CandiHap: a toolkit for haplotype analysis for sequence of samples and fast identification of candidate causal gene(s) in genome-wide association study

Xukai Li*李旭凯, Zhiyong Shi 石志勇, Qianru Qie 郄倩茹, Jianhua Gao 高建华, Xingchun Wang 王兴春 &

Yuanhuai Han 韩渊怀

Abstract

Genome-wide association study (GWAS) is widely used to identify genes involved in plants, animals and human complex traits. Generally, the identified SNP is not necessarily the causal variant, but it is rather in linkage disequilibrium (LD). One key challenge for GWAS results interpretation is to rapidly identify causal genes and provide profound evidence on how they affect the trait. Researches want to identify candidate causal variants from the most significant SNPs of GWAS in any species and on their local computer, while to complete these tasks are to be time-consuming, laborious and prone to errors and omission. To our knowledge, so far there is no tool available to solve the challenge for GWAS data very quickly. Based on the standard VCF (variant call format) format, CandiHap is developed to fast preselection candidate causal SNPs and gene(s) from GWAS by integrating LD result, SNP annotation, haplotype analysis and traits statistics of haplotypes. Investigators can specify genes or linkage regions based on GWAS results, linkage disequilibrium (LD), and predicted candidate causal gene(s). It supported Windows, Mac and Linux computers and servers in graphical interface and command line, and applied to any other plant, animal or bacteria species. The source code of CandiHap tool is freely available at https://github.com/xukaili/CandiHap

Introduction

With next generation sequencing (NGS), genome sequencing is becoming inexpensive and routine, and the obtention of large numbers of SNPs is convenient. Genome-wide association study (GWAS) has become established in medical, biological and agricultural research to elucidate the genetic basis of phenotypic traits such as disease or economically important features (Visscher et al. 2012; Visscher et al. 2017). SNP can alter a protein directly (non-synonymous SNPs, stop gained or stop lost SNPs, frameshift SNPs or SNPs in splice sites) or it can implicate gene expression if SNP are located in regulatory regions. From a huge number of genome-wide variants, a GWAS investigation generally identifies a few SNPs that are statistically significantly associated with some trait. As GWAS serves as initializations of future genetic and mechanism study of complex traits, one of the key challenges of GWAS data interpretation is to identify causal SNPs (the SNPs that affect trait) and provide profound evidence and hypothesis on the mechanism through which they affect the trait (McCarthy and Hirschhorn 2008).

There are some researches focusing on inferring candidate causal SNPs from the most significant (SNPs with *P* value below certain threshold) (Hindorff et al. 2009; Li et al. 2012) and prioritizing the most significant SNPs

by linkage disequilibrium (LD) analysis and functional SNP annotation (Adzhubei et al. 2010; Johnson et al. 2008; Kumar et al. 2009; Lee and Shatkay 2008; Mi et al. 2010; Saccone et al. 2010; Schmitt et al. 2010; Xu and Taylor 2009; Yuan et al. 2006; Yue et al. 2006). However, most existing tools are web-based tools or command-line for human studies, severely limiting those widely use. In fact, more researches want to identify candidate causal variants affect traits from the most significant SNP of GWAS in their own species (not only for human, but also for any species) and on their local computer (for maintain secrecy), while to complete these tasks are to be time-consuming, laborious and prone to errors and omission and not a convivial interface. A feasible proposal to address the above problems is to development a software to fast find candidate causal variants and gene(s). So far, there is no tool available to provide a solution for GWAS data very quickly. CandiHap aims to provide an open source to facilitate researchers to identify the candidate causal SNPs and gene(s) of traits and to guide future genetic and mechanism study.

Methods

Variant filtering and annotation

The variants of VCF file was further filtered using the VCFtools (Danecek et al. 2011) (ver. 0.1.15). The SNPs and indels were considered valid for the study if they met the following requirements: (1) two alleles only; (2) exclude sites on the basis of the proportion of missing data >0.9 (defined to be between 0 and 1, where 0 allows sites that are completely missing and 1 indicates no missing data allowed); (3) minor allele frequency ≥0.05; and (4) mean depth values ≥5. SNPs that did not meet these four criteria were excluded from the study. All identified SNPs that passed quality screening were further annotated with ANNOVAR (ver. 2015 Dec 14) based on the gene annotation of the reference genome (Wang et al. 2010). In practical application, users can adjust the above parameters for a study. When a VCF file is submitted, ANNOVAR is computed to rapidly categorize the effects of variants in the reference genome sequence. ANNOVAR annotates variants based on their genomic locations (annotated genomic locations can be intronic, exonic or intergenic) and predicts coding effects (mainly synonymous or non-synonymous amino-acid replacement). The process can be applied to any other plant, animal or bacteria species, by providing the genome file and its GFF (generic feature format) annotation file.

