

MIDAS: a graphical program for the analysis of interallelic linkage disequilibrium between multiallelic (and biallelic) markers.

Tom R. Gaunt 1a*, Santiago Rodriguez 1b#, Carlos Zapata 2, Ian N. M. Day 1c

1. Human Genetics Division, University of Southampton, School of Medicine, Duthie Building (MP 808), Southampton General Hospital, Tremona Road, Southampton SO16 6YD, UK, (a) Email Tom.Gaunt@soton.ac.uk (b) Email S.Rodriguez@soton.ac.uk (c) Email I.N.M.Day@soton.ac.uk

2. Departamento de Genética, Universidad de Santiago, Santiago de Compostela, Spain Email bfcazaba@usc.es

* Corresponding author. # These authors contributed equally.

Abstract

Various software tools are available for the analysis and visualisation of pairwise linkage disequilibrium across multiple single nucleotide polymorphisms. The HapMap project also presents these graphics within their website (based on their data). Most graphical representations use a colour to indicate magnitude of a statistic for each pairwise analysis. Whilst this is useful over long ranges of many markers, the total information available graphically is limited. We have developed an alternative program called MIDAS which calculates and displays an enhanced graphical representation of linkage disequilibrium. In addition to D' and significance values, haplotype frequencies (expected and estimated) are displayed and distance between markers is shown on a map. Statistics can be viewed by mouse-over for each pairwise analysis, and all data can be exported in tabular form for analysis in other programs. MIDAS is capable of analysing data from biallelic and multiallelic markers (including multi-locus haplotypes). This program therefore facilitates visualisation of patterns of linkage disequilibrium across genomic regions, analysis of the relationships between different markers and inferences about patterns of evolution and selection.

* MIDAS (Multiallelic Interallelic Disequilibrium Analysis Software) is available at: <http://www.oege.org/software/midas> and <http://www.genes.org.uk/software/midas>

Please also see HGM2006 posters 221 and 249 from these authors

Input Files

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

Marker name	M001	M002	S001	S002	S003
Position	1000	50000	55000	60000	70000
Alleles - "_" delimited (Columns tab-delimited)	122_110 118_92 122_110 110_108 110_108	139_137 141_141 137_135 137_137 137_137	2_2 2_2 1_1 1_2 2_2	2_2 1_1 1_2 2_2 2_2	1_1 1_2 1_1 1_1 1_1

Description

Data are input into MIDAS using a simple file format. This file can be prepared in a spreadsheet and exported as tab-delimited text. Alleles are separated by an underscore character, and missing data represented by a "?"

Data are plotted:
For SNP-SNP: a black square represents a frequency of 1 for each SNP, with two lines dividing the square into their respective allele proportions. The coloured squares then indicate the estimated haplotype frequency, with red indicating that it exceeds the expected (black square) area and blue that it is less.
For microsatellite data: for each allele combination the top left of the SNP-SNP plot is prepared. The black square shows expected haplotype frequency and the coloured square the observed haplotype frequency.
In both cases colour intensity indicates significance or magnitude of D' (depending on user setting).

An example dataset is shown (microsatellites and SNPs from the 11p chromosome region Zapata *et al* (2001) and Zapata *et al* (in prep) (subset of 50 samples)).
(a) Unzoomed view showing a grid of M_i vs M_j vs M_n (where M = marker). Distance between markers is represented by intensity of background colour.
(b) A microsatellite-microsatellite view.
(c) A microsatellite-SNP view.
(d) A SNP-SNP view.

Flowchart representing the use of MIDAS:
Data is imported, then the user selects the "Analyse LD" menu option. Hardy-Weinberg Equilibrium is tested to determine whether the user needs to be warned about any markers out of HWE. LD is calculated using the Hill algorithm (Hill, 1974). Finally, results are plotted. The user can then use the interactive features of the software to zoom in and out, navigate through the plots and view the statistics. Data can be exported in tabular format, and analyses can be saved for later viewing.

