Análise de Expressão Gênica Diferencial no Embranquecimento do Cabelo

Resumo

O objetivo foi identificar os genes cuja atividade (expressão) muda significativamente nos folículos capilares quando eles deixam de produzir pigmento (cabelos brancos). A análise busca entender a biologia molecular por trás do embranquecimento do cabelo. Foi usado Jupyter notebook para a análise.

Foi realizada uma análise de expressão diferencial utilizando um conjunto de dados de microarray disponível publicamente (GSE24009). Os dados foram pré-processados e analisados com Python e bibliotecas como pandas e scipy.

Este projeto está dividido em: *Resumo, Fonte de Dados e Pré-processamento, Análise Estatística e Interpretação dos Resultados, Modelo Exploratório, Conclusão e Limitações do Projeto.*

Fonte dos Dados e Pré-processamento

Os dados foram obtidos do **Gene Expression Omnibus (GEO)** (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24009) um repositório público do NCBI para dados de expressão gênica. O conjunto de dados utilizado foi o **GSE24009**, que contém dados de expressão de folículos capilares humanos.

Dados para instalação (https://ftp.ncbi.nlm.nih.gov/geo/series/GSE24nnn/GSE24009/matrix/).

Limpeza e Filtragem: O arquivo de dados brutos foi pré-processado com um script Python (data_loader) para:

- Remover metadados e linhas irrelevantes.
- Selecionar apenas as 12 amostras de folículos capilares (8 pigmentados e 4 não-pigmentados).
- Normalizar os dados com uma transformação logarítmica (log2) para estabilizar a variância.

Análise Estatística e Interpretação dos Resultados

O objetivo foi comparar os níveis de expressão de cada gene entre os dois grupos de folículos (pigmentados vs. não-pigmentados) para encontrar diferenças estatisticamente significativas.

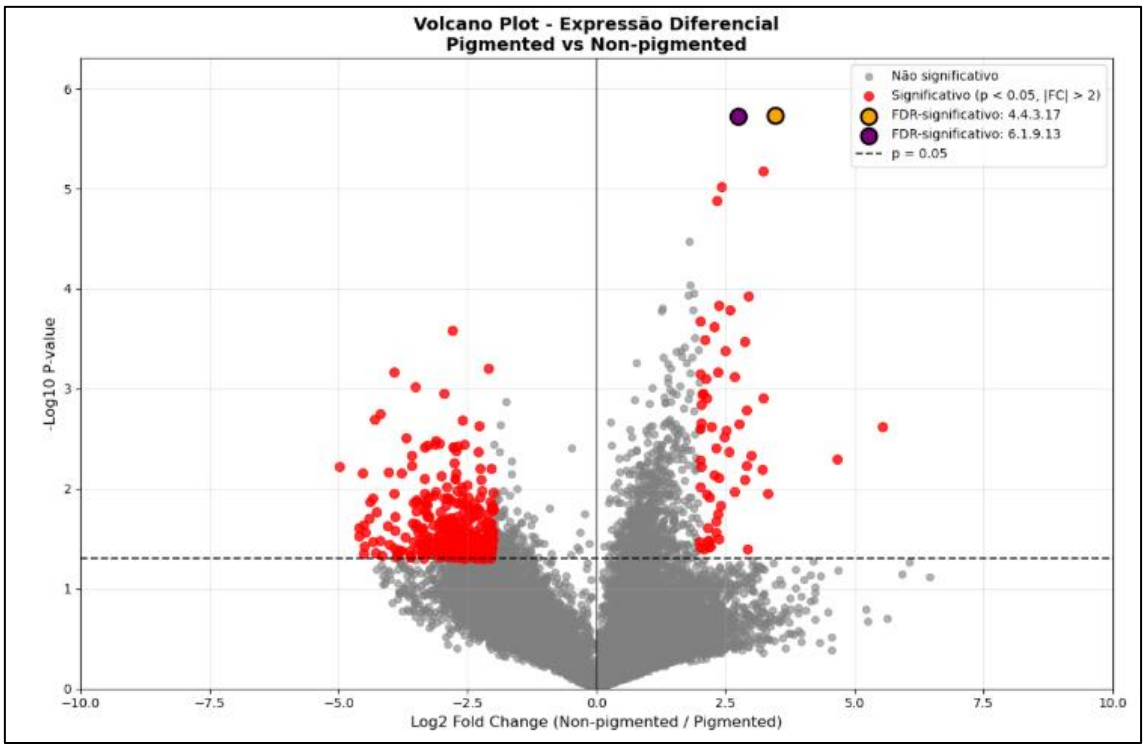
Foi utilizado o *Teste t de Student* para cada um dos 23.232 genes, com o intuito de verificar se a diferença de expressão entre as médias dos grupos era real ou aleatória.

