Published in final edited form as:

Methods. 2015 June 1; 0: 41–46. doi:10.1016/j.ymeth.2015.01.016.

On the design and analysis of next-generation sequencing genotyping for a cohort with haplotype-informative reads

Degui Zhi*, Nianjun Liu, and Kui Zhang*

Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL 35294, United States

Abstract

Next-generation sequencing (NGS) technologies, which can provide base-pair resolution genetic information for all types of genetic variations, are increasingly used in genetics research. However, due to the complex nature of NGS technologies and analytics and their relatively high cost, investigators face practical challenges for both design and analysis. These challenges are further complicated by recent methodological developments that make it possible to use haplotype information in sequencing reads. In light of these developments, we conducted comprehensive simulations to evaluate the effects of sequencing coverage, insert size of paired-end reads, and sample size on genotype calling and haplotype phasing in NGS studies. In contrast to previous studies that typically use idealized scenarios to tease out the effects of individual design and analytic decisions, we used a complete analytical pipeline from read mapping and variant detection to genotype calling and haplotype phasing so that we can assess the joint effects of multiple decisions and thus make more realistic recommendations to investigators. Consistent with previous studies, we found that the use of haplotype information in reads can improve the accuracy of genotype calling and haplotype phasing, and we also found that a mixture of short and long insert sizes of paired-end reads may offer even greater accuracy. However, this benefit is only clear in high coverage sequencing where variant detection is close to perfect. Finally, we observed that LD-based refinement methods do not always outperform single site based methods for genotype calling. Therefore, we should choose analytical methods that are appropriate to the sequencing coverage and sample size in order to use haplotype information in sequencing reads.

Keywords

Next-generation sequencing; Human genetics; Resequencing; Study design; Haplotype; Genotype calling; Simulation

1. Introduction

Recent advances in next-generation sequencing (NGS) technologies are transforming many aspects of biomedical research. For the human genetics and genomics community, NGS provides an increasingly affordable and effective way to capture all forms of genetic variants [1–3]. However, due to the complex nature of NGS technologies and their still relatively high cost, investigators face challenges in coming up with NGS study designs that can optimize the use of a tight budget. In addition, design challenges are closely correlated with analysis challenges: when NGS data is generated, what is the best strategy to produce high quality genotypes and haplotypes, which will serve as basis for all downstream analyses?

Current NGS informatics is converging on the procedures to infer genotypes and haplotypes [4,5] (Fig. 1). The starting points of NGS informatics are typically sequencing reads with corresponding quality scores at each site generated from sequencing machines. First, sequencing reads are mapped to the reference genomic sequence, forming a pileup alignment of reads over the genome, typically in the compressed BAM (binary version of the Sequence Alignment/Map, SAM) format. These initial alignments are subject to estimated alignment quality [6] or re-alignment around potential short insertion/deletion sites [7]. Second, geno-type likelihoods, summary statistics for the marginal probability of genotypes at individual sites, are calculated. These likelihoods are then integrated with a certain specification of the prior distribution in a Bayesian framework. Such statistical calculations are generally focused on single sites in either single or multiple samples [7–9]. Based on these calculations, a preliminary set of variant sites is identified and genotype calls at these sites are generated. Third, various additional information, including linkage disequilibrium (LD) information and haplotype information from sequencing reads, can be used to refine those preliminary variant and genotype calls and to phase haplotypes [10–13]. These methods are generally based on multiple sites and multiple samples.

The accuracy of the inferred genotypes and haplotypes for an NGS project depends on many factors of the study design. Most previous work on designing NGS studies had considered how to balance between sample size and sequencing coverage in order to maximize the power for inferences in population genetics and association analysis. A number of studies have shown that analyzing a large number of samples at low coverage is sufficient and yields highly accurate and precise inferences for population genetics [14,15]. Similarly, several investigations have shown that a higher power for sequencing-based association studies can be achieved by sequencing a large number of samples with low coverage [16–20]. In addition, in a less-commonly used pooled sequencing design, the effects of the number and size of pools on the power of sequencing based association studies have been assessed [16,21]. The effects of family samples on rare variants detection have also been studied [22–25].

