Title

InPhaDel: integrative shotgun and proximity-ligation sequencing to phase deletions with single nucleotide polymorphisms.

Journal

Nucleic acids research, 44(12)

ISSN

0305-1048

Authors

Patel, Anand
Edge, Peter
Selvaraj, Siddarth
et al.

Publication Date

2016-07-01

DOI

10.1093/nar/gkw281

Peer reviewed

Published online 21 April 2016

Nucleic Acids Research, 2016, Vol. 44, No. 12 ***e111***
doi: 10.1093/nar/gkw281

InPhaDel: integrative shotgun and proximity-ligation sequencing to phase deletions with single nucleotide polymorphisms

Anand Patel[1,2,*], Peter Edge[2], Siddarth Selvaraj[1], Vikas Bansal[3] and Vineet Bafna[1,2]

[1]Bioinformatics and Systems Biology Program, University of California, San Diego 9500 Gilman Drive, La Jolla, CA 92093, USA, [2]Department of Computer Science and Engineering, University of California, San Diego 9500 Gilman Drive, La Jolla, CA 92093, USA and [3]Department of Pediatrics, School of Medicine, University of California, San Diego 9500 Gilman Drive, La Jolla, CA 92093, USA

Received August 31, 2015; Revised March 13, 2016; Accepted April 6, 2016

ABSTRACT

Phasing of single nucleotide (SNV), and structural variations into chromosome-wide haplotypes in humans has been challenging, and required either trio sequencing or restricting phasing to population-based haplotypes. Selvaraj *et al.* demonstrated single individual SNV phasing is possible with proximity ligated (HiC) sequencing. Here, we demonstrate HiC can phase structural variants into phased scaffolds of SNVs. Since HiC data is noisy, and SV calling is challenging, we applied a range of supervised classification techniques, including Support Vector Machines and Random Forest, to phase deletions. Our approach was demonstrated on deletion calls and phasings on the NA12878 human genome. We used three NA12878 chromosomes and simulated chromosomes to train model parameters. The remaining NA12878 chromosomes withheld from training were used to evaluate phasing accuracy. Random Forest had the highest accuracy and correctly phased 86% of the deletions with allele-specific read evidence. Allele-specific read evidence was found for 76% of the deletions. HiC provides significant read evidence for accurately phasing 33% of the deletions. Also, eight of eight top ranked deletions phased by only HiC were validated using long range polymerase chain reaction and Sanger. Thus, deletions from a single individual can be accurately phased using a combination of shotgun and proximity ligation sequencing. InPhaDel software is available at: http://l337x911.github.io/inphadel/.

INTRODUCTION

Reference genomes are often represented as a haploid set of chromosomes, but humans and many other organisms are diploid. The two homologous chromosomes from a donor may differ from the haploid reference in the form of single nucleotide variations (SNVs), or structural variants (SVs), such as deletions. The variant sites are genotyped as heterozygous (only one chromosome differs from the reference), or homozygous (both chromosomes differ from the reference). Whole genome shotgun sequencing (WGS) accurately genotypes variants, mainly SNVs (1), but also SVs (2).

However, these technologies do not immediately extend to *phasing*, defined by the linking of alleles at heterozygous sites to the same chromosome. Unlinked alleles can indicate numerous possible genome interpretations, as shown in Figure 1A. Phased data is important for numerous biomedical applications. An important application of phasing is finding the causal variants for rare recessive Mendelian disorders. Phasing helps identify compound heterozygotes–mutant alleles that appear on different chromosomes (*in trans*) to knock out both copies of the same gene. In landmark studies, Ng *et al.* (3) and Roach *et al.* (4) found the cause of Miller Syndrome to be compound heterozygous single nucleotide mutations inactivating the *DHODH* gene. In the families analyzed, each parent was heterozygous for a mutation in *DHODH* and no unaffected siblings had compound heterozygous mutations. Additionally, two of the children in the quartet sequenced had compound heterozygous mutations in *DNAH5*, which explained why these children also presented with primary ciliary dyskinesia (4).

Compound heterozygous mutations are not limited to SNVs. Microarray analysis of large cohorts of schizophrenia and autism spectrum disorder (ASD) individuals have shown 17–21 megabase (Mb) sized *de novo* copy number variation (CNVs) to be a strong risk factor (5). In particular, loss of 22q11.2 has been shown to increase risk

*To whom correspondence should be addressed. Tel: +1 858 833 4978; Fax: +1 858 534 7029; Email: adp002@ucsd.edu

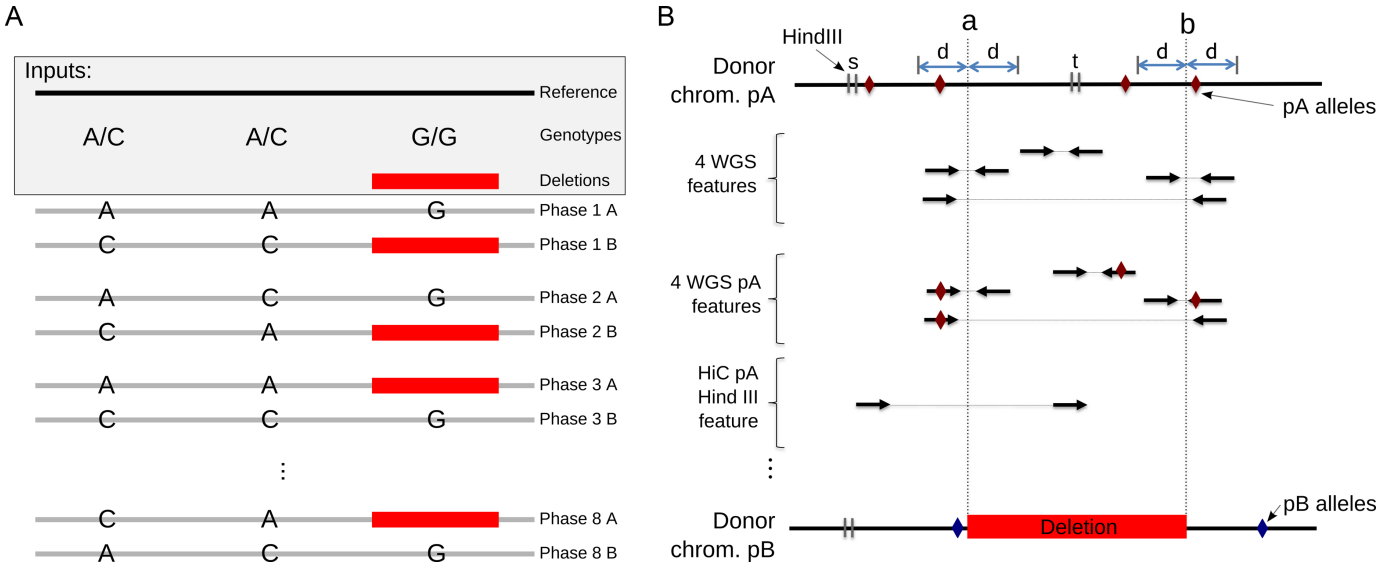

Figure 1. Phasing is necessary to identify heterozygous mutations are in *cis* or *trans*. (**A**) There are eight possible diploid genome interpretations given three genotypes, two heterozygous SNVs and one heterozygous deletion. (**B**) WGS and HiC reads mapping to intervals of length *d* around a putative deleted segment [*a*, *b*] are used as features to phase deletions or determine homozygosity. Examples of paired end (PE) reads fitting WGS, WGS *pA* and HiC HindIII cutsite features are shown, and an exhaustive list of features is given in Supplementary Data 5. A feature consists of the number of PE reads fitting the feature type, which is then normalized by interval length and million reads sequenced (RPKM).

of schizophrenia (6). The hemizygosity due to deletion, combined with other rare mutations within the locus, produces diverse phenotypes, including velo-cardio-facial syndrome and DiGeorge syndrome (6). Similarly, compound heterozygous CNV and rare variants have been implicated in partial loss of function in a number of genes suspected to explain ASD (5).

Hemizygosity can also be attributed to small deletions. For example, bi-allelic mutations in *ABCC6*, a 16.5 kilobase (kb) deletion of exons 23–29 compounded with more common mutations R1141X, R1164X and R1138W, are known to cause psuedoxanthoma elasticum (7). As the single nucleotide mutations overlap the deleted exons and additively depress *ABCC6*, the gene was easily implicated in the disease. The compound heterozygous mutations discovered in disease-related genes typically affect biophysical properties of both gene copies by altering protein function or dosage. In addition to compound heterozygosity, *cis* interactions, including long-range interactions, such as enhancer–promoter interactions can also affect expression of the same genes. In the absence of phasing, these are difficult to identify among the large numbers of mutations observed in any genomic study. Inferring whether the compound mutations act in *cis* or *trans* could greatly narrow down the list of candidate variations, and is only possible with long range phasing.

Variants in an individual can be phased to the two parental chromosomes, by assuming consistent Mendelian transmission and genotyping variants in the individual's parents (8). Similarly, in related individuals, identity-by-descent can be exploited for phasing (9). Furthermore, experimental and computational methods have been developed to phase variants to larger haplotypes in both related and unrelated individuals (10,11). The computational methods have been effective for estimating haplotype phase, when considering sets of common haplotypes (12–14). While the computational developments have focused on phasing single nucleotide polymorphisms (SNPs), the methods are amenable to phasing deletions (15). Also, many common deletions have been shown to have perfect or high linkage disequilibrium with nearby SNPs (16,17). These population-based phasing methods typically cannot span recombination hot-spots, and do not work for rare variants.

