T, Lenhard B, Eveno E et al.: Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol* 2004, **2**:e162.

30. Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pedersen JS, Hsu F, Hinrichs AS, Harte RA, Diekhans M, Clawson H, Bejerano G, Barber GP, Baertsch R, Haussler D, Kent WJ: The UCSC genome browser database: update 2007. *Nucleic Acids Res* 2007, **35**:D668-673.

31. Li WH: Molecular Evolution. *Sinauer Associates, Sunderland,* 1997, MA.

32. Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S,.Gingeras TR: Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res* 2005, **15**:987-997.

33. Liu WM, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho MH, Baid J, Smeekens SP: Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* 2002, **18**:1593-1599.

34. Bertone P, Stolc V, Royce TE, Rozowsky JS, Urban AE, Zhu X, Rinn JL, Tongprasit W, Samanta M, Weissman S, Gerstein M, Snyder M: Global identification of human transcribed sequences with genome tiling arrays. *Science* 2004, **306**:2242-2246.

35. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S,

Drenkow J, Piccolboni A, Bekiranov S, Helt G, Tammana H, Gingeras TR: Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res* 2004, **14**:331-342.

36. Khaitovich P, Kelso J, Franz H, Visagie J, Giger T, Joerchel S, Petzold E, Green RE, Lachmann M, Paabo S: Functionality of intergenic transcription: an evolutionary comparison. *PLoS Genet* 2006, **2**:e171.

37. Kondrashov FA, Koonin EV: Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *Trends Genet* 2003, **19**:115-119.

38. Kimura K, Wakamatsu A, Suzuki Y, Ota T, Nishikawa T, Yamashita R, Yamamoto J, Sekine M, Tsuritani K, Wakaguri H, Ishii S, Sugiyama T, Saito K, Isono Y, Irie R, Kushida N, Yoneyama T, Otsuka R, Kanda K, Yokoi T, Kondo H, Wagatsuma M, Murakawa K, Ishida S, Ishibashi T, Takahashi Fujii A, Tanase T, Nagai K, Kikuchi H, Nakai K et al.: Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res* 2006, **16**:55-65.

39. Xing Y, Lee C: Alternative splicing and RNA selection pressure--evolutionary consequences for eukaryotic genomes. *Nat Rev Genet* 2006, **7**:499-509.

40. Zhang J, Nielsen R, Yang Z: Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* 2005, **22**:2472-2479.

41. Bakewell MA, Shi P, Zhang J: More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc Natl Acad Sci U S A* 2007, **104**:7489-7494.

42. Stewart CB, Disotell TR: Primate evolution—in and out of Africa. *Curr Biol* 1998, **8**:R582-588.

43. Kaessmann H, Wiebe V, Paabo S: Extensive nuclear DNA sequence diversity among chimpanzees. *Science* 1999, **286**:1159-1162.

44. Yi S, Ellsworth DL, Li WH: Slow molecular clocks in Old World monkeys, apes, and humans. *Mol Biol Evol* 2002, **19**:2191-2198.

45. Kondrashov AS: Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat* 2003, **21**:12-27.

46. Chen WH, Wang XX, Lin W, He XW, Wu ZQ, Lin Y, Hu SN, Wang XN: Analysis of 10,000 ESTs from lymphocytes of the cynomolgus monkey to improve our understanding of its immune system. *BMC Genomics* 2006, **7**:82.

47. Wallace JC, Korth MJ, Paeper B, Proll SC, Thomas MJ, Magness CL, Iadonato SP, Nelson C, Katze MG: High-density rhesus macaque oligonucleotide microarray design

using early-stage rhesus genome sequence information and human genome annotations. *BMC Genomics* 2007, **8**:28.

48. Suzuki Y, Sugano S: Construction of a full-length enriched and a 5'-end enriched cDNA library using the oligo-capping method. *Methods Mol Biol* 2003, **221**:73-91.

49. Jurka J: Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* 2000, **16**:418-420.

50. Pruitt KD, Tatusova T, Maglott DR: NCBI Reference Sequence project: update and current status. *Nucleic Acids Res* 2003, **31**:34-37.

51. Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994, **22**:4673-4680.

52. Li WH: Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol* 1993, **36**:96-99.

53. Pamilo P, Bianchi NO: Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Mol Biol Evol* 1993, **10**:271-281.

54. Yang Z: PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 1997, **13**:555-556.

55. Kimura M: A simple method for estimating evolutionary rates of base substitutions

through comparative studies of nucleotide sequences. *J Mol Evol* 1980, **16**:111-120.

