Algorithm for identification of small genomic insertions

Next Generation Sequencing technologies are able to produce large amounts of sequence data at the cost of reducing the read length and increasing the error rate. One main challenge, under short read and high error rate, is to identify small insertions and deletions (from two to tens of nucleotides). There is a great demand for algorithms that can accurately identify small mutations. However, computational approaches to find these mutations using DNA fragment libraries or mate-pair libraries have so far been unsuccessful.

This technology provides an algorithm for identifying nucleotide insertions and deletions in a DNA fragment and enables identification and reconstruction of small insertion/deletion mutations from libraries of small DNA fragments.

Computational method more accurately identifies genomic insertions

This technology presents an algorithm capable of producing a list of candidates of small genomic insertions and their nucleotide sequences. Given a set of reads that are partially aligned to a reference genome, this method can find the possible positions of genomic insertions relative to the reference genome, through multiple statistical analyses. In a test, the chromosome X samples of 12 adult male patients with T-cell acute lymphoblastic leukemia have been analyzed and mutations in various genes, including small insertions and deletions, have been identified and verified via traditional Sanger sequencing. Continuing research is on further development of this method, particularly with the genomic data from cancer samples, obtained by Next Generation Sequencing systems.

Lead Inventor:

Raul Rabadan, Ph.D.

Applications:

- Identify small insertions and deletions associated with diseases, e.g. cancer
- Especially useful in finding inserted genomic sequences larger than two nucleotides and smaller than 20% of the length of the reads
- Can be used to re-sequencing organisms
This algorithm could make Next Generation Sequencing more precise and relevant, due to its ability to sift through errors.
This technology could potentially aid in drug discovery and development.
This algorithm could potentially apply to other data mining operations, e.g., for identifying small but persistent anomalies in weather, traffic, and purchasing patterns.

Advantages:

- The first method capable of identifying small genomic insertions using fragment libraries
- Can be used to improve the accuracy of Next Generation Sequencing technologies
- Can extract more information from only fragment libraries, without additional information

Patent Information:

Patent Pending
Tech Ventures Reference: IR 2642

Related Publications:


Inventors

Raul Rabadan