Trumpet plots: Visualizing The Relationship Between Allele Frequency And Effect Size In Genetic Association Studies

Lucia Corte¹,², Lathan Liou¹, Paul F. O'Reilly¹, Judit García-Gonzálezn

¹Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, New York City, NY 10029, USA
²Center for Excellence in Youth Education, Icahn School of Medicine, Mount Sinai, New York City, NY 10029, USA

Abstract

Recent advances in genome-wide association study (GWAS) and sequencing studies have shown that the genetic architecture of complex diseases and traits involves a combination of rare and common genetic variants, distributed throughout the genome. One way to better understand this architecture is to visualize genetic associations across a wide range of allele frequencies. However, there is currently no standardized or consistent graphical representation for effectively illustrating these results.

Here we propose a standardized approach for visualizing the effect size of risk variants across the allele frequency spectrum. The proposed plots have a distinctive trumpet shape, with the majority of variants having low frequency and small effects, while a small number of variants have higher frequency and larger effects. These plots, which we call ‘trumpet plots’, can help to provide new and valuable insights into the genetic basis of traits and diseases, and can help prioritize efforts to discover new risk variants. To demonstrate the utility of trumpet plots in illustrating the relationship between the number of variants, their frequency, and the magnitude of their effects in shaping the genetic architecture of complex diseases and traits, we generated trumpet plots for more than one hundred traits in the UK Biobank. To facilitate their broader use, we have developed an R package ‘TrumpetPlots’ and R Shiny application, available at https://juditgg.shinyapps.io/shinytrumpets/, that allows users to explore these results and submit their own data.

Research Area: Genetics and Genomics

Classifications: Genetics, Bioinformatics, Biomedical Science, Statistics and Probability

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
STATEMENT OF NEED

Visualizations are powerful tools that have helped the field of genetics to better understand and communicate complex findings. By using visual aids like Manhattan and Volcano plots, genetic variants identified through genome-wide association studies can be more easily pinpointed. With the advancement of genome-wide association and sequencing studies, a mounting number of significant genetic variants, both common and rare, are being discovered. To better understand the relationship between these variants, combining these findings into single visualizations help to observe the relationship between effect size and allele frequency, providing a clearer picture of the genetic architecture of different traits and diseases. However, there is currently no consistent method for illustrating these results. In this paper, we propose a standardized approach for visualizing the effect size of risk variants across the allele frequency spectrum, generate plots for over a hundred traits in the UK Biobank, and provide to the field a R package and R Shiny application to explore their own results.

Availability of supporting source code and requirements

- Project name:
  - R package available in project ‘TrumpetPlots’
    https://gitlab.com/JuditGG/trumpetplots
  - R shiny app and analyses in the UK Biobank available in project ‘freq_or_plots’
    https://gitlab.com/JuditGG/freq_or_plots
- Project home page: https://juditgg.shinyapps.io/shinytrumpets/
- Operating system(s): Platform independent.
- Programming language: R
- RRID: Not applicable
- License: MIT

Data Availability and implementation

All data used in this manuscript is publicly available. Rare variant associations are available in supplementary data table 2 of the original publication (DOI: https://doi.org/10.1038/s41586-021-04103-z), GWAS summary statistics are available in the website https://www.nealelab.is/uk-biobank/.

**List of abbreviations**

GWAS: genome-wide association study  
LD: linkage disequilibrium  
Log: logarithm or logarithmic

**Competing interests**

The authors declare no competing interests

**Authors’ contributions**

**LC**: Data curation, Formal Analysis, Investigation, Writing – original draft.  
**LL**: Software, Validation, Visualization.  
**PFO**: Conceptualization, Funding acquisition, Formal analysis, Supervision, Writing – review & editing.  
**JGG**: Conceptualization, Data curation, Formal Analysis, Investigation, Software, Validation, Supervision, Visualization, Writing – original draft, Writing – review & editing.

**Acknowledgements**

We thank the participants in the UK Biobank and the scientists involved in the construction of this resource. This work was supported by a grant from the National Institute of Health (R01MH122866) to PFO, by a 2022 NARSAD Young Investigator Grant (Number 30749) by the Brain & Behavior Research Foundation to JGG, and through the computational resources and staff expertise provided by Scientific Computing and the Data Ark (Data Commons) teams at the Icahn School of Medicine at Mount Sinai.
We would like to thank Dr. Shea Andrews for helpful discussions on several aspects of the project. We would also like to express our gratitude to the Center for Excellence in Youth Education (CEYE) program for their support and training, which enabled us to carry out this research. Without the invaluable assistance and dedication of CEYE staff, this project would not have been possible.

