Risk Quantification of Multigenic Conditions for SNP Array Based Direct-to-Consumer Genomic Services

Svetlana Bojić[1] and Stefan Mandić-Rajčević[2]

[1]Faculty of Biology, University of Belgrade, Serbia
ceca.bojic@gmail.com
[2]International Centre for Rural Health, Department of Health Sciences, University of Milan, Italy
stefan.mandicrajcevic@gmail.com

Abstract. Genome wide association studies (GWAS) are typically designed as case-control studies, collecting thousands of sick and healthy individuals, genotyping hundreds of thousands of SNPs, and documenting the SNPs which are more abundant in one group or the other. Direct-to-consumer genetic testing has opened the possibility for a regular person to receive data about his/her genotype, but the validity of risk assessment procedures and the final genetic risk estimate have been questioned. Many authors have discussed the advantage of use of the asymptotic Bayes factor (ABF) to measure the strength of SNP/trait associations, over the use of p-values. We propose a ABF based heuristic to filter-our and select SNP/trait associations to be used in multigenic risk assessment.

A raw genotype result from the 23andMe web service was merged with the GWAS catalog, and SNP/trait associations were filtered and selected using the R programming language together with free and publicly available databases.

From the initial 3195 SNP/trait associations, only 425 remained after the initial filters on descent, replicated findings, qualitative trait and availability of the number of cases and controls in the study. Selecting only one SNP/trait association from repeated studies and studies done with proxy SNPs left us with 377 SNP/trait associations available for multigenic risk assessment. After excluding the associations with unsatisfying ABF, only 300 SNP/trait associations remain for the multigenic risk assessment.

Whatever the link between SNP/trait associations and final DTC multigenic risk assessment for a given trait is, the final value of a risk score is heavily influenced by the number, as well as strength of evidence for individual SNP/trait pairs that are used for calculation. The ABF provides an unambiguous and simple criterion for ranking and including SNP/trait associations in multigenic risk assessment.

Keywords: GWAS, multigenic risk assessment, Bayes factor, SNP selection, direct to consumer.

1 Introduction

Genome wide association studies (GWAS) test for association between a disease or a quantitative trait and multiple genetic markers. These associations fall under the

F. Ortuño and I. Rojas (Eds.): IWBBIO 2015, Part II, LNCS 9044, pp. 264–275, 2015.

"common disease – common variant" hypothesis. The hypothesis suggests that the occurrence of common complex diseases is influenced by a moderate number of (interacting) disease alleles, called "casual variants". The most common type of markers are single nucleotide polymorphisms (SNPs). Nowadays, the SNP array technology offers the possibility to determine the state of even more than one million SNPs in one experiment. A typical GWAS is designed as a case-control study, collecting thousands of sick and healthy individuals, genotyping hundred thousand SNPs, and documenting the SNP-genotypes which are more abundant in one group or the other. They have been increasingly popular since almost a decade ago, but have been criticized, among other things, for lack of reproducibility or unclear utility for clinicians [1]. Nevertheless, GWAS have successfully identified many genetic variants contributing to the susceptibility for complex diseases.

To this day thousands of SNPs have been flagged as associated with hundreds of diseases, but the ability to predict one's disease status based solely on SNPs fails short of what would be expected, having in mind the high heritability of these diseases. One explanation could be that many phenotypes might be defined by a large group of SNPs with tiny effects, and present day GWAS are underpowered to detect them. For instance, an author estimated the overall number of SNPs which affect height to be 93,000 [2]. The naïve idea that it is enough to use the SNPs flagged as significantly associated with diseases or traits in GWAS has lead to low predictive value of risk assessments done in this way, as a large number of SNPs remain outside that scope.

Direct-to-consumer (DTC) genetic testing applies to the situations where genetic tests are marketed directly to the consumer via television, print advertisements or the Internet, as opposed to being ordered through healthcare providers such as physicians or genetic counselors. The upside of DTC genetic testing is that its growing market might promote awareness of genetic diseases and even more importantly allow patients/consumers to have a more proactive role in their health care, change their lifestyle, and organize their life better. There are, however, significant risks, as consumers might be misled by the results of invalid tests, or take important decisions regarding their health based on incomplete or misunderstood information [1].

