Summary of TCGA GBM analysis

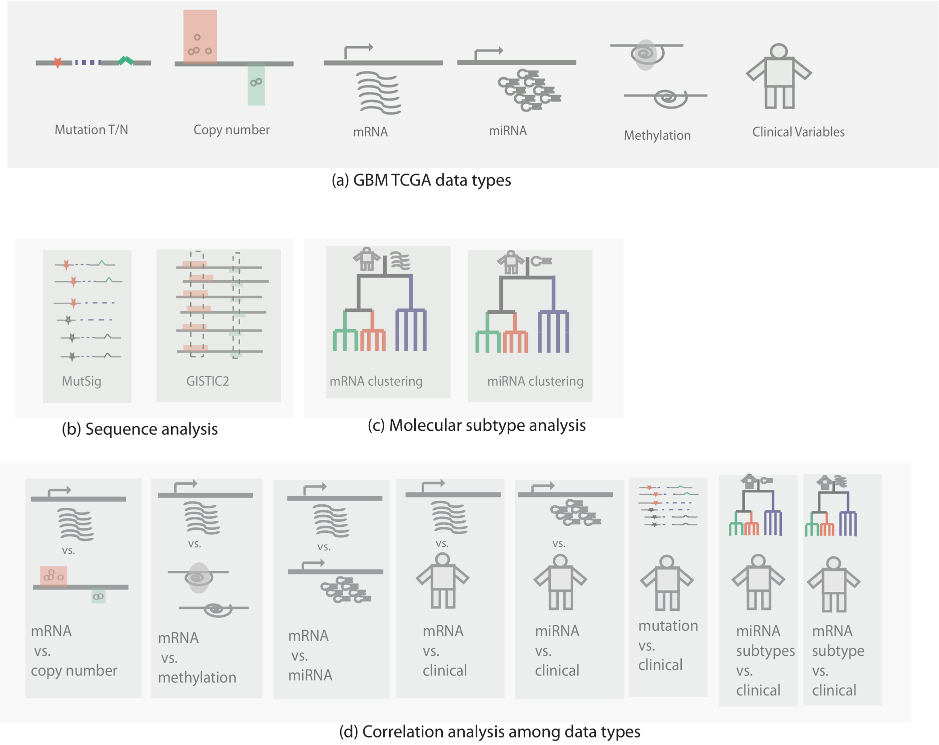
Lihua Zou

Jan 23rd, 2011

Summary

The analysis performed within firehose workspace *prod_2011_01_14_gbm_02*, can be grouped into three general categories: mutation and copy number analysis, molecular subtype clustering, and correlation analysis across data types. For an overview of workflow, please see Figure 1.

The analysis pipeline identified 106 significant mutated genes (q<0.1); and 10 significant genes with mutations from COSMIC. The molecular subtype analysis identified one mRNA subtype cluster significantly associated with *VITALSTATUS (p<0.00167).* A large number of DNA regions have copy amplification and deletion. No mutation gene and miRNA subtype clusters are found to be associated with clinical parameters. A list of mRNA genes is highly correlated with methylation, clinical variables and copy number change.



(a) GBM TCGA data types

(b) Sequence analysis

(c) Molecular subtype analysis

(d) Correlation analysis among data types

Figure 1: analysis overview

Result

Mutation and copy number analysis

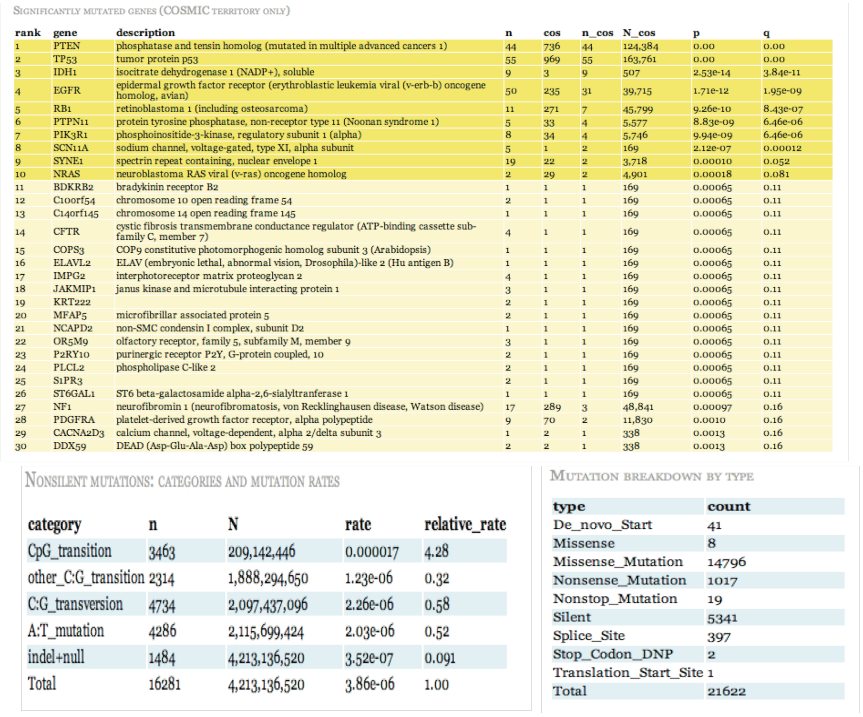
Mutation analysis: We use our in-house gene significance calling method (MutSig: unpublished; [16]) to call significant mutated genes. We identified 106 (q<0.1) significantly mutated genes from sequences of 169 individuals. There are 10 significant genes with mutation found previously from COSMIC. There are 362 genes with clustered mutations (<=3 amino acids apart). There are 21622 mutations after filtering mutations outside of gene sets and from zero-coverage samples [16]. There are 16281 non-silent mutations. The top ranked genes and breakdown of mutations by type and categories is shown in Figure 2.

