Metodologia

Desenho do estudo e objetivo

Realizamos uma revisão analítica, com curadoria estruturada de dados, para sintetizar descobertas e inovações tecnológicas em periodontite obtidas por abordagens de bioinformática (genômica, proteômica, transcriptômica, *bulk* e *single-cell,* metabolômica, metagenômica 16S e *shotgun*, metatranscritômica, arrays funcionais/filogenia e metodologias afins). O objetivo foi mapear biomarcadores, mecanismos de doença, ferramentas diagnósticas/prognósticas e aplicações translacionais, além de comparar o desempenho reportado (AUC, acurácia) entre plataformas e matrizes biológicas.

Estratégia de busca e triagem (Semantic Scholar/OpenAlex)

A busca semântica foi conduzida no Semantic Scholar e OpenAlex com a consulta: *"What are the main discoveries and technological innovations described in studies that applied bioinformatics approaches, including genomics, proteomics, transcriptomics, metabolomics, resistomics, phylogeny, and phenetics, to periodontitis? Please summarize findings on biomarkers, disease mechanisms, diagnostic/prognostic tools, and translational applications."*
A busca retornou os 494 artigos mais relevantes. A inclusão foi decidida por julgamento holístico, com base nos seguintes critérios: população com diagnóstico de periodontite (qualquer forma, segundo critérios clínicos), uso de pelo menos uma abordagem de bioinformática/ômica, estudos com amostras humanas, desenho elegível (originais, séries/cortes, transversais, caso-controle, ECR, revisões sistemáticas e meta-análises), relato de achados/avanços metodológicos, foco explícito em periodontite, integração bioinformática (além de métodos clínicos/ laboratoriais tradicionais) e publicação completa (excluídos resumos, cartas ou comunicações breves).

Extração de dados (LLM assistida) e variáveis

A extração foi realizada de modo híbrido (LLM assistida + checagens pontuais humanas), seguindo instruções padronizadas para cada coluna, abrangendo:

- Abordagens de bioinformática/ômicas (técnicas, ferramentas, bancos, pipelines).
- População/amostras (tamanho, estado de doença, tipo e local de coleta, desenho, status de tratamento).
- Biomarcadores e métricas (alvos, direção do efeito, significância, sensibilidade, especificidade, acurácia, AUC).
- Mecanismos de doença (vias, processos biológicos, interação hospedeiro–microbiota, enriquecimento funcional, PPI).
- Ferramentas preditivas (tipo de modelo, alvo preditivo, *features*, desempenho e validação).
- Inovações técnicas (algoritmos, *workflows*, integração multi-ômica, novos protocolos).
- Aplicações clínicas (diagnóstico, prognóstico, monitoramento, seleção terapêutica) e prontidão clínica.

Os resultados extraídos foram consolidados em planilha única (*.xlsx*), posteriormente processada no R.

Ambiente analítico e reprodutibilidade

As análises foram conduzidas em R com os pacotes: tidyverse, readxl, janitor, stringr, forcats, tidyr, ggpubr, ggalluvial, openxlsx (e uso opcional de ggbeeswarm). O *script* principal foi desenhado para reprodutibilidade e portabilidade em repositórios GitHub:

- Entrada padrão: data/periodontitis_bioinfo_omics_elicit_filled.xlsx
- Saída: outputs/ contendo figures/, tables/ e *workbook* .xlsx com múltiplas abas (dados brutos, estudos curados, contagens, desempenho e *top* estudos).
- Parâmetros via CLI: --in <arquivo.xlsx> e --out <dir>.

Limpeza e normalização dos dados

- Padronização textual: remoção de espaços redundantes e *trim* de campos de texto; *snake_case* de rótulos com janitor::clean_names.
- Tratamento de ausências: uso de *helpers* "seguros" para agregações (safe_median, safe_min, safe_max) que retornam NA quando aplicável.
- Seleção de desempenho: criação de auc_best por coalesce(auc_max, auc_min) para capturar o melhor valor reportado; accuracy_best derivada de accuracy_pct.

