PET-Tool: a software suite for comprehensive processing and managing of Paired-End diTag (PET) sequence data

Kuo Ping Chiu¹, Chee-Hong Wong², Qiongyu Chen³, Pramila Ariyaratne¹, Hong Sain Ooi¹, Chia-Lin Wei⁴, Wing-Kin Ken Sung¹,³ and Yijun Ruan*⁴

Address: ¹Information and Mathematical Sciences Group, Genome Institute of Singapore, 60 Biopolis Street, Genome #02-01, 138672, Singapore, ²Bioinformatics Institute, 30 Biopolis Street, Matrix #08-01, 138671, Singapore, ³Department of Computer Science, National University of Singapore, 3 Science Drive 2, 117543, Singapore and ⁴Genome Technology and Biology Group, Genome Institute of Singapore, 60 Biopolis Street, Genome #02-01, 138672, Singapore

Email: Kuo Ping Chiu - chiukp@gis.a-star.edu.sg; Chee-Hong Wong - wongch@bii.a-star.edu.sg; Qiongyu Chen - chenqy@comp.nus.edu.sg; Pramila Ariyaratne - ariyaratnep@gis.a-star.edu.sg; Hong Sain Ooi - ooih@gis.a-star.edu.sg; Chia-Lin Wei - weicl@gis.a-star.edu.sg; Wing-Kin Ken Sung - sungk@gis.a-star.edu.sg; Yijun Ruan* - ruanyj@gis.a-star.edu.sg

* Corresponding author

Published: 25 August 2006
Received: 27 June 2006
Accepted: 25 August 2006

Abstract

Background: We recently developed the Paired End diTag (PET) strategy for efficient characterization of mammalian transcriptomes and genomes. The paired end nature of short PET sequences derived from long DNA fragments raised a new set of bioinformatics challenges, including how to extract PETs from raw sequence reads, and correctly yet efficiently map PETs to reference genome sequences. To accommodate and streamline data analysis of the large volume PET sequences generated from each PET experiment, an automated PET data process pipeline is desirable.

Results: We designed an integrated computation program package, PET-Tool, to automatically process PET sequences and map them to the genome sequences. The Tool was implemented as a web-based application composed of four modules: the Extractor module for PET extraction; the Examiner module for analytic evaluation of PET sequence quality; the Mapper module for locating PET sequences in the genome sequences; and the ProjectManager module for data organization. The performance of PET-Tool was evaluated through the analyses of 2.7 million PET sequences. It was demonstrated that PET-Tool is accurate and efficient in extracting PET sequences and removing artifacts from large volume dataset. Using optimized mapping criteria, over 70% of quality PET sequences were mapped specifically to the genome sequences. With a 2.4 GHz LINUX machine, it takes approximately six hours to process one million PETs from extraction to mapping.

