Published in final edited form as:

Genet Epidemiol. 2016 February ; 40(2): 154–160. doi:10.1002/gepi.21949.

REGIONALLY SMOOTHED META-ANALYSIS METHODS FOR GWAS DATASETS

Ferdouse Begum¹, Monir H. Sharker², Stephanie L. Sherman³, George C. Tseng^{4,5}, and Eleanor Feingold^{4,5}

¹ Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, MD.

² Department of Information Science and Technology, University of Pittsburgh, PA

³ Department of Human Genetics, Emory University School of Medicine, Atlanta, GA.

⁴ Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, PA.

⁵ Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, PA.

Abstract

Genome-wide association studies (GWAS) are proven tools for finding disease genes, but it is often necessary to combine many cohorts into a meta-analysis to detect statistically significant genetic effects. Often the component studies are performed by different investigators on different populations, using different chips with minimal SNPs overlap. In some cases, raw data are not available for imputation so that only the genotyped SNP results can be used in meta-analysis. Even when SNP sets are comparable, different cohorts may have peak association signals at different SNPs within the same gene due to population differences in linkage disequilibrium or environmental interactions. We hypothesize that the power to detect statistical signals in these situations will improve by using a method that simultaneously meta-analyzes and smooths the signal over nearby markers. In this study we propose regionally smoothed meta-analysis (RSM) methods and compare their performance on real and simulated data.

Keywords

GWAS meta-analysis; sliding-window; simulation; window-based method

Introduction

Genome-wide association studies have popularly been used to map disease genes for human complex diseases. Large sample sizes are required to identify SNPs with moderate effect size. Though the number of GWAS has increased exponentially over the last decade, sample sizes of most of the individual studies are not large enough to identify SNPs with small to moderate effects. GWAS investigators have addressed this problem by integrating different

To whom correspondence should be addressed: Ferdouse Begum, PhD, Postdoctoral Fellow, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Tel: +1 410 502 3593, fbegum@jhsph.edu.

Conflict of Interest

None

studies using meta-analysis to improve power, but there are a number of unresolved methodological issues in the proper application of these methods [Begum, et al. 2012].

One of the methodological challenges identified in Begum et al. [Begum, et al. 2012] was the need for a method for meta-analysis of studies that have been performed on different chips. When different genotyping platforms are used, there is often limited overlap between SNP sets. Even when different data sets are genotyped on the same chip, but in different laboratories, differing quality control processes can lead to non-matching SNP sets. The standard practice for GWAS meta-analysis is to impute the non-overlapping and missing SNPs in each study, but in some cases, raw data are unavailable so that imputation cannot be performed. Moreover, if imputation is performed in regions where one or more datasets has low information content (sparse SNPs), the meta-analysis can give false negative results because it treats the uninformative imputed data identically to the genotyped data. Thus imputation can actually be misleading if SNP sets do not match. In this study we propose a set of methods to improve performance by performing meta-analysis of GWAS with non-overlapping SNPs incorporating regional smoothing.

There are many methods in the literature for combining results of a single GWAS across windows or groups of SNPs. The usual motivation for these methods is to increase biological significance and/or decrease the multiple testing burden. Different types of genomic region based GWAS have been proposed, such as gene-based methods [Curtis, et al. 2008; Hibar, et al. 2011; Li, et al. 2009; Ma, et al. 2013], haplotype-based methods [Iliadis, et al. 2012; Lorenz, et al. 2010; Wan, et al. 2013], pathway-based methods [Kanehisa and Goto 2000; Li, et al. 2012; Yu, et al. 2009; Zhang and Drabier 2012] and genomic region-based methods [Lin, et al. 2012; Sha, et al. 2009; Tang, et al. 2009]. In addition, there have been window-based GWAS, which motivated us to propose the methods described in this paper. In window-based GWAS, different window types and sizes have been considered, such as non-overlapping fixed windows, overlapping sliding windows and overlapping variable sized sliding windows [Clayton and Jones 1999; Huang, et al. 2007; Mathias, et al. 2006; Tang, et al. 2009]. Most of the studies aimed to capture the LD structure in the windows in finding associated disease SNPs. Depending on the principal components of genotypes, some studies tried to find an optimal window size [Tang, et al. 2009]. Window-based GWAS methods have been used application, but there is no optimal solution for the window size question.

