Article

The Role of SNP Interactions When Determining Independence of Novel Signals in Genetic Association Studies—An Application to *ARG1* and Bronchodilator Response

Ryan Walsh [1,†], Kirsten Voorhies [1,†], Merry-Lynn McDonald [2], Michael McGeachie [3], Joanne E. Sordillo [1], Christoph Lange [4], Ann Chen Wu [1] and Sharon M. Lutz [1,4,*]

1   Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA 02215, USA; Ryan_Walsh@harvardpilgrim.org (R.W.); Kirsten_Voorhies@harvardpilgrim.org (K.V.); rejoa@channing.harvard.edu (J.E.S.); Ann.Wu@childrens.harvard.edu (A.C.W.)
2   Division of Pulmonary, Allergy and Critical Care Medicine, Department of Medicine, University of Alabama at Birmingham, Birmingham, AL 35294, USA; mldonnelly@uabmc.edu
3   Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA; remmg@channing.harvard.edu
4   Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; clange@hsph.harvard.edu
*   Correspondence: smlutz@hsph.harvard.edu; Tel.: +1-617-867-4823; Fax: +1-617-867-4853
†   These authors contributed equally to this publication.

Abstract: Genome-wide association studies (GWAS) play a critical role in identifying many loci for common diseases and traits. There has been a rapid increase in the number of GWAS over the past decade. As additional GWAS are being conducted, it is unclear whether a novel signal associated with the trait of interest is independent of single nucleotide polymorphisms (SNPs) in the same region that has been previously associated with the trait of interest. The general approach to determining whether the novel association is independent of previous signals is to examine the association of the novel SNP with the trait of interest conditional on the previously identified SNP and/or calculate linkage disequilibrium (LD) between the two SNPs. However, the role of epistasis and SNP by SNP interactions are rarely considered. Through simulation studies, we examined the role of SNP by SNP interactions when determining the independence of two genetic association signals. We have created an R package on Github called gxgRC to generate these simulation studies based on user input. In genetic association studies of asthma, we considered the role of SNP by SNP interactions when determining independence of signals for SNPs in the *ARG1* gene and bronchodilator response.

Keywords: epistasis; SNP by SNP interactions; independence; GWAS; *ARG1*; bronchodilator response

1. Introduction

Genome-wide association studies (GWAS) play a critical role in identifying many loci for common diseases [1] as well as complex traits [2]. Over the past decade, the rapid increase in the number of GWAS provides an extraordinary opportunity to examine the potential impact of common and rare genetic variants on complex diseases [3]. Two independent GWAS may identify two different single nucleotide polymorphisms (SNPs) in the same gene or region that are both significantly associated with the trait of interest [4]. When determining whether these two signals are independent in GWAS, epistasis and SNP by SNP interactions are often not considered.

Epistasis is defined as the interaction between different genes or SNPs and refers to the departure from independence of the effect of different genetic loci on the disease or trait of interest [5,6]. There is a complex relationship with epistasis and linkage disequilibrium (LD) [7,8]. Multiple unobserved functional polymorphisms can lead to genotyped SNPs

that do not properly represent the causal variants [9] and high order LD can lead to spurious statistical epistatic associations [10]. Strong LD may suggest that detected epistasis between pairs of SNPs in different association studies need to be interpreted with caution [9,10].

Given epistasis, it is important to consider SNP by SNP interactions when determining whether genetic signals in GWAS are considered independent. For example, if an SNP is associated with the trait of interest in a GWAS, an investigator may want to determine if this SNP is an independent signal or purely the result of a correlated SNP that was previously associated with the trait of interest. In order to determine if the two signals are independent, the most popular approaches are to fit a regression of the trait of interest with the novel SNP conditional on the previously identified SNP and/or calculate LD between the two SNPs.

Through simulation studies, we examined the impact of SNP by SNP interactions when determining whether two signals are independent in a GWAS. We have created the R package, called gxgRC, which implements these simulation studies for user specified parameters. In addition, we considered the role of SNP by SNP interactions when determining independence of signals for SNPs in the *ARG1* gene and bronchodilator response in genetic association studies of asthma.

