Correlation between frequency of non-allelic homologous recombination and homology properties: evidence from homology-mediated CNV mutations in the human genome

Zhen Peng1,3, Weichen Zhou1, Wenqing Fu4, Renqian Du1, Li Jin1,2 and Feng Zhang1,2,*

1State Key Laboratory of Genetic Engineering and Ministry of Education Key Laboratory of Contemporary Anthropology and 2Collaborative Innovation Center for Genetics and Development, School of Life Sciences, Fudan University, Shanghai 200438, China, 3Department of Biology, Washington University in St. Louis, St. Louis, MO 63130, USA and 4Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA

*To whom correspondence should be addressed at: School of Life Sciences, Fudan University, 2005 Song Hu Road, Shanghai 200438, China. Tel: +86 2151630423; Fax: +86 2151630607; Email: zhangfeng@fudan.edu.cn

Abstract
Non-allelic homologous recombination (NAHR) is one of the key mechanisms of DNA rearrangement. NAHR occurring between direct homologous repeats can generate genomic copy number variation (CNV) and make significant contributions to both genome evolution and human diseases such as cancer. Intriguingly, previous observations on the rare CNVs at certain genomic disorder loci suggested that NAHR frequency could be dependent on homology properties. However, such a correlation remains unclear at the other NAHR-mediated CNV loci, especially the common CNVs in human populations. Different from the rare CNVs associated with genomic disorders, it is challenging to identify de novo NAHR events at common CNV loci. Therefore, our previously proposed statistic M was employed in estimating relative mutation rate for the NAHR-mediated CNVs in human populations. By utilizing generalized regression neural network and principal component analysis in studying 4330 CNVs ascertained in 3 HapMap populations, we identified the CNVs mediated by NAHR between paired segmental duplications (SDs) and further revealed the correlations between SD properties and NAHR probability. SD length and inter-SD distance were shown to make major contributions to the occurrence of NAHR, whereas chromosomal position and sequence similarity of paired SDs are also involved in NAHR. An integrated effect of SD properties on NAHR frequency was revealed for the common CNVs in human populations. These observations can be well explained by ectopic synapsis in NAHR together with our proposed model of chromosomal compression/extension/looping (CCEL) for homology mis-pairing. Our findings showed the important roles of SDs in NAHR and human genomic evolution.

Introduction
Homologous recombination (HR, generally referred to allelic HR) is a crucial process during meiotic division (1). In addition to occurring between allelic sequences, HR can also take place between non-allelic homologies (Fig. 1A and B) (1–3). These non-allelic HR (NAHR) events between direct homologies can
generate copy number variations (CNVs), which are one of the major sources of genetic variations in the human genome (4–8). Notably, gene duplication and subsequent NAHR between duplicated homologies is also a critical mechanism driving genome evolution of non-primate organisms (9,10).

It has been known that the genome characteristics such as GC content (11) and 13-mer sequence motif (e.g. CCNCCNTNNCCNC) (12) can affect the frequency of allelic HR. Different from allelic HR that can occur through the whole chromosome, NAHR is restricted to the loci with non-allelic homologous repeats. Therefore, homology properties can have specific effects on the incidence of NAHR. However, these roles of homology in NAHR remain unclear.

In the human genome, segmental duplications (SDs; >1 kb in length and >90% in sequence similarity) have been known as one of the main genetic substrates for NAHR (13). According to the NAHR mechanism (2) and our previous observations at the pathogenic loci of Y-linked AZFc and Potocki–Lupski syndrome-associated CNVs in 17p11.2 (14–16), we hypothesized that SD properties could potentially have strong influence on mutation probability of NAHR-mediated CNVs, thus making it possible to find the relationship between SD properties and NAHR frequencies. Utilizing multiple mathematic methods [including generalized regression neural network (GRNN) and principal component analysis (PCA)] to study NAHR-mediated CNVs in human populations, here we addressed the following important questions. Do the SD properties (e.g. length) correlate with mutation probability of NAHR-mediated CNVs? If so, what is the specific relationship between NAHR frequency and SD properties? What is the possible mechanism underlying this correlation?

