Archipelago method for variant set association test statistics.

Dylan Lawless*[1], Ali Saadat[2], Mariam Ait Oumelloul[2], and Jacques Fellay[2]

[1]Department of Intensive Care and Neonatology, University Children's Hospital Zürich, University of Zürich, Switzerland.
[2]Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Switzerland.

March 17, 2025

Abstract

Variant set association tests (VSAT), especially those incorporating rare variants via variant collapse, are invaluable in genetic studies. However, unlike Manhattan plots for single-variant tests, VSAT statistics lack intrinsic genomic coordinates, hindering visual interpretation. To overcome this, we developed the Archipelago method, which assigns a meaningful genomic coordinate to VSAT P values so that both set-level and individual variant associations can be visualised together. This results in an intuitive and information rich illustration akin to an archipelago of clustered islands, enhancing the understanding of both collective and individual impacts of variants. In a validation study using 504 East Asian subjects from the 1000 Genomes Project phase 3 (250 cases, 254 controls), we first conducted a single-variant GWAS using logistic regression with Firth correction. We then performed an equivalent case/control analysis by collapsing variants into protein pathways (defined via STRINGdb and PreoMCLustR) and testing associations with SKAT-O. The Archipelago plot is applicable in any genetic association study that uses variant collapse to evaluate both individual variants and variant sets, and its customizability facilitates clear communication of complex genetic data. By integrating at least two dimensions of genetic data into a single visualization, VSAT results can be easily read and aid in identification of potential causal variants in variant sets such as protein pathways.

GitHub repository: https://github.com/DylanLawless/archipelago.
CRAN submission: archipelago Version 0.0.1.9000.

*Addresses for correspondence: Dylan.Lawless@uzh.ch

1 Introduction

Variant set association tests (VSATs) are methods in which groups of variants are collapsed and analysed jointly to enhance an association signal. VSAT, particularly those incorporating rare variants (e.g. minor allele frequencies below 1%), have become indispensable in genetic association studies (1–5). Next-generation sequencing has enabled the comprehensive detection of rare variants, thereby complementing traditional single-variant tests (6). While Manhattan plots effectively visualise single-variant associations (7), a comparable graphical representation for VSAT results has been lacking.

To name a small few, VSAT methods include burden tests (e.g. CAST (8), CMC (9), weighted-sum (10)), adaptive burden tests (e.g. alpha-sum (11), VT (12), KBAC (13)) and variance component tests. Variance component approaches, such as the C(α) test (14; 15) and kernel machine regression methods (3), offer flexible strategies for assessing the combined effect of rare variants. In particular, the sequence kernel association test (SKAT) (16) and its extensions including SKAT-O (17), MSKAT (18), famSKAT, and RC-SKAT (19–22) have become widely adopted.

More recently, the variant-set test for association using annotation information (STAAR) framework (5; 23) has been introduced, utilising the aggregated Cauchy association test (ACAT) (24) to combine multiple annotation-based P values for each variant. Despite these methodological advances, the lack of a natural genomic coordinate for VSAT P values hinders their visual interpretation.

To address this gap, we introduce the Archipelago plot. By assigning a genomic coordinate to the VSAT P value based on the average of its constituent variants-this method facilitates the simultaneous visualisation of aggregated and individual variant effects. Notably, multiple types of VSAT, including SKAT-O and ACAT, have been highly successful and scalable to national biobank cohorts. For example, meta-analysis of TOPMed whole-genome sequencing data and UK Biobank whole-exome sequencing data, encompassing approximately 200,000 samples, has demonstrated the utility of these methods (25). Similarly, rare non-coding variation in complex human phenotypes has been investigated in 333,100 individuals from three cohorts: UK Biobank (N = 200,003), TOPMed (N = 87,652) and All of Us (N = 45,445) (26). Given their proven success, VSAT methods are likely to remain essential in future studies, and our Archipelago plot is designed to be applicable to such large-scale investigations.

2 Implementation

The Archipelago plot is a novel visualization technique developed to facilitate the interpretation of both (1) variant-set association testing (VSAT) P values (2) single variant P values in the context of the variant set. Its design is based on the commonly used Manhattan plot but accounts for the unique properties of VSAT statistics.

Figures 1 uses synthetic data to illustrate output that might be expected in rare variant association testing (RVAT) for case-control analysis. Briefly, 5,000 independent variants were simulated for GWAS results and 250 VSAT tests (20 variants per set). VSAT P values were drawn from a log-normal distribution using a minimal threshold (0.05/10), while GWAS P values were uniformly sampled between 0.05/100 and 1, with base pair positions and chromosome numbers assigned at random. This represented 250 VSAT P values and 5000 independent variants. As shown in **figure 1**, blue points indicate the joint VSAT P value as produced by SKAT-O or other statistical methods. Alternating yellow and orange points indicate the individual P values from each variant as produced by a regression or some other statistical method (e.g. R package SKAT: `$param$p.val.each`, P value for each single variant in a set-based test, or as calculated in single variant analysis). Alternating colors clarify the chromosome position. Lines connect the VSAT P value (blue) to its constituent individual variants to indicate variant set grouping. These edges are only shown for variants which are also in the significantly enrich VSAT P value group.

The precise protocol summary, definition, and algorithm can be found in **sections 7.6-7.8**. User setting of the R package can be read in **section 7.5**. The R code used in this manuscript can be found in the GitHub package repository. The main steps of the process are:

1. Within each variant set, get the average genome-wide genomic coordinate and assign it to the VSAT group P value.
2. Optionally, normalise the VSAT x-axis distribution for dispersion to prevent center clustering in dense datasets. Ranking of VSAT P values also provides priority in the event of overlaps.
3. Optionally, map the VSAT position to each individual variant from the set.

Since variant sets (such as protein pathways) have no specific single genomic coordinate, it is otherwise difficult to assign a logical x-axis position. Figure 2 illustrates the default format of raw VSAT P values, which will either be ranked by their P value strength or some other rank based on the variant set construction which results in an arbitrary x-axis distribution. The Archipelago plot assigns these same VSAT results into their x-axis location as the ranked average of the individual variant set genomic coordinate, thus representing an information rich illustration. For example, a variant set that is made from mostly chromosome 1 variants, will appear on the left, near its constituent variant P value points from chromosome 1. Genome-wide variant sets will appear near the center, however to prevent strong clustering in very dense datasets we also provide the method for ranked dispersion based on genomic coordinates for clearer illustration of the VSAT P value.