Administrator 1/24/11 11:07 AM

Comment: 169 are WGA now. There are 24 native samples for WES coming. Need to include this information. 19 of 20 WGS is available. 1 sample with incomplete checksum. Aaron is running co-cleaning on 14-15 WGS samples now.

Administrator 1/23/11 9:54 PM

Comment: Is this significance calculated only using the mutations seen in COSMIC?



Significantly mutated genes (COSMIC territory only)

rank	gene	description	n	cos	n_cos	N_cos	p	q
1	PTEN	phosphatase and tensin homolog (mutated in multiple advanced cancers 1)	44	736	44	124,384	0.00	0.00
2	TP53	tumor protein p53	55	969	55	163,761	0.00	0.00
3	IDH1	isocitrate dehydrogenase 1 (NADP+), soluble	9	3	9	507	2.53e-14	3.84e-11
4	EGFR	epidermal growth factor receptor (erythroblastic leukemia viral (v-erb-b) oncogene homolog, avian)	50	235	31	39,715	1.71e-12	1.95e-09
5	RB1	retinoblastoma 1 (including osteosarcoma)	11	271	7	45,799	9.26e-10	8.43e-07
6	PTPN11	protein tyrosine phosphatase, non-receptor type 11 (Noonan syndrome 1)	5	33	4	5,577	8.83e-09	6.46e-06
7	PIK3R1	phosphoinositide-3-kinase, regulatory subunit 1 (alpha)	8	34	4	5,746	9.94e-09	6.46e-06
8	SCN11A	sodium channel, voltage-gated, type XI, alpha subunit	5	1	2	169	2.12e-07	0.00012
9	SYNE1	spectrin repeat containing, nuclear envelope 1	19	22	2	3,718	0.00010	0.052
10	NRAS	neuroblastoma RAS viral (v-ras) oncogene homolog	2	29	2	4,901	0.00018	0.081
11	BDKRB2	bradykinin receptor B2	1	1	1	169	0.00065	0.11
12	C10orf54	chromosome 10 open reading frame 54	2	1	1	169	0.00065	0.11
13	C14orf145	chromosome 14 open reading frame 145	1	1	1	169	0.00065	0.11
14	CFTR	cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)	4	1	1	169	0.00065	0.11
15	COPS3	COP9 constitutive photomorphogenic homolog subunit 3 (Arabidopsis)	1	1	1	169	0.00065	0.11
16	ELAVL2	ELAV (embryonic lethal, abnormal vision, Drosophila)-like 2 (Hu antigen B)	1	1	1	169	0.00065	0.11
17	IMPG2	interphotoreceptor matrix proteoglycan 2	4	1	1	169	0.00065	0.11
18	JAKMIP1	janus kinase and microtubule interacting protein 1	3	1	1	169	0.00065	0.11
19	KRT222		2	1	1	169	0.00065	0.11
20	MFAP5	microfibrillar associated protein 5	2	1	1	169	0.00065	0.11
21	NCAPD2	non-SMC condensin I complex, subunit D2	1	1	1	169	0.00065	0.11
22	OR5M9	olfactory receptor, family 5, subfamily M, member 9	3	1	1	169	0.00065	0.11
23	P2RY10	purinergic receptor P2Y, G-protein coupled, 10	2	1	1	169	0.00065	0.11
24	PLCL2	phospholipase C-like 2	2	1	1	169	0.00065	0.11
25	S1PR3		2	1	1	169	0.00065	0.11
26	ST6GAL1	ST6 beta-galactosamide alpha-2,6-sialyltranferase 1	1	1	1	169	0.00065	0.11
27	NF1	neurofibromin 1 (neurofibromatosis, von Recklinghausen disease, Watson disease)	17	289	3	48,841	0.00097	0.16
28	PDGFRA	platelet-derived growth factor receptor, alpha polypeptide	9	70	2	11,830	0.0010	0.16
29	CACNA2D3	calcium channel, voltage-dependent, alpha 2/delta subunit 3	1	2	1	338	0.0013	0.16
30	DDX59	DEAD (Asp-Glu-Ala-Asp) box polypeptide 59	2	2	1	338	0.0013	0.16

