Statistical Applications in Genetics and Molecular Biology

Manuscript 1784

GENOVA: Gene Overlap Analysis of GWAS Results

Clara S. Tang, *Queensland Institute of Medical Research*
Manuel AR Ferreira, *Queensland Institute of Medical Research*

GENOVA: Gene Overlap Analysis of GWAS Results

Clara S. Tang and Manuel AR Ferreira

Abstract

In many published genome-wide association studies (GWAS), the top few strongly associated variants are often located in or near known genes. This observation raises the more general hypothesis that variants nominally associated with a phenotype are more likely to overlap genes than those not associated with a phenotype. We developed a simple approach – named GENe OVerlap Analysis (GENOVA) – to formally test this hypothesis. This approach includes two steps. First, we define largely independent groups of highly correlated SNPs (or "clumps") and classify each clump as intersecting a gene or not. Second, we determine how strongly associated each clump is with the phenotype and use logistic regression to formally test the hypothesis that clumps associated with the phenotype are more likely to intersect genes. Simulations suggest that the power of GENOVA is affected by at least three factors: GWAS sample size, the gene boundaries used to define gene-intersecting clumps and the *P*-value threshold used to define phenotype-associated clumps. We applied GENOVA to results from three recent GWAS meta-analyses of height, body mass index (BMI) and waist-hip ratio (WHR) conducted by the GIANT consortium. SNPs associated with variation in height were 1.44-fold more likely to be in or near genes than SNPs not associated with height ($P = 5x10^{-28}$). A weaker association was observed for BMI (1.09-fold, $P = 0.008$) and WHR (1.09-fold, $P = 0.014$). GENOVA is implemented in C++ and is freely available at https://genepi.qimr.edu.au/staff/manuelF/genova/main.html.

KEYWORDS: gene, enrichment, annotation, method

Author Notes: We thank Scott Gordon, Dixie Statham, Ann Eldridge, Marlene Grace, Lisa Bowdler, Steven Crooks, David Smyth, Harry Beeby, Anjali K. Henders, Dale R. Nyholt, Peter M. Visscher, Grant W. Montgomery, and Nicholas G. Martin, who were part of the original team that ascertained, genotyped, and processed SNP data for the twins and families who participated in the various QIMR cohorts. This work was funded by the Australian National Health and Medical Research Council (grants 613627 and 613679).

1. Introduction

In a recent genome-wide association study (GWAS) of platelet counts (PLT), we observed that single nucleotide polymorphisms (SNPs) with genome-wide significant association with PLT were often located within, or in close proximity to, known genes (Gieger, 2011). This informal observation, which has become increasingly more common as larger GWAS are carried out, raised the more general hypothesis that SNPs nominally associated with PLT were more likely to overlap genes than those not associated with PLT. Because, to our knowledge, there were no available methods at the time to formally test this hypothesis, we developed in that study a simple approach to test whether phenotype-associated SNPs are more often located in or near gene regions than those not associated with the phenotype. However, our approach was only summarised in that study and we did not describe its performance under different genetic models.

In this study, we (1) describe our GENe OVerlap Analysis (GENOVA) test in greater detail; (b) report a series of simulations conducted to assess its type-I error and power; and (c) apply GENOVA to the analysis of publicly available results from GWAS of height (Lango-Allen, 2010), body mass index (Speliotes, 2010) and waist-hip ratio (Heid, 2010).

2. Methods

2.1. GENOVA approach

We developed GENOVA to formally test the hypothesis that SNPs associated with a trait of interest are more likely to overlap genes than SNPs not associated with the trait. Our approach includes two steps, which are described below. Step 1 is performed to generate one of the two input files required to run the gene overlap test; this file is included in the GENOVA release and so most users can skip this step. Step 2 includes the actual gene overlap test.

