UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades de Pós-graduação em Bioinformática

THAIS SILVA TAVARES

CENÁRIO DE COOCORRÊNCIA DE VARIANTES DE NUCLEOTÍDEO ÚNICO (SNVS) DE DIFERENTES ORIGENS EVOLUTIVAS E CONSEQUÊNCIAS FUNCIONAIS NO GENOMA HUMANO.

Belo Horizonte–MG
2025

THAIS SILVA TAVARES

CENÁRIO DE COOCORRÊNCIA DE VARIANTES DE NUCLEOTÍDEO ÚNICO (SNVS) DE DIFERENTES ORIGENS EVOLUTIVAS E CONSEQUÊNCIAS FUNCIONAIS NO GENOMA HUMANO.

Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade Federal de Minas Gerais, como requisito parcial para a obtenção do título de Doutora em Bioinformática.

Orientador: Prof. Dr. Francisco Pereira Lobo
Coorientador: Prof. Dr. Renan Pedra de Souza

Belo Horizonte–MG
2025

043 Tavares, Thais Silva.
Cenário de coocorrência de Variantes de Nucleotídeo Único (SNVs) de diferentes origens evolutivas e consequências funcionais no genoma humano [manuscrito] / Thais Silva Tavares. – 2025.
181 f. : il. ; 29,5 cm.

Orientador: Prof. Dr. Francisco Pereira Lobo. Coorientador: Prof. Dr. Renan Pedra de Souza.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Bioinformática. 2. Mutação. 3. Polimorfismo de Nucleotídeo Único. 4. Hotspot de Doença. I. Lobo, Francisco Pereira. II. Souza, Renan Pedra de. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS
FOLHA DE APROVAÇÃO

Thais Silva Tavares

"Cenário de coocorrência de Variantes de Nucleotídeo Único (SNVs) de diferentes origens evolutivas e consequências funcionais no genoma humano"

Tese aprovada, em 22 de maio de 2025, pela banca examinadora constituída pelos Professores: Prof. Francisco Pereira Lobo - Orientador (Universidade Federal de Minas Gerais), Prof. Renan Pedra de Souza - Coorientador (Universidade Federal de Minas Gerais), Profa. Glória Regina Franco (Universidade Federal de Minas Gerais), Prof. Luiz Eduardo Vieira Del Bem (Universidade de São Paulo), Prof. Mateus Henrique Gouveia (National Human Genome Research Institute, USA) e Prof. Tetsu Sakamoto (Universidade Federal do Rio Grande do Norte).

Belo Horizonte, 22 de maio de 2025.

RAQUEL CARDOSO DE MELO MINARDI

Coordenadora do Programa Interunidades de Pós-Graduação em Bioinformática

Documento assinado eletronicamente por **Raquel Cardoso de Melo Minardi**, **Coordenador(a) de curso de pós-graduação**, em 30/05/2025, às 09:32, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4242023** e o código CRC **58D1851E**.

Referência: Processo nº 23072.230686/2025-80 SEI nº 4242023

AGRADECIMENTOS

Chegar até aqui foi uma jornada desafiadora e enriquecedora, e nada disso teria sido possível sem o apoio e a contribuição de muitas pessoas ao longo do caminho.

À minha mãe, minha maior inspiração. Agradeço não apenas por incentivar meu crescimento intelectual e profissional, mas também por ser um exemplo de mulher batalhadora, que sempre lutou pelos seus sonhos e nos proporcionou condições de estudo e uma base familiar sólida. Sua força e determinação me ensinaram a persistir diante das dificuldades e a nunca deixar de acreditar no meu potencial.

A mim mesma, por ter tido a coragem de me desafiar constantemente e buscar novos caminhos de conhecimento e amadurecimento profissional e pessoal. Minha transição da bioquímica para a bioinformática foi, sem dúvida, uma escolha audaciosa, mas extremamente acertada. Aprendi muito e encontrei minha verdadeira vocação científica na análise de dados biológicos em prol da saúde e evolução humana. O doutorado foi uma jornada árdua, mas incrivelmente enriquecedora. Ao longo desse percurso, conheci pessoas que contribuíram fortemente para minha evolução na informática, por meio de discussões, trocas de experiências e compartilhamento de conhecimento. Em especial, agradeço ao meu amigo Roger, que esteve ao meu lado desde o início do doutorado, sendo meu parceiro de trabalho, um grande apoio e incentivador.

Ao meu orientador, Chico Lobo, agradeço imensamente pela oportunidade de realizar minha pesquisa sob sua supervisão, mesmo sabendo que eu não tinha conhecimento prévio em bioinformática. Desde o começo, sabíamos que este doutorado seria um grande desafio para ambos, e sou grata pelas discussões enriquecedoras e pelo aprendizado ao longo dessa jornada. Sua abordagem rigorosa e seu incentivo ao pensamento analítico foram fundamentais para o meu crescimento científico, ajudando-me a desenvolver um olhar mais preciso e criterioso sobre a interpretação dos resultados. Esse aprendizado certamente me acompanhará ao longo de toda minha carreira.

Ao meu coorientador, Renan, por aceitar o desafio de me coorientar mesmo fora do seu departamento e por contribuir significativamente para a melhoria do meu trabalho. Seu olhar crítico e suas sugestões, especialmente na área de visualização de dados, foram essenciais para aprimorar minhas análises e expandir minha perspectiva científica.

Aos amigos do laboratório Zandora, Thieres e Maycon, por me acolherem desde o primeiro momento e me integrarem à equipe. Um agradecimento especial ao Henry, cuja parceria e apoio foram fundamentais. Sua escuta atenta, suas palavras acolhedoras e sua amizade tornaram essa jornada muito mais leve e especial.

Não posso deixar de agradecer novamente à professora Santuza e ao período que passei em seu laboratório durante o mestrado. Uma cientista admirável, de carreira inspiradora, com quem tive o privilégio de aprender muito sobre pesquisa científica, condução de projetos, métodos e protocolos de biologia molecular.

À minha família – irmãs, irmão e pai – que sempre acreditaram em mim, muitas vezes mais do que eu mesma. Obrigada por compreenderem meus momentos de “surtos” e ausências e por estarem sempre ao meu lado. O amor, o encorajamento e o suporte incondicional de vocês foram essenciais para que eu chegasse até aqui.

Aos meus amigos do grupo de patins, por compartilharem comigo momentos de alegria, carinho e descontração. Vocês foram fundamentais para manter minha saúde física e mental ao longo desses anos e me proporcionaram momentos inesquecíveis de leveza e felicidade.

Ao meu namorado, Fred, pelo encorajamento constante e pelo apoio nos momentos de desânimo. Suas palavras de incentivo e sua confiança em mim renovavam minha determinação e me lembravam do meu potencial sempre que eu duvidava de mim mesma. Obrigada por estar ao meu lado nessa caminhada.

Ao Programa de Pós-Graduação em Bioinformática pela estrutura e suporte oferecidos ao longo do meu doutorado.

Às instituições de fomento, pelo financiamento da minha bolsa de doutorado, viabilizando a realização desta pesquisa.

A todos que, de alguma forma, contribuíram para que eu chegasse até aqui, meu mais sincero agradecimento. Este trabalho é também um reflexo do apoio e da dedicação de cada um de vocês.

“Onde meus talentos e paixões encontram as necessidades do mundo, aí está meu lugar” — Aristóteles

RESUMO

A evolução do genoma e o surgimento de doenças genéticas são impulsionados por mutações, que ocorrem por diferentes mecanismos e impactam a variabilidade genética e a adaptação das espécies. Apesar de décadas de pesquisa sobre variações nas taxas e tipos de mutações em regiões genômicas, poucos estudos integraram análises comparativas de sua distribuição espacial no genoma humano. Neste estudo, realizamos uma avaliação em larga escala da distribuição e caracterização funcional de Variantes de Nucleotídeo Único (SNVs) pertencentes a cinco categorias: (i) polimorfismos neutros ou quase neutros; (ii) variantes raras; (iii) mutações somáticas associadas ao câncer; (iv) variantes germinativas patogênicas de relevância clínica; e (v) variantes benignas. Utilizamos dados dos bancos públicos COSMIC, ClinVar e HGDP para propor três modelos de distribuição mutacional no genoma humano. Nossos resultados revelaram uma sobreposição significativa entre diferentes classes de SNVs, sugerindo que certas regiões genômicas são mais propensas a acumular mutações, independentemente de sua origem germinativa ou somática. Notavelmente, Polimorfismos de Nucleotídeo Único (SNPs) e mutações recorrentes no câncer ocorrem frequentemente nos mesmos loci genômicos e compartilham os mesmos alelos com uma frequência superior ao esperado por acaso. Identificamos cinco assinaturas mutacionais associadas a esses sítios compartilhados (SBS1, SBS5, SBS6, SBS54 e SBS87), enquanto ilhas CpG e microssatélites explicam apenas uma pequena fração dessas mutações. Além disso, análises de enriquecimento funcional apontaram associações significativas entre as regiões sobrepostas e áreas genômicas-chave, como 13q12.12 e 19p13.3, bem como vias relacionadas à sinalização celular e interação da matriz extracelular, incluindo PI3K/AKT/mTOR e ECM-receptor interaction, ambas fundamentais para a progressão tumoral. Com base nesses achados, sugerimos a existência de loci críticos no DNA que atuam como hotspots mutacionais naturais, suscetíveis a mutações recorrentes em ambas as linhagens celulares. A identificação desses hotspots levanta questões sobre os mecanismos subjacentes à sua ocorrência e sua possível contribuição para a predisposição ao câncer.

Palavras-chave: Mutações humanas, Variantes de nucleotídeo único (SNVs), *Hotspots* mutacionais, Loci compartilhados germinativo-somático.

ABSTRACT

The evolution of the genome and the emergence of genetic diseases are driven by mutations, which arise through various mechanisms and shape genetic variability and species adaptation. Despite decades of research on variations in mutation rates and types across genomic regions, few studies have systematically integrated comparative analyses of their spatial distribution in the human genome. In this study, we conducted a large-scale assessment of the distribution and functional characterization of single nucleotide variants (SNVs) across five distinct categories: (i) neutral or nearly neutral polymorphisms, (ii) rare variants, (iii) cancer-associated somatic mutations, (iv) clinically relevant pathogenic germline variants, and (v) benign variants. Using data from public repositories such as COSMIC, ClinVar, and HGDP, we proposed three models to describe the mutational landscape of the human genome. Our findings reveal a significant overlap between different SNV classes, suggesting that certain genomic regions are inherently more prone to accumulating mutations, regardless of their germline or somatic origin. Notably, single nucleotide polymorphisms (SNPs) and recurrent cancer mutations frequently co-occur at the same genomic loci and share identical alleles more often than expected by chance. Five mutational signatures (SBS1, SBS5, SBS6, SBS54, and SBS87) were associated with these shared sites, while CpG islands and microsatellites accounted for only a minor fraction of the observed mutations. Furthermore, functional enrichment analyses identified significant associations between the overlapping regions and key genomic sites, such as 13q12.12 and 19p13.3, as well as pathways involved in cellular signaling and extracellular matrix interactions, including PI3K/AKT/mTOR and ECM-receptor interaction, both of which are crucial for tumor progression. Based on these findings, we propose the existence of critical DNA loci that function as natural mutational hotspots, recurrently affected in both germline and somatic lineages. The identification of these hotspots raises important questions regarding the underlying mechanisms driving their occurrence and their potential role in cancer predisposition.

Keywords: Human mutations, Single nucleotide variants (SNVs), Mutational hotspots, Germline-somatic shared loci

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES E SIGLAS

AF *Allele Frequency* (Frequência Alélica)

BED *Browser Extensible Data* (Formato de arquivo para anotações genômicas)

BER *Base Excision Repair* (Reparo por excisão de base)

CDS *Coding DNA Sequence* (Sequência de DNA codificante)

COSMIC *Catalogue of Somatic Mutations in Cancer* (Banco de dados de mutações somáticas no câncer)

CRISPR *Clustered Regularly Interspaced Short Palindromic Repeats* (Sistema de edição genética)

CpG *Citosina-fosfato-Guanina* (Dinucleotídeo associado à metilação do DNA)

DDR *DNA Damage Response* (Resposta ao dano no DNA)

dMMR *Deficient DNA Mismatch Repair* (Deficiência no reparo de incompatibilidades do DNA)

EBI *European Bioinformatics Institute* (Instituto Europeu de Bioinformática)

ECM *Extracellular Matrix* (Matriz Extracelular)

EMBL *European Molecular Biology Laboratory* (Laboratório Europeu de Biologia Molecular)

FASTQ Formato de arquivo para armazenamento de sequências de DNA/RNA com qualidade de leitura

FDR *False Discovery Rate* (Taxa de Descobertas Falsas)

GATK *Genome Analysis Toolkit* (Pacote de software para análise de dados genômicos)

HGDP *Human Genome Diversity Project* (Projeto de Diversidade do Genoma Humano)

HGF *Hepatocyte Growth Factor* (Fator de crescimento de hepatócitos)

HPV *Human Papillomavirus* (Papilomavírus Humano)

IGSR *International Genome Sample Resource* (Recurso de amostras do genoma internacional)

KEGG *Kyoto Encyclopedia of Genes and Genomes* (Base de dados de vias metabólicas e genéticas)

MANE *Matched Annotation from NCBI and EBI* (Anotação padronizada de genes)

MSI *Microsatellite Instability* (Instabilidade de microssatélites)

NCBI *National Center for Biotechnology Information* (Centro Nacional de Informação em Biotecnologia)

NER *Nucleotide Excision Repair* (Reparo por excisão de nucleotídeos)

NGS *Next-Generation Sequencing* (Sequenciamento de Nova Geração)

ORA *Over Representation Analysis* (Análise de super-representação)

PLINK Software para análise de dados genéticos de grande escala

SBS *Single Base Substitution* (Substituição de base única)

SNP *Single Nucleotide Polymorphism* (Polimorfismo de Nucleotídeo Único)

SNV *Single Nucleotide Variant* (Variante de nucleotídeo único)

UCSC *University of California, Santa Cruz* (Plataforma de bioinformática)

UMAP *Uniform Manifold Approximation and Projection* (Método de redução de dimensionalidade)

UTR *Untranslated Region* (Região não traduzida do RNA)

VCF *Variant Call Format* (Formato de arquivo para variantes genéticas)

VEP *Variant Effect Predictor* (Ferramenta para predição de efeitos de variantes)

WG *Whole Genome* (Genoma completo)

WGS *Whole Genome Sequencing* (Sequenciamento do genoma completo)

SUMÁRIO

1. INTRODUÇÃO

1.1. Mutações

As mutações no genoma humano são alterações na sequência do DNA que ocorrem espontaneamente ou em resposta a fatores externos. Esses eventos podem derivar de erros em processos naturais, como replicação, reparo e recombinação do DNA, ou da exposição a agentes mutagênicos ambientais, como radiações ionizantes, substâncias químicas e certos patógenos virais (Nesta *et al.*, 2021).

Enquanto estímulos endógenos e exógenos podem induzir alterações genômicas, os sistemas celulares de reparo (DDR - *DNA Damage Repair* - **Figura 1**) atuam para preservar a integridade do genoma. Esses mecanismos são essenciais para mitigar mutações potencialmente deletérias, as quais podem levar a doenças genéticas e processos carcinogênicos (Xiang *et al.*, 2023).

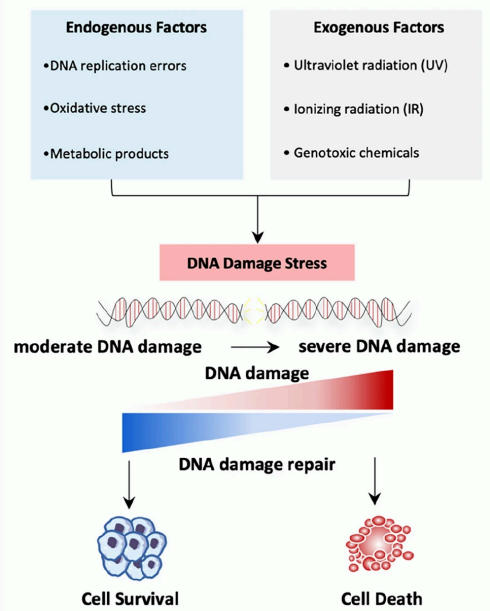


Figura 1: Fontes de alteração no DNA e seus desfechos. O DNA pode sofrer modificações na sequência de nucleotídeos por fatores endógenos, como erros de replicação, estresse oxidativo e produtos metabólicos, ou por fatores exógenos, como radiação ultravioleta (UV), radiação ionizante (IR) e produtos químicos genotóxicos. Enquanto danos moderados, geralmente causados por fatores endógenos, podem ser reparados, permitindo a sobrevivência celular, danos severos, mais comuns após exposições exógenas, frequentemente levam à morte celular. (Fonte: Xiang *et al.*, 2023)

Embora frequentemente associadas a processos patológicos, as mutações constituem a base fundamental da diversidade genética. Estudos populacionais demonstram que indivíduos carregam milhares de variantes genéticas, a grande maioria das quais são neutras ou moderadamente toleradas, contribuindo para a singularidade genética de cada organismo (1000 Genomes Project Consortium, 2015). Isto é, cada indivíduo carrega um número significativo de mutações que, muitas vezes, não têm consequências funcionais ou resultam em alterações moleculares sutis, como a produção de isoformas proteicas funcionalmente equivalentes. Essas mutações neutras ou de pequeno efeito são, inclusive, uma das principais fontes da diversidade genética intraespecífica, conferindo unicidade a cada organismo.

Sob uma perspectiva evolutiva, as mutações constituem uma das matérias-primas da evolução, sendo um dos principais mecanismos responsáveis pela introdução de novas variantes genéticas em uma população. A depender da interação com outros mecanismos evolutivos — como a seleção natural, o fluxo gênico e a deriva genética — mutações podem ser fixadas, eliminadas ou mantidas em equilíbrio ao longo das gerações. A grande maioria das mutações, especialmente as neutras, é amplamente tolerada e pode atingir frequências elevadas na população sem implicar efeitos patológicos.

No contexto deste trabalho, explorar as origens das mutações, seus tipos e consequências fenotípicas é fundamental para compreender os processos que moldam a diversidade genômica humana e sua relação com doenças genéticas. Tradicionalmente, as mutações podem ser classificadas de diferentes maneiras (Yu et al., 2024):

- **Quanto à origem:** germinativas (ocorrem nos gametas e, consequentemente, em todas as células de um organismo) ou somáticas (adquiridas durante a divisão celular de células somáticas e, consequentemente, restritas a linhagens específicas de células-filhas).
- **Quanto ao tipo:** variações estruturais como translocação, inversão, duplicação e outros, ou pequenas mutações como substituições de bases, inserções, deleções.
- **Quanto às consequências fenotípicas/funcionais:** neutras, benignas ou patogênicas, com impacto variando desde alterações mínimas até mudanças substanciais na função proteica.
- **Quanto à frequência alélica (AF):** variantes raras (usualmente definidas como aquelas com AF < 1%) ou comuns (AF ≥ 1%); classificação de grande relevância em estudos populacionais e evolutivos.

Essas classificações ajudam a compreender o papel das mutações em contextos biológicos e médicos, bem como em estudos evolutivos. As características e classificações mais relevantes para o presente trabalho serão discutidas nos tópicos subsequentes.

1.1.1. Origem das mutações

As mutações podem ter origens em diferentes fases do desenvolvimento celular. Quando surgem em células da linhagem reprodutiva (espermatozoides ou óvulos) são chamadas de mutações **germinativas** (**Figura 2**), alterando o material genético que será transmitido aos descendentes e resultando em indivíduos que possuem essa alteração em todas as células de todos os tecidos (Gundry *et al*. 2012, Yu, *et al*. 2024). Por outro lado, as mutações **somáticas** alteram o DNA em qualquer fase da divisão celular, desde a primeira segmentação do óvulo fecundado até as divisões celulares para autorrenovação tecidual. Embora essas mutações não sejam transmitidas para a próxima geração de indivíduos, as células com mutações somáticas transmitem a alteração para todas as suas células-filhas (Gundry *et al*. 2012, Yu *et al*. 2024).

Em uma perspectiva evolutiva, as variantes germinativas são as mais estudadas, uma vez que estas permitem investigar a evolução de linhagens de organismos ao longo do tempo, tendo amplo uso em estudos filogenéticos. Já as variantes somáticas são usualmente estudadas quando ocorrem recorrentemente em diferentes indivíduos e causam alterações fenotípicas de interesse, como no câncer (Perera *et al.* 2016, Cagan *et al*. 2022).

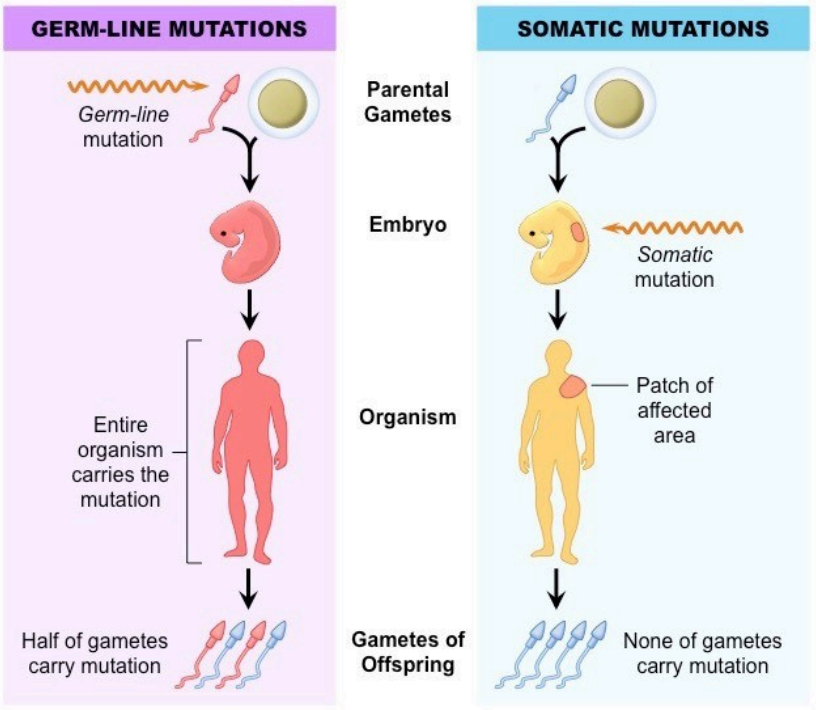


Figura 2: Diferenças entre mutações de origens germinativas ou somáticas. As mutações germinativas (*germline mutations*) ocorrem nas células reprodutivas dos pais (*parental gametes*), no óvulo ou espermatozóide, e alteram o material genético que será passado aos descendentes (são hereditárias). As mutações somáticas (*somatic mutations*) são uma alteração no DNA que ocorre após a concepção de uma pessoa em qualquer célula que não seja germinativa. As mutações somáticas não são hereditárias e acontecem de forma aleatória, sem que a mutação exista na história familiar da pessoa. (Fonte: *Centre for Health Effects of Radiological and Chemical Agents*[1], 2022)

1.1.2. Tipos de mutações

Os tipos de mutação são definidos por meio da análise das variações presentes entre o genoma de determinados indivíduos em comparação a uma sequência de referência. Elas podem ser divididas em duas categorias principais: variações estruturais (SVs) e pequenas mutações (SNVs e Indels) (**Figura 3**).

[1] https://chrc4veterans.uk/articles/articles/techniques-to-study-de-novo-mutations/

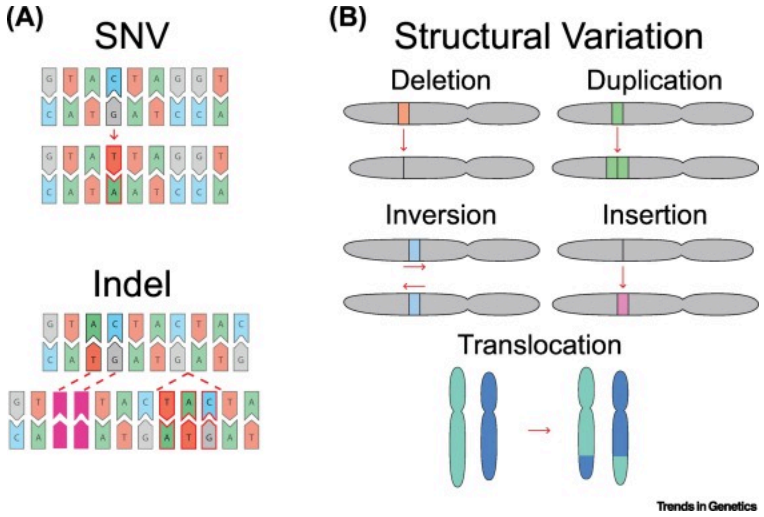


Figura 3: Representação dos principais tipos de variação genética. (A) Pequenas mutações incluem SNVs (*Single Nucleotide Variants*), que consistem na alteração de um único nucleotídeo, e Indels (inserções e deleções de pequenas sequências), onde nucleotídeos podem ser adicionados (inserção) ou removidos (deleção) do DNA. (B) As variações estruturais (*Structural Variations* - SVs) envolvem segmentos maiores de DNA (≥50 pares de bases), incluindo deleção (perda de um fragmento de DNA), duplicação (cópia de um segmento de DNA), inversão (reversão da orientação de um trecho dentro do cromossomo), inserção (adição de um segmento de DNA em uma nova posição) e translocação (transferência de um fragmento entre cromossomos distintos ou do mesmo cromossomo em regiões distintas). Fonte: Nesta *et al.,* 2021.

Resumidamente, as pequenas mutações incluem as Variantes de um Único Nucleotídeo (SNVs - *Single Nucleotide Variants*), que correspondem à substituição de uma única base no DNA, e os Indels (Inserções ou Deleções de Pequenas Sequências), que ocorrem quando nucleotídeos são removidos (deleção) ou adicionados (inserção) à sequência de DNA. Essas pequenas mutações podem afetar a função de um gene dependendo de sua localização no genoma (**Figura 3.A**). As variações estruturais (SVs – Structural Variations) envolvem segmentos maiores do DNA (≥50 pares de bases), podendo afetar extensas regiões do genoma. Nesse contexto, deleções e inserções referem-se à perda ou ganho de grandes fragmentos de DNA. A duplicação (duplication) ocorre quando um segmento do DNA é copiado e aparece repetidamente no genoma. A inversão (inversion) caracteriza-se pela reversão da orientação de um trecho do DNA dentro do cromossomo. Já a translocação (translocation) envolve a transferência de um fragmento de DNA de um cromossomo para outro, resultando na troca de segmentos entre cromossomos distintos (**Figura 3.B**) (Nesta *et al.*, 2021).

Cada tipo de mutação pode apresentar diferentes impactos, dependendo de fatores como da região do genoma onde ocorrem, da quantidade de nucleotídeos envolvidos, do aminoácido que codificam, no caso de genes codificadores, dentre outros. Em regiões codificadoras, quando o número de nucleotídeos inseridos ou deletados de certa região não é múltiplo de três, por exemplo, ocorre uma mudança na fase de leitura do código genético (*frameshift*) podendo levar a consequências mais graves. Mutações desta natureza também podem gerar um códon de parada prematuro, resultando na produção de uma proteína truncada, que frequentemente possui perda de função (Rodriguez-Murillo *et al.*, 2013).

Apesar da detecção de variantes por sequenciamento de próxima geração (NGS) estar se tornando cada vez mais comum e eficiente, este método ainda possui limitações. A detecção de *indels* a partir de *reads* pequenas, por exemplo, ainda representa um grande desafio para a bioinformática, gerando erros na delimitação dos nucleotídeos mutantes e na interpretação de suas consequências funcionais, enviesando conclusões científicas (Kim *et al*., 2017). Por outro lado, SNVs usualmente apresentam localizações genômicas bem definidas e uma detecção computacional altamente eficiente, representando um tipo de mutação que oferece uma abordagem mais precisa e confiável para identificar sítios genômicos específicos com evidências de mutação. Além disso, as SNVs representam uma das alterações mutacionais mais prevalentes no genoma humano, sendo aproximadamente 80% das diferenças genéticas entre dois indivíduos não-aparentados (Katsonis *et al.*, 2014, Trost *et al.* 2021). Esta classe de mutações desempenha um papel crucial na dinâmica da variabilidade genética e em suas implicações fenotípicas, incluindo processos evolutivos e doenças genéticas. Esses fenótipos exibem uma notável variabilidade influenciada pelo contexto genômico, origem e localização das SNVs no genoma (Katsonis *et al.*, 2014, Claussnitzer *et al*., 2020, Fedorova *et al.*, 2022).

1.1.3. Consequências fenotípicas e funcionais das SNVs

As SNVs podem levar a diferentes consequências fenotípicas, desde a ausência de efeitos negativos perceptíveis (fenótipos neutros ou benignos) até alterações funcionais significativas, resultando em doenças de penetrância variável (Claussnitzer *et al.*, 2020). Esse espectro de efeitos está diretamente relacionado a fatores internos e externos, como a localização da variante no genoma (regiões codificadoras ou não codificadoras), à relevância funcional da

região afetada, a composição genômica do indivíduo e o impacto do ambiente externo (Kumar *et al.*, 2023).

Considerando as mutações em regiões codificadoras, as substituições sinônimas (*synonymous*) que, por definição, não alteram o aminoácido codificado, foram inicialmente consideradas neutras, por não causarem alterações nas proteínas. Essas SNVs geralmente não afetam a estrutura primária das proteínas ou ocorrem em regiões genômicas cuja modificação é tolerada (Rentzsch *et al.*, 2019). Entretanto, alguns estudos já demonstram que substituições *synonymous* podem impactar a aptidão evolutiva, afetando, por exemplo, a abundância relativa de tRNAs ou modificando regiões regulatórias, como sítios doadores e aceptores de *splicing* (Shen *et al.*, 2022). Em outros casos, mesmo quando há alteração de um aminoácido, a variante pode ser considerada neutra ou benigna caso não comprometa significativamente a função ou a estabilidade proteica (Pal *et al.*, 2015).

Já as variantes patogênicas apresentam potencial de provocar consequências fenotípicas graves, sendo usualmente associadas a doenças monogênicas ou a fatores de risco para doenças genéticas complexas (Bhattacharya *et al.*, 2017; Claussnitzer *et al.*, 2020, Koprulu *et al.*, 2022; Pandey *et al.*, 2023). Em regiões codificadoras, mutações não-sinônimas (*missense*) podem alterar resíduos essenciais para a função enzimática ou estrutural da proteína (Pal *et al.*, 2015; Cheng *et al.*, 2023), enquanto variantes que introduzam um códon de parada prematuro (*stop gained*) interrompem a síntese proteica de maneira prematura, resultando em proteínas truncadas e frequentemente não funcionais (Karczewski *et al.*, 2020). Além disso, variantes que afetam regiões regulatórias ou de processamento do RNA em sua forma funcional (*splicing*) podem desestabilizar a expressão gênica ou alterar os padrões de processamento de transcritos, levando a manifestações fenotípicas severas (Perera *et al.*, 2016, Wu *et al.*, 2024). Diversos serviços de biologia computacional fornecem classificações categóricas detalhadas dessas e de outras consequências funcionais(e.g. *Ensembl*[2]).

A distinção entre variantes patogênicas e benignas não é absoluta. Algumas SNVs podem apresentar penetrância incompleta ou efeitos moderados, a depender de seu contexto genético e ambiental. A composição genômica dos indivíduos (arquitetura genômica) desempenha um papel crucial na determinação da patogenicidade de muitas mutações, uma vez que variantes em outras regiões genômicas podem amplificar ou atenuar os efeitos de mutações

[2] https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html#consequences

patogênicas. Diferenças populacionais e interações epistáticas também influenciam significativamente a manifestação fenotípica. No entanto, é importante destacar que tanto as bases de dados genômicas quanto os bancos clínicos apresentam viés de representatividade, tanto em relação às populações quanto às regiões genômicas contempladas. A maioria dos estudos genéticos e bancos de variantes patogênicas são baseados majoritariamente em populações de origem europeia, o que pode limitar a transferência e a acurácia clínica dessas informações para grupos sub-representados. Por exemplo, análises populacionais em biobancos revelaram que variantes classificadas como patogênicas em bancos clínicos frequentemente apresentam penetrância reduzida em populações geneticamente diversas, reforçando a necessidade de considerar múltiplos contextos ancestrais e funcionais na interpretação dos efeitos variantes (Ciesielski *et al.*, 2024). Nesse cenário, ferramentas de predição de impacto funcional, como algoritmos de *machine learning* e modelos de estrutura proteica, têm sido empregadas para classificar variantes de forma mais precisa e eficiente (Cheng *et al.*, 2023; Gromiha *et al*., 2024).

1.1.4. Manutenção da frequência alélica de SNVs no genoma humano e os perfis de doenças genéticas associadas

A frequência alélica de uma variante está fortemente relacionada a sua história evolutiva, bem como as suas consequências funcionais e fenotípicas. Substituições *missense* de alto impacto em proteínas essenciais, que comprometem funções críticas do organismo, tendem a ser mantidas em baixa frequência devido à ação da seleção purificadora e restrições evolutivas, uma vez que estas reduzem a aptidão evoutiva do organismo (*fitness*) (Drosou *et al.*, 2016, Omer *et al.*, 2017). Algumas variantes, em contraste, são selecionadas positivamente em genes onde há pressão seletiva para sua variação, em um fenômeno conhecido como seleção positiva Darwiniana; exemplos clássicos deste tipo de seleção são observados nos genes imunes de hospedeiros e nos genes de patogenicidade em parasitos (Hongo *et al.*, 2015). SNVs neutras usualmente apresentam flutuações aleatórias em sua frequência e podem atingir frequências elevadas, uma vez que exercerem pouco ou nenhum impacto na aptidão do indivíduo (Bush *et al.*, 2012; Claussnitzer *et al.*, 2020). As SNVs que ocorrem em AF mais elevada nas populações humanas, geralmente superior a 1% ou 5%, são classificadas como Polimorfismos de Nucleotídeo Único (SNPs) (Risch, 2000; Fedorova *et al.*, 2022). Os SNPs são amplamente utilizados em estudos de genética de populações e análises filogenéticas,

representando uma fração significativa do número total de alelos em uma população (Shakya *et al*. 2020, Fedorova *et al.*, 2022). Outro fator importante na determinação da frequência alélica de uma variante é a deriva genética, especialmente em populações que passaram por eventos como o efeito fundador e o efeito de gargalo de garrafa. Esses fenômenos reduzem a diversidade genética e podem, por acaso, aumentar a frequência de variantes raras, incluindo variantes patogênicas, independentemente da seleção natural. No espectro das doenças humanas genéticas, a relação entre frequência alélica e consequência fenotípica é particularmente evidente. Variantes comuns geralmente apresentam efeitos modestos e sinergísticos, estando associadas a doenças complexas germinativas, como a diabetes tipo 2, enquanto variantes raras, de maior impacto funcional, são frequentemente relacionadas a doenças monogênicas ou oligogênicas, como a fibrose cística. A fibrose cística, por exemplo, é causada por mutações no gene CFTR (*Cystic Fibrosis Transmembrane Conductance Regulator*) e exemplifica a forte relação entre variantes raras de alta penetrância e doenças mendelianas (Kerem, 1989; Bell et al., 2020). Em contraste, a diabetes tipo 2 envolve a contribuição de múltiplas variantes comuns, cada uma com pequeno efeito individual, mas que em conjunto influenciam o risco da doença, frequentemente moduladas por fatores ambientais e epigenéticos

A hipótese das doenças comuns/variantes comuns (CD/CV) postula que variantes comuns em uma população são responsáveis por distúrbios também comuns nessa mesma população (Reich *et al*., 2001). Contudo, variantes comuns tendem a apresentar efeitos mais modestos sobre o fenótipo em comparação com aquelas relacionadas a doenças raras monogênicas ou oligogênicas (Claussnitzer *et al.*, 2020). Por exemplo, um SNP com frequência de 20% em uma população dificilmente seria altamente deletério e causaria diretamente uma doença grave, dominante, e de alta penetrância, pois este resultaria em aproximadamente 20% da população apresentaria o fenótipo, o que seria evolutivamente improvável em um contexto de seleção natural. Em contraste, variantes de menor impacto, que afetam moderadamente a expressão gênica ou aumentam levemente o risco de doença, podem coexistir em frequências populacionais mais altas sem comprometer de forma evidente o *fitness* dos indivíduos (Bush *et al.*, 2012). Assim, ainda que uma variante seja patogênica, o tamanho de seu efeito funcional e sua penetrância são inversamente proporcionais a sua frequência alélica na população, conforme ilustrado na **Figura 4**.

Como consequência dos fatos expostos acima, variantes deletérias de tamanho de efeito elevado que apareçam em alta frequência em populações específicas possivelmente

apresentam este cenário devido a mecanismos evolutivos, como deriva genética ou endogamia. Esses processos podem levar à fixação de variantes prejudiciais em grupos populacionais isolados onde houve eventos históricos que restringiram a variabilidade genética. Nesse cenário, uma vez que a vasta maioria dos indivíduos dessa população possuiria a mesma variante patogênica, a aptidão evolutiva dos indivíduos seria a mesma, uma vez que esta é calculada em relação aos demais membros da população.

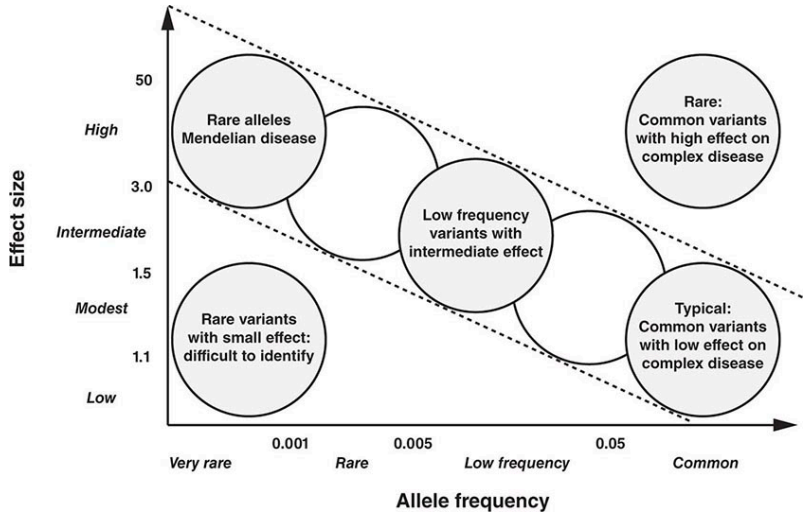


Figura 4. Relação entre frequência alélica e tamanho do efeito (penetrância) das variantes. Variantes podem ser classificadas de acordo com sua frequência alélica na população e o tamanho do efeito fenotípico. Variantes raras, como as associadas a doenças monogênicas, geralmente apresentam grandes efeitos e alta penetrância, mas baixa frequência alélica devido à seleção purificadora. Variantes comuns, por outro lado, frequentemente possuem impactos mais modestos e estão associadas a doenças complexas, como o câncer, onde múltiplas variantes com efeitos pequenos ou moderados contribuem para o fenótipo. Variantes de frequência intermediária também desempenham um papel importante em doenças oligogênicas e condições multifatoriais. Fonte: Basic medical key[3]

1.2. *Hotspots* mutacionais

Embora as mutações ocorram de maneira aleatória no genoma humano, isso não implica que todas as regiões do DNA tenham a mesma probabilidade de sofrer mutações. Em outras palavras, embora não seja possível prever o sítio exato onde a próxima mutação ocorrerá em uma célula, algumas regiões do genoma apresentam maior suscetibilidade à ocorrência de mutações, configurando uma distribuição não-uniforme. Essas regiões são historicamente conhecidas como *hotspots* mutacionais (Wheeler *et al.*, 2020).

Os *hotspots* mutacionais são influenciados por uma combinação de fatores, incluindo estrutura e função das sequências do DNA, características epigenéticas do *locus*, atividade de

[3] https://basicmedicalkey.com/principles-of-human-genetics/

reparo do DNA e interações com fatores indutores de mutação (Nesta *et al.*, 2021; Monroe *et al.*, 2022). Exemplos clássicos de *hotspots* no genoma humano incluem ilhas CpG, microssatélites, DNA centrômero, palíndromos ricos em AT e sequências que formam grampos de DNA (Nesta *et al.*, 2021). No caso das ilhas CpG, a metilação dos dinucleotídeos CG desempenha papel regulador na expressão gênica, mas também as torna propensas a desaminação espontânea. Elas são consideradas um dos principais marcadores epigenéticos, frequentemente encontradas próximas aos sítios de início de transcrição gênica em regiões promotoras (Wheeler *et al*., 2020). Essas regiões são suscetíveis à desaminação espontânea da citosina metilada (5-metilcitosina) em timina, gerando um pareamento incorreto T/G. Caso não ocorram reparos, tais falhas introduzem substituições C/G – T/A durante a replicação do DNA na célula-filha que herda a fita molde com a timina mutada (Nesta et al., 2021).

Essas regiões mais propensas a mutações podem estar associadas a processos evolutivos e doenças genéticas. Em humanos, *hotspots* estão sujeitos a intensa pressão seletiva e, em muitos casos, têm relação direta com doenças complexas, como o câncer (Kilgour *et al.*, 2021; Zou *et al.*, 2022). A maior suscetibilidade dessas regiões a mutações decorre de processos que incluem reparo dependente da transcrição e modificações epigenéticas, os quais moldam o padrão de mutações em diferentes áreas do genoma (Monroe et al., 2022).

Apesar do avanço no estudo de *hotspots* mutacionais, ainda há lacunas no entendimento sobre como mutações neutras, deletérias e benéficas coocorrem nesses sítios no genoma humano. Essa falta de conhecimento limita a compreensão dos padrões de ocorrência de mutações germinativas e somáticas, bem como sua relevância para a evolução e para a saúde humana. Estudos mais aprofundados podem elucidar melhor a contribuição dos *hotspots* mutacionais tanto para o desenvolvimento de doenças genéticas quanto para a adaptação evolutiva (Nesta *et al.*, 2021).

1.3. O papel da evolução na dinâmica mutacional e na diversidade genômica

Mutações são componentes essenciais do processo evolutivo, fornecendo a base primária de variação genética na qual a seleção natural atua. Essas alterações genéticas desempenham um papel crucial na diversificação das populações, além de possibilitar o desenvolvimento de características adaptativas. Contudo, mutações também são determinantes no surgimento de doenças genéticas (Omer *et al.*, 2017; Cvijović *et al.*, 2018).

A evolução atua sobre as mutações germinativas e somáticas, modulando sua frequência, impacto funcional e distribuição genômica. Além dos mecanismos evolutivos como a seleção natural, deriva genética e endogamia, fatores como *hotspots* mutacionais, marcadores epigenéticos e mecanismos de reparo de DNA moldam os padrões mutacionais ao longo das gerações. Enquanto algumas mutações são neutras ou conferem vantagens adaptativas, outras comprometem funções biológicas, podendo ser eliminadas pela pressão purificadora (Claussnitzer *et al.*, 2020, Monroe *et al.*, 2022, Kumar & Gerstein, 2023).

Além dos *hotspots* mutacionais, estudos recentes em *Arabidopsis thaliana* revelaram que certas regiões do genoma apresentariam menores taxas de mutação. Essa descoberta foi possível devido à capacidade de observar rapidamente várias gerações em um organismo de vida curta como *A. thaliana*, algo que é mais difícil de replicar em humanos devido ao longo tempo de geração (Monroe *et al.*, 2022). Esses estudos sugerem que mecanismos de reparo mais eficientes e marcas epigenéticas contribuem para proteger sequências críticas ou funcionais de mutações, o que resultaria em uma menor frequência mutacional nessas regiões. Isso contrasta com áreas mais neutras do genoma, onde as taxas de mutação são mais elevadas. Essa distribuição não-uniforme implica na associação entre pressão seletiva, relevância funcional de *loci* específicos e nível de expressão gênica.

Contudo, o trabalho de Monroe e colaboradores (2022) apresenta limitações importantes. Por exemplo, os autores não puderam avaliar a taxa de mutações não viáveis em genes essenciais (*housekeeping genes* - Wei *et al.*, 2018) que inviabilizam os estágios iniciais da embriogênese. Como o estudo se baseou apenas em mutações germinativas de plantas adultas e em mutações somáticas em tecidos específicos, não foi possível determinar se genes essenciais realmente sofreriam menos mutações ou se as células ou indivíduos com essas mutações não foram observados devido à inviabilidade. Assim, esses achados destacam que certas regiões do genoma são mais ou menos propensas a acumular mutações viáveis, mas não permitem descartar a ocorrência de mutações em regiões críticas inviáveis.

Essas observações reforçam um ponto importante: embora a ocorrência de mutações seja aleatória, sua distribuição ao longo do genoma não é uniforme. Esse conceito, já explorado em estudos de *hotspots* mutacionais há décadas, ganha novos contornos com as descobertas recentes, onde as regiões de ocorrência ou não ocorrência de mutações viáveis estão fortemente associadas a fatores epigenéticos. Assim, estes passam a ter papel importante também em processos evolutivos, confrontando a visão tradicional de que eventos

estocásticos sejam o mecanismo fundamental para a ocorrência de mutações e, consequentemente, para o próprio processo evolutivo (Futuyma 1986, Greenbaum *et al.,* 2014). Em humanos, padrões semelhantes poderiam ser observados, com a seleção purificadora reduzindo a frequência de variantes altamente deletérias e permitindo o acúmulo de variantes de menor impacto em regiões menos essenciais do genoma, abrindo discussão sobre o papel dos marcadores epigenéticos como possíveis moduladores da ocorrência de mutações também no genoma humano (Lynch *et al.*, 2016; Claussnitzer *et al.*, 2020, Monroe *et al.*, 2022).

Essa perspectiva evolutiva não somente reflete como o genoma mantém estabilidade ao longo do tempo, enquanto abriga diferentes graus de variação genética, mas amplia o nosso conhecimento sobre o entendimento de doenças humanas genéticas. Identificar regiões sob forte seleção purificadora e padrões mutacionais desiguais é fundamental para compreender por que algumas regiões do genoma permitem a emergência de variantes que causam doenças graves, enquanto outras regiões permitem o surgimento de polimorfismos neutros ou benignos.

1.4. Abordagens atuais para estudo de mutações

O estudo das mutações no genoma humano tem evoluído significativamente ao longo dos anos, mas a abordagem predominante ainda se concentra em analisar tipos específicos de mutações ou categorias isoladas de doenças. Por exemplo, estudos de genômica frequentemente abordam mutações germinativas ou somáticas separadamente, ou se limitam a um único contexto, como câncer, doenças raras ou comuns (Kumar *et al.*, 2023). Essa abordagem segmentada pode estar dificultando a compreensão holística da dinâmica das mutações no genoma e dos mecanismos subjacentes à sua ocorrência.

No entanto, estudos integrativos têm emergido como uma estratégia promissora para superar essas limitações. Ao combinar dados de diferentes classes de mutações e contextos fenotípicos, estes se estabelecem como métodos para uma compreensão mais abrangente. Alguns autores propõem integrar dados de variantes em diversas doenças para entender melhor os seus impactos funcionais, os padrões compartilhados entre mutações e fatores que modulam suas consequências (Kumar *et al.*, 2023). Esses estudos estão revelando fenômenos como a ocorrência de mutações germinativas e somáticas no mesmo *locus*, indicando a

existência de vulnerabilidades químicas no DNA ou assinaturas mutacionais comuns, como SBS1, SBS87 e DBS7 em momentos distintos (Meyerson *et al.*, 2020; Wang *et al.*, 2024).

Adotar abordagens integrativas não apenas permite mapear *hotspots* mutacionais compartilhados entre variantes somáticas e germinativas, mas também amplia o entendimento da evolução genômica e da patogênese de doenças humanas, promovendo avanços em campos como a medicina personalizada e nos estudos evolutivos.

1.5. Metodologias em bioinformática para estudo mutacional

1.5.1. Produção de arquivos de variantes: do sequenciamento ao VCF

O avanço observado na genômica nas últimas décadas foi impulsionado pelo desenvolvimento e maturação do sequenciamento massivamente paralelo de DNA e RNA, tradicionalmente conhecido como sequenciamento de nova geração (*Next-Generation Sequencing* - NGS) (Abdullah *et al.*, 2020). Este termo se refere a um conjunto de tecnologias que permitem a leitura em larga escala de fragmentos de DNA ou RNA, gerando grandes volumes de dados genômicos com alta precisão. O processamento computacional desses dados envolve diferentes ferramentas bioinformáticas (*pipeline*) e produz arquivos em formatos específicos que estruturam e armazenam as informações genéticas (Abdullah *et al.*, 2020). Essas *pipelines* podem ser conceitualmente divididas em quatro etapas principais: (1) preparação das amostras, (2) sequenciamento, (3) alinhamento das sequências ao genoma de referência e (4) detecção/chamada de variantes (**Figura 5**). Cada uma dessas etapas será descrita brevemente nos tópicos subsequentes.

