Comprehensive Profiling of the Osteosarcoma Genome

Jiayi M Sun[1,2], Jianhe Shen[2,3], Alex Yu[2,3], Joseph Luan[2,3], Horatiu Voicu[2,4], Rudy Guerra[3,5], Chris Man[2,3,4], Ching C Lau[1,2,3,4]

[1]Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, [2]Texas Children's Cancer Center, [3]Department of Pediatrics, Baylor College of Medicine, [4]Dan L. Duncan Cancer Center, [5]Department of Statistics, Rice University

Clinical Background

Osteosarcoma (OS) is the most common type of malignant bone cancer in young adults with approximately 400 cases diagnosed per year. The peak incidence is associated with adolescence with the tumor primarily manifesting in the long bones, such as the distal femur, with preferential metastasis to the lungs.

Site	SEER Incidence[a]					US Mortality[b]					5-Year Relative Survival[a] (%)	
	2005-2009		1975-2009 APC	1975-1992 APC	1993-2009 APC	2005-2009		1975-2009 APC	1975-1992 APC	1993-2009 APC	1975-1977	2002-2008
	# Cases	Rate				# Deaths	Rate					
All Sites	6,537	21.8	0.6*	1.2*	0.9*	3,408	3.2	-1.9*	-2.2*	-1.6*[B]	68.1	83.7[B]
Bone and Joint	462	1.5	0.4	0.8	-0.1	572	0.5	-1.0*	-2.7*	0.1[B]	51.0[B]	66.4[B]
Melanoma of the skin	401	1.3	1.3*	2.4*	1.3	34	0.0	-	-	-	76.2[B]	97.9[B]
Testis(males)	577	3.7	1.4*	2.1	1.8*	36	0.1	-	-	-	66.0[B]	95.3[B]
Brain and Other nervous	636	2.1	0.2	2.1*	0.5	539	0.5	-1.2*	-1.1*	-1.2*	65.3	77.1[B]
Thyroid	640	2.1	1.6*	1.1	2.8*	-	-	-	-	-	100.0	97.7
Hodgkin lymphoma	957	3.2	-0.8*	0.3	-0.6	73	0.1	-4.7*	-3.5*	-5.1*	89.0	96.6[B]
Non-Hodgkin lymphoma	552	1.8	1.7*	2.7*	1.6	230	0.2	-2.8*	-2.2*	-3.1*	47.7[B]	82.8[B]
Leukemia	927	3.1	0.9*	0.5	1.5	1,000	0.9	-2.1*	-1.9*	-2.2*	24.5	69.1[B]
Acute lymphocytic	529	1.8	1.3*	1.6	2.7*	396	0.4	-2.2*	-1.3*	-2.1*	30.0[B]	72.0[B]

Fig. 1. SEER cancer statistics within the 15-19 age group.

Neoadjuvant chemotherapy followed by tumor resection is currently the standard treatment. Prognostic markers such as tumor necrosis % as a response to chemotherapy guide the choice of post-operative chemotherapy. Patients with a good response of >90% tumor necrosis have a 5-year overall survival of ~70%. However, poor responders (<90% tumor necrosis), despite alterations of post-operative chemotherapy to include additional agents or dose-intensified chemotherapy, have a much lower 5-year overall survival of ~40%. Patients with metastasis have an even worse overall survival rate of ~20%.

Genomic Instability

Osteosarcoma has one of the most complex genomes among all cancers. Furthermore, its lower incidence relative to other cancers makes identifying prognostic biomarkers very difficult. Current literature has shown evidence of association between many levels of genomic control, such as miRNA and copy number, with osteosarcoma prognosis however these studies are performed on a small number of patients and specifically focus on certain levels of genomic control. A more comprehensive characterization of the genome will aid us in deducing a relationship between genomic aberrations and tumor biology and allow us to better understand the pathogenesis of osteosarcoma.

