Comprehensive Molecular Characterization of Gastric Adenocarcinoma

Supplementary Materials

Supplementary Material Table of Contents: Comprehensive Molecular Characterization of Gastric Adenocarcinoma

- continued -

Supplementary Material Table of Contents: Comprehensive Molecular Characterization of Gastric Adenocarcinoma

- continued -

Supplementary Material Table of Contents: Comprehensive Molecular Characterization of Gastric Adenocarcinoma

S1. Biospecimen collection, pathological data, and clinical data

S1 Section Authors:

Alex Boussioutas
Katherine S. Garman
Joseph E. Willis
Jihun Kim
M. Blanca Piazuelo
Adam J. Bass
Barbara G. Schneider
Margaret L. Gulley
Tara M. Lichtenberg
Kristen M. Leraas
Stephanie Weaver
Jay Bowen
John A. Demchok
Vésteinn Thorsson

Subsections:

S1.1 Biospecimen Collection, Quality Control, and Processing

Sample Acquisition:
Gastric tumours were collected and shipped to a central Biospecimen Core Resource (BCR) between 5/6/2010 and 10/17/2012. Before shipment, clinical information and sample descriptions were submitted to the BCR. Qualifying tumour samples were obtained from patients who had received no prior treatment for their disease (chemotherapy or radiotherapy). Specimens were shipped overnight from 13 tissue source sites (TSS) using a cryoport that maintained an average temperature of less than -180°C. TSSs contributing biospecimens included: Cureline, Inc.; Asterand, Inc.; ILSbio, LLC.; Indivumed GmbH; Greater Poland Cancer Center; Christiana Care Health Services, Inc.; University of North Carolina; International Genomics Consortium; Ontario Institute of Cancer Research; Roswell Park Cancer Institute; National Cancer Center Research Institute (Korea); University of Pittsburgh; and Analytical Biological Services, Inc.

In addition to tumour samples, each frozen primary tumour specimen had a companion normal tissue specimen (blood or blood components, including DNA extracted at the tissue source site). Adjacent non-tumour gastric tissue was also submitted for a subset of cases.

Cases were staged according to the American Joint Committee on Cancer (AJCC) staging system 7th Edition. Pathology quality control was performed on each tumour and adjacent normal tissue (if available) specimen from either a frozen section slide prepared by the BCR or from a frozen section slide prepared by the TSS. Hematoxylin and eosin (H&E) stained sections from each sample were subjected to independent pathology review to confirm that the tumour specimen was histologically consistent with gastric cancer and the adjacent tissue specimen contained no tumour cells. The percent tumour nuclei, percent necrosis, and other pathology annotations were also assessed. Tumour samples with ≥60% tumour nuclei and ≤20% necrosis were submitted for nucleic acid extraction. An exception to tumour nuclei criteria was made for diffuse tumour types.

Sample Processing:
DNA and RNA were extracted and quality was assessed at the central BCR. RNA and DNA were extracted from tumour and adjacent non-tumour tissue specimens using a modification of the DNA/RNA AllPrep kit (Qiagen). The flow-through from the Qiagen DNA column was processed using a *mir*Vana miRNA Isolation Kit (Ambion). This latter step generated RNA preparations that included RNA <200 nt suitable for miRNA analysis. DNA was extracted from blood using the QiaAmp DNA Blood Midi kit (Qiagen).

RNA samples were quantified by measuring Abs_{260} with a UV spectrophotometer and DNA quantified by PicoGreen assay. DNA specimens were resolved by 1% agarose gel electrophoresis to confirm high molecular weight fragments. A custom Sequenom SNP panel or the AmpFISTR Identifiler (Applied Biosystems) was utilized to verify that tumour DNA and germline DNA representing a case were derived from the same patient. Five hundred nanograms of each tumour and germline DNA were sent to Qiagen (Hilden, Germany) for REPLI-g whole genome amplification using a 100µg reaction scale. RNA was analyzed via the RNA6000 Nano assay (Agilent) for determination of an RNA Integrity Number (RIN), and only analytes with RIN ≥7.0 were included in this study. Only cases yielding a minimum of 6.9 µg of tumour DNA, 5.15 µg RNA, and 4.9 µg of germline DNA were included in this study.

Sample Qualification:
The BCRs received tumour samples with germline controls from a total of 618 cases, of which 343 samples qualified and were sent for further genomic analysis. Of the 275 samples that were disqualified, 114 cases failed pathology screening, 117 cases failed due to molecular criteria, 13 failed due to genotype mismatch between tumour and germline, 10 represented unacceptable histology, three had insufficient tumour, and 18 did not qualify for other reasons. Of the 114 that failed pathologic criteria, 111 failed for insufficient tumour nuclei (<60%), one failed for necrosis > than 20%, and two cases were disqualified for both insufficient tumour nuclei and necrosis. Regarding the exception made for the diffuse tumour type, in the final data set, 11 diffuse gastric tumours were included with tumour nuclei <60%. Of the 117 samples that failed molecular screening, the majority (n=97) had RNA integrity scores of < 7.0. Other causes of failure included low DNA or RNA yield, or evidence of low molecular weight DNA observed in the DNA gel, indicating DNA fragmentation.

Of the 343 samples that qualified based upon BCR pathology review and molecular characteristics, 295 samples were ultimately used for genomic analysis. The 48 samples that were not used in the final analysis

were redacted following independent pathology review (as described below), because of discovery of clinical disqualifiers (three patients with history of chemotherapy), unknown subject identity or case duplication. See flow diagram, Supplementary Figure S1.2, for sample qualification.

Of the qualifying cases, matched non-tumour stomach tissue was available from 73 cases. Samples with residual tumour tissue following extraction of nucleic acids were considered for proteomics analysis. When available, a 10 to 20 mg piece of snap-frozen tumour adjacent to the piece used for molecular sequencing and characterization was submitted to MD Anderson for reverse phase protein array (RPPA) analysis.

Microsatellite Instability Assay

Microsatellite instability (MSI) in qualified cases was evaluated by the BCR at Nationwide Children's Hospital. MSI-Mono-Dinucleotide Assay was performed to test a panel of four mononucleotide repeat loci (polyadenine tracts BAT25, BAT26, BAT40, transforming growth factor receptor type II), and three dinucleotide repeat loci (CA repeats in D2S123, D5S346, and D17S250). Two additional pentanucleotide loci (Penta D and Penta E) were included in this assay to evaluate sample identity. If variation in the number of microsatellite repeats was detected between tumour and matched non-neoplastic tissue or mononuclear blood cells, multiplex fluorescent-labeled PCR and capillary electrophoresis were used to confirm MSI. Equivocal or failed markers were re- evaluated by singleplex PCR. Tumour DNA was classified as microsatellite-stable (MSS) if zero markers were altered, low level MSI (MSI-L) if less than 40% of markers were altered, and high level MSI if greater than 40% of markers were altered. In the MSI-Mono-Dinucleotide Assay, samples were designated MSI-L if one or two markers were altered and MSI if three to seven markers were altered.

Individual markers were assigned a value of 1 through 6, based on the presence or absence of a MSI shift, allele homo/heterozygosity and loss of heterozygosity (LOH) if relevant. Markers that demonstrated MSI shift were classified as follows; 1= homozygous alleles, 2= heterozygous alleles with LOH, and 3= heterozygous alleles without LOH. Markers that did not demonstrate a MSI shift were classified as follows; 4= homozygous alleles, 5= heterozygous alleles with LOH, and 6= heterozygous alleles without LOH. Penta D and E markers were scored in the same manner as the MSI markers; however they did not contribute to MSI Class calculation.

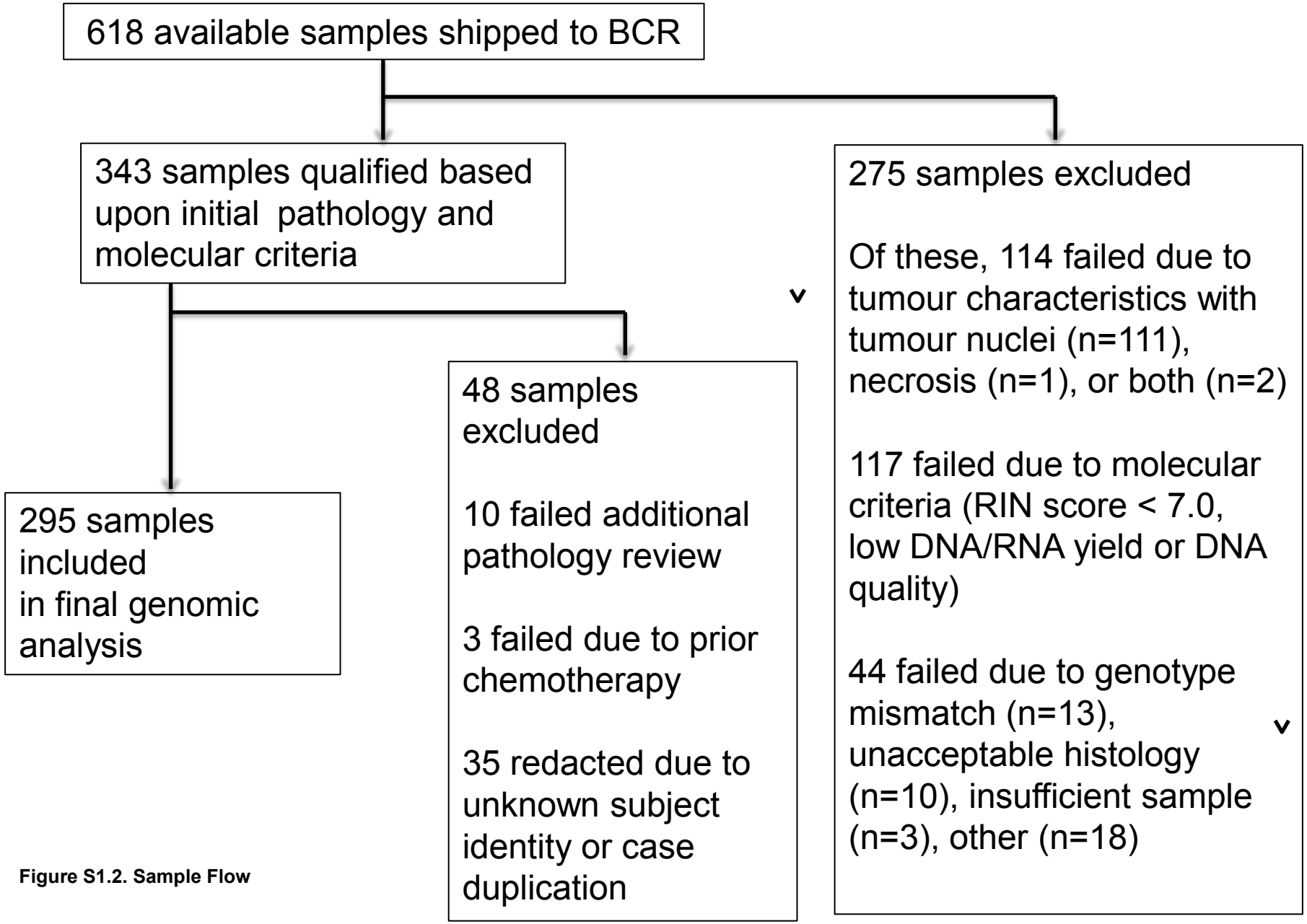


Figure S1.2. Sample Flow

Figure S1.2 Sample Flow Of the 618 samples shipped to the BCR, 295 were included in the final genomic analysis. For samples that did not meet inclusion criteria, causes of exclusion such as failure to meet quality metrics or pathology review are demonstrated here.

S1.3 text- Review of pathology, TNM stage, anatomic site, and tumour recurrence

Tumour TNM Stage (AJCC Stage) Review

All stage assignments were evaluated for consistency with the 7th edition of the AJCC Cancer Staging Manual, in order to fill missing entries, to identify errors, and to update stage annotations based on older editions of the manual. For cases that were submitted to the BCR with annotation from a prior edition, tumour characteristics were remapped to the 7th edition. The mapping was done using a table provided by the Collaborative Stage Data Collection System (http://web2.facs.org/cstage0204/stomach/Stomach_uam.html). Prior to this step, the number of positive lymph nodes was used to assign N0, N1, N2, N3, and for the 6th edition annotations, T was converted to be consistent with 7th edition definitions. Manual review was then performed, with a particular focus on cases where there were contradictions between mapped and tabulated stage values.

Anatomic Site Review

Anatomic location of tumours was verified through review of the primary de-identified pathology reports. Expert TCGA pathology review of tumour and surrounding tissue provided additional information about tumour location in 14 cases. Using this method, a location could be assigned for 97% (287 of 295) of tumours.

Pathology review

All cancers included in this study were reviewed by an Expert Pathologist Committee (EPC) that consisted of four experienced gastrointestinal pathologists (J.K., J.W., M.G., and M.P.). A centralized virtual pathology review system utilizing an Aperio slide scanner housed at the BCR at Nationwide Children's Hospital, Columbus OH, was constructed. Representative virtual slide images of each case were deposited on a web server and reviewed electronically by EPC members. Typically, two frozen sections flanking the tumour tissue from which all material was extracted for this study and one additional high-quality formalin-fixed paraffin-embedded tissue section were scanned. GCs were classified using the Lauren classification categories: (Intestinal, Diffuse, Mixed, Indeterminate), and the World Health Organization classification (Papillary, Tubular, Mucinous, Poorly Cohesive and other rare variants). All EPC members were well informed about the classification criteria before the review process. Two EPC members reviewed all cases without a pathological tumour classification from the tissue source site and one EPC member reviewed the cases that had a prior pathological tumour classification from the tissue source site. Uncommon cancer subtypes such as neuroendocrine carcinoma, malignant lymphoma, squamous cell carcinoma and gastrointestinal stromal tumour were identified and excluded after agreement by two EPC members. In most cases (71%), a consensus was reached after the first set of reviews. For cases with discrepant results, a tie-breaker reviewer was assigned. For rare cases upon which no consensus was reached after tie-breaker's review, all EPC members joined to reach consensus.

Review of Recurrence and Survival

Tissue source sites were asked to provide clinical follow-up data in periodic updates. Recurrence was defined as clinical or radiological evidence of recurrence. This was recorded as "YES" if there was definite recurrence or "NO" if there was no evidence at recurrence during the follow-up period. Cases where there were insufficient data to make a conclusion were set to "NA". Survival was determined by known date of death and data were censored based upon last known clinical follow-up. Clinical data are listed in Supplementary Table S11.1a.

S1.4 Correlation of Clinical Data with Molecular Subtypes

The division of pathological and clinical data by the four molecular subtypes is shown in Supplementary Table S1.5. In Supplementary Table S1.6, we highlight the statistically significant associations between Molecular Subtypes and clinical data categories. This table also includes associations with some other key variables discussed in the manuscript. Due to space limitations, only a limited set of sample divisions could be accommodated in the Tables. A more extensive set of significantly associated variables can be explored using the Regulome Explorer Web tool (Supplement S11.8, explorer.cancerregulome.org).

We also evaluated associations of Molecular Subtypes with both overall survival and tumour recurrence. All event times were defined with respect to time of surgical resection. Of the 295 participants, 57 were known to be deceased and 230 had follow-up (including deceased). The number of participants with greater than one year follow-up was 113; and 21, 8, 7, and 5 participants had more than two, three, four, and five years of follow-up, respectively. The median follow-up was 362.5 days. Recurrence information was available for 201 participants, with 31 experiencing recurrence. In Supplementary Figure S1.7, we show Kaplan-Meier curves for overall survival and for recurrence, for groups with defined TNM Stage, and for grouping by Molecular Subtype. Curves were drawn using the *survfit* function in R, and *p*-values were evaluated using Cox proportional hazards with a log-likelihood test (*coxph* function). The small number of samples and relatively short duration of follow-up for most cases posed limitations for drawing conclusions about survival and recurrence.

Supplementary Table S1.5 Fractional Division of Clinical Parameters by Molecular Subtype

Age*	**N**	**Mean**	**Range**	**%**						
FEMALE	112	67	34-90	38.4%						
MALE	180	65.27	39-90	61.6%						
	Total		**Molecular Subtype EBV**		**MSI**		**GS**		**CIN**	
Gender	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**	**n**	**%**
FEMALE	113	38.3%	5	4.4%	36	31.9%	22	19.5%	50	44.2%
MALE	182	61.7%	21	11.5%	28	15.4%	36	19.8%	97	53.3%
Lauren Class										
Diffuse	69	23.4	5	7.2%	6	8.7%	40	58.0%	18	26.1%
Intestinal	196	66.4	15	7.7%	48	24.5%	15	7.7%	118	60.2%
Mixed	19	6.4	3	15.8%	3	15.8%	3	15.8%	10	52.6%
Not specified	11	3.7	3	27.3%	7	63.6%	0	0.0%	1	9.1%
WHO Class										
Mixed	19	6.4	3	15.8%	3	15.8%	3	15.8%	10	52.6%
Mucinous	18	6.1	0	0.0%	7	38.9%	2	11.1%	9	50.0%
Papillary	22	7.5	1	4.5%	4	18.2%	2	9.1%	15	68.2%
Poor cohesive	69	23.4	5	7.2%	6	8.7%	40	58.0%	18	26.1%
Tubular	140	47.5	9	6.4%	35	25.0%	9	6.4%	87	62.1%
Not specified	27	9.2	8	29.6%	9	33.3%	2	7.4%	8	29.6%
Pathologic T										
T1A	1	0.3	0	0.0%	1	100.0%	0	0.0%	0	0.0%
T1B	10	3.4	1	10.0%	5	50.0%	0	0.0%	4	40.0%
T2	44	14.9	1	2.3%	9	20.5%	11	25.0%	23	52.3%
T3	155	52.5	15	9.7%	30	19.4%	28	18.1%	82	52.9%
T4	1	0.3	0	0.0%	0	0.0%	1	100.0%	0	0.0%
T4A	60	20.3	9	15.0%	8	13.3%	15	25.0%	28	46.7%
T4B	14	4.7	0	0.0%	7	50.0%	1	7.1%	6	42.9%
TX	10	3.4	0	0.0%	4	40.0%	2	20.0%	4	40.0%
Pathologic N										
N0	97	32.9	6	6.2%	26	26.8%	19	19.6%	46	47.4%
N1	64	21.7	6	9.4%	15	23.4%	8	12.5%	35	54.7%
N2	58	19.7	5	8.6%	7	12.1%	18	31.0%	28	48.3%
N3	65	22	9	13.8%	11	16.9%	11	16.9%	34	52.3%
NX	11	3.7	0	0.0%	5	45.5%	2	18.2%	4	36.4%
Pathologic M										
M0	273	92.5	23	8.4%	61	22.3%	52	19.0%	137	50.2%
M1	20	6.8	3	15.0%	2	10.0%	6	30.0%	9	45.0%
MX	2	0.7	0	0.0%	1	50.0%	0	0.0%	1	50.0%
AJCC stage										
Stage IA	8	2.7	1	12.5%	3	37.5%	0	0.0%	4	50.0%
Stage IB	24	8.1	1	4.2%	7	29.2%	4	16.7%	12	50.0%
Stage IIA	56	19	4	7.1%	14	25.0%	11	19.6%	27	48.2%
Stage IIB	60	20.3	5	8.3%	14	23.3%	12	20.0%	29	48.3%
Stage IIIA	40	13.6	2	5.0%	2	5.0%	10	25.0%	26	65.0%
STAGE_IIIB	57	19.3	8	14.0%	10	17.5%	10	17.5%	29	50.9%
STAGE_IIIC	14	4.7	2	14.3%	5	35.7%	2	14.3%	5	35.7%
STAGE_IV	20	6.8	3	15.0%	2	10.0%	6	30.0%	9	45.0%
X	16	5.4	0	0.0%	7	43.8%	3	18.8%	6	37.5%
Country of origin										
CANADA	3	1	0	0.0%	1	33.3%	0	0.0%	2	66.7%
GERMANY	39	13.2	1	2.6%	9	23.1%	6	15.4%	23	59.0%
KOREA_SOUTH	31	10.5	4	12.9%	11	35.5%	2	6.5%	14	45.2%
POLAND	32	10.8	4	12.5%	2	6.2%	6	18.8%	20	62.5%
RUSSIA	83	28.1	6	7.2%	26	31.3%	15	18.1%	36	43.4%
UKRAINE	39	13.2	6	15.4%	6	15.4%	14	35.9%	13	33.3%
UNITED_STATES	24	8.1	2	8.3%	5	20.8%	2	8.3%	15	62.5%
VIETNAM	44	14.9	3	6.8%	4	9.1%	13	29.5%	24	54.5%
Anatomical region										
ANTRUM	114	38.6	6	5.3%	31	27.2%	28	24.6%	49	43.0%
FUNDUS_BODY	116	39.3	16	13.8%	25	21.6%	18	15.5%	57	49.1%
GEJ_CARDIA	57	19.3	4	7.0%	5	8.8%	11	19.3%	37	64.9%
NA	8	2.7	0	0.0%	3	37.5%	1	12.5%	4	50.0%
Survival	**n**	**days**	**n**	**days**	**n**	**days**	**n**	**days**	**n**	**days**
Death	57	399.5	5	334.2	11	419.2	12	374.2	29	413.9
Tumour Recurrence	31	387.7	4	354	3	348.3	4	671.8	20	337

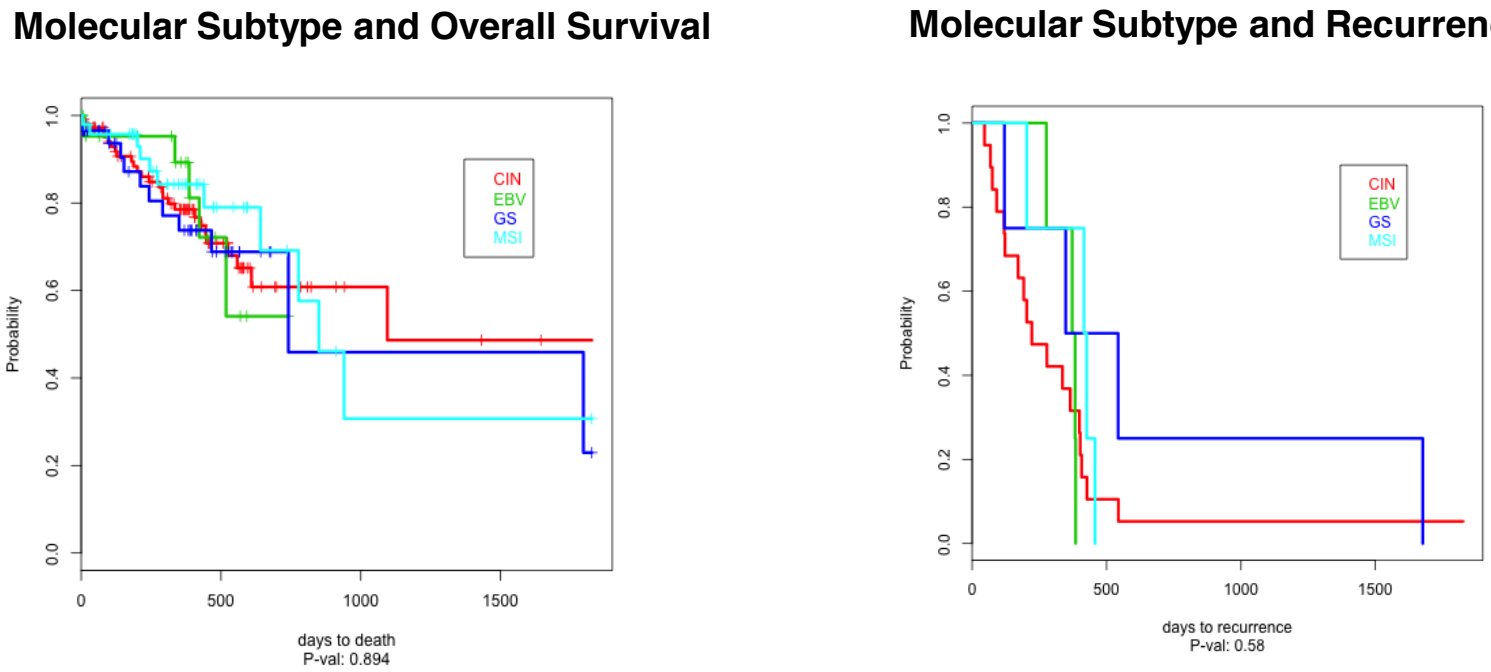
*Three missing fields for age

Supplementary Table S1.6 Statistical Association with Molecular Subtypes

Data Variable	All Subtypes	EBV	MSI	CIN	GS
Lauren Classification	6.57E-15	0.404	0.0117	4.99E-06	1.40E-16
Intestinal Subclass	0.646	0.845	0.294	0.51	0.526
Signet Ring	0.245	0.169	1	0.386	0.119
WHO Classification	0.00577	0.443	0.0139	1.49E-05	5.79E-15
Pathologic M	0.454	0.503	0.191	0.908	0.443
Pathologic N	0.831	0.486	0.0516	0.771	0.138
Pathologic T	0.831	0.345	0.852	0.496	0.707
TNM Stage	0.759	0.332	0.414	0.644	0.46
Anatomic Site	0.762	0.458	0.0547	0.0971	0.519
Neoplasm Cancer Status	0.091	0.29	0.0495	0.122	0.568
Residual Tumour	0.664	0.748	0.0209	0.195	0.961
Age at Initial Diagnosis	8.06E-08	0.241	5.02E-05	0.164	3.65E-07
Country	0.0481	0.635	0.0677	0.516	0.0174
Gender	0.00329	0.0368	0.00128	0.151	1
Race	0.925	1	0.323	0.959	0.89
MutSig rate Total	5.10E-34	0.104	2.78E-33	0.000112	3.34E-11
MLH1 epigen. Silenced	2.27E-30	0.0171	1.65E-30	2.46E-09	4.90E-05
CDKN2A epig. Silenced	1.41E-20	6.01E-14	2.95E-05	1.60E-08	0.0014
MSI status	1.49E-63	0.00256	1.75E-66	1.33E-23	4.06E-07
Mutation Rate Category	1.21E-53	0.267	6.27E-54	2.94E-18	5.68E-07
Gene Expression Clust	0.00874	5.55E-08	0.0316	1.48E-05	1.43E-11
MicroRNA Expr. Clust	2.56E-05	0.0696	0.00818	0.553	2.19E-05
Copy Number Cluster	4.10E-73	0.000884	1.32E-16	4.81E-70	1.15E-23
Methylation Cluster	0.00487	2.89E-35	2.06E-25	2.59E-16	0.00067
EBV present	7.64E-38	7.64E-38	0.00202	8.95E-09	0.00371
TP53 mutation	3.01E-18	1.13E-06	0.156	3.39E-16	2.59E-08
PIK3CA mutation	5.89E-22	9.15E-12	1.67E-06	1.55E-14	0.0889
KRAS mutation	0.00078	0.487	0.000122	0.00534	1
BRAF mutation	3.06E-09	0.637	1.68E-08	9.28E-06	0.0483
RHOA mutation	0.00641	0.637	1	0.0103	0.00394
ERBB2 amplification	8.76E-17	0.566	1.13E-07	5.55E-16	1.51E-06
ARHGAP-CLDN18 Rearr.	0.00235	0.614	1	0.0196	0.000996
ABSOLUTE Ploidy	6.60E-08	0.587	0.00769	7.62E-08	9.86E-06
ABSOLUTE Purity	0.0875	0.843	0.0301	0.637	0.0651
Est. Leukocyte Pcnt.	2.01E-08	0.00873	0.472	1.64E-08	7.79E-06
Percent Tumour Nuclei	0.00209	0.681	0.0734	0.206	0.000184
Percent Tumour Cells	0.00147	0.184	0.252	0.00476	0.000925
Pcnt Lymphocyte Infiltr.	0.776	0.502	0.435	0.604	0.748

Statistical Association with Molecular Subtypes. All associations with $p < 0.05$ are highlighted. The first column applies to subtype divisions overall, and the remaining columns to comparison between samples in the subgroup and other samples. Red denotes an elevated or greater than expected number in that subtype. Blue denotes a lower value or fewer than expected. Other statistically significant associations (such as for non-ordered data including all the molecular subtypes) are highlighted in gray. The following two dichotomous variables were ordered alphabetically to define directionality: Gender: Female, Male; Copy Number Cluster: High, Low. Significant associations for other variables can be found using the Regulome Explorer web tool (Supplement S11.8, explorer.cancerregulome.org)

Figure S1.7. Patient Survival and Tumour Recurrence, Kaplan-Meier Curves



Note: At the time of this analysis, the follow-up time for this cohort remains limited; hence any survival analyses are quite exploratory. We continue to accrue follow-up data and it is possible that with additional time such differences will become apparent.

Figure S1.7. Patient Survival and Tumour Recurrence, Kaplan-Meier curves

Figure S1.7. Patient Survival and Tumour Recurrence, Kaplan-Meier curves. Kaplan-Meier survival curves demonstrated separation by tumour stage. For the four molecular subtypes, no significant differences in survival or recurrence rates were found. The CIN subtype showed some evidence for an elevated rate of recurrence, but not at a statistically significant level.

Supplement 1.8: Relations to Country of Origin

Given the known global diversity associated with gastric cancer, country of origin was included in the analysis of clinical parameters by molecular subtype (Supplement Table S1.5). For the most part, the frequency of molecular subtype within any given country was similar to that of our entire cohort, but some exceptions were observed. More GS tumours were identified in samples from the Ukraine (35.9%,p=0.0094) and from Vietnam (29.5%,p=0.097) as compared to an overall GS molecular subtype prevalence of 19.7%. Samples from Russia had more than expected MSI (31.3%, p=0.018), contrasted with an overall MSI prevalence of 21.7%. The South Korean cohort also had a somewhat greater than average fraction of MSI (35.5%, p=0.06) and less than expected GS (6.5%,p=0.06).

To evaluate other potential differences in gastric cancers between geographic regions, the two East Asian countries (Vietnam and South Korea) were combined into an East Asian group of 75 patients and compared to the remainder of the cohort. In the East Asian group, 68% of tumours were antral/pyloric compared with only 8% of tumours at the GE junction/cardia. These percentages are in contrast with tumours from the United States and Canada where tumours were more commonly found at the GE junction/cardia (48%) and less commonly located at the gastric antrum/pylorus (37%). The East Asian group of patients also presented on average at a somewhat younger age (mean 64 years), compared with age of presentation from the other regions (mean 67 years and p-value 0.12). Vietnam had the lowest mean age of presentation (mean 61 years, p=0.03) and Germany the highest (mean 72 years, p=0.001). When molecular subtypes were assessed in the East Asian group, EBV subtype was similar to the overall group (9.6%), as were MSI (20%), CIN (50%), GS (20%); there was thus no evidence for association between the East Asian group and molecular subtype (p=0.97).

Then, we evaluated geographic differences, again comparing the East Asian group to the other regions, in somatic tumour mutation rates of genes that had been identified in the somatic mutation analysis. These genes (*TP53, KRAS, ARID1A, PIK3CA, ERBB3, PTEN, HLA-B, RNF4B, B2M, NF1, APC, CTNNB1, SMAD4, SMAD2, RASA1, ERBB2, BCOR, CDH1,* and *RHOA*) were assessed for mutations in the East Asian population using Regulome Explorer (Supplement S11.8). No significant differences for any of the genes of interest were identified on the basis of East Asian origin.

Comparing the East Asian group to other regions, we evaluated differences in pathway-level gene expression changes (Figure 5c, Supplement S11.2) in the context of geographical distribution. Pathway expression for tumours from the East Asia group was for the most part similar to that of the remainder of the cohort. Exceptions were elevated expression in East Asian patients of pathways related to regulation of telomerase (p_s=2.0x10^{-5}, see S11.2), and decreased expression of HIF-1-alpha transcription factor network (p_s=2.4x10^{-4}). Other pathways were seen to differ in expression between individual countries. For example, Beta 1 integrin cell-surface interaction pathways had decreased expression in the South Korean tumours (p_s=5.0x10^{-6}), but the pathways were elevated in Vietnamese tumours (p_s=5.0x10^{-6}). Since increased expression of this pathway is seen in the GS subtype, this trend could be reflecting the relatively less GS in the Korean samples, as discussed above. IL-12 mediated signaling events, which are elevated in EBV samples, also had increased expression in samples from Russia (p_s=3.2x10^{-6}).

Overall, these data do not identify strong biologic differences between tumours of East Asian origin compared to other tumours. However, further analysis with larger sample cohorts will be required to better delineate differences that may exist between GC tumours originating in different regions of the world and in patients of different ethnic backgrounds.

S2. Copy number analysis

S2 Section Authors:

Andrew D. Cherniack
Bradley A. Murray
Gordon Saksena
Yingchun Liu
Carrie Sougnez

Subsections:

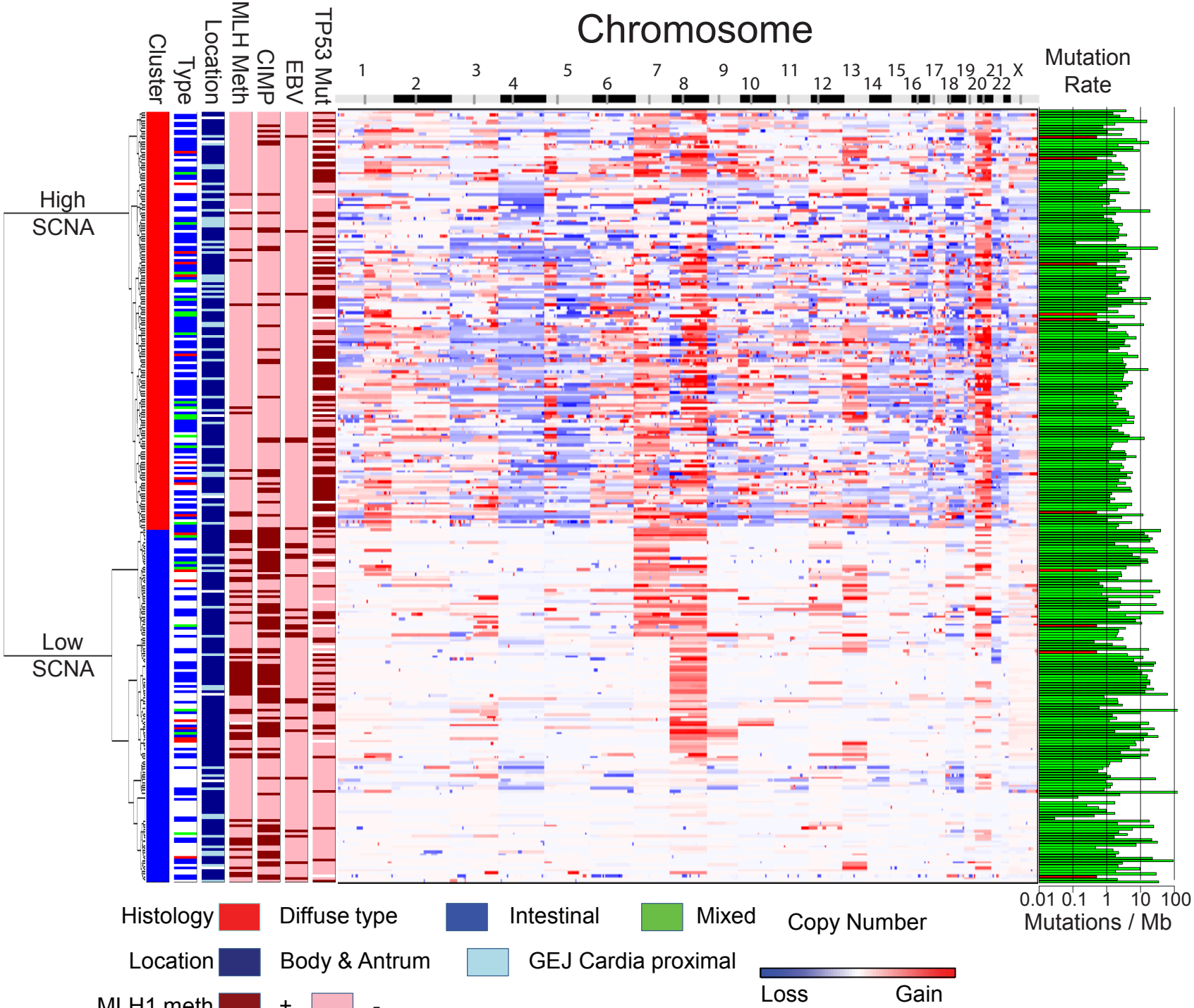
S2.1 Copy number analysis

SNP-based copy number analysis

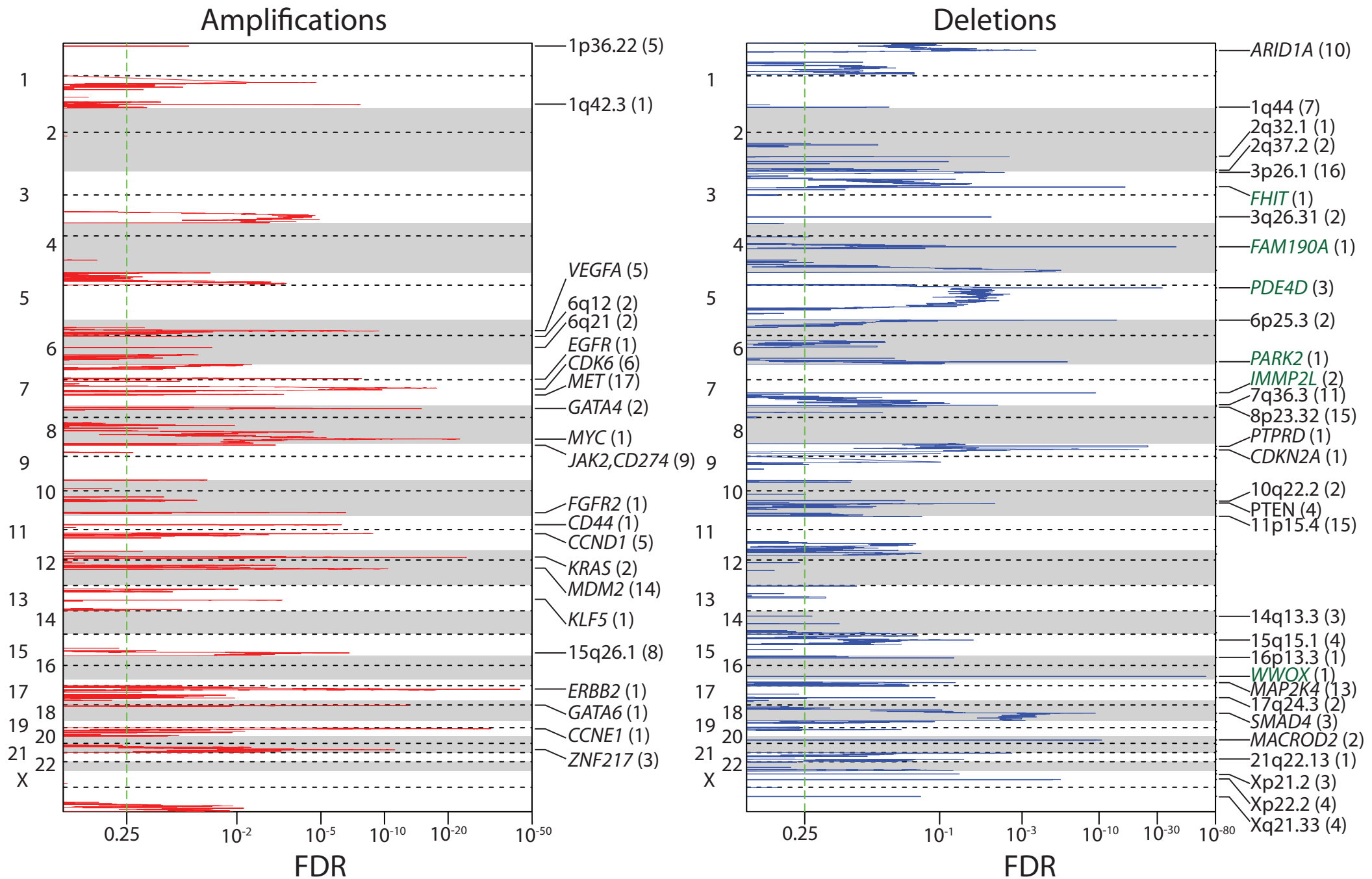
DNA from each tumour or germline sample was hybridized to Affymetrix SNP 6.0 arrays using protocols at the Genome Analysis Platform of the Broad Institute as previously described[1]. Briefly, from raw .CEL files, Birdseed was used to infer a preliminary copy-number at each probe locus[2]. For each tumour, genome-wide copy number estimates were refined using tangent normalization, in which tumour signal intensities are divided by signal intensities from the linear combination of all normal samples that are most similar to the tumour[3] (and Tabak B. and Beroukhim R. Manuscript in preparation). This linear combination of normal samples tends to match the noise profile of the tumour better than any set of individual normal samples, thereby reducing the contribution of noise to the final copy-number profile. Individual copy-number estimates then underwent segmentation using Circular Binary Segmentation[4]. As part of this process of copy-number assessment and segmentation, regions corresponding to germline copy-number alterations were removed by applying filters generated from either the TCGA germline samples from the ovarian cancer analysis or from samples from this collection. Segmented copy number profiles for tumour and matched control DNAs were analyzed using Ziggurat Deconstruction, an algorithm that parsimoniously assigns a length and amplitude to the set of inferred copy-number changes underlying each segmented copy number profile[5]. Significant focal copy number alterations were identified from segmented data using GISTIC 2.0[5]. For copy number based clustering, tumours were clustered based on thresholded copy number at reoccurring alteration peaks from GISTIC analysis (all_lesions.conf_99.txt file). Clustering was done in R based on Euclidean distance using Ward's method. In arm level analysis, chromosomal arms were considered altered if at least 66% of the arm was lost or gained with a log2 copy number change greater the 0.1. Allelic copy number, and purity and ploidy estimates were calculated using the ABSOLUTE algorithm[6].

