SCIENTIFIC REPORTS

OPEN

Effects of short indels on protein structure and function in human genomes

Received: 18 May 2017
Accepted: 24 July 2017
Published online: 24 August 2017

Maoxuan Lin, Sarah Whitmire, Jing Chen, Alvin Farrel, Xinghua Shi & Jun-tao Guo

Insertions and deletions (indels) represent the second most common type of genetic variations in human genomes. Indels can be deleterious and contribute to disease susceptibility as recent genome sequencing projects revealed a large number of indels in various cancer types. In this study, we investigated the possible effects of small coding indels on protein structure and function, and the baseline characteristics of indels in 2504 individuals of 26 populations from the 1000 Genomes Project. We found that each population has a distinct pattern in genes with small indels. Frameshift (FS) indels are enriched in olfactory receptor activity while non-frameshift (NFS) indels are enriched in transcription-related proteins. Structural analysis of NFS indels revealed that they predominantly adopt coil or disordered conformations, especially in proteins with transcription-related NFS indels. These results suggest that the annotated coding indels from the 1000 Genomes Project, while contributing to genetic variations and phenotypic diversity, generally do not affect the core protein structures and have no deleterious effect on essential biological processes. In addition, we found that a number of reference genome annotations might need to be updated due to the high prevalence of annotated homozygous indels in the general population.

Insertions and deletions (indels) are additions or deletions of one or more nucleotides in DNA sequence. Indels are highly abundant in human genomes, second only to single nucleotide polymorphisms (SNP), and make up 15–21% of human polymorphisms[1]. Indels in coding regions can result in two different types of variants, frameshift (FS) and non-frameshift (NFS). NFS indels consist of a multiple of three base pairs, introducing an insertion or deletion of one or more amino acids while keeping the rest of the protein sequence unchanged. In contrast, FS indels change the reading frame starting from the site of insertion/deletion, which can produce different protein sequences or lead to premature termination and the mRNA can be subjected to a surveillance pathway called non-sense-mediated mRNA decay (NMD)[2]. A rate of 2.94 indels (1–20 bp) and 0.16 structural variants (>20 bp) per generation was estimated based on whole genome sequencing of 250 families[3]. While regarded as an alternative of natural genetic variation to SNP, previous studies have demonstrated the role of indels in the development of a number of Mendelian diseases[4–6]. For example, cystic fibrosis, with an incidence rate of 1 in 3500 in North America, is caused by a three base-pair deletion within the *CFTR* (Cystic Fibrosis Transmembrane Conductance Regulator) gene[7,8]. Indels have also been implicated in diseases including acute myeloid leukemia[9,10] and other types of cancer[11].

With the advancement of sequencing techniques and cost reduction, a large number of personal genomes, both from healthy individuals and cancer patients have been sequenced, which sped up the process of building a comprehensive catalog of indel variants[1,12–22]. For example, the 1000 Genomes Project, the largest public catalogue of human variation and genotype data, has recently completed its final phase in 2015[23,24]. The project sequenced 2,504 individual genomes representing 26 diverse populations in Africa (AFR), the Americas (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS). The landscape of natural genetic variations and somatic mutations, including indels, has been investigated in an attempt to discover deleterious mutations[19–21,25–27]. Several machine-learning methods have been developed to predict the phenotypic effect of both FS[28–30] and NFS indels[5,30–32]. The disease-causing indels are generally derived from the Human Gene Mutation Database (HGMD), while the neutral indels are from the 1000 Genomes Project or curated from protein sequence databases. The structural effects of small NFS indels have also been investigated using protein isoform structures or highly homologous protein structures in Protein Data Bank (PDB)[33–35]. Results show that protein structures can

Department of Bioinformatics and Genomics, The University of North Carolina at Charlotte, Charlotte, NC, 28223, USA. Correspondence and requests for materials should be addressed to J.-t.G. (email: jguo4@uncc.edu)

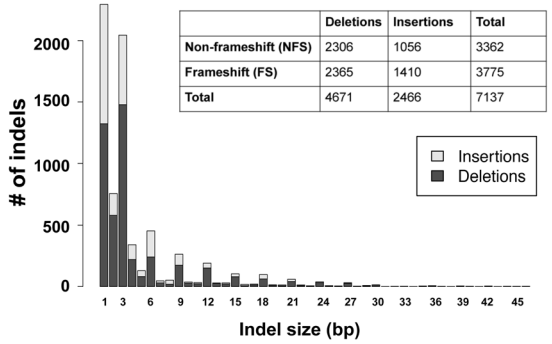

Figure 1. Indel size distribution.

tolerate small natural indels as the majority of indel residues are exposed to the solvent and about one-third of residues are in disordered state[35].

While efforts have been devoted to predictions of potential pathogenicity of small indels, there are no comprehensive studies of the effect of short coding indels on protein structure and function in a large number of human genomes. In this study, we focused on the analysis of short coding indels (<50 bp) in the 1000 Genomes Project to explore the role of these genetic variations in protein structure (for NFS indels) and function (for both FS and NFS indels), which can serve as background characteristics for studying disease-causing indels in various diseases. In addition, we identified a number of genes with homozygous FS and NFS indels that have very high frequency among the diverse populations, which may serve as basis for future reference genome updates.

