A Tool for Selecting SNPs for Association Studies Based on Observed Linkage Disequilibrium Patterns

Francisco M. De La Vega, Hadar I. Isaac, and Charles R. Scafe

Pacific Symposium on Biocomputing 11:487-498(2006)

A TOOL FOR SELECTING SNPS FOR ASSOCIATION STUDIES BASED ON OBSERVED LINKAGE DISEQUILIBRIUM PATTERNS

FRANCISCO M. DE LA VEGA[†], HADAR I. ISAAC, CHARLES R. SCAFE
Applied Biosystems, 850 Lincoln Centre Dr., Foster City, CA 94404, USA

The design of genetic association studies using single-nucleotide polymorphisms (SNPs) requires the selection of subsets of the variants providing high statistical power at a reasonable cost. SNPs must be selected to maximize the probability that a causative mutation is in linkage disequilibrium (LD) with at least one marker genotyped in the study. The HapMap project performed a genome-wide survey of genetic variation with about a million SNPs typed in four populations, providing a rich resource to inform the design of association studies. A number of strategies have been proposed for the selection of SNPs based on observed LD, including construction of metric LD maps and the selection of haplotype tagging SNPs. Power calculations are important at the study design stage to ensure successful results. Integrating these methods and annotations can be challenging: the algorithms required to implement these methods are complex to deploy, and all the necessary data and annotations are deposited in disparate databases. Here, we present the SNPbrowser™ Software, a freely available tool to assist in the LD-based selection of markers for association studies. This stand-alone application provides fast query capabilities and swift visualization of SNPs, gene annotations, power, haplotype blocks, and LD map coordinates. Wizards implement several common SNP selection workflows including the selection of optimal subsets of SNPs (e.g. tagging SNPs). Selected SNPs are screened for their conversion potential to either TaqMan® SNP Genotyping Assays or the SNPlex™ Genotyping System, two commercially available genotyping platforms, expediting the set-up of genetic studies with an increased probability of success.

1. Introduction

One problem researchers face when designing and executing human genetic studies with single nucleotide polymorphisms (SNPs) is the difficult task of selecting the most suitable set of the variants for the goal at hand in a cost-effective manner. This task is time-consuming and overwhelming due to the millions of SNPs currently listed on the public databases and the fact that relevant information is often distributed among multiple repositories which sometimes are difficult to access or the access is slow due to bandwidth and server load issues. Often this requires advanced algorithm development.

[†] To whom correspondence should be addressed. Email: delavefm@appliedbiosystems.com.

Furthermore, once a set of SNPs is selected, researchers lack a rapid way to obtain reliable, predictable assays for multiple SNPs that work together under the same experimental conditions.

To overcome these barriers, we developed the SNPbrowser Software, a freely available tool providing an intuitive interface to search a stand alone, embedded database that contains detailed information on millions of validated SNPs. Included in this SNP collection are over a million genome-wide distributed SNPs that the International HapMap Project recently genotyped,[2] as well as 160,000 intragenic SNPs previously validated by us in four populations using TaqMan SNP Genotyping Assays.[3,5] The depth of SNP and genomic information in the database together with the swift visual interface and embedded selection algorithms provides researchers greater flexibility when designing associations studies with an increased probability of success.

2. Methods

2.1. *SNP genotype and annotation data*

We previously genotyped DNA samples from 45 African-Americans, 46 Caucasians, 45 Chinese, and 45 Japanese, all unrelated individuals.[4] Over 160,000 TaqMan® SNP Genotyping Assays were used to genotype these samples. For the HapMap project dataset, we utilized genotypes from the public release 16 (Phase I data freeze) ignoring the children on the CEU and YRB trios[2]. Only SNPs having a unique mapping location on the NCBI b35 assembly and a minor allele frequency (MAF) of >5% were considered for further analysis. Gene annotation including HUGO names, exon and intron boundaries of all reported (RefSeq NM) and predicted (XM) transcripts were obtained from NCBI Entrez. Transcripts were coalesced into "supertranscript" constructs with boundaries delimited by the coordinates of the first and last base transcribed.

