UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE – UERN
CAMPUS AVANÇADO DE NATAL
CURSO DE CIÊNCIA DA COMPUTAÇÃO

VANDECLÉCIO LIRA DA SILVA

ANÁLISE EVOLUTIVA DOS ELEMENTOS REGULATÓRIOS QUE CONTROLAM O PADRÃO DE EXPRESSÃO DOS GENES HUMANOS

NATAL
2013

Vandeclécio Lira da Silva

Análise evolutiva dos elementos regulatórios que controlam o padrão de expressão dos genes humanos

Monografia apresentada à Universidade do Estado do Rio Grande do Norte – UERN - como requisito obrigatório para obtenção do título de Bacharel em Ciência da Computação.

ORIENTADOR: Wilfredo Blanco Figuerola

NATAL
2013

VANDECLÉCIO LIRA DA SILVA

Análise evolutiva dos elementos regulatórios que controlam o padrão de expressão dos genes humanos

Monografia apresentada à Universidade do Estado do Rio Grande do Norte – UERN - como requisito obrigatório para obtenção do título de Bacharel em Ciência da Computação.

Aprovado em ____/____/____.

Banca Examinadora

_____
Wilfredo Blanco Figuerola
UERN

_____
Sandro Jose de Souza
UFRN

_____
Adriana Takahashi
UERN

AGRADECIMENTOS

Primeiramente, gostaria de agradecer a Deus pela realização deste trabalho, pela força e sabedoria a mim concedida.

Aos meus pais, Valdetário e Célia, pela paciência, dedicação, compreensão e confiança. Ao meu irmão Vanderson e sua família, Renata e Marília, sempre me apoiando.

Aos meus orientadores Wilfredo Blanco e Sandro J. de Souza, por toda sabedoria transmitida e paciência que tiveram comigo.

A todos os colegas e amigos que compõem a UERN. Ao Instituto de Bioinformática e Biotecnologia – I2Bio, pelo apoio financeiro, a todos que fazem parte dele, em especial ao Jorge E. S. de Souza, pela sua ajuda e contribuições neste trabalho.

Ao ICe – UFRN e a todos que o compõem, em especial ao grupo de Bioinformática: Carol Corado, André Fonseca, Thayná Emília, IsabellaTanus e José E. Kroll.

E a todos que contribuíram diretamente e indiretamente para a realização deste trabalho, estes são meus agradecimentos.

Se algum de vocês tem falta de sabedoria,
peça-a a Deus, que a todos dá livremente,
de boa vontade; e lhe será concedida.

Tiago 1:5

RESUMO

Genômica e Bioinformática representam hoje os pilares em nossos esforços para compreender a biologia. Bioinformática, em particular, tem tirado proveito do desenvolvimento das tecnologias da informação para romper as barreiras da genética. O Projeto ENCODE, um consórcio internacional com o objetivo de identificar todos os elementos regulatórios no genoma humano, é um exemplo de como as tecnologias computacionais ajudam nas pesquisas genéticas nos dias atuais. Este trabalho, tem como objetivo utilizar ferramentas de genômica comparativa para estudar a evolução dos elementos reguladores identificados pelo projeto ENCODE, especialmente no contexto de elementos originados na linhagem humana. Os dados de alinhamento do genoma completo do humano/chimpanzé e do humano/gorila, obtidos a partir do UCSC Genome Browser, foram analisados para identificar sequências presentes apenas no genoma humano, representados por lacunas em ambos alinhamentos humano/chimpanzé e humano/gorila. Para realizar esta tarefa, foram desenvolvidos scripts escritos na linguagem Perl. Após identificadas as sequências presentes apenas no genoma humano, foram então pesquisados os elementos regulatórios do projeto ENCODE que estão presentes nestas sequências. Estes elementos regulatórios específicos de humano foram analisados no contexto dos seus genes regulados. Encontramos vários elementos regulatórios sendo enriquecidos num conjunto de sequências específicas de humano. Análises ontológicas nos permitiu concluir que um grupo de genes claramente associados com o sistema nervoso central, são reguladas por uma série de elementos de regulação que possivelmente foram originadas na linhagem humana. Os resultados obtidos a partir de todas estas análises permitiu descrever um cenário de evolução, levando em consideração a origem de elementos regulatórios específicos de humano.

Palavras-chave: Elementos regulatórios. ENCODE. Fatores de transcrição. DNA. Genes. Evolução.

ABSTRACT

Genomics and Bioinformatics represent today pillars in our efforts to understand Biology. Bioinformatics, in special, has capitalized on the development of information technologies to push the edge of genetics. The ENCODE Project, an international consortium with the goal to identify all the regulatory elements in the human genome, is an example of how computer technologies help genetics research in the present time. In this work, our goal is to use tools from comparative genomics to study the evolution of the regulatory elements identified by the ENCODE project, especially in the context of elements originated in the human lineage. Data from the human/chimpanzee and human/gorilla whole genome alignments (obtained from the UCSC genome browser) were analyzed to identify sequences only present in the human genome (represented by *gaps* in both human/chimpanzee and human/gorilla alignments). To carry out this task, we develop scripts written in Perl language. The sequences only present in the human genome were then searched for the presence of regulatory elements from the ENCODE project. These human-specific regulatory elements were analyzed in the context of their regulated genes. We found several regulatory elements being enriched in the set of human-specific sequences. Ontology analysis allowed us to conclude that group of genes clearly associated with the central nervous system presented a series of regulatory elements that were originated in the human lineage. The results obtained from all these analyses allowed us to depict an evolutionary scenario taking into consideration the origin of human-specific regulatory elements.

Keywords: Regulatory elements. ENCODE. Transcription factory. DNA. Genes. Evolution.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

BLAST	Basic Local Alignment Search Tool
ChIP-Seq	Chomatin Immunoprecipitation Sequencing
DBI	DataBase Interface
DD	Dicionário De Dados
DDBJ	DNA Data Bank Of Japan
DFD	Diagrama De Fluxo De Dados
DNA	Deoxyribonucleic acid NHGRI
EBI	European Bioinformatic Institute
ENCODE	ENCyclopedia Of DNA Elements
FT	Fator de Transcrição
I2Bio	Instituto de Bioinformática e Biotecnologia
NCBI	National Center of Biotechnology Information
RefSeq	Reference Sequence
RNA	Ribonucleic acid
SGBD	Sistema de Gerenciamento de Banco de Dados
SLFT	Sítio de Ligação dos Fatores de Transcrição
UCSC	University of California Santa Cruz

SUMÁRIO

1 INTRODUÇÃO

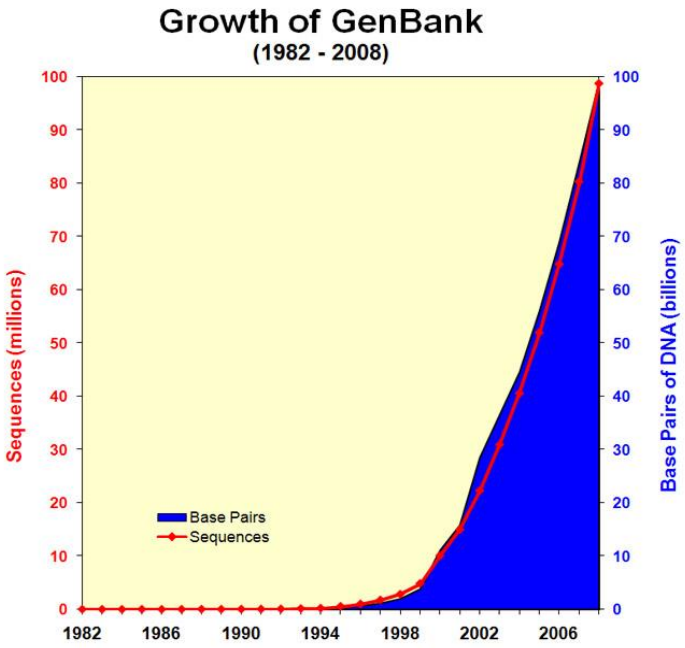
Atualmente, a biologia faz uso de muitas ferramentas computacionais para auxiliar em suas atividades, principalmente a área da bioinformática, que utiliza métodos e ferramentas da computação, matemática e estatística no desenvolvimento e análise de dados provenientes das suas pesquisas.

Podemos dizer que a bioinformática surgiu no momento que iniciou a utilização de ferramentas computacionais para realizar análise de dados genéticos. Com o desenvolvimento das tecnologias computacionais (hardware e software) nas últimas décadas, a bioinformática vem produzindo uma enorme quantidade de dados que precisam ser analisados. Antigamente, a análise destes dados era realizada manualmente pelos geneticistas, mas isso já se tornou uma tarefa muito exaustiva, sendo necessário cada vez mais, a utilização da computação para auxilia-los em seus trabalhos e também a necessidade de profissionais com um perfil capaz de utilizar e desenvolver ferramentas para analisar esses dados.

Um bom exemplo deste cenário de acumulo de dados, é o crescimento exponencial do (GenBank)[1] nas últimas décadas, como mostrado na Figura 1. Nele está contido milhares de dados, de diversas pesquisas em diversas áreas da biologia, disponibilizadas de forma gratuita.

[1] Banco de dados de sequências genéticas, administrada pelo National Center of Biotechnology Information (NCBI).

Figura 1: Crescimento do GenBank



Fonte: ("GenBank Statistics", 2008)

Nas últimas décadas, vem sendo realizado vários projetos voltados para o estudo do Ácido Desoxirribonucleico (ADN), comumente conhecido como *Deoxyribonucleic Acid* (DNA), um destes projetos realizado foi o *Encyclopedia Of DNA Elements* (ENCODE). O ENCODE foi um consórcio do *National Genome Research Institute* (NHGRI), criado em 2003, com o objetivo principal de identificar todos os elementos regulatórios presentes no genoma humano.

Nosso planeta é habitado por uma grande diversidade de organismos vivos, o motivo dessa grande diversidade é atribuída ao processo de evolução. A evolução é um processo natural que faz com que as populações de organismos se adaptem e se diversifiquem ao longo do tempo. O DNA de um organismo também sofre estas alterações ao longo do tempo, contribuindo ainda mais para o processo de evolução.

A evolução nos conta que o homem e o chimpanzé divergiram de um ancestral comum, sendo assim, compartilhamos o mesmo DNA que sofreu alteração ao longo do tempo. O nosso DNA é semelhante até mesmo com espécies filogeneticamente mais distante como o gorila. Mas muito ainda precisa ser descoberto de como nosso DNA evoluiu.

Estudar o DNA pode nos ajudar a entender como os seres humanos evoluíram. Os elementos regulatórios são objetos muito importantes neste processo de evolução do DNA, pois é com eles que podemos saber como nosso DNA funciona.

1.1 OBJETIVOS

Os objetivos do presente trabalho foram, realizar uma análise evolutiva dos elementos regulatórios do genoma humano, Avaliando a presença e ausência destes elementos regulatórios em espécies próximas, como o chimpanzé e o gorila. Afim de descobrir quais elementos regulatórios originaram-se na espécie humana.

O segundo objetivo é identificar os genes que são controlados pelos elementos regulatórios que possivelmente surgiram na espécie humana e correlaciona-los com o padrão de expressão do genes humano. Realizar análises de expressão e de ontologia para avaliar se esses genes estão relacionados a alguma característica adquirida durante a evolução da espécie humana.

1.2 ORGANIZAÇÃO

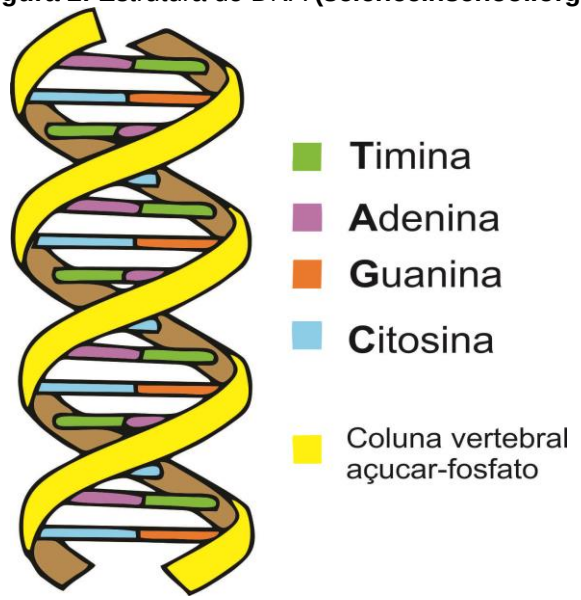
O presente trabalho está dividido em sete capítulos. No primeiro capítulo apresenta-se uma breve contextualização do projeto, expondo a problemática, os objetivos e os meios utilizados pra se alcançar os objetivos. Tendo em vista que área abordada não é comumente estudada na área da computação. No segundo capítulo é estudado uma contextualização da área na qual o trabalho foi desenvolvido, dando ênfase a conceitos da biologia. Ao entrar no terceiro capítulo, focamos a área da bioinformática, descrevendo métodos e ferramentas comumente utilizadas. No quarto capítulo, é realizada uma revisão da literatura, apresentando trabalhos relacionados a bioinformática, banco de dados e mais especificamente estudos evolutivos do genoma humano. A metodologia utilizada para chegar aos resultados finais é abordada no quinto capítulo, onde é descrito os *scripts* que foram produzidos, os dados que foram analisados e os dados que foram gerados. No sexto capítulo está descrito os

resultados que obteve-se através das análises computacionais e a interpretação destes resultados em um contexto biológico. E por fim, o sétimo capítulo contém as considerações finais e os trabalhos futuros.

2 BIOLOGIA MOLECULAR

Uma das coisas que torna um indivíduo único é o seu DNA, nele está codificado toda a informação genética que é passado entre as gerações. O DNA em seu formato de dupla hélice é formado por quatro bases de nucleotídeos agrupados em pares (Guanina – Citosina, Adenina – Timina)[1] e moléculas de açúcares e fosfato, ilustrado na Figura 2. As informações genéticas contidas no DNA, são responsáveis pela construção do Ácido Ribonucleico (ARN), mais conhecido como *Ribonuclic Acid* (RNA) e das proteínas. As proteínas, são responsáveis pela formação da maior parte da estrutura e atividade dos organismos, como por exemplo, musculo, cabelos, enzimas, entre outros.

Figura 2: Estrutura do DNA **(scienceinschool.org)**[2]



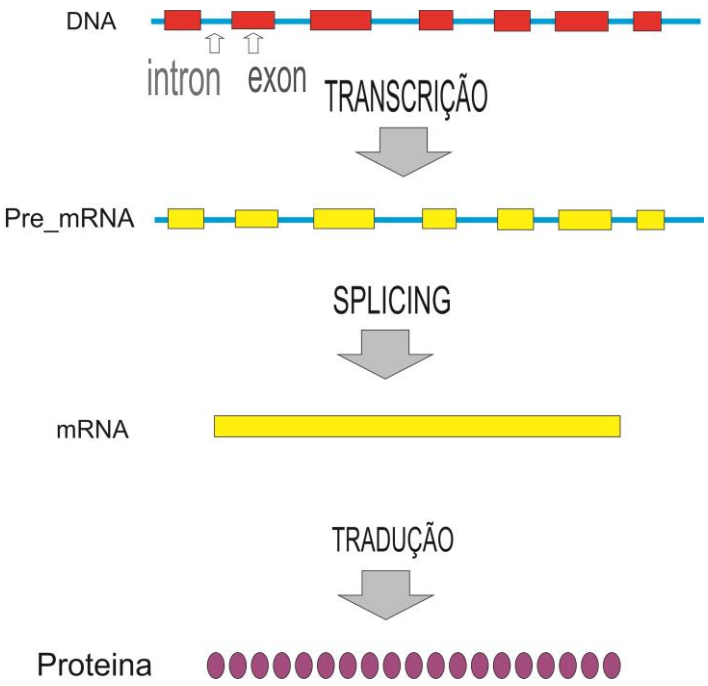
O dogma central da biologia molecular está baseado na interação de três agentes, o DNA, o RNA e as proteínas, ver Figura 3. Como já falamos no DNA está contido as informação genéticas de um indivíduo, mas nem todas as partes do DNA codificam informação. O DNA é formado por regiões codificadoras de informações,

[1] Bases de nucleotídeos representados pelas letras, G, C, A, T, respectivamente.
[2] Fonte original da figura.

chamadas de exons e por regiões não-codificadoras, chamadas de introns.

A informação genética que codifica uma determinada proteína ou característica de um indivíduo, está contida num seguimento de DNA chamado de gene. Para o DNA se transformar em RNA, ele passa por um processo chamado de transcrição. Nos seres eucariontes[3], este processo ocorre com a separação dos introns do DNA, através de um outro processo chamado de *splicing*, que separa os introns dos exons. Depois de separados, os exons se juntam e formam o RNA mensageiro maduro (mRNA), que depois é traduzida em uma proteína, Figura 3.

Figura 3: Dogma central da biologia molecular, processo de transformação do DNA em proteína



O fato dos introns não conterem informações codificadoras, isso não significa que eles não têm importância, pelo contrário, como revelado pelo projeto ENCODE, os introns exercem uma função muito importante, pois em algumas posições específica deles, estão contidos elementos responsáveis pela regulação gênica.

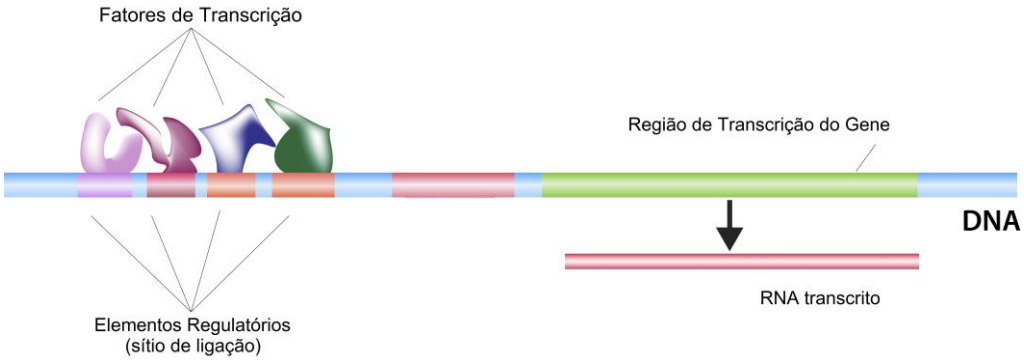
[3] Designação dada a organismos compostos por células que tem o núcleo envolvido por uma membrana chamada de carioteca.

2.1 REGULAÇÃO GÊNICA

Como vimos, os genes é um seguimento do DNA onde está codificada as informações genéticas de um indivíduo e para que esses genes sejam expressos ou inibidos, eles passam por um processo chamado de regulação gênica.

A regulação gênica é um processo que regula a ativação e desativação dos genes humano, regulando a intensidade que os genes serão expressos. Este processo ocorre quando proteínas chamadas de Fatores de Transcrição (FTs) que ligam-se ao DNA em posições especificas, muitas vezes localizadas próximas a segmentos de DNA que codificam proteínas (LESK; ANDRADE, 2008). Estas regiões do DNA são chamadas de Sítios de Ligação dos Fatores de Transcrição (SLFT), elas contêm sequências sinalizadoras que servem de sítios de ligação para os FTs, que podem atuar regulando a ativação ou inibição do mecanismo de transcrição que transforma o DNA em RNA (VARGAS, 2006), os sitos onde os FTs se ligam são chamados de elemento regulatórios. A Figura 4 ilustra mais claramente o processo da regulação gênica.

Figura 4: Processo de regulação gênica



3 BIOINFORMÁTICA

É difícil definir o que é a bioinformática, pois ela envolve várias linhas de conhecimento. A bioinformática é formada pela união entre biologia e ciência da computação (LESK; ANDRADE, 2008), mas ela também utiliza fundamentos de outras áreas, como matemática e estatística.

A bioinformática está sempre crescendo em paralelo com a computação. À medida em que as tecnologias computacionais avançam, mais ferramentas são criadas para serem utilizadas pela bioinformática. Com o surgimento dos sequenciadores automáticos de DNA, na segunda metade da década de 90, houve uma explosão na quantidade de sequências a serem armazenadas (PROSDOCINI et al., 2002). A computação contribuiu não somente com o armazenamento dos dados e com a capacidade de processamento, mas também com métodos matemáticos sofisticados necessários para se obter resultados (LESK; ANDRADE, 2008).

