Supplementary Information 1: a list of regular expression used in our “uninformative rule” filter.

```
HYPOTHETICAL PROTEIN
HYPOTHETICAL PROTEIN ¥S+
HYPOTHETICAL PROTEIN KIAA¥d+
HYPOTHETICAL PROTEIN FLJ¥d+
HYPOTHETICAL PROTEIN CGI¥-¥d+
HYPOTHETICAL PROTEIN DKFZP¥S+
HYPOTHETICAL PROTEIN, MGC:¥d+
HYPOTHETICAL GENE ¥S+
HA¥d{4}
HYPOTHETICAL [¥.¥d]+¥s*KDA PROTEIN
HYPOTHETICAL [¥.¥d]+¥s*KDA PROTEIN PRECURSOR
DKFZP¥S+ PROTEIN
UNKNOWN
PROTEIN FOR MGC:¥d+
PROTEIN FOR IMAGE:¥d+
R¥d{5}¥_¥d
R¥d{5}¥_¥d, PARTIAL PROTEIN
R¥d{5}¥_¥d, PARTIAL CDS
PRO¥d{4}
PRO¥d{4} PROTEIN
ORF ¥d
PUTATIVE KIAA¥d+ HOMOLOGUE
KIAA¥d+ GENE PRODUCT
KIAA¥d+
KIAA¥d+ PROTEIN
HSPC¥d+
HSPC¥d+ PROTEIN
CHROMOSOME ¥d+ OPEN READING FRAME ¥d+
C¥d+ORF¥d+
FLJ¥d+ PROTEIN
DJ¥d+[A-Z]¥d+(¥.¥d+)*
DJ¥d+[A-Z]¥d+(¥.¥d+)* PROTEIN
NOVEL PROTEIN
```

PUTATIVE NOVEL PROTEIN
CG¥d+ PROTEIN
CG¥d+ GENE PRODUCT
CG¥d+
CGI¥-¥d+ PROTEIN
CGI¥-¥d+
CDNA:? FLJ¥d+ FIS, CLONE ¥w+
BA¥d+[A-Z]¥d+[A-Z]?¥.¥d(¥.¥d)?
RIKEN CDNA .{10} GENE
MRNA, COMPLETE CDS, CLONE:¥d+(¥+¥d[A-Z])?¥-¥d+
MRNA, COMPLETE CDS, CLONE:SMAP¥d+¥-¥w+
BRAIN CDNA, CLONE MNCB-¥d+
.{10}RIK PROTEIN
UNNAMED PROTEIN PRODUCT
MY¥d{3}
MY¥d{3} PROTEIN
BRAIN MY¥d{3}
NPD¥d{3} PROTEIN
[A-Z][0-9]{2}[A-Z0-9]+¥.[0-9]+ PROTEIN
WUGSC:H_¥w+¥.¥w+ PROTEIN
expressed sequence ¥S+
putative
unknown
ORF
open reading frame ¥d+
DNA SEGMENT, CHR [0-9XY]+, WAYNE STATE UNIVERSITY ¥d+, EXPRESSED
DNA SEGMENT, CHR [0-9XY]+, KL MOHLKE ¥d+
DNA SEGMENT, CHR [0-9XY]+, BAYLOR ¥d+
PROTEIN HSPC¥d+
ORF PROTEIN
unknown protein, ¥d+¥-¥d+
HYPOTHETICAL [¥.¥d]+¥s*KDA PROTEIN ¥S+ IN CHROMOSOME ¥S+
EG:[0-9A-Z¥.]+ PROTEIN
GENOMIC DNA, CHROMOSOME ¥d+, P1 CLONE:¥S+
[^,]+, RIKEN FULL-LENGTH ENRICHED LIBRARY, CLONE:.{10}, FULL

INSERT SEQUENCE
ZK¥d+¥.¥d+ PROTEIN
EST ¥w+
B2 ELEMENT
hypothetical protein ¥S+