CODIGO DA PROVA: RP - 001/0001

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS
CONCURSO:

FOLHA DE RESPOSTA

Importante: O código da prova só será colocado na entrega da prova ao fiscal. As provas serão escaneadas e enviadas aos membros da banca avaliadora sem o nome do candidato.

Item A - Controle da expressão gênica com ênfase em métodos computacionais.

- O controle da expressão é uma das formas que buscamos para entender uma das perguntas mais fundamentais dentro da área de biologia celular e do desenvolvimento pois que é como o mesmo DNA de um organismo é capaz de dar origem a multiplos tipos celulares. isto é, é a partir do controle da expressão gênica que genes são ativados ou desativados ao longo da diferenciação celular e desenvolvimento.

O nosso entendimento deste processo está historicamente relacionado aos avanços em tecnologias para avaliar a propria expressão gênica e fatores fundamentais relacionados a como mensuramos o quanto, onde e como genes são expressos.

Inicialmente, métodos computacionais que nos ajudam a entender expressão gênica evoluem de acordo com as tecnologias que possuimos para extrair a informação biológica e quantifica-las. Assim desta forma, a bordagens iniciais usando métodos de quantificação da expressão genica com Southern blot e quantitative PCR, passaram a ser substituidos por métodos como microarray entre o final de 1990 e 2000 e então por métodos baseados em síntese com RNAseq, em especial bulk-RNAseq e a partir de meados 2010-2013 sendo empregado métodos de transcriptomica de células unicas. Portanto, houve a necessidade de se

Comissão Organizadora do Concurso
Gabinete da Direção

desenvolver métodos que atendessem esta demanda.
Para métodos de alto-rendimento onde é possivel interrogar a expressão gênica de forma genome-wide é necessário extrair a informação, normaliza-la e quantifica-la, buscando posteriormente métodos estatisticos para avalia-los.

Apesar de estar em desuso atualmente microarray deu o primeiro impulso com métodos computacionais voltados para expressão gênica, e um dos métodos desenvolvidos para analise de expressão diferencial foi o limma, baseado em regressão linear, que é suficiente para lidar com a variancia de dados derivado do sinal fluorescente derivado dos arrays. Ao passar para uso de bulk-RNA seq é possivel olhar de forma mais ampla para o conjunto de transcritos devido ao fato de não ser mais dependente de sondas com sequencias previamente conhecidas são usado métodos por sintese, especificamente sequencing-by-synthesis, com por exemplo illumina. Permitindo encontrar e quantificar não apenas regiões previamente anotadas mas também novas isoformas e regiões por exemplo não-codificante de proteina previamente desconhecidos.

Apartir do uso de microarray e bulk-RNA seq começamos a obter uma quantidade de dados massivo. E a partir desse momento foi possivel construir grandes redes gênicas regulatorias (GRN- gene regulatory network).

A construção dessas é baseado por exemplo em niveis de co-expressão, set-genes (grupos de genes de são conhecidos por atuarem em uma via em comum) e modelos de correlação e regressão. Combinado essas informações com informações prévias de anotação é possivel reconhecer genes regulados por exemplo pelo mesmo fator de transcrição.

Um grande desafio encontrado pela bioinformatica na analise de dados envolvidos na regulação gênica é a abundancia dos dados, e com isso quero dizer a quantidade

de dados que é necessário gerir e preparar. Dessa maneira mesmo se tratando de microarray e bulk-RNA seq é necessários organizar os dados de forma que possa extrair insights com relevância biológica e ao mesmo tempo ter ferramentas para verificar a qualidade e dispersão dos dados. Entre essas formas usando clusterização (clustering), isso é agrupar genes com perfil de expressão semelhantes. Para isso usando pacotes de bioinformática como clusterização hierárquica, sendo possível aplicar diferentes métodos, como inferência euclidiana ou então usando K-means. A partir disso é possível obter uma representação de genes com maior ou menor número de transcritos sendo up ou down regulados ao comparar diferentes grupos experimentais. Sendo essa representação ainda mais clara ao visualizarmos usando heatmaps, (gerados com pheatmaps package em R). Ao termos usado visão olística, em seguida por reduzir a dimensionalidade destes dados, uma vez que se trata de milhares de genes com alta variabilidade de abundância de transcritos, isso é feito ao usar Análise de Componentes Principais (PCA), normalizado para log2 (facilitando a visualização em escala) de tPm (transcritos por milhão). Ao usar essa abordagem é possível ver quais componente possuem maior variabilidade ou explicam a variabilidade e distância entre grupos, por exemplo ao comparar dois tipos celulares distintos, como neurônio motor e fibroblasto.

