Molecular BioSystems

PAPER

View Article Online
View Journal | View Issue

Cite this: *Mol. BioSyst.*, 2016, 12, 914

Characterization of sequence-specific errors in various next-generation sequencing systems†

Sunguk Shin and Joonhong Park*

Received 5th November 2015,
Accepted 4th January 2016

DOI: 10.1039/c5mb00750j

www.rsc.org/molecularbiosystems

Next-generation sequencing (NGS) is a popular method for assessing the molecular diversity of microbial communities without cultivation, for identifying polymorphisms in populations, and for comparing genomes and transcriptomes. However, sequence-specific errors (SSEs) by NGS systems can result in genome mis-assembly, overestimation of diversity in microbial community analyses, and false polymorphism discovery. SSEs can be particularly problematic due to rich microbial biodiversity and genomes containing frequent repeats. In this study, SSEs in public data from all popular NGS systems were discovered using a Markov chain model and hotspots for sequence errors were identified. Deletion errors were frequently preceded by homopolymers in non-Illumina NGS systems, such as GS FLX+. Substitution errors were often related to high GC contents and long G/C homopolymers in Illumina sequencing systems such as HiSeq. After removal of long G/C homopolymers in HiSeq, the average lengths of contigs and average SNP quality increased. SSEs were selectively removed from our mock community data by quality filtering, and a bias against specific microbes was identified. Our findings provide a scientific basis for filtering poor-quality reads, correcting deletion errors, preventing genome mis-assembly, and accurately assessing microbial community compositions and polymorphisms.

Introduction

Next-generation sequencing (NGS) systems are widely used for amplicon-based sequencing[1] and shotgun sequencing.[2] Most research conducted in the fields of genomics,[3] metagenomics,[4] and transcriptomics[5] requires cost-efficient and time-saving NGS technologies. Typical NGS platforms include GS FLX+ (FLX+)/GS Junior from Roche, Genome Analyzer (GA)/HiSeq/MiSeq from Illumina, and SOLiD/Ion PGMTM (PGM)/Ion Proton™ from Life Technologies. These NGS systems anchor DNA fragments to a solid surface, amplify the fragments, and sequence the amplified DNA in parallel. Although the basic principle is the same, there are differences in details of sequencing processes among the different NGS platforms. Illumina systems (MiSeq, GAII, and HiSeq platforms) use ddNTPs and washing steps, while non-Illumina systems (GS Junior, FLX+, and PGM platforms) generally utilize the sequential addition of dNTPs and apyrase.[6]

Many scientists and engineers within the fields of genomics and metagenomics are interested in reducing the errors that arise from NGS systems because even a low error rate can cause a significant number of sequencing errors due to the large-scale nature of sequence data. Such sequencing errors can result in genome mis-assembly,[7] overestimation of the diversity of microbial communities,[8] and misidentification of microbes.[9]

Sequence-specific errors (SSEs) are sequencing errors which are induced by sequence context. Some obvious SSEs have been easily discovered, and pioneering studies on nucleotide-dependent errors or SSEs have increased the accuracy and precision of NGS systems. For example, uneven and false signals from excessive dATP, T homopolymeric regions, primer-dimers, and loop structures can be prevented by 2′-deoxyadenosine-5′-*O*′-1-thiotriphosphate[10] and Sequenase.[11] In NGS systems from Roche, homopolymers were determined to increase insertion/deletion errors[12] and Ns,[9] and a software to correct such errors was developed.[9,13] In Illumina sequencing systems, a software[14] to correct nucleotide-dependent errors, such as the accumulation of T fluorophores,[15] was developed. There are a few studies to detect unknown sequential patterns of SSEs (Table 1). In Illumina systems, G-rich sequences,[16] inverted repeats,[17] and GGC[17] and GGT[18] sequences trigger errors. These studies are very important for distinguishing sequencing errors from true SNPs,[19,20] for filtering low-quality reads,[21] and correcting errors.[9]

However, more comprehensive and rigorous studies to discover unknown sequential patterns of errors are needed. First, more studies in non-Illumina platforms (GS Junior, FLX+, and PGM) are required. Second, methodologies[17,18] relevant to genome mapping may detect not SSEs but site-specific errors falsely induced by DNA damages/mutations similar to C-to-T transitions[22] or heterozygous

Department of Civil and Environmental Engineering, Yonsei University, Yonsei-ro 50, Seodaemoon-gu, Seoul, Republic of Korea. E-mail: parkj@yonsei.ac.kr; Tel: +82-2-2123-5798

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c5mb00750j

Table 1 Previous studies for detecting SSEs

Platform	Basic methodology	Major SSE pattern	Ref.
Illumina	Sequence context/shotgun sequencing	G-rich sequences	16
	Genome mapping/shotgun sequencing	Inverted repeats and GGC sequences	17
	Genome mapping/shotgun sequencing	GGT sequences	18
	Sequence context/shotgun sequencing	ACGGCGGT, *etc.*	19
FLX	Detecting error hotspots/amplicon sequencing	Homopolymer, a few other bases, and N	9

sites. Third, the read direction and mismatch fraction have been used to distinguish SSEs from putative SNPs. However, the methodology could not detect SSEs within inverted repeats in a previous study.[17] Fourth, the effects of sequential patterns of errors should be objectively assessed by exploring the error rates of reads containing SSEs. For example, some sequential patterns of SSEs, such as homopolymers and Ns, significantly increase the average error rates of the reads.[9,12] In particular, some SSEs relevant to lagging-strand dephasing[23] may significantly increase the average error rates of the reads. However, other sequential patterns of SSEs may induce a single sequencing error in each read slightly increasing the average error rate of the read. In this study, many sequential patterns of errors were identified using various approaches from various NGS platforms. A Markov chain model was used as a statistical method to discover biased sequential patterns of errors including inverted repeats regardless of the genomic position. The effects of sequential patterns of errors as filtering parameters were assessed by exploring the error rates of reads containing SSEs, microbial community structures, genome assembly quality, and SNP qualities. In addition, the sequential patterns of error hotspots were explored by amplicon sequencing.