Derivação de variáveis e regras de classificação

- Classe ômica (omics_class): mapeamento baseado em *regex* sobre omics_bioinfo_approaches, categorizando em: *Metagenomics (shotgun)*, *Metagenomics (16S)*, *Metatranscriptomics*, *Transcriptomics (single-cell)*, *Transcriptomics (bulk)*, *Proteomics*, *Metabolomics*, *Genomics/Functional arrays* e *Other/Methodological*.
- Uso de ML/DL (uses_ml_dl): *flag* binária por *regex* de termos como *machine learning*, *random forest*, SVM, *logistic regression*, *deep learning*, CNN, XGBoost/LightGBM e *neural*.
- Prontidão clínica (readiness): fator ordenado (Conceptual, Potential, Feasible, Promising, High accuracy, Pilot, Unspecified) a partir de *regex* aplicada ao texto livre de clinical_readiness.
- Aplicação clínica (app_group) para o diagrama alluvial: *Diagnosis/Screening*, *Monitoring/Treatment*, *Risk/Prognosis*, *Drug/Targeting* e *Other/Unspecified*, por *regex* no campo application.

Normalização de colunas multivalor e agrupamento de amostras

Campos com múltiplos itens separados por vírgula/ponto-e-vírgula foram "explodidos" (str_split + unnest). Tipos de amostras foram harmonizados em grupos: *Subgingival biofilm/plaque*, *Supragingival biofilm*, *Saliva*, *Gingival crevicular fluid (GCF)*, *Gingival tissue*, *Serum/Plasma*, *Oral microbiota (unspecified)* e *Other/NA*.

Sumários e métricas

1. Frequência por classe ômica e por tipo de amostra: contagens simples (*Tabelas* count_by_omics_class.csv, count_by_sample_type.csv), com percentuais calculados ad hoc no texto quando apropriado (p.ex., Figura 2).

2. Desempenho por classe ômica (performance_by_omics_class.csv): para cada omics_class, somamos o número de estudos com AUC e/ou acurácia reportadas, e computamos mediana, mínimo e máximo (ignorando NA).
3. Listas de destaque: *Top 10* estudos por AUC e por acurácia (com metadados de validação e prontidão).

Visualizações

- Figura 1 (*barras horizontais*): contagem de estudos por classe ômica (geom_col, coord_flip), *theme_minimal* customizado.
- Figura 2 (*barras horizontais*): contagem por tipo de amostra.
- Figura 3 (*heatmap* Ômica × Amostra): *tile* de AUC mediana por célula; quando auc_med é NA, a célula é exibida em cinza e anotada com n; quando disponível, anota-se "med=…" e n.
- Figura 4 (*dispersão/"beeswarm"* ou *jitter* de AUC por classe): pontos individuais por estudo; mediana destacada (ponto shape=23). *Fallback* automático para *jitter* quando ggbeeswarm não está instalado.
- Figura 5 (*alluvial* Ômica → Aplicação → Prontidão): fluxos por contagem (n) com ggalluvial, estratos rotulados e preenchimento por omics_class.

Todos os *plots* foram exportados em PNG (300 dpi) para outputs/figures/.

Análise quantitativa e regras de cálculo

- AUC: utilizamos o melhor valor reportado por estudo (auc_best = coalesce(auc_max, auc_min)), sem imputação.
- Acurácia: quando reportada como porcentagem, mantida em escala percentual.
- Agregação por classe: medianas de AUC são calculadas apenas sobre estudos com AUC disponível (estudos sem métrica não contribuem para o denominador da mediana).
- Percentuais descritivos (texto): com base nos totais observados em cada figura/tabela (p.ex., Figura 2: 33 ocorrências de matrizes; saliva e biofilme subgengival = 9 cada → 27,3% cada).
- Validação: o campo validation_status foi preservado do extrato (p.ex., validação externa, *cross-validation*, meta-análise, estudo piloto) e usado qualitativamente na interpretação.

Gestão de saídas e caderno de resultados

- Tabelas .csv com contagens e desempenho, além de um workbook Excel (tables/periodontitis_bioinfo_outputs.xlsx) contendo: *raw*, *studies_curated* (subconjunto padronizado), *count_by_omics*, *count_by_sample*, *perf_by_omics*, *top10_auc*, *top10_accuracy*.
- Mensagens de *log* registram diretórios de saída e arquivos gerados para rastreabilidade.

