Bioinformatics Strategies for Identifying Regions of Epigenetic Deregulation Associated with Aberrant Transcript Splicing and RNA-editing

Mia D. Champion[1], Ryan A. Hlady[2], Huihuang Yan[3], Jared Evans[3], Jeff Nie[3], Jeong-Heon Lee[4], James Bogenberger[5], Kannabiran Nandakumar[3], Jaime Davila[3], Raymond Moore[3], Asha Nair[3], Daniel O'Brien[3], Yuan-Xiao Zhu[5], K. Martin Kortum[5], Tamas Ordog[4,6], Zhiguo Zhang[7], Richard W. Joseph[8], A. Keith Stewart[5], Jean-Pierre Kocher[3], Eric Jonasch[9], Keith D. Robertson[2], Raoul Tibes[5] and Thai H. Ho[5]

[1]*Division of Biomedical Statistics and Informatics, Mayo Clinic, Scottsdale, AZ, U.S.A.*
[2]*Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic, Rochester, MN, U.S.A.*
[3]*Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, U.S.A.*
[4]*Translational Program Center for Individualized Medicine, Mayo Clinic, Rochester, MN, U.S.A.*
[5]*Division of Hematology and Oncology, Mayo Clinic, Scottsdale, AZ, U.S.A.*
[6]*Department of Physiology and Biomedical Engineering, Mayo Clinic, Rochester, MN, U.S.A.*
[7]*Department of Biochemistry and Molecular Biology, Mayo Clinic, Rochester, MN, U.S.A.*
[8]*Division of Hematology and Oncology, Mayo Clinic, Jacksonville, FL, U.S.A.*
[9]*Department of Genitourinary Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, U.S.A.*

Keywords: Epigenetic Modification, Splicing, RNA-editing, Co/Post-Transcriptional Processing, Bioinformatics, NMD, Alu Elements, Transcriptome, RNA-seq, Methylation, ChIP-seq, Methyl-CpG, DNAme.

Abstract: Epigenetic modifications are associated with the regulation of co/post-transcriptional processing and differential transcript isoforms are known to be important during cancer progression. It remains unclear how disruptions of chromatin-based modifications contribute to tumorigenesis and how this knowledge can be leveraged to develop more potent treatment strategies that target specific isoforms or other products of the co/post-transcriptional regulation pathway. Rapid developments in all areas of next-generation sequencing (DNA, RNA-seq, ChIP-seq, Methyl-CpG, etc.) have provided new opportunities to develop novel integration and data-mining approaches, and also allows for exciting hypothesis driven bioinformatics and computational studies. Here, we present a program that we developed and summarize the results of applying our methods to analyze datasets from patient matched tumor or normal (T/N) paired samples, as well as cell lines that were either sensitive or resistant (S/R) to treatment with an anti-cancer drug, 5-Azacytidine (http://sourceforge.net/projects/chiprnaseqpro/). We discuss additional options for user-defined approaches and general guidelines for simultaneously analyzing and annotating epigenetic and RNA-seq datasets in order to identify and rank significant regions of epigenetic deregulation associated with aberrant splicing and RNA-editing.

1 INTRODUCTION

Deregulation of epigenetic modifications either mimics the effects of genetic changes, or provides additional heritable alterations that contribute to the development and progression of many cancers (Feinberg *et al.*, 2006). Epigenetic regulation is not dependent upon simple binary states of a single type of modification nor is it a summation of the activities of regulating methyltransferases. Rather, it is dependent upon the interplay of both DNA and histone level modifications that make up complex "combinatorial codes" (Jin *et al.*, 2012; Cieślik and Bekiranov, 2014). Epigenetic modifications have a profound impact on co/post-transcriptional processes, which also have been identified as having

Champion M., Hlady R., Yan H., Evans J., Nie J., Lee J., Bogenberger J., Nandakumar K., Davila J., Moore R., Nair A., O'Brien D., Zhu Y., Kortüm K., Ordog T., Zhang Z., Joseph R., Stewart A., Kocher J., Jonasch E., Robertson K., Tibes R. and H. Ho T..
Bioinformatics Strategies for Identifying Regions of Epigenetic Deregulation Associated with Aberrant Transcript Splicing and RNA-editing.
DOI: 10.5220/0005248001630170
In *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms* (BIOINFORMATICS-2015), pages 163-170
ISBN: 978-989-758-070-3

