Small insertions and deletions (INDELs) in human genomes

Julienne M. Mullaney1,2, Ryan E. Mills4, W. Stephen Pittard5 and Scott E. Devine1,2,3,*

1Institute for Genome Sciences, 2Division of Endocrinology, Diabetes, and Nutrition, Department of Medicine and 3Greenebaum Cancer Center, University of Maryland School of Medicine, Baltimore, MD 21201, USA, 4Department of Pathology, Brigham and Women's Hospital, Boston, MA 02115, USA and 5Bimcore, Emory University, Atlanta, GA 30322, USA

Received September 3, 2010; Revised and Accepted September 10, 2010

In this review, we focus on progress that has been made with detecting small insertions and deletions (INDELs) in human genomes. Over the past decade, several million small INDELs have been discovered in human populations and personal genomes. The amount of genetic variation that is caused by these small INDELs is substantial. The number of INDELs in human genomes is second only to the number of single nucleotide polymorphisms (SNPs), and, in terms of base pairs of variation, INDELs cause similar levels of variation as SNPs. Many of these INDELs map to functionally important sites within human genes, and thus, are likely to influence human traits and diseases. Therefore, small INDEL variation will play a prominent role in personalized medicine.

INTRODUCTION

Next generation sequencing technologies have unlocked a new era of human genome sequencing. Human genomes are being sequenced at unprecedented rates (1–9), and the age of personalized medicine is rapidly approaching. In the near future, an individual’s genome will be sequenced to predict the future health of the individual and to develop personalized medical treatments that are tailored to work with the genetic variation that is detected. Variation profiles also will be used to predict a host of other human traits including height, weight, appearance, lifespan and intelligence. Genome-wide association studies (GWASs) have already begun to identify the predictive variants that are necessary for this enterprise, and many additional variants will be identified. Ultimately, personal genomes will be compared with databases of informative variants, genes and biological pathways to make predictions about human traits and diseases.

To fulfill this vision, it will be necessary to routinely detect all of the genetic variation that is present in a given human genome. This includes: (i) single nucleotide polymorphisms [SNPs (10–12)], (ii) small insertions and deletions (INDELs) ranging from 1 to 10,000 bp in length (13–15) and (iii) larger forms of structural variation (16–22). In this review, we focus on progress that has been made with detecting small INDELs in human genomes. Despite the fact that small INDELs are highly abundant in humans and cause a great deal of variation in human genes, they have received far less attention than SNPs and larger forms of structural variation. Small INDELs have been particularly challenging to detect, validate and genotype, and customized tools have been necessary to study this class of variation. Moving forward, new technologies will be needed to evaluate the impact of these INDELs on human traits and diseases.

INDEL VARIATION ON HUMAN CHROMOSOME 22

The first large-scale efforts to identify INDELs in the human genome were focused on human chromosome 22 (23,24). Mullikin et al. (23) analyzed ABI re-sequencing data from 31 humans to identify genetic variation on chromosome 22 and determined that 13% of the variants on this chromosome were INDEL polymorphisms. Dawson et al. (24) later identified dense clusters of both SNPs and INDELs on chromosome 22 by comparing the overlapping sequences of adjacent bacterial artificial chromosome (BAC) clones and P1-derived artificial chromosome (P1) clones. The BAC and P1 libraries that

*To whom correspondence should be addressed at: Institute for Genome Sciences, University of Maryland School of Medicine, 801 W. Baltimore Street, 615 BioPark II, Baltimore, MD 21201, USA. Tel: +1 4107062343; Fax: +1 4107061482; Email: sdevine@som.umaryland.edu

© The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org
were generated to sequence chromosome 22 were constructed from nine diverse diploid individuals and harbored 18 distinct haplotypes. Therefore, new genetic variation could be readily discovered by comparing the overlapping sequences of adjacent clones. A total of 10 051 SNPs and 2180 small INDELs were identified from the 358 overlapping regions that were compared [13.2 Mb of overlapping DNA (24)]. These studies suggested that small INDELs were likely to be abundant in human genomes (82% of the variants that were detected were SNPs and 18% were INDELs). Ninety-two percent of these variants were confirmed in follow-up validation studies (Table 1).

