SumStatsRehab: An Efficient Algorithm for GWAS Summary Statistics Assessment and Restoration

Puya Yazdi (puya@selfdecode.com)
SelfDecode.com
Manfred Grabherr
SelfDecode.com
Biljana Novković
SelfDecode.com
Umar Khan
SelfDecode.com
Varuna Bamunusinghe
SelfDecode.com
Andrew Terpolovsky
SelfDecode.com
Karatuğ Ozan Bircan
SelfDecode.com
Carlos Tello
SelfDecode.com
Lewis Cuthbertson
SelfDecode.com
Abdallah Mahmoud
SelfDecode.com
Madhuchanda Bose
SelfDecode.com
Mykyta Matushyn
SelfDecode.com

Research Article

Keywords: Bioinformatics, GWAS, Summary statistics, PRS, genetics

Posted Date: March 2nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1359902/v1

Abstract

Background: Generating polygenic risk scores for diseases and complex traits requires high quality GWAS summary statistic files. Often, these files can be difficult to acquire either as a result of unshared or incomplete data. To date, bioinformatics tools which focus on restoring missing columns containing identification and association data are limited, which has the potential to increase the number of usable GWAS summary statistics files.

Results: SumStatsRehab was able to restore rsID, effect/other alleles, chromosome, base pair position, effect allele frequencies, beta, standard error, and p-values to a better extent than any other currently available tool, with minimal loss.

Conclusions: SumStatsRehab offers a unique tool utilizing both functional programming and pipeline-like architecture, allowing users to generate accurate data restorations for incomplete summary statistics files. This in turn, increases the number of usable GWAS summary statistics files, which may be invaluable for less researched health traits.

Background

A major goal of modern precision medicine is to accurately predict individual health risks based on genetic data [1]. Alongside the advent of NGS technologies has come a plethora of discoveries linked to Genome-wide association studies (GWAS) [2]. The large amount of data generated by these studies has enabled researchers to apply statistical techniques in order to generate polygenic risk scores (PRS) [3].

These scores can be used to predict an individual's genetic risk of a particular health condition. An example of a method that can generate PRS is PUMAS. PUMAS uses trait-specific GWAS summary statistics files for training, in order to fine-tune its predictive model [4]. The core limitation of these techniques is the availability of high-quality GWAS summary statistics.

Summary statistics are used to convey key GWAS data such as variant ID (rsID), chromosome number (Chr), base pair position (BP), effect allele (EA), other allele (OA), minor allele frequency (MAF), t-statistics, p-value and standard error (StdErr). However, summary statistics from GWAS are often not shared, and there is no universally standardized format, even with regards to what data is reported and what is not [5–7]. GWAS summary statistic files are also often presented in a multitude of tabular formats, including plink, CTA, BOLT-LMM, GEMMA, Matrix eQTL, METAL, and VCF [6, 7]. As a result, some of the information needed for meta-analysis or downstream GWAS applications–rsID, Chr, BP, OA, EA, MAF, StdErr, Beta, p-value–may be missing from the files [6]. Additionally, variation in methodologies of genotyping arrays and quality control filters used by different research groups may contribute to missing SNP identification or association data [8].

Missing columns of data may influence the predictive power of techniques used to generate a PRS, or even render the GWAS file unusable. This can be a particular problem if there are a limited number of

high-power studies for a particular trait of interest, which is often the case with the majority of GWAS publications not publicly sharing their data [5]. For these files, restoration would be particularly prudent.

To date, there are no tools which restore this data perfectly. Murphy et al. recently developed MungeSumstats, an R software package, which manually standardizes and performs quality control on different GWAS summary statistic files [7]. This tool performs several quality control steps in order to ensure all key data is present and consistently formatted. Part of this quality control includes restoration of some incomplete data columns, though this is not the main function of the tool. Additionally, the quality of restoring RSID using MungeSumstats is limited by the most recently curated version of the SNPlocs database [9]. To our knowledge, no tool currently attempts to restore missing standard error, Beta, or p-values.

In order to address these issues, we developed SumStatsRehab. This tool is able to perform restoration of rsID, chromosome, base pair position, effect allele frequencies, back-calculation of t-statistics from p-values, beta value restoration, and standard error calculations and corrections. Once restored, the output is presented in a consistent tabular form. Additionally, SumStatsRehab can diagnose cases where critical data cannot be restored in a given GWAS summary statistics file, and can thus be used for both quality control and cleanup of files. In this paper, we describe SumStatsRehab, its features and utility. We also provide a comparison with the only current alternative for summary statistics restoration, MungeSumstats. The source code for SumStatsRehab is found at https://github.com/Kukuster/SumStatsRehab.

