BMC Genomics

RESEARCH ARTICLE   **Open Access**

Meta-analysis and genome-wide interpretation of genetic susceptibility to drug addiction

Chuan-Yun Li[1,2*], Wei-Zhen Zhou[2], Ping-Wu Zhang[3], Catherine Johnson[4], Liping Wei[2] and George R Uhl[4*]

Abstract

Background: Classical genetic studies provide strong evidence for heritable contributions to susceptibility to developing dependence on addictive substances. Candidate gene and genome-wide association studies (GWAS) have sought genes, chromosomal regions and allelic variants likely to contribute to susceptibility to drug addiction.

Results: Here, we performed a meta-analysis of addiction candidate gene association studies and GWAS to investigate possible functional mechanisms associated with addiction susceptibility. From meta-data retrieved from 212 publications on candidate gene association studies and 5 GWAS reports, we linked a total of 843 haplotypes to addiction susceptibility. We mapped the SNPs in these haplotypes to functional and regulatory elements in the genome and estimated the magnitude of the contributions of different molecular mechanisms to their effects on addiction susceptibility. In addition to SNPs in coding regions, these data suggest that haplotypes in gene regulatory regions may also contribute to addiction susceptibility. When we compared the lists of genes identified by association studies and those identified by molecular biological studies of drug-regulated genes, we observed significantly higher participation in the same gene interaction networks than expected by chance, despite little overlap between the two gene lists.

Conclusions: These results appear to offer new insights into the genetic factors underlying drug addiction.

Background

Twin and other classical genetic studies indicate that drug addiction is a complex brain disorder with strong genetic contributions [1,2]. Genetic association studies, including candidate gene studies and genome-wide association studies (GWAS), can provide insights into the genetic background of this neurobiological and behavioral disorder. Using these approaches, more than 800 publications during the past three decades have reported genomic loci and/or specific genetic variants that have been associated with susceptibility to drug addiction. It has been difficult to draw general inferences from these studies, however, because genetic association studies generated results that were sometimes inconsistent, many of these studies were modestly powered (especially when polygenic genetic architectures are considered), genomic controls are infrequent, and biases can be

detected in a number of analytic strategies. In this context, meta-analysis of genetic association studies may be particularly useful, especially when the underlying genetic architecture for the disorder is relatively straightforward [3-6]. In addition, although different addictive drugs have disparate pharmacological effects, there are also similarities after acute and chronic exposure such as acute rewarding and negative emotional symptoms upon drug withdrawal [7]. It would thus be interesting to elucidate the potential 'common' genetic backgrounds underlying those shared addictive actions, which might further help the development of effective treatments for a wide range of addictive disorders [7,8]. However, to date there have only been limited meta-analyses on drug addiction, mostly focused on candidate genes, and none on GWAS.

Although the number of genetic variations identified has increased rapidly, the understanding of how genetic variations contribute to disease susceptibility has lagged behind. Earlier studies mainly focused on nonsynonymous SNPs [9,10]. More recent studies have attempted to explain functional mechanisms of action of

* Correspondence: lichuanyun@gmail.com; guhl@intra.nida.nih.gov
[1]Laboratory of Bioinformatics and Genomic Medicine, Institute of Molecular Medicine, Peking University, Beijing, China
[4]Molecular Neurobiology Branch, NIH-IRP (NIDA), Suite 3510, 333 Cassell Drive Baltimore, Maryland, USA
Full list of author information is available at the end of the article

haplotypes that contain SNP and other regulatory variants [11-17]. A number of haplotypes defined by specific SNPs have been found to alter gene expression by modifying transcription factor binding sites [11], micro-RNA binding sites [12-15] and alternative splicing [16]. Others regulate signaling pathways [17]. However to date there have been only modest genome-scale efforts to study the molecular mechanisms of addiction-associated genetic variants. The relative contributions of different molecular mechanisms remain largely unknown.

Previous work has been spotty in seeking or reporting overlap between the genes identified by genetic association studies and genes identified by other molecular biologic approaches, such as animal models, cDNA microarrays and proteomics [18-20]. Our prior systematic assembly of data obtained by these "other" approaches [21] allows us to seek such overlaps in a systematic fashion.

Results

Meta-analysis of genetic association studies of drug addiction

First, we performed an integration and meta-analysis of candidate gene association studies of drug addiction. We retrieved 886 publications on candidate gene association studies of drug addiction from PubMed by keywords query and review paper curation (**See details in Methods**). Two hundred and twelve of these reports met our inclusion criteria, from which we extracted data on 506 allelic contrast tests for 286 genetic variants (Additional file 1). Thirty-five genetic variants were examined in case-control genotype comparisons from three or more independent datasets. We carried out meta-analyses of these 35 genetic variants under simple genetic models using both the random-effects model and fixed-effects model [22,23]. From these data, 12 genetic variants in 11 genes showed effects that reached statistical significance (Table 1, Additional file 2). We noted that most of these variants show comparatively weak genetic effects, with fixed effects summary odds ratios (OR) ranging from 0.52 to 2.34 (Table 1), typical results for studies on other highly heritable phenotypes with "common variants, common disease" design [3-6]. We further assessed the variants using criteria established by the HuGENet Road Map [24] that was recently proposed for assessing cumulative evidence from genetic association studies. Using these stringent criteria, six variants displayed moderate epidemiological credibility (Grade B, Table 1). A full list of the curated information is available online at http://karg.cbi.pku.edu.cn/karg2/. For each study, we extracted meta-data including over thirty demographic and experimental variables where available (Additional file 1).

Next, we retrieved 11 independent datasets of drug addiction GWAS [25-31] (**See Details in Methods**). Among them five datasets met our criteria for inclusion [25-27]. We integrated the five GWAS datasets using a new meta-analysis approach to select positive SNPs with significantly more GWAS support than expected by chance (**See Details in Methods**). Overall, 842 SNPs were supported by at least three items of positive evidence, with meta-false discovery rates less than 0.05 (Additional file 3).

