Algorithms for tagging and recognizing a large set of samples in highly parallel 454 sequencing

Vicario S[1], Calabrese C[3], Santamaria M[1], Simone D[2], Attimonelli M[2]

Motivation

Here is reported a bioinformatics pipeline which allows users sorting a large scale of different tagged samples, in highly parallel 454 sequencing. In this work we started from 52 human mitochondrial DNA samples, tagged with a PCR reaction and pooled in an amplicon library for a GS FLX Titanium Series sequencing. Our coverage requirements were about a eighth of a run (~ 81.7Mbp). The large number of samples to be sequenced in a small fraction of the plate, made us impossible to implement the approach proposed by Roche, that includes only 12 sequence tags (MIDs) usable for each of the 16 gasket regions. Thus we carried out a new system which allows us to design a number of tags sufficient for our aim, with minimal length and misidentification risk. For the subsequent bionformatics analysis of 454 sequence data, we developed an algorithm allowing to cluster all sample-specific reads in the same group, sorting each sequence by its specific tag. This python script, named “TagFinder”, scans, disregarding homopolymers, the 5’end of a read, searching for two types of strings: the sample-specific tag, 8 bases long and, downstream, the primer sequence variable in length.

Methods

Tag’s design We defined a set of tags, fitting with GS FLX Titanium Series requirements, taking into account the following parameters: tag length, presence/absence of homopolymers, spacing distance among sequence tags. We evaluated tag sets based on the rate of misidentification, assuming a binomial model. The binomial model was performed, in case of lack of homopolymer, considering a single sequencing substitution error percentage (p) equal to 0.68 [1]. We also defined the number of trials (n) equal to the tag length and the number of successes (k) & to the spacing distance. This approach was formalized with python script. Tag’s recognizing TagFinder sorts sequence data by their sample-specific tag and primer. The script performs assignments taking into account two error thresholds users defined, referred both to tag and primer matches and using the primers set and the tags, selected among the ones produced by the above described tags design algorithm. Furthermore, to evaluate the accuracy of this

[1] Consiglio Nazionale delle Ricerche - Istituto di Tecnologie Biomediche - Via Amendola 122D, I-70126 Bari, Italy [2] Department of Biochemistry and Molecular Biology "E. Quagliariello" - Bari, 70126, Italy [3] Unità di Genetica Medica, Policlinico Universitario S. Orsola-Malpighi, Università di Bologna, 40138 Bologna, Italy

approach, we calculated the number of true/false positives and false negatives on a subset of the samples, readily identifiable from the sequence itself, and considering the tag’s assignments only.

Results

The best set of designed tags is composed by 105 non homopolymeric octamers, with a spacing distance of 4 nucleotides, on the basis of which the percentage of misidentification is resulted to be 10-5. From this data set, we chose 52 tags, added to the 5’ end of sample-specific primers for the tagging with the PCR reaction. Upon preparing the amplicon library, the GS FLX Titanium Series sequencing was carried out. We obtained 121122 reads which were submitted to TagFinder. We obtained 101442 Assigned reads. In order to retrieve also those with the 5’end sequence tag cut (for a sequencing error), we applied TagFinder also to the reverse-complement data set of NotAssigned, taking advantage of the 3’end sequence tag complete. Thus, TagFinder assigned 110185 sequences (91%) to their specific tag, among which the number of false positives and false negatives, estimated on a subset of the obtained data, corresponds respectively to 1% and 4%.

Contact e-mail

saverio.vicario@ba.itb.cnr.it, claudia.calabrese23@gmail.com, dome.simone@gmail.com

Supplementary information

[1] Margulies M et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376-380.