an important role in cancer progression. For example, modulation of the levels of the histone modification H3K36me3 (trimethylated histone H3 lysine 36) either by overexpression, or by silencing of the SETD2 methyltransferase directly effects alternative splicing of associated exons (Luco *et al.*, 2010; Simon *et al.*, 2014). One of these alternatively spliced mRNAs, FGFR2, commonly undergoes an isoform switch from a "normal IIIb" FGFR2 transcript isoform to a mesenchymal "IIIc" form in ~90% of kidney renal clear cell carcinomas (ccRCC) (Zhao *et al.*, 2013). The mechanisms determining how epigenetic modifications influence differential mRNA splicing are unknown and may involve modified histone protein mediated recruitment of components of the splicing machinery, or on the DNA level, may involve a regulatory role of mobile sequence elements (e.g. Alu) that are known to mediate "exonization" (Makalowski *et al.*, 1994; Ast, 2004). Alu elements are reportedly enriched for DNA-methylation as well as active (e.g. H3K36me3) and repressive associated histone methylation marks (Huda *et al.*, 2010). It is known that in some cases, histone modifications involved in transposable element regulation serve as a "seed region" from which the marks can spread into adjacent genes (Kidwell and Lisch, 2000; Feschotte, 2008). Alu elements are also involved in mediating the post-transcriptional process of RNA-editing and if oriented in opposition, are a favored substrate for the ADAR enzymes (Athanasiadis *et al.*, 2004; Bazak *et al.*, 2014). Computational comparative studies have revealed that >90% of all A->I substitutions occur within Alu elements present in mRNAs (Kim *et al.*, 2004; Levanon *et al.*, 2004; Athanasiadis *et al.*, 2004), and there are over a hundred million sites (Bazak *et al.*, 2014). Although the functional implications of Alu-associated RNA editing are still largely unknown, it is known that these variations influence splice site modification (Rueter *et al.*, 1999), mRNA stability (Wang *et al.*, 2013), and may affect transport. In addition, RNA-editing sites have a known role in regulating cell proliferation during the progression of cancer (Paz *et al.*, 2007; Choudhury *et al.*, 2012; Chen *et al.*, 2013). Deeper understanding regarding how aberrant epigenomic modifications regulate gene expression and co/post-transcriptional changes during the development and progression of many cancers continues to unfold. Here, we present methods packaged as a Python program used for comparative analysis of paired patient matched tumor or normal (T/N) samples, as well as cell lines that were either sensitive or resistant (S/R) to treatment with the anti-cancer drug, 5-Azacytidine in order to: 1) identify regions exhibiting shifts in epigenetic modification peaks between two paired samples, 2) classify indicators of co/post-transcriptional mis-regulation associated with epigenetic deregulation as the abundance of aberrant splicing events and frequency of RNA editing variations within identified regions and, 3) assess the significance of differences for 1) and 2) between paired samples in order to prioritize regions for further clinical studies.

2 METHODS AND RESULTS

2.1 Source of Material for Comparative Studies of Epigenetic Deregulation and Co/Post-Transcriptional Modifications

Comparative epigenetic and transcriptional datasets can come from a number of sources. The two most commonly studied epigenetic modifications involve the binding of proteins (e.g. histones) to DNA and the methylation of cytosine nucleotides (e.g. CpG dinucleotide). Differential library preparations are used to characterize diverse types of histone methylation patterns that are commonly associated with repressive or enhanced states of chromatin and gene expression. Transcriptome datasets are typically generated by either microarray or RNA-seq technologies. RNA sequencing is more suitable for studies focused on assessing the effects of aberrant epigenetic regulation on transcriptional processing since it enables assessment of the presence and abundance of novel transcript isoforms, in addition to known transcripts (Zhao *et al.*, 2014). Read counts also provide a means to calculating more absolute levels of expression and reduce signal-to-noise ratios that are often problematic when assessing hybridization experiments (Zhao *et al.*, 2014). DNA sequencing (either at the genome or exome level) can be used to do comparative filtering from RNA variation datasets when identifying known or novel candidate RNA editing sites. With regard to parameters influencing studies of epigenetic regulation of co/post-transcriptional processes, we discuss below some of the key issues for dataset generation and processing.

