PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

# SmartPhase: Accurate and fast phasing of heterozygous variant pairs for genetic diagnosis of rare diseases

**Paul Hager**[1], **Hans-Werner Mewes**[2], **Meino Rohlfs**[3], **Christoph Klein**[3], **Tim Jeske**[1,3]*

**1** Institute of Bioinformatics and Systems Biology, Helmholtz Zentrum München GmbH, Neuherberg, Germany, **2** Technische Universität München, School of Life Sciences, Weihenstephan, Freising, Germany, **3** Department of Pediatrics, Dr. von Hauner Children's Hospital, University Hospital, LMU Munich, München, Germany

* tim.jeske@helmholtz-muenchen.de

## Abstract

There is an increasing need to use genome and transcriptome sequencing to genetically diagnose patients suffering from suspected monogenic rare diseases. The proper detection of compound heterozygous variant combinations as disease-causing candidates is a challenge in diagnostic workflows as haplotype information is lost by currently used next-generation sequencing technologies. Consequently, computational tools are required to phase, or resolve the haplotype of, the high number of heterozygous variants in the exome or genome of each patient. Here we present SmartPhase, a phasing tool designed to efficiently reduce the set of potential compound heterozygous variant pairs in genetic diagnoses pipelines. The phasing algorithm of SmartPhase creates haplotypes using both parental genotype information and reads generated by DNA or RNA sequencing and is thus well suited to resolve the phase of rare variants. To inform the user about the reliability of a phasing prediction, it computes a confidence score which is essential to select error-free predictions. It incorporates existing haplotype information and applies logical rules to determine variants that can be excluded as causing a recessive, monogenic disease. SmartPhase can phase either all possible variant pairs in predefined genetic loci or preselected variant pairs of interest, thus keeping the focus on clinically relevant results. We compared SmartPhase to WhatsHap, one of the leading comparable phasing tools, using simulated data and a real clinical cohort of 921 patients. On both data sets, SmartPhase generated error-free predictions using our derived confidence score threshold. It outperformed WhatsHap with regard to the percentage of resolved pairs when parental genotype information is available. On the cohort data, SmartPhase enabled on average the exclusion of approximately 22% of the input variant pairs in each singleton patient and 44% in each trio patient. SmartPhase is implemented as an open-source Java tool and freely available at http://ibis.helmholtz-muenchen.de/smartphase/.

OPEN ACCESS

**Citation:** Hager P, Mewes H-W, Rohlfs M, Klein C, Jeske T (2020) SmartPhase: Accurate and fast phasing of heterozygous variant pairs for genetic diagnosis of rare diseases. PLoS Comput Biol 16(2): e1007613. https://doi.org/10.1371/journal.pcbi.1007613

**Editor:** Mihaela Pertea, Johns Hopkins University, UNITED STATES

**Received:** April 4, 2019

**Accepted:** December 17, 2019

**Published:** February 7, 2020

**Data Availability Statement:** Data cannot be shared publicly because of European and national data protection regulations that require that genomic data must not be made public to protect the privacy of patients. Data are available from the Dr. von Hauner Children's Hospital, Medical Center of the LMU Munich for researchers who meet the criteria for access to confidential patient data. Please direct your requests to the Dr. von Hauner Children's Hospital (sekretariat.kinderklinik@med.uni-muenchen.de) or the Institutional Review

Board of the LMU (ethikkommission@med.uni-muenchen.de).

**Funding:** The work was supported by funding of The Leona M. and Harry B. Helmsley Charitable Trust, the Care-for-Rare Foundation, and BMBF (PID-NET, 01GM1517A). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

This is a *PLOS Computational Biology* Software paper.

## Introduction

Genetic defects are the source of a wide and diverse range of monogenic or Mendelian diseases that are individually rare but collectively common. So far, more than 5, 000 different disorders and traits are known that are caused by mutations in only one gene [1]. Genome as well as transcriptome sequencing is increasingly used to genetically diagnose patients suffering from a suspected monogenic rare disease [2–4]. However, detecting disease-causing variants among thousands of benign variants is a great challenge. Widely-used strategies and guidelines for variant prioritization are based on the predicted or known deleteriousness of a variant, its frequency in large scale sequencing studies and its segregation with the disease phenotype [5, 6]. Assuming autosomal recessive monogenic inheritance, the disease-causing variants are either homozygous or compound heterozygous with two heterozygous mutations together affecting both parental alleles of a gene locus [7]. Consequently, clinical workflows aim to detect with preference rare variants that are predicted to be harmful and are homozygous or compound heterozygous in the patient. The accurate determination of whether two heterozygous variants are located on the same or different parental alleles is a challenge that is faced by all diagnostic pipelines in the context of recessive monogenic diseases.