Haplotype assembly methods linking single nucleotide variants (SNV) that co-occur on WGS reads have been used to phase rare variants (18,19). While haplotype assembly methods can link many variants, they are unable to achieve chromosome wide haplotypes. Sequenced fragments for short-read technologies, such as Illumina are typically less than 1 kb, and distant variants cannot be linked. Alternate WGS sample preparation steps such as generating mate pair libraries (fragment sizes of 2–40 kb) (20) or isolating single clones (21–23), increase haplotype lengths, but are labor intensive and lower-throughput. Advanced WGS approaches, such as Complete Genomics's Long Fragment Read (LFR) and Illumina's Moleculo technologies, are capable of sequencing longer fragments and have been shown to generate haplotypes with median lengths of several hundred kilobases (24). These recent developments have the potential for producing chromosome wide haplotypes, but further investigation is needed.

Another sequencing approach, proximity ligation sequencing (HiC) involves a modification to sample preparation, and has been shown to successfully capture distant genome interactions. The HiC method has found many applications, beyond the original goal of determining 3D spatial organization of nuclear DNA (25). This includes finding long range interactions of regulatory elements with gene expression (26), and efficiently scaffolding large genomes (27).

In brief, the HiC method captures spatially proximal DNA fragments where paired ends (PEs) reads mapping to distant chromosomal locations link spatially proximal regions of chromosomes. A majority of the ends located <2 Mb apart are drawn from the same chromosome (*cis*) (28). As these PEs often span variants, HiC can link distant variants and generate longer haplotypes. Selvaraj *et al*. (29) coupled HiC to the haplotype assembly method HapCUT (18) to construct chromosome wide haplotypes for both mouse and human genomes.

In this paper, we develop a technique, *Integrative Phasing of Deletions* (INPHADEL), to phase deletions to SNVs using only WGS and HiC sequencing from a single human donor. Phasing SNVs and deletions is similar to the haplotype assembly problem. For example in Figure 1A, the donor has a heterozygous deletion spanning nucleotide G, and a distant heterozygous site A/C. A PE read spanning G and A will place the deletion in the haplotype containing allele C, which would support the Phase 1 in Figure 1A. We directly extend the method of Selvaraj *et al*. (29), by using it to phase SNVs into parental haplotypes A and B. INPHADEL then uses read evidence to phase previously called deletions to a parental haplotype A or B.

MATERIALS AND METHODS

Overview

INPHADEL takes as input, a list of heterozygous SNVs (genome positions), sequence of parent A alleles (pA), parent B alleles (pB) and deletion calls (represented by breakpoints (a, b)). INPHADEL reports a classification for each deletion as pA, pB, *homozygous*, or a fourth class denoted *inconsistent*, and explained below. Our method relies on accurate deletion calls, for which a number of methods have been developed (2,30–33). Since these calls are imperfect, if INPHADEL finds read evidence inconsistent with a heterozygous or homozygous deletion call, we place the deletion into a separate, *inconsistent* class. Finally, some deletions have zero reads supporting either allele—*missing read evidence*. These deletions were excluded from analysis by INPHADEL, since correct predictions in these cases would only arise by chance.

We used a combination of HiC and WGS read data to phase deletions. Because of the complexity of both data types and erroneous calls, applying simple phasing rules would result in low accuracy. For reliable phasing, we need to integrate different signals, including counts of discordant WGS reads and read depth changes in HiC and WGS coverage on one or both parental chromosomes, all of which can suggest or refute a deletion belongs to a particular class. A direct approach to integrate the signals is to frame the problem as a classification task and test a range of supervised learning techniques (see Supplementary Data 1 for diagram of class prediction). We demonstrated deletion phasing using K-Nearest Neighbors (NN), Support Vector Machines (INPHADEL-SVM) and Random Forest (INPHADEL-RF) methods. The learning methods were trained on deletions from NA12878 chromosomes 2, 3 and 4, and a simulated dataset. We then demonstrated performance using deletions on NA12878 chromosomes that were not used for training (see Supplementary Data 1 for diagram of training). Independent from the learning methods, the true deletion phasings for the NA12878 European individual were previously assigned using trio analysis (8).

SNV haplotypes in NA12878

An input to INPHADEL is SNV haplotypes. We used SNV haplotypes from NA12878, which were previously phased by (29) in human assembly hg18. For use in our analysis, we converted the positions to hg19 using liftOver (34). While the haplotypes on each chromosome could be used to phase deletions in NA12878, only haplotypes from chromosomes 2, 3 and 4 were used in the training procedure for learning models.

Simulating WGS and HiC data for a diploid genome

As HiC datasets with known phasing are not readily available, we used simulated data to improve training of INPHADEL. To generate simulated data, we started with human assembly hg19, and constructed a diploid assembly using the NA12878 SNV haplotypes. We used wgsim (35) to simulate WGS reads from the diploid sequences. Additionally, simulated HiC reads were generated by a custom HiC read shuffler (see Supplementary Data 2). HiC experiments create a highly distinct pattern of PE reads compared to WGS (28). In WGS (Illumina format), PE reads are expected to map concordantly with the lower position end mapping to the + strand, and the higher position end mapping to the − strand. In contrast, all four PE read combinations (+/−, −/+, −/−, +/+) are concordant in HiC experiments. Typically, +/− and −/+ read orientations are dominant when the ends map to <25 kb, which likely represent intrachromosomal interactions (28). Our HiC simulator was designed to produce a similar distribution of PE orientations at each distance as the reads generated from a real HiC experiment, and the pattern was confirmed empirically (Figure 2 and Supplementary Data 3). In addition, the simulator accounts for a majority of the PE reads having ends mapping within HindIII cut sites. The cut site bias results in non-uniform read coverage at >25 kb distances in Figure 2. For our HiC simulator, we chose not to simulate distant chromosomal spatial interactions as the effects are only observed at distances >25 kb. At a distance >25 kb on chromosome 20 there were only 2 million reads. This averages to a low 10.3 reads per 100 kb square bin and would not contribute greatly to deletion phasing. Most of the reads lie at distances <40 kb, and data simulated from the HiC shuffler and real HiC experiments appear identical at these distances (see right panel of Figure 2).

Simulating reads from reference chromosomes with deletions. Simulated data was only used to train models. We simulated six pairs of haploid chromosomes 2, 3 and 4, each containing at most 50 deletions. For each chromosome pair we randomly assigned deletions, (a_1, b_1), (a_2, b_2), …, (a_{50}, b_{50}), to classes pA, pB, homozygous or inconsistent. The location a_i of the ith deletion was randomly selected from non-centromere and non-telomere regions and the size, $b_i - a_i$, was randomly drawn according to the deletion length distribution from Mills *et al*. (2). In total, our simulated dataset