Materials and methods

Public data

To explore SSEs in various NGS platforms regardless of experimental conditions and avoid heterozygous genomes, mainly public microbial data from resequencing studies were downloaded. The public data were from GS Junior (SRA Accession: SRX111101), FLX+ (Accession: SRX111103), PGM (Accession: SRX111376), MiSeq (Accession: SRX111764), GAII (Accession: DRX000504), and HiSeq (Illumina Data Library Human Chr 21) platforms. In addition, public FLX amplicon sequencing data were obtained from a previous study.[24]

Mock community construction and environmental sampling

To assess changes in microbial community structures, a mock community was constructed from the genomes of the following 10 bacterial isolates that have been sequenced genome-wide (Bioproject PRJNA290408): *Arthrobacter chlorophenolicus* A6 (GenBank id gb: CP001341.1), *Chromobacterium violaceum* ATCC 12472 (gb: AE016825.1), *Corynebacterium glutamicum* ATCC 13032 (gb: BA000036.3), *Escherichia coli* ATCC 8739 (gb: CP000946.1), *Escherichia coli* W (gb: CP002185.1), *Klebsiella pneumoniae* KCTC 2242 (gb: CP002910.1), *Polaromonas naphthalenivorans* CJ2 (gb: CP000529.1), *Pseudomonas stutzeri* ATCC 17588 (gb: CP002881.1), *Roseobacter denitrificans* OCh 114 (gb: CP000362.1), and *Staphylococcus epidermidis* ATCC 12228 (gb: AE015929.1). These bacteria were selected to evaluate various sequences because their genome sizes and GC contents vary (Table S1 in the ESI†). As an environmental sample, DNA extracted directly from soil from Aewol Gotjawal, Jeju, Korea, was used. Genomic DNA was extracted using the PowerSoil DNA isolation kit (MoBio Laboratories, Inc., Carlsbad, CA, USA) and treated with RNase A (QIAGEN, Hilden, Germany) and protein precipitation solution (SolGent, Daejeon, Korea). The extracted DNA was quantified using a NanoDrop ND-1000 spectrophotometer (Thermo Scientific, DE, USA) and evenly mixed.

Markov chain analysis

To overcome the problems in previous studies, Markov chains were used to detect the relationship between errors and flanking sequences in various genomes. The underlying concept of Markov chains is that observers can detect bias in a word and evaluate significance from the number of occurrences of the word and smaller words contained in the word.[25]

$W = (w_1 w_2 \ldots w_m)$ is the word created by the concatenation of m nucleotides. In this study, Markov orders were assigned from 2 to 8. $N(W)$ is the observed count of a word in a sequence with a length of m. Under the Markov maximal order model, the expected count $E(W)$ of W is shown in eqn (1).

$$E(W) = \frac{N(w_1 w_2 \ldots w_{m-1}) N(w_2 w_3 \ldots w_m)}{N(w_2 w_3 \ldots w_{m-1})} \quad (1)$$

Having obtained a theoretical expectation for the count of a word, a statistical method to compare it to the real observed count in a statistically meaningful way is required. For this purpose, Z-value statistics were used.[26] Z values were calculated using eqn (2).

$$Z(W) = \frac{N(W) E(W)}{\mathrm{var}(W)} \quad (2)$$

in which var(W) represents the calculated variance of $N(W) - E(W)$. For large sequences and large counts, the variance of the maximal Markov model[27] can be well approximated by eqn (3).

$$\mathrm{var}(W) = E(W) \times [N(w_2 w_3 \ldots w_{m-1}) - N(w_1 w_2 \ldots w_{m-1})] \times [N(w_2 w_3 \ldots w_{m-1}) - N(w_2 w_3 \ldots w_m)] / N(w_2 w_3 \ldots w_{m-1})^2 \quad (3)$$

The Z value is a measure of the bias of a word. A large negative value signifies 'under-representation', and a large positive value represents 'over-representation' of the word in a NGS read. Perl scripts to select sequences with the highest and lowest Z values

as sequential motifs of errors (MC4SSE) are publicly available at http://markovchain4sse.sourceforge.net.

Computational analyses and filtering

After sequencing, the error rates of reads were determined by BLAST analysis with an optimized BLAST parameter setting (-max_target_seqs 1) to match against one reference sequence. Short sequences that flanked errors were selected from the BLAST results using Perl scripts to apply a Markov chain. To estimate the microbial community composition of the constructed mock community, MG-RAST[28] was used. After the default filtering process of MG-RAST, the microbial community structures were analyzed at the class level using an optimized parameter setting (annotation sources: Greengenes and SEED with maximal *e*-values of 1×10^{-30} and 1×10^{-5}, respectively) that confirmed the even mock community (DNA quantity). The filtered reads were assembled using Velvet assembly tools[29] into different *k*-mer lengths, and contigs of the best *k*-mer lengths producing the longest contig lengths were selected for analysis. To call SNPs from the environmental sample, Bowtie 2[30] and SAMtools/BCFtools[31] were used. Perl scripts to filter and correct reads containing sequential patterns of errors (SSECF) are publicly available at http://ssecf.sourceforge.net.

Results and discussion

Sequences flanking deletion errors

The highest/lowest *Z* values and the frequency of sequences flanking deletion errors are shown in Table 2. Except for the HiSeq platform, A⟨D⟩A and C⟨D⟩C had the lowest *Z* values at the mononucleotide level. Sequences with inverted repeats, such as TT⟨D⟩AA, TTT⟨D⟩AAA, AAA⟨D⟩TTT, and CAGG⟨D⟩CCTG had the lowest *Z* values. The highest-frequency sequences often contained homopolymers, such as AAAA⟨D⟩TAAA, TTTT⟨D⟩CTTT, TTTT⟨D⟩ATTT, TTTT⟨D⟩CTTT, TTTT⟨D⟩CCAG, TTTT⟨D⟩CAGC, TTTT⟨D⟩CACC, CGTA⟨D⟩TTTT, and AGGG⟨D⟩GGCT. The high *Z*-value sequences exhibited the following common sequential pattern: a homopolymer + a few nucleotides with bases other than the homopolymer + a few nucleotides with the same base as the base of the homopolymer; examples include AAAG⟨D⟩CAAT (the reference sequence: AAAGACAAT), GCCG⟨D⟩CCTG (reference: GCCGCCCTC), and TTTG⟨D⟩CAAC (reference: TTTGTCAAC) in GS

Table 2 The highest/lowest *Z* values and frequent sequences flanking deletion errors. ⟨D⟩ denotes deletion errors

