Impact of mutations within lincRNAs region in breast cancer gene expression programs

Abstract

The accumulation of genetic alterations that lead to cancer disease occur mainly in genomic regions that do not encode for proteins. These somatic and germline alterations impact on RNA transcripts potentially involved in genes regulation. Recently, some long intergenic noncoding RNAs (lincRNAs) have been shown to be differentially expressed in different cancer types with potential regulatory functions on cancer genes. Additionally, pan-cancer analysis showed that lincRNAs co-express with cancer genes in normal and tumor tissues and linked these lincRNAs with tumor suppressor or oncogenic functions.

However, genetic alterations leading to the differential expression of lincRNAs in cancer and that result in cancer genes dysregulation remains uninvestigated.

Hence, this project aims to identify lincRNAs with regulating function on oncogenes and tumor suppressor genes in breast cancer. First, I will investigate how lincRNAs and cancer gene expression levels correlate in normal and tumoral tissues. Then, I will evaluate the impact of genetic alterations within cancer genes associated lincRNAs regions on their putative cancer gene target. Finally, I will train a supervised machine-learning algorithm on a dataset of lincRNAs with known tumor suppressor and oncogenic activity to further predict tumor suppressor or oncogenic functions of new lincRNAs with unknown activity based on their functional mutation pattern.

This project aims to link breast-expressed lincRNAs with their potential oncogenic or tumor suppressor functions. We anticipate that it will allow the better understanding of the regulatory mechanisms of cancer genes by lincRNAs, with notable insights on the structural and functional elements within lincRNAs region that are important for genes regulation process.

Introduction

Cancer is a heterogeneous and widespread disease that can affect many different organs, with an overall worldwide annual incidence of 14 million new diagnosis and 8 million deaths. Some cancers are strongly related to gender, such as breast cancer affecting in most cases women (http://www.cancerresearchuk.org).

Cancerous cells are characterized by an aberrant state, in which a self-sufficiency in growth signal and an insensitivity to anti-growth signal allow to escape the cell cycle control and apoptosis. This can results in extensive cellular divisions and potential invasion of distant organs (Hanahan et al., 2000).

The development of cancerous cells usually results from the accumulation of several genetic alterations. These genetic alterations can occur in the germline and be inherited, but most commonly they have a somatic origin (Milholland et al., 2017). For example, they can be generated by pathogens, such as Rous Sarcoma Virus, or can be induced by an external factor, such as UV irradiation, smoking or diet (Jensen et al., 1964).

Most of these genetic alterations impair the expression of genes involved in cell proliferation and apoptosis. These genetic mutations can be classified in two categories: (i) overactivation of oncogenes, which stimulate cell division or repress apoptosis; and (ii) inactivation of tumor suppressor genes, which decrease the cell ability to control its cycle and regulate apoptosis. Both categories of mutations are associated with malignancy and their expression promotes the occurrence of cancer (Hanahan et al., 2000).

Recently, studies showed that most disease-associated somatic and germline alterations are localized in parts of the genome that do not encode for proteins (Maurano et al., 2012). Notably, lincRNAs, a class of non-coding transcripts longer than 200 nucleotides and transcribed from intergenic loci, was found to be abnormally expressed in tumor tissues (Cabanski et al., 2015; Li et al., 2016).

The number of differentially expressed lincRNAs exhibits a lot of variations across cancer types (coefficient of variability: 1.01), which indicates that some cancer type show a larger amount of differentially expressed lincRNAs as compared to others. Indeed, Cabanski *et al. (2015)* identified an average of 102 lincRNAs that are differentially expressed between tumor and normal tissues across 8 different cancer types, despite the lack of functional characterization of most lincRNAs and their low expression which makes them difficult to study (Derrien et al., 2012). Besides, 76% of the lincRNAs differentially expressed in malignant cells are unique to a single cancer type. Additionally, lincRNAs are not differentially expressed in more than 5 cancer types. Together, it suggests that some lincRNAs could be more important in certain cancer than others. For example, the lincRNAs *TCONS_00011854* is overexpressed in colorectal cancer (CRC) only and has an higher tumor expression than known CRC biomarkers such as *CCAT1*. Thus, this lincRNA is suggested to act as better biomarker for CRC than the previously used biomarkers (Cabanski et al., 2015).

