A new tool called DISSECT for analysing large genomic data sets using a Big Data approach

Oriol Canela-Xandri1, Andy Law1, Alan Gray2, John A. Woolliams1 & Albert Tenesa1,3

Large-scale genetic and genomic data are increasingly available and the major bottleneck in their analysis is a lack of sufficiently scalable computational tools. To address this problem in the context of complex traits analysis, we present DISSECT. DISSECT is a new and freely available software that is able to exploit the distributed-memory parallel computational architectures of compute clusters, to perform a wide range of genomic and epidemiologic analyses, which currently can only be carried out on reduced sample sizes or under restricted conditions. We demonstrate the usefulness of our new tool by addressing the challenge of predicting phenotypes from genotype data in human populations using mixed-linear model analysis. We analyse simulated traits from 470,000 individuals genotyped for 590,004 SNPs in ~4 h using the combined computational power of 8,400 processor cores. We find that prediction accuracies in excess of 80% of the theoretical maximum could be achieved with large sample sizes.
The astonishing rate at which genomic and genetic data are generated is rapidly propelling genomics and genetics research into the realm of 'Big Data'. This great opportunity is also becoming a big challenge, because success in extracting the useful information contained within these data will depend on our ability to analyse extremely large data sets with the most powerful statistical methods. The computational problems associated with 'Big Data' become critical when, for instance, fitting mixed-linear models (MLMs) and performing principal component analyses (PCA). These analyses are used in a wide range of fields ranging from predictive medicine and epidemiology, to animal and plant breeding, and pharmacogenomics. However, these calculations are so computationally expensive that, when applied to large data sets, one typically resorts to approximations, restricts the applicability to particular cases (for example, when the number of markers is small compared with the available sample size) or need at least one highly computationally demanding step such as performing the eigen decomposition of the relationship matrix. These workarounds are non-scalable and therefore could be impractical with increasingly large data sets.

As has been effectively demonstrated in other fields, such limitations can be overcome through moving to software capable of combining the computational power of thousands of processor cores distributed across the nodes of compute clusters and large supercomputers. To address this need, we developed DISSECT (http://www.dissect.ed.ac.uk), a new, highly scalable and freely available tool that is able to perform a large variety of genomic analyses across huge numbers of individuals. For increased versatility, the software also runs on single compute nodes (for example, desktops, regular workstations or single compute nodes on a cluster). Here we describe the methods underpinning our tool and demonstrate its usefulness by addressing the challenge of predicting phenotypes from genotype data in unrelated humans.

Phenotypic prediction is of central interest to disciplines such as quantitative genetics, animal breeding or human medicine and is one of the driving forces behind large-scale genotyping and sequencing projects in a wide range of species. Despite considerable efforts, predicting complex traits in unrelated humans has been an elusive goal. Accurate prediction of complex traits is expected to be strongly dependent on the availability of sufficiently large data sets and the capacity to analyze all these data together, which makes this an ideal challenge to showcase DISSECT's capabilities. We therefore simulate a cohort of half a million individuals and use DISSECT and the aggregated power of 8,400 processor cores to analyse it. We show that MLMs could be used to predict quantitative traits with increasing accuracy as the sample size of the training cohort increases and achieve over 80% of the theoretical maximum accuracy when the training cohort has 470,000 individuals. We study different scenarios where the genotyping array contains only ~20% of the quantitative trait nucleotides (QTNs) together with non-casual single-nucleotide polymorphisms (SNPs), contains all the QTNs together with non-casual SNPs or contains only the QTNs. The improvement in prediction accuracy obtained by including all the QTNs in the array is smaller than by removing the non-casual SNPs from the array, thus indicating the strong detrimental effect that the noise introduced by the non-casual SNPs has on prediction accuracy, even when using large sample sizes.

