Genome-Wide Association (GWA) Studies

Tun-Hsiang Yang, Mark Kon, Charles DeLisi

ABSTRACT

A host of data on genetic variation from the Human Genome and International HapMap projects, and advances in high-throughput genotyping technologies, have made genome-wide association (GWA) studies technically feasible. GWA studies help in the discovery and quantification of the genetic components of disease risks, many of which have not been unveiled before, and have opened a new avenue to understanding disease treatment, and prevention.

This chapter presents an overview of genome-wide association (GWA), an important tool for discovering regions of the genome that harbor common genetic variants to confer susceptibility for various diseases or health outcomes in the post-Human Genome Project era. A tutorial on how to conduct a GWA study and some practical challenges specifically related to the GWA design is presented, followed by a detailed GWA case study involving the identification of loci associated with glioma as an example and an illustration of current technologies.

I. INTRODUCTION

A significant scientific breakthrough in genomic research has been made in the first decade of the new millennium. The draft completion of the Human Genome Project

[1, 2] in 2001 is a major milestone in human genomics and biomedical sciences. It mapped the three billion nucleotide bases that make up the human genetic code, providing the foundation for studying genetic variations in the human genome, and showed that the DNA sequences of any two people are about 99.9% identical. The International HapMap Project [3] (http://www.hapmap.org/) which was completed in 2005 is another scientific landmark in the genomic research. It provides a catalog of common genetic variants, predominantly single nucleotide polymorphisms (SNPs), occurring in humans within and across populations in the world, and identifies chromosomal regions where genetic variants are shared. It further deepens our understanding of the genetic architecture of the human genome. The linkage disequilibrium (LD) map of the human genome provided by the HapMap project creates a valuable and useful genome-wide database of patterns of human genetic variation and also promotes the breakthrough in large-scale and high-throughput genotyping technological developments..

During the past five years, the accumulating knowledge about the correlation structure and frequency of common variants in the human genome combined with rapid advances in array technology have made GWA studies technically feasible The number of published GWA studies at $p \leq 5E\text{-}08$ has doubled within a year from June 2009 to June 2010 (N=439 through 6/2009, and N=904 for 165 traits through 6/2010 (NHGRI GWA Catalog, www.genome.gov/GWAStudies). In contrast to hypothesis-driven candidate-gene association studies, which largely rely on the understanding known and suspected pathology in a given trait, GWA studies systematically investigate genetic variation across the genome without the constraints of *a priori* hypotheses, and allows for the possibility of discovering associations in previously unsuspected pathways or in chromosomal regions of as yet undetermined function. This approach provides a comprehensive and unbiased examination of the common genetic basis of various complex traits. GWA studies

have expanded our understanding of the complexity and diversity of genetic variations in the human genome, and have led to pivotal discoveries of new genetic loci for a host of common human disorders, including cancer, type 2 diabetes mellitus, and autoimmune diseases [4].

II. HOW ARE GENOME-WIDE ASSOCIATION STUDIES CONUCTED?

As in other genetic association studies (such as candidate gene studies), genome-wide association compares the allele/genotype frequencies between groups that in principle differ in a single well defined phenotype; e.g. with and without a particular disease, looking for markers that are statistically significant correlates of phenotype.

A. Association study designs

The principal goal is to minimize systematic bias and maximize power. Two fundamentally different designs are used: population-based designs that collect unrelated individuals (such as case-control or cohort studies) and family-based designs that use families (such as trio or pedigree studies), but case-control studies are most typically used in GWA studies. For common diseases, population-based studies generally have higher statistical power; in addition, in late-onset diseases/disorders such as Alzheimer's disease, parents and siblings may not be available. On the other hand, although family-based design is generally more time- and resource-consuming, it is robust against population stratification and population admixture, and significant findings always imply both linkage and association.

B. DNA sample collection and genotyping technology

After appropriate samples are recruited, DNA is drawn from each participant, usually by either blood draw or buccal swab. Each person's complete set of DNA is then purified from the blood or buccal cells, placed on tiny chips and scanned on automated laboratory machines. The genotyping machines quickly survey each participant's genome for a dense set of strategically selected markers of genetic variation, including either single nucleotide polymorphisms (SNPs) or copy number polymorphisms (CNPs), or both.

The popular commercially available genotyping arrays for GWA studies include Illumina arrays (such as Human Hap550, Human Hap650, Infinium HD BeadChips, etc) and Affymetrix arrays (such as Genome-wide Human SNP Array 5.0, SNP Array 6.0, Human Mapping 500K Array Set, etc). In terms of the design in general, Affymetrix chips use the “random” design, in which the SNPs on the platform are randomly selected from the genome, without specific reference to the LD patterns. In contrast, Illumina chips use the “tagging” design, where SNPs are explicitly chosen to serve as surrogates for common variants in the HapMap data. The current genotyping platforms can accommodate up to 1 million or even more markers per chip per person.