The link between lincRNAs and malignancy could be tightly related to tumor suppressor or oncogenic functions, with possible regulatory functions on cancer genes. For example, ectopic expression of Sammson, a *trans*-acting lincRNA, facilitates oncogenic transformation of melanocytes into malignant melanoma. This lincRNA is specifically overexpressed in more than 90% of human melanoma and its expression level appears to correlate with its gene copy number. In contrast, targeted degradation of Sammson reduces cells transformation and decrease malignant cells survival, without affecting nearby genes expression level (Leucci et al., 2016).

The evidence that lincRNAs are differentially expressed in cancerous cells and their involvement in genes regulation has led researchers to investigate their role in cancer genes regulation. Recent pan-cancer studies mainly aimed to identify lincRNAs that are co-upregulated with cancer genes in tumor tissues (Cabanski et al., 2015; Wu et al., 2016; Liu et al., 2016), or lincRNAs that overlap known disease-associated single nucleotide polymorphisms (SNPs) (Chui et al., 2018). While such pan-cancer studies largely reported aberrant lincRNAs expressions in human cancer, they did not investigate how candidate lincRNAs genetic alterations impact on lincRNAs expression level and on the levels of their putative cancer gene target.

These genetic alterations that may impact gene expression levels are suggested to be located in different regions between *cis*-regulatory element, that regulate target genes located close to their site of synthesis, and *trans*-regulatory elements, that regulate target genes located at larger distance, including on others chromosomes. Genetic alterations within *cis*-acting elements are suspected to be mostly located in regions that contribute to transcription factors binding sites (TFBSs) (Poulos et al., 2015). On the other hand, genetic alterations that lie within *trans*-acting factors are suspected to be mostly located in regions that contribute to the TFBSs, to the secondary structure of the RNA, to the chromatin binding properties or to RNA-RNA interactions (Diederichs et al., 2016).

LincRNAs can be discriminated for *cis*- or *trans*-activity based on their histone modifications at their transcriptional start sites (TSS). Indeed, some lincRNAs show an enrichment of trimethylation of lysine 4 of histone H3 (H3K4me3) similar to protein-coding genes promoters. It is suggested that they are transcribed from promoter-like elements, and are therefore referred to as promoter-like lincRNAs (plincRNAs). On the other hand, a second class of lincRNAs is found to be transcribed from enhancer region. They exhibits a low level of H3K4me3 and a relative high level of H3K4me1, characteristic of enhancer elements. It has been established that elincRNAs are more weakly expressed than plincRNAs, but their expression is more strongly associated with an increase of expression of their neighboring protein-coding genes, and thus are mostly likely to act as *cis*-regulatory elements (Marques et al., 2013).

Here, we aim to investigate a role of *cis*- and *trans*-acting lincRNAs in cancer by detecting which are co-expressed with cancer genes in tumor tissues, and assessing the impact of genetic alterations in the lincRNA on the dysregulation of the co-expressed lincRNA and cancer gene expression level. We will also investigate the topology of mutated sites in lincRNAs regions and particularly the position of mutations along lincRNAs transcript, which might allows the identification of functional region in lincRNAs.

Research Plan

In the following, I propose to (i) annotate *cis*- and *trans*-acting breast-expressed lincRNAs co-expressed in normal and tumor samples with oncogenes and tumor suppressor genes, (ii) investigate the impact of genetic alterations within lincRNAs on their putative protein-coding target expression, and (iii) predict tumor suppressor and oncogenic lincRNAs association by functional mutation bias.

Because of their different process of gene regulation and their potential differences in different regions that exhibit genetic alterations, *cis*-acting lincRNAs and *trans*-acting lincRNAs will be investigated separately. In my master project I will carry out aims 1 and 2 focusing on *cis*-acting lincRNAs.

