Selection of Genes and Single Nucleotide Polymorphisms for Fine Mapping Starting From a Broad Linkage Region

An Windelinckx,[1] Robert Vlietinck,[2] Jeroen Aerssens,[3] Gaston Beunen,[1] and Martine A. I. Thomis[1]

[1] *Research Center for Exercise and Health, Department of Biomedical Kinesiology, Faculty of Kinesiology and Rehabilitation Sciences, Katholieke Universiteit Leuven, Leuven, Belgium*

[2] *Clinical Genetics Section, Department of Human Genetics, Faculty of Medicine, Katholieke Universiteit Leuven, Leuven, Belgium*

[3] *Department of Translational Medical Research, Tibotec bvba, Generaal De Witte Laan, Mechelen, Belgium*

Fine mapping of linkage peaks is one of the great challenges facing researchers who try to identify genes and genetic variants responsible for the variation in a certain trait or complex disease. Once the trait is linked to a certain chromosomal region, most studies use a candidate gene approach followed by a selection of polymorphisms within these genes, either based on their possibility to be functional, or based on the linkage disequilibrium between adjacent markers. For both candidate gene selection and SNP selection, several approaches have been described, and different software tools are available. However, mastering all these information sources and choosing between the different approaches can be difficult and time-consuming. Therefore, this article lists several of these in silico procedures, and the authors describe an empirical two-step fine mapping approach, in which candidate genes are prioritized using a bioinformatics approach (ENDEAVOUR), and the top genes are chosen for further SNP selection with a linkage disequilibrium based method (Tagger). The authors present the different actions that were applied within this approach on two previously identified linkage regions for muscle strength. This resulted in the selection of 331 polymorphisms located in 112 different candidate genes out of an initial set of 23,300 SNPs.

During the past decades, genetic epidemiologists have tried to identify genes and genetic variants that underlie variation in diseases. Some diseases show a Mendelian inheritance, but most are thought to result from a complex interplay between multiple genetic variants, each with a small to modest effect, and/or environmental factors.

To decipher the effect of genetic variants on a trait, two complementary designs are often used: linkage and association analyses. Linkage analyses focus on the co-segregation of a trait and a genetic variant in successive generations of related individuals. Such analyses result in a rather broad genetic region and one of the great challenges facing researchers today is how to analyse ('fine map') these regions further in order to detect the genetic variant(s) responsible for the linkage signal. Genetic association analyses aim to identify the association of a particular allele (or haplotype) with a trait or disease status. In the past, this approach has mainly focussed on candidate genes and specific polymorphisms therein. More recently, thanks to the declining costs of genotyping, genome-wide association analyses have become feasible.

One often used strategy for fine mapping a region found by linkage analysis is to focus on a limited number of candidate genes because (a) only a limited number of genes is present under the linkage peak, or (b) some well established candidate genes are present (Bergholdt et al., 2005; Curran et al., 2006; Lou et al., 2007; Lowe et al., 2007; Palmer et al., 2006). This approach has many advantages: (1) Limiting the number of candidate genes also implies limiting the number of polymorphisms to be assessed; (2) by using established candidate genes, the chance of finding a real causative allele (as opposed to a false positive result) is augmented; (3) restricting the number of polymorphisms diminishes the number of statistical tests, avoids (more or less) stringent correction procedures for multiple testing, as well as limits the chance of a false positive result; and (4) complementary to statistical evidence for association between a gene variant and an examined trait, a real causality can be proven by functional (physiological or biological) evidence for an association. This can be achieved by monitoring expression (at mRNA or protein level; in sera or tissue biopsies), examining animal models (e.g., transgenic mice) or specific cell lines by functional tests. For established candidate genes extensive

Received 7 September, 2007; accepted 2 October, 2007.

Address for correspondence: Martine Thomis, Research Center for Exercise and Health, Department of Biomedical Kinesiology, Faculty of Kinesiology and Rehabilitation Sciences, Katholieke Universiteit Leuven, Leuven, Belgium. E-mail: Martine.Thomis@faber.kuleuven.be

research has often already been done and these additional data may thus already be available.

However, disadvantages are also present. First, not all traits have already been studied to the extent that a list of candidate genes or functional studies are present (e.g., preterm birth; Pennell et al., 2007). Especially for complex traits that are influenced by multiple genes, selecting suitable candidate genes can be difficult. Second, by using this approach one will fail to detect new functions for known genes if their known function is not directly related to the studied trait. Third, this method *a priori* excludes unknown or uncharacterised genes. Moreover, since linkage regions are often very broad and may contain several hundreds of genes, screening all of these genes will be laborious and time consuming. To deal with these disadvantages, bioinformatics approaches have been developed to prioritize genes within a genomic region and aid in the selection of the most promising genes for further analyses (Adie et al., 2005, 2006; Aerts et al., 2006; De Bie et al., 2007; Lopez-Bigas & Ouzounis, 2004; Moses et al., 2006; Perez-Iratxeta et al., 2005; Rossi et al., 2006; Tiffin et al., 2006).

Once a list of candidate genes has been selected, the next step is to decide upon which variants within each of these genes should be genotyped. As most genes harbour tens to hundreds of single nucleotide polymorphisms (SNPs) and other mutations or repeats (depending e.g. on the size of the gene or the chromosomal region), a reduction of this number is often desired. Strategies identifying or prioritising SNPs can roughly be divided into two groups: those based on the linkage disequilibrium (LD) between markers (i.e. their tendency to segregate together), and those based on the probability of the SNP to be biologically relevant. Combinations of both LD-based and functional approaches are also possible.

Here we describe a two-step approach for the fine mapping of previously found linkage regions in which candidate genes are first prioritized using a bioinformatics approach, and then subsequently a number of SNPs are selected for a chosen list of top-ranked genes using a LD-based method. Additionally, we list several alternative in silico procedures and tools that provide similar or related information.

Materials and Methods

General Remarks

The proposed fine mapping strategy was born from the necessity to select SNPs for fine mapping purposes in our own studies. Since multiple alternatives (e.g., different software programs, databases, etc.) are present for every step in the process, we developed some criteria to guide us in selecting one possibility over another.

Simple and practical methods. As familiarisation always takes some time, we chose to develop a method that is easy to use, with a minimal number of different software programs or databases. Moreover, we selected programs that are publicly available and thereby preferred software programs that can be downloaded (to avoid dependence on network status). If these were not available, we selected web-based programs that provide direct output.

Easily adaptable and flexible methods. The choice of software was also guided by the need to be able to adapt the approach whenever necessary. Adding new information should be easy and alternative strategies should be testable.

Feasible within the available resources of the lab. Even though genotyping costs have decreased substantially over the last years and continue to do so, available financial resources for genotyping often constrain the number of SNPs to be genotyped. Our strategy was therefore also designed to optimise the balance between the number of SNPs selected and the information to be gained from genotyping these SNPs. Notwithstanding this, the strategy can be used for larger studies as well.