Results

Overview of DISSECT. Commonly used statistical analyses of genetic and genomic data are computationally intensive due to the requirement to perform different types of matrix operations. The computational requirements (that is, the compute and memory capacity that is required to perform linear algebra operations on these matrices) are usually a super-linear function of the number of markers and samples available. Therefore, the computational needs for the analysis of increasingly large data sets can rapidly surpass the computational capacity of single compute nodes. DISSECT is designed to overcome these compute and memory limitations by taking advantage of the aggregate power of the thousands of processor cores and memory that are distributed across clusters of networked compute nodes. For this purpose, it distributes the available data over the multiple nodes using a two-dimensional block-cyclic distribution scheme (Fig. 1 and Supplementary Note 1). This is convenient, because it achieves a good load balance by splitting the work reasonably evenly among the nodes available, maximizing efficiency and minimizing run time. At any given time, each node has access to only a small portion of the data on which it performs local computations. When the algorithm requires access to blocks of data currently held on other nodes, the nodes communicate to coordinate data redistribution (Supplementary Note 1). In addition, one node (the 'root' node) takes on the role of coordinating the work of all the other nodes and of performing small summarizing computations. If data sets are small enough to fit in the available memory of the root node, then the root node also handles the data input and output. However, when this is not possible, because the volume of data exceeds the available memory of the root node, then the processes of data loading or storing are also distributed across multiple nodes (Supplementary Note 1).

As the cores on a node cannot directly access data on the other nodes, the computational approach of distributing the data between nodes is necessarily more involved than parallelization of software that uses multiple cores within a single node. In addition, the distribution of workload introduces a relative loss of computational efficiency and scalability, because nodes need to communicate, with overheads determined by the speed of the network connection. Because of this, increasing the number of nodes does not guarantee a proportional reduction of computational time (Supplementary Fig. 1 shows the scalability of DISSECT as a function of the number of processor cores used). However, the broad applicability of this approach enables the analysis of data sets of sizes for which analysis is infeasible using the limited memory and computing capacity of a single compute node. Importantly, no mathematical approximation is required.

DISSECT was written in the programming language C++, using routines from the MPI and BLACS libraries, to handle the data distribution and the communication between nodes. The basic linear algebra computations are based on the ScaLAPACK libraries, which ensures optimal computational performance when using a performance-optimized implementation such as the Intel Math Kernel Library. DISSECT can be used on computer clusters the size of which may vary from a few tens or hundreds of processor cores to large supercomputers with hundreds of thousands of cores. The sole requirement is that an MPI implementation be available on the machine. Our software also allows the user to take full advantage of multi-core capabilities on more modest, single-node workstations with a performance similar to software designed for running on single compute nodes (Supplementary Fig. 2). DISSECT is as easy to use as other commonly used software such as GCTA or PLINK (see Supplementary Note 1) even when running on large supercomputers.

DISSECT implements several highly computational demanding analyses. Some of the most relevant are as follows: computing genetic relationship matrices; performing PCA for studying population structure in large data sets; fitting univariate MLMs;...
fitting bivariate MLMs, which greatly increase power to detect pleiotropic loci, but require a computational time that is roughly eight times bigger than fitting univariate MLMs to data sets of the same size; regional MLM fitting for studying the accumulated variance explained by the alleles within genomic regions, each region having similar computational cost regardless of the number of SNPs fitted but requiring an independent fit; and standard regression models with very large number of fixed effects (for example, fitting the markers of a whole chromosome as fixed effects when extremely large sample sizes are available).

DISSECT also allows other computationally less demanding analyses such as the prediction of individual phenotypes from estimated marker effects (that is, polygenic scores) or standard genome-wide association study (GWAS) analyses. Furthermore, it also implements optimized routines similar to those found in GEMMA based on performing the eigen decomposition of the genetic relationship matrix for MLM analysis. These routines allow DISSECT to run analyses much faster when the user wishes to fit several MLM in the same population (see the Supplementary Note 1 for a more detailed description of the analyses).