Type of deletion error	NGS system	Highest *Z* value	Most frequent sequence	Lowest *Z* value	Least frequent sequence
Mono-	GS Junior	T<D>A	T<D>A	A<D>A	C<D>C
	FLX+	C<D>G	T<D>C	C<D>C	G<D>G
	PGM	G<D>C	T<D>C	C<D>C	G<D>G
	MiSeq	C<D>G	C<D>G	C<D>C	C<D>A
	GAII	A<D>T	A<D>T	A<D>A	C<D>C
	HiSeq	C<D>T	A<D>G	C<D>A	T<D>G
Di-	GS Junior	AT<D>AA	TT<D>AT	TT<D>AA	GA<D>AG
	FLX+	AT<D>AA	TT<D>AT	TT<D>AA	GT<D>TC
	PGM	AC<D>TT	TT<D>CA	CC<D>TT	AG<D>GG
	MiSeq	GT<D>GG	TT<D>CA	AT<D>AT	CT<D>GG
	GAII	CT<D>AT	TA<D>TT	AA<D>TT	CT<D>AT
	HiSeq	CC<D>CC	CC<D>CC	CC<D>TA	TA<D>CG
Tri-	GS Junior	ATT<D>AAA, TAA<D>TTT	TTT<D>ATT, AAA<D>TAA	TTT<D>AAA, GGG<D>AAA	GAG<D>GTT, CTC<D>GAG, *etc.*
	FLX+	CGC<D>AAC, AAC<D>GTA	TTT<D>ATT, AAA<D>TAA	TAT<D>AAA, GTG<D>GTG	TTG<D>GTC, CAC<D>CAC, *etc.*
	PGM	TAA<D>TTT, CCC<D>TCC	TTT<D>CAG, AAA<D>TCA	AAA<D>TTT, AAA<D>ATC	TAG<D>GGA, CTC<D>CTA
	MiSeq	GTC<D>GGG, GGG<D>CTA	TTT<D>CAC, CGT<D>ATT	CGG<D>GTC, GGT<D>ATT	TAG<D>CTA, CTA<D>TAG
	GAII	CTA<D>TTG, TAC<D>AAT	ACT<D>ATT, CTA<D>TTG	CTA<D>TTT, AAA<D>TAG	TTC<D>AAG, CTT<D>GAA, *etc.*
	HiSeq	ACA<D>GGC, GCC<D>TGT	GCC<D>CCC, GCA<D>GGA	GCA<D>GGC, GCC<D>TGC	ACC<D>CAG, CGG<D>GGC, *etc.*
Tetra-	GS Junior	TAGG<D>CGGA, CAAG<D>CAAG	AAAA<D>TAAA, TTTT<D>CTTT	CAGG<D>CGGA, AAAA<D>TTTT	CAGG<D>CGGA, TCCG<D>CCTC, *etc.*
	FLX+	TCCG<D>CCTA, TAGG<D>CGGA	TTTT<D>ATTT, TTTT<D>CTTT	AAAC<D>GTAA, TATA<D>ATAT	TATA<D>ATAT, ACAA<D>AGTA, *etc.*
	PGM	TCCG<D>CCTA, TAGG<D>CGGA	TTTT<D>CCAG, TTTT<D>CAGC	TCCG<D>CCTG, AAAA<D>CCCA	TCAC<D>CTTG, TCAC<D>CCTA, *etc.*
	MiSeq	ACGT<D>ATTT, ATTT<D>CACA	TTTT<D>CACC, CGTA<D>TTTT	TTTT<D>CACA, ATTT<D>CACC	CGGG<D>CTAC, AGGG<D>CTAA, *etc.*
	GAII	CACT<D>ATTT, TACT<D>ATTG	TACT<D>ATTG, ACTA<D>TTGA	TACT<D>ATTT, CACT<D>ATTG	TACT<D>ATTT, CACT<D>ATTG, *etc.*
	HiSeq	GGCC<D>CCCG, GGCC<D>CCTC	AGCC<D>CCCT, AGGG<D>GGCT	CAGG<D>CCTG, GAGG<D>GGCT	CAGG<D>CCTG, GAGG<D>GGCT, *etc.*

Junior, FLX+, and PGM systems. Some minisatellites, such as CTCT⟨D⟩AGGT, had relatively high *Z* values in all of the NGS systems (not shown in Table 2). Deletion errors may frequently occur in homopolymers and minisatellites in all of the NGS systems investigated due to DNA slippage.

Sequences flanking insertion errors

As shown in Table 3, at the di-and trinucleotide levels, many inverted repeats, such as GG⟨I⟩CC, AAA⟨I⟩TTT, CGG⟨I⟩CCG, GGC⟨I⟩GCC, and GCC⟨I⟩GGC had the lowest *Z* values except for those observed in the GAII and HiSeq platforms. Insertion errors also frequently occurred in homopolymers and minisatellites in all of the NGS systems, such as AAAA⟨I⟩TGCC, AAAT⟨I⟩AAAA, GGCG⟨I⟩TTTT, ATCA⟨I⟩GGGG, ATTT⟨I⟩CCCC, ATCA⟨I⟩GGGG, and GAGA⟨I⟩GTGT (reference: GAGAGGTGT, not shown in Table 3).

Sequences flanking Ns

As shown in Table 4, Ns were not observed in the PGM system, and few Ns occurred in the MiSeq system. Based on our experience and data, Ns very seldom occur in public Illumina sequencing data. However, in the GS Junior and FLX+ systems, the most frequent sequences shared a common sequential pattern: a homopolymer + a few nucleotides with bases other than the homopolymer + one nucleotide with the same base as the homopolymer; examples include AAAC⟨N⟩CTAT (reference: AAACACTAT), AAAG⟨N⟩TAAA (reference: AAAGATAAA), and CCCTTGGGAC⟨N⟩TACTAGTTCT (reference: CCCTTGGGACCT ACTAGTTCT).

Sequences flanking substitution errors

The sequences with the highest/lowest *Z* values are shown in Table S2 in the ESI.† Obviously, common sequential patterns among the sequences were not identified. According to the BLAST results and the frequency of errors, the causes of most substitution errors may be incorrect reference sequences and/or variations in DNA. For example, in the MiSeq system, GGGG⟨S⟩GAAG nearly coincided with GGTGTCGGGGAGAAG CCCTGA (GenBank accession number gi: 288994861).

Error hotspots and their sequential patterns in the FLX system

The sequence motifs of Ns and insertion and deletion errors from FLX amplicon sequencing are summarized in Table 5. Substitution errors were not analyzed because most occurred at the ends of reads, regardless of their sequential patterns.

