Systematic Substitution Errors Found in...

mlelivelt

Systematic Substitution Errors Found in MiSeq (R) Data 31 janv. 2012 09:32

Systematic Homopolymer Errors in MiSeq® Sequencer Data

Mike Lelivelt & Yutao Fu, Ion Torrent by Life Technologies

Semiconductor sequencing by Ion Torrent® offers post-light DNA sequencing using unmodified nucleotides and naturally occurring DNA polymerase. This biochemistry more closely resembles the DNA replication process that has evolved over billions of years. Illumina's MiSeq® DNA sequencer relies on nucleotides that contain modified fluorescent molecules that must be both excited (via expensive light sources) for detecting the base and then completely removed from the growing polymer chain so nucleotide incorporation can continue. This leads to the following question: does the partial or incomplete removal of these fluorescent molecules impact the accuracy of the MiSeq® sequencer?

Observation: The MiSeq® sequencer tends to mis-call bases adjacent to a homopolymer region, especially stretches of Cs and Gs, thus falsely extending the homopolymer length and causing strand-specific substitution errors.

To illustrate the phenomenon, a small region of the *E. coil* genome was loaded into a standard genome browser, such as the Broad's IGV, using both public MiSeq® and Ion Torrent® *E. coli* DH10B datasets[1,2]. Figure 1 (A&B) shows a particular region of DH10B genome in which substitution errors are concentrated in MiSeq® reads, but not Ion Torrent® reads (such genomic regions aren't difficult to find in DH10B or K12). A careful review of this data reveals an interesting pattern of errors tending to replicate the previous (5' upstream) base calls. Following a homopolymer stretch in MiSeq®, the base that follows this stretch tends to be erroneous, andthose errors tends to retain the preceding

base from the homopolymer stretch. These substitution errors often fall to the last base of a homopolymer region - based on the direction of the read. For example, in a stretch of three G bases, the fourth base is often erroneously called a G. This strand-specific pattern is wide spread, and explains 49.9% and 51.8% of MiSeq® substitution errors overall in DH10B and K12, respectively. This dominant error profile that can be found so frequently next to homopolymer regions suggests a clear systematic bias within MiSeq® data.

Measuring the trend across the whole genome – conditional probability

In order to measure this trend across a larger genomic region, the conditional probability p(C|Cx) was measured across homopolymer regions across all bases sequenced. This conditional probability is the chance of observing another C base following runs of x Cs, and similarly p(G|Gx) for G. In other words, it is simply likelihood that the next base in a sequence will be called the same as the previous base. A baseline conditional probability is calculated based on the naturally occurring genomic sequence.The closer the conditional probability of each sequencing technology is to the genome background value, the less bias exists. The conditional probabilities were calculated for *E. coli* K12 and DH10B genomes for both MiSeq® and Ion Torrent® publicly available data sets. Bases not aligned to the genome were excluded as they are disregarded from the perspective of calling variants. For both K12 and DH10B, MiSeq® datasets showed much larger conditional probability deviations from genomic background than the matching Ion Torrent® datasets, as shown in Figure 2A & 2B. The deviation grows dramatically as homopolymer length increases out to 5 bases. Similar but less severe MiSeq® bias was also observed for bases A and T.

Sequencing *E. coli* is an excellent control substrate to test the accuracy of DNA sequencing. However, researchers often want to see sequencing performance measured on human DNA. The conditional probabilities of both C and G bases were measured from publicly available human amplicon data sets from both MiSeq® [3] and Ion Torrent PGM® sequencers. The same bias of overcalling G/Cbias was also observed in the MiSeq® human amplicon dataset (Figure 2C). The trend was especially noticeable in long stretches of both G and C bases.

Hypothesizing why such a trend exists?

