SUPPLEMENTAL INFORMATION

Variability in DNA methylation defines novel epigenetic subgroups of DLBCL associated with different clinical outcomes.

Nyasha Chambwe[1,2,3], Matthias Kormaksson[4,5], Huimin Geng[6], Subhajyoti De[7,8,9], Franziska Michor[10,11], Nathalie A. Johnson[12], Ryan D. Morin[13,14], David W. Scott[15], Lucy A. Godley[16], Randy D. Gascoyne[15,17], Ari Melnick[18,19], Fabien Campagne[1,2*] and Rita Shaknovich[19,20*]

[1]The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine; [2]Department of Physiology and Biophysics, Weill Cornell Medical College, New York, NY; [3]Tri-Instituitional Training Program in Computational Biology and Medicine, Weill Cornell Medical College, New York, NY; [4]Department of Public Health, Weill Cornell Medical Center, New York, NY; [5]IBM Research-Brazil, Rio de Janeiro, Brazil; [6]Department of Laboratory Medicine, University of California San Francisco; [7]Department of Medicine, University of Colorado School of Medicine, Aurora, CO; [8]Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO; [9]Molecular Oncology Program, University of Colorado Cancer Center; Aurora, CO; [10]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA; [11]Department of Biostatistics, Harvard School of Public Health, Boston, MA; [12]Department of Medicine, Jewish General Hospital, Montreal, Canada; [13]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, Canada; [14]Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, Canada; [15]Centre for Lymphoid Cancer, British Columbia Cancer Agency, Vancouver, Canada; [16]Department of Medicine, The University of Chicago, Chicago, IL; [17]Department of Pathology, University of British Columbia, Vancouver, British Columbia, Canada; [18]Department of Pharmacology, Weill Cornell Medical College, New York, NY; [19]Division of Hematology and Oncology, Weill Cornell Medical College, New York, NY; [20]Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, New York, NY

Table of Contents

Supplemental Materials and Methods

Sample Collection

140 diagnostic *de novo* DLBCL samples were collected from individuals that presented with de novo DLBCL at the British Columbia Cancer Agency (BCCA), Canada. Supplemental Table 1 presents detailed clinical and phenotypic characteristics of the study cohort. Normal Germinal Center B cells (NGCB) were obtained from leftover human tonsils after routine tonsillectomies performed at New York Presbyterian Hospital. All tissue collection was approved by the Weill Cornell Medical College Institutional Review Board and in accordance with the stipulations of the Helsinki treaties. Mononuclear cells were isolated using Histopaque density centrifugation. All washes were performed in PBS/ 2% Bovine Serum Albumine/ 2% EDTA. All antibodies were used at 1:100 dilution in cold PBS and staining was done for 10 min on ice, followed by 3 washes. B cell populations were separated using the AutoMACS system (Milteny Biotec, Auburn, CA) "posselD" program. Briefly, NGCB cells were separated by positive selection with CD77 (anti-CD77: Ab Serotec cat# MCA579 Batch 180510).

DNA Extraction and HELP Assay

Genomic DNA was extracted using the Qiagen Puregene Gentra cell kit (Qiagen, Valencia, CA). High molecular weight DNA was diluted in water and the quality was assessed using 1% agarose gel. The HELP assay was performed using our standard protocol[2]: 1 μg of genomic DNA was digested with HpaII and MspI (NEB, Ipswich, MA), adapters were ligated using T4 DNA Ligase followed by PCR amplification and labeling of HpaII and MspI digestion products. The PCR products were co-hybridized to custom NimbleGen HELP microarrays (NimbleGen, Inc. Madison, WI). **The Roche Nimblegen HG17 HELP array design used in this study interrogates 50,000 CpGs from 25,626 HpaII amplifiable fragments of ~14,000 genes[1,2]. The ~14000 genes are each represented with ten oligonucleotide probes (total 385,000 features), along with 2,000 random sequence probe controls as well as mitochondrial DNA probes (mitochondrial DNA is never methylated and is present at high copy numbers so that both HpaII and MspI fluorescence intensities are high and equal). There was no a priori selection of the interrogated genes, but criteria for selection of loci included: HpaII/MspI sites in the genome within 50-2000bp of each other and the ability to design uniquely mapping probes to those fragments[9].** The microarray design is documented in the Gene Expression Omnibus (GEO) Accession GPL6604. Data from this study is publicly available by accessing GEO accession GSE23967.

HELP Data Analysis

HELP data was processed using standard pipeline as outlined in the HELP analysis package[3] from the R Bioconductor suite. Probes with signal intensity less than 2.5 mean absolute deviation (MAD) were classified as failed and discarded from analysis. Intra- and inter-array normalization was performed by first subtracting the mean random probe intensity separately within the HpaII and MspI channels. Each channel was quantile normalized independently. Channel quantile normalized intensities were used to derive

the HELP log ratio, log(HpaII/MspeI), which was used for all subsequent analyses. Additional information can be found in

Quantifying Methylation Disruption

We derive a measure of methylation disruption in DLBCL in the following way (supplemental Figure 1) let $y_{ij} = log_2(HpaII/MspI)_{ij}$ denote the HELP Methylation log ratio for sample *i* at HELP fragment *j*. Further, define $z_j = \overline{log_2(HpaII/MspI)}_{.j}$ as the average methylation log ratio at HELP fragment *j* averaged across the 10 normal germinal center B cell (NGCB) control samples. We finally define $x_{ij} = y_{ij} - z_j$ as the methylation difference between sample *i* and the average NGCB methylation at probe set *j*. The methylation variability profile for sample *i* (MVP_i) is defined as the density function $f_i(x)$ of these differences $\left(x_{ij's}\right)$ across all loci represented on the array. We estimated the function $f_i(x)$ using the density() function in R[10] with bandwidth parameter 0.1.

We define the Methylation Variability Score (MVS) of sample i as the deviation of the sample's MVP to that of the expected MVP of an NGCB sample. More specifically, let $f_i(x)$ denote the MVP of patient i and let $g_1(x), \ldots, g_{10}(x)$ denote the MVPs of the 10 GCB samples. Then we define the Methylation Variability Score for patient i as

$MVS_i = \int [f_i(x) - \bar{g}(x)]^2 dx$ where $\bar{g}(x) = \frac{1}{10}\sum_{i=1}^{10} g_i(x)$.

Functional Clustering

To cluster DLBCLs based on their MVPs, we adapted an approach to cluster continuous data described by Ferreira et al.[11]. First we calculated the squared L_2-distance between two MVP functions $f_i(x)$ and $f_{i'}(x)$ for all pairs of patient samples (*i*,*i'*):

$$d(i,i') = \int [f_i(x) - f_{i'}(x)]^2 dx$$

This distance represents the squared difference in the area under the curve between two samples and is approximated using the Trapezoidal rule[11]. We perform unsupervised hierarchical clustering on the distance matrix of all pairwise L_2 distances using the Ward's hierarchical clustering in the base stats package of R[10].

Consensus Clustering

To determine the number of clusters in our study we performed consensus clustering using the same parameters that we used for our functional clustering. We used the L2 distance and hierarchical clustering with Ward's agglomeration method. We performed hierarchical clustering 1000 times on resampled subsets of the 140 samples (using 80% of samples as subset) and cut the dendogram at cluster numbers k=2,3……15. We note

that the plot of area under CDF change started plateau at K=6 as it was the smallest number that separated the 3 outlier MVPs into one distinct cluster.

Single locus quantitative DNA methylation assays

EpiTYPER assays (Sequenom, CA) were performed on bisulfite-converted DNA. EpiTYPER primers were designed so that the amplicons covered selected HpaII Amplifiable Fragments (HAF), as well as any other HpaII sites found up to 2kb upstream of the downstream site and up to 2kb downstream of the upstream site, in order to cover all possible alternative sites of digestion. Five randomly selected high variance genes (*p53AiP1, S100A9, B2M, CSF2, TREML2)* in 8 randomly selected DLBCL cases were epityped. MassARRAY and HELP showed high correlation (r^2=0.70 Supplemental Figure 5), indicating that change in log2 (HpaII/Msp1) HELP values of 1 is approximately equivalent to a 20% change in methylation. For technical validation primers were designed to cover genomic loci associated with the interrogated HAFs of interest. The primers were designed using Sequenom EpiDesigner beta software (http://www.epidesigner.com/). The primer sequences are available in supplemental Table 7.

Functional Enrichment Analysis

GO ontology enrichment was assessed using the DAVID Bioinformatics Resource[12,13]. We report enrichment of DAVID's pruned GO_FAT biological processes. GO process results are visualized using REVIGO treemap representation[14]. REVIGO prunes semantically similar terms and nominate a representative term for a cluster of similar terms. For comprehensiveness, we carried out pathway analysis for each gene signature using MetaCore from Thomson Reuters and Ingenuity Pathway Analysis (IPA) (Ingenuity® Systems, Redwood City, CA, www.ingenuity.com). We used the full set of genes represented on the array as a background gene list for enrichment testing. We used Ingenuity Pathway Analysis software (IPA) to identify molecular networks enriched for differentially methylated genes.

We assessed the significance of clinical and phenotypic class enrichment in the clusters using Fisher's exact test. We carried out enrichment analysis for each DNA methylation based cluster and experimentally derived targets of EZH2 from a previous ChIP-chip study in B cell[15]. We mapped the HELP and ChIP-chip probes to genes and consider the intersection set as the background for enrichment. We defined EZH2 targets as genes that had significant peaks called from the ChIP-chip experiment[15] and present in the HELP-ChIP intersection set. We carried out overrepresentation analysis using the ORA mode of GeneTrail[16]. P-values were calculated using the hypergeometric test and corrected for multiple testing using the Benjamini-Hochberg correction[13].

Gene Expression Profiling

Gene expression data was obtained from previous studies[5,17] for 52 DLBCL samples profiled for methylation in this study (GEO Accession: GSE23501) and 4 normal tonsil germinal center B cell samples (GEO accession: GSE15271). RNA extracted and purified from these samples was hybridized onto the Affymetrix chip (HG U133 plus 2.0). Raw (.CEL) files were downloaded from GEO, and processed together using the

Robust Multi Chip Average (RMA) method to derive log2 expression intensity for each probe[18]. RefSeq Custom CDF (version 15) was used to collapse probe intensities into a single value for each annotated RefSeq gene[19]. Differential expression analysis was carried out using a moderated t-test (limma package in R)[20]. Benjamini-Hochberg false discovery rate correction was applied to the p values for this test. We considered a gene significant if the adjusted p value was less than 0.05 and the magnitude of the log fold change |logFC| >=1.0, a two fold difference. Additionally 43/52 DLBCL samples and 19 flow-sorted centroblasts were also assayed by RNA-Seq.

Integrative analysis of methylation and expression

We used the results from the respective differential methylation and expression analysis to determine the association between DNA methylation and gene expression of specific genes. Methylation and expression data were integrated by performing a table "join" operation on RefSeq transcripts IDs using JMP (Version 10. SAS Institute Inc., Cary, NC, 1989-2007). Considering only genes covered on both the HELP platform and the Affymetrix HG133plus2 array, we counted how many RefSeq transcripts showed inverse correlation between expression and methylation as determined using adjusted p values and fold change for the limma tests for differences in average methylation and expression between normal and DLBCL clusters.

Supplemental Results

DLBLCs have a core set of functionally important aberrantly methylated genes

In order to understand which epigenetic events are common to all DLBCLs compared to NGCBs, we determined the fragments that are significantly differentially methylated between NGCBs and all DLBCLs studied. We found 157 fragments (200 genes) that were significantly differentially methylated between DLBLCs and NGCBs (supplemental Figure 14A, supplemental Table 2). 78 genes were hypermethylated in DLBCL compared to NGCB and 122 genes were hypomethylated in DLBCL relative to NGCB (supplemental table 2). The most significantly hypermethylated genes include *RGS22*, *BBS10*, *NID1*, *CDKN2B*-AS1, *SMARCA2*, and *SUSD5* while hypomethylated genes include FAM110B, *NKG7*, *IKZF4*, *ETFB*, *CLDND2* and *PEG3*. Cell adhesion molecules, such as the protocadherin gamma subfamily (*PCDHGA**, *PCDHGB**) and the cadherin-associated protein *CTNNA2*, are also commonly hypermethylated in most DLBCLs. The top network identified using Ingenuity Pathway analysis network algorithm, contains a set of genes involved in cell-mediated immune response *(CD3D, CD3G, CCR6, CCL17* and *STAT3*, supplemental Figure 14B). This network is also enriched in genes involved in cell differentiation and migration such as *ERRB3, HBEGF, BTG2, HOXB1* and *POU5F1 (OCT4)* (hypomethylated) and *STAT3* (hypermethylated). Integrative analysis of gene expression and methylation found 8 genes showing an inverse correlation between expression and methylation (supplemental Figure 14B). 1 gene, *UBE2J1* was hypermethylated and downregulated in DLBCLs compared to NGCBs (Figure 6B). 7 genes, *CD3D*, *VSTM3*, *NMB*, *FXYD2*, *GZMK*, *CALD1* and *RHOBTB3* were hypomethylated and up-regulated in DLBCLs (Figure 6B). Additionally, the inverse relationship was confirmed for expression using RNA-Seq data in a subset of cases, for example the over-expression of *CD3D*, *GZMK*, *VSTM3* and *CALD1* (supplemental Figure 14C).