Table 3 The highest/lowest *Z* values and frequent sequences flanking insertion errors. ⟨I⟩ denotes insertion errors

Type of insertion error	NGS system	Highest *Z* value	Most frequent sequence	Lowest *Z* value	Least frequent sequence
Mono-	GS Junior	A<I>T	G<I>C	A<I>A	C<I>C
	FLX+	C<I>A	A<I>T	A<I>A	A<I>A
	PGM	A<I>T	A<I>T	C<I>C	G<I>G
	MiSeq	C<I>G	C<I>G	G<I>G	T<I>A
	GAII	A<I>G	C<I>T	G<I>G	C<I>C
	HiSeq	C<I>T	A<I>G	C<I>G	T<I>T
Di-	GS Junior	GC<I>AA	TT<I>GC	GG<I>CC	TA<I>AT
	FLX+	GC<I>AA	AA<I>TG	GG<I>CC	TA<I>AA
	PGM	CG<I>CC	CG<I>TT	GG<I>CC	AG<I>GG
	MiSeq	AG<I>TC	TG<I>TT	TT<I>CT	CT<I>AG
	GAII	AG<I>TG	AC<I>TT	AA<I>TT	AG<I>TA
	HiSeq	GG<I>GG	GG<I>GG	AC<I>GT	CG<I>CG
Tri-	GS Junior	CAG<I>CTT, ATC<I>CAT	TTC<I>AGC, TTG<I>CTG	AAA<I>TTT, TTG<I>AAA	CTC<I>CAT, ATC<I>CAG, *etc.*
	FLX+	ATC<I>CAT, TAA<I>AGT	AAA<I>TCA, TTC<I>AGC	CTC<I>CAT, CGG<I>CCG	ATA<I>ACC, GCT<I>TGG, *etc.*
	PGM	GCG<I>TTT, GGC<I>GCT	AAT<I>AAA, GCG<I>TTT	GGC<I>GCC, GCC<I>GGC	TAG<I>GAG, ACT<I>TGG
	MiSeq	TTG<I>TTC, AGG<I>CGG	AGG<I>CGG, CTC<I>AAA	GTG<I>TTC, ATT<I>GCG	CCT<I>AGG, CTA<I>GGA
	GAII	TAC<I>TTG, GAA<I>GTG	TAC<I>TTG, CAA<I>GTA	CAC<I>TTG, CAA<I>GTG	TCT<I>TGA, CAA<I>GTT, *etc.*
	HiSeq	CGG<I>GGG, CCC<I>CCG	AGG<I>GGC, GCC<I>CCT	CGG<I>GGC, GCC<I>CCG	CCG<I>CAC, TGC<I>CAA, *etc.*
Tetra-	GS Junior	TCCG<I>CCTA, TAGG<I>CGGA	AAAA<I>TGCC, TTTC<I>AGCG	TCAA<I>CAGA, AAGG<I>CGGA	ACTC<I>GCAC, CCTC<I>GCAG, *etc.*
	FLX+	AAGA<I>TTTC, GAAA<I>TCTT	TTTC<I>AGCA, TTTC<I>AGCG	TCTG<I>TTGA, AAGA<I>TTTT	CCTG<I>TTGC, CAGG<I>GTTG, *etc.*
	PGM	TTCC<I>ATGG, CCGG<I>CCTA	AAAT<I>AAAA, GGCG<I>TTTT	CCCA<I>AGAA, TACG<I>TTTT	CCCA<I>AGAA, TCCA<I>AGAG, *etc.*
	MiSeq	ATCA<I>GGGG, ATTT<I>CCCC	TAGG<I>CGGA, ATCA<I>GGGG	TTTT<I>CCCC, AACC<I>ACAC	GCCG<I>CCTA, CGGA<I>AGGA, *etc.*
	GAII	TCAA<I>GTAA, ACAA<I>GTAT	TTAC<I>TTGA, TCAA<I>TGAA	ACAA<I>GTAA, TCAA<I>GTAT	ACAA<I>GTAA, TCAA<I>GTAT, *etc.*
	HiSeq	GAGG<I>GGCC, CGCA<I>GGAC	CAGG<I>GGCT, AGCC<I>CCTG	GGCA<I>GGGG, ACAC<I>CACA	ACAC<I>CACA, GCAC<I>CACC, *etc.*

Table 4 The highest/lowest *Z* values and frequent sequences flanking Ns

Type of N error	NGS system	Highest *Z* value	Most frequent sequence	Lowest *Z* value	Least frequent sequence
Mono-	GS Junior	T<N>T	C<N>G	T<N>G	T<N>G
	FLX+	T<N>T	C<N>G	C<N>T	T<N>G
	PGM	NA	NA	NA	NA
	MiSeq	G<N>A	C<N>G	C<N>C	A<N>A
	GAII	C<N>G	T<N>T	A<N>T	C<N>C
	HiSeq	T<N>T	T<N>T	T<N>G	G<N>C
Di-	GS Junior	AG<N>GA	AC<N>GA	AC<N>CG	CA<N>AT, *etc.*
	FLX+	AC<N>GA	AC<N>GA	TG<N>GA	AA<N>TT, *etc.*
	PGM	NA	NA	NA	NA
	MiSeq	AC<N>TA	TA<N>GA	GA<N>GA	GA<N>GA, *etc.*
	GAII	AA<N>AG	TT<N>TT	AA<N>AA	GG<N>CC
	HiSeq	AT<N>TA	AT<N>TA	GT<N>TA	CG<N>GC
Tri-	GS Junior	TCG<N>GAG, TAC<N>CAC	TTA<N>CTG, AAC<N>GAA	CGT<N>TGG, ACC<N>GGT	CGT<N>TGG, ACC<N>GGT, *etc.*
	FLX+	ACC<N>GTA, GCT<N>ACA	AAC<N>GAA, AAG<N>TAA	ACC<N>CGC, GCC<N>GTA	AGG<N>GCA, ATC<N>TAT, *etc.*
	PGM	NA	NA	NA	NA
	MiSeq	GCG<N>AAA, TCG<N>AAC	ATA<N>GAC, GAC<N>TAC	TCG<N>AAA, GCG<N>AAC	TCG<N>AAA, GCG<N>AAC, *etc.*
	GAII	CTC<N>ACC, TCT<N>ACT	TTT<N>TTT, TTC<N>TTT	AAA<N>AAA, AGT<N>CCA	TAG<N>CTA, AGG<N>CTA
	HiSeq	GAT<N>TAA, CGC<N>ATC	GAT<N>TAA, AAC<N>TCC	GAT<N>TAT, GAT<N>TAC	GAT<N>TAC, CAC<N>TCC, *etc.*
Tetra-	GS Junior	TAAC<N>CCGT, GAAC<N>TAAG	AAAC<N>CTAT, ATTA<N>CTGA	AAAC<N>CCAT, AAAC<N>TCTC	AAAC<N>CCAT, AAAC<N>TCTC, *etc.*
	FLX+	TAGG<N>TATA, TAGT<N>TTTG	AAAG<N>TAAA, GAAG<N>TAAA	AAAT<N>TAAA, AAAC<N>GCGA	TGGC<N>GCAT, TGCC<N>CAAA, *etc.*
	PGM	NA	NA	NA	NA
	MiSeq	CAAC<N>GACA, GAAC<N>GACG	CATA<N>GACG, GGAC<N>TACT	GAAC<N>GACA, CAAC<N>GACG	GAAC<N>GACA, CAAC<N>GACG, *etc.*
	GAII	GGAA<N>TTCG, AGGG<N>AACA	ATTC<N>TTTT, TTTC<N>TTTT	ATGG<N>TAGG, AAGA<N>CTTT	TTAG<N>ACGT, CCCG<N>ACCT, *etc.*
	HiSeq	AAAC<N>TCCT, CGCA<N>TCAC	CCAT<N>TAAA, TAAC<N>TCCA	CTAG<N>GCAA, CGCA<N>TCAG	CTAG<N>GCAA, CGCA<N>TCAG, *etc.*