C. Genotyping quality control

A battery of genotyping quality control procedures should be performed and checked after genotyping is completed, including marker completion rate, marker concordance, deviations from Hardy-Weinberg Equilibrium (HWE), sample completion rate, minor allele frequency (MAF), heterozygosity, gender concordance, duplicate sample detection, relatedness check, self-reported ethnicity concordance, and Mendelian consistency for markers and samples (if it is family-based study). In the population-based design, unexpected population structure can cause potential

bias due to population stratification when there is confounding due to correlated differences in both allele frequencies and disease risks across unobserved sub-populations. GWA studies therefore typically adjust for multiple random, unlinked markers as a surrogate for genetic variation across subpopulation using EIGENSTRAT software [5] . A principal component-based analysis to detect and correct for population stratification and false positive results from ethnic mixtures. (http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm). In addition, linkage agglomerative clustering based on pairwise identity-by-state (IBS) distance followed by multi-dimensional scaling (MDS) implemented in the PLINK toolset [6] (http://pngu.mgh.harvard.edu/~purcell/plink/) can be used to identify clusters of samples with more homogeneous genetic backgrounds for subsequent association tests.

A Quantile-Quantile (QQ) plot that compares observed order statistics of *p*-values against the expected order statistics of *p*-values under the null hypothesis is also useful for visualizing and to summarizing both systematic bias and evidence for association. Early departures from the expected *p*-values usually suggests systematic bias, whereas late departures suggest true association signals.

D. Statistical analysis

Because the purpose of the GWA studies is to analyze associations between thousands or millions of genetic markers at the genome-wide level and a disease or trait of interest without an *a priori* hypotheses, the initial association analysis examines marker-disease associations on a marker-by-marker basis from those who pass the quality control filtering. Depending on the assumption of genetic mode of inheritance, researchers may choose either allelic test, dominant, recessive, or co-dominant genotypic test, or trend test. Association analysis can be performed by

several existing programming packages, such as PLINK [6] and EIGENSTRAT [5]. PLINK is a free, open-source specifically developed for GWA studies that allows large-scale analyses in a computationally efficient manner for both population-based and family-based designs. EIGENSTRAT uses principal component analysis to model ancestry differences between cases and controls. The resulting correction minimizes spurious associations while maximizing power to detect true associations.

Because a large number of tests are conducted in GWA studies, stringent significance thresholds are essential to rule out false positive results. Several hundred thousand tests require thresholds of $p = 10^{-07}$ to control experiment-wide type I error for all common variants and $p = 5x10^{-8}$ for all variants [7-9]. A comprehensive analysis beyond single-marker analyses in the GWA setting is not yet feasible because it can introduce a large number of additional tests. For example, a combinatorial scan for all 2-way interaction on 1 million SNPs is barely feasible. A restriction to only a small subset of the data based on a specific rationale or hypothesis is more desirable, unless solutions on high-dimension data reduction and optimization are developed. SNPs or markers that are identified from the GWAS results can be further assigned to pathway analysis or enrichment analysis which could potentially be very useful for prioritizing genes and pathways within a biological context, which can be done with computational tools and pathway databases [10].

E. Validation and replication

If certain genetic markers are found to be significantly more frequent or less frequent in cases than in controls, the variations are said to be "associated" with the disease. The associated genetic variations can serve as powerful pointers to the region of the human genome where the causal locus resides. However, the significantly associated

variants themselves may not always directly cause the disease. In fact, in most cases they may just "tag along" with the actual causal variants due to the LD correlation structure in the human genome. Deeper analysis of the associated regions by sequencing is the best way to identify (a set of possible) causal variant(s), and filtering the list of highly associated variants using biologic annotation, including sequence context or known function (eQTLs), or conducting further *in vitro* experiments for functionality.

In order to represent credible genotype-phenotype associations observed in a GWA study, replication of the results is especially critical. That means finding the same marker or a marker in perfect or high LD with the prior marker [11].
With the increasing number consortia of multiple GWA studies, meta-analysis of multiple genome-wide studies conducted by different investigative groups, in different populations, using different genotyping technologies and different study designs) becomes an emerging approach of replication of GWA studies in the context of gene discovery [11], as illustrated below. Meta-analysis can increase the sample size effectively by combining different studies, which is especially powerful and useful in genetic association research, particularly when most of the common genetic variants contributing to complex diseases have only small to modest effects (odds ratio < 1.5). While a single GWA may not have sufficient statistical power to detect small effects, secondary analyses using a meta-analysis framework that pools information across studies provides an inexpensive and efficient way to accumulate evidence, which can also provide additional power for discovery of new associations by combing association signals across GWA studies, even when the original raw data are unavailable. For example, additional genetic loci with BMI [12] and lipid traits [13, 14] have recently been discovered by meta-analysis.

III. AN EXAMPLE – GLIOMA GENOME-WIDE ASSOCIATION STUDY

A case study involving the identification of loci which are associated with glioma using the GWA approach is presented here as a working example. We follow closely the presentation in [38].

A. Study samples

To identify risk variants for glioma, we conducted a principal component-adjusted genome-wide association study. 226 glioma patients were collected from The Cancer Genome Atlas (TCGA) [15] SNP data. The TCGA data portal contains clinical information, genomic characterization data of the tumor genomes and provides a platform for researchers to search, download, and analyze data sets generated by TCGA. Genotypes were determined using the Illumina Human Hap550 Array. We eliminated all samples for which more than 5% of the single nucleotide polymorphisms (SNPs) are missing, and eliminated all SNPs that (i) are determined in fewer than 95% of the samples, (ii) have MAF less than 5%, or (iii) have a HWE p-value of less than 10^{-6}. The procedure is outlined in Figure 1.