Our proposed regionally smoothed meta-analysis (RSM) methods work on genomic intervals genome-wide and provide the ranks of the genomic intervals depending on the significance level of the association as an output. We construct statistics in two stages. In the first stage, the whole genome is divided into fixed or sliding windows and we compute a summary of the window effect in each study individually. In the second stage, we then add an additional layer of meta-analysis to combine results for each window across. We consider varying window sizes and types as well as various methods for combining the results for the SNPs within the window.

Methods

First Stage

In the first stage we combine statistical evidence for association across SNPs within a regional window in a single GWAS study. We consider several different methods for doing this, most of which have been used before in similar methods. The standard gamble in choosing a method for combining evidence across SNPs is that some methods are more powerful to detect large effects of single SNPs (for example by choosing the smallest p -value in a window) and others are more powerful when there are several SNPs of smaller effect (for example by averaging across the window). Standard gene-based methods are often designed to detect the combination of multiple smaller effects (averaging-type statistics), but in our case we are somewhat more interested in single large effects (minimum p -value type statistics). The reason for this is that we are interested in the situation where there is a moderate-sized effect, but it may be evidenced in different SNPs in different studies. The SNP-combination methods we consider are described below.

(I) Fishers statistic (FS)—Fishers statistic is an averaging-type statistic that combines p -values by summing up the log scaled p -values and multiplying the sum by (-2) , assuming that all the p -values are independent. Under the null hypothesis, this statistic follows a chi-square distribution with two times the number of studies as the degrees of freedom. The equation of Fishers statistic is given below:

$$X^2 = -2 \sum_{i=1}^k \log(p_i) \sim \chi_{2k}^2,$$

where k is the number of studies to combine. In using this statistic to combine p -values for SNPs in a window, we know that the tests are not independent, so the asymptotic chi-square distribution does not hold. We use this statistic only as a scoring function, and do not assess statistical significance.

(II) Minimum p (MP)—To select the “best” p -value in a region, a widely used statistic is Tippet's minimum p -value [Tippett. 1931], which is defined as

$$X = \min_{1 \leq k \leq K} p_k$$

This statistic can be used directly as a scoring function, but we also consider the following.

(III) Derived Minimum p (DMP)—Tippet's minimum p -value follows Beta $(1;K)$ distribution under the null hypothesis. We used both the minimum p -value (above) and the p -value obtained from the Beta distribution as our summary statistics. Using the p -values obtained from the Beta distribution gives a statistic that can be interpreted as a p -value, which may behave more appropriately when used with our second-stage procedures.

(IV) Mean of log of p (MLP)—The other averaging-type statistic we considered is the arithmetic mean of the logarithm of the p -values in each window, which is given below:

$$X = \frac{1}{k} \sum_{i=1}^k \log(p_i)$$

Second Stage

In the second stage, our goal is to “sum” information across studies for each interval, so we use Fisher's statistic to combine p-values obtained in first stage for each window. We apply it directly to each of the scoring functions defined above. This statistic clearly does not follow the usual chi-square distribution. It can be used as a scoring function for ranking intervals, or p-values can be assigned by permutation if desired. All the steps described above are presented in Figure 1.

Window Type and Size

In the first stage of the RSM method we considered two different types of windows: fixed window meta-analysis (FWM) and sliding window meta-analysis (SWM). In general, fixed windows run more risk of missing a signal if it is split between windows, but a sliding window approach is more computationally intensive. We experimented with different window sizes: 50k, 75k, 100k, 300k, and 500k. The choice of window size and sliding or not also depends on the first-stage statistic. If we use the minimum p-value statistics in the first stage, the window choices have much less effect than if we use the averaging-type statistics such as Fisher's method or the mean log p-value.