Methods

Implementation

SumStatsRehab is written in Python3, and utilizes several native Linux executables. The key functions of SumStatsRehab are assessment, validation and restoration (Fig. 1). These functions can be implemented for chromosome, base pair position, rsID, effect allele, other allele, allele frequency, standard error, beta, and p-value. Each category of data in the input GWAS summary statistic file is assessed, validated, and restored independently. SumStatsRehab accepts GWAS summary statistic files with single nucleotide polymorphisms (SNPs) which reference human genome build 36, build 37, and build 38, and can output restored summary statistics files in reference builds 37 and 38. SumStatsRehab uses a .json header file to correctly read and interpret the columns in the input summary statistic file.

Assessment and validation of Summary Statistics Files

SumStatsRehab can be used to identify any invalid SNPs in a GWAS summary statistic file; invalid SNPs are those which are missing any core data such as variant ID. This enables users to determine the number and cause of missing or invalid SNPs (Fig. 2).

To demonstrate this command, we tested it on an example GWAS summary statistics file (GWAS blood pressure [10]). As shown, SumStatsRehab identified that less than 1% of entries for GW-significant SNPs were missing Fig. 2A), and that the majority of missing entries were rsIDs (Fig. 2B). The resultant plots derived using the diagnostic tools assess the number of invalid SNPs by significance level, showing the potential impact of the incomplete data columns on downstream calculations. The results of this diagnostic are used internally to guide and optimize restoration.

Restoration of Summary Statistics Files

SumStatsRehab only attempts to restore entries identified as invalid, with one exception. When either the base pair position or chromosome is invalid, SumStatsRehab restores both by looking up the rsID associated with that entry, and overwriting the chr and base pair position entries.

The extent of restoration possible is dependent on the inputs to SumStatsRehab. If only the summary statistics file is input, SumStatsRehab will be able to perform restoration of the p-values, betas and standard errors given two out of three of these values are present. The additional input of a dbSNP file in the target human genome reference build is optimal for restoration. SumStatsRehab preprocesses the dbSNP file, organizing it by rsID, chromosome, base pair position, alleles, ref/alt, and frequencies associated with each SNP, sorted by chromosome and base pair position, and by rsID. If the target build and the GWAS summary statistic file builds are different, an additional third input, the 'chain file' is needed for liftover from the summary statistic file build to the target build. With these inputs, SumStatsRehab is able to restore GWAS data files, and include EAF, missing t-statistics, rsID or chromosome numbers and base pair positions, effect allele (EA) and other allele (OA).

Preparation of test case files

To assess the utility of our tool and the extent of restoration it can achieve, we chose publicly available and complete summary statistics files from three different GWAS as test cases: : 1) blood pressure, 2) C-reactive protein, 3) allergies [10–12]. These files were preprocessed by removing one specific column of data per file at a time: rsID, chromosome number, base pair position, effect or other allele, allele frequency, p-value, beta, and standard error. After removing each column from the three different GWAS summary statistics files to generate a total of 9 test files per GWAS, with a total of 27 test files, we ran each file through both SumStatsRehab and the only current alternative, MungeSumstats.

In order to be run through MungeStumstats, test files required an extra round of extensive preprocessing. For the blood pressure test files, all columns were renamed in accordance with the MungeSumstats documentation [7]. For the GWAS allergies test files, all fields containing 'NA' had to be replaced with a placeholder dot, and all rows with any non-numeric value in BP fields had to be removed. In both cases, necessary preprocessing required manual deletion of SNPs, for which the missing or invalid data could be restored, to allow MungeSumstats to proceed with restoration for the remainder of the test files.

Additionally all non-traditionally formatted rsIDs e.g. “esv3584976”, were removed to prevent the automatic failure of the program.

Assessment of SumStatsRehab and comparison with MungeSumstats

To assess the restoration of both tools, two different metrics were used. For qualitative attributes, accuracy was assessed in terms of concordance with the original summary statistics file. This was used for the chromosome, base pair location, effect allele, and other allele columns. For quantitative attributes, we calculated the difference between the predicted values and the original, masked values using formula 1, which yields an accuracy score between 0 and 1. This was used to calculate the relative accuracy for the allele frequency, beta, standard error, and p-value columns, in order to account for floating point arithmetic and rounding errors.

