UNIVERSIDADE DE SÃO PAULO

Programa Interunidades de Pós-Graduação em Bioinformática

LUCAS FERREIRA DA SILVA

**LincRNAs fisicamente associados ao receptor de andrógeno
modificam o perfil de marcas da cromatina vizinha
e alteram a expressão gênica local**

SÃO PAULO

2017

LUCAS FERREIRA DA SILVA

**LincRNAs fisicamente associados ao receptor de andrógeno
modificam o perfil de marcas da cromatina vizinha
e alteram a expressão gênica local**

*Tese apresentada ao Programa Interunidades
de Pós-Graduação em Bioinformática da
Universidade de São Paulo para a obtenção do
Título de Doutor em Bioinformática.*

Área de Concentração: Bioinformática

Orientador: Dr. Sergio Verjovski-Almeida

Co-orientador: Dr. João Carlos Setubal

São Paulo

2017

Ficha catalográfica

daSilva, Lucas Ferreira

LincRNAs fisicamente associados ao receptor de andrógeno modificam o perfil de marcas da cromatina vizinha e alteram a expressão gênica local / Lucas Ferreira daSilva – São Paulo, 2017.

119p.

Tese (Doutorado). Programa Interunidades de Pós-graduação em Bioinformática da Universidade de São Paulo.

Orientador: Verjovski-Almeida, Sergio

Co-orientador: Setubal, João Carlos

1. lincRNA 2. Epigenética 3. RNA longo não-codificador 4. câncer

I. Universidade de São Paulo. Instituto de Matemática e Estatística.

AGRADECIMENTOS

Agradeço aos meus pais pelo incentivo durante todos esses anos. Pelo apoio e compreensão que foram gratuitamente oferecidos pelos dois. Agradeço minha mãe por ser a pessoa mais afetuosa e parcimoniosa que eu já conheci. Por possuir uma simplicidade tão intrigante que não poderia ser descrita em uma enciclopédia inteira. Ao meu pai por sempre ter valorizado e catalisado minha vontade criativa. Agradeço a ele por insistir em suas frases e metáforas que de tempos em tempos ricocheteiam vigorosamente na minha cabeça. Agradeço aos meus irmãos pela amizade cultivada durante todos esses anos de convivência.

Ao professor Sergio Verjovski pela oportunidade de participar do seu grupo e revigorar minha esperança em um momento de incerteza. Certamente seu pragmatismo, experiência, elogios e provocações incitaram minha vontade e ampliaram minha visão como pesquisador.

Ao professor Setubal pela coorientação e pela pronta disponibilização de sua infraestrutura computacional. Sem ele não seria possível realizar muitas das análises apresentadas aqui.

Ao professor Eduardo pelas suas críticas e sugestões durante o período inicial desse trabalho.

Aos diversos companheiros de laboratório; Alexandre, Julio Sosa, Ester, Bruno Souza, Bruno Silva, Diogo, Santiago, Bianca, Ana Ayupe, Mariana Carvalho, David, Huma, David Morales, Monete e Guilherme.

À Ana Paula pela amizade e por sempre me ajudar utilizando suas super habilidades de secretária.

Aos amigos mais próximos; a Adriana Andrade por ser minha oponente na sinuca e ao mesmo tempo uma grande amiga nos outros momentos da vida. Agradeço por me adiantar as novas tendências musicais do verão, mesmo sabendo que eu odiaria. Ao amigo Elton Vasconcelos e toda sua musicalidade, boas conversas e sugestões. Ao Amigo Dinar pelas contribuições feitas para o meu trabalho e por me acompanhar em diversos “programas de índio” (provavelmente gastarei algum tempo explicando essa expressão para ele...). Ao Murilo pelo seu senso de humor sutil e pelo sua forma de indagar que sempre me levou a repensar resultados. Ao aluno de pré-iniciação Vinícius por toda sua ajuda e por não enlouquecer precocemente durante o processo. Agradeço a Letícia pelas inúmeras discussões, amizade,

disponibilidade, conselhos e fornecimento de chá. Provavelmente duas pessoas juntas nunca arquitetaram tantas análises minuciosas que resultaram em tantos valores-p não significativos. Mas não desistiremos! Ao Felipe Beckedorff pela amizade e pela participação direta neste trabalho. Agradeço a ele por me introduzir no mundo do sequenciamento de nova geração e epigenética. Este trabalho não seria possível sem suas sugestões, ideias e trabalho duro.

Agradeço ao amigo Gilberto Kaihami pelas muitas horas de conversa sobre Python e bioinformática. Ao amigo André Anversa por compartilhar todo seu fascínio pelos acontecimentos do mundo material. À amiga Adriana Domarques pelos nanosegundos que demora para entender comentários deixados no ar e por sempre compartilhar visões humoradas de coisas corriqueiras. À Carol Parga pela companhia e pelas suas onomatopeias engraçadas.

Ao Diego por ser um amigo de longo data e por sua maneira única de encarar qualquer tipo de situação.

Em especial, agradeço a Maíra Nagai por tolerar diversos de meus raciocínios herméticos após o expediente de trabalho, mesmo quando ela estava com fome e cansada. Pelo seu constante incentivo, carinho, compreensão. Agradeço também pelos nossos longos diálogos costurados a partir de fragmentos da “gratuidade da coisa”. O apoio dela para a construção de todo esse trabalho é imprescindível.

Agradeço ao Instituto de Matemática e Estatística e ao Departamento de Bioquímica da USP.

Ao Instituto Butantan e às agências financiadoras FAPESP pelo auxílio financeiro ao laboratório e CAPES pela bolsa de pós-graduação.

RESUMO

A sinalização celular desencadeada na presença do hormônio andrógeno em células da próstata é dependente da ativação do receptor de andrógeno (AR), que é um fator de transcrição codificado pelo gene NR3C4. O AR quando não ligado ao hormônio andrógeno é encontrado no citoplasma, e a ligação do hormônio ao AR promove seu deslocamento para o núcleo. O AR se liga a motivos de DNA, promovendo a transcrição de determinados genes por meio de um complexo mecanismo de regulação ainda não totalmente conhecido. Neste trabalho exploramos a capacidade dos RNAs longos intergênicos não-codificadores (lincRNAs) se ligarem ao AR e contribuírem para a regulação transcricional de genes. Utilizamos a imunoprecipitação de RNA, seguida de sequenciamento em larga escala (RIP-Seq) para a identificação e quantificação de lincRNAs associados fisicamente ao AR na linhagem celular de próstata LNCaP. Em paralelo utilizamos dados de transcriptoma e de marcas epigenéticas para verificar se a interação física dos lincRNAs com o AR influenciaria a composição da cromatina e a expressão dos genes vizinhos desses ARA-lincRNAs (*Androgen Receptor Associated LincRNAs*). Como resultado deste estudo descrevemos pela primeira vez centenas de LincRNAs associados fisicamente ao receptor de andrógeno após o tratamento com o hormônio. Observamos que uma parte desses ARA-lincRNAs pode modificar *in cis* a expressão de genes codificadores de proteínas vizinhos e essa capacidade de regulação é reforçada pela presença de marcas epigenéticas que são características de reguladores *in cis*. Além disso, mostramos que as regiões da cromatina contendo domínios topologicamente associativos (TADs) que possuem ARA-lincRNAs apresentam fatores de transcrição, marcas epigenéticas e nível de transcrição gênica diferenciadas. Os resultados apresentados neste trabalho estendem a importância dos lincRNAs durante eventos regulatórios complexos e mostra pela primeira a atuação dessas moléculas em sinergia com um fator de transcrição modificando a cromatina e alterando a regulação gênica.

ABSTRACT

The cell signaling events triggered in the presence of the androgen hormone in prostate cells is dependent on the activation of the androgen receptor (AR), which is a transcription factor encoded by the NR3C4 gene. AR when not bound to the androgen hormone is found in the cytoplasm, and binding of the hormone to AR promotes its displacement to the nucleus. AR binds to DNA motifs, promoting the transcription of certain genes through a complex regulatory mechanism not yet fully understood. In this work we explore the ability of long non-coding intergenic RNAs (lincRNAs) to bind to AR and contribute to the transcriptional regulation of genes. We used immunoprecipitation of RNA, followed by large-scale sequencing (RIP-Seq) for the identification and quantification of lincRNAs physically associated with AR in the LNCaP prostate cell line. In parallel we used transcriptome data and data on epigenetic marks to verify whether the physical interaction of lincRNAs with AR would influence the chromatin composition and the expression of genes neighboring these ARA-lincRNAs (Androgen Receptor Associated LincRNAs). As a result of this study we described for the first time in the literature hundreds of LincRNAs physically associated with the androgen receptor after treatment with the hormone. We have observed that part of these ARA-lincRNAs can modify in cis the expression of neighboring protein coding genes and this regulatory ability is enhanced by the presence of epigenetic marks that are characteristic of in cis regulators. In addition, we have shown that chromatin regions containing topologically associating domains (TADs) that possess ARA-lincRNAs have different transcription factors, epigenetic marks and gene transcription levels. The results presented in this work extend the importance of the lincRNAs during complex regulatory events and shows for the first time the performance of these molecules in synergy with a transcription factor modifying the chromatin and altering gene regulation.

Lista de abreviaturas e siglas

3C: Captura da Conformação da Cromatina

AR : Receptor de Andrógeno

ARA: lincRNA – LincRNA associado ao Receptor de Andrógeno

ARE: Elementos Responsivos ao Andrógeno

ChIP-Seq: Sequenciamento da Imunoprecipitação da Cromatina

DHT: Dihidrotestosterona

DNase-Seq: Sequenciamento das regiões da cromatina sensíveis a DNaseI

eRNA: RNA de enhancer

FDR: False Discovery Rate

FPKM: Fragmentos Por Quilobase de Exons Por Milhão de Fragmentos Mapeados

GTF: General Transfer Format

Hi-C: Captura da Conformação da Cromatina em larga escala “all-vs-all”

IgG: Imunoglobulina G

LincRNA: RNA Longo Não Codificador Intergênico

LNCaP: Lymph Node Carcinoma of the Prostate

LncRNAs: RNA Longo Não Codificador

ncRNAs: RNA Não Codificador

NonA-lincRNAs: LincRNA Não associado ao Receptor de Andrógeno

PRC2: Polycomb Repressive Complex 2

PSA: Proteína antígeno específica de próstata

RIP-Seq: Sequenciamento de RNAs imunoprecipitados

TAD: Domínios Topologicamente Associativos

TSS: Sítio de Início da Transcrição

SUMÁRIO

UNIVERSIDADE DE SÃO PAULO	1
Programa Interunidades de Pós-Graduação em Bioinformática	1
LUCAS FERREIRA DA SILVA	1
SÃO PAULO.....	1
LUCAS FERREIRA DA SILVA	2
São Paulo	2
Ficha catalográfica	3
AGRADECIMENTOS	4
RESUMO	7
ABSTRACT	8
Lista de abreviaturas e siglas	9
3C: Captura da Conformação da Cromatina	9
SUMÁRIO	10
1- INTRODUÇÃO	12
1.1- Câncer de próstata	12
1.2- Receptor de andrógeno	14
1.3- A participação da cromatina na expressão gênica	18
1.4- RNAs longos não codificadores de proteínas	24
1.5- Bioinformática e epigenética	28
1.5.1- ChIP-Seq	28
1.5.2- DNase-Seq	30
1.5.3- RIP-Seq	31
1.5.4- Captura da conformação da cromatina	32
1.5.5- Aplicação do algoritmo Random Forest na exploração de dados epigenéticos	34
A equação do valor de predição é a seguinte,	36
2- Objetivos	38
2.1- Objetivo Geral	38
2.2- Objetivos específicos	38
3- Materiais e métodos	39
3.1- Obtenção do transcriptoma de células LNCaP estimuladas com andrógeno	39
3.2- Obtenção dos RNAs associados ao receptor de andrógeno em células LNCaP estimuladas com andrógeno	39

3.3-	Análise do transcriptoma de células LNCaP estimuladas com andrógeno	40
3.4-	Expandindo o catalogo de lincRNAs.....	42
3.5-	Descoberta de novos lincRNAs expressos em LNCaP	43
3.6-	Transcriptoma de tumores de próstata	43
3.7-	Análise de enriquecimentos de assinaturas genicas com a ferramenta GSEA	44
3.8-	Descoberta de lincRNAs associados ao receptor de andrógeno.....	44
3.9-	Determinação de um grupo de lincRNAs não associados ao receptor de andrógeno.....	45
3.10-	Análise dos genes vizinhos dos ARA-lincRNAs	46
3.11-	Análises de dados públicos de epigenética	47
3.12-	Extração de características epigenéticas nos TSSs dos lincRNAs associados ao Receptor de Andrógeno.....	48
7P:	Verdadeiros positivos	49
3.13-	Análise dos domínios topologicamente associativos.....	50
4-	Resultados	52
4.1-	Descoberta de novos lincRNAs no transcriptoma de próstata	52
4.2-	O estímulo com andrógeno promove mudanças na expressão gênica de LNCaP	57
4.3-	LincRNAs em amostras de pacientes com câncer de próstata	61
4.4-	Identificação de lincRNAs associados ao receptor de andrógeno	63
4.5-	Os ARA-lincRNAs podem regular os genes codificadores vizinhos	65
4.6-	O perfil epigenético dos ARA-lincRNAs e dos genes codificadores vizinhos.....	71
4.8-	Promotores de ARA-lincRNAs mostram semelhanças com regiões contendo <i>enhancers</i> ativos	86
5-	Discussão	88
6-	Conclusões	101
	Resumidamente, o estudo desta tese permite concluir que:	101
7-	Referências bibliográficas.....	103

1- INTRODUÇÃO

1.1- Câncer de próstata

O câncer de próstata está entre as doenças mais comumente encontradas em adultos. De acordo com a publicação *Global Burden of Disease Cancer*, 1,6 milhões de homens foram diagnosticados com a doença no mundo inteiro no ano de 2015 (Global Burden of Disease Cancer *et al.*, 2016). Em 2016, no Brasil, a projeção de novos casos de câncer de próstata é de aproximadamente 61.200 (INCA, 2016) e no ano de 2013 foram constatadas 13.772 mortes pela doença (SIM).

Uma das características que transformam o câncer de próstata em uma doença com um elevado nível de mortalidade é seu potencial de gerar metástases. A porcentagem de pacientes com câncer de próstata que são diagnosticados com metástases varia de 10% a 20%, uma vez que o aparecimento da metástase pode acontecer anos após o tratamento cirúrgico ou do tratamento radioterápico (Balk, 2003).

Um molécula bastante estudada, que possui várias implicações clínicas para essa doença é a proteína antígeno próstata-específico (PSA). O PSA é importante para o diagnóstico inicial e pode ser utilizada para o monitoramento de uma possível recorrência do câncer de próstata (Paller e Antonarakis, 2013; Roth *et al.*, 2016). O PSA é uma glicoproteína produzida pelas células epiteliais da próstata e o gene codificador dessa proteína é conhecido como calicreína-3 (*KLK3*) (Ménez *et al.*, 2008). No sangue de homens saudáveis é encontrada apenas uma pequena quantidade dessa proteína, porém, seu nível pode ser elevado em homens com câncer de próstata. Dentre os motivos responsáveis pelo aumento do PSA, podemos citar causas moleculares como

mutações da região promotora do gene (Lai *et al.*, 2007) e fisiológicas como o extravasamento da proteína a partir da ruptura de capilares e tecidos do lúmen da próstata (Catalona *et al.*, 1994).

Os hormônios andrógenos, que possuem como um importante representante a dihidrotestosterona (DHT), são cruciais para o entendimento do câncer de próstata. Em seres humanos, a DHT é sintetizada pela enzima α -reductase, a partir da conversão da testosterona gerada nos testículos e glândulas adrenais (Carson e Rittmaster, 2003).

O transporte e ligação da DHT nas células da próstata pode modular a expressão de genes codificadores de proteínas e desencadear a sinalização de processos decisivos para o crescimento e manutenção da próstata normal, bem como a progressão do câncer de próstata (Lamont e Tindall, 2011).

Um modelo de estudo muito utilizado para entender a resposta ao andrógeno fora das amostras de tumores de pacientes é a linhagem LNCaP. Essa linhagem celular é derivada de uma metástase do câncer de próstata encontrada no linfonodo supraclavicular de um homem de 50 anos (Horoszewicz *et al.*, 1983).

As células da linhagem LNCaP após serem estimuladas por andrógeno (ou com um análogo) se comportam de forma similar à de tumores de próstata responsivos à andrógeno. A similaridade desse comportamento se dá principalmente pelo desencadeamento de respostas específicas, como o aumento da proliferação celular e da produção de fosfatases (Horoszewicz *et al.*, 1983). Semelhanças como essas tornam a linhagem LNCaP um interessante modelo para o estudo de genes regulados pelo estímulo do hormônio andrógeno, uma vez que uma parte do conhecimento obtido por

meio dessa linhagem é útil para o estudo de modelos tumorais *in vivo* (Beltran *et al.*, 2011).

1.2- Receptor de andrógeno

A sinalização celular desencadeada na presença do hormônio andrógeno é dependente da ativação do receptor de andrógeno (AR), que é um fator de transcrição codificado pelo gene NR3C4 (Receptor Nuclear, subfamília 3, grupo C, gene 4) (Mangelsdorf *et al.*, 1995). O receptor de andrógeno pertence à família dos receptores nucleares de hormônios esteroides (Nuclear Receptors Nomenclature, 1999) e em células normais da próstata, na ausência de ligantes é predominantemente localizado no citoplasma (Roy *et al.*, 2001).

O receptor de andrógeno é uma proteína de 919 amino ácidos e possui um peso molecular de 110 kDa (Trapman *et al.*, 1988; van Laar *et al.*, 1989). A estrutura funcional do receptor de andrógeno é composta por três domínios principais, (i) Domínio N-terminal (resíduos de aminoácidos 1-555), (II) Domínio ligante de DNA (resíduos 555-623) e (III) Domínio ligante C-terminal (655-919). O domínio C-terminal é unido ao domínio ligante de DNA por uma região de dobradiça flexível (resíduos 623-665) (Mangelsdorf *et al.*, 1995).

O domínio N-terminal compreende mais da metade dos aminoácidos dessa proteína (resíduos 1-555) e até o momento não possui a sua estrutura tridimensional resolvida através de cristalografia de raio-X (Tan *et al.*, 2015). Alguns estudos apontam para o domínio N-terminal como uma estrutura capaz de interagir com diversas

proteínas, além de possuir uma estrutura similar àquela dos domínios intrinsecamente desordenados (McEwan e Gustafsson, 1997; Bevan *et al.*, 1999; McEwan, 2012).

O domínio N-terminal contém os motivos Tau-1 (FQNLF) e Tau-5 (WHTLF), que permitem a interação do domínio C-terminal com o domínio N-terminal (Callewaert *et al.*, 2006). Entre os efeitos da interação entre esses dois domínios estão a ativação de vários genes dependentes de andrógeno e o aumento da estabilidade dos dímeros de AR, o que gera uma diminuição na taxa de dissociação desses complexos do DNA (Doesburg *et al.*, 1997; Langley *et al.*, 1998).

O domínio ligante C-terminal é responsável pela interação com os ligantes (agonistas ou antagonistas) e possui sua estrutura cristalográfica bem resolvida (Cantin *et al.*, 2007).

O domínio ligante de DNA é uma região rica no aminoácido cisteína sendo esta uma característica altamente conservada na família dos hormônios receptores de esteroides. Essa região possui dois motivos *zinc-finger*, compostos por quatro cisteínas cada (Tan *et al.* 2015).

O receptor de andrógeno funciona como um dímero se ligando à alvos no DNA por meio de dois sítios hexaméricos (5'-AGAACA-3') que são separados por 3 pares de bases (Shaffer *et al.*, 2004). As sequências reconhecidas pelo receptor de andrógeno são idênticas àsquelas reconhecidas pelos receptores de estrógeno, glicocorticoide, progesterona e mineralocorticoide. Apesar da redundância do motivo de DNA reconhecido por esses diferentes receptores, algumas variações nessa sequência (ex: 5'-CCAAGA-3') podem ser reconhecidas especificamente pelo receptor de andrógeno. Essas sequencias são chamadas de elementos responsivos à andrógeno (AREs) e

ajudam a explicar a diferença na resposta celular, quando uma mesma célula é estimulada com esses diferentes hormônios (Claessens *et al.*, 1996). Apesar da existência de uma resposta particular para cada hormônio, ainda não é bem entendido como os outros receptores não reconhecem os AREs (Shaffer *et al.*, 2004).

Massie *et al.* (2007) realizou um ensaio de imunoprecipitação da cromatina usando anticorpo anti-AR, seguido da hibridização do DNA co-imunoprecipitado em um chip de microarranjo de DNA contendo sequências promotoras de 24.275 genes (ChIP-chip) e, com isso, identificou >1500 sítios no genoma humano ocupados pelo AR em células LNCaP. Curiosamente, mais de 50% dos sítios ocupados pelo receptor de andrógeno não possuíam os motivos comuns nos AREs (Massie *et al.*, 2007) sugerindo um complexo mecanismo de reconhecimento dos sítios alvos do AR.

O receptor de andrógeno quando não ligado ao hormônio andrógeno é encontrado no citoplasma associado ao *foldsome*. Esse complexo proteico citoplasmático é composto por diversas chaperonas e proteínas de choque térmico (HSPs), que possuem a função de estabilizar e manter o receptor de andrógeno apto a interagir com moléculas de hormônio que cheguem até a célula por meio da circulação sanguínea (Pratt e Toft, 1997; Heinlein e Chang, 2001). Dentre as proteínas presentes no *foldsome* é possível encontrar as bem estudadas HSP90, HSP70, HSP40 e a proteína auxiliadora de chaperona p23 (Reebye *et al.*, 2012).

A ligação do AR ao DHT promove uma mudança tridimensional na estrutura do AR o que permite sua interação com a chaperona HSP27. O AR ligado ao DHT recruta a proteína quinase p37, que rapidamente fosforila a chaperona HSP27. A forma ativa da HSP27 libera o AR do complexo por meio do deslocamento da HSP90 do *foldsome*. O

receptor de andrógeno, agora dissociado do *foldosome* é então translocado para o núcleo com ajuda da HSP27 fosforilada (Zoubeidi *et al.*, 2007).

No núcleo, o AR regula a expressão de diversos genes alvos pelo recrutamento de coativadores, corepressores e de proteínas do complexo remodelador e modificador da cromatina (Doesburg *et al.*, 1997; Shang *et al.*, 2002; Perissi *et al.*, 2010). Um das proteínas que está presente em parte das regiões ligadas pelo receptor de andrógeno é a FOXA1 (Jin *et al.* 2014). Esse fator de transcrição da família *forkhead* possui um domínio de ligação ao DNA semelhante ao de *linker-histones* (Cirillo *et al.*, 2002) e devido a sua capacidade de se ligar a porções bem compactas da cromatina, ele age pioneiramente, descompactando o DNA e facilitando a ligação do receptor de andrógeno aos seus alvos (Shim *et al.*, 1998).

Além de FOXA1, o fator de transcrição NKX3-1, que possui um domínio *homedomain* é outra importante proteína que se co-localiza com as regiões do DNA ligadas ao AR. Curiosamente, cerca de aproximadamente 90% dos sítios ocupados simultaneamente por AR e NKX3-1, também são ocupados por FOXA1 (Tan *et al.*, 2012). A atuação do NKX3-1 é crucial para o desenvolvimento da próstata normal e tumoral, sendo já descrito como importante regulador da inibição da apoptose nessas células (Bhatia-Gaur *et al.*, 1999; Shang *et al.*, 2002; Tan *et al.*, 2012). A presença dessa proteína nesses sítios ativa positivamente a transcrição dos genes alvos do AR e seu silenciamento decresce a expressão desses genes (Tan *et al.*, 2012).

Algumas proteínas quando associadas ao receptor de andrógeno podem modificar covalentemente as histonas presentes nas regiões alvo do AR, como é o caso da proteína quinase PKN1. Essa proteína fosforila a treonina 11 da histona 3 (H3T11P)

gerando uma marca característica de ativação da transcrição (Metzger *et al.*, 2008) que em seguida promove o recrutamento da lisina desmetilase KDM4C/JMJD2C, a enzima que cataliza a desmetilação da histona 3 na lisina 9 (H3K9me3), removendo esta marca típica de promotores no estado reprimido.