Software development

CandiHap is written in Perl 5 (v 5.26, https://www.perl.org), R (v 3.5, https://www.r-project.org) and Python 2.7 (https://www.python.org), which supported Windows, Mac and Linux computers and servers in graphical interface and command lines. Graphics are created by R. The graphical user interface is written in electron, which is freely available and registration is not required. Besides the graphical interface software, users can run CandiHap through command lines by using the Linux or Mac. For a given SNP that was found significant in a GWAS, runtime is ~1 min for a set of 400 samples and ~3 million SNPs. The CandiHap tool is freely available at https://github.com/xukaili/CandiHap.

General statistics

Using perl and R, the results provides and displays various statistics about the haplotypes such as annotation statistics, type of variations, number of varieties, varieties ID and its phenotype, average and SD (standard deviation) of phenotype. A boxplot of gene showed significant difference in the phenotype of each haplotype.

Methods of the graphical interface

Results

Process overview

Here, we propose the CandiHap local software, which supported Windows, Mac and Linux computers and servers in graphical interface and command lines. An overview of the process is presented Figure 1. Starting from a VCF file as entry point, the process first annotates the variants using an annotated reference genome to produce a new VCF file from which variants and genotyping data can be then mined and sent into a series of modules in charge of various processes. User has then the possibility to analyze variants either at the genome level or at the gene level. The GWAS result of genomic regions (Fig. 1a) and linkage disequilibrium (LD) can be defined by entering the limits, the application will loop and process these region genes.