Performance Measure of RSM Methods

It is important to note that we are not calculating correct p-values for any statistics. Instead we use the statistics only for ranking and focus on the ranks of the regions. This approach is tailored to the most common goal of genomic analysis, which is to achieve the most correct ranked list of results for follow-up analysis [Gail, et al. 2008a]. If one wanted to use these methods for hypothesis testing, permutation analysis would allow calculation of correct p-values.

Since we are applying these methods for ranking rather than for hypothesis testing, the performance needs to be judged based on the ranks each method produces. We consider a good procedure to be one that is able to put true-positive regions within the top 10 on a ranked list of regions (and the higher the better). It is also possible to calculate a positive prediction fraction [Gail, et al. 2008b] or other measure of ranking correctness.

Testing In Real Data

Methods

We applied this method to GWAS data for the phenotype of human meiotic recombination. This study included three populations: the Geneva Dental Caries Study (GDCS) [Shaffer, et al. 2011], the Autism Genetic Resource Exchange (AGRE) [Weiss LA 2008] and the Framingham Heart Study (FHS) [Dawber, et al. 1951]. The GDCS and AGRE samples were genotyped on the Illumina Human610-Quad Beadchip, and the FHS samples were genotyped on the Affymetrix 5.0 chip. GDCS and FHS genotype data are available at the

National Center for Biotechnology Information database of Genotype and Phenotype (dbGaP). AGRE data are available at the science program of autism speaks database (<https://research.agre.org/>). After quality control, final analysis was limited to autosomes only and a total of 551,227 SNPs, 520,018 SNPs, and 388,060 SNPs from GDCS, AGRE and FHS respectively were used.

Two-generation nuclear pedigrees with two or more children were used for this study: 171 from GDCS, 737 from AGRE, and 654 from FHS. Genotype data on each family were used to score recombination for the parents. The parents were then used as the subjects for the GWAS analyses. Sample sizes of each study are presented in Figure 2.

Recombination events in each nuclear family were called according to the method described in Chowdhury et al. [Chowdhury, et al. 2009]. Briefly, the method is as follows. First, a set of informative markers was identified in each family. A locus is informative if one parent is homozygous and another is heterozygous. Among two or more children, one is considered as reference child, and in a sibling pair a switch from one allele to another allele in a particular parental haplotype as we move along the chromosome indicates a recombination in that parent with heterozygous allele. For more accurate estimation, we used 5 or more consecutive markers to call a recombination event.

For the current experiments, we chose three recombination phenotypes: average recombination count (ARC), hotspot percent (HS_PCT), and non-hotspot count (NHS_CNT). Average recombination count (ARC) of a parent is calculated as the total number of recombination events in all children divided by the number of children. Hotspot percent (HS_PCT) is the percent of recombination events in historical hotspot areas. Non-hotspot count (NHS_CNT) is the average of the recombination event counts in non-hotspot areas.

To identify genes or SNPs associated with different aspects of recombination, we conducted GWAS for each phenotype. We used PLINK (29) to conduct all GWAS using an additive genetic model. We used the linear regression option in PLINK for the association tests, as all of our phenotypes are continuous. Before performing GWAS, we cleaned the data following standard GWAS data cleaning pipeline.

Since two of the data sets are from same array, it is easy to use existing meta-analysis methods with the overlapping SNPs. Here we include all three of the data sets in order to test our meta-analysis methods. The advantage of these datasets for testing our methods is that they contain known true positives. There are several well-established recombination genes such as *RNF212* and *PRDM9* [Chowdhury, et al. 2009; Kong, et al. 2014], and we can use this dataset to test for our ability to find (highly-rank) those gene regions for various phenotypes. SNPs in *RNF212* are known to be associated with total recombination, with those associations being somewhat different in males and females. SNPs in *PRDM9* are known to be associated with recombination in historical hotspots in both males and females. That association is strongest for the phenotype HS_PCT, but can also be observed in count of recombination in hotspots (not included in this paper) and count of recombination outside of hotspots (NHS_CNT). We tested the ability of all of our methods (including variations on

methods such as window size) to highly rank (and thus detect) these region/phenotype combinations. Moreover, a known problem with these datasets is that the FHS dataset, genotyped on the Affymetrix 5.0 chip, does not have any SNPs in the *PRDM9* gene, the gene that is highly associated with recombination in hotspots, so this provides a test of exactly the type of obstacle that we hope our method can overcome.