A proteína WDR5 (*WD repeat-containing protein 5*) é uma proteína essencial para a proliferação das células LNCaP e sua depleção bloqueia a ativação dos genes alvos do receptor de andrógeno. O trabalho de Kim *et al.* (2014) mostrou que WDR5 interage diretamente *in vitro* com H3TP11P e é co-localizada com regiões regulatórias dos genes regulados por andrógeno *in vivo* (Kim *et al.*, 2007). Esse mesmo trabalho mostrou que, em resposta ao tratamento com andrógeno, PKN1 e H3T11P facilitam o recrutamento de WDR5 associada ao complexo MLL1. O complexo MLL1 é responsável por trimetilar as histonas H3K4 dos promotores dos genes alvo do AR, gerando assim a ativação da transcrição desses genes (Kim *et al.*, 2014; Sun *et al.*, 2016).

1.3- A participação da cromatina na expressão gênica

A regulação da expressão gênica dos eucariotos é controlada por uma grande quantidade de elementos, que interagem em diferentes níveis de especificidade, organização e sinergia. Uma das maneiras pelas quais o programa transcricional é orquestrado nas células é por meio de processos epigenéticos, como é o caso das modificações na cromatina (Stricker *et al.*, 2016).

O processo de regulação da expressão gênica através das modificações na cromatina possui um caráter plástico e dinâmico, uma vez que essas marcas são

adicionadas, lidas e removidas por enzimas modificadoras da cromatina de uma forma altamente regulada (Falkenberg e Johnstone, 2014). As modificações nas histonas alteram a estrutura da cromatina de forma não-covalente, podendo gerar tanto mudanças na conformação de nucleossomos individuais, bem como modificações em diversos nucleossomos vizinhos simultaneamente gerando a compactação ou afrouxamento da cromatina nessas regiões (Kouzarides, 2007).

Além de atuarem no nível de compactação da cromatina, essas modificações também podem criar regiões que permitem proteínas com domínios especializados reconhecerem e recrutarem outros complexos modificadores de cromatina, fatores de transcrição, proteínas de replicação ou proteínas de reparo de DNA (Harmston e Lenhard, 2013).

Os mais bem documentados sítios de modificações de histonas são os resíduos de lisinas (K). As histonas H3K4 dimetiladas e trimetiladas (H3K4me2 e H3K4me3) são frequentemente associadas à promotores de genes ativos que estão presentes em regiões de eucromatina. Já as modificações H3K9me3, H3K27me3 e H4K20me3 são encontradas em regiões de heterocromatina e preferivelmente em genes transcricionalmente inativos (Barski *et al.*, 2007). Contudo, as metilações não são encontradas apenas nos promotores do genes. Muitas delas são vistas em regiões intergênicas e outras são frequentemente encontradas no corpo dos genes, como é o caso da H3K36me3, que possui grande sobreposição com exons que possuem *splicing* diferencial (Schwartz *et al.*, 2009; Luco *et al.*, 2010).

Além das metilações, as acetilações nas caudas das histonas também podem modular a regulação da transcrição gênica dos eucariotos. Uma marca pouco estudada,

porém fortemente presente em promotores transcritionalmente ativados, é a marca H4K5ac (Park *et al.*, 2013). Interessantemente, essa marca também é correlacionada com os genes que retomam rapidamente a transcrição logo após a divisão celular, indicando que essa marca contribui para um interessante mecanismo de propagação do programa celular entre células em divisão (Zhao *et al.*, 2011).

Um grande esforço, principalmente em áreas emergentes como a epigenética e transcriptômica têm se voltado para entender o funcionamento da regulação da transcrição gênica em tecidos e células normais, bem como a desregulação desses processos durante diferentes estados patológicos (Holoch e Moazed, 2015; Keung *et al.*, 2015)

Já é bem descrita na literatura a importância da contribuição dos promotores para a iniciação da transcrição genica, principalmente o seu papel no recrutamento de fatores que direcionam e estabilizam a maquinaria da RNA polimerase II (Butler e Kadonaga, 2002). Porém, nos eucariotos é bem comum a presença de regiões reguladoras auxiliares que não são situadas no promotores e a ausência dessas regiões adicionais gera a perturbação na expressão gênica (Lovén *et al.*, 2013). Essas regiões regulatórias distais são chamadas de *enhancers* ou módulos regulatórios *in cis* (Shlyueva *et al.*, 2014).

Enhancers ativos são tipicamente destituídos de nucleossomos; dessa forma, a região de DNA na qual o *enhancer* é localizado fica acessível e pode ser ligada por diferentes complexos. Um tipo de proteína comumente encontrada em regiões de *enhancers* são as proteínas que codificam os fatores de transcrição (Thurman *et al.*,

2012). Esses podem guiar para essas regiões co-ativadores e co-repressores importantes para a função regulatória dos *enhancers* (Shlyueva *et al.*, 2014).

Os nucleossomos situados na vizinhança de *enhancers* ativos parecem tipicamente conter histonas com modificações específicas. Entre as marcas relacionadas com elementos *enhancers* estão a mono-metilacção da lisina 4 da histona 3 (H3K4me1) e a acetilação da lisina 27 da histona 3 (H3K27ac) (Rada-Iglesias *et al.*, 2011). Além dos *enhancers* ativos existem aqueles *enhancers* que são denominados como *poised enhancers*. Os *poised enhancers* possuem algumas características de *enhancer*, como regiões sensíveis à DNase, presença da marca H3K4me1 e H3K4me2, porém, possuem uma quantidade baixa de H3K27ac e apresentam um aumento na marca repressora H3K27me3 (Calo e Wysocka, 2013). Alguns estudos sugerem que os *poised enhancers* necessitem de fatores adicionais para que se tornem ativos e isso é evidenciado pela sua incapacidade de gerar a transcrição nos genes vizinhos até que se transformem em *enhancers* ativos (Rada-Iglesias *et al.*, 2011).

Os *enhancers* podem regular a expressão gênica de seus alvos, mesmo situados a uma grande distância dos promotores desses genes. O que torna possível essa regulação é a presença de contatos (*loops*) entre a porção da cromatina contendo o *enhancer* e a porção contendo o promotor do gene regulado (Dowen *et al.*, 2014).

Em sua maioria, os organismos eucariotos possuem a cromatina composta por nucleossomos formados por cinco histonas mais comuns (H1, H2A, H2B, H3, H4), porém, variantes de histonas podem ser encontradas com diferentes frequências em diferentes organismos e contextos biológicos (Kamakaka e Biggins, 2005). Um exemplo de histona não convencional é a histona H2A.Z, que é uma variante da clássica histona H2A, que é

codificada pelo gene H2AFZ. A histona H2AZ é comumente encontrada em regiões da cromatina ligadas por fatores de transcrição, tais como: ER α , MYC e p53 (Gévry *et al.*, 2007; Gévry *et al.*, 2009), contudo, as metilações de DNA parecem ser infrequentes nessas regiões (Zilberman *et al.*, 2008). A variante de histona H2A.Z pode ser distribuída tanto no final do corpo de genes transcricionalmente competentes como em sobreposição com marcas H3K4me, localizadas nos *enhancers* de vários tipos celulares em diversos estágios de diferenciação (Sevilla e Binda, 2014).

Técnicas recentes de captura da conformação da cromatina (3C, Hi-C, ChIA-PET) tem indicado que a distribuição da cromatina no núcleo dos eucariotos é extremamente organizada e está diretamente relacionada com a regulação da expressão genica (Schmitt *et al.*, 2016). Dessa forma, a organização global do núcleo das células demonstra um elevado nível de compartimentalização e hierarquização, que se correlaciona com o programa transicional tecido-específico, além de indicar o estado de ativação da cromatina no genoma (eucromatina ou heterocromatina) (Rao *et al.*, 2014).

Os *topologically associating domains* (TAD) são pequenas estruturas que compartimentalizam a cromatina em densas regiões que possuem em média ~1 MB. Dentro dessas regiões é encontrado um elevado nível de contato entre as porções da cromatina que compartilham um mesmo TAD. Outra particularidade desses módulos estruturais da cromatina é a forte ausência de contatos nas regiões que demarcam as fronteiras ou bordas dos TADs (Rao *et al.*, 2014).

As bordas dos TADs são extremamente conservadas entre diferentes tipos celulares, contudo, o perfil de interações dentro das TADs é bastante variado entre os diferentes tipos celulares (Dixon *et al.*, 2012; Le *et al.*, 2013).

Apesar da literatura a respeito da estrutura tridimensional da cromatina estar em grande crescimento, ainda são pouco conhecidos os mecanismos que definem, criam e mantêm os TADs. Algumas proteínas como o CTCF, coesina e condensina possuem função associada com a formação de *loops* na cromatina e também com o desaparecimento das bordas de alguns TADs (Dixon et al. 2012; Ong & Corces 2014; Merkenschlager & Nora 2016). Essas proteínas podem ser encontradas em mais de 86% de todas as regiões que ancoram a base dos *loops* da cromatina (Rao et al. 2014).

Diferentes estudos mostraram que perturbações na cromatina, como por exemplo os rearranjos que acontecem em células do câncer, de alguma forma podem reestruturar os TADs causando alterações na dinâmica das interações de longa distância (Gröschel *et al.*, 2014). Esses eventos podem acontecer pela destruição da rede de contatos que existem no interior de um TAD, o que geraria uma perturbação na regulação intra-TAD específica. Outra consequência da rompimento da estrutura dos TADs seria a perda do isolamento de promotores situados no interior de um TAD. Esse fenômeno permitiria a ativação indesejada desses genes, por meio do contato dos promotores com *enhancers* externos que anteriormente estavam ilhados pelas bordas do TAD (Lupiáñez *et al.*, 2015).

Grande parte das regiões da cromatina que são ocupadas pelo receptor de andrógeno apresentam sobreposição com regiões que possuem características de *enhancers*. Esse acontecimento é um indicador da importância da ação direta do receptor de andrógeno nos *loci* de elementos regulatórios (Hsieh *et al.*, 2014).

A estrutura complexa da regulação dos eucariotos faz com que estudos que integrem dados de epigenética, como aqueles das marcas de histona, fatores de

transcrição e informações da própria dinâmica tridimensional da cromatina, sejam uma boa alternativa para um melhor entendimento da regulação gênica após o estímulo hormonal feito com andrógeno.

1.4- RNAs longos não codificadores de proteínas

Na última década ocorreram vários avanços nas técnicas de sequenciamento em larga escala que possibilitaram a investigação do transcriptoma e do genoma humano em alta resolução (Goodwin *et al.*, 2016). A integração dessas análises revelou que grande parte do genoma humano (> 86%) é transcrito em algum tecido, linhagem celular ou contexto biológico. Essa transcrição massiva do genoma foi nomeada como *transcrição* pervasiva e trouxe com ela um vasto repertório de RNAs não-codificadores de proteínas (ncRNAs) que não eram antes conhecidos (Djebali *et al.*, 2012; Hangauer *et al.*, 2013). Embora a literatura seja dominada por trabalhos de ncRNAs curtos (< 200 nt), nos últimos anos tem havido um aumento no número de trabalhos que têm reportado a funcionalidade e diversidade de ncRNAs longos ($\geq$ 200 nt) (lncRNAs) que mapeiam em regiões intrônicas e intergênicas (ENCODE, 2007; Dinger *et al.*, 2009; Kapranov *et al.*, 2010; Djebali *et al.*, 2012).

Os lncRNAs apresentam características semelhantes às dos mRNAs de genes codificadores de proteínas, tal como a presença da estrutura *cap* na ponta 5' do RNA, e possuem região promotora e marcas de RNA Polimerase II no sítio de início de transcrição (TSS) (Guttman *et al.*, 2009; Conley e King Jordan, 2012). Além disso, alguns trabalhos têm mostrado que esses transcritos podem apresentar uma cauda poli-

adenilada e podem se acumular diferentemente no compartimento nuclear ou citoplasmático das células (Kapranov *et al.*, 2007; Kapranov *et al.*, 2010).

Apesar de existir um grande repertório de lncRNAs no transcriptoma humano, estes apresentam de maneira geral uma baixa abundância de expressão quando comparados aos genes codificadores de proteínas. Outra particularidade dos lncRNAs é que a grande maioria apresenta expressão tecido específica (ENCODE, 2007; Cabili *et al.*, 2011).

Evidências têm apontado o envolvimento de lncRNAs com diversas doenças humanas (Taft *et al.*, 2010; Esteller, 2011; Wapinski e Chang, 2011), especialmente nos tumores de diversos tipos celulares (Prensner e Chinnaiyan, 2011; Reis e Verjovski-Almeida, 2012; Spizzo *et al.*, 2012).

De maneira geral, há fortes evidências que o mecanismo de ação dos lncRNAs está relacionado com interação entre os lncRNAs e proteínas (Khalil e Rinn, 2011), especialmente proteínas que compõem complexos remodeladores da cromatina que são ativadores ou repressores da expressão gênica (Khalil *et al.*, 2009).

Alguns trabalhos sugerem que além de interagir com complexos proteicos remodeladores da cromatina, os lncRNAs se ligam ao DNA formando um híbrido RNA/DNA que guia os complexos para sítios específicos do genoma onde executam sua função (Chu *et al.*, 2011; Mattick, 2012). Outros trabalhos indicam que os lncRNAs também podem promover ativação de determinados complexos proteicos através da indução de modificações alostéricas (Mattick, 2012).

Um exemplo muito bem caracterizado de interação física entre lncRNA-proteína é o que acontece com lncRNA *HOTAIR* (Rinn *et al.*, 2007). Esse gene é transcrito no locus

HOXC e regula a expressão do cluster de genes *HOXD* que controla a morfogênese dos membros superiores e inferiores dos vertebrados (Hérault *et al.*, 1998). No trabalho de Rinn *et al.* (Rinn *et al.*, 2007) foi observado que *HOTAIR* se liga ao *Polycomb repressive complex 2* (PRC2) e direciona esse complexo para a região promotora do locus *HOXD* (Rinn *et al.*, 2007). O complexo repressivo Polycomb 2 então trimetila a histona 3 no resíduo de lisina 27 (H3K27me3) resultando no silenciamento dos genes do cluster *HOXD*.

Por meio de sua porção 5', *HOTAIR* interage com a proteína SUZ12, uma integrante do PRC2, enquanto isso a porção 3' interage com a proteína LSD1, que é integrante do CoREST/REST *repressor complex*. O complexo CoREST/REST retira a dimetilação da histona H3 no resíduo de lisina 4 (H3K4me2) e isso se relaciona com a repressão da expressão gênica e com o silenciamento dos genes *HOXD* (Tsai *et al.*, 2010).

O lincRNA *Xist* é outro exemplo de um lncRNA que interage fisicamente com complexos proteicos e regula processos celulares cruciais, como é o caso da inativação de um dos cromossomos X em fêmeas. Durante esse evento, modificações epigenéticas que geram a inativação do cromossomo são primeiramente depositadas em porções do DNA que são co-localizadas com o locus do lincRNA *Xist*, no centro de inativação. Quando expresso, *Xist* se liga ao PRC2 e o guia para alvos específicos no DNA. À medida que a inativação progride, as regiões a serem desativadas se aproximam progressivamente do centro de inativação onde *Xist* é expresso. Dessa forma, a inativação do cromossomo X é um elegante exemplo de como a plasticidade da

cromatina é explorada pela ação coordenada do lncRNA *Xist* em parceria com o complexo PRC2 (Simon *et al.*, 2013).

Recentemente foi descoberta uma nova classe de ncRNAs que podem regular a transcrição de genes vizinhos; esses transcritos são conhecidos como RNAs de *enhancer* (eRNA) (Kim *et al.*, 2010). Os eRNAs em sua maioria não são poliadenilados, apresentam transcrição bidirecional e comportam a modificação H3K4me1 em seus *loci* (Kim *et al.*, 2010; Lai e Shiekhhattar, 2014). O aumento da expressão de alguns eRNAs após o estímulo com andrógeno (Wang *et al.*, 2011) , bem como a regulação da expressão do gene *KLK3* pelo eRNA (eRNA-KLK3) já estão documentados na literatura (Hsieh *et al.*, 2014). Dessa forma, esses trabalhos abrem o caminho para o entendimento do papel dessa nova classe de ncRNA durante a regulação do programa transcricional das células. Apesar da expressão dos eRNAs ser importante para a regulação de genes *in cis*, os lincRNAs não poliadenilados e sem transcrição bidirecional são também relacionados com a regulação da expressão de genes vizinhos e podem apresentar marcas semelhantes à dos *enhancers* (Orom *et al.*, 2010; Iltott *et al.*, 2014).

Alguns RNAs não codificadores já foram encontrados interagindo fisicamente com fatores de transcrição (Cassiday, 2002; Hsieh *et al.*, 2014). O *NRCP* é um dos lncRNAs que apresentam maior nível de expressão diferencial quando ovários tumorais são comparados com ovários normais (Rupaimoole *et al.* 2015). O fator de transcrição STAT1 se liga fortemente a esse lncRNA e o silenciamento de *NRCP* diminui a interação de STAT1 com a RNA polimerase II. O resultado da não interação entre STAT1 e *NRCP* é a diminuição da expressão dos genes da via glicolítica, que normalmente estariam regulados por STAT1. O não funcionamento dessa via propicia a respiração mitocondrial

nesses tecidos e favorece o desenvolvimento dos tumores (Rupaimoole et al. 2015). Apesar de existirem alguns exemplos de ncRNAs que interagem com fatores de transcrição, ainda é pequena a quantidade de trabalhos que procuram por esse tipo de associação, principalmente por meio de abordagens em larga escala.

Em resumo, têm se acumulado cada vez mais evidências de que os lncRNAs, juntamente com a arquitetura tridimensional da cromatina podem exercer papel crucial na regulação do funcionamento de células e tecidos (Nichols e Corces, 2015; Yang *et al.*, 2015). Sendo assim, é indispensável a utilização de metodologias que permitem o estudos de novos lncRNAs e de lncRNAs já anotados que possuem associação com proteínas parceiras, sejam elas fatores de transcrição e/ou componentes da maquinaria epigenética.

1.5- Bioinformática e epigenética

1.5.1- ChIP-Seq

Cada vez mais tecnologias tem possibilitado a compreensão do envolvimento da cromatina e das marcas de histonas para a diferenciação e função de diversos tipos celulares. Sabendo da importância desses elementos, tanto iniciativas específicas de certos grupos de laboratórios como laboratórios isolados geram dados que revelam a posição de nucleossomos (Segal e Widom, 2009), definem a conformação da cromatina (Wit e Laat, 2012) e determinam sítios na cromatina ocupados por fatores de transcrição (Farnham, 2009). Projetos maiores como o Roadmap Epigenomics Consortium e o Projeto ENCODE geraram perfis extensos de múltiplas marcas epignéticas em diversos

tecidos e linhagens celulares humanas (ENCODE, 2007; Roadmap Epigenomics *et al.*, 2015). Um dos principais métodos para a identificação de marcas epigenéticas de um genoma é a fragmentação do DNA genômico acompanhada de imunoprecipitação da cromatina que permanece ligada a uma proteína específica de interesse, com ou sem uma modificação química. Após isso é realizado o sequenciamento em larga escala da fração de DNA co-imunoprecipitado. Essa técnica é conhecida como Chromatin Immunoprecipitation – Sequencing (ChIP-Seq) (Furey, 2012).

A imunoprecipitação de fatores de transcrição em experimentos de ChIP-Seq é bastante utilizada para estudos epigenéticos em larga escala (Tan *et al.*, 2012; Jin *et al.*, 2014; Ong e Corces, 2014; Dixon *et al.*, 2015). Esse tipo de análise permite a captura das regiões da cromatina especificamente ligadas a um fator de transcrição selecionado após um dado estímulo. Além de permitir a visualização do mapa de ligações do fator de transcrição, essa abordagem auxilia na descoberta dos motivos de DNA envolvidos nessa interação (Farnham, 2009; Heinz *et al.*, 2010).

Outra variação comum que essa técnica permite é a imunoprecipitação de histonas contendo uma modificação química específica (O’Geen *et al.*, 2011). Esse tipo de abordagem permite a observação das marcas de histonas em diferentes contextos celulares, tais como o de divisão celular, reparo do DNA, durante a transcrição e repressão gênica (Roadmap Epigenomics *et al.*, 2015). Quando várias marcas de histona que foram obtidas em um mesmo contexto biológico são integradas, estados da cromatina mais sutis e específicos são revelados. Esses estados geralmente são compostos pela combinação de múltiplas marcas de histonas e podem ser modelados através de algoritmos de extração de sinais e modelos probabilísticos (Ernst *et al.*, 2011).

A análise desse tipo de dado é realizada por meio do alinhamento das sequências obtidas no genoma de referência escolhido. Entre os programas de alinhamento que permitem essa operação estão os alinhadores Bowtie (Langmead *et al.*, 2009) e STAR (Dobin *et al.*, 2013). Existem algumas ferramentas que possuem funções específicas para o tratamento de dados de ChIP-Seq, como por exemplo o programa HOMER (Heinz *et al.*, 2010) e MACS (Zhang *et al.*, 2008), que além de corrigirem alguns *bias* presentes nesse tipo de técnica podem anotar picos e motivos enriquecidos em cada experimento.

1.5.2- DNase-Seq

Outra importante técnica utilizada para a investigação de elementos regulatórios no genoma é o ensaio em larga escala da clivagem das regiões hipersensíveis à DNaseI (DNase-Seq) (Song e Crawford, 2010). A partir dos resultados de DNase-Seq é possível encontrar promotores, enhancers, silenciadores, *insulators* e outras regiões controladoras (Boyle *et al.*, 2008; González *et al.*, 2015). Esse tipo de visualização só é possível devido a sensibilidade à clivagem por DNaseI existente em regiões genômicas cuja ocupação por nucleossomos é diminuída (Casci, 2006).

Para a obtenção de dados de DNase-Seq o isolamento da cromatina é realizado e após isso é induzida sua digestão com a DNaseI. Os fragmentos digeridos são então separados por tamanho, amplificados e sequenciados. Os *reads* gerados através desse procedimento são alinhados contra o genoma de referência (Song e Crawford, 2010). A ferramenta HOMER também pode ser utilizada para identificar regiões em que os *reads* se aglomeram e formam picos. Com as informações desses picos é possível observar segmentos da cromatina que possuem grande hipersensibilidade a DNaseI, postulando

a existência de elementos regulatórios e descobrindo os motivos mais frequentes nessas regiões (González *et al.*, 2015).

1.5.3- RIP-Seq

É possível por meio de experimentos em larga escala determinar quais são os RNAs ligados a uma dada proteína. Uma das técnicas que oferecem essa possibilidade é a imunoprecipitação de RNA, seguida de sequenciamento (RIP-Seq) (Khalil *et al.*, 2009). Essa abordagem já foi utilizada por outros grupos para capturar RNAs interagindo com proteínas do Complexo Repressor Polycomb 2, proteína Argonauta e a proteína ligadora de RNA (RBP) Sm, que é envolvida com o *splicing* em *Drosophila* (Zhao *et al.*, 2010; Lu *et al.*, 2014; Krell *et al.*, 2016).

Resumidamente, no ensaio de RIP-Seq é utilizado um anticorpo com afinidade para uma proteína específica, que permite por meio da imunoprecipitação o isolamento e sequenciamento em larga escala das espécies de RNAs que previamente se encontravam ligadas a essa proteína e co-imunoprecipitaram (Zhao *et al.*, 2010). Para controlar o caráter ruidoso desse experimento é realizada uma imunoprecipitação de uma amostra semelhante, com um anticorpo sem especificidade para qualquer proteína (IgG). Os RNAs sequenciados que se ligaram ao IgG são utilizados como um controle negativo do experimento. Utilizando essa abordagem experimental, a quantidade de *reads* de cada transcrito de RNA imunoprecipitado junto a proteína de interesse é contrastada com a quantidade do mesmo RNA capturada através da imunoprecipitação do IgG (controle). Aqueles RNAs imunoprecipitados com a proteína de interesse, que

possuem um enriquecimento superior ao do anticorpo inespecífico são considerados “associados fisicamente à proteína de interesse”.

A identificação de RNAs significativamente ligados à proteína desejada pode ser realizada com o uso das mesmas ferramentas que calculam a expressão diferencial (Bonasio *et al.*, 2014; Chu *et al.*, 2015) ou por meio de ferramentas como o RIPSeeker (Li *et al.*, 2013).

1.5.4- Captura da conformação da cromatina

Técnicas que possibilitam a captura da conformação espacial da cromatina se tornaram cada vez mais populares na última década (Boyan e Giacomo, 2016). Algumas delas, como é o caso do 3C (**C**aptura da **C**onformação da **C**romatina) permitem a investigação do nível de contato entre diferentes partes da cromatina que foram pré-selecionadas. A pioneira abordagem de 3C foi muito importante para o desenvolvimento de novas técnicas que agora podem capturar a conformação da cromatina em larga escala (Wit e Laat, 2012). Essas novas abordagens possibilitaram a definição de mapas de alta resolução dos contatos que existem entre as diferentes partes da cromatina de um mesmo cromossomo e também a definição dos contatos presentes entre distintas regiões extracromossomais (Grob *et al.*, 2014).

