Zika CRISPR library analysis

Everett JK & Bushman FD

March 16, 2018

Contents

Data processing

8,115,291 filtered reads were acquired from the provided samples which employed a dual barcoding strategy. Two different barcode codes were used to track sample conditions (GTGCGTAA -> Uninfected and CTATTCAA -> Zika infected) while an additional 10 barcodes were used to track sample replicates.

The forward are reverse reads were combined into single combined reads with the PEAR software suite which requires a minimum of 30 NT overlap between paired-end reads. Examination of multiply aligned sequences revealed that the guide sequences could be readily excised by extracting sequences between the constant flanking vector sequences CGAAACACC and GTTTTAGAG. The first flanking sequence was found in 98.34% of sequences, the second flanking sequence was found in 92.59% sequences and both flanking sequences were found in 91.64% sequences. No mismatches were allowed while identifying the flanking sequence positions.

Guide sequence characterization

Guide sequence length distributions

Identified guide sequences show a distribution of lengths centered at 20 nucleotides (Figure 1). The tails of this distribution are shown in Figure 2 where guide sequences with lengths of 16-21 nucleotides were omitted. The 'Unknown' condition refers to reads with barcodes that did not map to one of the three experimental conditions either because of read quality or use of unknown barcodes.

Figure 1

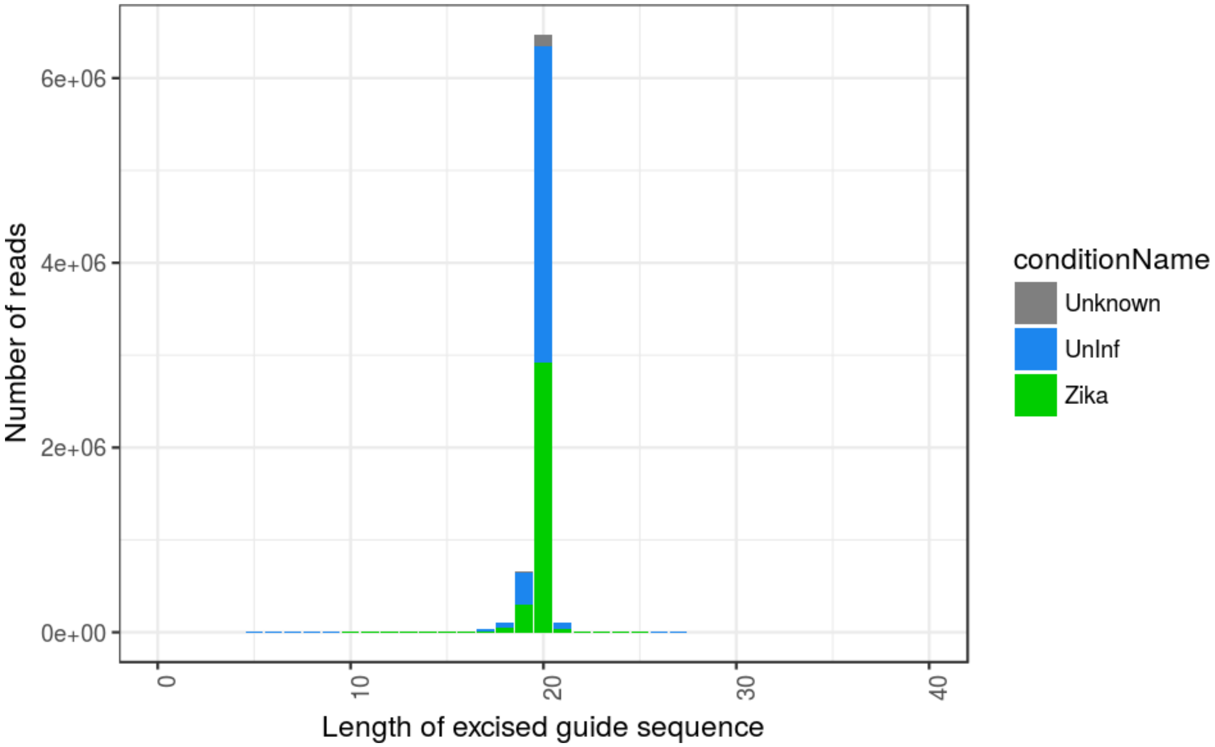
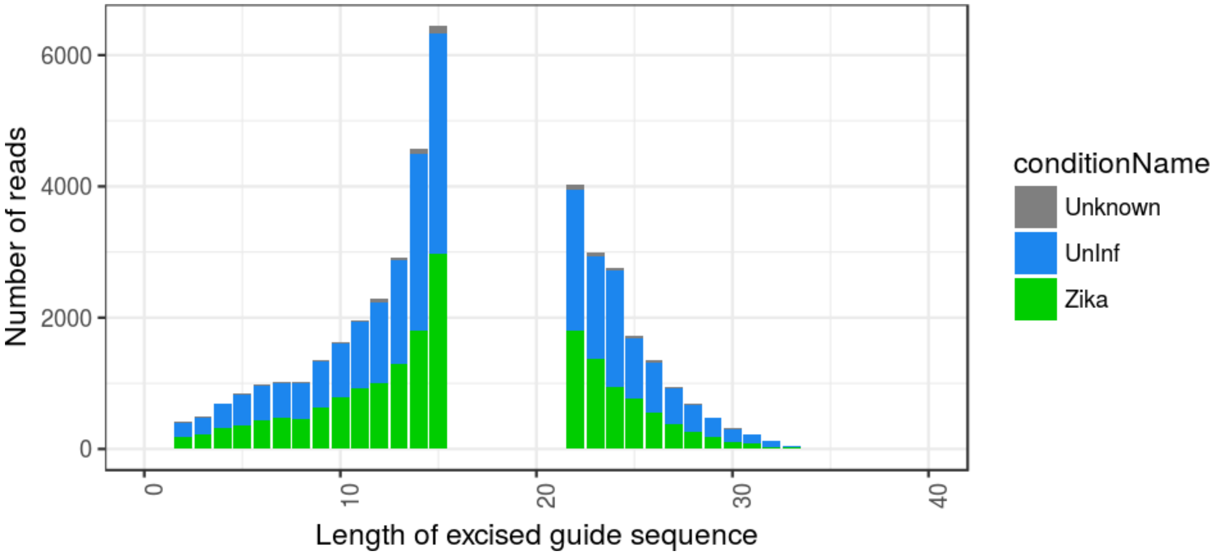