In order to adjust the potential confounding effects by ethnicity-specific SNP frequencies, we further confined our study sample to European-Americans only, which is the group from which the majority of samples were obtained. Glioma patient samples were identified by a two-step screening: (a) self-reporting of ancestry, and (b) computationally assisted stratification. The latter was carried out using the EIGENSTRAT package [5]. After screening, 179 TCGA samples remained. Of these, we used for confirmation only the 92 that were released after

August 2009, since the majority of the earlier samples have been already included in the Adult Glioma Study (AGS) [16].

The comparison group included normal European-American blood samples (n = 1366), which was downloaded from the Illumina iControlDB (iControls), an online data repository of genotype and phenotype data from individuals that can be used as controls in association studies. After applying the quality control procedures described above, 1306 control samples remained.

B. Association analysis

Association analysis was performed with the EIGENSTRAT package, under the null hypothesis of "no association between the glioblastoma multiform (GBM) SNP genotype and the control SNP genotype" based on an additive inheritance model [5]. To set the significance threshold for p we required that the probability of 1 or more false positives be less than 0.05; in particular, $1-e^{-Np} \le 0.05$ or $p \le 0.93\times10^{-7} \approx 10^{-7}$ with N, the total number of SNPs examined, taken as 550,000. At this level, few if any SNPs will be detected for typical glioma population sizes. The alternative is to accept a less stringent p-value, and to eliminate false discoveries by seeking confirmation in an independent study.

C. Meta-analysis

Various versions of meta-analysis can be used to combine p-values from two independent studies. Because the AGS and TCGA datasets differ wiely in the number of samples, we assigned them different weights [37] In particular

$$p_{12} = P\left(Z > \frac{W_1 Z_1 + W_2 Z_2}{\sqrt{W_1^2 + W_2^2}}\right) ; \; Z \sim Normal(0,1) \qquad (1)$$

where the weights (W_i) are proportional to the square root of the "total number of individuals", $Z_i = F^{-1}(1 - p_i)$, and F^{-1} (.) is an inverse standard normal CDF. The false discovery rate is estimated as the fused probability multiplied by the total number of SNPs, which is 300,000.

The procedure for calculating fused *p*-values begins with lists of SNPs that have *p*-values of less than 0.001 in each population. We walk down this list, calculating a combined *p* value (eq 1) for each pair, and accept all SNPs for which the false discovery rate (FDR) is less than 0.05 (or equivalently $p_{12} = 0.05/300{,}000 = 1.7 \times 10^{-7}$; see Table 1). For our results, when *p* exceeds 0.001 in either population, p_{12} no longer meets the required threshold, and the walk stops.[1]

D. Genes in Linkage Disequilibrium with SNPs

We use the coefficient of determination, R^2, to identify genes in strong linkage disequilibrium with the SNPs that we identified in the meta-analysis. R^2 is calculated based on the correlation between gene expression level and SNP genotype. Genes with R^2 greater then 0.8 are considered to be in strong LD with the SNP.

E. Relative risk

[1] As a practical matter, the walk can be stopped at more stringent *p* values without changing the main conclusions. In particular stopping AGS at $p = 10^{-5}$, and TCGA at 10^{-3}, while requiring that p_{12} pass the genomic significance level (1.7×10^{-7}), loses only 2 SNPs (rs12021720 and rs2810424), neither of which adds new genomic regions.

As indicated below, analysis of TCGA and AGS identifies 12 significant SNPs, 7 of which are new. One of the implications of additional SNPs is that the number of associated genes that can be used to estimate relative glioma risk increases combinatorially. Consequently we can expect higher prognostic reliability for individuals possessing a combination of risk alleles, although at some loss of population coverage. We consider here all combinations of two and three SNPs, while constraining our choices to SNPs that are more than a megabase (Mb) apart, in order to minimize redundant (disequilibrated) information. Specifically, the 12 SNPs are divided into 5 groups based on location. Chromosome 1 has 5 SNPs clustered together within 1 Mb, and chromosome 9 has 4 SNPs within 1 Mb around genes *CDKN2A/2B*. The remaining 3 SNPs are located on chromosomes 3, 5, and 7.

If we rule out combinations including any pair of SNPs that are within a single chromosome, we find 50 SNP pairs, and 88 SNP triplets. Statistical analyses were implemented using R (v2.7) and PLINK (v1.07) [6]. Combinations with odds ratios greater than three, along with *p*-values, are shown in Table 2, which also shows that SNP combinations from chromosomes 1 and 9 are associated with the highest relative risk.

F. Identification of Associated Pathways and Genes

The standard method for identifying altered processes is a pathway enrichment analysis, which can be carried out using a single population [17]. In this case pathways would be identified by showing that the number of significant SNPs/genes that occur in a particular pathway is above chance expectation. The procedure that we describe here requires multiple populations. The assignment of a SNP/gene to a particular pathway from a single population meets a significance threshold which is

loose enough to allow multiple assignments from that population, but not stringent enough for an acceptable FDR. The FDR is brought down to an acceptable level, as described below, when both populations assign the same gene(s) to the same pathway.

The procedure is as follows: (1) identify SNPs having a *p*-value < 10^{-3} in either the populations; (2) identify genes that include these SNPs, and (3) assign the genes thus obtained to KEGG pathways [18]. The detailed procedure by which assignments are made is explained elsewhere [38]

G. Results

i. Significant SNP candidates

Using TCGA datasets, we validated 4 of the 13 SNPs inferred by Wrensch et al based on the Adult Glioma study (Table 1, boldface) [16,38]. SNPs rs7530361 and rs501700, both at 1p21.2, were reported for the first time.