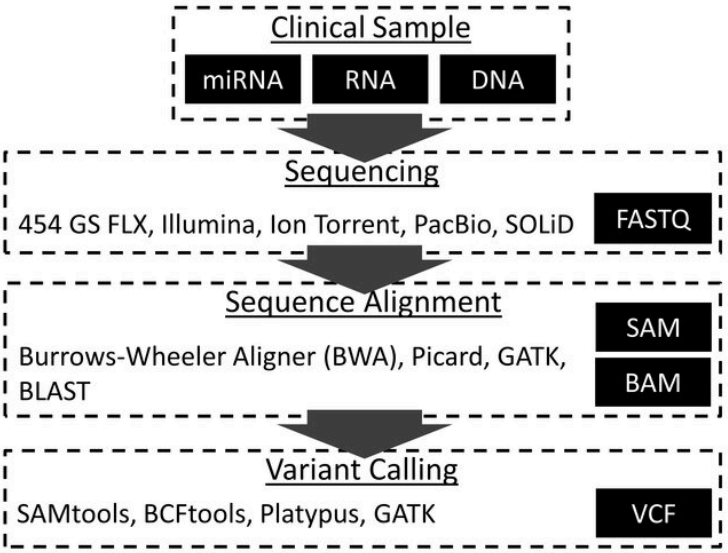


Figura 5. Pipeline de geração de dados genômicos NGS. A figura ilustra as etapas principais do sequenciamento de nova geração (NGS), desde a coleta da amostra clínica, passando pelo sequenciamento (gerando arquivos FASTQ), alinhamento das sequências ao genoma de referência (produzindo arquivos SAM/BAM), até o chamada de variantes, que resulta em arquivos VCF. Ferramentas específicas são utilizadas em cada etapa, como BWA, GATK, e SAMtools (Fonte: Abdullah *et al.*, 2020)

O NGS inicia com a preparação da amostra, que inclui o isolamento do DNA ou RNA, fragmentação e adição de adaptadores para a construção de bibliotecas. A amostra preparada é então sequenciada, atualmente, por plataformas como Illumina, PacBio ou Oxford Nanopore. Os resultados brutos do sequenciamento são usualmente armazenados como arquivos no formato **FASTQ** (**Figura 6**), que contêm as leituras (*reads*) das sequências de nucleotídeos e informações de qualidade para cada base (Cock *et al.,* 2010).

Figura 6. Exemplo de arquivo formato FASTQ. O arquivo contém quatro linhas por sequência. A primeira linha identifica o nome da sequência. A segunda linha apresenta a sequência de DNA. A terceira linha é um separador e a quarta linha contém as pontuações de qualidade, que indicam a confiabilidade da leitura de cada base. (Fonte: (Gen)coded[4] 2025)

[4] https://gencoded.com/index.php/2020/05/20/fastq-format-an-overview/

Os arquivos FASTQ são processados para mapear as sequências ao genoma de referência, utilizando ferramentas como BWA (Li, & Durbin, 2009) ou Bowtie (Langmead *et al.*, 2009). O alinhamento associa cada *read* à sua posição mais provável no genoma de referência. O resultado do processo de mapeamento é armazenado em arquivos no formato **SAM** (*Sequence Alignment/Map*) - **Figura 7**, que pode ser convertido para o formato comprimido **BAM**, mais eficiente para manipulação e armazenamento (Hosseini *et al.*, 2016).