Authors’ information

Lucia Corte
Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, New York City, NY 10029, USA and Center for Excellence in Youth Education, Icahn School of Medicine, Mount Sinai, New York City, NY 10029, USA
Email: lc105@wellesley.edu
ORCID ID: https://orcid.org/0009-0005-0963-4880

Lathan Liou
Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, New York City, NY 10029, USA
Email: lathan.liou@icahn.mssm.edu
ORCID ID: https://orcid.org/0000-0002-8066-5947

Paul F. O’Reilly
Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, New York City, NY 10029, USA.
email: paul.oreilly@mssm.edu
ORCID ID: https://orcid.org/0000-0001-7515-0845

Judit García-González,
Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, New York City, NY 10029, USA.
email: judit.garciagonzalez@mssm.edu,
ORCID ID: https://orcid.org/0000-0001-6245-740X
Background

Results visualization is an essential tool for interpreting complex data. By using visual representations such as graphs, charts, and plots, researchers can quickly identify patterns, trends, and outliers that may not be detected in tables of raw data. Visualizations help to gain a more intuitive and comprehensive understanding of the data, especially when dealing with large and complex datasets. Furthermore, visualizing results can facilitate communication of research findings to a broader audience, including non-experts [1]. Visual representations are often more accessible and engaging, making it easier for others to understand and appreciate the significance of the research conducted.

In the field of genetics, the use of visualizations has revolutionized the interpretation and communication of research findings. Over the past two decades, visualizations such as Manhattan plots [2], which display the results of genome-wide association studies (GWAS), software packages like haploview [3] that analyze and visualize linkage disequilibrium (LD) patterns of GWAS associated loci, and Volcano plots that assess patterns of differential gene expression [4], have all played crucial roles in illustrating and sharing the key summaries of data that have advanced the field. These and other visualizations [5] have allowed the genetics field to more easily identify candidate causal variants, relevant genes, and potential outliers that may not be apparent in tables of GWAS summary statistics or differential expression results.

More recent advances in GWAS and sequencing studies [6]–[8] have resulted in the identification of an increasing number of significant genetic variants, including both common [8] and rare variants [5,6]. Researchers are now starting to combine these findings into single visualizations to observe the relationship between effect size and allele frequency across the full range of significantly associated variants. Given that the number of risk-conferring variants, their frequency in a population, and their effect size can vary across different diseases and traits, using these plots can provide a better and instantaneous understanding of their relative genetic architecture. Recent studies on height [9], schizophrenia [10] and coronary artery disease [11], [12] have already included this full range as the main figure, highlighting the utility of this type of visualization. However, a formal and consistent method for illustrating these results has not yet emerged.

The aim of this work is to introduce an R package and R shiny application to illustrate the distribution of risk variants across a wide range of allele frequencies. We term the resulting plots ‘trumplet plots’, due to their trumpet-like shape. To demonstrate their utility, we generated
trumpet plots for over one hundred continuous traits available in the UK Biobank [13], illustrating the distribution of risk variants across an effect allele frequency range between 0.00001 and 1. These plots are available at https://juditgg.shinyapps.io/shinytrumpets/ and we illustrate a single trumpet plot combining the results of all of these in Figure 1.

We propose that trumpet plots are valuable representations of genetic associations across the full allele frequency spectrum that can help researchers to better understand the genetic architecture of traits and diseases and potentially aid in study design and the prioritisation of investments to discover new variants that contribute to disease.

Methods

In the following sections, we will explain the various decisions we made when creating trumpet plots. These decisions include selecting the appropriate scale to represent allele frequencies, deciding whether to use the full GWAS summary statistics or independent GWAS variants, addressing issues related to the reporting of rare variant association tests, determining whether to include power curves, and considering the effect size sign of the variants included. By carefully considering each of these factors, we aimed to create informative and visually appealing trumpet plots that illustrate the effect size of genetic associations across a wide range of allele frequencies.