Risk assessment is often defined as the determination of quantitative or qualitative value of risk related to a concrete situation and a recognized threat, and is also an essential part of genetic testing and counseling. It must be calculated as accurately as possible to enable the clinician and the patient to make sound health-related decisions. The calculation of genetic risk should incorporate all available information at a particular point in time. It should be considered an ongoing process of analysis of estimates [3]. As for DTC risk assessment, two trends dominate the industry [4]. First, practiced for instance by the web services Promethease, Interpretome, or LiveWello limits itself to listing SNP-wise (allele or genotype) based risk, as extracted from dbSNP database [5]. This information might be enriched with ClinVar entry providing the basis for dividing those SNP-wise risks into "good news" and "bad news" [6], with unquestionable scientific background, but yet not coherent enough to direct potential health related intervention. Higher level of DTC risk assessment implies combining SNP-wise evaluated risks into some form of multigenic risk score for a given trait. This has proven to be tricky, because we are far from knowing how these

SNPs interact shaping the influence on any phenotype. The three leading companies (23andMe, deCODE, Navigenics) use both absolute lifetime risk and relative risk compared to the general population, as obtained from odds ratios for individual SNPs under multiplicative effect assumption. Both approaches require high level of transparency regarding the exact method used, since there are numerous sources of potential bias (starting with inappropriate specification of disease prevalence in the control population, to unclear assumptions about prior odds), not to mention the high variability of the risk estimates even when well defined workflow is strictly followed [7]. These sources of compromised validity have been recognized by the Federal Drug and Administration agency, which resulted in the (in)famous ban of the 23andMe health reports [8].

In the frequentist inference, an association of a marker with a disease is evaluated based on whether or not the *p*-value of an association test is less than a significance level. The markers that have significant *p*-values are tested for replication in subsequent studies. A typical significance level for GWA studies is determined by Bonferroni procedure, and it lies somewhere around $5x10^{-7}$, although other levels have emerged and have been evaluated in the literature [9, 10]. However, this significance level is used in practice independently of the sample size of studies, the effect size of associated markers, and the power to detect association [10]. For many common diseases, the effect size in studies is small, having an odds ratio in the range of 1.1-1.5. Therefore, markers with true associations can have *p*-values greater than the significance level, or even be ranked far away from the top, if based solely on *p*-values [11].

In Bayesian hypothesis testing, Bayes factor (BF) is often used and reported to measure the strength of association. The BF is a ratio of the probabilities of the observed data under the alternative H_1 and the null H_0 hypotheses, in which all the parameters are treated as random and averaged out with respect to their prior distributions [12, 13]. Many authors have discussed the advantage of BF over *p*-values for genetic association studies [14-16]. It measures the strength of association by integrating the significance of association with the power to detect it, while the *p*-value only measures the significance of association [15, 16]. When the sample size is large enough (thousands of cases and thousands of controls), large BF values ($\log_{10}BF > 5$) strongly support observed small *p*-values and therefore the association. On the other hand, small BF values ($\log_{10}BF < 0$) can be used to exclude the markers with small *p*-values as false positives with high confidence [15]. In addition, to compare results across studies, the BF is a better measure that the *p*-value, since it integrates both the significance of association and the power (sample size and effect size), as demonstrated by simulating different studies with different sample sizes and genetic effects, showing that different studies had different BFs, while the *p*-values remained the same [15]. The more widespread use of the Bayes factor has been hampered by the need for prior distributions to be specified for all of the unknown parameters in the model, and the need to evaluate multidimensional integrals, a complex computational task. Still, Wakefield [14, 17] has proposed the asymptotic Bayes factor (ABF) which avoids each of these requirements.

Whatever is the link between SNP/trait associations and final DTC multigenic risk assessment for a given trait, the final value of any multigenic risk score is heavily

influenced by the number, as well as strength of evidence for individual SNP/trait pairs that are used for calculation. This is where ABF based inference can provide unambiguous and simple criterion for the ranking of SNP/trait associations, which embeds all the relevant information at the same time. In the present work we demonstrate the usefulness of ABF values in selecting SNP/trait associations for DTC multigenic risk assessment.