Haplotypes can either be resolved experimentally during sequencing or inferred computationally afterwards [8]. Several technologies for haplotype-resolved genome sequencing have been developed but are seldom used in a clinical setting because of their prohibitive cost and complexity. Computational tools for phasing use sequencing data of family members, reads spanning multiple variants or reference haplotype panels. Sequencing data of parents or other family members is most informative for phasing but might not always be available and cannot be used for variants that are heterozygous in both parents and the child. Using reads spanning multiple variants requires no additional data, but the length of the underlying reads limits the number of variants that can be phased. Panel-based phasing methods are useful for common variants but fail for rare variants which are the focus when diagnosing rare diseases. The combination of different phasing strategies is promising as it can compensate for the disadvantages of the individual approaches.

Existing phasing tools offer limited utility for clinical purposes because they are designed to phase complete chromosomes instead of genetic loci of interest or incorporate only one phasing strategy. phASER improves the phasing range of read-based phasing by incorporating RNA sequencing reads in addition to DNA sequencing reads, but it does not perform pedigree-based phasing [9]. WhatsHap combines read-based phasing with pedigree-based phasing but offers no options to restrict phasing to pre-selected variants or genomic regions [10]. The user would either have to accept unnecessarily long runtimes for phasing complete chromosomes or trim the sequencing data to the regions of interest before each execution which would require additional time and storage resources. Neither option is feasible in a clinical setting especially when dealing with large cohorts of thousands of patients with few regions of interest. Furthermore, none of these phasing tools are able to label pairs of heterozygous variants as clinically irrelevant by using the fact that the genotypes of healthy parents contradict the potential pathogenicity of the pair.

To overcome these limitations, we developed SmartPhase, a ready-to-use phasing tool tailored for clinical workflows to improve the analysis of potential compound heterozygous variant pairs in terms of simplicity, speed and accuracy. SmartPhase is able to flexibly use available trio sequencing information and read information of DNA as well as RNA sequencing data.

Additionally, it informs about the confidence of its predictions and implements rules to logically exclude variant constellations that cannot be disease-causing.

## Design and implementation

To fully take advantage of the breadth of phasing informative data generated in clinical research, SmartPhase is able to combine trio phasing, read-based phasing, additional logical rules and GATK physical phasing to resolve as many variant combinations as non-pathogenic as possible. Furthermore, SmartPhase focuses on diagnostically relevant genomic loci in its input while providing a comprehensive bitflag and confidence score system to intuitively represent its results in its output.

### Trio phasing

If parental genotypes are provided, all patient heterozygous variants are examined for the possibility of using the parental variant calls to allocate a variant to either the maternal or paternal chromosome. If the pedigree information allows the phase to be determined, a confidence score of 1.0 is given.

Generally, it is not possible to assign *de novo* mutations to the correct allele by trio information. However, if the inherited allele is present in both parents, with one being heterozygous and the other being homozygous for said allele, it is probabilistically assigned to the homozygous parent with a confidence score of 0.66.

### Read-based phasing

If DNA or RNA sequencing reads are provided, SmartPhase uses multi-variant spanning reads to aid in the resolution of local haplotypes. If a read spans two variant positions and contains both variants, this constitutes an evidence that both variants lie on the same allele in *cis* configuration. If two reads span both variant positions, each containing only one of the variants, this is an evidence that the variants lie on opposing chromosomes in *trans* configuration. To ensure the creation of accurate haplotypes, SmartPhase ignores reads that are not mapped, not part of a proper pair in case of paired-end reads, marked as duplicate reads, or not part of a primary alignment. Further, the user can choose to ignore reads whose mapping quality is lower than that of a defined threshold.

In order to inform about the quality of the inferred haplotype for two variants $v_1$ and $v_2$, a confidence score, Confidence($v_1$, $v_2$), is computed by the formula

$$\frac{\left| \overbrace{\sum_{i=1}^{n_{trans}} (1 - \underbrace{\frac{\frac{\overbrace{\sum_{k=1}^{l_{v_1}} 10^{-q_k/10}}^{\text{Corrected } v_1}}{l_{v_1}} + \frac{\overbrace{\sum_{k=1}^{l_{v_2}} 10^{-q_k/10}}^{\text{Corrected } v_2}}{l_{v_2}}}{2}}_{\text{Average Inverse Phred}})}^{\text{Trans Subscore}} - \overbrace{\min(2\sum_{j=1}^{n_{cis}}(1 - \underbrace{\frac{\frac{\overbrace{\sum_{k=1}^{l_{v_1}} 10^{-q_k/10}}^{\text{Corrected } v_1}}{l_{v_1}} + \frac{\overbrace{\sum_{k=1}^{l_{v_2}} 10^{-q_k/10}}^{\text{Corrected } v_2}}{l_{v_2}}}{2}}_{\text{Average Inverse Phred}}), n)}^{\text{Cis Subscore}} \right|}{n+2} \tag{1}$$

where $n$ is the number of reads overlapping two variant positions, $n_{trans}$ is the number of reads supporting a *trans* configuration where each variant is present in only half of the reads, $n_{cis}$ is the number of reads containing both variants, $q_k$ is the Phred quality score of a read at a particular position $k$, and $l$ represents the length of the variant allele being examined in this read (either $v_1$ or $v_2$). The confidence score was designed to summarize the strength of the evidence behind a read-based phasing call. For a thorough explanation of how the confidence score is calculated, see Section 1.1 of S1 Appendix. In order to differentiate high quality phasing calls from low quality phasing calls, we derived 0.34 as a threshold for the confidence score as explained in Section 1.2 of S1 Appendix.

