ABSTRACT

Motivation: Unravelling the genetic architecture of complex traits requires large amounts of data, sophisticated models and large computational resources. The lack of user-friendly software incorporating all these requisites is delaying progress in the analysis of complex traits.

Methods: Linkage disequilibrium and linkage analysis (LDLA) is a high-resolution gene mapping approach based on sophisticated mixed linear models, applicable to any population structure. LDLA can use population history information in addition to pedigree and molecular markers to decompose traits into genetic components. Analyses are distributed in parallel over a large public grid of computers in the UK.

Results: We have proven the performance of LDLA with analyses of simulated data. There are real gains in statistical power to detect quantitative trait loci when using historical information compared with traditional linkage analysis. Moreover, the use of a grid of computers significantly increases computational speed, hence allowing analyses that would have been prohibitive on a single computer.

Availability: The authors have implemented LDLA within the freely available GridQTL software (www.gridqtl.org.uk).

Contact: jules.hernandez@ed.ac.uk

1 INTRODUCTION

Most studied phenotypes, such as height, weight or disease risk, are the outcome of complex interactions between multiple environmental and genetic variables. Thus, the task of unravelling the genetic architecture of those traits requires large amounts of data, sophisticated models capable of optimally using all available information, and sufficient computational power. Scientific progress has been hampered by the lack of user-friendly software that combines sophisticated models and computational power. GridQTL is our contribution to advancements in this area (Seaton et al., 2006). It is a web application running genetic analyses on a computer grid. GridQTL will replace QTL Express (Seaton et al., 2002), a widely used package to map Quantitative Trait Loci (QTL) in structured populations. Currently, GridQTL contains a set of independent analyses, some inherited from QTL Express and others developed anew, and administrative tools as indicated by each of the tabs in Figure 1.

QTL Express consisted of five desktops and a server, and therefore repeatedly experienced ‘clogging’ due to usage overload. This limitation prompted the migration of QTL Express to the grid, which can provide several orders of magnitude more computer power but also requires a new set of technologies to be implemented to manage resources efficiently. The grid has been defined as a heterogeneous set of large computational resources geographically dispersed (Foster, 2005). For the purposes of GridQTL, the grid currently includes the National Grid Service (NGS) of UK (www.grid-support.ac.uk), the Edinburgh Compute and Data Facility (Egrid) (www.ecdf.ed.ac.uk) and a local Condor pool (http://www.cs.wisc.edu/condor).

Although well-thought experimental designs have been, and will continue to be, powerful tools to map genes underlying complex traits, they are not always feasible (e.g. in humans), are always expensive and lengthy (because it is necessary to set up an experimental population, arrange matings, wait for offspring to be born, etc.), results may not apply to general populations (e.g. different alleles segregating in different populations) and mapping resolution is usually poor (i.e. large confidence intervals around the best QTL location because of the limited number of recombination events in short pedigrees). Therefore, a direct analysis of extant populations is desirable.
In this article, we will describe in detail the linkage disequilibrium and linkage analysis (LDLA) module within GridQTL. Like the other analyses in GridQTL, LDLA has been designed to search for Quantitative Trait Loci (QTL), which are chromosomal regions likely to contain genes related to the trait under investigation. The difference is that it does so in general populations by partitioning phenotypic variance into several genetic and environmental components, assuming random QTL effects and contrasting the nested hypotheses of presence versus absence of QTL at user-specified chromosomal locations (George et al., 2000). The choice of a random QTL effects model is appropriate because, unlike in experimental crosses between inbred lines, there is no prior knowledge on the number of QTL alleles segregating in the sample. It can be applied to any population because it is based on flexible mixed linear models which can account for any type of pedigree whilst estimating genetic components of variance (additive, dominant, epistatic) based on identity-by-descent (IBD) covariances between individuals at putative QTL locations. Variances are estimated via restricted maximum likelihood (REML) (Patterson and Thompson, 1971), using the specialized software ASReml (Gilmour et al., 2002).