2 Methodology

Our subject was a European 30-year old Caucasian man, from the Balkan peninsula. He had received his raw genotype data from the 23andMe web service [18] and expressed interest in gaining knowledge regarding his own genetic risks.

2.1 Importing Data and Associating SNPs with GWAS

Raw data were a tab delimited text file, downloaded from the 23andMe website, with four columns: rsid (SNP marker ID), chromosome, genomic position and (unphased) genotype.

The data.frame with raw genotypes was merged with data from the National Human Genome Research Institute (NHGRI) GWAS catalog [19] using the ***gwascat*** *R* package [20]. This way we had several additional columns, describing all the GWAS that have implied some kind of trait association with our SNPs.

The rows with the "strongest risk allele" entry from a GWAS study matching any of the genotyped alleles for a particular SNP were retained for the future analysis, representing the body of evidence for this subject's genotype/trait risk estimation.

2.2 Filtering SNP-Trait Associations

After reducing our *data.frame* to only the SNPs with "risk" alleles, we applied several filters to account for the subject's origin and to rule out studies which had insufficient information by our criteria. Filters were created using regular expressions, defined as sequences of characters which form a search pattern, implemented through the ***stringr*** *R* package [21].

First we filtered out studies which were not done on European population, then studies which were not replicated. Finally we excluded studies which had no odds ratios (OR) and/or confidence intervals (CI).

Before calculating the ABF, number of cases and controls in each study were extracted using regular expressions.

2.3 Asymptotic Bayes Factor Calculation

Prior probability of H_1 was set to be equal for all the SNPs, dependent only on the number of SNPs investigated in the particular study. The logic behind such approach was that, for given trait, only about 100-1000 SNPs are truly expected to be associated, so

the probability that a given SNP (on a given platform) is one of them approximates to 500/number of SNPs printed on the platform. That gave us prior probabilities in the range of 10^{-4} to 10^{-5}, which was in accordance with previous studies [22].

Asymptotic Bayes factor was calculated for each SNP/trait association, according to the formula:

$$ABF_0 = \sqrt{\frac{V+W}{V}} exp\left(-\frac{z^2 W}{2(V+W)}\right) \quad (1)$$

where V is the estimated variance of the parameter θ, W is the variance of the prior probability of null hypothesis, $z^2 = \hat{\theta}^2/V$ is the usual Wald statistic, and $\exp(\theta)$ is estimated odds ratio [14]. Subscript "0" is there for this form of ABF summarizes the evidence for/against null hypothesis, whilst its reciprocal value $(1/\mathrm{ABF}_0)$ then supports the working hypothesis - the evidence for the association. The crucial property of the relationship between the power of the study (represented with V of the estimated parameter θ) and evidence pro H_1 is that, providing all other elements, except V, of the equation (1) are fixed, the evidence for H_1 would grow as the power decreases, but only until it reaches its maximum at $V = W/(z^2 - 1)$. After that, ABF for H_1 would decline, since the power is not sufficient to provide strong evidence. This contrasts strongly with the behavior of the *p*-values, where very small departures from H_0 would produce small *p*-values when the power is high [14].

Parameters of prior variances were set as advised by Kraft and Evangelou [23, 24]. The usage of ABF guaranteed comparability of the "amount of evidence" for each SNP/trait association across studies, and allowed for the filtering procedures that followed [16]. We assumed autosomal-dominant model of inheritance at this point for the sake of simplicity.

2.4 Redundancy Check

SNP/trait associations were checked for redundancy, and repeated SNP/trait associations, as well as the trait association with SNPs that are in LD (defined as $R^2>0.8$) and/or physically closer than 500bp were discarded, keeping only the SNP/trait pairs with the highest ABF within their "redundancy cluster". We used SNAP web based tool [25] for identification of those "proxy" SNPs, and the search was conducted over SNPs in 1000GenomesPilot1 data.