To ensure that VSAT points do not cluster and heavily overlap, ranking is applied before finding the variant set average based on the VSAT P value. That is, the x-axis position of the strongest association will have priority. Although overlap is unlikely in genome-wide testing, it may occur where variant sets are constructed from a narrow

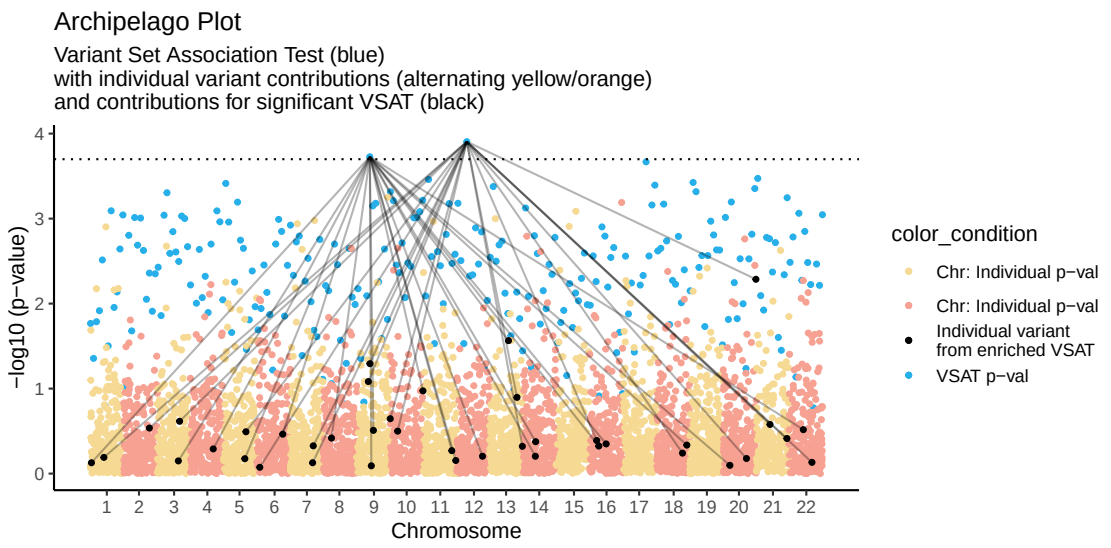


Figure 1: Synthetic data for human autosomal chromosomes 1-22. Contains 250 variant sets with 5000 individual variants. Each significantly enriched VSAT P value (blue) is mapped to its constituent individual P values (yellow and orange points). Significant threshold based on the number of VSAT.

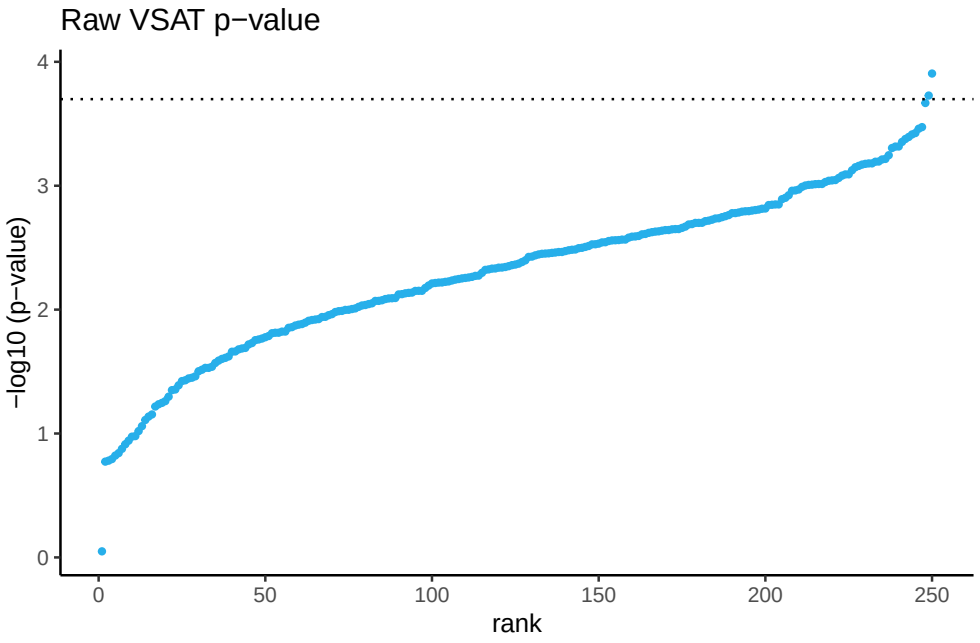


Figure 2: Raw VSAT P values have no natural x-axis position and must be ranked on association strength or some arbitrary ranking. The same VSAT result shown in **figure 1** (blue) are shown here without the use of the Archipelago method

.

set of genomic locations. In all figures a significant P value threshold was used which would typically be derived based on the number of independent tests, such as the number of variant sets. In these examples we used the arbitrarily chosen P value $= \frac{.05}{250} = .0002$ indicating 250 variant set tests.

We include a set of supplemental **figures S1a**, **S1b**, and **S1c** which contain the the same dataset as **figure 1**. However, each version illustrates the decreasing levels of clarification information to demonstrate the layers of annotation. **Figure S1a** shows the original plot without the figure legend. **Figure S1b** drops the edge highlights for the significantly enriched VSAT to reveal all connections, which can be difficult to read in high density plots. **Figure S1c** next drops the individual variant P value highlighting. Conversely, **figure S2** adds an additional layer of information by adding two colours for the significantly enriched variant sets which is useful for when there are multiple enriched variant sets but otherwise may be distracting. We demonstrate a set of sparse plots in **figures S3a**, **S3b**, and **S3c**. These smaller dataset examples use synthetic data to represent 500 qualifying variants, and VSAT of 20 genes/variants per variant set (25 VSAT P values).

Customisation settings are described in section 7.9. **Figure S4** shows an example of customised colors. **Figure S5** shows all customisable elements (title, subtitle, colors, color labels, critical threshold line, genomic coordinate, show title and subtitle, and show legend). **Figures S6 - S7** shows all 16 color themes.

3 Expansion with additional information layers

The variant-set test for association using annotation information (STAAR) method has special features since each individual variant P value (pSTAAR) is itself made from an aggregated P value derived from the set of annotation weighted P value using the ACAT method (5; 23; 24). Therefore, a third level of information hierarchy is present. In the basic Archipelago plot we have two layers: (1) VSAT P value and (2) individual P value. The third pSTAAR level might be illustrated by in a sub-plot which focuses on the significantly enriched VSAT group. This sub-plot is then free to show x-axis genomic coordinates, and y-axis P value which is used for the Cauchy distribution aggregation of $f(p_n)$.

To demonstrate the source of annotation based ACAT P values, **figure 3** panel [1] shows the Archipelago plot with two significantly enriched variant sets (blue), 40 contributing individual variants (black). Panel [2] then shows the ACAT P values for 4 annotation layers for each one of the individual (black) variants. To clarify that each single variant position will have four annotation P values, panel [3] overlays line/group box to show the spread of ACAT P values for each single variant. We note that highly correlated annotation layers are likely since many prediction tools relay on population genetics frequency as an indicator of variant effect quantification. Therefore, it would be reasonable to also show a further plot of the correlation between the annotation layers. While this is a simplified and specific example, we image that other similar

applications will exist for VSAT with third layers of information that consist of x-axis genomic coordinate and y-axis test statistic.

