Genomics Data 3 (2015) 57–60

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Whole transcriptome RNA sequencing data from blood leukocytes derived from Parkinson's disease patients prior to and following deep brain stimulation treatment

Lilach Soreq [a], Nathan Salomonis [b], Alessandro Guffanti [c], Hagai Bergman [a,d], Zvi Israel [e], Hermona Soreq [f,d,*]

[a] Department of Medical Neurobiology, The Hebrew University — Hadassah Medical School, Jerusalem 91120, Israel
[b] Department of Pediatrics, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA
[c] Genomnia srl, Lainate, Milan, Italy
[d] The Edmond and Lily Safra Center for Brain Sciences (ELSC), Israel
[e] The Center for Functional and Restorative Neurosurgery, Department of Neurosurgery, Hadassah University Hospital, Jerusalem, Israel
[f] Department of Biological Chemistry, The Life Sciences Institute, The Hebrew University of Jerusalem, Jerusalem, Israel

ARTICLE INFO

Article history:
Received 7 November 2014
Received in revised form 17 November 2014
Accepted 17 November 2014
Available online 22 November 2014

Keywords:
Alternative Splicing
Deep Brain Stimulation
Leukocytes
Parkinson's Disease
RNA sequencing

ABSTRACT

Recent evidence demonstrates the power of RNA sequencing (RNA-Seq) for identifying valuable and urgently needed blood biomarkers and advancing both early and accurate detection of neurological diseases, and in particular Parkinson's disease (PD). RNA sequencing technology enables non-biased, high throughput, probe-independent inspection of expression data and high coverage and both quantification of global transcript levels as well as the detection of expressed exons and junctions given a sufficient sequencing depth (coverage). However, the analysis of sequencing data frequently presents a bottleneck. Tools for quantification of alternative splicing from sequenced libraries hardly exist at the present time, and methods that support multiple sequencing platforms are especially lacking. Here, we describe in details a whole RNA-Seq transcriptome dataset produced from PD patient's blood leukocytes. The samples were taken prior to, and following deep brain stimulation (DBS) treatment while being on stimulation and following 1 h of complete electrical stimulation cessation and from healthy control volunteers. We describe in detail the methodology applied for analyzing the RNA-Seq data including differential expression of long noncoding RNAs (lncRNAs). We also provide details of the corresponding analysis of in-depth splice isoform data from junction and exon reads, with the use of the software AltAnalyze. Both the RNA-Seq raw (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42608) and analyzed data (https://www.synapse.org/#!Synapse:syn2805267) may be found valuable towards detection of novel blood biomarkers for PD.

Specifications	
Organism/cell line/tissue	Human/blood leukocytes
Sex	Male
Sequencer or array type	SOLiD-3
Data format	Raw and analyzed (primary RNA-Seq analysis level).
Experimental factors	Parkinson's disease pre-deep brain stimulation (DBS) treatment, and post treatment on and off electrical stimulation as well as age and gender-matched control samples.
Experimental features	From each patient, 3 blood samples were taken: one day prior to DBS neurosurgery which included bi-lateral implantation of microelectrodes into the subthalamic nucleus (upon hospitalization), and several weeks post-DBS 1) while being on electrical stimulation and following 1 h of electrical stimulation cessation. The leukocytes were fractionated from each sample.
Consent	All the samples were taken under informed consent and under the Hadassah Ethics committee approval.
Sample source location	Jerusalem, Israel, Human.

Direct link to deposited data

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42608
https://www.synapse.org/#!Synapse:syn2805267

* Corresponding author.
E-mail address: hermona.soreq@mail.huji.ac.il (H. Soreq).

http://dx.doi.org/10.1016/j.gdata.2014.11.009

Experimental design, materials and methods

Experimental design

Blood leukocyte samples from three PD patients were collected one day prior to DBS treatment, and a few weeks later upon clinical symptoms stabilization. Sampling occurred under turned on electrical stimulation and following 1 h of electrical stimulation cessation (which re-induced the disease motor symptoms). Samples were also collected from age and gender matched apparently healthy control (HC) volunteers (Fig. 1).

