Genome resequencing and identification of variations by Illumina Genome Analyzer Reads

Jun Wang
Beijing Genomics Institute at Shenzhen
Ruiqiang Li
Beijing Genomics Institute at Shenzhen
Yingrui Li
Beijing Genomics Institute at Shenzhen
Xiaodong Fang
Beijing Genomics Institute at Shenzhen
Binxiao Feng
Beijing Genomics Institute at Shenzhen
Jun Li
Beijing Genomics Institute at Shenzhen

Method Article

Keywords: resequencing, SNP, indel, structural variations, Illumina Genome Analyzer

Posted Date: November 18th, 2008

DOI: https://doi.org/10.1038/nprot.2008.238

Abstract

Introduction

Illumina Genome Analyzer \(GA), as one of the new-generation sequencing instruments, can produce large amount \(typically 3G per paired-end run) of short reads \(30-50bp in length) in several days. The high-throughput is suitable for genome resequencing, which requires deep coverage to detect genomic polymorphisms such as single nucleotide polymorphisms \(SNP's), insertion/deletion events \(indels) and structural variations \(SV's). However, the new technology also challenges bioinformatics tools due to intensive computing resource requirements. Traditional alignment and mutation detection pipelines are mainly designed for capillary sequencing, which are not appropriate for new-generation sequencing technology as they may consume unacceptable memory or disk space and cost too long time to finish. Therefore, we have developed a new protocol, which takes full use of characteristics of Illumina GA reads, for genome resequencing and mutation detection. The new protocol runs over ten times faster than traditional method and achieves a high accuracy on detecting polymorphisms.

Equipment

FASTQ files which are generated by Illumina Genome Analyzer \(GA) and data analysis pipeline. The FASTQ files contain read sequences and quality scores in ASCII characters. Hardware requirements: CPU: 64-bit CPUs RAM: 16G main memory or larger HDD: 1T free hard disk space or larger Software requirements: Operation system: A Linux system \(kernel version >=2.6.9) or other up-to-date POSIX systems C compiler: GNU Compiler Collection \(version >=4.2.1) or other compilers that are compatible with ANSI C standard. Perl: version >= 5.7.6 Python: version >=2.4

Procedure

Read Alignment 1. Concatenate chromosome sequences of reference assembly of NCBI build v36 into a single reference sequence file in FASTA format \(hereafter, referred to as "Reference.fasta"). 2. Align Illumina GA reads from each Illumina GA lane to the reference sequence by software SOAP1. To take use of more data, we recommend to add command line parameter "-c 52" to trim low-quality bases in alignment process. Typically, the command lines are as follows: For single-end \(SE) data: soap -a <Reads.fastq> -d <Reference.fasta> -o <Alignment.soap> -p <# of parallel processes> -c 52 -s 12 For paired-end \(PE) data: soap -a <Reads1.fastq> -b <Reads2.fastq> -d <Reference.fasta> -o <PEalignment.soap> -2 <SEalignment.soap> -p <# of parallel processes> -c 52 -s 12 -m <maximum insert size> -x <minimum insert size> 3. Sort alignment result of each lane, first by chromosome names lexicographically, then by mapping coordinates on each chromosome numerically. This could be done by various kinds of sorting tools. 4. Merge the sorted alignment results of all lanes into a single file, keeping the alignments sorted. 5. Split the sorted and merged alignments, chromosome by chromosome, into different files, each file comprising alignments of a single chromosome. **Building Consensus Sequence** 6. Download flat ASN files of dbSNP2 database and allele frequency information if possible,