Table 5 Error hotspots and their sequence motifs in FLX amplicon sequencing. Positions and sequence motifs of Ns (N), insertion errors (I), and deletion errors (D). The bold letters represent the bases at the positions of the errors. The underlined letters denote the homopolymers in front of the Ns and deletion errors. The additional files of Gilles's study (Gilles *et al.*, 2011) were analyzed. There was no significant peak of Ns in sequence 2 (NA)

Sequence	Error type	Position (bp)	Motif	Erroneous sequence
1	N	210	CCAAAAACG**A**GGAGG	CCAAAAACG**N**GGAGG
	I	207	CC**AAAAA**CGAGG	CC**AAAAAA**CGAGG
	D	365	GGGTTTGG**TTTTT**G	GGGTTTGG**TTTT**G
2	N	NA	NA	NA
	I	151	TT**AAAA**CTTT	TT**AAAAA**CTTT
	D	413	CGCGCG**TTT**CG	CGCGCG**TT**CG
3	N	256	CGAAA**C**AGG	CGAAA**N**GG
	I	37	TTAAAGC**TTTTTT**GAAA	TTAAAGC**TTTTTTT**GAAA
	D	32	TTAAAGC**TTTTTT**GAAA	TTAAAGC**TTTTT**GAAA
4	N	301	AACCCCCG**C**GG	AACCCCCG**N**GG
	I	291	TTA**C**GAACCCCC	TTA**C**GAACCCCC
	D	313	GCCGGG**CCC**GG	GCCGGG**CC**GG
5	N	274	GCCCCC**G**CT	GCCCCC**N**T
	I	324	AAA**C**CCCTG	AAA**C**CCCCTG
	D	16	TTTTTGCG**TTTT**G	TTTTTGCG**TTT**G

Many Ns, insertion errors, and deletion errors contained the following similar sequential patterns: a homopolymer + a few nucleotides with bases other than the homopolymer + one nucleotide with the same base as the preceding homopolymer; a homopolymer; a homopolymer + a few nucleotides with bases other than the homopolymer + a few nucleotides with the same base as the preceding homopolymer, respectively. As shown in sequence 2 of Table 5, deletion errors also occurred in minisatellites.

Inverted repeats, GGC and GGT sequences, and GC contents

Inverted repeats, GGC sequences, and GGT sequences have been reported to trigger sequencing errors in the Illumina

Table 6 Error rates of all the reads and reads with GGC sequences, GGT sequences, or inverted repeats. Inverted repeats contain 0–6 bp gaps

		Reads containing inverted repeats			Reads containing GGC sequences			Reads containing GGT sequences		
NGS system	All reads (%)	4 bp (%)	5 bp (%)	≥6 bp (%)	0 (%)	1–2 (%)	≥3 (%)	0 (%)	12 (%)	≥3 (%)
GS Junior	0.60	0.59	0.61	0.70	1.69	1.23	0.58	1.51	1.06	0.58
FLX+	0.85	0.83	0.86	0.95	1.36	1.17	0.84	1.23	1.30	0.84
PGM	2.68	2.68	2.69	2.76	3.02	2.72	2.50	2.84	2.68	2.58
MiSeq	0.95	0.92	1.00	1.04	0.32	0.47	1.03	0.72	0.76	0.99
GAII	0.57	0.56	0.57	0.62	0.51	0.59	0.82	0.56	0.59	0.64
HiSeq	0.63	0.61	0.66	0.52	0.59	0.57	1.14	0.68	0.57	1.05

systems.[17,18] To assess the effects of these sequences on the error rates of reads in various NGS platforms, the error rates of reads containing these sequences were evaluated (Table 6). The error rates of reads containing inverted repeats were not significantly higher than the error rate of all reads. However, long inverted repeats may induce sequencing errors because they produce hairpins. Manual inspection revealed that long inverted repeats (≥6 bp) resulted in nearby insertion/deletion errors although, at the middle of the inverted repeats, the insertion/deletion error rates might be lower than in other general sequences according to the Markov chain results. Reads containing very long inverted repeats (≥10 bp) were rarely extended at the ends of the repeat. The error rates of reads containing more than 3 GGC or GGT sequences increased 0.08–0.71% than those of reads not containing GGC or GGT sequence. However, the error rates of the reads containing 1–2 GGC or GGT sequences were lower than those of reads containing 0 GGC or GGT sequences in HiSeq. According to the Markov chain results and thorough manual analysis, the positions of the GGC and GGT sequences were not directly related to the errors. However, GC-rich sequences frequently contained many GGC and GGT sequences, and A/T was often substituted for G/C. By contrast, in the GS Junior, FLX+, and PGM platforms, reads containing more than 3 GGC or GGT sequences did not have higher error rates than reads containing fewer GGC or GGT sequences. In the GS Junior, FLX+, and PGM systems, reads not containing GGC or GGT sequences frequently contained many A/T homopolymers and insertion/deletion errors.

