bioRxiv preprint doi: https://doi.org/10.1101/834580; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SNPector: SNP inspection tool for diagnosing gene pathogenicity and drug response in a naked sequence

Peter T. Habib[1*], Alsamman M. Alsamman[2], Ghada A. Shereif[3] 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

[3] Faculty of Pharmacy, October 6 University, Giza, Egypt

* To whom correspondence should be addressed: p.habib911@gmail.com or a.hamwieh@cgiar.org

Abstract

Due to the ability to diagnose diseases early and evaluate the effectiveness of medicinal drugs, single nucleotide polymorphism (SNP) identification receives significant interest. Detection and diagnosis of genetic variation through skill-less computational tools would help researchers reducing the severity of such health complications and improving the well-tailored therapies using discovered and previously known information. We introduce SNPector, which is a standalone SNP inspection software could be used to diagnose gene pathogenicity and drug reaction in naked genomic sequences. It identifies and extracts gene-related SNPs, and reports their genomic position, associated phenotype disorder, associated diseases, linkage disequilibrium, in addition to various drug reaction information. SNPector detects and verifies the existence of an SNP in a given DNA sequence based on different clinically relevant SNP databases such as NCBI Clinvar database , Awesome database , and PharmGKB and generates highly informative visualizations of the recovered information.

Introduction

In recent years, the number of cases of genetically originated diseases has increased, alarming the world and sparking interest in the development of precision medicine using molecular biomarkers. Single nucleotide polymorphism (SNP), the most common genetic difference among individuals, occurs in the human genome. These randomized modifications in DNA bases cause alterations in protein sequence residues of amino acids, thus altering their functions which lead to different disease conditions in individuals (1) . Several of these SNPs have been identified as disease-related genetic markers that have been used to recognize genes responsible for a particular disease in humans (2).

Distinguishing the evidence and the interpretation of a rich range of markers will be necessary to relate the major alterations in the SNPs and to discover their connection in the progression of disease . Clarification of the phenotypic-associative mechanisms for these

bioRxiv preprint doi: https://doi.org/10.1101/834580; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

variations is therefore vital for comprehending the sub-atomic subtleties of disease start and for developing novel therapeutic methods (3,4).

Although SNPs may exist in various areas of the gene, such as promoters, introns, 5′-and 3′ UTRs, to date, most research has focused on disease-associated SNPs (daSNPs) in coding regions or exons, especially non-synonymous SNPs, which may alter the biochemical ability of encoded proteins. In turn, altering gene promoters impact gene expression by changing transcription, binding transcription factor, methylation of DNA and modifications of histones. As a consequence, changes in gene expression, their impact on disease susceptibility, and drug responses can differ depending on the location of the SNP (5–7).

With the expansion of genetic variants, different software could be used to generate new knowledge to support disease diagnosis and drug response studies and to develop new biomarkers for disease identification and drug customization. In this regard, a number of software applications have been developed in the last few years to classify, prioritize and evaluate the impact of genomic variants.

For example, the Ensemble Variant Effect Predictor offers access to a large range of genomic annotations, with a variety of frameworks that answer different needs, with easy setup and evaluation methods (8) . Similarly, SnpEff categorizes the results of genome sequence variations , annotate variants according to their genomic location and estimates the coding effects. Depending on genome annotation, it is possible to predict coding effects such as non-synonymous or synonymous substitution of amino acids, stop codon gains or losses,start codon gains or losses, or frame changes (9) .

On the other hand, PolyPhen-2 assesses the potential impact of the genetic substitution of amino acids on the basis of physical, evolutionary comparative factors and model structural changes. Based on these profiles, the probability of a missene mutation becoming dangerous is measured on the basis of a combination of all these properties (10) . In like manner, SIFT calculates whether the substitution of amino acids affects protein activity, based on the homology of sequences and the physical properties of amino acids. It may be used for non-synonymous polymorphisms and laboratory-induced missense mutations that naturally occur, to effectively classify the effects of SNPs as well as other types, including multiple nucleotide polymorphisms (MNPs) (11) .

Moreover, Phyre2 is a web-based suite of tools for predicting and analyzing protein structure, function and mutations. It has sophisticated remote homology identification methods to build 3D models, anticipate ligand binding sites, and evaluate the effect of amino acid variants, e.g. non-synonymous SNPs (12) . Missense 3D uses the user-provided UniProt ID of the query protein, wild-type residue and substitution and other information to generate PDB residue mapping and predict the substitution effect on the 3D protein structure (13) .