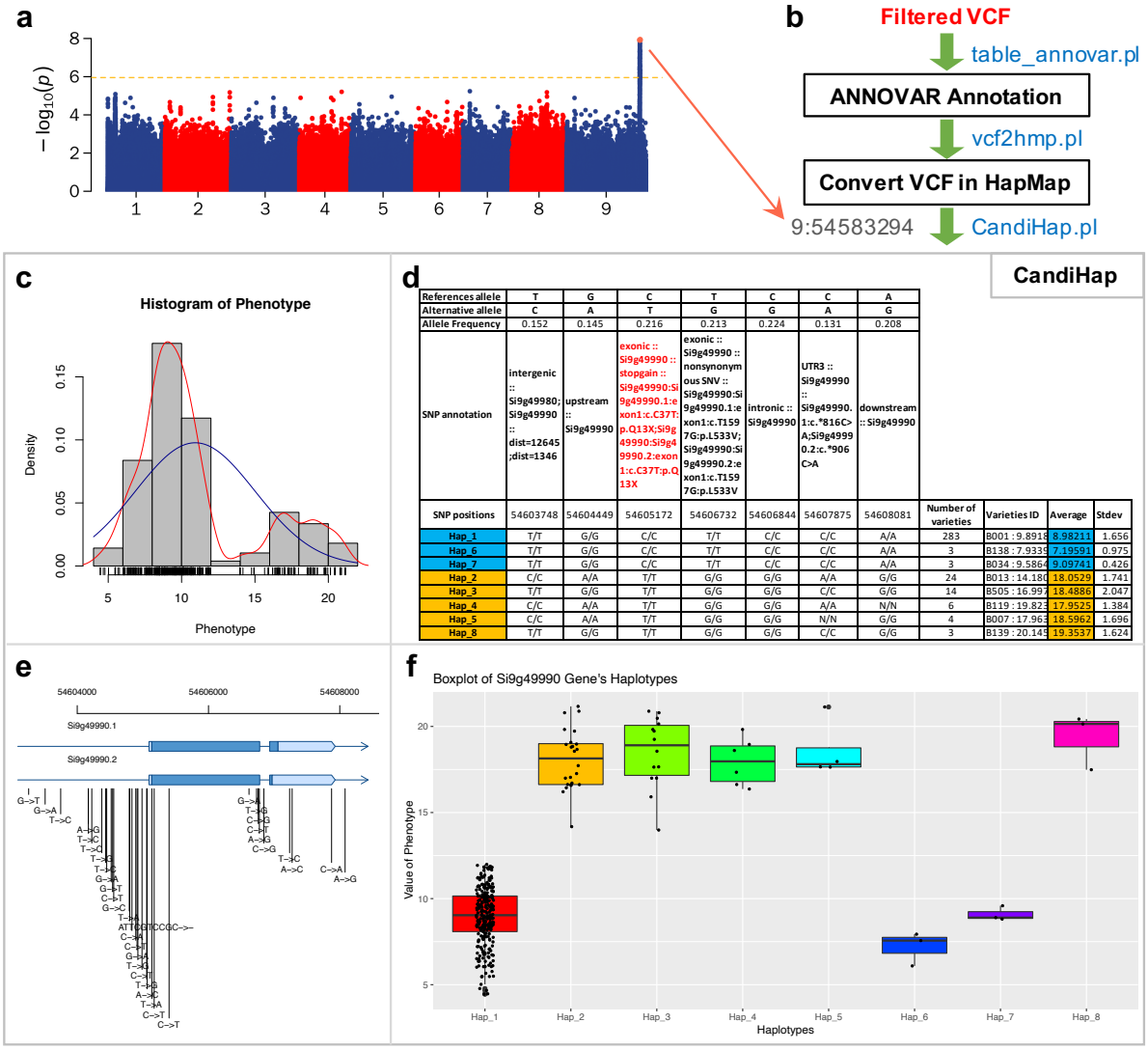


Fig. 1 Overview of the CandiHap process. (a) A GWAS result. (b) General schema of the process. (c) The histogram of phenotype. (d) The statistics of haplotypes. (e) Gene structure and SNPs of key gene. (f) Boxplot of key gene's haplotypes.

Input data and running procedure

The CandiHap implements a three-stage analysis (Fig. 1b). The first stage is to annotate the VCF file for GWAS by ANNOVAR (table_annovar.pl). The second stage is to convert the txt result of annovar to hapmap format (vcf2hmp.pl). The third stage required input data of hapmap file, GFF file of your reference genome, the phenotype data, the linkage disequilibrium (LD), and a most significant SNPs position of GWAS result. If users only want to run one gene, they can input the vcf, phenotype, gff and gene ID. Besides the graphical interface software, users can run CandiHap through command lines by using the Linux, Mac and DOS.

There are mainly three steps included in the CandiHap analytical through command lines, and the test data files can freely download at https://github.com/xukaili/CandiHap.

1. To annotate the vcf by ANNOVAR:

1.1 gffread test.gff -T -o test.gtf

1.2 gtfToGenePred -genePredExt test.gtf si_refGene.txt

1.3 retrieve_seq_from_fasta.pl --format refGene --seqfile genome.fa si_refGene.txt --outfile si_refGeneMrna.fa

1.4 table_annovar.pl test.vcf ./ --vcfinput --outfile test --buildver si --protocol refGene --operation g -remove

2. To convert the txt result of annovar to hapmap format (0.1 means the minor allele frequency (MAF)):
   perl vcf2hmp.pl test.vcf test.si_multianno.txt 0.1
3. To run CandiHap:
   perl GWAS_LD2haplotypes.pl ./test.gff ./haplotypes.hmp ./Phenotype.txt 50kb 9:54583294
   Or to run CandiHap by one gene:
   perl CandiHap.pl ./haplotypes.hmp ./Phenotype.txt ./test.gff Si9g49990

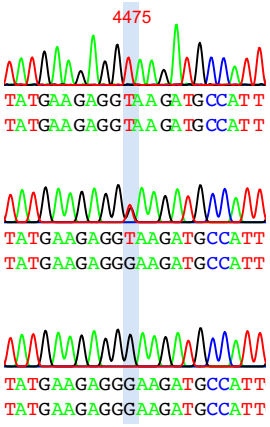
For the graphical user interface, …….

Output and analyzing a GWAS investigation

The output includes a txt file of haplotypes with detailed information and three pdf files of figures (Fig. 1c-f). The result of haplotypes includes References allele, Alternative allele, Allele Frequency, SNP annotation, SNP positions and haplotypes (Fig. 1d). The information for each haplotype also includes Number of varieties, Varieties ID and its phenotype, Average and SD of phenotype (Fig. 1d).