Biological pathways affected by DNA methylation changes

We asked which biological pathways were represented in the genes that compose the different DLBCL cluster signatures. The following sections describe the biological process ontology, pathways and networks found over-represented in each cluster signature.

Cluster A: With only 49 probesets corresponding to 38 genes, the signature for Cluster A contains key molecules involved in B cell differentiation and in immune response, particularly immune signaling (supplemental Figure 7). Of particular interest is the hypermethylation of cytokine mediated signaling pathway genes STAT3, TNFRSF1A and KRAS. Other genes involved in cell surface receptor signaling such as CD2, CD3D, CD3G, NMB, DTX1, CCR6 and CD274 are differentially methylated in Cluster A. Ingenuity pathway analysis reveals that the top biological function in cluster A is inflammatory response and one of the top networks contains CCR6, CD274 and STAT3 molecules. Cytokine-mediated signaling also is detected as a GO Biological process. Thus cluster A reveals epigenetic deregulation of key molecules involved in immune response and also interaction with microenvironment.

Cluster B: The signature was enriched in genes contributing to multicellular organismal homeostasis, but no more specific pathway was detected after adjusting for multiple tests (supplemental Figure 8).

Cluster C: An Ingenuity analysis suggests a deregulation of a network of genes interacting with DLX5, a homeobox transcription factor (Supplemental Figure 9). Genes in this network are primarily involved in embryonic and organ development and in tissue specification. Cluster C is also characterized by hypermethylation of many developmental transcription factors: of note many members of homeobox gene family (*HOXA10-A9 HOXD8, SATB2, TLX3, ESX1, POU3F4, MSX1* (hyper) and HOXB1 (hypo)) and forkhead box family genes (*FOXA1, FOXA2, FOXF2, FOXG1, FOXL1, FOXQ1*). Other key cell fate commitment cell differentiation genes include hypermethylated *WNT2, STAT3, SOX11, POU3F4* and *GDNF*. IPA top canonical pathways include IL-9 signaling and signaling through JAK1 and JAK3. Aberrantly methylated genes HNFalpha/FOXA1, HNFbeta/FOXa2 and PCK1 play a role in the in regulation of gluconeogenesis and may reflect changing metabolic requirements in neoplastic cells.

Cluster D: We found that the tricarbonic acid (TCA) cycle is one of the top canonical pathways in cluster D. Of note, IDH2 belongs to this pathway is significantly hypomethylated in clusters D and E, and F. IDH1 and IDH2 mutations in AML are associated with hypermethylation[21]. Cluster D also contains aberrantly methylated genes involved in cell adhesion, particularly proto-cadherins, as well as WNT signaling genes such as CTBP2, SMARCA2, SMARCAL1, CTTNA2, WNT2 WNT2B and WNT8A (supplemental Figure 10).

Cluster E: A unique feature of Cluster E is the aberrant methylation of Ephrin signaling genes characterized by the hypermethylation of EPHA5 and PIK3CG, and the

hypomethylation EPHB1, the tyrosine-protein kinase FYN, GRB7, GNAO1, PXN and ephexin (Supplemental Figure 11).

Clusters D and E: Recent reports indicate that the epigenetic dysregulation of JMJD4 in DLBCLs may perturb the balance between inhibitory DNA methylation marks and H3K27Me marks. Both clusters D and E revealed hypomethylation of JMJD4. Hypomethylation of IDH2 and JMJD4 did not seem to have a significant effect on gene expression in this cohort.

Cluster F: the signature of Cluster F is the largest, with over 7,000 genes differentially methylated from NGCB controls. Ingenuity network analysis showed that the top deregulated network included genes involved in cellular growth and proliferation, hematological system development and function and the inflammatory response centered on hypomethylated IL-4 (Supplemental Figure 12). Most processes that contribute to a malignant phenotype are enriched in this cluster such as regulation of apoptotic processes, aberrant methylation of cell cycle genes and those that regulate them, as well as most signal transduction pathways associated with cancer (AKT signaling, inhibition of ERK, or AMPK signaling).

Supplemental References

1. Khulan B, Thompson RF, Ye K, et al. Comparative isoschizomer profiling of cytosine methylation: the help assay. *Genome Res.* 2006;16(8):1046–55.

2. Shaknovich R, Figueroa ME, Melnick A. Help (hpaii tiny fragment enrichment by ligation-mediated pcr) assay for dna methylation profiling of primary normal and malignant b lymphocytes. *Methods Mol. Biol.* 2010;632:191–201.

3. Thompson RF, Reimers M, Khulan B, et al. An analytical pipeline for genomic representations used for cytosine methylation studies. *Bioinformatics*. 2008;24(9):1161–7.

4. Figueroa ME, Lugthart S, Li Y, et al. Dna methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer Cell*. 2010;17(1):13–27.

5. Shaknovich R, Geng H, Johnson NA, et al. Dna methylation signatures define molecular subtypes of diffuse large b-cell lymphoma. *Blood*. 2010;116(20):e81–9.

6. Nischal S, Bhattacharyya S, Christopeit M, et al. Methylome profiling reveals distinct alterations in phenotypic and mutational subgroups of myeloproliferative neoplasms. *Cancer Res.* 2013;73(3):1076–85.

7. Heuck CJ, Mehta J, Bhagat T, et al. Myeloma is characterized by stage-specific alterations in dna methylation that occur early during myelomagenesis. *J. Immunol.* 2013;190(6):2966–75.

8. Leshchenko V V, Kuo PY, Shaknovich R, et al. Genomewide dna methylation analysis reveals novel targets for drug development in mantle cell lymphoma. *Blood*. 2010;116:1025–1034.

9. Suzuki M, Greally JM. Dna methylation profiling using hpaii tiny fragment enrichment by ligation-mediated pcr (help). *Methods*. 2010;52(3):218–22.

10. R Development Core Team, Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2006.

11. Ferreira L, Hitchcock DB. A comparison of hierarchical methods for clustering functional data. *Commun. Stat. - Simul. Comput.* 2009;38(9):1925–1949.

12. Huang DW, Sherman BT, Lempicki R a. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat. Protoc.* 2009;4(1):44–57.

13. Huang DW, Sherman BT, Lempicki R a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009;37(1):1–13.

14. Supek F, Bošnjak M, Škunca N, Šmuc T. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS One*. 2011;6(7):e21800.

15. Velichutina I, Shaknovich R, Geng H, et al. Ezh2-mediated epigenetic silencing in germinal center b cells contributes to proliferation and lymphomagenesis. *Blood*. 2010;116(24):5247–55.

16. Backes C, Keller A, Kuentzer J, et al. Genetrail--advanced gene set enrichment analysis. *Nucleic Acids Res.* 2007;35(Web Server issue):W186–92.

17. Caron G, Le Gallou S, Lamy T, Tarte K, Fest T. Cxcr4 expression functionally discriminates centroblasts versus centrocytes within human germinal center b cells. *J. Immunol.* 2009;182(12):7595–602.

18. Irizarry R a, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*. 2003;4(2):249–64.

19. Dai M, Wang P, Boyd AD, et al. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Res.* 2005;33(20):e175.

20. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3:Article3.

21. Figueroa ME, Abdel-Wahab O, Lu C, et al. Leukemic idh1 and idh2 mutations result in a hypermethylation phenotype, disrupt tet2 function, and impair hematopoietic differentiation. *Cancer Cell*. 2010;18(6):553–67.

Supplemental Tables

Supplemental Table 1. (.xls) Detailed clinical and phenotypic characteristics of patient cohort

Supplemental Table 2. (.xls) Cluster Signatures

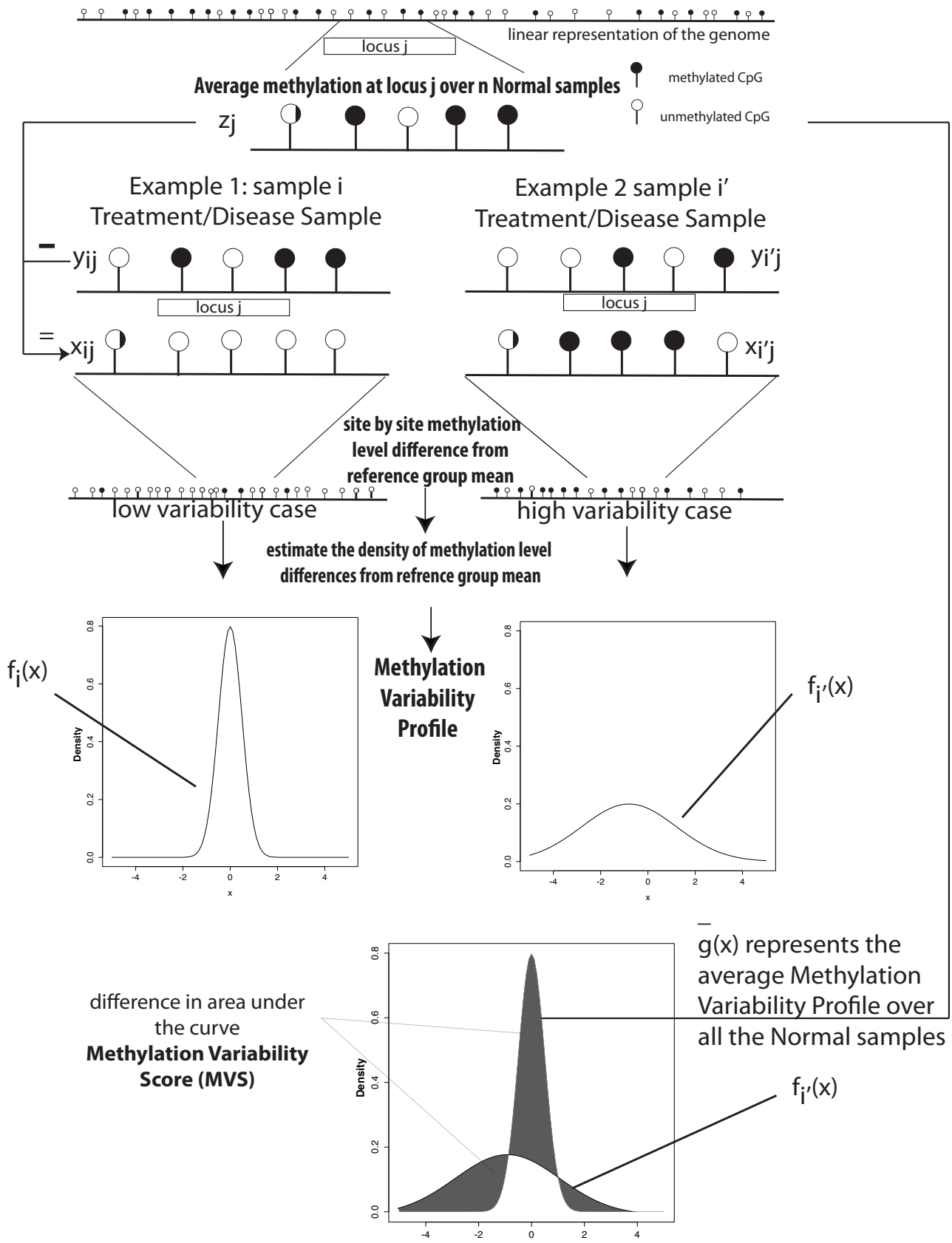
Supplemental Table 3. (.xls) Aberrantly methylated EZH2 target genes

Supplemental Table 4. (.xls) Broad amplification and deletion regions called by the GISTIC algorithm