Joint analysis of data, as reported in (our Bio direct paper), rather than sequential analysis of two or more populations can increase the power to detect genetic associations [19]. In particular using eq (1) as described in Methods, we found 12 SNPs (Table 1), confirmed by AGS and TCGA at an FDR <0.05, one of which was previously confirmed by Wrensch et al. [16] and Shete et al. [20]. Of the 11 remaining, 4 were reported by Shete et al.; the other 7 are reported for the first time. The 12 SNPs are distributed over five genomic regions: 5q15.33, 9q21.3, 1p21.2, 3q26.2 and 7p15.3. Two of these, 5q15.33 and 9q21.3, have been reported in previous studies [16, 20]. The 12 candidates are in strong linkage disequilibrium with 25 genes, 8 of which are previously known to be associated with cancer are indicated in boldface in Table 1. An additional SNP of interest is rs12341266 at

9q32, which has an FDR of 0.06 and is in the glioma associated gene, *RGS3*.

ii. Genes identified by conserved pathway analysis

We identified 49 pathways that contain genes associated with loosely defined AGS or TCGA SNPs. Thirty-six of them do not meet the hypergeometric test at a p value of 0.001 (an FDR of 0.05 divided by 49), leaving 13 invariant pathways; i.e. pathways that are relevant to both populations. Each of the 13 pathways has 1 common gene (Table 3) from the two groups. There are 5 such genes – *FHIT, GABRG3, PRKG1, DCC, ITGB8* – each of which occurs in more than one of these pathways.

iii. Genes in Strong Linkage Disequilibrium with SNP candidates

The SNP candidates occur within, or are in strong linkage disequilibrium with, 25 genes (Table 1). Eight of which are cancer associated. The latter are *TERT* [16, 21, 22], *SLC6A18* [21], *CLPTM1L* [21, 22], *CDKN2A/2B* [16, 23, 24]*, SASS6* [25], *ITGB8* [26], and *MACC1* [27] (Table 1)**.** Five of the genes, *TERT*, *SLC6A18, CLPTM1L,* and *CDKN2A/2B*, were previously shown to be associated with glioma by other GWA studies.

As explained below, we have predicted by a combination of GWA and pathway analysis, 4 additional genes, which are identified in the literature as cancer related. The detail literature citations and the type of cancers that associated with these genes are discussed in discussion section. We therefore predict 29 glioma associated genes, 12 of them known by previous studies to be cancer related. It is useful to ask for the probability that as many as12 cancer related genes in a set of 29 would be found by chance. If we use the fraction of OMIM genes that are cancer related as an

estimate of the background frequency of cancer genes in the disease genes population, the probability that 29 genes have 12 cancer associated genes by chance is 1.4E-06. The fraction of OMIM genes that are cancer related is 0.1 (750 cancer associated gene in 7,381 OMIM genes).

Each of the 8 cancer related genes listed above plays one or more key roles in processes known to be altered during tumor initiation and development [28]. For example, *MACC1* is a growth pathway regulator influencing angiogenesis and processes related to metastasis [27]; CDKN2A is a well studied cell cycle regulator [24] and a known tumor suppressor whose loss results in a diminished ability to regulate growth and predisposition to cancer [23]; ITGB8 has been implicated in activities related to metastasis, including adhesion and migration [26]; and the telomerase enzyme (TERT) is linked to unlimited replication [16]. It is worth noting that CDKN2A/2B are in strong linkage disequilibrium with rs1412829 at 9p21.3, which has now been identified in 3 independent studies and should therefore be considered an unusually high confidence gene marker.

IV. POTENTIAL CHALLENGES IN GENOME-WIDE ASSOCIATION STUDIES

While GWA studies open a new avenue of discovering and understanding of the common genetic variation of the human genome in diseases and health, the assessments of the overall evidence deriving from GWA studies remains a complex endeavor. The GWA studies also create some open challenges as the field is still under development and much of the literature remains exploratory.

A. **From statistics to functionality**

Although statistically compelling associations have been identified, many association signals identified in GWA studies are not localized to intervals that include a gene, unlike Mendelian human diseases whose genetics is understood that functional rare mutations with large effects act through altering or truncating gene products. However, there is growing evidence that a sizeable proportion, perhaps the majority, of the functional variants that underlie GWA studies exert their effects through gene regulation rather than changing gene products [29]. For example, a SNP (rs6983267) in the 8q24 locus implicated in multiple cancer pathogeneses identified in GWA studies is located in a gene desert that is >300 kilo-bases (kb) away from the most neighboring annotated *MYC* proto-oncogene; recently studies have shown that the region harboring this risk allele is a transcriptional enhancer that interacts with the *MYC* gene [30, 31]. How to translating mere statistically association signals to biological relevance of the precise variants that have a causal role in conferring the disease susceptibility remains unclear at present, but more research towards a deeper understanding of the vast regulatory regions within the human genome and functional studies will be the future direction.

B. Investigations of complex interactions

Given the fact that common complex diseases are multi-factorial with each factor contributing a small effect, it is possible that what really counts is not the main effects of the genes but complex gene-gene or gene-environment interactions. How to proceed with the investigations of gene-gene interactions or gene-environment interactions in GWA studies is an important question with no straightforward answers.

