BIOINFORMATICS APPLICATIONS NOTE

Vol. 26 no. 1 2010, pages 127–129
doi:10.1093/bioinformatics/btp615

Sequence analysis

inGAP: an integrated next-generation genome analysis pipeline

Ji Qi†, Fangqing Zhao†, Anne Buboltz and Stephan C. Schuster*

Pennsylvania State University, Center for Comparative Genomics and Bioinformatics, University Park, Pennsylvania 16802, USA

Received on July 10, 2009; revised on October 20, 2009; accepted on October 23, 2009

Advance Access publication October 30, 2009

Associate Editor: Alex Bateman

ABSTRACT

Summary: We develop a novel mining pipeline, Integrative Next-generation Genome Analysis Pipeline (inGAP), guided by a Bayesian principle to detect single nucleotide polymorphisms (SNPs), insertion/deletions (indels) by comparing high-throughput pyrosequencing reads with a reference genome of related organisms. inGAP can be applied to the mapping of both Roche/454 and Illumina reads with no restriction of read length. Experiments on simulated and experimental data show that this pipeline can achieve overall 97% accuracy in SNP detection and 94% in the finding of indels. All the detected SNPs/indels can be further evaluated by a graphical editor in our pipeline. inGAP also provides functions of multiple genomes comparison and assistance of bacterial genome assembly.

Availability: inGAP is available at http://sites.google.com/site/nextgengenomics/ingap

Contact: scs@bx.psu.edu

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

1 INTRODUCTION

The rapid development of promising new parallel sequencing technologies, known as Roche/454, Illumina and ABI/SOLID (Mardis, 2008a, b), has dramatically changed the nature of genetic studies, covering a wide range in high-throughput *de novo* genome sequencing, from microorganisms to living or ancient mammals and human genome re-sequencing and producing hundreds of thousands of reads with continuously decreasing cost. These huge amount of reads can be used for varied purposes (Mardis, 2008a, b), among which single nucleotide polymorphism (SNP) identification is a common interest as these technologies provide the highest resolution.

Most current methods (Trapnell and Salzberg, 2009) focus on data from only specific sequencing platform and use only either Illumina and ABI/SOLID data, like SOAP (Li *et al*., 2008b) and MAQ (Li *et al*., 2008a), or Roche/454 data (Brockman *et al*., 2008). None of them can integrate sequencing data from different platforms. The pipeline we describe here, Integrative Next Generation Genome Analysis Pipeline (inGAP), is designed for this purpose. Its workflow is shown in Figure 1.

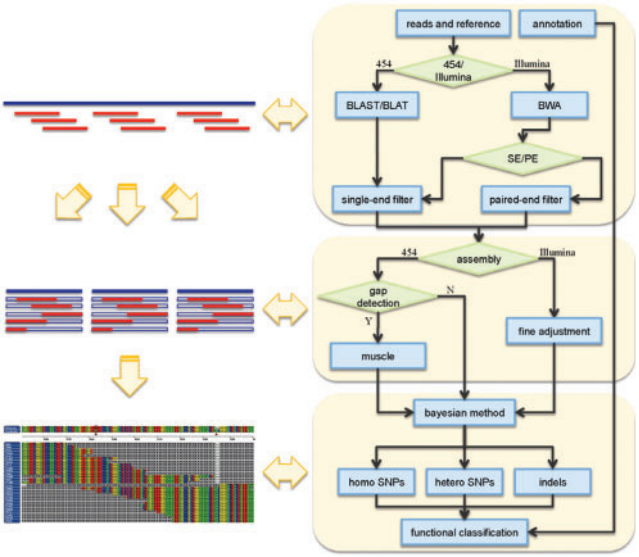

Fig. 1. A description of SNP/indel calling workflow of inGAP. First, assigning reads to a reference genome. Second, precise multiple alignments are performed for gapped regions. Third, a Bayesian algorithm is used to call SNPs and indels.

inGAP can detect SNPs and indels by comparing sequence data generated by either Roche/454 and/or Illumina sequencing technologies, with a reference sequence, regardless read lengths and numbers. Furthermore, it can deal with various genomes from prokaryotes to eukaryotes and detect genetic variations from highly divergent reads against distantly related reference genomes. Extensive evaluations on both simulated and real datasets show that inGAP detects 97% and 94% of SNP and indels, respectively. Additionally, we have incorporated genome assembly and multiple genome alignment softwares into inGAP. To make inGAP user-friendly, all detected nucleotide changes can be searched and further edited, and genome assembly and multiple genome alignment operations can be completed using a graphical viewer.