Results and Discussion

Distribution of short coding indels. There are a total of 769,743 short coding indels in the 2,504 human genomes, where raw indels were first called based on the numbers of reads supporting reference and alternative alleles and the genotypes of these indels were further refined by considering SNPs genotypes and haplotype structure[23,24]. While some coding indels are rare variants, 209 homozygous indels (72 NFS and 137 FS) were found in over 50% of the 2,504 individuals (Supplementary Table S1). Among them, 1 NFS and 61 FS homozygous indels appear in all 2,504 genomes (Supplementary Tables S2 and S3). These high frequency homozygous indels should be a point of interest for human reference genome updates as suggested in structural variants studies[23]. The number of unique indels is 7,137 (if the same indel occurs in multiple genomes, it only counts as one unique indel). There are slightly more FS indels (3,775) than NFS indels (3,362). About 37% (1240/3362) of the NFS indels and 32.5% (1226/3775) of the FS indels have more than 1% of allele frequency.

There are about twice as many deletion indels (4,671) than the insertion indels (2,466) (Fig. 1). Short coding indels are highly enriched. Insertions and deletions of one to three nucleotides represent about 70% of all unique coding indels. Except for the single nucleotide indel, which has the highest occurrence, there are more NFS indels (multiple of three nucleotides) than FS indels in each of the three nucleotides window (Fig. 1). Since two out of three codon positions result in FS indels, the number of FS indels is smaller than expected, which is not surprising as FS indels are considered more deleterious, and mutants with such indels are more likely to be removed from population through purifying selection[36,37]. The actual number of FS indels could be even smaller after updates of the human reference genome in the future since a number of homozygous FS indels appear in every individual genome (Supplementary Table S3). In addition, in some cases a second FS indel may rescue a potential deleterious variant of the first FS indel by correcting the open reading frame. For example, a 2 bp insertion at position 74,836,315 on one individual's *ARID3B* gene (ENSG00000179361) can be rescued by a 1 bp insertion at the next position 74,836,316. However, if the two FS positions are so far away, it makes it a different protein sequence between the two variant sites. Gene *CLTCL1* (ENSG00000070371) on one genome is such an example. It has a 1 bp insertion at position 19,189,003 and a 10 bp deletion at position 19,170,999. Even though the combination of these two FS indels results in a 9 bp deletion, a relatively larger piece of the protein sequence involving several exons is changed. A list of genes with at least two FS indels on one individual's same gene is shown in Supplementary Table S4. Not only can an FS indel introduce premature stop codon, NFS can also introduce a premature stop codon, we found a total of eight such unique cases (Supplementary Table S5).

The distribution of the unique indels on each chromosome is shown in Fig. 2. While the chromosomes are generally numbered from the largest to the smallest, the protein coding genes are unevenly distributed across the chromosomes. For example, chromosome 19, one of the smallest chromosomes, has the highest gene density of all human chromosomes[38,39]. The next highest gene dense chromosomes are 17, 22, 16, and 11, while the lowest density chromosomes are 13 and 18[39]. Therefore the number of unique indels on each chromosome is closely related to its number of protein coding genes (Fig. 2). Coding indels are enriched in N- or C-terminal regions (Fig. 3). It is not surprising to observe that there are more N-terminal indels in the NFS cases and more C-terminal indels in the FS cases. In both situations, a majority of the protein sequences are not changed and indels should have minimal effects on the structure and function of affected proteins.

Principal component analyses (PCA) of indel patterns revealed the clustering of 26 populations into their respective five super populations (Fig. 4). There are no clear differences between the results from all indels (Fig. 4A) and homozygous indels (Fig. 4B). The first two principal components PC1 and PC2 explain about 77%

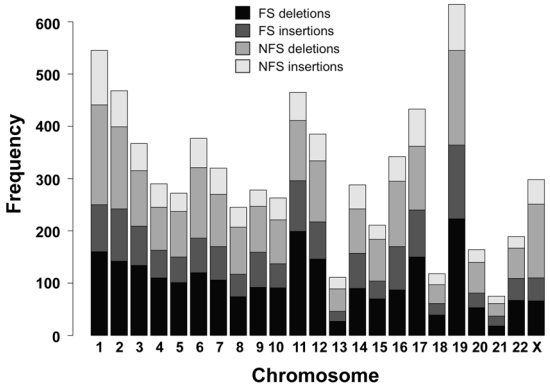

Figure 2. Number of unique indels on each chromosome.