Conclusion: The speed, accuracy, and comprehensiveness have proved that PET-Tool is an important and useful component in PET experiments, and can be extended to accommodate other related analyses of paired-end sequences. The Tool also provides user-friendly functions for data quality check and system for multi-layer data management.
Background
Tag-based sequencing strategies such as Serial Analysis of Gene Expression (SAGE) are efficient for analyzing DNA fragments in transcriptome characterization and genome annotation studies [1-3]. However, the information content in each SAGE tag based on an anchored restriction enzyme recognition site within the DNA segment is limited, and the mapping of SAGE tags to genome sequences for transcript identification can be ambiguous. Despite the recent improvements in tagging 5’ terminal signatures of cDNA [4,5] to determine transcription start sites (TSS), the most significant advance in this field is the simultaneous tagging of 5’ and 3’ terminal signatures of DNA fragments subjected to study. In this effort, we first developed an intermediate approach that precisely extracts separate 5’ and 3’ terminal tags from cDNA fragments for sequencing [6]. With this new capability, we proceeded to design and develop a cloning strategy, called Gene Identification Signature (GIS) analysis, which covalently links the 5’ and 3’ signatures of each full-length transcript into a Paired-End diTag (PET) structure [7]. In a GIS-PET experiment, most of the PETs are 36bp in length (18bp for the 5’ signature tag and 18bp for the 3’ signature tag); and multiple PETs can be concatenated together to form longer stretches of DNA fragments for efficient high-throughput sequencing. An average sequencing read (700–800bp) of a GIS-PET library clone can reveal 10–15 PET units, which is equivalent to 30 conventional cDNA sequencing reads for 15 cDNA clones analyzed from both ends. The PET sequences can then be accurately mapped to the reference genome sequences and precisely demarcate the boundaries of transcription units in the genome landscape. With this combined efficiency and accuracy of GIS-PET, a mammalian transcriptome can be thoroughly analyzed using hundreds of thousands high quality transcript sequences by a modest sequencing effort as further demonstrated in the comprehensive characterization of mouse transcriptomes [8]. The PET-based DNA analysis strategy has also been applied to characterize genomic DNA fragments generated by chromatin immunoprecipitation (ChIP) enriched for specific binding targets by given DNA-binding proteins, and whole genome ChIP-PET data has provided global maps of transcription factor binding sites for p53 in the human genome [9] and Oct4 and Nanog in the mouse genome [10]. PET-based DNA analyses (GIS-PET and ChIP-PET) promise to play a significant role in the post-genome efforts to identify all functional elements in the human genome [11], and there is no inherent limit for the PET-based approach to be applied to other DNA analyses, such as analyses of epigenetic elements.

To fully appreciate the potential of PET-based sequencing analyses, we have to develop sophisticated informatics capabilities to manage the large volume of specific PET sequences generated from each of the GIS-PET and ChIP-PET experiments. There is a battery of new bioinformatics challenges around how to accurately identify and extract PET sequences embedded in raw sequence reads, how to specifically and efficiently map the paired 5’ and 3’ signatures of PET sequences in complex genomes such as the human and mouse genome sequences; and how to be user-friendly in managing the immense amount of data generated from GIS-PET and ChIP-PET experiments for effective data mining and analysis. Based on the paired end nature of PET sequences generated from GIS-PET and ChIP-PET experiments, the issues are far more complicated than those related to SAGE-like mono-tags and therefore cannot be handled by available software packages previously developed for SAGE analysis [12-15].

To accommodate and process PET sequence data, we developed a complete software suite called PET-Tool that is designed to provide complete solutions starting from extracting PET sequences from raw sequencing reads, to mapping the PET sequences to the reference genomes. Here in this study, we describe the architecture design, technical details of implementation, utility, and robustness of PET-Tool by analyzing four datasets generated from two GIS-PET libraries and two ChIP-PET libraries.

The architecture of PET-Tool
PET-Tool is designed to provide complete solutions for processing and managing the PET data generated from GIS-PET and ChIP-PET experiments. In these experiments, either full-length cDNA or genomic DNA enriched by ChIP are converted into PET structures that are further concatenated and cloned into plasmid vector for sequencing analysis [7,9]. The core functions of the Tool are to extract PET sequences from raw DNA sequence reads and map the PET sequences to the genome sequences. In addition, we want the Tool to be able to manage large volume of PET data generated from each PET experiment and provide user-friendly analytic functions to evaluate the quality of each PET dataset. The design of PET Tool is comprised of four modules: Extractor, Examiner, Mapper, and ProjectManager (Figure 1). In the PET-Tool system, Extractor uploads raw sequence files and de-convolutes the PET sequence units embedded in each raw sequence read to generate PET sequences, which are stored in a relational database. Examiner provides an analytical capability for users to examine and validate the PET extraction results. It provides the basic statistics of PETs in each project, library, and plate of sequences. It also presents graphic dissection for each of the input sequence reads and highlights the sequence sections with various color codes to help users to distinguish vector flanking regions, spacer sequences between PET units, and the PET sequences themselves. This ability allows users to identify any potential irregularities in the sequence, and adjust extraction parameters. The Mapper module is to map the
quality PET sequences to the corresponding genome sequences. For efficient mapping of large volumes of PET sequences, we used a newly developed alignment approach that was based on compressed suffix array (CSA), in which the entire genome sequence assembly was indexed as a reference database, and the 5' and 3' signatures of a PET sequence were matched to the genomic index [unpublished results]. The ProjectManager module organizes the data and analysis results in a hierarchical order, in which multiple projects can be managed at various levels of organisms, libraries, raw DNA sequences (in plate and single well format), PETs, and the attributes of each PET.

