UNIVERSIDADE FEDERAL DO PARANÁ

SETOR DE CIÊNCIAS BIOLÓGICAS

ALESSANDRO MAX

INTEGRAÇÃO DE DADOS DE GENÔMICA FUNCIONAL E GENÉTICA DE POPULAÇÕES PARA INTERPRETAR POLIMORFISMOS NÃO CODIFICADORES ASSOCIADOS AOS NÍVEIS DE EXPRESSÃO DOS GENES DO SISTEMA COMPLEMENTO

CURITIBA

Março / 2015

ALESSANDRO MAX

INTEGRAÇÃO DE DADOS DE GENÔMICA FUNCIONAL E GENÉTICA DE POPULAÇÕES PARA INTERPRETAR POLIMORFISMOS NÃO CODIFICADORES ASSOCIADOS AOS NÍVEIS DE EXPRESSÃO DOS GENES DO SISTEMA COMPLEMENTO

Projeto de Pesquisa apresentado ao trabalho de conclusão de curso para obtenção do titulo de Graduação em Ciências Biológicas da Universidade Federal do Paraná.

Orientadora Profª. Drª. Angelica B. Winter Boldt

Co-orientador: Dr. Rodrigo Coutinho de Almeida

Curitiba

Março/2015

AGRADECIMENTOS

A minha querida orientadora Profª. Drª. Angelica B. Winter Boldt, pela oportunidade de estar trabalhando nesta área maravilhosa da ciência, por todo aprendizado, incentivo e contribuição na conclusão deste trabalho, e por me instruir a seguir na área da bioinformática.

Ao meu co-orientador Drº. Rodrigo C. de Almeida, pelo grande aprendizado e contribuição na realização deste trabalho, e pelo momento oportuno que chegou a esta universidade.

A minha carinhosa Família, pelo incentivo e apoio.

Ao grupo LGMH.

A todos aqueles pesquisadores que produziram tal imensidão de trabalhos, ainda inexplorada em toda sua magnitude. *"Se vi mais longe foi por estar de pé sobre ombros de gigantes." Isaac Newton*

Ao ser celestial, uno, Deus, e seu enigmático caminho a que me enveredo.

RESUMO

Polimorfismos (SNPs) associados a variações na expressão gênica (ao nível de mRNA) são denominados eQTLs (expression Quantitative Trait Loci). O projeto ENCODE (Encyclopedia of Non Coding Elements) têm expandido o volume de dados sobre elementos reguladores do DNA, possibilitando sua associação com os eQTLs. Dentre os fenótipos complexos possivelmente associados a eQTLs, estão os níveis séricos de proteínas da via das lectinas do sistema complemento,como a lectina ligante de manose (gene *MBL2*) e ficolinas (genes *FCN1, FCN2* e *FCN3*. Embora vários SNPs dos respectivos genes estejam associados com a modulação dos níveis séricos de seus produtos, assim como a uma maior susceptibilidade a doenças infecciosas e autoimunes, a estratificação da população e o desequilíbrio de ligação (DL) a outros SNPs dificultam saber se são causais. Para ajudar a esclarecer esta questão, elaborou-se uma abordagem integrativa de genômica funcional e genética de populações, começando pela reunião de eQTLs identificados em bancos de dados para *FCN1, FCN2, FCN3* e *MBL2*, junto a SNPs em DL na população eurodescendente. Estes foram ordenados de acordo com escores Z normalizados, construídos com base na frequência de sobreposição a elementos reguladores, dentro de cada bloco de DL. Para gerar novas hipóteses sobre o mecanismo de ação dos eQTLs, avaliou-se a topologia da cromatina e distribuição de elementos reguladores-chave, como grau de metilação da lisina 4 da histona H3, nos sítios contendo eQTLs com escores mais elevados. Para *FCN1*, identificaram-se 406 SNPs (188 eQTLs e 213 SNPs em DL), dos quais 97,5% associados a pelo menos um elemento regulador da expressão gênica em células sanguíneas (P<0,003), distribuídos em 16 blocos de desequilíbrio de ligação em uma sequência que engloba os genes *FCN2, OLFM1* e *COL5A1*, além de *FCN1* e vários lincRNAs. Trinta e sete destes SNPs apresentaram escores Z mais elevados, dos quais onze estão sobrepostos a proteínas de ligação, e três a RNAs não codificantes, em domínios topológicos específicos da cromatina. Também se observou uma correlação negativa entre os níveis de expressão de *FCN1* e seu gene vizinho, *OLFM1*, associados a 70 dos 406 SNPs, mas não houve correlação alguma entre a expressão de *FCN1* e do seu gene parálogo, a *FCN2*. Para *FCN2*, identificou-se apenas quatro SNPs, e para *FCN3*, dois SNPs em uma região de intensa atividade regulatória. Para *MBL2,* foram encontrados 9 SNPs em 6 blocos de DL, associados à regulação da expressão deste gene no fígado. Tomados em conjunto, os resultados nos levam a sugerir um forte papel regulatório da região intergênica destes genes, contextualizado por interações de longa distância da cromatina que podem caracterizar importantes regiões intensificadoras da regulação gênica. A partir dessa abordagem integrativa, foi possível priorizar SNPs que podem ser candidatos funcionais em um painel de referência para futuros estudos de associação com foco em genes do complemento, em doenças complexas.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Muitos polimorfismos de nucleotídeo único (SNPs, do inglês *single nucleotide polymorphism*) que contribuem para a variação na expressão dos níveis de mRNA de genes já foram detectados no genoma humano (WESTRA, FRANKE *et al.*, 2013). Mais de 90% destas variantes, conhecidas como eQTLs (do inglês *expression quantitative trait loci*), e de SNPs identificados em estudos de associação de amplitude genômica (GWAS, do inglês *Genome Wide Association Study*), estão localizados em regiões não codificantes (MAHER, 2012), sendo provável que o mecanismo subjacente seja regulatório. Abordagens de genômica funcional têm fornecido informações em volume crescente, capazes de esclarecer os possíveis efeitos dos SNPs intrônicos e intergênicos associados a um fenótipo (RITCHIE *et al*., 2015). Estas abordagens são dadas por recentes avanços tecnológicos nos métodos de sequenciamento do genoma, transcriptoma, metiloma e relações metabólicas. Um objetivo essencial da análise destes dados é a identificação de modelos eficazes que predizem características e resultados fenotípicos, elucidando importantes biomarcadores e gerando *insights* sobre as bases genéticas de características complexas. Entretanto, o sucesso na compreensão do componente genético de fenótipos complexos tem sido modesto, e ainda há uma grande necessidade de estratégias avançadas e práticas de análise para aproveitar totalmente a utilidade destes dados de alto rendimento. Com a finalidade de priorizar SNPs possivelmente causais, reduzindo o número de falsas associações e ampliando hipóteses sobre o mecanismo regulatório subjacente às associações encontradas com a expressão gênica, têm-se buscado a integração de dados, combinando estudos genômicos de associação, análise de expressão gênica, evidências experimentais da atividade regulatória e o cenário tridimensional da cromatina (RITCHIE *et al.*, 2015).

A imunidade inata envolve um conjunto de moléculas capazes de identificar padrões moleculares associados à superfície de patógenos (PAMPs), distinguindo-os de auto-antígenos, assim como, de um modo semelhante, reconhecer padrões moleculares associados a células apoptóticas (ACAMPs), os quais não são expostos em células saudáveis (FRANC *et al.,* 1999). Várias destas moléculas fazem parte do sistema

complemento, que confere a primeira linha de defesa da imunidade inata, mas também atua estimulando e orientando a resposta adaptativa (FEARON *et al*., 1996). O complemento compreende mais de 35 proteínas plasmáticas e de membrana celular, cuja ativação proteolítica promove a opsonização e eliminação de patógenos. Além disso, está envolvido na eliminação de substâncias potencialmente perigosas, tais como detritos celulares e células danificadas (MATZINGER, 2002). O processo de ativação do sistema complemento se dá por três vias: clássica, alternativa e das lectinas . A via clássica inicia-se pela ligação de C1q a superfície de bactérias, na proteína C reativa ou a complexos antígeno-anticorpo. A via alternativa é ativada pela hidrólise espontânea de C3 ou pelo reconhecimento de PAMPs/ACAMPs pela properdina (MURPHY; TRAVERS; WALPORT, 2010a). A via das lectinas, descrita mais recentemente, é iniciada pela atividade proteolítica desencadeada pelas serinas proteases (MASPs) associadas à lectina ligante de manose (MBL), colectina 11 ou às ficolinas (FCN-1, FCN-2 e FCN-3), proteínas que também atuam como receptores de reconhecimento de padrões do tipo PAMPs e ACAMPs (PRRs) (BELTRAME *et al.,* 2015).

Variações nas concentrações circulantes das ficolinas e MBL já foram associadas a polimorfismos presentes na região promotora e mutações estruturais dos genes (BELTRAME *et al.,* 2015; GARRED et al. 2009; BOLDT *et al.*, 2006). A deficiência de MBL está associada a uma maior suscetibilidade a infecções por vírus, bactérias, fungos e protozoários, assim como a doenças autoimunes como a artrite reumatoide juvenil e lúpus eritematoso sistêmico. Por outro lado, níveis elevados de MBL são mais frequentes em certas doenças com intensa atividade inflamatória, como a artrite reumatoide de início tardio e febre reumática (ADDOBBATI *et al*., 2015). No caso de ficolina 2, baixos níveis circulantes da proteína foram associados com histórico de infecções respiratórias repetitivas em crianças. Haplótipos associados a estes níveis também foram associados à susceptibilidade à febre reumática (MESSIAS-REASON *et al*., 2009). Baixos níveis de ficolina 1 no soro foram recentemente associados com enterocolite necrosante fatal em recém-nascidos, e uma variação no promotor distal do gene, com susceptibilidade à artrite reumatoide (SCHLAPBACH *et al*., 2010; VAN DER CRUYSSEN *et al.*, 2007). A deficiência de *FCN3* é causada por uma deleção no exon 5, e reduz drasticamente a concentração sérica de ficolina 3 (GARRED et al., 2009).

Apesar destes resultados, não se sabe ainda se os SNPs não codificadores associados realmente alteram a expressão dos genes investigados, ou se as associações encontradas se devem a SNPs causais, em desequilíbrio de ligação com os primeiros. No presente trabalho, pretende-se ajudar a esclarecer esta questão, por meio da elaboração de uma abordagem integrativa de dados de genômica funcional e de genética de populações, sobre os polimorfismos não codificadores encontrados na região genômica que integra estes genes.

2. OBJETIVOS

2.1 Objetivo principal.

Propor uma estratégia de seleção de SNPs não codificadores, candidatos a uma associação causal com a expressão dos genes *FCN1*, *FCN2*, *FCN3* e *MBL2* da via das lectinas do complemento, com base em anotações de genômica funcional e em padrões de desequilíbrio de ligação (DL) entre estes SNPs, na população eurodescendente.

2.2 Objetivos específicos

a. Identificar eQTLs listados em bancos de dados públicos *online* para *FCN1*, *FCN2*, *FCN3* e *MBL2*;
b. Identificar SNPs correlacionados aos eQTLs, devido ao DL em populações de origem europeia;
c. Identificar as anotações de genômica funcional correspondentes a cada polimorfismo, disponíveis no banco de dados do projeto ENCODE (THE ENCODE PROJECT CONSORTIUM, 2013);
d. Atribuir escores Z para sumarizar estas anotações;
e. Verificar se os escores Z médios diferem para eQTLs e SNPs selecionados e para SNPs não selecionados, numa mesma região genômica;
f. Selecionar os SNPs com os maiores escores Z, dentre SNPs e eQTLs agrupados de acordo com os padrões de DL e de sinal de escore, na população eurodescendente;

g. Detalhar estes SNPs com base em dados de genética de populações e genômica funcional;
h. Avaliar possíveis associações físicas entre estes SNPs e os domínios topológicos da cromatina;
i. Gerar hipóteses sobre o mecanismo biológico subjacente às associações identificadas.

3. REVISÃO DA LITERATURA

3.1 Sistema Complemento

O processo de ativação do sistema complemento compreende três vias: clássica, alternativa e das lectinas (Figura 1). A via clássica inicia-se pela ligação de C1q a superfície de bactérias, a proteína C reativa ou a complexos antígeno-anticorpo. A via alternativa é ativada pela hidrólise espontânea de C3 ou pelo reconhecimento de PAMPs pela properdina. A via das lectinas é iniciada pela atividade proteolítica desencadeada pelas serinas proteases (MASP-1 e MASP-2) associadas a colectinas como a lectina ligante de manose (MBL) ou a ficolinas (FCN-1, FCN-2 e FCN-3), as quais reconhecem padrões de açúcares ou resíduos acetilados, respectivamente, presentes nos patógenos. Nesta via, além de realizar auto-ativação, MASP-1 ativa MASP-2, que cliva C4 e C2 em C4b2b, formando a enzima C3 convertase (HÉJA *et al.*, 2012). A cascata proteolítica desencadeada pelo processo de ativação permite uma grande amplificação e rapidez da resposta inflamatória, pois cada ativação de um elemento pré-formado pode gerar, rapidamente, múltiplas enzimas e complexos enzimáticos ativados (ABBAS; LICHTMAN; PILLAI, 2008), sendo capazes de ativar células do sistema imune para secretar citocinas pró-inflamatórias, tais como o interferon (IFN) -γ e o fator de necrose tumoral (TNF) -α (REN *et al*., 2014).

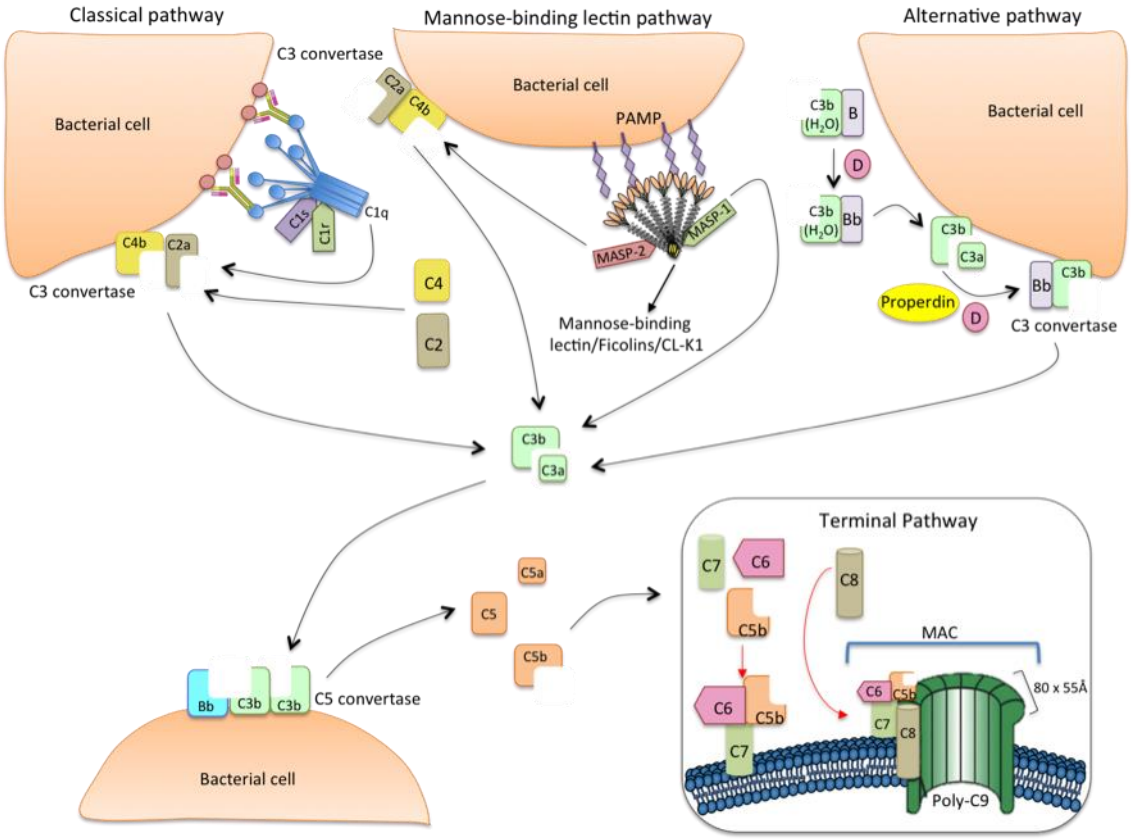

FIGURA 1 - As três vias de ativação do complemento, proteínas envolvidas e produtos.

NOTA: Subsequente à ativação da via clássica pela ligação de C1q na superfície de agente ativador, C1s cliva C4, que se liga de forma covalente à superfície do patógeno, e em seguida, cliva C2, levando à formação do complexo C4b2a, a C3 convertase da via clássica. A ativação da via das lectinas ocorre através da ligação de oligômeros de MBL, oligômeros de ficolina ou heteromorfos de colectina (CL-K1 + CL-L1), complexados a homodímeros de serina-proteases associadas a MBL-1 e 2 (MASP-1 e MASP-2, respectivamente), a vários grupos de hidratos de carbono ou acetilados sobre a superfície de patógenos (PAMPs: padrões moleculares associados a patógenos). Como C1s, MASP-2 induz a formação da C3 convertase, C4b2a, porém sua ativação é dependente de MASP-1. MASP-1 também cliva C2 e C3. A ativação da via alternativa depende de baixo grau de hidrólise espontânea de C3 no plasma que leva à formação de C3b. Este C3b liga o Fator B (homólogo a C2) para formar o complexo C3bB. A clivagem do Fator B pelo Fator D forma a convertase da via alternativa C3, C3bBb. A properdina estabiliza este complexo. A C3 convertase cliva C3 m em fragmentos C3b, que se ligam covalentemente ao lado do local de ativação do complemento (levando a opsonização). Esse processo amplifica a cascata e promove a fagocitose, bem como a resposta inflamatória e da imunidade adaptativa. Além disso, há a formação da C5 convertase que leva a clivagem de C5 em C5a e C5b e a formação do complexo de ataque à membrana (MAC). C5a atua como

uma potente anafilotoxina e C5b forma um complexo com C6 e C7, o qual é inserido na membrana celular. Posteriormente, moléculas de C8 e C9 (80 x 55 Å) ligam-se a este complexo, resultando em um complexo de ataque à membrana (C5b-9). As três vias convergem para esta via terminal comum, culminando com a lise celular e morte (ABBAS; LICHTMAN; PILLAI, 2008).