References:

1. McCarroll, SA, et al. Integrated detection and population genetic analysis of SNPs and copy number variation. Nat Genet. 40:1166-1174 (2008).
2. Korn, JM, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet. 40: 1253-1260 (2008).
3. The Cancer Genome Atlas Research Network, Integrated genomic analyses of ovarian carcinoma. Nature 474:609-615 (2011).
4. Olshen, AB, et al. Circular binary segmentation for the analysis of array based DNA copy number data. Biostatistics 5: 557-572 (2004).
5. Mermel, CH, et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy number alteration in human cancers. Genome Bio. l12:R41 (2011).
6. Carter, SL, et al. Absolute quantification of somatic DNA alterations in human cancer. Nature 47: 609-615 (2011).

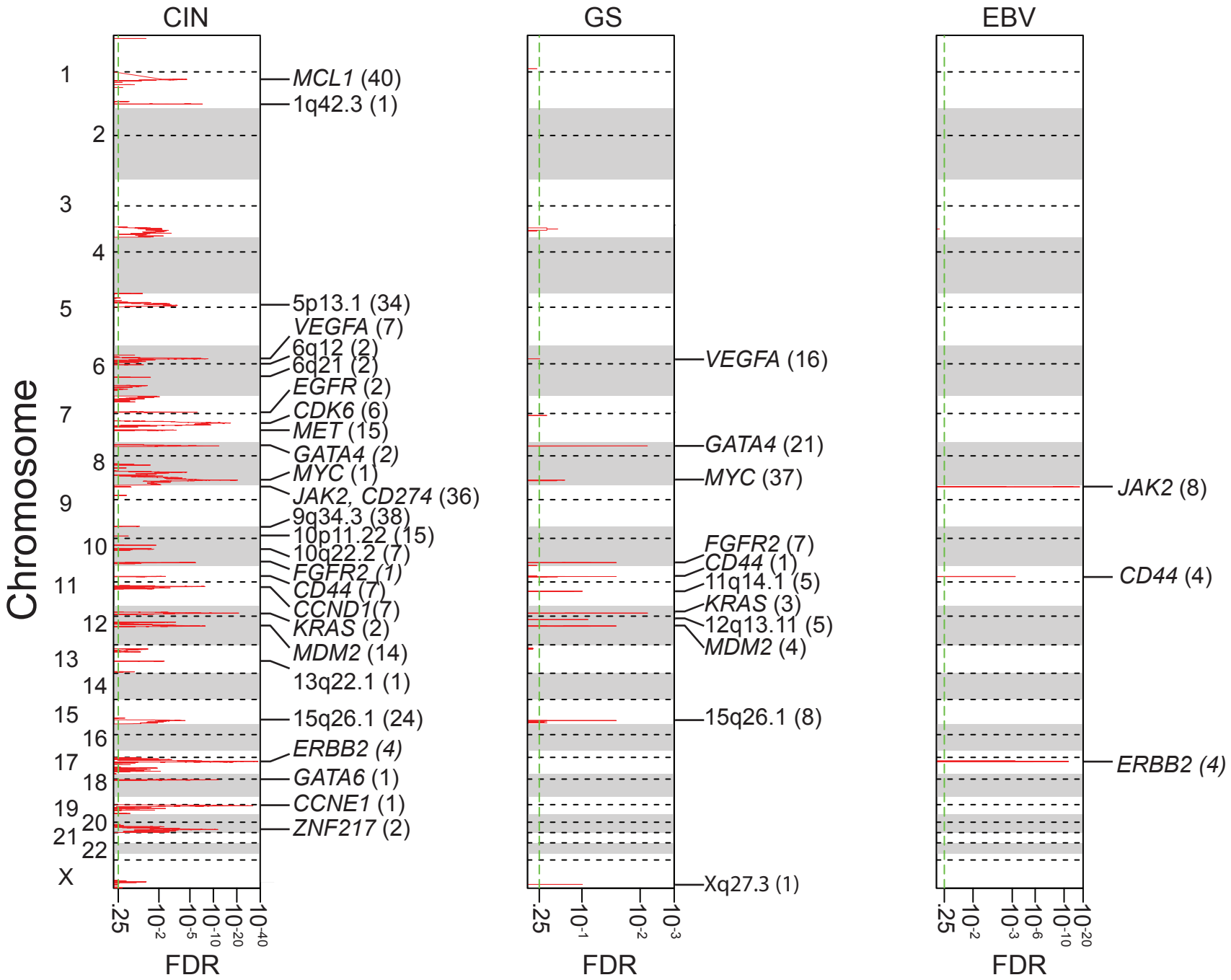


S2.2 . Somatic Copy Number Alterations. In the heatmap, SCNAs in tumors (vertical axis) are plotted by chromosomal location (horizontal axis). Tumors were hierarchical clustered by significantly reoccurring copy number alterations identified by GISTIC 2.0 analysis of the entire dataset. Vertical side bars (left) show the division of two major copy number cluster groups, histology, location, MLH methylation, CIMP and EBV status. Bar graphs (right) show the mutation rates in tumors.

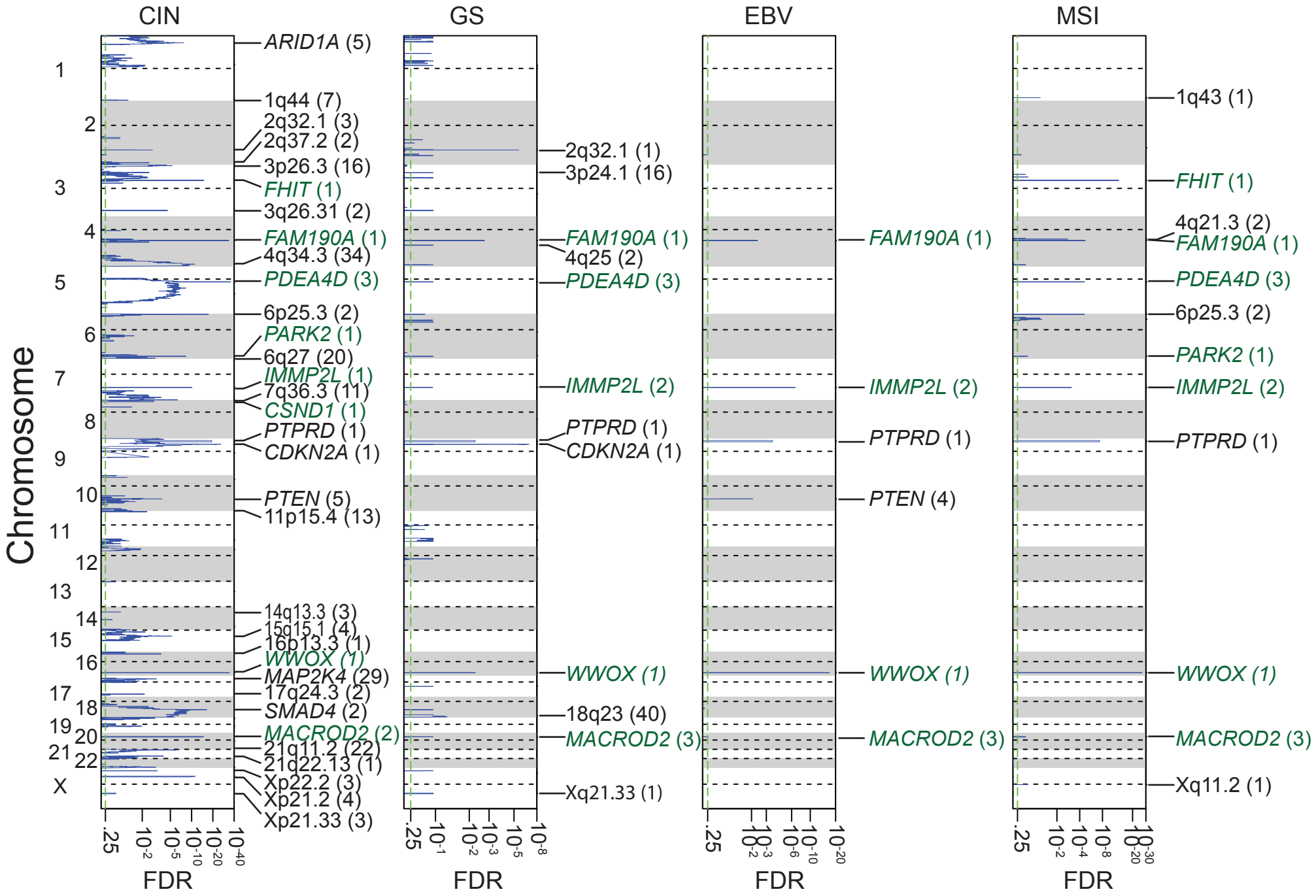


S2.3. GISTIC 2.0 amplifications and deletions. Chromosomal locations of peaks of significantly recurring focal amplifications (red) and deletions (blue) are plotted by false discovery rates. Annotated peaks have an FDR < .25 and encompass 16 or fewer genes. Peaks are annotated with candidate driver oncogenes, tumor suppressors, fragile site genes (green) or by cytoband. The number of genes within each peak is shown next to driver genes or cytobands.

S2.4 The data file- GISTIC 2.0 peaks can be found on the TCGA Stomach Adenocarcinoma publication page at https://tcga-data.nci.nih.gov/docs/publications/stad_2014/.

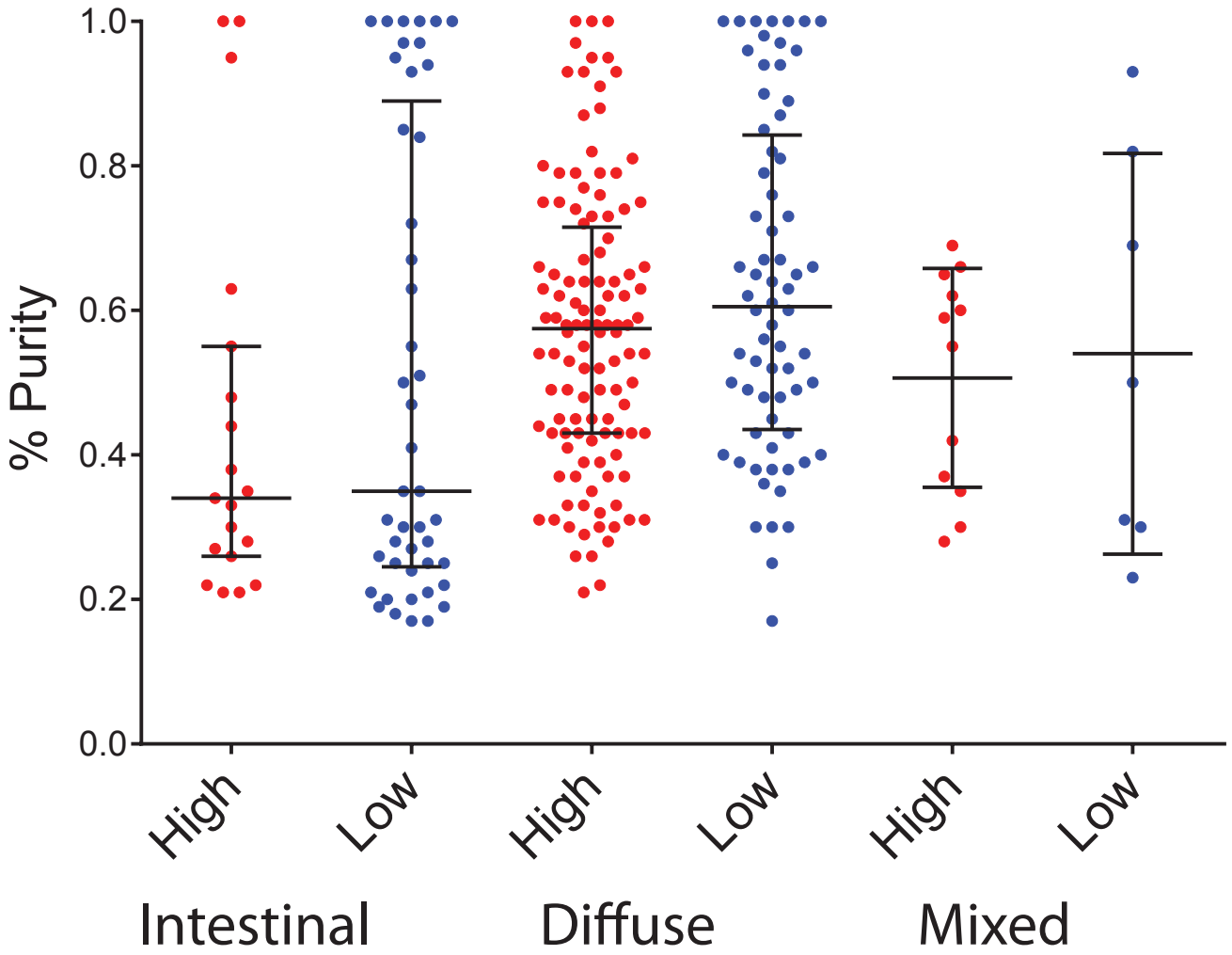


S2.5 GISTIC 2.0 analyses of focal amplifications in molecular subtypes. Chromosomal locations of peaks of significantly recurring focal amplifications are plotted by false discovery rates. Annotated peaks have an FDR < .25 and encompass 40 or fewer genes. Peaks are annotated with candidate driver oncogenes or by cytoband. The number of genes within each peak is shown next to driver genes or cytobands.

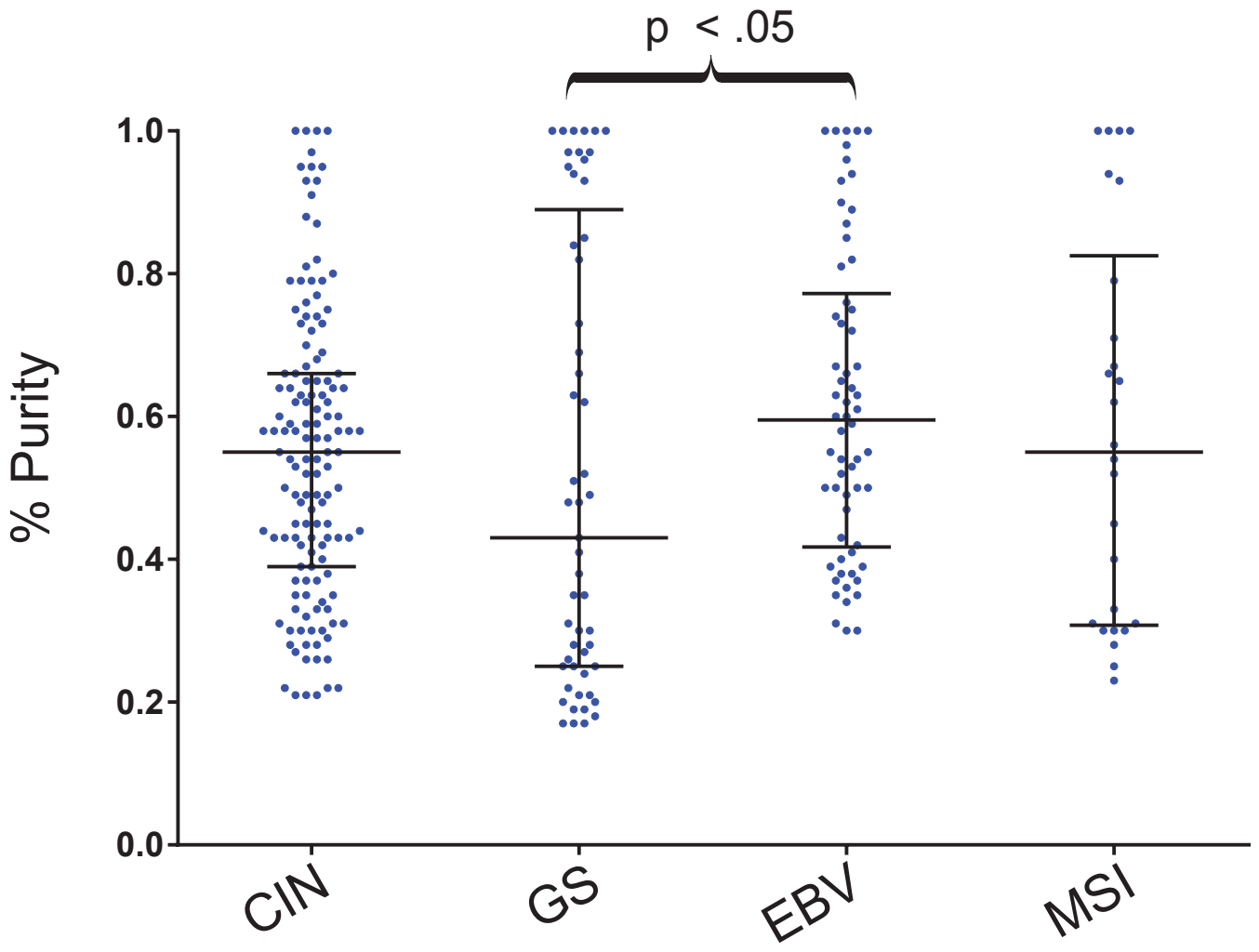


S2.6 GISTIC 2.0 analyses of focal deletions in molecular subtypes.

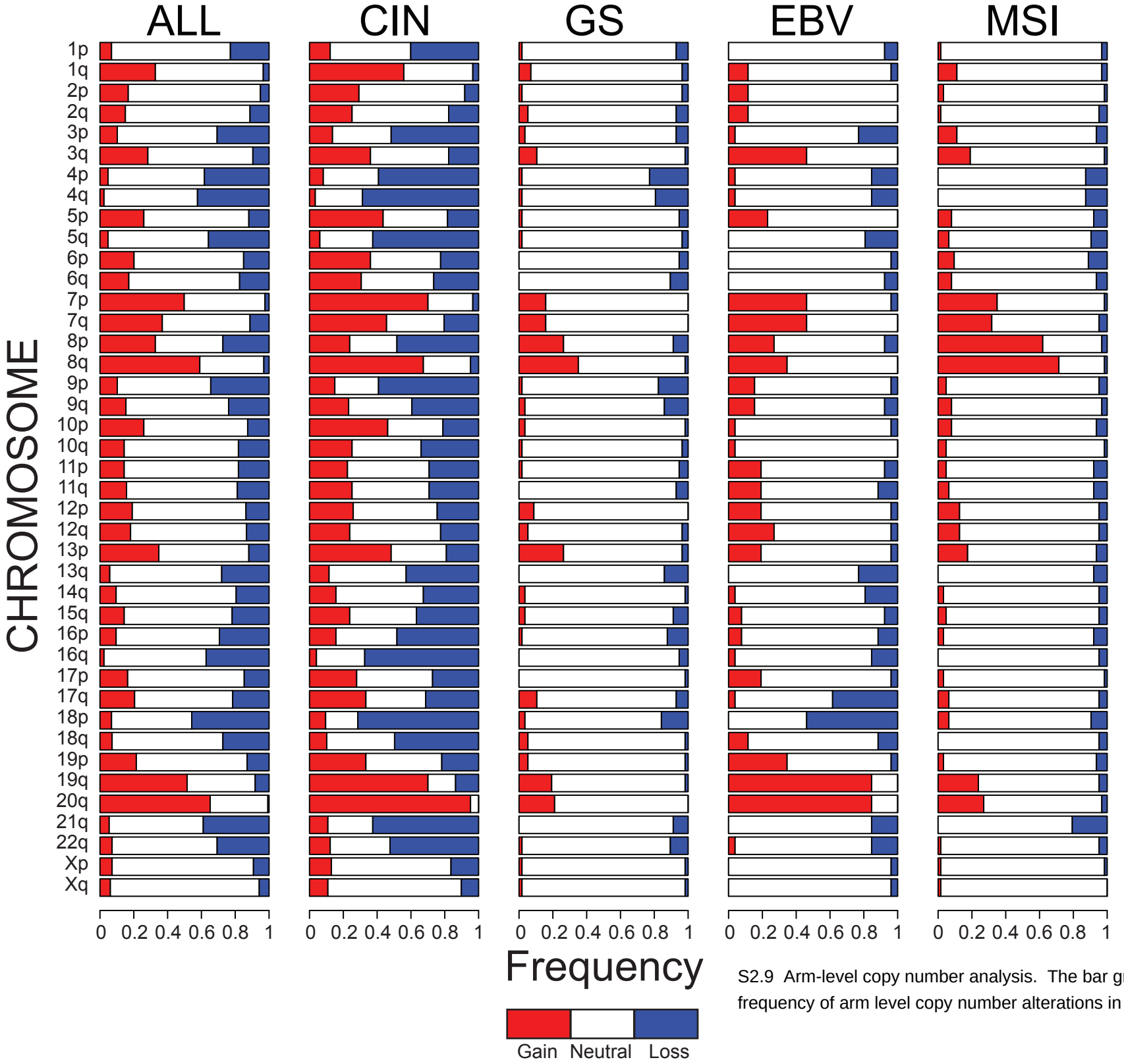
S2.6 GISTIC 2.0 analyses of focal deletions in molecular subtypes. Chromosomal locations of peaks of significantly recurring focal deletions are plotted by false discovery rates. Annotated peaks have an FDR < .25 and encompass 40 or fewer genes. Peaks are annotated with candidate driver tumor suppressors, fragile site genes (green) or by cytoband. The number of genes within each peak is shown next to driver genes or cytobands.



S2.7 Tumor purity of copy number clusters in histology classes. Tumor purities using the ABSOLUTE algorithm.



S2.8 Tumor purity by molecular subtypes. Tumor purities were estimated by using the ABSOLUTE algorithm. Significant Mann-Whitney p values are shown.



S2.9 Arm-level copy number analysis. The bar graphs show the frequency of arm level copy number alterations in molecular subtypes.

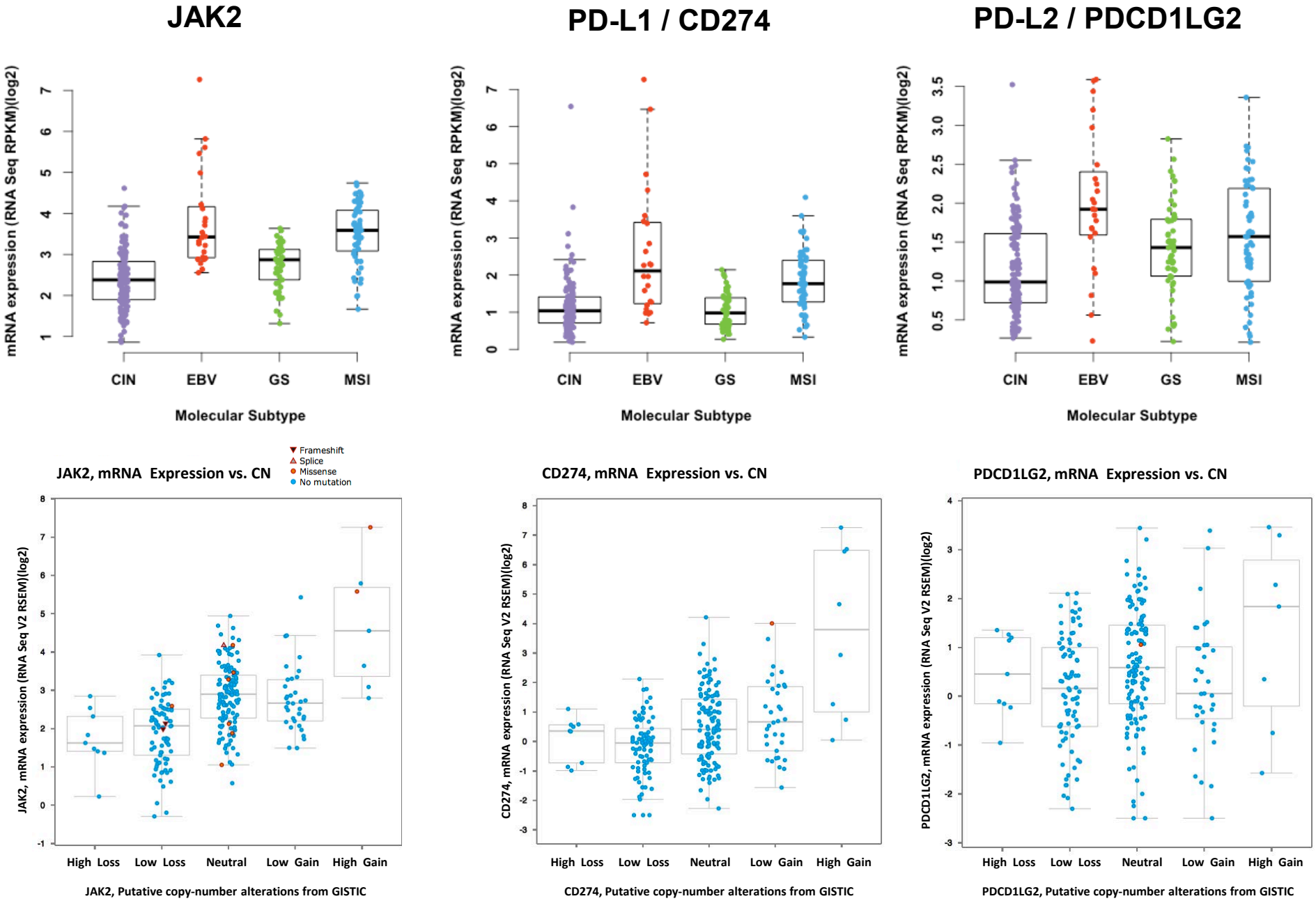


Figure S2.10. JAK2, PD-L1, and PD-L2 gene expression and copy number.

S2.10 JAK2, PD-L1, PD-L2 gene expression and copy number. Top graphs show in each molecular subtype the expression of JAK2, PD-L1, and PDL-2, which are all contained within the 9p24.1 focal amplification. Bottom graphs show expression (y axis) graphed by GISTIC 2.0 estimated levels of copy number. Low level losses or gains are estimated changes of one copy, high level losses or gains are changes estimated to be of two or more copies.

S3. DNA Sequencing

S3 Section Authors:

Amaro Taylor-Weiner
Carrie Sougnez
Vésteinn Thorsson
Reanne Bowlby
Angeliki Pantazi
Katayoon Kasaian

Subsections:

S3.1 Supplementary Materials: DNA sequencing Variant Calling

Sequencing methods

Whole exome sequencing was performed as previously described[1]. Briefly, 0.5-3 micrograms of DNA from each sample was used for library preparation, which included shearing and ligation of sequencing adaptors. Exome capture was performed using the Agilent SureSelect Human All Exon 5Mb kit. Captured DNA was sequenced using the Illumina HiSeq platform, and paired-end sequencing reads were generated for each sample. Initial alignment and quality control were performed using the Picard and Firehose pipelines at the Broad Institute.

Picard generated a single BAM file for each sample that included reads, calibrated quantities, and alignments to the genome. Firehose represents a set of tools for analyzing sequencing data from tumour and matched normal DNA. The pipeline performed quality control, local realignment, mutation calling, small insertion and deletion identification, and coverage calculations, among other analyses. Complete details of this pipeline have been published[2] or can be found online at www.broadinstitute.org/cancer/cga.

Variant Calling and Significance Analysis

Somatic mutations and short insertions/deletions (indels) were called and post-filtered using MuTect, Strelka and Indel locator[5,6,7]. These were then annotated to genes, variant severity, and transcript using Oncotator (http://www.broadinstitute.org/cancer/cga/oncotator). Variants were filtered against a panel of normals as described[5]. We assembled our mutation file for significance analysis by combining MuTect-called SNVs with Strelka and Indel locator called indels and split the file by mutation rate (Supp. 3.2). We limited our significance analysis of the Hypermutator set by removing indels below 30% allele fraction or alternate allele count four. Mutation significance was assessed using the MutSigCV algorithm[3]. In brief, this algorithm takes into account recurrence of mutations, nucleotide context, gene-expression, replication time, and somatic background mutation rate. Genes with a q-value less than 0.1 were deemed significant.

References:

1. TCGA Network. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012 Sep 27;489(7417):519-25

2. Stransky, N. et al. The mutational landscape of head and neck squamous cell carcinoma. Science 333, 1157-60 (2011).

3. Lawrence, MS. et al. Mutational heterogeneity in cancer and the search for novel cancer-associated genes. Nature 2013 Jun 16.

4. Futreal, P.A. et al. A census of human cancer genes. Nature reviews cancer 4, 177-83 (2004).

5.Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 2013;31:213-9.

6. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumour-normal sample pairs. Bioinformatics 2012;28:1811-7.

7. Chapman, M.A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* 471, 467-72 (2011).

Supplement S3.2. Determining thresholds for mutation rate categories

A plot of total mutation rate for samples, sorted in decreasing order (yielding ranks 1 through 289) indicated that mutation rates fell into three distinct populations, which we termed Ultramutated (UM), Hypermutated (HM) and Standard mutation rate (SM), in order of decreasing mutation rate (Supp. Fig. S3.3a). The term "Ultramutated" will be used only to aid this discussion – in the manuscript, the ultramutated group was considered a subset of hypermutated, and SM is referred to as non-hypermutated. In order to identify the boundaries between these regions, we made use of the observation that rates were approximately linear in the HM and SM regions. We then determined the values at which the transitions occurred between these regions of constant slope. Polynomial spline fitting of the values in the rate vs. rank plot (Supp. Fig. 3.3a) was used to estimate the transition points.

Fitting with third-order polynomial splines (R package pspline) gave a good representation of the rate vs rank curve (Supp. Fig. 3.3b, e.g.) and provided a framework for numerical estimation. First derivative (slope) estimates were found to be approximately constant in HM and SM regions but not in UM (Supp. Fig. 3.3c). The mean slope was -0.7941 in HM (estimated using ranks 20-60) and -0.0553 in SM (ranks 80-200). The corresponding standard deviations were 0.308 and 0.0442, respectively. To determine transition points between regions, we determined the point at which the fitted curve deviated by a specified number of standard deviations from these means. Supp. Fig. 3.3d shows the transition from HM (higher rank) to UM (lower rank), defined as where the curve deviated from the mean by more than 3 standard deviations. The spline fit, combined with root-finding (uniroot function in R) gave 11.41 as the interpolated rank at which this occurs. This rank corresponded to 63.6767 mutations/Mb in the spline fit (Supp. Fig. 3.Xa). Similarly, Supp. Fig.3.3e shows the transition from SM to HM. The crossover point, at rank 73.90, is 6 standard deviations below the SR mean, and corresponds to 11.4354 mutations/Mb.

To summarize, hypermutated samples are defined as those with mutation rates greater than 11.4354 mutations/Mb (74 samples). Those samples with lower mutation rates are termed non-hypermutated (215 samples). Among the hypermutated samples, we detected a distinct group of 11 samples with mutation rates greater than 67.6767 mutations/Mb which were excluded from the MutSigCV analysis. Supplementary figure S3.3e shows the resulting classification for the 100 most highly mutated samples, using the nomenclature of the analysis described herein.

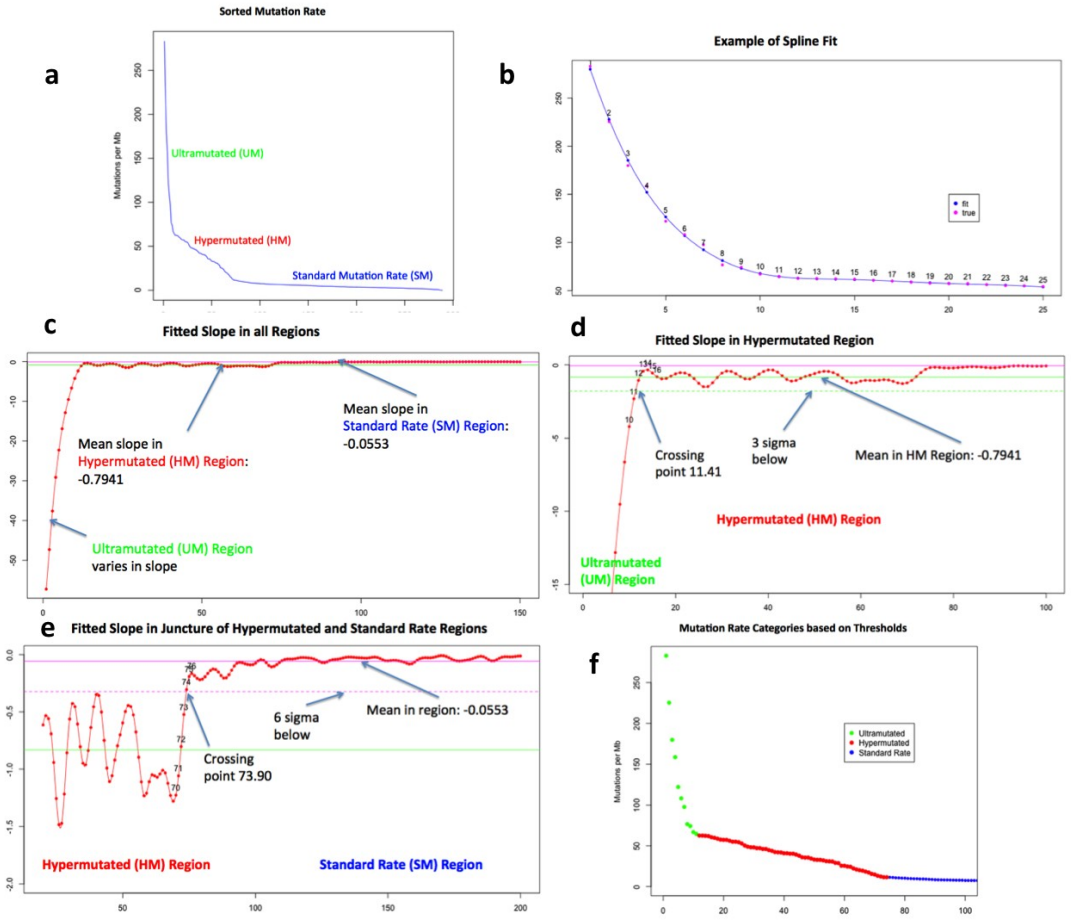


Figure S3.3: **Determining Thresholds for Mutation Rate Categories. a)** Sorted total mutation rates from high (left) to low (right) labeled by three implied populations: Ultramutated (UM), Hypermutated (HM), and Standard Mutation Rate (SM). In the manuscript, the first two categories were combined to create the category hypermutated, and SM was called non-hypermutated. **b)** Third-order spline fit shown in a selected range. True mutation rates are in blue and fitted rates in magenta. **c)** First derivatives (slopes) in the HM and SM regions were approximately constant. The green horizontal line indicates the mean slope in the HM region, magenta the mean slope in SM. **d)** Transition from HM to UM region. Solid lines are as in panel c), while the green dotted line is 3 standard deviations below the HM mean. **e)** Transition from SM to HM region. Solid lines are as in panel c), and the magenta dotted line is 6 standard deviations below the SM mean. **f)** Mutation rate categories after application of thresholds.

S3.4 Description of mutation validation

The mRNA sequencing data were used to validate mutations in genes in our standard-mutator significant gene list (S 3.5). We selected candidate mutations for validation by filtering the significant genes by expression, selecting only genes with median RPKM greater than one. Next we limited candidate mutations to coding regions. These 778 events in 18 genes were then compared against mRNA-Seq SNP and indel calls made by extended-SNVMix2 and Trans-ABySS v1.4.6 respectively (www.bcgsc.ca/platform/bioinfo/software). Of 712 events, 586 (82%) were verified; the remaining 66 events were among patients with no mRNA-Seq data. When we removed the three genes with the lowest mutation rate (CASC3 51%, MUC6 49% and PTPRC 0), 542 out of 600 (90%) events were verified.

Additionally, we performed validation of the somatic status of the genes we identified as subject to statistically significantly recurrent mutations (in both the analysis of the standard mutator groups and hypermutated groups of tumours). Validation was performed on DNA from both tumour and matched germline samples using multiplexed PCR amplification using a microfluidic PCR platform (Fluidigm Access Array) followed by sequencing of the PCR products on Illumina MiSeq instruments as

performed earlier for the TCGA Urothelial Carcinoma study[1]. Mutations were only considered for validation if there was adequate power based on the allele fraction of the candidate site in the discovery data and coverage in the validation data. Analysis of sequencing data resulted in removal of two genes initially considered to be novel significantly recurrently mutated genes; the gene *PGM5* in hypermutated tumours and *CASC3* in the non-hypermutated tumours. Following exclusion of mutations of these two genes, 95% of evaluable candidate mutations of the genes in the significant gene lists were validated as somatic alterations.

1 Cancer Genome Atlas Research, N. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315-322, doi:10.1038/nature12965 (2014).

	No RNA data	No RNA evidence	RNA evidence	Verification Rate
ERBB2	1	0	15	100.00%
RHOA	1	0	16	100.00%
SMAD2	0	0	11	100.00%
TP53	14	5	116	95.87%
CTNNB1	2	1	23	95.83%
ARID1A	7	8	102	92.73%
RNF43	9	4	51	92.73%
KRAS	4	2	22	91.67%
CDH1	3	3	24	88.89%
SMAD4	2	3	23	88.46%
PIK3CA	5	9	60	86.96%
APC	3	10	43	81.13%
BCOR	5	5	19	79.17%
RASA1	1	5	12	70.59%
EIF2C4	1	3	5	62.50%
CASC3	1	17	18	51.43%
MUC6	3	27	26	49.06%
PTPRC	4	24	0	0.00%
Total	66	126	586	82.30%
*Total	58	58	542	90.33%

*Excludes genes CASC3, MUC6, and PTPRC

Table S3.4a: Gene-specific mutation verification rates with mRNA-Seq data. Mutations are those from the standard mutator gene list after filtering genes by median RPKM less than one. Splice site events and events in the UTR are excluded.

S3.5 The data file listing the MutSig significantly mutated genes can be found on the TCGA Stomach Adenocarcinoma publication page at:
https://tcga-data.nci.nih.gov/docs/publications/stad_2014/.

S3.6. Low-pass whole genomes sequencing methods

Library construction. Between 500 and 700 ng of each gDNA sample was sheared into fragments of ~250 bp, using a Covaris E220 ultrasonicator; fragments were converted to a pair-end Illumina library using KAPA Bio kits with Caliper (PerkinElmer) robotic NGS Suite according to manufacturers' protocols. All libraries were sequenced on HiSeq2000 using one sample per lane, with the paired-end 2 x 51bp setup. Tumours and matching normal DNA samples were usually loaded on the same flowcell. Average sequence coverage was ~6.19X, read quality was 38.04 and 92.31% of the reads were mapped. Raw data were converted to the FASTQ format, and BWA alignment was used to generate bam files.

Identification of copy number variants. To characterize somatic copy number alterations in the tumour genome, we applied BIC-seq[1], an algorithm we developed previously. Briefly, we first counted the uniquely aligned reads in fixed-size, non-overlapping windows along the genome. Given these bins with read counts for tumour and matched normal genomes, BIC-seq attempts to iteratively combine neighbouring bins with similar copy numbers. Whether the two neighbouring bins should be merged is based on Bayesian Information Criteria (BIC), a statistical criterion measuring both the fit and complexity of a statistical model. Segmentation stops when no merging of windows improves BIC, and the boundaries of the windows are reported as a final set of copy number breakpoints. Segments with copy ratio differences smaller than 0.1 (log2 scale) between tumour and normal genomes were merged in the post-processing step to avoid excessive refinement of altered regions with high read counts.

Detection of structural rearrangements with BreakDancer and Meerkat. We used two algorithms, BreakDancer[2] and Meerkat[3], to detect structural variation. The first step in BreakDancer requires a configuration file of each bam file for each tumour pair with the bam2cfg.pl perl module of the program. The perl module BreakDancerMax.pl is then run on the configuration file to call structural variants in the tumour and control files. The set of structural variants from each tumour sample was compared to the set from its matched normal, to remove germline variants. Structural variations were also detected by Meerkat, which requires at least two discordant read pairs supporting each event and at least one read covering the breakpoint junction. Variants detected from tumour genomes were filtered by the variants from all normal genomes to remove germ-line events and were also removed if both breakpoints fell into simple repeats or satellite repeats. The final set of variants fit one of these criteria: (i) the read identified to span the breakpoint junction hit the predicted breakpoint region uniquely, according to a BLAT search, or (ii) the mate of the read spanning the breakpoint junction was mapped near the predicted breakpoint.

Validation of rearrangement hits. A subset of all detected structural variants was cross-confirmed by both Breakdancer and Meerkat. In Meerkat, we also sought a read that spanned the breakpoint, to provide the precise location at which the break occurred. RNA-seq fusion data was used to assess if a novel junction between two genes could be found. Finally, in a number of cases that were detected by only one method and/or a split read was not found, we PCR-amplified the junction fragment and subjected it to Sanger sequencing.

Pathogen detection: To detect bacteria and viruses and to examine the physical status of the bacterial/viral genome, a custom pipeline PathWatch was used. As the first step, the pipeline performed computational subtraction of sequences mapped previously to the human genome. Then it used BWA aligner to map the remaining set of non-human sequences to the set of bacterial and viral reference genomes obtained from the NCBI RefSeq database (ftp://ftp.ncbi.nih.gov/refseq/release/microbial/ and ftp://ftp.ncbi.nih.gov/refseq/release/viral/ respectively). Reads that aligned to the genomes of multiple species were filtered out.
The percentage of covered pathogen genome and count of pathogen sequencing reads normalized by the length of the pathogen genome and total number of non-human reads in the sample were calculated. To consider a given sample positive for the pathogen, we chose an empirical threshold of 1kb of pathogen genome to be covered.