The most exciting feature of LDLA is that it can also account for some forms of population history. In essence, what LDLA is trying to do is to use all recombination events since population foundation. Modelling the history of a population allows researchers the opportunity to estimate genetic relationships among pedigree founders more accurately. In traditional linkage analysis (LA), pedigree founders are considered unrelated and non-inbred, however this is not correct when a pedigree is a snapshot of the last few generations in a population.

In LA, information comes from linkage disequilibrium (LD) between markers and QTL within pedigrees. Here, strong LD is expected over medium to long distances, generally covering a few tens of centiMorgans (cM), thus being ideal for coarse location of novel QTL using limited resources, e.g. typing a few hundred markers over the entire genome. Fine QTL mapping, i.e. loosely defined as achieving narrow confidence intervals spanning less than tens of centiMorgans, thus being ideal for coarse location, generally covering a few generations in a population.

2 METHODS

2.1 IBD

The crucial parameters in LDLA are IBD probabilities between all alleles at a hypothetical QTL position, denoted by y. Two alleles are IBD when they are exact copies of the same ancestral allele (Malécot, 1948). The higher y between any two QTL alleles, the more similar their effects are expected to be. Hence, y is proportional to the knowledge of grandparental origin of alleles. Phased markers beyond the closest bracketing pair do not provide additional information to infer y. If phased markers are not found then this method reverts to expected relationships relative to a pedigree. For example, assume a pedigree founder has a heterozygous marker locus, AB, 0.1 recombination rate units (denoted by c) away from the QTL position. If this individual has two progeny, which have unambiguously inherited the same marker allele, let us say A, then the probability of them having also inherited IBD alleles at the QTL would be (1−c)^2 = 0.82. Otherwise, if the progeny unambiguously inherited different alleles at the marker, i.e. A and B, from the same parent then y, at the QTL diminishes to (1−c)c = 0.18. If these progeny had no informative markers or, equivalently, c = 0.5 then the probability of both inheriting the same QTL allele is 0.5, i.e. the expected IBD sharing between two full-sibs regardless of marker information.

Table 2 in Pong-Wong et al. (2001) contains a full set of y, probabilities between pedigree founders and their offspring given marker information, and Equation (1) and Table 1 show how the remaining y, probabilities can be calculated recursively. For example, y, between alleles from individuals i and j, where i is not a founder and is older than j, can be calculated recursively via the y, probabilities between i and the parents of j. This recursive method requires marker phases to be known, and therefore is unsuitable for the first two generations in a pedigree when genotypes but not haplotypes are available. This deterministic method of calculating y, is much faster than and almost as accurate as a stochastic method called LOKI developed by Heath (1997) (Sorensen et al., 2001). It is less accurate than LOKI when a low-density SNP panel is used, however not when microsatellites or any type of markers at high density are used.

2.2 Pedigree IBD

We implemented a deterministic method to calculate y, based on Pong-Wong et al. (2001). This method is applied recursively, starting with the oldest individuals in the pedigree, and taking into account the pattern of haplotype transmission from parents to offspring. Here, a haplotype consists of the closest marker, or marker bracket, to the QTL for which phase is known. Phase is equivalent to the knowledge of grandparental origin of alleles. Phased markers beyond the closest bracketing pair do not provide additional information to infer y. If phased markers are not found then this method reverts to expected relationships relative to a pedigree. For example, assume a pedigree founder has a heterozygous marker locus, AB, 0.1 recombination rate units (denoted by c) away from the QTL position. If this individual has two progeny, which have unambiguously inherited the same marker allele, let us say A, then the probability of them having also inherited IBD alleles at the QTL would be (1−c)^2 = 0.82. Otherwise, if the progeny unambiguously inherited different alleles at the marker, i.e. A and B, from the same parent then y, at the QTL diminishes to (1−c)c = 0.18. If these progeny had no informative markers or, equivalently, c = 0.5 then the probability of both inheriting the same QTL allele is 0.5, i.e. the expected IBD sharing between two full-sibs regardless of marker information.

