Benefits of Random-Priming: Exhaustive Survey of a cDNA Library From Lung Tissue of a SARS Patient

Hongkai Wu, Jinwen Wang, Riqiang Deng, Ke Xing, Yuanyan Xiong, Junfeng Huang, Xionglei He, and Xunzhang Wang*

State Key Laboratory of Biocontrol, Sun Yat-sen University, Xingangxi Road, Guangzhou, People's Republic of China

The severe acute respiratory syndrome (SARS) leads to severe injury in the lungs with multiple factors, though the pathogenesis is still largely unclear. This paper describes the particular analyses of the transcriptome of human lung tissue that was infected by SARS-associated coronavirus (SARS-CoV). Random primers were used to produce ESTs from total RNA samples of the lung tissue. The result showed a high diversity of the transcripts, covering much of the human genome, including loci which do not contain protein coding sequences. 10,801 ESTs were generated and assembled into 267 contigs plus 7,659 singletons. Sequences matching to SARS-CoV RNAs and other pneumonia-related microbes were found. The transcripts were well classified by functional annotation. Among the 7,872 assembled sequences that were identified as from human genome, 578 non-coding genes were revealed by BLAST search. The transcripts were mapped to the human genome with the restriction of identity = 100%, which found a candidate pool of 448 novel transcriptional loci where EST transcriptional signal was never found before. Among these, 13 loci were never reported to be transcriptional by other detection methods such as gene chips, tiling arrays, and paired-end ditags (PETs). The result showed that random-priming cDNA library is valid for the investigation of transcript diversity in the virus-infected tissue. The EST data could be a useful supplemental source for SARS pathology researches. ***J. Med. Virol. 83:574–586, 2011.***
© 2011 Wiley-Liss, Inc.

KEY WORDS: SARS-associated coronavirus; random primer; cDNA library; transcript diversity

INTRODUCTION

During the winter of 2002–2003, a newly emerged infectious disease, the severe acute respiratory syndrome (SARS), appeared in Guangdong Province, China. With a novel coronavirus, the SARS-associated coronavirus, as the etiologic agent [Drosten et al., 2003; Ksiazek et al., 2003], the disease spread rapidly across the world, infecting individuals in several countries with a fatality rate of ~10% [Gu and Korteweg, 2007]. In this study, a cDNA library was generated from the autopsy lung tissue of a SARS-associated coronavirus (SARS-CoV) infected patient by using random primers. cDNA cloning is a fundamental technology for researches in gene discovery and transcriptome analysis [Harbers, 2008]. However, standard cDNA missed many transcripts that lack poly(A) tails [Cheng et al., 2005; Kapranov et al., 2007]. Reverse transcription reactions through alternative random primers of six nucleotides would have more tendency to clone transcripts lacking poly(A) tails, such as ncRNAs and ploy(A)–mRNAs, with more chances to find novel transcripts as compared to cDNA library generated by oligo (dT) primers. In addition, the data presented in this paper is also a human cDNA library generated from SARS-CoV infected lung tissue, which might contribute to the human genomic analyses and the SARS pathological researches.

SARS-CoV infection was reported to affect both host cell transcription and translation, and might cause

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: Natural Science Foundation of China; Grant number: 30970139; Grant sponsor: Science and Technology Project of Guangdong Province; Grant number: 03-33200-4202224; Grant sponsor: Natural Science Foundation of Guangdong Province; Grant number: 4203388; Grant sponsor: Subject of State Key Laboratory of Biocontrol, Sun Yat-sen University; Grant number: SKLBC08B01.

*Correspondence to: Xunzhang Wang, 135# Xingangxi Road, Guangzhou 510275, People's Republic of China.
E-mail: wangxunz@hotmail.com; wxz@mail.sysu.edu.cn

Accepted 15 November 2010
DOI 10.1002/jmv.22012
Published online in Wiley Online Library (wileyonlinelibrary.com).

© 2011 WILEY-LISS, INC.

inflammation, alter immune and stress responses, and modify the coagulation pathways [Enjuanes et al., 2006]. SARS-CoV infection has been investigated by several measures. Human peripheral blood mononuclear cells [Ng et al., 2004; Yu et al., 2005], the human hepatoma cell line [Tang et al., 2005], the animal model of macaques [de Lang et al., 2007], and young and aged mice [Baas et al., 2008] that were infected with SARS-CoV have been Microarray and RT-qPCT analyzed to investigate the up- and down-regulated genes and to explain the pathogenesis at the early time point after SARS-CoV infection, or in the acute severe phase and convalescent phase to understand the gene expression profiles. Another analysis systematically profiled host responses of 40 clinically well-described SARS patients during phases of illness from the onset of symptoms to discharge or a fatal outcome by sampling peripheral blood from the patients [Cameron et al., 2007].

The SARS-CoV library was then thoroughly studied and the gene products were characterized through GO annotation, the Gene Ontology Annotation project of the European Bioinformatics Institute [Harris et al., 2004], which describes the biological process, cellular component, and molecular function of generic cells. GO annotation has been successfully used to analyses of differentially expressed genes before and after the virus infection [Leu et al., 2007]. The knowledge on GO also provides a systematic inquiry and functional classification of etiologies of multifactorial diseases, increasing information that could improve the planning and the treatment of these illnesses [Philip-Couderc et al., 2004; Prabakaran et al., 2004].

Many libraries that were sampled from human lung tissue of fetuses, tumor patients, or healthy adults are available in the expressed sequence tag (EST) division of GenBank, dbEST [Boguski et al., 1993]. The amount of ESTs in dbEST (release 062008) that have been generated from human tissues currently exceeds 800 million. Through GO annotation, the transcripts of the SARS-CoV infected lung library was categorized by gene functions and the abundances of the categories were compared with those of other libraries that were from human lung tissue and were downloaded from dbEST.

MATERIALS AND METHODS

Lung Sample and RNA Isolation

A piece of lung tissue of a SARS patient was excised and frozen in liquid nitrogen immediately after his death. The research of the autopsy sample (kindly provided by professor Hongzhang Huang, Second Affiliated Hospital of Sun Yat-sen University, Guangzhou, China) was after informed consent by the patient's family members and in accordance with Helsinki principles. Total RNA was extracted from this tissue sample using Trizol reagent (Gibco–BRL, Carlsbad, CA). RNAse-free DNAse was used to remove genomic DNA contamination from RNA preparation.

cDNA Library Construction

The first-strand cDNA was generated from total RNA by using random hexamer primers and SuperScript™ II (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. The second-strand cDNA was synthesized by *E. coli* DNA polymerase I and primers were created by RNase H using the first-strand cDNA as template. Residual nicks were then repaired by *E. coli* DNA ligase and the frayed termini of the double-stranded cDNA were polished by bacteriophage T4 DNA polymerase. Finally, purified (Qiaquick PCR Purification Kit, Qiagen, Hilden, Germany) double-stranded cDNA was ligated into pGEM-T-Easy (Promega, Madison, WI) and was transformed into *E. coli* cell DH10B.

EST Sequencing

Di-deoxy terminator sequencing was performed from the 3′ end of the cDNA clones using M13 forward (5′-GTTTTCCCAGTCAC-3′) sequencing primer in a 96-well format via cycle sequencing. After thermal cycling, sequencing reactions were processed and analyzed on a MegaBACE capillary sequencer.

Sequence Data Processing

Vector sequences at both sides of ESTs were trimmed according to the result of BlastN searches against the pGEM-T-Easy vector sequence. ESTs were then assembled by CAP3 [Huang and Madan, 1999] with default parameters. Contigs and singlets were referred to as assembled sequences or transcripts. GMAP [Wu and Watanabe, 2005] was used to map assembled sequences to human genome assembles (NCBI Build 36). Unmapped sequences were Blast at NCBI nt database. Human exogenous sequences were then excluded. 7,872 assembled sequences were defined as human sequences, and the concerned ESTs, that is, ESTs from human genome but excluding those from ribosome and mitochondria, were submitted to NCBI under accession numbers: GD255082–GD264846.

Analysis of Protein Coding and Non-Coding Genes

Human RNA and protein sequences were collected from the NCBI RefSeq dataset (Release 26) [Pruitt et al., 2007]. BlastN and BlastX analyses were performed (E-value $< 1E-10$) between the 7,872 assembled human sequences and these two datasets. The numbers of best-matching sequences in a range of identity and coverage were counted by a PERL script. Immune related sequences were collected according to the annotations of best-matching sequences. BlastN program was also used to compare the 7,872 assembled sequences with non-coding genes stored in miRBase [Griffiths-Jones et al., 2006] database (Release 13.0, miRNA hairpins dataset) and RNAdb (released date: September 2006) [Pang et al., 2005]. Matched sequences were reported at a cut-off E-value of $1E-7$.