Com essa imensa quantidade de novos dados de captura de conformação da cromatina foi possível destacar o papel da arquitetura da cromatina e da dinâmica criada por ela durante o controle da expressão gênica (Harmston e Lenhard, 2013) e da replicação dos eucariotos (Naumova *et al.*, 2013).

Uma das variações do 3C que permite estudar o mapa global de interações da cromatina é o Hi-C (Belton *et al.*, 2012). O primeiro passo experimental realizado nesse

método é o *cross-linking* entre a cromatina e as proteínas próximas. Essa etapa introduz um “congelamento” no estado atual da cromatina permitindo que as interações entre os *loci* e proteínas sejam conservadas. O segundo passo do Hi-C consiste em fragmentar o DNA por digestão e ligar a ponta dos fragmentos genômicos vizinhos usando adaptadores específicos. Essa etapa permite que a co-localização dos fragmentos seja revelada, pois em seguida o *cross-linking* entre as proteínas e os DNAs ligados pelos adaptadores presentes nas pontas dos fragmentos é revertida. Dessa forma, apenas os fragmentos que permaneceram ligados a um mesmo complexo proteico têm suas pontas unidas, e são especificamente capturados porque durante a ligação dos adaptadores é incorporado na junção um nucleotídeo biotinilado que serve para a captura.

Na última etapa, após a degradação das proteínas *em cross-linking* com o DNA, o sequenciamento de todos os fragmentos capturados é realizado e assim os segmentos da cromatina que possuíam proximidade dentro do núcleo podem ser anotados (Belton et al. 2012).

Além da elucidação dos mapas de contatos entre a cromatina, esses dados podem ser utilizados para a procura dos domínios topologicamente associativos (TAD) e de suas subestruturas, conhecidas também como sub-TADs (Boyan e Giacomo, 2016). Algumas abordagens computacionais ainda permitem a reconstrução por inteiro do arranjo tridimensional dos cromossomos dentro do núcleo. Para realizarem esse feito, essas ferramentas usam a frequência dos contatos de cromatina obtidos pelo ensaio de Hi-C, juntamente com a modelagem de propriedades físico-química desses polímeros (Lesne et al., 2014; Di Stefano et al., 2016).

1.5.5- Aplicação do algoritmo Random Forest na exploração de dados epigenéticos

Random Forest é um algoritmo de aprendizagem supervisionada que utiliza diversos sub-classificadores (árvores) em conjunto para gerar previsões mais acuradas (*ensemble learning*) (Dietterich, 2000). O algoritmo Random Forest é flexível e pode ser utilizado tanto em problemas de classificação como em problemas de regressão (Breiman, 2001).

Nos últimos anos tem se tornado cada vez mais difundida a utilização desse algoritmo para a análise de dados epigenéticos (Zilberman *et al.*, 2008; Heyn *et al.*, 2013; Rajagopal *et al.*, 2013; Dixon *et al.*, 2015; Polak *et al.*, 2015; Kim *et al.*, 2016). Um dos principais fatores que favorecem esse crescimento é a facilidade de recuperação e da avaliação da contribuição individual de cada variável para o processo de aprendizado dos dados de treinamento. Por esta razão decidimos utilizar este algoritmo para nossas análises, como descrito mais adiante.

Para realizar a tarefa de aprendizado, o algoritmo Random Forest cria uma floresta contendo diversas árvores de decisão. Cada uma das árvores recebe uma quantidade específica e aleatória dos dados originais do treinamento. As variáveis utilizadas para o treinamento também são aleatoriamente distribuídas para cada uma das árvores. Isso faz com que cada um dos sub-preditores (árvores) aprenda com um subgrupo de dados diferente, cujas variáveis utilizadas para a separação também sejam diferentes. Essa heurística faz com que o algoritmo Random Forest seja bastante resistente ao conhecido problema de *overfitting*.

Durante processo iterativo de criação dos galhos de uma árvore de decisão, o algoritmo Random Forest percorre os valores de uma variável e tenta minimizar o resultado da função de $GINI_{SPLIT}$. Dessa forma, o valor utilizado como ponto para a criação dos novos galhos procura diminuir a impureza dos componentes presentes nos novos subconjuntos gerados.

A equação abaixo formaliza o mecanismo de criação de novos galhos em uma árvore:

$$GINI_{SPLIT}(T) = \frac{N^1}{N} \times Gini(T_1) + \frac{N^2}{N} \times Gini(T_2), \quad (1)$$

na qual:

T = Conjunto de valores $T = \{v_1, v_2, \dots, v_n\}$ que possui correspondência com uma coleção de classes $C = \{j_1, j_2, \dots, j_n\}$.

T_1 e T_2 = Subconjuntos gerados por uma ramificação no ponto de corte S . O valor de S deve seguir o seguinte critério: $S > \min(v)$ e $S < \max(v)$.

N^1 e N^2 = Número de elementos nos subconjuntos T_1 e T_2 .

N = Número total de elementos dentro da coleção.

A equação utilizada para calcular o $Gini(T_1)$ e $Gini(T_2)$ é dada por:

$$Gini(T) = 1 - \sum_{j=1}^n (p_j)^2, \quad (2)$$

na qual:

p_j = Frequência da classe j no conjunto T_i

A etapa de classificação pode ser esquematizada na figura abaixo (Fig. 0), em uma floresta simplificada na qual se usou apenas 3 árvores preditoras.

Figura 0 – Esquema representando o processo de classificação pós-treinamento em um modelo simplificado com 3 árvores. Cada árvore em marrom, obtida na etapa de treinamento, recebe um mesmo valor V como entrada, para o qual pretende-se determinar a probabilidade de que ele pertença a uma das quatro diferentes classes representadas pelas cores (azul, amarela, vermelha e verde). Quando a entrada V a ser predita atinge um nó que não pode ser mais ramificado, o valor de probabilidade de V pertencer a cada classe é atribuído com base na ocupação final daquele nó durante o processo de treinamento (Criminisi et al. 2011).

A equação do valor de predição é a seguinte,

$$p(c|v) = \frac{1}{T} \sum_{t=1}^T p_t(c|v), \quad (3)$$

na qual:

$p(c|v)$: A probabilidade de um valor v pertencer a um classe c

T = Um conjunto de árvores presente na floresta

$(c|v)$ = A probabilidade de v **pertencer a** c para cada árvore t contida no conjunto de árvores T

No final da predição, o valor de classificação final é dado pela somatória das probabilidades encontradas em cada uma das 3 árvores presentes no modelo.

O valor da contribuição de cada variável V é calculado utilizando uma somatória do valor de saída da função $GINI_{SPLIT}$ de V , para cada vez que V foi utilizado em uma das árvores durante o treinamento. Sendo assim, a importância de uma variável é diretamente dependente do grau de pureza que ela proporciona durante a criação de novos galhos (Menze *et al.*, 2009)

2- Objetivos

2.1- Objetivo Geral

- Identificar em larga escala LincRNAs associados fisicamente ao receptor de andrógeno na linhagem LNCaP estimulada com o hormônio andrógeno e caracterizar um possível padrão de regulação da expressão gênica determinado por esses lincRNAs.

2.2- Objetivos específicos

- Identificar os LincRNAs existentes no transcriptoma da linhagem LNCaP tratada com o hormônio andrógeno e nos transcriptomas de tumores de pacientes.
- Identificar lincRNAs associados fisicamente ao receptor de andrógeno após a indução hormonal.
- Identificar as marcas epigenéticas da cromatina presentes nos sítios de transcrição dos lincRNAs associados fisicamente ao receptor de andrógeno após o estímulo com o hormônio e identificar as marcas e o nível de transcrição dos genes codificadores de proteína vizinhos aos lincRNAs.

3- Materiais e métodos

3.1- Obtenção do transcriptoma de células LNCaP estimuladas com andrógeno

A linhagem celular LNCaP foi obtida do ATCC (<https://www.atcc.org>) e cultivada por 24 horas em meio RPMI 1640 (Sigma) com privação de andrógeno e com 10 % (peso/volume) de soro fetal bovino tratado com carvão ativado (Sigma). Após 24 horas, as células foram mantidas por mais 24 horas em um meio contendo 1 nM do andrógeno sintético R1881 (tratamento) ou em um meio contendo um volume de etanol (veículo) equivalente ao que acomodaria o andrógeno (controle). O RNA poly(A)⁺ foi extraído usando o kit de isolamento de mRNA *FastTrack MAG maxi* (Invitrogen) seguindo o protocolo indicado pelo manual do fabricante, exceto por um tratamento estendido de 1 h com 2x DNaseI (Invitrogen). As bibliotecas fita-específicas de cDNA foram preparadas usando o kit TrueSeq RNA (versão 1, rev A) e sequenciadas com o equipamento Illumina HiSeq 2000 usando reads *paired-end* de comprimento de 75 pares de base.

Foram sequenciadas duas amostras das células tratadas com o andrógeno e duas tratadas apenas com o veículo. Essas amostras geraram de 80-90 milhões de *mate pairs* por amostra.

3.2- Obtenção dos RNAs associados ao receptor de andrógeno em células LNCaP estimuladas com andrógeno

As células da linhagem LNCaP foram cultivada por 24 horas em meio RPMI 1640 (Sigma) com privação de andrógeno e com 10 % (peso/volume) de soro fetal bovino

tratado com carvão ativado (Sigma). Após 24 horas, as células foram mantidas por mais 6 horas em um meio contendo 1 nM do andrógeno sintético R1881.

A imunoprecipitação nativa (sem *crosslink*) dos RNAs ligados ao receptor de andrógeno foi realizada usando o kit Magna RIP RNA-Binding Protein Immunoprecipitation kit (Millipore) seguindo as recomendações no manual do fabricante. A lavagem das amostras foi realizada utilizando o *RIP Wash Buffer* (Millipore) e após isso o lisado celular foi adicionado a um sistema de esfera-anticorpo. Os anticorpos conjugados com as esferas (*beads*) foram os seguintes: IgG de soro normal de coelho (Millipore # 12-370) ou Anti-Receptor de andrógeno (Millipore # 06-680). O RNA foi extraído utilizando Trizol e TURBO DNase (Ambion) a 37 graus por 30 minutos. A purificação foi realizada com o kit RNeasy Micro KIT (Qiagen) e quantificado com o kit RiboGreen (Invitrogen).

Para cada amostra de RNA do RIP nativo foi preparada uma biblioteca usando o kit SMARTer Ultra Low Input RNA Kit for Sequencing v3 (Clontech). As bibliotecas foram então sequenciadas usando o equipamento Illumina NextSeq500 com *paired-end reads* de 150 pares de base de comprimento. Para cada anticorpo (anti-AR ou IgG) foram geradas duas amostras possuindo em média 30 milhões de *reads* cada.

3.3- Análise do transcriptoma de células LNCaP estimuladas com andrógeno

A ferramenta Trimomatic (Bolger *et al.*, 2014) foi utilizada para remover adaptadores oriundos do processo de criação das bibliotecas, bem como durante a remoção das regiões dos *reads* cujos valores de qualidade média eram menores do que

15 (probabilidade de erro de leitura > 3.1%) em uma janela deslizante (Direção: 5' para 3') de 4 nucleotídeos. Os *reads* com comprimento menor que 20 nucleotídeos foram descartados e os restantes foram alinhados contra a versão do genoma humano hg19, com a ferramenta TopHat2 (Kim *et al.*, 2013). A montagem dos transcritos foi realizada através da ferramenta Cufflinks 2.2 (Trapnell *et al.*, 2012) utilizando os parâmetros *default*. Como referência para o alinhamento e para a montagem foi utilizado um arquivo de anotações de transcritos expandido. Esse arquivo GTF continha transcritos de diferentes banco de anotações, principalmente de transcritos anotados como lincRNAs, como detalhado no item 3.4 abaixo. As montagens geradas por cada uma das amostras foram combinadas em um único arquivo utilizando a ferramenta Cuffmerge (Trapnell *et al.*, 2012).

A abundância dos RNAs foi estimada em unidades de FPKM (Fragmentos Por Quilobase de Exons Por Milhão de Fragmentos Mapeados) através da ferramenta Cuffdiff v 2.2 .

A expressão diferencial entre as amostras de células LNCaP tratadas com andrógeno e tratadas com veículo foi realizada como descrito em Love *et al* (2014) (Love *et al.*, 2014). Resumidamente, a quantidade de *reads* de cada locus foi contada com a ferramenta HTSeq (Anders *et al.*, 2015) e a expressão diferencial foi calculada com a ferramenta DESeq2 (Love *et al.*, 2014) considerando um FDR (*False Discovery rate*) < 5 % . A ferramenta DESeq2 utiliza um modelo regressivo que tenta estimar a dispersão da contagem de *reads* de cada gene. Isso é possível através do compartilhamento da informação da dispersão dos genes de diferentes faixas de expressão. Dessa maneira, o valor de dispersão (estimado) e a quantidade de *reads* servem como parâmetros de

entrada para o teste de expressão diferencial. Como os lincRNAs possuem valores de expressão menores que o de genes codificadores de proteínas, a utilização dessa abordagem facilita a identificação dos lincRNAs diferencialmente expressos e diminui a taxa de falsos positivos.

3.4- Expandindo o catalogo de lincRNAs

Para expandir o catálogo padrão de lincRNAs comumente utilizado em estudos da expressão genica em larga escala (GENCODE V19), foi gerado um arquivo de referência contendo todos os genes codificadores de proteína e lincRNAs anotados em diversos outros repositórios públicos.

Os arquivos utilizados para a criação dessa nova referência foram: GENCODE V19 (Harrow *et al.*, 2012) (contém todos os genes codificadores de proteínas anotados e diversos lincRNAs), repositório de lincRNAs identificados no projeto Human Body Map 2 (Cabili *et al.*, 2011), repositório de lincRNAs expressos em diversos tecidos humanos (Hangauer *et al.*, 2013) e os RNAs não codificadores associados ao câncer de próstata (*prostate cancer associated ncRNA transcripts* - PCATs) (Prensner *et al.*, 2011).

Os repositórios de transcritos descritos acima foram processados e *parseados* por meio de scripts escritos na linguagem Python. Esses scripts procuravam por inconsistências nos diferentes arquivos BED/GTF, ordenavam e mesclavam as informações desses transcritos em um único arquivo GTF (*Gene Transfer Format*).

A fusão deu origem a um arquivo GTF que foi utilizado como um referência expandida para a montagem de transcritos e para a expressão diferencial.

3.5- Descoberta de novos lincRNAs expressos em LNCaP

Para encontrar novos lincRNAs específicos da linhagem LNCaP foi utilizado o pipeline mostrado na figura 1A. O pipeline consistiu em alinhar os *reads* no genoma hg19 e montar os transcritos com a ferramenta Cufflinks. Nesse passo, o arquivo GTF expandido foi utilizado como referência para a descoberta e montagem de transcritos.

Todas as saídas contendo transcritos montados pelo programa Cufflinks foram fundidas em um único arquivo através da ferramenta Cuffmerge. O arquivo gerado por essa união contém diversas anotações sobre os transcritos encontrados. Uma anotação importante para este trabalho foi a “Transflag = ‘u’”, que identifica se um transcrito não foi encontrado no GTF referência e se sua localização genômica é intergênica. Todos os genes possuindo essa notação foram selecionados como possíveis novos lincRNAs. Após isso, aqueles candidatos com FPKM < 1 ou menores que 200 nucleotídeos foram removidos. Por último, utilizamos o programa CPC (Kong et al. 2007) (coding potential calculator - versão 0.9-r2) para remover aqueles transcritos com algum potencial de tradução. Os novos transcritos gerados foram integrados ao GTF expandido para a utilização nas análises de RIP-Seq.

3.6- Transcriptoma de tumores de próstata

Dados públicos pareados de amostras tumorais e normais da próstata de 14 homens chineses (Ren *et al.*, 2012) foram obtidos por meio do código de acesso E-

MTAB-567 na plataforma arrayexpress (www.ebi.ac.uk/arrayexpress/). Após isso, esses dados foram analisados utilizando a mesma metodologia adotada para quantificação do transcriptoma da linhagem LNCaP. O arquivo GTF expandido, com adição dos novos lincRNAs encontrados no RNA-Seq de LNCaP foi utilizado para a descoberta dos genes diferencialmente expressos nessas amostras.

3.7- Análise de enriquecimentos de assinaturas genicas com a ferramenta GSEA

A ferramenta de procura por assinaturas genicas GSEA (Subramanian *et al.*, 2005) foi utilizada para buscar módulos enriquecidos entre os genes diferencialmente expressos de células LNCaP e nas amostras de pacientes. O software GSEA em sua versão desktop foi instalado localmente, juntamente com o banco de assinaturas MSigDB em sua versão 5.2 (Subramanian *et al.*, 2005). Durante a análise foram considerados significativos apenas módulos com valor de FDR < 5%.

3.8- Descoberta de lincRNAs associados ao receptor de andrógeno

O controle de qualidade das amostras de RIP-Seq imunoprecipitadas com Anti-AR ou IgG foi realizado adotando os mesmos critérios da filtragem do transcriptoma de LNCaP.

Durante a etapa de alinhamento de *reads* e montagem dos transcritos foram utilizadas a referência hg19 e o arquivo GTF expandido (com adição dos novos lincRNAs encontrados no transcriptoma de LNCaP). Para a identificação dos RNAs que se associaram significativamente ao receptor de andrógeno foi utilizada a ferramenta

DESeq2. Essa estratégia permitiu calcular, para cada transcrito, a razão entre as contagens de reads obtidas na amostra imunoprecipitada com anti-AR e na amostra imunoprecipitada com IgG de soro de animal não-imune.

A lista final de RNAs associados ao receptor de andrógeno continha apenas aqueles lincRNAs cujo Log_2 da razão de *reads* anti-AR/ IgG era positivo e que possuíam valor de FDR menor que 10%.

Nesse trabalho, somente os RNAs da classe dos lincRNAs (RNAs longos não codificadores intergênicos) foram utilizados para as análises de associação RNA-AR que se seguem. Os mRNAs, rRNAs, snoras, ncRNAs antisense e pseudogenes foram descartados, mesmo se apresentassem associação significativa com o AR.

3.9- Determinação de um grupo de lincRNAs não associados ao receptor de andrógeno.

Com o objetivo de estudar de forma rigorosa os lincRNAs associados ao receptor de andrógeno (nomeados de ARA-lincRNAs, ou *Androgen Receptor Associated* lincRNAs) foi criado um grupo controle de lincRNAs não associados ao andrógeno (NonA-lincRNAs, ou *Non-Associated* lincRNAs).

Durante o processo de criação do grupo dos NonA-lincRNAs foi exigido que os lincRNAs candidatos possuísem $\text{FPKM} > 0.1$ em ambas as condições (Anti-AR e IgG) e não fossem significativamente associados ao receptor de andrógeno ou ao IgG (ou seja, lincRNAs cujo log_2 da razão Anti-AR/IgG ≈ 0). Esses lincRNAs foram ordenados de forma crescente usando o valor absoluto do Log_2 da razão da contagem média de reads Anti-AR/IgG. Para manter os dois grupos (ARA-lincRNAs e NonA-lincRNAs) com uma

quantidade equivalente de lincRNAs, o grupo final de NonA-lincRNAs foi composto pelos primeiros 619 lincRNAs da lista de candidatos. Dessa forma, os NonA-lincRNAs escolhidos possuíam uma baixa razão do número de *reads* entre as duas condições de imunoprecipitação.

3.10- Análise dos genes vizinhos dos ARA-lincRNAs

Para investigar a possível associação funcional entre os ARA-lincRNAs e os genes codificadores de proteína vizinhos, procurou-se por genes codificadores de proteínas expressos, que se encontravam nas proximidades dos loci dos ARA-lincRNAs. Para ser considerado vizinho de um ARA-lincRNA, um gene codificador deveria ser o gene com o TSS mais próximo do locus do ARA-lincRNA, além de possuir a expressão maior do que 1 (FPKM >1) nos dados de RNA-Seq de células LNCaP tratadas com andrógeno ou tratadas com o etanol veículo. Caso a segunda restrição não fosse cumprida, o próximo gene codificador de proteína era interrogado até que um vizinho fosse determinado.

A proximidade das coordenadas dos pares lincRNAs-genes codificador foram calculadas por meio da utilização da ferramenta BEDTools (Quinlan e Hall, 2010), usando o comando *bedtools closest*. Apenas a posição inicial do *Transcription Start Site* (TSS) dos lincRNAs e dos genes codificadores de proteínas foi utilizada. A busca foi feita tanto em genes codificadores situados em regiões acima (*upstream*) do TSS dos lincRNAs como em genes situados abaixo (*downstream*) do TSS.

A busca de termos GO (*Gene Ontology*) nos genes codificadores vizinhos de ARA-lincRNAs ou de NonA-lincRNAs foi realizada por meio da ferramenta de anotação

funcional DAVID (versão 6.8) (<https://david.ncifcrf.gov>) (Huang *et al.*, 2007), e como critério estatístico utilizou-se o teste hipergeométrico com o limite de FDR < 5% (Wei Huang *et al.*, 2007).

3.11- Análises de dados públicos de epigenética

Dados de ChIP-Seq para as marcas de histona H3T11P, H3K27me3, H3K4me1, H3K4me2, H3K4me3, H3K36me3, H3K27ac, H4K5ac e H2A.Z e dos fatores de transcrição p300, FoxA1, MED12, NKX3-1, PolIII, AR, FOXF1, RUNX1, EZH2, WDR5, AP4, CTCF e LSD1, juntamente com sequências provenientes de regiões sensíveis ao corte da DNaseI foram obtidos nos repositórios de dados públicos presentes na tabela do anexo 1 (no CD que está em anexo). Todos os ensaios de ChIP-Seq e DNase-Seq são pertencentes à linhagem LNCaP tratada com o hormônio andrógeno.

Todos os arquivos no formato FASTQ passaram pelo mesmo controle de qualidade utilizado para os *reads* do RNA-Seq/RIP-Seq. Os *reads* que passaram no controle de qualidade foram alinhados no genoma hg19 utilizando a ferramenta Bowtie2 com os parâmetros *default*. Para encontrarmos aquelas regiões enriquecidas em cada uma das amostras utilizamos o programa *findpeaks* do pacote HOMER (v.4.8) (Heinz *et al.*, 2010). As amostras foram todas processadas sem a utilização em paralelo de valores de *input* como controle de comparação de enriquecimento. Dessa forma o valor de contagem de *reads* do background foi estimado localmente pelo próprio programa, usando uma janela deslizante em torno de cada pico. O filtro de FDR menor que < 1% foi utilizado para a descoberta de regiões genômicas enriquecidas pelas marcas.

3.12- Extração de características epigenéticas nos TSSs dos lincRNAs associados ao Receptor de Andrógeno

A contagem de *reads* de ChIP-Seq/DNase-seq de cada uma das 22 marcas foi realizada no TSS de cada lincRNA e de seu respectivo gene codificador de proteína vizinho por meio do comando *coverage*, da ferramenta BedTools. Como entrada para o cálculo de cobertura, os arquivos de alinhamento gerados pela ferramenta Bowtie2 foram utilizados.

O TSS foi definido como o intervalo entre os 500 nucleotídeos acima da coordenada de início do gene (*upstream* ao TSS) e os 500 nucleotídeos abaixo (*downstream* ao TSS). Dessa forma, cada gene no arquivo BED possuía um intervalo de 1000 nucleotídeos em torno do TSS. O número total de reads de cada experimento de ChIP-Seq/DNase-Seq foi utilizado para normalizar a contagem final dentro desses TSSs.

A etapa de treinamento do algoritmo random forest (veja detalhamento abaixo) foi realizada utilizando dois grupos. Um grupo composto por ARA-lincRNAs e o genes codificadores de proteína vizinhos e o outro grupo composto por NonA-lincRNAs e os genes codificadores de proteínas assinalados como vizinhos.

Para cada par lincRNA-Gene codificador de proteína do treinamento, um único vetor contendo a quantidade de reads no TSS do lincRNA e no TSS do gene codificador vizinho expresso foi gerado. Como cada uma das 22 marcas foi medida duas vezes, uma no TSS do lincRNA e outra no gene codificador de proteína vizinho, o tamanho total do vetor de um dado par foi igual a 46.

