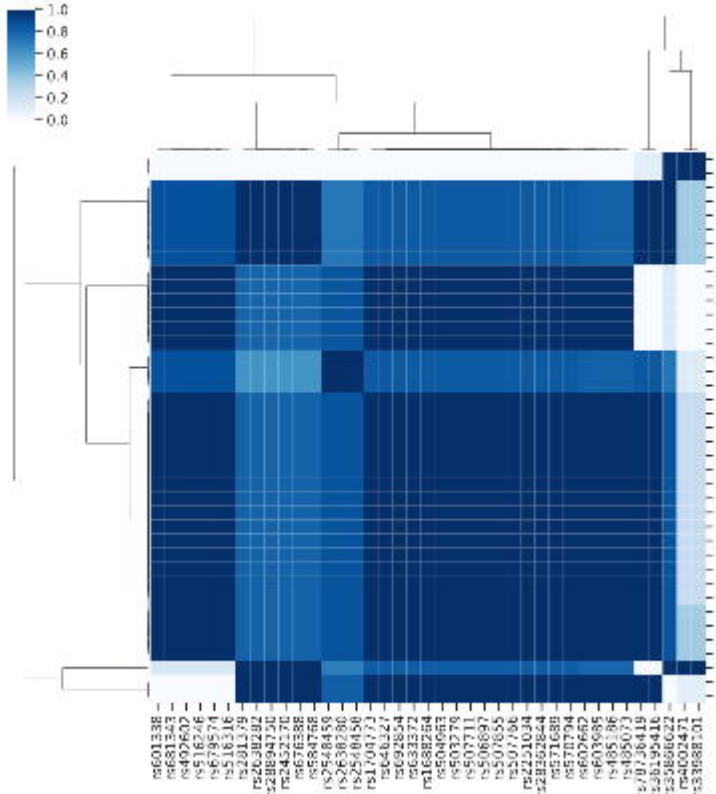
[28] Kelley, Lawrence A., et al. "The Phyre2 web portal for protein modeling, prediction and analysis." Nature protocols 10.6 (2015): 845.

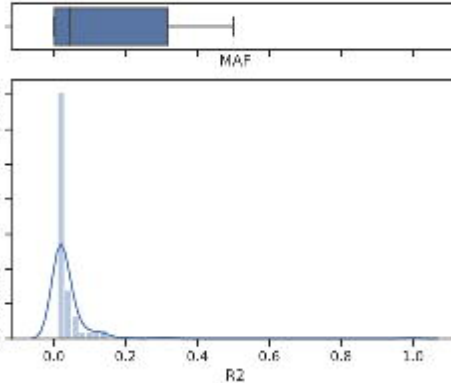
[29] Machiela, Mitchell J., and Stephen J. Chanock. "LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants." Bioinformatics 31.21 (2015): 3555-3557.

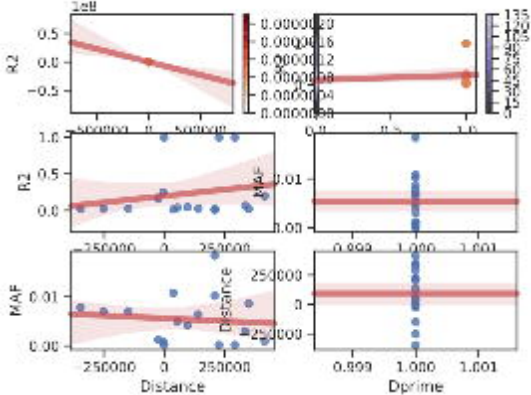
[30] Altschul, Stephen F., et al. "Basic local alignment search tool." Journal of molecular biology 215.3 (1990): 403-410.

