Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

Genetics and Population Analysis

GWAlpha: Genome-Wide estimation of additive effects (Alpha) based on trait quantile distribution from pool-sequencing experiments

Alexandre Fournier-Level[1]*, Charles Robin[1] and David J. Balding[1,2]

[1]School of BioSciences and Centre for Systems Genomics, The University of Melbourne, Parkville 3010, Australia, [2]School of Mathematics and Statistics, The University of Melbourne, Parkville 3010, Australia.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Sequencing pools of individuals (Pool-Seq) is a cost-effective way to gain insight into the genetics of complex traits, but as yet no parametric method has been developed to both test for genetic effects and estimate their magnitude. Here, we propose GWAlpha, a flexible method to obtain parametric estimates of genetic effects genome-wide from Pool-Seq experiments.

Results: We showed that GWAlpha powerfully replicates the results of GWAS from model organisms. We perform simulation studies that illustrate the effect on power of sample size and number of pools and test the method on different experimental data.

Availability: GWAlpha is implemented in python, designed to run on Linux operating system and tested on Mac OS. It is freely available at https://github.com/aflevel/GWAlpha.

Contact: afournier@unimelb.edu.au

Supplementary information: Manual available at https://github.com/aflevel/GWAlpha/raw/master/GWAlpha_manual.pdf.

1 Introduction

Progress in sequencing technology has enabled genome-wide association studies (GWAS) in potentially any organism with segregating genetic diversity, even those lacking prior genomic resources. Nonetheless, developing association panels involving extensive genotyping and maintaining living genotypes to repeatedly measure phenotypes remains limiting in most organisms. Individual-based phenotyping performed on fully sequenced genotypes such as inbred lines is not always technically possible or biologically relevant. As an alternative, the sequencing of pools of individuals contrasting for a given phenotype is straightforward (Schlötterer et al., 2014). This requires no prior knowledge on the genetic make-up of the sampled population and reduces the sequencing time and effort to a limited set of pools representing the diversity of the trait. However, even if a substantial corpus of work has focused on the calling of genetic variants in Pool-Seq experiments or obtaining robust allele frequency estimates (Cao et al., 2014; Edwards et al., 2012), statistical models to measure the magnitude of the genetic effects underlying complex trait variation are scarce. The existing methods have predominantly applied non-parametric contingency tests contrasting the count of alleles across pools (Kofler et al., 2011; Magwene et al., 2011; Yang et al., 2015). There is a lack of test statistics that measure the effect size of genome-wide polymorphisms in the Pool-Seq context. Here, we propose a flexible parametric test to infer the size of genetic effects from the allele frequency in pools covering the entire range of variation for a quantitative trait.

2 Model

Consider a population of individuals measured for a given trait Y and binned into k pools based on their trait values. $Y \in [\min Y; \max Y]$ is first inverse-quantile transformed into $Y' \in [0;1]$. For each pool i encompassing all observations with trait values in $[y'_{i-1}; y'_i]$, we observe $Q_k = \{q_1, ..., q_k\}$ the distribution of a specific allele across the k pools summing to 1. GWAlpha estimates the parameters of the distribution of the q_i, both for a specified allele and for all alternative alleles at the locus, combined into a single allelic class. It assumes

the Beta distribution, which is tractable yet flexible enough to capture the relevant features of the true distribution. Specifically, we assume that the transformed phenotype Y' associated with a random copy of the allele has a Beta distribution with parameters $\Theta = (\theta_1; \theta_2)$. It follows that the expected allele fraction for the i-th pool is: $Pr(Allele|\Theta) = cdf_{Beta}(y'_i, \Theta) - cdf_{Beta}(y'_{i-1}, \Theta)$. The parameters Θ of the distribution of the allele can be alternatively estimated by least-square estimation solving: $\underset{\Theta}{argmin} \sum_{i=1}^{k} (q_i - Pr(Allele|\Theta))^2$ or maximising the likelihood: $L(\Theta|Q_k) \propto \prod_{i=1}^{k} f(q_i|\Theta)$. The distribution of the alternative allele states is modeled identically. Finally, the test statistic is obtained as: $\hat{\alpha} = W(\frac{\hat{\mu}_{Allele} - \hat{\mu}_{Alternative}}{\sigma_Y})$ which compares how $\hat{\mu}_{Allele}$ the mean of the estimated distribution for the allele deviates from $\hat{\mu}_{Alternative}$ the mean of the estimated distribution of the alternative states scaled by σ_Y the observed standard deviation of the trait, and where $W = 2\sqrt{p_{Allele}(1 - p_{Allele})}$ is a default penalisation for low allele frequency which can be set to 1 (no penalisation). The distribution of the α's is modeled using a normal distribution with parameters estimated through maximum-likelihood; the cumulative density function of this normal distribution is used to calculate empirical p-values.