Table 1 Genes and polymorphisms showing significant summary odds ratio (OR) of the addiction susceptibility from random/fixed-effects meta-analyses using allelic contrasts

Gene Name	Polymorphism*	Model (Major allele > Minor Allele)	Cases vs. Controls (independent samples)	Fixed Effects OR (95% CI)	Random Effects OR (95% CI)	Heterogeneity *p*-Value	I-Square	Grade**
SLC4A7	rs3278	G > A	1410 vs. 906 (3)	2.34 (1.599-3.420)	2.28 (1.555-3.333)	0.51	0	B
DRD4	48-bp repeat	Other > 7/8 repeats	2324 vs. 1932 (6)	1.44 (1.155-1.804)	1.48 (1.000-2.197)	0.06	52	C
DRD2/ ANKK1***	Taq1A	A2 > A1	6312 vs. 7424 (20)	1.30 (1.192-1.410)	1.38 (1.096-1.733)	< 0.0001	84	C
BDNF	rs6265	G > A	2530 vs. 4126 (9)	1.31 (1.165-1.451)	1.38 (1.056-1.790)	< 0.0001	80	C
CCK	-45 C/T	C > T	860 vs. 2002 (6)	1.34 (1.089-1.650)	1.34 (1.083-1.646)	0.62	0	B
FAAH	rs324420	P > T	498 vs. 1570 (3)	1.38 (1.014-1.875)	1.32 (0.807-2.171)	0.24	28	B
OPRM1	rs1799971	A > G	2846 vs. 4072 (9)	1.24 (1.090-1.410)	1.31 (0.958-1.790)	< 0.0001	80	C
COMT	rs4680	Val > Met	862 vs. 1594 (3)	0.76 (0.634-0.923)	0.82 (0.644-1.051)	0.71	0	B
CNR1	(AAT)n	14 repeats > Other	2304 vs. 2144 (8)	0.76 (0.658-0.878)	0.75 (0.619-0.906)	0.17	32	B
HNMT	rs35953316	Thr > Ile	1540 vs. 1306 (3)	0.76 (0.598-0.975)	0.72 (0.444-1.179)	0.04	70	C
OPRK1	rs702764	A > G	292 vs. 246 (3)	0.62 (0.431-0.901)	0.62 (0.412-0.944)	0.99	0	B
OPRM1	C691G	C > G	796 vs. 786 (3)	0.52 (0.416-0.647)	0.61 (0.330-1.095)	0.0025	83	C

*Variants were ranked based on the summary ORs. **Degree of 'epidemiological credibility' based on published protocols (A, strong; B, modest; C, weak; see Methods for more details). ***Researchers previously associated the polymorphism Taq 1A to the *DRD2* gene. However, the polymorphism sits in an exon of the *ANKK1* gene.

We combined the findings identified by candidate gene association studies and GWAS into a list of 849 SNPs in 843 haplotypes. Since many of the genetic susceptibility SNPs may provide genetic 'tag markers', while these tag SNPs were generally designed to detect linkage disequilibrium blocks and functional SNPs may be easily left out in most GWAS platforms [32], we thus used the whole-genome linkage disequilibrium data identified by HapMap [33] to expand the list into 1,907 SNPs by adding SNPs that displayed strong linkage disequilibrium with these genetic marker SNPs in all three HapMap populations.

Genome-wide analysis of possible molecular mechanisms of the addiction susceptibility factors

We mapped the 1,907 SNPs to putative functional elements in the human genome. As summarized in Table 2 and detailed in Additional file 4, we identified a total of 124 putative functional SNPs in 70 of the haplotype blocks identified herein. Only 26 of these putative functional SNPs, in 23 haplotypes, were non-synonymous. One SNP was located in splicing junctions. Four lay in putative transcription factor binding sites. Two lay in potential microRNA target sites. By integrating data from high-throughput studies that have correlated human genotypes with levels of gene expression (**See Details in Methods**), we found that 24 SNPs in two haplotypes were strongly correlated with differential expression of at least one human gene, one haplotype also contain SNP located in transcription factor binding sites, providing a possible explanation for the observed correlations (Additional file 4).

Additional evidence for functional roles for many of these SNPs came from studies of apparent effects of natural selection. A total of 31 SNP in 26 haplotypes displayed evidence for negative selection. Thirty-four SNPs in 19 haplotypes displayed evidence for positive selection. Signals of recent positive selection provide information about the adaptation of humans to local conditions and have been implicated in phenotypic variations [34]. Thus, the 6 genes located in these regions of positive selection may be of particular interest in studying addiction vulnerabilities.

We estimated the magnitude of the contributions of different molecular mechanisms to the effects of addiction susceptibility. We compared observed values to those that would be obtained by chance based on 10,000 *Monte Carlo* simulations (**See Methods for details**). The categories of 'synonymous SNP' ($p = 0.001$) and 'non-synonymous SNP' ($p = 0.001$) showed nominally significant over-representation, consistent with the conventional idea that SNPs in coding regions may play important roles in disease susceptibility. In addition, the data suggest regulatory SNPs that modify transcription factor, microRNA binding or alternative splicing sites, may also contribute to addiction susceptibility in addition to those played by non-synonymous SNPs and other allelic variants.