Results

Our first tested phenotype was HS_PCT and we looked at the rank of the *PRDM9* gene. Performance of different methods is listed in Table 1. When we considered a smaller window size such as 50k, the gene was split in two windows. Ranks of the two intervals are then listed in the table. Sometimes due to the positioning of the window on the gene, it may split in two or more windows even if the windows are longer. The results showed that use of MP and DMP statistics in first stage performed very well irrespective of the window size. The MLP statistic performed worse with increasing window size. FS was also poorer with a bigger window size.

Our second phenotype of interest was NHS_CNT and we again looked at the ranks for the *PRDM9* gene. This is a more challenging test for the methods than HS_PCT, because the effect size of the gene on this phenotype is smaller. Table 2 shows the ranks of the *PRDM9* gene for different window sizes and for different statistics in different stages. The result showed that DMP statistics gave lowest (best) rank for the *PRDM9* gene. In fact, only DMP performed well enough that the gene would be likely to be detected in a GWAS, although MP is close.

Table 3 shows the results for the *RNF212* gene for the ARC phenotype. Again, the effect size is quite large and all methods perform well. FS was the best-performing method, followed by MP. Among the window sizes, 100k performed best.

Testing in simulated data

Methods

To evaluate and compare the performance of our proposed methods, we carried out a series of simulations. To test the power, type I error, and computation time of our methods, we performed two simulations: one on synthetic data and the other on permuted data of real GWAS. In the synthetic data, we applied both fixed window and sliding window methods and compared computation time in different settings. In the permuted real data, in addition to investigating comparative performance of our methods, we looked at the proportion of times truly identified the true signal among multiple replications. Both of the simulation schemes are explained below.

Synthetic data

Data generation—We generated five GWAS datasets. For each dataset, we first simulated genome-wide p-values under the null hypothesis of no genome wide association. To simulate true positive regions, we substituted corresponding segments from our null data sets with 500kb and 150kb true positive regions from two real GWAS data sets. Details of

the simulated data sets are presented in the supplementary document. Once the five datasets were generated, we randomly dropped SNPs from each of the five according to the scheme described in the supplement in order to mimic the effect of having different SNP sets in each dataset. To compare our proposed methods, we tracked the rank of the truly associated regions, and the time required to complete the analysis in each method.

Results—We tested both fixed window and sliding window methods and all four statistics choices in the first stage of the RSM method. We recorded the rank of the positively associated two regions (the one with minimum p-value is assigned the highest rank). The rankings of these two positively associated regions, as obtained using different methods, are listed in Table S1. When we considered a smaller window size such as 50k, the regions were split into more windows. The results show that use of FS, MP and DMP statistics in the first stage performed very well irrespective of the window size. The MLP statistic performed worse for smaller window size compared to larger window size and overall, compared to other methods. Similarly for the sliding window method, region one was detected by all methods since the association was stronger and the LD block was larger compared to the region two. The region two was also detected by FS, MP and DMP as a first or 2nd rank which is better than the performance of MLP irrespective of the window size.

Computation Time—The computation time for RSM depends on several factors. There were 4 statistics to choose in the first stage and 5 window sizes totaling 20 possible methods. To get a comparative computation time assessment, we assigned the methods to twenty different nodes each with 50GB memory in a high performance cluster; one method for each node. The computation can also be performed on a computer with lower memory size (e.g. laptops with 8GB memory). Computation time for all 20 cases using fixed window and sliding window is listed in the supplementary document (Table S1 and Table S2). Irrespective of window size, the FS statistic took the maximum time. The MLP statistic took the minimum time. Time decreased proportionately with the window size.

Permuted real data

Data generation—We selected one phenotype from our original recombination data sets. We varied the number of studies to be included in the meta-analysis between two to six GWAS in each meta-analysis, as presented in Table S5. We considered 50 replicates of such meta-analyses. We randomly selected the percentage of SNPs in each dataset we would include in each of the replications. The complete simulation scheme is described in Table S6.