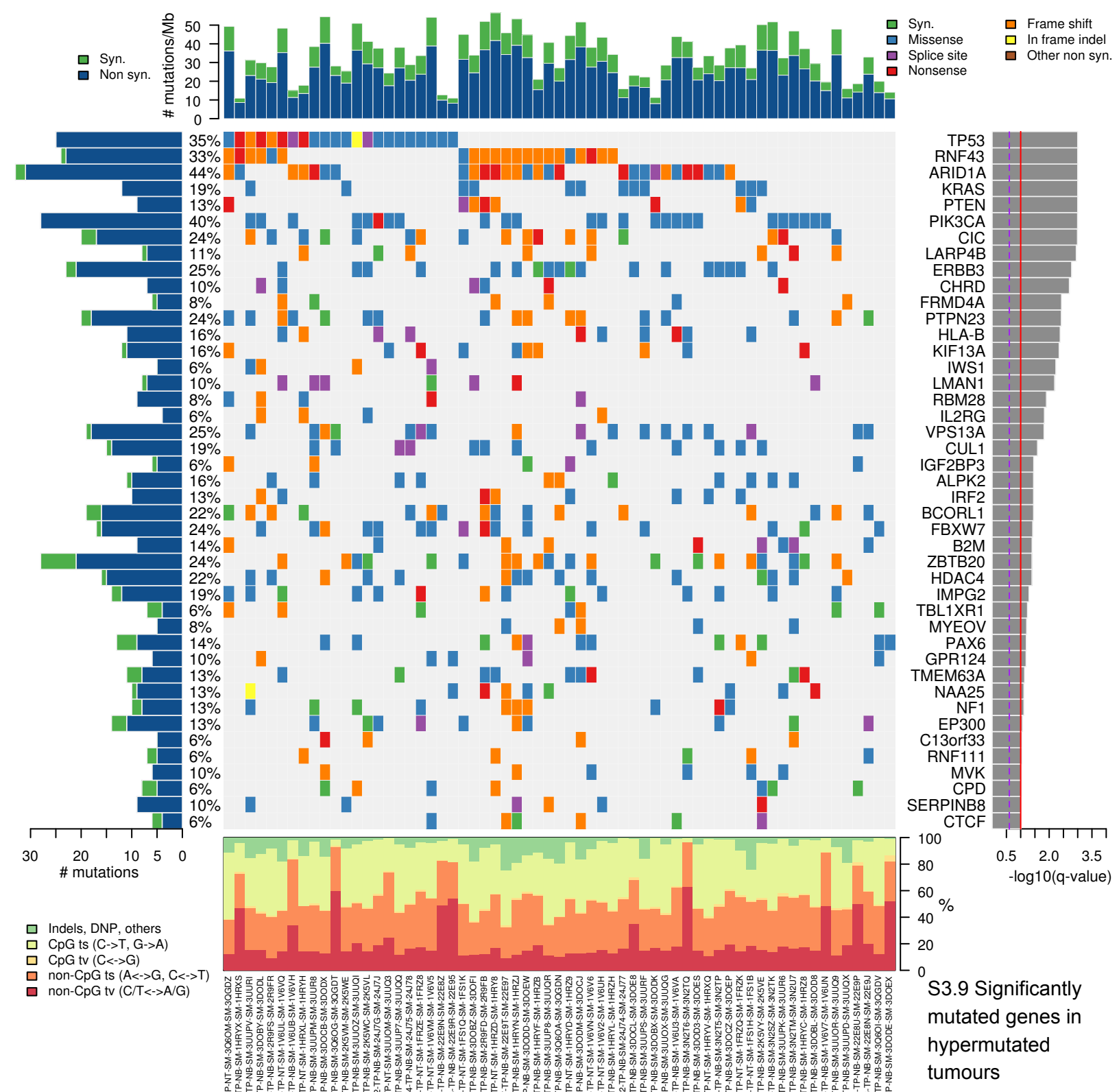
Evaluation of putative viral integration: To assess possible virus integration into the host genome, the pipeline used the advantage of paired-end (PE) sequencing technology and searched for the clusters of discordant read pairs where one mate is aligned to the human genome and the second mate mapped to the viral sequence. As an input, the original set of all PE reads, mapped and unmapped to the human genome was used, and two subsets of reads were generated: ends represented by human sequences and their unmapped mates. Then such unmapped reads were aligned against the specific viral genome identified in the previous step.

Clusters of discordant read pairs were calculated. To determine the putative presence of a cluster, we used an empirical cutoff of 3 discordant read pairs within the same integration region. To assess the precise site of a candidate integration event at nucleotide resolution, the pipeline searched for the chimeric viral-human reads. Soft-clipped reads, i.e. reads in which only a portion of a read had been mapped to the human genome, were filtered from the original PE dataset and were aligned by BLAT to the virus genome. As the pipeline operates in each step with only filtered subsets of reads, it is efficient in terms of the required time and computational resources, keeping the same precision as previously published methods.

1. Xi, R. *et al.* Copy number variation detection in whole-genome sequencing data using the
2. Bayesian information criterion. *Proc Natl Acad Sci U S A* **108**, E1128-36 (2011).
3. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* **6**, 677-81 (2009).
4. Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell* **153**, 919-29 (2013).

S3.7 The data file- In-frame rearrangement fusion list can be found on the TCGA Stomach Adenocarcinoma publication page at https://tcga-data.nci.nih.gov/docs/publications/stad_2014/.

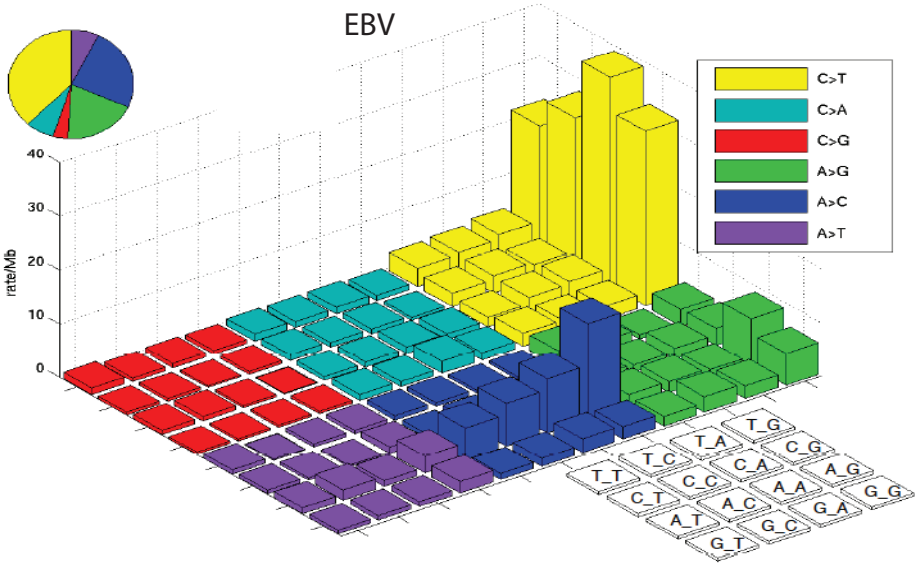
S3.8 The data file- low-pass structural rearrangements can be found on the TCGA Stomach Adenocarcinoma publication page at https://tcga-data.nci.nih.gov/docs/publications/stad_2014/.

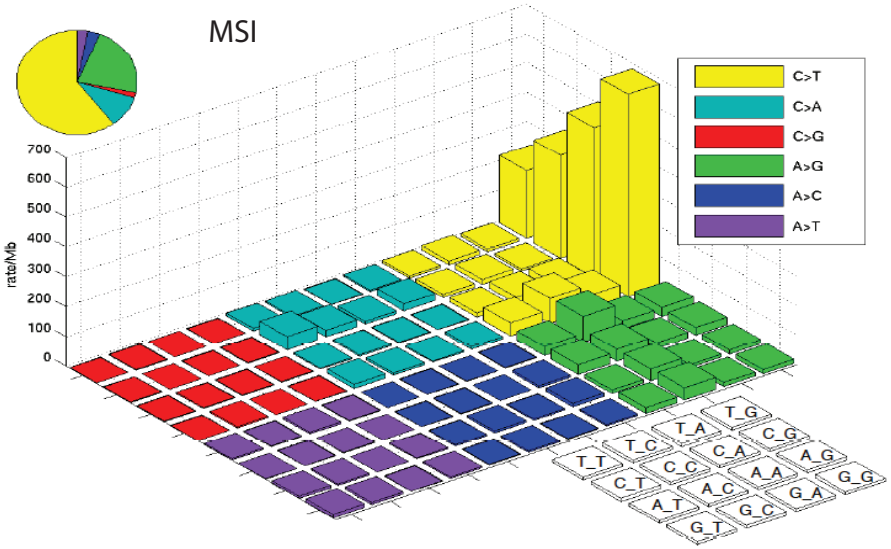


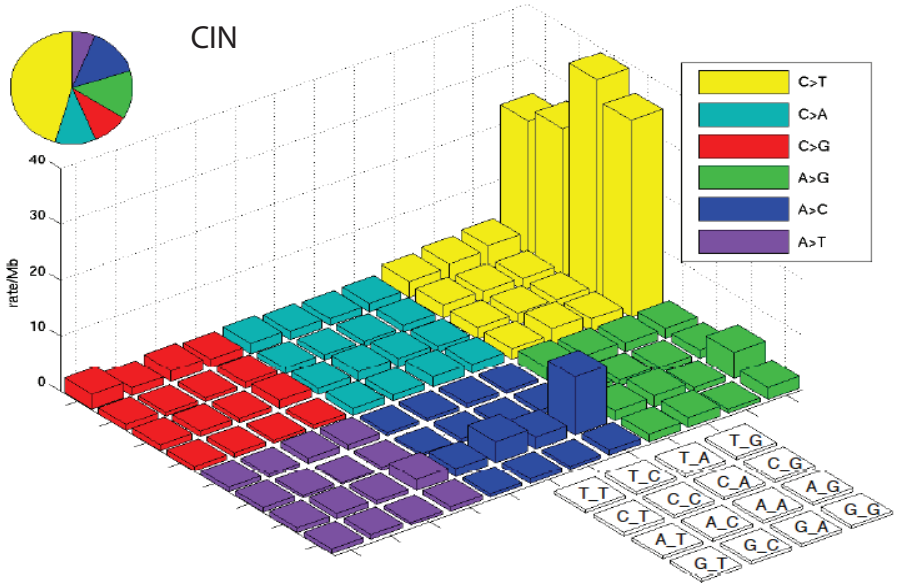
S3.9 Significantly mutated genes in hypermutated tumours

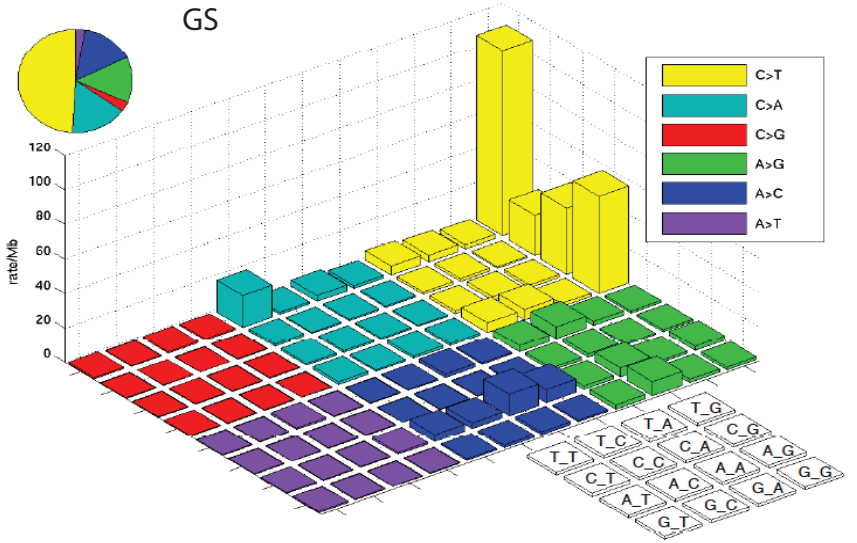
S3.9. The center matrix displays individual mutations in patient samples, color coded by mutation type, for the significantly mutated genes in the hypermutated tumours. The rate of mutations is shown at the top. The bar plot on the left shows the percentage of tumours with at least one mutation in each gene. The bar plot on the right displays q-values for each mutation. The panel below shows the spectrum of mutations in each tumour above.

Figure S.3.10. Mutation types across four molecular subtypes









S3.10. Mutation types across four molecular subtypes. Each bin shows the percentage of a particular mutation type out of all mutation types in the context of three bases: 5' base, mutated base, 3'base. There are 96 mutation types in total.

S4. DNA Methylation

S4 Section Authors:

Toshinori Hinoue
Hui Shen
Daniel J. Weisenberger
Moiz S. Bootwalla
Peter W. Laird

Subsections:

S4. DNA Methylation

S4.1 Methods

Array-based DNA methylation assay

We used two Illumina Infinium DNA methylation platforms (Illumina, San Diego, CA), HumanMethylation27 (HM27) and HumanMethylation450 (HM450), to obtain DNA methylation profiles of 295 gastric adenocarcinoma samples and 27 adjacent non-malignant stomach tissue samples. Ten control cell line technical replicates were also included in the assay to monitor technical variations, with two on the HM27 platform and eight on the HM450 platform; 47 tumours and 25 adjacent non-malignant samples were analyzed on the HM27 platform, and 248 tumours and 2 adjacent non-malignant samples were analyzed on the HM450 platform. The HM27 array targets 27,578 CpG sites located in proximity to the transcription start sites of 14,475 consensus coding sequencing (CCDS) in the NCBI Database (Genome Build 36). The HM450 assay analyzes DNA methylation status of up to 482,421 CpG sites throughout the genome. It covers 99% of RefSeq genes with multiple probes per gene and 96% of CpG islands from the UCSC database and their flanking regions. The assay probe sequences and information for each interrogated CpG site on both Infinium DNA methylation platforms are available from Illumina (www.illumina.com).

The DNA methylation score for each locus is presented as a beta (β) value (β = (M/(M+U)) in which M and U indicate the mean methylated and unmethylated signal intensities for each locus, respectively. β-values range from zero to one, with scores of zero indicating no DNA methylation and scores of one indicating complete DNA methylation. A detection *P* value accompanies each data point and compares the signal intensity difference between the analytical probes and a set of negative control probes on the array. Any data point with a corresponding *P* value greater than 0.05 is deemed not to be statistically significantly different from background and is thus masked as "NA" in the Level 3 data packages as described below. Further details on the Illumina Infinium DNA methylation assay technology have been described previously[1,2].

Sample and data processing

We performed bisulfite conversion on 1 µg of genomic DNA from each sample using the EZ-96 DNA Methylation Kit (Zymo Research, Irvine, CA) according to the manufacturer's instructions. We assessed the amount of bisulfite-converted DNA and completeness of bisulfite conversion using a panel of MethyLight-based quality control (QC) reactions as previously described[3]. All the TCGA samples passed our QC tests and entered the Infinium DNA methylation assay pipeline. Bisulfite-converted DNAs were whole-genome-amplified (WGA) and enzymatically fragmented prior to hybridization to BeadChip arrays. BeadArrays were scanned using the Illumina iScan technology to produce IDAT files. Raw IDAT files for each sample were processed with the R/Bioconductor package *methylumi*. TCGA DNA methylation data packages were then generated using the *EGC.tools* R package which was developed internally and is publicly available on GitHub (https://github.com/uscepigenomecenter/EGC.tools).

TCGA Data Packages

The data levels and the files contained in each data level package are described below and are present on the TCGA Data Portal website (http://tcga-data.nci.nih.gov/tcga/). A DESCRIPTION file outlining the methods used to generate the Level 1, 2 and 3 files is also provided in the AUX directory. Please note that as continuing updates of genomic databases and data archive revisions frequently become available, the data packages on TCGA Data Portal are updated accordingly.

HM27: Level 1 data contain raw IDAT files (two per sample) as produced by the iScan system and as mapped by the SDRF. These IDAT files can be directly processed by the R/Bioconductor package *methylumi*. We provided a comma-separated value (CSV) disease-mapping file (STAD.mappings.csv) in the AUX directory to facilitate this process. Level 2 data contain background-corrected methylated (M) and unmethylated (U) summary intensities as extracted by the R/Bioconductor package *methylumi*. Non-detection probabilities (*P* values) were computed as the minimum of the two values (one per allele) for the empirical cumulative density function of the negative control probes in the appropriate colour channel. Background correction was performed via normal-exponential deconvolution. Level 3 data contain β-values for each interrogated locus with annotations for the HUGO Gene Nomenclature Committee (HGNC) gene symbol, chromosome (UCSC hg19, Feb 2009), and CpG coordinate (UCSC hg19, Feb 2009). Probes having a SNP within 10bp of the interrogated CpG site or having a repeat element [as defined by RepeatMasker and Tandem Repeat Finder Masks (UCSC hg19, Feb 2009) contained in the *BSgenome.Hsapiens.UCSC.hg19* R package] that covers 15 bp from the interrogated CpG site were

masked as “NA” across all samples. Furthermore, probes with a detection *P* value greater than 0.05 in a given sample were masked as “NA” on that array. Probes that were mapped to multiple sites on hg19 were annotated as “NA” for chromosome and 0 for CpG coordinate.

HM450: Level 1 data contain raw IDAT files (two per sample) as produced by the iScan system and as mapped by the SDRF. These IDAT files can be directly processed by the R/Bioconductor package *methylumi*. We provided a comma-separated value (CSV) disease-mapping file (STAD.mappings.csv) in the AUX directory to facilitate this process. Level 2 data contain background-corrected methylated (M) and unmethylated (U) summary intensities as extracted by the R/Bioconductor package *methylumi*. Non-detection probabilities (*P* values) were computed as the minimum of the two values (one per allele) for the empirical cumulative density function of the negative control probes in the appropriate color channel. Background correction was performed via normal-exponential deconvolution. Multiple-batch archives had the intensities in each of the two channels multiplicatively scaled to match a reference sample (sample with R/G ratio of the normalization control probes closest to 1.0). Level 3 data contain β-value calculations with annotations for HGNC gene symbol, chromosome (UCSC hg19, Feb 2009), and genomic coordinate (UCSC hg19, Feb 2009) for each targeted CpG/CpH site on the array. Probes having a common SNP (Minor Allele Frequency > 0.01, per dbSNP build 135 via the UCSC snp135common track) within 10 bp of the interrogated CpG site or having a 15 bp from the interrogated CpG site which overlaps with a repetitive element (as defined by RepeatMasker and Tandem Repeat Finder Masks contained in the *BSgenome.Hsapiens.UCSC.hg19* R package) were masked as “NA” across all samples, and probes with a non-detection probability (*P* value) greater than 0.05 in a given sample were masked as “NA” on that array. Probes that were mapped to multiple sites on hg19 were annotated as “NA” for chromosome and 0 for CpG/CpH coordinate.

The following data archives were used for the analyses described in this manuscript.
jhu-usc.edu_STAD.HumanMethylation27.Level_3.1.2.0
jhu-usc.edu_STAD.HumanMethylation27.Level_3.2.2.0
jhu-usc.edu_STAD.HumanMethylation450.Level_3.1.6.0
jhu-usc.edu_STAD.HumanMethylation450.Level_3.2.6.0
jhu-usc.edu_STAD.HumanMethylation450.Level_3.3.6.0
jhu-usc.edu_STAD.HumanMethylation450.Level_3.4.6.0
jhu-usc.edu_STAD.HumanMethylation450.Level_3.5.6.0
jhu-usc.edu_STAD.HumanMethylation450.Level_3.6.6.0
jhu-usc.edu_STAD.HumanMethylation450.Level_3.7.6.0
jhu-usc.edu_STAD.HumanMethylation450.Level_3.8.6.0
jhu-usc.edu_STAD.HumanMethylation450.Level_3.9.6.0

HM27 and HM450 data merging

A total of 25,978 probes were shared between the HM27 and HM450 platforms. Among these, 2,597 probes were masked in the Level 3 data due to SNPs, repeats, and non-unique mapping to the genome as described above (n=23,381 remaining). We observed batch- and platform-specific effects with the technical replicates. To minimize systematic platform-specific effects (dye bias, background level, etc.), we fitted a LOESS regression model between the two platforms using cell line control technical replicates. We normalized the HM450 data against the HM27 data with this fitted model on M values, stratified by the number of CpGs in the probe (CpG=1,2,3,4,5,6+). M value is the $\log_2$ ratio of Methylated (M) intensity and Unmethylated (U) intensity ($=\log_2(M/U)$) and better satisfies the linearity assumption[4]. The M values were then transformed back to β- values.

Unsupervised clustering analysis of DNA methylation data

We removed probes that showed high technical variances after the platform correction based on standard deviation (>0.15) across technical replicates (n=23,322 remaining probes). We also removed probes which had any “NA”-masked data points and probes that were designed for sequences on X and Y chromosomes. Furthermore, to capture cancer-specific DNA hypermethylation events, we selected 3,732 CpG sites that are located in high CpG density regions (top 20% of the sites with the highest observed/expected CpG ratio around their 3 kb regions) and are not highly methylated in adjacent non-malignant stomach tissues (mean β-value <0.5). However, a clustering analysis can be strongly confounded by the purity of tumour samples. To minimize the potential influence of variable levels of tumour purity in our sample set on our clustering result, we dichotomized the data using a β-value of >0.3 as a threshold for positive DNA methylation. The dichotomization not only ameliorated the effect of tumour sample purity on the clustering, but also removed a great portion of residual batch/platform effects that are mostly reflected in small variations near the two ends of the range of β-values. We then

performed consensus clustering using 1,315 CpG sites that were methylated with this threshold in more than 5% of the tumour samples. The optimal number of clusters was assessed based on 80% resampling over 1,000 iterations of hierarchical clustering for K=2,3,4,5,6 using the binary distance metric for clustering and Ward's method for linkage as implemented in the R/Bioconductor *ConsensusClusterPlus* package. The heatmap shown in Figure 2a was generated based on the original β-values to visualize 1,315 CpG sites used in the clustering. The CpG sites were arranged based on the order of unsupervised hierarchal clustering of the dichotomous data using the binary distance metric and Ward's linkage method.

DNA hypermethylation frequency

We previously identified 12,862 CpG sites that were constitutively unmethylated in 12 different normal tissue types as a part of the TCGA Pan-Cancer project[5]. We dichotomized the β-values in the tumours at 0.3. For each locus, tumours with a β-value of 0.3 or greater were designated as methylated and tumours with a β-value of lower than 0.3 were designated as unmethylated. We then calculated the percentage of loci that were methylated among the loci investigated in each tumour. DNA hypermethylation frequencies in 4,923 tumours consisting of 12 different tumour types were previously calculated[5]. The upper and lower ends of the box correspond to the 25th and 75th quartiles, respectively. The line within the box identifies the median. The whiskers above and below the box extend to at most 1.5 times the interquartile range.

Identification of epigenetically silenced genes

We identified genes that were silenced by DNA methylation and made dichotomous epigenetic silencing calls for each sample for the corresponding genes. We generated two sets of silencing calls independently; one based on the HM450 DNA methylation data only (Data file S4.4) and the other based on the HM27-HM450 merged DNA methylation data set (Data file S4.5).

We first removed DNA methylation probes overlapping with SNPs, repeats or designed for sequences on X and Y chromosomes and non-CpG sites. The remaining probes were mapped against UCSC Genes using the *GenomicFeatures* R/Bioconductor package. Probes that were located either in a promoter region (defined as the 3 kb region spanning from 1,500 bp upstream to 1,500 bp downstream of the transcription start sites) or in a gene body were identified. Level 3 RNA-seq RPKM data on 29,699 genes were $\log_2$ transformed [$\log_2$ (RPKM+1)] and used to assess the expression levels associated with DNA methylation changes. DNA methylation and gene expression data were merged by Entrez Gene IDs. For the epigenetic silencing calls based on the HM450 data, a total of 220 tumours had both DNA methylation and expression data and we examined ~21,000 genes for this analysis. For the epigenetic silencing calls based on the HM27-HM450 merged data, we examined ~14,000 genes in 262 tumours. We removed the CpG sites that were methylated in adjacent non-malignant stomach tissues (mean β-value >0.3). We then dichotomized the DNA methylation data using a β-value of >0.3 as a threshold for positive DNA methylation, and further eliminated CpG sites methylated in fewer than 5% of the tumour samples. For each probe/gene pair, we applied the following algorithm: 1) Organize the tumours as either methylated (β ≥0.3) or unmethylated (β <0.3); 2) Compute the mean expression in the methylated and unmethylated groups; 3) Compute the standard deviation of the expression in the unmethylated group. We then selected probes for which the mean expression in the methylated group was lower than 1.28 standard deviations (bottom 10%) of the mean expression in the unmethylated group. Furthermore, we selected from the remaining probes in which >80% of the tumour samples in the methylated group had expression levels lower than the mean expression in the unmethylated group. We labeled each individual tumour sample as epigenetically silenced for a specific probe/gene pair selected from above if: a) it belonged to the methylated group and b) the expression of the corresponding gene was lower than the mean of the unmethylated group of samples. If there were multiple probes associated with the same gene, a sample identified as epigenetically silenced at more than half the probes for the corresponding gene was also labeled as epigenetically silenced at the gene level. The complete list of 769 genes (derived from the HM450 data) and 249 genes (derived from the HM27-HM450 merged data) identified as epigenetically silenced are provided in Data file S4.4 and Data file S4.5, respectively.

CDKN2A (*p16INK4*) epigenetic silencing calls were made separately using the exon level RNA-seq data. *p16INK4* DNA methylation status was assessed in each sample based on the probe (cg13601799) located in the *p16INK4* promoter CpG island. Note that the DNA methylation level at this locus was measured only in the HM450 platform, as the HM27 array does not contain a probe specific for this locus within the *p16INK4* gene. *p16INK4* expression was determined by the $\log_2$(RPKM+1) level of its first exon (chr9:21974403-21975132). The epigenetic silencing calls for each sample were made manually by evaluating a scatter plot showing an inverse association between DNA methylation and expression. We

incorporated the *p16INK4* epigenetic silencing calls into the HM450 silencing call list as described above (Data file S4.4).

Notably, a further analysis of the 769 epigenetically silenced genes revealed that nearly half of the silencing events occurred specifically in the EBV-positive molecular subgroup (Figure S4.2). We identified 526 genes that were significantly more frequently silenced (FDR-adjusted P <0.01, Fisher's exact test) in the EBV-positive subgroup compared with all the other groups (Table S4.3).

Statistics

Statistical analysis and data visualization were carried out using the R/Biocoductor software packages (http://www.bioconductor.org). Cancer-specific DNA hypermethylation was assessed based on unpaired analyses, since matched non-malignant tissues were available for fewer than 10% of the tumour samples.

Section References

1. Bibikova, M. Genome-wide DNA methylation profiling using Infinium assay. Epigenomics 1, 177-200 (2009).
2. Bibikova, M. et al. High density DNA methylation array with single CpG site resolution. Genomics 98, 288-295 (2011).
3. Campan, M. et al. MethyLight. Methods Mol Biol 507, 325-337 (2009).
4. Du, P. et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11, 587 (2010).
5. Shen, H. et al. Comprehensive cross-cancer comparison of DNA methylation profiles. (Manuscript under review)

Figure S4.2

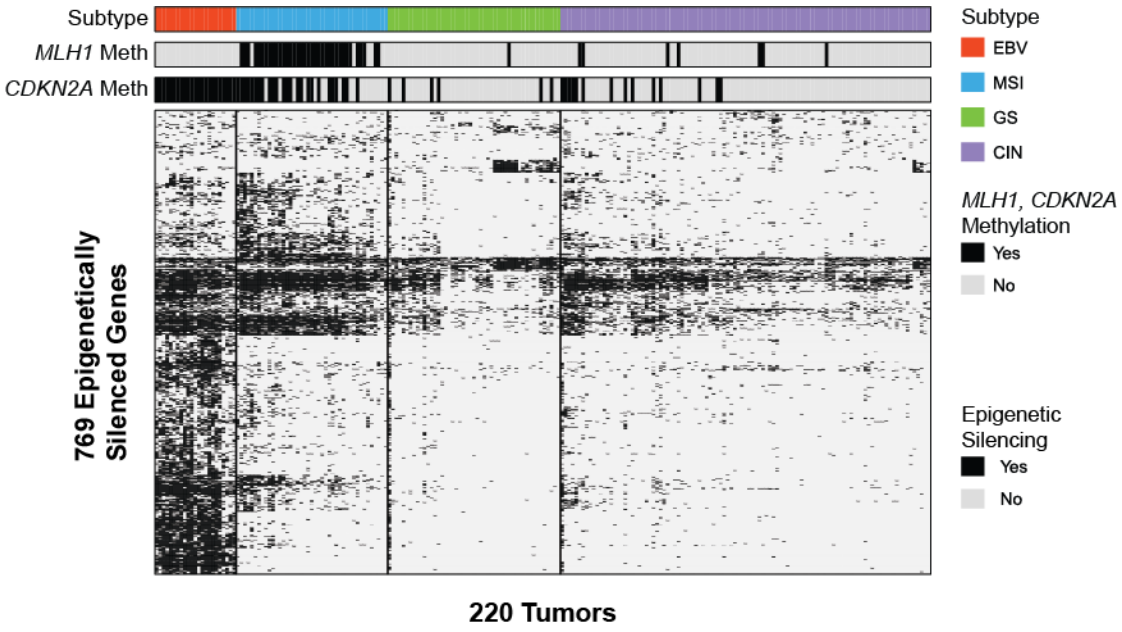


Figure S4.2 Clustering and heatmap representation of dichotomous epigenetic silencing calls. We identified 769 genes that were silenced by DNA methylation in at least 5% of the tumours. The epigenetic silencing calls of each gene in 220 STAD tumour samples are indicated by using a black (yes) and gray (no) color scheme. Genes are ordered based on a hierarchical clustering with Ward's method on the Jaccard Distance, a distance measure that best suits binary data. *CDKN2A* and *MLH1* DNA methylation status of each tumour sample are also shown as bars above the heatmap. Samples are grouped according to the four major molecular subtypes indicated as a vertical colour bar at the top.

S4.3 Table- A list of genes significantly more frequently silenced by DNA methylation in EBV tumours compared with non-EBV tumours can be found on the TCGA Stomach Adenocarcinoma publication page at https://tcga-data.nci.nih.gov/docs/publications/stad_2014/.

S4.4 The data file- Epigenetic silencing calls based on HM450 data set can be found on the TCGA Stomach Adenocarcinoma publication page at https://tcga-data.nci.nih.gov/docs/publications/stad_2014/.

S4.5 The data file- Epigenetic silencing calls based on HM27-HM450 merged data set can be found on the TCGA Stomach Adenocarcinoma publication page at https://tcga-data.nci.nih.gov/docs/publications/stad_2014/.

Figure S4.6

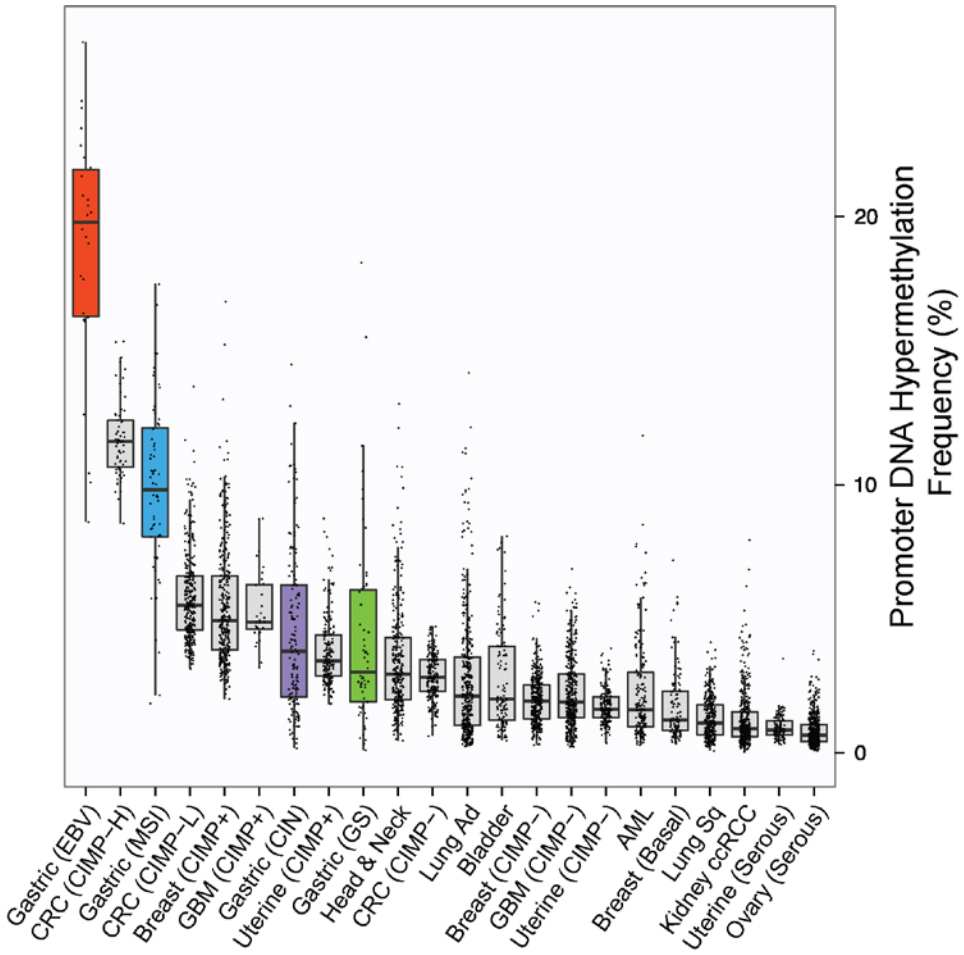


Figure S4.6 DNA hypermethylation frequencies across 10 tumour types. Box and jitter plots show DNA hypermethylation frequencies, calculated as the percentage of CpG sites methylated among 12,862 constitutively unmethylated sites in normal tissues, for 10 different neoplastic tissue types studied through TCGA (see Supplemental Methods S4.1). Tumour types/subtypes are organized from left to right in descending order of median frequencies of DNA hypermethylation.

S5. RNA Sequencing

S5 Section Authors:

Reanne Bowlby
Andrew J. Mungall

Subsections:

Supplement S5.1 - Messenger RNA library construction, sequencing and analysis

Two micrograms of total RNA samples were arrayed into a 96-well plate, and polyadenylated (PolyA+) messenger RNA (mRNA) was purified using the 96-well MultiMACS mRNA isolation kit on the MultiMACS 96 separator (Miltenyi Biotec, Germany) with on-column DNaseI-treatment according to the manufacturer's instructions. The eluted polyA+ mRNA was ethanol- precipitated and resuspended in 10µL of DEPC treated water with 1:20 SuperaseIN (Life Technologies, USA). Double-stranded cDNA was synthesized from the purified polyA+ RNA using the Superscript Double-Stranded cDNA Synthesis kit (Life Technologies, USA) and random hexamer primers at a concentration of 5µM. The cDNA was quantified in a 96-well format using PicoGreen (Life Technologies, USA) and VICTOR3V spectrophotometry (PerkinElmer, Inc. USA). The cDNA quality was checked on a random sampling using the High Sensitivity DNA chip Assay (Agilent). The cDNA was fragmented by a Covaris E210 (Covaris, USA) ultrasonicator for 55 seconds, using a Duty cycle of 20% and Intensity of 5. Plate-based libraries were prepared following the British Columbia Cancer Agency, Genome Sciences Centre (BCGSC) paired-end (PE) protocol on a Biomek FX robot (Beckman-Coulter, USA). Briefly, the cDNA was purified in 96-well format using Ampure XP SPRI beads, and was subject to end-repair and phosphorylation by T4 and Klenow DNA polymerases, and T4 polynucleotide kinase respectively in a single reaction, followed by cleanup using Ampure XP SPRI beads and 3' A-tailing by Klenow fragment (3' to 5' exo minus). After cleanup using Ampure XP SPRI beads, Picogreen quantification was performed to determine the amount of Illumina PE adapters used in adapter ligation reaction. The adapter-ligated products were purified using Ampure XP SPRI beads, then PCR-amplified with Phusion DNA Polymerase (Thermo Fisher Scientific Inc. USA) using Illumina's PE primer set, with cycle conditions of 98°C for 30sec followed by 10-15 cycles of 98°C for 10 sec, 65°C for 30 sec and 72°C for 30 sec, and finally 72°C for 5min. The PCR products were purified using Ampure XP SPRI beads, and checked with a Caliper LabChip GX for DNA samples using the High Sensitivity Assay (PerkinElmer, Inc. USA). PCR products with a desired size range were purified using a 96-channel size selection robot developed at the BCGSC, and the DNA quality was assessed and quantified using an Agilent DNA 1000 series II assay and Quant-iT dsDNA HS Assay Kit using Qubit fluorometer (Invitrogen), then diluted to 8nM. The final concentration was verified by Quant-iT dsDNA HS Assay prior to Illumina HiSeq2000 PE 75 base sequencing.

Alignment and coverage analysis of RNA-seq data

Using BWA (Burrows-Wheeler Aligner) version 0.5.7, we aligned chastity-passed reads (i.e. reads for which the bases have a ratio of the highest of the four (base type) intensities to the sum of highest two of ≥0.6) to an extended human reference genome consisting of hg19/GRCh37 plus exon junction sequences constructed from all known transcript models in RefSeq, EnsEMBL and UCSC genes[1]. We used default BWA parameter settings but disabled Smith-Waterman alignment. After alignment, the reads that aligned to exon junctions were repositioned in the genome as large-gapped alignments, using repositioning software developed in-house. We removed adapter dimer sequences and soft-clipped reads that contained adapter sequences. The unambiguously aligned, filtered reads were then analyzed by custom gene coverage analysis software to calculate the coverage over the total collapsed exonic regions in each gene as annotated in EnsEMBL (version 59), and RPKM values were calculated to represent the normalized expression level of exons and genes.

References

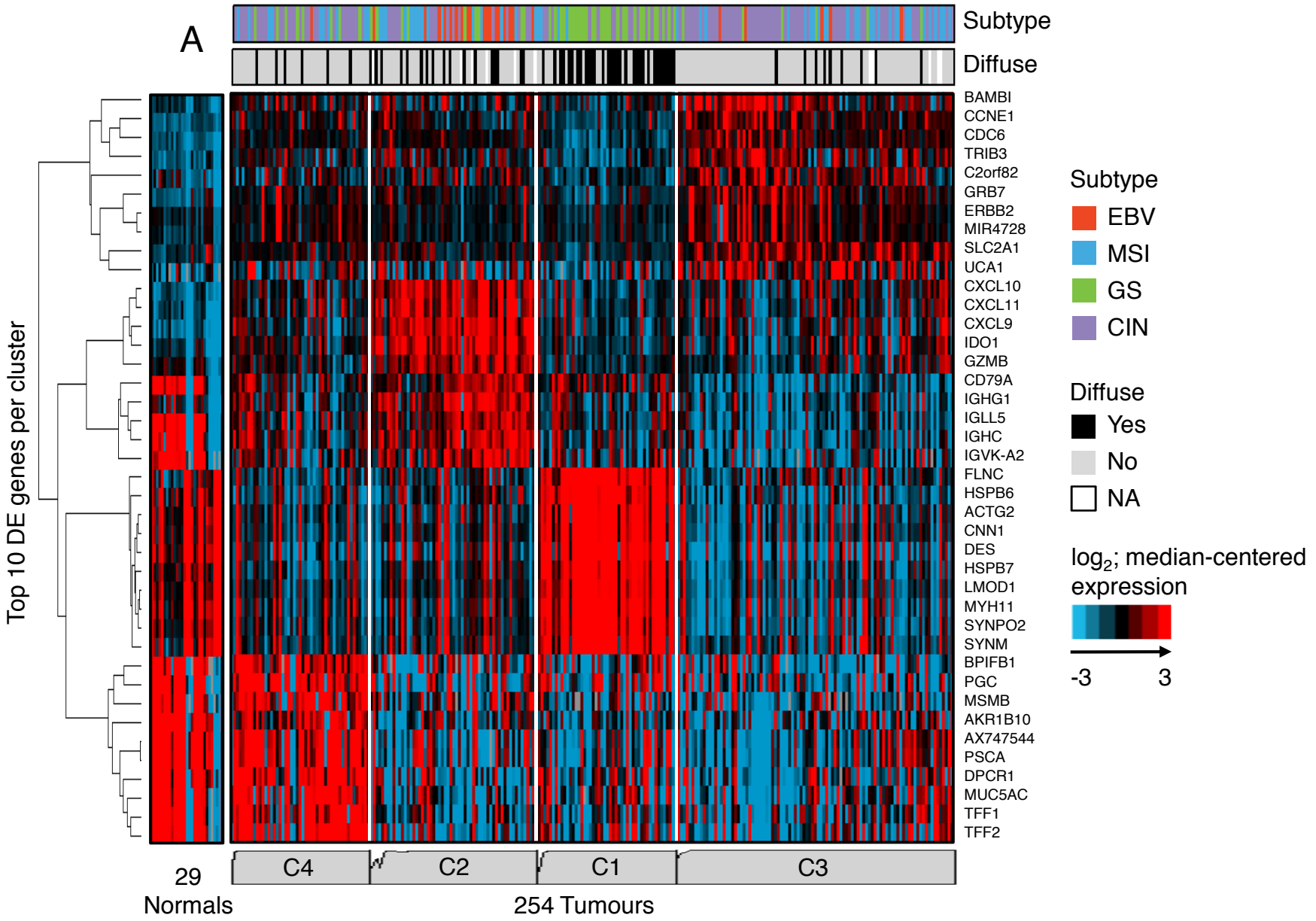
1. Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. 2008. *Biotechniques* 45(1):81-94 [PMID:18611170]

Supplement S5.2 NMF mRNA expression clustering

For 262 mRNA-Seq expression datasets, we first removed genes expressed at or below a noise threshold of ≤0.2 reads per kilobase (of gene model) per million mapped reads (RPKM) in at least 75% of samples. We then checked for outliers and batch effects using our principal component (PC) analysis tool *BLISS* [Batch anaLysIS Suite; http://www.bcgsc.ca/platform/bioinfo/software/bliss]. We removed two samples that were outliers in PC2 (TCGA-D7-6527, TCGA-BR-6710) in which *PGC* was over-expressed. The *PGC* gene encodes Progastricsin, a major component of the gastric mucosa. PC1 had a significant sample separation that was largely driven by high expression of mitochondrial (MT) genes. While this may reflect gastric cancer (GC) biology, we excluded MT genes from further analyses.