**A GENOME-WIDE STUDY OF SMALL INDEL VARIATION IN HUMANS**

One of the earliest genome-wide INDEL discovery efforts in humans was conducted by Weber et al. (13). Similar to the Dawson et al. (24) study, most of the INDELs in this study were identified by comparing overlapping sequences from adjacent BAC clones that were sequenced for the human genome project. Three other methods also were used, and a total of 2000 INDELs were identified (Table 1). Most of these INDELs (96%) were 2–16 bp in length, and the largest INDEL was 55 bp in length. These INDELs were genotyped extensively in three panels of diverse humans to both validate the INDELs and to examine the allelic frequencies of these polymorphisms. The INDELs also were genotyped in chimpanzees to examine the ancestral alleles. These studies confirmed that INDEL variation is the second most abundant form of genetic variation in humans after SNPs (79% of the variants detected were SNPs and 21% of the variants were INDELs).

**INDEL DISCOVERY BY TARGETED RE-SEQUENCING OF HUMAN GENES AND ENCODE REGIONS**

Bhangale et al. (14) re-sequenced 330 genes in diverse humans to identify 2393 small INDELs ranging from 1 to 543 bp in length (Table 1). The genes that were targeted for re-sequencing were involved in several medically relevant biological pathways, including lipid metabolism and DNA repair. The authors discovered that heterozygous INDELs produce a characteristic signature on an ABI chromatogram in which the two superimposed INDEL alleles cause a transition from a high quality region at the beginning of the chromatogram to a poor quality region that abruptly begins at the INDEL. Bhangale et al. found that this was a highly reproducible signature for identifying heterozygous INDELs from polymerase chain reaction (PCR) amplicons. The authors developed a computational package called PolyPhred that automates INDEL detection and also detects SNPs. In a follow-up study, Bhangale et al. (25) used PolyPhred ver. 6.0 to identify INDELs in the ENCODE regions of the human genome, and identified 1126 additional small INDELs (Table 1). A similar INDEL detection pipeline called PolyScan has been developed by Chen et al. (26).

**LARGE-SCALE INDEL DISCOVERY PROJECTS IN HUMAN POPULATIONS**

Mills et al. (15) developed a custom variation discovery pipeline that detects SNPs, INDELs, and transposon insertions from ABI trace data. The pipeline was used to discover 415 436 unique INDEL variants from DNA re-sequencing traces that originally were generated for SNP discovery projects [including the HapMap project (Table 1) (10–12)]. In a follow-up study, almost two million non-redundant INDELs were identified from DNA re-sequencing data (R.E. Mills et al., submitted for publication). The INDELs that were discovered ranged from 1 to 10 000 bp in length, and were distributed throughout the human genome on all 24 chromosomes. Several classes of INDELs were identified, including repeat expansions, transposon insertions and random sequences. PCR-based validation studies indicated that the Mills et al. (15) pipeline has an accuracy of 97%.

Kidd et al. (22) also reported 796 273 small INDELs in a project that was designed to detect structural variation in eight diverse humans (Table 1). In addition to the structural variants that were detected in fosmids by paired-end sequencing, SNPs and small INDELs also were discovered from the ABI traces that were generated. A comparison of the Mills et al.’s INDELs [dbSNP handle: DEVINE_LAB (15, R.E. Mills et al., submitted for publication)] with those discovered by Kidd et al. [dbSNP handle: HSVAR (22)] revealed minimal overlap between the two collections (only 10% of the DEVINE_LAB INDELs were present in the HSVAR collection). This low level of overlap suggests that small INDEL discovery is far from complete in human genomes. The 1000 Genomes Project is likely to close this gap by discovering a very large number of INDELs in the next few years (9).