Implementation
Pet-Tool is implemented for both UNIX and LINUX. The web-based user interface is implemented in Perl/CGI and hosted by Apache web server. The interface of the Tool can be accessed by any web-browser that supports the current web standards.

Data storage is facilitated by a combination of flat file system and mySQL based Relational Database Management System (RDBMS). The mySQL database was used for efficient and fast PET data storage, tracking, retrieving, and interfacing with back-end programs through Perl:DBI module. We also applied mySQL to host various statistical data and mapping results. Flat files were used for storage of uploaded sequence data, with the positional indices of all sequences stored in mySQL database for quick sequence retrieval. Back-end programs were implemented in Perl and C languages. Compressed Suffix Array (CSA) programs were implemented in C language for high efficiency and robust performance of advanced data structures. Programs for PET sequence extraction, statistic computation, data retrieval/storage, web-interaction and other non-intensive tasks were implemented in Perl. Minimum hardware requirements include Pentium III processor, CPU of 500 MHz, 256 Mega byte RAM, and 20 Giga-byte hard disk drive. A regular 500 MHz machine would take about two days to process a library of one million PETs. If a computer was equipped with 2.4 GHz processor, the same job could be done in a few hours.

Results and discussion
The current settings of PET-Tool can handle GIS-PET for transcriptome analysis and ChIP-PET data for whole genome localization of transcription factor binding sites. We have successfully applied PET-Tool to more than 45 GIS-PET and ChIP-PET libraries. To demonstrate the data
processing workflow, and the functionalities and performance of PET-Tool, we analyzed two GIS-PET libraries and two ChIP-PET libraries in this study.

Datasets used for analysis in this study
The two GIS-PET libraries (SHC12 and SHC13) used in this study were derived from human cancer cells MCF7 and HCT116, respectively; and the two ChIP-PET libraries (SHC16 and SHC19) were from chromatin immunoprecipitated DNA enriched for STAT1 binding sites in human HeLa cells with and without interferon-γ treatment, respectively. For each of the PET libraries, more than 50,000 plasmid clones were sequenced from one direction. In total, over 300,000 high quality sequence reads were generated for these 4 libraries (Table 1).

PET-Tool procedure to process the PET library sequences
The use of PET-Tool starts from the ProjectManager module, from which new project and library IDs were assigned for each new sequence dataset. In the PET-Tool system, a project can host multiple libraries, and a library can contain from a few to hundreds of thousands of sequences. The two GIS-PET libraries presented in this study were assigned library IDs SHC012 and SHC013 under the project name "Human Transcriptome", while the two ChIP-PET libraries were assigned as SCH016 and SCH019 under the project name "Human TFBS" (transcription factor binding sites). Once the entry for a library is created, the library sequence file is uploaded and the PET sequences are extracted from raw DNA sequences via the Extractor module. The quality of PET sequences can be examined using the Examiner module, and high quality PET sequences can be mapped to the corresponding genome sequences using the Mapper module (Figure 2).

### Table 1: Statistics of PET characteristics

| Library Type | Library ID | GIS-PET | ChIP-PET |
|--------------|------------|---------|----------|
|              | SHC012     | SHC013  | (combined) | SHC016 | SCH019 | (combined) |
| Raw sequence reads | 74,537 | 53,758 | 128,295 | 89,359 | 82,941 | 172,300 |
| Spacer-defined PETs | 741,799 | 363,963 | 1,105,762 | 777,038 | 845,045 | 1,622,083 |
| PET/sequence read | 10 | 6.8 | 8.6 | 8.7 | 10.2 | 9.4 |
| Rejected poor PETs | 157,175 | 83,623 | 240,798 | 51,161 | 81,510 | 132,671 |
| Rejection rate % * | 21.1 | 22.3 | 22 | 6.6 | 9.7 | 2.2 |
| Total high-quality PETs | 584,624 | 280,340 | 864,964 | 725,877 | 763,535 | 1,489,412 |
| Total unique PETs | 135,757 | 145,138 | 280,895 | 640,844 | 582,253 | 1,223,097 |
| Redundancy % ** | 76.8 | 48.2 | 62.5 | 11.7 | 23.7 | 17.7 |
| 5' AT content (%) | 31.71 | 32.89 | 32.30 | 58.06 | 57.85 | 57.96 |
| 3' AT content (%) | 61.49 | 62.18 | 61.84 | 57.98 | 57.57 | 57.78 |