Genetic association findings and molecular biology findings form significantly more gene interactions

The 124 functional SNPs identified belong to 50 genes. These addiction susceptible genes are enriched in several functional categories such as focal adhesion (hypergeometric test, p-Value = 0.02) that had been previously reported to be involved in drug addiction [21]. We compared these findings to findings from molecular biological studies extracted from the Knowledgebase for Addiction Related Genes (KARG) [21]. In KARG, 348 genes are linked to addiction susceptibility by at least two independent lines of molecular biologic evidence such as results from animal mutagenesis, microarray

Table 2 Functional categories of addiction susceptibility SNPs

Functional Categories	Vulnerable SNP Number	Haplotype Number	*Monte Carlo* *p*-values
SNPs Introducing Non-Synonymous Mutations	26	23	0.001*
SNPs Introducing Synonymous Mutations	25	21	0.001*
SNPs Introducing Stop Codon Gain	0	0	1.00
SNPs Introducing Stop Codon Lost	0	0	1.00
SNPs Introducing ORF Frame Shift	0	0	1.00
SNPs Introducing Altered Splicing Junction	1	1	0.92
SNPs Introducing Altered TF Binding Sites	4	4	0.83
SNPs Introducing Altered miRNA Targets	2	2	0.28
SNPs Correlated With Differentially Gene Expression	24	2	0.42
SNPs Under Positive Selection	34	19	0.99
SNPs Under Negative Selection	31	26	0.05*
Functional Addiction Susceptibility SNPs	124	70	0.63
All SNPs in The Positive Haplotypes	1907	843	1.00

mRNA profiling and proteomics profiling. Only four genes were common between the two genetic association findings and the molecular biology findings (Official Symbol: FAAH, OPRM1, OPRK1, BDNF), consistent with previously observed modest overlaps between genetic and molecular biology findings in studies of other diseases [35].

We set out to explain this difference with further analysis. Because of the different nature of genetic experiments and molecular biology experiments, could they have discovered different genes in the same molecular network underlying addiction? We hypothesized that the genes identified by genetic studies and those by molecular biology studies may interact more frequently than expected by chance. Indeed, gene interaction enrichment analyses (**See Details in Methods**) revealed that genes identified by these two types of studies interact with each other more than expected by chance. The addiction susceptibility genes formed interactions with 37.2% (89/239) of the addiction-related genes identified by molecular biology studies that had known interaction data (*Monte Carlo p*-value < 0.0001). This result thus provides one explanation for the differences between the genes identified through genetics and those identified through molecular biologic and molecular pharmacologic approaches.

Development of an updated version of KARG database

We make all of our new data publicly available in an updated version of a comprehensive knowledgebase for addiction-related genes, KARG [21], available at http://karg.cbi.pku.edu.cn/karg2/.

Discussion

In this study, we collected genetic association studies published in the field of drug addiction for meta-analyses. The power of such meta-analyses is linked to the relatively simple model of the underlying genetic architecture that they presuppose: that SNP genotype results from different samples with differences in genetic background will provide association with drug addiction with the same phase. The significant convergence that such analyses provide herein does support roles for genetic variants with these properties in some aspects of individual differences in susceptibility to dependence. However, recent analyses also provide evidence for roles in addiction susceptibility for more "recent" variants raised after population divergences, which are less likely to be identified by such meta-analytic procedures. Besides the 'common' genetic background identified, it is also interesting to evaluate susceptible variants for different addictive drugs. However, currently the number of available allelic contrast tests data was too limited to perform such an analysis. In the future we will continue to integrate new data toward a better understanding of drug addiction. In addition, recent re-sequencing efforts using next-generation deep sequencing technology support larger effects for at least some rarer variants in both Mendelian [36-40] and complex diseases [41,42], which would also be missed by the current analyses. Nevertheless, the interesting findings from these meta-analyses is complementary to recently published gene-based approach that was used to analyze primary GWAS data in ways that allow for substantial allelic and locus heterogeneities [25-27]. This study also provided an opportunity to study the relationship between addiction susceptible genes identified by traditional genetic association studies and rare addiction causal variants linked by "common disease, rare variants" approaches, when more genomic re-sequencing efforts become available [43-45].

Over 800 candidate gene association studies have been published in this field, but only 212 (24%) of these reports met our inclusion criteria. Some papers published 20 years ago were missing raw genotype and allelic distribution data and had inconsistent use of genetic markers. In addition, since the number of available allelic contrast tests was limited, we combined all data regardless of the types of addictive drugs and the racial/ethnic composition of the group studied. The heterogeneity of the datasets was high: even after our comprehensive meta-analysis, the results were still Grades B and C, according to the criteria of the Human Genome Epidemiology Network (HuGENet) (Table 1). Protocols such as those proposed by HuGENet [24] could standardize data collection and reporting and allow for improved meta-analyses in the future.

Compared to candidate gene association studies, GWAS provide hypothesis-free, genome-wide view of possible genetic susceptibility factors underlying drug addiction [25-31]. When we compare the addiction susceptible genetic variants linked by candidate gene association studies and GWAS, we found that the GWA arrays included probes for three polymorphisms showing significant summary odds ratio of the addiction susceptibility (rs6265, rs1799971 and rs4680). Among these polymorphisms, only rs1799971 show some suggestive significance in methamphetamine abusers of Japanese (p-Value = 0.0465) [26]. Consistent with meta-analyses in Alzheimer disease, schizophrenia, major depressive disorder and Parkinson disease [3-6], it seems some important candidate genes have received inordinate attention in candidate-gene based association study, while the GWA studies with hypothesis-free design might not support many *a prior* hypothesis. On the other hand, GWAS provide more opportunities for traditional candidate-gene based association study to improve the experimental designs by avoiding potential

biases from subjectively selection of candidate genes in the beginning of the study.

We were able to tentatively link 124 of the identified susceptibility variations to potential functional mechanisms (Additional file 4). We expanded the genetics tag SNPs using haplotype data to detect the most likely nearly functional SNPs and genes. Besides fitting with the conventional idea that SNPs in coding regions may play important roles in disease susceptibility, the analyses presented here suggest that regulatory SNPs may also play important roles in addiction susceptibility. It will be interesting to study why and how natural selection shaped these *cis*-regulatory factors that potentially modulate addiction susceptibility.