**INDEL DISCOVERY IN PERSONAL GENOMES**

Small INDELs also have been discovered in most of the personal human genomes that have been sequenced thus far (Table 1) (1–8). Both the number of INDELs reported and the range of INDEL sizes that have been detected vary greatly among these genomes (Table 1). For example, 823 396 INDELs were discovered in the Venter genome (1), 222 718 INDELs were found in the Watson genome (2), and 135 262 INDELs were discovered in the Han Chinese genome (3). It seems unlikely that INDEL variation is genuinely this different among three humans. The INDELs discovered in the Venter genome ranged from 1 to 82 711 bp in length (1), whereas those discovered in the Han Chinese genome were limited to 1–3 bp in length (3). Thus, some of these differences appear to be due to differences in INDEL discovery among these projects. The sequencing platforms and analysis approaches that were employed also were quite variable. For example, the Venter genome was sequenced with traditional ABI reads (1), the Watson genome was sequenced with 454 reads (2) and the Han Chinese genome was sequenced with Illumina reads (3). The validation rates for these approaches also varied considerably. In most cases, very few INDELs were evaluated in validation studies, and in many cases, only a small subset was examined. For example, although 342 965 small INDELs were reported in
the Korean genome [SJK (7)], only nine of these INDELs were validated by PCR and sequencing (for a validation rate of 9/9 or 100%, Table 1) (7). For the Yoruban genome, no INDEL validation was performed (Table 1) (4).

Several other approaches also have been developed (29–32), and most of these approaches have INDEL detection capabilities (Table 2). Although all of these approaches initially attempt to map either the whole trace or a segment of the trace to the reference genome, accurate mapping is often limited by the presence of INDELS [particularly if they are larger than 1–3 bp (27–32)]. In some cases, sequence reads that initially do not map well to the genome are re-mapped and aligned using programs that can support gapped or split-read alignments (30). Paired-end reads also are used to detect INDELS: one read serves to map the pair of reads, whereas the second read is subjected to gapped alignments and INDEL detection (28–30). In this instance, the first read is not allowed to vary as much as the second read, and variation is undoubtedly missed when both reads contain INDELS. In other cases, all traces are mapped and aligned to the genome using algorithms that perform gapped alignments (29,32). Although simulation studies suggest that several of these methods have high levels of sensitivity, the false negative rates for these approaches range from 10 to 35% (Table 2). Thus, up to a third of the small INDELS in personal genomes are being missed with these approaches.

**FUNCTIONALLY IMPORTANT INDELS**

Like SNPs and structural variation, INDELS are of great interest because they can alter human traits and can cause human diseases. One of the most common genetic diseases in humans, cystic fibrosis, is frequently caused by a coding INDEL polymorphism within the *CFTR* gene that eliminates a single amino acid (33). The Mills et al. (15) study described above revealed that coding INDELS are prevalent in human populations. A total of 262 coding INDELS were identified in the 36 healthy humans that were examined. More than half of these coding INDELS (160/262 or 61.1%) were multiples of 3 bp, and thus, maintained the open reading frames in the 36 healthy humans that were examined. More than half of these coding INDELS (160/262 or 61.1%) were multiples of 3 bp, and thus, maintained the open reading frames in the 36 healthy humans that were examined. More than half of these coding INDELS (160/262 or 61.1%) were multiples of 3 bp, and thus, maintained the open reading frames in the 36 healthy humans that were examined. More than half of these coding INDELS (160/262 or 61.1%) were multiples of 3 bp, and thus, maintained the open reading frames in the 36 healthy humans that were examined. More than half of these coding INDELS (160/262 or 61.1%) were multiples of 3 bp, and thus, maintained the open reading frames in the 36 healthy humans that were examined. More than half of these coding INDELS (160/262 or 61.1%) were multiples of 3 bp, and thus, maintained the open reading frames in the 36 healthy humans that were examined.