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGAT *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGG   *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCT     * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCA     *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC         * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGC       * NM:i:1
```

Figura 7. Exemplo de arquivo SAM. O arquivo SAM (*Sequence Alignment/Map*) armazena informações sobre o alinhamento de sequências ao genoma de referência. O cabeçalho (@HD e @SQ) especifica a versão do formato e o genoma de referência utilizado. Cada linha subsequente representa uma leitura alinhada, com os seguintes dados: (1) Nome da leitura: identificador único da sequência (ex.: r001); (2) Flag de alinhamento: indica propriedades, como alinhamento direto ou reverso; (3) Nome do genoma de referência: identifica o cromossomo ou sequência-alvo; (4) Posição inicial: local no genoma onde o alinhamento começa; (5) Mapeamento CIGAR: descreve correspondências, inserções, deleções e *soft clipping*; e (6) Colunas adicionais: incluem pontuação de qualidade e anotações específicas. (Fonte: Hosseini *et al.*, 2016)

Após o alinhamento, ocorre a identificação de variantes (chamada de variantes, *variant call*) em relação ao genoma de referência. No caso da detecção de pequenas variantes, como SNVs e pequenos indels, essa etapa é realizada por softwares como GATK, FreeBayes ou Samtools,. O resultado dessa análise é geralmente armazenado no formato **VCF** (*Variant Call Format*) - **Figura 8** (Abdullah *et al.*, 2020). O VCF contém informações detalhadas, incluindo a posição genômica, tipo de variante, genótipos, qualidade da chamada e anotações adicionais. Esse formato é usualmente preferido por ser inequívoco, escalável e flexível, permitindo a inclusão de informações extras no campo "INFO". Além disso, o VCF é eficiente, podendo armazenar milhões de variantes em um único arquivo (Danecek *et al.*, 2011). Os arquivos VCF gerados também podem ser filtrados para remover chamadas de baixa qualidade, bem como podem ser anotados para incluir informações funcionais, como impacto biológico e associações clínicas. Ferramentas como o VEP (*Variant Effect Predictor* - McLaren *et al.,* 2016) e o ANNOVAR (Wang *et al.,* 2010) são utilizadas para enriquecer os dados com informações de bancos de dados biológicos.

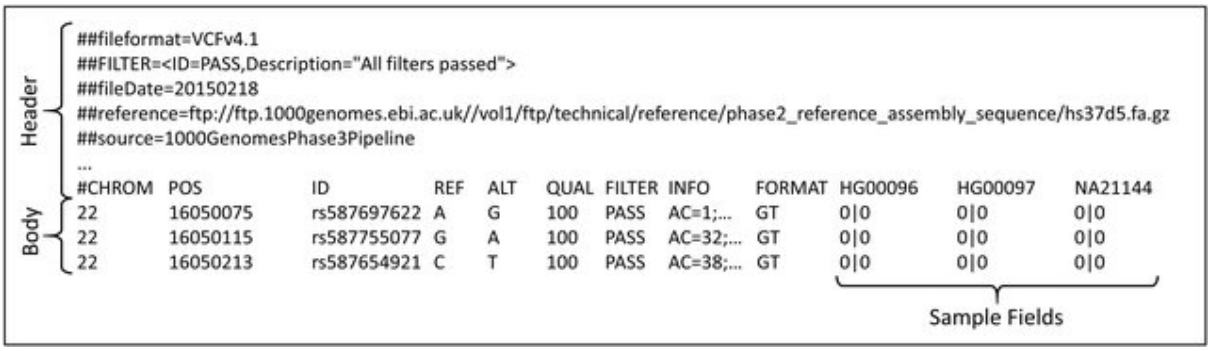


Figura 8. Exemplo de arquivo VCF. O formato VCF (*Variant Call Format*) é usado para armazenar informações sobre variantes genômicas. O arquivo é dividido em duas seções principais: o cabeçalho (*Header*) e o corpo (*Body*). O cabeçalho inclui metadados, como a versão do formato, filtros aplicados e referências genômicas utilizadas. O corpo lista as variantes em colunas: CHROM (cromossomo onde a variante está localizada), POS (posição da variante no cromossomo), ID (identificadores conhecidos, como rsID), REF/ALT (bases de referência e alternativas), QUAL (qualidade da variante), FILTER (filtros aplicados), INFO (informações adicionais), FORMAT (campos de genótipo) e os campos de amostras (Sample Fields), que fornecem dados específicos para cada indivíduo analisado. (Fonte: Abdullah *et al.*, 2020)

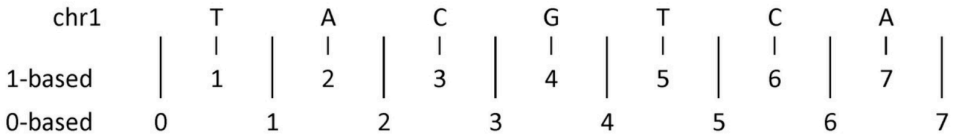
Outros dois formatos de arquivos comumente utilizados na análise de variantes são popularmente conhecidos como BED e GFF. O formato **BED** (*Browser Extensible Data*) é empregado para definir intervalos genômicos de interesse, como regiões-alvo, *motifs* ou *hotspots* mutacionais. Esse formato contém no mínimo três tipos de informação: cromossomo, posição inicial e posição final; opcionalmente, outras informações, tais como nome da região e direção (*strand*) também podem ser incluídos. Sua estrutura facilita a manipulação de grandes conjuntos de dados e a integração com ferramentas de visualização (Niu *et al*., 2022). Os formatos **GFF** (*General Feature Format*) e **GTF** (*Gene Transfer Format*) são utilizados para anotações estruturais e funcionais do genoma, como regiões regulatórias, genes, exons e transcritos.

1.5.2. Sistemas de anotação de coordenadas genômicas

Sistemas de coordenadas genômicas são utilizados para descrever as posições e regiões do genoma nos arquivos gerados a partir do sequenciamento de DNA. Esses sistemas especificam a numeração das bases em sequências de DNA e determinam como intervalos de nucleotídeos são definidos. Existem dois sistemas principais utilizados em bioinformática: o sistema de coordenadas baseado em 1 (*1-based*) e o sistema baseado em 0 (*0-based*) (*The SAM/BAM Format Specification Working Group*[5], 2024).

[5] https://samtools.github.io/hts-specs/SAMv1.pdf

No sistema de coordenadas 1-*based*, a primeira base de uma sequência é numerada como "1" (**Figura 9**). Neste sistema, um intervalo é definido como fechado, ou seja, inclui ambas as bases delimitadoras. Por exemplo, a região que abrange da terceira à sétima base em uma sequência é representada como, incluindo as bases 3, 4, 5, 6 e 7. Este sistema é adotado em formatos como SAM, VCF, GFF e Wiggle. Já no sistema de coordenadas 0-*based*, a contagem começa em "0" para a primeira base da sequência. Os intervalos neste sistema são definidos como meio-abertos, ou seja, incluem a base inicial, mas excluem a base final. Assim, a região correspondente à terceira até a sétima base em uma sequência é especificada como [2, 7), incluindo as bases 2, 3, 4, 5 e 6, mas excluindo a base 7. Este sistema é adotado em formatos como BAM, BED e PSL.



	1-based	0-based
Indicate a single nucleotide	chr1:4-4 G	chr1:3-4 G
Indicate a range of nucleotides	chr1:2-4 ACG	chr1:1-4 ACG
Indicate a single nucleotide variant	chr1:5-5 T/A	chr1:4-5 T/A

Figura 9. Sistemas de coordenadas genômicas 1-*based* e 0-*based*. Comparação dos sistemas de coordenadas genômicas 1-*based* e 0-*based*, usados para especificar a localização de nucleotídeos em uma sequência de DNA. No sistema 1-*based*, a contagem começa em 1, enquanto no sistema 0-*based* a contagem começa em 0. As diferenças entre os sistemas são ilustradas para a indicação de um único nucleotídeo, um intervalo de nucleotídeos e uma variante de nucleotídeo único (SNV). O sistema 1-*based* é usado em formatos como VCF, enquanto o 0-*based* é usado em formatos como BED. (Fonte: https://www.biostars.org/p/84686/).

1.5.3. Uso de bancos de dados biológicos públicos para pesquisas em bioinformática

Os bancos de dados genômicos públicos são ferramentas organizadas que armazenam e disponibilizam informações genômicas, transcriptômicas e proteômicas, essenciais para o desenvolvimento de pesquisas em bioinformática. Esses repositórios possibilitam a integração e análise de grandes volumes de dados biológicos, permitindo descobertas científicas em larga escala e avanços em áreas como medicina personalizada, biologia evolutiva e genômica de doenças. Entre suas principais vantagens estão a acessibilidade gratuita, a padronização de formatos e a integração de dados de diferentes estudos, populações e contextos biológicos (Branco *et al.*, 2021).

Os bancos de dados biológicos podem ser classificados em primários e secundários, de acordo com o tipo de informação que armazenam. Bancos de dados primários, como o NCBI *Sequence Read Archive* (SRA), armazenam dados experimentais brutos, como sequências de DNA e RNA obtidas a partir do sequenciamento de amostras biológicas. Em contraste, bancos de dados secundários, como o Ensembl e o UniProt, integram, processam e anotam informações extraídas de dados primários, fornecendo dados adicionais sobre a funcionalidade de genes, proteínas e variantes genômicas (Mehmood *et al.*, 2014).

Entre os bancos de dados secundários, COSMIC, ClinVar e HGDP se destacam como repositórios importantes para pesquisas envolvendo mutações de diferentes naturezas. O **COSMIC** (*Catalogue Of Somatic Mutations In Cancer*) é um dos maiores repositórios de mutações somáticas recorrentes relacionadas ao câncer. Ele compila dados de variantes encontradas em tumores humanos, como mutações pontuais, alterações estruturais e *indels*, associando-as a tipos específicos de câncer, *genes driver* e efeitos funcionais. Esse banco de dados é amplamente utilizado para identificar padrões de mutações somáticas e mecanismos genéticos subjacentes à oncogênese (Tate *et al.*, 2019).

O **ClinVar** é um banco público que centraliza informações sobre variantes humanas e suas relações com fenótipos clínicos. Entre suas funcionalidades destaca-se a classificação de variantes em uma escala semi-quantitativa das suas consequências funcionais com base em evidências clínicas e funcionais: patogênicas, provavelmente patogênicas, benignas ou de significado incerto. Este banco é extensivamente utilizado em estudos clínicos e genômicos, auxiliando na interpretação e identificação de fatores genéticos associados a doenças genéticas (Landrum *et al.*, 2020).

O **HGDP** (*Human Genome Diversity Project*) é uma referência para estudos populacionais, reunindo dados de variantes germinativas de populações humanas isoladas de diferentes regiões do mundo. Ele fornece informações sobre a diversidade genética humana, sendo de extrema importância para investigações sobre evolução, demografia e seleção natural em populações humanas, além de possibilitar estudos sobre variantes específicas a certas populações (Bergström *et al.*, 2020).

Além de armazenar dados, muitos destes bancos públicos oferecem ferramentas integradas que permitem realizar análises diretamente nas plataformas, como comparações de sequências, análises estatísticas e visualizações. Essas ferramentas tornam o processo

analítico mais acessível e eficiente, especialmente para pesquisadores com recursos computacionais limitados.

1.5.4. Ferramentas de manipulação e análise de arquivos genômicos

Com o advento das tecnologias de sequenciamento de nova geração (NGS), a produção de dados biológicos cresceu exponencialmente, deslocando o foco da coleta para a análise e interpretação de dados. Nesse contexto, a bioinformática desempenha um papel central fornecendo ferramentas computacionais capazes de processar e interpretar informações complexas. Conjuntos de ferramentas integradas, tais como Bioconductor e tidyomics, têm facilitado a comparação de sequências, a identificação de variantes e a inferência de funções gênicas e proteicas (Mehmood *et al.*, 2014, Niu *et al.*, 2020, Hutchison *et al.*, 2024).

Estudos genômicos dependem do uso de ferramentas computacionais capazes de manipular e analisar grandes volumes de dados em formatos padronizados, como VCF, BED e BAM. Essas ferramentas permitem tanto a execução de análises robustas como a extração de informações funcionais e estatísticas dos conjuntos de dados. A seguir, são apresentadas algumas das ferramentas utilizadas neste trabalho, bem como uma breve descrição das suas principais funcionalidades.

O **Bedtools** é uma das ferramentas mais versáteis para a manipulação de arquivos BED. Suas funções permitem realizar operações como interseções, subtrações e testes estatísticos entre conjuntos de intervalos genômicos (Quinlan *et al.*, 2010). Alguns exemplos de operações possíveis incluem: *intersect* - identifica regiões compartilhadas entre dois conjuntos de coordenadas genômicas (**Figura 10**); *subtract* - remove regiões específicas de um conjunto de intervalos, permitindo a exclusão de áreas previamente definidas, como regiões não codificantes ou mal anotadas; *fisher* - realiza o teste exato de Fisher para determinar se a sobreposição entre dois conjuntos de coordenadas genômicas é maior ou menor do que o esperado ao acaso.

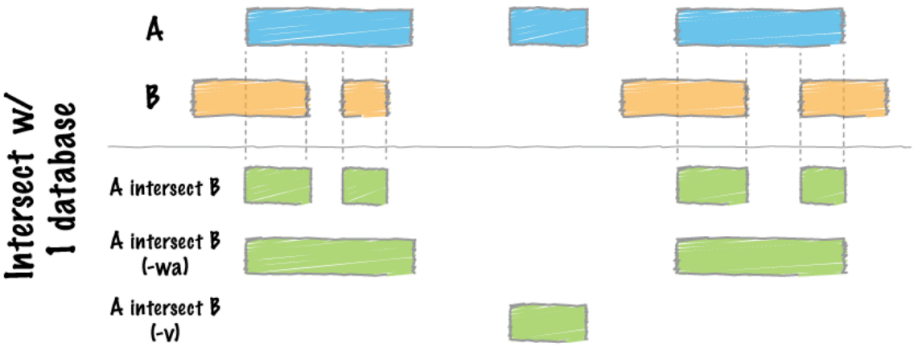


Figura 10. Diagrama ilustrativo do comando intersect do Bedtools, utilizado para identificar sobreposições entre conjuntos de intervalos genômicos. A e B representam dois conjuntos distintos de intervalos genômicos. A intersect B: Retorna as regiões de A que intersectam com B. A intersect B (-wa): Retorna todas as regiões de A que possuem interseção com B, preservando as coordenadas originais de A. A intersect B (-v): Retorna as regiões de A que não apresentam interseção com B.

O **Bcftools** é especializado na manipulação e análise de arquivos VCF, que armazenam informações sobre variantes genômicas (Danecek *et al.*, 2011). Ele oferece uma ampla gama de funcionalidades, incluindo: *filtragem de variantes* - permite filtrar variantes com base em critérios como qualidade, profundidade de cobertura e presença em intervalos específicos definidos em arquivos BED; *concatenação e integração* - combina múltiplos arquivos VCF, útil para integrar variantes de diferentes amostras ou experimentos; *busca de variantes*: realiza buscas rápidas em arquivos VCF indexados para identificar variantes de interesse.

O **VEP**, desenvolvido pelo *Ensembl*, é uma ferramenta para a anotação de variantes genômicas. Ele fornece informações sobre as consequências funcionais de variantes, como alterações na sequência proteica ou impacto em elementos regulatórios (McLaren *et al.*, 2016). Além disso, o VEP suporta a integração de *plugins*, como o AlphaMissense, que prediz a patogenicidade de variantes *missense* com base em modelos de aprendizado de máquina (*deep learning*, aprendizado profundo) (Cheng *et al.*, 2023). Especificamente, o *plugin* AlphaMissense fornece uma pontuação quantitativa de patogenicidade, uma escala contínua que estima a probabilidade de uma variante ser patogênica. Esse *plugin*fornece também uma classificação categórica funcional que classifica variantes nas classes "*likely pathogenic*", "*likely benign*" ou "*ambiguous*", auxiliando assim na priorização de variantes para estudos clínicos.

O **PLINK2** é uma ferramenta utilizada em estudos de genética de populações e genômica que oferece diversas funcionalidades (Chang *et al.*, 2015). O PLINK2 utiliza como *input* arquivos em formato PLINK (.ped/.map ou .bed/.bim/.fam) que podem ser produzidos usando a própria

ferramenta. Dentre as funcionalidades mais utilizadas estão a de *filtragem por desequilíbrio de ligação (LD)* que identifica e remove variantes em LD elevado, com base em parâmetros como tamanho de janela e valor de r^2, e *análises de PCA* (Análise de Componentes Principais) para capturar a estrutura populacional em dados genômicos, diversidade genética e controle de estratificação populacional.

O **SigProfiler Signature** é um algoritmo disponível como uma ferramenta online no banco de dados COSMIC e foi projetado para identificar assinaturas mutacionais em arquivos VCF (Díaz-Gay *et al.*, 2023). As assinaturas mutacionais representam padrões característicos de mutações associados a diferentes processos biológicos causadores de mutações, tais como exposição a agentes mutagênicos, erros no reparo de DNA ou falhas na replicação (Alexandrov *et al.*, 2020). Essas assinaturas são derivadas de análises de variantes somáticas em amostras genômicas e ajudam a identificar os mecanismos subjacentes às mutações. Um exemplo de assinatura é a *Single Base Substitution* (SBS), que analisa substituições de nucleotídeos únicos no contexto de seus trinucleotídeos adjacentes (i.e., as bases antes e depois da mutação). O programa *SigProfiler Signature* categoriza padrões de SBS em assinaturas associadas a processos como exposição à luz ultravioleta (caracterizada por transições C>T quando em um contexto de dímeros de pirimidinas) ou defeitos em vias de reparo, como o reparo por excisão de nucleotídeos (NER). A ferramenta permite explorar graficamente a contribuição de cada assinatura mutacional nas amostras analisadas.

1.6. Justificativa e importância do presente estudo

Os tópicos abordados nesta introdução destacaram conceitos fundamentais sobre mutações, suas origens, tipos e consequências fenotípicas e funcionais, além do papel dos mecanismos evolutivos na modulação da variação genética. O uso de bancos de dados públicos, como COSMIC, ClinVar e HGDP, e ferramentas computacionais para manipulação e análise de dados são componentes centrais para a análise e interpretação dos grandes volumes de dados genômicos.

Neste contexto, o presente trabalho buscou integrar informações desses bancos de dados para investigar não uniformidade e não aleatoriedade de ocorrência das mutações de diferentes origens e com diferentes consequências no genoma humano. Por meio da análise de cinco conjuntos distintos de variantes, classificadas de acordo com suas propriedades biológicas,

como origem (germinativa ou somática), consequências fenotípicas (benignas/neutras ou patogênicas), frequência (comum ou rara) e penetrância (baixa ou alta), nosso objetivo foi avaliar a coocorrência de variantes nesses perfis e determinar se essas interseções ocorrem de maneira significativa, tanto para o genoma completo quanto para regiões codificadoras. Além disso, por meio da caracterização funcional dessas variantes e a avaliação dos subconjuntos que compartilham a mesma posição genômica, buscamos evidências sobre os processos subjacentes à dinâmica mutacional no genoma humano.

Essa abordagem integrativa visou ampliar a compreensão dos padrões de distribuição de mutações de diferentes classificações e suas implicações funcionais, oferecendo um guia para estudos futuros e contribuindo para a interpretação da complexa relação entre variação genética, evolução e doenças humanas, além de indícios acerca dos mecanismos responsáveis pela coocorrência de mutações de diferentes classes.

2. OBJETIVOS

2.1. Objetivo Geral

Investigar a distribuição espacial e a coocorrência de diferentes classes de Variantes de Nucleotídeo Único (SNVs) no genoma humano, explorando possíveis padrões de não uniformidade e não aleatoriedade associados a diferentes perfis de SNVs e seus impactos funcionais e fenotípicos.

2.2. Objetivos Específicos

2.2.1. Caracterizar funcionalmente os conjuntos de SNVs produzidos, identificando regiões genômicas, tipos de mutações e impacto funcional das variantes.

2.2.2. Comparar a sobreposição de *loci* genômicos entre classes distintas de SNVs e avaliar estatisticamente se essas interseções ocorrem mais frequentemente do que o esperado ao acaso.

2.2.3. Avaliar SNVs que coocorrem entre os conjuntos sobrepostos, diferenciando aquelas que compartilham o mesmo nucleotídeo das que ocupam a mesma posição com alelos diferentes, e testar a significância desses fenômenos.

2.2.4. Propor modelos de distribuição espacial das mutações no genoma humano para descrever padrões entre a ocorrência de diferentes classes de SNVs.

2.2.5. Gerar uma base de discussão sobre os padrões de distribuição de SNVs no genoma humano, contribuindo para estudos futuros sobre dinâmica mutacional, evolução e doenças genéticas.

2.2.6. Caracterizar o subconjunto *SNPs vs. COSMIC* e avaliar as assinaturas mutacionais responsáveis pelo padrão de ocorrência semelhante entre variantes germinativas e somáticas.

2.2.7. Investigar se existe correlação entre a frequência de ocorrência de mutações por genes entre os conjuntos SNPs e COSMIC.

2.2.8. Realizar análises de enriquecimento funcional de vias biológicas utilizando o WebGestalt para identificar mecanismos subjacentes às mutações.

3. MATERIAL E MÉTODOS

3.1. Padronização dos Polimorfismos de Nucleotídeo Único (SNPs) em humanos

O banco de dados *Human Genome Diversity Project* - HGDP contém informações sobre populações humanas isoladas classificadas em sete superpopulações: África, América, Ásia Central e do Sul, Leste Asiático, Europa, Oriente Médio e Oceania. SNVs da linhagem germinativa (versão GRCh38) foram obtidas do HGDP por meio das etapas descritas a seguir. Inicialmente, variantes bialélicas e de alta qualidade (parâmetro "PASS" do VCF) foram selecionadas utilizando o software BCFtools versão 1.10.2 (Danecek *et al.*, 2021). Especificamente, o filtro PASS foi aplicado para selecionar apenas variantes que passaram por todos os filtros de qualidade, excluindo aquelas marcadas com LOW_VQSLOD e ExcHet, que indicam baixa qualidade ou excesso de heterozigosidade, respectivamente (Bergström *et al.*, 2020).

Os SNVs autossômicos foram ajustados para Desequilíbrio de Ligação (LD) utilizando o parâmetro 'indep-pairwise' no PLINK2, com *window size* de 200kb, *step size* de 25 SNVs e r2 *value* de 0,4, seguindo os métodos descritos por Borda *et al.*, 2020. Além disso, indivíduos e SNVs com mais de 10% de dados ausentes foram removidos utilizando os parâmetros 'mind 0.1' e 'geno 0.1' no PLINK2, respectivamente (Borda *et al.*, 2020).

A redução da dimensionalidade dos dados foi realizada inicialmente por meio de uma *Principal Components Analysis* (PCA) no conjunto de variantes de alta qualidade definido na etapa anterior, utilizando as configurações padrão do software PLINK2 (Chang *et al.*, 2015). Por meio do PCA, amostras *outlier* de mRNA foram identificadas e o arquivo foi processado para a remoção de amostras de transcriptoma. O resultado da análise PCA foi então utilizado como entrada na biblioteca umap do R para calcular o *Uniform Manifold Approximation and Projection* - UMAP e melhor visualizar a estrutura populacional ao buscar indivíduos isolados nas superpopulações pré-definidas pelo banco de dados (Sakaue *et al.*, 2020). Os parâmetros 'n_neighbors 30', 'min_dist 0.1' e 'spread 5' foram configurados seguindo os métodos descritos por Diaz-Papkovich *et al.*, 2021. Após todas as etapas de processamento de dados, foi definido o limiar para variantes polimórficas como aquelas com frequência alélica global igual ou maior do que 1% (AF $\geq$ 0,01) e que esteja presente em pelo menos duas das sete superpopulações isoladas.

3.2. Identificação de SNVs raras

Por meio do conjunto de dados HGDP, foi estabelecido um conjunto de SNVs germinativas raras, removendo todas aquelas consideradas comuns na população global. Para isso, filtramos as SNVs com AF menor que 1% (AF<0,01) no conjunto do HGDP inteiro considerando todas as populações pré definidas. O processo de filtragem também envolveu a retenção somente das variantes de alta qualidade (PASS) e bialélicas do DNA autossômico. SNVs em LD foram filtradas utilizando o parâmetro 'indep-pairwise' no PLINK2, com *window size* de 200kb, *step size* de 25 SNVs e r2 *value* de 0,4, seguindo os métodos descritos por Borda *et al.*, 2020.

3.3. Mutações somáticas recorrentes no câncer

O banco de dados *Catalogue Of Somatic Mutations In Cancer* - COSMIC (Bamford *et al.*, 2004) fornece dois arquivos VCF separados contendo variantes localizadas em diferentes regiões: um para variantes de regiões codificadoras do genoma humano e outro de variantes em regiões não codificadoras. Os dois arquivos foram utilizados para produzir um único arquivo, usando o *merged* do software BCFTools (Danecek *et al.*, 2021), que foi usado como fonte de mutações de nucleotídeo único somáticas recorrentemente encontradas no câncer. Como etapas de processamento de dados, foram removidas variantes complexas (não SNVs), variantes que resultaram em mais de uma isoforma devido a consequências transcricionais, retendo apenas uma ocorrência, além de variantes redundantes usando um *script* Python personalizado.

3.4. Variantes germinativas clinicamente relevantes e curadas

SNVs humanas de significância clínica (versão GRCh38) foram obtidas no banco de dados público ClinVar, um repositório de variantes humanas com curadoria e vários metadados fenotípicos e funcionais (Landrum *et al.*, 2020). O software BCFtools foi utilizado para todos os procedimentos de filtragem (Danecek *et al.*, 2021). As variantes foram filtradas com base em sua origem, consequência clínica e *status* de revisão, sendo separadas em dois conjuntos distintos conforme os seguintes critérios: I) ClinVar benignas [*Origin*: Germline, *clinical consequence*: benign/likely-benign, and *review status:* 4-stars: practice guideline, 3-stars:

reviewed by expert panel, 2-stars: criteria provided, multiple submitters, no conflicts]. II) ClinVar patogênicas [*Origin*: Germline, *clinical consequence*: pathogenic/likely-pathogenic, and *review status*: 4-stars: practice guideline, 3-stars: reviewed by expert panel, 2-stars: criteria provided, multiple submitters, no conflicts]. Após a filtragem, variantes complexas (não-SNVs) e variantes redundantes foram removidas.

3.5. Filtragem do subconjunto de dados genômicos para SNVs em regiões codificadoras (CDS - *coding regions*)

Foi obtido um conjunto de coordenadas exônicas do MANE (*Matched Annotation from NCBI and EBI*) (Morales *et al.*, 2022) representando o conhecimento biológico canônico de cada *locus* codificador de proteína no genoma humano. Essas coordenadas foram usadas para construção de subconjuntos dos arquivos VCF anotados gerados em etapas anteriores para variantes encontradas em regiões CDS. Analisamos as taxas de transição (TI) e transversão (TV), bem como a razão TI_TV, para cada um dos conjuntos de coordenadas CDS, comparando os alelos alternativos com a referência. Essas análises foram realizadas no R Studio usando scripts personalizados.

3.6. Anotação com En'sembl Variant Effect Predictor (VEP)

O *Ensembl Variant Effect Predictor* - VEP (McLaren *et al.*, 2016) foi utilizado para realizar a anotação de variantes. Cada um dos cinco diferentes conjuntos de dados foi processado usando VEP, tanto os arquivos de coordenadas do WG quanto os de regiões CDS. A gravidade das consequências foi categorizada[1] com base na região genômica, consequências de codificação mais graves e tamanho do efeito esperado do impacto funcional, conforme descrito abaixo:

- Abordagem 1: Categorização das consequências codificadoras por regiões genômicas: *1) Splicing site: splice donor variant, splice acceptor variant, splice region variant, splice donor 5th base, splice donor region, splice polypyrimidine tract. 2) Intron: Intron variant. 3) Intergenic: regulatory region, TF binding site, intergenic variant, upstream gene variant, downstream gene variant. 4) Exon: synonymous variant, missense variant, inframe insertion, inframe deletion, stop gained, frameshift variant, coding sequence variant, start retained variant, start lost, stop lost, stop retained*

variant, incomplete terminal codon variant. 5) UTR: 5 prime UTR variant, 3 prime UTR variant. 6) OTHERS: transcript ablation, transcript amplification, protein altering variant, mature miRNA variant, NMD transcript variant, non coding, transcript exon variant, non coding transcript variant, TFBS ablation / amplification, feature elongation, feature truncation, coding transcript, variant, sequence variant.

- *Abordagem 2: Categorização por consequências codificadoras mais sev*[6]*eras: 1) Stop gained, 2) missense variant, 3) synonymous variant, 4) start and stop lost 5) stop retained 6) splice variant.*
- Abordagem 3: Categorização das consequências codificadoras pelo tamanho dos impactos funcionais: *1) HIGH: transcript ablation, splice acceptor variant, splice donor variant, stop gained, frameshift variant, stop lost, start lost, transcript amplification, feature elongation, feature truncation, 2) MODERATE: inframe insertion, inframe deletion, missense variant, protein altering variant, 3) LOW: splice donor 5th base variant, splice region variant, splice donor region variant, splice polypyrimidine tract variant, incomplete terminal codon variant, start retained variant, stop retained variant, synonymous variant.*

O *plugin* AlphaMissense, integrado ao VEP, foi incluído na anotação para relatar a previsão da patogenicidade das variantes *missense* (Cheng *et al.*, 2023). Este *plugin* é compatível com as versões de referência genômica GRCh37 (hg19) e GRCh38 (hg38).

3.7. Avaliação de interseções de coordenadas genômicas

Para avaliar a significância da sobreposição entre as classes distintas de variantes, foi utilizado o teste exato de Fisher conforme implementado no BEDTools (Quinlan *et al.*, 2010). Foram usadas as configurações padrão da ferramenta junto com três arquivos de entrada: dois arquivos BED contendo as coordenadas genômicas dos SNVs de diferentes classes obtidas neste estudo e um arquivo descrevendo o espaço total no qual as coordenadas podem ser observadas. No caso dos testes WG, esse arquivo corresponde aos tamanhos dos cromossomos autossômicos (GRCh38). Quanto aos dados de CDS obtidos pelo subconjunto MANE, inicialmente foi usada a anotação VEP para obter a coordenada CDS de cada variante

[6] critérios estabelecidos conforme parâmetros do site ensembl: https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html

anotada por meio de um *script* Perl personalizado. As informações de comprimento do CDS do arquivo gff descrevendo o subconjunto MANE foram reunidas.

A interseção destes conjuntos foi avaliada usando três arquivos de entrada: dois arquivos BED contendo as coordenadas das variantes e um arquivo de texto contendo o tamanho de cada CDS ou, no caso de WG, de cada cromossomo. Para ambas as análises, foi calculada a proporção de interseções e avaliada a associação entre as coordenadas genômicas de cada par de conjuntos de dados com base no valor de *odds ratio.*

O teste cria uma tabela 2×2 baseada nas seguintes categorias: o número de interseções (in_a/in_b), os números de intervalos únicos para cada arquivo (in_a/not_b ou not_a/in_b) e, a partir de uma estimativa do número de todos os intervalos possíveis, calculada utilizando o tamanho de cada cromossomo (WG) ou gene (CDS), o número estimado de coordenadas não sobrepostas (not_a/not_b).

	in_b	**not_in_b**
in_a	x_{11}: A ∩ B	x_{12}: A - B
not_in_a	x_{21}: B - A	x_{22}:

Os arquivos BED da interseção entre *SNPs vs. COSMIC* foram produzidos usando a função *intersect* do BEDtools para WG. O teste de Fisher do BEDtools foi realizado entre as coordenadas genômicas do subconjunto *SNPs vs. COSMIC* e das ilhas CpG ou com as coordenadas dos microssatélites para o genoma completo. As coordenadas das ilhas CpG e microssatélites foram obtidas no banco de dados do *UCSC genome browser*.

3.8. Análise comparativa de alelos sobrepostos

Foi realizada uma análise comparativa dos alelos encontrados nas regiões de interseção entre os dois conjuntos de coordenadas genômicas de diferentes classes de variantes para avaliar se os alelos alternativos eram idênticos ou diferentes. Calculamos a proporção de alelos idênticos e, em seguida, uma análise de padrões de substituição de nucleotídeos foi realizada, onde os alelos substituídos foram avaliados (por exemplo, T>G, A>C, C>G, etc.), comparando os alelos alternativos com a referência. As análises foram realizadas para todos os subconjuntos de SNVs que coocorreram nas mesmas posições.

Para avaliar se a frequência de alelos idênticos é maior do que o esperado ao acaso, foram construídas tabelas de contingência para cada subconjunto de SNVs sobrepostas, e as frequências esperadas foram calculadas com base em um modelo de aleatoriedade. O Teste Exato de Fisher (*fisher.test()*) do pacote stats do R foi aplicado utilizando os dados dessas tabelas. Considerando que, além do alelo de referência, existem três possibilidades de nucleotídeos (alelos diferentes) e uma possibilidade de alelo idêntico, e assumindo que 1) todos os nucleotídeos possuem uma frequência de 25% no genoma humano; 2) as probabilidades de trocas entre os nucleotídeos são as mesmas (transições *versus* transversões), os cálculos das frequências esperadas assumem uma chance de 25% de coincidência (alelo idêntico) e 75% de discordância (alelos diferentes), caso as variantes fossem distribuídas aleatoriamente.

- A probabilidade de um alelo ser **igual** (ao acaso) é: P (igual) = $\frac{1}{4}$ = 0,25
- A probabilidade de um alelo ser **diferente** (ao acaso) é: P (diferente) = $\frac{3}{4}$ = 0,75

Portanto, as frequências esperadas foram calculadas considerando essas proporções. Para cada interseção:

Esperados iguais = Interseção × P (igual)
Esperados diferentes = Interseção × P (diferentes)

Essas análises foram realizadas no R Studio usando scripts personalizados.

3.9. Assinaturas mutacionais

Novos arquivos VCF foram produzidos a partir do VCF original fornecido pelo COSMIC (*merged* dos VCFs de regiões codificadoras e não codificadoras) por meio de filtragem usando o BEDtools *intersect*, selecionando apenas SNVs localizados nas coordenadas de interesse do subconjunto *SNPs vs. COSMIC,* identificadas a partir da interseção entre os conjuntos de SNVs COSMIC e SNPS. O VCF resultante foi então usado como *input* para avaliar as assinaturas mutacionais através da versão *web* da ferramenta *Cosmic SigProfiler Assignment*, assinaturas de referência "*Cosmic v3.4*" e parâmetro "*Whole Genome Sequencing"* para identificar e atribuir assinaturas mutacionais conhecidas

3.10. Análise de Enriquecimento Funcional (ORA)

A análise de enriquecimento funcional foi conduzida utilizando o método de Over-Representation Analysis (ORA) por meio da ferramenta WebGestalt. Os genes foram selecionados com base em mutações coincidentes em termos de posição genômica, sendo consideradas ocorrências únicas, resultando em um total de 6.149 genes analisados.

As categorias avaliadas incluíram vias metabólicas do KEGG e localizações genômicas baseadas em bandas citogenéticas. A significância estatística foi avaliada utilizando o método de correção para múltiplos testes pelo False Discovery Rate (FDR), adotando FDR < 0,05 como critério para identificar categorias significativamente enriquecidas. Os arquivos gerados pela análise de enriquecimento foram utilizados para a produção de gráficos utilizando o R.

4. RESULTADOS

4.1. Descrição dos conjuntos de dados

Neste trabalho, integramos cinco conjuntos de dados genômicos para representar diferentes classes de Variantes de Nucleotídeo Único (SNVs). A adoção dessa abordagem integrativa e em larga escala foi utilizada para elucidar o cenário de distribuição espacial das variantes de diferentes origens, frequências alélicas e consequências fenotípicas ao longo do genoma humano e relacioná-las com possíveis implicações evolutivas e doenças humanas.

Os dados de coordenadas genômicas foram selecionados para cinco conjuntos de SNVs, tanto para o genoma completo (WG) quanto para regiões codificadoras (CDS) (vide Metodologia). Os arquivos VCF provenientes dos bancos de dados HGDP, COSMIC e Clinvar foram processados para produzir arquivos BED descrevendo coordenadas genômicas que representassem cinco diferentes classes de SNVs: 1. SNPs (HGDP), 2. Rares (HGDP), 3. Recorrentes no câncer (COSMIC), 4. Benign (benignas ClinVar), 5. Pathogenic (Patogênicas ClinVar).

De posse das informações dos bancos de dados e estratégias de filtragem, definimos os cinco conjuntos de SNVs em função de atributos como a origem da variante (germinativa ou somática), a consequência clínica/fenotípica (benigna ou patogênica), a frequência alélica (elevada, intermediária ou baixa) e o tamanho de efeito/penetrância (grande ou pequeno) (**Tabela 1**).

Tabela 1. Características dos conjuntos de coordenadas genômicas avaliadas.

	Origem da variante	**Consequência clínica esperada**	**Frequência alélica esperada**	**Tamanho de efeito esperado**
SNPs_HGDP	Germinativa	Neutra / Benigna	Elevada	Pequeno
Rares_HGDP	Germinativa	Patogênica ou Neutra	Baixa	Grande ou Pequeno
COSMIC	Somática	Patogênica	Intermediária	Intermediário
Clvr_Benign	Germinativa	Neutra / Benigna	Intermediária	Pequeno
Clvr_Pathogenic	Germinativa	Patogênica	Baixa	Grande

Após o processamento e etapas de filtragem dos arquivos VCF, observamos uma redução significativa no volume de dados, como era esperado (**Tabela 2**).

Tabela 2. Resumo das coordenadas genômicas de SNVs para WG e CDS. *Input:* Número de ocorrências nos arquivos VCF de entrada. *Output*: Contagem de coordenadas genômicas de SNVs nos arquivos BED de saída após o processamento, tanto para WG quanto para CDS. Os dados foram obtidos de três bases de dados (COSMIC, ClinVar e HGDP), resultando em cinco conjuntos de dados: COSMIC, Benignas, Patogênicas, SNPs e Rares.

	HGDP		**COSMIC**		**ClinVar**	
Input	75.385.302		**Coding** 37.261.323	**No coding** 31.368.583	1.303.558	
Output *WG*	**SNPs** 2.140.477	**Rares** 50.239.810	10.587.546		**Benign** 76.971	**Pathogenic** 16.313
Output *CDS*	16.040	496.791	3.662.887		56.794	13.098

4.1.1. Identificação de variantes polimórficas e estudo de estrutura populacional

O banco de dados HGDP apresentava inicialmente 75.385.302 variantes, classificadas como sendo de alta qualidade em relação à chamada de variantes, obtidas de 929 indivíduos provenientes do HGDP (**Tabela 2**). Este conjunto representa variantes germinativas de 54 diferentes populações humanas isoladas de sete principais superpopulações continentais: África, América, Ásia Central e Sul, Leste da Ásia, Europa, Oriente Médio e Oceania.

O conjunto de coordenadas de SNVs polimórficas, gerado a partir dos dados do HGDP, foi definido por variantes que apresentavam frequência alélica (*Allelic Frequency,* AF) igual ou superior a 1% na população global e que estivessem presentes em pelo menos duas das sete superpopulações isoladas (definidas a seguir por PCA-UMAP). Esse critério foi adotado para evitar a inclusão de mutações restritas a uma única população, que poderiam estar presentes em alta frequência devido a mecanismos evolutivos como deriva genética, efeito fundador ou endogamia. Tais eventos ocorrem, por exemplo, quando um grupo pequeno e isolado de ancestrais se reproduz internamente ao longo de gerações, fixando mutações raras — e, eventualmente, patogênicas — em comparação com o alelo ancestral (Greenbaum *et al*., 2014). Nesse contexto, variantes deletérias associadas a doenças podem persistir em algumas populações. Portanto, a seleção de SNVs com AF elevada em duas ou mais populações reduz a chance de incluir essas variantes deletérias, permitindo que o conjunto seja classificado como representativo de sítios polimórficos neutros ou quase neutros.

Foi realizada uma análise para definir o limiar de AF utilizado na identificação de polimorfismos. Os gráficos da **Figura 11** ilustram o número de SNVs encontradas nas diferentes superpopulações humanas, definidas pelo banco de dados, considerando dois pontos de corte de AF (1% e 5%) em função do número de populações nas quais cada variante

está presente. Como esperado, observou-se uma diminuição acentuada no número de variantes polimórficas à medida que o limiar de AF aumentava em função do decréscimo no número de variantes compartilhadas entre múltiplas populações. A maior redução no número de variantes foi observada ao aplicar o critério de presença em pelo menos duas populações, sugerindo que muitas variantes com alta frequência global são, na verdade, um reflexo de um viés em uma única população, uma vez que estas não são observadas em outras superpopulações. Essa distribuição restrita pode ser explicada por eventos evolutivos como a deriva genética ou por efeito fundador, que levariam à fixação dessas variantes em populações específicas ao longo do tempo. Valores de AF entre 1% e 5% são frequentemente adotados em estudos de genética de populações. Neste estudo, optou-se por um limiar de AF ≥ 1% na população global, pois este critério já seria suficientemente restritivo ao exigir elevada AF global e presença em pelo menos duas populações isoladas ao mesmo tempo em que minimiza a perda desnecessária de dados informativos.

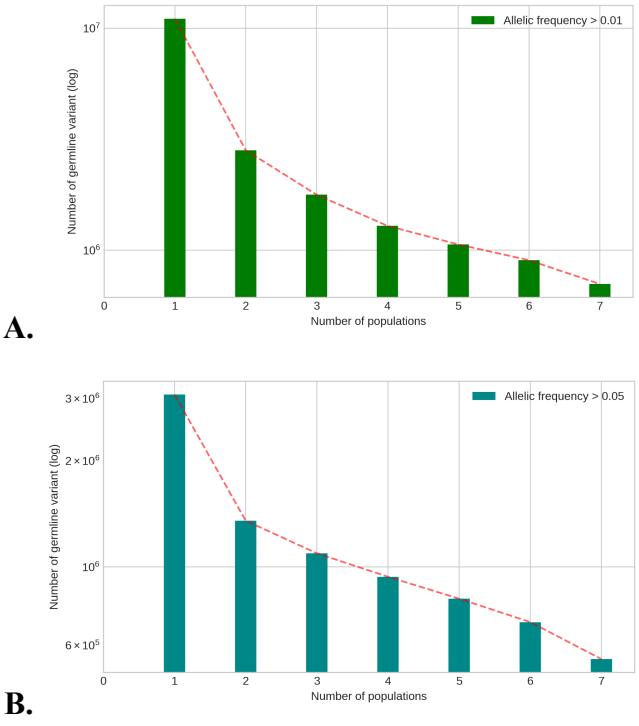


Figura 11. Definição do limiar de AF para SNVs polimórficas. Número de SNPs encontrados em quantidades crescentes de populações humanas distintas considerando uma frequência alélica de 1% (**A**) e 5% (**B**) para o genoma humano completo, variando pelo número de populações 1 a 7 onde a condição é verdadeira. Eixo Y em escala logarítmica.

Para analisar a variação genética e as relações populacionais dos indivíduos previamente classificados em sete superpopulações pelo banco de dados, aplicamos uma abordagem

combinando duas técnicas de redução de dimensionalidade: PCA e UMAP (Sakaue *et al.*, 2020; Diaz-Papkovich *et al.*, 2021). Inicialmente, o PCA foi empregado para reduzir a dimensionalidade dos dados genéticos, capturando a variação genética global e destacando os componentes principais (PCs) mais informativos. Em seguida, aplicamos o UMAP sobre esses PCs, permitindo a preservação de relações locais entre os indivíduos e a identificação de padrões de agrupamento genético, além de revelar subestruturas populacionais sutis que poderiam não ser evidentes no PCA isolado.

Duas amostras africanas *outliers*, HGDP01029 e HGDP00992, foram identificadas por meio do PCA (**Figura 12**). Após a revisão do banco de dados, constatou-se que essas amostras, junto com outras 33, eram provenientes de mRNA, totalizando 35 amostras de transcriptoma (**Figura 13**). A detecção de variantes a partir do sequenciamento de DNA oferece uma visão mais abrangente do genoma, incluindo regiões não transcritas ou de baixa expressão. Em contraste, a análise de variantes derivadas de mRNA é limitada aos mRNAs expressos no momento da coleta, resultando em menor cobertura de SNVs em comparação ao DNA, especialmente em tecidos onde certos genes não estão ativos. Por essa razão, as 35 amostras de mRNA foram descartadas, restando um total de 894 amostras obtidas a partir de tecnologias de sequenciamento de alta cobertura, como *PCR-free high coverage* ou *High coverage WGS*.

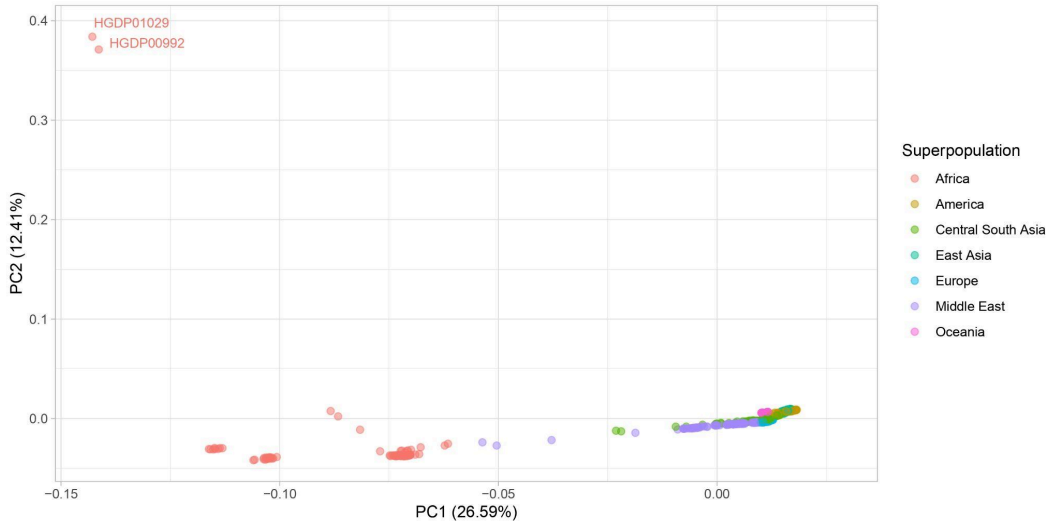


Figura 12. Visualização das amostras do HGDP por meio de análise de componentes principais (PCA). Os dois primeiros componentes principais (PC1 e PC2) explicam 26,59% e 12,41% da variação genética, respectivamente, totalizando 39% da variância capturada. Foram identificadas duas amostras africanas *outliers*, HGDP01029 e HGDP00992.

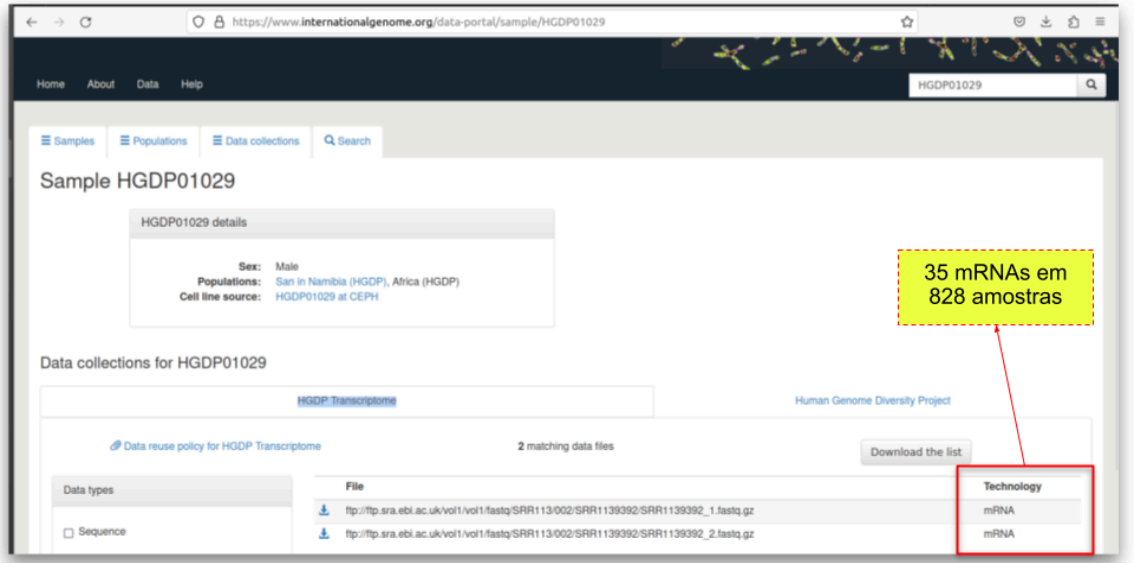


Figura 13. Identificação e avaliação das amostras *outliers* por meio do banco de dados IGSR (*The International Genome Sample Resource*). A figura mostra a página do *IGSR*[7] onde a amostra HGDP01029, assim como a HGDP00992, é listada com a técnica de sequenciamento de mRNA (transcriptoma), em vez de sequenciamento do genoma completo. Essa identificação pode explicar por que essas amostras apresentaram comportamento discrepante na análise genômica sendo consideradas *outliers*.

Após a exclusão dos dois *outliers* e das demais 33 amostras de transcriptoma, o PCA foi repetido, revelaram um padrão de estruturação populacional parcialmente consistente com a literatura (**Figura 14**). A população africana (n=88) destacou-se claramente dos demais grupos, refletindo sua maior diversidade genética e histórico evolutivo distinto, conforme reportado por Tishkoff *et al*. (2009). No entanto, as demais superpopulações, incluindo América (n=51), Ásia Central/Sul (n=181), Leste Asiático (n=193), Europa (n=137), Oriente Médio (n=153) e Oceania (n=25), exibiram sobreposição significativa nos componentes principais. Essa limitação pode ser atribuída a fatores inter-relacionados ao próprio método do PCA. Primeiro, sua natureza linear, que projeta os dados em eixos ortogonais, tende a comprimir relações não lineares complexas, como gradientes contínuos de miscigenação entre populações geograficamente próximas (Reich *et al.*, 2008). Segundo, o desequilíbrio no tamanho amostral entre grupos (variando de 25 a 193 indivíduos) pode introduzir distorções na representação da variância, fazendo com que populações menores, como a da Oceania, tenham menor influência na definição dos componentes principais (Patterson *et al.*, 2006, Elhaik, 2022). Terceiro, a presença de regiões genômicas sob seleção natural ou com desequilíbrio de ligação extenso pode dominar artificialmente os primeiros PCs, mascarando variações demográficas mais sutis (Novembre & Stephens, 2008; McVean, 2009). Dessa

[7]Link de acesso ao banco de dados IGSR: https://www.internationalgenome.org/data-portal/data-collection/hgdp

forma, os padrões observados devem ser interpretados com cautela, pois inferências baseadas exclusivamente no PCA podem ser afetadas por distorções estruturais e pela incapacidade do método de representar gradientes contínuos de miscigenação e agrupamentos não lineares. Complementamos a análise com o UMAP, um método não linear particularmente eficaz em preservar relações locais entre indivíduos e menos sensível a desequilíbrios amostrais, conforme estabelecido por Sakaue *et al.* (2022).

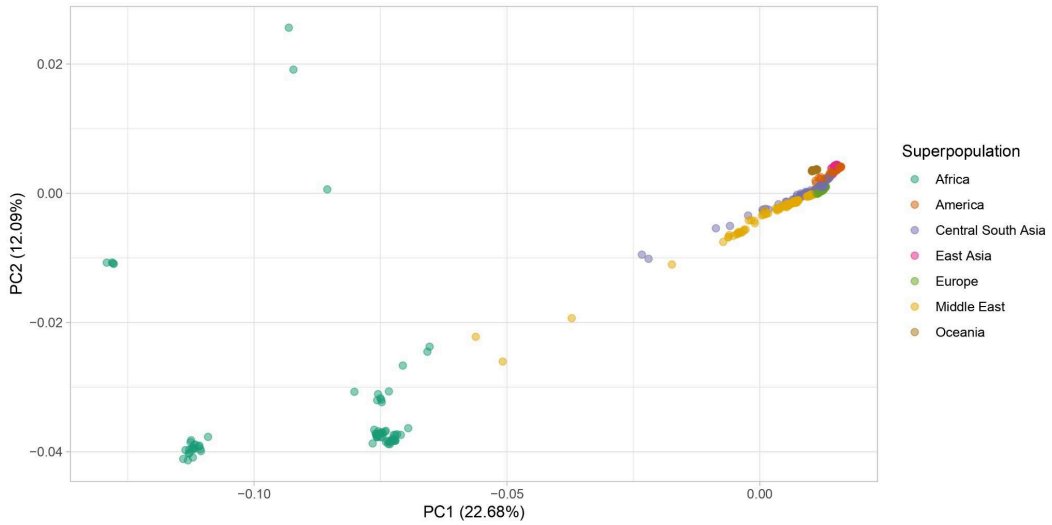


Figure 14. Análise de Componentes Principais (PCA) das amostras do HGDP após filtragem de outliers e exclusão de amostras de transcriptoma. Distribuição das amostras genômicas agrupadas por superpopulações: África, América, Ásia Central e Sul, Leste Asiático, Europa, Oriente Médio e Oceania. O eixo PC1 explica 22,68% da variação genética, enquanto o eixo PC2 explica 12,09%. A separação das superpopulações reflete a diversidade genética global e evidencia as diferenças populacionais entre os continentes.

A aplicação do UMAP após o PCA no cenário global resultou em agrupamentos populacionais mais definidos, evidenciando as relações de vizinhança entre indivíduos e permitindo uma separação mais clara entre as superpopulações (**Figura 15**). Isso ocorre porque o PCA, além de reduzir a dimensionalidade dos dados, minimiza ruídos e preserva as direções de variação global, enquanto o UMAP enfatiza relações locais, melhorando a distinção entre populações geneticamente próximas. Para garantir uma representação mais rigorosa das superpopulações, selecionamos apenas os indivíduos que melhor se agrupavam com seus pares populacionais dentro do espaço UMAP. A seleção foi baseada na proximidade dentro dos agrupamentos, removendo amostras com padrões de agrupamento inconsistentes que poderiam refletir ruído genético ou sinal de miscigenação recente. Essa estratégia teve como objetivo assegurar que a amostra final fosse mais representativa, conferindo maior robustez às análises subsequentes.

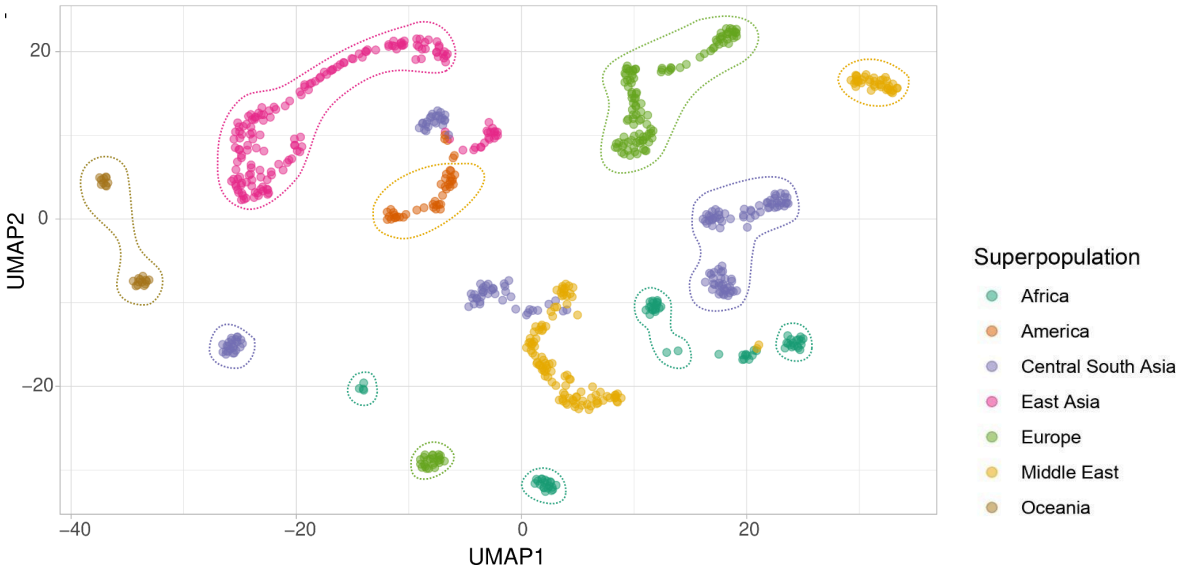


Figura 15. Visualização dos dados do HGDP por PCA-UMAP. Representação bidimensional por PCA-UMAP das sete superpopulações globais: África, América, Ásia Central e Sul, Leste Asiático, Europa, Oriente Médio e Oceania. Os círculos ao redor dos pontos indicam indivíduos considerados não miscigenados, escolhidos para representar a variabilidade genética característica de suas respectivas populações. A separação dos *clusters* reflete a diversidade genética e a estrutura populacional global, enquanto as distâncias e agrupamentos locais são otimizados para preservar as relações de proximidade genética.

Após a filtragem, a distribuição final dos indivíduos por população foi: África: 62; América: 39; Ásia Central e do Sul: 111; Leste Asiático: 153; Europa: 135; Oriente Médio: 39; e Oceania: 23 indivíduos.

Com base no limiar estabelecido para definição dos SNPs (AF > 1% em pelo menos duas populações previamente definidas no banco de dados e reavaliadas pelas análises PCA-UMAP), além das demais etapas de filtragem, observamos uma redução significativa no número de SNVs: aproximadamente 65% para o genoma completo e 99% para as regiões CDS. O processamento e a filtragem dos dados resultaram em 2.140.477 SNPs no conjunto WG e 16.040 SNPs nas regiões CDS (**Tabela 2** - SNPs).

4.1.2. Outras classes de SNVs: raras, associadas ao câncer e de relevância médica

Um conjunto de SNVs raras, também construído a partir dos dados do HGDP, foi definido como as variantes com AF menor do que 1% (AF<0.01). Ressaltamos que, embora ambos os conjuntos do HGDP (SNPs e Rares) derivarem do mesmo banco de dados, eles são

mutuamente excludentes, dados os critérios de corte utilizando AF em ambos os casos, uma vez que consideramos somente variantes bialélicas.

Espera-se que uma fração substancial de SNVs prejudiciais ou patogênicas sejam raras devido à ação da seleção natural. Variantes patogênicas de tamanho de efeito considerável tendem a ser encontradas em baixa frequência, uma vez que comprometem a aptidão evolutiva do indivíduo que as carrega. Assim, uma parte substancial das doenças genéticas raras pode ser atribuída a SNVs de baixa frequência. Por outro lado, variantes raras também podem representar mutações novas (*de novo*) no contexto evolutivo, podendo ter surgido na gametogênese que resultou no indivíduo que a possui ou em gerações recentes de seus antepassados e, portanto, ainda não disseminadas entre a população (Karczewski *et al.*, 2020). Dentre as 75.385.302 variantes do HGDP, identificamos 50.239.810 variantes raras para WG, resultando em uma redução de 33,35%; e 496.791 para CDS, representando uma redução de 99,34% (**Tabela 2**).

