

High-throughput identification of structural variations from sequence trace data

Daniel C Koboldt, Raymond D Miller

Department of Genetics, Washington University School of Medicine, 4566 Scott Avenue, Box 8232, St. Louis, MO 63110, United States of America

Introduction

Structural variants have recently become a popular area of research in the biomedical research community. In particular, reports of widespread copy number variation (Iafrate et al. 2004; Sebat et al. 2004) and large deletions (McCarroll et al 2005; Conrad et al 2005) revealed that structural variation affects a substantial portion of the human genome and is likely to play a significant role in human disease. While comparative genome hybridization (CGH) and SNP genotyping technologies have been adapted to detect structural variants, doing so at the nucleotide level remains a challenge. We have developed a method, based on the BLAST algorithm, that allows for the high-throughput identification of structural variation from sequence trace data.

Methods & Materials

Each set of sequence traces was used to create a BLAST database against which the reference genome was aligned using WU-BLAST. The BLAST results were parsed and filtered to identify reads that appear to span a molecular lesion or "breakpoint". Alignment positions and orientations were analyzed to characterize the size, type, and precise molecular boundaries of the apparent structural variant.

Results

We tested the method using data from the model organism *C. briggsae*. From a database of 13,632 traces we identified 966 breakpoint reads representing 901 unique arrangement events. The results were further analyzed with CrossMatch and filtered for indel events. Some 208 were identified (46 read-through insertions and 164 deletions). We chose a test set of 12 indels, and designed PCR assays for them. Nine of the 12 PCR assays were successful under uniform conditions and each of the nine reflected the predicted indel.

Conclusions

Our method is both efficient and sensitive, and when applied to the wealth of human sequence trace data, should offer a powerful approach for identification of structural variants in the human genome.