<table>
<thead>
<tr>
<th>Study</th>
<th>Year</th>
<th>No. of INDELs</th>
<th>Individual(s)</th>
<th>Method</th>
<th>INDEL size range (bp)</th>
<th>Validation study</th>
<th>Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human populations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mullikin et al. (23)</td>
<td>2000</td>
<td>NR</td>
<td>31</td>
<td>ABI trace mapping</td>
<td>1–82 711</td>
<td>PCR</td>
<td>84–100</td>
</tr>
<tr>
<td>Dawson et al. (24)</td>
<td>2001</td>
<td>2180</td>
<td>9</td>
<td>BAC overlap</td>
<td>NR</td>
<td>PCR/RFLP/invader</td>
<td>92</td>
</tr>
<tr>
<td>Weber et al. (13)</td>
<td>2002</td>
<td>2000</td>
<td>NR</td>
<td>Various</td>
<td>2–55</td>
<td>PCR</td>
<td>58</td>
</tr>
<tr>
<td>Bhangale et al. (14)</td>
<td>2005</td>
<td>2393</td>
<td>137</td>
<td>PCR/sequencing</td>
<td>1–543</td>
<td>ND</td>
<td>N/A</td>
</tr>
<tr>
<td>Bhangale et al. (25)</td>
<td>2006</td>
<td>1126</td>
<td>ENCODE</td>
<td>PCR/sequencing</td>
<td>1 to 30</td>
<td>Manual inspection</td>
<td>100</td>
</tr>
<tr>
<td>Mills et al. (15)</td>
<td>2006</td>
<td>415 436</td>
<td>36</td>
<td>ABI trace mapping</td>
<td>1–9989</td>
<td>PCR</td>
<td>97</td>
</tr>
<tr>
<td>Kidd et al. (22)</td>
<td>2008</td>
<td>796 273</td>
<td>8</td>
<td>ABI trace mapping</td>
<td>1–100</td>
<td>ND</td>
<td>96</td>
</tr>
<tr>
<td>R.E. Mills et al. (submitted for publication)</td>
<td>2010</td>
<td>1.96 million</td>
<td>79</td>
<td>ABI trace mapping</td>
<td>1–10 000</td>
<td>PCR</td>
<td>97</td>
</tr>
<tr>
<td>Personal human genomes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Levy et al. (1)</td>
<td>2007</td>
<td>823 396</td>
<td>Venter</td>
<td>ABI</td>
<td>1</td>
<td>PCR</td>
<td>84–100</td>
</tr>
<tr>
<td>Wheeler et al. (2)</td>
<td>2008</td>
<td>222 718</td>
<td>Watson</td>
<td>454</td>
<td>2–38 896</td>
<td>PCR</td>
<td>70</td>
</tr>
<tr>
<td>Wang et al. (3)</td>
<td>2008</td>
<td>133 262</td>
<td>Han Chinese</td>
<td>Illumina/SOAP</td>
<td>1–3</td>
<td>PCR</td>
<td>90–100</td>
</tr>
<tr>
<td>Bentley et al. (4)</td>
<td>2008</td>
<td>400 000</td>
<td>Yoruban</td>
<td>Illumina/ELAND/MAQ</td>
<td>1–16</td>
<td>ND</td>
<td>N/A</td>
</tr>
<tr>
<td>Levy et al. (5)</td>
<td>2008</td>
<td>726</td>
<td>AML</td>
<td>Illumina/Custom</td>
<td>1–30</td>
<td>PCR</td>
<td>93</td>
</tr>
<tr>
<td>Ahn et al. (7)</td>
<td>2009</td>
<td>342 965</td>
<td>Korean (SJ)</td>
<td>Illumina/MAQ</td>
<td>1–26</td>
<td>PCR</td>
<td>100</td>
</tr>
<tr>
<td>Kim et al. (6)</td>
<td>2009</td>
<td>170 202</td>
<td>Korean (AK1)</td>
<td>Illumina/Alpheus</td>
<td>1–29</td>
<td>Re-sequencing</td>
<td>100</td>
</tr>
<tr>
<td>Schuster et al. (8)</td>
<td>2010</td>
<td>NR</td>
<td>African</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

NR, not reported; ND, not done; N/A, not applicable.