Formula 1. Where x_o and x_r are the original and restored values, and k is a fudge factor/an error term, which is different for each column. For allele frequency column: k = 2, for beta column: k = 6, for standard error column: k = 4, for p-value column: k = 3.

$$1 - min(k|x_0 - x_r|, 1)$$

The overall accuracy for each column was calculated as the average of the accuracy metrics for each entry. These results were then used to assess and compare the restoration process of both SumStatsRehab and MungeSumstats.

We did not use any accuracy metrics with respect to evaluating restoration of rsIDs, as the rsID restoration is dependent on the publication timeframe of the GWAS. For earlier GWAS, rsID names do not correspond well to more current dbSNPs databases; the differences in rsID may not reflect differences in accuracy of restoration but differences in dbSNP versions.

Results

Restoration using SumStatsRehab

SumStatsRehab was able to successfully restore rsID, effect/other alleles, chromosome, base pair position, effect allele frequencies, beta, standard error, and p-values for all 27 test files. These restorations occurred without any SNP loss (Table 1). As the original CRP file was missing variant ID data, Chr and BP restoration could not be assessed and was input as N/A. SumStatRehab managed to restore on average 97.61% of rsIDs accurately (Fig. 3). The 2.39% discrepancy in rsID accuracy may be attributed to variations found in the earlier dbSNP dataset version used in the original GWAS’. Restoration accuracy was also greater than 94% for chromosome, base pair position, other allele, standard error, and p-values,

while EA, EAF, and beta had restoration accuracies of 77.61%, 72.3% and 54.2%, respectively (Fig. 3). The lower accuracy of EA is in part due to using different SNP databases with a greater number of alternate allele possibilities than the older databases which these studies reference. The reduced restoration accuracy of EAF is a function of population-dependent differences in EAF; restored EAFs are a naive approximation of the EAF as we don't know the population-specific composition of the GWAS samples. Beta restoration at 54.2% can be attributed to unsigned standard error values, leading to inverse beta scores which prevented an exact match (supplementary information). Caution should be taken to only attempt beta restoration when signed standard error data is provided.

Table 1
Total % of SNPs removed per GWAS summary statistics file for restoration runs by SumStatsRehab and MungeSumstats.

	MungeSumStats			**SSrehab**		
	Allergies	Blood pressure	CRP	Allergies	Blood pressure	CRP
rsID	6.00%	12.05%	11.40%	0%	0%	0%
Chr	6.50%	Fail	N/A	0%	0%	N/A
BP	6.50%	Fail	N/A	0%	0%	N/A
EA	8.70%	Fail	Fail	0%	0%	0%
OA	6.50%	Fail	Fail	0%	0%	0%
MAF	N/A	N/A	N/A	0%	0%	0%
t-statistics	N/A	N/A	N/A	0%	0%	0%
p-value	N/A	N/A	N/A	0%	0%	0%
StdErr	N/A	N/A	N/A	0%	0%	0%

Comparison of SumStatsRehab and MungeSumstats

As MungeSumstats is the only current alternative tool for data restoration, we also attempted restoration runs with MungeSumstats on the prepared test files. This comparison was performed using all 27 previously described test files derived from 3 GWAS, in order to compare the efficiency of restoring 9 data categories (Table 2). We assessed the comparison on the basis of restoration accuracy and data loss.

Table 2
Comparison of supported restoration categories for SumStatsRehab and MungeSumstats

Column restored	MungeSumstats	SumStatsRehab
rsID	Yes	Yes
Chr	Yes	Yes
BP	Yes	Yes
EA	Yes	Yes
OA	Yes	Yes
MAF	No	Yes
t-statistics	No	Yes
p-value	No	Yes
StdErr	No	Yes

MungeSumstats was initially unable to restore any of the test files associated with the blood pressure and CRP GWAS'. Despite significant preprocessing to prevent automatic failures, MungeSumstats still failed to restore all columns with the exception of rsID for these files (Table 1).