To explain the modest overlap between genetic association findings and other molecular biology findings at the gene level, we identified abundant evidence for interactions between the sets of genes identified in these two ways. Thus, at the level of network analysis, there was good consistency between the genetic and molecular biologic results. This new insight should continue to motivate communication between geneticists and molecular biologists as they study addiction from different perspectives.

Conclusions

In this study, we report the first comprehensive meta-analysis of genetic association studies in drug addiction. We curated and integrated 212 candidate gene association studies and 5 GWAS. 843 vulnerable haplotypes were identified. We estimated the magnitude of the contributions of different molecular mechanisms to the effects of addiction susceptibility in one of the first 'post-GWAS' global attempts. We further found that at the levels of gene interaction networks, there was in fact good consistency between the genes identified by association studies and those identified by molecular biological studies of drug-regulated genes.

We have made all new data and knowledge publicly available by updating the KARG database [21]. Our study thus provides a 'dynamic' approach. We hope that this approach, as it stands, will provide a basis for meta-analyses of GWAS results of other diseases under the simple genetic architectures postulated herein, as well as a basis for consideration of meta-analytic approaches to more complex architectures in which the focus might be on genes in which variants that display differing frequencies in individuals with different genetic backgrounds are likely to be located. Such analyses could conceivably integrate both the idea of more population-specific variants with the rare variants that are being identified in disease and control samples through re-sequencing efforts.

Methods

Figure 1 shows the overall pipeline of our meta-analyses of addiction-associated genetic variations, genome-wide analysis of the molecular mechanisms of implicated SNPs, and the pathways and gene interaction networks that might involve these genetic factors.

Meta-analyses of candidate genetic association studies of drug addiction

To identify the candidate genetic association studies, we performed a search for all abstracts deposited in PubMed database (National Center for Biotechnology Information; NCBI) using the keywords '("addiction" OR "abuse*") AND ("genetic*" AND "association*")'. To identify publications using different phenotype terms other than 'addiction' or 'abuse', we also identify candidate genetic association studies from published reviews selected from PubMed query under the keywords '("addiction" OR "abuse*" OR "dependen*") AND "genetic*" AND Review[ptyp]'. The combined approach resulted in 886 articles. All 886 abstracts were retrieved from PubMed database and manually curated by two independent reviewers. Only those studies that met the following criteria were included in further analyses: (i) It must represent an assessment of association between a polymorphic genetic marker (including SNP and microsatellite markers) and drug addiction phenotypes. Only studies focused exclusively on case-control or population-based designs were included. Studies on markers with more than three alleles (which are generally more difficult to determine unequivocally across different laboratories) or those with an otherwise complex allelic architecture were not considered for meta-analysis. (ii) The study must be published in a peer-reviewed English scientific journal as original research articles. This explicitly excludes studies reported only in the form of an abstract. This yielded 212 papers eligible for inclusion in this study (Figure 1).

From each publication, full text of the original papers were downloaded and manually curated to extracted meta-data, such as publication information ('PubMed ID', 'First Author', 'Title', 'Year' and 'Study Method'), sample information ('Study Design', 'Sample Size,' 'Age', 'Gender Ratio' and 'Ethnic Group'), drug information ('Addictive Drug' and 'Behavior Description'), genotype information ('Gene ID', 'SNP/Marker ID', 'Primary Significance Report', 'Detailed Genotypes in Case and Control' and 'HWE P-value') and curation information ('Curator' and 'Date') (Additional file 1). A full list for the curated information is online available at http://karg.cbi.pku.edu.cn/karg2/.