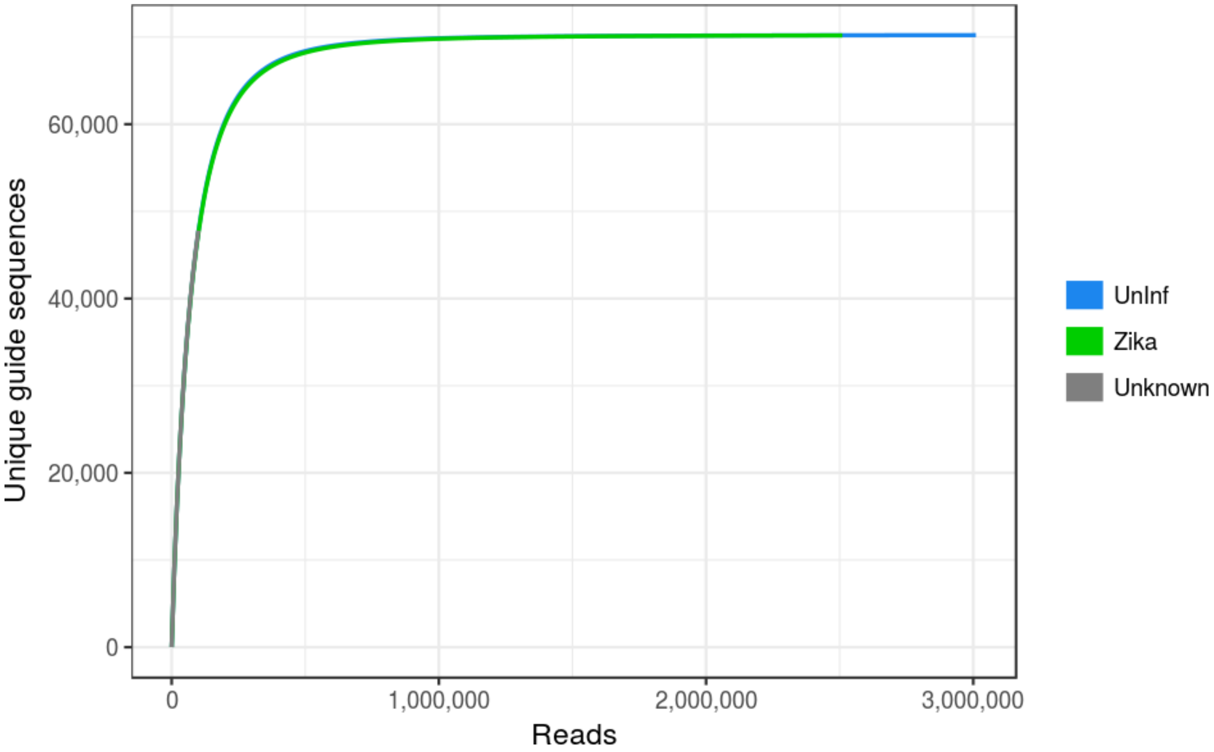
Figure 2



Guide sequence statistics

19.71% of the identified 20 NT guide sequences were found in the library of 70,297 expected guide sequences while 87.02% of the reads mapped to the expected sequences. 99.93% of the expected guide sequences were identified in the sequencing experiments. *Only reads that correspond to the expected guide sequences were considered in the following analysis.* Rarefaction analysis (Figure 3) shows that the identification of unique guide sequences in each condition approaches saturation.

Figure 3



Guide sequence enrichment

Enrichment of specific guide sequences

The number of reads for each unique guide sequence were tallied and Fisher's exact tests (one tailed towards Zika enrichment) were calculated for each guide sequence. The read frequencies of each guide (n reads for guide G in condition C / all reads in condition C) were calculated and are shown as a bivariate plot in Figure 4 revealing that guides targeting Z3HAV1 and FBXW2 appear to be highly enriched in the uninfected condition.