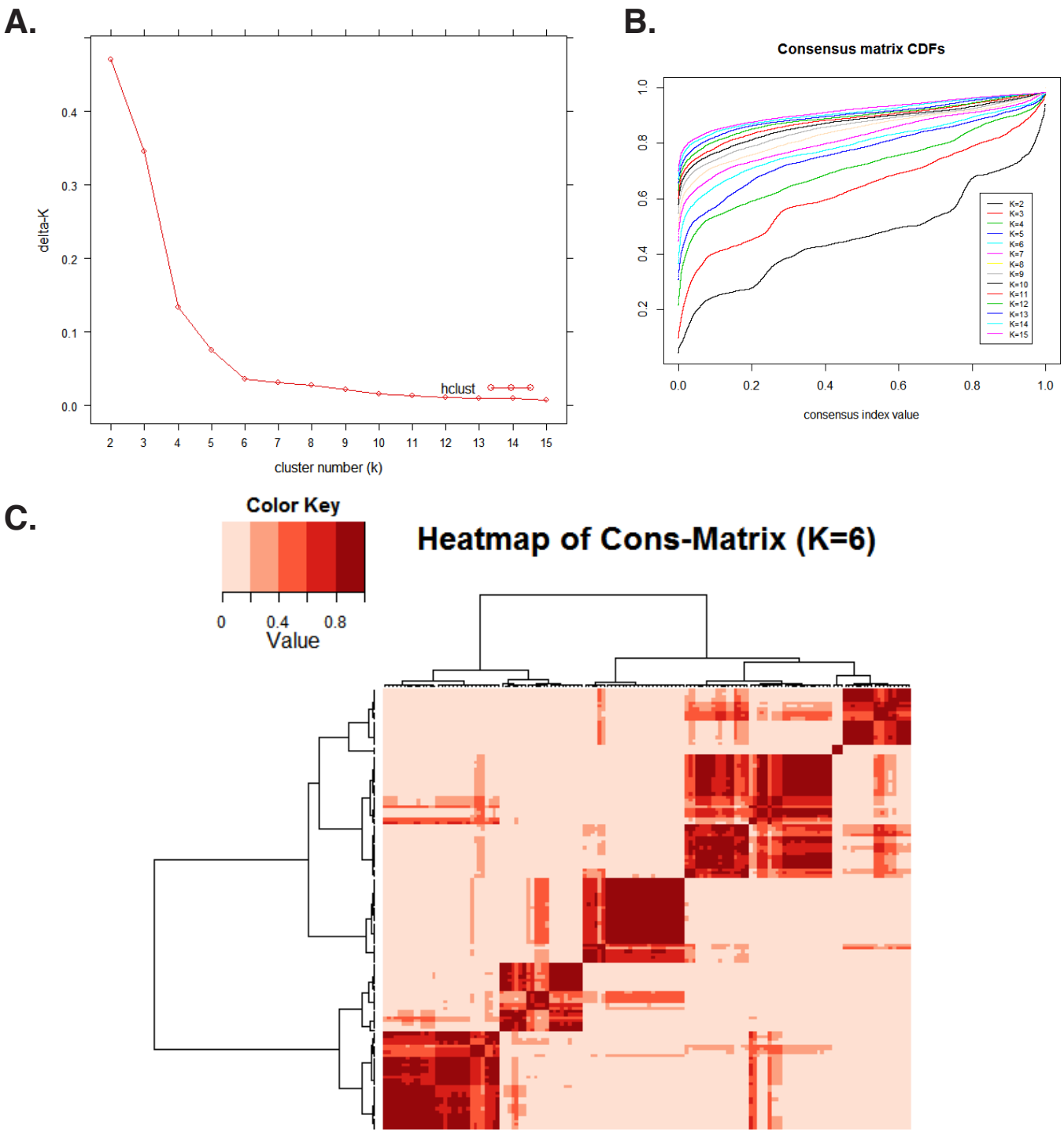
Supplemental Table 5. (.xls) Summary of methylation and expression inversely correlated RefSeq transcripts

Supplemental Table 6. (.xls) Genes with an inverse relationship between methylation and expression.

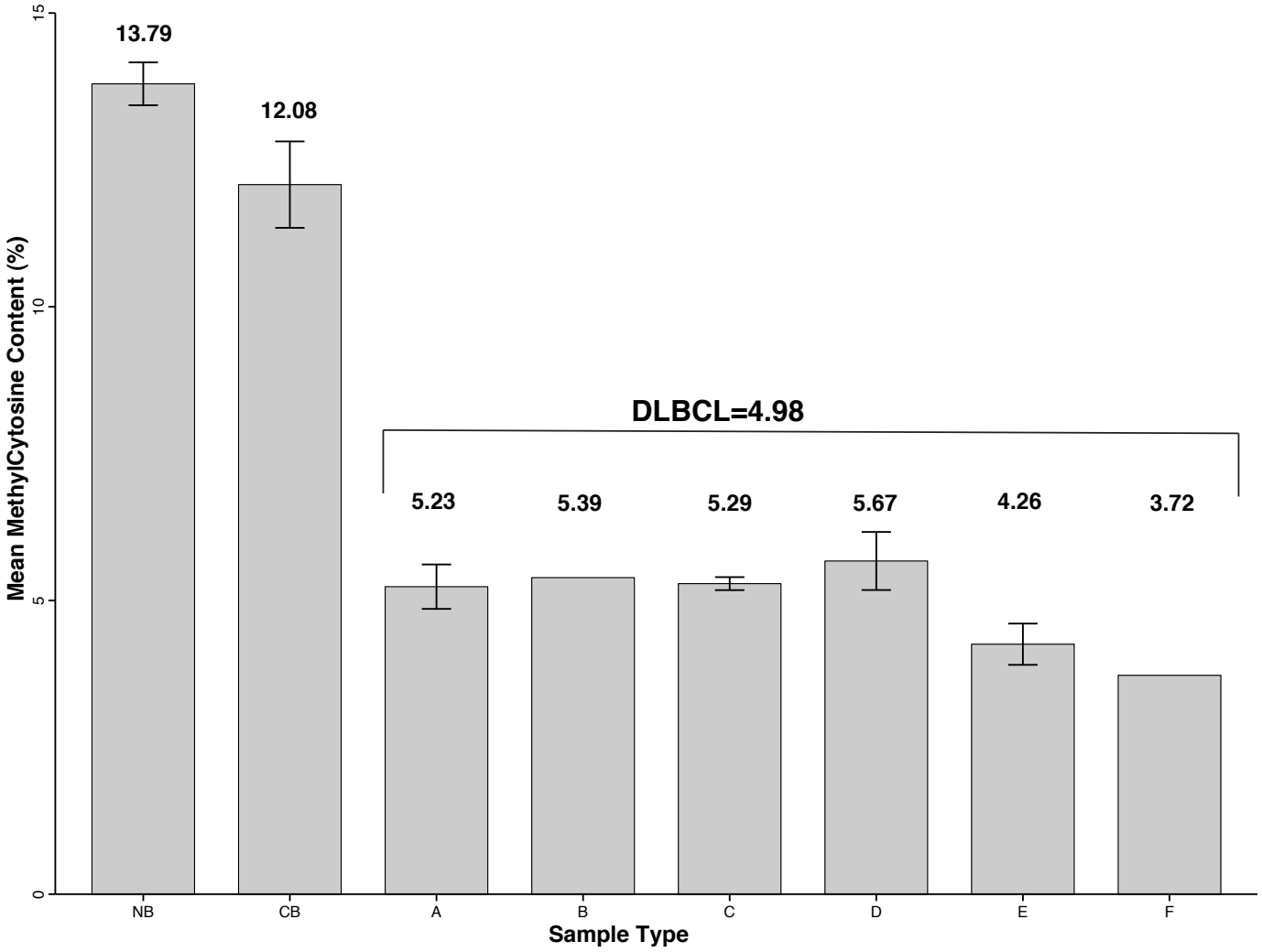
Supplemental Table 7. (.xls) Mass array primers.



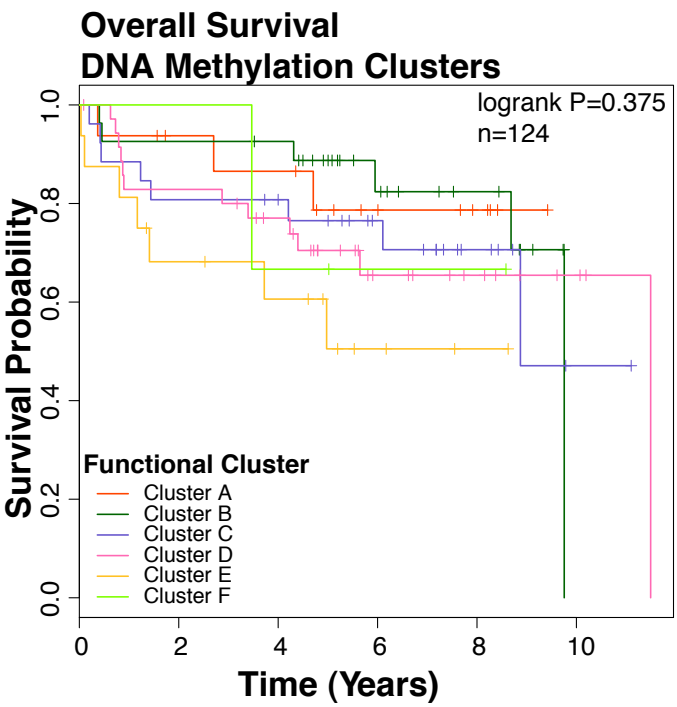
Supplemental Figure 1. Approach to measure methylation variability. The average methylation across normal/control samples is calculated for all loci covered by this array platform. For each disease/treatment sample, the difference in methylation at each particular locus is calculated. The density function (histogram) that describes differences from normal is termed the methylation variability profile (MVP). The methylation variability score (MVS) is the estimate of the area under the density curve between a given sample MVP and the average normal MVP (calculated using the trapezoidal rule). A high methylation variability score indicates greater methylation differences compared to the average normal methylation profile.

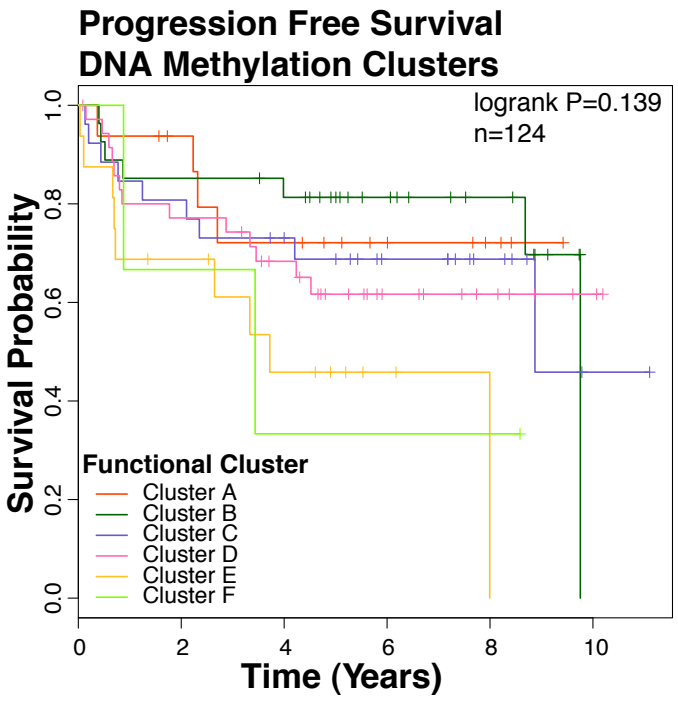


Supplemental Figure 2. Consensus clustering diagnostic plots (A) Consensus matrix CDF for k=2-15 (B) Change in AUC (delta k) for consensus matrix CDFs as k varies from 2 to 15. (C) Heatmap for consensus matrix (k=6)