3 Results and Discussion

3.1 Simulation

3 The GWAlpha method was tested by simulating 10000 diallelic SNP genotypes for either 100, 200 or 500 individuals and randomly selecting either one or ten SNPs as additive QTL to generate a phenotype with heritability h^2=0.5. The data were then converted into synchronised genotype files (SYNC files) by assigning the individuals into two to ten pools based on phenotypes and assuming a 40X Poisson-distributed coverage. Each simulation condition was replicated 500 times, and GWAlpha was performed using maximum-likelihood estimation.

4 With a single SNP affecting the phenotype, GWAlpha recaptured the causal SNP as the top candidate in over 95% of the simulations when including three pools or more. When ten SNPs affect the phenotype with random effect sizes, a minimum of 26% of the causal SNPs (average over 500 replicates) were recaptured in the top 100 candidates when 100 individuals were sampled in two pools; and 58% with 500 individuals in five pools (Suppl. Data 1).

5 The results show that a large sample size which is non-limiting in pool-seq yield accurate estimates of alpha (Suppl. Data 2); leading to the specific detection of associated SNPs (Suppl. Data 3) and showing limited bias due to allele frequency and coverage differences (Suppl. Data 4). With a sample size of 500 individuals, the best suggested pool number was determined to be five.

3.2 Data Analysis

6 We tested maximum-likelihood-based GWAlpha with four GWAS datasets in two different organisms (Brachi et al., 2015; Magwire et al., 2012; Baxter et al., 2010; Battlay et al., 2016), converting the genotypes and phenotypes into synchronised genotype files with a 200X uniform coverage (Suppl. Data 5). For the *Drosophila melanogaster* datasets, including within-population variability with limited population structure, the two associations identified in Magwire et al. (2012) were recaptured as the two top peaks with four sequencing pools (Suppl. Figure 5). The major association with the *Cyp6g1* gene identified in Battlay et al. (2016) was recaptured with five pools as second top SNP. For the *Arabidopsis thaliana* datasets, gathering across-population variability, the major association with the *AtHKT1;1* gene identified in Baxter et al. (2010) was recaptured as the top candidate using five pools (Suppl. Figure 6). The second and third highest associations from Brachi et al. (2015) were identified with only three pools while the top association linked to SNPs with frequency <10% was not detected. The results obtained with GWAlpha are consistent with those of GWAS, showing satisfactory power to detect associated SNPs. However, higher correlation among SNPs led to potentially spurious association peaks in the more structured *A. thaliana* sample, making GWAlpha more suitable to identify genetic variants segregating within unstructured populations.

4 Conclusion

7 GWAlpha provides a parametric estimation of genetic effect, enabling straightforward comparison across populations or phenotypes. It provides a middle ground between costly individual-based GWAS and two extreme pools contingency test. GWAlpha outperforms contingency tests in most scenarios, unless replicated experiments enable the use of the more powerful Cochran-Mantel-Haenszel test. The generic input format and flexible python implementation allow a straightforward integration to other genomic analysis pipelines with reasonable speed and memory usage (Suppl. Data 6) and is suitable for any organism, irrespective of the extent of resources available.

5 Acknowledgements

8 We thank the Victorian Life Science Computation Initiative for its support.

6 Funding

9 Work funded by a Human Frontier in Science Program fellowship to AFL.

10 *Conflict of Interest:* none declared.

7 References

Battlay,P. et al. (2016) Genomic and Transcriptomic Associations Identify a New Insecticide Resistance Phenotype for the Selective Sweep at the Cyp6g1 Locus of Drosophila melanogaster. G3 Genes|Genomes|Genetics, g3.116.031054.

Baxter,I. et al. (2010) A Coastal Cline in Sodium Accumulation in Arabidopsis thaliana Is Driven by Natural Variation of the Sodium Transporter AtHKT1;1. Plos Genet., 6.

Brachi,B. et al. (2015) Coselected genes determine adaptive variation in herbivore resistance throughout the native range of Arabidopsis thaliana. Proc. Natl. Acad. Sci., 112, 4032–4037.

Cao,C.-C. et al. (2014) Quantitative group testing-based overlapping pool sequencing to identify rare variant carriers. BMC Bioinformatics, 15, 195.

Edwards,M.D. et al. (2012) High-resolution genetic mapping with pooled sequencing. BMC Bioinforma. 2012 136, 13, 9828–9832.

Kofler,R. et al. (2011) PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics, 27, 3435–6.

Magwene,P.M. et al. (2011) The Statistics of Bulk Segregant Analysis Using Next Generation Sequencing. PLoS Comput. Biol., 7, e1002255.