Em 1975, métodos de sequenciamento de DNA foram desenvolvidos por F.Sanger e independentemente por A. Maxam e W. Gilbert, (LESK; ANDRADE, 2008). Os sequenciadores automáticos de DNA tiveram uma grande contribuição para o avanço da bioinformática. O primeiro aparelho para sequenciamento de DNA foi produzido pela empresa, Applied Biosystems e foi chamado de ABI. Ela utiliza o método de Sanger e tem a capacidade de sequenciar aproximadamente 40 KB de DNA por corrida, KB equivale a mil bases nitrogenadas, e tinha um custo de 0.5KB/US$[1]. Com o passar dos anos novos aparelhos da nova geração dos sequenciadores foram criados, como o 454-Roche, que é capaz de sequenciar 0,5 GB por corrida como o custo de 50KB/UR$[2], A Solexa – Ilumina capaz de sequenciar aproximadamente 90 GB por corrida, com um custo de 9MB/UR$ e o SOLID – ABI capaz de sequenciar aproximadamente 100GB de DNA por corrida, com um custo de 10MB/UR$. Podemos observar que quanto mais moderno os sequenciadores são maior é a quantidade de dados que eles podem produzir.

Além do sequenciamento do genoma humano, também foram sequenciados

[1] US$ - unidade monetária para Dólar americano.
[2] UR$ - unidade monetária para Peso uruguaio.

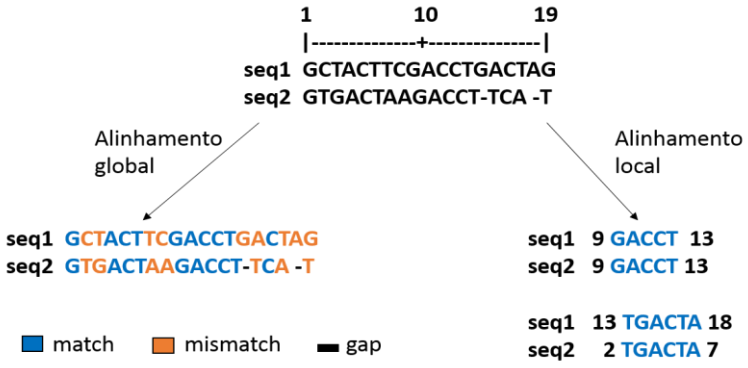
genomas de várias outras espécies de animais, como primatas, peixes, insetos, plantas, bactérias, dente outras espécies.

3.1 ALINHAMENTO

Depois de sequenciado o genoma de várias espécies de organismos, algo interessante realizado, foi o alinhamento dessas sequências de DNA. O alinhamento de sequências possui uma diversidade de aplicações na bioinformática, sendo considerada uma das operações mais importantes desta área (PROSDOCINI et al., 2002). O alinhamento de sequências entre espécies nada mais é que a comparação entre sequências de DNA ou Proteínas de organismos da mesma espécie ou de espécies distintas, assim identificando o nível de similaridade destas sequências (FASSLER; COOPER, 2008). A partir do alinhamento de sequências foi possível descobrir que o DNA do chimpanzé é 99% similar ao do homem (WILDMAN, 2003).

Existem dois tipos de alinhamento, o alinhamento global e o local. O alinhamento global, realiza a comparação entre duas sequências de DNAs ou proteínas ao longo do toda a sua extensão, quando o alinhamento encontra similaridade entre as bases nucleotídicas, dar-se o nome de *match*, quando são diferentes são chamados e *mismatch* e quando ocorre alguma falha chama-se de *gaps*. O alinhamento local, procura similaridades entre duas sequências, mas não ao longo de sua extensão, e sim, através de pequenos pedaços localmente. O principal programa utilizado para alinhamento global é a Ferramenta de Busca por Alinhamento Local Básico, do inglês, *Basic Local Alignment Search Tool* (BLAST), a Figura 5 ilustra mais claramente o funcionamento dos alinhamentos global e local.

Figura 5: Alinhamento global e local



3.2 BANCO DE DADOS

Como vimos, estão sendo gerados muitos dados oriundos das pesquisas com bioinformática, então é necessário armazenar esses dados de uma forma segura e organizada, para isso, utiliza-se os bancos de dados. O grande investimento do setor público nas pesquisas em bioinformática, proporcionou a criação dos grandes bancos de dados públicos, dentre os quais podemos citar: GeneBank, o banco de dados de sequências de DNA e Proteinas, o banco de dados europeu de sequências de DNA – *European Bioinformatics Institute* (EBI) e o banco de dados japonês de sequências de DNA – *DNA Data Bank of Japan* (DDBJ). Os três bancos de dados trocam informação entre si diariamente, mantendo seus dados sempre atualizados (BENSON et al., 2012).

Existem basicamente dois tipos de bancos de sequências, os bancos de dados primários e os secundários. Os bancos de dados primários são formados por resultados de pesquisas publicados com alguma interpretação, mas sem que haja uma análise por trabalhos anteriores, ocorrendo a redundância de dados, como por exemplo o GenBank. Já os bancos secundários, se preocupam com a redundância dos dados, evitando a redundância de suas informações, como por exemplo o banco de dados de Sequência de Referencia – *Reference Sequence* (RefSeq).

3.3 PERL

O desenvolvimento dos scripts para nossas análises, será feito utilizando a linguagem Perl, por ela ser amplamente usada na área da bioinformática. O Perl é uma linguagem de programação estável e multiplataforma (Perl, 2001). A popularidade do uso do Perl pelos bioinformatas, deve-se ao fato da linguagem possuir funções muito eficientes para a manipulação de textos, processamento de cadeias (*strings*) e na utilização de expressões regulares. Esta facilidade de manipulação de strings é muito útil para os usuários nesta área, tendo em vista que, os arquivos utilizados por eles, são arquivos de strings, principalmente os arquivos de alinhamento (EDWARDS, 2009).

4 TRABALHOS RELACIONADOS

Hoje em dia sabemos que a informação genética de um indivíduo não está apenas associada a região codificadora do DNA, mas também estão relacionadas com as regiões não codificadoras, pois nelas existem elementos que controlam o mecanismo de transcrição (CONSORTIUM, 2012).

O maior projeto que trabalhou diretamente com os elementos regulatórios foi o projeto ENCODE, que tinha como objetivo principal encontrar todos os elementos regulatórios presentes no genoma humano.

Entre as contribuições do projeto ENCODE encontramos o mapeamento de regiões de transcrição e os sítios de ligação dos fatores de transcrição (CONSORTIUM, 2012). Para fazer isto, utilizou alguns métodos bioquímicos, como por exemplo, o método de Chomatin Immunoprecipitation Sequencing (ChIP-Seq), utilizado para descobrir os SLFT (LANDT et al., 2012). Em geral, o projeto revela uma nova perspectiva sobre a organização e regulação dos genes e do genoma humano, mostrando que grande parte do DNA que antes pensava-se não ter importância, na realidade é responsável pela regulação gênica, controlando o mecanismo de transcrição do DNA, através de suas regiões regulatórias.

Com os resultados do projeto ENCODE, vários trabalhos vêm sendo desenvolvidos utilizando seus resultados, como por exemplo o desenvolvimento de uma nova abordagem de um banco de dados, o RegulomeDB, que orienta a interpretação das variações de regulação do genoma humano (BOYLE et al., 2012).

Apesar de já se saber a décadas que os elementos regulatórios têm um papel importante na evolução humana, poucos trabalhos foram realizados nesta área, mas este cenário vem mudando. Encontramos alguns trabalhos como o (ARBIZA et al., 2013), que utiliza sequência de todo genoma e dados de ChIP-Seq do projeto ENCODE, para demostrar que a seleção natural tem profunda influência nos sítios de ligação dos fatores de transcrição desde a divergência do homem com o chimpanzé, utilizando um novo método probabilístico, chamado de INSIGHT, para medir a influência da seleção dos elementos regulatórios.

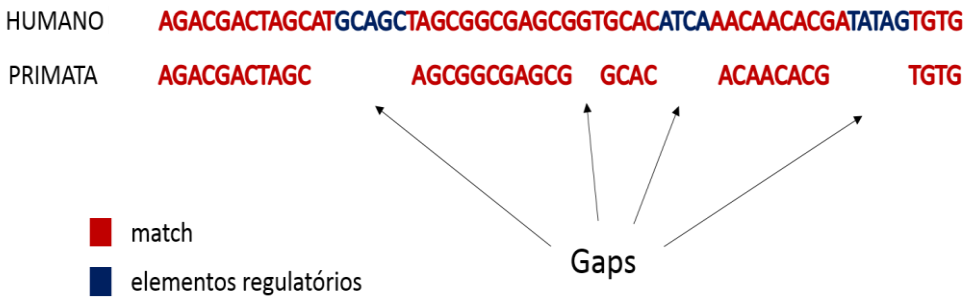
Estudos utilizando uma espécie de mosca, *Drosophila*, são comumente realizados pelos biólogos, devido a facilidade de manuseio, rápida reprodução e fácil criação. Hoje em dia o genoma da Drosophila é utilizado como modelo para realização de estudos biológicos, como por exemplo, a investigação da variabilidade dos SLFTs presentes no genoma humano e na linhagem da *Drosophila* (SPIVAKOV et al., 2012).

A evolução humana é um tema que fascina muitos pesquisadores, e várias pesquisas são desenvolvidas sobre esse tema, como por exemplo o desenvolvimento de um *framework* de comparação genômica para alinhar genomas filogeneticamente distantes e detectar de forma abrangente elementos não gênicos conservados (HILLER et al., 2013).

5 METODOLOGIA

O primeiro passo tomado para realizar a análise evolutiva dos elementos regulatórios, foi identificar os seguimentos de DNA existentes apenas no genoma humano e procurar dentro desses seguimentos a existência de elementos regulatórios, para isso, foi utilizado dados do projeto ENCODE e dados de alinhamento entre as espécies humana, chimpanzé e gorila. Primeiro, foi procurado no arquivo de alinhamento, regiões do DNA humano que não se alinhavam com a do chimpanzé, na qual foi chamado de "*gap*", o arquivo de alinhamento está descrito na próxima subseção, este arquivo contém informações sobre o tamanho dos *gaps*, então quando encontrado um *gap* com tamanho superior a 20 bases nucleotídicas (b)[1] é verificado nos dados do ENCODE se existe algum elemento regulatório, nesta região do DNA humano, ilustrado na Figura 6. Esta mesma análise é realizada utilizando os dados de alinhamento do gorila.

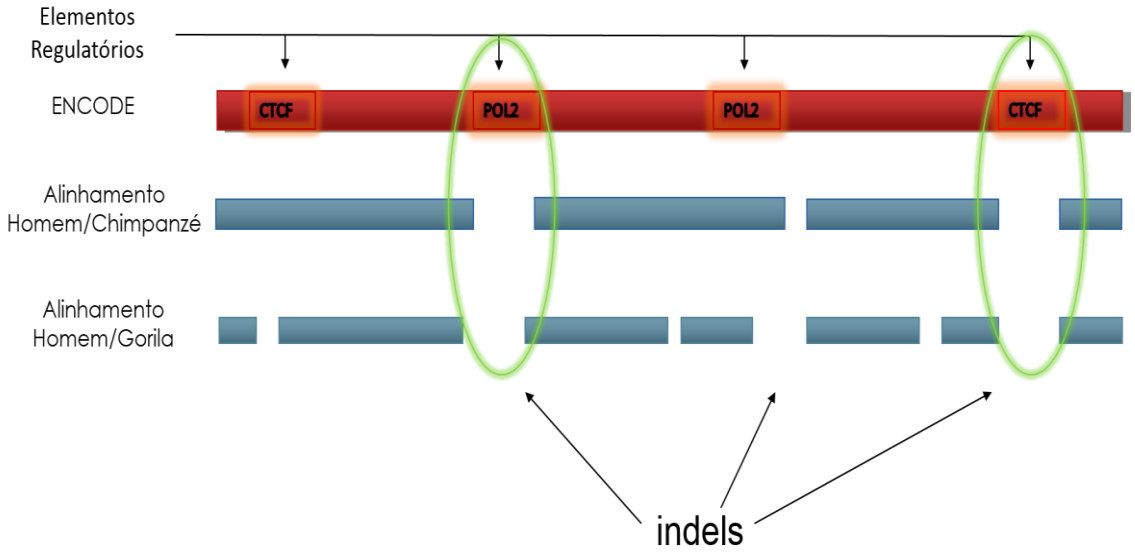
Figura 6: Busca por elementos regulatórios dentro dos *gaps*



A segunda etapa, foi realizar a comparação entre os resultados da análise do chimpanzé, com os resultados da análise do gorila, com objetivo de encontrar SLFT que não existissem em ambas as espécies de primata. Isto nos mostraria que se no DNA da espécie humana existem determinados SLFT, mas não existe no chimpanzé nem no gorila, isto nos leva a inferir que ao longo da evolução, naquela região do DNA humano, ou do chimpanzé ou do gorila, ocorre uma inserção ou deleção de sequências de DNA, chamados de *indel* na biologia molecular, ver Figura 7.

[1] "b" é uma unidade de medida utilizada para contar a quantidade de bases de nucleotídeos (A, T, G, C).

Figura 7: Identificação dos *indels* com presença de elementos regulatórios



Depois de mapeados os SLFT que foram encontrados apenas na espécie humana, iniciou a identificação dos genes que estão próximos a estes SLFT. Para cada SLFT que foi identificado, é observado se na distância de 2.000B há existência de algum gene, se existir, as informações dos genes, descritos na subseção a seguir, são gravados em arquivos de texto. A busca pelos genes é realizada em ambas as fitas do DNA afim de identificar todos os genes. Foi utilizado uma base de dados do RefSeq, para identificar a localização dos genes. Após identificados os genes, armazenamos estas informações em um banco de dados para manipulá-los melhor.

O objetivo de identificar os genes que são regulados pelos elementos regulatórios identificados apenas no humano, foi para que pudéssemos avaliar a existência de alguma característica genética que seja predominante da espécie humana ou que tenha contribuído para processo de evolução. Para isso, foi feito dois tipos de análises biológicas sobre estes genes: expressão e ontologia.

O primeiro tipo de análise é a verificação do padrão de expressão desses genes. A expressão de um gene é um valor quantitativo, utilizado para medir o quanto um gene está sendo ativado em um determinado lugar e situação. O padrão de expressão define o quanto um determinado gene está sendo expresso num determinado tecido em uma situação normal. Para verificar o padrão de expressão dos genes humano foi utilizado um banco de dados com essas informações, disponibilizado pelo Instituto de Bioinformática e Biotecnologia (I2Bio).

O segundo tipo de análise realizado foi da ontologia dos genes. A ontologia de um gene representa o domínio na qual os (produtos de um gene)[2] pertence. A ontologia pode ser dividida em três domínios: componente celular, função molecular e processo biológico. (STEVENS, 2000).

Para realizar a análise da ontologia dos genes, foi utilizado um programa desenvolvido por Jorge E. S. de Souza, pesquisador principal do I2Bio. O programa verifica a ontologia de uma lista de genes passados como dado de entrada, classificando os produtos dos genes de acordo com seu domínio e atribuindo um "valor-p" aos produtos dos genes com maior significância. O valor-p é o valor probabilidade dos resultados da ontologia terem ocorrido ao acaso dentro de uma Distribuição Normal, então quanto menor o valor-p mais significativos serão os resultados.

5.1 DADOS DE ENTRADA

Para dar-se início as atividades, era necessário primeiramente obter os dados de entrada a serem analisados e saber como eles estavam estruturados. Os dados de entrada utilizados foram, os dados de alinhamento do genoma humano com o do chimpanzé, dados de alinhamento do genoma humano com o do gorila e os dados do projeto ENCODE relacionados aos elementos regulatórios. Todos os dados de entrada foram obtidos através do site da *University of California Santa Cruz* (UCSC) ("UCSC Genome Browser Home", 2013).

Os dados de alinhamento estão contidos em um arquivo no formato ".chain" que descreve as posições do alinhamento que foram emparelhados e também os *gaps* em ambas as sequências das espécies. A descrição do arquivo pode ser encontrada na página do UCSC (UCSC, 2013).

A versões utilizadas dos alinhamentos[3] do chimpanzé foi a "panTro4" e alinhamento do gorila foi a "gorGor3", disponível em (UCSC, 2013).

---

[2] Material bioquímico, RNA ou Proteína, resultante da expressão de um gene.
[3] Os arquivos de alinhamento são descritos no DFD da próxima subseção como (Alinhamento Homem-Chimpanzé/Homem-Gorila).

Para encontrar a localização dos SLFT, foi utilizado os dados do projeto ENCODE. Através do site do UCSC (“UCSC Genome Browser Home”, 2013) foi obtido o arquivo wgEncodeRegTfbsClusteredV2.bed.gz[4], onde, nele está contido dados dos elementos regulatórios identificados pelo projeto ENCODE através do método ChIP-Seq.

Outro tipo de dado utilizado, foram os dados de RefSeq, para identificar quais genes estão próximos dos elementos regulatórios encontrados apenas na espécie humana. Esta base de dados foi obtida em colaboração com o I2Bio, disponibilizados em seu cluster em Ribeirão Preto, que utiliza o Sistema de Gerenciamento de Banco de Dados (SGBD), MySQL.

5.2 MODELAGEM

A análise dos dados foi realizada a partir do desenvolvimento de *scripts* utilizando a linguagem Perl. A linguagem Perl dá suporte a vários paradigmas de programação, na qual, no presente trabalho foi optado por utilizar o paradigma estruturado.

O desenvolvimento de um software inicia-se bem antes da programação, para um software chegar a um produto final é necessário passar por algumas etapas antes do desenvolvimento, umas delas é a modelagem. Antes do iniciar a modelagem foi realizado um levantamento de requisitos, que se baseou no aprofundamento no conhecimento em biologia molecular e em bioinformática.

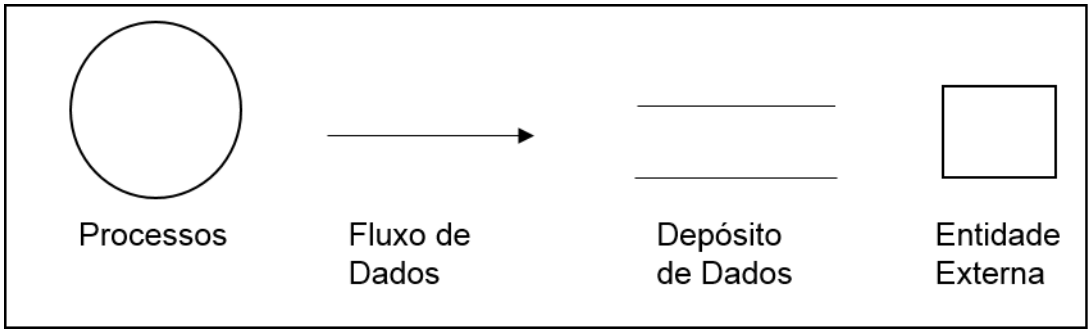
A abordagem estruturada utiliza uma metodologia top-down e a divisão do sistema em módulos ou submódulos, (REZENDE, 2006). Algumas técnicas de analise estruturada foram utilizadas, como o Diagrama de Fluxo de Dados (DFD), o Dicionário de Dados (DD) e a especificação de processos.

5.2.1 Diagramas de Fluxo de Dados

[4] Descrito no DFD como “Elementos Regulatórios (ENCODE)”.

Os DFDs são muito utilizados na modelação funcional da análise estruturada. Neles o sistema é modelado como uma rede de processos ou funções, ligados por fluxos de dados e depósitos de dados. Geralmente os DFDs são compostos por processos, fluxo de dados, depósitos de dados e entidades externas, representados na notação descrita na Figura 8. Os DFDs também são apresentados por níveis, que aprofundam quando há a necessidade de modelar os processos.