FONTE: BELTRAME *et al.,* 2015.

3.2 MBL e FCNs: moléculas iniciadoras da via das lectinas do complemento

A MBL reconhece oligossacarídeos como a manose e a N-acetil-D-glicosamina (GlcNAc) na superfície de microorganismos, ativando a via das lectinas do complemento, atuando como opsonina e mediando fagocitose. A estrutura molecular da MBL é caracterizada pela presença de uma sequência similar a do colágeno, correspondente ao sítio de ativação do complemento, e um domínio de reconhecimento de carboidratos (CRD), através do qual a MBL interage com os diferentes microorganismos. A MBL humana é composta por oligômeros de até 6 subunidades, sendo cada monômero formado por 3 cadeias polipeptídicas idênticas que dão origem a 3 sítios de CRD na região C-terminal (cálcio-dependente). Na região N-terminal, as cadeias interagem através de pontes dissulfeto formando o sítio de reconhecimento colagenoso (DOMMET *et al.*, 2006). A concentração sérica de MBL é profundamente afetada por mutações estruturais no gene, sendo também influenciada por polimorfismos na região promotora (BOLDT *et al.*, 2006). Na região promotora, foram identificadas pelo menos três mutações pontuais que influenciam os níveis circulantes de MBL (figura 2). Estes polimorfismos estão associados a uma maior suscetibilidade a infecções por vírus, bactérias, fungos e protozoários (DOMMET *et al*., 2006), afetando inclusive a resposta ao tratamento contra a infecção por vírus da hepatite C (ALVES PEDROSO *et al*., 2008). A deficiência de MBL tem um possível efeito protetor contra infecções por organismos intracelulares como o *Mycobacterium leprae*, que se utilizam de C3 ativado e receptores para C3b para infectar o hospedeiro (MESSIAS-REASON *et al*., 2007).

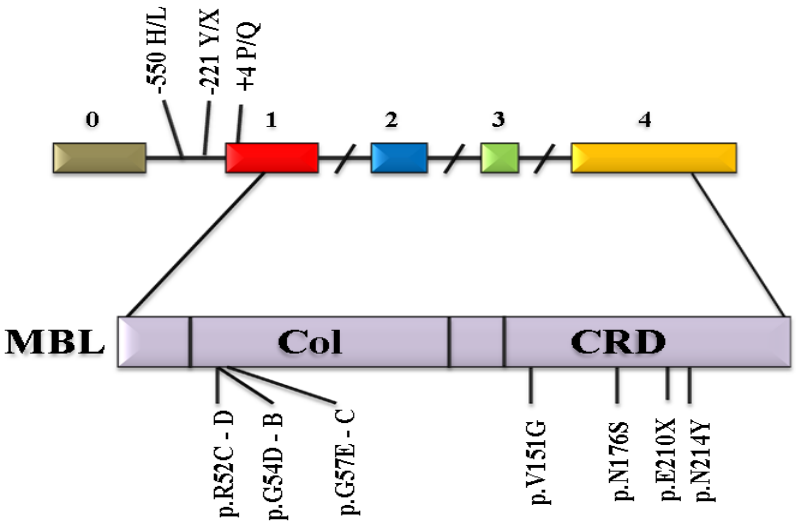

FIGURA 2 - Estrutura gênica, polimorfismos do gene *MBL2* e relação com os domínios da proteína.

LEGENDA: Exons estão representados pelas caixas coloridas. Exon 0 é não traduzido. MBL – lectina ligante de manose, Col – colágeno, CRD – Domínio de reconhecimento de carboidrato.
FONTE: modificado de BELTRAME *et al*., 2015.

Nos seres humanos, há três tipos de ficolina: ficolina-1 (M-ficolina), ficolina-2 (L-ficolina), e ficolina-3 (H-ficolina ou antígeno Hakata). Assim como *MBL*, a ficolina-2 tem a habilidade de ativar a via das lectinas do complemento após o reconhecimento de GlcNAc, mas também de ácido lipoteicoico (componente de parede celular de bactérias Gram-positivas e de grupos acetilados na superfície de patógenos como *Streptococcus pneumoniae* e *Staphylococcus aureus*). Além disso, ficolina-2 participa na eliminação de debris celulares e está envolvida na gravidade da nefropatia por IgA (THIEL e GADJEVA, 2009a). Também é capaz de se ligar à glicoproteína E1 do vírus da hepatite C (HCV), possivelmente contribuindo para eliminação do patógeno (ZHAN; ZHANG, 2007). O gene *FCN2* contém oito exons e é polimórfico (figura 3). Variações nas concentrações circulantes de *FCN-2* já foram associadas com a presença de 3 polimorfismos na região promotora, rs3124952, rs3124953 e rs17514136 (-986 A>G, -602 A>G, -4 A>G) e um polimorfismo no exon 8, rs7851696 (+6424 G>T Ala258Ser, G>C Ala258Pro). Haplótipos compostos por polimorfismos do promotor estão associados a baixos níveis circulantes da proteína, os quais foram associados com infecções respiratórias repetitivas em crianças (STENGAARD-PEDERSEN *et al*., 2003). Haplótipos

responsáveis por níveis normais desta proteína no plasma foram associados com a resistência à hanseníase (MESSIAS-REASON *et al.*, 2009). Também tem sido relatado que a ficolina-2 pode contribuir para a progressão da nefropatia pela imunoglobulina IgA (ROOS *et al.,* 2006) e que os polimorfismos do promotor podem influenciar na recorrência da doença de Behcet entre os portadores da variante de suceptibilidade *HLA*B51* (CHEN *et al.*, 2006). Um polimorfismo do promotor (rs17514136) foi associado à manifestação mais grave de lupus eritematoso, em cujas lesões é observada elevada deposição de produtos da ativação do complemento (ADDOBBATI *et al*., 2015).

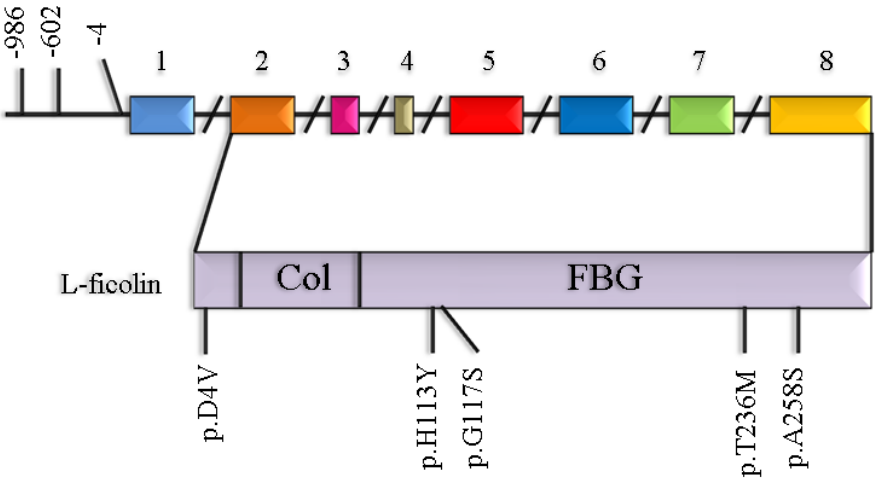

FIGURA 3 - Estrutura gênica, polimorfismos do gene *FCN2* e relação com os domínios da proteína.

LEGENDA: Exons estão representados pelas caixas coloridas. Col – colágeno, FBG – Domínio de fibrinogênio.

FONTE: modificado de BELTRAME *et al*., 2015.

Ao contrário de MBL, FCN-2 e FCN-3, que são proteínas séricas, FCN-1 (também conhecida como ficolina-M) é expressa em células mononucleares fagocíticas, atuando como receptor de superfície para patógenos. Existem evidências de que FCN-1, assim como FCN-2, interagem com carboidratos de diferentes tipos de bactérias e que FCN-1 aumenta a fagocitose desses microorganismos por monócitos (HONORE *et al*., 2007). Tanto *FCN1* quanto *FCN2* estão localizados em 9q34. O gene *FCN1* contém nove exons (Figura 4). A sequência de aminoácidos de *FCN-1* apresenta cerca de 80% de

homologia com a de *FCN-2*. Haplótipos do promotor também foram recentemente associados com a regulação dos níveis desta proteína (MUNTHE-FOG *et al.,* 2012). Os SNPs rs10120023 e rs10117466 (-542A-144C, respectivamente) apresentaram efeito protetor contra a hanseníase em Euro-brasileiros, e uma tendência para associação positiva da variante rs17039495 (-399A) foi detectada em afro-brasileiros (BOLDT *et al*., 2013). Em pacientes com artrite reumatoide inicial, níveis plasmáticos elevados de ficolina 1 foram correlacionados com um aumento na gravidade da doença, enquanto níveis baixos servem como preditor de remissão da doença (AMMITZOLL *et al*., 2013). Como se trata de um assunto bastante recente, o impacto do polimorfismo de *FCN1* em doenças humanas ainda foi pouco explorado.

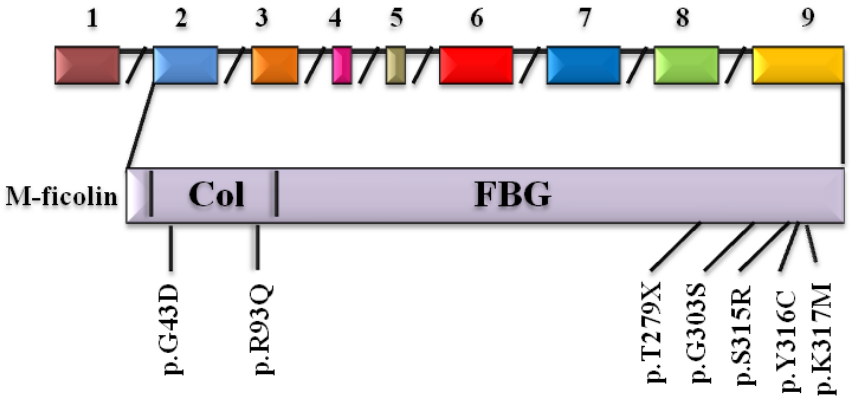

FIGURA 4 - Estrutura gênica, polimorfismos do gene *FCN1* e relação com os domínios da proteína.

LEGENDA: Exons estão representados pelas caixas coloridas. Exon 1 não é traduzido. Col – colágeno, FBG – Domínio de fibrinogênio.

FONTE: modificado de BELTRAME *et al*., 2015.

A concentração média da ficolina 3 (também conhecida como ficolina-H) no soro é de cerca de 18,4 µg/ml (KRARUP *et al*., 2005), comparado a 3,7 µg/ml para ficolina 2 (KILPATRICK *et al*., 1999), variando em até 10 vezes (3-54 µg/ml) (MUNTHE-FOG *et al*., 2008). O gene *FCN3* (Figura 5) ocorre em 1p35.3, é altamente conservado e apresenta cinco substituições não sinônimas, com frequências abaixo de 5%: p.Leu12Val, p.Leu117fs. Esta última produz uma proteína truncada, sem capacidade de ativação do complemento, e está associada com infecções repetitivas (HUMMELSHOJ et al. 2008; MUNTHE-FOG *et al.*, 2008).

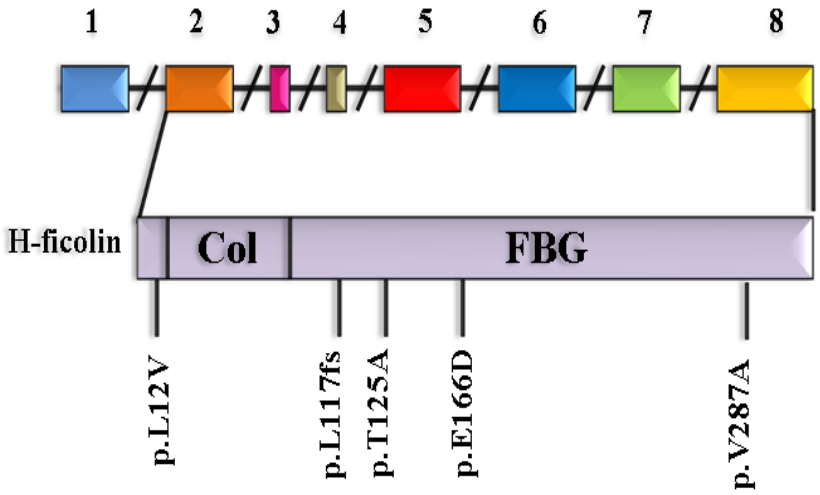

FIGURA 5 - Estrutura gênica, polimorfismos do gene *FCN*3 e relação com os domínios da proteína.

LEGENDA: Exons estão representados pelas caixas coloridas. Exon 1 não é traduzido. Col – colágeno, FBG – Domínio de fibrinogênio.

FONTE: modificado de BELTRAME *et al*., 2015.

Na circulação, a MBL e as ficolinas são encontradas formando um complexo com quatro proteínas relacionadas estruturalmente, as serina proteases associadas a MBL (MASPs) 1, 2 e 3, MAp44 e MAp19, sendo que estas duas últimas são versões truncadas de MASP-1 e MASP-2, respectivamente.

3.2 eQTLs (expression quantitative trait loci)

Desde a conclusão do Projeto Genoma Humano (LANDER *et al*., 2001) até a finalização do Projeto 1000 Genomas (1000 Genomes Project Consortium et al. 2015), acumulou-se uma riqueza de conhecimento sem precedentes sobre as variações da sequência do DNA humano. Contudo, muito pouco deste conhecimento foi traduzido para a compreensão de doenças humanas. O mapeamento de loci que podem influenciar a expressão gênica (eQTL) é uma das abordagens mais promissoras para preencher esta lacuna (COOKSON *et al*., 2009). Em termos práticos, o ponto de partida para o mapeamento de eQTLs é a medição da expressão global de genes numa célula ou tecido alvo, provenientes de vários indivíduos. Esta informação é o substrato para investigar os efeitos de polimorfismos sobre a expressão de genes (COOKSON *et al*., 2009). O uso da tecnologia de microarranjos e RNA-seq para medir a expressão de

milhares de genes simultaneamente, tem sido a principal força motriz para o mapeamento sistemático de eQTLs (BREM *et al*., 2002; SUN e HU, 2013).

Os eQTLs são separados na literatura entre *cis* ou *trans*, dependendo da distância física entre o gene que regulam. Convencionalmente, as variantes dentro de 1 Mb (megabase) do gene são chamados cis-eQTLs, enquanto aquelas a 5 Mb ou mais, ou em um cromossomo diferente, são considerados trans-eQTLs. Enquanto os cis-eQTLs são encontrados com mais frequência no genoma humano, trans-eQTLs são mais raros (NICA; DERMITZAKIS, 2013).

3.3 Genômica Funcional

Enquanto a genômica se concentra nos aspectos estáticos do genoma, por meio do sequenciamento do DNA, a genômica funcional se concentra no estudo dos mecanismos envolvidos na regulação da transcrição gênica e da tradução, incluindo aspectos relacionados com a função do genoma e sua variação, tais como análise das mutações, bem como as medidas da atividade molecular, com o objetivo de entender a relação entre o genoma de um organismo e seu fenótipo. O termo “genômica funcional” já foi descrito como "o estudo da função do genoma completo, incluindo genes e elementos não-gênicos, bem como ácidos nucleicos e proteínas codificadas pelo DNA" (PEVSNER 2009).

Historicamente, cada tipo de dado tem sido considerado de forma independente ao se olhar para as relações com os processos biológicos. De fato, a maior parte da etiologia genética de características complexas ainda permanece sem explicação, o que pode ser, ao menos em parte, devido ao foco em estudos com dados únicos e restritivos. Várias abordagens foram desenvolvidas para identificar as variantes que são capazes de desempenhar um papel biológico importante, mas a maioria concentra esforços na interpretação de SNPs em regiões transcritas (ADZHUBEI *et al.,* 2010), os quais podem alterar os códons do DNA, que por sua vez podem levar a alteração de um aminoácido, da estrutura e função da proteína. No entanto, a grande maioria dos SNPs identificados em associação a alguma doença estão localizados em regiões não transcritas, sendo provável que o mecanismo subjacente seja regulatório (MAHER, 2012). Por esta razão, abordagens genômicas

sistêmicas têm sido desenvolvidas e crescentemente utilizadas (RITCHIE *et al*., 2015), permitindo alcançar respostas para um interrogatório mais completo e informativo sobre as associações genótipo-fenótipo, do que uma análise que utiliza apenas um único tipo de dado. Ao se combinar várias camadas de dados, pode-se compensar a informação restritiva ou pouco confiável de dados isolados, pois múltiplas fontes de evidências que apontam para o mesmo sentido são menos propensas a levar a falsos positivos. O modelo biológico completo só é susceptível de ser descoberto se os diferentes níveis de regulação genética forem considerados em uma análise integrada e robusta. Incluir a referência aqui (não lembro de ter uma referencia para isso… procurei e não achei… acho q eu sou a referencia hahahaha)

A amplitude de dados a respeito das variações genéticas e suas relações tem aumentado exponencialmente. Atualmente, muitos projetos visando a identificação sistemática de elementos funcionais em grande escala estão em vigor, gerados a partir de uma combinação de ensaios bioquímicos seguidos de sequenciamentos completos do genoma, do transcriptoma, metiloma e relações metabólicas (COOPER; SHENDURE, 2011). Os estudos que geram esses dados são realizados pelos consórcios responsáveis pela "Enciclopédia de elementos regulatórios do DNA" (The ENCODE Project Consortium, 2013), pelo projeto "Anotação funcional do genoma mamífero" (FANTOM5) (FORREST *et al*., 2014), "Projeto 1000 Genomas" (1000 Genomes Project Consortium, 2010) e pelo "Mapa rodoviário do epigenoma" (Roadmap Epigenomics Consortium, 2015). Os estudos forneceram dados para as anotações funcionais do genoma, incluindo acetilação das histonas e metilação, estados de segmentação da cromatina, os locais de hipersensibilidade à DNase, locais de ligação de fatores de transcrição, sequências conservadas de intensificadores e silenciadores de transcrição. Esses recursos foram disponibilizados ao público através de navegadores da web, tais como o navegador UCSC (KENT et al., 2009) e Ensembl Genome (FLICEK *et al.,* 2014). Um objetivo essencial da análise destes dados é a identificação de modelos eficazes que predizem características e resultados fenotípicos, elucidando importantes biomarcadores e gerando novas hipóteses sobre as bases genéticas de características complexas (RITCHIE, M. D., *et al.*, 2015).