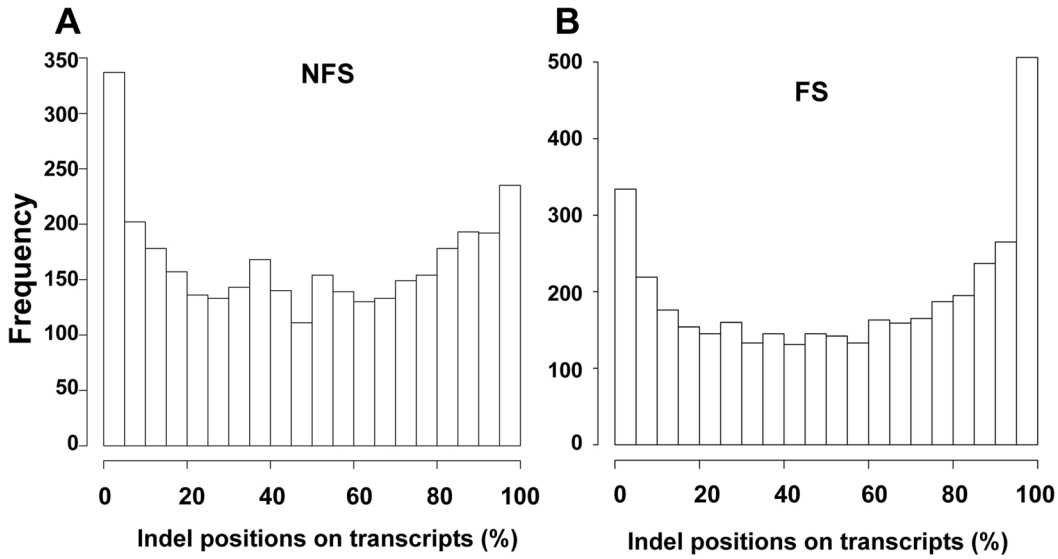

Figure 3. Relative positions of NFS (**A**) and FS (**B**) indels on proteins.

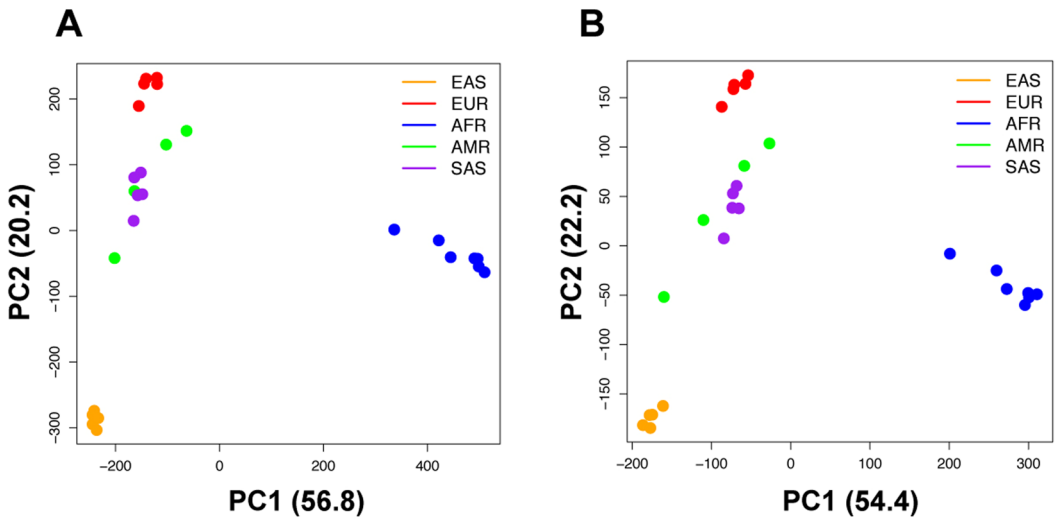

Figure 4. PCA analysis of indel patterns in 26 populations. (**A**) All indels; (**B**) Homozygous indels only

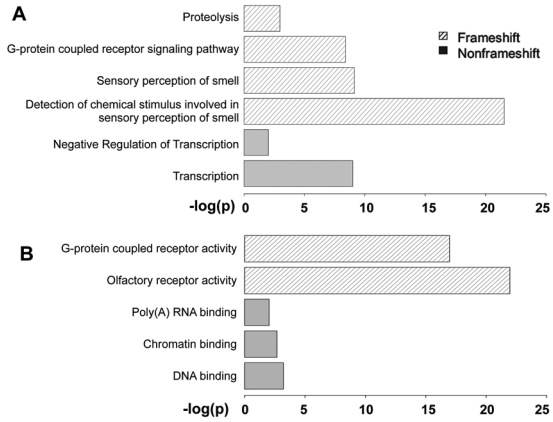

Figure 5. Gene enrichment analysis of genes with NFS or FS indels. (**A**) Significantly enriched categories in terms of Biological Process; (**B**) Significantly enriched categories in terms of Molecular Function.

of the variations of the indel patterns. In both cases, AFR can be clearly separated from other super populations by the first principal component, and the second principal component further separates EAS and EUR from the other two ancestries: SAS and AMR. There is some overlap between AMR and SAS based on the first two principal components. Our indel PCA results are consistent with the broad patterns from structural variants (SV, defined as DNA variants of more than 50 bp) PCA analyses[23]. The data suggest that each super population has its own distinct patterns of indels, which may potentially contribute to the phenotypic differences among the populations. For example, an FS indel on *GPR142* (ENSG0000257008) was only found in AFR super populations and another indel on *LGR6* (ENSG0000133067) has different frequencies in AMR, AFR and SAS with zero occurrences in EUR and EAS. Recent report on global reference for human genetic variants revealed similar results[24]. About 762,000 rare variants (<0.5% in full population) were found frequently in at least one population (>5%) and populations with higher numbers of variants were geographically separated. This is especially true for the AFR populations[24].