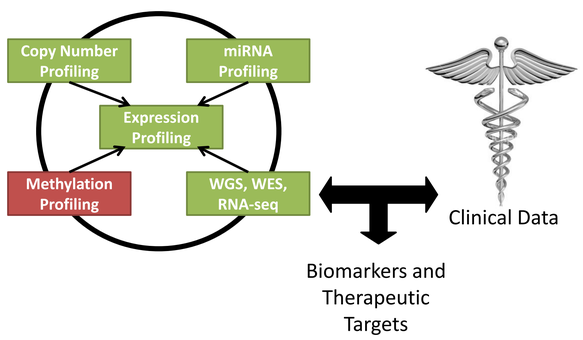
Fig.2: Spectral karyogram of OS (Left) vs normal (Right)[3]

Sample Profile

Profiling	Platform	Number of Samples
Copy Number	Affymetrix 6.0 SNP	86
mRNA	Affymetrix Human Exon 1.0 ST Array	85
microRNA	Life Technologies TLDA Card	85
Methylation	Illumina Infinium 450K Array	83
Whole Genome Sequencing (WGS)		48
Whole Exome Sequencing (WES)		86
RNAseq		86

Samples were collected from the Children's Oncology Group (COG), Texas Children's Cancer Center (TCCC) and other collaborating institutions. All samples have matched clinical outcome data from COG's Statistics and Data Center. All patients in this study have been treated with the same protocol (pre-operative chemotherapy, tumor resection, post-operative therapy). All tissue DNA and RNA have gone through full QC/QA evaluation histopathologically and molecularly.

Method Diagram



DNA Methylation Background

DNA methylation occurs through the action of DNA methylatransferases (DNMT) catalyzing the addition of a methyl group to the C5 position of a CpG dinucleotide. CpG dinucleotides occur with lower than expected frequency in the genome. However in certain regions within gene promoters , there are long regions several hundreds of base pairs in length of repeating CpG dinucleotides, called CpG islands. Methylation of these regions has a well known effect of gene transcriptional silencing.

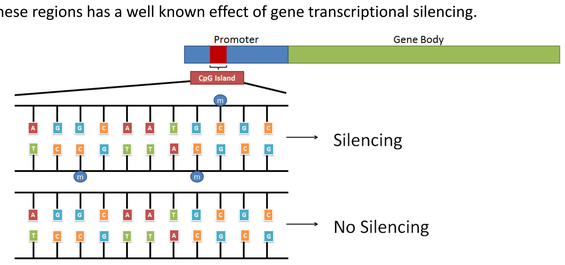


Fig. 3. DNA methylation mechanism.

Human Methylation Array

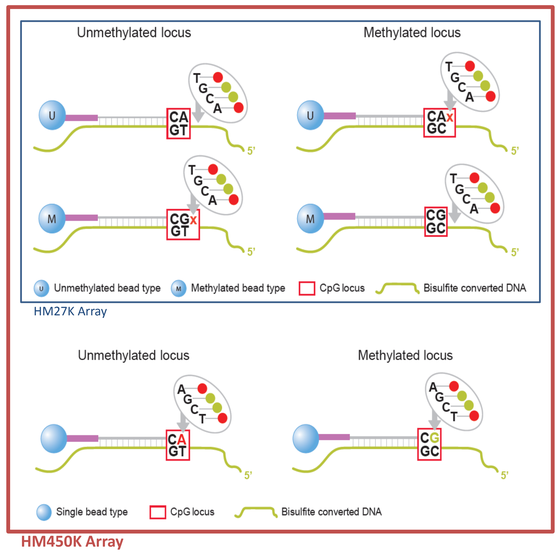


Fig. 4. Probe technologies for Illumina methylation arrays. Each probe of the 450K array is annotated with several different categories of genomic information including where the probe is located within the gene (5', gene body, etc.).