**Results**

The statistic M estimates the relative mutation rate of CNVs in human populations

As for the genomic disorder-associated CNVs that are absent or very rare in human populations, the NAHR frequency at these CNV loci can be directly calculated by identifying de novo mutation events in large patient cohorts (17,18). However, it is challenging to estimate the mutation rates for common NAHR-mediated CNVs, the majority of which can be inherited. Therefore, we employ our previously proposed statistic M to evaluate the ‘relative mutation rate’ of human common CNVs (19). The statistic M utilizes the phylogeny information in human populations and estimates CNV mutation rates from population data. Briefly, for each CNV locus, we tried to construct its evolutionary history by establishing a series of plausible phylogenetic trees through its flanking SNPs. Then, we can trace CNV mutation events along these plausible phylogenetic trees according to the principle of maximum parsimony. At a given evolutionary time, the average number of CNV mutation events (M) normalized by sample size can serve as a proxy for CNV mutation rate. More methodological details were previously described (19). The M data have been obtained for the 4330 CNVs in 3 HapMap populations (19,20).

In this study, we matched the above 4330 HapMap CNVs to the SD pairs in the human genome (21) and identified 62 putative NAHR-mediated CNVs (Supplementary Material, Table S1). Therefore, the statistic M measures the relative mutation rate of CNVs and reflects the NAHR frequency at these NAHR-mediated CNV loci.
SD properties can affect M at human NAHR-mediated CNV loci

Based on the previous observations at certain pathogenic loci in the human genome (14,16,21,22), the following properties of paired SDs were hypothesized to affect NAHR frequency: (a) distance, (b) alignment length, (c) sequence similarity between paired SDs and (d) their chromosomal position (Fig. 1C). Accordingly, we introduced four parameters (D, A, F and R) to represent these SD properties, respectively (see Supplementary Material, Materials and Methods). The potential roles of 13-mer recombination hotspot motif in NAHR have also been studied (Supplementary Material, Table S1) (12).

In this study, we tested multiple mathematical tools to investigate whether SD properties are related to M of NAHR-associated CNVs. These methods included: GRNN, linear regression (Supplementary Material, Note and Table S2), nonlinear regression (Supplementary Material, Note and Table S3) and decision tree (Supplementary Material, Note, Table S4 and Fig. S3). Among them, the GRNN was suggested to be an appropriate method by the parameter of mean squared error (MSE).

Generally, the network gets better approximation when more artificial neurons are available in the hidden layer. However, redundant hidden-layer neurons will result in over-fitting and deteriorate the generalization ability of GRNN (23). Therefore, MSE is usually used to describe the performance of GRNN approximation and generalization. In this study, an MSE of <0.05 indicates good performance. By analyzing how the approximation and generalization of a GRNN change with the spread constant, one can conclude whether the observed correlation between network input and output is reliable. The GRNNs that we employed were three-layer feed-forward artificial neural networks (Supplementary Material, Fig. S1). Three GRNNs were constructed for three HapMap populations (YRI, CEU or CHB+JPT) as training sets, respectively (see Supplementary Material, Note and Table S3) and decision tree (Supplementary Material, Note, Table S4 and Fig. S3). Among them, the GRNN was suggested to be an appropriate method by the parameter of mean squared error (MSE).

We initially proposed Assumption I: there is no significant correlation between SD properties and M in the SD-mediated NAHR events. Given that there is no significant correlation between input and output of the training set (i.e. the training set is highly noisy), the GRNN with only one hidden layer in its feed-forward network needs a large number of hidden-layer neurons to well approximate input–output correlation (23,24). The spread constant of a GRNN is negatively correlated with the number of hidden-layer neurons. For the GRNN with too many hidden-layer neurons, the spread constant should be much smaller than the typical Euclidean distance between input vectors. Therefore, we have Conclusion I: for the training set, if and only if the spread constant of GRNN is far less (<20% typical distance) than the typical Euclidean distance between input vectors, the MSE of training set is small enough (<0.05).

