PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes

Lucía Conde1, Juan M. Vaquerizas1, Hernán Dopazo1, Leonardo Arbiza1, Joke Reumers2, Frederic Rousseau2, Joost Schymkowitz2 and Joaquín Dopazo1,3,*

1Department of Bioinformatics, Centro de Investigación Príncipe Felipe (CIPF), Valencia, 46013, Spain, 2Switch laboratory, Flanders Interuniversity Institute for Biotechnology. (VIB), Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium and 3Functional Genomics Node, INB, CIPF Valencia 46013, Spain

Received February 14, 2006; Revised February 23, 2006; Accepted March 3, 2006

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are the simplest and most frequent type of DNA sequence variation among individuals and constitute one of the most powerful tools in the search for disease susceptibility genes, drug response-determining genes and the like (1,2). With the introduction of large-scale genotyping techniques the bottleneck in this type of experiments has moved towards the management and analysis of the data generated. In this context, one of the topics which has become a problem is the step of the selection of optimal set of SNPs. It improves the functionality of PupaSNP and PupasView programs and implements new facilities such as the analysis of user’s data to derive haplotypes with functional information. A new estimator of putative effect of polymorphisms has been included that uses evolutionary information. Also SNPeffect database predictions have been included. The PupaSuite web interface is accessible through http://pupasuite.bioinfo.cipf.es and through http://www.pupasnp.org.

OUTLINE OF THE PROGRAM

PupaSuite combines the functionality of PupaSNP (6) and PupasView (5) with new algorithms and visualisation procedures for functional haplotype prediction. The PupaSNP and PupasView programs are part of the pipeline of genotyping of the Spanish National Genotyping Center (CeGen; http://www.cegen.org/). Both tools combined bear an average of 60 SNP designs per day.
correspond to two common types of analysis: genes probably related to a disease because they are functionally related (e.g. they belong to a pathway affected in the disease), or genes present in a chromosomal region linked to a disease. PupaSuite can also directly analyse lists of SNPs. In these three cases a list of SNPs with their putative functional effect is reported. In the case of chromosomal regions it is also possible to find haplotype blocks (10). For the list of SNPs, in addition to their putative functional effect, it is possible to retrieve information on MAF in different populations from dbSNP (8) [as annotated in the Ensembl (11)], as well as LD parameters and haplotype blocks.

In addition to the analysis of lists of SNPs there is another new option: Functional haplotypes. This option (see below) allows the user to test their own SNP data and to find haplotypes (12) with the functional SNPs (5,6) and the tag SNPs (13) highlighted. Case-control studies can also be performed at this stage. The option Display and Filter SNPs for a single gene implements new functionalities in an environment a la PupasView (5). More information is presented in a graphical intuitive format (Figure 1). This option allows the sequential and interactive application of filters based on functionality, conservation, MAF and the like (5) thus permitting an easy selection of a set of optimal SNPs for a particular gene.

CRITERIA TO SELECT SNPS AS A GOOD CANDIDATES FOR GENOTYPING

Here three important features of a SNP have been taken into account in order to be considered as an optimal candidate for genotyping purposes: MAF, LD with respect to other candidates (5) and putative functional effect. MAF values were taken from the Ensembl (11), which maps dbSNP (8) data onto the corresponding chromosomal coordinates. LD are were taken from the Ensembl (11), which maps dbSNP (8) [as annotated in the Ensembl (11)], as well as LD parameters and haplotype blocks.

FUNCTIONAL HAPLOTYPES

In addition to using already available data, the users can input their own data to use the predictions on possible functional effects in combination with haplotype analysis. This possibility can be used through the Functional haplotypes

EVOLUTION AT WORK: THE SELECTIVE STRENGTHS ON CSNPS

The combined effect of all the selective pressures causes the preservation of the functionally relevant parts of the genes. Under this perspective, comparative and evolutionary studies have been used to predict the putative functional effect of SNPs (19,23) although these have mainly ignored the underlying phylogeny. Here we present another more accurate estimator of functional effect, based on sequence comparison, but taking into account phylogenetic information (24). The selective pressures acting at a codon-level where non-synonymous cSNPs are found were evaluated by means of two alternative approaches: codon-based maximum likelihood (ML) models (25) implemented in PAML (26), and likelihood-ratio (SLR) method (27) for testing deviations of neutrality.