| Breakdown of rejected PETs | SHC012 | SHC013 | SCH016 | SCH019 |
|----------------------------|--------|--------|--------|--------|
| Length < 34                | 43,022 | 7,942  | 26,343 | 16,581 |
| -27.40%                    | -9.50% | -51.10%| -20.40%|
| Length > 40                | 17,377 | 7,967  | 24,794 | 16,581 |
| -11.10%                    | -9.50% | -48.90%| -20.40%|
| Contain N                  | 43,673 | 26,901 | N. A.  | N. A.  |
| -27.80%                    | -32.20%|
| No AA-tail at 3' end       | 41,983 | 20,264 | N. A.  | N. A.  |
| -26.70%                    | -24.20%|
| PolyA(9) in 3' tag         | 433    | 99     | N. A.  | N. A.  |
| -25%                       | N. A.  | N. A.  |
| PolyT(9) in 5' tag         | 116    | 118    | N. A.  | N. A.  |
| PolyT(9) in 3' tag         | 639    | 323    | N. A.  | N. A.  |
defined, each of the library sequence files in FASTA format was uploaded (Figure 3) and stored in the flat file system. In the Extractor module, DNA sequences were read through, all sequence sections resided in between two spacer sequences were distinguished, and potential PET sequences were extracted. From the 128,295 raw sequence reads of the two GIS-PET libraries, 1,105,762 PETs were generated; while from the 172,300 raw sequences of the two ChIP-PET libraries, 1,622,083 PETs were generated. The overall efficiency of PET production is 8.6 PETs per sequence read for the two GIS-PET libraries and 9.4 PETs per sequence read for the two ChIP-PET libraries.

The spacer-defined raw PETs were then subjected to serial steps of filtering to exclude incorrect PETs due to imperfect molecular reactions during the molecular cloning process. It is known that the TypeII restriction enzymatic cleavage, DNA end polishing, and ligation reactions have a certain level of slippage, and the combination of these reactions would contribute to deviation of actual PET lengths from the predicted PET lengths by one to several nucleotides [7]. Hence, we have set an empirical range (34–40bp) around the expected size (36bp) for true ditags. Other ditags that were either shorter than 34 bp or longer than 40 bp were considered experimental artifacts.

Figure 2
Process flow of PET data Analysis in PET-Tool. Experimental information for each PET library was entered into the system through the ProjectManager functions. High quality DNA sequences in the FASTA files of PET libraries were uploaded through Extractor. Extracted PETs and all related information were stored in a MySQL database. Virtually, PETs were organized in a hierarchical order from project to library, plate, well, and individual PET. Each unique PET was assigned a unique ID and accorded with copy number (count). The quality of PET extraction was evaluated using Examiner. Errors that occurred in any steps in PET extraction or PET library construction and sequencing could be identified for correction. After PET extraction was validated, PET sequences were mapped to the genome sequences, and the mapping coordinates for each of the PETs were reported in table format.
PET sequences with low complexity (homopolymer stretches of more than 8 consecutive same nucleotides such as As or Ts, etc) were also removed because these PETs lack sufficient specificity for mapping to reference genome sequences. As an indication of PET orientation, we kept an "AA" residue of the cDNA polyA tail in the PET sequences at 3' end in GIS-PET libraries. Therefore, if any GIS-PET ditags did not contain the AA tail at the 3' end, these questionable PETs were also removed. After these layers of filtering, 864,964 high quality PETs were collected for the two GIS-PET libraries and 1,489,412 high quality PETs for the ChIP-PET libraries. Redundant PETs were collapsed into unique PETs. The copy numbers for each of the unique PETs reflect the abundance level of the PET in a given library. In total, 135,757 unique PETs were collected for SCH012, 145,138 for SCH013, 640,844 for SCH016, and 582,253 for SCH019 (Table 1).