Using the logarithmic vs linear scale to represent allele frequencies

In the representation of allele frequencies, the range of values can vary greatly between the smallest and largest frequency. When these associations are plotted on a linear scale, the rare variants can be obscured or difficult to distinguish. To address this issue, we recommend using a logarithmic (log) scale (we use log base 10) for allele frequencies in trumpet plots. Compared to a linear scale, the log scale uses increments that represent a relative increase or decrease, rather than a fixed value increase or decrease. The log scale compresses the allele frequencies that are most common, which results in a more even distribution of values across the scale. This scale of visualization facilitates the identification of important patterns and trends.

Identification of independent significant variants to enhance the interpretation of trumpet plots

Genetic association studies involve testing up to millions of genetic variants for their association with a particular trait or disease. However, many of these variants are correlated with each other due to their physical proximity on the genome, which is known as linkage disequilibrium (LD).
This means that many variants nearby to causal variants often show significant associations with the trait under study, due only to their correlation rather than any biological involvement with the trait.

Two methods that can be used to identify independent significant variants in a GWAS are clumping and conditional analysis [14], [15]. Clumping involves selecting a subset of independent significant variants by choosing a lead variant for each LD cluster and then discarding all other variants in that cluster. The lead variant is typically the one with the strongest association with the trait of interest. Conditional analysis, on the other hand, involves identifying independent significant variants after performing a joint analysis of multiple variants together. In this approach, the effect of one variant is conditioned on the effect of other variants, meaning that the association between the trait of interest and one variant is evaluated after accounting for the effect of other variants. This can either be performed as a joint analysis (e.g. a regression) with multiple variants in one model or else as an iterative process, where the lead variant and each other variant in the region are tested jointly one-by-one. By considering the effects of multiple variants jointly in single models, conditional analysis helps to identify independent signals more accurately than clumping, which relies only on correlations between variants to indirectly infer independent signals.

Since the correlation between variants can make it challenging to interpret trumpet plots, we recommend plotting only independent significant variants. These are variants that represent distinct genetic associations with the trait of interest.

**Variant-level associations for low allele frequencies**

While GWAS are a valuable tool for detecting common genetic variants associated with complex traits or diseases, they have limited power to identify associations with rare variants. To address this limitation, sequencing studies [16]–[18], such as whole exome sequencing and whole genome sequencing, have been used to detect variant-level associations with low allele frequencies.

To further enhance the statistical power of rare-variant associations, a commonly used strategy is to aggregate the rare variants detected into functional genetic units, such as genes, and perform collective variation analysis (e.g. gene burden tests [19]–[21]). However, we found that the reporting of rare variant association analyses varies across studies. Some studies report
results at the variant level, while others only report results of the functional unit in which rare
variants were aggregated. This makes comparing results across independent studies, and
determining the functional significance of specific variants, challenging.

We therefore encourage the reporting of results at the variant level, so that they can be included
in visualizations of allele frequency in relation to effect size, such as trumpet plots, aiding study
comparisons and variant interpretation. Although the analysis of genetic variants at low
frequency is expected to improve with the availability of biobank-scale samples and the
development of new methods to reduce biases in association tests, caution should still be
exercised when inspecting associations with rare variants, as these can suffer from instability
and low power, particularly in the relation to binary traits.

Statistical power considerations
GWAS require careful consideration of statistical power [22], which depends on various factors,
including the allele frequency and effect size of variants, represented by the x- and y-axis of
trumpet plots, respectively. Common variants, usually defined as having allele frequency greater
than 1%, tend to have higher power in association studies because common causal variants are
more likely to be present in the sample (either genotyped or imputed), and because their
relatively balanced number of alleles is akin to having a larger sample size. Variants with larger
effect sizes have higher power because their effects are further from the null hypothesis of zero
effect.

We therefore recommend incorporating power curves into trumpet plots, since they visually
represent the statistical power across the allele frequency spectrum for a given sample size and
effect size [23]. Moreover, power curves can aid in identifying parts of the association testing
space in which the power to detect significant associations is low.

Two alternative approaches to illustrate the joint distribution of allele frequency and effect size
One approach to illustrate the relationship between allele frequency and effect size is to plot
only positive effects i.e the allele effect for each variant that increases the value of the
phenotype. In this case, the effect sizes are always positive, and both the allele and sign of the
association regression coefficients (betas or Odds Ratios, ORs) need to be flipped (to the other
allele) if they are reported as negative to ensure that the effect size is greater than zero. If this
‘flipping’ is required, then the allele frequency of the other allele should be reported, which will
be 1 minus the original allele frequency. In this case, the allele frequencies of the plot range from 0 to 1.