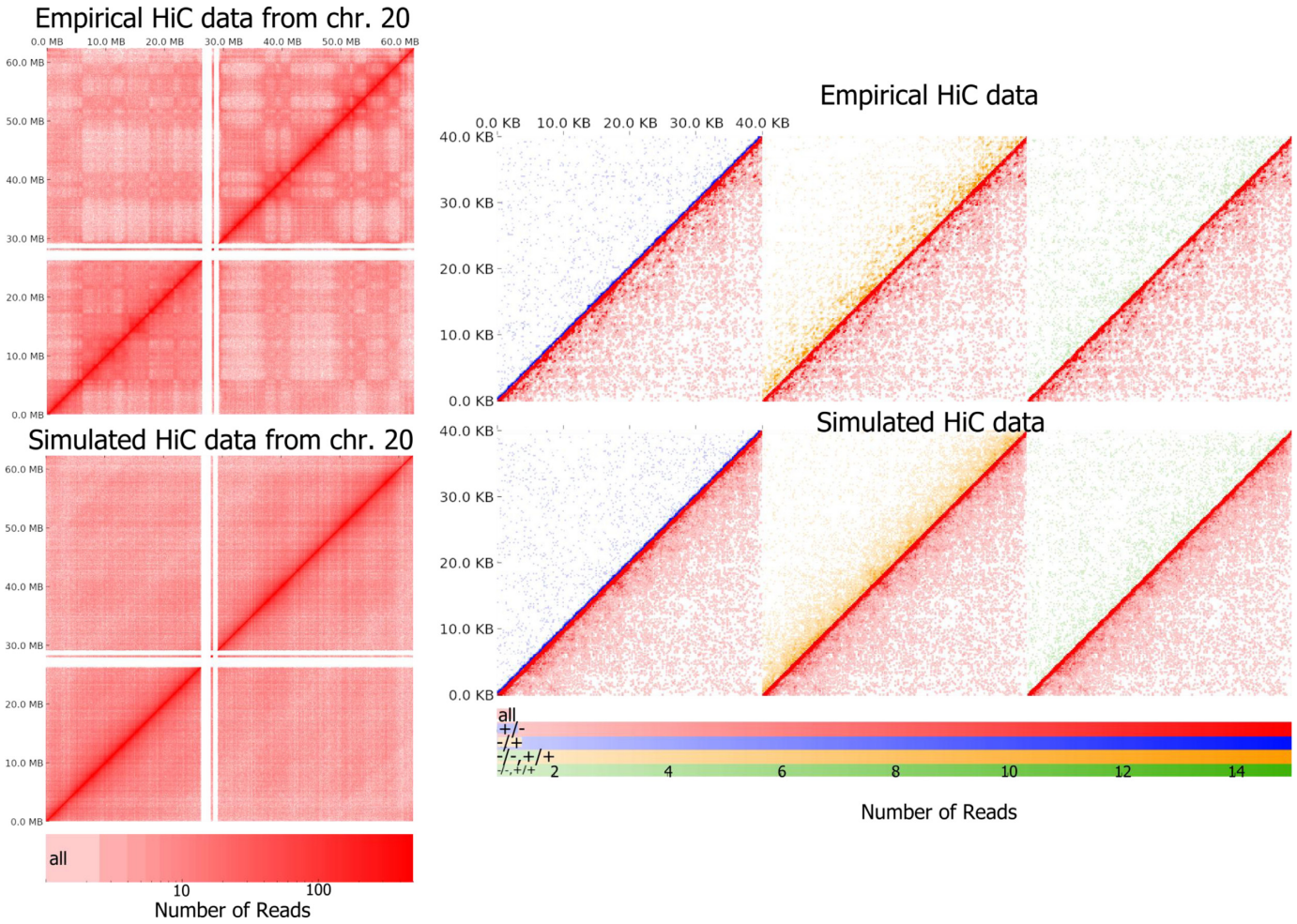

Figure 2. Simulated HiC reads have the same distribution of counts as the real HiC dataset. The visualization shows counts across 62 Mb length chromosome 20 with 100 by 100 kb bins (left panels). The x-axis marks the binned positions of the left-most mapped read ends and the y-axis marks the binned position of the right-most mapped read ends. Top left shows real HiC data from chromosome 20 (29). Bottom left shows similar read counts, but was generated by shuffling the HiC reads mapping to chromosome 20. Within 40 kb windows, the simulation and real data have indistinguishable read count distributions (right panels). The simulation preserves read count distributions of each orientation type (28). The close-up is a read count average over 100 windows with 200 by 200 bp binning. In the right panels, above the diagonal are read counts for PE orientations, +/− (blue), −/+ (yellow) and +/+, −/− (green). Below the diagonal, is the total read count (red). The top right shows real HiC data from chromosome 20 (29). The bottom right shows the same pattern, but was generated by shuffling HiC reads mapping to chromosome 20.

comprised of 256 homozygous, 263 *pA*, 263 *pB* deletions. The simulated dataset also included 268 inconsistent deletion examples, for a total of 1050 simulated deletion phasings.

Depending on the deletion class, the corresponding haploid chromosome pair may have deletions on both copies (homozygous), one copy (heterozygous *pA* or *pB*), or no copies (inconsistent). For each haploid chromosome pair, we simulated 100 bp WGS mapped reads to 81X depth of coverage using wgsim (35) at default settings 500 bp fragment size and 50 standard deviation, and HiC mapped reads to 41X depth of coverage using our HiC read shuffler (see Supplementary Data 2 and Figure 2). The total simulated read counts mapped for the chromosome pairs is listed in Supplementary Data 3. As the simulated chromosome matched the reference except at deletions, each position of the simulated chromosome aligned to a unique position on the reference chromosome. Thus, the reference starting position for each simulated read was known. Simulated reads with a starting reference position p where $a_i - 75 \leq P < a_i$ for some deletion i would form split read mappings and were considered unmapped.

Deletion calls and phasings in NA12878 genome

We analyzed INPHADEL on the well-sequenced European individual, NA12878, which has been previously studied by both HapMap Consortium and 1000 Genomes Project (36). We identified a high confidence set of 421 deletion variants of size >1 kb from a combination of previously made SV calls (2), a split-read alignment method (Bansal *et al.*, unpublished data), and visual inspection of read alignments in Integrative Genome Viewer (IGV) (37) (see Supplementary additional files). The set was identified by removing deletion calls with no supporting evidence in the form of discordantly mapping read pairs or reduced read depth. Additionally, the visual inspection was used to ver-

ify a highly accurate set of deletion calls for training classifiers. Since the parents of NA12878 were also sequenced, the phase of each heterozygous NA12878 deletion was inferred (8) from transmittance of parental deletion genotypes (see Supplementary Data 4). For example, a heterozygous deletion in NA12878 that is only observed in one parent is phased to the same chromosomal haplotype shared between NA12878 and the parent. Furthermore, the phasings were confirmed by manual inspection using Savant Genome Browser (38). Thus, we started with a final high-confidence set of 421 deletion variants.

For our method, we used 2.57 billion PE 101 bp reads from WGS (2,36) and 1.15 billion PE 100 bp reads from HiC (29) generated from NA12878. After mapping (29), the WGS and HiC data amount to 81 and 41× coverage of the genome, respectively. As mentioned earlier, scaffolded phased SNPs from NA12878 were obtained from Selvaraj *et al*. (29).

Training versus test data

Simulated data was only used for training models. Out of the 1050 simulated deletions generated, 171 were missing allele specific read evidence and were excluded from training.

From the NA12878 sample, we used 99 deletions annotated on the chromosomes 2, 3 and 4 for training. These deletions comprised of 25 homozygous, 60 heterozygous and 14 inconsistent. The inconsistent deletions were selected from the pool of Yoruba deletion polymorphisms analyzed by the 1000 Genomes Project (2) where the copy number for each European individual in the NA12878 trio was 2. Among these 99 deletions, there were 22 deletions missing reads supporting either pA or pB, and were excluded from training.

In all, our training dataset contained simulated and NA12878 deletions from chromosomes 2, 3 and 4, which amounted to 261 heterozygous pA, 273 heterozygous pB, 159 homozygous and 263 inconsistent examples. The remaining 336 deletions on NA12878 non-training chromosomes (test chromosomes) were used for a final independent evaluation of the learning methods.

InPhaDel deletion classification

Our approach relied on supervised learning of classes $C = [$ heterozygous pA, heterozygous pB, homozygous, inconsistent $]$ to predict deletion phasings. As with all supervised learning procedures, the key first step is to represent each of the n deletions with 'feature vectors' $\mathbf{f}_1, \mathbf{f_2}, \ldots, \mathbf{f_n}$ and corresponding class labels $l_1, l_2, \ldots, l_n$ where $l_i \in C$. Once a model is learned, the model can then predict the class label for a new object represented by $\mathbf{f}$. For our purpose, the feature vector comprises of read counts from WGS and HiC datasets that distinguish deletions calls belonging to classes in C (see Figure 1B). Deletions with no reads specific to pA or pB were excluded from training and testing.

InPhaDel defining feature vectors

When PE reads from whole genome sequencing are sampled from a donor genome and mapped to a reference, they can reveal clues about heterozygous, homozygous and inconsistent deletions (32,33,39,40). We define PE read mapping as *concordant* if the distance between PEs is consistent with insert size distribution, and *discordant* otherwise. Likewise, the PE reads mapping with +/− orientation are denoted as *normally* oriented, while PE reads mapping with +/+ or −/− orientation are denoted as *identically* oriented.

Consider a heterozygous deletion of an interval (a, b). We expect half of the expected number of concordant mapping reads within the deleted segment (a, b) (41). Correspondingly, a homozygous deletion of an interval (a, b), would have zero concordant mapping reads within then deleted segment. Therefore, we count the number of concordant and normally oriented reads mapping between a and b as a feature (see Figure 1B for illustration of some features and Supplementary Data 5 for a complete list). Second, for some window of size d ($d = 1000$ bp), consider normally oriented but discordant PE reads with the ends mapping to the intervals $[a - d, a)$ and $[b, b + d)$, respectively. These reads are indicative of deletions. Similarly normally oriented PE reads with ends mapping in the intervals $[a - d, a)$, $[a, a + d)$, or the intervals $[b - d, b)$, $[b, b + d)$ also provide clues to the deletion phasing. We use the counts of these four sets of PE reads as four features supporting a deletion call.

Next, we filter for WGS reads supporting the two haplotypes. Reads where one of the PEs maps to an allele in pA are filtered into a WGS pA subset, and reads overlapping pB go into a WGS pB subset. For each of the two subsets, we create four features reporting the following: (i) counts of normally oriented concordant reads mapping to the interval $[a, b)$; (ii) counts of normally oriented, discordant PE reads with ends falling in segments $[a - d, a)$ and $[b, b + d)$ respectively; (iii) counts of normally oriented PE reads with ends in $[a - d, a)$ and $[a, a + d)$; and, (d) counts of normally oriented PE reads with ends in $[b - d, b)$ and $[b, b + d)$, for a total of eight features supporting phasing.

Similar to WGS, we filter for HiC reads overlapping phased variants and assign the filtered HiC reads into pA and pB subsets. For HiC, we count the number of identically oriented PE reads with PEs mapping in $[a, b)$ as a separate feature from the number of opposingly (+/− and −/+) oriented PE reads. Correspondingly, identically and opposingly oriented PE reads with ends mapping to $[a - d, a)$ and $[a, a + d)$ are counted as two separate features, and identically and opposingly oriented PE reads with ends mapping to $[b - d, b)$ and $[b, b + d)$ provide additional two features. Unlike WGS reads, HiC reads mapping in $[a - d, a)$ and $[b, b + d)$ are not necessarily discordant and thus not informative of which allele a deletion appears. These HiC reads create six additional features for each chromosome for a total of 12 features.

Recall that a majority of HiC reads, but not all, map close to HindIII cut sites. To incorporate this additional signal, consider closest cut sites s and t, where $s < a < t < b$. The number of identically oriented HiC pA subset reads mapping $[s - \frac{d}{2}, s + \frac{d}{2})$ and $[t - \frac{d}{2}, t + \frac{d}{2})$ is counted as a single feature. Likewise, we add a feature for HindIII supported reads around breakpoint b. Another two features come from reads in opposite orientation, and finally, a similar set of 4 features is obtained from HindIII supported HiC pB reads, for a total of eight features. In total, there are

20 allele-specific features derived from HiC data. All read counts are normalized to reads per kilobase per million total reads mapped to each chromosome (RPKM).