To conclude the effect and possible phenotype of SNP, these software and web applications require minimum information such as SNP genomic position, SNP ID, allele form, and/or gene name. Acquiring these information require using different computational tools , extensive time and some analysis skills. Most of the time, only gene sequences are available in which the SNPs are hidden without any additional information.

Int this regard, we introduce SNPector, which is a standalone SNP inspection software could be used to diagnose gene pathogenicity and drug reaction in naked genomic sequences. It identifies and extracts gene-related SNPs, and reports their genomic position, associated phenotype disorder, associated diseases, linkage disequilibrium, in addition to various drug

bioRxiv preprint doi: https://doi.org/10.1101/834580; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

reaction information. It detects and verifies the existence of an SNP in a given DNA sequence based on different clinically relevant SNP databases such as NCBI Clinvar database (14)□, Awesome database (15)□, and PharmGKB (16)□. Lastly, it connects identified SNPs, related diseases and drugs, and produces numerous visualization figures to explain these relationships with the support of different Python modules.

Design and implementation

SNPector was written using Pythpn3 programming language as a standalone package and could be run on different operating systems platforms supported with Python 3.x compilers. To achieve user-friendly usage, the SNPector can be operated from a console through simple command line (**Figure 1**) .

SNPector use different SNPs record information collected from NCBI (159,184 record), Awesome (1,080,551 record), and PharmGKB (17) (3,932 record). Ldlink is an online tool can be used to assess linkage imbalance (LD) throughout ancestral populations and is a popular method to exploring population-specific genetic framework and functionally navigation disease susceptibility areas (18) . An Application Program Interface (API) has been programmed to download an LDhap file containing linkage disequilibrium statistics and potentially functional variants for a query variant resulted from sequence.

SNPector starts by running BLAST (19) software locally to find out the genomic location of a given DNA sequence on human genome. If successfully, it retrieves SNPs record located within query genomic range using NCBI ClinVar database. According to retrieved records from database , the detected SNPs in user-provided queries are marked as wild or mutated. Additionally, more information regarding detecting SNPs records will be retrieved from different implemented databases. These information will be used to generate different illustration figures.

If process is successfully finished, SNPector will generate four different files; (A) Text file contains the output BLAST result, where the genomic location of the user-defined sequences is predicted, (B) Tab delimited file contains SNPs retrieved NCBI database located in the same regions, (C) Two files regarding these specific SNPs information retrieved from Awesome and PharmGKB databases, (D) different figures depicting SNPs with a similar mutation effect to the detected SNPs located on other genomic regions , SNP linkage disequilibrium, the relationship between SNP, drug, and phenotype (**Figure 2**).

```
python3 scan_dna.py -blastoff -modsearch -circoson -network -download -vis GivenSequence.fasta
  (A)      (B)        (C)        (D)       (E)      (F)       (G)     (H)        (I)
```

Figure (1): The SNPector command line structure. A) Python3 compiler, B) scan_dna.py : The program main script, C) -blaston / -blastoff: in order to, initiate BLAST procedure for provided

bioRxiv preprint doi: https://doi.org/10.1101/834580; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

sequence against the genome to locate where the sequence is situated, if the blastoff is chosen the will use previous blast results, D) -modesearch / -modescan: Figure out SNPs that are located in the scope of query using different modes, E) -circoson : Draw circos figure to illustrate where SNP with same properties/effect are located, F) -networkon : in oreder to link between SNPs, diseases and drugs and produces network HTML file, G) -download: The activatation the API to download data for identified SNPs from LDlink database, H) -vis: in order to produce different figures and plots, I) GivenSequence.fasta : The user-provided sequence in fasta file format. Any of the previous parameter can be deactivated when replaced with -off.

Results and discussion

SNPector can collect and retrieve information from the user-provided DNA sequence in the simplest way possible. By integrating different databases into SNPector, it is possible to detect the fluctuations in the abundance of SNPs in query through the comparison with known variants of human genome. Such steps are accompanied by the use of online and verified sources to gather previously published details regarding target genomic regions and to generate highly informative visualizations of the recovered information.

