Hi-C Reveals The Proximity of sQTLs and Their Target Genes in 3D Genome of Cancer Cells

1 Abstract

2 Introduction

Single nucleotide polymorphisms (SNPs) are the most frequent genetic variants in humans and represent a valuable resource for investigating the genetic basis of diseases [17]. Genome-wide association studies (GWAS) have found abundant SNPs associated with various traits and diseases, but most of risk loci lack clear molecular mechanisms [2, 3]. Expression quantitative trait locus (eQTL) studies have been employed to identify SNPs that may influence the expression levels of genes, thereby contributing to the phenotype outcome [9, 28, 8]. However, only a moderate proportion of GWAS-identified loci are strong eQTLs [24], which might be partly due to the small sample sizes, the tissues studied, and a focus on overall gene level expression measurements without consideration of transcript isoforms [27].

Alternative splicing (AS) is a molecular mechanism that produces multiple distinct transcript isoforms from a single gene, increasing the structural transcript variation and proteome diversity. The number of different transcripts in the human transcriptome vastly exceeds the number of protein-coding genes [4]. On average, there are seven different transcript variants encoded by each protein-coding gene [10]. The invention of RNA sequencing greatly facilitated the identification of AS at the genomic level [11]. In human, alternative splicing can occur in more than 90% of genes in a cell type-, condition- or species-specific manner, which is thought to extensively increase the number of proteins over the number of genes in a genome [22, 1].

Increasing evidence has demonstrated that Alternative Splicing (AS) events could undergo modulation by genetic variants. The identification of splicing quantitative trait loci (sQTLs), genetic variants that affect AS events, might represent an important step toward fully understanding the contribution of genetic variants in disease development. AS process is often altered in cancer cells to produce aberrant proteins that drive the progression of cancer. A better understanding of the misregulation of alternative splicing will shed light on the development of novel targets for pharmacological interventions of cancer.

In cancer, aberrant splicing patterns are frequently observed in cancer initiation, progress, prognosis and therapy and known to contribute to carcinogenesis, dedifferentiation and metastasis [18, 21]. RNA sequencing have previously identified a number of cancer-specific transcript isoforms [6]. For

example, an alternatively spliced transcript isoform of the gene encoding spleen tyrosine kinase is frequently expressed in breast cancer cells but never in matched normal tissues [23].

In cancer, the splicing process is commonly disrupted, resulting in both functional and non-functional end-products. Cancer-specific splicing events are known to contribute to disease progression; however, the dysregulated splicing patterns found on a genome-wide scale have until recently been less well-studied.

Available evidence reveals that at least 20% of disease-causing single base-pair mutations (SNPs) affect splicing [7]. Common genetic variation that affects splicing regulation, referred to as splicing quantitative trait loci (sQTLs), can lead to differences in alternative splicing between individuals, consequently influence disease susceptibility and drug response [14]. Thus, the identification of sQTLs, especially in cancer tissues, might represent an important step toward fully understanding the contribution of genetic variants in tumorigenesis and development.

Single nucleotide polymorphisms (SNPs) are the most frequent genetic variants in humans and represent an invaluable resource to understand the genetic basis of diseases [17]. Genome-wide association studies (GWAS) have found abundant SNPs associated with various traits, cancers, and diseases, but most of risk loci lack clear molecular mechanisms [3]. Genome-wide sQTL mapping has been achieved for multiple species ranging from human to Arabidopsis Thaliana [13].

Gene regulation and alternative splicing (AS) are important for cells and tissues to function. It has been studied from two independent aspects at the genomic level, the identification of splicing quantitative trait loci (sQTLs) and identification of long-range chromatin interactions. It is important to understand their relationship, such as whether sQTLs regulate alternative splicing of genes through physical chromatin interaction.

We are interested in understanding whether sQTLs regulate their target genes through physical chromatin interactions. Our data will possibly demonstrate the close spatial proximity between sQTLs and their target genes among multiple human primary tissues, cancer tissues, and cell lines.

This has been previously shown in eQTLs. Chromatin interactions have been shown to be one of the main mechanisms underlying eQTLs both in cell lines and primary tissues.

Although chromatin interactions have been widely believed as one of the main mechanisms underlyinge sQTLs, we are unaware of any direct evidence of this for tissues. It is well known that sQTLs are tissue specific [3]. Moreover, Schmitt et al. [23] recently identified hotspots of local chromatin

interactions from Hi-C data, called frequently interacting regions (FIREs). FIREs are bins that frequently interact with nearby regions < 200Kb, and they display strong tissue specificity. It is unclear how much overlap exists between tissue-specific FIREs and tissue-specific sQTLs.

2.1 Related Work

Joint analyses of multi-tissue Hi-C and eQTL data demonstrate close spatial proximity between eQTLs and their target genes by [26] and [5] discuss work similar to ours.