As an example, we investigated a GWAS result of foxtail millet (Unpublished). The result of this GWAS includes ~3679 K GWAS SNP *P*-values and 531 SNPs with *P*-value < 9.42×10^{-7}, and LD is 50 kb. We studied all SNPs that are in the LD 50 kb region of 9:54583294, that is the most significant SNPs (*P*-value = 1.23×10^{-8}). CandiHap identified one candidate causal gene (*Si9g49990*) (Fig. 1d). SNP 9:54605172 is in LD (50 kb), which is with genome-wide significance in the original GWAS (*P*-value = 1.03×10^{-7}), and it is a stop gain (Fig. 1d). The boxplot of *Si9g49990* showed significant difference in the phenotype of each haplotype between Hap 1, 2, 6 and Hap 3, 4, 5, 7, 8, 9 (Fig. 1f). The results of other genes in the LD region are not shown because of limited space. User can run the test data at https://github.com/xukaili/CandiHap/tree/master/test_data to check those results.

a



b

References allele	CA	G	G	G	A	Number of samples	Sample ID
Alternative allele	C	C	C	T	AG		
SNP positions	4243	4271	4273	4475	4887		
Hap_1	CA/CA	G/G	G/G	T/T**	A/A	36	B1, B10, B11, B14, B15, B16, B17, B18, B19, B20, B21, B22, B23, B24, B25, B26, B28, B29, B3, B30, B31, B32, B33, B34, B35, B36, B37, B38, B4, B41, B42, B5, B6, B7, B8, B9
Hap_5	CA/CA	G/C	G/G	T/T**	A/A	6	B12, B13, B27, B39, B40, B44
Hap_9	CA/CA	C/C**	G/G	T/T**	A/A	1	B43
Hap_2	CA/CA	G/G	G/G	G/T	A/A	24	W-1, W-11, W-13, W-18, W-20, W-21, W-22, W-31, W-32, W-34, W-35, W-37, W-41, W-43, W-45, W-46, W-48, W-8, W-9, W101, W54, W72, W86, W97
Hap_4	CA/CA	G/C	G/C	G/T	A/A	11	W-14, W-15, W-17, W-23, W-24, W-25, W-26, W-47, W-6, W85, W91
Hap_7	CA/CA	G/G	G/C	G/T	A/A	2	W-3, W98
Hap_8	CA/CA	G/G	G/C	G/T	A/AG	1	W-7
Hap_11	CA/CA	G/G	C/C**	G/T	AG/AG**	1	W71
Hap_12	CA/CA	G/G	G/G	G/T	AG/AG**	1	W-40
Hap_3	CA/CA	G/G	G/G	G/G	A/A	11	GBS, W-10, W-16, W-28, W-33, W-36, W-38, W-39, W-44, W87, W99
Hap_6	CA/CA	G/C	G/C	G/G	A/A	4	W-30, W-4, W-5, W93
Hap_10	C/C**	G/C	G/C	G/G	A/A	1	W79
Hap_13	CA/CA	G/G	G/C	G/G	A/A	1	W-12

Fig. 1 The haplotype analysis in Sanger ab1 files. (a) PeakTrace of ab1 images of three genotypes. (b) The statistics of haplotypes.

Discussion

In order to solve the challenge for GWAS data interpretation, our CandiHap tool is developed to identify candidate causal SNPs and gene(s) from GWAS by integrating linkage disequilibrium (LD) analysis, SNP annotation, haplotype analysis and traits statistics of haplotypes. CandiHap is a flexible and user-friendly toolkit, that provides a rapidly solution form GWAS result to candidate causal gene(s), and it will help researchers to derive candidate causal gene(s) for complex traits study. At the time of writing, there is no tool that performs the same function as CandiHap.

The CandiHap could be widely used for the available GWAS investigations. CandiHap supported Windows, Mac and Linux computers and servers in graphical interface and command line, and applied to any other plant, animal or bacteria species. It should be noted that CandiHap is not intended to be used to predict true causal SNPs and gene(s) since for complex traits. So the outputs of CandiHap are candidate causal SNPs and gene(s). An important application of the CandiHap results is to allow investigators to test ‘a priori’ hypothesis concerning pathways by using candidate causal SNPs as the practical starting point.

Intergenic SNPs are SNPs that are located at least 5 kb up- or downstream of a gene. In general, they are not associated with a gene and not located in a known regulatory region. We set a strict default parameter in CandiHap. The parameter limit mapping SNPs to 2000 bp upstream and 500 bp downstream of gene. The default settings ensure that the result is based on the association signals in gene(s) and with statistical significance. Users may also adjust the parameter in ‘CandiHap.pl’.