Many tools, however, provide SNPs annotation, but they are still limited to the information provided (**Table 1**) . SNPector, on the other hand, provides a new technique that extracts SNP from a naked sequence with no prior information. In addition, other benefit of SNPector is to annotate the discovered SNPs based on various known databases. Moreover, SNPector provides user with more deeper and visualization figures, highlighting other SNPs with similar mutation effect on protein phosphorylation, ubiquitination, methylation, or sumoylation sites, and predicts substrates of N-acetyltransferase.

Additionally, SNPector provides the ability to visualize obtained information about the linkage disequilibrium of detected SNPs using various python packages such as Matplotlib (20), generating a number of figures that summarize vast amounts of previously published data indicating SNPs allelic segregation, association, minor allele frequency. **Figure 2** shows an example of illustrations that can be generated through SNPector.

Table 1 : Comparison between SNPector and published SNP annotation tool.

Software	SNPector	Ensembl VEP	PolyPhen-2	Missense3D	SIFT	SnpEff	Phyre2
SNP detection from sequence	Yes	No	No	No	No	No	No
Disease and drug annotation	Yes	No	No	No	No	No	No
SNPs relationship analysis	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

bioRxiv preprint doi: https://doi.org/10.1101/834580; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

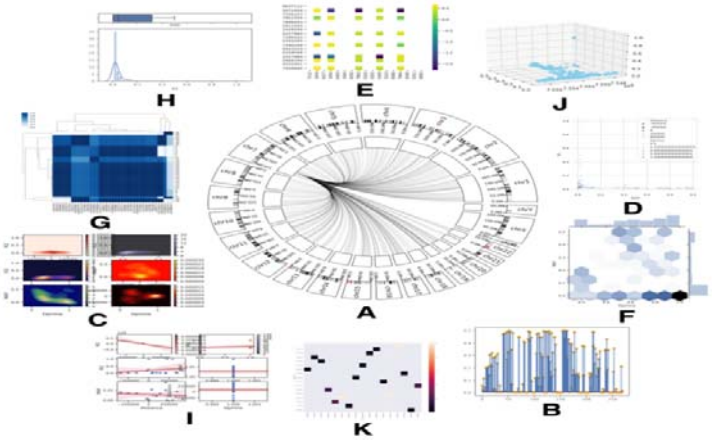


Figure 2: (A) Circos illustrate where other SNPs that have same proprieties are located. (B) Lollipop that show values each with head to be more distinguishable specially in lower values. (C) Counter Plot between two value creating this colored shade in which more contrast means higher value. (D) Numerical schematic show the distribution between four values by plotting and scaling color contrast according to other to values. (E) Heat map between SNP linkage disequilibrium matrix to show how each two SNPs are linked. (F) Marginal plot combine between column graph and plot both show the relationship between two values. (G) Dendogram with heatmap which show 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.

Conclusion

One of the currently growing medical research paradigms is the diagnosis of genetic virulence that accumulates in our genome causing catastrophic health problems. Detection and diagnosis of genetic variation through skill-less computational tools would help researchers reducing the severity of such health complications and improving the well-tailored therapies using discovered and previously known information.

SNPector provides and detects all available information about the disease-related SNPs in the given query with minimum user-provided information. It connects between different available information and produce various illustrations depicting SNP related diseases and treatment network, linked disequilibrium, minor allele frequency, similar SNPs with the same mutation effect and other information.

References

bioRxiv preprint doi: https://doi.org/10.1101/834580; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Chaudhary R, Singh B, Kumar M, Gakhar SK, Saini AK, Parmar VS, et al. Role of single nucleotide polymorphisms in pharmacogenomics and their association with human diseases. Drug Metab Rev. 2015;47(3):281–90.

2. Kong J, Zhu J, Keyser UF. Single molecule based SNP detection using designed DNA carriers and solid-state nanopores. Chem Commun. 2017;53(2):436–9.

3. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2013;42(D1):D1001--D1006.

4. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011;187(2):367–83.

5. Schirmer MA, Lüske CM, Roppel S, Schaudinn A, Zimmer C, Pflüger R, et al. Relevance of Sp binding site polymorphism in WWOX for treatment outcome in pancreatic cancer. JNCI J Natl Cancer Inst. 2016;108(5).

6. Fan H, Liu D, Qiu X, Qiao F, Wu Q, Su X, et al. A functional polymorphism in the DNA methyltransferase-3A promoter modifies the susceptibility in gastric cancer but not in esophageal carcinoma. BMC Med. 2010;8(1):12.