Para encontrar a gora genes diferencialmente expressos aplica-se métodos estatísticos com teste-t ao se comparar dois grupos experimentais ou ANOVA ao se comparar mais de dois grupos experimentais entre si.

Uma vez que temos uma visão geral do perfil da expressão gênica utilizando os métodos bioinformática de recurso para detectar e quantificar a expressão gênica, podemos buscar entender por exemplo como ou quais são os mecanismos que

levarão estes genes a serem diferencialmente expressos.

Assim, precisamos agora usar métodos em bioinformática que nos ajudem a identificar dentro dessa rede de genes GRN possuem assinaturas moleculares semelhantes ou distintas. É possível fazer isso olhando para além do gene body (exon-intron) do gene, isso é para regiões reguladoras, como TSS (transcription start site), promotor proximal, promotor distal, enhancers e silencer regions, além de sítios de ligação de fatores de transcrição.

Métodos de bioinformática são capazes de prever ~~identificar~~ motivos específicos baseado em dados experimentais previamente identificado ao buscar sequências consenso de fatores de transcrição, sequências conservadas baseadas em dados de alinhamentos multiplos como Clustalw e aplicando modelos como Hidden-Markov chain. Estas abordagens geram softwares como JASPAR, HOMER, MEME, TRANSFAC para a identificação de regiões de putativos binding-sites de fatores de transcrição e daí maquinário para início da transcrição. Um grande desafio neste campo é o grande número de falso-positivos ~~encontrados~~ encontrados ao usar estes métodos. Portanto é necessário combinar e alimentar métodos de machine learning supervisionado, Forest trees e neural network para reconhecerem padrões vindos de abordagens ~~que~~ que visam estabelecer por exemplo a acessibilidade da cromatina como ATAC-seq e métodos de busca identificar regiões de interação Proteina-RNA, como Chip-seq e cut-and-run.

~~Recalcitrância~~ É fundamental ressaltar que métodos de bioinformática e métodos experimentais devem ser usados em conjunto. Assim como um único método de bioinformática por exemplo predição de fatores de transcrição não é suficiente para explicar a uma ~~regulação~~ mudança de expressão por exemplo.

Dentre maneira uma abordagem de bioinformática desenvolvida para integrar as diferentes camadas de mecanismos de regulação gênica é a multi-omica, onde é possivel integrar dados de expressão gênica derivada de RNA-seq, junto com dados de acessibidade da cromatina disponiveis para transposase (ATAC-seq), mais dados de chip-seq para analisar a ligação de um fator de transcrição ou modificação especifica de histona ex H3K4me3. Ao combinarmos estes dados temos a possibilidade de encontrar fatores em comum correlacionados entre si. isso é possivel ao usar modelos de variaveis-latentes como uma forma de redução de dimensionalidade para em seguida combinar estes dados em matriz de fatorização. seguindo o workflow para clusterização e identificação de fatores que correlacionam e convergem em mostrar que mecanismos em comum atuam para fazer com que um gene ou um grupo de genes sejam ativamente expressos.

Até o momento olhamos e buscamos entender redes regulatorias com métodos que fornece geral fazem uma média da expressão gênica derivada de um grupo de células, usando por exemplo - bulK-RNAseq. No entanto, como mencionado anteriormente entre meados de 2010-2013 tentativas de se obter dados de expressão gênica derivadas de células únicas cresceram e tiveram um boom a partir de meados de 2017, com abordagens com drop-seq exemplo 10x chromium e Smart-seq. A bioinformatica aplicada a redes de regulação usando dados derivados de transcriptoma e ATAC-seq de células únicas nos levam a ter maior resolução, a nivel de quais redes regulatórias estão sendo utilizadas por exemplo durante a diferenciação de uma célula tronco germinativa humana até estágios de gastulação, sendo possivel traçar a mudança das

redes regulatórias pareado em mudança de expressão gênica e mudança de isoformas ou de transcritos derivados de splicing diferencial, usando ferramentas como Monocle e RNA velocity.

É claro que a combinação de dados multi-omicos deve ser integrada com análises como identificar os tipos celulares identificados assim como estados celulares virados de scRNAseq usando por exemplo Cluster profiler em R.