For all variants with case-control genotype data available in three or more independent samples, we
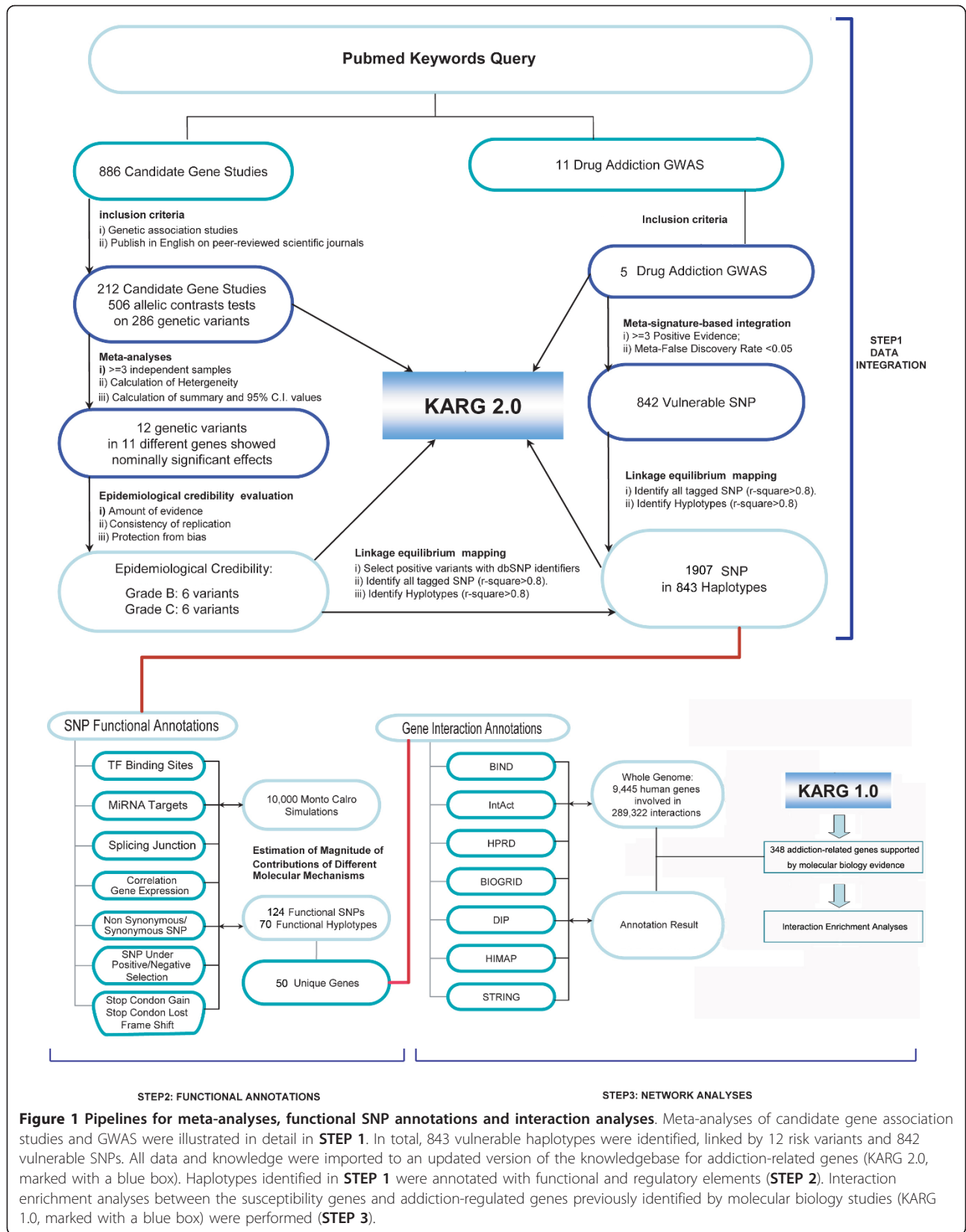
Figure 1 Pipelines for meta-analyses, functional SNP annotations and interaction analyses. Meta-analyses of candidate gene association studies and GWAS were illustrated in detail in **STEP 1**. In total, 843 vulnerable haplotypes were identified, linked by 12 risk variants and 842 vulnerable SNPs. All data and knowledge were imported to an updated version of the knowledgebase for addiction-related genes (KARG 2.0, marked with a blue box). Haplotypes identified in **STEP 1** were annotated with functional and regulatory elements (**STEP 2**). Interaction enrichment analyses between the susceptibility genes and addiction-regulated genes previously identified by molecular biology studies (KARG 1.0, marked with a blue box) were performed (**STEP 3**).

calculated crude ORs (Odds Ratios) and 95% C.I. values of the addiction susceptibility from the allelic distributions for each study following the published protocols [4]. Summary ORs and 95% C.I. values of the addiction susceptibility were then calculated using both the fixed-effects model and DerSimonian & Laird random-effects model [22]. We further graded the epidemiological credibility of these genetic associations according to the criteria of the Human Genome Epidemiology Network (HuGENet) [24]. Details of the grading system followed Ioannidis *et al* [46]. Briefly, each meta-analyzed association is graded on the basis of the amount of evidence, consistency of replication, and protection from bias, following the published protocols [4].

Meta-analyses of GWAS

On the basis of PubMed query under the keywords 'addiction AND association* AND genome' followed by manually curation, we identified 7 GWAS on drug addiction, containing 11 independent samples. Five of them met our inclusion criteria: i) genetic association studies with case-control design, ii) published in peer-reviewed English scientific journals, iii) the original case-control genotype data is available (raw data available with adequate ethnic approval) and iv) the genotype data are generated by comparable genotyping platforms and arrays with density designs. Detailed raw data of the five GWAS datasets came from the Molecular Neurobiology Branch, NIH-IRP (NIDA) led by Dr. George Uhl, including i) 500 K SNP genotype data from 560 African-American poly-substance abusers who reported dependence on at least one illegal substance and 360 controls [25]; ii) 500 K SNP genotype data from 420 European-American poly-substance abusers who reported dependence on at least one illegal substance and 320 controls [25]; iii) 500 K SNP genotype data from 140 methamphetamine abusers of ethnic Han Chinese origin, with 240 Han Chinese controls [26]; iv) 500 K SNP genotype data from 100 methamphetamine abusers of Japanese origin, with 100 Japanese controls [26] and v) 100 K SNP genotype data from 120 alcohol-dependent individuals and 160 unrelated unaffected controls with European-American ethnicities [27]. Initial data analyses were performed and statistical tests were conducted to assess the susceptibility of each SNP marker [25].

We assumed that results from different GWAS should share a significant intersection of addiction vulnerable SNPs which would be genetic factors underlying drug addiction in general, regardless of addictive drug types and population demographics [32]. We thus implemented a "meta-signature" approach following the "meta-signature" method that Oncomine used to identify common gene-expression signatures [47]. Briefly, (i) Five

GWAS as described in the previous paragraph were selected for meta-signature study; (ii) Significant thresholds (T) were chosen to define positive SNPs in the 5 selected GWAS; (iii) Positive SNPs were selected in each GWAS result; (iv) Positive SNPs were sorted by the number of GWAS positive findings in which they are present; (v) the numbers of positive SNPs with 1~5 supporting GWAS were tallied as (T_1, T_2, T_3, T_4, T_5); (vi) 10,000 random permutations were performed, in which the actual p-values were randomly assigned to SNPs within each GWAS, so that the positive SNPs in each GWAS change at random, but the number of positive SNPs remained the same. This simulation generated distributions about the number of positive SNPs with 1~5 supporting GWAS, with the means of these distributions tallied as (E_1, E_2, E_3, E_4, E_5); (vii) the significance of intersection for the real data was assessed by the minimum meta-false discovery rate (*mFDR*) calculated as $mFDR = \text{Minimum} ([E_i]/[T_i])$ for i = 1 to j, $1 < j <= 5$. If $mFDR < 0.05$, a meta-signature was defined as those SNPs that are significantly identified (p-value < T) in at least j of 5 independent GWAS, where j is equal to i when mFDR was defined. The p-Value threshold (T) with 0.05 and 0.01 were calculated respectively and significant results were combined for further study. On the basis of the HapMap Linkage Disequilibrium data compiled from genotype data (HapMap data release rel#21 NCBI B35) [33], we further expanded this list using SNP pairs with strong linkage disequilibrium (r^2> = 0.8) in all three HapMap populations. The protocol was implemented in Perl.