Table 2 in Pong-Wong et al. (2001) contains a full set of y, probabilities between pedigree founders and their offspring given marker information, and Equation (1) and Table 1 show how the remaining y, probabilities can be calculated recursively. For example, y, between alleles from individuals i and j, where i is not a founder and is older than j, can be calculated recursively via the y, probabilities between i and the parents of j. This recursive method requires marker phases to be known, and therefore is unsuitable for the first two generations in a pedigree when genotypes but not haplotypes are available. This deterministic method of calculating y, is much faster than and almost as accurate as a stochastic method called LOKI developed by Heath (1997) (Sorensen et al., 2001). It is less accurate than LOKI when a low-density SNP panel is used, however not when microsatellites or any type of markers at high density are used.

2.3 Historical IBD

This variable measures long-term genetic relationship among QTL alleles with respect to an ancestral founder population. Because a pedigree reflects the whole history of a population, from foundation to present generation, LDLA models history from population founders to pedigree founders. The simplest mathematical model assumes a Fisher–Wright population in which genetic drift is the only evolutionary force acting, generations are discrete, size is constant and population founders are assumed unrelated, non-inbred and in linkage equilibrium. This population model...
The same probability was estimated as 0.4701 and 0.4703 with the methods H&HS are more accurate than R at estimating given haplotype information and correct history parameters M&G and R over both M&G and H&HS is that it only requires genotypes instead P among population founders (Handy et al. 2006). The real advantage of R over both M&G and H&HS is that it only requires genotypes instead of haplotypes. This makes R the fastest method to compute. However, given haplotype information and correct history parameters M&G and H&HS are more accurate than R at estimating y3. However, this extra accuracy contributes only marginally to the overall power of QTL detection (Hernández-Sánchez et al. 2006). The advantage of H&HS over M&G is that we have developed theory to incorporate mutation, migration and population size fluctuations in the model, but these are yet to be implemented in LDLA.

A fourth method (Meuwissen and Goddard, 2007) will also be implemented shortly. This latter approach differs from M&G in that a population is assumed to be in mutation-drift equilibrium, thus precluding the need to estimate the age of a population. The code for the latter method was provided by Meuwissen, which is in agreement with the open source spirit of GridQTL.

All of these methods impute y2 at a putative QTL given identity-by-state (IBS), or homozygosity, at several linked markers simultaneously, under a specific population history model. Intuitively, two identical haplotypes are more likely to contain IBD alleles at a QTL than two completely different haplotypes. Nonetheless, the strength of this statement depends, in addition to haplotype similarity, on the type of history modelled, and the genetic distances between loci. For example, markers unlinked to the QTL do not provide information, and y3 increases with T given Np and P0.

The three methods implemented to calculate y2 render very similar results under ideal conditions, e.g. true T, Np, P0 and known phases. For example, y2 at a QTL given a homozygous marker 1 cM away, and initial allele frequency among population founders of 0.5 was predicted to be 0.473 with M&G whilst simulations gave 0.468 (Table 2 in Meuwissen and Goddard, 2001). The same probability was estimated as 0.4701 and 0.4703 with the methods of H&HS and R, respectively. We decided to keep all three methods because: (i) M&G was the most robust in some circumstances; (ii) H&HS and R can be readily extended to account for other population histories; and (iii) R was the only method capable of working with genotypes (the other two required haplotypes).

2.4 Combining pedigree and historical IBD

Mathematically, the combined estimate of y3 and y4 for any pair of alleles was done as follows

\[ y = y_3 + \left(1 - y_3\right) y_4 \]  \hspace{1cm} (1)

(Meuwissen and Goddard, 2001). This equation gives a strong weight to y3, as it should, because its estimation is free from population history assumptions, and hence robust to incorrect population models. In contrast, y3 relies on population assumptions, i.e. a Fisher–Weight model with specific Np, T and P0 values, and therefore it should only add to the similarity explained by the pedigree. When a complete pedigree is known across all generations, i.e. T=0, or when y3 = 1, y4 should add nothing to y.