Nonsilent mutations: categories and mutation rates

category	n	N	rate	relative_rate
CpG_transition	3463	209,142,446	0.000017	4.28
other_C:G_transition	2314	1,888,294,650	1.23e-06	0.32
C:G_transversion	4734	2,097,437,096	2.26e-06	0.58
A:T_mutation	4286	2,115,699,424	2.03e-06	0.52
indel+null	1484	4,213,136,520	3.52e-07	0.091
Total	16281	4,213,136,520	3.86e-06	1.00

Mutation breakdown by type

type	count
De_novo_Start	41
Missense	8
Missense_Mutation	14796
Nonsense_Mutation	1017
Nonstop_Mutation	19
Silent	5341
Splice_Site	397
Stop_Codon_DNP	2
Translation_Start_Site	1
Total	21622

Figure 2: gbm_mutsig

Copy number analysis: We used GISTIC2 [11] to perform copy number analysis to identify genomic regions showing amplification and deletion. The significantly amplified region and deleted region are shown in Figure 3.

Administrator 1/23/11 9:33 PM

Comment: The report of GISTIC2 is largely incomplete. Can someone give me a quick intuition of what is G-score and how is the cutoff chosen to define significant amplification/deletion?

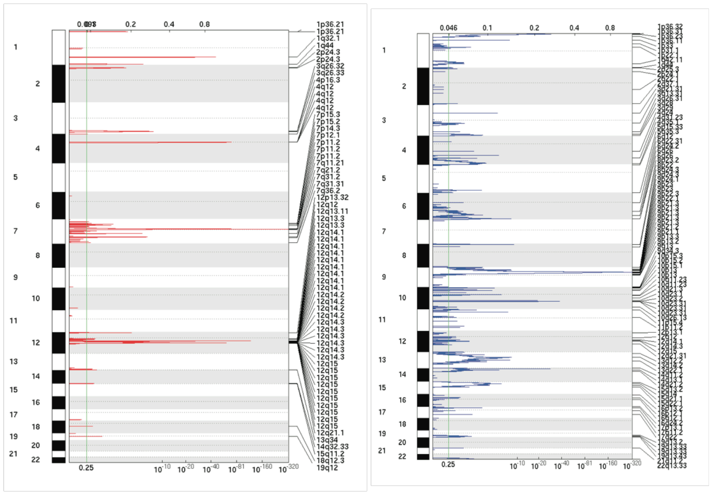


Figure 3: gbm_gistic2

Molecular subtype clustering

*mRNA subtypes***:** We applied consensus non-negative matrix factorization method to identify molecular subtypes based on mRNA expression [14]. We select 1500 most variable genes and applied Consensus NMF clustering method to classify 440 samples. Our analysis identified 3 subtypes. "Core samples" representative of each cluster were identified based on positive silhouette width [14]. Core samples indicate higher similarity to their own class than to any other classes. We used core samples to select differentially expressed marker genes (p<=0.05) for each subtype by comparing the subclass versus the other subclasses based on student's t-test. In addition, we also applied an alternative consensus hierarchical clustering methods [15] using 440 samples and 1500 genes to identify 4 molecular subtypes (Figure 4).

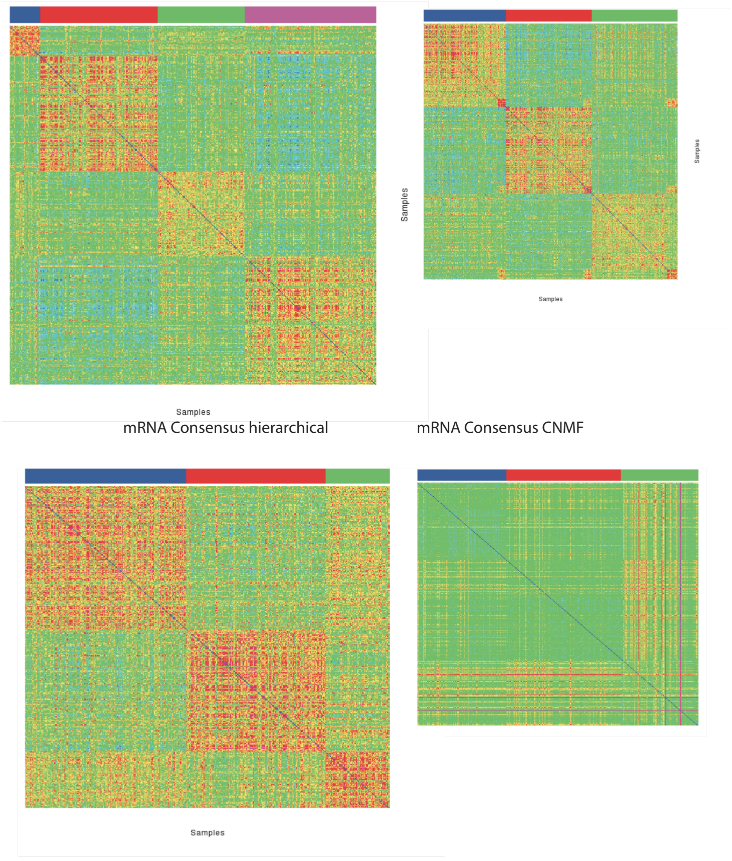
Administrator 1/23/11 10:00 PM

Comment: This is reasonable! An alternative and arguably more powerful approach is to use PAM analysis developed by Tibshironi et al which is based on a modified t-test by adding a Bayesion factor to the denominator.