Variants are directly phased with their immediate neighbor using the *cis* and *trans* subscores, calculated by summing the inverse Phred score corrected evidence counts. This basic strategy permits the creation of seed haplotypes that can be locally extended to neighboring variants if overlapping reads exist. SmartPhase can elongate these haplotypes by applying this strategy within paired-end reads as the two paired-end reads must come from the same haplotype, regardless of physical distance or disjointedness. SmartPhase also leverages RNA sequencing reads that connect distant variants due to the read spanning exon-exon boundaries. To calculate the confidence of non-directly phased variants, the confidence scores of all directly phased variants on the shortest path are multiplied together. This helps to represent the growing uncertainty of phase calls as haplotype blocks increase in size and the distance between variants increases.

## Phasing intervals

If both reads as well as parental variants are provided, the local haplotype blocks created by read-based phasing are combined using variants that were pedigree phased. Any contradictions between pedigree phasing and read-based phasing are resolved according to their confidence scores. All variant pairs not phased by direct evidences again have their confidence scores calculated by taking the product of all directly phased linking variants on the shortest possible path between the variants in the pair.

## Innocuous labeling

If parental genotype information is given, certain variant pair constellations can be designated as *innocuous* based on the assumption that the parents of the patient are healthy. *Innocuous* variants are those variants that are deemed to be clinically irrelevant as all variant combinations they partake in have been deduced to be non-disease-causing in Section 1.3 of S1 Appendix. Variant pairs are labeled as *innocuous* if one of the variants is homozygous in a parent or if mother, father, and child all possess the same heterozygous genotype for one of the variants in the pair.

## GATK physical phasing

If a variant is not visible in the alignment of the reads as given by the provided mapping file, this variant was most likely called as a result of read realignment done by the used variant calling program. As a consequence, these variants are designated as not found by SmartPhase. As the HaplotypeCaller (HC) tool of the Genome Analysis Toolkit (GATK) [11] is currently one of the most widely-used variant calling algorithms, we implemented the ability to incorporate phasing information returned by HC when no variant evidences were found within the reads and the variant could not be phased by trio phasing. If HC calls variants through read rearrangement, these variants are usually physically phased at the same time and the used local haplotype information is provided in the resulting variant file. The phase of otherwise missing variants is adopted from the variant files and given a confidence score of 1.0.

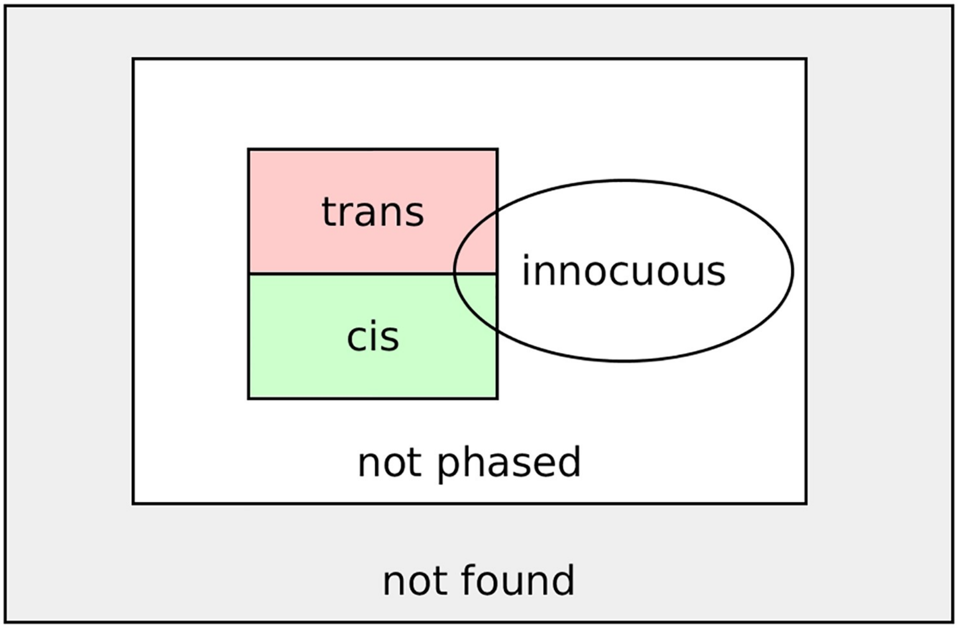

**Fig 1. Visualization of the bit flag system.** If a variant pair could be phased, it is either labeled as *cis* or *trans*. Additionally, it can be labeled as *innocuous*. If a variant pair could not be phased, there was either too little evidence for calling *cis* or *trans* or one of both variant alleles could not be found in the mapped reads.

https://doi.org/10.1371/journal.pcbi.1007613.g001

## Input & output

SmartPhase resolves haplotypes in genomic intervals of interest. Genomic intervals can either be directly defined by the user, or are generated by creating regions enveloping potential compound heterozygous variant pairs of interest. SmartPhase accepts up to two variant specifying files encompassing all variants and those that have been filtered to be deemed clinically relevant. The all variants file is used to create haplotype blocks and usually corresponds to the result of variant calling. The filtered variants file is optional, but can be used to narrow the scope of the output as only those variants specified in this file are printed in the final result. These variants generally constitute the set of variants that were filtered for clinical relevance according to allele frequency, predicted functional impact and other criteria. As many mapping files containing DNA or RNA sequencing reads as desired can be provided to be used during read-based phasing.