Recently, several methods have been developed that use the haplotype information within individual sequencing reads to improve genotype calling and phasing from NGS data [10,12,26]. The haplotype information in sequencing reads depends on the length and insert size of those reads. Therefore, it is of great interest to assess the effect of length and insert size of sequencing reads on these newly developed methods. Delaneau et al. conducted a

simulation study to assess the effect of different length and insert size of sequencing reads and sequencing error rate on haplotype phasing based on their own method of SHAPEIT2 [26]. They used SHAPEIT2 for phasing and found that the use of a mixture of insert sizes and longer reads achieved the best accuracy for phasing [26]. However, their simulations focused solely on the performance of SHAPEIT2 for haplotype phasing. Also, their simulations were based on that genotypes were already available, thus ignored the complicated aspects of upstream analyses of variant site detection and genotype calling which may be relevant to phasing accuracy.

In this study, we conducted comprehensive simulations to evaluate the design strategies in light of the new methods that can leverage the haplotype information in reads. We considered the effects of the following factors: sequencing coverage, insert size for paired-end library, combinations of insert size, and analytical methods on genotype calling and haplotype phasing in next-generation sequencing studies. Our simulation started with a human reference sequence and involved (1) a population-genetics based variant implementation, (2) sequencing reads generation with realistic error models, and (3) an examination of the complete bioinformatics pipeline from read mapping, variant detection, preliminary genotype calling to the final stage of genotype refinement and haplotype phasing. Through this simulation we investigated the effects of various design and analytic choices on the precision and sensitivity of variant detection as well as the accuracy for genotype calling and haplotype phasing. Our investigation focused on the settings of genome sequencing of a cohort with 100 individuals. We compared the performance of Samtools [9], which is a single site based, commonly used method for genotype calling from NGS data, Thunder [13], which is a multiple site and multiple sample based, commonly used method for genotype calling and haplotype phasing from NGS data, and HapSeq2 [12], which uses haplotype information in sequencing reads to improve Thunder in terms of accuracy for genotype calling and haplotype phasing.

2. Methods

2.1. Simulation

In our simulation setting, a primary parameter of interest is the insert size of paired-end sequencing reads. The insert size, as a result of DNA library preparation, is the distance between the farther ends of a pair of reads including the lengths of both reads. All reads were of length 100 bp, a commonly used read length from Illumina. We used single reads as baseline. We simulated two sets of paired-end reads with different insert sizes: the short one with an insert size of 250 bp and the long one with an insert size of 1000 bp. We also simulated two additional sets of paired-end reads with mixtures of insert sizes: one with 25%/75% and the other with 50%/50% of 250 bp and 1000 bp inserts.

To evaluate the performance on various typical sequencing depths of coverage, we considered the following settings: extremely low coverage of 0.5 $\times$, as considered in [27], low coverage of 4 $\times$, as used in the 1000 Genomes Project [28], and high coverage of 30 $\times$ [29]. Noticing that intermediate sequencing depths were adopted since the kickoff of the 1000 Genomes Project due to the drop of the price of sequencing, we also simulated sequencing depths of 10 $\times$ and 20 $\times$. For simplicity, we focused on the sample size of $n = 100$,

as this is a typical sample size of whole genome population profiling. The combinations of insert size and coverage are listed in Table 1.

To provide a realistic evaluation, we used a 1 Mb region of real genomic sequence from 40 megabase (Mb) to 41 Mb of chromosome 20 of hg19, as the background reference genomic sequence. We used COSI [30] to simulate mutation patterns of 3000 chromosomes according to the population genetic model similar to the recent European population history. Point substitutions are implanted to the hg19 sequencing at the loci simulated by COSI to generate realistic population genomic sequences. For sample size $n = 100$, we generated 10 independent random subsets with replacement of $2n$ chromosomes and paired them into n diploid genomes. Note that the real reference sequence from chromosome 20 was only used as the ancestral state but all genetic variations were created on top of the reference sequence by simulation methods implemented in COSI. Therefore, each random dataset will have distinct genealogies and LD structures. Using the same set of genomes we generated reads with different sequencing parameters to enhance the comparability across different sequencing parameters. For each of the 10 random samples of genomes, we used ART (Q Version 1.5.0) [31] to generate random reads of specified read length, insert size, and sequencing coverage. Sequencing error rate was set to 0.5% with default indel rates.