Figure Legends

Figure 1

Classification of the 1245 Non-RefSeq transcripts. Transcripts shorter than 300 bp after masking the repetitive sequences were categorized as junk sequences. The remaining sequences were BLAST-searched against all public human cDNA sequences for the forward strand. Homologous sequences to the unannotated human cDNAs were classified as orphan transcripts for the forward strand and anti-transcript for the reverse strand. The remaining 947 clones were mapped on the human genome sequence and arranged according to the annotation from the UCSC genome browser (hg18). The transcripts that overlapped with the genic regions including UTR were classified as intronic transcripts, and the transcripts that were mapped more than 5 kb away from the genic region were classified as intergenic transcripts.

Figure 2

The proportion of the expressed transcripts in the RefSeq homologs (control) and unidentified transcripts. Cerebrum, cerebellum, liver, and testis of a male macaque were used for the microarray experiments with duplicated hybridizations. The transcripts were classified into no expression (blue), expressed in 1–3 tissues (grey), or expressed in all tissues (red).

Figure 3

Sequence conservation of the brain-expressed and testis-expressed transcripts between humans and macaques. For the RefSeq homologs (control), the non-synonymous (K_a) and synonymous (K_s) substitution rates were estimated using the Li-Pamilo-Bianchi method [48]. The substitution rates in the intergenic and intronic transcripts were estimated using Kimura's two parameter methods [55]. The heights of the boxes represent the lower and upper quartile points, and the whiskers show the minimum and maximum points.

Figure 4

Distribution of transcript expression levels of the RefSeq homologs (blue) and the intergenic transcripts (red). Only the transcripts that were determined as significantly expressed on the microarray are presented in the figure. Log-transformed signal intensity in the tissue with the highest expression was shown. The intergenic transcripts showed significantly lower expression levels than the RefSeq homologs.

Figure 5

RT-PCR gel images for the expression of the intergenic transcripts in the human (H) and the macaque (Q) brain. Transcript names indicate whether the expression was detected by the microarray experiments (red) or not (blue). Expected PCR products are marked by the white arrows.

Figure 6

Pattern of the unidentified exons. The closed boxes represent exons in the genomes. Unidentified exons in macaques are presented as blue boxes. Intergenic regions and introns are depicted by thick and thin horizontal lines, respectively. (A) extended exons. (B) novel exons. These exons were further classified into internal (right panel) and external (left panel) exons. The number of genes in each category is shown on the left of each schema. The number of unidentified exons that have not been found even in the EST sequences is shown in parentheses.

Figure 7

Genealogical relationship (phylogeny of genes) among the humans (H), cynomolgus macaque (C), and rhesus macaque (R). The common ancestor of the two macaques is indicated by the letter O. The time of speciation between the two macaques is shown by the dashed line. Note that the tree is unrooted.

Table 1

Summary of cDNA clones

Library	# of isolated clones	# of full-sequenced clones
Brain: Parietal Lobe (QnpA)	8063 (5890)	649 (336)
Brain: Frontal Lobe (QflA)	13,215 (9286)	2493 (1768)
Brain: Temporal Lobe (QtrA)	6797 (6039)	1078 (862)
Brain: Occipital Lobe (QorA)	5458 (4518)	634 (606)
Brain Stem (QbsA, B)	2776 (1993)	359 (301)
Brain: Medulla Oblongata (QmoA)	4485 (3645)	1146 (912)
Brain: Cerebellar Cortex (QccE)	11,734 (9028)	731 (563)
Testis (QtsA)	10,867 (8510)	2316 (2175)
Liver (Qlv)	22,326 (20,833)	0 (0)
Total	85,721 (69,742)	9407 (7523)
Averaged Length		1882 bp

*Numbers of genes with the RefSeq homologs are shown in parentheses.

Table 2

Number of expressed transcripts in the unknown macaque transcripts

	Unidentified transcripts[a]	**Intergenic transcripts**[b]
Cerebrum	321	54
Cerebellum	417	58
Liver	139	13
Testis	241	52
All tissues	74	10
Any tissue	544	137
Total	1024	231

[a] Transcripts that have no homology to the public human cDNA sequences.

[b] Transcripts that were mapped more than 5 kb away from the annotated genic regions on the human genome (see Fig. 1).

