Figure S1. MeDIP enrichment analysis. All samples were subjected to a quantitative PCR assay using specific and non-specific primers for DNA methylation. Non-specific primers analyzed known global methylated and unmethylated genes. Specific primers for DNA methylation enrichment analysis included TSH2B, positive methylation control and GAPDH as a negative methylation control. meDNA, methylated DNA; un/ume, unmethylated; MeDIP, methylated DNA immunoprecipitation.

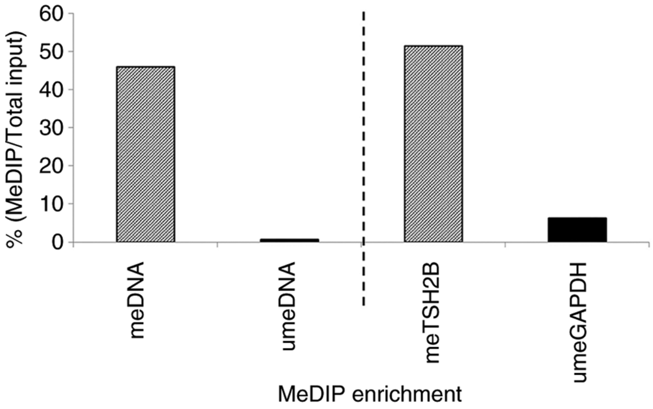


Figure S2. Venn diagram shows differentially methylated regions in head and neck squamous cell carcinoma tumor samples distributed by anatomical subsite.

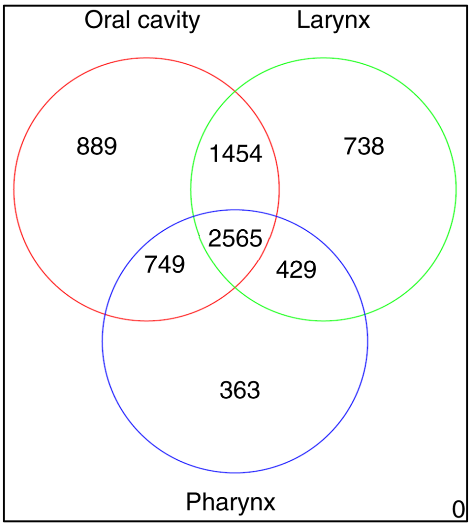
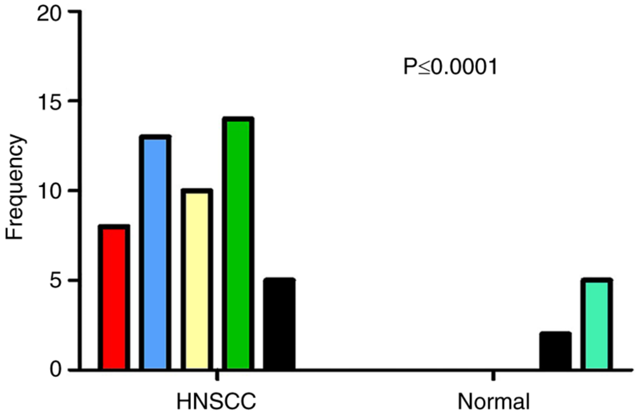


Figure S3. Concurrent aberrant DNA methylation of candidate genes in HNSCC vs normal samples. The red, blue, ivory, green, black and turquoise bars represents number of samples with five, four, three, two, one and zero candidate genes methylated, respectively. HNSCC, head and neck squamous cell carcinoma.



Data S1. Bioinformatics pipeline for peak discovery algorithm. Detailed bioinformatics pipeline or peak discovery algorithm using CHARM bioinformatics package within the R statistical programming.

Promoter DNA methylation patterns in oral, laryngeal, and oropharyngeal anatomical regions pipeline analysis

Nitesh Turaga

Pipeline Description

1. Separate files/Samples into Cancer and Control
2. Analyze Transcription start sites and CpG Islands separately, ignore other features.
3. Transcription Start Sites

   ### For one status

   1. Get the frequency of genes within each sample.
   2. Among all the samples, select for genes with frequency greater than 20%(based on the sample size).
   3. Re-annotate with the peaks file for the selected genes with high fre

   ### Between status types

   1. Get common genes between both status types (based on ncbi gene id - no duplicates).
   2. Remove from either status, making them mutually exclusive (based on gene symbol).