Computational performance. We performed MLM and PCA analyses using simulated cohorts (Supplementary Methods) of different sample size (N; Fig. 2), to demonstrate the computational capabilities of DISSECT. We selected these two examples, because they are very computationally demanding analyses, requiring a running time of \( O(N^3) \). The analyses were run on the UK National Supercomputing Service (ARCHER), a supercomputer with 4,920 computer nodes containing 9,840 processors with 12 cores each (that is, a total of 118,080 cores available). DISSECT was able to fit, after eight iterations, an MLM to a sample of 470,000 individuals and 590,004 SNPs in less than 4 h.
using the aggregated power of 8,400 cores and a total of ~16 TB of memory (~2 GB of memory per core; Supplementary Fig. 3). The running time included estimation of the variances using REML,24,25, best linear predictions of the individual’s genetic values and best linear predictions of SNP effects18,26. If we disregard the computational overhead of communication between nodes, we can roughly estimate the computational time required by a computer with one core, to complete the analysis by multiplying the number of used cores with the computation time (core hours). In this situation, the MLM fit would need 3.6 years (Fig. 2a). Performing a PCA for 108,000 individuals and 590,004 SNPs required ~2 h using 1,920 cores. That is, around ~4,000 core hours, which would be equivalent to ~160 days of computation on a single core (Fig. 2b). All these results show both the high computational demands required for performing these analyses and the ability of DISSECT to perform them.

**Prediction results with huge sample sizes.** We tested the accuracy of phenotypic prediction from genotype data when large numbers of individuals are available. To this end, more than half a million SNP genotypes for half a million individuals were simulated based on linkage disequilibrium patterns and allele frequencies from the Hapmap CEU population. Then, we simulated several quantitative traits by using both different heritabilities ($h^2$) and numbers of QTNs. We first assumed a situation where only ~20% of the QTNs were in the genotyping array. In each case, we divided the cohort into two subsets: one for training the models and another for validating the predictions (Supplementary Methods). Predictions were based on the effects of all available SNPs estimated jointly from the MLM fit. As expected, prediction accuracy increased with the heritability of the trait and the size of the training data set (Fig. 3). The MLM efficiently captured the effects of large numbers of genotyped and ungenotyped QTNs. Simulated traits determined by 10,000 QTNs (Fig. 3) gave very similar results to traits determined by 1,000 QTNs (Supplementary Fig. 4). Importantly, high accuracies were only achieved when large numbers of individuals were used to train the prediction model. For instance, training the MLM with 470,000 individuals yielded correlations of 0.72, 0.57 and 0.30 for traits with 10,000 QTNs and heritabilities of 0.7, 0.5 and 0.2, respectively. That is, between 86% and 68% of the theoretical maximum, which is the square root of the heritability. We compared our results against predictions obtained from SNP effects computed using the BOLT-LMM software8, which is able to estimate variance components with large sample sizes on single compute nodes by performing approximations. The analyses with BOLT-LMM required up to ~14 days to analyse a sample with 470,000 individuals using 8 threads in a single compute node. Compared with DISSECT, there was a significant decrease in the prediction accuracy of BOLT-LMM as the sample size of the training set increased (Supplementary Fig. 5). BOLT-LMM was designed and developed in the context of GWAS testing, where each marker effect is estimated independently, which could explain the loss of prediction accuracy that we observed.

We investigated why—even when training the models with this extremely large sample sizes—the limit of prediction accuracy was still not close to the theoretical maximum. As the estimation of QTN effects appeared to be very accurate (Supplementary Fig. 6), we hypothesized that the loss in accuracy might be a consequence of the improper QTN tagging by markers in the array. Under this hypothesis we expected that an array that included all the QTNs would substantially improve prediction accuracy.