When the MSE of training sets ranges from 0.01 to 0.03, the spread constant is ~0.2 (Supplementary Material, Fig. S2). Considering that the typical Euclidean distance between input vectors is 0.75 for YRI as the training set, 0.79 for CEU and 0.75 for CHB+JPT, we concluded that the spread constant is not far less than the typical distance between input vectors. Therefore, both Conclusion I and Assumption I are false, suggesting that the correlation between SD properties and M does exist.

The correlation between SD properties and M is consistent across human populations

GRNN was also used to test whether the correlation between SD properties and M is robust among different human populations.

We proposed Assumption II: different populations do not have consistent correlations between SD properties and M.

According to Assumption II, the correlation obtained from one population is not applicable to other populations, and we have Conclusion II: when using one population as training set and the other two as test sets, GRNN always performs badly in predicting the M of test sets, i.e. the MSE of test sets would never be smaller than 0.05.

In Supplementary Material, Figure S2, the minimum MSE of a test set is almost always smaller than 0.05 except for one outlier—0.0507 is the minimum MSE for CEU as the test set when CHB+JPT is the training set (Supplementary Material, Fig. S2C). However, even for this outlier, the minimum MSE is still very close to 0.05. Therefore, a GRNN trained by one population is able to effectively predict M in the other two populations. Conclusion II and Assumption II are false, suggesting that different HapMap populations have similar correlations between SD properties and M.

Even when these three HapMap populations share similar correlations, outlier samples that do not fit such a correlation could still exist and should be excluded before concluding the correlation. According to the optimal spread constants of GRNN (I to III) of 0.08, 0.14 and 0.16, respectively (Supplementary Material, Fig. S4), the outliers were identified (see Supplementary Material, Materials and Methods and Table S1).

Relative contributions of SD properties to M: D ≈ A > R > F

We used PCA to investigate the contributing weights of the SD properties (represented by D, A, R and F) in NAHR frequency (represented by M) (see Supplementary Material, Materials and Methods and Table 1). The Bartlett’s test of sphericity (significance = 0.000 < 0.001) indicates that PCA is applicable. The Kaiser–Meyer–Olkin (KMO) test showed that the measure of sampling adequacy (MSA) is 0.650 > 0.5, suggesting that our PCA is applicable to approximately estimate the weights of different parameters (25).

For the unrotated component matrix, only one principal component (PC) is extracted (eigenvalue >1). According to this PC, we conclude that the weights of D and A are similar and larger than those of R. The weight of F is the smallest. However, the extracted PC only explains ~58% of variance. Therefore, we obtained a rotated component matrix to better explain these parameters.

For the rotated component matrix, three PCs are extracted (eigenvalues > 1), which cumulatively explain ~85% of the variance. PC1 has the highest eigenvalue and explains ~34% of variance, which mainly represents the effects of D and A. PC2 and PC3 mainly represent R and F, respectively. Taking these observations together, we conservatively estimate the magnitude of weights as D ≈ A > R > F.

The correlation between SD properties and M

To investigate the correlation between SD properties and M, we utilized GRNN to build a black box model and observed how output variable changed with input variables. We used SD properties as input variables and M as output variable to train GRNN (IV), of which the spread constant is the average of the three optimal spread constants. The optimal spread constants of GRNN (I–III) are 0.08, 0.14 and 0.16, respectively (Supplementary Material, Fig. S4). Therefore, 0.13 is set as the spread constant of GRNN (IV). The training set of GRNN (IV) is the sum of YRI, CEU and CHB+JPT. The MSE of GRNN (IV) is 0.004, suggesting a good approximation.
We defined a new parameter, input variable coordinate (IVC), as the index of combination of input variables (see Supplementary Material, Materials and Methods). A certain IVC value represents a unique combination of D, A, R, and F. While IVC varies from 1 to 10 000 with step size of 1, we observed the outputs of GRNN and thus investigated how M changes with SD properties (Supplementary Material, Fig. S5). The combinations with ‘distance × alignment length’ (see Supplementary Material, Materials and Methods and Fig. S5) were excluded.

To obtain the overall correlation between one input parameter (for instance, D) and M, we calculated the average M of the samples with the same D and plotted the average M versus D. The overall influence of A, R, and F on M was analyzed in the same way (Fig. 2).