Materials and methods

Subject recruitment

Three PD patients and three matched control volunteers were recruited to the study. All the participants signed informed consent forms prior to inclusion. The samples were collected from the PD patients' pre- and post-bilateral sub-thalamic (STN)-DBS neurosurgery while being on stimulation and following a short 1-hour of stimulation cessation. All of the study volunteers that passed the study exclusion criteria signed informed consent forms prior to inclusion in the study (clinical parameters of the recruited volunteers are given under [1]). To control for potential variability in the leukocyte expression profiles that might stem from other factors (such as infections, or other diseases), volunteers were assessed for their clinical overall background and state up to 1 year earlier (e.g past hospitalizations, high fever) and subsequently fulfilled detailed medical history questionnaires. Exclusion criteria for participant patients included depression and past and current DSM Axis I and II psychological disorders (SM), chronic inflammatory disease, coagulation irregularities, previous malignancies or cardiac events, or any surgical procedure up to one year pre-DBS. Potential volunteers that did not fulfill these inclusion criteria were excluded from the study. All patients went through bilateral STN-DBS electrode implantation (Medtronics, USA) and were under dopamine replacement therapy (DRT) both pre- and post-DBS (on significantly reduced dosage post-DBS, with *t*-test $p < 0.01$), the last medication administered at least five hour pre-sampling. The clinical severity of the disease was assessed by a neurologist using the Unified PD Rating Scale (UPDRS) [2]. Controls were recruited among Hadassah hospital staff and researchers at the Edmond J. Safra Campus (Jerusalem). All study volunteers underwent stringent filtering prior to inclusion in the study. The exclusion criteria for the healthy control volunteers included smoking, chronic inflammatory diseases, drug/alcohol usage, major depression, previous cardiac events, fever within up to 3 months prior to inclusion in the study and past year hospitalizations.

Blood sample collection

To reduce expression profile variability due to time differences in the sample collection, blood collection was conducted within a fixed range of hours (10 AM–14 PM). The samples were collected in 9 ml blood using 4.5 ml EDTA (anti-coagulant) tubes.

Leukocyte fractionation

The collected venous blood was filtered using the LeukoLock fractionation and stabilization kit (Ambion, Applied Biosystems, Inc., Foster City, CA) up to 15 post-extraction minutes to enhance inspection accuracy. To ensure high RNA quality, the leukocyte-enriched samples were immediately incubated in RNALater (Ambion) (http://www.affymetrix.com/support/technical/technotes/blood_technote.pdf). Stabilized filters and serum samples were stored at −80 °C.

RNA extraction

RNA extraction followed the manufacturer's (Life Technologies) alternative protocol instructions for using the LeukoLock filters. Briefly, cells were flushed (TRI-Reagent Ambion) into 1-bromo-3-chloropropane-containing 15 ml tubes and centrifuged. 0.5 and 1.25 volume water and ethanol were added to the aqueous phase. Samples were filtered through spin cartridges, stored in pre-heated 150 μl EDTA; RNA was quantified in Bioanalyzer 2100. Determination of RNA quality and quantity was conducted using the Eukaryote Total RNANano 6000 kit (Agilent). RNA was frozen and stored in −80 °C immediately after production.

cDNA libraries preparation and sequencing

RNA quality was assessed by running the samples on Agilent RNA 6000 Nano-gels (#5067-1511). For each library Ribosomal RNA of 5 μg total RNA was removed using Invitrogen's RiboMinus kit (#A10837-08) and then samples were concentrated using the RiboMinus Concentration Module (Invitrogen). Ribosomal RNA removal was verified by RNA 6000 Nano gel analysis. Library construction was conducted according to SOLiD Whole Transcriptome Analysis Kit (PN4425680) protocol, fragmentation (by RNase-III) was verified on Agilent RNA 6000 Pico Kit (#5067-1513) and 150 ng fragmented RNA was used for further protocols. cDNA samples were run on 4% Agarose gels, 150–250 base pairs (bp) sized fragments were cut and extracted using Qiagen Min-Elute Gel-Extraction Kit (#28604), and gels were dissolved by intensive vortex and not by heating. Libraries were amplified for 12 cycles using bar-coded primers supplied in SOLiD Transcriptome Multiplexing Kit (Ambion, #4427046). Libraries were quantified using the Kapa ABI SOLiD Library Quantification Kit (KK4833) and diluted for final analysis on Agilent High Sensitivity DNA Kit (#5067-4626). 500 μM libraries were used for emulsion PCR according to Applied Biosystems SOLiD-3 System Template Bead Preparation Guide (4407421) to prepare for sequencing on the SOLiD-3 platform.

The general workflow of RNA-Seq analysis is under Fig. 2.