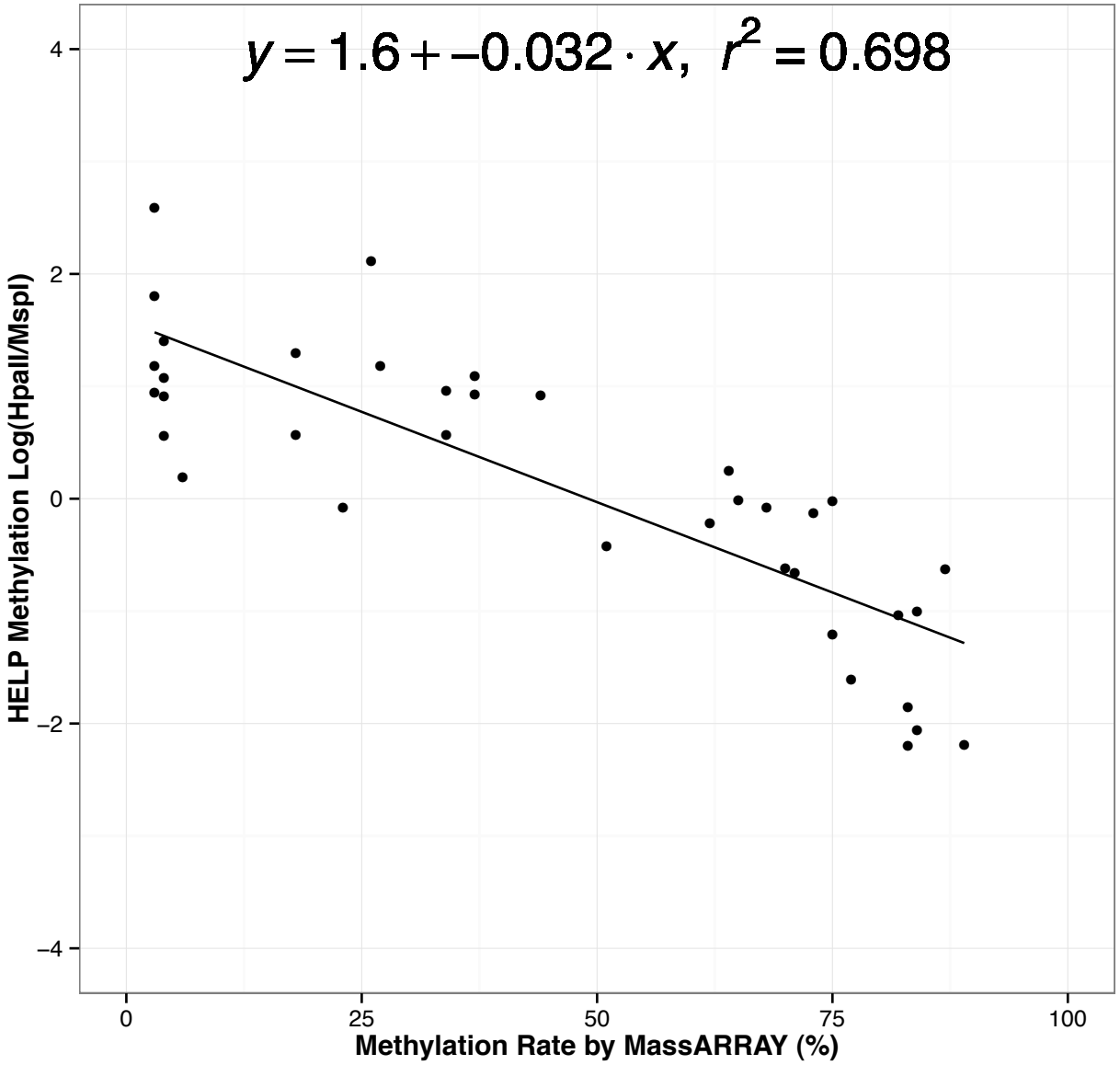


Supplemental Figure 3. Genome-wide quantification of 5-methylcytosine (5-mC). 5-mC content was measured in normal Naïve B cells (NB), normal Germinal Center B cells (Centroblasts) (CB) and DLBCL Clusters A-F using liquid chromatography–mass spectrometry (LC-MS). Mean and standard error of 5-mC content are depicted in the bar graph.

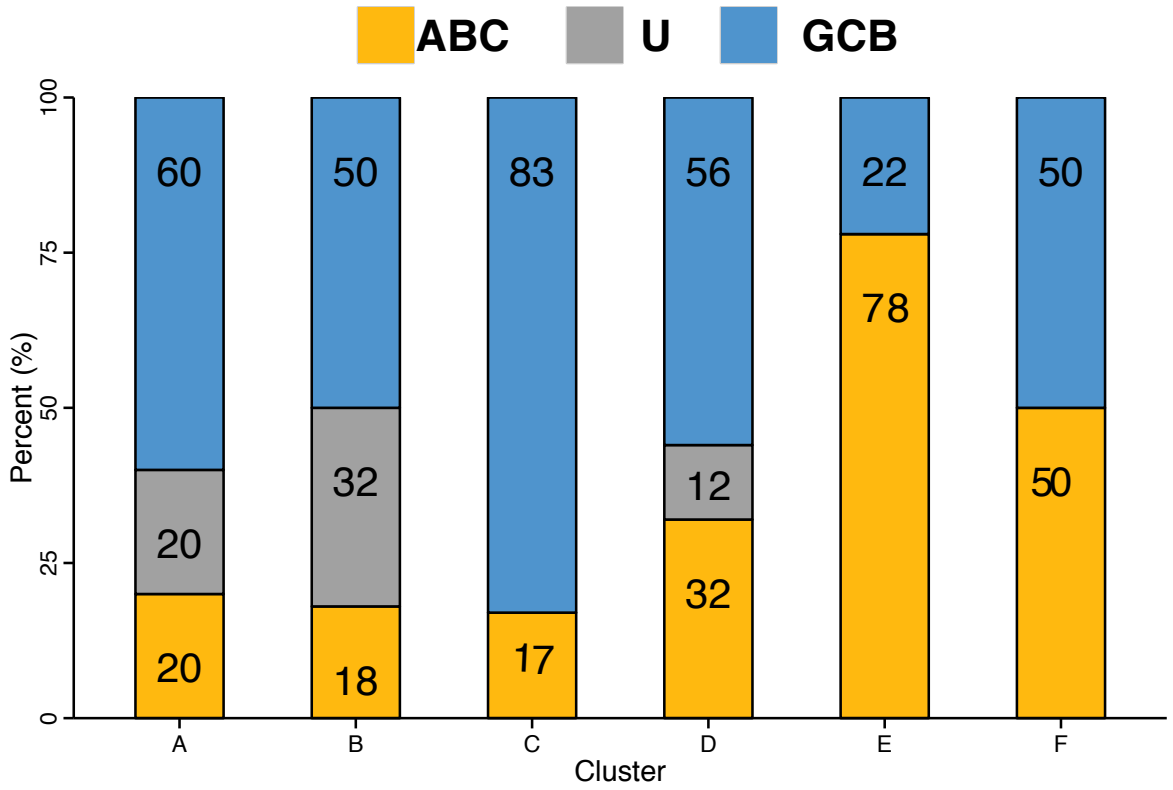




Supplemental Figure 4. Kaplan-Meier curves for (left) Overall Survival (OS) and (right) Progression Free Survival (PFS) for novel DLBCL Clusters. The log rank test p value for cluster association with survival is reported in the top right corner. n represents the number of patients that underwent R-CHOP therapy in this cohort with available follow up data.

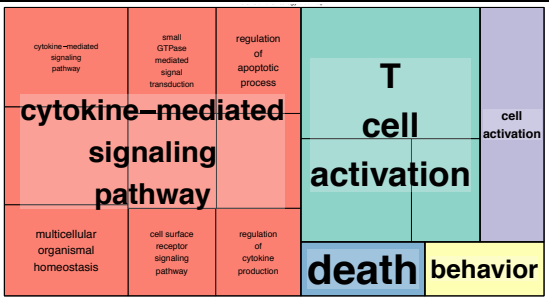


Supplemental Figure 5. Technical Validation of HELP using MassARRAY. Scatter plot showing methylation signal from HELP (y-axis) and MassARRAY (x-axis). Correlation of between the two platforms is 0.698. One unit change in HELP log ratio corresponds to approximately 20% change in methylation rate (%) as measured by MassARRAY.



Supplemental Figure 6. Distribution of Gene Expression based DLBCL subtypes for DNA methylation based clusters. Barplot representing the frequency (%) of the gene-expression based DLBCL subtypes for each DNA methylation defined cluster (n=80). Numbers represent the % frequency of a given COO class in that cluster.

A. GO Biological Process



B.

GeneGO Process Network	pValue	Ratio
Immune response_Antigen presentation	0.000	0.045
Signal transduction_ERBB-family signaling	0.018	0.038
Inflammation_IL-2 signaling	0.020	0.036
Signal transduction_CREM pathway	0.022	0.034
Inflammation_Inflammasome	0.026	0.031
Immune response_T helper cell differentiation	0.034	0.027
Development_Hemopoiesis, Erythropoietin pathway	0.038	0.025
Inflammation_Innate inflammatory response	0.048	0.022
Immune response_TCR signaling	0.053	0.021
Inflammation_Jak-STAT Pathway	0.055	0.021

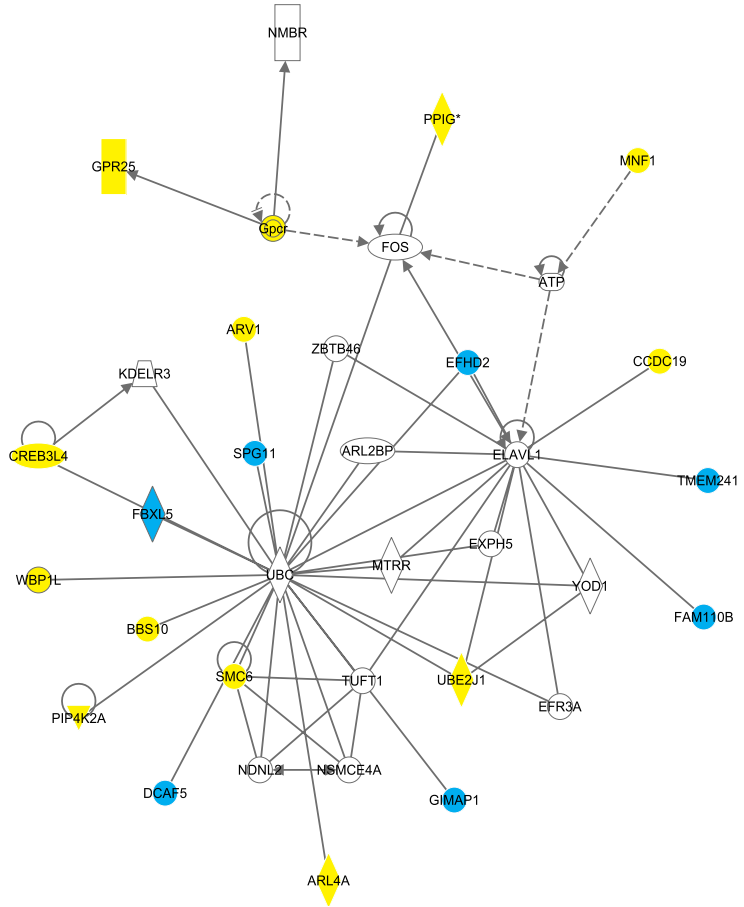
C.

GeneGO Pathway Map	pValue	Ratio
Immune response_IL-9 signaling pathway	0.002	0.111
Immune response_Inhibitory action of Lipoxins on pro-inflammatory TNF-alpha signaling	0.003	0.100
Some pathways of EMT in cancer cells	0.006	0.069
Immune response_IL-23 signaling pathway	0.036	0.111
wtCFTR and deltaF508 traffic / Late endosome and Lysosome (norm and CF)	0.040	0.100
Immune response_Oncostatin M signaling via JAK-Stat in mouse cells	0.044	0.091
Immune response_IL-27 signaling pathway	0.044	0.091
Development_Angiotensin signaling via STATs	0.044	0.091
Immune response_Oncostatin M signaling via JAK-Stat in human cells	0.048	0.083
Regulation of degradation of wt-CFTR	0.048	0.083

D.

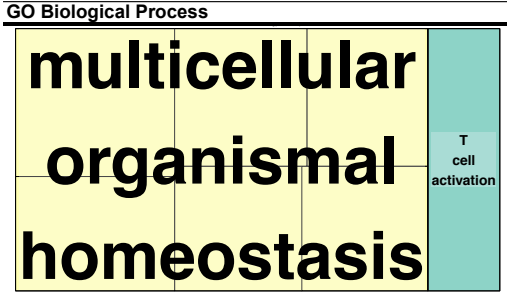
Ingenuity Canonical Pathways	-log(B-H pValue	Ratio
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	1.370	0.015
FLT3 Signaling in Hematopoietic Progenitor Cells	1.370	0.040
Regulation of IL-2 Expression in Activated and Anergic T Lymphocytes	1.370	0.034
T Cell Receptor Signaling	1.370	0.028
Phospholipase C Signaling	1.370	0.015
Type I Diabetes Mellitus Signaling	1.370	0.025
Colorectal Cancer Metastasis Signaling	1.370	0.016
PKCθ Signaling in T Lymphocytes	1.370	0.021
IL-6 Signaling	1.370	0.024
Glucocorticoid Receptor Signaling	1.350	0.014

E. Ingenuity Top Network



Supplemental Figure 7. Cluster A Functional Enrichment Summary. The signature gene list was used as input to various functional annotation resources to determine biological functionality of differentially methylated genes. Highlighted terms represent categories that pass the FDR threshold filter ($q < 0.05$). (A) Enriched GO Biological Processes (BP) from the GO_FAT resource in DAVID. Statistically significant (ease score < 0.05) processes are visualized using the treemap representation from REVIGO. (B) Top 10 most significantly enriched GeneGO (Metacore) Process Networks. (C) Top 10 most significantly enriched GeneGO (Metacore) Pathway Maps. (D) Most enriched Ingenuity Canonical Pathways. (E) Ingenuity top network represent the highest scoring enrichment for signature genes. Genes represented in blue indicate hypomethylated and yellow hypermethylated in DLBCL compared to GCB cells. Edges represent known interactions (curated) between two genes.