O sucesso na compreensão da arquitetura genética e genômica de fenótipos complexos tem sido modesto, entretanto, métodos para identificação de SNPs que se sobrepõem a elementos reguladores já têm sido utilizados com sucesso para identificar loci específicos que têm um papel funcional (JARINOVA *et al*., 2009; LANDERS *et al*., 2009). Em um estudo recente onde foram analisados 5.694 SNPs presentes no catálogo de GWAS, associados a 470 fenótipos, constatou-se que 44,8% de todos os SNPs se sobrepõem a algum elemento regulatório, 13,1% estão sobrepostos por mais de um tipo de elemento funcional, 4,7% se sobrepõem a regiões codificantes, 3,1% se sobrepõem em parte a um exon não codificante, 36,3% se sobrepõem a um pico de DNAase I em pelo menos uma linha celular, e 19,9% se sobrepõem a pelo menos uma das proteínas de ligação avaliadas em algum tipo celular (SCHAUB *et al*., 2012).

3.4 Projeto ENCODE

O Projeto ENCODE consiste em uma enciclopédia de pesquisa pública que visa identificar e integrar todos os elementos funcionais do genoma humano. O projeto envolve um consórcio mundial de grupos de pesquisa, cujos dados gerados podem ser acessados através de bancos de dados públicos, e destina-se como um recurso abrangente para permitir que a comunidade científica compreenda melhor como o genoma pode afetar a saúde humana (BERNSTEIN, 2012). Tomados em conjunto, estes dados mostram as regiões genômicas que são transcritas em RNA, quais as regiões mais prováveis de exercer controle sobre os genes em diferentes tipos celulares, e as regiões que estão associadas com uma grande variedade de proteínas reguladoras.

Os ensaios primários utilizados no ENCODE são ChIP-seq (imunoprecipitação da cromatina seguida por sequenciamento), ensaios de detecção de sítios com hipersensibilidade à DNase I, RNA-seq (sequenciamento de RNA) e ensaios de metilação do DNA. A técnica de ChiP-seq é utilizada para identificar a localização das interações entre proteínas e o DNA, como fatores de transcrição, histonas e seus fatores de modificação. Experimentos ChIP-seq já foram conduzidos para um total de 119 fatores de transcrição e outras proteínas de ligação ao DNA. A técnica consiste na geração de reações cruzadas entre o DNA e as proteínas associadas, seguida pela

quebra do DNA por sonicação e da imunoprecipitação dos complexos de DNA-proteína de forma seletiva, utilizando-se o anticorpo específico para a proteína reguladora de interesse. Em seguida, os fragmentos de DNA associados a estas proteínas são purificados e sequenciados para identificar o local de interação da proteína *in vivo.* Atualmente, um total de 147 tipos celulares já foram estudados, utilizando-se uma ampla variedade de ensaios experimentais (The ENCODE Project Consortium 2013).

De acordo com os últimos dados do ENCODE, aproximadamente 90% dos SNPs do genoma humano são encontradas fora das regiões codificadoras de proteínas (MAHER, 2012). Além disso, tem sido descrito que 80,4% do genoma possui ao menos um papel funcional em algum tipo de célula, seja através de um transcrito de RNA, fatores de transcrição ou modificação de cromatina. Também foi demonstrado que 99% do genoma encontra-se dentro de um raio de no máximo 1,7 kb de pelo menos um dos eventos bioquímicos identificados pelo ENCODE (BERNSTEIN *et al*., 2012). Estes resultados demonstraram que a regulação gênica é muito mais complexa do que se acreditava anteriormente (PENNISI, 2012).

3.5 Interações topológicas da cromatina

A organização espacial do genoma humano é conhecida por desempenhar um papel importante no controle da regulação gênica (BICKMORE, 2013). A descoberta de uma organização tridimensional formada por domínios topológicos como blocos de construção arquitetônicos nos cromossomos metazoários (LIEBERMAN *et al*, 2009), aumentou a compreensão sobre a estrutura e função do genoma. Essa nova camada de organização do genoma de grande escala fornece insights sobre a maneira pela qual elementos regulatórios poderiam interagir entre longas distâncias, levando intensificadores distantes a uma estreita proximidade do gene, influenciando na sua regulação.

A primeira técnica de detecção de interações cromossômicas (3C) e suas adaptações posteriores (4C, 5C e 6C) eram limitadas à escolha de um locus alvo específico de cada vez para estudo, tornando os estudos genômicos inviáveis. A técnica Hi-C (Figura 6) permitiu a avaliação de interações cromossômicas em larga escala, compreendendo todo o genoma (LIEBERMAN *et al*., 2009).

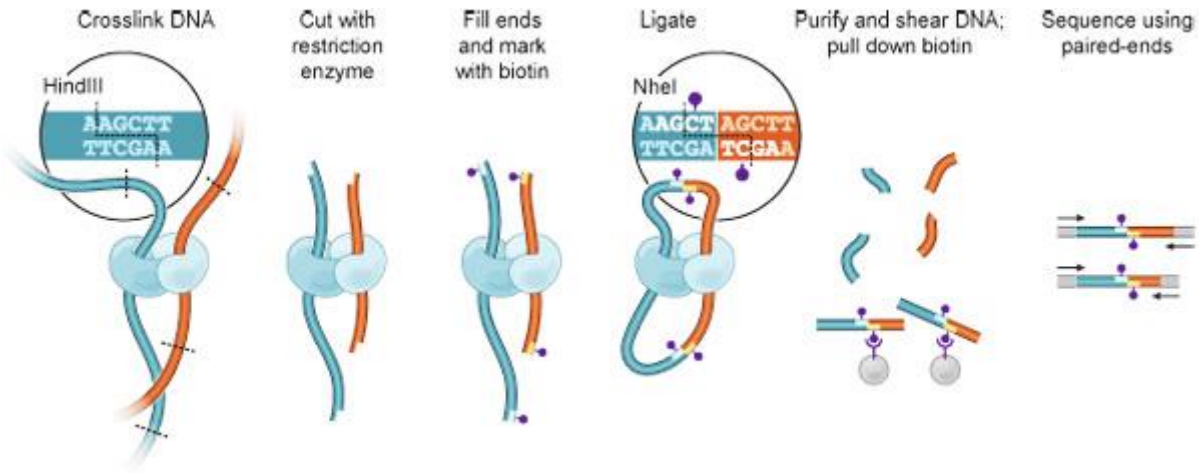

FIGURA 6 – Técnica Hi-C.

NOTA: As células são tratadas com formaldeído, resultando em ligações covalentes entre segmentos próximos da cromatina (fragmentos de DNA: azul e vermelho; proteínas, que podem mediar tais interações azul claro). A cromatina é digerida com uma enzima de restrição (HindIII). As extremidades coesivas resultantes são preenchidas com nucleotídeos, um dos quais é biotinilado (ponto roxo). Desta forma, os locais de restrição de HindIII são perdidos e repostos por sítios reconhecidos por Nhel. O DNA purificado é cortado, e junções biotiniladas são isoladas utilizando esferas de estreptavidina. Os fragmentos que interagem são identificados por sequenciamentos e emparelhados para a construção da matriz.

FONTE: VAN BERKUM et al., 2010.

Atualmente, o mecanismo organizador dos domínios topológicos melhor caracterizado envolve as interações de longo alcance mediado por proteínas isoladoras CTCF (“CCCTC-binding factor”, uma proteína de dedos de zinco) e do complexo de coesinas, necessário para a coesão da cromatina (PHILLIPS-CREMINS *et al.*, 2013). Juntos, CTCF e coesinas exercem seus efeitos sobre a formação ou estabilização das alças de cromatina de longo alcance (RUDAN *et al*., 2015). Locais com fortes intensidades de ligação ao CTCF são altamente conservados entre diferentes espécies, enquanto que sítios de ligação de baixa intensidade apresentam mais variações nos motivos de ligação (RUDAN *et al*., 2015). Polimorfismos associados a alguma doença por GWAS ocupam preferencialmente regiões de interações da cromatina, revelando a

importância das interações topológicas na manutenção do fenótipo (ROLAND et al., 2014).

4. MATERIAIS E MÉTODOS

4.1 Casuística

Para a análise de desequilíbrio de ligação entre os SNPs selecionados, foram utilizados os dados de genotipagem disponíveis no banco de dados do projeto HapMap para a população CEU (euro-descendente do estado de Utah, com amostra composta por 60 indivíduos) e no banco de dados do projeto 1000 Genomes para a população europeia (EUR), composta pela reunião das seguintes subpopulações: GBR (britânica, com amostra composta por 94 indivíduos), IBS (ibérica, com amostra composta por 107 indivíduos), FIN (finlandesa, com amostra composta por 100 indivíduos), TSI (toscanos da Itália, com amostra composta por 110 indivíduos) e CEU (com amostra composta por 103 indivíduos).

4.2 Recursos e Dados

4.2.1 Polimorfismos

Para este estudo, foram selecionados SNPs que se encontram em regiões não codificantes, intrônicas e intergênicas, de genes da via das lectinas do complemento (*FCN1, FCN2, FCN3* e *MBL*), os quais estão associados com a variação da expressão destes genes (eQTLs), ou em DL com estes. Para estes SNPs, foi realizada uma análise integrativa multi-etapa de genética de populações e genômica funcional, para anotá-los de acordo com o desequilíbrio de ligação e com os elementos investigados no projeto ENCODE e verificar se são, de fato, os mais prováveis a influenciar a regulação gênica nessas regiões, além de gerar hipóteses sobre o mecanismo subjacente a essas associações. Os eQTLs foram coletados a partir do banco de dados Blood eQTL browser (WESTRA., FRANKE., et al., 2013) (http://genenetwork.nl/bloodeqtlbrowser/) e GTEX portal (GTEx Consortium, 2013) (http://www.gtexportal.org/home/).

4.2.2 Anotações Funcionais

Os dados de elementos reguladores foram coletados a partir do HaploReg v4 (atualizado em 05/11/2015) (WARD; KELLIS, 2012), uma ferramenta para explorar anotações funcionais provenientes do projeto ENCODE. Foram utilizados dados de modificação de histonas, sítios de hipersensibilidade a DNAse (regiões de cromatina aberta), estado de segmentação da cromatina (promotores, intensificadores, insuladores e regiões de transcrição) e ligação de fatores de transcrição. Para visualizar esses dados de forma integrada, foi utilizada a ferramenta da web UCSC Genome Browser (http://genome.ucsc.edu) (KENT *et al.,* 2009).

4.2.3 Testes de desequilíbrio de ligação

A hipótese de desequilíbrio de ligação (DL) entre os eQTLs selecionados foi testada com o programa HAPLOVIEW v.4.2 (http://broadinstitute.org/haploview/haploview) (BARRETT *et al*., 2005), a partir dos dados do projeto HapMap para a população eurodescendente do estado de Utah (USA). A mesma hipótese foi avaliada entre as variantes dos eQTLs selecionados e os SNPs genotipados pelo projeto 1000 Genomes (1000 Genomes Project Consortium, 2010) para a população europeia como um todo, utilizando a ferramenta online HaploReg v4 (WARD; KELLIS, 2012).

4.2.4 Análises estatísticas e construção dos gráficos

Para as análises estatísticas e construção de gráficos, foi utilizado o programa R versão 3.2.2 (https://www.r-project.org), usando os testes de Wilkoxon-Mann-Whitney/ Kruskal-Wallis, uma vez que a distribuição dos valores Z não é paramétrica.

Para gerar os gráficos da topologia da cromatina, foi utilizado o programa Juicebox (http://www.aidenlab.org/juicebox/) (SUHAS *et al*., 2014), o qual gera uma matriz de visualização dos dados de Hi-C, uma técnica de captura tridimensional da cromatina, para as linhagem celulares linfoblastoide (GM12878, in situ Mbol, primary+replicate) e fibroblastos de pulmão (IMR90, in situ+combined).

4.3 Construção dos escores Z e seleção dos SNPs

O escore Z é um valor relativo à proporção de elementos funcionais sobrepostos a um certo SNP. Portanto, permite inferir o grau com que a variante pesquisada afeta a expressão de um gene. A construção destes escores foi baseada no conceito simples de construção de um *ranking*, onde cada variante é pontuada de acordo com a sua frequência de sobreposição a diferentes elementos funcionais anotados no ENCODE, como regiões de modificação de histonas, sítios de hipersensibilidade a DNAse (regiões de cromatina aberta), estado de segmentação da cromatina (promotores, intensificadores e insuladores), ligação de proteínas e motivos de ligação a fatores de transcrição. De forma simples, cada anotação funcional por linhagem celular somou 1 peso ao escore. Foram analisadas todas as 147 linhagens celulares disponíveis no ENCODE. Para detectar enriquecimento funcional dos SNPs e gerar escores para tecidos específicos, os dados foram filtrados no Excel. Posteriormente, os escores foram normalizados em escores Z, utilizando o programa de estatística R (R DEVELOPMENT CORE TEAM, 2008).

Um grande desafio na interpretação dos resultados de estudos de associação vem do assim chamado “efeito carona”, ou seja: do fato de que as associações encontradas possam ser devidas a outras variantes, em desequilíbrio de ligação com as pesquisadas. Em outras palavras, as variantes associadas ao fenótipo podem não apresentar efeito funcional sobre o mesmo, mas por sua vez, variantes em desequilíbrio de ligação podem revelar-se funcionalmente significativas. No caso das eQTLs, pode-se comparar os escores Z de SNPs que ocorrem em blocos de DL, refinando os resultados e restringindo o número de possíveis associações verdadeiras. Os SNPs com os maiores escores Z em cada bloco de desequilíbrio de ligação, além dos SNPs sobrepostos a fatores de transcrição em linhagem celular específica e os SNPs sobrepostos a algum RNA não codificante (microRNAs e RNAs longos não codificantes) foram selecionados como os mais prováveis de desempenhar papel funcional.

4.4 Fluxograma do trabalho

A seleção de SNPs funcionais foi atingida por meio de etapas sequenciais com o propósito de triagem, integração e visualização dos dados de forma informativa (figura

7). É importante ressaltar que ainda não existe um método padronizado para predizer com precisão as associações verdadeiras sem a necessidade de experimentos *in vitro*, pois os mecanismos de regulação para a expressão destes genes ainda são desconhecidos. Esta análise *in silico* tem o viés de priorizar variantes detectadas nas regiões mais enriquecidas com elementos funcionais, baseando-se no pressuposto de que estas são as mais prováveis de desempenhar um papel sobre o fenótipo.

Iniciamos com a seleção dos SNPs relacionados como eQTLs dos genes em estudo, a partir dos bancos de dados Blood eQTLs browser (WESTRA *et al*., 2013) e GTEX (GTEx Consortium, 2013). Em seguida, avaliou-se a possibilidade de desequilíbrio de ligação entre os mesmos, possibilitando detectar variantes causais ligadas em haplótipos. O próximo passo consistiu na avaliação das anotações de genômica funcional para cada SNP, atribuindo o valor 1,0 a cada elemento sobreposto ao SNP, para a construção de um escore. Neste trabalho, o escore não foi ponderado, sendo somado apenas 1 para cada anotação, por tipo celular.

As anotações foram triadas por tecidos específicos, para restringir as análises ao tecido onde foram detectadas as associações com expressão gênica, reportadas na literatura. No caso de um grande número de variantes associadas localizadas proximamente, os SNPs foram testados para avaliar se estão mais enriquecidos com elementos funcionais, quando comparados aos SNPs em entorno e no mesmo perímetro do bloco de DL.

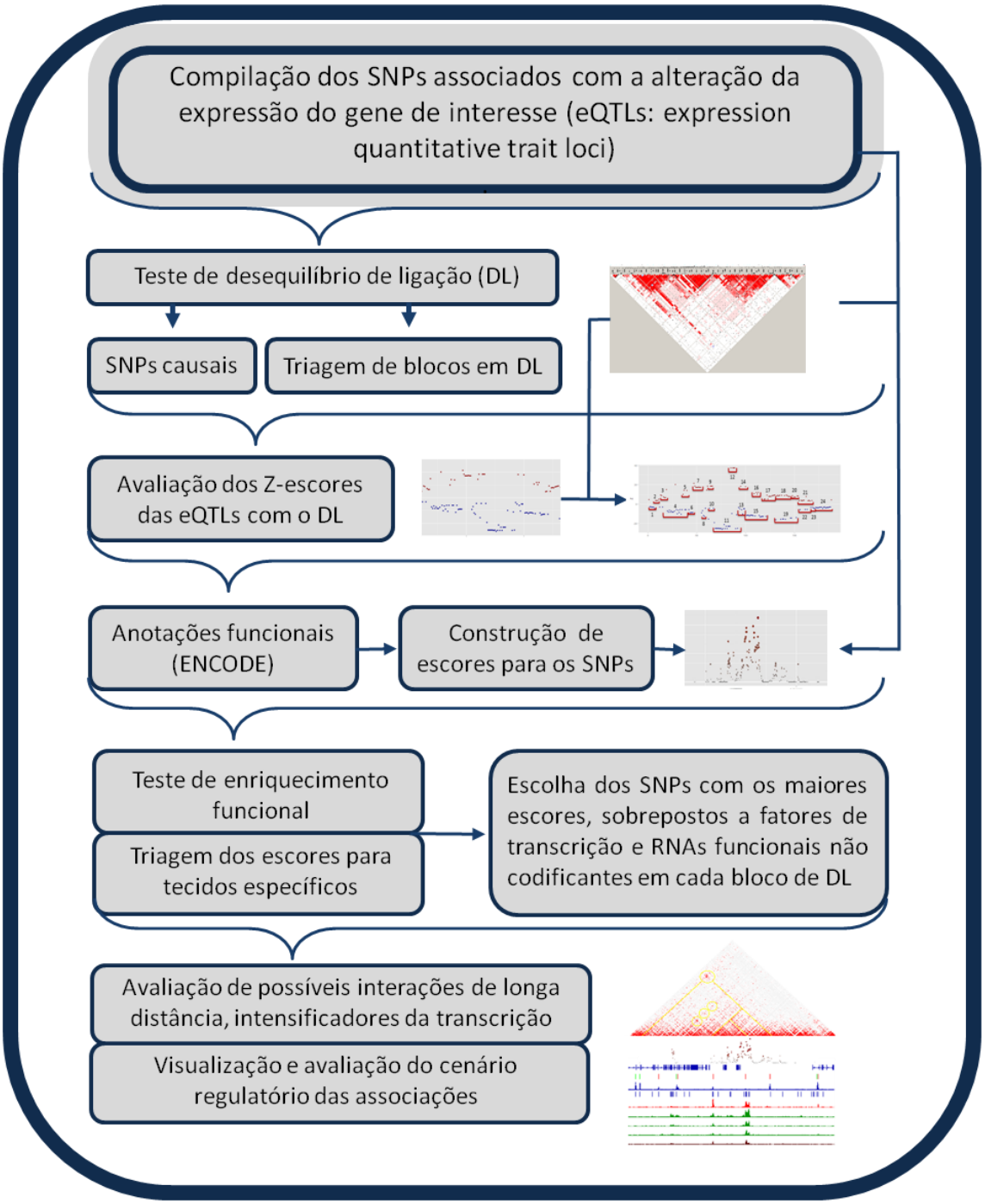

FIGURA 7 - Fluxograma para a seleção de polimorfismos capazes de alterar a regulação gênica.