MungeSumstats removed 12.05% and 11.4% of SNPs, for the blood pressure and C-reactive protein files respectively, relative to the 0% loss achieved by SumStatsRehab (Table 1). Removed entries included rows where all entries were correct. For the allergy GWAS test files, MungeSumstats removed up to 8.7% of SNP entries, while restoring chromosome number, base pair position, rsID and allele columns. MungeSumstats does not restore missing allele frequencies, standard error, beta values and p-values; we were thus unable to compare our tool against MungeSumstats for these test cases. For all runs on the GWAS test files, SumStatsRehab had greater restoration accuracy than MungeSumstats for all categories other than effect allele (Fig. 3).

Comparison of Computational Load

We also compared SumStatsRehab and MungeSumstats on the basis of execution time and memory usage. MungeSumstats had significantly lower run time for all restorations. When running tests sequentially or in parallel, the average execution time of MungeSumstats was around 6 minutes 53 seconds. Execution times for SumStatsRehab were significantly higher, with an average of 18 minutes 5 seconds, when running up to 6 tests in parallel, and 35 minutes when running all tests sequentially. The current implementation of SumStatsRehab runs processes sequentially, although its architecture leaves

room for introducing parallelism in the future updates, while MungeSumstats architecture currently allows parallelization.

To compare memory usage, as SumStatsRehab can only be run sequentially, we chose to also run MungeSumstats sequentially to allow for equitable comparison. MungeSumstats used 12–16 GB of RAM during execution with rare drops into the 1–5 GB range, with a peak memory usage of 25 GB while unsuccessfully trying to restore the blood pressure GWAS effect allele test file. MungeSumstats' system cache usage was 1–2 times the size of the input file, which varied between 0.5–1.5 GB. In contrast, SumStatsRehab uses a maximum of 800 MB of RAM, and cache equivalent to the unpacked input file size.

Discussion

A recent workshop set up to outline the best practice of sharing and standardizing GWAS summary statistics recommended that the following should be mandatory when sharing this data: a form of variant identifier, p-value, effect allele, other allele, effect allele frequency, effect and standard error [13]. While these recommendations will assist researchers going forward, there still remain thousands of incomplete GWAS summary statistics which may be able to provide insightful information for lesser studied health traits had they adhered to the aforementioned suggestions.

One potential avenue to address this issue lies in the restoration of incomplete columns of data. SumStatsRehab was able to restore this data in GWAS summary statistics, including the chromosome, base pair position, rsID, allele frequency, effect and other alleles, beta, p-value and standard error, more accurately and with less loss than any other currently available tool. These entries are key to generating robust polygenic risk score (PRS) models.

References to GWAS variants and their incorporation into PRS often relies upon correct identification of these variants by rsID [4, 14, 15]; SumStatsRehab is able to accurately restore rsID entries by overcoming several challenges associated with standardizing and inferring rsIDs. As SNP databases are updated, many rsIDs have been renamed across the different reference builds and versions, such that several different identifiers may refer to the same SNPs [16]. This may pose an issue when combining multiple GWAS which utilize different reference builds [16]. By rewriting rsID data using a specified reference database version, SumStatsRehab allows users to combine GWAS summary statistics which were generated using different databases. The ability to reliably combine GWAS as a result of proper rsID updating and restoration allows the identification of variants that were individually deemed insignificant in individual studies, but may play some role in disease or trait determination with increased power [17, 18].

Beyond rsID, Standard error, p-values and beta are also important in the estimation of effect sizes when generating polygenic risk scores [14]. This data is often omitted from non-standardized publicly available GWAS summary statistic files, rendering the files unusable. To date, no tool currently allows for any of these three types of data to be restored. However, by utilizing the relationship between these three types

of data [19], SumStatsRehab is able to closely predict the missing values, given that at least two of the three types of data are present; with the ability to restore this data at an accuracy > 99% for all 3 test files. To our knowledge, SumStatsRehab is the only currently available tool which allows users to restore this type of data.

Before data can be restored, first the researcher must identify whether there are any issues with the dataset. While in most cases, incomplete identification and association data such as rsID or standard error may be obvious, this is not always the case. Manually identifying incomplete data entries can be a time-consuming and labor-intensive process, with GWAS summary statistic files now typically containing more than 8 million genetic variants [13]. The diagnose command implemented within the SumStatsRehab workflow provides visual aids, considerably increasing the speed at which issues with these files can be identified, and reducing the overall time spent processing data.