2.2. Variant detection, genotype calling, and haplotype phasing

The variant calling pipeline is shown in Fig. 1, which is a simplified version of the best practice pipeline of GATK (<https://www.broadinstitute.org/gatk/about/#typical-workflows>) [4]. The reads were first mapped to the 1 Mb region of the reference genomic sequence using BWA (Version 0.7.5a-r405) [32]. The preliminary variants were then called using Samtools mpileup (Version 0.1.19+) [9] with default parameters. Thunder and HapSeq2 were only used for genotype calling and haplotype phasing on the variants called by Samtools. The preliminary variants set and the corresponding read count and/or genotype, likelihood data were supplied to Thunder [13]. Similarly, the preliminary variants set and the BAM files were supplied to and HapSeq2 [10,12]. Thunder was also implemented in our HapSeq2 program, which can be freely downloaded from <http://www.ssg.uab.edu/hapseq>. The default parameters were used to generate a refined set of genotypes and haplotypes.

2.3. Measures of success for genotype calling and phasing

For variant detection and preliminary genotype calling of Samtools, the primary goal is variant discovery. The success of this step was measured by sensitivity and precision. Sensitivity is defined as the proportion of true variant sites among n samples that are detected. Precision is defined as the proportion of variant sites detected that are true variant sites. We used internal scripts to generate VCF formatted files for the true variant sites derived from the simulated genomic sequences. We used VCFtools (v0.1.9.0) [33] and used the ‘-diff’ subcommand to match the variant sites across VCF files to calculate the sensitivity and precision.

For the refinement step using Thunder and HapSeq2, the primary goals are to derive the most accurate genotypes and haplotypes. We used genotype concordance to measure the accuracy of genotype calling, which is defined as the percent of genotype calls that are

same as the underlying true genotype over the variant sites that were detected by Samtools. We used the number of switch errors to measure the accuracy of haplotype phasing. Both geno-type concordance and switch errors were calculated using VCFtools (v0.1.9.0).

3. Results

3.1. Variant detection

For the NGS bioinformatics pipeline shown in Fig. 1, the primary goal of this middle step is variant discovery. While accurate geno-types are still preferred and will serve as the starting point for the final step, errors in genotype calling in this step can be refined in the final step. However, errors in site detection will be propagated in the final step and may not be corrected.

As shown in Fig. 2, we have the following observations. First, rather unsurprisingly, higher coverage sequencing always results in higher sensitivity and precision, and at sequencing coverage of 30× both the sensitivity and precision approach 100%. Second, it is interesting to see that the sensitivity of Samtools using default setting is determined by the sequencing coverage regardless of paired-end read configuration: with the sensitivity above 40% in 0.5×, 80% in 4×, 92% in 10×, 98.7% in 20×, and about 99.8% in 30×. Third, the precision with paired-end reads is higher than single reads. However, the difference diminishes with the read coverage approaches 20–30×. Given that Samtools is a single site based method, these differences can be attributed to the effect of read alignment: single reads are more difficult to align and thus have the lower alignment quality.

3.2. Genotype calling

The primary goal of genotype calling is to infer the most accurate genotype over the variant sites that are already defined by the variant detection step.

In our simulations, we compared three methods in terms of genotype concordance (Fig. 3). While it is commonly believed that LD-based refinement methods will improve genotype calling [5], it is actually not always supported by our results. We found that genotype concordances of LD-based refinement methods are clearly better than Samtools in only all settings of 4× and 10× coverage and paired-end settings of 0.5× coverage. These are the exact situations where read coverage is relatively low, but discovered sites are with very high precision. In such situations, Samtools will suffer from little available genotype likelihood data direct from the reads, and LD-based refinement methods can help by offering “additional coverage” by borrowing information across adjacent sites.

The LD-refinement methods fail to improve genotype calling for different reasons with the coverage of 0.5× and 20–30×. For 0.5× coverage, low precision of variant detection results in many false-positive variant sites, which will actually block the information flow of the Hidden Markov Model (HMM) in LD-based refinement methods such as Thunder and HapSeq2. Indeed, the genotype concordance with LD-based refinement is 0.741 for Thunder and 0.772 for HapSeq2, respectively. Using the set of true variant sites, geno-type concordance after LD-based refinement is much improved: with 0.849 for Thunder and 0.861 for HapSeq2 (Fig. 3b), respectively. For the 30× coverage, the strong genotype

likelihood of reads is able to identify very rare variants singletons and doubletons, while these mutations may be erroneously overwritten by LD-refinement methods when sampling chromosome segments from others.