Candidate Gene Selection

Several software programs to select candidate genes can be found in literature and are summarized in Table 1. Most of these are largely based on keyword similarity to known disease genes or phenotypes. For example, *Geneseeker* (van Driel et al., 2003; 2005) combines keyword search results from positional, expression, and phenotypic databases from both human and mouse. Tiffin et al. (2005) developed a method based on expression profiles within tissues related to a disease and combined these with clinical and molecular data based on eVOC (a controlled vocabulary for unifying gene expression data) anatomy ontology (Kelso et al., 2003). *Suspects* (Adie et al., 2006) compares functional annotations (Gene Ontology [GO] terminology), InterPro (a database of protein families, domains and functional sites; Apweiler et al., 2001; Mulder et al., 2005; 2007) domains, gene expression data between candidate genes, and genes that are already known to influence the disease. Similarly, *TOM* (Transcriptomics of OMIM [Online Mendelian Inheritance in Man]; Rossi et al., 2006) combines data on gene mapping, expression profiling and GO terminology. It can either look for expression similarities between candidate genes and genes known to influence a trait, or between two *bona fide* linkage regions identified for the trait of interest. *G2D* (Candidate Genes to Inherited Diseases; Perez-Iratxeta et al., 2002, 2007; Perez-Iratxeta et al., 2005) combines the extraction of relations between phenotypes and gene functions in sequence, disease, and literature databases with sequence similarity searches. *Genesniffer* (Moses et al., 2006) interrogates NCBI's Gene, OMIM, and PubMed together with Jackson's Mouse Genome Informatics database, by means of a list of user-specified disease-specific keywords. Additionally homologues of each gene are

Table 1

Sources of Input Data for Each Candidate Gene Selection/Prioritization Method (Adapted from Tiffin et al., 2006)

	Used data sources													URL
	NCBI gene	OMIM (or related databases)	PubMed abstracts	Homologues	Orthologus mouse genes	Expression data	Sequence data	GO annotation	Protein data	Pathway data	Transcriptional motifs/cis regulatory modules	Binding data	eVOC	
GeneSeeker		X	X		X	X			X					http://www.cmbi.ru.nl/GeneSeeker/
eVOC annotation			X										X	not available
SUSPECTS						X	X	X	X					http://www.genetics.med.ed.ac.uk/suspects
PROSPECTR							X							http://www.genetics.med.ed.ac.uk/prospectr
TOM		X						X						http://www-micrel.deis.unibo.it/~tom
G2D		X	X	X			X	X						http://www.ogic.ca/projects/g2d_2
Genesniffer	X	X	X	X										http://www.genesniffer.org
DGP									X					http://cgg.ebi.ac.uk/services/dgp/index.html
ENDEAVOUR			X			X	X	X	X	X	X	X		http://www.esat.kuleuven.be/endeavour
Kernel-based data fusion			X			X	X	X	X	X	X	X		not yet available

identified by BLAST, and scored for content of their Gene, OMIM, PubMed, and Jackson entries.

Other methods are based on sequence similarity between genes and/or proteins. The Disease Gene Prediction (DGP) program (Lopez-Bigas & Ouzounis, 2004) compares conservation, phylogenetic extent, protein length, and paralogy between the candidate genes and known disease genes. Similarly, *Prospectr* (Adie et al., 2005) differentiates between genes likely and unlikely to be involved in disease by means of sequence-based features, such as gene length, protein length, and the percent identity of homologues in other species. For a review and a comparison of most of these methods, and an example of their use, see Tiffin et al. (2006), who have applied these methods to the selection of candidate genes for Type 2 diabetes and obesity.

Recently, an effort was made to develop a program that incorporates most of the above mentioned sources of information in one candidate gene prioritization tool. *ENDEAVOUR* (Aerts et al., 2006) software prioritizes candidate genes based on their similarity to genes known to influence the trait of interest. Multiple heterogeneous data sources are used by the program (literature, functional annotation (GO), micro array expression, EST expression, protein domains, protein-protein interactions, pathway membership, *cis*-regulatory modules, transcriptional motif, and sequence similarity) and integrated into a global ranking using order statistics. Very recently (De Bie et al., 2007), the same research group developed a new method based on the same principle as the ENDEAVOUR software, namely prioritising genes based on their similarity to known genes. They use a new kernel-based method for data fusing and show that is has a better performance and computational efficiency than ENDEAVOUR.

Out of these programs we favour the ENDEAVOUR program (Aerts et al., 2006) for a number of reasons. First, the program makes use of several possible inputs for training and candidate genes. One can search not only on gene identifiers (HUGO name, ENSEMBL-id, ...), but also on GO terms, KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway numbers, OMIM classification, or chromosomal position. This means that no additional programs or databases have to be scanned to get this information. Second, all gene lists and models can be saved for later use. This makes it possible to adjust analyses rather easily. Third, ENDEAVOUR mines multiple heterogeneous data sources. However, one can choose to use all data sources or only a subset. This makes it possible to deselect data sources that do not provide additional information, thereby reducing analysis time. Moreover, the results for each data source are provided, allowing the assessment of the specific contributions of each data source to the overall ranking. Finally, the program is freely accessible and regularly updated. A drawback of the ENDEAVOUR program

is that the results are highly dependent on the initial selection of the genes in the training set. This implies some *a priori* knowledge about the examined trait. However, the multiple input possibilities (GO, OMIM, ...) should make it feasible to design a set of possible training genes for most traits. Moreover, to reduce the dependency on the selected training genes, we suggest the use of several different training sets and to combine the results in a global ranking statistic.

Alternative candidate gene selection strategies are obviously also possible. Oliveira et al. (2005) identified a linkage peak for Parkinson disease, and suggested fine mapping using a 'genomic convergence' method, whereby genes that underlie a linkage peak, and are differentially expressed between cases and controls in a relevant tissue, are tested for association. They selected this method because of the limited understanding of the biological systems involved in Parkinson disease. A similar approach was suggested by Rodd et al. (2007). Other researchers did not use available programs, but developed their own bioinformatics approach (Dash et al., 2006; Wilson et al., 2006; Yang et al., 2005)

SNP selection

For the selection of SNPs within candidate genes of interest, several approaches have been suggested, mostly based on SNP tagging and/or predicted functionality. A large body of literature about this topic is present and it is out of the scope of this article to review all methods in detail. Excellent reviews describing and comparing SNP selection procedures have been published and we refer the reader to these articles (Bhatti et al., 2006; Chi et al., 2006; Halldörsson et al., 2004; Ke et al., 2005; Stram, 2005). An updated list of tagging and functional SNP selection software can be found in supplementary Table 2.

SNP tagging programs need genotypic information in order to select tagSNPs. As the purpose of fine mapping is to determine which SNPs should be genotyped, this genotypic information needs to be imported from some other genotyped population. For this reason, the HapMap (The International HapMap Consortium, 2003, 2005) project was started. Several studies have already shown that tagSNPs selected based on the HapMap data are transferable to other populations with similar ancestral backgrounds (Conrad et al., 2006; de Bakker et al., 2006; Ke et al., 2004; Montpetit et al., 2006; Mueller et al., 2005; Smith et al., 2006). A useful feature of the HapMap project is that genotypes for different populations can be downloaded and filtered according to several criteria, such as SNP rs-number, minor allele frequency in the selected population, validation status, and so on. By doing this, the number of polymorphisms entering the tagging process can be reduced, and the analyses will become less computer intensive.

We prefer *Tagger* (de Bakker et al., 2005) to identify tagSNPs based on LD between adjacent polymorphisms, based on the fact that *Tagger* is implemented in *Haploview* (Barrett et al., 2004), and genotype information of the CEPH families of the HapMap can therefore be easily used as input data. *Haploview* can also be downloaded.

If the number of tagSNPs still exceeds the number of polymorphisms to be genotyped, a further selection can be made based on the informativeness of the tagSNP (i.e., the number of other SNPs it tags; de Bakker et al., 2005), validation status, or the initial ranking of the candidate gene the SNPs reside in. For example, it could be decided that all available tagSNPs will be genotyped in the top 5 genes, the 3 to 5 most informative tagSNPs will be genotyped for the top 20 genes, and only 1 or 2 tagSNPs will be genotyped for the remaining genes.

Other approaches described in the present literature propose genotyping of all the coding SNPs (Bergholdt et al., 2005) or a selection of (validated) intragenic SNPs (Curran et al., 2006; Moses et al., 2006), sequencing the gene and comparing discordant subjects (Dash et al., 2006; Lowe et al., 2007), using evenly spaced SNPs throughout the gene (Lou et al., 2007; Palmer et al., 2006), or combinations of the previous methods (Hinks et al., 2006; Nicolae et al., 2005; Wang et al., 2007; Wilson et al., 2006; Yang et al., 2005). This field is clearly evolving rapidly, but it is beyond the scope of this manuscript to go into detail on each of these methods.

Most, but not all, of the previously described approaches, identified significant associations between one or more polymorphisms and the observed trait. However, as far as we are aware, no articles have compared the efficacy of the different SNP selection procedures on a single dataset. The increasing availability of genome wide, high-density SNP panels might make these analyses, albeit retrospectively, possible.