The other approach, which we recommend, allows for both positive (risk allele in the context of disease phenotypes) and negative (protective allele in the context of disease phenotypes) effect sizes and always corresponds to the minor allele. In this case, the effect size of the allele can have either a positive or negative value, and the allele frequencies of the plot range from 0 to 0.5.

**Practical example: Generating trumpet plots for 129 traits in the UK Biobank**

We examined all continuous UK Biobank traits with available GWAS analyses performed by Benjamin Neale’s group ([https://www.nealelab.is/uk-biobank/](https://www.nealelab.is/uk-biobank/)) and searched whether rare variant associations were available for the same trait (by UK Biobank Field ID) in the exome sequencing analysis performed by the Regeneron team [13].

Common variant associations were extracted from the Neale’s group GWAS summary statistics. For each GWAS, we extracted the independent variants using COJO GCTA (--cojo-slct command), and a random subset of 4,000 unrelated individuals with European ancestry from the UK Biobank as LD reference panel. We selected independent variants with minor allele frequency >0.01 and association P-value < 5x10^{-8} within a 100Kb window.

Rare variants association results were extracted from the supplementary data table (SD2) of the Regeneron study [13]. This study reports results for both burden tests (which typically aggregated variants and indels) and individual rare variant level tests. To ensure that effect sizes reported in our analyses corresponded to individual rare variants, we extracted only results for ‘singleton variants’ with predicted loss of function - including stop-gain, frameshift, stop-lost, start-lost and essential splice variants - and deleterious missense variants. Interactive plots illustrating the relationship between allele frequency (x-axis) and odds ratio (y-axis) were plotted using our R package TrumpetPlots ([https://gitlab.com/JuditGG/trumpetplots](https://gitlab.com/JuditGG/trumpetplots)), which uses the R packages data.table, ggplot and ggplotly (**Figure 1**).
Figure 1. Distribution of allele frequencies and effect sizes for genetic associations across 129 continuous traits from the UK Biobank. Power curves for statistical power of 0.5 (blue), 0.7 (purple) and 0.9 (pink) were constructed for the median GWAS sample size (N=351,550).

R Shiny application

We developed a user-friendly web application called *shinytrumpets* to visualize trumpet plots for our UK Biobank results, as well as any other genetic association results that can be uploaded by the user. With *shinytrumpets*, researchers with no knowledge of R programming can easily upload and visualize their own datasets.

If a user uploads their own results, *shinytrumpets* prompts them to upload the input data files and specify the sample size used for the study, such as the GWAS sample size. This information is used to perform power calculations for the visualization. *Shinytrumpets* offers an intuitive interface for users to explore and download trumpet plots.

Discussion

Visual representations of genetics and genomics results, such Manhattan [2], Q-Q plots [2], haploview [3] or Volcano plots [4], have been helpful in interpreting research findings and
identifying patterns, trends, and outliers that may not be easily apparent in tables of raw data. These visualizations have revolutionized the interpretation and communication of research findings relating to the identification of GWAS associated loci, putatively causal genes, and potential outliers.

In this manuscript, we introduce a new R package and shiny application to illustrate the distribution of risk variants across a wide range of allele frequencies, which we coin 'trumpet plots'. We illustrate the distribution of variant effect sizes across the allele frequency range (from 0.00001 to 1) for over 100 continuous traits available in the UK Biobank, and propose that these plots are valuable representations of genetic associations that can help researchers better understand the genetic architecture of traits and diseases and prioritize certain study designs (e.g. sequencing or GWAS) to discover new variants that contribute to disease.

One important consideration when interpreting the trumpet plots we constructed for the UK Biobank is that they only represent individuals of European ancestry. The relationship between effect size and allele frequency can be affected by population genetic differences [24], [25] and as such, one interesting application of trumpet plots could be to compare the joint distribution of allele frequencies and effect sizes across different ancestries to identify similarities and differences for further investigation. Insights about the similarities and differences across populations in the relationship between effect size and allele frequency could have important implications for disease risk prediction and prevention strategies.

In conclusion, we emphasize the significance of data visualization in the genetics field and present a novel R package and shiny application for visualizing the relationship between allele frequency and effect size in association studies. We hope that the proposed ‘trumplet plots’ will provide a valuable representation of genetic associations and will enhance the interpretation of the association results across the allele frequency spectrum.
References