C. Sufficiency of common variants to account for genetic bases of complex traits

The current technology for GWAS studies consider common genetic variants, predominantly SNPs, as possible targets for association with a trait or a phenotype, and do not capture information about rare variants. However, not only SNPs, there are also others forms of genetic variations that could account for disease risk. For example, recently, genomic copy number variations (CNVs) have begun investigated in several GWA studies. CNVs are defined as gains or losses of repeats of DNA sequences consisting of between kilo- to mega-base pairs. CNVs have been detected in locations covering about 12% of the human genome [32, 33]. As technology and knowledge surrounding CNVs continue to improve, CNVs have become a significantly more mainstream in GWA studies [34, 35]. However, in addition to SNPs and CNVs, there are also other types of structural variations and epigenetics in the human genome and it is unclear how much each type of genetic variation contributes to inherited risk and the relative proportion of rare versus common variants. The use of new technologies for assaying DNA sequences can provide important and additional insights about the roles of different types of genetic variants in human disease or health. For example, the 1000 Genomes Project [36] launched in 2008 has used the next-generation sequencing technique to provide a comprehensive resource on human genetic variation with at least 1% across most of the genome and down to 0.5% or lower within genes. The 1000 Genomes Project will map not only the SNPs but also will produce a high-resolution map of structural variants, including rearrangements, deletions or duplications of segments of the human genome.

References

1. Collins FS, Morgan M, Patrinos A: **The Human Genome Project: lessons from large-scale biology**. *Science* 2003, **300**(5617):286-290.
2. Roberts L, Davenport RJ, Pennisi E, Marshall E: **A history of the Human Genome Project**. *Science* 2001, **291**(5507):1195.
3. 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.
4. 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.
5. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies**. *Nature genetics* 2006, **38**(8):904-909.
6. 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**. *American journal of human genetics* 2007, **81**(3):559-575.
7. Dudbridge F, Gusnanto A: **Estimation of significance thresholds for genomewide association scans**. *Genet Epidemiol* 2008, **32**(3):227-234.
8. Pe'er I, Yelensky R, Altshuler D, Daly MJ: **Estimation of the multiple testing burden for genomewide association studies of nearly all common variants**. *Genet Epidemiol* 2008, **32**(4):381-385.
9. Wakefield J: **A Bayesian measure of the probability of false discovery in genetic epidemiology studies**. *American journal of human genetics* 2007, **81**(2):208-227.
10. Cantor RM, Lange K, Sinsheimer JS: **Prioritizing GWAS results: A review of statistical methods and recommendations for their application**. *American journal of human genetics*, **86**(1):6-22.
11. Kraft P, Zeggini E, Ioannidis JP: **Replication in genome-wide association studies**. *Stat Sci* 2009, **24**(4):561-573.
12. Willer CJ, Speliotes EK, Loos RJ, Li S, Lindgren CM, Heid IM, Berndt SI, Elliott AL, Jackson AU, Lamina C *et al*: **Six new loci associated with body mass index highlight a neuronal influence on body weight regulation**. *Nature genetics* 2009, **41**(1):25-34.

13. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, Pramstaller PP, Penninx BW, Janssens AC, Wilson JF, Spector T *et al*: **Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts**. *Nature genetics* 2009, **41**(1):47-55.
14. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, Kaplan L, Bennett D, Li Y, Tanaka T *et al*: **Common variants at 30 loci contribute to polygenic dyslipidemia**. *Nature genetics* 2009, **41**(1):56-65.
15. **Comprehensive genomic characterization defines human glioblastoma genes and core pathways**. *Nature* 2008, **455**(7216):1061-1068.
16. Wrensch M, Jenkins RB, Chang JS, Yeh RF, Xiao Y, Decker PA, Ballman KV, Berger M, Buckner JC, Chang S *et al*: **Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility**. *Nature genetics* 2009, **41**(8):905-908.
17. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES *et al*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles**. *Proc Natl Acad Sci U S A* 2005, **102**(43):15545-15550.
18. Kanehisa M: **The KEGG database**. *Novartis Found Symp* 2002, **247**:91-101; discussion 101-103, 119-128, 244-152.
19. Skol AD, Scott LJ, Abecasis GR, Boehnke M: **Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies**. *Nature genetics* 2006, **38**(2):209-213.
20. Shete S, Hosking FJ, Robertson LB, Dobbins SE, Sanson M, Malmer B, Simon M, Marie Y, Boisselier B, Delattre JY *et al*: **Genome-wide association study identifies five susceptibility loci for glioma**. *Nature genetics* 2009, **41**(8):899-904.
21. Kang JU, Koo SH, Kwon KC, Park JW, Kim JM: **Gain at chromosomal region 5p15.33, containing TERT, is the most frequent genetic event in early stages of non-small cell lung cancer**. *Cancer Genet Cytogenet* 2008, **182**(1):1-11.
22. Stacey SN, Sulem P, Masson G, Gudjonsson SA, Thorleifsson G, Jakobsdottir M, Sigurdsson A, Gudbjartsson DF, Sigurgeirsson B, Benediktsdottir KR *et al*: **New common variants affecting susceptibility to basal cell carcinoma**. *Nature genetics* 2009, **41**(8):909-914.
23. Bisio A, Nasti S, Jordan JJ, Gargiulo S, Pastorino L, Provenzani A, Quattrone A, Queirolo P, Bianchi-Scarra G, Ghiorzo P *et al*: **Functional analysis of CDKN2A/p16INK4a 5'-UTR variants predisposing to melanoma**. *Hum Mol Genet*, **19**(8):1479-1491.
24. Sherr CJ: **Cancer cell cycles**. *Science* 1996, **274**(5293):1672-1677.