Considerações de qualidade, vieses e limitações

- Extração LLM-assistida pode incorrer em erros de associação (p.ex., métricas, amostra, validação). Mitigamos com normalizações conservadoras, campos explícitos de NA, e checagens pontuais contra o texto original quando necessário.
- Heterogeneidade entre estudos (desenho, *endpoints*, *pipelines* analíticos, amostras) impede meta-análise formal; portanto, nossas comparações de AUC/acurácia são descritivas.
- Ausência de métricas em várias classes (p.ex., metabolômica, 16S, metatranscritômica, *bulk* transcriptomics) reflete foco exploratório e/ou falta de validação padronizada; esses estudos entram nas contagens, mas não nos sumários de desempenho.
- Classificação via *regex* (ômica, aplicação, prontidão) privilegia sensibilidade; ambiguidade residual foi categorizada como *Other/Unspecified* ou *Unspecified* para não inflar categorias específicas.
- Reprodutibilidade: o *script* é parametrizável (entradas/saídas), contém *fallbacks* e evita falhas em presença de NA — permitindo replicação/atualização com novos *datasets*.

Resultados detalhados:

A análise da literatura revelou uma predominância de estudos baseados em proteômica (n=16; 21,3%), refletindo a centralidade dessa plataforma para investigar a periodontite, dado o papel das proteínas na mediação da inflamação, na resposta imune do hospedeiro e na composição de painéis biomarcadores clinicamente acionáveis. Em seguida, observa-se um volume expressivo de transcriptômica de célula única (n=12; 16,0%), tecnologia em franca expansão que expõe a heterogeneidade celular do periodonto e a dinâmica transcricional de subpopulações imunes e estromais em microambientes inflamatórios. As abordagens do microbioma também ocupam lugar de destaque: metagenômica shotgun (n=11; 14,7%) e metagenômica 16S (n=8; 10,7%) sustentam a identificação de táxons-chave e a caracterização do desequilíbrio disbiótico; de modo complementar, a metatranscritômica (n=10; 13,3%) acrescenta uma camada funcional ao revelar a atividade gênica das comunidades microbianas in situ. A transcriptômica em larga escala (bulk), RNA-seq e microarranjos, mantém relevância histórica (n=8; 10,7%), ancorando comparações de expressão gênica global na doença periodontal. Já a metabolômica (n=6; 8,0%) traz sinais bioquímicos do hospedeiro e do microbioma, ampliando a leitura de vias metabólicas ligadas à inflamação crônica. Embora menos representadas, genômica/arrays funcionais (n=3; 4,0%) e abordagens metodológicas diversas (n=1; 1,3%) completam o panorama, indicando oportunidades de expansão para essas frentes. Em conjunto, os dados sugerem que o campo tem se apoiado em plataformas maduras (proteômica e transcriptômica), enquanto cresce a adoção de tecnologias emergentes e de maior

resolução, notadamente single-cell e metatranscritômica, capazes de aprofundar a caracterização integrada entre microbiota, metabolismo e resposta imune na periodontite.

