BMC Bioinformatics

BioMed Central

Software

Open Access

MIDAS: software for analysis and visualisation of interallelic disequilibrium between multiallelic markers

Tom R Gaunt*†[1], Santiago Rodriguez†[1], Carlos Zapata[2] and Ian NM Day[1]

Address: [1]Human Genetics Division, University of Southampton, School of Medicine, Duthie Building (MP 808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, UK and [2]Departamento de Genética, Universidad de Santiago, Santiago de Compostela, Spain

Email: Tom R Gaunt* - Tom.Gaunt@soton.ac.uk; Santiago Rodriguez - S.Rodriguez@soton.ac.uk; Carlos Zapata - bfcazaba@usc.es; Ian NM Day - I.N.M.Day@soton.ac.uk

* Corresponding author †Equal contributors

Published: 27 April 2006

BMC Bioinformatics 2006, **7**:227 doi:10.1186/1471-2105-7-227

Received: 19 December 2005
Accepted: 27 April 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/227

© 2006 Gaunt et al; licensee BioMed Central Ltd.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Various software tools are available for the display of pairwise linkage disequilibrium across multiple single nucleotide polymorphisms. The HapMap project also presents these graphics within their website. However, these approaches are limited in their use of data from multiallelic markers and provide limited information in a graphical form.

Results: We have developed a software package (MIDAS – Multiallelic Interallelic Disequilibrium Analysis Software) for the estimation and graphical display of interallelic linkage disequilibrium. Linkage disequilibrium is analysed for each allelic combination (of one allele from each of two loci), between all pairwise combinations of any type of multiallelic loci in a contig (or any set) of many loci (including single nucleotide polymorphisms, microsatellites, minisatellites and haplotypes). Data are presented graphically in a novel and informative way, and can also be exported in tabular form for other analyses. This approach facilitates visualisation of patterns of linkage disequilibrium across genomic regions, analysis of the relationships between different alleles of multiallelic markers and inferences about patterns of evolution and selection.

Conclusion: MIDAS is a linkage disequilibrium analysis program with a comprehensive graphical user interface providing novel views of patterns of linkage disequilibrium between all types of multiallelic and biallelic markers.

Availability: Available from http://www.genes.org.uk/software/midas and http://www.sgel.humgen.soton.ac.uk/midas

Background

Gametic disequilibrium (widely known as linkage disequilibrium or LD) is a genetic phenomenon which occurs when alleles at different loci are non-randomly associated in a given population. This correlation between polymorphisms is caused and/or influenced by their shared history of mutation and recombination, and by many other factors including genetic drift, population growth, admixture or migration, population structure, the ages of the polymorphisms, the physical distance separating them and the effects of selective pressure [1]. The characterization of LD is an important issue in both evolutionary and medical genetics, since it is informative in association mapping of trait or disease loci, and an indicator of the interaction between genes, the relative influence of different evolu-

tionary forces in the generation/disruption of genetic variability, and the genetic history of populations [2].

The theory of estimation of LD has been substantially developed in recent years. Relevant advances have been made in the knowledge of the properties of LD coefficients and LD statistical tests, which are used respectively to measure the magnitude and to estimate the significance of LD. LD is said to exist when the frequency of a haplotype observed in a population sample is significantly greater or lesser than the frequency expected from the product of the allele frequencies, the magnitude of LD correlating with such difference. There are a variety of measures and statistical tests available for the estimation of LD (D', ρ, r, r^2, d, d^2, and chi-square and Fisher exact tests, being the most used LD coefficients and statistical tests), and many programs exist for that purpose (including Haploview, 2LD, Arlequin, GDA, DNAsp, ALLASS, DISEQ, DMAP, etc., reviewed in [3] and [4]). Some software, such as GOLD [5], GOLDsurfer [6] and Haploview [7], also include graphical displays enabling quick overviews of large regions. However, most packages are intended for use with single-nucleotide polymorphism (SNP) data in a pairwise fashion. This focus on biallelic markers makes both LD estimation and graphical representation straightforward compared with multiallelic markers such as microsatellites.

The analysis of LD between a pair of multiallelic loci represents a conceptual difference in relation to the analysis of LD between a pair of biallelic loci. In both instances, LD can be analysed at two different levels. One is the overall LD between the pair of loci, and the other is the interallelic LD between each of the alleles at the first locus and each of the alleles at the second one.

The magnitude and the significance of both overall and interallelic LD are the same for pairwise analyses involving two biallelic loci. This does not apply, however, for LD between multiallelic loci. Given a pair of multiallelic loci with k and l alleles respectively, there are $k \times l$ possible interallelic associations. In theory, pairwise combinations of alleles at different loci can differ in parameters such as magnitude, significance and patterns of LD. This has been confirmed empirically in the characterization of interallelic LD between pairs of dinucleotide repeat loci spanning human chromosome 11p15 [2], and in the analysis of LD between the *TH01* microsatellite and *IGF2* SNP haplotypes in the context of the identification of microsatellite loci tagging haplotypes relevant to association mapping of complex disease traits [8,9]. The analysis of interallelic associations is therefore necessary for a complete description of LD between multiallelic loci [2].

Despite the existence of alternative estimation theory [2,10,11], LD between multiallelic loci is often estimated by pooling alleles into two groups in order to reduce the system to a two-allele two-locus model. This approach does not allow the analysis of all possible interallelic associations. In contrast, it reduces the LD between multiallelic loci to a single estimate of overall LD. It has been shown that the overall measure obtained by pooling alleles of multiallelic loci tends to underestimate LD, may complicate discrimination among the evolutionary forces generating LD in populations, and may decrease the success of association mapping of trait or disease loci ([12] and references therein). In addition to the number of alleles, the magnitude and the power to detect LD depend on other factors, including the sample size, the statistical tests and coefficients used, the allele frequencies and the sign of the association [12]. This latter issue has been shown to be of special importance. A sign-based LD estimation method recently developed for multiallelic systems [2], has been shown to considerably increase both the statistical power and the accuracy of estimation of the intensity of LD [2,13]. On the other hand, the task of presenting a graphical overview of interallelic disequilibrium between alleles of multiallelic markers is rather more challenging than for biallelic markers (with colour intensity indicating the magnitude of linkage disequilibrium between a pair of markers) and has not been previously attempted.