Addressing Candidate Novel Transcriptional Loci

Four thousand eleven transcripts with lengths longer than 200 bp and having no significant similarity with NCBI RefSeq human datasets (E-value cut off: 1E−10) were mapped to UCSC [Zweig et al., 2008; Kuhn et al., 2009] human genome (March 2006 assembly) by DNA BLAT search. The loci having best BLAT identities were manually explored through UCSC genome browser, with the configurations remained to the default values, and the tracks of "Human mRNAs" (data last updated: 21 March, 2009), "Human ESTs" (data last updated: 05 December, 2008), "Spliced ESTs," "H-Inv," "UniGene," "SIB Alt-Splicing," "GIS PET RNA" and all the tracks in the categories of "Pilot ENCODE Transcription" [Birney et al., 2007] set to "full" option. All the tracks of tiling arrays and "RNA-seq" in the categories of "Expression" also set to "full" option, the parameters of "Row signal" were set to "hide." These tracks record known transcriptional signals detected by different methods. The genome transcriptional signal at each locus was viewed and recorded through the track display in the browser. A locus was identified as novel if all the tracks showed no record.

Gene Ontology Annotation

Other Six cDNA libraries (Table I) of human lung tissue were downloaded from the NCBI dbEST public database and were assembled by CAP3 the same way as the SARS-CoV library was assembled. Each of the seven libraries was searched against EBI IPI (human 3.32) [Kersey et al., 2004] dataset using BlastX program (E-value < 1E−10). In the Blast result, only the query sequences with length larger than 100 bp were reserved, which were then annotated by obtaining Gene Ontology (GO) IDs of the best-matching sequences from the EBI human GOA (V53.0) [Camon et al., 2004]. The stand along program GOBU [Lin et al., 2006] and the WEGO [Ye et al., 2006] web service were exploited to categorize the GO IDs based on the GO terms and hierarchical structure [Ashburner et al., 2000]. These two programs led to the same GO result, that is, these two programs output the same relative GO ID counts of the assembled sequences falling to the main GO categories and to the second level GO categories. During the comparison of the annotation between the transcripts in the SARS-CoV library and in the other six libraries, only those second level GO categories with significant relative count differences were considered and discussed where the relative counts of the SARS-CoV library were the highest or the lowest among the seven libraries. The statistical significance of the relative count differences between the assembled sequences in the SARS-CoV library and in the other six libraries was determined by using Wilcoxon Signed-Rank Test with P-value < 0.05.

RESULTS

Generation and Assembly of cDNA Sequences

A cDNA library was generated from the autopsy lung tissue of a SARS patient, with a total of 10,801 DNA clones that ranged from 40 to 791 bp in length after vector fragments were manually clipped. Low complexity DNA sequences were not removed to preserve as much non-coding region information as possible in the random primer library.

The sequences were assembled by using CAP3, which extended the maximal clone size to 6,193 bp and assembled 3,142 ESTs into 267 unique clusters, generating 7,926 unique sequences. The sequence length distribution, before and after assembly, is shown in Figure 1.

Mapping to Human Genome and Identifying Foreign Sequences

Among the 7,926 assembled sequences, 7,264 sequences (91.65%) aligned to human genome when mapping the transcripts to human genome assembles (NCBI Build 36) by using the program GMAP with the cut off parameters of coverage $\geq$ 80% and identity $\geq$ 90%. Lowering the GMAP cut off parameters to coverage $\geq$ 50% and identity $\geq$ 80% resulted in 7,675 matched sequences (96.83%). And more matched sequences (7,861 sequences, 99.18%) were obtained when the GMAP cut off parameters were lowered to coverage $\geq$ 10% and identity $\geq$ 45%.

Even in the loosest condition (coverage $\geq$ 10% and identity $\geq$ 45%), there were still 65 assembled sequences (0.82%) that did not match to human genome. These sequences were further BlastN searched against NCBI nt database and were divided into several groups (Table II) based on the similarity hits on sequences. Among these, 11 sequences matched to human sequences, which could be attributed to incomplete or alternate genome assembles; 10 SARS-CoV RNA sequences were

TABLE I. Other cDNA Libraries of Human Lung Tissue Used in This Study for Comparison Analysis

Library name	dbEST ID	Description[a]	EST number
MGC101	dbEST:10453	Epidermoid carcinoma	9,166
MGC69	dbEST:5608	Large cell carcinoma, undifferentiated	9,748
dbEST16438	dbEST:16438	Fetal fibroblast	3,339
UICFEC1	dbEST:10395	Normal lung from adult and from fetal day 64, day 87, week 19, and week 42	12,971
UICFDU1	dbEST:10398	Adult primary lung epithelial cells	12,742
dbEST249	dbEST:249	Male, 72 years old	13,244

[a]All these cDNA libraries were generated by oligo-dT priming.

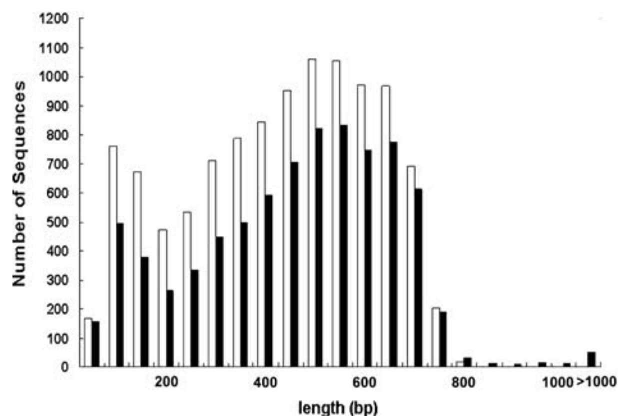

Fig. 1. The length distributions for ESTs before and after assembly. The lengths of assembled ESTs (black bars) were prolonged comparing with that of ESTs before assembly (open bars), with the longest ESTs exceeding 6 kbp.

found, which were the clustering of 162 ESTs that accounted for 1.50% of the total cDNA library; 12 sequences were found highly similar with pneumonia-related microbes (Table III); and some sequences had no significant E-value (>0.001) to any known sequences in the nt database, which were defined as non-information sequences.

The remaining 7,861 assembled sequences that matched to human genome plus the 11 sequences matching to human sequences, which were composed of 7,614 singletons and 258 contigs from 2,980 ESTs, were considered as transcripts from human lung tissue and were used for further analyses. The concerned ESTs, that is, ESTs from human genome but excluding those from ribosome (810 ESTs) and mitochondria (19 ESTs), have been deposited in public databases [GenBank accession numbers: GD255082–GD264846].

Similarity Search Against NCBI Refseq Database and Immune Related Genes

The 7,872 assembled sequences from human lung tissue were further compared to the sequence collection that was annotated as species *Homo sapiens* in the NCBI RefSeq dataset (Release 26). Distribution of identity and coverage values of the best BlastN and BlastP alignments is shown in Figure 2.

When cut-off E-value was set to 1E−10, 2,878 (36.56%) and 1,110 (14.10%) assembled sequences hit on RefSeq RNA and protein datasets, respectively. Among the 2,878 assembled sequences matching to the RefSeq RNA dataset, 305 matched with length coverage greater than 80% (Fig. 2a). And in the 1,110 assembled sequences matching to the protein datasets, 917 aligned with length coverage from 20% to 80% and percentage identity from 40% to 100% (Fig. 2b). The result demonstrated that, by comparing to human known or predicted sequences in NCBI Refseq dataset, many transcripts in the library were novel.

Result from similarity searching also indicated that many transcripts were highly similar to known genes related with cytokines, chemokines, adhesion molecules, and immune cells, and similar to genes associated with viral infection, for example, CCL5 (28 ESTs), IL28RA (28 ESTs), DQB1 (8 ESTs), VNN1 (6 ESTs), IL12RB1 (5 ESTs), and SIGLEC10 (5 ESTs), which mainly related with anti-inflammatory, pro-inflammatory, and immune regulation (Supplementary Table SI).