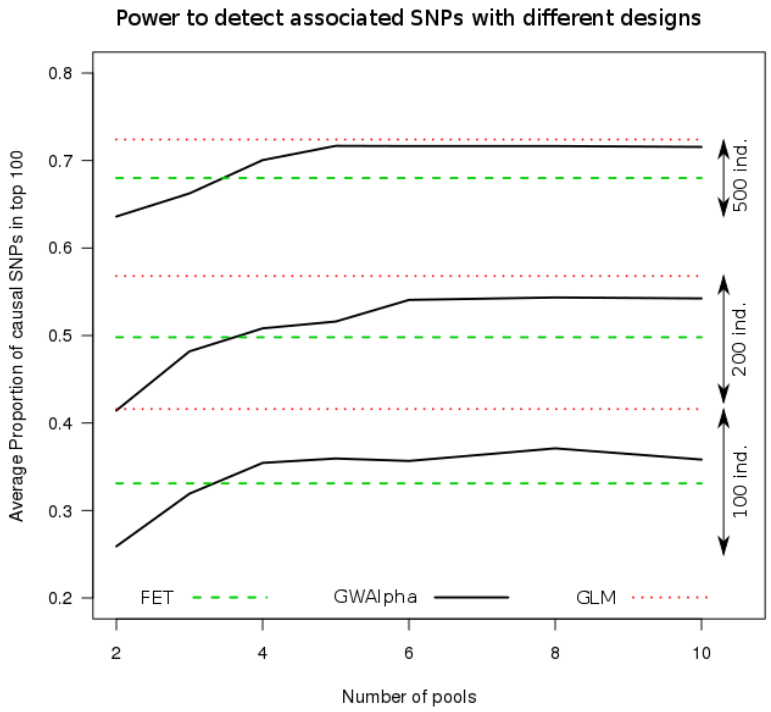
Magwire,M.M. et al. (2012) Genome-Wide Association Studies Reveal a Simple Genetic Basis of Resistance to Naturally Coevolving Viruses in Drosophila melanogaster. PLoS Genet., 8, e1003057.

Schlötterer,C. et al. (2014) Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. Nat. Rev. Genet., 15, 749–763.

Yang,J. et al. (2015) Extreme-phenotype genome-wide association study (XP-GWAS): a method for identifying trait-associated variants by sequencing pools of individuals selected from a diversity panel. Plant J., 84, 587–596.

Supplementary Data 1: Power analysis

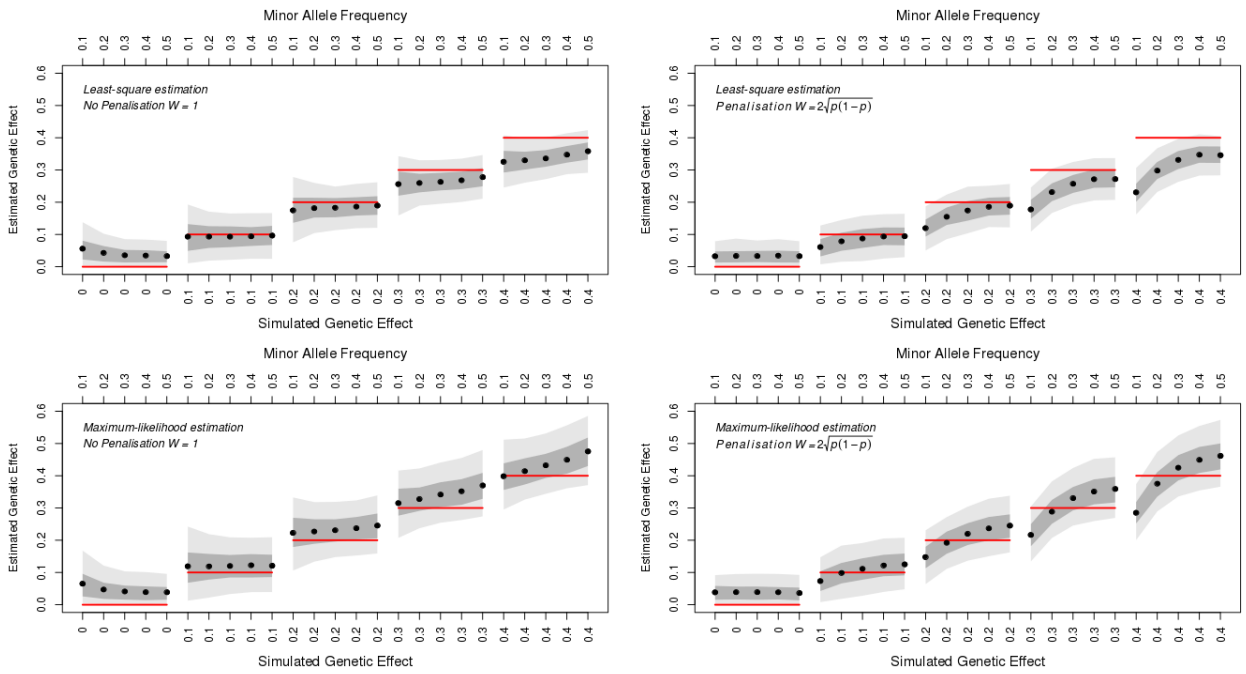
Power was assessed as the proportion of causal SNPs detected in the top 100 associations using 500 simulations for each combination of pools and individuals number and compared to a Fisher's Exact Test (FET) performed using two extreme pools each covering 1/3 of the distribution (situation where power is maximal, Magwene *et al.* 2010) and to a General Linear Models (GLM) using individual sequencing data and testing SNP effect through ANOVA. In most cases, GWAlpha showed significantly improved power to identify causal SNPs compared to FET (Suppl. Figure 1). For a population size of 500 individuals, the power of GWAlpha using 5 sequencing pools almost matched the one of GLM.



