BMC Research Notes

BioMed Central
The Open Access Publisher

LDGIdb: a database of gene interactions inferred from long-range strong linkage disequilibrium between pairs of SNPs

BMC Research Notes 2012, **5**:212 doi:10.1186/1756-0500-5-212

Ming-Chih Wang (carlw@gate.sinica.edu.tw)
Feng-Chi Chen (fcchen@nhri.org.tw)
Yen-Zho Chen (a14.a13@msa.hinet.net)
Yao-Ting Huang (ythuang@cs.ccu.edu.tw)
Trees-Juen Chuang (trees@gate.sinica.edu.tw)

ISSN 1756-0500

Article type Data Note

Submission date 28 October 2011

Acceptance date 2 May 2012

Publication date 2 May 2012

Article URL http://www.biomedcentral.com/1756-0500/5/212

LDGIdb: a database of gene interactions inferred from long-range strong linkage disequilibrium between pairs of SNPs

Ming-Chih Wang[1]
Email: carlw@gate.sinica.edu.tw

Feng-Chi Chen[2,3,4*]
[*] Corresponding author
Email: fcchen@nhri.org.tw

Yen-Zho Chen[1]
Email: a14.a13@msa.hinet.net

Yao-Ting Huang[5]
Email: ythuang@cs.ccu.edu.tw

Trees-Juen Chuang[1*]
[*] Corresponding author
Email: trees@gate.sinica.edu.tw

[1] Genomics Research Center, Academia Sinica, Taipei 11529, Taiwan

[2] Division of Biostatistics and Bioinformatics, Institute of Population Health Sciences, National Health Research Institutes, Miaoli County 350, Taiwan

[3] Department of Life Science, National Chiao-Tung University, Hsinchu 300, Taiwan

[4] Department of Dentistry, China Medical University, Taichung 404, Taiwan

[5] Department of Computer Science and Information Engineering, National Chung Cheng University, Chia-yi County 600, Taiwan

Abstract

Background

Complex human diseases may be associated with many gene interactions. Gene interactions take several different forms and it is difficult to identify all of the interactions that are potentially associated with human diseases. One approach that may fill this knowledge gap is to infer previously unknown gene interactions via identification of non-physical linkages between different mutations (or single nucleotide polymorphisms, SNPs) to avoid hitchhiking effect or lack of recombination. Strong non-physical SNP linkages are considered to be an indication of biological (gene) interactions. These interactions can be physical protein interactions, regulatory interactions, functional compensation/antagonization or many other forms of interactions. Previous studies have shown that mutations in different genes can be

linked to the same disorders. Therefore, non-physical SNP linkages, coupled with knowledge of SNP-disease associations may shed more light on the role of gene interactions in human disorders. A user-friendly web resource that integrates information about non-physical SNP linkages, gene annotations, SNP information, and SNP-disease associations may thus be a good reference for biomedical research.

Findings

Here we extracted the SNPs located within the promoter or exonic regions of protein-coding genes from the HapMap database to construct a database named the Linkage-Disequilibrium-based Gene Interaction database (LDGIdb). The database stores 646,203 potential human gene interactions, which are potential interactions inferred from SNP pairs that are subject to long-range strong linkage disequilibrium (LD), or non-physical linkages. To minimize the possibility of hitchhiking, SNP pairs inferred to be non-physically linked were required to be located in different chromosomes or in different LD blocks of the same chromosomes. According to the genomic locations of the involved SNPs (i.e., promoter, untranslated region (UTR) and coding region (CDS)), the SNP linkages inferred were categorized into promoter-promoter, promoter-UTR, promoter-CDS, CDS-CDS, CDS-UTR and UTR-UTR linkages. For the CDS-related linkages, the coding SNPs were further classified into nonsynonymous and synonymous variations, which represent potential gene interactions at the protein and RNA level, respectively. The LDGIdb also incorporates human disease-association databases such as Genome-Wide Association Studies (GWAS) and Online Mendelian Inheritance in Man (OMIM), so that the user can search for potential disease-associated SNP linkages. The inferred SNP linkages are also classified in the context of population stratification to provide a resource for investigating potential population-specific gene interactions.

Conclusion

The LDGIdb is a user-friendly resource that integrates non-physical SNP linkages and SNP-disease associations for studies of gene interactions in human diseases. With the help of the LDGIdb, it is plausible to infer population-specific SNP linkages for more focused studies, an avenue that is potentially important for pharmacogenetics. Moreover, by referring to disease-association information such as the GWAS data, the LDGIdb may help identify previously uncharacterized disease-associated gene interactions and potentially lead to new discoveries in studies of human diseases.