Por fim, os SNPs selecionados como principais candidatos funcionais foram analisados individualmente e de forma integrativa, a partir do cenário regulador da região genômica em foco. Quando as associações estão localizadas a longas distâncias da região promotora, podem estar envolvidas na formação de alças na cromatina. Estas interações foram visualizadas através da comparação com matrizes da topologia

do DNA, geradas por técnicas de captura da conformação tridimensional da cromatina, e pela presença de CTCF (do inglês *CCCTC-binding factor*) nestas regiões, a principal proteína de ligação caracterizada por mediar a formação de domínios topológicos. Os SNPs detectados em regiões de interação da cromatina são mais prováveis a interferir na atividade de intensificadores da transcrição.

5. RESULTADOS E DISCUSSÃO

5.1 Resultados para *FCN1*

Foram analisados 188 cis eQTLs associados à expressão do gene *FCN1* em células sanguíneas (APÊNDICE A) (WESTRA *et al*., 2013), localizados em um perímetro de 382.674 pares de base (pb) ao entorno de *FCN1*, a partir do exon 2 do gene *COL5A1* (collagen, type V, alpha 1) até o exon 1 do gene *OLFM1* (olfactomedin 1) (9q34.3). Do total, 112 eQTLs (60%) regulam positivamente a expressão de *FCN1* (escore Z positivo) e 76 (40%), negativamente (escore Z negativo). Também foram encontrados 83 eQTLs para *FCN1,* a partir do banco de dados GTEX (GTEx Consortium, 2013), todos incluídos entre os 188 eQTLs do banco de dados Blood eQTL browser (WESTRA *et al*., 2013).

Os haplótipos com os eQTls detectados por Westra e colaboradores (2013) foram construídos com referência aos dados do projeto HapMap (THE INTERNATIONAL HAPMAP CONSORTIUM, 2010) para uma população eurodescendente, utilizando-se o programa Haploview (BARRETT *et al*., 2005). Contudo, os dados deste projeto estão desatualizados quanto ao número de variantes, não representando o total dos SNPs em DL, para cada bloco. Para ajustar os blocos em DL aos dados atuais, os eQTLs foram testados para DL com outros SNPs avaliados no projeto 1000 Genomes (1000 GENOMES PROJECT CONSORTIUM, 2010) para a população europeia, obtendo-se mais 218 SNPs em DL com os 188 eQTLs ($r^2 \geq 0,8$), totalizando 406 SNPs. Este passo é necessário para não deixar de analisar possíveis SNPs candidatos a desempenhar um papel no fenótipo.

Embora os 188 eQTLs estejam estatisticamente associados a variações quantitativas do mRNA de *FCN1*, ainda não se tem evidências diretas de que interfiram, de fato, na expressão gênica. Tais SNPs podem não apresentar efeito

funcional, mas por sua vez, as 218 variantes em desequilíbrio de ligação com estes, podem revelar-se funcionalmente importantes. Isso torna difícil identificar, com precisão, os SNPs que têm uma ligação biológica com o fenótipo. Para contornar este problema, os eQTLs e SNPs ligados a estes foram agrupados pela sua localização genômica e grau de DL nas populações europeias investigadas pelo projeto HapMap e 1000Genomes. Em seguida, foi possível comparar os valores Z, por bloco de DL.

A região investigada apresenta 16 blocos de DL na população europeia (figura 8.b). Cada um destes blocos foi associado a um ou mais de 24 agrupamentos de eQTLs com escores Z, positivos ou negativos (figura 8.a). Presumindo-se que cada agrupamento contenha ao menos um SNP funcional, e que os demais SNPs estejam associados à expressão, unicamente devido ao DL com este, assume-se que, dentre os 188 eQTLs distribuídos nos 24 agrupamentos, existam no mínimo 24 SNPs funcionais influenciando o fenótipo. Alguns eQTLs que não foram encontrados em desequilíbrio de ligação com outros SNPs, foram avaliados separadamente.

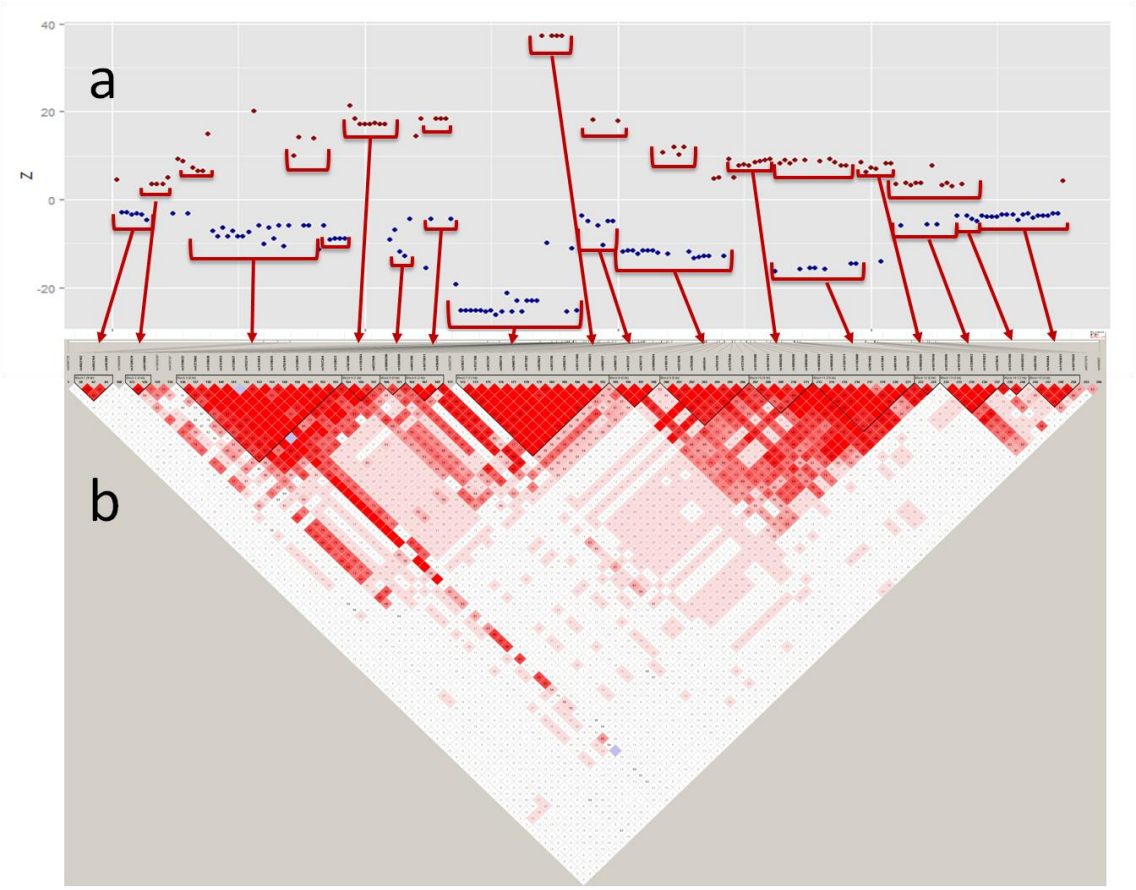

FIGURA 8 - Separação dos eQTLs em agrupamentos, de acordo com o escore Z e grau de desequilíbrio de ligação.

NOTA: Os valores de Z apresentaram o mesmo sinal (valor Z semelhante) para os SNPs, em cada bloco de DL. A partir deste resultado, foram separados 24 agrupamentos de eQTLs. Oito agrupamentos com valores positivos de Z estão sobrepostos a agrupamentos com valores negativos de Z, no LD plot em “b”.

LEGENDA: a) pontos vermelhos: Z escores positivos (associados com a ativação do gene); pontos azuis: Z escores negativos (associados com a repressão da expressão gênica); b) tons de rosa = LOD≥2 e D’<1; vermelho = LOD≥2 e D’=1 desequilibrio de ligação absoluto; branco = D’=0 sem desequilíbrio de ligação (LOD =Log of the likelihood odds ratio score).

FONTE: a) o autor; b) Programa Haploview versão 4.2, com a população euro-descendente do estado americano de Utah (BARRETT *et al*., 2005).

5.1.2 Anotações funcionais

Os SNPs foram avaliados para a sobreposição a elementos reguladores detectados no projeto ENCODE. A partir disso, foram construídos os escores. Cerca de 396 dos 406 SNPs apresentaram-se sobrepostos a algum traço funcional detectado em pelo menos uma linhagem celular, e 11 localizados sobrepostos a um sítio reconhecido por uma proteína de ligação ao DNA.

Para averiguar se os 406 SNPs estão mais enriquecidos para fatores reguladores, em comparação com os SNPs em entorno, comparou-se as médias dos seus escores com as dos demais SNPs, localizados no mesmo perímetro da região cromossômica investigada (6788 SNPs em 382.674 pb). Isto foi feito para dois diferentes tipos de escores: (1) escores obtidos com base na frequência de sobreposição dos SNPs aos elementos funcionais anotados em todas as linhagens celulares analisadas no projeto ENCODE, e (2) escores obtidos com base apenas nas anotações em linhagens celulares de células sanguíneas (APÊNDICE A). A média dos escores do tipo (1) foi maior para o grupo de 6788 SNPs, exceto para proteínas reguladoras. Já o grupo incluindo apenas os 406 SNPs apresentaram médias muito superiores para linhagens sanguíneas excetuando-se as proteínas reguladoras e motivos de ligação (para os quais não há especificidade reconhecida) (tabela 1). Os eQTLs e SNPs em DL para *FCN1* estão mais concentrados em regiões enriquecidas para elementos funcionais em células sanguíneas (figura 9.a). Embora estas regiões também sejam ocupadas pelos demais SNPs, estes últimos também estão concentrados em outras regiões de regulação gênica, que parecem específicas de outras linhagens celulares (figura 9.b). Este resultado era esperado, uma vez que a maior parte dos eQTLs avaliados para *FCN1*, regulam sua expressão em células sanguíneas, sendo este gene sabidamente expresso em monócitos e neutrófilos (GARRED *et al.,* 2009).

TABELA 1 – Média dos escores em cada variável de anotações funcionais do projeto ENCODE.

Valores médios de Z		**Células sanguíneas**				
	eQTLs #	*Demais SNPs*	P	*eQTLs #*	*Demais SNPs*	P
	N= 406	N= 6788		N= 406	N= 6788	
Estados da Cromatina	5,12	**6,15**	0,0046	**1,62**	0,98	2,94E-05
Marcações da Cromatina	26,06	**34,31**	0,002	**4,41**	2,55	1,59E-06
DNAse	1,24	1,49	0,0851	**0,27**	0,18	0,00285
Proteínas reguladoras	0,32	0,53	0,342	0,12	0,11	0,9486
Motivos de ligação	3,59	3,93	0,0895	*	*	*
Total ± Desvio-padrão	36,33 ± 3,16	**46,41 ± 4,22**	0,0224	**9,26 ± 1,38**	5,58 ± 1,16	8,98E-05

NOTA: utilizou-se o Teste de Mann-Whitney-Wilcoxon para avaliar a diferença entre os escores obtidos usando os dados derivados de todas as linhagens celulares estudadas pelo ENCODE (“Total de linhagens celulares”) e os restritos a células sanguíneas, para os 406 SNPs avaliados para *FCN1* e para os 6788 demais SNPs, localizados na mesma região.

LEGENDA: # inclui SNPs em DL com os EQTLs *os motivos de ligação não foram incluídos no grupo de anotações funcionais restritas a células sanguíneas, devido a não apresentarem distinção entre tipos celulares pelo método PWM. Em vermelho: resultados altamente significativos. Em negrito: valores significativamente superiores de médias dos escores. N= número de SNPs.

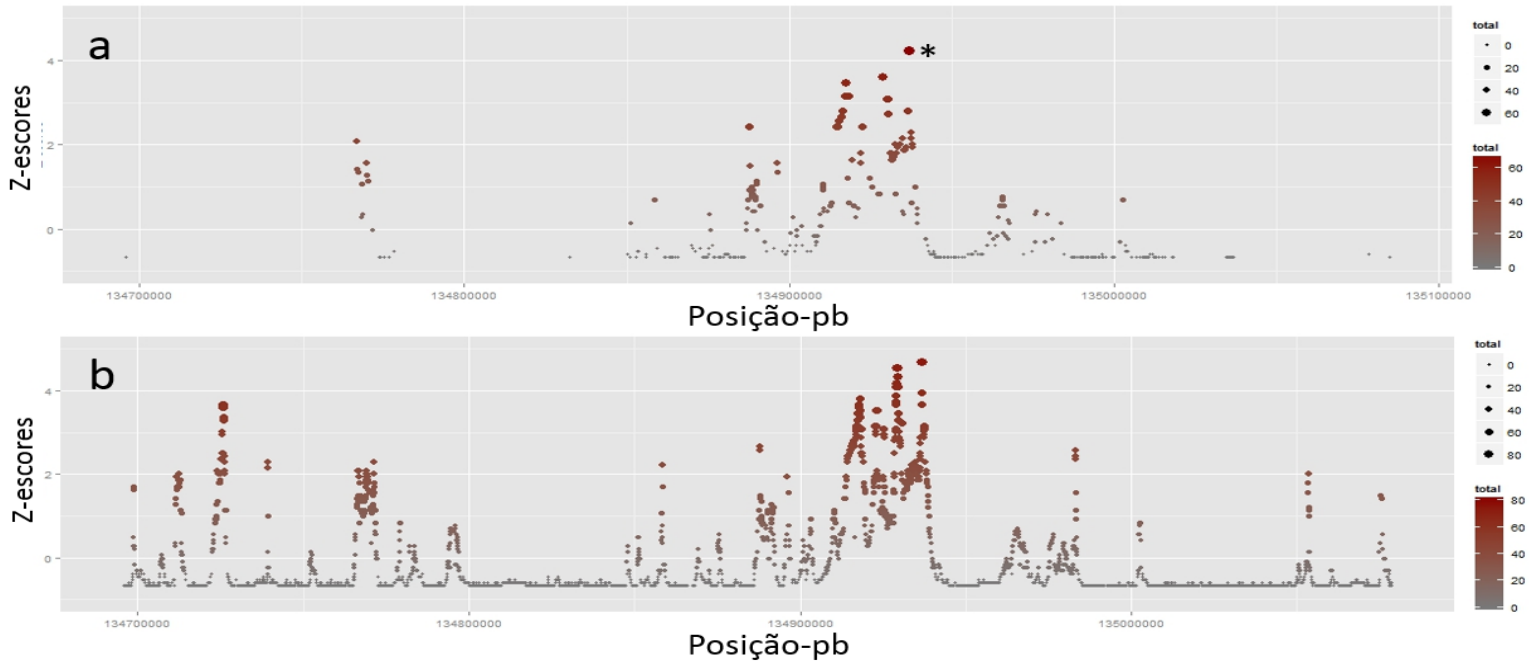

FIGURA 9 - Distribuição dos escores obtidos pela densidade de elementos reguladores em células sanguíneas, de acordo com a localização do SNP.

NOTA: Os escores foram normalizados através da estatística Z. a - Distribuição de escores Z dos 406 SNPs em DL, associados à expressão de *FCN1*. b - Distribuição de escores Z dos outros 6788 SNPs, encontrados na mesma região.

LEGENDA: A gradação de cores e tamanho dos pontos é referente ao valor do escore Z. Os maiores escores estão representados em pontos maiores de tonalidade mais vermelha. * rs1105176, com o maior escore, de 68, sobreposto a 10 proteínas de ligação em células sanguíneas.

FONTE: O autor, utilizando o pacote ggplot2 (WICKHAM 2009) através do software R versão 3.2.2 (https://www.r-project.org).

5.1.3 Seleção dos principais candidatos funcionais

Para cada um dos 24 agrupamentos de eQTLs e SNPs em desequilíbrio de ligação, foram selecionados os SNPs com os maiores escores funcionais, além dos que ocorrem sobrepostos a uma região reconhecida por uma proteína reguladora ou por um RNA não codificante (RNA longos intergênicos não codificadores - lincRNA e microRNAs ou miRNAs), totalizando 37 SNPs como principais candidatos a desempenhar um papel funcional (tabela 2).