Suppl Figure 1: Power to detect causal SNPs among the top 100 associations using GWAlpha with varying number of pools (x-axis), a General Linear Model based on individual sequencing data and a Fisher's Exact Test between two extreme pools each representing 1/3 of the trait distribution.

Supplementary Data 2: Accuracy of the alpha estimates

The accuracy of the alpha estimates was assessed using each of the proposed estimation methods (Least-square or Maximum-likelihood, with or without penalisation) for a range of simulated genetic effects from 0 to 0.4 units of standard deviation and minor allele frequencies from 0.1 to 0.5. Each condition was simulated 1000 times with 500 individuals distributed in 5 pools and an environmental/error standard deviation for the trait of 0.5. The results are presented in Suppl. Figure 2.



Suppl. Figure 2: Accuracy of the estimated genetic effects (alphas) for different simulated values of genetic effects and minor allele frequencies. Red lines represent the actual simulated genetic effect; black dots represent the mean estimated effect; the dark grey envelop the top/bottom 25%-tile; and the light grey envelop the top/bottom 5%-tile.

The Root Mean Squared Error values across all simulated genetic effects and allele frequency suggested the Maximum-likelihood estimation with penalisation is the most accurate method (mean RMSE=0.167, Suppl.

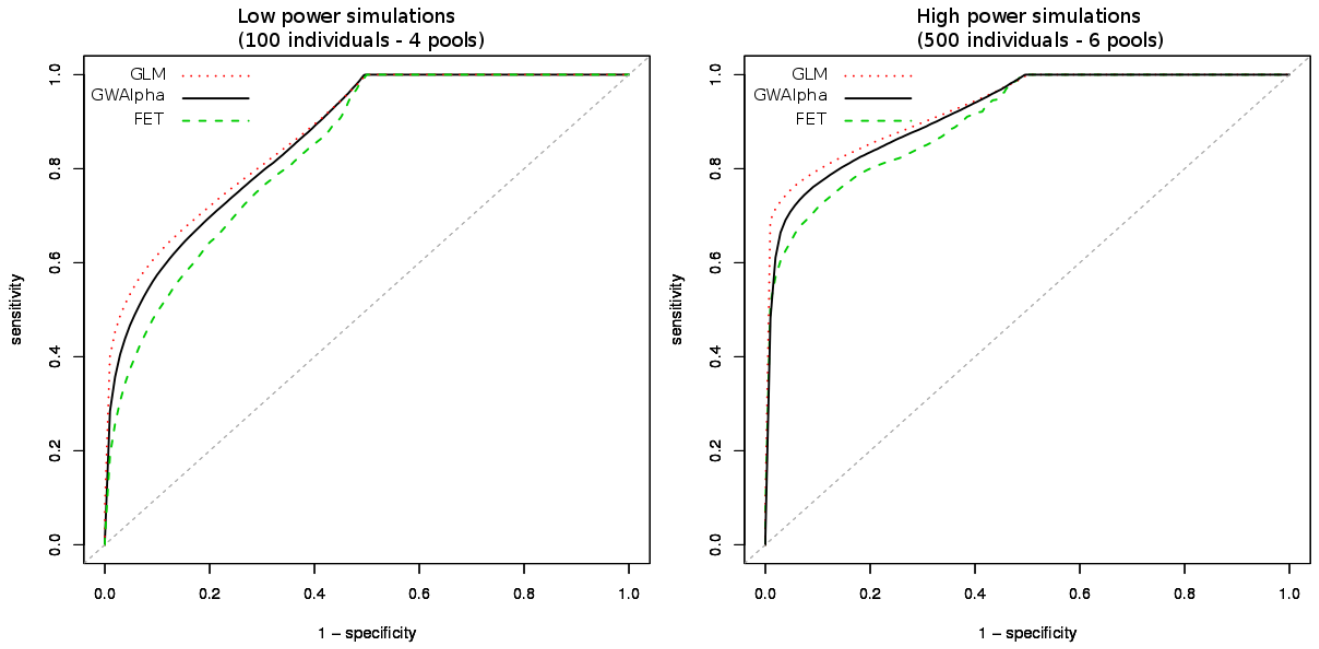
			No Penalisation					With Penalisation				
			Minor Allele frequency									
			0.1	0.2	0.3	0.4	0.5	0.1	0.2	0.3	0.4	0.5
Least-square	Genetic Effect	0	0.06	0.16	0.265	0.366	0.467	0.071	0.168	0.267	0.367	0.468
		0.1	0.056	0.116	0.211	0.308	0.405	0.054	0.127	0.216	0.309	0.407
		0.2	0.097	0.05	0.124	0.217	0.313	0.045	0.061	0.133	0.218	0.313
		0.3	0.166	0.074	0.055	0.138	0.226	0.089	0.053	0.059	0.134	0.231
		0.4	0.231	0.137	0.053	0.065	0.147	0.138	0.107	0.052	0.065	0.158
						mean	0.18				mean	0.172
Maximum -likelihood	Genetic Effect	0	0.062	0.157	0.26	0.362	0.462	0.067	0.163	0.262	0.362	0.464
		0.1	0.071	0.1	0.187	0.282	0.382	0.051	0.112	0.195	0.283	0.378
		0.2	0.141	0.064	0.086	0.17	0.26	0.069	0.05	0.095	0.172	0.26
		0.3	0.225	0.14	0.071	0.076	0.144	0.128	0.104	0.063	0.076	0.152
		0.4	0.306	0.223	0.144	0.078	0.072	0.193	0.184	0.138	0.078	0.073
						mean	0.181				mean	0.1669