Table 3. Number of genes under positive selection out of 1499 non-duplicated genes determined using the branch-site test of positive selection

Lineage[a]	$P \leq 0.05$	$P \leq 0.01$
H-O	39	14
C-O	15	10
R-O	22	21
Between the macaques (C-O + R-O)	37	32
All lineages (H-O + C-O + R-O)	74	33

[a]H: human; C: cynomolgus macaque; R: rhesus macaque; O: cynomolgus-rhesus ancestor (see Fig. 7).

Table 4. Divergence among the human, cynomolgus, and rhesus genes

Model without ancestral polymorphisms (Raw data)

Lineage[a]	K_a (± S.E.)	K_s (± S.E.)
H-O	1.06×10^{-2} (3.20×10^{-4})	6.82×10^{-2} (1.07×10^{-3})
C-O	1.02×10^{-3} (4.79×10^{-5})	3.04×10^{-3} (1.20×10^{-4})
R-O	4.98×10^{-4} (3.36×10^{-5})	2.50×10^{-3} (1.15×10^{-4})

Model with ancestral polymorphisms

	$2tu$[b] (±S.E.)	$4N_eu$[b] (±S.E.)
Raw data	2.13×10^{-3} (2.24×10^{-4})	3.27×10^{-3} (2.52×10^{-4})
PCR error corrected	1.81×10^{-3} (2.12×10^{-4})	3.11×10^{-3} (2.40×10^{-4})

[a] H: human, C: cynomolgus macaque, R: rhesus macaque, O: cynomolgus-rhesus ancestor (see Fig. 7).

[b] t: time after speciation; u: mutation rate; N_e: effective population size of the cynomolgus-rhesus ancestor.

Additional data files

Additional file 1

File format: XLS

Title: Expression of novel macaque transcripts

Description: Significance of gene expression in the *M. fascicularis* oligonucleotide microarray analysis is shown.

Additional file 2

File format: XLS

Title: Unidentified UTR regions in the macaque cDNAs.

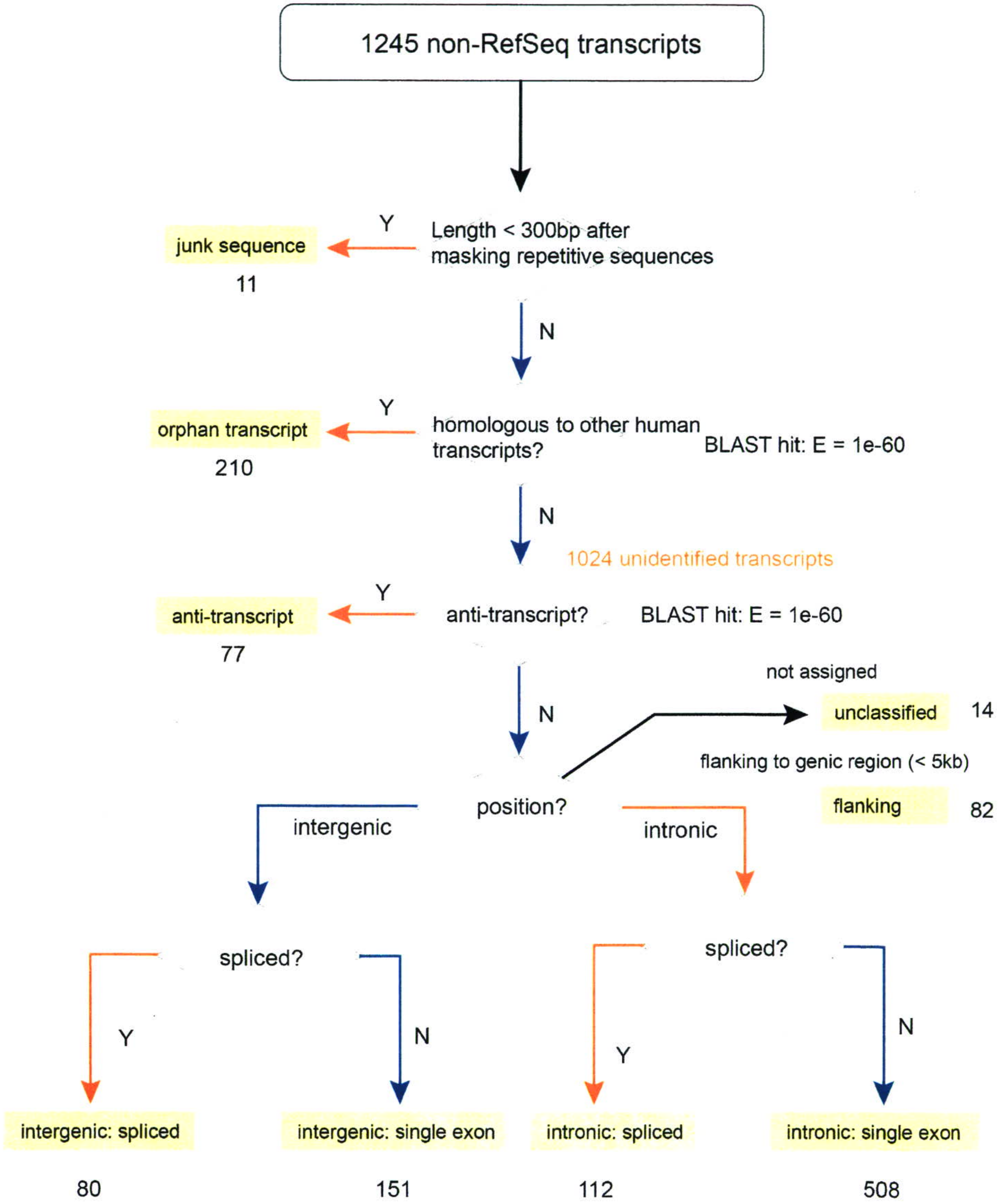
Description: The regions of macaque cDNAs that did not show homology to human cDNAs are listed.