Candidate Novel Transcriptional Loci Identified by UCSC Human Genome Browser

Excluding the transcripts matching to NCBI RefSeq human datasets (E-value cut off: 1E−10), the remaining 4,011 transcripts with lengths over 200 bp were mapped to UCSC human genome (March 2006 assembly) by using BLAT search. The transcriptional information at each best matching locus was analyzed with UCSC genome browser, which found 1,364 best matching loci (BLAT identity > 97%) that lacked EST or mRNA signal records, or found 448 best matching loci that lacked EST or mRNA signal records when BLAT identity was set to 100%. Considering other transcriptional detection methods, such as gene chips, tilling arrays, and paired-end ditags (PETs) [Ng et al., 2005], there were still 37 novel loci (BLAT identity > 97%) or 13 novel loci (BLAT identity = 100%) that had never been detected as transcriptional. Even after the great amount of reads obtained by deep-sequencing technology (RNA-seq) [Wang et al., 2009] were taken into consideration by setting UCSC genome browser options for related "RNA-seq" tracks to "full," there were still three novel loci (identity > 97) that had never been detected as transcriptional by UCSC genome browser and other transcriptional detection methods (Table IV).

The result showed a potential extended covering of the transcriptional region in the human genome, which could be due to the use of random primers in generating the cDNA library.

TABLE II. Classification of the 65 Transcripts That Did Not Match to Human Genome

Classes[a]	Number of assembled sequences	Number of ESTs
Transcripts matched to human sequences from nt database	11	11
SARS-CoV RNA sequences	10	162
Sequences similar to pneumonia-related microbes	12	13
Non-information transcripts[b]	32	32
Total	65	218

[a]The classification was based on the annotations of the best matching homologs during BlastN searches of the 65 transcripts against the NCBI nt database.
[b]The class "Non-information transcripts" includes sequences that did not match to any known sequences (E-value >0.001).

TABLE III. Transcripts With Significant Similarity to Pneumonia-Related Microbes

Accession	Sequence length (bp)	Nt accession	Species	Score	Coverage (%)	E-Value	Identity (%)
a0_000908	138	EF061771.1	*M. hominis*	183	80	1E−43	96
da0_004802	128	AF443616.3	*M. hominis*	226	100	1E−56	99
b0_000821	135	AF443616.3	*M. hominis*	232	99	1E−57	98
b0_000840	212	AF125581.1	*M. arginini*	347	100	2E−92	93
Contig267[a]	388	AY737013.1	*M. arginini*	695	100	0.0	99
de0_010065	626	AE004091.2	*P. aeruginosa*	325	72	2E−34	92
de0_012036	427	AF440524.1	*P. aeruginosa*	187	78	4E−44	72
de0_003139	44	AY956411.1	*Stenotrophomonas maltophilia*	50.1	56	4E−04	100
de0_007533	112	AJ746243.1	*S. maltophilia*	141	91	5E−31	92
de0_001693	217	BX640425.1	*Bordetella parapertussis*	150	95	7E−35	76
db0_001118	181	CP000408.1	*Streptococcus suis*	269	100	1E−69	90
db0_003991	129	AF269487.1	*Staphylococcus epidermidis*	129	98	2E−27	82

[a]The contig was assembled from 2 ESTs.

Non-Coding Genes in the Library

Assembled sequences of the library were BlastN searched against miRBase database (Release 13.0), a collection of microRNA sequences with experimental evidence in related organisms. The result showed many matches with high similarity to pri-miRNA sequences in the database, for example, 17, 5, and 5 transcripts matched to miRNA hairpins hsa-mir-1268, hsa-mir-566 (in *H. sapiens*), and ptr-mir-566 (in *pan troglodytes*) respectively, with coverage of over 95% or 100% (Table V). Among these, hsa-mir-566 and ptr-mir-566 belong to miRNA gene family mir-566.

Assembled sequences were also searched against RNAdb (released date: September 2006), a comprehensive mammalian non-coding RNA (ncRNA) database, which is a collection of several datasets that were constructed by different methods and materials. The result showed a total of 573 assembled sequences with high similarity to the known sequences registered in RNAdb. Table VI summarized the distribution of the assembled sequences matching to each dataset.

A Venn was made to show the above alignment results (Fig. 3). The diagram indicated that 1,110 assembled sequences belonged to the class of protein coding sequences which matched to RefSeq protein dataset and 578 assembled sequences belonged to the class of non-coding sequences which matched to sequences from miRBase or RNAdb. Among these, 281 assembled sequences shared between the two classes, which matched to both coding and non-coding datasets by locally aligning to sequences in the datasets with significant similarity.

Functional Classification Adopting the GO Annotation

The assembled sequences were then searched against EBI IPI (human 3.32) human protein dataset by BlastP (E-value cut off: 1E−10). And with the help of the EBI human GOA (v53.0) file, GO annotation IDs of the best-matching homologs in the IPI dataset were assigned to the transcripts, so as to study the provisional gene function distribution in the ontology categories by

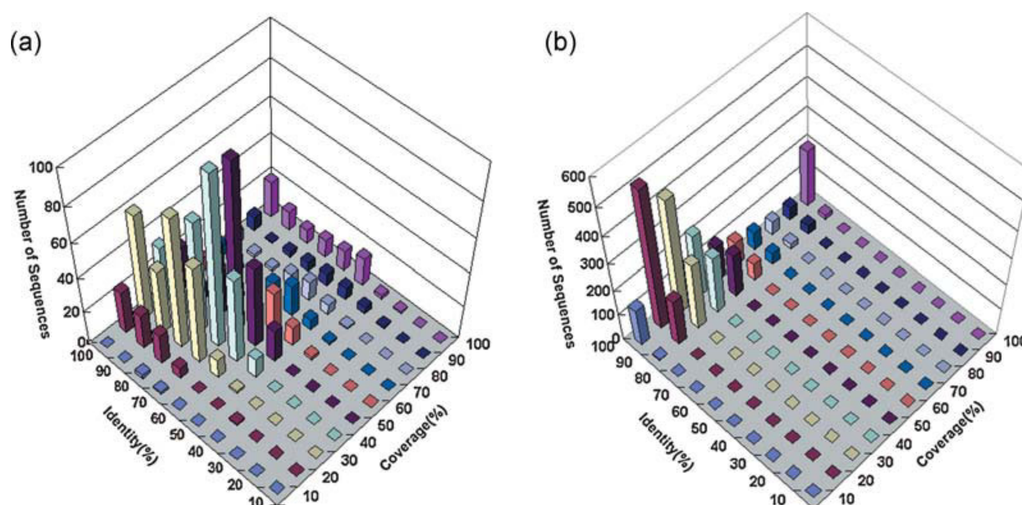

Fig. 2. Statistics of BLAST searches. **a**: BlastP search of assembled sequences against NCBI refseq human protein dataset. **b**: BlastN search of assembled sequences against NCBI refseq human RNA dataset.

TABLE IV. Novel Transcriptional Loci Addressed by UCSC Human Genome Browser

Blat identity (%)	Amounts of loci lacking transcriptional signals					
	Without EST signal	Without EST + method1[a] signals	Without EST + method2[b] signals	Without EST + method3[c] signals	Without EST + method1 + method2 signals	Without EST + method1 + method2 + method3 signals
>97	1,364	443	69	15	37	3
>98	1,318	425	64	15	35	3
>99	1,179	378	56	11	31	2
=100	448	139	21	3	13	1

[a]Method1—gene chips or tiling arrays.
[b]Method2—paired-end ditags (PETs).
[c]Method3—RNA-Seq.

TABLE V. Transcripts With Significant Similarity to MicroRNAs

Accession	Length (bp)	MicroRNA	MiR length (bp)	Coverage (%)	E-Value	Identity (%)
a0_001869	342	hsa-mir-1268	52	100	2E−11	90.38
Contig135	2,438			100	1E−10	90.38
Contig161	1,589			100	7E−11	90.38
Contig167	1,344			100	6E−11	90.38
Contig203	2,565			100	1E−10	90.38
Contig223	1,208			100	5E−11	90.38
Contig241	1,696			100	3E−13	92.31
Contig68	1,673			100	8E−11	90.38
da0_001367	504			100	9E−14	92.31
de_178	570			100	3E−11	90.38
de0_00060	569			100	3E−11	90.38
de0_003191	683			100	5E−16	94.23
de0_003984	701			100	3E−11	90.38
de0_004380	601			100	1E−13	92.31
de0_005202	701			100	3E−11	90.38
de0_006649	602			100	1E−13	92.31
de0_012147	743			98.08	1E−10	90.2
Contig135	2,438	hsa-mir-566	94	95.74	1E−28	92.22
Contig223	1,208			95.74	2E−26	91.11
Contig232	1,131			95.74	2E−26	91.11
Contig251	814			95.74	1E−26	91.11
de0_008873	649			95.74	9E−27	91.11
Contig135	2,438	ptr-mir-566	93	96.77	1E−28	92.22
Contig223	1,208			96.77	2E−26	91.11
Contig232	1,131			96.77	2E−26	91.11
Contig251	814			96.77	1E−26	91.11
de0_008873	649			96.77	9E−27	91.11

calculating the percentages or the relative counts of the transcripts that fell into each ontology category.