In this work, we have developed an integrated LD analysis software (MIDAS: Multiallelic Interallelic Disequilibrium Analysis Software) that computes interallelic LD from genotypic data incorporating the latest advances in the theory of estimation of LD, and represents graphically the intensity and significance of pairwise non-random associations between any combination of microsatellites, SNPs, haplotypes or other multi-allelic markers.

Implementation

MIDAS was written in the Python programming language v2.4 [14], using the Tkinter module for generating a graphical user interface (GUI). The Tkinter "Canvas" widget was used for plotting of graphical data, whilst other Tkinter widgets were used for creation of menus, buttons and other aspects of the interface. All modules used were part of the standard Python distribution [14] and include: Tkinter, tkFileDialog, math, copy, cPickle, os and webbrowser. The program reads and writes standard tab-delimited text files, and has an additional option to save a binary analysis file (using the cPickle module) which stores all variables and allows the user to reload a previous analysis.

Figure 1 shows a flow-chart representing the program structure. Data (raw genotypes, with marker IDs and positions) are imported, and then the user selects analysis

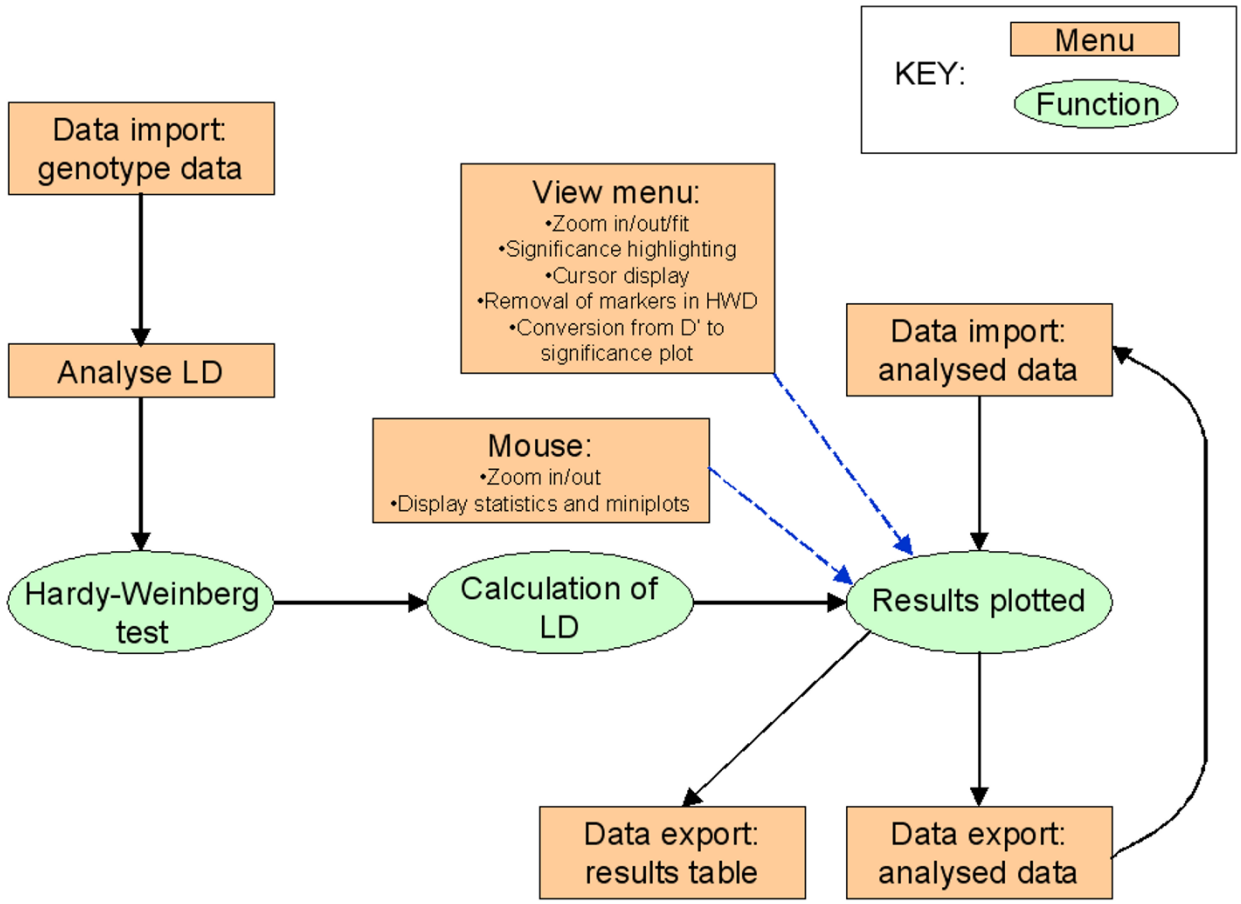

Figure 1
Flow-chart of MIDAS from the users perspective. Rectangles indicate user inputs, ovals indicate program functions.

(this is a separate step to enable incorporation of different analyses in future versions). Analysis begins with an assessment of Hardy-Weinberg equilibrium (HWE) as previously described [15,16]. Markers out of HWE are flagged for highlighting in the final outputs. The next step is estimation of LD. Finally, the results of the analysis are plotted (figure 2).

The program has been designed for simple installation and use by any computer user, and requires only the prior installation of the standard Python distribution [14] to function on a Microsoft® Windows® 2000/XP computer. Operation is mouse and menu-driven with optional hotkeys for scroll and zoom. Input files can be prepared in most spreadsheet programs and exported as tab-text. Results output is tab-text format and can be imported into most spreadsheet programs.

All parts of the program were scripted *de novo*, but the algorithm for LD calculation was based on previous programs developed by two of the authors (CZ and SR).