Additional file 3

File format: DOC

Title: Divergence among the human, cynomolgus, and rhesus genes (dataset I: without a duplication filtering).

Description: Estimation of gene divergence using 2655 human-rhesus-cynomolgus alignment is shown.


1245 non-RefSeq transcripts
Y
Length < 300bp after masking repetitive sequences
junk sequence
11
N
Y
homologous to other human transcripts?
BLAST hit: E = 1e-60
orphan transcript
210
N
1024 unidentified transcripts
Y
anti-transcript?
BLAST hit: E = 1e-60
anti-transcript
77
N
not assigned
unclassified
14
flanking to genic region (< 5kb)
flanking
82
position?
intergenic
intronic
spliced?
Y
N
intergenic: spliced
80
intergenic: single exon
151
spliced?
Y
N
intronic: spliced
112
intronic: single exon
508

Additional file 4

File format: XLS

Title: LRT (likelihood ratio test) statistics for the test of positive selection

Description: Candidate genes under positive selection using branch-site test of positive selection are given with log-likelihood ratio.

Additional file 5

File format: XLS

Title: Specific probes for known genes in the *M.fascicularis* microarray

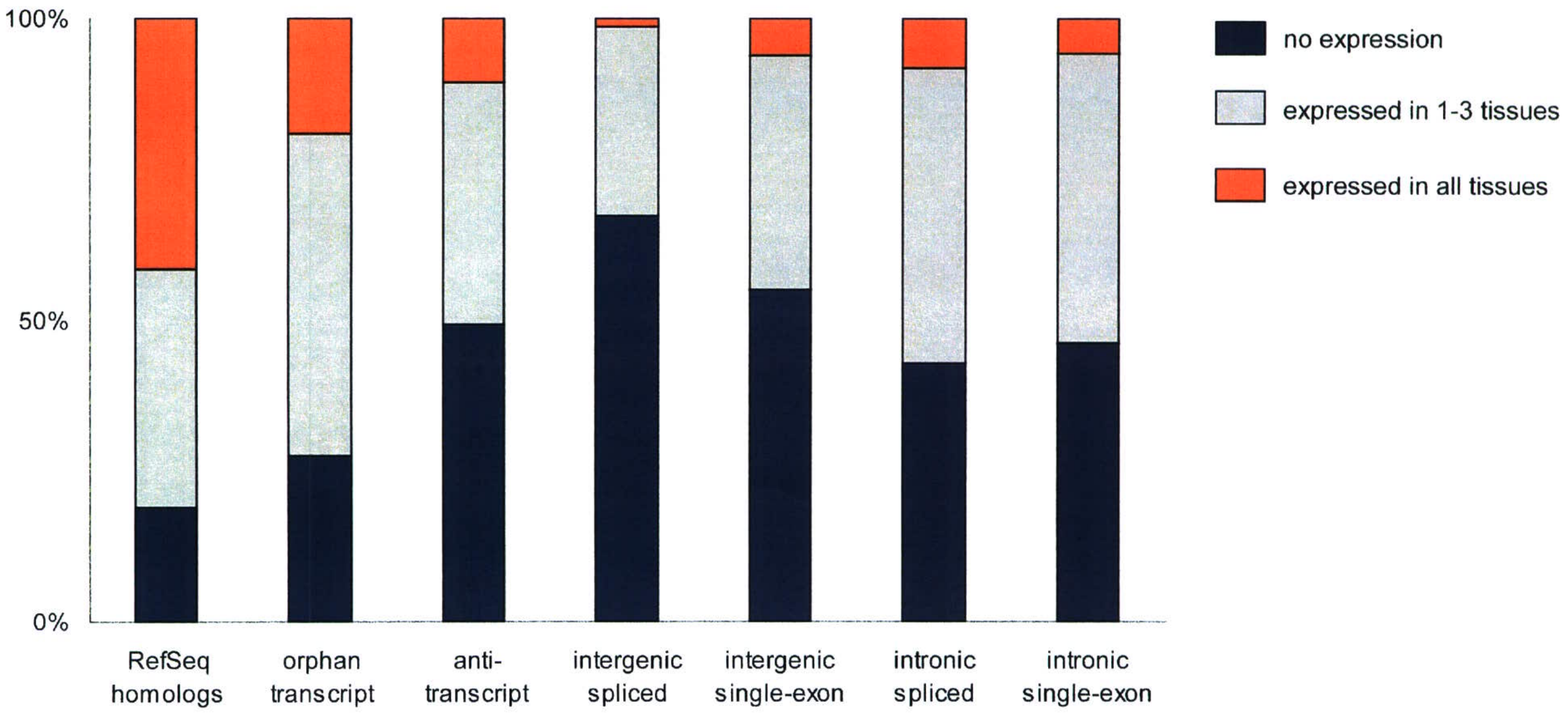
Description: These genes are not found in the previously published macaque oligonucleotide microarray.