Background

Gene interactions are usually inferred from biological interactions such as protein-protein interactions (PPIs) [1-3], co-expression of genes [4,5], co-localization of proteins [6,7], co-evolution of proteins [8,9], and shared gene-phenotype associations [10]. Gene interactions that are implicated in human disorders are of particular interest [11]. Recently, it has been proposed that the associations between mutations and human disorders can be evaluated at the systems level [11-13]. This concept is based on observations that mutations in different genes can be linked to the same disorders, and that multiple mutations in the same genes can be associated with different diseases [11]. In other words, a human disorder may be the outcome of a molecular system where mutations in different genes are interconnected via a variety of gene interactions. Single nucleotide polymorphisms (SNPs) are frequently associated with human phenotypes, and SNPs in different genes that are strongly correlated

with each other may be important for gene interactions. Therefore, exploring the linkages between SNPs may offer new insights into the biological interactions in the human molecular system. A database that stores information about non-physical SNP linkages and possible SNP-disease associations may be helpful for exploring the role of gene interactions in human disorders.

Here we infer potential gene interactions on the basis of long-range linkage disequilibrium (LRLD) between SNPs. We term these potential interactions "linkage disequilibrium-based gene interactions" (LDGIs), where two genes are considered to be connected if the SNPs located in these two genes are subject to strong linkage disequilibrium (LD; usually measured by r^2 or D' [14]). Theoretically, LD should be observed between SNPs that are physically close to each other owing to the hitchhiking effect or lack of recombination [15]. In this study, however, we consider only the SNP pairs (designated as LRLD-SNP pairs) that are subject to strong LD ($r^2 \geq 0.8$) but are located in different LD blocks (or different chromosomes) to minimize the possibilities of accidentally linked SNPs or physical linkage, and thus increase the probability that the associations between the LRLD-linked SNPs/genes are functionally meaningful. To facilitate research based on these inferred SNP linkages (and potential gene interactions), we constructed a user-friendly database, the LDGIdb, to store the information. The LDGIdb also contains information about disease-associated SNPs/genes, such as the associations identified in genome-wide association studies (GWAS) [16] and those recorded in Online Mendelian Inheritance in Man (OMIM) database [17]. Users can thus search for LDGIs that involve disease-associated SNPs/genes, and identify potentially uncharacterized disease-associated gene interactions for further studies.

Findings

Construction of LDGIs

The data analysis workflow is shown in Figure 1. We first extracted human haplotypes from the HapMap Phase II and III data [18], which were generated using the PHASE software [19]. Only the SNPs that are located within the promoter or exonic regions of protein-coding genes (with reference to the Ensembl annotations [20]) were considered. Note that the promoter regions encompass 2 kb sequences upstream of the transcriptional start sites, and exonic regions include coding regions (CDSs) and untranslated regions (UTRs). In view of population stratification, we clustered the individuals examined in the HapMap Phase II and III projects into subpopulations using the PLINK package (version 1.07) [21] (Table 1). Here we consider only the subpopulations that contain at least 20 individuals. For each subpopulation, we calculated LD scores (i.e., r^2 and D' [14]) for all combinations of SNP pairs. Two SNPs were considered to be a long-range LD-linked SNP pair (designated as an "LRLD-SNP pair") if they satisfied both of the following criteria: (1) to avoid the inclusion of accidentally linked SNPs, an LRLD-SNP pair had to be subject to a strong LD ($r^2 \geq 0.8$); (2) to minimize the probability of hitchhiking or lack of recombination, the two SNPs had to be located in different chromosomes or be separated by at least one recombination hotspot retrieved from the International HapMap Project. The latter criterion may considerably decrease the probability that the identified LRLD-SNP pairs belong to the same "LD blocks" (or "haplotype blocks", which represent regions where recombination events occur rarely, and consequently LD is maintained) even if they are located in the same chromosomes. Accordingly, we identified 801,340 LRLD-SNP pairs, which contained 94,876 SNPs (Table 1). Genes connected by these LRLD-SNP pairs were considered human LD-based gene

interactions (LDGIs). The LDGIdb is composed of a collective total of about 646,203 gene linkages, which contain 21,240 genes (Table 1). Since population stratification was also considered, the LDGIdb also provides potential population-specific gene interactions, which may be useful for investigations of population-specific traits/diseases.