Overall, the data analysis was done using *R* language and environment [26], with additional packages from the Comprehensive *R* Archive Network (CRAN) and Bioconductor [27].

3 Results

3.1 Filtering the SNPs for Multigenic Risk Assessment

The initial raw genotype file consisted of approximately 700,000 SNPs. Of that number, around 600,000 passed the genotype quality control. Merging with the NHGRI

GWAS database yielded 9172 SNP/trait associations. In the next step we excluded all the SNPs where the subject did not have at least one risk allele, leading to 3195 SNP/trait associations.

When we filtered out non-European sample studies, non-replicated findings, and studies not reporting OR and CI, we were left with 438 SNP/trait associations. ***Figure 1*** shows a flow-chart of the filters applied and the number of SNP/trait associations left after each filter was applied.

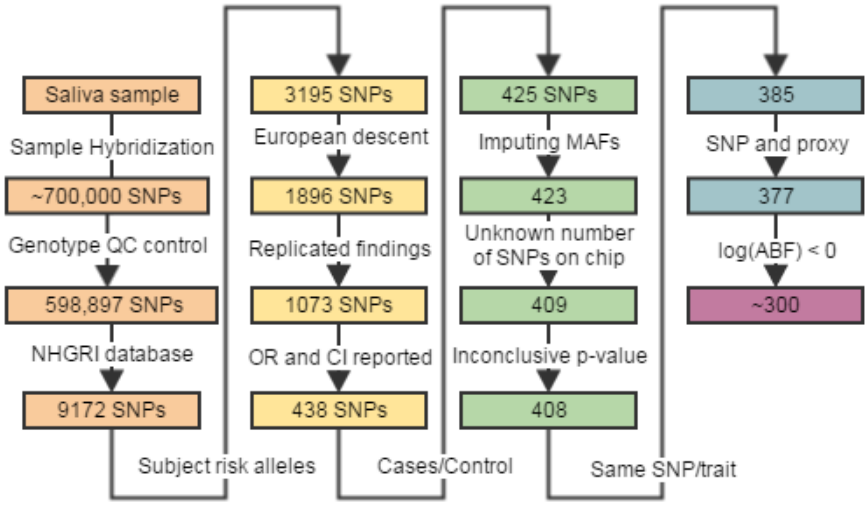

Fig. 1. Flow-chart of the filtering and selection process of SNPs

In the next step, missing MAFs were imputed where possible from the dbSNP database [5], and entries without reported/imputable MAFs were excluded. Rows with no inferable number of SNPs on the chip used, and samples with inconclusive p-value from the original studies were left out.

3.2 Asymptotic Bayes Factor

ABF was calculated for the remaining 408 SNP/trait associations. ***Figure 2*** shows the Manhattan plot of genomic coordinates of SNP/trait associations along the X-axis and the base 10 logarithm of the respective ABF on the Y-axis, together with the proposed cut-off values for $\log_{10}$(ABF) of 5 (dashed line). Twelve percent of SNP/trait associations had $\log_{10}$(ABF) values above cut-off, meaning there really is strong evidence to support those associations. Sixty seven percent had $\log_{10}$(ABF) between 5 and 0, meaning there is enough evidence that those SNP/trait associations should be taken into consideration for the multigenic risk assessment. It is interesting to note that as many as 21% of SNP/trait associations fell below the second cut-off level of 0,