2. Materials and Methods

In the following simulation scenarios, we examined the impact of SNP by SNP interactions when determining the independence of two SNPs by regressing the trait of interest Y with the SNP X_1 conditional on the SNP X_2. We generated 1000 subjects for 5000 simulations using a significance level of $5*10^{-8}$. SNP X_1 is generated from a binomial distribution with a binary genetic coding (i.e., dominant or recessive model) and $P(X_1 = 1) = 0.5$. SNP X_2 is generated from a logistic regression based on X_1 such that:

$$logit(P(X_2 = 1)) = \gamma_0 + \gamma_1 X_1 \tag{1}$$

where $\gamma_0 = 0$ and $\gamma_1 = 0.3$. While X_1 and X_2 assume a binary genetic coding for simplicity, the results are generalizable to an additive genetic coding (i.e., $X_j = 0, 1, 2$ for j = 1, 2). The continuous outcome Y is generated from a normal distribution with a variance of 1 and a mean such that

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_I X_1 X_2 \tag{2}$$

where $\beta_0 = 0$, $\beta_1 = 0.3$ or 0, $\beta_2 = 0.3$ or 0, and β_I varies from 0.3 to 1 by 0.05. We considered additional simulation scenarios for different values for γ_0, γ_1 in Equation (1) and β_0, β_1, β_2 in Equation (2). However, we observed similar results to the presented simulations scenarios; therefore, the results are not shown here.

After the data were simulated using Equations (1) and (2), we then fit 3 algorithms and tested the following null hypotheses for each algorithm:

Algorithm 0: Fitting $E[Y] = \delta_0 + \delta_1 X_1$, we tested $H_0 : \delta_1 = 0$ to determine if the SNP X_1 is associated with the trait of interest Y.

Algorithm 1: Fitting $E[Y] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$, we tested $H_0 : \alpha_1 = 0$ to determine if the SNP X_1 is associated with the trait of interest Y conditional on the SNP X_2.

Algorithm 2: Fitting $E[Y] = \varphi_0 + \varphi_1 X_1 + \varphi_2 X_2 + \varphi_I X_1 X_2$, we tested $H_0 : \varphi_I = 0$ to determine if there is an interaction of the 2 SNPs on the trait of interest Y.

Then, we determined when the following scenarios were true.

Scenario 1: Rejected Algorithm 0 $H_0 : \delta_1 = 0$, Algorithm 1 $H_0 : \alpha_1 = 0$, and Algorithm 2 $H_0 : \varphi_I = 0$.

Scenario 2: Rejected Algorithm 0 $H_0 : \delta_1 = 0$ and Algorithm 1 $H_0 : \alpha_1 = 0$, but failed to reject H_0 for Algorithm 2.

Scenario 3: Rejected Algorithm 0 $H_0 : \delta_1 = 0$ and Algorithm 2 $H_0 : \varphi_I = 0$, but failed to reject H_0 for Algorithm 1.

Scenario 4: Rejected Algorithm 0 $H_0 : \delta_1 = 0$, but failed to reject H_0 for Algorithms 1 and 2.