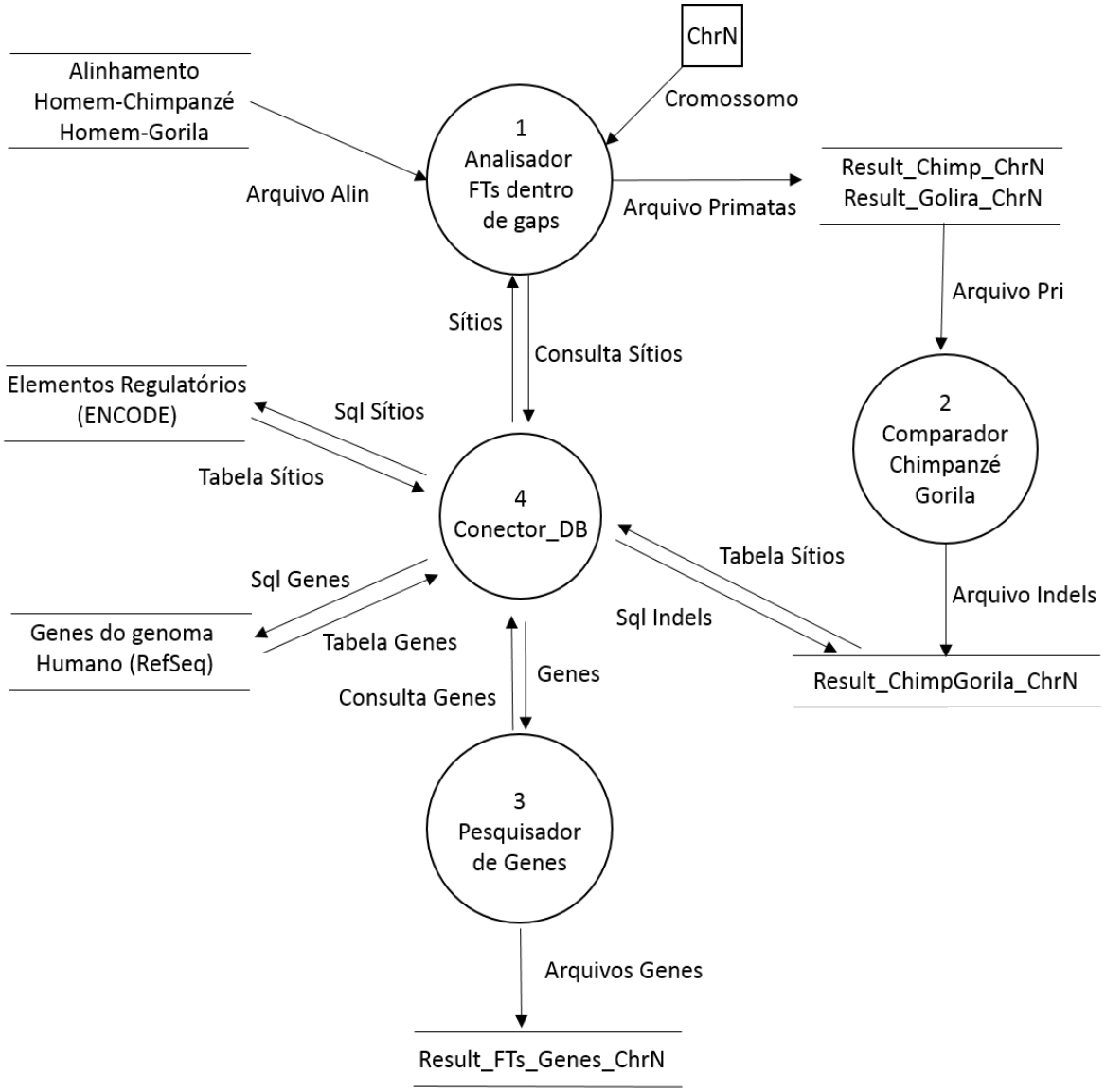
Figura 8: Notação dos componentes do DFD



Para a modelagem do sistema que realiza a identificação dos SLFT que estão presentes apenas na espécies humana, foram criados dois DFDs.

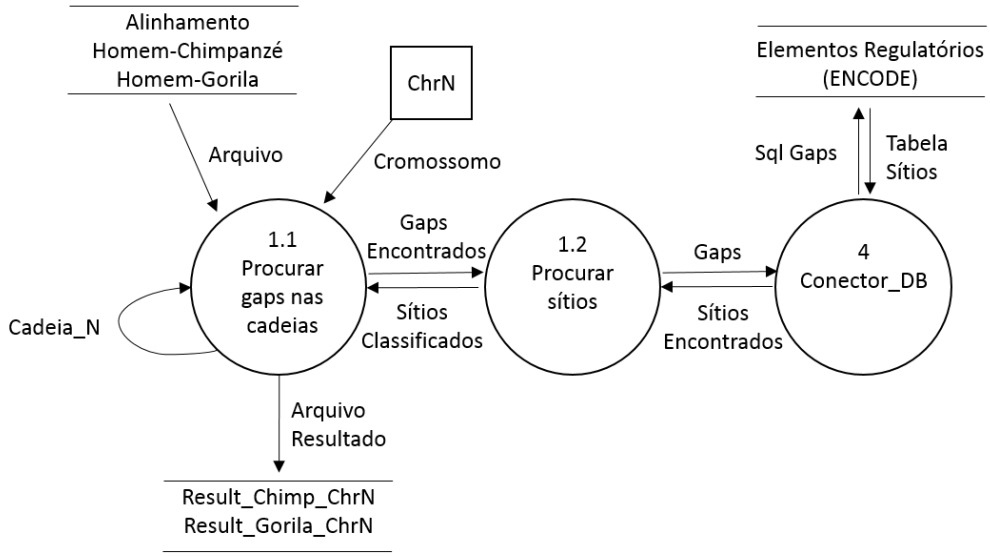
O primeiro para descrever o sistema como um todo mostrando a interação de cada Processo. Foi criado 3 processos (Analisador, Comparador, Pesquisador e o Conector_DB), que interage com 6 Depósitos de Dados (Alinhamento Homem-Chimpanzé/Homem-Gorila, Result_Chimp_ChrN/Result_Corila_ChrN, Elementos Regulatórios, Genes do genoma Humano, Result_ChimpGorila_ChrN, Result_FTs_Genes_ChrN) e todo sistema sofre ação de uma única entidade externa, o cromossomo que está sendo analisado (ChrN), ver Figura 9,

Figura 9: Diagrama de Fluxo de Dados de todo sistema



O segundo DFD está em um nível abaixo do primeiro, representa o Processo 1 Analisador, já que ele é mais complexo que os demais processos, este DFD é composto por 3 processos (Procurar *gaps*, Procurar sítios), 3 depósitos de dados (Alinhamento, Elementos Regulatórios e Result_Chimp_ChrN/Result_Gorila_ChrN) e um entidade externa ChrN, que interagem entre si através dos fluxos de dados, ver Figura 10.

Figura 10: Diagrama de Fluxo de Dados do Processo do Analisador



5.2.2 Dicionário de Dados

No DD está descrito todos os eventos do sistema, listando todas as entradas, saídas, componentes de depósitos de dados, cálculos e processos que ocorrem nos fluxos de dados.

5.2.2.1 Fluxo de Dados do DFD geral

Nome:	Cromossomo
Descrição:	Define o cromossomo que será analisado
Conteúdo:	Identificador do cromossomo
Origem:	Enviados por parâmetro
Destino:	Todos os processos do sistema

Nome:	Arquivo Alin
Descrição:	Abre o arquivo de alinhamento e o armazena em uma lista de dados
Conteúdo:	Nome do arquivo de alinhamento, podendo ser do chimpanzé ou do gorila
Origem:	Enviado por parâmetro na hora da execução do programa
Destino:	Processo 1 Analisador

Nome:	Consulta Sítios
Descrição:	Envia uma requisição sql para o banco de dados

Conteúdo:	Abertura de conexão, query com a posição de um gap
Origem:	Processo 1 Analisador
Destino:	Processo 4 Conector_DB

Nome:	Sítios
Descrição:	Retorna o resultado da busca no banco
Conteúdo:	Sitio de ligação dentro do intervalo do gap
Origem:	Processo 4 Conector_DB
Destino:	Processo 1 Analisador

Nome:	Arquivo Primatas
Descrição:	Grava os resultados no arquivo de saída
Conteúdo:	Listagem dos sítios de ligação presentes nos gaps
Origem:	Processo 1 Analisador
Destino:	Result_Chimp_chrN / Result_Gorila_chrN

Nome:	Arquivo Pri
Descrição:	Ler os arquivos de resultado com os sítios de ligação presentes no chimpanzé e no gorila
Conteúdo:	Nome do arquivo de resultados do chimpanzé e do gorila
Origem:	Enviados por parâmetro na hora da execução do script
Destino:	Processo 2 Comparador

Nome:	Arquivo Indels
Descrição:	Grava o resultado da comparação entre chimpanzé e gorila
Conteúdo:	Lista de indels contendo sítios de ligação encontrados apenas no genoma humano
Origem:	Processo 2 Comparador
Destino:	Result_ChimpGorila_chrN

Nome:	Sql Sítios
Descrição:	Faz uma busca no banco pelos sítios de ligação dentro do intervalo dos gaps
Conteúdo:	Intervalo do gap
Origem:	Processo 4 Conector_DB
Destino:	ENCODE

Nome:	Tabela Sítios
Descrição:	Retorna os sítios de ligação que estão dentro do intervalo do gap

Conteúdo:	Sítios de ligação
Origem:	ENCODE
Destino:	Processo 4 Conector_DB

Nome:	Sql Genes
Descrição:	Envia posição do sitio de ligação
Conteúdo:	Query procurando por genes
Origem:	Processo 4 Conector_DB
Destino:	RefSeq

Nome:	Tabela Genes
Descrição:	Retorna os genes que estão próximos a um determinado sítio de ligação
Conteúdo:	Genes encontrados próximos aos sítios de ligação
Origem:	RefSeq
Destino:	Processo 4 Conecto_DB

Nome:	Consulta Genes
Descrição:	Procura os genes próximos aos sítios de ligação
Conteúdo:	Posição dos sítios de ligação
Origem:	Processo 3 Pesquisador
Destino:	Processo 4 Conector_DB

Nome:	Genes
Descrição:	Retorna os genes que foram encontrados próximos aos sítios de ligação
Conteúdo:	Genes próximos aos sítios de ligação
Origem:	Processo 4 Conector_DB
Destino:	Processo 3 Pesquisador

Nome:	Arquivo Genes
Descrição:	Grava no arquivo de saída Result_FTs_Genes_chrN
Conteúdo:	Sítios de ligação com os genes que estão próximos
Origem:	Processo 3 Pesquisador
Destino:	Result_FTs_Genes_chrN

5.2.2.2 Deposito de Dados do DFD geral

Nome:	ChrN

Descrição:	Cromossomo alvo
Conteúdo:	ID do cromossomo
Fluxo de entrada	Enviado por parâmetro
Fluxo de saída	Todo o sistema

Nome:	Alinhamentos Homem-Chimpanzé / Homem-Gorila
Descrição:	Arquivo de alinhamento
Conteúdo:	Posições dos gaps
Fluxo de entrada	Processo 1 Analisador
Fluxo de saída	Processo 1 Analisador

Nome:	Result_Chimp_chrN / Result_Gorila_chrN
Descrição:	Arquivo de saída do Processo 1 Analisador
Conteúdo:	Sítios de ligação que estão contidos totalmente ou parcialmente nos gaps para cada cromossoma
Fluxo de entrada	Processo 1 Analisador
Fluxo de saída	Processo 2 Comparador

Nome:	ENCODE
Descrição:	Dados do projeto ENCODE
Conteúdo:	Elementos regulatórios identificados através do ChIPseq
Fluxo de entrada	Processo 4 Conector_DB
Fluxo de saída	Processo 4 Conector_DB

Nome:	RefSeq
Descrição:	Dados dos genes do genoma humano
Conteúdo:	Posição onde estão localizados os genes
Fluxo de entrada	Processo 4 Conector_DB
Fluxo de saída	Processo 4 Conector_DB

Nome:	Result_ChimpGorila_chrN
Descrição:	Resultado da comparação dos resultados do chimpanzé e do gorila
Conteúdo:	Sítios de ligação encontrados apenas na espécie humana
Fluxo de entrada	Processo 4 Conector_DB, Processo 2 Comparador
Fluxo de saída	Processo 4 Conector_DB

Nome:	Result_FTs_Genes_ChrN
Descrição:	Genes encontrados próximos aos sítios de ligação encontrados apenas na espécie humana

Conteúdo:	Posição dos sítios de ligação, posição dos genes
Fluxo de entrada	Processo 4 Conector_DB, Processo 3 Pesquisador
Fluxo de saída	Processo 4 Conector_DB

5.2.2.3 Fluxo de Dados do Processo 1 Analisador

Nome:	Cromossomo
Descrição:	Define o cromossomo que será analisado
Conteúdo:	Identificador do cromossomo
Origem:	Enviados por parâmetro;
Destino:	Todo os Processos do sistema

Nome:	Arquivo
Descrição:	Abre o arquivo de alinhamento e armazena em um array
Conteúdo:	Nome do arquivo de alinhamento, podendo ser do chimpanzé ou do gorila
Origem:	Enviado por parâmetro na hora da execução do programa
Destino:	Processo 1.1 procura gaps

Nome:	Cadeia_N
Descrição:	Passa o ID das cadeias mais conservadas
Conteúdo:	Identificador das cadeias
Origem:	Enviado por parâmetro na hora da execução do Processo
Destino:	Processo 1.1 Procura gaps

Nome:	Gaps Encontrados
Descrição:	Envia a posição do gap encontrado
Conteúdo:	Posição do genoma humano onde foi encontrado um gap
Origem:	Processo 1.1 Procura gaps
Destino:	Processo 1.2 Procura sítios

Nome:	Sítios Classificados
Descrição:	Retorna os sítios classificados pelo status
Conteúdo:	Lista com os sítios encontrados dentro do intervalo do gap
Origem:	Processo 1.2 Procura sítios
Destino:	Processo 1.1 Procura gaps

Nome:	Arquivo Resultado
Descrição:	Grava os resultados no arquivo de saída

Conteúdo:	Listagem dos sítios de ligação presentes nos gaps
Origem:	Processo 1.1 Procura gaps
Destino:	Result_Chimp_chrN / Result_Gorila_chrN

Nome:	Gaps
Descrição:	Abre a conexão com o banco de dados e envia a posição do gap através de querys
Conteúdo:	Posição do gap
Origem:	Processo 1.2 Procura sítios
Destino:	Conector_DB

Nome:	Sítios Encontrados
Descrição:	Retorna os sítios de ligação presentes no intervalo da posição do gap enviado
Conteúdo:	Listagem dos sítios de ligação presentes nos gaps
Origem:	Conector_DB
Destino:	Processo 1.2 Procura SLFT

Nome:	Sql Gaps
Descrição:	Faz uma busca no banco por SLFT dentro do intervalo do gap
Conteúdo:	Intervalo do gap
Origem:	Conector_DB
Destino:	ENCODE

Nome:	Tabela Sítios
Descrição:	Retorna os SLFT que estão dentro do intervalo do gap enviado
Conteúdo:	Posição do SLFT
Origem:	ENCODE
Destino:	Conector_DB

5.2.2.4 Deposito de Dados do DFD Processo 1 Analisador

Nome:	ChrN
Descrição:	Cromossomo alvo
Conteúdo:	ID do cromossomo
Fluxo de entrada	Enviado por parâmetro
Fluxo de saída	Todo o sistema

Nome:	Alinhamentos Homem-Chimpanzé / Homem-Gorila
Descrição:	Arquivo de alinhamento
Conteúdo:	Posições dos gaps
Fluxo de entrada	Processo 1.1 Procura gaps
Fluxo de saída	Processo 1.1 Procura gaps

Nome:	Result_Chimp_chrN / Result_Gorila_chrN
Descrição:	Arquivo de saída do Processo 1 Analisador
Conteúdo:	SLFT que estão contidos totalmente ou parcialmente nos gaps para cada cromossoma
Fluxo de entrada	Processo 1.1 Procura gaps
Fluxo de saída	Processo 1.1 Procura gaps

5.2.3 Especificação de Processos

5.2.3.1 Processo 1 Analisador

O primeiro *script* desenvolvido foi o parserTFinDB.pl[5], este script é responsável por procurar *gaps* nos arquivos de alinhamento e identificar se dentro da posição dos *gaps* encontrados, existe algum STFT do genoma humano. O *script* é executado separadamente para o chimpanzé e para o gorila.

Devido a necessidade de grande processamento a execução do *script* foi dividida por cromossomos, ou seja o *script* analisa um cromossomo por vez, utilizando as máquinas do cluster do I2Bio foi possível executar o *script* em máquinas diferentes com cromossomos diferentes ao mesmo tempo. Devido a limitações fornecidas pelos dados de alinhamento não foi possível realizar uma análise 100% precisa dos alinhamentos, sendo necessário realizar uma escolha manual das cadeias de DNA mais conservadas para serem analisadas. A seguir uma breve descrição do pseudocódigo do parserTFinDB.pl

Como o *script* realiza a análise para um cromossomo de uma espécie alvo por vez, é necessário passar alguns valores por parâmetro na hora da execução. Os

[5] Descrito no DFD como Processo 1 Analisador

parâmetros de entrada são:

- O arquivo de alinhamento da espécie alvo;
- O identificador do cromossomo alvo, por exemplo: chr1;
- A quantidade de cadeias mais conservadas no arquivo de alinhamento;
- IDs das cadeias utilizadas;

5.2.3.1.1 *Pseudocódigo*

- ABRIR e LER o arquivo de alinhamento;
- ABRIR arquivo de saída, para escrita;
- CONECTAR com o banco de dados onde está o arquivo do projeto ENCODE com os elementos regulatórios;
- PROCURAR no arquivo de alinhamento a primeira cadeia passada por parâmetro, até ter encontrado todas as cadeias;
- SE encontrado uma cadeia;
    - PROCURAR os *gaps* maiores que 20b;
    - SE encontrado um *gap* maior que 20b;
    - CHAMAR a função Pesquisa SLFT, está função é responsável por:
        - PROCURA os elementos regulatórios que estão presente no intervalo do *gap*, através do Processo ConnectionDB.pm[6];
        - CONECTAR com o banco de dados que contém os dados do elementos regulatórios;
        - SE encontrado o SLFT;
        - IMPRIMIR a posição do *gap* e os elementos regulatórios presentes no *gap*;
        - ATRIBUIR um *status* para esse elemento regulatório, 1 caso ele esteja totalmente presente no *gap* ou 2 caso ele esteja parcialmente presente no *gap*;
- EXECUTAR até acabar os *gaps* para aquela cadeia, então passa para próxima cadeia informada até não existir mais cadeias;

O produto destes *script*s resultará em 24 arquivos de textos referentes aos 24 cromossomos para cada espécie analisada, chimpanzé e gorila. Os arquivos estão armazenados em arquivos tabulados, estruturados da seguinte forma:

[6] Descrito no DFD como Conector_DB

chainID – identificador da cadeia;

startGap – posição onde inicia o *gap* no genoma humano;

endGap – posição onde termina o *gap* no genoma humano;

startSite – posição onde inicia o SLFT;

endSite – posição onde termina o SLFT;

nameTF – nome do FT;

status – *status* que em que o elemento regulatório se encontra, total ou parcial, 1 ou 2 respectivamente. 1 o sitio está totalmente dentro do *gap*, 2 o sitio está parcialmente dentro do *gap*;

rate – porcentagem do SLFT dentro do *gap*;

5.2.3.2 Processo 2 Comparador

Depois de ter identificado os fatores de transcrição para cada cromossomo de cada espécie, foi criado o *script* comparatorChimpGorila.pl[7] que realiza a comparação entre as duas espécies de animais, chimpanzé e gorila e identifica quais SLFT são comuns em ambas as espécies, indicando que estes SLFT estão presentes apenas na espécie humana, tendo em vista que eles estão localizados em *gaps* comuns entre o chimpanzé e gorila.

A execução deste código também foi dividido por cromossomos e recebe os seguintes parâmetros de entrada:

- Um arquivo resultante do *script* anterior, de um cromossomo alvo do chimpanzé (Result_Chimp_ChrN);
- Um arquivo resultante do *script* anterior, de um cromossomo alvo do gorila (Result_Chimp_ChrN);
- O identificador do cromossomos que está sendo comparado (ChrN);

5.2.3.2.1 Pseudocódigo

- ABRIR e LER os arquivos os arquivos resultados dos chimpanzé e do gorila para o cromossomo alvo;
- CRIAR o arquivo de saída;
- LER o arquivo do chimpanzé linha por linha;

[7] Descrito no DFD como Processo 2 Comparador

- VERIFICAR os sítios de ligação ausentes no chimpanzé;
- PROCURAR no arquivo do gorila;
- Se o mesmo sitio de ligação também estiver ausente;
    - GRAVAR no arquivo de saída e passa pro próximo sitio de ligação do chimpanzé;
- SE não existir;
    - GRAVAR também, mas com um código de *status* diferente;
- REALIZAR a busca vai sendo feita até percorrer todo arquivo do chimpanzé;
- REALIZAR o mesmo procedimento para o gorila;
- FIXAR em um sitio do gorila;
- PERCORRE todo arquivo do chimpanzé;
- PROCURAR os SLFT existentes apenas nos resultados do gorila;
- SALVAR e FECHAR arquivo de saída;

No final de todas as comparações serão gerados 24 arquivos tabulados de resultados, estruturados da seguinte forma:

nameTF – nome do fator de transcrição;

startSite – posição onde inicia o sítio de ligação;

endSite – posição onde termina o sitio de ligação;

statusC – *status* do sítio no chimpanzé. Pode ser (1, 2, -1). (1) se o sítio estiver num *gap* naquela espécie. (2) se o sítio estiver parcialmente num *gap* naquela espécie. (-1) se o sítio também existir no genoma do chimpanzé;

statusG – *status* do sítio no gorila. Pode ser (1, 2, -1). (1) se o sítio estiver num *gap* naquela espécie. (2) se o sítio estiver parcialmente num *gap* naquela espécie. (-1) se o sítio também existir no genoma do gorila;

idchrom – identificador do cromossomo;

Com os dados da comparação entre chimpanzé e gorila foi possível identificar quais elementos regulatórios estão ausentes em ambos os genomas. Então o próximo passo foi identificar os genes que estão próximos a estes elementos regulatórios, para isto foi desenvolvido o *script* searchGenesNextTF.pl, descrito abaixo:

5.3.2.3 Processo 3 Pesquisador

Tendo em vista que estes *script* necessita menos poder de processamento e para evitar o árduo trabalho de ficar executando o código para os 24 cromossomos, o

próprio script trata de chamar os arquivos dos 24 cromossomos e executá-los.