A. **GO Biological Process**



B.

GeneGO Process Network	pValue	Ratio
Immune response_TCR signaling	0.009	0.032
Immune response_Antigen presentation	0.014	0.027
Inflammation_IL-13 signaling pathway	0.021	0.042
Signal transduction_NOTCH signaling	0.042	0.018
Immune response_T helper cell differentiation	0.046	0.027
Cell adhesion_Cell junctions	0.074	0.021
Neurophysiological process_Corticoliberin signaling	0.098	0.045
Cell cycle_Mitosis	0.101	0.017
Immune response_IL-5 signalling	0.103	0.043
Proliferation_Lymphocyte proliferation	0.113	0.016

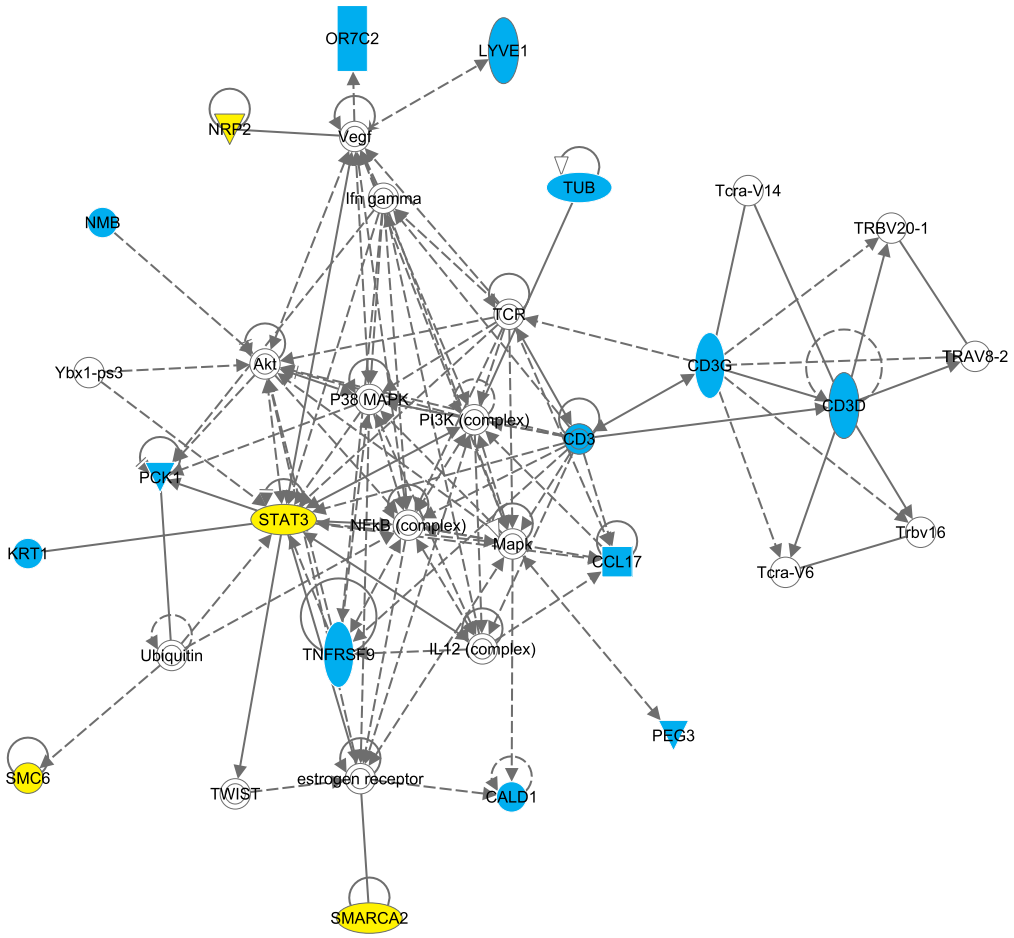
C.

GeneGO Pathway Map	pValue	Ratio
Immune response_IL-13 signaling via JAK-STAT	0.006	0.083
Transcription_Transcription factor Tubby signaling pathways	0.025	0.200
Ganglioside Metabolism p2	0.035	0.143
Immune response_IL-23 signaling pathway	0.045	0.111
Immune response_Oncostatin M signaling via JAK-Stat in mouse cells	0.055	0.091
Immune response_IL-27 signaling pathway	0.055	0.091
Development_Angiotensin signaling via STATs	0.055	0.091
Immune response_Oncostatin M signaling via JAK-Stat in human cells	0.060	0.083
Role of Nicotine-induced Leptin resistance in hypothalamus in development of obesity	0.065	0.077
Lacto-series GSL Metabolism / Human version	0.065	0.077

D.

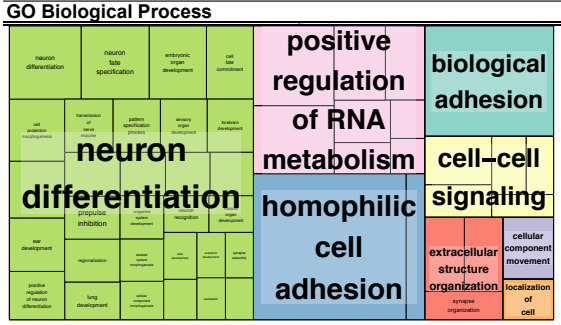
Ingenuity Canonical Pathways	-log(B-H p-value	Ratio
Glucocorticoid Receptor Signaling	1.080	0.017
Hematopoiesis from Pluripotent Stem Cells	0.592	0.032
Nur77 Signaling in T Lymphocytes	0.592	0.032
Calcium-induced T Lymphocyte Apoptosis	0.592	0.029
CCR5 Signaling in Macrophages	0.592	0.021
Cytotoxic T Lymphocyte-mediated Apoptosis of Target Cells	0.592	0.024
Regulation of IL-2 Expression in Activated and Anergic T Lymphocytes	0.592	0.023
OX40 Signaling Pathway	0.592	0.022
CTLA4 Signaling in Cytotoxic T Lymphocytes	0.559	0.020
T Cell Receptor Signaling	0.559	0.018

E. **Ingenuity Top Network**



Supplemental Figure 8. Cluster B Functional Enrichment Summary. The signature gene list was used as input to various functional annotation resources to determine biological functionality of differentially methylated genes. Highlighted terms represent categories that pass the FDR threshold filter ($q < 0.05$). (A) Enriched GO Biological Processes (BP) from the GO_FAT resource in DAVID. Statistically significant (ease score < 0.05) processes are visualized using the treemap representation from REVIGO. (B) Top 10 most significantly enriched GeneGO (Metacore) Process Networks. (C) Top 10 most significantly enriched GeneGO (Metacore) Pathway Maps. (D) Most enriched Ingenuity Canonical Pathways. (E) Ingenuity top network represent the highest scoring enrichment for signature genes. Genes represented in blue indicate hypomethylated and yellow hypermethylated in DLBCL compared to GCB cells. Edges represent known interactions (curated) between two genes.

A. GO Biological Process



B.

GeneGO Process Network	pValue	Ratio
Cell adhesion_Cadherins	0.000	0.067
Development_Neurogenesis in general	0.000	0.063
Development_Neurogenesis_Synaptogenesis	0.005	0.050
Reproduction_Gonadotropin regulation	0.010	0.045
Reproduction_GnRH signaling pathway	0.011	0.048
Reproduction_Progesterone signaling	0.015	0.042
Signal transduction_NOTCH signaling	0.028	0.038
Reproduction_Feeding and Neurohormone signaling	0.038	0.038
Regulation of metabolism_Bile acid regulation of lipid metabolism and negative FXR-dependent regulation of bile acids concentration	0.040	0.056
Cell adhesion_Attractive and repulsive receptors	0.040	0.040

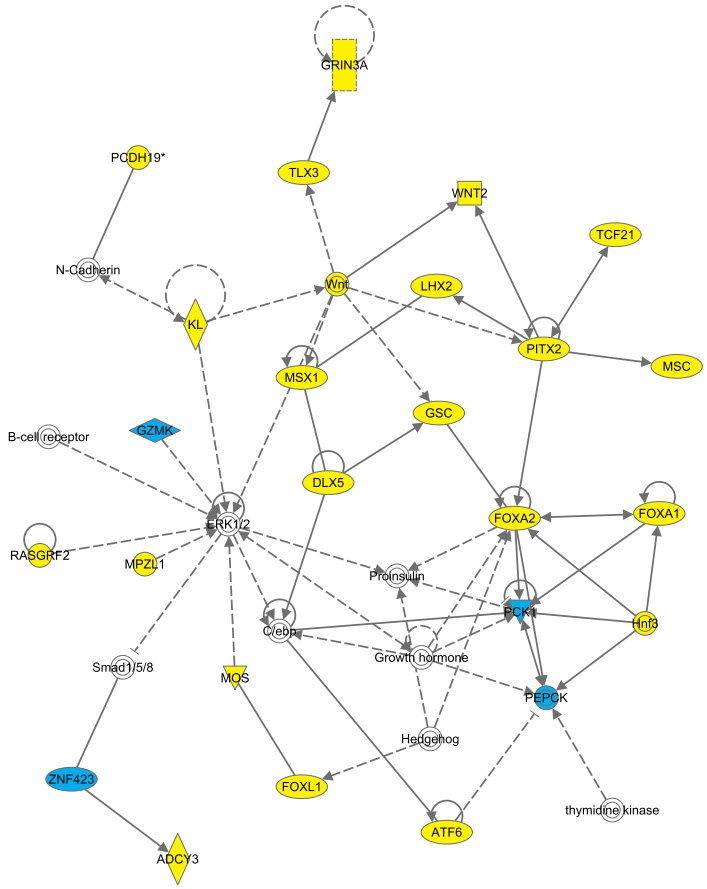
C.

GeneGO Pathway Map	pValue	Ratio
Regulation of metabolism_Bile acids regulation of glucose and lipid metabolism via FXR	0.000	0.108
Cell adhesion_Endothelial cell contacts by junctional mechanisms	0.001	0.115
Signal transduction_Activin A signaling regulation	0.002	0.091
Apoptosis and survival_Cytoplasmic/mitochondrial transport of proapoptotic proteins Bid, Bmf and Bim	0.003	0.088
Immune response_IL-9 signaling pathway	0.003	0.083
Cytokine production by Th17 cells in CF	0.004	0.077
Immune response_Inhibitory action of Lipoxins on pro-inflammatory TNF-alpha signaling	0.006	0.065
Apoptosis and survival_HTR1A signaling	0.008	0.060
Signal transduction_PKA signaling	0.008	0.059
Some pathways of EMT in cancer cells	0.008	0.059

D.

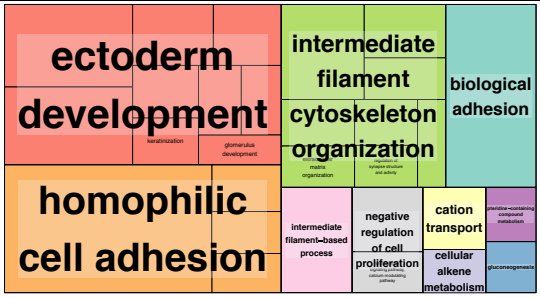
Ingenuity Canonical Pathways	-log(B-H pValue	Ratio
IL-9 Signaling	1.03	0.100
Role of JAK1 and JAK3 in γc Cytokine Signaling	1.03	0.075
Amyotrophic Lateral Sclerosis Signaling	1.03	0.051
Neuropathic Pain Signaling In Dorsal Horn Neurons	1.03	0.056
G-Protein Coupled Receptor Signaling	1.03	0.037
Regulation of the Epithelial-Mesenchymal Transition Pathway	1.03	0.042
Role of Oct4 in Mammalian Embryonic Stem Cell Pluripotency	1.03	0.089
Leptin Signaling in Obesity	1.03	0.060
Pancreatic Adenocarcinoma Signaling	1.03	0.050
CNTF Signaling	0.87	0.073

E. Ingenuity Top Network



Supplemental Figure 9. Cluster C Functional Enrichment Summary. The signature gene list was used as input to various functional annotation resources to determine biological functionality of differentially methylated genes. Highlighted terms represent categories that pass the FDR threshold filter ($q < 0.05$). (A) Enriched GO Biological Processes (BP) from the GO_FAT resource in DAVID. Statistically significant (ease score < 0.05) processes are visualized using the treemap representation from REVIGO. (B) Top 10 most significantly enriched GeneGO (Metacore) Process Networks. (C) Top 10 most significantly enriched GeneGO (Metacore) Pathway Maps. (D) Most enriched Ingenuity Canonical Pathways. (E) Ingenuity top network represent the highest scoring enrichment for signature genes. Genes represented in blue indicate hypomethylated and yellow hypermethylated in DLBCL compared to GCB cells. Edges represent known interactions (curated) between two genes.