ASpedia: Alternative Splicing Encyclopedia of Human

3 Experiments and Datasets

Are SQTLs occur often near TAD boundaries?

We obtain sQTLs for cancer cells from CancerSplicingQTL database [20] `http://www.cancersplicingqtl-hust.com/#/sqtls`, which compiles genome-wide sQTL data from publications []. We intersect gene names in this database with gene names or IDs in the Ensembl database [25] and select genes that are associated with a unique range in the Ensembl database to produce a collection S of sQTLs.

- Prostate cancer: A high-resolution 3D epigenomic map reveals insights into the creation of the prostate cancer transcriptome. `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118629` In this paper, normal prostate (RWPE1) and prostate cancer cells (C42B and 22Rv1). sQTL is in CancerSplicingQTL database under PRAD (Prostate Adenocarcinoma) [20].

- LAML (Acute Myeloid Leukemia): is of K562 cell type. Hi-C data for K562 cell is obtained in [15], sQTL is in CancerSplicingQTL database [20].

- CML (Chronic Myeloid Leukemia): is of KBM7 cell type. Hi-C data for K562 cell is obtained in [15], sQTL is in CancerSplicingQTL database [20].

- Multiple Myeloma (MM): 3D genome of multiple myeloma reveals spatial genome disorganization associated with copy number variations. GEO GSE87585. Could not find sQTL database.

- Breast Cancer (BRCA): *Information will be added!*
- Glioblastoma (GBM): *Information will be added!*
- Dorsolateral Prefrontal Cortex (DLPFC): sQTL is available in Supplementary files of Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci paper [19]. Hi-C data is available in [16].
- Hippocampus: Hi-C data is available in A Compendium of Chromatin Contact Maps Reveals Spatially Active Regions in the Human Genome [16].
- Some drugs also target 3D genome organization as in [12] on HT1080 cells (fibrosarcoma(SARC)). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM3304 Cool extension files are interaction matrices. sQTLs are available under xxx in CancerSplicingQTL

Some research questions we are interested in answering:

- Relationship between chromatin interaction frequency and number of sQTLs. What about TADs? We implement variety of regression models such as: 1- Negative binomial regression with covariates xxx and yyy.

  Negative binomial regression is implemented in statsmodels package in Python.

  Relationships between gene orthologs as well.

  Gene Ontology
- Does sQTLs appear more frequently in TAD boundaries?
- How is median-survival time is correlated with xxx?
- How is phenotype impacted?
- How does splicing type impact?
- Cis-trans difference
- ddd

Negative binomial regression is implemented in statsmodels package in Python.

References

[1] Nuno Barbosa-Morais, Manuel Irimia, Qun Pan, Hui Xiong, Serge Gueroussov, Leo Lee, Valentina Slobodeniuc, Claudia Kutter, Stephen Watt, Recep Colak, Taehyung Kim, Christine Misquitta, Michael Wilson, Philip Kim, Duncan Odom, Brendan Frey, and Benjamin Blencowe. The evolutionary landscape of alternative splicing in vertebrate species. *Science (New York, N.Y.)*, 338:1587–93, 12 2012.

[2] Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, Daniel Suveges, Olga Vrousgou, Patricia L Whetzel, Ridwan Amode, Jose A Guillen, Harpreet S Riat, Stephen J Trevanion, Peggy Hall, Heather Junkins, Paul Flicek, Tony Burdett, Lucia A Hindorff, Fiona Cunningham, and Helen Parkinson. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, 11 2018.

[3] Jiang Chang, Jianbo Tian, Yang Yang, Rong Zhong, Jiaoyuan Li, Kan Zhai, Juntao Ke, Jiao Lou, Wei Chen, Beibei Zhu, Na Shen, Yi Zhang, Yajie Gong, Ying Zhu, Danyi Zou, Xiating Peng, Kun Huang, and Xiaoping Miao. A rare missense variant in tcf7l2 associates with colorectal cancer risk by interacting with a gwas-identified regulatory variant in the myc enhancer. *Cancer Research*, 78(17):5164–5172, 2018.

[4] Sarah Djebali, Carrie Davis, Angelika Merkel, Alexander Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi Marinov, Jainab Khatun, Brian Williams, Chris Zaleski, Joel Rozowsky, Maik Röder, Felix Kokocinski, Rehab Abdelhamid, and Thomas Gingeras. Landscape of transcription in human cells. *Nature*, 489:101–8, 09 2012.

[5] Geet Duggal, Hao Wang, and Carl Kingsford. Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic Acids Research*, 42(1):87–96, 10 2013.

[6] Luisa Escobar-Hoyos, Katherine Knorr, and Omar Abdel-Wahab. Aberrant rna splicing in cancer. *Annual Review of Cancer Biology*, 3(1):167–185, 2019.