Application

In a previous study from our laboratory (Huygens et al., 2004; Huygens et al., 2005), a linkage analysis was performed to investigate the possible linkage between genes encoding key proteins from the myostatin pathway and isometric and concentric knee strength. Three regions were found to be significantly/suggestively linked with a quantitative trait locus for knee muscle strength: 12q12-14 (LOD score 3.4), 12q22-23 (LOD score 2.7), and 13q14.2 (LOD score 2.7). Besides the initial candidate genes from the myostatin pathway, several other interesting muscle-associated genes are located within these regions. As an illustration of the methodology that we applied to select candidate genes and SNPs therein, we present here the results from our analyses in the linkage regions observed on chromosome 12 — a similar two-step procedure was applied to the chromosome 13 region. A graphical representation of the methodology used is shown in Figure 1.

First, in an attempt to reduce the number of candidate genes, we identified the 1-LOD confidence interval of the linkage regions on chromosome 12.

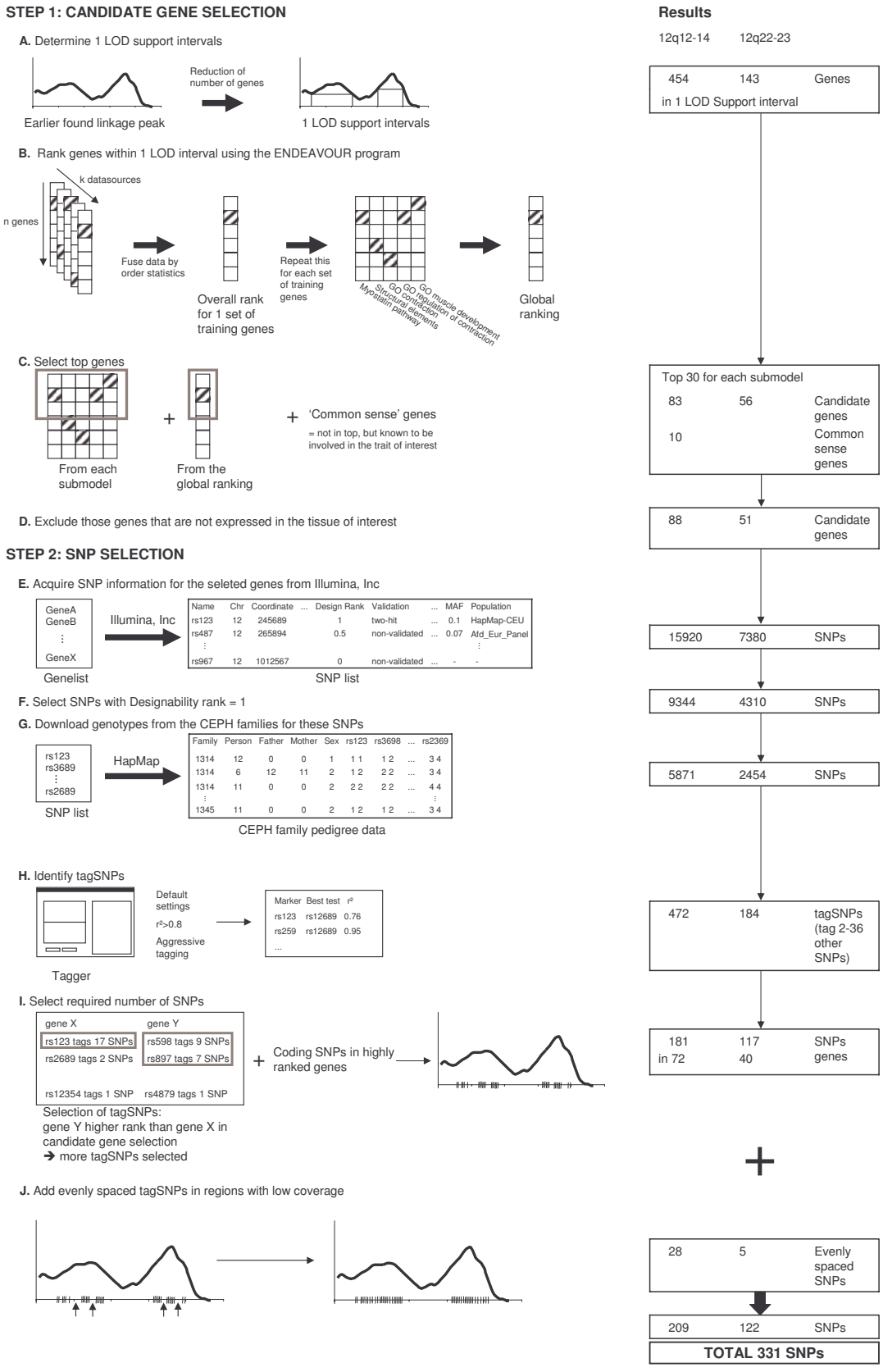

Figure 1

Graphical overview of the two step fine mapping approach, together with results for the application of the approach on two previously identified linkage regions on chromosome 12 for muscle strength.

The genes within these intervals were then prioritised using ENDEAVOUR software (version 1.37.02.01). Five different training sets were used and subsequently combined into a global ranking: (1) candidate genes from the original myostatin pathway (Huygens et al., 2004; Huygens et al., 2005), (2) structural elements of muscle (actin and myosin related genes, together with troponin, titin, and nebulin), (3) GO term 'contraction', (4) GO term 'muscle development', and (5) GO term 'regulation (negative and positive) of contraction'. The top 30 genes of the global ranking, and of each sub-model, were further investigated regarding expression levels and functionality, using available online databases (e.g., Genecards, Entrez Gene, Ensembl). Genes that were not expressed in muscle were excluded from further analysis. Additionally, some genes known to have a role in muscle strength, but not prioritised by ENDEAVOUR, were added to the analyses ('common sense' genes). A description of the top 10 candidate genes and of the 'common sense' genes, together with their ranking, can be found in supplementary Tables 3 and 4 for the 12q12-14 and the 12q22-23 region, respectively. The resulting genes were further processed in the SNP selection process.

Since the Illumina GoldenGate Custom Panels genotyping method was to be used in our analyses, a list of the candidate genes was sent to Illumina, in San Diego, California, in order to get a list of possible SNPs within these genes, including codes for the feasibility of assay development, validation status, and minor allele frequency for a number of selected populations. From this list, all SNPs with a designability rank of 1 (high success rate of assay development), and a minor allele frequency of at least 0.05, were selected, and CEPH (Utah residents with ancestry from northern and western Europe) family genotype data for these SNPs were downloaded from the HapMap website (release #20, January 2006, based on NCBI build 35; The International HapMap Consortium, 2003, 2005). These data were then used to determine tagSNPs using *Tagger* (de Bakker et al., 2005), implemented in the *Haploview* software (Barrett et al., 2004). In addition to the default settings, aggressive tagging was selected. From these tagSNPs, a subset was then chosen based on validation status and initial ranking of the candidate gene. Finally, nonsynonymous SNPs were added for the top genes. For the regions with a low coverage, additional tagSNPs were identified and evenly spaced tagSNPs were selected in order to span those regions.

Results

Two linkage regions on chromosome 12 were analysed using the two-step fine mapping approach: one on 12q12-14 encompassing ca. 40cM or 53Mb and one on 12q22-23 of ca. 15cM or 10Mb. Within these regions respectively, 454 and 143 genes were analysed.

ENDEAVOUR analysis and exclusion of nonexpressed genes resulted in a list of, respectively, 78 and 51 genes. An additional 10 'common sense' genes, all located in or around the 12q12-14 region, were added.

Within these genes, Illumina identified 15,920 and 7380 SNP for the 12q12-14 and the 12q22-23 regions, respectively. Of these SNPs respectively 9344 and 4310 had a designability rank of 1, and genotype data was present in the HapMap for 5871 and 2454 of these SNPs. SNP selection resulted in a total of 331 SNPs, of which 298 are located in 112 genes (range 1-9 SNPs per gene), and 33 are spaced in the genomic regions in between the genes. A graphical representation of the localisation of the SNPs is shown in Figure 2. A similar two-step procedure was applied to the chromosome 13 region, in which 52 tagSNPs were selected (not included in Figure 2).

