Journal of Next Generation Sequencing & Applications

Fox et al., Next Generat Sequenc & Applic 2014, 1:1
DOI: 10.4172/2469-9853.1000106

Editorial Open Access

Accuracy of Next Generation Sequencing Platforms

Edward J Fox[1], **Kate S Reid-Bayliss**[1], **Mary J Emond**[2] and **Lawrence A Loeb**[1*]

[1]*Departments of Pathology and Biochemistry, University of Washington, USA*

[2]*Department 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

Rec Date: Apr 30, 2014, **Acc date:** Jun 26, 2014, **Pub date:** Jun 28, 2014

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 arising during cluster amplification, sequencing cycles, and image analysis result in approximately 0.1–1% of bases being called incorrectly (Table 1).

Commercial Platform	Most Frequent Error Type	Error Frequency
Capillary sequencing	single nucleotide substitutions	10-1
454 GS Junior	Deletions	10-2
PacBio RS	CG deletions	10-2
Ion Torrent PGM	Short deletions	10-2
Solid	A-T bias	2×10-2
Illumina MiSeq	single nucleotide substitutions	10-3
Illumina HiSeq2000	single nucleotide substitutions	10-3
Tag-based methods:		
SafeSeq	single nucleotide substitutions	1.4×10-5
CircleSeq	single nucleotide substitutions	7.6×10-6
Duplex Sequencing	Single nucleotide substitutions	5×10 -8

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].

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 can, thus, 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 Lou 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, these 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 5×10^{-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.

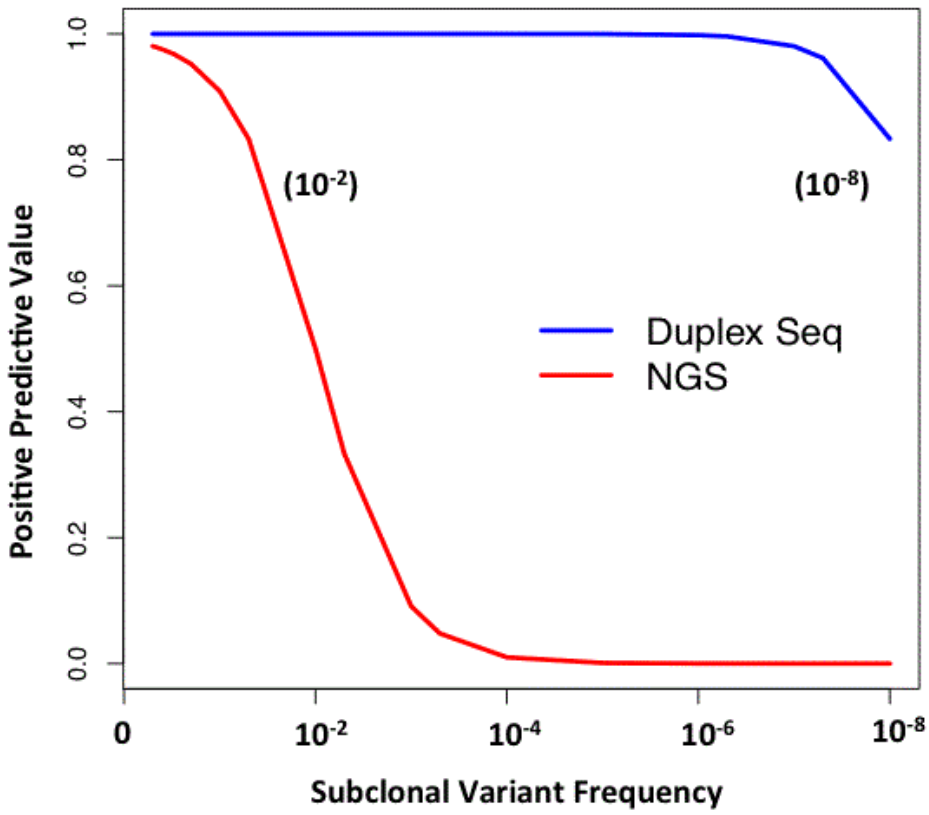

Figure 1: Comparison of the probability that an observed variant is real [54] for subclonal variants using Q30-filtered reads of an Illumina HiSeq2500 (NGS) versus Duplex Sequencing. Error Frequencies of each approach is given in parenthesis. PPV (Positive Predictive Value)=(Expected Number of true positives)/(Expected 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].

