Accuracy of Next Generation Sequencing Platforms

Edward J Fox1, Kate S Reid-Bayliss1, Mary J Emond2 and Lawrence A Loeb1*  
1Departments of Pathology and Biochemistry, University of Washington, USA  
2Department of Biostatistics, University of Washington, USA  
*Corresponding author: Lawrence A Loeb, Departments of Pathology and Biochemistry, University of Washington, USA, Tel: 1-206-543-0556; Fax: 1-206-543-3967; E-mail: eddiefox@uw.edu

Abstract

Next-generation DNA sequencing has revolutionized genomic studies and is driving the implementation of precision diagnostics. The ability of these technologies to disentangle sequence heterogeneity, however, is limited by their relatively high error rates. A Several single molecule barcoding strategies have been propose to reduce the overall error frequency. A Duplex Sequencing additionally exploits the fact that DNA is double-strand, with one strand reciprocally encoding the sequence information of its complement, and can eliminate nearly all sequencing errors by comparing the sequence of individually tagged amplicons derived from one strand of DNA with that of its complementary strand. This method reduces errors to fewer than one per ten million nucleotides sequenced.

Keywords: Next-generation DNA sequencing; Precision medicine; Accuracy; Duplex sequencing

Introduction

Mutation drives evolution and underlies many diseases, most prominently cancer [1]. Of the newly developed genomic technologies, next-generation DNA sequencing (NGS), in particular, has revolutionized the scale of study of biological systems [2] and has already started to enter the clinic where it is expected to enable a more personalized approach to patient care [3].Unlike conventional sequencing techniques, which simply report the average genotype of an aggregate of molecules, NGS digitally tabulates the sequence of individual DNA fragments, thereby offering the unique ability to detect minor variants within heterogeneous mixtures [4]. Already, NGS has been used to characterize exceptional diversity within microbial [5,6], viral [7-9], and tumor cell populations [10-12], and many low frequency, drug-resistant variants of therapeutic importance have been identified [13,14]. NGS has also revealed previously underappreciated intra-organismal mosaicism in both the nuclear [15] and mitochondrial genomes [16]. This somatic heterogeneity, along with that underlying adaptive immunity [17], is an important factor in determining the phenotypic variability of disease.

In theory, DNA subpopulations of any size should be detectable via ‘deep sequencing’ of a sufficient number of molecules. However, a fundamental limitation of standard NGS is the high frequency with which bases are scored incorrectly due to artifacts introduced during sample preparation and sequencing [18]. For example, amplification bias during PCR of heterogeneous mixtures can result in skewed populations [19]. Additionally, polymerase mistakes, such as base misincorporations and rearrangements due to template switching, can result in incorrect variant calls. Furthermore, errors arise during cluster amplification, sequencing cycles, and image analysis result in approximately 0.1–1% of bases being called incorrectly (Table 1).

<table>
<thead>
<tr>
<th>Commercial Platform</th>
<th>Most Frequent Error Type</th>
<th>Error Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capillary sequencing</td>
<td>single nucleotide substitutions</td>
<td>10^{-5}</td>
</tr>
<tr>
<td>454 GS Junior</td>
<td>Deletions</td>
<td>10^{-2}</td>
</tr>
<tr>
<td>PacBio RS</td>
<td>CG deletions</td>
<td>10^{-2}</td>
</tr>
<tr>
<td>Ion Torrent PGM</td>
<td>Short deletions</td>
<td>10^{-2}</td>
</tr>
<tr>
<td>Solid</td>
<td>A-T bias</td>
<td>2×10^{-2}</td>
</tr>
<tr>
<td>Illumina MiSeq</td>
<td>single nucleotide substitutions</td>
<td>10^{-3}</td>
</tr>
<tr>
<td>Illumina HiSeq2000</td>
<td>single nucleotide substitutions</td>
<td>10^{-3}</td>
</tr>
<tr>
<td>Tag-based methods</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SafeSeq</td>
<td>single nucleotide substitutions</td>
<td>1.4×10^{-5}</td>
</tr>
<tr>
<td>CircleSeq</td>
<td>single nucleotide substitutions</td>
<td>7.6×10^{-6}</td>
</tr>
<tr>
<td>Duplex Sequencing</td>
<td>Single nucleotide substitutions</td>
<td>5×10^{-4}</td>
</tr>
</tbody>
</table>

Table 1: Comparison of the primary error frequencies of DNA sequencing platforms and tag-based error correction methodologies

For a genetically homogenous sample, the effects of these base miscalls can be mitigated by establishing a consensus sequence from high-coverage sequencing reads.

However, when rare genetic variants are sought, this base call error frequency presents a profound barrier and has limited the use of deep sequencing in a variety fields that require the highly accurate disentangling of subpopulations within complex (heterogeneous or mixed) biological samples, including metagenomics [20,21], forensics [22], paleogenomics [23] and human genetics [4,24]. Furthermore, for many applications, such as the prenatal screening for fetal aneuploidy [25,26], detection of circulating tumor DNA [27], and monitoring response to chemotherapy with nucleic acid-based serum biomarkers [28], a level of detection well below 1 in 10,000 is highly desirable;
Unfortunately, the high frequency of erroneous base calls inherent to standard NGS imposes a practical limit of detection of approximately 1 in 100. These technical shortcomings have also limited the elucidation of mechanism by which genomes, and DNA itself, have evolved [29-31], where bioinformatics analyses have been used to reconstruct phylogenetic relationships [32-35].