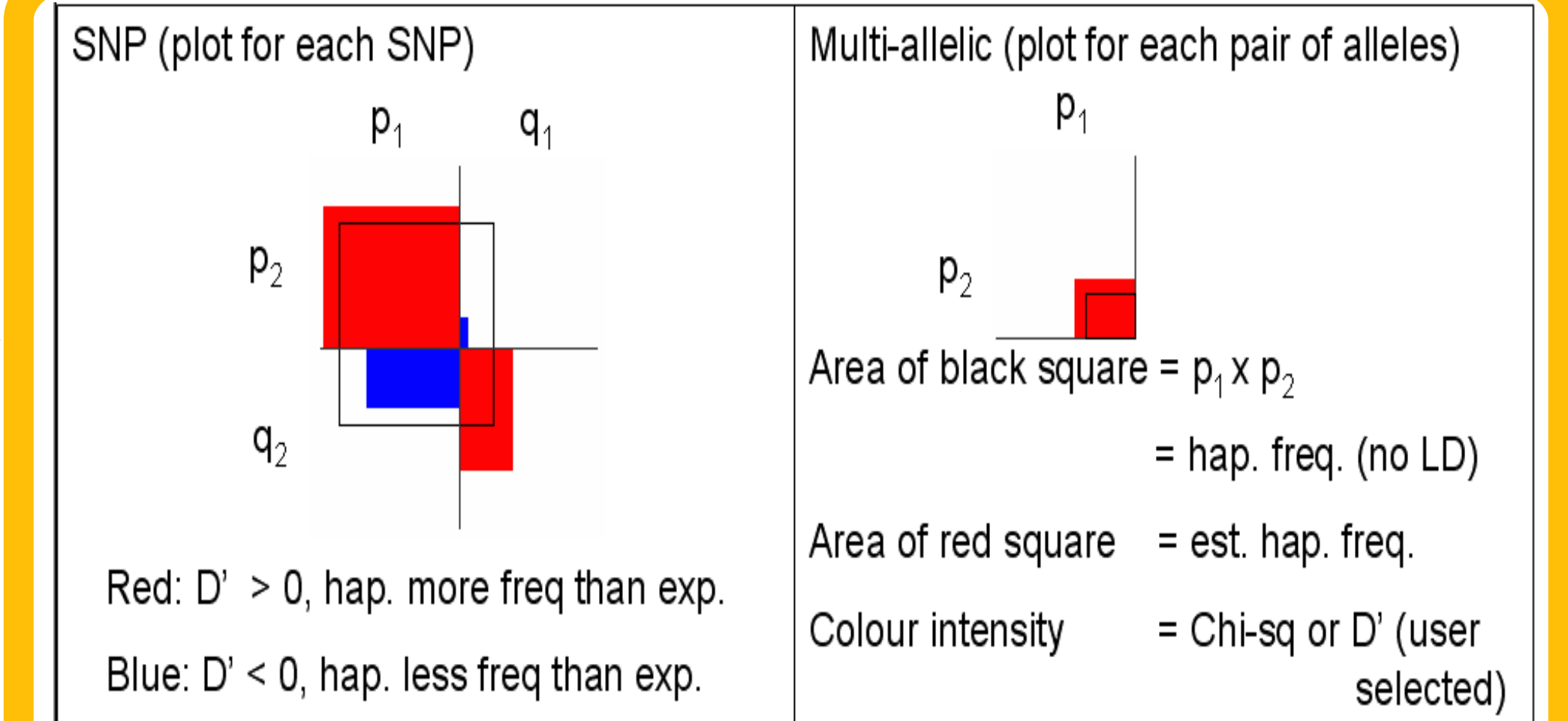
An example analysis of linkage disequilibrium between a complex minisatellite (*INS* VNTR) and a tetranucleotide microsatellite (*TH01*) (O'Dell *et al*, 1999; Gu *et al*, 2002; Rodriguez *et al*, 2004). Here the complex relationships between alleles can be observed, suggesting that the VNTR mutates more rapidly than the microsatellite (one *TH01* allele associates with multiple *INS* VNTR alleles).

Linkage disequilibrium analysis between a complex microsatellite near the *GH1* (growth hormone) locus and both SNP and SNP haplotypes (Day *et al*, 2004). This demonstrates the different lineages within the haplotypic and microsatellite data, with "large" and "small" microsatellite groups being marked by SNP alleles and haplotypes.