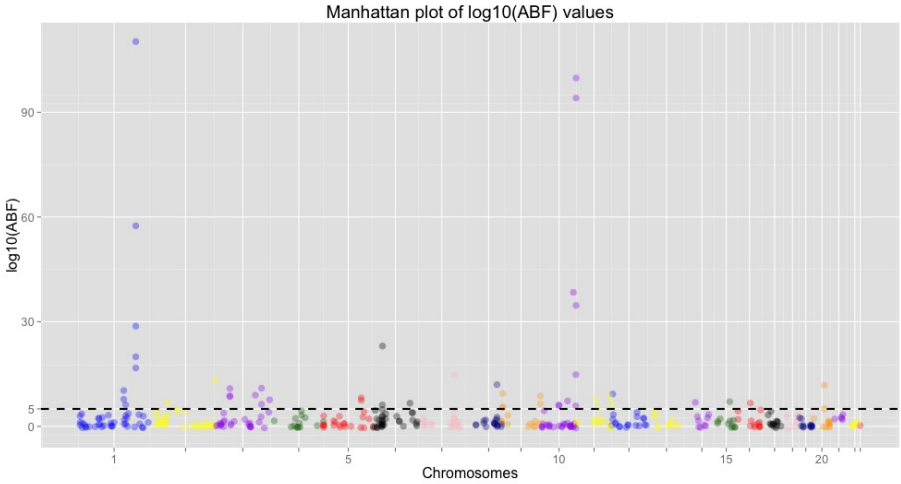

Fig. 2. Manhattan plot of $\log_{10}$(ABF) values

offering no firm evidence for those associations. Still, for illustration purposes these "unsupported" SNP/trait associations were not excluded from further analysis.

3.3 Removing Redundant SNP/Trait Association Based on the ABF

It is a common practice to explore indicated SNP/trait associations in several independent GWA studies, as to pursue verification through replication. However, when constructing the multigenic risk assessment profile only one estimation of OR and CI per SNP should be used in the calculation. A criteria should be defined to consistently select the "correct" OR and CI. ***Table 1*** reports 23 SNP/trait associations, each targeted by 2 studies, as well as the estimated odds ratios. The OR from the SNP/trait association with highest ABF from every pair was kept for the future analysis and bolded in the table. In this step the initial data.frame was reduced by another 23 rows.

Some studies explore the SNP/trait association using a proxy SNP instead of the original. A proxy SNP is usually defined as a SNP in LD with the original SNP (R2 > 0.8). ***Table 2*** reports original SNPs and their proxies for the same trait, along with the their respective odds ratios and confidence intervals. The OR and CI of the SNP/trait association selected using the ABF to remain in the analysis is bolded in the table.

Based on the previous steps, we have reduced the initial data.frame from 408 to 378 SNP/trait associations using the ABF. Having in mind that another 21% of SNP/trait association would not pass the strength of evidence criteria considering their ABF, our final data.frame would consist of only around 300 rows. These SNP/trait associations should pose a high enough reliability to be used downstream in any established form of multigenic risk assessment.

Table 1. Selected ORs and CI based on the ABF from pairs of studies targeting the same SNP/trait association

rsID	Trait	OR[CI] - 1	OR [CI] - 2
rs10737680	Age-related macular degeneration	3.11[2.76-3.51]	**2.43[2.39-2.47]**
rs2075650	Age-related macular degeneration	1.37[1.22-1.54]	**1.23[1.13-1.34]**
rs429608	Age-related macular degeneration	2.16[1.84-2.53]	**1.74[1.68-1.79]**
rs9621532	Age-related macular degeneration	**1.61[1.37-1.89]**	1.41[1.27-1.57]
rs2294008	Bladder cancer	**1.15[1.10-1.20]**	1.13[1.09-1.17]
rs4973768	Breast cancer	1.14[1.09-1.19]	**1.10[1.08-1.12]**
rs6651252	Crohn's disease	1.23[1.17-1.30]	**1.19[1.13-1.25]**
rs7702331	Crohn's disease	**1.12[1.07-1.17]**	1.09[1.05-1.13]
rs12700667	Endometriosis	**1.20[1.13-1.27]**	1.18[1.11-1.25]
rs6457327	Follicular lymphoma	1.69[1.43-2.00]	**1.47[1.27-1.72]**
rs9298506	Intracranial aneurysm	1.35[1.22-1.49]	**1.28[1.20-1.38]**
rs6719884	Myasthenia gravis	1.35[1.19-1.53]	**1.35[1.19-1.52]**
rs7078160	Orofacial clefts	**1.38[1.21-1.58]**	1.36[1.21-1.53]
rs7590268	Orofacial clefts	1.42[1.23-1.64]	**1.42[1.26-1.59]**
rs8001641	Orofacial clefts	**1.35[1.14-1.61]**	1.31[1.13-1.51]
rs11782652	Ovarian cancer	**1.24[1.16-1.33]**	1.19[1.12-1.26]
rs3018362	Paget's disease	1.52[1.36-1.70]	**1.45[1.34-1.56]**
rs11672691	Prostate cancer	**1.11[1.02-1.20]**	1.08[1.04-1.12]
rs10488631	Systemic sclerosis	**1.50[1.35-1.67]**	1.35[1.20-1.51]
rs7583877	Type 1 diabetes nephropathy	1.29[1.18-1.40]	**1.29[1.17-1.42]**
rs11739663	Ulcerative colitis	**1.15[1.09-1.21]**	1.07[1.03-1.12]
rs4728142	Ulcerative colitis	**1.10[1.07-1.14]**	1.07[1.03-1.11]
rs6017342	Ulcerative colitis	**1.23[1.19-1.27]**	1.20[1.15-1.26]