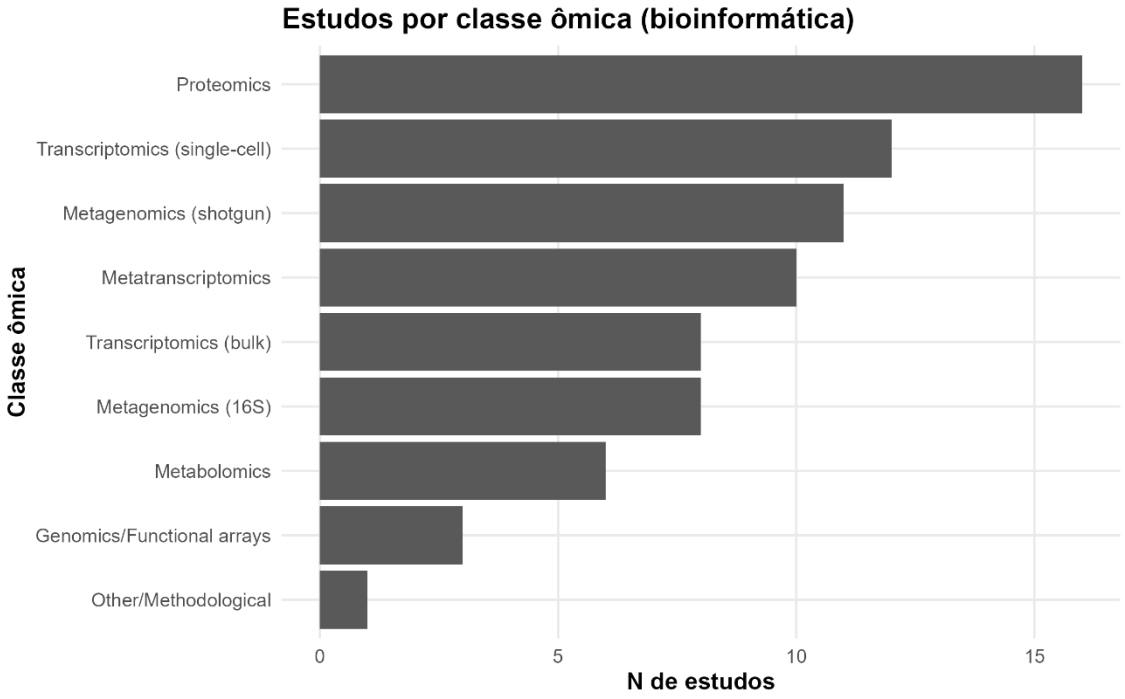

Figura 1. Distribuição dos estudos sobre periodontite segundo a classe ômica empregada em análises de bioinformática. Observa-se predomínio de abordagens proteômicas (n=16), seguidas por transcriptômica de célula única (n=12), metagenômica shotgun (n=11), metatranscritômica (n=10), metagenômica baseada em 16S e transcriptômica bulk (n=8 cada), metabolômica (n=6), genômica/arrays funcionais (n=3) e outras metodologias (n=1).

A literatura mostrou predominância de saliva e biofilme/plaque subgengival, com 9 estudos cada (27,3% + 27,3% = 54,5%), refletindo a combinação entre viabilidade operacional (saliva, coleta não invasiva e baixo custo) e proximidade do sítio patológico (biofilme subgengival, espelho direto do microambiente disbiótico). Em seguida aparecem fluido crevicular gengival – GCF e tecido gengival com 4 estudos cada (12,1% + 12,1% = 24,2%), matrizes que oferecem leitura fina do eixo hospedeiro–microbioma (mediadores inflamatórios no GCF; perfis celulares e moleculares no tecido). As categorias Other/NA (3 estudos; 9,1%), soro/plasma (2; 6,1%), microbiota oral não especificada (1; 3,0%) e biofilme supragengival (1; 3,0%) foram menos representadas, sinalizando oportunidade de diversificação para biomarcadores sistêmicos e fontes complementares. Em conjunto, 78,8% dos estudos concentram-se nas quatro matrizes principais (saliva, subgengival, GCF e tecido), evidenciando a preferência por amostras

de alta relevância biológica e/ou acessibilidade, sem excluir a expansão para matrizes alternativas conforme crescem os objetivos translacionais (Figura 2).

Figura 2. Tipos de amostras utilizados nos estudos incluídos. Distribuição das matrizes biológicas (n = 33 ocorrências): saliva e biofilme/plaque subgengival foram as mais frequentes (9 cada; 27,3% + 27,3% = 54,5%), refletindo, respectivamente, viabilidade operacional (coleta não invasiva e baixo custo) e proximidade do sítio patológico (espelho do microambiente disbiótico). Em seguida, fluido crevicular gengival – GCF e tecido gengival (4 cada; 12,1% + 12,1% = 24,2%) oferecem leitura fina do eixo hospedeiro–microbioma (mediadores inflamatórios no GCF; perfis celulares/moleculares no tecido). As categorias Other/NA (3; 9,1%), soro/plasma (2; 6,1%), microbiota oral não especificada (1; 3,0%) e biofilme supragengival (1; 3,0%) foram menos representadas, indicando oportunidade de diversificação para biomarcadores sistêmicos e fontes complementares. (Barras exibem contagens absolutas; percentuais indicados no texto.)

A análise integrativa evidenciou que, embora a distribuição dos estudos entre plataformas ômicas tenha sido heterogênea, alguns grupos apresentaram desempenhos notavelmente elevados. A proteômica foi a abordagem mais representada com cinco estudos, dos quais três reportaram métricas de acurácia diagnóstica (AUC), alcançando valores entre 0,88 e 0,97, com mediana de 0,97, refletindo robustez e aplicabilidade na identificação de biomarcadores moleculares. Esse desempenho posiciona a proteômica como a técnica com maior consistência entre as plataformas avaliadas.