Primary processing of SOLiD RNA-Seq reads

RNA-Seq reads (.csfasta files) and quality scores (.qual files) were obtained using the SOLiD instrument software: SOLiD-3. System

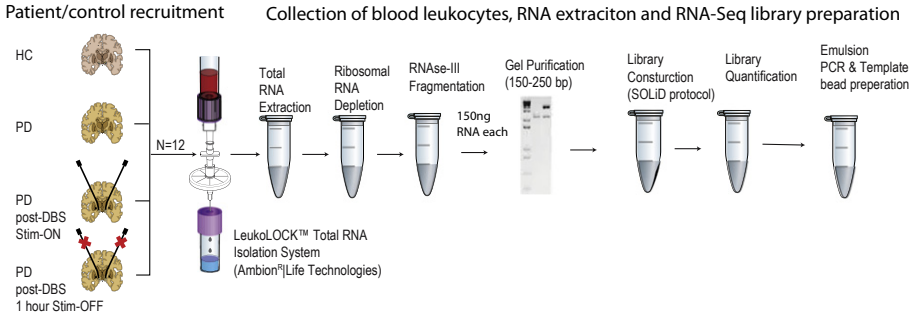

Fig. 1. General experimental flow: blood samples from 3 Parkinson's disease (PD) patients pre-Deep Brain Stimulation (DBS) and post-DBS on electrical stimulation and following 1 h of complete electrical stimulation cessation were collected as well as from 3 age- and gender- matched healthy control (HC) volunteers. From each sample, The leukocytes were filtered, RNA was produced and cDNA library prepared for sequencing by SOLiD 3 system.

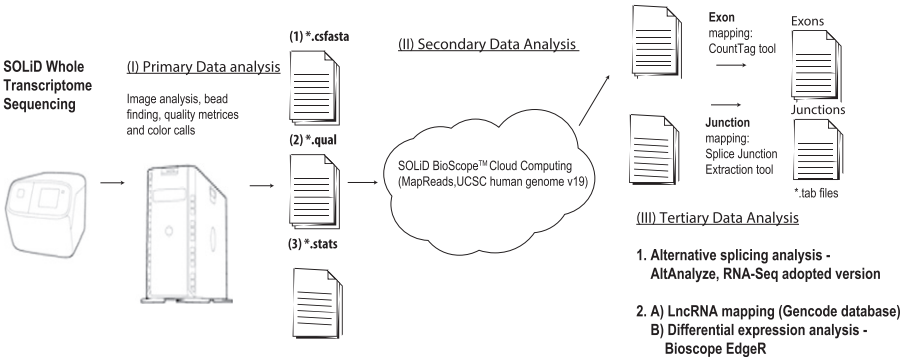

Fig. 2. Whole transcriptome sequencing on the produced libraries yielded .csfasta, .qual and .stat files for each sequenced sample. Mapping to the human genome (UCSC database) was conducted through cloud computing to produce both exon and junction level quantification files. AltAnalyze software was used to analyze these files separately and combined and detect high confidence splice isoform alteration between the different clinical states. Long noncoding RNAs (lncRNAs) were detected in the sequenced libraries through mapping to the GENCODE catalog (version 7), and differential expression analysis conducted for the lncRNAs using the Bioconductor project EdgeR program.

software analysis was used for all the primary data analyses including image analysis, bead finding, quality metrics and color calls. The software applications used to set and control data analysis included SOLiD software suite under license agreement. The suit included: Instrument Control Software (ICS), SOLiD Experimental Tracking System (SETS), and SOLiD Analysis Tools (SAT) V3.0. Job management by the Job Manager used the Corona-Lite v4.0 platform. Sequencing was run on the Applied Biosystems SOLiD 3 System. Images of each cycle were analyzed, data clustered and normalized. For each tag, a sequential (sequence-ordered) set of color space calls was produced. Quality metrics were produced through normalization. Two probe sets were used to maximize the fraction of "mappable" amplified beads, read length and sequencing throughput for sequencing of the 50-bp reads. Five rounds of primers (A, B, C, D and E) were used to sequence template by ligation of di-base labeled probes. As the libraries were size fragmented, the set of primers used was specific to the P1 Adaptor. For each library three types of raw data files were created: .csfasta (the sequenced reads in color space), .qual and .stats. The quality values given in the .qual files (estimate of confidence given for each color call), q for a particular call, is mathematically related to its probability of error (p), and its SOLiD q values are similar for those generated by Phred and the KB basecaller for capillary electrophoresis (described in detail under [3]). The algorithm relies on training (calibration) to a large set of control data and color calls for which the correct call is known. In the SOLiD-3 system, the correct call is determined by mapping the read to a known reference sequence. Further details regarding additional RNASeq analysis pipelines are detailed under [1].