O banco de dados COSMIC foi o escolhido como fonte de mutações somáticas de nucleotídeo único, recorrentes no câncer. Esse banco compreende um dos mais amplos e mais utilizados recursos de pesquisa na genética do câncer. Esse recurso compila e organiza informações sobre variantes de células tumorais, oferecendo uma base de dados para uma ampla variedade de tumores e tipos de câncer. O COSMIC abrange mais de 30 tipos de câncer, com foco em neoplasias sólidas, como câncer de pulmão, mama, cólon, melanoma e outros, além de mutações em genes supressores de tumor, oncogenes e genes de reparo de DNA. Sendo, portanto, uma boa escolha como fonte representativa de mutações somáticas relacionadas ao câncer. Após o processamento das variantes do COSMIC, que inicialmente totalizavam 68.629.906 SNVs, obtivemos 10.587.546 coordenadas genômicas para WG, correspondendo a uma redução de aproximadamente 84,57%, e para 3.662.887 para CDS, representando uma redução de 90,17% (**Tabela 2**).

O ClinVar é um repositório gratuito e curado de informações sobre a significância clínica de variantes. O objetivo principal dele é catalogar as variantes e classificá-las de acordo com sua relevância clínica, ou seja, se e como essas variantes estão relacionadas a doenças, condições genéticas ou respostas a tratamentos. Os dados do ClinVar foram processados para extrair apenas variantes germinativas com *status*[8] de revisão confiável, categorizando-as como patogênicas ou benignas. As SNVs patogênicas foram classificadas com base em evidências

[8] https://www.ncbi.nlm.nih.gov/clinvar/docs/review_guidelines/

clínicas e funcionais associadas a uma condição clínica específica. Essas variantes estão tipicamente relacionadas a doenças hereditárias de alta penetrância, desempenhando um papel causal ou fortemente contributório na manifestação da condição clínica em questão. Em contraste, as variantes benignas foram classificadas como aquelas que não são esperadas para causar doenças ou ter impacto negativo significativo na saúde, conforme a classificação do banco de dados. Apesar de serem consideradas neutras do ponto de vista funcional, o termo "neutro" foi utilizado neste trabalho para o conjunto de variantes polimórficas obtidas a partir do HGDP, a fim de diferenciá-las do conjunto de variantes definidas como benignas e provenientes do ClinVar. Após o processamento das 1.303.558 variantes iniciais do ClinVar, observamos uma redução de 94,09% para SNVs benignas no WG, resultando em 76.971 variantes, e de 98,74% para SNVs patogênicas, resultando em 16.313 variantes. Para CDS, a redução foi de 95,64% para SNVs benignas, resultando em 56.794, e de 98,99% para SNVs patogênicas, totalizando 13.098 variantes (**Tabela 2**).

4.2. Caracterização dos conjuntos de dados

4.2.1. Padrão de localização das SNVs de diferentes classes mostram viéses de enriquecimento para algumas regiões genômicas

O perfil de anotação funcional dos conjuntos de variantes, obtidos a partir do software VEP, fornece informações sobre o perfil de distribuição e o contexto genômico das variantes. Nas variantes do conjunto WG, SNVs em regiões intrônicas compreendem a maior fração dos dados, aproximadamente 60%, em três conjuntos: COSMIC, SNPs e Rares (**Figura 16-A, Tabela 3**). Em contraste, observou-se um enriquecimento de SNVs em regiões exônicas nos conjuntos Benign e Pathogenic do ClinVar. O conjunto Pathogenic compreendeu cerca de 81% de SNVs em regiões exônicas, enquanto o conjunto Benign atingiu aproximadamente 78% (**Figura 16-A, Tabela 3**). Os resultados encontrados refletem a natureza dos dados do ClinVar, que estão predominantemente associados a condições clinicamente relevantes de alta penetrância, muitas das quais são causadas por mutações em genes codificadores de proteínas (Landrum *et al.*, 2014).

Comparativamente, o conjunto COSMIC apresentou 36% de SNVs em regiões exônicas, enquanto os conjuntos SNPs e Rares do HGDP apresentaram aproximadamente 1%. Essa distribuição desigual entre os conjuntos possivelmente reflete tanto as características

intrínsecas de cada banco de dados quanto às diferenças nas regiões genômicas em que essas variantes estão mais frequentemente localizadas. Embora não tenhamos investigado mais a fundo as razões para estas diferenças, acreditamos que o conjunto COSMIC possui uma fração substancial de suas entradas em regiões exônicas devido também ao método de sequenciamento direcionado de exomas, um método muito utilizado na genômica do câncer.

A análise estatística, utilizando o teste qui-quadrado de Pearson para independência revelou uma associação significativa entre as regiões genômicas e os conjuntos de dados analisados no WG. A estatística de teste foi de 21.977 com 20 graus de liberdade ($p = 2{,}2e\text{-}16$), indicando uma heterogeneidade na distribuição das SNVs entre as diferentes classes. Isso sugere, conforme observamos qualitativamente, que alguns conjuntos de dados são enriquecidos em regiões genômicas específicas, o que pode influenciar as comparações subsequentes.

Para mitigar potenciais vieses de coocorrência causados pela super-representação de SNVs exônicas nos conjuntos Pathogenic e Benign, realizamos uma análise utilizando somente o subconjunto de SNVs localizadas em regiões CDS. Utilizamos o projeto MANE (*Matched Annotation from NCBI and EMBL-EBI*), que define um conjunto representativo de transcritos canônicos para genes codificadores de proteínas, para definir qual seria a consequência funcional de uma variante que afeta múltiplos transcritos de um mesmo gene. Essa abordagem garantiu uma redução substancial nas variantes intrônicas e intergênicas, resultando em uma distribuição mais uniforme entre os conjuntos, permitindo comparações mais robustas e equitativas. Após a filtragem, SNVs em regiões CDS passaram a compreender aproximadamente 99% de cada conjunto analisado (**Figura 16-B, Tabela 3**). A partir desse ponto, todos os resultados subsequentes desta seção (**4.2.**) referem-se ao conjunto de CDS, salvo menção contrária.

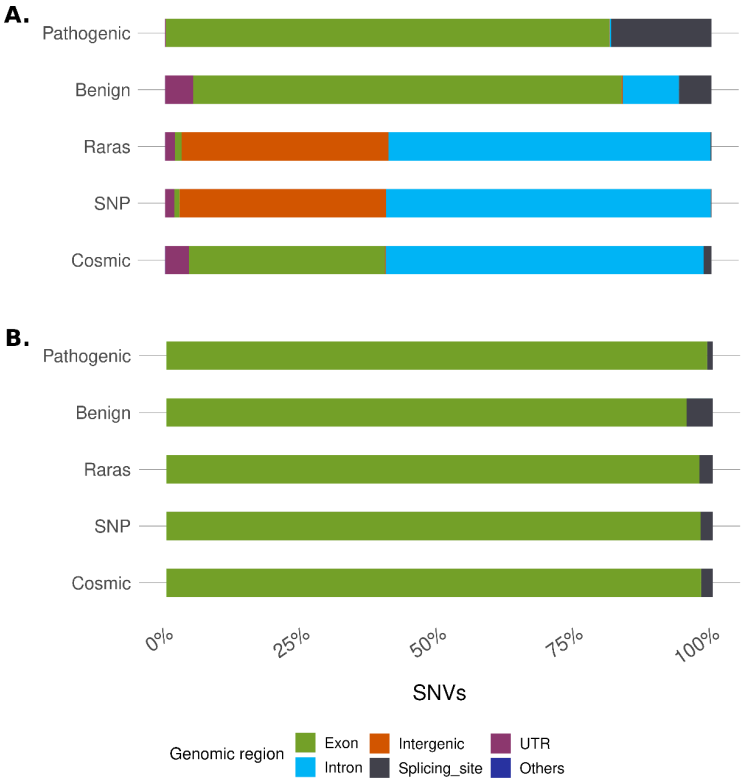


Figure 16. Localização genômica de SNVs nos cinco conjuntos de dados: SNVs Pathogenic, Benign, RarEs, SNPs e COSMIC. As barras representam cada conjunto de dados e proporção de SNVs por distribuição de região genômica para o (A) WG e (B) para regiões CDS após filtragem mostrando o padrão de distribuição de SNVs, indicando uma filtragem bem-sucedida de SNVs para CDS. O teste qui-quadrado de Pearson para independência mostrou uma associação significativa entre a região genômica e os conjuntos de dados WG (p = 2,2e-16).

Tabela 3. Tabelas de contingência da quantidade de SNVs em cada região genômica. A tabela apresenta as contagens absolutas de SNVs em diferentes regiões genômicas, permitindo uma comparação entre WG e CDS.

	Genomic region	Benign	Pathogenic	Cosmic	Rares	SNPs
WG	**Exon**	56691	13078	3737233	529859	17691
	Intergenic	173	8	27857	18433205	759658
	Intron	7469	48	5881335	27434542	1191835
	Others	2	0	977	814	28
	Splicing_site	4292	2913	140117	61434	2516
	UTR	3759	28	515291	834841	34816
CDS	**Exon**	56010	13032	3638854	492084	15859
	Intergenic	0	0	0	0	0
	Intron	0	0	0	0	0
	Others	0	0	0	0	0
	Splicing site	784	66	24033	4707	181
	UTR	0	0	0	0	0

4.2.2. Razão Transição/Transversão (Ti/Tv) com diferentes consequências funcionais e evolutivas

A razão transição/transversão (Ti/Tv) é uma métrica clássica utilizada em estudos genômicos como indicador de conservação evolutiva e ferramenta de controle de qualidade (J. Wang *et al.*, 2015). Em condições sem pressão seletiva, a razão Ti/Tv esperada seria 0,5, assumindo igual probabilidade de transições e transversões. No entanto, devido a fenômenos bioquímicos, como a desaminação espontânea de citosina metilada para timina (uma mutação de transição), a razão aumenta para cerca de 2,0 em regiões genômicas gerais e atinge valores próximos a 3,0 em éxons, refletindo a maior conservação e as pressões seletivas que atuam sobre essas regiões funcionais, uma vez que transições usualmente causam mutações sinônimas que não alteram o aminoácido codificado, especialmente para mutações na terceira base do códon, enquanto transversões tem uma chance maior de causar mutações não-sinônimas (Liu *et al.*, 2012; Guo *et al.*, 2017).

Altos valores de Ti/Tv, como os observados em regiões exônicas de variantes benignas/neutras, indicam a atuação de seleção purificadora, um processo que reduz a frequência de mutações potencialmente deletérias ao longo das gerações. Em contraste, razões mais baixas, geralmente associadas a variantes patogênicas, sugerem maior prevalência de transversões disruptivas que podem impactar a funcionalidade proteica.

Nos cinco conjuntos de SNVs analisados em CDS, a razão Ti/Tv variou amplamente, em um padrão que reflete as diferentes características funcionais e evolutivas de cada grupo. Os conjuntos Benign e SNPs, associados a variantes neutras ou benignas, apresentaram os maiores valores de Ti/Tv (4,38 e 3,35, respectivamente), indicando uma predominância de transições que tendem a preservar a integridade genômica. Em contrapartida, os conjuntos Rares, Pathogenic e COSMIC, onde espera-se uma maior prevalência de variantes patogênicas, exibiram valores progressivamente mais baixos (2,87, 2,19 e 1,71, respectivamente), sugerindo uma maior ocorrência de transversões potencialmente deletérias (**Tabela 4**). Esses resultados ressaltam como a razão Ti/Tv reflete não apenas a natureza funcional das variantes em cada conjunto, mas também as forças seletivas e evolutivas que moldam os diferentes perfis de SNVs.

Tabela 4: Contagens absolutas e porcentagens de transições (Ti) e transversões (Tv) para cada conjunto de dados analisado, bem como a razão Ti/Tv. Transições são substituições de purina por purina (A<> G) ou pirimidinas por pirimidina (C<>T), enquanto transversões são trocas de purina por bases de pirimidina (A<>C, A<>T, G<>C, G<>T).

Dataset	Mutation type		Ti_Tv_Ratio
	Transitions	Transversion	
Pathogenic	9000 (68.72%)	4097 (31.28%)	2.19
Benign	46238 (81.41%)	10555 (18.59%)	4.38
Rares	368677 (74.21%)	128114 (25.79%)	2.87
SNPs	12360 (77.06%)	3680 (22.94%)	3.35
COSMIC	2312710 (63.14%)	1350176 (36.86%)	1.71

4.2.3. Caracterização funcional das SNVs

A anotação funcional dos cinco conjuntos de SNVs revelou diferenças marcantes em suas consequências funcionais, destacando suas distintas propriedades biológicas e potenciais impactos fenotípicos. No conjunto Pathogenic, predominam variantes de alto impacto, com aproximadamente 51,1% resultando em *stop-gains*, que interrompem prematuramente a tradução proteica, e 47,1% identificadas como *missense*, capazes de alterar a sequência de aminoácidos (**Figura 17 -** *Pathogenic***, Tabela 5**). Essas consequências estão associadas a doenças monogênicas e fatores de risco para condições complexas, conforme documentado na literatura (Krumm *et al.*, 2015; Pal *et al*., 2015; von Stedingk *et al.*, 2021; Fu *et al.*, 2022; Koprulu *et al.*, 2022; Leggatt *et al.*, 2023; Venizelos *et al.*, 2023).

Em contraste, variantes Benign exibiram predominantemente efeitos sinônimos (*synonymous* - 71,1%), indicando uma baixa chance de impacto funcional significativo, enquanto 27,3% foram *missense*, com impacto funcional potencialmente baixo a moderado, refletindo a natureza não deletéria das variantes benignas no ClinVar (**Figura 17 -** *Benign*).

O conjunto Rares apresentou uma predominância de variantes *missense* (59,5%), o que pode refletir a influência da seleção purificadora sobre variantes de baixa frequência com potencial deletério (Wright, 2005; Quintana-Murci, 2016, Kumar *et al.*, 2023). Além disso, 38,2% das variantes foram *synonymous*, possivelmente representando variantes que surgiram *de novo* com menor impacto funcional (**Figura 17 -** *Rares*).

O conjunto SNPs mostrou uma distribuição equilibrada entre variantes *missense* (49,7%) e *synonymous* (48,6%), com uma menor incidência de variantes de alto impacto funcional, consistindo com sua classificação como variantes neutras ou quase neutras (**Figura 17 -** *SNPs*).

O conjunto COSMIC apresentou uma alta prevalência de variantes *missense* (69,4%), com aproximadamente 69% exibindo impacto funcional moderado, alinhando-se com a natureza multifatorial do câncer, que frequentemente envolve interações sinérgicas entre múltiplas mutações (Matlak *et al.,* 2019). A menor prevalência de variantes *synonymous* (24,9%) reflete a maior propensão de mutações somáticas a impactar proteínas e promover características tumorais. O entendimento de que variantes que influenciam doenças complexas têm efeitos menores e muitas vezes com ação aditivas e sinergística quando comparados com o tamanho do efeito observado nos genes que causam doenças monogênicas/oligogênicas destaca a necessidade de análises multifacetadas para compreender a etiologia dessas condições (Claussnitzer *et al.*, 2020).

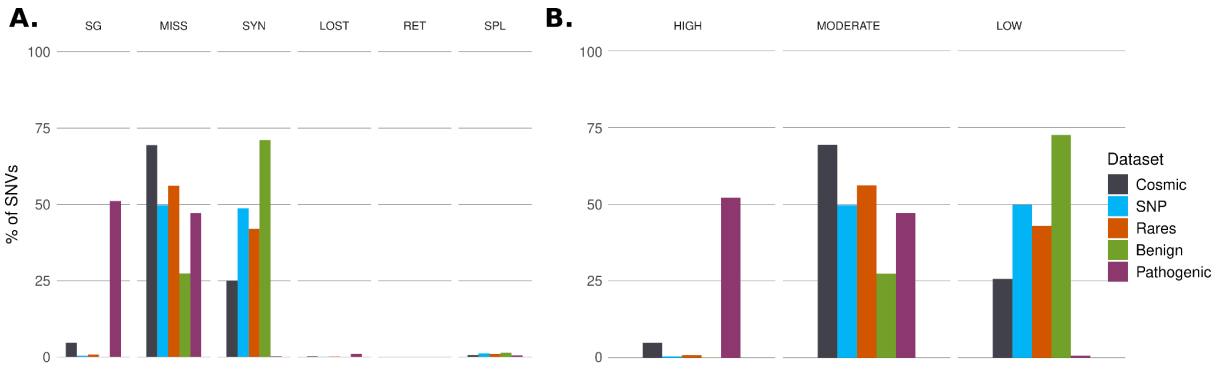


Figura 17: Consequências de codificação mais graves de SNVs e tamanho do impacto funcional em cada conjunto de dados analisado. (A) Proporção de SNVs para as seis consequências de codificação mais graves. SG (*Stop-gained)*: Esta mutação introduz um códon de parada prematuro, levando a uma proteína truncada e frequentemente não funcional. MISS (*Missense variant*): Uma única alteração de nucleotídeo resulta em um aminoácido diferente, afetando potencialmente a função da proteína. SYN (*Synonymous variant*): Uma mutação que não altera a sequência de aminoácidos e é frequentemente considerada neutra. LOST (*Start lost* e *Stop lost*): Mutações que levam à perda do códon de início ou parada, impactando o comprimento da proteína e potencialmente a função. RET (*Stop retained variant*): Esta mutação retém o códon de parada, levando a uma proteína alongada com função alterada. SPL (*Splice region variant*): Uma mutação que afeta os locais de splicing, levando a um splicing anormal. A figura ilustra a distribuição de SNVs entre essas consequências de codificação. **(B)** Tamanho do impacto funcional pela codificação de consequências de regiões genômicas para SNVs em CDS para os cinco conjuntos de dados. HIGH - A variante é prevista para ter um impacto significativo na proteína, potencialmente levando ao truncamento da proteína, perda de função ou ativação de decaimento mediado por *nonsense*. MODERATE - Uma variante que não é esperada para interromper severamente a função da proteína, mas pode alterar sua eficácia. LOW - Provavelmente benigna ou com impacto mínimo na função da proteína

Tabela 5: Contagem absoluta de SNVs de regiões CDS para as consequências de codificação por conjunto de dados. A tabela fornece a contagem detalhada das consequências de codificação observadas em cada conjunto de dados.

Consequence type	Pathogenic	Benign	SNPs	Rares	Cosmic
missense variant	6172	15553	7976	295730	2543087
splice region variant	66	784	181	4707	24033

start and stop lost	136	11	13	767	6887
stop gained	6695	33	54	5341	172708
stop retained variant	0	16	7	155	1203
synonymous variant	29	40397	7809	190091	914969

4.2.4. A análise de variantes *missense* utilizando o AlphaMissense revela padrões de impacto funcional

As variantes *missense* podem ter uma ampla gama de impactos funcionais, variando desde substituições que não afetam significativamente o enovelamento, atividade ou estabilidade de proteínas, até substituições que são experimentalmente demonstradas como fatores causais de doenças monogênicas. Visando explorar os potenciais impactos da significativa quantidade de variantes *missense* observadas em todos os conjuntos de dados avaliados, essas variantes foram estratificadas e classificadas com maior resolução utilizando o AlphaMissense. Esse modelo estatístico utiliza um algoritmo de aprendizado profundo treinado com estruturas proteicas preditas pelo modelo AlphaFold2 para calcular uma pontuação (*score*) de patogenicidade de SNVs *missense* (Cheng *et al.*, 2023). Os resultados do AlphaMissense variam entre zero e um, sendo que valores mais altos indicam uma maior probabilidade de a mutação ser patogênica.

Recuperamos a pontuação de patogenicidade das saídas do VEP e avaliamos suas distribuições nos cinco conjuntos de dados de SNVs (**Figura 18-A**). Essa análise quantitativa revelou um padrão semelhante à distribuição de dados de impacto funcional em categorias qualitativas. O conjunto de dados Pathogenic exibiu uma distribuição assimétrica e enviesada à esquerda (**Figura 18-A -** *Patogênico*), com uma frequência maior de variantes patogênicas com pontuações mais altas, consistentes com condições de alta penetrância, como doenças monogênicas e oligogênicas.

Em contraste, variantes do conjunto Benign mostraram uma distribuição assimétrica oposta, enviesada para a direita, com uma concentração de SNVs de baixas pontuações de patogenicidade (**Figura 18-A -** *Benigno*). Essa distribuição reflete o impacto funcional reduzido das variantes benignas e está em concordância com as curadorias do ClinVar, que classificam essas variantes como funcionalmente neutras. Este resultado já havia sido relatado na validação do AlphaMissense (Cheng *et al.*, 2023).

Para variantes COSMIC, observamos uma distribuição bimodal, com um dos picos de

variantes apresentando pontuações baixas e um pico com pontuações altas (**Figura 18-A -** *COSMIC*). Embora não tenhamos avaliado mais a fundo esse resultado, o mesmo é compatível com a existência de mutações *passenger*, as quais teriam um menor impacto funcional, e mutações *driver* patogênicas que desempenham papéis críticos na oncogênese (Ostroverkhova *et al.,* 2023).

O conjunto de SNPs exibe uma distribuição enviesada à direita (**Figura 18-A -** *SNPs*), com uma concentração de variantes mostrando baixas pontuações de patogenicidade, um perfil semelhante ao observado para variantes benignas no ClinVar. Este resultado sugere que a maioria dessas variantes seja neutra/quase neutra, estando consoante com a utilização dos filtros restritivos aplicados.

Variantes raras também exibiram uma distribuição enviesada para a direita (**Figura 18-A - Raras**), embora com um pequeno pico de SNVs com pontuações altas. O conjunto Rares representa uma gama de pontuações consistentes com SNVs de baixa frequência na população humana, pois pode incluir variantes deletérias com frequências reduzidas devido à seleção purificadora e variantes *de novo* com consequências fenotípicas neutras ou quase neutras.

A análise qualitativa das variantes *missense* utilizando o AlphaMissense ofereceu uma visão complementar sobre a diversidade entre os diferentes conjuntos de dados. O conjunto Pathogenic apresentou uma predominância de variantes *missense* classificadas como *likely pathogenic* (78,5%), enquanto o conjunto Benign exibiu 89,5% de variantes *likely benign*, consistente com a curadoria do ClinVar (**Figura 18-B** e **Tabela 6*)***. Os conjuntos Benign e SNPs apresentaram padrões semelhantes, com aproximadamente 90% das SNVs categorizadas como *likely benign*, refletindo sua natureza predominantemente neutra ou quase neutra de uma perspectiva evolutiva. Por outro lado, variantes do conjunto Rares exibiram uma distribuição com 80% classificadas como *likely benign*, mas também revelaram um pico moderado (11,5%) de variantes com altas pontuações relatadas como *likely pathogenic*. O conjunto COSMIC, por sua vez, mostrou uma distribuição mais heterogênea, com cerca de 60% das variantes sendo classificadas como *likely benign* e 29% como *likely pathogenic*. Essa distribuição complexa reflete a natureza multifatorial do câncer, no qual mutações *driver* com alto impacto coexistem com mutações *passenger* neutras ou de impacto baixo.

Esses resultados ressaltam a importância de se considerar não apenas a presença de variantes *missense*, mas também seus potenciais impactos funcionais em estudos genômicos relacionados à saúde humana. Além disso, a aplicação do AlphaMissense provou ser uma

ferramenta valiosa para prever a patogenicidade de variantes *missense*, que podem apresentar uma ampla gama de possíveis impactos, destacando sua utilidade para o estudo de variantes sem caracterização experimental e para o diagnóstico de doenças genéticas idiopáticas.

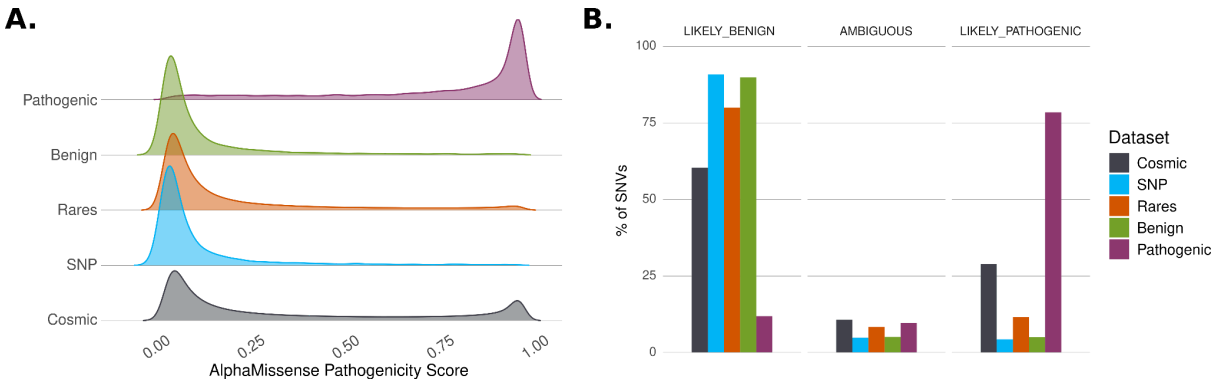


Figura 18. Avaliação de patogenicidade das SNVs missense de regiões CDS por meio do AlphaMissense nos cinco conjuntos de dados - Benigno, COSMIC, Patogênico, Raro e SNPs. (A) Distribuição dos *scores* de patogenicidade para variantes missense. O eixo y do gráfico *ridgeline* representa a densidade de probabilidade das pontuações de patogenicidade para cada conjunto de dados. A altura das curvas de densidade em qualquer ponto dado indica a probabilidade relativa de encontrar uma pontuação de patogenicidade dentro do intervalo correspondente de valores. **(B)** Classificação qualitativa da patogenicidade das SNVs *missense*. Com base na pontuação contínua fornecida pelo AlphaMissense, variando de 0 a 1, que prevê a patogenicidade para cada variante com *missense*. As variantes são classificadas em uma das três categorias: *likely pathogenic* (provavelmente patogênica), likely benign (provavelmente benigna) e ambiguous (ambígua). A classificação é baseada nos seguintes limites: *likely benign* apresentam *am pathogenicity* < 0,34; *likely pathogenic* possuem am pathogenicity > 0,564 ou pontuações intermediárias de patogenicidade (0,34 ≤ *am pathogenicity* ≤ 0,564), são consideradas incertas pelo modelo (*ambiguous*). Esses limites foram definidos para atingir 90% de precisão para variantes patogênicas e benignas.

Tabela 6: Contagem absoluta de SNVs de regiões CDS para as consequências de codificação por conjunto de dados.

Dataset	Categoria	Contagem de SNVs	Porcentagem (%)
Rares	*ambiguous*	21916	8.352
	likely benign	210120	80.075
	likely pathogenic	30365	11.571
SNP	*ambiguous*	341	4.801
	likely benign	6456	90.903
	likely pathogenic	305	4.294
COSMIC	*ambiguous*	244551	10.693
	likely benign	1381746	60.420
	likely pathogenic	660568	28.885
Benign	*ambiguous*	710	5.089
	likely benign	12548	89.949

	likely pathogenic	692	4.960
Pathogenic	*ambiguous*	565	9.605
	likely benign	697	11.849
	likely pathogenic	4620	78.544

4.3. Estatística global dos conjuntos de coordenadas de SNVs

4.3.1. Modelos do cenário mutacional no genoma humano

O genoma humano apresenta um cenário dinâmico, onde mutações ocorrem e exercem diferentes graus de influência sobre funções biológicas e fenótipos de doenças. Compreender a distribuição espacial dessas mutações — sejam elas patogênicas ou benignas — é essencial para auxiliar na compreensão dos mecanismos biológicos subjacentes aos distúrbios genéticos e avançar na medicina de precisão. Neste estudo, foram propostos três modelos para explorar os padrões de distribuição espacial das mutações no genoma.

O **Modelo 1** sugere que mutações de diferentes classes ocorram em *loci* distintos (**Figura 19**). Este modelo nos permite explorar a hipótese de que a separação espacial no genoma pode ser correlacionada com distintos resultados funcionais ou caminhos evolutivos. Já o **Modelo 2** propõe que, embora as mutações ocorram nas mesmas posições genômicas, os alelos diferem, sugerindo uma interação diferenciada entre o tipo de mutação e os efeitos específicos do alelo. Por fim, o **Modelo 3** sugere que mutações benignas e patogênicas não apenas compartilham *loci*, mas também envolvem alelos idênticos, levantando questões sobre os contextos diferenciais ou mecanismos regulatórios que modulam o impacto dessas mutações no organismo.

Esses modelos foram desenvolvidos para abordar lacunas no entendimento do contexto genômico das mutações e suas implicações coletivas. A comparação sistemática desses cenários em diferentes conjuntos de dados genômicos busca elucidar aspectos como estabilidade genômica, existência de *hotspots* mutacionais e o papel dessas dinâmicas na saúde humana e nas doenças. Nas seções subsequentes, detalhamos os experimentos escolhidos para testar esses modelos e discutimos as implicações dos resultados no contexto mais amplo da pesquisa genômica mutacional.

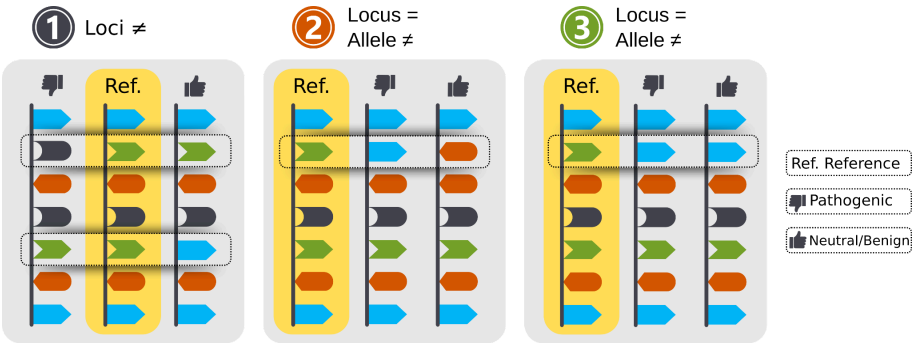


Figure 19. Modelos de ocorrência espacial de mutações no genoma humano. Três modelos são propostos para a distribuição de mutações ao comparar variantes patogênicas e neutras em relação ao alelo de referência. Modelo 1: Variantes patogênicas e neutras ocorrem em *loci* distintos. Modelo 2: Variantes patogênicas e neutras ocorrem no mesmo *locus*, mas apresentam alelos diferentes. Modelo 3: Variantes patogênicas e neutras ocorrem no mesmo *locus* e compartilham o mesmo alelo.

4.3.2. Avaliação estatística da coocorrência de SNVs ao longo do genoma humano

Após a construção e caracterização dos cinco conjuntos de SNVs, realizamos uma análise de coocorrência genômica entre os SNVs dos cinco conjuntos de variantes utilizados em nosso estudo para avaliar os padrões de sobreposição entre as diferentes classes de variantes. Para tal, empregamos o teste exato de Fisher conforme implementado no BEDTools, uma ferramenta amplamente utilizada para avaliar associações entre intervalos genômicos e determinar se a sobreposição observada é significativamente maior ou menor do que o esperado ao acaso (Quinlan *et al.*, 2010). Esse método é particularmente útil em pesquisas de doenças complexas, como o câncer, onde identificar associações entre mutações em diferentes conjuntos de dados pode revelar padrões biológicos importantes (Chen *et al.*, 2016; Hall *et al.*, 2018; Luo *et al.*, 2020; Yadav *et al.*, 2022; Baca *et al.*, 2022; Arthur *et al.*, 2024).

Os resultados das análises foram expressos em valores de p (*p-values*) e razões de chances (*odds ratios*, OR), permitindo avaliar tanto a significância estatística da sobreposição entre os conjuntos de coordenadas genômicas quanto a magnitude e direção da associação. Consideramos significativo um *p-value* menor que 0,05, indicando que a sobreposição entre os conjuntos é estatisticamente diferente do esperado ao acaso. O OR, por sua vez, quantifica a intensidade e direção da associação.

- OR > 1: A sobreposição ocorre com maior frequência do que o esperado, sugerindo uma coocorrência não aleatória.
- OR < 1: A sobreposição ocorre com menor frequência do que o esperado, indicando exclusividade entre os conjuntos.

Entre as dez comparações possíveis, oito mostraram interseções significativas entre os conjuntos de coordenadas de SNVs. As únicas exceções foram observadas nos conjuntos *SNPs vs. Rares*, (que são, por definição, mutuamente excludentes), e *SNPs vs. Pathogenic*, que também apresentou ausência de sobreposição (**Tabela 7**). A ausência de interseção entre SNPs e Pathogenic é consistente com as características intrínsecas desses conjuntos, embora não tenhamos investigado mais profundamente esse resultado: enquanto SNPs representam variantes polimórficas de alta frequência com baixo impacto funcional, variantes do conjunto Pathogenic são definidas por seus grandes efeitos deletérios e baixa frequência alélica.

Tabela 7. Resultados completos do teste exato de Fisher. Tabelas de contingência para o número de sobreposições e intervalos únicos entre dois arquivos em DNA autossômico para CDS e WG. O teste exato de Fisher, juntamente com o fornecimento de valores *p* e odds ratios (OR) para interseções entre quaisquer dois conjuntos de coordenadas genéticas ou genômicas (A e B), produz quatro números adicionais: o número de interseções (in_A/in_A), as contagens de intervalos exclusivos para cada arquivo (in_A/not_B ou not_A/in_B) e, com base em uma estimativa do número total de intervalos possíveis calculados usando o tamanho de cada gene, o número estimado de coordenadas não sobrepostas (not_A/not_B). OR indica a força da associação. Os resultados são mostrados para cinco conjuntos de dados: SNPs, Rares, COSMIC, Benign e Pathogenic em regiões CDS e WG. O teste entre SNPs e Rares não foi realizado porque os conjuntos de dados são mutuamente excludentes. Os testes de regiões CDS são mais robustos, particularmente para conjuntos de dados derivados do ClinVar, devido ao padrão de distribuição distinto de SNVs observado para todo o genoma.

CDS							
Dataset A	**Dataset B**	***p***	**OR**	**in_A/in_B**	**in_A/not_B**	**not_A/in_B**	**not_A/not_B**
SNPs	Cosmic	0	5.652	9575	6465	3653035	13940452
SNPs	Benign	0	66.480	2728	13312	54066	17539421
SNPs	Pathogenic	0.013	0.335	4	16036	13094	17580393
Rares	Cosmic	0	1.315	126723	370068	3536158	13576578
Rares	Benign	0	16.240	17777	479014	39017	17073719
Rares	Pathogenic	0.026	0.885	328	496463	12770	17099966
Cosmic	Benign	0	0.699	20372	3642515	36422	4550829
Cosmic	Pathogenic	1.1e-225	0.557	4036	3658851	9062	4578189
Benign	Pathogenic	5.6e-24	2.150	192	56602	12906	8180438
WG							
SNPs	Cosmic	0	2.622	34725	887571	10552821	707275263
SNPs	Benign	0	24.755	2367	919929	74604	717753480
SNPs	Pathogenic	1.6e-16	0.048	1	922295	16312	717811772
Rares	Cosmic	0	0.558	428910	49810900	10158636	658351934
Rares	Benign	0	5.604	22804	50217006	54167	668456403
Rares	Pathogenic	3.5e-169	0.300	360	50239450	15953	668494617

Cosmic	Benign	0	17.910	25666	19528706	51306	699144702
Cosmic	Pathogenic	0	13.900	4566	19549806	11748	699184260
Benign	Pathogenic	0	112.648	194	76778	16120	718657288

Dentro dos oito testes pareados com sobreposições, cinco apresentaram OR na mesma direção para os conjuntos de WG e CDS (**Tabela 8**). As exceções são as comparações COSMIC vs. Pathogenic e COSMIC vs. Benign, que têm associação positiva ao comparar coordenadas WG e negativas para os conjuntos de dados CDS, além de Rares vs. COSMIC que revelaram o padrão oposto.

Tabela 8. Resultados resumidos do teste exato de Fisher e representação em cores para associações positivas ou negativas. O teste exato de Fisher forneceu valores de *p* ($p < 0{,}001$ - veja valores completos na **Tabela 7**) e de OR para interseções entre dois conjuntos (Dataset 1 e 2) de coordenadas genômicas. Para evidenciar as diferenças de tendência de associação entre cada teste, valores que apresentaram OR > 1 e associação positiva foram destacados em verde, enquanto os valores de OR < 1 e associação negativa foram destacados em vermelho.

Dataset 1	Dataset 2	CDS	WG
SNPs	Cosmic	5.652	1.920
SNPs	Benign	66.480	22.493
SNPs	Pathogenic	0.335	0.082
Rares	Cosmic	1.315	0.558
Rares	Benign	16.240	5.604
Rares	Pathogenic	0.885	0.300
Cosmic	Benign	0.699	17.910
Cosmic	Pathogenic	0.557	13.900
Benign	Pathogenic	2.150	112.648

Conforme descrevemos anteriormente, os conjuntos COSMIC, Benign e Pathogenic compartilham um forte viés para variantes em regiões CDS (**Figura 15**). Desta forma, os testes do WG para esses três conjuntos provavelmente apresentariam um valor significativo de OR artificialmente elevado, dado o tamanho do genoma completo e a maior ocorrência de variantes destes conjuntos em regiões CDS. Em outras palavras, como o tamanho do genoma humano é muito maior do que as suas regiões codificadoras, a abundância de variantes em regiões de CDS nesses três conjuntos aumenta a probabilidade de que as variantes desses conjuntos coincidam, mesmo quando a análise é realizada no contexto do WG. Assim, o viés para CDS nesses conjuntos pode influenciar os resultados do WG, levando a valores de OR

significativos mesmo que a sobreposição real no genoma completo não seja tão expressiva. Por essa razão, consideramos os dados CDS mais robustos para avaliar as estatísticas de variantes que coocorram entre os conjuntos de dados, particularmente para testes envolvendo estes conjuntos.

Por outro lado, embora os conjuntos COSMIC e Rares compartilhem um viés intrônico, os testes estatísticos realizados no genoma completo (WG) não detectaram uma sobreposição significativa entre esses conjuntos, como observado para o CDS. Esse resultado pode refletir a heterogeneidade das variantes intrônicas, uma vez que regiões intrônicas ocupam uma proporção maior do genoma e possuem maior dispersão de mutações, reduzindo a probabilidade de coocorrência. Além disso, diferenças nos padrões específicos de mutação para CDS e WG, bem como algum viés intrínseco dos conjuntos de dados, podem ter contribuído para essa ausência de sobreposição significativa.

Avaliando os resultados dos testes alguns padrões gerais foram observados para WG e CDS:

- Associações Negativas (OR < 1) foram observadas em quatro comparações (*SNPs vs. Pathogenic*, *Rares vs. Pathogenic*, *COSMIC vs. Benign*, *COSMIC vs. Pathogenic*). Essas associações indicam que essas classes de SNVs são predominantemente exclusivas em seus *loci* genômicos, sugerindo que compartilham pressões seletivas distintas ou diferentes contextos funcionais. Esses resultados se alinham com o **Modelo 1** proposto (**Figura 19**), onde mutações neutras e patogênicas ocorrem em *loci* distintos.

- Associações Positivas (OR > 1) foram observadas em cinco comparações (*SNPs vs. COSMIC*, *SNPs vs. Benign*, *Rares vs. COSMIC*, *Rares vs. Benign*, *Benign vs. Pathogenic*). Esses resultados sugerem que essas classes compartilham *loci* genômicos mais frequentemente do que seria esperado ao acaso. O OR variou de 1,3 a 66, dependendo dos conjuntos comparados, com o maior valor observado entre SNPs e Benign (**Tabela 8**). Essas associações são consistentes com os **Modelos 2** ou **3**, onde variantes distintas compartilham a mesma posição genômica.

4.3.3. Análise comparativa de alelos das SNVs significativamente sobrepostas

Para os subconjuntos de SNVs que coocorrem nos mesmos *loci*, obtidos a partir dos testes que apresentaram associação positiva de sobreposição genômica (OR > 1), avaliamos também se

essas variantes compartilham alelos idênticos com maior frequência do que o esperado ao acaso (Métodos - seção 3.8). Essa análise visa determinar se os padrões observados são mais bem explicados pelo **Modelo 2** ou **Modelo 3** (**Figura 19**).

A análise foi realizada utilizando o teste exato de Fisher, aplicado para verificar a associação entre alelos idênticos e diferentes em uma tabela de contingência 2×2. Sob a hipótese nula, a frequência de alelos iguais e diferentes segue um padrão aleatório, indicando independência entre as categorias. Valores de OR e *p*-value foram utilizados para interpretar os resultados. Um *p*-value pequeno ($p<0.05$) indica que os alelos iguais ocorrem significativamente mais frequentes do que o esperado, enquanto o OR revela a força da associação:

- Um OR > 1 Frequência de alelos iguais maior do que o esperado.
- Um OR = 1 Frequência de alelos iguais compatível com o acaso.
- Um OR < 1 Frequência de alelos iguais menor do que o esperado.

Os resultados revelaram que, na maioria das comparações entre conjuntos com sobreposição significativa de *loci*, os alelos idênticos ocorrem com maior frequência do que o esperado, com exceção da comparação entre os conjuntos Benign e Pathogenic, onde foi observada 100% de dissimilaridade (**Tabela 9**).

Tabela 9. Resultados do teste exato de Fisher aplicado à análise da frequência de alelos idênticos entre variantes sobrepostas para CDS. Para cada comparação, são apresentados os números de alelos iguais e diferentes observados (*Observed Equal* e *Different*), as frequências esperadas de alelos iguais e diferentes (*Expected Equal* e *Different*), o odds Ratio (OR) calculado e o *p*-value (*p*) correspondente. Os valores de odds Ratio indicam a força da associação entre alelos idênticos, enquanto os *p*-values indicam a significância estatística dessa associação.

Dataset A	Dataset B	Observed equal	Observed different	Expected equal	Expected different	OR	*p*
SNPs	COSMIC	9358	217	2393.75	7181.25	129.34	0
SNPs	Benign	2728	0	682.00	2046.00	Inf	0
Rares	COSMIC	97619	29104	31680.75	95042.25	10.06	0
Rares	Benign	17454	323	4444.25	13332.75	162.11	0
Benign	Pathogenic	0	192	48.00	144.00	0.00	0

A associação mais forte foi encontrada entre os conjuntos SNPs e Benign (OR = ∞, p = 0), onde 100% das variantes sobrepostas compartilham alelos idênticos. De maneira surpreendente, a comparação entre SNPs e COSMIC demonstrou uma associação significativa

entre os alelos observados (OR = 129.34, p = 0), com 97.73% das variantes sobrepostas compartilhando alelos idênticos. A análise entre os conjuntos Rares e Benign também apresentou uma forte associação (OR = 162.11, p = 0), indicando que 98.18% das variantes sobrepostas possuem alelos idênticos.

Por outro lado, a ausência de alelos idênticos na interseção entre Benign e Pathogenic (OR = 0) demonstra que essas categorias possuem alelos completamente distintos nas posições avaliadas, podendo refletir diferentes mecanismos funcionais e contextos evolutivos.

Esses resultados destacam que a identidade de alelos entre variantes sobrepostas não ocorre de maneira aleatória, mas segue padrões específicos, sugerindo que o **Modelo 3** (alelos idênticos) explica melhor a distribuição nas comparações *SNPs vs. Benign, SNPs vs. COSMIC, Rares vs. COSMIC* e *Rares vs. Benign*, enquanto o **Modelo 2** (alelos diferentes) é mais compatível com o padrão observado para *Benign vs. Pathogenic*.

4.3.4. Cenário de coocorrência mutacional de SNVs de diferentes origens evolutivas e consequências funcionais no genoma humano

Os resultados obtidos nos testes Fisher descritos anteriormente, tanto para avaliar as posições das SNV quanto para a análise comparativa de alelos, serão explorados individualmente nesta sessão.

SNPs vs. Benign

A análise revelou 2.728 variantes coocorrendo nas mesmas coordenadas CDS, uma interseção 66 vezes maior do que o esperado ao acaso (**Tabela 7** e **8**). Todas as variantes sobrepostas compartilham alelos idênticos (100% de identidade, OR = ∞), configurando o padrão mais significativo observado em nossas análises. Esse resultado indica que as variantes das classes SNPs e Benign não apenas ocorrem nas mesmas posições genômicas, mas também compartilham os mesmos alelos com alta frequência, o que seria melhor explicado pelo **Modelo 3 - Figura 19,** caso um dos conjuntos representasse variantes patogênicas.

A força dessa associação, juntamente com a alta sobreposição de *loci* e a identidade alélica, valida as metodologias empregadas na seleção desses conjuntos. Como esperado, ambos os conjuntos são enriquecidos em variantes neutras ou quase neutras, refletindo a natureza

funcional similar dessas categorias e confirmando a robustez das abordagens analíticas utilizadas para o estabelecimento dos diferentes conjuntos, funcionando como um controle positivo das nossas análises.

Pathogenic vs. Benign

A análise revelou uma sobreposição significativa de 192 variantes compartilhando coordenadas CDS, com um OR de 2,1, indicando que a coocorrência entre os loci dessas variantes ocorre aproximadamente duas vezes mais do que o esperado ao acaso (**Tabela 7** e **Tabela 8**). No entanto, todas as variantes apresentaram dissimilaridade completa de alelos (100%, OR = 0, **Tabela 9**), sugerindo uma clara separação alélica entre essas classes.

Variantes classificadas como benignas no ClinVar são aquelas que, com base nas evidências disponíveis, não foram associadas a doenças ou condições patológicas. Espera-se que muitas dessas variantes benignas tenham sido definidas a partir de uma variante patogênica, de alelo distinto, que ocorre no mesmo locus. Assim, a sobreposição observada pode refletir um artefato inerente ao processo de curadoria adotados para definir esses conjuntos de variantes pelo banco de dados ClinVar.

Esses resultados reforçam a distinção funcional entre esses dois grupos de SNVs em relação à sua associação com doenças. A dissimilaridade alélica observada apoia a robustez dessas classificações, refletindo variantes de alta penetrância para doenças versus variantes neutras ou quase neutras. O cenário observado é mais bem explicado pelo **Modelo 2** de coocorrência mutacional (**Figura 19**), no qual mutações ocorrem nos mesmos *loci*, mas com alelos diferentes. Avaliados de maneira conjunta, estes resultados também provêm um importante controle positivo das nossas análises.

Rares vs. Benign

A análise revelou que SNVs raras e benignas coocorrem 16,2 vezes mais frequentemente do que o esperado, com uma interseção significativa de 17.777 loci (**Tabela 7** e **Tabela 8**). Além disso, 98% das variantes sobrepostas compartilham alelos idênticos, indicando uma forte associação alélica entre essas categorias (**Tabela 9**).

Assim como no padrão observado entre SNPs e Benign, espera-se que o conjunto Rares contenha uma fração substancial de variantes germinativas neutras ou de baixo impacto funcional (**Figura 17** e **Figura 18**), provavelmente incluindo mutações *de novo* não sujeitas a pressões seletivas significativas (Rentzsch *et al.*, 2019). A elevada frequência de alelos idênticos sugere que a coocorrência entre variantes raras e benignas, ambas de origem germinativa, não é aleatória, mas sim influenciada por processos evolutivos ou mecanismos mutacionais compartilhados. Esses resultados são mais bem explicados pelo **Modelo 3** de coocorrência mutacional (**Figura 19**), no qual mutações compartilham tanto loci quanto alelos.

Rares vs. COSMIC

A análise inicial revelou uma sobreposição significativa de 1,3 vezes maior do que o esperado ao acaso entre SNVs raras e somáticas relacionadas ao câncer (**Tabela 7** e **Tabela 8**) com uma similaridade alélica de 77% (**Tabela 9**). Esses resultados sugerem uma possível distribuição não aleatória entre mutações germinativas raras e mutações somáticas do câncer, potencialmente associadas a mecanismos mutacionais compartilhados. Este cenário seria mais bem representado pelo **Modelo 3** (**Figura 19**), onde mutações compartilham tanto loci quanto alelos.

Entretanto, foi observada uma discrepância entre os resultados de CDS e WG: enquanto a associação foi positiva para CDS (OR = 1,3), observou-se uma associação negativa para WG (OR = 0.558). Considerando a abordagem inicial adotada neste estudo, que foi propositalmente mais permissiva com o conjunto de variantes raras para evitar perda desnecessária de dados, o desequilíbrio de ligação (LD) pode ter influenciado os resultados. O LD tende a aumentar a frequência de sobreposições aparentes devido a blocos de variantes correlacionadas em regiões genômicas próximas, mesmo que estas variantes não compartilhem uma relação funcional ou evolutiva direta. Para investigar essa possibilidade, as variantes raras foram filtradas por LD, e os testes de Fisher foram repetidos.

Após a filtragem, a associação tornou-se negativa tanto para WG quanto para CDS (WG OR = 0,67; CDS OR = 0,522; **Tabela 10**), eliminando sobreposições infladas e permitindo uma avaliação mais precisa da coocorrência verdadeira. Além disso, para os outros testes envolvendo variantes raras, a direção das associações permaneceu consistente após a

filtragem por LD, mas o número de sobreposições foi reduzido, reforçando o papel do LD como um fator de confusão em análises genômicas.

Tabela 10. Resultados do teste exato de Fisher para testes pareados com conjunto Rares_LD (SNVs raras em equilíbrio de ligação). O teste exato de Fisher, juntamente com o fornecimento de valores *p* e odds ratios (OR) para interseções entre

CDS							
Dataset A	**Dataset B**	***p***	**OR**	**in_A/in_B**	**in_A/not_B**	**not_A/in_B**	**not_A/not_B**
Rares_LD	Cosmic	0	0.522	39337	93512	3623529	4493760
Rares_LD	Benign	0	11.153	8339	124510	48455	8068834
Rares_LD	Pathogenic	1.6e-36	0.267	57	132792	13041	8104248
WG							
Rares_LD	Cosmic	0	0.670	136384	13536348	10451162	694626486
Rares_LD	Benign	0	8.462	10844	13661888	66127	705011521
Rares_LD	Pathogenic	1.8e-65	0.203	64	13672668	16249	705061399

Esses resultados destacam a importância de balancear permissividade e rigor metodológico ao trabalhar com variantes raras. Embora a abordagem inicial tenha priorizado a inclusão de variantes, a filtragem por LD demonstrou ser essencial para refinar as análises e interpretar os padrões de coocorrência com maior precisão, enfatizando a complexidade dos cenários mutacionais e reforçando a necessidade de estratégias metodológicas customizadas para minimizar a ocorrência de viéses causados por *e.g.* artefatos estatísticos.

SNPs vs. COSMIC

A análise revelou que variantes SNPs e somáticas relacionadas ao câncer (COSMIC) coocorrem 5,6 vezes mais frequentemente do que o esperado ao acaso (**Tabela 7** e **Tabela 8**). Aproximadamente 60% das SNVs polimórficas (9.575 de 16.040) compartilham as mesmas coordenadas genômicas nas regiões CDS que as variantes somáticas do COSMIC. Além disso, 97,7% das SNVs sobrepostas apresentam o mesmo alelo alternativo (OR = 129,34), mesmo sendo variantes de origens distintas (germinativa e somática).

A distribuição não aleatória observada entre variantes germinativas polimórficas e mutações somáticas do câncer indica a presença de *hotspots* mutacionais, sugerindo a existência de mecanismos mutacionais compartilhados. Esse fenômeno é melhor representado pelo **Modelo 3** (**Figura 19**), no qual mutações de diferentes contextos funcionais, sejam elas neutras

polimórficas ou patogênicas associadas a doenças, ocorrem em um mesmo locus e compartilham os mesmos alelos.

Entre os conjuntos analisados, SNPs foi a única classe de variantes germinativas que apresentou uma associação positiva significativa com variantes somáticas relacionadas ao câncer, tanto para CDS quanto para WG, destacando-se como um subconjunto relevante para análises mais aprofundadas. O baixo viés de distribuição para WG e a consistência das associações observadas reforçam a importância dessas variantes no contexto genômico.

De forma mais ampla, as sobreposições significativas observadas sugerem que um ou mais mecanismos mutacionais comuns podem estar contribuindo para a ocorrência de diferentes classes de SNVs com maior ou menor frequência do que o esperado. A sobreposição entre SNPs e COSMIC é particularmente relevante por ser a única associação positiva que envolve variantes com origens distintas (germinativa e somática), diferentes consequências funcionais (neutras/benignas e patogênicas) e impactos evolutivos (SNPs e mutações relacionadas ao câncer). Propõe-se que as variantes sobrepostas ocorram em *hotspots* mutacionais devido a mecanismos moleculares compartilhados, cuja natureza ainda é desconhecida.

4.4. Padrões de substituição de nucleotídeo único dos subconjuntos de SNVs

A partir dos resultados de coocorrência das variantes, produzimos subconjuntos de SNVs que apresentaram associações positivas ou negativas (seção 4.3.2). A análise dos padrões de substituição de nucleotídeo único revelou diferenças significativas entre os subconjuntos, refletindo possíveis processos biológicos subjacentes que moldam essas mutações.

As substituições mais frequentes identificadas foram C>T e sua complementar G>A, predominantemente em *loci* com alelos alternativos compartilhados, que compõem a maioria dos casos analisados (**Figura 20.C**, **Figura 21.C**). Este padrão é consistente com o processo de desaminação espontânea de citosinas metiladas em dinucleotídeos CpG, um mecanismo bioquímico amplamente descrito em genomas humanos (Guo *et al.*, 2017).

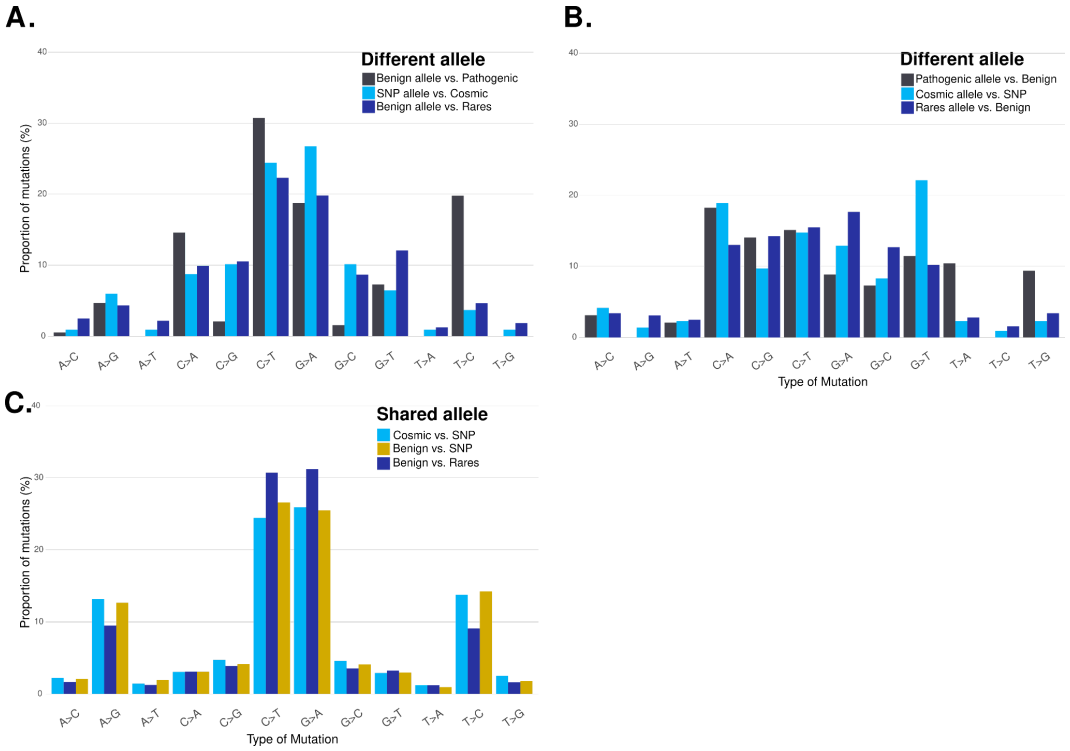


Figura 20. Padrão de substituições de nucleotídeo único em subconjuntos dos testes com associação <u>positiva</u> pelo teste de Fisher. (A e B) Substituições para posições com alelos diferentes nos pares de conjuntos analisados (e.g., Benign vs. Pathogenic, SNPs vs. COSMIC). **(C)** Padrões de alelos compartilhados entre subconjuntos (e.g., SNP vs. COSMIC, Benign vs. SNPs).

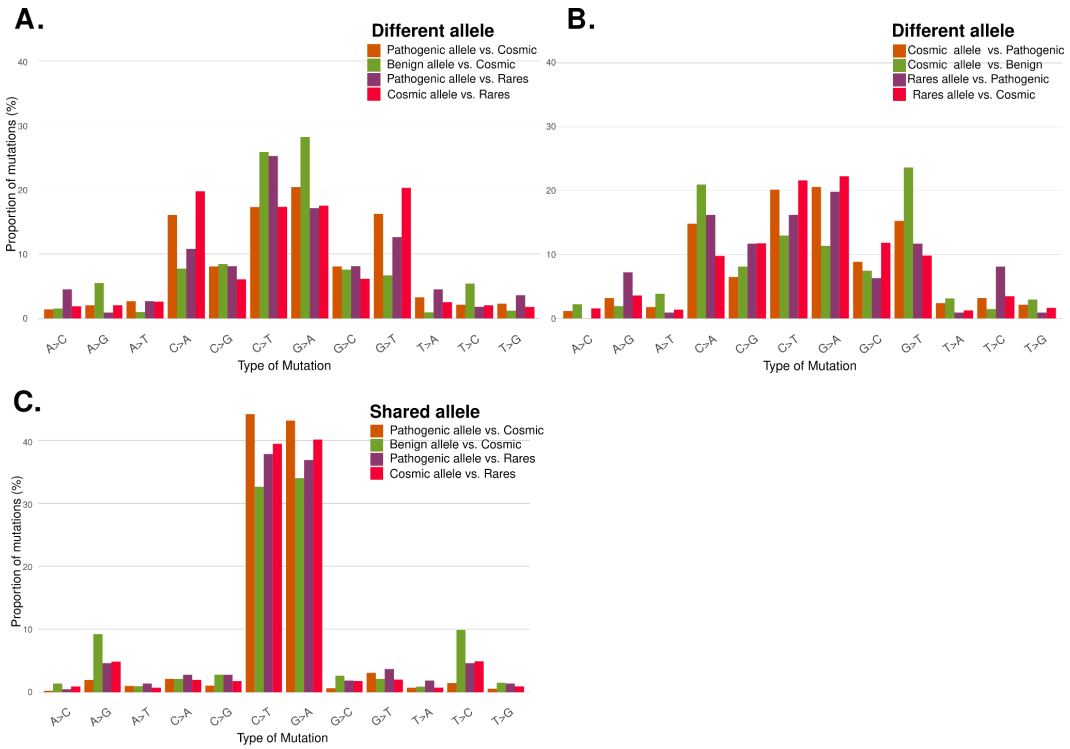


Figura 21. Padrão de substituições de nucleotídeo único em subconjuntos dos testes com associação <u>negativa</u> pelo teste de Fisher. (A e B) Substituições para posições com alelos diferentes nos pares de conjuntos analisados (e.g., Pathogenic vs. COSMIC). **(C)** Alelos compartilhados entre os subconjuntos (e.g., COSMIC vs. Rares).

Nos testes com associações positivas, as substituições C>T e G>A foram dominantes para quase todos, tanto em *loci* com alelos diferentes (**Figura 20.A** e **20.B**), mas principalmente para alelos compartilhados (**Figura 20.C**). No entanto, as proporções dessas substituições variaram entre os subconjuntos.

Nos testes sem associação significativa de sobreposição, padrões mais heterogêneos foram observados, especialmente em *loci* com alelos diferentes (**Figura 21.A** e **21.B**). Entretanto, *loci* com alelos compartilhados (**Figura 21.C**) mantiveram a predominância de C>T/G>A, reforçando a universalidade do processo de desaminação citosina-metilada em *loci* genômicos.

De modo geral, os padrões observados em *loci* com alelos compartilhados destacaram a maior proporção de substituições C>T/G>A em ambos os cenários (associação positiva e negativa). Em contraste, *loci* com alelos diferentes apresentaram maior heterogeneidade, com destaque para substituições como G>T e C>A em alelos patogênicos (**Figura 20.B** e **21.B**).