Amount of DNA methylation is called the **beta-value**. This is a measure of the probe intensity. Full methylation corresponds to a beta-value of 1 and no methylation to a beta-value of 0.

$$\text{Beta-value} = \frac{\text{Intensity}_{\text{Methylated}}}{\text{Intensity}_{\text{Methylated}} + \text{Intensity}_{\text{Unmethylated}}}$$

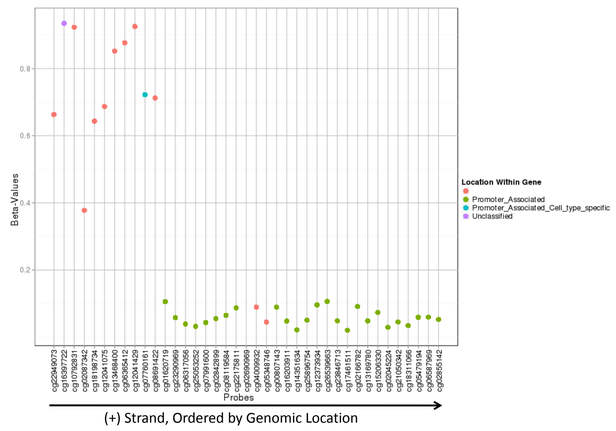


Fig. 5. Probe level beta-value distribution of TP53.

Methylation Analysis

To represent the beta-value of genes in a concise and summarized way, we performed circular binary segmentation of the methylation probes using Partek.

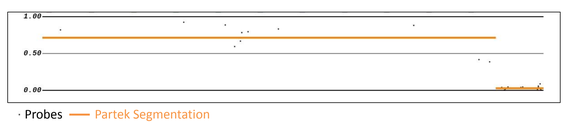


Fig. 6. Segmentation of methylation probe values.

Methylation of CpG islands within promoter regions has a silencing effect on gene transcription through the recruitment of methyl-binding proteins (MBP) to reshape the structure of the DNA into a more compact form, making the sequences less accessible to transcription machinery. To analyze the role of promoter methylation, we extracted curated promoter regions from the Eukaryotic Promoter Database (EPD) and overlap the regions with our segmentation values. This database includes promoter values of ~9700 genes. This database imports information both from ENSEMBL and the Database of Transcriptional Start Sites (DBTSS), an experimentally curated databse for TSSs.

Fig. 7. Overlap of EPD promoter regions on segmentation of methylation probe values.

For EPD regions which did not overlap with a segment, the beta-value of the closest probe annotated with the same gene association was chosen. We preferentially selected probes within CpG islands followed by promoter-associated probes.

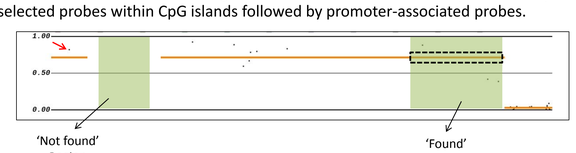


Fig. 8. Example of selecting methylation value for promoters.

This method allowed us to simplify the data to make the analysis more gene-centric to better compare with the other platforms. Additionally, segmentations allows us to identify regions with more consistent beta-value. We performed our analysis with both segment and promoter datasets due to less than half of known genes being included in the EPD.

Methylation Analysis

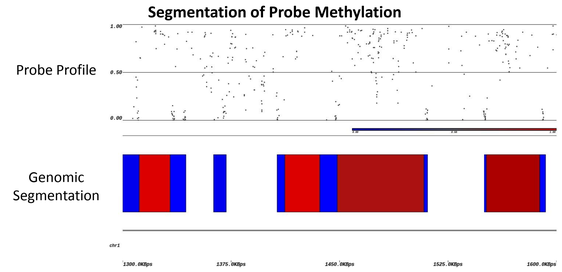


Fig. 9. Example of genomic segmentation in chr1, color is representative of beta-value with blue values closer to 0 and red values closer to 1.