Administrator 1/24/11 11:29 AM

Comment: We might need to compare the result from two clustering approaches. Maybe need to compare with public reference GBM gene sets.

miRNA subtypes: We used similar approach to identify molecular subtypes based on miRNA expression. We select 150 most variable miRNAs. We applied CNMF consensus clustering to 415 samples and identified 3 subtypes. We also applied consensus hierarchical clustering to 415 samples to identify 3 subtypes using the 150 most variable miRNAs (Figure 4).



mRNA Consensus hierarchical

mRNA Consensus CNMF

miRNA consensus hiearchical

miRNA consensus CNMF

Figure 4: gbm_subtypes

Correlation across data types

Mutation vs. clinical: We examined the association between the status of the 98 significantly mutated genes and clinical *VITALSTATUS* of 167 samples. We used the chi-square test to calculate the significance of association. No single mutated gene is found to be significantly associated with *VITALSTATUS*.

*Molecular subtypes vs. clinical***:** We found significant association between the four subtype clusters identified by *CONSENSUS_MRNA_CLUSTERING* and clinical feature '*VITALSTATUS*' (chi-square test p-value < 0.00167). However, we didn't found significant association between the 3 subtypes identified by *CNMFCLUSTERING_MRNA* and clinical feature *VITALSTATUS*. The P value by *Chi-*

Administrator 1/24/11 11:00 AM

Comment: Howis the 98 genes chosen? Besides single mutation, it will also be interesting to test co-mutation and combination of mutations for clinical association as well. I will work to include this feature into NetSig!

Administrator 1/23/11 10:22 PM

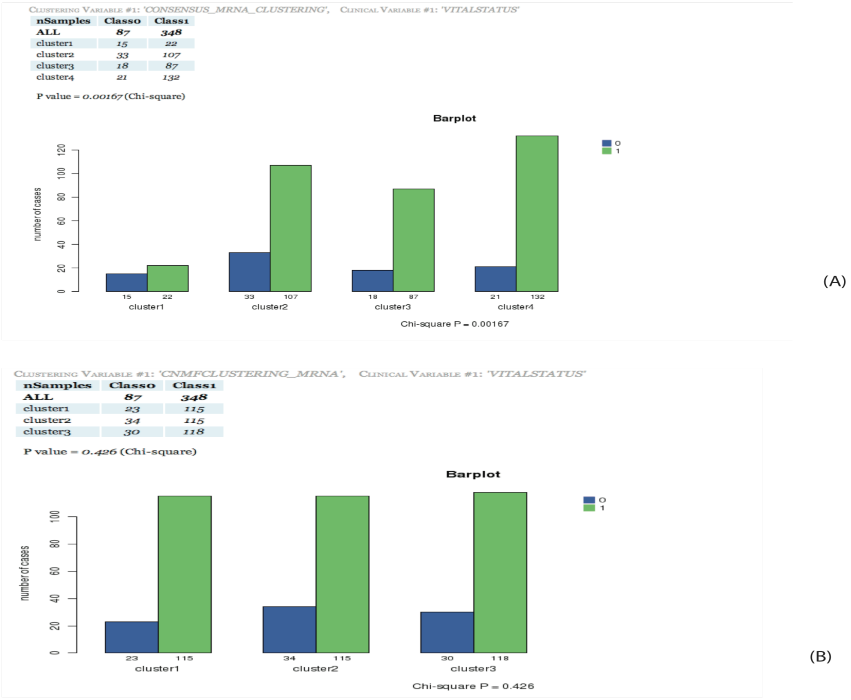
Comment: When any expected value in contingency table is smaller than 5, probably should use Fisher's exact test to estimate p-value.

square test is 0.426. The significant association is mostly driven by the smallest mRNA cluster by *CONSENSUS_MRNA_CLUSTERING* (Figure 5).

Administrator 1/23/11 9:46 PM

Comment: Again, we need to compare the two clustering result to confirm if this real.

We didn't find significant association between *CNMFCLUSTERING_MIRNA* and clinical feature *VITALSTATUS* (*Chi-square pval*=0.868); also no association found between *CONSENSUS_MIRNA_CLUSTERING* and clinical feature *VITALSTATUS* (chi-square p-value=0.358) [1].