We created the NMF input matrix for 260 tumour samples using the top 25% most-variant genes, by ranking expressed genes having a mean RPKM of at least 10 by the coefficient of variation. We then removed six additional samples by iteratively creating an input matrix and removing *BLISS* outliers, until no further outliers were identified. To ensure that NMF solutions improved through this filtering process, at each iteration we ran NMF and checked for a cophenetic score above 0.98, and below 0.7 for the randomized data. The final NMF run used 254 tumour samples and 1,577 genes. Rank survey profiles for cophenetic and silhouette width, and consensus membership heatmaps (data not shown), suggested a four-cluster solution.

For mRNA-Seq data, we generated unsupervised consensus clustering results with NMF v0.5.02 in R v2.12.0, with the default Brunet algorithm, and 30 iterations for the rank survey and clustering runs. With the NMF output for mRNA-Seq data, we generated abundance heatmaps from the clustering input matrix as follows. The top differentially expressed genes in each cluster were used to filter the RNA-seq RPKM matrix for visualization. We reordered columns in each matrix into the NMF output order. Finally, we used Cluster v3.0 (bonsai.hgc.jp/~mdehoon/software/cluster) to log-transform and median-center each row, then to reorder rows using hierarchical clustering with Pearson correlation distance metric and average linkage.



Supplementary Figure S5.3: Unsupervised NMF consensus clustering of mRNA-Seq data. The heatmap shows normalized abundance for 254 tumour and 29 adjacent non-malignant tissue samples, for 40 discriminatory genes (i.e. top 10 differentially expressed genes in each cluster). Tumour samples (columns) are ordered based on a four-group NMF solution; adjacent non-malignant tissue samples are shown for contrast. Genes (rows) are ordered by hierarchical clustering of log-transformed, median-centered, RPKM data. In the profile below the heatmap, high silhouette widths show typical cluster members, while low silhouette widths show atypical group members.

Supplement S5.4 Gene fusion detection

RNA-seq libraries were assembled with ABySS (version 1.3.2 – http://www.bcgsc.ca/platform/bioinfo/software/abyss/releases) using k-mer values of 26 to 50 (for 50bp reads) and 38 to 74 (for 75bp reads) as previously described[1]. The contigs from assemblies were filtered, merged, aligned and post-processed using the Trans-ABySS pipeline (version 1.4.6 - http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss/releases) with parameters: -G hg19 g (gene model, g: TCGA.hg19.gaf.v3_0.txt); -f = 0.95 (minimum fusions fraction); -r –T /reference/genomes/human/hg19.fa (realign against hg19); -l = 500 (maximum subsequence length); -U (only count unique spanning reads and breakpoint pairs). Contigs were aligned against the reference genome using GMAP[2] with the following parameters: -d hg19 (genome database = hg19); -x 10 (threshold for the amount of unaligned sequence that is required before searching for the remaining sequence). Potential fusion candidates were identified as contigs that could not be mapped to a single unique location for at least 95% of the sequence. Such contig alignments with the following characteristics were considered candidates: both alignments had percentage of identity of at least 98%, one of the alignments did not reside entirely within its partner in terms of genomic coordinates, the alignments did not overlap by more than 5% in terms of contig coordinates, the alignments did not overlap in terms of genome coordinates, and the total coverage of the two alignments, in terms of contig sequence, was at least 90%. To further filter the candidate events from contig alignments, we used alignment of sequence reads to both contigs and the genome. Reads were aligned to the contigs using BWA 0.6.2-r126 and the parameter = bwasw (BWA-SW for long queries). Reads were aligned to the reference genome with exon-exon junctions using BWA 0.5.7, and reads that mapped across exon junctions were repositioned to their original genomic positions. Candidate fusion cases were then filtered by requiring at least two reads spanning the contig breakpoint with at least 4 flanking base pairs on either side, and at least two read pairs flanking the genomic breakpoint and pointing towards each other.

Partial and internal tandem duplications (PTDs and ITDs) were also reported from the RNA-seq data using the Trans-ABySS pipeline.

Candidate gene fusions, partial and internal tandem duplications identified from the RNA-seq data were compared with those of low-pass whole genome sequencing (WGS). Orthogonally verified structural variants were determined by enriching as follows:
1) All events in which the breakpoint in the RNA-seq data lay within 50,000 bp of the WGS genomic breakpoint coordinates, relative to hg19, were kept.
2) Those events identified in step 1 using the same gene symbol (or NA for Not Annotated) were kept.
3) Those events in step 2 that represent the same structural variant type (i.e. duplication, deletion, inversion, translocation) were kept.
Briefly, there were 95 tumours assayed with both low-pass WGS and RNA-seq, 44 of which have at least one fusion event detected by both data types. In total, 170 events were identified by this approach (Table S5.4) including the CLDN18-ARHGAP gene fusions (Table S5.6). For each of the 170 events, we surveyed all STAD RNA-seq datasets to determine the frequency in this cohort.

References

1. Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJ. 2009. *Bioinformatics* 25(21):2872-7. [PMID:19528083]
2. Wu T.D. and Watanabe C.K. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21(9):1859-75. [PMID:15728110]

Supplementary Data File S5.4a The data file- *Overlap list of RNA and whole genome sequencing events* can be found on the TCGA Stomach Adenocarcinoma publication page at https://tcga-data.nci.nih.gov/docs/publications/stad_2014/.

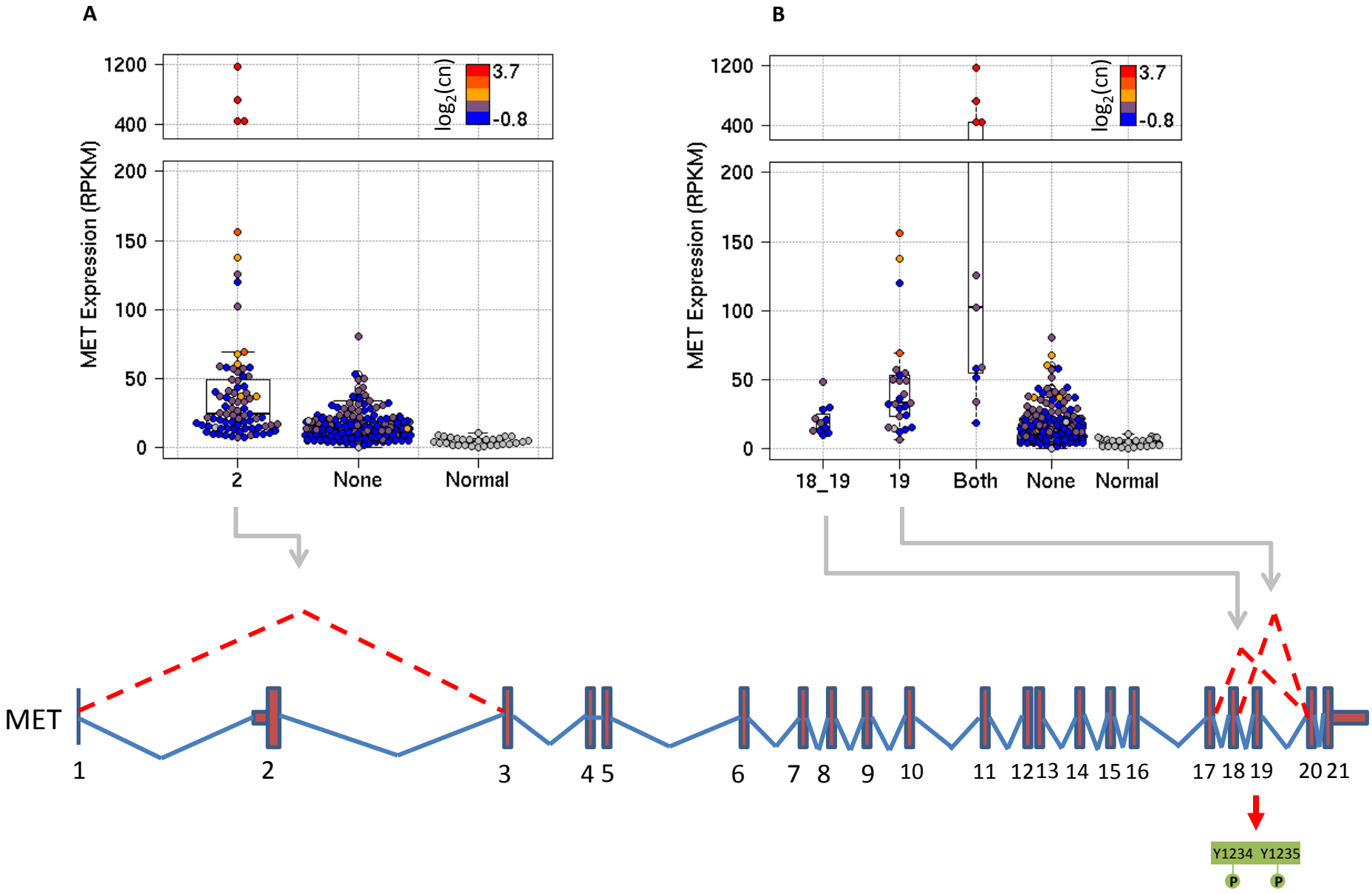


Figure S5.5: RPKM expression of MET in samples with and without exon 2 (A) or exon 18 and/or 19 (B) skipping in MET. The color scale shows $\log_2$ copy number (cn). Normal samples are in gray and have no copy number information. 82 samples have evidence for skipped exon 2, 36 for skipped exon 19 and 22 for skipped exons 18 and 19. MET exon 19 harbours the tyrosine protein kinase domain (green).

Supplementary Table S5.6 –CLDN18-ARHGAP [26/6] gene fusions identified in low-pass whole genome sequence and/or RNA sequence data

ID	Breakpoints (WGS)	Breakpoints (RNA-Seq)	5' gene	3' gene	Frame
TCGA-BR-8384-01A	No WGS sample	3:137749947|5:142393645	CLDN18	ARHGAP26	IN
TCGA-BR-8284-01A	No WGS sample	3:137749947|5:142393645	CLDN18	ARHGAP26	IN
TCGA-BR-6453-01A	No WGS sample	3:137749947|5:142393645	CLDN18	ARHGAP26	IN
TCGA-D7-8579-01A	No WGS sample	3:137749947|5:142393645	CLDN18	ARHGAP26	IN
TCGA-D7-8576-01A	No WGS sample	3:137749947|5:142393645	CLDN18	ARHGAP26	IN
TCGA-CG-4469-01A	3:137750512|5:142377782	3:137749947|5:142393645	CLDN18	ARHGAP26	IN
TCGA-CG-4462-01A	No WGS sample	3:137749949|5:142393645	CLDN18	ARHGAP26	IN
TCGA-BR-A4IU-01A	No WGS sample	3:137749949|5:142393647	CLDN18	ARHGAP26	IN
TCGA-BR-8367-01A	No WGS sample	3:137749953|5:142393653	CLDN18	ARHGAP26	IN
TCGA-BR-6852-01A	No WGS sample	3:137749947|5:142393645	CLDN18	ARHGAP26	NA[#]
TCGA-B7-5816-01A	3:137750740|5:142292613	3:137749947|5:142292764	CLDN18	ARHGAP26	IN
TCGA-B7-5816-01A	No WGS sample	3:137750565|5:142292834	ARHGAP26	CLDN18	OUT[$]
TCGA-D7-A4Z0-01A	No WGS sample	3:137749947|X:11272827	CLDN18	ARHGAP6	IN
TCGA-BR-8588-01A	No WGS sample	3:137749947|X:11272827	CLDN18	ARHGAP6	IN

[#] No assembled contig. A probe sequence derived from the fusion junction was aligned to sequence reads and 2 reads identified.

[$] Reciprocal fusion product

S5.7- Differentially expressed genes. We used SAMseq (samr v2.0, R 3.0.2) two-class unpaired analyses with an FDR threshold of 0.05 to identify genes that were differentially expressed. For each run on a pair of sample groups, we first reduced the number of genes by removing those with median less than 5 RPKM in both groups, and those for which the Wilcoxon BH adjusted P-value between the two groups was greater than 0.05. This subset of genes was submitted to SAMseq. Each run generated a pair of files: genes 'up' and 'down'. We then ranked the genes by a median-based fold change, and generated a figure showing up to 10 of the largest fold changes in each direction.

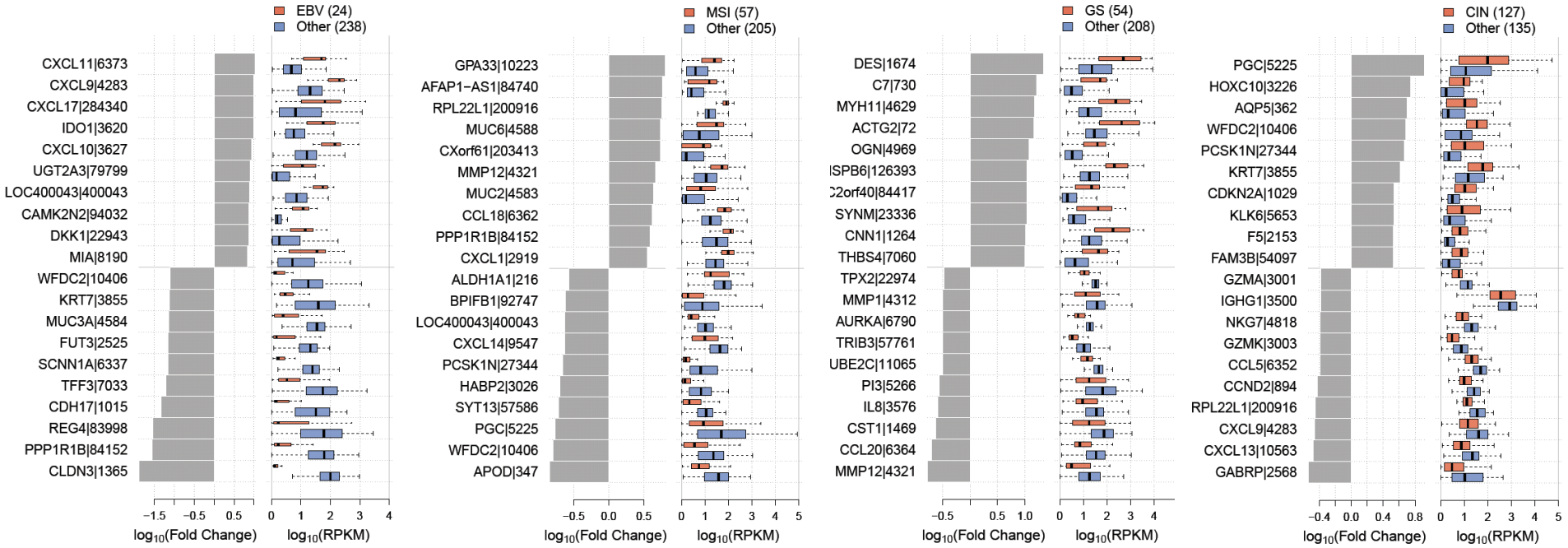
S5.7a The data file- Differentially expressed genes of multiple subtype combinations can be found on the TCGA Stomach Adenocarcinoma publication page at https://tcga-data.nci.nih.gov/docs/publications/stad_2014/.

Figure S5.8. Genes that were differentially abundant between the four molecular subtypes. Refer to pages 56-57.

S5.9 The data file- Top 20 least variable genes by coefficient of variation can be found on the TCGA Stomach Adenocarcinoma publication page at https://tcga-data.nci.nih.gov/docs/publications/stad_2014/.

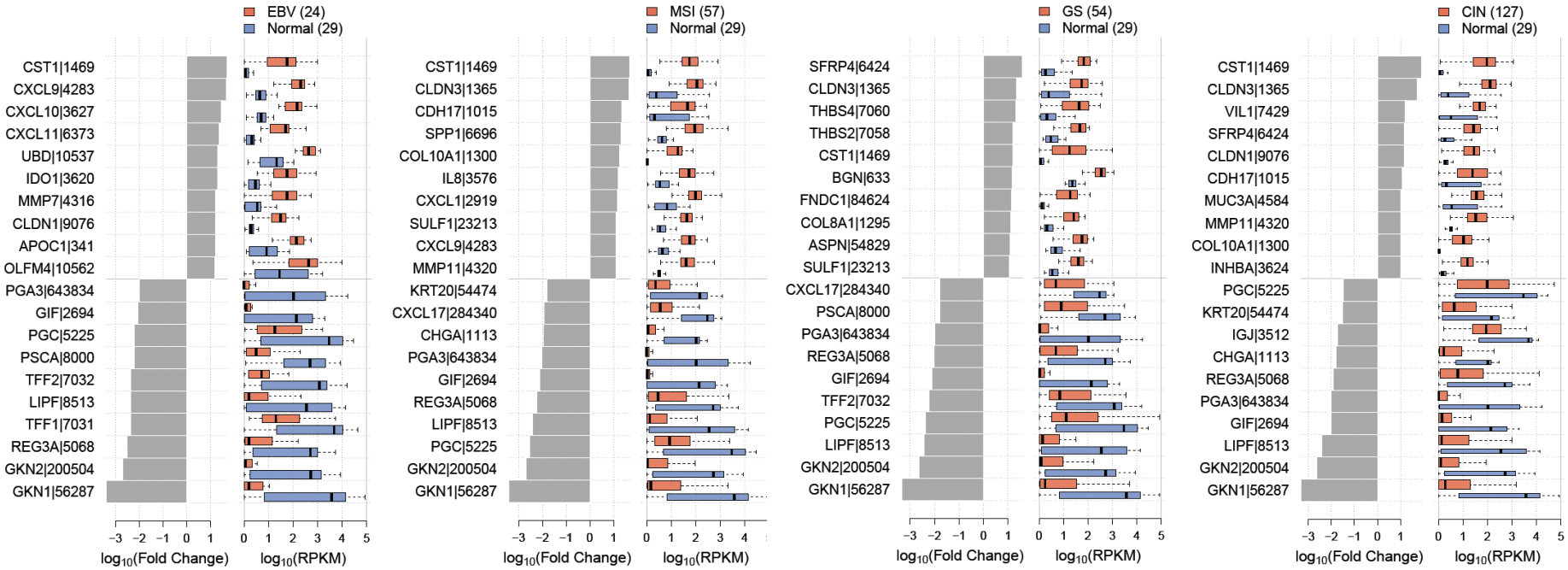
S5.8 figure- Differentially expressed genes

A



Supplementary Figure S5.8. Genes that were differentially abundant between the four molecular subtypes. a) Tumours in one subtype versus all other tumours. b) Each tumour subtype versus 29 adjacent non-malignant tissue samples. Left: median-based log10 fold change. Right: distributions of RPKM abundance, log10 scale, with black vertical lines showing medians. Up to 10 of the largest positive and negative fold changes satisfying FDR ≤ 0.05 are shown.

B



Supplementary Figure S5.8. Genes that were differentially abundant between the four molecular subtypes. a) Tumours in one subtype versus all other tumours. b) Each tumour subtype versus 29 adjacent non-malignant tissue samples. Left: median-based log10 fold change. Right: distributions of RPKM abundance, log10 scale, with black vertical lines showing medians. Up to 10 of the largest positive and negative fold changes satisfying FDR ≤ 0.05 are shown.

S6. miRNA Sequencing

S6 Section Authors:

Andrew Chu
A. Gordon Robertson
Andrew J. Mungall
Reanne Bowlby

Subsections:

Supplement S6.1 - miRNA library construction, sequencing and analysis

Two micrograms of total RNA per tumour sample was arrayed into 96-well plates, with controls as described below. RNA entering library construction was required to have at least a quality (RNA Integrity Number) >7.0 on the BCR submission documentation. Total RNA was mixed with oligo(dT) MicroBeads and loaded into a 96-well MACS column, which was then placed on a MultiMACS separator (Miltenyi Biotec, Germany). The separator's strong magnetic field allows beads to be captured during washes. From the flow-through after poly(A) selection for messenger RNA transcripts, small RNAs, including microRNAs (miRNA), were recovered by ethanol precipitation. Flow-through RNA quality was checked for a subset of 12 samples using an Agilent Bioanalyzer RNA Nano chip. miRNA-seq libraries were constructed using a plate-based protocol developed at the British Columbia Genome Sciences Centre (BCGSC). Negative controls were added at three stages: elution buffer was added to one well when the total RNA was loaded onto the plate, water to another well just before ligating the 3' adapter, and PCR reagents to a third well before PCR. A 3' adapter was ligated using a truncated T4 RNA ligase2 (NEB Canada, cat. no. M0242L) with an incubation of 1 hour at 22^{O}C. This adapter is an adenylated, single-strand DNA with the sequence 5'/5rApp/ ATCTCGTATGCCGTCTTCTGCTTGT /3ddC/, which selectively ligates miRNAs. An RNA 5' adapter was then added, using a T4 RNA ligase (Ambion USA, cat. no. AM2141) and ATP, and was incubated at 37^{O}C for 1 hour. The sequence of the single strand RNA adapter is 5'-GUUCAGAGUUCUACAGUCCGACGAUCUGGUCAA-3'.

When ligation was complete, first-strand cDNA was synthesized using Superscript II Reverse Transcriptase (Invitrogen, cat. no. 18064 014) and RT primer (5'-CAAGCAGAAGACGGCATACGAGAT-3'). This cDNA was the template for the final library PCR, into which we introduced index sequences to enable libraries to be identified from a sequenced pool that contains multiple libraries. Briefly, a PCR reagent mix was made with the 3' PCR primer (5'-CAAGCAGAAGACGGCATACGAGAT-3'), Phusion Hot Start High Fidelity DNA polymerase (NEB Canada, cat. no. F-540L), buffer, dNTPs and DMSO. The mix was distributed evenly into a new 96-well plate. A Biomek FX robot (Beckman Coulter, USA) was used to transfer the PCR template (first-strand cDNA) and indexed 5' PCR primers into the reagent mix plate. Each indexed 5' PCR primer, (5'-AATGATACGGCGACCACCGACAGNNNNNNGTTCAGAGTTCTACAGTC CGA-3'), contained a unique six- nucleotide 'index' (shown here as N's), and was added to each well of the 96-well PCR reagent plate. PCR thermocycling conditions were 98°C for 30 sec, followed by 15 cycles of 98°C for 15 sec, 62°C for 30 sec and 72°C for 15 sec, and finally a 5 min incubation at 72^{O}C.

Quality was checked across the whole plate using a Caliper LabChip GX DNA chip (Perkin Elmer). PCR products were pooled, then size-selected to remove larger cDNA fragments and smaller adapter contaminants, using an in-house 96-channel automated size-selection robot. After size-selection, each pool was ethanol-precipitated, quality-checked using an Agilent Bioanalyzer DNA1000 chip and quantified using a Qubit fluorometer (Invitrogen, cat. no. Q32854). Each pool was then diluted to a target concentration for cluster generation and was loaded into a single lane of an Illumina GAIIx or HiSeq 2000 flow cell. Clusters were generated, and lanes were sequenced with a 31 bp main read for the insert and a 7 bp read for the index.

Preprocessing, alignment and annotation of miRNA

Briefly, the sequence data were separated into individual samples based on the index read sequences, and the reads underwent an initial QC assessment. Adapter sequence was then trimmed off, and the trimmed reads for each sample were aligned to the NCBI GRCh37-lite reference genome. Below we describe these steps in more detail.

Routine QC assessed a subset of raw sequences from each pooled lane for the abundance of reads from each indexed sample in the pool, the proportion of reads that possibly originated from adapter dimers (i.e. a 5' adapter joined to a 3' adapter with no intervening biological sequence) and for the proportion of reads that map to human miRNAs. Sequencing error was estimated by a method originally developed for Serial Analysis of Gene Expression[2].

Libraries that passed this QC stage were preprocessed for alignment. While the size-selected miRNAs varied somewhat in length, typically they were ~21 bp long, and so were shorter than the 31 bp read length. Given this difference, each read sequence extended some distance into the 3' sequencing adapter. Because this non-biological sequence could interfere with aligning the read to the reference genome, 3' adapter sequence was identified and removed (trimmed) from a read. The adapter-trimming algorithm identified as long an adapter sequence as possible, allowing a number of mismatches that depended on the adapter length found. A typical sequencing run yielded several million reads; using only the first (5') 15 bases of the 3' adapter in trimming made processing efficient, while minimizing the chance that a miRNA read would match the adapter sequence.

The algorithm first determined whether a read sequence should be discarded as an adapter dimer by checking whether the 3' adapter sequence occurred at the start of the read. For reads passing this stage, the algorithm then searched for an exact 15 bp match anywhere within the read sequence. If a match was not found, the algorithm then started the search from the 3' end, allowing up to 2 mismatches. If the full 15 bp was not found, decreasing lengths of adapter were checked, down to the first 8 bases, allowing one mismatch. If a match was still not found, from 7 bases down to 1 base was checked, with an exact match required. Finally, the algorithm trimmed 1 base off the 3' end of a read if the sequence matched the first base of the adapter. This was based on two considerations; First, it was preferable to get a perfect alignment than an alignment that had a potential one-base mismatch. Second, if only 1 base of adapter was found in the read sequence, the read was likely too long to be from a miRNA and the effect of the trimming on its alignment would not affect this sample's overall miRNA profiling result.

After each read was processed, a summary report was generated containing the number of reads at each read length. Because the shortest mature miRNA in miRBase v16 is 15 bp, any trimmed read that was shorter than 15 bp was discarded; remaining reads were submitted for alignment to the reference genome.

Burrows-Wheeler alignment(s)[3] for each read were checked with a series of three filters. A read with more than 3 alignments was discarded as ambiguous. For TCGA quantification reports, only perfect alignments with no mismatches were used. Based on comparing expression profiles of test libraries (data not shown), reads that failed the Illumina base-calling chastity filter (The chastity of a base call is the ratio of the intensity of the greatest signal divided by the sum of the two greatest signals) were retained, while reads that have soft-clipped CIGAR strings (a series of operation lengths plus the operations that store sequence alignment information) were discarded.

For reads retained after filtering, each coordinate for each read alignment was annotated using the reference databases (Table S6.1), using a requirement of a minimum 3 bp overlap between the alignment and an annotation. In annotating reads, we addressed two potential issues. First, a single read alignment could overlap feature annotations of different types; second, a read could have up to three alignment locations, and each alignment location could overlap a different type of feature annotation. By considering heuristically-determined priorities (Table S6.1), we resolved the first issue by giving each alignment a single annotation. We resolved the second by collapsing multiple annotations to a single annotation, as follows:

If a read had more than one alignment location, and the annotations for these were different, we used the priorities from Table S6.1 to assign a single annotation to the read, as long as only one alignment was to a miRNA. When there are multiple alignments to different miRNAs, the read was flagged as cross-mapped[1], and all of its miRNA annotations were preserved, while all of its non-miRNA annotations were discarded. This ensured that all annotation information about ambiguously mapped miRNAs was retained, and allowed annotation ambiguity to be addressed in downstream analyses. Note that

we considered miRNAs to be cross-mapped only if they mapped to different miRNAs, not to functionally identical miRNAs that are expressed from different locations in the genome. Such cases are indicated by miRNA miRBase names, which can have up to 4 separate sections separated by "-", e.g. hsa-mir-26a-1. A difference in the final (e.g. '-1') section denotes functionally equivalent miRNAs expressed from different regions of the genome, and we considered only the first 3 sections (e.g. 'hsa-mir-26a') when comparing names. As long as a read mapped to multiple miRNAs for which the first 3 sections of the name were identical (e.g. hsa-mir-26a-1 and hsa-mir-26a-2), the read was treated as if it mapped to only one miRNA, and was not flagged as cross-mapped.

From the profiling results for a tumour type, for a minimum of approximately 100 samples, we identified the depth of sequencing required to detect the miRNAs that were expressed in a sample by considering a graph of the number of miRNAs detected in a sample as a function of the number of reads aligned to miRNAs. For the current work, a library from a sequenced pool was required to have at least 1,000,000 reads mapped to miRBase annotations. For any sequencing run that failed to meet this threshold, we sequenced the sample again to achieve at least the minimum number of miRNA-aligned reads.

Finally, for each sample, the reads that corresponded to particular miRNAs were summed and normalized to a million miRNA-aligned reads to generate the quantification files that were submitted to the DCC. Quantification files included information on variable 5' and 3' read alignment locations, which can reflect isoforms, adapter trimming and RNA degradation.

Table S6.2. Annotation priorities that are used to resolve multiple database matches for a single alignment location and multiple alignment locations for a read.

Priority	Annotation type	Database
1	mature strand	miRBase v16
2	star strand	
3	precursor miRNA	
4	stemloop, from 1 to 6 bases outside the mature strand, between the mature and star strands	
5	"unannotated", any region other than the mature strand in miRNAs where no star strand is annotated	
6	snoRNA	UCSC small RNAs, RepeatMasker
7	tRNA	
8	rRNA	

9	snRNA	
10	scRNA	
11	srpRNA	
12	Other RNA repeats	
13	coding exons with zero annotated CDS region length	UCSC knownGenes
14	3' UTR	
15	5' UTR	
16	coding exon	
17	intron	
18	LINE	UCSC RepeatMasker
19	SINE	
20	LTR	
21	Satellite	
22	RepeatMasker DNA	
23	RepeatMasker Low complexity	
24	RepeatMasker Simple Repeat	
25	RepeatMasker Other	
26	RepeatMasker Unknown	

References

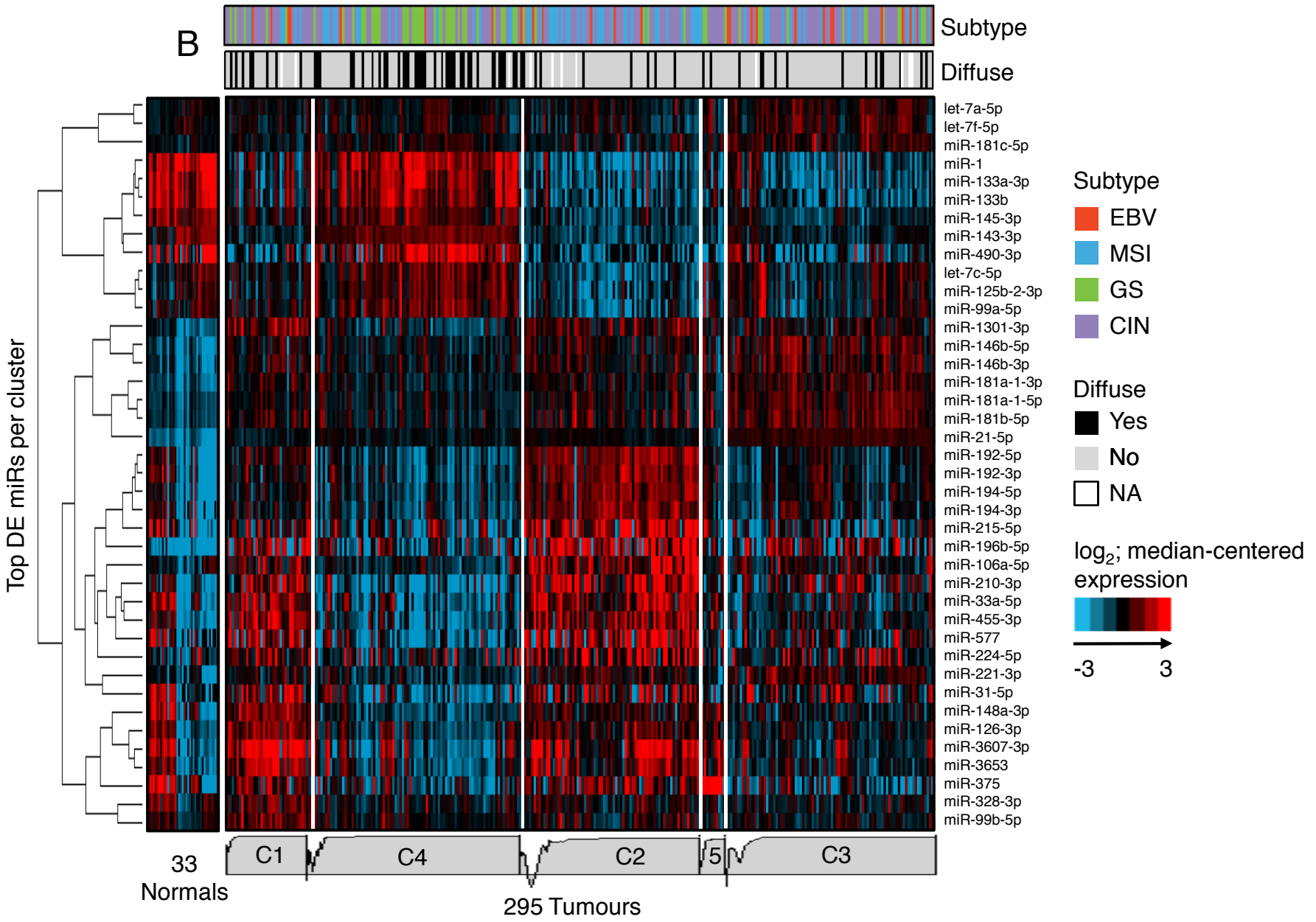
1. de Hoon MJ, Taft RJ, Hashimoto T, Kanamori-Katayama M, Kawaji H, Kawano M, Kishima M, Lassmann T, Faulkner GJ, Mattick JS, Daub CO, Carninci P, Kawai J, Suzuki H, Hayashizaki Y. Cross-mapping and the identification of editing sites in mature microRNAs in high-throughput sequencing libraries. *Genome Res*. 2010. 20(2):257-64. [PMID:20051556]
2. Khattra J, Marra MA. Large-scale production of SAGE libraries from microdissected tissues, flow-sorted cells and cell lines. *Genome Res*. 2007. 17(1):108-16. [PMID:17135571]
3. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009. 25(14):1754-60. [PMID:19451168]

Supplement S6.3. NMF expression clustering

For miRNA-seq data, read counts for 295 tumour samples were normalized to RPM, i.e. to reads per million reads aligned to miRBase 5p and 3p strands. Strands corresponding to miRNAs that had been removed from v19 miRBase (miRNA.dead) were eliminated from the data matrix. 5p and 3p strands were ranked by RPM variance across the samples, and the most variant 25% (304 MIMATs) were retained.

We generated unsupervised consensus-clustering results with NMF v0.5.06 in R v2.12.0, with the default Brunet algorithm, and 200 iterations for the rank survey and clustering runs. Given NMF outputs for miRNA-Seq data, we generated abundance heatmaps from the clustering input matrix as follows. The top differentially expressed miRNAs in each cluster were used to filter the miRNA-seq RPM matrix for visualization. We reordered columns in the matrix into the NMF output order. Finally, we used Cluster v3.0 (bonsai.hgc.jp/~mdehoon/software/cluster) to log-transform and median-center each row, then to reorder rows using hierarchical clustering with Pearson correlation distance metric and average linkage.

Figure S6.4: Unsupervised NMF consensus clustering of miRNA-Seq data



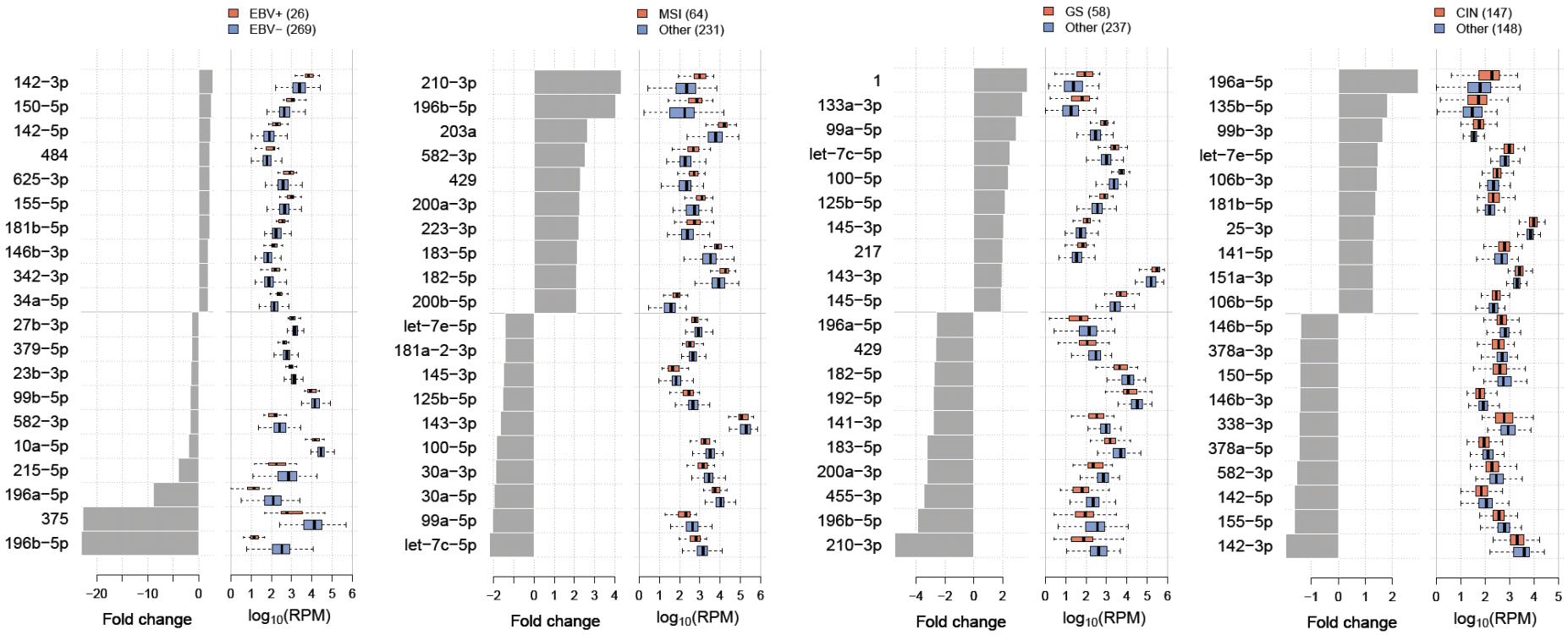
Supplementary Figure Legend S6.4: Unsupervised NMF consensus clustering of miRNA-Seq data. The heatmap shows normalized abundance for 295 tumour and 33 adjacent non-malignant samples, for 40 discriminatory 5p or 3p strands (i.e. top differentially expressed miRNAs in each cluster). Tumour samples (columns) were ordered based on a five-group NMF solution. miRNAs (rows) were ordered by hierarchical clustering of log-transformed, median-centered, RPM data.

Supplement S6.5. Differentially expressed miRs.
We used SAMseq (samr v2.0, R 3.0.2) two-class unpaired analyses with an FDR threshold of 0.05 to identify miRs that were differentially expressed. Each run generated a pair of files: miRs 'up' and 'down'. We filtered each file by removing miRs with median expression less than 50 RPM in both of the input sample groups, and miRs for which the Wilcoxon BH adjusted P-value was greater than 0.05; then ranked the filtered results by a median-based fold change, and generated a figure showing up to 10 of the largest fold changes in each direction.

Supplementrary Data File S6.7 The data file *Differentially expressed miRs* can be found on the TCGA Stomach Adenocarcinoma publication page at https://tcga-data.nci.nih.gov/docs/publications/stad_2014/.