5.1.4 Contexto regulatório

A grande maioria dos eQTLs estão localizados a longas distâncias de *FCN1* (até 213803 pb a 3' e 160496 pb a 5'), sendo provável que sua influência sobre o gene seja resultante de longas interações topológicas da cromatina, pela formação de alças mediando o contato de intensificadores com a região promotora ou com sítios reguladores adjacentes. Para caracterizar o contexto e influência das associações sobre a formação de intensificadores, foram avaliados dados de HI-c (técnica de captura da conformação tridimensional da cromatina) e anotações funcionais de fatores de transcrição, CTCF, modificações de histonas e estados de segmentação da cromatina.

Um quadro geral da topologia do DNA flanqueando *FCN1* é mostrado na figura 10. Foram detectados dois potenciais domínios topológicos de longa distância (figura 10.a; 1, 2 e 3) em regiões com grande densidade de fatores reguladores e de SNPs associados à expressão de *FCN1*, na linhagem celular linfoblastoide (GM12878). Os sítios de ligação ao CTCF flanqueiam exatamente os domínios topológicos, e a densidade de atividade reguladora mostrou-se diretamente relacionada à formação das interações da cromatina. A maioria das eQTLs com altos escores estão localizadas nestas regiões, realçando a hipótese de que estejam influenciando intensificadores da transcrição.

TABELA 2 - Principais polimorfismos candidatos para a regulação de *FCN1*.

Bloco de DL #	SNPs	Nomenclatura genômica	MAF %	∑ Escore	Proteínas de ligação ao DNA	RNAs não codificantes
1	rs4842158	*9:g.137658722G>A*	14	29	5	
	rs4842157	*9:g.137658608A>G*	31	38	5	
2	rs62571442	*9:g.137742124G>A,T,C*	42	1	0	miRNA
	rs72502717	*9:g.137742597A>G*	12	0	0	miRNA
3	rs56044219	*9:g.137750200_137750202delAAC*	05	19	11	
4	rs3128632	*9:g.137767197G>A*	42	14	0	
	rs17549193	*9:g.137779026C>T*	28	16	0	
5	rs4521835	*9:g.137779306T>A,C,G*	41	22	0	
6	rs73664188	*9:g.137779607T>C*	11	43	0	
7	rs11103564	*9:g.137779875T>C*	29	30	0	
8	rs10858290	*9:g.137784268A>G*	30	5	0	
	rs4842188	*9:g.137788134C>T*	41	28	0	
9	rs4842189	*9:g.137788288T>C*	30	31	0	
10	rs4842192	*9:g.137795604C>T*	40	10	0	
11	**rs28909976**	*9:g.137809989delA*	37	53	0	
12	rs12377780	*9:g.137808961C>A,T*	35	53	0	
	rs7857015	*9:g.137811241A>G*	36	32	1	
13	rs2026335	*9:g.137822287T>C*	31	52	6	
	rs75735785	*9:g.137822288G>A*	05	52	6	
14	rs10858294	*9:g.137824613T>G*	33	21	1	
15	rs2274844	*9:g.137834828C>T*	42	2	0	lincRNA
	rs1105176	*9:g.137828650T>A,G*	48	68	10	
	rs10776912	*9:g.137829353C>T*	48	18	1	
16	rs6537966	*9:g.137826183C>A*	25	37	0	
17	rs34242617	*9:g.137848975C>A*	28	2	1	
	rs7037945	*9:g.137857630G>A*	30	19	0	
18	rs11103604	*9:g.137857446G>A,T*	28	20	0	
19	rs4841951	*9:g.137857282A>G*	35	17	0	
	rs7046707	*9:g.137857344C>T*	35	19	0	
20	rs7046492	*9:g.137867534T>A,C*	36	14	0	
	rs7851241	*9:g.137871481C>T*	28	14	0	
21	rs11103612	*9:g.137875536G>C*	45	11	0	
22	rs914392	*9:g.137894409T>C*	06	19	0	
23	rs7019767	*9:g.137895838T>G*	34	2	0	
24	rs10858313	*9:g.137902809C>T*	09	1	1	
*	rs6537952	*9:g.137764457G>A*	32	3	0	
*	rs2989726	*9:g.137809374A>G*	28	57	0	

NOTA: os SNPs foram selecionados a partir dos maiores escores, sobreposição a sítios reconhecidos por proteínas de ligação e RNAs não codificantes.

LEGENDA: *SNPs não encontrados em nenhum bloco de DL. Em negrito: SNP previamente investigado por BOLDT et al. (2012). Com exceção deste, nenhum dos SNPs sequenciados por HUMMELSHOJ et al. (2008) consta na lista. MAF: frequência do alelo menos comum na população europeia, segundo dados do 1000 Genomes. lincRNA: RNA longo intergênico não codificador. # blocos de DL enumerados abaixo:

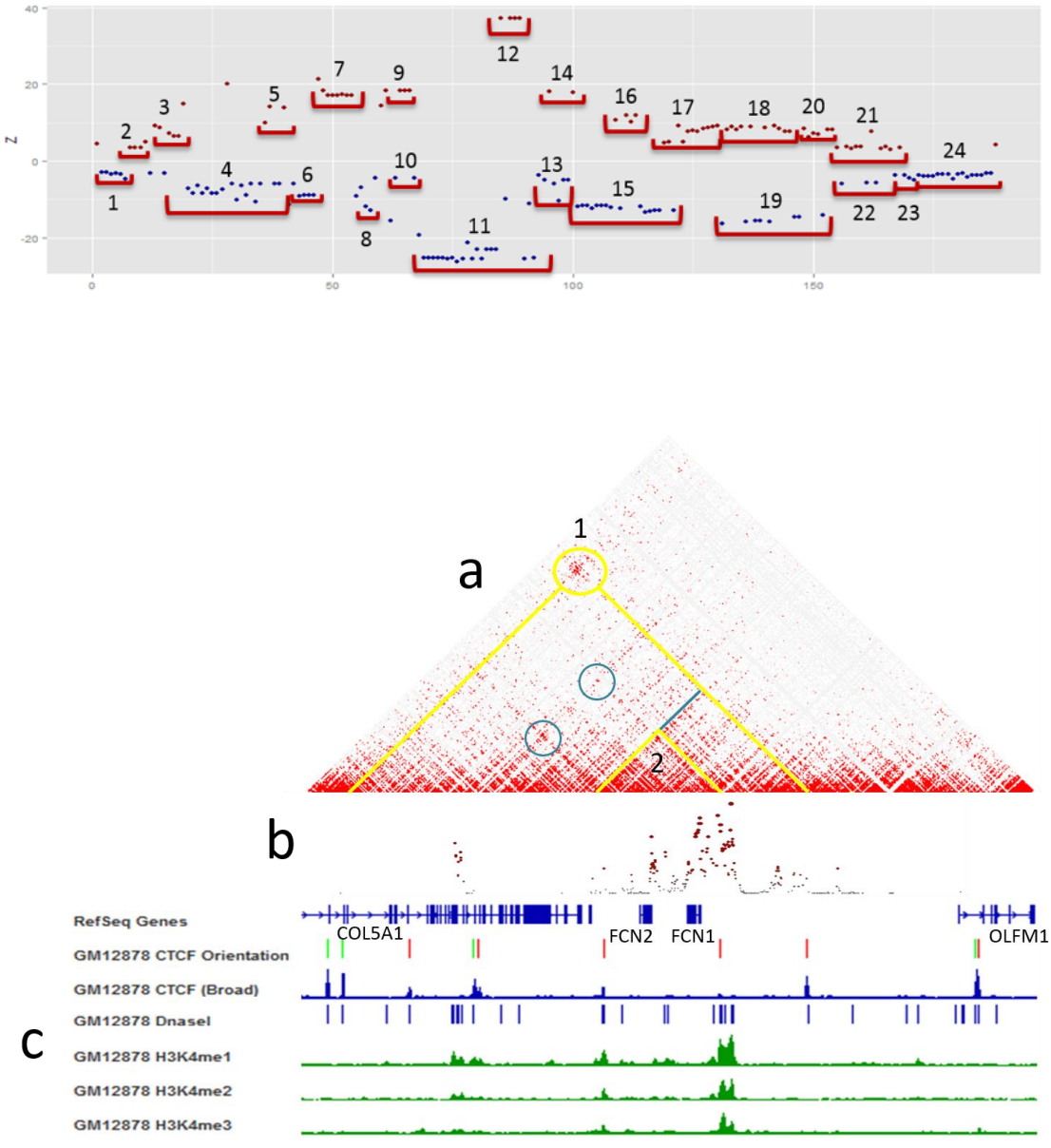

FIGURA 10 - Contexto regulatório dos SNPs associados à modulação da expressão gênica de *FCN1* em linhagem celular linfoblastoide (GM12878).

NOTA: Os números 1 e 2 (a) representam os principais domínios topológicos da cromatina nas regiões adjacentes a *FCN1* e *FCN2*, enquanto a linha e os círculos em azul evidenciam possíveis regiões de contato não calculados pelo programa Juicebox. Os 406 SNPs, caracterizados em DL com as 188 eQTLs, estão dispostos em um gráfico (b) de acordo com o seu escore, gerado a partir do número de anotações funcionais correspondentes, em células sanguíneas. É importante notar que as regiões de interação da cromatina (a) estão caracterizadas por (c) sítios de ligação ao CTCF (modificador conformacional da cromatina) e a fortes picos de modificação de DNAse (c, em azul), associados a regiões de cromatina aberta em GM12878. As marcações de histonas H3K4me1, H3K4me2 e H3K4me (c, em verde) também ocorrem nas regiões de interação. Estas estão associados à metilação da lisina 4 da histona H3 e são encontradas em promotores ativos e intensificadores distais (BANNISTER; KOUZARIDES, 2011).

FONTE: O autor, através do programa Juicebox (http://www.aidenlab.org/juicebox/) (SUHAS et al., 2014), para a linhagem celular linfoblastoide (GM12878, in situ Mbol, primary+replicate).

Dentre os 37 SNPs selecionados (tabela 2), 11 estão sobrepostos a sítios de reconhecimento de proteínas de ligação ao DNA em tecidos sanguíneos, dois estão sobrepostos a micro RNAs (miRNA), e um sobre um RNA intergênico longo não codificante (lincRNA), os demais estão sobrepostos a sítios de hipersensibilidade a DNAse e modificação de histonas. Cada SNP selecionado foi avaliado individualmente, dos quais três foram selecionados como exemplo.

O SNP rs56044219, com o escore 19, é uma variante indel de 3 nucleotídeos (*9:g.137750200_137750202delAAC*), cuja frequência do alelo menos comum (MAF) varia de 4% (média das populações do sul da Ásia) a 8% (populações africanas) (dados acessados em 20.11.2015, disponíveis em: http://grch37.ensembl.org/Homo_sapiens/Variation/Explore?db=core;r=9:137749700-137750702;v=rs56044219;vdb=variation;vf=113746125). Ele está localizado numa região de ligação a 11 proteínas de ligação detectados em sangue, sítio de hipersensibilidade a DNAse e uma região intensificadora (figura 11). Esta região delimita o domínio topológico que engloba *FCN1* e *FCN2* (figura 10.a.2).

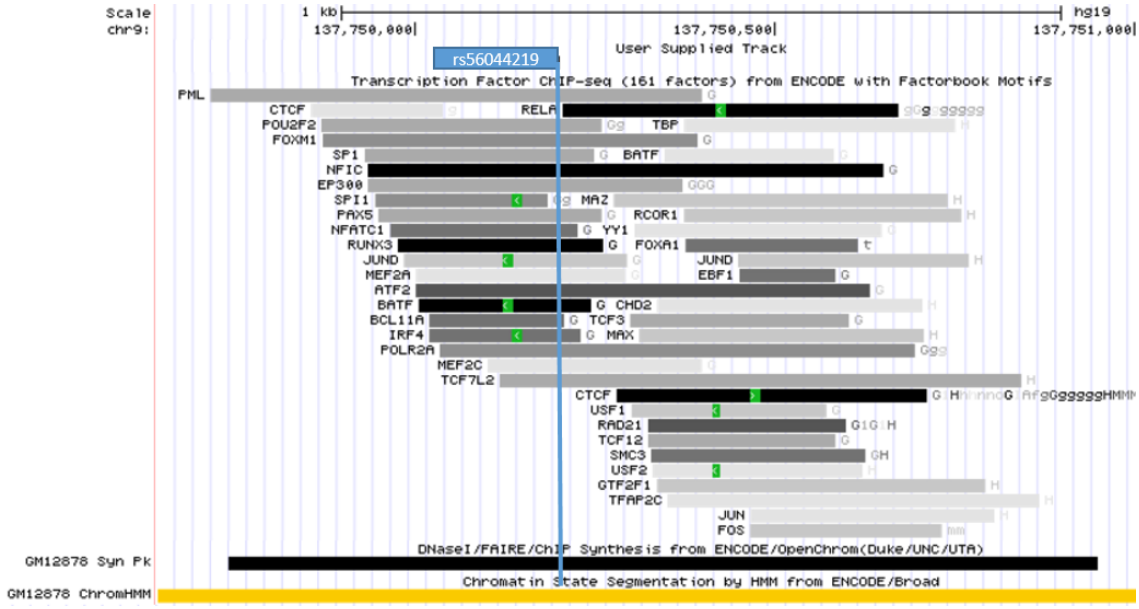

FIGURA 11 - Contexto regulatório do SNP rs56044219.

LEGENDA: A gradação de cor é proporcional à intensidade do sinal observado por ChIp-Seq. As regiões verdes representam o motivo de ligação de maior afinidade para a proteína correspondente. A região em preto (GM12878 Syn Pk) demarca um sitio de hipersensibilidade a DNAse, característico de uma região de cromatina aberta. As proteínas de ligação sobrepostas ao SNP, com o maior sinal (representadas em preto) são: NFIC (*Nuclear factor 1 C-type*), RUNX3 (*Runt-related transcription factor 3*) e BATF (*Basic leucine zipper transcription factor, ATF-like*).

FONTE: UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr9%3A137749950-137750452&hgsid=454209933). Acesso em: 10.11.2015

Outro polimorfismo associado ao domínio topológico que engloba *FCN1* e *FCN2* é o SNP rs1105176, com o escore 68. Este SNP é trialélico: *9:g.134936804T>A* ou *9:g.134936804T>G*, embora o alelo *A* seja muito raro, e o *G* ocorra com alta frequência em todas as populações humanas, de 32% em populações ameríndias a 55% no Leste Asiático (dados acessados em 20.11.2015, disponíveis em: http://www.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=9:134936304-134937304;v=rs1105176;vdb=variation;vf=101927328). A frequência similarmente elevada e altas taxas de heterozigosidade em todas as populações investigadas podem ser devidas à seleção balanceadora global. Este SNP está sobreposto a 10 proteínas de

ligação, sítios de hipersensibilidade a DNAse e um intensificador (figura 11). Uma DNA polimerase marca o início de transcrição para um lincRNA (*TCONS_00015773*) nesta região, um possível efetor da regulação gênica.

Por sua vez, o SNP rs2274844 (*9:g.137834828C/T)*, ocorre em altas frequências em todas as populações, de 41% na população europeia a 51% na população africana (dados acessados em 20.11.2015, disponíveis em: http://grch37.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=9:137834328-137835328;v=rs2274844;vdb=variation;vf=104442572). Este SNP está sobreposto ao exon 3 do lincRNA *TCONS_00015773* (figura 12), e pode estar influenciando a sua função. Os lincRNAs podem desempenhar diferentes papéis na regulação de genes e outros processos celulares, embora a função da maioria seja ainda, desconhecida (ULITSKY; BARTEL, 2013). *TCONS_00015773* possui uma estrutura secundária bem complexa, entretanto, ainda não é possível determinar se possui alguma função biológica (disponível em: http://lncipedia.org/db/transcript/lnc-FCN2-1:1).

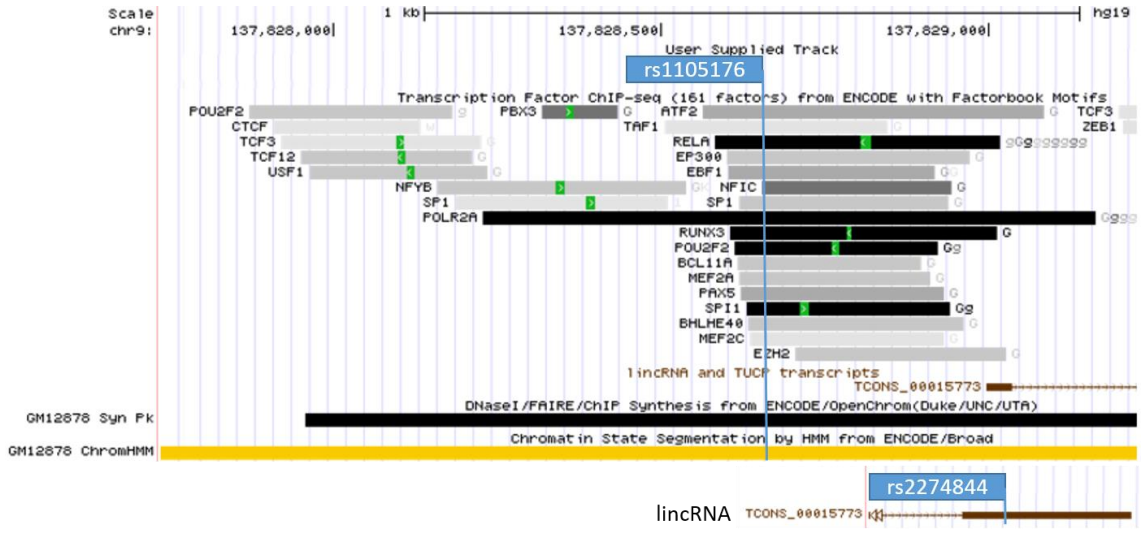

FIGURA 12 - Contexto regulatório dos SNPs rs1105176 (apresentando o maior escore, de 68) e rs2274844, sobreposto ao exon 3 do lincRNA *TCONS_00015773*.

LEGENDA: A gradação de cor é proporcional a força de detecção do sinal observado por chip-seq. As regiões verdes representam o motivo de ligação de maior pontuação para a proteína correspondente. As proteínas de ligação sobrepostas ao SNP, com o maior sinal (representadas em preto) são: POLR2A (*DNA-directed RNA polymerase II subunit RPB1*), RELA

(*nuclear factor NF-kappa-B p65*), RUNX3 (*Runt-related transcription factor 3*), POU2F2 (octamer transcription factor) e SPI1 (*Transcription factor PU.1*).