We then stratified the genotype concordance comparison for common and rare variants (Fig. 4). Here we used a strict definition of rare variants: a variant was considered as rare if its MAF $\leq 1\%$ and considered as common if its MAF $> 1\%$. For this definition, rare variants only included singletons and doubletons with the sample size of 100 in our simulations. Therefore, we expected that the genotype calling and phasing for rare variants are not benefitted much from HMM sampling procedure in LD-refinement methods. Indeed, Samtools outperforms Thunder and HapSeq2 in all simulation settings in terms of genotype concordance for rare variants while both Thunder and HapSeq2 have higher genotype concordance rates than Samtools for common variants in most of simulation settings.

Finally, it is worth noting that in all simulation settings, HapSeq2 has higher genotype concordance than Thunder, indicating the use of haplotype information of reads does improve the accuracy for genotyping calling for LD-based refinement methods. This benefit even extends to very rare variants.

3.3. Haplotype phasing

The results of our simulations with different coverage and insert size of sequencing reads are shown in Figs. 5 and 6. First, it is apparent that, in all situations simulated, HapSeq2 has smaller number of switch errors than Thunder, indicating the use of haplotype information of reads does improve the accuracy for haplotype phasing. Moreover, this improvement is true for both common and rare variants, with more pronounced improvements for common variants, as shown in Figs. 5 and 6. This is our primary result as it directly demonstrates the benefit of using haplotype information in reads for haplotype phasing. Furthermore, it is clear that the improvement is larger with higher coverage (Fig. 6): the improvement is about 1–1.5-fold, 2-fold, and 2–3-fold in the 0.5 $\times$, 4 $\times$, and 30 $\times$ coverage settings, respectively. Finally, although the magnitude of improvement is mainly due to better phasing of common variants, it is remarkable to see that the phasing of very rare variants is also improved by HapSeq2. The improvement of rare variant phasing is miniscule without paired end reads, indicating that the MCMC procedure in HapSeq2 that uses longer range haplotype information in reads is essential for better phasing of rare variants.

We focused on our comparison of phasing with various insert sizes and the same high coverage of 30 $\times$, in which accuracies for both variant detection and genotype calling are close to 100% (Fig. 2 and Fig. 3) in order to rule out other factors in the comparison. Using the paired *t*-test to compare results over 10 replication data sets, we find that HapSeq2 in the setting 14 with 75% of reads with the insert size of 250 bp and 25% of reads with the insert size of 1000 bp performs better than in the setting 12 with 100% of reads with the insert size of 250 bp (*p*-value = 0.024). While there is no significant difference of HapSeq2 in the setting 14 and the setting 15 with 50% of reads with the insert size of 250 bp and 50% of reads with the insert size of 1000 bp.

Further, a direct interpretation of results in Fig. 5 and Fig. 6 is challenging as many factors that can affect accuracies for variant detection and genotype calling and such accuracies may fluctuate from data sets to data sets. Nonetheless, several patterns are emerged from Fig. 5 and Fig. 6. The higher coverage does not always result in the lower switch errors. However, this is largely the artifact of higher sensitivity in variant detection in high coverage settings. It is thus remarkable to see that HapSeq2 still outperforms Thunder in spite of the imperfect variant detection. Interestingly, Thunder has the smallest switch error when the coverage is 4× while HapSeq2 has the smallest switch error for all depths of coverage from 4× to 30×. Also, the use of paired-end reads improves the performance of HapSeq2 but has little effect on Thunder. For a given coverage, the performance of Thunder for haplotype phasing with different insert size of sequencing reads is similar since Thunder does not use the haplotype information of reads.

4. Discussion

In this work, we conducted comprehensive simulations to evaluate the effects of sequencing coverage and the insert size of one or a combination of paired-end libraries on the accuracy of genotype calling and phasing from NGS data. We evaluated the performances in variant detection, genotype calling, and phasing in the context of a complete NGS bioinformatics pipeline. Our focus was on strategies that can improve LD-based refinement methods for genotype calling and haplotype phasing by adding haplotype information in sequencing reads. Such work has been largely ignored by previous work on designing NGS studies.

In terms of variant detection, we confirmed that higher coverage sequencing always results in the higher sensitivity and precision and the use of paired-end reads improves the precision in variant detection. Using Samtools (mpileup with default setting), the precision of variant detection is very high (>98%) for all coverage settings even for low-coverage of 0.5×. Interestingly, we observed that the sensitivity of Samtools is determined by the sequencing coverage regardless of paired-end read configuration.