Figure 4

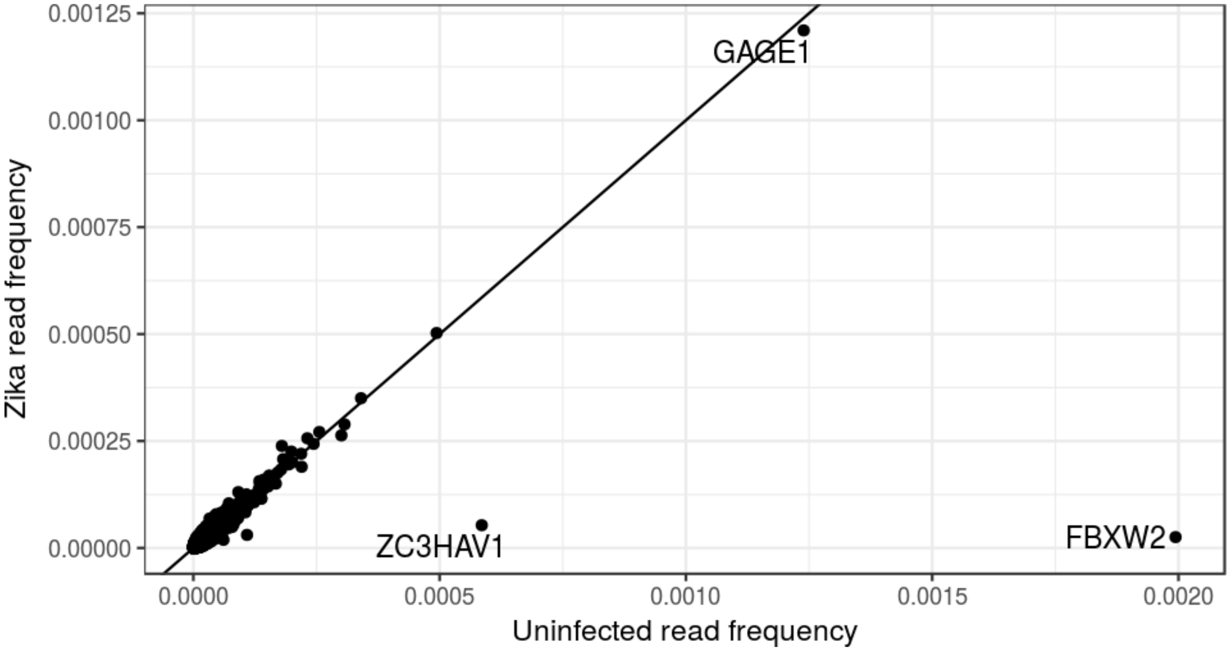


Figure 5 focuses on the lower left corner of Figure 4 and colors the guide sequence data points by their Fisher's exact statistics. Sequences enriched in the Zika infected condition can be grouped by different components of the Fisher's exact test results. From least to most conservative, enriched sequences can be considered to be those with a p-value ≤ 1e-4 (arbitrary value), those with a odds ratio of ≤ 1/2 or those with an odds ratio 95% confidence interval upper limit (OR_95CI_UL) ≤ 1/2. Table 1 shows the number of enriched guides obtained with each cut-off.