2 METHODS

As most popular technologies in genome sequencing, Roche/454 and Illumina are quite different on reads length and reads numbers they produce. The read length from Roche/454 is much longer than that from Illumina and is more difficult to handle.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

inGAP maps Illumina reads to reference genomes through BWA (Li and Durbin, 2009) by default. For Roche/454 data, reads mapping and gap opening are performed by BLASTN (Altschul *et al*., 1990) (for high-divergent mapping) or BLAT (Kent, 2002) (for close related mapping), MUSCLE (Edgar, 2004) is then applied to these mapped reads to obtain detailed multiple alignment for further assembly as shown in Figure 1.

After Roche/454 and/or Illumina reads are assembled, consistency between reads and reference are checked for initially indentifying candidates of SNPs and then evaluated by a Bayesian method. SNPs candidates passing the evaluation are classified into synonymous/non-synonymous/non-coding ones when annotation information is available (see Supplementary Material for details).

3 RESULTS

3.1 Detecting SNPs and indels from simulated datasets

To test the performance of our approach, we used inGAP to simulate 75 bp Illumina reads with different coverage (from 5× to 100×) and various levels of divergence (from 0.1% to 1%) from a *Helicobacter pylori J99* genome (NC_000921). Simulated results are shown in Figure 2. We also incorporated 1% substitutions and 0.2% indels to mimic sequencing errors (see Supplementary Fig. 1). We first tested whether genetic divergence between the target and reference genomes could affect the performance of SNP calling. Approximate 0.4 million 75 bp reads (20×) with 0.1–1% point mutations and 0.02–0.2% indels (ranging from 1 bp to 10 bp) were simulated from the complete genome of *H.pylori J99*. As shown in Figure 2A, with the increase of genetic divergence, the sensitivity of SNP and indel detectionusing inGAP remain relatively constant as compared with those using MAQ. Under a lower divergence level (0.1% point mutations and 0.02% indels), both inGAP and MAQ can identify >98% SNPs, and inGAP can also identify >94% indels with a high accuracy (99.4%, Fig. 2B). We then used this divergence rate to evaluate the performance of inGAP and MAQ under different coverage (Fig. 2C and D). inGAP performs slightly better than MAQ on the sensitivity of detecting SNPs, but exhibits a much higher PPV. Compared to MAQ, inGAP can detect short indels from single-end date sets. Moreover, it also outperforms MAQ in SNP prediction, because MAQ tends to falsely predict SNPs from gapped regions. We also used simulated 454 reads to evaluate the performance of inGAP, and found a more promising result as longer reads can span certain short repeat regions. As shown in Figure 2E and F, SNP discovery rate reaches 94% when using 5× coverage 454 data and sharply increases to 98% under 10× coverage. It should be noted that inGAP could identify 98% of short indels (1–10 bp) with 99% accuracy using 454 reads.

By further exploring the undetected SNPs, we found that nearly half of them located in repeat regions and the other half was missed due to a low quality of multiple sequence alignment. SNPs located in repeat regions can be partially recalled by using less stringent filtering parameters (e.g. minimum alignment identity or matched read length) or longer reads. SNPs missed by false alignments can be recovered by manually editing the problematic alignment. inGAP provides a user-friendly graphical interface for checking and editing predicted SNPs.

Moreover, owing to its robustness and flexibility in mapping more divergent reads, we extended inGAP to assemble repetitive element from fragmented short reads. Supplementary Figure 2 illustrated an assembly of a 3.5 Kb LINE/RTE element from 454 sequenced mammoth genome sequences, where the RTE-2_MD from the opossum genome was used as a reference. In this way, users can easily build a consensus sequence for each type of repetitive elements. As described in our early study (Zhao *et al*., 2009), we have successfully built the consensus sequences for various types of interspersed repeats in the mammoth genome.