Os arquivos de entrada são chamados dentro do próprio código, são eles:

- O arquivo resultante da comparação do chimpanzé e gorila;
- O banco de dados que contém os genes do genoma humano;

5.3.2.3.3 Pseudocódigo

- ABRIR e LER o arquivo de comparação das espécies analisadas;
- CRIAR o arquivo de saída;
- CRIAR a conexão com o banco de dados;
- PROCURAR no arquivo resultado da comparação o elemento regulatórios onde o statusC e statusG sejam iguais a 1 ou misto com 2, até chegar ao final do arquivo;
- SE encontrado;
    - BUSCAR no banco de dados os genes que estão à frente dele 2.000b na fita negativa do DNA e os que estão 2.000b depois dele na fita positiva, através da função selectDB do Processo Connection_DB;
- GRAVAR no arquivo de saída junto com outras informações;
- ANALISAR para todos os outros cromossomos automaticamente;

A saída dele, também resultará na criação de 24 arquivos tabulados organizados da seguinte forma:

idChrom – identificador do cromossomo;
nameTF – nome do fator de transcrição;
startSite – posição do início do sitio de ligação;
endSite – posição final do sitio de ligação;
startGene – posição do início do gene;
endGene – posição final do sitio de ligação;
nameGene – nome do transcrito;
name2Gene – nome no gene;

O processo 4 é apenas um módulo desenvolvido para realizar a conexão e consultas no banco de dados, utilizando a interface de banco de dados do Perl (Perl DBI) (Perl, 2001).

6 RESULTADOS E DISCUSSÕES

Através da análise evolutiva dos elementos regulatórios, foram identificados 27.285 SLFT que possivelmente surgiram na espécie humana. Observou-se que o CTCF foi o FT com o maior número de sítios de ligação encontrados unicamente na espécie humana, contabilizando 1181 sítios, isso deve-se ao fato do CTCF também ter o maior número de sítios em todo o genoma humano, então foi normalizada a ocorrência dos sítios de ligação de cada FT encontrados unicamente na espécie humana e dividindo pelo total de sítios de ligação de cada FT presente em todo genoma humano, mostrando a percentagem dos sítios de ligação de cada FT presente apenas no genoma humano. Com os valores normalizados, foi verificado que o FT com o maior percentual de sítios de ligação, foi o XRCC4. A lista com todos os FT e o número de sítios encontrados pode ser encontrada no APÊNDICE A. Depois de mapeados os 27.285 SLFTs, realizou-se uma busca dos genes que estão próximos a estes SLFTs, na qual foi identificado 853 genes diferentes.

Feito a primeira parte dos objetivos do presente trabalho, que é realizar uma análise evolutiva dos elementos regulatórios, foi iniciado a análise do segundo objetivo, que é correlacionar a existência dos elementos regulatórios identificados em nossa análise evolutiva com o padrão de expressão dos genes humanos.

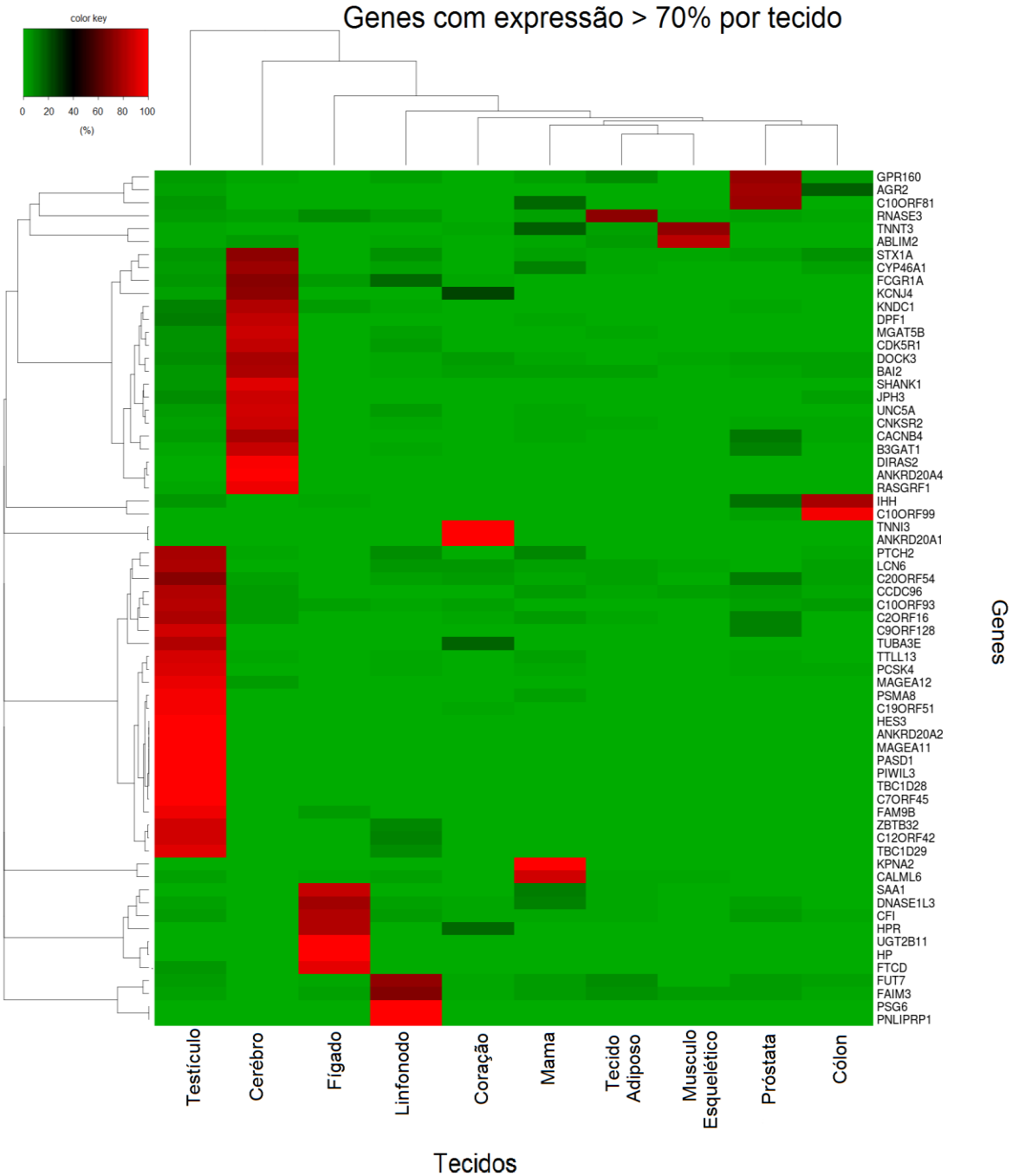
6.1 ANÁLISE DA EXPRESSÃO

Realizou-se uma análise da expressão dos 853 genes nos tecidos de testículo, musculo esquelético, próstata, linfonodo, fígado, coração, cólon, mama, cérebro e tecido adiposo. Devido à base de dados utilizada para verificar a expressão dos genes estar passando por um processo de atualização, encontrou-se dados de expressão para 611 genes nos 10 tipos de tecidos citados acima.

Para melhor identificar em qual tecido determinado gene estava sendo mais expresso, os resultados de expressão foram normalizados, verificando a porcentagem da expressão de cada gene em cada um dos 10 tecidos (para maiores detalhes ver

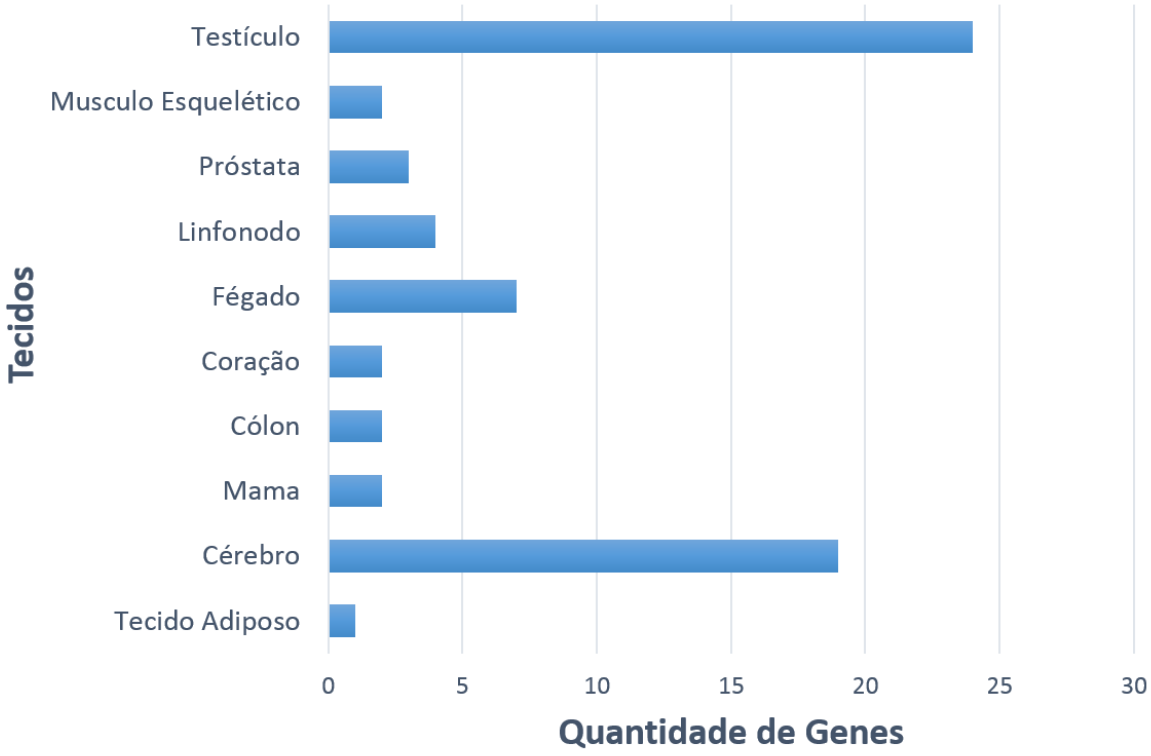
APÊNDICE C). Verificou-se então, quais genes apresentavam um valor de expressão acima de 70% em um determinado tecido. Foram selecionados 66 genes com pelo menos 70% de expressão em um determinado tecido. O Gráfico 1 mostra o nível de expressão dos 66 genes selecionados, onde as colorações verde e vermelha indicam baixo e alto valor de expressão, respectivamente. Através do dendrograma é possível visualizar o agrupamento de determinados genes com níveis de expressão próximas em determinados tecidos.

Gráfico 1: Percentagem da expressão dos 66 genes em cada tecido



O Gráfico 2 apresenta a quantidade de genes que tem o nível de expressão acima de 70% em um tecido específico. Observa-se que dos 66 genes com nível de expressão acima de 70%, 19 deles sendo expressos em cérebro e 24 em testículos. A maior ocorrência de genes expressos em testículos é um fato já observado dentro da biologia, mas o que nos chama mais atenção é a ocorrência de genes sendo expressos em cérebro, uma indicação de que os elementos que regulam estes genes tem uma relevância na história evolutiva da espécie humana.

Gráfico 2: Contagem de genes por tecido, com expressão acima de 70%



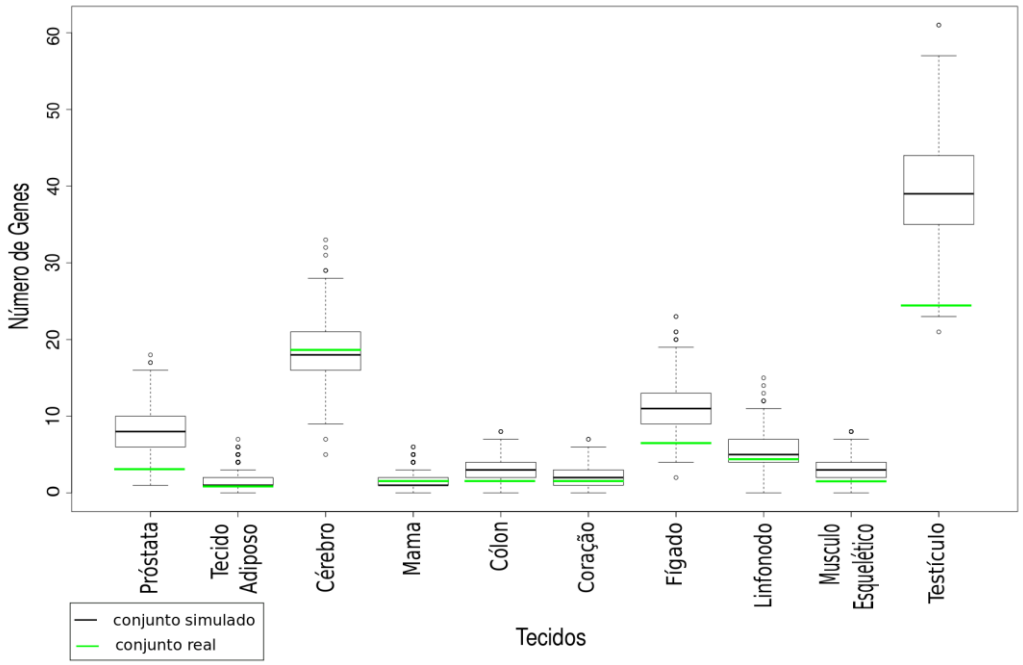
Para termos uma maior confiabilidade da distribuição observada dos genes expressos em cada tecido, foi realizada uma simulação probabilística chamada de Simulação de Monte Carlo. A Simulação de Monte Carlo é um método probabilístico que realiza uma série de simulações de cenários aleatórios, de modo a fornecer uma distribuição estatística dos resultados calculados.

Nesta simulação realizou-se a análise de expressão para 611 genes escolhidos aleatoriamente e verificado o número de genes com expressão acima de 70 % em um determinado tecido. Esta simulação foi realizada 1.000 vezes, utilizando em cada simulação uma nova lista de 611 genes gerados aleatoriamente.

Observou-se que a média do número de genes com expressão acima 70% era de 18,53 genes em cérebro e 39,43 genes em testículo. O Gráfico 3 mostra a distribuição de genes obtidas através da Simulação de Monte Carlo, calculando os valores da mediana, do quartil superior, do quartil superior e dos valores máximos e mínimos das simulações através do modelo gráfico de boxplot, (THEUS, 2009), nele também é apresentado os valores obtidos do conjunto real dos dados, representados com uma linha verde.

Também foi observado que das 1000 simulações realizadas, foi encontrado em 400 simulações um grupo com mais de 19 genes com expressão acima de 70% em cérebro e 994 simulações com um grupo com mais de 24 genes com expressão acima de 70% em testículo. Esses números indicam que o padrão de expressão do grupo de 611 genes não está enviesado para testículo (p=0.99) e cérebro (p=0.4). Para testículo, há de fato uma depleção de genes no conjunto real de 611 genes (p=0.01).

Gráfico 3: Nº de genes com expressão > 70% obtido através da Simulação de Monte Carlo



Para melhor entender as funções dos 853 genes selecionados na presente análise, analisou-se a sua ontologia através da utilização dos dados provenientes da iniciativa “Gene Ontology”. A análise da ontologia realizada abrangeu apenas os domínios de função molecular e processo biológico. A análise da ontologia foi

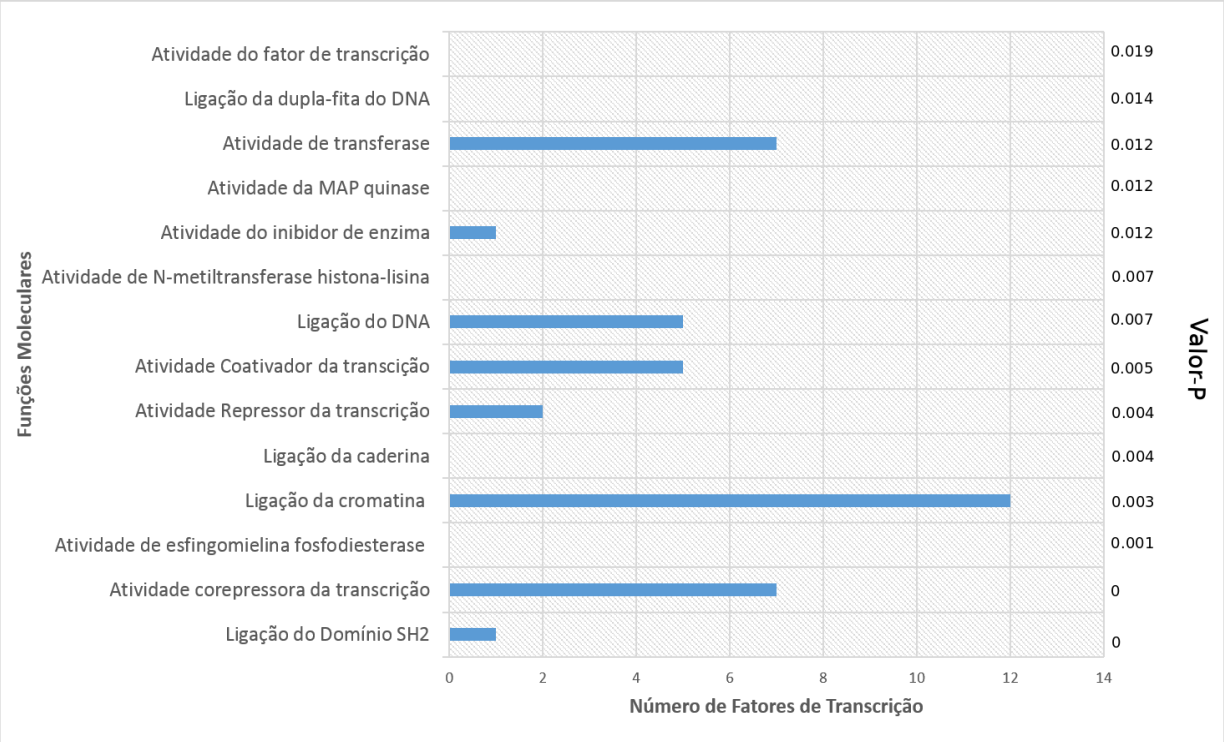
realizada com os 853 genes encontrados nas analise dos SLFTs. Observou-se no domínio de função molecular, um enriquecimento de genes para funções relacionadas ao mecanismo de transcrição (para maiores detalhes ver APÊNDICE D). Observou-se também que no domínio de processos biológicos há um enriquecimento de genes para processos relacionados ao desenvolvimento do tubo neural, cromatina, e principalmente para regulação da transcrição (para maiores detalhes, ver APÊNDICE E).

6.2 ANÁLISE DE ONTOLOGIA

Foi realizado também a análise de ontologia dos genes de cada um dos 30 primeiros FTs com maior número de genes (ver APÊNDICE B), afim de encontrar alguma relação entre as ontologias que tem um valor-p menor que 0.01 e a ontologia dos genes de cada um dos 30 primeiros FTs.