The phases and confidence scores of all heterozygous variant combinations within the given or created intervals are the results of SmartPhase. To fully capture the complexity of phased variant pairs, *innocuous* pairs, and variants that were not found in the mapping data, we developed a bitwise flag system to efficiently store all necessary information in a single number. The classification of variant pairs according to the defined criteria is visualized in Fig 1. Table A of S1 Appendix shows the possible combinations of bits and the corresponding final flag.

## Results

We validated SmartPhase on simulated as well as real clinical whole-exome sequencing (WES) data to show its accuracy and compare its performance to WhatsHap. We benchmarked both tools using the runtime, the number of *innocuous* and phased pairs, and the proportion of incorrect predictions. As variant pairs phased or labeled as *innocuous* are equally informative for diagnostic workflows, we summarize these by the term "*cleared* pairs". We refer to them as *confidently cleared* pairs after removing low quality calls with a confidence score below 0.34 as derived in Section 1.2 of S1 Appendix. In the context of clinical diagnosis, we are especially interested in variant pairs that can be excluded as being non-disease-causing. Variant pairs can be designated as being clinically non-relevant by *innocuous* labeling or by being on the same allele through confident *cis* calls which we sum up as *non-pathogenic* pairs.

## Comparison of SmartPhase to WhatsHap on simulated data

As described in Section 2 of S1 Appendix, we simulated WES data of the widely used CEU and YRI trio and phased their heterozygous variants in genes on chromosome 1 and 19 using SmartPhase and WhatsHap. We generated a set of 26, 638 potential heterozygous variant pairs distributed over 2, 922 genes with 4.21 heterozygous variants per gene on average (see Table B of S1 Appendix). Table 1 shows an overview of the main results of the benchmark (complete data in S1 Table).

Phasing in *read only* mode results in low amounts of phased pairs for both phasing tools. In comparison to WhatsHap, SmartPhase clears 2.8%–3.6% less pairs because variant alleles are often not found in the mapped reads as reported in the variant file. Variant calling tools, like the GATK HaplotypeCaller, rearrange read alignments internally before calling variants. As SmartPhase does not realign reads in contrast to WhatsHap, it performs worse in areas of uncertain mapping when only read information is provided. The results for phasing based on read and trio information demonstrate the power of SmartPhase, as the combination of *innocuous* labeling and phasing clears all input variant pairs. This corresponds to 10.9%–20.4% more variant pairs cleared in comparison to WhatsHap.

SmartPhase generated error-free predictions in the combined *read & trio* mode and all errors in *read only* mode are labeled as low quality. WhatsHap has an average error-rate of 1.03% (0.21%–1.89%) with a remarkable increase of the error rate in combined *read & trio* mode. Although this number is quite low, it can have detrimental consequences if even only one pair is wrongly predicted as being in *cis* configuration when in reality it is disease-causing. This emphasizes how crucial a confidence score is in generating trustworthy and accurate predictions.

Another advantage of SmartPhase is its runtime which is on average five times faster than WhatsHap, independent of using only read or both read and trio information. While the

**Table 1. Benchmark results of SmartPhase and WhatsHap.**

| Scenario | | | | SmartPhase | | | | | | WhatsHap | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Trio** | **Chr** | **Pairs** | **Mode** | ***Confidently cleared* in %** | ***Innocuous*** | **Phased** | **Err** | **LQ** | **Time in S** | **Phased in %** | **Phased** | **Err** | **Time in S** |
| CEU | 1 | 6, 783 | *read only* | 12.1 | 0 | 953 | 14 | 130 | 49.3 | 14.9 | 1, 010 | 5 | 274.3 |
| CEU | 19 | 4, 531 | *read only* | 17.2 | 0 | 891 | 4 | 110 | 31.0 | 20.8 | 940 | 2 | 157.6 |
| YRI | 1 | 9, 186 | *read only* | 14.3 | 0 | 1, 462 | 4 | 146 | 62.4 | 17.4 | 1, 595 | 7 | 288.0 |
| YRI | 19 | 6, 138 | *read only* | 17.2 | 0 | 1, 198 | 3 | 141 | 37.2 | 20.3 | 1, 248 | 3 | 169.6 |
| CEU | 1 | 6, 783 | *read & trio* | 100.0 | 4, 700 | 2, 083 | 0 | 0 | 51.3 | 79.6 | 5, 399 | 74 | 270.2 |
| CEU | 19 | 4, 531 | *read & trio* | 100.0 | 3, 087 | 1, 444 | 0 | 0 | 31.1 | 85.2 | 3, 860 | 66 | 159.7 |
| YRI | 1 | 9, 186 | *read & trio* | 100.0 | 5, 618 | 3, 568 | 0 | 0 | 63.0 | 88.0 | 8, 085 | 153 | 282.7 |
| YRI | 19 | 6, 138 | *read & trio* | 100.0 | 3, 739 | 2, 399 | 0 | 0 | 38.3 | 89.1 | 5, 470 | 101 | 168.1 |