Para realizar a extração das marcas mais importantes nesses TSSs foi utilizada a implementação do algoritmo Random Forest do pacote Scikit-Learn (0.18.1) (Pedregosa

et al., 2012). O modelo foi inicialmente treinado com diferentes números de árvores (100, 500, 1000, 10000) utilizando a opção “*out of bag score*”, que permite estimar individualmente o acerto de cada árvore em uma parcela de dados que não foi utilizada durante a construção da árvore. Em todos os modelos testados, o critério “*gini*” foi utilizado para determinar a separação dos galhos das árvores. Os valor de predição de cada modelo treinado foi calculado usando o método de validação cruzada, conhecido como *five-fold cross validation*, no qual os valores da predição foram calculados com o índice de acurácia (IA),

No qual:

$$IA = (TP + TN) / (TP + TF + FN + FP) \quad (4)$$

TP: Verdadeiros positivos

TN: Verdadeiros negativos

FP: Falsos positivos

FN: Falsos negativos

Os valores de *gini index score* obtidos durante o treinamento foram utilizados como medidas da contribuição de uma dada marca para a classificação final dos pares. Um filtro de significância foi criado com o objetivo de estabelecer um limite para que um dado valor de *gini index score* fosse considerado relevante. Para isso, uma distribuição nula de *gini index score* foi criada através de mil permutações nos grupos de ARA-lincRNAs e NonA-lincRNAs. Dessa forma, a distribuição de *gini index score* obtida nesse

processo foi atribuída aos valores esperados por acaso, uma vez que os dois grupos foram a cada permutação propositalmente misturados.

Utilizando a distribuição nula dos valores de *gini index score* como um guia, foram consideradas significantes apenas aquelas marcas com valores situados em regiões que correspondiam a um intervalo de confiança maior do que 95% ($z\text{-score} \geq 1.5$) em relação a distribuição de valores esperados por pura sorte.

3.13- Análise dos domínios topologicamente associativos

Os arquivos BED já processados que continham as coordenadas dos *topological associating domains* da linhagem celular LNCaP foram obtidos através do banco de dados GEO (id: GSE73782) (Taberlay *et al.*, 2016). Os TADs presentes nesse arquivo foram gerados utilizando uma resolução de 100 kb, ou seja, o genoma inteiro foi sequencialmente segmentado em regiões de 100 kb e a frequência de contatos de um segmento de 100 kb era dada pela soma de todas as interações que possuem âncoras dentro desse limite de coordenadas. Para determinar a frequência de contatos intercromatina na linhagem LNCaP, Taberlay *et al* processaram as amostras de Hi-C com a ferramenta Fit-Hic (Ay *et al.*, 2014). A etapa de anotação dos TADs foi realizada por meio de uma metodologia que utiliza o HMM (Modelo oculto de Markov) proposto por Dixon *et al* (Dixon *et al.*, 2012).

Para atribuir uma TAD a um dado lincRNA foi realizada uma busca de sobreposição de coordenadas utilizando a ferramenta *intersect* do pacote BedTools.

A ferramenta *intersect* também foi utilizada para a contagem de picos de ChIP-Seq/DNAse significativos dentro das coordenadas dos TADs.

Para todos os lincRNAs que eram situados dentro de algum TAD, a distância entre o TSS desse lincRNA e da borda mais próxima foi calculada. Uma vez que um TAD possui duas bordas (5' e 3') foi considerada apenas a distância entre o lincRNA e a borda mais próxima. A distância final foi apresentada em termos de porcentagem e para isso a coordenada do centro do TAD foi utilizada como a distância máxima (100%), enquanto a distância mínima (0%) representava uma sobreposição entre o início do TSS do lincRNA e uma das bordas. As distancias foram medidas pela ferramenta *closest* do pacote BedTools.

Durante o trabalho foi utilizado o *wrapper* em python, pyBedTools, para todas as medidas realizadas com a ferramenta BedTools (Dale *et al.*, 2011).

4- Resultados

4.1- Descoberta de novos lincRNAs no transcriptoma de próstata

Na literatura existe uma grande quantidade de trabalhos que utilizam ferramentas como o microarranjo e o RNA-Seq para a identificação de genes responsivos ao andrógeno (DePrimo *et al.*, 2002; Ngan *et al.*, 2009; Cai *et al.*, 2011), porém, apenas uma pequena quantidade deles foca na expressão dos lincRNAs (Hsieh *et al.*, 2014; Wan *et al.*, 2016). Com o objetivo de investigar a participação dos lincRNAs no programa de regulação mediado pelo receptor de andrógeno (AR) após a exposição ao hormônio, células da linhagem LNCaP foram tratadas por 24 horas com andrógeno sintético (R1881) e foram sequenciadas gerando-se um conjunto de dados de RNA-Seq fita-específico, contendo RNAs selecionados pela cauda poliA+. Foi utilizado como controle para esse experimento células da mesma linhagem tratadas apenas com o veículo (etanol) durante 24 horas. As quatro amostras sequenciadas (2 do tratamento e 2 controles) tiveram uma quantidade média de *reads* de 80-90 milhões cada amostra, que são números próximos do recomendado pelo projeto ENCODE para a investigação de transcriptomas completos (Encode_Consortium, 2011).

Fig1. Pipeline para identificar os lincRNAs no transcriptoma da próstata. Uma ilustração esquemática do procedimento usado para a obtenção do transcriptoma da linhagem celular LNCaP (Tratada e não tratada com andrógeno). Geramos dados de RNA-Seq com *reads* de 75 pb, pareados (paired-end, PE), fita-específicos, a partir de RNA-Poli-A+. Em seguida os dados foram alinhados contra o genoma, e montados usando como referência o catálogo de linRNAs expandido, que contem dados dos 4 bancos indicados. Além da obtenção do perfil transcricional dos lincRNAs e mRNAs já catalogados foi possível a descoberta de novos lincRNAs (*Transcritos Abundantes em Próstata*).

Com o intuito de obter o transcriptoma das células LNCaP tratadas e não tratadas com o hormônio andrógeno, foi utilizado o *pipeline* descrito na Figura 1. Além de analisar a abundância dos transcritos (mRNAs e lincRNAs) já depositados em algum banco de dados, o objetivo dessa abordagem era anotar também lincRNAs ainda não catalogados. Essa tarefa foi conduzida através de uma estratégia que envolvia o alinhamento e montagem dos *reads* obtidos no sequenciamento em larga escala, como descrito abaixo.

Após a montagem do transcriptoma foram encontrados 12060 genes codificadores de proteína com o valor de FPKM médio > 1 na condição tratada ou controle. A obtenção dos lincRNAs expressos em LNCaP envolveu uma etapa adicional que possibilitou a descoberta de novos transcritos. Aqueles transcritos intergênicos que se encontravam fora das referências utilizadas durante a montagem foram tratados como possíveis novos lincRNAs. Para serem posteriormente considerados lincRNAs, esses transcritos intergênicos deveriam ter o valor de FPKM médio > 1 na condição tratada ou controle, deveriam possuir mais de 200 nucleotídeos de comprimento e não deveriam apresentar potencial codificador. Por meio dessa metodologia foi identificado um conjunto confiável de lincRNAs expressos na linhagem LNCaP, que compreende 6996 loci únicos que codificam 7525 isoformas de lincRNAs. Da quantidade total de loci únicos, 3012 (43 %) são lincRNAs já descritos em alguma referência anterior e 3984 (57 %) são prováveis novos lincRNAs (Figura 2A).

Fig 2. Descrição geral dos transcritos expressos na próstata (A) Diagrama circular com o número de transcritos expressos que pertencem a mRNAs, lincRNAs anotados e lincRNAs novos. **(B)** Diagrama circular mostrando a quantidade de lincRNAs expressos anotados em cada uma das referências (GENCODE v19 lincRNAs (Harrow *et al.*, 2012), Cabili *et al.* (Cabili *et al.*, 2011), PCATs e lincRNAs do trabalho de Hangauer *et al.* (Hangauer *et al.*, 2013)).

Dos 3042 loci que pertencem a lincRNAs anotados em referências da literatura, 681 são provenientes das anotações do GENCODE v19 (Harrow *et al.*, 2012), 462 são lincRNAs gerados por Cabili *et al.* (Cabili *et al.*, 2011), 88 são lincRNAs PCATs e 1811 são outros lincRNAs, ambos pertencentes ao catálogo de Hangauer *et al.* (Hangauer *et al.*, 2013) (Figura 2B).

Foi observado que a expressão média dos lincRNAs foi menor do que a dos genes codificadores de proteínas, sendo os valores médios respectivamente 1.2 e 3.2 FPKM (Figura 3A). Os lincRNAs possuem também uma menor quantidade de exons quando comparados com os genes codificadores de proteínas, uma vez que a quantidade média de exons respectivamente foi de 2.3 e 7.9 (Figura 3B).

Fig 3. Expressão e número médio de exons dos lincRNAs e mRNAs (A) Gráfico em violino representado os valores de expressão dos lincRNAs e mRNAs que possuem FPKM >1 após o tratamento da linhagem LNCaP com andrógeno. **(B)** Média da quantidade exons dos genes codificadores e dos lincRNAs.

O nível de expressão, juntamente com a quantidade de exons dos lincRNAs encontrados neste trabalho estão de acordo com dados anteriores presentes na literatura (Cabili *et al.*, 2011; Harrow *et al.*, 2012).

Os dados obtidos através do RNA-seq dessas amostras acrescentam uma grande quantidade de informações relevantes, principalmente a respeito de antigos e novos lincRNAs presentes no transcriptoma da próstata estimulada com o hormônio andrógeno. Além disso, esses dados reforçam a importância da utilização de referências expandidas para o entendimento dos lincRNAs.

4.2- O estímulo com andrógeno promove mudanças na expressão gênica de LNCaP

Utilizando o transcriptoma de próstata obtido na etapa anterior, foi possível identificar os transcritos de genes codificadores de proteínas e lincRNAs que foram diferencialmente expressos após o tratamento com o hormônio. A saída do programa DESeq2 apontou para uma grande variedade de genes afetados pelo tratamento com o andrógeno. A classe dos genes codificadores de proteínas foi a mais afetada pelo tratamento, portanto, foi possível detectar um total de 1000 genes codificadores de proteínas diferencialmente expressos; desses, 606 apresentaram expressão aumentada após o tratamento e 394 tiveram a expressão diminuída (Figura 4A). O número de lincRNAs com expressão diferencial significativa após o tratamento foi de 893, entre os quais 700 lincRNAs apresentaram a expressão aumentada e 193 expressão diminuída (Figura 4B).

Fig 4. Perfil dos genes diferencialmente expressos em LNCaP tratada com andrógeno. Gráfico em vulcão mostrando os **(A)** genes codificadores de proteínas e **(B)** lincRNAs diferencialmente expressos após o tratamento com o hormônio andrógeno. As análises foram realizadas utilizando a ferramenta DESeq e foram considerados significantes aqueles genes com o *p-value* ajustado de < 0.05 . Os genes foram considerados *up-regulated* caso o Log2 do *fold-change* (razão da média de reads do tratado/controle) fosse > 2 (vermelho) e *down-regulated* caso o *fold-change* fosse < -2 (verde).

Interessantemente, dos 894 lincRNAs diferencialmente expressos, um total de 524 foram lincRNAs novos identificados no nosso estudo. Essa grande quantidade de lincRNAs novos diferencialmente expressos reforça a ideia de que a abordagem de sequenciamento de grande profundidade utilizada nesse trabalho está apta a documentar novos lincRNAs presentes na próstata estimulada (Tabela 1).

Tabela 1. Sumário dos lincRNAs mostrando a relação dos lincRNAs expressos (FPKM >1) e significativamente diferencialmente expressos (responsivos ao andrógeno, p-value < 0.05) após o tratamento da linhagem LNCaP com hormônio andrógeno.

	Total expressos	Não significativos	LincRNAs responsivos ao andrógeno		Total
			<i>up-regulated</i>	<i>down-regulated</i>	
LincRNAs anotados	3012	2643	268	101	369
Novos LincRNAs	3984	3460	432	92	524
Total	6996	6103	700	193	893

Vários genes codificadores de proteínas como *FASN*, *NDRG1*, (Crea *et al.*, 2014; Sakurai *et al.*, 2015), *TMPRSS2*, *KLK2*, *KLK3* e *KLK4*, que conhecidamente são estimulados por andrógeno (Kim *et al.*, 2007; Israel *et al.*, 2010; Lawrence *et al.*, 2010; Bueno e Colomer, 2015) tiveram a expressão significativamente aumentada quando as amostras controle e tratadas foram comparadas.

Também foram detectados alguns lincRNAs descritos em trabalhos anteriores como aumentados após o tratamento, tais como o lincRNA PCAT18 e os lincRNAs com expressão diminuída DRAIC e PCAT29 (Crea *et al.* 2014; Sakurai *et al.* 2015).

Para checar se os genes diferencialmente expressos possuíam alguma assinatura transcricional já descrita na literatura, foi conduzida uma análise de enriquecimento de módulos gênicos (*gene set enrichment analysis* - GSEA) utilizando-se todos os genes

codificadores de proteínas expressos como entrada. Ao final dessa análise observou-se que o perfil do transcriptoma de células tratadas com o andrógeno apresentou semelhança ($p < 0.01$) com o perfil de genes responsivos ao andrógeno (*hallmark genes in androgen response*) (Figura 5A) e ao perfil de genes que são comuns encontrados em câncer de próstata (Figura 5B).

Fig 5 - Enriquecimento de perfis nos genes diferencialmente expressos em LNCaP tratada com hormônio. Perfis encontrados pelo GSEA utilizando como entrada todos os genes codificadores de proteínas expressos. Os perfis **(A)** *Hallmark in Androgen Response* e **(B)** *Prostate Cancer* foram similares aos genes diferencialmente expressos após o estímulo com andrógeno ($p\text{-value} < 0.01$). Um intervalo qualquer na linha verde representa o valor atual da somatória do ES (*Enrichment Score*) dos genes utilizados como entrada, dado um perfil existente. O ES aumenta quando os valores de expressão de um gene de entrada (representado pela linha vertical em preto) estão correlacionados com o seu valor no perfil testado. O ES é penalizado quando o gene tem um comportamento inverso ao esperado ou quando um gene do perfil testado não está presente dentre os genes utilizados como entrada (ausência de uma linha preta vertical em uma dada posição). As cores vermelhas e azuis representam, respectivamente, genes com a expressão aumentada ou diminuída. A pontuação final é dada como o mais alto valor obtido no gráfico de ES (linha verde). O $p\text{-value}$ é computado calculando a significância do valor de ES obtido contra uma distribuição de valores de ES gerados por permutações dos genes de entrada.

4.3- LincRNAs em amostras de pacientes com câncer de próstata

Numerosos estudos demonstraram que os lincRNAs contribuem para a progressão e desenvolvimento do câncer (Wapinski e Chang, 2011; Ferreira *et al.*, 2012). Utilizamos os dados de RNA-Seq públicos de câncer de próstata (Ren *et al.*, 2012) para analisarmos nessas amostras a expressão dos lincRNAs encontrados no transcriptoma de LNCaP. Para isso, os dados de RNA-Seq das 28 amostras pareadas (tumoral e não-tumoral) de 14 pacientes foram mapeados no genoma humano (hg19) e a expressão diferencial dos lincRNAs entre os tecidos tumorais e não-tumorais dos 14 pacientes foi calculada.

Fig 6. Perfil dos lincRNAs diferencialmente expressos em amostras de 14 pacientes com câncer de próstata. Gráfico em vulcão mostrando os lincRNAs diferencialmente expressos quando os tecidos não-tumorais de próstata são comparados com tecidos de próstata tumoral. As análises foram realizadas utilizando a ferramenta DESeq e foram considerados significantes aqueles genes com o *p*-valor ajustado de < 0.05 . Os genes foram considerados *up-regulated* caso o Log_2 do *fold-change* (razão da média de reads do tumor/controle) fosse > 2 (vermelho) e *down-regulated* caso o *fold-change* fosse < -2 (verde).

Dos 6996 lincRNAs dados que foram detectados como expressos (FPKM > 1) no sequenciamento da LNCaP, identificamos 1294 lincRNAs que foram expressos em pelo menos 75 % das amostras de câncer dos 14 pacientes. Desses 1294 lincRNAs, 426 foram diferencialmente expressos na comparação entre as amostras câncer de próstata e de tecido não-tumoral (*fold-change* > |2| e com um FDR ≤ 0.05 obtido pela ferramenta DESeq2), sendo que 416 desses genes tiveram a expressão significativamente aumentada, enquanto apenas 10 apresentaram expressão diminuída (Figura 6).

Tabela 2. Sumário dos lincRNAs do transcriptoma de 14 amostras de câncer de próstata de pacientes, contendo a relação de genes expressos (FPKM>1) e diferencialmente expressos entre o tecido de próstata tumoral e não tumoral.

	Total expressos	Não significativos	Transcriptoma dos lincRNAs do câncer de próstata		Total
			up-regulated	down-regulated	
LincRNAs anotados	3012	2729	276	7	283
Novos LincRNAs	3984	3841	140	3	143
Total	6996	6570	416	10	426

Foram encontrados na lista dos genes diferencialmente expressos os lincRNAs *PCA3*, *PVT1* e *GAS5*, que recentemente foram associados ao câncer de próstata em trabalhos isolados (Ferreira *et al.*, 2012; Cui *et al.*, 2016; Song *et al.*, 2016).

No total de 1294 genes dados como expressos tanto nos dados de amostras de pacientes com câncer de próstata como nos dados de LNCaP estimulada com hormônio, 68 foram diferencialmente expressos nas duas situações. Entre os lincRNAs

diferencialmente expressos em ambos os conjuntos de dados, 23 pertencem aos novos lincRNAs que foram encontrados na linhagem LNCaP sequenciada neste trabalho (Tabela 3).

Tabela 3. Sumário dos lincRNAs do transcriptoma de câncer de próstata contendo a relação dos genes diferencialmente expressos nas linhagens LNCaP tratadas com Andrógeno e nos tumores de próstata dos pacientes.

	Intersecção com lincRNAs diferencialmente expressos em LNCaP	Transcriptoma dos lincRNAs do câncer de próstata nos pacientes			Total
		Não significantes	<i>up-regulated</i>	<i>down-regulated</i>	
LincRNAs catalogados	Aumentados em andrógeno	45	29	2	31
LincRNAs catalogados	Diminuídos em andrógeno	30	13	1	14
Novos LincRNAs	Aumentados em andrógeno	13	13	0	13
Novos LincRNAs	Diminuídos em andrógeno	4	10	0	10
Total		92	65	3	68

Os resultados obtidos por meio da integração de dados dos tumores de próstata de pacientes e da linhagem LNCaP reafirmam o potencial dos novos lincRNAs e lincRNAs previamente anotados como importantes candidatos para o estudo tanto da resposta ao estímulo como para o entendimento do câncer de próstata.

4.4- Identificação de lincRNAs associados ao receptor de andrógeno

Para identificar RNAs que potencialmente interagem com o receptor de andrógeno e que poderiam de alguma forma mediar o programa de regulação desse fator de

transcrição, foi realizado um experimento de imunoprecipitação do RNA, seguido de sequenciamento em larga escala (Zhao *et al.*, 2010). O experimento foi conduzido com a utilização de um anticorpo contra o receptor de andrógeno e um anticorpo controle, o IgG de um animal não-imunizado. Utilizamos os RNAs que estavam ligados ao IgG como um controle de ligações inespecíficas, criando assim uma amostra *background* para a comparação com os RNAs que foram imunoprecipitados com o anticorpo contra o AR. As amostras de RNAs imunoprecipitados foram obtidas de células LNCaP tratadas com o andrógeno sintético (R1881) por 6 horas.

O RNA-Seq de duas replicatas biológicas independentes (duas replicatas com anti-AR e duas replicatas com IgG) foram sequenciadas e mapeadas no genoma humano (hg19). Essas bibliotecas geraram por volta de 30 milhões de *reads* e a taxa de alinhamento de cada amostra ficou em torno de 60 - 65%.

Para identificar os lincRNAs associados fisicamente ao receptor de andrógeno a contagem de *reads* das amostras foi realizada utilizando a ferramenta HTSeq e os lincRNAs enriquecidos na amostra ligada ao AR foram encontrados utilizando a ferramenta DESeq2 com filtro de FDR < 10%.

Durante o processo de mapeamento e contagem do experimento de RIP-Seq foi utilizada a referência expandida do transcriptoma. O objetivo era a criação um catálogo mais abrangente dos lincRNAs que poderiam se ligar ao AR. No total, foram encontrados 619 lincRNAs significativamente associados ao receptor de andrógeno (ARA-lincRNAs), dentre esses ARA-lincRNAs, 267 são de novos lincRNAs encontrados no RNA-Seq das células LNCaP estimuladas com o andrógeno.

Os resultados do sequenciamento dos RNA ligados fisicamente ao receptor de andrógeno revelaram, pela primeira vez, como é extenso o repertório de lincRNAs associados ao receptor de andrógeno em células induzidas com o hormônio andrógeno. Esse tipo de descoberta abre o caminho para a investigação de centenas de moléculas não traduzidas, que eram desconhecidas até agora, mas que parecem se ligar ao AR e contribuir para a sinalização celular gerada pela indução hormonal (tabela do anexo 2, ver CD em anexo).

4.5- Os ARA-lincRNAs podem regular os genes codificadores vizinhos

Assumindo que os lincRNAs podem atuar *in cis* regulando os genes vizinhos (Ilott *et al.*, 2014), e para testar esta hipótese no nosso conjunto de dados, foi atribuído um gene codificador de proteína vizinho para cada um dos lincRNAs enriquecidos no experimento de RIP-Seq. Para que um gene codificador de proteína fosse assinalado como um vizinho de um lincRNA, o gene codificador deveria ter o valor médio de FPKM maior que 1 nas amostras de LNCaP tratadas com andrógeno ou no controle com etanol. Adicionalmente, o gene codificador de proteína deveria ser o gene mais próximo do locus do respectivo lincRNAs. Dessa forma, cada ARA-lincRNA formou um par com um gene codificador de proteína de sua vizinhança.

Para as análises subsequentes foi criado um grupo de lincRNAs cujos integrantes não apresentavam associação com o receptor de andrógeno (*non-associated* lincRNAs ou NonA-lincRNAs). Esse grupo foi criado com o objetivo de servir como um grupo controle e foi povoado utilizando os seguintes critérios: (1) O lincRNA candidato deveria ter a média de FPKM maior que 0.1 nas amostras de RNAs capturados no experimento de RIP-Seq com o anticorpo anti-AR ou com o IgG. (2) Os lincRNAs não poderiam ser

contidos dentro do grupo dos lincRNAs significativamente enriquecidos no experimento de RIP-Seq, ou seja, o $\log_2$ da razão anti-AR/IgG deveria estar distribuído em torno de zero, o que corresponde à uma abundância similar do lincRNA tanto no RIP-Seq com o anticorpo anti-AR quanto no RIP-Seq com o IgG. Para balancear o número de NonA-lincRNAs com o dos 619 ARA-lincRNAs, o $\log_2$ da razão do número de *reads* encontrados no anticorpo contra o AR e na amostra com IgG foi calculado e os lincRNAs foram ordenados de forma crescente em função do $\log_2$ da razão. Com isso, os primeiros 619 lincRNAs com os $\log_2$ da razão anti-AR/IgG distribuídos em torno de zero e mais próximos de zero foram utilizados para criar um grupo de tamanho balanceado de NonA-lincRNAs.

A criação do grupo de NonA-lincRNAs permitiu uma série de comparações diretas com o grupo de ARA-lincRNAs (Figura-7).

Figura 7 – Um grande número de lincRNAs são fisicamente associados ao receptor de andrógeno. O sequenciamento de RNAs co-imunoprecipitados com o receptor de andrógeno revelou 619 lincRNAs significativamente enriquecidos (barra verde) nas amostras capturadas com o anticorpo contra AR em relação às amostras com o IgG controle (FDR < 10% DESeq2). A barra em preto na horizontal representa os 619 lincRNAs escolhidos como não significativamente associados ao receptor de andrógeno (NonA-lincRNAs) e serviram como um grupo controle para diversas análises. O eixo-Y representa a densidade de lincRNAs para um determinado valor da razão (AR/IgG).

Após a definição de um grupo controle, foi realizada uma busca por termos de ontologia de genes (*Gene Ontology terms* - GO) que poderiam estar enriquecidos com genes codificadores de proteínas vizinhos de ARA-lincRNAs ou de NonA-lincRNAs. Para isso, a ferramenta *DAVID tool* de busca de termos GO foi utilizada em duas buscas diferentes. Uma das buscas continha os genes codificadores de proteínas vizinhos de ARA-lincRNAs e a segunda com os genes codificadores vizinhos dos NonA-lincRNAs. Curiosamente, apenas os genes codificadores de proteínas vizinhos de ARA-lincRNAs exibiram termos estatisticamente significantes (Teste hipergeométrico com FDR < 5%) (Figura-8).