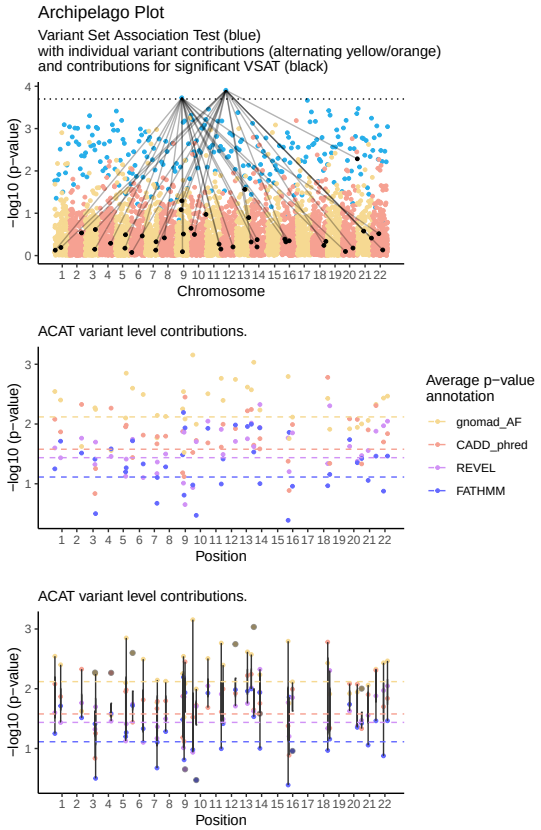


Figure 3: Panel [1] Archipelago plot as per figure 1 with two significantly enriched variant sets (blue), 40 contributing individual variants (black). Panel [2] ACAT P values based on 4 annotation layers for each one of the individual (black) variants. Horizontal lines show the average annotation P value as an estimate of the hierarchy of contribution. Panel [3] overlays line/group box to show the spread of ACAT P values for each single variant.

4 Validation

We validated the archipelago method reproducibly (**supplemental, section 7**) using the public dataset of 504 East Asian individuals from the 1000 Genomes Project phase 3 (version 5, hg19)http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. Since 1KG data lack a common disease phenotype, we simulated a binary trait (250 cases, 254 controls) under a heritability of 0.8, then applied standard QC steps (e.g. splitting multi-allelic variants, normalising genotypes, LD pruning, KING-based relatedness checks) to retain 500 samples and 1,224,104 SNPs. A single-variant

GWAS was run via logistic regression with Firth correction, and a variant-set association test (VSAT) was conducted using protein pathways from ProteoMCLustR (941 pathways) and SKAT-O. **Tables 1 - 2** illustrate the resulting pathway- and SNP-level outputs; the archipelago plot (**figure 4**) integrates both in a single genomic view.

The enriched variant signals displayed on the bottom of **figure 4** were assessed based on their allele frequency. In this dataset, top GWAS signals involved highly common variants (n = 67, mean allele frequency 0.333, min = 0.13 , max = 0.49), whereas the top VSAT signal (set_ID 532) included 261 variants with a mean frequency of 0.151 (min = 0.01 , max = 0.48), reflecting a different genetic architecture captured by pathway-based collapse.

set_ID	P
1	0.755726849903356
2	0.288273411609041
3	0.540978825441736
4	0.785789910303476

Table 1: The output of SKAT-O with pathway IDs (set_ID) and P value (df_pathway_sim.csv).

set_ID	CHR	BP	P
1	1	160289049	0.25598
1	1	160295829	0.559735
1	1	160300635	0.920975

Table 2: Output of the GWAS data using Plink after mapping the pathway IDs (set_ID) to SNPs (df_snp_id.csv). CHR, chromosome; BP, position, P, P value.

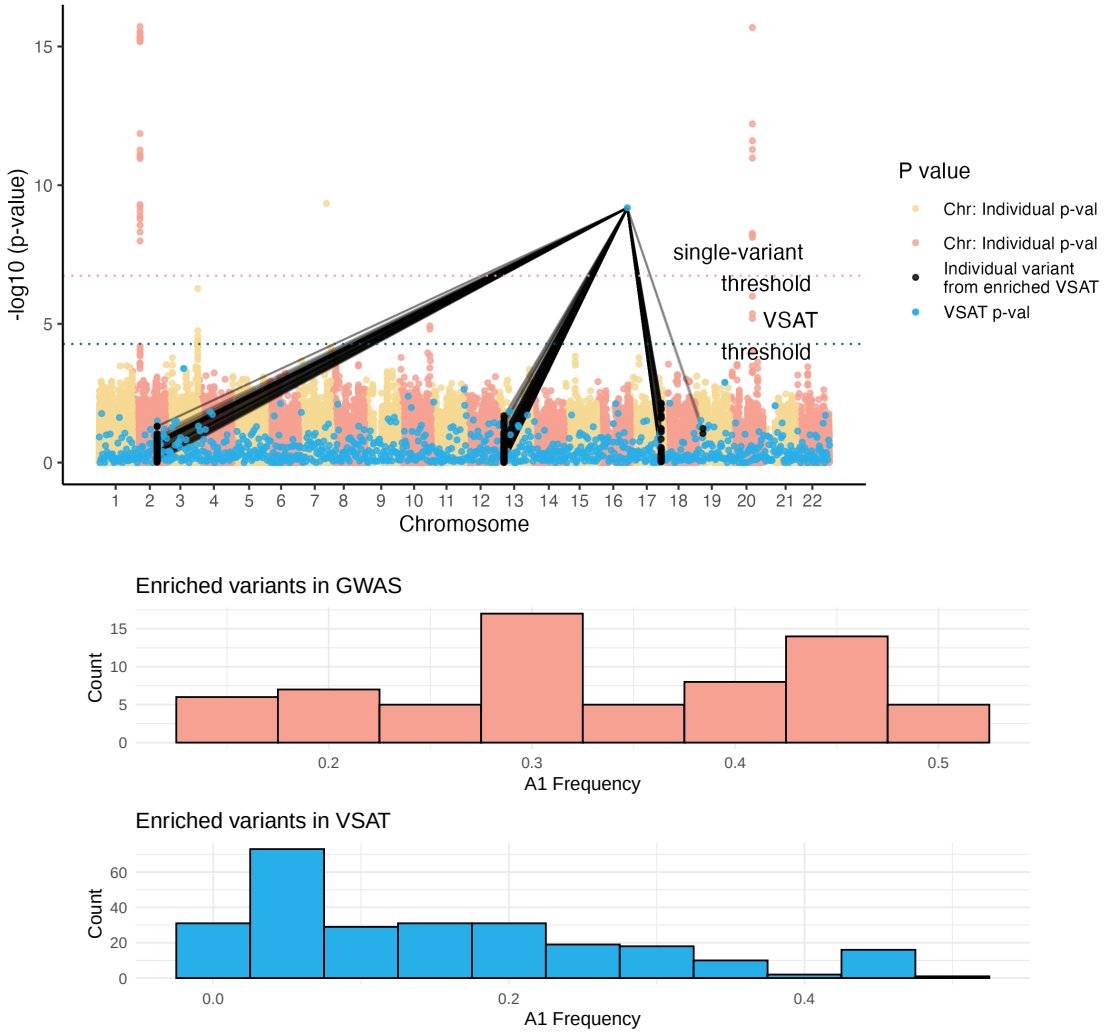


Figure 4: Archipelago plot of validation study results. Data from 504 East Asian individuals (1000 Genomes Project phase 3, hg19) were processed following standard QC, LD pruning and variant normalisation. A simulated binary phenotype (250 cases, 254 controls) was analysed via single-variant GWAS using Firth-corrected logistic regression, and by collapsing variants into protein pathways (defined by ProteoM-CLustR and mapped with Ensembl Biomart) for pathway-level association testing with SKAT-O. The plot integrates pathway-level VSAT P values with corresponding SNP-level associations, providing a unified genomic visualisation of both collective and individual variant effects. On the bottom, variant allele frequencies are shown for the enriched variants found by GWAS and by VSAT (collapse within protein pathways).