A. GO Biological Process



B.

GeneGO Process Network	pValue	Ratio
Cell adhesion_Cadherins	0.000	0.096
Cytoskeleton_Intermediate filaments	0.001	0.127
Development_Skeletal muscle development	0.007	0.082
Proliferation_Negative regulation of cell proliferation	0.072	0.056
Development_Regulation of angiogenesis	0.077	0.055
Development_Keratinocyte differentiation	0.085	0.083
Development_Neuromuscular junction	0.100	0.059
Development_EMT_Regulation of epithelial-to-mesenchymal transition	0.102	0.051
Cell cycle_G0-G1	0.115	0.073
Signal Transduction_BMP and GDF signaling	0.134	0.059

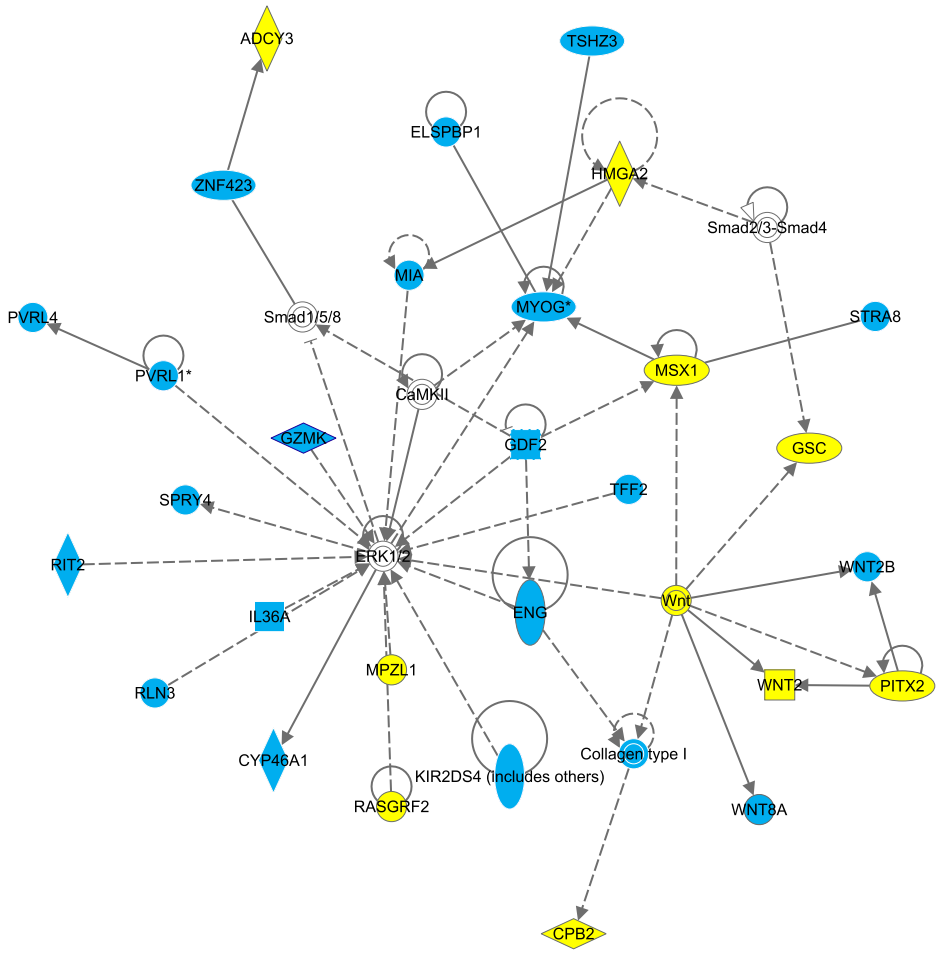
C.

GeneGO Pathway Map	pValue	Ratio
Cytoskeleton remodeling_Keratin filaments	0.000	0.200
Development_Hedgehog and PTH signaling pathways in bone and cartilage development	0.006	0.188
Folic acid metabolism	0.026	0.182
Tricarbonic acid cycle	0.036	0.154
CCR4-dependent immune cell chemotaxis in asthma and atopic dermatitis	0.041	0.143
Chemotaxis_CCR4-induced chemotaxis of immune cells	0.041	0.143
Mechanism of action of CCR4 antagonists in asthma and atopic dermatitis (Variant 1)	0.041	0.143
Development_Keratinocyte differentiation	0.044	0.088
Immune response_Antigen presentation by MHC class I	0.047	0.133
Ascorbate metabolism	0.069	0.333

D.

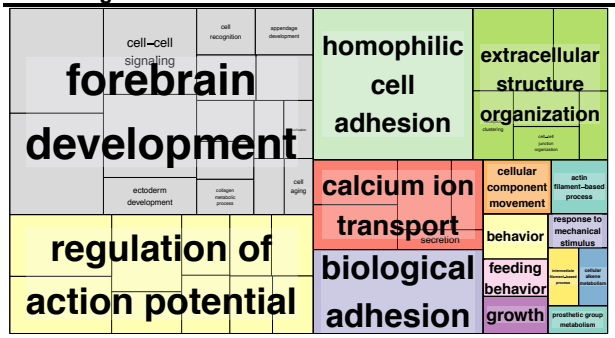
Ingenuity Canonical Pathways	-log(B-H pValue	Ratio
Thioredoxin Pathway	1.600	0.375
Vitamin-C Transport	0.693	0.136

E. Ingenuity Top Network



Supplemental Figure 10. Cluster D Functional Enrichment Summary. The signature gene list was used as input to various functional annotation resources to determine biological functionality of differentially methylated genes. Highlighted terms represent categories that pass the FDR threshold filter ($q < 0.05$). (A) Enriched GO Biological Processes (BP) from the GO_FAT resource in DAVID. Statistically significant (ease score < 0.05) processes are visualized using the treemap representation from REVIGO. (B) Top 10 most significantly enriched GeneGO (Metacore) Process Networks. (C) Top 10 most significantly enriched GeneGO (Metacore) Pathway Maps. (D) Most enriched Ingenuity Canonical Pathways. (E) Ingenuity top network represent the highest scoring enrichment for signature genes. Genes represented in blue indicate hypomethylated and yellow hypermethylated in DLBCL compared to GCB cells. Edges represent known interactions (curated) between two genes.

A. GO Biological Process



B.

GeneGO Process Network	pValue	Ratio
Development_Skeletal muscle development	0.000	0.194
Development_Neurogenesis in general	0.000	0.172
Cell adhesion_Cadherins	0.000	0.161
Cytoskeleton_Intermediate filaments	0.000	0.198
Development_Neurogenesis_Axonal guidance	0.001	0.139
Muscle contraction	0.001	0.150
Development_Neurogenesis_Synaptogenesis	0.001	0.144
Signal transduction_Neuropeptide signaling pathways	0.001	0.148
Neurophysiological process_Transmission of nerve impulse	0.003	0.132
Cytoskeleton_Regulation of cytoskeleton rearrangement	0.003	0.137

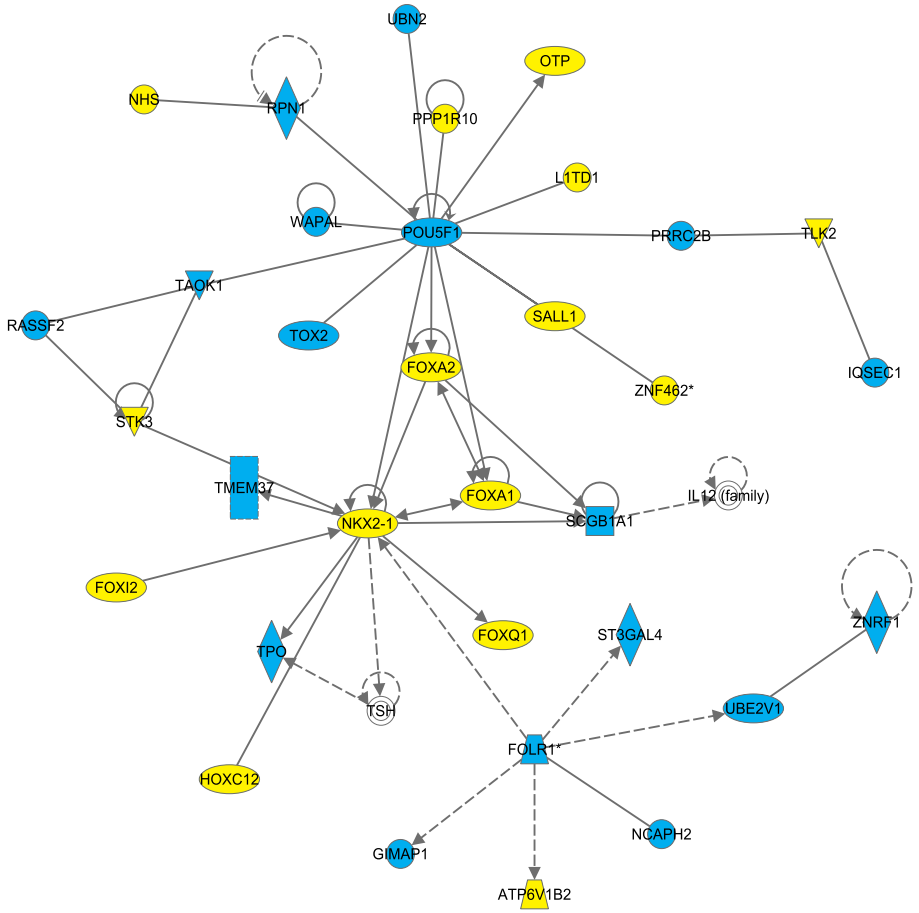
C.

GeneGO Pathway Map	pValue	Ratio
ENaC regulation in airways (normal and CF)	0.000	0.212
Cell adhesion_Ephrin signaling	0.000	0.200
G-protein signaling_Regulation of cAMP levels by ACM	0.000	0.200
Cytoskeleton remodeling_Keratin filaments	0.000	0.222
Neurophysiological process_Receptor-mediated axon growth repulsion	0.000	0.178
Cell adhesion_Tight junctions	0.000	0.194
Transport_ACM3 in salivary glands	0.001	0.167
Nicotine signaling (general scheme)	0.001	0.238
Development_ACM2 and ACM4 activation of ERK	0.001	0.163
Airway smooth muscle contraction in asthma	0.001	0.143

D.

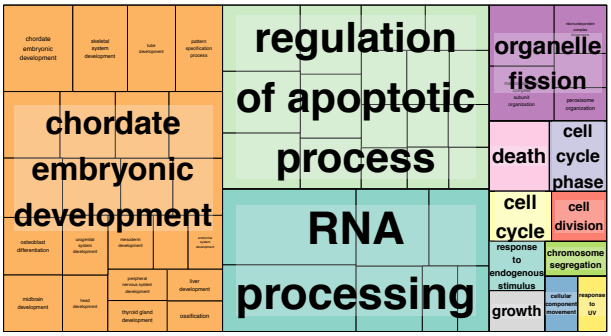
Ingenuity Canonical Pathways	(B-H pV	Ratio
Neuropathic Pain Signaling In Dorsal Horn Neurons	0.156	0.139
Thioredoxin Pathway	0.156	0.375
Pregnenolone Biosynthesis	0.156	0.231

E. Ingenuity Top Network



Supplemental Figure 11. Cluster E Functional Enrichment Summary. The signature gene list was used as input to various functional annotation resources to determine biological functionality of differentially methylated genes. Highlighted terms represent categories that pass the FDR threshold filter ($q < 0.05$). (A) Enriched GO Biological Processes (BP) from the GO_FAT resource in DAVID. Statistically significant (ease score < 0.05) processes are visualized using the treemap representation from REVIGO. (B) Top 10 most significantly enriched GeneGO (Metacore) Process Networks. (C) Top 10 most significantly enriched GeneGO (Metacore) Pathway Maps. (D) Most enriched Ingenuity Canonical Pathways. (E) Ingenuity top network represent the highest scoring enrichment for signature genes. Genes represented in blue indicate hypomethylated and yellow hypermethylated in DLBCL compared to GCB cells. Edges represent known interactions (curated) between two genes.

A. GO Biological Process



B.