Step 1: define largely independent groups of highly correlated SNPs (or "clumps"). The goal of step 1 is two-fold: (A) to identify and group SNPs that are in high linkage disequilibrium (LD) with each other but in low LD with SNPs that belong to other groups - we refer to these groups of SNPs as clumps; and (B) to classify each clump as intersecting a gene or not.

To identify clumps of SNPs (Step 1A), we downloaded the August 2010 version of the 1000 Genomes Project data for the European population, including 280 founders and 11.9 million SNPs (The 1000 Genomes Project Consortium, 2010). We restricted our analysis to 7.8 million autosomal SNPs with a minor allele frequency (MAF) > 1%. We then grouped SNPs located within 1,000 kb of

each other into individual clumps based solely on pair-wise LD. This was performed by modifying the --clump routine implemented in PLINK (Purcell, 2007) to use two r^2 thresholds (rather than a single threshold): an exclusion (0.1) and an inclusion (0.8) threshold. For example, starting with the first SNP on chromosome 1 (index SNP for the first clump), we (i) identified nearby SNPs with $r^2 \geq 0.8$ with that index SNP and added those to the same LD clump; and (ii) identified nearby SNPs with an r^2 between 0.1 and 0.8 with that first index SNP and excluded them from subsequent analyses. The remaining nearby SNPs had an $r^2 < 0.1$ with the index SNP of that first clump. We then moved on to the first of those remaining SNPs, which became the index SNP of the second LD group, and repeated (i) and (ii) above. This procedure was repeated until all 7.8 million SNPs had been either grouped (1.7 million) or excluded from subsequent analyses. As a result, we created 487,293 largely independent groups of highly correlated SNPs, with a mean number of SNPs per clump of 3.5 (range 1 to 1584).

To determine whether a given clump intersected a gene or not (Step 1B), we recorded the lowest and highest genomic coordinate for the clump, which corresponded to the physical position of the first and last SNP in the group. The average segment length across the 487,293 clumps was 7 kb (range 0 to 1386 kb). Then, for each clump, we determined whether the corresponding chromosomal segment overlapped or intersected to any extent (i.e. at least 1 bp) the sequence of a known gene, using the first and last exonic base of the longest isoform +/- 50 kb as the gene boundaries (based on the February 2009 UCSC assembly, GRCh37/hg19). Of 487,293 clumps, 299,776 (62%) intersected at least one gene. In the simulation study described below, we assess the impact of using a stricter (e.g. +/- 0 kb) or more liberal (e.g. +/- 100 kb) gene boundary definition to determine whether a clump intersected a known gene. A text file that lists all 487,293 clumps is included with the GENOVA release and is one of two input files required to carry out the gene overlap test.

Step 2: gene overlap test. This is the main analytical step: the goal is to test whether clumps associated with the trait are more likely to intersect genes than those that do not. Two input files are required: the first is the clump file described above, which lists clumps of SNPs defined based solely on LD data and indicates whether a clump intersects a gene or not. The second file must contain genome-wide association results, one row per SNP, including SNP name and trait association *P*-value. Based on the SNP results provided, GENOVA will then (A) determine how strongly associated each clump is with the trait and (B) formally test the hypothesis that clumps associated with the trait are more likely to intersect genes.

To summarise how strongly associated with the trait the clumps are (Step 2A), we assign to each clump a single trait association *P*-value, specifically the *P*-value for the SNP showing the most significant association amongst all SNPs tested in that clump. Because SNPs within a clump are in high LD with each other, association *P*-values are expected to be very similar across SNPs that belong to the same clump. We then classify each clump based on the *P*-value as being associated with the trait ($P < 0.05$) or not; this information is stored as a binary variable. In the simulation study described below, we assess the impact on power of using a stricter (e.g. 0.01 or 0.005) *P*-value threshold to determine if clumps are associated with the trait.

Lastly, we use logistic regression to test the hypothesis that trait-association level (i.e. associated or not associated) is a significant predictor of whether a clump intersects a gene or not (Step 2B). We include in the regression model potential confounders, namely the minor allele frequency and imputation confidence metric for the index SNP of the clump, clump length and number of SNPs in the clump.