25. Leidel S, Delattre M, Cerutti L, Baumer K, Gonczy P: **SAS-6 defines a protein family required for centrosome duplication in C. elegans and in human cells**. *Nat Cell Biol* 2005, **7**(2):115-125.
26. Culhane AC, Quackenbush J: **Confounding effects in "A six-gene signature predicting breast cancer lung metastasis"**. *Cancer Res* 2009, **69**(18):7480-7485.
27. Boardman LA: **Overexpression of MACC1 leads to downstream activation of HGF/MET and potentiates metastasis and recurrence of colorectal cancer**. *Genome Med* 2009, **1**(4):36.
28. Hanahan D, Weinberg RA: **The hallmarks of cancer**. *Cell* 2000, **100**(1):57-70.
29. Ku CS, Loy EY, Pawitan Y, Chia KS: **The pursuit of genome-wide association studies: where are we now?** *Journal of human genetics*, **55**(4):195-206.
30. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, Beckwith CA, Chan JA, Hills A, Davis M *et al*: **The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer**. *Nature genetics* 2009, **41**(8):882-884.
31. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, Kivioja T, Bjorklund M, Wei G, Yan J, Niittymaki I *et al*: **The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling**. *Nature genetics* 2009, **41**(8):885-890.
32. Perry GH, Ben-Dor A, Tsalenko A, Sampas N, Rodriguez-Revenga L, Tran CW, Scheffer A, Steinfeld I, Tsang P, Yamada NA *et al*: **The fine-scale and complex architecture of human copy-number variation**. *American journal of human genetics* 2008, **82**(3):685-695.
33. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W *et al*: **Global variation in copy number in the human genome**. *Nature* 2006, **444**(7118):444-454.
34. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PI, Maller JB, Kirby A *et al*: **Integrated detection and population-genetic analysis of SNPs and copy number variation**. *Nature genetics* 2008, **40**(10):1166-1174.
35. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K *et al*: **Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs**. *Nature genetics* 2008, **40**(10):1253-1260.
36. Kuehn BM: **1000 Genomes Project promises closer look at variation in human genome**. *Jama* 2008, **300**(23):2715.
37. Liptak, T. (1958). **On the combination of independent tests**. *Magyar Tud. Akad. Mat. Kutato Int. Kozl.* **3**: 171–197

38. Yang TH, Kon M, Hung JH, DeLisi C. **Combinations of Newly Confirmed Glioma-Associated Loci Link Regions on Chromosomes 1 and 9 to Increased Disease Risk. Submitted for publication. (**http://www.biology-direct.com/imedia/2001713445537683_article.pdf)

Figure 1. **Subjects and single-SNP exclusion schema for genome-wide association studies.**