Training classifiers

We trained *K* Nearest Neighbors, SVM and Random Forest learning techniques on simulated and NA12878 deletion phasings restricted to chromosomes 2, 3 and 4. Nested cross validation was used to select the best performing model, while minimizing generalization error and parameter optimization bias (42) (see Supplementary Data 1 for diagram of parameter optimization procedure). The nested cross validation consisted of an inner loop, which chose the best parameters for learning a model, and an outer loop, which independently assessed the performance of the model–model selection. In the inner loop, the deletions are divided into five approximately equal-sized subsets where one subset is used to test models with different parameters built on the remaining four deletion subsets. The parameters supplying the greatest accuracy across the five sets is then used in the outer loop. In the outer loop, all the training deletions are randomly partitioned into five approximately equally sized subsets. Each partition was used to test the performance of a parameterized model selected by an inner cross validation on the remaining four partitions. This procedure results in five models, with potentially different parameters. For SVM and Nearest Neighbors, the parameters did not differ and accuracies were similar. Random Forest had differing parameters with similar accuracies (maximal difference 5%), and the most general parameters were selected as the best. Lastly, a final model was built using all the training chromosome deletions with the best parameter set (see Supplementary Data 1 for diagram of training).

We repeated the above training procedure for each learning technique. For *K* Nearest Neighbors, the brute-force algorithm was used and parameters $K = 2, 4, 8, 16, 32$ were tested in the inner cross validation. For SVM, a linear kernel SVM (43) is trained for each pair of classes, constituting a total of six models. The final prediction for a deletion is the class that received the most votes out of the six SVM models (44). For the linear kernel SVM, regularization coefficients $C = 1, 10, 100$ were tested in the inner cross validation. For Random Forest (45), forests were trained with number of tree estimators 10, 20, 50 and100, and tree max depth of 2, 5, 10 and 20. The training procedure selects the best performing parameters using the inner cross validation. The scikit-learn python package (46) was used for *K* Nearest Neighbors, SVM and Random Forest training, prediction and measuring accuracy. For Random Forest (45), normalized mean decrease impurity is used to compute the relative feature importance from the ensemble of decision trees as implemented in the scikit-learn package (46).

Performance

Each parameterized model's accuracy is measured by the fraction of correctly classified predictions. For each learning technique, the 5-fold nested cross validation procedure yielded five models with accuracies on five test sets from the training chromosomes, which estimated the training accuracy for the technique. In addition, a final model is generated using all the deletions on training chromosomes, and assessed on deletions from test chromosomes. Thus, we have two measures for performance, (i) average test accuracy estimated from our training procedure, and (ii) test accuracy from test chromosomes that were withheld from any training. In addition to a class prediction, the models also reported the log probability that the deletion belongs to each class. A good classifier should have high accuracy, and assign lower log probability to misclassified predictions.

RESULTS

Phasing accuracy from training procedure and simulations

SNV phasing produces two set of alleles *pA* and *pB*, one for each homologous chromosome. Thus, phasing of deletion calls corresponds to classifying each deletion as *pA*, *pB*, *homozygous* (on both chromosomes) or *inconsistent* (read mappings do not support a deletion call).

We first attempted classification with training restricted to high confidence deletions on chromosomes 2, 3 and 4 of the NA12878 genome. Our training procedure on these deletions, which used nested cross validation (see 'Materials and Methods' section and Supplementary Data 1), reported five best parameterized models with five test accuracies for each classification method: Nearest Neighbors, Support Vector Machines (InPhaDel-SVM) and Random Forest (InPhaDel-RF). When only the NA12878-specific deletions were used for training, each method resulted in cross validation models with different optimal parameters, and highly variable test accuracies (see Figure 3A). The low and variable test accuracies indicated the number of deletions were insufficient for appropriately training models, and assessing the performance of these classification methods. While using the first 10 chromosomes for training resulted in a better training model selection, there remained few test chromosomes to confidently estimate independent accuracy.

To boost the number of deletions used for nested cross validation training without sacrificing NA12878 deletions used for testing, we simulated deletions on the same three chromosomes, and simulated corresponding HiC and WGS data. On the combined set of 956 training deletions and using all 32 features, InPhaDel-SVM and InPhaDel-RF averaged phasing test accuracies of 92.8 and 95.8% respectively (see Figure 3B). These methods outperformed the Nearest Neighbors method, which had an average accuracy of 84.0%.

Each learning method involves selecting parameters for training models. The nested cross validation used in our training procedure, independently optimizes for selecting the best parameters, and evaluates models trained with the best parameters (see 'Materials and Methods' section and Supplementary Data 1). In the learning procedure, five models with optimal parameters are trained on distinct subsets of the training deletions and evaluated on deletions from training chromosomes that were not used for training the specific model. Since the parameters were optimized on a distinct subset of training deletions, it is possible for the models to have different parameters for the same learning

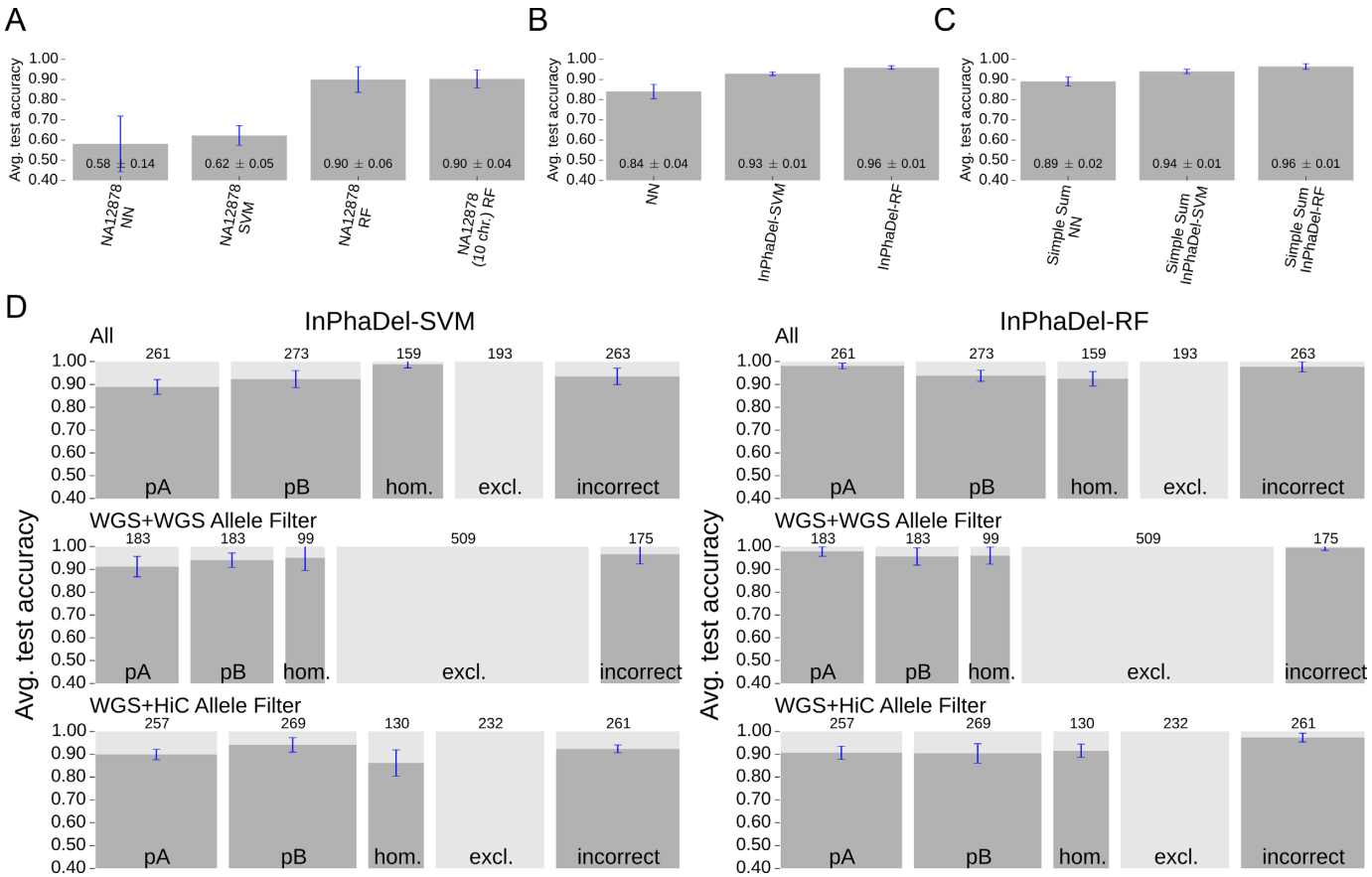

Figure 3. Training on simulated and NA12878 deletions results in accurate and consistent classifiers. Panels (**A**–**C**) show accuracies estimated from training models for nearest neighbors, INPHADEL-SVM and INPHADEL-RF. In the training procedure, accuracy is measured as the fraction of correct predictions made on a test during the outer loop of nested cross validation. The error bars show standard deviation between the outer five tests. Panel A shows accuracies for models trained using only deletions from NA12878 training chromosomes. These models had highly variable and low accuracy. Panel (B) shows models trained using deletions from NA12878 chromosomes 2, 3 and 4, and simulated chromosomes resulted in high accuracy with lower standard deviation than panel (A). Panel (C) shows models trained using deletions from NA12878 training chromosomes and simulated chromosomes, when using only simple sum features (see Supplementary Data 6). Panels (A and B) used 32 distinct features for training, while simple sum features in panel (C) only used 6 to achieve a similar level of accuracy. Panel (D) compares accuracy of INPHADEL-SVM and INPHADEL-RF models trained on WGS allele-specific features, HiC allele specific features, and all features, when accuracy is separated by class. Deletions are classified as *pA*, *pB*, homozygous (hom.), excluded (excl.), or inconsistent. Note, deletions missing reads supporting either allele are excluded from analysis. The width of each column is proportional to the number of examples in each class. While the overall test accuracies are similar, there are fewer deletions with WGS allele-specific support than HiC allele-specific support.