Figura 8 – Genes codificadores de proteínas na vizinhança dos lincRNAs associados ao receptor de andrógeno apresentam termos GO enriquecidos. O eixo-Y representa os termos GO enriquecidos dos genes codificadores vizinhos de ARA-lincRNAs. Os termos foram ordenados pelo valores de significância ($-\log(p\text{-value})$) obtidos no teste-hipergeométrico. Todos os termos exibidos na figura foram considerados enriquecidos quando utilizado um filtro de $FDR < 5\%$. O tamanho do círculo azul representa a quantidade de genes de entrada que foram enriquecidos em uma dada categoria. Os genes codificadores de proteínas vizinhos de NonA-lincRNAs não apresentaram termos GO significativamente enriquecidos.

Essa análise de categorias GO mostrou que os genes codificadores de proteínas na vizinhança dos ARA-lincRNAs são enriquecidos para termos como o de “organização da cromatina” e “adesão celular”, ambos relevantes para o desenvolvimento e manutenção do câncer de próstata (Mason *et al.*, 2002; Prensner *et al.*, 2013; Gu *et al.*, 2015).

Termos relacionados a “transcrição de DNA” e “processamento de RNA” também foram encontrados, evidenciado que alguns genes codificadores vizinhos de lincRNAs

estão relacionados com processos anabólicos que precedem a divisão celular e se relacionam com o papel do andrógeno na próstata (White, 2005; Lin *et al.*, 2012). Os genes codificadores de proteínas vizinhos de NonA-lincRNAs não apresentaram nenhum termo significativamente enriquecido.

Para investigar se a presença dos ARA-lincRNAs poderia modificar a expressão dos genes codificadores vizinhos, foram mantidos nas análise apenas aqueles ARA-lincRNAs cuja expressão estava aumentada após o tratamento com o hormônio andrógeno (*Fold-change* positivo no experimento de RNA-Seq de LNCaP estimuladas com andrógeno). Usando os 258 ARA-lincRNAs com expressão aumentada, separamos os pares ARA-lincRNAs e vizinhos codificadores em dois grupos, um contendo codificadores cuja expressão foi aumentada com o tratamento com andrógeno (*fold-change* positivo) (n= 177) e outro em que a expressão foi diminuída após o tratamento com o andrógeno (*fold-change* negativo) (n= 81). Para definir esses grupos foram utilizados os valores de *fold-change* obtidos no experimento de RNA-Seq de LNCaP estimulada com hormônio (Figura 9A).

A)

LincRNAs expressed in same direction
of neighbor protein coding genes

LincRNAs expressed in opposite direction
of neighbor protein coding genes

B)

Figura 9 - Genes codificadores vizinhos de ARA-lincRNAs possuem um maior grau de modificação na expressão após o tratamento com hormônio andrógeno (A) Esquema utilizado para separar ARA-lincRNAs ativados por andrógeno (setas azuis) que possuíam genes codificadores vizinhos com a expressão aumentada após o tratamento com andrógeno (seta vermelha esquerda), daqueles com expressão diminuída após o tratamento com andrógeno (seta vermelha direita). **(B)** Distribuição cumulativa do *fold-change* dos genes vizinhos com a expressão aumentada (linha preta esquerda) e diminuída (linha preta direita) após o tratamento com andrógeno. Nessa análise consideramos apenas aqueles genes com valores de *fold-change* dentro do percentil 33 ao 66, como feito por González et al (González *et al.*, 2015). Note que no gráfico da esquerda os valores da razão $\log_2$ (androgeno/control) são positivos, indicando ativação da transcrição do codificador de proteína na presença do hormônio, enquanto na direita os valores da razão são negativos, indicando inibição da transcrição do codificador de proteína na presença do hormônio. A diferença estatística dos dois grupos (ARA-lincRNAs e Nona-lincRNAs) foi computada utilizando um teste-t.

Esses resultados mostraram que o aumento no nível de expressão dos genes codificadores vizinhos de ARA-lincRNAs ativados por andrógeno foi significativamente maior que o dos genes codificadores vizinhos de NonA-lincRNAs ($p < 0.05$ em um teste-t) (Figura 9B, esquerda). Os valores de expressão dos genes codificadores vizinhos de ARA-lincRNAs cujos *fold-changes* diminuíram após o tratamento com andrógeno foram significativamente menores que os valores dos genes codificadores vizinhos de NonA-lincRNAs ($p < 0.05$ em um teste-t) (Figura 9B, direita).

Esses dados demonstram que uma parte dos genes codificadores de proteínas possui modificação em sua expressão na presença de um vizinho ARA-lincRNA. Esse tipo de fenômeno pode indicar que as células LNCaP estimuladas com andrógeno podem explorar a ligação dos ARA-lincRNAs ao receptor de andrógeno para modular *in cis* a expressão dos genes codificadores de proteínas vizinhos.

4.6- O perfil epigenético dos ARA-lincRNAs e dos genes codificadores vizinhos

Com a lista dos lincRNAs fisicamente associados ao receptor de andrógeno disponível, procurou-se a existência de assinaturas epigenéticas que poderiam diferenciar os ARA-lincRNAs e seus respectivos genes codificadores vizinhos daqueles lincRNAs que não interagem fisicamente com o receptor de andrógeno (NonA-lincRNAs).

Existe uma grande quantidade de trabalhos na literatura que mostram como é importante a contribuição dos diversos fatores epigenéticos para a regulação de eventos de expressão gênica (Barski *et al.*, 2007; Capurso *et al.*, 2012; Conley e King Jordan,

2012; Calo e Wysocka, 2013). Pensando nisso, utilizamos dados públicos de ChiP-Seq de várias marcas de histonas, fatores de transcrição, Polimerase II e sítios de hipersensibilidade à DNaseI obtidos com a linhagem de LNCaP após o tratamento com andrógeno. Foi utilizado esse enorme repertório de dados de marcas epigenéticas na tentativa de identificar se no sítio de transcrição dos ARA-lincRNAs haveria uma composição característica da cromatina e se os genes codificadores vizinhos dos ARA-lincRNAs também possuíam uma assinatura da cromatina distinta.

Os genes codificadores de proteínas vizinhos dos ARA-lincRNAs foram novamente divididos em dois grupos. O primeiro grupo foi formado com os genes codificadores que possuíam *fold-change* de expressão positivo na presença de andrógeno e um outro grupo com os vizinhos codificadores que apresentavam o *fold-change* negativo nos experimento de RNA-seq na linhagem LNCaP induzida por andrógeno.

Para capturar o enriquecimento das marcas e fatores de transcrição em cada gene, foi definida uma janela de 500 nt +/- que era centralizada no TSS dos lincRNAs e do respectivo gene codificador vizinho. Utilizando as 22 marcas epigenéticas, computou-se a quantidade normalizada de *reads* dentro de cada uma das regiões de TSS dos ARA-lincRNAs, NonA-lincRNAs e dos codificadores vizinhos de cada um desses grupos.

A frequência de *reads* para cada marca no TSS dos lincRNAs e dos seus genes codificadores vizinhos foram utilizadas como variáveis de entrada para um modelo de aprendizado que utiliza o algoritmo Random-Forest. O objetivo desejado ao utilizar esse algoritmo nessa etapa foi de capturar quais marcas possuem maior importância para a separação dos lincRNAs associados fisicamente ao receptor de andrógenos dos

lincRNAs que não são associados. A Figura 10 mostra resumidamente como se deu o processo da extração de variáveis epigenéticas mais relevantes presentes nos ARA-lincRNAs.

Figura 10 – Pipeline utilizado para a extração das variáveis epigenéticas com maior relevância para a separação de ARA-lincRNAs e genes codificadores vizinhos. (A)

Sequenciamento e mapeamento dos lincRNAs co-imunoprecipitados com o receptor de andrógeno no experimento de RIP-Seq. Os lincRNAs significativamente ligados ao receptor de andrógeno foram chamados de ARA-lincRNAs e um grupo controle foi estabelecido e chamado de NonA-lincRNAs. **(B)** Atribuição de um gene codificador de proteína vizinho para cada lincRNA e a contagem de *reads* para cada experimento de ChIP-Seq/DNase-Seq dentro do TSS (+/- 500 pares de base) dos pares lincRNA-Gene codificador vizinho **(C)** Utilizando o algoritmo Random Forest foi realizada uma etapa de aprendizado e extração das variáveis epigenéticas de maior relevância para separação dos ARA-lincRNAs e NonA-lincRNAs.

Para o grupo de ARA-lincRNAs cujos genes codificadores vizinhos tinham *fold-change* positivo após o tratamento da linhagem LNCaP com andrógeno, obtivemos os valores de acurácia médio do modelo Random Forest igual a 0.629 (Tabela 4). No grupo de ARA-lincRNAs cujos vizinhos tinham o *fold-change* negativo, o valor de acurácia médio foi de 0.60 (Tabela 5). Esses valores foram obtidos por um modelo de validação cruzada conhecido como K-fold (Fushiki, 2011). Para esse trabalho o valor de K foi igual a 5, sendo o valor final da predição dado pela média do índice de acurácia obtidos em cada particionamento dos dados. O modelos utilizando 1000 árvores foram aqueles que obtiveram a maior taxa de acurácia.

Tabela 4. Valores de scores obtidos para diferentes números de árvores durante o treinamento do modelo *Random Forest* para os ARA-lincRNAs com vizinhos com o *fold-change* positivo.

Tipo do Score*	Número de árvores	Valor médio do score
OOB score	100	0.610169
IA	100	0.598413
OOB score	500	0.598870
IA	500	0.615476
OOB score	1000	0.635593
IA	1000	0.629444
OOB score	10000	0.618644
IA	10000	0.615397

* Tipos de score computados: OOB, *out of bag score*; IA, índice de acurácia (ver Métodos)

Tabela 5. Valores de scores obtidos para diferentes números de árvores durante o treinamento do modelo *Random Forest* para os ARA-lincRNAs com vizinhos com o *fold-change* negativo.

Tipo do Score*	Número de árvores	Valor médio do score
OOB score	100	0.598765
IA	100	0.590809
OOB score	500	0.641975
IA	500	0.597059
OOB score	1000	0.623457
IA	1000	0.597426
OOB score	10000	0.629630
IA	10000	0.578676

* Tipos de score computados: OOB, *out of bag score*; IA, índice de acurácia (ver Métodos)

Para acessar a relevância de cada variável epigenética para a classificação de ARA-lincRNAs e NonA-lincRNAs, foram utilizados os valores de Gini-index médios obtidos após o treinamento do modelo Random Forest com 1000 árvores.

Como os valores de acurácias foram mais baixos do que o esperado, foi implementada uma maneira de filtrar a tabela de Gini-index deixando apenas os valores que estavam acima de um dado limiar considerado seguro. O limiar de significância foi estabelecido através da criação de um valor nulo de Gini-index, gerado pelo embaralhamento de lincRNAs e NonA-lincRNAs, seguido da obtenção dos valores de Gini-index (Descrito com mais detalhes nos materiais e métodos) (Capurso et al. 2012).

Com o resultado dessa análise foi possível identificar que existe um enriquecimento de marcas de acessibilidade da DNase e de Pol II tanto nos ARA-lincRNAs que são vizinhos de genes codificadores de proteínas com expressão aumentada como nos ARA-lincRNAs que diminuem a expressão após o tratamento com o hormônio andrógeno (Figura 11).

A)

LincRNAs expressed in same direction
of neighbor protein coding genes

LincRNAs expressed in opposite direction
of neighbor protein coding genes

B)

Figura 11 – Promotores de ARA-lincRNAs apresentam uma assinatura epigenética específica. **(A)** Esquema utilizado para separar ARA-lincRNAs ativados por andrógeno (setas azuis) que possuíam genes codificadores vizinhos com a expressão aumentada após o tratamento com andrógeno (seta vermelha, esquerda), daqueles com expressão diminuída após o tratamento com andrógeno (seta vermelha, direita). **(B)** Ranking das 10 características epigenéticas mais relevantes para a classificação dos ARA-lincRNAs. As variáveis foram organizadas pelo potencial discriminatório dado pelo Gini index obtido pelo algoritmo Random Forest. Uma seta em azul indica que a variável epigenética teve sua significância acima do limiar do modelo nulo. Uma seta apontando para cima indica que a média da quantidade de *reads* daquela marca foi maior nos ARA-lincRNAs do que em NonA-lincRNAs. Uma seta apontando para baixo significa que a média da quantidade de *reads* daquela marca foi menor nos ARA-lincRNAs do que em NonA-lincRNAs. A linha pontilhada vertical indica o limite de significância obtido pelo modelo nulo de Gini-index.

Analisando separadamente os dois grupos, foi possível observar que a marca de histona H3K27ac, que é uma importante marca presente em enhancers ativos (Creyghton *et al.*, 2010), foi enriquecida no promotor de ARA-lincRNAs vizinhos de genes codificadores de proteína com o *fold-change* positivo no experimento de RNA-seq com a linhagem LNCaP induzida. Curiosamente, analisando o enriquecimento de marcas no TSS dos genes codificadores com o *fold-change* positivo que são vizinhos de ARA-lincRNAs, observamos que eles exibem enriquecimento para a variante de histona H2A.Z (Figura 12, esquerda).

A variante de histona H2A.Z já foi descrita como um importante fator para o estabelecimento e manutenção da interação entre enhancers e promotores. Além disso, participa no recrutamento da Polimerase II para regiões contendo eRNAs em processo de transcrição (Brunelle *et al.*, 2015).

Figura 12 – Promotores de genes codificadores de proteínas vizinhos de ARA-lincRNAs apresentam o enriquecimento de algumas marcas epigenéticas. (A) Esquema utilizado para separar ARA-lincRNAs ativados por andrógeno (setas azuis) que possuíam genes codificadores vizinhos com a expressão aumentada após o tratamento com andrógeno (seta vermelha, esquerda), daqueles com expressão diminuída após o tratamento com andrógeno (seta vermelha, direita). **(B)** Ranking das 10 características epigenéticas nos promotores de genes codificadores de proteínas vizinhos de ARA-lincRNAs, que possuem maior relevância para a classificação dos ARA-lincRNAs. As variáveis foram organizadas pelo potencial discriminatório dado pelo Gini-index obtido pelo algoritmo Random Forest. Uma seta em vermelho indica que a variável epigenética teve sua significância acima do limiar do modelo nulo. Uma seta apontando para cima indica que a média da quantidade de *reads* de uma dada marca foi maior nos promotores de genes codificadores vizinhos de ARA-lincRNAs. Uma seta apontando para baixo significa que a média da quantidade de *reads* de uma dada marca foi menor nos genes codificadores vizinhos de ARA-lincRNAs. A linha pontilhada vertical indica o limite de significância obtido pelo modelo nulo de Gini-index.

Os genes codificadores de proteínas vizinhos de ARA-lincRNAs, cujos *fold-changes* eram negativos no experimento de RNA-Seq com a linhagem LNCaP estimulada com hormônio, tiveram diminuição da ocupação dos co-fatores de transcrição WDR5 e NKX3-1, ambos importantes para o direcionamento da proliferação e sobrevivência celular (Figura 12, direita). Após o estímulo hormonal, esses co-fatores tendem a ocupar os mesmos sítios genômicos ocupados pelo receptor de andrógeno (Tan *et al.*, 2012; Kim *et al.*, 2014).

As diferenças epigenéticas observadas no TSS dos ARA-lincRNAs e dos genes codificadores de proteínas vizinhos sugerem que alguns dos elementos associados à cromatina podem contribuir para a regulação de genes codificadores de proteínas na presença de lincRNAs associados fisicamente ao receptor de andrógeno.

4.7- A expressão gênica e o perfil epigenético dentro dos domínios topologicamente associados são modificados na presença de ARA-lincRNAs

Para investigar o envolvimento da arquitetura genômica nas possíveis modificações transcricionais e epigenéticas causadas pela presença de lincRNAs associados ao receptor de andrógeno após o estímulo hormonal, foram utilizados dados que demarcavam os domínios topologicamente associativos (Topologically Associating Domains ou TAD) da linhagem LNCaP.

As coordenadas de todos os ARA-lincRNAs com *fold-change* positivo foram comparadas com as coordenadas dos TADs. Dessa forma foi possível definir a qual TAD

cada um ARA-lincRNAs pertencia. Usando essa informação, foi analisada a composição das marcas epigenéticas nos TADs contendo ARA-lincRNAs e NonA-lincRNAs. Após estimar o número de picos significativos dentro dessas TADs, algumas marcas tiveram uma maior quantidade de picos representada dentro das TADs que possuíam ARA-lincRNAs. Picos das marcas de histona H3K27ac, H4K5ac, H3K4me1, H3K36me3 e do fator de transcrição LSD1 foram significativamente maiores em TADs povoadas por lincRNAs que se associam ao receptor de andrógeno (p -valor < 0.05 teste-t) (Figura 13).

Figura 13 – TADs contendo ARA-lincRNAs são enriquecidas com um conjunto específico de marcas epigenéticas. (A) Diagrama esquemático representando TADs com ARA-lincRNAs e NonA-lincRNAs. **(B)** *Boxplot* da quantidade (eixo-Y) de picos significativos, de cada uma das marcas indicadas no eixo Y, dentro de TADs contendo ARA-lincRNAs (caixas verdes) e de TADs contendo NonA-lincRNAs (caixas cinzas). No gráfico estão representados os dados de todas as marcas significativamente diferentes entre os TADs dos dois grupos (p-valor < 0.05, teste-t).

Uma metodologia semelhante foi utilizada para investigar se a expressão dos genes codificadores de proteínas dentro das TADs contendo ARA-lincRNAs foi alterada após o tratamento com andrógeno. O resultado dessa análise mostrou que os genes codificadores dentro de TADs contendo ARA-lincRNAs mostram um aumento de expressão médio maior do que a os genes codificadores cujos lincRNAs não se associam ao receptor de andrógeno (NonA-lincRNAs) (P-valor < 0.05, teste-t) (Figura 14).

Figura 14- Genes codificadores de proteínas dentro de TADs contendo ARA-lincRNAs tem maior aumento de expressão após o tratamento com andrógeno. *Boxplot* mostrando o Log_2 (média do FPKM andrógeno/controlado) da expressão dos genes codificadores de proteínas após o tratamento com andrógeno. As caixas verdes representam os valores dos genes codificadores que estão dentro de TADs que possuem ARA-lincRNAs. As caixas cinzas representam genes codificadores que estão dentro de TADs contendo NonA-lincRNAs. A diferença dos dois grupos foi dada como significativa após a aplicação de um teste-t (p-valor < 0.05).

As bordas das TADs são importantes regiões que flanqueiam os TADs. Essas regiões possuem uma pequena quantidade de contatos com a cromatina fora de seu domínio e são conhecidamente ligadas por vários grupos de proteínas que possuem

função de modelar a arquitetura da cromatina (Van Bortle *et al.*, 2014). Para examinar se existia uma tendência dos ARA-lincRNAs de se situarem mais próximos das bordas dos TADs, foi calculada a distância relativa dos promotores dos ARA-lincRNAs e NonA-lincRNAs da borda mais próxima. Esse resultado mostrou que os ARA-lincRNAs estão mais próximos das bordas das TADs do que os NonA-lincRNAs utilizados como controle (p-valor < 0.05 em um teste-t) (Figura 15).

Figura 15 – Promotores de ARA-lincRNAs são mais próximos das bordas dos TADs do que promotores de NonA-lincRNAs. O eixo-X representa a posição relativa do TSS dos lincRNAs em relação a borda mais próxima do TAD. O centro do TAD foi utilizado como a maior distância possível entre o promotor do lincRNA e as bordas. Os ARA-lincRNAs exibiram uma maior proximidade das bordas do que os NonA-lincRNAs (p-valor < 0.05 em um teste-t).

Após utilizar os TADs como um ponto de referência para entender as modificações locais causadas pela existência de um ARA-lincRNA, foi possível mostrar que, de fato, existem modificações específicas que parecem ser mais frequentes dentro de domínios da cromatina enriquecidos com ARA-lincRNAs. Dessa forma, o perfil epigenético enriquecido com marcas de ativação dentro dos TADs contendo ARA-lincRNAs de

alguma forma poderia justificar o aumento médio da expressão de genes codificadores dentro desses TADs.

A proximidade dos ARA-lincRNAs das bordas das TADs sugere que a organização desses elementos não parece ser aleatória. Existe a possibilidade de que os ARA-lincRNAs utilizem alguma particularidade dessas regiões de borda para desempenharem seu papel regulatório.

4.8- Promotores de ARA-lincRNAs mostram semelhanças com regiões contendo *enhancers* ativos

Para acrescentar evidências sobre a capacidade de regulação *in cis* dos ARA-lincRNAs, foi calculada a similaridade das marcas presentes nos promotores dessas moléculas com as marcas dos *enhancers* ativos. Para isso, a razão entre a contagem de reads de H3K27ac e de H3K27me3 foi mensurada. Essa medida poderia apontar se os ARA-lincRNAs se assemelhavam mais com *enhancers* ativos ou com *poised enhancers* (*enhancers* em potencial, porém desativados) (Heinz *et al.*, 2015). O resultado dessa análise indicou que os promotores de ARA-lincRNAs exibem uma razão de H3K27ac/H3K27me3 maior que a dos NonA-lincRNAs (p-valor < 0.05 em um teste de Kolmogorov-smirnov) e portanto possuem uma característica mais similar ao de *enhancers* ativos (Figura 16).

Figura 16 – ARA-lincRNAs são mais similares a *enhancers* ativos do que os NonA-lincRNAs. Distribuição cumulativa da razão de *reads* para as marcas H3K27ac/H3K27me3 nos promotores de ARA-lincRNAs (verde) e NonA-lincRNAs (preto). Os ARA-lincRNAs mostraram uma distribuição da razão de H3K27ac/H3K27me3 significativamente maior do que a dos NonA-lincRNAs (teste de Kolmogorov-smirnov com um p-valor < 0.05)

5- Discussão

Apesar do estudo e caracterização dos lincRNAs estar em franca expansão, ainda não é possível estimar a magnitude da importância de cada um dos RNAs não codificadores para os diversos mecanismos celulares (Engreitz *et al.*, 2016; Khurana *et al.*, 2016).

Particularidades dos lincRNAs como (1) o baixo nível de conservação de sequências, quando comparados aos genes codificadores de proteínas, (2) a fraca modularidade em suas estruturas secundárias e (3) a quantidade pequena de lincRNAs validados experimentalmente tornam ainda mais complicada a caracterização funcional dessas moléculas (Guttman *et al.*, 2010; Ching *et al.*, 2015). Ainda que essas características dificultem o processo de entendimento dos lincRNAs, a grande tecido-especificidade apresentada por eles pode conduzir o olhar dos pesquisadores para processos específicos que acontecem em determinados tecidos, sob certas condições. Neste trabalho foi realizada uma procura por novos lincRNAs e lincRNAs já anotados que foram expressos na linhagem LNCaP de câncer de próstata tratada com hormônio andrógeno por 24 horas. Racionalmente, esse tipo de prospecção parece ser vantajosa no estudo dos lincRNAs, uma vez que o contexto pode auxiliar na delimitação do papel destas moléculas.

Para que fosse possível descrever lincRNAs com participação exclusiva no processo de resposta ao hormônio andrógeno, o catalogo de lincRNAs foi expandido com a integração de diferentes bancos de dados de lincRNAs. Apesar dessa iniciativa ter aumentado o catalogo inicial de lincRNAs, é bem possível que vários bancos de dados contendo essas moléculas não foram integrados neste trabalho e poderiam diminuir o

número de lincRNAs chamados aqui de “novos lincRNAs”. Desta forma, é reforçada aqui a necessidade e importância da criação e manutenção de repositórios de lincRNAs, que ofereçam anotações atualizadas e expandidas de repertórios de trabalhos anteriores.

Tanto para os lincRNAs previamente anotados como também para aqueles considerados novos, foram encontradas algumas características aparentemente comuns nesse tipo de molécula. Assim, a quantidade menor de exons e o baixo nível de expressão quando comparado aos genes codificadores mostram que os lincRNAs de células LNCaP estimuladas não apresentam particularidades nessas características.

A resposta ao hormônio andrógeno modifica a expressão de uma enorme quantidade de genes codificadores de proteínas, RNAs não codificadores intrônicos e até de eRNAs (Louro *et al.*, 2007; Hsieh *et al.*, 2014), porém, nunca foi feita uma procura tão minuciosa como a desse trabalho por lincRNAs que possuem a expressão modificada nesta condição.