The GO annotations of the CoV-infected library were contrasted with the GO annotations of other six human lung cDNA libraries downloaded from NCBI dbEST public database (Table I). MGC101 and MGC69 [Gerhard et al., 2004] were the libraries from the lung tissues related with carcinoma. Libraries dbEST16438 [Suzuki et al., 2004] and UICFEC1 [Bonaldo et al., 1996] contained ESTs mainly from the fetal lung tissues. Libraries UICFDU1 [Bonaldo et al., 1996] and dbEST249 [Hillier et al., 1996] were ESTs from normal adult lung tissues. These six libraries were assembled and functional annotated in the same way as the CoV-infected lung ESTs were processed.

For the three main GO categories (biological process, molecular function, and cellular component), the relative counts of GO annotation IDs were set to 100%. Since

TABLE VI. Statistics of Transcripts with Significant Similarity to Different Datasets of RNAdb[a]

Datasets of RNAdb	rnaz	ncrnascan	evofold	fantom3	asoverlaps	snorna	hinv	combinedlit	pirna
Number of matched transcripts	56	1	26	42	18	0	11	12	447

[a]For dataset “pirna” the cut off values of coverage = 100%, identity = 100%, and E-value < 1E−7 were used; for other datasets, the cut off values of coverage > 90%, identity > 90%, and E-value < 1E−7 were used.

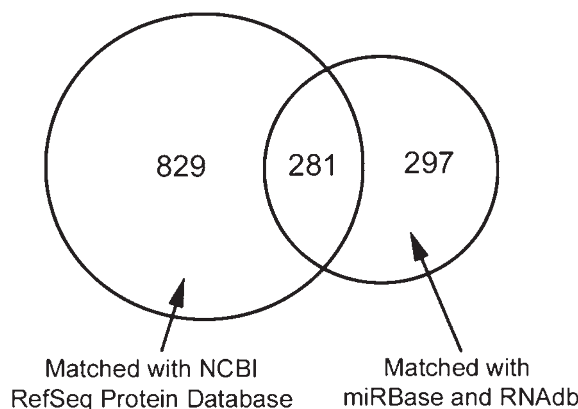

Fig. 3. Venn diagram of the transcripts with similarities to sequences in coding and non-coding datasets, showing the coding sequence class and the non-coding sequence class, and an overlapping of 281 transcripts that shared between the two classes.

a transcript could match to homologs of different functions, therefore it was assigned different GO annotation IDs, that is, the transcript might fall into more than one second-level GO categories, the relative counts of the second-level GO categories inside a main category might add up over 100% owing to the great amount of tentative multiple function genes in the main category. The CoV-infected library showed a significant small sum up value (257.91%, P-value $= 0.028$) of the relative counts of the second-level GO categories in the biological process main category, implying less tentative multiple function genes in the CoV-infected lung tissue concerning biological process.

The relative counts of the CoV-infected library were found to be significantly different from the relative counts of the other six human lung cDNA libraries (P-value < 0.05 by Wilcoxon Signed-Rank Test) in 31 second-level GO categories (Table VII, a full vision of Table VII was provided in Supplementary Table SIII). Among these, categories "virion" (6.89%), "extracellular region part" (7.66%), "virion part" (6.89%), "synapse part" (0.38%), "synapse" (1.14%), "motor activity" (2.26%), "auxiliary transport protein activity" (0.26%), "molecular transducer activity" (11.02%), "viral reproduction" (1.18%), and "multicellular organismal process" (17.15%) exhibited significantly higher relative counts in the CoV-infected library (highest among the seven libraries). While in many categories with important functions, such as "enzyme regulator activity" (2.36%), "translation regulator activity" (0.52%), "immune system process" (2.36%), "metabolic process" (65.08%), "developmental process1" (16.56%), "growth" (1.18%), "response to stimulus" (6.21%), "localization" (17.15%), and "biological regulation" (19.23%), the CoV-infected library showed significantly lower relative counts (lowest among the seven libraries).

Statistics of the annotation analysis (Table VIII) indicated that the library tended to contain more novel transcripts (78.5%) comparing to other lung tissue libraries (5.4–71.2%). In the CoV-infected library, transcripts matching to IPI proteins accounted for only 21.45%, and the sequences with GO assignments accounted for only 25.18% of the matched sequences. While in the other libraries, transcripts matching to IPI proteins accounted for 28.7–94.5%, and the sequences with GO assignments accounted for 59–78% of the matched sequences. The result suggested that a random-primer-generated library is valid for transcript diversity.

DISCUSSION

RT-PCR has been used to detect SARS-CoV in lung autopsy material in several cases [Nicholls et al., 2003, 2006; Ding et al., 2004]. Although most positive results have been achieved by sequence-specific primers, PCR strategy using random primers in conjunction with the microarray was reported effective in detecting RNA viruses from potato tissue culture plantlets and human respiratory specimens [Wang et al., 2002; Agindotan and Perry, 2007]. The library used in this study was a reverse-transcribed cDNA library generated with random hexamer primers from total RNA of SARS-CoV infected lung tissue. There were 162 ESTs in the library that were identified as SARS-CoV sequences, which accounted for about 1.5% of the total cDNAs, indicating a relatively high viral load in the sample.

This study also detected sequences from several pneumonia-associated mycoplasma and bacteria. Clones were found matching with high similarity to the sequences of the species *M. hominis*, *Mycoplasma arginini*, *Pseudomonas aeruginosa*, and *Stenotrophom onasmaltophilia*, implying the presence of co-infections in this case, which were probably related to longer disease durations. Co-infections have been reported in several SARS cases [Nicholls et al., 2003; Chong et al., 2004; Hwang et al., 2005]. SARS-CoV seems to impair the phagocytic capacity of macrophages, which may render SARS patients prone to secondary pulmonary infections [Tseng et al., 2005]. And invasion of the viral pneumonia damaged lungs by pathogenic bacteria may result in secondary bacterial pneumonia [Strauss and Strauss, 2002].

Although clinical evaluations of immune response have been carried out at different time points in the course of the SARS infection, the observations cannot be confirmed because the same immune mediators were not sought or found by different investigators [Chen and Subbarao, 2007]. The library from this study provided a set of active genes that might be associated with the host immune response (Additional file 1: Table SII) in lung at the death stage of SARS infection, for example, a large number of transcripts associated with cytokines, chemokines, and their receptors. Interestingly, transcripts associated with the hepatitis A virus cellular receptor 2, the HIV-1 binding protein, and the virus-induced signaling adapter were identified, suggesting the sharing of some mechanisms among SARS-CoV infection and the infections of other RNA viruses.

