BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 14 2010, pages 1704–1707
doi:10.1093/bioinformatics/btq269

Genome analysis

Iterative Correction of Reference Nucleotides (iCORN) using second generation sequencing technology

Thomas D. Otto[1,*], Mandy Sanders[1], Matthew Berriman[1] and Chris Newbold[1,2,*]

[1]Parasite Genomics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA and [2]Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, OX3 9DS, UK

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: The accuracy of reference genomes is important for downstream analysis but a low error rate requires expensive manual interrogation of the sequence. Here, we describe a novel algorithm (Iterative Correction of Reference Nucleotides) that iteratively aligns deep coverage of short sequencing reads to correct errors in reference genome sequences and evaluate their accuracy.

Results: Using *Plasmodium falciparum* (81% A+T content) as an extreme example, we show that the algorithm is highly accurate and corrects over 2000 errors in the reference sequence. We give examples of its application to numerous other eukaryotic and prokaryotic genomes and suggest additional applications.

Availability: The software is available at http://icorn.sourceforge.net

Contact: tdo@sanger.ac.uk; cnewbold@hammer.imm.ox.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 8, 2010; revised on April 23, 2010; accepted on May 19, 2010

1 INTRODUCTION

Although there are now over 5000 whole genome sequences in the public databases, their level of accuracy varies considerably. The aspiration set by the Human Genome Project was for a maximum of one error per 10 kb of finished sequence (International Human Genome Sequencing Consortium, 2001). However, the true error rate varies significantly from this figure depending on the nature of the sequence (base composition, repeats, etc.) both in human and in other organisms. Even to achieve this error rate, expensive manual finishing is required to ensure that each base is covered by at least 2 clones and has a cumulative Phred score of at least 40 (Ewing and Green, 1998). The 'Gold Standard' for quality involves the manual inspection of each base by an experienced finisher. This is a major expense within a genome project. For example, nine chromosomes of *Plasmodium falciparum* were completed at the Wellcome Trust Sanger Institute (Hall *et al*., 2002), by the equivalent of approximately seven finishers working for up to 5 years. Despite this and subsequent efforts since publication, as we show here, many errors are still present. Even in genomes described as completed or finished, the underlying quality at each base is unknown and the error rate can be variable genome-wide. Therefore, rapidly fixing errors, highlighting regions that are error-prone and quantifying accuracy genome-wide is a priority that will significantly benefit the end user.

*To whom correspondence should be addressed.

So far, few methods exist to correct genomic errors automatically. There are algorithms to improve base calling (Gajer *et al*., 2004) or to detect frameshifts by protein homology or by sequence analysis. New assembly software like Mira (http://www.chevreux.org/projects_mira.html) has also been developed that allows hybrid assemblies with different sequencing technologies. This can both assemble mixed Sanger/454 data and improve the homopolymer length errors in 454 technologies using high Illumina read coverage. To date, however, no methods exist that can accurately detect and correct base errors and small indels in genome sequences.

We have developed an algorithm that uses deep coverage of sequence reads produced using Illumina's Genome Analyser platform, mapped iteratively to a reference genome, in a way that allows confident sequence correction.

2 METHODS

Due to their short length, mapping reads from second generation sequencing platforms is highly susceptible to single base errors or small indels. Small corrections made to a reference can, therefore, improve the mapability of short reads and, conversely, introducing small errors in a reference will markedly reduce mapability. We have made use of this fact in developing a new methodology to automatically correct base errors and short insertions or deletions (indels) of up to 3 bp. In an iterative process, short reads are mapped against the genome and high-quality discrepancies and indels are identified and corrected. In each iteration, we compare the coverage of perfectly mapping reads at each corrected base before and after correction. Corrections that reduce the read coverage at that position are rejected. In this way, we evaluate whether each potential correction is accurate or not. We repeat the iterations until no new corrections are called.

2.1 Data

For the *P.falciparum* reference sequence, we used 3D7 version 2.1.4 (ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/falciparum/3D7/3D7.version2.1.4/). All Illumina data were produced within the Sanger sequencing facility. The protocol to obtain PCR-free data is described in Kozarewa *et al*. (2009). Preparation of other samples can be seen in Quail *et al*. (2008).