Aim 1 - Annotation of breast cancer associated *cis*- and *trans*-acting lincRNAs

To identify *cis*- and *trans*-acting lincRNAs involved breast cancer, I propose to retrieve and curate breast-expressed *cis*- and *trans*-acting lincRNAs, to detect those co-expressed with a cancer genes in normal sample, and to investigate pairs of lincRNAs and cancer genes that are differentially expressed in tumor samples and that may contribute to oncogenes or tumor suppressor genes regulation.

First, I will identify breast-expressed lincRNAs. To determine these genes expression, I will align RNA-sequencing reads from Michigan Cancer Foundation-7 (Mcf-7) cell line (Soule et al., 1973) retrieved from ENCODE (Encode Consortium, 2012). I will use GENCODE annotations for lincRNAs to determine which lincRNAs have evidences of expression in breast tissues. Since the RNA is randomly fragmented and sequenced from the 3'-end in RNA-sequencing, it does not allow to locate exact transcription start site (TSS) of lincRNA genes. To localize TSS, I will use publically available Cap Analysis of Gene Expression (CAGE) from Mcf-7, which is a method that capture the 5'-end of transcribed and capped RNAs and that provides a single base-pair resolution map of TSS (Haberle et al., 2015). It is used to infer the exact position of lincRNAs TSS.

Then, I will discriminate these lincRNAs into enhancer and promoter associated lincRNAs (elincRNAs and plincRNAs, respectively). I will colocalize DNase I hypersensitivity sites (DHSs), with CAGE tags, as DHSs have been shown to map with regulatory elements such as promoters and enhancer (Boyle et al., 2008). It will allow the identification of regulatory regions that are transcribed in Mcf-7 cells. I will use the Chromatin Immunoprecipitation sequencing (ChIP-seq) from Mcf-7 to detect which of these regions correspond to a promoter or enhancer. I will then intersect these regions with lincRNAs TSS to detect which lincRNAs are associated to enhancer and promoters.

The co-expression analysis of lincRNAs and cancer genes for normal breast samples will be performed using breast sample expression data from Genotype-Tissue Expression (GTEx), that correspond to 290 donors. From GTEx data, I will focus on lincRNAs and cancer genes expression levels, that I will retrieve by their gene-id. To do so, I will use GENCODE annotations for breast-expressed lincRNAs and the Cancer Gene Census (CGC), which is a database of cancer genes annotated as oncogenes or tumor suppressor genes.

To identify co-expressed lincRNAs and cancer genes, I will compute the correlation of their gene expression levels. To define that a co-expression is significant for a lincRNA and a cancer gene, I will evaluate, by the resulted p-value, whether the frequency of their expression level correlation is significantly different from the frequency of correlations of this lincRNA expression level with all gene expression levels. Additionally, I will use the method of Amar *et al. (2013)* to shuffle cancer genes expression levels among samples. I will recompute the correlation for each shuffled set of samples and evaluate the distribution of correlation differences. It is expected few differences in correlation between shuffled set for co-expressed genes.

I will infer a biological function for lincRNAs that are significantly co-expressed with a cancer gene. To do so, I will functionally characterize each co-expressed mRNA, that will be subject to KEGG, Gene Ontology and Biocarta annotations. I will then investigate whether some of these mRNA share a pathway. It may indicate that their co-expressed lincRNAs are part of the regulation of the same pathway. Complemented by the following mutations analysis, it will allow to identify the most frequently altered pathway in breast cancer.

Finally, I will investigate pairs of lincRNAs and cancer genes that are differentially expressed in tumor samples, and that may contribute to breast cancer. To do so, I will use data from The Cancer Genome Atlas (TCGA), from which I will focus on breast cancer cases that have available at least the Transcriptome profiling data, Copy Number Variations (CNVs) data and Simple Nucleotide Variations (SNVs) data. It represents more than one thousand cases of breast cancer. I will retrieve the

gene expression levels in tumor sample for the co-expressed pairs of lincRNAs and cancer genes previously identified in normal samples, and identify pairs for which there is a differential expression both for the lincRNA and the cancer gene in tumor samples. I will then compute their correlation and investigate pairs for which correlations are similar (i.e same direction) between the two samples. I will characterize these lincRNAs with several features, such as their difference in length, in expression level, in number and size of introns/exons or in evolutionary conservation. I expect to find specific genomic features for lincRNAs co-expressed with tumor suppressor compared to lincRNAs co-expressed with oncogenes.