Estimation of LD

Given two multiallelic loci, A and B, we estimated the LD for each pair of alleles defining a two-locus haplotype. The accurate computation of all possible interallelic associations requires that each of the two-locus haplotypes defining an interallelic combination represents only the observed and expected counts for the pair of alleles under consideration. This is not attained when alleles (and therefore haplotype counts) are pooled arbitrarily. MIDAS computes interallelic disequilibrium between multiallelic loci following an approach previously described [2] which avoids both losing and pooling interallelic information. Both this approach and its underlying theory have been applied and discussed in detail previously ([2] and references therein). In brief, if locus A has k alleles A_i (i = 1,....., k) and locus B has l alleles B_j (j = 1,....., l), then the complete array of possible two-locus haplotypes was partitioned into $k \times l$ separate 2×2 contingency tables by

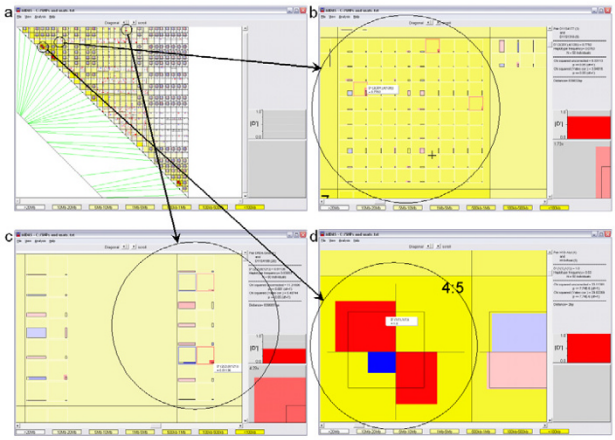

Figure 2
Screenshots of MIDAS. (a) A region of chromosome 11 showing 30 markers. Green lines indicate relative position of markers. Yellow intensity indicates distance between pairwise markers. Placing the mouse over a feature provides details. (b) A pairwise plot for two microsatellites (zoomed in). Significant results are boxed in red (D' ≥ 0) or blue (D'<0). Placing the mouse cursor over an allele pair provides details and statistics and also plots that pair at the bottom right of the screen. Magnitude of |D'| is also plotted (middle right). (c) A SNP/microsatellite pair (zoomed in). This is identical to the microsatellite/microsatellite plot, but with only two alleles in one dimension. (d) A SNP/SNP pair (zoomed in). The plot is oriented to place the most frequent alleles for both SNPs in the top left. Statistics can be observed by placing the mouse over an allele pair. For SNPs a magnified plot is not shown, but the |D'| graph is still used (middle right).

collapsing the data into A_i vs. not-A_i ($A_{\bar{i}}$) at the A locus, and B_j vs. not-B_j ($B_{\bar{j}}$) at the B locus. Estimates of two-locus haplotype frequencies were obtained from genotype data by the Hill method, an expectation-maximisation (EM) algorithm [17]. The magnitude of disequilibrium between pairs of alleles at different loci was measured by $D'_{ij} = D_{ij}/D_{max}$, where $D_{ij} = X_{ij} - p_i q_j$, p_i and q_j are the frequencies of alleles i and j, respectively, X_{ij} is the observed frequency of the haplotype A_iB_j and $D_{max} = \min [p_i(1 - q_j),(1 - p_i)q_j]$ when $D_{ij} > 0$ or $D_{max} = \min [p_i q_j,(1 - p_i)(1 - q_j)]$ when $D_{ij} < 0$ [10,18,19]. Significance test of the null hypothesis of random association between pairs of alleles at the two loci ($D_{ij} = 0$) was tested by $X^2_{ij} = nD^2_{ij} / p_i(1-p_i)q_j(1-q_j)$, which approximates a χ^2 distribution with one degree of freedom, where n is the number of individuals sampled [20,21]. Yates's correction was also computed.

Estimation of the magnitude and significance of pairwise LD involving biallelic loci was performed in the same way, but considering that p_i and q_j are the frequencies of the commonest alleles for each biallelic locus. This establishes a homogeneous criterion for the construction of 2 × 2 contingency tables, (i.e., consideration of haplotype A_iB_j as the one constituted by the two more frequent alleles). This criterion was uniformly followed for the estimation of the observed haplotype frequency and for computation of pairwise LD magnitude and significance in all SNP/SNP analyses. This criterion is consistent with respect to the sign-based LD estimation method recently developed for multiallelic systems [2,13] and establishes consistency of some biological basis. By placing the most frequent allele of each of a pair of SNPs in the top left of the 2 × 2 table then if the minor alleles coincide on some haplotype, the display shows a 'main diagonal' excess (D' positive) whereas if minor allele at locus A predominate with major allele of locus B (and vice versa) the display shows a minor diagonal pattern (D' negative). When |D'| = 1, the haplotype patterns depicted (either three or two of the possible four) give information which is relevant to their possible history not fully evident from D' nor from r^2 nor any other coefficient (see figure 5).

For pairwise analyses involving two multiallelic loci, haplotype A_iB_j was considered to comprise the two alleles of interest. This is also consistent with the sign-based LD estimation method [2,13] in most situations, except in rare circumstances when haplotype A_iB_j is constituted by one allele with frequency higher than 0.5 and another allele with frequency lower than 0.5. For pairwise analyses involving one multiallelic locus and one biallelic locus, both LD estimation and representation were performed twice for each microsatellite allele of interest: A_iB_j was considered to comprise the microsatellite allele of interest and the commonest allele at the biallelic locus in one analysis, and the microsatellite allele of interest and the rarest allele at the biallelic locus in the other.

For users that wish to perform the analysis of multiallelic markers by dichotomising the marker into most common allele versus all other alleles combined we provide data in the output file to indicate that analysis for each combination of markers. This is provided in rows where there is a "Y" for "MostFreq1" (first marker) and "Y" for "MostFreq2" (second marker). For users who wish to collapse multiallelic markers to biallelic markers in other ways the software will accept that data in the form of input files with multiallelic markers recoded as if they were SNPs. However, it should be noted that no dichotomization represents the actual overall LD between two multiallelic loci, but only one of the possible interallelic associations.