Secondary RNA-Seq analysis: mapping of RNA-Seq reads to exons and junctions

The analysis stage included aligning the reads to the human genome — UCSC human genome version 19 February 2009 GRCh37 (the reference genome in this case, using the database file homo_sapiens.GRCh37.56.dna.toplevel.fa) and generation of base space sequences. Each library was mapped using the Mapreads function of SOLiD™ BioScope™ (v1.3) cloud computing software (life technologies, applied biosystems, Carlsbad, California). All the alignment and mapping were done on ordered genomes.

Tertiary RNA-Seq analysis: exon and junction expression quantification

Junction mapping was conducted using the Bioscope splice junction extractor tool through cloud computing and exon level mapping conducted using the count tag tool. 2 mappings were run for each sample: 1) exon level — to receive exon quantification and 2) junction level quantification. The output of this step included .gff and .sam files. For both cases, merging of count reads that employed discontinuous word pattern search algorithms was performed. The scripts used for mapping are also included in the public Synapse depository.

De-novo splicing prediction, poly-A predictions and UCSC exon annotation

Following the identification of all the transcript structural event types by parsing of both the Ensembl and UCSC databases, the cases of intron retention were identified first by searching for regions that span two adjacent exons in at least one additional transcript for each gene. The remaining exons were clustered based on overlapping genomic coordinates (e.g. alternate 5′ ends or 5′ start-sites). Each exon was annotated as corresponding to exon block and region number within the corresponding transcript. All possible pair-wise transcript comparisons for each gene were then performed to identify exon pairs that show evidence of alternative exon-cassettes, alternative 3′ or 5′ splice-sites or alternative-N or -C terminal exons. All transcript exon pairs were considered (except for those adjacent to a retained intron) by comparing the exon block-ID and region-IDs of an exon and its neighboring exons to the exon blocks and regions in the compared transcripts, and a custom heuristic assigned the appropriate annotation based on these transcript comparisons. This process is repeated when importing RNA-Seq reads to identify novel splicing events. In addition to all de-novo splicing annotations, additional splicing annotations (e.g., alternative cassette-exon, alternative promoter) are imported from the UCSC genome database and linked to existing exon blocks and regions based on genomic coordinate overlap. We have further incorporated alternative Poly-A predictions using the polyA database [4]. Further details are given under AltAnalyze online documentation (www.altanalyze.org) [5].

Finding alternatively spliced exons

Through AltAnalyze, we analyzed exons and junctions separately as well as combined these results to find high confidence alternatively spliced exons and differentially expressed genes. We used the splicing index method (originally developed for exon arrays), ASPIRE (originally developed for splice junction arrays [6]) as well as linear regression calculations (originally developed for splice junction arrays) on pairs of reciprocal junction pairs (that can either include or exclude the exon found in between them) [7]. The value calculated by splicing index is the change in exon-inclusion (delta I, dI, or splicing index fold change).

$\mathrm{NI}(probeset\_i) = \left(\frac{\mathrm{PI}}{\mathrm{GE}}\right)$ where PI is the intensity of the exon and GE is the gene level expression value in that sample group to obtain a normalized intensity (NI) for each exon. Comparison of the $\log_2$ values of the NI between two sample groups arrives at a splicing-index (SI) value [8,9].

$$\mathrm{SI}(probeset\_i) = \log_2\left(\frac{\mathrm{NI(PI)sample1}}{\mathrm{NI(PI)sample2}}\right).$$

In ASPIRE, similarly to the splicing-index method, for each reciprocal junction, an estimate of overall junction ratio differences is calculated by comparing the expression of the two reciprocal junctions being measured (e.g., E1–E2 versus E1–E3) (non-log), between the baseline and experimental groups. These ASPIRE dI can be calculated from the comparison of inclusion and exclusion junctions or an inclusion exon and an exclusion junction. The linear regression algorithm (based on previously described approach [7]) also uses the same reciprocal exon/junction features as ASPIRE. To derive the slope for each of the two biological conditions (control and experimental), the constitutive corrected expression of all samples for both reciprocal junctions is plotted against each other to calculate a slope for each of the two biological groups using the least squares method. In each case, the slope is forced through the origin of the graph (model = y ~ x − 1 as opposed to y ~ x). The final linear regression score is the $\log_2$ ratio of the slope of the baseline group divided by the experimental group. This ratio is analogous to a $\log_2$ fold change, where 1 is equivalent to a 2-fold change.