Métricas Chave:

- P-value**:** Mede a significância estatística. Um valor abaixo de 0.05 foi utilizado como critério para identificar genes relevantes.
- Log2(Fold Change): Mede a magnitude da diferença de expressão. Um valor de 1.0 significa que a expressão do gene dobrou; -1.0 significa que foi reduzida pela metade.

Resultados Antes da Correção (P-values originais):

Para visualizar os resultados, foi gerado um Volcano Plot. Este gráfico é ideal para identificar os genes que são, ao mesmo tempo, estatisticamente significativos (alto no eixo Y) e biologicamente relevantes (longe do centro no eixo X). Os pontos vermelhos no gráfico representam os genes que atendem a esses critérios. Genes significativos p< 0.05, |FC| > 2, estão em vermelho abaixo.



Resultados Após a Correção de Testes Múltiplos:

Foi aplicada a correção FDR aos p-values originais de todos os genes. Alpha de do FDR foi 0.05. Estes genes estão indicados no gráfico Volcano Plot acima. Abaixo estão os genes que passaram pela correção.

Genes significativos após FDR: 2

	gene	p_value	t_statistic	log2_fold_change	mean_pigmented	mean_non_pigmented	n_pigmented	n_non_pigmented	p_adj	significant_fdr
13406	4.4.3.17	0.000002	12.746728	3.464585	0.172857	1.908232	8	4	NaN	True
15954	6.1.9.13	0.000002	9.880745	2.745708	0.235194	1.577492	8	4	NaN	True

Interpretação dos Resultados com a correção:

O ID 4.4.3.17 é RNA não codificante (lncRNA/host gene) que hospeda o microRNA MIR646HG é um regulador pós-transcricional importante, ligando-se a sinais de estresse oxidativo e proliferação celular; pode regular negativamente genes críticos para a sobrevivência dos melanócitos durante estresse oxidativo ou envelhecimento.

O ID 6.1.9.13 é do gene MS4A6E e ele é da família MS4A associada a imunidade/micróglia; pode refletir sinalização imune local no folículo.

Interpretação dos Resultados ordenados pelo p_value:

Selecionei os 20 genes mais significativos antes da correção para uma análise mais detalhada. Os valores com o log2_fold_change negativos se relaciona com o grupo pigmentado e o positivo ao não pigmentado. Segue os IDs desses genes ordenados pelo p_value.

Genes significativos: 426
O log2_fold_change quando é negativo ele é sobre o grupo pigmentado

	gene	p_value	log2_fold_change	significant	significant_fdr
13406	4.4.3.17	0.000002	3.464585	True	True
15954	6.1.9.13	0.000002	2.745708	True	True
8259	2.2.10.18	0.000007	3.226997	True	False
12826	4.3.2.1	0.000010	2.430278	True	False
10516	3.2.4.1	0.000013	2.333746	True	False
16497	6.3.10.7	0.000119	2.945242	True	False
9794	3.1.14.13	0.000148	2.368070	True	False
3223	10.3.22.2	0.000165	2.578150	True	False
6349	12.2.11.21	0.000212	2.015599	True	False
11012	3.3.4.20	0.000238	2.276702	True	False
1589	1.4.15.14	0.000262	-2.800970	True	False
5851	12.1.10.9	0.000321	2.100227	True	False
11002	3.3.4.11	0.000339	2.874257	True	False
16642	6.3.17.19	0.000418	2.496112	True	False
9117	2.3.6.18	0.000626	-2.095474	True	False
66	1.1.12.1	0.000687	2.351377	True	False
7293	12.4.10.2	0.000687	-3.929574	True	False
23104	9.4.4.13	0.000709	2.010254	True	False
10960	3.3.22.13	0.000766	2.682945	True	False
14523	5.3.1.12	0.000789	2.115064	True	False

A partir da tabela de resultados, foi possível identificar os genes mais impactantes relacionados ao estudo:

1. ID 4.4.3.17, BC016338 (Gene correlacionado MIR646HG)

- RNA não codificante, hospedeiro do microRNA miR-646, que regula genes ligados ao estresse oxidativo e apoptose.
- Possível papel no embranquecimento: modula a sobrevivência dos melanócitos, principais células pigmentares do folículo. Alterações podem aumentar a suscetibilidade ao dano oxidativo.

