Poster H-64

Precisely Mapping Structural Variations in the Human Genome by Splitting Shotgun Reads

Authors:
Zemin Ning (*The Wellcome Trust Sanger Institute*)

Short Abstract: We have developed a new computational method to detect structural variations by splitting shotgun reads against the reference sequence. A total number of 7,293 structural variants have been identified: 2,500 deletions, 2,358 insertions and 2,435 VNTRs, using 44 million shotgun reads from 10 different individuals

Long Abstract:

There is a growing interest in studying structural variations among humans. Hundreds of submicroscopic copy-number variants and inversions have been identified and it was reported and the sequences involved contained sometimes entire genes and their regulatory regions, stretching millions of DNA bases. However, the array CGH studies1, reported in the literature lacked the sequence leve [l] precision on breakpoints and also the survey was only on a small fraction of the sequence. The in silico strategy using pair-ending fosmids [2] achieved higher resolution, but it still cannot, in most cases, provide exact loci for breakpoints, neither for a solution to detect variants less than 5 kb. Identification and characterization of all types of structural variations remain a great challenge in the genomic community.

We have developed a computational method to precisely identity structural variants across the genome by aligning shotgun reads against the reference human sequence. As individual reads covering the boundaries of variation regions are split, this enables us to pinpoint the exact breakpoint loci as well as to extract sequences involved between the boundaries if applicable. DNA samples for the shotgun reads used in this analysis were from 10 different human individuals which gave a wealth of resources and diversity in studying structural variations in the human genome. Two types of the reads were used, flow sorting whole chromosome shotgun (WCS) reads sequenced by the International HAPMAP project and whole genome shotgun (WGS) reads by Celera. There are 5 individuals from each data set. Among the HAPMAP data of 16,593,859 reads, there are 13,398,084 reads were sequenced by the Sanger Institute, covering 10 chromosomes (1, 6, 9, 10, 11, 12, 13, 20, 22, X, respectively). The reads by other sequencing centers were from other chromosomes, except chromosomes 4 and 5. The Celera reads accounted for 27.45 million and the DNA samples were taken from 5 individuals with one dominant individual Celera_HuBB making 70% of the total reads (19.39 million). Before aligned to the reference sequence, quality clipping and vector screening were carried out to remove low quality bases and vector segments of possible contaminations. We used SSAHA2 [3] as the alignment tool to place genomic reads on the human genome assembly (NCBI35, May 2004).

A total number of 7,293 pieces of structural variants have been identified: 2,500 deletions, 2,358 insertions and 2,435 VNTRs, using 44 million shotgun reads from 10 different human individuals. Compared with one existing database [4] of structural variations, there are 230

exact matches among ~900 L1 retrotransposon polymorphisms. For the detected variants, we also selected 300 segments for experimental validation by PCR.

[1] Sebat, J. et al. (2004) Large-scale copy number polymorphism in the human genome. Science 305: 525–528.
[2] Tuzun E. et al. (2005) Fine-scale structural variation of the human genome. Nat. Genet. 37: 727–732.
[3] Ning, Z., Cox, A.J. & J.C. Mullikin. (2001) SSAHA: a fast search method for large DNA databases. Genome Res. 11: 1725-1729.
[4] Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P. (2006) dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum Mutat. 27: 323-329.