2.2. *SNP screening for genotyping assay development*

All the SNPs in our database were passed through the high-throughput design pipelines for both TaqMan SNP Genotyping Assays, and the SNPlex Genotyping System[6]. SNPs that passed the design rules of either platform and thus are candidates for the development of good assays where flagged and subsequent analyses were performed separately for each subset. In the case of TaqMan, assay designs (primers and probes) were uploaded into our TaqMan Predesigned database for immediate commercial availability. In the case of the

SNPlex System, since it is a multiplexed assay format, we perform a pre-screen for the "single-plex" part of the pipeline; when users submit an actual design request, a few SNPs may still be lost at the final design due to multiplexing rules.[6]

2.3. *Analysis of linkage disequilibrium*

We constructed metric maps scaled to the strength of LD that can guide the selection of SNPs for association studies. Linkage disequilibrium units (LDUs) define a metric coordinate system where locations are additive and distances are proportional to the allelic association between markers[10]. The LDMAP software v0.9 (available at: http://cedar.genetics.soton.ac.uk/public_html/helpld.html) was applied separately to each chromosome and population to construct the corresponding LDU maps. Haplotype blocks were estimated by a rule-based algorithm which uses the D' confidence interval,[7] optimized through a dynamic programming algorithm,[11] or dynamically by LDUs, as user-defined intervals with a very small distance in this coordinate system (the default value is 0.3, which returns similar blocks to previous methods[7]).

2.4. *Selection of minimum informative subsets of SNPs*

We utilized three algorithms[9] to select minimum informative subsets of SNPs or tag-SNPs: (i) simple genotype correlation between samples (allowing for one item of missing data); (ii) pair-wise r^2; and (iii) haplotype R^2. First, minimum sets of tag SNPs were selected on a chromosome-wide basis at three thresholds of pair-wise r^2 or haplotype R^2 through the use of a block-free dynamic programming algorithm framework.[8] The output of these calculations was included in the software database. Alternatively, we implemented an on-the-fly selection of tagging SNPs where the user has a greater choice of parameters which is suitable for smaller regions.

2.5. *Power calculations for case/control studies*

We calculated power for a fixed sample size of cases and controls on a per gene basis. For each gene, power is calculated using a haplotype based test, for each of the common haplotypes in the window, and entering in the calculation the empirically observed average LD on the gene region. Using a multiplicative genetic model with relative risk ratio of 3 and prevalence of 1.5%, power is calculated for each haplotype and a frequency weighted average is provided as the summary. This is repeated separately for each population, for three settings

of sample sizes of cases and controls (250/250, 500/500, 1000/1000), and assuming a disease allele frequency of either 10 or 20%. The resulting estimated power is visualized using a color scale ranging from 0.5 to 1.0 displayed as a background to each gene region.[4]

2.6. *Downloading, installing, and updating SNPbrowser*

SNPbrowser is developed using the Microsoft® Visual C++ IDE and compiler, and currently is available only as a native Windows application requiring a system with 512 Mbytes of RAM. However, the software can be readily used with the MacOS platform with Microsoft Virtual PC, a commercially available emulation environment. The latest version of SNPbrowser is always freely available for download at http://www.allsnps.com/snpbrowser/. Once installed, the software checks for updated versions either automatically or manually.

3. Results

3.1. *SNPbrowser Software embedded database*

When SNPbrowser is launched, the user has the option to select which reference database they want to utilize: either the maps and data derived from the HapMap Project[2], or the gene-centric maps obtained by Applied Biosystems (AB) by typing 160,000 SNPs in four populations[3,5]. After the user selection, SNP and gene annotations, their physical coordinates on the NCBI b35 assembly, genotypes on the corresponding reference populations, and the results of a series of LD analysis, tagging SNP, and power calculations pipelines performed offline are loaded from a set of binary files distributed with the application into a highly compressed and indexed embedded database maintained in memory. The SNPbrowser database also includes a set of *metric* LD maps[10], which are empirically derived from the patterns of allelic association observed on the hundreds of millions of genotypes analyzed, and provide information on how to best position SNPs across the genes or regions of interest in a study[1].

3.2. *Visualization and query tools*

The SNPbrowser main interface is a visualization panel consisting of a chromosome map viewer representing the location in the physical map of SNPs, and their relationship to annotated human genes and exons. Researchers studying a particular gene or a set of genes can easily pan and zoom to the region of the genome of interest. For example, investigators studying a

candidate gene for type 2 diabetes, calpain 10, can type the gene HUGO name (CAPN10) into the search box and quickly see that the gene spans about 31 kb and just 0.5 LDUs in the Caucasian population. The intron/exon structure of the gene is readily apparent, as is the haplotype block structure and the location of SNPs along the chromosomal axis (Fig. 1).