2. ID 6.1.9.13, MS4A6E (Gene)

- Membro da família MS4A, associada à sinalização imunológica e envelhecimento celular (já descrito em estudos de Alzheimer).
- Possível papel: pode refletir um estado de inflamação crônica e envelhecimento do folículo, favorecendo perda de melanócitos.

3. ID 2.2.10.18, PAX4

- Fator de transcrição essencial na diferenciação celular, principalmente em pâncreas endócrino.
- Possível papel: pode influenciar diferenciação de células progenitoras no folículo, inclusive melanócitos. Alterações poderiam reduzir a capacidade regenerativa pigmentária.

4. ID 4.3.2.1, AK024514(SUZ12)

- Subunidade central do complexo PRC2, responsável por metilação de histonas e silenciamento gênico.
- Possível papel: alterações epigenéticas podem desregular genes ligados à pigmentação e manutenção de melanócitos.

5. ID 3.2.4.1, TRPV5

- Canal de cálcio altamente seletivo, regulado pela vitamina D.
- Possível papel: o cálcio é crítico na sinalização de melanócitos. Disfunções em TRPV5 podem prejudicar a resposta celular ao estresse e comprometer a produção de melanina.

6. ID 6.3.10.7, AL049325(KRIT1)

- Relacionado à estabilidade vascular e resposta ao estresse oxidativo.
- Possível papel: manutenção inadequada do microambiente vascular do folículo pode comprometer nutrição e oxigenação dos melanócitos.

7. ID 3.1.14.13, BC015133(SOX6)

- Fator de transcrição envolvido em diferenciação neuronal, muscular e possivelmente melanocítica.
- Possível papel: alterações podem prejudicar a renovação de células pigmentares, acelerando o embranquecimento.

8. ID 10.3.22.2 H200000013

- Não encontrado gene relacionado.

9. ID 12.2.11.21, SARA1

- GTPase envolvida em tráfego intracelular e transporte de vesículas.
- Possível papel: pode afetar o transporte de proteínas essenciais para melanogênese (ex.: tirosinase), prejudicando a pigmentação.

10. ID 3.3.4.20, PLEKHB2

- Relacionado à organização da membrana plasmática e sinalização celular.
- Possível papel: pode influenciar adesão e comunicação entre células do folículo, impactando sobrevivência dos melanócitos.

11. ID 1.4.15.14, LY6H

- Proteína de superfície associada a sinalização neuronal e imunomodulação.
- Possível papel: pode refletir ligação entre sistema nervoso e pigmentação capilar, já que o estresse neural influencia fortemente o embranquecimento.

12. ID 12.1.10.9, GPRASP1

- Regulador do tráfego de receptores acoplados à proteína G (GPCRs).
- Possível papel: muitos receptores envolvidos em resposta a hormônios e estresse são GPCRs; sua desregulação pode afetar vias que controlam melanogênese.

13. ID 3.3.4.11, AL137552(GDPD5)

- Participa no metabolismo lipídico e pode induzir apoptose.
- Possível papel: pode aumentar a morte dos melanócitos sob condições de estresse, um dos mecanismos centrais do embranquecimento.

14. ID 6.3.17.19, NM_018194(HHAT)

- Enzima essencial da via Hedgehog, fundamental para crescimento e regeneração do folículo piloso.
- Possível papel: alterações nesta via podem prejudicar o nicho de células-tronco, reduzindo repigmentação após ciclos do cabelo.

15. ID 2.3.6.18, NM_207006(FAM83A)

- Regula proliferação celular via EGFR/MAPK.
- Possível papel: pode influenciar a proliferação de melanócitos e células-tronco do folículo; disfunções aceleram a perda da reserva pigmentária.

16. ID 1.1.12.1, ELL3

- Fator de elongação da RNA polimerase II, envolvido em transcrição gênica.
- Possível papel: pode modular a expressão de genes-chave da melanogênese, afetando a produção de melanina.

17. ID 12.4.10.2, SIGLEC6

- Receptor de lectina ligado a processos de imunidade inata.
- Possível papel: pode favorecer respostas inflamatórias locais no folículo, prejudicando melanócitos.