Not all pedigree members may have genotypes. In fact, genotypes may be available only in the last few generations of pedigree, and a variable number of markers may have been genotyped per individual. However, accurate y3 predictions are based on the IBS pattern of several relatively close markers and, thus, would be poorly predicted among ungenotyped (or poorly genotyped) individuals. So, in order to combine y3 and y4 probabilities, it is useful to classify individuals into three categories: pre-base, base and post-base. The most important category is the base. Pre-base individuals are all ancestors of base individuals. Post-base individuals are all descendants of base individuals. Base individuals fulfill two requirements: (i) they must have a minimum proportion of genotyped markers which is defined by the user; and (ii) they cannot have base ancestors. Next, we calculate y3 and y4 for pre-base and base individuals and apply Equation (1) to obtain y. Second, we obtain y3 among post-base individuals recursively from y among base individuals. Note that in this case, y3 already incorporates y3 through y from previous generations. This protocol ensures that the recursive process starts from highly informative individuals (base) as opposed to starting by default from pedigree founders, which could have no marker information. There can be other ways of combining both sources of information. Other LDLA initiatives (e.g. Lee and Van der Werf, 2008; Meuwissen and Goddard, 2001) generally assume pedigree founders are all base individuals independently of whether they are genotyped or not (or may trim the pedigree so that all base individuals have genotypes).

2.5 How to use LDLA

LDLA requires the following information: pedigree, genotypes or haplotypes, inter-marker distances and map order, traits, explanatory variables (optional) and historical parameters. The upload and check button starts the process of data checking and recoding. As part of the checking procedure, LDLA detects inconsistencies in sex, individuals phenotyped and/or genotyped but not in pedigree, repeated entries and Mendelian inheritance errors in genotypes.

Mendelian errors can reduce the power of QTL detection, or even generate false positive results (Xu et al., 2002). Therefore, it is crucial to clean the data from all detectable errors before proceeding with further analysis. Our program can scan multiple generations simultaneously, thus being able to detect errors among ancestors and descendants even if some intermediate generations have not been genotyped.

Missing alleles will be recovered whenever possible under two conditions. First, there must be no Mendelian errors in the data, and second, there is only one possible genotype. We attempt to recover missing alleles given genotypes in family trios and entire full-sib families. Once the data have been recorded and contain no observable errors, the next set of model options appears (Fig. 2). The first one is a box with all recorded traits (T), covariates (X1) and factors (X2). Only single trait analyses are possible at present. Different models can be built by ticking a trait and different combinations of fixed effects, covariates and random effects. The next option is whether to save all generated IBD matrices, called G matrices, for each location tested. These matrices are large and dense and, therefore, potentially cause problems with storage and transmission of data over a network. Hence, the default is not to save them. Next, users can select one of three different methods to estimate y3, i.e. M&G, H&HS or R.

Currently, only single QTL models are implemented. Both additive and dominant effects can be estimated. Two QTL models that allow for epistatic interactions are currently being developed and tested. Despite a recent negative report on the chances of detecting epistasis in general populations (Hill et al., 2008), other studies show some hope (Lee and Van der Werf, 2008). Testing for the presence of a QTL can be done at regular intervals or at pre-specified locations. The former requires a number being typed in the tab labelled Every—cM. The locations for the latter can be directly entered through the web interface [At—cM (Hand)] or uploaded from a text file [At—cM (File)]. Information about the history of the population can be entered through the box named Demographic History. The history file must contain five parameters: Np for males and females, T, a random mating system and the minimum percentage of fully genotyped loci for an individual to be classified as base. Both Np parameters are assumed constant over time. The random mating system can be chosen with an integer: 0 for haploids, 1 for diploids monochrome with selling, 2 like 1 without selling, 3 diploid diocious and 4 like 3 but with a hierarchical mating typical in animal breeding where...
a few males are mated to many females. In this last case, the ratio must be fixed, so for instance if 10 males are mated to 50 different females each, then \( N_e \) for each sex must be exactly these numbers. Finally, a panel named ASReml Instructions allows experienced users to input complex models not feasible with the options in Figure 2. For example, if there is a block error structure as in plant breeding trials or if random regressions or splines are to be used.