such as HapMap3 data for human resequencing. Transform dbSNP information of each chromosome into a tab-delimited plain text file \(hereafter, referred to as “chrN.dbSNP.txt”), which looks like: chr1 201979756 1 1 0 0.161 0 0 0.839 rs568 The columns from left to right mean: name of chromosome, coordinate on the chromosome, whether the SNP has external allele frequency information \(1 is true, 0 is false), whether the SNP is a validated dbSNP \(1 is true, 0 is false), whether the SNP is actually an indel \(1 is true, 0 is false), frequency of A, frequency of C, frequency of T, frequency of G, refSNP ID. For dbSNP sites that do not have allele frequency information, the frequencies can be arbitrarily determined as any positive values, which only imply what alleles have already been deposited in the database. 7. Based on the alignment result, build consensus sequence for each chromosome by SoapSNP. Typically, the command line are as follows: SoapSNP -i <Alignment.soap.sort.chrN> -d <chrN.fasta> -o <chrN.consensus> -r 0.00005 –e 0.0001 -t -u -L <Maximum Read Length> -M <chrN.mat> -s <chrN.dbSNP.txt> -2 The result of SoapSNP “chrN.consensus” has 17 columns: chromosome ID, coordinate on chromosome, reference genotype, consensus genotype, quality score of consensus genotype, best allele, average quality score of best allele, count of uniquely mapped best allele, count of all mapped best allele, second best allele, average quality score of second best allele, count of uniquely mapped second best allele, count of all mapped second best allele, sequencing depth of the site, rank sum test p_value, average copy number of nearby region, whether the site is a dbSNP \(0 is false; 1 is true). These kinds of information would be used in SNP extraction process. **SNP extraction** 8. Extract potential SNP sites from each consensus files \(“chrN.consensus”), where the called genotypes are different from reference one, and the reference genotypes are not “N”. For each potential SNP sites, record its distance to the neighboring potential SNP. 9. Filter the raw SNP dataset to get the confident SNP’s that satisfy the following criteria: a) the quality score is larger than 20; b) each of the allele is supported by more than a certain number of reads \(depending on the average sequencing depth, typically 2 for 20X data); c) the overall depth should be less than average plus 3 *standard deviation of sequencing depth; d) the average copy number of nearby region of the SNP is less than 2; e) each allele is supported by at least one paired-end reads; f) the site is at least 5bp away from another SNP. **Identification of indels** 10. Align all paired-end reads to the reference genome using SOAP in gap alignment mode: soap -a <Reads1.fastq> -b <Reads2.fastq> -d <Reference.fasta> -o <PEalignment.soap> -2 <SEalignment.soap> -p <# of parallel processes> -c 52 -s 12 -m <maximum insert size> -x <minimum insert size> -g 3 -e 5 11. Extract all gapped alignments from the SOAP result and merge them into a single file. 12. Sort all gapped alignments, first by chromosome names lexicographically, and then by coordinates on each chromosome. 13. Extract coordinates, sizes and numbers of supporting reads of potential indels from the alignments. 14. Select only one representative indel according to their numbers of supporting reads in each 10bp window. Thus indels are at least 10bp away from each other. 15. Filter all potential indels to get the confident ones that satisfy the following criteria: a) the indels are supported by at least 3 reads; b) number of ungapped alignment that cross the indels are no more than twice that of gapped reads. **Structural Variation Detection** 16. Extract all abnormally aligned paired-end reads, which have unexpected orientations and/or unexpected span size. Normally mapped read pairs are aligned in a forward-reverse pattern, i.e. the upstream read of a mapped pair is on the forward strand, and the downstream one is on the reverse strand. 17. Cluster abnormally aligned paired-end reads that have same

alignment orientation and similar coordinates \(distance between coordinates of two read pairs is smaller than paired-end insert size) on both ends of read pairs. 18. Fit all abnormal paired-end clusters into alignment models \(Figure 1) and call potential structural variations \(SVs). 19. Merge redundant called SVs. 20. Filter potential structural variations to get confident ones which satisfy the following criteria: a) each of the paired-end clusters that support the SV comprise at least 4 read pairs; b) the SV is not conflict with other SVs, otherwise they should be merged into a complex structural variation case.

Timing

Resequencing a human genome with 20 folds \(20X) data requires around 100 2.0 GHz 64-bit CPUs to run a week.

Critical Steps

Step 2: When aligning reads to the reference genome by SOAP, please set proper seed size \(parameter “-s”) for certain read lengths \(see SOAP manual). Current Illumina GA generally produces reads that are longer than 27bp. A seed size of 12 is appropriate. On trimming issue, if the data quality is good enough, we may not set -c 52 to align reads with relatively lower quality as they compose merely a very small proportion of all data, and they need much more CPU hours to align. **Step 7:** The program SoapSNP would create a file “chrN.mat” to store a quality calibration matrix. While rerunning the program, we may reuse the matrix by replace this parameter by “-I <chrN.mat>”, which would directly read the matrix to the memory. It is strongly recommended that using dbSNP information to refine SNP call in this step. Specify “-s <chrN.dbSNP.txt> -2” to set larger prior probability for alternative allele at known dbSNP sites. When processing monoploid chromosomes, such as chrX and chrY in a human male, specify SoapSNP command line switch “-m” to make sure all genotypes are called as homozygous. **Step 14:** Some indels may have ambiguous coordinates as they are in a short range of local repeats. This may cause difficulty in alignment process, and a single indel may be reported to locate at two or more coordinates. For those indels which are close to each other \(distance <10bp), we only choose the one which have the most supporting reads as the “representative” indels. **Step 19:** In all alignment models of Figure 1, some may share same pattern of abnormally mapped paired-end clusters. For example, a single deletion and a duplication both will result in forward-reverse mapped read pairs with unexpected large insert size. In this situation, we should merge the “simple” SV \(which has less kinds of supporting PE clusters) into a more complex SV.

References

1. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. _Bioinformatics_ **24**, 713-4 \(2008). 2. Sherry, S. T. et al. dbSNP: the NCBI database of genetic variation. _Nucleic Acids Res_ **29**, 308-11 \(2001). 3. Frazer, K. A. et al. A second generation human haplotype map of over 3.1 million SNPs. _Nature_ **449**, 851-61 \(2007).