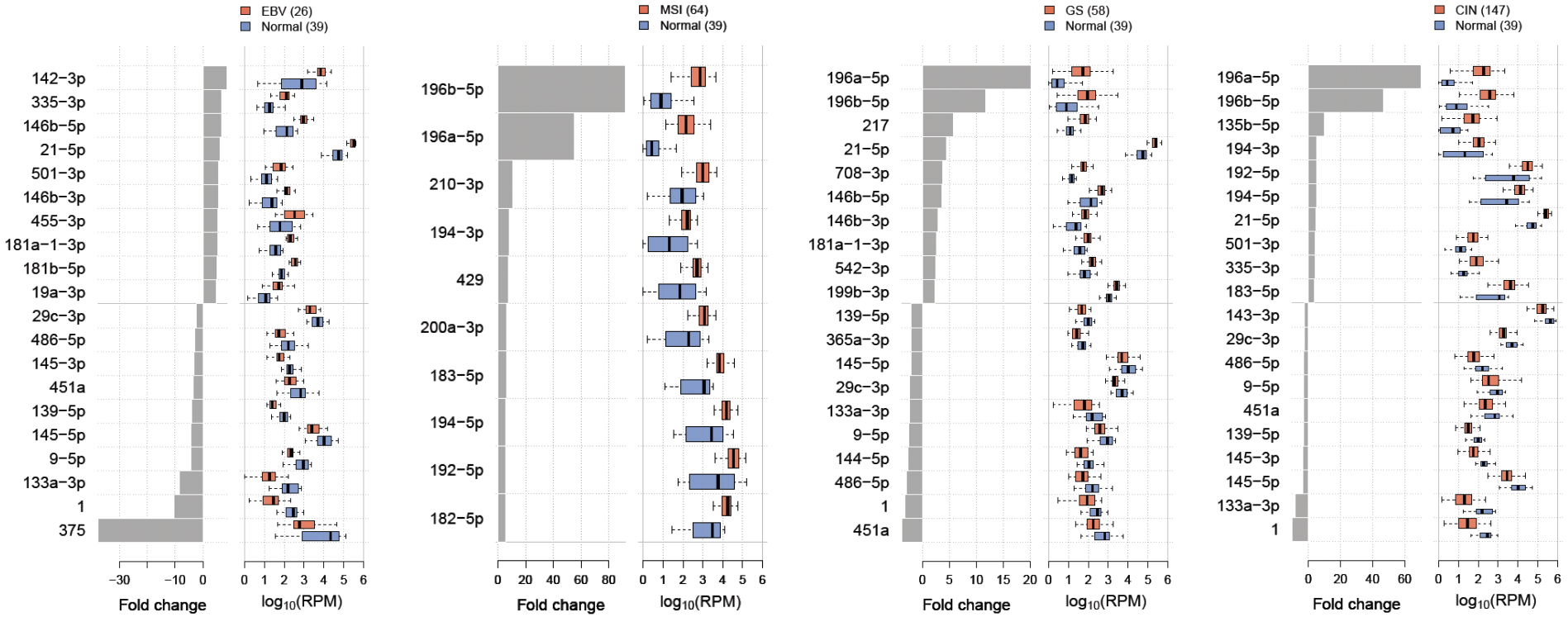
Figure S6.6a. miRs that are differentially abundant between the four molecular subtypes

A



Supplementary Figure S6.6. miRs that are differentially abundant between the four molecular subtypes. a) Tumours in one subtype versus all other tumours. b) Each tumour subtype versus 29 adjacent non-malignant tissue samples. Left: median-based fold change, linear scale. Right: distributions of RPM abundance, log10 scale, with black vertical lines showing medians. Up to 10 of the largest positive and negative fold changes satisfying FDR ≤ 0.05 are shown.

B



Supplementary Figure S6.6. miRs that are differentially abundant between the four molecular subtypes. a) Tumours in one subtype versus all other tumours. b) Each tumour subtype versus 29 adjacent non-malignant tissue samples. Left: median-based fold change, linear scale. Right: distributions of RPM abundance, log10 scale, with black vertical lines showing medians. Up to 10 of the largest positive and negative fold changes satisfying FDR ≤ 0.05 are shown.

S7. Reverse-Phase Protein Array

S7 Section Authors:

Yiling Liu
Wenbin Liu
Sang-Bae Kim
Gordon B. Mills
Ju-Seog Lee

Subsections:

S7.1 REVERSE PHASE PROTEIN ARRAY (RPPA)

Methods.
Protein was extracted using RPPA lysis buffer (1% Triton X-100, 50 nmol/L Hepes (pH 7.4), 150 nmol/L NaCl, 1.5 nmol/L MgCl2, 1 mmol/L EGTA, 100 nmol/L NaF, 10 nmol/L NaPPi, 10% glycerol, 1 nmol/L phenylmethylsulfonyl fluoride, 1 nmo/L Na3VO4, and aprotinin 10 Ag/mL) from human tumours and RPPA was performed as described previously[1-5]. Lysis buffer was used to lyse frozen tumours by Precellys homogenization. Tumour lysates were adjusted to 1 µg/µL concentration as assessed by bicinchoninic acid assay (BCA) and boiled with 1% SDS. Tumour lysates were manually diluted in fivefold serial dilutions with lysis buffer. An Aushon Biosystems 2470 arrayer (Burlington, MA) printed 1,056 samples on nitrocellulose-coated slides (Grace Bio-Labs). Slides were probed with 191 validated primary antibodies (see table below) followed by corresponding secondary antibodies (Goat anti-Rabbit IgG, Goat anti-Mouse IgG or Rabbit anti-Goat IgG). Signal was captured using a DakoCytomation catalyzed system and DAB colorimetric reaction. Slides were scanned in CanoScan 9000F. Spot intensities were analyzed and quantified using ArrayPro (http://www.mediacy.com/index.aspx?page=ArrayPro), to generate spot signal intensities (Level 1 data). The software SuperCurveGUI[3,5], available at http://bioinformatics.mdanderson.org/Software/supercurve/) was used to estimate the EC50 values of the proteins in each dilution series (in log2 scale). Briefly, a fitted curve ("supercurve") was plotted with the signal intensities on the Y-axis and the relative log2 concentration of each protein on the X-axis using the non-parametric, monotone increasing B-spline model[1]. During the process, the raw spot intensity data were adjusted to correct spatial bias before model fitting. A QC metric[5] was returned for each slide to help determine the quality of the slide: if the score was less than 0.8 on a 0-1 scale, the slide was omitted. In most cases, the staining was repeated to obtain a high quality score. If more than one slide was stained with an antibody, the slide with the highest QC score was used for analysis (Level 2 data). Protein measurements were corrected for loading as described[3,5,6] using median-centering across antibodies (level 3 data). In total, 191 antibodies and 255 samples were used (255 of which were represented in the extended sample set) and 224 of which were represented in the core sample set in Figure 1). Final selection of antibodies was also driven by the availability of high quality antibodies that consistently passed a strict validation process as previously described[7]. These antibodies were assessed for specificity, quantification and sensitivity (dynamic range) in their application for protein extracts from cultured cells or tumour tissue. Antibodies were labeled as validated or used with caution based on the degree of validation by criteria previously described[7].

Raw data (level 1), SuperCurve nonparameteric model fitting on a single array (level 2), and loading corrected data (level 3) were deposited at the TCGA Data Coordinating Center (DCC).

Unsupervised hierarchical clustering analysis
We used ConsensusClusterPlus package in R v2.13.2 to identify robust subtypes of gastric adenocarcinoma based on protein expression[8]. The consensus clusters were obtained from 1,000 resampling iterations of the hierarchical clustering, by randomly selecting a fraction of the samples and of the highly variable protein features (standard deviation > 0.3).

Unsupervised hierarchical clustering analysis revealed three robust RPPA clusters (**Figure S7-1**). These three RPPA subtypes were significantly correlated with histology (Pearson's chi–squared test, P=0.002) as well as with the subtypes/clusters defined by other genomic data including DNA methylation clusters (P=0.006), *MHL1* hypermethylation (P=0.006), MSI (0.001), microRNA clusters (P=3.9 $\times 10^{-15}$), and mRNA clusters (P=7.4$\times 10^{-8}$). RPPA cluster 1 (reactive) were associated with the diffuse subtype and expressed high levels of CAV1, MYH11, and RICTOR, likely reflecting activation of stromal cells in the tumour microenvironment, and showed strong concordance with mRNA cluster 1 and microRNA cluster 4. Cluster 2 (invasive) was associated with low expression of CTNNB1 and CDH1 that play key roles in cell adhesion, suggesting that tumours in cluster 2 might have increased potential of invasion and metastasis (i.e., an EMT phenotype). Cluster 3 (proliferative) was characterized by high expression of proteins involved in cell proliferation such as PCNA, CCNB1, TIGAR, MTOR, and FOXM1.

Supervised analysis
Supervised analysis of the RPPA profiling data identified 45 protein features that were significantly associated (Student's t-test, $P < 0.001$) with molecular subtypes of gastric cancer (Figure S7.2). The EBV subtype had elevated expression of CASP7, PCNA, BAX, SYK, and LCK while the MSI subtype had elevated expression of CLDN7, VHL, and CCNB1. Likewise, expression of KIT, MYC, AKT, and PRKCA was highly elevated in the GS subtype. Phosphorylation of EGFR (py1068) was elevated in the CIN subtype,

consistent with higher amplification of EGFR in the CIN subtype. Expression of p53, suggestive of increased levels of DNA damage and the presence of *TP53* mutation, was also highly elevated in the CIN subtype.

References

1. Tibes R, Qiu Y, Lu Y, et al: Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells.
Molecular Cancer Therapeutics 5:2512-2521, 2006
2. Liang J, Shao SH, Xu Z-X, et al: The energy sensing LKB1-AMPK pathway regulates p27kip1 phosphorylation mediating the decision to enter autophagy or apoptosis. Nat Cell Biol 9:218-224, 2007.
3. Hu J, He X, Baggerly KA, et al: Non-parametric quantification of protein lysate arrays.
Bioinformatics 23:1986-1994, 2007.
4. Hennessy BT, Lu Y, Poradosu E, et al: Pharmacodynamic Markers of Perifosine Efficacy.
Clinical Cancer Research 13:7421-7431, 2007.
5. Coombes K, Neeley S, Joy C, et al: SuperCurve: SuperCurve Package. R package version 1.4.1. 2011.
6. Gonzalez-Angulo A, Hennessy B, Meric-Bernstam F, et al: Functional proteomics can define prognosis and predict pathologic complete response in patients with breast cancer. Clin Proteomics 8:11, 2011
7. Hennessy B, Lu Y, Gonzalez-Angulo A, et al: A Technical Assessment of the Utility of Reverse Phase Protein Arrays for the Study of the Functional Proteome in Non-microdissected Human Breast Cancers. Clin Proteomics 6:129-151, 2010.
8. Wilkerson MD, Hayes DN: ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 26:1572-3, 2010

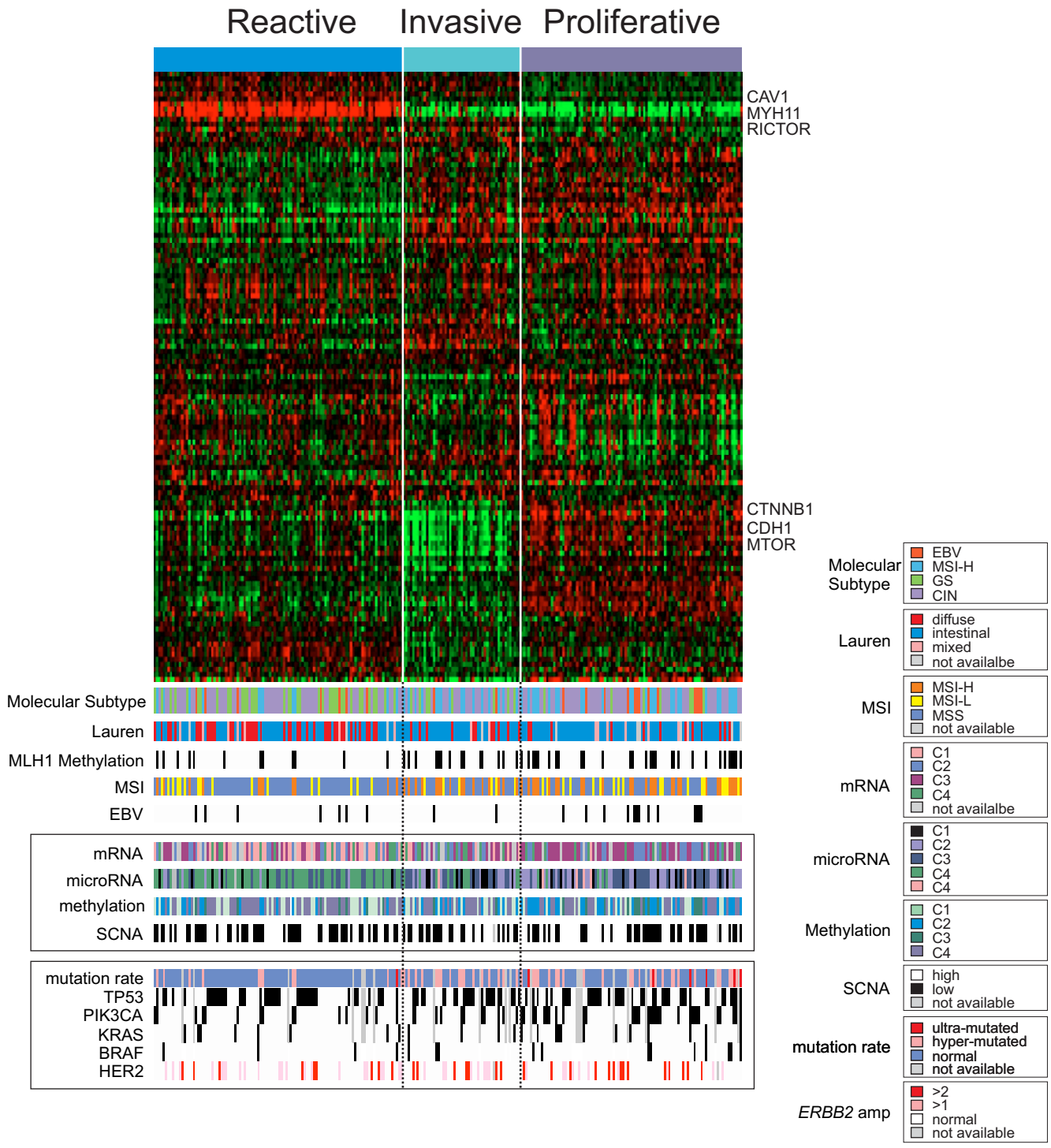


Figure S7.2. Unsupervised hierarchical clustering of RPPA data.

Three clusters were identified by consensus cluster algorithm. Annotation bars are provided at the bottom of the heatmap.

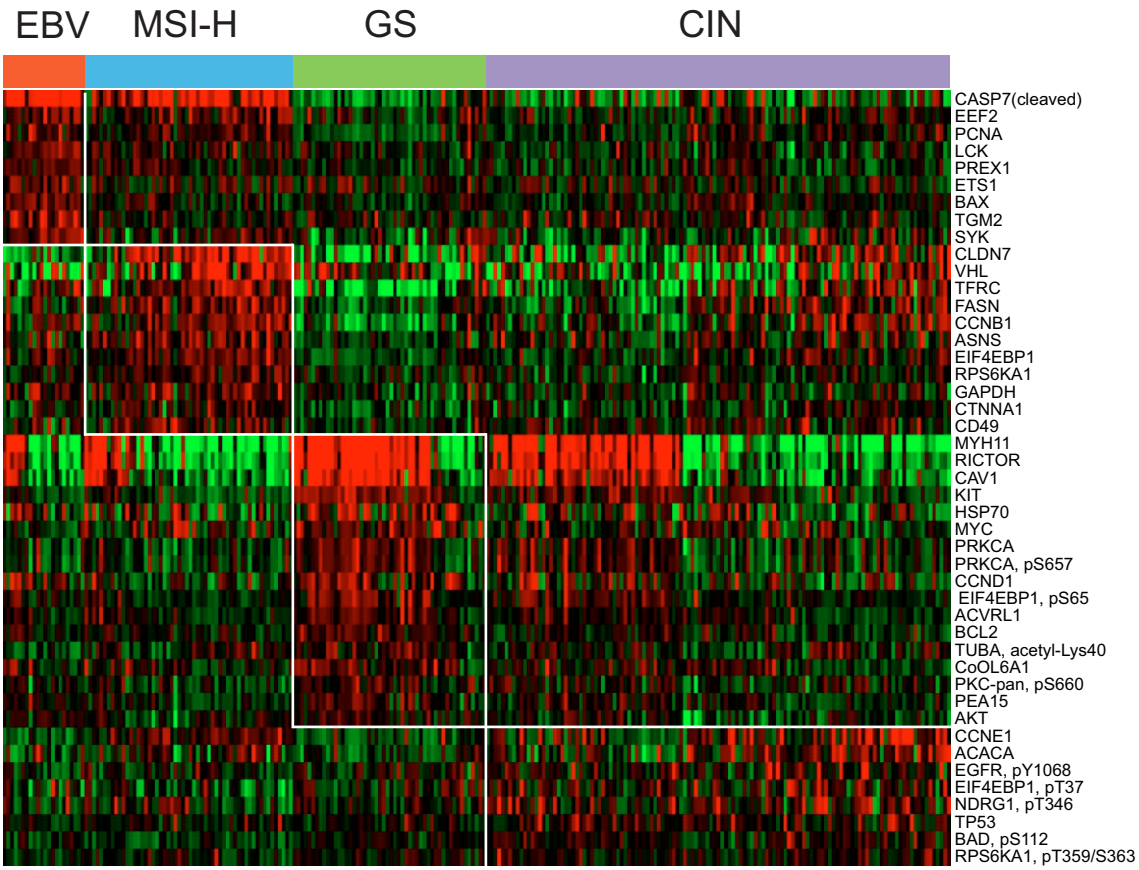


Figure S7.3. Supervised clustering of RPPA data according to molecular subtypes.

S7.4 The data file- List of antibodies used for sample profiling by RPPA can be found on the TCGA Stomach Adenocarcinoma publication page at https://tcga-data.nci.nih.gov/docs/publications/stad_2014/.

S8. Batch Effects Analysis

S8 Section Authors:

Rehan Akbani
Shiyun Ling
Arvind Rao
John N. Weinstein

Subsections:

Supplement S8: Batch effects analysis for TCGA gastric cancer data sets

S8.1 Supplemental Methods:

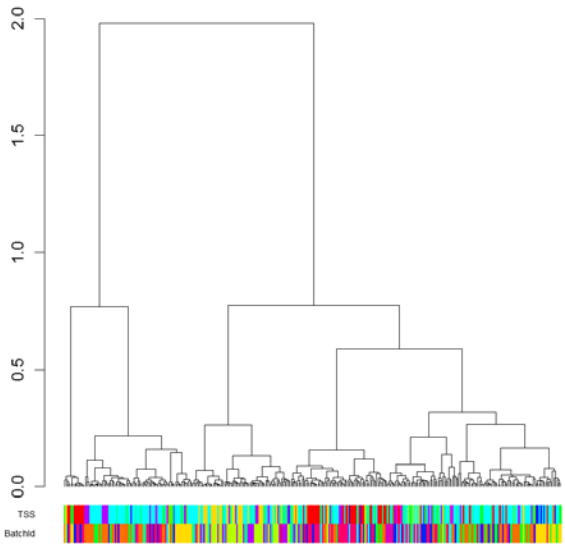
We used hierarchical clustering and Principal Components Analysis (PCA) to assess batch effects in the gastric cancer data sets. Four different data sets were analyzed: miRNA sequencing (Illumina HiSeq), DNA methylation (Infinium HM450 microarray), mRNA sequencing (Illumina HiSeq), and protein expression (Reverse Phase Protein Array (RPPA)). All of the data sets were at TCGA level 3, since that's the level on which most of the analyses in the paper are based. We assessed batch effects with respect to two variables; batch ID and Tissue Source Site (TSS). Detailed results and batch effects analysis of other TCGA data sets can be found at: http://bioinformatics.mdanderson.org/tcgabatcheffects

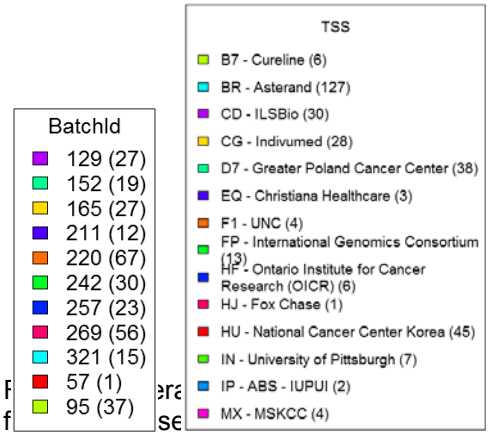
For hierarchical clustering, we used the average linkage algorithm with 1 minus the Pearson correlation coefficient as the dissimilarity measure. We clustered the samples and then annotated them with colored bars at the bottom. Each color corresponded to a batch ID or a TSS. For PCA, we plotted the first four principal components, but only plots of the first two components are shown here. To make it easier to assess batch effects, we enhanced the traditional PCA plot with centroids. Points representing samples with the same batch ID (or TSS) were connected to the batch centroid by lines. The centroids were computed by taking the mean across all samples in the batch. That procedure produced a visual representation of the relationships among batch centroids in relation to the scatter within batches. The results for the four data sets follow.

miRNA (RNA-seq Illumina HiSeq)

Figures S8.2-S8.4 show clustering and PCA plots for miRNA seq data. miRNAs with zero values were removed and the read counts were $\log_2$-transformed before generating the figures. The figures show a small batch effect by batch ID, where a slight dichotomy can be observed (batches 95, 129, 152). However, the magnitude of the batch effect wasn't too great, so we didn't think that it warranted batch effect correction for the type of analyses done in this paper. The trade off with batch effects correction algorithms is the possibility of losing important biological variation in the data, along with the technical variation.

Legends





NA expression

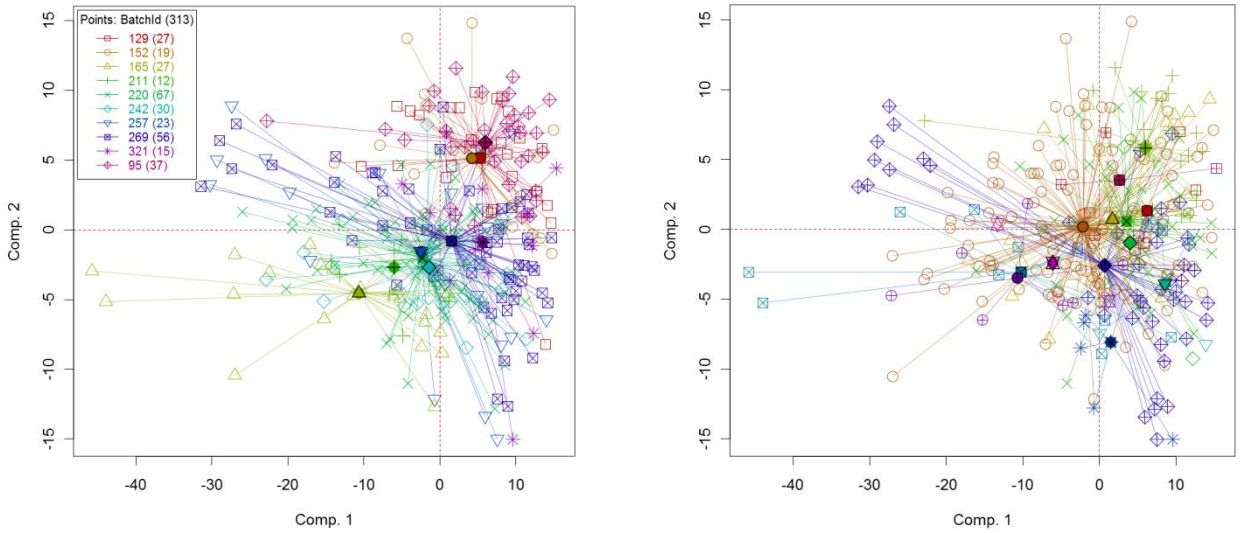


Fig. S8.3. PCA: First two principal components for miRNA expression from miRNA-seq data, with samples connected by centroids according to batch ID.

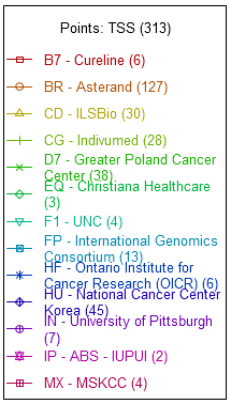
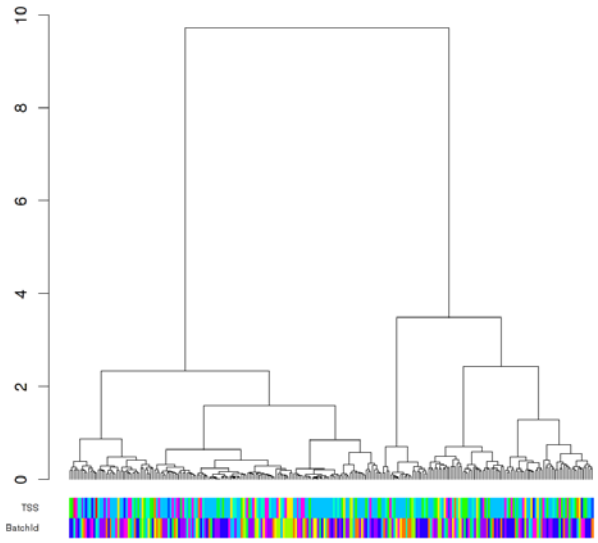


Fig. S8.4. PCA: First two principal components for miRNA expression from miRNA-seq data, with samples connected by centroids according to TSS.

DNA Methylation (Infinium HM450 microarray)

Figures S8.5-S8.7 show clustering and PCA plots for the Infinium DNA methylation platform. None of the batches or tissue source sites stood apart from the others, indicating no serious batch effects were present.



Legends

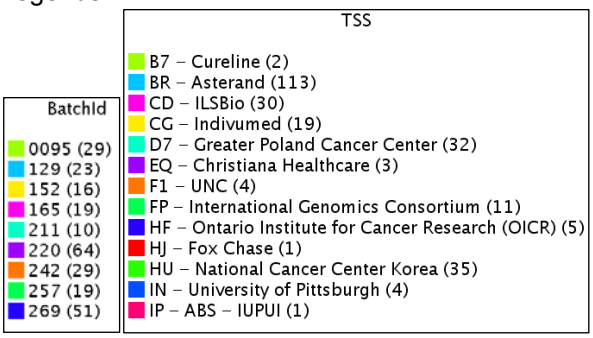


Fig. S8.5. Hierarchical clustering plot for DNA methylation data.

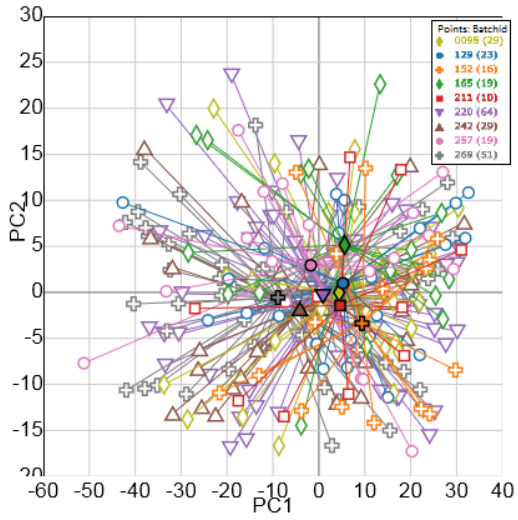


Fig. S8.6. PCA for DNA methylation, with samples connected by centroids according to batch ID.

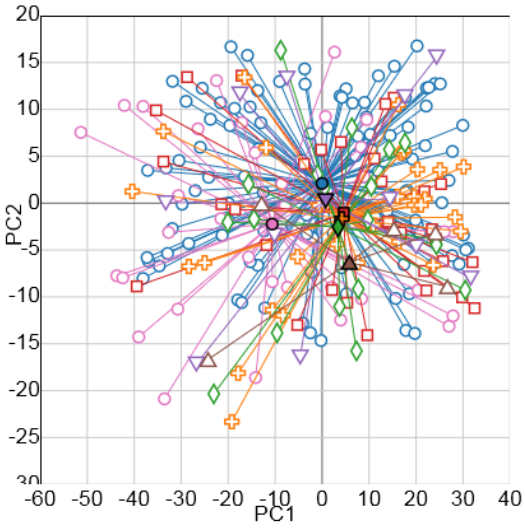
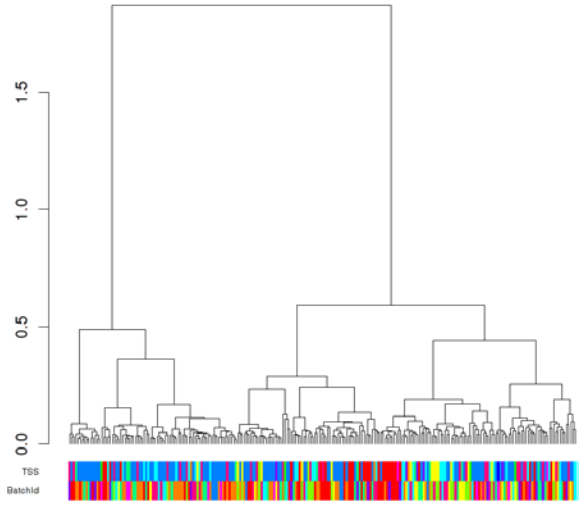


Fig. S8.7. PCA for DNA methylation, with samples connected by centroids according to TSS.

RNASeqV2 (RNA-Seq Illumina HiSeq)

Figures S8.8-S8.10 show clustering and PCA plots for the RNA-seq platform. Genes with zero values were removed and the values were $\log_2$-transformed before generating the figures. Once again, none of the batches or tissue source sites stood apart from the others, indicating no major batch effects were present.



Legends

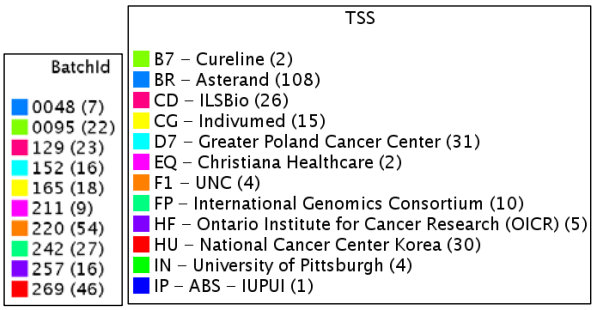


Fig. S8.8. Hierarchical clustering for mRNA expression from RNA-seq data

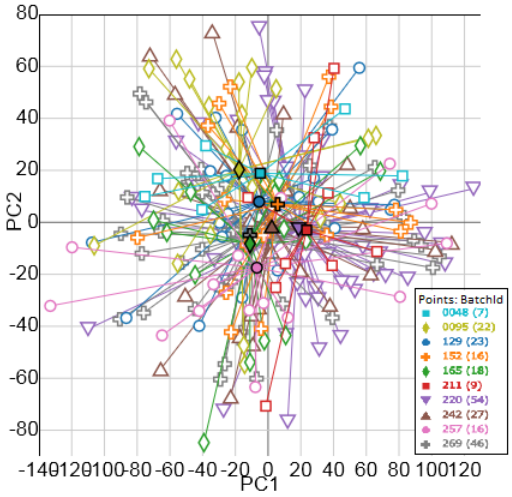
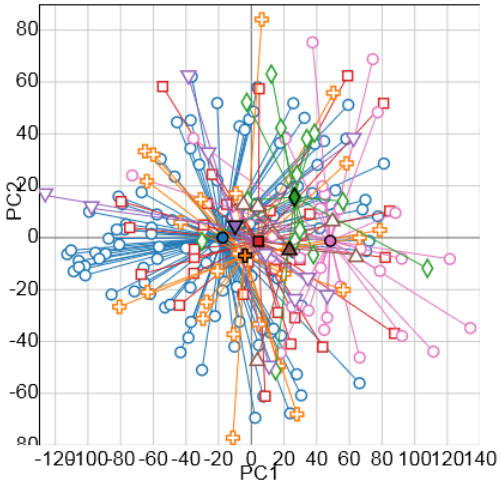


Fig. S8.9. PCA: First two principal components for RNA-seq, with samples connected by centroids according to batch ID.



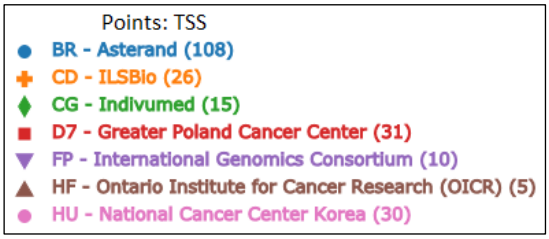
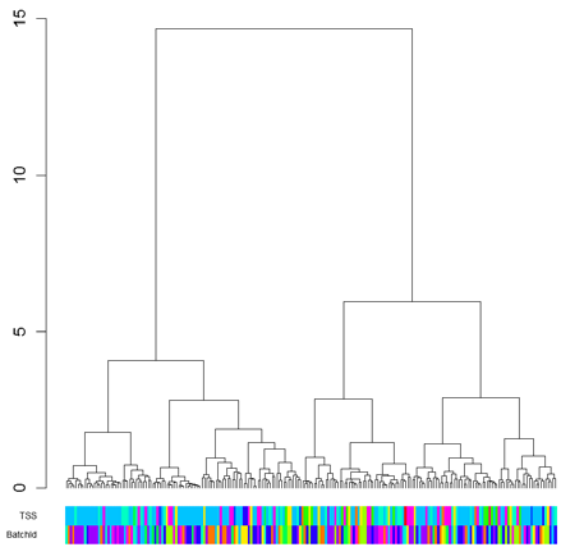


Fig. S8.10. PCA: First two principal components for RNA-seq, with samples connected by centroids according to TSS.

Protein expression (RPPA)

Figures S8.11-S8.13 show clustering and PCA plots for the Reverse Phase Protein Array (RPPA) platform. None of the batches or TSSs stood apart from the others, indicating no major batch effects were present.



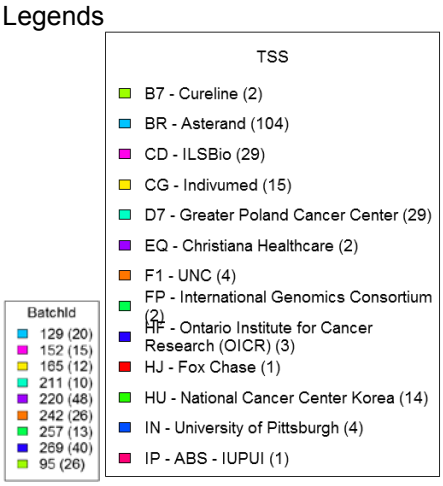


Fig. S8.11. Hierarchical clustering for protein expression data

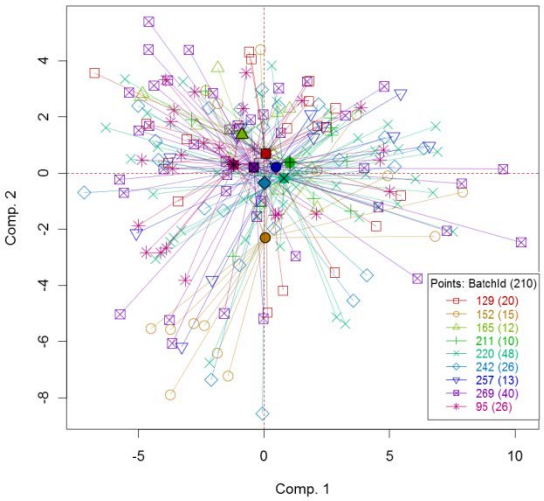


Fig. S8.12. PCA: First two principal components for protein expression data, with samples connected by centroids according to batch ID.

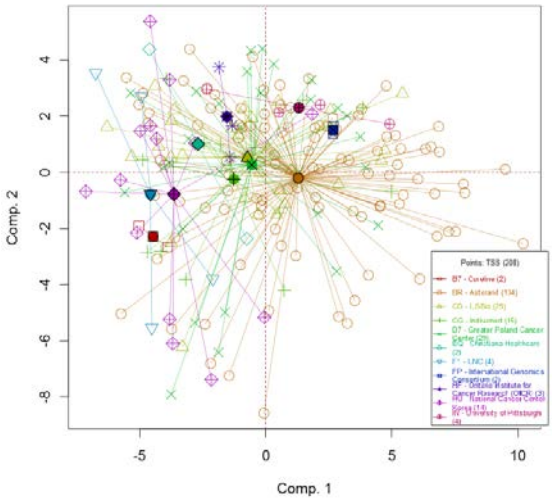


Fig. S8.13. PCA: First two principal components for protein expression data, with samples connected by centroids according to TSS.

S8.14 Conclusions

Batch effects were analyzed in four different data sets. miRNA data showed a small batch effect in samples from the batches 95, 129, 152. However, the batch effects weren't considered strong enough to warrant algorithmic batch effects correction, since that often removes useful biology along with batch effects. DNA methylation, mRNA seq and protein expression data didn't show any major batch effects.

S9. Microbiome Analysis

S9 Section Authors:

Reanne Bowlby
M. Constanza Camargo
Andy Chu
Margaret L. Gulley
Joonil Jung
Andrew J. Mungall
Akinyemi I. Ojesina
Michael Parfenov
Chandra Sekhar Pedamallu
Charles S. Rabkin
Barbara G. Schneider
Vésteinn Thorsson.

Subsections:

S9.1 Microbial Detection in mRNA-Seq

To detect microbial transcripts in mRNA-Seq data, we developed an implementation of Bloom filters[1] termed *BioBloom*, which efficiently tests whether or not a read pair derives from a target microbe. We divided each 75bp mRNA read sequence into three adjacent 25bp tiles and classified the tiles against 45 Bloom filters that we generated from human, bacterial, viral, fungal, and vector sequences downloaded from the National Center for Biotechnology Information (ftp://ftp.ncbi.nlm.nih.gov/genomes/). A read pair was considered a positive match when 2 of the 3 tiles from each read in the pair matched a given filter. Read pairs matching multiple filters were removed from further analyses.
We applied *BioBloom* to 237 TCGA STAD tumours (samples sequenced on Illumina GAII instruments with 50bp reads were excluded from this analysis) and 29 adjacent non-malignant tissue mRNA-Seq datasets. Counts were normalized by millions of quality reads sequenced (*number of reads mapped to the microbe*10^6/number of chastity-passed reads*). We identified 24 EBV-positive tumours (10%) that had between 4 and 300 normalized counts, averaging 88. These counts were 209-fold higher than those of EBV-negative samples, including adjacent non-malignant tissue, which had an average of 0.4 normalized counts ranging from 0 to 1.3.
Samples with more than zero read-pairs classified as EBV by *BioBloom* went on to EBV gene-expression analysis. For this analysis, BWA-0.5.7 was used to align reads to a custom-created reference based on the NCBI EBV type 1 complete genome and gene annotations (NC_007605.1). Reads with an alignment spanning exon-exon junctions were then transformed into large gapped genomic alignments using JAGuaR. Reads with a mapping quality of 10 or greater were included in the gene expression quantification analysis. Results were normalized to reads per kilobase of exon per million reads mapped to the EBV transcriptome (*number of reads mapped to the gene * (1000/gene length) * (10^6/total EBV reads aligned)*).

S9.2 Microbial Detection in miRNA-Seq data

We extracted unaligned reads from .bam files aligned to the human reference genome (GRCh37) and re-aligned these to a meta-genome constructed from multiple bacterial, viral and fungal sequences retrieved from ftp://ftp.ncbi.nlm.nih.gov/genomes/, and Illumina adapter sequences. Reads that aligned to only one microbial reference genome were summed to produce a total count for each microbe for each sample, and totals were normalized against the human miRNA read counts for that library (*number of reads mapped to the microbe*10^6/number of reads aligned to human miRNAs*). Reads from TCGA placental controls were processed in the same way, and a Wilcoxon test was used to compare distributions of normalized counts between STAD and control samples.
This independent analysis of miRNA libraries constructed from the same total RNA samples revealed that 26 EBV-positive STAD tumour samples (1 had no mRNA-Seq data) contained an average of >500,000 depth-normalized reads that mapped to known EBV miRNAs, while EBV- negative samples contained <200 reads. Results from these two RNA platforms were in complete agreement with EBV-positive calls from whole exome sequence and/or low-pass whole genome sequence data.

S9.3 Identification of a chimaeric human-EBV transcript

In the 24 tumour samples identified as EBV-positive, we searched for evidence of chimaeric transcripts in the mRNA-Seq data. *De novo* assembly of the mRNA-Seq data from the tumour sample TCGA-FP-7998 revealed a single 507bp contig representing a potential human-EBV chimaeric gene transcript (Figure S9.4). Performing BLAST against the National Center for Biotechnology Information non-redundant nucleotide database revealed that bases 1 to 410 of this contig had perfect homology with the complete genome sequence of EBV (human herpesvirus 4) type II strain AG876 (DQ279927.1). Bases 404 to 507 of the sequence contig showed 100% identity to human chromosome 9 (Genome Reference Consortium GRCh37 positions chr9:5431895-5431985 and 5436568-5436583) corresponding to the plasminogen receptor, C-terminal lysine transmembrane (*PLGRKT)* gene positions 261 to 362 of NM_018465.3. The contig represents a gene fusion between exon 3 of *PLGRKT* and EBV gene *BHLF1* at an AG/GT splice site (AC/CT on the reverse strand, Figure S9.4). A six-frame translation of the contig revealed a stop codon in the BHLF1 portion of the peptide sequence. However, there is evidence that BHLF1 has a non-coding function[2] and thus we cannot exclude the possibility that the fusion product was functional.
In support of the *de novo*-assembled chimaeric transcript, we identified 15 x 75bp reads spanning the fusion breakpoint. Of note, a 7bp sequence (AC^CTGAA) at the fusion junction was shared between EBV and

human genome loci and thus microhomology-mediated end joining[3] may explain the mechanism for the fusion transcript.