5 Discussion

We have introduced the Archipelago plot as a novel method for interpreting VSAT results alongside individual variant P values, as produced from SKAT-O or some other statistical method after variant collapse (1; 2; 4). In contrast to traditional Manhattan plots that are restricted to single-variant associations with clear genomic coordinates, the Archipelago plot assigns a representative genomic location to the VSAT P value based on the average of its constituent variants (7). This enables a direct comparison of the aggregated signal with individual variant effects within a familiar framework.

The absence of a natural genomic coordinate for VSAT P values has previously hindered the integrated visual interpretation of variant set and single-variant results. By linking the VSAT P value to the genomic positions of the individual variants, the Archipelago plot offers an intuitive depiction of the relationship between variant set effects and the underlying individual signals. This approach may assist in identifying potential causal variants or pathways and can be readily extended to accommodate additional layers of information, such as annotation-based P values derived using methods like ACAT. We have seen that this method could be applied to national biobank scale genomic studies with relative ease (25; 26).

While the method was developed with rare-variant testing in mind, its customisable design permits application to any genetic association study that utilises variant collapse. Overall, the Archipelago plot provides a modest yet useful tool for the clear communication of complex genetic data, contributing to improved interpretability in large-scale VSAT analyses.

6 Author contributions

Dylan Lawless designed and wrote the work and performed analysis. Ali Saadat and Mariam Ait Oumelloul wrote the work. Jacques Fellay lead project management.

7 Author declarations

The authors have declared no competing interest.

References

[1] Andriy Derkach, Jerry F. Lawless, and Lei Sun. Pooled Association Tests for Rare Genetic Variants: A Review and Some New Results. *Statistical Science*, 29 (2), May 2014. ISSN 0883-4237. doi: 10.1214/13-STS456. URL http://arxiv.org/abs/1205.4079. arXiv:1205.4079 [stat].

[2] Seunggeung Lee, Gonçalo R. Abecasis, Michael Boehnke, and Xihong Lin. Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal of Human Genetics*, 95(1):5–23, July 2014. ISSN 00029297. doi: 10.1016/j.ajhg.2014.06.009. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929714002717.

[3] Nicholas B. Larson, Jun Chen, and Daniel J. Schaid. A review of kernel methods for genetic association studies: LARSON et al.. *Genetic Epidemiology*, 43(2):122–136, March 2019. ISSN 07410395. doi: 10.1002/gepi.22180. URL https://onlinelibrary.wiley.com/doi/10.1002/gepi.22180.

[4] Gundula Povysil, Slavé Petrovski, Joseph Hostyk, Vimla Aggarwal, Andrew S. Allen, and David B. Goldstein. Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nature Reviews Genetics*, 20(12):747–759, December 2019. ISSN 1471-0056, 1471-0064. doi: 10.1038/s41576-019-0177-4. URL https://www.nature.com/articles/s41576-019-0177-4.

[5] Zilin Li, Xihao Li, Hufeng Zhou, Sheila M Gaynor, Margaret Sunitha Selvaraj, Theodore Arapoglou, Corbin Quick, Yaowu Liu, Han Chen, Ryan Sun, et al. A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nature Methods*, 19(12):1599–1611, December 2022. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-022-01640-x. URL https://www.nature.com/articles/s41592-022-01640-x.

[6] Dan L. Nicolae. Association Tests for Rare Variants. *Annual Review of Genomics and Human Genetics*, 17(1):117–130, August 2016. ISSN 1527-8204, 1545-293X. doi: 10.1146/annurev-genom-083115-022609. URL https://www.annualreviews.org/doi/10.1146/annurev-genom-083115-022609.

[7] Stephen D Turner. qqman: an r package for visualizing gwas results using qq and manhattan plots. *Biorxiv*, page 005165, 2014.

[8] Stephan Morgenthaler and William G. Thilly. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1-2):28–56, February 2007. ISSN 00275107. doi: 10.1016/j.mrfmmm.2006.09.003. URL https://linkinghub.elsevier.com/retrieve/pii/S0027510706002740.

[9] Bingshan Li and Suzanne M. Leal. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *The American Journal of Human Genetics*, 83(3):311–321, September 2008. ISSN 00029297. doi: 10.1016/j.ajhg.2008.06.024. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929708004084.

[10] Bo Eskerod Madsen and Sharon R. Browning. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genetics*, 5(2):e1000384,

February 2009. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000384. URL https://dx.plos.org/10.1371/journal.pgen.1000384.

[11] Fang Han and Wei Pan. A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants. *Human Heredity*, 70(1):42–54, 2010. ISSN 0001-5652, 1423-0062. doi: 10.1159/000288704. URL https://www.karger.com/Article/FullText/288704.

[12] Alkes L. Price, Gregory V. Kryukov, Paul I.W. de Bakker, Shaun M. Purcell, Jeff Staples, Lee-Jen Wei, and Shamil R. Sunyaev. Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *The American Journal of Human Genetics*, 86(6):832–838, June 2010. ISSN 00029297. doi: 10.1016/j.ajhg.2010.04.005. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929710002077.

[13] Dajiang J. Liu and Suzanne M. Leal. A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. *PLoS Genetics*, 6(10): e1001156, October 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1001156. URL https://dx.plos.org/10.1371/journal.pgen.1001156.

[14] Benjamin M. Neale, Manuel A. Rivas, Benjamin F. Voight, David Altshuler, Bernie Devlin, Marju Orho-Melander, Sekar Kathiresan, Shaun M. Purcell, Kathryn Roeder, and Mark J. Daly. Testing for an Unusual Distribution of Rare Variants. *PLoS Genetics*, 7(3):e1001322, March 2011. ISSN 1553-7404. doi: 10.1371/journal.pgen.1001322. URL https://dx.plos.org/10.1371/journal.pgen.1001322.

[15] Jerzy Neyman. C (α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 1–21, 1979.

[16] Michael C. Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. *The American Journal of Human Genetics*, 89(1):82–93, July 2011. ISSN 00029297. doi: 10.1016/j.ajhg.2011.05.029. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929711002229.

[17] S. Lee, M. C. Wu, and X. Lin. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775, September 2012. ISSN 1465-4644, 1468-4357. doi: 10.1093/biostatistics/kxs014. URL https://academic.oup.com/biostatistics/article-lookup/doi/10.1093/biostatistics/kxs014.

[18] Seunggeun Lee, Mary J. Emond, Michael J. Bamshad, Kathleen C. Barnes, Mark J. Rieder, Deborah A. Nickerson, David C. Christiani, Mark M. Wurfel, and Xihong Lin. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies.