7. Rintisch C, Heinig M, Bauerfeind A, Schafer S, Mieth C, Patone G, et al. Natural variation of histone modification and its impact on gene expression in the rat genome. Genome Res. 2014;24(6):942–53.

8. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol. 2016;17(1):122.

9. 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 (Austin). 2012;6(2):80–92.

10. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013;76(1):7–20.

11. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4.

12. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10(6):845.

13. Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? J Mol Biol. 2019;431(11):2197–212.

14. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res.

bioRxiv preprint doi: https://doi.org/10.1101/834580; this version posted November 25, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2015;44(D1):D862--D868.

15. Yang Y, Peng X, Ying P, Tian J, Li J, Ke J, et al. AWESOME: a database of SNPs that affect protein post-translational modifications. Nucleic Acids Res. 2018;47(D1):D874--D880.

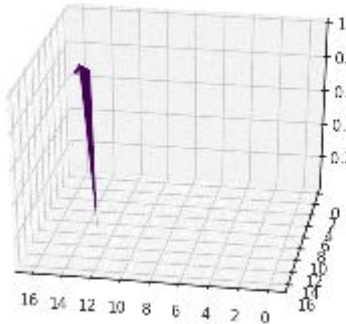
16. Thorn CF, Klein TE, Altman RB. PharmGKB: the pharmacogenomics knowledge base. In: Pharmacogenomics. Springer; 2013. p. 311–20.

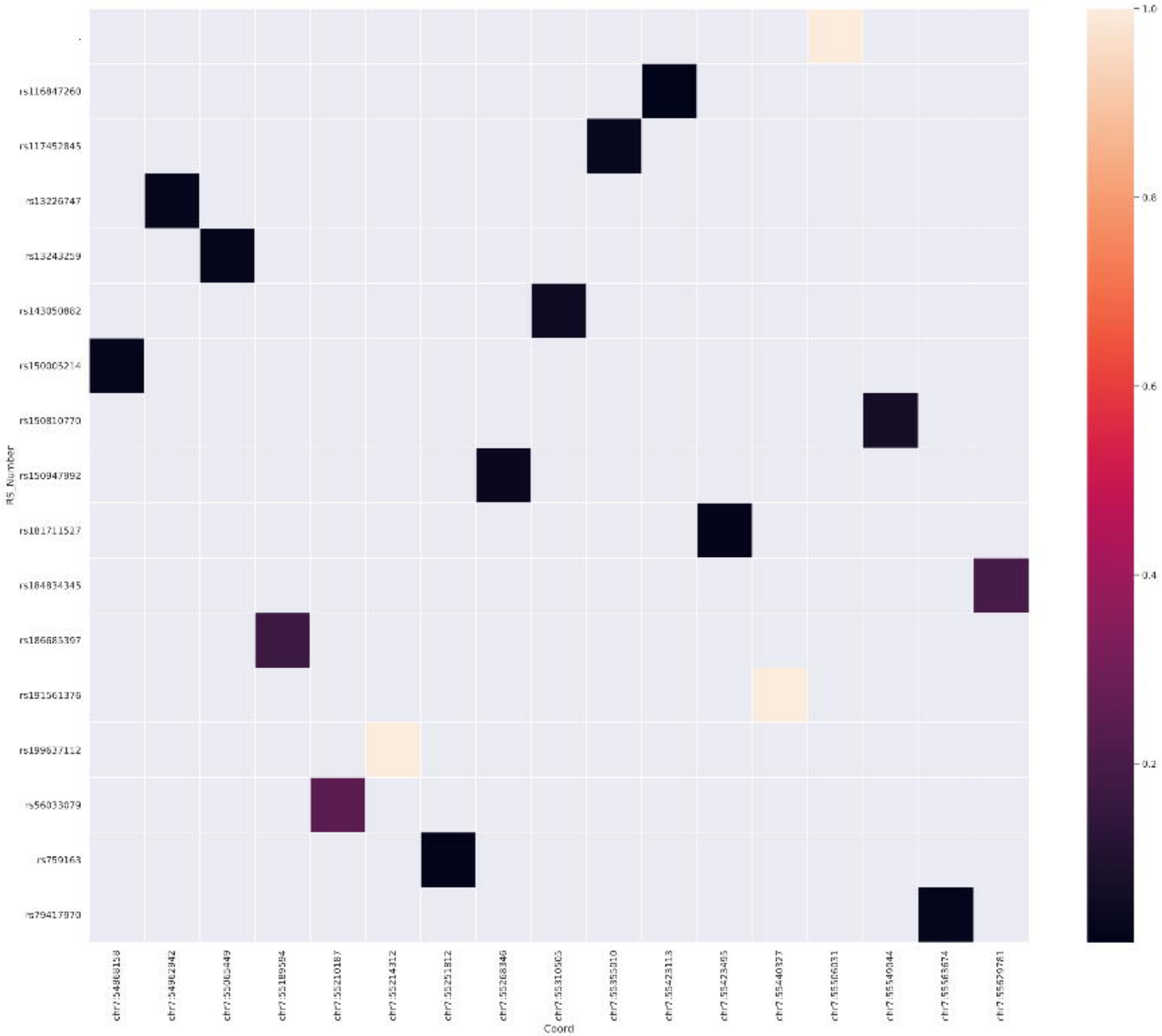
17. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, et al. PharmGKB: the pharmacogenetics knowledge base. Nucleic Acids Res. 2002;30(1):163–5.