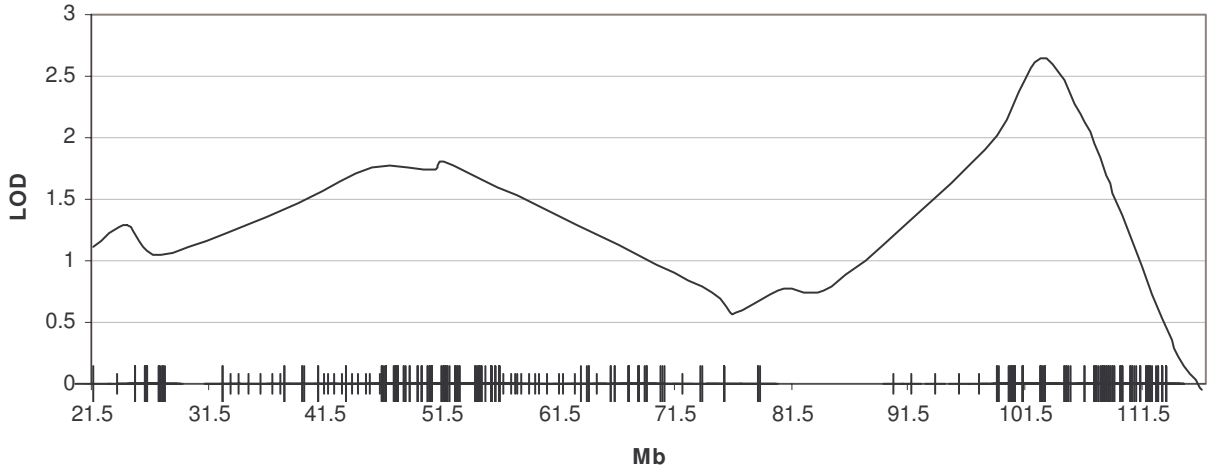

Figure 2

Location of the 331 selected SNPs in relation to the original microsatellite based linkage peak. High bars denote polymorphisms selected within candidate genes. Small bars are evenly spaced tagSNPs selected in-between genes.

Discussion

Selecting polymorphisms for follow-up fine mapping of a linkage peak is a great challenge for researchers within the field of genetic epidemiology. Several different approaches exist, and different software tools are available for most methods. In an attempt to streamline our fine mapping analysis efforts, we developed a flexible two-step fine mapping approach, consisting of candidate gene selection in a genomic region of interest, followed by a SNP selection within these genes. Additionally, we listed several alternatives in silico tools useful in these analyses. Most of these methods are not new, but as far as we know, a recent overview is not present in current literature.

Application of our fine mapping approach to two previously identified linkage regions for muscle strength, shows that when SNPs are selected within prioritised candidate genes, the polymorphisms are evenly distributed under the linkage curve, with a higher density where LOD score peaks are present. Therefore, we believe our chances of finding variants that influence our trait have been markedly improved by our fine mapping approach, but actual linkage and association analyses will be needed to show whether this assumption is true.

Some advantages of our two-step approach are worth mentioning. First, it is very flexible. One can easily alter the number of genes selected after prioritization or the number of SNPs after tagging. Also, the broad input and output possibilities of the ENDEAVOUR software (Aerts et al., 2006) make flexible analyses possible. The researcher can choose to use all possible data sources in the analyses, or select specific data sources to speed up the process. Additionally if all data sources are chosen, besides the global ranking, the results for each data source separately can also be obtained. This makes it possible to assign weights to each data source, depending on the importance or the amount of information that is presumably available in the data source. The new kernel-based approach developed by the same research group (De Bie et al., 2007) already incorporates several weighting schemes in the program. Second, the number of necessary programs and databases has been kept as small as possible. With the use of ENDEAVOUR, the HapMap website, *Haploview* and some basic data management software (e.g., Microsoft Access or Excel, depending on the magnitude of the data), it should be possible to perform the basic analyses presented in this article. Additional databases or online resources can be accessed for further information, but are not absolutely required. Third, the spread of the selected polymorphisms over the linkage region and the use of family data make it possible to use both linkage and association analyses for follow-up analyses. Additionally, a combined linkage and association approach can be applied to identify the specific polymorphisms responsible for the linkage signal.

Yet, our fine mapping approach also has some limitations. First, as was already pointed out, most candidate gene selection/prioritization tools select genes based on their similarity to genes known to influence the trait of interest (training genes; Adie et al., 2005; 2006; Aerts et al., 2006; De Bie et al., 2007; Perez-Iratxeta et al., 2005; Perez-Iratxeta et al., 2007; Rossi et al., 2006). The selection of an appropriate set of training genes is, however, highly dependent on the knowledge of the field being studied, and on the amount of useful information in publicly available databases. If no candidate genes for a specific trait have been identified, results from related phenotypes can be used, or the physiology or biology of the studied system can be examined.

Second, as an intermediate step, we selected a subset of markers with a high probability of assay success based on the data provided by Illumina (designability rank = 1). Tagging was then performed on the polymorphisms for which genotypes were available in the HapMap for the CEPH subjects, reducing the number of SNPs to approximately 35% of the original number of SNPs identified by Illumina. An alternative approach would have been to select tagSNPs first, and then assess if these could be genotyped with the Illumina platform. Even better would be to take both the patterns of LD and the likely genotyping success into account in the same analysis. At present, several of the tagSNP selection programs, such as the recent version of Haploview (Barrett et al., 2004), Multipop-tagselect (Howie et al., 2006), Tagger (de Bakker et al., 2005), TAGster (Xu et al. 2007), mPopTag (Xu et al. 2007) and SNPselector (Xu et al., 2005b), provide the option to include design scores and use the information during the SNP prioritization process.

To evaluate the influence of selecting the tagSNPs on the polymorphisms with designability rank 1, rather than on all the SNPs, we calculated the percentage of the alleles that were captured by our list of tagSNPs. For example, for the 12q22-23 region, the selected SNPs capture 48% of the alleles of the other SNPs at $r^2 > .8$, for the designability rank 1 SNPs as well as for all the SNPs. We would, however, like to point out that the coverage per gene will vary extensively, as the original design of the study intended an alternative coverage dependent on the initial ranking of the gene. For example, in the gene that ranked fifth on the global ranking, 8 tagSNPs were selected, which capture 62% of all alleles with $r^2 > .8$. In contrast, in the gene that ranked 30th, only one tagSNP was selected. Therefore only 13% of the alleles were captured at $r^2 > .8$.

Moreover, we believe that the availability of genotyped and polymorphic SNPs in the HapMap is a more limiting factor for inclusion in the final tagSNP set than is the design score. For example, for the 22q22-23 region, 7380 SNPs were identified

by Illumina in our candidate genes. Of these SNPs, only 3047 (41%) had genotypes available in the #20 release of the HapMap, and of these 1099 were monomorphic. Thus, only 1948 (26.4%) of the originally identified SNPs were included in the tagging process. Of these SNPs, 350 had a designability rank of 0 or 0.5. So an additional 5% of SNPs were lost when we decided to include only the SNPs with a designability rank of 1.

This use of the CEPH family genotype data from the HapMap project also results in a third shortcoming in our method. Since these represent only an approximation of the European population, minor allele frequencies (MAF) can differ when compared to other populations. The use of the stringent cut off of MAF > .05 during the tagging process may result in the exclusion of SNPs that have higher frequencies in our population. Vice versa, SNPs that have a MAF >.05 for the CEPH population may have a lower MAF in our population. This, again, may result in genotyping of SNPs that have low/no informativity in the sample that will be genotyped. Fourth, by using this MAF cut off in combination with a LD-based tagging method, only more common variants are captured. Additionally, the LD structure within the entire human genome is highly variable. Depending on the extent of the LD, a different number of tagSNPs will thus have to be selected to cover the same physical region. Finally, most of the SNPs were selected based on their ability to serve as a proxy for the surrounding SNPs, and not based on their possible functionality. As suggested by Bhatti et al. (2006), an approach combining both LD-based and functionality-based selection of polymorphisms for further analyses could be more suitable.