Functional enrichment analysis. To investigate possible associations between short coding indels and functional categories of the affected proteins, we applied DAVID v6.8, the "Database for Annotation, Visualization and Integrated Discovery" and performed functional enrichment analysis[40]. The categories are analyzed based on Gene Ontology (GO)'s Biological Process and Molecular Function annotations respectively and the significantly enriched categories were selected using an FDR threshold of 0.05[41]. We observed different enrichment patterns between genes with FS and NFS indels (Fig. 5). In terms of biological process, the top three significantly enriched categories in FS related genes are all olfactory-related: detection of chemical stimulus-smell, sensory perception of smell, and G-protein coupled receptor signaling pathway (Fig. 5A). In NFS cases, transcription-related biological processes are highly enriched (Fig. 5A). Results from molecular function enrichment analysis are consistent with corresponding biological process data (Fig. 5B).

Olfactory receptor activity and G-protein coupled receptor (GPCR) activity are the two significantly enriched GO functional categories in FS cases. Further analysis revealed a big overlap of genes between these two categories, 81.4% of the analyzed genes involved in GPCR activity also have the same GO terms in olfactory receptor activity. In other words, the majority of the GPCR-related genes make up the olfactory receptor activity. The genetic variation in human olfactory receptors, one of the largest gene families in humans, has been linked to phenotypic diversity[42]. The sense of smell is a complex process and requires a large number of olfactory receptors to differentiate minute differences among thousands of combinations of chemicals with differing concentrations[43]. Enrichment of olfactory-related genes for FS indels have been reported previously from investigation of genetic variation in an individual human exome and a systematic survey of loss of function (LoF) variants in human protein-coding genes[6,28,44]. Another study comparing human and chimpanzee olfactory receptor gene repertoires suggested that these genes are under relaxed selection, which may explain the relatively large number of variants in olfactory genes[45].

The NFS enrichment analysis indicated an overrepresentation of transcription-related coding indels from 2,504 individual genomes. This is consistent with previous studies that demonstrated high variations in transcription-related genes and their potential link to phenotypic diversity[46,47]. Ribeiro-dos-Santos *et al.* characterized transcription-related genes that have been the target of positive evolutionary forces[46]. In addition to describing a similar enrichment of transcription-related indels and their possible role in positive selection, Chen *et al.* suggested that these indels may contribute to the diversity of RNA and protein levels in humans, which gives rise to our unique traits[47]. The effects of these transcription-related NFS indels on protein structure and function are discussed in the next subsection. We also performed PCA analysis using genes with transcription-related indels only. There are 405 unique genes with 496 unique indels (322 deletions and 174 insertions) in transcription-related coding regions. Similar to the full indel data analysis, PCA analysis showed similar distinct clustering into the five super populations (Fig. 6). The AMR populations are less separated from SAS and EAS in the homozygous transcription-related indel analysis (Fig. 6B) than the all transcription-related indel analysis (Fig. 6A). This may be caused by a combination of two factors: a low number of homozygous

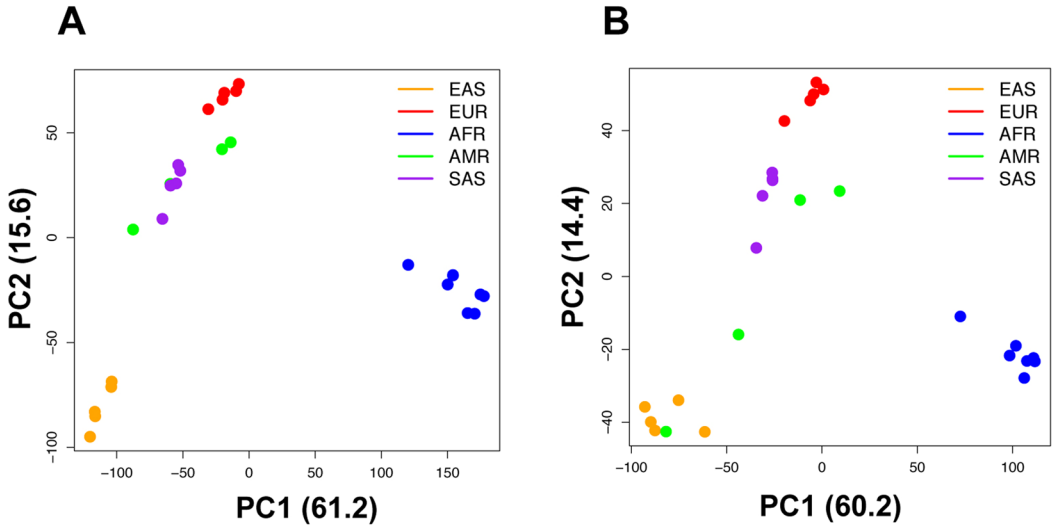

Figure 6. PCA analysis of transcription-related indel patterns in 26 populations. (**A**) All indels; (**B**) Homozygous indels only

transcription-related indels and the admixture of populations as discussed in investigation of human structural variants[23].

Since small coding indels in the 1000 Genome Project have different allele frequency, one interesting question is if there are any differences in functional enrichment between common ($\geq$ 1% allele frequency, about 32–37% of all indels) and rare indels (<1% allele frequency). Functional enrichment analysis showed similar results between the common and rare indels, *i.e.* NFS indels are significantly enriched ($p\text{-value} < 0.05$) in transcription-related genes and FS indels are enriched in olfactory-related genes and activities (Supplementary Figure S1).