E a partir disso combinar com informações de gene ontology, usando DAVID ou exemplo ou IPA (Quiagen). A análise de gene ontology se faz crucial ao usar tanto bulk RNA seq quanto SC-RNAseq. Ou mesmo a combinação paralela de SC-RNAseq mais SC-ATAC-seq.

Para concluir se faz necessário pontuar que a integração de multiplas omicas e a predição e alimentação de modelos e ferramentas de bioinformática por métodos experimentais é a chave para entendermos redes de regulação gênica, assim como a integração de dados de interação proteina-proteina derivados de ~~yeast hybrid assay~~ ensaio duplo hibrido de leveduras e predições derivadas deste dados, assim como de imuno precipitação seguida de mass espec. (espectrometria de massas). Uma ver combinada a informação vinda destes métodos softwares de produção e visualização de redes como cytoscape podem ser usados para visualizar redes de regulatórias envolvendo multiplas camadas de regulação molecular revelados ao utilizar abordagens de bioinformática.

- tópico 7 - Bioinformática aplicada a análise epigenômica.

- A pergunta de como a informação contida em uma célula leva a multitude de tipos celulares é chave para guiar o campo da epigenômica. Porque a epigenética por si é o conjunto de modificações no DNA possuindo informação biológica sem alterar a sequencia de nucleotídeos que compõe o DNA.
Desta forma ao analisar o conjunto de modificações no DNA desde metilação em citosina até mesmo pós-modificação em caudas de histonas como acetilação metilação ubiquitinação e fosforilação traz informação da acessibilidade da informação contida no DNA e portanto fundamental para que circuitos genéticos e programas celulares sejam utilizados. Para por exemplo a diferenciação ou reprogramação (ipscs) de tipos celulares seja obtida.
Métodos de bioinformática para acessar por exemplo citosina metiladas, em especial dinucleotídeos CpG são aplicados ao utilizar métodos de sequenciamento de alto rendimento por exemplo Bissulfite sequencing usando illumina paired-end após tratar o DNA com bissulfite, levando a ter no final do processo a conversão citosina para timina. E a partir disso, é possivel identificar citosinas metiladas, uma vez que citosinas metiladas não são convertidas para timina. Ao produzir bibliotecas de BS-seq ou RR-BSseq (reduced representation bissulfite sequencing é possivel ter uma visão geral do metiloma. Isso gerou ao uso de softwares como para analise e qualidade e limpeza dos reads como FASTQC, seguida de normalização, analise de componentes principais e análises especificas para metilação

como methkit e Bismarck para analise de reads metilados e não metilados. A partir disso, busca-se mapear onde estão os dinucleotides CpG ao longo do genoma, assim como cpG island contendo de 200 - 3000 repetições.

Uma vez que se identifica sitios metilados e não metilados o mapeamento usando por exemplo com UCSC genome Browser ou IGV reads metilados ao redor de por exemplo gene body (exon-intron). Uma vez que a metilação de citosina está relacionada com regiões não acessiveis a maquinaria que controla a transcrição, esta abordagem de bioinformática pode ser associada a inativação da expressão gênica ou na ausencia de metilação de citosina com região aptas a serem transcricionalmente ativa.

É fundamental salientar que um único metodo de bioinformática sozinho não leva a uma compreensão abrangente de epigenômica. Por isso é necessário combinar metodos como ATAC-seq (cromatina acessivel ao ataque de transposase TN5 - seguida de sequenciamento). Com esta abordagem é possivel detectar regiões de cromatina em estado de eucromatina (aberta) e disponivel a transposição por TN5, e apartir disso com a inclusão de adaptares illumina, por exemplo, é possivel amplificar essa fragmento-to pelo PCR em seguida sequenciados. Ao mapear os reads derivados da analise seguida de analise de qualidade e remoção de adaptadores e normalização, é possivel identificar regiões de cromatina acessivel. E portanto, a partir desse momento identificar passando pelas anotações e informações previas se esta região é uma região regulatória (promotor/enhancer) que pode conter sitios putativos para ligação de fatores de transcrição. É possivel fazer uma

busca por este sítios utilizando preditores de fatores de transcrição e elementos regulatórios como MEME, TRANSFAC e JASPAR por exemplo.