American journal of human genetics, 91(2):224–237, August 2012. ISSN 1537-6605 0002-9297. doi: 10.1016/j.ajhg.2012.06.007. Place: United States.

[19] Han Chen, James B. Meigs, and Josée Dupuis. Sequence Kernel Association Test for Quantitative Traits in Family Samples. *Genetic Epidemiology*, 37(2): 196–204, February 2013. ISSN 07410395. doi: 10.1002/gepi.21703. URL https://onlinelibrary.wiley.com/doi/10.1002/gepi.21703.

[20] Iuliana Ionita-Laza, Seunggeun Lee, Vladimir Makarov, Joseph D Buxbaum, and Xihong Lin. Family-based association tests for sequence data, and comparisons with population-based association tests. *European Journal of Human Genetics*, 21(10):1158–1162, October 2013. ISSN 1018-4813, 1476-5438. doi: 10.1038/ejhg.2012.308. URL http://www.nature.com/articles/ejhg2012308.

[21] Iuliana Ionita-Laza, Seunggeun Lee, Vlad Makarov, Joseph D. Buxbaum, and Xihong Lin. Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants. *The American Journal of Human Genetics*, 92(6):841–853, June 2013. ISSN 00029297. doi: 10.1016/j.ajhg.2013.04.015. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929713001766.

[22] Diptavo Dutta, Laura Scott, Michael Boehnke, and Seunggeun Lee. Multi-SKAT: General framework to test for rare-variant association with multiple phenotypes: DUTTA et al.. *Genetic Epidemiology*, 43(1):4–23, February 2019. ISSN 07410395. doi: 10.1002/gepi.22156. URL https://onlinelibrary.wiley.com/doi/10.1002/gepi.22156.

[23] Xihao Li, Zilin Li, Hufeng Zhou, Sheila M Gaynor, Yaowu Liu, Han Chen, Ryan Sun, Rounak Dey, Donna K Arnett, Stella Aslibekyan, et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature Genetics*, 52(9):969–983, September 2020. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-020-0676-4. URL https://www.nature.com/articles/s41588-020-0676-4.

[24] Yaowu Liu, Sixing Chen, Zilin Li, Alanna C. Morrison, Eric Boerwinkle, and Xihong Lin. Acat: A fast and powerful p value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*, 104 (3):410–421, March 2019. ISSN 00029297. doi: 10.1016/j.ajhg.2019.01.002. URL https://linkinghub.elsevier.com/retrieve/pii/S0002929719300023.

[25] Xihao Li, Corbin Quick, Hufeng Zhou, Sheila M Gaynor, Yaowu Liu, Han Chen, Margaret Sunitha Selvaraj, Ryan Sun, Rounak Dey, Donna K Arnett, et al. Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nature genetics*, 55(1):154–164, 2023.

[26] Gareth Hawkes, Robin N Beaumont, Zilin Li, Ravi Mandla, Xihao Li, Christine M Albert, Donna K Arnett, Allison E Ashley-Koch, Aneel A Ashrani, Kathleen C Barnes, et al. Whole-genome sequencing in 333,100 individuals reveals

rare non-coding single variant and aggregate associations with height. *Nature Communications*, 15(1):8549, 2024.

8 Supplemental

8.1 Validation methods

We performed a validation test using a popular reference dataset consisting of 504 EAS individuals from 1000 Genomes Project phase 3 version 5 (1KG 3v5), genome build human_g1k_v37.fasta (hg19) (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/). This dataset contained around 1 million variants from the ancestral backgrounds: CHB: Han Chinese in Beijing, China; JPT: Japanese in Tokyo, Japan; CHS: Southern Han Chinese; CDX: Chinese Dai in Xishuanagbanna, China; and KHV: Kinh in Ho Chi Minh City, Vietna. The result of this study is shown in **figure 4**.

To streamline reproducibility we processed this dataset using the popular tutorial from the Laboratory of Complex Trait Genomics (Kamatani Lab) in the Deparment of Computational Biology and Medical Sciences at the Univerty of Tokyo (https://github.com/Cloufield/GWASTutorial). The data was processed in preparation for a typical GWAS study: we selected only autosomal variants, split multi-allelic variants, variants were normalized, remove duplicated variants, selected only SNP (ATCG), selected 2% rare SNPs (`plink --mac 2 --max--maf 0.01 --thin 0.02`, selected 15% common SNPs (`plink --maf 0.01 --thin 0.15`), converted to plink bed format and merged to a single file, and randomly added some missing data points. A prepared version of this processed data is also available from the authors repository in zipped Plink format. Since this public data set (1KG 3v5) as no shared disease phenotype, we followed the same resource to simulate the phenotype file. The simulation parameters were: number of simulation replicate(s) = 1 (Default = 1), heritability of liability = 0.8 (Default = 0.1), disease prevalence = 0.5 (Default = 0.1), number of cases = 250, and number of controls = 254. The QC steps included running `--r2` with the following filters: `--ld-window: 10`, `--ld-window-kb: 1000`, `--ld-window-r2: 0.2`, and `--r2 to plink_results.ld`. We then included `--geno 0.02`, `--hwe 1e-6`, `--mind 0.02`. High-LD regions were flagged (`high-ld-hg19.txt`) and pruning was done with `--indep-pairwise 500 50 0.2`. For subject kinship, KING-cutoff 0.0884 was followed by projection (related and unrelated samples).

We performed the single-variant GWAS as a logistic regression with firth correction for a simulated binary trait under the additive mode, with `--maf 0.01`, five PCs, and the simulated phenotypes. Next, we used the default protein pathway gene sets from ProteoMCLustR (https://github.com/DylanLawless/ProteoMCLustR). This consists of a pre-computer set of pathways after cluster the full human genome set of STRINGdb with PPI evidence confidence score 0.7 and contains 941 distinct pathways. We mapped these pathways ID with the variant coordiates (bim file) using the gene coordiates available for this genome build from ensembl biomart `mart_export_GRCh37p13.txt` (https://grch37.ensembl.org/biomart/martview/) for the dataset: human genes (GRCh37.p13), with attributes: Gene stable ID, Gene start (bp), Gene end (bp), Chromosome/scaffold name, Gene name. Our repository also includes `mart_export_GRCh38p14.txt` from http://mart.ensembl.org/

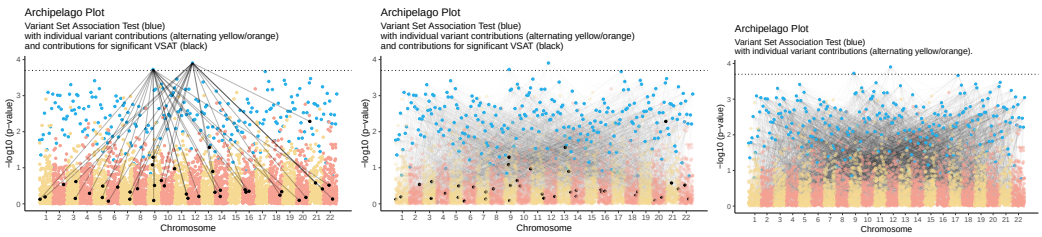
biomart/martview/.