Clustering Variable #1: '*CONSENSUS_MRNA_CLUSTERING*', Clinical Variable #1: '*VITALSTATUS*'

nSamples	Class0	Class1
ALL	***87***	***348***
cluster1	*15*	*22*
cluster2	*33*	*107*
cluster3	*18*	*87*
cluster4	*21*	*132*

P value = *0.00167* (Chi-square)

Clustering Variable #1: '*CNMFCLUSTERING_MRNA*', Clinical Variable #1: '*VITALSTATUS*'

nSamples	Class0	Class1
ALL	***87***	***348***
cluster1	*23*	*115*
cluster2	*34*	*115*
cluster3	*30*	*118*

P value = *0.426* (Chi-square)

Figure 5: correlate subtypes with clinical variable *VITALSTATUS*.

*miRNA/mRNA vs. clinical***:** we performed association analysis between 556 miRNAs and 6 clinical features of 415 samples. The 6 clinical features are as following: *PATIENTTUMORRECURRENCESTATUS, KARNOFSKYPERFORMANCESCORE, HISTOLOGICALTYPE, VITALSTATUS, NEOADJUVANTTHERAPY, GENDER*. 556 genes are used based on a statistical selection criteria at P value <= 0.01. The numbers

Administrator 1/24/11 11:15 AM

Comment: Ask Gordon for more clinical parameters (tier1). For high mutation samples, neeed to check if treated or mismatch (dbGAP).

Administrator 1/23/11 10:22 PM

Comment: Which statistical criteria?

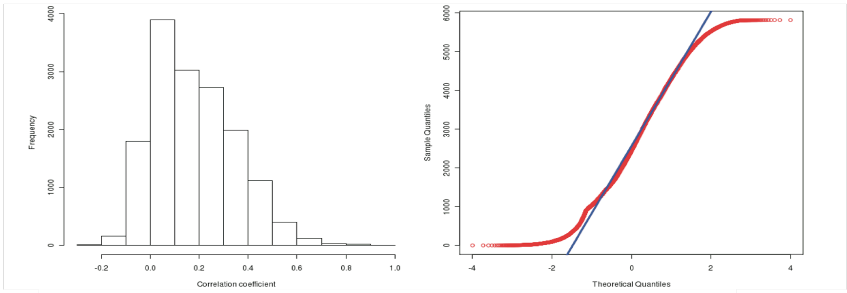
of genes that are significantly associated with each clinical feature are linked in reference [4].

We also performed association analysis between 18699 mRNAs and the same 6 clinical features of 435 samples. The numbers of genes that are significantly associated with each clinical feature are linked in reference [1].

mRNA/miRNA expr vs. copy number: we calculated the pearson correlation between expression intensity and log2 copy number (the gene-by-sample copy number data is obtained using CNTools package of bioconductor). The correlation distribution and significantly correlated mRNA genes are shown in Figure 6. The correlation distribution and significantly correlated miRNA genes are shown in Figure 7.

Administrator 1/23/11 10:22 PM

Comment: From the correlation distribution, it looks about 50% expression variation are explained by copy number (if we assume cor>0.35 is significant). Perhaps we can do a regression analysis to get a more accurate estimate about this.



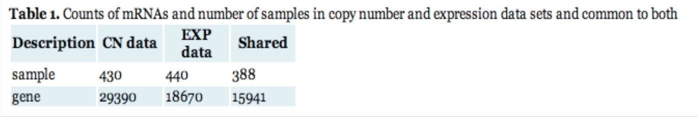


Table 1. Counts of mRNAs and number of samples in copy number and expression data sets and common to both

Description	CN data	EXP data	Shared
sample	430	440	388
gene	29390	18670	15941

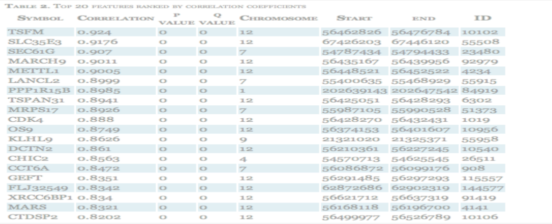
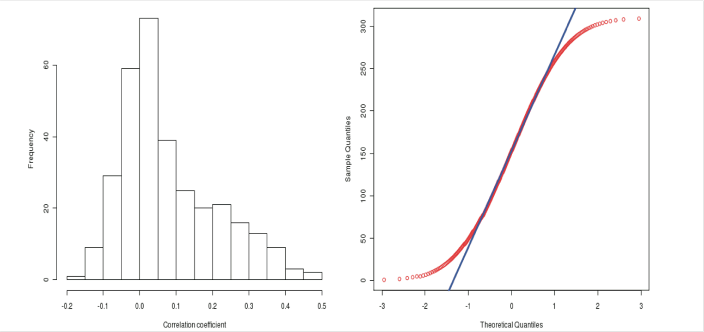