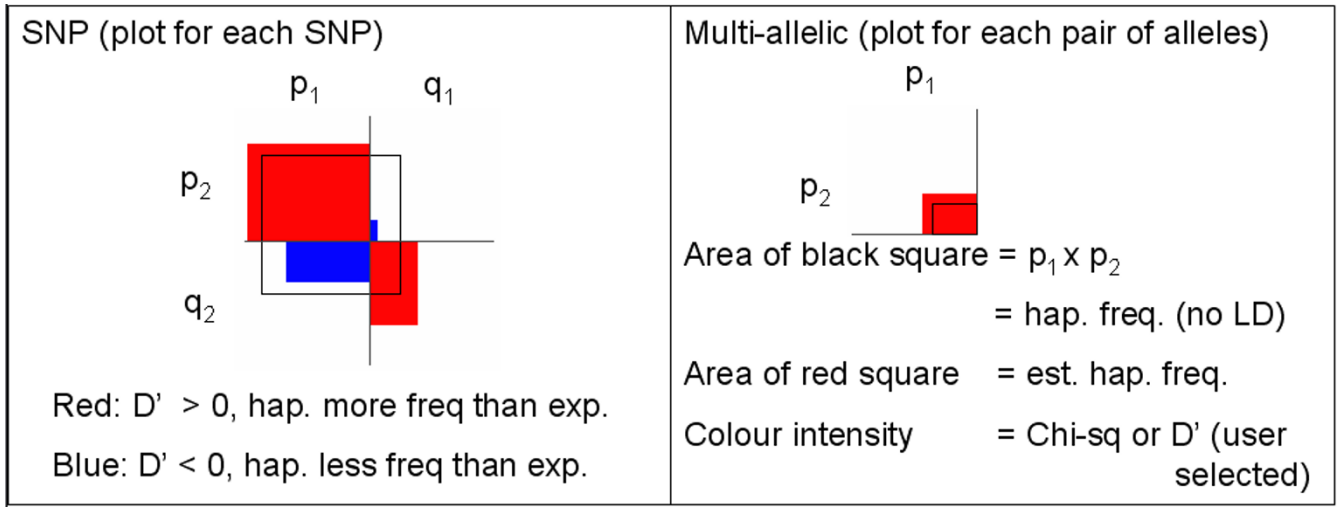

Figure 3

Representation of haplotype frequencies and LD in MIDAS. For SNPs the expected (under no LD) haplotype frequencies for each allele combination are plotted with an unfilled, black rectangle divided into four quadrants by two lines. The estimated haplotype frequencies are represented by solid red or blue rectangles. Where a coloured rectangle exceeds the size of the black rectangle it is coloured red, indicating an excess of that haplotype (D' ≥ 0). The opposite situation is indicated by a blue rectangle (D'<0). For multi-allelic markers the principle is the same, but a separate plot is shown for each combination of alleles, i.e. locus 1 allele i/allele not-i, locus 2 allele j/allele not-j.

The MIDAS output file provides D', r^2, expected and estimated haplotype frequencies, allele frequencies, χ^2 and distance between markers.

Results and discussion

An example dataset is shown in figure 2, comprising a set of microsatellites and SNPs from the 11p chromosome region [2] and Zapata *et al* (in prep) (subset of 50 samples). Figure 2a shows the typical unzoomed view, with pairs of markers in a grid of 1 to n columns and 1 to n rows. The plot begins at top left with 1 versus 2 and continues to n-1 versus n at bottom right. The distance between markers is represented by the intensity of background colour (closer = darker yellow). The image can be zoomed and scrolled, and positioning the mouse over a plotted result gives the statistics, a magnified plot and a plot of |D'| (as shown in the right-hand panel of figure 2b–d).

For pairwise SNP analyses the LD is represented as in figure 2d. The vertical and horizontal lines split the black square into four rectangles, the areas of which represent the expected haplotype frequencies for each allele combination (upper left is A_iB_j frequency = $A_i \times B_j$) (figure 3). Each quadrant then has a coloured rectangle to represent the "observed" (i.e. estimated using EM) haplotype frequency. The dimensions of the rectangle are in proportion to the two allele frequencies it represents, and its colour intensity represents the significance (by χ^2) of LD or the magnitude of D' (user option on the View menu, figures in this paper show use of the significance option). Blue rectangles represent a less frequent haplotype than expected (D' < 0), and red a more frequent haplotype than expected (D' > 0). Alleles are re-ordered to ensure that the most common alleles for each marker are represented by the top-left quadrant.

For multiallelic versus biallelic or multiallelic the plot is slightly different (figure 2c). For each marker combination there are multiple pairs of vertical and horizontal lines (matching the upper left quadrant of the biallelic display). Each pair represents one allele combination, with the black rectangle indicating the expected haplotype frequency and the coloured rectangle the "observed" (i.e. estimated using EM) haplotype frequency (figure 3). The colour scheme is the same as for biallelic markers.

The markers are arrayed with marker 1 versus marker 2 at top left and marker n-1 versus marker n at bottom right, forming a right-angle triangle of plots (figure 2a). To the bottom left of the display is a line parallel to the long side of this triangle representing a map of the genomic region in which the markers are situated. Each marker is represented by a green line from their relative position on this map to the row and column in which they are plotted (figure 2a). Placing the mouse over this line (or the circle at its end) gives marker name and position.

A typical session involves preparation of an input file of genotypes (figure 4) using any mainstream spreadsheet

Markers should be in position order.
One sample per line, one marker per column.

Marker name →
Position →
Alleles - "_" delimited →
(Columns tab-delimited)

M001	M002	S001	S002	S003
1000	50000	55000	60000	70000
122_110	139_137	2_2	2_2	1_1
118_92	141_141	2_2	1_1	1_2
122_110	137_135	1_1	1_2	1_1
110_108	137_137	1_2	2_2	1_1
110_108	137_137	2_2	2_2	1_1

Figure 4
The format of a MIDAS input file. Data are raw genotypes in a tab-delimited text file. Row 1 contains marker names, row 2 contains positions. Markers should be sorted in position order for clarity. Alleles should be delimited by an underscore ("_"), and can be any valid letter or number. Where numbers are used, ensure that the same number of digits are used for all alleles (eg 094, 098, 102) to preserve size order in the alphanumeric sort. There must be no more than one blank line at the end of the data and all null values must be coded as "?_?".

program and exporting as tab-text format. MIDAS is then run by double-clicking the script file. The window shows basic instructions – briefly: (1) "Open genotype file" from "Open" on the "File" menu, then (2) Select "Analysis" – "LD and haplotypes". Zoom can be operated by mouse-click, key-stroke ("i" and "o") or menu, while scrolling can be operated by cursor key or scroll-bar. At minimum zoom ("View" – "fit to screen" eg figure 2a) the user can rapidly spot statistically significant results and patterns. Placing the mouse cursor over a feature displays its statistics and detail. The high levels of zoom shown in figures 2b,2c and 2d enable the user to analyse the LD in more detail, and read the statistics by using mouse-over. Export functions include a tab-text file of all results (which can be opened by any mainstream spreadsheet program) and a binary file format that stores all analysis variables and can be used to store the whole analysis for future use. The latter format speeds up viewing of previous analyses with regions containing many markers. Finally a postscript export option is available to save graphical view, although standard screen captures (using Alt-PrintScreen in Microsoft® Windows®) are adequate for most uses.