We then performed the VSAT with pathway-level variant collapse by running SKAT-O with the `set_ID` assigned to these ProteoMCLustR pathways (e.g. pathway 1,...,n). After QC in GWAS this cohort contained 500 samples and in VSAT we had 945 sets with 1224104 total SNPs. We saved the results from both the Plink single-variant GWAS and the pathway-level SKAT-O. Since this is not a disease cohort and the phenotype is simulated, we automatically increased the association strength of the top VSAT set to provide an enrichment signal. This result is similar to what we have observed in other studies of this size but which do not have public genetic data. Running archipelago simply requires providing two such inputs (i.e. VSAT results, GWAS results) and any custom theme settings: `archipelago_plot(df_pathway_sim, df_snp_id)`.

The enriched variant signals displayed in **figure 4** were assessed based on their allele frequency. For the GWAS, the allele frequency of enriched variants (n = 67) was: mean = 0.333 , median = 0.315 , min = 0.133 , max = 0.49 which demonstrates that these individual SNP associations were due to highly common variants. For the top VSAT hit (set_ID 532), there were 261 variants. The allele frequency of variants was: mean = 0.151 , median = 0.121 , min = 0.011 , max = 0.478. This demonstrates the scenarios where rare variants within a collapsed set can have a different association than the GWAS region, although this is not always true.

8.2 Figure layers

Figures S1a, **S1b**, and **S1c** contain the the same dataset as **figure 1** however, each version illustrates the decreasing levels of clarification information to demonstrate the layers of annotation. **Figure S1a** shows the original plot without the figure legend. **Figure S1b** drops the edge highlights for the significantly enriched VSAT to reveal all connections, which can be difficult to read in high density plots. **Figure S1c** next drops the individual variant p-value highlighting.



(a) (b) (c)

Figure S1: Illustration of information layers used to reduce complexity.

8.3 Variant set edge color

Figure S2 adds an additional layer of information to that seen in **figure S1a** by adding two colours for the significantly enriched variant sets which is useful for when there are multiple enriched variant sets but otherwise may be distracting.

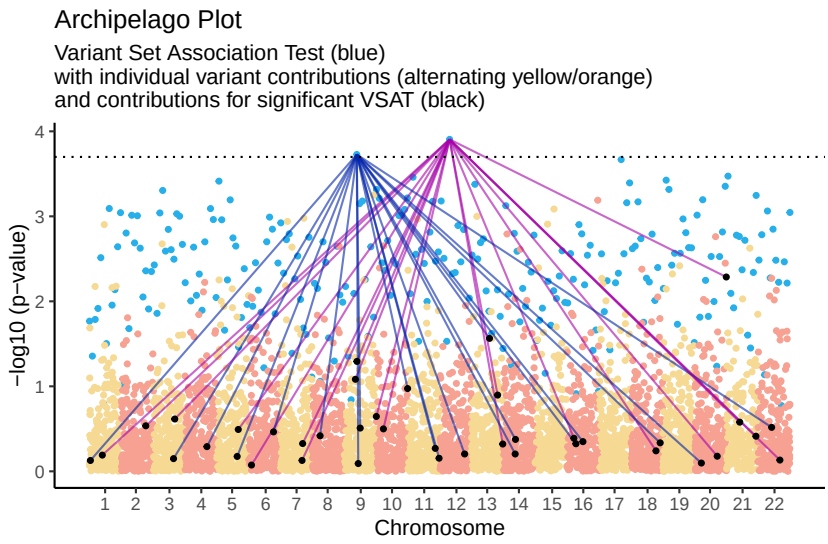


Figure S2: Colored VSAT edges to indicate separate variant sets from two significantly enriched VSAT p-values.

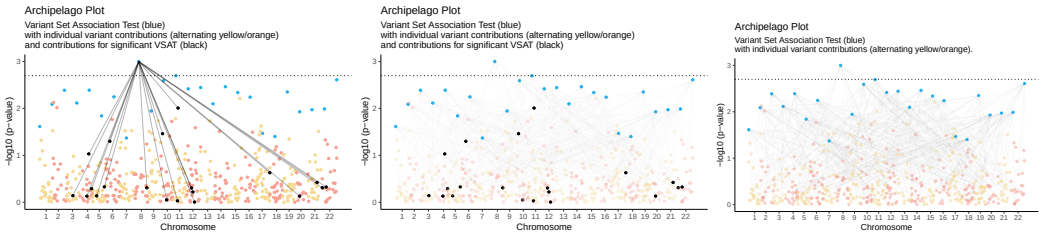
8.4 Sparse plots

We demonstrate a set of sparse plots in **figures S3a**, **S3b**, and **S3c**. These smaller dataset examples use synthetic data to represent 500 qualifying variants, and VSAT of 20 genes/variants per variant set (25 VSAT p-values). **Figure S3a** illustrates a small variant set of 500 individual variants in 25 variant sets. **Figure S3b** and **figure S3c** continue by dropping information layers sequentially as previously described in the dense plot examples.

8.5 R package user settings

```
# Install
install.packages("~/archipelago_0.0.0.9000.tar.gz", repos = NULL,
    ↪  type = "source")

# Load
library(archipelago)
```



(a) (b) (c)

Figure S3: Illustration of information layers used to reduce complexity in a sparse dataset.

```
# Load test data for df1 and df2
data("vsat_pval")
data("variant_pval")

# Import user data
# df1 <- read.csv(file="../data/vsat_pval.txt")
# df2 <- read.csv(file="../data/variant_pval.txt")

# Import user data
df1 <- read.csv(file="../data/vsat_pval.txt")
df2 <- read.csv(file="../data/variant_pval.txt")

# Use default settings
archipelago_plot(df1, df2)

# 8 color themes:
# 'retro', 'metro', 'summer', 'messenger', 'sunset', 'alice', '
   ↪ yawn', 'lawless'
output_path = "../output/archipelago_plot_custom_color.pdf"
archipelago_plot(df1,
                 df2,
                 color_theme = 'alice',
                 output_path = output_path)

# Custom everything
color_labels <- c("Label_1", "Label_2" ,"Label_3","Label_4")
custom_colors = c("#9abfd8", "#cac1f3", "#371c4b", "#2a5b7f") #
   ↪ alice theme colors
plot_title <- "Title"
plot_subtitle <- "Subtitle"
crit_val_VSAT = .05/300 # P-value threshold line
point_size = .5 # geom_point size
output_path = "../output/archipelago_plot_custom_everything.pdf"
better_space = "TRUE"
```

```
archipelago_plot <- function(df1, df2,
                             plot_title = "Archipelago␣Plot",
                             add_title = FALSE,
                             plot_subtitle = "Variant␣Set␣
                                ↪ Association␣Test\nwith␣
                                ↪ individual␣variant␣
                                ↪ contributions\nand␣
                                ↪ contributions␣for␣significant␣
                                ↪ VSAT",
                             add_subtitle = FALSE,
                             chr_ticks = TRUE,
                             show_legend = TRUE,
                             color_theme = NULL,
                             custom_colors = NULL,
                             color_labels = c("Chr:␣Individual␣p-
                                ↪ val", "Chr:␣Individual␣p-val" ,
                                ↪ "Individual␣variant\nfrom␣
                                ↪ enriched␣VSAT","VSAT␣p-val"),
                             crit_val_VSAT = NULL,
                             crit_val_single_variant = NULL,
                             point_size = 1,
                             point_size_large = 1,
                             fig_width = 8,
                             fig_height = 4,
                             raw_fig_width = 8,
                             raw_fig_height = 4,
                             output_path = "archipelago_plot",
                             output_raw = "archipelago_vsat_raw_
                                ↪ plot",
                             file_type = "png",
                             alpha_point = 1,
                             alpha_seg = 0.3,
                             better_space = FALSE,
                             legend_position = "right"
                             )


# Print every color theme plot for manual
# List of color themes
color_themes <- c('retro', 'metro', 'summer', 'messenger', '
   ↪ sunset', 'alice', 'yawn', 'lawless')

# Loop over each color theme
for (color_theme in color_themes) {
  # Define output path
  output_path <- paste0("../output/archipelago_plot_", color_
     ↪ theme, ".pdf")
```

```
  # Generate and save plot
  archipelago_plot(df1,
                   df2,
                   color_theme = color_theme,
                   output_path = output_path,
                   show_legend = FALSE)
}
```