Over 98% of variants can be detected with 20–30×. For 4×–10× and for 0.5× paired-end reads, though only 46–92% variants were detected, the quality of the detected variants is sufficient for the downstream LD-based refinement to produce more accurate genotypes.

In terms of genotype calling, we found that HapSeq2 is more accurate than Thunder in all simulation settings, confirming our previous conclusions [12]. The new information we found is that the improvement in genotype calling of HapSeq2 over Thunder is sizeable in low-coverage setting, but saturated at high coverage (20×–30×). Also, the improvement is more pronounced for rare variants (singletons and doubletons in our simulations). All these results confirmed that the haplotype information in sequencing reads is helpful for genotype calling from NGS data. However, we also observed that LD-based methods may not always outperform single site based methods for genotype calling, a deviation from common wisdom. This is largely due to that, for singletons and doubletons, LD-based refinement methods cannot leverage the information of adjacent sites from other samples.

In terms of haplotype phasing, we found that HapSeq2 substantially outperforms Thunder across all simulation settings. For high coverage settings of 30× with paired-end reads, the improvement is as high as threefold for all variants and fivefold for common variants in terms of the length of switch error free fragments. The best improvement was observed with mixed insert sizes for the coverage of 30×. All these results indicate that effective use of haplotype information in reads does indeed make more accurate haplotypes.

For the 30× coverage, our results are grossly consistent with the result of Delaneau et al. [26], which found that the best phasing performance of their SHAPEIT2 program is achieved when using a mixture of varying insert size that is closest to the distribution of distances between heterozygote sites. It is worth noting that the use of a mixture of varying insert sizes that is closest to the distribution of distances between heterozygote sites is often impractical, as the distribution may not be known *a priori* and it is costly to generate a library of arbitrary number of insert sizes. Our conclusion is that generating a mixed library with 75% of reads with a short insert size of 250 bp and 25% of reads with a long insert size of 1000 bp seems is able to offer a desirable design for genotyping a cohort with NGS. Compared to the simulation study of Delaneau et al. [26], which only focused on an idealized case where high coverage data with genotypes that are already available or can be accurately inferred by other methods, our evaluation provides the first evidence that the use of a mixture of varying insert sizes can improve the accuracy for haplotype phasing in realistic settings. In addition, we made a statistical assessment using the paired *t*-test and thus our results are more conclusive.

One of the strengths of our study is that we can offer a realistic evaluation for the types of designs and analyses that investigators would actually face. Unlike previous evaluations, our study is based on a complete pipeline of bioinformatics analysis of NGS data. Our results revealed that erroneous variant detection can create major problems in downstream analysis such as genotype calling and haplotype phasing using LD-based refinement methods. Indeed, 0.5× was proposed as a theoretical recommendation [27] but was rarely used in practice. Our simulation revealed that the challenge for this extreme low coverage might be from the low sensitivity and precision for variant detection. Therefore, investigators should be careful with the variant detection step, and we recommend that certain levels of quality assessment should be done before downstream analyses. On the other hand, one of the limitations of our approach is in that it may be difficult to tease out the effects of choices made in individual analysis steps. This should be the subject of future work.

Another strength of our study is that we stratified the analyses for common and rare variants. We chose a strict definition of rare variants ($MAF \leq 1\%$) so rare variants are singletons or doubletons in our simulations. With our simulation settings, we were able to provide direct evidences that singletons and doubletons are indeed not suited for LD-based refinement methods. We also added to the limited literature on the evaluation of the phasing of rare variants in population sequencing settings and showed phasing of rare variants is extremely challenging. We showed that HapSeq2 is indeed possible to improve the phasing of rare variants, though better methods for which are still highly desired.

Admittedly, one of the limitations of our simulation strategy is that the read mapping results might be a little inflated. We used a small genome region, chr20:40 M–41 M, for our evaluation, which is a practical strategy (e.g., Le and Durbin [34] used the 43 M–48 M region of chr20 for their evaluation). However, not using the whole genome as reference in read mapping may artificially boost the read mapping quality, as it overlooks the facts that reads from this region may be mapped to other regions, and reads from other regions may be mapped to this region. Nonetheless, this 1 M region is actually with a very decent mappability for 100 bp reads with an average of 0.97 and a standard deviation of 0.15. So we expected that the read mapping results using the whole genome as the reference would be similar to what have been presented in our simulations. We also found that the genome-wide average mappability of 100 bp reads is 0.95 with standard deviation 0.17, which is only slightly lower than the mappability of the region used in our simulations. Therefore, we would expect our results are generalizable to most of the genome regions, and recommend caution when dealing with regions with low mappability.