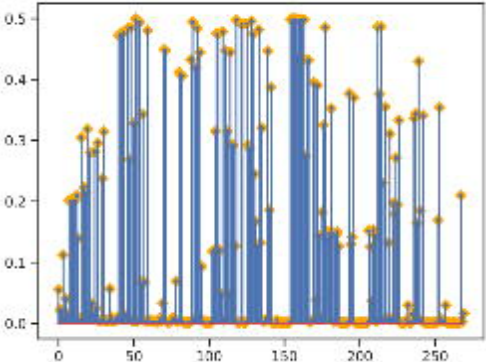

.
rs116847260
rs117452845
rs13226747
rs13243259
rs143050862
rs150005214
rs150810770
rs150947992
rs101711527
rs184834345
rs186685307
rs191561376
rs199637112
rs56033079
rs759163
rs79417970
RS_Number
chr7:54908158
chr7:54962942
chr7:55005449
chr7:55189584
chr7:55210187
chr7:55214312
chr7:55251812
chr7:55268346
chr7:55310905
chr7:55355810
chr7:55423113
chr7:55423465
chr7:55440327
chr7:55506831
chr7:55540044
chr7:55568674
chr7:55629781
Coord
1.0
0.8
0.6
0.4
0.2

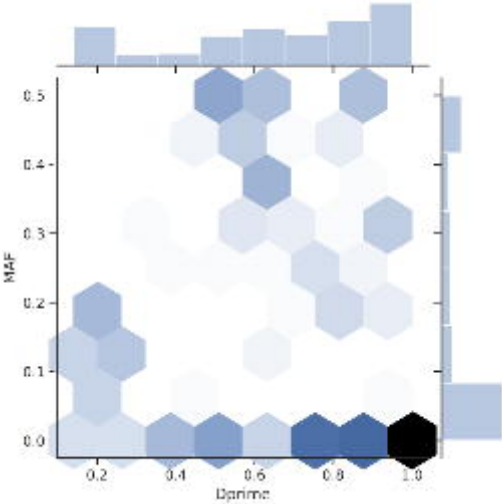

R2
MAF
Dprime
Dprime

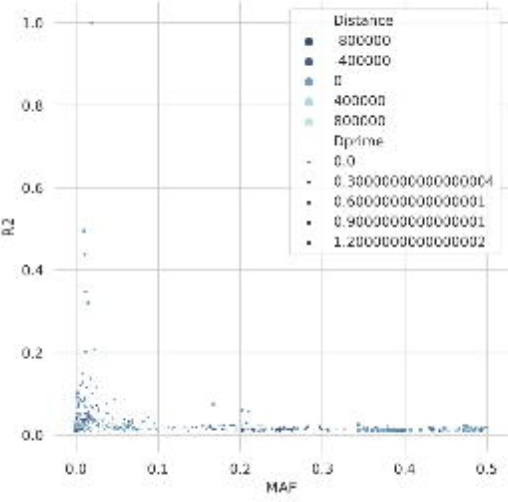

1.0
0.8
0.6
0.4
0.2
0.0
36
38
35
34
33
32
30
31
5
4
3
2
0
1
21
22
23
27
26
25
24
20
19
18
17
16
15
14
13
12
10
11
6
9
7
8
37
28
29
rs601338
rs681343
rs492602
rs516246
rs679574
rs516316
rs201379
rs2638282
rs2889475
rs2452170
rs676388
rs584768
rs2548459
rs2638280
rs2548458
rs1704773
rs646327
rs692854
rs633372
rs1688264
rs504963
rs503279
rs507711
rs506697
rs507655
rs507766
rs2251034
rs2836284
rs571689
rs570794
rs602662
rs603985
rs485186
rs485073
rs7873649
rs3619541
rs3586622
rs4002471
rs3388101