Effects of short NFS indels on protein structure. Due to the low number of matches of coding indels to known protein structures in PDB (79 NFS deletions and 12 NFS insertions)[33], the secondary structure types of the remaining NFS indels were predicted as described in Methods. These coding NFS indels are depleted in the two regular major secondary structure types, helix (11%) and strand (9%), and highly enriched in coil conformation (80%) when compared to the background secondary structure type distribution as we reported previously (helix: 36%, strand: 20.8%, and coil: 43.2%, p-value of chi-square test $<2.2 \times 10^{-16}$)[35] (Fig. 7A). There is no clear difference between NFS deletions and NFS insertions regarding their effect on secondary structures of proteins. Disordered residue prediction showed a similar pattern to that of secondary structure types for these NFS indels. Only about 20% of the NFS indel residues are predicted as ordered while about 60% of the residues are predicted as disordered (Fig. 7B). These results are consistent with our previous structural analysis of "natural" indels in PDB and the published work by Zhao *et al.*, which showed a depleted regular secondary structure types (helix and strand) and highly enriched in disorder and coil conformation[31,35].

Since transcription-related genes are enriched in NFS indels, we examined the secondary structure types and disorder prediction of these indels to see if there are any significant differences between transcription-related and all NFS indels. In terms of secondary structure types, there are more coil types and fewer helix and strand conformations (p-value of chi-square test is 0.002) (Fig. 7A,C). Moreover, the disorder prediction is significantly different in transcription-related NFS indels compared to all NFS indels (p-value of chi-square test $<2.2 \times 10^{-16}$). There are more disordered residues in transcription-related NFS indels (Fig. 7B,D). The above results suggest that these transcription-related indels may keep the core of transcription-related proteins intact while introducing variations at the coil regions, providing differences in DNA binding affinity/specificity and contributes to phenotypic diversity[48,49]. Binding differences have been shown to correlate well with differences in gene expression, which is a driving force in the evolution of organisms and plays an important role in phenotypic diversity[50–52].

The structural effects were also compared between common and rare NFS indels. High frequency common NFS indels tend to have slightly more coil/disordered residues and fewer ordered residues including helix and strand secondary structure types than the rare NFS indels (Fig. 8). There are bigger differences between insertion and deletion indel types in common indels than those in rare indels, especially in homozygous NFS indels (Supplementary Figures S2 and S3). However, caution should be taken about these small differences, as they could be well within the prediction errors in protein secondary structure and disorder predictions.

Taken together, NFS indels from the general population genomes tend to locate on non-core structural segments and may have minimal effect on protein structural integrity. The deletion and insertion of small fragments in the coil region may result in differences in binding affinity and gene expression, which in turn can drive evolution and contribute to the diversity of phenotypes[48,49].

Conclusion

Accurate prediction of structural and functional effects of indels, the second largest type of genetic variation in human genomes, is of paramount importance in interpretation of variation in genomes from various diseases.

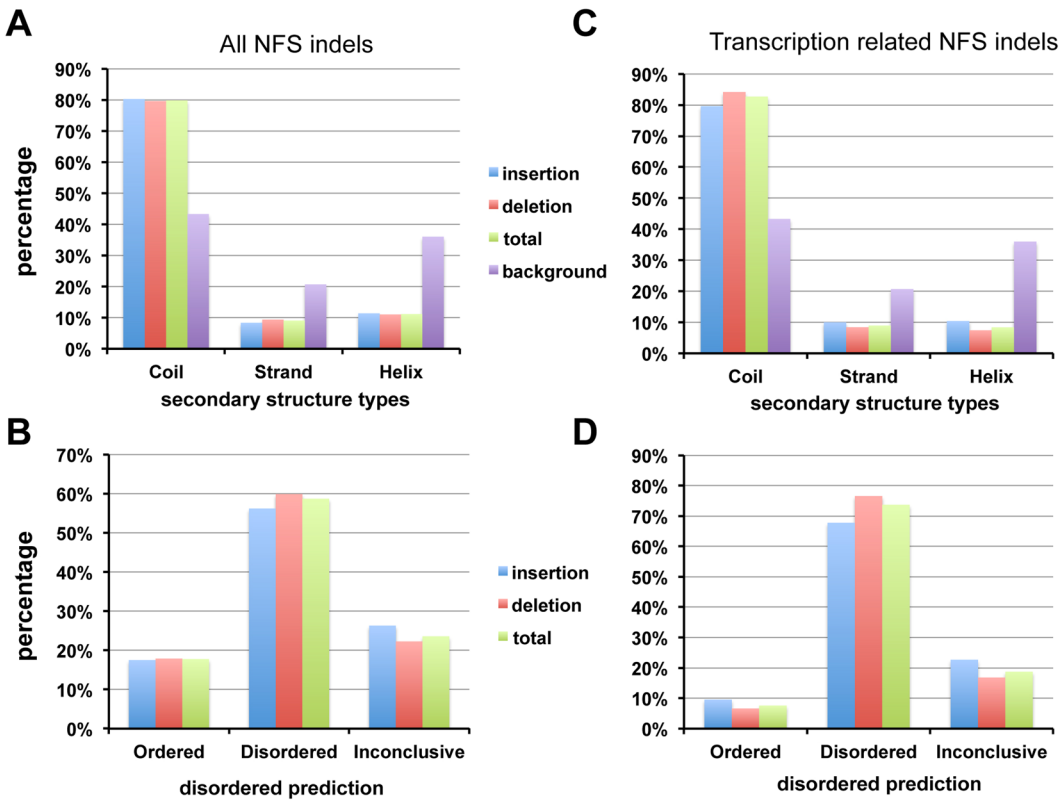

Figure 7. Secondary structure and residue disorder types for NFS indels. (**A**) Distribution of secondary structure types of all NFS indels; (**B**) Distribution of secondary structure types of transcription-related NFS indels; (**C**) Distribution of residue disorder of all NFS indels; (**D**) Distribution of residue disorder of transcription-related NFS indels. A residue in an indel is considered "disordered" or "ordered" if both IUPred and DisProt agree; otherwise it is annotated as "inconclusive".