In summary, our results have implications on designing and analyzing NGS studies. In designing NGS studies, we should include paired-end reads, which are not only useful for the read mapping but also for subsequent genotype calling and haplotype phasing. Also, in terms of insert sizes of paired-end reads, it is recommended to devote the majority (e.g., 75%) of sequencing capacity to routine short inserts and reserve some sequencing capacity for longer inserts (e.g., 1000 bp). In analyzing NGS data, we should use methods that can use haplotype information in reads to obtain more accurate genotypes and haplotypes. In addition, we should choose appropriate analytical methods according to the sequencing coverage and sample size. For example, Samtools performs best in terms of genotype calling when the sample size is small and coverage is very low. We then recommend using HapSeq2 in a mode that will not refine genotype calling but only focus on haplotype phasing in such situations.

In our simulation study, we varied insert size of paired-end reads but fixed the length of reads as 100 bp, which is the most commonly used read length from Illumina. With recent advances of “third generation” sequencing technologies, very long sequencing reads but with much higher base error rates, such as those from PacBio [35] and Oxford Nanopore [36] are becoming more available. It is our future work to assess the effects of such long sequencing reads on genotype calling and phasing from NGS data.

Acknowledgments

This work is supported by NIH Grants R01 HG008115 to Kui Zhang and Degui Zhi and R01 GM081488 to Nianjun Liu.

References

1. Bentley DR, et al. Nature. 2008; 456(7218):53–59. [PubMed: 18987734]
2. Metzker ML. Nat. Rev. Genet. 2010; 11(1):31–46. [PubMed: 19997069]
3. Pushkarev D, Neff NF, Quake SR. Nat. Biotechnol. 2009; 27(9):847–850. [PubMed: 19668243]
4. McKenna A, et al. Genome Res. 2010; 20(9):1297–1303. [PubMed: 20644199]
5. Nielsen R, et al. Nat. Rev. Genet. 2011; 12(6):443–451. [PubMed: 21587300]

6. Li H. *Bioinformatics*. 2011; 27(8):1157–1158. [PubMed: 21320865]
7. DePristo MA, et al. *Nat. Genet.* 2011; 43(5):491–498. [PubMed: 21478889]
8. Li H, et al. *Bioinformatics*. 2009; 25(16):2078–2079. [PubMed: 19505943]
9. Li H. *Bioinformatics*. 2011; 27(21):2987–2993. [PubMed: 21903627]
10. Zhi D, et al. *Bioinformatics*. 2012; 28(7):938–946. [PubMed: 22285565]
11. Menelaou A, Marchini J. *Bioinformatics*. 2013; 29(1):84–91. [PubMed: 23093610]
12. Zhang K, Zhi D. *Bioinformatics*. 2013; 29(19):2427–2434. [PubMed: 23943637]
13. Li Y, et al. *Genet. Epidemiol.* 2010; 34(8):816–834. [PubMed: 21058334]
14. Alex, Buerkle. *Mol. Ecol.* 2013; 22(11):3028–3035. [PubMed: 23174005]
15. Fumagalli M. *PLoS One*. 2013; 8(11):e79667. [PubMed: 24260275]
16. Kim SY, et al. *Genet. Epidemiol.* 2010; 34(5):479–491. [PubMed: 20552648]
17. Pasaniuc B, et al. *Nat. Genet.* 2012; 44(6):631–635. [PubMed: 22610117]
18. Li B, Leal SM. *PLoS Genet.* 2009; 5(5):e1000481. [PubMed: 19436704]
19. Sampson J, et al. *Genet. Epidemiol.* 2011
20. Sims D, et al. *Nat. Rev. Genet.* 2014; 15(2):121–132. [PubMed: 24434847]
21. Liang WE, Thomas DC, Conti DV. *Genet. Epidemiol.* 2012
22. Peng G, et al. *Proc. Natl. Acad. Sci. U.S.A.* 2013; 110(10):3985–3990. [PubMed: 23426633]
23. Feng BJ, et al. *PLoS One*. 2011; 6(8)
24. Shi G, Rao DC. *Genet. Epidemiol.* 2011; 35(6):572–579. [PubMed: 21618604]
25. Wijsman EM. *Hum. Genet.* 2012; 131(10):1555–1563. [PubMed: 22714655]
26. Delaneau O, et al. *Am. J. Hum. Genet.* 2013; 93(4):687–696. [PubMed: 24094745]
27. Pasaniuc B, et al. *Nat. Genet.* 2012
28. Abecasis GR, et al. *Nature*. 2012; 491(7422):56–65. [PubMed: 23128226]
29. Meyer M, et al. *Science*. 2012; 338(6104):222–226. [PubMed: 22936568]
30. Schaffner SF, et al. *Genome Res.* 2005; 15(11):1576–1583. [PubMed: 16251467]
31. Huang W, et al. *Bioinformatics*. 2012; 28(4):593–594. [PubMed: 22199392]
32. Li H, Durbin R. *Bioinformatics*. 2009; 25(14):1754–1760. [PubMed: 19451168]
33. Danecek P, et al. *Bioinformatics*. 2011; 27(15):2156–2158. [PubMed: 21653522]
34. Le SQ, Durbin R. *Genome Res.* 2011; 21(6):952–960. [PubMed: 20980557]
35. Carneiro MO, et al. *BMC Genomics*. 2012; 13:375. [PubMed: 22863213]
36. Schneider GF, Dekker C. *Nat. Biotechnol.* 2012; 30(4):326–328. [PubMed: 22491281]