Figure 1 Process of identification of LRLD-SNP pairs and LDGIs

Table 1 Identified LRLD-SNP pairs and LDGIs (with $r^2 \geq 0.8$)

Population	# Individuals	#LRLD-SNP pairs	#Affected SNPs	# LDGIs	# Affected genes
PLINK ($P < 0.01$)					
Phase II					
CEU	30	66343	44756	23425	14444
Phase III					
CEU cluster 1	27	34940	28817	18569	11644
CEU cluster 2	40	30353	28082	15333	11308
CHD cluster 1	22	44513	29109	24675	11934
CHD cluster 2	31	30981	27563	15425	11161
JPT + CHB cluster 1	28	34672	28024	18212	11401
JPT + CHB cluster 2	22	48626	29360	29751	12014
LWK	23	42305	23398	35808	10545
MKK cluster 1	28	28924	22740	21185	10056
MKK cluster 2	24	54795	25086	47884	11203
MKK cluster 3	21	98150	27043	89007	11952
MKK cluster 4	22	63718	25598	53623	11465
YRI cluster 1	29	19116	20874	13011	9246
YRI cluster 2	28	22127	21138	16279	9390
PLINK ($P < 0.001$)					
Phase II					
CEU	48	61699	43945	21112	14181
JPT + CHB	20	117251	44754	61435	14792
YRI cluster 1	22	86494	38786	65698	13775
YRI cluster 2	21	98239	39389	75825	13950
Phase III					
ASW	25	18880	21631	11922	9432
CEU cluster 1	62	29924	27834	14997	11208
CEU cluster 2	41	30586	27967	15302	11269
CHD cluster 1	33	31456	27684	15283	11224
CHD cluster 2	21	31931	27598	15821	11215
CHI cluster 1	23	48723	30147	29355	12304
CHI cluster 2	21	56926	30705	37185	12543
JPT+CHB cluster 1	28	34509	28021	19004	11404
JPT+CHB cluster 2	30	33894	28111	17153	11424
JPT+CHB cluster 3	27	36917	28193	19958	11495
JPT+CHB cluster 4	23	44475	29077	25505	11928
JPT+CHB cluster 5	61	32011	27861	15491	11229

LWK cluster 1	21	61580	24684	54472	11194
LWK cluster 2	33	15850	19800	10436	8729
MEX	25	36330	29249	22032	11998
MKK cluster 1	41	17272	20819	10688	8997
MKK cluster 2	26	30057	22480	24739	10124
MKK cluster 3	25	52057	25081	44570	11104
MKK cluster 4	27	36459	23314	28450	10368
MKK cluster 5	24	41993	24359	33284	10779
TSI cluster 1	32	31021	28490	16191	11517
TSI cluster 2	30	32289	28501	16722	11535
YRI cluster 1	31	18261	20636	12029	9185
YRI cluster 2	37	15825	19577	9889	8602
YRI cluster 3	22	45628	23996	39632	10846
Sum		801340	94876	646203	21240

Calculation of r^2 and D' values

Let P_A and P_B be the major allele frequencies at SNP_1 and SNP_2, respectively. Define P_a and P_b as the minor allele frequencies at SNP_1 and SNP_2, respectively. Let P_{AB} be the haplotype frequency of observing both A and B alleles at these two loci. Define $D = P_{AB} - P_A P_B$. The LD scores, r^2 and D' [14], between SNP_1 and SNP_2 can be computed by

$$r^2 = \frac{(P_{AB} - P_A P_B)^2}{P_A(1-P_A)P_B(1-P_B)} = \frac{D^2}{P_A(1-P_A)P_B(1-P_B)} \text{ and } D' = \begin{cases} \dfrac{D}{\min(P_A P_B, P_a P_b)}, & \text{if } D < 0; \\ \dfrac{D}{\min(P_A P_b, P_a P_B)}, & \text{if } D > 0. \end{cases},$$