MAF
35
30
25
20
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0
R2


1e8
R2
MAF
Distance
Dprime


0.5
0.4
0.3
0.2
0.1
0.0
0
50
100
150
200
250


0.5
0.4
0.3
0.2
0.1
0.0
MAF
0.2
0.4
0.6
0.8
1.0
Dprime

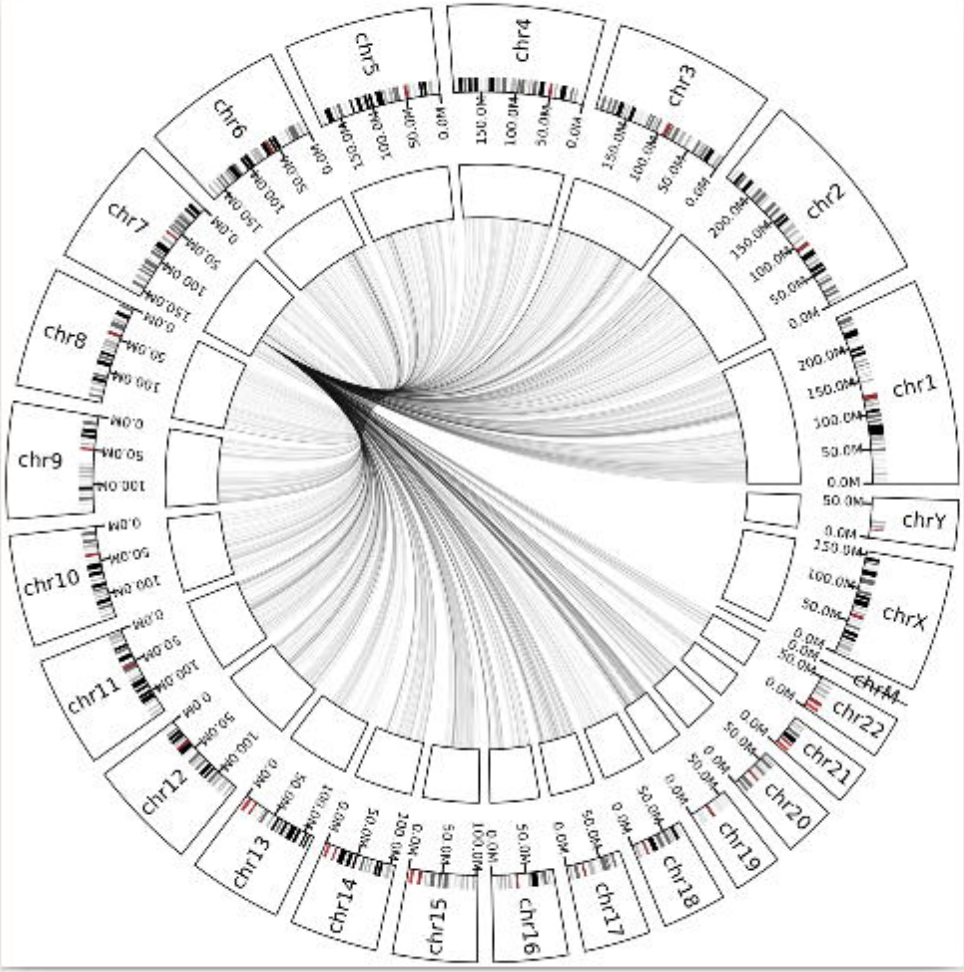

Distance
800000
400000
0
400000
800000
Dprime
0.0
0.3000000000000004
0.6000000000000001
0.9000000000000001
1.2000000000000002
R2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
MAF