*Defined as a study examining a randomly selected set of INDEL variants predicted in real humans and then measuring the confirmation rate of the predicted variants using an assay such as PCR-RFLP, PCR sequencing, microarray genotyping or similar assay.

13% of the variants were INDELS.

BAC/PAC libraries (each representing a human).

SSAHA-DIP, which was used to detect the INDELS, has a validation rate of 96% (46).

Number of humans or libraries (each library representing a human).

The pipeline that was developed and tested above (15) was used in this study.

64/34 (84%) of 100–1000 bp homozygous insertions and 15/15 (100%) of 1–8 bp INDELs examined were confirmed.

Coding INDELS only.

Reported as ‘0.4 million’ (4).

24/26 (93%) of the candidates that were examined successfully were validated.

9/9 (100%) of the INDELS examined were validated.
abolish gene function. Although the 3 bp in-frame \textit{CFTR} coding INDEL abolishes gene function and produces cystic fibrosis, other in-frame INDELS might have less severe effects on protein function.

Coding INDELS subsequently have been found in several of the personal genome sequences that have been examined. For example, 739 coding INDELS were identified in the Venter genome (1,34) and 345 coding INDELS were identified in the Watson genome (2). Although it is not clear whether all of the personal genomes that have been sequenced thus far harbor coding INDELS, it does not appear that attempts were made to identify coding INDELS in all cases (Table 1). Exome re-sequencing projects likewise have identified a large number of coding INDELS in typical humans (35), suggesting that most, if not all humans, carry a substantial genetic load of coding INDELS. Thus, coding INDELS are likely to have a major impact on human biology and diseases.

INDELS that occur in other functionally important regions of genes also could be envisioned to affect gene function. For example, DNA insertions within the promoter region of the \textit{FMR1} gene cause Fragile X syndrome in humans (36). Once the trinucleotide expansion reaches a critical size threshold, the methylation patterns of the promoter are altered, leading to changes in \textit{FMR1} gene expression. Mills et al. (15) identified 3829 small INDELS that map to the promoter regions of RefSeq genes. In addition to repeat expansions as outlined for \textit{FMR1}, INDELS that occur within transcription factor binding sites or enhancers also could be envisioned to diminish or abolish gene expression. INDELS also have the capacity to dramatically alter the phasing and spacing of DNA sequences within promoters. For example, a 5 bp INDEL could rotate a transcription factor binding site to the opposite face of the DNA helix, and a 100 bp INDEL could increase the spacing between two binding sites such that productive interactions are disrupted. Thus, INDELS within gene promoter regions might help explain the differences in gene expression that have been observed in diverse humans (37).

### Mobile Genetic Elements

Mobile genetic elements produce a great deal of small INDEL variation in humans (38,39). In particular, active \textit{Alu}, \textit{L1} and SVA retrotransposons continue to produce new insertions when they ‘jump’ from one place to another in the human genome (38,39). When they jump, \textit{Alu} elements cause new insertions of \(\sim 300\) bp, whereas \textit{L1} and SVA elements cause insertions that range from tens of base pairs in length up to \(3\) kb (SVA) or \(6\) kb (L1) in length (38). Recent studies indicate that \textit{Alu} and \textit{L1} elements produce millions of new insertions worldwide in human populations (40–45). New insertions, by definition, cause rare insertion alleles because they initially occur in only a single human. Many of these new insertions occur in human genes (including the exons of these genes) and such insertions can affect gene function (38,40). The full-length source elements that produce new \textit{L1} insertions are themselves highly variable among diverse humans, suggesting that some genomes might produce more new \textit{L1} insertions than others (41). Nevertheless, endogenous retrotransposons appear to cause rare insertion alleles in most, if not all, human genomes and these insertions can influence human biology (38,40).