Set working directory

```
rm(list = ls(all=T))
mypath = "~/Documents/TestRun/Charm Analysis/charmData/Nimblegen_Originals/Bianca_Project/Peaks_Files/"
setwd(mypath)
library(plyr)
library(BiocGenerics)
```

```
## Loading required package: parallel
##
## Attaching package: 'BiocGenerics'
##
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
##
## The following object is masked from 'package:stats':
##
##     xtabs
##
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, as.vector, cbind,
##     colnames, duplicated, eval, evalq, Filter, Find, get,
##     intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rep.int, rownames, sapply, setdiff, sort,
##     table, tapply, union, unique, unlist
```

Pipeline Steps

Initialization step

```
# Preprocess for easy file reading. Make lists
files.control = list.files(file.path(mypath,"control"),pattern = ".txt",full.names=T)
files.cancer = list.files(file.path(mypath,"cancer"),pattern = ".txt",full.names=T)
```

Read in each sample for from status

```
# Select for TSS within each sample read.sample =
function(filename,feature) {
        f = read.csv(as.character(filename),sep="\t",comment.char = "#") f =
        f[f$FEATURE_TRACK == feature,]
        f = f[f$Peaks > 1.999,] # Peaks cutoff return(f)
        }



gene.frequency.in.status = function(list.of.files,feature) { f =


        read.sample(list.of.files[1],feature)
        f.gene = data.frame(table(as.character(f$Name))) f.gene$sample =
        gsub(".+/","",list.of.files[1])

        for (i in 2:length(list.of.files)) {
                f = read.sample(list.of.files[i],feature)
                f.next.gene = data.frame(table(as.character(f$Name))) f.next.gene$sample =
                gsub(".+/","",list.of.files[i]) f.gene = rbind(f.gene,f.next.gene)
                 }
           # get
        overall.f.gene = data.frame(table(f.gene$Var1)) cutoff =
        round((20/100)*length(list.of.files))
        overall.f.gene = overall.f.gene[overall.f.gene$Freq>cutoff,]

        return(overall.f.gene)
        }



annotate = function(gene.list,list.of.files){
        f = read.sample(list.of.files[1], feature = "transcription_start_site") f.genes = f[which(gene.list %in%
        f$Name),] ################################################
        # use this only for controls
        f.genes = f.genes[,c("DATA_INDEX","CHROMOSOME","DATA_START","DATA_END","Peaks","FEATURE_TRACK",
                              "FEATURE_STRAND","FEATURE_START","FEATURE_END","SHORTEST_DISTANCE_FROM_FEATURE_TO_DATA
                              _POINT

                              ,"CENTER_TO_CENTER_DISTANCE_FROM_FEATURE_TO_DATA_POINT","ncbi_gene_id"
                              ,"Name","description","synonyms")]
        ################################################
        for(i in 2:length(list.of.files)){
                f.next = read.sample(list.of.files[i], feature = "transcription_start_site") f.next.genes = f.next[which(gene.list %in%
                f.next$Name),] ################################################
                # Use this only for controls
        f.next.genes = f.next.genes[,c("DATA_INDEX","CHROMOSOME","DATA_START","DATA_END","Peaks","FEATURE_TR
ACK",
                     "FEATURE_STRAND","FEATURE_START","FEATURE_END","SHORTEST_DISTANCE_FROM_FEATURE_TO_DATA_PO
INT"
                     ,"CENTER_TO_CENTER_DISTANCE_FROM_FEATURE_TO_DATA_POINT","ncbi_gene_id"
                     ,"Name","description","synonyms")]
        ##############################################
        f.genes = rbind(f.genes,f.next.genes)
        }
    return(f.genes)
    }
```

```
# Get overall.genes over frequency in both cancer and control
cancer.genes = gene.frequency.in.status(files.cancer,feature ="transcription_start_site")
normal.genes = gene.frequency.in.status(files.control,feature = "transcription_start_site")

# Annotate
cancer.list = cancer.genes$Var1
normal.list = normal.genes$Var1
cancer.probes.with.high.freq.genes = annotate(cancer.list,files.cancer)
control.probes.with.high.freq.genes = annotate(normal.list,files.control)




# Get common genes
cancer.no.dup = subset(cancer.probes.with.high.freq.genes,!duplicated(cancer.probes.with.high.freq.genes$ncb
i_gene_id))
table(cancer.no.dup$CHROMOSOME)
```

```