A transcriptômica de célula única também apresentou resultados expressivos: em três estudos, dois reportaram AUC, ambos com valores elevados (mediana 0,92), destacando

o poder dessa tecnologia para capturar a heterogeneidade celular do periodonto e caracterizar respostas específicas em microambientes inflamatórios.

As análises metagenômicas por shotgun (quatro estudos) forneceram evidências pontuais de alto desempenho, com AUC relatada em um estudo (0,86) e acurácia associada de 94,4%, indicando seu potencial para mapear de forma abrangente a diversidade e função microbiana ligada à periodontite.
Por outro lado, abordagens tradicionais como genômica/arrays funcionais (dois estudos) exibiram resultados mais modestos, com AUC única de 0,73, sugerindo limitações na sensibilidade em comparação às tecnologias emergentes. Da mesma forma, metabolômica, metagenômica 16S, metatranscriptômica e bulk transcriptomics não apresentaram resultados de acurácia reportados, o que pode refletir tanto a natureza exploratória desses estudos quanto a ausência de validações padronizadas em contexto clínico.
Quando analisado o conjunto, observa-se que apenas 33,3% das classes ômicas (3/9) apresentaram métricas robustas de desempenho (proteômica, scRNA-seq e metagenômica shotgun), mas essas responderam por 62,5% dos valores de AUC disponíveis. Em contraste, 66,7% das classes (6/9) ainda não reportaram métricas de acurácia, o que evidencia a necessidade de maior padronização e validação.
Em síntese, os dados apontam para uma tendência em que plataformas de alta resolução e aplicabilidade clínica (proteômica e single-cell transcriptomics) lideram em desempenho, enquanto estratégias clássicas ou exploratórias permanecem sub-representadas. Essa assimetria reforça tanto os avanços recentes em biologia de sistemas aplicados à periodontite quanto os desafios para consolidar outras abordagens como ferramentas diagnósticas e prognósticas (Figura 3).