ESTs matching with RNA datasets accounted for 36.56%, while ESTs matching with NCBI RefSeq human

TABLE VII. Gene Ontology Comparison of the SARS-CoV Library against Other Six Human Lung Tissue Libraries

Categories	Libraries (brief description)							P-value
	SARS-CoV Inf	MGC101 (epidermoid carcinoma)	MGC69 (large cell carcinoma)	dbEST16438 (fetal fibroblast)	UICFEC1 (normal, adult + fetal)	UICFDU1 (normal, adult)	dbEST249 (normal, 72 year old)	
Cellular component	261 (100%)	2,735 (100%)	1,962 (100%)	962 (100%)	1,846 (100%)	1,376 (100%)	951 (100%)	
Cell	239 (91.57%)↓[a]	2,620 (95.79%)	1,894 (96.53%)	917 (95.32%)	1,721 (93.22%)	1,306 (94.91%)	872 (91.69%)	0.028
Virion	18 (6.89%)↑	1 (0.03%)	0 (0.00%)	0 (0.00%)	3 (0.16%)	1 (0.07%)	5 (0.52%)	0.027
Envelope	8 (3.06%)↓	126 (4.60%)	413 (21.04%)	43 (4.46%)	85 (4.60%)	81 (5.88%)	47 (4.94%)	0.027
Macromolecular complex	39 (14.94%)↓	779 (28.48%)	772 (39.34%)	237 (24.63%)	503 (27.24%)	411 (29.86%)	294 (30.91%)	0.028
Organelle	131 (50.19%)↓	1,850 (67.64%)	1,439 (73.34%)	633 (65.80%)	1,163 (63.00%)	910 (66.13%)	576 (60.56%)	0.028
Extracellular region part	20 (7.66%)↑	88 (3.21%)	54 (2.75%)	42 (4.36%)	96 (5.20%)	31 (2.25%)	50 (5.25%)	0.028
Organelle part	68 (26.05%)↓	836 (30.56%)	818 (41.69%)	263 (27.33%)	517 (28.00%)	422 (30.66%)	282 (29.65%)	0.028
Virion part	18 (6.89%)↑	1 (0.03%)	0 (0.00%)	0 (0.00%)	3 (0.16%)	1 (0.07%)	5 (0.52%)	0.027
Synapse part	1 (0.38%)↑	4 (0.14%)	1 (0.05%)	0 (0.00%)	1 (0.05%)	0 (0.00%)	3 (0.31%)	0.027
Cell part	239 (91.57%)↓	2,620 (95.79%)	1,894 (96.53%)	917 (95.32%)	1,721 (93.22%)	1,306 (94.91%)	872 (91.69%)	0.028
Synapse	3 (1.14%)↑	8 (0.29%)	3 (0.15%)	3 (0.31%)	7 (0.37%)	0 (0.00%)	4 (0.42%)	0.028
Molecular function	381 (100%)	3,129 (100%)	2,058 (100%)	1,077 (100%)	2,106 (100%)	1,593 (100%)	1,121 (100%)	
Motor activity	10 (2.62%)↑	35 (1.11%)	16 (0.77%)	16 (1.48%)	19 (0.90%)	16 (1.00%)	11 (0.98%)	0.028
Auxiliary transport protein activity	1 (0.26%)↑	0(0.00%)	0 (0.00%)	2 (0.18%)	3 (0.14%)	2 (0.12%)	0 (0.00%)	0.026
Chaperone regulator activity	0 (0.00%)↓	2 (0.06%)	2 (0.09%)	0 (0.00%)	5 (0.23%)	4 (0.25%)	1 (0.08%)	0.043
Enzyme regulator activity	9 (2.36%)↓	148 (4.72%)	63 (3.06%)	60 (5.57%)	86 (4.08%)	60 (3.76%)	54 (4.81%)	0.028
Translation regulator activity	2 (0.52%)↓	81 (2.58%)	55 (2.67%)	25 (2.32%)	43 (2.04%)	41 (2.57%)	30 (2.67%)	0.027
Molecular transducer activity	42 (11.02%)↑	176 (5.62%)	108 (5.24%)	81 (7.52%)	122 (5.79%)	84 (5.27%)	76 (6.77%)	0.028
Biological process	338 (100%)	2,746 (100%)	1,943 (100%)	975 (100%)	1,901 (100%)	1,403 (100%)	1,020 (100%)	
Cell killing	0 (0.00%)↓	13 (0.47%)	3 (0.15%)	0 (0.00%)	5 (0.26%)	2 (0.14%)	1 (0.09%)	0.043
Immune system process	8 (2.36%)↓	108 (3.93%)	66 (3.39%)	37 (3.79%)	94 (4.94%)	59 (4.20%)	79 (7.74%)	0.028
Metabolic process	220 (65.08%)↓	1,856 (67.58%)	1,474 (75.86%)	669 (68.61%)	1,331 (70.01%)	1,027 (73.20%)	743 (72.84%)	0.028
Viral reproduction	4 (1.18%)↑	11 (0.40%)	12 (0.61%)	2 (0.20%)	5 (0.26%)	8 (0.57%)	6 (0.58%)	0.028
Multicellular organismal process	58 (17.15%)↑	373 (13.58%)	218 (11.21%)	131 (13.43%)	289 (15.20%)	165 (11.76%)	133 (13.03%)	0.028
Developmental process	56 (16.56%)↓	678 (24.69%)	331 (17.03%)	237 (24.30%)	442 (23.25%)	294 (20.95%)	205 (20.09%)	0.028
Growth	4 (1.18%)↓	62 (2.25%)	23 (1.18%)	14 (1.43%)	32 (1.68%)	18 (1.28%)	14 (1.37%)	0.043
Locomotion	0 (0.00%)↓	21 (0.76%)	3 (0.15%)	7 (0.71%)	7 (0.36%)	2 (0.14%)	4 (0.39%)	0.028
Rhythmic process	0 (0.00%)↓	12 (0.43%)	3 (0.15%)	5 (0.51%)	5 (0.26%)	5 (0.35%)	1 (0.09%)	0.028
Response to stimulus	21 (6.21%)↓	322 (11.72%)	225 (11.58%)	128 (13.12%)	247 (12.99%)	178 (12.68%)	151 (14.80%)	0.028
Localization	58 (17.15%)↓	59 (21.63%)	533 (27.43%)	232 (23.79%)	357 (18.77%)	251 (17.89%)	184 (18.03%)	0.028
Establishment of localization	53 (15.68%)↓	514 (18.71%)	489 (25.16%)	202 (20.71%)	317 (16.67%)	232 (16.53%)	163 (15.98%)	0.028
Maintenance of localization	0 (0.00%)↓	18 (0.65%)	14 (0.72%)	7 (0.71%)	6 (0.31%)	10 (0.71%)	8 (0.78%)	0.027
Biological regulation	65 (19.23%)↓	876 (31.90%)	431 (22.18%)	317 (32.51%)	581 (30.56%)	37 7(26.87%)	277 (27.15%)	0.028

[a]Only those second level GO categories with significant relative count differences were considered and discussed where the relative counts of the SARS-CoV library were the highest or the lowest among the seven libraries. Difference significances were tested by the statistic method of Wilcoxon Signed-Rank test with P-value < 0.05.

TABLE VIII. Statistical Comparison of Transcripts Matching to Annotated Sequences Among the Seven Libraries

Libraries	EST numbers	Numbers of assembled sequence (ESTs per. assembled sequence)	Numbers of sequences matched to IPI proteins (/sequence number %)	Numbers of matched sequences with GO assignments (/matched number %)
SARS-CoV Inf	10,594	7,872 (1.35)	1,689 (21.45%)	425 (25.18%)
MGC101	9,166	5,250 (1.75)	4,446 (84.69%)	3,391 (76.27%)
MGC69	9,748	5,366 (1.82)	2,929 (54.58%)	2,249 (76.78%)
UICFEC1	12,971	7,455 (1.74)	3,509 (47.07%)	2,335 (66.54%)
dbEST16438	3,339	1,598 (2.09)	1,511 (94.56%)	1,181 (78.16%)
UICFDU1	12,742	7,383 (1.73)	2,861 (38.75%)	1,766 (61.73%)
dbEST249	13,244	7,244 (1.83)	2,081 (28.73%)	1,229 (59.06%)

datasets accounted for only 14.10% in the library from this study (Fig. 2), which was consistent with the fact that less than 2% of the human genome is translated into protein, yet more than 40% of the genome is thought to be transcribed into RNA [Cheng et al., 2005]. It was reported that a large proportion of the transcripts in human and mouse were unique to the largely unstudied poly(A−) fractions of the transcriptome [Cheng et al., 2005; Kiyosawa et al., 2005]. In addition, recent studies revealed that protein-coding loci are more transcriptionally complex than previously thought, and there is often no clear distinction between splice variants and overlapping and neighboring genes [Carninci et al., 2005; Birney et al., 2007]. This might explain to some extent why, in the present study, most ESTs matching to known proteins with coverage less than 70% (Fig. 2).

It appears that much more of the human genome is transcribed than was previously appreciated. Genomic tiling arrays revealed a large number of novel transcribed regions throughout the ENCODE region from different cell types and NB4 cells treated with different conditions [Rozowsky et al., 2006]. 5′-RACE and 3′-RACE sequencing uncovered transcription in previous non-transcribed regions of the human genome [Wu et al., 2008]. These two studies focused on the poly(A+) RNAs and transcripts were obtained by specific primers. The present study revealed 448 novel transcriptional loci lacking EST or mRNA signal records (Table IV). Among these, at least 37 (BLAT identity $> 97\%$) or 13 (BLAT identity $= 100\%$) loci were not detected before as transcriptional by other methods such as gene chips, tiling arrays, and PETs, indicating the benefit of random primer method in transcriptome studies.