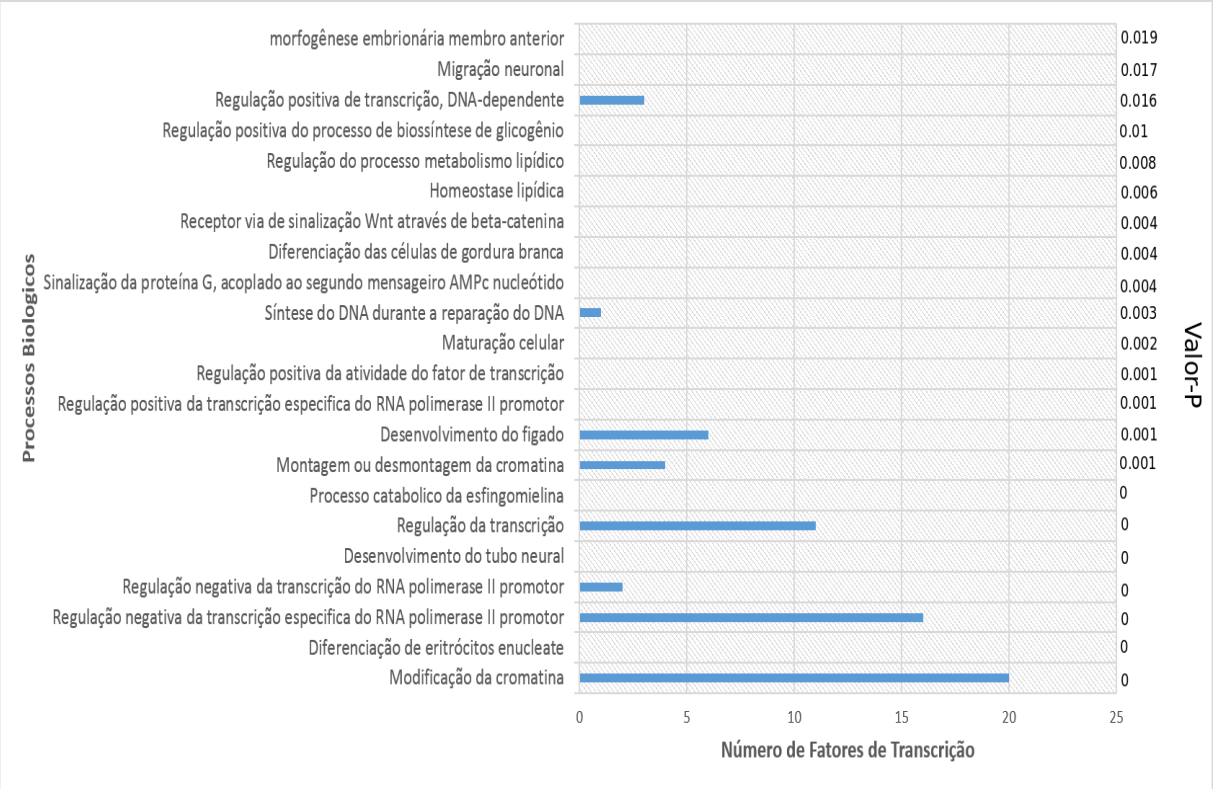
Observou-se uma ocorrência significativa de enriquecimento da função molecular de formação da cromatina e de funções relacionadas a transcrição. O Gráfico 4, mostra que a função molecular de construção da cromatina aparece nos genes de 12 dos 30 primeiros FTs.

Gráfico 4: Ocorrência das principais funções moleculares nos genes dos 30 primeiros FTs



No domínio “processos biológicos” também foi observado maior ocorrência nos processos de modificação da cromatina e processos relacionados a transcrição. Sendo observado no Gráfico 5 que o processo biológico de modificação da cromatina aparece nos genes de 20 dos 30 primeiros FTs e a ocorrência do processo de regulação positiva da transcrição específica de RNA polimerase II promotor encontrada nos genes de 16 dos 30 primeiros FTs.

Gráfico 5: Ocorrência dos principais processos biológicos nos genes dos 30 primeiros FTs



Os resultados de ontologia indicam claramente que genes relacionados ao controle de transcrição estão enriquecidos no grupo de 853 genes com SLFT específicos da espécie humana. Sabe-se que alterações em genes dessa categoria funcional tem um efeito mais pronunciado na evolução de novos tratos fenotípicos. Isso deve-se ao fato de que tais genes estão no topo hierárquico de vias de controle de regulação gênica. Devido a isso, eles são chamados de genes maestro.

7 CONCLUSÃO

Na análise evolutiva realizada dos elementos regulatórios, foram identificados 27.285 SLFT que possivelmente surgiram na espécie humana em algum momento de sua evolução, deste sua divergência com o chimpanzé.

Contudo, analisando os dados obtidos, já é possível observar uma significância dos genes regulados pelos elementos regulatórios identificados na análise evolutiva. Avaliando o nível de expressão dos genes encontrados, observamos uma maior ocorrência de genes que têm um valor de expressão significante nos tecidos de cérebro e testículo. A maior ocorrência de genes expressos em testículos é um fato observado na biologia, mas o que nos chama atenção é a ocorrência de genes sendo expressos em cérebro, este resultado observado tem uma grande significância evolutiva, tendo em vista que uma das características que nos diferencia dos primatas é a nossa capacidade de raciocínio. Observamos também utilizando a Simulação de Monte Carlo, uma depleção na quantidade de genes expressos em testículo na espécie humana.

Outro resultado muito interessante observado foi da ontologia dos 853 genes encontrados, que nos mostra o enriquecimento de genes para os processos biológicos de desenvolvimento do tubo neural e da migração neural. Outros processos biológicos e funções moleculares também foram encontradas, relacionados a formação da cromatina e da regulação da transcrição, mas a presença delas é compreensível, já que elas estão localizadas em um nível biológico mais propicio a sofrem maior influência das ações da evolução.

Nossas analises mostram claramente que um grupo de genes que estão associados com o sistema nervoso central, são reguladas por uma série de elementos de regulação que possivelmente foram originadas na linhagem humana, mas sendo necessário pra trabalhos futuros, a realização de experimentos biológicos com o DNA do chimpanzé e do gorila para comprovar a ausência dos SLFTs em seus DNAs, comprovando o surgimento dos elementos regulatórios na linhagem humana.

REFERÊNCIAS

ARBIZA, L. et al. Genome-wide inference of natural selection on human transcription factor binding sites. **Nature Genetics**, v. 45, n. 7, p. 723–729, 9 jun. 2013.

BENSON, D. A. et al. GenBank. **Nucleic Acids Research**, v. 41, n. D1, p. D36–D42, 27 nov. 2012.

BOYLE, A. P. et al. Annotation of functional variation in personal genomes using RegulomeDB. **Genome Research**, v. 22, n. 9, p. 1790–1797, 5 set. 2012.

CONSORTIUM, T. E. P. An integrated encyclopedia of DNA elements in the human genome. **Nature**, v. 489, n. 7414, p. 57–74, 6 set. 2012.

EDWARDS, D. **Bioinformatics tools and applications**. New York; London: Springer, 2009.

FASSLER, J.; COOPER, P. **BLAST® Help, NCBI Help Manual**. Disponível em: <http://www.ncbi.nlm.nih.gov/books/NBK62051/>. Acesso em: 17 jul. 2013.

GenBank Statistics. Disponível em: <http://www.ncbi.nlm.nih.gov/genbank/genbankstats-2008/>. Acesso em: 24 abr. 2013.

HILLER, M. et al. Computational methods to detect conserved non-genic elements in phylogenetically isolated genomes: application to zebrafish. **Nucleic Acids Research**, 27 jun. 2013.

LANDT, S. G. et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. **Genome Research**, v. 22, n. 9, p. 1813–1831, 5 set. 2012.

LESK, A. M.; ANDRADE, A. E. **Introdução à bioinformática**. Porto Alegre: Artmed, 2008.

Perl: how to program. Upper Saddle River, N.J: Prentice Hall, 2001.

PROSDOCINI, F. et al. BIOTENCNOLOGIA, Ciência & Desenvolvimento. **Bioinformática: Manual do Usuario**, v. V, n. 29, dez. 2002.

REZENDE, D. A. **Engenharia de software e sistemas de informação**. Rio de Janeiro: Brasport, 2006.

SCIENCEINSCHOOL. **scienceinschool.org**. Disponível em: <http://www.scienceinschool.org/print/2331>. Acesso em: 27 ago. 2013.

SPIVAKOV, M. et al. Analysis of variation at transcription factor binding sites in Drosophila and humans. **Genome Biology**, v. 13, n. 9, p. R49, 2012.

STEVENS, R. Ontology-based knowledge representation for bioinformatics. **Briefings in Bioinformatics**, v. 1, n. 4, p. 398–414, 1 jan. 2000.

THEUS, M. **Interactive graphics for data analysis: principles and examples**. Boca Raton: CRC Press, 2009.

UCSC, G. B. **UCSC Genome Browser: Chain Format**. Disponível em: <http://genome.ucsc.edu/goldenPath/help/chain.html>. Acesso em: 23 jul. 2013.

UCSC, G. B. **Alignment gorila**. Disponível em: <http://hgdownload.soe.ucsc.edu/goldenPath/hg19/vsGorGor3/>. Acesso em: 23 jul. 2013.

UCSC Genome Browser Home. Disponível em: <http://genome.ucsc.edu/>. Acesso em: 18 jul. 2013.

VARGAS, R. B. **Identification of Regulatory Elements using Comparative Genomics and Phylogenetic Footprinting**. Tese Graduação—Vitória-Brasil: UFES, 11 ago. 2006.

WILDMAN, D. E. Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: Enlarging genus Homo. **Proceedings of the National Academy of Sciences**, v. 100, n. 12, p. 7181–7188, 2 jun. 2003.

APÊNDICE A – Tabela SLFT para cada fator de transcrição

Genes	Número de SLFT encontrados apenas na espécie humana	Número de SLFT encontrados em todo genoma humano	Valor normalizado
CTCF	1181	112155	1,05%
Pol2	1102	76623	1,44%
Pol2-4H8	825	58996	1,40%
Egr-1	770	34044	2,26%
TAF1	699	34042	2,05%
HEY1	588	30300	1,94%
Rad21	582	72543	0,80%
E2F6_(H-50)	575	26422	2,18%
c-Myc	555	35736	1,55%
CEBPB	552	81423	0,68%
ZBTB7A_(SC-34508)	519	19926	2,60%
CTCF_(SC-5916)	504	50505	1,00%
PU.1	496	55128	0,90%
ELF1_(SC-631)	479	36883	1,30%
FOXA1_(C-20)	442	66091	0,67%
HA-E2F1	426	18580	2,29%
YY1	426	23591	1,81%
CTCF_(C-20)	425	42804	0,99%
YY1_(C-20)	412	32804	1,26%
USF-1	409	44846	0,91%
TBP	371	34675	1,07%
ZNF263	366	25944	1,41%
MafK_(ab50322)	364	58318	0,62%
PAX5-C20	358	23404	1,53%
Sin3Ak-20	331	20932	1,58%
p300	319	70914	0,45%
SP1	317	35415	0,90%
GABP	309	17563	1,76%
NRSF	302	27386	1,10%
SMC3_(ab9263)	298	38434	0,78%
JunD	287	48954	0,59%
SETDB1	285	19140	1,49%
CCNT2	268	17115	1,57%
NFKB	261	32291	0,81%
FOXA1_(SC-101058)	260	34272	0,76%
EBF1_(C-8)	247	27567	0,90%
STAT3	235	52438	0,45%
TCF12	229	21601	1,06%
FOXA2_(SC-6554)	229	32864	0,70%
CTCFL_(SC-98982)	215	12906	1,67%
Max	215	30219	0,71%
TAL1_(SC-12984)	208	22683	0,92%
USF1_(SC-8983)	205	19812	1,03%
BCL3	199	9600	2,07%
GATA-2	195	53976	0,36%
GR	192	26559	0,72%
POU2F2	191	17331	1,10%

Oct-2	191	17443	1,09%
c-Fos	191	45176	0,42%
c-Jun	190	42843	0,44%
KAP1	186	19427	0,96%
E2F6	185	11933	1,55%
SUZ12	181	6124	2,96%
Znf143_(16618-1-AP)	181	18296	0,99%
MafK_(SC-477)	181	32268	0,56%
TCF4	171	24353	0,70%
NF-YA	167	8335	2,00%
HDAC2_(SC-6296)	167	22838	0,73%
EBF	165	24996	0,66%
IRF1	162	11498	1,41%
TAF7_(SQ-8)	161	8307	1,94%
ETS1	161	9735	1,65%
NF-YB	157	8958	1,75%
FOSL2	154	19704	0,78%
AP-2gamma	151	17957	0,84%
HMGN3	150	10966	1,37%
PAX5-N19	150	14330	1,05%
GATA3_(SC-268)	150	28108	0,53%
MafF_(M8194)	150	28433	0,53%
HNF4A_(H-171)	145	22696	0,64%
BATF	142	24594	0,58%
GATA-1	141	21375	0,66%
E2F1	137	4591	2,98%
USF2	132	17717	0,75%
ERalpha_a	132	21170	0,62%
Pol2(phosphoS2)	131	11243	1,17%
RXRA	130	16566	0,78%
ZEB1_(SC-25388)	125	8283	1,51%
SRF	124	14967	0,83%
CHD2_(N-1250)	121	9312	1,30%
Pbx3	120	7410	1,62%
GATA2_(CG2-96)	118	12118	0,97%
Nrf1	114	6382	1,79%
RFX5_(N-494)	108	15412	0,70%
Pol2(b)	104	13984	0,74%
TR4	101	3418	2,95%
Mxi1_(bHLH)	97	11566	0,84%
AP-2alpha	93	11800	0,79%
eGFP-JunD	91	20332	0,45%
MEF2A	90	16689	0,54%
HNF4G_(SC-6558)	88	15615	0,56%
E2F4	86	5520	1,56%
NANOG_(SC-33759)	82	10022	0,82%
IRF4_(M-17)	81	14224	0,57%
BCL11A	79	14968	0,53%
STAT1	78	11330	0,69%
FOSL1_(SC-183)	75	8341	0,90%
Ini1	73	6949	1,05%
ZBTB33	68	3583	1,90%

BCLAF1_(M33-P5B11)	66	5098	1,29%
SIX5	66	5136	1,29%
p300_(N-15)	63	18973	0,33%
ATF3	61	5459	1,12%
GTF2F1_(RAP-74)	58	9450	0,61%
SP2_(SC-643)	57	3004	1,90%
THAP1_(SC-98174)	53	2580	2,05%
RPC155	52	2609	1,99%
TFIIIC-110	48	3284	1,46%
BRCA1_(C-1863)	48	5225	0,92%
MEF2C_(SC-13268)	39	6775	0,58%
BHLHE40	38	3864	0,98%
HNF4A	36	7473	0,48%
ELK4	33	4742	0,70%
eGFP-FOS	32	8298	0,39%
eGFP-JunB	32	9845	0,33%
eGFP-GATA2	31	8098	0,38%
STAT2	26	2721	0,96%
CtBP2	24	2536	0,95%
PRDM1_(Val90)	24	3543	0,68%
NF-E2	22	2032	1,08%
BAF155	22	4260	0,52%
IRF3	20	1574	1,27%
POU5F1_(SC-9081)	18	3134	0,57%
p300_(F-4)	16	2736	0,58%
ZZZ3	15	839	1,79%
SREBP1	14	1637	0,86%
BDP1	13	595	2,18%
GTF2B	13	1939	0,67%
SIRT6	12	1706	0,70%
ZNF274	11	1123	0,98%
NELFe	9	478	1,88%
BRF2	9	818	1,10%
Brg1	9	1816	0,50%
NF-E2_(H-230)	8	713	1,12%
ERRA	7	645	1,09%
HSF1	7	984	0,71%
WHIP	7	1137	0,62%
BAF170	6	1316	0,46%
GRp20	5	420	1,19%
SPT20	4	271	1,48%
eGFP-HDAC8	4	285	1,40%
GCN5	3	128	2,34%
SREBP2	3	209	1,44%
BRF1	3	272	1,10%
XRCC4	2	24	8,33%
Pol3	2	181	1,10%
PGC1A	2	429	0,47%
eGFP-NR4A1	1	116	0,86%

APÊNDICE B - Tabela de genes encontrados por FT

Fatores de Transcrição	Nº de Genes
Pol2	222
Egr-1	202
Pol2-4H8	187
E2F6_(H-50)	177
ZBTB7A_(SC-34508)	174
HEY1	169
TAF1	164
CTCF	132
ELF1_(SC-631)	122
HA-E2F1	107
Sin3Ak-20	96
TBP	96
YY1	95
CCNT2	90
c-Myc	89
YY1_(C-20)	83
GABP	81
CTCF_(C-20)	80
PAX5-C20	75
CTCFL_(SC-98982)	72
CTCF_(SC-5916)	72
Rad21	71
CEBPB	67
ZNF263	67
NFKB	66
USF1_(SC-8983)	66
SP1	65
E2F6	61
USF-1	60
ETS1	57
HMGN3	48
Znf143_(16618-1-AP)	48
EBF1_(C-8)	47
IRF1	47
NRSF	47
TCF12	47
Nrf1	46
PU.1	46
SMC3_(ab9263)	45
FOXA1_(C-20)	44
Max	44
E2F1	43
Oct-2	41
POU2F2	41
Pol2(phosphoS2)	39
TAF7_(SQ-8)	39
ZEB1_(SC-25388)	39
JunD	38

AP-2gamma	37
NF-YB	34
CHD2_(N-1250)	33
E2F4	33
HDAC2_(SC-6296)	33
USF2	33
EBF	32
Mxi1_(bHLH)	32
p300	32
SRF	31
TCF4	31
RFX5_(N-494)	30
Pbx3	29
RXRA	29
NF-YA	28
MafK_(ab50322)	27
c-Fos	26
PAX5-N19	26
Pol2(b)	26
FOXA1_(SC-101058)	25
HNF4A_(H-171)	25
BCLAF1_(M33-P5B11)	21
SP2_(SC-643)	21
AP-2alpha	20
ATF3	20
THAP1_(SC-98174)	20
GATA-1	19
GR	19
HNF4G_(SC-6558)	19
NANOG_(SC-33759)	19
BCL3	18
FOSL2	18
Ini1	18
BRCA1_(C-1863)	17
FOXA2_(SC-6554)	17
MEF2A	17
SIX5	17
MafK_(SC-477)	16
ZBTB33	16
GATA3_(SC-268)	15
SUZ12	15
ELK4	14
FOSL1_(SC-183)	14
GTF2F1_(RAP-74)	14
BCL11A	13
IRF4_(M-17)	13
STAT1	13
BATF	12
IRF3	12
MafF_(M8194)	12
SETDB1	12
TAL1_(SC-12984)	12

GATA-2	11
KAP1	11
STAT3	11
BHLHE40	10
eGFP-JunD	10
ERalpha_a	10
RPC155	10
TFIIIC-110	10
c-Jun	9
HNF4A	7
STAT2	6
MEF2C_(SC-13268)	5
TR4	5
BAF155	4
NF-E2_(H-230)	4
PRDM1_(Val90)	4
SREBP1	4
Brg1	3
CtBP2	3
eGFP-FOS	3
eGFP-GATA2	3
ERRA	3
NF-E2	3
p300_(N-15)	3
BDP1	2
GATA2_(CG2-96)	2
GTF2B	2
HSF1	2
SREBP2	2
WHIP	2
eGFP-JunB	1
GCN5	1
GRp20	1
NELFe	1
POU5F1_(SC-9081)	1
SIRT6	1
SPT20	1
ZNF274	1
ZZZ3	1

APÊNDICE C – Tabela de expressão dos 66 genes com expressão > 70%

Genes	Normal prostate	Adipose	Brain	Breast	Colon	Heart	Liver	Lymph node	Skeletal muscle	Testes
ABLIM2	0	4	4	1	0	0	0	1	84	1
AGR2	78	0	0	0	18	0	0	0	0	2
ANKRD20A1	0	0	0	0	0	100	0	0	0	0
ANKRD20A2	0	0	0	0	0	0	0	0	0	100
ANKRD20A4	0	0	100	0	0	0	0	0	0	0
B3GAT1	9	0	86	0	0	0	0	1	0	1
BAI2	1	2	80	2	2	2	0	1	0	5
C10ORF81	77	0	0	16	0	0	0	0	0	5
C10ORF93	2	0	3	0	3	2	2	1	0	83
C10ORF99	2	0	0	0	97	0	0	0	0	0
C12ORF42	0	0	0	0	0	0	0	9	0	89
C19ORF51	0	0	0	0	0	1	0	0	0	97
C20ORF54	11	2	2	1	2	3	0	2	0	72
C2ORF16	9	1	3	3	0	1	0	0	0	80
C7ORF45	0	0	0	0	0	0	0	0	0	99
C9ORF128	9	0	0	0	0	0	0	0	0	89
CACNB4	12	0	79	1	1	0	0	0	0	4
CALML6	0	1	0	89	0	0	1	2	1	2
CCDC96	3	1	3	3	1	0	0	0	2	82
CDK5R1	0	0	85	0	0	0	0	4	0	6
CFI	3	1	0	1	1	1	82	3	0	3
CNKSR2	1	1	88	1	1	0	0	1	0	2
CYP46A1	0	1	77	9	2	0	0	2	0	4
DIRAS2	0	0	98	0	0	0	0	0	0	0
DNASE1L3	2	1	0	9	0	1	78	2	0	2
DOCK3	2	0	79	1	2	3	0	1	1	7
DPF1	0	0	85	1	0	0	0	0	0	11
FAIM3	4	6	0	4	1	1	2	71	4	2
FAM9B	0	0	0	0	0	0	3	0	0	96
FCGR1A	0	0	72	0	0	1	3	15	0	5
FTCD	0	0	0	0	0	0	93	0	0	5
FUT7	4	7	0	4	2	1	1	74	0	4
GPR160	76	7	1	2	4	0	0	2	0	3
HES3	0	0	0	0	0	0	0	0	0	100
HP	0	0	0	0	0	0	99	0	0	0
HPR	0	0	0	0	0	16	81	0	0	0
IHH	13	0	0	0	79	0	1	0	0	5
JPH3	0	0	88	0	2	0	0	0	0	7
KCNJ4	0	0	73	0	0	24	0	0	0	2
KNDC1	1	0	82	0	0	0	4	1	0	9

