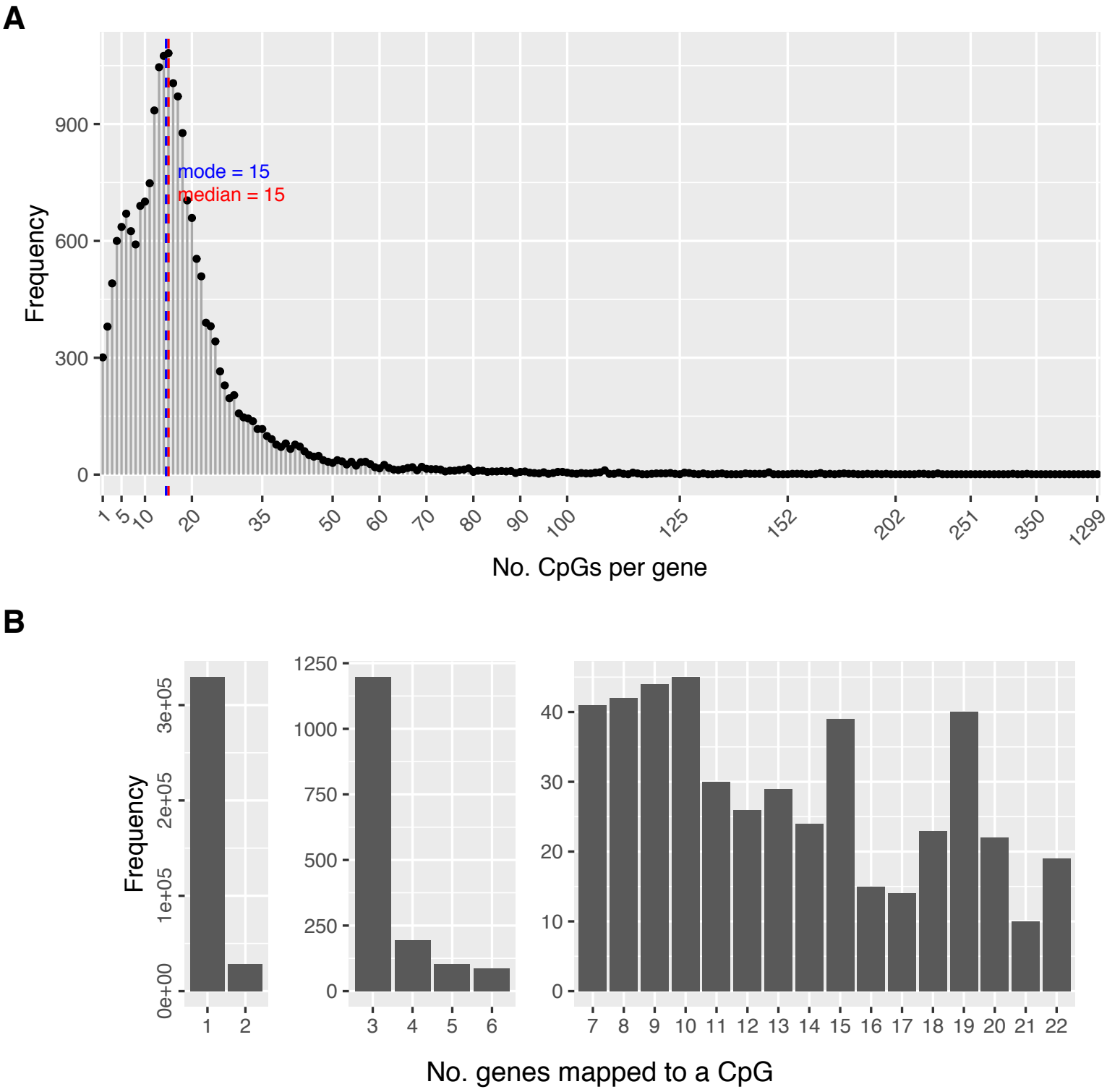
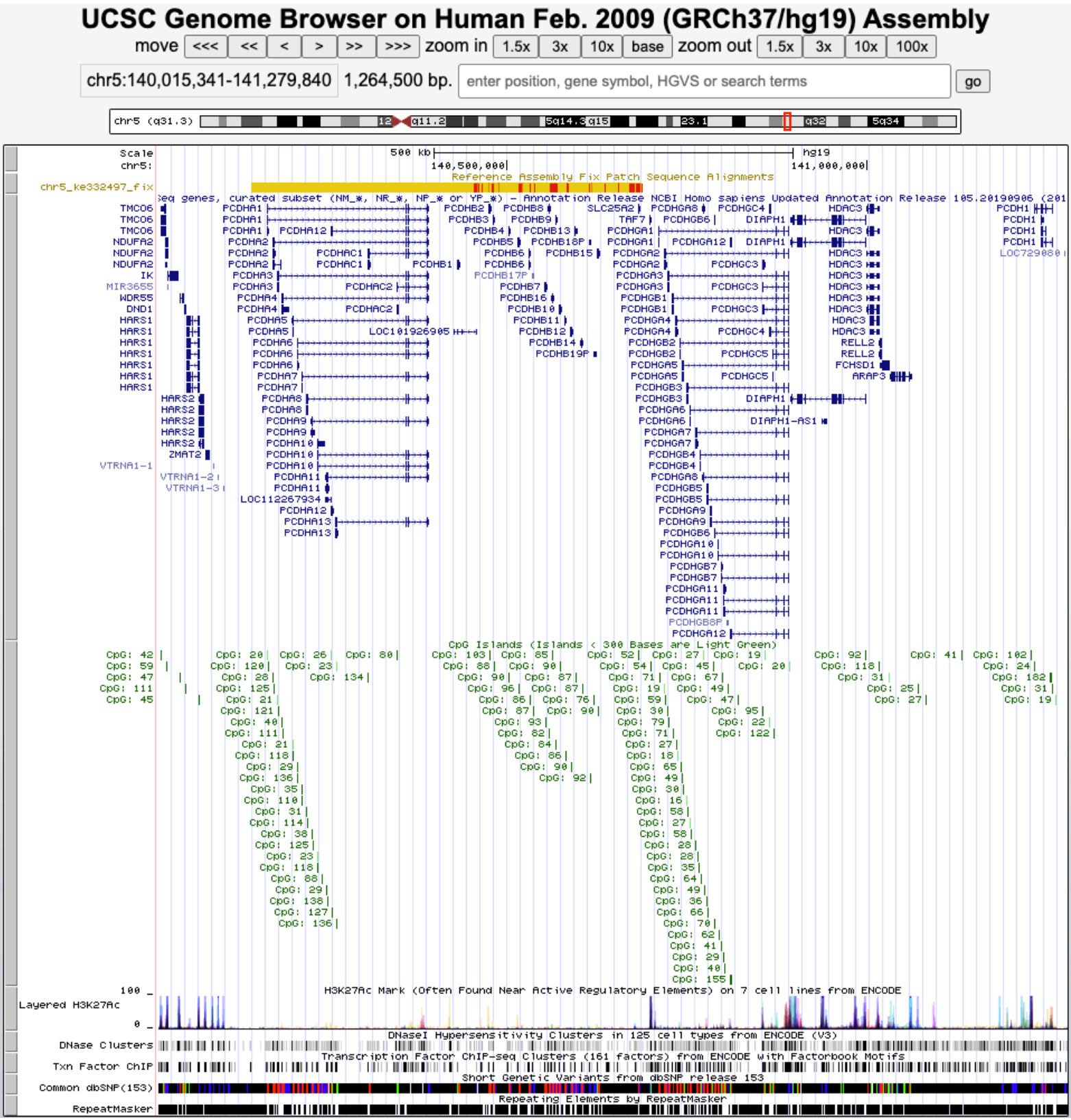
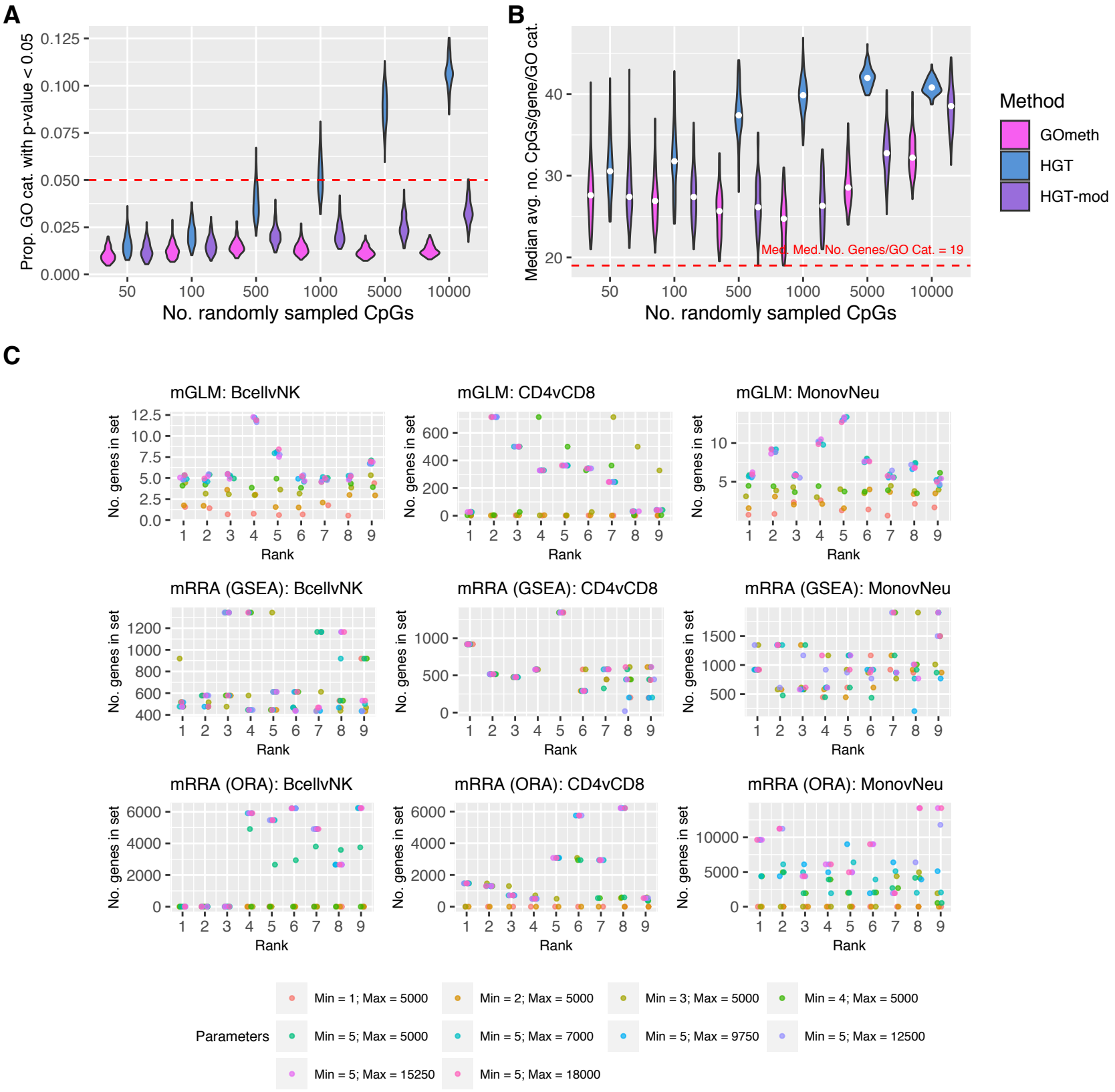

A
Frequency
900
600
300
0
mode = 15
median = 15
1
5
10
20
35
50
60
70
80
90
100
125
152
202
251
350
1299
No. CpGs per gene
B
Frequency
3e+05
2e+05
1e+05
0e+00
1250
1000
750
500
250
0
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
No. genes mapped to a CpG