Nowadays, a great amount of short reads obtained by the deep-sequencing technology (RNA-seq) is covering more of the human transcriptome. After taking into consideration of the RNA-seq, at least 15 novel transcriptional loci (BLAT identity $> 97\%$) were still found where EST or RNA-seq records were lacking (Table IV). Additionally, RNA-seq libraries downloaded from NCBI were mapped to the transcripts of the SARS-CoV library and the "reads of match per kilo base" was calculated. The result showed that novel ESTs (1,364 transcripts without EST signal, see Table IV) in the library exhibited relatively poor matches (Supplementary Table SIV), indicating that novel ESTs in the library had less chance of presence in the datasets obtained by deep-sequencing technologies.

This study also revealed 281 assembled sequences that matched with both protein sequence data and ncRNA sequence data (Fig. 3), which might result from the parallel production of protein-coding sequences, regulatory ncRNAs, and introns. The latter, perhaps as a major source of regulatory ncRNAs, accounted for at least 30% of the human genome [Mattick and Makunin, 2006]. Some of the ncRNAs including microRNAs found in this study might be related with the pathology of SARS-CoV infection and worth for further studying. Recent reports provided strong evidence that microRNAs and other ncRNAs are linked to the pathological conditions of some human diseases [Costa, 2007]. Viral transcripts were also regulated by Host-encoded miRNAs. Human miR-32 has been reported to restrict the replication of primate foamy virus type 1 (PFV-1) [Lecellier et al., 2005]. The 3′ portion of HIV-1 mRNAs was shown to be redundantly targeted for repression by a cluster of human miRNAs [Huang et al., 2007]. It was demonstrated that IFNα/β up regulated several cellular miRNAs that are capable of inhibiting hepatitis C virus (HCV) replication and infection. Down regulation of miR-122 in response to IFNβ further contributes to the antiviral effects [Pedersen et al., 2007].

In addition, function annotation and comparison of the SARS-CoV library with other six libraries also revealed the SARS-CoV infected lung library a significantly higher relative GO annotation ID counts of categories "virion" and "virion part," which belonged to human endogenous retrovirus (HERV) genes, mainly HERV-H, HERV-K, and HERV-W being identified in this study. Genome sequencing revealed that 8% of the human genome consists of HERVs [Medstrand et al., 2002; Bannert and Kurth, 2004]. Different endogenous retroviruses also exhibit differential responsiveness to environmental signals and activation of the immune system [Taruscio and Mantovani, 2004]. It was reported that the expression of HERVs was affected by cytokines [Katsumata et al., 1999; Schneider et al., 2001] and steroids [Ono et al., 1987]. Recent report showed that HIV-1 infection leads to HERV expression and stimulation of a HERV-specific CD8+ T cell response [Garrison et al., 2007]. The high level HERVs expression in this case might imply the steroid hormone therapy in the severe patient.

According to the function annotation results, synapse-associated proteins were also over expressed in the SARS-CoV infected library (0.38% in "synapse part" and 1.14% in "synapse"). Proteins in these categories were lipid-binding proteins in the synaptic vesicle, such as synaptotagmin and synaptoporin, and the postsynaptic element, such as cholinergic receptor. Since a line of evidence showed that steroid hormone increased synapse density [Jelks et al., 2007] or up-regulated the expression of synaptic proteins [Crispino et al., 1999; Rune et al., 2002; Kretz et al., 2004], the high level expression of synaptic proteins in this case should have resulted from the high dose corticosteroids therapy in the severe patient.

"Motor activity" was another category that was over expressed in the SARS-CoV infected library (2.62%). These proteins were isoforms of myosims, Dynein heavy chains, and Kinesin-like proteins. Since many studies have indicated that viruses exploit cytoskeletal dynamics and viral particles recruit molecular motors in order to hitchhike rides to different subcellular sites [Fackler and Krausslich, 2006; Greber and Way, 2006; Radtke et al., 2006], the high level expression of motor proteins in this case might suggest the active intracellular transportation of SARS-CoV particles.

Besides "motor activity," categories "molecular transducer activity" and "multicellular organismal process" also exhibited the largest relative counts in the SARS-CoV infected library (11.02% and 17.15%, respectively) comparing to other libraries. Many sequences matching with proteins involved in cytoskeletal dynamics and endocytosis were found (Supplementary Table SII). The result was consistent with the result of a recent study that SARS-CoV entered cells through direct fusion with the plasma membrane as well as a clathrin- and caveolae-independent endocytic mechanism mediated by cholesterol- and sphingolipid lipid raft microdomains [Wang et al., 2008] and also consistent with the results of other studies that cytoskeleton reorganization and cellular signal transduction as well as endocytosis were inextricably linked [Balklava and Grant, 2005; Pelkmans, 2005; Pelkmans et al., 2005]. The result suggested the active processes of viral endocytosis and viral particle spread in the SARS-CoV infected tissue.

Although one should remember that the above function annotation results have been obtained by comparison between the SARS-CoV library and other six libraries that were generated with a different methodology and therefore hard to compare, the comparison did provide us with helpful and valuable information, which could be useful for future researches on related fields.

The results from the present study demonstrated that ESTs generated by random primers had favorable advances for revealing the transcript diversity in human transcriptome. In other words, the random primer library provided complex information of the total RNA in a specific tissue. Exhaustive survey of the cDNA library generated with random primers from lung tissue of a SARS patient obtained at least five important results: (1) foreign sequences including the SARS-CoV virus and other agents of accompanying infections were identified; (2) many sequences matching with significant similarity to the known or predicted ncRNAs were revealed; (3) revealed a majority of sequences matching to human protein sequences with significant similarity, but with relatively low coverage, and sequences matching to both human protein sequences and ncRNAs simultaneously; (4) lots of sequences were mapped to novel transcriptional regions in human genome with 100% identity; (5) the library tended to contain more novel transcripts comparing to other lung tissue libraries (Table VIII).

ACKNOWLEDGMENTS

Thanks to Guangdong Center for Disease Control and Prevention for providing RNA sample and Beijing Genomics Institute for providing sequencing platform. We are grateful to Prof. Weihua Zheng, Guangdong provincial people's hospital, for her useful discussion.

REFERENCES

Agindotan B, Perry KL. 2007. Macroarray detection of plant RNA viruses using randomly primed and amplified complementary DNAs from infected plants. Phytopathology 97:119–127.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, Consortium GO. 2000. Gene ontology: Tool for the unification of biology. Nat Genet 25:25–29.

Baas T, Roberts A, Teal TH, Vogel L, Chen J, Tumpey TM, Katze MG, Subbarao K. 2008. Genomic analysis reveals age-dependent innate immune responses to severe acute respiratory syndrome coronavirus. J Virol 82:9465–9476.

Balklava Z, Grant BD. 2005. The regulation of endocytosis by kinases: Cell biology meets genomics. Genome Biol 6:245.

Bannert N, Kurth R. 2004. Retroelements and the human genome: New perspectives on an old relation. Proc Natl Acad Sci USA 101:14572–14579.

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaoz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Loytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, Batzoglou S, Goldman N, Hardison RC,

Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Xu M, Haidar JN, Yu Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrimsdottir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447:799–816.

Boguski MS, Lowe TM, Tolstoshev CM. 1993. dbEST—Database for "expressed sequence tags." Nat Genet 4:332–333.

Bonaldo MF, Lennon G, Soares MB. 1996. Normalization and subtraction: Two approaches to facilitate gene discovery. Genome Res 6: 791–806.

Cameron MJ, Ran L, Xu L, Danesh A, Bermejo-Martin JF, Cameron CM, Muller MP, Gold WL, Richardson SE, Poutanen SM, Willey BM, DeVries ME, Fang Y, Seneviratne C, Bosinger SE, Persad D, Wilkinson P, Greller LD, Somogyi R, Humar A, Keshavjee S, Louie M, Loeb MB, Brunton J, McGeer AJ, Kelvin DJ. 2007. Interferon-mediated immunopathological events are associated with atypical innate and adaptive immune responses in patients with severe acute respiratory syndrome. J Virol 81:8692–8706.

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R. 2004. The Gene Ontology Annotation (GOA) Database: Sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res 32:D262–D266.

Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impiombato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta, di Bernardo G, Down D, Engstrom T, Fagiolini P, Faulkner M, Fletcher G, Fukushima CF, Furuno T, Futaki M, Gariboldi S, Georgii-Hemming M, Gingeras P, Gojobori TR, Green T, Gustincich RE, Harbers S, Hayashi M, Hensch Y, Hirokawa TK, Hill N, Huminiecki D, Iacono L, Ikeo M, Iwama K, Ishikawa A, Jakt T, Kanapin M, Katoh A, Kawasawa M, Kelso Y, Kitamura J, Kitano H, Kollias H, Krishnan G, Kruger SP, Kummerfeld A, Kurochkin SK, Lareau IV, Lazarevic LF, Lipovich D, Liu L, Liuni J, McWilliam S, Madan S, Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schonbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y. 2005. The transcriptional landscape of the mammalian genome. Science 309:1559–1563.

Chen J, Subbarao K. 2007. The immunobiology of SARS*. Annu Rev Immunol 25:443–472.

Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, Sementchenko V, Piccolboni A, Bekiranov S, Bailey DK, Ganesh M, Ghosh S, Bell I, Gerhard DS, Gingeras TR. 2005. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science 308:1149–1154.

Chong PY, Chui P, Ling AE, Franks TJ, Tai DY, Leo YS, Kaw GJ, Wansaicheong G, Chan KP, Ean Oon LL, Teo ES, Tan KB, Nakajima N, Sata T, Travis WD. 2004. Analysis of deaths during the severe acute respiratory syndrome (SARS) epidemic in Singapore: Challenges in determining a SARS diagnosis. Arch Pathol Lab Med 128:195–204.

Costa FF. 2007. Non-coding RNAs: Lost in translation? Gene 386:1–10.

Crispino M, Stone DJ, Wei M, Anderson CP, Tocco G, Finch CE, Baudry M. 1999. Variations of synaptotagmin I, synaptotagmin IV, and synaptophysin mRNA levels in rat hippocampus during the estrous cycle. Exp Neurol 159:574–583.

de Lang A, Baas T, Teal T, Leijten LM, Rain B, Osterhaus AD, Haagmans BL, Katze MG. 2007. Functional genomics highlights differential induction of antiviral pathways in the lungs of SARS-CoV-infected macaques. PLoS Pathog 3:e112.

Ding Y, He L, Zhang Q, Huang Z, Che X, Hou J, Wang H, Shen H, Qiu L, Li Z, Geng J, Cai J, Han H, Li X, Kang W, Weng D, Liang P, Jiang S. 2004. Organ distribution of severe acute respiratory syndrome (SARS) associated coronavirus (SARS-CoV) in SARS patients: Implications for pathogenesis and virus transmission pathways. J Pathol 203:622–630.

Drosten C, Gunther S, Preiser W, van der Werf S, Brodt HR, Becker S, Rabenau H, Panning M, Kolesnikova L, Fouchier RA, Berger A, Burguiere AM, Cinatl J, Eickmann M, Escriou N, Grywna K, Kramme S, Manuguerra JC, Muller S, Rickerts V, Sturmer M, Vieth S, Klenk HD, Osterhaus AD, Schmitz H, Doerr HW. 2003. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med 348:1967–1976.

Enjuanes L, Almazan F, Sola I, Zuniga S. 2006. Biochemical aspects of coronavirus replication and virus-host interaction. Annu Rev Microbiol 60:211–230.

Fackler OT, Krausslich HG. 2006. Interactions of human retroviruses with the host cell cytoskeleton. Curr Opin Microbiol 9:409–415.

Garrison KE, Jones RB, Meiklejohn DA, Anwar N, Ndhlovu LC, Chapman JM, Erickson AL, Agrawal A, Spotts G, Hecht FM, Rakoff-Nahoum S, Lenz J, Ostrowski MA, Nixon DF. 2007. T cell responses to human endogenous retroviruses in HIV-1 infection. PLoS Pathog 3:e165.

Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, Schuler G, Klein SL, Old S, Rasooly R, Good P, Guyer M, Peck AM, Derge JG, Lipman D, Collins FS, Jang W, Sherry S, Feolo M, Misquitta L, Lee E, Rotmistrovsky K, Greenhut SF, Schaefer CF, Buetow K, Bonner TI, Haussler D, Kent J, Kiekhaus M, Furey T, Brent M, Prange C, Schreiber K, Shapiro N, Bhat NK, Hopkins RF, Hsie F, Driscoll T, Soares MB, Casavant TL, Scheetz TE, Brown-stein MJ, Usdin TB, Toshiyuki S, Carninci P, Piao Y, Dudekula DB, Ko MS, Kawakami K, Suzuki Y, Sugano S, Gruber CE, Smith MR, Simmons B, Moore T, Waterman R, Johnson SL, Ruan Y, Wei CL, Mathavan S, Gunaratne PH, Wu J, Garcia AM, Hulyk SW, Fuh E, Yuan Y, Sneed A, Kowis C, Hodgson A, Muzny DM, McPherson J, Gibbs RA, Fahey J, Helton E, Ketteman M, Madan A, Rodrigues S, Sanchez A, Whiting M, Madari A, Young AC, Wetherby KD, Granite SJ, Kwong PN, Brinkley CP, Pearson RL, Bouffard GG, Blakesly RW, Green ED, Dickson MC, Rodriguez AC, Grimwood J, Schmutz J, Myers RM, Butterfield YS, Griffith M, Griffith OL, Krzywinski MI, Liao N, Morin R, Palmquist D, Petrescu AS, Skalska U, Smailus DE, Stott JM, Schnerch A, Schein JE, Jones SJ, Holt RA, Baross A, Marra MA, Clifton S, Makowski KA, Bosak S, Malek J. 2004. The status, quality, and expansion of the NIH full-length cDNA project: The Mammalian Gene Collection (MGC). Genome Res 14:2121–2127.

Greber UF, Way M. 2006. A superhighway to virus infection. Cell 124:741–754.

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: MicroRNA sequences, targets and gene nomenclature. Nucleic Acids Res 34:D140–D144.

Gu J, Korteweg C. 2007. Pathology and pathogenesis of severe acute respiratory syndrome. Am J Pathol 170:1136–1147.

Harbers M. 2008. The current status of cDNA cloning. Genomics 91: 232–242.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo

MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258–D261.

Hillier LD, Lennon G, Becker M, Bonaldo MF, Chiapelli B, Chissoe S, Dietrich N, DuBuque T, Favello A, Gish W, Hawkins M, Hultman M, Kucaba T, Lacy M, Le M LeN, Mardis E, Moore B, Morris M, Parsons J, Prange C, Rifkin L, Rohlfing T, Schellenberg K, Bento Soares M, Tan F, Thierry-Meg J, Trevaskis E, Underwood K, Wohldman P, Waterston R, Wilson R, Marra M. 1996. Generation and analysis of 280,000 human expressed sequence tags. Genome Res 6:807–828.

Huang X, Madan A. 1999. CAP3: A DNA sequence assembly program. Genome Res 9:868–877.

Huang J, Wang F, Argyris E, Chen K, Liang Z, Tian H, Huang W, Squires K, Verlinghieri G, Zhang H. 2007. Cellular microRNAs contribute to HIV-1 latency in resting primary CD4+ T lymphocytes. Nat Med 13:1241–1247.

Hwang DM, Chamberlain DW, Poutanen SM, Low DE, Asa SL, Butany J. 2005. Pulmonary pathology of severe acute respiratory syndrome in Toronto. Mod Pathol 18:1–10.

Jelks KB, Wylie R, Floyd CL, McAllister AK, Wise P. 2007. Estradiol targets synaptic proteins to induce glutamatergic synapse formation in cultured hippocampal neurons: Critical role of estrogen receptor-alpha. J Neurosci 27:6903–6913.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, Bell I, Cheung E, Drenkow J, Dumais E, Patel S, Helt G, Ganesh M, Ghosh S, Piccolboni A, Sementchenko V, Tammana H, Gingeras TR. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316:1484–1488.

Katsumata K, Ikeda H, Sato M, Ishizu A, Kawarada Y, Kato H, Wakisaka A, Koike T, Yoshiki T. 1999. Cytokine regulation of env gene expression of human endogenous retrovirus-R in human vascular endothelial cells. Clin Immunol 93:75–80.

Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. 2004. The International Protein Index: An integrated database for proteomics experiments. Proteomics 4:1985–1988.

Kiyosawa H, Mise N, Iwase S, Hayashizaki Y, Abe K. 2005. Disclosing hidden transcripts: Mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. Genome Res 15:463–474.