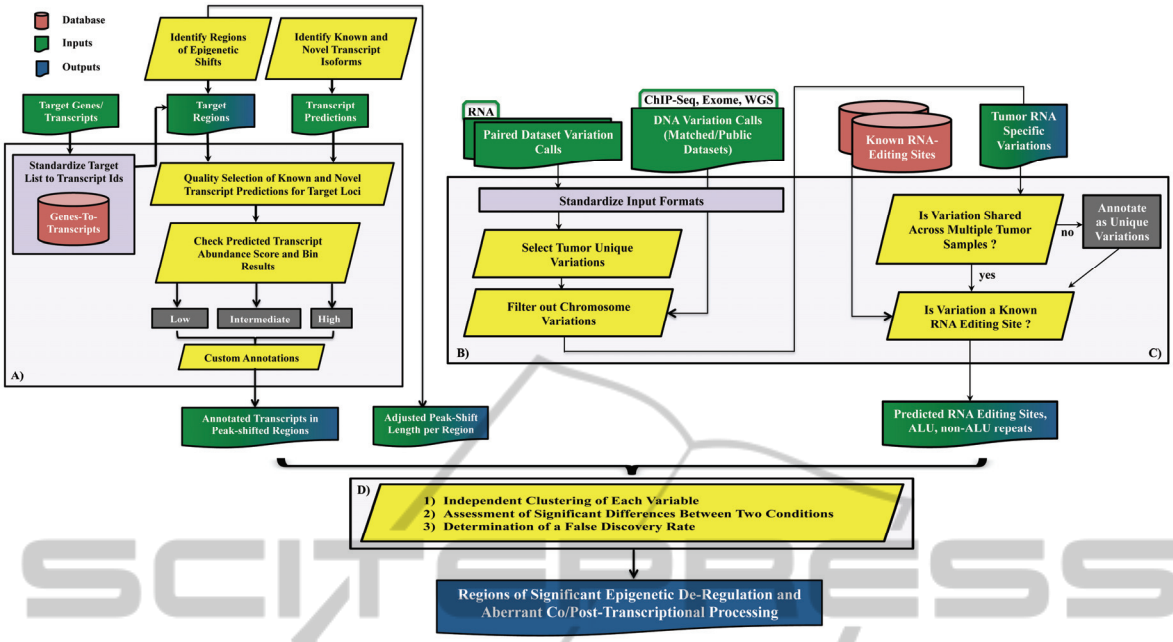

Figure 1: Workflow to expedite identification of diagnostic biomarkers or clinical targets (Champion, 2014).

2.2 Complying with the "Garbage-In-Garbage-Out" Concept

Each additional data source used for broad comparative analysis increases the demand to comply with the Garbage-In-Garbage-Out (GIGO) concept. The percentage and distribution of millions of sequencing reads successfully aligned to a reference genome greatly impacts predictions of peak enrichment (ChIP-seq), differential transcript isoform modeling and abundance (RNA-seq), and sequence variation identification (DNA and RNA-seq). Standardizing thresholds for sequencing quality, number of mismatches allowed per alignment, and placement of reads mapping to multiple genomic locations are steps to introducing consistency that reduces noise when comparing diverse datasets. Along these lines, community established "best practices" optimize each independent workflow, which translates to high quality comparative datasets when interpreting convergence of diverse variables (The Encode Project Consortium, 2011; Landt *et al.*, 2012; The Broad Institute of MIT and Harvard, 2014). Different categories of ChIP-seq tag enrichments, or "peaks", determine which algorithms or parameter settings are best for optimization of true signal prediction. Narrow peaks are characteristic of sequence specific transcription factor binding or RNA polymerase II transcription start site specificity whereas broader peak domains are characteristic of most histone marks that span a nucleosome sized region or larger chromatin domain (Pepke *et al.*, 2014). In our studies (Figure 1A), SICER was used to identify extended domains of ChIP enrichment by adjusting the window scan with gaps allowed parameter to the recommended size of approximately one nucleosome and linker (~200 bps) (Zang *et al.*, 2009). Number of read pairs from each peak region in IP and that from the corresponding region in input was normalized to a library size of 10 million (FPTM) and the input-subtracted FPTM values were used for differential binding analysis. False Discovery Rates (FDRs) were determined from poisson p-values and enrichment predictions were further filtered according to a threshold cutoff (FDR<0.01). For our genome-wide DNA methylation studies, we used the Infinium Human Methylation450 BeadChip (Illumina, 2014) and normalized results using subset quantile within-array normalization (SWAN) (Maksimovic *et al.*, 2012). Since there are known gender specific epigenetic modifications, many studies typically remove all X/Y associated data points in order to analyze them separately. However, we recommend doing this as a final step in the analysis process since the inclusion of these data points allow for a more accurate p-value adjustment for multiple testing. Comparisons of epigenetic deregulation and co/post-transcriptional processes are at the gene level, such that the total region of

epigenetic modification affecting identified transcript isoforms, sequence elements, and variations is a summation of all significant peaks within the ORF (Figure 1A).