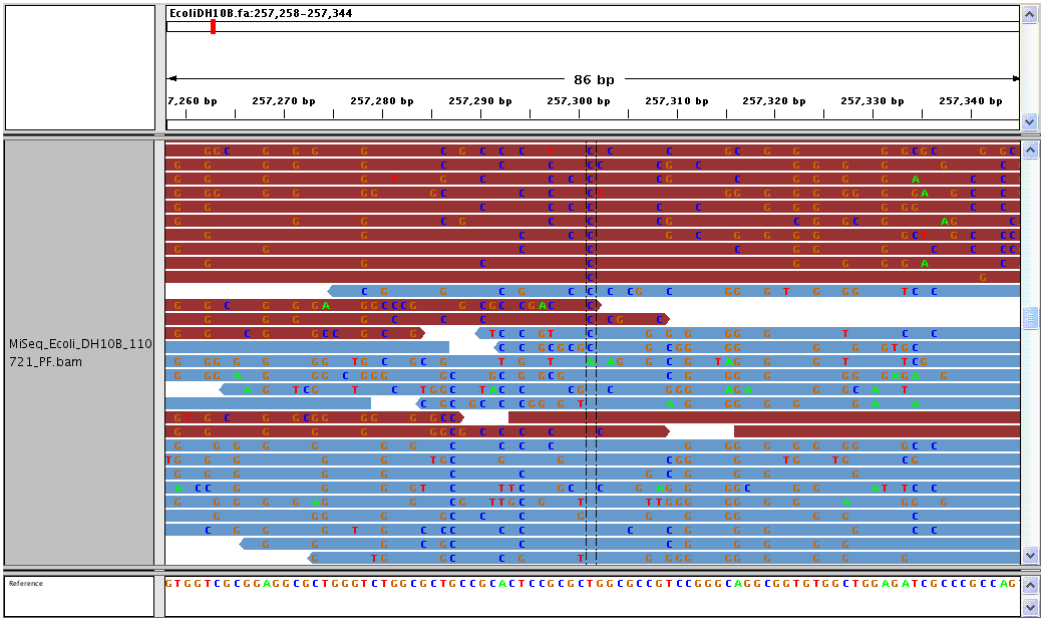
One hypothesis that fits nicely with the above observations: the fluorescent dyes used by MiSeq® sequencing are subject to incomplete cleavage and may accumulate across a homopolymer region. If fluorescent dyes are retained from the previous flow and then excited in the subsequent flow, a wrong base call is more likely to be produced. Such a symptom may be unique to reversible terminator based sequencing chemistry. Independent studies on Illumina sequencing data already suggested filtering out “variants flanked by homopolymers of length greater than 3 or surrounded by greater than 6 identical bases”[4].

Conclusion

Light-based DNA sequencing from MiSeq® Sequencer is only as good as the ability for the technology to uniformly remove the fluorescent moiety from the growing polymer at each flow cycle. The data available from multiple public releases of MiSeq® data suggest that MiSeq® data contains systematic errors associated with homopolymer regions and the impact of these errors is larger in the MiSeq® platform relative to the Ion Torrent PGM® platform especially in G/C homopolymer regions.

Figure 1. IGV screenshots of MiSeq® (panel A - top) and Ion Torrent® (panel B - bottom) datasets for a genomic region of *E. coli* DH10B

Systematic Substitution Errors Found in...