However, we were aware of most of these potential disadvantages from the start of our study, and regard this SNP selection only as an intermediate stage to identify genes associated with muscular strength. Additional genotyping within such associated genes will undoubtedly be necessary to identify the causal SNPs within these genes.

We sought to design an approach for both gene and SNP prioritization that keeps the balance between an in depth and comprehensive search, and practically applicable in silico methods for routine analysis. We acknowledge that other methods exist, and appreciate that individual researchers may choose other tools, or make adaptations to the presented methods. Our major objective here has been to address a number of important issues encountered during the selection of candidate genes and SNPs for genotyping, which we believe may also be useful for other researchers in their quest to identify the genes underlying a certain trait or disease.

Acknowledgements

An Windelinckx is funded by the Research Fund of the K.U.Leuven (OT/04/44). The fine mapping phase of the Leuven Genes for Muscular Strength Study is funded by OT/04/44 and the Research Foundation — Flanders (G.0496.05).

Online resources

http://www.esat.kuleuven.be/endeavour

http://www.hapmap.org

http://www.genecards.org

http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene

http://www.ensembl.org

* other online resources are indicated in tables 1 and 2

References

Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J., & Pickard, B. S. (2005). Speeding disease gene discovery by sequence based candidate prioritization. *BMC.Bioinformatics, 6,* 55.

Adie, E. A., Adams, R. R., Evans, K. L., Porteous, D. J., & Pickard, B. S. (2006). SUSPECTS: Enabling fast and effective prioritization of positional candidates. *Bioinformatics, 22,* 773–774.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L. C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., & Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nature biotechnology, 24,* 537–544.

Ao, S. I., Yip, K., Ng, M., Cheung, D., Fong, P. Y., Melhado, I., & Sham, P. C. (2005). CLUSTAG: Hierarchical clustering and graph methods for selecting tag SNPs. *Bioinformatics, 21,* 1735–1736.

Apweiler, R., et al. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research, 29,* 37–40.

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics*, *21*, 263–265.

Bergholdt, R., Nerup, J., & Pociot, F. (2005). Fine mapping of a region on chromosome 21q21.11-q22.3 showing linkage to type 1 diabetes. *Journal of Medical Genetics, 42,* 17–25.

Bhatti, P., Church, D. M., Rutter, J. L., Struewing, J. P., & Sigurdson, A. J. (2006). Candidate Single Nucleotide Polymorphism Selection using Publicly Available Tools: A Guide for Epidemiologists. *American Journal of Epidemiology, 164,* 794–804.

Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., & Nickerson, D. A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics, 74,* 106–120.

Chi, P. B., Duggal, P., Kao, W. H., Mathias, R. A., Grant, A. V., Stockton, M. L., Garcia, J. G.,

Ingersoll, R. G., Scott, A. F., Beaty, T. H., Barnes, K. C., & Fallin, M. D. (2006). Comparison of SNP tagging methods using empirical data: Association study of 713 SNPs on chromosome 12q14.3-12q24.21 for asthma and total serum IgE in an African Caribbean population. *Genetic Epidemiology, 30,* 609–619.

Conde, L., Vaquerizas, J. M., Dopazo, H., Arbiza, L., Reumers, J., Rousseau, F., Schymkowitz, J., & Dopazo, J. (2006). PupaSuite: Finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic Acids Research, 34,* W621-W625.

Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., & Pritchard, J. K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature Genetics, 38,* 1251–1260.

Curran, S., Powell, J., Neale, B. M., Dworzynski, K., Li, T., Murphy, D., & Bolton, P. F. (2006). An association analysis of candidate genes on chromosome 15 q11-13 and autism spectrum disorder. *Molecular Psychiatry, 11,* 709–713.

Dash, D. P., Silvestri, G., & Hughes, A. E. (2006). Fine mapping of the keratoconus with cataract locus on chromosome 15q and candidate gene analysis. *Molecular vision., 12,* 499–505.

Davidovich, O., Kimmel, G., & Shamir, R. (2007). GEVALT: An integrated software tool for genotype analysis. *BMC.Bioinformatics, 8,* 36.

de Bakker, P. I., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., & Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nature Genetics, 37,* 1217–1223.

de Bakker, P. I. W., et al. (2006). Transferability of tag SNPs in genetic association studies in multiple populations. *Nature Genetics, 38,* 1298–1303.

De Bie, T., Tranchevent, L. C., van Oeffelen, L. M., & Moreau, Y. (2007). Kernel-based data fusion for gene prioritization. *Bioinformatics, 23,* i125-i132.

De La Vega, F. M., Isaac, H. I., & Scafe, C. R. (2006). A tool for selecting SNPs for association studies based on observed linkage disequilibrium patterns. *Pacific Symposium on Biocomputing,* 487–498.

Ding, K., Zhang, J., Zhou, K., Shen, Y., & Zhang, X. (2005). htSNPer1.0: Software for haplotype block partition and htSNPs selection. *BMC.Bioinformatics, 6,* 38.

Eyheramendy, S., Marchini, J., McVean, G., Myers, S., & Donnelly, P. (2007). A model-based approach to capture genetic variation for future association studies. *Genome Research, 17,* 88–95.

Grover, D., Woodfield, A. S., Verma, R., Zandi, P. P., Levinson, D. F., & Potash, J. B. (2007). QuickSNP: An automated web server for selection of tagSNPs. *Nucleic Acids Research, 35,* W115-W120.

Halldörsson, B. V., Istrail, S., & De La Vega, F. M. (2004). Optimal Selection of SNP Markers for Disease Association Studies. *Human Heredity, 58,* 190–202.

Halperin, E., Kimmel, G., & Shamir, R. (2005). Tag SNP selection in genotype data for maximizing SNP prediction accuracy. *Bioinformatics, 21,* i195-i203.

Hampe, J., Schreiber, S., & Krawczak, M. (2003). Entropy-based SNP selection for genetic association studies. *Human Genetics, 114,* 36–43.

He, J. & Zelikovsky, A. (2006). MLR-tagging: Informative SNP selection for unphased genotypes based on multiple linear regression. *Bioinformatics, 22,* 2558–2561.

Hemminger, B. M., Saelim, B., & Sullivan, P. F. (2006). TAMAL: An integrated approach to choosing SNPs for genetic studies of human complex traits. *Bioinformatics, 22,* 626–7.

Hinks, A., Barton, A., John, S., Shephard, N., & Worthington, J. (2006). Fine mapping of genes within the IDDM8 region in rheumatoid arthritis. *Arthritis Research & Therapy, 8,* R145.

Howie, B., Carlson, C., Rieder, M., & Nickerson, D. (2006). Efficient selection of tagging single-nucleotide polymorphisms in multiple populations. *Human Genetics, 120,* 58–68.

Huygens, W., Thomis, M. A., Peeters, M. W., Aerssens, J., Janssen, R., Vlietinck, R., & Beunen, G. (2004). Linkage of myostatin pathway genes with knee strength in humans. *Physiological Genomics, 17,* 264–270.

Huygens, W., Thomis, M. A. I., Peeters, M. W., Aerssens, J., Vlietinck, R., & Beunen, G. P. (2005). Quantitative trait loci for human muscle strength: Linkage analysis of myostatin pathway genes. *Physiological Genomics, 22,* 390–397.

Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di, G. G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., Twells, R. C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S. C., Clayton, D. G., & Todd, J. A. (2001). Haplotype tagging for the identification of common disease genes. *Nature Genetics, 29,* 233–237.

Ke, X. & Cardon, L. R. (2003). Efficient selective screening of haplotype tag SNPs. *Bioinformatics, 19,* 287–288.

Ke, X., Miretti, M. M., Broxholme, J., Hunt, S., Beck, S., Bentley, D. R., Deloukas, P., & Cardon, L. R. (2005). A comparison of tagging methods and their tagging space. *Human Molecular Genetics, 14,* 2757–2767.