KPNA2	0	0	0	100	0	0	0	0	0	0
LCN6	0	2	0	2	2	5	0	5	1	80
MAGEA11	0	0	0	0	0	0	0	0	0	100
MAGEA12	0	0	4	0	0	0	0	0	0	95
MGAT5B	0	1	87	0	0	0	0	2	0	6
PASD1	0	0	0	0	0	0	0	0	0	100
PCSK4	1	0	0	1	1	0	0	1	0	91
PIWIL3	0	0	0	0	0	0	0	0	0	100
PNLIPRP1	0	0	0	0	0	0	0	100	0	0
PSG6	0	0	0	0	0	0	0	100	0	0
PSMA8	0	0	0	2	0	0	0	0	0	97
PTCH2	0	0	1	8	1	0	0	7	0	79
RASGRF1	0	0	96	0	0	0	0	0	0	1
RNASE3	2	73	2	2	1	0	7	4	1	3
SAA1	0	1	0	11	0	0	86	0	0	0
SHANK1	0	0	92	0	0	0	0	0	0	5
STX1A	2	1	73	2	6	0	0	6	1	5
TBC1D28	0	0	0	0	0	0	0	0	0	100
TBC1D29	0	0	0	0	0	0	0	7	0	92
TNNI3	0	0	0	0	0	99	0	0	0	0
TNNT3	0	2	0	18	0	1	0	0	74	1
TTLL13	1	0	1	2	0	0	0	1	0	90
TUBA3E	0	0	0	0	0	17	0	0	0	82
UGT2B11	0	0	0	0	0	0	100	0	0	0
UNC5A	0	0	89	1	0	0	0	3	0	4
ZBTB32	0	0	0	0	0	0	0	8	0	89

APÊNDICE D – Funções moleculares onde os 853 estão mais enriquecidos

Valor-P	ID Gene Ontology	Nº Genes	Ontologia	Genes
p: 0.001	GO:0004767	3	sphingomyelin phosphodiesterase activity	(SMPDL3B SMPD3 SMPD4)
p: 0.003	GO:0003682	15	chromatin binding	(CHD4 TLE1 POLE MSL3 CDYL POLD1 SP3 CDYL2 KAT5 SMARCC2 ZEB1 NSD1 SKIL CHD5 CTNNB1)
p: 0.004	GO:0016564	13	transcription repressor activity	(ZFHX3 KAT5 HES1 CTBP1 SP3 HMX1 ZNF148 ATXN1 DNAJB6 ID2 HIVEP1 CTBP2 FOXN3)
p: 0.004	GO:0045296	4	cadherin binding	(NEO1 PSEN1 CDK5R1 CTNNB1)
p: 0.005	GO:0003713	16	transcription coactivator activity	(GMEB1 ZEB1 SMARCD2 SERTAD2 SRCAP CTNNB1 NCOA4 EP300 MED16 ZFX VGLL1 LPIN2 CREBBP ARID1B KAT5 SMARCC2)
p: 0.007	GO:0003677	66	DNA binding	(ZNF618 CHD5 PHF21A VEZF1 ZNF702P NTHL1 HP1BP3 ZFPM1 MLL3 HDAC3 TBL1XR1 NAT14 CHD4 SMARCC2 ZNF592 SRCAP RREB1 NUCB1 STRBP NPAS3 UBTF ZNF512B TOX3 POLE ZNF644 DNASE1L3 SMARCA4 HIVEP1 ZNF630 ZNF319 ZFAT FUS GLYR1 POLD1 HES1 PIAS4 ZNF358 FBXL19 TCEA1 SALL3 SETD2 ZNF114 TWIST1 RYBP KIF22 ZNF443 MAZ RPS27 ZFX ZNF516 HES3 CXXC4 DOT1L TERT TIGD2 HIST2H2BF ZBTB32 SKIL WHSC1 PRR12 ZNF451 ZNF512 ARID1B SP3 DNAJB6 CDK9)
p: 0.007	GO:0018024	4	histone-lysine N-methyltransferase activity	(MLL3 SETD2 WHSC1 DOT1L)
p: 0.012	GO:0004707	3	MAP kinase activity	(MAPK12 MAPK11 MAPK10)
p: 0.012	GO:0004857	4	enzyme inhibitor activity	(CRIM1 TWIST1 PRPSAP2 DGKZ)
p: 0.012	GO:0016740	63	transferase activity	(MLL3 WHSC1 PRKACA CDC42BPB DGKE PRKCE DGKZ KAT5 POLD1 NAT8L TERT DYRK2 SETD2 MAPK11 IRAK3 CDYL GRK5 CHSY1 FTCD DOT1L MBOAT1 SRC SBK1 STK11 B4GALNT4 CREBBP ITPKA DAPK3 HMBS PGS1 EP300 MAP4K4 NSUN5 CDK9 GSTT2 POLE ZDHHC14 TNK2 CMPK1 WEE1 MAPK10 MAPK12 CHST3 ROR2 TRMT1 NME6 TLK1 TLK2 SETD1B PRKACB STRADA ADRBK2 LPCAT1 NSD1 PORCN NAT14 ZDHHC9

				PIGG B3GAT1 INSR MAST2 CAMKK1 STK24)
p: 0.014	GO:0003690	7	double-stranded DNA binding	(ZEB1 SP3 NTHL1 CTNNB1 ZNF148 USF2 CEBPG)
p: 0.019	GO:0003700	51	transcription factor activity	(PBX3 ONECUT3 RCOR2 IRX2 PPARA CREBBP NFATC1 MSL3 LHX4 HMX1 SOX1 ZNF148 ZNF444 SOX8 MLLT10 LCOR HIC1 POU3F1 CTBP1 PPARG ELK1 TCF7L2 RORA GSC2 IRX4 NFYB TCF3 FOXN3 NFIC HEY1 CBFA2T3 MAFG ZSCAN4 LMX1B KLF3 USF2 FOXD4 EP300 ZNF138 FOXK2 SPIB KLF10 MEIS3P1 GMEB1 VAX1 TCF25 NFIB NR2F6 CTNNB1 CEBPG ZEB1)
p: 0.022	GO:0010843	7	promoter binding	(ZFHX3 TBL1XR1 CTNNB1 NFATC1 TCF3 PPARG HDAC3)
p: 0.024	GO:0050681	4	androgen receptor binding	(KAT5 CTNNB1 NCOA4 NSD1)
p: 0.025	GO:0019904	9	protein domain specific binding	(PPARA TNNI3 SKI ATN1 CTBP1 IRS2 DLGAP4 CHMP1A SKIL)
p: 0.027	GO:0004402	5	histone acetyltransferase activity	(CDYL SRCAP CREBBP EP300 KAT5)
p: 0.042	GO:0008601	3	protein phosphatase type 2A regulator activity	(PPP2R2A PPP2R5C SET)
p: 0.047	GO:0008022	11	protein C-terminus binding	(SLC6A4 BLOC1S2 ATXN1 DGKZ SLC9A3R2 FOXN3 HRAS CTBP1 CTNNB1 MLLT4 USP7)
p: 0.049	GO:0003746	3	translation elongation factor activity	(EEFSEC ABTB1 TCEA1)
p: 0.051	GO:0008415	10	acyltransferase activity	(LPCAT1 CDYL ZDHHC9 ZDHHC14 MBOAT1 KAT5 PORCN GCAT NAT8L NAT14)
p: 0.053	GO:0016818	3	hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides	(SMARCA4 CHD5 CHD4)
p: 0.065	GO:0005097	5	Rab GTPase activator activity	(TBC1D28 TBC1D2B TBC1D1 TBC1D29 TBC1D3H)
p: 0.069	GO:0019903	3	protein phosphatase binding	(CTNNB1 IRS2 INSR)
p: 0.077	GO:0008168	8	methyltransferase activity	(MLL3 SETD1B NSUN5 TRMT1 DOT1L WHSC1 NSD1 SETD2)

p: 0.084	GO:0008270	105	zinc ion binding	(MEX3D ZFPM1 CREBBP ZNF630 ZFAT ZNF516 ZNF138 ZNF644 RNF182 ZNF592 HIVEP1 DGKE MLL3 TRMT1 ZFAND2B ZDHHC9 ZFYVE28 ZCCHC14 PLEKHM1P BMI1 LMX1B PHF21A GMEB1 POLE LMCD1 KLF10 U2AF1 MAZ ZNF618 SALL3 MORC4 ZNF512 SPIRE2 CXXC4 ATXN7L3 ZNF702P POLD1 PPARG DGKZ TP53I3 HELZ ZC3H12C ASTL NPEPL1 ZNF358 MLLT6 ZDHHC14 CBFA2T3 SKI HIC1 PHF21B ZFX RREB1 PPARA KLF3 ZNF512B LHX4 RYBP CHMP1A SP3 LUC7L2 EP300 TCEA1 UBR3 MLLT10 ZBTB32 ABLIM2 CHD4 PDLIM2 CHD5 VEZF1 FUS ZNF443 CCS ZNF148 KAT5 POLR2H ZNF114 ZNF444 PHF10 ZNF451 APOBEC3G LIMD2 ZEB1 ACAP3 RNF157 NSD1 NR2F6 ZNF319 RORA ZSCAN4 SQSTM1 WHSC1 PRKCE IKBKG NAIP DPF1 CDC42BPB ADAM11 ZFHX3 TRAF3 PHF19 PIAS4 FBXL19 RPS27)
p: 0.089	GO:0043565	27	sequence-specific DNA binding	(CEBPG FOXD4 IRX2 ONECUT3 HMX1 VAX1 ELK1 SPIB NR2F6 POU3F1 ZNF148 SOX1 NFYB ZEB1 FOXN3 LHX4 IRX4 MEIS3P1 RORA FOXK2 PBX3 GSC2 MAFG ZFHX3 USF2 LMX1B PPARA)
p: 0.08	GO:0046966	3	thyroid hormone receptor binding	(NSD1 MED16 TRIP12)
p: 0.097	GO:0046332	3	SMAD binding	(CREBBP SKIL SKI)
p: 0.099	GO:0016563	12	transcription activator activity	(ELK1 RREB1 USF2 SP3 NFATC1 SMARCA4 TBL1XR1 PPARA TCF3 POU3F1 NAT14 PPARG)
p: 0.113	GO:0016798	4	hydrolase activity, acting on glycosyl bonds	(NTHL1 ATHL1 SMPDL3B GBA)
p: 0.116	GO:0003707	4	steroid hormone receptor activity	(RORA PPARG PPARA NR2F6)
p: 0.117	GO:0032403	5	protein complex binding	(SKIL INSR PRKACA GNAO1 ATP5D)
p: 0.123	GO:0008134	9	transcription factor binding	(TCF7L2 HDAC3 CEBPG CTBP1 TCF3 CTNNB1 ZEB1 LCOR TLE1)
p: 0.135	GO:0046872	109	metal ion binding	(ZBTB32 MLLT10 ATXN7L3 SP3 ASTL KLF10 CBFA2T3 HIC1 ZNF644 ZNF451 MAZ PIAS4 BMI1 RNF157 ZNF592

				ZNF358 NTHL1 ZFAT TRMT1 FUS SALL3 ZNF512B ZNF630 CHSY1 APOBEC3G KLF3 TCEA1 ZFHX3 CXXC4 MEX3D VEZF1 CHD4 PDLIM2 GMEB1 P4HA2 B4GALT7 ZFAND2B MLL3 PPARG CREBBP RREB1 ACAP3 UBR3 RYBP PPARA ZDHHC9 ZNF516 NR2F6 TRAF3 DGKZ HELZ PHF10 NSD1 RPS27 INSR MORC4 ZDHHC14 PLEKHM1P CYP46A1 EP300 DGKE ZNF512 LHX4 ZNF114 ZFYVE28 KAT5 FBXL19 EIF2C2 ZNF319 LEPREL2 RORA ZNF702P CHD5 ZEB1 RNF182 ZCCHC14 ZNF138 ZNF148 ZSCAN4 ZNF618 PDF ZNF443 WHSC1 TYW1B LMCD1 PHF21A LIMD2 SQSTM1 PHF19 PHF21B POLE CCS PRKCE POLD1 NPEPL1 LUC7L2 MLLT6 IKBKG ABLIM2 HIVEP1 NAIP LMX1B ZFX MGAT5B DPF1 ZFPM1 B3GAT1 ZNF444 U2AF1)
p: 0.137	GO:0004674	17	protein serine/threonine kinase activity	(PRKACB MAPK10 PRKACA MAST2 SBK1 MAP4K4 MAPK11 MAPK12 TLK2 TLK1 WEE1 IRAK3 STK11 STK24 CDC42BPB DYRK2 DAPK3)
p: 0.142	GO:0016757	9	transferase activity, transferring glycosyl groups	(GALNT10 RFNG MGAT5B LARGE FUT7 DPAGT1 ALG12 B3GNT7 B4GALT7)
p: 0.144	GO:0048306	3	calcium-dependent protein binding	(STX1A TNNI3 TNNT3)
p: 0.145	GO:0019901	7	protein kinase binding	(TNNI3 IRS2 DVL1 TCF7L2 SKI CDKN2D PRKACA)
p: 0.151	GO:0016566	3	specific transcriptional repressor activity	(PPARG TCF7L2 NACC1)
p: 0.16	GO:0003704	3	specific RNA polymerase II transcription factor activity	(VEZF1 ZNF148 ZBTB32)
p: 0.175	GO:0019899	6	enzyme binding	(COTL1 LUC7L2 RAC1 PRKCE CCS ZFHX3)
p: 0.177	GO:0004871	15	signal transducer activity	(IKBKG ADRBK2 TRAF3 CTNNB1 PASD1 ARHGEF11 IRS2 DVL1 NPAS3 GRK5 GNAO1 RTN2 PRKCE CREBBP RGS19)
p: 0.17	GO:0004221	5	ubiquitin thiolesterase activity	(USP27X USP7 USP34 USP42 USP18)
p: 0.189	GO:0004715	3	non-membrane spanning protein tyrosine kinase activity	(WEE1 SRC TNK2)

p: 0.191	GO:0051287	3	NAD or NADH binding	(GPD1L CTBP1 CTBP2)
p: 0.193	GO:0003924	12	GTPase activity	(RRAS2 RAP2C EIF5 HRAS EEFSEC DIRAS2 TUBA3E TUBB2A RAC1 GNAO1 ARHGAP5 RASD2)
p: 0.194	GO:0003705	3	RNA polymerase II transcription factor activity, enhancer binding	(ZFHX3 USF2 NFIX)
p: 0.197	GO:0030528	6	transcription regulator activity	(NEO1 HES3 NPAS3 VGLL1 ZFX TWIST1)
p: 0.209	GO:0008092	3	cytoskeletal protein binding	(FARP1 EPB41L4B PKD2L1)
p: 0.216	GO:0008013	3	beta-catenin binding	(PSEN1 TBL1XR1 TCF7L2)
p: 0.221	GO:0003824	5	catalytic activity	(HDDC2 CDYL2 HP ISOC2 HPR)
p: 0.236	GO:0005089	4	Rho guanyl-nucleotide exchange factor activity	(ARHGEF10L FARP1 RASGRF1 ARHGEF11)
p: 0.245	GO:0004252	8	serine-type endopeptidase activity	(TPSD1 HP CELA2A RHBDL3 ELANE PCSK4 HPR CFI)
p: 0.248	GO:0004519	4	endonuclease activity	(ZC3H12C NTHL1 RNASE3 EIF2C2)
p: 0.269	GO:0030955	7	potassium ion binding	(KCNJ4 SLC12A7 HCN2 SLC9A7 KCNQ1 ATP1B1 ATP1A1)
p: 0.26	GO:0008234	4	cysteine-type peptidase activity	(USP34 USP42 USP18 USP27X)
p: 0.286	GO:0005096	8	GTPase activator activity	(TBC1D3H ARHGAP11B TBC1D1 TBC1D2B ARHGAP5 ARHGEF11 ARHGAP8 ARHGDIA)
p: 0.28	GO:0005244	8	voltage-gated ion channel activity	(CACNA1H CLCN2 CACNA1D KCNQ1 KCNJ4 CACNG4 CACNB4 HCN2)
p: 0.318	GO:0005515	233	protein binding	(TMEM8A ISOC2 SQSTM1 AHNAK U2AF1 DAPK3 CDK9 GSC2 DGKZ PHF21B RASGRF1 CACNB4 LRRC8D TSPYL2 SERTAD2 SETD2 IL3RA DGKE QKI ADRA2C HIC1 FGFRL1 POLR2H KIF22 RTN4R CNKSR2 PIGG PHF21A ATXN7L3 MACF1 PODXL2 SCUBE3 NFYB KDELR2 FCGR1A RBM3 UNC5A SNX16 TNK2 CBFA2T3 GABARAP SAMD4B TWIST1 IKBKG GPS1 RORA FXYD1 RCOR2 TUBB2A MEX3D AES UBTF STON1 NACC1 PLEKHA7 EIF4H IPO13 KPNA2 ZNF592 NTHL1 LRRC45 TERT ZNF148 PDLIM2 ATP1A1 POLD1

