Proc. of the Second Intl. Conf. on Advances in Bio-Informatics and Environmental Engineering - ICABEE 2015

ISBN: 978-1-63248-043-9 doi: 10.15224/ 978-1-63248-043-9-127

# SnvDMiR: Associating the genomic proximity of genetic variants with deregulated miRNAs and differentially methylated regions

Zapp A[1], Helms V[1], and Hamed M[1, *]

***Abstract*—Although next generation sequencing of diseased traits has unraveled thousands of DNA alterations, the functional relevance of most of these mutations and how they relate to other epigenetic mechanisms are still poorly understood. Here, we present SnvDMiR as a freely–available R pipeline that conducts combinatorial proximity analysis between disease–associated SNVs, deregulated miRNAs, and differentially methylated regions (DMRs) to identify genomically adjacent SNV-miRNA pairs as well as SNV-DMR pairs. These variants could be further investigated as putative candidates for driving pathogenic processes in diseases. We demonstrated the usefulness of the SnvDMiR pipeline by applying it on a published set of breast cancer-related mutations, deregulated miRNAs, and DMRs. Our pipeline characterized potential driver mutations that are predicted to have damaging effects on related protein functions. Availability: http://gepard.bioinformatik.uni-saarland.de/software**

***Keywords*—Single Nucleotide Variant (SNV), somatic mutation, deregulated miRNA, differentially methylated region (DMR), genomic proximity, epigenetics, data integration, breast cancer.**

## I. Introduction

To further our understanding of human oncogenesis, high-throughput sequencing of tumor genomes has uncovered thousands of DNA alterations such as somatic mutations of single nucleotide variants (SNVs) that may be important for tumor initiation or progression (1-7). Nevertheless, it remains a pressing challenge to determine which mutations are key drivers for tumor pathophysiology and which ones are passengers with no functional effects. To address this need, several approaches have been presented to characterize driver missense mutations (4, 8-10). Most straightforward is the annotation of non-synonymous mutations in oncogenes or tumor suppressors. In contrast, relatively little attention has been paid to cases where driver mutations could be in close genomic proximity to disease-related genes, miRNAs, or methylated CpG sites.

DNA methylation is an epigenetic mechanism that is being increasingly recognized to play an important role in the regulation of gene expression and is used as epigenetic marker for different disease pathways (11-14). DNA methylation typically occurs in a CpG dinucleotide context that is often grouped in clusters called CpG islands. DNA methylation profiling unravels differentially methylated regions (DMRs) that are in principle CpG sites altered during disease or oncogenic processes (15). Hypermethylation of CpG islands located in promoter regions, for example, is involved in gene silencing at the transcriptional level (16) and often leads to a high rate of C to T mutations at these sites (17).

MicroRNAs (miRNAs) are small, non-coding RNAs that function as post-transcriptional regulators of mRNA expression. A miRNA can target a plethora of mRNAs, creating a post-transcriptional regulatory network (18) that has a critical role not only in cellular functions (19) but also in pathological processes (20) especially in human cancerogenesis (18, 21-23). A considerable amount of literature has been published on miRNA-related mutations and on the impact of somatic mutations on miRNA functions. These studies have reported that genetic variants within miRNAs or their target sites can alter miRNA function in cancers (24-28) and have been associated with cancer risk, treatment efficacy and patient prognosis (24), as well as genomic phenotypes (29).

The recent availability of disease-related genomic data such as somatic mutations, associated DMRs and miRNAs calls for the development of integrative genomic proximity-based approaches to better understand the functional relevance of most of these mutations and how they relate to epigenetic marks. To this end, this study presents SnvDMiR, a freely–available R pipeline that is able to conduct combinatorial proximity analysis between disease–associated SNVs, deregulated miRNAs, and DMRs to identify genomically adjacent SNV-miRNA pairs as well as SNV-DMR pairs. We have demonstrated these features on breast cancer-related datasets [Hamed et al. 2015, accepted in BMC Genomics] and we review these here as an example to confirm the functionality of our tool. The matched SNVs suggested putative driver mutations that could play a critical role in breast cancerogenesis.