Differences in alignment methods are also fundamental to ensuring "best practices" of variant calling in DNA versus RNA sequencing datasets (The Broad Institute of MIT and Harvard, 2014). Correct processing of RNA splice junctions, avoidance of using soft-clipped bases, and specialized confidence thresholds minimizes false positive or negative variation calls. In addition, RNA-seq library preparation protocols for creating cDNA from RNA commonly use random hexamers for the priming step, thus increasing the likelihood of errors in the terminal 6bp of the read. A combination of existing GATK parameters (e.g. FisherStrandFilter, ReadPosRankSumTest (The Broad Institute of MIT and Harvard, 2014)) provide best practice filtering approaches instead of customized methods that likely remove true positive RNA editing sites (RVboost (Wang *et al.*, 2014)). Aggressive removal of shared sequence variations between RNA and comparative DNA datasets is a first step to identifying known and novel RNA editing sites (Figure 1B). In addition to variations identified using DNA sequencing generated from same individual/cell line collections, DNA SNPs were also identified from public population databases (1000 Genomes, HapMap, dbSNP, BGI200/Danish, ESP6500 European and African datasets) (dbSNP, 2014; HapMap, 2010; ESP6500, 2014; 1000 Genomes, 2014). Mapping identified RNA sequencing variations to existing or customized RNA editing databases and resources expedites identification of known RNA editing sites (Champion, 2014; Ramaswami and Li, 2014; Kiran and Baranov, 2010; Li *et al.*, 2009) (Figure 1C). Computational prioritization of candidate novel RNA editing sites includes evaluation of the proximity of identified RNA sequencing variations to splice sites or paired Alu elements in opposite orientation.

RNA-seq alignment methods, such as Tophat (Trapnell *et al.*, 2010), have been developed to handle mapping of reads spanning exon-exon splice junctions (Pepke *et al.*, 2014). Cufflinks uses a bipartite graphing method to assess a "minimum-cost-maximum-matching" of bundled fragments from Tophat read alignments in order to build a parsimonious set of transcript models (Trapnell *et al.*, 2010). Cufflinks then also provides an estimation of expression levels using established methods (Li *et al.*, 2010; Jiang and Wong, 2009). We also binned predicted transcript isoforms according to their scored minor FPKM/major FPKM ratio in order to identify and compare differences in transcript isoform abundances (Figure 1A). Unlike other available algorithms, cufflinks also allows for the identification of novel as well as known transcripts, and exhibits superior estimates of accuracy as measured by the median value of relative errors in percentage across all genes when compared to other methods (Nicolae *et al.*, 2011). Novel transcript predictions are essential for studies of aberrant co/post-transcriptional processes. Identifying novel transcripts associated with epigenetic deregulation is useful for the discovery of "isoform-switching" events that are associated with drug resistance or cancer progression, similar to the mesenchymal "IIIc" FGFR2 isoform abundant in ccRCC. Correlating the abundance of novel transcripts identified with regions of epigenetic deregulation is useful for assessing levels of aberrant transcription (e.g. "transcriptional-noise"), and exploring possible regulatory roles of the Nonsense-Mediated-Decay (NMD) pathway during cancer progression (Gardner, 2011; Frischmeyer and Dietz, 1999). In addition to evidence that epigenetic deregulation mediates aberrant splicing, these processes also affect the rate of transcription (Veloso *et al.*, 2014; Eswaran *et al.*, 2013); although, the functional implications of differences in transcript isoform abundances with regards to tumorigenesis or drug response are largely unknown. Differences in "methodological-flow" can also be used to expedite specific outputs from these types of studies. For example, transcriptome-profiling studies aimed to characterize global regulatory shifts in transcript splicing or abundance would be better done using a transcriptome-to-peak analysis workflow. Conversely, identification of diagnostic biomarkers or potential clinical targets is expedited by starting with epigenetic deregulated regions (Figure 1A) in order to identify aberrant target transcript isoforms, or RNA-editing site variations associated with phenotype progression.