				CRIPAK APOBEC3G VAMP2 CTBP2 PHF19 CDC42BPB CBLN1 SYNPO2L MLLT10 WEE1 FUS MAST2 LRRC8C IHH SHANK1 CDYL2 FBXL19 CYTH3 POLE BSG CHD4 RYBP PPP4R1 CDK5R1 ZFYVE28 PANX2 PRRG2 RBMX MLL3 NFKBIE C9orf163 IRS4 ABTB1 HIVEP1 TMEM8B STK24 RFC2 NCOA4 MAZ ZCCHC14 PTCH1 SOX8 SBK1 PRPF8 LMCD1 CBWD2 TOM1 HSD17B14 MLLT6 PSMD2 COPG SLC9A7 HCN2 KIAA1609 SF3B14 CHD5 TSSC1 PTBP1 TIMM23 SP3 PNLIPRP1 DOCK3 P4HA2 NSD1 ELK1 BMI1 FAHD1 PARD3 GGA1 GBA CUL4B ICAM4 C20orf54 TRAF3 COBRA1 EIF2C2 TCF25 LPHN1 PHF10 NEO1 PXMP2 LCN1 DPH2 IGSF8 LRRC4B KCNQ1 STRADA RGS19 WAC CAB39 CYTH1 CSF2RA ID2 SRC SLITRK4 TCEA1 CCNF LRRC37A3 PDXDC1 VANGL1 DUSP7 LAT2 ELL TLE2 PPP2R5C ZFAND2B NFATC1 ZBTB32 COMMD3 WHSC1 RPS27 ARHGEF11 GRK5 PFAS MAPK10 MAGEA11 RNF182 TCF3 ZFPM1 ARID1B LAIR1 SMARCC2 MAPK11 DPF1 TLK2 PSMG2 ATP1B1 PORCN KIF13A PIAS4 SSX2IP SLC25A6 HOMER3 FRMPD1 EFNA2 DBNDD1 EXD3 RTN4RL2 STK11 SGOL1 LRRC56 FBXL14 AGR2 WDR4 CCNK NCK2 SHCBP1 KLF10 SKIL CXorf40A PLXND1 PLP2 GRID1 SMPD4 TULP1 PPP2R2A RRAS2 MAGIX CPSF7 PODNL1 SMARCD2 TLK1 LMX1B BMP8B MAPK12 IGF2BP2 RNF157)
p: 0.331	GO:0051015	3	actin filament binding	(TULP1 UXT MACF1)
p: 0.341	GO:0031402	6	sodium ion binding	(SLC9A7 SLC6A8 SLC38A10 HCN2 ATP1A1 ATP1B1)
p: 0.361	GO:0003702	6	RNA polymerase II transcription factor activity	(USF2 LCOR FOXK2 SOX8 TCF7L2 SPIB)
p: 0.361	GO:0003743	3	translation initiation factor activity	(EIF4H EIF2C2 EIF5)
p: 0.392	GO:0004672	3	protein kinase activity	(POLR2H CDK5R1 STRADA)
p: 0.425	GO:0005525	16	GTP binding	(RAP2C EEFSEC DOCK3 EIF5 RRAS2 DIRAS2 NBR2 TUBA3E RAC1 RAB23 RASD2 ARHGAP5 HRAS INSR TUBB2A GNAO1)

p: 0.444	GO:0004386	5	helicase activity	(CHD4 SRCAP SMARCA4 CHD5 HELZ)
p: 0.44	GO:0008017	3	microtubule binding	(GABARAP MACF1 UXT)
p: 0.451	GO:0003779	12	actin binding	(ABLIM2 INF2 MACF1 CLMN DBN1 SYNPO2L TNNI3 COTL1 TNNT3 DIAPH1 SPIRE2 BCL7B)
p: 0.463	GO:0030145	7	manganese ion binding	(DYRK2 NPEPL1 INSR GALNT10 B3GAT1 STK11 B4GALT7)
p: 0.485	GO:0005215	12	transporter activity	(GRID1 AQP7 SLC25A6 SYPL2 CPNE7 LCN6 LCN10 SYT7 HIATL2 SLC12A7 LCN1 CRABP2)
p: 0.493	GO:0005085	5	guanyl-nucleotide exchange factor activity	(DOCK3 ARHGEF10L ARHGEF11 FARP1 KNDC1)
p: 0.496	GO:0000166	79	nucleotide binding	(DAPK3 RAP2C ADRBK2 SKI EIF5 PFAS TLK1 DGKZ SMARCA4 KIF13A POLE MAPK11 IRAK3 CMPK1 PRKCE MRPL23 EEFSEC STK24 CHD4 ATP1A1 STRADA NAIP PRKACB SETD1B CDK9 ACSF3 MAPK12 CDC42BPB SKIL NME6 HRAS DYRK2 MAP4K4 RBMX CAMKK1 ATP6V1A SBK1 HCN2 MAST2 TNK2 ITPKA RBM3 U2AF1 RRAS2 POLD1 KIF22 IGF2BP2 STK11 CPSF7 GRK5 EIF4H HELZ CBWD2 SRC PTBP1 RAB23 CBWD3 MSI2 RFC2 PGS1 HSPA12A WEE1 SF3B14 TUBA3E ROR2 RASD2 INSR TLK2 RAC1 GNAO1 SRCAP DCAKD TUBB2A CBWD5 CHD5 MAPK10 DIRAS2 DGKE PRKACA)
p: 0.511	GO:0003723	23	RNA binding	(MAZ CPSF7 SETD1B IGF2BP2 RBMX MEX3D APOBEC3G PUSL1 PIWIL3 ATXN1 TRMT1 RPL19 FUS PTBP1 U2AF1 QKI MRPL23 EIF4H PRPF8 DIS3L2 SF3B14 TRUB2 RBM3)
p: 0.516	GO:0031072	3	heat shock protein binding	(DNAJB12 DNAJB6 DNAJC30)
p: 0.558	GO:0016779	3	nucleotidyltransferase activity	(POLD1 TERT POLE)
p: 0.568	GO:0016829	4	lyase activity	(PDXDC1 NTHL1 FTCD MLYCD)
p: 0.578	GO:0008565	3	protein transporter activity	(KPNA2 TIMM23 IPO13)
p: 0.582	GO:0008289	4	lipid binding	(PLEKHA4 CRABP2 GPIHBP1 PPARA)
p: 0.628	GO:0005506	6	iron ion binding	(NTHL1 TYW1B PDF SLC25A28 P4HA2 LEPREL2)
p: 0.629	GO:0016853	4	isomerase activity	(EBP TRUB2 TXNDC5 PUSL1)

p: 0.649	GO:0004930	13	G-protein coupled receptor activity	(GPRC5B GPR157 GPR160 GPR162 GPR27 CELSR1 AVPR2 OPRL1 LPHN1 GPR173 LPAR1 ADRA2C P2RY8)
p: 0.668	GO:0003674	21	molecular_function	(GNB1L NBR2 TMEM129 C16orf57 MAGEA12 ABHD11 JPH3 KNDC1 GPKOW LMCD1 FRAT2 CENPV NTN3 GGA1 TMEM8A DCUN1D1 SMA5 RFNG BRD3 PMS2L2 CUL4B)
p: 0.685	GO:0000287	17	magnesium ion binding	(MAPK12 IRAK3 TNK2 FOXK2 MAST2 PRPSAP2 NME6 FAHD1 ATP1A1 CDC42BPB PRKACB ZC3H12C WEE1 SMPD4 SMPD3 STK11 DYRK2)
p: 0.695	GO:0015293	4	symporter activity	(SLC6A6 SLC12A7 SLC6A4 SLC6A8)
p: 0.715	GO:0046982	7	protein heterodimerization activity	(TCF3 CEBPG PPARG IKBKG STX1A USF2 IRAK3)
p: 0.749	GO:0017124	3	SH3 domain binding	(SIRPA QKI DOCK3)
p: 0.781	GO:0005125	5	cytokine activity	(CD70 CMTM3 BMP8B CMTM4 IL34)
p: 0.781	GO:0008083	5	growth factor activity	(IL34 MDK OSGIN1 FGF17 BMP8B)
p: 0.797	GO:0005529	5	sugar binding	(BSG CLC SFTPA1 LPHN1 GALNT10)
p: 0.806	GO:0008233	16	peptidase activity	(ELANE RHBDL3 PSEN1 TPSD1 CFI USP42 USP27X ASTL USP18 CELA2A PCSK4 USP7 PSMA8 USP34 IHH NPEPL1)
p: 0.822	GO:0051082	3	unfolded protein binding	(DNAJC30 UXT DNAJB12)
p: 0.832	GO:0003676	8	nucleic acid binding	(HELZ EXD3 RNASE3 PHF19 ZCCHC14 GPKOW SSX7 ZC3H12C)
p: 0.842	GO:0005524	53	ATP binding	(GRK5 SRC CDK9 ATP1A1 PGS1 ADRBK2 ATP6V1A SRCAP ROR2 PRKACA INSR RFC2 SBK1 STK11 MORC4 MAPK10 WEE1 KIF22 PRKACB MAST2 IRAK3 MAP4K4 PRKCE TNK2 MAPK12 KIF13A STK24 TLK2 CDC42BPB CHD4 DCAKD STRADA PFAS CAMKK1 DAPK3 CBWD3 CBWD5 CBWD2 HELZ CMPK1 PMS2L2 SMARCA4 ACSF3 TLK1 DYRK2 MAPK11 HSPA12A CHD5 ATP5D NME6 DGKZ ITPKA DGKE)
p: 0.848	GO:0008201	3	heparin binding	(MDK ELANE PTCH1)
p: 0.874	GO:0042803	11	protein homodimerization activity	(TCF3 USF2 TP53I3 IRAK3 APOBEC3G GPD1L SLC9A7 TERT G6PD SLC6A4 CBLN1)
p: 0.886	GO:0005488	15	binding	(SFTPA1 FARP1 LCN6 NCAPG2 GLYR1 TRIP12 TP53I3 SLC25A28 SEL1L3

				GMEB1 EPB41L4B GPD1L LCN10 G6PD GFOD1)
p: 0.889	GO:0005216	4	ion channel activity	(FXYD1 GRID1 HTR3D CHRNA4)
p: 0.893	GO:0003735	4	structural constituent of ribosome	(MRPL23 RPS27 RPL19 MRPL41)
p: 0.912	GO:0016787	31	hydrolase activity	(CLC SMPD3 POLD1 PDF ZC3H12C EIF2C2 APOBEC3G CHD4 CHD5 PNLIPRP1 SERHL SRCAP EXD3 FAHD1 ATP6V1A RNASE3 PNPLA7 LPIN2 HDAC3 INPP5A ACOT1 DIS3L2 SMARCA4 ATP1A1 HELZ DUSP7 DNASE1L3 ABHD11 SSU72 SMPD4 LPPR3)
p: 0.94	GO:0016491	11	oxidoreductase activity	(CTBP2 TP53I3 GLYR1 TYW1B AKR1C1 SETD2 NECAB2 G6PD CTBP1 GFOD1 HSD17B14)
p: 0.982	GO:0005198	3	structural molecule activity	(TUBA3E TUBB2A COPG)
p: 0.995	GO:0016874	5	ligase activity	(TRIP12 PFAS ACSF3 TTLL13 UBR3)
p: 0.996	GO:0042802	5	identical protein binding	(BLOC1S2 DVL1 ARHGDIA SMARCA4 ATXN1)
p: 0.998	GO:0004872	33	receptor activity	(MED16 GPR160 FGFRL1 INSR GPR27 UNC5A AVPR2 IL3RA IL17RD LAIR1 GPR173 P2RY8 LRP5 TNFRSF18 FCGR1A KDELR2 LSR GPR162 PLXND1 SEMA3F CSF2RA GPRC5B OPRL1 RTN4R TSPO PTCH1 GRID1 HTR3D NEO1 LPAR1 ROR2 ADRA2C RTN4RL2)
p: 0.998	GO:0005509	23	calcium ion binding	(CACNB4 LTBP2 MACF1 CACNA1H CACNG4 LPCAT1 SCUBE3 NOTCH2NL SYT7 NUCB1 CDK5R1 CELSR1 PKD2L1 DNASE1L3 EGFL7 PRRG2 RHBDL3 NECAB2 CALML6 SFTPA1 GALNT10 CACNA1D FAHD1)
p: 0	GO:0003714	14	transcription corepressor activity	(HDAC3 TLE1 NSD1 RYBP SKIL ATN1 TBL1XR1 ZEB1 PIAS4 ZBTB32 AES LMCD1 SKI RCOR2)
p: 0	GO:0042169	6	SH2 domain binding	(SQSTM1 SRC SHCBP1 INSR LAT2 ARHGAP5)

APÊNDICE E – Processos biológicos onde os 853 genes estão mais enriquecidos

Valor-P	ID Gene Ontology	Nº Genes	Ontologia	Genes
p: 0.001	GO:0001889	7	liver development	(CEBPG HES1 LSR SP3 HDAC3 EP300 CTNNB1)
p: 0.001	GO:0006333	7	chromatin assembly or disassembly	(CHD5 KAT5 SMARCC2 CHD4 MSL3 CDYL CDYL2)
p: 0.001	GO:0010552	8	positive regulation of specific transcription from RNA polymerase II promoter	(CTNNB1 TCF3 HES3 USF2 HES1 PPARG TCF7L2 SMARCA4)
p: 0.001	GO:0051091	5	positive regulation of transcription factor activity	(PPARG CEBPG CTNNB1 SMARCA4 EP300)
p: 0.002	GO:0048469	5	cell maturation	(PPARG CTNNB1 SOX8 HES1 IHH)
p: 0.003	GO:0000731	3	DNA synthesis during DNA repair	(CDKN2D POLD1 POLE)
p: 0.004	GO:0007188	4	G-protein signaling, coupled to cAMP nucleotide second messenger	(GRK5 AVPR2 PRKACB GCGR)
p: 0.004	GO:0050872	3	white fat cell differentiation	(CTBP2 CTBP1 PPARG)
p: 0.004	GO:0060070	4	Wnt receptor signaling pathway through beta-catenin	(TBL1XR1 DVL1 CTNNB1 TCF7L2)
p: 0.006	GO:0055088	3	lipid homeostasis	(USF2 IRS2 PPARG)
p: 0.008	GO:0019216	3	regulation of lipid metabolic process	(LSR PPARG IRS2)
p: 0.016	GO:0045893	9	positive regulation of transcription, DNA-dependent	(SMARCC2 NSD1 TCF3 ATXN7L3 NFYB ARHGEF11 NCOA4 SERTAD2 KAT5)
p: 0.017	GO:0001764	6	neuron migration	(VAX1 PSEN1 SOX1 CDK5R1 TWIST1 LMX1B)
p: 0.019	GO:0035115	3	embryonic forelimb morphogenesis	(ZNF358 TWIST1 CRABP2)
p: 0.01	GO:0045725	3	positive regulation of glycogen biosynthetic process	(IRS2 DYRK2 INSR)
p: 0.022	GO:0040008	6	regulation of growth	(BRMS1L CAPRIN2 OSGIN1 KAT5 TMEM8B PTCH1)
p: 0.022	GO:0050771	3	negative regulation of axonogenesis	(RTN4R PSEN1 ARHGDIA)
p: 0.024	GO:0046320	3	regulation of fatty acid oxidation	(CAB39 STRADA STK11)
p: 0.025	GO:0006297	3	nucleotide-excision repair, DNA gap filling	(RFC2 POLE POLD1)

p: 0.025	GO:0050772	3	positive regulation of axonogenesis	(SKIL ARHGDIA METRN)
p: 0.026	GO:0031016	3	pancreas development	(TCF7L2 IHH CTNNB1)
p: 0.034	GO:0050852	3	T cell receptor signaling pathway	(PSEN1 IKBKG CACNB4)
p: 0.036	GO:0070555	3	response to interleukin-1	(SRC AES IRAK3)
p: 0.038	GO:0006468	25	protein amino acid phosphorylation	(TNK2 CDK5R1 MAPK11 CTBP1 ROR2 STRADA STK11 PRKCE CDC42BPB MAP4K4 DAPK3 TLK1 PRKACB DYRK2 CDK9 SRC MAPK10 TLK2 GRK5 WEE1 IRAK3 STK24 CAMKK1 MAST2 ADRBK2)
p: 0.039	GO:0006916	13	anti-apoptosis	(NME6 SQSTM1 CDKN2D LHX4 TERT TNFRSF18 NAIP PSEN1 FAIM3 TCF7L2 HDAC3 ARHGDIA TXNDC5)
p: 0.03	GO:0007528	3	neuromuscular junction development	(AGRN CACNB4 DVL1)
p: 0.041	GO:0007163	3	establishment or maintenance of cell polarity	(PARD3 MACF1 CDC42BPB)
p: 0.042	GO:0007265	6	Ras protein signal transduction	(RREB1 MAPK11 RRAS2 MAPK12 HRAS SRC)
p: 0.046	GO:0001569	3	patterning of blood vessels	(PLXND1 CTNNB1 IHH)
p: 0.047	GO:0045596	3	negative regulation of cell differentiation	(IHH TWIST1 SKIL)
p: 0.052	GO:0032313	5	regulation of Rab GTPase activity	(TBC1D3H TBC1D29 TBC1D28 TBC1D1 TBC1D2B)
p: 0.055	GO:0006357	12	regulation of transcription from RNA polymerase II promoter	(FOXK2 SRCAP SMARCC2 MED16 ONECUT3 SPIB CEBPG GSC2 SMARCD2 MED13L VEZF1 CHD4)
p: 0.056	GO:0016055	7	Wnt receptor signaling pathway	(FRAT2 MACF1 PORCN LRP5 PIAS4 AES TLE2)
p: 0.063	GO:0051789	4	response to protein stimulus	(KLF10 SETD2 ID2 INSR)
p: 0.065	GO:0048538	3	thymus development	(CACNB4 CTNNB1 PSEN1)
p: 0.068	GO:0045892	6	negative regulation of transcription, DNA-dependent	(HMX1 SMARCC2 FOXN3 TBL1XR1 CHMP1A DNAJB6)
p: 0.069	GO:0030521	4	androgen receptor signaling pathway	(NCOA4 CTNNB1 MED16 KAT5)
p: 0.06	GO:0001702	3	gastrulation with mouth forming second	(NSD1 LRP5 CTNNB1)
p: 0.079	GO:0001934	3	positive regulation of protein amino acid phosphorylation	(PSEN1 IL34 INSR)
p: 0.083	GO:0006874	5	cellular calcium ion homeostasis	(FXYD1 TNNI3 SYPL2 ELANE CACNB4)

p: 0.086	GO:0030512	3	negative regulation of transforming growth factor beta receptor signaling pathway	(SKIL BAMBI SKI)
p: 0.091	GO:0006139	5	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	(PRPSAP2 CMPK1 EXD3 POLD1 POLE)
p: 0.092	GO:0045944	17	positive regulation of transcription from RNA polymerase II promoter	(SOX8 SQSTM1 TCF3 LMX1B USF2 PPARA ELK1 ZNF148 IHH ZEB1 TBL1XR1 RORA EP300 NFIC CREBBP AGRN TCEA1)
p: 0.093	GO:0009411	3	response to UV	(ELANE CDKN2D POLD1)
p: 0.093	GO:0030335	5	positive regulation of cell migration	(RRAS2 IRS2 ARHGAP5 ROR2 INSR)
p: 0.094	GO:0042733	3	embryonic digit morphogenesis	(CTNNB1 LRP5 RAB23)
p: 0.095	GO:0030324	5	lung development	(EP300 CLCN2 SP3 HES1 CTNNB1)
p: 0.102	GO:0006511	9	ubiquitin-dependent protein catabolic process	(USP27X PSMA8 USP34 USP7 UBR3 CUL4B USP18 SQSTM1 USP42)
p: 0.109	GO:0001756	3	somitogenesis	(ROR2 EP300 PSEN1)
p: 0.109	GO:0042593	4	glucose homeostasis	(PTCH1 PPARG INSR TCF7L2)
p: 0.113	GO:0001843	3	neural tube closure	(TWIST1 PTCH1 CELSR1)
p: 0.115	GO:0006486	5	protein amino acid glycosylation	(FUT7 ST6GALNAC2 PSEN1 B3GNT7 LARGE)
p: 0.115	GO:0007254	3	JNK cascade	(ROR2 GPS1 MAPK10)
p: 0.115	GO:0043433	3	negative regulation of transcription factor activity	(TCF7L2 ID2 CEBPG)
p: 0.11	GO:0009887	8	organ morphogenesis	(EP300 TLE1 RFNG AES HRAS HEY1 TLE2 LHX4)
p: 0.121	GO:0007507	8	heart development	(CTNNB1 TNNI3 EP300 IRX4 ID2 TCF25 PSEN1 ZFPM1)
p: 0.124	GO:0007411	5	axon guidance	(VAX1 EFNA2 UNC5A CDK5R1 NTN3)
p: 0.124	GO:0043410	3	positive regulation of MAPKKK cascade	(INSR HRAS CTNNB1)
p: 0.125	GO:0006605	3	protein targeting	(HOMER3 GABARAP LTBP2)
p: 0.134	GO:0006928	7	cell motion	(IGSF8 ARHGEF11 STRBP MACF1 ARHGDIA PALM RAC1)
p: 0.135	GO:0045165	3	cell fate commitment	(PPARG HES1 SOX8)
p: 0.13	GO:0030308	6	negative regulation of cell growth	(PPARG CAPRIN2 TSPYL2 CDKN2D OSGIN1 SERTAD2)
p: 0.146	GO:0007205	3	activation of protein kinase C activity by G-protein coupled receptor protein signaling pathway	(PARD3 DGKE DGKZ)