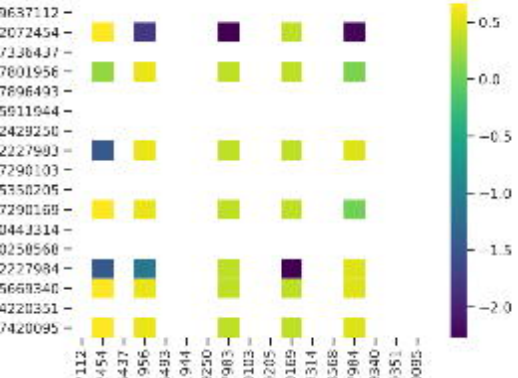
Circos(rs1800072)_Fig1080549.pdf

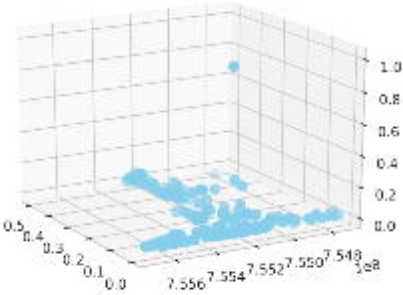
1 of 1

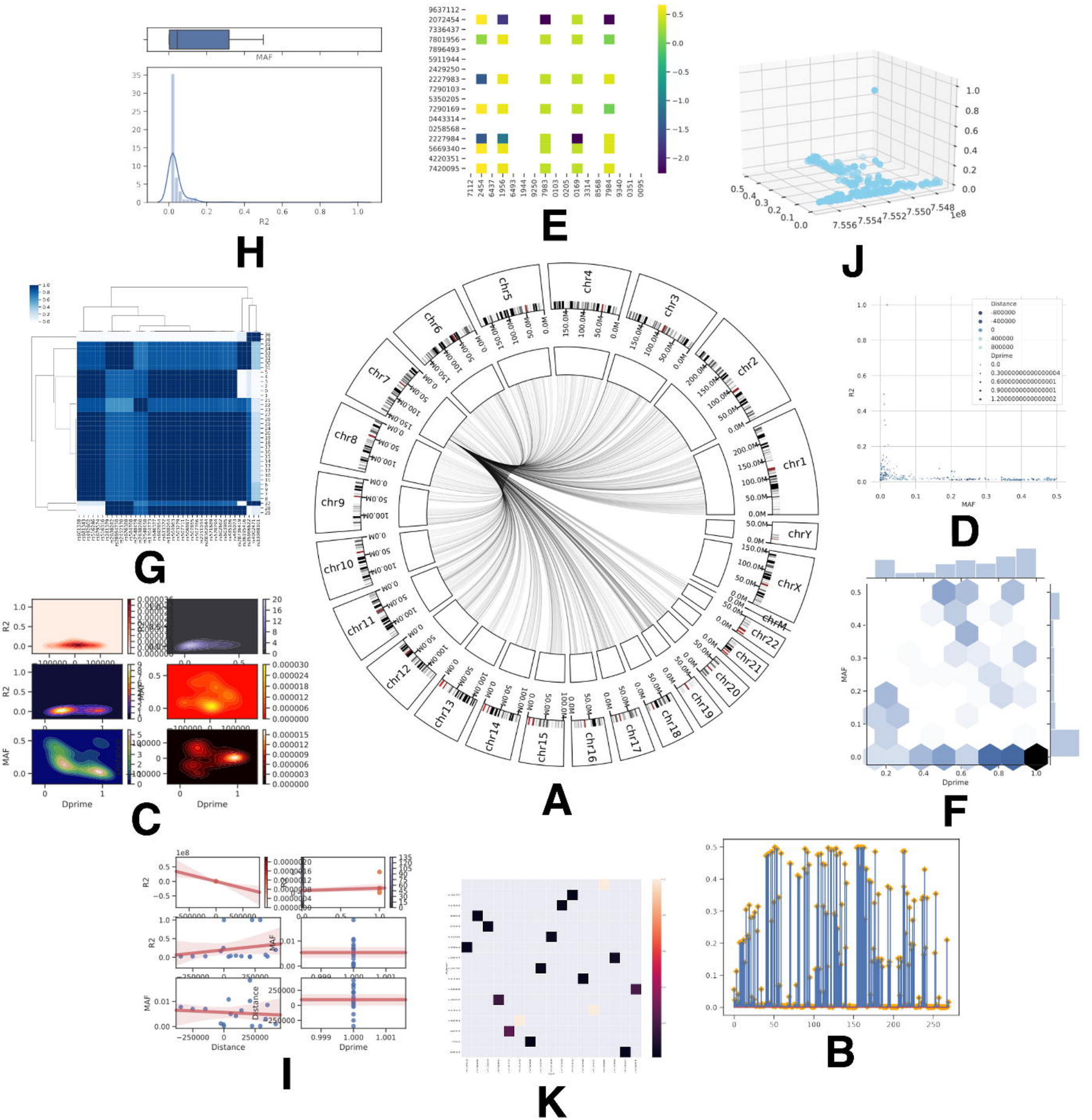
100.37%

Thumbnails




0637112
2072454
7338437
7801956
7896493
5811944
2429250
2227983
7290103
5350205
7290169
0443314
0250560
2227984
5669340
4220351
7420095
0.5
0.0
−0.5
−1.0
−1.5
−2.0


1.0
0.8
0.6
0.4
0.2
0.0
0.5
0.4
0.3
0.2
0.1
0.0
7.556
7.554
7.552
7.550
7.548
1e8


H
E
J
G
A
D
C
F
I
K
B
MAF
R2
Dprime
Distance
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
chrY
chrM