18. ID 9.4.4.13, AK024599(RABGEF1)

- Regula endocitose e resposta imune.
- Possível papel: pode afetar a degradação/reciclagem de receptores ligados à pigmentação ou desencadear inflamação crônica.

19. ID 3.3.22.13, CSF3

- Também chamado de G-CSF, estimula proliferação e sobrevivência de células da medula óssea.
- Possível papel: pode estar ligado a inflamação sistêmica ou local no folículo. Inflamação crônica é um fator chave na morte de melanócitos.

20. ID 5.3.1.12, GSG1

- Associado a diferenciação celular em tecidos reprodutivos, mas pouco caracterizado.
- Possível papel: possivelmente um gene acessório no contexto capilar, mas pode refletir vias gerais de proliferação/diferenciação.

Agrupando as funções:

Estresse Oxidativo & Apoptose

- MIR646HG, KRIT1, GDPD5 → modulam sobrevivência dos melanócitos.

Epigenética & Diferenciação

- SUZ12, SOX6, PAX4, GSG1 → regulam identidade e regeneração celular.

Sinalização & Crescimento Folicular

- HHAT, FAM83A, ELL3, GPRASP1, TRPV5 → controlam comunicação e renovação do folículo.

Inflamação & Imunidade

- MS4A6E, SIGLEC6, CSF3, RABGEF1, LY6H → refletem inflamação crônica e envelhecimento celular.

Tráfego & Homeostase Celular

- SARA1, PLEKHB2 → mantêm equilíbrio intracelular e membranas.

Site usado para checagem de estudo e características do gene, https://biit.cs.ut.ee/gprofiler/convert e https://www.ensembl.org/Homo_sapiens/Tools/Blast?db=core;expand_form=true;tl=bWAj2Es0id5INNW1-11262466

Site da tabela dos nomes dos genes relacionados com os IDs, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?view=data&acc=GPL3877&id=11482&db=GeoDb_blob92 .

Para mais detalhes dos genes e possíveis correlações com o tema usei uma LLM (ChatGPT, Gemini e DeepSeek), para aprofundar e ter mais acurácia na análise.

Modelo Exploratório

O modelo é exploratório, não preditivo robusto. Importante deixar claro que a ideia é apenas demonstrar que existe separação possível com poucos genes, não criar um biomarcador pronto. Para avaliar se os genes identificados poderiam discriminar entre cabelos pigmentados e não pigmentados, foi construído um modelo de regressão logística. O modelo utilizou quatro genes como variáveis de entrada: os dois genes que permaneceram significativos após correção por FDR (MIR646HG e MS4A6E), acrescidos de dois genes com maiores valores absolutos de log2 fold change (IDs 12.2.20.2 (3xSSC) e 5.3.15.15 (H200007402)) não apresentaram anotação clara em bases de dados como Ensembl. Embora estatisticamente relevantes, limitam a exploração desses candidatos no contexto do embranquecimento capilar.

O pipeline incluiu normalização por StandardScaler e regularização L2 (C=0.1), com o objetivo de reduzir o risco de sobreajuste dado o número reduzido de amostras. A avaliação do desempenho foi realizada por validação cruzada leave-one-out (LOOCV), apropriada para conjuntos de dados pequenos (n=12). Além disso, foi aplicada Repeated Stratified K-Fold (3 divisões repetidas 20 vezes) como verificação adicional de robustez.