## 
##          chr1   chr1_random         chr10         chr11         chr12 
##          1987             0           743          1254           980 
##         chr13  chr13_random         chr14         chr15         chr16 
##           317             0           635           651           783 
##         chr17  chr17_random         chr18         chr19  chr19_random 
##          1087             9           264          1169             5 
##          chr2         chr20         chr21  chr21_random         chr22 
##          1193           478           223             0           390 
##          chr3   chr3_random          chr4   chr4_random          chr5 
##          1043             0           707             0           846 
##          chr6          chr7          chr8          chr9          chrX 
##          1009           901           646           767           727 
##   chrX_random          chrY   chr2_random   chr5_random  chr16_random 
##             0            82             0             0             0 
##   chr6_random  chr15_random   chr9_random  chr11_random  chr22_random 
##             0             0             1             0             0
```

```
head(cancer.no.dup)
```

```
##   DATA_INDEX CHROMOSOME DATA_START DATA_END Peaks         FEATURE_TRACK
## 1      66098       chr1     883773   883928 2.689 transcription_start_site
## 2      66098       chr1     883773   883928 2.689 transcription_start_site
## 4      66099       chr1     888768   889017 2.682 transcription_start_site
## 5          3       chr1     936559   937295 3.109 transcription_start_site
## 6      57426       chr1    1099854  1099983 2.590 transcription_start_site
## 8          8       chr1    1159043  1159794 4.135 transcription_start_site
##   FEATURE_STRAND FEATURE_START FEATURE_END
## 1              -        884542      884542
## 2              +        885829      885829
## 4              +        891739      891739
## 5              +        938709      938709
## 6              +       1104939     1104939
## 8              -       1157310     1157310
##   SHORTEST_DISTANCE_FROM_FEATURE_TO_DATA_POINT
## 1                                          614
## 2                                        -1901
## 4                                        -2722
## 5                                        -1414
## 6                                        -4956
## 8                                        -1733
##   CENTER_TO_CENTER_DISTANCE_FROM_FEATURE_TO_DATA_POINT ncbi_gene_id
## 1                                                  691        26155
## 2                                                -1978       339451
## 4                                                -2846        84069
## 5                                                -1782         9636
## 6                                                -5020       254173
## 8                                                -2108        51150
##      Name                                          description
## 1   NOC2L nucleolar complex associated 2 homolog (S. cerevisiae)
## 2  KLHL17                             kelch-like 17 (Drosophila)
## 4 PLEKHN1 pleckstrin homology domain containing, family N member 1
## 5   ISG15                            ISG15 ubiquitin-like modifier
## 6  TTLL10          tubulin tyrosine ligase-like family, member 10
## 8    SDF4                               stromal cell derived factor 4
##         synonyms
## 1  DKFZp564C186
## 2   RP11-54O7.6
## 4 DKFZp434H2010
## 5          G1P2
## 6      FLJ36119
## 8         Cab45
```

```
dim(cancer.no.dup)
```

```
## [1] 18897    15
```

```
cancer.no.dup[grep("SMAD1",cancer.no.dup$Name),]
```

```
##             DATA_INDEX CHROMOSOME DATA_START   DATA_END Peaks
##   41592         114921            chr4   146620835 146620995 2.087
##                               FEATURE_TRACK FEATURE_STRAND FEATURE_START
                                  FEATURE_END
##   41592  transcription_start_site                      +          146623406        146623406
##             SHORTEST_DISTANCE_FROM_FEATURE_TO_DATA_POINT
##   41592                                                 -2411
##             CENTER_TO_CENTER_DISTANCE_FROM_FEATURE_TO_DATA_POINT
            ncbi_gene_id
##  41592                                                          -2496               4086
##               Name                        description synonyms
##  41592 SMAD1           family member 1              BSP1
   SMAD
```

```
normal.no.dup = subset(control.probes.with.high.freq.genes,!duplicated(control.probes.with.high.freq.genes$n
cbi_gene_id))
table(normal.no.dup$CHROMOSOME)
```

```