Based on the previous steps, we have reduced the initial *data.frame* from 408 to 378 SNP/trait associations using the ABF. Having in mind that another 21% of SNP/trait association would not pass the strength of evidence criteria considering their ABF, our final *data.frame* would consist of only around 300 rows. These SNP/trait associations should pose a high enough reliability to be used downstream in any established form of multigenic risk assessment.

Table 2. Selected ORs and CIs based on the ABF from paris of studies targeting the SNP/trait associations in LD

SNP-proxy pairs	R^2	Trait	OR[CI] - 1	OR[CI] - 2
rs6666258 rs13376333	1	Atrial fibrillation	1.18[1.13-1.23]	**1.52[1.40-1.64]**
rs646776 rs599839	0.90	Coronary heart disease	1.14[1.09-1.19]	**1.11[1.08-1.15]**
rs10490924 rs3793917	0.95	Age-related macular degeneration	**3.67[3.33-4.05]**	3.4[2.94-3.94]
rs10801555 rs1061147	1	Age-related macular degeneration	**2.33[2.08-2.63]**	1.4[1.32-1.48]
rs1329424 rs1061147	0.93	Age-related macular degeneration	**1.88[1.68-2.10]**	1.4[1.32-1.48]
rs1219648 rs2981579	0.97	Breast cancer	1.32[1.22-1.42]	**1.27[1.24-1.29]**
rs10801555 rs1329424	0.93	Age-related macular degeneration	2.33[2.08-2.63]	**1.88[1.68-2.10]**
rs4474514 rs995030	0.86	Testicular cancer	**3.07[2.29-4.13]**	2.26[1.95-2.61]

4 Discussion

Our work focuses on the possibility of performing multigenic risk assessment using the raw genotype data from a DTC genomic service, in this case the 23andMe web service [18], and utilizing only the publicly and freely available databases and data analysis tools. We present an overview of the advantages of using the ABF over *p*-values for the selection of "significant" SNP/trait associations, and we illustrate its multipurposeness in developing the heuristic for including only the meaningful SNP/trait associations in multigenic risk assessment.

The flow-chart (see ***Figure 1***) depicts how the number of useful SNP/trait associations rapidly reduces when using just the basic filters on descent, insist on replicated findings, and select only SNP/trait associations for qualitative traits (reporting ORs and CIs), even before checking for redundancies.

It is necessary to have in mind that all the SNP/trait associations reported in GWAS had been selected based on *p*-values only, and had passed stringent *p*-value criteria. Nevertheless, the use of *p*-values only has been characterized as problematic by many authors [14, 15], and an alternative has been offered: using $\log_{10}(ABF)$ and two cut-off criteria for the strength of evidence: $\log_{10}(ABF) > 5$ and $\log_{10}(ABF) > 0$ [15]. In ***Figure 2*** we demonstrate how the SNP/trait associations would rank under the ABF criteria, showing that more than 20% of SNP/trait associations would not satisfy the minimum criteria of log10(ABF) > 0, even though their *p*-values were significant enough for reporting in GWAS. It could be claimed that these SNP/trait

associations cannot be trusted enough to enter a multigenic risk score. Indeed, all the studies that had $\log_{10}(\text{ABF}) < 0$ were seriously overpowered, and the use of evidence from them in downstream analysis would be misleading, since, as argued above, such conclusion would not account for the adverse effect the high power has on the strength of evidence for the true association [14].