Menu options include file operations, analysis, options to customise the view and a help option, which provides simple information and instructions for usage.

Application of MIDAS

Evolutionary relationships between SNPs

The SNP/SNP plots (which can be seen at low levels of zoom) provide a quick way of inferring evolutionary relationships between markers. Figure 5 shows how three different type of plot provide this type of information.

LD between a multiallelic microsatellite and several SNPs

We have previously described the association between allele groups of a highly polymorphic microsatellite in the Growth Hormone/chorionic somatomammotrophin (*GH*/*CSH*) gene region on chromosome 17 (*CSH1*.01) and phenotypes of the metabolic syndrome [22]. For these analyses we dichotomised the microsatellite on the basis of size and distribution [22]. Figure 6 shows the interallelic LD between SNPs in the *GH*/*CSH* gene region and the *CSH1*.01 microsatellite. These analyses confirm the validity of our dichotomisation, indicating two major clades of alleles within the microsatellite. However, in most cases the analysis of all alleles of a multiallelic marker rather than a dichotomisation of alleles provides the maximum information, with no necessity to make biological assumptions.

LD between input haplotypes and other markers

Figure 6 also demonstrates the potential to input haplotype data and analyse it as a multi-allelic marker. In this case a 4-SNP haplotype is analysed for LD with a multiallelic microsatellite (figure 6f). This approach can provide an overview of how biallelic or multiallelic markers interact with haplotypes, and also how haplotypes in two different regions interact with each other.

LD between two multiallelic tandem repeat loci

Our work in the *IGF2-INS-TH* region of chromosome 11 has identified associations between SNPs, haplotypes and other markers and obesity [23-28]. Two multi-allelic markers within this region (included in our haplotype analyses) are the insulin gene VNTR (*INS* VNTR) and the tyrosine hydroxylase tetranucleotide microsatellite

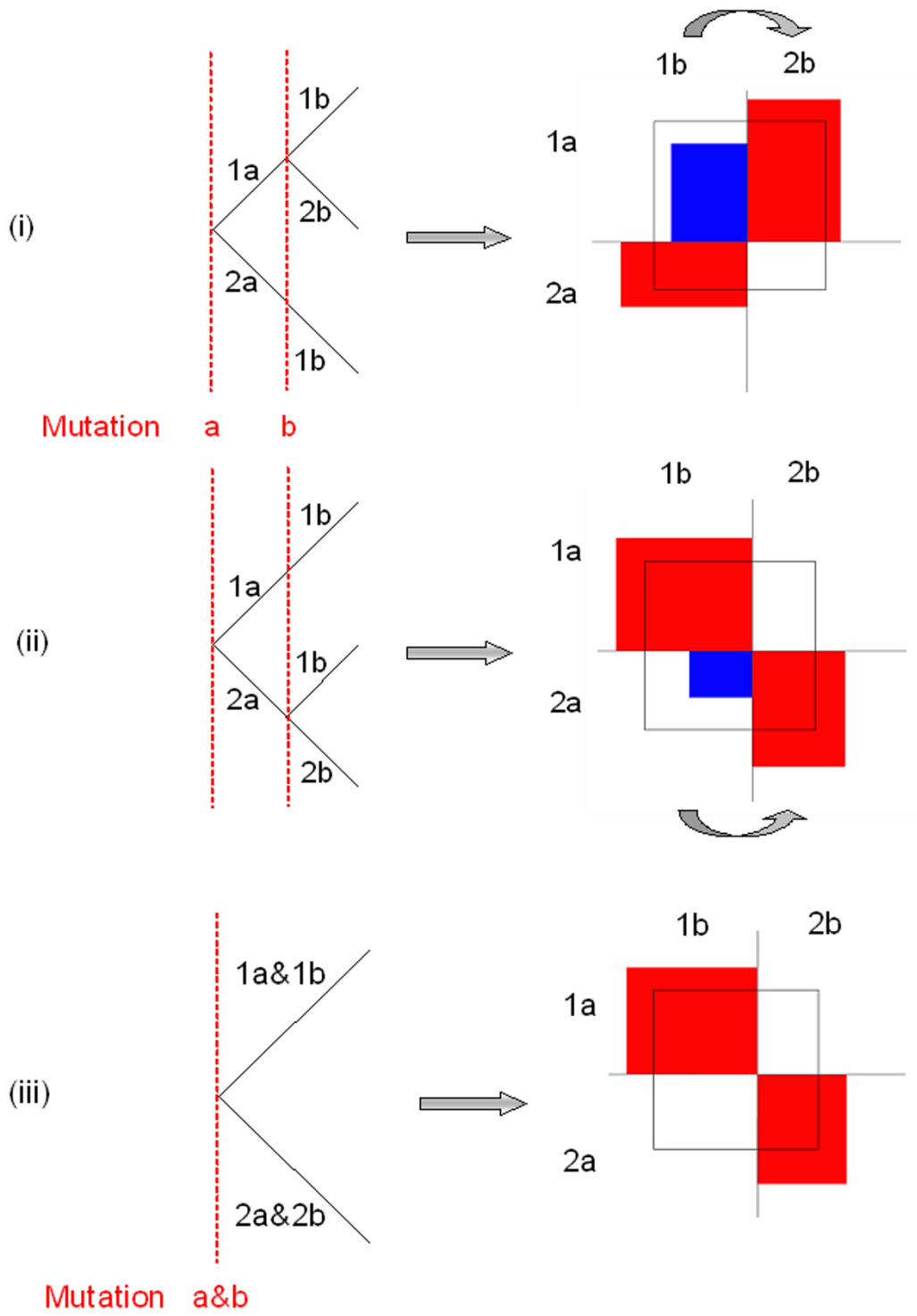

Figure 5
Use of MIDAS SNP/SNP plots to infer evolutionary history. The haplotype on which a SNP first arose is indicated by the estimated frequency of the haplotype carrying the most frequent alleles at both loci. (i) If this is less than expected, it implies that the SNP 2 arose on the haplotype carrying the common allele at SNP 1 (i.e. D'<0). (ii) If it is more common than expected then SNP 2 arose on the haplotype carrying the rare allele at SNP 1 (i.e. D' $\geq$ 0). (iii) If only two haplotypes are observed then perfect LD exists (r^2 = 1). This may arise through bottlenecks, selection or simultaneous occurrence.