##
##           chr1          chr10          chr11          chr12          chr13
##           1797            472            714            635            234
##          chr14          chr15          chr16          chr17          chr18
##            420            429            370            595            191
##          chr19           chr2          chr20          chr21          chr22
##            262           1095            192            124            116
##           chr3           chr4           chr5           chr6           chr7
##            941            638            653            659            562
##           chr8           chr9           chrX           chrY   chr13_random
##            392            444            284             44              0
##   chr17_random    chr6_random    chrX_random    chr1_random   chr19_random
##             11              1              0              0              0
##   chr21_random    chr3_random    chr5_random    chr9_random
##              0              0              0              1
```

```
head(normal.no.dup)
```

```
##       DATA_INDEX CHROMOSOME DATA_START DATA_END Peaks
## 1          32387       chr1     861090   861217 2.157
## 2          32388       chr1     867686   868029 2.482
## 4          32390       chr1     885728   893410 4.064
## 8          32391       chr1     896609   898064 3.055
## 13         32391       chr1     896609   898064 3.055
## 25         32394       chr1     934947   935192 2.295
##                  FEATURE_TRACK FEATURE_STRAND FEATURE_START
FEATURE_END
## 1  transcription_start_site              +        861120      861120
## 2  transcription_start_site              +        871145      871145
## 4  transcription_start_site              +        895966      895966
## 8  transcription_start_site              +        901876      901876
## 13 transcription_start_site              -        894679      894679
## 25 transcription_start_site              -        935552      935552
##    SHORTEST_DISTANCE_FROM_FEATURE_TO_DATA_POINT
## 1                                             0
## 2                                         -3116
## 4                                         -2556
## 8                                         -3812
## 13                                        -1930
## 25                                          360
##    CENTER_TO_CENTER_DISTANCE_FROM_FEATURE_TO_DATA_POINT
   ncbi_gene_id
## 1                                                    33      148398
## 2                                                 -3287          NA
## 4                                                 -6397      339451
## 8                                                 -4539       84069
## 13                                                -2657       26155
## 25                                                  482       57801
##       Name                                          description
## 1    SAMD11                sterile alpha motif domain containing 11
## 2      cmpl
## 4    KLHL17                               kelch-like 17 (Drosophila)
## 8  PLEKHN1 pleckstrin homology domain containing, family N member 1
## 13   NOC2L    nucleolar complex associated 2 homolog (S. cerevisiae)
## 25    HES4                    hairy and enhancer of split 4 (Drosophila)
##          synonyms
## 1       MGC45873
## 2
## 4     RP11-54O7.6
## 8  DKFZp434H2010
## 13  DKFZp564C186
## 25       bHLHb42
```

```
dim(normal.no.dup)
```

```
## [1] 12276    15
```

```
normal.no.dup[grep("HOXA9",normal.no.dup$Name),]
```

```
##    [1]  DATA_INDEX
##    [2]  CHROMOSOME
##    [3]  DATA_START
##    [4]  DATA_END
##    [5]  Peaks
##    [6]  FEATURE_TRACK
##    [7]  FEATURE_STRAND
##    [8]  FEATURE_START
##    [9]  FEATURE_END
##   [10]  SHORTEST_DISTANCE_FROM_FEATURE_TO_DATA_POINT
##   [11]  CENTER_TO_CENTER_DISTANCE_FROM_FEATURE_TO_DATA
        _POINT
##   [12]  ncbi_gene_id
##   [13]  Name
##   [14]  description
##   [15]  synonyms
## <0 rows> (or 0-length row.names)
```

```
common_genes = BiocGenerics::intersect(normal.no.dup$Name,cancer.no.dup$Name)
tail(common_genes,100)
```

```
##     [1] "KRTAP19-1"   "KRTAP19-4"   "KRTAP19-5"   "KRTAP6-3"    "KRTAP6-1"
##     [6] "KRTAP20-4"   "KRTAP20-2"   "KRTAP20-3"   "KRTAP21-1"   "KRTAP7-1"
##    [11] "KRTAP11-1"   "KRTAP19-8"   "IFNAR2"      "DONSON"      "CRYZL1"
##    [16] "KCNE2"       "FAM165B"     "KCNE1"       "RCAN1"       "TTC3"
##    [21] "DSCR3"       "NCRNA00114"  "PSMG1"       "LCA5L"       "SH3BGR"
##    [26] "C21orf88"    "FAM3B"       "ABCG1"       "TFF1"        "AGPAT3"
##    [31] "TRPM2"       "KRTAP10-5"   "KRTAP10-11"  "KRTAP12-2"   "KRTAP12-1"
##    [36] "C21orf29"    "ITGB2"       "C21orf67"    "C21orf70"    "NCRNA00162"
##    [41] "LOC642852"   "psiTPTE22"   "UFD1L"       "CDC45"       "ZNF280A"
##    [46] "RTDR1"       "RAB36"       "C22orf43"    "SLC2A11"     "C22orf13"
##    [51] "SNRPD3"      "PIWIL3"      "HPS4"        "SRRD"        "TFIP11"
##    [56] "TTC28AS"     "XBP1"        "RFPL1"       "ASCC2"       "SF3A1"
##    [61] "CCDC157"     "C22orf27"    "RNF185"      "LIMK2"       "C22orf24"
##    [66] "RFPL3S"      "SYN3"        "TOM1"        "MB"          "APOL3"
##    [71] "MYH9"        "TXN2"        "EIF3D"       "PVALB"       "TMPRSS6"
##    [76] "PLA2G6"      "LOC400927"   "JOSD1"       "ENTHD1"      "GRAP2"
##    [81] "TNRC6B"      "SLC25A17"    "XPNPEP3"     "ST13"        "ARFGAP3"
##    [86] "PACSIN2"     "LDOC1L"      "NCRNA00207"  "SMC1B"       "MLC1"
##    [91] "MOV10L1"     "LMF2"        "NCAPH2"      "RPL23AP82"   "ASMTL"
##    [96] "XG"          "XGPY2"       "ARSE"        "NLGN4X"      "VCX3A"
```