4.5. Análises do subconjunto *SNPs vs. COSMIC*

4.5.1. Caracterização do subconjunto *SNPs vs. COSMIC*

O conjunto de variantes polimórficas foi o único conjunto germinativo que apresentou uma associação significativa com as SNVs somáticas do câncer, tanto no conjunto de CDS como no conjunto WG. A partir dessa associação, produzimos dois novos subconjuntos de SNVs representando a interseção entre SNPs e COSMIC: um para variantes sobrepostas no WG e outro para aquelas localizadas exclusivamente em CDS. Esses conjuntos compreendem 34.725 e 9.575 *loci*, respectivamente.

No subconjunto *SNPs vs. COSMIC* do WG, a maioria das SNVs sobrepostas (70%) ocorre em regiões intrônicas, indicando um ponto de partida promissor para estudos de variantes não codificadoras (**Figura 22**). Variantes intrônicas, apesar de não afetarem diretamente a sequência proteica, podem impactar elementos regulatórios, como enhancers, ou processos como *splicing* alternativo, desempenhando um papel crítico em algumas doenças genéticas e no câncer (Barbosa *et al.*, 2020).

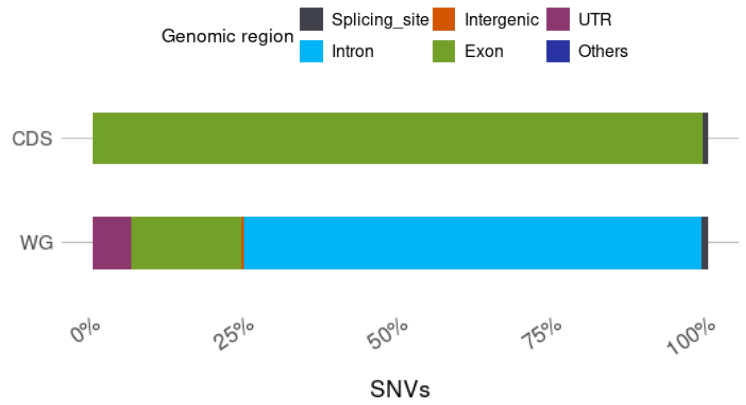


Figura 22. Proporção de SNVs do CDS e WG por regiões genômicas no subconjunto *SNP vs. COSMIC*. A distribuição das variantes é apresentada de acordo com a classificação genômica em intrônicas, exônicas, intergênicas, UTRs, splicing sites e outras regiões genômicas. No WG, observa-se uma predominância de SNVs em regiões intrônicas, representando aproximadamente 70% do total, enquanto, no CDS, a maior parte das variantes está localizada em regiões exônicas.

Em CDS, as mutações *synonymous* e *missense* compreendem as classes mais abundantes observadas sobreposição *SNP vs. COSMIC*, refletindo um impacto funcional predominantemente moderado ou baixo, assim como o observado para as variantes *synonymous* e *missense* do WG (**Figura 23**, **Tabelas 9** e **10**).

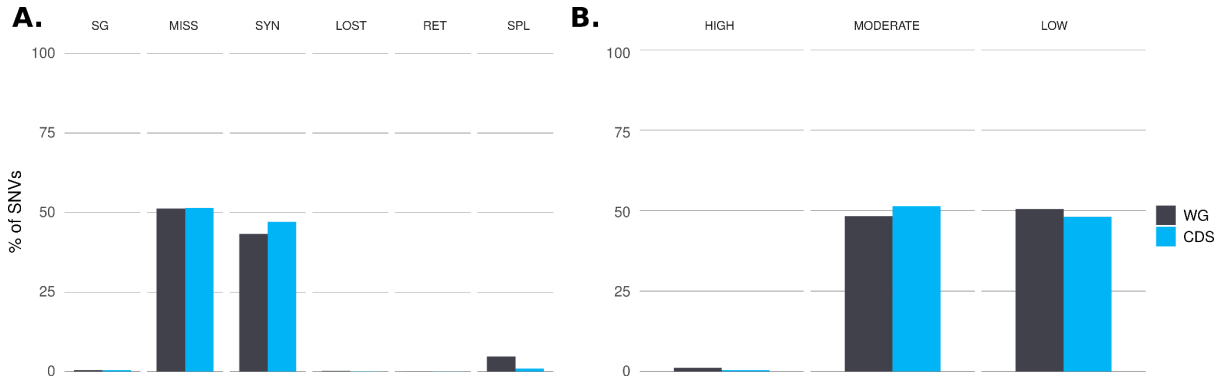


Figura 23. Consequências de codificação e impacto funcional das SNVs no subconjunto *SNP vs. COSMIC*. **(A)** Proporção de SNVs para diferentes tipos de consequências de codificação: SG (*Stop-gained*), MISS (*Missense variant*), SYN (*Synonymous variant*), LOST (*Start lost e Stop lost*), RET (*Stop retained variant*), SPL (*Splice region variant*). As variantes *missense* e *synonymous* são predominantes em ambos os conjuntos (WG e CDS). **(B)** Impacto funcional das SNVs categorizado em três classes: HIGH (impacto significativo, como perda de função), MODERATE (impacto moderado) e LOW (impacto mínimo ou benigno). Variantes de impacto moderado e baixo são mais frequentes em ambos os conjuntos.

Tabela 11. Contagem e proporção de SNVs do WG para consequências de codificação no subconjunto *SNP vs. COSMIC*.

Consequence type	Count	Percent (%)
3 prime UTR variant	2652	4.451

5 prime UTR variant	857	1.438
downstream gene variant	30	0.050
intergenic variant	65	0.109
intron variant	41778	70.133
mature miRNA variant	4	0.006
missense variant	5463	9.170
non coding transcript exon variant	2821	4.735
regulatory region variant	11	0.018
splice acceptor variant	24	0.040
splice donor 5th base variant	29	0.048
splice donor region variant	98	0.164
splice donor variant	30	0.050
splice polypyrimidine tract variant	458	0.768
splice region variant	504	0.846
start lost	14	0.023
stop gained	68	0.114
stop lost	6	0.010
stop retained variant	7	0.011
synonymous variant	4561	7.656
TF binding site variant	1	0.001
upstream gene variant	88	0.147

Tabela 12. Contagem e proporção de SNVs do CDS para consequências de codificação no subconjunto *SNP vs. COSMIC.*

Consequence type	Count	Percent
missense variant	4926	51.446
splice region variant	89	0.929
start lost	4	0.041
stop gained	40	0.417
stop lost	2	0.020
stop retained variant	4	0.041
synonymous variant	4510	47.101

A análise dos *scores* de patogenicidade das SNVs *missense* sobrepostas revelou um pico de pontuações baixas (**Figura 24**). Variantes *missense* de baixo impacto no conjunto SNPs são consistentes com o perfil de SNVs neutras ou quase neutras quando analisadas em ambientes biológicos normais, dado que apresentam alta AF mundial. No entanto, no COSMIC, que

representa variantes recorrentes no câncer, mesmo SNVs com baixo *score* de patogenicidade, por estarem inseridas em um microambiente altamente heterogêneo, sua classificação como variantes patogênicas não pode ser desconsiderada, mesmo que de baixa penetrância. Em tumores, essas variantes podem atuar como mutações *passenger*, que, apesar de não impulsionarem diretamente a oncogênese como mutações *driver*, podem contribuir para a progressão tumoral ao interagir com vias alteradas ou modular a resposta celular ao ambiente neoplásico (McFarland *et al.*, 2017; Kumar *et al.*, 2020). Assim, a presença recorrente dessas variantes no COSMIC sugere que sua patogenicidade deve ser avaliada no contexto tumoral, onde efeitos indiretos e de penetrância variável podem desempenhar papéis críticos na evolução do câncer.

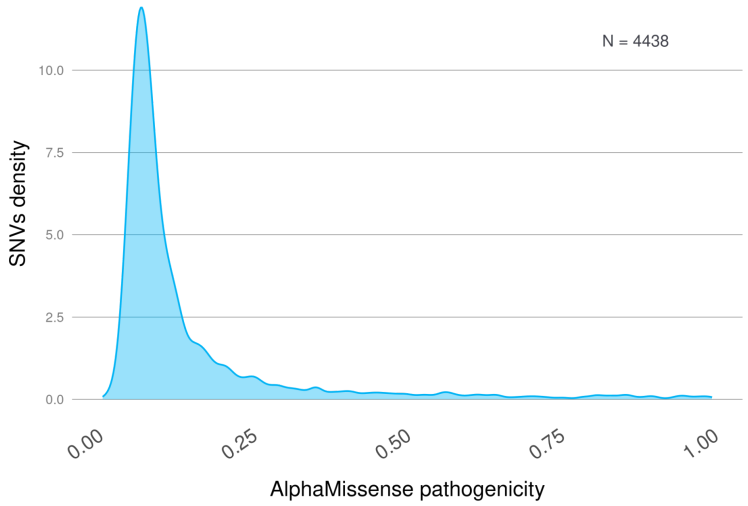


Figura 24. Distribuição das pontuações de patogenicidade do AlphaMissense para SNVs missense do CDS do subconjunto *SNPs vs. COSMIC*. O gráfico apresenta a densidade das pontuações de patogenicidade atribuídas pelo AlphaMissense para 4438 variantes *missense* compartilhadas entre os conjuntos SNPs e COSMIC. A maioria das variantes apresenta pontuações de patogenicidade baixas (<0.25), indicando um impacto funcional menor ou nulo, consistente com o perfil de mutações neutras ou *passenger*. A presença de uma cauda longa à direita sugere a existência de uma pequena proporção de variantes com pontuações mais altas, possivelmente associadas a impactos mais significativos, como mutações driver em contextos específicos.

Adicionalmente, identificamos uma pequena fração de mutações que potencialmente representam *drivers* no câncer, como mutações *stop-gained*, *start-lost* e *stop-lost* (**Tabelas 11** e **12**). Essas mutações de perda de função podem ter implicações importantes em genes relacionados ao câncer e poderiam ser exploradas mais profundamente com preditores de mutações *driver* (Ostroverkhova *et al.*, 2023). No contexto do conjunto de SNPs, essas mutações podem refletir outros aspectos biológicos além de sua relação com o câncer. Por exemplo, essas variantes podem ter surgido em genes ou regiões onde a pressão seletiva é

atenuada, como em genes não essenciais ou redundantes, o que explicaria sua retenção em populações humanas (Rentzsch *et al.*, 2019, Ostroverkhova *et al.*, 2023). Em outra perspectiva, algumas dessas variantes poderiam influenciar indiretamente a função gênica por meio de efeitos cis-regulatórios, como a regulação da transcrição ou o *splicing* (Barbosa *et al.*, 2020, Kumar *et al.*, 2020). Outra possibilidade é que essas mutações representem variação genética que, embora atualmente neutra, possa fornecer uma base adaptativa em condições futuras ou específicas, contribuindo para a diversidade genética e a resiliência populacional (Quintana-Murci, 2016, Karczewski *et al.*, 2020). Um exemplo clássico desse fenômeno é a mutação responsável pela anemia falciforme, que oferece resistência à malária em portadores heterozigotos. Algo semelhante poderia ser verdadeiro para algumas variantes do conjunto SNPs, onde sua persistência no genoma poderia ser um reflexo de sua importância evolutiva em condições específicas.

De maneira geral, esses resultados ressaltam a relevância do subconjunto *SNPs vs. COSMIC* para investigações sobre *hotspots* mutacionais e possíveis mecanismos biológicos compartilhados entre variantes germinativas e somáticas, servindo como um conjunto base para análises de assinaturas mutacionais e de enriquecimento biológico realizadas posteriormente.

4.5.2. Assinaturas mutacionais associadas do subconjunto *SNPs vs. COSMIC*

Assinaturas mutacionais representam padrões específicos de mutações associados a processos mutagênicos particulares ou deficiências em mecanismos de reparo de DNA (Alexandrov *et al.*, 2020, Koh *et al.,* 2021). Para explorar mecanismos potencialmente associados à coocorrência de mutações entre SNPs e SNVs somáticas do câncer, avaliamos as assinaturas mutacionais relacionadas às variantes encontradas simultaneamente nos conjuntos ao subconjunto *SNPs* vs. *COSMIC*. O algoritmo *SigProfiler assignment* implementado no COSMIC v3.4 (Díaz-Gay *et al.*, 2023, Sondka *et al.*, 2024) foi aplicado aos subconjuntos *SNPs vs. COSMIC* do WG e do CDS.

Três assinaturas mutacionais foram predominantes no conjunto WG: SBS5, contribuindo com 73,01% das mutações, SBS1, com 15,48%, e SBS54, com 11,52% (**Figura 25**). Interessantemente, a assinatura SBS5 está associada a processos de origem desconhecida e frequentemente vinculada a mecanismos endógenos relacionados ao envelhecimento celular.

A SBS1, por sua vez, é associada à desaminação espontânea de 5-metilcitosina, enquanto a SBS54 está relacionada à instabilidade de microssatélites (MSI) e deficiência no reparo do DNA (dMMR) (Yaacov *et al.*, 2023, Yaacov *et al.*, 2024).

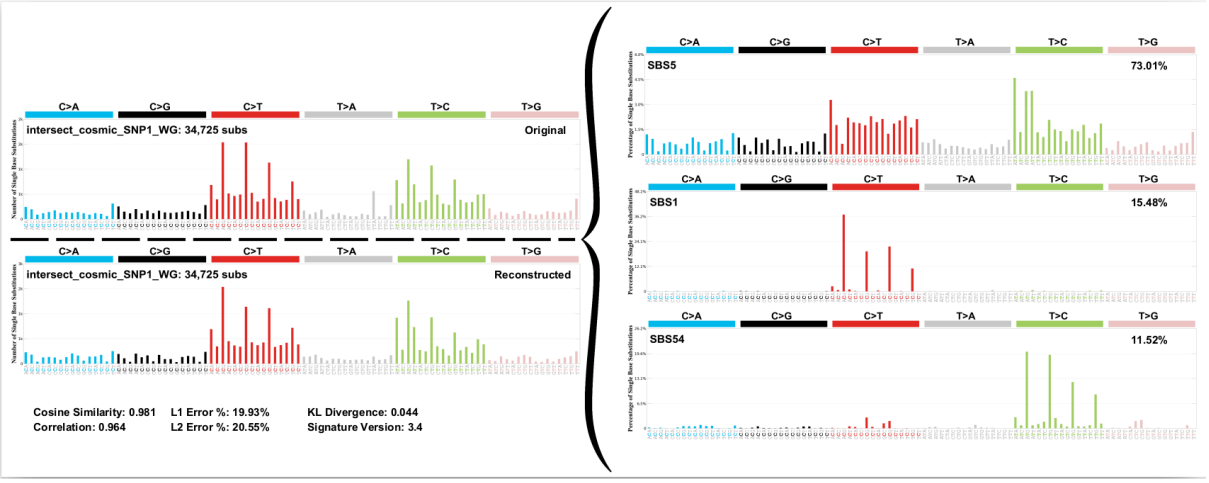


Figura 25. Assinatura mutacional no subconjunto *Cosmic vs. SNP* para WG. A análise conduzida com o SigProfiler Assignment identificou SBS5, SBS1 e SBS54 como predominantes no conjunto de dados WG. A SBS5 corresponde a 73,01% das mutações e está associada a processos de origem desconhecida, enquanto a SBS1, representando 15,48%, está ligada à desaminação espontânea de 5-metilcitosina. A SBS54 (11,52%) está associada à instabilidade de microssatélites (MSI) e à deficiência no reparo do DNA (dMMR). Uma similaridade cosseno de 0,981 e uma correlação de 0,964 indicam um forte alinhamento entre as assinaturas do nosso subconjunto e as assinaturas de referência do COSMIC. Os erros L1 e L2 (19,93% e 20,55%) estão dentro dos limites aceitáveis, apoiando um ajuste adequado, enquanto uma divergência KL de 0,044 confirma uma semelhança próxima na distribuição.

As métricas fornecidas pelo *SigProfiler Assignment* indicaram uma forte similaridade entre as assinaturas do subconjunto e as assinaturas COSMIC de referência. Os valores observados foram:

- *Cosine similarity* = 0,981
- *Correlation* = 0,964
- L1 Error % = 19,93%
- L2 Error % = 20.55%
- KL Divergence = 0.044

O *cosine similarity* mede o quanto a assinatura mutacional obtida a partir dos dados fornecidos se assemelha às assinaturas de referência no catálogo COSMIC, em uma escala de 0 a 1. Um valor de 0.981 indica uma semelhança muito alta, o que sugere que a assinatura reconstruída a partir das amostras fornecidas se assemelha bastante às assinaturas COSMIC usadas na comparação. A *correlation* entre a assinatura observada e as assinaturas de referência também está em uma escala de 0 a 1, com 1 indicando uma correlação perfeita.

Uma correlação de 0.964 também indica uma forte correspondência entre as assinaturas identificadas em seus dados e as assinaturas COSMIC, suportando a ideia de que as assinaturas encontradas são consistentes com o catálogo. O erro L1 mede a diferença absoluta entre as assinaturas identificadas e as assinaturas de referência. Em outras palavras, ele quantifica o erro de ajuste para cada ponto da assinatura em relação ao catálogo. Um erro de 19.93% significa haver uma discrepância moderada entre a assinatura reconstruída e as assinaturas COSMIC de referência. Isso indica que o modelo conseguiu capturar a maioria das características, mas há uma margem de variação na precisão da reconstrução. O erro L2 mede a diferença quadrática média entre a assinatura reconstruída e a assinatura de referência. Valores mais baixos indicam menor erro. Um erro de 20.55% é semelhante ao L1 e sugere uma precisão razoável, embora haja um desvio residual na reconstrução. A divergência KL (*Kullback-Leibler Divergence*) mede a diferença entre a distribuição da assinatura mutacional da amostra e as assinaturas de referência. Valores mais próximos de 0 indicam uma melhor correspondência. Um valor de 0.044 é muito baixo, sugerindo que a distribuição das mutações da amostra fornecida é muito semelhante à distribuição da assinatura de referência.

A análise limitada às regiões CDS revelou um perfil mutacional similar, porém com algumas assinaturas a mais (**Figura 26**). As assinaturas SBS5 (32,27%), SBS1 (12,87%) e SBS54 (15,3%) permaneceram presentes, embora outras assinaturas específicas também tenham sido observadas: SBS87 (21,51%) e SBS6 (13,04%). As assinaturas exclusivas do conjunto CDS indicam processos mutacionais mais específicos, possivelmente mascarados no WG devido à predominância de mutações em regiões intrônicas e intergênicas.

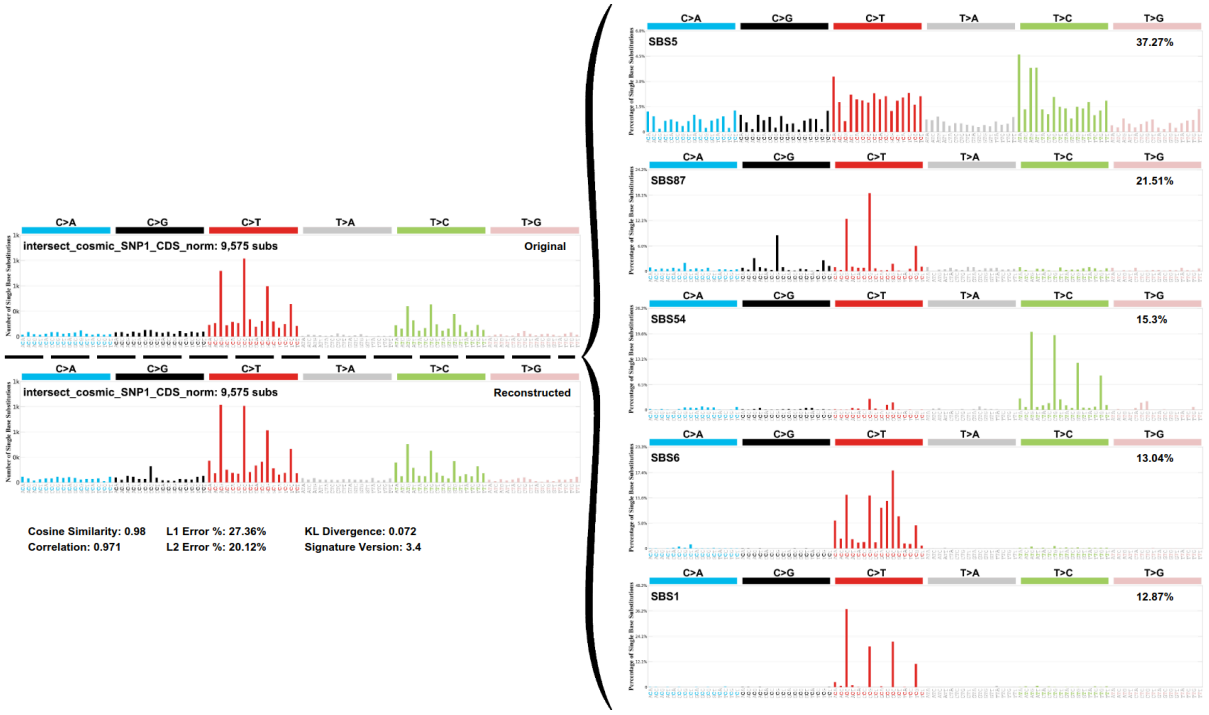


Figure 26. Perfis de assinaturas mutacionais identificados no subconjunto CDS de SNVs sobrepostas entre COSMIC e SNPs. A SBS5 representa 32,27% das mutações, enquanto a SBS1, foi responsável por 12,87%. Outras assinaturas emergiram, incluindo SBS87 (21,51%), SBS54 (15,3%) e SBS6 (13,04%). A similaridade de cosseno de 0,98 e a correlação de 0,971 indicam uma forte semelhança entre as assinaturas do subconjunto e as assinaturas de referência COSMIC. Os erros L1 e L2 (27,36% e 20,12%) reforçam a adequação do modelo, enquanto a divergência KL de 0,072 confirma uma correspondência próxima entre as distribuições.

As métricas de similaridade para CDS também indicaram uma boa correspondência com as assinaturas de referência:

- Cosine similarity = 0,98
- Correlation = 0,971
- L1 Error % = 27,36%
- L2 Error % = 20,12%
- KL Divergence = 0.072

A ausência das assinaturas mutacionais SBS87 e SBS6 no conjunto WG, mas sua detecção quando a análise é restringida às regiões CDS, pode estar associada a um efeito de diluição devido à alta prevalência da assinatura SBS5 no WG. Essa predominância pode estar relacionada a regiões intrônicas e intergênicas, que representam aproximadamente 70% das variantes do nosso conjunto. O SigProfiler Signature baseia-se na decomposição matemática de padrões mutacionais para identificar assinaturas presentes em um conjunto de dados. No entanto, quando uma assinatura altamente prevalente, como SBS5, domina a composição mutacional, ela pode mascarar assinaturas de menor frequência, como SBS87 e SBS6,

dificultando sua identificação. Isso ocorre porque o modelo prioriza componentes com maior representatividade estatística, reduzindo a detecção de assinaturas menos abundantes. Ao restringir a análise ao CDS, a influência da assinatura predominante (SBS5) é reduzida, permitindo que SBS87 e SBS6 se tornem detectáveis. Esse resultado sugere que essas assinaturas estão particularmente associadas a mutações em CDS e que, no WG, a assinatura SBS5 seja a mais abundante, possivelmente devido a seu predomínio em regiões não codificantes do genoma.

Na **Tabela 13**, são descritas de maneira resumida as cinco assinaturas mutacionais observadas neste trabalho (SBS1, SBS5, SBS87, SBS54 e SBS6). A tabela destaca os mecanismos mutacionais principais, os contextos biológicos associados, as implicações clínicas e as referências relevantes, fornecendo uma visão geral dos padrões mutacionais e sua relação com os processos biológicos subjacentes, sintetizando o papel de cada assinatura no cenário mutacional identificado, para auxiliar na interpretação dos resultados apresentados.

Tabela 13. Resumo das assinaturas mutacionais observadas no subconjunto *SNPs vs. COSMIC*. A tabela descreve as cinco assinaturas mutacionais identificadas (SBS1, SBS5, SBS87, SBS54 e SBS6), incluindo seus mecanismos mutacionais principais, contextos biológicos, associações clínicas e referências relevantes. Essas assinaturas refletem processos endógenos e exógenos que contribuem para padrões mutacionais distintos, com implicações na estabilidade genômica e no desenvolvimento de doenças, como o câncer. *Siglas: CpG (dinucleotídeo citosina-fosfato-guanina), ALL (leucemia linfoblástica aguda), MSI (instabilidade de microssatélites), dMMR (deficient DNA mismatch repair), MLH1 (gene da mutL homólogo 1, envolvido no sistema de reparo de DNA por incompatibilidade).*

Assinatura	Mecanismo Principal	Contexto Biológico	Associações Clínicas	Referências
SBS1	Desaminação de CpG	Relógio biológico, idade	Presente em vários cânceres e envelhecimento	Alexandrov *et al.*, 2020; Wang *et al*., 2024
SBS5	Etiologia desconhecida	Relógio biológico	Associada a TP53 e instabilidade genômica	Yamamoto *et al.,* 2024; Yaacov *et al.*, 2024
SBS87	Exposição a tiopurinas	Quimioterapia (ALL)	Induzida por mercaptopurina/ tioguanina	Li *et al.*, 2020; van der Ham *et al.*, 2024
SBS54	Artefato de sequenciamento ou MSI/dMMR	Reparo defeituoso	Cânceres com MSI/dMMR, especialmente colorretais	Ji *et al.*, 2023; Farmanbar *et al.*, 2023
SBS6	Instabilidade MSI	Hipermetilação de MLH1	Síndrome de Lynch e tumores MSI-H	Farmanbar *et al.*, 2023; Jayakrishnan *et al.*, 2024

Os resultados obtidos neste trabalho dialogam com os achados por Wang *et al.* (2024), que também observaram uma prevalência significativa de SBS1 (13,87%) e SBS87 (10,02%) em variantes somáticas compartilhando *loci* com mutações germinativas. Esses autores destacaram o impacto de processos endógenos, como a desaminação de citosinas, e exógenos, como exposições terapêuticas, na geração de padrões mutacionais complexos.

4.5.3. Interseção entre o subconjunto *SNPs vs. COSMIC* (WG) e coordenada de Ilhas CpG ou microssatélites

O padrão de substituições de nucleotídeos únicos foi avaliado em todos os subconjuntos produzidos pelas interseções dos testes pareados (**Figuras 20** e **21**). As substituições mais frequentes identificadas foram C>T e sua complementar G>A, especialmente nos *loci* com alelos idênticos compartilhados entre diferentes conjuntos. No subconjunto *SNPs vs. COSMIC* do WG, essas substituições corresponderam a aproximadamente 24% e 27% do total identificado, respectivamente (**Figura 20**). Esse padrão sugere que a desaminação espontânea de citosinas metiladas, um mecanismo associado às ilhas CpG, possa estar contribuindo significativamente para a prevalência dessas mutações.

As Ilhas CpG são *hotspots* mutacionais clássicos, frequentemente vinculados à assinatura SBS1, que reflete a desaminação espontânea de 5-metilcitosina para timina. Essa assinatura foi observada tanto no WG quanto no CDS neste trabalho (**Figuras 25** e **26**), embora não seja a assinatura mais abundante em nenhum destes conjuntos. Já os microssatélites, associados às assinaturas SBS6 e SBS54, estão relacionados a defeitos no reparo de DNA (*deficient DNA mismatch repair* - dMMR) e instabilidade de microssatélites (MSI) (Alexandrov *et al.*, 2020; Ji *et al.*, 2023). Considerando essas associações, realizamos uma investigação adicional para avaliar a coocorrência entre as variantes do subconjunto *SNPs vs. COSMIC* e regiões de microssatélites, visando explorar se características de instabilidade genômica contribuem para os padrões mutacionais compartilhados.

Aplicamos o teste exato de Fisher utilizando o BEDTools, seguido do cálculo da proporção de coocorrência entre o subconjunto *SNPs vs. COSMIC* e as ilhas CpG ou os microssatélites. Para as ilhas CpG, observamos um OR elevado (7,0), indicando que a sobreposição entre essas regiões e o subconjunto analisado ocorreu com uma frequência maior do que a esperada ao acaso. No entanto, apenas cerca de 5% das mutações (1623 de 34.725) puderam ser atribuídas a mecanismos associados às ilhas CpG (**Tabela 14** -

Subset_SNP_intersect_COSMIC vs. CpG Island). Esses resultados estão de acordo com a baixa prevalência da assinatura SBS1, cuja origem é a desaminação espontânea de citosinas metiladas (**Figuras 25** e **26**). Assim, embora um número significativo de SNVs coocorra em ilhas CpG, a desaminação de citosinas nessas regiões não parece ser o principal fator responsável pelas mutações observadas, sugerindo a existência de outros hotspots mutacionais.

Para os microssatélites, a análise revelou uma coocorrência 2 vezes maior do que o esperado ao acaso entre essas regiões e o subconjunto SNPs vs. COSMIC (**Tabela 14** - *Subset_SNP_intersect_COSMIC vs. Microsatellite*). No entanto, apenas 0,1% das mutações (45 de 34.725) puderam ser atribuídas à presença de microssatélites. Esses dados estão de acordo com a baixa contribuição das assinaturas mutacionais SBS6 e SBS54. A assinatura SBS6, associada à instabilidade de microssatélites (MSI), não foi detectada no WG devido à sua baixa prevalência, enquanto a SBS54, embora identificada, apresentou uma frequência ainda menor do que a SBS1. Assim, embora microssatélites também possam influenciar os padrões mutacionais observados, seu impacto no subconjunto analisado é minoritário.

Tabela 14. Resultados do teste exato de Fisher para coocorrência entre o subconjunto *SNPs vs. COSMIC* e ilhas CpG ou microssatélites. Os valores fornecidos incluem odds ratio (OR) e p-values (< 0,001), avaliando a associação entre dois conjuntos de coordenadas genômicas (Dataset 1 e Dataset 2). O OR > 1 indica uma coocorrência maior do que o esperado ao acaso, enquanto valores de p significativos refletem associação estatística robusta.

				WG			
Dataset A	**Dataset B**	***p***	**OR**	**in_A/in_B**	**in_A/not_B**	**not_A/in_B**	**not_A/not_B**
Subset SNP intersect Cosmic	CpG Island	0	7.029	1623	33102	25240	3618606
Subset SNP intersect Cosmic	Microsatellite	4e-37	2.261	45	34680	37839	65939551

4.5.4. Análise de enriquecimento de genes mutados no subconjunto *SNPs vs. COSMIC*

Nossas análises revelaram uma intrigante sobreposição de *loci* genômicos entre variantes germinativas de alta frequência alélica (AF ≥ 1%) e mutações somáticas associadas ao câncer, que em sua maioria compartilham o mesmo alelo alternativo. Os mecanismos responsáveis

por esse fenômeno ainda não foram esclarecidos. Para expandir o conhecimento sobre esse padrão mutacional, investigamos as regiões genômicas, vias e genes mais afetados por essas mutações compartilhadas, realizando uma análise de enriquecimento funcional (Over-Representation Analysis - ORA) baseada em bandas citogenéticas e vias KEGG.

A interseção entre coordenadas de SNPs e mutações COSMIC revelou 6.149 genes afetados por variantes nos dois cenários (**Figura 27.A**). As regiões 13q12.12, 19p13.3, 19p13.12, 5q31.3, 22q13.33, 6q27 e o segmento chrHG2308_PATCH foram significativamente enriquecidas (**Figura 27.B**). Em particular, o enriquecimento das regiões 19p13.12 (ratio = 1.7, FDR = 0,018) e 19p13.3 (ratio = 1.46, FDR = 0,002) aponta para um possível papel do braço curto do cromossomo 19 como um reservatório de SNPs em sítios de ocorrência de mutações germinativas-somáticas.

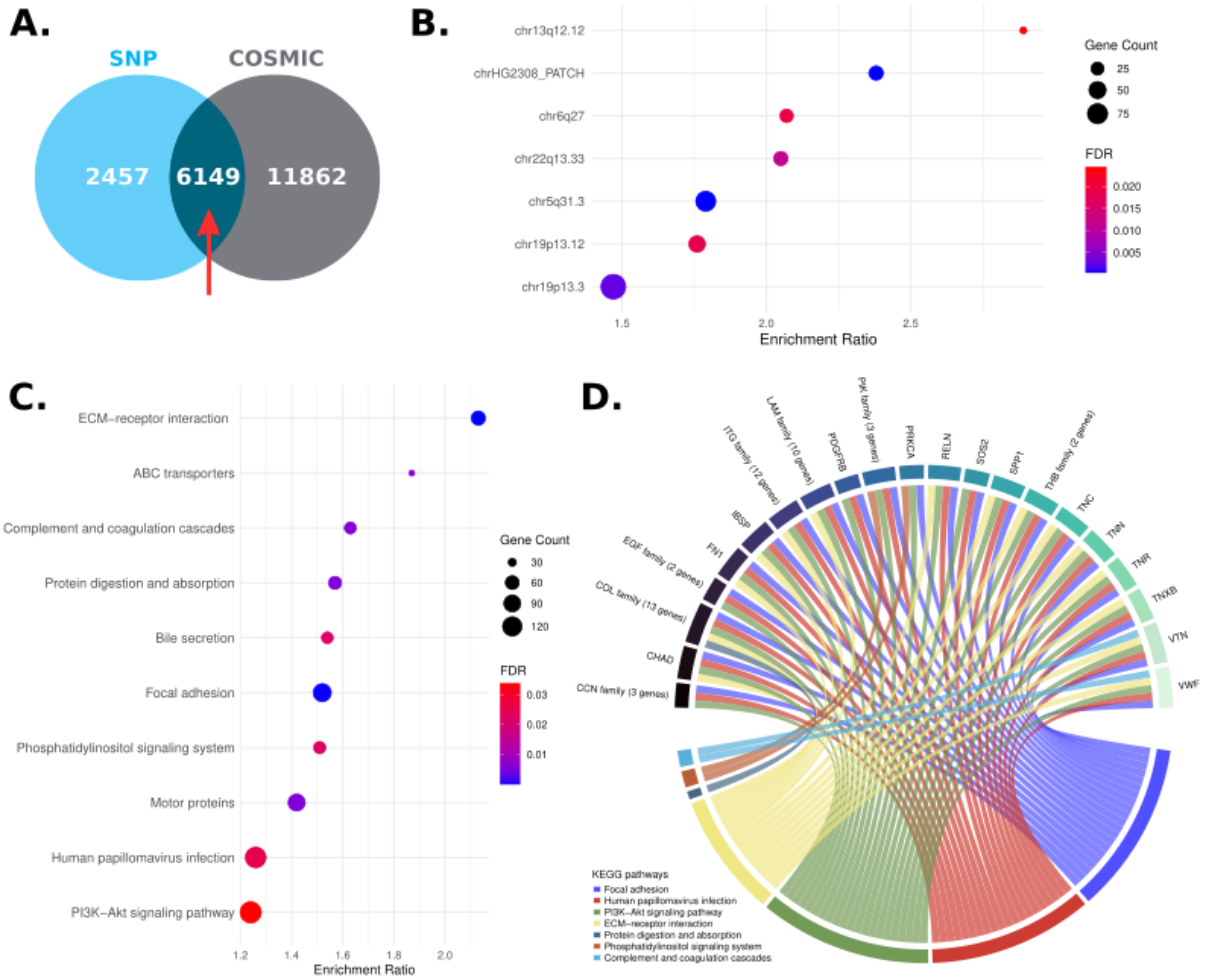


Figura 27. Análise de enriquecimento funcional (Over-Representation Analysis – ORA) de SNPs em sítios compartilhados com mutações somáticas do câncer. (A) Diagrama de interseção entre genes contendo variantes germinativas e somáticas. O conjunto de dados COSMIC apresentou 18.011 genes com mutações somáticas, enquanto o conjunto SNPs revelou 8.606 genes contendo variantes germinativas de alta frequência

alélica (AF ≥ 1%), resultando em uma interseção de 6.149 genes. (B) Análise de enriquecimento de bandas citogenéticas. O eixo Y representa as regiões genômicas significativamente enriquecidas (FDR < 0,05). O eixo X indica a razão de enriquecimento. O tamanho dos pontos reflete a contagem de genes em cada região, enquanto a escala de cores representa o nível de significância estatística, variando de azul (mais significativo) a vermelho (menos significativo). (C) Análise de enriquecimento de vias KEGG, destacando as vias biologicamente relevantes e significativamente enriquecidas (FDR < 0,05). (D) Diagrama de interconectividade entre genes/famílias multigênicas e vias enriquecidas. O gráfico circular exibe genes enriquecidos compartilhados entre três ou mais vias KEGG. As cores representam as diferentes vias, e as conexões indicam os genes presentes em múltiplas vias.

A análise de vias KEGG indicou um enriquecimento significativo em múltiplos processos biológicos (**Figura 27.C**). As vias *ECM-receptor interaction*, *PI3K-Akt signaling pathway, Human papillomavirus infection* e *Focal adhesion* foram as mais interconectadas, compartilhando um grande número de genes enriquecidos (**Figura 27.D**). Essas vias estão envolvidas em regulação da adesão celular, remodelação da matriz extracelular, sinalização proliferativa e interação tumoral com o microambiente—processos classicamente reconhecidos como essenciais para a progressão do câncer (Hanahan et al., 2000). O enriquecimento foi impulsionado por genes pertencentes a famílias multigênicas, como COL (colágenos), ITG (integrinas), LAM (lamininas), TNC (tenascinas) e PIK (família PI3K) presentes em pelo menos três vias KEGG enriquecidas. Esses genes são diretamente envolvidos na adesão célula-célula, comunicação intercelular e transdução de sinais pró-oncogênicos, influenciando a progressão tumoral e a plasticidade celular (Levental *et al.*, 2009; Banerjee *et al.*, 2022).

Além dessas, outras vias significativamente enriquecidas incluem *ABC transporters, Complement and coagulation cascades, Inositol phosphate metabolism, Protein digestion and absorption, Bile secretion, Phosphatidylinositol signaling system* e *Motor proteins*, muitas das quais estão associadas a processos metabólicos e homeostáticos.

5. DISCUSSÃO

Os genomas são tradicionalmente representados como um elemento estático: um único genoma-referência para cada espécie, de maneira análoga aos holótipos. Entretanto, e de maneira análoga ao que ocorre com os holótipos em relação à variação fenotípica total da espécie que representam, os dados de um único genoma certamente não capturam as suas diversas variações no tempo e espaço. Na dimensão temporal, os genomas se alteram nas populações ao longo das gerações de acordo com os princípios gerais da genética de populações, tais como seleção, migração e deriva genética. Entretanto, as mutações são o componente que fornece o substrato molecular para que todos esses fatores possam existir. Outra dimensão de variação genômica ocorre nos corpos dos indivíduos multicelulares, uma vez que mutações somáticas e germinativas possuem diferenças substanciais quanto ao seu papel evolutivo e em doenças com componente genético. Outra importante variação espacial nos genomas se dá em suas próprias coordenadas genômicas, as quais podem mutar diferencialmente em função de diversos componentes internos e externos. Atualmente, o genoma humano é o único arcabouço disponível para avaliar, de maneira coordenada e coletiva, como os diferentes tipos de variantes genéticas se distribuem nessas dimensões espaço-temporais.

Dessa forma, compreender a dinâmica mutacional requer não apenas a análise das forças evolutivas que atuam ao longo do tempo, mas também a investigação de como diferentes tipos de variantes se distribuem e interagem no genoma. A integração de múltiplas fontes de dados permite ampliar essa perspectiva, possibilitando a caracterização dos padrões mutacionais em larga escala e a identificação de mecanismos subjacentes à sua ocorrência. Os bancos de dados públicos COSMIC, ClinVar e HGDP forneceram os conjuntos genômicos necessários para o desenvolvimento deste trabalho. Essas fontes possibilitaram a integração em larga escala de informações sobre SNVs de diferentes classificações — germinativas e somáticas, neutras e patogênicas — viabilizando as análises dos padrões de distribuição e coocorrência dessas variantes no genoma humano.

Alguns conjuntos de variantes exigiram apenas etapas de filtragem para obtenção do perfil desejado, uma vez que integram bancos de dados curados, como ClinVar e COSMIC. No caso das SNVs polimórficas provenientes do HGDP, que refletem a diversidade genética global, adotamos critérios restritivos para selecionar variantes de alta frequência alélica (AF $\geq$ 1%) e com consequência funcional esperada neutra. O HGDP disponibiliza dados sobre a

diversidade genética humana de populações isoladas, ou seja, grupos com contato histórico limitado com outras populações. Nosso objetivo foi identificar variantes frequentes na população global, garantindo que estivessem presentes com essa mesma frequência alélica em pelo menos duas superpopulações isoladas. Essa abordagem visou minimizar a influência de fatores evolutivos como deriva genética e efeito fundador, que podem levar à fixação de variantes deletérias em uma população. Dado que o HGDP já categoriza indivíduos em superpopulações, poderíamos simplesmente ter adotado essa classificação prévia. No entanto, optamos por uma abordagem mais criteriosa, realizando uma análise exploratória da estrutura genética para confirmar se as populações de fato apresentavam baixo grau de miscigenação.

Entre as metodologias amplamente utilizadas para caracterizar a estrutura genética de populações humanas, destacam-se PCA, UMAP e a abordagem combinada PCA-UMAP (Sakaue *et al*., 2020, Diaz-Papkovich *et al.,* 2021). A escolha do método deve considerar a pergunta de pesquisa a fim de garantir inferências robustas. O PCA é uma abordagem clássica baseada na decomposição da matriz de variância-covariância dos dados genéticos, permitindo capturar as principais direções de variação genética. É amplamente utilizado para visualizar relações entre indivíduos e inferir ancestralidade, pois reflete gradientes contínuos de mistura genética. No entanto, o PCA pode ter limitações na captura de relações não lineares, especialmente ao analisar populações isoladas (Elhaik*,* 2022). O UMAP, por sua vez, é um método não linear projetado para preservar relações locais entre as amostras, projetando dados de alta dimensionalidade em um espaço reduzido de duas dimensões. Sua principal vantagem está na capacidade de detectar agrupamentos finos, permitindo a identificação de subestruturas populacionais ocultas. No entanto, quando aplicado diretamente a dados genéticos, o UMAP pode apresentar sensibilidade à densidade local, resultando em exagero das distâncias entre populações e distorções na estrutura genética global. Diante dessas considerações, a abordagem PCA-UMAP para redução de dimensionalidade e análise exploratória da estrutura genética surge como uma solução intermediária, combinando a estabilidade do PCA com a flexibilidade do UMAP (Sakaue *et al.*, 2020). Primeiro, o PCA é aplicado para reduzir a dimensionalidade e remover ruídos dos dados genéticos, preservando apenas os PCs mais informativos. Essa etapa minimiza o impacto de variações raras e evita que o UMAP amplifique diferenças populacionais que possam ser estatisticamente irrelevantes. Em seguida, o UMAP é aplicado sobre os PCs mais informativos, proporcionando uma visualização mais clara das relações genéticas, sem introduzir padrões artificiais de variação.

Para populações com baixa miscigenação, como os dados do HGDP, o PCA-UMAP pode ser suficiente para revelar uma estrutura genética mais sutil, embora não quantifique diretamente a ancestralidade. Em contrapartida, para populações altamente miscigenadas, é necessário um método de inferência de ancestralidade mais robusto, como ADMIXTURE ou abordagens baseadas em haplótipos. Assim, quando o objetivo é visualizar padrões de agrupamento populacional, PCA seguido de UMAP pode ser suficiente. No entanto, se a intenção for quantificar e inferir ancestralidade individual, é imprescindível o uso de análises mais detalhadas (Sakaue *et al.*, 2020). No presente estudo, a metodologia PCA-UMAP foi adotada para visualizar e validar a estrutura global das superpopulações previamente definidas pelo HGDP, além de identificar indivíduos que melhor se agrupam com seus pares populacionais e estabelecer um limiar robusto para a seleção de SNPs. O critério que utilizamos para a seleção de SNVs polimórficos considerou frequência alélica mínima de 1% em pelo menos duas populações isoladas, visando minimizar o impacto de mecanismos evolutivos como efeitos fundadores e deriva genética. Assim, a escolha do PCA-UMAP mostrou-se mais adequada do que o PCA isolado, pois permitiu capturar relações genéticas complexas que não seriam evidentes em um espaço puramente linear, demonstrando-se uma ferramenta eficaz para refinar a análise da diversidade genética global.

A redução significativa no número de variantes observada após o processamento e a filtragem dos arquivos VCF de todos os conjuntos de dados foi consistente com as expectativas e ocorreu em todos os conjuntos analisados. Para o conjunto de dados WG, as reduções variaram entre aproximadamente 33% a 93%, enquanto, para o conjunto de dados CDS, as reduções foram mais acentuadas, sempre superiores a 90% e chegando a 99% em alguns casos. Esses resultados refletem tanto a eficiência das etapas de filtragem quanto as diferenças intrínsecas entre as regiões codificadoras e não codificadoras do genoma.

Regiões não codificadoras usualmente apresentam maior tolerância à acumulação de SNVs, enquanto regiões codificadoras estão sujeitas a uma pressão seletiva mais intensa devido à importância funcional dos genes para a viabilidade celular e reprodutiva (Karczewski *et al.*, 2020; Monroe *et al.*, 2022). Assim, a alta proporção de variantes em regiões intrônicas observada nos conjuntos COSMIC, SNPs e Raras reflete a menor pressão seletiva nessas regiões, permitindo o acúmulo de variantes neutras ou de baixo impacto funcional. Por outro lado, o enriquecimento de SNVs em regiões exônicas nos conjuntos Benign e Pathogenic, do ClinVar, está diretamente relacionado à relevância funcional dessas variantes, frequentemente associadas às condições clínicas. Além disso, os padrões de localização genômica observados

também refletem os vieses dos métodos de construção dos conjuntos de dados. Os conjuntos COSMIC, SNPs e Raras derivam de bancos que reportam variantes de diferentes projetos de sequenciamento, abrangendo genomas de células normais e tumorais, representando de forma mais ampla o genoma completo. Já os conjuntos Benign e Pathogenic, do ClinVar, resultam de dados observados e curados em contextos clínicos que priorizam variantes em regiões codificadoras com maior relevância para a saúde humana estudadas na genética médica clássica (doenças mono e oligogênicas).

A aplicação do filtro para regiões CDS foi essencial para reduzir vieses e uniformizar a distribuição genômica das SNVs entre os conjuntos. Esse ajuste possibilitou análises comparativas mais robustas, evitando que vieses na distribuição genômica superestimassem ou subestimassem os resultados estatísticos, uma vez que a posição genômica foi um parâmetro crucial das análises estatísticas realizadas.

Os resultados da razão transição/transversão (Ti/Tv) evidenciaram diferenças significativas entre os conjuntos de SNVs, refletindo suas características funcionais e evolutivas. Conjuntos associados a variantes benignas ou neutras, como SNP e Benign, apresentaram razões mais altas (3,35 e 4,38, respectivamente), indicando uma predominância de transições, que geralmente estão associadas à maior conservação genômica (Liu *et al.*, 2012; J. Wang *et al.*, 2015). Em contraste, conjuntos relacionados a variantes patogênicas, como Pathogenic, Rares e COSMIC, exibiram razões mais baixas (2,19, 2,87 e 1,71, respectivamente), sugerindo uma maior frequência de transversões, frequentemente associadas à alterações mais disruptivas na funcionalidade gênica (Guo *et al.*, 2017; Sun *et al.*, 2019). A razão Ti/Tv, portanto, serviu como um indicador útil para compreender os processos evolutivos e o impacto funcional das variantes, além de refletir a pressão seletiva diferencial entre variantes neutras e variantes com consequências fenotípicas deletérias (J. Wang *et al.*, 2015).

A caracterização funcional das SNVs revelou padrões distintos entre os conjuntos analisados que também refletem suas origens e consequências evolutivas. O conjunto Pathogenic apresentou a maior proporção de variantes de impacto elevado, consistente com mutações associadas a doenças de alta penetrância. O conjunto COSMIC também exibiu variantes de perda de função com alto impacto, frequentemente atribuídas a mutações *driver* no estabelecimento do câncer. Nos conjuntos COSMIC e Rares, a predominância de variantes *missense* também sugere um impacto funcional relacionado a fenótipos deletérios. Enquanto variantes associadas a doenças monogênicas tendem a causar alterações proteicas severas,

variantes relacionadas a doenças complexas, como o câncer, apresentam padrões mais heterogêneos, com efeitos moderados e multifatoriais (Matlak *et al.*, 2019). Em contraste, os conjuntos Benign e SNPs exibiram uma maior proporção de variantes *synonymous* de baixo impacto funcional, sugerindo que a maioria dessas variantes é funcionalmente neutra, com reduzido potencial de alteração proteica.

As variantes *missense* predominaram em todos os conjuntos, sendo a consequência funcional mais prevalente. A análise dessas variantes utilizando o AlphaMissense revelou padrões consistentes com as propriedades biológicas esperadas de cada conjunto de SNVs. O conjunto Pathogenic apresentou alta frequência de variantes *missense* com *scores* elevados de patogenicidade, refletindo sua associação com doenças de alta penetrância. Esse padrão é consistente com a presença predominante de variantes deletérias severas que comprometem a funcionalidade proteica essencial (Cheng *et al.*, 2023). Em contraste, o conjunto Benign apresentou predominantemente variantes com *scores* baixos, corroborando sua classificação como funcionalmente neutras, em conformidade com as anotações do ClinVar.

A distribuição bimodal observada no conjunto COSMIC reflete a complexidade das mutações somáticas no câncer. A coexistência de mutações *driver*, que desencadeiam e promovem o crescimento tumoral, e mutações *passenger*, que não possuem impacto direto na progressão do câncer, demonstra os processos evolutivos que moldam o genoma tumoral, como mutação acumulativa e seleção clonal. Esses resultados estão alinhados com estudos que destacam a heterogeneidade genômica como uma característica fundamental das células tumorais (Hess *et al.*, 2019).

Os conjuntos SNPs e Rares apresentaram perfis de patogenicidade consistentes com suas dinâmicas evolutivas. O conjunto SNPs, com predominância de variantes *likely benign*, reflete o papel das variantes comuns na manutenção da estabilidade funcional e genética das populações humanas. Já as variantes do conjunto Rares, mais heterogêneas, incluem tanto variantes *de novo* quanto aquelas sujeitas à seleção purificadora (Rentzsch *et al.*, 2019). A fração de variantes raras com altos *scores* de patogenicidade (~11,5%) reforça a relevância desse conjunto para o estudo de doenças genéticas raras e novas.

As variantes *missense*, embora representem alterações únicas na sequência de DNA, exibem uma ampla gama de impactos funcionais. A utilização do AlphaMissense foi essencial para identificar esses padrões e oferecer uma resolução funcional detalhada, especialmente em conjuntos complexos como o COSMIC.

Após a caracterização, todos os conjuntos de dados apresentaram algum grau de heterogeneidade em relação à localização genômica, às consequências funcionais e à penetrância das variantes. Essa heterogeneidade reflete as diferentes origens biológicas e os critérios de curadoria adotados por cada banco de dados utilizado. Contudo, cada conjunto demonstrou predominantemente um perfil específico, alinhado aos padrões inicialmente esperados, reforçando a qualidade das etapas de processamento e filtragem realizadas neste trabalho.

Por exemplo, variantes patogênicas apresentaram maior concentração em regiões exônicas, as quais estão usualmente associadas a impactos funcionais elevados, enquanto variantes benignas e SNPs exibiram distribuições mais equilibradas, consistentes com variantes neutras ou quase neutras. Essa separação clara entre os conjuntos de dados evidencia sua representatividade e adequação para os objetivos deste estudo, fornecendo uma base sólida para análises comparativas e estatísticas robustas.

Os resultados das análises estatísticas de coocorrência entre os diferentes conjuntos de SNVs forneceram evidências relevantes sobre os padrões de distribuição mutacional no genoma humano. Os modelos propostos neste estudo ajudaram a interpretar os cenários observados, destacando diferenças nos padrões mutacionais subjacentes às interações genômicas e potenciais mecanismos moleculares envolvidos.

O Modelo 1, que propõe a ocorrência de mutações em loci distintos, foi indicado por associações negativas e mostrou-se mais aplicável às seguintes comparações: *Rares vs. COSMIC, SNPs vs. Rares, SNPs vs. Pathogenic, Rares vs. Pathogenic, COSMIC vs. Benign* e *COSMIC vs. Pathogenic*. Para *Rares vs. COSMIC*, a influência do desequilíbrio de ligação (LD) foi evidente, com o padrão inicial de associação positiva para CDS sendo revertido após a filtragem por LD. A complexidade dos cenários mutacionais e a necessidade de considerar fatores estruturais, como correlações genômicas, ao interpretar padrões de coocorrência, são evidenciados. Na comparação *SNPs vs. Rares*, a ausência de sobreposição é explicada pela metodologia de filtragem adotada, que define os conjuntos como mutuamente excludentes com base em frequência alélica e bialelismo. Já para *SNPs vs. Pathogenic* e *Rares vs. Pathogenic*, a baixa ou inexistente interseção pode refletir diferenças nos mecanismos mutacionais subjacentes. Variantes patogênicas de alta penetrância são frequentemente sujeitas a forte seleção purificadora, enquanto SNPs e variantes raras neutras ou de baixo impacto podem se acumular em regiões menos críticas. Por fim, as comparações *COSMIC vs.*

Benign e *COSMIC vs. Pathogenic* podem refletir os diferentes mecanismos mutacionais associados às origens germinativa e somática, bem como às consequências funcionais divergentes, indicando que as pressões seletivas e os processos moleculares diferem significativamente entre os dois contextos.

O Modelo 2, que postula a ocorrência de mutações em loci compartilhados, mas com alelos distintos, foi mais adequado para explicar a comparação entre Pathogenic vs. Benign. Essa análise revelou 100% de dissimilaridade alélica, o que pode refletir diferentes pressões seletivas atuantes — onde alelos patogênicos de alta penetrância estão associadas a efeitos fenotípicos severos, enquanto alelos benignos, funcionalmente neutras, tendem a ser mais permissivos. Alternativamente, fatores metodológicos relacionados à curadoria do ClinVar podem ter contribuído para essa configuração.

O Modelo 3, que propõe a ocorrência de mutações neutras e patogênicas em *loci* compartilhados com alelos idênticos, mostrou-se o mais consistente para explicar comparações como *SNPs vs. Benign* e *Rares vs. Benign*. Essas associações apresentaram altas frequências de alelos idênticos, sugerindo uma possível relação funcional ou evolutiva entre os conjuntos. A sobreposição inesperada entre *SNPs vs. COSMIC*, também compatível com o Modelo 3, caracteriza-se por uma alta similaridade alélica (97,7%). Essa comparação destacou-se por ser a única a apresentar associação positiva, tanto para o conjunto de dados WG quanto para o conjunto CDS, envolvendo conjuntos de origens distintas (germinativa e somática). Esses resultados sugerem a existência de *hotspots* mutacionais, uma vez que variantes de diferentes origens e consequências funcionais compartilham *loci* específicos em frequências significativamente maiores que as esperadas. Tal padrão reforça a hipótese de que mecanismos moleculares comuns, ainda não identificados, podem estar direcionando a ocorrência dessas variantes em regiões genômicas específicas, o que representa um ponto de interesse para investigações futuras.

A abordagem adotada para estimar a frequência esperada de alelos idênticos baseou-se em duas premissas simplificadas: (1) a suposição de que todos os nucleotídeos possuem frequências iguais (~25%) no genoma humano e (2) a hipótese de que as substituições ocorrem com igual probabilidade entre todas as bases, sem considerar as diferenças entre transições e transversões. Embora essas simplificações ofereçam um modelo estatístico de referência, elas não refletem completamente os processos biológicos subjacentes à distribuição das variantes genômicas. Em termos reais, as probabilidades de substituição não

são equiprováveis, uma vez que as transições ocorrem com maior frequência do que as transversões (Liu *et al.*, 2012; Guo *et al.*, 2017). Além disso, mutações do tipo C>T e G>A são particularmente comuns, especialmente em regiões ricas em ilhas CpG, onde a desaminação espontânea da 5-metilcitosina representa um dos principais mecanismos de hotspot mutacional (Alexandrov *et al.*, 2020). No entanto, apesar dessas limitações, a análise da proporção de alelos idênticos demonstrou que mais de 97% das variantes eram iguais ou diferentes entre os conjuntos, e conforme esperado, o Teste Exato de Fisher indicou que a grande maioria das variantes apresentou uma frequência de alelos idênticos acima do esperado pelo acaso. A consistência nos achados reforça que, apesar de potenciais refinamentos metodológicos para incorporar taxas de mutação dependentes do contexto genômico, o modelo adotado para nosso estudo foi suficiente para avaliar os padrões de similaridade entre as variantes dos conjuntos analisados.

De maneira geral, nossos achados reforçam a importância de modelos teóricos na interpretação de padrões mutacionais no genoma humano, oferecendo uma base conceitual para compreender as origens genéticas e as consequências funcionais de variantes germinativas e somáticas. Além disso, destacamos também a relevância de abordagens metodológicas rigorosas, como a filtragem por LD e a distinção entre variantes em WG e CDS, para garantir inferências robustas e reduzir possíveis vieses. As associações observadas sugerem a presença de mecanismos mutacionais que desempenham um papel importante na configuração da arquitetura genômica, influenciando tanto a diversidade genética humana quanto a predisposição a doenças humanas. Em conjunto, nossos resultados destacam a complexidade dos processos evolutivos e funcionais que moldam o genoma, apontando para novas direções em estudos sobre *hotspots* mutacionais e mecanismos moleculares subjacentes.

A caracterização do subconjunto *SNPs vs. COSMIC* evidencia também sua relevância para investigações sobre as interações entre mutações germinativas e somáticas. Foi observada uma alta proporção de SNVs intrônicas no WG, além de uma predominância de mutações *synonymous* e *missense* nos subconjuntos de WG e CDS, refletindo a heterogeneidade funcional das variantes sobrepostas. As mutações *synonymous*, de impacto funcional baixo, podem representar tanto variantes germinativas neutras no contexto evolutivo como mutações *passenger* com impacto limitado em COSMIC. Por outro lado, variantes *missense* com baixa pontuação de patogenicidade no AlphaMissense sugerem que estas podem representar polimorfismos neutros no conjunto de SNPs e de mutações *passenger* no conjunto COSMIC.

Variantes *missense* com pontuações mais altas poderiam representar mutações com impacto funcional mais direto, alinhadas a fenótipos deletérios. A identificação de mutações *stop-gained*, *start-lost* e *stop-lost* no subconjunto *SNPs vs. COSMIC*, embora em menor proporção, sugere que essas variantes também possam atuar como mutações *driver* em genes relacionados ao câncer, contribuindo diretamente para processos oncogênicos. No contexto germinativo, essas mutações poderiam ocorrer em genes não essenciais ou redundantes, permitindo sua manutenção na população sem consequências fenotípicas significativas. Alternativamente, é possível que essas variantes exerçam efeitos cis-regulatórios indiretos, influenciando elementos regulatórios próximos e modulando padrões de expressão gênica, embora esses aspectos não tenham sido diretamente explorados no presente estudo. Outra alternativa é que mutações em outras regiões do genoma possam atuar de maneira compensatória aos efeitos deletérios das mutações com patogenicidade maior.