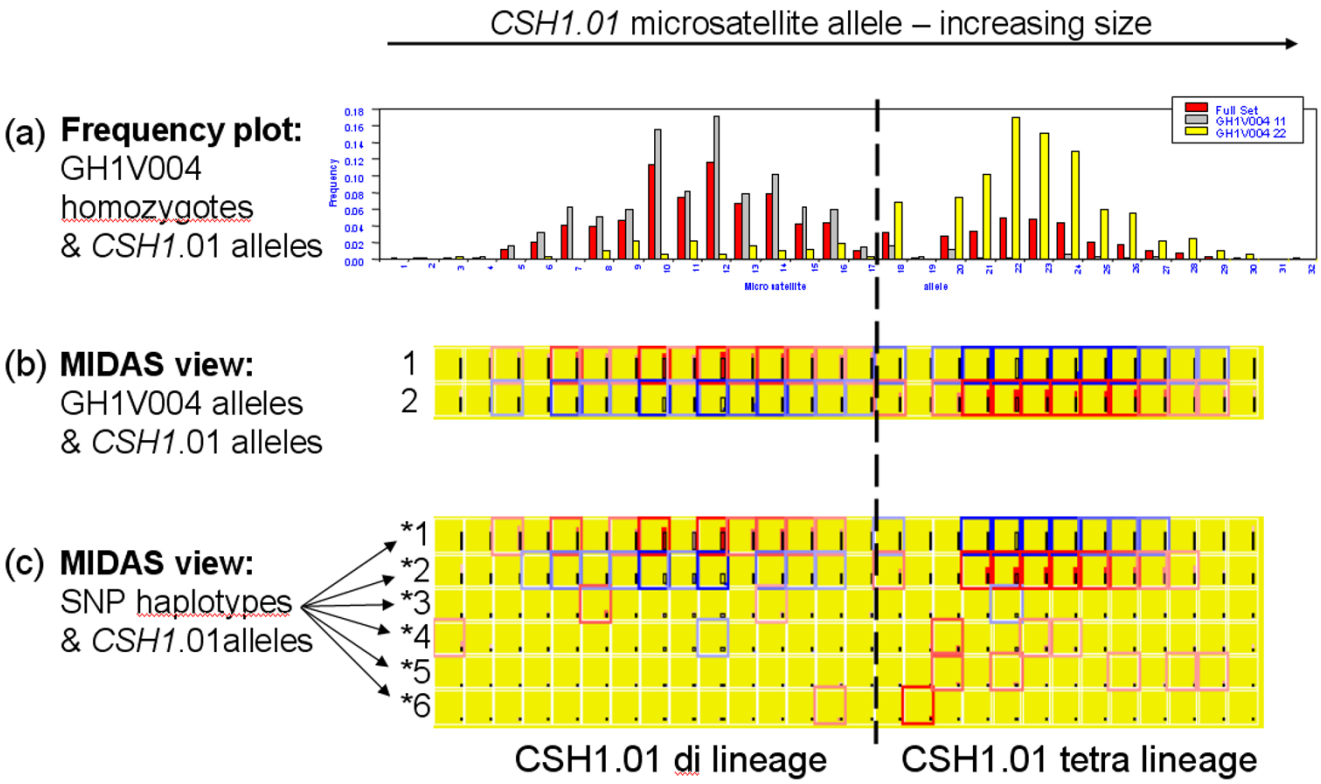

Figure 6

LD between a complex microsatellite and SNPs. (a) Previous work [22] indicated SNP alleles in LD with two size ranges of the *CSH1*.01 microsatellite. The lower size range has dinucleotide spacing, the upper has tetranucleotide spacing. This suggested two major lineages. (b) Plotting interallelic LD between a SNP (GH1V004) and the *CSH1*.01 microsatellite demonstrates clear LD with the two lineages. The common SNP alleles associate with the lower size range and the rare SNP alleles associate with the upper size range. Results are boxed in red where the haplotype frequency is significantly higher than expected (D' $\geq$ 0) and blue where it is significantly lower (D'<0). (c) SNP haplotypes (four SNPs, including GH1V004) confirm these findings and demonstrate the ability of MIDAS to handle haplotype data as a multi-allelic marker.

(*TH*01). Figure 7 shows the interallelic linkage disequilibrium between these two markers, with the patterns indicating which alleles are associated and also suggesting that the VNTR mutates more rapidly than the microsatellite (one *TH*01 allele associates with multiple *INS* VNTR alleles).

Regions of "perfect" LD

Recent work using data from HapMap [29,30] and Celera has indicated the presence of extended regions of perfect LD (where only two major haplotypes exist) [31]. Figure 8a shows the characteristic MIDAS pattern for this type of region, with only two haplotypes existing for each pair of SNPs (data from HapMap [29,30]). Figure 8b and 8c demonstrates an alternative approach which we have developed (SNPFrequencyViewer) to rapidly scan for these regions, in which extended regions of isofrequent SNPs correlate with regions of perfect LD (data from HapMap [29,30]). These can then be confirmed and examined in more detail using MIDAS. Admixture of two populations highly differentiated in these genomic regions is one possible explanation, selection another.

Conclusion

MIDAS is a new program that presents the novel approach of analysing and graphically representing the interallelic linkage disequilibrium (LD) between multiple pairs of bi- and multiallelic markers. The graphical representation of LD incorporates information on expected haplotype frequency (under no LD), estimated haplotype frequency and D' or significance. Distance information and statistics are also presented in the interface. This enables rapid visual interpretation and inference of evolutionary and functional relationships between SNPs and microsatellites across large genomic regions. Applications to data-sets we have analysed previously demonstrate the effectiveness of viewing patterns in the data graphically rather than numerically.