2.2. Implementation

A free C++ implementation of GENOVA, which includes steps 1B, 2A and 2B above, is available at https://genepi.qimr.edu.au/staff/manuelF/genova/main.html. Step 1A can be carried out in PLINK, although most users will not need to perform this step (nor 1B), as summarised above.

2.3. Simulation study

We performed a number of simulations to investigate the type-I error rate and power of GENOVA. In these simulations, we used real genotype data obtained with the Illumina 610 K array for 3,000 unrelated individuals genotyped as part of four QIMR projects described in detail elsewhere (Medland, 2009). We restricted our analysis to 38,490 SNPs with a MAF > 1% and located on chromosome 1, as it would be prohibitively slow to perform extensive simulations based on the full genome-wide data. These SNPs mapped to 5,269 clumps, of which 3,815 (72%) intersected a gene and 1,454 (28%) did not.

A single quantitative phenotype was simulated for the 3,000 individuals under five different genetic models, which differed in the degree of association between gene-intersecting and trait-associated clumps.

In model 1, no single SNP contributed to inter-individual variation in the simulated phenotype. This represents the null model that was used to assess type-I error rate: because no single SNP contributed to the phenotype, clumps that by chance are associated with the phenotype are expected to intersect genes at the same rate as clumps that are not associated with the phenotype. Specifically, we

expect ~191 (3,815 x 0.05) clumps that intersect a gene to be significantly associated with the simulated phenotype at $P < 0.05$ and, similarly, ~73 (1,454 x 0.05) clumps that do not intersect a gene to be associated ($P < 0.05$) with the phenotype. We simulated 10,000 datasets under model 1. In this situation, GENOVA should yield a significant result only as expected by chance.

In models 2 to 5, we progressively increased the number of clumps that contributed to the phenotype from 36 (model 2) to 54 (model 3), 70 (model 4) and 88 (model 5). In each model, each causal SNP (one per clump) explained 0.3% of the phenotypic variance; for example, in model 2, 10.8% of the variance of the simulated phenotype was explained by the 36 clumps. Importantly, for a clump to be selected to contribute to the phenotype (i.e. to be causal), it had to intersect a gene (i.e. located in or within 50 kb of the gene). In this way, we specifically manipulated the degree of association between gene-intersecting and trait-associated clumps. For example, in model 2, by selecting 36 clumps that intersected a gene as causal, we ensured that clumps significantly associated with the phenotype at a $P < 0.05$ level were 1.2-fold more likely to intersect a gene than clumps not significantly associated with the phenotype. Similarly, for models 3 to 5, this figure was 1.3-fold, 1.4-fold and 1.5-fold, respectively. We simulated 1,000 datasets for each model. Under these models, GENOVA is expected to provide progressively increased power to detect a significant overlap between trait-associated clumps and gene-intersecting clumps.

We also assessed the impact on power of using stricter or more liberal gene boundary definitions when determining whether a clump intersected a gene or not (Step 1B above). Specifically, we investigated whether power was affected by expanding known gene boundaries by 0, 25, 50 or 100 kb when considering four different simulation models which differed in how close the simulated causal SNPs (n=70, as in model 4 above) were to a known gene. All 70 causal SNPs were located in or within 0 kb (model 4A), 25 kb (model 4B), 50 kb (model 4C) or 100 kb (model 4D) of a gene. Intuitively, we expected GENOVA to perform better when the gene boundary definition used for analysis matched the gene boundary definition used to simulate the phenotype.