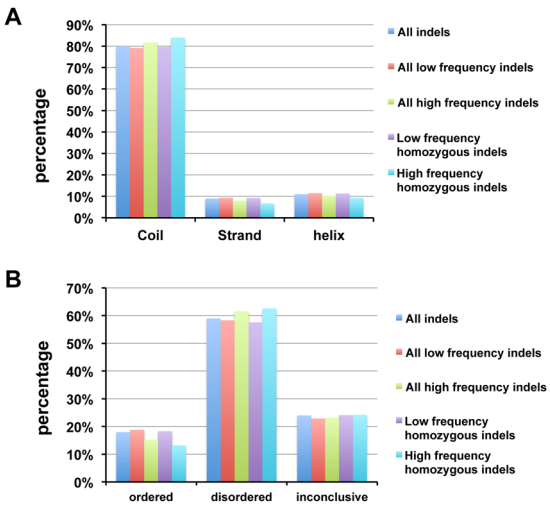

Figure 8. Comparisons of structural types between high and low allele frequency NFS indels. (**A**) Distribution of secondary structure types; (**B**) Distribution of residue disorder. A residue in an indel is considered "disordered" or "ordered" if both IUPred and DisProt predictions agree. Otherwise it is annotated as "Inconclusive".

In this study, we analyzed all short indels in coding regions on chromosome 1–22 and chromosome X from the 1000 Genomes Projects to establish baseline characteristics of short coding indels in general, non-diseased populations. We found that these short NFS and FS indels are more likely to occur in N- and C-terminal regions and

assume coil or disordered conformations. For the functional effects, FS indels are highly enriched in olfactory receptors while NFS indels are mainly associated with transcription-related functionalities.

FS coding indels are considered more deleterious as they change protein sequences and may result in loss-of-function variants for essential proteins. It is not surprising that the number of short FS coding indels is smaller than expected, as deleterious mutants are more likely to be removed from the population by natural selection. FS indels found in healthy individuals generally are less deleterious and contribute to phenotypic diversity through different ways. First, a second FS indel may rescue potential deleterious effect of the first FS indel by correcting the open reading frame (see Results section). Secondly, the protein with an FS indel might be non-essential or has other similar proteins to carry out the same function. Thirdly, a heterozygous FS indel has a normal copy of the gene to carry out the essential function unless the variant is dominant. Lastly, mis-annotations on the human reference genome also contribute to some of the FS indel cases, especially for the 100% frequency of homozygous FS indels (Supplementary Table S3).

Methods

Dataset. Raw variant call format (.vcf) files of the 1000 Genomes Project phase 3, including variant calls of chromosome 1 to 22 and chromosome X, were downloaded from the 1000 Genomes Project at http://www.1000genomes.org. All variants were annotated based on the coordinates of these variants with Variant Effect Predictor[53]. Since the goal of this study was to study the effect of indels on protein structure and function, we only selected insertions and deletions in coding regions, including frameshift and non-frameshift indels.

Indel distribution and gene enrichment analysis. Indels' distribution in 26 populations was analyzed. In counting the number of unique indels, the same indel occurring in multiple genomes was counted as 1. The relative frequency of each indel of a gene in each population was calculated, and the indel population patterns were visualized using PCA to identify if geographical and ancestral backgrounds can account for the distribution of coding indels. We also performed PCA on the homozygous only indels.

To investigate the functional categories of genes affected by these small indels, we applied DAVID 6.8, (the Database for Annotation, Visualization and Integrated Discovery) to perform functional enrichment analysis[40]. Lists of FS genes and NFS genes were analyzed separately. A cutoff of 0.05 was set for FDR (False discovery rate) to identify the significantly enriched functional categories.

Protein structural analysis. To avoid redundancy, only protein sequence derived from the longest transcript was selected, which was downloaded from Ensembl's FTP site (http://grch37.ensembl.org/info/data/ftp/index.html). Since FS indels change the amino acid sequences starting at the indel sites, we only performed structural analysis on NFS indels. The protein sequences with indels were first blasted against protein sequences *pdbaanr* with known structures in Protein Data Bank (PDB)[33,54]. The alignments that had E-values less than 0.001 with at least 80% sequence identity and 50% coverage were selected. The secondary structure types of the deletion sequences with a reference protein structure were assigned using DSSP[55]. For protein sequences with indels that did not have corresponding protein structures available and insertion sequences that did not have corresponding secondary structures, the secondary structure types were predicted with RaptorX-SS8, an 8-class secondary structure prediction method[56]. Each indel residue was assigned to one of four secondary structure states, helix, strand, coil and disordered. DSSP program was used to assign three secondary structure states: helix, strand and coil following the widely used convention, H (α-helix), G (3_{10}-helix) and I (π-helix) from DSSP as helix type; E (extended strand) and B (residue in isolated β-bridge) states as strand type and all the other states from DSSP are considered as coil[35]. The disordered residues were defined by comparing the "ATOM" and "SEQRES" records in PDB file. If a residue or a fragment appeared in "SEQRES", but is missing from the "ATOM" record in a PDB file, this residue or fragment was considered disordered or unstructured[57]. Disorder predictions of indel residues were performed using IUPred[58] and DisProt[59].