technique. For example, the Nearest Neighbor method predicts phasings using a maximum vote for k nearest neighbors found in the training dataset. In our training procedure, we optimized across 2, 4, 8, 16 and 32 neighbors. For training on deletion phasings from three NA12878 chromosomes and simulations, one model used four neighbors, two models used two neighbors and two models used eight neighbors. Since the accuracy of the models was similar, we selected the largest number of neighbors for final training across all training deletions. For parameter selection of linear Support Vector Machines, we optimized across the regularization coefficients 1, 10 and 100. Larger coefficients correspond to defining larger margins of separation between deletion phasings, by permitting for more incorrect deletion phasings in training. In our training procedure, a coefficient of 100 was found to be optimal for all five models, and was used for the final training. Lastly, Random Forest method uses training deletions to create a forest of decision trees, where the maximum vote for a phasing across decision trees is the final predicted phasing. For Random Forest, we optimized across the number of decision trees, 10, 20, 50 and 100, and the maximum depth of each decision tree, 2, 5, 10 and 20 (16 parameter sets total). Only two of the five models had the same parameters; maximum tree depth of 10, and number of decision trees 50. The other optimized parameters ranged between maximum depth of 10–20 and 20–100 decision trees. Although, the parameters were different, the evaluated accuracy of on training deletions were within 1.1% of the mean accuracy across the five models. Thus, we used a maximum tree depth of 10, and 50 decision trees for the final training, however any of these other optimal parameters could be used to achieve similar results.

In the above classification, each deletion was represented by 32 features drawn from the HiC and WGS data. Alternatively, these features can be summarized into six read count features distinguishing the possible phasings (see Supple-

mentary Data 6). We found each method yields similar performance when training used the summarized six features. Nearest Neighbors, INPHADEL-SVM and INPHADEL-RF had average test accuracies of 89.0, 93.9 and 96.3%, respectively (see Figure 3C). While the summary six features resulted in good test accuracy, analyzing other feature subsets informed the importance of specific data sources.

For example, WGS is insufficient for phasing distant single nucleotide variants (29), and we expected allele-specific features from WGS reads to be insufficient for accurate deletion phasing. To test the power of HiC, we repeated the training procedure using two feature subsets: WGS unfiltered plus WGS *pA* and *pB* filtered data, and unfiltered WGS plus HiC *pA* and *pB* filtered data. Respectively, the subsets had 12 and 24 features (see Supplementary Data 5). INPHADEL-SVM and INPHADEL-RF trained on allele-specific feature subsets had high phasing accuracy (see Figure 3D). INPHADEL-RF trained on only WGS data had an accuracy of 97.3 ± 1.1%, which was higher than INPHADEL-SVM's 94.1 ± 1.9% accuracy on the same feature subset. This corroborated INPHADEL-RF was a better method for phasing deletions. The most drastic difference between WGS and HiC allele-specific feature subsets is the number of deletions with missing allele-specific reads. The HiC feature subset had 43.3% more deletions with allele supporting reads than the WGS feature subset. While HiC data covers more deletions, the phasing accuracy is worse. INPHADEL-RF trained on only WGS data had 4.7% higher accuracy than INPHADEL-RF trained on WGS unfiltered plus HiC *pA* and *pB* filtered data. Of all the correct predictions for simulated deletions between the INPHADEL-RF HiC and WGS feature subset models, 32.1% were phased only by HiC compared to 3.2% by WGS alone (see Figure 4).

Finally, our method depends on accurate SNP phasing and SV calls. We tested the tolerance of INPHADEL-RF to increasing errors in deletion breakpoints *a*, *b*. Starting with the true *pA* and *pB* deletion call set, we modified *a* in decrements of 100 bp and *b* in increments of 100 bp for each deletion to create erroneous call sets. To compare calls made from different deletion error sets, we compute accuracy as the fraction of correctly phased deletions over all deletions, including deletions missing read evidence. INPHADEL-RF retained greater than 73% accuracy when the errors for *a* and *b* were $<$200 bp (see Supplementary Data 7). Regardless of deletion size, the INPHADEL-RF had the largest loss of accuracy when the error in each end-point exceeded 500 bp. Additionally, there is higher tolerance of breakpoint errors for larger deletions. For example, deletions with 400 bp error and size between 5 and 10 kb are called with 67% accuracy, whereas deletions with size greater than 10 kb are called with a 92% accuracy.

Evaluation of models on NA12878 chromosomes withheld from training

In our training procedure, we restricted training classifiers to only NA12878 high-confidence deletions from chromosomes 2, 3 and 4 that were annotated by Bansal *et al*. (unpublished data), and Mills *et al*. (2) on the 1000 Genomes Project data. The remaining 336 high-confidence deletions on other chromosomes were intentionally left out from our training procedure to independently evaluate the accuracy of the final trained models. Of these test chromosomes deletions, only 256 had allele-specific evidence, and were used to evaluate the methods.

The best performing method on these test chromosomes was INPHADEL-RF (see Figure 5), which achieved 85.9 ± 4.3% accuracy (95% binomial confidence interval estimated using normal approximation). In comparison, the accuracy was 70.7 ± 5.6% and 31.6 ± 5.7% for INPHADEL-SVM and Nearest Neighbors, respectively. In comparison to the simpler Nearest Neighbors classifier, the more complex automated feature learning used by INPHADEL-SVM and INPHADEL-RF was necessary for accurate deletion phasing.

INPHADEL-RF was trained on a mixture of deletions from three NA12878 chromosomes and simulated data. Random Forest when trained using only deletions from three NA12878 chromosomes had an accuracy of 79.3 ± 5.0%. Similarly, Random Forest trained using only simulated deletions had an accuracy of 77.7 ± 4.6% on NA12878 chromosomes. Even though the performance is worse than INPHADEL-RF, the result is surprising since Nearest Neighbors and SVM had much lower accuracies when trained exclusively on simulated data (see Supplementary Data 8).

In general, the actual accuracy, which is measured by evaluation on independent deletions from NA12878 test chromosomes, is lower than the accuracy estimated from the training procedure. For example, INPHADEL-RF had an accuracy of 85.9 ± 4.3% when evaluated on NA12878 test chromosomes, and 95.8 ± 1.1% when evaluated by the training procedure (see Figure 3). Random Forest trained using only deletions from 3 NA12878 chromosomes had an estimated 89.8% accuracy from the training procedure, whereas Random Forest trained on exclusively simulated instances had an estimated 97.7% accuracy. As apparent from the accuracy estimated from training, estimation of accuracy in the training procedure was 10% higher than evaluation on test chromosomes, and simulated instances were easily phased. Likewise, INPHADEL-RF training included numerous simulated deletions, which were likely easier to phase than deletions on NA12878 chromosomes and boosted training procedure accuracy.

We also performed Random Forest training using deletions from the first 10 NA12878 chromosomes, and found similar training accuracy to INPHADEL-RF (see Figure 3A). Since 10 NA12878 chromosomes were used for training, only 109 deletions remained on NA12878 chromosomes used for testing. Random Forest trained on 10 chromosomes had an accuracy of 88.1 ± 6.1% on the training-independent NA12878 test chromosomes (see Supplementary Data 9). While the accuracy is greater than INPHADEL-RF, the 95% confidence interval is much larger due to the smaller set of test deletions. Thus, the simulated deletions used in training of INPHADEL-RF improved the confidence of an accurate and reliable Random Forest model on phasing an independent test set of deletions.

Additionally, for each deletion INPHADEL-RF assigns a probability estimate for each possible phasing. As shown in Figure 5, INPHADEL-RF assigns higher probabilities to