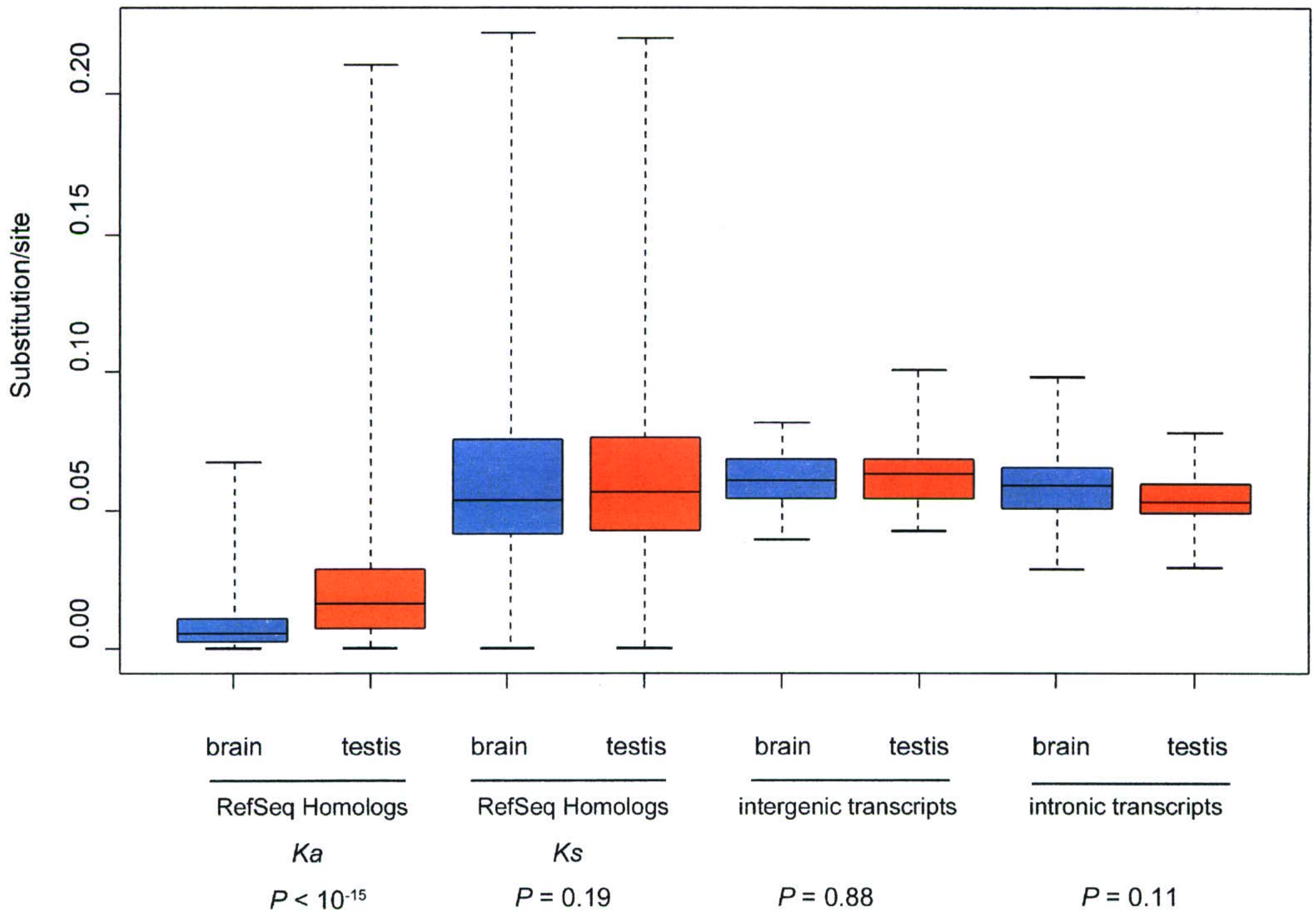
Additional file 6