[7] Nuno André Faustino and Thomas A. Cooper. Pre-mrna splicing and human disease. *Genes Development*, 17(4):419–437, 2003.

[8] J. Gong, J. Tian, J. Lou, X. Wang, J. Ke, J. Li, Y. Yang, Y. Gong, Y. Zhu, D. Zou, X. Peng, N. Yang, S. Mei, R. Zhong, J. Chang, and X. Miao. A polymorphic myc response element in kbtbd11 influences colorectal cancer risk, especially in interaction with an myc-regulated snp rs6983267. *Annals of Oncology*, 29(3):632–639, 2020/12/17 2018.

[9] Xingyi Guo, Weiqiang Lin, Jiandong Bao, Qiuyin Cai, Xiao Pan, Mengqiu Bai, Yuan Yuan, Jiajun Shi, Yaqiong Sun, Mi-Ryung Han, Jing Wang, Qi Liu, Wanqing Wen, Bingshan Li, Jirong Long, Jianghua Chen, and Wei Zheng. A comprehensive cis-eqtl analysis revealed target genes in breast cancer susceptibility loci identified in genome-wide association studies. *The American Journal of Human Genetics*, 102(5):890–903, 2020/12/17 2018.

[10] Jennifer Harrow, Adam Frankish, Jose M. Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L. Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt, Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J. Hubbard. Gencode: The reference human genome annotation for the encode project. *Genome Research*, 22(9):1760–1774, 2012.

[11] Daejin Hyung, Jihyun Kim, Soo Young Cho, and Charny Park. ASpedia: a comprehensive encyclopedia of human alternative splicing. *Nucleic Acids Research*, 46(D1):D58–D63, 11 2017.

[12] Omar Kantidze, Artem Luzhin, Ekaterina Nizovtseva, Alfiya Safina, Maria Valieva, Arkadiy Golov, Artem Velichko, Alexander Lyubitelev, Alexey Feofanov, Katerina Gurova, Vasily Studitsky, and Sergey Razin. The anti-cancer drugs curaxins target spatial genome organization. *Nature Communications*, 10:1441, 03 2019.

[13] Waqas Khokhar, Musa A. Hassan, Anireddy S. N. Reddy, Saurabh Chaudhary, Ibtissam Jabre, Lee J. Byrne, and Naeem H. Syed. Genome-wide identification of splicing quantitative trait loci (sqtls) in diverse

ecotypes of arabidopsis thaliana. *Frontiers in Plant Science*, 10:1160, 2019.

[14] Emilie Lalonde, Kevin C.H. Ha, Zibo Wang, Amandine Bemmo, Claudia L. Kleinman, Tony Kwan, Tomi Pastinen, and Jacek Majewski. Rna sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Research*, 21(4):545–554, 2011.

[15] S.S.P. Rao, M.H. Huntley, Neva Durand, E.K. Stamenova, Ivan Bochkov, J.T. Robinson, A.L. Sanborn, I. Machol, Arina Omer, E.S. Lander, and E.L. Aiden. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 162:687–688, 08 2015.

[16] Anthony D. Schmitt, Ming Hu, Inkyung Jung, Zheng Xu, Yunjiang Qiu, Catherine L. Tan, Yun Li, Shin Lin, Yiing Lin, Cathy L. Barr, and Bing Ren. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Reports*, 17(8):2042 – 2059, 2016.

[17] Barkur S Shastry. Snps: impact on gene function and phenotype. *Methods in molecular biology (Clifton, N.J.)*, 578:3—22, 2009.

[18] Anita Sveen, S Kilpinen, A Ruusulehto, RA Lothe, and Rolf Skotheim. Aberrant rna splicing in cancer; expression changes and driver mutations of splicing factor genes. *Oncogene*, 35, 08 2015.

[19] Atsushi Takata, Naomichi Matsumoto, and Tadafumi Kato. Genome-wide identification of splicing qtls in the human brain and their enrichment among schizophrenia-associated loci. *Nature Communications*, 8:14519, 02 2017.

[20] Jianbo Tian, Zhihua Wang, Shufang Mei, Nan Yang, Yang Yang, Juntao Ke, Ying Zhu, Yajie Gong, Danyi Zou, Xiating Peng, Xiaoyang Wang, Hao Wan, Rong Zhong, Jiang Chang, Jing Gong, Leng Han, and Xiaoping Miao. CancerSplicingQTL: a database for genome-wide identification of splicing QTLs in human cancer. *Nucleic Acids Research*, 47(D1):D909–D916, 10 2018.

[21] Bi-Dar Wang and Norman Lee. Aberrant rna splicing in cancer and drug resistance. *Cancers*, 10:458, 11 2018.