Scenario 5: Failed to reject Algorithm 0 $H_0 : \delta_1 = 0$

3. Results

In Figure 1, $\beta_1 = 0.3$ and $\beta_2 = 0.3$ in Equation (2) for the plot on the left and $\beta_1 = 0$ and $\beta_2 = 0$ in Equation (2) for the plot on the right. For both plots and simulations, when a stronger interaction between the two SNPs is generated (i.e., β_I closer to 1 in Equation (2)), the majority of simulations concluded scenario 1: rejecting Algorithm 0 $H_0 : \delta_1 = 0$, Algorithm 1 $H_0 : \alpha_1 = 0$, and Algorithm 2 $H_0 : \varphi_I = 0$. These simulations show that there can be a significant association between the SNP X_1 and the trait of interest Y in Algorithm 0, and the SNP X_1 is still significantly associated with the trait of interest Y when conditioning on the SNP X_2 in Algorithm 1. However, there is a significant interaction between the two SNPs in Algorithm 2. This shows that if a researcher were to use Algorithm 1 to conclude that the two SNPs are independent since the SNP X_1 is significantly associated with the trait of interest Y conditional on the SNP X_2, a false conclusion would be reached because there is a significant interaction of the two SNPs on the trait Y in Algorithm 2 and as generated by the data using Equation (2), such that $\beta_I \neq 0$. These simulations demonstrate that it is not sufficient to consider independence of two genetic signals by considering Algorithm 1: $E[Y] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$ and testing $H_0 : \alpha_1 = 0$. One needs to also consider if there is a significant SNP by SNP interaction by fitting Algorithm 2: $E[Y] = \varphi_0 + \varphi_1 X_1 + \varphi_2 X_2 + \varphi_I X_1 X_2$ and testing $H_0 : \varphi_I = 0$.
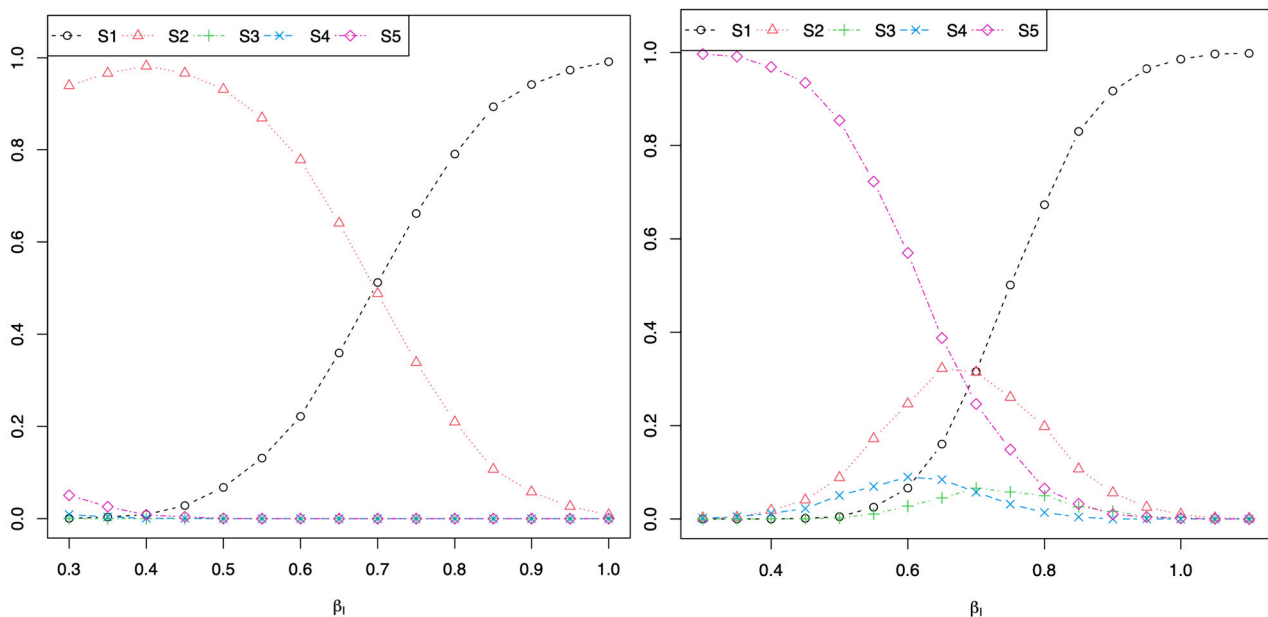

Figure 1. $\beta_1 = 0.3$ and $\beta_2 = 0.3$ in Equation (2) for the plot on the left and $\beta_1 = 0$ and $\beta_2 = 0$ in Equation (2) for the plot on the right, where both y-axes are the proportion of simulations where the null hypothesis was rejected. For the plot on the left, when the interaction was simulated to be weaker (i.e., β_I closer to 0.3 in Equation (2)), the majority of simulations concluded scenario 2: rejecting Algorithm 0 $H_0 : \delta_1 = 0$ and Algorithm 1 $H_0 : \alpha_1 = 0$, but failing to reject H_0 for Algorithm 2 (i.e., there was not a signification SNP by SNP interaction). For the plot on the right, when the interaction was simulated to be weaker (i.e., β_I closer to 0.3 in Equation (2)), the majority of simulations concluded scenario 5: failing to reject Algorithm 0 $H_0 : \delta_1 = 0$ (i.e., the SNP X_1 was not associated with the trait of interest Y). For both plots, when a stronger interaction between the 2 SNPs was simulated (i.e., β_I closer to 1 in Equation (2)), the majority of simulations concluded scenario 1: rejecting Algorithm 0 $H_0 : \delta_1 = 0$, Algorithm 1 $H_0 : \alpha_1 = 0$, and Algorithm 2 $H_0 : \varphi_I = 0$.