We compared SmartPhase (SP) to WhatsHap (WH) using the number of phased pairs, the number of incorrect phased pairs (Err), and the runtime on the same processing node, measured in seconds, as benchmark parameters. Variant pairs that were both labeled as *innocuous* and phased were only counted as *innocuous*. Pairs that were phased with a confidence score below 0.34 are counted as low quality (LQ) pairs.

https://doi.org/10.1371/journal.pcbi.1007613.t001

absolute difference is minor for the limited, simulated data, the runtime becomes particularly relevant in a clinical setting where variant pairs from all chromosomes of hundreds of patients must be phased.

## Validation of SmartPhase on clinical WES data

We validated SmartPhase on a cohort of clinical WES data that consists of 121 trio and 800 singleton patients without parental genotype information. As detailed in Section 3 of S1 Appendix, we selected a set of 116, 613 potential compound heterozygous variant pairs after filtering for rare and protein-altering heterozygous autosomal variants. On average, we identified 126.62 ± 161.25 variant pairs per individual.

**Overall performance of SmartPhase.** To evaluate the overall performance of SmartPhase on real data, we applied it to all 116, 613 variant pairs identified in the 921 individuals of the cohort with a runtime of 190 minutes for all patients or 12 seconds per patient. The results with and without physical phasing for both singleton and trio patients are shown in Fig 2 (complete data in S2 Table).

For singleton patients, in median 25.91% of all variant pairs can be cleared by being phased (see Fig 2(a)). Ignoring low quality phase predictions with a confidence score below 0.34