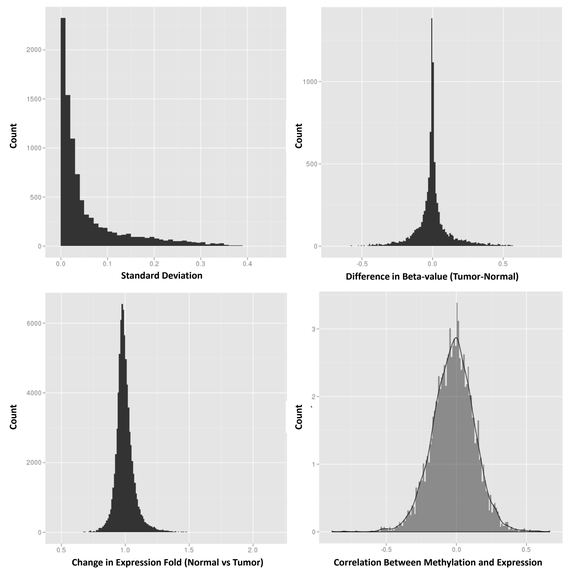


Fig. 10. Profile of promoter segmentation including all tumor and normal samples.

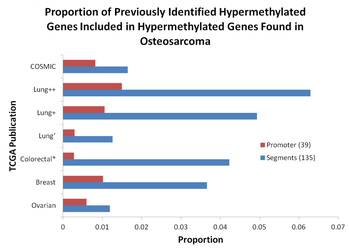
Promoters which lie in multiple categories are more likely to be associated with tumor biology.

Standard Deviation: Highly variant segments indicates varying methylation values which may be more relevant to tumor biology. Hypermethylated (high-methylation) genes leading to tumor-suppressors is one method of tumor pathogenesis.

Difference in Beta-value (Tumor-Normal): Large differences may translate to expression or silencing of a gene that is normally oppositely regulated.

Expression Fold Change: Analysis of matched HuEx expression data between tumor and normal samples detects genes which have altered transcription rates, possibly related to tumor biology.

Correlation Between Methylation and Expression: Negative correlation between beta-values and expression of a promoter region supports current knowledge about the silencing effect of methylation on gene transcription.



We compare genes that were in the promoter regions found in several of the above categories with known databases, including TCGA publications and found high overlap with the COSMIC database.

Cluster Analysis

To identify whether osteosarcoma may include subtypes similar to those seen in other cancers, we used non-negative matrix factorization (NMF) to perform unsupervised clustering on all tumor samples. We were able to identify four clusters.

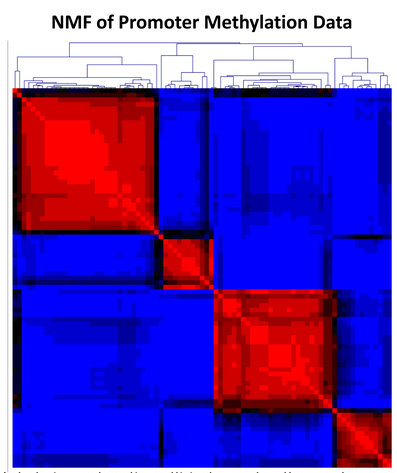


Fig. 11. The best clustering cophenetic coefficient was when the samples were separated into four clusters.

Cluster Analysis

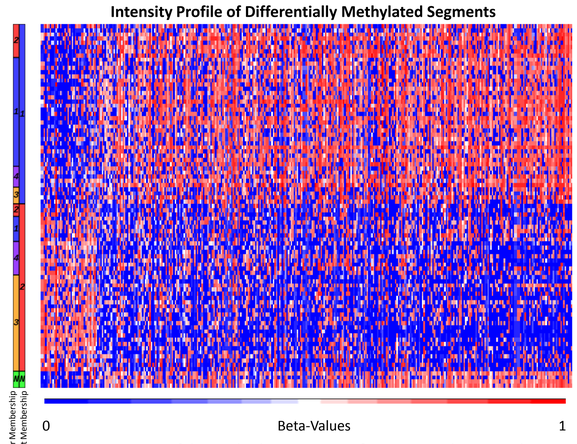


Fig. 12. Comparison of the overlap between NMF of promoter and segment data. A majority of samples in cluster 1 for both promoter and segment memberships overlap. Also, a majority of samples in cluster 3 of the promoter memberships overlaps with cluster 2 in the segment memberships.