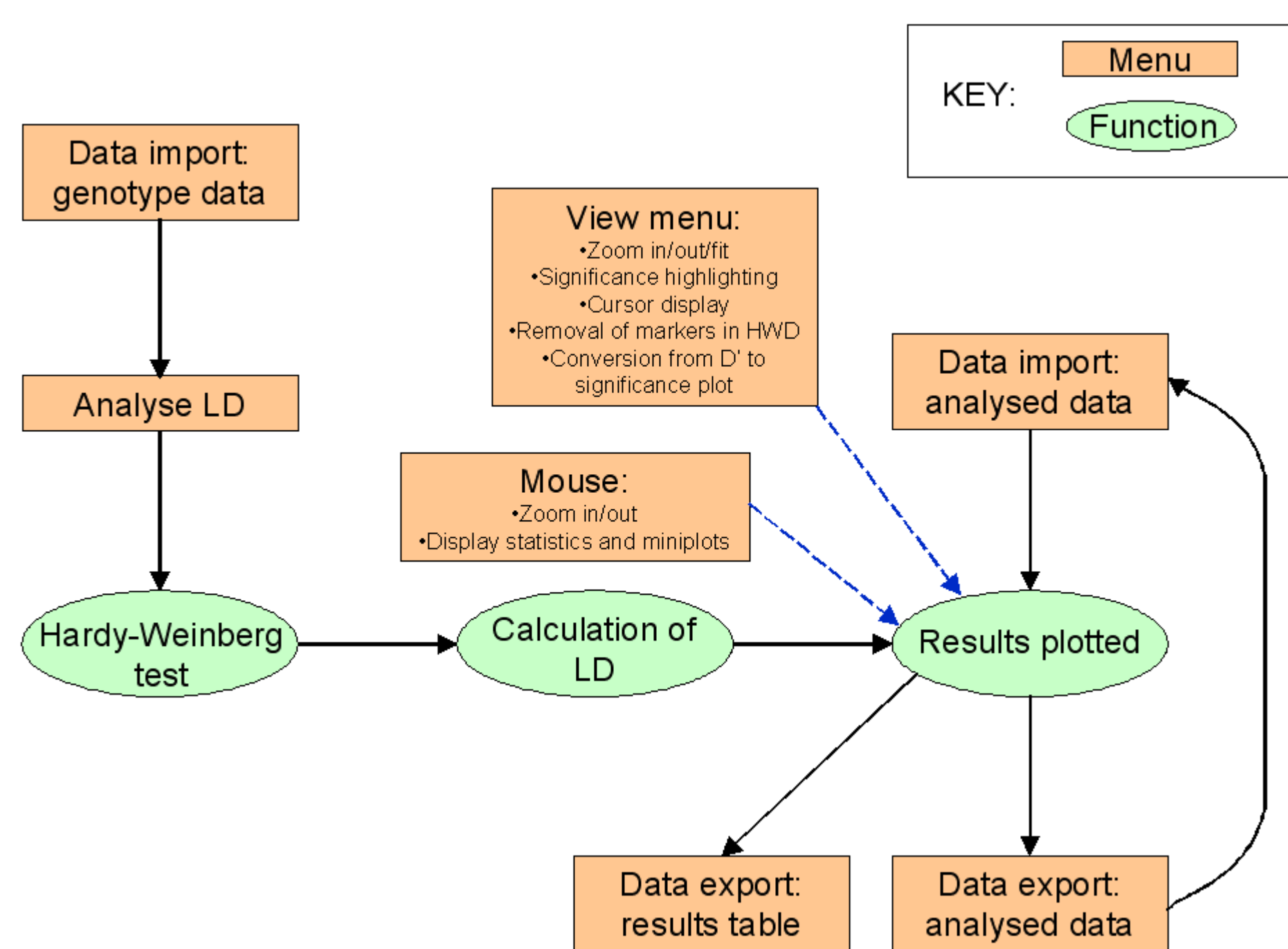
Regions of perfect linkage disequilibrium (Gaunt *et al*, 2006). This MIDAS view of HapMap (The International HapMap Project, 2003, www.hapmap.org) data around the *BRCA1* locus on chromosome 17 demonstrates the usefulness of the broad view. Here you can see perfect LD as pairs with only two haplotypes. A view of SNP major allele frequency across this region confirms that the majority of SNPs are equifrequent.

Evolutionary history (Gaunt *et al*, 2006). The layout of the SNP-SNP haplotype plots is informative for evolutionary history of the markers, suggesting which mutation arose first.