Tanto os genes codificadores de proteínas quanto lincRNAs exibiram uma tendência geral de aumento da expressão após o estímulo. Isso sugere que a resposta ao andrógeno aumenta a complexidade do transcriptoma que precede o estímulo. É também possível pensar que o aumento da complexidade do transcriptoma possa ser atribuído a transcrição de genes que pertencem ou auxiliam a extensa maquinaria necessária para a progressão do ciclo celular que é recrutada após o estímulo hormonal.

Interessantemente, a maior parte dos lincRNAs diferencialmente expressos após o tratamento são lincRNAs que ainda não estavam anotados nos repositórios utilizados neste trabalho. Isso sugere que a resposta ao andrógeno pela célula LNCaP conta com

a expressão de uma enorme quantidade de lincRNAs, que são altamente especializados na regulação da resposta a esse hormônio. A ideia de que os lincRNAs adicionam camadas *customizadas* de regulação nas respostas celulares dos eucariotos (Ulitsky e Bartel, 2013) pode ser reforçada pela observação de outros fenômenos semelhantes ao estudado aqui. Ensaio em que diferentes tipos celulares são estimulados com moléculas indutoras (Ex: partículas de antígenos ou indução hormonal) mostram que o estímulo indutor realmente pode suscitar a expressão de um transcriptoma de lincRNAs altamente orientado e específico para o dado evento (Ilott *et al.*, 2014; Miano *et al.*, 2015; Sun *et al.*, 2015).

Dentre os genes com expressão diferencial obtidos neste trabalho, foi observada a expressão diferencial de alguns genes codificadores e lincRNAs conhecidamente regulados durante a resposta ao andrógeno (Kim *et al.*, 2007; Israel *et al.*, 2010; Lawrence *et al.*, 2010; Crea *et al.*, 2014; Sakurai *et al.*, 2015). Isso sugere que os dados de transcriptoma gerados neste trabalho foram coerentes com os resultados da literatura e são válidos para o estudo da regulação durante a indução dessas células. Outro resultado que suporta essa ideia é a similaridade dos genes diferencialmente expressos neste trabalho com algumas assinaturas presentes em bancos de dados que possuem condições experimentais semelhantes as utilizadas para a obtenção do presente transcriptoma. Um desses exemplos é a assinatura *hallmark genes in androgen response* (Liberzon *et al.*, 2015).

O catálogo expandido, que continha os lincRNAs das diferentes bibliotecas públicas e também os novos lincRNAs descobertos pelo sequenciamento da LNCaP induzida, foi utilizado para uma procura de lincRNAs diferencialmente expressos em

dados de pacientes com tumor de próstata. O resultado dessa busca revelou a expressão diferencial de 23 lincRNAs nas amostras de tumores de pacientes, que ainda não estavam anotados na literatura e que aparentemente podem ter importância para o entendimento do tumor de próstata. Além dos transcritos que não estavam anotados, lincRNAs dos diferentes repositórios foram encontrados diferencialmente expressos nesses tumores. Esses resultados são animadores, uma vez que a descoberta de moléculas obliquamente reguladas em tumores de diferentes pacientes é de extrema importância. De maneira geral, essa compilação poderá abrir o caminho para a investigação do papel individual desses lincRNAs durante o desenvolvimento e manutenção dos tumores de próstata. Por outro lado, a extensão do catálogo de lincRNAs presentes nesses tumores poderá impulsionar a criação de painéis gênicos mais robustos. Painéis como esse possibilitariam tanto a distinção do grau da patologia de um dado paciente como poderiam prever a possibilidade de uma recidiva após a prostatectomia (Mortensen *et al.*, 2015; Albitar *et al.*, 2016).

Apesar da sobreposição entre os genes diferencialmente expressos não ser tão grande entre a linhagem celular e os tumores de pacientes, ainda sim um total de 68 lincRNAs foram encontrados diferencialmente expressos tanto na linhagem tratada como nos tumores de pacientes. Dentre os fatores que poderiam explicar a baixa correspondência dos genes diferencialmente expressos entre tumores de pacientes e a linhagem estimulada com andrógeno encontra-se a heterogeneidade transcricional existente entre os tumores de diferentes pessoas (Marusyk e Polyak, 2011), a diferença da concentração de hormônio andrógeno nos pacientes e na linhagem LNCaP

(Schulman *et al.*, 2010) e até mesmo a baixa cobertura dos dados públicos de sequenciamento dos tumores utilizados.

Vários trabalhos tem voltado a atenção para compreensão dos mecanismos que permitem a interação entre os RNAs não codificadores e a maquinaria regulatória que é composta por diversos complexos proteicos (Khalil e Rinn, 2011). Muitos desses estudos mostraram que os lincRNAs podem se ligar fisicamente a determinadas proteínas, principalmente aquelas que pertencem à maquinaria de edição, leitura e remodelamento da cromatina (Beckedorff *et al.*, 2013; Simon *et al.*, 2013; Yang *et al.*, 2015). Apesar do volume de evidências experimentais que suportam a importância da associação entre lincRNAs e proteínas ter aumentado, ainda não estão disponíveis na literatura trabalhos que mostrem a importância da associação entre lincRNAs e fatores de transcrição, principalmente durante fenômenos regulatórios complexos, como é o caso da indução da linhagem LNCaP com o hormônio andrógeno.

O trabalho de Hsieh *et al* (2014) apresentou um isolado exemplo de um RNA de *enhancer* que pode se associar fisicamente ao receptor de andrógeno. Um dos resultados interessantes mostrados por esse grupo é que o silenciamento desse eRNA afeta a expressão do gene codificador vizinho KLK3 (Hsieh *et al.*, 2014). Esses achados abrem espaço para a investigação funcional não só dos eRNAs, mas também de lincRNAs que poderiam participar da modulação da expressão gênica e alterar a composição da cromatina após a indução feita com hormônio andrógeno.

Nosso trabalho possibilitou a descoberta de 619 lincRNAs significativamente associados ao receptor de andrógeno (ARA-lincRNAs). Desses, 267 foram representados por novos lincRNAs não previamente anotados na literatura.

A migração do receptor de andrógeno para o núcleo é um fenômeno bem caracterizado, que ocorre após o tratamento com o hormônio andrógeno (Brinkmann *et al.*, 1999; Mahajan *et al.*, 2007). Por isso, espera-se que os ARA-lincRNAs encontrados em nosso trabalho representem, em quase sua totalidade, RNAs enriquecidos no núcleo e que poderiam estar envolvidos em processos de regulação da expressão genica e modificação da cromatina.

Um sequenciamento dos RNAs ligados ao receptor de andrógeno sem o estímulo por andrógeno foi realizado em paralelo ao experimento de RIP-Seq de nosso trabalho. Esses dados não foram apresentados na presente tese, porém indicaram que menos de 10 lincRNAs estão significativamente ligados ao receptor de andrógeno em células não tratadas. Curiosamente, alguns desses lincRNAs já são descritos na literatura como sendo citoplasmáticos.

A detecção das várias particularidade da cromatina e dos promotores de genes vizinhos dos ARA-lincRNAs só foi possível após a criação de um grupo controle contendo os NonA-lincRNAs. O processo de criação desse grupo tentou manter o número de elementos similar ao dos ARA-lincRNAs, ao mesmo tempo que capturava lincRNAs que estavam presentes no experimento de RIP-Seq mas que não possuíam enriquecimento significativo nas amostra de IgG ou AR. Essa abordagem tentou evitar que as diferenças entre os grupo não fossem artificialmente causadas. Esse tipo de erro poderia acontecer caso os NonA-lincRNAs fossem escolhidos aleatoriamente, sem levar em consideração o aparecimento desses lincRNAs nas amostras de RIP-Seq e RNA-Seq.

Inicialmente, o grupo contendo NonA-lincRNAs foi criado utilizando os lincRNAs associados significativamente ao IgG. Essa escolha aparentemente óbvia foi descartada

devido a possibilidade desses lincRNAs interagirem inespecificamente com o IgG por algum mecanismo ou característica física desconhecidos. Esse tipo de problema foge do escopo deste trabalho e provavelmente modificaria as interpretações dos resultados aqui obtidos.

Frequentemente a função dos lincRNAs é associada com a regulação *in cis* de genes codificadores de proteínas vizinhos (Hsieh *et al.*, 2014; Ilott *et al.*, 2014; Quinn e Chang, 2016). Utilizando essa premissa, foram definidos possíveis genes codificadores para cada um dos ARA-lincRNAs encontrado no experimento de RIP-Seq. Curiosamente, foi observado que os genes codificadores de proteínas vizinhos de ARA-lincRNAs apresentam termos que poderiam sugerir a participação sinérgica dos lincRNAs em determinadas categorias de eventos celulares (Ex: organização da cromatina, adesão celular, transcrição de DNA e processamento de RNA), que seriam necessários para que a células respondam efetivamente após estimuladas com o hormônio andrógeno.

Além do receptor de andrógeno promover o aumento da transcrição dos genes alvos, a interação do receptor com alguns genes pode gerar a repressão transcricional dos mesmos (Prescott *et al.*, 2007; Grosse *et al.*, 2012). Em geral, o fenômeno de repressão por meio de fatores de transcrição são menos referenciados que aqueles que geram a indução da expressão. Porém, o balanço dessas duas faces da regulação feita por fatores de transcrição possui similar importância para entendimento da resposta ao andrógeno (Zhao *et al.* 2012).

Sabendo que alguns genes codificadores podem aumentar ou diminuir a expressão na presença do hormônio andrógeno, os pares ARA-lincRNAs e codificadores

vizinho foram separados seguindo essa premissa. Os resultados indicaram que a magnitude do efeito regulatório após a indução com andrógeno é maior nos genes codificadores vizinhos de ARA-lincRNAs. Esse efeito sugere que a ligação dos lincRNAs vizinhos ao receptor de andrógeno possa de alguma forma intensificar o efeito regulatório nesses loci.

Os valores de predição obtidos pelo algoritmo *Random Forest* na busca por diferenças nos promotores dos pares ARA-lincRNAs e vizinhos codificadores apresentaram uma taxa de acerto de 0,60 – 0,62, acima da taxa de eventos aleatórios (taxa de 0,5 ou menos). Além disso, acrescentamos à análise um limiar de significância, o que parece aumentar a confiabilidade dessa metodologia para a identificação de marcas epigenéticas envolvidas com o programa regulatório dos ARA-lincRNAs.

A busca por diferenças nos promotores dos pares ARA-lincRNAs e vizinhos codificadores mostrou que os promotores de ARA-lincRNAs além de apresentarem maior enriquecimento de marcas para Polimerase II, se destacam também por possuírem maior depleção de histonas (maior abundância de marcas de acessibilidade à DNase). A abundância de marcas de DNase é predominante em regiões da cromatina que desempenham papel regulatório, principalmente por servirem como uma região que aglomera vários fatores que auxiliam na regulação de genes *in cis* (Song e Crawford, 2010; Rajagopal *et al.*, 2013; González *et al.*, 2015).

Os ARA-lincRNAs que possuíam vizinhos com a expressão aumentada após a indução com andrógeno apresentaram enriquecimento para a marca de histona H3K27ac, que normalmente é presente em elementos *enhancers*. A semelhança dos promotores de alguns lincRNAs com as regiões de *enhancers* já está descrita na

literatura (Ilott *et al.*, 2014; Brazão *et al.*, 2016), sendo que a expressão dos lincRNAs com esse perfil foi relacionada com a regulação *in cis* dos genes vizinhos. É possível que a deposição das marcas H3K27ac seja acelerada na presença da interação receptor de andrógeno com os ARA-lincRNAs. Em contrapartida, a proteína histona acetiltransferase p300 que é comumente encontrada nos *enhancers* (Calo e Wysocka, 2013) não foi diferencialmente enriquecida nos loci dos ARA-lincRNAs. Esse fato pode sugerir que a deposição dessa marca nos promotores dos ARA-lincRNAs aconteça através de outra histona acetiltransferase.

A marca H4K5ac, que é importante para a reativação pós-mitótica da transcrição de genes (Zhao *et al.*, 2011), foi vista enriquecida nos promotores dos ARA-lincRNAs que ativam os genes vizinhos. Isso poderia mostrar que ativação de genes vizinhos guiada por ARA-lincRNAs é importante durante a mitose das células estimuladas pelo andrógeno e também para que as células filhas modulem esses ARA-lincRNAs imediatamente após a mitose.

Os promotores de genes codificadores vizinhos de genes ARA-lincRNAs apresentaram enriquecimento da variante de histona H2A.Z. O papel dessa variante de histona é pouco conhecido tanto quando presente nos promotores como no corpo dos genes (Valdés-Mora *et al.*, 2012). Contudo, alguns trabalhos mostram uma anti-correlação da variante H2A.Z com a metilação do DNA, o que seria um efeito coerente com o de genes transcritionalmente ativos após o tratamento com o hormônio. Existem trabalhos que mostram que a incorporação desse tipo de histona nos promotores pode regular a acessibilidade dessas regiões aos fatores de transcrição, o que poderia criar

uma situação favorável para o acontecimento da expressão gênica (Marques *et al.*, 2010).

Os promotores de genes codificadores vizinhos de ARA-lincRNAs, cuja expressão foi diminuída após o tratamento com o andrógeno, apresentaram duas modificações importantes. Nesses promotores, os fatores de transcrição WDR5 e NKX3.1, que trabalham sinergicamente com o receptor de andrógeno (Gurel *et al.*, 2010; Kim *et al.*, 2014), foram encontrados em nossos dados com menor frequência. Uma possível especulação é que a ligação do receptor de andrógeno a esses ARA-lincRNAs poderia evitar o recrutamento ou bloquear a interação dessas proteínas com esses promotores. Essa obstrução da maquinaria de expressão genica poderia explicar a diminuição da transcrição desses genes codificadores vizinhos de ARA-lincRNAs após a ocupação por andrógeno.

Curiosamente, a vizinhança dos ARA-lincRNAs não foi diferencialmente enriquecida com marcas do receptor de andrógeno. Isso indica que a frequência da ligação do receptor de andrógeno não é modificada na presença dos ARA-lincRNAs. Dessa forma, é possível que, apesar da interação do receptor com os sítios regulados não sofrer alteração, a associação adicional do receptor com um ARA-lincRNA criaria um *scaffold* que recrutaria outros elementos que se ligam a cromatina nessas regiões.

Na literatura existe um exemplo similar ao da hipótese levantada acima. Nele, os autores descrevem o papel da interação entre o receptor de andrógeno e os lincRNAs PCGM1 e PRNCR1 na linhagem tumoral de próstata C2R22RV1. Esses lincRNAs se ligam fisicamente ao receptor de andrógeno e contribuem com a regulação *in trans* dos genes alvos do receptor. Ensaios que promoviam o *knockdown* desses lincRNAs

mostraram que a expressão dos genes regulados pelo receptor de andrógeno foi perturbada, porém, o receptor de andrógeno continuou ocupando normalmente os sítios de regulação dos seus alvos (Yang *et al.*, 2013).

A arquitetura da cromatina nos organismos eucariotos tem papel fundamental para a expressão gênica. Os contatos que acontecem entre diferentes regiões da cromatina não são aleatórios, muito pelo contrário, a cromatina nos eucariotos apresenta um elevado nível de setorização e hierarquização (Rao *et al.*, 2014; Nichols e Corces, 2015; Boyan e Giacomo, 2016).

Em nosso trabalho investigamos a existência de modificações epigenéticas que iriam além daquelas encontradas nos promotores dos ARA-lincRNAs e de seus genes codificadores vizinhos; para isso usamos a estrutura da cromatina obtida por meio de dados de Hi-C. Notamos que genes codificadores co-localizados no mesmo TAD que ARA-lincRNAs possuem em geral maior aumento de expressão após o tratamento com o andrógeno. Esses resultados propõem que a forte rede de contatos existente na cromatina compartimentalizada nos TADs poderia expandir localmente os efeitos gerados pela interação dos ARA-lincRNAs com o receptor de andrógeno.

Dessa forma, as consequências da interação entre ARA-lincRNAs não se limitariam apenas à modificação da expressão gênica dos genes codificadores de proteínas. No interior dos TADs contendo ARA-lincRNAs vimos uma maior concentração das marcas de histona H3K27ac, H4K5ac, H3K4me1, H3K36me3. Todas essas marcas se relacionam com um perfil de aumento da expressão gênica dentro do TAD. Surpreendentemente, a ocupação da proteína LSD1 foi aumentada no interior dos TADs contendo ARA-lincRNAs. Durante a resposta ao hormônio andrógeno, LSD1 se

colocaliza com o receptor de andrógeno e tem o papel de desmetilar a histona H3K9 (não presente entre os dados de ChIP-Seq) levando a de-repressão dos genes alvo do receptor (Metzger *et al.*, 2005). A proteína LSD1, como descrito na literatura, tem o potencial de se ligar a lincRNAs, como é o caso da sua interação com o lincRNA HOTAIR (Tsai *et al.*, 2010). Essas evidências poderiam sugerir que a interação física entre ARA-lincRNAs e receptor de andrógeno promoveria a ligação de LSD1 aos loci ocupados pelo receptor. Uma vez ligado ao receptor de andrógeno, o ARA-lincRNA poderia se ligar à LSD1, de forma similar ao que acontece com lincRNA HOTAIR (Figura 17) (Tsai *et al.*, 2010).

Outra descoberta interessante é a de que os ARA-lincRNAs apresentam uma maior tendência a se situarem próximos das bordas dos TADs. Esse resultado é intrigante, uma vez que as regiões das bordas dos TADs apresentam uma grande quantidade de complexos proteicos que definem e mantem a estrutura da TAD organizada. Diversos trabalho tem mostrado que o rompimento das bordas de um TAD pode causar um espalhamento de marcas repressoras dentro dos TADs (Barkess e West, 2012; Van Bortle *et al.*, 2014; Ali *et al.*, 2016).

Apesar dos picos de CTCF não terem sido mais abundantes nos TADs contendo ARA-lincRNAs é possível que outras proteínas organizadoras das estruturas dos TADs sejam abundantes na presença de ARA-lincRNAs.

Uma outra característica interessante encontrada nos promotores dos ARA-lincRNAs que foram descritos nesse trabalho foi a maior similaridade de seus promotores com promotores de *enhancers* ativos. A existência dessa similaridade aponta que, no

geral, ARA-lincRNAs estão envolvidos com papel regulatório *in cis* mais frequentemente do que lincRNAs que não se ligam ao receptor de andrógeno.

Figura 17- Esquema da possível ação dos ARA-lincRNAs dentro dos TADs contendo ARA-lincRNAs. O esquema propõe que um ARA-lincRNA depois de transcrito (linhas azuis) se ligue ao receptor de andrógeno (verde) e sirva como um ancoradouro para outras enzimas remodeladoras da cromatina como por exemplo a LSD1 (roxo). Esses complexos regulariam a transcrição de um gene codificador de proteína (vermelho) por meio da ativação de promotores e de enhancers distais, que estão em contato com o locus do ARA-lincRNA em questão, dentro de uma unidade TAD.

6- Conclusões

Resumidamente, o estudo desta tese permite concluir que:

O estudo minucioso do transcriptoma da linhagem de câncer de próstata LNCaP tratada com hormônio andrógeno permitiu a descoberta de lincRNAs desconhecidos e de lincRNAs que ainda não estavam relacionados com a resposta dessa linhagem ao hormônio andrógeno.

A integração entre dados de transcriptoma da linhagem celular LNCaP e dados de tumores de pacientes com câncer de próstata permitiu a descoberta de lincRNAs que podem ter papel semelhante no tumores e nas células estimuladas com andrógeno.

Centenas de LincRNAs são associados fisicamente ao receptor de andrógeno (ARA-lincRNAs) após o tratamento com o hormônio andrógeno.

ARA-lincRNAs são vizinhos de genes codificadores de proteína que desempenham processos importantes para que a célula LNCaP responda ao estímulo hormonal.

Genes codificadores de proteínas vizinhos de ARA-lincRNAs possuem a expressão gênica diferenciada após o estímulo com o hormônio andrógeno.

Os promotores dos ARA-lincRNAs e dos seus genes codificadores vizinhos apresentam particularidades quando comparado com genes codificadores que não possuem ARA-lincRNAs em sua vizinhança.

ARA-lincRNAs apresentam algumas características conhecidas de reguladores *in cis*, sendo essas características mais acentuadas nos ARA-lincRNAs do que nos lincRNAs que não interagem com o receptor de andrógeno.

Os ARA-lincRNAs se localizam mais próximos das bordas dos TADs do que os lincRNAs que não se associam com o receptor de andrógeno.

ARA-lincRNAs parecem modificar a composição da cromatina e a expressão gênica dentro dos TADs.

7- Referências bibliográficas

- Albitar, M., W. Ma, L. Lund, F. Albitar, K. Diep, H. A. Fritsche and N. Shore (2016). "Predicting Prostate Biopsy Results Using a Panel of Plasma and Urine Biomarkers Combined in a Scoring System." Journal of Cancer **7**: 297-303.
- Ali, T., R. Renkawitz and M. Bartkuhn (2016). "Insulators and domains of gene expression." Current Opinion in Genetics & Development **37**: 17-26.
- Anders, S., P. T. Pyl and W. Huber (2015). "HTSeq-A Python framework to work with high-throughput sequencing data." Bioinformatics **31**: 166-169.
- Ay, F., T. L. Bailey and W. S. Noble (2014). "Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts." Genome research **24**: 999-1011.
- Balk, S. P. (2003). "Biology of Prostate-Specific Antigen." Journal of Clinical Oncology **21**: 383-391.
- Barkess, G. and A. G. West (2012). "Chromatin insulator elements: establishing barriers to set heterochromatin boundaries." Epigenomics **4**: 67-80.
- Barski, A., S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev and K. Zhao (2007). "High-Resolution Profiling of Histone Methylations in the Human Genome." Cell **129**: 823-837.
- Beckedorff, F. C., A. C. Ayupe, R. Crocci-Souza, M. S. Amaral, H. I. Nakaya, D. T. Soltys, C. F. M. Menck, E. M. Reis and S. Verjovski-Almeida (2013). "The Intronic Long Noncoding RNA ANRASSF1 Recruits PRC2 to the RASSF1A Promoter, Reducing the Expression of RASSF1A and Increasing Cell Proliferation." PLoS Genetics **9**: e1003705.
- Belton, J.-M., R. P. McCord, J. H. Gibcus, N. Naumova and Y. Zhan (2012). "Hi-C: A comprehensive technique to capture the conformation of genomes." Methods **58**: 268-276.
- Beltran, H., D. S. Rickman, K. Park, S. S. Chae, A. Sboner, T. Y. MacDonald, Y. Wang, K. L. Sheikh, S. Terry, S. T. Tagawa,

- R. Dhir, J. B. Nelson, A. de la Taille, Y. Allory, M. B. Gerstein, S. Perner, K. J. Pienta, A. M. Chinnaiyan, Y. Wang, C. C. Collins, M. E. Gleave, F. Demichelis, D. M. Nanus and M. A. Rubin (2011). "Molecular characterization of neuroendocrine prostate cancer and identification of new drug targets." Cancer discovery **1**: 487-495.
- Bevan, C. L., S. Hoare, F. Claessens, D. M. Heery and M. G. Parker (1999). "The AF1 and AF2 domains of the androgen receptor interact with distinct regions of SRC1." Molecular and cellular biology **19**: 8383-8392.
- Bhatia-Gaur, R., A. A. Donjacour, P. J. Sciavolino, M. Kim, N. Desai, P. Young, C. R. Norton, T. Gridley, R. D. Cardiff, G. R. Cunha, C. Abate-Shen and M. M. Shen (1999). "Roles for Nkx3.1 in prostate development and cancer." Genes & development **13**: 966-977.
- Bolger, A. M., M. Lohse and B. Usadel (2014). "Trimmomatic: A flexible trimmer for Illumina sequence data." Bioinformatics **30**: 2114-2120.
- Bonasio, R., E. Lecona, V. Narendra, P. Voigt, F. Parisi, Y. Kluger and D. Reinberg (2014). "Interactions with RNA direct the Polycomb group protein SCML2 to chromatin where it represses target genes." eLife **3**: e02637.
- Boyan, B. and C. Giacomo (2016). "Organization and function of the 3D genome." Nature Review Genetics **17**: 661 - 678.
- Boyle, A. P., S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey and G. E. Crawford (2008). "High-Resolution Mapping and Characterization of Open Chromatin across the Genome." Cell **132**: 311-322.
- Brazão, T. F., J. S. Johnson, J. Müller, A. Heger, C. P. Ponting and V. L. J. Tybulewicz (2016). "Long noncoding RNAs in B-cell development and activation." Blood **128**: e10-e19.
- Breiman, L. (2001). "Random forests." Machine Learning **45**: 5-32.
- Brinkmann, A. O., L. J. Blok, P. E. de Ruiter, P. Doesburg, K. Steketee, C. A. Berrevoets and J. Trapman (1999).