### Figure 1: Comparison of the probability that an observed variant is real [54] for subclonal variants using Q30-filtered reads of an Illumina HiSeq2500 (NGS) versus Duplex Sequencing Error Frequencies of each approach is given in parenthesis. PPV (Positive Predictive Value) = (Expected Number of true positives)/(Expected Total Number of Positive Calls). Note that the PPV is 0.50 for NGS when the variant frequency at a single position is ~1/100, i.e., any variant call has a 50/50 chance of being real hen the frequency of real variants equals the frequency of mistakes invalidity [62].

Although biochemical protocols [36-39] and bioinformatics [10,40-43] have improved sequencing accuracy, the ability to confidently resolve subpopulations below 1% has remained problematic [44]. Laird and colleagues demonstrated that it was possible to significantly reduce the frequency of variant miscalls by covalently linking individual DNA molecules to unique tags prior to amplification [45,46]. This ‘barcoding’ technique allows many artifactual variations in the sequence to be identified as due to technical error [47-52], as all amplicons derived from a particular individual starting molecule carry the same unique specific tag and, thus, can be collapsed to a consensus sequence representing that of the original DNA strand. An alternative to single-stranded tagging based on shear-points is the circle sequencing methodology developed by Schmitt et al., which utilizes the strand-displacement activity of Phi29’s DNA polymerase to generate multiple copies of circularized DNA molecules in tandem prior to amplification [53]. After sequencing, these linked copies are collapsed to a consensus sequence, thereby eliminating many artifactual errors. Though significant improvements, single-strand approaches all (Table 1) still exhibit error frequencies greater than the estimated frequency of variation of many biological systems. The mutation rate of normal cells, for example, is estimated to range from $10^{-9}$ to $10^{-11}$ mutations/per nucleotide/per cell division [54,55].

Schmitt et al., highlighted a conceptual shortcoming of initial tag-based methods, and of next-generation sequencing platforms in general, in that use is made of sequence data derived from a single strand of DNA [56]. As a consequence, artifactual variants introduced during the initial rounds of PCR amplification become fixed and are indistinguishable from true variants, since the sequence information of the complementary strand is not taken into account. Damage to DNA from oxidative cellular processes, or generated ex vivo during tissue processing and DNA extraction [57,58], is a particular concern, as such damage can result in frequent copying errors by DNA polymerases. For example, the most thoroughly studied DNA lesion arising from oxidative damage, 8-oxoguanine, incorrectly pairs with adenine during copying with an overall efficiency greater than that of correct pairing with cytosine, and can, thus, contribute a large frequency of artifactual G:C→T:A mutations [59]. Similarly, deamination of cytosine to form uracil is a common event, which leads to inappropriate pairing with adenine during polymerase extension, thus producing artifactual C:G→T:A mutations, at a frequency approaching 100% [60]. Significantly, DNA damage and the resulting sequencing artifacts occur in strand-specific patterns.

Schmitt et al. recognized that these types of errors could be resolved by exploiting the fact that DNA naturally exists as a double-stranded entity, with one molecule reciprocally encoding the sequence information of its complement. Using this insight and the arising sequencing methodology, termed Duplex Sequencing, Schmitt et al., demonstrated that it is possible to identify and eliminate nearly all sequencing errors by comparing the sequence of individually tagged amplicons derived from one strand of DNA with that of its complementary strand; a base sequenced at a given position is scored only if the read data from each of the two strands match perfectly. The method has a theoretical background error rate of less than one artifactual error per 109 nucleotides and has been used to detect variants at a frequency of $5 \times 10^{-8}$.

In principle, Duplex Sequencing can be used with any NGS platform and can call sequence variants when present in an excess of 10 million wild-type sequences [53,56,61]. In contrast, with an error rate of approximately $10^{-2}$, the probability of accurately distinguishing a true subclonal variant from a sequencing artifact in an excess of 100 wild-type molecules with NGS is approximately 50%, using standard (Q30)–filtered reads (Figure 1). A real variant at or below these frequencies cannot be resolved by increasing sequencing depth at a single position, as the proportion of errors will not change. Duplex Sequencing, thus, offers an improvement of nearly 5-orders of magnitude over standard Q30-filtered sequencing and 3-orders of magnitude over other tag-based methods. Thus by exploiting the redundant sequence information contained in the complementary strand of a double-stranded DNA molecule, Duplex Sequencing has dramatically increased the precision and power of NGS. Its application will likely improve our understanding of the substructure of biological systems, including human cancers, help to pinpoint mechanisms of mutation generation, modify the catalog of rare variants, dramatically improve our ability to accurately deconvolute complex biological admixtures, and offer the diagnostic accuracy required for the implementation of precision medicine.

### References