GeneGO Process Network	pValue	Ratio
Cytoskeleton_Regulation of cytoskeleton rearrangement	0.000	0.530
Cell cycle_G1-S Growth factor regulation	0.000	0.523
Development_Neurogenesis in general	0.000	0.521
Cell cycle_Mitosis	0.000	0.520
Transcription_mRNA processing	0.000	0.519
Cell cycle_G1-S Interleukin regulation	0.000	0.539
Development_Skeletal muscle development	0.000	0.514
Proliferation_Positive regulation cell proliferation	0.000	0.480
Cell adhesion_Attractive and repulsive receptors	0.000	0.497
Signal transduction_NOTCH signaling	0.000	0.475

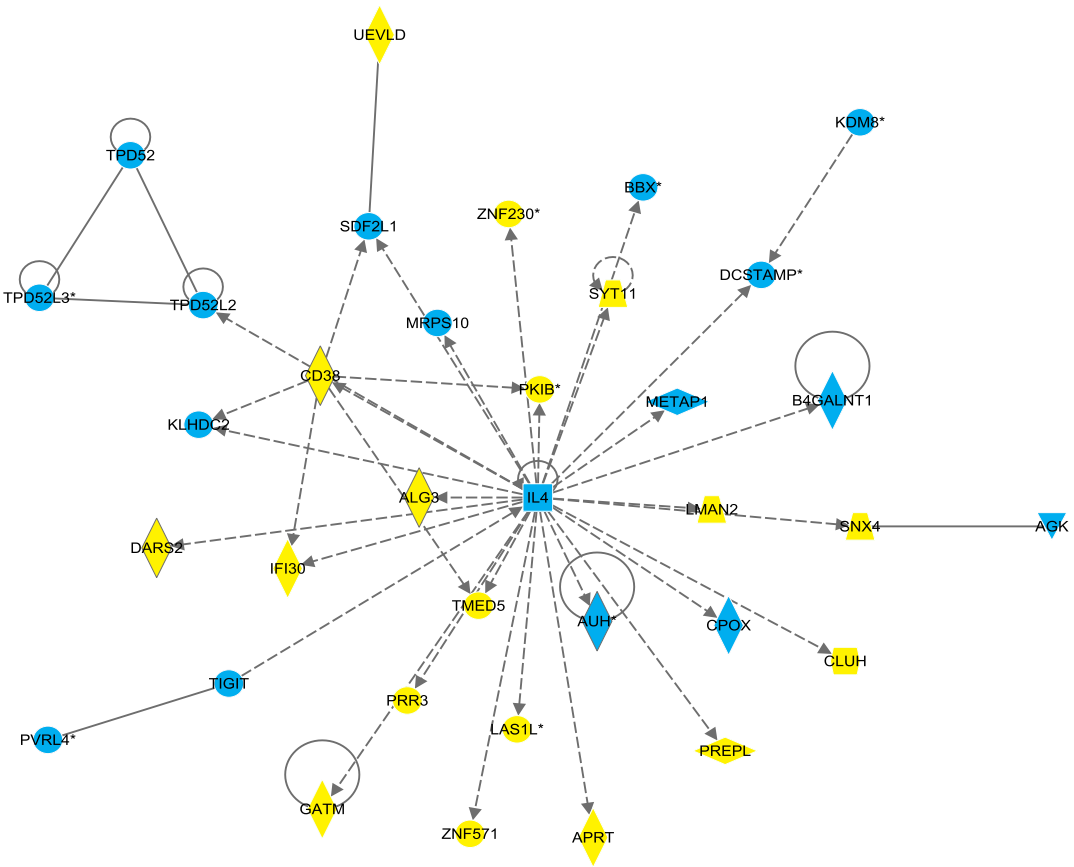
C.

GeneGO Pathway Map	pValue	Ratio
Cytoskeleton remodeling_TGF, WNT and cytoskeletal remodeling	0.000	0.459
Development_Beta-adrenergic receptors transactivation of EGFR	0.000	0.649
Development_PIP3 signaling in cardiac myocytes	0.000	0.574
Translation _Regulation of EIF2 activity	0.000	0.615
Signal transduction_Erk Interactions: Inhibition of Erk	0.000	0.647
Neurophysiological process_Receptor-mediated axon growth repulsion	0.000	0.578
Protein folding and maturation_POMC processing	0.000	0.667
Signal transduction_AKT signaling	0.000	0.581
Development_Angiotensin signaling via PYK2	0.000	0.581
Development_Endothelin-1/EDNRA transactivation of EGFR	0.000	0.565

D.

Ingenuity Canonical Pathways	-log(B-H pValue	Ratio
Axonal Guidance Signaling	4.340	0.401
CXCR4 Signaling	2.750	0.440
Huntington's Disease Signaling	2.750	0.421
Breast Cancer Regulation by Stathmin1	2.690	0.430
Glioblastoma Multiforme Signaling	2.690	0.433
Molecular Mechanisms of Cancer	2.690	0.381
Cardiac Hypertrophy Signaling	2.490	0.406
AMPK Signaling	2.420	0.383
PI3K/AKT Signaling	2.400	0.417
IL-1 Signaling	2.400	0.439

E. Ingenuity Top Network

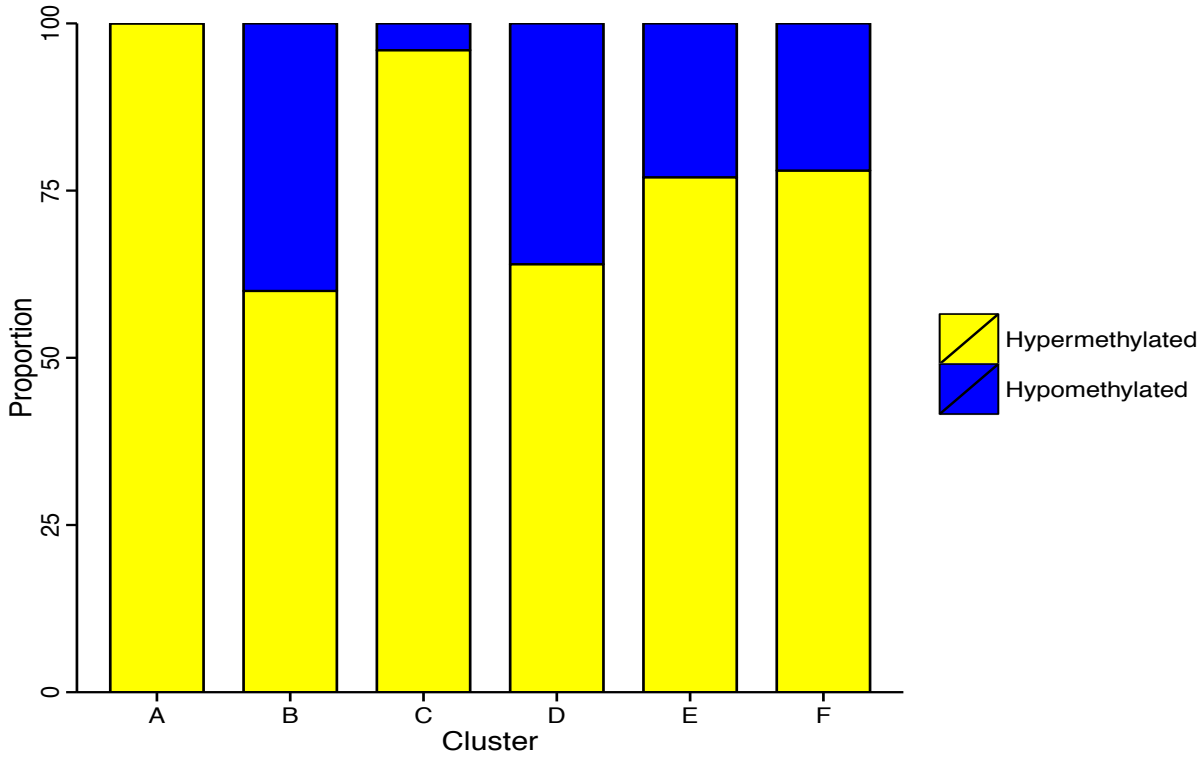


Supplemental Figure 12. Cluster F Functional Enrichment Summary. The signature gene list was used as input to various functional annotation resources to determine biological functionality of differentially methylated genes. Highlighted terms represent categories that pass the FDR threshold filter ($q < 0.05$). (A) Enriched GO Biological Processes (BP) from the GO_FAT resource in DAVID. Statistically significant (ease score < 0.05) processes are visualized using the treemap representation from REVIGO. (B) Top 10 most significantly enriched GeneGO (Metacore) Process Networks. (C) Top 10 most significantly enriched GeneGO (Metacore) Pathway Maps. (D) Most enriched Ingenuity Canonical Pathways. (E) Ingenuity top network represent the highest scoring enrichment for signature genes. Genes represented in blue indicate hypomethylated and yellow hypermethylated in DLBCL compared to GCB cells. Edges represent known interactions (curated) between two genes.

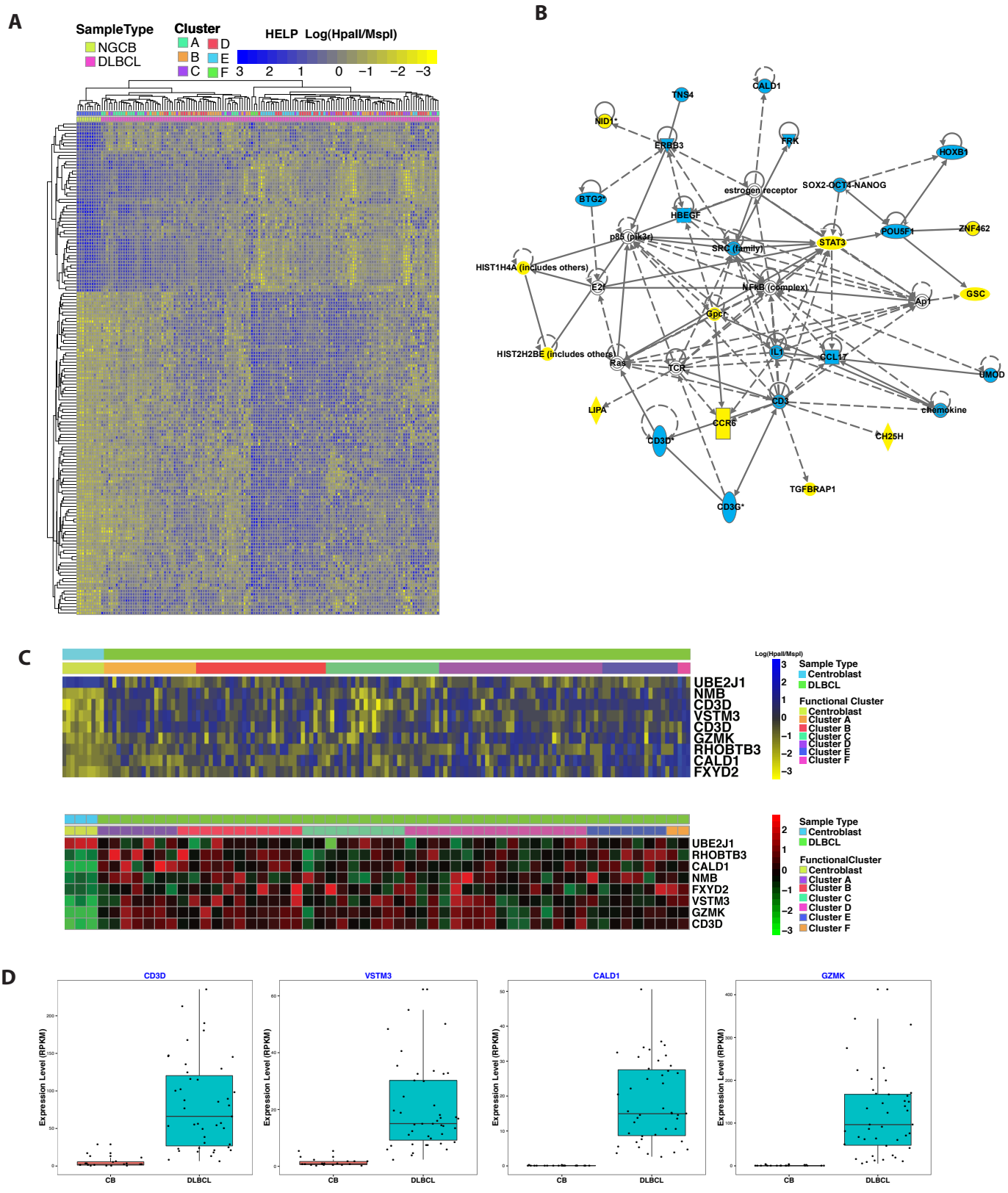
A

Cluster	# Genes Observed	# Genes Expected	Q-value	Enrichment
Cluster A	5	8	7.30E-02	-
Cluster B	5	7	1.86E-01	-
Cluster C	93	65	**1.20E-08**	up
Cluster D	47	33	**8.73E-05**	up
Cluster E	225	153	**7.19E-24**	up
Cluster F	809	842	**1.79E-03**	down