Table 2. Top 20 features ranked by correlation coefficients

Symbol	Correlation	P value	Q value	Chromosome	Start	end	ID
TSFM	0.924	0	0	12	56462826	56476784	10102
SLC35E3	0.9176	0	0	12	67426203	67446120	55508
SEC61G	0.907	0	0	7	54787434	54794433	23480
MARCH9	0.9011	0	0	12	56435167	56439956	92979
METTL1	0.9005	0	0	12	56448521	56452522	4234
LANCL2	0.8999	0	0	7	55400635	55468929	55915
PPP1R15B	0.8985	0	0	1	202639143	202647542	84919
TSPAN31	0.8941	0	0	12	56425051	56428293	6302
MRPS17	0.8926	0	0	7	55987105	55990528	51373
CDK4	0.888	0	0	12	56428270	56432431	1019
OS9	0.8749	0	0	12	56374153	56401607	10956
KLHL9	0.8626	0	0	9	21321020	21325371	55958
DCTN2	0.861	0	0	12	56210361	56227245	10540
CHIC2	0.8563	0	0	4	54570713	54625545	26511
CCT6A	0.8472	0	0	7	56086872	56099176	908
GEFT	0.8351	0	0	12	56291485	56297293	115557
FLJ32549	0.8342	0	0	12	62872686	62902319	144577
XRCC6BP1	0.834	0	0	12	56621712	56637319	91419
MARS	0.8321	0	0	12	56168118	56196700	4141
CTDSP2	0.8202	0	0	12	56499977	56526789	10106

Figure 6: correlate mRNA expression and copy number



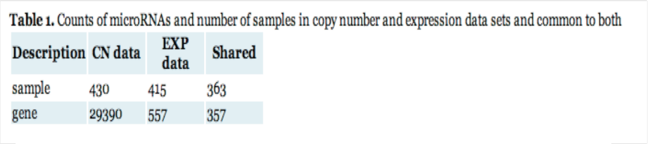


Table 1. Counts of microRNAs and number of samples in copy number and expression data sets and common to both