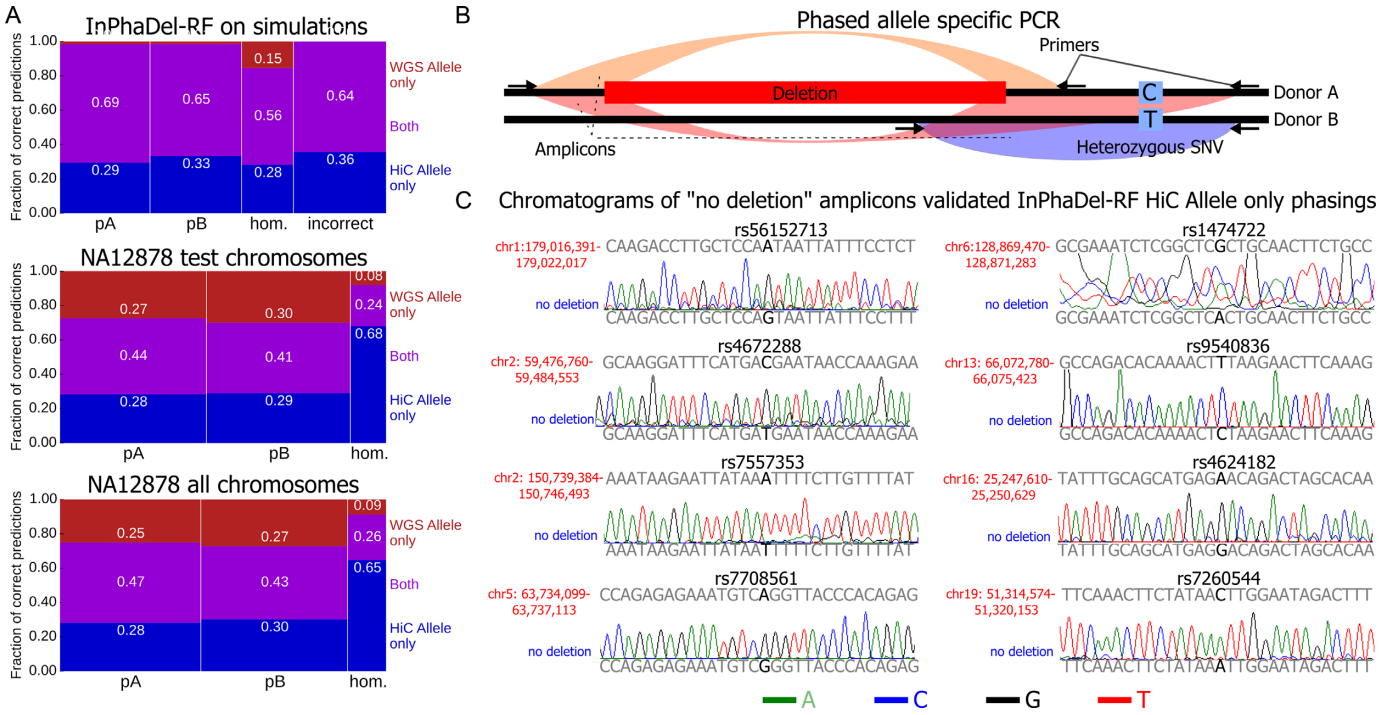

Figure 4. HiC contributed to correctly predicting 33% of the deletion classes in the simulated dataset. (**A**) The contribution of HiC reads remained consistent between deletions in the NA12878 genome and the simulated deletions. WGS reads corrected HiC incorrect phasings for deletions in the NA12878 genome. The plot shows the breakdown of the correct phasings made by only INPHADEL-RF HiC feature subset model (red), WGS feature subset model (blue) or both (magenta). (**B**) For each deletion, four primers (black arrows) were designed to carry out three PCRs. One reaction verifies the presence of a deletion with the expected breakpoints (red box on reference) by amplifying the reference sequence marked in orange. The second PCR encompasses the deletion and corresponding variant (marked in red). A third PCR encompasses the non-deleted sequence and variant (marked in blue). (**C**) Sanger sequencing of amplicons generated from non-deleted sequence corroborated the predicted phase. Chromatograms for the top eight ranked deletions classified using only HiC data are shown.

deletions that were correctly phased, and lower probabilities to incorrectly phased deletions.

Evaluating INPHADEL-RF on independent HiC data

To address possible bias when using a single source of HiC data, we re-evaluated phasing our high-confidence deletion calls on test chromosomes using INPHADEL-RF on a deeper coverage HiC dataset generated by Rao *et al.* (47). Even though there were 84.3% more reads in the Rao *et al.* (47) HiC data compared to (29) (2.12 billion compared to 1.15 billion in Selvaraj), only 7.2% more deletions had allele-specific read evidence. Also, INPHADEL-RF using the Rao *et al.* data had a lower 79.0 ± 4.8% accuracy for phasing deletions on test chromosomes. Specifically, INPHADEL-RF had difficulty in accurately predicting homozygous deletions (see Table 1). The phasing accuracy on only heterozygous deletions was 84.4 ± 4.8%, which is similar to INPHADEL-RF on the lower coverage HiC dataset. Additionally, a similar number of correct phasings were called incorrectly when using the other HiC data. There were 15 deletions that were correctly phased using the low-coverage HiC data that were incorrectly phased by the high-coverage HiC data, with eight correct phasings missing evidence in the high-coverage HiC data. Similarly, there were 11 deletions correctly phased using the high-coverage HiC data that were incorrectly phased by the low coverage HiC data. Also, nine correct phasings in the high-coverage HiC data were not covered by the low-coverage HiC data. In all, the deeper HiC coverage did not correctly phase more deletions, and INPHADEL-RF retains reasonable accuracy when phasing HiC data generated by different laboratories.

Evaluation of INPHADEL-RF on deletion phasings by other technologies

Recently, Pendleton *et al.* (48) utilized Illumina WGS, single-molecule sequencing and single-molecule genome maps to reconstruct the NA12878's diploid genome. The reconstruction included identifying 1323 deletions with size >1 kb. In addition, their reconstruction used trio analysis to infer the homozygous, maternal or paternal phasing for 336 deletions (see Supplementary additional file). While 407 of their deletions overlapped with our 421 high-confidence deletions, only 194 deletions were annotated with parent transmittance. Assuming the scaffold pair (*pA*, *pB*) is assigned either (maternal, paternal) or (paternal, maternal) for each chromosome, 93% of the shared deletions had the same phasing (see Table 2). The few deletions that differed in phasings are likely due to mismatching of parental transmittance to chromosome scaffolds. Overall, the high concordance between the two deletion phasing sets derived using orthogonal sequencing technologies indicates the deletion calls and phasings used for training and testing of INPHADEL-RF are highly accurate.

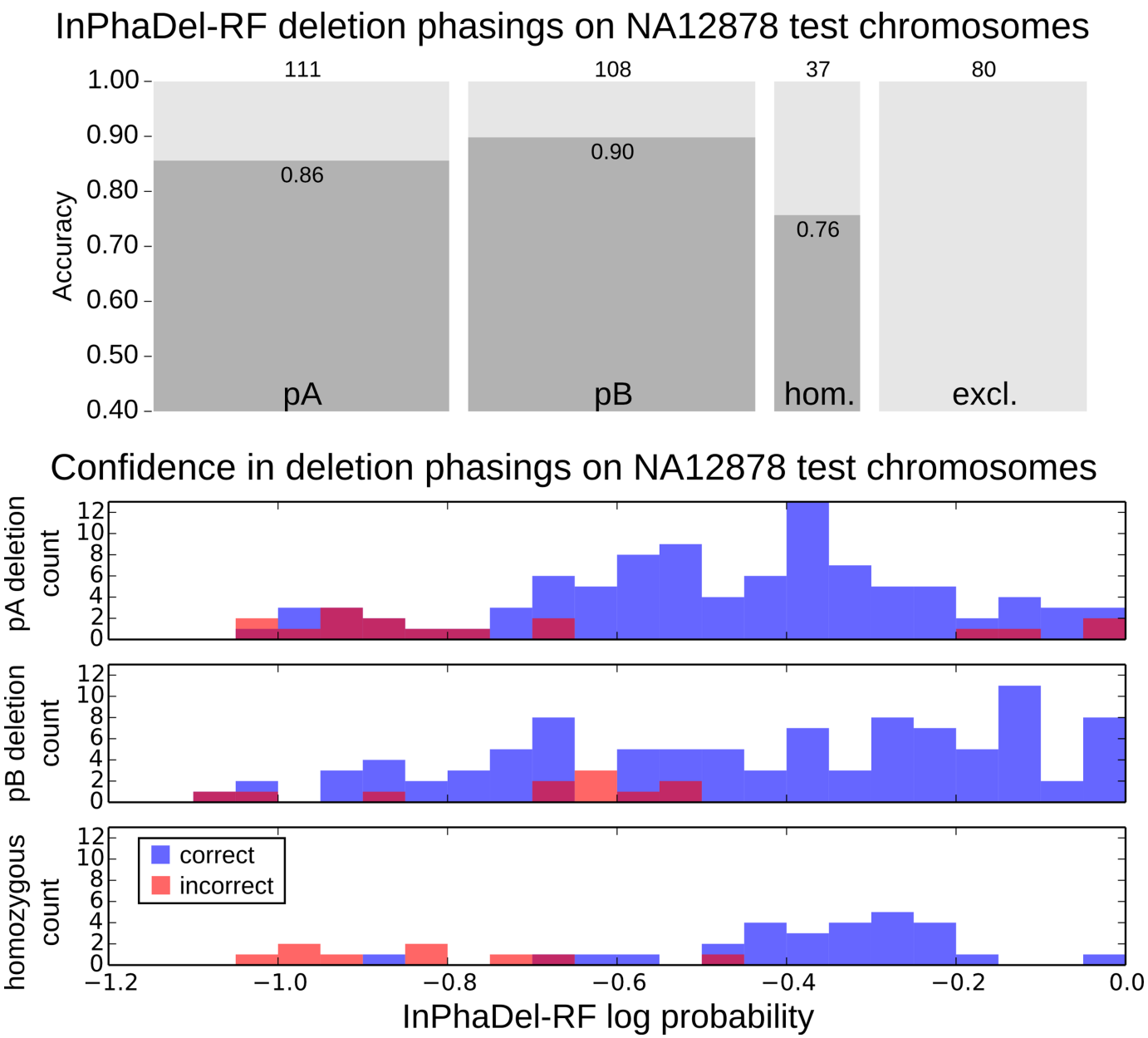

Figure 5. INPHADEL-RF accurately phased deletions on NA12878 test chromosomes. The top panel shows INPHADEL-RF accuracy for each phasing type. The bottom panel shows INPHADEL-RF has weaker confidence in deletions that were incorrectly phased. The blue histogram shows the log probabilities INPHADEL-RF assigned to correctly predicted phasings, while the red histogram shows the log probabilities assigned to incorrectly predicted phasings.