FONTE: UCSC Genome browser (https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr9%3A137827782-137835296&hgsid=454209933). Acesso em: 10.11.2015

5.1.5 Associações compartilhadas entre *FCN1* e *OLFM1*

Cerca de 70 eQTLs para *FCN1* também foram detectados como eQTLs para *OLFM1* (WESTRA *et al*., 2013), gene vizinho de *FCN1,* localizado a 160 kb (também em 9q34.3). A função exata de *OLFM1* (olfactomedin 1) ainda não é conhecida, mas devido a sua expressão abundante no cérebro, sugere-se que possa ter um papel essencial no tecido nervoso (HILL *et al*., 2015). Foi observada uma forte correlação negativa entre estes eQTLs (coeficiente de correlação de Spearman: r = -0.91, p = 2.2E-16); ou seja, todos os eQTLs que ativam a expressão de *FCN1* (Z positivo), reprimem a de *OLFM1* (Z negativo), e vice-versa (figura 13). Este resultado leva-nos a sugerir que estes genes compartilham elementos reguladores, que atuam de forma oposta sobre sua expressão em células sanguíneas. Aparentemente, o SNP que integra um eQTL intensificador para *FCN1*, simultaneamente integra um eQTL silenciador para *OLFM1,* e vice-versa. Uma vez que estes genes são mais expressos em tecidos distintos, é possível que níveis elevados de um deles, retroalimentem o silenciamento da expressão do gene vizinho, através dos mesmos elementos reguladores. Para testar esta hipótese, seria necessário verificar se estes SNPs também representam eQTLs no tecido cerebral, onde *OLFM1* é mais expresso. O estudo dos mecanismos que explicam estas associações opostas com a expressão de genes vizinhos, ainda é incipiente. Possivelmente, o estudo sistemático de eQTLs compartilhadas entre um ou mais genes poderiam revelar novas associações desta espécie, como objeto de pesquisa para futuros estudos funcionais.

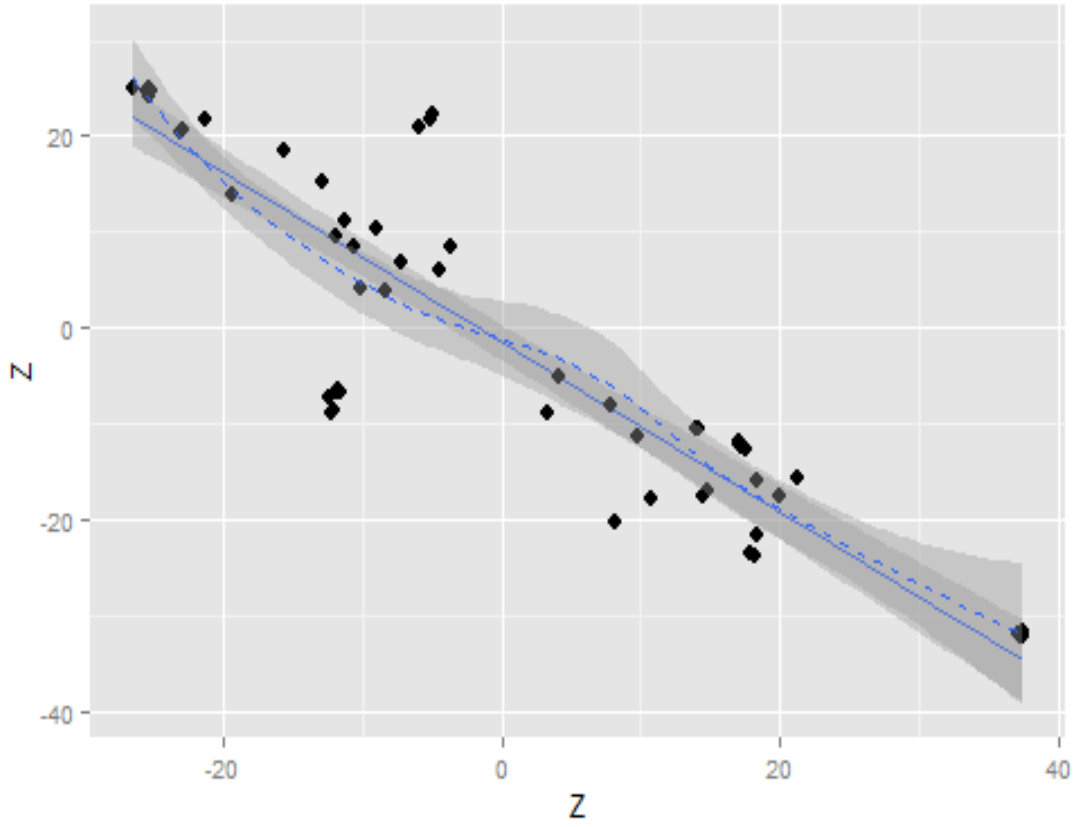

FIGURA 13 - Correlação negativa entre os valores de Z para as 70 eQTLs compartilhadas entre os genes *FCN1* e *OLFM1*.

NOTA: coeficiente de correlação de Spearman: r = -0.91, p = 2.2E-16.

LEGENDA: A linha em azul representa a reta de regressão linear. Em pontilhado, uma linha não-paramétrica de tendência ("smooth" ou suavização) para a visualização da dispersão.

FONTE: O autor, utilizando o pacote ggplot2 (WICKHAM 2009) através do software R versão 3.2.2 (https://www.r-project.org).

5.2 Resultados para *FCN2*

O gene *FCN2* é expresso em grandes quantidades no fígado e glândula adrenal, de onde a proteína é liberada no sangue, em forma solúvel. A partir do banco de dados Blood eQTLs browser foram encontrados 3 cis-eQTLs e 2 trans-eQTLs para *FCN2* (APÊNDICE B). Devido ao DL com os SNPs do banco de dados 1000 Genomes para a população europeia, estes eQTLs foram reunidos a mais 28 SNPs, somando 33 SNPs divididos em 4 blocos, sendo que dois cis-eQTLs encontravam-se no mesmo bloco.

5.2.1 Anotações funcionais

Nenhum dos SNPs foi encontrado em sobreposição a fatores reguladores detectados em linhagens sanguíneas, portanto, os escores correspondentes foram construídos apenas sobre as anotações oriundas do total de linhagens celulares (tabela 4). Apenas um dos SNPs (rs6537977) foi encontrado em sobreposição a proteínas de ligação.

A partir dos escores finais gerados pelo número de anotações, foram selecionados 4 SNPs com os maiores escores (tabela 3), cada um referente a um bloco em DL previamente averiguado ($r^2 \geq 0,8$). Nenhum destes SNPs foi previamente investigado com relação a uma possível associação com as concentrações séricas da ficolina 2 (MUNTHEFOG *et al.* 2007).

TABELA 3 - SNPs selecionados como os principais candidatos a influenciarem a regulação gênica de FCN2.

SNPs	Nomenclatura genômica	MAF %	∑ Escore	Proteínas de ligação ao DNA
rs12891687	*14:g.84763688A/G*	46	2	0
rs16872085	*8:g.105957540A/G*	09	5	0
rs6537977	*9:g.137915572A>G*	26	86	6
rs11103552	*9:g.137754467A>G*	17	14	0

NOTA: a seleção foi feita a partir dos maiores escores e sobreposição a sítios reconhecidos por proteínas de ligação ao DNA. Nenhum dos SNPs sequenciados por HUMMELSHOJ et al. (2008) consta na lista.

LEGENDA: MAF frequência do alelo menos comum na população europeia (segundo dados do projeto 1000 Genomes).

O SNP rs12891687 (*14:g.84763688A>G)* ocorre em frequências altas em todas as populações menos na africana, de 46% na população europeia e de 3% na população africana (dados acessados em 20.11.2015, disponíveis em:http://grch37.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=14:84763188-84764188;v=rs12891687;vdb=variation;vf=110611323). Este SNP atua como um trans-eQTL para *FCN2*, estando localizado em 9q34.3, e ocorre em DL com

rs11159647, o qual foi associado com a suscetibilidade para a doença de Alzheimer em um estudo genômico de associação (GWAS) (BERTRAM *et al*., 2008), embora não ocorra sobreposto a elementos reguladores. A única evidência de um possível efeito regulador deste trans-eQTL assenta-se na alteração de dois motivos de ligação a fatores de transcrição detectados por PWM (do inglês: *Position weight matrix*), e por sua localização no único intron de um lincRNA (código *TCONS_00022821*) (CABILI *et al*., 2011). rs12891687 pode estar alterando uma sequência regulatória presente no intron. É possível que este lincRNA tenha um efeito sobre a regulação de *FCN2*, ou sobre outro gene ou via que afete *FCN2*.

O SNP rs16872085 (*8:g.105957540A/G)* é um trans-eQTL localizado em 8q22.3, e ocorre em frequências baixas na maioria das populações, de 9% na população europeia e 27% no leste asiático (dados acessados em 20.11.2015, disponíveis em: http://grch37.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=8:105957040-105958040;v=rs16872085;vdb=variation;vf=111100872#population_freq_EAS).
Este SNP foi detectado em um GWAS, relacionado à parada cardíaca súbita (AOUIZERAT *et al*., 2011). A única evidência regulatória para este SNP é a alteração de 5 motivos de ligação a fatores de transcrição detectados por PWM.

O SNP rs6537977 (*9:g.137915572A>G*), em DL com os eQTLs rs7852219 e rs7043210, possui frequências altas na maioria das populações, sendo de 26% na população europeia e 10% no leste asiático (dados acessados em 20.11.2015, disponíveis em: http://grch37.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=9:137915072-137916072;v=rs6537977;vdb=variation;vf=106304204). Este SNP foi detectado com o maior escore (86), e está sobreposto aos sítios reconhecidos por seis proteínas de ligação, sítios de hipersensibilidade a DNAse e sobre um intensificador. rs7852219 (*9:g.137916810C>T*) também é predito como um eQTL para o gene vizinho *OLFM1*, localizado a 180 kb.

O eQTL rs11103552 (*9:g.137754467A>G*), com a frequência de 17% na população europeia (dados acessados em 20.11.2015, disponíveis em: http://grch37.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=9:1377539

67-137754967;v=rs11103552;vdb=variation;vf=109110544), apresentou escore 14, estando localizado em locais de modificação de histonas, um intensificador e um motivo para um fator de transcrição detectado por PWM.

5.3 Resultados para *FCN3*

O gene *FCN3* é transcrito em maiores quantidades no pulmão. Foram encontrados apenas dois eQTLs para *FCN3*, o trans-eQTL rs7226677 localizado em 18p11.32, a partir do banco de dados Blood eQTL browser, e o cis-eQTL rs71636790 localizado em 1p36.11, associado com a expressão gênica da *FCN3* em 13 tecidos, segundo dados do Portal Gtex (GTEX CONSORTIUM, 2013) (acessado em 10.09.2015: http://www.gtexportal.org/home/gene/FCN3). Apenas o SNP rs62073531 ocorre em DL com rs7226677, mas 53 outros SNPs estão ligados a rs71636790, no mesmo bloco de DL. Para estes, foram construídos os escore Z (APÊNDICE C).

5.3.1 Anotações funcionais e contexto regulatório

O trans-eQTL rs7226677 (*18:g.278796A>G*), com a frequência de 13% na população europeia (dados acessados em 20.11.2015, disponíveis em: http://grch37.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=18:278296-279296;v=rs7226677;vdb=variation;vf=106943160), já foi associado ao transtorno bipolar e à depressão (LIU *et al.,* 2010). Devido a este fato, é presumível que este SNP tenha um forte papel funcional, entretanto, a única evidência para tanto, é a alteração de 6 motivos de ligação para fatores de transcrição detectados por PWM.

O SNP rs71636790 (*1:g.27214245A>G*) ocorre em frequências baixas em todas as populações, sendo de 13% em europeus e 2% no leste asiático (dados acessados em 20.11.2015, disponíveis em: http://grch37.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=1:27213745-27214745;v=rs71636790;vdb=variation;vf=116185683). Este SNP, além dos 53 SNPs que ocorrem em DL, está localizado em uma região muito rica em fatores reguladores, em todos os tipos celulares estudados, sendo provável que caracterize uma região central na regulação de um grande conjunto de genes (figura 14).

FIGURA 14 - Contexto regulatório da sequência onde são encontrados os SNPs em DL com o eQTL rs71636790.

NOTA: Os SNPs selecionados estão localizados a aproximadamente 600.000 pb de *FCN3*, em 1p36.11. A região revela forte atividade reguladora para todos os tipos celulares estudados pelo ENCODE, dos quais nove estão representados.

LEGENDA: As cores representam o estado da cromatina em nove linhagens celulares estudadas pelo ENCODE (GM12878, H1-hESC, K562, HepG2, HUVEC, HMEC, HSMM, NHEK e NHLF). Em vermelho escuro e claro estão representados regiões de promotor ativo e fraco, respectivamente. Em amarelo escuro e claro, regiões de intensificador forte e fraco, respectivamente. Em verde escuro e claro, regiões de transcrição e alongamento transcricional forte e fraco, respectivamente. Em azul claro estão representadas regiões de contato da cromatina, caracterizadas pela ligação de CTCF (ERNST; KELLIS, 2010).

FONTE: UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr1%3A27117426-27267725&hgsid=454209933). Acesso em: 10.11.2015

Embora apenas rs71636790 tenha sido identificado como eQTL para *FCN3*, todos os demais 53 SNPs em DL são eQTLs para ao menos um outro gene (dados acessados em 05.11.2015, disponíveis em http://www.broadinstitute.org/mammals/haploreg/haploreg.php). Para aumentar a precisão na identificação do SNP com maior influência causal sobre a expressão de *FCN3,* foram selecionados os dois SNPs com maior frequência de sobreposição a fatores de transcrição e regiões com forte atividade reguladora. O SNP rs1883660 (figura 15) (*1:g.27191411A>T*) tem frequências relativamente baixas em todas as populações, de 13% em europeus e 2% no leste asiático (dados acessados em 20.11.2015, disponíveis em:http://grch37.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=1:27190911-27191911;v=rs1883660;vdb=variation;vf=104092787). Possui o maior escore, de 607. O SNP rs71636795 (*1:g.27241025A>G*) (figura 16), com frequências iguais a rs1883660, apresentou o escore 151.

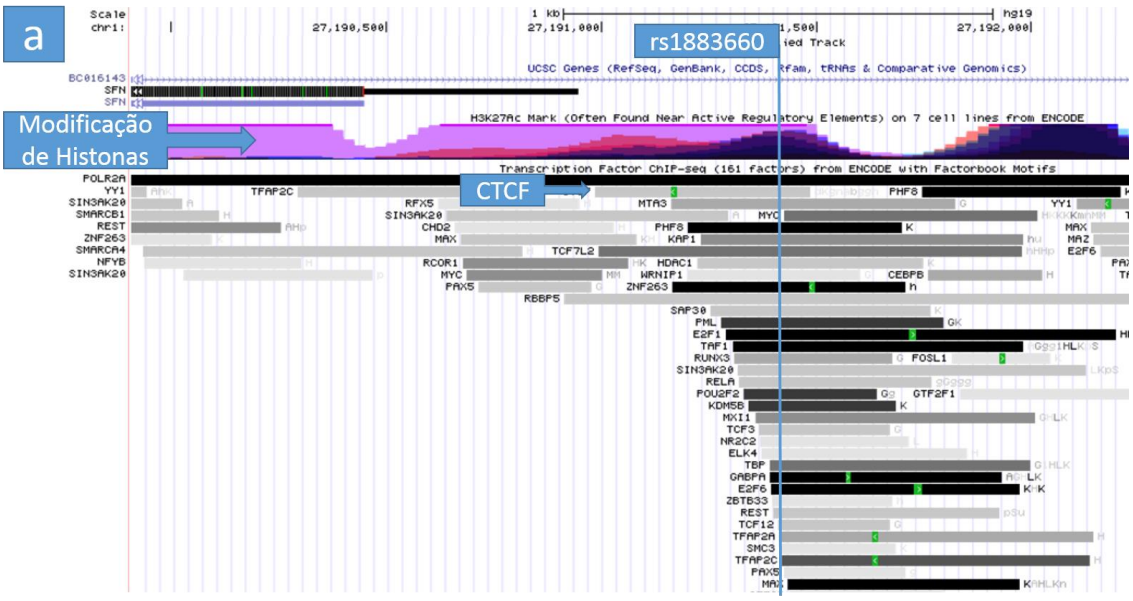

FIGURA 15 - Contexto regulatório do SNP rs1883660.

NOTA: é mostrada a sobreposição de SNPs a sítios de ligação de proteínas ao DNA e sítios de modificação de histonas H3K27Ac (acetilação da lisina 27 da histona H3, relacionada à intensificação da transcrição). Em destaque, a proteína CTCF, importante fator de modificação conformacional da cromatina.

LEGENDA: A gradação de cor é proporcional à força de detecção do sinal observado por ChIp-seq. As regiões verdes representam o motivo de ligação de maior pontuação para a proteína correspondente. As proteínas de ligação sobrepostas ao SNP, com o maior sinal (representadas em preto) são: POLR2A (*DNA-directed RNA polymerase II subunit RPB1*), ZNF263, PHF8 (*PHD finger protein 8*), E2F1 (*Transcription factor E2F1*), TAF1, GABPA (*GA-binding protein alpha*) e E2F6.

FONTE: UCSC UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr1%3A27191035-27191786&hgsid=454209933). Acesso em: 10.11.2015

FIGURA 16 - Contexto regulatório do SNP rs71636795.

NOTA: é mostrada a sobreposição do SNP rs71636795 a sítios de ligação de proteínas ao DNA e sobre sítios de modificação de histonas H3K27Ac (acetilação da lisina 27 da histona H3, relacionada à intensificação da transcrição).