Pathway Analysis

A. Group 1 vs Group 2 Significant Pathways (Segment Membership)

#	Maps	-log(pValue)	pValue	Ratio
1	DNA damage: NHEJ mechanisms of DSBs repair		1.500e-5	4/74
2	Cell cycle: Sister chromatid cohesion		2.779e-5	4/72
3	Cell cycle: Transition and termination of DNA replication		7.493e-5	4/28
4	Development: S1P1 receptor signaling via beta-arrestin		1.635e-4	4/34
5	Transcription: Receptor-mediated HIF regulation		2.811e-4	4/39
6	Translation: Regulation of EIF2 activity		2.811e-4	4/39
7	Cytoskeleton remodeling: TGF, WNT and cytoskeletal remodeling		2.909e-4	5/111
8	Development: PIP3 signaling in cardiac myocytes		5.801e-4	4/47
9	Cell cycle: Role of Nek in cell cycle regulation		2.248e-3	3/32
10	DNA damage: DNA-damage-induced responses		2.327e-3	3/3

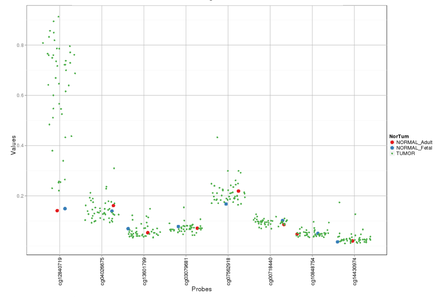
B. Group 3 vs All Other Groups Significant Pathways (Promoter Membership)

#	Maps	-log(pValue)	pValue	Ratio
1	Development: Role of CDK5 in neuronal development		3.395e-5	4/24
2	Immune response: IL-10 signaling pathway		3.789e-4	3/26
3	Development: Thrombopoietin-regulated cell processes		1.921e-3	3/45
4	Immune response: IL-13 signaling via PI3K-ERK		2.603e-3	3/50
5	G-protein signaling: Proinsulin C-peptide signaling		2.912e-3	3/52
6	Cell cycle: Chromosome condensation in prometaphase		5.899e-3	3/71
7	Development: EGFR signaling via PIP3		7.057e-3	3/73
8	Immune response: IL-23 signaling pathway		8.309e-3	3/75
9	Immune response: Antigen presentation by MHC class I		1.036e-2	3/78
10	Possible influence of low doses of Arsenite on glucose uptake in muscle		1.036e-2	3/78

Future Work

Focal Methylation Analysis

Segmentation analysis is beneficial for the purpose of summarizing the 450,000 probes on the methylation array into gene-specific values for better comparison across platforms. There may be areas within the genome where probes are more sparse such as below, where there may be changes related to tumor biology.



Alternative Methods of Epigenetic Analysis

Methylation data has a relatively different distribution from other microarray data, known as a beta-distribution. We are currently working to test the use of different methods such as recursively partituioned mixture models (RPMM) for clustering analysis.

Clinical Data

We are in the processes of combining the clinical with the genomic data to identify clinical significance in focal genes as well as looking at the broad methylation spectrum for clinical features specific to the genomic subgroups we have identified.

Integration of Data

Previously, we have analyzed the other genomic platforms (copy number, expression, and miRNA) with similar methods, looking for both focal differences as well as clustering methods to identify genomic subgroups within the data. We are currently looking at more integrative approaches such as the 'cluster-of-clusters' method which combines the clustering membership from several sources into an overall membership. We are also looking into other methods which search for mutually exclusive genes indicating strong evidence for the importance of that genes in tumor pathogenesis.

Acknowledgement