Fig. 1.
Pipeline for variant detection, genotype calling, and phasing.

Fig. 2. Performance of variant detection using Samtools. An average of 4509 true variants were simulated in each data set. The error bars represent the standard deviation among 10 replicates. The *x*-axis represents the index of simulated settings in Table 1.

Fig. 3.

Genotype concordance rates for simulated data sets: (A) all simulated data sets; (B) for 0.5× settings. y-Axis is the genotype concordance between the called genotypes and the true genotypes. The error bars represent the standard deviation among 10 replicates. The x-axis represents the index of simulated settings in Table 1. Genotype concordance rates from LD-based refinement methods using the true variant sets are much more higher than using the low-sensitivity variant sets detected by Samtools.

Fig. 4. Genotype concordance rates for (A) common variants; and (B) rare variants. y-Axis is the genotype concordance between the called genotypes and the true genotypes. The x-axis represents the index of simulated settings in Table 1.

Fig. 5.

The number of switch errors for (A) all variants; (B) common variants; and (C) rare variants. y-Axis is the total number of switch errors. The x-axis represents the index of simulated setting in Table 1.

Fig. 6.

Improvement of haplotype phasing using haplotype information in reads. The y-axis is the mean number of switch errors of Thunder divided by that of HapSeq2. The x-axis represents the index of simulated settings in Table 1.

Table 1

Simulation settings. All reads are with length of 100 bp.

Data set	Read paired?	Insert size if paired	Coverage
1	No	NA	0.5×
2	Yes	250 bp	0.5×
3	Yes	1000 bp	0.5×
4	Yes	250 bp (75%) and 1000 bp (25%)	0.5×
5	Yes	250 bp (50%) and 1000 bp (50%)	0.5×
6	No	NA	4×
7	Yes	250 bp	4×
8	Yes	1000 bp	4×
9	Yes	250 bp (75%) and 1000 bp (25%)	4×
10	Yes	250 bp (50%) and 1000 bp (50%)	4×
11	No	NA	10×
12	Yes	250 bp	10×
13	Yes	1000 bp	10×
14	Yes	250 bp (75%) and 1000 bp (25%)	10×
15	Yes	250 bp (50%) and 1000 bp (50%)	10×
16	No	NA	20×
17	Yes	250 bp	20×
18	Yes	1000 bp	20×
19	Yes	250 bp (75%) and 1000 bp (25%)	20×
20	Yes	250 bp (50%) and 1000 bp (50%)	20×
21	No	NA	30×
22	Yes	250 bp	30×
23	Yes	1000 bp	30×
24	Yes	250 bp (75%) and 1000 bp (25%)	30×
25	Yes	250 bp (50%) and 1000 bp (50%)	30×