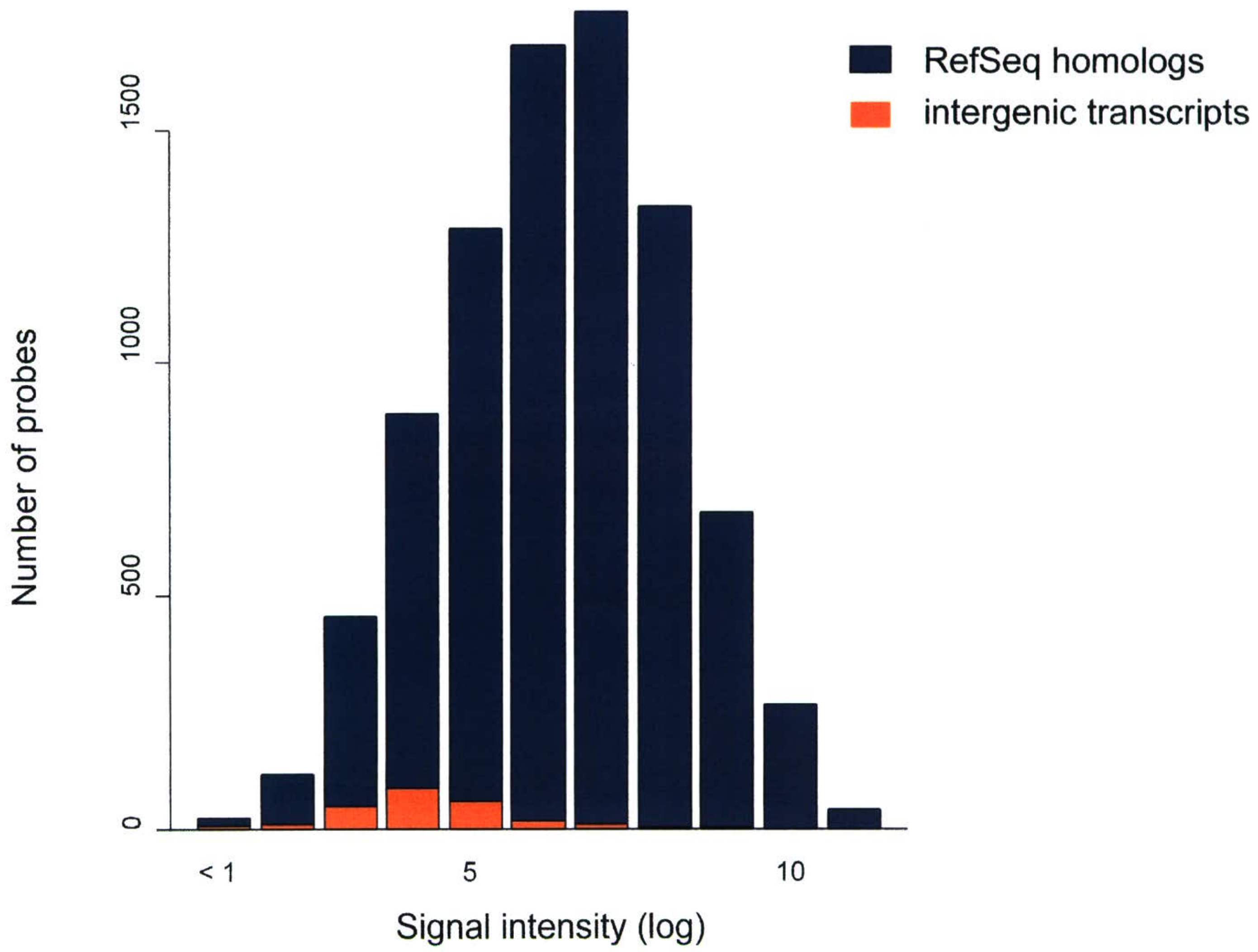
File format: XLS