8.6 Protocol Summary

The Archipelago Plot protocol takes as input two datasets containing variant-set association testing (VSAT) p-values and single variant p-values. It first orders the variant p-values by chromosome and genomic coordinate, then averages the coordinates within each variant set. The protocol subsequently normalizes the VSAT x-axis distribution (optional) to avoid center clustering in dense datasets, prioritizing VSAT p-values in case of overlaps. It can optionally map the VSAT position to each variant from the set. The protocol then creates a ggplot visualization, which includes an option to highlight individual variant contributions from significant VSATs.

8.7 Protocol Formal Definition

The formal definition of the Archipelago Plot Protocol comprises the following steps: **Data Preparation**: The data, including variant-set association testing (VSAT) p-values and single variant p-values, is sorted and averaged within each variant set. **Normalization**: Optionally, the VSAT x-axis distribution is normalized to avoid center clustering in dense datasets. Ranking is done to prioritize VSAT p-values in the event of overlaps. **Mapping**: Optionally, the position of each VSAT is mapped to each variant from the set. **Visualization**: A plot is created using ggplot2, which includes options to add a title, subtitle, chromosome ticks, a legend, and custom colors. **Highlighting**: The plot can be made to highlight individual variant contributions from significant VSATs. **Output**: The plot is saved to a specified path and returned by the function.

8.8 Protocol Algorithm

The core of the algorithm can be broken down into the following steps: **Pre-processing**: Load required libraries (ggplot2 and dplyr). Determine the critical value if it's not provided. Data ordering and coordinate calculation: Sort the single variant p-values by chromosome and genomic coordinate, and create an index to order the chromosomes. Calculate the position for each data point, essentially ordering the chromosomes end to end, and create mid-points for each chromosome for labeling. **Merging and preparation**: Merge the VSAT p-values with the single variant p-values, then calculate the sum and average position of each variant set group. Replace missing positions with the average position of their variant set. **Normalization**: If chosen,

normalize the VSAT x-axis distribution. This is done by sorting the variant set p-values by their position, then assigning evenly spaced numbers across the range of all position values. **Data coloring and splitting**: Assign colors to the different groups in the data and split the data into variant sets and individual variants. **Highlighting**: Define a condition to highlight individual variant contributions from significant VSATs and apply it to the data. **Plotting**: Create the plot with ggplot2, add color scaling, labels, theme, and optionally title, subtitle, and legend. **Finalization**: Save the plot to a specified output path and return the plot from the function.

The R code is relatively easy to modify or rewrite. Alternative versions could be written based on the following algorithm:

```
Input: D1 (VSAT p-values), D2 (single variant p-values)
Output: Archipelago Plot

Begin
  1. Load required libraries.
  2. Determine critical value if not provided.
  3. Sort D2 by chromosome and genomic coordinate, create index.
  4. Calculate position for each variant in D2, create chromosome
     ↪  mid-points.
  5. Merge D1 and D2, calculate sum and average position of each
     ↪ variant set.
  6. Replace NA positions with the average position of their
     ↪ variant set.
  7. Normalize x-axis distribution of D1 (optional).
  8. Assign colors to data groups, split data into variant sets
     ↪ and individual variants.
  9. Highlight individual variant contributions from significant
     ↪ VSATs.
  10. Create ggplot, add scaling, labels, theme.
  11. Optionally add title, subtitle, legend.
  12. Save and return plot.
End
```

8.9 Customisation

Figure S4 shows an example of customised colors. Figure S5 shows all customisable elements. Figure S6 - S7 shows all 16 color themes.

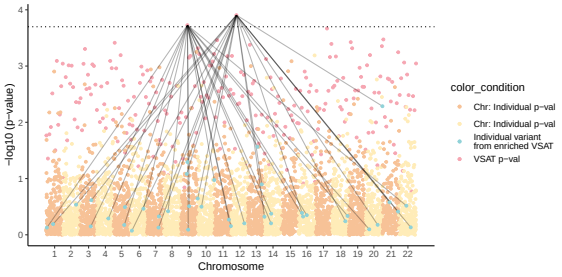


Figure S4: Custom colors specified by hex values, color names, etc.

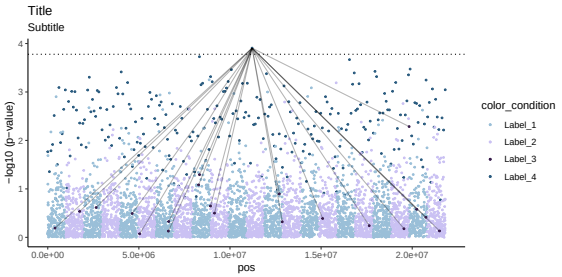


Figure S5: Custom title, subtitle, colors, color labels, critical threshold line, genomic coordinate, show title and subtitle, and show legend.