To our knowledge, SumStatsRehab represents the only tool which currently solely addresses data restoration to this extent. Alternative tools which also attempt partial data restoration, such as MungeSumstats, are primarily focused on data standardization. In order to assess the restoration quality of SumStatsRehab, we performed a comparison between this tool and MungeSumstats with regards to data restoration.

One of the major benefits of restoring non-standard GWAS summary statistic files using SumStatsRehab is the ability to specify column names using a JSON file input. This in turn allows users to utilize files with non-standardized layouts and headers. In comparison, MungeSumstats requires column headers to be in a specific format based on the standard input format of IEU GWAS VCF files. When this is not the case, it is unable to identify the effect allele column causing the process to fail.

One area in which MungeSumstats outperformed SumStatsRehab was in the restoration of effect alleles. This may be explained by the fact that when running SumStatsRehab, we employed the latest version of the dbSNP dataset. More recent dbSNP datasets contain additional alternative alleles reducing the likelihood of an exact match between the newly restored file and the older original dataset which may have relied on an earlier dbSNP dataset release. In contrast, MungeSumstats relies on the most up-to-date SNPlocs dataset. These curated datasets inevitably lag behind the release of NCBI’s dbSNP datasets, and contain fewer SNPs. However, SumStatsRehab allows users to implement any build and version of the dbSNP database. So while in this instance the accuracy of restoration may have been reduced, the overall quality of the data may have been improved. Additionally, as SumStatsRehab utilizes only the preprocessed dbSNPs, it can be easily deployed as part of a pipeline without external dependencies.

For all other tested data categories, SumStatsRehab outperformed MungeSumstats in data restoration. However, it must be noted that if the researcher's goal is performing meta analyses or other types of analysis which require standardized data, MungeSumstats is a more appropriate tool.

Conclusion

Overall, SumStatsRehab offers a highly maintainable tool which can easily be optimized for specific use cases with minimal modifications. This tool incorporates functional programming in addition to pipeline-like architecture, to define a flexible framework that is suitable for working on massive scales with various cloud computing platforms with minimal to no refactoring. The combined effect of this is a unique bioinformatics offering which allows users interested in generating PRS a method to increase the likelihood of being able to use GWAS summary statistics for any given health trait.

Declarations

Ethics

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available at the following web link: https://github.com/Kukuster/SumStatsRehab#supplementary-information

Competing interests

The authors declare no competing interests.

Funding

All authors are paid employees of SelfDecode.

Authors' contributions

P.Y., M.M., and B.N. conceived and co-ordinated the study; M.M., L.C., A.M., K.B., A.T., V.B., U.K., M.G., M.B., C.T., B.N., and P.Y. contributed to the development and assessment of the SumStatsRehab tool; M.M., M.B., and L.C. wrote the manuscript; C.T., B.N., P.Y., and L.C. edited and approved the final manuscript.

Acknowledgement

Not applicable.

References

1. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).

2. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).
3. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
4. Zhao, Z. *et al.* PUMAS: fine-tuning polygenic risk scores with GWAS summary statistics. *Genome Biol.* **22**, 257 (2021).
5. Thelwall, M. *et al.* Is useful research data usually shared? An investigation of genome-wide association study summary statistics. *PLoS ONE* **15**, (2020).
6. Lyon, M. S. *et al.* The variant call format provides efficient and robust storage of GWAS summary statistics. *Genome Biol.* **22**, 32 (2021).
7. Murphy, A. E., Schilder, B. M. & Skene, N. G. MungeSumstats: a Bioconductor package for the standardization and quality control of many GWAS summary statistics. *Bioinformatics* **37**, 4593–4596 (2021).
8. Jiang, Y. *et al.* Proper conditional analysis in the presence of missing data: Application to large scale meta-analysis of tobacco use phenotypes. *PLOS Genet.* **14**, e1007452 (2018).
9. Pagès, H. *SNPlocs.Hsapiens.dbSNP144.GRCh37: SNP locations for Homo sapiens (dbSNP Build 144)*. (2017).
10. Wojcik, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514 (2019).
11. Ferreira, M. A. *et al.* Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat. Genet.* **49**, 1752–1757 (2017).
12. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
13. MacArthur, J. A. L. *et al.* Workshop proceedings: GWAS summary statistics standards and sharing. *Cell Genomics* **1**, 100004 (2021).
14. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).
15. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Primer* **1**, 1–21 (2021).
16. Iperen, E. P. A. van, Hovingh, G. K., Asselbergs, F. W. & Zwinderman, A. H. Extending the use of GWAS data by combining data from different genetic platforms. *PLOS ONE* **12**, e0172082 (2017).
17. Wang, M. & Xu, S. Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity* **123**, 287–306 (2019).
18. Asif, H. *et al.* GWAS significance thresholds for deep phenotyping studies can depend upon minor allele frequencies and sample size. *Mol. Psychiatry* **26**, 2048–2055 (2021).
19. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).