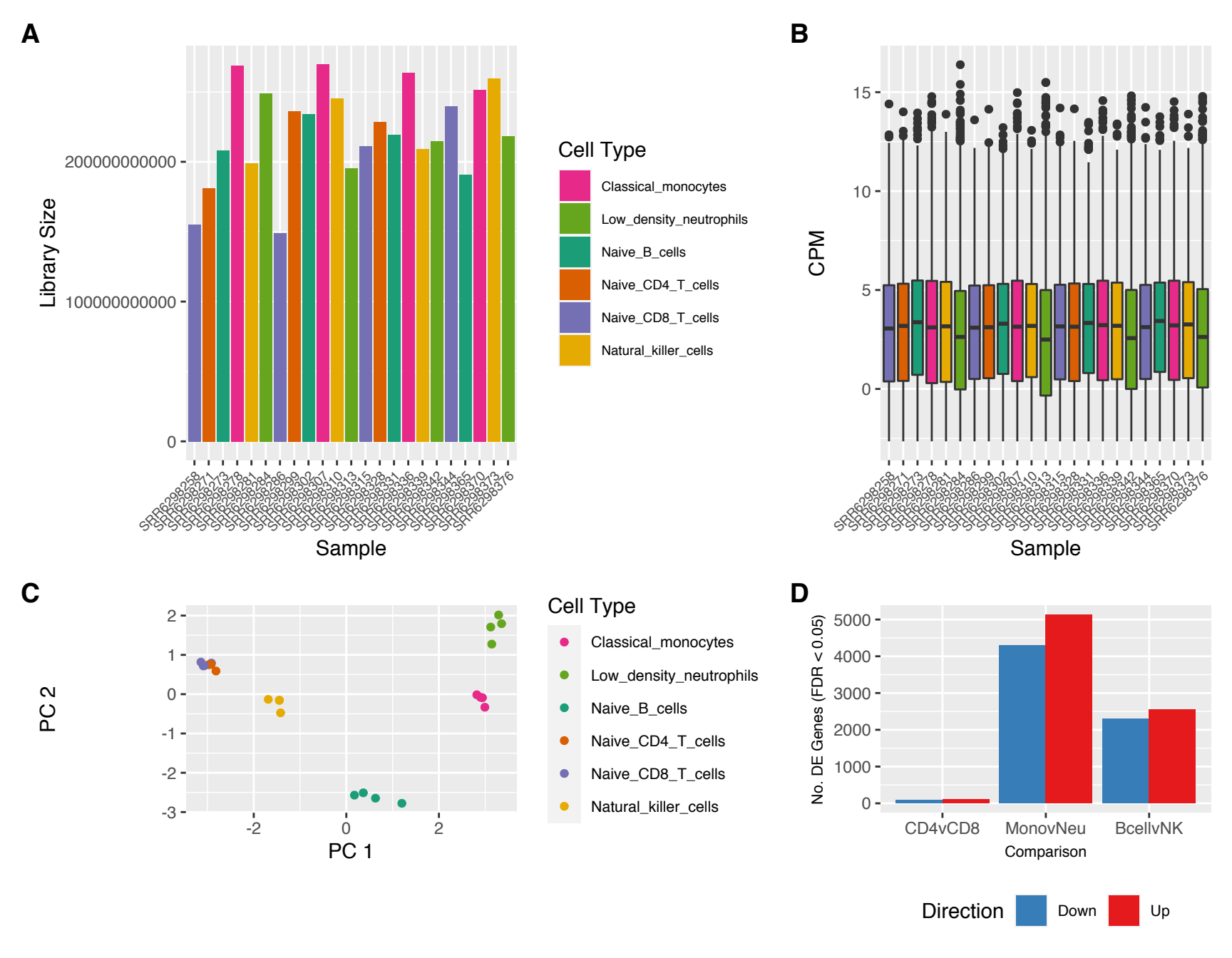
Supplementary Figure 1. **Array design bias for the Illumina HumanMethylation 450K BeadChip. (A)** Frequency plot of the numbers of CpGs measuring methylation across each gene for the 450K array. The most extreme value is 1299 CpGs measuring methylation for a single gene. The median is 15 and the mode is 15. **(B)** Split bar chart showing the numbers of genes annotated to each CpG (multi-gene bias). While the majority of CpGs are annotated to only one gene, there is still a large number annotated to 2 or more genes.


UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x
chr5:140,015,341-141,279,840 1,264,500 bp. enter position, gene symbol, HGVS or search terms go
chr5 (q31.3)
Scale
chr5:
500 kb
hg19
140,500,000
141,000,000
Reference Assembly Fix Patch Sequence Alignments
chr5_ke332497_fix
Seq genes, curated subset (NM_*, NR_*, NP_* or YP_*) - Annotation Release NCBI Homo sapiens Updated Annotation Release 105.20190906 (201
TMCO6
NDUFA2
IK
MIR3655
WDR55
DND1
HARS1
HARS2
ZMAT2
VTRNA1-1
VTRNA1-2
VTRNA1-3
PCDHA1
PCDHA2
PCDHAC1
PCDHAC2
PCDHA3
PCDHA4
PCDHA5
PCDHA6
PCDHA7
PCDHA8
PCDHA9
PCDHA10
PCDHA11
PCDHA12
PCDHA13
LOC112267934
LOC101926905
PCDHB1
PCDHB2
PCDHB3
PCDHB4
PCDHB5
PCDHB6
PCDHB7
PCDHB8
PCDHB9
PCDHB10
PCDHB11
PCDHB12
PCDHB13
PCDHB14
PCDHB15
PCDHB16
PCDHB17P
PCDHB18P
PCDHB19P
SLC25A2
TAF7
PCDHGA1
PCDHGA2
PCDHGA3
PCDHGA4
PCDHGA5
PCDHGA6
PCDHGA7
PCDHGA8
PCDHGA9
PCDHGA10
PCDHGA11
PCDHGA12
PCDHGB1
PCDHGB2
PCDHGB3
PCDHGB4
PCDHGB5
PCDHGB6
PCDHGB7
PCDHGB8P
PCDHGC3
PCDHGC4
PCDHGC5
PCDHGB6
DIAPH1
DIAPH1-AS1
HDAC3
RELL2
FCHSD1
ARAP3
PCDH1
LOC729680
CpG Islands (Islands < 300 Bases are Light Green)
100
Layered H3K27Ac
0
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Txn Factor ChIP
Transcription Factor ChIP-seq Clusters (161 factors) from ENCODE with Factorbook Motifs
Common dbSNP(153)
Short Genetic Variants from dbSNP release 153
RepeatMasker
Repeating Elements by RepeatMasker

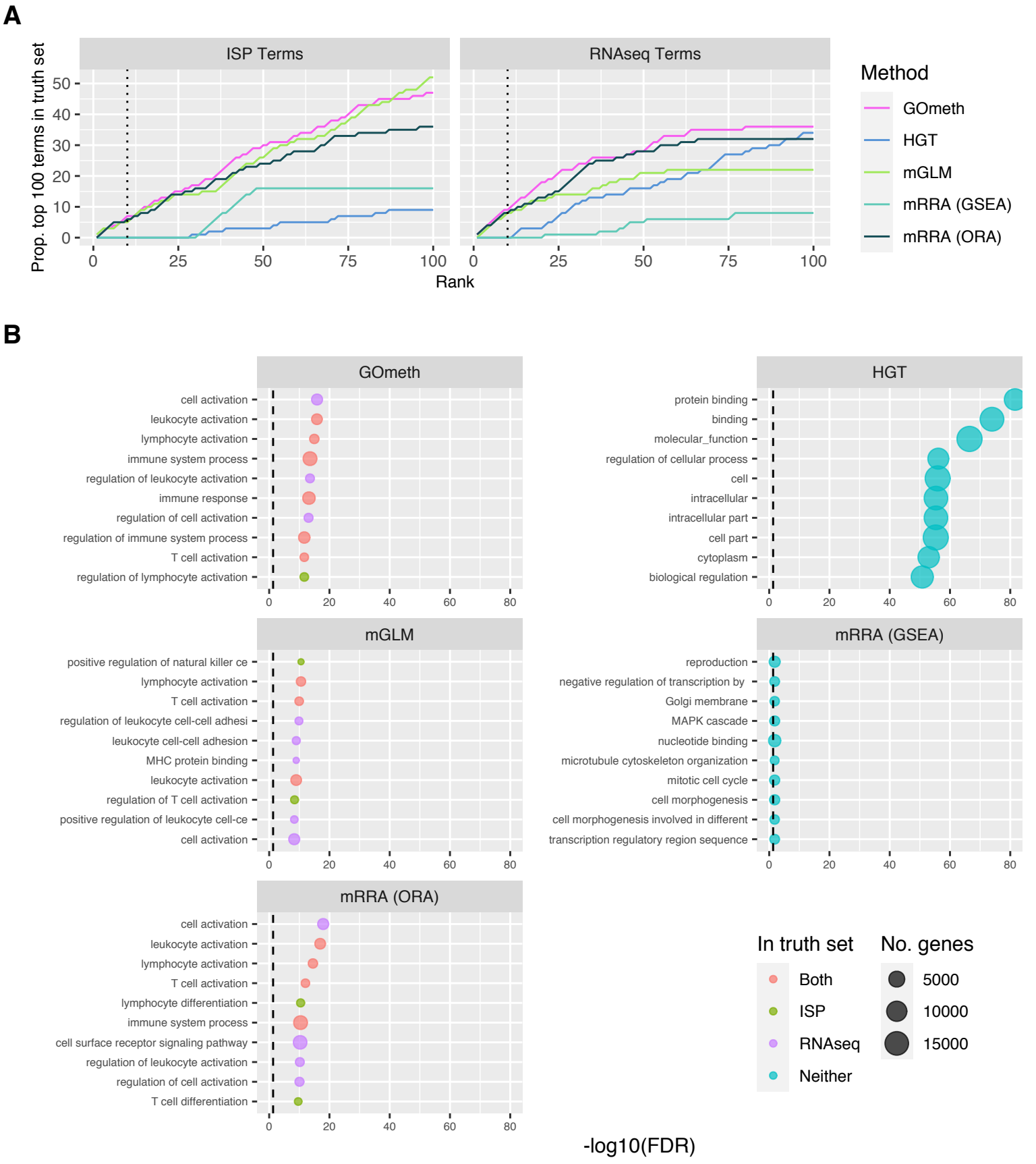
Supplementary Figure 2. UCSC genome browser snapshot of the protocadherin gamma cluster.