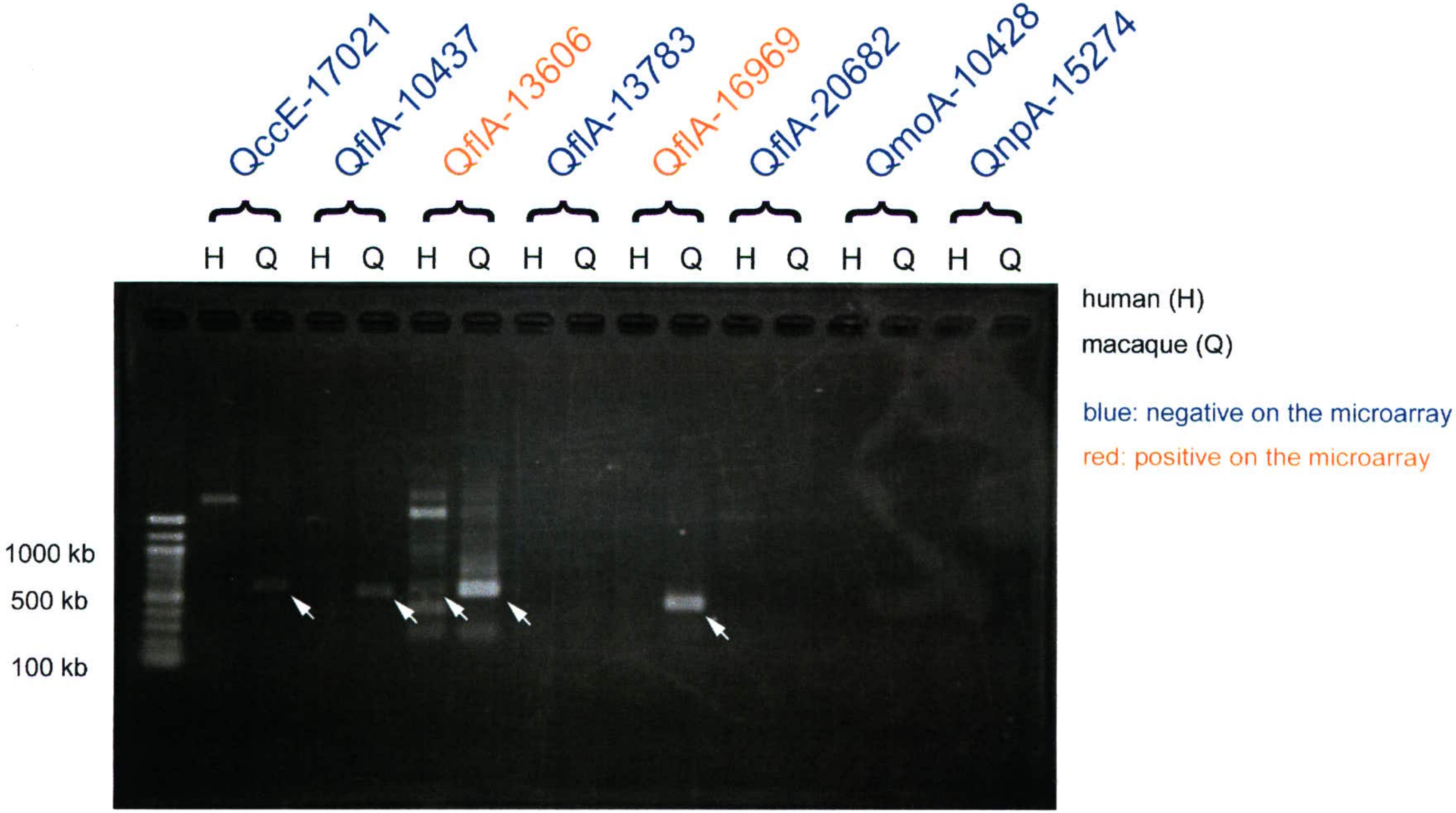
Title: Primer sequences for RT-PCR.