Figures

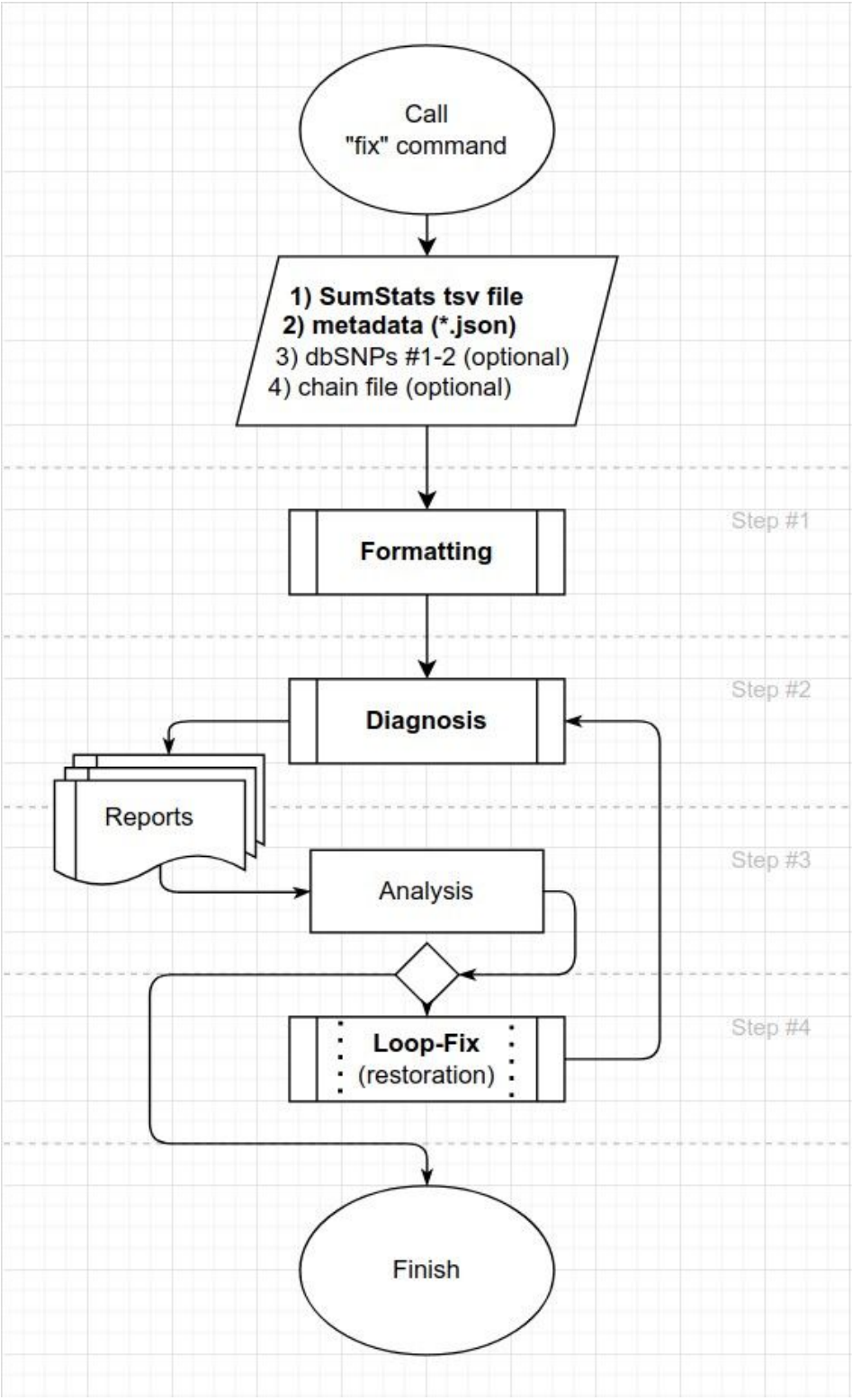

Figure 1

SumStatsRehab implementation pipeline.