2.3 Bioinformatics Methods to Integrate Epigenetic and Co/Post-Transcriptional Datasets

Historically, methods that integrate epigenetic and microarray expression datasets were developed to characterize transcriptional regulons. There are many available tools and methods available for comparative analysis via correlative clustering of significant shifts in epigenetic modification patterns

with differential gene expression (e.g. Rcade, TransView (Bioconductor Software Packages, 2014)). However, unique to our approach is the inclusion of additional variables of co/post-transcriptional mis-regulation associated with changes in epigenetic modifications. Although important biological relationships between the variables used to evaluate epigenetic influences on transcriptional processes exist, we find that pairwise Spearman correlations (R Development Core Team, 2011) between most of the variables assessed in our studies are not significant (Figure 2, Frequency of non-Alu (A.) and Alu (B.) repeats, C. Frequency of RNA editing sites, D. Adjusted Peak Length Shift, E. ORF size, F. Frequency of novel transcript isoforms). An exception is the expected positive correlation between the distribution of Alu elements (B.) and RNA editing sites (C.) within a given region (Figure 2).

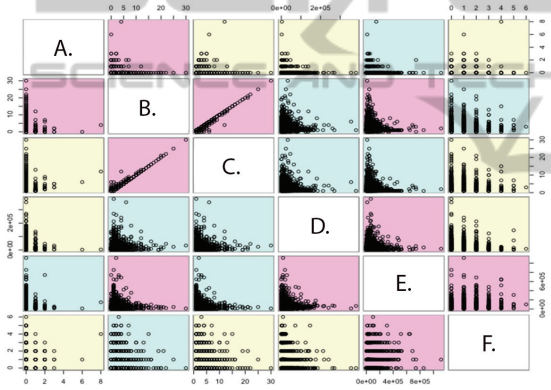

Figure 2: RNA-editing sites and *Alu*-elements are the only two variables that are significantly correlated in regions of epigenetic modification shifts.

Therefore, application of any multi-variable co-clustering approaches would likely be inadequate and unnecessarily exhaust computational resources. Rather, we find that first clustering independent variables followed by a comparative assessment of significant differences between two conditions (e.g. T/N, R/S) across a given region using the Student unpaired t-Test followed by estimation of FDR using a beta-uniform mixture model (R Development Core Team, 2011), provides biologically meaningful results and is useful for ranking the identified regions of aberrant epigenetic modification and co/post-transcriptional deregulation (Figure 3). Finally, functional clustering of candidate clinical targets identified by our comparative studies into predicted interaction networks and biological pathways identified several genes regulating cell

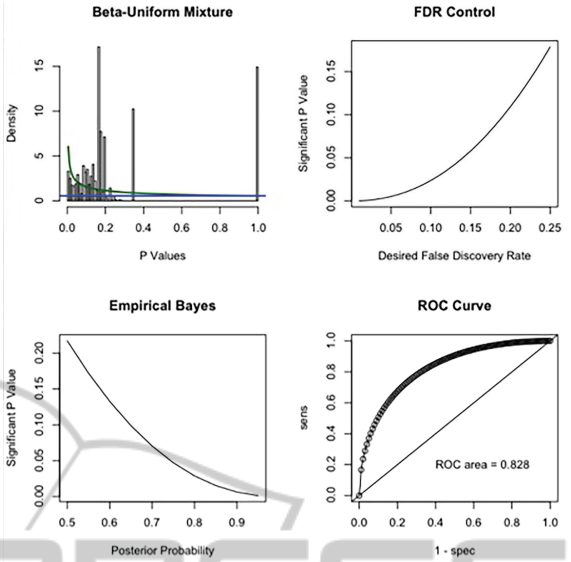

Figure 3: P-values equal to or less than 0.01 are within the range of desired false discovery rate as estimated by a beta-uniform mixture model. Significant p-values were used to select regions for further functional analysis using network interaction and biological pathway prediction algorithms.

cycle progression and mRNA processing, which further supports the validity of our methodologies and provides an additional level of prioritization for future diagnostic biomarker or clinical target development (Table 1).

3 CONCLUSIONS AND PERSPECTIVES

Advanced sequencing techniques and well-considered bioinformatics methods provide unprecedented opportunities for in depth comparative studies of paired (T/N or R/S) datasets in order to understand the regulatory roles of epigenetic modifications on co/post-transcriptional processes, and how deregulation of these functional relationships promotes drug resistance and contributes to the progression of cancer. Our studies using a developed program to identify regions exhibiting significant epigenetic modification changes and aberrant co/post-transcriptional processing exemplifies one of the workflows we presented and provides evidence that studies such as these yield meaningful results of potential high impact for subsequent diagnostic biomarker or therapeutic target design endeavors.

Table 1. Genes exhibiting significant epigenetic modification shifts associated with aberrant co/post-transcriptional processing in a 5-Azacytidine resistant human erythroleukemia cell line, but not in a myeloid progenitor cell line, cluster into predicted interaction networks and functional biological pathways.