**Prediction accuracy when all QTNs are genotyped.** To test whether poor tagging of the QTNs explained the loss of accuracy, we assumed that all previously used tagging SNPs and all simulated QTNs were included in the genotyping array. Our results showed that the prediction accuracy for traits with 10,000 QTNs increased only slightly (Fig. 4). For traits with 1,000 QTNs the results were very similar to those of 10,000 QTNs (Supplementary Fig. 7). This suggests that under our genetic model one might not approach the theoretical limit of prediction accuracy even when training the models with 470,000 individuals and the genotyping array or resequencing included all QTNs27. We then hypothesized that if we were able to discriminate causal from non-causal variants, then prediction accuracy would improve. Hence, we repeated our experiments but now assuming that only the QTNs were included in the genotyping array. Prediction accuracy increased significantly (Supplementary Fig. 8),

---

**Figure 3 | Prediction accuracy of MLM as a function of sample size and heritability.** Correlation between true (P) and predicted phenotypes (A) as a function of cohort size for a trait determined by 10,000 QTNs. Black, blue and red curves represent heritabilities of 0.2, 0.5 and 0.7, respectively. Constant dashed lines indicate the theoretical maximum achievable for each heritability. Error bars are two times the s.d. over 6 replicas (470,000 individuals case has only 1 replica).

**Figure 4 | Prediction accuracy when all QTNs were genotyped.** Correlation between true (P) and predicted phenotypes (A) as a function of the cohort size when the trait is determined by 10,000 QTNs. Black, blue and red curves represent traits with heritabilities of 0.2, 0.5 and 0.7, respectively. Solid lines are the correlations obtained when all QTNs were genotyped. Dotted lines are the correlations obtained when only ~20% of QTNs were genotyped. Constant dashed lines indicate the maximum theoretical correlation for each heritability.
Methods

Simulations. We used the HAPGEN 2 software\(^\text{31}\) to simulate half a million individuals—based on linkage disequilibrium patterns and allele frequencies of 2,543,887 SNPs available in the Hapmap 2 (release 22) CEU population\(^\text{32}\)—from which we generated subsets of 20, 40, 60, 80, 120, 300 and 500 thousand individuals. From each subset of data, we used 90% of the individuals for training the models and the rest for validating the predictions, except for the subset including 500,000 individuals where we used 470,000 individuals for training and 30,000 for validation. We simulated traits that were determined by 1,000 and 10,000 randomly distributed QTNs, respectively. The QTNs were randomly distributed across the genome and their combined effects explained 20, 30 and 70% of the phenotypic variation. That is, we simulated heritabilities ($h^2$) of 0.2, 0.5 and 0.7. The QTNs effects were the same for all data subsets. Six replicates were performed for each trait heritability and genetic architecture, except for the subset including 500,000 individuals. Each replica assumed different QTNs with different effects drawn at random. The phenotypes were simulated using DISSECT, which assumes an additive genetic model for the selected QTNs (see Supplementary Note 1).

MLM and prediction. MLM analyses were performed using DISSECT. The software and its source code are freely available (http://www.dissect.ed.ac.uk). For our first set of analyses we excluded all SNPs not present on the Illumina Human OmniExpress BeadChip. That gave us a set of 590,004 SNPs which included only ~20% of the QTNs available within the simulated data set. Later, we investigated the effect of having the QTNs in the genotyping array and included the remaining ~80% of QTNs to the genotyping array.

The model fitted was:

$$y_i = \mu + \sum_{j=1}^{m} z_{ij} e_j + e_i,$$

where $\mu$ is the mean term and $e_i$ the residual. $z_{ij}$ is the standardized genotype of individual $i$ at marker $j$. The vector of random SNP effects $s$ is distributed as $N(0, \mathbf{I} \sigma_e^2)$. $\sum_{j=1}^{m} z_{ij} e_j$ is the total genetic effect for individual $i$. The phenotypic variance–covariance matrix is $\mathbf{V} = \mathbb{V} = \mathbf{I} \sigma_e^2 + \mathbf{I} \sigma^2$, SNP effects were estimated using the equation:\(^\text{26}\)