A
Prop. GO cat. with p-value < 0.05
0.125
0.100
0.075
0.050
0.025
0.000
50
100
500
1000
5000
10000
No. randomly sampled CpGs
B
Median avg. no. CpGs/gene/GO cat.
40
30
20
Med. Med. No. Genes/GO Cat. = 19
No. randomly sampled CpGs
Method
GOmeth
HGT
HGT-mod
C
mGLM: BcellvNK
mGLM: CD4vCD8
mGLM: MonovNeu
mRRA (GSEA): BcellvNK
mRRA (GSEA): CD4vCD8
mRRA (GSEA): MonovNeu
mRRA (ORA): BcellvNK
mRRA (ORA): CD4vCD8
mRRA (ORA): MonovNeu
No. genes in set
Rank
Parameters
Min = 1; Max = 5000
Min = 2; Max = 5000
Min = 3; Max = 5000
Min = 4; Max = 5000
Min = 5; Max = 5000
Min = 5; Max = 7000
Min = 5; Max = 9750
Min = 5; Max = 12500
Min = 5; Max = 15250
Min = 5; Max = 18000

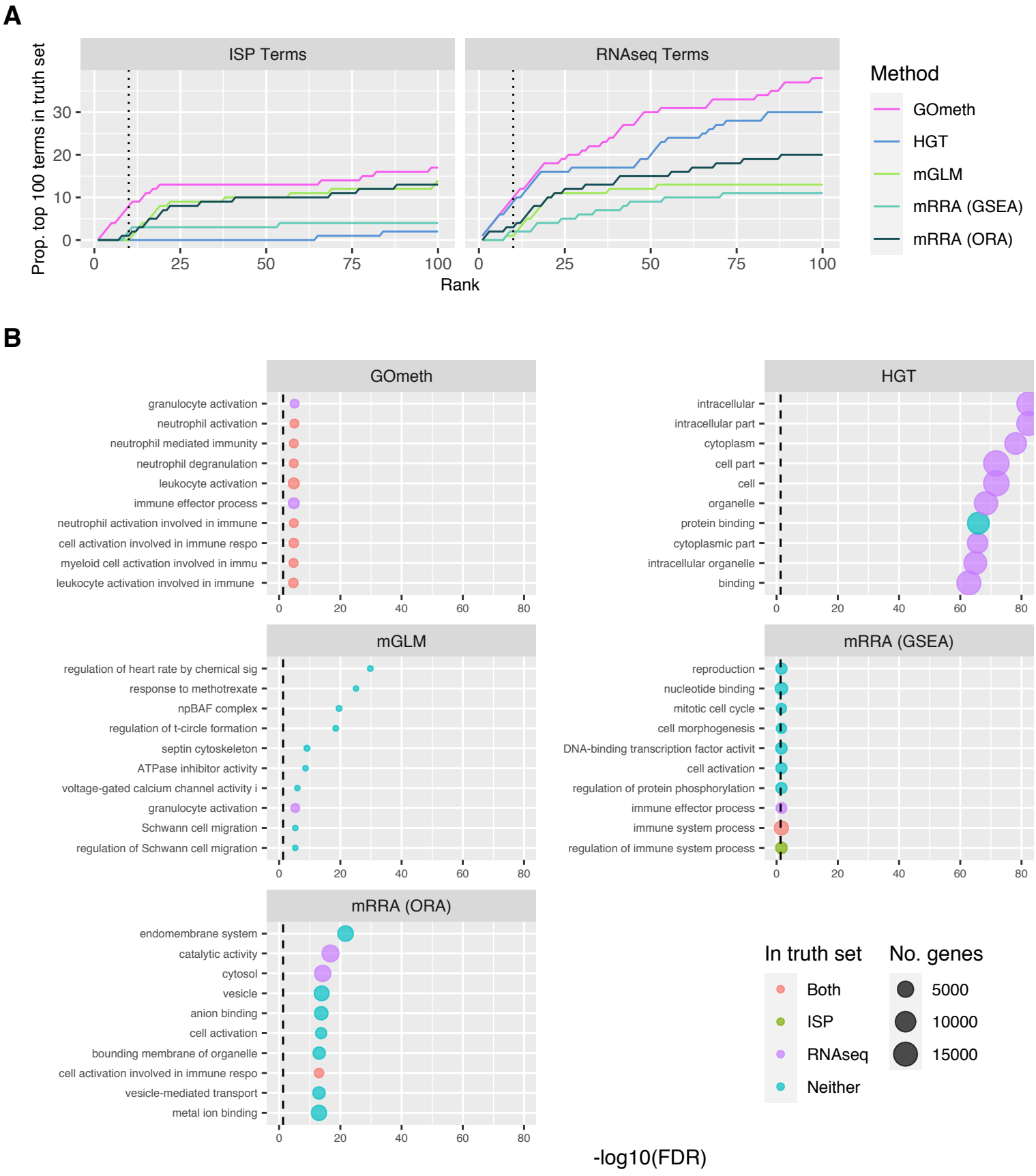
Supplementary Figure 3. Evaluation of false discovery rate control for 450K data. (A) Type I error rates across 100 simulations for varying numbers of randomly sampled CpGs. **(B)** Median average numbers of CpGs per gene for GO categories with an unadjusted p-value < 0.05. The hypergeometric test is biased towards GO categories with more CpGs per gene on average. GOmeth = adjust for probe-number and multi-gene bias; HGT = hypergeometric test; HGT-mod = adjust for probe-number bias only. **(C)** Gene set testing was performed on the results of the three blood cell type comparisons: CD4 T-cells vs. CD8 T-cells, monocytes vs. neutrophils and B-cells vs. NK cells, using the *MethylGSA* methods: mGLM, mRRA (GSEA) and mRRA (ORA), with several combinations of minimum and maximum gene set size parameters. When the minimum gene set size is set to less than five, mGLM and mRRA (ORA) highly rank very small gene sets containing the number of genes equal to the minimum size parameter. mRRA is also biased towards highly ranking very large gene sets, if they are not filtered out. mGLM = methylglm; mRRA (GSEA) = methylRRA using gene set enrichment analysis; mRRA (ORA) = methylRRA using over-representation analysis.


A
Library Size
200000000000
100000000000
0
Sample
Cell Type
Classical_monocytes
Low_density_neutrophils
Naive_B_cells
Naive_CD4_T_cells
Naive_CD8_T_cells
Natural_killer_cells
B
CPM
15
10
5
0
Sample
C
PC 2
2
1
0
-1
-2
-3
-2
0
2
PC 1
Cell Type
Classical_monocytes
Low_density_neutrophils
Naive_B_cells
Naive_CD4_T_cells
Naive_CD8_T_cells
Natural_killer_cells
D
No. DE Genes (FDR < 0.05)
5000
4000
3000
2000
1000
0
CD4vCD8
MonovNeu
BcellvNK
Comparison
Direction
Down
Up

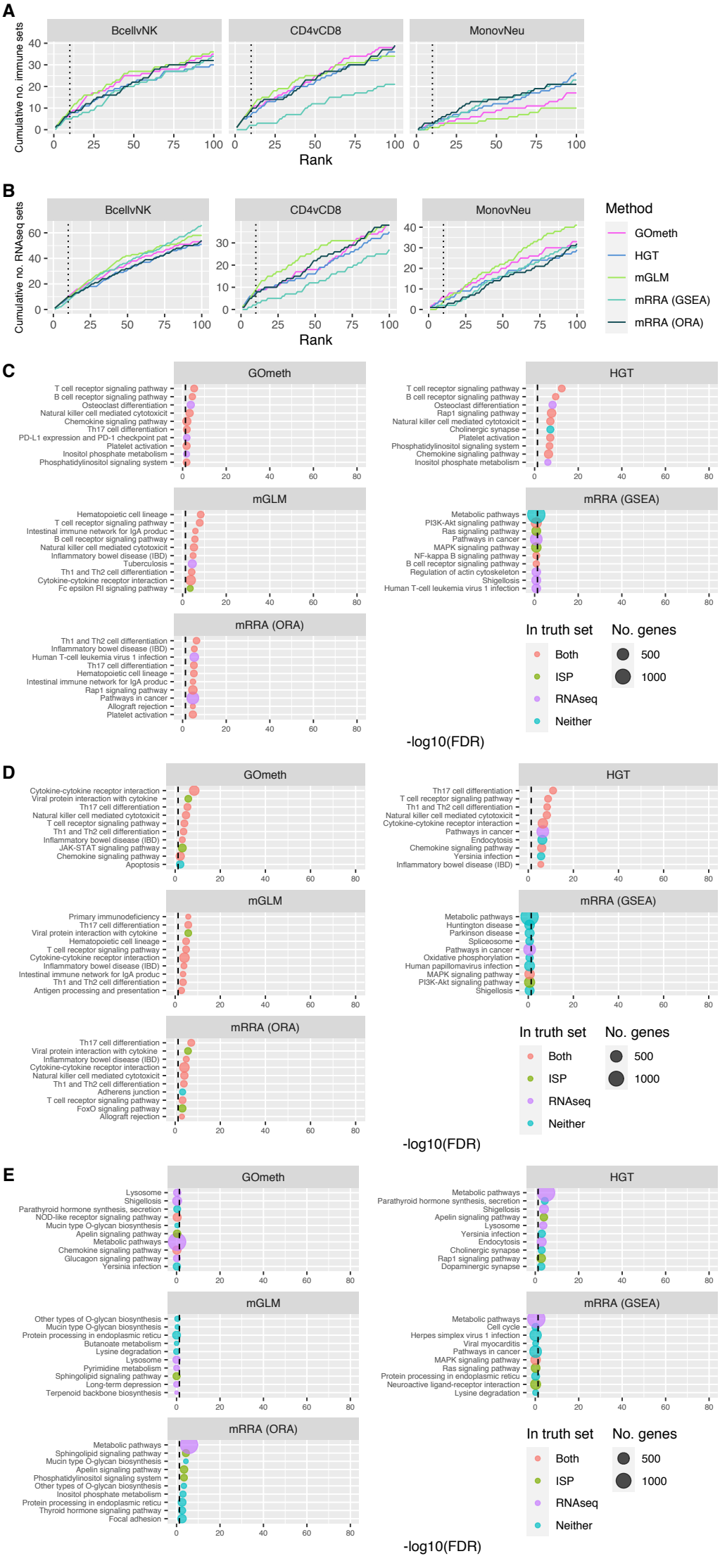
Supplementary Figure 4. Characteristics of sorted blood cell RNAseq data and analysis. **(A)** Library sizes (length-scaled TPM) for each of the samples after Salmon mapping and quantification. **(B)** Distribution of counts per million (CPM) after filtering and normalisation. **(C)** Multidimensional scaling plot of the filtered and normalised RNAseq data. **(D)** Numbers of differentially expressed genes with an adjusted p-value < 0.05, for each cell type comparison: CD4 T-cells vs. CD8 T-cells, monocytes vs. neutrophils and B-cells vs. NK cells. The blue bar is the number of significantly down-regulated genes, and the red bar is the number that are significantly up-regulated; e.g. ~2,300 genes are down-regulated and ~2500 are up-regulated in B-cells, compared to NK cells.


A
ISP Terms
RNAseq Terms
Prop. top 100 terms in truth set
Rank
Method
GOmeth
HGT
mGLM
mRRA (GSEA)
mRRA (ORA)
B
GOmeth
cell activation
leukocyte activation
lymphocyte activation
immune system process
regulation of leukocyte activation
immune response
regulation of cell activation
regulation of immune system process
T cell activation
regulation of lymphocyte activation
HGT
protein binding
binding
molecular_function
regulation of cellular process
cell
intracellular
intracellular part
cell part
cytoplasm
biological regulation
mGLM
positive regulation of natural killer ce
lymphocyte activation
T cell activation
regulation of leukocyte cell-cell adhesi
leukocyte cell-cell adhesion
MHC protein binding
leukocyte activation
regulation of T cell activation
positive regulation of leukocyte cell-ce
cell activation
mRRA (GSEA)
reproduction
negative regulation of transcription by
Golgi membrane
MAPK cascade
nucleotide binding
microtubule cytoskeleton organization
mitotic cell cycle
cell morphogenesis
cell morphogenesis involved in different
transcription regulatory region sequence
mRRA (ORA)
cell activation
leukocyte activation
lymphocyte activation
T cell activation
lymphocyte differentiation
immune system process
cell surface receptor signaling pathway
regulation of leukocyte activation
regulation of cell activation
T cell differentiation
In truth set
Both
ISP
RNAseq
Neither
No. genes
5000
10000
15000
-log10(FDR)