LEGENDA: A gradação de cor é proporcional a força de detecção do sinal observado por ChIp-Seq. As regiões verdes representam o motivo de ligação de maior pontuação para a proteína correspondente. As proteínas de ligação sobrepostas ao SNP, com o maior sinal (representadas em preto) são: POLR2A (*DNA-directed RNA polymerase II subunit RPB1*), EP300 (*E1A binding protein p300*), FOXA1 (*Forkhead box protein A1*), FOXA2 (*Forkhead box protein A2*) e HNF4G (*Hepatocyte nuclear factor 4 gamma*).

FONTE: UCSC UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr1%3A27240775-27241275&hgsid=454209933). Acesso em: 10.11.2015

Ao averiguar possíveis alças da cromatina em fibroblastos de pulmão (IMR90) ligando a região dos polimorfismos com *FCN3*, detectou-se um domínio topológico, calculado usando o programa Juicebox (SUHAS *et al*., 2014) na região 1:27247414-27647413 (figura 17). Fatores de transcrição sobrepostos a SNPs no mesmo bloco de DL, localizados próximos, podem favorecer a formação deste domínio, levando intensificadores próximos ao promotor de *FCN3*.

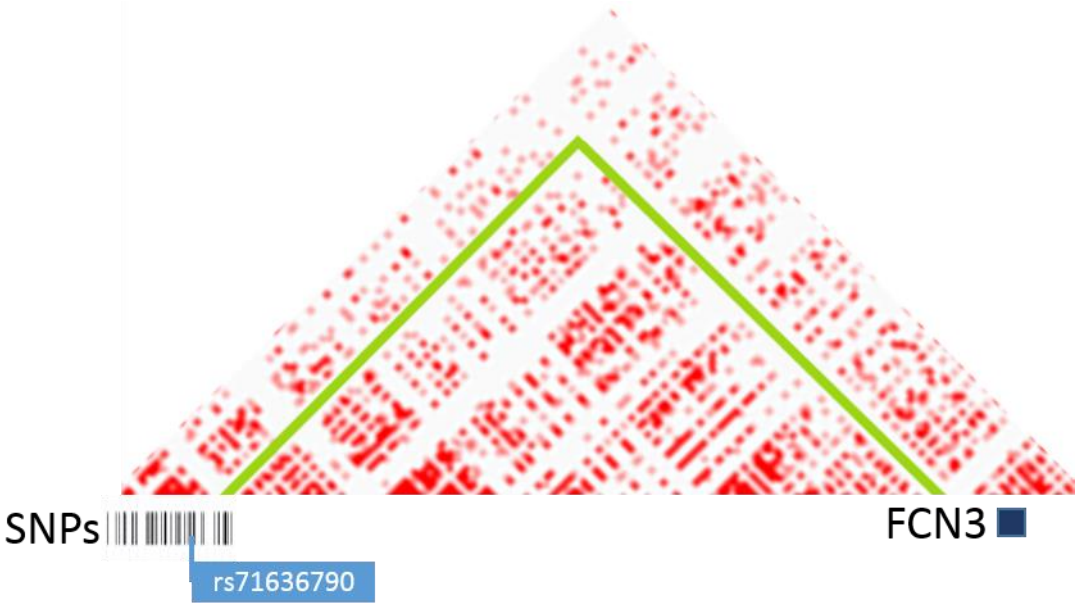

FIGURA 17 - Domínio topológico marcando regiões nas adjacências de *FCN3* e dos 53 SNPs em DL com o eQTL rs71636790.

NOTA: O domínio topológico, representado em verde, foi calculado pelo programa Juicebox entre as posições 1:27247414-27647413.

FONTE: O autor, usando o programa Juicebox (SUHAS *et al*., 2014), para a linhagem celular de fibroblastos de pulmão (IMR90, in situ+combined).

5.4 Resultados para *MBL2*

O gene *MBL2* é expresso em maiores quantidades no figado, mas também pode ser encontrado expresso em baixas quantidades no nervo tibial. Foram encontrados 29 eQTLs associados à expressão de *MBL2* segundo o portal Gtex (GTEx Consortium,

2013) (http://www.gtexportal.org/home/gene/MBL2), 20 das quais para o fígado, e 20 para o nervo tibial (APÊNDICE D). Para a análise subsequente, foram selecionados apenas os 20 eQTLs para *MBL2* no fígado, devido à sua maior expressão neste tecido. Estes eQTLs apresentaram-se em DL com 32 SNPs, os quais foram divididos em 6 blocos.

5.4.1 Anotações funcionais e contexto regulatório

A construção dos escores foi realizada a partir de anotações funcionais oriundas de células do fígado. A partir dos maiores escores e dos SNPs sobrepostos a proteínas de ligação, foram selecionados 9 SNPs como principais candidatos funcionais (tabela 4).

TABELA 4 - SNPs selecionados como os principais candidatos funcionais sobre a regulação gênica de *MBL2*, a partir dos maiores escores e proteínas de ligação ao DNA

Blocos de DL	SNPs	Nomenclatura genômica	MAF %	∑ Escore	Proteínas de ligação ao DNA
1	rs930507	*10:g.54528298G>C*	19	18	8
	rs930508	*10:g.54528298G>C*	19	15	5
	rs930509	*10:g.54528353C>G*	20	11	1
2	rs7096206	*10:g.54531685G>C*	22	13	2
	rs10824796	*10:g.54534613A>G*	22	28	19
3	rs2165811	*10:g.54535020T>C*	28	11	0
4	rs111492012	*10:g.54538913delA*	27	5	0
5	rs11003137	*10:g.54539897G>A*	20	6	0
6	rs7093304	*10:g.54546586A>T*	30	2	0

LEGENDA: MAF frequência do alelo menos comum na população europeia (segundo dados do projeto 1000 Genomes).

O SNP rs930507 (*10:g.52768506G>C*), do primeiro bloco de DL, é uma mutação sinônima localizada no último códon do primeiro exon de *MBL2*, e já foi associado ao aumento do transporte de sódio-lítio em eritrócitos, observado em pacientes com hipertensão arterial (MORRISON *et al*., 2008) (figura 18).

FIGURA 18 - Contexto regulatório do SNP rs930507.

NOTA: rs930507 forma um agrupamento próximo a rs930508 e rs930509, sobrepostos à região reconhecida por 8 diferentes proteínas reguladoras em hepatócitos.

LEGENDA: A gradação de cor é proporcional a força de detecção do sinal obcervado por chip-seq. As regiões verdes representam o motivo de ligação de maior pontuação para a proteína correspondente. A proteína de ligação com o maior sinal (representadas em preto) é a EP300 (*E1A binding protein p300*).

FONTE: UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr10%3A54528016-54528516&hgsid=454209933). Acesso em: 10.11.2015

O SNP rs10824796 (*10:g.54534613A>G*), apresentando o maior escore entre as eQTLs, de 29, está sobreposto ao local de ligação de 19 fatores de transcrição, em um sítio marcado como intensificador e próximo a um sítio de modificação de histonas H3K4Me3, característico de promotor ativado (figura 19). O contexto regulatório torna este SNP, ainda não estudado, um forte candidato para estudos de associação.

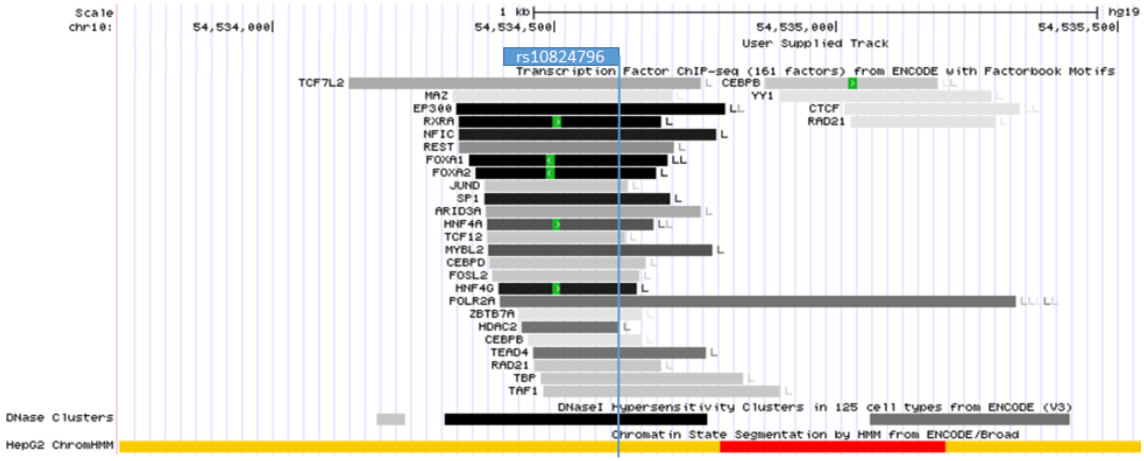

FIGURA 19 - Contexto regulatório do SNP rs10824796.

NOTA: Este SNP está localizado sobreposto à sequência reconhecida por, pelo menos, 19 proteínas reguladoras detectadas em hepatócidos.

LEGENDA: A gradação de cor é proporcional à força de detecção do sinal observado por ChIp-Seq. As regiões verdes representam o motivo de ligação de maior afinidade para a proteína correspondente. As proteínas de ligação sobrepostas ao SNP, com o maior sinal (representadas em preto) são: EP300 (*E1A binding protein p300*), RXRA (*Retinoid X receptor alpha: RXR-alpha*), FOXA1 (*Forkhead box protein A1*), FOXA2 (*Forkhead box protein A2*) e HNF4G (*Hepatocyte nuclear factor 4 gamma*).

FONTE: UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&position=chr10%3A54534363-54534863&hgsid=454209933). Acesso em: 10.11.2015

Ainda não foram realizados testes de HI-c para tecido de fígado. Considerando-se que os domínios topológicos são pouco variáveis entre diferentes tecidos, ou que, quando variáveis, possam representar alvos específicos na regulação de cada tecido, a avaliação da topologia da região genômica de *MBL2* foi testada para resultados de HI-c em linhagem celular linfoblastoide (GM12878). A avaliação revelou um domínio topológico englobando a maioria dos eQTLs, calculado pelo programa Juicebox (SUHAS *et al*., 2014), localizado entre 10:54380001-54555000 (figura 20.a.1).

O eQTL rs11003137 (*10:g.54539897G>A*) (escore 6), com a frequência de 20% em europeus (dados acessados em 20.11.2015, disponíveis em: http://grch37.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=10:545393

97-54540397;v=rs11003137;vdb=variation;vf=109019039), não está sobreposto a nenhum sítio reconhecido por proteínas reguladoras no fígado, mas encontra-se sobreposto ao de CTCF (figura 20.b), em linhagem celular de colo do útero (HeLa-S3) e osteoblastos (Osteobl), numa localização que flanqueia o domínio topológico principal (figura 20.a.1). Entretanto, em fígado (HepG2) esta região é marcada por modificação de histonas característica de um promotor ativo (H3K4Me3).

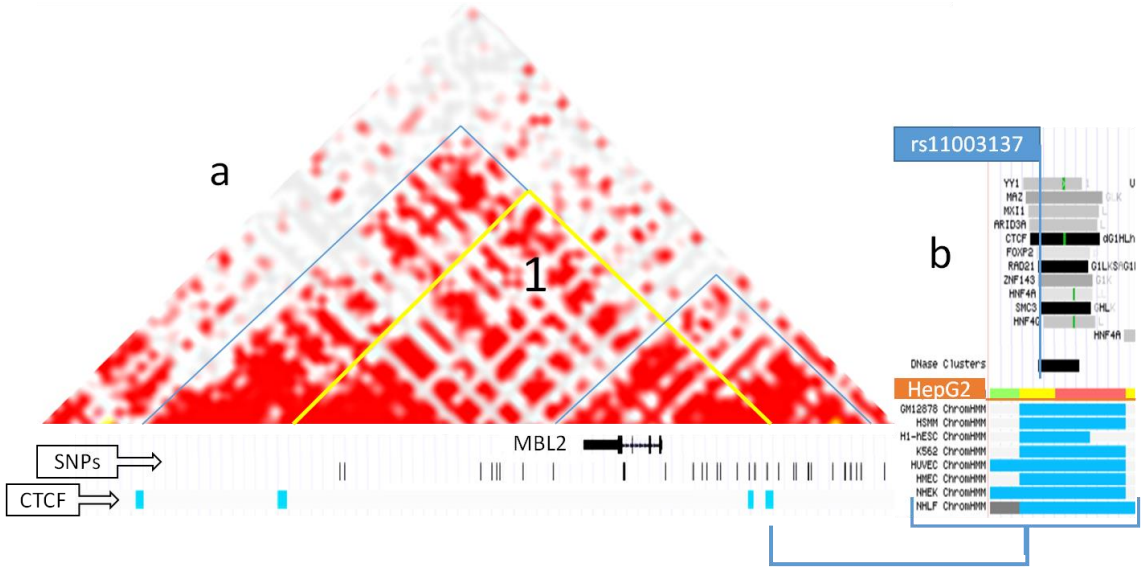

FIGURA 20 - Contexto regulatório dos SNPs associados à modulação da expressão gênica de *MBL2.*

NOTA: O gráfico da topologia (a) foi feito para linhagem celular linfoblastoide (GM12878). O domínio topológico calculado pelo programa Juicebox (SUHAS *et al*., 2014) é representado por linhas amarelas (a.1). Outros possíveis domínios são representados por linhas azuis. O SNP rs11003137 (b) está localizado em região de interação da cromatina, sobreposto a uma região isoladora (em azul claro), dentre outros fatores de transcrição, em 8 linhagens celulares, contudo, esta sob um promotor ativo em HepG2 (hepatócitos), representado em vermelho claro (b).

FONTE: O autor, através do programa Juicebox (http://www.aidenlab.org/juicebox/) (SUHAS et al., 2014), para a linhagem celular linfoblastoide (GM12878, in situ Mbol, primary+replicate).

CONCLUSÃO

- A integração dos dados de Genômica Funcional com os de desequilíbrio de ligação, oriundos de estudos de Genética de Populações, permitiu explorar o contexto regulatório de genes que codificam proteínas iniciadoras da via das lectinas do complemento.
- Para *FCN1*, identificaram-se 406 SNPs (188 eQTLs e 213 SNPs em DL com eles), dos quais 97,5% associados a pelo menos um elemento regulador da expressão gênica em células sanguíneas (P<0,003), distribuídos em 16 blocos de desequilíbrio de ligação em uma sequência que engloba os genes *FCN2, OLFM1* e *COL5A1*, além de *FCN1* e vários lincRNAs.
- Trinta e sete destes SNPs apresentaram escores mais elevados, dos quais onze estão sobrepostos a proteínas de ligação, e três a RNAs não codificantes.
- SNPs de FCN1 com os maiores escores estão posicionados sobre domínios topológicos da cromatina segundo dados de Hi-c, que coincidem com a intensidade de metilação da lisina 4 da histona H3, regiões susceptíveis à DNAse e reconhecidas pelo CTCF.
- Houve uma correlação negativa entre os níveis de expressão de *FCN1* e seu gene vizinho, *OLFM1,* associados a 70 das 188 eQTLs.
- Para o gene *FCN2,* identificou-se quatro SNPs como os principais candidatos funcionais, cada qual pertencente a um de quatro blocos de DL.
- Para *FCN3*, foram selecionados dois SNPs sobrepostos a fatores de transcrição em uma região de intensa atividade regulatória.
- Para *MBL2,* foram encontrados 9 SNPs em 6 blocos de DL, a maioria incluída em um domínio topológico da cromatina, como principais candidatos funcionais para a regulação da expressão deste gene no fígado.
- O emprego de dados referentes à conformação tridimensional da cromatina teve um papel importante na visualização do contexto

regulatório das associações, a maioria das quais estão localizadas a grandes distâncias do gene.

- Grande parte dos SNPs sobrepostos a sítios com evidência funcional encontravam-se relacionados a regiões de interação da cromatina, comprovando a importância relativa à consideração destes dados em uma análise integrativa.
- SNPs associados a mais de um gene, como no caso de *FCN1* e *OLFM1*, podem indicar que tais genes compartilham os mesmos sítios regulatórios, e gerar insights sobre o mecanismo por trás das vias regulatórias.

REFERÊNCIAS BIBLIOGRÁFICAS

ABBAS, A.K.; *et al*. Imunologia Celular e Molecular. 7 ed. **Elsevier,** 2012.

ADDOBBATI, C.; SILVA, J.A.; TAVARES, N.A.C.; MONTICIELO, O. Ficolin Gene Polymorphisms in Systemic Lupus Erythematosus and Rheumatoid Arthritis. **Human genetics**, 2015.

ABECASIS, G.R.; ALTSHULER, D.; AUTON, A.; BROOKS, L.D.; DURBIN, R.M.; GIBBS, R.A.; HURLES, M.E.; MCVEAN, G.A. A map of human genome variation from population-scale sequencing. **Nature**, v.467, p. 1061-1073, 2010.

ABZHUBEI, I.A.; SCHMIDT, S.; PESHKIN, L.; RAMENSKY, V.E.; GERASIMOVA, A.; BORK, P.; KONDRASHOV. A.S.; SUNYAEV, S.R. A method and server for predicting damaging missense mutations. **Nat Methods**, v. 7, p. 248–249, 2010.

ALVES PEDROSO, M.L.; BOLDT, A.B.; PEREIRA-FERRARI, L.; STEFFENSEN, R.; STRAUSS, E.; JENSENIUS, J.C.; IOSHII, S.O.; MESSIAS-REASON, I. Mannan-binding lectin MBL2 gene polymorphism in chronic hepatitis C: association with the severity of liver fibrosis and response to interferon therapy. **Clinical and Experimental Immunology**, v. 152, p. 258-264, 2008.

AMMITZBOLL, C.G.; STEFFENSEN, R.; NIELSEN, H.J.; THIEL, S.; STENGAARD-PEDERSEN, C.; BØGSTED, m.; JENSENIUS, J.C.;. Polymorphisms in the MASP1 Gene Are Associated with Serum Levels of MASP-1, MASP-3, and MAp44. **PLoS One**, v. 8(9), p. 73317, 2013.

AOUIZERAT, BE., *et al*. GWAS for discovery and replication of genetic loci associated with sudden cardiac arrest in patients with coronary artery disease. **BMC Cardiovasc Disord**. 2011.