## II. Implementation

SnvDMiR is a computational pipeline implemented in R. (Fig. 1). Based on lists of genomic variants, deregulated miRNAs, differentially methylated sites, and user defined parameters (configurations), SnvDMiR investigates whether the significantly deregulated miRNAs and differentially methylated sites are in close genomic vicinity to the provided

Proc. of the Second Intl. Conf. on Advances in Bio-Informatics and Environmental Engineering - ICABEE 2015

ISBN: 978-1-63248-043-9 doi: 10.15224/ 978-1-63248-043-9-127

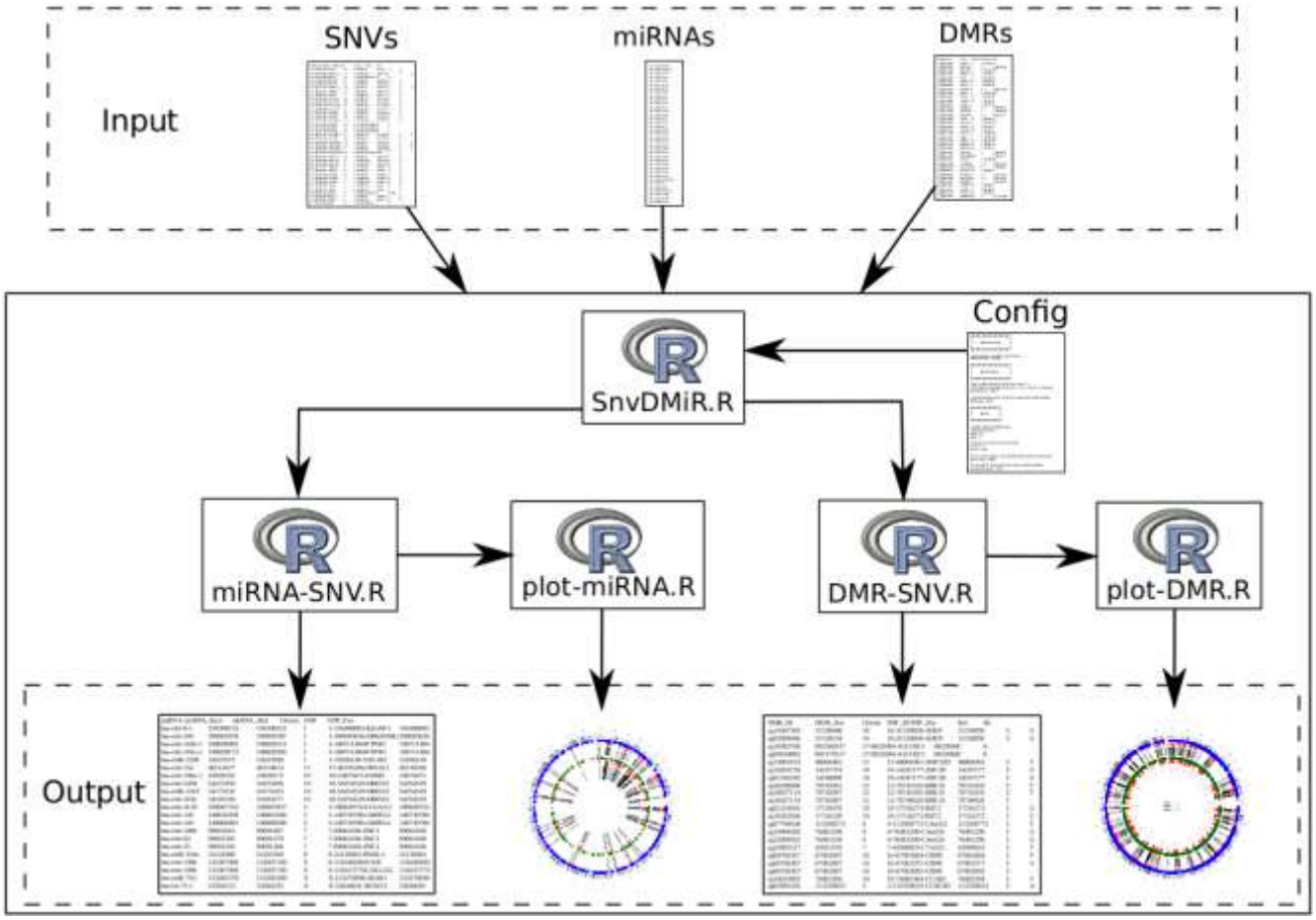


**Figure 1. The data model of the SnvDMiR pipeline**. A schematic diagram describing data processing and integration of genetic variants, deregulated miRNAs and DMRs.

genomic variants and outputs matching entries in tabular and ideogram plots. The user needs only to run the main script *SnvDMiR.R* which in turn loads the required libraries/packages, carries out the analysis on the input data, and visualizes the matched entries in genomic ideograms with circular layouts.