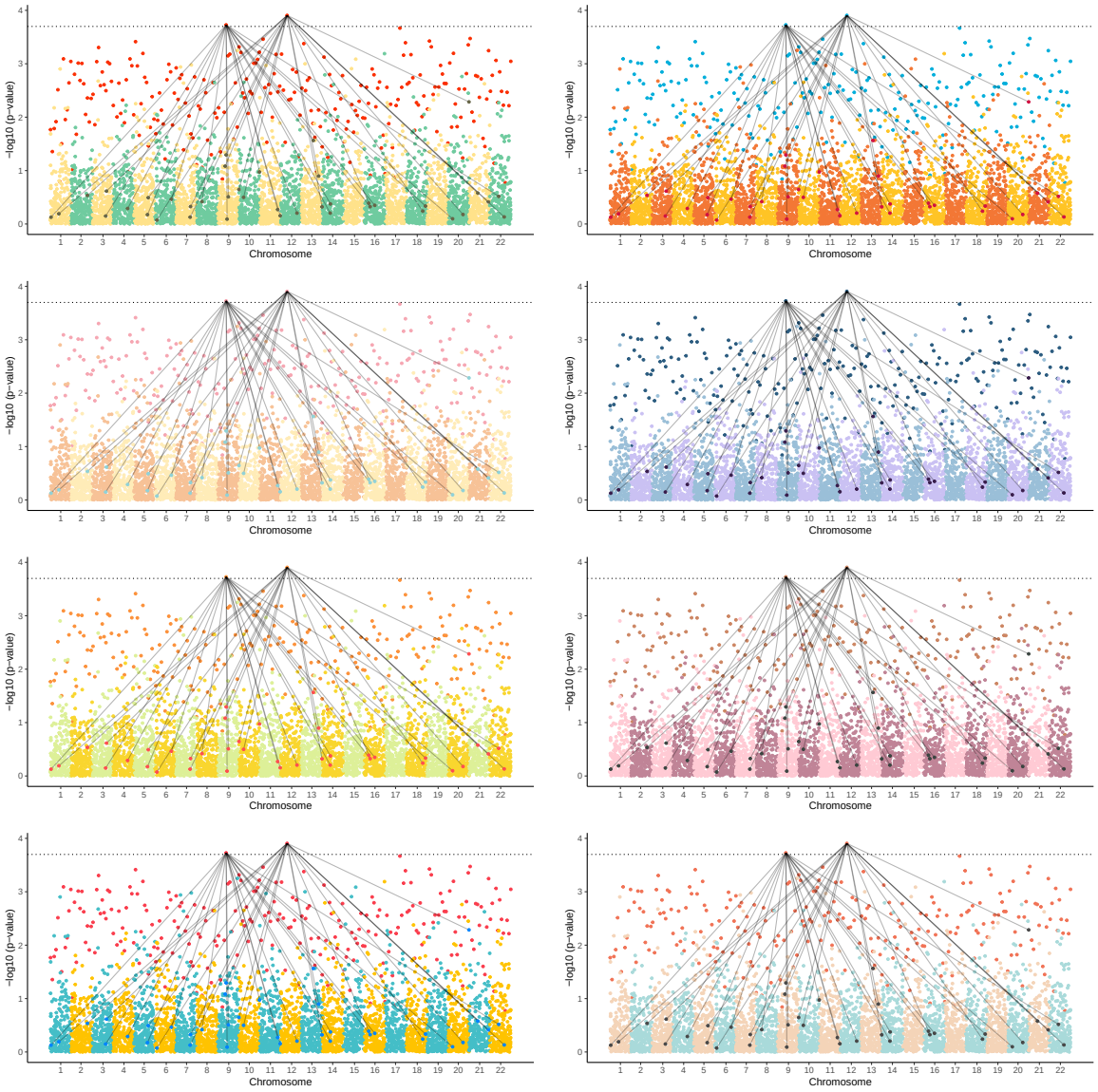


Figure S6: Color themes left to right: ‘retro’, ‘metro’, ‘alice’, ‘buckley’, ‘summer’, ‘romance’, ‘messenger’, ‘meme’.

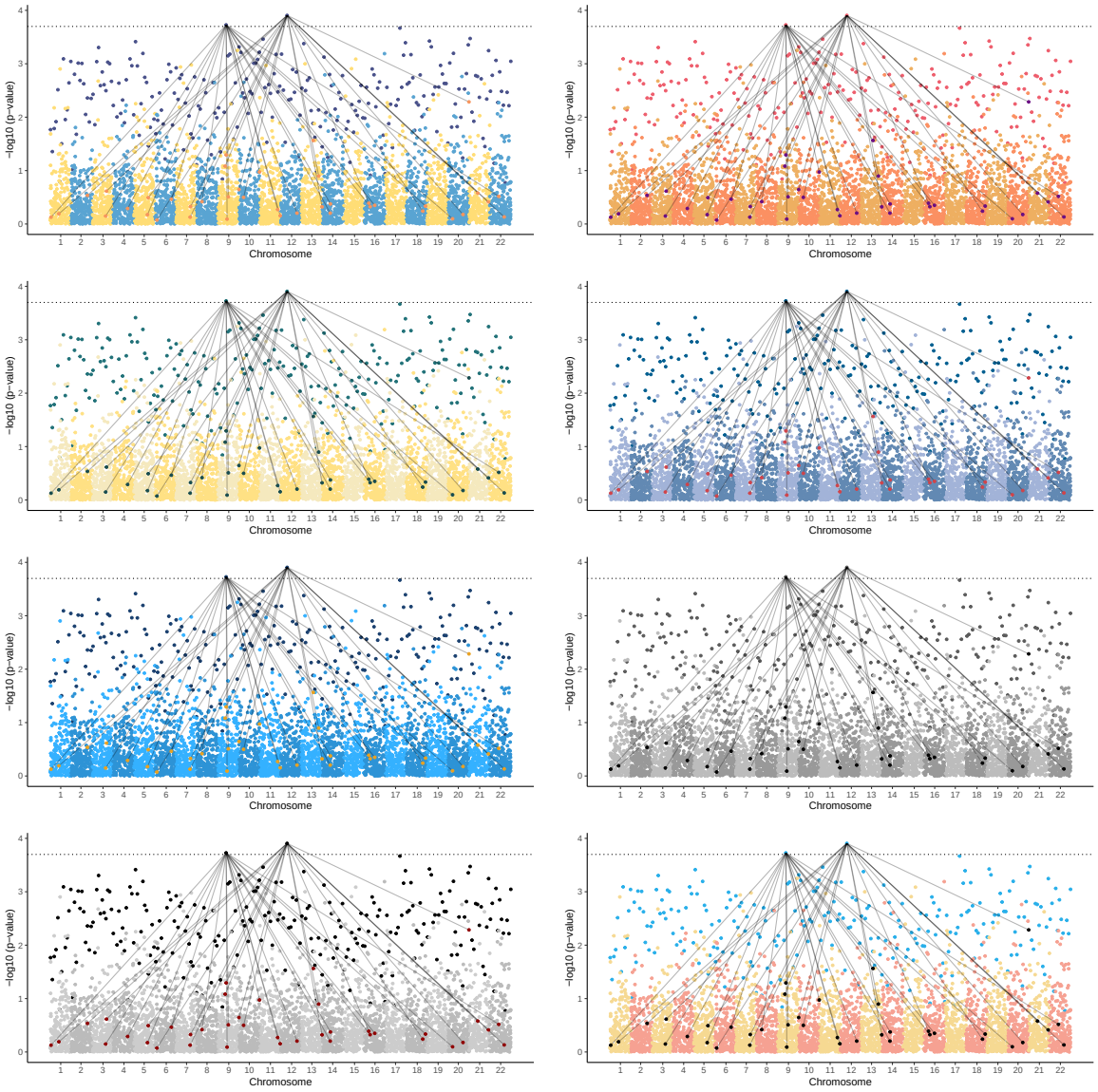


Figure S7: Color themes left to right: ‘pagliacci’, ‘sunset’, ‘ambush’, ‘saiko’, ‘sunra’, ‘yawn’, ‘caliber’, ‘lawless’.

Figure S8: Archipelago plot logo.