In the future, CandiHap will be regularly updated, and extended to fulfill more functions with more user-friendly options.

References

Adzhubei IA et al. (2010) A method and server for predicting damaging missense mutations Nat Methods 7:248-249 doi:10.1038/nmeth0410-248

Danecek P et al. (2011) The variant call format and VCFtools Bioinformatics (Oxford, England) 27:2156-2158 doi:10.1093/bioinformatics/btr330

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits Proceedings of the National Academy of Sciences of the United States of America 106:9362-9367 doi:10.1073/pnas.0903103106

Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap Bioinformatics (Oxford, England) 24:2938-2939 doi:10.1093/bioinformatics/btn564

Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm Nature protocols 4:1073-1081 doi:10.1038/nprot.2009.86

Lee PH, Shatkay H (2008) F-SNP: computationally predicted functional SNPs for disease association studies Nucleic acids research 36:D820-D824 doi:10.1093/nar/gkm904

Li M-X, Yeung JMY, Cherny SS, Sham PC (2012) Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets Human Genetics 131:747-756 doi:10.1007/s00439-011-1118-2

McCarthy MI, Hirschhorn JN (2008) Genome-wide association studies: potential next steps on a genetic journey Hum Mol Genet 17:R156-R165 doi:10.1093/hmg/ddn289

Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium Nucleic acids research 38:D204-D210 doi:10.1093/nar/gkp1019

Saccone SF et al. (2010) SPOT: a web-based tool for using biological databases to prioritize SNPs after a genome-wide association study Nucleic acids research 38:W201-W209 doi:10.1093/nar/gkq513

Schmitt AO, Aßmus J, Bortfeldt RH, Brockmann GA (2010) CandiSNPer: a web tool for the identification of candidate SNPs for causal variants Bioinformatics (Oxford, England) 26:969-970 doi:10.1093/bioinformatics/btq068

Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery Am J Hum Genet 90:7-24 doi:10.1016/j.ajhg.2011.11.029

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation Am J Hum Genet 101:5-22 doi:10.1016/j.ajhg.2017.06.005

Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data Nucleic acids research 38:e164-e164 doi:10.1093/nar/gkq603

Xu Z, Taylor JA (2009) SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies Nucleic acids research 37:W600-W605 doi:10.1093/nar/gkp290

Yuan H-Y et al. (2006) FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization Nucleic acids research 34:W635-W641 doi:10.1093/nar/gkl236

Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies BMC Bioinformatics 7:166-166 doi:10.1186/1471-2105-7-166

Acknowledgments

We thank Yibo Li (Huazhong Agricultural University, China) for fruitful discussion and all our colleagues and friends in the Shanxi Agricultural University who helped us test the tool and provided the valuable suggestions. We thank the anonymous reviewers from the Molecular Breeding journal for improving it through their valuable comments and suggestions. We are grateful to people that interact regularly for the improvements of the system.

Funding

This work was funded by the Youth Fund Project on Application of Basic Research Project of Shanxi Province (201901D211362), the Scientific and Technologial Innovation Programs of Shanxi Agricultural University (2017YJ27) and Graduate Education Innovation Project of Shanxi Province (2019SY228).

Author information

Xukai Li and Zhiyong Shi contributed equally to this work.

Affiliations

College of Life Sciences, Shanxi Agricultural University, Taigu, 030801, China

Xukai Li, Zhiyong Shi, Jianhua Gao, Xingchun Wang

College of Agriculture, Shanxi Agricultural University, Taiyuan, 030801, China

Qianru Qie, Yuanhuai Han

Contributions

Xukai Li sourced funding; Xukai Li designed research; QQ performed lab work; Xukai Li and Zhiyong Shi provided project resources and undertook project management; Xukai Li and Qianru Qie wrote the manuscript; all the authors read and approved the final manuscript.

Corresponding authors

Correspondence to Xukai Li.

