PGP Statement of Work

Background

The Personal Genome Project (PGP) aims to annotate and publish the complete genomes and medical records of group of volunteers.

GenomeQuest is a commercial informatics company providing genomic data management solutions to industry and academia for searching, managing, mining, and sharing genomic information discoveries.

The purpose of this pilot study is to establish the scientific foundation for the PGP to leverage the scalability of the GQ bioinformatics platform. The project is designed to build a working relationship between the PGP team and the GQ team and to demonstrate that the GQ platform is able to quickly and accurately map the Illumina reads back to the reference exons and accurately detect SNPs over the covered areas.

On success of the pilot, the goal is for the PGP team and GQ to expand the scope of their collaboration to provide a platform for advanced informatics services to the PGP team and the broader research community the PGP wishes to serve.

Statement of Work (SOW)

During the first step in the PGP project, the exons from 10 individuals were sequenced using an Illumina machine. These exons were extracted by a unique method relying on the usage of padlock probes to capture a specific region of the genome.

In the initial stage of the project, GQ and the PGP bioinformatics group will have a back-and-forth dialogue to work out the best combination of sensitivity/specificity parameters to be used in the alignment and SNP detection. Thereafter, those parameters will be used to run up to 10 Illumina datasets and identify the SNPs for them.

Services description

1. Data source

The Illumina reads are delivered as flat-text files containing the sequence information (the query data-bases). The query databases will contain 2 million reads per sample, each 36 bp in length.

2. Data set cleaning

a. Trimming (optional): The query databases will be compared against a known primer sequence used in the experiment (the adaptor and/or the padlock probes). The adaptor is expected at the 3' end of the read and can vary in length. The comparison algorithm will produce a local alignment allowing one mismatch, insertion or deletion. This information will be used to remove the adaptor from the reads (the trimmed query data-base). All reads smaller than 18 nucleotides after removing the adaptor sequence will be discarded.

b. Quality score filtering (optional): to be defined through discussions between GenomeQuest and the PGP team.

3. Data set re-annotation

The alignment algorithm will try to align the entire read allowing one mismatch, insertion or deletion.

a. The trimmed query database will be aligned against the human transcript databases (refseq,genbank) using the HS3 alignment suite. Each read will be annotated with the number of alignments it had on the mRNAs. The sequence ID and description of the best match (associated gene names) will be attached to each individual read.

b. All reads not mapping to the transcripts will be compared against the non-masked human genome using the HS3 alignment suite. Each read will be annotated with the number of alignments it had on the genome, and the chromosomal positions.

c. All reads not mapping to the transcript and genome databases will be compared against the most up-to-date bacterial and viral divisions of Genbank. Reads will be classified as (a) bacterial, (b) viral, or (c) still unknown. As well, the sequence ID and description of the best match will be attached to each individual read (the annotated unknown read database).

4. Target mapping

The trimmed query database will be aligned against the relevant database of human exons. The best database to be used will be defined together with the PGP team. The alignment will be done using the HS3 suite using a gapped alignment method. The best 10 alignments for each read will be kept. Reads for which the 2 best alignments have a difference of less than 2 mismatches will be discarded. The exact parameters for the alignment will be defined after some preliminary testing.

5. SNP detection

The alignments will be used to detect the SNPs over the exons using the GQ SNP detection workflow.

PGP Deliverables

1. The Illumina reads as fastQ files containing the sequence information. Reads coming from separate samples should be labeled as such, or ideally come in separate files.

2. The primer / adaptor sequence mentioned in step 2 of the services description.
3. Results of the matching MAQ run in order to compare the results.
4. Exon database used to design the padlock probes

GenomeQuest Deliverables

1. Workflow details and gross statistics, (e.g., total counts of reads that move through the workflow at each step).
2. The trimmed and annotated query databases in flat-text FASTA format.
3. Three (3) GQ Seats to use the GQ Live application (said seats to remain available throughout the term of this Collaboration)
4. Access to the database of reads with their exon alignment inside the GenomeQuest web application.
5. Walkthrough of the GenomeQuest application to demonstrate the results of the workflow and to highlight GenomeQuest sorting, filtering, grouping, and high-throughput analysis capabilities for further examination of workflow results.
6. Exon coverage statistics for each exon hit by reads in the experiment.