SNP functional annotations

Coordinates of the SNPs were retrieved from NCBI dbSNP database (Build 130) [48]. Genomic coordinates of 3' UTR, 5' UTR, intron region, intergenic regions, synonymous, non-synonymous, and splicing sites were retrieved from the UCSC Genome Browser Database (NCBI36/hg18) [49]. Regulatory elements including transcription factor binding sites and experimentally validated and putative miRNA targets were retrieved from TransFac [50], Argonaute [51], TarBase [14] and PicTar [52]. Information for SNPs under negative selection or positive selection was retrieved from published data [34,49]. The correlation between SNPs and gene expression were retrieved from high-throughput studies correlating human gene expression and genotypes. The full text papers of 11 such studies were manually curated for fulfillment of inclusion criteria of (i) neuro-pathologically normal samples, (ii) association design and iii) available statistics data. In all, four studies met the inclusion criteria [53-56]. A total list of 33,731 significantly correlated SNP-expression pairs was identified, involving 22,178 SNPs and 3,640 transcripts [53-56].

Then, for the addiction vulnerable SNPs, We estimated the magnitude of the contributions of different molecular mechanisms to the effects of addiction susceptibility. We further compared observed values to those that would be obtained by chance based on 10,000 *Monte Carlo* simulations. Briefly, the positive SNPs were randomly selected from all tag SNPs, but the number of positive SNPs remained the same. Then, for each SNP list, we performed the identical pipelines to estimate the contributions of different molecular mechanisms to the effects of addiction susceptibility. Perl and R scripts were implemented to integrate the datasets, annotate addiction vulnerable SNPs and perform statistical tests.

Functional enrichment analyses

Information about gene interactions comes from seven interaction databases including IntAct [57], BIND [58], HPRD [59], BioGRID [8], HiMAP, DIP and STRING [60]. We annotated all addiction susceptibility genes using these data. 10,000 *Monte Carlo* simulations were performed to estimate the distribution for testing the enrichment for interactions between addiction susceptibility genes and addiction-related genes identified by molecular biology studies, in which addiction susceptibility gene lists were randomly created from human genome, followed by the identical analyses pipelines for gene interaction annotations and calculations. *Monte Carlo p*-values < 0.05 were considered to be a sign for interaction enrichment between the two datasets to a statistically significant degree. We performed functional enrichment test for addiction susceptibility genes using KOBAS [61] and DAVID [62], following published protocols [21]. Functional categories with *p*-values < 0.05 were considered enriched in addiction susceptibility genes to a statistically significant degree.

Development of an updated version of KARG database

We updated KARG with the new data and knowledge discussed above. Cross-references to key external databases were included to integrate functional information about the genes, such as gene annotations [49], Gene Ontology (GO) annotations [63], interacting proteins [8,58,59] and functional domain annotations [64]. We enhanced the web-based user interface of the database using PHP and queries of the database using PHP/SQL query script.

Additional material

Additional file 1: Description of meta-data. Features integrated for each item of evidence.

Additional file 2: Forest plots of meta-analyses. Forest plots of meta-analyses using allelic contrasts for variations showing significant summary Odds Ratios (OR).

Additional file 3: Vulnerable SNPs identified by Meta-analyses of public GWAS. Meta-analyses of five genome-wide association studies (GWAS) identified 842 vulnerable SNPs for drug addiction.

Additional file 4: Functional annotations of addiction susceptibility SNPs. Addiction susceptibility variants and items of evidence.

Author details

[1]Laboratory of Bioinformatics and Genomic Medicine, Institute of Molecular Medicine, Peking University, Beijing, China. [2]Center for Bioinformatics, National Laboratory of Protein Engineering and Plant Genetic Engineering, College of Life Sciences, Peking University. Beijing, China. [3]Department of Neurology, the Johns Hopkins School of Medicine, Baltimore, Maryland, USA. [4]Molecular Neurobiology Branch, NIH-IRP (NIDA), Suite 3510, 333 Cassell Drive Baltimore, Maryland, USA.

Authors' contributions

CYL, LW and GU conceived and designed the experiments. CYL performed most of the experiments. CYL, WZZ, PWZ and CJ analyzed the data and performed the statistical analysis. CYL, LW and GU wrote the paper. All authors read and approved the final manuscript.