To confirm the relationship between the GC content and substitution error rates, substitution error rates according to the GC content were examined (Table 7). In the GS Junior, FLX+, and PGM systems, reads containing ≥40% and <60% GC content had the lowest substitution error rates, whereas in the MiSeq, GAII, and HiSeq systems, the substitution error rates of reads containing ≥60% and <80% GC content increased 0.59–2.06% than those of reads containing ≥20% and <40% GC content. The relationships between the GC content and substitution errors were also explored because G in particular has been reported to preferentially incur an incomplete deprotection and fluorophore removal step.[16] As indicated in Table 6, a high C content as well as a high G content increased the substitution error rate in the Illumina systems. These results may coincide with high G→T and C→A substitution error rates in a previous study.[32]

Error rates of reads containing homopolymers

Ns and insertion/deletion errors were related to the presence of homopolymers, particularly in the GS Junior, FLX+, and PGM systems. The error rates according to the number and average length of homopolymers (≥4 bp) are shown in Fig. 1. The error rates significantly increased as the number and the length of homopolymers increased in all NGS systems examined except for GS Junior. However, the Illumina sequencing systems demonstrated dramatically higher error rates for reads containing long homopolymers, which were related to substitution errors instead of insertion/deletion errors. The error rates of the reads according to the longest homopolymers were calculated and are shown in Table S3 in the ESI.† Reads containing long homopolymers corresponded to high error rates in the PGM systems. The error rates of the reads according to the longest G or C homopolymers were also analyzed. As shown in Table 8, the error rates of reads with G or C homopolymers greater than 8 bp in length in the Illumina systems were generally greater than 3%. G/C homopolymers may induce errors in most positions of reads containing G/C homopolymers.

Error rates of reads containing sequences with high *Z* values

To test the applicability of sequences with high *Z* values for filtering erroneous reads, TAGGNCGGA was selected as a sequence with a

Table 7 Substitution error rates and GC /G /C content. GC content is the percentage of nitrogenous bases on a DNA molecule that are either guanine or cytosine

	GC content (%)			G content (%)			C content (%)		
NGS system	≥20% and <40%	≥40% and <60%	≥60% and <80%	≥10% and <20%	≥20% and <30%	≥30% and <40%	≥10% and <20%	≥20% and <30%	≥30% and <40%
GS Junior	0.24	0.13	0.65	0.46	0.24	0.29	0.47	0.24	0.29
FLX+	0.29	0.24	1.01	0.28	0.24	0.39	0.29	0.24	0.40
PGM	0.47	0.37	0.47	0.47	0.36	0.33	0.41	0.36	0.44
MiSeq	0.32	0.83	2.38	0.35	0.83	1.56	0.44	0.82	1.56
GAII	0.56	0.56	1.15	0.60	0.63	0.82	0.59	0.65	0.81
HiSeq	0.33	0.73	1.55	0.41	0.61	1.13	0.43	0.61	1.11

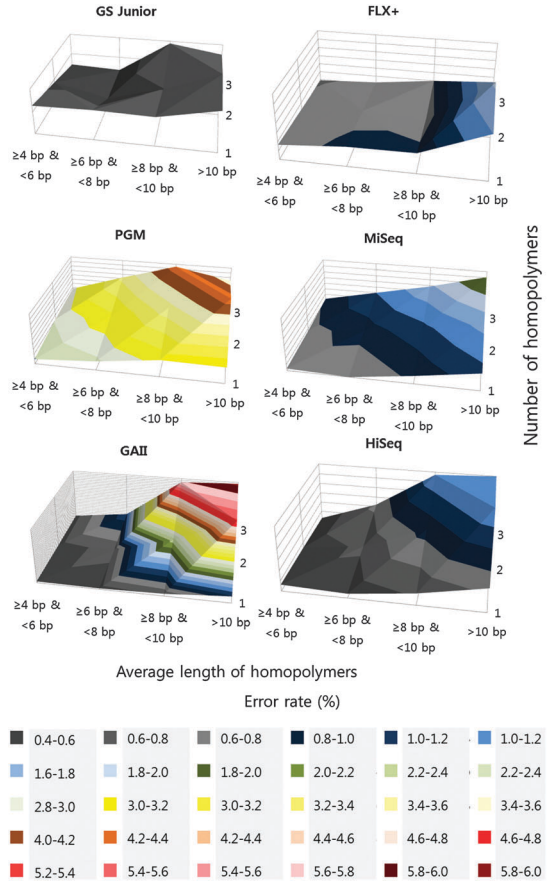

Fig. 1 Error rates according to the number and average length of homopolymers (≥4 bp). The colored regions correspond to different error rates.

Table 8 Error rates and lengths of the longest G or C homopolymers. "Length of a G or C homopolymer" indicates the length of the longest G or C homopolymer in each read

	Length of a G or C homopolymer				
NGS system	<4 bp (%)	≥4 bp and ≤5 bp (%)	≥6 bp and ≤7 bp (%)	≥8 bp and ≤9 bp (%)	>9 bp (%)
GS Junior	0.59	0.59	0.86	1.08	1.52
FLX+	0.79	0.84	1.01	1.18	1.27
PGM	2.64	2.76	3.23	3.77	4.58
MiSeq	0.62	0.97	1.85	3.43	2.81
GAII	0.54	0.70	1.45	5.00	4.22
HiSeq	0.55	0.83	1.26	2.64	4.15

high Z value in GS Junior, FLX+, and PGM (Table 2). As shown in Table S4 in the ESI,† the error rates of reads containing the TAGGNCGGA sequence were 0.22–0.77% higher than the error rates of all reads in GS junior, FLX+, and PGM. Surprisingly, the error rate of reads containing the TAGGNCGGA sequence was high in MiSeq due to a high substitution error rate (2.35%). In the MiSeq data, reads containing TAGGNCGGA sequence might have high GC contents or contain G/C homopolymers in the data. However, the TAGGNCGGA sequence did not significantly increased the average error rates of the entire reads in GS Junior, FLX+, and PGM.

Quality filtering and sequential patterns of errors

The metagenomes of the constructed mock community were sequenced using the FLX and HiSeq platforms, which are the most frequently used NGS platforms. The microbial community compositions were analyzed after filtering using different quality scores to test the quality score filtering bias against reads containing sequential patterns of errors. As major contributors to SSEs, homopolymers (≥10 bp) and G/C homopolymers (≥8 bp) were selected for FLX and HiSeq analysis, respectively. In the FLX system, the fractions of reads containing homopolymers (≥10 bp) were 0.049% and 0.024% using quality filtering scores of 20 and 30, respectively. In the HiSeq system, the fractions of reads containing G or C homopolymers (≥8 bp) were 0.013% and 0.006% after using quality filtering Phred scores of 20 and 30, respectively. The ratios of reads containing homopolymers and G/C homopolymers were significantly altered by quality filtering in the FLX system (chi-square = 9.26 with 1 df, $P \leq 0.01$) and HiSeq system (chi-square = 390.8 with 1 df, $P \leq 0.001$), respectively. In the FLX system, the RNA and protein database compositions of Gammaproteobacteria and Bacilli, respectively, decreased when a quality filtering score of 30 rather than 20 was used (Fig. 2), and the overall microbial community structure using the protein database statistically varied with the quality filtering score (chi-square = 79.1 with 6 df, $P \leq 0.001$). The composition of the mock community was not significantly changed by the quality filtering process while using the RNA database with the FLX platform, likely due to the small number of reads. The percentage of Gammaproteobacteria was 31% and 24% after quality filtering using scores of 20 and 30, respectively. In the HiSeq system, the Bacilli fraction increased in