Analysis of protein binding domain composition

Identification of protein domains that were predictably disrupted by alternative splicing changes in each clinical state was conducted through AltAnalyze. To identify alternative protein domains, RNA-Seq and microarrays probe-sets sequences were used to identify which proteins align to, or are missing from, transcripts for each disease, treatment or stimulation cessation spliced gene transcript, and specifically for each spliced isoform.

Quantification and differential expression analysis of long non-coding RNAs (lncRNAs)

As part of the ENCODE project, the GENCODE consortium manually curated 9277 human lncRNAs. The human long non-coding RNA database bed coordinates of Gencode (version 7) [10] were downloaded from the GENCODE lncRNA data page of the CRG Bioinformatics and Genomics Group [http://big.crg.cat/bioinformatics_and_genomics/lncrna_data] and complemented with other non-coding transcript information available from the EnsEMBL BioMart version 0.7 query interface to the EnsEMBL Genes 72 — GRCh37.p11 database (www.ensembl.org). Genome coordinates in .BED format (corresponding to the mapped reads) used the Lifescope Lifetech 2.5.1 software and UCSC hg19 masked reference database as obtained by the original .sam files with SAMtools, SAMtools view and bedtools bamToBed. These read bed files were intersected with the genome coordinates of the above-mentioned lncRNAs using the bedtools intersectBed program, requiring a 90% overlap of each sequence read with a target lncRNA. Lists of sequence tags corresponding to lncRNAs were obtained by intersection of the bed tools. For differential expression analysis, the read count information of all of the detected lncRNAs was first filtered. LncRNAs that did not present read count in >3 libraries, or did not exist in EnsEMBL were filtered out. The remaining lncRNAs (overall, 6430 leukocyte expressed lncRNAs) were analyzed using the Bioconductor edge-R [11] software (version 3.0.1) to detect differential expression between the patients in different clinical stages and the control volunteers. This analysis module is particularly suitable to use on small numbers of replicate samples. The results were annotated using the BioMart integrated annotation database query interface [12] using the human genome reference consortium assembly build version 37 (GRch37, hg19) and GENCODE version 7.

Biological pathways analysis

Gene Ontology (GO)-Elite pathway analysis [13] was conducted through AltAnalyze program, on the analyzed output result files of the detected splice events at both the junction and exon levels as well as the David tool EASE [14].

Conflict of interest

The authors declare no conflict of interests.

Acknowledgments

The authors are grateful to the patient and healthy control volunteers who made this study possible by donating their blood cells and time.

Funding

The contribution of the Edmond and Lily Safra Center of Brain Science and Ms Ellaine B Miller and family towards this study is acknowledged. L.S. thanks HUJI sources for PhD fellowship support.

References

[1] L. Soreq, et al., Long non-coding RNA and alternative splicing modulations in Parkinson's leukocytes identified by RNA sequencing. PLoS Comput. Biol. 10 (3) (2014) e1003517.
[2] S. Fahn, R. Elton, Members of the UPDRS development committee, Unified Parkinson's Disease Rating Scale. 1987.
[3] B. Ewing, P. Green, Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 8 (3) (1998) 186–194.
[4] H. Zhang, et al., PolyA_DB: a database for mammalian mRNA polyadenylation. Nucleic Acids Res. 33 (Database issue) (2005) D116–D120.
[5] N. Salomonis, et al., Alternative splicing regulates mouse embryonic stem cell pluripotency and differentiation. Proc. Natl. Acad. Sci. 107 (23) (2010) 10514.
[6] J. Ule, et al., Nova regulates brain-specific splicing to shape the synapse. Nat. Genet. 37 (8) (2005) 844–852.
[7] C.W. Sugnet, et al., Unusual intron conservation near tissue-regulated exons found by splicing microarrays. PLoS Comput. Biol. 2 (1) (2006) e4.
[8] K. Srinivasan, et al., Detection and measurement of alternative splicing using splicing-sensitive microarrays. Methods 37 (4) (2005) 345–359.
[9] P.J. Gardina, et al., Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. BMC Genomics 7 (2006) 325.
[10] T. Derrien, et al., The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. 22 (9) (2012) 1775–1789.
[11] M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 (1) (2010) 139–140.
[12] S. Durinck, et al., BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21 (16) (2005) 3439–3440.
[13] A.C. Zambon, et al., GO-Elite: a flexible solution for pathway and ontology over-representation. Bioinformatics 28 (16) (2012) 2209–2210.
[14] D.A. Hosack, et al., Identifying biological themes within lists of genes with EASE. Genome Biol. 4 (10) (2003) R70.