**Evaluation of PET quality using examiner**

When the automated process of PET extraction was complete, the PET sequences with related information were organized in mySQL tables. The PET extraction results can be viewed using Examiner. To add analytical functions for evaluating the library sequencing quality, any particular sequencing plate and individual sequencing wells can be viewed for PET extraction results. Furthermore, in the Examiner module, the pattern of PETs and spacers embedded in the original sequence reads can be displayed with color codes at the nucleotide level (examples of screen-
shots are shown in Figure 4). This level of detail in validating the quality of PET library sequences and the accuracy of PET extraction results provide a great convenience for users.

**Comparison of PET sequences derived from GIS-PET and ChIP-PET libraries**

Although the methods used to generate GIS-PET and ChIP-PET were similar, the starting DNA materials were rather different. GIS-PETs were derived from cDNA, while ChIP-PETs were derived from ChIP enriched genomic DNA fragments. It appears that the quality of GIS-PETs is lower than that of ChIP-PETs. About 22% of GIS-PET sequences as opposed to 8.2% of ChIP-PET sequences were rejected after quality filtering. There are several reasons contributing to higher error rates for GIS-PETs. One of the major differences between GIS-PET and the ChIP-PET was the inclusion of AA-tail as a 3’ directional indicator at the end of 3’ signature for each GIS-PET sequence. We observed that 30% of the rejected GIS-PETs lacked the appropriate AA-tail. We also observed that the AT content in GIS-PETs was significantly polarized, at 31% for the 5’ tag region and 61% for the 3’ tag region. This observation is in consistent with our knowledge that in transcripts or
cDNAs, 5' UTR (untranslated region) is GC rich and 3' UTR is AT rich [16]. In contrast, the 3'-prone polarization of AT content was not observed in ChIP-PET sequences because the ChIP DNA fragments were generated by randomly shearing of genomic DNA.

**Mapping of PET sequences to reference genome sequences**

Once the PET extraction results were confirmed, the PET sequences were ready for mapping to the human genome sequences. Using the Mapper module, each of the PET sequences was split into a 5' tag and a 3' tag, and the tags were independently searched for matches in of the human genome assembly (hg17). When mapped correctly to the genome sequences, a PET sequence from nucleotide position 1–18 should be aligned to the 5' boundary and 19- last to the 3' boundary of the corresponding DNA segment on a chromosomal locus. However, considering that both reverse transcriptase-derived non-templated nucleotide incorporation and TyIII restriction enzyme slip-pase could lead to ambiguities at the two ends of individual tags in a PET structure [6], we mandated a minimum 16-nucleotide contiguous match for the 5' (from nucleotide position 1 to 19) and 3' (from 18 to the last) tags of PET to accommodate most possible variations. The mapped tags were then mated based on the criteria that the paired 5' and 3' signatures of a PET sequence must be on the same chromosome, in the correct order and orientation (5'→3'), and within appropriate genomic distance. Human genes encoded in chromosomes contain intron/exon structures, and many of them could span large distance of the genome for (hundreds of thousands of base pairs). To accommodate most genes but set a limit to reduce non-specific matches, we decided on an arbitrary cutoff of one million base pairs between 5' and 3' tags of GIS-PETs. For the ChIP-PET sequences, since the ChIP enriched DNA fragments were ranged from less than 100 base pairs to the up-limit of 4000 base pairs as estimated by DNA electrophoresis on agarose gel, the cutoff for ChIP-PET mapping span was set at 6000 bp. This criterion for pairing within a defined size limit greatly increased the mapping specificity of PET to genome sequences, and non-specific mapping of individual tags were most likely not fit in such mating requirement (Figure 5). The final mapping and pairing results were reported in a mySQL table format that included the PET nucleotide sequences, the count (copy number) for each of the PET sequences, nucleotide positions of 5' and 3' matches, and genome coordinates to which the PET sequences mapped (Figure 6). These data provide detailed information, not only where PET sequences mapped, but also the frequency for each of PETs occurred in each library.