Suppl. Table 1: RMSE of each estimation method for a range of simulated genetic effects and minor allele frequency.

Table 1) and is also the most accurate under the null hypothesis of no genetic effect. However, each estimation method showed to have better accuracy under specific conditions as reported in Suppl. Table 1.

Supplementary Data 3: False positive and false negative detection rates

The difference in rate of false positives and false negative was assessed for the General Linear Model (GLM, based on individual genotype data) and the Fisher's Exact Test (FET, based on two extreme pools) using Receiver-Operator Characteristic (ROC) curves (Suppl. Figure 3) and compared in terms of Area Under the Curve (AUC). We used one set of 5000 simulations in low power conditions (100 individuals in 4 pools) compared to 5000 simulations in high power conditions (500 individuals in 6 pools). Each simulation was generated with 10 random genetic effects among 10000 SNPs contributing a total heritability of h^2=0.5. The AUC obtained with GWAlpha ranged from 0.86 under the low power conditions to 0.93 under high power conditions and complemented the results of the power analysis showing GWAlpha as an intermediate between the General Linear Model and the Fisher's Exact Test, converging to the power of GLM test pending a sufficiently big sample size.



Suppl. Figure 3: ROC curves for low and high power simulation for three classes of test (General Linear Model, GWAlpha and Fisher's Exact Test, respectively).

The ROC curve showed that pending a big sample size, GWAlpha is not significantly less specific than GLM (for a threshold of 100 most associated SNPs using a t-test, p-val>0.05). GWAlpha is thus not more prone to false positive than GLM. Increasing the sample size from 100 to 500 individual strongly increased sensitivity without decreasing significantly specificity therefore not inflating the false positive rate.

Supplementary Data 4: Effect of allele frequency and sequencing coverage on power