References

1. Loeb LA (2011) Human cancers express mutator phenotypes: origin, consequences and targeting. Nat Rev Cancer 11: 450-457.
2. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26: 1135-1145.
3. Schwartz WB, Wolfe HJ, Pauker SG (1981) Pathology and probabilities: a new approach to interpreting and reporting biopsies. N Engl J Med 305: 917-923.
4. Druley TE, Vallania FL, Wegner DJ, Varley KE, Knowles OL, et al. (2009) Quantification of rare allelic variants from pooled genomic DNA. Nat Methods 6: 263-265.
5. LaTuga MS, Ellis JC, Cotton CM, Goldberg RN, Wynn JL, et al. (2011) Beyond bacteria: a study of the enteric microbial consortium in extremely low birth weight infants. PLoS One 6: e27858.
6. Hyman RW, Herndon CN, Jiang H, Palm C, Fukushima M, et al. (2012) The dynamics of the vaginal microbiome during infertility therapy with in vitro fertilization-embryo transfer. J Assist Reprod Genet 29: 105-115.
7. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, et al. (2011) The human gut virome: inter-individual variation and dynamic response to diet. Genome Res 21: 1616-1625.
8. Nasu A, Marusawa H, Ueda Y, Nishijima N, Takahashi K, et al. (2011) Genetic heterogeneity of hepatitis C virus in association with antiviral therapy determined by ultra-deep sequencing. PLoS One 6: e24907.
9. Yang J, Yang F, Ren L, Xiong Z, Wu Z, et al. (2011) Unbiased parallel detection of viral pathogens in clinical samples by use of a metagenomic approach. J Clin Microbiol 49: 3463-3469.
10. Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, et al. (2008) Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc Natl Acad Sci U S A 105: 13081-13086.
11. Landau DA, Carter SL, Stojanov P, McKenna A, Stevenson K, et al. (2013) Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. Cell 152: 714-726.
12. Costello M, Pugh TJ, Fennell TJ, Stewart C, Lichtenstein L, et al. (2013) Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res 41: e67.
13. Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, et al. (2012) Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature 481: 506-510.

14. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW (2007) Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. Genome Res 17: 1195-1201.
15. Carlson CA, Kas A, Kirkwood R, Hays LE, Preston BD, et al. (2011) Decoding cell lineage from acquired mutations using arbitrary deep sequencing. Nat Methods 9: 78-80.
16. Ameur A, Stewart JB, Freyer C, Hagström E, Ingman M, et al. (2011) Ultra-deep sequencing of mouse mitochondrial DNA: mutational patterns and their origins. PLoS Genet 7: e1002028.
17. Boyd SD, Marshall EL, Merker JD, Maniar JM, Zhang LN, et al. (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. Sci Transl Med 1: 12ra23.
18. Glenn TC (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour 11: 759-769.
19. Kanagawa T (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). J Biosci Bioeng 96: 317-323.
20. Lecroq B, Lejzerowicz F, Bachar D, Christen R, Esling P, et al. (2011) Ultra-deep sequencing of foraminiferal microbarcodes unveils hidden richness of early monothalamous lineages in deep-sea sediments. Proc Natl Acad Sci U S A 108: 13177-13182.
21. Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, et al. (2011) Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature 480: 368-371.
22. Tillmar AO, Dell'Amico B, Welander J, Holmlund G (2013) A universal method for species identification of mammals utilizing next generation sequencing for the analysis of DNA mixtures. PLoS One 8: e83761.
23. Schubert M, Ermini L, Der Sarkissian C, Jónsson H, Ginolhac A, et al. (2014) Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat Protoc 9: 1056-1082.
24. Out AA, van Minderhout IJ, Goeman JJ, Ariyurek Y, Ossowski S, et al. (2009) Deep sequencing to reveal new variants in pooled DNA samples. Hum Mutat 30: 1703-1712.
25. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR (2008) Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci U S A 105: 16266-16271.
26. Chiu RW, Akolekar R, Zheng YW, Leung TY, Sun H, et al. (2011) Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study. BMJ 342: c7401.
27. Beck J, Urnovitz HB, Mitchell WM, Schutz E (2010) Next generation sequencing of serum circulating nucleic acids from patients with invasive ductal breast cancer reveals differences to healthy and nonmalignant controls. Mol Cancer Res 8: 335-342.
28. Leary RJ, Kinde I, Diehl F, Schmidt K, Clouser C, et al. (2010) Development of personalized tumor biomarkers using massively parallel sequencing. Sci Transl Med 2: 20ra14.
29. Di Mauro E, Saladino R, Trifonov EN (2014) The path to life's origins. Remaining hurdles. J Biomol Struct Dyn 32: 512-522.
30. Frenkel ZM, Trifonov EN (2012) Origin and evolution of genes and genomes. Crucial role of triplet expansions. J Biomol Struct Dyn 30: 201-210.
31. Sobolevsky Y, Guimarães RC, Trifonov EN (2013) Towards functional repertoire of the earliest proteins. J Biomol Struct Dyn 31: 1293-1300.
32. Gerhardt GJ, Takeda AA, Andrighetti T, Sartor IT, Echeverrigaray SL, et al. (2013) Triplet entropy analysis of hemagglutinin and neuraminidase sequences measures influenza virus phylodynamics. Gene 528: 277-281.
33. Yang Y, Zhang Y, Jia M, Li C, Meng L (2013) Non-degenerate graphical representation of DNA sequences and its applications to phylogenetic analysis. Comb Chem High Throughput Screen 16: 585-589.
34. Huang T, Zhang J, Xu ZP, Hu LL, Chen L, et al. (2012) Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches. Biochimie 94: 1017-1025.
35. Wu G, Yan S (2008) Prediction of mutations engineered by randomness in H5N1 neuraminidases from influenza A virus. Amino acids 34: 81-90.
36. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, et al. (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat Methods 6: 291-295.
37. Vandenbroucke I, Van Marck H, Verhasselt P, Thys K, Mostmans W, et al. (2011) Minor variant detection in amplicons using 454 massive parallel pyrosequencing: experiences and considerations for successful applications. Biotechniques 51: 167-177.
38. Vandenbroucke I, Eygen VV, Rondelez E, Vermeiren H, Baelen KV, et al. (2008) Minor Variant Detection at Different Template Concentrations in HIV-1 Phenotypic and Genotypic Tropism Testing. Open Virol J 2: 8-14.
39. Flaherty P, Natsoulis G, Muralidharan O, Winters M, Buenrostro J, et al. (2012) Ultrasensitive detection of rare mutations using next-generation targeted resequencing. Nucleic Acids Res 40: e2.
40. Muralidharan O, Natsoulis G, Bell J, Newburger D, Xu H, et al. (2012) A cross-sample statistical model for SNP detection in short-read sequencing data. Nucleic Acids Res 40: e5.
41. Zagordi O, Klein R, Däumer M, Beerenwinkel N (2010) Error correction of next-generation sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Res 38: 7400-7409.
42. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, et al. (2008) A large genome center's improvements to the Illumina sequencing system. Nat Methods 5: 1005-1010.
43. Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, et al. (2010) A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res 20: 273-280.
44. Klco JM, Spencer DH, Miller CA, Griffith M, Lamprecht TL, et al. (2014) Functional heterogeneity of genetically defined subclones in acute myeloid leukemia. Cancer Cell 25: 379-392.
45. Miner BE, Stöger RJ, Burden AF, Laird CD, Hansen RS (2004) Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. Nucleic Acids Res 32: e135.
46. McCloskey M, Stöger R, Hansen RS, Laird CD (2007) Encoding PCR products with batch-stamps and barcodes. Biochem Genet 45: 761-767.
47. Casbon JA, Osborne RJ, Brenner S, Lichtenstein CP (2011) A method for counting PCR template molecules with application to next-generation sequencing. Nucleic Acids Res 39: e81.
48. Jabara CB, Jones CD, Roach J, Anderson JA, Swanstrom R (2011) Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc Natl Acad Sci U S A 108: 20166-20171.
49. Kivioja T, Vähärautio A, Karlsson K, Bonke M, Enge M, et al. (2011) Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods 9: 72-74.
50. Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. Proc Natl Acad Sci U S A 108: 9530-9535.
51. Shiroguchi K, Jia TZ, Sims PA, Xie XS (2012) Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. Proc Natl Acad Sci U S A 109: 1347-1352.
52. Fu GK, Hu J, Wang PH, Fodor SP (2011) Counting individual DNA molecules by the stochastic attachment of diverse labels. Proc Natl Acad Sci U S A 108: 9026-9031.
53. Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, et al. (2013) High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. Proc Natl Acad Sci U S A 110: 19872-19877.
54. Jackson AL, Loeb LA (1998) The mutation rate and cancer. Genetics 148: 1483-1490.
55. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, et al. (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328: 636-639.
56. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, et al. (2012) Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A 109: 14508-14513.

57. Lindahl T, Wood RD (1999) Quality control by DNA repair. Science 286: 1897-1905.
58. Preston BD, Albertson TM, Herr AJ (2010) DNA replication fidelity and cancer. Semin Cancer Biol 20: 281-293.
59. Shibutani S, Takeshita M, Grollman AP (1991) Insertion of specific bases during DNA synthesis past the oxidation-damaged base 8-oxodG. Nature 349: 431-434.
60. Stiller M, Green RE, Ronan M, Simons JF, Du L, et al. (2006) Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. Proc Natl Acad Sci U S A 103: 13578-13584.
61. Kennedy SR, Salk JJ, Schmitt MW, Loeb LA (2013) Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet 9: e1003794.
62. Greenberg RS (2005) Medical epidemiology. (4thedn) Lange Medical Books/McGraw-Hill, New York.