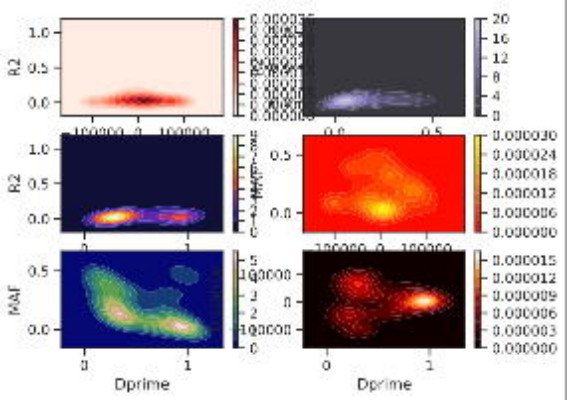
18. Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics. 2015;31(21):3555–7.

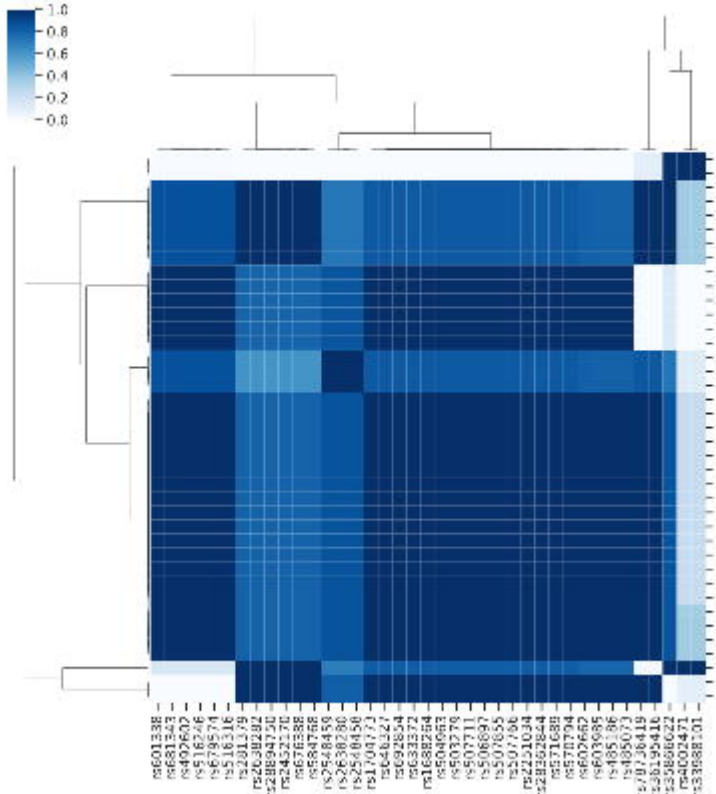
19. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.

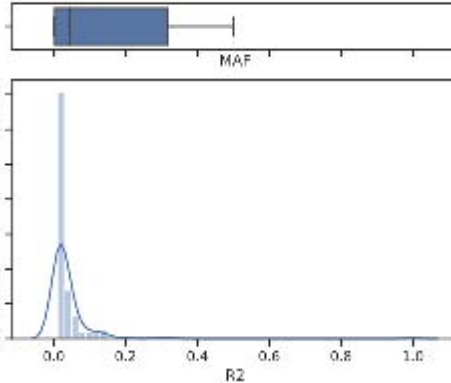
20. Habib, Peter Tharwat, Alsamman Mahmoud Alsamman, and Aladdin Hamwieh. "BioAnalyzer: Bioinformatic Software of Routinely Used Tools for Analysis of Genomic Data." Biotechnology 10 (2019): 33-41.