```
# Exclusive genes from both sets of Cancer and Normal files
cancer.genes.exclusive = cancer.no.dup[!(cancer.no.dup$Name %in% common_genes), ]
dim(cancer.genes.exclusive)
```

```
## [1] 7248   15
```

```
table(cancer.genes.exclusive$CHROMOSOME)
```

```
##
##         chr1  chr1_random        chr10        chr11        chr12
##          307            0          277          559          363
##        chr13 chr13_random        chr14        chr15        chr16
##           95            0          220          230          421
##        chr17 chr17_random        chr18        chr19 chr19_random
##          519            1           81          952            5
##         chr2        chr20        chr21 chr21_random        chr22
##          168          318          116            0          291
##         chr3  chr3_random         chr4  chr4_random         chr5
##          151            0          104            0          229
##         chr6         chr7         chr8         chr9         chrX
##          375          358          261          334          472
##  chrX_random         chrY  chr2_random  chr5_random chr16_random
##            0           41            0            0            0
##  chr6_random chr15_random  chr9_random chr11_random chr22_random
##            0            0            0            0            0
```

```
cancer.genes.exclusive[grep("CDKN2A",cancer.genes.exclusive$Name),]
```

```
##       DATA_INDEX CHROMOSOME DATA_START DATA_END Peaks
## 71913     116631       chr9   21966746 21967303 3.437
##                    FEATURE_TRACK FEATURE_STRAND FEATURE_START
                   FEATURE_END
## 71913 transcription_start_site              -      21965038      21965038
##       SHORTEST_DISTANCE_FROM_FEATURE_TO_DATA_POINT
## 71913                                        -1708
##       CENTER_TO_CENTER_DISTANCE_FROM_FEATURE_TO_DATA_POINT
      ncbi_gene_id
## 71913                                                 -1986         1029
##         Name
## 71913 CDKN2A
##                                                          description
## 71913 cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)
##       synonyms
## 71913      ARF
```

```
write.table(cancer.genes.exclusive,file = "objs/cancer.genes.exclusive.txt",sep="\t")


normal.genes.exclusive = normal.no.dup[!(normal.no.dup$Name %in% common_genes), ]
dim(normal.genes.exclusive)
```

```
## [1] 627  15
```

```
table(normal.genes.exclusive$CHROMOSOME)
```

```
##
##          chr1        chr10        chr11        chr12        chr13
##           118            6           19           18           12
##         chr14        chr15        chr16        chr17        chr18
##             5            8            8           28            8
##         chr19         chr2        chr20        chr21        chr22
##            45           70           32           17           17
##          chr3         chr4         chr5         chr6         chr7
##            49           35           36           26           19
##          chr8         chr9         chrX         chrY  chr13_random
##             7           11           28            4            0
##  chr17_random  chr6_random  chrX_random  chr1_random  chr19_random
##             1            0            0            0            0
##  chr21_random  chr3_random  chr5_random  chr9_random
##             0            0            0            0
```

```
normal.genes.exclusive[grep("SMAD1",normal.genes.exclusive$Name),]
```

```
##  [1] DATA_INDEX
##  [2] CHROMOSOME
##  [3] DATA_START
##  [4] DATA_END
##  [5] Peaks
##  [6] FEATURE_TRACK
##  [7] FEATURE_STRAND
##  [8] FEATURE_START
##  [9] FEATURE_END
## [10] SHORTEST_DISTANCE_FROM_FEATURE_TO_DATA_POINT
## [11] CENTER_TO_CENTER_DISTANCE_FROM_FEATURE_TO_DATA
         _POINT
## [12] ncbi_gene_id
## [13] Name
## [14] description
## [15] synonyms
## <0 rows> (or 0-length row.names)
```

```
write.table(normal.genes.exclusive,file = "objs/normal.genes.exclusive.txt",sep="\t")
```