Ethics declarations

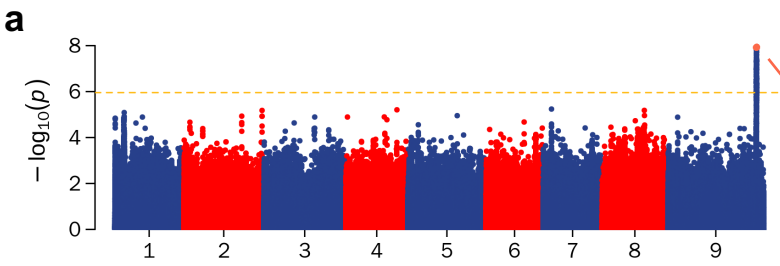
Conflict of interest

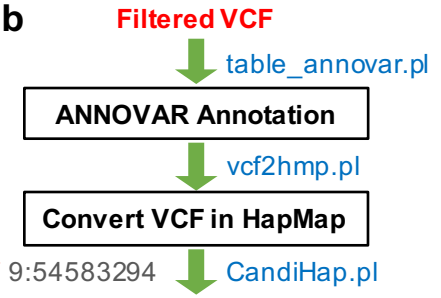
The authors declare that they have no conflict of interest.

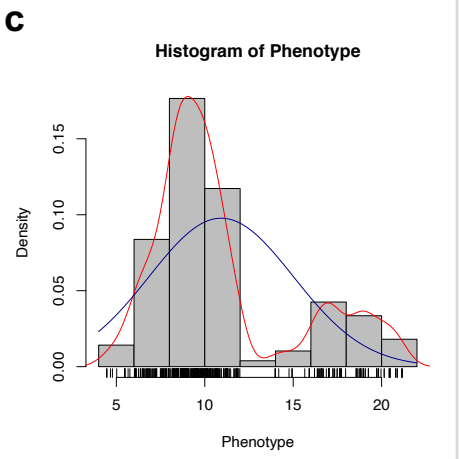
Additional information

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.