For matching miRNAs and somatic variants, the genomic coordinates of the significantly deregulated miRNAs were downloaded from miRBase (30). Then, SnvDMiR searches for the miRNA sequences in a predefined genomic window (default is 250kb (31)) around each somatic variant. The window size can be set in the configuration file attached with the SnvDMiR script. The matched miRNA-SNV pairs, where the miRNAs occur within the window around the SNV location, are extracted into the som-miRNA-matches.txt file in the output folder.

The second part of the SnvDMiR functionality is to explore whether differentially methylated regions (usually CpG islands) are in the vicinity of somatic mutations. To this end, our tool tests the occurrence of the SNV within a certain genomic distance (default is 3kb) from the genomic coordinates of the differentially methylated sites. The default setting of the predefined distance in the configuration file (3kb) was based on the maximum considered length of typical CpG islands, that is, 500bp (17) ≤ CpG islands ≤ 3kb (32). Moreover, the user has the option to investigate only the C->A, C->G, and C->T SNVs instead of all mutations via setting the parameter *filter_mutations* in the configuration file. The matched entries are also exported to som-DMR-matches.txt file in the output folder.

Finally, SnvDMiR utilizes the circlize R package (33) to efficiently plot the related ideogram and flexibly visualize the matched entries in a circular layout as well as the entire input data (all SNVs and either all miRNAs or all DMRs) as genomic background. This helps to better understand the genomic patterns behind the matched entries. Documentation demonstrating the use of the SnvDMiR tool, sample input files, and the user manual are provided with the tool. We encourage users to adopt the output to their needs using self-defined parameters.

Alexander Zapp [1], Volkhard Helms[1], and Mohamed Hamed[1].
[1] Center for Bioinformatics, Saarland University
66041 Saarbrucken, Germany

Proc. of the Second Intl. Conf. on Advances in Bio-Informatics and Environmental Engineering - ICABEE 2015

ISBN: 978-1-63248-043-9 doi: 10.15224/ 978-1-63248-043-9-127

# III. Case study

## A. *Application on breast cancer*

In a recent work on breast cancer (Hamed et al. 2015), we processed DNA methylation, miRNA expression, and somatic mutation datasets for 131 tumor samples and 20 control samples of healthy tissues downloaded from the TCGA portal (34). The differential analysis of the DNA promoter methylation and miRNA expression data determined 2623 differentially methylated gene promoters and 121 differentially expressed miRNAs, respectively. In order to scrutinize the functional relevance of the somatic mutations and how they relate to other epigenetic mechanisms (such as DNA methylation and deregulation of miRNAs), we applied the SnvDMiR pipeline to these somatic mutations, the deregulated miRNAs, and the differentially methylated regions.

SnvDMiR tested whether the significantly differentially expressed miRNAs are in genomic vicinity to the respective somatic variants by assuming that deregulation of miRNA expression due to carcinogenesis may somehow be related to the associated nearby somatic variants. We searched for the coding sequences of the deregulated miRNAs in a genomic window of 250 kb around the somatic variants as previously described in (31). We detected 21 cases of physical genomic proximity between somatic variants and the deregulated miRNAs (Hamed et al. 2015). They are mostly located in chromosomes 1, 7, and 19 (Fig. 2). These 21 cases encompass 15 distinct mutations and 20 distinct dysregulated miRNAs. To test the significance of these cases, we performed 1000 Wilcoxon tests against random SNV positions considering the same mutation frequency for each chromosome. The deregulated miRNAs identified in the 21 cases were significantly closer to their somatic SNVs pairs in comparison to random SNV positions (*p-value* equal to 0.001). We also checked whether the non-deregulated miRNAs (925 miRNAs) are in genomic proximity to the 15 somatic mutations involved in the 21 cases as well. We found that 52 non-deregulated miRNAs (5.6%) were in vicinity to only 8 mutations so that the other 7 mutations are exclusively associated with the deregulated miRNAs (Hamed et al. 2015).

Similarly, we analyzed the somatic mutations that mainly occurred at differentially methylated CpG sites in promoter regions. Overall we identified 347 pairs of SNV- differentially methylated promoter regions (Hamed et al. 2015). These were mostly located on chromosomes 1, 5, and X (Fig. 3). To address how changes in methylation levels caused by tumorigenesis correlate with mutation rates of different mutation genotypes, we separately analyzed the cases of up- and down-methylated promoters. 234 cases involved up-methylated genes, whereas only 113 were associated with down-methylated genes. Generally, mutations in the promoter regions of up-methylated genes occur at a remarkably higher rate than its peers in down-methylated genes especially the C->T genotypes since methylated cytosines are prone to thymine transitions via deamination. This result is in line with the findings of Xia et al. (17) who examined the relationship between DNA methylation and mutation rate.