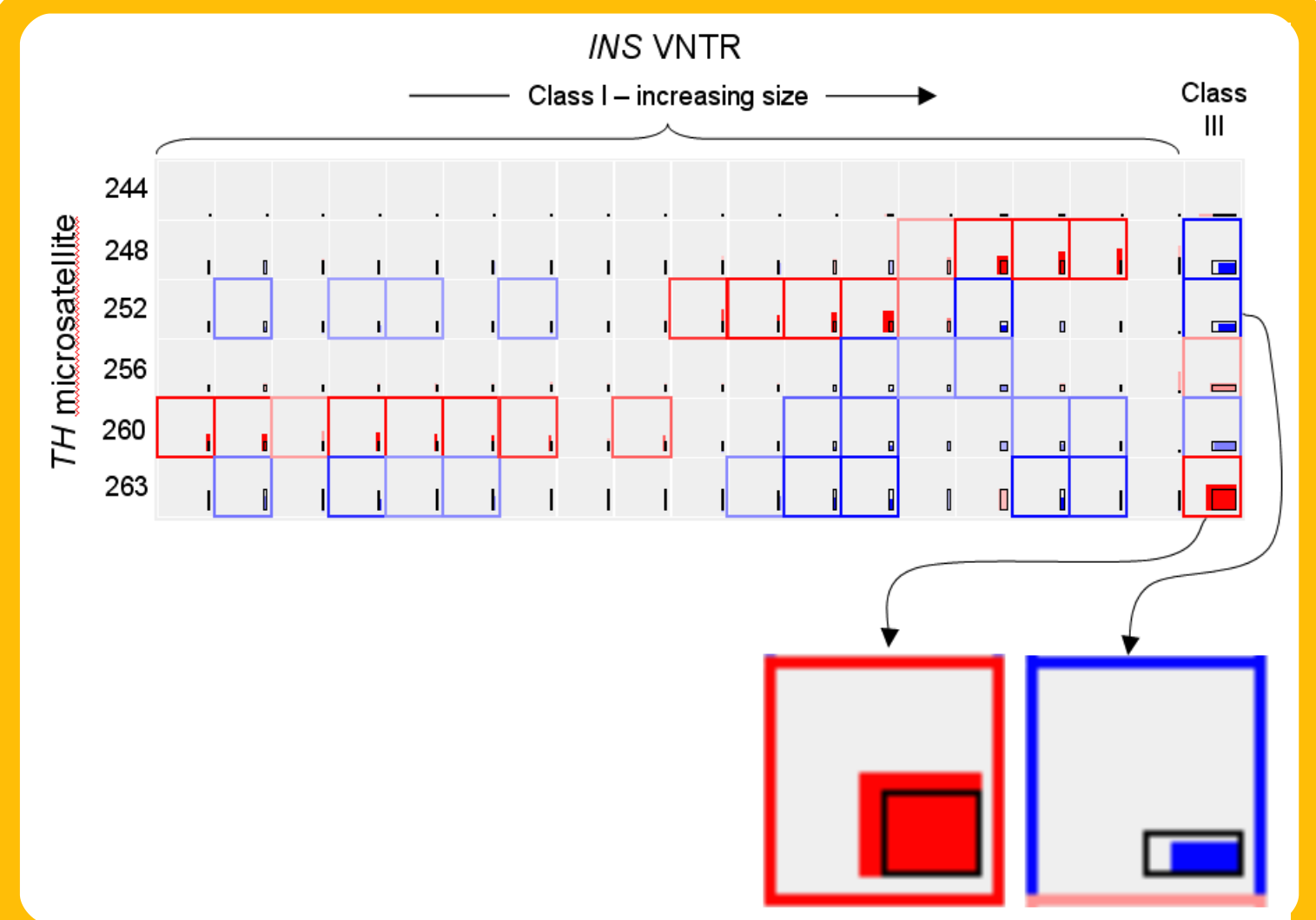
How the data are plotted



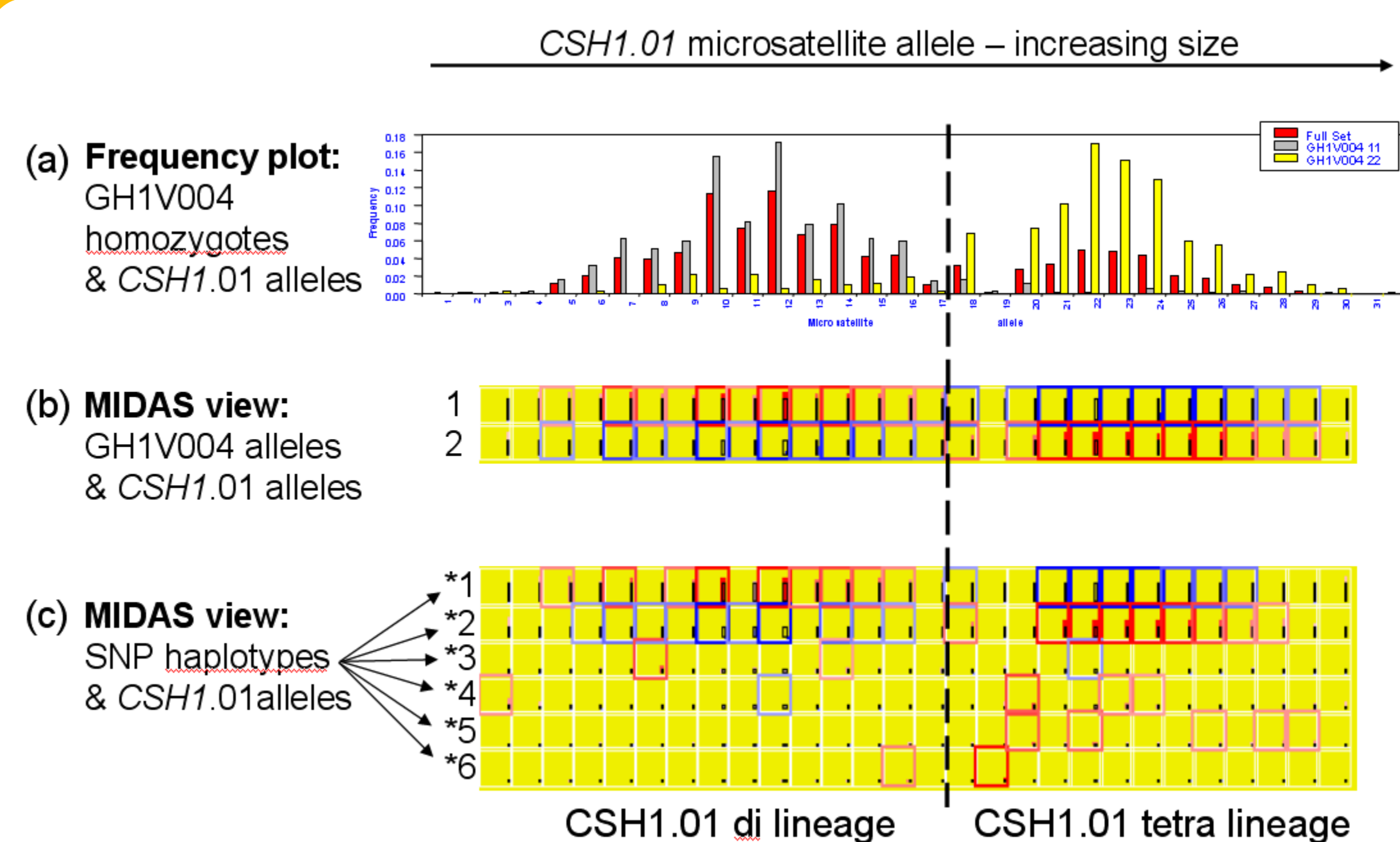
Program schematic – how you use MIDAS



LD between *INS* VNTR and *TH01*



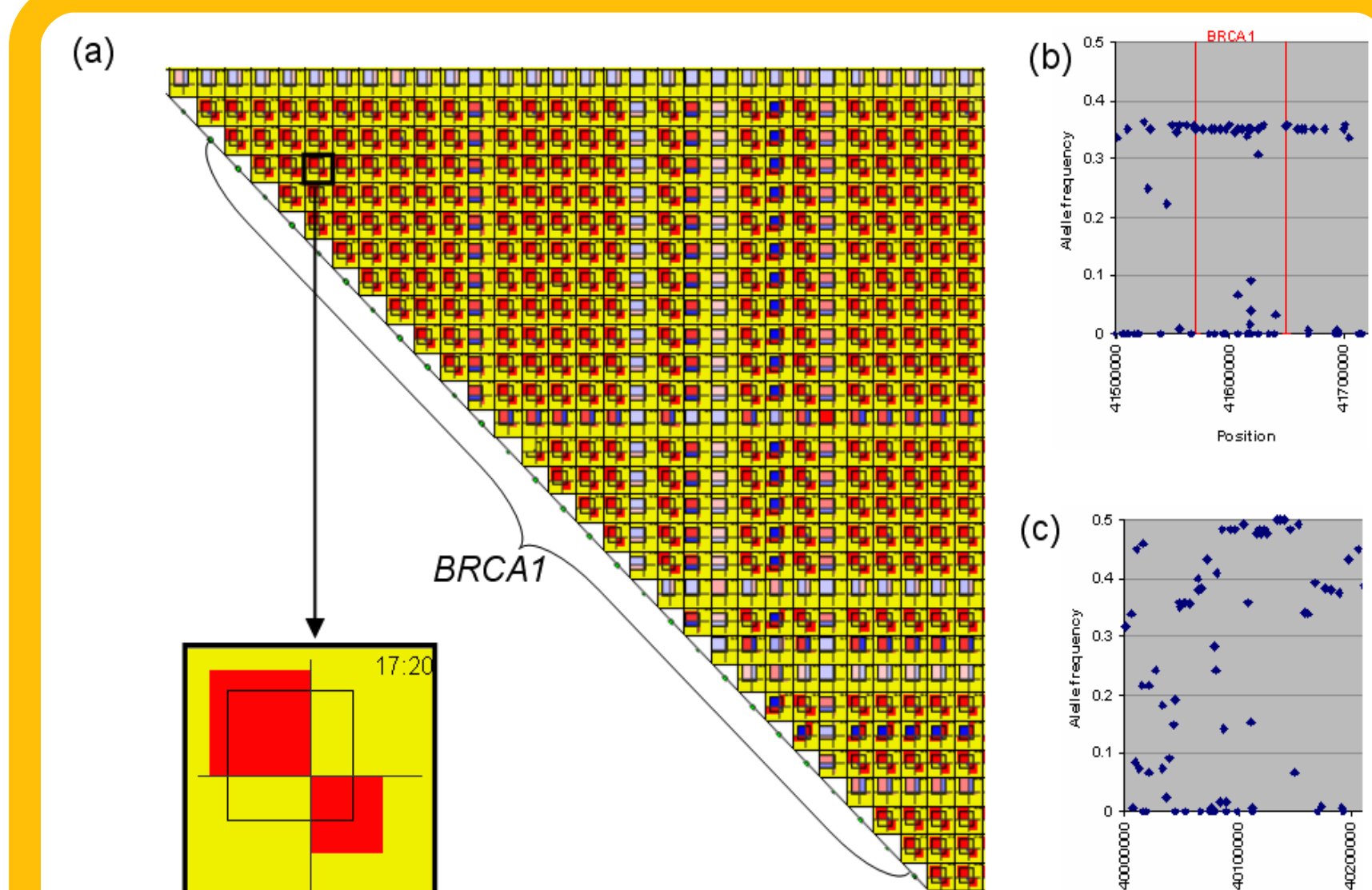
LD between haplotype and microsatellite



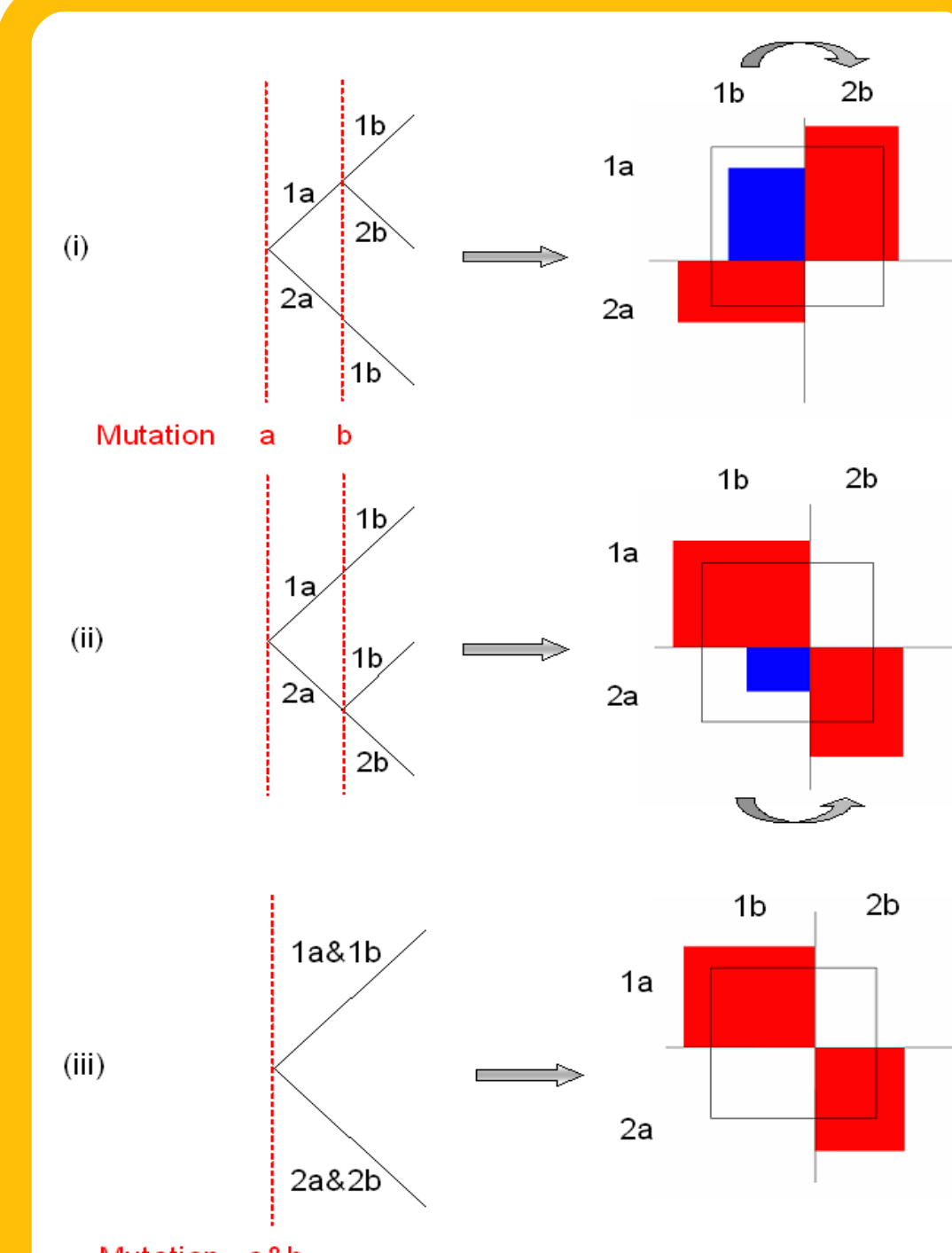
Download

<http://www.genes.org.uk/software/midas>

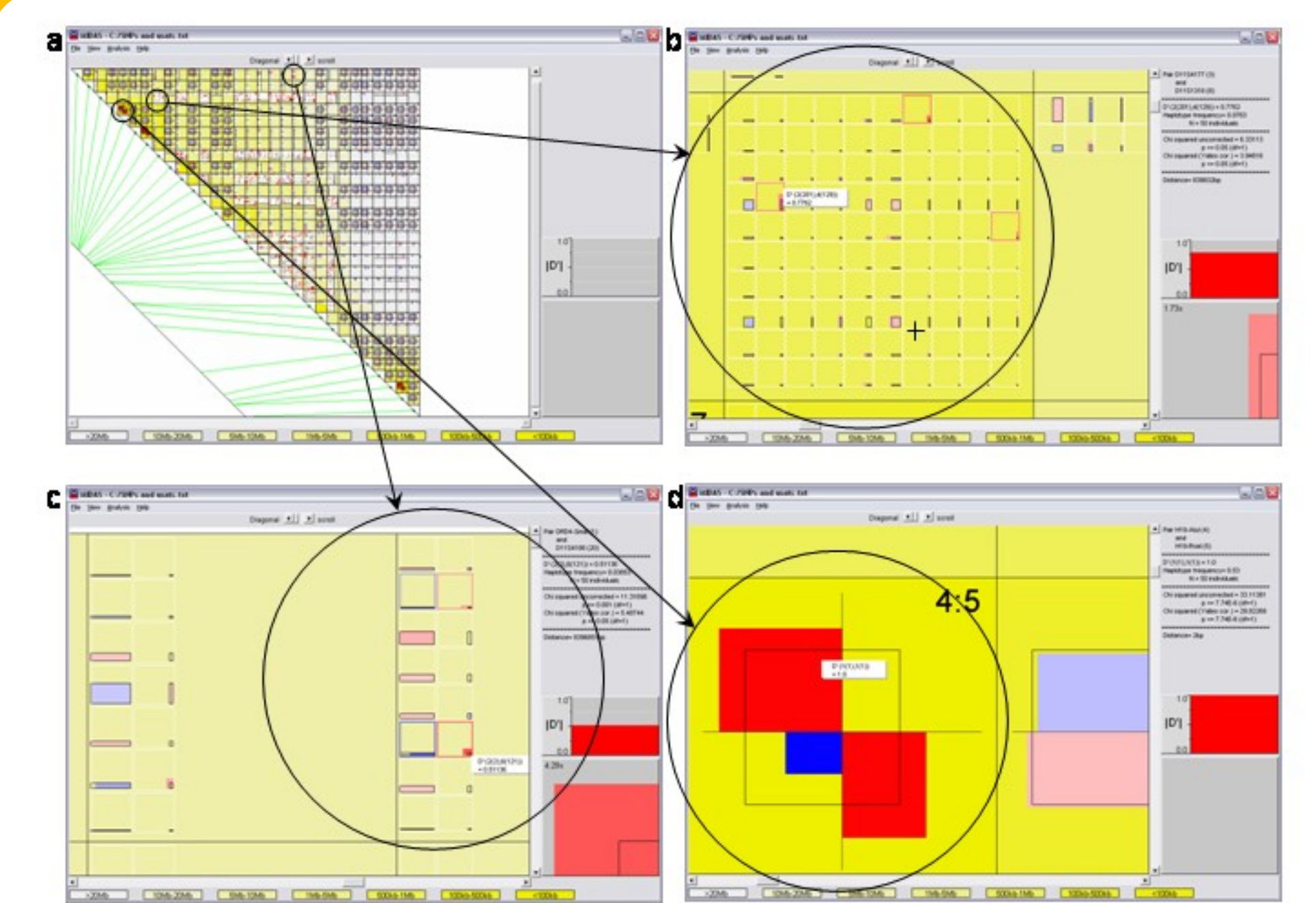
Regions of perfect LD



Evolutionary history