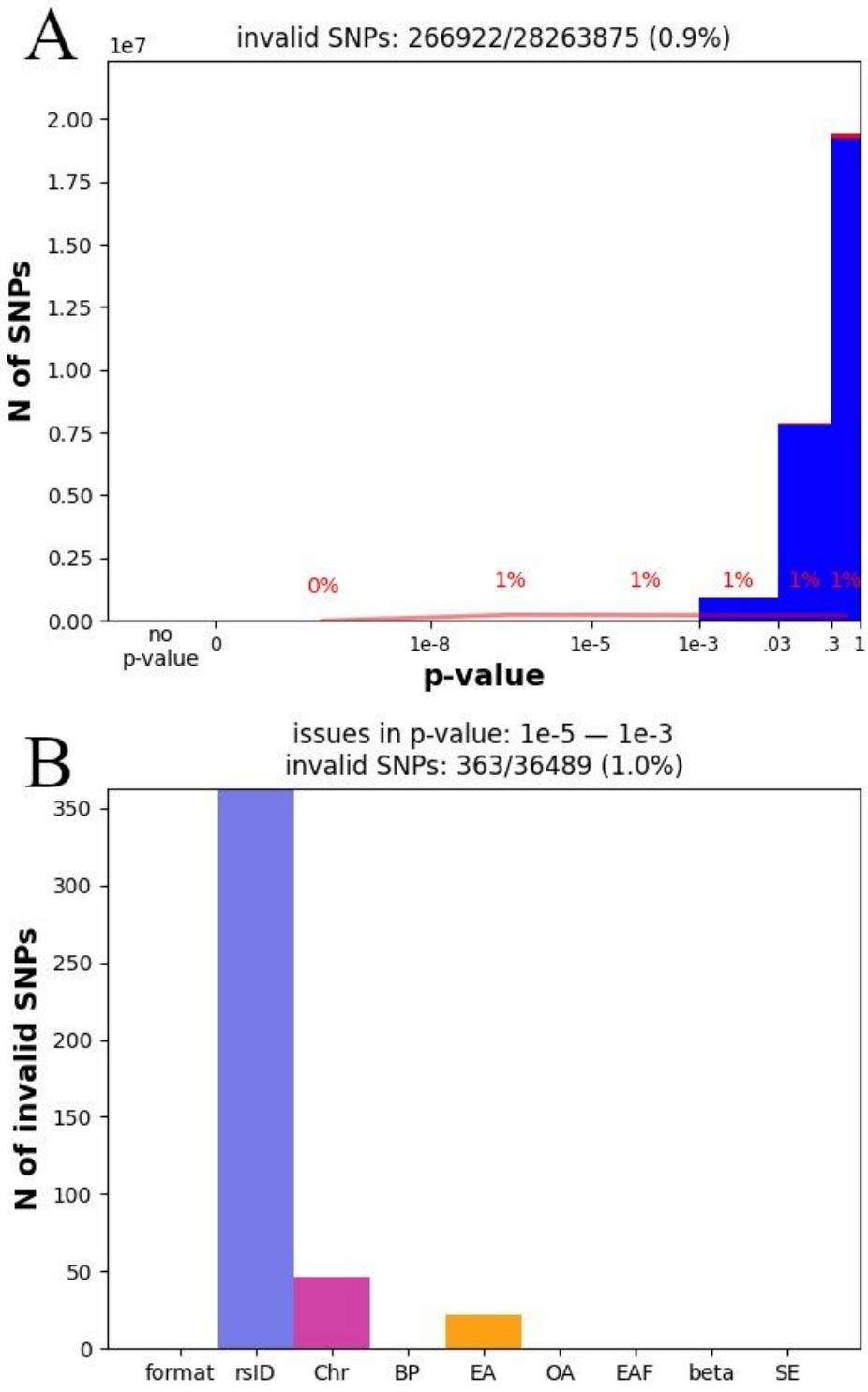

Figure 2

(A) stacked histogram plot - the core plot produced by the “diagnosis” command. It maps the number of invalid SNPs against p-value, allowing assessment of the distribution of invalid SNPs by significance. Valid SNPs are shown as blue, and invalid SNPs are shown as red. **(B)** One of several bar charts produced by the “diagnosis” command. This plot is generated for each bin of the stacked histogram plot (A) and reports the number of issues that invalidate the SNP entries in a particular bin.