4. Data Analysis

To illustrate the effect of SNP by SNP interactions when determining conditional independence of genetic signals, we considered SNPs in chromosome 6 [*ARG1*], which has previously been associated with bronchodilator response (BDR) in asthma [11]. In the CAMP (*N* = 560) [12], CARE (*N* = 206) [13,14], and LODO (*N* = 126) [15] cohorts, we used weighted least squares regression to examine the conditional effect of four SNPs in the *ARG1* gene on BDR among subjects of European ancestry (total *N* = 892) adjusting for cohort, age, sex, body mass index (BMI) as a categorical variable (obese, overweight, vs. normal/underweight), and genetic ancestry. We picked these covariates based on other studies of BDR [16–18], as BDR has been found to differ depending on age [19,20], sex [21] and BMI [22]. Three of these SNPs are common variants that have previously been associated with BDR (rs2781659, rs2781663, rs2781665) [11] and one SNP (rs185631674) is a rare variant in the region.

Based on previous studies [11], rs2781659 had the most significant association with BDR in *AGR1*. In our cohorts, rs2781659 was associated with BDR adjusting for cohort, age, sex, BMI as a categorical variable (obese, overweight, vs. normal/underweight), and genetic ancestry (minor allele frequency (MAF) = 0.32, Beta = −0.15, sd = 0.05, *p*-value = 0.0037). In order to determine if the association of rs2781659 with BDR is independent of the other three SNPs, we considered the following algorithms:

Algorithm 1: $E[Y] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_C C^T$ and $H_0 : \alpha_1 = 0$

Algorithm 2: $E[Y] = \varphi_0 + \varphi_1 X_1 + \varphi_2 X_2 + \varphi_I X_1 X_2 + \varphi_C C^T$ and $H_0 : \varphi_I = 0$,

where *C* is a vector of the covariates: cohort, age, sex, BMI as a categorical variable (obese, overweight, vs. normal/underweight), and genetic ancestry.

As seen in Table 1, the association between rs2781659 and BDR was still significant when conditioning on rs2781663 (*p* = 0.003) and rs185631674 (*p* = 0.004) but not when conditioning on rs2781665 (*p* = 0.78). The SNP by SNP interaction on BDR was marginally significant for rs2781663 (*p* = 0.06) and rs2781665 (*p* = 0.06), and significant for rs185631674 (*p* = 0.03). However, rs2781659 is in LD with rs2781663 (r^2 = 0.995) and rs2781665 (r^2 = 0.891) and rs185631674 is a rare variant in a study with a small sample size (*N* = 892).

Table 1. Algorithm 1 considers the association of rs2781659 with bronchodilator response (BDR) conditioning on the SNPs in column 1 and Algorithm 2 considers the interaction of rs2781659 with the SNPs in column 1 on BDR. MAF denotes the minor allele frequency and SD denotes the standard deviation in the table below.

				Algorithm 1			Algorithm 2		
SNP	Position (Hg38)	r^2	MAF	Beta	SD	*p*-Value	Beta	SD	*p*-Value
rs2781663	131571207	0.995	0.32	−0.15	0.05	0.003	0.14	0.08	0.06
rs2781665	131572107	0.891	0.31	−0.18	0.63	0.78	0.14	0.08	0.06
rs185631674	131570984	0.002	0.001	−0.15	0.05	0.004	−1.75	0.83	0.03

This shows that if a researcher were to only consider Algorithm 1, where the association of rs2781659 with BDR is still significant conditioning on rs2781663 and rs185631674, the researcher would falsely conclude that there is more than one independent genetic signal with BDR. This false conclusion would not be reached with the SNP by SNP interaction considered in Algorithm 2 as well as the LD as measured by r^2. This also shows that special consideration needs to be given to rare variants.

5. Discussion

Through simulation studies and a data analysis of SNPs in *ARG1* with BDR, we demonstrate that it is not sufficient to consider independence of two genetic signals by considering Algorithm 1: $E[Y] = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$ and testing $H_0 : \alpha_1 = 0$. One needs to also consider whether there is a significant SNP by SNP interaction by fitting Algorithm 2: $E[Y] = \varphi_0 + \varphi_1 X_1 + \varphi_2 X_2 + \varphi_I X_1 X_2$ and testing $H_0 : \varphi_I = 0$ and/or calculating an

estimate of LD, such as r^2 or D'. Also, prior knowledge, for example protein–protein interactions or biological pathways, should be considered when examining SNP by SNP interactions [23].