In the following step we examined the common situation where the same SNP/trait association is examined in different studies, each of these studies proposing different OR estimate. As shown in ***Table 1***, when selecting which study to take into account for multigenic risk assessment, the answer is often not the one with larger OR. That is paradoxal, but only at first glance, since we should remember that ABF is influenced more by the precision of the estimate of OR, not as much by OR itself [17]. The situation is similar when selecting between studies where the original and a proxy SNP/trait association has been explored, as shown in ***Table 2***. An important benefit in the method we use is the prevention of an overestimation of the "true" odds ratio.

The filtering and selection procedure (see ***Figure 1***) has "approved" only around 300 SNP/trait association for the further downstream use in multigenic risk assessment. We must take into consideration the fact that, at this point, only SNPs with sufficiently low p-values are indexed in the NHGRI database. The threshold *p*-value has been decided by each study's investigators, meaning the SNP/trait associations available to us had already been pre-filtered. Since a standard procedure for the selection of a threshold value might not be well defined [9, 10], there is an immense loss of information, and this might prove critical for the downstream procedures that are relying on the large initial set of SNPs to work on, such as building predictive models [28].

Furthermore, the Authors are aware that the decision to work only with GWAS studies which had the string "European" in the population description field is questionable. Our subject was from the Balkans, and as pointed out [29], that might imply quite different referent minor allele frequencies, LD patterns as well as disease prevalence and corrupt validity of risk assessment. Still, facing the scarcity of studies from the Balkan peninsula, we compromised on this.

Another issue was that is that the only output readily available to us, in terms of effect sizes, was the allelic odds ratio. Since risk ratios are considered more intuitive and precise [30], and can be approximated with odds ratios only under the "rare disease" assumption, we would suggest the NHGRI database report also the risk ratios, as well as the number of cases and controls with a specific risk allele, which would reduce the potential error due to data mining. Until it becomes common practice, we propose the use of an available R package ***orsk*** [31] for RR estimation from incomplete GWAS based data, with perhaps additional MAF based constraints when choosing the optimal solution for RR.

Even with a well established method for the filtering and selection of SNP/trait associations supported by enough evidence to be included into a multigenic risk score, a question remains on how a trait is defined, as different GWAS can explore the association of SNPs with a "general" trait, or just a characteristic (a "subtrait") falling under the general trait in question. A prospective direction of improvement in conveying the risk information might be MeSH term enrichment analysis of the genotype profile as a

whole, and some tools have already been developed for this aim [32, 33]. In this scenario, the proper selection of SNPs for which we are positive there is an association with a trait plays and even more crucial role.

In conclusion, the price of DTC genomic services has gone from several thousand down to several hundreds of dollars in the last 10 years, dropping even bellow 100$ in the past few years [1, 18], and it is easy to see the use of having a genetic report of a patient at your doctor's disposal. We have demonstrated the utility of the ABF in developing the criteria for selecting SNP/trait associations supported by strong enough evidence for the downstream multigenic risk assessment. Future steps to improve this process include the update of the information reported in the NHGRI database, using risk ratios instead of odds ratios, and MeSH term enrichment. An accurate genetic report is crucial to correctly interpret genetic information, but also to influence a range of health-promoting behaviors.