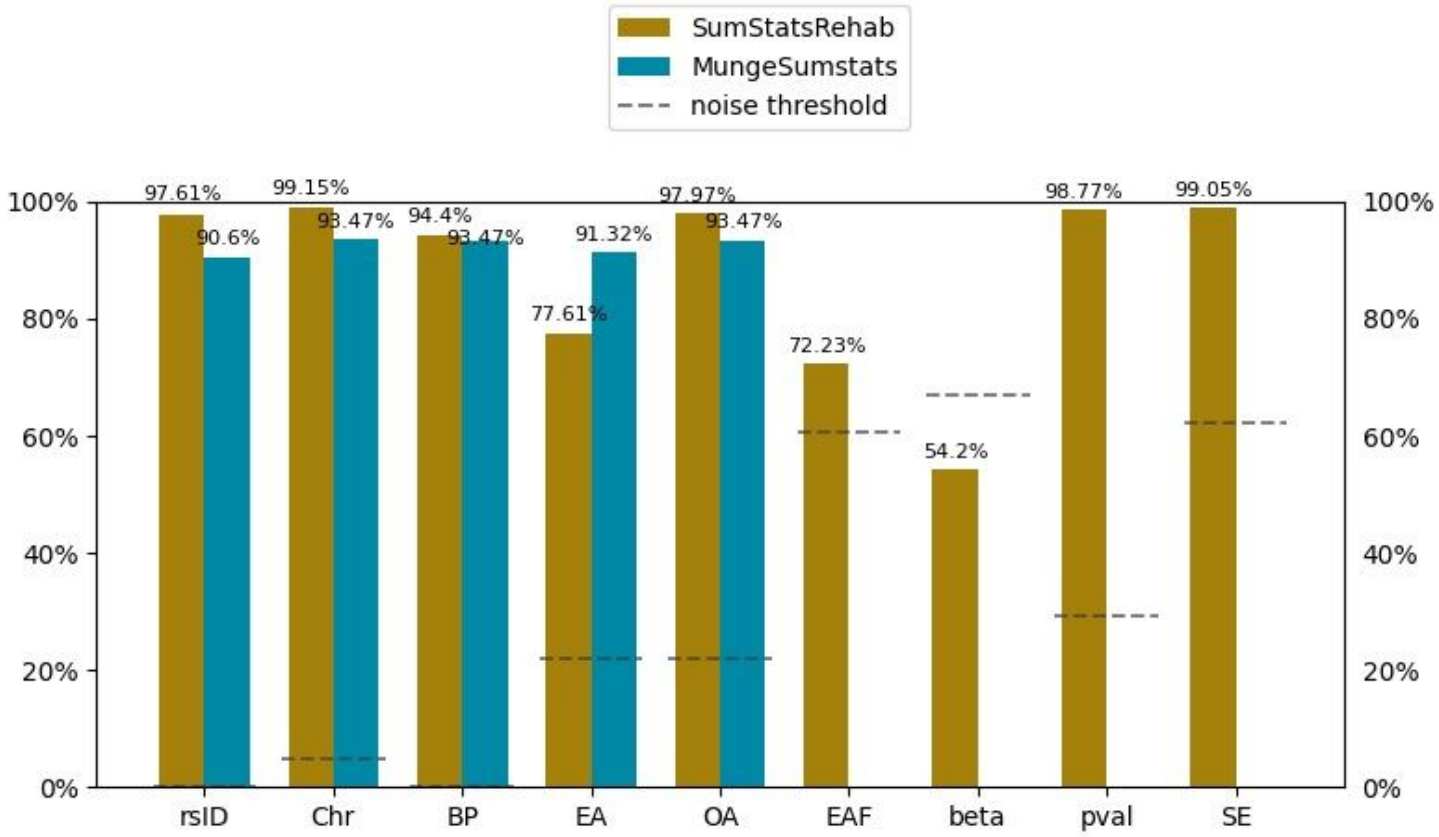

Figure 3

Comparison of SumStatsRehab and MungeSumstats average restoration accuracy. The noise threshold represents the expected level of accuracy achieved by restoring with a random value, and is intended to correct for correct restoration by chance.