References

1. Uhl GR: **Molecular genetic underpinnings of human substance abuse vulnerability: likely contributions to understanding addiction as a mnemonic process.** *Neuropharmacology* 2004, **47(Suppl 1)**:140-147.
2. Crabbe JC: **Genetic contributions to addiction.** *Annu Rev Psychol* 2002, **53**:435-462.
3. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE: **Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database.** *Nat Genet* 2007, **39**:17-23.
4. Allen NC, Bagade S, McQueen MB, Ioannidis JP, Kavvoura FK, Khoury MJ, Tanzi RE, Bertram L: **Systematic meta-analyses and field synopsis of genetic association studies in schizophrenia: the SzGene database.** *Nat Genet* 2008, **40**:827-834.
5. Lopez-Leon S, Janssens AC, Gonzalez-Zuloeta Ladd AM, Del-Favero J, Claes SJ, Oostra BA, van Duijn CM: **Meta-analyses of genetic studies on major depressive disorder.** *Mol Psychiatry* 2008, **13**:772-785.
6. Evangelou E, Maraganore DM, Ioannidis JP: **Meta-analysis in genome-wide association datasets: strategies and application in Parkinson disease.** *PLoS ONE* 2007, **2**:e196.
7. Nestler EJ: **Is there a common molecular pathway for addiction?** *Nat Neurosci* 2005, **8**:1445-1449.
8. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bahler J, Wood V, *et al*: **The BioGRID Interaction Database: 2008 update.** *Nucleic Acids Res* 2008, **36**:D637-640.
9. Li CY, Yu Q, Ye ZQ, Sun Y, He Q, Li XM, Zhang W, Luo J, Gu X, Zheng X, Wei L: **A nonsynonymous SNP in human cytosolic sialidase in a small Asian population results in reduced enzyme activity: potential link with severe adverse reactions to oseltamivir.** *Cell Res* 2007, **17**:357-362.
10. Huang W, Payne TJ, Ma JZ, Li MD: **A functional polymorphism, rs6280, in DRD3 is significantly associated with nicotine dependence in European-American smokers.** *Am J Med Genet B Neuropsychiatr Genet* 2008, **147B**:1109-1115.

11. Zhou Z, Zhu G, Hariri AR, Enoch MA, Scott D, Sinha R, Virkkunen M, Mash DC, Lipsky RH, Hu XZ, *et al*: **Genetic variation in human NPY expression affects stress response and emotion.** *Nature* 2008, **452**:997-1001.

12. Jensen KP, Covault J, Conner TS, Tennen H, Kranzler HR, Furneaux HM: **A common polymorphism in serotonin receptor 1B mRNA moderates regulation by miR-96 and associates with aggressive human behaviors.** *Mol Psychiatry* 2008.

13. Feng J, Sun G, Yan J, Noltner K, Li W, Buzin CH, Longmate J, Heston LL, Rossi J, Sommer SS: **Evidence for X-chromosomal schizophrenia associated with microRNA alterations.** *PLoS One* 2009, **4**:e6121.

14. Sethupathy P, Collins FS: **MicroRNA target site polymorphisms and human disease.** *Trends Genet* 2008.

15. Sun G, Yan J, Noltner K, Feng J, Li H, Sarkis DA, Sommer SS, Rossi JJ: **SNPs in human miRNA genes affect biogenesis and function.** *Rna* 2009, **15**:1640-1651.

16. Hishimoto A, Liu QR, Drgon T, Pletnikova O, Walther D, Zhu XG, Troncoso JC, Uhl GR: **Neurexin 3 polymorphisms are associated with alcohol dependence and altered expression of specific isoforms.** *Hum Mol Genet* 2007, **16**:2880-2891.

17. Dreses-Werringloer U, Lambert JC, Vingtdeux V, Zhao H, Vais H, Siebert A, Jain A, Koppel J, Rovelet-Lecrux A, Hannequin D, *et al*: **A polymorphism in CALHM1 influences Ca2+ homeostasis, Abeta levels, and Alzheimer's disease risk.** *Cell* 2008, **133**:1149-1161.

18. Editorials: **An unnecessary battle.** *Nature* 2008, **454**:137-138.

19. Arguello PA: **Mental health: drop ideological baggage in favour of best tools.** *Nature* 2008, **454**:824.

20. Abbott A: **Psychiatric genetics: The brains of the family.** *Nature* 2008, **454**:154-157.

21. Li CY, Mao X, Wei L: **Genes and (common) pathways underlying drug addiction.** *PLoS Comput Biol* 2008, **4**:e2.

22. DerSimonian R, Laird N: **Meta-analysis in clinical trials.** *Control Clin Trials* 1986, **7**:177-188.

23. Fleiss JL: **The statistical basis of meta-analysis.** *Stat Methods Med Res* 1993, **2**:121-145.

24. Ioannidis JP, Gwinn M, Little J, Higgins JP, Bernstein JL, Boffetta P, Bondy M, Bray MS, Brenchley PE, Buffler PA, *et al*: **A road map for efficient and reliable human genome epidemiology.** *Nat Genet* 2006, **38**:3-5.

25. Liu QR, Drgon T, Johnson C, Walther D, Hess J, Uhl GR: **Addiction molecular genetics: 639,401 SNP whole genome association identifies many "cell adhesion" genes.** *Am J Med Genet B Neuropsychiatr Genet* 2006, **141B**:918-925.

26. Uhl GR, Drgon T, Liu QR, Johnson C, Walther D, Komiyama T, Harano M, Sekine Y, Inada T, Ozaki N, *et al*: **Genome-wide association for methamphetamine dependence: convergent results from 2 samples.** *Arch Gen Psychiatry* 2008, **65**:345-355.

27. Johnson C, Drgon T, Liu QR, Walther D, Edenberg H, Rice J, Foroud T, Uhl GR: **Pooled association genome scanning for alcohol dependence using 104,268 SNPs: validation and use to identify alcoholism vulnerability loci in unrelated individuals from the collaborative study on the genetics of alcoholism.** *Am J Med Genet B Neuropsychiatr Genet* 2006, **141B**:844-853.

28. Drgon T, Montoya I, Johnson C, Liu QR, Walther D, Hamer D, Uhl GR: **Genome-wide association for nicotine dependence and smoking cessation success in NIH research volunteers.** *Mol Med* 2009, **15**:21-27.

29. Uhl GR, Liu QR, Drgon T, Johnson C, Walther D, Rose JE: **Molecular genetics of nicotine dependence and abstinence: whole genome association using 520,000 SNPs.** *BMC Genet* 2007, **8**:10.

30. Uhl GR, Liu QR, Drgon T, Johnson C, Walther D, Rose JE, David SP, Niaura R, Lerman C: **Molecular genetics of successful smoking cessation: convergent genome-wide association study results.** *Arch Gen Psychiatry* 2008, **65**:683-693.