References

1. Mitchell, J.A., Fun, J., McCray, A.T.: Design of Genetics Home Reference: a new NLM consumer health resource. Journal of the American Medical Informatics Association 11(6), 439–447 (2004)
2. Goldstein, D.B.: Common genetic variation and human traits. N. Engl. J. Med. 360(17), 1696–1698 (2009)
3. Baptista, P.V.: Principles in genetic risk assessment. Ther. Clin. Risk Manag. 1(1), 15–20 (2005)
4. Regaldo, A.: How a Wiki Is Keeping Direct-to-Consumer Genetics Alive (2014), http://www.technologyreview.com/featuredstory/531461/how-a-wiki-is-keeping-direct-to-consumer-genetics-alive/
5. Sherry, S.T., et al.: dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 29(1), 308–311 (2001)
6. Landrum, M.J., et al.: ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research, gkt1113 (2013)
7. Yang, Q., et al.: Using lifetime risk estimates in personal genomic profiles: estimation of uncertainty. The American Journal of Human Genetics 85(6), 786–800 (2009)
8. Szoka, B. FDA Just Banned 23andMe's DNA Testing Kits, and Users Are Fighting Back (2013), http://www.huffingtonpost.com/berin-szoka/fda-just-banned-23andmes-_b_4339182.html
9. Jannot, A.-S., Ehret, G., Perneger, T.: P<5*10-8 has emerged as a standard of statistical significance for genome-wide association studies. Journal of Clinical Epidemiology (2015)
10. Zheng, G., Yuan, A., Jeffries, N.: Hybrid Bayes factors for genome-wide association studies when a robust test is used. Computational Statistics & Data Analysis 55(9), 2698–2711 (2011)
11. Zaykin, D.V., Zhivotovsky, L.A.: Ranks of Genuine Associations in Whole-Genome Scans. Genetics 171(2), 813–823 (2005)
12. Wang, L., et al.: Bayes Factor Based on a Maximum Statistic for Case-Control Genetic Association Studies. Journal of Agricultural, Biological, and Environmental Statistics 17(4), 568–582 (2012)
13. Kass, R.E., Raftery, A.E.: Bayes Factors. Journal of the American Statistical Association 90(430), 773–795 (1995)

14. Wakefield, J.: Bayes factors for genome-wide association studies: comparison with P-values. Genet. Epidemiol. 33(1), 79–86 (2009)
15. Sawcer, S.: Bayes factors in complex genetics. Eur. J. Hum. Genet. 18(7), 746–750 (2010)
16. Stephens, M., Balding, D.J.: Bayesian statistical methods for genetic association studies. Nat. Rev. Genet. 10(10), 681–690 (2009)
17. Wakefield, J.: A Bayesian measure of the probability of false discovery in genetic epidemiology studies. The American Journal of Human Genetics 81(2), 208–227 (2007)
18. 23andMe. 23andMe Web Service (2014), https://www.23andme.com/
19. Welter, D., et al.: The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research 42(D1), D1001–D1006 (2014)
20. Carey, V.: gwascat: structuring and querying the NHGRI GWAS catalog (2013)
21. Wickham, H.: stringr: modern, consistent string processing. The R Journal 2(2), 38–40 (2010)
22. Burton, P.R., et al.: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447(7145), 661–678 (2007)
23. Evangelou, E., Ioannidis, J.P.: Meta-analysis methods for genome-wide association studies and beyond. Nat. Rev. Genet. 14(6), 379–389 (2013)
24. Kraft, P., Zeggini, E., Ioannidis, J.P.: Replication in genome-wide association studies. Statistical Science: A Review Journal of the Institute of Mathematical Statistics 24(4), 561 (2009)
25. Johnson, A.D., et al.: SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 24(24), 2938–2939 (2008)
26. Team, R.C.: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2012) ISBN 3-900051-07-0
27. Gentleman, R.C., et al.: Bioconductor: open software development for computational biology and bioinformatics. Genome Biology 5(10), R80 (2004)
28. Wei, Z., et al.: From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet. 5(10), e1000678 (2009)
29. Barrett, J.: Why prediction is a risky business (2010), http://genomesunzipped.org/2010/08/why-prediction-is-a-risky-business.php
30. Cummings, P.: The relative merits of risk ratios and odds ratios. Arch. Pediatr. Adolesc. Med. 163(5), 438–445 (2009)
31. Wang, Z.: Converting Odds Ratio to Relative Risk in Cohort Studies with Partial Data Information. Journal of Statistical Software 55(5) (2013)
32. Nikaido, I., Tsuyuzaki, K., Morota, G.: meshr: Tools for conducting enrichment analysis of MeSH. R package version 1.2.4
33. Tsuyuzaki, K., et al.: How to use MeSH-related Packages (2014)