2.2 Iterative Correction of Reference Nucleotides Implementation

An overview of Iterative Correction of Reference Nucleotides (iCORN) is given in Figure 1. The program itself is hosted at http://icorn.sourceforge.net/. Short reads are first mapped with SSAHA2 (Ning *et al*., 2001) against the genome sequence that is to be corrected (although another mapping algorithm could be used, e.g. Li *et al*., 2008). Standard Illumina mapping

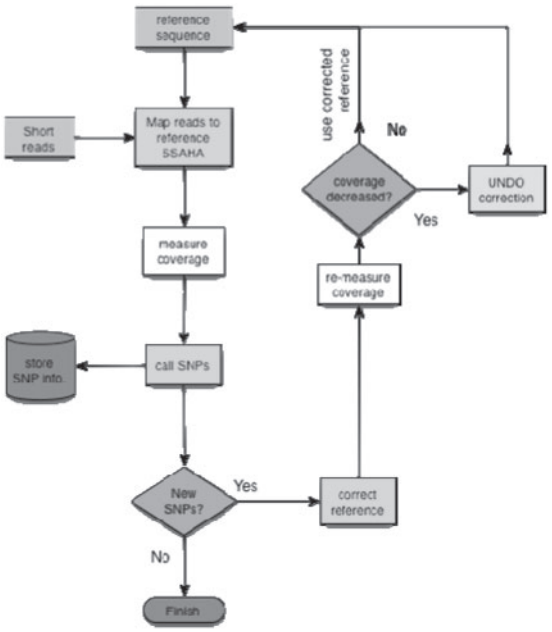

Fig. 1. Flow chart of iCORN.

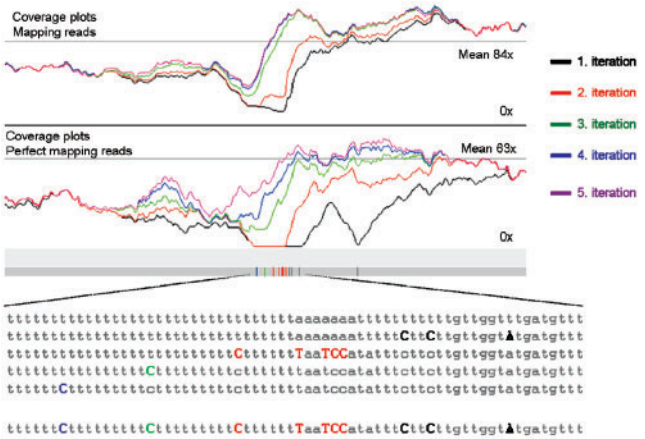

Fig. 2. Example of correction of a region of chromosome one of *P.falciparum* 3D7. The upper plot shows the coverage per iteration of the SSAHA mapping. The lower plot represent the coverage of the perfect mapping reads SNP-o-matic (http://snpomatic.sourceforge.net/). The vertical bars show the positions of the corrections. The actual corrections made at each iteration are shown in the multiple sequence alignment below.

values are used with the 'paired' option when reads are paired. Read pairs that do not map within the correct insert size constraint, map to different chromosomes or are in the wrong orientation, are ignored. Using the SSAHA pileup pipeline (ftp://ftp.sanger.ac.uk/pub/zn1/ssaha_pileup/), single nucleotide polymorphisms (SNPs) and short indels (1–3 bp) are called from the remaining read pairs. Note that, each 'SNP' or 'indel' called by the software refers to potential sequencing errors or sample heterogeneity. A SNP is accepted if it has a SSAHA SNP quality of at least 60. Short indels are called if they occur in at least 30% of the reads with a minimum read coverage of at least 5. These parameters are the standard values but can be changed. The called SNP and indel errors are corrected in the genome sequence and saved as a new version.

To evaluate the corrections, the coverage of each base before correction is compared to that after correction using SNP-o-matic (Manske and Kwiatkowski, 2009) that only maps reads mapping perfectly over their whole length. If read coverage of a corrected base goes down, the change is rejected and the original sequence is restored. If there is no change in coverage, we assume that this region may have additional errors and accept the correction.

The procedure is repeated, using the newly corrected genome sequence as the reference and continues to iterate until no new errors can be found. The algorithm returns all changes, including coverage statistics, in GFF format [visible with Artemis (Carver *et al*., 2008)] or as Gap4 feature file.

3 RESULTS

To calibrate the SSAHA2 alignment score threshold to detect real base errors but minimize false positives, all calls on a single chromosome at different calling thresholds were manually inspected by an experienced professional finisher. This involved interrogating the capillary reads and their quality scores in a GAP4 database. Using a SSAHA SNP score of 60 (reflecting a base coverage of ≥20) resulted in all of the corrections being confirmed. This score was subsequently adopted for future analyses.

We first applied our algorithm to the genome sequence of *P.falciparum*: a reference sequence whose low complexity results from an extremely biased base composition (19% G+C content) and presents a challenge to short read alignment algorithms due to its exceptionally low information content. For the analysis, we used 28 million 36 bp paired-end and 20 million 76 bp paired-end Illumina reads. The mean coverage obtained by mapping read pairs with the correct fragment size was 82.9× and we used a minimum coverage of 20-fold to call changes (see Section 2). An example of a corrected region can be seen in Figure 2. The coverage plots show that the amount of mapping reads and perfectly mapping reads increase with each iteration.