The analysis starts after clicking on the Start button and the status of the grid usage can be viewed after ticking the box named Display Grid Activity. At this stage, all chromosomal positions tested are distributed as independent jobs in the grid. A progress bar will show how much of the current analysis is finished, how much is being analysed and how much is queuing. Figure 3 shows an example in which only the clusters in Oxford, Leeds and Manchester were used, but this changes from run to run according to a clever scheduling procedure to minimize queuing time (Section 3). Users have the option to cancel the whole analysis if it takes too long or they suspect there is something wrong with the process. Users will be notified with a message when the analysis finishes. If they logout, the message will appear after logging in again. Once they click OK, two plots will appear on their screen. The first plot is for the QTL profile in terms of likelihood ratio test (LRT) between the hypotheses of presence versus absence of QTL in terms of position (cM on the x-axis) (as in Figure 4). The second will show all genetic heritabilities, i.e. polygenic, additive and dominant QTL, if the corresponding model term has been requested. Finally, users will be able to download a zipped file with all information.

### 2.6 Additional information

In addition to QTL profile, heritabilities and REML estimates of variances and fixed effects, our application gives users further information. For each trait, the mean, phenotypic SD, number of missing records and frequency distribution (with 1 SD bin size centred around the mean) are given. The number of ancestors for each individual in the pedigree is also calculated. This number is 0 for founders and a multiple of 2 for non-founders, because every time a single parent is missing, a new dummy parent, who is also a new founder, is introduced in the pedigree. In the absence of birthdates, this statistic can be used to group individuals in discrete generations. The distribution of the number of different mates per father and mother are also calculated, e.g. the number of fathers (mothers) with one mate, with two different mates, etc. This distribution gives information about the mating structure, and half-sib family sizes (common in some animal and plant species). The marker information consists of estimates of heterozygosity and allele frequencies per marker, and the average (and SD) inter-marker distance (in cM).

### 2.7 The grid

As part of the GridQTL project, the LDLA portlet was ported on the grid. The LDLA is parallelized as a set of independent tasks (each task is the analysis at a different hypothetical QTL location) to be distributed on the available computational resources. These resources are a local Condor pool (connecting four local computers with six CPU in total) and the NGS of UK with four 256 CPU high-performance clusters (HPC) located in the universities of Oxford, Leeds, Manchester and in the Rutherford Appleton Laboratory (Oxfordshire), plus an additional 128 CPU HPC at the University of Oxford and a 1456 CPU HPC at the University of Edinburgh.

To distribute tasks efficiently over this heterogeneous set of resources that has not central administration, we developed the scheduler SWARM (scheduling with a request multiplication). SWARM aims to minimize queuing time for distributed analyses on the grid, and restarts tasks which fail in order to achieve a robust service (J.A. Grunchec et al., submitted for publication). SWARM also provides dynamic queuing statistics to aid users in monitoring progress.

Data security is dealt with as follows. First, the GridQTL administrator has a personal electronic certificate required to access user accounts in the grid. Second, asymmetric encryption and decryption techniques ensure data protection during transfer between server and HPCs. Third, data are also protected during transfer between user and server using secure http connections. Fourth, only the GridQTL administrator has access to the server.
A beta version of LDLA was released in March 2008. The objectives of this beta release were to familiarize users with data formats, get their feedback and suggestions, identify and eliminate code glitches and assess the computational requirements and performance of LDLA in a grid. Table 1 shows some of the increases in computational speed due to the grid. Some users required hundreds of CPU hours per full scan (Table 1). Grid time is the total grid computing time required to complete each analysis. User time is the waiting time between submitting an analysis and obtaining results.

### 3 RESULTS

#### 3.1 Simulated data analysis

A single QTL explaining ~5% of the phenotypic variance (polygenic and error variances set to 1, QTL variance set to 0.1) was simulated at 7 cM from the start of a chromosomal segment of 10 cM. An ideal Fisher–Wright population was modelled with parameters \( T = 100 \), \( N_e = 100 \) and \( P_0 \) randomly drawn from a uniform distribution. Trait, pedigree and genotypes were recorded for three further generations. Eleven SNPs were genotyped at positions 0, 1.5, 2.5 and then every cM until 10.5 and the QTL search was performed every cM starting at 0 cM. Figure 4 shows the average QTL profiles from 100 simulations for LA, LDLA using the correct population parameters and the two closest markers to each QTL position to estimate \( y_h \). LDLA as before but using an incorrect estimate of 50 for \( N_e \). On average, LDLA is more powerful than LA when the correct population history is modelled (see also Hernández-Sánchez et al., 2006; Lee and Van der Werf, 2008; Meeuwissen and Goddard, 2001). However, Zhao et al. (2007) found that, under simple genetic models, LDLA was not more powerful than marker–trait association analyses. Several things are worth noting in Figure 4. First, the profile obtained with LA is flatter than those obtained with LDLA, due to the limited amount of recombination events captured in a short pedigree of three generations. Second, the LA profile is also lower at the maximum LRT than any of the LDLA profiles. Third, the best LDLA profile (highest LRT at 7 cM, the true QTL position) was obtained using the correct population parameters. Fourth, the LDLA profile located the QTL better than LA even using incorrect population parameters. LDLA is known to be robust to some departures from the idealized model (Hernández-Sánchez et al., 2006). However, a full study of robustness of LDLA is still required.