Lastly, we assessed the impact on power of using a stricter *P*-value threshold (0.01 or 0.005, instead of 0.05) when determining whether a clump was associated with the trait or not (Step 1A above). We anticipated that stricter thresholds would provide improved power when the overlap between clumps that associate with the trait and clumps that intersect genes is restricted to a small number of very significant clumps. In contrast, when the overlap extends to a large number of clumps that associate modestly with the trait, then we would expect more liberal thresholds to perform better. We simulated SNP and phenotype data under two models that reproduce these two contrasting scenarios. In model 6A, a total of 18 clumps contributed to the phenotype, each explaining

1% of the phenotypic variance. On the other hand, in model 6B, 180 clumps contributed to the phenotype, each explaining 0.1% of the variance. We estimated the power of GENOVA using three different *P*-value thresholds (0.05, 0.01 and 0.005) under these two models.

2.4. Analysis of publicly available GWAS results

To illustrate the applicability of GENOVA, we downloaded the publicly available summary statistics from three recently published GWAS meta-analysis of height (Lango-Allen, 2010), body mass index (BMI) (Speliotes, 2010) and waist-hip ratio (WHR) (Heid, 2010) conducted by the GIANT consortium. In these studies, sample size varied considerably between SNPs and so, to ensure comparable power across different regions of the genome, we restricted our analysis to SNPs available for >130,000 individuals for height (1,800,462 SNPs), >120,000 individuals for BMI (1,838,860 SNPs), and >74,000 individuals for WHR (1,798,977 SNPs). We then applied GENOVA to each set of GWAS results as described above to test whether SNPs associated with each trait were more likely to intersect known genes than SNPs not associated with the trait.

3. Results

3.1. Type-I error and power of GENOVA

To assess the performance of GENOVA, we simulated a quantitative phenotype under five different models for 3,000 individuals with genotype data available for 38,490 SNPs on chromosome 1, representing 5,269 largely independent groups of highly correlated SNPs (or “clumps”).

In model 1, no single clump contributed to inter-individual variation in the simulated phenotype; therefore, clumps that by chance were associated with the phenotype were expected to intersect genes at the same rate as clumps that were not associated with the phenotype. Indeed, when we analysed 10,000 datasets simulated under this null model, we observed significant GENOVA results only as expected by chance (**Table 1**), indicating that the type-I error rate of GENOVA is close to the expected nominal levels.

Table 1. Type-I error rate for GENOVA.

	Nominal α level			
	0.1000	0.0500	0.0100	0.0050
Type-I error rate	0.1133	0.0590	0.0124	0.0065

Next, we investigated the power of GENOVA to detect a significant overlap between clumps associated with a trait and clumps intersecting genes. To this end, SNP and phenotype data were simulated such that clumps significantly associated with the phenotype at a $P < 0.05$ level were 1.2-fold (model 2) to 1.5-fold (model 5) more likely to intersect a gene than clumps not significantly associated with the phenotype. Power to detect a significant ($\alpha = 0.05$) overlap increased steadily from 26% for model 2 to 85% for model 5 (**Figure 1**, orange line). Power was lower for the same four models when using smaller sample sizes (**Figure 1**, red and blue lines). This demonstrates that, as expected, power increases with increasing GWAS sample size. These results thus confirm that GENOVA can be efficiently used to test the hypothesis that SNPs associated with a trait are more likely to be in or near genes, particularly when considering results from large GWAS.