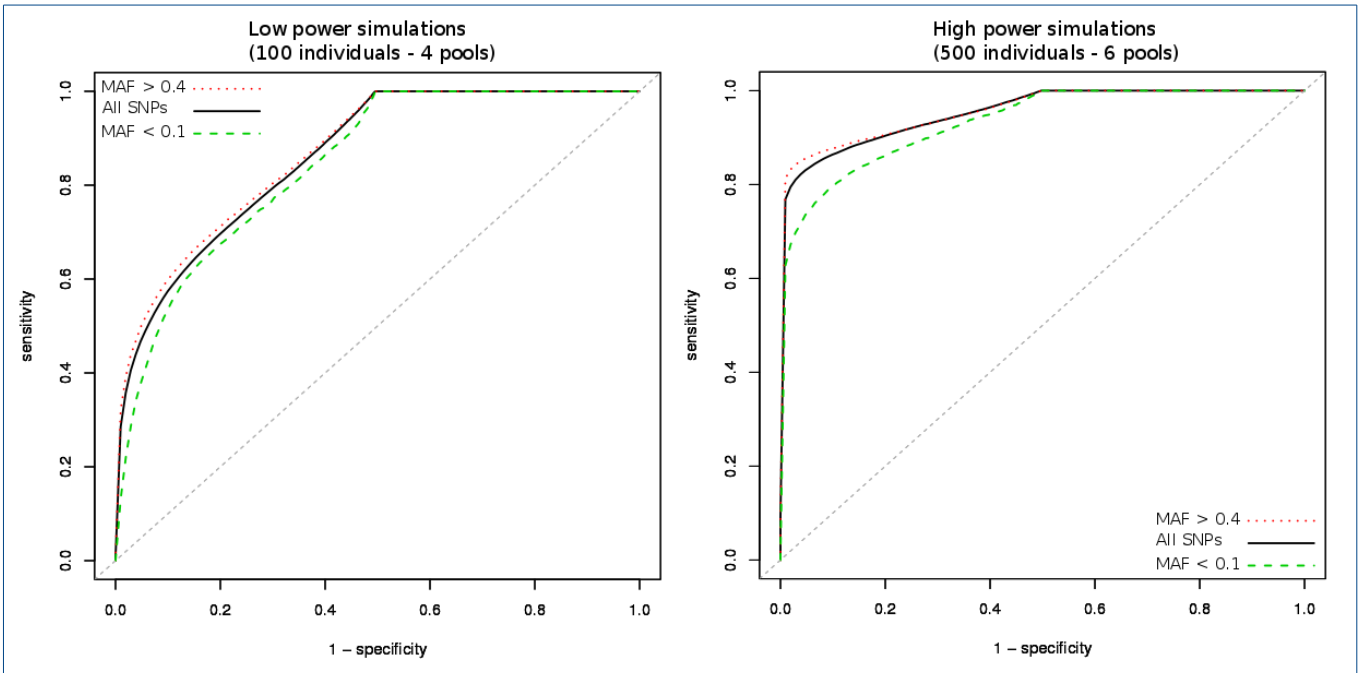
5000 simulations were performed under the low power and high power conditions described in Suppl. Data 3 but increasing the variance in sequencing coverage of the simulated data by generating a log-normal distributed coverage of mean 40 and standard deviation 40. The effect of the minor allele frequency and the sequencing coverage of the minor allele was tested on the association rank of the causal SNP obtained through GWAlpha using the following linear model: rank = allele frequency + coverage + allele frequency * coverage; and modeling all variables as fixed effects. The ANOVA tables for the high and low power conditions models are presented in Suppl. Table 2. Since both models have identical degrees of freedom (3 for the parameters and 49997 for the residual), the test statistics can be directly compared.

	Low power (100 individuals in 4 pools)			High power (500 individuals in 6 pools)		
Adj. R2	0.002929			0.003793		
	Estimate	Std. Error	Pr(>|t|)	Estimate	Std. Error	Pr(>|t|)
Intercept	2580.718	55.152	<2E-016	1155.797	40.454	<2E-016
MAF	-1450.942	182.12	1.66E-015	-1442.701	379.762	0.000145
COV	-3.915	1.806	0.0301	-5.856	1.76	0.000877

Suppl. Table 2: ANOVA table for the effect of minor allele frequency and coverage on the ranking of causal SNPs determined through GWAlpha.

Based on the significance of their individual effects, both allele frequency and coverage have an effect on the ranking of causative SNPs. Both these effects are in direction of higher allele frequency and greater coverage lead to lower (more significant) ranking. However, the small R2 values support the fact that these effects are marginal in explaining causative SNP rankings. In addition, the estimated effects of frequency and coverage were similar for both high and low power conditions.

We also analysed the capacity of the GWAlpha model to identify true positives and rule out false positive using a ROC approach for different allele frequencies (Suppl. Figure 4). While relatively little differences were observed under the low power conditions with AUC values ranging from 0.84 to 0.86, the high power conditions showed



Suppl. Figure 4: ROC curves for low and high power simulation for all SNPs (solid black line), SNPs with minor allele frequency (MAF) greater than 40% (dotted red) or lesser than 10% (green dashed).