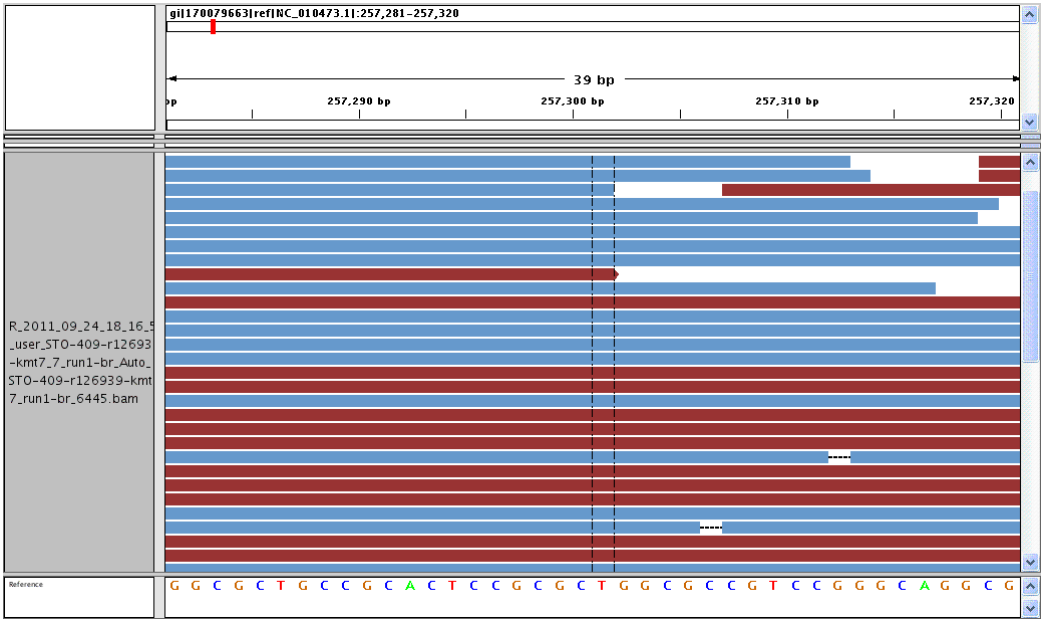
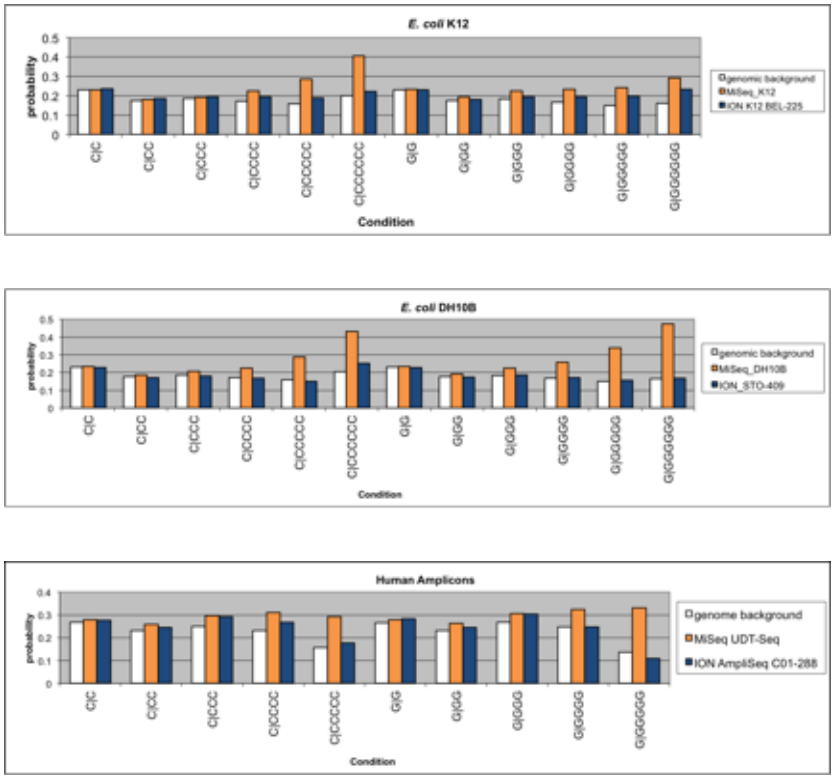


Figure 2. Conditional probabilities p(C|Cx) and p(G|Gx) for increasing homopolymer length x. Panels A (top) for E. coli K12, B (middle) for E. coli DH10B, and C (bottom) for Human Amplicons. For panel C, only the genomic background distribution for Ion targeted genome regions is shown. MiSeq background from UDT-Seq was calculated independently, as the composition of the amplicons between the panels is different. The difference between the background probabilities was not significantly different.



Methods:

BAM files for *E.coli* data were downloaded from Ion Community (http://lifetech-it.hosted.jivesoftware.com/community/torrent_dev, except for a K12 run from an R&D Torrent Server) and Illumina web pages (http://www.illumina.com/systems/miseq/scientific_data.ilmn). MiSeq® human amplicon reads were downloaded from NCBI SRA page for run SRR385941, converted to fastq format, and mapped to target regions to produce a BAM file. Both BWA and a C program in Torrent Suite, TMAP, were used for

mapping, and produced similar observations downstream. The parameters used for TMAP were:

tmap mapall -n10 -f targets.fasta -r SRR385941.fastq map1 map2 map3|samtools view -bS - | samtools sort - results

, and for BWA:

bwa aln -q 15 targets.fasta SRR385941.fastq > SRR385941.sai

bwa samse targets.fasta SRR385941.sai SRR385941.fastq > results.bam

The bam files were first parsed first using a C++ program alignStats, also available as a binary in Torrent Suite, with the following parameters:

alignStats -i <bam_file> -p 1

The resulting Default.sam.parsed files were used for conditional probabilities.