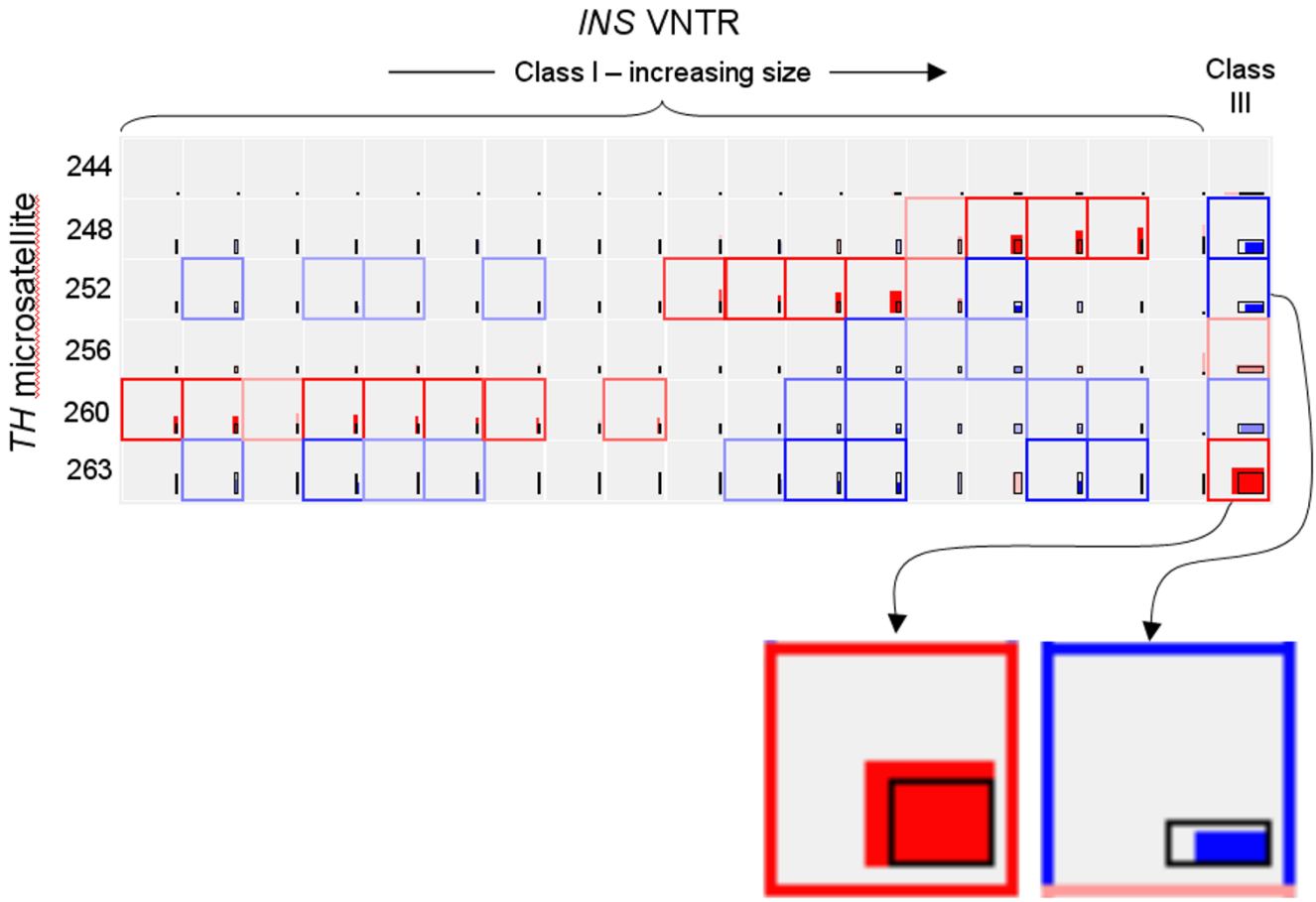

Figure 7
LD between the *INS* VNTR and the *TH*01 microsatellite. Each *TH*01 allele associates with a size range of VNTR alleles (256 and 263 associate with the class III alleles). This infers a greater rate of mutation in the VNTR because there is a wider range of allele sizes in the VNTR dimension significantly associated with *TH*01 alleles than *vice versa*. Close-ups of individual allele plots are shown to indicate the magnitude of effect – black rectangle indicates expected haplotype frequency under no LD, coloured rectangle indicates the estimated haplotype frequency.

Availability and requirements

• Project name: MIDAS: Multiallelic Interallelic Disequilibrium Analysis Software.

• Project home page: http://www.genes.org.uk/software/midas

• Operating system(s): Microsoft® Windows® 2000/XP

• Programming language: Python 2.4/Tkinter

• Other requirements: Python 2.4 or later [14] must be installed before MIDAS

• License: MIDAS licence supplied with program

• Any restrictions to use by non-academics: royalty-free use allowed within terms of licence

Authors' contributions

MIDAS was written in Python 2.4 with a Tkinter graphical interface by TRG with design suggestions and testing by all authors. Algorithms for estimating haplotype frequencies from Hill method and for LD were adapted from a BASIC code program developed by CZ (who also suggested some statistical improvements). The manuscript was drafted by TRG and SR with inputs from all authors.

Acknowledgements

TRG is funded by a BHF (British Heart Foundation) Intermediate Fellowship (FS/05/065/19497), SR by a HOPE (Wessex Medical Trust) fellowship and work in our laboratory by the Medical Research Council (UK) (Programme Grant G9800748).

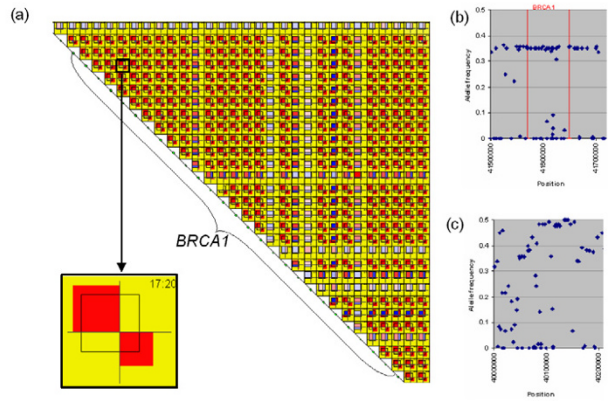

Figure 8
Visualisation of regions of perfect LD. Marker pairs can have either two or three haplotypes present when D' = 1. Most programs do not distinguish between these graphically, despite the potential biological importance. (a) The *BRCA1* region on chromosome 17. MIDAS shows only two haplotypes for many SNPs (perfect LD, $r^2 = 1$) using HapMap data [29,30]. (b) Allele frequencies from HapMap data [29,30] show that SNPs in regions with only two haplotypes share the same minor allele frequency (MAF) for many SNPs (eg *BRCA1* region on chromosome 17) compared to (c) nearby regions which have a mixture of MAFs. Viewing MAF may therefore be a quick way to find regions of perfect LD, which can then be checked with MIDAS.