Under the first approximation, an a priori statistical distribution describing the variation of \( \omega = dN/dS \) among sites is assumed for a number \( k \) of different classes of sites with \( \omega_k \) values at a proportion \( p_k \) of the sequences representing the effects of purifying selection \( (0 < \omega_0 < 1) \), neutral evolution \( (\omega_1 = 1) \), and positive selection \( (\omega_2 > 1) \) (25). The method involves two main steps: first, the adjustment by maximum likelihood of the evolutionary parameters to the sequences of the species compared considering two different models; and second, the use of the Bayes theorem to compute the posterior probability that each site belongs to a specific site class \( \omega_k \) defined under an a priori distribution (28). Two different models (M2a and M8) were evaluated by maximum likelihood on the sequences (29).

Under the sitewise likelihood-ratio method (SLR) a site-by-site approach to test for neutrality is used. In contrast to similar approaches developed previously (30), SLR uses the entire alignment of the sequence to determine parameters common to all sites, such as evolutionary distances. Using this approach there is no need to specify a model of how \( \omega \) varies along the sequence. A correction for multiple testing in order to obtain statistical confidence for inferences on deviations from neutrality on each site is also performed.

SNPEFFECT DATABASE

The SNPeffect database (9) describes the effect of coding non-synonymous SNPs on several phenotypic properties of human proteins using either sequence-based or structural bioinformatics tools. Molecular phenotypes are grouped in three categories: structure and dynamics, functional sites and cellular processing. Next to various external tools SNPeffect uses algorithms developed at the collaborating research groups, among which Tango (20) to predict \( \beta \)-aggregation regions in protein sequences and FoldX (21) to predict the stability change caused by the single amino acid variation.

SNP EFFECT DATABASE

The SNPeffect database (9) describes the effect of coding non-synonymous SNPs on several phenotypic properties of human proteins using either sequence-based or structural bioinformatics tools. Molecular phenotypes are grouped in three categories: structure and dynamics, functional sites and cellular processing. Next to various external tools SNPeffect uses algorithms developed at the collaborating research groups, among which Tango (20) to predict \( \beta \)-aggregation regions in protein sequences and FoldX (21) to predict the stability change caused by the single amino acid variation.

FUNCTIONAL HAPLOTYPES

In addition to using already available data, the users can input their own data to use the predictions on possible functional effects in combination with haplotype analysis. This possibility can be used through the Functional haplotypes
Figure 1. Output with the graphic representation of SNPs with putative functional effect in the gene BRCA2, along with LD maps.
option. Data must be provided to the program in linkage pedigree format (pre MAKEPRED, http://pupasuite.bioinfo.cipf.es/html/help/index.html). The PupaSuite estimates blocks by three methods: Confidence intervals (10), Four gamete rule (31) and Solid Spine of LD (14) and reconstruct haplotypes using the EM algorithm (12) as implemented in Haploviev (14). The haplotypes found in this way are represented with the corresponding functional information on all the SNPs included in it and all the LD values. This representation provides a very intuitive picture of the possible functional impact of any haplotypes beyond the individual effect of each SNP. For case/control data a chi-square test is performed and the corresponding P-value for the allele frequencies in cases versus control is reported. The combination of functional haplotype information with case/control tests allows to easily ascribe cases to haplotypes with functional alterations.

DISCUSSION

We have presented an integrated resource for helping in the selection of optimal sets of SNPs oriented to large-scale genotyping assays. The program merges the functionalities of other two previous resources, PupaSNP (6) and PupasView (5), and expand the capabilities of the program with new information and new facilities. The SNPeffect database (9) as well as a new, unpublished prediction method has been included to improve the estimation of the putative pathological effect of SNPs. Moreover, in addition to use publicly available data on SNPs, users can analyse their own experiments. What is novel and unique to tools of this type is the possibility of analysing functionally haplotypes, beyond the classical analysis one-SNP-at-a-time which ignores interactions between the mutations.

The usefulness of this type of resources is proven by the use made by the CeGen in its pipeline of genotyping. The previous tools, which have been running for more than two years, have now an approximate average of 60 daily SNP designs (http://bioinfo.cipf.es/webalizer/pupasnp and http://bioinfo.cipf.es/webalizer/pupasview).

ACKNOWLEDGEMENTS

This work is supported by grants from Fundacio La Caixa, Fundacion BBVA, MEC BIO2005-01078 and NRC Canada-SEPOCT Spain. The Functional Genomics node (INB) is supported by Genoma Espana. LC is supported by fellowship from the CeGen (Genoma Espana). Funding to pay the Open Access publication charges for this article was provided by Genome Espana.

Conflict of interest statement. None declared.

REFERENCES