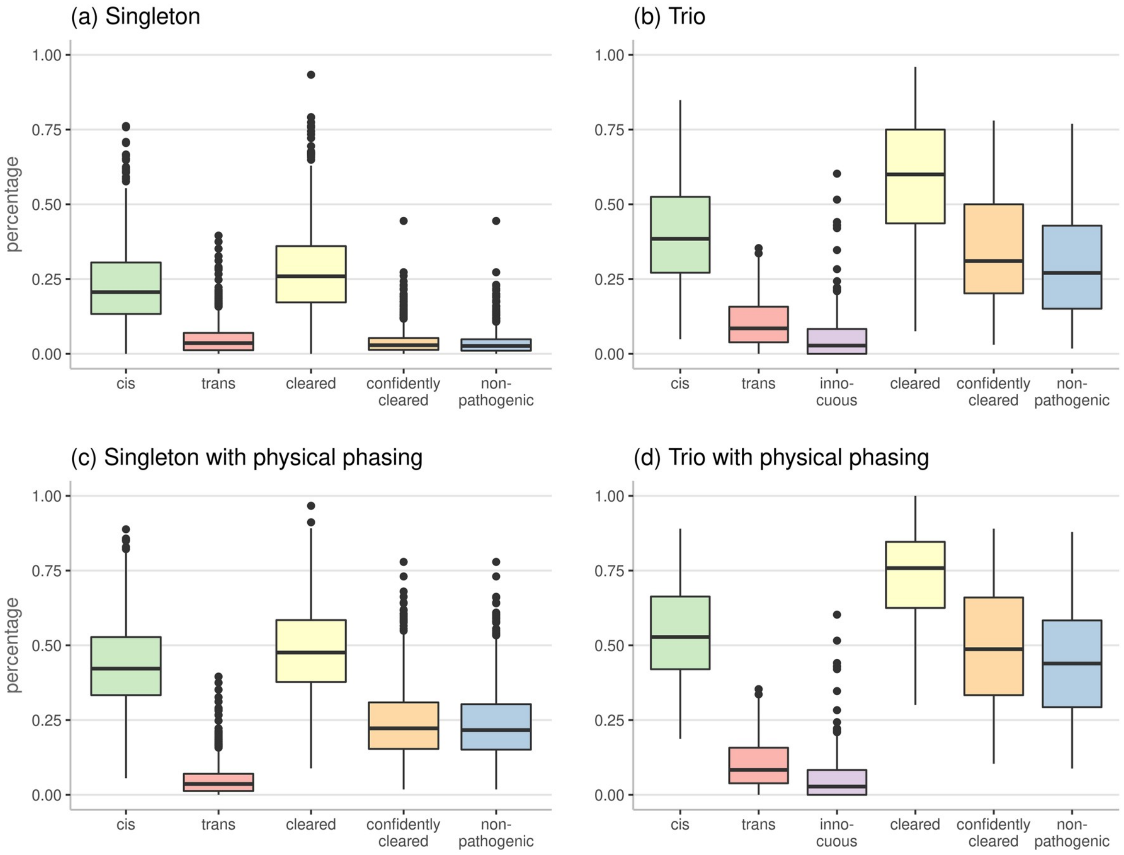

**Fig 2. Boxplots showing the distribution of relative amounts of pairs labeled as *cis*, *trans*, and *innocuous* (only for trio phasing) as well as the percentages of pairs that are *cleared*, *confidently cleared* after removing low quality phasing predictions, and pairs that can be excluded as being *non-pathogenic*.** The plots show results for SmartPhase using only read information for 800 singleton patients (a), using both trio and read phasing for 121 trio patients (b) and the results for the same individuals using physical phasing information provided by the HaplotypeCaller of GATK (c) & (d).

https://doi.org/10.1371/journal.pcbi.1007613.g002 results in 2.88% phased pairs. The median fraction of *non-pathogenic* variants is slightly lower at 2.62%.

For trio patients, in median 60.00% of all input pairs are cleared (see Fig 2(b)). The percentage is lower than in the simulated data where all variant pairs could be cleared. As real WES data is imperfect due to failures in exome capturing or low coverage, some genotypes are missing in one or both of the parents which may make a variant pair impossible to resolve. In our simulated data, the largest fraction of pairs is cleared due to *innocuous* labeling. In our clinical data, it corresponds to the smallest fraction of *cleared* pairs. This is due to the preceding filtering for rare variants which are unlikely to be homozygous in one parent or heterozygous in both parents (see Section 3.2 of S1 Appendix). Removing low quality phase predictions results in 31.03% of the pairs being cleared. In median, 27.06% of all input variant pairs are *non-pathogenic*.

To enhance the power of SmartPhase it is able to incorporate phasing information generated by the HaplotypeCaller of GATK. For both phasing of singleton and trio patients, Fig 2(c) and 2(d) show a noticeable increase in *cis* calls indicating that GATK physical phasing mostly informs about variants being on the same allele, as expected. The percentage of pairs that can be considered clinically irrelevant increases from 2.62% to 21.64% for singletons and from 27.06% to 43.91% for trio patients.

**Comparison of SmartPhase to WhatsHap.** In order to extend the comparison of SmartPhase and WhatsHap to real data, we applied both tools in *read only* and in combined *read & trio* mode to 21, 066 variant pairs in 121 trio patients of the clinical WES cohort. On the reduced data set, WhatsHap required 90 hours to phase the patients making an analysis of the entire cohort of 921 patients not realistic as it would take approximately 585 hours or more than 28 days. Reducing the input files to the regions of interest would shorten the runtime but requires additional preprocessing steps consuming time and storage resources also resulting in an unrealistic scenario. Fig 3 shows the distribution of the percentages of *cleared* pairs for each trio patient for both tools in both modes (complete data in S3 Table).

The performance in *read only* mode is in the same range for SmartPhase and WhatsHap with 50.0% and 62.5% in median. When removing low confidence predictions of SmartPhase with a confidence score below 0.34 the percentage of *cleared* pairs decreases to a median of 19.4% with some outliers still showing a similar performance to WhatsHap.

Examining a subset of 4, 701 variant pairs that can be phased using parental genotype information, SmartPhase generated phase predictions for 1, 577 pairs with 246 pairs being phased confidently. All of these confident phasing predictions are consistent with parental genotypes. For the 1, 331 pairs with a confidence score below 0.34, only 63 predictions are inconsistent with parental genotypes. For 1, 491 pairs, the aligned reads do not contain at least one of both variants. WhatsHap in comparison generated phase predictions for 3, 058 pairs with 34 erroneous calls. Even though WhatsHap phased markedly more variant pairs, it is not possible to filter out incorrect calls. This can have serious clinical effects, especially when considering that 15 of the 34 errors are *cis* calls that are in fact compound heterozygous according to the parental genotypes.

In combined *read & trio* mode SmartPhase increased the percentage of *cleared* variant pairs considerably to a median value of 75.8% before and 48.7% after confidence score filtering. WhatsHap performed remarkably worse when parental genotype data is provided. The phasing rate drops from 62.5% to 25.0% because WhatsHap ignores variants when genotypes are missing in the parents or contradicting Mendelian inheritance is observed.