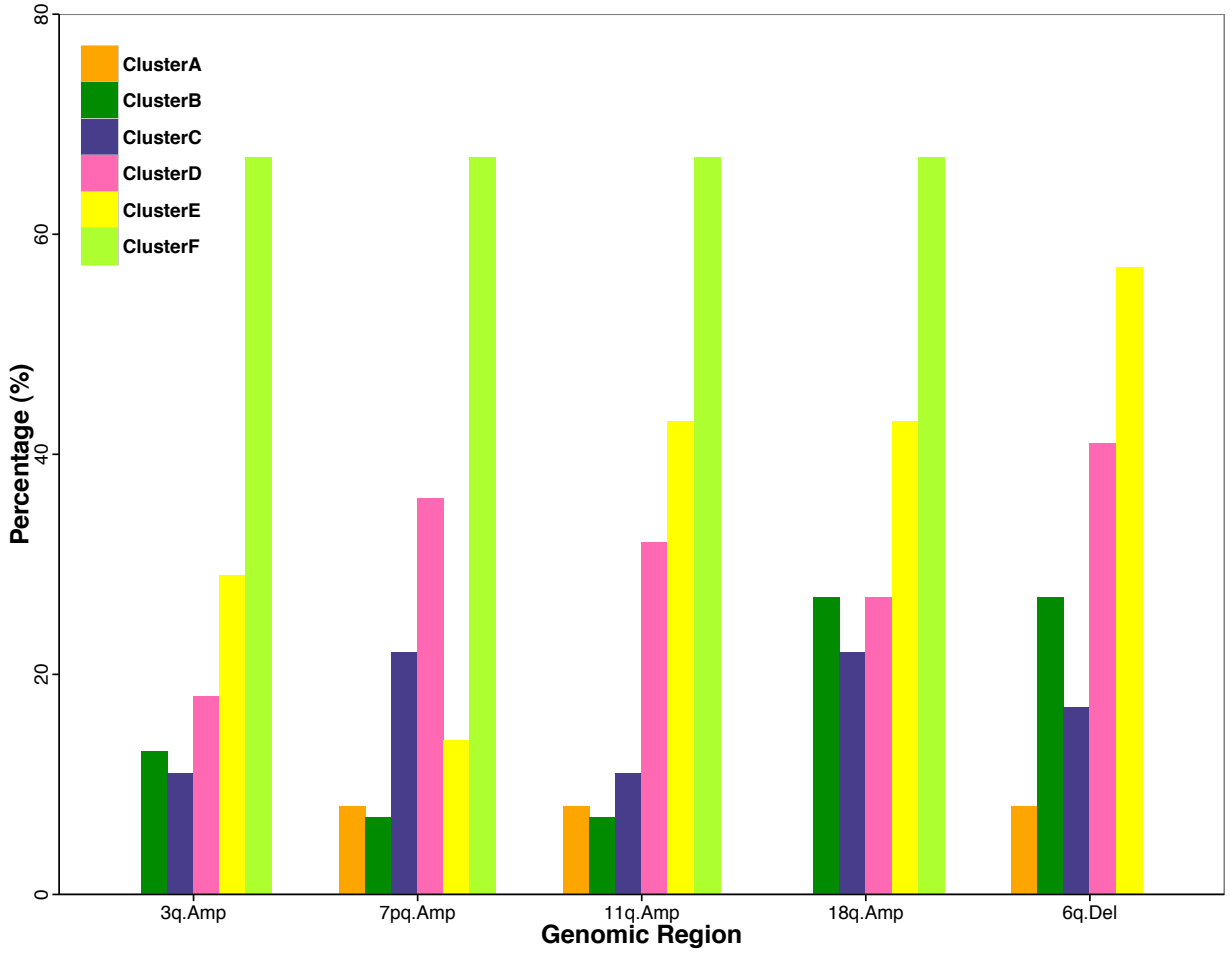
B



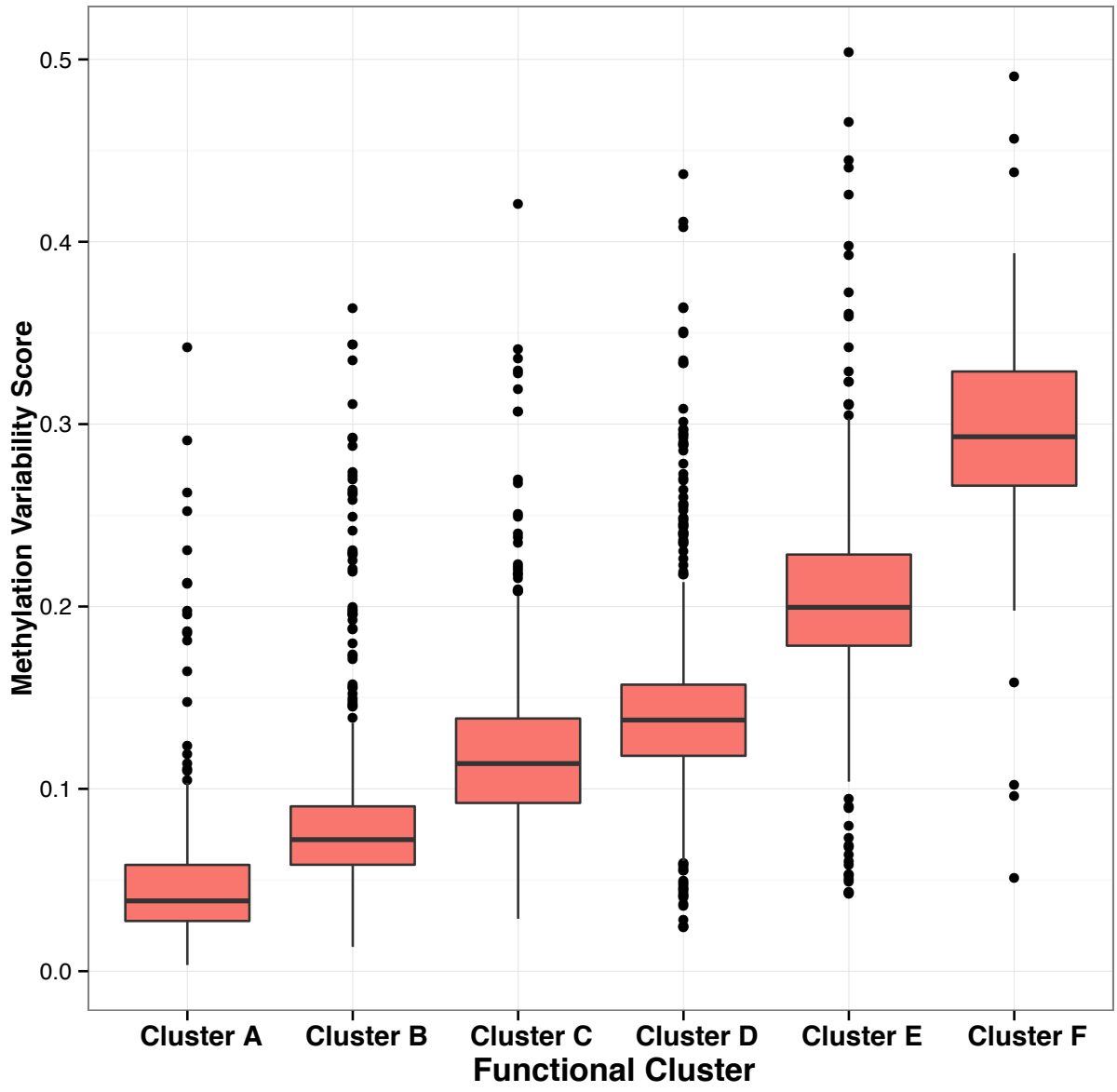
Supplemental Figure 13. EZH2 Target Enrichment. Cluster signature overlap with experimentally defined targets of EZH2. (A) Hypergeometric test results for statistical enrichment for EZH2 target genes in each cluster. (B) Proportion of cluster EZH2 targets that gain and lose methylation in DLBCL clusters.



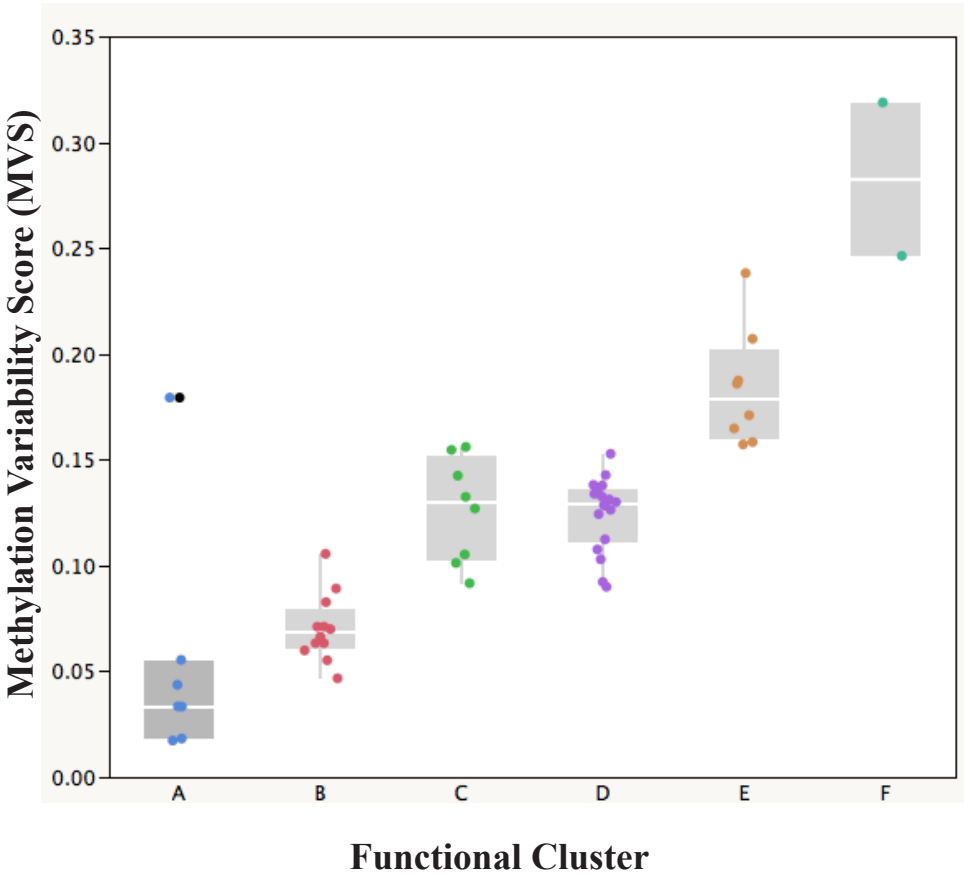
Supplemental Figure 14. DNA methylation changes across all DLBCLs. (A) Heatmap showing differentially methylated HELP fragments between Normal and DLBCL samples from moderated t-test (LIMMA $q < 0.05$ and $\log |FC| >= 1.5$). (B) Ingenuity network analysis for differentially methylated genes between GCB and DLBCL. Genes shaded in yellow are hypermethylated in DLBCL and blue are hypomethylated in DLBCL. (C) Heatmap representation of genes inversely correlated between methylation and expression. Each row represents a probeset, and column a sample. Annotation bars indicate the Sample Type Normal Germinal Center B Cell (NGCB) or DLBCL, as well as the functional cluster identity of the sample. The top heatmap represents methylation data. Yellow shows relative hypermethylation and blue relative hypomethylation. The bottom heatmap represents expression data measured on a microarray platform. Red indicates over-expressed genes, and green under-expressed genes.



Supplemental Figure 15. Broad amplification and deletion regions in DLBCL. Frequency of GISTIC called genetic abnormalities in DLBCL clusters. (A) Regions with significant differences across clusters are shown (fisher's exact test p <= 0.1). (B) Regions with significant differences when comparing Clusters B, D and E versus Cluster A and C (fisher's exact test p <= 0.1, Cluster F excluded from analysis).



Supplemental Figure 16. Methylation Variability in Copy Number Neutral Regions. Boxplots depicting methylation variability score (MVS) (y-axis) by Functional cluster (x-axis). The MVS for each cluster was calculated using HELP fragments that mapped to copy number neutral regions (all other genomic regions without GISTIC called amplifications or deletions.



Supplemental Figure 17. Methylation Variability Scores for High Tumor Purity Samples. Boxplots depicting methylation variability score (MVS) (y-axis) by Functional cluster (x-axis) for the subset of samples with tumor purity >=90% (n=55).