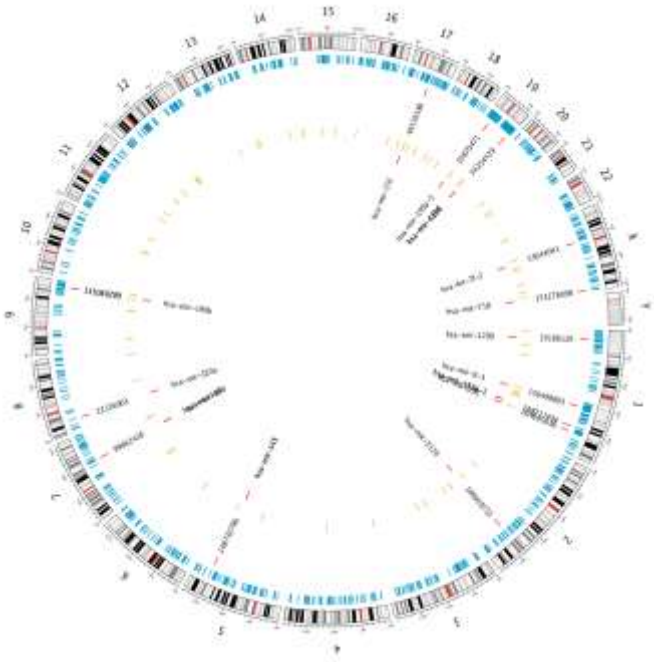

***Figure 2. Proximity analysis of the somatic mutations with the dysregulated miRNAs.*** *Ideogram plots showing the genomic distribution for the 21 cases of deregulated miRNAs adjacent to somatic mutations. The outer blue circle represents the subset of the breast cancer SNVs and the next highlighted red lines depict the SNVs matched to the 21 cases. The inner yellow circle shows the input deregulated set of miRNAs, whereas the next highlighted red lines refer to the adjacent (matched) deregulated miRNAs (20 miRNAs where one miRNA is matched to 2 SNVs).*

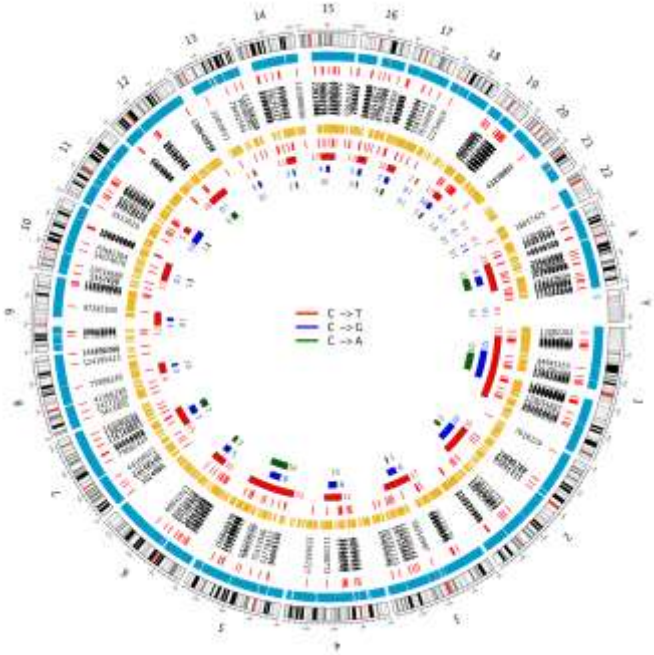


***Figure 3. Proximity analysis of the somatic mutations with differentially methylated regions.*** *Ideogram plots showing the genomic distribution for the 347 cases of SNVs occurring in the promoter regions of differentially methylated genes. The outer blue circle represents the entire set of breast cancer SNVs and the next highlighted red lines depict the SNVs matched to the identified cases. The inner yellow circle shows the entire set of differentially methylated genes, whereas the next highlighted red lines refer to the identified cases adjacent to the SNVs. The plot illustrates also the count of the three considered types of mutations (C->T, C->G and C->A).*