No entanto, graças a outras abordagens de bioinformática podemos buscar validar experimentalmente estes sítios putativos. A abordagem funciona da seguinte forma, ao usarmos preditores podemos encontrar enriquecimento de putativos binding sites na região upstream do gene body, dentro de uma janela genômica de 3KB (3 mil pares de bases), um exemplo é Kappa B sites, desta forma o pesquisador pode direcionar o experimento de Chip-seq para usar o anticorpo anti-NFKB (p65). Ao usar Chip-seq como sugerido pelo nome imunoprecipitação da cromatina seguida de sequenciamento. Desta maneira no imunoprecipitar o DNA genômico com anti-NFKB, o pesquisador deve fragmentar o DNA, por exemplo por sonicação e pegar fragmentos que são enriquecidos na proteína de interesse. A partir desse momento, se amplica por PCR os fragmentos que foram enriquecidos na imunoprecipitação, prepara os fragmentos para biblioteca illumina e analisa as qualidades das reads com Fastqc e em seguida normaliza pelo background do amostra controle sem imunoprecipitação. Ao fazer isso, agora é possível validar. se os putativos sítios de KB sites contem a interação DNA-Proteína predita.

Ao combinar metiloma, ATAC-seq e Chipseq é possível ter uma visão com preensiva de modificação, epigenomicas controlada por exemplo a resposta a inflamação ou processos de diferenciação celular ou mesmo fatores chave que são suficientes

UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS BIOMÉDICAS

para por exemplo reprogramar uma célula diferenciada e se tornar pluripotente novamente como ocorre com células tronco de pluripotência induzida

tópico 9 - Aplicação da bioinformática no estudo das patologias humanas.

A bioinformática é tem um papel fundamental no entendimento e busca de terapias para patologias humanas.

Com o avanço e barateamento de sequenciamento ~~do avanço e~~ de genomas inteiros ~~sendo~~ (WGS) assim como de WES, hoje é possível sequenciar o genoma ou exoma de milhares de doadores saudáveis ou com patologias, sejam elas familiares ou esporádicas. A partir do sequenciamento desses genoma e exomas inteiros é possível montalos usando o método de re-sequencing usando do short reads Illumina ou long-reads (PACBIO HiFi, Oxford Nanopore) para encontrar variantes de nucleotideos únicos, indels ou variantes estruturais.

Com estes dados é possível usar PLINK para conduzir estudos de GWAS (genome-wide association studies). Desta forma encontrar o odds ratio de uma determinada variante genica ser benéfica ou patogenica dentro de uma população ou coorte coorte.

A partir disso é possível testar ~~variante~~ variante associadas desde doenças ~~de~~ genicas de frequencia comum até doenças raras. Portanto ferramentas como GATK e SAMtools são essenciais para encontrar variantes no processo de variant calling.

Uma vez que variantes foram associadas com uma determinada patologia. É possível estratificar estes pacientes, o que vem se tornando fundamental para a utilização de novas terapias que são desenvolvidas de forma personalizada. Um exemplo é a estratificação de pacientes

com a doença neurodegenerativa rara esclerose lateral amiotrófica. Devido a natureza de ter multiplos genes e variantes relacionadas, descobrir usando métodos de bioinformática que terapias com drogas específicas são efetivas para genes e variantes específicos.

É importante ressaltar que a bioinformática pode revelar marcadores clínicos, muito antes dos pacientes de fato manifestarem seus sintomas/fenótipos. Um exemplo é huntington, onde é possivel detectar a expansão somática de poly Q, sendo alimentado antes do pacientes manifestarem sintomas ao usar amplicon seq seguido de mapeamento devendo detectando a expansão.

Para o cancer, hoje se aplica exome-seq e WGS para detectar variantes potencialmente causadoras como driver mutações, mas vai além ao empregar RNA scRNA-seq combinado com a transcriptoma espacial.

Ao usar abordagens com Seurat(R) é possivel revelar a heterogeneidade celular desde tipos e estados celulares, e para comparalos usando MAST(R), revelando populações complexas sendo visualizadas com redução de dimensionalidade com UMAP ou t-SNE.

Porém, ao introduzirmos scRNA-seq perdemos uma informação fundamental que é a espacial onde estas células estão, e por exemplo se tratando de cancer onde está o infiltrado de celulas imunes. Ao usar Visium 10x para gene wide transcriptome temos a resolução de spots ou ao usar MERFISH com alvos direcionados baseado em marcadores únicos derivados de scRNA-seq é possivel do pacote SPAPROS identificar

transcritos únicos para MER fish. MER Fish, ao contrário de Visum resolve transcriptoma espacial com resolução sub-celular.
Ao encontrar cluster únicos, resolvidos espacialmente e com essa resolução celular única, é possível selecionar marcadores únicos para drug discovery/ baseado em estrutura.