- "Mechanisms of androgen receptor activation and function." The Journal of steroid biochemistry and molecular biology **69**: 307-313.
- Brunelle, M., A. Nordell Markovits, S. Rodrigue, M. Lupien, P. É. Jacques and N. Gévry (2015). "The histone variant H2A.Z is an important regulator of enhancer activity." Nucleic Acids Research **43**: 9742-9756.
- Bueno, M. J. and R. Colomer (2015). A Fatty Acid Synthase Inhibitor Shows New Anticancer Mechanisms. EBioMedicine. **2**: 778-779.
- Butler, J. E. F. and J. T. Kadonaga (2002). "The RNA polymerase II core promoter: a key component in the regulation of gene expression." Genes & development **16**: 2583-2592.
- Cabili, M., C. Trapnell, L. Goff, M. Koziol, B. Tazon-Vega, A. Regev and J. L. Rinn (2011). "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses." Genes and Development **25**: 1915-1927.
- Cai, C., H. H. He, S. Chen, I. Coleman, H. Wang, Z. Fang, S. Chen, P. S. Nelson, X. S. Liu, M. Brown and S. P. Balk (2011). "Androgen receptor gene expression in prostate cancer is directly suppressed by the androgen receptor through recruitment of lysine-specific demethylase 1." Cancer cell **20**: 457-471.
- Callewaert, L., N. Van Tilborgh and F. Claessens (2006). "Interplay between Two Hormone-Independent Activation Domains in the Androgen Receptor." Cancer Research **66**.
- Calo, E. and J. Wysocka (2013). "Modification of enhancer chromatin: what, how, and why?" Molecular cell **49**: 825-837.
- Cantin, L., F. Faucher, J.-F. Couture, K. P. de Jésus-Tran, P. Legrand, L. C. Ciobanu, Y. Fréchette, R. Labrecque, S. M. Singh, F. Labrie and R. Breton (2007). "Structural characterization of the human androgen receptor ligand-binding domain complexed with EM5744, a rationally designed steroidal ligand bearing a bulky chain directed

- toward helix 12." The Journal of biological chemistry **282**: 30910-30919.
- Capurso, D., H. Xiong and M. R. Segal (2012). "A histone arginine methylation localizes to nucleosomes in satellite II and III DNA sequences in the human genome." BMC genomics **13**: 630.
- Carson, C. and R. Rittmaster (2003). "The role of dihydrotestosterone in benign prostatic hyperplasia." Urology **61**: 2-7.
- Casci, T. (2006). "DNase I assay gets a face lift." Nat Rev Genet **7**: 590-590.
- Cassiday, L. A. (2002). "Having it both ways: transcription factors that bind DNA and RNA." Nucleic Acids Research **30**: 4118-4126.
- Catalona, W. J., J. P. Richie, F. R. Ahmann, M. A. Hudson, P. T. Scardino, R. C. Flanigan, J. B. deKernion, T. L. Ratliff, L. R. Kavoussi and B. L. Dalkin (1994). "Comparison of digital rectal examination and serum prostate specific antigen in the early detection of prostate cancer: results of a multicenter clinical trial of 6,630 men." The Journal of urology **151**: 1283-1290.
- Ching, T., J. Masaki, J. Weirather and L. X. Garmire (2015). "Non-coding yet non-trivial: a review on the computational genomics of lincRNAs." BioData Min **8**: 44.
- Chu, C., Z. Fang, X. Hua, Y. Yang, E. Chen, A. W. Cowley, M. Liang, P. Liu and Y. Lu (2015). "deGPS is a powerful tool for detecting differential expression in RNA-sequencing studies." BMC Genomics **16**: 455.
- Chu, C., K. Qu, F. L. Zhong, S. E. Artandi and H. Y. Chang (2011). "Genomic Maps of Long Noncoding RNA Occupancy Reveal Principles of RNA-Chromatin Interactions." Molecular Cell **44**: 667-678.
- Cirillo, L. A., F. R. Lin, I. Cuesta, D. Friedman, M. Jarnik and K. S. Zaret (2002). "Opening of compacted chromatin by early

- developmental transcription factors HNF3 (FoxA) and GATA-4." Mol Cell **9**(2): 279-289.
- Claessens, F., P. Alen, A. Devos, B. Peeters, G. Verhoeven and W. Rombauts (1996). "The Androgen-specific Probasin Response Element 2 Interacts Differentially with Androgen and Glucocorticoid Receptors." Journal of Biological Chemistry **271**: 19013-19016.
- Conley, A. B. and I. King Jordan (2012). "Epigenetic regulation of human cis-natural antisense transcripts." Nucleic Acids Research **40**: 1438-1445.
- Crea, F., A. Watahiki, L. Quagliata, H. Xue, L. Pikor, A. Parolia, Y. Wang, D. Lin, W. L. Lam, W. L. Farrar, T. Isogai, R. Morant, S. Castori-Eppenberger, K. N. Chi, Y. Wang and C. D. Helgason (2014). "Identification of a long non-coding RNA as a novel biomarker and potential therapeutic target for metastatic prostate cancer." Oncotarget **5**: 764-774.
- Creyghton, M. P., A. W. Cheng, G. G. Welstead, T. Kooistra, B. W. Carey, E. J. Steine, J. Hanna, M. A. Lodato, G. M. Frampton, P. A. Sharp, L. A. Boyer, R. A. Young and R. Jaenisch (2010). "Histone H3K27ac separates active from poised enhancers and predicts developmental state." Proceedings of the National Academy of Sciences of the United States of America **107**: 21931-21936.
- Cui, M., L. You, X. Ren, W. Zhao, Q. Liao and Y. Zhao (2016). "Long non-coding RNA PVT1 and cancer." Biochemical and Biophysical Research Communications **471**: 10-14.
- Dale, R. K., B. S. Pedersen and A. R. Quinlan (2011). "Pybedtools: A flexible Python library for manipulating genomic datasets and annotations." Bioinformatics **27**: 3423-3424.
- DePrimo, S. E., M. Diehn, J. B. Nelson, R. E. Reiter, J. Matese, M. Fero, R. Tibshirani, P. O. Brown and J. D. Brooks (2002). "Transcriptional programs activated by exposure of human prostate cancer cells to androgen." Genome biology **3**: RESEARCH0032.

- Di Stefano, M., J. Paulsen, T. G. Lien, E. Hovig and C. Micheletti (2016). "Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization." Scientific reports **6**: 35985.
- Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. MCS '00: Proceedings of the First International Workshop on Multiple Classifier Systems: 1-15.
- Dinger, M. E., P. P. Amaral, T. R. Mercer and J. S. Mattick (2009). "Pervasive transcription of the eukaryotic genome: Functional indices and conceptual implications." Briefings in Functional Genomics and Proteomics **8**: 407-423.
- Dixon, J. R., I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget, A. Y. Lee, Z. Ye, A. Kim, N. Rajagopal, W. Xie, Y. Diao, J. Liang, H. Zhao, V. V. Lobanenko, J. R. Ecker, J. A. Thomson and B. Ren (2015). "Chromatin architecture reorganization during stem cell differentiation." Nature **518**: 331-336.
- Dixon, J. R., S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu and B. Ren (2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions." Nature **485**: 376-380.
- Djebali, S., C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, N. S. Bar, P. Batut, K. Bell, I. Bell, S. Chakraborty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, E. Falconnet, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, O. J. Luo, E. Park, K. Persaud, J. B. Preall, P. Ribeca, B. Risk, D. Robyr, M. Sammeth, L. Schaffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters,

- H. Wang, J. Wrobel, Y. Yu, X. Ruan, Y. Hayashizaki, J. Harrow, M. Gerstein, T. Hubbard, A. Reymond, S. E. Antonarakis, G. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo and T. R. Gingeras (2012). "Landscape of transcription in human cells." Nature **489**(7414): 101-108.
- Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras (2013). "STAR: Ultrafast universal RNA-seq aligner." Bioinformatics **29**: 15-21.
- Doesburg, P., C. W. Kuil, C. A. Berrevoets, K. Steketee, P. W. Faber, E. Mulder, A. O. Brinkmann and J. Trapman (1997). "Functional in vivo interaction between the amino-terminal, transactivation domain and the ligand binding domain of the androgen receptor." Biochemistry **36**: 1052-1064.
- Dowen, J. M., Z. P. Fan, D. Hnisz, G. Ren, B. J. Abraham, L. N. Zhang, A. S. Weintraub, J. Schuijers, T. I. Lee, K. Zhao and R. A. Young (2014). "Control of cell identity genes occurs in insulated neighborhoods in mammalian chromosomes." Cell **159**: 374-387.
- ENCODE, C. (2007). "Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project." Nature **447**: 799-816.
- Encode_Consortium (2011). "Standards, Guidelines and Best Practices for RNA-Seq V.1.0." https://genome.ucsc.edu/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf.
- Engreitz, J. M., J. E. Haines, E. M. Perez, G. Munson, J. Chen, M. Kane, P. E. McDonel, M. Guttman and E. S. Lander (2016). "Local regulation of gene expression by lncRNA promoters, transcription and splicing." Nature **539**(7629): 452-455.
- Ernst, J., P. Kheradpour, T. S. Mikkelsen, N. Shores, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis and B. E. Bernstein (2011).

- "Mapping and analysis of chromatin state dynamics in nine human cell types." Nature **473**.
- Esteller, M. (2011). "Non-coding RNAs in human disease." Nature reviews. Genetics **12**: 861-874.
- Falkenberg, K. J. and R. W. Johnstone (2014). "Histone deacetylases and their inhibitors in cancer, neurological diseases and immune disorders." Nature Reviews Drug Discovery **13**: 673-691.
- Farnham, P. J. (2009). "Insights from genomic profiling of transcription factors." Nature Reviews Genetics **10**: 605-616.
- Ferreira, L. B., A. Palumbo, K. D. de Mello, C. Sternberg, M. S. Caetano, F. L. de Oliveira, A. F. Neves, L. E. Nasciutti, L. R. Goulart and E. R. P. Gimba (2012). "PCA3 noncoding RNA is involved in the control of prostate-cancer cell survival and modulates androgen receptor signaling." BMC cancer **12**: 507.
- Furey, T. S. (2012). "ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions." Nat Rev Genet **13**(12): 840-852.
- Fushiki, T. (2011). "Estimation of prediction error by using K-fold cross-validation." Statistics and Computing **21**(2): 137-146.
- Gévry, N., H. M. Chan, L. Laflamme, D. M. Livingston and L. Gaudreau (2007). "p21 transcription is regulated by differential localization of histone H2A.Z." Genes & development **21**: 1869-1881.
- Gévry, N., S. Hardy, P.-E. Jacques, L. Laflamme, A. Sotelis, F. Robert and L. Gaudreau (2009). "Histone H2A.Z is essential for estrogen receptor signaling." Genes & development **23**: 1522-1533.
- Global Burden of Disease Cancer, C., C. Fitzmaurice, C. Allen, R. M. Barber, L. Barregard, Z. A. Bhutta, H. Brenner, D. J. Dicker, O. Chimed-Orchir, R. Dandona, L. Dandona, T. Fleming, M. H. Forouzanfar, J. Hancock, R. J. Hay, R. Hunter-Merrill, C. Huynh, H. D. Hosgood, C. O. Johnson, J. B. Jonas, J. Khubchandani, G. A. Kumar, M. Kutz, Q. Lan,

H. J. Larson, X. Liang, S. S. Lim, A. D. Lopez, M. F. MacIntyre, L. Marczak, N. Marquez, A. H. Mokdad, C. Pinho, F. Pourmalek, J. A. Salomon, J. R. Sanabria, L. Sandar, B. Sartorius, S. M. Schwartz, K. A. Shackelford, K. Shibuya, J. Stanaway, C. Steiner, J. Sun, K. Takahashi, S. E. Vollset, T. Vos, J. A. Wagner, H. Wang, R. Westerman, H. Zeeb, L. Zoeckler, F. Abd-Allah, M. B. Ahmed, S. Alabed, N. K. Alam, S. F. Aldhahri, G. Alem, M. A. Alemayohu, R. Ali, R. Al-Raddadi, A. Amare, Y. Amoako, A. Artaman, H. Asayesh, N. Atnafu, A. Awasthi, H. B. Saleem, A. Barac, N. Bedi, I. Bensenor, A. Berhane, E. Bernabe, B. Betsu, A. Binagwaho, D. Boneya, I. Campos-Nonato, C. Castaneda-Orjuela, F. Catala-Lopez, P. Chiang, C. Chibueze, A. Chittheer, J. Y. Choi, B. Cowie, S. Damtew, J. das Neves, S. Dey, S. Dharmaratne, P. Dhillon, E. Ding, T. Driscoll, D. Ekwueme, A. Y. Endries, M. Farvid, F. Farzadfar, J. Fernandes, F. Fischer, G. H. TT, A. Gebru, S. Gopalani, A. Hailu, M. Horino, N. Horita, A. Hussein, I. Huybrechts, M. Inoue, F. Islami, M. Jakovljevic, S. James, M. Javanbakht, S. H. Jee, A. Kasaeian, M. S. Kedir, Y. S. Khader, Y. H. Khang, D. Kim, J. Leigh, S. Linn, R. Lunevicius, H. M. El Razek, R. Malekzadeh, D. C. Malta, W. Marcenes, D. Markos, Y. A. Melaku, K. G. Meles, W. Mendoza, D. T. Mengiste, T. J. Meretoja, T. R. Miller, K. A. Mohammad, A. Mohammadi, S. Mohammed, M. Moradi-Lakeh, G. Nagel, D. Nand, Q. Le Nguyen, S. Nolte, F. A. Ogbo, K. E. Oladimeji, E. Oren, M. Pa, E. K. Park, D. M. Pereira, D. Plass, M. Qorbani, A. Radfar, A. Rafay, M. Rahman, S. M. Rana, K. Soreide, M. Satpathy, M. Sawhney, S. G. Sepanlou, M. A. Shaikh, J. She, I. Shiue, H. R. Shore, M. G. Shrimme, S. So, S. Soneji, V. Stathopoulou, K. Stroumpoulis, M. B. Sufiyan, B. L. Sykes, R. Tabares-Seisdedos, F. Tadese, B. A. Tedla, G. A. Tessema, J. S. Thakur, B. X. Tran, K. N. Ukwaja, B. S. Uzochukwu, V. V. Vlassov, E. Weiderpass, M. Wubshet Terefe, H. G. Yebyo, H. H. Yimam, N. Yonemoto, M. Z.

- Younis, C. Yu, Z. Zaidi, M. E. Zaki, Z. M. Zenebe, C. J. Murray and M. Naghavi (2016). "Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study." JAMA Oncol **388**: 1459-1544.
- González, A. J., M. Setty and C. S. Leslie (2015). "Early enhancer establishment and regulatory locus complexity shape transcriptional programs in hematopoietic differentiation." Nature genetics **47**: 1249-1259.
- Goodwin, S., J. D. McPherson and W. R. McCombie (2016). "Coming of age: ten years of next-generation sequencing technologies." Nat Rev Genet **17**: 333-351.
- Grob, S., M. Schmid and U. Grossniklaus (2014). "Hi-C Analysis in Arabidopsis Identifies the KNOT, a Structure with Similarities to the flamenco Locus of Drosophila." Molecular Cell **55**: 678-693.
- Gröschel, S., M. A. Sanders, R. Hoogenboezem, E. De Wit, B. A. M. Bouwman, C. Erpelinck, V. H. J. Van Der Velden, M. Havermans, R. Avellino, K. Van Lom, E. J. Rombouts, M. Van Duin, K. Döhner, H. B. Beverloo, J. E. Bradner, H. Döhner, B. Löwenberg, P. J. M. Valk, E. M. J. Bindels, W. De Laat and R. Delwel (2014). "A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in Leukemia." Cell **157**: 369-381.
- Grosse, A., S. Bartsch and A. Baniahmad (2012). "Androgen receptor-mediated gene repression." Molecular and Cellular Endocrinology **352**: 46-56.
- Gu, L., S. C. Frommel, C. C. Oakes, R. Simon, K. Grupp, C. Y. Gerig, D. Bar, M. D. Robinson, C. Baer, M. Weiss, Z. Gu, M. Schapira, R. Kuner, H. Sultmann, M. Provenzano, I. P. o. E. O. P. Cancer, M.-L. Yaspo, B. Brors, J. Korbel, T. Schlomm, G. Sauter, R. Eils, C. Plass and R. Santoro (2015). "BAZ2A (TIP5) is involved in epigenetic alterations in prostate cancer

- and its overexpression predicts disease recurrence." Nat Genet **47**: 22-30.
- Gurel, B., T. Z. Ali, E. A. Montgomery, S. Begum, J. Hicks, M. Goggins, C. G. Eberhart, D. P. Clark, C. J. Bieberich, J. I. Epstein and A. M. De Marzo (2010). "NKX3.1 as a marker of prostatic origin in metastatic tumors." The American journal of surgical pathology **34**: 1097-1105.
- Guttman, M., I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn and E. S. Lander (2009). "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals." Nature **458**: 223-227.
- Guttman, M., M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander and A. Regev (2010). "Ab initio reconstruction of transcriptomes of pluripotent and lineage committed cells reveals gene structures of thousands of lincRNAs." Nature biotechnology **28**: 503-510.
- Hangauer, M. J., I. W. Vaughn and M. T. McManus (2013). "Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs." PLoS Genet **9**(6): e1003569.
- Harmston, N. and B. Lenhard (2013). "Chromatin and epigenetic features of long-range gene regulation." Nucleic acids research **41**: 7185-7199.
- Harrow, J., A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. Van Baren, M. Brent, D. Haussler, M. Kellis, A.

- Valencia, A. Reymond, M. Gerstein, R. Guigó and T. J. Hubbard (2012). "GENCODE: The reference human genome annotation for the ENCODE project." Genome Research **22**: 1760-1774.
- Heinlein, C. A. and C. Chang (2001). "Role of Chaperones in Nuclear Translocation." Endocrine **14**: 143-150.
- Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh and C. K. Glass (2010). "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities." Molecular cell **38**: 576-589.
- Heinz, S., C. E. Romanoski, C. Benner and C. K. Glass (2015). "The selection and function of cell type-specific enhancers." Nature reviews. Molecular cell biology **16**: 144-154.
- Hérault, Y., J. Beckers, T. Kondo, N. Fraudeau and D. Duboule (1998). "Genetic analysis of a Hoxd-12 regulatory element reveals global versus local modes of controls in the HoxD complex." Development (Cambridge, England) **125**: 1669-1677.
- Heyn, H., S. Moran, I. Hernando-Herraez, S. Sayols, A. Gomez, J. Sandoval, D. Monk, K. Hata, T. Marques-Bonet, L. Wang and M. Esteller (2013). "DNA methylation contributes to natural human variation." Genome Research **23**: 1363-1372.
- Holoch, D. and D. Moazed (2015). "RNA-mediated epigenetic regulation of gene expression." Nature reviews. Genetics **16**: 71-84.
- Horoszewicz, J. S., S. S. Leong, E. Kawinski, J. P. Karr, H. Rosenthal, T. M. Chu, E. A. Mirand and G. P. Murphy (1983). "LNCaP model of human prostatic carcinoma." Cancer research **43**: 1809-1818.
- Hsieh, C.-L., T. Fei, Y. Chen, T. Li, Y. Gao, X. Wang, T. Sun, C. J. Sweeney, G.-S. M. Lee, S. Chen, S. P. Balk, X. S. Liu, M. Brown and P. W. Kantoff (2014). "Enhancer RNAs participate in androgen receptor-driven looping that

- selectively enhances gene activation." Proceedings of the National Academy of Sciences of the United States of America **111**: 7319-7324.
- Huang, D. W., B. T. Sherman, Q. Tan, J. R. Collins, W. G. Alvord, J. Roayaei, R. Stephens, M. W. Baseler, H. C. Lane and R. A. Lempicki (2007). "The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists." Genome Biol **8**(9): R183.
- Ilott, N. E., J. A. Heward, B. Roux, E. Tsitsiou, P. S. Fenwick, L. Lenzi, I. Goodhead, C. Hertz-Fowler, A. Heger, N. Hall, L. E. Donnelly, D. Sims and M. A. Lindsay (2014). "Long non-coding RNAs and enhancer RNAs regulate the lipopolysaccharide-induced inflammatory response in human monocytes." Nature Communications **5**: 3979.
- INCA (2016). "INCA - CÂNCER - Tipo - Próstata (<http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/prostata>)."
- Israel, B., D. Medical and B. Avenue (2010). "Gene Expression in Castration Resistant Prostate Cancer." **69**: 6027-6032.
- Jin, H. J., J. C. Zhao, L. Wu, J. Kim and J. Yu (2014). "Cooperativity and equilibrium with FOXA1 define the androgen receptor transcriptional program." Nat Commun **5**: 3972.
- Kamakaka, R. T. and S. Biggins (2005). "Histone variants: deviants?" Genes & development **19**: 295-310.
- Kapranov, P., J. Cheng, S. Dike, D. A. Nix, R. Dutttagupta, A. T. Willingham, P. F. Stadler, J. Hertel, J. Hackermüller, I. L. Hofacker, I. Bell, E. Cheung, J. Drenkow, E. Dumais, S. Patel, G. Helt, M. Ganesh, S. Ghosh, A. Piccolboni, V. Sementchenko, H. Tammanna and T. R. Gingeras (2007). "RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription." Science **316**.
- Kapranov, P., G. S. Laurent, T. Raz and F. Oszolák (2010). "The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA." BMC.

- Keung, A. J., J. K. Joung, A. S. Khalil and J. J. Collins (2015). "Chromatin regulation at the frontier of synthetic biology." Nature reviews. Genetics **16**: 159-171.
- Khalil, A. M., M. Guttman, M. Huarte, M. Garber, A. Raj, D. Rivea Morales, K. Thomas, A. Presser, B. E. Bernstein, A. van Oudenaarden, A. Regev, E. S. Lander and J. L. Rinn (2009). "Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression." Proceedings of the National Academy of Sciences of the United States of America **106**: 11667-11672.
- Khalil, A. M. and J. L. Rinn (2011). "RNA–protein interactions in human health and disease." Seminars in Cell & Developmental Biology **22**: 359-365.
- Khurana, E., Y. Fu, D. Chakravarty, F. Demichelis, M. A. Rubin and M. Gerstein (2016). "Role of non-coding sequence variants in cancer." Nat Rev Genet **17**(2): 93-108.
- Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley and S. L. Salzberg (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." Genome Biology **14**: R36.
- Kim, J.-Y., T. Banerjee, A. Vinckevicius, Q. Luo, J. B. Parker, M. R. Baker, I. Radhakrishnan, J.-J. Wei, G. D. Barish and D. Chakravarti (2014). "A Role for WDR5 in Integrating Threonine 11 Phosphorylation to Lysine 4 Methylation on Histone H3 during Androgen Signaling and in Prostate Cancer." Molecular Cell **54**: 613-625.
- Kim, K., K. Jang, W. Yang, E.-Y. Choi, S.-m. Park, M. Bae, Y.-j. Kim and J. K. Choi (2016). "Chromatin structure-based prediction of recurrent noncoding mutations in cancer." Nature genetics **48**: 1321-1326.
- Kim, K.-H., A. Dobi, S. Shaheduzzaman, C. L. Gao, K. Masuda, H. Li, A. Drukier, Y. Gu, V. Srikantan, J. S. Rhim and S. Srivastava (2007). "Characterization of the androgen receptor in a benign prostate tissue-derived human prostate epithelial cell line: RC-165N/human telomerase reverse

- transcriptase." Prostate cancer and prostatic diseases **10**: 30-38.
- Kim, T.-K., M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, E. Markenscoff-Papadimitriou, D. Kuhl, H. Bito, P. F. Worley, G. Kreiman and M. E. Greenberg (2010). "Widespread transcription at neuronal activity-regulated enhancers." Nature **465**: 182-187.
- Kouzarides, T. (2007). "Chromatin Modifications and Their Function." Cell **128**: 693-705.
- Krell, J., J. Stebbing, C. Carissimi, A. F. Dabrowska, A. D. Giorgio, A. E. Frampton, V. Harding, V. Fulci, G. Macino, T. Colombo and L. Castellano (2016). "TP53 regulates miRNA association with AGO2 to remodel the miRNA-mRNA interaction network." Genome Research **26**: 331-341.
- Lai, F. and R. Shiekhattar (2014). Enhancer RNAs: The new molecules of transcription. Current Opinion in Genetics and Development, NIH Public Access. **25**: 38-42.
- Lai, J., M.-A. Kedda, K. Hinze, R. L. G. Smith, J. Yaxley, A. B. Spurdle, C. P. Morris, J. Harris and J. A. Clements (2007). "PSA/KLK3 ARE1 promoter polymorphism alters androgen receptor binding and is associated with prostate cancer susceptibility." Carcinogenesis **28**: 1032-1039.
- Lamont, K. R. and D. J. Tindall (2011). "Minireview: Alternative Activation Pathways for the Androgen Receptor in Prostate Cancer." Molecular Endocrinology **25**: 897-907.
- Langley, E., J. A. Kemppainen and E. M. Wilson (1998). "Intermolecular NH₂-/Carboxyl-terminal Interactions in Androgen Receptor Dimerization Revealed by Mutations That Cause Androgen Insensitivity." Journal of Biological Chemistry **273**: 92-101.
- Langmead, B., C. Trapnell, M. Pop and S. L. Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biology **10**.