[22] Eric Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, kkk Mayr, Stephen Kingsmore, Gary Schroth, and Christopher Burge. Al-

ternative isoform regulation in human tissue transcriptomes. *Nature*, 456:470–6, 12 2008.

[23] Lei Wang, Lindsay Duke, Peter S. Zhang, Ralph B. Arlinghaus, W. Fraser Symmans, Aysegul Sahin, Richard Mendez, and Jia Le Dai. Alternative splicing disrupts a nuclear localization signal in spleen tyrosine kinase that is required for invasion suppression in breast cancer. *Cancer Research*, 63(15):4724–4730, 2003.

[24] Harm-Jan Westra, Marjolein J Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W Christiansen, Benjamin P Fairfax, Katharina Schramm, Joseph E Powell, Alexandra Zhernakova, Daria V Zhernakova, Jan H Veldink, Leonard H Van den Berg, Juha Karjalainen, Sebo Withoff, AndréG Uitterlinden, Albert Hofman, Fernando Rivadeneira, Peter A C 't Hoen, Eva Reinmaa, Krista Fischer, Mari Nelis, Lili Milani, David Melzer, Luigi Ferrucci, Andrew B Singleton, Dena G Hernandez, Michael A Nalls, Georg Homuth, Matthias Nauck, Dörte Radke, Uwe Völker, Markus Perola, Veikko Salomaa, Jennifer Brody, Astrid Suchy-Dicey, Sina A Gharib, Daniel A Enquobahrie, Thomas Lumley, Grant W Montgomery, Seiko Makino, Holger Prokisch, Christian Herder, Michael Roden, Harald Grallert, Thomas Meitinger, Konstantin Strauch, Yang Li, Ritsert C Jansen, Peter M Visscher, Julian C Knight, Bruce M Psaty, Samuli Ripatti, Alexander Teumer, Timothy M Frayling, Andres Metspalu, Joyce B J van Meurs, and Lude Franke. Systematic identification of trans eqtls as putative drivers of known disease associations. *Nature Genetics*, 45(10):1238–1243, 2013.

[25] Andrew D Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, Jyothish Bhai, Konstantinos Billis, Sanjay Boddu, José Carlos Marugán, Carla Cummins, Claire Davidson, Kamalkumar Dodiya, Reham Fatima, Astrid Gall, Carlos Garcia Giron, Laurent Gil, Tiago Grego, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G Izuogu, Sophie H Janacek, Thomas Juettemann, Mike Kay, Ilias Lavidas, Tuan Le, Diana Lemos, Jose Gonzalez Martinez, Thomas Maurel, Mark McDowall, Aoife McMahon, Shamika Mohanan, Benjamin Moore, Michael Nuhn, Denye N Oheh, Anne Parker, Andrew Parton, Mateus Patricio, Manoj Pandian Sakthivel, Ahamed Imran Abdul Salam, Bianca M Schmitt, Helen Schuilenburg, Dan Sheppard, Mira Sycheva, Marek Szuba, Kieron Taylor, Anja Thormann, Glen Threadgold, Alessandro Vullo, Brandon Walts, An-

drea Winterbottom, Amonida Zadissa, Marc Chakiachvili, Bethany Flint, Adam Frankish, Sarah E Hunt, Garth IIsley, Myrto Kostadima, Nick Langridge, Jane E Loveland, Fergal J Martin, Joannella Morales, Jonathan M Mudge, Matthieu Muffato, Emily Perry, Magali Ruffier, Stephen J Trevanion, Fiona Cunningham, Kevin L Howe, Daniel R Zerbino, and Paul Flicek. Ensembl 2020. *Nucleic Acids Research*, 48(D1):D682–D688, 11 2019.

[26] Jingting Yu, Ming Hu, and Chun Li. Joint analyses of multi-tissue hi-c and eqtl data demonstrate close spatial proximity between eqtls and their target genes. *BMC Genetics*, 20, 12 2019.

[27] Xiaoling Zhang, Roby Joehanes, Brian H Chen, Tianxiao Huan, Saixia Ying, Peter J Munson, Andrew D Johnson, Daniel Levy, and Christopher J O'Donnell. Identification of common genetic variants controlling transcript isoform variation in human whole blood. *Nature Genetics*, 47(4):345–352, 2015.

[28] Danyi Zou, Jiao Lou, Juntao Ke, Shufang Mei, Jiaoyuan Li, Yajie Gong, Yang Yang, Ying Zhu, Jianbo Tian, Jiang Chang, Rong Zhong, Jing Gong, and Xiaoping Miao. Integrative expression quantitative trait locus–based analysis of colorectal cancer identified a functional polymorphism regulating slc22a5 expression. *European Journal of Cancer*, 93:1–9, 2020/12/17 2018.