Data retrieval

HapMap Phase II (release 22) and III (release 2) haplotype data and the corresponding recombination hotspot information were retrieved from the International HapMap Project [22]. The human protein-coding genes were downloaded from the Ensembl genome browser (release 53). The human PPI data (designated as "collected PPIs" in the LDGIdb) were collected from seven experiment-supported PPI databases: HPRD [23], DIP [24], MINT [25], IntAct [26], REACTOME [27], BioGRID [28], and MIPS [29]. The extracted PPI collection included a total of 76,955 interactions. The CRG (Centre for Genomic Regulation) human interactomes (designated as "CRG PPIs" in the LDGIdb) were downloaded from Bossi and Lehners' study [30], which comprised 80,922 interactions. Human gene co-expression data were downloaded from the TMM database [4], which contained 203,043 high-confidence co-expression links that were observed in at least three microarray data sets. The biological interactions inferred from the above databases (i.e., collected PPIs, CRG PPIs, and co-expression links) were integrated into the LDGIdb for comparison with LDGIs. If an LDGI was not found in any of the databases, it was considered to be a potentially uncharacterized gene interaction. The GWAS [16] data were downloaded on August 23rd, 2011 [31]. For LRLD-linked genes, more detailed information was provided including protein domain descriptions (according to Interpro [32], SMART, and PFAM), KEGG pathways [33], and

disease association information (OMIM, HIV interaction, and the Genetic Association Database [34]), which were all downloaded from the DAVID knowledgebase [35].

Web interface

Users can search for LRLD-SNP pairs and LDGIs (which are linked by LRLD-SNP pairs) by setting three adjustable parameters: HapMap data source (Phase II or III), *P* value for PLINK population clustering ($P < 0.01$ or $P < 0.001$), and r^2 value for linkage disequilibrium (≥ 0.8, ≥ 0.9, or 1) (Figure 2A). Note that we only considered population clusters containing at least 20 individuals (Table 1). Also note that LDLR-SNP pairs with $r^2 = 1$ are subject to a "complete" LD. The LDGIdb supports four types of queries. Users can search for LRLD-SNP pairs/LDGIs by specifying the types of genomic location of LRLD-linked SNPs, SNP ID, gene accession number(s), or genomic coordinates (Figure 2B). GWAS-related LRLD-SNP pairs are also provided (Figure 2C). As shown in Figure 2D, the LRLD-SNP pairs/LDGIs are categorized, according to the types of genomic location of the linked SNPs, into promoter-promoter, promoter-UTR, promoter-CDS, CDS-CDS, CDS-UTR and UTR-UTR interactions. The CDS-related LDGIs are further categorized according to whether the LD-linked SNPs are nonsynonymous or synonymous (Figure 2D). Therefore, the user can choose LRLD-SNP pairs that occur in different genomic regions and that (in the case of coding SNPs) represent changes at the RNA or protein levels (the user can choose more than one type of interaction). The user can further select one or more population of interest to retrieve population-specific LDGIs. The results are downloadable (Figure 2E). For simplicity, the web interface displays only the first 10 records of each query (Figure 2F). The user can find detailed information of allele combinations of LRLD-linked SNPs and genomic regions where the linked SNPs are located in the results (Figure 2G). For the identified LDGIs, the interface also provides human PPI data collected from eight experiment-supported databases (i.e., collected PPIs and CRG PPIs) and high-confidence co-expression interactions for comparison. More detailed information of LDGI genes is also provided, including protein domain annotations, biological pathways, and disease associations.

Figure 2 The LDGIdb interface. (A) The three adjustable parameters. Users can search for LRLD-SNP pairs and LDGIs by setting the three adjustable parameters: HapMap Phase (II or III), *P* value of PLINK population clustering ($P < 0.01$ or $P < 0.001$), and r^2 value for linkage disequilibrium (≥ 0.8, ≥ 0.9, and 1). (B) Types of queries. Users can query by selecting the genomic types of the LRLD-linked SNP loci (D) and the population of interest (E), SNP accession number, gene accession number, or the coordinates of the genomic region of interest. (C) GWAS-related LRLD-SNP pairs. (F) and (G) are results. Users can download all records by clicking on the button (F). The first 10 records are displayed (G). If the linked SNP(s) is located within alternatively spliced genomic regions or overlapping genes, a LRLD-SNP pair record appears more than once with different genomic types or gene accession numbers in the downloaded file