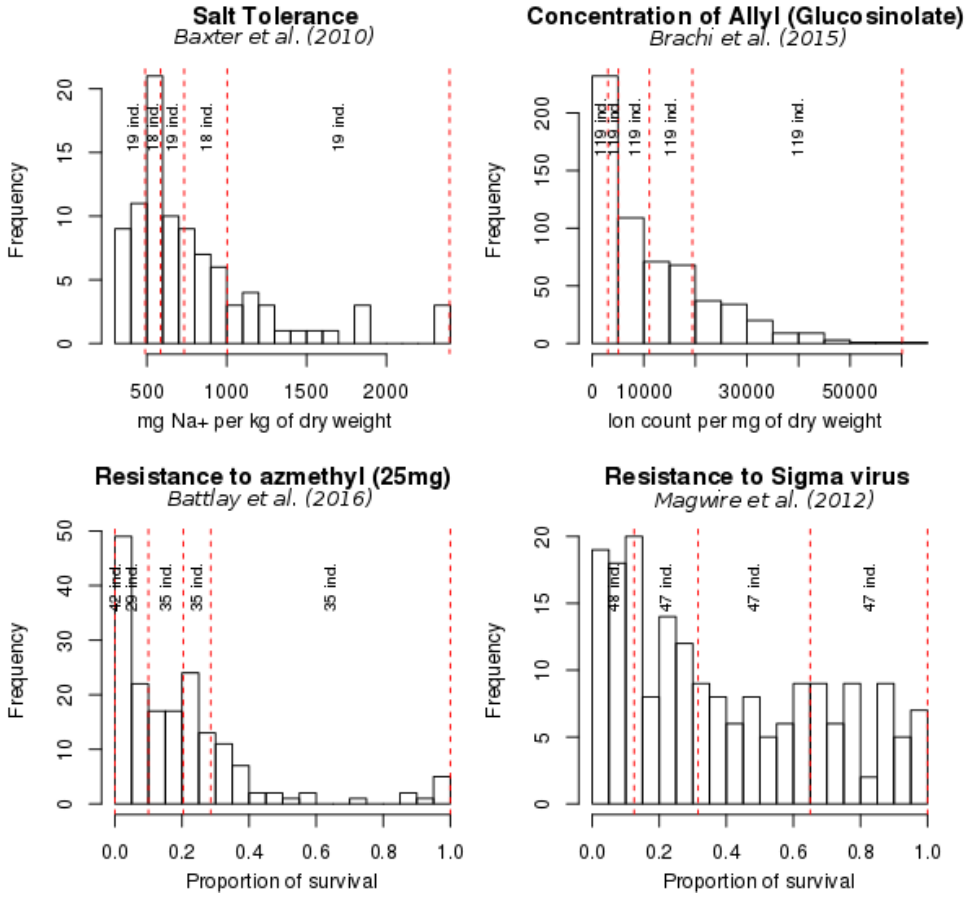
significant differences with AUC values ranging from 0.93 to 0.96. Under the high power conditions, low minor allele frequencies significantly decreased sensitivity to detect true positive as well as decreasing specificity (for a threshold of 100 most associated SNPs using a t-test, p-val>0.05). Moreover, when comparing the analysis in Suppl. Figure 3 and 4 which only differ by an increased variance in sequencing coverage in Suppl. Figure 4, an increase in sensitivity and in AUC was observed, suggesting high variance in sequencing depth is actually beneficial to the overall detection power. We further compared the detection power of GWAlpha with respect to GLM (individual based) and FET (2 pools representing 1/3 of the distribution each) either for rare (< 0.1) or balanced (>0.4) frequencies. We used the same framework as in Suppl. Data 1 but using 5000 simulations with increased coverage variance as described above and scored the proportion of causative SNPs ranked among the top 100 SNPs in the simulations. The results confirmed a loss of power to detect low frequency alleles, however this loss of power was less pronounced for GWAlpha compared to the alternative methods.

	Low power			High power		
	All	Low frequency (<0.1)	High frequency (>0.4)	All	Low frequency (<0.1)	High frequency (>0.4)
GWAlpha	0.28	0.13	0.39	0.77	0.63	0.81
GLM	0.4	0.11	0.45	0.78	0.55	0.83

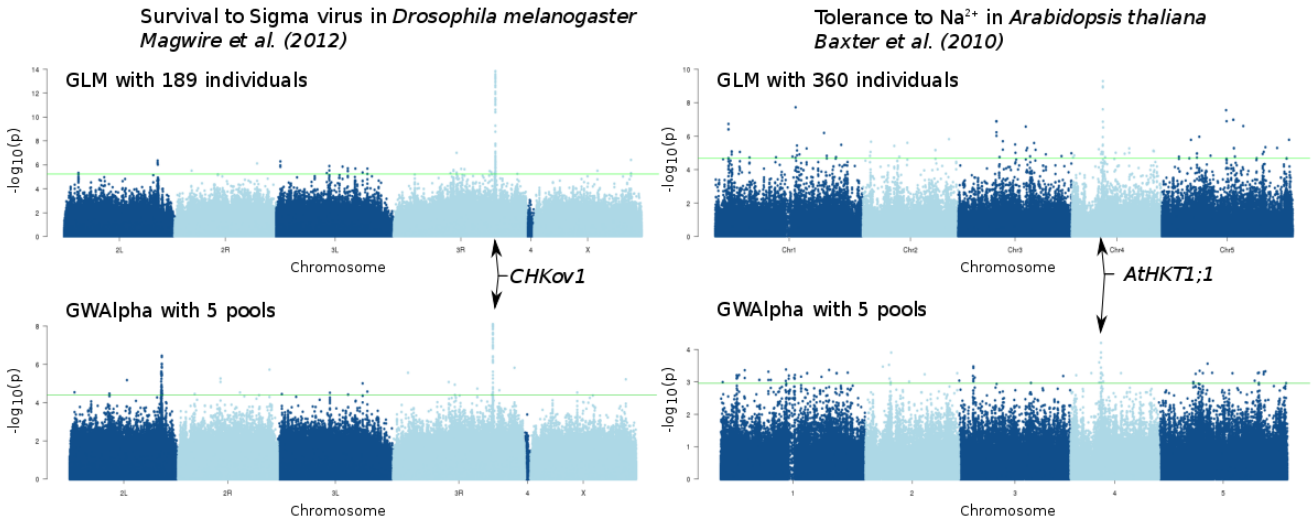
Suppl. Table 3: Proportion of causative SNPs found in the top 100 associations using GWAlpha or alternative methods for different range of frequencies.