References:

1. Burton H. Bloom (1970) Space/time trade-offs in hash coding with allowable errors *Communications of the ACM* **13**(7);422-426
2. Lin, Z et al. Whole-genome sequencing of the Akata and Mutu Epstein-Barr virus strains. *J. Virol*. 2013 Jan;87(2):1172-82.
3. McVey, M. and Lee, S.E. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. Trends Genet 2008 Nov;24(11):529-38

Figure S9.4 EBV-human chimeric transcript

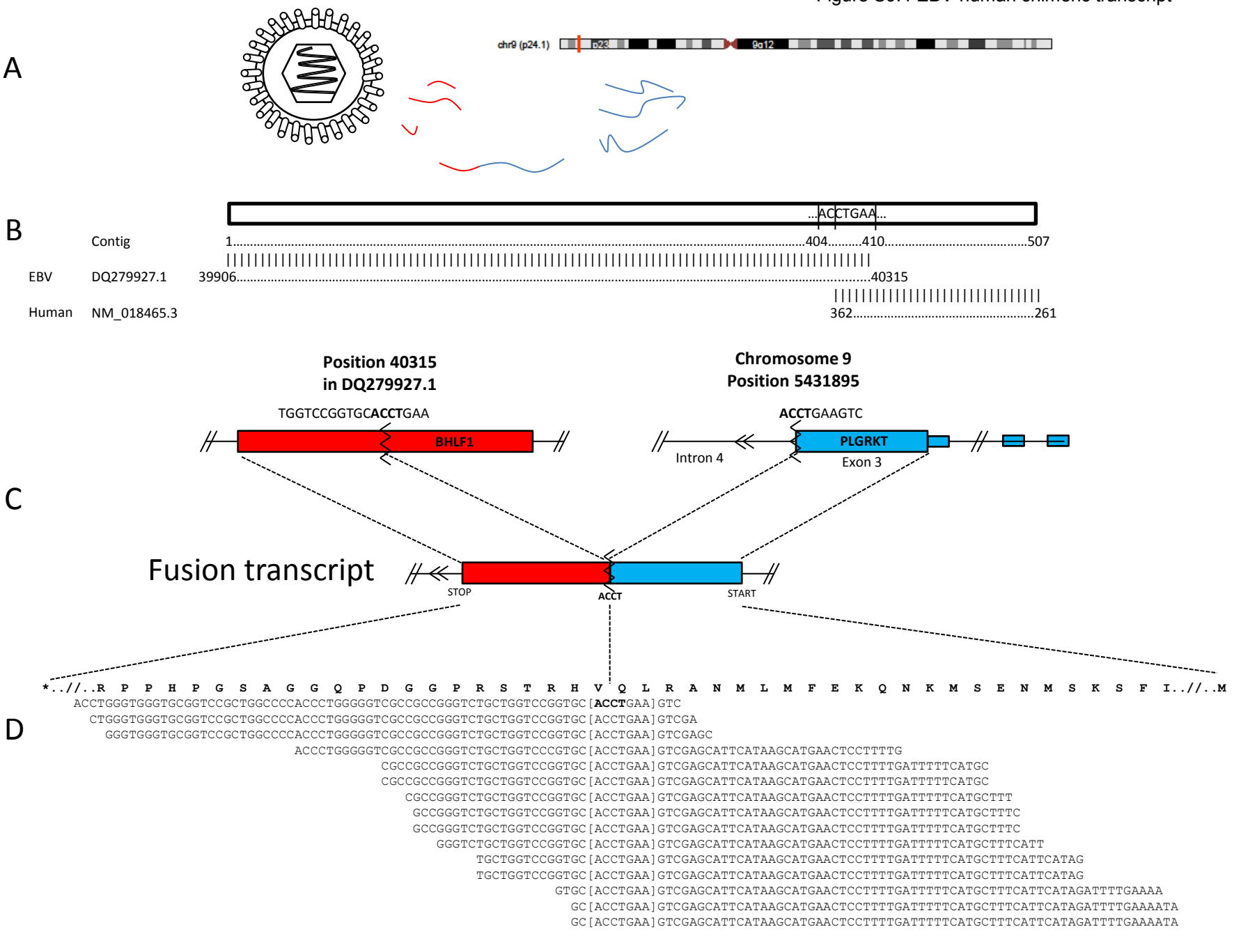


Figure S9.4 – Identification of a chimaeric human-EBV transcript

Human and EBV transcripts, including chimaeras can be found in the mRNA-Seq data (A). A de novo assembly of the mRNA-Seq data from the tumour sample TCGA-FP-7998 revealed a single 507bp contig representing a potential human-EBV chimaeric gene transcript. Bases 1 to 410 of this contig exhibited perfect homology with the complete genome sequence of Epstein Barr Virus (EBV or human herpesvirus 4) strain AG876 (DQ279927.1). Bases 404 to 507 of the sequence contig showed 100% identity to human chromosome 9 (Genome Reference Consortium GRCh37 positions chr9:5431895-5431985 and chr9:5436568-5436583) corresponding to the plasminogen receptor, C-terminal lysine transmembrane (PLGRKT) gene positions 261 to 362 of NM_018465.3 (B). The contig represents a gene fusion between exon 3 of PLGRKT and EBV gene BHLF1 at an AG/GT splice site (AC/CT on the reverse strand, panel C). Six-frame translation of the contig revealed a stop codon (*) in the BHLF1 portion of the peptide sequence. In support of the assembled chimaeric transcript contig, we identified 15 x 75bp reads spanning the fusion breakpoint (D). A 7 base-pair sequence exhibiting microhomology is shown in square brackets and contains the splice site AC^CT (reverse strand) in bold type.

S9.5 Epstein-Barr virus and *Helicobacter pylori* detection methods in WGS and WES

The PathSeq[1] algorithm was used to perform computational subtraction of human reads, followed by alignment of residual reads to human reference genomes and microbial reference genomes (which includes bacterial, viral, archaeal, and fungal sequences - downloaded from NCBI in June, 2012). These alignments resulted in the identification of reads mapping with Epstein-Barr virus (EBV; also referred to as human herpes virus 4, HHV4) and *Helicobacter pylor*i in whole genome sequencing (WGS) and whole exome sequencing (WES) data.
In brief, human reads were subtracted by first mapping reads to a database of human genomes (downloaded from NCBI in November 2011) using BWA[2] (Release 0.6.1, default settings), Megablast (Release 2.2.25, cut-off E-value 10^{-7}, word size 16) and Blastn[3] (Release 2.2.25, cut-off E-value 10^{-7}, word size 7, nucleotide match reward 1, nucleotide mismatch score -3, gap open cost 5, gap extension cost 2). Only sequences with perfect or near perfect matches to the human genome were removed in the subtraction process. In addition, low complexity and highly repetitive reads were removed using Repeat Masker[4] (version open-3.3.0, libraries dated 2011-04-19).

To identify EBV and *H. pylori* reads, the residual reads were aligned with Megablast to a database of microbial and human reference genomes. Raw read counts were calculated using the reads that were mapped to EBV and *H. pylori* with at least 90% identity and 90% query coverage.

Using the raw read counts, the abundance metric of a given microbe in a sample was calculated as

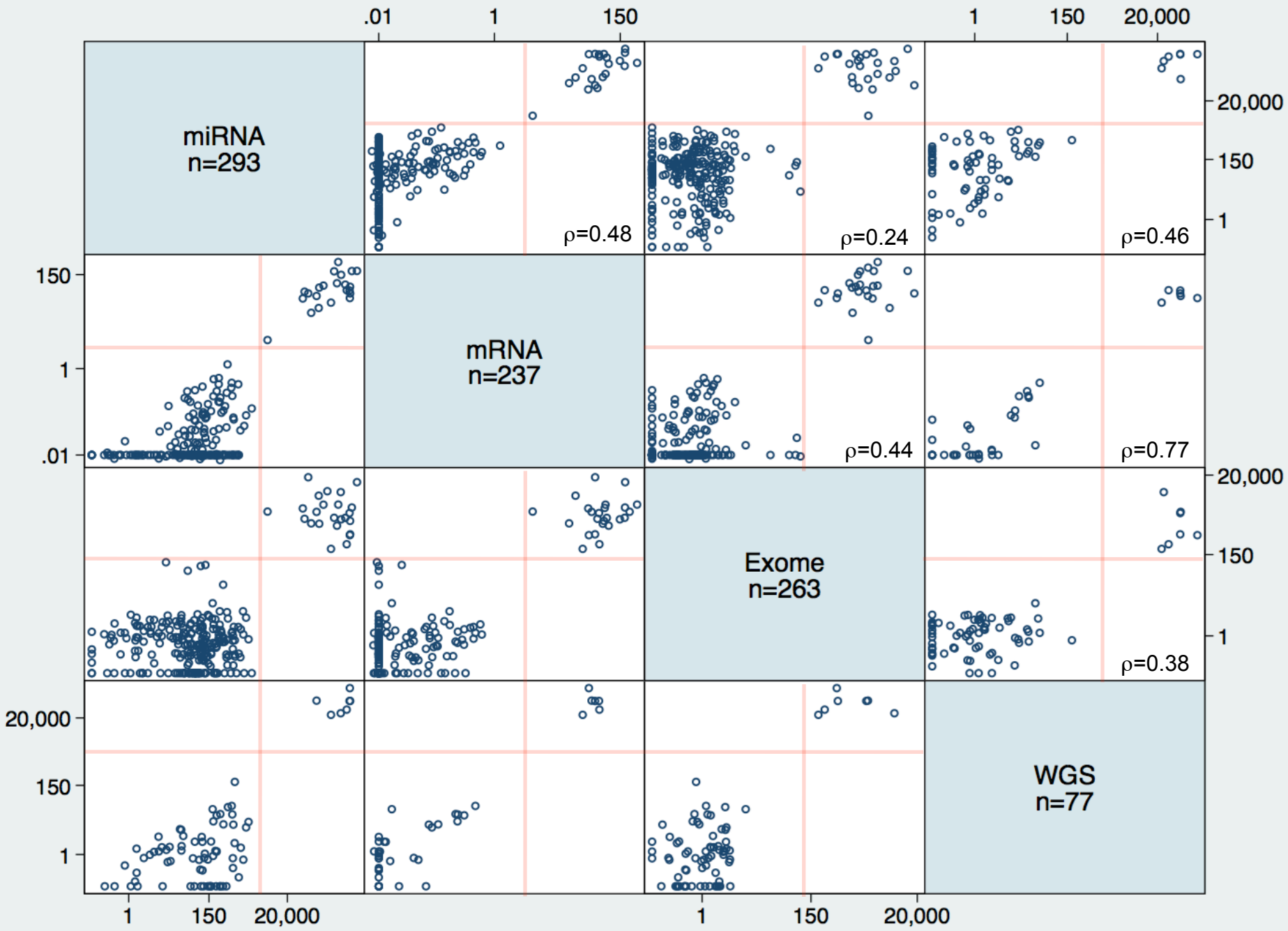
$$Abundance\ \ metric = \left(\frac{\#\,reads\ \ mapped\ \ to\,the\,microbe}{\left(\frac{\#\,reads\ \ mapped\ \ to\,human\,in\,the\,sample}{Average\ \#\,reads\ \ mapped\ \ to\,human\,in\,the\,sample\ cohort} \right) * \left(\frac{Genome\ \ size\,of\ \ the\,microbe}{Average\ \ \ genome\ \ size\,of\ \ the\,microbes\ \ in\,that\ kingdom} \right)} \right)$$

Samples were considered to be EBV positive if the abundance metric exceeded 1000 by WGS or 100 by WES.

1. Kostic, A.D. et al. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat Biotechnol 29, 393-6 (2011).
2. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-60 (2009).
3. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25, 3389-402 (1997).
4. Smit, AFA. et al. RepeatMasker Open-3.0. 1996-2010 <http://www.repeatmasker.org>.

S9.6 text-Sequencing-based determination of tumour EBV status.

All 295 gastric tumour samples were analyzed for the presence of viral nucleic acids by at least two sequencing platforms. EBV read counts were determined by miRNA (n=293), mRNA (n=237), whole exome (n=263) and whole genome (WGS; n=77) sequencing and normalized to human sequence counts, as described above (S9.1, S9.2 and S9.5). Numbers of EBV reads in individual samples were bimodally distributed, with distinct separation of a minority of tumours having much higher counts for each platform (Figure S9.7). Empiric cut-offs (shown as red lines) of 5000 for miRNA, 4 for mRNA, 100 for exome and 1000 for WGS had perfect concordance for identifying 26 (9%) EBV-positive samples among the tumours analyzed. Quantitative counts were moderately correlated, with Spearman rank correlation coefficients (ρ) ranging from 0.24 to 0.77 (p-values < .001).



S9.7 figure-Pairwise comparisons of normalized EBV read counts by four sequencing platforms

Figure 9.8. Transcription profiling of the EBV genome.

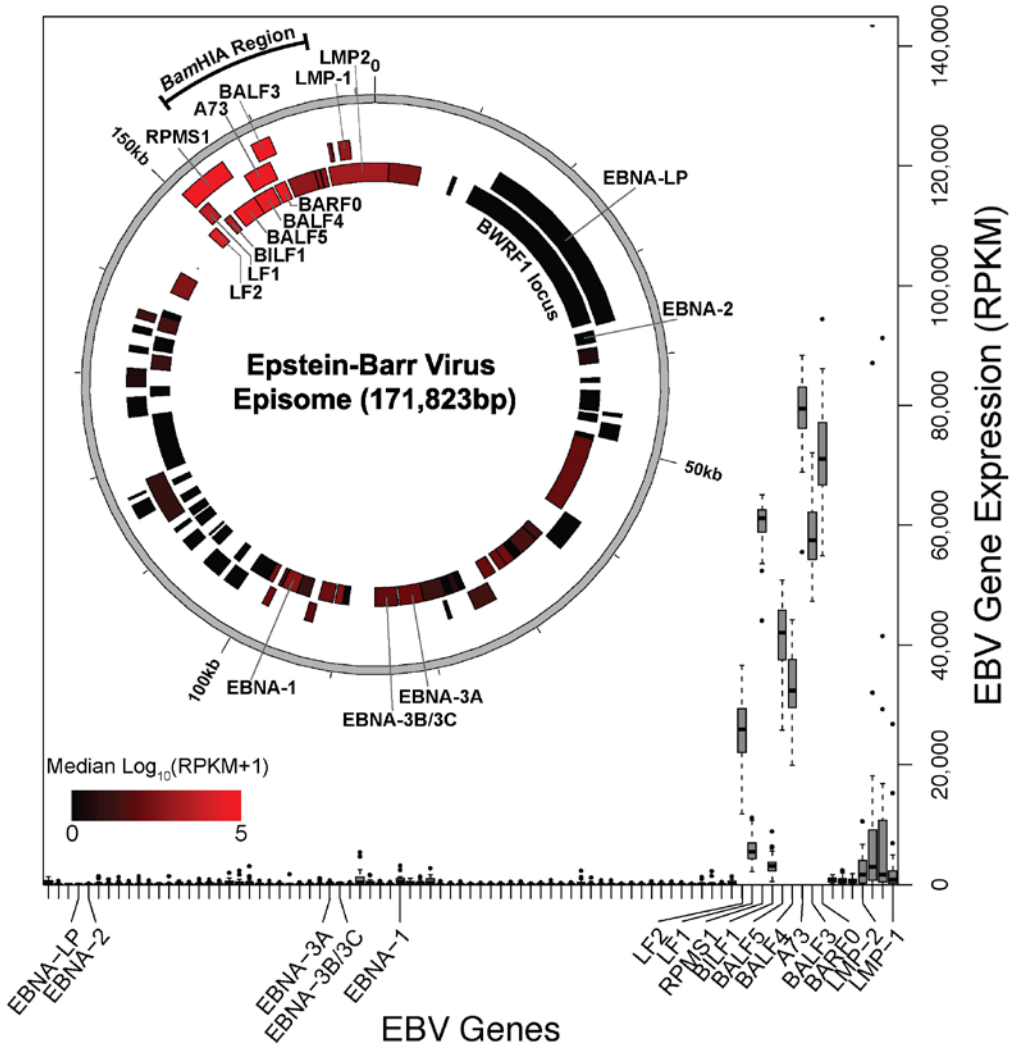


Figure S9.8 Transcription profiling of the EBV genome. Box plots displaying expression level (RPKM) across 25 samples (vertical axis) of each transcript from the EBV genome (horizontal axis). Genes are arranged from left to right according to their EBV genomic coordinate (NCBI: NC_007605.1). *Insert*: Diagram showing the locations and transcription levels [median $\log_{10}$(RPKM+1)] of the EBV genes mapped to the viral DNA episome.

S10. Clustering Analysis

S10 Section Authors:

Christina Curtis
Vésteinn Thorsson
Ronglai Shen

Subsections:

Supplement S10.1 Molecular Subtpype Definitions obtained through Integrative Clustering- Overview and Flowchart

The overarching goal of integrative clustering was to utilize the multi-dimensional molecular/genomic data from our tumour cohort to identify robust classes of GC. Following identification of these robust classes using multi-dimensional data, we then sought to develop a simpler classification model to enable assigning GC tumours into molecular subgroups (Manuscript Figure 1b). The integrative clustering methods (outlined in the flowchart below) identify groups of tumours that show similar features as assayed by the multiple platforms used in this study: somatic DNA copy number aberrations, somatic mutation, CpG methylation, mRNA, miRNA, and protein expression. Thus, the process by which we identified molecular subtypes was based on a composite analysis across distinct data types, each providing a view of the molecular features of GC. By considering measurements across different platforms, we obtained robust groupings beyond what could be achieved by analyzing a single platform. Two integrative clustering methods were used, and they are complementary in their approach. The first method (described in subsection S10.2) is similar to a previously described technique1 and begins with cluster analysis of each of the molecular platforms individually (as described in S2-S7). Integrative analysis of these diverse platform-specific cluster assignments was then performed to identify groups of tumours that shared features across multiple data platforms. The second technique, termed iCluster+ (summarized in S10.3), has also been applied to other large cancer genomics studies[3,4,5] and takes as input features from multiple platforms (without first performing clustering for each individual data platform) and performs joint clustering with simultaneous feature reduction to identify structure within the dataset. Despite these differences in approach, the resulting sample groupings were quite comparable (summarized in S10.4).

Our goal was to identify a small group of markers that could be used to assign tumours to a subgroup, in order to facilitate the classification of GC cases in the setting of clinical care, where it is not practical to obtain comprehensive molecular data. Notably, the groupings developed through our integrative analysis were found to be strongly correlated with specific features of the samples. Both methods yielded tumour groups that were predominantly EBV-positive or MSI-H, and the remaining sample groups had distinctly high or low overall degree of copy number derangement. Therefore, we used the following features: EBV, MSI and high or low aneuploidy to classify each of the 295 tumours in our dataset into one of four molecular subtypes, as described in the classification schema in Figure 1b. Analyses described in this manuscript were thus performed with these four molecular subtypes in order to generate results that could most readily be applied to future patient samples and to guide development of new clinical approaches to treat this disease. Additionally, we showed that the molecular subtypes were reflected in an analysis of principal components of the tumour samples (S10.5).

The procedure used to arrive at the molecular classification is illustrated on the next page, including references to the corresponding supplementary text.

Heterogeneous Tumour Set from 295 Patients

Comprehensive Molecular Measurements with Six Platforms
Supplements S2-S7

Platform-Specific Subtypes
Supplements S2-S7

Clustering of Platform-Specific Subtypes
Supplement S10.2

Clustering of Molecular Data with iCluster
Supplement S10.3

Integrative clustering to identify groups of similar samples

4 clusters
Supplement S10.2

3-5 clusters
Supplement S10.3

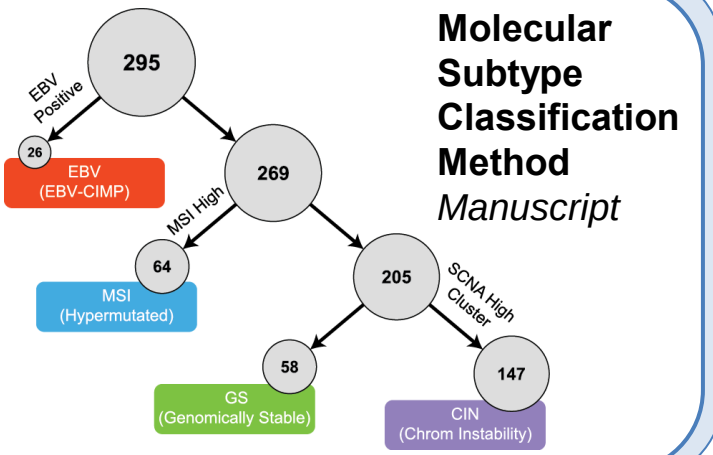
Evaluate cluster robustness and stability

Enriched for EBV positive, MSI-High, copy-number (SCNA)
Supplement S10.2

Enriched for EBV positive, MSI-High, copy-number (SCNA)
Supplement S10.3

Identify salient features of clusters

Construct decision tree using salient features



Molecular Subtype Classification Method
Manuscript

Evaluation

Cross comparison of subtypes
Supplement S10.4
PCA analysis and subtypes
Supplement S10.5

Supplement S10.2 Integrative clustering by platform-specific subtypes

As described in the preceding sections, subtype discovery was first performed through an analysis of cluster assignments in which each data platform was analyzed independently. The clustering approach for each of these platforms is described in the respective sections for each of the molecular platforms. For gene expression, four subtypes were identified (Supplement S5); for miRNA expression, five (Supplement S6); for protein expression, three (Supplement S7); for copy number, two (Supplement S2); and for DNA methylation, four, of which two can be jointly characterized as "non-CIMP" (Figure 2a, Supplement S4). In addition, mutation rates were found to fall into three categories (Supplement S3). Thus, a total of 20 platform-specific subtypes (PSSs) were identified.

We began by constructing a matrix with 20 rows, one for each PSS, and a column for each of the 295 samples. Each matrix element indicated whether a sample was a member of the PSS of that row (1) or not (0). If a PSS was not identified, the corresponding entry was set to "NA". There were 214 samples for which we had complete subtype information. Clustering was then performed on rows and columns independently. Clustering of rows identified which groups of PSSs were similar. Clustering of columns utilized information on sample similarity, in terms of their shared subtypes, in order to identify integrative tumour subtypes. Notably, no information other than PSS membership was used. Following the identification of these integrative clusters, we identified additional data elements, such as specific mutations and clinical attributes that were associated with individual clusters.

Clustering of subtype assignments (rows) To compare subtypes, we scored the 2-way contingency table for a pair of PSSs using Fisher's exact test. We then took the negative of the (base 10) logarithm of the Fisher *p*-value, and multiplied that by +1 for positive associations and -1 for negative associations. Hence, the resulting score was large and positive if the two subtypes were positively correlated, and highly negative if the subtypes were negatively correlated (in the case of continuously-valued variables), or if they tended to be mutually exclusive (for dichotomous variables). If there was no association, the score was near zero. For example, there was a strong association between Gene Expression Cluster 2 and Gastric EBV-CIMP. The fraction of Gastric EBV-CIMP cases that were also in Gene Expression Cluster 2 was 20/26 (76.9%), and the fraction of Gene expression Cluster 2 cases that were also in Gastric EBV-CIMP was 20/59 (33.9%). The null model of no association was rejected at *p*-value of 9.47×10^{-10}, corresponding to an association score of $-log_{10}(9.47 \times 10^{-10})=9.02$. A heatmap illustrating the score for this subtype pair and all other pairs is displayed in Sup. Fig. S10.6a.

Average-linkage agglomerative clustering was done using a distance defined from this association score, yielding the dendrogram shown in Supp. Fig. S10.6b. Notably, the DNA and expression-based molecular subtypes were found to group into five triplets:

A: High Copy Number, Standard Mutation Rate and non-CIMP DNA Methylation,
B: Gene Expression Cluster 1, MicroRNA Cluster 4, and RPPA Cluster 1;
C: Gene Expression Cluster 2, GASTRIC-EBV-CIMP, and MicroRNA Cluster 3.
D: Gene Expression Cluster 3, MicroRNA Cluster 2, and RPPA Cluster 3; and finally
E: GASTRIC-CIMP, Low Copy Number, and Hypermutated.

These triplets were also found using a different choice of distance, based on the Jaccard score. The remaining PSSs (MicroRNA cluster 5, MicroRNA cluster 1, Gene Expression 4, Ultramutated, RPPA Cluster 2) were all among the last clusters to be incorporated into the agglomerative clustering procedure, and were not consistently grouped by the two methods (data not shown).

Clustering of samples (columns)

To compare two samples, one can count the co-occurrence of their subtype assignments. For example, TCGA-BR-4187 and TCGA-BR-4279 were both non-CIMP, microRNA Cluster 4, Low Copy Number, and Standard Mutation Rate, but differed in their RPPA and Gene Expression clustering assignments. As such, 4 of 6 assignments matched, and this ratio could be used as the basis of a comparison score for clustering. However, when the possible subtypes were relatively few for a given data type (e.g. copy number), there was a greater chance of co-occurrence. Thus, a weighting scheme that accounts for the inherent probability of co-occurrence was desirable.

Here, we used the inverse frequencies of each of the 20 PSSs as a weight when adding co-occurrences. The score was converted to a distance prior to clustering (maximum score minus score) and the calculation was done on the 214 samples for which we had complete subtype information. We used the *ConsensusClusterPlus* R-package[6] with 1000 resamplings of 80% of tumour samples, using k-means on the distance matrix. To determine an appropriate number of clusters, we looked at the change in the Cumulative Distribution Function (CDF), and the cluster co-occurrence over variation in the number of clusters[6]. The transition to the four-cluster solution led to a substantially lesser increase in CDF difference (between adjacent k values) than did subsequent steps (Supp. Fig. S10.6c). The four-cluster solution also provided a relatively clean separation in the clustered consensus matrix (Supp. Fig. S10.6d), and had 54, 56, 74 and 30 samples in the clusters assigned numbers 1 through 4 by *ConsensusClusterPlus*.

To further evaluate the stability of the clustering of platform-specific subtype results, we performed leave-one-out validation in which one of each of the six data platforms was omitted, and the calculation was otherwise performed identically. In each case, we evaluated the recovered fraction, defined as the fraction of the integrated cluster that was found in the five-data-type calculation (four clusters in six experiments, for a total of 24 values in all). The mean recovered fraction was 85%, and the per-cluster average recovery ranged from 83% to 89%. In nearly all cases, 3 out of 4 clusters were reproduced with a recovery fraction of 73.2% or better, with the exception being the experiment excluding the mutation rate category (64.8%). The recovery rate of the remaining clusters varied from 56% to 91%.

The four-cluster integrated subtype solution is shown in Supp. Fig. S10.6e and cluster assignments are provided in Supplementary Data File S11.1a. The integrated clusters were next compared with all available data and were each found to have strong associations with specific molecular signatures: Cluster 1 with MSI-H and MLH1 methylation (45/49 MSI-H tumours were in Cluster 1, p=2.1x10^{-32}), Cluster 2 with diffuse tumours (31/56 diffuse tumours were in Cluster 2, p=3.0x10^{-7}), Cluster 3 with *TP53* mutations (25/38 samples in Cluster 3 had *TP53* mutations, p=4.9x10^{-4}), and Cluster 4 with EBV (19/20 EBV positive samples were in Cluster 4, p=1.5x10^{-18}). These and other associated variables are displayed above the clustered matrix in Supp. Fig. S10.6e. The strength of the integrated cluster associations with EBV and MSI-H supports their use in defining molecular subtypes in this manuscript (Figure 1b).

S10.3 Integrative clustering using iCluster

As a second means to identifying subgroups of GC, we utilized iCluster, which formulates the problem of subgroup discovery as a joint multivariate regression of multiple data types with reference to a set of common latent variables that represent the underlying tumour subtypes[2, 7, 8]. Unlike the first integrative approach (from S10.2), platform-specific clustering was not performed. A penalized likelihood approach was used for estimation, and a Monte Carlo Newton–Raphson algorithm was employed to maximize the penalized log-likelihood. Due to the computational intensity of the parameter-tuning procedure, the current implementation of iCluster+ takes as input up to four data types.

Data processing

Data processing methods were similar to those previously described[2,3] and are outlined below. For somatic mutation data, the mutation MAF file was used. A gene-by-sample matrix of binary values (1-mutated, 0-wildtype) was generated for clustering. The top 1000 mutated genes ranked by the Mutsig analysis were included for clustering. Segmented somatic copy number profiles (after removal of CNVs) were obtained, and dimension reduction was performed to obtain non-redundant copy number regions, as

previously described[3,4]. For the methylation data, the median absolute deviation was employed to select the top 4,000 most variable CpG sites based on the β-value for input to the clustering framework. For mRNA and miRNA sequence data, lowly expressed genes were excluded based on median-normalized counts, and variance filtering led to 361 mRNAs and 145 miRNAs for clustering. mRNA and miRNA expression features were combined as a single data type, representing transcriptomic measurements. For the RPPA (proteomic) data, 121 antibodies were employed in downstream analyses. Given that iCluster+ can accept four data types, and five were available, we built two models (A and B), including either the transcriptome (mRNA+miRNA) data or the RPPA data. Supp. Fig. S10.7 shows that the results were highly comparable for model A (transcriptome) versus model B (proteome), indicating the robustness of this approach and lack of sensitivity to a particular data type.

Model selection

To determine the optimal combination of the penalty parameter values, a large search space was required. We employed an efficient sampling method that utilized uniform design (UD)[9], such that for a given K, we determined the penalty parameter vector that minimized a Bayesian information criterion. A theoretical advantage of the uniform design over an exhaustive grid search is the uniform space-filling property that avoids wasteful computation at close-by points.

The number of clusters (K) was estimated. We computed a deviance ratio metric, where K was chosen to maximize the deviance ratio. As shown in Supp. Fig. S10.7A, for model A an "elbow" point was noted at K=3, beyond which point the increase in the deviance ratio diminished, increasing again at K=5. For model B (Supp. Fig. S10.7B), an elbow point was similarly noted at K=3. While these results suggested that K=3 represented an optimal solution, it was of interest to compare this with the alternate K=5 cluster solution, which provided greater granularity. Cluster assignments are provided in Supplementary Data File S11.1. A heatmap representation of the associated subtype assignments for all data platforms and the association with other sample attributes (top panel) is shown in Supp. Fig. S10.8 for K=3 and K=5. Both the 3- and 5-cluster solutions showed an association with the Lauren classification and were enriched for diffuse histologic type tumours (iClust1). For the K=3 and K=5 solutions, iClust3 and iClust5, respectively, were enriched for Gastric-CIMP, MLH1 methylated, and MSI-H samples. In contrast, only the K= 5 solution revealed the clear separation of an EBV-positive group (iClust2), consistent with other molecular features of these data. Hence, the K=5 solution was used in comparisons with the alternative clustering strategies. Importantly, these distinct integrative clustering approaches both robustly partitioned gastric tumour samples on the basis of EBV-positivity, MSI status, and SCNAs, supporting their use in defining the four molecular subtypes presented throughout the text (Supp. Fig. 1b).

Supplement S10.4. Cross-comparison of subtypes

A comparison of cluster assignments based on the 4-cluster integrative clustering by platform-specific subtypes with the K=3 and K=5 iCluster results is shown in Supp. Fig. S10.9, as is the cross-tabulation of each of these two approaches with the 4 molecular subtypes defined in Supp. Fig. S10.6e. From tables A and C, we see that the EBV molecular subtype has almost complete overlap with the integrative cluster from the both procedures (Clusters C1, and iClust2, respectively). Similarly, the MSI molecular subtype has a very strong overlap with integrative clusters. CIN is found most commonly in cluster C3 from the platform-specific subtypes and iCluster 3. Table B shows that overall there is good overlap between sample groupings obtained from the two integrative clustering methods. A primary difference between the 5-cluster solution from the iCluster analysis and the 4-cluster solution from the integrative clustering based upon platform-level cluster assignments is that the aneuploidy/CIN group of tumours is split into two subgroups using the iCluster approach.

Supplement S10.5. Subtypes in the context of Principal Component Analysis of tumour samples

In order to investigate the validity of the molecular subtypes without reference to integrative clustering or to the results thereof, we examined the correspondence between the subtypes and the first few principal components of the tumour sample set. The principal components are the main directions of variance in the multi-dimensional space in which the samples reside. The dimensionality of the space is equal to the length of the data vector of any given sample. To simplify the calculation, we use the vector of 20 values described in S10.2. In Figure S10.10, projections into the first few principal components are shown for the data, and each tumour data point is coloured according to subtype. The first two principal

components, PC1, and PC2, have strong contributions from mutation rate and copy number, and this is reflected in the separation of CIN and MSI in this subspace (panel A). Looking into dimension three (panel B: PC3 vs PC2), we see the emergence of EBV samples from other samples. This implies that the main directions of variance in the data have good correspondence with the molecular subtypes.

References

1. The Cancer Genome Atlas Research Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature.* 490:61-70.
2. Shen, R., A.B. Olshen, and M. Ladanyi. (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 25(22): p. 2906-12.
3. The Cancer Genome Atlas Research Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. Nature. 489:519-25.
4. The Cancer Genome Atlas Research Network. (2013) Integrated genomic characterization of endometrial carcinoma. *Nature.* 497:67-73.
5. Curtis, C., Shah, S., Rueda, OM, et al. Integrated Genomic and Transcriptomic Architecture of 2,000 breast tumors reveals novel subgroups. (2012) *Nature*. 486:346-52.
6. Wilkerson MD, Hayes DN: ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. (2010) *Bioinformatics* 26:1572-3.
7. Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, Powers RS, Ladanyi M, and Shen R**.** (2013) Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*. 12:4245-4250.
8. Shen R, Wang S, Mo Q. (2012) Sparse integrative clustering of multiple omics data sets. *Annals of Applied Statistics*. 7:269-294.
9. Fang, K. and Y. Wang (1994). *Number-theoretic methods in statistics*. 1st ed. Monographs on statistics and applied probability. London; New York: Chapman & Hall. xii, 340.

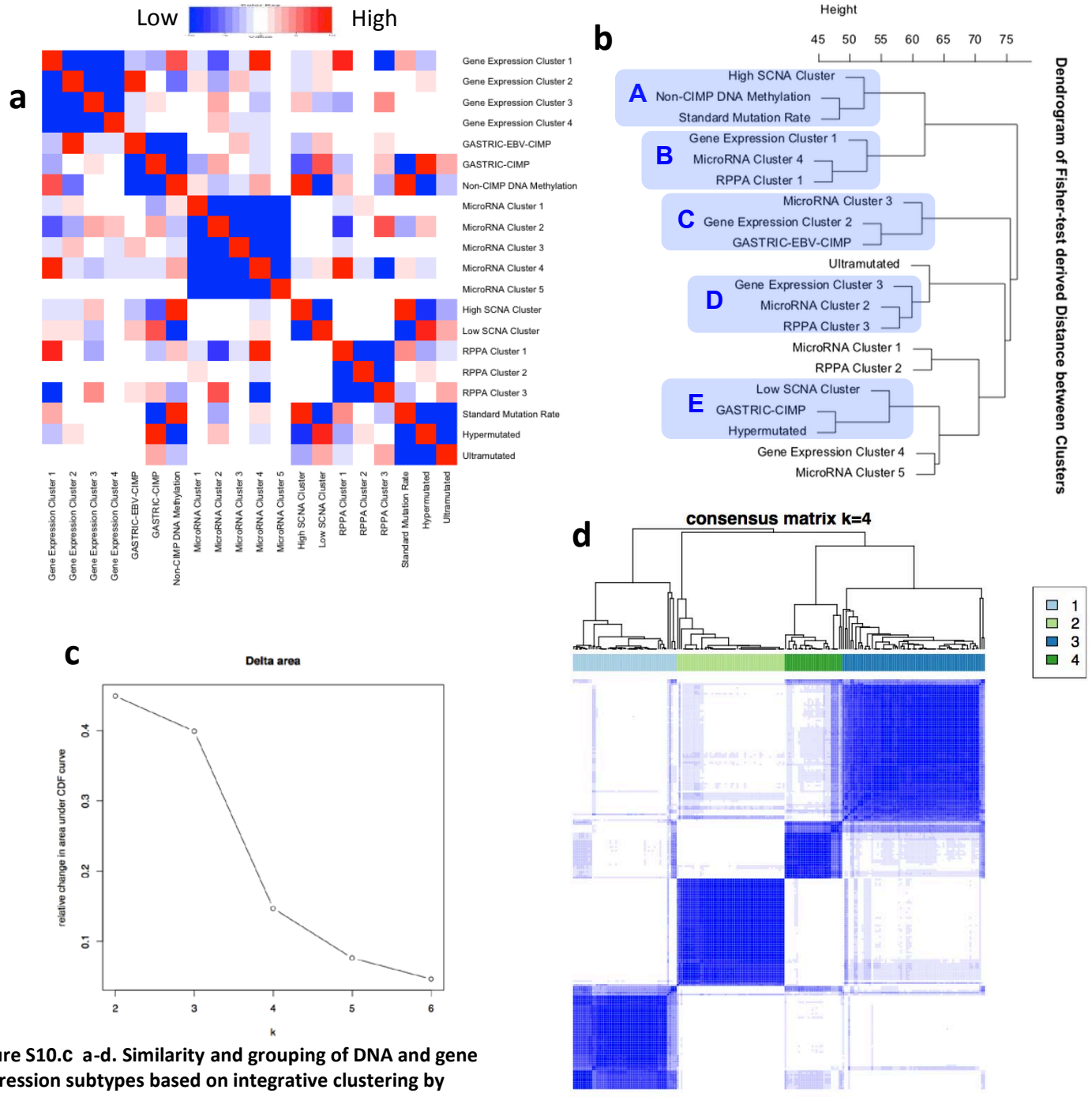
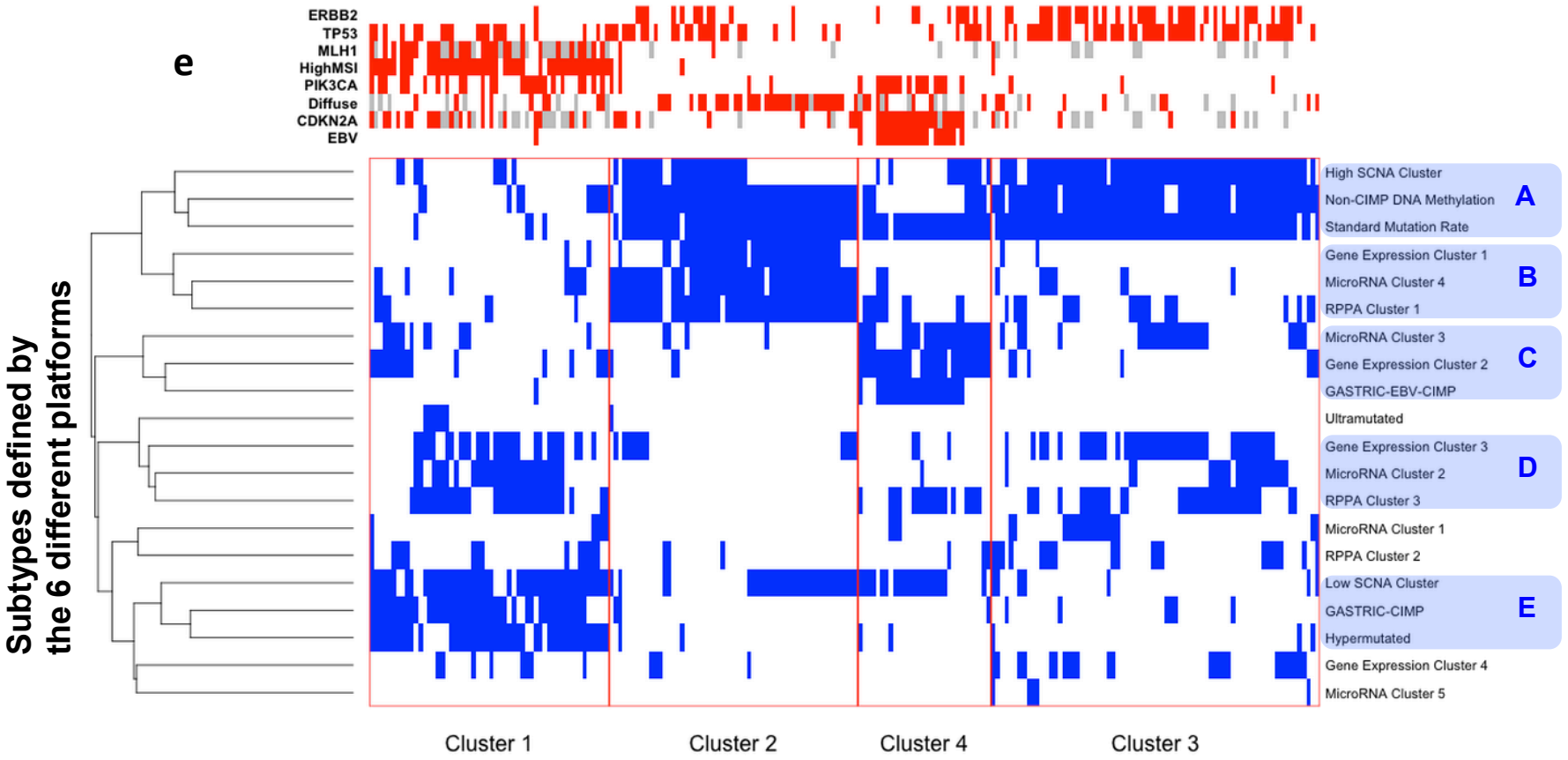


Figure S10.c a-d. Similarity and grouping of DNA and gene expression subtypes based on integrative clustering by platform-specific subtypes

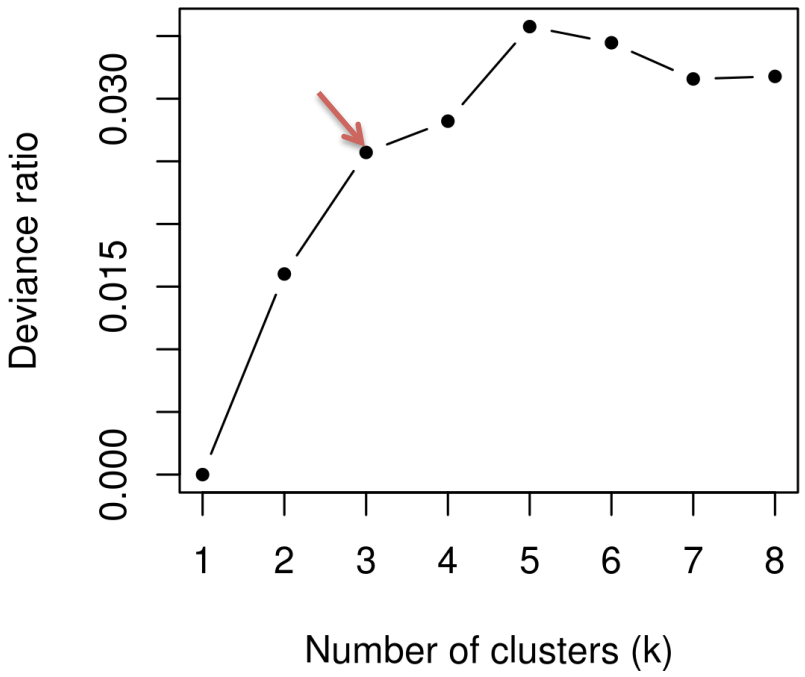
Figure S10.6e Platform-Specific Subtype Membership and the Four-Cluster Consensus



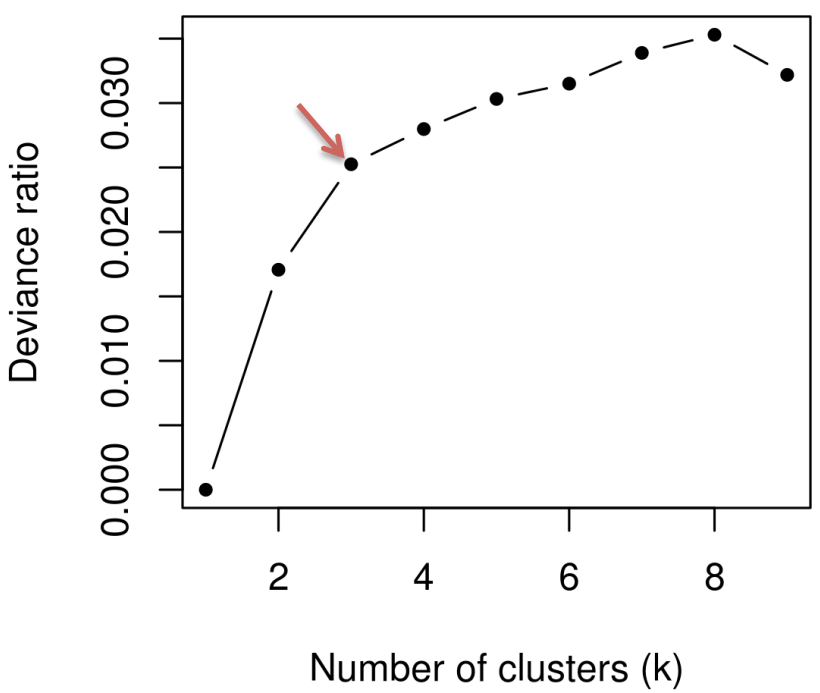
Gastric-CIMP	Diffuse	EBV	Aneuploid (ERBB2 amp)
Hypermut	Expression-1	EBV-CIMP	*TP53* mutations
MSI-H	MicroRNA-4	Expression-2	Non-CIMP
PIK3CA mut.	RPPA-1	Low SCNAs	
Low SCNAs			

Supplementary Figure S10.6. Similarity and grouping of DNA and gene expression subtypes based on integrative clustering by platform-specific subtypes. a) Similarity among platform-specific subtypes. Each of the 20 subtypes, represented by rows and columns, was obtained from a single molecular platform. The heatmap shows a similarity score based on the Fisher p-value for the contingency table comparing the two platform-specific subtypes in the corresponding row and column. A red cell corresponds to two subtypes that are highly similar: they tend to have many tumour samples in common. A blue cell corresponds to subtypes that are highly dissimilar. b) Subtype similarity dendrogram. The distance metric was based on the score shown in panel a. The highlighted subtype triplets were consistently grouped when the comparison metric was varied. Note that these are groupings of similar platform-specific subtype classifications, not groupings of tumour samples. c) Relative change in area under CDF. ConsensusClusterPlus plot comparison of samples using weighted co-occurrence. d) Consensus matrix for the four-cluster solution from ConsensusClusterPlus. e) Integrative subtype assignments. The red box defines a subtype membership matrix, in which rows correspond to platform-specific subtypes, and columns represent 214 tumour samples. Blue indicates subtype membership. Rows are arranged as in panel b, and columns are arranged according to the similarity of sample pairs based on platform-specific molecular subtype membership. Integrated subtypes are numbered 1 through 4. Above the membership matrix are dichotomous attributes of samples, each shown in red: non-silent mutations in PIK3CA; non-silent mutations in TP53; MSI-H; diffuse, as opposed to intestinal histologic types; and EBV-positivity. Unassigned values are indicated in gray. Below each integrative cluster, data elements enriched in that subgroup are indicated.