As assinaturas mutacionais observadas no *subconjunto* SNPs vs. COSMIC refletem processos que interligam o genoma germinativo e somático. A predominância de SBS1, SBS54, SBS5 no subconjunto WG (15,48%, 11,52% e 73,01%, respectivamente) destaca a relevância de processos endógenos na acumulação de mutações. A SBS1, amplamente associada à desaminação espontânea ou enzimática de 5-metilcitosina para timina, segue um padrão semelhante a um relógio biológico, onde o número de mutações aumenta com a idade (Alexandrov et al., 2020). A SBS54, inicialmente considerada um artefato de sequenciamento, foi identificada como uma assinatura mutacional relevante em contextos de deficiência no reparo de DNA (dMMR) e instabilidade de microssatélites (MSI) (Ji *et al.*, 2023; Farmanbar *et al.*, 2023). Sua presença significativa neste estudo reforça a hipótese de que a assinatura não seja meramente técnica, mas reflete padrões biológicos reais. Já a SBS5, cuja etiologia permanece desconhecida, é caracterizada por sua onipresença em quase todos os tipos celulares e cânceres, sendo também associada a mutações somáticas pós-zigóticas precoces (Yamamoto *et al.,* 2024). Estudos recentes demonstraram uma ligação entre SBS5 e mutações no gene TP53, conhecido como "guardião do genoma", implicando seu papel na progressão tumoral e instabilidade genômica (Yaacov *et al.*, 2024).

No subconjunto CDS, além das assinaturas SBS1, SBS54 e SBS5 (32,27%, 15,3% e 12,87%, respectivamente), identificamos também duas assinaturas adicionais: SBS87 (21,51%) e SBS6 (13,04%). A SBS87, tipicamente associada ao uso de tiopurinas no tratamento da leucemia linfoblástica aguda (ALL), também foi observada em posições compartilhadas entre variantes germinativas e somáticas (Wang *et al.*, 2024, Li *et al.,* 2020; van der Ham *et al.*,

2024). Um estudo recente relatou a presença de SBS87 e SBS1 em variantes somáticas que compartilham loci com mutações germinativas (Wang *et al.*, 2024), reforçando a hipótese de que processos de reparo de DNA e danos específicos em regiões CpG podem influenciar ambos os contextos genômicos. Embora a SBS87 seja caracteristicamente somática e associada a quimioterápicos, sua presença em loci compartilhados no estudo de Wang et al. (2024) sugere que mecanismos como erros de replicação, deficiências no reparo de DNA, influência do timing de replicação, estabilidade estrutural do genoma e possíveis exposições ambientais podem contribuir para a ocorrência de mutações germinativas.

A SBS6 é amplamente associada a MSI e tumores relacionados à síndrome de Lynch. Estudos mostram que esta assinatura é mediada por defeitos epigenéticos, como a hipermetilação do gene MLH1, e fatores ambientais que exacerbam o estresse genômico (Jayakrishnan *et al.*, 2024; Pužar Dominkuš *et al.*, 2023). A sobreposição de SNPs e mutações somáticas associadas à SBS6 indica que regiões instáveis, como microssatélites, poderiam servir como *hotspots* de mutação, influenciados por predisposições genéticas e alterações estruturais no genoma.

Os resultados deste estudo demonstraram que ilhas CpG coocorrem nas mesmas coordenadas genômicas do que o subconjutno *SNPs vs. COSMIC* aproximadamente 7 vezes mais frequentemente do que o esperado ao acaso, explicando cerca de 5% das variantes compartilhadas. Microssatélites, por sua vez, apresentaram uma coocorrência 4 vezes maior do que o esperado, embora expliquem apenas 0,2% das mutações. Portanto, as ilhas CpG e os microssatélites podem atuar como locais de fragilidade genômica germinativas-somáticas, predispondo à ocorrência de mutações subsequentes, como as associadas à SBS1 e SBS6. Entretanto, a coocorrência com ilhas CpG e regiões de microssatélite são claramente insuficientes para explicar todas as coocorrências observadas, sugerindo que mecanismos adicionais possivelmente contribuem para o fenômeno que observamos.

As assinaturas SBS1 e SBS5, predominantes no subconjunto SNPs vs. COSMIC para WG e CDS, foram recentemente relacionadas ao timing de replicação (RT) celular. A SBS1, frequentemente encontrada em regiões de replicação tardia (LRR) em células normais, migra para regiões de replicação precoce (ERR) no câncer, refletindo alterações genômicas e epigenéticas associadas à transformação maligna. Essa mudança pode ser atribuída a diversos fatores, incluindo: i) redistribuição de ilhas CpG metiladas, que se concentram em ERR em células cancerosas; ii) deficiência no reparo por excisão de bases (BER), cuja eficiência é

reduzida devido à alta carga mutacional e ciclos rápidos de replicação em tumores; iii) seleção positiva de células ERR durante a evolução tumoral, com eliminação de células associadas a LRR; e iv) interação de mutações SBS1 com fatores de transcrição, como CEBPβ, que inibem o reparo em regiões específicas. Esses fatores promovem uma reorganização do padrão mutacional, resultando na predominância de SBS1 em ERR no câncer. Em contraste, a SBS5 mantém sua associação com ERR tanto em células normais quanto cancerosas, sugerindo uma assinatura estável, menos influenciada por alterações malignas (Yaacov *et al.*, 2023). A estabilidade observada na SBS5, aliada à sua abundância no WG e CDS, destaca seu potencial como marcador robusto de processos intrínsecos compartilhados entre genomas germinativos e somáticos. Por outro lado, a flexibilidade da SBS1 em contextos distintos reflete a influência de alterações específicas do microambiente tumoral e processos evolutivos associados ao câncer.

Biologicamente, a sobreposição significativa entre SNPs germinativos e mutações somáticas relacionadas ao câncer sugere a presença de mecanismos compartilhados que afetam ambos os genomas. A identificação de assinaturas como SBS1, SBS5, SBS87, SBS54 e SBS6 em *loci* compartilhados reforça a hipótese de que *hotspots* mutacionais são moldados por interações entre características genômicas intrínsecas, como ilhas CpG e microssatélites, e fatores externos ou defeitos no reparo de DNA. O papel da SBS5 no contexto de variantes germinativas é especialmente relevante, considerando que essa assinatura está associada a processos celulares endógenos ainda pouco compreendidos. Investigações futuras que explorem o mecanismo molecular subjacente à SBS5 podem esclarecer sua contribuição para a formação de *hotspots* mutacionais e para a interação entre genomas germinativos e somáticos.

Os resultados da análise de enriquecimento funcional forneceram percepções complementares sobre os padrões mutacionais observados no subconjunto *SNPs vs. COSMIC*. Embora os mecanismos responsáveis pela sobreposição mutacional entre SNPs e COSMIC ainda sejam desconhecidos, a presença de *hotspots* compartilhados no braço curto do cromossomo 19 e em genes da via PI3K/AKT/mTOR levanta questões importantes. Estudos prévios indicam que variantes germinativas na região 19p13.3 podem aumentar a taxa de mutações somáticas no gene PTEN, um dos principais supressores tumorais envolvidos na regulação negativa da via PI3K/AKT/mTOR (Carter *et al.*, 2017). A ativação constitutiva dessa via está associada a proliferação celular descontrolada, inibição da apoptose e progressão tumoral. A presença de variantes germinativas em genes dessa via já foram relacionados à susceptibilidade familiar

ao câncer de pulmão (Lin *et al.*, 2020). No entanto, os mecanismos responsáveis pela ocorrência de SNPs altamente frequentes nesses mesmos *loci* genômicos durante a fase germinativa ainda não foram explorados (Carter *et al.*, 2017, Meyerson *et al*., 2020, Wang *et al.*, 2024)

Entre os genes enriquecidos na via PI3K-AKT, identificamos MTOR, RPTOR, MLST8 e STK11, que modulam diretamente a atividade da via (Panwar *et al*., 2023). Além disso, genes da família G-protein subunit beta/gamma (GNB3, GNB5, GNG8 e GNGT2) foram detectados, sugerindo uma possível regulação da via mTOR, embora com menor impacto do que genes previamente descritos, como GNA11 e GβL (Carter *et al*., 2017; Panwar *et al*., 2023). Também observamos enriquecimento nos genes TSC2 e TP53, reguladores negativos da via mTOR. O complexo TSC1/TSC2 recebe sinais de múltiplas cascatas, incluindo PI3K/AKT, MAPK/ERK e TNF-α, para controlar a ativação de mTORC1. Já TP53, conhecido como o "guardião do genoma", inibe mTORC1 via ativação de AMPK, PTEN e TSC2, reduzindo a proliferação celular descontrolada (Panwar *et al*., 2023). Outros genes enriquecidos incluem PIK3R1 e PIK3R2, responsáveis pela regulação da ativação da PI3K em resposta a fatores de crescimento e oncogenes (Glaviano *et al*., 2023). Além disso, proto-oncogenes frequentemente mutados em câncer, como KIT, MET, MYC e RET, foram identificados, assim como fatores de crescimento e receptores EGFR, FGFR, IGF1R, PDGFRB, GHR e HGF.

As vias ECM-receptor interaction e Focal adhesion, também enriquecidas, estão relacionadas à remodelação tecidual e migração celular em tumores. A matriz extracelular (ECM) regula a adesão celular e sinalização, afetando diretamente a interação tumoral com o microambiente. A via Focal adhesion, por sua vez, conecta a ECM ao citoesqueleto, modulando a adesão celular, sinalização mecânica e resistência ao estresse, processos frequentemente desregulados no câncer (Chen *et al*., 2024). Outro achado relevante foi o enriquecimento da via Human papillomavirus (HPV) infection, associada à ativação da via PI3K/AKT/mTOR, promovendo proliferação celular, evasão da apoptose e resistência terapêutica (Letafati *et al.*, 2024; Wang *et al.*, 2024). A associação dessa via sugere um possível papel da sinalização PI3K/AKT/mTOR na predisposição a mutações somáticas induzidas por infecção viral.

Apesar dos resultados apresentados não elucidarem diretamente a causa das mutações nos sítios compartilhados entre SNPs e SNVs somáticas, eles evidenciam um fenômeno intrigante que pode estar associado a mecanismos específicos de *hotspots* genômicos ainda inexplorados

e oferecem caminhos para investigações futuras. A correlação entre assinaturas mutacionais (SBS1, SBS5, SBS87, SBS54 e SBS6), vias enriquecidas (PI3K/AKT/mTOR, ECM-receptor interaction, Focal adhesion e Human papillomavirus infection) e regiões genômicas específicas (13q12.12, 19p13.3, 19p13.12, 5q31.3, 22q13.33, 6q27 e o segmento chrHG2308_PATCH) sugere que fatores estruturais, restrições genômicas e pressão seletiva podem estar envolvidos nesse padrão. Isso reforça a necessidade de investigações adicionais para compreender como SNPs germinativos e mutações somáticas frequentemente observadas no câncer compartilham *loci* mais do que o esperado aleatoriamente e quais mecanismos podem mediar esse fenômeno.

Nossos achados no campo dos *hotspots* mutacionais podem ser utilizados para otimizar o uso de variantes recorrentes como biomarcadores em diagnóstico, prognóstico e terapias direcionadas, especialmente no câncer. A sobreposição entre variantes germinativas e somáticas sugere que determinadas regiões genômicas são intrinsecamente instáveis, independentemente da linhagem celular, possibilitando novas estratégias de predição de risco e medicina personalizada. A integração dessas informações com inteligência artificial e análise de dados ômicos pode melhorar a identificação de variantes críticas, permitindo a estratificação mais precisa de pacientes e o desenvolvimento de terapias direcionadas, como inibidores específicos e edição genética via CRISPR. Essa abordagem tem o potencial de revolucionar o monitoramento precoce e a intervenção clínica, promovendo tratamentos mais eficazes e personalizados para doenças genéticas complexas. A longo prazo, esses avanços podem redefinir a forma como as doenças genéticas são diagnosticadas e tratadas, aproximando a medicina de um modelo cada vez mais preditivo e personalizado.

6. CONCLUSÕES

A vasta disponibilidade de dados genômicos humanos têm impulsionado significativamente o avanço do nosso entendimento sobre mutações de diferentes contextos biológicos e evolutivos. O genoma humano é um dos recursos mais completos e bem curados para a análise dos padrões de ocorrência de mutações de distintas categorias, como germinativas e somáticas, polimórficas e raras, patogênicas e benignas. Contudo, apesar da qualidade e riqueza desses dados, análises integrativas que sistematicamente caracterizem essas mutações de forma comparativa, revelando os mecanismos subjacentes aos seus padrões de distribuição, permanecem escassas.

Este estudo, ao caracterizar comparativamente cinco conjuntos distintos de SNVs, propõe e contribui para a ampliação de pesquisas que abordem a lacuna existente em análises integrativas. As propriedades biológicas e os padrões genômicos identificados nos conjuntos analisados mostraram-se consistentes com o conhecimento disponível na literatura, reforçando a confiabilidade dos dados utilizados e destacando sua relevância para futuras investigações científicas.

Pela primeira vez, foram propostos, avaliados e estabelecidos modelos que refletem o cenário comparativo de distribuição mutacional entre diferentes classes de mutações. Entre os achados mais relevantes, destaca-se a associação positiva identificada entre SNVs germinativas polimórficas e somáticas relacionadas ao câncer, representando um avanço significativo na compreensão dos padrões mutacionais no genoma humano. Essa associação permaneceu significativa tanto para CDS quanto para WG, com a maioria das variantes sobrepostas compartilhando o mesmo alelo alternativo. Além disso, a análise funcional dessas variantes indica que, em sua maioria, possuem impacto funcional menor em relação às variantes *driver* típicas do câncer, sugerindo que esses alelos compartilhados são mais provavelmente mutações *passenger* ou *drivers* fracos, ao invés de *drivers* primários no câncer. Acredita-se que essas SNVs sobrepostas possam apresentar uma natureza dual: enquanto podem ser neutras no contexto germinativo, podem adquirir potencial patogênico como mutações somáticas em células cancerosas, dependendo do contexto genômico. Isso inclui fatores como a acumulação de mutações adicionais ao longo do tempo, interações epistáticas e alterações em genes *driver*.

As ilhas CpG, *hotspots* mutacionais conhecidos, e regiões de fragilidade genômica, como microssatélites, explicam apenas uma pequena fração dessas mutações sobrepostas. Esses

achados sugerem que um mecanismo compartilhado, ainda desconhecido, pode estar impulsionando a coocorrência de SNPs e mutações recorrentes no câncer, potencialmente contribuindo para as assinaturas mutacionais SBS1, SBS5, SBS6, SBS54 e SBS87. Interessantemente, esse mecanismo parece estar associado a ocorrência de mutações germinativas e somáticas em genes do braço curto do cromossomo 19 e da via PI3K/Akt/mTOR.

7. PERSPECTIVAS

Este estudo fornece evidências acerca da sobreposição entre variantes germinativas e mutações somáticas em regiões específicas do genoma humano, sugerindo a existência de *hotspots* mutacionais compartilhados ainda não descritos anteriormente. Essa observação levanta novas hipóteses sobre a organização funcional do genoma e a predisposição intrínseca de determinadas regiões à mutação, abrindo caminho para investigações futuras sobre os mecanismos moleculares que regulam essa instabilidade.

No contexto da medicina de precisão, a identificação de regiões recorrentes de mutação pode aprimorar a interpretação clínica de variantes genéticas, auxiliando na distinção entre mutações benignas e patogênicas. Esses achados também podem informar o desenvolvimento de modelos preditivos de risco genético mais sensíveis e específicos, além de apoiar estratégias de triagem genômica populacional.

Além disso, os dados gerados neste estudo oferecem uma base conceitual e metodológica para aplicações futuras em terapia gênica, especialmente com o uso de ferramentas de edição como o CRISPR-Cas9. Ao compreender melhor os contextos genômicos mais vulneráveis a mutações, torna-se possível minimizar efeitos *off-target* e melhorar a segurança de abordagens terapêuticas baseadas em edição genômica.

Por fim, este trabalho reforça a importância da ciência básica na construção de conhecimento sólido, capaz de sustentar avanços tecnológicos e aplicações clínicas, e destaca a necessidade de aprofundar a integração entre dados populacionais, funcionais e evolutivos para uma compreensão mais abrangente da genética humana.

Desta forma, para elucidar os mecanismos subjacentes aos *hotspots* mutacionais compartilhados entre variantes germinativas e somáticas, abordagens experimentais direcionadas devem investigar quais fatores biológicos e estruturais tornam essas regiões vulneráveis a mutações recorrentes nos dois cenários. Como continuidade, sugere-se a investigação direcionada aos mecanismos mutacionais responsáveis por SNPs e SNVs somáticas recorrentes no câncer ocorrerem nas mesmas posições genômicas e compartilharem o mesmo alelo por meio de abordagens experimentais integradas, combinando análises computacionais, genômicas funcionais e validações *in vitro/in vivo*.

Dando continuidade, propomos as seguintes abordagens:

- **Epigenéticas e arquitetura da cromatina**

Determinar se a acessibilidade da cromatina e as modificações epigenéticas influenciam a vulnerabilidade dessas regiões a mutações.

1. ATAC-seq (*Assay for Transposase-Accessible Chromatin-sequencing*) - Avaliar se regiões abertas da cromatina são mais propensas a estas mutações "germinativo-somáticas".
2. ChIP-seq (*Chromatin Immunoprecipitation-sequencing*) - Investigar modificações de histonas (ex.: H3K4me3, H3K27ac) para determinar se estes *loci* mutacionais comuns coincidem com regiões ativamente reguladas.
3. Hi-C - Mapear a organização tridimensional do genoma e verificar se os *hotspots* compartilham interações com regiões regulatórias distantes.

- **Efeito das variantes na expressão gênica**

Determinar se variantes em *hotspots* germinativo-somáticos influenciam a expressão gênica.

1. eQTL (*Expression Quantitative Trait Loci*) - Integrar dados do projeto GTEx (Genotype-Tissue Expression Project) para identificar se SNPs associados aos *hotspots* mutacionais germinativo-somáticos afetam a expressão gênica.

- **Mecanismos de reparo de DNA e propensão à mutação**

Avaliar se as regiões compartilhadas entre SNPs e COSMIC estão associadas a falhas nos mecanismos de reparo de DNA.

1. Testes de sensibilidade a estresse genotóxico (Radiação UV, Agentes Alquilantes) - Expor células a danos no DNA e avaliar taxas de mutação nas regiões *hotspots*.
2. Edição de bases (base editing) via CRISPR - Introduzir mutações pontuais nos *hotspots* para testar se são mais propensos a escape do reparo de DNA.

8. REFERÊNCIAS BIBLIOGRÁFICAS

1. Abdullah, T., Ahmet, A. (2020). Extracting Insights: A Data Centre Architecture Approach in Million Genome Era. In: Hameurlain, A., Tjoa, A.M. (eds) Transactions on Large-Scale Data- and Knowledge-Centered Systems XLVI. Lecture Notes in Computer Science(), vol 12410. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-62386-2_1

2. Alan F. Wright, "Genetic variation: Polymorphisms and mutations" in Encyclopedia of Life Sciences, Hoboken, NJ, USA:Wiley, 2005. https://doi.org/10.1038/npg.els.0005005

3. Alekseenko, I.V., Kuzmich, A.I., Pleshkan, V.V. et al. The cause of cancer mutations: Improvable bad life or inevitable stochastic replication errors?. Mol Biol 50, 799–811 (2016). https://doi.org/10.1134/S0026893316060030

4. Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., PCAWG Mutational Signatures Working Group, Getz, G., … PCAWG Consortium (2020). The repertoire of mutational signatures in human cancer. Nature, 578(7793), 94–101. https://doi.org/10.1038/s41586-020-1943-3

5. Akdemir, K.C., Le, V.T., Kim, J.M. et al. Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. Nat Genet 52, 1178–1188 (2020). https://doi.org/10.1038/s41588-020-0708-0

6. Ambalavanan, A., Chang, L., Choi, J. et al. Human milk oligosaccharides are associated with maternal genetics and respiratory health of human milk-fed children. Nat Commun 15, 7735 (2024). https://doi.org/10.1038/s41467-024-51743-6

7. Arthur, T.D., Nguyen, J.P., D'Antonio-Chronowska, A. et al. Complex regulatory networks influence pluripotent cell state transitions in human iPSCs. Nat Commun 15, 1664 (2024). https://doi.org/10.1038/s41467-024-45506-6

8. Baca, S.C., Singler, C., Zacharia, S. et al. Genetic determinants of chromatin reveal prostate cancer risk mediated by context-dependent gene regulation. Nat Genet 54, 1364–1375 (2022). https://doi.org/10.1038/s41588-022-01168-y

9. Bamford, S., Dawson, E., Forbes, S. et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 91, 355–358 (2004). https://doi.org/10.1038/sj.bjc.6601894

10. Banerjee, S., Lo, W. C., Majumder, P., Roy, D., Ghorai, M., Shaikh, N. K., Kant, N., Shekhawat, M. S., Gadekar, V. S., Ghosh, S., Bursal, E., Alrumaihi, F., Dubey, N. K., Kumar, S., Iqbal, D., Alturaiki, W., Upadhye, V. J., Jha, N. K., Dey, A., & Gundamaraju, R. (2022). Multiple roles for basement membrane

proteins in cancer progression and EMT. European journal of cell biology, 101(2), 151220. https://doi.org/10.1016/j.ejcb.2022.151220

11. Bhattacharya, R., Rose, P. W., Burley, S. K., & Prlić, A. (2017). Impact of genetic variation on three dimensional structure and function of proteins. PloS one, 12(3), e0171355. https://doi.org/10.1371/journal.pone.0171355

12. Barbosa, P., Savisaar, R., Carmo-Fonseca, M., & Fonseca, A. (2022). Computational prediction of human deep intronic variation. GigaScience, 12, giad085. https://doi.org/10.1093/gigascience/giad085

13. Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J. F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., Durbin, R., … Tyler-Smith, C. (2020). Insights into human genetic variation and population history from 929 diverse genomes. Science (New York, N.Y.), 367(6484), eaay5012. https://doi.org/10.1126/science.aay5012

14. Branco, I., & Choupina, A. (2021). Bioinformatics: new tools and applications in life science and personalized medicine. Applied microbiology and biotechnology, 105(3), 937–951. https://doi.org/10.1007/s00253-020-11056-2

15. Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. PLoS computational biology, 8(12), e1002822. https://doi.org/10.1371/journal.pcbi.1002822

16. Cagan, A., Baez-Ortega, A., Brzozowska, N., Abascal, F., Coorens, T. H. H., Sanders, M. A., Lawson, A. R. J., Harvey, L. M. R., Bhosle, S., Jones, D., Alcantara, R. E., Butler, T. M., Hooks, Y., Roberts, K., Anderson, E., Lunn, S., Flach, E., Spiro, S., Januszczak, I., Wrigglesworth, E., … Martincorena, I. (2022). Somatic mutation rates scale with lifespan across mammals. Nature, 604(7906), 517–524. https://doi.org/10.1038/s41586-022-04618-z

17. Cain, J. A., Montibus, B., & Oakey, R. J. (2022). Intragenic CpG Islands and Their Impact on Gene Regulation. Frontiers in cell and developmental biology, 10, 832348. https://doi.org/10.3389/fcell.2022.832348

18. Carter, H., Marty, R., Hofree, M., Gross, A. M., Jensen, J., Fisch, K. M., Wu, X., DeBoever, C., Van Nostrand, E. L., Song, Y., Wheeler, E., Kreisberg, J. F., Lippman, S. M., Yeo, G. W., Gutkind, J. S., & Ideker, T. (2017). Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. Cancer discovery, 7(4), 410–423. https://doi.org/10.1158/2159-8290.CD-16-1045

19. Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4, 7. https://doi.org/10.1186/s13742-015-0047-8

20. Chen, K., Zhang, J., Guo, Z. et al. Loss of 5-hydroxymethylcytosine is linked to gene body hypermethylation in kidney cancer. Cell Res 26, 103–118 (2016). https://doi.org/10.1038/cr.2015.150

21. Chen, Y., Chang, L., Hu, L., Yan, C., Dai, L., Shelat, V. G., Yarmohammadi, H., & Sun, J. (2024). Identification of a lactylation-related gene signature to characterize subtypes of hepatocellular carcinoma using bulk sequencing data. Journal of gastrointestinal oncology, 15(4), 1636–1646. https://doi.org/10.21037/jgo-24-405

22. Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P., & Avsec, Ž. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science (New York, N.Y.), 381(6664), eadg7492. https://doi.org/10.1126/science.adg7492

23. Ciesielski, T.H., Sirugo, G., Iyengar, S.K. et al. Characterizing the pathogenicity of genetic variants: the consequences of context. npj Genom. Med. 9, 3 (2024). https://doi.org/10.1038/s41525-023-00386-5

24. Claussnitzer, M., Cho, J.H., Collins, R. et al. A brief history of human disease genetics. Nature 577, 179–189 (2020). https://doi.org/10.1038/s41586-019-1879-7

25. Cock, P. J., Fields, C. J., Goto, N., Heuer, M. L., & Rice, P. M. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic acids research, 38(6), 1767–1771. https://doi.org/10.1093/nar/gkp1137

26. Copson, E. R., Maishman, T. C., Tapper, W. J., Cutress, R. I., Greville-Heygate, S., Altman, D. G., Eccles, B., Gerty, S., Durcan, L. T., Jones, L., Evans, D. G., Thompson, A. M., Pharoah, P., Easton, D. F., Dunning, A. M., Hanby, A., Lakhani, S., Eeles, R., Gilbert, F. J., Hamed, H., … Eccles, D. M. (2018). Germline BRCA mutation and outcome in young-onset breast cancer (POSH): a prospective cohort study. The Lancet. Oncology, 19(2), 169–180. https://doi.org/10.1016/S1470-2045(17)30891-4

27. Crisafulli G. (2024). Mutational Signatures in Colorectal Cancer: Translational Insights, Clinical Applications, and Limitations. Cancers, 16(17), 2956. https://doi.org/10.3390/cancers16172956

28. Cvijović, I., Good, B. H., & Desai, M. M. (2018). The Effect of Strong Purifying Selection on Genetic Diversity. Genetics, 209(4), 1235–1278. https://doi.org/10.1534/genetics.118.301058

29. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics (Oxford, England), 27(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

30. Danecek P, Bonfield JK, et al. Twelve years of SAMtools and BCFtools. Gigascience (2021) 10(2):giab008

31. Díaz-Gay, M., Vangara, R., Barnes, M., Wang, X., Islam, S. M. A., Vermes, I., Duke, S., Narasimman, N. B., Yang, T., Jiang, Z., Moody, S., Senkin, S., Brennan, P., Stratton, M. R., & Alexandrov, L. B. (2023). Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. Bioinformatics (Oxford, England), 39(12), btad756. https://doi.org/10.1093/bioinformatics/btad756

32. Diaz-Papkovich, A., Anderson-Trocmé, L., & Gravel, S. (2021). A review of UMAP in population genetics. Journal of human genetics, 66(1), 85–91. https://doi.org/10.1038/s10038-020-00851-4

33. Elhaik, E. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. Sci Rep 12, 14683 (2022). https://doi.org/10.1038/s41598-022-14395-4

34. Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. Nature reviews. Genetics, 8(8), 610–618. https://doi.org/10.1038/nrg2146

35. Farmanbar, A., Kneller, R. & Firouzi, S. Mutational signatures reveal mutual exclusivity of homologous recombination and mismatch repair deficiencies in colorectal and stomach tumors. Sci Data 10, 423 (2023). https://doi.org/10.1038/s41597-023-02331-8

36. Fedorova, L., Khrunin, A., Khvorykh, G., Lim, J., Thornton, N., Mulyar, O. A., Limborska, S., & Fedorov, A. (2022). Analysis of Common SNPs across Continents Reveals Major Genomic Differences between Human Populations. Genes, 13(8), 1472. https://doi.org/10.3390/genes13081472

37. Fu, R., Zhang, J. T., Chen, R. R., Li, H., Tai, Z. X., Lin, H. X., Su, J., Chu, X. P., Zhang, C., Qiu, Z. B., Chen, Z. H., Tang, W. F., Dong, S., Yang, X. N., Zhang, G. Q., Zhao, G. P., Wu, Y. L., & Zhong, W. Z. (2022). Identification of heritable rare variants associated with early-stage lung adenocarcinoma risk. Translational lung cancer research, 11(4), 509–522. https://doi.org/10.21037/tlcr-21-789

38. Futuyma, D. J. (1986). Evolutionary Biology 2nd edn, Sinauer

39. Glaviano, A., Foo, A.S.C., Lam, H.Y. et al. PI3K/AKT/mTOR signaling transduction pathway and targeted therapies in cancer. Mol Cancer 22, 138 (2023). https://doi.org/10.1186/s12943-023-01827-6

40. Guo, C., McDowell, I. C., Nodzenski, M., Scholtens, D. M., Allen, A. S., Lowe, W. L., & Reddy, T. E. (2017). Transversions have larger regulatory effects than transitions. BMC genomics, 18(1), 394. https://doi.org/10.1186/s12864-017-3785-4

41. Greenbaum, G., Templeton, A. R., Zarmi, Y., & Bar-David, S. (2014). Allelic richness following population founding events--a stochastic modeling framework incorporating gene flow and genetic drift. PloS one, 9(12), e115203. https://doi.org/10.1371/journal.pone.0115203

42. Gromiha, M. M., Pandey, M., Kulandaisamy, A., Sharma, D., & Ridha, F. (2024). Progress on the development of prediction tools for detecting disease causing mutations in proteins. Computers in

biology and medicine, 185, 109510. Advance online publication. https://doi.org/10.1016/j.compbiomed.2024.109510

43. Hall, A. W., Battenhouse, A. M., Shivram, H., Morris, A. R., Cowperthwaite, M. C., Shpak, M., & Iyer, V. R. (2018). Bivalent Chromatin Domains in Glioblastoma Reveal a Subtype-Specific Signature of Glioma Stem Cells. Cancer research, 78(10), 2463–2474. https://doi.org/10.1158/0008-5472.CAN-17-1724

44. Hanahan, D., & Weinberg, R. A. (2000). The hallmarks of cancer. Cell, 100(1), 57–70. https://doi.org/10.1016/s0092-8674(00)81683-9

45. Hosseini, M., Pratas, D., & Pinho, A. J. (2016). A Survey on Data Compression Methods for Biological Sequences. Information, 7(4), 56. https://doi.org/10.3390/info7040056

46. Hutchison, W.J., Keyes, T.J., The tidyomics Consortium. et al. The tidyomics ecosystem: enhancing omic data analyses. Nat Methods 21, 1166–1170 (2024). https://doi.org/10.1038/s41592-024-02299-2

47. Jayakrishnan, R., Kwiatkowski, D. J., Rose, M. G., & Nassar, A. H. (2024). Topography of mutational signatures in non-small cell lung cancer: emerging concepts, clinical applications, and limitations. The oncologist, 29(10), 833–841. https://doi.org/10.1093/oncolo/oyae091

48. Ji, X., Wang, E., & Cui, Q. (2023). Deciphering gene contributions and etiologies of somatic mutational signatures of cancer. Briefings in bioinformatics, 24(2), bbad017. https://doi.org/10.1093/bib/bbad017

49. Kato, G., Piel, F., Reid, C. et al. Sickle cell disease. Nat Rev Dis Primers 4, 18010 (2018). https://doi.org/10.1038/nrdp.2018.10

50. Karczewski, K.J., Francioli, L.C., Tiao, G. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature 581, 434–443 (2020). https://doi.org/10.1038/s41586-020-2308-7

51. Kilgour, J. M., Jia, J. L., & Sarin, K. Y. (2021). Review of the Molecular Genetics of Basal Cell Carcinoma; Inherited Susceptibility, Somatic Mutations, and Targeted Therapeutics. Cancers, 13(15), 3870. https://doi.org/10.3390/cancers13153870

52. Koh, G., Degasperi, A., Zou, X. et al. Mutational signatures: emerging concepts, caveats and clinical applications. Nat Rev Cancer 21, 619–637 (2021). https://doi.org/10.1038/s41568-021-00377-7

53. Koprulu, M., Zhao, Y., Wheeler, E., Dong, L., Rocha, N., Li, C., Griffin, J. D., Patel, S., Van de Streek, M., Glastonbury, C. A., Stewart, I. D., Day, F. R., Luan, J., Bowker, N., Wittemans, L. B. L., Kerrison, N. D., Cai, L., Lucarelli, D. M. E., Barroso, I., McCarthy, M. I., … Savage, D. B. (2022). Identification of Rare Loss-of-Function Genetic Variation Regulating Body Fat Distribution. The Journal of clinical endocrinology and metabolism, 107(4), 1065–1077. https://doi.org/10.1210/clinem/dgab877

54. Krumm, N., Turner, T. N., Baker, C., Vives, L., Mohajeri, K., Witherspoon, K., Raja, A., Coe, B. P., Stessman, H. A., He, Z. X., Leal, S. M., Bernier, R., & Eichler, E. E. (2015). Excess of rare, inherited truncating mutations in autism. Nature genetics, 47(6), 582–588. https://doi.org/10.1038/ng.3303

55. Kumar, S., Warrell, J., Li, S., McGillivray, P. D., Meyerson, W., Salichos, L., Harmanci, A., Martinez-Fundichely, A., Chan, C. W. Y., Nielsen, M. M., Lochovsky, L., Zhang, Y., Li, X., Lou, S., Pedersen, J. S., Herrmann, C., Getz, G., Khurana, E., & Gerstein, M. B. (2020). Passenger Mutations in More Than 2,500 Cancer Genomes: Overall Molecular Functional Impact and Consequences. Cell, 180(5), 915–927.e16. https://doi.org/10.1016/j.cell.2020.01.032

56. Kumar, S., & Gerstein, M. (2023). Unified views on variant impact across many diseases. Trends in genetics : TIG, 39(6), 442–450. https://doi.org/10.1016/j.tig.2023.02.002

57. Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research, 42(Database issue), D980–D985. https://doi.org/10.1093/nar/gkt1113

58. Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., Lyoshin, V., Maddipatla, Z., Maiti, R., Mitchell, J., O'Leary, N., Riley, G. R., Shi, W., Zhou, G., Schneider, V., Maglott, D., … Kattman, B. L. (2020). ClinVar: improvements to accessing data. Nucleic acids research, 48(D1), D835–D844. https://doi.org/10.1093/nar/gkz972

59. Langmead, B., Trapnell, C., Pop, M. et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10, R25 (2009). https://doi.org/10.1186/gb-2009-10-3-r25

60. Leggatt, G., Cheng, G., Narain, S., Briseño-Roa, L., Annereau, J. P., Genomics England Research Consortium, Gast, C., Gilbert, R. D., & Ennis, S. (2023). A genotype-to-phenotype approach suggests under-reporting of single nucleotide variants in nephrocystin-1 (NPHP1) related disease (UK 100,000 Genomes Project). Scientific reports, 13(1), 9369. https://doi.org/10.1038/s41598-023-32169-4

61. Letafati, A., Taghiabadi, Z., Zafarian, N. et al. Emerging paradigms: unmasking the role of oxidative stress in HPV-induced carcinogenesis. Infect Agents Cancer 19, 30 (2024). https://doi.org/10.1186/s13027-024-00581-8

62. Levental, K. R., Yu, H., Kass, L., Lakins, J. N., Egeblad, M., Erler, J. T., Fong, S. F., Csiszar, K., Giaccia, A., Weninger, W., Yamauchi, M., Gasser, D. L., & Weaver, V. M. (2009). Matrix crosslinking forces tumor progression by enhancing integrin signaling. Cell, 139(5), 891–906. https://doi.org/10.1016/j.cell.2009.10.027

63. Li, B., Brady, S. W., Ma, X., Shen, S., Zhang, Y., Li, Y., Szlachta, K., Dong, L., Liu, Y., Yang, F., Wang, N., Flasch, D. A., Myers, M. A., Mulder, H. L., Ding, L., Liu, Y., Tian, L., Hagiwara, K., Xu, K., Zhou,

X., … Zhang, J. (2020). Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. Blood, 135(1), 41–55. https://doi.org/10.1182/blood.2019002220

64. Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England), 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

65. Lin, H., Zhang, G., Zhang, X. C., Lian, X. L., Zhong, W. Z., Su, J., Chen, S. L., & Wu, Y. L. (2020). Germline variation networks in the PI3K/AKT pathway corresponding to familial high-incidence lung cancer pedigrees. BMC cancer, 20(1), 1209. https://doi.org/10.1186/s12885-020-07528-3

66. Liu, Q., Guo, Y., Li, J., Long, J., Zhang, B., & Shyr, Y. (2012). Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC genomics, 13 Suppl 8(Suppl 8), S8. https://doi.org/10.1186/1471-2164-13-S8-S8

67. Luo, Z., & Farnham, P. J. (2020). Genome-wide analysis of HOXC4 and HOXC6 regulated genes and binding sites in prostate cancer cells. PloS one, 15(2), e0228590. https://doi.org/10.1371/journal.pone.0228590

68. Luongo, F., Colonna, F., Calapà, F., Vitale, S., Fiori, M. E., & De Maria, R. (2019). PTEN Tumor-Suppressor: The Dam of Stemness in Cancer. Cancers, 11(8), 1076. https://doi.org/10.3390/cancers11081076

69. Matlak, D., & Szczurek, E. (2017). Epistasis in genomic and survival data of cancer patients. PLoS computational biology, 13(7), e1005626. https://doi.org/10.1371/journal.pcbi.1005626

70. McFarland, C. D., Yaglom, J. A., Wojtkowiak, J. W., Scott, J. G., Morse, D. L., Sherman, M. Y., & Mirny, L. A. (2017). The Damaging Effect of Passenger Mutations on Cancer Progression. Cancer research, 77(18), 4763–4772. https://doi.org/10.1158/0008-5472.CAN-15-3283-T

71. McLaren, W., Gil, L., Hunt, S.E. et al. The Ensembl Variant Effect Predictor. Genome Biol 17, 122 (2016). https://doi.org/10.1186/s13059-016-0974-4

72. Mehmood MA, Sehar U, Ahmad N (2014) Use of Bioinformatics Tools in Different Spheres of Life Sciences. J Data Mining Genomics Proteomics 5: 158. doi:10.4172/2153-0602.1000158

73. Meyerson, W., Leisman, J., Navarro, F. C. P., & Gerstein, M. (2020). Origins and characterization of variants shared between databases of somatic and germline human mutations. BMC bioinformatics, 21(1), 227. https://doi.org/10.1186/s12859-020-3508-8

74. Monroe, J. G., Srikant, T., Carbonell-Bejerano, P., Becker, C., Lensink, M., Exposito-Alonso, M., Klein, M., Hildebrandt, J., Neumann, M., Kliebenstein, D., Weng, M. L., Imbert, E., Ågren, J., Rutter, M. T., Fenster, C. B., & Weigel, D. (2022). Mutation bias reflects natural selection in Arabidopsis thaliana. Nature, 602(7895), 101–105. https://doi.org/10.1038/s41586-021-04269-6

75. Morales, J., Pujar, S., Loveland, J. E., Astashyn, A., Bennett, R., Berry, A., Cox, E., Davidson, C., Ermolaeva, O., Farrell, C. M., Fatima, R., Gil, L., Goldfarb, T., Gonzalez, J. M., Haddad, D., Hardy, M., Hunt, T., Jackson, J., Joardar, V. S., Kay, M., … Murphy, T. D. (2022). A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. *Nature*, *604*(7905), 310–315. https://doi.org/10.1038/s41586-022-04558-8

76. Nesta, A. V., Tafur, D., & Beck, C. R. (2021). Hotspots of Human Mutation. Trends in genetics: TIG, 37(8), 717–729. https://doi.org/10.1016/j.tig.2020.10.003

77. Niu J., Denisko D., Hoffman M. (2022). The Browser Extensible Data (BED) format (PDF). https://samtools.github.io/hts-specs/BEDv1.pdf

78. Omer, S., Harlow, T. J., & Gogarten, J. P. (2017). Does Sequence Conservation Provide Evidence for Biological Function?. Trends in microbiology, 25(1), 11–18. https://doi.org/10.1016/j.tim.2016.09.010

79. Ostroverkhova, D., Przytycka, T. M., & Panchenko, A. R. (2023). Cancer driver mutations: predictions and reality. Trends in molecular medicine, 29(7), 554–566. https://doi.org/10.1016/j.molmed.2023.03.007

80. Pal, L. R., & Moult, J. (2015). Genetic Basis of Common Human Disease: Insight into the Role of Missense SNPs from Genome-Wide Association Studies. Journal of molecular biology, 427(13), 2271–2289. https://doi.org/10.1016/j.jmb.2015.04.014

81. Pandey, P., & Alexov, E. (2023). Most monogenic disorders are caused by mutations altering protein folding free energy. Research square, rs.3.rs-3442589. https://doi.org/10.21203/rs.3.rs-3442589/v1

82. Panwar, V., Singh, A., Bhatt, M. et al. Multifaceted role of mTOR (mammalian target of rapamycin) signaling pathway in human health and disease. Sig Transduct Target Ther 8, 375 (2023). https://doi.org/10.1038/s41392-023-01608-z

83. Pathan, N., Deng, W.Q., Di Scipio, M. et al. A method to estimate the contribution of rare coding variants to complex trait heritability. Nat Commun 15, 1245 (2024). https://doi.org/10.1038/s41467-024-45407-8

84. Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. PLoS genetics, 2(12), e190. https://doi.org/10.1371/journal.pgen.0020190

85. Perera, D., Poulos, R. C., Shah, A., Beck, D., Pimanda, J. E., & Wong, J. W. (2016). Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. Nature, 532(7598), 259–263. https://doi.org/10.1038/nature17437

86. Pužar Dominkuš, P., & Hudler, P. (2023). Mutational Signatures in Gastric Cancer and Their Clinical Implications. Cancers, 15(15), 3788. https://doi.org/10.3390/cancers15153788

87. Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics (Oxford, England), 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

88. Quintana-Murci L. (2016). Understanding rare and common diseases in the context of human evolution. Genome biology, 17(1), 225. https://doi.org/10.1186/s13059-016-1093-y

89. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic acids research, 47(D1), D886–D894. https://doi.org/10.1093/nar/gky1016

90. Reich, D. E., & Lander, E. S. (2001). On the allelic spectrum of human disease. Trends in genetics : TIG, 17(9), 502–510. https://doi.org/10.1016/s0168-9525(01)02410-6

91. Reich, D., Price, A. & Patterson, N. Principal component analysis of genetic data. Nat Genet 40, 491–492 (2008). https://doi.org/10.1038/ng0508-491

92. Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405: 847–856

93. Sakaue, S., Hirata, J., Kanai, M. et al. Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction. Nat Commun 11, 1569 (2020). https://doi.org/10.1038/s41467-020-15194-z

94. Seplyarskiy, V.B., Sunyaev, S. The origin of human mutation in light of genomic data. Nat Rev Genet 22, 672–686 (2021). https://doi.org/10.1038/s41576-021-00376-2

95. Shakya, M., Ahmed, S.A., Davenport, K.W. et al. Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. Sci Rep 10, 1723 (2020). https://doi.org/10.1038/s41598-020-58356-1

96. Shen, X., Song, S., Li, C. et al. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. Nature 606, 725–731 (2022). https://doi.org/10.1038/s41586-022-04823-w

97. Shteinberg, M., Haq, I. J., Polineni, D., & Davies, J. C. (2021). Cystic fibrosis. Lancet (London, England), 397(10290), 2195–2211. https://doi.org/10.1016/S0140-6736(20)32542-3

98. Sondka, Z., Dhir, N. B., Carvalho-Silva, D., Jupe, S., Madhumita, McLaren, K., Starkey, M., Ward, S., Wilding, J., Ahmed, M., Argasinska, J., Beare, D., Chawla, M. S., Duke, S., Fasanella, I., Neogi, A. G., Haller, S., Hetenyi, B., Hodges, L., Holmes, A., … Teague, J. (2024). COSMIC: a curated database of somatic variants and clinical data for cancer. Nucleic acids research, 52(D1), D1210–D1217. https://doi.org/10.1093/nar/gkad986

99. Sun, H., & Yu, G. (2019). New insights into the pathogenicity of non-synonymous variants through multi-level analysis. Scientific reports, 9(1), 1667. https://doi.org/10.1038/s41598-018-38189-9

100. Tan H. (2017). The association between gene SNPs and cancer predisposition: Correlation or causality?. EBioMedicine, 16, 8–9. https://doi.org/10.1016/j.ebiom.2017.01.047

101. Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting, L., Ramshaw, C. C., Rye, C. E., Speedy, H. E., … Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic acids research, 47(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

102. Tenorio, J., Nevado, J., González-Meneses, A., Arias, P., Dapía, I., Venegas-Vega, C. A., Calvente, M., Hernández, A., Landera, L., Ramos, S., SOGRI Consortium, Cigudosa, J. C., Pérez-Jurado, L. A., & Lapunzina, P. (2020). Further definition of the proximal 19p13.3 microdeletion/microduplication syndrome and implication of PIAS4 as the major contributor. Clinical genetics, 97(3), 467–476. https://doi.org/10.1111/cge.13689

103. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 526, 68–74 (2015). https://doi.org/10.1038/nature15393

104. Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J. M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., Pretorius, G. S., … Williams, S. M. (2009). The genetic structure and history of Africans and African Americans. Science (New York, N.Y.), 324(5930), 1035–1044. https://doi.org/10.1126/science.1172257

105. Tomasetti, C., & Vogelstein, B. (2015). Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science (New York, N.Y.), 347(6217), 78–81. https://doi.org/10.1126/science.1260825

106. Trost, B., Loureiro, L. O., & Scherer, S. W. (2021). Discovery of genomic variation across a generation. Human molecular genetics, 30(R2), R174–R186. https://doi.org/10.1093/hmg/ddab209

107. van der Ham, C.G., Suurenbroek, L.C., Kleisman, M.M. et al. Mutational mechanisms in multiply relapsed pediatric acute lymphoblastic leukemia. Leukemia 38, 2366–2375 (2024). https://doi.org/10.1038/s41375-024-02403-7

108. Venizelos, A., Sorbye, H., Elvebakken, H., Perren, A., Lothe, I. M. B., Couvelard, A., Hjortland, G. O., Sundlöv, A., Svensson, J., Garresori, H., Kersten, C., Hofsli, E., Detlefsen, S., Vestermark, L. W., Ladekarl, M., Tabaksblat, E. M., & Knappskog, S. (2023). Germline pathogenic variants in patients with high-grade gastroenteropancreatic neuroendocrine neoplasms. Endocrine-related cancer, 30(10), e230057. https://doi.org/10.1530/ERC-23-0057

109. Viansone, A., Pellegrino, B., Omarini, C., Pistelli, M., Boggiani, D., Sikokis, A., Uliana, V., Zanoni, D., Tommasi, C., Bortesi, B., Bonatti, F., Piacentini, F., Cortesi, L., Camisa, R., Sgargi, P., Michiara, M., &

Musolino, A. (2022). Prognostic significance of germline BRCA mutations in patients with HER2-POSITIVE breast cancer. Breast (Edinburgh, Scotland), 65, 145–150. https://doi.org/10.1016/j.breast.2022.07.012

110. von Stedingk, K., Stjernfelt, K. J., Kvist, A., Wahlström, C., Kristoffersson, U., Stenmark-Askmalm, M., Wiebe, T., Hjorth, L., Koster, J., Olsson, H., & Øra, I. (2021). Prevalence of germline pathogenic variants in 22 cancer susceptibility genes in Swedish pediatric cancer patients. Scientific reports, 11(1), 5307. https://doi.org/10.1038/s41598-021-84502-4

111. Wan, J.C.M., Stephens, D., Luo, L. et al. Genome-wide mutational signatures in low-coverage whole genome sequencing of cell-free DNA. Nat Commun 13, 4953 (2022). https://doi.org/10.1038/s41467-022-32598-1

112. Wang, H., Guo, M., Wei, H. et al. Targeting p53 pathways: mechanisms, structures and advances in therapy. Sig Transduct Target Ther 8, 92 (2023). https://doi.org/10.1038/s41392-023-01347-1

113. Wang, H., Liu, C., Jin, K., Li, X., Zheng, J., & Wang, D. (2024). Research advances in signaling pathways related to the malignant progression of HSIL to invasive cervical cancer: A review. Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie, 180, 117483. https://doi.org/10.1016/j.biopha.2024.117483

114. Wang, J., Raskin, L., Samuels, D. C., Shyr, Y., & Guo, Y. (2015). Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics (Oxford, England), 31(3), 318–323. https://doi.org/10.1093/bioinformatics/btu668

115. Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research, 38(16), e164. https://doi.org/10.1093/nar/gkq603

116. Wang, M., Li, S. C., & Shen, B. (2024). Elevated incidence of somatic mutations at prevalent genetic sites. Briefings in bioinformatics, 25(2), bbae065. https://doi.org/10.1093/bib/bbae065

117. Wheeler, L. C., Wing, B. A., & Smith, S. D. (2020). Structure and contingency determine mutational hotspots for flower color evolution. Evolution letters, 5(1), 61–74. https://doi.org/10.1002/evl3.212

118. Witt, K. E., & Huerta-Sánchez, E. (2019). Convergent evolution in human and domesticate adaptation to high-altitude environments. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 374(1777), 20180235. https://doi.org/10.1098/rstb.2018.0235

119. World Health Organization (WHO). Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019. WHO; 2020. Accessed December 11, 2020. https://www.who.int/health-topics/cancer#tab=tab_1

120. Wu, H., Lin, JH., Tang, XY. et al. Combining full-length gene assay and SpliceAI to interpret the splicing impact of all possible SPINK1 coding variants. Hum Genomics 18, 21 (2024). https://doi.org/10.1186/s40246-024-00586-9

121. Xiang, Z., Liu, H. & Hu, Y. DNA damage repair and cancer immunotherapy. GENOME INSTAB. DIS. 4, 210–226 (2023). https://doi.org/10.1007/s42764-023-00098-1

122. Yaacov, A., Rosenberg, S. & Simon, I. Mutational signatures association with replication timing in normal cells reveals similarities and differences with matched cancer tissues. Sci Rep 13, 7833 (2023). https://doi.org/10.1038/s41598-023-34631-9

123. Yaacov, A., Ben Cohen, G., Landau, J., Hope, T., Simon, I., & Rosenberg, S. (2024). Cancer mutational signatures identification in clinical assays using neural embedding-based representations. Cell reports. Medicine, 5(6), 101608. https://doi.org/10.1016/j.xcrm.2024.101608

124. Yadav, S., Virk, R., Chung, C.H. et al. Lipid exposure activates gene expression changes associated with estrogen receptor negative breast cancer. npj Breast Cancer 8, 59 (2022). https://doi.org/10.1038/s41523-022-00422-0

125. Yamamoto, K., Lee, Y., Masuda, T., Ozono, K., Iwatani, Y., Watanabe, M., Okada, Y., & Sakai, N. (2024). Functional landscape of genome-wide postzygotic somatic mutations between monozygotic twins. DNA research : an international journal for rapid publication of reports on genes and genomes, 31(5), dsae028. https://doi.org/10.1093/dnares/dsae028

126. Yu, Z., Coorens, T.H.H., Uddin, M.M. et al. Genetic variation across and within individuals. Nat Rev Genet (2024). https://doi.org/10.1038/s41576-024-00709-x

127. Zou, H., Pan, T., Gao, Y., Chen, R., Li, S., Guo, J., Tian, Z., Xu, G., Xu, J., Ma, Y., & Li, Y. (2022). Pan-cancer assessment of mutational landscape in intrinsically disordered hotspots reveals potential driver genes. Nucleic acids research, 50(9), e49. https://doi.org/10.1093/nar/gkac028

128. Zou, Z., Tao, T., Li, H. et al. mTOR signaling pathway and mTOR inhibitors in cancer: progress and challenges. Cell Biosci 10, 31 (2020). https://doi.org/10.1186/s13578-020-00396-1

Mutational landscape of the human genome: co-occurrence of single nucleotide variants (SNVs) of different evolutionary origins and functional consequences

Thais Silva Tavares[1]; Daniela Souza Lima Barbosa[2]; Carolina Silva de Carvalho[1]; Maycon Douglas de Oliveira[1]; Luigi Marchionni[3]; Renan Pedra de Souza[1]; Francisco Pereira Lobo[1‡]

[1]Institute of Biological Sciences, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

[2]Pontifícia Universidade Católica de Minas Gerais PUC, Belo Horizonte, MG, Brazil

[3]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine, New York, NY, USA

[‡]Corresponding author:

Francisco Pereira Lobo: franciscolobo@gmail.com

ABSTRACT

Background

Mutations play a key role in genome evolution and the development of disease. Despite decades of research on variations in mutation rates and types across genomic regions, few studies have systematically integrated comparative analyses of their spatial distribution in the human genome.

