Bioinformatic workflow for Chr16 characterization using proteomic shotgun and transcriptomic RNA-seq experiments

Maríana B. Monteiro¹, Daniel Tabas-Madrid², Elizabeth Guruceaga¹, Salvador Martínez-Bartolomé³, Juan Pablo Albar³, Fernando Corrales¹, Alberto Pascual-Montano², Víctor Segura¹

(¹ Unit of Proteomics, Genomics and Bioinformatics, Center for Applied Medical Research, CIMA, University of Navarra, ProteoRed-ISCIII, Pamplona, Spain.
² Functional Bioinformatics Group, National Center for Biotechnology (CNB-CSIC), Madrid, Spain.
³ Proteomics Facility, Centro Nacional de Biotecnología-Consejo Superior de Investigaciones Científicas (CNB-CSIC), ProteoRed-ISCIII, Madrid, Spain.

Spanish team of the Human Proteome Project (Sp-HPP) marked the annotation and data analysis of Chr16 genes as one of the priorities for its bioinformatics group. Four cell lines (CCD18, MCF7, Ramos and Jurkat) were selected to cover the chromosome 16 proteome, and shotgun proteomic and microarray transcriptomic experiments were performed to characterize their molecular profiles. The information available in ENSEMBL, UniprotKB and GPMDB databases was used for the analysis of the generated data. In order to improve the identification of coding genes and properly define the so-called “missing” proteins (coding proteins that have not been detected by any experimental procedure according to the criteria established by the HPP consortium) we propose the incorporation of public RNA-seq experiments provided by the Iluminated Human Body Map database for the bioinformatic analysis pipeline of Chr16. A new annotation map of the Chr16 was developed based on gene and protein databases in addition to the information obtained from RNA-Seq data analysis.

I. UPDATE OF CHROMOSOME 16 ANNOTATION

Chromosome 16 contains 886 protein coding genes (ENSEMBL v70) including 669 known and 187 missing proteins. The number of missing proteins is calculated as the difference between the total protein coding genes in Chr16 and the number of protein coding genes with experimental evidence at the level of protein in at least one of the following databases, according to the criteria adopted within the C-HPP: (1) naProProt (release Nov 2012), gold proteins mapping 642 protein coding genes; (2) Human Peptide Atlas (release Dec 2012), canonical proteins mapping 423 protein coding genes; (3) GPMdb (release 26 Nov 2012), green proteins mapping 630 protein.

II. BIOINFORMATIC ANALYSIS OF RNA-SEQ DATA FROM HUMAN BODY MAP

Public datasets from the Iluminated Human Body Map Project (HBM) were used. HBM integrates transcription profiles from high throughput sequencing data of 16 human tissues. All the selected samples were processed using the same pipeline: (1) downloaded sra files were converted into fastx files and the quality of the samples was verified using FastQC; (2) preprocessing of reads included elimination of contaminant adapter substrings with Sickle and quality-based trimming using Sickle; (3) alignment of reads to the human genome (hg19) was performed using Tophat2 mapper; (4) transcript assembly and quantification using FPKM of genes and transcripts was carried out with Cufflinks2; (5) annotation of the gene locus level of protein in at least one of the following protein coding genes with experimental evidence at the level of protein coding genes in Chr16 and the number of missing proteins is calculated as the difference between the total protein coding genes in Chr16 and the number of protein coding genes with experimental evidence at the level of protein in at least one of the following databases, according to the criteria adopted within the C-HPP: (1) naProProt (release Nov 2012), gold proteins mapping 642 protein coding genes; (2) Human Peptide Atlas (release Dec 2012), canonical proteins mapping 423 protein coding genes; (3) GPMdb (release 26 Nov 2012), green proteins mapping 630 protein. All the selected samples were processed using the same pipeline: (1) downloaded sra files were converted into fastx files and the quality of the samples was verified using FastQC; (2) preprocessing of reads included elimination of contaminant adapter substrings with Sickle and quality-based trimming using Sickle; (3) alignment of reads to the human genome (hg19) was performed using Tophat2 mapper; (4) transcript assembly and quantification using FPKM of genes and transcripts was carried out with Cufflinks2; (5) annotation of the gene locus was performed using Cuffmerge with Ensembl v70 and ENCODE v15 as references. Further analysis and graphical representations were performed using R/Bioconductor.

III. EVALUATION OF HBM RESULTS

The number of transcripts detected in the HBM ranged from 39870 isoforms recovered in liver sample to 160710 transcript assemblies obtained in lymph nodes sample. The mean number of transcripts for the 16 tissues was 88578, including known and novel transcripts. The mean number of known isoforms in Ensembl70 in the tissues sequenced in the HBM was 29101 transcripts. Similar calculation was performed for novel isoforms and we found a mean number of redundant novel transcripts for the 16 tissues of 10518 transcripts.

Interestingly, the distribution of FPKM in the 16 tissue samples of HBM comparing known and missing protein genes show statistically significant decrease (p-value<0.05) in the expression of genes coding for Chr16 missing proteins compared to expression levels of genes coding for Chr16 known proteins in most of the analysed tissues.

IV. INTEGRATION OF PROTEOMIC AND TRANSCRIPTOMIC PROFILES

To characterize in detail the proteome of Chr16, four cell lines were selected: MCF7, CCD18, Jurkat and Ramos. Shotgun proteomics were conducted in parallel by 6 independent laboratories resulting in 24 2D-LC-MS/MS and gel-LC-MS/MS experiments. Raw MS and MS/MS data were translated to mascot general file (mgf) format and searched against the UniprotsSwissProt human database (release 2012_06, June 13) using an in-house Mascot Server. False Discovery Rates (FDR ≤1% at the protein level) for protein identification were manually calculated.

The number of protein coding genes detected in all cell lines was 100, and the number of protein detected in only one of the four cell lines analysed was 112, suggesting that a significant number of Chr16 proteins must be searched in specific tissues/cell lines.

The detected transcripts were divided into different categories based on the quantities of the global FPKM distribution: very low, low, medium and high express genes. For Chr16 protein coding genes, 33 (3.72 %) were low or very low expressed in all samples and only 7 genes (0.79 %) were highly expressed. The expression index for the remaining genes was tissue dependent.

To characterize in detail the proteome of Chr16, four cell lines were selected: MCF7, CCD18, Jurkat and Ramos. Shotgun proteomics were conducted in parallel by 6 independent laboratories resulting in 24 2D-LC-MS/MS and gel-LC-MS/MS experiments. Raw MS and MS/MS data were translated to mascot general file (mgf) format and searched against the UniprotsSwissProt human database (release 2012_06, June 13) using an in-house Mascot Server. False Discovery Rates (FDR ≤1% at the protein level) for protein identification were manually calculated.

The detected transcripts were divided into different categories based on the quantities of the global FPKM distribution: very low, low, medium and high express genes. For Chr16 protein coding genes, 33 (3.72 %) were low or very low expressed in all samples and only 7 genes (0.79 %) were highly expressed. The expression index for the remaining genes was tissue dependent.