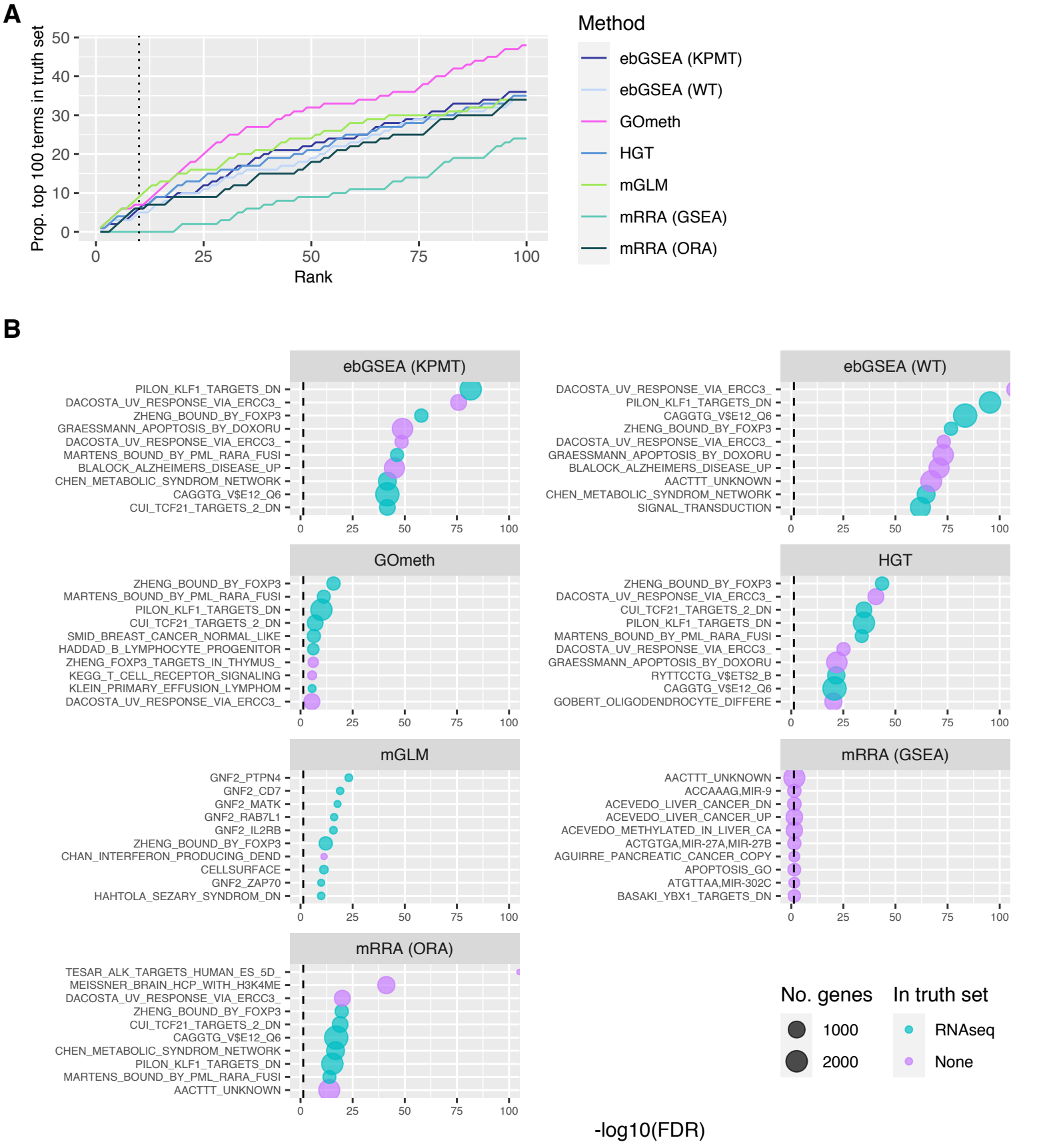
Supplementary Figure 5. Comparison of gene set testing performance for CD4 T cells vs. CD8 T-cells (GO). (A) Cumulative number of GO terms, as ranked by various methods, that are present in each truth set for the CD4 T-cells vs. CD8 T-cells comparison. ISP Terms = immune-system process child terms truth set; RNAseq Terms = top 100 terms from RNAseq analysis of the same cell types. **(B)** Bubble plots of the top 10 GO terms as ranked by various gene set testing methods for the CD4 T-cells vs. CD8 T-cells comparison. The size of the bubble indicates the relative number of genes in the set. The colour of the bubble indicates whether the term is present in either RNAseq (purple) or ISP (green) truth sets, both (red) or neither (blue). GOmeth = GOmeth using top 5000 most significant probes; HGT = hypergeometric test; mGLM = methylglm; mRRA (GSEA) = methylRRA using gene set enrichment analysis; mRRA (ORA) = methylRRA using over-representation analysis.


A
ISP Terms
RNAseq Terms
Prop. top 100 terms in truth set
Rank
Method
GOmeth
HGT
mGLM
mRRA (GSEA)
mRRA (ORA)
B
GOmeth
granulocyte activation
neutrophil activation
neutrophil mediated immunity
neutrophil degranulation
leukocyte activation
immune effector process
neutrophil activation involved in immune
cell activation involved in immune respo
myeloid cell activation involved in immu
leukocyte activation involved in immune
HGT
intracellular
intracellular part
cytoplasm
cell part
cell
organelle
protein binding
cytoplasmic part
intracellular organelle
binding
mGLM
regulation of heart rate by chemical sig
response to methotrexate
npBAF complex
regulation of t-circle formation
septin cytoskeleton
ATPase inhibitor activity
voltage-gated calcium channel activity i
granulocyte activation
Schwann cell migration
regulation of Schwann cell migration
mRRA (GSEA)
reproduction
nucleotide binding
mitotic cell cycle
cell morphogenesis
DNA-binding transcription factor activit
cell activation
regulation of protein phosphorylation
immune effector process
immune system process
regulation of immune system process
mRRA (ORA)
endomembrane system
catalytic activity
cytosol
vesicle
anion binding
cell activation
bounding membrane of organelle
cell activation involved in immune respo
vesicle-mediated transport
metal ion binding
In truth set
Both
ISP
RNAseq
Neither
No. genes
5000
10000
15000
-log10(FDR)

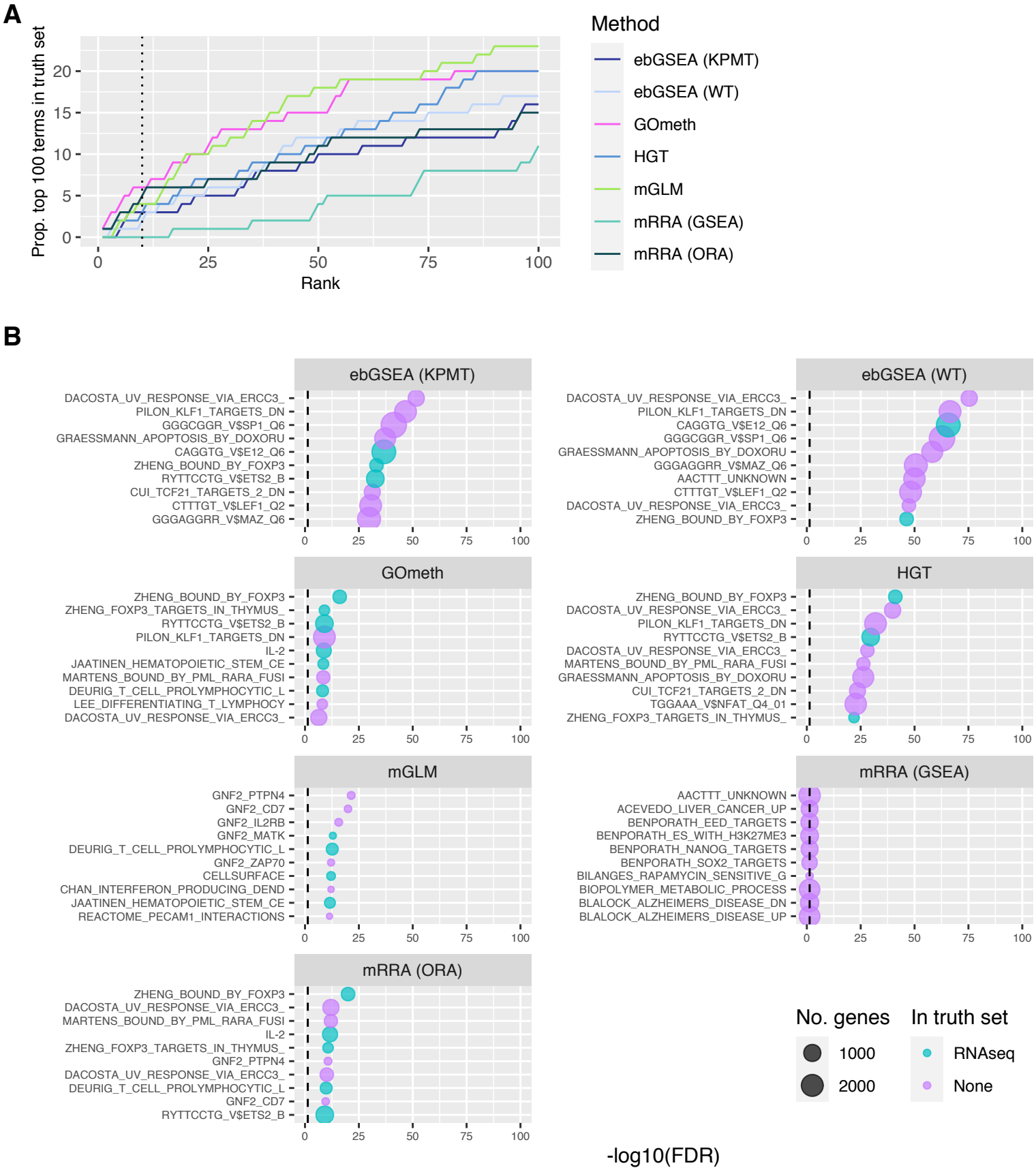
Supplementary Figure 6. Comparison of gene set testing performance for monocytes vs. neutrophils (GO). (A) Cumulative number of GO terms, as ranked by various methods, that are present in each truth set for the monocytes vs. neutrophils comparison. ISP Terms = immune-system process child terms truth set; RNAseq Terms = top 100 terms from RNAseq analysis of the same cell types. **(B)** Bubble plots of the top 10 GO terms as ranked by various gene set testing methods for the monocytes vs. neutrophils comparison. The size of the bubble indicates the relative number of genes in the set. The colour of the bubble indicates whether the term is present in either RNAseq (purple) or ISP (green) truth sets, both (red) or neither (blue). GOmeth = GOmeth using top 5000 most significant probes; HGT = hypergeometric test; mGLM = methylglm; mRRA (GSEA) = methylRRA using gene set enrichment analysis; mRRA (ORA) = methylRRA using over-representation analysis.


A
BcellvNK
CD4vCD8
MonovNeu
Cumulative no. immune sets
Rank
B
Cumulative no. RNAseq sets
Method
GOmeth
HGT
mGLM
mRRA (GSEA)
mRRA (ORA)
C
GOmeth
HGT
mGLM
mRRA (GSEA)
mRRA (ORA)
T cell receptor signaling pathway
B cell receptor signaling pathway
Osteoclast differentiation
Natural killer cell mediated cytotoxicit
Chemokine signaling pathway
Th17 cell differentiation
PD-L1 expression and PD-1 checkpoint pat
Platelet activation
Inositol phosphate metabolism
Phosphatidylinositol signaling system
Rap1 signaling pathway
Cholinergic synapse
Phosphatidylinositol signaling system
Hematopoietic cell lineage
Intestinal immune network for IgA produc
Inflammatory bowel disease (IBD)
Tuberculosis
Th1 and Th2 cell differentiation
Cytokine-cytokine receptor interaction
Fc epsilon RI signaling pathway
Metabolic pathways
PI3K-Akt signaling pathway
Ras signaling pathway
Pathways in cancer
MAPK signaling pathway
NF-kappa B signaling pathway
B cell receptor signaling pathway
Regulation of actin cytoskeleton
Shigellosis
Human T-cell leukemia virus 1 infection
Human T-cell leukemia virus 1 infection
Allograft rejection
In truth set
Both
ISP
RNAseq
Neither
No. genes
500
1000
-log10(FDR)
D
Cytokine-cytokine receptor interaction
Viral protein interaction with cytokine
JAK-STAT signaling pathway
Apoptosis
Primary immunodeficiency
Antigen processing and presentation
Huntington disease
Parkinson disease
Spliceosome
Oxidative phosphorylation
Human papillomavirus infection
Adherens junction
FoxO signaling pathway
E
Lysosome
Shigellosis
Parathyroid hormone synthesis, secretion
NOD-like receptor signaling pathway
Mucin type O-glycan biosynthesis
Apelin signaling pathway
Metabolic pathways
Glucagon signaling pathway
Yersinia infection
Endocytosis
Dopaminergic synapse
Other types of O-glycan biosynthesis
Protein processing in endoplasmic reticu
Butanoate metabolism
Lysine degradation
Pyrimidine metabolism
Sphingolipid signaling pathway
Long-term depression
Terpenoid backbone biosynthesis
Cell cycle
Herpes simplex virus 1 infection
Viral myocarditis
Neuroactive ligand-receptor interaction
Thyroid hormone signaling pathway
Focal adhesion