Results—We tested all four statistics (FS, MP, DMP and MLP) for fixed windows and looked at the rank of the known gene on chromosome 5. Results of 50 replications are presented in Table 4. Among the 4 statistics, DMP worked best irrespective of the window sizes, followed by MP. MLP worked better for smaller window size and FS worked better for 100k window size. Overall, DMP and MP performed better than MLP and FS.

Discussion

In this study we demonstrated that GWAS results can successfully be combined across studies with disparate SNP sets using a two-stage smoothing and meta-analysis procedure. We considered different methods for the smoothing stage and different window types and sizes. The performances of the methods were compared on simulated data set as well as on three phenotypes in real datasets.

We tested four different statistics for the first stage. FS and MLP are statistics that more or less sum or average results across SNPs, while MP and DMP choose the best p-value in each window. Results on real data show that MP worked best across phenotypes, even when the effect size was relatively smaller such as the association between *NHS_CNT* and *PRDM9*. For *PRDM9* and *RNF212*, window size 100k worked best. But in general which window size will perform best will also depend on the linkage disequilibrium (LD) blocks or gene size. A window size that can capture the whole gene or LD block may yield lowest rank.

In our synthetic simulated data, we demonstrated that FS, DMP, and MP all perform well. This may be because these simulated data sets are different than our real data sets in the null regions. In particular, the synthetic dataset has no linkage disequilibrium in the null regions. To overcome this limitation, we considered simulation based on permuted real data. The permuted real data showed that DMP worked best in general followed by MP.

The choice of fixed window vs. sliding window and the choice of smoothing statistics are linked. If we choose MP or DMP in the first stage, it may not necessary to choose sliding window. Instead window size choice will be critical. Program run time for fixed windows is faster than for sliding windows, so this is a further advantage to these methods. The first stage is much more computationally intensive than the second stage. Computing time in the first stage varies greatly among the choices of window sizes. Smaller window size and smaller size of increments takes longer time compared to the larger window size and larger size of increments in sliding window method.

In the second stage we showed results for FS only. We also investigated the adaptively weighted statistic (AWS), which is computationally more challenging. The AWS [Li 2011] uses optimal weights for each study and is defined by $X = \min_{w \in W} pU(u(W))$, where w is the weight and $u(w)$ is the observed statistic for the weight function

$U(w) = -\sum_{i=1}^k w_i \log(p_i)$ and $W = \{w | w_i \in \{0,1\}\}$. Compared to FS, AWS provides better power for a range of alternative hypotheses. This method might be useful in the second stage if computational limitations can be overcome.

In summary, computational time will vary based on the number of studies and choice of window sizes, choice of methods in first and second stages and window types. With regard to sliding windows vs. fixed, sliding will always have a statistical advantage, because it avoids the problem of accidentally splitting a signal in half (as in our real data example), but fixed windows may be preferable for computational reasons. Meta analysis of less than ten studies, fixed window method can be done in regular PCs. But more than 10 studies and sliding window will make the computation time consuming, but use of high performance

computing facilities can resolve that problem very easily. Using MP statistic in the first stage with 100k-window size and fixed window method could be a starting place.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

National Institutes of Health (NIH) (R01MH077159, RC2HL101715); NIH (R01HD38979, R01DE14899); NIH (R01HD038979); Funding for open access charge: University of Pittsburgh. We are thankful to COPDGene consortium for providing data for method validation.