$$\mathbf{a} = \sigma_e^2 \mathbf{V}^{-1} (y - \mathbf{1}).$$

SNP effects were used as an input for DISSECT, to predict phenotypes on the validation cohort. DISSECT computes the prediction for individual $i$ as a sum of the product of the SNP effects and the number of reference alleles of the corresponding SNPs:

$$\hat{y}_i = \sum_{j=1}^{l} s_j \mathbf{a}_j,$$

where $s_j$ is the number of copies of the reference allele at SNP $j$ of individual $i$, $l$ is the number of SNPs used for the prediction and $\mathbf{a}_j$ is the effect of SNP $j$ estimated from the MLM analyses or obtained from BOLT-LMM software\(^\text{5}\), $\mathbf{a}_j$ and $\mathbf{a}_j$ are the mean and the s.d. of the reference allele in the training population.

Code availability. DISSECT source code is freely available under GPLv3 license at the url: http://www.dissect.ed.ac.uk.

References

1. Marx, V. Biology: the big challenges of big data. Nature 498, 255–260 (2013).
2. Matilainen, K., Mäntysaari, A. E., Lidauer, M. H., Strandén, I. & Thompson, R. Employing a Monte Carlo algorithm in Newton-type methods for restricted maximum likelihood estimation of genetic parameters. PLoS ONE 8, e60821 (2013).
3. Abraham, G. & Inouye, M. Fast principal component analysis of large-scale genome-wide data. PLoS ONE 9, e93766 (2014).
4. Aulchenko, Y. S., de Koning, D. J. & Haley, C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics 177, 577–585 (2007).
5. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44, 821–824 (2012).
6. Kang, H. M. et al. Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. 42, 348–354 (2010).
7. Zhang, Z. et al. Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42, 355–360 (2010).
8. Loh, P. R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. 47, 284–290 (2015).
9. Lippert, C. et al. FaST linear mixed models for genome-wide association studies. Nat. Methods 8, 833–835 (2011).
10. De los Campos, G., Gianola, D. & Allison, D. B. Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nat. Rev. Genet.* **11**, 880–886 (2010).

11. Lango Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).

12. Schrodi, S. J. et al. Genetic-based prediction of disease traits: prediction is very difficult, especially about the future. *Front. Genet.* **5**, 162 (2014).

13. Wray, N. R. et al. Pitfalls of predicting complex traits from SNPs. *Nat. Rev. Genet.* **14**, 507–515 (2013).

14. Meeuwissen, T. H. E., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).

15. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).

16. Meeuwissen, T. H. E. Accuracy of breeding values of ‘unrelated’ individuals predicted by dense SNP genotyping. *Genet. Sel. Evol.* **41**, 35 (2009).

17. Blackford, L. S. et al. ScalAPACK Users’ Guide (Society for Industrial and Applied Mathematics, 1997).

18. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

19. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 599–575 (2007).

20. Korte, A. et al. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* **44**, 1066–1071 (2012).

21. Cebamanos, L., Gray, A., Stewart, I. & Tenesa, A. Regional heritability advanced complex trait analysis for GPU and traditional parallel architectures. *Bioinformatics* **30**, 1177–1179 (2014).

22. Nagamine, Y. et al. Localising loci underlying complex trait variation using Regional Genomic Relationship Mapping. *PLoS ONE* **7**, e46501 (2012).

23. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–1528 (2007).

24. Gilmour, A. R., Thompson, R. & Cullis, B. R. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**, 1440–1450 (1995).

25. Lee, S. H. & van der Werf, J. H. J. An efficient variance component approach implementing an average information REML suitable for combined LD and linkage mapping with a general complex pedigree. *Genet. Sel. Evol.* **38**, 25–43 (2006).

26. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer, 1998).

27. Daetwyler, H. D., Villanueva, B. & Woolliams, J. A. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS ONE* **3**, e3395 (2008).

28. Collins, R. What makes UK Biobank special? *Lancet* **379**, 1173–1174 (2012).

29. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).

30. Chatterjee, N. et al. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45**, 400–405 (2013).

31. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304–2305 (2011).

32. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1302 (2005).