As presented in the four PET libraries in this study, more than 70% of PETs were mappable. For example, of the total unique PETs generated for Library SCH012, 102,660
PETs (75.6%) were mapped accurately to the human genome assembly hg17 (Table 2). The majority of the mapped PETs (83,089/102,660; 80%) mapped to single location in the genome. The rest of the PETs mapped to two or more locations in the genome, which may reflect the mapping of PETs to duplicated gene segments and repetitive elements in the genome. It is noteworthy that more than 92% of the mapped ChIP-PETs found a unique position in the genome. As we reported in previous studies [7,9], most PET sequences were accurate in demarcating the corresponding DNA elements in the genome. For example, over 98% of the GIS-PETs mapped accurately to 5' and 3' boundaries of known gene transcripts.

**Conclusion**

We have developed a comprehensive computation program package, PET-Tool, to accommodate demands for automated processing of large volume of PET sequences generated by PET-based experiments. We demonstrated the utility of PET-Tool by analyzing four PET libraries and more than 2.7 millions PET sequences, and proved that PET-Tool can accurately and efficiently dissect PET concatemer sequences, extract, organize PET sequences in a relational database for convenient evaluation of sequence quality and overall experimental integrity, and specifically map the PET sequences to the corresponding reference genome sequences.

**Availability and requirements**

Project name: PET-Tool; Project home page: http://www.gis.a-star.edu.sg/PET_Tool

Operating system(s): UNIX and LINUX; Programming language: Perl and C languages.

PET-Tool is free for non-commercial use. The complete package of PET-Tool is available in DVD format to be sent upon request, and downloadable from the PET-Tool home page. For users who would like to understand more
of the PET methodology, a detailed experimental protocol and a user manual are also available at the PET-Tool website.

Abbreviations
SAGE: Serial Analysis of Gene Expression
ChIP: chromatin immunoprecipitation
PET: Paired-End diTag
GIS: Gene Identification Signature
CSA: Compressed Suffix Array
TSS: Transcription Start Site
PAS: Poly-Adenylation Site

Authors’ contributions
CLW and YR conceptualized the framework of PET-Tool. KPC was overall responsible for detailed design and implementation of the programs, and involved in coding some of the Perl programs, as well as data curation. CHW was responsible for mySQL database design and coding of related Perl programs. WKS and QC developed the CSA programs used for PET mapping. PA and HSO helped to modify the programs. CLW provided user inputs and evaluated the prototype. WKS provided computational advice to improve the overall performance. KPC and YR wrote the paper.

Acknowledgements
The authors want to thank Mr. Charlie Lee for participation in webpage design, and Drs. Patrick Ng and Guillaume Bourque for invaluable suggestions. This work is supported by the Agency for Science, Technology and Research (A*STAR) of Singapore and the NIH/NHGRI (1R01HG003521-01).

References
1. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: Serial analysis of gene expression. Science 1995, 270:484-487.
2. Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW: Using the transcriptome to annotate the genome. Nature Biotechnol 2002, 20:508-512.
3. Wang TL, Maierhofer C, Speicher MR, Lengauer C, Vogelstein B, Kinzler KW, Velculescu VE: Digital karyotyping. PNAS USA 2002, 99:16156-16161.
4. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, Kodzius R, Watahiki A, Nakamura M, Arakawa T, Fukuda S, Sasaki D, Podhajska A, Harbers M, Kawaji J, Carninci P, Hayashizaki Y: Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. PNAS USA 2003, 100:15776-15781.
5. Hashimoto SI, Suzuki Y, Kasai Y, Morohoshi K, Yamada T, Sese J, Morishita S, Sugano S, Matsushima K: 5' end SAGE for the analysis of transcriptional start sites. Nature biotechnology 2004, 22:1146-1149.