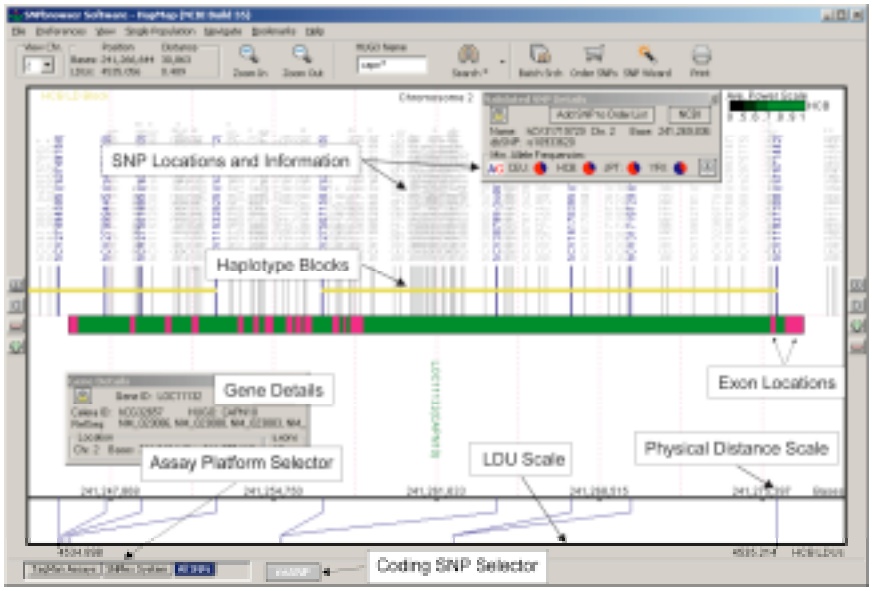

Figure 1. The SNPbrowser Software allows visualization of extensive gene and genomic information, including the physical and LD maps, intron/exon structure, the locations and allele frequencies of SNPs, and putative haplotype blocks in four different populations.

Vertical blue lines represent SNPs validated, either by the HapMap project or AB and for which genotypes are available, while grey lines represent SNPs corresponding to over 5 million putative SNP deposited at public databases. The user can select between by clicking the bottom tabs to display the SNPs which can be developed as assays for the SNPlex Genotyping System or as TaqMan SNP Genotyping assays.[6] By clicking the "All SNPs" tab, the union of both sets is displayed.

The vertical lines representing SNPs connect to their locations on the LDU coordinates shown on the bottom horizontal axis, in many cases coalescing together into a single position when LD is extensive (i.e. a haplotype block[7]). By clicking and dragging with the mouse any interval can quickly be measured in both base pairs or interpolated LDUs (see distance box upper left, Fig 1).

Finally, overall statistical power of the full SNP map, estimated per gene for a pre-selected genetic model, assumed disease allele frequency, and sample size,[4] is shown color coded within the intronic regions (scale is visible at the upper right corner).

Searches can be performed with a variety of terms, including gene name, RefSeq transcript ID, NCBI ID, SNP ID, assembly base-pair range, or Linkage Mapping Set microsatellite marker set intervals. For most of these identifiers, batch searches are also allowed. Since SNPbrowser database is loaded into RAM memory, searches are almost instantaneous, which is an advantage over web-based tools. The batch search feature allows users to quickly search genes in big candidate lists and to explore interactively the results of various selection scenarios.

3.3. *Selection of Evenly Spaced Markers*

SNPbrowser Software provides a number of SNP selection "wizards" where researchers can define a region and select SNPs at a given density, based on either LDU or kilobase (kb) distances. When selecting SNPs by spacing, the wizards also allow researchers to prioritize the SNPs that are included in the set based on criteria such as minor allele frequency (MAF) and type of SNP. For example, with a few clicks, researchers can configure the software to include only SNPs with a MAF of more than 10% in the CEPH population and for which a validated SNP assay is available (Figure 2).

Another typical use case for study design is the candidate region study, where the researchers already performed a linkage study and the goal is to perform fine mapping of an implicated chromosomal region to find the disease gene. For example, choosing an arbitrary region on chromosome 4, and searching for validated SNPs spaced at least 20 kb across the region, the SNPbrowser Software identified 33 appropriate SNPs and indicated that it was possible to achieve this spacing across the entire region (Figure 2). If validated SNPs had not been available, a red indicator bar would replace the green indicator bar in the bottom right-hand corner of the read-out window. The slider allows researchers to modify the spacing or MAF parameters to quickly visualize the level of coverage that is possible in the region given their other requirements. Alternatively, SNPs can be selected to try to achieve an even spacing of 0.5 LDUs on the metric LD map for CEPH by simply going back and changing the density parameters. In this case the wizard selects only 6 SNPs due to the extensive LD, although there is one interval at the 3'-end of the gene (red bar) where the LDU distance was greater than this value, suggesting the