Ke, X., Durrant, C., Morris, A. P., Hunt, S., Bentley, D. R., Deloukas, P., & Cardon, L. R. (2004). Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Human Molecular Genetics, 13,* 2557–2565.

Kelso, J., Visagie, J., Theiler, G., Christoffels, A., Bardien, S., Smedley, D., Otgaar, D., Greyling, G., Jongeneel, C. V., McCarthy, M. I., Hide, T., & Hide, W. (2003). eVOC: A Controlled Vocabulary for Unifying Gene Expression Data. *Genome Research, 13,* 1222–1230.

Liu, Z. & Lin, S. (2005). Multilocus LD measure and tagging SNP selection with generalized mutual information. *Genetic Epidemiology, 29,* 353–364.

Liu, Z., Lin, S., & Tan, M. (2006). Genome-wide tagging SNPs with entropy-based Monte Carlo method. *Journal of computational biology, 13,* 1606–1614.

Lopez-Bigas, N. & Ouzounis, C. A. (2004). Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Research, 32,* 3108–3114.

Lou, X. Y., Ma, J. Z., Sun, D., Payne, T. J., & Li, M. D. (2007). Fine mapping of a linkage region on chromosome 17p13 reveals that GABARAP and DLG4 are associated with vulnerability to nicotine dependence in European-Americans. *Human Molecular Genetics, 16,* 142–153.

Lowe, C. E., Cooper, J. D., Brusko, T., Walker, N. M., Smyth, D. J., Bailey, R., Bourget, K., Plagnol, V., Field, S., Atkinson, M., Clayton, D. G., Wicker, L. S., & Todd, J. A. (2007). Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nature Genetics*. *39*, 1074–1082

Magi, R., Kaplinski, L., & Remm, M. (2006). The whole genome tagSNP selection and transferability among HapMap populations. *Pacific Symposium on Biocomputing,* 535–543.

Maniatis, N., Collins, A., Xu, C. F., McCarthy, L. C., Hewett, D. R., Tapper, W., Ennis, S., Ke, X., & Morton, N. E. (2002). The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis. *Proceedings of the National Academy of Sciences of the United States of America, 99,* 2228–2233.

McCauley, J. L., Kenealy, S. J., Margulies, E. H., Schnetz-Boutaud, N., Gregory, S. G., Hauser, S. L., Oksenberg, J. R., Pericak-Vance, M. A., Haines, J. L., & Mortlock, D. P. (2007). SNPs in Multi-species Conserved Sequences (MCS) as useful markers in association studies: A practical approach. *BMC.Genomics, 8,* 266.

Montpetit, A., Nelis, M., Laflamme, P., Magi, R., Ke, X., Remm, M., Cardon, L., Hudson, T. J., & Metspalu, A. (2006). An Evaluation of the Performance of Tag SNPs Derived from HapMap in a Caucasian Population. *PLoS Genetics, 2,* e27.

Moses, E. K., Fitzpatrick, E., Freed, K. A., Dyer, T. D., Forrest, S., Elliott, K., Johnson, M. P., Blangero, J., & Brennecke, S. P. (2006). Objective prioritization of positional candidate genes at a quantitative trait locus for pre-eclampsia on 2q22. *Molecular Human Reproduction, 12,* 505–512.

Mueller, J. C., Lohmussaar, E., Magi, R., Remm, M., Bettecken, T., Lichtner, P., Biskup, S., Illig, T., Pfeufer, A., Luedemann, J., Schreiber, S., Pramstaller, P., Pichler, I., Romeo, G., Gaddi, A., Testa, A., Wichmann, H. E., Metspalu, A., & Meitinger, T. (2005). Linkage disequilibrium patterns and tagSNP transferability among European populations. *American Journal of Human Genetics, 76,* 387–398.

Mulder, N. J., et al. (2007). New developments in the InterPro database. *Nucleic Acids Research, 35,* D224-D228.

Mulder, N. J., et al. (2005). InterPro, progress and status in 2005. *Nucleic Acids Research, 33,* D201-D205.

Nicolae, D., Cox, N. J., Lester, L. A., Schneider, D., Tan, Z., Billstrand, C., Kuldanek, S., Donfack, J., Kogut, P., Patel, N. M., Goodenbour, J., Howard, T., Wolf, R., Koppelman, G. H., White, S. R., Parry, R., Postma, D. S., Meyers, D., Bleecker, E. R., Hunt, J. S., Solway, J., & Ober, C. (2005). Fine mapping and positional candidate studies identify HLA-G as an asthma susceptibility gene on chromosome 6p21. *American Journal of Human Genetics, 76,* 349–357.

Nicolas, P., Sun, F., & Li, L. M. (2006). A model-based approach to selection of tag SNPs. *BMC.Bioinformatics, 7,* 303.

Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *American Journal of Human Genetics, 74,* 765–769.

Oliveira, S. A., Li, Y. J., Noureddine, M. A., Zuchner, S., Qin, X., Pericak-Vance, M. A., & Vance, J. M. (2005). Identification of Risk and Age-at-Onset Genes on Chromosome 1p in Parkinson Disease. *The American Journal of Human Genetics, 77,* 252–264.

Palmer, N. D., Langefeld, C. D., Campbell, J. K., Williams, A. H., Saad, M., Norris, J. M., Haffner, S. M., Rotter, J. I., Wagenknecht, L. E., Bergman, R. N., Rich, S. S., & Bowden, D. W. (2006). Genetic Mapping of Disposition Index and Acute Insulin Response Loci on Chromosome 11q: The Insulin Resistance Atherosclerosis Study (IRAS) Family Study. *Diabetes, 55,* 911–918.

Pennell, C. E., Jacobsson, B., Williams, S. M., Buus, R. M., Muglia, L. J., Dolan, S. M., Morken, N. H., Ozcelik, H., Lye, S. J., & Relton, C. (2007). Genetic epidemiologic studies of preterm birth: Guidelines for research. *American journal of obstetrics and gynecology, 196,* 107–118.

Perez-Iratxeta, C., Bork, P., & Andrade, M. A. (2002). Association of genes to genetically inherited diseases using data mining. *Nature Genetics, 31,* 316–319.

Perez-Iratxeta, C., Bork, P., & Andrade-Navarro, M. A. (2007). Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Research, 35,* W212-W216.

Perez-Iratxeta, C., Wjst, M., Bork, P., & Andrade, M. A. (2005). G2D: A tool for mining genes associated with disease. *BMC.Genet, 6,* 45.

Phuong, T. M., Lin, Z., & Altman, R. B. (2006). Choosing SNPs using feature selection. *Journal of Bioinformatics and Computational Biology, 4,* 241–257.

Qin, Z. S., Gopalakrishnan, S., & Abecasis, G. R. (2006). An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria. *Bioinformatics, 22,* 220–225.

Rinaldo, A., Bacanu, S. A., Devlin, B., Sonpar, V., Wasserman, L., & Roeder, K. (2005). Characterization of multilocus linkage disequilibrium. *Genetic Epidemiology, 28,* 193–206.

Riva, A. & Kohane, I. S. (2001). A web-based tool to retrieve human genome polymorphisms from public databases. *Proceedings of the AMIA Symposium,* 558–562.

Riva, A. & Kohane, I. S. (2002). SNPper: Retrieval and analysis of human SNPs. *Bioinformatics, 18,* 1681–1685.

Riva, A. & Kohane, I. (2004). A SNP-centric database for the investigation of the human genome. *BMC Bioinformatics, 5,* 33.

Rodd, Z. A., Bertsch, B. A., Strother, W. N., Le-Niculescu, H., Balaraman, Y., Hayden, E., Jerome, R. E., Lumeng, L., Nurnberger, J. I., Jr., Edenberg, H. J., McBride, W. J., & Niculescu, A. B. (2007). Candidate genes, pathways and mechanisms for alcoholism: An expanded convergent functional genomics approach. *Pharmacogenomics Journal, 7,* 222–256.