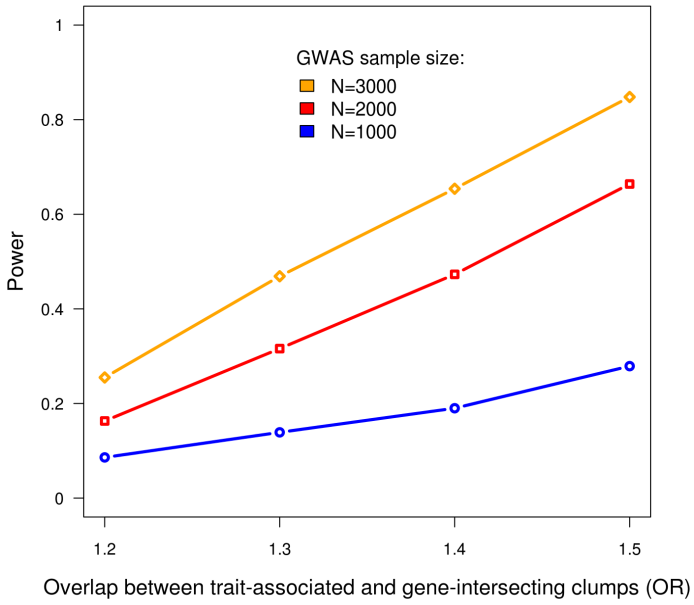

Figure 1. Power of GENOVA as a function of the degree of overlap between trait-associated clumps and gene-intersecting clumps. SNP and phenotype data were simulated such that clumps significantly associated with the phenotype at a $P < 0.05$ level were 1.2-fold to 1.5-fold more likely to intersect a gene than clumps not significantly associated with the phenotype. Data were simulated for 3000 (orange diamonds), 2000 (red squares) or 1000 (blue circles) individuals. Power was estimated as the proportion of 1,000 simulated datasets with a GENOVA $P < 0.05$.

In models 1 to 5, we expanded known gene boundaries arbitrarily by 50 kb when determining whether a clump intersected a gene or not. Next, we assessed the impact on power of expanding gene boundaries by stricter or more liberal intervals. When all SNPs that contributed to variation in the simulated quantitative phenotype were located strictly between known gene boundaries (i.e. between the start and the end of the coding region; model 4A), power of GENOVA was greatest (97%) when an equally strict gene boundary definition was used to determine whether a clump intersected a gene or not (**Figure 2,** blue line). In this model, power progressively dropped as known gene boundaries were expanded by 25 kb (85%), 50 kb (69%) or 100 kb (50%) to include nearby non-coding regions.

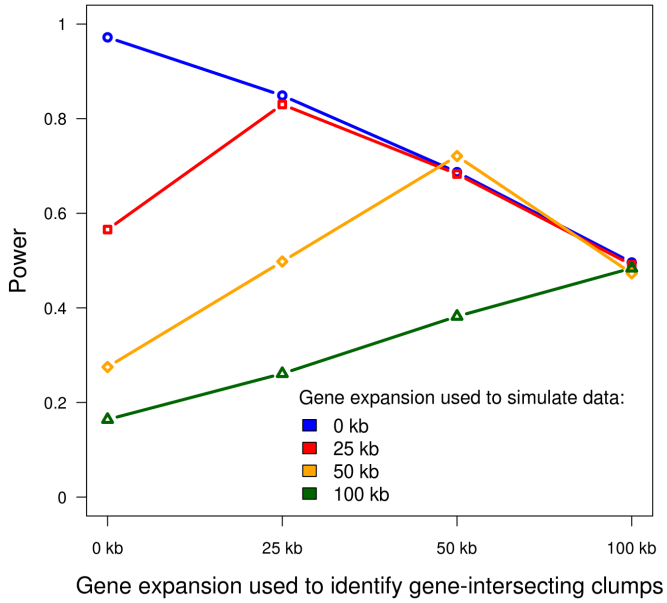

Figure 2. Power of GENOVA as a function of the interval used to expand known gene boundaries when defining gene-intersecting clumps. SNP and phenotype data were simulated under four models which differed in how close the simulated causal SNPs were to a known gene: in or within 0 kb (blue circles), 25 kb (red squares), 50 kb (orange diamonds) or 100 kb (green triangles) of a gene. We then analysed each dataset with GENOVA expanding known gene boundaries by 0, 25, 50 or 100 kb (*x*-xis). Power was estimated as the proportion of 1,000 simulated datasets with a GENOVA $P < 0.05$.