Figure 5

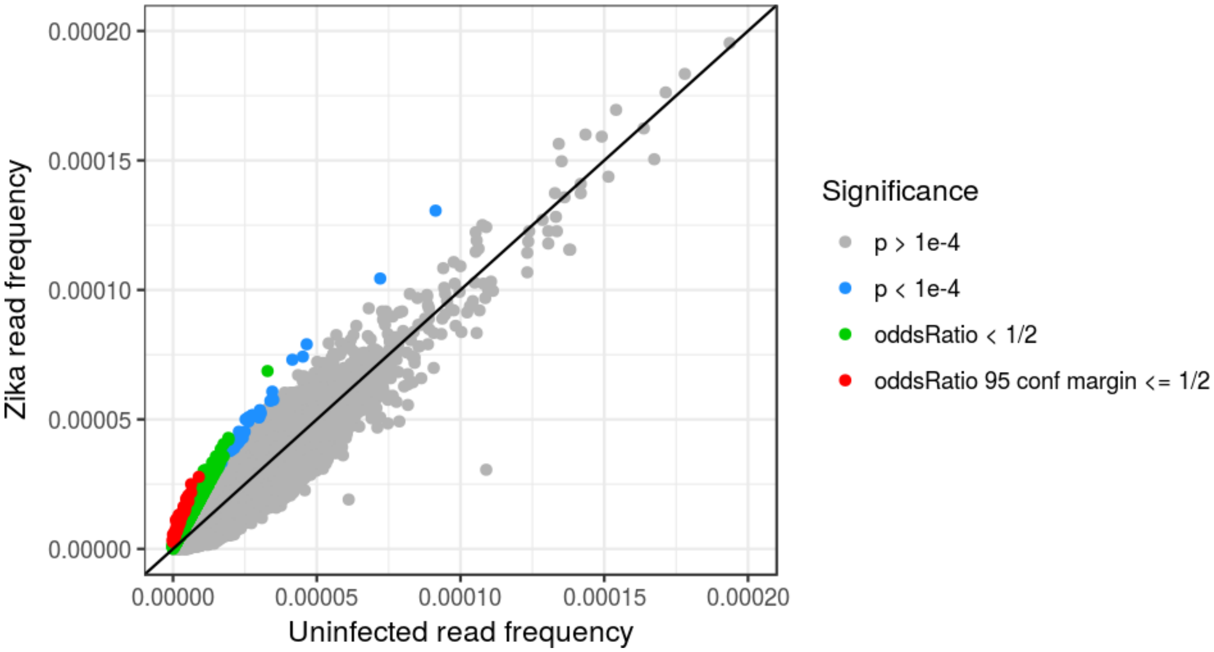


Table 1

	pval ≤ 1e-4	odds ratio ≤ 1/3	odds ratio ≤ 1/4	OR_95CI_UL ≤ 1/2	OR_95CI_UL ≤ 1/3
Guide seqs	181	379	146	62	7

Enrichment of guides targeting specific genes

Alternatively, the enrichment of guides targeting specific genes can be evaluated (Figures 6 & 7) and the number of unique genes enriched can be tallied (Table 2).