Biological Pathway	p-value	FDR	Frequency of Genes in Pathway
Mitotic Metaphase and Anaphase	0	<1.00E-03	10
Mitotic Prometaphase	0	<5.00E-04	8
Processing of Capped Intron-Containing Pre-mRNA	0	3.33E-04	8
RNA transport	0.0001	1.48E-02	7
PLK1 signaling events	0.0003	2.06E-02	4
IL6-mediated signaling events	0.0003	2.22E-02	4
Deadenylation-dependent mRNA decay	0.0004	2.24E-02	4
mRNA surveillance pathway	0.0004	2.15E-02	5
Role of Calcineurin-dependent NFAT signaling in lymphocytes	0.0005	1.97E-02	4
Regulation of retinoblastoma protein	0.001	3.95E-02	4
IL2 signaling events mediated by STAT5	0.001	3.87E-02	3
Mitotic G2-G2/M phases	0.0013	4.66E-02	5
IFN-alpha signaling pathway	0.0015	4.74E-02	2
EPO signaling pathway	0.0016	4.91E-02	3

ACKNOWLEDGEMENTS

We would like to thank Ian Davis, W. Kimryn Rathmell, Kathryn E. Hacker and Jeremy M. Simon for assistance with genotyping of tissue. We would like to thank Amylou Dueck for advice regarding statistical analysis of preliminary studies.

The results published here are in whole or part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov.

FINANCIAL SUPPORT

T.H.H. is supported by funding from the ASCO Young Investigator Award from the Kidney Cancer Association, the Action to Cure Kidney Cancer Organization, the MD Anderson Hematology-Oncology Fellowship, a Mayo Clinic CR5 grant, Mayo Clinic Center for Individualized Medicine Epigenomics Translational Program and a Kathryn H. and Roger Penske Career Development Award to Support Medical Research. This work is supported in part by the Mayo Clinic Center for Individualized Medicine Epigenomics Translational Program

REFERENCES

1000Genomes. Accessed November, 2014. Available from: http://www.1000genomes.org/data#DataAccess.

Ast, G., 2004. How did alternative splicing evolve?, *Nat Rev Genet*, vol. 5, no. 10, pp. 773-782.

Athanasiadis, A, Rich, A & Maas, S., 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. , *PLoS Biol*, vol. 2, p. e391.

Bazak, L, Haviv, A, Barak, M, Jacob-Hirsch, J, Deng, P, Zhang, R, Isaacs, FJ, Rechavi, G, Li, JB, Eisenberg, E & Levanon, EY., 2014. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes, *Genome Res.*, vol. 24, pp. 365-376.

Bioconductor Software Packages. Accessed November, 2014. Available from : <master.bioconductor.org/packages/release/bioc.

Champion, MD., Accessed November, 2014. *ChIP-RNA-seqPRO: A strategy for identifying regions of epigenetic deregulation associated with aberrant transcript splicing and RNA-editing sites.* Available from: <http://sourceforge.net/projects/chiprnaseqpro/>.

Chen, L, Li, Y, Lin, CH, Chan, TH, Chow, RK, Song, Y, Liu, M, Yuan, YF, Fu, L, Kong, KL, Qi, L, Li, Y & Zhang N, TA, Kwong DL, Man K, Lo CM, Lok S, Tenen DG, Guan XY., 2013. Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma., *Nature Medicine*, vol. 19, no. 2, pp. 209-16.

Choudhury, Y, Tay, FC, Lam, DH, Sandanaraj, E, Tang, C, Ang, BT & Wang, S., 2012. Attenuated adenosine-to-inosine editing of microRNA-376a* promotes invasiveness of glioblastoma cells., *J Clin Invest*, vol. 122, no. 11, pp. 4059-76.

Cieślik, M & Bekiranov, S., 2014. Combinatorial epigenetic patterns as quantitative predictors of chromatin biology., *BMC Genomics*, vol. 15, p. 76.

The ENCODE Project Consortium, 2011. A users guide to the encyclopedia of DNA elements (ENCODE)., *PLoS Biol*, vol. 9, no. 4, p. e1001046.

dbSNP. *version 137*. Accessed November, 2014. Available from: <http://www.ncbi.nlm.nih.gov/snp/>

ENCODE. *DataStandards*. Available from: <https://genome.ucsc.edu/encode/protocols/dataStandards/>.