Data availability. The data used in this study were downloaded from the 1000 Genomes Project.

References

1. Mullaney, J. M., Mills, R. E., Pittard, W. S. & Devine, S. E. Small insertions and deletions (INDELs) in human genomes. *Human molecular genetics* **19**, R131–136, doi:10.1093/hmg/ddq400 (2010).
2. Brogna, S. & Wen, J. Nonsense-mediated mRNA decay (NMD) mechanisms. *Nature structural & molecular biology* **16**, 107–113, doi:10.1038/nsmb.1550 (2009).
3. Kloosterman, W. P. *et al.* Characteristics of de novo structural changes in the human genome. *Genome Res* **25**, 792–801, doi:10.1101/gr.185041.114 (2015).
4. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet* **12**, 745–755, doi:10.1038/nrg3031 (2011).
5. Bermejo-Das-Neves, C., Nguyen, H. N., Poch, O. & Thompson, J. D. A comprehensive study of small non-frameshift insertions/deletions in proteins and prediction of their phenotypic effects by a machine learning method (KD4i). *BMC Bioinformatics* **15**, 111, doi:10.1186/1471-2105-15-111 (2014).
6. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828, doi:10.1126/science.1215040 (2012).
7. Kosorok, M. R., Wei, W. H. & Farrell, P. M. The incidence of cystic fibrosis. *Statistics in medicine* **15**, 449–462, doi:10.1002/(SICI)1097-0258(19960315)15:5<449::AID-SIM173>3.0.CO;2-X (1996).
8. Collins, F. S. *et al.* Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science* **235**, 1046–1049 (1987).
9. Falini, B. *et al.* Cytoplasmic nucleophosmin in acute myelogenous leukemia with a normal karyotype. *N Engl J Med* **352**, 254–266, doi:10.1056/NEJMoa041974 (2005).
10. Nakao, M. *et al.* Internal tandem duplication of the flt3 gene found in acute myeloid leukemia. *Leukemia* **10**, 1911–1918 (1996).
11. Ye, K. *et al.* Systematic discovery of complex insertions and deletions in human cancers. *Nature medicine* **22**, 97–104, doi:10.1038/nm.4002 (2016).