After 6 iterations, no new corrections were called. We found a total of 1906 base errors and 368 indels. In the first iteration, 81% of the base errors were corrected. After the corrections, the coverage of 84 827 more bases pairs increased to at least 5 and 87 952 additional reads were perfectly mapped. For most chromosomes, the single base error detection drops to zero after the fifth iteration (Supplementary Table S1). Intitially, we found 208 sites (SNP score ≥60) that appear to be heterozygous in this haploid organism, using a cutoff of 15% of calls of an alternative base. Visual inspection of the areas in which these heterozygous calls occurred, however, revealed that the majority (75%) were roughly symmetrically distributed around homopolymeric tracts. The remainder appears to be strand specific as they only occur in one read direction and are clustered in general in sequences that are rich in T and G bases. These calls were not present in the original capillary sequence data and we believe them to be hitherto unreported systematic errors occurring during Illumina sequencing (Supplementary Fig. S1).

Ninety-six percent of the genome is covered by a read depth of ≥20, so that we were unable at this level of confidence to correct the remaining 4% of the genome. These regions are mostly telomeric and non-unique.

To test the accuracy of our algorithm, we randomly introduced approximately one error per 50 kb into the 3D7 sequence 2.1.4, inserting a total of 457 errors and used the Illumina reads to correct this altered genome using iCORN. In the first iteration, 435 (96%) of the errors were found (Supplementary Table S2). As the errors were generated randomly, they were not clustered

Table 1. Application of iCORN to prokaryotic and eukaryotic genome projects in various stages of completion

Organism	Sequence quality	Sequencing method	Genome size (Mb)	SNPs	Indels	Number rejected	Genome covered		New mappable reads	Iterations
							Before (%)	After (%)		
Plasmodium falciparum 3D7	A	Capillary	23	1906	368	30	97.20	97.56	24698	6
Echinococcus multilocularis	B	Capillary	110	5508	2520	2140	48.89	49.11	1023315	5
Leishmania major	B	Capillary	33	594	1061	122	98.52	98.62	313	6
Leishmania infantum	B	Capillary	32	2770	1878	320	89.26	89.72	5629	8
Plasmodium ovale	B	Capillary	21	1431	238	1081	91.27	91.42	6368	4
Plasmodium berghei	B	454	18	25976	33860	5639	88.65	95.38	140788	7
Plasmodium berghei	B	Capillary	22	1901	3818	538	97.18	97.48	23805	7
Chlaymiadia trachomatis	B	Capillary	1.0	487	16	18	99.86	99.997	9734	4
Clostridium difficile	B	454	4.1	61	1652	32	99.30	99.43	1708	6
Streptococcus pneumoniae	B	RNAseq	2.0	13	5	1	64.23	64.23	6	3
Streptococcus suis BM402	A	Capillary	2.1	2	1	0	98.84	98.85	15	2
Streptococcus suis P1_7	A	Capillary	2.0	0	0	0	99.7626		0	1
Salmonella Dublin Strain	B	454	5.0	13	45	18	96.84	96.85	207	7
Yersinia enterocolitica	B	Capillary	5.0	25	235	6	99.96	99.97	131796	3

Sequence quality: 'A' indicates manually finished and published genomes and 'B' indicates a draft assembly. SNPs and Indels shows the total number called between the first and last iteration. Rejected indicates the total number of changes that were rejected because they decreased the total of perfectly mapping reads at that location. Percent genome covered indicates how many bases are covered at least five times by perfectly mapping reads, before and after the correction. New mapable reads indicates the additional number of reads that could be mapped by SSAHA between the first and last iteration. Further information can be found in Supplementary Table S3.

and could be found quickly. The random distribution also explains that 4% of the introduced errors cannot be found, as 4% of the genome is not covered sufficiently to be corrected, (Supplementary Table S1).

We further evaluated the performance of iCORN by manually inspecting the capillary chromatograms of called errors in chromosomes 5, 9 and 14. Of 174 corrected errors, 1 was rejected. This region comprised a string of 45 As with a G in the middle and was re-sequenced following polymerase chain reaction (PCR) amplification. This confirmed the presence of the G that had been erroneously corrected by iCORN to an A. We suspect that this may be due to the fact that polyA sequences are over represented in Illumina data because of occasional edge effects on the slide. Finally, we designed an additional 96 PCR products over regions with correction. Eighty-eight out of 96 PCR reactions were successful and in no case did the PCR product sequence disagree with the changes called by iCORN.

We next went on to assess the utility of this approach to correct the homopolymer errors that can occur using 454 technology (Droege and Hill, 2008). We applied it to a 454 assembly of 310 242 reads (fragment size 3 kb) from *P.berghei*. The contigs from the assembly were corrected with ~50 million 76 bp PCR-free paired-end Illumina reads. After 6 iterations, 25 976 SNPs and 33 860 indels were called (Table 1). Figure 3A shows a typical example where multiple frameshifts due to homopolymer errors are corrected after just two iterations. Figure 3B shows similar data from the correction of a 454 assembly of *Clostridium difficile* using deep Illumina coverage. In both cases more indels than SNPs are called due to homopolymer errors.