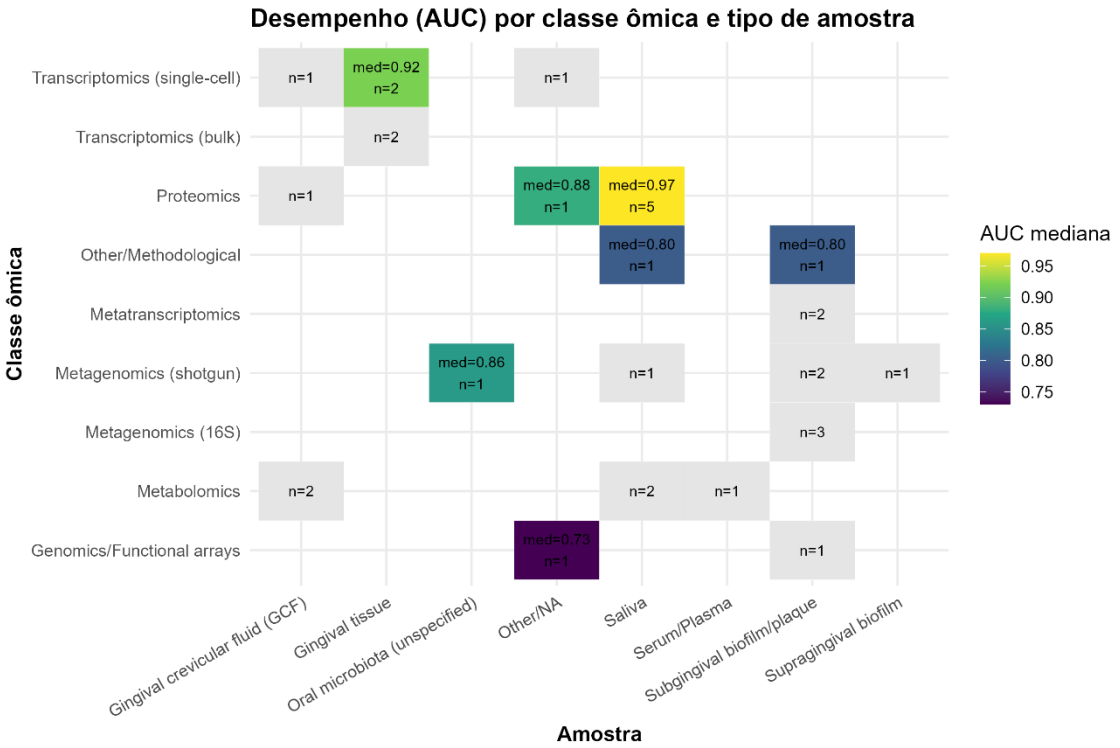

Figura 3. Relação entre abordagens ômicas, tipos de amostras e desempenho (AUC mediana). Apenas três classes (proteômica, transcriptômica de célula única e metagenômica shotgun) reportaram valores de AUC, representando 33,3% das plataformas analisadas mas respondendo por 62,5% dos valores de desempenho disponíveis. A proteômica (n=5) destacou-se com AUC entre 0,88–0,97 (mediana 0,97), seguida pela transcriptômica single-cell (n=3; mediana 0,92) e pela metagenômica shotgun (n=4; AUC=0,86; ACC=94,4%). Em contraste, classes como metabolômica, metagenômica 16S, metatranscriptômica e bulk transcriptomics não apresentaram métricas de acurácia, refletindo limitações na validação clínica ou caráter exploratório dos estudos. Os resultados reforçam que plataformas de maior resolução celular e molecular concentram os desempenhos mais elevados, enquanto abordagens clássicas permanecem subexploradas no contexto da periodontite.

A análise dos valores de desempenho (AUC) estratificados por classe ômica evidenciou diferenças marcantes na acurácia diagnóstica entre as abordagens avaliadas (Figura 4). A proteômica apresentou os maiores valores, com AUC variando de 0,88 a 0,97 (mediana 0,97), sustentada por validações em coortes independentes e meta-análises, refletindo sua maturidade e aplicabilidade translacional. Em seguida, a transcriptômica de célula única demonstrou desempenho igualmente elevado (AUC = 0,922), confirmando o potencial das tecnologias de alta resolução celular em capturar assinaturas moleculares robustas relacionadas à periodontite.

A metagenômica shotgun também se destacou com AUC consistente em 0,86, complementada por valores de acurácia acima de 94%, evidenciando sua utilidade na caracterização microbiana e no desenvolvimento de modelos preditivos. Abordagens mais exploratórias, como análises metodológicas isoladas e arrays funcionais/genômicos, apresentaram desempenho inferior (AUC de 0,80 e 0,73, respectivamente), refletindo tanto limitações no desenho dos estudos quanto menor maturidade tecnológica.

Por outro lado, classes amplamente investigadas como metabolômica, metagenômica 16S, metatranscritômica e bulk transcriptomics não apresentaram métricas reportadas de AUC, restringindo-se a análises conceituais ou descritivas. Essa ausência ressalta a necessidade de padronização metodológica e validação em larga escala para que essas abordagens possam alcançar níveis comparáveis de aplicabilidade clínica.

Em síntese, os resultados reforçam que proteômica, transcriptômica single-cell e metagenômica shotgun concentram os desempenhos mais elevados e consistentes, configurando-se como plataformas promissoras para a translação de biomarcadores em ferramentas diagnósticas. Já as demais tecnologias permanecem em estágio exploratório, carecendo de estudos com maior robustez metodológica e validação clínica.