presence of a recombination hotspot around this location. If desirable, the user can request the wizard to fill this "gap" with as many non validated SNPs (gray vertical lines) as required. All this process can be carried out in seconds. Since the selection process is carried out selecting a particular genotyping platform (selected by the platform tabs), additional time is saved by not having to go back and refill gaps created by SNPs that cannot be converted to a given assay format. Furthermore, the SNPbrowser wizard can also take into account SNPs for which the user already developed assays, and fill gaps around them.

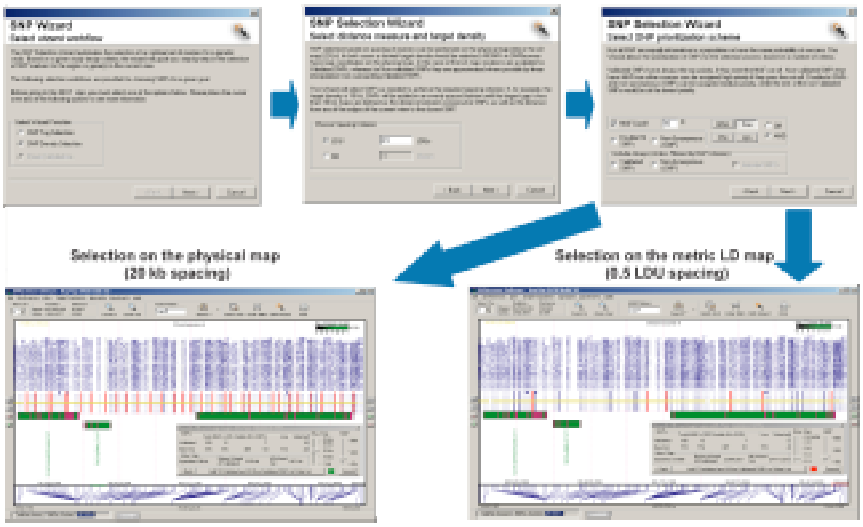

Figure 2. The SNP Selection wizard in density mode allows researchers to find SNPs at even LDU or kb intervals. In this example, a search on chromosome 4 for validated SNPs in the CEPH population with MAF of at least 10% and a gap no larger than either 20 kb or 0.5 LDU was carried out, and generated a set of either 37 (kb) or 6 (LDU) selected SNPs (shown in red).

3.4. *Selection of tagging SNPs*

Because SNPs that show high LD result in chromosomal segments in which a limited number of haplotypes are found in a population (i.e. a haplotype block[7]), it is possible to select a small subset of SNPs that distinguish, or "tag", the common haplotypes previously found in a gene or region. This eliminates a large number of SNPs from the study that would only provide redundant information. In principle, this reduction in markers brings down the cost and time necessary to conduct a study retaining good statistical power.[4] An example of this strategy is provided by inspecting the BRCA1 gene, which is

covered by a single, continuous block of LD in all four populations studied (i.e., all SNPs within the gene fall in the same location on the LD map). Therefore, although there are a vast number of SNPs in the gene, and including 37 validated SNPs, the SNPbrowser Software Tagging SNP wizard reveals that only 8 SNPs are actually required to retain most of the haplotype diversity of the gene observed in the reference samples (using a haplotype r^2 metric threshold of 0.99; see Fig. 3).

3.5. *Selecting Coding SNPs*

SNPbrowser Software also makes it easy to include putatively functional coding SNPs (cSNPs) in association studies. SNPs that result in non-synonymous codon changes and consequently, amino-acid substitutions (or premature stop codons) in the gene's protein product that can potentially affect its function[12], also referred to as non-synonymous cSNPs (nsSNPs). By simply clicking on the "nsSNP" button only this type of variants are visualized. If cSNPs are the study focus, it is possible to limit the search at two points. First, the Density Wizard includes a checkbox to make selecting cSNPs the search priority. Second, the Shopping Basket has one-click functionality that will add only the cSNPs to the cart.

3.6. *Implementing the study*

Selected SNPs can be added to a working list of markers (or "shopping basket") by either simply clicking on the results bar of the SNP wizards, manually adding individual markers with the right click option, or by invoking the "shopping basket" window and adding markers from the current view in many forms. There are two separate shopping baskets: one for each TaqMan and SNPlex platforms. In the case of TaqMan, assay availability and previous performance validation is indicated. The contents of each basket can be exported and saved for use in a future session. Finally, once the researcher has identified the ideal set of SNPs for an association study, genotyping reagents can be easily be obtained. The user can also export the list of SNPs from the shopping basket to a text file, including a number of the annotations maintained in the software internal database.