Description	CN data	EXP data	Shared
sample	430	415	363
gene	29390	557	357

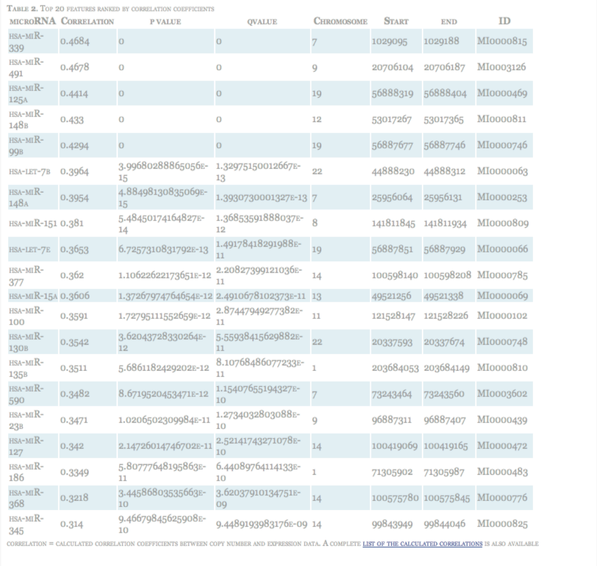


Table 2. Top 20 features ranked by correlation coefficients

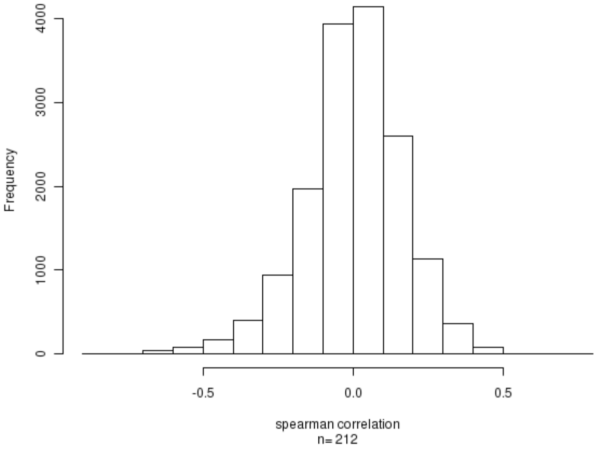
microRNA	Correlation	p value	qvalue	Chromosome	Start	end	ID
hsa-miR-339	0.4684	0	0	7	1029095	1029188	MI0000815
hsa-miR-491	0.4678	0	0	9	20706104	20706187	MI0003126
hsa-miR-125a	0.4414	0	0	19	56888319	56888404	MI0000469
hsa-miR-148b	0.433	0	0	12	53017267	53017365	MI0000811
hsa-miR-99b	0.4294	0	0	19	56887677	56887746	MI0000746
hsa-let-7b	0.3964	3.99680288865056e-15	1.32975150012667e-13	22	44888230	44888312	MI0000063
hsa-miR-148a	0.3954	4.88498130835069e-15	1.3930730001327e-13	7	25956064	25956131	MI0000253
hsa-miR-151	0.381	5.48450174164827e-14	1.36853591888037e-12	8	141811845	141811934	MI0000809
hsa-let-7e	0.3653	6.7257310831792e-13	1.49178418291988e-11	19	56887851	56887929	MI0000066
hsa-miR-377	0.362	1.10622622173651e-12	2.20827399121036e-11	14	100598140	100598208	MI0000785
hsa-miR-15a	0.3606	1.37267974764654e-12	2.4910678102373e-11	13	49521256	49521338	MI0000069
hsa-miR-100	0.3591	1.72795111552659e-12	2.87447949277382e-11	11	121528147	121528226	MI0000102
hsa-miR-130b	0.3542	3.62043728330264e-12	5.55938415629882e-11	22	20337593	20337674	MI0000748
hsa-miR-135b	0.3511	5.6861182429202e-12	8.10768486077233e-11	1	203684053	203684149	MI0000810
hsa-miR-590	0.3482	8.6719520453471e-12	1.15407655194327e-10	7	73243464	73243560	MI0003602
hsa-miR-23b	0.3471	1.0206502309984e-11	1.2734032803088e-10	9	96887311	96887407	MI0000439
hsa-miR-127	0.342	2.14726014746702e-11	2.52141743271078e-10	14	100419069	100419165	MI0000472
hsa-miR-186	0.3349	5.80777648195863e-11	6.44089764114133e-10	1	71305902	71305987	MI0000483
hsa-miR-368	0.3218	3.44586803535663e-10	3.62037910134751e-09	14	100575780	100575845	MI0000776
hsa-miR-345	0.314	9.46679845625908e-10	9.4489193983176e-09	14	99843949	99844046	MI0000825

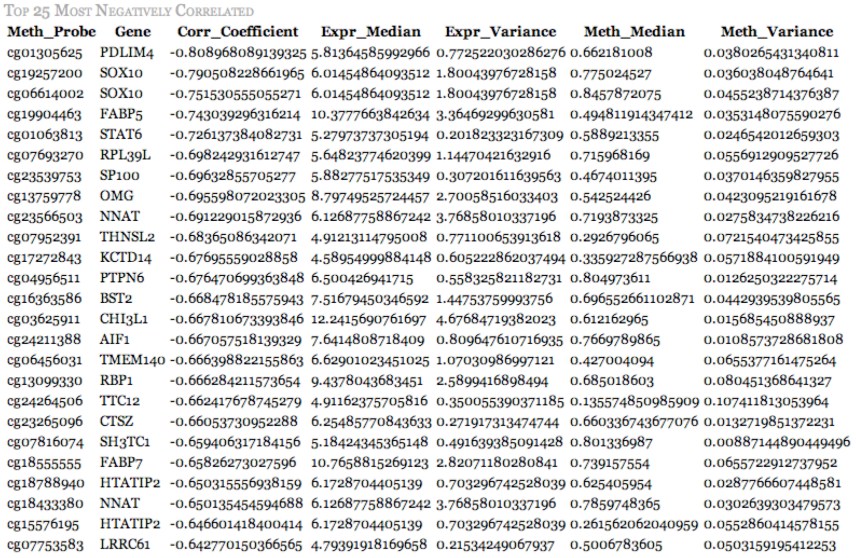
correlation = calculated correlation coefficients between copy number and expression data. A complete list of the calculated correlations is also available

Figure 7: correlate miRNA expression with copy number

mRNA vs. methylation: we calculated the spearman correlation between mRNA and methylation. The result is shown in Figure 8.

Administrator 1/24/11 11:34 AM
Comment: Need to address Clinical assay vs. research asaay.