On the other hand, when the simulated causal SNPs were located in or within 25 kb of a gene (model 4B; **Figure 2**, red line), power was greatest when known gene boundaries were expanded by 25 kb (83%), being lower for either stricter (0 kb, 57%) or more liberal (e.g. 49% for 100 kb) gene boundary

definitions. Similar results were observed when the simulated causal SNPs were located in or within 50 kb (model 4C; **Figure 2**, yellow line) or 100 kb (model 4D; **Figure 2**, green line) from a gene: power was greatest when the gene boundary definition used in the GENOVA test matched that used to simulate the genetic data. Across the four models tested (4A to 4D), expanding gene boundaries by either 25 kb or 50 kb resulted in better overall performance.

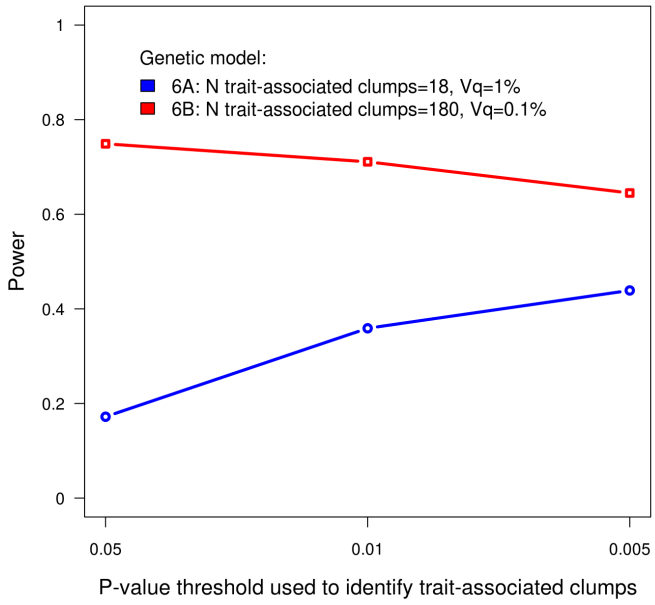

Figure 3. Power of GENOVA as a function of the *P*-value threshold used to identify trait-associated clumps. SNP and phenotype data were simulated under two models which differed in the number of SNPs that contributed to phenotypic variation: 18 SNPs, each explaining 1% of the variance (blue circles) or 180 SNPs each explaining 0.1% of the variance (red squares). We then analysed each dataset with GENOVA using a *P*-value threshold of 0.05, 0.01 or 0.005 (*x*-xis) to define trait-associated clumps. Power was estimated as the proportion of 1,000 simulated datasets with a GENOVA $P < 0.05$.

Lastly, we assessed the impact on power of using a stricter *P*-value threshold (for example 0.005 instead of 0.05) when determining whether a clump was associated with the trait or not. When data were simulated such that the overlap between clumps that associate with the trait and clumps that intersect genes was restricted to a small number of very significant clumps (n=18, each explaining 1% of the phenotypic variation, model 5A), power was greatest when using a stricter *P*-value threshold (e.g. 44% vs 17%, for 0.005 vs 0.05 respectively;

Figure 3, blue line). In contrast, when there were a large number of clumps that overlapped genes but were modestly associated with the trait (n=180, each explaining 0.1% of the phenotypic variation, model 5B), power was greatest when using a more liberal *P*-value threshold (e.g. 64% vs 75%, for 0.005 vs 0.05 respectively; **Figure 3**, red line).

3.2. Analysis of publicly available GWAS results

We applied GENOVA to results from three recent GWAS meta-analyses of height, BMI and waist-hip ratio (WHR) conducted by the GIANT consortium. In these analyses, we expanded known gene boundaries by 50 kb to define gene-intersecting clumps and used a *P*-value threshold of 0.05 to define trait-associated clumps. Results are summarised in **Figure 4**. Clumps significantly associated with variation in height were 1.44-fold more likely to intersect a gene than clumps not associated with height ($P = 5 \times 10^{-28}$). A significant but weaker overlap was also identified for BMI (OR=1.09, $P = 0.008$) and WHR (OR=1.09, $P = 0.014$).