Rossi, S., Masotti, D., Nardini, C., Bonora, E., Romeo, G., Macii, E., Benini, L., & Volinia, S. (2006). TOM: A web-based integrated approach for identification of candidate disease genes. *Nucleic Acids Research, 34,* W285-W292.

Sham, P. C., Ao, S. I., Kwan, J. S., Kao, P., Cheung, F., Fong, P. Y., & Ng, M. K. (2007). Combining functional and linkage disequilibrium information in the selection of tag SNPs. *Bioinformatics, 23,* 129–131.

Smith, E. M., Wang, X., Littrell, J., Eckert, J., Cole, R., Kissebah, A. H., & Olivier, M. (2006). Comparison of linkage disequilibrium patterns between the HapMap CEPH samples and a family-based cohort of Northern European descent. *Genomics, 88,* 407–414.

Stram, D. O. (2005). Software for tag single nucleotide polymorphism selection. *Human Genomics, 2,* 144–151.

Stram, D. O., Haiman, C. A., Hirschhorn, J. N., Altshuler, D., Kolonel, L. N., Henderson, B. E., & Pike, M. C. (2003). Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Human Heredity, 55,* 27–36.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature, 437,* 1299–1320.

The International HapMap Consortium (2003). The International HapMap Project. *Nature, 426,* 789–796.

Tiffin, N., Kelso, J. F., Powell, A. R., Pan, H., Bajic, V. B., & Hide, W. A. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Research, 33,* 1544–1552.

Tiffin, N., Adie, E., Turner, F., Brunner, H. G., van Driel, M. A., Oti, M., Lopez-Bigas, N., Ouzounis, C., Perez-Iratxeta, C., Andrade-Navarro, M. A., Adeyemo, A., Patti, M. E., Semple, C. A. M., & Hide, W. (2006). Computational disease gene identification: A concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Research, 34,* 3067–3081.

van Driel, M. A., Cuelenaere, K., Kemmeren, P. P., Leunissen, J. A., & Brunner, H. G. (2003). A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *European Journal of Human Genetics, 11,* 57–63.

van Driel, M. A., Cuelenaere, K., Kemmeren, P. P. C. W., Leunissen, J. A. M., Brunner, H. G., & Vriend, G. (2005). GeneSeeker: Extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Research, 33,* W758-W761.

Wang, L., Hauser, E. R., Shah, S. H., Pericak-Vance, M. A., Haynes, C., Crosslin, D., Harris, M., Nelson, S., Hale, A. B., Granger, C. B., Haines, J. L., Jones, C. J., Crossman, D., Seo, D., Gregory, S. G., Kraus, W. E., Goldschmidt-Clermont, P. J., & Vance, J. M. (2007). Peakwide mapping on chromosome 3q13 identifies the kalirin gene as a novel candidate gene for coronary artery disease. *American Journal of Human Genetics, 80,* 650–663.

Wang, L., Liu, S., Niu, T., & Xu, X. (2005). SNPHunter: A bioinformatic software for single nucleotide polymorphism data acquisition and management. *BMC.Bioinformatics, 6,* 60.

Weale, M. E., Depondt, C., Macdonald, S. J., Smith, A., Lai, P. S., Shorvon, S. D., Wood, N. W., & Goldstein, D. B. (2003). Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: Implications for linkage-disequilibrium gene mapping. *American Journal of Human Genetics, 73,* 551–565.

Wilson, S. G., Adam, G., Langdown, M., Reneland, R., Braun, A., Andrew, T., Surdulescu, G. L., Norberg, M., Dudbridge, F., Reed, P. W., Sambrook, P. N.,

Kleyn, P. W., & Spector, T. D. (2006). Linkage and potential association of obesity-related phenotypes with two genes on chromosome 12q24 in a female dizygous twin cohort. *European Journal of Human Genetics, 14,* 340–348.

Xu, H., Gregory, S. G., Hauser, E. R., Stenger, J. E., Pericak-Vance, M. A., Vance, J. M., Zuchner, S., & Hauser, M. A. (2005a). SNPselector: A web tool for selecting SNPs for genetic association studies. *Bioinformatics, 21,* 4181–4186.

Xu, H., Gregory, S. G., Hauser, E. R., Stenger, J. E., Pericak-Vance, M. A., Vance, J. M., Zuchner, S., & Hauser, M. A. (2005b). SNPselector: A web tool for selecting SNPs for genetic association studies. *Bioinformatics, 21,* 4181–4186.

Xu, Z., Kaplan, N. L., & Taylor, J. A. (2007). Tag SNP selection for candidate gene association studies using HapMap and gene resequencing data. *European Journal of Human Genetics*. advance online publication June 13, 2007 doi:10.1038/sj.ejhg.5201875

Yang, Q., Lai, C. Q., Parnell, L., Cupples, L. A., Adiconis, X., Zhu, Y., Wilson, P. W. F., Housman, D. E., Shearman, A. M., D'Agostino, R. B., & Ordovas, J. M. (2005). Genome-wide linkage analyses and candidate gene fine mapping for HDL3 cholesterol: The Framingham Study. *Journal of Lipid Research, 46,* 1416–1425.

Yuan, H. Y., Chiou, J. J., Tseng, W. H., Liu, C. H., Liu, C. K., Lin, Y. J., Wang, H. H., Yao, A., Chen, Y. T., & Hsu, C. N. (2006). FASTSNP: An always up-to-date and extendable service for SNP function analysis and prioritization. *Nucleic Acids Research, 34,* W635-W641.

Zhang, J., Rowe, W. L., Struewing, J. P., & Buetow, K. H. (2002a). HapScope: A software system for automated and visual analysis of functionally annotated haplotypes. *Nucleic Acids Research, 30,* 5213–5221.

Zhang, K., Deng, M., Chen, T., Waterman, M. S., & Sun, F. (2002b). A dynamic programming algorithm for haplotype block partitioning. *Proceedings of the National Academy of Sciences of the United States of America, 99,* 7335–7339.

Zhang, K., Qin, Z. S., Liu, J. S., Chen, T., Waterman, M. S., & Sun, F. (2004). Haplotype Block Partitioning and Tag SNP Selection Using Genotype Data and Their Applications to Association Studies. *Genome Research, 14,* 908–916.

Zhang, K. & Jin, L. (2003). HaploBlockFinder: Haplotype block analyses. *Bioinformatics, 19,* 1300–1301.