Proc. of the Second Intl. Conf. on Advances in Bio-Informatics and Environmental Engineering - ICABEE 2015

ISBN: 978-1-63248-043-9 doi: 10.15224/ 978-1-63248-043-9-127

### B. *Validation and assessment*

In order to validate the results obtained by the SnvDMiR pipeline, we examined which of the above somatic mutations, which were identified on the basis of their vicinity to either dysregulated miRNAs or differentially methylated genes, could potentially drive tumor cell proliferation in breast cancer. For this, we applied the CHASM tool (1) to distinguish between driver and passenger somatic mutations. As training set, we used the breast cancer labeled data (BRCA) curated from the COSMIC database (35) and provided by CHASM. We identified nine putative driver mutations (three from miRNA cases and six from differentially methylated gene cases) suggesting their causative role in tumorigenesis (Hamed et al. 2015). All these nine mutations are missense and lead to an amino acid substitution. Next, we analyzed the possible impact of the resulting amino acid substitution on structure and function of the respective protein using the PolyPhen (36) and SIFT (37) prediction tools. Interestingly, both methods predicted damaging effects of these mutations on protein function confirming their putative role in driving cancer (Hamed et al. 2015). This strongly supports the usefulness of the SnvDMiR pipeline in integrating the genetic mutations with deregulated miRNAs and DMRs to identify putative driver mutations that may open up new avenues for novel therapeutic drugs.

## IV. Conclusion

In this paper, we presented SnvDMiR, a freely–available R pipeline that examines the genomic proximity between somatic mutations, disease-related miRNAs, and DMRs in order to identify putative driver mutations that could possibly play an important role in disease pathways. We demonstrated the usefulness of the SnvDMiR pipeline by applying it to a dataset of breast cancer-related mutations, deregulated miRNAs, and DMRs. Further analysis revealed that almost half of the identified mutations were predicted to have deleterious effects on related protein functions and therefore might be valid targets for new drugs. Especially when combined with experimental validation, our proximity pipeline could promote important insights on disease genomic data to develop new therapeutic strategies and thus better treatment. Finally, SnvDMiR is an extendible pipeline that can be applied on various diseases-related datasets and can be further expanded to study cellular functions where such multi-dimensional genomic data are available.

### *Acknowledgment*

MH thanks the german academic exchange service (DAAD) and the graduate school of computer science, saarbrucken, Germany. MH was supported by DFG SFB 1027.

### *References*

[1]. Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, et al. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer research. 2009;69(16):6660-7.
[2]. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, et al. Patterns of somatic mutation in human cancer genomes. Nature. 2007;446(7132):153-8.
[3]. Jones S, Zhang X, Parsons DW, Lin JC-H, Leary RJ, Angenendt P, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. science. 2008;321(5897):1801-6.
[4]. Kaminker JS, Zhang Y, Waugh A, Haverty PM, Peters B, Sebisanovic D, et al. Distinguishing cancer-associated missense mutations from common polymorphisms. Cancer research. 2007;67(2):465-73.
[5]. Parsons DW, Jones S, Zhang X, Lin JC-H, Leary RJ, Angenendt P, et al. An integrated genomic analysis of human glioblastoma multiforme. Science. 2008;321(5897):1807-12.
[6]. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The consensus coding sequences of human breast and colorectal cancers. science. 2006;314(5797):268-74.
[7]. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007;318(5853):1108-13.
[8]. Torkamani A, Schork NJ. Prediction of cancer driver mutations in protein kinases. Cancer research. 2008;68(6):1675-82.
[9]. Barnholtz-Sloan J, Sloan AE, Land S, Kupsky W, Monteiro AN. Somatic alterations in brain tumors. Oncology reports. 2008;20(1):203-10.
[10]. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome research. 2001;11(5):863-74.
[11]. Zhong C, Hou Z, Huang J, Xie Q, Zhong Y. Mutations and CpG islands among hepatitis B virus genotypes in Europe. BMC Bioinformatics. 2015;16(1):38.
[12]. Gu Y, Liu G-H, Plongthongkum N, Benner C, Yi F, Qu J, et al. Global DNA methylation and transcriptional analyses of human ESC-derived cardiomyocytes. Protein & cell. 2014;5(1):59-68.
[13]. Das PM, Singal R. DNA methylation and cancer. Journal of Clinical Oncology. 2004;22(22):4632-42.
[14]. Esteller M, Herman JG. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. The Journal of pathology. 2002;196(1):1-7.
[15]. Li S, Garrett-Bakelman FE, Akalin A, Zumbo P, Levine R, To BL, et al. An optimized algorithm for detecting and annotating regional differential methylation. BMC bioinformatics. 2013;14(Suppl 5):S10.
[16]. Tahara T, Arisawa T, Shibata T, Yamashita H, Yoshioka D, Hirata I. Effect of promoter methylation of multidrug resistance 1 (MDR1) gene in gastric carcinogenesis. Anticancer research. 2009;29(1):337-41.
[17]. Xia J, Han L, Zhao Z. Investigating the relationship of DNA methylation with mutation rate and allele frequency in the human genome. BMC genomics. 2012;13(Suppl 8):S7.
[18]. Volinia S, Galasso M, Costinean S, Tagliavini L, Gamberoni G, Drusco A, et al. Reprogramming of miRNA networks in cancer and leukemia. Genome research. 2010;20(5):589-99.
[19]. Friedman RC, Farh KK-H, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome research. 2009;19(1):92-105.
[20]. Taft RJ, Pang KC, Mercer TR, Dinger M, Mattick JS. Non‐coding RNAs: regulators of disease. The Journal of pathology. 2010;220(2):126-39.
[21]. Esquela-Kerscher A, Slack FJ. Oncomirs—microRNAs with a role in cancer. Nature Reviews Cancer. 2006;6(4):259-69.
[22]. Medina PP, Slack FJ. microRNAs and cancer: an overview. Cell cycle. 2008;7(16):2485-92.
[23]. Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, Yi M, et al. Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. Cancer cell. 2006;9(3):189-98.
[24]. Ryan BM, Robles AI, Harris CC. Genetic variation in microRNA networks: the implications for cancer research. Nature Reviews Cancer. 2010;10(6):389-402.