M reaches its maximum when D and/or A is of medium value (Fig. 2A and B). R obtains its minimum M when it is of medium value (Fig. 2C). F tends to be positively correlated with M when F is >0.6 (Fig. 2D). The relationship between SD properties and M is non-monotonic.

### Discussion

Based on the previous observations on several pathogenic CNVs associated with genomic disorders, it has been expected that longer homology could provide more substrates for NAHR, i.e. homology length is likely to be positively correlated with NAHR frequency (26). In addition, longer distance between paired homologies was hypothesized to bring less chance for mis-pairing of non-allelic homologies to recombine; therefore, the distance between paired homologies could be negatively correlated with NAHR frequency. Notably, the above-mentioned expectations have been further supported by the recent findings at more genomic disorder loci (16,18). However, the relationship between NAHR and homology properties in human common CNVs (relatively smaller than rare pathogenic ones) has not been addressed yet before our genomic analysis in this study.

Generally, our findings in common CNVs are consistent with the previous observations on pathogenic CNVs. For example, we found that the inter-SD distance is negatively correlated with NAHR frequency when this distance is longer than 50 kb (Fig. 2A). Consistently, the genomic disorder-associated CNVs are much larger than 50 kb and show a similar negative correlation between NAHR frequency and inter-SD distance (18). In addition, the previously observed positive correlation between NAHR frequency and homology sequence similarity at genomic disorder loci (18) is also supported by the human common CNVs, especially when the sequence similarity is >96% (Fig. 2D).

Our observations on NAHR-mediated common CNVs also reveal some new relationships between NAHR frequency and homology properties. For example, when the inter-SD distance is shorter than ~20 kb, it is positively correlated with NAHR frequency (Fig. 2A). Therefore, our observations on NAHR in the loci with common and relatively small CNVs can potentially provide additional information for the correlation between SD properties and NAHR frequency, which could complement with the previous reports using larger pathogenic CNVs (14,16,22).

### Chromosomal compression/extension/looping model: chromosome with elasticity

Here, we propose a qualitative model of ‘chromosomal compression/extension/looping’ (CCEL) to explain the relationship between SD properties and M. This new model was based on the elasticity theory (27,28), in which a chromosome is not a lithe string but like an elastic beam. When a chromosome is compressed or extended or bent, the stress is generated as a barrier against SD mis-pairing in the NAHR events.

NAHR can happen in different ways (Fig. 3). The first is inter-chromatid NAHR by chromosomal compression/extension (inter-C/E), in which one chromatid is compressed and/or the other chromatid is extended to achieve SD mis-pairing between chromatids (Fig. 3B). The second is inter-chromatid NAHR by chromosomal looping (inter-L), in which a loop is formed outside of the region between paired SDs on one chromatid, and non-allelic SDs on different chromatids could mis-pair (Fig. 3C). The third is inter- and/or intra-chromatid NAHR by chromosomal looping (intra/intra-L), in which a loop is formed within the region between paired SDs on one chromatid, so that non-allelic SDs on the same chromatid or from the different chromatids could mis-pair (Fig. 3D).
Based on the CCEL model for NAHR, we expect that there are following physical factors influencing NAHR.

First, the chromosome stress. The chromosome deformation can generate stress because of chromosomal elasticity. Bigger deformation comes with higher stress, which inhibits NAHR. For inter-C/E, the Young’s modulus \( E \) of chromosome strongly influences the stress, so that longer distance between paired SDs comes with higher stress in the chromosome. For inter-L and inter/intra-L, the flexural strength of the chromosome strongly influences stress. Because the support span \( L \) of a beam negatively correlates with the load \( P \) with a given flexural strength \( \sigma \) (\( \sigma = cP/L \), where \( c \) is a constant) (30), longer distance between paired SDs (equivalent to longer support span) comes with lower stress (equivalent to smaller load). Specifically, for inter/intra-L, too large alignment length of paired SDs would dramatically decrease the effective length of the bending region (effective length = distance – alignment length), resulting in high stress in the chromosome. However, this effect may not be crucial when the alignment length is not that large.