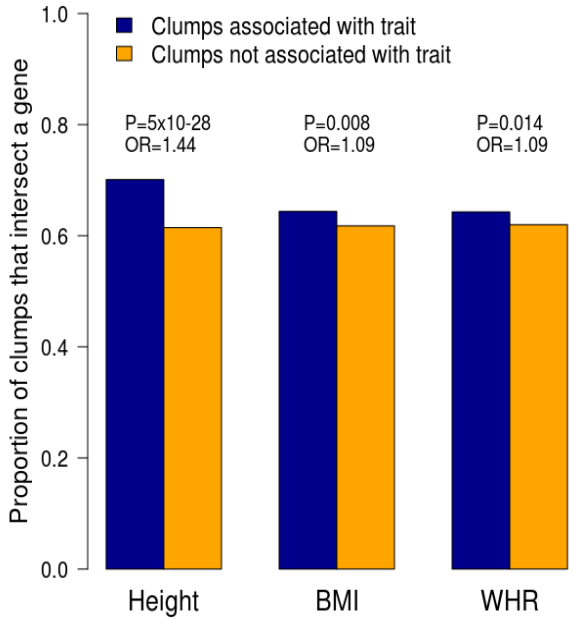

Figure 4. GENOVA results for published GWAS meta-analyses of height, body mass index (BMI) and waist-hip ratio (WHR). Known gene boundaries were expanded by 50 kb to define gene-intersecting clumps and a *P*-value threshold of 0.05 was used to define trait-associated clumps.

4. Discussion

Our simulations suggest that the power of GENOVA is affected by at least three factors: GWAS sample size, the gene boundaries used to define gene-intersecting clumps and the *P*-value threshold used to define phenotype-associated clumps.

For most common, complex diseases, GWAS that use small sample sizes will inevitably have low power to detect true associations between SNPs and disease status. As such, even if for a given disease there is a true underlying overlap between gene-intersecting clumps and disease-associated clumps, the distribution of disease association *P*-values will be very similar between clumps that intersect genes and clumps that do not intersect genes (approaching a null distribution), and so GENOVA will necessarily have low power to distinguish these two groups. On the other hand, as GWAS sample size increases so does the power to detect true SNP-disease associations and, consequently, GENOVA can more efficiently detect *P*-value differences between clumps that intersect genes and those that do not, if such an overlap is indeed present. The implications of this intuitive result are two-fold. First, GENOVA should be applied to adequately powered GWAS, namely to results obtained by large GWAS meta-analyses. Second, care should be taken to ensure that the sample size used is comparable across all SNPs analysed. Sample size can vary widely across SNPs in many meta-analyses and so applying a simple sample size SNP inclusion filter should ensure that power to detect SNP-disease associations is homogeneous across the genome.

Another factor that influences the power of GENOVA is the extent to which known gene boundaries are expanded when defining gene-intersecting clumps, such that nearby non-coding regions (e.g. promoters or enhancers) that may have important regulatory effects are also considered. Our simulations demonstrate that, as expected, power is greatest when gene boundaries are expanded just enough to include all nearby causal SNPs. Expanding boundaries by too little or too much resulted in decreased power. A recent GWAS of gene expression levels suggested that 90% of SNPs that influence the expression level of a gene fall within 15 kb of the gene (Pickrell, 2010). Therefore, in practice, 15 kb can be a biologically meaningful interval to expand known gene boundaries when using GENOVA.

Lastly, our simulations also suggest that the *P*-value threshold used to identify phenotype-associated clumps impacts the power of the test. Specifically, when there were a large number of genetic variants that contributed only modestly to the phenotypic variance, a liberal threshold ($P = 0.05$) provided greater power than a stricter threshold ($P = 0.005$). As this is the most likely genetic model for most common, complex traits or diseases, we recommend applying a *P*-value

threshold of 0.05 to define phenotype-associated clumps when using GENOVA. We also explored the performance of a slightly modified GENOVA test that was not threshold-based but, instead, modeled the full distribution of clump phenotype-association *P*-values (vintile-based test). However, the threshold-based test drastically outperformed the vintile-based test across all models tested (not shown), and so we opted for only implementing the former.