ISBN: 978-1-63248-043-9 doi: 10.15224/ 978-1-63248-043-9-127

[25]. Sethupathy P, Collins FS. MicroRNA target site polymorphisms and human disease. Trends in genetics. 2008;24(10):489-97.
[26]. Bhattacharya A, Ziebarth JD, Cui Y. Systematic analysis of microRNA targeting impacted by small insertions and deletions in human genome. PloS one. 2012;7(9):e46176.
[27]. Mendell JT, Olson EN. MicroRNAs in stress signaling and human disease. Cell. 2012;148(6):1172-87.
[28]. Bhattacharya A, Ziebarth JD, Cui Y. SomamiR: a database for somatic mutations impacting microRNA function in cancer. Nucleic acids research. 2012:gks1138.
[29]. Abelson JF, Kwan KY, O'Roak BJ, Baek DY, Stillman AA, Morgan TM, et al. Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science. 2005;310(5746):317-20.
[30]. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic acids research. 2011;39(suppl 1):D152-D7.
[31]. Keller A, Leidinger P, Bauer A, ElSharawy A, Haas J, Backes C, et al. Toward the blood-borne miRNome of human diseases. nature methods. 2011;8(10):841-3.
[32]. Fatemi M, Pao MM, Jeong S, Gal-Yam EN, Egger G, Weisenberger DJ, et al. Footprinting of mammalian promoters: use of a CpG DNA methyltransferase revealing nucleosome positions at a single molecule level. Nucleic acids research. 2005;33(20):e176-e.
[33]. Gu Z, Gu L, Eils R, Schlesner M, Brors B. circlize implements and enhances circular visualization in R. Bioinformatics. 2014:btu393.
[34]. TCGAPortal. Nationl Human Genome Research Institutehttps://tcga-data.nci.nih.gov/tcga/.
[35]. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic acids research. 2010:gkq929.
[36]. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature methods. 2010;7(4):248-9.
[37]. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research. 2003;31(13):3812-4.

About Author (s):

**Alex Zapp** is persuading his post-graduate degree in CBI center for Bioinformatics, Saarland University, Saarbrucken, Germany. His research is based on developing integrative approaches to combine the somatic mutations with other genetic and epigenetic data.

**Volkhard Helms** is a professor for bioinformatics at Saarland University, Saarbrucken, Germany. His research group focuses on biomolecular interactions, and mechanisms of bacterial resistance and of stem cell differentiation.

**Mohamed Hamed** is a senior PhD student in the graduate school of computer science, Center for Bioinformatics (CBI) Saarbrucken, Germany. He is developing and applying computational approaches to study the regulatory machinery of stem cell differentiation process as well as the pathogenicity of complex diseases.