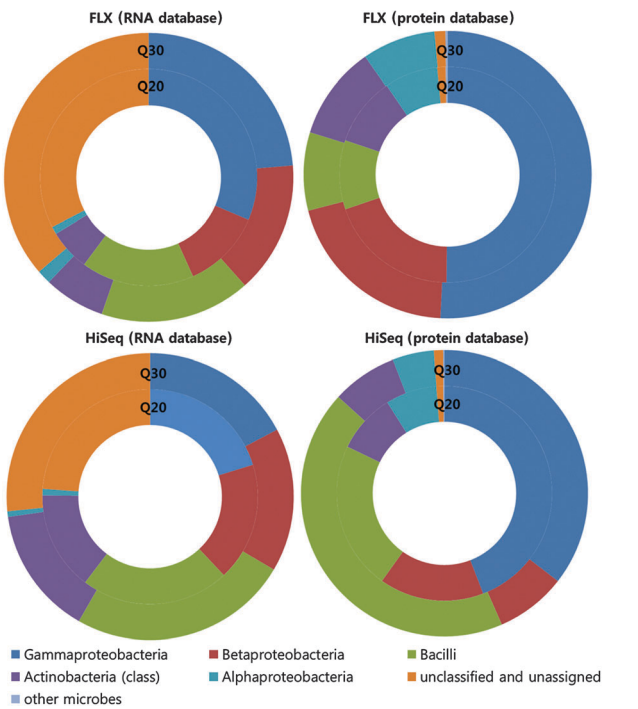

Fig. 2 Effects of different (Q30 and Q20) quality filtering processes on the estimation of microbial community structures using NGS methods.

Table 9 Removal of G/C homopolymers and its effects on assembly. Paired-end reads containing less than 70% bases with a PHRED score over 20 in HiSeq were filtered. Reads containing G or C homopolymers (≥8/10/12 bp) were removed. "Mock" and "Env" denote the mock community sample and environmental sample, respectively

Data	Sequence	*k*-mer	N50	The longest contig (bp)	Average length of contigs (bp)	Total length (bp)
Mock without removal	52310999	61	27628	904075	1900.63	51480516
		67	32761	904087	2063.55	51636093
		71	39582	487763	2058.86	51815387
Mock after removal (≥8 bp)	52217188	61	25390	904075	1868.68	51506362
		67	29571	904087	2029.00	51660454
		71	33207	487763	2035.71	51829252
Mock after removal (≥10 bp)	52289810	61	27403	904075	1903.93	51476579
		67	32742	904087	2067.11	51632376
		71	39908	487763	2074.32	51797868
Mock after removal (≥12 bp)	52302712	61	27096	904075	1902.51	51478166
		67	35004	904087	2075.95	51622743
		71	39750	487763	2066.31	51808599
Env without removal	15183392	67	254	3506	270.50	9805546
		71	262	3716	278.68	4796346
		77	275	1874	293.10	1141926
Env after removal (≥8 bp)	15136735	67	254	3542	270.65	9757610
		71	262	3680	278.74	4770888
		77	275	1874	292.61	1135318
Env after removal (≥10 bp)	15170371	67	254	3542	270.99	9785768
		71	262	3716	278.84	4785727
		77	275	1874	293.20	1141705
Env after removal (≥12 bp)	15178020	67	254	3542	270.66	9802427
		71	262	3716	278.90	4792582
		77	275	1874	293.75	1139737

Table 10 Removal of G/C homopolymers and effects on SNP calling. Paired-end reads containing less than 70% bases with a PHRED score over 20 in HiSeq were filtered. Reads containing G or C homopolymers (≥10 bp) were removed. Reads were aligned using Bowtie 2 to contigs that had been assembled using Velvet (*k*-mer = 71). "Env" denotes an environmental sample

Data	SNPs	Homozygous SNPs/heterozygous SNPs	Average SNP quality	Average total depth
Env without removal	17 826	0.2359	20.92	6.613
Env after removal	17 833	0.2353	21.03	6.591

the RNA and protein databases after using a quality filtering score of 30 compared with 20, and the overall microbial community composition using both the RNA (chi-square = 92.5 with 5 df, $P \leq 0.001$) and protein (chi-square = 1021506.3 with 6 df, $P \leq 0.001$) databases was significantly changed by the quality filtering scores. Interestingly, Bacilli (*Staphylococcus epidermidis*) in the mock community had the lowest genomic GC content, as shown in Table S1 in the ESI.† GC-rich microbes might be selectively removed in the Illumina systems by quality filtering.

Assembly, SNP calling, and sequential patterns of errors

Filtered reads of the mock community and environmental sample in HiSeq were assembled. As shown in Table 9, after the removal of reads containing G/C homopolymers (≥10 bp), the average lengths of the contigs were slightly higher in both the mock community and the environmental sample. SNPs were called from the environmental sample in Table 10. The average SNP quality was increased slightly after the removal of reads containing G/C homopolymers.

Conclusions

Many sequential patterns of errors in popular NGS systems were newly identified and reconfirmed in this study. For example, deletion errors exhibited an identical sequential pattern in GS Junior, FLX+, and PGM. In the MiSeq, GAII, and HiSeq systems, substitution errors were frequent in GC-rich reads and reads containing G/C homopolymers. Interestingly, the various NGS systems were roughly divided into two categories with regard to the sequential patterns of errors: the Illumina sequencing systems and the other sequencing systems. The Illumina systems exhibited high substitution error rates in reads containing G/C homopolymers and in GC-rich reads. These substitution errors might be attributable to the accumulation of the ddGTP/ddCTP remaining in clusters of the GC-rich DNA fragment after the washing steps in the Illumina systems because G and C have triple hydrogen bonds, whereas A and T have double hydrogen bonds,[33] similar to the accumulation of T dyes in the old Illumina system.[15] The substitution errors might also be due to lagging-strand dephasing.[31] For example, GC-rich strands that could not synthesize bases in previous cycles might incorporate G/C during an A/T cycle, generating substitution errors. The other sequencing systems tended to produce frequent Ns and insertion/deletion errors after or in homopolymers. The deletion errors and Ns may be due to the presence of excessive by-products in the other systems.[8] However, more detailed experimental results are required to identify the exact causes.