A



B



C

	Model B		
Model A	1	2	3
1	3	88	0
2	55	0	0
3	1	0	53

D

	Model B				
Model A	1	2	3	4	5
1	0	0	0	0	59
2	0	0	0	26	0
3	0	0	36	0	0
4	2	33	6	0	1
5	42	0	2	0	3

Figure S10.7- robustness of iCluster results to different data inputs and model selection

Supplementary Figure S10.7: **Robustness of iCluster results to different data inputs and model selection.** Cluster membership results were highly comparable, regardless of whether mRNA+miRNA (transcriptome) or RPPA (proteome) features were included with somatic mutation, copy number, and CpG methylation data. a) The deviance ratio is plotted versus the number of clusters for Model A, which includes transcriptome data (mRNA + miRNA), where an "elbow" point at K=3 is noted, beyond which the increase in the deviance ratio diminished. b) As in A, but for Model B, which included RPPA data rather than transcriptome data. c) The cluster membership assignments were highly concordant regardless of whether transcriptome (Model A) or proteome data (Model B) were included for the K = 3 solutions. d) The cluster membership assignments for Models A and B were also highly concordant for the K = 5 solutions.

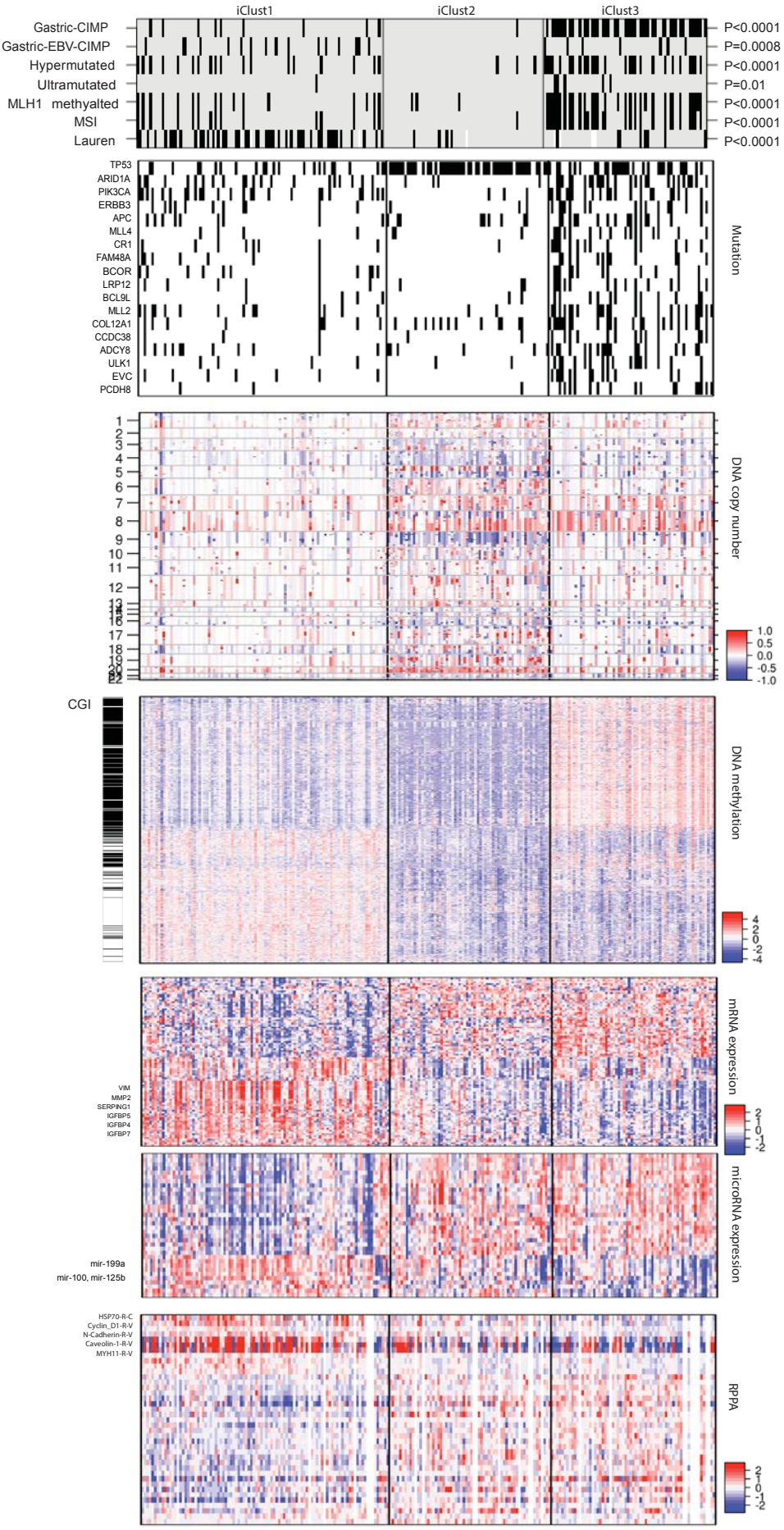


Figure S10.8: Heatmap representation of iCluster subtype assignments for K=3 solution

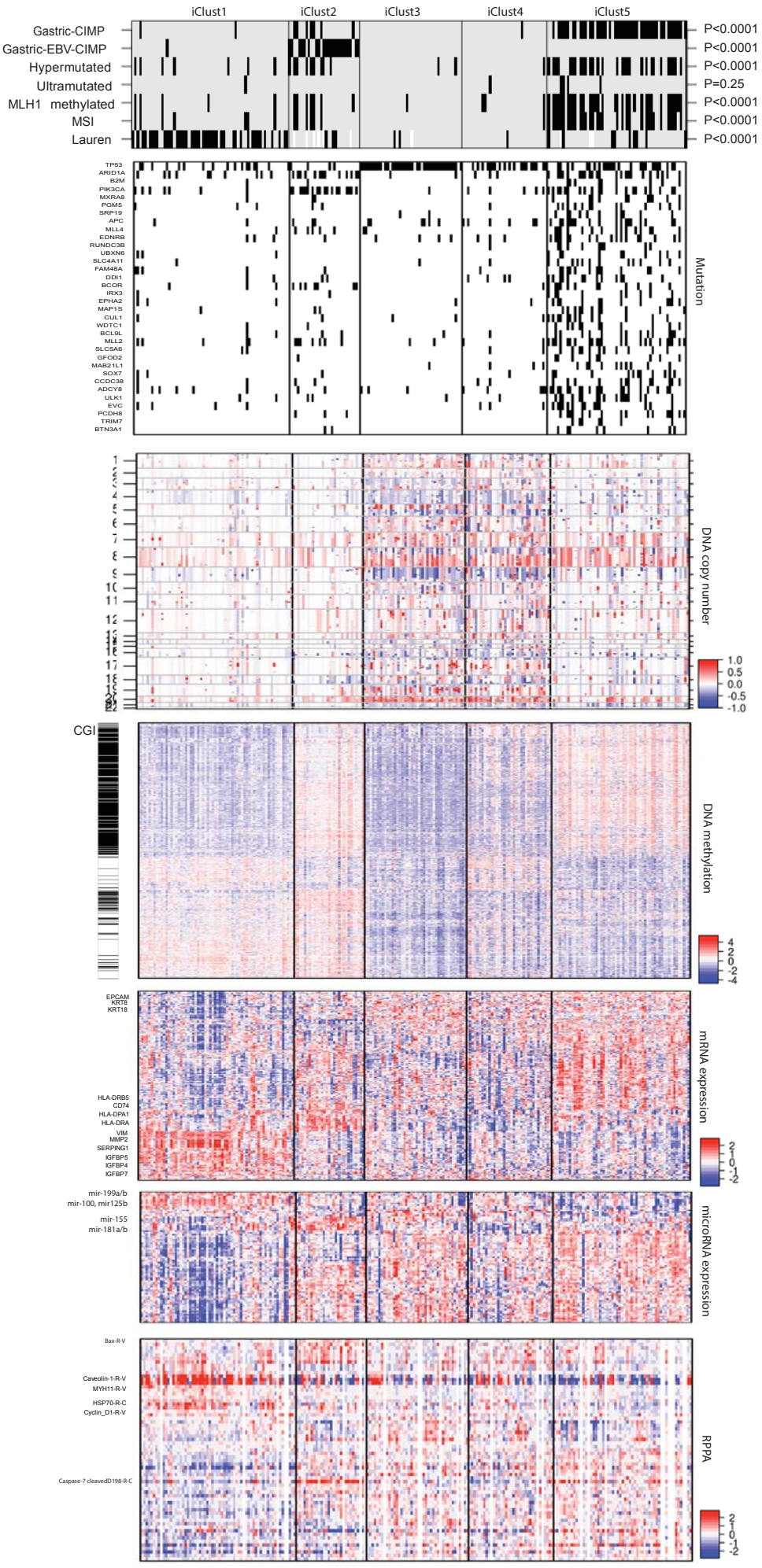


Figure S10.8: Heatmap representation of iCluster subtype assignments for K=5 solution

Supplementary Figure S10.8: **Heatmap representation of iCluster subtype assignments.** Heatmap illustrating the iCluster subtypes for all data platforms based on clustering of mutation, copy number, methylation, and transcriptome (mRNA + mRNA) features (note that RPPA data were not included here, but yield comparable results when exchanged for transcriptome data). The association between cluster membership and other sample attributes is illustrated in the top panel. a) Results for the K = 3 solution and b) for the K = 5 solution.

A

	EBV	MSI	CIN	GS
C1	1	45	4	4
C2	0	3	26	27
C3	0	1	67	6
C4	19	0	6	5

B

	iClust1	iClust2	iClust3	iClust4	iClust5
C1	4	7	1	2	34
C2	35	0	9	6	3
C3	13	0	21	21	8
C4	5	19	4	0	1

C

	iClust1	iClust2	iClust3
C1	13	2	33
C2	40	10	3
C3	15	37	11
C4	19	4	6

D

	iClust1	iClust2	iClust3	iClust4	iClust5
EBV	1	20	0	0	0
MSI	7	6	0	1	33
CIN	17	2	42	30	20
GS	39	1	0	4	4

E

	iClust1	iClust2	iClust3
EBV	15	0	6
MSI	16	1	30
CIN	27	61	23
GS	40	2	6

Figure S10.9 Comparison of cluster membership using different integrative clustering approaches

Supplementary Figure S10.9. Comparison of cluster membership using different integrative clustering approaches. Cross-tabulation of cluster membership assignments comparing the integrative clustering approaches (iCluster K=5, iCluster K=3, and the four-cluster solution from clustering of platform-specific subtypes) with each other and with the molecular subtypes defined in the manuscript. a) Integrative clustering by platform-specific subtypes four-cluster solution (Row, C1 through C4) compared with molecular subtype (Column). b) Integrative clustering by platform-specific subtypes four cluster solution (Row) compared with iCluster K=3 sample membership (Column) c) Integrative clustering by platform-specific subtypes four cluster solution (Row) compared with iCluster K=5 sample membership (Column). d) Molecular subtype (Row) compared with iCluster K= 5 sample membership (Column). 3) Molecular subtype (Row) compared with iCluster K= 3 sample membership (Column).

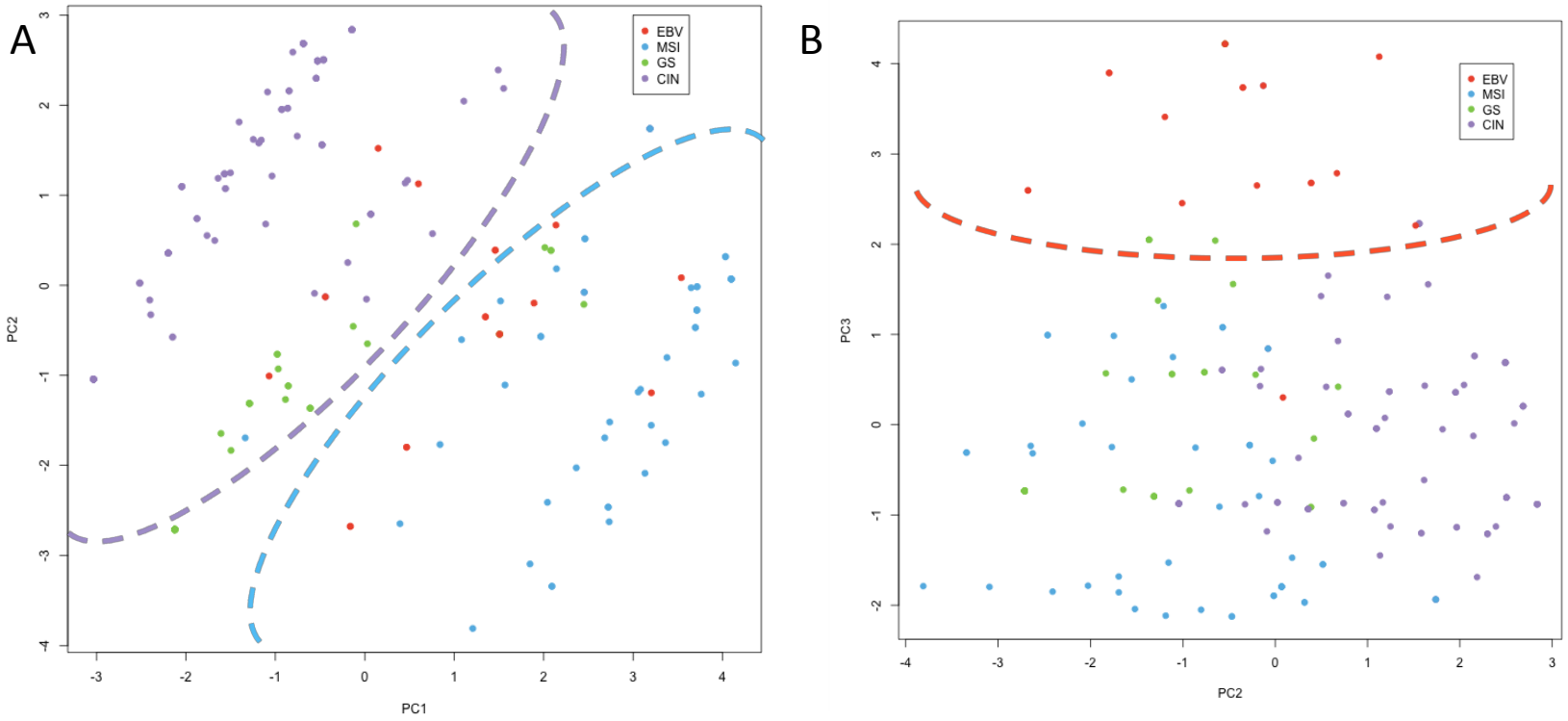


Figure S10.мл – Principal Components and Molecular Subtypes

Supplementary Figure S10.10: **Principal Components and Molecular Subtypes.** Projection of samples into the subspace of the first two (panel A) and the second two (panel B) principal components is shown. The points represent samples and are colored by molecular subtype. In panel A, dotted lines have been inserted to denote regions that are predominantly CIN (upper left) and MSI (lower right). In panel B, the dotted line encloses a region that is almost exclusively EBV.

S11. Data Integration, Pathway Analysis, and Resources for Data Exploration

S11 Section Authors:

Vésteinn Thorsson
Sheila Reynolds
Christina Curtis
Sam Ng
Hsin-Ta Wu
Max Leiserson
Benjamin Raphael
Richard Kreisberg
Brady Bernard
Hector Rovira
Spring Liu
Michael Noble
Da Yang
Wei Zhang
Nils Weinhold
Nikolaus Schultz

Subsections:

S11.9 text- Firehose Analysis
S11.10 text- MIRACLE analysis
S11.11 figure- MIRACLE miR-RNA regulatory network for epigenetic silencing
S11.12 figure- MIRACLE DNA methylation and expression of miR-9
S11.13 figure- MIRACLE DNA methylation and expression of miR-196b
S11.14 figure- Somatic mutations recurrently altered in receptor tyrosine kinases
S11.15 figure- Oncoprint of cell cycle genes

Supplement S11.1. Master Patient Table and Feature Matrix

STAD Feature Matrix. To facilitate the identification of associations among the diverse clinical and molecular data in this study, a "feature matrix" (FM) was constructed by integrating values from all data types. Each row in the FM represents one of the 295 tumor samples or 91 matched- normal samples, and the columns contain all available clinical, sample, and molecular data for each sample: mRNA and microRNA expression levels, protein levels, copy number alterations, DNA methylation levels, and somatic mutations. Each column in the FM represents a single clinical, sample, or molecular data element, and the individual data values may be numerical (continuous or discrete) or categorical, as appropriate. Missing values are indicated within the FM by "NA", and the number of non-NA data values varies significantly across the different data types (columns). Overall, approximately 77% of the matrix elements are non-NA (94% for the tumor samples). Data were retrieved from the DCC on Jan 20, 2014 and further processed into columns as follows.

Clinical and sample data (567 features): DCC clinical and sample data were processed into a matrix. Assignments from EPC review (Supplement S1) were used for anatomic site, histology, and TNM AJCC Stage. Additionally, columns were included for ABSOLUTE calls for tumor purity (Supplement S2), estimated leukocyte percentage (Supplement S4), and MLH1 and CDKN2A epigenetic silencing (Supplement S4). Cluster assignments obtained via clustering of cluster assignments (Supplement S10) and iCluster (Supplement S10) were added, as were results of unsupervised clustering for each of the individual molecular data types: SCNA (Supplement S2), RNAseq (Supplement S5), miRNA-seq (Supplement S6), DNA methylation (Supplement S4), and RPPA (Supplement S7). Mutation rates and rate categories (Supplement S3) were included, as were fusion events such as ARHGAP-CLDN18 (Supplement S3). In addition, variables were generated that contrast pairs of subgroups in a non-dichotomous classification.

Molecular Data. Gene expression (22,277 features): Gene level RPKM values from RNA-seq (Supplement S5) were log2 transformed, and filtered to remove low-variability genes (bottom 25% removed, based on interdecile range). *MicroRNA expression* (697 features): The summed and normalized microRNA quantification files (Supplement S6) were log2 transformed, and filtered to remove low-variability microRNAs. (An initial filter removed any microRNA not observed in at least 6 samples, and a second filter removed the bottom 25% by interdecile range.) *Somatic copy number alterations*: Copy number and focal copy number changes were obtained for peaks identified by GISTIC as described above (Supplement S2, 188 features). *DNA methylation* (19,711 features): Probe-specific Level 3 β-values were obtained as described above (Supplement S4). We started with 26,258 probes in common between the two methylation platforms, and then removed the bottom 25% based on interdecile range. RPPA (189 features) (Supplement S7). *Somatic mutations* (8,220): The Mutations Annotation Format file (Supplement S3), was used to generate a binary indicator vector indicating whether a particular nonsilent mutation is present in a specific sample. Mutation features found in fewer than five tumor samples were removed.

The Synapse platform[1], by Sage Bionetworks (www.sagebase.org), was used during the development of this project for distributing versioned data to project researchers and as a staging area for assembling files into the Feature Matrix.

STAD Master Patient Table. Key variables, including those discussed in the manuscript, were extracted from the FM to create **Supplementary Data File 11.1a: Master Patient Table**. This file can be found on the TCGA Stomach Adenocarcinoma publication page at
https://tcga-data.nci.nih.gov/docs/publications/stad_2014/

[1]*Enabling transparent and collaborative computational analysis of 12 tumor types within The Cancer Genome Atlas*; Larsson Omberg, Kyle Ellrott, Yuan Yuan, Cyriac Kandoth, Chris Wong, Michael R Kellen, Stephen H Friend, Josh Stuart, Han Liang & Adam A Margolin; Nature Genetics **45**, 1121–1126 (2013).

S11.2 NCI-PID Pathway Expression Associated with Molecular Subtypes.

In order to gain insight into the underlying differences between the four main molecular subtypes identified, we performed pathway-level analysis of the mRNA expression differences across different stratifications of the tumour and adjacent non-tumour tissue samples in this dataset. In this analysis, pathways were defined as lists of genes. Specifically, we used gene lists describing the 224 pathways from the NCI-PID pathway database[1]. Given a stratification of the samples into two non-overlapping subgroups A and B (which may represent, for example, EBV tumours vs. adjacent non-tumour samples, or MSI tumours vs. CIN tumours), a p-value was computed for each gene using the non-parametric Kruskal-Wallis one-way analysis of variance by ranks. This p-value estimates the statistical significance that the expression of gene X is elevated or reduced in subgroup A relative to subgroup B. For each pathway, the gene-level p-values were log-transformed and summed (using an approach based on Fisher's combined statistic[2]) to obtain a pathway-level composite score:

$$S = -\sum_{i=1}^{n} \log\ (p_i)$$

The significance of this score, p_s, was then estimated empirically by similarly scoring 10000 randomly generated pathways for each NCI-PID pathway, using the same distribution of pathway sizes and gene membership. Finally, a heatmap was created using the absolute value of $\log(p_s)$, with the sign (+ or -) indicating whether the pathway was elevated (or reduced) in subset A relative to B.

References:

1. Schaefer, Carl F; Anthony Kira, Krupa Shiva, Buchoff Jeffrey, Day Matthew, Hannay Timo, Buetow Kenneth H (Jan 2009). "PID: the Pathway Interaction Database". Nucleic Acids Res. 37 (Database issue): D674–9. doi:10.1093/nar/gkn653. PMC 2686461. PMID 18832364

2. Fisher, R.A. Questions and answers #14. The American Statistician 2, 30-31 (1948).

figure 11.2a- Heatmap of relative pathway expression levels for all contrasts among molecular subtypes and normals

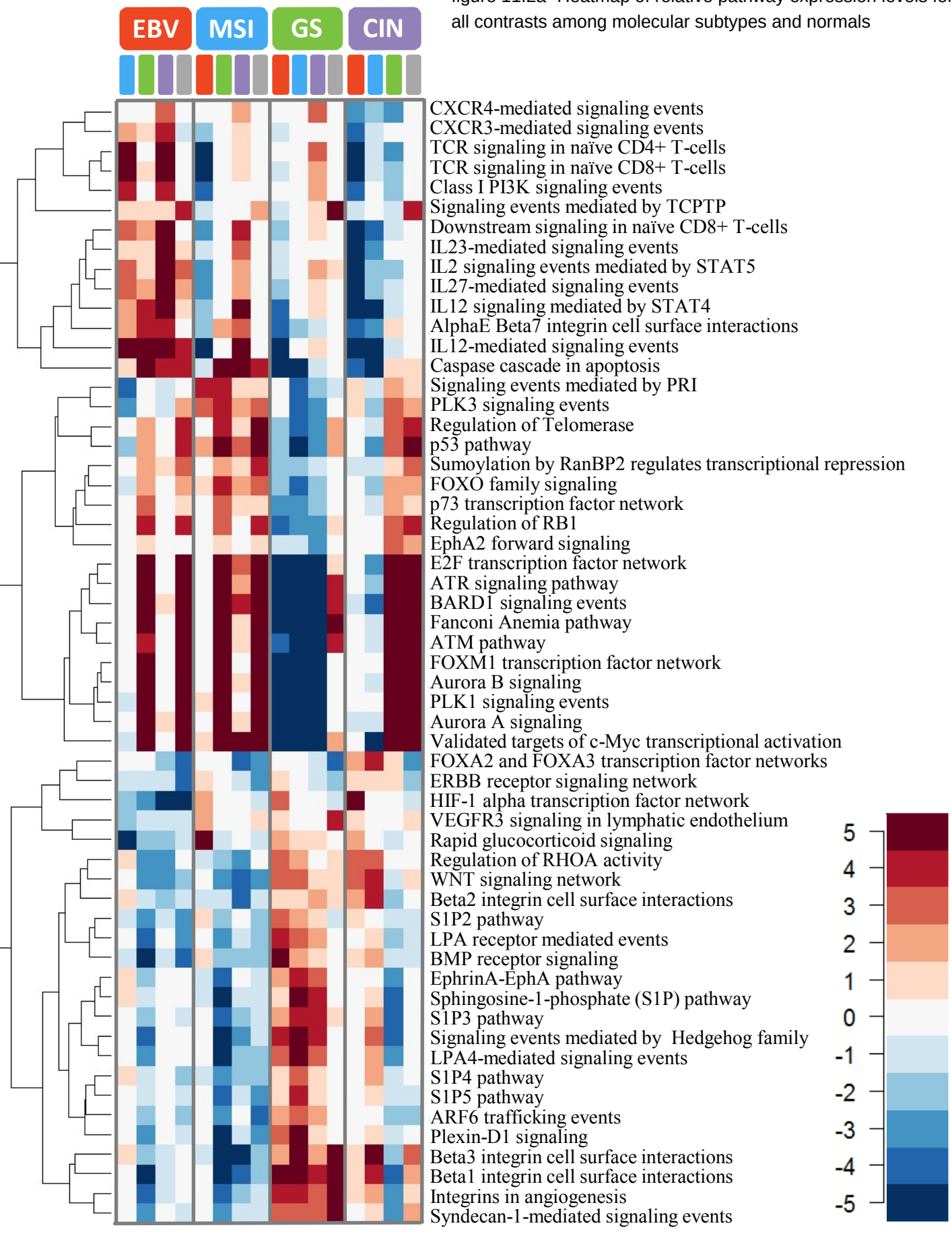


Figure S11.2a: Each molecular subtype was compared with every other molecular subtype as well as with the adjacent non-tumour samples. This pairwise comparison resulted in 16 columns, with each column indicating which pathways were elevated (or reduced) when comparing the two subsets indicated by the colors at the top of the heatmap. For example, the first column (on the left) is a comparison between the EBV and MSI subsets, in which we see that the genes involved in the TCR signaling pathways (rows 3 and 4) were expressed at much higher levels in the EBV samples than in the MSI samples. The reverse is seen in column 5, in which the comparison is reversed: MSI relative to EBV. Within each group of four columns, the rightmost column is the comparison to the non-tumour samples (indicated by a grey box at the top).

S11.3 Characterization of *RHOA* mutations and *CLDN18-ARHGAP* fusions led to a predicted activation of the RHOA- ROCK signaling pathway.

Frequent mutations in the *RHOA* gene and also gene fusions involving ARHGAP26 and ARHGAP6 were identified in the genome-stable (GS) molecular subgroup. To evaluate whether these specific events may be activating or inactivating RHO signaling, we employed the PARADIGM-SHIFT algorithm, which assigns a pathway impact score for each event[1]. Pathway signatures for the ARHGAP26 and ARHGAP6 fusion events were analyzed jointly with the RHOA mutations to provide clues about the impact of these mutations and fusion events.

Within the GS group, 50 samples had available copy number and expression data required to run PARADIGM-SHIFT analysis, with 6 *RHOA* mutations and 8 *CLDN18-ARHGAP* fusions occurring in this set. PARADIGM parameters were trained on the complete cohort of 258 samples with available copy number and expression data. The RHOA-ROCK signaling pathway was constructed from MetaCore™, and RHOA mutation neighborhoods (the network of interactions surrounding RHOA) were selected in a supervised fashion by selecting features based on Fisher score. PARADIGM-SHIFT (P-Shift) scores for RHOA, which reflect the discrepancy in upstream versus downstream pathway signals, were calculated as the difference in inferred activity between the two runs of PARADIGM: one in which only the connections with the upstream regulators were retained (R-run) and one in which only the connections with the downstream targets were retained (T-run). The accuracy of the model was then assessed by using the absolute P-Shift score as a classifier to predict the presence of an alteration (*RHOA* mutation or *ARHGAP* fusion). The model was able to predict alteration status with an average area under the curve of 0.62 across 5-fold cross-validation, suggesting that the PARADIGM-SHIFT model was able to distinguish samples altered in this pathway.

When the distribution of P-Shift scores for samples with alterations in either *RHOA* or *ARHGAP* were compared to samples without either of these alterations, an enrichment of positive P-Shift scores was identified, indicating gain-of-function (GOF) on average through the RHOA signaling pathway (Supplemental Figure S11.4a). The significance of this aggregated GOF score was determined by running a background model in which the selected network topology was fixed, but the data were permuted, thus assigning random genes to the surrounding network neighborhood of the RHOA protein. Using this background model, the GOF-aggregated score was found to have a p-value of 0.047 (Supplemental Figure S11.4b). Altogether, these findings suggest that the signaling consequences of RHOA mutations or ARHGAP fusions lead to GOF, based on the discrepancy of up- vs. down- stream activity signals.

PARADIGM-SHIFT was run on the complete cohort to determine the functional impact of alterations on the network, and its network was viewed with a CircleMap[2] display (Supplemental Figure S11.4c). As expected from prior knowledge, RHOA activation was found to be mediated through the transcription factor STAT3. The pattern of expression for downstream targets IRF1 and IL1B mirrored the profile of P-Shift score concordant with RHOA pathway activation in the samples with alterations. Interestingly, downstream targets IFNG and PLA2G4A appeared to be active in the case of either RHOA mutation or ARHGAP fusion. This concordance suggests that different alterations in the RHOA pathway may not be equivalent, leading to slightly different phenotypes. Additionally, the presence of samples with high P-Shift scores in the non-altered set suggests that there may be additional events within the GS subgroup that lead to RHOA signaling activation.

1. Ng, S. *et al.* PARADIGM-SHIFT predicts the function of mutations in multiple cancers using pathway impact analysis. *Bioinformatics* **28**, i640-i646 (2012).
2. Wong, C.K. *et al.* The UCSC Interaction Browser: multidimensional data views in pathway context. *Nucleic Acids Res* **41**, W218-24 (2013).

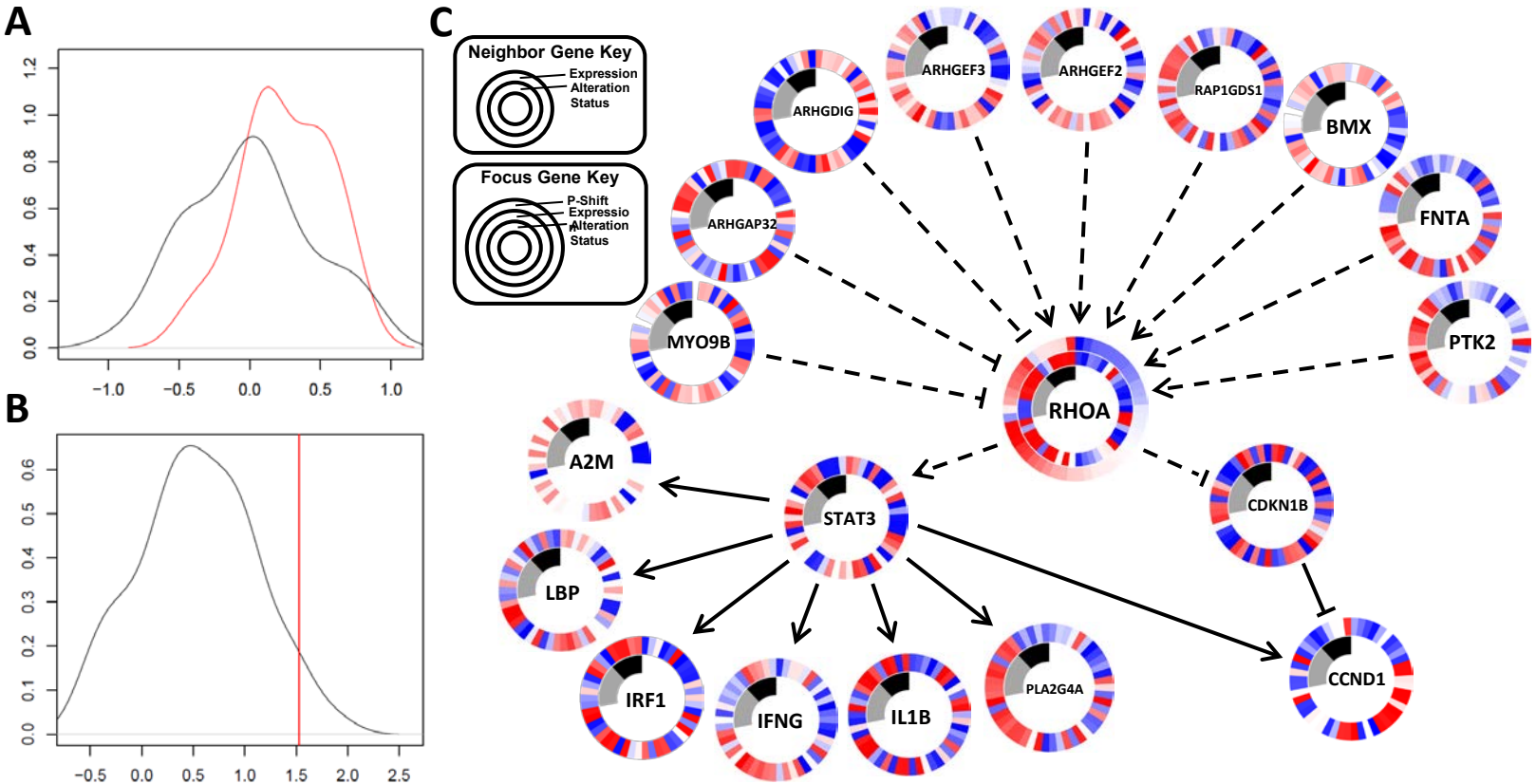


Figure S11.4. PARADIGM-SHIFT analysis of RhoA pathway alterations, RHOA mutations and ARHGAP fusions. (A) Comparison of the distribution of P-Shifts for samples with alterations (red) and samples without (black). (B) Distribution of t-statistics of the difference in P-Shift scores between samples with alterations versus samples without. Red line shows t-statistic based on actual data. (C) Circlemap display of mutation neighborhood selected for RHOA. Solid lines indicate transcriptional regulation and dashed lines indicate protein regulation. Samples were sorted first by the RHOA alteration status (Black: RHOA mutation, Grey: ARHGAP fusion), then by P-Shift score.

S.11.5. HotNet Analysis

We used HotNet2[1] to identify subnetworks of a protein-protein interaction network that contained genes with significant number of mutations. HotNet2 identifies significantly mutated subnetworks according to both scores of individual genes and the topology of interactions between genes. HotNet2 uses a localized heat diffusion process to model mutations on the topology of interactions, and a partitioning procedure that distinguishes the directionality of heat flow on the network. Hotnet2 improves our previous HotNet algorithm[2] that was used in earlier TCGA publications. For this analysis, the score, or heat, assigned to each gene (node in the network) is the fraction of samples that have a mutation in the gene, using mutation data described below.

HotNet2 uses a permutation test similar to the one employed by HotNet, in which the gene scores are permuted among the measured genes. A two-stage statistical test was used to assess the significance of the list of subnetworks obtained by HotNet. In the first stage, a *p*-value for the number of subnetworks in the list was computed. In the second stage, we derived an estimate of the false discovery rate (FDR) of the list of subnetworks. Finally, we tested whether the identified subnetworks were significantly enriched for genes in known pathways and protein complexes.

Mutation Data and Interaction Network for HotNet Analysis

We analyzed the non-silent mutations (single nucleotide variants and small indels) from the MAF file in 289 STAD samples. We also included focal driver copy number aberrations from GISTIC2 output via Firehose (Supplement S11.9), fusion genes (Table S5.6), rearrangements (Table S3.7) and splicing events (Figure S5.5). We removed 74 samples identified as hypermutators in Supplementary Table S11.1a. We removed genes with non-silent mutations in < 2% of samples and also genes with mutations in > 3% of samples that with MutSigCV q>0.25 (Table S3.5). After this processing, the dataset included 217 samples and 879 genes with mutations.

We also ran HotNet2 separately on each of the four subtypes: GS, CIN, EBV and MSI. For each subtype, we removed genes with non-silent mutations in < 2% of samples of subtypes CIN and MSI due to their high mutation rate, and in > 8% of samples that with MutSigCV q>0.25 in each subtype. Following this filtering, GS contained 55 samples and 4290 mutated genes; CIN contained 140 samples and 2674 mutated genes; EBV contained 22 samples and 2193 mutated genes; MSI contained 64 samples and 4868 mutated genes.

We built a protein-protein interaction network by combining high-quality protein-protein interactions from HINT[3] with recent high-throughput interactions from the HI-2012[4] set of protein-protein interactions. The network contained a total of 9859 proteins and 40705 interactions.