Results

We conducted a large-scale assessment of the distribution and functional characterization of single nucleotide variants (SNVs) across five distinct categories: (i) neutral or nearly neutral polymorphisms, (ii) rare variants, (iii) cancer-associated somatic mutations, (iv) clinically

relevant pathogenic germline variants, and (v) benign variants. Our findings reveal a significant overlap between different SNV classes, suggesting that certain genomic regions are inherently more prone to accumulate mutations, regardless of their germline or somatic origin. Single nucleotide polymorphisms (SNPs) and recurrent mutations in cancer have been extensively studied to uncover relationships that advance our understanding of cancer biology. Surprisingly, we showed that SNPs and recurrent cancer mutations frequently occur at the same genomic loci and share identical alleles more often than expected by chance. Five mutational signatures (SBS1, SBS5, SBS6, SBS54, and SBS87) were associated with these shared sites, whereas CpG islands and microsatellites accounted for only a minor fraction of the observed mutations. Furthermore, functional enrichment analyses revealed significant associations between the overlapping regions and key genomic sites, such as 13q12.12 and 19p13.3, as well as pathways involved in cellular signaling and extracellular matrix interactions, including PI3K/AKT/mTOR and ECM-receptor interactions, both of which are crucial for tumor progression.

Conclusions

We propose the existence of critical DNA loci that function as natural mutational hotspots that are recurrently affected in both germline and somatic genomes. The identification of these hotspots raises important questions regarding the underlying mechanisms driving their occurrence and their potential role in cancer predisposition.

Keywords: Human mutations, Single nucleotide variants (SNVs), Mutational hotspots, Germline-somatic shared loci

INTRODUCTION

Mutations are a major component of the evolutionary process, providing the primary source of genetic variation upon which natural selection acts. Over evolutionary time, these genetic changes can drive the diversification of populations and species, together with the development of complex traits through adaptation. However, mutations also play a critical role in the emergence of genetic diseases. While some mutations can be neutral or provide adaptive advantages, others can disrupt biological functions. Understanding the dual nature of mutations — both as catalysts for evolution and contributors to disease — offers crucial insights into the intricate dynamics of human genetics and health [1, 2].

Mutations may be classified in different ways according to their origin, functional consequences, and other aspects. A major distinction is germline and somatic mutations. Germinative mutations, defined as those that emerge *de novo* during meiotic gametogenesis, are transmitted to offspring via sexual reproduction. In contrast, somatic mutations are found in somatic tissues and are inherited by daughter cells through mitotic divisions [3]. From a functional perspective, mutations can be classified as either pathogenic, defined as those causally associated with disease phenotypes, or neutral/benign, when they are not found to cause diseases or are even beneficial in some conditions [4]. From a population genetics evolutionary perspective, germline mutations can be classified as either polymorphic, when they are observed in a substantial fraction of individuals of a population, or rare, if not polymorphic. Rare variants may also play a crucial role in the pathogenesis of both complex and Mendelian diseases by potentially contributing to the unexplained heritability commonly observed in complex diseases with a strong genetic

component [5]. Access to an abundance of human genomic data has transformed our comprehension of mutation types. The human genome, one of the most thoroughly curated genomic resources available, offers an unparalleled framework for exploring the distribution and frequency of diverse mutation classes [6–9].

Single nucleotide variants (SNVs) are among the most prevalent mutational alterations in the human genome and play crucial roles in the genetic variability of our species and its phenotypic implications. These phenotypes exhibit notable variability influenced by the genomic background, origin, and location of SNVs [10]. Some germline mutations, for example, might contribute to a wide range of hereditary conditions with varying degrees of frequency, penetrance, and expressivity. These diseases range from classical monogenic diseases such as sickle cell anemia [11] and cystic fibrosis [12] to complex polygenic traits [13, 14]. Noninheritable SNVs, which originate in somatic cells and are recurrently detected in different individuals exhibiting pathogenic alterations, also hold considerable medical and scientific importance, especially in association with complex diseases, as observed in some cancer cases [15, 16]. Single nucleotide polymorphisms (SNPs) are germline SNVs that have a high allele frequency and are commonly defined as SNVs with frequencies greater than 1% in the studied population [17]. Typically, SNPs are considered neutral variants of no medical significance, as their higher population frequencies suggest that these variants do not have considerable negative consequences for individual fitness. In fact, some SNPs may even be adaptive [18]. Thus, while some SNVs may be considered neutral or even beneficial in certain populations from an evolutionary perspective, others have an adverse impact on evolutionary fitness and tend to be less frequent [19].

The impossibility of scientifically predicting the locus where the next mutation will occur (randomness) in a cell does not necessarily mean that these mutations have an equal probability of occurrence in all nucleotides that compose the genome (uniformity). Some regions have a relatively high probability of mutation and have historically been termed mutational hotspots [20]. Some factors, such as transcription-coupled DNA repair and epigenetic modifications, have already been identified as influencing the frequency of mutations in different DNA regions [21]. In the human genome, mutational hotspots are regions highly susceptible to selective pressure and, in many cases, are related to complex genetic diseases, such as cancer [22, 23]. Classic examples of hotspots include CpG islands, microsatellites, centromeric DNA, AT-rich palindromes, and hairpin-forming sequences, among others [24]. The greater susceptibility of these DNA areas to mutations can be attributed to many mutation-inducing factors, together with the structure of the DNA sequence and the enzymes involved in DNA repair, replication, and modification [24].

Surprisingly, few studies have investigated the spatial distribution of the different classes of mutations in the human genome from evolutionary and functional perspectives (neutral/beneficial *versus* deleterious, somatic *versus* germline, and common *versus* rare) [24]. Most disease-related genomics studies focus on a single type of mutation or human disease, whether cancer, rare, or common, leading to disconnected approaches; however, there is an increasing need to promote integrative studies that encourage the exchange of ideas and methods for a more comprehensive and cohesive understanding of these different sets of diseases [9]. To our knowledge, few recent studies have evaluated the genomic overlap between two types of mutations: cancer-associated somatic mutations and germline mutations. A significant number of shared variants have been observed, defined as

mutations that occur at the same locus and/or involve the same nucleotide [3, 25]. This co-occurrence phenomenon for both types of mutations has been attributed to chemical vulnerabilities in the DNA or to specific mutational signatures, such as SBS1, SBS87, and DBS7. Statistical analyses that interpret these patterns within the framework of known biochemical processes could offer mechanistic models of mutagenesis in humans.

In this study, we integrated information from public databases to identify new evidence of the nonuniformity and nonrandomness of mutations in the human genome by assessing the co-occurrence of variants of five different profiles. Our analysis provides insights into the distribution patterns of different mutation types, serving as a guide for future studies. The databases HGDP, COSMIC, and ClinVar harbor human variants with distinct biological properties, such as origin (germline *versus* somatic), phenotypic consequences (neutral/beneficial *versus* pathogenic), frequency (polymorphic *versus* rares) and penetrance (low *versus* high). COSMIC focuses on somatic mutations found in different types of human cancers, compiling data that help better understand the genetic mechanisms underlying cancer development [6]. The HGDP (Human Genome Diversity Project) offers a genetic framework for individuals from isolated populations worldwide and is crucial for investigating genomic diversity among human populations [8]. ClinVar is a curated catalog of germline variants together with their associations with clinical conditions, providing a valuable data source for physicians, researchers, and patients to assess the clinical relevance of certain variants [7].

Through these three databases, we produced five different sets of variants: clinically relevant germline variants i) neutral/benign or ii) pathogenic; iii) cancer-related somatic

variants; and iv) germline variants classified as either polymorphic or v) rare variants. We evaluated the existence of genomic loci overlapping these five distinct classes of SNVs and assessed whether this overlap was greater than expected by chance. Additionally, we functionally characterized the variant sets and evaluated the subsets of SNVs that showed intersection in each comparison performed, both those sharing the same nucleotide (shared variants - [3]) and those revealing intersection of position but involving different alleles.

As a result, we report the occurrence of common mutational hotspots for different types of mutations, whether germline (neutral or pathogenic) or cancer-related somatic mutations. Thus, we propose that mutations of different classes generally occur in intrinsic regions of DNA that are more prone to mutational events. Additionally, we suggest the existence of one or more common mechanisms related to SBS1, SBS5, SBS6, SBS54, and SBS87, which are responsible for the spatial signature or mutational control in the human genome. Interestingly, this mechanism appears to be associated with the occurrence of germline and somatic mutations in genes on the short arm of chromosome 19 and the PI3K/Akt/mTOR pathway. This finding reinforces the idea of nonuniformity and nonrandomness in mutation occurrence throughout the entire human genome.

METHODS

Data acquisition

We downloaded VCF data from databases that represent distinct types of SNVs in the human genome regarding their tissue of occurrence (somatic *versus* germline), phenotypic

consequences (pathogenic *versus* neutral/benign) and frequency across human populations (rares *versus* polymorphisms).

Identification of single nucleotide polymorphisms (SNPs)

The Human Genome Diversity Project (HGDP) [8] provides data on variants that represent the genomic diversity of 929 individuals stemming from 54 isolated human populations across seven major superpopulations: Africa, America, Central and South Asia, East Asia, Europe, the Middle East, and Oceania. We obtained germline SNVs, on the basis of the human genome assembly version GRCh38, from HGDP through the following filtering pipeline: transcriptome samples were discarded, resulting in 894 samples; high-quality ("PASS" VCF parameter) and biallelic variants were filtered via BCFtools version 1.10.2 [26] - specifically, the PASS filter was applied to select only variants that passed all quality filters, excluding those marked with the LOW_VQSLOD and ExcHet filters, which indicate low quality or excess heterozygosity, respectively [8]; all autosomal SNVs were pruned for linkage disequilibrium (LD) via the 'indep-pairwise' parameter in PLINK2 [27], with a window size of 200 kb, a step size of 25 SNVs, and a r2 value of 0.4, following [28]. Additionally, individuals and variants with more than 10% missing data were removed via the 'mind 0.1' and 'geno 0.1' parameters in PLINK2, respectively [28].

We reduced the data dimensionality by first conducting a principal component analysis (PCA) on the set of high-quality variants resulting from the previous step via PLINK2 with default settings. Afterward, we used the output of the PCA as input for the *umap* R library to compute the uniform manifold approximation and projection (UMAP) and better visualize the population structure when searching for isolated individuals. We set the parameters 'n_neighbors 30', 'min_dist 0.1' and 'spread 5' following [29]. After all the data processing steps, we defined polymorphic variants as those with global allele frequency (AF) values ≥

0.01 (1% allele frequency) and those present in at least two of the seven isolated superpopulations previously defined through PCA-UMAP analysis.

Identification of Rare Single Nucleotide Variants

Through the HGDP dataset, we established a set of rare SNVs by excluding common germline SNVs. To achieve this, we filtered all SNVs with AFs less than 1% (AF<0.01) across the entire HGDP dataset, considering all predefined HGDP populations. The filtering process also involved retaining only high-quality (PASS) and biallelic autosomal variants.

Identification of Somatic SNVs that are recurrently mutated in cancer

The Catalog of Somatic Mutations In Cancer (COSMIC) database [6] provides two separate VCF files containing variants (genome assembly version GRCh38) of two different types: one for variants occurring in coding regions of the human genome and another for noncoding regions. We used BCFTools [30] to merge the variants from both files, using the merged results as a source of recurrent single nucleotide somatic mutations found in cancer. As data processing steps, we removed complex variants (non-SNVs), variants that resulted in more than one isoform due to transcriptional consequences, retained only one occurrence, and removed redundant variants via a custom Python script.

Identification of clinically relevant germline SNVs

Human variants with clinical significance (genome assembly version GRCh38) were obtained from the public database ClinVar, a curated repository of human variants with clinical relevance and various phenotypic and functional metadata [7]. We utilized BCFtools

[30] for all filtering procedures. Variants were filtered on the basis of their origin, clinical consequence, and review status and were divided into two distinct datasets according to the following criteria: A) ClinVar benign [*Origin*: Germline, *clinical consequence*: benign/likely benign, and *review status:* 4 stars: practice guideline, 3 stars: reviewed by expert panel, 2 stars: criteria provided, multiple submitters, no conflicts]. B) ClinVar pathogenic [*Origin*: Germline, *clinical consequence*: pathogenic/likely pathogenic, and *review status*: 4 stars: practice guidelines, 3 stars: reviewed by expert panels, 2 stars: criteria provided, multiple submitters, no conflicts]. After filtering, complex variants (non-SNVs) and redundant variants were removed.

Variant annotation

We used the Ensembl Variant Effect Predictor (VEP) [31] for variant annotation as follows. Each of the five distinct datasets was processed via VEP. We included the alphamissense plugin to predict the functional consequences of missense variants [32]. The severity of consequences was categorized on the basis of genomic region, most severe coding consequence, and expected effect size of functional impact, as described below:

Approach 1: Coding consequence categorization by genomic region: *1) splicing site: splice donor variant, splice acceptor variant, splice region variant, splice donor 5th base, splice donor region, and splice polypyrimidine tract. 2) Intron: Intron variant. 3) Intergenic: regulatory region, TF binding site, intergenic variant, upstream gene variant, and downstream gene variant. 4) Exons: synonymous variant, missense variant, inframe insertion, inframe deletion, stop gained, frameshift variant, coding sequence variant, start retained variant, start lost, stop lost, stop retained variant, and incomplete terminal codon variant. 5) UTR: 5 prime UTR variant, 3 prime UTR variant. 6) OTHERS: transcript ablation, transcript amplification, protein altering variant, mature miRNA variant, NMD transcript variant, noncoding, transcript exon variant, noncoding transcript variant, TFBS*

ablation/amplification, feature elongation, feature truncation, coding transcript, variant, and sequence variant. We applied Pearson's chi-square test to evaluate the associations between genomic regions and datasets. A contingency table was created for the whole genome (WG), and the frequency of each genomic region across the datasets was recorded. Pearson's chi-square test was then used to determine if the distribution of genomic regions differed significantly between datasets, with a p value < 0.05 indicating a significant association. The analysis was conducted via R (version 4.3.0). The null hypothesis (H_0) was that there was no association between the genomic regions and the datasets ($p \geq 0.05$). A significant p value ($p < 0.05$) indicates that the distribution of genomic regions varies significantly between datasets, suggesting that some datasets are enriched or underrepresented in specific genomic regions.

Approach 2: The most severe coding consequences: *1) stop gained, 2) missense variant, 3) synonymous variant, 4) start and stop lost, 5) stop retained, and 6) splice variant*

Approach 3: Coding consequence categorization by the magnitude of functional impact: *1) HIGH: transcript ablation, splice acceptor variant, splice donor variant, stop gained, frameshift variant, stop lost, start lost, transcript amplification, feature elongation, feature truncation; 2) MODERATE: inframe insertion, inframe deletion, missense variant, and protein altering variant; 3) LOW: splice donor 5th base variant, splice region variant, splice donor region variant, splice polypyrimidine tract variant, incomplete terminal codon variant, start retained variant, stop retained variant, and synonymous variant.*

Subsetting genomic data for variants in coding regions

We obtained a set of exonic coordinates representing the canonical biological knowledge of each protein-coding locus in the human genome from MANE (matched annotation from NCBI and EBI) [33]. These coordinates were used to subset the annotated VCF files

generated in previous steps for variants found on translated exons (CDS). We analyzed the transition (TI) and transversion (TV) rates and the TI_TV ratio for each CDS coordinate set, comparing the alternative alleles with the reference. These analyses were carried out in R via custom scripts.

Evaluation of Genomic and CDS Coordinate Intersections

We used Fisher's exact test as implemented in BEDTools ("fisher" subcommand) to assess the significance of overlap among distinct classes of variants [34]. We used the tool's default settings together with two input files: 1) BED files containing the genomic coordinates of the SNVs from different classes obtained in this study and 2) a text file describing the autosomal chromosome sizes (GRCh38). For the CDS data obtained from the MANE subset, we initially used VEP annotation to obtain the CDS coordinate of each annotated variant via a custom Perl script. We gathered the CDS length information from the GFF file describing the MANE subset. We evaluated the intersection of these datasets by using three input files: two BED files containing the CDS coordinates of variants and a text file containing the length of each CDS. For both analyses, we calculated the proportion of intersections and evaluated the association between the genomic coordinates of each pair of datasets on the basis of the odds ratio value.

Overlapping allele comparative analysis

We performed a comparative analysis of alleles in the intersection regions between two sets of genomic coordinates representing different variant classes to determine whether alternative alleles were identical or distinct. Additionally, we analyzed nucleotide substitution patterns by evaluating the changes from the reference allele (e.g., T>G, A>C, C>G, etc.)

and comparing alternative alleles to the reference allele. These analyses were conducted for all subsets of SNVs co-occurring at the same positions.

To assess whether the frequency of identical alleles was higher than expected by chance, we constructed contingency tables for each subset of overlapping SNVs. Expected frequencies were calculated via a randomness model, assuming a 25% chance of allele identity (1/4) and a 75% chance of allele identity (3/4) for randomly distributed variants. Fisher's exact test (fisher.test()) from the R stats package was applied to these data to evaluate statistical significance. The expected frequencies were calculated as follows: identical alleles = intersection × P(0.25), discordant alleles = intersection × P(0.75).

Coordinates for the CpG islands were obtained from the UCSC Genome Browser database. All analyses were conducted in R via custom scripts.

Mutational signatures

To perform mutational signature analysis, we used a VCF file filtered from the original VCF provided by COSMIC, retaining only SNVs located at coordinates of interest identified from the intersection between COSMIC and SNP variants. This intersection was generated via the BEDTools intersect function and a BED file containing overlapping SNV coordinates from both datasets. The resulting VCF was then analyzed for mutational signatures via the website version of the COSMIC SigProfiler assignment tool, with COSMIC v3.4 reference signatures and whole-genome sequencing parameters to identify and assign known mutational signatures [35, 36].

Functional enrichment analysis

The functional enrichment analysis was conducted via the Over-Representation Analysis (ORA) method via the WebGestalt tool. Genes were selected on the basis of mutations that coincided with the genomic position, with each occurrence considered only once, resulting in the analysis of a total of 6,149 genes.

The evaluated categories included KEGG metabolic pathways and genomic locations on the basis of cytogenetic bands. Statistical significance was assessed via a multiple testing correction method based on the false discovery rate (FDR), with an FDR < 0.05 adopted as the criterion for identifying significantly enriched categories. The files generated from the enrichment analysis were used to produce graphs via R.

RESULTS AND DISCUSSION

We started our characterization of the mutational landscape of the human genome by gathering high-quality SNV mutations from curated datasets. Specifically, we collected SNV data from three databases encompassing a rich set of human genomic variants, namely, COSMIC, HGDP and ClinVar, as described below.

Dataset Description

We started our analysis with 75,385,302 variants classified as high-quality regarding variant calling that were obtained from 929 individuals sourced from the Human Genome Diversity Project (HGDP) [8]. This dataset represents germline variants from 54 diverse human populations across seven major superpopulations: Africa, America, Central Asia, South Asia, East Asia, Europe, the Middle East, and Oceania. To visualize the genetic patterns of individuals belonging to the seven superpopulations evaluated, a population structure

approach combining two techniques was applied: principal component analysis (PCA) and uniform manifold approximation and projection (UMAP). Initially, PCA was used to reduce the dimensionality of the data and capture global variation (**Figure 1-A**). Subsequently, UMAP was applied to preserve complex structures and identify proximity relationships in this lower-dimensional space (**Figure 1-B**). We considered UMAP a better resource to define isolated populations from the superpopulations. We proceed by generating two datasets of germline SNVs from the HGDP data by applying stringent criteria (see “Methods”). The first consists of polymorphic variants, defined as those with a global AF equal to or greater than 1% and present in at least two of the seven isolated populations established by PCA-UMAP. This criterion was adopted to avoid the inclusion of single population-specific mutations, which might be present at high frequency due to evolutionary mechanisms such as genetic drift, founder effects, or inbreeding [37]. This resulted in a set of 2,140,477 polymorphic SNVs across the whole genome (**Supplementary Table 1** - HGDP: *SNPs*). The second dataset consisted of rare variants, defined as those with an AF less than 1% when all 894 individuals from the HGDP dataset were considered. This set resulted in a total of 50,239,810 rare SNVs (**Supplementary Table 1** - HGDP: *Rares*).

The COSMIC database was used as a source of somatic SNV mutations recurrently found in cancer. This database provides two separate VCF files with variants from two different classes: one for variants occurring in coding regions of the human genome (37,261,323 entries) and another for noncoding regions (31,368,583 entries). Our final processing pipeline to gather only one SNV per locus detected 10,587,546 sites (**Supplementary Table 1** - *COSMIC*).

The initial ClinVar dataset consisted of a VCF file containing 1,303,558 variants. We processed this file to extract only germline variants with reliable review status, categorizing them as either pathogenic (16,313 SNVs) or benign (76,971 SNVs) (**Supplementary Table 1** – *Pathogenic* - *Benign*). Despite being arguably neutral from an evolutionary perspective,

we refer to these variants as benign, whereas the term 'neutral' used in this article is reserved for the set of polymorphic variants from HGDP (SNPs).

Exonic bias in clinically relevant variants

We used VEP to annotate the SNVs from the different datasets and obtain an unbiased profile of their distribution and genomic context across the human genome. We observed a greater proportion of SNVs in intronic regions (approximately 60%) for three datasets: COSMIC, SNPs and Rare (**Figure 2.A**, **Supplementary Table 2**). In contrast, we observed an enrichment of SNVs in exonic regions for the Benign and Pathogenic datasets from ClinVar. The Pathogenic dataset comprised approximately 81% exonic SNVs, whereas the Benign dataset reached approximately 78% exonic SNVs (**Figure 2.A, Supplementary Table 2**). Owing to the association of ClinVar variants with high-penetrance, clinically relevant conditions, this database predominantly represents variants occurring in coding regions, as the majority of these conditions are causally associated with protein-coding genes [7]. In comparison, the COSMIC database showed approximately 36% of variants within exons, whereas for the SNPs and Rare datasets, this value was only 1%. Pearson's chi-square test for independence revealed a significant association between the genomic region and datasets in the whole-genome (WG) analysis (χ^2 = 21,977.078; degrees of freedom = 20; p = 2.2e-16). This finding indicates that the distribution of SNVs in the genomic regions varies significantly between datasets, suggesting heterogeneity in the represented sets. Some sets may be enriched for specific genomic regions, potentially influencing downstream comparisons.

To remove potential co-occurrence biases caused by the overrepresentation of exonic SNVs in the pathogenic and benign datasets, we subset all datasets via the MANE project (Matched Annotation from NCBI and EBI), which defines a representative set of transcripts for protein-coding genes of the human genome. As expected, this procedure generated a

substantial reduction in the quantities of intronic and intergenic variants, as well as an increase in the similarity of distribution patterns across all datasets, with exonic SNVs comprising approximately 99% of them (**Figure 2.B, Supplementary Table 2**).

Distinct transition/transversion ratios of variants with diverse functional and evolutionary consequences

The transition/transversion ratio (Ti/Tv) is a classic metric used in genetic studies to measure genomic conservation and as a quality control tool [38]. In a scenario with no selective pressure and assuming an equal likelihood of transitions and transversions, the Ti/Tv ratio for the complete human genome would be approximately 0.5. However, owing to biochemical phenomena such as the spontaneous deamination of methylated cytosine to thymine (i.e., a transition), the expected ratio increases to approximately two. Transversions are also less common in exons, as these mutations are potentially more disruptive because of their higher probability of causing nonsynonymous amino acid changes and potentially impacting protein functionality [39, 40]. Consequently, the Ti/Tv ratio is even more pronounced in exonic regions (approximately three), which underscores the selective pressures acting upon these regions [38, 41]. In general, high Ti/Tv ratios reflect a conservative evolutionary force that minimizes protein alterations, whereas lower ratios suggest a higher incidence of mutations that are likely to disrupt protein function. As such, this metric is a useful proxy for comparing mutational profiles across datasets and understanding the underlying mutation patterns.

We computed the Ti/Tv ratios for the five CDS datasets and found high variability across datasets that largely reflects the expected variant classes found within each dataset. Specifically, the “benign” and “SNP” datasets, which are expected to contain benign/neutral variants, presented the highest ratios (4.38 and 3.35, respectively), suggesting selection for genomic conservation. In contrast, the datasets expected to contain pathogenic variants

("Rares", "Pathogenic" and "COSMIC") had lower values (2.87, 2.19 and 1.71, respectively), indicating a prevalence of potentially disruptive transversions (**Supplementary Table 3**). Together, these results suggest that the variants from the distinct datasets are likely to have distinct functional and evolutionary consequences.

Functional annotation of SNVs

The functional annotation of the five datasets provided further evidence of their distinctive functional roles. The analysis of the "Pathogenic" dataset revealed a predominance of two main functional consequences of potentially severe impacts: approximately 51.1% of these variants resulted in the premature insertion of stop codons (**Figure 3-A -** *Pathogenic* and **Supplementary Table 4**), and 47.1% of the variants were identified as missense. Such modifications are considered to have a high to moderate impact on cellular functionality according to VEP (**Figure 3-B -** *Pathogenic*), and the associations of those two coding consequences with various monogenic diseases or risk factors for complex diseases are well documented in the literature [42–48]. In contrast with the pathogenic variants, those classified as benign in ClinVar predominantly exhibited synonymous functional effects, accounting for 71.1% of the total (**Figure 3-A -** *Benign*). An additional 27.3% constituted missense variants, suggesting that, in this context, such changes may have a low to moderate functional impact (**Figure 3-B -** *Benign*). The "Rare" dataset also contains a predominance of missense variants (59.5%) (**Figure 3** - *Rare*), which might reflect deleterious variants of low frequency due to purifying selection [9, 49, 50]. This dataset also includes 38.2% synonymous variants with a potentially lower impact, possibly representing *de novo* variants in the human population. The SNPs exhibited a more balanced distribution of functional consequences, with 49.7% being missense and 48.6% being synonymous (**Figure 3-A** - *SNPs*). As expected, the "SNPs" dataset also displayed a lower incidence of variants with high functional impact (**Figure 3-B** - *SNPs*). The COSMIC dataset has a

prevalence of missense variants (69.4%) (**Figure 3-A -** *COSMIC*). Furthermore, 69% of the COSMIC variants have moderate effects, aligning with the multifactorial genetic nature of cancer, which involves interactions among multiple mutations, typically with synergistic effects (epistasis) [51] (**Figure 3-B -** *COSMIC*). Additionally, 24.9% of the COSMIC variants are synonymous. The understanding that variants influencing complex diseases have smaller effects than those observed in rare monogenic/oligogenic diseases highlights the need for a multifaceted analysis to comprehend the etiology of these conditions [10]. This detailed overview of the functional characteristics of variants highlights the complexity of genotype–phenotype interactions and the high prevalence of missense variants across all datasets. These findings also reinforce the importance of integrative approaches in investigating the underlying mechanisms of genetic and complex diseases.

Analysis of missense variants via AlphaMissense reveals patterns of functional impact

Missense variants may have a wide range of causal impacts, ranging from substitutions that do not significantly affect protein folding, activity, or stability to substitutions that are major causal factors for monogenic diseases. AlphaMissense is a tailored deep learning model that uses AlphaFold2-predicted structures to accurately compute a pathogenicity score of missense SNVs [32]. We recovered this pathogenicity score from the VEP outputs and evaluated their distributions across the five SNV datasets (**Figure 4-A**). This quantitative analysis revealed a similar pattern to the distribution of functional impact data across qualitative categories. The Pathogenic dataset exhibited a left-skewed distribution (**Figure 4-A** - *Pathogenic*), with a higher frequency of pathogenic variants with higher AlphaMissense scores, which is consistent with conditions of high penetrance, such as monogenic and oligogenic diseases. In contrast, benign variants from ClinVar showed the opposite distribution: a right-skewed distribution, with a concentration of variants with low

pathogenicity scores (**Figure 4-A** - *Benign*). This result has already been reported in the validation of AlphaMissense [32].

For COSMIC variants, we observed a bimodal distribution, with a peak of variants with low scores and a peak of high scores (**Figure 4-A** - *COSMIC*). This distribution may reflect the presence of passenger and driver mutations. Passenger mutations occur in genomic loci intrinsically prone to mutations during cancer progression but do not contribute to the establishment of cancer and are not under positive selection. In contrast, genes considered "drivers" have a high phenotypic impact, as they are involved in cancer development and progression. These genes contain mutations that confer a selective advantage to cancer cells by, e.g., promoting uncontrolled growth and metastatic capabilities. Mutations in driver genes are often associated with significant changes in protein function, directly contributing to the malignant characteristics of cancer cells [52]. Therefore, these results are in line with the genomic properties of complex diseases and of recurrent somatic mutations in cancers.

The SNPs exhibited a right-skewed distribution (**Figure 4-A** - *SNPs*), with a concentration of variants showing low pathogenicity scores, a profile similar to that observed for benign variants in ClinVar. Therefore, despite the greater quantity of missense variants among SNPs compared with the Benign dataset, our findings suggest that most of these variants are expected to be neutral/quasi neutral This finding is also in agreement with the restrictive filters applied to generate the SNP database to remove loci with evidence of selection or founder effects. Rare variants also displayed a right-skewed distribution (**Figure 4-A** - *Rares*), albeit with a small peak of variants with smaller pathogenicity scores. The Rares set represents a range of scores consistent with low-frequency SNVs in the human population, as it might include both deleterious variants with reduced frequencies due to natural selection and *de novo* variants with neutral or quasi neutral phenotypic consequences.

A detailed analysis of missense variants via AlphaMissense revealed distinct patterns of functional impacts of SNVs, providing a broader view of the genetic diversity of different

sets. The Pathogenic set exhibited predominantly likely pathogenic missense variants (78.5%), in contrast to the Benign set, which presented 89.5% likely benign variants, which is consistent with the curation conditions of ClinVar (**Figure 4-B** and **Supplementary Table 5**). The Benign and SNP sets showed similar patterns, with ~90% of SNVs being likely benign, as both sets represent neutral or nearly neutral variants from an evolutionary perspective. Although Rares also exhibited a distribution of low scores and 80% of likely benign variants, a moderate peak (11.5%) of high-scoring variants reported as likely pathogenic. COSMIC also showed a more complex distribution, with ~60% of the SNVs being likely benign and ~29% being likely pathogenic. These findings emphasize the importance of considering not only the presence of missense variants but also their potential functional impacts on human health via genomic studies. Additionally, the application of AlphaMissense was important for predicting the pathogenicity of missense SNVs that may present a wide range of possible impacts, highlighting its potential contribution to the study of variants and genetic disease diagnosis.

Co-occurrence landscape of SNVs with distinct evolutionary origins and functional consequences

The human genome is a complex landscape where mutations occur and exert varying degrees of influence on biological functions and disease phenotypes. Understanding the spatial distribution of these mutations — whether pathogenic or benign — is essential for deciphering the underlying mechanisms of genetic disorders and advancing precision medicine. In this study, we introduce three potential models of spatial mutation distribution, providing a framework to investigate their distribution and functional impacts.

Model 1 hypothesizes that mutations of different classes occur at distinct loci (**Figure 5**). This model allows us to explore the hypothesis that spatial separation in the genome can be correlated with distinct functional outcomes or evolutionary paths. Model 2 proposes that

while mutations occur at the same genomic positions, the alleles differ, suggesting a nuanced interplay between mutation type and allele-specific effects. Finally, Model 3 suggests that both benign and pathogenic mutations not only share loci but also involve identical alleles, raising questions about the differential context or regulatory mechanisms that might dictate the impact of a mutation on the organism.

The development of these models is driven by the need to better understand the genomic context of mutations and their potential collective implications. By systematically comparing these models across various genomic datasets, our research aims to increase the number of new studies on key aspects of genomic stability, mutation hotspots, and their roles in human health and disease. In the following, we discuss the experiments designed to test these models and the implications of our findings in the broader context of genetic research and clinical practice.

After we produced well-characterized and robust sets of SNVs, we proceeded by evaluating their genomic co-occurrence patterns. For this purpose, an analytical approach was employed via Fisher's exact test, as implemented in the BEDTools suite [34]. We generated contingency tables for all pairs of comparisons. The outputs of these tests could result in three distinct models, as depicted in **Figure 5**. Among the ten possible pairwise combinations tested from the five sets of coordinates, eight revealed intersections of the gene coordinates of SNVs, except for the tests between SNPs and Rares, as these are mutually exclusive, and between SNPs and Pathogenic (**Table 1**). Although we did not further investigate this result, the absence of intersection of the SNP and Pathogenic sets is probably due to the distinct nature of the two sets, since one is composed mainly of SNVs with large effects and low allele frequencies (Pathogenic) and the other contains polymorphic variants with low impact and high allele frequencies (SNPs).

Within the eight pairwise tests with overlaps, in five comparisons, we observed that the odds ratio changed in the same direction for the WG and CDS datasets (**Table 1**). The exceptions

are the COSMIC vs. Pathogenic and COSMIC vs. Benign comparisons, which have positive associations when WG coordinates are compared with negative associations for the CDS datasets, whereas Rares vs. COSMIC shows the opposite pattern. The COSMIC, Benign and Pathogenic datasets share a strong bias for variants within coding regions (**Figure 2**). The results of the genomic coordinates for these datasets are likely to have a greater overlap of variants in coding regions, which could bias downstream conclusions. On the other hand, COSMIC and Rares, despite sharing an intronic bias, did not show significant overlap for WG, which is a less consistent result. For those reasons, we considered the CDS data more robust for evaluating the statistics of variants that co-occur across datasets.

All nine CDS tests revealed a significant association (p values < 0.05, **Supplementary Table 6**). Four tests indicated a negative association with odds ratios < 1 (SNPs vs. Pathogenic, Rares vs. Pathogenic, Cosmic vs. Benign, Cosmic vs. Pathogenic), suggesting that these mutations occur in different positions more frequently than expected by chance. For these comparisons, we argue that Model 1 better describes this pattern of nonco-occurrence (**Figure 5**).

Conversely, five pairwise comparisons (SNPs vs. Cosmic, SNPs vs. Benign, Rares vs. Cosmic, Rares vs. Benign, Benign vs. Pathogenic) revealed positive associations, with odds ratios greater than 1 and a significant number of intersections, suggesting that these mutation classes co-occur at the same genomic loci more frequently than expected by chance (odds ratios ranging from 1.3 to 66, **Table 1** and **Supplementary Table 6**). We further examined whether variants sharing genomic positions in these comparisons also shared more identical alleles than expected by chance (Methods, section "*Overlapped allele comparative analysis*") to determine whether Model 2 or 3 better explains the distribution of these SNVs (**Figure 5**). The results revealed that identical alleles occurred more frequently than expected by chance in most comparisons with significant positive associations, except for the Benign vs. Pathogenic sets. The strongest association was observed between SNPs and Benign (OR = ∞, p = 0), with 100% of overlapping variants sharing identical alleles.

Similarly, SNPs and COSMIC variants showed a significant positive association (OR = 129.34, p = 0), with 97.73% identical alleles. The Rares vs. Benign comparison also exhibited a strong association (OR = 162.11, p = 0), with 98.18% identical alleles. Conversely, no identical alleles were found between Benign and Pathogenic (OR = 0), indicating distinct mutational signatures at the evaluated loci.

SNPs vs. Benign - We found 2728 variants co-occurring in the same CDS coordinates, an intersection approximately 66 times greater than expected by chance, marking this as the most significant result in our analyses (**Table 1** and **Supplementary Table 6**). Furthermore, our analyses revealed not only a substantial intersection of genomic coordinates but also the highest fraction of shared alleles among the overlapping loci from the different databases (100% of identity, **Supplementary Table 7**). The significant odds ratio, the substantial overlap and high allele similarity observed between SNPs and benign variants validate our methodology, which can be represented by *Model 3* - **Figure 5**. Together, these observations provide a robust validation of our strategy to select these variants, as these two sets are expected to be enriched in neutral/quasi neutral variants.

Pathogenic vs. Benign - The co-occurrence of Pathogenic and Benign variants yielded a significant odds ratio value of 2.1 (**Table 1**), with an intersection of 192 and, in contrast with the previous results, 100% allele dissimilarity (**Supplementary Table 7**). Variants classified as benign in ClinVar are those that, on the basis of available evidence, have not been associated with diseases or pathological conditions. Therefore, many of these benign variants are expected to be defined in opposition to a pathogenic and distinct variant that occurs at the same locus. Thus, although these two datasets occur approximately two times more than what is expected by chance, we argue that this is likely to be an artifact caused by the procedures used to define these variant sets by ClinVar. In summary, these results highlight the clear distinction between these two groups of SNVs regarding their association with diseases. The high allele dissimilarity observed reinforces the robustness of classifying

these variants as pathogenic with high penetrance or neutral/quasi neutral, respectively, and supports *Model 2* of the mutation landscape.

Rares vs. Benign - Rare and Benign have genomic coordinates that co-occur 16.2 times more frequently than expected by chance (**Table 1**), with an intersection of 17,777 loci and 98% shared alleles (**Supplementary Table 7**). Like the SNPs and Benign datasets, the Rare dataset contains a high proportion of predicted neutral or low-impact germline variants likely to be *de novo* mutations with no significant selective pressure (**Figure 3-B** and **Figure 4-B**) [53]. This suggests that the co-occurrence of Rares and Benign, both of which are germline variants, is not random and may be influenced by shared evolutionary processes or genetic mechanisms, which can be better explained by *Model 3* - **Figure 5**.

SNPs vs. COSMIC – The co-occurrence of variants from the SNP and COSMIC databases is 5.6 times greater than expected by chance (**Table 1**). Interestingly, approximately 97.7% of the overlapping SNVs shared the same alternative allele (**Supplementary Table 7**). A high overlap was also observed in this paired test, where 9,575 of the 16,040 polymorphic SNVs, approximately 60%, co-occur at the same gene coordinates as cancer-related somatic mutations. The nonrandom distribution of both polymorphic and cancer-related mutations points to the presence of mutational hotspots and possibly shared mutational mechanisms, yet unidentified. This landscape can be better represented by *Model 3* - **Figure 5**, where neutral and pathogenic mutations significantly occur at the same locus while sharing the same alleles. SNPs were the only germline set that showed a significant association with cancer-related somatic SNVs for both the CDS and WG. In this association, the majority of the overlapping SNVs in the WG were intronic, accounting for approximately 70%, suggesting a promising starting point for studies on noncoding variants (**Supplementary Table 9**). Additionally, synonymous and missense mutations were the most abundant mutation types in the SNP vs. COSMIC co-occurrence for both the WG and the CDS (**Supplementary Tables 9 and 10**). Pathogenicity scores for missense SNVs showed a peak of low scores, together with a high number of synonymous SNVs (**Supplementary**

Figure 2). These low-pathogenicity missense SNVs may represent passenger mutations in cancer. Owing to the mutational, biochemical, and histological heterogeneity of tumors, passenger mutations, although they do not directly contribute to cancer development, such as driver mutations, may, in specific situations, indirectly contribute to the establishment or modulation of the tumor microenvironment and are thus considered “weak drivers” [9, 54]. A slight proportion of mutations that could represent driver mutations in cancer, such as stop-gained, start-lost, and stop-lost mutations, were also observed in the overlap (**Supplementary Tables 9 and 10**). In general, loss-of-function mutations might indicate driver mutations, which could be better approached by cancer driver mutation predictors [55].