Aim 2 - How genetic alteration within lincRNAs can impact the level of the associated protein-coding target?

Given the putative involvement of lincRNAs in *cis*- and *trans*-regulation of genes as indicated by recents studies, we propose that genetic alterations that lie within lincRNAs regions may be directly related to variations in expression of the protein-coding gene targeted by the lincRNA. Because of the different process of gene regulation by *cis*-acting and *trans*-acting lincRNAs (Metzger et al., 2016), I expect that genetic alterations that impact *cis*-acting lincRNAs (elincRNAs) lie in different regions than genetic alterations that impact *trans*-acting lincRNAs (plincRNAs). Because of this, *cis*-acting lincRNAs and *trans*-acting lincRNAs will be investigated separately.

I will use CNVs and SNVs data from TCGA to identify genetic alterations that are located in regions of previously identified breast cancer-associated lincRNAs.
I will compute the recurrence of each CNV and SNV across samples to identify those that have a relatively high frequency and that are likely to have a functional impact. I will then investigate the position of these genetic alterations in order to detect functional elements and domains within lincRNAs region.

I will also investigate whether the breast cancer-associated lincRNAs contain any known breast cancer risk loci identified in previous GWAS of breast cancer (Welter et al., 2013). It will allow to associated lincRNAs in which are located breast cancer risk loci with the SNP-associated trait. To do so, I will retrieve from the GWAS Catalog breast-cancer associated SNPs that I will mapp on genomic regions. I will then investigate whether some SNPs are located within lincRNAs regions. Finally, I will investigate others parameters such as the relation of genetic alterations within lincRNAs regions to the variation of lincRNA and target protein-coding gene expression level, or the proportion of each type of alterations (SNVs, CNVs) across the population.

Aim 3 - Identification of tumor suppressor and oncogenic lincRNAs by functional mutation bias

Cancer pathway are mostly altered by mutations that impact expression of oncogenes and tumor suppressor genes. As these two classes of genes have opposite functions on cell cycle control, they might have different regulatory mechanisms. I suggest that these differences in regulatory mechanisms may result in different mutated genomic regions between oncogenic and tumor suppressor regulatory elements, which will be reflected by a different impact on the phenotype.

I propose to investigate differences in impact of mutated sites between lincRNAs that regulate tumor suppressor genes or oncogenes, and to use those potential differences to predict whether a lincRNA is associated with oncogenic or tumor suppressor function.
To do so, I will use a method that aim to analyze the functional impact of tumor somatic mutations and that is able to investigate both coding and non-coding genomic regions. This method, referred to as OncoDriveFML, is able to compute the functional mutations (FM) bias for a genomic element and asses its impact on several targets, such as the RNA secondary structure or transcription factor binding sites (TFBSs) (Mularoni et al., 2016).
I will train a supervised machine-learning algorithm on a dataset of lincRNAs with known tumor suppressor and oncogenic activity. The oncogenic lincRNA activity will be defined from the literature. In the training phase, the algorithm will learn to predict whether a lincRNA acts as tumor suppressor or oncogene, basing its decision on general properties of genes (Libbrecht et al., 2015). After training the algorithm on lincRNAs with known activity, I will use it to predict oncogenic activity (tumor suppressor versus oncogene) on lincRNAs from a dataset of somatic mutations in breast cancer retrieve from TCGA and the genomic coordinates of lincRNAs obtained from GENCODE. The algorithm will compute the FM bias for each lincRNA and assign it to one of the classes.