Next, we used INPHADEL-RF to predict phasings for the Pendleton *et al*. (48) deletion calls, in conjunction with our previously compiled SNV haplotypes, WGS and HiC read data. For these new deletion calls and assigned phasings, INPHADEL-RF concurred with 72% of the Pendleton *et al*. (48) deletion phasings with allele-specific read data on non-training chromosomes (see Table 2). Of note, INPHADEL-RF identified 18% of the deletions as inconsistent. The inconsistent deletion labeling arises from complex read mappings that do not fit a specific phasing, and are a limitation of short read sequencing. Single molecule sequencing as used by Pendleton *et al*. (48) generates longer reads, and can more accurately identify deletions that confuse short read sequencing technologies.

Importance of features in INPHADEL-RF

The Random Forest approach produces an ensemble of decision trees, which can be used to report the relative importance of features (see 'Materials and Methods' section). WGS features were found to be slightly more important than HiC features for deletion phasing. The relative importance of WGS features was 0.579 compared to 0.421 for HiC based features (see Supplementary Data 10). Features as used in previous structural variation analysis (32,33,39,40) to call deletions had the highest importance. Since WGS

Table 1. Analysis of INPHADEL-RF on deep coverage HiC data from Rao *et al.* (47)

Rao *et al.* HiC data	pA 123	pB 107	hom. 27	inc. 15	excl. 64
High confidence deletions					
pA	96 (0.35)	8 (0.03)	0 (0.00)	6 (0.02)	7 (0.03)
pB	14 (0.05)	93 (0.34)	1 (0.00)	6 (0.02)	1 (0.00)
hom.	13 (0.05)	6 (0.02)	26 (0.10)	3 (0.01)	56 (0.21)
Rao *et al.* HiC data	pA	pB	hom.	inc.	excl.
Selvaraj *et al.* HiC data					
pA	92 (0.34)	8 (0.03)	0 (0.00)	1 (0.00)	2
pB	15 (0.06)	90 (0.33)	0 (0.00)	2 (0.01)	1
hom.	1 (0.00)	1 (0.00)	24 (0.09)	0 (0.00)	5
inc.	3 (0.01)	1 (0.00)	0 (0.00)	10 (0.04)	0
excl.	12	7	3	2	56

The confusion matrices shows InPhaDel-RF deep coverage HiC predictions on NA12878 test chromosome deletions compared to high confidence deletion phasings, and to InPhaDel-RF low coverage HiC predictions (29). Parenthesis show fraction of deletions divided by total shared deletions, or fraction of deletions divided by total deletions with InPhaDel-RF predictions.

Table 2. Comparison of Pendleton *et al.* (48) deletion phasings from single-molecule technologies to our high confidence deletion phasings and INPHADEL-RF predictions

Pendleton *et al.* phasings	pA 142	pB 122	hom. 101	disjoint
High confidence phasings				
pA	68 (0.35)	5 (0.03)	1 (0.01)	68
pB	7 (0.04)	65 (0.34)	1 (0.01)	73
hom	0 (0.00)	0 (0.00)	47 (0.24)	88
disjoint	67	52	52	
Pendleton *et al.* phasings	pA 121	pB 100	hom. 76	
INPHADEL-RF				
pA	69 (0.36)	5 (0.03)	6 (0.03)	
pB	6 (0.03)	57 (0.30)	2 (0.01)	
hom	0 (0.00)	0 (0.00)	11 (0.06)	
inc	18 (0.09)	15 (0.08)	2 (0.01)	
excl.	28	23	55	

Deletions on all chromosomes were used to compare high confidence deletion phasings with Pendleton et al. deletions, whereas only deletions on NA12878 test chromosomes were used to compare InPhaDel-RF predictions. Also, shown are the counts for Pendleton deletion phasings that weren't in our high confidence deletion (disjoint), and *vice* versa. Parenthesis show fraction of deletions divided by total shared deletions, or fraction of deletions divided by total deletions with InPhaDel-RF predictions.

has more uniform coverage and less noise, the features had higher rank than the more numerous HiC features. Notably, HiC allele-specific features were more important than WGS allele-specific features. Surprisingly, features recruiting only reads from HindIII cutsites had the least importance. We expected the feature to be important since the proximity ligation protocol accumulates HiC reads closest to HindIII cutsites. However, the low importance is likely due to HiC allele-specific read depth features, which include read pairs regardless of ends mapping to cutsites.

Importance of HiC data for phasing deletions

Additionally, we directly analyzed the importance of HiC data for phasing. Similar to the simulation results, HiC read data was responsible for correctly predicting 33.0% of the deletions on NA12878 test chromosomes (see Figure 4). On the NA12878 chromosomes, WGS was also important for correctly predicting 26.4% deletion phasings. The increase in contribution of WGS in NA12878 phasing in comparison to simulated phasing is likely due to low or erroneous mapping of HiC reads.

To confirm the HiC findings, we chose to experimentally validate eight top ranked NA12878 deletions that were correctly phased with adjacent SNVs. Since the deletions were supported by only HiC read data, the distance between the deletion breakpoints and nearest SNV ranged from 800 to 4000 bp. Typical, DNA fragments in WGS studies are $<$800 bp, which explains the lack of WGS read support in predicting the deletion phasing. To validate the long range phased deletions, we used AmBre (49) to design primers for three polymerase chain reaction (PCR) reactions (see Supplementary Data 11 and 12). The first reaction validated the presence of a deletion. The remaining reactions amplified products encompassing the nearest adjacent SNV and either the deletion breakpoints or non-deleted sequence. The PCR products containing a SNV and non-deleted sequence were subsequently Sanger sequenced and the deletion phasing made by the INPHADEL-RF was confirmed in all 8 cases (see Figure 4).

DISCUSSION

Our approach addresses the question of determining if heterozygous deletions act in *cis* or *trans* with other heterozygous variants. We formulate the task as a multiclass classification problem where deletions can be phased to either chromosome (*pA* or *pB*), homozygous or unsupported by the read data. We used Nearest Neighbors, SVM and Random Forest learning methods to solve this problem and found INPHADEL-RF provided superior prediction accuracy. We demonstrated accuracy on the NA12878 European individual, who was part of a trio analyzed by whole genome sequencing (1000 Genomes Project (2,36)). The data on trios provides a truth set of deletions with known phasing or known homozygosity to assess performance. INPHADEL-RF achieved 86% accuracy on NA12878 deletions that were not used in training. In addition, INPHADEL-RF reported lower confidence in phasings that were incorrect.

To robustly train and evaluate INPHADEL-RF required simulation of WGS and HiC reads sets on chromosomes with known deletions. To our knowledge, there is no known HiC read simulator. We developed a method for simulating HiC reads using a read shuffling approach (see Supplementary Data 2). The simulations not only improved robustness for learning models, but also facilitated analyzing the capabilities of HiC and WGS for haplotype reconstruction. Using simulations, we show that INPHADEL-RF tolerates breakpoint errors up to 200 bp with 73% accuracy and then accuracy deteriorates for larger errors depending on deletion size. Most importantly, simulations show that HiC read data contributes to uniquely phase 32.1% of deletions compared to 3.2% that are uniquely phased by WGS data. HiC enables linkages between distant mutations, in our case SNPs and deletions. The same concept had been previously applied to phase distant SNPs (29).

The HiC method is becoming more popular, and researchers are generating WGS and HiC Illumina data for different individuals. While more data would corroborate our results on the importance of HiC data, and breakpoint error tolerance for phasing deletions, it may obviate the need for simulating read data to train models. For example, linear Support Vector Machines performed poorly on NA12878 test chromosomes, even though the accuracy estimated from training was similar to INPHADEL-RF. INPHADEL-SVM likely over optimized on the simulated data used for training, and thus performed poorly on NA12878 deletion phasings. We trained Support Vector Machines on exclusively deletions from the first 10 NA12878 chromosomes, and the accuracy on NA12878 test chromosomes increased from $70.7 \pm 5.6\%$ to $82.6 \pm 7.1\%$. Thus, using more empirical data could benefit training of some models.

The focus of this paper has been on deletions—the primary category of structural variations analyzed by next generation sequencing. The next step is to phase duplications. Our problem formulation, which works well for deletions is not ideal for duplications. Duplications encompass multiple copy number states and may or may not occur in tandem. Thus, there are numerous phasing possibilities for a duplication event. To handle the additional phasings requires more data and introduces new kinds of errors. For example, a region duplicated to a homologous chromosome should be phased to both chromosomes *pA* and *pB*. Further investigation is needed to address phasing of high CNVs, and other computational approaches may be more valuable. In any direction taken, the distant interactions provided by HiC will clearly be informative.

CONCLUSION

To better understand variation in the human genome, computational methods were developed to call deletions (2) and phasing SNPs (18,29) from WGS data. While proximity ligation methods (HiC) were originally used to investigate spatial nuclear DNA organization, they are also a powerful tool for phasing variants. In our approach, HiC data is responsible for phasing 33% of deletions. Phasing deletions with SNPs is important for identifying genetic causes of rare diseases (3,4) and neurological disorders (5,6). CNVs have also been associated with increased risk for schizophrenia and ASD. The observed diverse phenotypes in neurological disorders could be explained by compound heterozygous CNVs and rare variants. Our results show 86% of deletions with allele-specific read data for an individual can be accurately phased to SNP haplotypes using only shotgun and proximity-ligation sequencing from the same individual.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We also like to acknowledge Jesse Dixon's help for extracting genomic DNA from NA12878 derived cell line for use in validation experiments.