A PERL script ‘countpattern.pl’ (appendix 1) was used to search aligned portions of sequencing reads recorded in Default.sam.parsed files for HxM patterns, in which Hx is a homopolymer with x(x>=1) base of H (H=A,T,C or G), and M is a monomer base. The counts were tallied in Excel and conditional probabilities of homopolymer extension p(Hx+1|Hx) were inferred as the ratio between counts of {HxM|H==M} and {HxM}.

Appendix:

1. countpattern.pl:

```
#!/usr/bin/perl

$pattern = $ARGV[1]?$ARGV[1]:'CG';

$patternlen = $ARGV[2]?$ARGV[2]:length($pattern);

open in, "$ARGV[0]";

while (<in>)

{if (/>(.*)/) {push @s,$1;$s=$1;}

else {s/ //g; $s{$s}.=uc($_)}

}

close in;

for (@s)

{$seq = uc($s{$_});

pos $seq=0;

$count=0;

while ($seq=~/$pattern/g)

{$count++;

pos($seq) = pos($seq)-$patternlen+1;

}

$len = length($seq);

$len++ if $len==$patternlen-1;

print "$_\t",int($count/($len-$patternlen+1)*10000+0.5)/100,"\t$count\n";

}
```

__END__

#Example usage:

for a in C CC CCC CCCC CCCCC CCCCCC G GG GGG GGGG GGGGG GGGGGG;
do for b in A C G T;do echo -ne “$a\t$b\t” >> output.txt; countpattern.pl
sample.fasta ab|cut -f3 >> output.txt; done; done;

References:

1. http://www.illumina.com/systems/miseq/scientific_data.ilmn

2. http://lifetech-it.hosted.jivesoftware.com/community/torrent_dev (except a K12 dataset which is not yet released)

3. Olivier Harismendy, Richard B Schwab, Lei Bao, Jeff Olson, Sophie Rozenzhak, Steve K Kotsopoulos, Stephanie Pond, Brian Crain, Mark S Chee, Karen Messer, Darren R Link and Kelly A Frazer. Detection of low prevalence somatic mutations in solid tumors with ultra-deep targeted sequencing.

Genome Biology 2011, 12:R124

4. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK, Ding L. SomaticSniper: Identification of Somatic Point Mutations in Whole Genome Sequencing Data. Bioinformatics (2011) epub Dec 6, doi: 10.1093/bioinformatics/btr665

Balises : miseq_homopolymer

flxlex

Systematic Substitution Errors Found in MiSeq (R) Data 30 janv. 2012 11:27

Interesting results...

As DH10B is a substrain of E. coli K12, perhaps what you call 'K12' is E. coli K12 substrain MG1655? Could you please (please) make the run for this strain available? This would finally make a comparison with 454 E. coli data (always generated from MG1655) possible...

mlelivelt

Systematic Substitution Errors Found in MiSeq (R) Data 31 janv. 2012 23:43

K12 (MG1655) data is now available

http://lifetech-it.hosted.jivesoftware.com/docs/DOC-2492

markus_g

Systematic Substitution Errors Found in MiSeq (R) Data 31 janv. 2012 09:51

Hi,

Is this maybe related to this here?

Nakamura K, Oshima T, Morimoto T, et al. Sequence-specific error profile of Illumina sequencers. Nucleic acids research. 2011;39(13):e90.

The MIRA assembler could already handle these errors long before that paper was published.

lek2k

Systematic Substitution Errors Found in MiSeq (R) Data 31 janv. 2012 15:33

Hey thanks for the NAR reference. Interesting read..

yfu

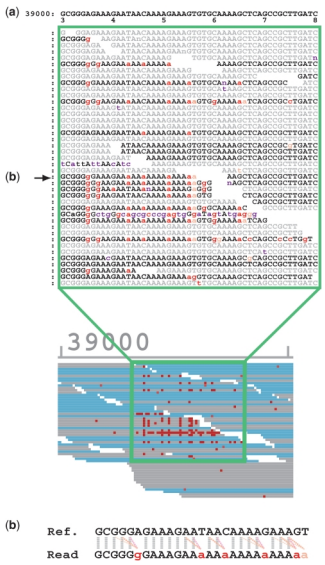
Re: Systematic Substitution Errors Found in MiSeq (R) Data 3 févr. 2012 07:48

Many thanks for the reference Markus. It does seem to be related to what we reported here, and more importantly, it represents knowledge from another independent party.