Discussion and future development

Here we propose a new resource for studies of potential human gene interactions (i.e., LDGIs) based on haplotype data. In LDGIs, the linked genes are located in different chromosomes or LD blocks but are connected by one or more exonic/promoter SNP pairs that are subject to strong linkage disequilibrium ($r^2 \geq 0.8$, ≥ 0.9, or 1). We suggest that this LRLD approach and the LDGIdb can be potentially applied to the following areas. First, LDGIs may represent potential uncharacterized gene interactions, in which the functional associations

between the LDGI genes may not be explicitly indicated in other biological networks. Second, although we constructed the LDGIdb using SNP data in this study, the LRLD approach can actually be expanded to include other types of genomic variants such as copy number variation and insertion/deletion. Third, given enough haplotype information, population-specific LDGIs/LRLD-SNP pairs may be identified for more focused studies, particularly in the field of pharmacogenetics. Fourth, the correlation between the LDGIs/LRLD-SNP pairs and disease-associated SNPs such as those identified in GWAS studies can be explored. For example, SNP rs393152, which is associated with Parkinson's disease [36], forms an LRLD-SNP pair with rs12185268. Interestingly, rs12185268 was demonstrated to be connected to the same disease [37] two years after the publication (i.e., Ref #36) of the association of rs393152 with the disease. Another example is the LRLD-SNP pair: rs9858542–rs3197999. The two SNPs in this pair were shown to be related, respectively, to the Crohn's disease [38-41] and the ulcerative colitis [42,43]. These examples show that two SNPs that are associated with the same (or related) human diseases/traits can be identified by our approach. Moreover, there are also cases in which GWAS SNPs and their LDGI partners are associated with the same (or related) human diseases. For example, the GWAS SNP rs5215 in *KCNJ11* is known to be associated with Type II diabetes [44,45]. This SNP forms an LRLD-SNP pair with rs757110, which is located within the CDS of *ABCC8*. Mutations and deficiencies in the protein encoded by *ABCC8* have been suggested to be associated with hyperinsulinemic hypoglycemia of infancy and non-insulin-dependent diabetes mellitus type II [46,47]. The above examples suggest that the LRLD-SNP linkages may reflect biological interactions in the human molecular system and have the potential to detect previously uncharacterized gene interactions. As disease-association data accumulate, the LDGIdb may become an increasingly powerful tool by which to identify potentially uncharacterized disease-associated gene interactions, contributing to network-based studies of human diseases. Notably, however, since the majority of HapMap SNPs are relatively common variants, the linkages of rare alleles may not be represented in LDGIdb.

This study actually examined whether observed non-physical SNP linkages occur simply by chance or whether they are biologically meaningful. The above examples suggest that the inferred LDGIs may be functionally relevant. One interesting question is what are the molecular mechanisms underlying the inferred gene interactions. For the CDS-CDS LDGIs that involve only nonsynonymous changes, the functional association is speculated to result from direct or indirect protein-level interactions. Of course, the LDGIs may also represent adventitious linkages or false positives that result from unknown population substructures. Meanwhile, the biological meanings of the LDGIs that involve UTR SNPs (i.e., CDS-UTR and UTR-UTR linkages) or synonymous SNPs (i.e., nonsynonymous-synonymous and synonymous-synonymous linkages) may be more subtle. These potential interactions may be associated with translational regulation. Specifically, 5′UTRs may contain multiple sequence features that are involved in translational regulation, including upstream open reading frames, secondary structures, internal ribosome entry sites, and iron regulatory protein binding sites [48]. The disruption of these functional elements may cause changes in the efficiency of protein translation. On the other hand, 3′UTRs are known to be the major binding target of microRNAs, which can also suppress protein expression [49]. In addition, 3′UTRs may harbor protein-interacting secondary structures or the signals of nonsense-mediated decay or polyadenylation [48], both of which can affect the efficiency of protein translation. Meanwhile, synonymous coding SNPs are known to affect mRNA stability and splicing, leading to changes in the corresponding protein products [50]. Since both the UTR and synonymous SNPs may affect protein abundance, dosage imbalance and unidentified, indirect protein interactions may be possible explanations for the observed linkages.

Availability and requirements

Project name: LDGIdb project

Availability: LDGIdb is freely accessible at http://LDGIdb.genomics.sinica.edu.tw. Operating systems: Platform independent

Programming language: Javascript, CSS, PHP

Other requirements: None

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

TJC conceived and designed the study. FCC, YTH and TJC conducted the analyses. MCW and YZC built the web server. FCC and TJC wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We especially thank Shaou-Yen Liu and the GRC Information group for technical assistance and the HapMap III project team for providing information about phasing data. This work was supported by the National Science Council, Taiwan (under grants NSC99-2628-B-001-008-MY3 (to T.-J.C.) and National Health Research Institutes intramural funding (to F.-C.C.)

References

1. Barabasi AL, Oltvai ZN: **Network biology: understanding the cell's functional organization.** *Nature reviews* 2004, **5**(2):101–113.

2. Benyamini H, Friedler A: **Using peptides to study protein-protein interactions.** *Future Med Chem* 2010, **2**(6):989–1003.

