Revealing multiple nucleotide polymorphysms and indels in homopolymeric regions in human exome data

T. Nagy[1], A. Guffanti[2], K. McElreavey[3], S. Juhos[4]
[1] Omixon Ltd, Budapest, Hungary
[2] Genomnia Ltd, Milan Italy
[3] Pasteur Institute, Paris, France

www.omixon.com

Introduction

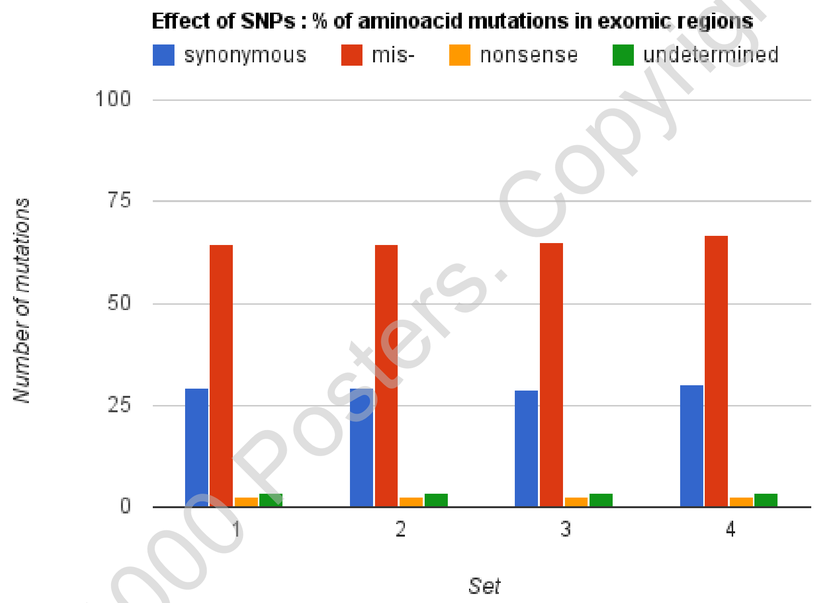
While detecting single nucleotide polymorphisms (SNPs) is the common goal of many studies, double or multiple nucleotide polymorphisms (DNPs and MNPs i.e. mutation of two or three adjacent nucleotides) are generally less well researched. Using next-generation sequencing it is relatively difficult to detect MNPs, insertions or deletions (indels) that are in homopolymeric or repetitive regions. Short-read aligners usually tolerate only a few SNPs in a read and have difficulties finding adjacent mismatches in color space data. Insertions and deletions are notoriously hard to detect in homopolymers. Using a new evolution-based read-alignment approach we were looking for these sort of variations in human exome data.