ESP6500. Accessed November, 2014. Available from: <evs.gs.washington.edu/EVS/>.

Eswaran, J, Horvath, A, Godbole, S, Reddy, SD, Mudvari, P, Ohshiro, K, Cyanam, D, Nair, S , Fuqua, SAW, Polyak, K, Florea, LD, Kumar, R., 2013. RNA sequencing of cancer reveals novel splicing alterations, *Scientific Reports*, vol. 3, p. 1689.

Feinberg, AP, Ohlsson, R & Henikoff, S., 2006. The epigenetic progenitor origin of human cancer, *Nature*, vol. 7, pp. 21-33.

Feschotte, C., 2008. Transposable elements and the evolution of regulatory networks., *Nat Rev Genet*, vol. 9, pp. 397-405.

Frischmeyer, PA & Dietz, HC., 1999. Nonsense-mediated mRNA decay in health and disease., *Hum Mol Genet*, vol. 8, no. 10, pp. 1893-900.

Gardner, LB., 2011. Nonsense mediated RNA decay regulation by cellular stress; implications for tumorigenesis, *Mol Cancer Res*, vol. 8, no. 3, pp. 295-308.

HapMap. Accessed November, 2014. Available from: <http://hapmap.ncbi.nlm.nih.gov/>

Huda, A, Mariño-Ramírez, L & Jordan, IK., 2010. Epigenetic histone modifications of human transposable elements: genome defense versus exaptation., *Mob DNA.* , vol. 1, no. 1.

Illumina. *Infinium Human Methylation450 Bead Chip*. Accessed November, 2014. Available from: <http://res.illumina.com/documents/products/datasheets/datasheet_humanmethylation450.pdf>.

The Broad Institute of MIT and Harvard. *GATK Best Practices*. Available from: <https://www.broadinstitute.org/gatk/guide/best-practices>

Jiang, H & Wong, WH., 2009. Statistical inferences for isoform expression in RNA-Seq, *Bioinformatics*, vol. 25, no. 8, pp. 1026-1032.

Jin, B, Ernst, J, Tiedemann, RL, Xu, H, Sureshchandra, S, Kellis, M, Dalton, S, Liu, C, Choi, JH & Robertson, KD., 2012. Linking DNA methyltransferases to epigenetic marks and nucleosome structure genome-wide in human tumor cells., *Cell Rep*, vol. 2, no. 5, pp. 1411-24.

Kidwell, MG & Lisch, DR., 2000. Transposable elements and host genome evolution, *Trends Ecol Evol*, vol. 15, pp. 95-99.

Kim, DD, Kim, TT, Walsh, T, Kobayashi, Y, Matise, TC, Buyske, S & Gabriel, A., 2004. Widespread RNA editing of embedded alu elements in the human transcriptome., *Genome Res.*, vol. 14, no. 9, pp. 1719-25.

Kiran, A & Baranov, PV., 2010. DARNED: a DAtabase of RNa EDiting in humans., *Bioinformatics*, vol. 26, no. 14, pp. 1772-6.

Landt, SG, Marinov, GK, Kundaje, A, Kheradpour, P, Pauli, F, Batzoglou, S, Bernstein, BE, Bickel, P, Brown, JB & Cayting P, CY, DeSalvo G, Epstein C, Fisher-Aylor KI, Euskirchen G, Gerstein M, Gertz J, Hartemink AJ, Hoffman MM, Iyer VR, Jung YL, Karmakar S, Kellis M, Kharchenko PV, Li Q, Liu T, Liu XS, Ma L, Milosavljevic A, Myers RM, Park PJ, Pazin MJ, Perry MD, Raha D, Reddy TE, Rozowsky J, Shoresh N, Sidow A, Slattery M, Stamatoyannopoulos JA, Tolstorukov MY, White KP, Xi S, Farnham PJ, Lieb JD, Wold BJ, Snyder M., 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia., *Genome Res.*, vol. 22, no. 9, pp. 1813-31.

Levanon, E, Eisenberg, Y, Yelin, E, Nemzer, R, Hallengger, M, Shemesh, R, Fligelman, ZY, Shoshan, A, Pollock, SR & Sztybel, D., 2004. Systematic identification of abundant A-to-I editing sites in the human transcriptome, *Nat Biotechnol*, vol. 22, pp. 1001-1005.

Li, B, Ruotti, V, Stewart, RM, Thomson, JA & Dewey, CN., 2010. RNA-Seq gene expression estimation with read mapping uncertainty, *Bioinformatics*, vol. 26, no. 4, pp. 493-500.