Some patterns of SSEs can be used as filtering parameters. In particular, sequential patterns of substitution errors in the Illumina NGS systems might be too ambiguous to be corrected, but they can be filtered using the GC content or G/C homopolymers. Filtering processes might remove erroneous reads

and prevent mis-assembly and fake SNPs in environmental samples and plant genomes. According to Fig. 1, homopolymers greater than ~6 bp in the PGM systems should also be filtered out for microbial 16S rRNA gene amplicon sequencing at the species level (3%). Error rates can vary according to experimental conditions, but Table S3 (ESI†), Fig. 1 and Table 7 should be useful for determining filtering criteria. However, other sequences with high Z values may induce only nearby sequencing errors, and they can be hardly used as filtering parameters. These filtering parameters should be used carefully. Strict filtering processes can also affect the estimation of the microbial community composition (Fig. 2). The patterns of SSEs were selectively removed by a quality filtering process. In particular, GC-rich microbes might be selectively removed in the Illumina systems by quality filtering. To accurately assess the structures of microbial communities, SSEs should be corrected,[9] and low scores should be used for quality filtering.

Another interesting issue concerning SSEs is related to the identification of polymorphisms in the GS Junior, FLX+, and PGM systems. Sequence-specific deletion errors can be confused with SNPs. Our newly discovered patterns of deletion errors in the GS Junior, FLX+, and PGM systems will be helpful for distinguishing true SNPs from sequence-specific deletion errors. Our findings provide a scientific basis for identifying true polymorphisms, filtering reads of poor quality, improving assembly work and SNP calling, and accurately assessing microbial community compositions and species identification.

Acknowledgements

The authors thank Dr. James Cole for his helpful comments. This work was supported by the Korea Ministry of Environment *via* a grant from The GAIA Project, and by the National Research Foundation (NRF) of Korea *via* a grant (No. 2011-0030040) funded by the Korea government (MSIP).

References

1 N. Mancuso, B. Tork, P. Skums, L. Ganova-Raeva, I. Măndoiu ons and A. Zelikovshy, *In Silico Biol.*, 2011, **11**, 237–249.
2 J. W. Davey, P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen and M. L. Blaxter, *Nat. Rev. Genet.*, 2011, **12**, 499–510.
3 T. Werner, *Briefings Bioinf.*, 2010, **11**, 499–511.
4 M. B. Scholz, C. C. Lo and P. S. Chain, *Curr. Opin. Biotechnol.*, 2012, **23**, 9–15.
5 J. A. Martin and Z. Wang, *Nat. Rev. Genet.*, 2011, **12**, 671–682.
6 B. Gharizadeh, A. Ohlin, P. Mölling, A. Bäckman, B. Amini, P. Olcén and P. Nyrén, *Mol. Cell. Probes*, 2003, **17**, 203–210.
7 D. R. Kelley and S. L. Salzberg, *Genome Biol.*, 2010, **11**, R28.
8 V. Kunin, A. Engelbrektson, H. Ochman and P. Hugenholtz, *Environ. Microbiol.*, 2010, **12**, 118–123.
9 S. Shin and J. Park, *Nucleic Acids Res.*, 2014, **42**, e51.
10 M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén and P. Nyrén, *Anal. Biochem.*, 1996, **242**, 84–89.
11 B. Gharizadeh, J. Eriksson, N. Nourizad, T. Nordström and P. Nyrén, *Anal. Biochem.*, 2004, **330**, 272–280.
12 M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. I. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley and J. M. Rothberg, *Nat.*, 2005, **437**, 376–380.
13 C. Quince, A. Lanzén, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read and W. T. Sloan, *Nat. Methods*, 2009, **6**, 639–641.
14 M. Kircher, U. Stenzel and J. Kelso, *Genome Biol.*, 2009, **10**, R83.
15 N. Whiteford, T. Skelly, C. Curtis, M. E. Ritchie, A. Löhr, A. W. Zaranek, I. Abnizova and C. Brown, *Bioinformatics*, 2009, **25**, 2194–2199.
16 J. C. Dohm, C. Lottaz, T. Borodina and H. Himmelbauer, *Nucleic Acids Res.*, 2008, **36**, e105.
17 K. Nakamura, T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M. C. Linak, A. Hirai, H. Takahashi, M. Altaf-Ul-Amin, N. Ogasawara and S. Kanaya, *Nucleic Acids Res.*, 2011, **39**, gkr344.
18 F. Meacham, D. Boffelli, J. Dhahbi, D. I. Martin, M. Singer and L. Pachter, *BMC Bioinf.*, 2011, **12**, 451.
19 M. Allhoff, A. Schönhuth, M. Martin, I. G. Costa, S. Rahmann and T. Marschall, *BMC Bioinf.*, 2013, **14**, S1.
20 V. Bansal, O. Harismendy, R. Tewhey, S. S. Murray, N. J. Schork, E. J. Topol and K. A. Frazer, *Genome Res.*, 2010, **20**, 537–545.
21 M. L. Davey, E. Heegaard, R. Halvorsen, M. Ohlson and H. Kauserud, *New Phytol.*, 2012, **195**, 844–856.
22 G. K. Alderton, *Nat. Rev. Cancer*, 2013, **13**, 220–221.
23 M. L. Metzker, *Nat. Rev. Genet.*, 2010, **11**, 31–46.
24 A. Gilles, E. Meglécz, N. Pech, S. Ferreira, T. Malausa and J. F. Martin, *BMC Genomics*, 2011, **12**, 245.
25 I. J. Good, *J. R. Stat. Soc. Ser. B*, 1967, **29**, 399–431.
26 S. Schbath, *J. Comput. Biol.*, 1997, **4**, 189–192.
27 S. Schbath, B. Prum and É. Turckheim, *J. Comput. Biol.*, 1995, **2**, 417–437.
28 F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening and R. A. Edwards, *BMC Bioinf.*, 2008, **9**, 386.
29 D. R. Zerbino and E. Birney, *Genome Res.*, 2008, **18**, 821–829.
30 B. Langmead and S. L. Salzberg, *Nat. Methods*, 2012, **9**, 357–359.
31 H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and 1000 Genome Project Data Processing Subgroup, *Bioinformatics*, 2009, **15**, 2078–2079.
32 J. A. Sleep, A. W. Schreiber and U. Baumann, *BMC Bioinf.*, 2013, **14**, 367.
33 W. L. Jorgensen and J. Pranata, *J. Am. Chem. Soc.*, 1990, **112**, 2008–2010.