Figure 6

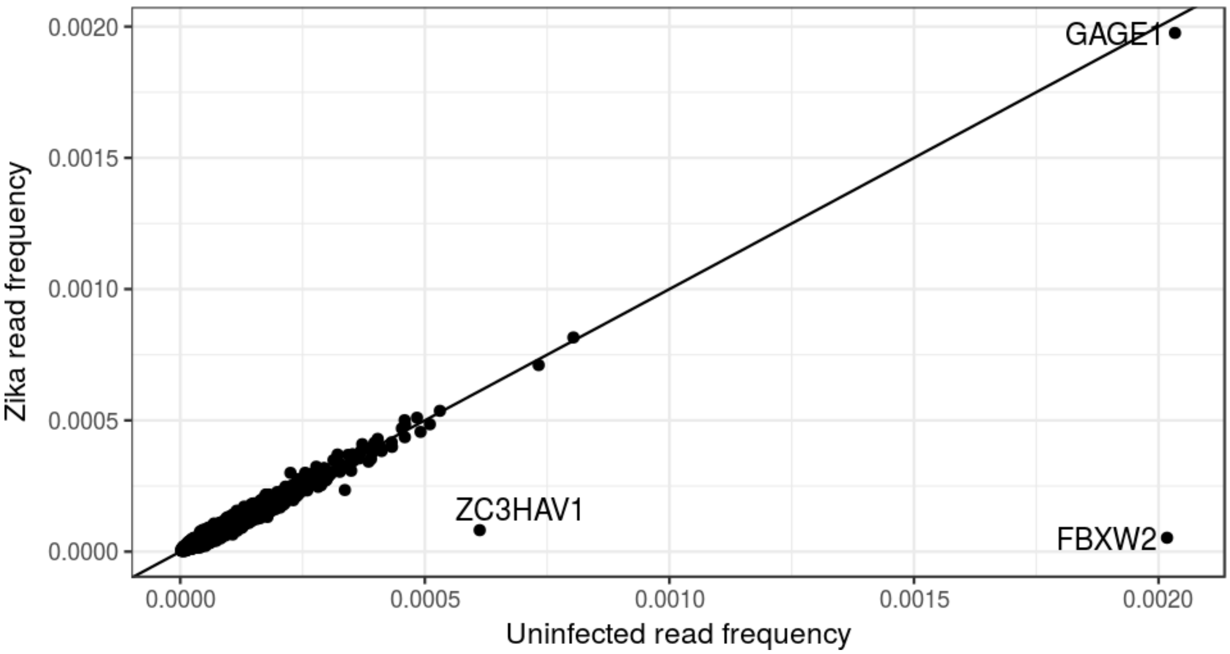


Figure 7

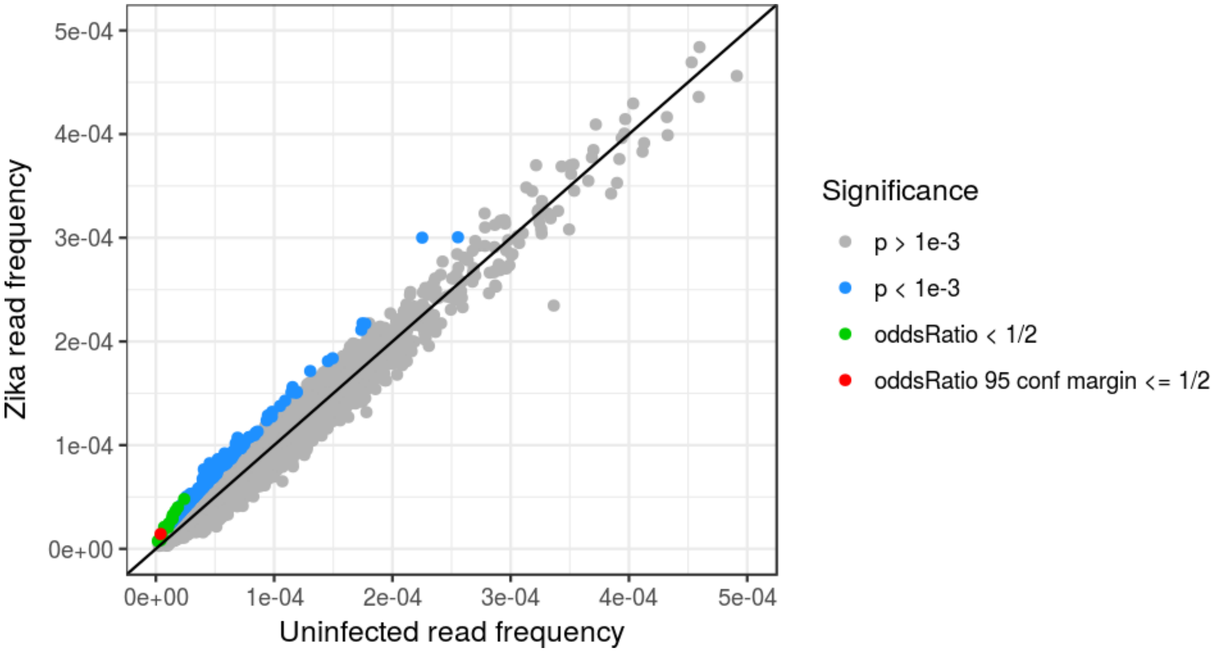


Table 2

	pval ≤ 1e-3	odds ratio ≤ 1/2	odds ratio ≤ 1/3	OR_95CI_UL ≤ 1/2
Targeted genes	214	15	4	1

Data files

The full output of the Fisher's exact tests are available on-line (UPenn file sharing service).

[Enrichment of guide sequences]
[Enrichment of targeted genes]

GO term enrichment

Table 3 below details the result of a GO term enrichment analysis (Fatigo) using the gene names from the guide sequence enrichment where a cut-off of Fisher's exact odd ratio of $\leq 1/4$ was used (Table 1, 146 genes). The full analysis output which includes the enriched gene names can be [downloaded here]. Fatigo is a free, web-based tool (http://babelomics.bioinfo.cipf.es) for studying gene enrichment. Further use of this tool with the gene names provided in the linked data files may prove insightful.