References

- Begum F, Ghosh D, Tseng GC, Feingold E. Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic acids research*. 2012; 40(9):3777–84. [PubMed: 22241776]
- Chowdhury R, Bois PR, Feingold E, Sherman SL, Cheung VG. Genetic analysis of variation in human meiotic recombination. *PLoS Genet*. 2009; 5(9):e1000648. [PubMed: 19763160]
- Clayton D, Jones H. Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet*. 1999; 65(4):1161–9. [PubMed: 10486335]
- Curtis D, Vine AE, Knight J. A simple method for assessing the strength of evidence for association at the level of the whole gene. *Advances and applications in bioinformatics and chemistry : AABC*. 2008; 1:115–20. [PubMed: 21918610]
- Dawber TR, Meadors GF, Moore FE Jr. Epidemiological approaches to heart disease: the Framingham Study. *American journal of public health and the nation's health*. 1951; 41(3):279–81.
- Gail MH, Pfeiffer RM, Wheeler W, Pee D. Probability of detecting disease-associated single nucleotide polymorphisms in case-control genome-wide association studies. *Biostatistics*. 2008a; 9(2):201–15. [PubMed: 17873152]
- Gail MH, Pfeiffer RM, Wheeler W, Pee D. Probability that a two-stage genome-wide association study will detect a disease-associated snp and implications for multistage designs. *Annals of human genetics*. 2008b; 72(Pt 6):812–20. [PubMed: 18652601]
- Hibar DP, Stein JL, Kohannim O, Jahanshad N, Saykin AJ, Shen L, Kim S, Pankratz N, Foroud T, Huentelman MJ. Voxelwise gene-wide association study (vGeneWAS): multivariate gene-based association testing in 731 elderly subjects. *NeuroImage*. 2011; 56(4):1875–91. others. [PubMed: 21497199]
- Huang BE, Amos CI, Lin DY. Detecting haplotype effects in genomewide association studies. *Genet Epidemiol*. 2007; 31(8):803–12. [PubMed: 17549762]
- Iliadis A, Anastassiou D, Wang X. Fast and accurate haplotype frequency estimation for large haplotype vectors from pooled DNA data. *BMC genetics*. 2012; 13:94. [PubMed: 23110720]
- Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*. 2000; 28(1):27–30. [PubMed: 10592173]
- Kong A, Thorleifsson G, Frigge ML, Masson G, Gudbjartsson DF, Villemoes R, Magnusdottir E, Olafsdottir SB, Thorsteinsdottir U, Stefansson K. Common and low-frequency variants associated with genome-wide recombination rate. *Nat Genet*. 2014; 46(1):11–6. [PubMed: 24270358]
- Li J, Tseng GC. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics*. 2011; 5(2A):994–1019.
- Li M, Wang K, Grant SF, Hakonarson H, Li C. ATOM: a powerful gene-based association test by combining optimally weighted markers. *Bioinformatics*. 2009; 25(4):497–503. [PubMed: 19074959]

- Li MX, Kwan JS, Sham PC. HYST: a hybrid set-based test for genome-wide association studies, with application to protein-protein interaction-based association analysis. *American journal of human genetics*. 2012; 91(3):478–88. [PubMed: 22958900]
- Lin YC, Hsiao CL, Hsieh AR, Lian Ie B, Fann CS. Using maximal segmental score in genome-wide association studies. *Genetic epidemiology*. 2012; 36(6):594–601. [PubMed: 22807216]
- Lorenz AJ, Hamblin MT, Jannink JL. Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. *PLoS ONE*. 2010; 5(11):e14079. [PubMed: 21124933]
- Ma L, Clark AG, Keinan A. Gene-based testing of interactions in association studies of quantitative traits. *PLoS genetics*. 2013; 9(2):e1003321. [PubMed: 23468652]
- Mathias RA, Gao P, Goldstein JL, Wilson AF, Pugh EW, Furbert-Harris P, Dunston GM, Malveaux FJ, Toggias A, Barnes KC. A graphical assessment of p-values from sliding window haplotype tests of association to identify asthma susceptibility loci on chromosome 11q. *BMC Genet*. 2006; 7:38. others. [PubMed: 16774684]
- Sha Q, Tang R, Zhang S. Detecting susceptibility genes for rheumatoid arthritis based on a novel sliding-window approach. *BMC proceedings*. 2009; 3(Suppl 7):S14. [PubMed: 20018003]
- Shaffer JR, Wang X, Feingold E, Lee M, Begum F, Weeks DE, Cuenco KT, Barmada MM, Wendell SK, Crosslin DR. Genome-wide association scan for childhood caries implicates novel genes. *Journal of dental research*. 2011; 90(12):1457–62. others. [PubMed: 21940522]
- Tang R, Feng T, Sha Q, Zhang S. A variable-sized sliding-window approach for genetic association studies via principal component analysis. *Annals of human genetics*. 2009; 73(Pt 6):631–7. [PubMed: 19735491]
- Tippett, LHC. *The Methods in Statistics*. Williams and Norgate, Ltd.; 1931.
- Wan, X.; Yang, C.; Yang, Q.; Zhao, H.; Yu, W. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE. ACM*; 2013. HapBoost: A fast Approach to Boosting Haplotype Association Analyses in Genome-Wide Association Studies..
- Weiss LA SY, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MA, Green T, Platt OS, Ruderfer DM, Walsh CA, Altshuler D, Chakravarti A, Tanzi RE, Stefansson K, Santangelo SL, Gusella JF, Sklar P, Wu BL, Daly MJ. Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med*. 2008; 358(7):737–9. [PubMed: 18184953]
- Yu K, Li Q, Bergen AW, Pfeiffer RM, Rosenberg PS, Caporaso N, Kraft P, Chatterjee N. Pathway analysis by adaptive combination of P-values. *Genetic epidemiology*. 2009; 33(8):700–9. [PubMed: 19333968]
- Zhang F, Drabier R. IPAD: the Integrated Pathway Analysis Database for Systematic Enrichment Analysis. *BMC bioinformatics*. 2012; 13(Suppl 15):S7.