Table 2: PET mapping to the genome

| Library Type | Library ID | GIS-PET | CHIP-PET |
|--------------|-----------|---------|----------|
|              | SHC012    | SHC013  | SHC016   | SHC019   |
| Unique PETs  |           |         |          |          |
| Un-mapped PET|           |         |          |          |
| (un-mapping rate %) |           |         |          |          |
| Mapped PET   |           |         |          |          |
| (mapping rate %) |           |         |          |          |
| PETs mapped to one location (PET1) | | | |
| (PET1 rate % to unique PETs) | | | |
| (PET1 rate % to mapped PETs) | | | |
| (PET1 tag total counts) | | | |
| (PET1 redundancy %)* | | | |
| PET1 mapped to known genes | | | |
| (PET1 rate % to known genes) | | | |

PETs mapped to multiple locations (% to all mapped PETs)

| PET2 (%) | 7.67 |
| PET3 (%) | 2.33 |
| PET4 (%) | 1.2 |
| PET5 (%) | 0.64 |
| PET6 (%) | 0.43 |
| PET7 (%) | 0.31 |
| PET8 (%) | 0.18 |
| PET9 (%) | 0.16 |
| PET10 (%) | 0.11 |
| PET1+ (%) | 1.38 |

* PET1 Redundancy % = (1 - Total unique PETs/Total high quality PETs) × 100. The number states the percentage of PET tags that are redundant in the category.
6. Wei CL, Ng P, Chiu KP, Wong CH, Ang CC, Lipovich L, Liu ET, Ruan Y: 5' long serial analysis of gene expression (LongSAGE) and 3' LongSAGE for transcriptome characterization and genome annotation. PNAS USA 2004, 101:11701-11706.

7. Ng P, Wei CL, Sung WK, Chiu KP, Lipovich L, Ang CC, Gupta S, Shahab A, Ridwan A, Wong CH, Liu E, Ruan Y: Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. Nature Methods 2005, 2:105-111.

8. The FANTOM Consortium: The transcriptional landscape of the mammalian genome. Science 2005, 309:1559-1563.

9. Wei CL, Wu Q, Vega V, Chiu KP, Ng P, Zhang T, Shahab A, Ridwan A, Fu Y, Weng Z, Liu J, Kuznetsova VA, Sung K, Lim B, Liu E, Chan QY, Ng HH, Ruan Y: A global mapping of p53 transcription factor binding sites in the human genome. Cell 2006, 124:207-219.

10. Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J, Wong KY, Sung KW, Lee CWH, Zhao X-D, Chiu K-P, Lipovich L, Kuznetsova VA, Robson P, Stanton LW, Wei CL, Ruan Y, Lim B, Ng HH: The Oct4 and Nanog transcription network that regulates pluripotency in mouse embryonic stem cells. Nature Genetics 2006, 38:431-440.

11. The ENCODE Project Consortium: The ENCODE (ENCyclopedia of DNA Elements) Project. Science 2004, 306:636-640 [http://www.genome.gov/Pages/Research/ENCODE/]

12. Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B, Kinzler KW: Gene expression profiles in normal and cancer cells. Science 1997, 276:1268-1272.

13. van Kampen AHC, van Schaik BDC, Pauws E, Michiels EMC, Ruiters JM, Canon HN, Versteeg R, Heisterkamp SH, Leunissen JAM, Baas F, van der Mee M: USAGE: a web-based approach towards the analysis of SAGE data. Bioinformatics 2000, 16:899-905.

14. Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Autschul SF: SAGEmap: A public Gene Expression Resource. Genome Research 2000, 10:1051-1060.

15. Bala P, Georgantas RW 3, Sudhir D, Shankar K, Vrushaben-dra BM, Civin CI, Pandey A: TAGmapper: a web-based tool for mapping SAGE tags. Gene 2005, 364:123-9.

16. Louie E, Ott J, Majewski J: Nucleotide frequency variation across human genes. Genome Research 2003, 13:2594-2601.

Publish with BioMed Central and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:
http://www.biomedcentral.com/info/publishing_adv.asp