HotNet2 Analysis

HotNet2 identified 6 subnetworks containing at least 3 genes (p < 0.001) with a corresponding FDR = 0.439. Two subnetworks contained core genes in the p53 signalling pathway and ErbB pathway (Table S11.6, Figure S11.7a). Another subnetwork contained *RHOA*, a master regulator of actin organization, focal adhesion and cell motility and upregulated in various cancer types. Moreover, *RHPN2* in the subnetwork may function normally in a Rho pathway to limit stress fiber formation and/or increase the turnover of F-actin structures in the absence of high levels of RhoA activity[5].

Another subnetwork contained SMAD4, and interacting genes. SMAD4 has been reported as a tumour suppressor gene during gastric carcinoma progression. Several copy number deletions and

inactivating mutations shown in SMAD4 indicate the functional loss of SMAD4.

In the analysis of the subtype GS, HotNet2 identified 4 subnetworks containing at least 6 genes ($p \leq 0.019$) with a corresponding FDR = 0.597. One subnetwork contained *RHOA* and *PKN2* (Figure S11.7c). *RHOA* has been reported to regulate *PKN2* and control entry into mitosis and exit from cytokinesis[6]. In addition, the other subnetwork contained four cadherin family genes *CDH1, CDH2, CDH3, CDH5*, *CTNNA1* and *PTPRM* (Figure S11.7b). Somatic mutations and deletions of *CDH1* gene have related to poor survival of patients with gastric cancer[7]. Moreover, *CDH1* and *CTNNA1* mutations might be important, because their germline mutations underlie Hereditary Diffuse Gastric Cancer (HDGC)[8] (Table S11.6).

In the analysis of the CIN subtype, HotNet2 identified 5 subnetworks containing at least 5 genes ($p \leq 0.002$) with a corresponding FDR = 0.41. A candidate subnetwork contained core genes in the p53 signalling pathway, i.e. *CDKN2A, CDK6* and *TP53* (Table S11.6).

In the analysis of the EBV subtype, HotNet2 identified 4 subnetworks containing at least 6 genes ($p < 0.001$) with a corresponding FDR = 0.517. We observed that a subnetwork containing *PIK3CA, KRAS, and NRAS* was mutated in 18 of the 22 samples. In addition, an exon-skipping event on *MET* and an *ITGB4* mutation in another subnetwork alter genes from a list reported to be biomarkers of gastric cancer[9] (Table S11.6).

In the analysis of the MSI subtype, HotNet2 identified 2 subnetworks containing at least 5 genes ($p \leq 0.329$) with a corresponding FDR = 1. Thus, these results were not statistically significant. A subnetwork with 8 genes contains the core of MHC class I including *B2M*, *HLA-B* and *HLA-E*, and *CD8A* a co-receptor of T-cells that interacts with MHC class I genes (Figure S11.7d). In addition, the *HFE* gene is related to gastric cancer due to its association with iron overload[10] (Table S11.6).

Pathway Enrichment

To focus attention on subnetworks with known biological function, we computed the overlap between the genes in candidate subnetworks and known pathways from the KEGG database [11]. Subnetworks returned by HotNet2 in STAD and in the CIN subtype had statistically significant (corrected $p \leq 0.05$) overlap with at least one KEGG pathway. Of those 4 subnetworks in the GS subtype, 3 had statistically significant overlap with one KEGG pathway, including cell adhesion, TGF-beta signalling, and complement and coagulation cascade pathways. In the EBV subtype, 2 subnetworks were enriched in at least one KEGG pathway. In MSI, the subnetwork containing B2M showed statistically significant overlap with the KEGG antigen processing and presentation pathway (Table S11.6).

All results can be reached in the link: http://compbio.cs.brown.edu/public/stad/

References

1. Leiserson, M., Vandin, F., Dobson, J., Wu, H.-T., Papoutsaki, A., Niu, B., McLellan, M., Lawrence, M., Gonzalez-Perez, A., Tamborero, D., Ryslik, G., Cheng, Y., Lopez-Bigas, N., Getz, G., Ding, L., Raphael, B. Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations across Pathways and Protein Complexes. *In preparation.*

2. Vandin, F., Upfal, E. and Raphael B., Algorithms for Detecting Significantly Mutated Pathways in Cancer, J Comput Biol. 2011 Mar;18(3):507-22.

3. Das, J. & Yu, H. HINT: High-quality protein interactomes and their applications in understanding

human disease. *BMC systems biology* **6**, 92 (2012).

4. Khurana, E., Fu, Y., Chen, J. & Gerstein, M. Interpretation of genomic variants using a unified biological network approach. *PLoS computational biology* **9**, e1002886 (2013).

5. Peck JW, Oberst M, Bouker KB, Bowden E, Burbelo PD. The RhoA-binding protein, rhophilin-2, regulates actin cytoskeleton organization. Journal of Biological Chemistry, 277, 43924-43932 (2002).

6. Schmidt A, Durgan J, Magalhaes A, Hall A. Rho GTPases regulate PRK2/PKN2 to control entry into mitosis and exit from cytokinesis. The EMBO Journal. p26, 1624 – 1636 (2007).

7. Corso G, Carvalho J, Marrelli D, Vindigni C, Carvalho B, Seruca R, Roviello F, Oliveira C. Somatic mutations and deletions of the E-cadherin gene predict poor survival of patients with gastric cancer. Journal of Clinical Oncology. vol. 31 no. 7 868-875 (2013).

8. Majewski IJ, Kluijt I, Cats A, Scerri TS, de Jong D, Kluin RJ, Hansford S, Hogervorst FB, Bosma AJ, Hofland I, Winter M, Huntsman D, Jonkers J, Bahlo M, Bernards R. An α-E-catenin (CTNNA1) mutation in hereditary diffuse gastric cancer. J. Pathol., 229: 621–629 (2013).

9. Guo T, Fan L, Ng WH, Zhu Y, Ho M, Wan WK, Lim KH, Ong WS, Lee SS, Huang S, Kon OL, Sze SK. Multidimensional Identification of Tissue Biomarkers of Gastric Cancer. J. Proteome Res. 11 (6), pp 3405–3413 (2012).

10. Agudo A et al. Hemochromatosis (HFE) gene mutations and risk of gastric cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC) study. Carcinogenesis. 34 (6): 1244-1250 (2013).

11. Kanehisa, M. and S. Goto, KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 28(1): p. 27-30 (2000).

Tables and Figures

SUBNETWORKS gene (# mutations)	**KEGG PATHWAYS ENRICHMENTS (corrected p-value)**
STAD	
TP53 (107), CDKN2A (47), CDK6 (31), MTA1 (19), PTEN (7)	**Cell cycle (0.011), Pathways in cancer (0.015), p53 signaling pathway (2.34e-06),**
ERBB2 (61), EGFR (23), CD44 (22), ERBB3 (10), ANK1 (6), ITPR3 (6), PTPRZ1 (6), SPTB (6), TLN1 (6), TRPC6 (6), ANKS1B (5), FER (5), IRS4 (5)	
SMAD4 (54), APBB2 (5), HDLBP (5)	
MYC (52), EP400 (6), SMC4 (5)	
RHOA (13), RHPN2 (10), FAM65B (6)	
FBLN2 (5), HSPG2 (5), NID1 (5)	
GS	
CDH1 (21), CDH2 (4), CTNNA1 (3), CDH5 (2), PTPRM (2), CDH3 (1)	**Cell adhesion molecules (CAMs) (2.17e-06)**
RHOA (8), FAM65B (3), MPRIP (2), PKN2 (2), CIT (1), RHPN2 (1)	
ACVR2A (3), MAGI2 (3), INHBA (2), ACVR1B (1), DSCAML1 (1), PLCL2 (1)	**TGF-beta signaling pathway (0.022)**
C3 (3), CR1 (2), CFB (1), CFH (1), CRP (1), ITGAM (1), ITGAX (1)	**Complement and coagulation cascades (7.43e-05)**
CIN	
TP53 (99), CDKN2A (40), CDK6 (32), MTA1 (21), PDZD2 (10), LAMA4 (8), WRN (6)	**Cell cycle (0.03), Pathways in cancer (0.027), p53 signaling pathway (0.0091),**
ERBB2 (53), EGFR (21), CD44 (17), TRPC4 (8), ITPR2 (7), ITPR3 (6), ANK1 (5), ANKS1B (4), FER (4), PTPRZ1 (4), SPTB	

(4), SUPT6H (4), TLN1 (4), TNS3 (4), TRPC5 (4), IRS4 (3), SH2B3 (3)	
CELSR3 (8), SPTBN4 (8), BAHCC1 (6), CLSTN1 (6), SCAPER (4), TRIP12 (4)	
THSD7B (9), HECTD1 (6), IGSF1 (5), CNTNAP4 (4), RANBP10 (3)	
DSCAML1 (5), PLXNB3 (4), CUL9 (3), MAGI3 (3), PLCL2 (3)	
EBV	
PIK3CA (18), ITIH1 (1), KRAS (1), NRAS (1), PLCE1 (1), SHOC2 (1)	**Neurotrophin signaling pathway (0.027), Insulin signaling pathway (0.048), VEGF signaling pathway (0.019), T cell receptor signaling pathway (0.027), Fc epsilon RI signaling pathway (0.011), B cell receptor signaling pathway (0.027)**
MET (8), GIPC1 (1), ITGA6 (1), ITGB4 (1), NRP1 (1), PLEC (1)	
JAK2 (3), CSF2RB (1), GHR (1), MPL (1), PTPN2 (1), SOCS3 (1), STAT5B (1)	**Jak-STAT signaling pathway ()**
COL1A2 (1), COL5A1 (1), DCN (1), F2 (1), IGF1 (1), IGF2 (1), IGFALS (1), IGFBP5 (1), THBS1 (1)	
MSI	
ERBB3 (35), RASA1 (20), ARHGAP5 (17), DOK2 (5), ANXA6 (4)	
B2M (23), HLA-B (13), CD8A (5), HFE (5), KLRD1 (5), TFRC (5), HLA-E (4), KLRC3 (3)	**Natural killer cell mediated cytotoxicity (0.0007),** **Graft-versus-host disease (0.0039),** **Antigen processing and presentation (2.13e-08)** **Cell adhesion molecules (CAMs) (0.037)**

Table S11.6: The candidate subnetworks identified by HotNet, and corresponding KEGG pathways with significant overlap with each subnetwork. For the pathways, we list the name of the pathway and the (multiple hypothesis corrected) *p*-value of the hypergeometric enrichment test.

Figure S11.7a: The oncoprint, transcript annotation, copy number aberrations, and protein-protein interaction network of the candidate subnetwork contained the ErbB pathway.

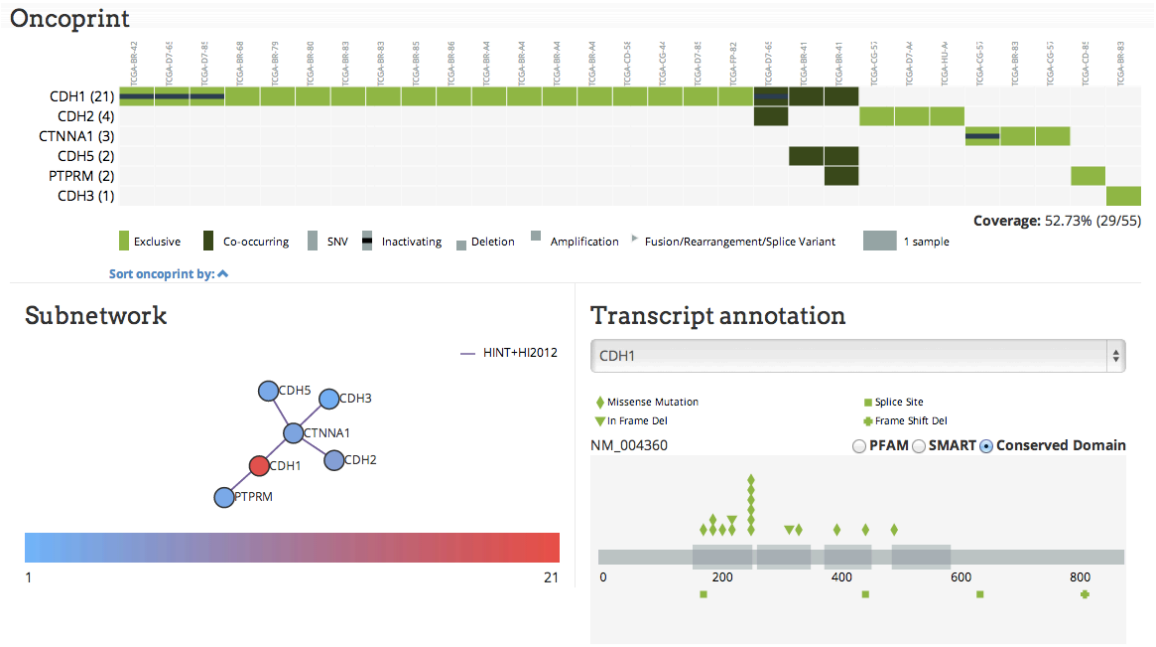


Figure S11.7b: The oncoprint, transcript annotation and protein-protein interaction network of the candidate subnetwork contain the cadherin gene family in the subtype GS; 29 of 55 GS samples were mutated.

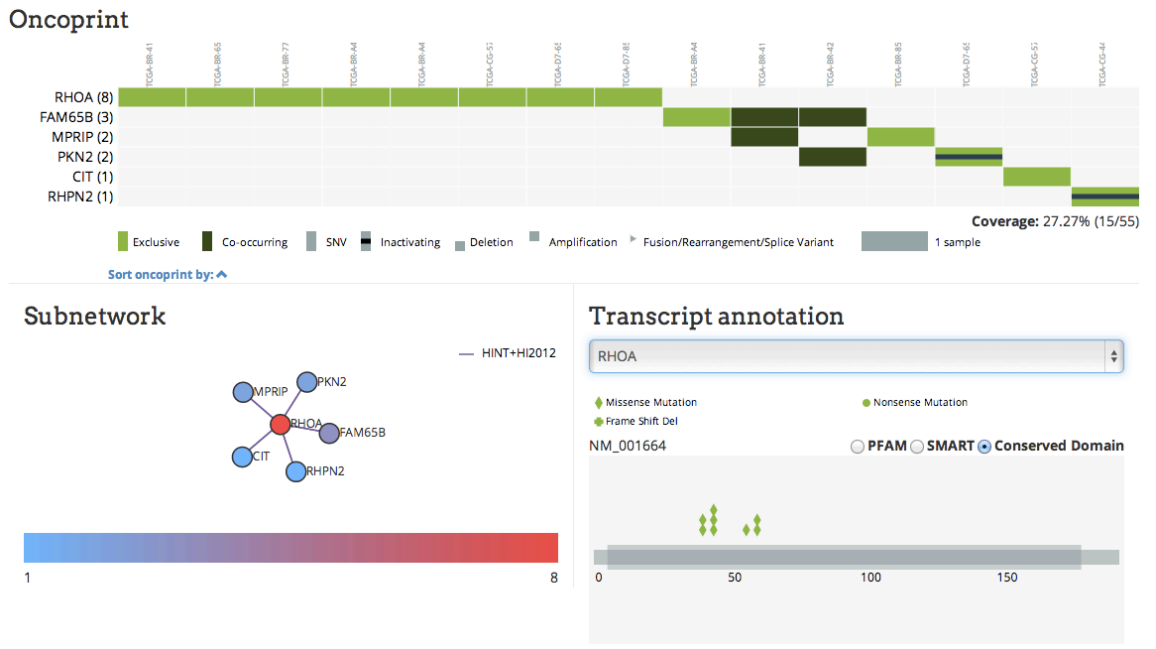


Figure S11.7c: The oncoprint, transcript annotation and protein-protein interaction network of the candidate subnetwork contained the RhoA signaling pathway in the subtype GS; 15 of 55 GS samples were mutated.

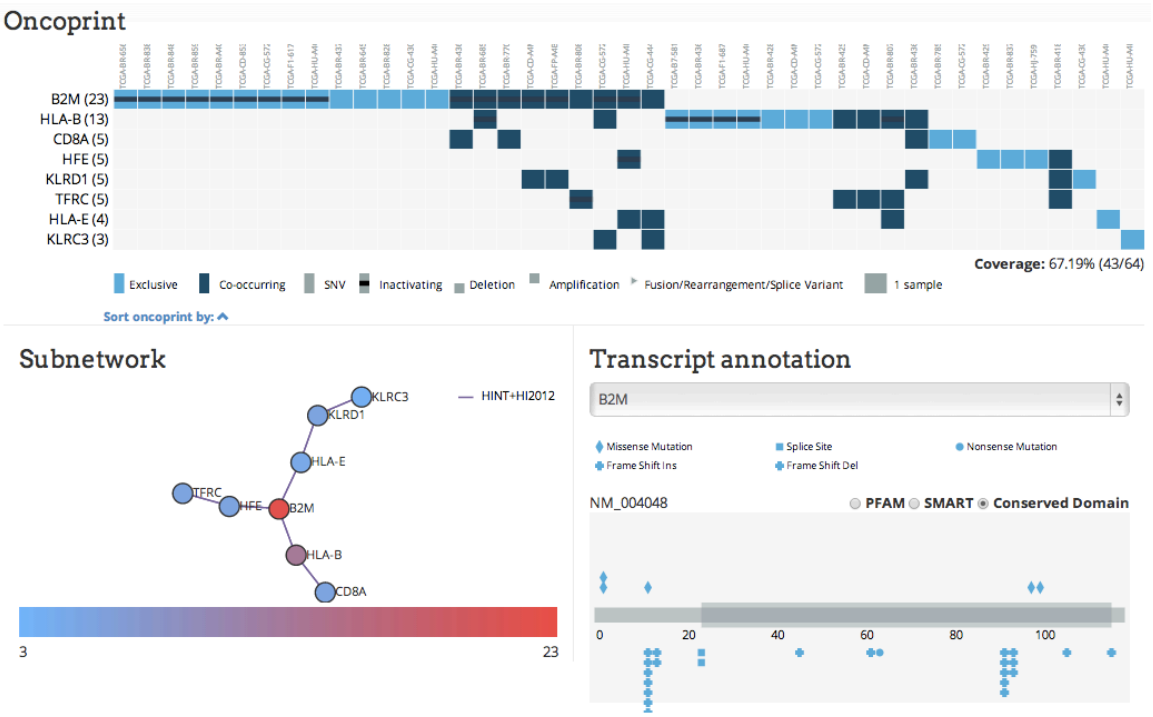


Figure S11.7d: The oncoprint, transcript annotation and protein-protein interaction network of the candidate subnetwork contained MHC class I in the subtype MSI; 43 of 64 MSI samples were mutated.

Supplement S11.8: All-by-all Pairwise Associations, Regulome Explorer, and GeneSpot

Statistical association among the diverse data elements in this study was evaluated by comparing pairs of columns in the feature matrix (Supplement S11.1). Hypothesis testing was performed by testing against null models for absence of association, yielding a *p*-value. *P*-values for the association between and among clinical and molecular data elements were computed according to the nature of the data levels for each pair: discrete vs. discrete (Fisher's exact test); discrete vs. continuous (ANOVA *F*- test, equivalently *t*-test for binary vs. continuous) or continuous vs. continuous (*F*-test). Ranked data values were used in each case. To account for multiple- testing bias, the *p*-value was adjusted using the Bonferroni correction.

In order to allow researchers to further explore potentially interesting relationships in this dataset, including primary data, the statistically significant pairs of associations were loaded into the Regulome Explorer web application, which is designed to allow researchers to explore associations among multiple data types in cancer genomics (http://explorer.cancerregulome.org). Prior to loading, a *p*-value threshold was chosen specific to each pair of data types (e.g. clinical data vs. gene expression data) in such a way as to strike a balance between making potentially interesting associations available to queries by the tool, while still allowing the tool to be responsive, since the number of loaded graph edges (each corresponding to a statistically significant relationship) is in the millions.

All identified pairwise relationships, including those described in this manuscript can be found at http://explorer.cancerregulome.org·

To allow researchers to explore this dataset in the context of other TCGA data types and to provide additional plotting and querying capabilities, the data have been made available in the GeneSpot web application at www.genespot.org. This software tool for systems biology provides a way to view TCGA data from a gene-centric point-of-view.

S11.9 Firehose Analysis

The Broad Institute GDAC Firehose provides TCGA preprocessed data and analysis pipelines to the cancer research community. It generates regular standard runs as well as customized runs to coordinate with the TCGA analysis working groups (AWGs). The Stomach Adenocarcinoma (STAD) AWG runs were based on the AWG data.

The AWG run generated Nozzle[1] reports for the analyses of copy number[2], mutation[3,4], methylation, mRNA and miRNA expression, and protein quantification. In addition to the analysis reports of individual data types, the AWG runs also generated reports of the associations between multiple selected data types, including clinical data.

Reports of the GDAC STAD AWG runs in line with the data freeze for manuscript submission are reflected on the Firehose webpage
at http://gdac.broadinstitute.org/runs/awg_stad__2013_09_30/

References

1. Gehlenborg, N., Noble, M., Getz, G., Chin, L., & Park, P. Nozzle: a report generation toolkit for data analysis pipelines. *Bioinformatics* doi:**10**.1093 (2013)
2. Mermel, C., Schumacher, S., Hill, B., Meyerson, M., Beroukhim, R. & Getz, G. Gistic2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology **12**:R41 (2011)
3. Cibulskis, K., Lawrence, M., Carter, S., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. & Getz, G. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature Biotechnology **31**, 213-219 (2013)
4. Lawrence, M *et al*. Mutational heterogeneity in cancer and the search for new cancer- associated genes. Nature **499**, 214-218 (2013)

S11.10. MIRACLE Analysis

We used MIRACLE (**M**aster m**I**R**R**na **A**nalysis for **C**ancer mo**L**ecular subtyp**E**)[1] to identify a miRNA regulatory network driven by epigenetically silenced miRNAs in the EBV subtype. MIRACLE identifies a miRNA-gene regulatory network by integrating the DNA methylation, miR-seq and RNA-seq data.

miR-seq, DNA methylation, and RNA-seq Data for MIRACLE Analysis

MiRNA isoform expression files were downloaded from Synapse. The expression levels of 3p/5p mature miRNA were summarized by the MIRACLE pipeline. Briefly, read numbers mapped to the same miRNA isoform (based on MIMAT ID) were summed up regardless of their sequence variations. The MIMAT IDs were further converted to miRNA mature product names according to miRBase V19 annotations[2]. The numbers of reads that were mapped to the precursors, stemloops and unannotated/retired miRNAs were summed up as "Precursor/Stemloop/Unannotated" in each tumour sample. We also identified the DNA methylation probes which represent the methylation status of miRNA promoters by mapping the ~450k HumanMethylation450 probes with 1kb region regions of miRNAs. This analysis identified 3439 probes. Among the 289 STAD samples, 98 cases had complete HumanMethylation450 microarray, miR-seq and RNA-seq data. We first integrated DNA methylation data with microRNA expression data to identify the microRNAs that are probably directly regulated by DNA methylation. We ran MIRACLE using the EBV subtype versus the other three non-EBV subtypes. Spearman rank correlation was used to analyze the correlation between miRNA expression level and DNA methylation status of the same miRNA. A refined subset of epigenetically silenced miRs was then selected according to two criteria: 1) each miRNA was significantly inversely correlated with DNA methylation (FDR<1%) and 2) the miRNA was also significantly altered (FDR<1%) between the EBV subgroup and other subgroups. After we identified the epigenetically silenced miRNAs, we performed the regulatory network prediction as previously described[1]. The miRNA seed binding information was from TargetScan[3].

Epigenetically silenced miRNAs and predicted network

MIRACLE identified nine miRNAs that are potentially regulated by DNA methylation (Figure S11.11). Further network analysis revealed the miRNA-regulatory network, including seven epigenetically silenced miRNAs and 83 predicted targets. Among the most regulated miRNAs is miR-9, which has been reported as a potential tumour suppressing miRNA in multiple cancer types[4]. miR-9 has recently been established to be epigenetically silenced in gastric cancer[4]. The DNA methylation of the miR-9 gene and its reverse association with 3p and 5p miRNA-products can be seen in Figure S11.12. Pathway enrichment analysis on the targets predicted to be regulated by epigenetically silenced miRNAs yielded DNA repair, leukocyte activation and cell adhesion pathways.

References:

1. Yang, D., et al., *Integrated analyses identify a master microRNA regulatory network for the mesenchymal subtype in serous ovarian cancer.* Cancer cell, 2013. 23(2): p. 186-199.
2. Kozomara, A. and S. Griffiths-Jones, *miRBase: annotating high confidence microRNAs using deep sequencing data.* Nucleic acids research, 2014. 42(1): p. D68-73.
3. Lewis, B.P., et al., *Prediction of mammalian microRNA targets.* Cell, 2003. 115(7): p.787-98.
4. Tsai, K.W., et al., *Aberrant hypermethylation of miR-9 genes in gastric cancer.* Epigenetics: official journal of the DNA Methylation Society, 2011. 6(10): p. 1189-97

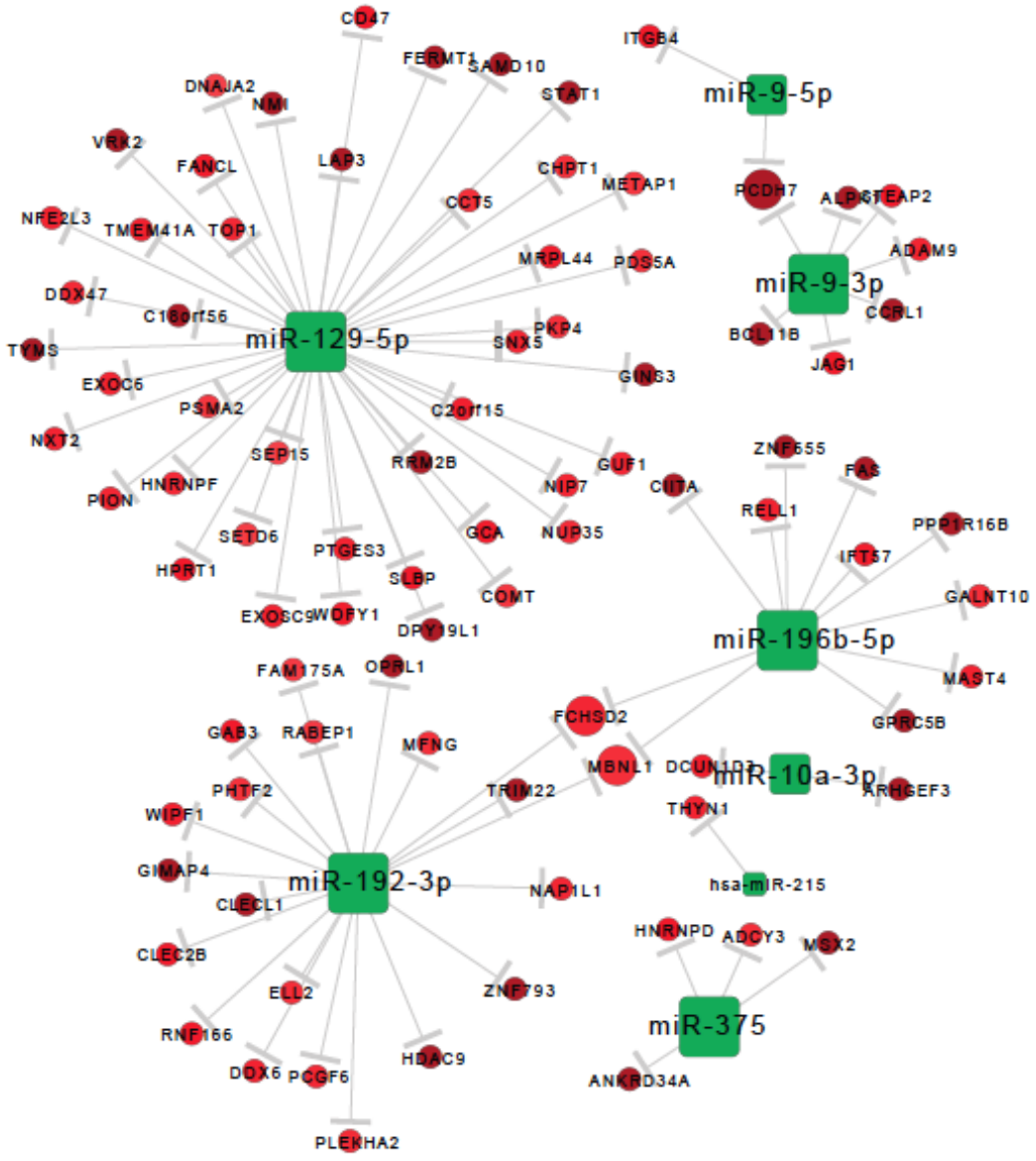


Figure S11.11. miR-RNA regulatory network for epigenetically silenced miRNAs.

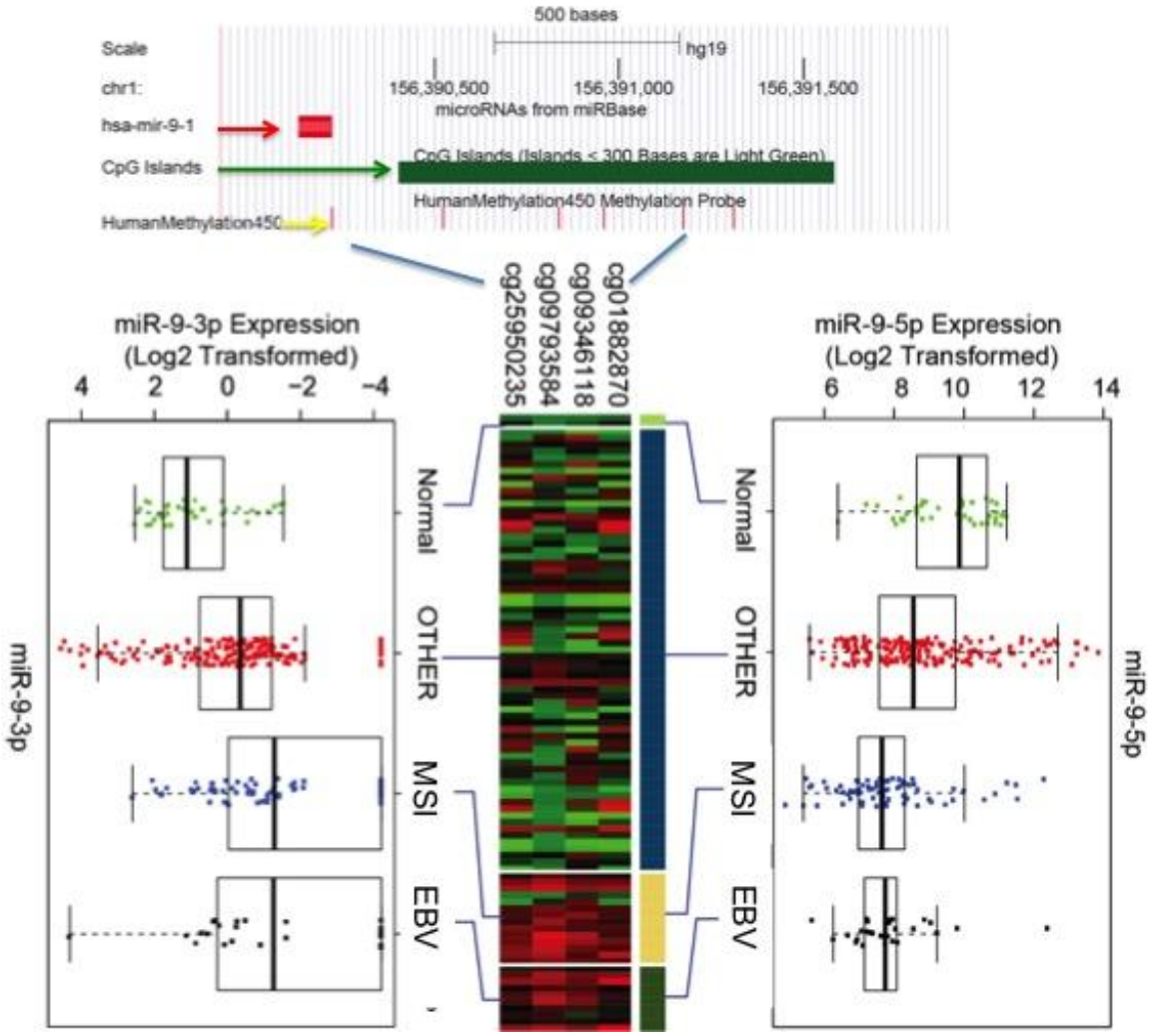


Figure S11.12. DNA methylation and expression of miR-9

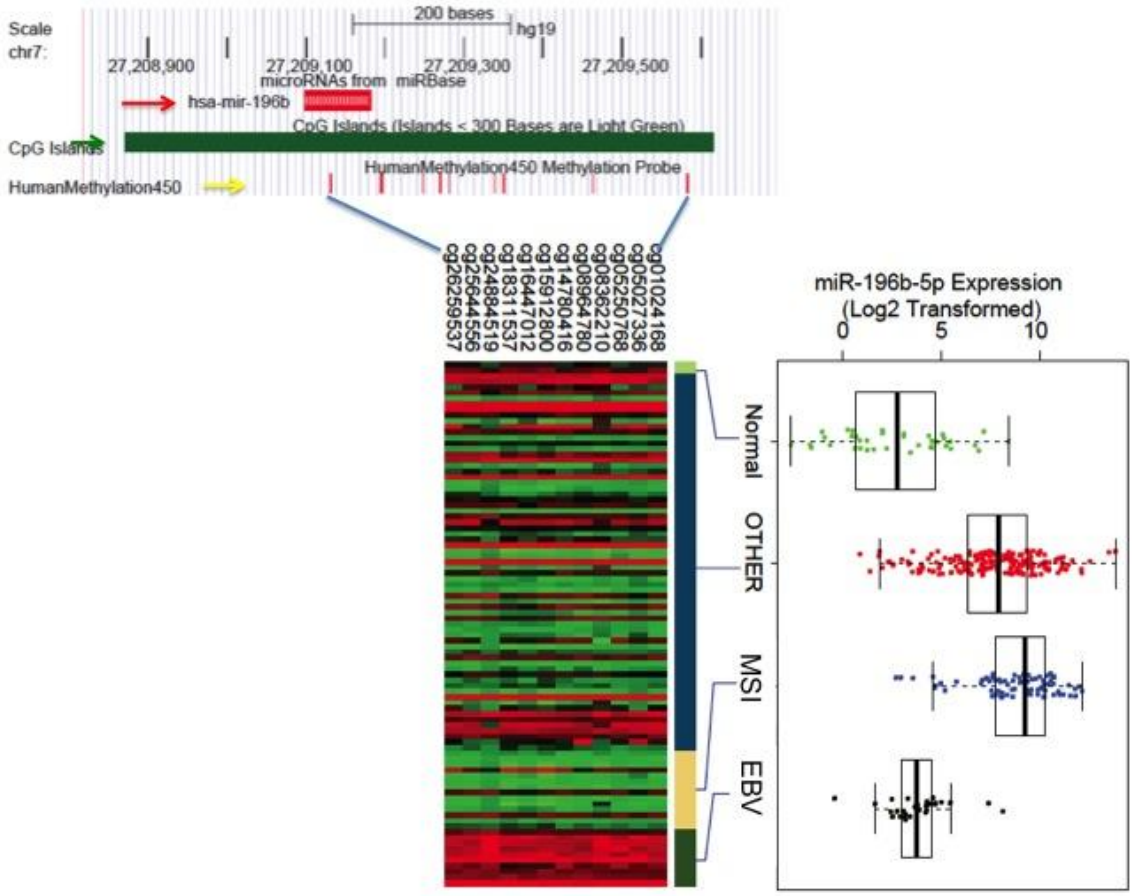
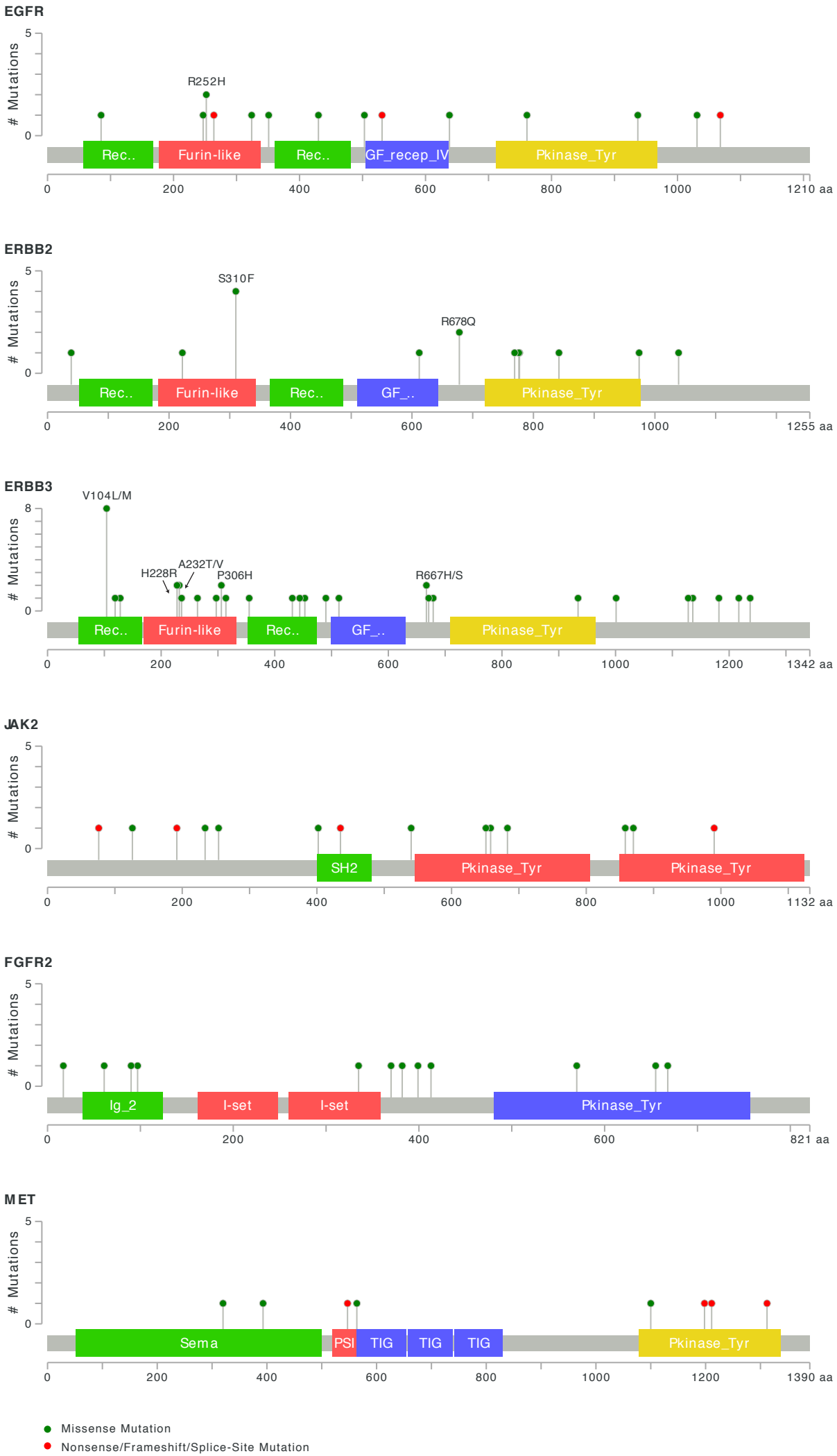


Figure S11.13. DNA methylation and expression of miR-196b



S11.14 figure- Somatic mutations recurrently altered in receptor tyrosine kinases

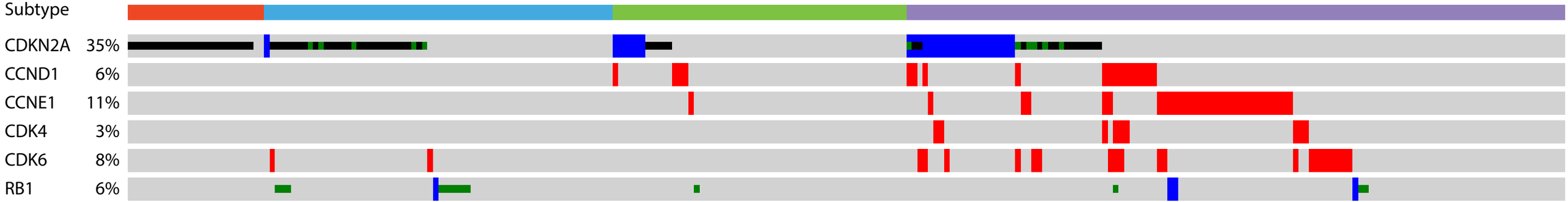


figure 11.15- Oncoprint of cell cycle genes

S12. TCGA Funding Sources

S12. TCGA Funding Sources

This work was supported by the following grants from the United States National Institutes of Health: U54 HG003067 (E. Lander), U24 CA143799 (P. Spellman), U24 CA143835 (I. Shmulevich), U24 CA143840 (C. Sander), U24 CA143843 (D. Wheeler), U24 CA143845 (G. Getz), U24 CA143848 (C. Perou), U24 CA143858 (D. Haussler), U24 CA143866 (M. Marra), U24 CA143867 (M. Meyerson), U24 CA143882 (P. Laird), U24 CA143883 (G. Mills), and U24 CA144025 (R. Kucherlapati).