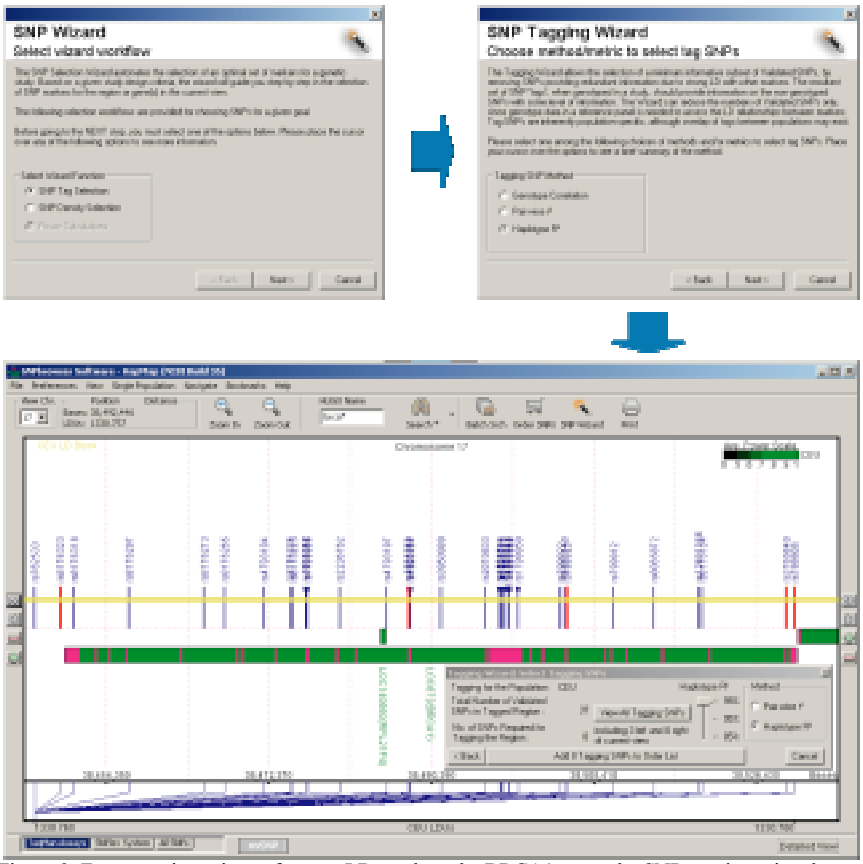

Figure 3. For genes in regions of strong LD, such as the BRCA1 gene, the SNP tagging wizard allows the selection of minimum subsets of SNPs (shown in red) retaining a threshold value of the selected quality metric (either pair-wise r^2 or haplotype R^2).

4. Discussion

Since there is no single SNP selection approach that can serve all the requirements of different types of studies, the SNPbrowser Software offers researchers a choice of methods for picking markers suited to a wide range of objectives and disease characteristics. Two basic paradigms for selecting SNP markers are supported: 1) selection of evenly spaced markers on the physical or metric linkage disequilibrium (LD) maps[10], and 2) selection of non-redundant subsets of haplotype "tagging" SNPs.[9] Furthermore, the tool immediately indicates the SNPs for which genotyping assays are viable and available from commercial sources. This means that researchers can get promptly started in

their study after identifying an optimal SNP set. Previously, identifying the most efficient and highly informative SNP set for a multi-megabase region (e.g. a candidate region from a previous linkage study performed with microsatellites) was extremely time-consuming. With the SNPbrowser wizards it only takes a few seconds, for example, to get a list of evenly spaced, highly-informative SNPs across the region of interest either on the physical (kb) or metric LD (LDU) maps. A metric LD map, expressed in LD units (LDUs) calculated by the LDMAP software,[10] places SNPs on a coordinate system where distances between SNPs are additive and directly related to the degree of LD between them. For example, SNPs in perfect LD (completely correlated) have zero distance between them, whereas SNPs with no significant correlation are separated by over three LDUs in this map. Analogous to the genetic map expressed in centi-Morgans commonly used for selecting markers for linkage studies in families, the LD map can be used to efficiently position markers for population-based disease association studies.[1]

Normally, the HapMap database would be preferred due to the depth of coverage, but often the AB maps could be useful, for example, if the study involves African-Americans (The HapMap Project did not genotype samples of this population). Although it is always preferable to utilize validated SNPs when designing genetic studies, there may be circumstances when it would be desirable to include SNPs present in the public databases but that have not been validated, e.g. by the HapMap project. SNPbrowser allows displaying the complete SNP complement for the visible region that can be converted to commercially available genotyping assays, whether validated or not, making it easy to select additional SNPs that can be used to fill gaps left by the validation projects.