In the main text, Nakamura *et al.* referred to homopolymers in particular:

Interestingly, the mismatches tended to appear after a sequence of identical base calls in all SSE regions. In other words, the mismatched base was often similar to a preceding reference base.

Linked below is their Figure 4, exactly the same observation as in the MiSeq® datasets .



Also in Table 5 (linked below), Nakamura *et al.* reported 49-62% of mismatches involving homopolymers, on par with the 49.9% and 51.8% numbers we found for DH10B and K12:

Table 5.

Percentage of mismatches in SSE regions that match the reference base positioned 1–5 bp before the mismatch position

Species	1	2	3	4	5
Bacillus subtilis	61.2	19.7	7.4	3.5	1.9
Mycobacterium bovis	61.6	22.3	7.7	3.4	1.7
Staphylococcus aureus	48.9	20.9	9.7	5.5	3.5
Bordetella pertussis	54.4	20.6	8.7	4.3	2.7

Inspired by this paper, I went back to the MiSeq datasets, and found its homopolymer problem is indeed associated with upstream GGC motifs. GGC (or GCC on the reverse strand) is present in almost every genomic region where substitution errors are concentrated. I bet it is near the example lek2k showed in his blog too.

The authors offered an alternative hypothesis to our dye cleavage model. Essentially they proposed difficulties for labeled nucleotides to get on instead of for reverse terminators and fluorophores to get off. Nonetheless the apparent error mode is persistent in all sequencers with the same chemistry.

Power of open access and crowd sourcing!

lek2k

Systematic Substitution Errors Found in MiSeq (R) Data 31 janv. 2012 15:32

A similar but independent analysis performed on the MiSeq Bacillus cereus genome.

http://biolektures.wordpress.com/2012/02/01/are-miseq-miscalls-influenced-by-preceding-homopolymers/

qingqing.zhang@lifetech.com

Systematic Substitution Errors Found in MiSeq (R) Data 31 janv. 2012 23:16

Hi lek2k, nice blog.

I have a quesion. Is this kind of error strand specific? From the above figure1A, it seems this kind of error occur at both directions.

However, from the figure of Bacillus, it shows only in one direction, strand specific.

I'm a little confused why it should be strand specific. I think only position matters.

ypaquet

Systematic Substitution Errors Found in MiSeq (R) Data 1 févr. 2012 07:34

Qingqing,

It is strand-specific in the sense that this type of error seems to be influenced by the preceding stretch of Cs or Gs (and possibly other contextual factors), so it depends on the direction of synthesis and thus whether the sequenced template is from the plus or minus strand. The errors on the other strand in Figure 1A above are not necessarily linked to the same error bias (more typical of errors at the end of reads, for example).

mlelivelt

Re: Systematic Substitution Errors Found in MiSeq (R) Data 1 févr. 2012 08:28

We'd had a couple of requests for the BAM files from MiSeq that we created. Here are links to those files hosted on box.net.

Open Source. Open Comment. Let us know what anyone finds.

BWA BAM

TMAP BAM

lek2k

Systematic Substitution Errors Found in MiSeq (R) Data 1 févr. 2012 22:00

I was going to say exactly what ypaquet wrote. Probably not as elegantly (smiley face).

qingqing.zhang@lifetech.com

Systematic Substitution Errors Found in MiSeq (R) Data 1 févr. 2012 22:03

Thank you, lek2k and ypaquet:)

Simon Cawley

Re: Systematic Substitution Errors Found in MiSeq (R) Data 19 févr. 2012 11:56

There's a recent paper from Lior Pachter's group further exploring this topic - see

http://www.biomedcentral.com/1471-2105/12/451/abstract

They describe a new statistical method for detection of systematic errors and using it they find something similar to what is described in this thread.

Its a good read, check it out. I copy below a couple of the salient figures:

Figure 3 - Sequence motifs at systematic error sites

(a) The motif around systematic errors reveals a strong enrichment for instances preceded by an occurrence of GG and for the error to occur at locations where the reference genome is T. (b) Categorized by the nucleotide at the error location. The number of systematic errors in each subset is denoted by n.