12. Levy, S. *et al*. The diploid genome sequence of an individual human. *PLoS Biol* **5**, e254, doi:10.1371/journal.pbio.0050254 (2007).
13. Wheeler, D. A. *et al*. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876, doi:10.1038/nature06884 (2008).
14. Wang, J. *et al*. The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65, doi:10.1038/nature07484 (2008).
15. Bentley, D. R. *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59, doi:10.1038/nature07517 (2008).
16. Kim, J. I. *et al*. A highly annotated whole-genome sequence of a Korean individual. *Nature* **460**, 1011–1015, doi:10.1038/nature08211 (2009).
17. Ahn, S. M. *et al*. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* **19**, 1622–1629, doi:10.1101/gr.092197.109 (2009).
18. Schuster, S. C. *et al*. Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943–947, doi:10.1038/nature08795 (2010).
19. Nik-Zainal, S. *et al*. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54, doi:10.1038/nature17676 (2016).
20. Mills, R. E. *et al*. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res* **16**, 1182–1190, doi:10.1101/gr.4565806 (2006).
21. Mills, R. E. *et al*. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res* **21**, 830–839, doi:10.1101/gr.115907.110 (2011).
22. Montgomery, S. B. *et al*. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res* **23**, 749–761, doi:10.1101/gr.148718.112 (2013).
23. Sudmant, P. H. *et al*. An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81, doi:10.1038/nature15394 (2015).
24. Genomes Project, C. *et al*. A global reference for human genetic variation. *Nature* **526**, 68–74, doi:10.1038/nature15393 (2015).
25. Kidd, J. M. *et al*. Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64, doi:10.1038/nature06862 (2008).
26. Weber, J. L. *et al*. Human diallelic insertion/deletion polymorphisms. *American journal of human genetics* **71**, 854–862, doi:10.1086/342727 (2002).
27. Bhangale, T. R., Rieder, M. J., Livingston, R. J. & Nickerson, D. A. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Human molecular genetics* **14**, 59–69, doi:10.1093/hmg/ddi006 (2005).
28. Hu, J. & Ng, P. C. Predicting the effects of frameshifting indels. *Genome Biology* **13**, R9, doi:10.1186/gb-2012-13-2-r9 (2012).
29. Folkman, L. *et al*. DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels. *Bioinformatics* **31**, 1599–1606, doi:10.1093/bioinformatics/btu862 (2015).
30. Douville, C. *et al*. Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST-Indel). *Hum Mutat* **37**, 28–35, doi:10.1002/humu.22911 (2016).
31. Zhao, H. *et al*. DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. *Genome Biology* **14**, R23, doi:10.1186/gb-2013-14-3-r23 (2013).
32. Hu, J. & Ng, P. C. SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. *PLoS One* **8**, e77940, doi:10.1371/journal.pone.0077940 (2013).
33. Berman, H. M. *et al*. The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
34. Studer, R. A., Dessailly, B. H. & Orengo, C. A. Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochem J* **449**, 581–594, doi:10.1042/BJ20121221 (2013).
35. Kim, R. & Guo, J. T. Systematic analysis of short internal indels and their impact on protein folding. *BMC Struct Biol* **10**, 24, doi:10.1186/1472-6807-10-24 (2010).
36. Taylor, M. S., Ponting, C. P. & Copley, R. R. Occurrence and consequences of coding sequence insertions and deletions in Mammalian genomes. *Genome Res* **14**, 555–566, doi:10.1101/gr.1977804 (2004).
37. de la Chaux, N., Messer, P. W. & Arndt, P. F. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evol Biol* **7**, 191, doi:10.1186/1471-2148-7-191 (2007).
38. Grimwood, J. *et al*. The DNA sequence and biology of human chromosome 19. *Nature* **428**, 529–535, doi:10.1038/nature02399 (2004).
39. Gilbert, N. *et al*. Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers. *Cell* **118**, 555–566, doi:10.1016/j.cell.2004.08.011 (2004).
40. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44–57, doi:10.1038/nprot.2008.211 (2009).
41. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29 (2000).
42. Hasin-Brumshtein, Y., Lancet, D. & Olender, T. Human olfaction: from genomic variation to phenotypic diversity. *Trends Genet* **25**, 178–184, doi:10.1016/j.tig.2009.02.002 (2009).
43. Fushimi, K., Osumi, N. & Tsukahara, T. NSSRs/TASRs/SRp38s function as splicing modulators via binding to pre-mRNAs. *Genes Cells* **10**, 531–541 (2005).
44. Ng, P. C. *et al*. Genetic variation in an individual human exome. *PLoS Genetics* **4**, e1000160, doi:10.1371/journal.pgen.1000160 (2008).
45. Gilad, Y., Man, O. & Glusman, G. A comparison of the human and chimpanzee olfactory receptor gene repertoires. *Genome Res* **15**, 224–230, doi:10.1101/gr.2846405 (2005).
46. Ribeiro-dos-Santos, A. M., da Silva, V. L., de Souza, J. E. & de Souza, S. J. Populational landscape of INDELs affecting transcription factor-binding sites in humans. *BMC Genomics* **16**, 536, doi:10.1186/s12864-015-1744-5 (2015).
47. Chen, F. C., Chen, C. J., Li, W. H. & Chuang, T. J. Human-specific insertions and deletions inferred from mammalian genome sequences. *Genome Res* **17**, 16–22, doi:10.1101/gr.5429606 (2007).
48. Song, W. Y. & Guo, J.-T. Investigation of arc repressor DNA-binding specificity by comparative molecular dynamics simulations. *Journal of biomolecular structure & dynamics* **33**, 2083–2093, doi:10.1080/07391102.2014.997797 (2015).
49. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003).
50. Dowell, R. D. Transcription factor binding variation in the evolution of gene regulation. *Trends Genet* **26**, 468–475, doi:10.1016/j.tig.2010.08.005 (2010).
51. Williams, R. B., Chan, E. K., Cowley, M. J. & Little, P. F. The influence of genetic variation on gene expression. *Genome Res* **17**, 1707–1716, doi:10.1101/gr.6981507 (2007).
52. Kasowski, M. *et al*. Variation in transcription factor binding among humans. *Science* **328**, 232–235, doi:10.1126/science.1183621 (2010).
53. McLaren, W. *et al*. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070, doi:10.1093/bioinformatics/btq330 (2010).
54. Wang, G. & Dunbrack, R. L. Jr. PISCES: a protein sequence culling server. *Bioinformatics* **19**, 1589–1591 (2003).

55. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
56. Wang, Z., Zhao, F., Peng, J. & Xu, J. Protein 8-class secondary structure prediction using conditional neural fields. *Proteomics* **11**, 3786–3792, doi:10.1002/pmic.201100196 (2011).
57. Vucetic, S. *et al*. DisProt: a database of protein disorder. *Bioinformatics* **21**, 137–140 (2005).
58. Dosztanyi, Z., Csizmok, V., Tompa, P. & Simon, I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* **21**, 3433–3434, doi:10.1093/bioinformatics/bti541 (2005).
59. Sickmeier, M. *et al*. DisProt: the Database of Disordered Proteins. *Nucleic Acids Res* **35**, D786–793, doi:10.1093/nar/gkl893 (2007).

Acknowledgements

This work was supported by the National Institutes of Health [R15GM110618 to J.G]; and National Science Foundation [DBI1356459 to J.G, DGE-1523154 and IIS-1502172 to X.S].

Author Contributions

J.T.G. conceived the study and designed the experiment. A.F. and X.S. provided the initial data retrieval and analysis. M.L., S.W. and J.C. carried out the experiments and performed data analysis. M.L., S.W. and J.T.G. wrote the manuscript. M.L., S.W., X.S., A.F. and J.T.G. revised the manuscript. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-09287-x

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.