References

1. Ardlie KG, Kruglyak L, Seielstad M: **Patterns of linkage disequilibrium in the human genome.** *Nat Rev Genet* 2002, **3:**299-309.
2. Zapata C, Rodríguez S, Visedo G, Sacristán F: **Spectrum of nonrandom associations between microsatellite loci on human chromosome 11p15.** *Genetics* 2001, **158:**1235-1251.
3. Jorde LB: **Linkage disequilibrium and the search for complex disease genes.** *Genome Res* 2000, **10:**1435-1444.
4. Mueller JC: **Linkage disequilibrium for different scales and applications.** *Brief Bioinform* 2004, **5:**355-364.
5. Abecasis GR, Cookson WO: **GOLD – graphical overview of linkage disequilibrium.** *Bioinformatics* 2000, **16:**182-183.
6. Pettersson F, Jonsson O, Cardon LR: **GOLDsurfer: three dimensional display of linkage disequilibrium.** *Bioinformatics* 2004, **20:**3241-3243.
7. Barrett JC, Fry B, Maller J, Daly MJ: **Haploview: analysis and visualization of LD and haplotype maps.** *Bioinformatics* 2005, **21:**263-265.
8. Rodríguez S, Gaunt TR, O'Dell SD, Chen XH, Gu D, Hawe E, Miller GJ, Humphries SE, Day IN: **Haplotypic analyses of the IGF2-INS-TH gene cluster in relation to cardiovascular risk traits.** *Hum Mol Genet* 2004, **13:**715-725.
9. Rodríguez S, Gaunt TR, Dennison E, Chen XH, Syddall HE, Phillips DI, Cooper C, Day IN: **Replication of IGF2-INS-TH*5 haplotype effect on obesity in older men and study of related phenotypes.** *Eur J Hum Genet* 2006, **14:**109-116.
10. Weir BS, Cockerham CC: **Testing hypothesis about linkage disequilibrium with multiple alleles [abstract].** *Genetics* 1978, **88:**633.
11. Karlin S, Piazza A: **Statistical methods for assessing linkage disequilibrium at the HLA-A, B, C loci.** *Ann Hum Genet* 1981, **45:**79-94.
12. Zapata C, Carollo C, Rodriguez S: **Sampling variance and distribution of the D' measure of overall gametic disequilibrium between multiallelic loci.** *Ann Hum Genet* 2001, **65:**395-406.
13. Zapata C, Nunez C, Velasco T: **Distribution of nonrandom associations between pairs of protein loci along the third chromosome of Drosophila melanogaster.** *Genetics* 2002, **161:**1539-1550.
14. **The Python Programming Language** [http://www.python.org]
15. Curie-Cohen M: **Estimates of inbreeding in a natural population: a comparison of sampling properties.** *Genetics* 1982, **100:**339-358.
16. Robertson A, Hill WG: **Deviations from Hardy-Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients.** *Genetics* 1984, **107:**703-718.
17. Hill WG: **Estimation of linkage disequilibrium in randomly mating populations.** *Heredity* 1974, **33:**229-239.
18. Lewontin RC: **The interaction of selection and linkage. I. General considerations; heterotic models.** *Genetics* 1964, **49:**49-67.
19. Hedrick PW: **Gametic disequilibrium measures: proceed with caution.** *Genetics* 1987, **117:**331-341.
20. Weir BS: **Inferences about linkage disequilibrium.** *Biometrics* 1979, **35:**235-254.
21. Day IN, Chen XH, Gaunt TR, King TH, Voropanov A, Ye S, Rodriguez S, Syddall HE, Sayer AA, Dennison EM, Tabassum F, Barker DJ, Cooper C, Phillips DI: **Late life metabolic syndrome, early growth, and common polymorphism in the growth hormone and placental lactogen gene cluster.** *J Clin Endocrinol Metab* 2004, **89:**5569-5576.
22. O'Dell SD, Miller GJ, Cooper JA, Hindmarsh PC, Pringle PJ, Ford H, Humphries SE, Day IN: **Apal polymorphism in insulin-like growth factor II (IGF2) gene and weight in middle-aged males.** *Int J Obes Relat Metab Disord* 1997, **21:**822-825.
23. O'Dell SD, Bujac SR, Miller GJ, Day IN: **Associations of IGF2 Apal RFLP and INS VNTR class I allele size with obesity.** *Eur J Hum Genet* 1999, **7:**821-827.
24. Gu D, O'Dell SD, Chen XH, Miller GJ, Day IN: **Evidence of multiple causal sites affecting weight in the IGF2-INS-TH region of human chromosome 11.** *Hum Genet* 2002, **110:**173-181.
25. Gaunt TR, Cooper JA, Miller GJ, Day IN, O'Dell SD: **Positive associations between single nucleotide polymorphisms in the IGF2 gene region and body mass index in adult males.** *Hum Mol Genet* 2001, **10:**1491-1501.
26. Rodriguez S, Gaunt TR, Dennison E, Chen XH, Syddall HE, Phillips DI, Cooper C, Day IN: **Replication of IGF2-INS-TH(*)5 haplotype effect on obesity in older men and study of related phenotypes.** *Eur J Hum Genet* 2006, **14:**109-116.
27. **The International HapMap Project.** *Nature* 2003, **426:**789-796.
28. Altshuler D, Brooks LD, Chakravarti A, Collins FS, Daly MJ, Donnelly P: **A haplotype map of the human genome.** *Nature* 2005, **437:**1299-1320.
29. Lawrence R, Evans DM, Morris AP, Ke X, Hunt S, Paolucci M, Ragoussis J, Deloukas P, Bentley D, Cardon LR: **Genetically indistinguishable SNPs and their influence on inferring the location of disease-associated variants.** *Genome Res* 2005, **15:**1503-1510.
30. Costas J, Salas A, Phillips C, Carracedo A: **Human genome-wide screen of haplotype-like blocks of reduced diversity.** *Gene* 2005, **349:**219-225.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral

31. **SNPFrequencyViewer** [http://www.genes.org.uk/software/snpfrequencyviewer]

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp

BioMedcentral