FUNDING

NIH [1R01HG007836,1R01GM114362 to A.P., V.B.]; NSF [ABI-1458059, IIS-1318386]; NIH [1R21HG007430 to V.B.]. Funding for open access charge: NIH [1R01HG007836].

Conflict of interest statement. V.B. is a partner in Digital Proteomics, which licenses and sells computational tools for analyzing mass spectrometry data. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. D.P. was not involved in the research presented here.

REFERENCES

1. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
2. Mills,R.E., Walter,K., Stewart,C., Handsaker,R.E., Chen,K., Alkan,C., Abyzov,A., Yoon,S.C., Ye,K. *et al.* (2011) Mapping copy number variation by population-scale genome sequencing. *Nature*, **470**, 59–65.

3. Ng,S.B., Buckingham,K.J., Lee,C., Bigham,A.W., Tabor,H.K., Dent,K.M., Huff,C.D., Shannon,P.T., Jabs,E.W., Nickerson,D.A. *et al.* (2010) Exome sequencing identifies the cause of a mendelian disorder. *Nat. Genet.*, **42**, 30–35.
4. Roach,J.C., Glusman,G., Smit,A.F., Huff,C.D., Hubley,R., Shannon,P.T., Rowen,L., Pant,K.P., Goodman,N., Bamshad,M. *et al.* (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, **328**, 636–639.
5. Yu,T.W., Chahrour,M.H., Coulter,M.E., Jiralerspong,S., Okamura-Ikeda,K., Ataman,B., Schmitz-Abe,K., Harmin,D.A., Adli,M., Malik,A.N. *et al.* (2013) Using whole-exome sequencing to identify inherited causes of autism. *Neuron*, **77**, 259–273.
6. Stone,J.L., O'Donovan,M.C., Gurling,H., Kirov,G.K., Blackwood,D.H., Corvin,A., Craddock,N.J., Gill,M., Hultman,C.M., Lichtenstein,P. *et al.* (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, **455**, 237–241.
7. Ringpfeil,F., McGuigan,K., Fuchsel,L., Kozic,H., Larralde,M., Lebwohl,M. and Uitto,J. (2006) Pseudoxanthoma elasticum is a recessive disease characterized by compound heterozygosity. *J. Invest. Dermatol.*, **126**, 782–786.
8. Conrad,D.F., Andrews,T.D., Carter,N.P., Hurles,M.E. and Pritchard,J.K. (2006) A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.*, **38**, 75–81.
9. Kong,A., Masson,G., Frigge,M.L., Gylfason,A., Zusmanovich,P., Thorleifsson,G., Olason,P.I., Ingason,A., Steinberg,S., Rafnar,T. *et al.* (2008) Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.*, **40**, 1068–1075.
10. Menelaou,A. and Marchini,J. (2013) Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics*, **29**, 84–91.
11. Browning,S.R. and Browning,B.L. (2011) Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.*, **12**, 703–714.
12. Clark,A.G. (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.*, **7**, 111–122.
13. Howie,B.N., Donnelly,P. and Marchini,J. (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.
14. Stephens,M. and Donnelly,P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.*, **73**, 1162–1169.
15. Su,S.Y., Asher,J.E., Jarvelin,M.R., Froguel,P., Blakemore,A.I., Balding,D.J. and Coin,L.J. (2010) Inferring combined CNV/SNP haplotypes from genotype data. *Bioinformatics*, **26**, 1437–1445.
16. McCarroll,S.A., Hadnott,T.N., Perry,G.H., Sabeti,P.C., Zody,M.C., Barrett,J.C., Dallaire,S., Gabriel,S.B., Lee,C., Daly,M.J. *et al.* (2006) Common deletion polymorphisms in the human genome. *Nat. Genet.*, **38**, 86–92.
17. Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T., McVean,G.A. *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
18. Bansal,V. and Bafna,V. (2008) HapCUT: an efficient and accurate algorithm for the haplotype assembly problem. *Bioinformatics*, **24**, i153–159.
19. Halldorsson,B.V., Bafna,V., Lippert,R., Schwartz,R., De La Vega,F.M., Clark,A.G. and Istrail,S. (2004) Optimal haplotype block-free selection of tagging SNPs for genome-wide association studies. *Genome Res.*, **14**, 1633–1640.
20. Levy,S., Sutton,G., Ng,P.C., Feuk,L., Halpern,A.L., Walenz,B.P., Axelrod,N., Huang,J., Kirkness,E.F., Denisov,G. *et al.* (2007) The diploid genome sequence of an individual human. *PLoS Biol.*, **5**, e254.
21. Zhang,K., Zhu,J., Shendure,J., Porreca,G.J., Aach,J.D., Mitra,R.D. and Church,G.M. (2006) Long-range polony haplotyping of individual human chromosome molecules. *Nat. Genet.*, **38**, 382–387.
22. Fan,H.C., Wang,J., Potanina,A. and Quake,S.R. (2011) Whole-genome molecular haplotyping of single cells. *Nat. Biotechnol.*, **29**, 51–57.
23. Lo,C., Liu,R., Lee,J., Robasky,K., Byrne,S., Lucchesi,C., Aach,J., Church,G., Bafna,V. and Zhang,K. (2013) On the design of clone-based haplotyping. *Genome Biol.*, **14**, e100.
24. Peters,B.A., Kermani,B.G., Alferov,O., Agarwal,M.R., McElwain,M.A., Gulbahce,N., Hayden,D.M., Tang,Y.T., Zhang,R.Y., Tearle,R. *et al.* (2015) Detection and phasing of single base de novo mutations in biopsies from human in vitro fertilized embryos by advanced whole-genome sequencing. *Genome Res.*, **25**, 426–434.
25. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
26. Sanyal,A., Lajoie,B.R., Jain,G. and Dekker,J. (2012) The long-range interaction landscape of gene promoters. *Nature*, **489**, 109–113.
27. Burton,J.N., Adey,A., Patwardhan,R.P., Qiu,R., Kitzman,J.O. and Shendure,J. (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.*, **31**, 1119–1125.
28. Jin,F., Li,Y., Dixon,J.R., Selvaraj,S., Ye,Z., Lee,A.Y., Yen,C.A., Schmitt,A.D., Espinoza,C.A. and Ren,B. (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**, 290–294.
29. Selvaraj,S., Dixon,J.R., Bansal,V. and Ren,B. (2013) Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat. Biotechnol.*, **31**, 1111–1118.
30. Rausch,T., Zichner,T., Schlattl,A., Stutz,A.M., Benes,V. and Korbel,J.O. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333–i339.
31. Rimmer,A., Phan,H., Mathieson,I., Iqbal,Z., Twigg,S.R., Wilkie,A.O., McVean,G. and Lunter,G. (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.*, **46**, 912–918.
32. Chen,K., Wallis,J.W., McLellan,M.D., Larson,D.E., Kalicki,J.M., Pohl,C.S., McGrath,S.D., Wendl,M.C., Zhang,Q., Locke,D.P. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*, **6**, 677–681.
33. Ye,K., Schulz,M.H., Long,Q., Apweiler,R. and Ning,Z. (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.
34. Rhead,B., Karolchik,D., Kuhn,R.M., Hinrichs,A.S., Zweig,A.S., Fujita,P.A., Diekhans,M., Smith,K.E., Rosenbloom,K.R., Raney,B.J. *et al.* (2010) The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.
35. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
36. DePristo,M.A., Banks,E., Poplin,R., Garimella,K.V., Maguire,J.R., Hartl,C., Philippakis,A.A., del Angel,G., Rivas,M.A., Hanna,M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
37. Robinson,J.T., Thorvaldsdottir,H., Winckler,W., Guttman,M., Lander,E.S., Getz,G. and Mesirov,J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.
38. Fiume,M., Williams,V., Brook,A. and Brudno,M. (2010) Savant: genome browser for high-throughput sequencing data. *Bioinformatics*, **26**, 1938–1944.
39. Korbel,J., Urban,A., Affourtit,J., Godwin,B., Grubert,F., Simons,J., Kim,P., Palejev,D., Carriero,N., Du,L. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.
40. Zeitouni,B., Boeva,V., Janoueix-Lerosey,I., Loeillet,S., Legoix-Ne,P., Nicolas,A., Delattre,O. and Barillot,E. (2010) SVDetect: a tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics*, **26**, 1895–1896.
41. Medvedev,P., Stanciu,M. and Brudno,M. (2009) Computational methods for discovering structural variation with next-generation sequencing. *Nat. Methods*, **6**(Suppl. 11), 13–20.
42. Cawley,G.C. and Talbot,N.L. (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.*, **11**, 2079–2107.
43. Vapnik,V. (1995) *The Nature of Statistical Learning Theory*, Springer, NY.
44. Knerr,S., Personnaz,L. and Dreyfus,G. (1990) Single-layer learning revisited: a stepwise procedure for building and training a neural

network. In: Soulié,FF and Hérault,J (eds). *Neurocomputing*, Springer, Berlin Heidelberg, pp. 41–50.
45. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
46. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
47. Rao,S.S., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
48. Pendleton,M., Sebra,R., Pang,A.W., Ummat,A., Franzen,O., Rausch,T., Stutz,A.M., Stedman,W., Anantharaman,T., Hastie,A. *et al.* (2015) Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods*, **12**, 780–786.
49. Patel,A., Schwab,R., Liu,Y.T. and Bafna,V. (2014) Amplification and thrifty single-molecule sequencing of recurrent somatic structural variations. *Genome Res.*, **24**, 318–328.