Screenshots



Summary

MIDAS is a graphical linkage disequilibrium analysis program that incorporates analysis of interallelic linkage disequilibrium between multiallelic and biallelic markers. The program generates statistics to enable quantitative analyses, but also provides an important overview of LD across broad genomic regions, enabling qualitative interpretations of large datasets.

Statistics presented include D' , r^2 , allele frequencies, haplotype frequencies, significance (χ^2) and pairwise distances. The program is written in the Python programming language, enabling easy adaptation to other uses.

The MIDAS homepage is at <http://www.genes.org.uk/midas>.

References and Acknowledgements

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).

References

- Gaunt TR, Rodriguez S, Zapata C, Day IN. MIDAS: software for analysis and visualisation of interallelic disequilibrium between multiallelic markers. *BMC Bioinformatics*. 2006 Apr 27;7(1):227
- Day IN, Chen XH, Gaunt TR, King TH, Voronovan 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.
- Hill WG: Estimation of linkage disequilibrium in randomly mating populations. *Heredity* 1974, 33: 229-239.
- 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.
- 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.
- Rodriguez S, Gaunt TR, O'Dell SD, Chen XH, Gu D, Hawe E *et al.*: Haplotype analyses of the IGF2-*INS*-*TH* gene cluster in relation to cardiovascular risk traits. *Hum Mol Genet* 2004, 13: 715-725.
- The International HapMap Project. *Nature* 2003, 426:789-796.
- The Python Programming Language. <http://www.python.org>. 2005.
- Zapata C, Rodriguez S, Visedo G, Sacristán F: Spectrum of nonrandom associations between microsatellite loci on human chromosome 11p15. *Genetics* 2001, 158: 1235-1251.