Supplementary Data 5: Repeatability of GWAS applying GWAlpha to empirical data

In each of the four GWAS reanalysed, each individual i was assigned to a pool k bound by $[y'_{q_{k-1}}; y'_{q_k}]$ based on its trait quantile position y'_i if $y'_{q_{k-1}} < y'_i \leqslant y'_{q_k}$. Since the boundaries of the pools are defined using the inverse-quantile function for the trait distribution, GWAlpha is less sensitive to the initial trait distribution. Each individual is assigned to a single pool, and conversely all individuals with identical trait value are assigned in the same pool. As a consequence, if numerous individuals have the same phenotype (as in Battlay et al. (2016) with all 42 individuals in pool 1 showing no survival), the number of individuals may be different across pools (Suppl. Figure 5).



Suppl Figure 5: Trait variation and design of the pools used in the data analysis. All datasets were analysed using five pools with the exception of Magwire et al. (2012) where association could be repeated using four pools only.

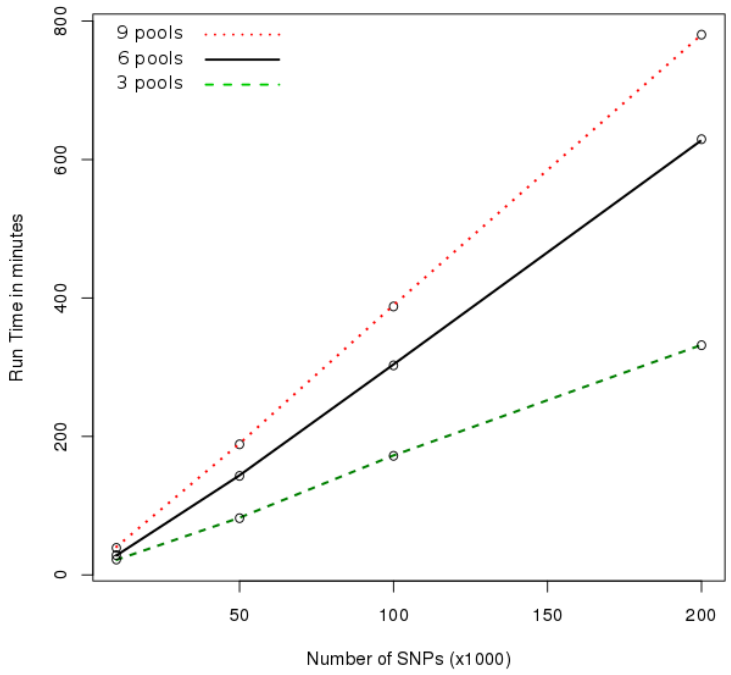


Suppl Figure 6: Convergence of the association signal between GWAlpha using 5 or 4 pools and individual-based GLM was tested in different organisms using two published data sets (Baxter et al., 2010; Magwire et al., 2012). The p-value threshold correspond to the empirical thousand-ile for each dataset.

To further validate the coherence between GWAlpha and GWAS performed using GLM, correlations between alphas and betas (the linear predictors of genetic effect from GWAlpha and GLM, respectively) were computed. All GWAlpha/GLM correlations were positive and significant using Pearson's rho. The correlation coefficients were the highest for the *D. melanogaster* datasets: 0.62 for the Magwire *et al.* (2012) dataset and 0.57 for the Battlay *et al.* (2016) dataset; and slightly lower for the *A. thaliana* datasets: 0.38 for the Brachi *et al.* (2015) dataset and 0.45 for the Baxter *et al.* (2010).

Supplementary Data 6: Memory usage and Performance

The run time and performance of GWAlpha was tested on a Linux architecture operating with Ubuntu 15.04. 12 GWAlpha simulations were run for each SNP and pools number combination, setting a memory usage limit of 1.25GB for each simulation and the average run time is presented in Suppl. Figure 7. Furthermore, GWAlpha was able to run 5 million SNPs with 9 pools under this memory usage setting without exceeding the memory limit. However, since running 5 million SNP on 9 pools is expected to take ~130 hours to complete using 1.25GB of memory, we suggest computing the model in parallel as implemented in the GWAlpha.sh script.



Suppl Figure 7: Run time in minute to complete a GWAlpha test for different number of SNPs and pools.

As an indication for the examples reported in Data Analysis section, using 12 parallel processes on a 12 core processor of 4GB each, the A. thaliana dataset (214,566 SNPs) was analysed in less than 45 minutes and the D. melanogaster dataset (4,438,427 SNPs), in less than 15 hours.