Figure 1.
Steps of the RSM methods.

Figure 2.
Sample size distribution of different studies.

Table 1

Ranks of *PRDM9* gene for HS_PCT phenotype

Stages		Window size				
First Stage	Second Stage	50k	75k	100k	300k	500k
Fishers statistics (FS)	Fishers statistics (FS)	1,74 *	6,20	1	6	20
Minimum p (MP)	Fishers statistics (FS)	1	1	1	1	1
Derived minimum p (DMP)	Fishers statistics (FS)	1,7	1,6	1,58	1	1
Mean of log of p (MLP)	Fishers statistics (FS)	1	3	2	8	16

* Ranks of the two windows into which the gene was split

Table 2

Ranks of *PRDM9* gene for NHS_CNT phenotype

Stages		Window size				
First Stage	Second Stage	50k	75k	100k	300k	500k
Fishers statistics (FS)	Fishers statistics (FS)	217	85	90	115	259
Minimum p (MP)	Fishers statistics (FS)	20	13	10	4	139
Derived minimum p (DMP)	Fishers statistics (FS)	8	7	1	2	56
Mean of log of p (MLP)	Fishers statistics (FS)	62	89	19	104	243

Table 3

Ranks of *RNF212* gene for ARC phenotype

Stages		Window size				
First Stage	Second Stage	50k	75k	100k	300k	500k
Fishers statistics (FS)	Fishers statistics (FS)	1	2	1	3	3
Minimum p (MP)	Fishers statistics (FS)	2	2	2	3	3
Derived minimum p (DMP)	Fishers statistics (FS)	4,6 *	4,6	3	3,4	3,7
Mean of log of p (MLP)	Fishers statistics (FS)	2,4	3,4	2	2	3

* Ranks of the two windows into which the gene was split

Table 4

Simulation scheme 2 results of 50 replications

Window size	Statistics	Statistics in 2 stages			
		FS_FS	MP_FS	DMP_FS	MLP_FS
50k window size	True positive as top hit	11/50 (22%)	34/50 (68%)	38/50 (76%)	30/50 (60%)
	True positive among top ten hits	27/50 (54%)	45/50 (90%)	48/50 (96%)	41/50 (82%)
100k window size	True positive as top hit	15/50 (30%)	35/50 (70%)	40/5 (80%)	15/50 (30%)
	True positive among top ten hits	32/50 (64%)	39/50 (78%)	44/50 (88%)	31/50 (62%)
300k window size	True positive as top hit	0/50 (0%)	35/50 (70%)	39/50 (78%)	1/50 (2%)
	True positive among top ten hits	24/50 (48%)	42/50 (84%)	45/50 (90%)	14/50 (28%)