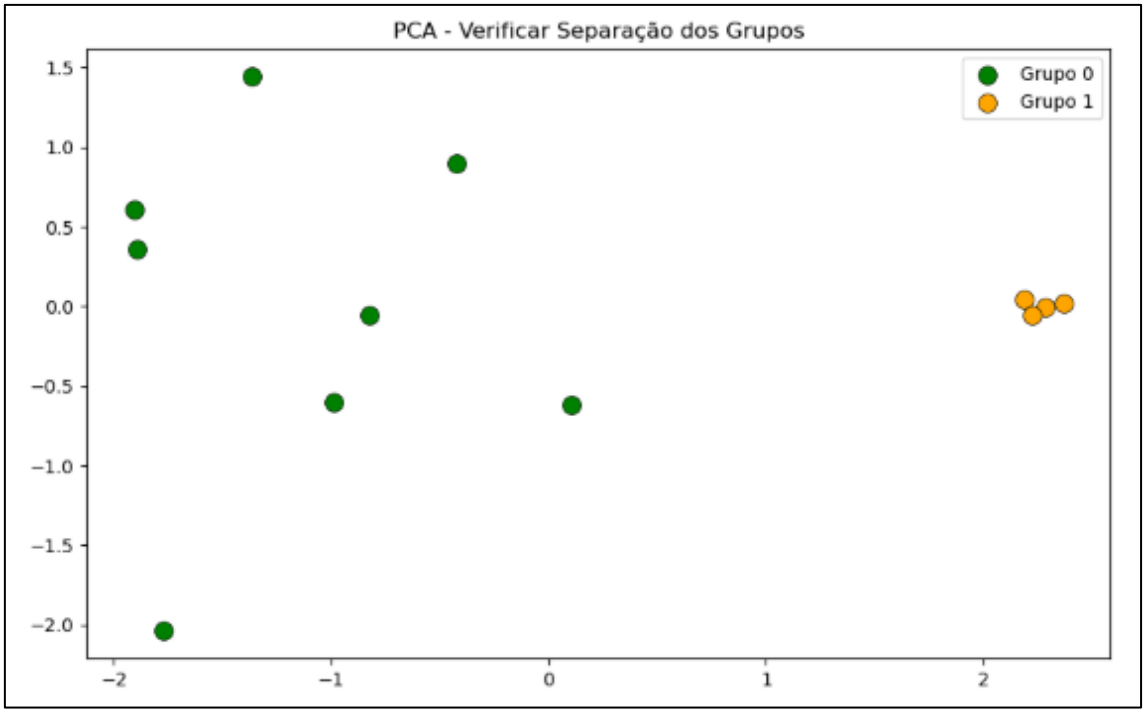
Os resultados mostraram que o modelo foi capaz de separar parcialmente os grupos, com acurácia média acima do esperado para classificações aleatórias. Isso sugere que os genes analisados possuem poder discriminativo preliminar, embora limitado pelo baixo tamanho amostral e pela ausência de validação externa.

```
Acurácia média: 1.000
Acurácia por amostra: [1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1.]
```

```
LOOCV: 1.000
Repeated K-Fold: 0.833 (±0.236)
```

Gráfico PCA (Análise de Componentes Principais):

A análise de componentes principais (PCA) foi realizada utilizando os genes incluídos no modelo logístico. Este resultado reforça a ideia de que assinaturas multigênicas, mesmo em pequena escala, podem fornecer discriminação útil entre os grupos. Grupo 0 é o pigmentado e o 1 o não pigmentado.