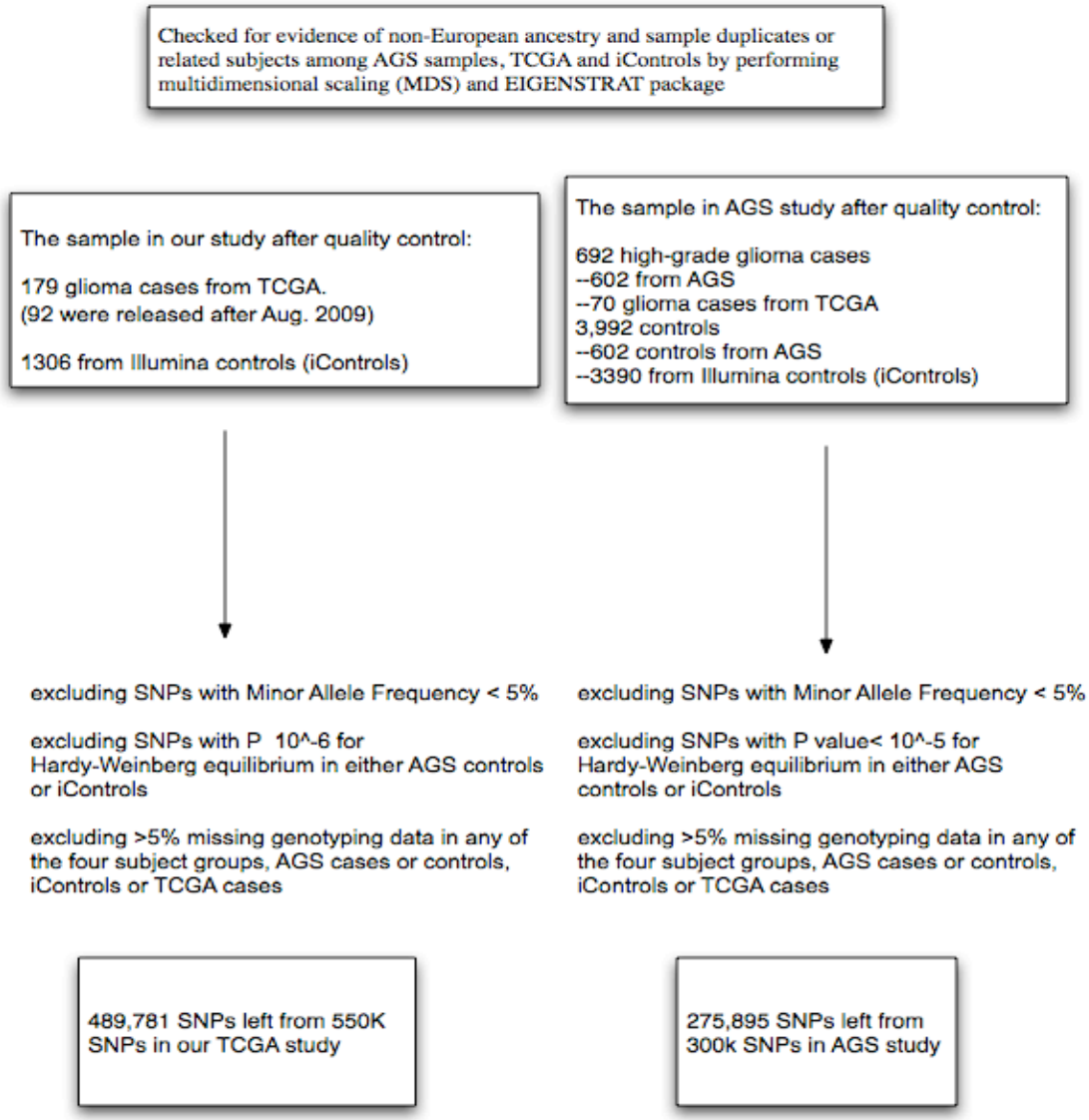

Table 1. **Concordant SNPs recovered from TCGA and AGS data, and associated genes.** Concordant SNPs (FDR < 0.05) recovered from TCGA (*n* = 97) and AGS data (*n* = 692), and associated genes.

SNP	chr	gene	location	left_gene	right_gene	Genes in LD with SNP (R^2)	AGS (P1)	TCGA (P2)	FDR
rs2736100+	**5**	**TERT**	5p15.33	**SLC6A18**	**CLPTM1L**	NA	**5.30E-13**	**2.66E-04**	**7.38E-09**
rs1412829*+	**9**	**CDKN2A/2B**	9p21.3	LOC100130239	LOC729983	**CDKN2A**(1.0); **CDKN2B**(1.0); C9orf53(0.87)	**3.40E-08**	**3.26E-03**	**1.27E-03**
rs2157719	9	CDKN2A/2B	9p21.3	LOC100130239	LOC729983	**CDKN2A**(0.97); **CDKN2B**(0.97); C9orf53(0.93); RP11-145E5.4(0.97); LOC100130239(0.97)	6.10E-08	8.00E-03	5.40E-03
rs1063192+	9	CDKN2A/2B	9p21.3	**CDKN2A**	LOC100130239	**CDKN2A**(0.97); **CDKN2B**(0.97); C9orf53(0.93); RP11-145E5.4(0.97); LOC100130239(0.97)	9.20E-08	8.31E-03	7.95E-03
rs4977756+	9	CDKN2A/2B	9p21.3	LOC100130239	LOC729983	**CDKN2A**(0.97); **CDKN2B**(0.97); C9orf53(0.93); RP11-145E5.4(0.97); LOC100130239(0.97)	4.20E-07	1.12E-02	3.90E-02
rs7530361	**1**	**SLC35A3**	1p21.2	LOC730081	HIAT1	SLC35A3(1.0); CCDC76(0.95); HIAT1(1.0); LRRC39(0.95); **SASS6**(1.0); BRI3P1(0.95); LOC730081(1.0)	**6.50E-07**	**2.19E-06**	**4.29E-05**
rs501700	**1**	**HIAT1**	1p21.2	SLC35A3	**SASS6**	DBT(0.95); RTCD1(0.89); SLC35A3(1.0); CCDC76(0.95); HIAT1(1.0); LRRC39(0.94); **SASS6**(1.0); BRI3P1(0.94); LOC730081(1.0)	**7.10E-07**	**5.99E-06**	**9.72E-05**
rs1920116	3	LRRC31	3q26.2	LRRIQ4	KRT18P43	MYNN(0.89); LRRC31(1); ARPM1(0.85); KRT18P43(1) LRRC34(1)	1.40E-06	2.88E-03	2.81E-02
rs506044	1	SASS6	1p21.2	**SASS6**	LRRC39	DBT(1.0); RTCD1(0.89); SLC35A3(0.95); CCDC76(1.0); HIAT1(1.0); LRRC39(1.0); **SASS6**(1.0); BRI3P1(0.95); LOC730081(0.94)	2.10E-06	2.45E-06	1.57E-04
rs640030	1	SASS6	1p21.2	HIAT1	CCDC76	DBT(1.0); RTCD1(0.89); SLC35A3(0.95); CCDC76(1.0); HIAT1(1.0); LRRC39(1.0); **SASS6**(1.0); BRI3P1(0.95); LOC730081(0.94)	2.40E-06	2.57E-06	1.86E-04
rs687513	1	SASS6	1p21.2	**SASS6**	LRRC39	DBT(0.95); RTCD1(0.90); SLC35A3(0.94); CCDC76(1.0); HIAT1(1.0); LRRC39(1.0); **SASS6**(1.0); BRI3P1(0.95); LOC730081(0.90)	2.90E-06	3.91E-06	3.03E-04
rs3779505	7	**ITGB8**	7p15.3	**MACC1**	LOC100130234	**ITGB8**(1.0)	3.00E-06	5.67E-04	1.35E-02

\+ reported by Shete et, al.
* reported by Wrensch et, al. and validated on Mayo Clinic population

Table 2. **Pairwise and triplet SNP combinations with odds ratios greater than 3.** Numbers in parenthesis are single SNP odds ratios. Last column is the Wald test *p*-value for the odds ratio of the combination. This is an unadjusted *p*-value, with an 0.05 multiple testing adjusted threshold of $p = 0.05/(50+88) = 3.6 \times 10^{-4}$. Freq denotes the combined frequency of the given combination in the case and control populations as a whole.