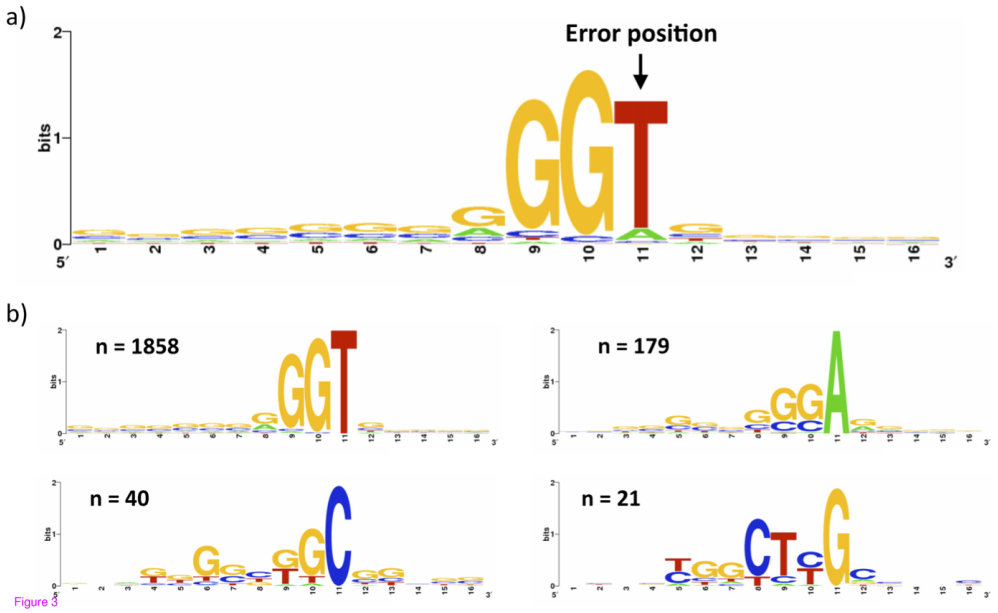
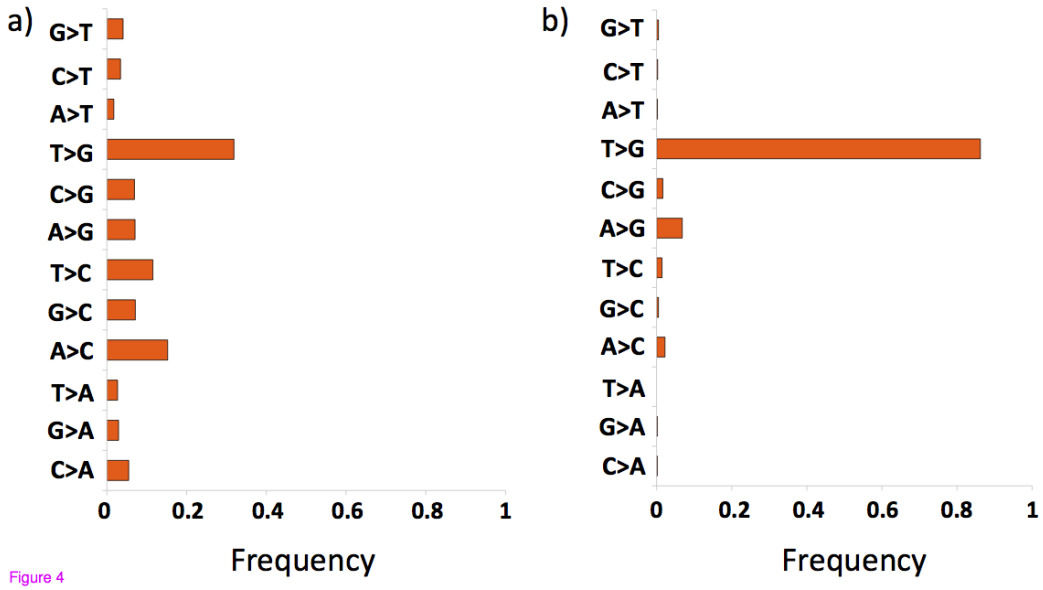


Figure 4 - Base substitutions of systematic errors

Frequency of different base substitutions in (a) all errors (b) systematic errors.



Edited, forgot to add in the image for figure 4 in the original post

flxlex

Systematic Substitution Errors Found in MiSeq (R) Data 22 févr. 2012 07:46

Hi,

I just came over this blog post from 2010. It also shows a homopolymer issue, but then particularly towards the ends of Illumina reads...

http://ivory.idyll.org/blog/jul-10/illumina-read-phenomenology

But, this is older data...