Sometimes nsSNPs are included because they are referenced in the literature, and other times adding nsSNPs to the study may increase its power because in some instances an nsSNP can be indeed a causative variant for the phenotype under investigation. It is important to note that non-coding SNPs such as those in regulatory regions or splice junctions can also influence the trait of interest and thus cannot be completely ignored, but these are difficult to identify or predict. Further, if their penetrance is high, cSNPs may not occur in sufficient frequency in the population to be informative in a study with a typical sample size. Ultimately, most researchers find that it is most productive to include a mix of nsSNPs and surrogate marker SNPs with high minor allele frequencies.

In summary, SNPbrowser is a free tool that allows researchers to easily select SNPs for genetic association or other types of studies involving human

SNPs. Its main advantages include: Ease of use; swift interaction and searches; informative visualization; intuitive wizards that automate the most common selection workflows; no need to be online to access the data; completeness in terms of data and selection algorithms, enabling rapid experimental cycles by considering an assay platform conversion potential from the beginning. The software also includes extensive online help describing in detail additional features and facilities that due to length limitations cannot be discussed in this manuscript. The extensive and detailed information available through the SNPbrowser Software solves many of the major challenges that researchers face when designing human association studies, including visualizing complete genomic information in their region or gene of interest, leveraging the extensive reference genotype datasets becoming available from the HapMap project, identifying the best set of SNPs for their studies, and easily obtaining reliable assays that correspond to those SNPs.

Acknowledgments

We are very grateful to Andrew Collins, for providing the LDMAP software, Bjarni Halldórsson and Ross Lippert, who provided the block-free tagging SNP selection pipeline and haplotype phasing code, and Derek Gordon, who provided power calculation algorithms. We acknowledge the valuable support and feedback provided by Joanna Curlee, Pius Brzoska, Dennis Gilbert, Toinette Hartshorne, Fiona Hyland, Michael Rhodes, Katherine Rogers, Leila Smith, Eugene Spier, Rob Tarbox, Fenton Williams, and Trevor Woodage.

References

1. Collins, A., Lau, W. and De La Vega, F.M. *Hum Hered* 2004; **58**:2-9.
2. Consortium, T.I.H. *Nature* 2003; **426**:789-796.
3. De La Vega, F.M., Dailey, D., Ziegle, J., Williams, J., Madden, D. and Gilbert, D.A. *Biotechniques* 2002; **Suppl**:48-50, 52, 54.
4. De La Vega, F.M., Gordon, D., Su, X., Scafe, C., Isaac, H., Gilbert, D. and Spier, E.G. *Hum Hered* 2005; **60**:In Press.
5. De La Vega, F.M., Isaac, H., Collins, A., Scafe, C.R., Halldorsson, B.V., Su, X., Lippert, R.A., Wang, Y., Laig-Webster, M., Koehler, R.T., et al. *Genome Res* 2005; **15**:454-462.
6. De la Vega, F.M., Lazaruk, K.D., Rhodes, M.D. and Wenz, M.H. *Mutat Res* 2005; **573**:111-135.
7. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. *Science* 2002; **296**:2225-2229.

8. Halldorsson, B.V., Bafna, V., Lippert, R., Schwartz, R., De La Vega, F.M., Clark, A.G. and Istrail, S. *Genome Res* 2004; **14**:1633-1640.
9. Halldorsson, B.V., Istrail, S. and De La Vega, F.M. *Hum Hered* 2004; **58**:190-202.
10. Maniatis, N., Collins, A., Xu, C.F., McCarthy, L.C., Hewett, D.R., Tapper, W., Ennis, S., Ke, X. and Morton, N.E. *Proc Natl Acad Sci U S A* 2002; **99**:2228-2233.
11. Schwartz, R., Halldorsson, B.V., Bafna, V., Clark, A.G. and Istrail, S. *J Comput Biol* 2003; **10**:13-19.
12. Thomas, P.D. and Kejariwal, A. *Proc Natl Acad Sci U S A* 2004; **101**:15398-15403.