Second, the attraction between paired SDs. Generally, high similarity and long alignment length make strong attraction between paired SDs. Strong attraction can facilitate mis-pairing between SDs, thus facilitating NAHR.

Third, the spatial distance between paired SDs. Shorter distances between paired SDs on the same or different chromatid (s) bring about more chances for SDs to mis-pair for NAHR. In addition, as the telomeres are attached to the nuclear envelope during the stage of synapsis in Meiosis I (31) and two chromatids of the same chromosome are linked by the centromere, the centromeres and telomeres are ‘fixed’ ends on the chromosomes. Therefore, the spatial distance between paired SDs on different chromatids is shorter when the SDs are closer to either centromere or telomere.

When \( D \) (representing inter-SD distance; Fig. 2A) is very small, it is possibly getting hard for a chromatid to form a loop because a small loop means large flexural stress. Instead, the chromatid only needs to be slightly compressed or extended to allow SD mis-pairing. Therefore, inter-C/E could be the major way of NAHR when inter-SD distance is extremely short. When \( D \) increases, large compressive/tensile stress can inhibit inter-C/E, whereas it is easier to form a big loop with small flexural stress. Therefore, inter-L and inter/intra-L could become the major ways.
of NAHR. Here, larger D comes with smaller flexural stress, thus resulting in larger M (for example, the positive correlation between D and M when D is <0.7). When D further increases, the effect of spatial distance becomes dominant, and thus larger D results in smaller M.

When A (representing SD alignment length; Fig. 2B) is small, the role of A in enhancing the attraction between paired SDs is dominant. Therefore, larger A achieves stronger attraction between paired SDs, thus resulting in larger M. However, the large A could also have reverse effects. First, given that NAHR happens through inter/intra-L and that D keeps constant, larger A results in shorter effective length of the bending region, so that the flexural stress within the bending region becomes bigger, and thus NAHR is less likely to happen (smaller M). Second, large A means that D cannot be too small, so that A’s effect of increasing the attraction between paired SDs could be attenuated by D’s effect on spatial distance between paired SDs.

As for the locations of SD pairs in a chromosome, small R means that SDs are close to centromere, whereas large R means that they are close to telomere (Fig. 2C). Owing to the connection of sister chromatids at centromere (28) and the dynamic bouquet organization of telomeres during meiotic recombination (1), the spatial distance between paired SDs on different chromatids is short when SDs are close to centromere or telomere, thus facilitating inter-chromatid NAHR and resulting in big M.

It is expected that larger F brings about stronger attraction between paired SDs, so that M is bigger. Generally, this expectation fits Figure 2D, especially when F is >0.6. However, when F is <0.6 (i.e. the sequence similarity is <96%), M is generally small. This result may reflect the requirement of extremely high similarity between repeats for NAHR to occur (4). For example, in the NAHR-mediated CNV loci of genomic disorders, the sequence similarity between homologous is generally >97% (2).

The 13-mer motif density of SDs and its effect on M

The 13-mer sequence motif (CCNCCNTNNCCNC) is significantly associated with the homologous recombination hotspots in human genome (12). Therefore, we also investigated the density of this 13-mer motif located in SDs and its potential roles in facilitating NAHR. However, the addition of 13-mer motif density into the previous four parameters (D, A, F and R) did not improve the modeling performance (Supplementary Material, Figs S6 and S7). With the 13-mer motif density, the optimal spread constants are basically the same as those without the motif density (Supplementary Material, Figs S4 and S7). Notably, the minimum MSEs of test sets are even slightly higher than those without the motifs (Supplementary Material, Figs S2 and S6). Therefore, the density of the 13-mer motifs in SDs seems to have a limited contribution to NAHR. Consistently, a recent study on rare pathogenic CNVs did not reveal a statistically significant correlation between 13-mer motif density and NAHR frequency (18).

The possible explanations for the observed limited effect of 13-mer motif density on NAHR are as follows. The process of both allelic HR and NAHR can be divided into three steps: (i) homology search, (ii) the identified homologies approaching to each other and (iii) the completion of recombination mediated by recombination-related proteins (32). During these processes,
NAHR is mainly different from allelic HR in the first two steps, whereas the last step of recombination is similar for both allelic HR and NAHR. As for the 13-mer hotspot motif, it may only affect the last step mediated by recombination-related proteins. Therefore, the 13-mer motif may not be a significant factor specific for NAHR.