BANNISTER, A. J; KOUZARIDE, T. Regulation of chromatin by histone modifications. **Cell Research**, v. 21, n.3, p.381-395, 2011.

BERTRAM, L., *et al*. Genome-wide association analysis reveals putative Alzheimer's disease susceptibility loci in addition to APOE. **Am J Hum Genet**, v. 83(5), p. 623-32, 2008.

BELTRAME, M. H. et al. The lectin pathway of complement and rheumatic heart disease. **Frontiers in pediatrics**, v. 2, n. January, p. 148, jan. 2014.

BERNSTEIN, BE.; STAMATOYANNOPOULOS, J.A.; COSTELLO, J.F.; REN, B.; MILOSAVLJEVIC, A.; MESSNER, A.; KELLIES, M.; MARRA, M.A.; BEAUDET, A.L.; ECKER, J.R.; FARNHAM, P.J.; HIRST, M.; LANDER, E.S.; MIKKELSEN, T.S.; THOMSON, J.A. The NIH Roadmap Epigenomics Mapping Consortium. **Nat Biotechnol,** v. 28, p. 1045-1048, 2010.

BERNSTEIN, BE.; BIRNEY, E.; DUNHAM, I.; GREEN, ED.; GUNTER, C.; SNYDER, M. "An integrated encyclopedia of DNA elements in the human genome". **Nature,** v. 489(7414), p. 57–74, 2012.

BICKMORE, W.A. (2013). The spatial organization of the human genome. **Annu.Rev. Genomics Hum. Genet**. v. 14, p. 67–84, 2013.

BOLDT, A. B.; LUTY, A.; GROBUSCH, M. P.; DIETZ, K.; DZEING, A.; KOMBILA, M.; KREMSNER, P. G.; KUN, J. F. Association of a new mannose-binding lectin variant with severe malaria in Gabonese children. **Genes and Immunity.** v.7, p.393 - 400, 2006.

BOLDT A.B.; SANCHEZ M.I.; STAHLKE E.R.; STEFFENSEN R.; THIEL S.; JENSENIUS J.C.; PREVEDELLO F.C.; MIRA M.T.; KUN J.F.; MESSIAS- REASON I.J. Susceptibility to Leprosy is Associated with M-ficolin Polymorphisms. **J Clin Immunol**, v. 33, n.1, p.210-219, 2013.

BREM, R.B.; YVERT. G.; CLINTON, R.; KRUGLYAK, L. Genetic dissection of transcriptional regulation in budding yeast. **Science.** v. 296, p. 752–755, 2002.

CHEN, X.; KATOH, Y.; NAKAMURA, K.; OYAMA, N.; KANEKO, F.; ENDO, Y.; *et al*. Single nucleotide polymorphisms of Ficolin 2 gene in Behçet's disease. **J Dermatol Sci**, v. 43, p. 201-205, 2006.

COOKSON, W.; LIANG, L.; ABECASIS, G.; MOFFATT, M.; LATHROP, M. Mapping complex disease traits with global gene expression. **Nature Review Genetics.** v. 10, p. 184–194, 2009.

COOPER, G.M.; SHENDURE, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. **Nature Review Genetics**. v. 12, p. 628-640, 2011.

DOMMETT, R. M.; KLEIN, N.; TURNER, M. W. Mannose-binding lectin in innate immunity: past, present and future. **Tissue Antigens**. v.68, n.3, p.193-209, 2006.

ERNST, J.; KELLIS, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. **Nature Biotechnology.** v.28, p. 817-25, 2010.

FEARON, D.T.; LOCKSLEY, R.M.;. The instructive role of innate immunity in the acquired immune response. **Science**, v. 272, p. 50-53, 1996.

FLICEK, P.; *et al*. Ensembl 2014. **Nucleic Acids Research,** v. 42, p. 749-755, 2014.

FORREST, A.R.; KAWAJI, H.; REHLI, M.; BAILLIE, J.K.; HOON, M.J.; LASSMANN, T.; ITOH, M.; SUMMERS, K.M.; SUZUKI, H.; DAUD, C.O.; KAWAI, J.; HEUTINK, P.; HIDE, W.; FREEMAN, T.C.; LENHARD, B.; BAJIC, V.B.; TAYLOR, M.S.; MAKEEV, V.J.; SANDELIN, A.; HUME, D.A.; CARNINCI, P.; HAYASHIZAKI, Y. A promoter-level mammalian expression atlas. **Nature**, v. 507, p. 462-470, 2014.

FRANC, N.C.; WHITE, K.; EZEKOWITZ, R.A. Phagocytosis and development: back to the future. **Current Opinion in Immunology,** v. 11, p. 47-52, 1999.

GARRED, P.; HONORÉ, C.; MA, Y. J.; et al. The genetics of ficolins**. Journal of Innate Immunity**, v. 2, p. 3–16, 2009.

GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. **Nat Genet**, v. 6, p. 580-5, 2013.

HÉJA, D. et al. Revised mechanism of complement lectin-pathway activation revealing the role of serine protease MASP-1 as the exclusive activator of MASP-2. **Proceedings of the National Academy of Sciences of the United States of America**, v. 109, n. 26, p. 10498–503, 26 jun. 2012.

HILL. S.E.; DONEGAN. R.K.; NGUVEN. E.; DESAI. T.M.; LIEBERMAN. R.L. Molecular Details of Olfactomedin Domains Provide Pathway to Structure-Function Studies. **PLoS One**. v. 10, 2015.

HOLMSKOV, U.; THIEL, S.; JENSENIUS, J.C. Collections and ficolins: humoral lectins of the innate immune defense. **Annual Review of Immunology**. v. 21, p.547–578, 2003.

HONORE, C.; *et al*. The innate immune component ficolin 3 (Hakata antigen) mediates the clearance of late apoptotic cells. **Arthritis Rheum**, v. 56, n. 5, p. 1598-1607, 2007.

HUMMELSHOJ, T.; MUNTHE-FOG, L.; MADSEN, H.O.; GARRED, P. Functional SNPs in the human ficolin (FCN) genes reveal distinct geographical patterns. **Molecular Immunology**. v.45, p.2508–2520, 2008.

JARINOVA, O.; STEWART, AF.; ROBERTS, R.; WELLS, G.; LAU, P.; NAING, T.; BUERKI, C.; MCLEAN, BW.; COOK, RC.; PARKER, JS.; *et al*. Functional analysis of the chromosome 9p21.3 coronary artery disease risk locus. **Arterioscler Thromb Vasc Biol**, v. 29, p. 1671–1677, 2009.

JIN, F., *et al.* A high-resolution map of the three-dimensional chromatin interactome in human cells. **Nature**. v. 503(7475), p. 290-4, 2013.

KASOWSKI, M.; GRUBERT, F.; HEFFELFINGER, C.; HARIHARAN, M.; ASABERE, A.; WASZAK, SM.; HABEGGER, L.; ROZOWSKY, J.; SHI M, URBAN, AE.; *et al*. Variation in transcription factor binding among humans. **Science**, v. 328, p. 232–235, 2010.

KENT, W.J.; KAROLCHIK, D.; ANGIE, S. The UCSC Genome Browser. **Curr Protoc Bioinformatics**, 2009.

KILPATRICK, D. C.; FUJITA, T.; MATSUSHITA, M. P35, an opsonic lectin of the ficolin family, in human blood from neonates, normal adults, and recurrent miscarriage patients. **Immunol.Lett**, v. 67, n. 2, p. 109-112, 1999.

KRARUP, A. et al. Effect of capsulation of opportunistic pathogenic bacteria on binding of the pattern recognition molecules mannan-binding lectin, L-ficolin, and H-ficolin. **Infect.Immun**, v. 73, n. 2, p. 1052-1060, 2005.

LANDER, E.; LINTON, L.; BIRREN, B.; NUSBAUM, C.; ZODY, M.; BALDWIN, J.; DEVON, K.; DEWAR, K.; DOYLE, M.; FITZHUGH, W.; *et al*. Initial sequencing and analysis of the human genome. **Nature.** v. 409, p. 860–921, 2001.

LANDERS, JE.; MELKI, J.; MEININGER, V.; GLASS, JD.; VAN DEN BERG, LH.; VAN, ES MA.; SAPP, PC.; VAN VUGHT, PW.; MCKENNA-YASEK, DM.; BLAUW, HM.; *et al*. Reduced expression of the Kinesin-Associated Protein 3 (KIFAP3) gene increases survival in sporadic amyotrophic lateral sclerosis. **Proc Natl Acad Sci**, v. 106, p. 9004–9009, 2009.

LIEBERMAN, A.E.; BERKUM, V.N.L.; WILLIAMS, L.; IMAKAEV, M.; RAGOCZY, T.; TELLING, A.; AMIT, I.; LAJOIE, B.R.; SABO, P.J.; DORSCHENER, M.O.; *et al*. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. **Science**, v. 326, p. 289–293, 2009.

LIU, Y. *et al.* Meta-analysis of genome-wide association data of bipolar disorder and major depressive disorder. **Molecular Psychiatry**. v. 16, p. 2-4, 2011.

LONSDALE, J., *et al*. The Genotype-Tissue Expression (GTEx) project. **Nature** Genetics, v. 45, p. 580–585, 2013.

MAHER, B. ENCODE: The human encyclopaedia. **Nature, v.** 489 (7414), p. 46–8, 2012.

MATZINGER, P. The danger model: a renewed sense of self. **Science**, v. 296, p. 301-305, 2002.

MESSIAS-REASON I. J.; BOLDT A. B.; MORAES BRAGA A. C.; VON ROSEN SEELING S. E.; DORNELLES L.; PEREIRA-FERRARI L.; KREMSNER P. G.; KUN J. F. The association between mannan-binding lectin gene polymorphism and clinical leprosy: new insight into an old paradigm. **Journal of Infectious Diseases**, v.196, p.1379 - 1385, 2007.

MESSIAS-REASON, I.; KREMSNER, P. G.; KUN, J. F. Functional haplotypes that produce normal ficolin-2 levels protect against clinical leprosy. **Journal of Infectious Diseases**, v.199, p. 801–804, 2009.

MORRISON, A. C.; BOERWINKLE, E.; TURNER, S. T; FERRELL, R. E. Regional association-based fine mapping for sodium-lithium countertransport on chromosome 10. **American Journal of Hypertension**. v. 21, n.1, p. 117-121, 2008.

MUNTHE-FOG, L. *et al.* Variation in FCN1 affects biosynthesis of ficolin-1 and is associated with outcome of systemic inflammation. **Genes Immun**. v.13, n.7, p.515-522, 2012.

MUNTHE-FOG, L. *et al*. Characterization of a polymorphism in the coding sequence of FCN3 resulting in a Ficolin-3 (Hakata antigen) deficiency state. **Molecular Immunology**, v. 45, n. 9, p. 2660-2666, 2008.

MURPHY, K.; TRAVERS, P.; WALPORT, M. Imunobiologia de Janeway. Porto Alegre: **Artmed**, p. 61- 81, 2010a.

NICA, A. C.; DERMITZAKIS, E. T. Expression quantitative trait loci: Present and future, **Philosophical transactions of the Royal Society of London**. Series B, Biological sciences. v. 368, n. 1620, p. 20120362, 2013.

NG, P.C.; HENIKOFF, S. Predicting amino acid changes that affect protein function. **Nucleic Acids Res**, v. 31, p. 3812–3814, 2003.

PENNISI, E. Genomics. ENCODE project writes eulogy for junk DNA. **Science**, v. 337 (6099), p. 1159-1161, 2012.

PHILLIPS-CREMINS, J.E.; SAURIA, M.E.G.; SANYAL, A.; GERASIMOVA, T.I.; LAJOIE, B.R.; BELL, J.S.K.; ONG, C.T.; HOOKWAY, T.A.; GUO, C.; SUN, Y. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. **Cell**. V. 153, p. 1281–1295, 2013.

PEVSNER. J.; Bioinformatics and functional genomics. 2ed. Hoboken, 2009.

REN, T.; QIU, Y.; WU, W.; FENG, X.; YE, S.; WANG, Z.; TIAN, T.; HE, Y.; YU, C.; ZHOU, Y. Activation of adenosine A3 receptor alleviates TNF-α-induced inflammation through inhibition of the NF-κB signaling pathway in human colonic epithelial cells. **Mediators of inflammation**. 2014.

ROADMAP EPIGENOMICS CONSORTIUM. Integrative analysis of 111 reference human epigenomes. **Nature.** V.518, p. 317-30, 2015.

ROLAND, J.; *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. **Nature Comunications,** v. 6, 2014.

RUDAN, M.V.; BARRINGTON, C.; HENDERDON, S.; ERNST, C.; DUNCAN, T.; ODOM. TANAY, A.; HADJUR, S. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. **Cell Reports,** v. 10, p. 1297–1309, 2015.

ROOS, A.; RASTALDI, M.P.; CALVARESI, N.; OORTWIJN, B.D.; SCHLAGWEIN, N.; VAN GIJLSWIJK-JANSSEN, D.J.; *et al*. Glomerular activation of the lectin pathway of complement in IgA nephropathy is associated with more severe renal disease. **J Am Soc Nephrol,** v. 17, p. 1724-1734, 2006.

SACCONE, SF.; BOLZE, R.; THOMAS, P.; QUAN, J.; MEHTA, G.; DEELMAN, E.; TISCHFIELD, JA.; RICE, JP. A web-based tool for using biological databases to prioritize SNPs after a genome-wide association study. **Nucleic Acids Res,** v. 38, p. 201–209, 2010.

SCHAUB, M.A.; BOYLE, A.P.; KUNDAJE, A.; SNYDER, S. Linking disease associations with regulatory information in the human genome. **Genome Res**, v. 22 (9), p. 1748-1759, 2012.

STENGAARD-PEDERSEN, K.; THIEL, S.; GADJEVA, M.; MOLLER-KRISTENSEN, M.; SORENSEN, R.; JENSEN, L. T.; SJOHOLM, A. G.; FUGGER, L.; JENSENIUS, J. C. Inherited deficiency of mannan-binding lectin-associated serine protease 2. **New England Journal of Medicine**, v.349, p.554 - 560, 2003.

SUHAS. S.P.; *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. **Cell**, v. 159, 2014.

SUN, W.; HU, Y. EQTL mapping using RNA-seq data. **Statistics in biosciences***.* v. 5, n. 1, p. 198-219, 2013.

THE ENCODE PROJECT CONSORTIUM. An integrated encyclopedia of DNA elements in the human genome. **Nature**, v. 489(7414), p. 57–74, 2013.

THE INTERNATIONAL HAPMAP CONSORTIUM. Integrating common and rare genetic variation in diverse human populations. **Nature**. v. 467, p. 52-58, 2010.

THIEL, S.; GADJEVA, M. Humoral pattern recognition molecules: mannan-binding lectin and ficolins. **Adv Exp Med Biol,** v.653, p.58-73, 2009a.

VAN BERKUM, N. L.; LIEBERMAN-AIDEN, E.; WILLIAMS, L.; IMAKAEV, M.; GNIRKE, A.; MIRNY, L. A.; DEKKER, J.: LANDER, E. S. 'Hi-C: A method to study the Three-dimensional architecture of Genomes'**. J Vis Exp**. v. 39, 2010.

VAN DER CRUYSSEN B.; NUYTINCK L.; BOULLART L.; *et al.* Polymorphisms in the ficolin 1 gene (FCN1) are associated with susceptibility to the development of rheumatoid arthritis. **Rheumatology (Oxford),** v.46, n.12, p.1792-1795, 2007.

WARD, L.D.; KELLIS, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. **Nucleic Acids Res**, v. 40, p. 930-4, 2012.

WESTRA, H.; FRANKE, L.; *et al*. Systematic identification of trans eQTLs as putative drivers of known disease associations. **Nature**, v. 45, p. 1238-1245, 2013.

WICKHAM, H. **ggplot2: elegant graphics for data analysis.** New York. Springer Publishing Company. v. 1, p. 1-182, 2009.

XIE, D.; BOYLE, AP.; WU, L.; ZHAI, J.; KAWLI, T.; SNYDER, M. Dynamic trans-acting factor colocalization in human cells. **Cell**, v. 155(3), p.713-24, 2013.

ULITSKY, I.; BARTEL. D.P. lincRNAs: Genomics, Evolution, and Mechanisms. **Cell**. v. 154, p. 26–46, 2013.

ZHAN, L.; ZHANG, X.L. Binding of L-ficolin to HCV glycoprotein E1 stimulates opsonophagocytosis to targets cells expressed E1 by macrophages. **Molecular Immunology**, v. 44, n.1-3, p. 262, 2007.

1000 GENOMES PROJECT CONSORTIUM. A map of human genome variation from population-scale sequencing. **Nature**. v.467, p. 1061-73, 2010.

1000 GENOMES PROJECT CONSORTIUM; AUTON, A.; BROOKS, L.D.; DURBIN, R.M.; GARRISON, E.P.; KANG, H.M.; KORBEL, J.O.; MARCHINI, J.L.; MCCARTHY, S.; MCVEAN,

G.A.; ABECASIS, G.R. A global reference for human genetic variation. **Nature**. v. 526, n. 7571, p. 68-74, 2015.

APÊNDICE

APÊNDICE A – Tabelas referentes às análises para *FCN1*, contendo: lista de eQTLs; eQTLs compartilhados entre *FCN1* e *OLFM1*; escores Z dos eQTLs e de SNPs em DL; médias de escores Z. Disponíveis online em: http://preview.tinyurl.com/TABELAS-FCN1

APÊNDICE B – Tabelas referentes às análises para *FCN2*, contendo: lista de eQTLs; escores Z dos eQTLs e dos SNPs em DL. Disponíveis em: http://tinyurl.com/TABELAS-FCN2

APÊNDICE C – Tabelas referentes às análises para *FCN3*, contendo: lista de eQTLs; escores Z dos eQTLs e de SNPs em DL. Disponíveis em: http://tinyurl.com/TABELAS-FCN3

APÊNDICE D - Tabelas referentes às análises para *MBL2*, contendo: lista de eQTLs; escores dos eQTLs e de SNPs em DL. Disponíveis em: http://tinyurl.com/TABELAS-MBL2