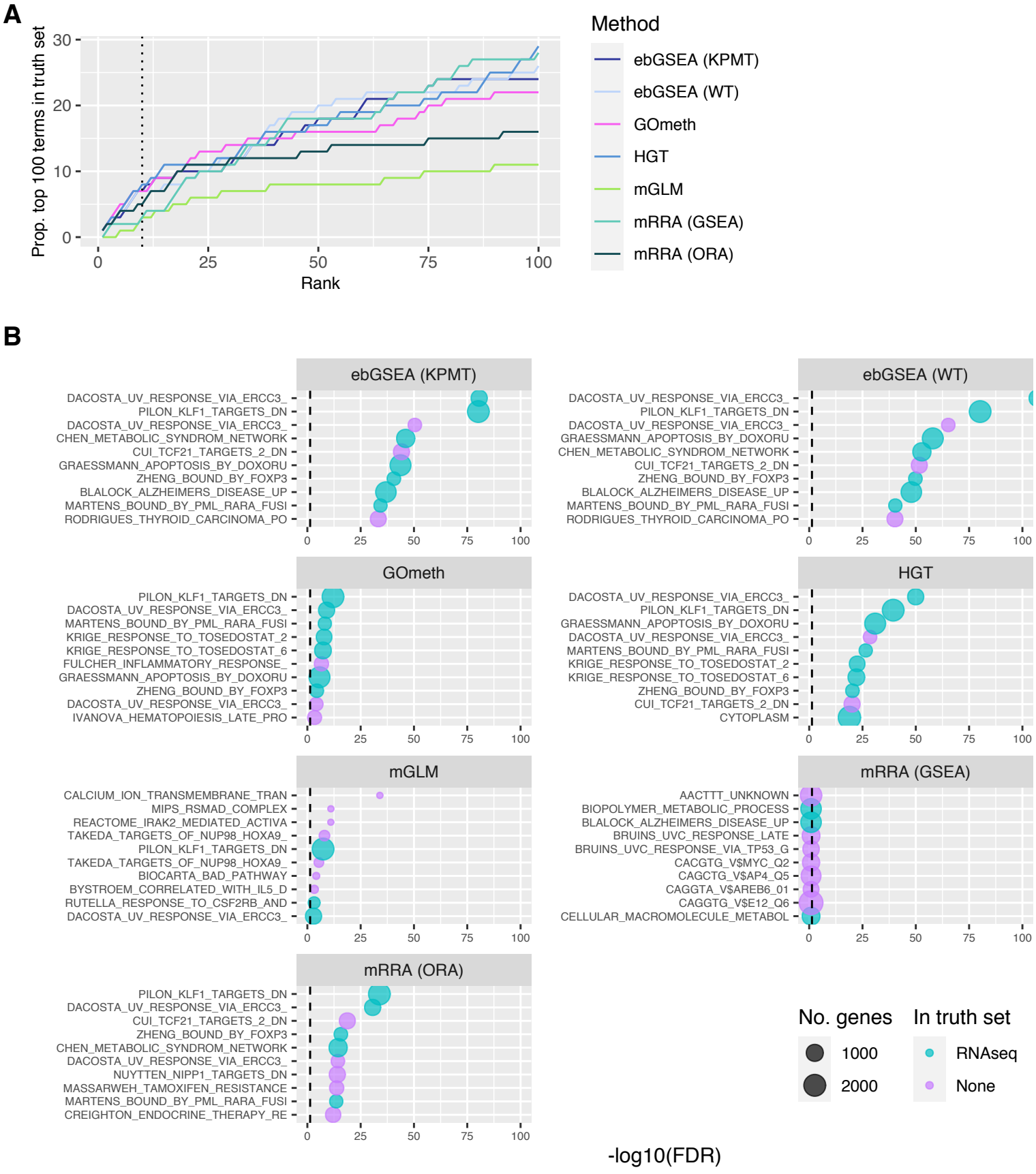
Supplementary Figure 7. Comparison of gene set testing performance (KEGG). (A) Cumulative number of KEGG pathways, as ranked by various methods, that are present in the immune truth set for all comparisons. Immune sets = all pathways belonging to the following categories: Immune system, Immune disease, Signal transduction, Signaling molecules and interaction. **(B)** Cumulative number of KEGG pathways, as ranked by various methods, that are present in the RNAseq truth set for all comparisons. RNAseq Terms = top 100 KEGG pathways from RNAseq analysis of the same cell types. **(C)** Bubble plots of the top 10 KEGG pathways as ranked by various gene set testing methods for the B-cell vs. NK comparison. The size of the bubble indicates the relative number of genes in the set. The colour of the bubble indicates whether the term is present in either RNAseq (purple) or ISP (green) truth sets, both (red) or neither (blue). **(D)** Bubble plots of the top 10 KEGG pathways as ranked by various gene set testing methods for the CD4 T-cells vs. CD8 T-cells comparison. **(E)** Bubble plots of the top 10 KEGG pathways as ranked by various gene set testing methods for the monocytes vs. neutrophils comparison. GOmeth = GOmeth using top 5000 most significant probes; HGT = hypergeometric test; mGLM = methylglm; mRRA (GSEA) = methylRRA using gene set enrichment analysis; mRRA (ORA) = methylRRA using over-representation analysis.


A
Prop. top 100 terms in truth set
Rank
Method
ebGSEA (KPMT)
ebGSEA (WT)
GOmeth
HGT
mGLM
mRRA (GSEA)
mRRA (ORA)
B
ebGSEA (KPMT)
PILON_KLF1_TARGETS_DN
DACOSTA_UV_RESPONSE_VIA_ERCC3_
ZHENG_BOUND_BY_FOXP3
GRAESSMANN_APOPTOSIS_BY_DOXORU
MARTENS_BOUND_BY_PML_RARA_FUSI
BLALOCK_ALZHEIMERS_DISEASE_UP
CHEN_METABOLIC_SYNDROM_NETWORK
CAGGTG_V$E12_Q6
CUI_TCF21_TARGETS_2_DN
ebGSEA (WT)
DACOSTA_UV_RESPONSE_VIA_ERCC3_
PILON_KLF1_TARGETS_DN
CAGGTG_V$E12_Q6
ZHENG_BOUND_BY_FOXP3
DACOSTA_UV_RESPONSE_VIA_ERCC3_
GRAESSMANN_APOPTOSIS_BY_DOXORU
BLALOCK_ALZHEIMERS_DISEASE_UP
AACTTT_UNKNOWN
CHEN_METABOLIC_SYNDROM_NETWORK
SIGNAL_TRANSDUCTION
GOmeth
ZHENG_BOUND_BY_FOXP3
MARTENS_BOUND_BY_PML_RARA_FUSI
PILON_KLF1_TARGETS_DN
CUI_TCF21_TARGETS_2_DN
SMID_BREAST_CANCER_NORMAL_LIKE
HADDAD_B_LYMPHOCYTE_PROGENITOR
ZHENG_FOXP3_TARGETS_IN_THYMUS_
KEGG_T_CELL_RECEPTOR_SIGNALING
KLEIN_PRIMARY_EFFUSION_LYMPHOM
DACOSTA_UV_RESPONSE_VIA_ERCC3_
HGT
ZHENG_BOUND_BY_FOXP3
DACOSTA_UV_RESPONSE_VIA_ERCC3_
CUI_TCF21_TARGETS_2_DN
PILON_KLF1_TARGETS_DN
MARTENS_BOUND_BY_PML_RARA_FUSI
DACOSTA_UV_RESPONSE_VIA_ERCC3_
GRAESSMANN_APOPTOSIS_BY_DOXORU
RYTTCCTG_V$ETS2_B
CAGGTG_V$E12_Q6
GOBERT_OLIGODENDROCYTE_DIFFERE
mGLM
GNF2_PTPN4
GNF2_CD7
GNF2_MATK
GNF2_RAB7L1
GNF2_IL2RB
ZHENG_BOUND_BY_FOXP3
CHAN_INTERFERON_PRODUCING_DEND
CELLSURFACE
GNF2_ZAP70
HAHTOLA_SEZARY_SYNDROM_DN
mRRA (GSEA)
AACTTT_UNKNOWN
ACCAAAG,MIR-9
ACEVEDO_LIVER_CANCER_DN
ACEVEDO_LIVER_CANCER_UP
ACEVEDO_METHYLATED_IN_LIVER_CA
ACTGTGA,MIR-27A,MIR-27B
AGUIRRE_PANCREATIC_CANCER_COPY
APOPTOSIS_GO
ATGTTAA,MIR-302C
BASAKI_YBX1_TARGETS_DN
mRRA (ORA)
TESAR_ALK_TARGETS_HUMAN_ES_5D_
MEISSNER_BRAIN_HCP_WITH_H3K4ME
DACOSTA_UV_RESPONSE_VIA_ERCC3_
ZHENG_BOUND_BY_FOXP3
CUI_TCF21_TARGETS_2_DN
CAGGTG_V$E12_Q6
CHEN_METABOLIC_SYNDROM_NETWORK
PILON_KLF1_TARGETS_DN
MARTENS_BOUND_BY_PML_RARA_FUSI
AACTTT_UNKNOWN
No. genes
1000
2000
In truth set
RNAseq
None
-log10(FDR)