Significance of the project

Several lincRNAs have recently been highlighted for their involvement in *cis*- and *trans*-regulation of cancer genes. However, the impact of mutations within lincRNAs regions on the co-expressed cancer genes expression level is still under current research.
With this project, I will gain new insights on the *cis*- and *trans*-regulation of oncogenes and tumor suppressor by lincRNAs and on their dysregulation by genetic alteration within lincRNAs region.
The annotations of *cis*- and *trans*-acting lincRNAs and the detection of their co-expression with cancer genes in breast normal and tumor samples will allow to identify lincRNAs associated with breast cancer.
The identification of mutations within these breast cancer-associated lincRNAs will allow to detect structural and functional elements within lincRNAs sequence that may be involved in the regulation of cancer genes.
Finally, a trained and supervised machine-learning algorithm will allow to predict lincRNAs association with oncogenic or tumor suppressor functions.
Overall, this project will allow a better comprehension of the involvement of lincRNAs in breast cancer. It will also lead to the identification of functional elements within lincRNAs, which still remains undetermined.

Timetable:

Aim 1: 4 months
Aim 2: 6 months
AIm 3: 6 months

Wordcount:

Abstract	: 255
Introduction	: 994
Research plan	: 1484
Significance of the project	: 178
Total	: 2911

References

Amar, David, Hershel Safer, and Ron Shamir. 2013. “Dissection of Regulatory Networks That Are Altered in Disease via Differential Co-Expression.” PLoS Computational Biology 9 (3). https://doi.org/10.1371/journal.pcbi.1002955.

Boyle, Alan P, Sean Davis, Hennady P Shulha, Paul Meltzer, Elliott H Margulies, Zhiping Weng, Terrence S Furey, and Gregory E Crawford. 2008. “NIH Public Access.” Cell 132 (2):311–22. https://doi.org/10.1016/j.cell.2007.12.014.High-Resolution.

Cabanski, Christopher R., Nicole M. White, Ha X. Dang, Jessica M. Silva-Fisher, Corinne E. Rauck, Danielle Cicka, and Christopher A. Maher. 2015. “Pan-Cancer Transcriptome Analysis Reveals Long Noncoding RNAs with Conserved Function.” RNA Biology 12 (6):628–42. https://doi.org/10.1080/15476286.2015.1038012.

Chiu, Hua-Sheng, Sonal Somvanshi, Ektaben Patel, Ting-Wen Chen, Vivek P Singh, Barry Zorman, Sagar L Patil, et al. 2018. “Pan-Cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context.” Cell Rep 23 (1):297–312.e12. https://doi.org/10.1016/j.celrep.2018.03.064.

Derrien, Thomas, Rory Johnson, Giovanni Bussotti, Andrea Tanzer, Sarah Djebali, Hagen Tilgner, Gregory Guernec, et al. 2012. “The GENCODE v7 Catalog of Human Long Noncoding RNAs : Analysis of Their Gene Structure , Evolution , and Expression,” 1775–89. https://doi.org/10.1101/gr.132159.111.

Diederichs, Sven, Lorenz Bartsch, Julia C Berkmann, Karin Fröse, Jana Heitmann, Caroline Hoppe, Deetje Iggena, et al. 2016. “The Dark Matter of the Cancer Genome: Aberrations in Regulatory Elements, Untranslated Regions, Splice Sites, Non-coding RNA and Synonymous Mutations.” EMBO Molecular Medicine 8 (5):442–57. https://doi.org/10.15252/emmm.201506055.

Encode Consortium, North Carolina, and Chapel Hill. 2013. “For Junk DNA.” Nature 489 (7414):57–74. https://doi.org/10.1038/nature11247.An.

Haberle, Vanja, Alistair R R Forrest, Yoshihide Hayashizaki, Piero Carninci, and Boris Lenhard. 2015. “CAGEr: Precise TSS Data Retrieval and High-Resolution Promoterome Mining for Integrative Analyses.” Nucleic Acids Research 43 (8). https://doi.org/10.1093/nar/gkv054.

Hanahan, D, and R A Weinberg. 2000. “The Hallmarks of Cancer.” Cell 100 (1):57–70. https://doi.org/10.1007/s00262-010-0968-0.

Jensen, Fc, and Aj Girardi. 1964. “Infection of Human and Simian Tissue Cultures with Rous Sarcoma Virus.” Proceedings of the … 52:53–59. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC300571/.