### The Future of INDEL Discovery in Personal Human Genomes

Moving forward, there are several major questions. First, exactly how many small INDELS are present in the average human genome? And how does this number differ from one human to the next? The numbers that have been reported thus far are quite variable (Table 1). It appears that variability in INDEL discovery is at the root of these differences rather than true differences in the number of INDELS. Genuine differences between two individuals might be of great interest. Do some genomes actually carry fewer INDELS than others? To answer these questions, it will be necessary to define a standardized size range for INDEL discovery that will be applied to all genomes (1 to 100 bp? 1 to 10,000 bp?). In this regard, it might be useful to compare two or more of the genomes that have already been sequenced (Table 1) using a single uniform approach. Another major issue is INDEL validation. Most of the validation studies that have been conducted thus far have sampled only a very limited number of INDELS, and as a consequence, the true rates of INDEL discovery are hard to estimate. The false negative rates that have been reported (Table 2) indicate that many of the INDELS in personal genomes are being missed with current detection methods. What are these missing INDELS and can we afford to overlook them? Can the detection methods be modified to detect more of these INDELS?

And the biggest question of all: which small INDELS affect human traits and diseases and how will we find these INDELS? Several experimental approaches could be taken to answer this question. First, the small INDELS that have been discovered thus far (Table 1; totaling several million in dbSNP) could be genotyped and integrated into the human HapMap in the same way that SNPs were genotyped for the HapMap (10–12). Common INDELS that map to exons, promoters and other functionally important sites could be given the

---

**Table 2.** Computational approaches for Next generation trace mapping and INDEL detection

<table>
<thead>
<tr>
<th>Program</th>
<th>Reference</th>
<th>INDEL discovery (%)</th>
<th>Validation Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOAP</td>
<td>Li et al. (27)</td>
<td>NR</td>
<td>PCR&lt;sup&gt;a&lt;/sup&gt; 90–100&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>MAQ</td>
<td>Li et al. (28)</td>
<td>&lt;2</td>
<td>PCR&lt;sup&gt;a&lt;/sup&gt; 100&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>BWA</td>
<td>Li and Durbin (29)</td>
<td>NR</td>
<td>ND</td>
</tr>
<tr>
<td>Pindel</td>
<td>Ye et al. (30)</td>
<td>2</td>
<td>20–31.4</td>
</tr>
<tr>
<td>Bowtie</td>
<td>Langmead et al. (31)</td>
<td>NR</td>
<td>ND</td>
</tr>
<tr>
<td>BFAST</td>
<td>Homer et al. (32)</td>
<td>&lt;3</td>
<td>10–20</td>
</tr>
</tbody>
</table>

NR, not reported; ND, not done; N/A, not applicable.

<sup>a</sup>Defined as a study examining a randomly selected set of INDEL variants predicted in real humans and then measuring the confirmation rate of the predicted variants using an assay such as PCR-RFLP, PCR sequencing, microarray genotyping or similar assay.

<sup>b</sup>From the Han Chinese genome sequencing project (3).

<sup>c</sup>From the Korean Genome Project (7), nine coding INDELS were examined.
highest priority. Once integrated into the HapMap, these INDELs would greatly expand the number of high-scoring variants that are identified in GWASs. As efforts turn to whole genome re-sequencing as an approach for identifying causative variants, it will become even more critical to detect INDELs fully and accurately. The biggest and perhaps most interesting challenge ahead will be to develop a comprehensive framework of variants, genes and biological pathways that can be used to predict human health.

ACKNOWLEDGEMENTS

We thank Shari Corin for critical review of the manuscript and for helpful discussions. We also thank Anup Mahurkar, Todd Creasy and Umar Farooq for helpful discussions.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by the National Human Genome Research Institute, National Institutes of Health (grant number R01HG002898 to S.E.D.).

REFERENCES