Top 25 Most Negatively Correlated

Meth_Probe	Gene	Corr_Coefficient	Expr_Median	Expr_Variance	Meth_Median	Meth_Variance
cg01305625	PDLIM4	-0.808968089139325	5.81364585992966	0.772522030286276	0.662181008	0.0380265431340811
cg19257200	SOX10	-0.790508228661965	6.01454864093512	1.80043976728158	0.775024527	0.036038048764641
cg06614002	SOX10	-0.751530555055271	6.01454864093512	1.80043976728158	0.8457872075	0.0455238714376387
cg19904463	FABP5	-0.743039296316214	10.3777663842634	3.36469299630581	0.494811914347412	0.0353148075590276
cg01063813	STAT6	-0.726137384082731	5.27973737305194	0.201823323167309	0.5889213355	0.0246542012659303
cg07693270	RPL39L	-0.698242931612747	5.64823774620399	1.14470421632916	0.715968169	0.0556912909527726
cg23539753	SP100	-0.69632855705277	5.88277517535349	0.307201611639563	0.4674011395	0.0370146359827955
cg13759778	OMG	-0.695598072023305	8.79749525724457	2.70058516033403	0.542524426	0.0423095219161678
cg23566503	NNAT	-0.691229015872936	6.12687758867242	3.76858010337196	0.7193873325	0.0275834738226216
cg07952391	THNSL2	-0.68365086342071	4.91213114795008	0.771100653913618	0.2926796065	0.0721540473425855
cg17272843	KCTD14	-0.67695559028858	4.58954999884148	0.605222862037494	0.335927287566938	0.0571884100591949
cg04956511	PTPN6	-0.676470699363848	6.500426941715	0.558325821182731	0.804973611	0.0126250322275714
cg16363586	BST2	-0.668478185575943	7.51679450346592	1.44753759993756	0.696552661102871	0.0442939539805565
cg03625911	CHI3L1	-0.667810673393846	12.2415690761697	4.67684719382023	0.612162965	0.015685450888937
cg24211388	AIF1	-0.667057518139329	7.6414808718409	0.809647610716935	0.7669789865	0.0108573728681808
cg06456031	TMEM140	-0.666398822155863	6.62901023451025	1.07030986997121	0.427004094	0.0655377161475264
cg13099330	RBP1	-0.666284211573654	9.4378043683451	2.5899416898494	0.685018603	0.080451368641327
cg24264506	TTC12	-0.662417678745279	4.91162375705816	0.350055390371185	0.135574850985909	0.107411813053964
cg23265096	CTSZ	-0.66053730952288	6.25485770843633	0.271917313474744	0.660336743677076	0.0132719851372231
cg07816074	SH3TC1	-0.659406317184156	5.18424345365148	0.491639385091428	0.801336987	0.00887144890449496
cg18555555	FABP7	-0.65826273027596	10.7658815269123	2.82071180280841	0.739157554	0.0655722912737952
cg18788940	HTATIP2	-0.650315556938159	6.1728704405139	0.703296742528039	0.625405954	0.0287766607448581
cg18433380	NNAT	-0.650135454594688	6.12687758867242	3.76858010337196	0.7859748365	0.0302639303479573
cg15576195	HTATIP2	-0.646601418400414	6.1728704405139	0.703296742528039	0.261562062040959	0.0552860414578155
cg07753583	LRRC61	-0.642770150366565	4.79391918169658	0.21534249067937	0.5006783605	0.0503159195412253

Figure 8: correlate expression and methylation

Method

Data description: We included for the analysis high-throughput sequencing data of 169 individuals; mRNA expression data of 440 individuals; miRNA expression data of 415 individuals.

Mutsig: There are 45684 total number of mutations in the 169 input MAF generated at Broad Institute. There are 21836 noncoding mutations after removing 23848 noncoding mutations. There are 21828 mutations after collapsing adjacent/redundant mutations. We removed 178 mutations outside of gene sets; 28 “impossible” mutations in gene-patient-category bins of zero coverage.

GISTIC2: (unpublished)

Consensus clustering using mRNA/miRNA: We performed clustering using the median based integrated expression data generated from Affymetrix HT-HG-U133A genechips, Affymetrix Human Exon 1.0 ST GeneChips, and custom designed Agilent 244k feature Gene Expression Microarrays. If a gene was only assayed on one platform, this measurement was used. If the gene was assayed on two platforms, the average of the two measurements was used; if the gene was assayed on all platforms the median measurement was used. We used the average silhouette width calculation for selecting the robust clusters.

Administrator 1/23/11 9:59 PM
Comment: What would happen if clustering on each platform separately and then combine later?

For clustering analysis of miRNA DATA, we used the mean row subtraction of expression data, we filtered the data to 150 most variable miRNAs. Consensus NMF clustering of 415 samples and 150 miRNAs identified 3 subtypes, with the stability of the clustering increasing for k = 2 to k = 8 and the average silhouette width calculation for selecting the robust clusters.

Reference:

[1] gbm_correlate_expr_clinical_report.pdf
[2] gbm_correlate_expr_cnv_report.pdf
[3] gbm_correlate_expr_methylation_report.pdf
[4] gbm_correlate_miRNA_clinical_report.pdf
[5] gbm_correlate_miRNAcnmfconsensusclustering_clinical_report.pdf
[6] gbm_correlate_miRNAconsensusclustering_clinical_report.pdf
[7] gbm_correlate_miRNAexpr_miRNAcnv_report.pdf
[8] gbm_correlate_mRNAcnmfconsensusclustering_clinical_report.pdf
[9] gbm_correlate_mRNAconsensusclustering_clinical_report.pdf
[10] gbm_correlate_mutation_clinical_report.pdf
[11] gbm_gistic2_report.pdf
[12] gbm_miRNAcnmfconsensusclustering_report.pdf
[13] gbm_miRNAconsensusclustering_report.pdf
[14] gbm_mRNAcnmfconsensusclustering_report.pdf
[15] gbm_mRNAconsensusclustering_report.pdf
[16] gbm_mutsig_report.pdf
[17] gbm_targetmir_report.pdf
[18] gbm_pathwayenrich_report.pdf