#### 3.2 Performance of the grid

A beta version of LDLA was released in March 2008. The objectives of this beta release were to familiarize users with data formats, get their feedback and suggestions, identify and eliminate code glitches and assess the computational requirements and performance of LDLA in the grid. Table 1 shows some of the increases in computational speed due to the grid. Some users required hundreds of CPU hours per full scan (Table 1). Grid time is the total grid computing time required to complete each analysis. User time is the waiting time between submitting an analysis and obtaining results.

<table>
<thead>
<tr>
<th>Date and hour</th>
<th>Grid time (h)</th>
<th>User time (min)</th>
<th>Speed-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>27/10/08 10:45</td>
<td>39.7</td>
<td>53.1</td>
<td>44.9</td>
</tr>
<tr>
<td>30/10/08 07:26</td>
<td>28.1</td>
<td>11.4</td>
<td>147.3</td>
</tr>
<tr>
<td>30/10/08 07:42</td>
<td>28.1</td>
<td>11.7</td>
<td>144.8</td>
</tr>
<tr>
<td>30/10/08 08:19</td>
<td>29</td>
<td>12</td>
<td>144.8</td>
</tr>
<tr>
<td>30/10/08 09:29</td>
<td>29.5</td>
<td>13.9</td>
<td>217</td>
</tr>
<tr>
<td>01/11/08 08:57</td>
<td>31</td>
<td>14.4</td>
<td>129.2</td>
</tr>
<tr>
<td>05/11/08 10:49</td>
<td>33.7</td>
<td>14.3</td>
<td>141.8</td>
</tr>
<tr>
<td>05/11/08 11:19</td>
<td>35.4</td>
<td>18.5</td>
<td>114.6</td>
</tr>
<tr>
<td>08/11/08 01:20</td>
<td>286.5</td>
<td>190.9</td>
<td>90</td>
</tr>
<tr>
<td>08/11/08 21:49</td>
<td>29</td>
<td>14.8</td>
<td>117.9</td>
</tr>
</tbody>
</table>

Example with 10 of the largest successful LDLA runs. Date and hour, GMT at which LDLA was used; Grid time, hours of grid computation; User time, users’ waiting time in minutes; Speed-up, grid time/user time (in the same time unit).

The ratio between grid and user times is the speed-up factor which ranges from 45 to 147. Thus, the comparison is between the grid as a whole and an average CPU from the grid (equivalent to a single desktop). These results were collected at different times over different days so performance in the grid was assessed in both busy and idle periods. We have demonstrated that significant reductions in user time are to be expected when running LDLA across multiple locations in the grid. The grid will allow users to run analyses that are prohibitive in single desktops, such as complete genome scans with high-density marker maps and thousands or hundreds of thousands of records.

### 4 DISCUSSION

#### 4.1 Key contributions

The LDLA module in GridQTL allows QTL analyses in general populations using either linkage information alone or combined with LD information. LDLA uses a computer grid to submit jobs in parallel, hence greatly reducing the burden of computationally expensive analyses. Moreover, LDLA allows modelling different population histories (through varying \( N_e \), \( T \) and \( P_0 \)) using a user-defined amount of marker information. Hence, LDLA should become a powerful modelling tool for geneticists. There have been other attempts at producing software to perform LDLA, notably by S.H. Lee (personal communication) but, to our knowledge, our work is the first offering a web-based version linked to a grid of computers, with data storage capacity for multiple users. At present, we devote one server to host the main GridQTL portal, and a desktop to host the beta version of LDLA for development purposes. Eventually, LDLA will be transferred to the main GridQTL portal.