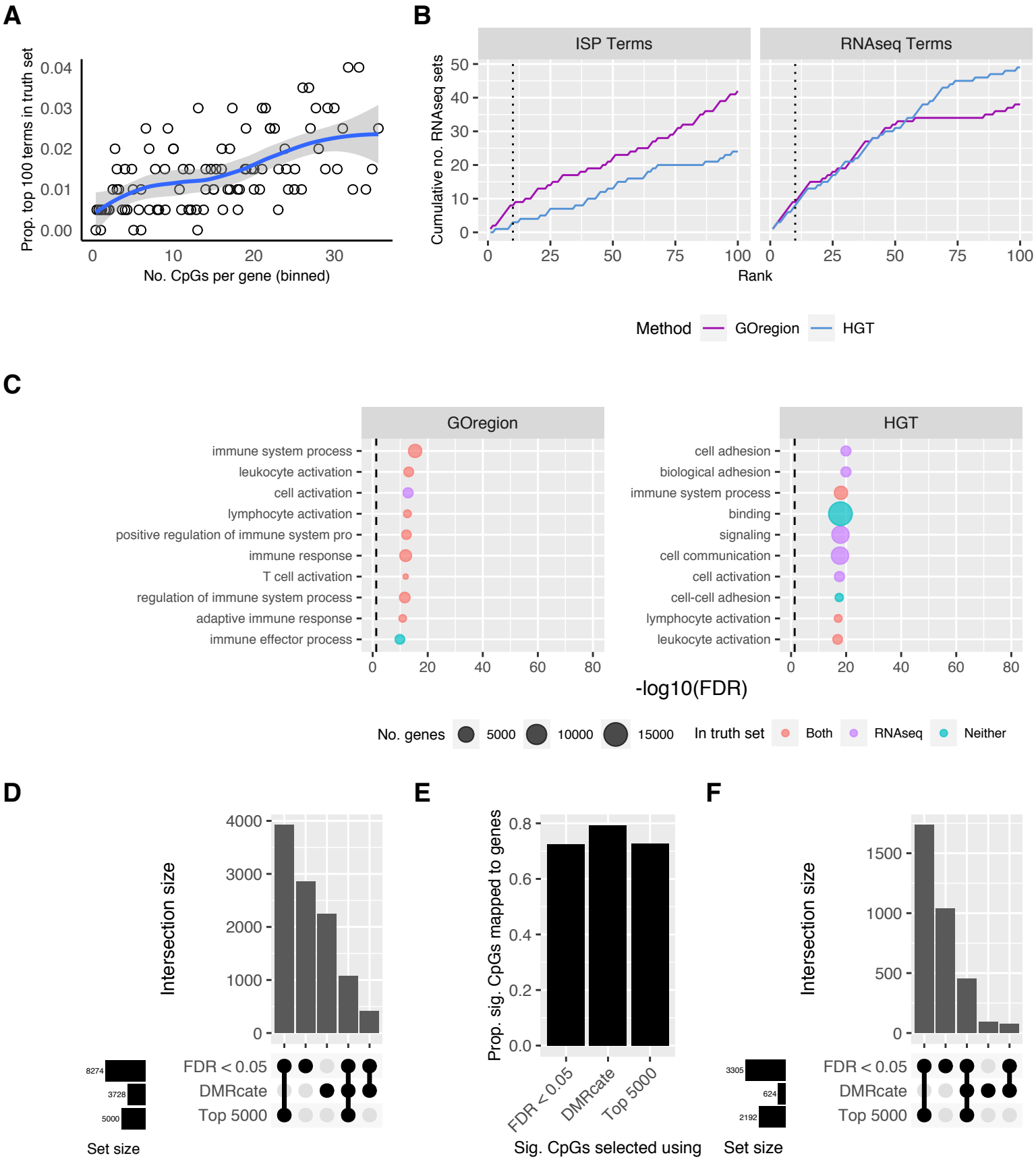
Supplementary Figure 8. Comparison of gene set testing performance for B-cells vs. NK cells (BROAD). (A) Cumulative number of BROAD MSigDB gene sets, as ranked by various methods, that are present in the RNAseq truth set for the B-cells vs. NK comparison. RNAseq truth = top 100 BROAD MSigDB gene sets from RNAseq analysis of the same cell types. **(B)** Bubble plots of the top 10 BROAD MSigDB gene sets as ranked by various gene set testing methods for the B-cells vs. NK comparison. The size of the bubble indicates the relative number of genes in the set. The colour of the bubble indicates whether the gene set is present in the RNAseq truth set (purple) or not (blue). ebGSEA (KPMT) = ebGSEA using Known Population Median Test; ebGSEA (WT) = ebGSEA using Wilcoxon Test; GOmeth = GOmeth using top 5000 most significant probes; HGT = hypergeometric test; mGLM = methylglm; mRRA (GSEA) = methylRRA using gene set enrichment analysis; mRRA (ORA) = methylRRA using over-representation analysis.


A
Prop. top 100 terms in truth set
Rank
Method
ebGSEA (KPMT)
ebGSEA (WT)
GOmeth
HGT
mGLM
mRRA (GSEA)
mRRA (ORA)
B
ebGSEA (KPMT)
DACOSTA_UV_RESPONSE_VIA_ERCC3_
PILON_KLF1_TARGETS_DN
GGGCGGR_V$SP1_Q6
GRAESSMANN_APOPTOSIS_BY_DOXORU
CAGGTG_V$E12_Q6
ZHENG_BOUND_BY_FOXP3
RYTTCCTG_V$ETS2_B
CUI_TCF21_TARGETS_2_DN
CTTTGT_V$LEF1_Q2
GGGAGGRR_V$MAZ_Q6
ebGSEA (WT)
DACOSTA_UV_RESPONSE_VIA_ERCC3_
PILON_KLF1_TARGETS_DN
CAGGTG_V$E12_Q6
GGGCGGR_V$SP1_Q6
GRAESSMANN_APOPTOSIS_BY_DOXORU
GGGAGGRR_V$MAZ_Q6
AACTTT_UNKNOWN
CTTTGT_V$LEF1_Q2
DACOSTA_UV_RESPONSE_VIA_ERCC3_
ZHENG_BOUND_BY_FOXP3
GOmeth
ZHENG_BOUND_BY_FOXP3
ZHENG_FOXP3_TARGETS_IN_THYMUS_
RYTTCCTG_V$ETS2_B
PILON_KLF1_TARGETS_DN
IL-2
JAATINEN_HEMATOPOIETIC_STEM_CE
MARTENS_BOUND_BY_PML_RARA_FUSI
DEURIG_T_CELL_PROLYMPHOCYTIC_L
LEE_DIFFERENTIATING_T_LYMPHOCY
DACOSTA_UV_RESPONSE_VIA_ERCC3_
HGT
ZHENG_BOUND_BY_FOXP3
DACOSTA_UV_RESPONSE_VIA_ERCC3_
PILON_KLF1_TARGETS_DN
RYTTCCTG_V$ETS2_B
DACOSTA_UV_RESPONSE_VIA_ERCC3_
MARTENS_BOUND_BY_PML_RARA_FUSI
GRAESSMANN_APOPTOSIS_BY_DOXORU
CUI_TCF21_TARGETS_2_DN
TGGAAA_V$NFAT_Q4_01
ZHENG_FOXP3_TARGETS_IN_THYMUS_
mGLM
GNF2_PTPN4
GNF2_CD7
GNF2_IL2RB
GNF2_MATK
DEURIG_T_CELL_PROLYMPHOCYTIC_L
GNF2_ZAP70
CELLSURFACE
CHAN_INTERFERON_PRODUCING_DEND
JAATINEN_HEMATOPOIETIC_STEM_CE
REACTOME_PECAM1_INTERACTIONS
mRRA (GSEA)
AACTTT_UNKNOWN
ACEVEDO_LIVER_CANCER_UP
BENPORATH_EED_TARGETS
BENPORATH_ES_WITH_H3K27ME3
BENPORATH_NANOG_TARGETS
BENPORATH_SOX2_TARGETS
BILANGES_RAPAMYCIN_SENSITIVE_G
BIOPOLYMER_METABOLIC_PROCESS
BLALOCK_ALZHEIMERS_DISEASE_DN
BLALOCK_ALZHEIMERS_DISEASE_UP
mRRA (ORA)
ZHENG_BOUND_BY_FOXP3
DACOSTA_UV_RESPONSE_VIA_ERCC3_
MARTENS_BOUND_BY_PML_RARA_FUSI
IL-2
ZHENG_FOXP3_TARGETS_IN_THYMUS_
GNF2_PTPN4
DACOSTA_UV_RESPONSE_VIA_ERCC3_
DEURIG_T_CELL_PROLYMPHOCYTIC_L
GNF2_CD7
RYTTCCTG_V$ETS2_B
No. genes
1000
2000
In truth set
RNAseq
None
-log10(FDR)

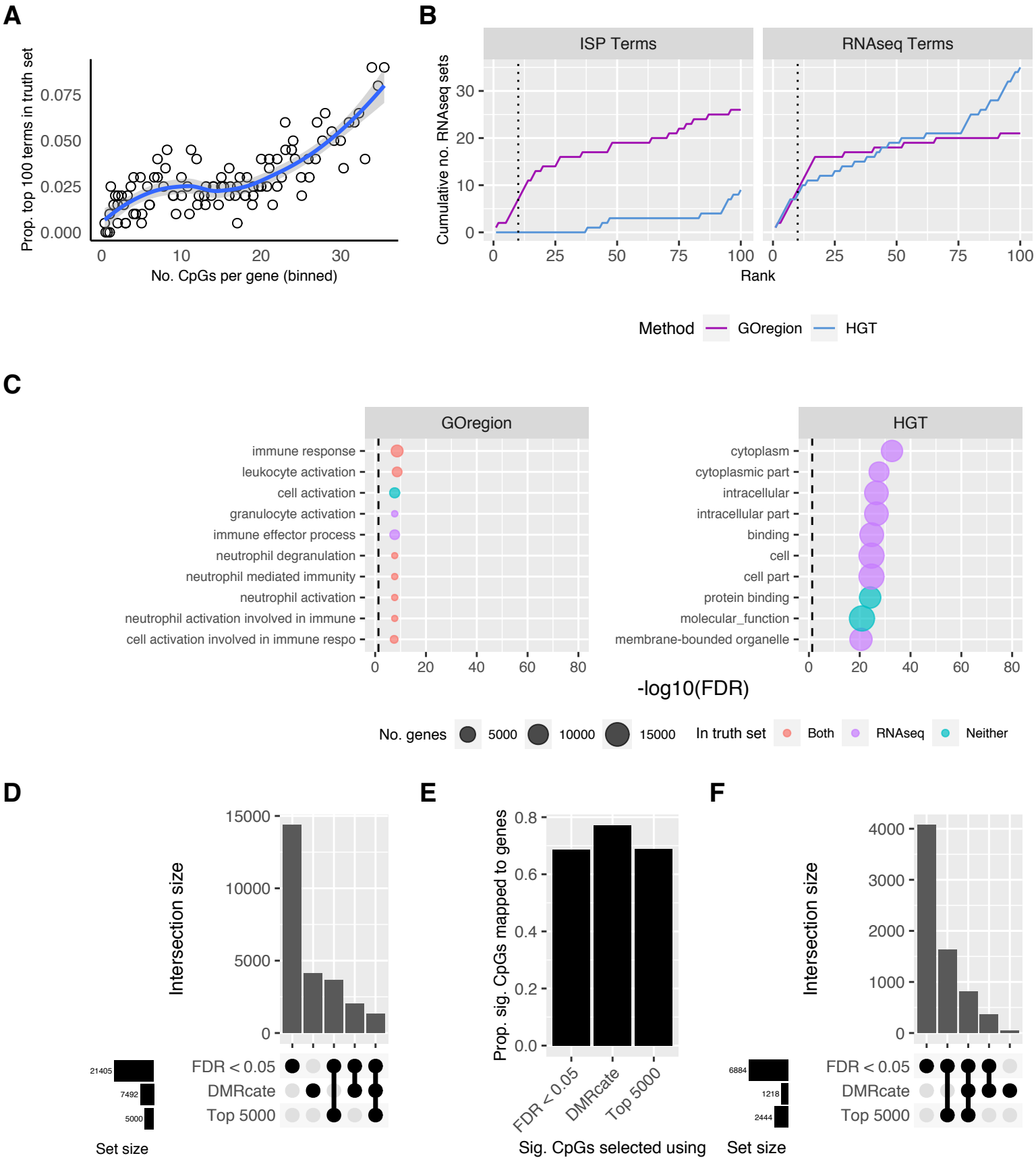
Supplementary Figure 9. Comparison of gene set testing performance for CD4 T-cells vs. CD8 T-cells (BROAD). (A) Cumulative number of BROAD MSigDB gene sets, as ranked by various methods, that are present in the RNAseq truth set for the CD4 T-cells vs. CD8 T-cells comparison. RNAseq truth = top 100 BROAD MSigDB gene sets from RNAseq analysis of the same cell types. **(B)** Bubble plots of the top 10 BROAD MSigDB gene sets as ranked by various gene set testing methods for the CD4 T-cells vs. CD8 T-cells comparison. The size of the bubble indicates the relative number of genes in the set. The colour of the bubble indicates whether the gene set is present in the RNAseq truth set (purple) or not (blue). ebGSEA (KPMT) = ebGSEA using Known Population Median Test; ebGSEA (WT) = ebGSEA using Wilcoxon Test; GOmeth = GOmeth using top 5000 most significant probes; HGT = hypergeometric test; mGLM = methylglm; mRRA (GSEA) = methylRRA using gene set enrichment analysis; mRRA (ORA) = methylRRA using over-representation analysis.