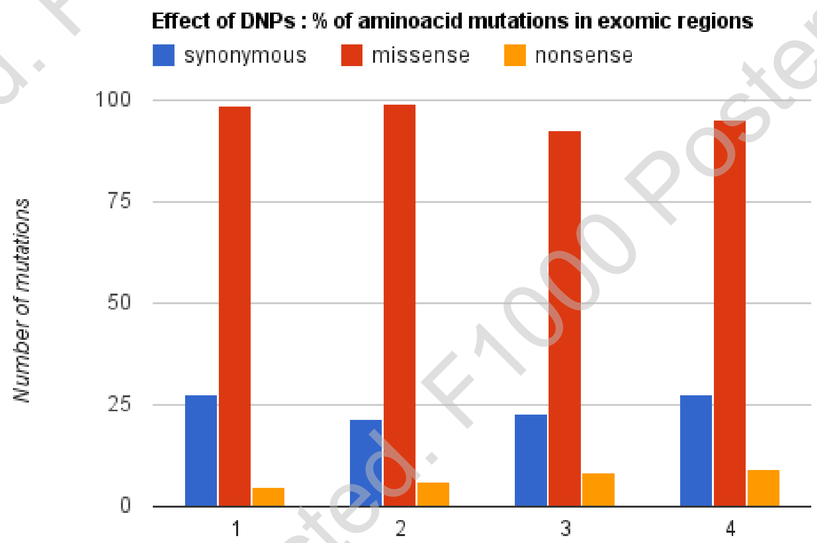
Methods and Results

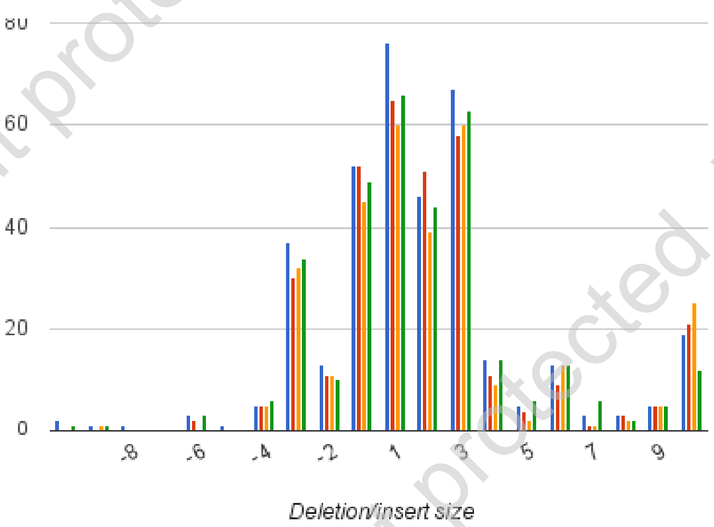
Four human samples were sequenced using the Agilent full exome enrichment kit and SOLiD NGS sequencing. The resulting 4x160M 50bp long color-space reads were aligned with a sensitive method described elsewhere [Csűrös et al. 2010] that is applicable for finding DNPs and indels.
Aligned reads were re-aligned around known indels using GATK [DePristo 2011] with snpDB v132 and variants were called also using GATK. Low quality variant calls (qual<50) were discarded.

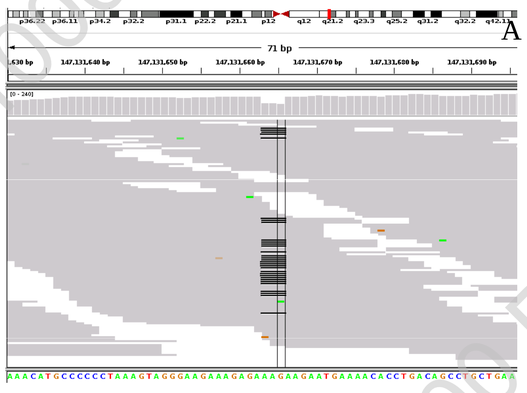
Set	Number of SNPs	Number of DNPs
1	44607	141
2	37481	127
3	35606	109
4	35606	119

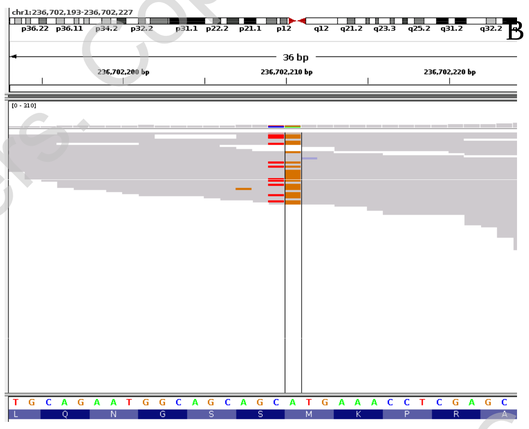
Set	Total number of indels	Indels in homopolymers
1	621	193
2	479	152











A: Heterozygous deletion in a repetitive region
B: Putative double nucleotide substitution changing two codons on one chromosome and only one on the other

Conclusions

We were able to find double nucleotide substitutions in the human exome using NGS data with the appropriate alignment methods. These mutations are showing similar characteristics to those found by other studies [Rosenfeld 2010] in similar data. Most of the putative MNPs are causing missense mutations. Insertions and deletions in homopolymeric regions contribute significantly to the total number of indels

References:

M. Csűrös, S. Juhos, A. Bérces: Fast mapping and precise alignment of AB SOLiD color reads to reference DNA WABI'10 Proceedings of the 10th international conference on Algorithms in bioinformatics Springer-Verlag, 2010

J. A. Rosenfeld, A. K. Malhotra1, T. Lencz: Novel multi-nucleotide polymorphisms in the human genome characterized by whole genome and exome sequencing . Nucl. Acids Res. (2010) 38 (18): 6102-6111. doi: 10.1093/nar/gkq408

DePristo, M., Banks, E., Poplin, R., Garimella, K., Maguire, J., Hartl., C., Philippakis, A., del Angel, G., Rivas, M.A, Hanna, M., McKenna, A., Fennell, T. Kernytsky, A., Sivachenko, A, Cibulskis, K., Gabriel, S., Altshuler, D. and Daly, M. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics. 2011 Apr; 43(5):491-498.