#### 4.2 LDLA assumptions

LDLA assumes a Fisher–Wright population model. Strong deviations from this model may be detrimental to LDLA. Moderate deviations within the general model, i.e. supplying incorrect population parameters \( N_e \), \( T \) and \( P_0 \) for analysis, do not have a major effect on performance. Nevertheless, a thorough robustness analysis in this area is urgently required. At present, our recommendation is that LDLA results should be compared with LA results. LA is robust when using incorrect population parameters. LDLA is known to be robust to some departures from the idealized model (Hernández-Sánchez et al., 2006). However, a full study of robustness of LDLA is still required.

![Fig. 4. Results from simulated data analysis. One SE around the mean LRT is given for LDLA and LA. LDLA with bad \( T \) or \( N \) had similar SE.](image-url)
to incorrect population assumptions and therefore can provide a base line against which LDLA can be compared. A large significant result obtained with LDLA and a very insignificant result obtained with LA may indicate a false positive result. However, we have not observed this behaviour in either simulations or real data analyses. Our experience is that LDLA modifies LA profiles, sometimes rising above them, but sometimes going below.

4.3 Future developments

There are a number of features we plan to add to the current LDLA to enhance its performance.

1. Simultaneous searching for two QTL.
2. Haplotypeing.
3. Inclusion of other random effects via the interface panel.
4. Bivariate trait analysis.
5. Selection of optimum number of markers to predict \( y_h \).
6. Extension to other population models.
7. Inclusion of \( y_h \) prediction in populations at equilibrium.

First, searching for two QTL simultaneously will allow users to fit epistatic interactions in their models. Epistasis must be present in most if not all complex traits (Carlborg and Haley, 2004), and simulation work has shown that detecting it is possible with LDLA (Lee and van der Werf, 2008). Second, haplotypes are essential for estimating \( y_h \) with methods H&HS or M&G Haplotypeing can be very computationally intensive, therefore we are developing stand-alone software to be run independently of LDLA. We use the minimum recombinant paradigm approach of Qian and Beckmann (2002) to obtain haplotypes from genotypes and pedigrees. Third, random effects other that polygenic and associated to QTL effects can be important in some circumstances. For example, maternal effects can greatly affect traits measured early in life. Fourth, bivariate trait analysis can increase power of QTL detection when traits are correlated. However, analysing more than a few traits simultaneously with mixed models can be technically problematic and other approaches may be more suitable. Fifth, we have observed that the accuracy of \( y_h \) prediction increases with the number of markers used until it asymptotes to a maximum value. It would be ideal to be able to predict the appropriate number of markers to use in estimation of \( G \) matrices for each hypothetical QTL location. Sixth, no population follows a strict Fisher–Wright model. Migration, mutation and variation in population size are common deviations from that model. Although we have developed statistical tools to account for some of these effects (Hill and Hernández-Sánchez, 2007), the question arises of which models are likely to apply to a sufficiently large number of populations to be worthwhile implementing. And seventh, a new method to predict \( y_h \) in populations at equilibrium will be implemented (Meuwissen and Goddard, 2007). The main advantage of this method is that estimation of \( T \) is no longer required. However, the question of whether populations are likely to be in equilibrium or not due to the disturbing forces of evolution remains.

4.4 Conclusions

We offer the scientific community a powerful tool to perform LDLA in complex traits in extant populations. LDLA takes advantage of population history information to increase the power and resolution of QTL detection with respect to traditional LA. Moreover, genome-wide scans in large datasets are now possible given the computational power of the grid. The beta version of LDLA can be found at http://cleopatra.cap.ed.ac.uk/gridsphere.

ACKNOWLEDGEMENTS

We thank professor J. Pemberton for proof reading and all LDLA users for their input.

Funding: Biotechnology and Biological Sciences Research Council of UK (BB/B/1695X to GridQTL project); Research Councils UK (GR/T727983/01 to S.K.).

Conflict of Interest: none declared.

REFERENCES

LDLA in a computer grid