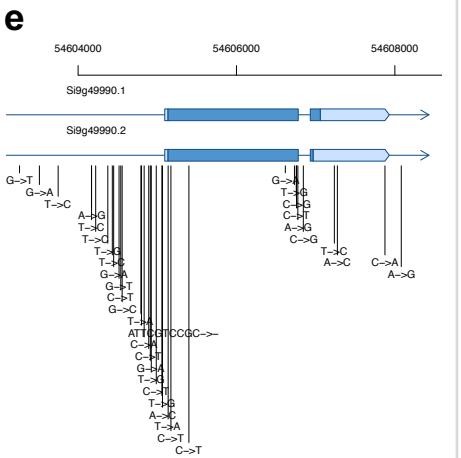


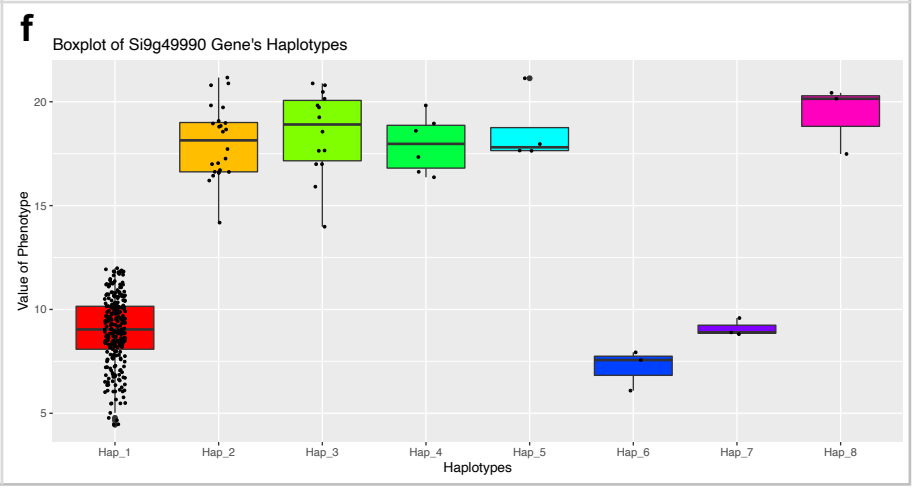
CandiHap

d

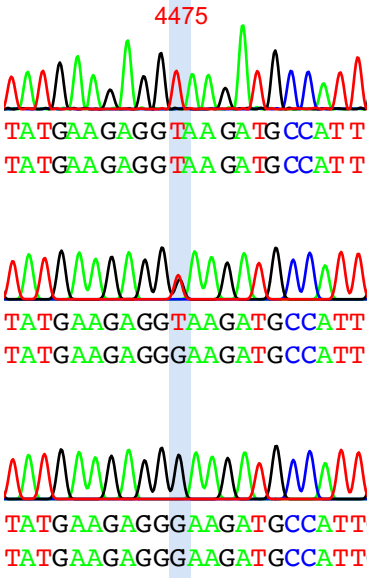
References allele	T	G	C	T	C	C	A
Alternative allele	C	A	T	G	G	A	G
Allele Frequency	0.152	0.145	0.216	0.213	0.224	0.131	0.208
SNP annotation	intergenic :: Si9g49980; Si9g49990 :: dist=12645 ;dist=1346	upstream :: Si9g49990	exonic :: Si9g49990 :: stopgain :: Si9g49990:Si9g49990.1:exon1:c.C37T:p.Q13X;Si9g49990:Si9g49990.2:exon1:c.C37T:p.Q13X	exonic :: Si9g49990 :: nonsynonymous SNV :: Si9g49990:Si9g49990.1:exon1:c.T1597G:p.L533V; Si9g49990:Si9g49990.2:exon1:c.T1597G:p.L533V	intronic :: Si9g49990	UTR3 :: Si9g49990 :: Si9g49990.1:c.*816C>A;Si9g49990.2:c.*906C>A	downstream :: Si9g49990

SNP positions	54603748	54604449	54605172	54606732	54606844	54607875	54608081	Number of varieties	Varieties ID	Average	Stdev
Hap_1	T/T	G/G	C/C	T/T	C/C	C/C	A/A	283	B001 : 9.8918	8.98211	1.656
Hap_6	T/T	G/G	C/C	T/T	C/C	C/C	A/A	3	B138 : 7.9339	7.19591	0.975
Hap_7	T/T	G/G	C/C	T/T	C/C	C/C	A/A	3	B034 : 9.5864	9.09741	0.426
Hap_2	C/C	A/A	T/T	G/G	G/G	A/A	G/G	24	B013 : 14.180	18.0529	1.741
Hap_3	T/T	G/G	T/T	G/G	G/G	C/C	G/G	14	B505 : 16.997	18.4886	2.047
Hap_4	C/C	A/A	T/T	G/G	G/G	A/A	N/N	6	B119 : 19.823	17.9525	1.384
Hap_5	C/C	A/A	T/T	G/G	G/G	N/N	G/G	4	B007 : 17.963	18.5962	1.696
Hap_8	T/T	G/G	T/T	G/G	G/G	C/C	G/G	3	B139 : 20.145	19.3537	1.624





a



b

References allele	CA	G	G	G	A	Number of samples	Sample ID
Alternative allele	C	C	C	T	AG		
SNP positions	4243	4271	4273	4475	4887		
Hap_1	CA/CA	G/G	G/G	T/T**	A/A	36	B1, B10, B11, B14, B15, B16, B17, B18, B19, B20, B21, B22, B23, B24, B25, B26, B28, B29, B3, B30, B31, B32, B33, B34, B35, B36, B37, B38, B4, B41, B42, B5, B6, B7, B8, B9
Hap_5	CA/CA	G/C	G/G	T/T**	A/A	6	B12, B13, B27, B39, B40, B44
Hap_9	CA/CA	C/C**	G/G	T/T**	A/A	1	B43
Hap_2	CA/CA	G/G	G/G	G/T	A/A	24	W-1, W-11, W-13, W-18, W-20, W-21, W-22, W-31, W-32, W-34, W-35, W-37, W-41, W-43, W-45, W-46, W-48, W-8, W-9, W101, W54, W72, W86, W97
Hap_4	CA/CA	G/C	G/C	G/T	A/A	11	W-14, W-15, W-17, W-23, W-24, W-25, W-26, W-47, W-6, W85, W91
Hap_7	CA/CA	G/G	G/C	G/T	A/A	2	W-3, W98
Hap_8	CA/CA	G/G	G/C	G/T	A/AG	1	W-7
Hap_11	CA/CA	G/G	C/C**	G/T	AG/AG**	1	W71
Hap_12	CA/CA	G/G	G/G	G/T	AG/AG**	1	W-40
Hap_3	CA/CA	G/G	G/G	G/G	A/A	11	GBS, W-10, W-16, W-28, W-33, W-36, W-38, W-39, W-44, W87, W99
Hap_6	CA/CA	G/C	G/C	G/G	A/A	4	W-30, W-4, W-5, W93
Hap_10	C/C**	G/C	G/C	G/G	A/A	1	W79
Hap_13	CA/CA	G/G	G/C	G/G	A/A	1	W-12