**Summary.** The validation on simulated and clinical data confirms that SmartPhase is characterized by a fast and highly accurate performance. We demonstrated that using a confidence score threshold of 0.34 generates error-free predictions for the complete simulated data

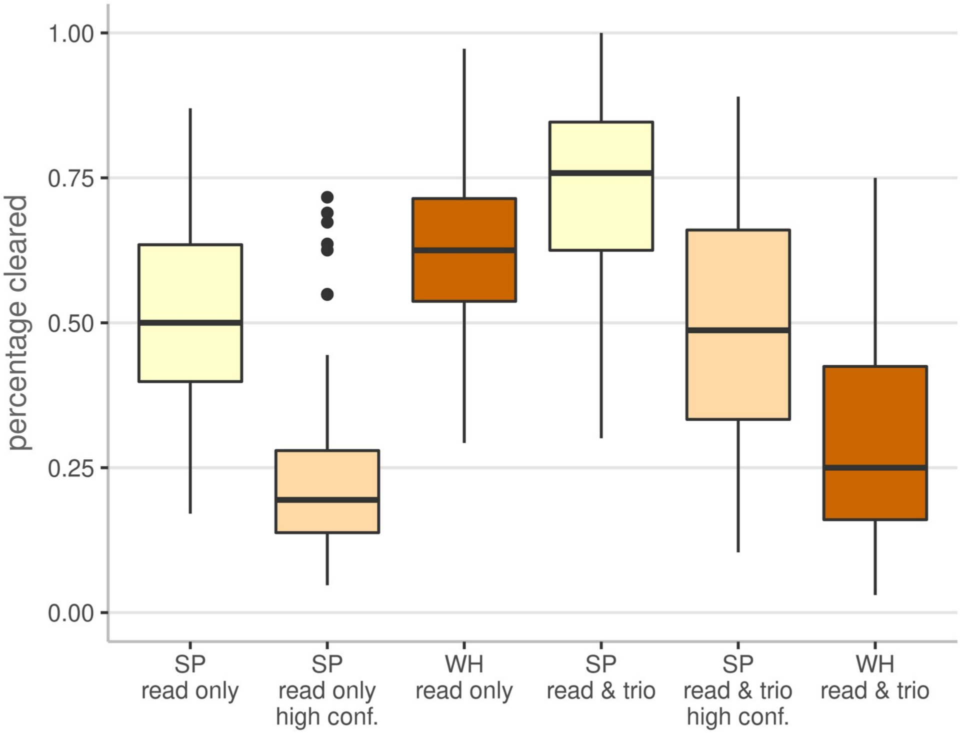

**Fig 3. Boxplots showing the percentage of *cleared* pairs for SmartPhase (SP) and WhatsHap (WH) in *read only* and in combined *read & trio* mode on the 21, 066 variant pairs identified in the 121 trio patients of the clinical WES data cohort.**

https://doi.org/10.1371/journal.pcbi.1007613.g003

set and a sub-set of the clinical data. We showed that SmartPhase outperforms WhatsHap when both read and trio information is provided due to *innocuous* labeling and WhatsHap's strict handling of variants with missing genotypes in the parents or those violating Mendelian rules. If only read-based phasing is possible, SmartPhase clears less pairs than WhatsHap as SmartPhase takes the provided read alignments as is without performing any realignment. However, as SmartPhase compensates by incorporating phasing information generated by GATK HC, it can approximate WhatsHap's read-based phasing performance. As GATK HC is the currently prevailing variant caller, physical phasing information of realigned reads can be exploited in most pipelines. Not performing computationally expensive read realignment allows SmartPhase to be markedly faster which is crucial when considering rapidly growing clinical patient cohorts. As the average read length of sequencing techniques is constantly increasing, read mapping will become more and more accurate without the need for realignment, putting SmartPhase at a clear advantage.

As a stand-alone Java application, SmartPhase can be seamlessly incorporated into any clinical workflow without requiring further installations or downloads. Taken together, SmartPhase greatly simplifies the selection of potential compound heterozygous variant pairs as disease candidates and reduces the search space for pathogenic compound heterozygous variant pairs considerably. The resulting speed up of the analysis of clinical sequence data is helpful for all patients, even if their disease is not caused by a compound heterozygous variant.

## Availability and future directions

The source code of SmartPhase, its documentation, a minimum test data set, and the complete validation pipeline that generates and evaluates simulated data are provided in S1 Code and

S1 Text. Additionally, all files together with the simulated data can be found at http://ibis.helmholtz-muenchen.de/smartphase/. Comprehensive results of SmartPhase using simulated data and for each of the 921 individuals of the cohort are available in S1, S2 and S3 Tables.

Besides the demonstrated use of SmartPhase in detecting and filtering compound heterozygous variants, it can further be used for the analysis of multi-nucleotide variants that have been shown to play an important role in human diseases [12]. As the detection of multi-nucleotide variants requires phasing of nearby variants, SmartPhase is perfectly suited for this clinical task. SmartPhase was designed to be easy to use in any existing clinical sequencing data workflow. To increase the usability even more, we plan to make SmartPhase available as a module for analysis platforms like Galaxy [13] or KNIME4NGS [14]. Beyond that, the phasing efficacy of SmartPhase can be improved by the integration of panel-based phasing methods enabling the connection of haplotype blocks by phasing of common variants.

### Ethics statement

The underlying studies that generated human exome sequencing data that were used for the manuscript were approved by the Ethics Commission of the Medical Faculty of the LMU Munich. The reference numbers of the ethics votes are 346-11, 381-11, 387-11, 438-11, 486-11, 303-12, 187-13 BB, 353-13, 66-14, 501-14 and 806-16. The consent was obtained in written form.