Although our simulation study was based on SNP data for a single chromosome, limited tests performed at the genome-wide scale provided very comparable results (not shown), indicating that our findings can be generalized to GWAS. An additional potential caveat of the proposed approach is that by taking the minimum of *P*-values across SNPs in a clump, larger clumps are potentially more likely to be classified as trait-associated than smaller clumps. However, because SNPs in a clump are highly correlated with each other (all have an $r^2 > 0.8$ with the index SNP), this multiple testing problem is minimized, resulting in a type-I error rate that is close to the expected nominal levels. An additional area for future development is the extension of GENOVA to formerly test for an overlap between results from two separate GWAS (are SNPs significantly associated with trait A in a GWAS more likely to intersect regions significantly associated with trait B in an independent GWAS?).

When applied to three large GWAS meta-analyses conducted by the GIANT consortium (Lango-Allen, 2010; Speliotes, 2010; Heid, 2010), GENOVA found that SNPs associated with variation in height were 1.44-fold more likely to intersect genes than SNPs not associated with height, a highly significant overlap. This result is consistent with the observation in the original study that the 180 most significantly associated height variants clustered near biologically relevant genes (Lango-Allen 2010). We note, however, that excluding height SNPs with a $P < 5x10^{-6}$ from the GENOVA analysis had only a minor impact on the results (1.43-fold enrichment instead of 1.44), suggesting that SNP-clustering near genes extends beyond the most significantly associated loci. Results for BMI revealed a significant but much weaker overlap between associated SNPs and gene regions (1.09-fold), despite this meta-analysis including a comparable number of samples as for height. We speculate that this difference may indicate that height variants are truly more often located in or near genes than are BMI variants (which may reflect, for example, different underlying gene regulatory mechanisms) or, perhaps more plausibly, that these two traits have a distinct genetic architecture, such that despite being based on comparable sample sizes, the height meta-analysis has greater power than the BMI meta-analysis to detect true SNP-trait associations. For example, height may have a larger proportion of loci that explain e.g. >0.05% of the phenotypic variance than is the case for BMI. Some support for this

hypothesis is given by the observation that a 5-fold increase in the number of loci with a $P < 5x10^{-6}$ was observed in the original height discovery analysis (207 loci, based on 133,653 samples) as compared to the equivalent BMI analysis (42 loci, based on 123,865 samples).

In conclusion, GENOVA implements a simple strategy to test the hypothesis that SNPs associated with a phenotype are more likely to intersect genes than SNPs not associated with the phenotype. This information increases confidence that genes located near associated variants are indeed enriched for causal genes, which is a first step towards characterising the biological significance of GWAS results.

References

Gieger C, Radhakrishnan A, Cvejic A, Tang W, Porcu E, Pistis G, Serbanovic-Canic J et al. (2011) New gene functions in megakaryopoiesis and platelet formation. *Nature* 480: 201–208.

Heid IM, Jackson AU, Randall JC, Winkler TW, Qi L, Steinthorsdottir V, Thorleifsson G et al. (2010) Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet* 42(11): 949-60.

Lango-Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, Willer CJ et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467 (7317): 832-8.

Medland SE, Nyholt DR, Painter JN, McEvoy BP, McRae AF, Zhu G, Gordon SD, Ferreira MA, Wright MJ, Henders AK, Campbell MJ, Duffy DL, Hansell NK, Macgregor S, Slutske WS, Heath AC, Montgomery GW, Martin NG (2009) Common variants in the trichohyalin gene are associated with straight hair in Europeans. *Am J Hum Genet* 85(5):750-5.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 464(7289):768-72.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3): 559-75.

Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, Allen HL et al. (2010) Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet* 42(11):937-48.

The 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073.