APPENDIX

Supplementary Table 1

List of Available SNP Tagging or Functional SNP Selection Software

Name	Web url	Reference
TagSNP selection		
FESTA	http://www.sph.umich.edu/csg/qin/FESTA/	Qin et al., 2006
Genecap	on request	Eyheramendy et al., 2007
Genedigger	http://www.genedigger.de/	Hampe et al., 2003
Genotype2LDblock	http://cgi.uc.edu/cgi-bin/kzhang/genotype2LDBlock.cgi	N/A
Hapblock	http://www-hto.usc.edu/msms/HapBlock/	Zhang et al., 2004
Haploblockfinder	http://cgi.uc.edu/cgi-bin/kzhang/haploBlockFinder.cgi	Zhang et al., 2002b; Zhang & Jin, 2003
Haploview	http://www.broad.mit.edu/mpg/haploview/	Barrett et al., 2004
Hapscope	http://lpg.nci.nih.gov/lpg_small/protocols/HapScope/	Zhang et al., 2002a
HCLUST	http://wpicr.wpic.pitt.edu/WPICCompGen/hclust.htm	Rinaldo et al., 2005
htSNP and htSNP2	http://www-gene.cimr.cam.ac.uk/clayton/software/stata/	Johnson et al., 2001
htSNP Finder	http://htsnp.stanford.edu/	Phuong et al., 2006
HTSNPER	http://www.chgb.org.cn/htSNPer/htSNPer.html	Ding et al., 2005
Ldmap	http://cedar.genetics.soton.ac.uk/public_html/helpld.html	Maniatis et al., 2002
Ldselect	http://droog.gs.washington.edu/ldSelect.html	Carlson et al., 2004
MLR-tagging	http://alla.cs.gsu.edu/~software/tagging/tagging.html	He & Zelikovsky, 2006
MultiPop-TagSelect	http://droog.gs.washington.edu/multiPopTagSelect.html	Howie et al., 2006
REAPER	http://bioinfo.ebc.ee/download/	Magi et al., 2006
SNPbrowser	http://marketing.appliedbiosystems.com/mk/get/SNP_LANDING	De La Vega et al., 2006
SNPspD	http://gump.qimr.edu.au/general/daleN/SNPSpD/	Nyholt, 2004
SNPtagger	http://www.well.ox.ac.uk/~xiayi/haplotype/	Ke & Cardon, 2003
STAMPA (implemented in GEVALT)	http://acgt.cs.tau.ac.il/gevalt/	Davidovich et al., 2007; Halperin et al., 2005
Tagger	http://www.broad.mit.edu/mpg/tagger/	de Bakker et al., 2005
tagIT	http://www.genome.duke.edu/resources/computational/software/	Weale et al., 2003
Tag'n'Tell	http://snp.cgb.ki.se/tagntell/	N/A
TagSNP	http://www-rcf.usc.edu/~stram/tagSNPs.html	Stram et al., 2003
TagSNP	http://www-rcf.usc.edu/~lilei/tagsnp.html	Nicolas et al., 2006
TagSNPfinder	http://www.stat.osu.edu/~statgen/SOFTWARE/tagSNPfinder/	Liu et al., 2006; Liu & Lin, 2005
TAGster	http://www.niehs.nih.gov/research/resources/software/tagster/index.cfm	Xu et al. 2007
mPopTag	http://www.niehs.nih.gov/research/resources/software/mpoptag/index.cfm	Xu et al., 2007
QuickSNP	http://bioinformoodics.jhmi.edu/quickSNP.pl	Grover et al., 2007
CLUSTAG	http://hkumath.hku.hk/web/link/CLUSTAG/CLUSTAG.html	Ao et al., 2005
Functionality-based SNP selection		
SNPselector	http://snpselector.duhs.duke.edu/hqsnp36.html	Xu et al., 2005a
TAMAL	http://neoref.ils.unc.edu/tamal/	Hemminger et al., 2006
PupaSuite	http://pupasuite.bioinfo.cipf.es/	Conde et al., 2006
SNPper	http://snpper.chip.org/	Riva & Kohane, 2001; 2002; 2004
SNPHunter	http://www.hsph.harvard.edu/ppg/software.htm	Wang et al., 2005
FastSNP	http://fastsnp.ibms.sinica.edu.tw/pages/input_CandidateGeneSearch.jsp	Yuan et al., 2006
BinCONs	http://zoo.nhgri.nih.gov/binCons/index.cgi	McCauley et al., 2007
WCLUSTAG	http://bioinfo.hku.hk/wclustag/	Sham et al., 2007

Note: Supplementary Table 2 was adapted from Bhatti et al. (2006) and Halldörsson et al. (2004). Additional programs were added based on a PubMed search using the search terms: 'SNP selection', 'SNP tagging', 'tagSNPs' and on related articles from other references. Additionally the Alphabetic List of Genetic Analysis Software (http://linkage.rockefeller.edu/soft/) was consulted.

Supplementary Table 3

Top 10 Candidate Genes for Each Submodel and the Global Ranking From Endeavour Prioritization for 12q12-14

Name or gene description	Contraction	Myostatin pathway	Structural elements	Muscle development	Regulation of muscle contraction	Global ranking
Congenital myopathy	**2**	**4**	**2**	**1**	**2**	**1**
Carbohydrate metabolism	**1**	**5**	**3**	**3**	**3**	**2**
Structural component of muscle	**7**	*	**1**	**4**	**4**	**3**
Structural component of muscle	10	**2**	**4**	**14**	**1**	**4**
Structural component of muscle	11	*	**5**	**10**	45	**5**
Contributes to the development and growth of multiple mammalian tissues including skeletal muscle	**5**	56	**6**	20	**5**	**6**
DEAD (Asp-Glu-Ala-Asp) box polypeptide	19	48	**7**	17	**7**	**7**
Mediates gene silencing	33	60	**10**	**7**	11	**8**
Anion channel	**8**	**8**	61	45	**9**	**9**
Transcription factor	36	**6**	30	15	49	**10**
Proliferation inhibitor	13	74	19	**8**	34	11
Involved in muscle maturation	37	44	11	**2**	77	17
Signaling protein	32	16	38	83	**10**	18
ATP-ase activity	26	129	**8**	16	**8**	19
Homologue involved in mouse posterior limb patterning/ cell cycle control	44	**3**	43	86	23	20
Cytoplasmic mRNA decay	62	96	**9**	23	24	23
Protein kinase	25	252	24	13	**6**	40
Integrin family	43	159	15	**5**	98	41
Homeobox protein	67	**9**	58	103	94	42
Calcium Channel	**4**	66	158	22	95	45
Involved in muscle dystrophies	**3**	251	48	32	15	46
Transcription factor activity	77	41	113	**6**	114	48
Homeobox protein	165	**1**	89	124	179	86
Homeobox protein	116	**7**	114	186	151	88
Involved in the growth and differentiation of neural cells	143	148	193	**9**	86	90
Suppressor of apoptosis	**6**	250	201	89	79	98
Sodium channel	**9**	181	83	207	251	133
Transcription factor activity	359	**10**	274	326	279	261
Additional common sense genes						
VDR: Associated with variability in muscle strength	194	92	226	274	170	188
IGFBP6: Regulation of cell growth, IGF1 pathway	250	110	336	132	149	190
CYP27B1: Defects in CYP27B1 are a cause of vitamin D-dependent rickets type, characterized by muscle weakness	218	165	293	261	240	244
ACVR1B: Transforming growth factor beta receptor activity, myostatin signalling	52	170	69	255	116	108
ACVRL1: Activin pathway	85	39	37	265	134	560
YAF2: Muscle specific YY1 cofactor	209	220	202	253	395	269

Note: *not scored because included in training set.

Supplementary Table 3

Top 10 Candidate Genes for Each Submodel and the Global Ranking from Endeavour Prioritization for 12q22-23

Name or gene description	Contraction	Myostatin pathway	Structural elements	Muscle development	Regulation of muscle contraction	Global ranking
Structural component of muscle, striated muscle contraction	*	**2**	**2**	**1**	**2**	**1**
Structural component of muscle	*	**5**	**1**	**3**	**1**	**2**
Calcium pump	**1**	**1**	**3**	13	**3**	**3**
Regulator of somatic growth and cellular proliferation	**5**	**6**	**6**	**2**	**6**	**4**
Heat shock protein	**6**	**9**	**5**	**7**	**4**	**5**
Potassium channel	**3**	15	**4**	**9**	**5**	**6**
Limb pattern development during embryogenesis, transcription factor activity	**9**	**3**	**10**	**6**	**9**	**7**
Role in cellular proliferation	14	**10**	**7**	11	**10**	**8**
Open reading frame	15	**7**	17	**8**	15	**9**
Glycolysis/ gluconeogenesis	**8**	17	16	10	16	**10**
Starch and sucrose metabolism	27	13	**8**	26	**7**	11
Ribosomal protein	23	35	11	**4**	12	12
Actin polymerisation	**4**	44	18	17	25	15
Developmental regulation, limb pattern	**7**	22	19	49	17	16
Protein kinase	47	36	**9**	16	11	17
Structural protein	28	84	12	**5**	**8**	20
Ca ion binding	2	100	22	19	21	22
May directly link growth factor receptors and other signalling proteins	**10**	68	47	27	61	32
Neurogenesis/ transcription factor activity/ cell differentiation	59	**4**	103	30	98	51
Transcription regulation	86	**8**	132	123	131	107
Additional common sense genes						
PPP1CC: Regulation of glycogen metabolism, muscle contractility and protein synthesis.	**					
PPP1R12A: Contraction of smooth muscle	**					
CSRP2: Regulation of cell differentiation, muscle development	**					

Note: *not scored because included in training set

**no ENDEAVOUR ranking because these genes are located slightly outside the 1-LOD support interval.