In addition, the 13-mer motif may only account for a small portion of recombination hotspots in the human genome. There are still unknown recombination motifs remaining to uncover. Therefore, the 13-mer motif itself may not be able to reveal a full view of recombination hotspots.

**GRNN versus other mathematical tools**

In addition to GRNN, other mathematical tools, including linear regression, nonlinear regression and decision tree (Supplementary Material, Note), were also employed in our preliminary analyses. However, none of them worked as well as GRNN. For example, the linear model is not eligible to analyze nonlinear correlation. Some nonlinear models require large number of parameters, thus result in over-parameterization. The method of decision tree can only generate discrete values as output, which results in lower resolution. Consequently, GRNN is the only tool that successfully catches the majority of the nonlinear and/or linear correlation and has good resolution when analyzing both intermediate and extreme cases. In addition, only one parameter (the spread constant) other than the input factors is required to be set in GRNN. Therefore, we use the GRNN as our major analyzing method.

In summary, we established a black box model using GRNN and successfully revealed the relationships between the frequencies of SD-mediated NAHR events and SD properties in the NAHR-mediated CNVs of human populations. These observations can be well explained by ectopic synopsis in NAHR together with our proposed model of CCEL for homology mis-pairing. Our findings will be meaningful to the future researches on both NAHR-mediated human CNVs and repeat-driven genome evolution.

**Materials and Methods**

**Relative mutation rate of human CNVs estimated in HapMap populations**

In our previous study, a genome-wide analysis for relative mutation rate of CNV was conducted using the phylogeny information of CNV-flanking SNPs in human populations (19). A statistic, M, which is positively correlated with mutation rate, was estimated for 4330 CNVs in 3 HapMap populations: YRI, CEU and CHB+JPT (19,20). We used M as the measurement of relative CNV mutation rate in this study.

NAHR-mediated CNVs between paired SD homologies

The long genomic repeats, SDs, are major substrates for NAHR in the human genome. To screen for the CNVs mediated by NAHR between paired SDs, we first obtained the human SD data of the Eichler lab (21) via UCSC Genome Browser (http://genome.ucsc.edu). These data include coordinate, alignment length and fraction match (i.e. sequence similarity) of paired SDs. The human genome assembly of hg18 is used for the coordinate consistency of the SD and CNV data. The NAHR events between paired SDs in direct orientation can generate CNVs (2). Therefore, we only collected the direct SDs on the same arm of each human chromosome. Then, we matched these SD pairs with the 4330 CNVs according to their coordinates (Supplementary Material, Fig. S8). Owing to the limited resolution of CNV genotyping technologies (20), the CNVs with both breakpoints located within paired SDs or their 1-kb flanking region are most likely generated by NAHR (Supplementary Material, Fig. S8).

**SD properties and parameters**

Based on the experimental observations in previous studies (14,16,21,22), the following four properties of paired SDs, including alignment length, distance, sequence similarity (measured by fraction match) and chromosomal position, were hypothesized to affect NAHR frequency and related CNV mutation rate. In order to make the raw data suitable for further mathematical analysis, we introduced four parameters (A, D, F and R) with values approximately ranging from 0 to 1 to measure these four SD properties, respectively.

\[
\text{distance} = \frac{(SD_2\text{start} + SD_2\text{end})}{2} - \frac{(SD_1\text{start} + SD_1\text{end})}{2} \tag{1}
\]

\[
D = \frac{\log(\text{distance}) - 3}{2.41} \tag{2}
\]

\[
A = \frac{\log(\text{alignment length}) - 3}{1.95} \tag{3}
\]

\[
\text{position} = \frac{SD_1\text{start} + SD_1\text{end} + SD_2\text{start} + SD_2\text{end}}{4} \tag{4}
\]

\[
R = \begin{cases} 
|\text{position} - \text{Centromere}| & (\text{position} > \text{Centromere}) \\
|\text{position} - \text{Centromere}| & (\text{position} < \text{Centromere}) 
\end{cases} \tag{5}
\]

\[
F = \frac{\text{fraction match} - 0.9}{0.1} \tag{6}
\]

The distance is log-transformed and then normalized to the interval of (0, 1). The minimum of distance is \(-10^3\) bp, and the maximum is \(-10^{6.41}\) bp; therefore, log-transformed distance is subtracted by 3 and then divided by 2.41 in order to get D. Similarly, alignment length is also log-transformed and then normalized to (0, 1) in order to get A.