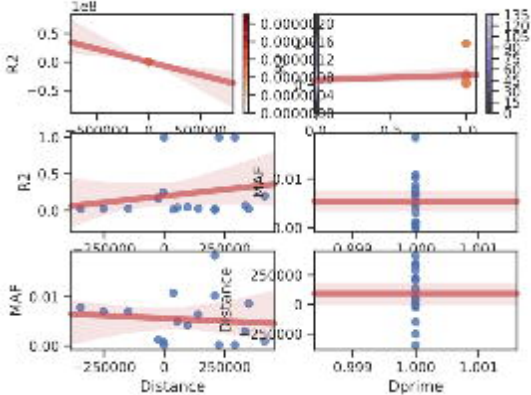

1.0
0.8
0.6
0.4
0.2
0
2
4
6
8
10
12
14
16
16 14 12 10 8 6 4 2 0

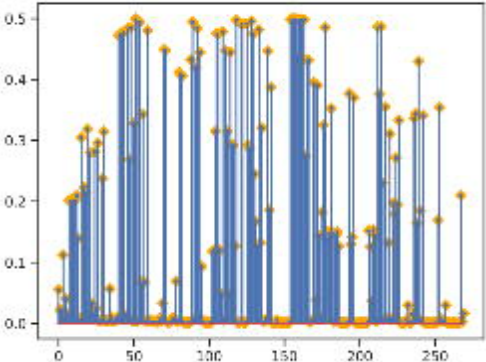

.
rs116847260
rs117452845
rs13226747
rs13243259
rs143050862
rs150005214
rs150810770
rs150947992
rs101711527
rs184834345
rs186685307
rs191561376
rs199637112
rs56033079
rs759163
rs79417970
RS_Number
chr7:54908158
chr7:54962942
chr7:55005449
chr7:55189584
chr7:55210187
chr7:55214312
chr7:55251812
chr7:55268346
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

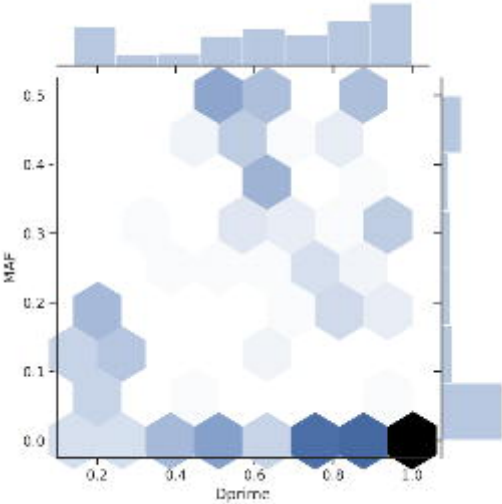

R2
MAF
Dprime
Dprime

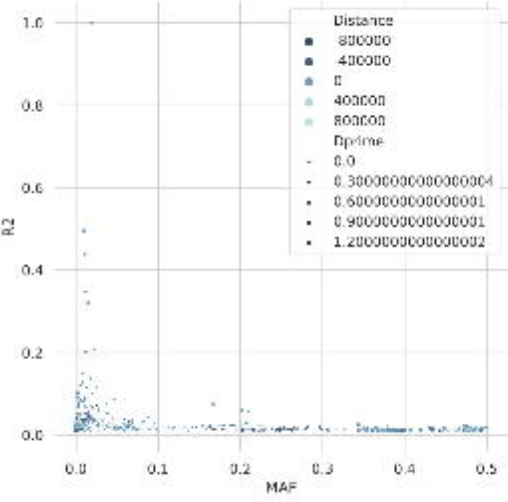

1.0
0.8
0.6
0.4
0.2
0.0
36
38
35
34
33
32
30
31
5
4
3
2
0
1
21
22
23
27
26
25
24
20
19
18
17
16
15
14
13
12
10
11
6
9
7
8
37
28
29
rs601338
rs681343
rs492602
rs516246
rs679574
rs516316
rs201379
rs2638282
rs2889475
rs2452170
rs676388
rs584768
rs2548459
rs2638280
rs2548458
rs1704773
rs646327
rs692854
rs633372
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