31. Drgon T, Zhang PW, Johnson C, Walther D, Hess J, Nino M, Uhl GR: **Genome wide association for addiction: replicated results and comparisons of two analytic approaches.** *PLoS One* 2010, **5**:e8832.

32. Kreek MJ, Bart G, Lilly C, LaForge KS, Nielsen DA: **Pharmacogenetics and human molecular genetics of opiate and cocaine addictions and their treatments.** *Pharmacol Rev* 2005, **57**:1-26.

33. 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**:851-861.

34. Voight BF, Kudaravalli S, Wen X, Pritchard JK: **A map of recent positive selection in the human genome.** *PLoS Biol* 2006, **4**:e72.

35. Mirnics K, Levitt P, Lewis DA: **Critical appraisal of DNA microarrays in psychiatric genomics.** *Biol Psychiatry* 2006, **60**:163-176.

36. Mao X, Cai T, Luo J, Wei L: **KOBAS server: a web-based platform for automated annotation and pathway identification.** 2008.

37. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, *et al*: **Targeted capture and massively parallel sequencing of 12 human exomes.** *Nature* 2009, **461**:272-276.

38. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.

39. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M, *et al*: **Analysis of genetic inheritance in a family quartet by whole-genome sequencing.** *Science* 2010, **328**:636-639.

40. Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler DA, *et al*: **Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy.** *N Engl J Med* 2010, **362**:1181-1191.

41. Van Vlierberghe P, Palomero T, Khiabanian H, Van der Meulen J, Castillo M, Van Roy N, De Moerloose B, Philippe J, Gonzalez-Garcia S, Toribio ML, *et al*: **PHF6 mutations in T-cell acute lymphoblastic leukemia.** *Nat Genet* 2010, **42**:338-342.

42. Cirulli ET, Goldstein DB: **Uncovering the roles of rare variants in common disease through whole-genome sequencing.** *Nat Rev Genet* 2010, **11**:415-425.

43. McClellan J, King MC: **Genetic heterogeneity in human disease.** *Cell* 2010, **141**:210-217.

44. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, *et al*: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461**:747-753.

45. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB: **Rare variants create synthetic genome-wide associations.** *PLoS Biol* 2010, **8**:e1000294.

46. Ioannidis Jp, Fau-Boffetta P, Boffetta P, Fau-Little J, Little J, Fau-O'Brien TR, O'Brien Tr, Fau-Uitterlinden AG, Uitterlinden Ag, Fau-Vineis P, Vineis P, Fau-Balding DJ, Balding Dj, Fau-Chokkalingam A, Chokkalingam A, Fau-Dolan SM, Dolan Sm, Fau-Flanders WD, Flanders Wd, Fau-Higgins JPT, *et al*: **Assessment of cumulative evidence on genetic associations: interim guidelines.** 2008.

47. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, Barrette TR, Ghosh D, Chinnaiyan AM: **Mining for regulatory programs in the cancer transcriptome.** *Nat Genet* 2005, **37**:579-583.

48. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, *et al*: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2008, **36**:D13-21.

49. Karolchik D, Kuhn RM, Baertsch R, Barber GP, Clawson H, Diekhans M, Giardine B, Harte RA, Hinrichs AS, Hsu F, *et al*: **The UCSC Genome Browser Database: 2008 update.** *Nucleic Acids Res* 2008, **36**:D773-779.

50. Wingender E, Dietze P, Karas H, Knuppel R: **TRANSFAC: a database on transcription factors and their DNA binding sites.10.1093/nar/24.1.238.** *Nucl Acids Res* 1996, **24**:238-241.

51. Shahi P, Loukianiouk S, Bohne-Lang A, Kenzelmann M, Kuffer S, Maertens S, Eils R, Grone HJ, Gretz N, Brors B: **Argonaute–a database for gene regulation by mammalian microRNAs.** *Nucleic Acids Res* 2006, **34**:D115-118.

52. Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N: **Combinatorial microRNA target predictions.** *Nat Genet* 2005, **37**:495-500.

53. Duan S, Huang RS, Zhang W, Bleibel WK, Roe CA, Clark TA, Chen TX, Schweitzer AC, Blume JE, Cox NJ, Dolan ME: **Genetic architecture of transcript-level variation in humans.** *Am J Hum Genet* 2008, **82**:1101-1113.

54. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, Nath P, *et al*: **A survey of genetic human cortical gene expression.** *Nat Genet* 2007, **39**:1494-1499.

55. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, *et al*: **Relative impact of nucleotide and copy number variation on gene expression phenotypes.** *Science* 2007, **315**:848-853.

56. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, *et al*: **Population genomics of human gene expression.** *Nat Genet* 2007, **39**:1217-1224.

57. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, Orchard S, Vingron M, Roechert B, Roepstorff P, Valencia A, *et al*: **IntAct: an open source molecular interaction database.** *Nucleic Acids Res* 2004, **32**: D452-455.

58. Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, Betel D, Bobechko B, Boutilier K, Burgess E, *et al*: **The Biomolecular Interaction Network Database and related tools 2005 update.** *Nucleic Acids Res* 2005, **33**:D418-424.

59. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, Bala P, Shivakumar K, Anuradha N, Reddy R, Raghavan TM, *et al*: **Human protein reference database–2006 update.** *Nucleic Acids Res* 2006, **34**:D411-414.

60. von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B: **STRING: a database of predicted functional associations between proteins.** *Nucleic Acids Res* 2003, **31**:258-261.

61. Mao X, Cai T, Olyarchuk JG, Wei L: **Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary.** *Bioinformatics* 2005, **21**:3787-3793.

62. Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4**:P3.

63. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25**:25-29.

64. Mulder N, Apweiler R: **InterPro and InterProScan: Tools for Protein Sequence Classification and Comparison.** *Methods Mol Biol* 2007, **396**:59-70.