Rare vs. COSMIC – This comparison revealed a significant overlap of SNV gene coordinates, 1.3 times greater than expected by chance, with an allele similarity of 77% (**Table 1**, **Supplementary Table 7**). Similar to the SNP vs. COSMIC case, we hypothesize that a nonrandom distribution may exist between rare germline mutations and cancer-related somatic mutations, potentially associated with yet unidentified shared mutational mechanisms. This case would also be better represented by *Model 3* (**Figure 5**). However, this analysis revealed an intriguing result of a relatively weak positive association for CDS and a negative association for WG, despite the shared genomic position bias between the WG datasets. To further explore our findings, we filtered the rare variants by linkage disequilibrium and repeated the tests. After LD filtering, we observed a negative association for both CDS and WG in the *Rare vs. COSMIC* comparison (WG OR = 0.67, CDS OR = 0.522), whereas the direction of association for the other pairwise tests remained consistent. This suggests that LD patterns may influence the statistical outcome in certain comparisons, reinforcing the complexity of these mutational landscapes.

Together, these significant overlaps suggest that one or more nonrandom mutational mechanisms may drive the greater than expected co-occurrences of SNVs from the different classes. Of special interest is the overlap between SNPs and COSMIC, as this is the only

significant association of SNVs for both CDS and WG (**Table 1**), which have distinct origins (somatic and germline), functional effects (neutral/benign and pathogenic) and evolutionary consequences (SNP and cancer disease). Overall, we propose that the overlapping variants may occur in mutational hotspots because of shared yet unknown molecular mechanisms.

Low Intersection between SNVs of subsets and CpG Islands

We further investigated the subsets of SNVs that share genomic coordinates by evaluating their co-occurrence with CpG islands, arguably one of the main types of mutational hotspots. CpG islands are also important epigenetic markers that are often found near gene transcription start sites in promoter regions [20]. They are susceptible to the spontaneous deamination of methylated cytosine (5-methylcytosine) to thymine, generating a T/G mismatch. When not repaired, such mismatches introduce C/G–T/A substitutions during DNA replication in daughter cells that inherit the template strand harboring the mutated thymine [24]. These spontaneous deaminations are responsible for the highest single nucleotide substitution rate of C>T and their complementary G>A in both germline and somatic lineages [3].

First, we evaluated the pattern of single nucleotide substitutions across all subsets. The most frequent substitutions identified were C>T and the complementary G>A (**Figure 6** and **Supplementary Figure 1**), particularly for nucleotides shared between the different sets (**Figure 6-C**, **Supplementary Figure 1-C**, and **Supplementary Table 8** - *Shared*). With respect to the *SNP vs. COSMIC* overlapping subset, C>T/G>A substitutions accounted for approximately 25% of the total substitutions identified (**Supplementary Table 8**). This finding raises the question of whether the mechanism of spontaneous deamination associated with CpG islands significantly contributes to the prevalence of these mutations. To investigate this, we used Fisher's exact test implemented in BEDTools and calculated the proportion of genomic coordinate co-occurrence between the SNP vs. COSMIC overlaps

and CpG islands. Although we observed a high odds ratio of 7, indicating that co-occurrence was more frequent than expected by chance, only about 5% (1623 of 34725 SNVs) of the mutations could be attributed to CpG island-associated mechanisms (**Supplementary Table 6 -** *Subset_SNP_intersect_COSMIC vs. CpG Island*). This finding suggests that while a significant number of SNVs co-occur in CpG islands, cytosine deamination in these regions is unlikely to be the primary driver of these mutations, indicating the presence of other mutational hotspots.

Signatures of mutational overlap between SNPs and cancer-related mutations

Mutational signatures are specific mutation patterns linked to particular mutagenic processes or DNA repair deficiencies [56, 57]. In this study, we identified SBS5 (73.01%), SBS1 (15.48%) and SBS54 (11.52%) as predominant signatures in WG (**Figure 7**) and likely enriched them in intronic and intergenic regions, which may have masked less prevalent signatures. When limiting the analysis to CDS regions, in addition to SBS5 (32.27%), SBS1 (12.87%) and SBS54 (15,3%), two additional signatures emerged: SBS6 (13,04%) and SBS87 (21,51%) (**Figure 8**). The SBS5 signature was the most abundant for both WG and CDS sets.

SBS1 and SBS5 are ubiquitous mutational signatures that exhibit a clock-like pattern, with mutation numbers increasing with age [58]. SBS1 is associated with the deamination of 5-methylcytosine, a process prevalent in CpG islands. In our investigation, the co-occurrence of the SNP vs. COSMIC subset with CpG islands was observed to be 7 times greater than expected. However, this accounts for only approximately 5% of the overlapping SNVs, indicating that deamination does not explain the majority of these mutations. Although the etiology of SBS5 remains unknown, it has been shown to contribute to TP53 mutations, compromising its tumor suppressor role and driving genomic instability, immune evasion, treatment resistance, and metastasis [58, 59]. Previously associated with early

postzygotic somatic mutations in monozygotic twins [60], SBS5 is now observed in SNPs overlapping cancer mutations, suggesting a shared mechanism affecting both germline and somatic contexts. SBS6 and SBS54 have been linked to microsatellite instability (MSI) and DNA repair deficiency (dMMR) and have been reported in MSI-H cancers such as colorectal and gastric, Lynch syndrome, and other tumors [61–66]. Microsatellite instability regions, often affected by SBS6 and SBS54, may harbor germline variants that influence their susceptibility to subsequent somatic mutations. The presence of SNPs in these regions could serve as an indicator of genomic fragility. A fourfold higher-than-expected co-occurrence between microsatellites and SNPs vs. COSMIC regions was observed, suggesting that microsatellites can influence SBS6 and SBS54 occurrence, although this accounts for only 0.1% of the mutations (45 overlapping SNVs - **Supplementary Table 6 -** *Subset_SNP_intersect_COSMIC vs. Microsatellites*). SBS87, a somatic mutational signature typically induced by chemotherapy with thiopurines and experimentally validated [67, 68], exhibited a 21.51% correspondence in the SNPs vs. COSMIC subset, which includes variants present in both the somatic and germline sets. SBS87, along with SBS1, has previously been reported to be highly prevalent in somatic variants that share loci with germline mutations [25].

These findings, marking the first association of all five signatures with germline mutations, underscore a potential shared mechanism influencing germline and somatic contexts and support a nonrandom mutational distribution across uncharacterized hotspots.

Germline-somatic hotspots in the human genome showed an evidence of enrichment in cancer-related regions and pathways

Our analyses revealed an intriguing overlap of genomic loci between germline variants with high allele frequency (AF ≥ 1%) and somatic mutations associated with cancer, with most of these variants sharing the same alternative allele. The mechanisms responsible for this

phenomenon remain unexplained. To expand our understanding of this mutational pattern, we investigated the genomic regions, pathways, and genes most affected by these shared mutations through an Over-Representation Analysis (ORA) based on cytogenetic bands and KEGG pathways.

The intersection between SNP coordinates and COSMIC mutations revealed that 6,149 genes were affected in both contexts (**Figure 9-A**). The regions 13q12.12, 19p13.3, 19p13.12, 5q31.3, 22q13.33, and 6q27 and the chrHG2308_PATCH segment were significantly enriched (**Figure 9-B, Supplementary Table 13**). In particular, the enrichment of 19p13.12 (ratio = 1.7, FDR = 0.018) and 19p13.3 (ratio = 1.46, FDR = 0.002) suggested that the short arm of chromosome 19 may serve as a reservoir of SNPs at loci prone to somatic mutations.

KEGG pathway analysis revealed significant enrichment in multiple biological processes (**Figure 9-C, Supplementary Table 14**). The most relevant pathways included ECM-receptor interaction, the PI3K-Akt signaling pathway, Human papillomavirus infection, and Focal adhesion, all of which were highly enriched (**Figure 9-D**). These pathways play key roles in regulating cell adhesion, extracellular matrix remodeling, proliferative signaling, and tumor-microenvironment interactions—critical processes in cancer progression. This association is driven by genes from multigenic families, such as COL (collagens), ITG (integrins), LAM (laminins), and TNC (tenascins), which regulate cell adhesion dynamics and intercellular communication. Other significantly enriched pathways included ABC transporters, complement and coagulation cascades, inositol phosphate metabolism, Protein digestion and absorption, Bile secretion, Phosphatidylinositol signaling system, and Motor proteins, all of which are linked to metabolic and homeostatic processes.

The enrichment analysis results provided new perception into the mutational patterns observed in the SNPs vs. COSMIC subset. While the mechanisms driving the overlap between SNPs and somatic mutations remain unknown, the presence of shared hotspots in

the short arm of chromosome 19 and genes from the PI3K/AKT/mTOR pathway raises important questions. Previous studies have indicated that germline variants in the 19p13.3 region are associated with an increased somatic mutation rate in PTEN, one of the key tumor suppressor genes involved in the negative regulation of the PI3K/AKT/mTOR pathway [69]. Constitutive activation of this pathway is associated with uncontrolled cell proliferation, apoptosis inhibition, and tumor progression. Moreover, germline variants in genes from this pathway have been linked to familial susceptibility to lung cancer [70]. However, the causes and mechanisms responsible for the presence of highly frequent SNPs at these genomic loci during germline development remain unexplored [3, 25, 69].

Among the enriched genes in the PI3K-AKT pathway, we identified MTOR, RPTOR, MLST8, and STK11, all of which directly modulate pathway activity [71]. Additionally, genes from the G protein subunit beta/gamma family (GNB3, GNB5, GNG8, and GNGT2) were detected, suggesting a potential role in mTOR regulation, albeit with a lesser impact than genes previously described, such as GNA11 and GβL [69, 71]. We also observed enrichment of TSC2 and TP53, which act as negative regulators of the mTOR pathway. The TSC1/TSC2 complex integrates signals from multiple cascades, including the PI3K/AKT, MAPK/ERK, and TNF-α cascades, to control mTORC1 activation. Moreover, TP53, known as the "guardian of the genome," inhibits mTORC1 by activating AMPK, PTEN, and TSC2, thereby preventing uncontrolled cell proliferation [71]. Other enriched genes included PIK3R1 and PIK3R2, which regulate PI3K activation in response to growth factors and oncogenes [72]. Additionally, frequently mutated proto-oncogenes in cancer, such as KIT, MET, MYC, and RET, along with growth factors and the receptors EGFR, FGFR, IGF1R, PDGFRB, GHR, and HGF, have been identified. The genes enriched in these pathways are highly relevant in the tumorigenic context and are often associated with a predisposition to cancer owing to germline mutations. However, from the perspective of highly frequent germline variants with low functional impact, the underlying mechanisms remain unclear.

The enriched ECM-receptor interaction and Focal adhesion pathways are involved in tissue remodeling and cell migration in tumors. The extracellular matrix (ECM) regulates cell adhesion and signaling, affecting tumor-microenvironment interactions. The Focal adhesion pathway connects the ECM to the cytoskeleton, modulating cell adhesion, mechanical signaling, and stress resistance—processes frequently dysregulated in cancer [73]. Another relevant finding was the enrichment of the Human papillomavirus (HPV) infection pathway, which is associated with PI3K/AKT/mTOR activation and promotes cell proliferation, apoptosis evasion, and therapy resistance [74, 75]. The association of this pathway with SNPs and somatic mutations at shared loci suggests a possible role of PI3K/AKT/mTOR signaling in the predisposition to somatic mutation induced by viral infection.

The presented results highlight an intriguing phenomenon that may be linked to specific, yet unexplored, genomic hotspot mechanisms causing mutations in loci shared between SNPs and somatic cancer mutations. The correlation between specific mutational signatures (SBS1, SBS5, SBS87, SBS54, and SBS6), enriched pathways (PI3K/AKT/mTOR, ECM-receptor interaction, Focal adhesion, and Human papillomavirus infection), and enriched genomic regions (13q12.12, 19p13.3, 19p13.12, 5q31.3, 22q13.33, 6q27, and the chrHG2308_PATCH segment) not only points to genomic patterns that warrant thorough investigation but also suggests that such interactions may be fundamental to understanding the mechanisms of cancer predisposition, particularly in the context of germline mutations.

CONCLUSION

The vast availability of human genomic data has substantially advanced our understanding of various mutation types, with the human genome serving as one of the most comprehensive and well-curated resources for analyzing patterns of mutation occurrence across different categories: 1) somatic *versus* germline variants; 2) polymorphic *versus* rare variants; and 3) pathogenic *versus* benign variants. However, the availability of high-quality

data describing human mutations of distinct evolutionary origins and with a variety of functional consequences is limited. This contrasts with the lack of integrative analyses to systematically characterize such mutations in a comparative way to unveil potential mechanisms underlying their distribution patterns. Our characterization of the five distinct sets of SNVs largely agrees with currently known biology, including their expected functional consequences and genomic distribution. Therefore, these findings provide strong evidence that our data provide a reliable representation of these distinct classes of mutations.

Interestingly, our search for biases in co-occurrence also revealed significant statistics for the overlap of somatic SNVs associated with cancer and germline polymorphisms, which, to our knowledge, has not been previously reported. Importantly, the positive association for this overlap remained significant considering both CDS and WG, and the vast majority of the overlapping coordinates shared the same allele in both datasets. Furthermore, the functional analysis of these variants revealed a lower impact relative to the overall distribution of COSMIC variants, suggesting that these shared alleles are more likely passenger mutations or weak drivers rather than primary driver mutations in cancer. However, these overlapping SNVs may exhibit a dual nature. While they may be neutral as germline variants, they could acquire pathogenic potential as somatic mutations in cancer cells, depending on the genomic context. This includes factors such as the age-related accumulation of additional mutations, epistatic interactions, and alterations in driver genes [76].

CpG islands, one of the most well-known mutational hotspots, and fragile genomic regions such as microsatellites account for only a small fraction of these overlapping mutations. This finding suggests that an as-yet-unknown shared mechanism may drive the co-occurrence of SNPs and cancer-related mutations, potentially contributing to the SBS1, SBS5, SBS6, SBS54, and SBS87 mutational signatures. The greater-than-expected recurrence of mutations at the same loci—resulting in identical alleles in both somatic and germline cells—and the predominance of SBS5 in the mutational signatures of these overlapping SNVs further support our hypothesis. We propose that shared yet unidentified mechanisms

may underlie the occurrence of germline and polymorphic SNVs, as well as cancer-associated somatic SNVs, within specific unexplored hotspots. Moreover, this mechanism is intriguingly associated with the occurrence of germline and somatic mutations in genes from the short arm of chromosome 19 and the PI3K/Akt/mTOR pathway.

LIST OF ABBREVIATIONS

AF	Allele Frequency
BED	Browser Extensible Data
CDS	Coding DNA Sequence
COSMIC	Catalogue of Somatic Mutations in Cancer
CpG	Cytosine-phosphate-Guanine
DDR	DNA Damage Response
dMMR	Deficient DNA Mismatch Repair
FDR	False Discovery Rate
HGDP	Human Genome Diversity Project
KEGG	Kyoto Encyclopedia of Genes and Genomes
MANE	Matched Annotation from NCBI and EBI
MSI	Microsatellite Instability
ORA	Over Representation Analysis
SBS	Single Base Substitution
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
UMAP	Uniform Manifold Approximation and Projection
UTR	Untranslated Region
VCF	Variant Call Format
VEP	Variant Effect Predictor

WG Whole Genome

DECLARATIONS

Ethics approval and consent to participate

Not applicable. This study does not involve human participants, human data, or human tissue.

Consent for publication

Not applicable. This manuscript does not contain any individual person's data.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

XXXXXXXX

Authors' contributions

TST conceived and designed the study, conducted the project, performed all data analyzes, interpreted the results, and wrote the manuscript. FPL contributed to the initial idea, provided theoretical and technical support, and assisted with some analyses and interpretations. DL, MDO, LM, and RP contributed to data analysis. CSC provided the HGDP dataset and contributed to the analysis. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Federal University of Minas Gerais for their support and contributions to this study.

REFERENCES

1. Akdemir KC, Le VT, Kim JM, Killcoyne S, King DA, Lin Y-P, et al. Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. Nat Genet. 2020;52:1178–88.

2. Seplyarskiy VB, Sunyaev S. The origin of human mutation in light of genomic data. Nat Rev Genet. 2021;22:672–86.

3. Meyerson W, Leisman J, Navarro FCP, Gerstein M. Origins and characterization of variants shared between databases of somatic and germline human mutations. BMC Bioinformatics. 2020;21:227.

4. Alekseenko IV, Kuzmich AI, Pleshkan VV, Tyulkina DV, Zinovyeva MV, Kostina MB, et al. The cause of cancer mutations: Improvable bad life or inevitable stochastic replication errors? Mol Biol. 2016;50:799–811.

5. Pathan N, Deng WQ, Di Scipio M, Khan M, Mao S, Morton RW, et al. A method to estimate the contribution of rare coding variants to complex trait heritability. Nat Commun. 2024;15:1245.

6. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer. 2004;91:355–8.

7. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar:

public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42 Database issue:D980-5.

8. Bergström A, McCarthy SA, Hui R, Almarri MA, Ayub Q, Danecek P, et al. Insights into human genetic variation and population history from 929 diverse genomes. Science. 2020;367:eaay5012.

9. Kumar S, Gerstein M. Unified views on variant impact across many diseases. Trends Genet. 2023;39:442–50.

10. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. Nature. 2020;577:179–89.

11. Kato GJ, Piel FB, Reid CD, Gaston MH, Ohene-Frempong K, Krishnamurti L, et al. Sickle cell disease. Nat Rev Dis Primers. 2018;4:18010.

12. Shteinberg M, Haq IJ, Polineni D, Davies JC. Cystic fibrosis. Lancet. 2021;397:2195–211.

13. Copson ER, Maishman TC, Tapper WJ, Cutress RI, Greville-Heygate S, Altman DG, et al. Germline BRCA mutation and outcome in young-onset breast cancer (POSH): a prospective cohort study. Lancet Oncol. 2018;19:169–80.

14. Viansone A, Pellegrino B, Omarini C, Pistelli M, Boggiani D, Sikokis A, et al. Prognostic significance of germline BRCA mutations in patients with HER2-POSITIVE breast cancer. Breast. 2022;65:145–50.

15. Perera D, Poulos RC, Shah A, Beck D, Pimanda JE, Wong JWH. Differential DNA repair underlies mutation hotspots at active promoters in cancer genomes. Nature. 2016;532:259–63.

16. Cagan A, Baez-Ortega A, Brzozowska N, Abascal F, Coorens THH, Sanders MA, et al. Somatic mutation rates scale with lifespan across mammals. Nature. 2022;604:517–24.

17. Risch NJ. Searching for genetic determinants in the new millennium. Nature. 2000;405:847–56.

18. Witt KE, Huerta-Sánchez E. Convergent evolution in human and domesticate adaptation to high-altitude environments. Philos Trans R Soc Lond B Biol Sci. 2019;374:20180235.

19. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nat Rev Genet. 2007;8:610–8.

20. Wheeler LC, Wing BA, Smith SD. Structure and contingency determine mutational hotspots for flower color evolution. Evol Lett. 2021;5:61–74.

21. Monroe JG, Srikant T, Carbonell-Bejerano P, Becker C, Lensink M, Exposito-Alonso M, et al. Mutation bias reflects natural selection in Arabidopsis thaliana. Nature. 2022;602:101–5.

22. Kilgour JM, Jia JL, Sarin KY. Review of the molecular genetics of basal cell carcinoma; Inherited susceptibility, somatic mutations, and targeted therapeutics. Cancers (Basel). 2021;13:3870.

23. Zou H, Pan T, Gao Y, Chen R, Li S, Guo J, et al. Pan-cancer assessment of mutational

landscape in intrinsically disordered hotspots reveals potential driver genes. Nucleic Acids Res. 2022;50:e49.

24. Nesta AV, Tafur D, Beck CR. Hotspots of human mutation. Trends Genet. 2021;37:717–29.

25. Wang M, Li SC, Shen B. Elevated incidence of somatic mutations at prevalent genetic sites. Brief Bioinform. 2024;25.

26. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.

27. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015;4:7.

28. Borda V, Alvim I, Mendes M, Silva-Carvalho C, Soares-Souza GB, Leal TP, et al. The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture. Proc Natl Acad Sci USA. 2020;117:32557–65.

29. Sakaue S, Hirata J, Kanai M, Suzuki K, Akiyama M, Lai Too C, et al. Dimensionality reduction reveals fine-scale structure in the Japanese population with consequences for polygenic risk prediction. Nat Commun. 2020;11:1569.

30. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:giab008.

31. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol. 2016;17.

32. Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science. 2023;381:eadg7492.

33. Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022;604:310–5.

34. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.

35. Díaz-Gay M, Vangara R, Barnes M, Wang X, Islam SMA, Vermes I, et al. Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. Bioinformatics. 2023;39.

36. Sondka Z, Dhir NB, Carvalho-Silva D, Jupe S, Madhumita, McLaren K, et al. COSMIC: a curated database of somatic variants and clinical data for cancer. Nucleic Acids Res. 2024;52:D1210–7.

37. Greenbaum G, Templeton AR, Zarmi Y, Bar-David S. Allelic richness following population founding events – A stochastic modeling framework incorporating gene flow and genetic drift. PLoS One. 2014;9:e115203.

38. Wang J, Raskin L, Samuels DC, Shyr Y, Guo Y. Genome measures used for quality control are dependent on gene function and ancestry. Bioinformatics. 2015;31:318–23.

39. Guo C, McDowell IC, Nodzenski M, Scholtens DM, Allen AS, Lowe WL, et al.

Transversions have larger regulatory effects than transitions. BMC Genomics. 2017;18.

40. Sun H, Yu G. New insights into the pathogenicity of non-synonymous variants through multi-level analysis. Sci Rep. 2019;9:1667.

41. Liu Q, Guo Y, Li J, Long J, Zhang B, Shyr Y. Steps to ensure accuracy in genotype and SNP calling from Illumina sequencing data. BMC Genomics. 2012;13 Suppl 8:S8.

42. Krumm N, Turner TN, Baker C, Vives L, Mohajeri K, Witherspoon K, et al. Excess of rare, inherited truncating mutations in autism. Nat Genet. 2015;47:582–8.

43. Pal LR, Moult J. Genetic basis of common human disease: Insight into the role of missense SNPs from genome-wide association studies. J Mol Biol. 2015;427:2271–89.

44. von Stedingk K, Stjernfelt K-J, Kvist A, Wahlström C, Kristoffersson U, Stenmark-Askmalm M, et al. Prevalence of germline pathogenic variants in 22 cancer susceptibility genes in Swedish pediatric cancer patients. Sci Rep. 2021;11:5307.

45. Fu R, Zhang J-T, Chen R-R, Li H, Tai Z-X, Lin H-X, et al. Identification of heritable rare variants associated with early-stage lung adenocarcinoma risk. Transl Lung Cancer Res. 2022;11:509–22.

46. Koprulu M, Zhao Y, Wheeler E, Dong L, Rocha N, Li C, et al. Identification of rare loss-of-function genetic variation regulating body fat distribution. J Clin Endocrinol Metab. 2022;107:1065–77.

47. Leggatt G, Cheng G, Narain S, Briseño-Roa L, Annereau J-P, Genomics England Research Consortium, et al. A genotype-to-phenotype approach suggests under-reporting of single nucleotide variants in nephrocystin-1 (NPHP1) related disease (UK 100,000 Genomes Project). Sci Rep. 2023;13:9369.

48. Venizelos A, Sorbye H, Elvebakken H, Perren A, Lothe IMB, Couvelard A, et al. Germline pathogenic variants in patients with high-grade gastroenteropancreatic neuroendocrine neoplasms. Endocr Relat Cancer. 2023;30.

49. Wright AF. Genetic Variation: Polymorphisms and Mutations. In: Encyclopedia of Life Sciences. Chichester: John Wiley & Sons, Ltd; 2005.

50. Quintana-Murci L. Understanding rare and common diseases in the context of human evolution. Genome Biol. 2016;17:225.

51. Matlak D, Szczurek E. Epistasis in genomic and survival data of cancer patients. PLoS Comput Biol. 2017;13:e1005626.

52. Hess JM, Bernards A, Kim J, Miller M, Taylor-Weiner A, Haradhvala NJ, et al. Passenger Hotspot Mutations in Cancer. Cancer Cell. 2019;36:288-301.e14.

53. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019;47:D886–94.

54. McFarland CD, Yaglom JA, Wojtkowiak JW, Scott JG, Morse DL, Sherman MY, et al. The damaging effect of passenger mutations on cancer progression. Cancer Res. 2017;77:4763–72.

55. Ostroverkhova D, Przytycka TM, Panchenko AR. Cancer driver mutations: predictions and reality. Trends Mol Med. 2023;29:554–66.

56. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.

57. Koh G, Degasperi A, Zou X, Momen S, Nik-Zainal S. Mutational signatures: emerging concepts, caveats and clinical applications. Nat Rev Cancer. 2021;21:619–37.

58. Yaacov A, Ben Cohen G, Landau J, Hope T, Simon I, Rosenberg S. Cancer mutational signatures identification in clinical assays using neural embedding-based representations. Cell Rep Med. 2024;5:101608.

59. Wang H, Guo M, Wei H, Chen Y. Targeting p53 pathways: mechanisms, structures, and advances in therapy. Signal Transduct Target Ther. 2023;8:92.

60. Yamamoto K, Lee Y, Masuda T, Ozono K, Iwatani Y, Watanabe M, et al. Functional landscape of genome-wide postzygotic somatic mutations between monozygotic twins. DNA Res. 2024;31.

61. Ji X, Wang E, Cui Q. Deciphering gene contributions and etiologies of somatic mutational signatures of cancer. Brief Bioinform. 2023;24.

62. Farmanbar A, Kneller R, Firouzi S. Mutational signatures reveal mutual exclusivity of homologous recombination and mismatch repair deficiencies in colorectal and stomach tumors. Sci Data. 2023;10:423.

63. Yaacov A, Rosenberg S, Simon I. Mutational signatures association with replication timing in normal cells reveals similarities and differences with matched cancer tissues. Sci Rep. 2023;13:7833.

64. Pužar Dominkuš P, Hudler P. Mutational signatures in gastric cancer and their clinical implications. Cancers (Basel). 2023;15.

65. Crisafulli G. Mutational signatures in colorectal cancer: Translational insights, clinical applications, and limitations. Cancers (Basel). 2024;16:2956.

66. Jayakrishnan R, Kwiatkowski DJ, Rose MG, Nassar AH. Topography of mutational signatures in non-small cell lung cancer: emerging concepts, clinical applications, and limitations. Oncologist. 2024;29:833–41.

67. Li B, Brady SW, Ma X, Shen S, Zhang Y, Li Y, et al. Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia. Blood. 2020;135:41–55.

68. van der Ham CG, Suurenbroek LC, Kleisman MM, Antić Ž, Lelieveld SH, Yeong M, et al. Mutational mechanisms in multiply relapsed pediatric acute lymphoblastic leukemia. Leukemia. 2024;38:2366–75.

69. Carter H, Marty R, Hofree M, Gross AM, Jensen J, Fisch KM, et al. Interaction landscape of inherited polymorphisms with somatic events in cancer. Cancer Discov. 2017;7:410–23.

70. Lin H, Zhang G, Zhang X-C, Lian X-L, Zhong W-Z, Su J, et al. Germline variation networks in the PI3K/AKT pathway corresponding to familial high-incidence lung cancer pedigrees. BMC Cancer. 2020;20:1209.

71. Panwar V, Singh A, Bhatt M, Tonk RK, Azizov S, Raza AS, et al. Multifaceted role of mTOR (mammalian target of rapamycin) signaling pathway in human health and disease. Signal Transduct Target Ther. 2023;8:375.

72. Glaviano A, Foo ASC, Lam HY, Yap KCH, Jacot W, Jones RH, et al. PI3K/AKT/mTOR signaling transduction pathway and targeted therapies in cancer. Mol Cancer. 2023;22:138.

73. Chen Y, Chang L, Hu L, Yan C, Dai L, Shelat VG, et al. Identification of a lactylation-related gene signature to characterize subtypes of hepatocellular carcinoma using bulk sequencing data. J Gastrointest Oncol. 2024;15:1636–46.

74. Letafati A, Taghiabadi Z, Zafarian N, Tajdini R, Mondeali M, Aboofazeli A, et al. Emerging paradigms: unmasking the role of oxidative stress in HPV-induced carcinogenesis. Infect Agent Cancer. 2024;19:30.

75. Wang H, Liu C, Jin K, Li X, Zheng J, Wang D. Research advances in signaling pathways related to the malignant progression of HSIL to invasive cervical cancer: A review. Biomed Pharmacother. 2024;180:117483.

76. Ciesielski TH, Sirugo G, Iyengar SK, Williams SM. Characterizing the pathogenicity of genetic variants: the consequences of context. NPJ Genom Med. 2024;9:3.

FIGURES

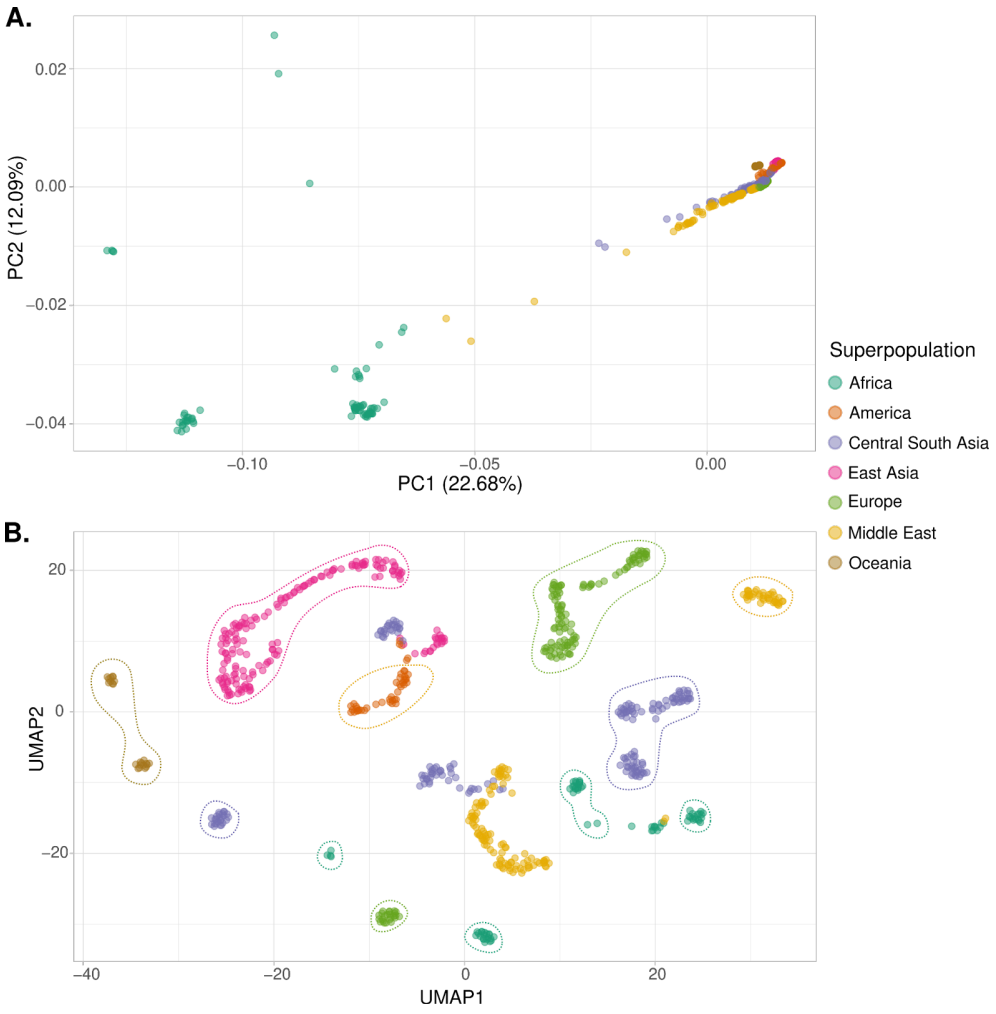


Figure 1. Visualization of HGDP data via PCA-UMAP.

Two-dimensional representation of the seven global superpopulations—Africa, America, Central and South Asia, East Asia, Europe, the Middle East, and Oceania—obtained through PCA-UMAP. (A) The first two principal components and (B) a two-dimensional UMAP embedding of the seven global populations. The circles around the points indicate individuals considered nonadmixed and were selected to represent the characteristic genetic variability of their respective populations. The separation of clusters reflects global genetic diversity and population structure, while local distances and groupings are optimized to preserve genetic proximity relationships.

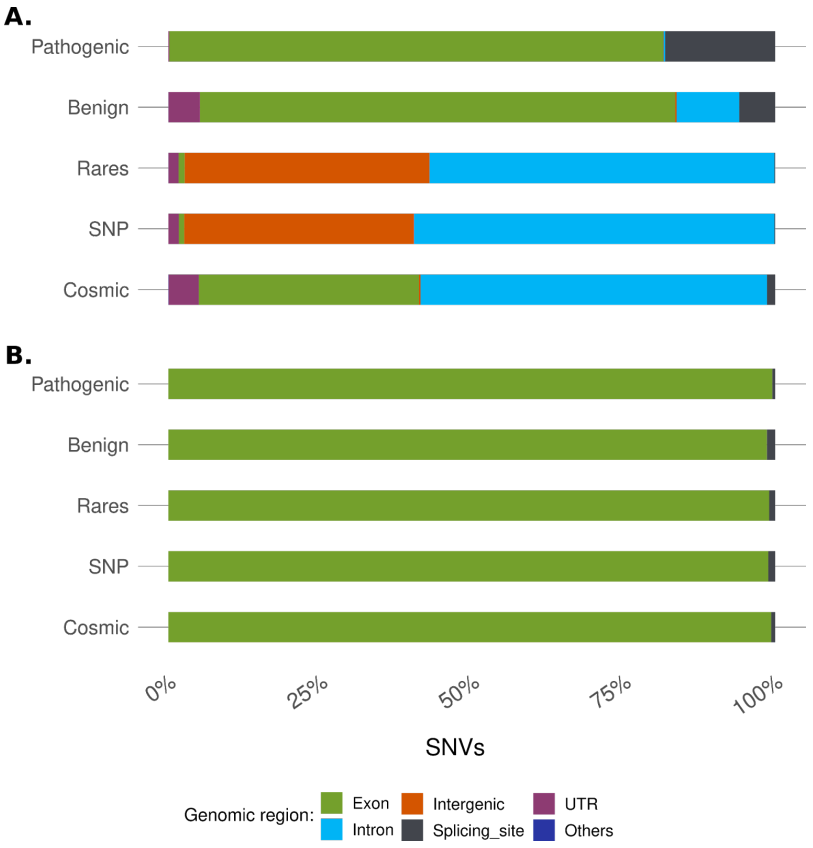


Figure 2. Genomic localization of SNVs across the five datasets: Pathogenic, Benign, Rare, SNPs, and COSMIC. The bars represent each dataset and the proportion of SNVs distributed across genomic regions for (A) WG and (B) CDS after filtering, highlighting the distribution pattern of SNVs and demonstrating successful filtering for CDS. Pearson's chi-square test for independence revealed a significant association between genomic regions and the WG datasets (χ^2 = 21,977.078; p = 2.2e-16).

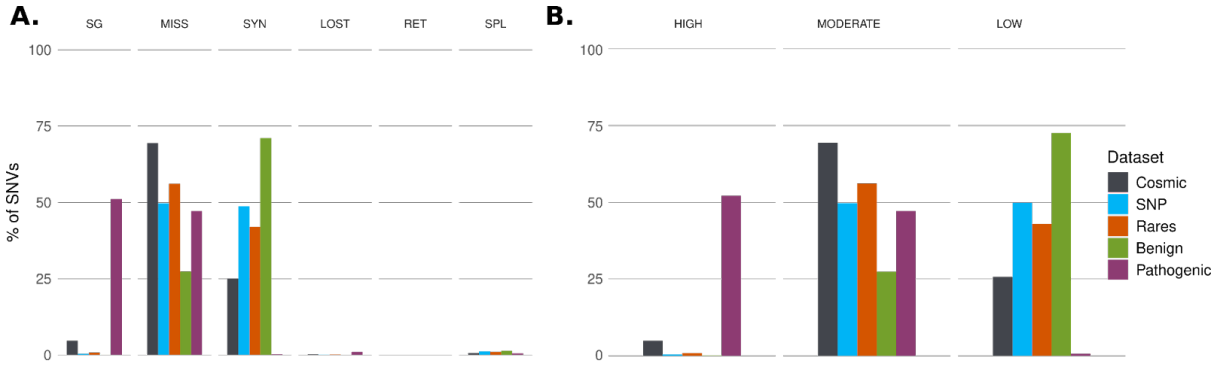


Figure 3: Severe coding consequences of SNVs and functional impact magnitude across each analyzed dataset. (A) Proportion of SNVs for the most severe coding

consequences: SG (stop-gained): This mutation introduces a premature stop codon, resulting in a truncated, often nonfunctional protein. MISS (missense variant): A single nucleotide change results in a different amino acid, potentially affecting protein function. Synonymous variant (SYN): A mutation that does not alter the amino acid sequence and is often considered neutral. LOST (Start lost and Stop lost): Mutations that result in the loss of start or stop codons, impacting protein length and potentially its function. RET (stop-retained variant): This mutation retains the stop codon, leading to an elongated protein with altered function. SPL (splice region variant): A mutation affecting splice sites, resulting in abnormal splicing. The figure illustrates the distribution of SNVs among these coding consequences. (B) Functional impact magnitude by coding consequences of genomic regions for SNVs in the CDS across the five datasets. HIGH - The variant is predicted to have a significant effect on the protein, potentially causing truncation, loss of function, or nonsense-mediated decay activation. MODERATE - A variant not expected to severely disrupt protein function but may alter its efficacy. LOW - Likely benign or with minimal impact on protein function.

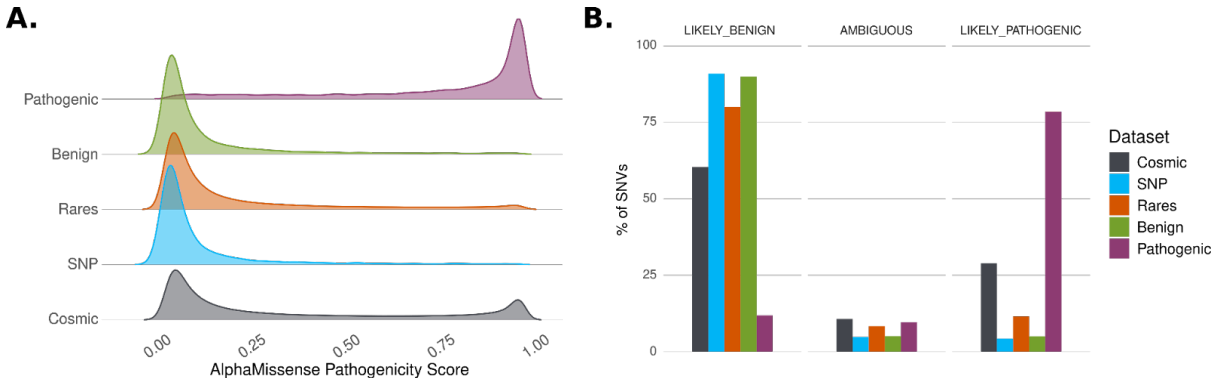


Figure 4. Pathogenicity assessment of missense SNVs in CDS regions via AlphaMissense across the five datasets: Benign, COSMIC, Pathogenic, Rare, and SNPs datasets. (A) Distribution of pathogenicity scores for missense variants. The y-axis of the ridgeline plot represents the probability density of the pathogenicity scores for each dataset. The height of the density curves at any given point indicates the relative likelihood of observing a pathogenicity score within the corresponding range of values. (B) Qualitative

classification of missense SNV pathogenicity. On the basis of the continuous pathogenicity score provided by AlphaMissense, which ranges from 0 to 1, the pathogenicity of each missense variant is categorized into one of three classes: likely_pathogenic, likely_benign, or ambiguous. Variants are classified as likely_benign if am_pathogenicity < 0.34, likely_pathogenic if am_pathogenicity > 0.564, and those with intermediate pathogenicity scores (0.34 ≤ am_pathogenicity ≤ 0.564) are considered uncertain by the model (ambiguous). These thresholds were defined to achieve 90% accuracy for both pathogenic and benign variants.

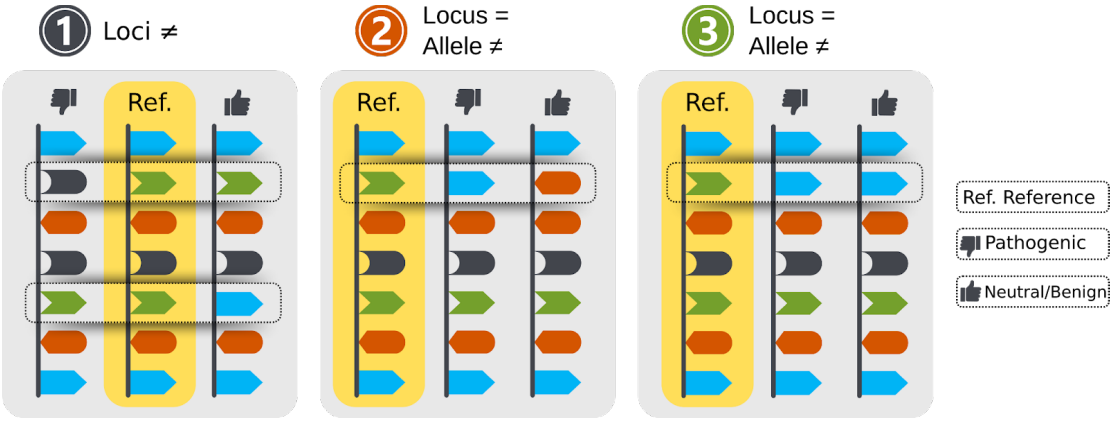
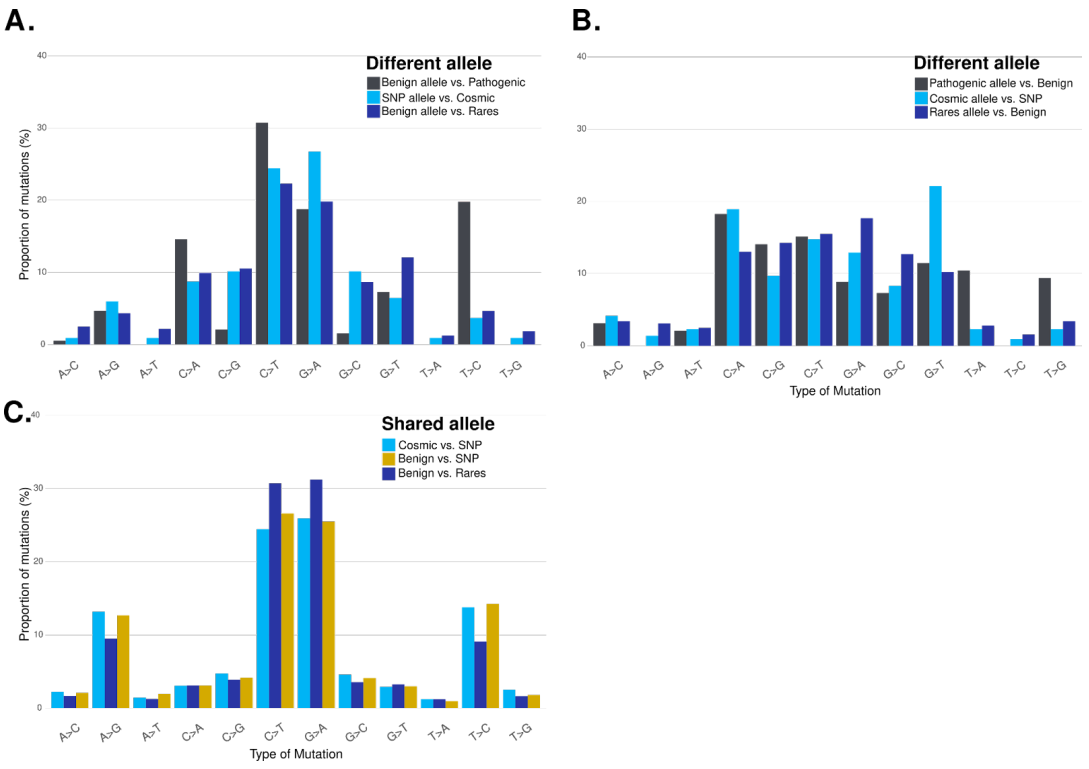


Figure 5. Models of spatial mutation occurrence in the human genome. Three models are proposed to describe the distribution of mutations when comparing pathogenic and neutral variants relative to the reference allele. Model 1: Pathogenic and neutral variants occur at distinct loci. Model 2: Pathogenic and neutral variants occur at the same locus but have different alleles. Model 3: Pathogenic and neutral variants occur at the same locus and share the same allele.



1138

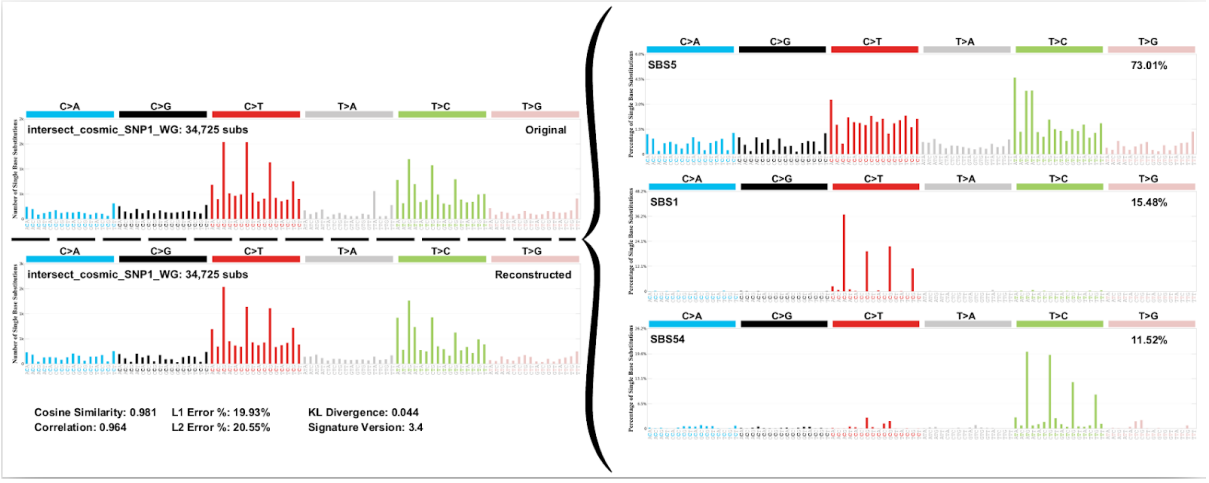
1139 **Figure 6. Single-nucleotide substitution patterns in subsets with positive associations**

1140 **identified by Fisher's exact test.** (A and B) Substitutions at positions with different alleles

1141 in the analyzed dataset pairs (e.g., Benign vs. Pathogenic, SNPs vs. COSMIC). (C) Patterns

1142 of alleles shared between subsets (e.g., SNP vs. COSMIC, Benign vs. SNPs).

1143



1144

1145 **Figure 7. Mutational signature reconstructed in the subset of overlapping SNVs**

1146 **between Cosmic and SNP for WG.** An analysis conducted with the SigProfiler Assignment

1147 identified SBS5, SBS1 and SBS54 as predominant in the WG dataset. SBS5 accounts for

73.01% of mutations and is associated with processes of unknown origin, whereas SBS1, representing 15.48%, is linked to the spontaneous deamination of 5-methylcytosine. SBS54 (11.52%) is linked to microsatellite instability (MSI) and DNA repair deficiency (dMMR). A cosine similarity of 0.981 and correlation of 0.964 indicate strong alignment between our subset's signatures and the COSMIC reference signatures. The L1 and L2 errors (19.93% and 20.55%, respectively) fall within acceptable limits, supporting a reasonable fit, whereas a KL divergence of 0.044 confirms a close distributional resemblance.

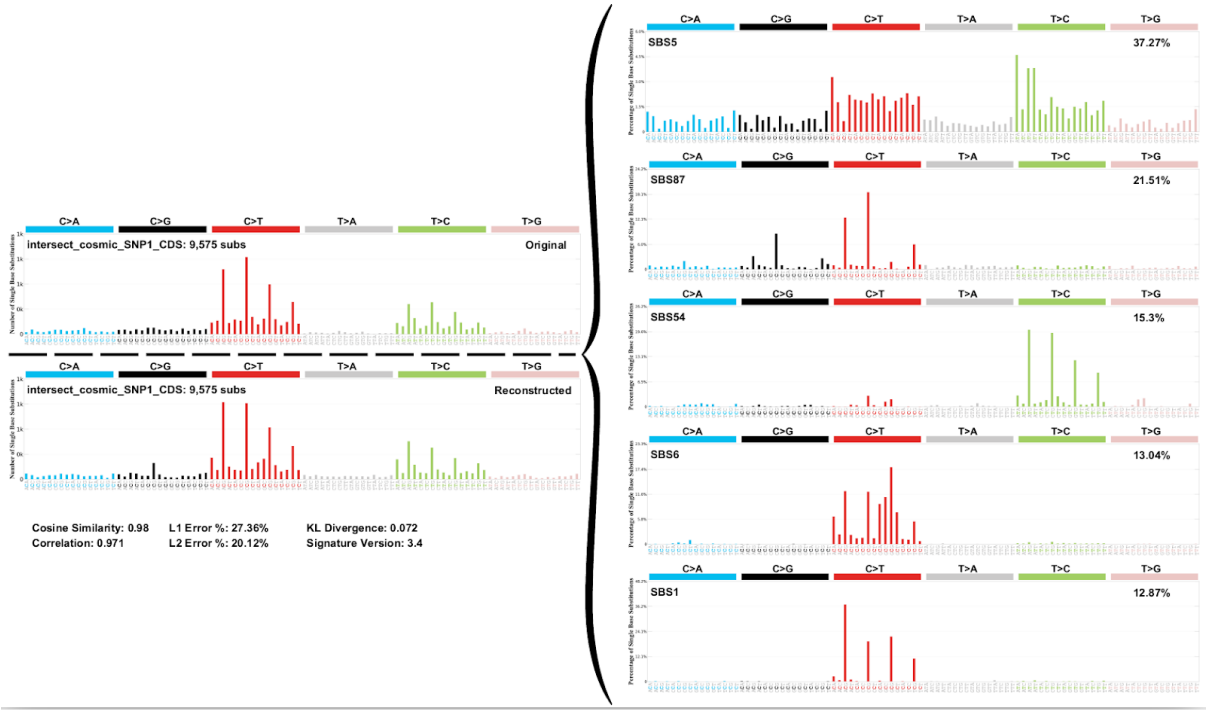


Figure 8. Mutational signature profiles identified in the CDS subset of overlapping SNVs between COSMIC and SNPs. SBS5 accounted for 32.27% of the mutations, whereas SBS1 accounted for 12.87%. Other signatures include SBS87 (21.51%), SBS54 (15.3%), and SBS6 (13.04%). A cosine similarity of 0.98 and a correlation of 0.971 indicate strong resemblance between the subset signatures and the COSMIC reference signatures. The L1 and L2 errors (27.36% and 20.12%, respectively) further support the model's robustness, whereas a KL divergence of 0.072 confirms a close match between the distributions.

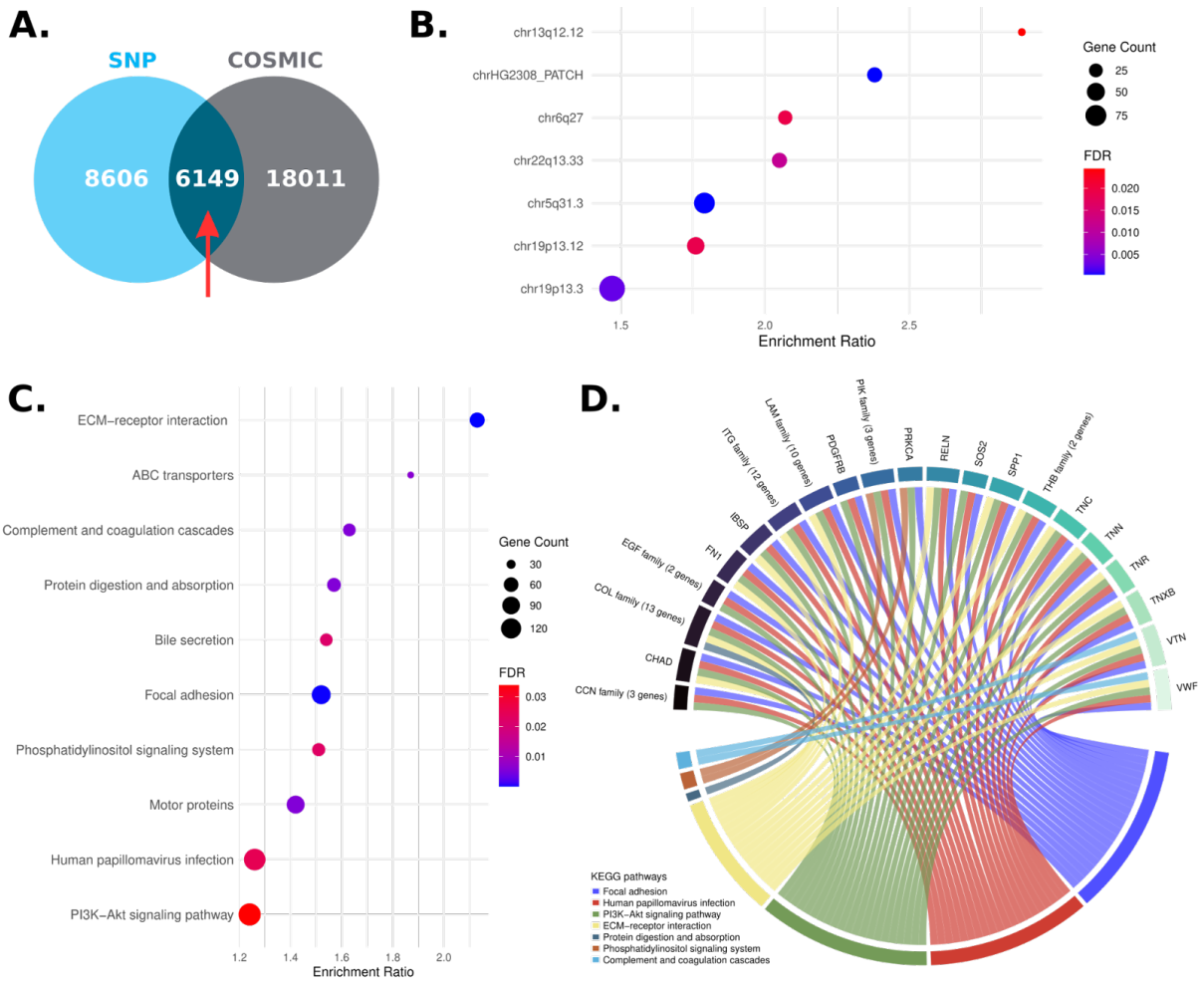
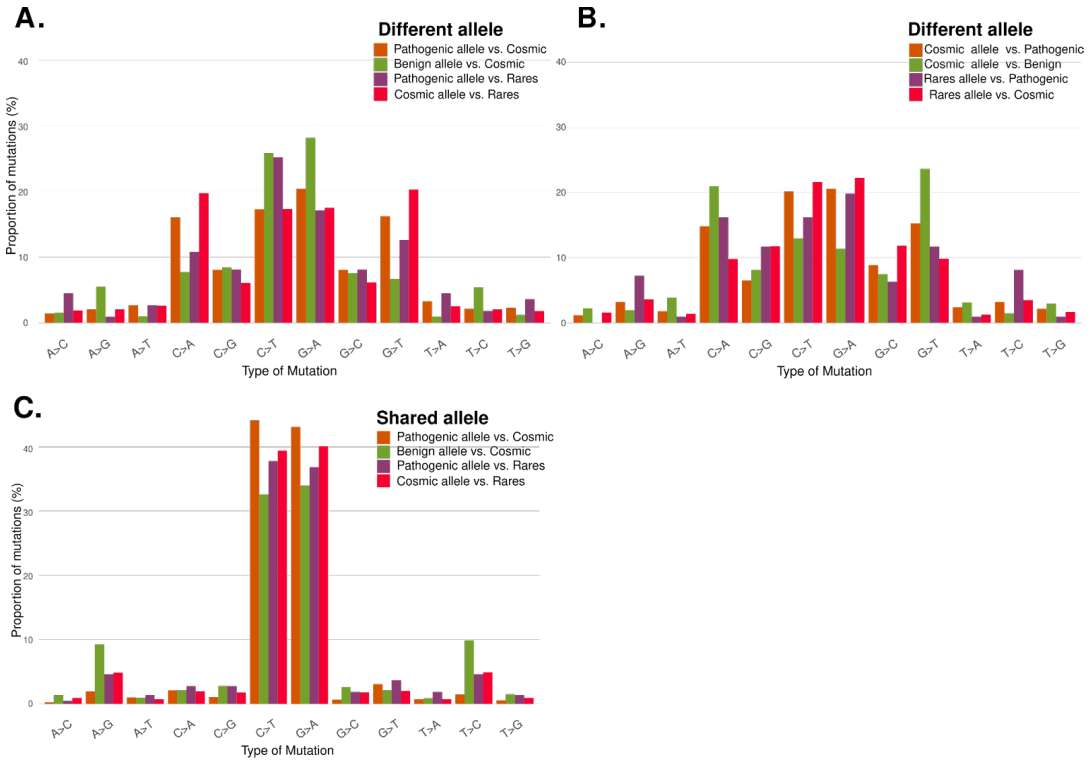


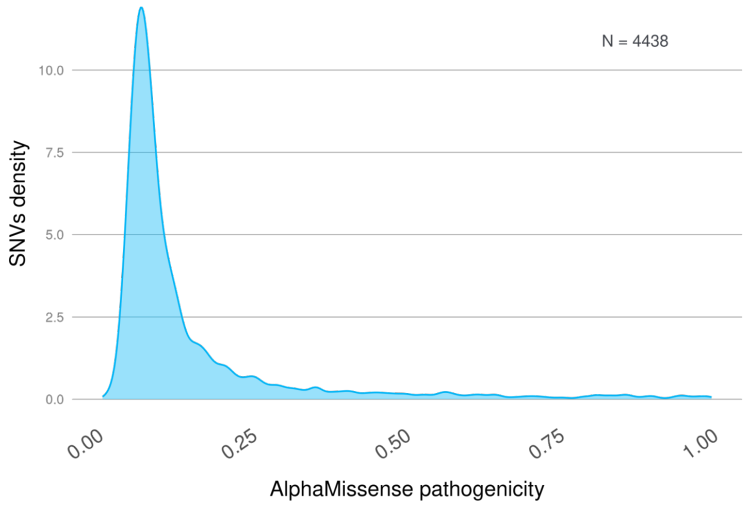
Figure 9. Functional enrichment analysis (Over-Representation Analysis – ORA) of SNPs at loci shared with cancer somatic mutations. (A) Intersection diagram of genes containing germline and somatic variants. The COSMIC dataset contains 18,011 genes with somatic mutations, whereas the SNPs dataset comprises 8,606 genes with high allele frequency (AF ≥ 1%) germline variants, resulting in an intersection of 6,149 genes. (B) Cytogenetic band enrichment analysis. The Y-axis represents significantly enriched genomic regions (FDR < 0.05), whereas the X-axis shows the enrichment ratio. The dot size reflects the number of genes within each region, and the color scale represents statistical significance, ranging from blue (most significant) to red (less significant). (C) KEGG pathway enrichment analysis highlighting biologically relevant and significantly enriched pathways (FDR < 0.05). (D) Gene-pathway interconnectivity diagram. The circular plot illustrates

genes/families shared across three or more KEGG pathways. The different colors represent different pathways, and the connections indicate genes enriched in multiple pathways.

SUPPLEMENTARY FIGURES



Supplementary Figure 1. Single-nucleotide substitution patterns in subsets with negative associations identified by Fisher's exact test. (A and B) Substitutions at positions with different alleles in the analyzed dataset pairs (e.g., Pathogenic vs. COSMIC). (C) Shared alleles between subsets (e.g., COSMIC vs. Rare).



Supplementary Figure 2. Distribution of AlphaMissense pathogenicity scores for missense SNVs in the CDS region of the SNP subset vs. the COSMIC subset. The plot shows the density of pathogenicity scores assigned by AlphaMissense for 4,438 missense variants shared between the SNPs and COSMIC datasets. Most variants presented low pathogenicity scores (<0.25), indicating minimal or no functional impact, which is consistent with the profile of neutral or passenger mutations. The presence of a long right tail suggests that a small proportion of variants with higher scores are potentially associated with more significant impacts.

TABLES

Table 1. Summary of Fisher's exact test results and color-coded representation for positive or negative associations. Fisher's exact test provided p values ($p < 0.001$; see full values in Supplementary Table 6) and odds ratios (OR) for intersections between two sets of genomic coordinates (Datasets 1 and 2). To highlight differences in association trends for each test, values with OR > 1 indicating positive associations are highlighted in green, whereas values with OR < 1 indicating negative associations are highlighted in red.

Dataset 1	Dataset 2	OR (CDS)	OR (WG)
SNPs	Cosmic	5.652	1.920
SNPs	Benign	66.480	22.493

SNPs	Pathogenic	0.335	0.082
Rares	Cosmic	1.315	0.558
Rares	Benign	16.240	5.604
Rares	Pathogenic	0.885	0.300
Cosmic	Benign	0.699	17.910
Cosmic	Pathogenic	0.557	13.900
Benign	Pathogenic	2.150	112.648

SUPPLEMENTARY TABLES

Supplementary Table 1. Summary of SNV genomic coordinates postprocessing for WG and CDS.

Number of SNVs in the input VCF files and genomic coordinate entries in the output BED files after processing through the pipeline developed in this study for both the WG and the CDS. Data were obtained from three databases (COSMIC, ClinVar, and HGDP), resulting in five datasets: COSMIC, Benign, Pathogenic, SNPs, and Rares.

	HGDP		Cosmic		ClinVar	
Input	75,385,302		**Coding** 37,261,323	**No coding** 31,368,583	1,303,558	
Output *WG*	**SNPs** 922,296	**Rares** 50,239,810	10,587,546		**Benign** 76,971	**Pathogenic** 16,313
Output *CDS*	16,040	496,791	3,662,887		56,794	13,098

Supplementary Table 2. Contingency tables of SNVs in each genomic region for the whole genome (WG) and for protein-coding regions (CDS) across each dataset - Benign, Pathogenic, Cosmic, Rares and SNPs. The table presents the absolute counts of SNVs in different genomic regions, enabling a comparison between the comprehensive WG analysis and the targeted examination of CDS in all studied datasets.

EXCEL FILE

Supplementary Table 3: Absolute counts and percentages of transitions (Ti) and transversions (Tv) for each CDS analyzed dataset, as well as the Ti_Tv ratio. Transitions are interchanges of two-ring purines (A<> G) or one-ring pyrimidines (C<>T), whereas transversions are interchanges of purines for pyrimidine bases (A<>C, A<>T, G<>C, G<>T).

	Mutation type		
Dataset	**Transitions**	**Transversion**	**Ti_Tv_Ratio**
Pathogenic	9000 (68.72%)	4097 (31.28%)	2.19
Benign	46238 (81.41%)	10555 (18.59%)	4.38
Rares	368677 (74.21%)	128114 (25.79%)	2.87
SNPs	12360 (77.06%)	3680 (22.94%)	3.35
Cosmic	2312710 (63.14%)	1350176 (36.86%)	1.71

Supplementary Table 4: Absolute count of coding consequences for SNVs in CDS regions, categorized by dataset. The table provides a detailed breakdown of the types of coding consequences observed in each dataset.

Consequence type	Pathogenic	Benign	SNPs	Rares	Cosmic
missense_variant	6172	15553	7976	295730	2543087
splice_region_variant	66	784	181	4707	24033
start and stop_lost	136	11	13	767	6887
stop_gained	6695	33	54	5341	172708
stop_retained_variant	0	16	7	155	1203
synonymous_variant	29	40397	7809	190091	914969

Supplementary Table 5. Categorization of CDS SNV datasets by the AlphaMissense pathogenicity score. The number of missense SNVs classified into one of three categories: likely pathogenic, likely benign, or ambiguous.

Dataset	Category	Count	Percent (%)
Rares	ambiguous	21916	8.352
	likely_benign	210120	80.075
	likely_pathogenic	30365	11.571
SNP	ambiguous	341	4.801
	likely_benign	6456	90.903

	likely_pathogenic	305	4.294
Cosmic	ambiguous	244551	10.693
	likely_benign	1381746	60.420
	likely_pathogenic	660568	28.885
Benign	ambiguous	710	5.089
	likely_benign	12548	89.949
	likely_pathogenic	692	4.960
Pathogenic	ambiguous	565	9.605
	likely_benign	697	11.849
	likely_pathogenic	4620	78.544

Supplementary Table 6. Complete results of Fisher's exact test. Contingency tables for the number of overlaps and unique intervals between two files in autosomal DNA for the CDS and WG regions. Fisher's exact test provides p values and odds ratios (OR) for intersections between two sets of genetic or genomic coordinates (A and B), along with four additional values: the number of intersections (in_A/in_B), the counts of unique intervals for each file (in_A/not_B or not_A/in_B), and the estimated number of nonoverlapping coordinates (not_A/not_B), which are calculated on the basis of the total possible intervals derived from each gene's size. OR indicates the strength of the association. The results are presented for five datasets: SNPs, Rare, COSMIC, Benign, and Pathogenic, across CDS and WG regions. The test between SNPs and Rare was not performed, as these datasets are mutually exclusive. Tests conducted on CDS regions are more robust, particularly for datasets derived from ClinVar, due to the distinct SNV distribution patterns observed across the WG.

CDS							
Dataset A	**Dataset B**	***p*-value**	**OR**	**in_A/in_B**	**in_A/not_B**	**not_A/in_B**	**not_A/not_B**
SNPs	Cosmic	0	5.652	9575	6465	3653035	13940452
SNPs	Benign	0	66.480	2728	13312	54066	17539421
SNPs	Pathogenic	0.013	0.335	4	16036	13094	17580393
Rares	Cosmic	0	1.315	126723	370068	3536158	13576578
Rares	Benign	0	16.240	17777	479014	39017	17073719
Rares	Pathogenic	0.026	0.885	328	496463	12770	17099966

Cosmic	Benign	0	0.699	20372	3642515	36422	4550829
Cosmic	Pathogenic	1.1e-225	0.557	4036	3658851	9062	4578189
Benign	Pathogenic	5.6e-24	2.150	192	56602	12906	8180438
WHOLE GENOME							
SNPs	Cosmic	0	2.622	34725	887571	10552821	707275263
SNPs	Benign	0	24.755	2367	919929	74604	717753480
SNPs	Pathogenic	2e-08	0.048	1	922295	16312	717811772
Rares	Cosmic	0	0.558	428910	49810900	10158636	658351934
Rares	Benign	0	5.604	22804	50217006	54167	668456403
Rares	Pathogenic	3.5e-169	0.300	360	50239450	15953	668494617
Cosmic	Benign	0	17.910	25666	19528706	51306	699144702
Cosmic	Pathogenic	0	13.900	4566	19549806	11748	699184260
Benign	Pathogenic	0	112.64	194	76778	16120	718657288
Subset SNP intersect Cosmic	CpG Island	0	7.029	1623	33102	25240	3618606
Subset SNP intersect Cosmic	Microsatellites	4e-37	2.261	45	34680	37839	65939551

Supplementary Table 7.Description of each subset obtained after performing Fisher's exact tests. This table includes the total number of SNVs and the percentages of equal (shared) and different alleles between the CDS datasets. The subsets are derived from significant intersections of genomic coordinates, showing the frequency of shared and unique alleles among different datasets.

Tested datasets	Alleles	Total of SNVs	
		Equal	Different
SNPs vs. Cosmic	SNPs	0	217 (2.27%)
	Cosmic	0	217 (2.27%)
	Shared	9358 (97.73%)	0
SNPs vs. Benign	SNPs	0	0
	Benign	0	0
	Shared	2728 (100%)	0
SNPs vs. Pathogenic	SNPs	0	4 (100%)
	Pathogenic	0	4 (100%)
Pathogenic vs. Benign	Pathogenic	0	192 (100%)
	Benign	0	192 (100%)

Pathogenic vs. Cosmic	Pathogenic	0	1127 (27.92%)
	Cosmic	0	1127 (27.92%)
	Shared	2909 (72.08%)	0
Benign vs. Cosmic	Benign	0	2740 (13.45%)
	Cosmic	0	2740 (13.45%)
	Shared	17632 (86.55%)	0
Pathogenic vs. Rares	Pathogenic	0	111 (33.84%)
	Rares	0	111 (33.84%)
	Shared	217 (66.16%)	0
Benign vs. Rares	Benign	0	323 (1.82%)
	Rares	0	323 (1.82%)
	Shared	17454 (98.18%)	0
Rares vs. Cosmic	Rares	0	29104 (22.97%)
	Cosmic	0	29104 (22.97%)
	Shared	97619 (77.03%)	0

Supplementary Table 8. Count of single nucleotide substitutions for all subsets obtained after Fisher's exact test. This table provides a detailed breakdown of the number and percentage of each type of single nucleotide substitution (e.g., A>G, C>T) within the subsets that showed significant intersections of genomic coordinates.

EXCEL FILE

Supplementary Table 9. Count and proportion of WG SNVs by coding consequence in the SNP vs. COSMIC subset.

Consequence_type	Count	Percent (%)
3_prime_UTR_variant	11936	1.294
5_prime_UTR_variant	2811	0.304
downstream_gene_variant	35334	3.831
intergenic_variant	247380	26.822
intron_variant	514536	55.788
mature_miRNA_variant	13	0.001
missense_variant	3558	0.385
non_coding_transcript_exon_variant	22869	2.479
regulatory_region_variant	33197	3.599
splice_acceptor_variant	60	0.006
splice_donor_5th_base_variant	74	0.008
splice_donor_region_variant	187	0.020
splice_donor_variant	103	0.011

splice_polypyrimidine_tract_variant	730	0.079
splice_region_variant	894	0.096
start_lost	10	0.001
stop_gained	37	0.004
stop_lost	14	0.001
stop_retained_variant	5	0.0005
synonymous_variant	3302	0.358
TF_binding_site_variant	2409	0.261
upstream_gene_variant	42837	4.644

Supplementary Table 10. Count and proportion of CDS SNVs by coding consequence in the SNP vs. COSMIC subset.

Consequence_type	Count	Percent
missense_variant	4926	51.446
splice_region_variant	89	0.929
start_lost	4	0.041
stop_gained	40	0.417
stop_lost	2	0.020
stop_retained_variant	4	0.041
synonymous_variant	4510	47.101

Supplementary Table 11. Mutational matrix SBS in the subset WG of overlapping SNPs vs. COSMIC.

EXCEL FILE

Supplementary Table 12. Mutational matrix SBS in the subset CDS of overlapping SNPs vs. COSMIC.

EXCEL FILE

Supplementary Table 13. Complete list of genes present in the significantly enriched cytogenetic bands identified in the Over-Representation Analysis (ORA). The table includes all cytogenetic bands with FDR < 0.05, along with their respective mapped genes.

EXCEL FILE

Supplementary Table 14. Complete list of genes enriched in the KEGG pathways identified in the Over-Representation Analysis (ORA). The table presents all significantly enriched pathways (FDR < 0.05), along with the mapped genes associated with each pathway.

EXCEL FILE