There are some potential limitations for our simulation studies and data analysis. For the data analysis, all subjects are of European origin. A more diverse population could provide varying results, which should be considered for future analyses. Additionally, our data analysis considered genetic association studies of asthma but there is opportunity to explore the role of SNP by SNP interactions in conditional analyses by examining other diseases or traits. The sample size for the data analysis was relatively moderate ($n = 892$). Since the power to detect an interaction is substantially smaller than detecting a main effect, it should be noted that a larger sample size may be needed or the study may be underpowered to detect SNP by SNP interactions. While we have only presented two simulation studies here, we have created an R package on Github called gxgRC (https://github.com/SharonLutz/gxgRC (accessed on 12 December 2020)) [24] to generate similar simulations based on user input.

Understanding the role of epistasis and SNP by SNP interactions is important for the development of pharmacogenetic tests and personalized medicine. To date, studies in asthma pharmacogenetics have not resulted in clinical practice changes; however, exploring the role of SNP by SNP interactions has the potential to increase the likelihood of translatable findings.

References

1. Castaldi, P.J.; Dy, J.; Ross, J.; Chang, Y.; Washko, G.R.; Curran-Everett, D.; Williams, A.; Lynch, D.A.; Make, B.J.; Crapo, J.D.; et al. Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax* **2014**, *69*, 415–422. [CrossRef] [PubMed]
2. Nicolae, D.L.; Gamazon, E.; Zhang, W.; Duan, S.; Dolan, M.E.; Cox, N.J. Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS. *PLoS Genet.* **2010**, *6*, e1000888. [CrossRef]
3. Hindorff, L.A.; Sethupathy, P.; Junkins, H.A.; Ramos, E.M.; Mehta, J.P.; Collins, F.S.; Manolio, T.A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 9362–9367. [CrossRef] [PubMed]
4. Hind, J.; Lisboa, P.; Hussain, A.; Al-Jumeily, D. A Novel Approach to Detecting Epistasis using Random Sampling Regularisation. *IEEE/ACM Trans. Comput. Biol. Bioinform* **2019**. [CrossRef] [PubMed]
5. Cordell, H.J. Epistasis: What it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **2002**, *11*, 2463–2468. [CrossRef]
6. Li, P.; Guo, M.; Wang, C.; Liu, X.; Zou, Q. An overview of SNP interactions in genome-wide association studies. *Breifings Funct. Genom.* **2014**, *14*, 143–155. [CrossRef] [PubMed]
7. Platt, A.; Vilhjálmsson, B.J.; Nordborg, M. Conditions under which genome-wide association studies will be positively misleading. *Genetics* **2010**, *186*, 1045–1052. [CrossRef]
8. Zan, Y.; Forsberg, S.K.G.; Carlborg, Ö. On the Relationship Between High-Order Linkage Disequilibrium and Epistasis. *G3* **2018**, *8*, 2817–2824. [CrossRef]

9. Wood, A.R.; Tuke, M.A.; Nalls, M.A.; Hernandez, D.G.; Bandinelli, S.; Singleton, A.B.; Melzer, D.; Ferrucci, L.; Frayling, T.M.; Weedon, M.N. Another explanation for apparent epistasis. *Nature* **2014**, *514*, E3–E5. [CrossRef] [PubMed]

10. de Los Campos, G.; Sorensen, D.A.; Toro, M.A. Imperfect Linkage Disequilibrium Generates Phantom Epistasis (& Perils of Big Data). *G3* **2019**, *9*, 1429–1436. [CrossRef]

11. Litonjua, A.A.; Lasky-Su, J.; Schneiter, K.; Tantisira, K.G.; Lazarus, R.; Klanderman, B.; Lima, J.J.; Irvin, C.G.; Peters, S.P.; Hanrahan, J.P.; et al. ARG1 is a novel bronchodilator response gene: Screening and replication in four asthma cohorts. *Am. J. Respir. Crit. Care Med.* **2008**, *178*, 688–694. [CrossRef] [PubMed]

12. The Childhood Asthma Management Program (CAMP). Design, rationale, and methods. Childhood Asthma Management Program Research Group. *Control. Clin. Trials* **1999**, *20*, 91. [CrossRef]