Leucci, Eleonora, Roberto Vendramin, Marco Spinazzi, Patrick Laurette, Mark Fiers, Jasper Wouters, Enrico Radaelli, et al. 2016. “Melanoma Addiction to the Long Non-Coding RNA SAMMSON.” Nature 531 (7595). Nature Publishing Group:518–22. https://doi.org/10.1038/nature17161.

Li, Chengyun, Geyu Liang, Wenzhuo Yao, Jing Sui, Xian Shen, Yanqiu Zhang, Shumei Ma, et al. 2016. “Differential Expression Profiles of Long Non-Coding RNAs Reveal Potential Biomarkers for Identification of Human Gastric Cancer.” Oncology Reports 35 (3):1529–40. https://doi.org/10.3892/or.2015.4531.

Libbrecht, Maxwell W, and William Stafford Noble. 2017. “Machine Learning in Genetics and Genomics.” Nature Reviews Genetics 16 (6):321–32. https://doi.org/10.1038/nrg3920.Machine.

Liu, Yining, and Min Zhao. 2016. “LnCaNet: Pan-Cancer Co-Expression Network for Human lncRNA and Cancer Genes.” Bioinformatics 32 (10):1595–97. https://doi.org/10.1093/bioinformatics/btw017.

Marques, Ana C., Jim Hughes, Bryony Graham, Monika S. Kowalczyk, Doug R. Higgs, and Chris P. Ponting. 2013. “Chromatin Signatures at Transcriptional Start Sites Separate Two Equally Populated yet Distinct Classes of Intergenic Long Noncoding RNAs.” Genome Biology 14 (11). https://doi.org/10.1186/gb-2013-14-11-r131.

Maurano, Matthew T, Richard Humbert, Eric Rynes, Robert E Thurman, Hao Wang, Alex P Reynolds, Richard Sandstrom, et al. 2012. “Science.” Science 337 (6099):1190–95. https://doi.org/10.1126/science.1222794.Systematic.

Metzger, Brian P.H., Fabien Duveau, David C. Yuan, Stephen Tryban, Bing Yang, and Patricia J. Wittkopp. 2016. “Contrasting Frequencies and Effects of Cis- and Trans-Regulatory Mutations Affecting Gene Expression.” Molecular Biology and Evolution 33 (5):1131–46. https://doi.org/10.1093/molbev/msw011.

Milholland, Brandon, Xiao Dong, Lei Zhang, Xiaoxiao Hao, Yousin Suh, and Jan Vijg. 2017. “Differences between Germline and Somatic Mutation Rates in Humans and Mice.” Nature Communications 8 (May). Nature Publishing Group:1–8. https://doi.org/10.1038/ncomms15183.

Mularoni, Loris, Radhakrishnan Sabarinathan, Jordi Deu-Pons, Abel Gonzalez-Perez, and Núria López-Bigas. 2016. “OncodriveFML: A General Framework to Identify Coding and Non-Coding Regions with Cancer Driver Mutations.” Genome Biology 17 (1). Genome Biology:1–13. https://doi.org/10.1186/s13059-016-0994-0.

Poulos, Rebecca C., Mathew A. SLoane, Luke B. Hesson, and Jason W. H. Wong. 2001. “The Search for.” Health (San Francisco) 2 (1):19–30. https://doi.org/10.1079/AHRR200112.

Soule, H. D., J. Vazquez, A. Long, S. Albert, and M. Brennan. 1973. “A Human Cell Line From a Pleural Effusion Derived From a Breast Carcinoma 2.” JNCI: Journal of the National Cancer Institute 51 (5):1409–16. https://doi.org/10.1093/jnci/51.5.1409.

Welter, Danielle, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, et al. 2014. “The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations.” Nucleic Acids Research 42 (D1):1001–6. https://doi.org/10.1093/nar/gkt1229.

Wu, Wenting, Erin K. Wagner, Yangyang Hao, Xi Rao, Hongji Dai, Jiali Han, Jinhui Chen, Anna Maria V. Storniolo, Yunlong Liu, and Chunyan He. 2016. “Tissue-Specific Co-Expression of Long Non-Coding and Coding RNAs Associated with Breast Cancer.” Scientific Reports 6 (May). Nature Publishing Group:1–13. https://doi.org/10.1038/srep32731.