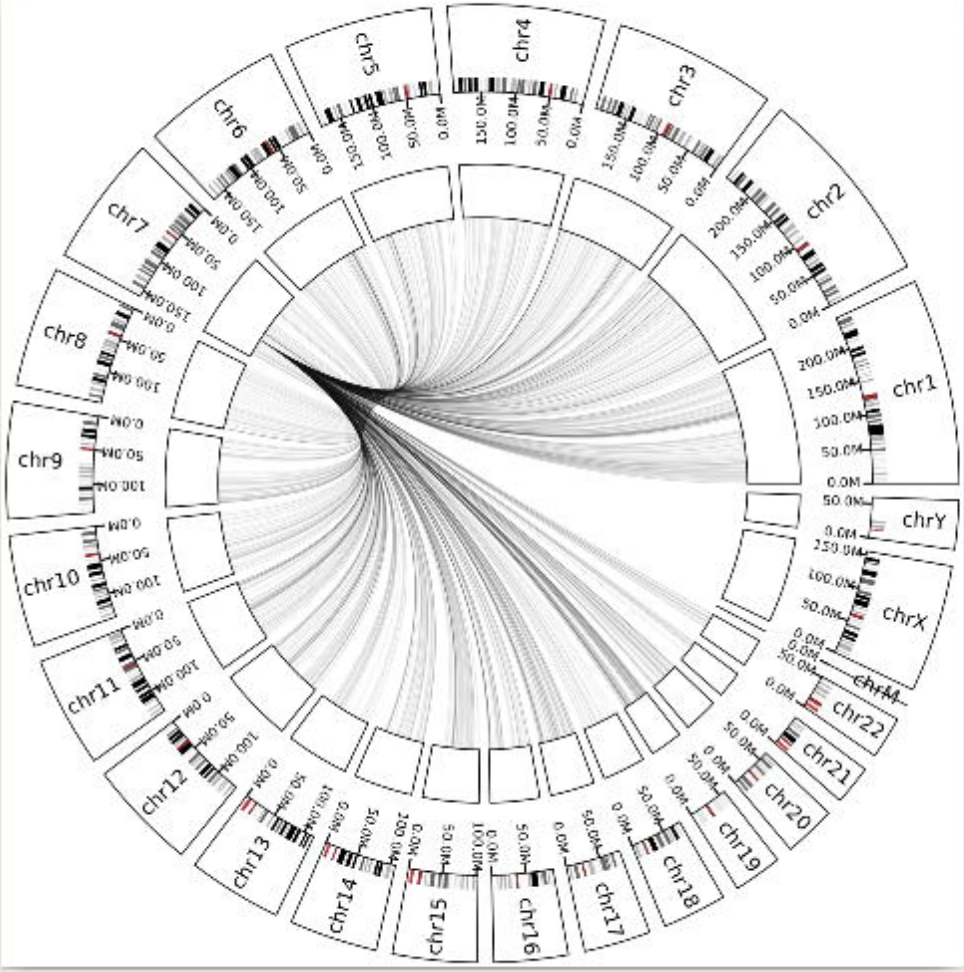

Distance
800000
400000
0
400000
800000
Dprime
0.0
0.3000000000000004
0.6000000000000001
0.9000000000000001
1.2000000000000002
R2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
MAF

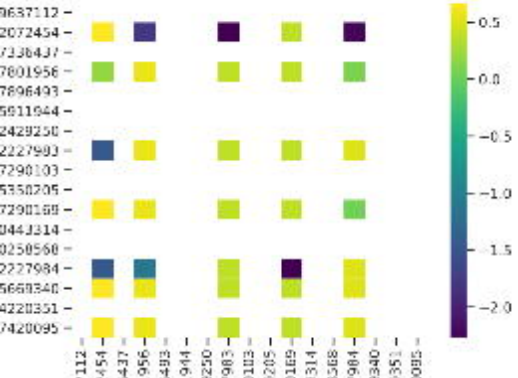
Circos(rs1800072)_Fig1080549.pdf

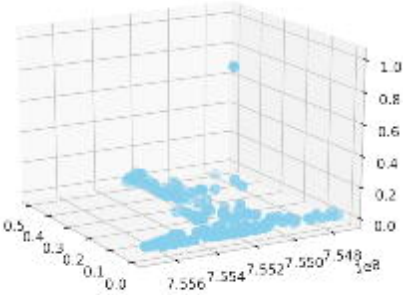
1 of 1

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