A
Prop. top 100 terms in truth set
30
20
10
0
0
25
50
75
100
Rank
Method
ebGSEA (KPMT)
ebGSEA (WT)
GOmeth
HGT
mGLM
mRRA (GSEA)
mRRA (ORA)
B
ebGSEA (KPMT)
DACOSTA_UV_RESPONSE_VIA_ERCC3_
PILON_KLF1_TARGETS_DN
DACOSTA_UV_RESPONSE_VIA_ERCC3_
CHEN_METABOLIC_SYNDROM_NETWORK
CUI_TCF21_TARGETS_2_DN
GRAESSMANN_APOPTOSIS_BY_DOXORU
ZHENG_BOUND_BY_FOXP3
BLALOCK_ALZHEIMERS_DISEASE_UP
MARTENS_BOUND_BY_PML_RARA_FUSI
RODRIGUES_THYROID_CARCINOMA_PO
ebGSEA (WT)
DACOSTA_UV_RESPONSE_VIA_ERCC3_
PILON_KLF1_TARGETS_DN
DACOSTA_UV_RESPONSE_VIA_ERCC3_
GRAESSMANN_APOPTOSIS_BY_DOXORU
CHEN_METABOLIC_SYNDROM_NETWORK
CUI_TCF21_TARGETS_2_DN
ZHENG_BOUND_BY_FOXP3
BLALOCK_ALZHEIMERS_DISEASE_UP
MARTENS_BOUND_BY_PML_RARA_FUSI
RODRIGUES_THYROID_CARCINOMA_PO
GOmeth
PILON_KLF1_TARGETS_DN
DACOSTA_UV_RESPONSE_VIA_ERCC3_
MARTENS_BOUND_BY_PML_RARA_FUSI
KRIGE_RESPONSE_TO_TOSEDOSTAT_2
KRIGE_RESPONSE_TO_TOSEDOSTAT_6
FULCHER_INFLAMMATORY_RESPONSE_
GRAESSMANN_APOPTOSIS_BY_DOXORU
ZHENG_BOUND_BY_FOXP3
DACOSTA_UV_RESPONSE_VIA_ERCC3_
IVANOVA_HEMATOPOIESIS_LATE_PRO
HGT
DACOSTA_UV_RESPONSE_VIA_ERCC3_
PILON_KLF1_TARGETS_DN
GRAESSMANN_APOPTOSIS_BY_DOXORU
DACOSTA_UV_RESPONSE_VIA_ERCC3_
MARTENS_BOUND_BY_PML_RARA_FUSI
KRIGE_RESPONSE_TO_TOSEDOSTAT_2
KRIGE_RESPONSE_TO_TOSEDOSTAT_6
ZHENG_BOUND_BY_FOXP3
CUI_TCF21_TARGETS_2_DN
CYTOPLASM
mGLM
CALCIUM_ION_TRANSMEMBRANE_TRAN
MIPS_RSMAD_COMPLEX
REACTOME_IRAK2_MEDIATED_ACTIVA
TAKEDA_TARGETS_OF_NUP98_HOXA9_
PILON_KLF1_TARGETS_DN
TAKEDA_TARGETS_OF_NUP98_HOXA9_
BIOCARTA_BAD_PATHWAY
BYSTROEM_CORRELATED_WITH_IL5_D
RUTELLA_RESPONSE_TO_CSF2RB_AND
DACOSTA_UV_RESPONSE_VIA_ERCC3_
mRRA (GSEA)
AACTTT_UNKNOWN
BIOPOLYMER_METABOLIC_PROCESS
BLALOCK_ALZHEIMERS_DISEASE_UP
BRUINS_UVC_RESPONSE_LATE
BRUINS_UVC_RESPONSE_VIA_TP53_G
CACGTG_V$MYC_Q2
CAGCTG_V$AP4_Q5
CAGGTA_V$AREB6_01
CAGGTG_V$E12_Q6
CELLULAR_MACROMOLECULE_METABOL
mRRA (ORA)
PILON_KLF1_TARGETS_DN
DACOSTA_UV_RESPONSE_VIA_ERCC3_
CUI_TCF21_TARGETS_2_DN
ZHENG_BOUND_BY_FOXP3
CHEN_METABOLIC_SYNDROM_NETWORK
DACOSTA_UV_RESPONSE_VIA_ERCC3_
NUYTTEN_NIPP1_TARGETS_DN
MASSARWEH_TAMOXIFEN_RESISTANCE
MARTENS_BOUND_BY_PML_RARA_FUSI
CREIGHTON_ENDOCRINE_THERAPY_RE
0
25
50
75
100
-log10(FDR)
No. genes
1000
2000
In truth set
RNAseq
None

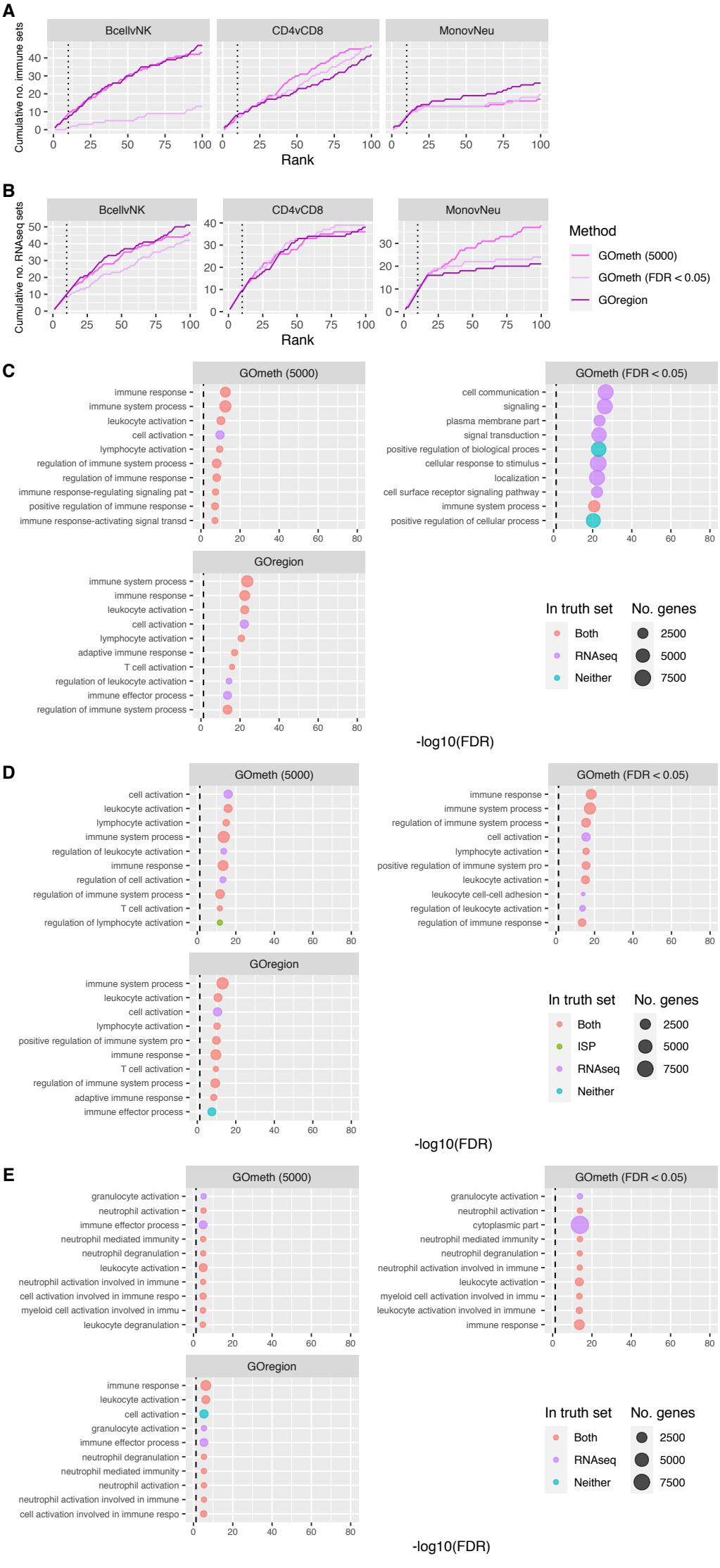
Supplementary Figure 10. Comparison of gene set testing performance for monocytes vs. neutrophils (BROAD). (A) Cumulative number of BROAD MSigDB gene sets, as ranked by various methods, that are present in the RNAseq truth set for the monocytes vs. neutrophils comparison. RNAseq truth = top 100 BROAD MSigDB gene sets from RNAseq analysis of the same cell types. **(B)** Bubble plots of the top 10 BROAD MSigDB gene sets as ranked by various gene set testing methods for the monocytes vs. neutrophils comparison. The size of the bubble indicates the relative number of genes in the set. The colour of the bubble indicates whether the gene set is present in the RNAseq truth set (purple) or not (blue). ebGSEA (KPMT) = ebGSEA using Known Population Median Test; ebGSEA (WT) = ebGSEA using Wilcoxon Test; GOmeth = GOmeth using top 5000 most significant probes; HGT = hypergeometric test; mGLM = methylglm; mRRA (GSEA) = methylRRA using gene set enrichment analysis; mRRA (ORA) = methylRRA using over-representation analysis.


A
Prop. top 100 terms in truth set
0.04
0.03
0.02
0.01
0.00
0
10
20
30
No. CpGs per gene (binned)
B
ISP Terms
RNAseq Terms
Cumulative no. RNAseq sets
50
40
30
20
10
0
0
25
50
75
100
Rank
Method
GOregion
HGT
C
GOregion
HGT
immune system process
leukocyte activation
cell activation
lymphocyte activation
positive regulation of immune system pro
immune response
T cell activation
regulation of immune system process
adaptive immune response
immune effector process
cell adhesion
biological adhesion
immune system process
binding
signaling
cell communication
cell activation
cell-cell adhesion
lymphocyte activation
leukocyte activation
0
20
40
60
80
-log10(FDR)
No. genes
5000
10000
15000
In truth set
Both
RNAseq
Neither
D
Intersection size
4000
3000
2000
1000
0
FDR < 0.05
DMRcate
Top 5000
8274
3728
5000
Set size
E
Prop. sig. CpGs mapped to genes
0.8
0.6
0.4
0.2
0.0
FDR < 0.05
DMRcate
Top 5000
Sig. CpGs selected using
F
Intersection size
1500
1000
500
0
FDR < 0.05
DMRcate
Top 5000
3305
624
2192
Set size