Table 3. Results of GO term enrichment analysis

term	adj_pvalue
potassium ion transmembrane transport(GO:0071805)	0.0000572
cellular potassium ion transport(GO:0071804)	0.0000572
central nervous system neuron differentiation(GO:0021953)	0.0025968
cartilage development(GO:0051216)	0.0055568
bone morphogenesis(GO:0060349)	0.0055568
multicellular organismal signaling(GO:0035637)	0.0059206
regulation of heart contraction(GO:0008016)	0.0059206
chondrocyte differentiation(GO:0002062)	0.0059907
regulation of ossification(GO:0030278)	0.0059907
positive regulation of endocytosis(GO:0045807)	0.0063220
connective tissue development(GO:0061448)	0.0063220
membrane repolarization(GO:0086009)	0.0064770
regulation of nucleocytoplasmic transport(GO:0046822)	0.0064770
action potential(GO:0001508)	0.0066223
osteoblast differentiation(GO:0001649)	0.0067475
insulin receptor signaling pathway(GO:0008286)	0.0067475
adult behavior(GO:0030534)	0.0086532
positive regulation of receptor-mediated endocytosis(GO:0048260)	0.0088666
appendage morphogenesis(GO:0035107)	0.0089555
limb morphogenesis(GO:0035108)	0.0089555
JAK-STAT cascade(GO:0007259)	0.0098654
bone development(GO:0060348)	0.0098654
endochondral bone morphogenesis(GO:0060350)	0.0098654
regulation of intracellular protein transport(GO:0033157)	0.0100003
regulation of osteoblast differentiation(GO:0045667)	0.0100003
cellular response to acid chemical(GO:0071229)	0.0114147
appendage development(GO:0048736)	0.0140081
limb development(GO:0060173)	0.0140081
purine nucleoside biosynthetic process(GO:0042451)	0.0156432
purine ribonucleoside biosynthetic process(GO:0046129)	0.0156432
T cell differentiation in thymus(GO:0033077)	0.0156716

term	adj_pvalue
regulation of carbohydrate metabolic process(GO:0006109)	0.0156778
ATP biosynthetic process(GO:0006754)	0.0160348
negative regulation of membrane potential(GO:0045837)	0.0187657
skeletal system morphogenesis(GO:0048705)	0.0194374
regulation of protein import into nucleus(GO:0042306)	0.0194770
telencephalon development(GO:0021537)	0.0194770
tissue homeostasis(GO:0001894)	0.0194770
negative regulation of protein transport(GO:0051224)	0.0198969
ribonucleoside biosynthetic process(GO:0042455)	0.0200920
potassium ion export(GO:0071435)	0.0203379
hepatocyte growth factor receptor signaling pathway(GO:0048012)	0.0203379
regulation of receptor-mediated endocytosis(GO:0048259)	0.0203379
regulation of endocytosis(GO:0030100)	0.0211839
myeloid leukocyte differentiation(GO:0002573)	0.0214160
glycosyl compound biosynthetic process(GO:1901659)	0.0214413
regulation of protein localization to nucleus(GO:1900180)	0.0214413
thymic T cell selection(GO:0045061)	0.0214413
negative regulation of multicellular organism growth(GO:0040015)	0.0214413
embryonic skeletal system morphogenesis(GO:0048704)	0.0214413
nucleoside biosynthetic process(GO:0009163)	0.0214413
positive regulation of osteoblast differentiation(GO:0045669)	0.0214413
regulation of myeloid cell differentiation(GO:0045637)	0.0214413
regulation of cation channel activity(GO:2001257)	0.0214413
interleukin-6-mediated signaling pathway(GO:0070102)	0.0222686
cellular sodium ion homeostasis(GO:0006883)	0.0222686
endoderm development(GO:0007492)	0.0230002
potassium ion import(GO:0010107)	0.0232817
polysaccharide metabolic process(GO:0005976)	0.0242950
positive regulation by host of viral transcription(GO:0043923)	0.0242950
cartilage development involved in endochondral bone morphogenesis(GO:0060351)	0.0257231
cognition(GO:0050890)	0.0262692
chondrocyte development(GO:0002063)	0.0267392
tongue development(GO:0043586)	0.0281886
positive regulation of response to external stimulus(GO:0032103)	0.0283422
myelination(GO:0042552)	0.0283422
dorsal/ventral pattern formation(GO:0009953)	0.0283422
response to acid chemical(GO:0001101)	0.0283422
positive regulation of neuron projection development(GO:0010976)	0.0285477
purine ribonucleoside triphosphate biosynthetic process(GO:0009206)	0.0285477
purine nucleoside triphosphate biosynthetic process(GO:0009145)	0.0285477
regulation of epidermal growth factor receptor signaling pathway(GO:0042058)	0.0303760
regulation of leukocyte differentiation(GO:1902105)	0.0303760
multicellular organismal homeostasis(GO:0048871)	0.0303760
protein localization to nucleus(GO:0034504)	0.0321778
neuronal action potential(GO:0019228)	0.0321778
regulation of ERBB signaling pathway(GO:1901184)	0.0321778
embryonic limb morphogenesis(GO:0030326)	0.0321778
embryonic appendage morphogenesis(GO:0035113)	0.0321778
regulation of carbohydrate biosynthetic process(GO:0043255)	0.0321778
ensheathment of neurons(GO:0007272)	0.0321778