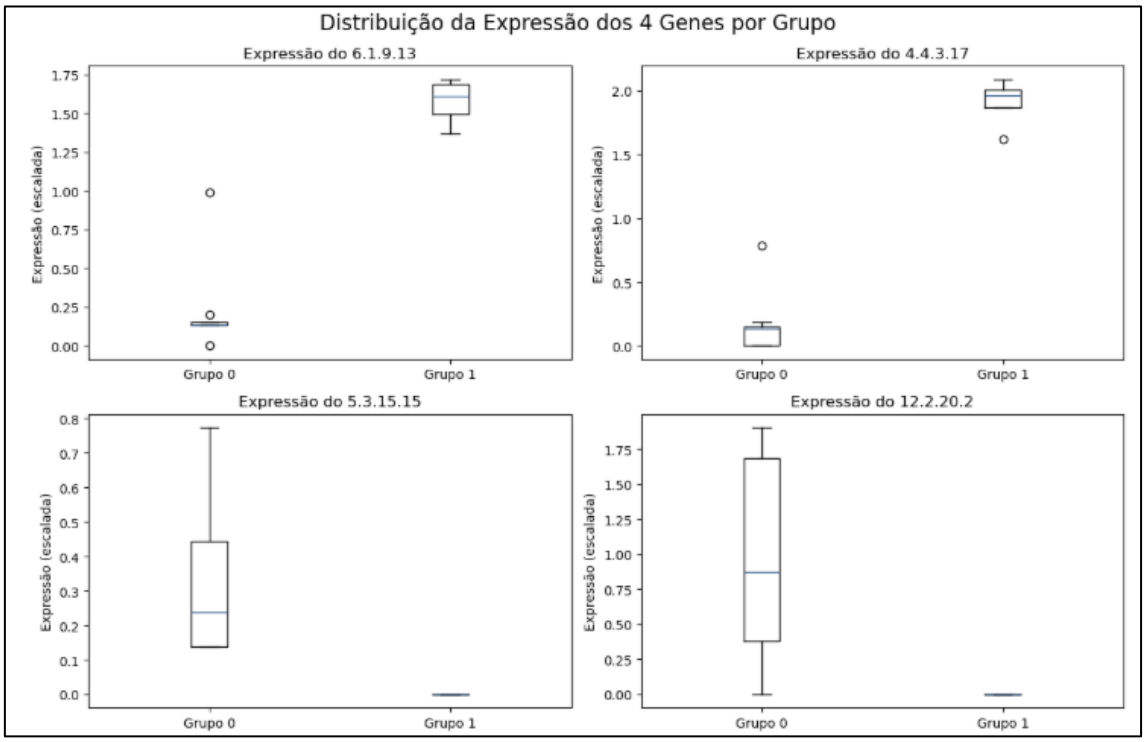


Boxplot de cada gene do modelo:

Foram construídos boxplots individuais para os quatro genes incluídos no modelo (os dois genes significativos após FDR e os dois com maiores valores absolutos de log2 fold change). Os gráficos mostram a distribuição de expressão entre os grupos pigmentado e não pigmentado, permitindo uma visualização direta das diferenças.

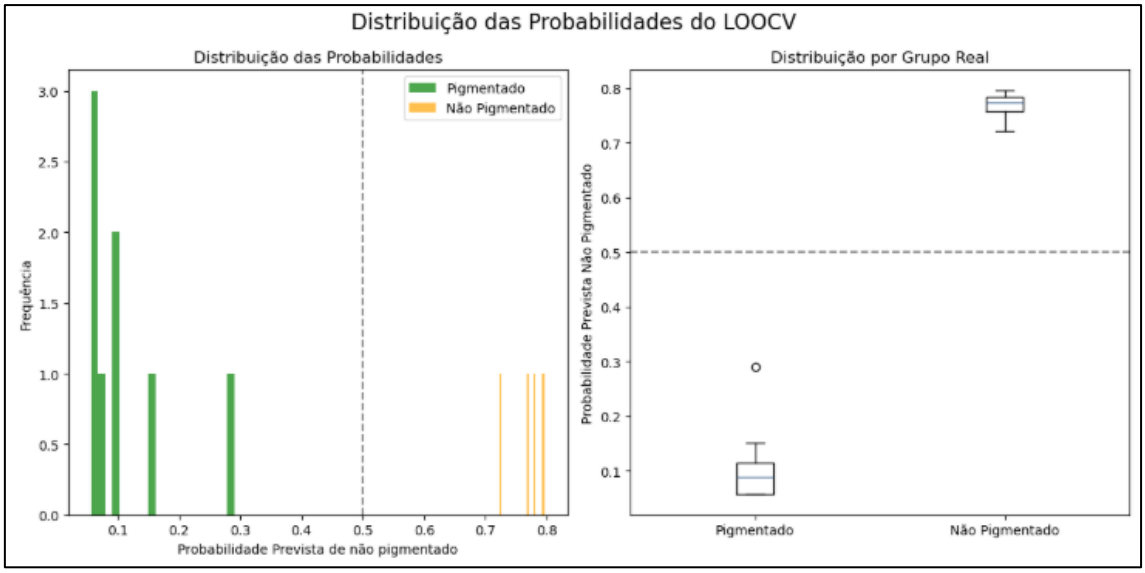
ID 6.1.9.13 MS4A6E, ID 4.4.3.17 MIR646HG