## Supporting information

**S1 Appendix. Additional information on methodology and validation.** The first part provides details on the confidence score formula, the derivation of the confidence score threshold, *innocuous* labeling and the list of all possible bitflags generated by SmartPhase. The second part describes the generation of simulated WES data, the configuration of SmartPhase and WhatsHap, the selection of candidate variant pairs. The third part describes the processing of clinical WES data, the filtering of candidate variant pairs and gives general information on the performed validations.
(PDF)

**S1 Table. Results of SmartPhase and WhatsHap on simulated data.** Results of our benchmark of SmartPhase on simulated data.
(XLSX)

**S2 Table. Results of SmartPhase on clinical WES data.** Results of our application of SmartPhase to the cohort of clinical WES data.
(XLSX)

**S3 Table. Results of SmartPhase and WhatsHap on trio patients.** Results of our comparison of SmartPhase to WhatsHap on trio patients of the clinical WES data cohort.
(XLSX)

**S1 Code. Source code files of SmartPhase.** In addition to the source code, the archive file contains the scripts of the validation pipeline and a minimum test data set.
(ZIP)

**S1 Text. Documentation of SmartPhase.** The documentation gives instructions on how to run SmartPhase and how to interpret its results.
(PDF)

## Author Contributions

**Conceptualization:** Paul Hager, Hans-Werner Mewes, Tim Jeske.

**Data curation:** Meino Rohlfs, Tim Jeske.

**Formal analysis:** Paul Hager, Tim Jeske.

**Funding acquisition:** Christoph Klein.

**Investigation:** Paul Hager, Tim Jeske.

**Methodology:** Paul Hager, Tim Jeske.

**Project administration:** Hans-Werner Mewes, Christoph Klein, Tim Jeske.

**Resources:** Paul Hager, Tim Jeske.

**Software:** Paul Hager, Tim Jeske.

**Supervision:** Hans-Werner Mewes, Christoph Klein, Tim Jeske.

**Validation:** Paul Hager, Tim Jeske.

**Visualization:** Paul Hager, Tim Jeske.

**Writing – original draft:** Paul Hager, Tim Jeske.

**Writing – review & editing:** Paul Hager, Hans-Werner Mewes, Meino Rohlfs, Christoph Klein, Tim Jeske.

## References

1. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Research. 2015; 43(D1):D789–D798. https://doi.org/10.1093/nar/gku1205 PMID: 25428349
2. Wright CF, FitzPatrick DR, Firth HV. Paediatric genomics: diagnosing rare disease in children. Nature Reviews Genetics. 2018; 19(5):253–268. https://doi.org/10.1038/nrg.2017.116 PMID: 29398702
3. Cummings BB, Marshall JL, Tukiainen T, Lek M, Donkervoort S, Foley AR, et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Science Translational Medicine. 2017; 9(386):eaal5209. https://doi.org/10.1126/scitranslmed.aal5209 PMID: 28424332
4. Kremer LS, Bader DM, Mertes C, Kopajtich R, Pichler G, Iuso A, et al. Genetic diagnosis of Mendelian disorders via RNA sequencing. Nature Communications. 2017; 8:15824. https://doi.org/10.1038/ncomms15824 PMID: 28604674
5. Ng SB, Nickerson DA, Bamshad MJ, Shendure J. Massively parallel sequencing and rare disease. Human Molecular Genetics. 2010; 19(R2):R119–R124. https://doi.org/10.1093/hmg/ddq390 PMID: 20846941
6. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine. 2015; 17(5):405–423. https://doi.org/10.1038/gim.2015.30 PMID: 25741868
7. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nature Reviews Genetics. 2013; 14(10):681–691. https://doi.org/10.1038/nrg3555 PMID: 23999272
8. Choi Y, Chan AP, Kirkness E, Telenti A, Schork NJ. Comparison of phasing strategies for whole human genomes. PLOS Genetics. 2018; 14(4):e1007308. https://doi.org/10.1371/journal.pgen.1007308 PMID: 29621242
9. Castel SE, Mohammadi P, Chung WK, Shen Y, Lappalainen T. Rare variant phasing and haplotypic expression from RNA sequencing with phASER. Nature communications. 2016; 7:12817. https://doi.org/10.1038/ncomms12817 PMID: 27605262
10. Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Klau GW, et al. WhatsHap: fast and accurate read-based phasing. bioRxiv. 2016; p. 085050.

11. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, der Auwera GAV, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018; p. 201178.

12. Kaplanis J, Akawi N, Gallone G, McRae JF, Prigmore E, Wright CF, et al. Exome-wide assessment of the functional impact and pathogenicity of multi-nucleotide mutations. bioRxiv. 2018; p. 258723.

13. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research. 2018; 46(W1):W537–W544. https://doi.org/10.1093/nar/gky379 PMID: 29790989

14. Hastreiter M, Jeske T, Hoser J, Kluge M, Ahomaa K, Friedl MS, et al. KNIME4NGS: a comprehensive toolbox for Next Generation Sequencing analysis. Bioinformatics. 2017; p. btx003.