3. Khan SH, Ahmad F, Ahmad N, Flynn DC, Kumar R: **Protein-protein interactions: principles, techniques, and their potential role in new drug development.** *J Biomol Struct Dyn* 2011, **28**(6):929–938.

4. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P: **Coexpression analysis of human genes across many microarray data sets.** *Genome Res* 2004, **14**(6):1085–1094.

5. Ramani AK, Li Z, Hart GT, Carlson MW, Boutz DR, Marcotte EM: **A map of human protein interactions derived from co-expression of human mRNAs and their orthologs.** *Mol Syst Biol* 2008, **4:**180.

6. Cooper WN, Hesson LB, Matallanas D, Dallol A, von Kriegsheim A, Ward R, Kolch W, Latif F: **RASSF2 associates with and stabilizes the proapoptotic kinase MST2.** *Oncogene* 2009, **28**(33):2988–2998.

7. Murphy DM, Buckley PG, Das S, Watters KM, Bryan K, Stallings RL: **Co-localization of the oncogenic transcription factor MYCN and the DNA methyl binding protein MeCP2 at genomic sites in neuroblastoma.** *PLoS One* 2011, **6**(6):e21436.

8. Tillier ER, Charlebois RL: **The human protein coevolution network.** *Genome Res* 2009, **19**(10):1861–1871.

9. Zill OA, Scannell D, Teytelman L, Rine J: **Co-evolution of transcriptional silencing proteins and the DNA elements specifying their assembly.** *PLoS Biol* 2010, **8**(11):e1000550.

10. Jiang X, Liu B, Jiang J, Zhao H, Fan M, Zhang J, Fan Z, Jiang T: **Modularity in the genetic disease-phenotype network.** *FEBS Lett* 2008, **582**(17):2549–2554.

11. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL: **The human disease network.** *Proc Natl Acad Sci U S A* 2007, **104**(21):8685–8690.

12. Lee DS, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabasi AL: **The implications of human metabolic network topology for disease comorbidity.** *Proc Natl Acad Sci U S A* 2008, **105**(29):9880–9885.

13. Park J, Lee DS, Christakis NA, Barabasi AL: **The impact of cellular networks on disease comorbidity.** *Mol Syst Biol* 2009, **5:**262.

14. Wall JD, Pritchard JK: **Haplotype blocks and linkage disequilibrium in the human genome.** *Nature reviews* 2003, **4**(8):587–597.

15. Stephan W, Song YS, Langley CH: **The hitchhiking effect on linkage disequilibrium between linked neutral loci.** *Genetics* 2006, **172**(4):2647–2663.

16. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci U S A* 2009, **106**(23):9362–9367.

17. **OMIM:** [http://omim.org/].

18. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, *et al*: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449**(7164):851–861.

19. Stephens M, Donnelly P: **A comparison of bayesian methods for haplotype reconstruction from population genotype data.** *Am J Hum Genet* 2003, **73**(5):1162–1169.

20. **Ensembl genome browser:** [http://www.ensembl.org/index.html].

21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, *et al*: **PLINK: a tool set for whole-genome association and population-based linkage analyses.** *Am J Hum Genet* 2007, **81**(3):559–575.

22. **HapMap:** [http://hapmap.ncbi.nlm.nih.gov/].

23. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, *et al*: **Human protein reference database--2009 update.** *Nucleic Acids Res* 2009, **37**(Database issue):D767–D772.

24. Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D: **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.** *Nucleic Acids Res* 2002, **30**(1):303–305.

25. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database.** *FEBS Lett* 2002, **513**(1):135–140.

26. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J, *et al*: **The IntAct molecular interaction database in 2010.** *Nucleic Acids Res* 2010, **38**(Database issue):D525–D531.

27. Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, *et al*: **Reactome: a knowledge base of biologic pathways and processes.** *Genome Biol* 2007, **8**(3):R39.

28. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: a general repository for interaction datasets.** *Nucleic Acids Res* 2006, **34**(Database issue):D535–D539.

29. Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Mark P, Stumpflen V, Mewes HW, *et al*: **The MIPS mammalian protein-protein interaction database.** *Bioinformatics (Oxford, England)* 2005, **21**(6):832–834.

30. Bossi A, Lehner B: **Tissue specificity and the human protein interaction network.** *Mol Syst Biol* 2009, **5:**260.

31. **GWAS:** [http://www.genome.gov/gwastudies/#1].

32. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R, *et al*: **New developments in the InterPro database.** *Nucleic Acids Res* 2007, **35**(Database issue):D224–D228.

33. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, *et al*: **KEGG for linking genomes to life and the environment.** *Nucleic Acids Res* 2008, **36**(Database issue):D480–D484.

34. Becker KG, Barnes KC, Bright TJ, Wang SA: **The genetic association database.** *Nat Genet* 2004, **36**(5):431–432.

35. Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC, *et al*: **DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists.** *Nucleic Acids Res* 2007, **35**(Web Server issue):W169–W175.

36. Simon-Sanchez J, Schulte C, Bras JM, Sharma M, Gibbs JR, Berg D, Paisan-Ruiz C, Lichtner P, Scholz SW, Hernandez DG, *et al*: **Genome-wide association study reveals genetic risk underlying Parkinson's disease.** *Nat Genet* 2009, **41**(12):1308–1312.

37. Do CB, Tung JY, Dorfman E, Kiefer AK, Drabant EM, Francke U, Mountain JL, Goldman SM, Tanner CM, Langston JW, *et al*: **Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease.** *PLoS Genet* 2011, **7**(6):e1002141.

38. Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447**(7145):661–678.

39. Parkes M, Barrett JC, Prescott NJ, Tremelling M, Anderson CA, Fisher SA, Roberts RG, Nimmo ER, Cummings FR, Soars D, *et al*: **Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility.** *Nat Genet* 2007, **39**(7):830–832.

40. Franke A, McGovern DP, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, Lees CW, Balschun T, Lee J, Roberts R, *et al*: **Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci.** *Nat Genet* 2010, **42**(12):1118–1125.

41. Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, *et al*: **Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease.** *Nat Genet* 2008, **40**(8):955–962.

42. Barrett JC, Lee JC, Lees CW, Prescott NJ, Anderson CA, Phillips A, Wesley E, Parnell K, Zhang H, Drummond H, *et al*: **Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region.** *Nat Genet* 2009, **41**(12):1330–1334.

43. McGovern DP, Gardet A, Torkvist L, Goyette P, Essers J, Taylor KD, Neale BM, Ong RT, Lagace C, Li C, *et al*: **Genome-wide association identifies multiple ulcerative colitis susceptibility loci.** *Nat Genet* 2010, **42**(4):332–337.

44. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G, *et al*: **Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes.** *Nat Genet* 2008, **40**(5):638–645.

45. Cho YM, Kim TH, Lim S, Choi SH, Shin HD, Lee HK, Park KS, Jang HC: **Type 2 diabetes-associated genetic variants discovered in the recent genome-wide association studies are related to gestational diabetes mellitus in the Korean population.** *Diabetologia* 2009, **52**(2):253–261.

46. Mannikko R, Flanagan SE, Sim X, Segal D, Hussain K, Ellard S, Hattersley AT, Ashcroft FM: **Mutations of the same conserved glutamate residue in NBD2 of the sulfonylurea receptor 1 subunit of the KATP channel can result in either hyperinsulinism or neonatal diabetes.** *Diabetes* 2011, **60**(6):1813–1822.

47. Zhou K, Bellenguez C, Spencer CC, Bennett AJ, Coleman RL, Tavendale R, Hawley SA, Donnelly LA, Schofield C, Groves CJ, *et al*: **Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes.** *Nat Genet* 2011, **43**(2):117–120.

48. Chatterjee S, Pal JK: **Role of 5′- and 3′-untranslated regions of mRNAs in human diseases.** *Biol Cell* 2009, **101**(5):251–262.

49. Bartel DP: **MicroRNAs: target recognition and regulatory functions.** *Cell* 2009, **136**(2):215–233.

50. Chamary JV, Parmley JL, Hurst LD: **Hearing silence: non-neutral evolution at synonymous sites in mammals.** *Nat Rev Genet* 2006, **7**(2):98–108.