13. Sorkness, C.A.; Lemanske, R.F., Jr.; Mauger, D.T.; Boehmer, S.J.; Chinchilli, V.M.; Martinez, F.D.; Strunk, R.C.; Szefler, S.J.; Zeiger, R.S.; Bacharier, L.B.; et al. Long-term comparison of 3 controller regimens for mild-moderate persistent childhood asthma: The Pediatric Asthma Controller Trial. *J. Allergy Clin. Immunol.* **2007**, *119*, 64–72. [CrossRef]

14. Zeiger, R.S.; Szefler, S.J.; Phillips, B.R.; Schatz, M.; Martinez, F.D.; Chinchilli, V.M.; Lemanske, R.F., Jr.; Strunk, R.C.; Larsen, G.; Spahn, J.D.; et al. Response profiles to fluticasone and montelukast in mild-to-moderate persistent childhood asthma. *J. Allergy Clin. Immunol.* **2006**, *117*, 45–52. [CrossRef] [PubMed]

15. American Lung Association Asthma Clinical Research Centers. Clinical trial of low-dose theophylline and montelukast in patients with poorly controlled asthma. *Am. J. Respir. Crit. Care Med.* **2007**, *175*, 235–242. [CrossRef] [PubMed]

16. Spear, M.L.; Hu, D.; Pino-Yanes, M.; Huntsman, S.; Eng, C.; Levin, A.M.; Ortega, V.E.; White, M.J.; McGarry, M.E.; Thakur, N.; et al. A genome-wide association and admixture mapping study of bronchodilator drug response in African Americans with asthma. *Pharmacogenom. J.* **2019**, *19*, 249–259. [CrossRef] [PubMed]

17. Mak, A.C.Y.; White, M.J.; Eckalbar, W.L.; Szpiech, Z.A.; Oh, S.S.; Pino-Yanes, M.; Hu, D.; Goddard, P.; Huntsman, S.; Galanter, J.; et al. Whole-Genome Sequencing of Pharmacogenetic Drug Response in Racially Diverse Children with Asthma. *Am. J. Respir. Crit. Care Med.* **2018**, *197*, 1552–1564. [CrossRef]

18. Sordillo, J.E.; McGeachie, M.; Lutz, S.M.; Lasky-Su, J.; Tantisira, K.; Tsai, C.H.; Dahlin, A.; Kelly, R.; Wu, A.C. Longitudinal analysis of bronchodilator response in asthmatics and effect modification of age-related trends by genotype. *Pediatr. Pulmonol.* **2019**, *54*, 158–164. [CrossRef]

19. Wu, A.C.; Tantisira, K.; Li, L.; Schuemann, B.; Weiss, S.; Childhood Asthma Management Program Research Group. Repeatability of response to asthma medications. *J. Allergy Clin. Immunol.* **2009**, *123*, 385–390. [CrossRef]

20. McGeachie, M.J.; Yates, K.P.; Zhou, X.; Guo, F.; Sternberg, A.L.; Van Natta, M.L.; Wise, R.A.; Szefler, S.J.; Sharma, S.; Kho, A.T.; et al. Patterns of Growth and Decline in Lung Function in Persistent Childhood Asthma. *N. Engl. J. Med.* **2016**, *374*, 1842–1852. [CrossRef] [PubMed]

21. Tuomisto, L.E.; Ilmarinen, P.; Lehtimäki, L.; Tommola, M.; Kankaanranta, H. Immediate bronchodilator response in FEV1 as a diagnostic criterion for adult asthma. *Eur. Respir. J.* **2019**, *53*. [CrossRef] [PubMed]

22. McGarry, M.E.; Castellanos, E.; Thakur, N.; Oh, S.S.; Eng, C.; Davis, A.; Meade, K.; LeNoir, M.A.; Avila, P.C.; Farber, H.J.; et al. Obesity and bronchodilator response in black and Hispanic children and adolescents with asthma. *Chest* **2015**, *147*, 1591–1598. [CrossRef] [PubMed]

23. Fang, G.; Wang, W.; Paunic, V.; Heydari, H.; Costanzo, M.; Liu, X.; Liu, X.; Vandersluis, B.; Oately, B.; Steinbach, M.; et al. Discovering genetic interactions bridging pathways in genome-wide association studies. *Nat. Commun.* **2019**, *10*, 1–18. [CrossRef] [PubMed]

24. GitHub—SharonLutz/gxgRC. Available online: https://github.com/SharonLutz/gxgRC (accessed on 12 December 2020).