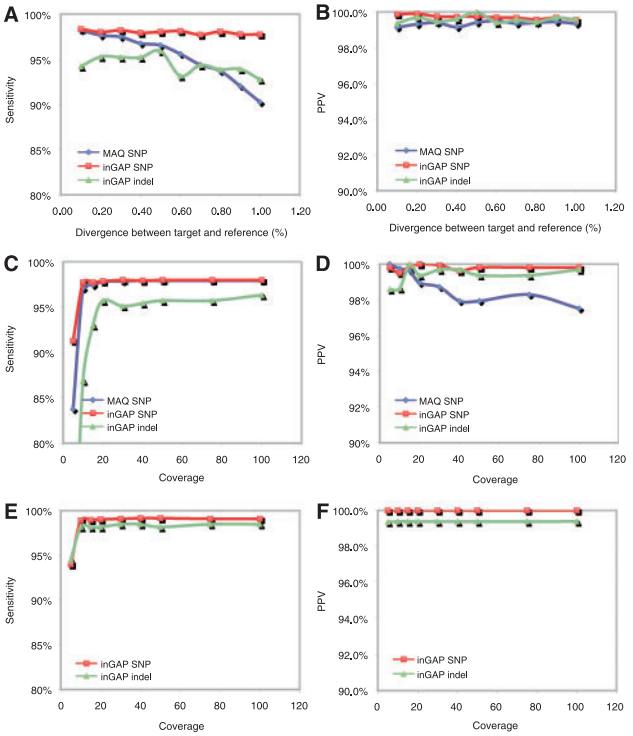

Fig. 2. Performance comparison between MAQ and inGAP on simulated datasets. (**A**) Sensitivity on SNP/indel calling on different levels of divergent Illumina reads. Green line shows the indels (1–10 bp) identified by inGAP. (**B**) Positive predictive value (PPV) comparison between MAQ and inGAP based on different divergent Illumina reads. (**C**) Performance on simulated Illumina reads with coverage ranging from 5× to 100×. (**D**) PPVs on simulated Illumina reads under different sequence coverage. (**E**) Performance on simulated 454 reads with coverage ranging from 5× to 100×. (**F**) PPVs on simulated 454 reads under different sequence coverage.

3.2 Application of inGAP in large-scale eukaryotic genomes

Various real datasets have been used to evaluate the performance of inGAP. The application of inGAP on eight strains of *Salmonella* Typhi is shown in Supplementary Table 1 (Holt *et al*., 2008).

We used a combined data from both Roche/454 and Illumina sequencing technologies to investigate crossover and gene conversion in yeast meiosis (Qi *et al*., 2009). We identified over 46 000 single nucleotide differences between the two budding yeast strains, from which 91 crossovers and 21 gene conversions have been detected in four meiotic products of one tetrad.

To handle even larger reference genome (e.g. human chromosomes), we suggest BWA as the reads aligner, which enable inGAP to map 10 million 35 bp Illumina reads on human chromosome 1 within 30 min and 2 Gb memory on a 8-core DELL machine.

4 DISCUSSION

inGAP is the first platform that allows users to evaluate the genetic variation of a sample, that contains multiple types of next-generation sequencing data. It can also help with completing genome assembly and comparative genome analysis. inGAP outperforms other software for the following aspects. (1) It does not have any read length restriction. It can handle 454 sequencing and/or Illumina sequencing and/or Sanger sequencing datasets. (2) Besides SNPs, it can detect most small indels in either single- or paired-end datasets. (3) It has a strong capability to identify variants based on a relatively divergent reference genome, which brings it to a much wider application other than re-sequencing projects. (4) It provides a user-friendly graphic interface, through which users can browse, search, check, classify and even edit the identified variants.

ACKNOWLEDGEMENTS

We greatly appreciate Dr Webb Miller and Aakrosh Ratan (Penn State University) for thoughtful readings of the manuscript. We thank Huabin Hou (Wenzhou Medical College, China) for suggestions and software testing.

Funding: Gordon and Betty Moore Foundation (to S.C.S.).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Brockman,W. *et al*. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.*, **18**, 763–770.

Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Holt,K.E. *et al*. (2008) High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. *Nat. Genet.*, **40**, 987–993.

Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

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

Li,R. *et al*. (2008b) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.

Mardis,E.R. (2008a) Next-generation DNA sequencing methods. *Ann. Rev. Genomics Hum. Genet.*, **9**, 387–402.

Mardis, E.R. (2008b) The impact of next-generation sequencing technology on genetics. *Trends Genet.*, **24**, 133–141.

Qi,J. *et al*. (2009) Characterization of meiotic crossovers and gene conversion by whole-genome sequencing in Saccharomyces cerevisiae. *BMC Genomics*, **10**, 475.

Trapnell,C. and Salzberg,S.L. (2009) How to map billions of short reads onto genomes. *Nat. Biotechnol.*, **27**, 455–457.

Zhao,F. *et al*. (2009) Tracking the past: interspersed repeats in an extinct Afrotherian mammal, Mammuthus primigenius. *Genome Res.*, **19**, 1384–1392.