Kretz O, Fester L, Wehrenberg U, Zhou L, Brauckmann S, Zhao S, Prange-Kiel J, Naumann T, Jarry H, Frotscher M, Rune GM. 2004. Hippocampal synapses depend on hippocampal estrogen synthesis. J Neurosci 24:5913–5921.

Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, Emery S, Tong S, Urbani C, Comer JA, Lim W, Rollin PE, Dowell SF, Ling AE, Humphrey CD, Shieh WJ, Guarner J, Paddock CD, Rota P, Fields B, DeRisi J, Yang JY, Cox N, Hughes JM, LeDuc JW, Bellini WJ, Anderson LJ. 2003. A novel coronavirus associated with severe acute respiratory syndrome. N Engl J Med 348:1953–1966.

Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hinrichs AS, Harte RA, Giardine B, Fujita P, Diekhans M, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ. 2009. The UCSC Genome Browser Database: Update 2009. Nucleic Acids Res 37:D755–D761.

Lecellier CH, Dunoyer P, Arar K, Lehmann-Che J, Eyquem S, Himber C, Saib A, Voinnet O. 2005. A cellular microRNA mediates antiviral defense in human cells. Science 308:557–560.

Leu JH, Chang CC, Wu JL, Hsu CW, Hirono I, Aoki T, Juan HF, Lo CF, Kou GH, Huang HC. 2007. Comparative analysis of differentially expressed genes in normal and white spot syndrome virus infected Penaeus monodon. BMC Genomics 8:120.

Lin WD, Chen YC, Ho JM, Hsiao CD. 2006. GOBU: Toward an integration interface for biological objects. J Inf Sci Eng 22:19–129.

Mattick JS, Makunin IV. 2006. Non-coding RNA. Hum Mol Genet 15:R17–R29.

Medstrand P, van de Lagemaat LN, Mager DL. 2002. Retroelement distributions in the human genome: Variations associated with age and proximity to genes. Genome Res 12:1483–1495.

Ng LF, Hibberd ML, Ooi EE, Tang KF, Neo SY, Tan J, Murthy KR, Vega VB, Chia JM, Liu ET, Ren EC. 2004. A human in vitro model system for investigating genome-wide host responses to SARS coronavirus infection. BMC Infect Dis 4:34.

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

Nicholls JM, Poon LL, Lee KC, Ng WF, Lai ST, Leung CY, Chu CM, Hui PK, Mak KL, Lim W, Yan KW, Chan KH, Tsang NC, Guan Y, Yuen KY, Peiris JS. 2003. Lung pathology of fatal severe acute respiratory syndrome. Lancet 361:1773–1778.

Nicholls JM, Butany J, Poon LL, Chan KH, Beh SL, Poutanen S, Peiris JS, Wong M. 2006. Time course and cellular localization of SARS-CoV nucleoprotein and RNA in lungs from fatal cases of SARS. PLoS Med 3:e27.

Ono M, Kawakami M, Ushikubo H. 1987. Stimulation of expression of the human endogenous retrovirus genome by female steroid hormones in human breast cancer cell line T47D. J Virol 61:2059–2062.

Pang KC, Stephen S, Engstrom PG, Tajul-Arifin K, Chen W, Wahlestedt C, Lenhard B, Hayashizaki Y, Mattick JS. 2005. RNAdb—A comprehensive mammalian noncoding RNA database. Nucleic Acids Res 33:D125–D130.

Pedersen IM, Cheng G, Wieland S, Volinia S, Croce CM, Chisari FV, David M. 2007. Interferon modulation of cellular microRNAs as an antiviral mechanism. Nature 449:919–922.

Pelkmans L. 2005. Viruses as probes for systems analysis of cellular signalling, cytoskeleton reorganization and endocytosis. Curr Opin Microbiol 8:331–337.

Pelkmans L, Fava E, Grabner H, Hannus M, Habermann B, Krausz E, Zerial M. 2005. Genome-wide analysis of human kinases in clathrin- and caveolae/raft-mediated endocytosis. Nature 436:78–86.

Philip-Couderc P, Pathak A, Smih F, Dambrin C, Harmancey R, Buys S, Galinier M, Massabuau P, Roncalli J, Senard JM, Rouet P. 2004. Uncomplicated human obesity is associated with a specific cardiac transcriptome: Involvement of the Wnt pathway. FASEB J 18: 1539–1540.

Prabakaran S, Swatton JE, Ryan MM, Huffaker SJ, Huang JT, Griffin JL, Wayland M, Freeman T, Dudbridge F, Lilley KS, Karp NA, Hester S, Tkachev D, Mimmack ML, Yolken RH, Webster MJ, Torrey EF, Bahn S. 2004. Mitochondrial dysfunction in schizophrenia: Evidence for compromised brain metabolism and oxidative stress. Mol Psychiatry 9:684–697.

Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:D61–D65.

Radtke K, Dohner K, Sodeik B. 2006. Viral interactions with the cytoskeleton: A hitchhiker's guide to the cell. Cell Microbiol 8: 387–400.

Rozowsky J, Wu J, Lian Z, Nagalakshmi U, Korbel JO, Kapranov P, Zheng D, Dyke S, Newburger P, Miller P, Gingeras TR, Weissman S, Gerstein M, Snyder M. 2006. Novel transcribed regions in the human genome. Cold Spring Harb Symp Quant Biol 71:111–116.

Rune GM, Wehrenberg U, Prange-Kiel J, Zhou L, Adelmann G, Frotscher M. 2002. Estrogen up-regulates estrogen receptor alpha and synaptophysin in slice cultures of rat hippocampus. Neuroscience 113:167–175.

Schneider PM, Witzel-Schlomp K, Rittner C, Zhang L. 2001. The endogenous retroviral insertion in the human complement C4 gene modulates the expression of homologous genes by antisense inhibition. Immunogenetics 53:1–9.

Strauss JH, Strauss EG. 2002. Minus-strand RNA viruses. Viruses and human disease. San Diego: Academic Press. p 151–153.

Suzuki Y, Yamashita R, Shirota M, Sakakibara Y, Chiba J, Mizushima-Sugano J, Nakai K, Sugano S. 2004. Sequence comparison of human and mouse genes reveals a homologous block structure in the promoter regions. Genome Res 14:1711–1718.

Tang BS, Chan KH, Cheng VC, Woo PC, Lau SK, Lam CC, Chan TL, Wu AK, Hung IF, Leung SY, Yuen KY. 2005. Comparative host gene transcription by microarray analysis early after infection of the Huh7 cell line by severe acute respiratory syndrome coronavirus and human coronavirus 229E. J Virol 79:6180–6193.

Taruscio D, Mantovani A. 2004. Factors regulating endogenous retroviral sequences in human and mouse. Cytogenet Genome Res 105: 351–362.

Tseng CT, Perrone LA, Zhu H, Makino S, Peters CJ. 2005. Severe acute respiratory syndrome and the innate immune responses: Modulation of effector cell function without productive infection. J Immunol 174:7977–7985.

Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, DeRisi JL. 2002. Microarray-based detection and genotyping of viral pathogens. Proc Natl Acad Sci USA 99:15687–15692.

Wang H, Yang P, Liu K, Guo F, Zhang Y, Zhang G, Jiang C. 2008. SARS coronavirus entry into host cells through a novel clathrin- and caveolae-independent endocytic pathway. Cell Res 18:290–301.

Wang Z, Gerstein M, Snyder M. 2009. RNA-Seq: A revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63.

Wu TD, Watanabe CK. 2005. GMAP: A genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21:1859–1875.

Wu JQ, Du J, Rozowsky J, Zhang Z, Urban AE, Euskirchen G, Weissman S, Gerstein M, Snyder M. 2008. Systematic analysis of transcribed loci in ENCODE regions using RACE sequencing reveals extensive transcription in the human genome. Genome Biol 9:R3.

Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, Wang J, Li S, Li R, Bolund L. 2006. WEGO: A web tool for plotting GO annotations. Nucleic Acids Res 34:W293–W297.

Yu SY, Hu YW, Liu XY, Xiong W, Zhou ZT, Yuan ZH. 2005. Gene expression profiles in peripheral blood mononuclear cells of SARS patients. World J Gastroenterol 11:5037–5043.

Zweig AS, Karolchik D, Kuhn RM, Haussler D, Kent WJ. 2008. UCSC genome browser tutorial. Genomics 92:75–84.