- Lawrence, M. G., J. Lai and J. A. Clements (2010). "Kallikreins on steroids: Structure, function, and hormonal regulation of prostate-specific antigen and the extended kallikrein locus." Endocrine Reviews **31**: 407-446.
- Le, T. B. K., M. V. Imakaev, L. A. Mirny and M. T. Laub (2013). "High-resolution mapping of the spatial organization of a bacterial chromosome." Science (New York, N.Y.) **342**: 731-734.
- Lesne, A., J. Riposo, P. Roger, A. Cournac and J. Mozziconacci (2014). "3D genome reconstruction from chromosomal contacts." Nat Methods **11**: 1141-1143.
- Li, Y., D. Y. Zhao, J. F. Greenblatt and Z. Zhang (2013). "RIPSeeker: A statistical package for identifying protein-associated transcripts from RIP-seq experiments." Nucleic Acids Research **41**: 1-18.
- Liberzon, A., C. Birger, H. Thorvaldsdóttir, M. Ghandi, J. P. Mesirov and P. Tamayo (2015). "The Molecular Signatures Database Hallmark Gene Set Collection." Cell Systems **1**: 417-425.
- Lin, C. Y., J. Lovén, P. B. Rahl, R. M. Paranal, C. B. Burge, J. E. Bradner, T. I. Lee and R. A. Young (2012). "Transcriptional amplification in tumor cells with elevated c-Myc." Cell **151**: 56-67.
- Louro, R., H. I. Nakaya, P. P. Amaral, F. Festa, M. C. Sogayar, A. M. da Silva, S. Verjovski-Almeida and E. M. Reis (2007). "Androgen responsive intronic non-coding RNAs." BMC biology **5**: 4.
- Love, M. I., W. Huber and S. Anders (2014). "Beginner's guide to using the DESeq2 package." Genome biology **15**: 550.
- Lovén, J., H. A. Hoke, C. Y. Lin, A. Lau, D. A. Orlando, C. R. Vakoc, J. E. Bradner, T. I. Lee and R. A. Young (2013). "Selective inhibition of tumor oncogenes by disruption of super-enhancers." Cell **153**: 320-334.
- Lu, Z., X. Guan, C. A. Schmidt and A. G. Matera (2014). "RIP-seq analysis of eukaryotic Sm proteins identifies three major

- categories of Sm-containing ribonucleoproteins." Genome biology **15**: R7.
- Luco, R. F., Q. Pan, K. Tominaga, B. J. Blencowe, O. M. Pereira-Smith and T. Misteli (2010). "Regulation of alternative splicing by histone modifications." Science (New York, N.Y.) **327**: 996-1000.
- Lupiáñez, D. G., K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel and S. Mundlos (2015). "Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions." Cell **161**: 1012-1025.
- Mahajan, N. P., Y. Liu, S. Majumder, M. R. Warren, C. E. Parker, J. L. Mohler, H. S. Earp and Y. E. Whang (2007). "Activated Cdc42-associated kinase Ack1 promotes prostate cancer progression via androgen receptor tyrosine phosphorylation." Proceedings of the National Academy of Sciences of the United States of America **104**: 8438-8443.
- Mangelsdorf, D. J., C. Thummel, M. Beato, P. Herrlich, G. Schutz, K. Umesono, B. Blumberg, P. Kastner, M. Mark, P. Chambon and R. M. Evans (1995). "The nuclear receptor superfamily: the second decade." Cell **83**(6): 835-839.
- Marques, M., L. Laflamme, A. L. Gervais and L. Gaudreau (2010). "Reconciling the positive and negative roles of histone H2A.Z in gene transcription." Epigenetics **5**: 267-272.
- Marusyk, A. and K. Polyak (2011). "Tumor heterogeneity: causes and consequences." Biochim Biophys Acta **1805**: 1-28.
- Mason, M. D., G. Davies and W. G. Jiang (2002). "Cell adhesion molecules and adhesion abnormalities in prostate cancer." Critical Reviews in Oncology/Hematology **41**: 11-28.
- Massie, C. E., B. Adryan, N. L. Barbosa-Morais, A. G. Lynch, M. G. Tran, D. E. Neal and I. G. Mills (2007). "New androgen

- receptor genomic targets show an interaction with the ETS1 transcription factor." EMBO reports **8**: 871-878.
- Mattick, J. S. (2012). "RNA driving the epigenetic bus." The EMBO Journal **31**.
- McEwan, I. J. (2012). "Intrinsic disorder in the androgen receptor: identification, characterisation and drugability." Mol. BioSyst. **8**: 82-90.
- McEwan, I. J. and J.-Å. Gustafsson (1997). "Interaction of the human androgen receptor transactivation function with the general transcription factor TFIIF." Biochemistry **94**: 8485-8490.
- Ménez, R., S. Michel, B. H. Muller, M. Bossus, F. Ducancel, C. Jolivet-Reynaud and E. A. Stura (2008). "Crystal structure of a ternary complex between human prostate-specific antigen, its substrate acyl intermediate and an activating antibody." Journal of molecular biology **376**: 1021-1033.
- Menze, B. H., B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich and F. A. Hamprecht (2009). "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data." BMC Bioinformatics **10**: 213.
- Metzger, E., M. Wissmann, N. Yin, J. M. Müller, R. Schneider, A. H. F. M. Peters, T. Günther, R. Buettner and R. Schüle (2005). "LSD1 demethylates repressive histone marks to promote androgen-receptor-dependent transcription." Nature **437**: 25-28.
- Metzger, E., N. Yin, M. Wissmann, N. Kunowska, K. Fischer, N. Friedrichs, D. Patnaik, J. M. G. Higgins, N. Potier, K.-H. Scheidtmann, R. Buettner and R. Schüle (2008). "Phosphorylation of histone H3 at threonine 11 establishes a novel chromatin mark for transcriptional regulation." Nature Cell Biology **10**: 53-60.
- Miano, V., G. Ferrero, S. Reineri, L. Caizzi, L. Annaratone, L. Ricci, S. Cutrupi, I. Castellano, F. Cordero and M. De Bortoli (2015). "Luminal long non-coding RNAs regulated by

- estrogen receptor alpha in a ligand-independent manner show functional roles in breast cancer." Oncotarget **7**: 1-16.
- Mortensen, M. M., S. Høyer, A.-S. Lynnerup, T. F. Ørntoft, K. D. Sørensen, M. Borre and L. Dyrskjød (2015). "Expression profiling of prostate cancer tissue delineates genes associated with recurrence after prostatectomy." Scientific reports **5**: 16018.
- Naumova, N., M. Imakaev, G. Fudenberg, Y. Zhan, B. R. Lajoie, L. A. Mirny and J. Dekker (2013). "Organization of the mitotic chromosome." Science (New York, N.Y.) **342**: 948-953.
- Ngan, S., E. A. Stronach, A. Photiou, J. Waxman, S. Ali and L. Buluwela (2009). "Microarray coupled to quantitative RT-PCR analysis of androgen-regulated genes in human LNCaP prostate cancer cells." Oncogene **28**: 2051-2063.
- Nichols, M. H. and V. G. Corces (2015). "A CTCF Code for 3D Genome Architecture." Cell **162**: 703-705.
- Nuclear Receptors Nomenclature, C. (1999). "A unified nomenclature system for the nuclear receptor superfamily." Cell **97**(2): 161-163.
- O'Geen, H., L. Echipare and P. J. Farnham (2011). Using ChIP-Seq Technology to Generate High-Resolution Profiles of Histone Modifications. The American Journal of Medicine. **119**: 265-286.
- Ong, C.-T. and V. G. Corces (2014). "CTCF: an architectural protein bridging genome topology and function." Nature Reviews Genetics **15**: 234-246.
- Orom, U. A., T. Derrien, M. Beringer, K. Gumireddy, A. Gardini, G. Bussotti, F. Lai, M. Zytnicki, C. Notredame, Q. Huang, R. Guigo and R. Shiekhattar (2010). "Long noncoding RNAs with enhancer-like function in human cells." Cell **143**: 46-58.
- Paller, C. J. and E. S. Antonarakis (2013). "Management of biochemically recurrent prostate cancer after local therapy: evolving standards of care and new directions." Clinical advances in hematology & oncology : H&O **11**: 14-23.

- Park, C., H. Rehrauer and I. M. Mansuy (2013). "Genome-wide analysis of H4K5 acetylation associated with fear memory in mice." BMC Genomics **14**: 539.
- Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay (2012). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research **12**: 2825-2830.
- Perissi, V., K. Jepsen, C. K. Glass and M. G. Rosenfeld (2010). "Deconstructing repression: evolving models of co-repressor action." Nature Reviews Genetics **11**: 109-123.
- Polak, P., R. Karlic, A. Koren, R. Thurman, R. Sandstrom, M. S. Lawrence, A. Reynolds, E. Rynes, K. Vlahovicek, J. A. Stamatoyannopoulos and S. R. Sunyaev (2015). "Cell-of-origin chromatin organization shapes the mutational landscape of cancer." Nature **518**: 360-364.
- Pratt, W. B. and D. O. Toft (1997). "Steroid Receptor Interactions with Heat Shock Protein and Immunophilin Chaperones ¹." Endocrine Reviews **18**: 306-360.
- Prensner, J. R. and A. M. Chinnaiyan (2011). The emergence of lncRNAs in cancer biology. Cancer Discovery. **1**: 391-407.
- Prensner, J. R., M. K. Iyer, O. A. Balbin, S. M. Dhanasekaran, Q. Cao, J. C. Brenner, B. Laxman, I. A. Asangani, C. S. Grasso, H. D. Kominsky, X. Cao, X. Jing, X. Wang, J. Siddiqui, J. T. Wei, D. Robinson, H. K. Iyer, N. Palanisamy, C. A. Maher and A. M. Chinnaiyan (2011). "Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression." Nature Biotechnology **29**: 742-749.
- Prensner, J. R., M. K. Iyer, A. Sahu, I. A. Asangani, Q. Cao, L. Patel, I. A. Vergara, E. Davicioni, N. Erho, M. Ghadessi, R. B. Jenkins, T. J. Triche, R. Malik, R. Bedenis, N. McGregor, T. Ma, W. Chen, S. Han, X. Jing, X. Cao, X. Wang, B. Chandler, W. Yan, J. Siddiqui, L. P. Kunju, S. M.

- Dhanasekaran, K. J. Pienta, F. Y. Feng and A. M. Chinnaiyan (2013). "The long noncoding RNA SchLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex." Nature genetics **45**: 1392-1398.
- Prescott, J., U. Jariwala, L. Jia, J. P. Cogan, A. Barski, S. Pregizer, H. C. Shen, A. Arasheben, J. J. Neilson, B. Frenkel and G. A. Coetzee (2007). "Androgen receptor-mediated repression of novel target genes." The Prostate **67**: 1371-1383.
- Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." BIOINFORMATICS APPLICATIONS NOTE **26**: 841-84210.
- Quinn, J. J. and H. Y. Chang (2016). "Unique features of long non-coding RNA biogenesis and function." Nature Publishing Group **17**: 47-62.
- Rada-Iglesias, A., R. Bajpai, T. Swigut, S. A. Brugmann, R. A. Flynn and J. Wysocka (2011). "A unique chromatin signature uncovers early developmental enhancers in humans." Nature **470**: 279-283.
- Rajagopal, N., W. Xie, Y. Li, U. Wagner, W. Wang, J. Stamatoyannopoulos, J. Ernst, M. Kellis and B. Ren (2013). "RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State." PLoS Computational Biology **9**.
- Rao, S. S. P., M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander and E. L. Aiden (2014). "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." Cell **159**: 1665-1680.
- Reebye, V., L. Querol Cano, D. N. Lavery, G. N. Brooke, S. M. Powell, D. Chotai, M. M. Walker, H. C. Whitaker, R. Wait, H. C. Hurst and C. L. Bevan (2012). "Role of the HSP90-Associated Cochaperone p23 in Enhancing Activity of the Androgen Receptor and Significance for Prostate Cancer." Molecular Endocrinology **26**: 1694-1706.

- Reis, E. M. and S. Verjovski-Almeida (2012). "Perspectives of long non-coding RNAs in cancer diagnostics." Frontiers in Genetics **3**: 32.
- Ren, S., Z. Peng, J. H. Mao, Y. Yu, C. Yin, X. Gao, Z. Cui, J. Zhang, K. Yi, W. Xu, C. Chen, F. Wang, X. Guo, J. Lu, J. Yang, M. Wei, Z. Tian, Y. Guan, L. Tang, C. Xu, L. Wang, X. Gao, W. Tian, J. Wang, H. Yang, J. Wang and Y. Sun (2012). "RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings." Cell Res **22**(5): 806-821.
- Rinn, J. L., M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal and H. Y. Chang (2007). "Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs." Cell **129**: 1311-1323.
- Roadmap Epigenomics, C., A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shores, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker,

- M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang and M. Kellis (2015). "Integrative analysis of 111 reference human epigenomes." Nature **518**(7539): 317-330.
- Roth, J. A., R. Gulati, J. L. Gore, M. R. Cooperberg and R. Etzioni (2016). "Economic Analysis of Prostate-Specific Antigen Screening and Selective Treatment Strategies." JAMA Oncol **2**(7): 890-898.
- Roy, A. K., R. K. Tyagi, C. S. Song, Y. Lavrovsky, S. C. Ahn, T. S. Oh and B. Chatterjee (2001). "Androgen receptor: structural domains and functional dynamics after ligand-receptor interaction." Annals of the New York Academy of Sciences **949**: 44-57.
- Sakurai, K., B. J. Reon, J. Anaya and a. Dutta (2015). "The lncRNA DRAIC/PCAT29 Locus Constitutes a Tumor-Suppressive Nexus." Molecular Cancer Research **13**: 828-838.
- Schmitt, A. D., M. Hu and B. Ren (2016). "Genome-wide mapping and analysis of chromosome architecture." Nature Reviews Molecular Cell Biology **17**.
- Schulman, C. C., J. Irani, J. Morote, J. A. Schalken, F. Montorsi, P. L. Chlosta and A. Heidenreich (2010). "Testosterone Measurement in Patients with Prostate Cancer." European Urology **58**: 65-74.
- Schwartz, S., E. Meshorer and G. Ast (2009). "Chromatin organization marks exon-intron structure." Nature Structural & Molecular Biology **16**: 990-995.
- Segal, E. and J. Widom (2009). "From DNA sequence to transcriptional behaviour: a quantitative approach." Nature Reviews Genetics **10**: 443-456.
- Sevilla, A. and O. Binda (2014). "Post-translational modifications of the histone variant H2AZ." Stem cell research **12**: 289-295.
- Shaffer, P. L., A. Jivan, D. E. Dollins, F. Claessens and D. T. Gewirth (2004). "Structural basis of androgen receptor

- binding to selective androgen response elements." Proceedings of the National Academy of Sciences of the United States of America **101**: 4758-4763.
- Shang, Y., M. Myers and M. Brown (2002). "Formation of the androgen receptor transcription complex." Molecular cell **9**: 601-610.
- Shim, E. Y., C. Woodcock and K. S. Zaret (1998). "Nucleosome positioning by the winged helix transcription factor HNF3." Genes & Development **12**: 5-10.
- Shlyueva, D., G. Stampfel and A. Stark (2014). "Transcriptional enhancers: from properties to genome-wide predictions." Nature Reviews Genetics **15**: 272-286.
- SIM "DATASUS - SIM - Sistema de Informação sobre Mortalidade." <http://datasus.saude.gov.br/sistemas-e-aplicativos/eventos-v/sim-sistema-de-informacoes-de-mortalidade>.
- Simon, M. D., S. F. Pinter, R. Fang, K. Sarma, M. Rutenberg-Schoenberg, S. K. Bowman, B. A. Kesner, V. K. Maier, R. E. Kingston and J. T. Lee (2013). "High-resolution Xist binding maps reveal two-step spreading during X-chromosome inactivation." Nature **504**: 465-469.
- Song, L. and G. E. Crawford (2010). "DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells." Cold Spring Harbor Protocols **2010**: pdb.prot5384-pdb.prot5384.
- Song, W., K. Wang, R.-J. Zhang, Q.-X. Dai and S.-B. Zou (2016). "Long noncoding RNA GAS5 can predict metastasis and poor prognosis: a meta-analysis." Minerva medica **107**: 70-76.
- Spizzo, R., M. I. Almeida, A. Colombatti and G. A. Calin (2012). "Long non-coding RNAs and cancer: a new frontier of translational research?" Oncogene **31**: 4577-4587.
- Stricker, S. H., A. Köferle and S. Beck (2016). "Non-genetic factors contribute to many cellular functions, traits and phenotypes." Nature Publishing Group.

- Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov (2005). "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles." Proc Natl Acad Sci U S A **102**(43): 15545-15550.
- Sun, M., S. S. Gadad, D. S. Kim and W. L. Kraus (2015). "Discovery, Annotation, and Functional Analysis of Long Noncoding RNAs Controlling Cell-Cycle Gene Expression and Proliferation in Breast Cancer Cells." Molecular Cell **59**: 698-711.
- Sun, S., X. Zhong, C. Wang, H. Sun, S. Wang, T. Zhou, R. Zou, L. Lin, N. Sun, G. Sun, Y. Wu, B. Wang, X. Song, L. Cao and Y. Zhao (2016). "BAP18 coactivates androgen receptor action and promotes prostate cancer progression." Nucleic Acids Res **44**: 8112-8128.
- Taberlay, P. C., J. Achinger-Kawecka, A. T. L. Lun, F. A. Buske, K. Sabir, C. M. Gould, E. Zotenko, S. A. Bert, K. A. Giles, D. C. Bauer, G. K. Smyth, C. Stirzaker, S. I. O'Donoghue and S. J. Clark (2016). "Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations." Genome Research **26**: 719-731.
- Taft, R. J., K. C. Pang, T. R. Mercer, M. Dinger and J. S. Mattick (2010). Non-coding RNAs: Regulators of disease. Journal of Pathology. **220**: 126-139.
- Tan, M. E., J. Li, H. E. Xu, K. Melcher and E.-I. Yong (2015). "Androgen receptor: structure, role in prostate cancer and drug discovery." Acta Pharmacologica Sinica **36**: 3-23.
- Tan, P. Y., C. W. Chang, K. R. Chng, K. D. S. A. Wansa, W.-K. Sung and E. Cheung (2012). "Integration of regulatory networks by NKX3-1 promotes androgen-dependent prostate cancer survival." Molecular and cellular biology **32**: 399-414.
- Thurman, R. E., E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H.

- Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kuttyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford and J. A. Stamatoyannopoulos (2012). "The accessible chromatin landscape of the human genome." Nature **489**(7414): 75-82.
- Trapman, J., P. Klaassen, G. G. Kuiper, J. A. van der Korput, P. W. Faber, H. C. van Rooij, A. Geurts van Kessel, M. M. Voorhorst, E. Mulder and A. O. Brinkmann (1988). "Cloning, structure and expression of a cDNA encoding the human androgen receptor." Biochemical and biophysical research communications **153**: 241-248.
- Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn and L. Pachter (2012). "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks." Nature protocols **7**: 562-578.
- Tsai, M.-C., O. Manor, Y. Wan, N. Mosammaparast, J. K. Wang, F. Lan, Y. Shi, E. Segal and H. Y. Chang (2010). "Long noncoding RNA as modular scaffold of histone modification complexes." Science (New York, N.Y.) **329**: 689-693.
- Ulitsky, I. and D. P. Bartel (2013). "lincRNAs: Genomics, Evolution, and Mechanisms." Cell **154**: 26-46.
- Valdés-Mora, F., J. Z. Song, A. L. Statham, D. Strbenac, M. D. Robinson, S. S. Nair, K. I. Patterson, D. J. Tremethick, C. Stirzaker and S. J. Clark (2012). "Acetylation of H2A.Z is a key epigenetic modification associated with gene

- deregulation and epigenetic remodeling in cancer." Genome research **22**: 307-321.
- Van Bortle, K., M. H. Nichols, L. Li, C. T. Ong, N. Takenaka, Z. S. Qin and V. G. Corces (2014). "Insulator function and topological domain border strength scale with architectural protein occupancy." Genome Biol **15**(6): R82.
- van Laar, J. H., J. Bolt-de Vries, M. M. Voorhorst-Ogink and A. O. Brinkmann (1989). "The human androgen receptor is a 110 kDa protein." Molecular and cellular endocrinology **63**: 39-44.
- Wan, X., W. Huang, S. Yang, Y. Zhang, H. Pu, F. Fu, Y. Huang, H. Wu, T. Li and Y. Li (2016). "Identification of androgen-responsive lncRNAs as diagnostic and prognostic markers for prostate cancer." Oncotarget **7**(37): 60503-60518.
- Wang, D., I. Garcia-Bassets, C. Benner, W. Li, X. Su, Y. Zhou, J. Qiu, W. Liu, M. U. Kaikkonen, K. a. Ohgi, C. K. Glass, M. G. Rosenfeld and X.-D. Fu (2011). "Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA." Nature **474**: 390-394.
- Wapinski, O. and H. Y. Chang (2011). "Long noncoding RNAs and human disease." Trends in Cell Biology **21**: 354-361.
- Wei Huang, D., B. T. Sherman, Q. Tan, J. R. Collins, W. Gregory Alvord, J. Roayaei, R. Stephens, M. W. Baseler, H. Clifford Lane and R. A. Lempicki (2007). "The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists." Genome Biology **8**.
- White, R. J. (2005). "RNA polymerases I and III, growth control and cancer." Nature reviews. Molecular cell biology **6**: 69-78.
- Wit, E. D. and W. D. Laat (2012). "A decade of 3C technologies- insights into nuclear organization." Genes & development: 11-24.
- Yang, F., X. Deng, W. Ma, J. B. Berletch, N. Rabaia, G. Wei, J. M. Moore, G. N. Filippova, J. Xu, Y. Liu, W. S. Noble, J. Shendure and C. M. Disteche (2015). "The lncRNA Firre

anchors the inactive X chromosome to the nucleolus by binding CTCF and maintains H3K27me3 methylation." Genome Biology **16**: 52.

- Yang, L., C. Lin, C. Jin, J. C. Yang, B. Tanasa, W. Li, D. Merkurjev, K. A. Ohgi, D. Meng, J. Zhang, C. P. Evans and M. G. Rosenfeld (2013). "lncRNA-dependent mechanisms of androgen-receptor-regulated gene activation programs." Nature **500**: 598-602.
- Zhang, Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nussbaum, R. M. Myers, M. Brown, W. Li and X. S. Liu (2008). "Model-based Analysis of ChIP-Seq (MACS)." Genome Biology **9**: R137.
- Zhao, J., T. K. Ohsumi, J. T. Kung, Y. Ogawa, D. J. Grau, K. Sarma, J. J. Song, R. E. Kingston, M. Borowsky and J. T. Lee (2010). "Genome-wide Identification of Polycomb-Associated RNAs by RIP-seq." Molecular Cell **40**: 939-953.
- Zhao, R., T. Nakamura, Y. Fu, Z. Lazar and D. L. Spector (2011). "Gene bookmarking accelerates the kinetics of post-mitotic transcriptional re-activation." Nature Cell Biology **13**: 1295-1304.
- Zilberman, D., D. Coleman-Derr, T. Ballinger and S. Henikoff (2008). "Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks." Nature **456**: 125-129.
- Zoubeidi, A., A. Zardan, E. Beraldi, L. Fazli, R. Sowery, P. Rennie, C. Nelson and M. Gleave (2007). "Cooperative Interactions between Androgen Receptor (AR) and Heat-Shock Protein 27 Facilitate AR Transcriptional Activity." Cancer Research **67**.