SNP Combinations			+RISK ALLELE	Freq	OR[eq2]	*p*-value
*rs1412829 (1.58)		#rs7530361 (1.89)	11	5.45E-02	3.31	3.58E-07
*rs1412829 (1.58)		#rs501700 (1.90)	11	5.51E-02	3.09	1.95E-06
*rs1412829 (1.58)		#rs506044 (1.96)	11	5.47E-02	3.23	5.15E-07
*rs1412829 (1.58)		#rs640030 (1.95)	11	5.42E-02	3.28	4.30E-07
*rs1412829 (1.58)		#rs687513 (1.93)	11	5.52E-02	3.18	7.32E-07
*rs2157719 (1.49)		#rs7530361 (1.89)	11	5.51E-02	3.2	6.83E-07
*rs2157719 (1.49)		#rs506044 (1.96)	11	5.54E-02	3.12	9.64E-07
*rs2157719 (1.49)		#rs640030 (1.95)	11	5.49E-02	3.16	8.12E-07
*rs2157719 (1.49)		#rs687513 (1.93)	11	5.59E-02	3.07	1.35E-06
*rs1063192 (1.42)		#rs7530361 (1.89)	11	5.60E-02	3.12	1.13E-06
*rs1063192 (1.42)		#rs506044 (1.96)	11	5.63E-02	3.05	1.60E-06
*rs1063192 (1.42)		#rs640030 (1.95)	11	5.59E-02	3.08	1.35E-06
*rs4977756 (1.60)		#rs7530361 (1.89)	11	5.35E-02	4.28	3.14E-10
*rs4977756 (1.60)		#rs501700 (1.90)	11	5.44E-02	4.17	5.57E-10
*rs4977756 (1.60)		#rs506044 (1.96)	11	5.36E-02	4.18	4.46E-10
*rs4977756 (1.60)		#rs640030 (1.95)	11	5.31E-02	4.24	3.66E-10
*rs4977756 (1.60)		#rs687513 (1.93)	11	5.41E-02	4.1	6.86E-10
rs2736100 (0.63)	#rs7530361 (1.89)	rs1920116 (0.68)	212	5.01E-02	4.3	5.02E-10
rs11823971 (1.45)	*rs1412829 (1.58)	#rs7530361 (1.89)	211	5.21E-02	3.04	5.05E-06
rs11823971 (1.45)	*rs1412829 (1.58)	#rs506044 (1.96)	211	5.26E-02	3.01	4.67E-06

+ Denotes alleles in which significant shifts occur. 11 denotes significant shift in the minor alleles for both SNPs. 212 denotes significant shifts in major, minor major; 211, significant shifts in major, minor, minor.

# denotes SNP on chromosome 1

* denotes SNP on chromosome 9 in gene CDKN2A/2B

Table 3. Pathways that contain significant SNPs ($p < 10^{-3}$) inferred from both AGS and TCGA samples

PATHWAY*	AGS_SNP	GENE	TCGA_SNP	GENE
Purine metabolism (p=3.50E-04)**				
Small cell lung cancer (p=4.35E-04) **				
Non-small cell lung cancer (p=2.6E-04) **	rs7617530	FHIT	rs13059601	FHIT
Neuroactive ligand-receptor interaction (p=8.00E-04) **	rs1011455	GABRG3		
	rs4887546	GABRG3		
	rs1011456	GABRG3	rs12904325	GABRG3
Vascular smooth muscle contraction (p=3.48E-04) **	rs4400745	PRKG1		
Gap junction (p=1.30E-04) **				
Long-term depression (p=6.95E-04) **				
Olfactory transduction (p=3.47E-04) **	rs4466778	PRKG1	rs1922139	PRKG1
Axon guidance (p=3.91E-04) **			rs11082983	DCC
Pathways in cancer (p=2.13E-03)			rs11872471	DCC
Colorectal cancer (p=8.69E-05) **	rs1145245	DCC	rs12604940	DCC
Focal adhesion (p=1.95E-03)	rs3779505	ITGB8		
ECM-receptor interaction (p=8.69E-04) **	rs2301727	ITGB8		
Cell adhesion molecules (CAMs) (p=1.74E-04) **	rs3807936	ITGB8		
Regulation of actin cytoskeleton (p=1.56E-03)				
Hypertrophic cardiomyopathy (HCM) (p=1.22E-03)				
Arrhythmogenic right ventricular cardiomyopathy (ARVC) (p=9.12E-04) **				
Dilated cardiomyopathy (p=1.04E-03)	rs2158250	ITGB8	rs3779505	ITGB8

* p = Probability of the gene overlap in two independent populations. Multiple testing adjusted threshold of $p = .05/49 = 10^{-3}$

** Pathways with $p < 10^{-3}$