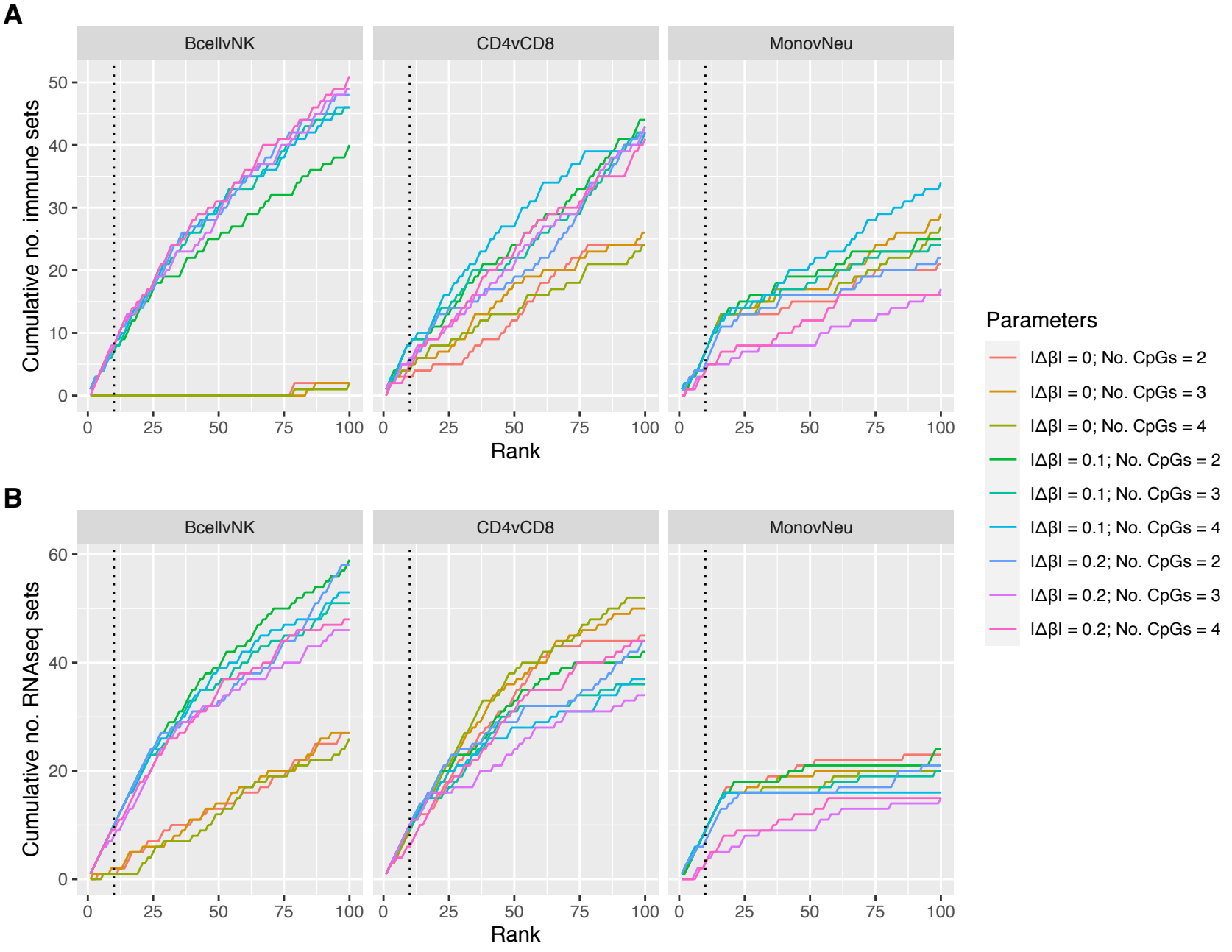
Supplementary Figure 11. Evaluation of the performance of GOregion for CD4 T-cells vs. CD8 T-cells. (A) Bias plot showing that genes that have more CpGs measuring methylation are more likely to have a differentially methylated region. This plot is produced from EPIC array sorted blood cell type data, comparing CD4 T-cells vs. CD8 T-cells. **(B)** Cumulative number of GO terms, as ranked by GOregion and a simple hypergeometric test (HGT), that are present in each truth set for the CD4 T-cells vs. CD8 T-cells comparison. ISP Terms = immune-system process child terms truth set; RNAseq Terms = top 100 terms from RNAseq analysis of the same cell types. **(D)** Bubble plots of the top 10 GO terms as ranked by GOregion and a simple HGT for the CD4 T-cells vs. CD8 T-cells comparison. The size of the bubble indicates the relative number of genes in the set. The colour of the bubble indicates whether the term is present in either RNAseq (purple) or ISP (green) truth sets, both (red) or neither (blue). **(E)** Upset plot showing the characteristics of the CpGs selected as "significant" for the CD4 T-cells vs. CD8 T-cells comparison by a probe-wise differential methylation analysis using a significance cut off (FDR < 0.05), the top 5000 CpGs as ranked by the probe-wise analysis (Top 5000) or the CpGs underlying the filtered *DMRcate* regions (DMRcate). The probe-wise analysis with FDR < 0.05 identified over 8,000 CpGs as "significant" and had the most unique CpGs. Although DMRcate identified the fewest "significant" CpGs (~3,700), over 2,000 were unique to that approach. **(F)** Proportion of "significant" CpGs that are annotated to genes as identified by the three different strategies. **(G)** Upset plot showing the characteristics of the genes that "significant" CpGs are annotated to, as identified by the three different strategies, for the CD4 T-cells vs. CD8 T-cells comparison. CpGs identified by the probe-wise analysis with FDR < 0.05 map to over 3,000 genes. The CpGs identified by DMRcate mapped to ~600, several of which are unique to this approach.


A
Prop. top 100 terms in truth set
0.075
0.050
0.025
0.000
0
10
20
30
No. CpGs per gene (binned)
B
ISP Terms
RNAseq Terms
Cumulative no. RNAseq sets
30
20
10
0
0
25
50
75
100
Rank
Method
GOregion
HGT
C
GOregion
HGT
immune response
leukocyte activation
cell activation
granulocyte activation
immune effector process
neutrophil degranulation
neutrophil mediated immunity
neutrophil activation
neutrophil activation involved in immune
cell activation involved in immune respo
cytoplasm
cytoplasmic part
intracellular
intracellular part
binding
cell
cell part
protein binding
molecular_function
membrane-bounded organelle
0
20
40
60
80
-log10(FDR)
No. genes
5000
10000
15000
In truth set
Both
RNAseq
Neither
D
Intersection size
15000
10000
5000
0
FDR < 0.05
DMRcate
Top 5000
21405
7492
5000
Set size
E
Prop. sig. CpGs mapped to genes
0.8
0.6
0.4
0.2
0.0
FDR < 0.05
DMRcate
Top 5000
Sig. CpGs selected using
F
Intersection size
4000
3000
2000
1000
0
FDR < 0.05
DMRcate
Top 5000
6884
1218
2444
Set size

Supplementary Figure 12. Evaluation of the performance of GOregion for monocytes vs. neutrophils. (A) Bias plot showing that genes that have more CpGs measuring methylation are more likely to have a differentially methylated region. This plot is produced from EPIC array sorted blood cell type data, comparing monocytes vs. neutrophils. **(B)** Cumulative number of GO terms, as ranked by GOregion and a simple hypergeometric test (HGT), that are present in each truth set for the monocytes vs. neutrophils comparison. ISP Terms = immune-system process child terms truth set; RNAseq Terms = top 100 terms from RNAseq analysis of the same cell types. **(D)** Bubble plots of the top 10 GO terms as ranked by GOregion and a simple HGT for the monocytes vs. neutrophils comparison. The size of the bubble indicates the relative number of genes in the set. The colour of the bubble indicates whether the term is present in either RNAseq (purple) or ISP (green) truth sets, both (red) or neither (blue). **(E)** Upset plot showing the characteristics of the CpGs selected as "significant" for the monocytes vs. neutrophils comparison by a probe-wise differential methylation analysis using a significance cut off (FDR < 0.05), the top 5000 CpGs as ranked by the probe-wise analysis (Top 5000) or the CpGs underlying the filtered *DMRcate* regions (DMRcate). The probe-wise analysis with FDR < 0.05 identified over 21,000 CpGs as "significant" and had the most unique CpGs. Although DMRcate identified fewer CpGs (~7500), ~3500 of them were unique to that approach. **(F)** Proportion of "significant" CpGs that are annotated to genes as identified by the three different strategies. **(G)** Upset plot showing the characteristics of the genes that "significant" CpGs are annotated to, as identified by the three different strategies, for the monocytes vs. neutrophils comparison. CpGs identified by the probe-wise analysis with FDR < 0.05 map to over 6,000 genes. The CpGs identified by DMRcate mapped to ~1200, several of which are unique to this approach.


A
BcellvNK
CD4vCD8
MonovNeu
Cumulative no. immune sets
Rank
B
Cumulative no. RNAseq sets
Method
GOmeth (5000)
GOmeth (FDR < 0.05)
GOregion
C
GOmeth (5000)
immune response
immune system process
leukocyte activation
cell activation
lymphocyte activation
regulation of immune system process
regulation of immune response
immune response-regulating signaling pat
positive regulation of immune response
immune response-activating signal transd
GOmeth (FDR < 0.05)
cell communication
signaling
plasma membrane part
signal transduction
positive regulation of biological proces
cellular response to stimulus
localization
cell surface receptor signaling pathway
immune system process
positive regulation of cellular process
GOregion
immune system process
immune response
leukocyte activation
cell activation
lymphocyte activation
adaptive immune response
T cell activation
regulation of leukocyte activation
immune effector process
regulation of immune system process
In truth set
Both
RNAseq
Neither
No. genes
2500
5000
7500
-log10(FDR)
D
GOmeth (5000)
cell activation
leukocyte activation
lymphocyte activation
immune system process
regulation of leukocyte activation
immune response
regulation of cell activation
regulation of immune system process
T cell activation
regulation of lymphocyte activation
GOmeth (FDR < 0.05)
immune response
immune system process
regulation of immune system process
cell activation
lymphocyte activation
positive regulation of immune system pro
leukocyte activation
leukocyte cell-cell adhesion
regulation of leukocyte activation
regulation of immune response
GOregion
immune system process
leukocyte activation
cell activation
lymphocyte activation
positive regulation of immune system pro
immune response
T cell activation
regulation of immune system process
adaptive immune response
immune effector process
In truth set
Both
ISP
RNAseq
Neither
No. genes
2500
5000
7500
-log10(FDR)
E
GOmeth (5000)
granulocyte activation
neutrophil activation
immune effector process
neutrophil mediated immunity
neutrophil degranulation
leukocyte activation
neutrophil activation involved in immune
cell activation involved in immune respo
myeloid cell activation involved in immu
leukocyte degranulation
GOmeth (FDR < 0.05)
granulocyte activation
neutrophil activation
cytoplasmic part
neutrophil mediated immunity
neutrophil degranulation
neutrophil activation involved in immune
leukocyte activation
myeloid cell activation involved in immu
leukocyte activation involved in immune
immune response
GOregion
immune response
leukocyte activation
cell activation
granulocyte activation
immune effector process
neutrophil degranulation
neutrophil mediated immunity
neutrophil activation
neutrophil activation involved in immune
cell activation involved in immune respo
In truth set
Both
RNAseq
Neither
No. genes
2500
5000
7500
-log10(FDR)

Supplementary Figure 13. Comparison of probe-wise and region-wise analyses in the gene set testing context. (A) Cumulative number of GO terms, as ranked by various analysis strategies, that are present in the immune truth set for all comparisons. ISP Terms = immune-system process child terms truth set. **(B)** Cumulative number of GO terms, as ranked by various methods, that are present in the RNAseq truth set for all comparisons. RNAseq Terms = top 100 terms from RNAseq analysis of the same cell types. **(C)** Bubble plots of the top 10 GO terms as ranked by various analysis strategies for the B-cell vs. NK comparison. The size of the bubble indicates the relative number of genes in the set. The colour of the bubble indicates whether the term is present in either RNAseq (purple) or ISP (green) truth sets, both (red) or neither (blue). **(D)** Bubble plots of the top 10 GO terms as ranked by various analysis strategies for the CD4 T-cells vs. CD8 T-cells comparison. **(E)** Bubble plots of the top 10 GO terms as ranked by various analysis strategies for the monocytes vs. neutrophils comparison. GOmeth (5000) = GOmeth using top 5000 most significant probes; GOmeth (FDR < 0.05) = GOmeth using all probes significant at FDR < 0.05; GOregion = GOregion using all selected region coordinates.


A
BcellvNK
CD4vCD8
MonovNeu
Cumulative no. immune sets
0
10
20
30
40
50
0
25
50
75
100
Rank
B
BcellvNK
CD4vCD8
MonovNeu
Cumulative no. RNAseq sets
0
20
40
60
0
25
50
75
100
Rank
Parameters
|Δβ| = 0; No. CpGs = 2
|Δβ| = 0; No. CpGs = 3
|Δβ| = 0; No. CpGs = 4
|Δβ| = 0.1; No. CpGs = 2
|Δβ| = 0.1; No. CpGs = 3
|Δβ| = 0.1; No. CpGs = 4
|Δβ| = 0.2; No. CpGs = 2
|Δβ| = 0.2; No. CpGs = 3
|Δβ| = 0.2; No. CpGs = 4

Supplementary Figure 14. Effect of DMR filtering on gene set testing. (A) Cumulative number of GO terms, as ranked by GOregion analysis of DMRs filtered using various parameters, that are present in the immune truth set for all comparisons. ISP Terms = immune-system process child terms truth set. **(B)** Cumulative number of GO terms, as ranked by GOregion analysis of DMRs filtered using various parameters, that are present in the RNAseq truth set for all comparisons. RNAseq Terms = top 100 terms from RNAseq analysis of the same cell types. $|\Delta\beta|$ = mean methylation difference across region; No. CpGs = number of CpGs underlying region.