HapMap Phasing data

Extracting SNPs located within promoter or exonic regions

PLINK

Performing clustering of subpopulations for the individuals examined in the HapMap project

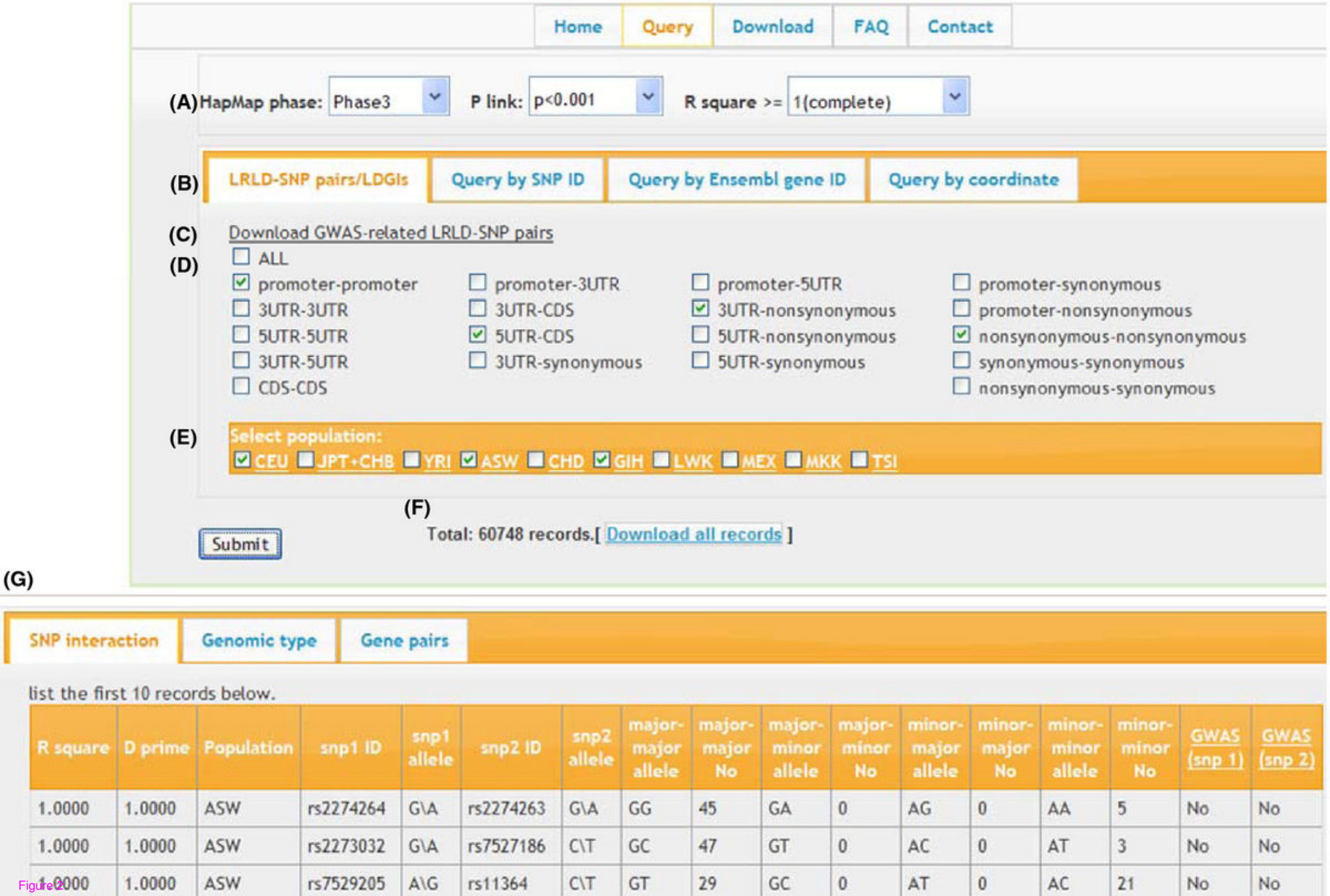
Calculating LD scores (r^2) for any two extracted SNPs, if two SNPs satisfy the following rules, they are linked:
① they are subject to $r^2 \geq 0.8$
② they are not belonged to the same LD block

LRLD-SNP pairs

Extracting LDGIs: if two genes are linked by LRLD-SNP pairs

LDGIs

Figure 1



R square	D prime	Population	snp1 ID	snp1 allele	snp2 ID	snp2 allele	major-major allele	major-major No	major-minor allele	major-minor No	minor-major allele	minor-major No	minor-minor allele	minor-minor No	GWAS (snp 1)	GWAS (snp 2)
1.0000	1.0000	ASW	rs2274264	G\A	rs2274263	G\A	GG	45	GA	0	AG	0	AA	5	No	No
1.0000	1.0000	ASW	rs2273032	G\A	rs7527186	C\T	GC	47	GT	0	AC	0	AT	3	No	No
1.0000	1.0000	ASW	rs7529205	A\G	rs11364	C\T	GT	29	GC	0	AT	0	AC	21	No	No

Figure 2