p: 0.147	GO:0006917	10	induction of apoptosis	(TRAF3 KLF10 PPARG PRKCE IKBKG ZNF443 DAPK3 NACC1 DPF1 CD70)
p: 0.149	GO:0001501	7	skeletal system development	(CTNNB1 KLF10 ANKH TWIST1 EBP IHH ROR2)
p: 0.149	GO:0007243	6	protein kinase cascade	(DAPK3 PRKACB MAPK11 MAP4K4 SRC PRKACA)
p: 0.154	GO:0008654	4	phospholipid biosynthetic process	(MBOAT1 DGKE PGS1 LPCAT1)
p: 0.157	GO:0007519	3	skeletal muscle tissue development	(EP300 SKIL IGSF8)
p: 0.157	GO:0030097	4	hemopoiesis	(EBP CTNNB1 BMI1 ZBTB32)
p: 0.15	GO:0007242	15	intracellular signaling cascade	(NFATC1 GNB1L LAT2 DGKZ CDC42BPB PLEKHM1P DVL1 PRKCE MAPK12 TLK1 DGKE MLL3 SQSTM1 TLK2 PSEN1)
p: 0.162	GO:0006936	6	muscle contraction	(GNAO1 KCNQ1 FXYD1 SLC6A8 HCN2 CACNA1H)
p: 0.164	GO:0016481	8	negative regulation of transcription	(TLE1 NACC1 HES3 RCOR2 ATXN1 ID2 COBRA1 PIAS4)
p: 0.167	GO:0007050	7	cell cycle arrest	(MAPK12 SKIL HRAS MACF1 TCF7L2 STK11 CDKN2D)
p: 0.171	GO:0030326	3	embryonic limb morphogenesis	(PSEN1 SKI PTCH1)
p: 0.173	GO:0006865	3	amino acid transport	(SLC38A10 SLC43A2 SLC6A6)
p: 0.173	GO:0030900	4	forebrain development	(CTNNB1 PSEN1 GNAO1 SOX1)
p: 0.176	GO:0001570	3	vasculogenesis	(EGFL7 TNNI3 QKI)
p: 0.177	GO:0032526	3	response to retinoic acid	(CDKN2D PTCH1 MICB)
p: 0.193	GO:0016337	4	cell-cell adhesion	(COL6A2 PSEN1 CTNNB1 ICAM4)
p: 0.197	GO:0000079	3	regulation of cyclin-dependent protein kinase activity	(CDKN2D CCNK CDK5R1)
p: 0.198	GO:0031175	3	neuron projection development	(RASGRF1 GNAO1 CDK5R1)
p: 0.19	GO:0006886	10	intracellular protein transport	(STON1 KDELR2 GGA1 STX1A TIMM23 COPG TOM1 IPO13 KPNA2 TLK1)
p: 0.202	GO:0008033	4	tRNA processing	(TRUB2 TYW1B PUSL1 TRMT1)
p: 0.209	GO:0001568	3	blood vessel development	(EGFL7 TCF7L2 PSEN1)
p: 0.213	GO:0051260	4	protein homooligomerization	(AKR1C1 SCUBE3 MUC20 NACC1)
p: 0.216	GO:0006469	3	negative regulation of protein kinase activity	(CRIPAK PSEN1 DVL1)
p: 0.222	GO:0006915	20	apoptosis	(CTNNB1 HRAS TSPO SLC25A6 SQSTM1 MFSD10 NME6 EP300 TWIST1 TNFRSF18 UNC5A ARHGEF11 RYBP DNASE1L3 RAC1 DYRK2 RASGRF1 DAPK3 MRPL41 NAIP)
p: 0.223	GO:0006836	3	neurotransmitter transport	(STX1A SLC6A6 SLC6A8)

p: 0.238	GO:0006417	4	regulation of translation	(EIF2C2 NCK2 QKI PIWIL3)
p: 0.245	GO:0007219	3	Notch signaling pathway	(HEY1 HES1 NOTCH2NL)
p: 0.247	GO:0008544	5	epidermis development	(PTCH2 PPARA CRABP2 POU3F1 PTCH1)
p: 0.257	GO:0007067	9	mitosis	(KIF22 WEE1 CCNK CENPV NCAPG2 TUBB2A CCNF SGOL1 RCC2)
p: 0.269	GO:0044419	14	interspecies interaction between organisms	(SLC25A6 KAT5 CTBP1 EP300 SRCAP MICB EIF4H USP7 CRTC3 CREBBP KPNA2 APOBEC3G SRC IKBKG)
p: 0.26	GO:0006814	7	sodium ion transport	(SLC6A8 SLC9A7 SLC38A10 ATP1B1 SLC12A7 ATP1A1 HCN2)
p: 0.277	GO:0007049	21	cell cycle	(STK11 CENPV CUL4B MRPL41 FOXN3 CHMP1A SGOL1 MAPK12 RCC2 WEE1 CDKN2D GPS1 NCAPG2 TLK2 TLK1 SMPD3 PARD3 CCNF EP300 STRADA CCNK)
p: 0.28	GO:0006367	4	transcription initiation from RNA polymerase II promoter	(POLR2H MED16 CDK9 MAZ)
p: 0.296	GO:0007399	16	nervous system development	(IGSF8 IGSF9 CRIM1 CBLN1 DPF1 RFNG RAB23 DBN1 CYP46A1 EP300 HEY1 METRN NAIP FGF17 AHNAK MDK)
p: 0.302	GO:0051291	3	protein heterooligomerization	(HRAS SCUBE3 CTNNB1)
p: 0.308	GO:0001701	6	in utero embryonic development	(MAFG ELL HES1 LMX1B HES3 SOX8)
p: 0.308	GO:0030182	3	neuron differentiation	(DAPK3 LMX1B TUBB2A)
p: 0.309	GO:0007389	3	pattern specification process	(ZEB1 RFNG PTCH1)
p: 0.314	GO:0043627	3	response to estrogen stimulus	(PPARG CRIPAK CTNNB1)
p: 0.315	GO:0008624	5	induction of apoptosis by extracellular signals	(RAC1 ARHGEF11 PSEN1 SQSTM1 RASGRF1)
p: 0.328	GO:0009653	5	anatomical structure morphogenesis	(WHSC1 RAC1 IGF2BP2 GSC2 TWIST1)
p: 0.329	GO:0009790	4	embryonic development	(PSEN1 LSR UBR3 CDK5R1)
p: 0.331	GO:0008285	13	negative regulation of cell proliferation	(CDKN2D PPARG STK11 ROR2 HRAS PPP2R5C KLF10 CTBP1 NCK2 CTBP2 SKI ZEB1 FGFRL1)
p: 0.332	GO:0043434	3	response to peptide hormone stimulus	(PNLIPRP1 BSG IRS2)
p: 0.333	GO:0006368	3	RNA elongation from RNA polymerase II promoter	(ELL POLR2H CDK9)
p: 0.354	GO:0009755	3	hormone-mediated signaling	(PRKACB PRKACA GCGR)

p: 0.355	GO:0007626	3	locomotory behavior	(RASD2 NPAS3 ANKH)
p: 0.35	GO:0035023	4	regulation of Rho protein signal transduction	(ARHGEF11 ARHGEF10L RASGRF1 FARP1)
p: 0.361	GO:0008283	12	cell proliferation	(CD70 TSPO CDK9 LMX1B RPS27 CREBBP CBFA2T3 CDK5R1 FRAT2 TCF7L2 IGSF8 RASGRF1)
p: 0.365	GO:0016477	3	cell migration	(BAMBI NCK2 CTHRC1)
p: 0.375	GO:0006629	11	lipid metabolic process	(HSD17B14 PPARA ACOT1 LRP5 ACSF3 G6PD LPIN2 GBA CPNE7 PNPLA7 CYP46A1)
p: 0.377	GO:0006950	5	response to stress	(SQSTM1 MAPK11 ZNF443 SNN MAP4K4)
p: 0.377	GO:0014070	4	response to organic cyclic substance	(PTCH1 GNAO1 CTNNB1 PPARG)
p: 0.379	GO:0008217	3	regulation of blood pressure	(ATP1A1 PPARG GCGR)
p: 0.37	GO:0016044	3	membrane organization	(COPG TXNDC5 VAMP2)
p: 0.381	GO:0006396	3	RNA processing	(ATXN1 TRUB2 RBM3)
p: 0.387	GO:0007010	3	cytoskeleton organization	(CDC42BPB ABLIM2 PALM)
p: 0.399	GO:0006366	6	transcription from RNA polymerase II promoter	(NFATC1 RREB1 DVL1 CCNK NFIX MSL3)
p: 0.39	GO:0006821	3	chloride transport	(FXYD1 SLC12A7 CLCN2)
p: 0.401	GO:0045941	6	positive regulation of transcription	(PPARA PPARG CREBBP USF2 BLOC1S2 ELK1)
p: 0.413	GO:0032355	3	response to estradiol stimulus	(INSR PTCH1 IHH)
p: 0.415	GO:0001558	3	regulation of cell growth	(FGFRL1 ARHGEF11 CRIM1)
p: 0.424	GO:0006334	4	nucleosome assembly	(TSPYL2 SET HP1BP3 HIST2H2BF)
p: 0.428	GO:0007018	4	microtubule-based movement	(TUBB2A KIF13A TUBA3E KIF22)
p: 0.434	GO:0051384	3	response to glucocorticoid stimulus	(PNLIPRP1 INSR EP300)
p: 0.437	GO:0006813	7	potassium ion transport	(KCNQ1 HCN2 SLC12A7 ATP1B1 ATP1A1 KCNJ4 SLC9A7)
p: 0.458	GO:0006968	3	cellular defense response	(VEZF1 ZNF148 FAIM3)
p: 0.477	GO:0005975	9	carbohydrate metabolic process	(B4GALT7 B3GAT1 INSR GBA G6PD SMA5 CHST3 ATHL1 GPD1L)
p: 0.478	GO:0006816	5	calcium ion transport	(CACNB4 CACNG4 CACNA1H CACNA1D CHRNA4)
p: 0.486	GO:0007264	8	small GTPase mediated signal transduction	(DIRAS2 RAP2C ARHGAP5 RGS19 RAC1 TNK2 RASD2 RAB23)

p: 0.487	GO:0007275	37	multicellular organismal development	(IGSF9 DBN1 QKI PIWIL3 BMP8B UNC5A SEMA3F STRBP TWIST1 FOXD4 DVL1 KLF3 VANGL1 RREB1 METRN HMX1 LRP5 RYBP EGFL7 PLXND1 OSGIN1 SMPD3 MDK HEY1 CELSR1 FRAT2 ID2 LMX1B HIC1 MSL3 TLE1 NOTCH2NL CATSPERG AES ROR2 CLC VAX1)
p: 0.49	GO:0016192	7	vesicle-mediated transport	(CHMP1A COPG KDELR2 GGA1 CYTH1 VAMP2 CYTH3)
p: 0.507	GO:0006812	3	cation transport	(HCN2 SLC9A7 PKD2L1)
p: 0.522	GO:0006952	3	defense response	(HCP5 CX3CL1 HP)
p: 0.52	GO:0008219	5	cell death	(PSEN1 FUS ATN1 ATXN1 LMX1B)
p: 0.549	GO:0007160	3	cell-matrix adhesion	(TMEM8B CTNNB1 RAPH1)
p: 0.554	GO:0007420	4	brain development	(VAX1 SBK1 CDK5R1 IRS2)
p: 0.566	GO:0008284	10	positive regulation of cell proliferation	(HES1 IL34 HRAS FGF17 INSR PDF BLOC1S2 NACC1 ID2 LRP5)
p: 0.567	GO:0042493	8	response to drug	(PTCH1 HDAC3 PPARG ATP1A1 EP300 AQP7 GRK5 GNAO1)
p: 0.603	GO:0006260	5	DNA replication	(SET RFC2 NFIC NFIX NFIB)
p: 0.604	GO:0045087	4	innate immune response	(FCGR1A PPARG CFI APOBEC3G)
p: 0.613	GO:0007154	3	cell communication	(INPP5A ZCCHC14 SNX16)
p: 0.614	GO:0006412	5	translation	(ABTB1 MRPL23 MRPL41 EIF4H PDF)
p: 0.626	GO:0006935	5	chemotaxis	(HRAS CMTM4 CX3CL1 PLP2 CMTM3)
p: 0.628	GO:0006461	4	protein complex assembly	(PARD3 PTCH2 CREBBP SLC9A3R2)
p: 0.63	GO:0001666	5	response to hypoxia	(EP300 PPARA CREBBP ATP1B1 CHRNA4)
p: 0.649	GO:0006355	40	regulation of transcription, DNA-dependent	(ZNF443 IRX4 MLLT6 MEIS3P1 NR2F6 CREBBP ELK1 SOX1 ZNF630 NPAS3 HIC1 HEY1 VAX1 LHX4 NFIC ZNF138 RREB1 HES1 FOXD4 ZNF444 MLL3 IRX2 ZNF114 LMX1B RORA NFIB NFATC1 VGLL1 EP300 SSX7 ZFHX3 PBX3 POU3F1 NFIX CBFA2T3 ZSCAN4 MAFG SP3 CRABP2 PPARA)
p: 0.654	GO:0042742	3	defense response to bacterium	(MICA RNASE3 DEFB131)
p: 0.677	GO:0007204	3	elevation of cytosolic calcium ion concentration	(SAA1 OPRL1 LPAR1)
p: 0.681	GO:0008152	18	metabolic process	(NTHL1 ACSF3 CDYL2 ATP1A1 NAT8L ATHL1 NAT14 LPCAT1 ISOC2 GBA CDYL CENPV UGT2B11 PNPLA7 PGS1 PIGG SMPDL3B FAHD1)
p: 0.683	GO:0007218	3	neuropeptide signaling pathway	(CELSR1 LPHN1 BAI2)
p: 0.686	GO:0006457	5	protein folding	(DNAJB12 DNAJB6 ALG12 UXT DNAJC30)

p: 0.689	GO:0042981	4	regulation of apoptosis	(SKIL BLOC1S2 TRAF3 PSMG2)
p: 0.693	GO:0030036	4	actin cytoskeleton organization	(ARHGEF11 INF2 HRAS DIAPH1)
p: 0.708	GO:0051301	8	cell division	(CCNF CENPV WEE1 SGOL1 CHMP1A NCAPG2 RCC2 CCNK)
p: 0.713	GO:0007267	9	cell-cell signaling	(CD70 MLLT4 FGF17 ADRA2C IHH HCN2 EFNA2 DLGAP4 KLF10)
p: 0.719	GO:0006811	19	ion transport	(FXYD1 CACNB4 SLC38A10 SLC12A7 SLC25A28 KCNQ1 SLC6A8 PLP2 CLCN2 CHRNA4 CACNA1D KCNJ4 GRID1 CACNA1H ATP5D ATP6V1A ATP1B1 HTR3D CACNG4)
p: 0.724	GO:0008380	8	RNA splicing	(SF3B14 RBMX QKI CPSF7 FUS POLR2H U2AF1 PTBP1)
p: 0.739	GO:0007417	3	central nervous system development	(ATN1 CELSR1 ZEB1)
p: 0.749	GO:0006974	8	response to DNA damage stimulus	(MICA PSEN1 TLK2 EPC2 TLK1 NTHL1 CUL4B POLE)
p: 0.769	GO:0007517	3	muscle organ development	(CACNA1H ZFHX3 MAPK12)
p: 0.76	GO:0006414	3	translational elongation	(EEFSEC RPS27 RPL19)
p: 0.795	GO:0006897	3	endocytosis	(TOM1 LRP5 HRAS)
p: 0.79	GO:0006979	3	response to oxidative stress	(MICB PSEN1 CHRNA4)
p: 0.81	GO:0008150	20	biological_process	(TMEM129 BCL7B RBMX MACF1 NBR2 BRD3 PMS2L2 KNDC1 SMA5 DCUN1D1 TMEM8A FAM189B MUC17 MAGEA12 GPKOW PLEKHA4 LMCD1 C16orf57 COTL1 ABHD11)
p: 0.828	GO:0001525	3	angiogenesis	(EGFL7 HDAC3 VEZF1)
p: 0.842	GO:0030154	16	cell differentiation	(OSGIN1 CTBP2 PIWIL3 VAX1 CTBP1 BMP8B CATSPERG NOTCH2NL ROR2 CBFA2T3 RFNG SQSTM1 IGSF9 DBN1 STRBP MDK)
p: 0.845	GO:0006810	11	transport	(SPIRE2 LCN10 LCN6 QKI SYT7 LCN1 ANKH CRABP2 NDUFB7 FAM21A SYPL2)
p: 0.868	GO:0055114	16	oxidation reduction	(CYP46A1 CTBP1 CCS GPD1L TECR TYW1B G6PD CTBP2 AKR1C1 GLYR1 GFOD1 LEPREL2 HSD17B14 P4HA2 SETD2 TP53I3)
p: 0.87	GO:0006281	4	DNA repair	(CUL4B EPC2 CHRNA4 KIF22)
p: 0.87	GO:0006397	6	mRNA processing	(SSU72 RBMX SF3B14 PTBP1 QKI CPSF7)
p: 0.887	GO:0055085	16	transmembrane transport	(SLC25A28 SLC12A7 CACNA1D AQP7 SLC25A6 POM121 CLCN2 HIATL2 HCN2 CACNA1H SLC43A2 TIMM23 KCNQ1 MFSD7 SLC9A7 MFSD10)

p: 0.907	GO:0006464	3	protein modification process	(TRIP12 B4GALT7 TTLL13)
p: 0.907	GO:0007268	4	synaptic transmission	(CBLN1 CTNNB1 STX1A GABARAP)
p: 0.909	GO:0007165	59	signal transduction	(PPP4R1 ARHGAP11B GPR173 RRAS2 RAPH1 LPAR1 MLLT4 PRKACB TLE2 TRAF3 GPRC5B OPRL1 ARHGAP8 LRP5 DIRAS2 CHRNA4 PKD2L1 CD70 CRABP2 IRS2 GRK5 ARHGAP5 ITPKA IRS4 AGRN ADRBK2 P2RY8 RAB23 ADRA2C RTN2 ROR2 STK24 PTCH1 RASD2 GNAO1 PLXND1 AVPR2 GPR160 MDK NPAS3 CREBBP NCK2 IRAK3 FGF17 NR2F6 PPP2R5C TNFRSF18 PPP2R2A GPR27 GPR162 RORA RHBDL3 MAPK10 PASD1 PPARG UNC5A FCGR1A TLE1 RAP2C)
p: 0.924	GO:0007283	7	spermatogenesis	(CDYL CATSPERG PTCH2 QKI AQP7 PIWIL3 STRBP)
p: 0.936	GO:0006508	12	proteolysis	(PCSK4 ASTL NPEPL1 CFI HP ELANE LCN1 TPSD1 IHH HPR ADAM11 CELA2A)
p: 0.964	GO:0006955	10	immune response	(LAIR1 ZEB1 LAT2 FAIM3 CD70 CX3CL1 SQSTM1 CEBPG MICB IKBKG)
p: 0.979	GO:0007601	3	visual perception	(PRCD TULP2 PRPF8)
p: 0.995	GO:0007155	12	cell adhesion	(PODXL2 ARHGAP5 HES1 TMEM8B SSX2IP MLLT4 TMEM8A NEO1 RAC1 SIRPA CELSR1 CX3CL1)
p: 0.996	GO:0019941	7	modification-dependent protein catabolic process	(PIAS4 FBXO36 TRIP12 FBXL19 DCUN1D1 FBXL14 TBL1XR1)
p: 0.999	GO:0015031	7	protein transport	(KIF13A RAB23 POM121 PSEN1 GABARAP SNX16 CHMP1A)
p: 0	GO:0000122	22	negative regulation of transcription from RNA polymerase II promoter	(PHF21A LCOR TWIST1 VAX1 CTBP1 ZBTB32 TCF25 ATN1 NSD1 ZFHX3 PIAS4 ZNF148 HIVEP1 TLE1 LMCD1 ID2 SKI RYBP KLF10 CTNNB1 ZEB1 NFIC)
p: 0	GO:0006685	3	sphingomyelin catabolic process	(SMPD4 SMPDL3B SMPD3)
p: 0	GO:0010553	12	negative regulation of specific transcription from RNA polymerase II promoter	(KAT5 PPARG HEY1 HDAC3 HES3 AES SKIL TCF7L2 HES1 SMARCA4 TBL1XR1 PPARA)
p: 0	GO:0016568	23	chromatin modification	(SMARCD2 MLL3 KAT5 CHD4 DOT1L EPC2 TLK2 TLK1 TSPYL2 TBL1XR1 PHF21A ATXN7L3 NSD1 WHSC1 SRCAP DAPK3 ARID1B HDAC3 SETD2 CHD5 BMI1 SETD1B MSL3)
p: 0	GO:0021915	5	neural tube development	(HES3 ZNF358 HES1 DVL1 TCF7L2)
p: 0	GO:0043353	3	enucleate erythrocyte differentiation	(ID2 SP3 CEBPG)

p: 0	GO:0045449	74	regulation of transcription	(CTBP1 ZNF451 MBD3L3 SETD1B NAT14 SETD2 ZNF512B CHD4 NCOA4 GMEB1 ARID1B CDK9 RYBP TBL1XR1 MED13L ZNF516 PHF21A ZFAT CHMP1A KLF3 ZNF644 ZNF358 SMARCA4 DPF1 ELL ZNF618 LPIN2 NSD1 BRMS1L MSL3 WHSC1 ZFX CTBP2 ZNF148 EIF2C2 MLLT10 BMI1 VEZF1 ZFPM1 ATXN7L3 CCNK TSPYL2 ZNF512 SRCAP CHD5 PHF19 IRF2BP2 KLF10 MAZ CRTC3 ZNF702P TWIST1 IKBKG ZBTB32 AES KAT5 TIGD2 SOX8 MED16 SMARCC2 HDAC3 CDYL TLE2 PHF10 SMARCD2 NEO1 ZNF319 HIVEP1 LCOR ZNF592 SALL3 EPC2 UBTF TCF25)
p: 1	GO:0007186	17	G-protein coupled receptor protein signaling pathway	(GPR160 LPAR1 GPR157 GPR173 ARHGEF11 GPR27 AVPR2 INSR GNAO1 GNB1L GPRC5B P2RY8 GCGR ADRA2C RGS19 OPRL1 GPR162)
p: 1	GO:0050896	4	response to stimulus	(PRCD LCN1 TULP1 PRPF8)