term	adj_pvalue
axon ensheathment(GO:0008366)	0.0321778
ribonucleoside triphosphate biosynthetic process(GO:0009201)	0.0321778
monovalent inorganic cation homeostasis(GO:0055067)	0.0321778
negative regulation of JAK-STAT cascade(GO:0046426)	0.0321778
purine nucleoside monophosphate biosynthetic process(GO:0009127)	0.0324413
purine ribonucleoside monophosphate biosynthetic process(GO:0009168)	0.0324413
embryonic skeletal system development(GO:0048706)	0.0325762
regulation of T cell differentiation in thymus(GO:0033081)	0.0326390
glycogen metabolic process(GO:0005977)	0.0330644
proximal/distal pattern formation(GO:0009954)	0.0330644
glucan metabolic process(GO:0044042)	0.0330644
cellular glucan metabolic process(GO:0006073)	0.0330644
energy reserve metabolic process(GO:0006112)	0.0330644
regulation of myeloid leukocyte differentiation(GO:0002761)	0.0333749
regulation of fatty acid oxidation(GO:0046320)	0.0333749
positive regulation of nucleocytoplasmic transport(GO:0046824)	0.0333749
protein targeting to nucleus(GO:0044744)	0.0340208
positive regulation of receptor internalization(GO:0002092)	0.0340208
protein import into nucleus(GO:0006606)	0.0340208
ventricular cardiac muscle cell action potential(GO:0086005)	0.0340208
single-organism nuclear import(GO:1902593)	0.0340208
phagocytosis(GO:0006909)	0.0340208
cellular ketone metabolic process(GO:0042180)	0.0340208
fatty acid transmembrane transport(GO:1902001)	0.0340208
pallium development(GO:0021543)	0.0340208
regulation of Wnt signaling pathway(GO:0030111)	0.0340208
protein secretion(GO:0009306)	0.0340208
nucleoside triphosphate biosynthetic process(GO:0009142)	0.0340208
regulation of T cell differentiation(GO:0045580)	0.0340208
receptor-mediated endocytosis(GO:0006898)	0.0342549
cell-cell junction organization(GO:0045216)	0.0347227
regulation of membrane repolarization(GO:0060306)	0.0360858
transmission of nerve impulse(GO:0019226)	0.0388752
ribonucleoside monophosphate biosynthetic process(GO:0009156)	0.0388752
replacement ossification(GO:0036075)	0.0406298
modulation by host of viral transcription(GO:0043921)	0.0406298
modulation by host of symbiont transcription(GO:0052472)	0.0406298
endochondral ossification(GO:0001958)	0.0406298
negative regulation of ion transmembrane transport(GO:0034766)	0.0412631
T cell selection(GO:0045058)	0.0412631
modulation of transcription in other organism involved in symbiotic interaction(GO:0052312)	0.0412631
histone lysine methylation(GO:0034968)	0.0439046
cellular response to interleukin-6(GO:0071354)	0.0439046
locomotory behavior(GO:0007626)	0.0449622
regulation of chemotaxis(GO:0050920)	0.0457025
regulation of catenin import into nucleus(GO:0035412)	0.0477621
nucleoside monophosphate biosynthetic process(GO:0009124)	0.0477621