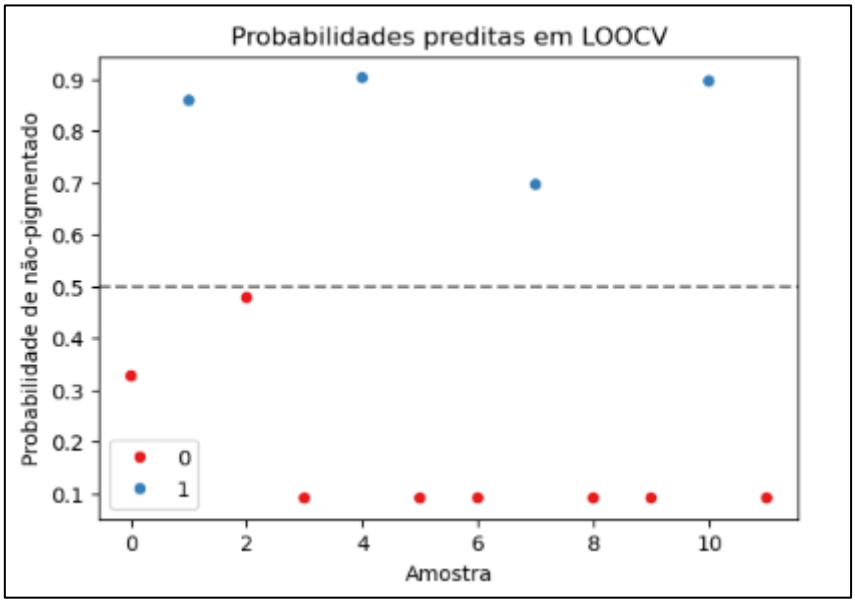
ID 5.3.15.15 (H200007402), ID 12.2.20.2 (3xSSC)



Gráficos de distribuição de probabilidades do LOOCV:

A validação cruzada leave-one-out (LOOCV) foi utilizada para estimar a probabilidade de cada amostra pertencer ao grupo pigmentado ou não pigmentado. O gráfico de distribuição de probabilidades mostrou que, em geral, o modelo atribuiu valores de probabilidade mais altos ao grupo correto, ainda que com variações entre amostras individuais. Essa tendência indica que o modelo é capaz de aprender padrões discriminativos, mesmo com número reduzido de genes e amostras. Contudo, a sobreposição em alguns casos também evidencia a limitação do modelo em termos de poder preditivo, devendo ser interpretado como uma análise exploratória.





Conclusão

Este estudo investigou a expressão diferencial de genes em amostras de cabelos pigmentados e não pigmentados, buscando compreender os possíveis mecanismos moleculares associados ao embranquecimento capilar.

A análise estatística envolveu teste t, cálculo de p-value, t-statistic e log2 fold change, além de correção por múltiplos testes via FDR. Apenas dois genes (MIR646HG e MS4A6E) mantiveram-se significativos após a correção, indicando alta robustez estatística, ainda que não estejam diretamente ligados à melanogênese clássica.

De modo complementar, foi realizada uma análise dos 20 genes com menor p-value (sem correção FDR), o que permitiu observar potenciais candidatos adicionais, relacionados a processos como regulação epigenética (SUZ12, SOX6), inflamação e resposta imune (CSF3, SIGLEC6) e sinalização celular (HHAT, FAM83A, TRPV5).

Por fim, um modelo logístico utilizando quatro genes (os dois genes confirmados pelo FDR e dois genes com maiores valores absolutos de log2 fold change) foi testado, demonstrando que a combinação de marcadores pode melhorar a separação entre os grupos pigmentado e não pigmentado. Isso sugere que o embranquecimento capilar pode ser mais bem explicado por assinaturas multigênicas, em vez de alterações isoladas.

Em conjunto, os resultados apontam que o processo de perda de pigmentação capilar pode estar ligado não apenas a genes diretamente da melanogênese, mas também a mecanismos de envelhecimento celular, regulação epigenética, estresse oxidativo e inflamação local.

Limitações do Projeto

- Tamanho amostral reduzido: A análise foi realizada com um número limitado de amostras (8 pigmentadas e 4 não-pigmentadas), o que impactou o poder estatístico.
- Correção múltipla conservadora: Após ajuste por FDR, apenas dois genes permaneceram significativos, o que garante rigor estatístico, mas reduz a quantidade de candidatos exploráveis.
- Análise funcional aprofundada: Embora a análise tenha identificado os principais genes, seria necessário expandir a análise para todas as vias biológicas e genes, usando ferramentas de enriquecimento funcional automatizadas.
- Modelo logístico restrito: o modelo foi exploratório e treinado em número pequeno de genes, sem validação cruzada extensiva ou replicação em outros conjuntos de dados.

- Ausência de variáveis clínicas: fatores como idade, estresse, doenças metabólicas (ex.: diabetes) e hábitos de vida não foram incluídos, mas podem impactar fortemente a expressão gênica e o embranquecimento capilar.
- Anotação incompleta de sondas: Parte dos IDs analisados (como 3xSSC e H200007402) não possuem correspondência clara em genes humanos conhecidos, o que limita a interpretação funcional dos resultados.

.