Li, JB, Levanon, EY, Yoon, JK, Aach, J, Xie, B, Leproust, E, Zhang, K, Gao, Y & Church, GM., 2009. Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing., *Science*, vol. 324, no. 5931, pp. 1210-3.

Luco, RF, Pan, Q, Tominaga, K, Blencowe, BJ, Pereira-Smith, OM & Misteli, T., 2010. Regulation of alternative splicing by histone modifications., *Science*, vol. 327, no. 5968, pp. 996-1000.

Makalowski, W, Mitchell, GA & Labuda, D., 1994. Alu sequences in the coding regions of mRNA: a source of protein

variability. , *Trends Genet*, vol. 10, pp. 188-193.

Maksimovic, J, Gordon, L & Oshlack, A., 2012. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips., *Genome Biol*, vol. 13, no. 6, p. R44.

Nicolae, M, Mangul, S, Măndoiu, II & Zelikovsky, A., 2011. Estimation of alternative splicing isoform frequencies from RNA-Seq data., *Algorithms Mol Biol*, vol. 6, no. 1, p. 9.

Paz, N, Levanon, EY, Amariglio, N, Heimberger, AB, Ram, Z, Constantini, S, Barbash, ZS, Adamsky, K, Safran, M, Hirschberg, A, Krupsky, M, Ben-Dov, I, Cazacu, S, Mikkelsen, T, Brodie, C, Eisenberg, E & Rechavi, G., 2007. Altered adenosine-to-inosine RNA editing in human cancer., *Genome Res.*, vol. 17, no. 11, pp. 1586-95.

Pepke, S, Wold, BJ & Mortazavi, A., 2014. Computation for ChIP-seq and RNA-seq studies, *Nat Methods*, vol. 6, no. 11, pp. S22-S32.

Ramaswami, G & Li, JB., 2014. RADAR: a rigorously annotated database of A-to-I RNA editing, *Nucleic*

Acids Res. , vol. 42, no. (Database Issue), pp. D109-13.

Rueter, SM, Dawson, TR & Emeson, RB., 1999. Regulation of alternative splicing by RNA editing, *Nature*, vol. 399, pp. 75-80.

Simon, JM, Hacker, KE, Singh, D, Brannon, AR, Parker, JS, Weiser, M, Ho, TH, Kuan, PF, Jonasch, E, Furey, TS, Prins, JF, J.D., L, Rathmell, WK & Davis, IJ., 2014. Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects., *Genome Res.*, vol. 24, no. 2, pp. 241-50.

R Development Core Team. *R: A language and environment for statistical computing*. Available from: <http://www.r-project.org>

Trapnell, C, Williams, BA, Pertea, G, Mortazavi, A, Kwan, G, van Baren, MJ, Salzberg, SL, Wold, BJ & Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation., *Nat Biotechnol*, vol. 28, no. 5, pp. 511-5.

Veloso, A, Kirkconnell, KS, Magnuson, B, Biewen, B, Paulsen, MT, Wilson, TE & Ljungman, M., 2014. Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications, *Genome Res.*, vol. 24, no. 6, pp. 896-905.

Wang, C, Davila, JI, Baheti, S, Bhagwate, AV, Wang, X, Kocher, JP, Slager, SL, Feldman, AL, Novak, AJ, Cerhan, JR, Thompson, EA & Asmann, YW., 2014. RVboost: RNA-Seq variants prioritization using a boosting method., *Bioinformatics*.

Wang, IX, So, E, Devlin, JL, Zhao, Y, Wu, M & Cheung, VG., 2013. ADAR regulates RNA editing, transcript stability, and gene expression., *Cell Rep*, vol. 5, no. 3, pp. 849-60.

Zang, C, Schones, DE, Zeng, C, Cui, K, Zhao, K & Peng, W., 2009. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data, *Bioinformatics*, vol. 25, pp. 1952-1958.

Zhao, Q, Caballero, OL, Davis, ID, Jonasch, E, Tamboli, P, Yung, WK, Weinstein, JN & Yao, J., 2013. Tumor-specific isoform switch of the fibroblast growth factor receptor 2 underlies the mesenchymal and malignant phenotypes of clear cell renal cell carcinomas., *Clin Cancer Res.*, vol. 19, no. 9, pp. 2460-72.

Zhao, S, Fung-Leung, W-P, Bittner, A, Ngo, K & Liu, X., 2014. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells, *PLoS One*, vol. 9, no. 1, p. e78644.