In addition to these four main SD properties, the density of 13-mer motif in SDs and its role in NAHR was also investigated. The data of NAHR-associated CNVs and their mediating SD pairs are shown in Supplementary Material, Table S1.

**Generalized regression neural network**

GRNN is a type of artificial neural networks (ANNs), which are especially effective in solving nonlinear problems (33). The basic unit of ANN is an artificial neuron, and the structure of an artificial neuron is shown in Supplementary Material, Figure S9.

In this study, the GRNNs are three-layer feed-forward networks (Supplementary Material, Fig. S1). According to Hornik et al. (34), any continuous multivariate function can be approximated by a three-layer feed-forward network. Generally, the more neurons in the hidden layer, the better the network fits the target function. However, when there are too many hidden-layer neurons, over-fitting would occur and the network’s generalization ability would deteriorate. For a GRNN, the number of
hidden-layer neurons negatively correlates with the spread constant (33). Therefore, it is crucial to determine a proper spread constant by analyzing the MSE in test sets. Usually, with an optimal spread constant, the MSE of test set reaches the minimum and the MSE of training set keeps an acceptable low level. In this study, an MSE of <0.05 indicates good approximation for a training set or good generalization for a test set.

We used the Neural Network Toolbox of MATLAB (R2012b). A, D, F and R are input variables, and M is output variable. We constructed three GRNNs (I–III) using one population as the training set and the other two as the test sets (Supplementary Material, Fig S2).

Excluding outliers
First, we choose an optimal spread constant in order to minimize the average MSE of test sets for each of the three GRNNs, thus avoiding over-fitting. Second, we define the samples with M residuals of which the absolute values are bigger than the square root of MSE as potential outliers. Third, if a sample is a potential outlier in at least two of the three approximation or generalization processes, it is identified as an outlier.

Principal component analysis
We employed PCA in assessing the contributing weights of various SD properties in mediating CNV mutations via NAHR. Here, we use Bartlett’s test of sphericity and KMO test to confirm whether PCA is applicable (35). For Bartlett’s test of sphericity, the null hypothesis is that the sample correlation matrix comes from a multivariate normal population in which the variables of interest are independent. If the null hypothesis is rejected, the analysis is applicable. For KMO test, the closer the MSA is to 1, the more applicable PCA is. If MSA is <0.5, the data are not appropriate for analysis. Correlation matrix is used in PCA, and we estimated the weights of D, A, R and F according to unrotated and rotated component matrix; the rotation method is varimax. The PCA tool of SPSS 20 was used.

Input variable coordinate
IVC is defined as follows:

\[
IVC = 1000(10D - 1) + 100(10A - 1) + 10(10R - 1) + 10F
\]

\[
D \in X
\]

\[
A \in X
\]

\[
R \in X
\]

\[
F \in X
\]

Therefore, a combination of D, A, R and F values can be represented by an IVC value. For example, IVC = 6341 means that D = 0.7, A = 0.4, R = 0.5 and F = 0.1.

Supplementary Material
Supplementary Material is available at HMG online.

Conflict of Interest statement. None declared.

Funding
This work was supported by National Basic Research Program of China (2012CB944600 and 2011CBA00401), National Natural Science Foundation of China (81222014, 31171210 and 31000552), Shu Guang Project (12SG08), Shanghai Pujiang Program (10PJ1400300) and Recruitment Program of Global Experts.

References
16. Liu, P., Lacaria, M., Zhang, F., Withers, M., Hastings, P.J. and Lupski, J.R. (2011) Frequency of nonallelic homologous recombination is correlated with length of homology: evidence...