Finally, we applied iCORN to a series of other eukaryotic and prokaryotic genome projects in various stages of completion (Table 1). For finished bacterial genomes, very few or no corrections were made. For those in draft assembly, it was possible to call a number of errors in relatively few iterations. This is presumably because bacteria generally have higher coverage, are shorter and have a less complex genome structure than eukaryotes. All errors in *Yersinia enterocolitica* and *Streptococcus suis* were confirmed by manual inspection of the trace files.

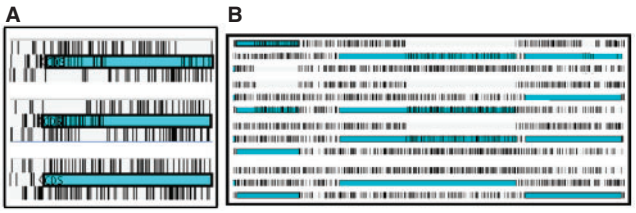

Fig. 3. Examples of corrections of homopolymer length errors in assemblies from 454 sequencing. Details of the reads used can be found in Table 1. Figures are Artemis screen shots that show the three different reading frames in the direction of the gene. Black vertical lines are stop codons. Filled coloured boxes denote open reading frames. (**A**) Correction of a region of an assembly of *P.berghei* 454 reads. (**B**) Correction of a region of a 454 assembly of *C.dificile*.

4 DISCUSSION

Here, we have shown that iterative mapping of short reads can correct errors remaining in a reference genome with great accuracy. Critical to the success of this approach is the use of two different mapping strategies during the iterations. High-quality discrepancies called using SSAHA2 are introduced into the genome and only confirmed if a separate mapping of perfectly aligning reads along their whole length using SNP-o-matic does not decrease coverage at the altered sites. Iterative mapping approaches have been used before to derive a consensus genome sequence from metagenomic sequencing data (Dutilh *et al.*, 2009) but since this derives from aggregated sequences from an unknown number of starting

genotypes, the resulting consensus represents no single genome and hides much of the diversity present in the original sequence pool.

We have also shown that, after very few iterations, iCORN is efficient at correcting homopolymer errors that are often present in 454 data, thus potentially improving the ability to combine assemblies constructed using different sequencing technologies.

We have explored the use of iCORN to 'morph' a reference genome into a closely related genotype using deep short read coverage. Although this approach may produce erroneous sequence changes, we have found that it has been very successful in improving the mapping of assembled contigs from a new genotype onto a reference genome.

Finally, with third generation sequencing technology on the horizon, bringing gigabase coverage from much longer read lengths but with an increase in error rates, the use of additional Illumina reads and algorithms such as iCORN may be of considerable use in first-pass error correction.

ACKNOWLEDGEMENTS

We would like to thank Stephen Bentley and Nicholas Thomson for Illumina sequences to test iCORN and Heidi Hauser and Danielle Walker for manually checking the corrections in *Y.enterocolitica* and *S.suis*.

Funding: Wellcome Trust (grant number WT085775/Z/08/Z); European Union 6th Framework Program grant to the BioMalPar Consortium (grant number LSHP-LT-2004-503578).

Conflict of Interest: none declared.

REFERENCES

Carver,T. *et al*. (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, **24**, 2672–2676.

Droege,M. and Hill,B. (2008) The Genome Sequencer FLX™ System—longer reads, more applications, straight forward bioinformatics and more complete data sets. *J. Biotechnol*., **136**, 3–10.

Dutilh,B.E. *et al*. (2009) Increasing the coverage of a metapopulation consensus genome by iterative read mapping and assembly. *Bioinformatics,* **25**, 2878–2881.

Ewing,B. and Green,P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*., **8**, 186–194.

Gajer,P. *et al*. (2004) Automated correction of genome sequence errors. *Nucleic Acids Res*., **32**, 562–569.

Hall,M. *et al* (2002) Sequence of *Plasmodium falciparum* chromosomes 1, 3-9 and 13. *Nature*, **419**, 527–531

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Kozarewa,I. *et al*. (2009) Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. *Nat. Methods*, **6**, 291–295.

Li,H. *et al*. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Manske,H.M. and Kwiatkowski,D.P. (2009) SNP-o-matic. *Bioinformatics*, **25**, 2434–2435.

Ning,Z. *et al*. (2001) SSAHA: a fast search method for large DNA databases. *Genome Res*., **11**, 1725–1729.

Quail,M.A. *et al*. (2008) A large genome centre's improvements to the Illumina sequencing system. *Nat. Methods,* **5**, 1005–1010.