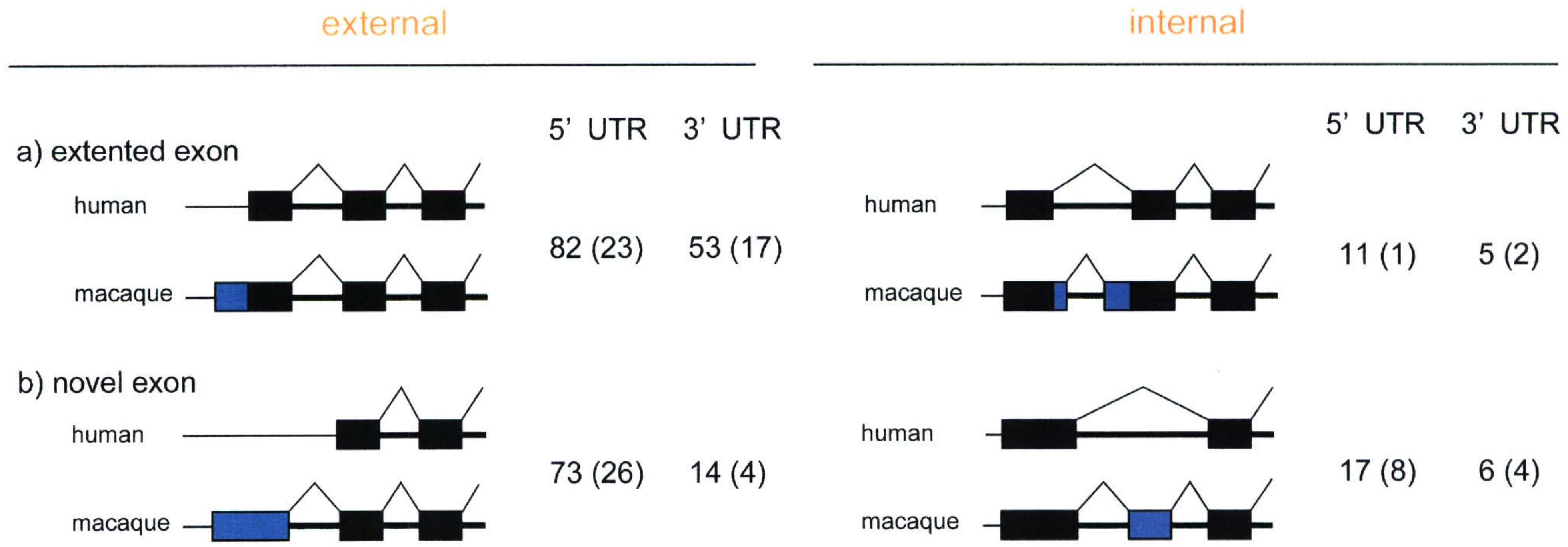
Description: The list shows the primer sequences that were used for RT-PCR.


100%
50%
0%
RefSeq homologs
orphan transcript
anti-transcript
intergenic spliced
intergenic single-exon
intronic spliced
intronic single-exon
no expression
expressed in 1-3 tissues
expressed in all tissues


Substitution/site
0.00
0.05
0.10
0.15
0.20
brain
testis
RefSeq Homologs
Ka
$P < 10^{-15}$
brain
testis
RefSeq Homologs
Ks
$P = 0.19$
brain
testis
intergenic transcripts
$P = 0.88$
brain
testis
intronic transcripts
$P = 0.11$


RefSeq homologs
intergenic transcripts
Number of probes
0
500
1000
1500
< 1
5
10
Signal intensity (log)


QccE-17021
QflA-10437
QflA-13606
QflA-13783
QflA-16969
QflA-20682
QmoA-10428
QnpA-15274
H Q H Q H Q H Q H Q H Q H Q H Q
human (H)
macaque (Q)
blue: negative on the microarray
red: positive on the microarray
1000 kb
500 kb
100 kb


external
internal
a) extented exon
b) novel exon
human
macaque
human
macaque
5' UTR
3' UTR
82 (23)
53 (17)
73 (26)
14 (4)
human
macaque
human
macaque
5' UTR
3' UTR
11 (1)
5 (2)
17 (8)
6 (4)