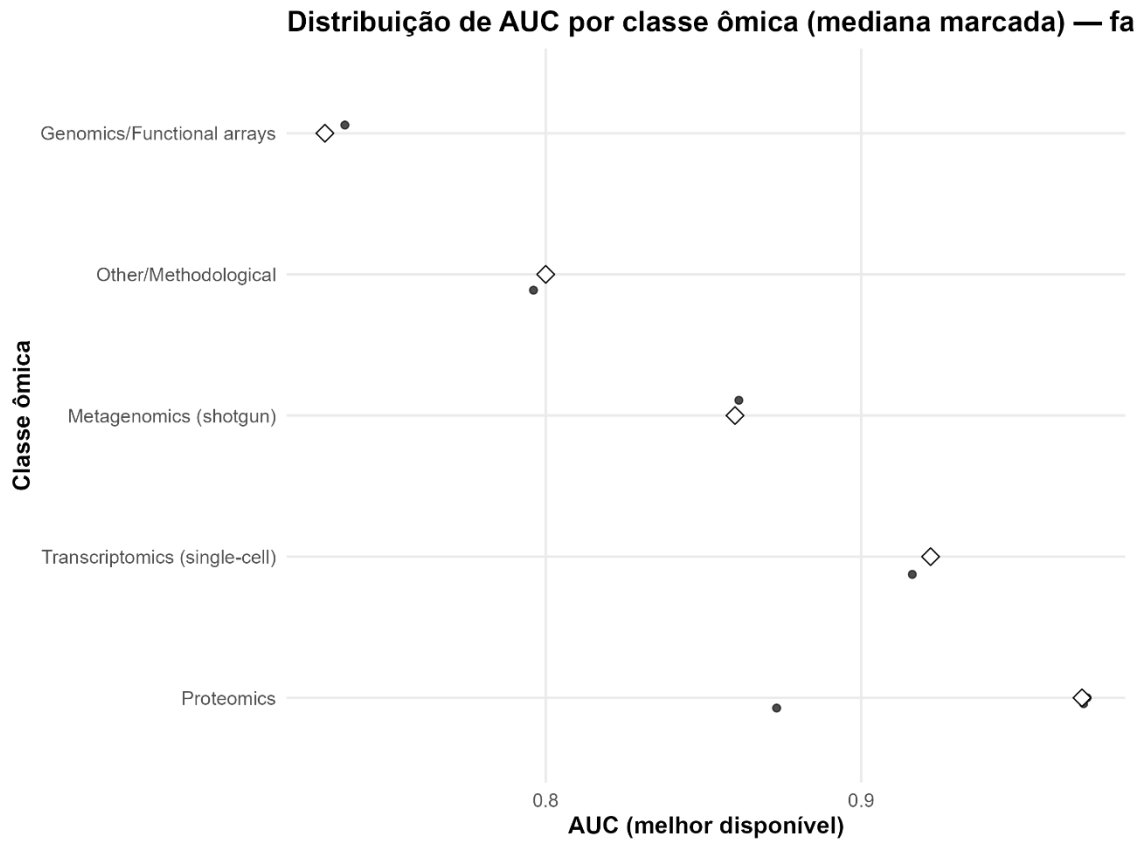

Figura 4. Distribuição dos valores de desempenho (AUC) reportados em estudos de periodontite segundo a classe ômica. Observa-se superioridade das abordagens proteômicas (AUC 0,88–0,97; mediana 0,97), seguidas pela transcriptômica de célula única (AUC 0,922) e pela metagenômica shotgun (AUC 0,86). Estratégias metodológicas isoladas e arrays genômicos/funcionais apresentaram desempenho mais modesto (AUC 0,80 e 0,73, respectivamente). Classes como metabolômica, metagenômica 16S, metatranscritômica e transcriptômica bulk não reportaram valores de AUC, permanecendo em estágios conceituais ou descritivos.

A Figura 5 integra, em um fluxo Ômica → Aplicação → Prontidão, como as plataformas têm sido convertidas em uso clínico e quão maduras elas estão. Observa-se uma clara concentração das trilhas em Diagnosis/Screening, alimentada sobretudo por proteômica, metagenômica (shotgun e 16S) e transcriptômica de célula única; esse eixo reúne desde pipelines proteômicos validados em saliva com *machine learning* até classificadores microbiômicos multi-coorte e modelos diagnósticos baseados em scRNA-seq. Os caminhos para Monitoring/Treatment aparecem em menor volume — com exemplos como 16S para acompanhamento pós-terapia e dispositivos ponto-de-cuidado que combinam marcadores de hospedeiro e patógenos — enquanto Risk/Prognosis e Drug/Targeting surgem como trilhas específicas e menos frequentes (p.ex., modelos com SNPs e priorização de alvos via RNA-seq). No eixo de prontidão clínica, predomina um bloco Conceitual/Potential (ancorado por revisões, meta-análises e estudos exploratórios em metabolômica, transcriptômica *bulk* e metatranscritômica), seguido por um subconjunto Promising/Feasible (impulsionado por proteômica com validação

independente, scRNA-seq e monitoramento 16S) e por pontos isolados de High accuracy/Pilot, notadamente scRNA-seq com *deep learning* (AUC ≈ 0,92) e metagenômica+ML com validação interna/externa. Em síntese, o panorama indica que plataformas mais consolidadas e acessíveis sustentam a aplicação diagnóstica já em patamares Promising/Feasible, enquanto tecnologias emergentes ampliam a profundidade mecanística e delineiam a transição para integração multi-ômica com validação clínica mais ampla.

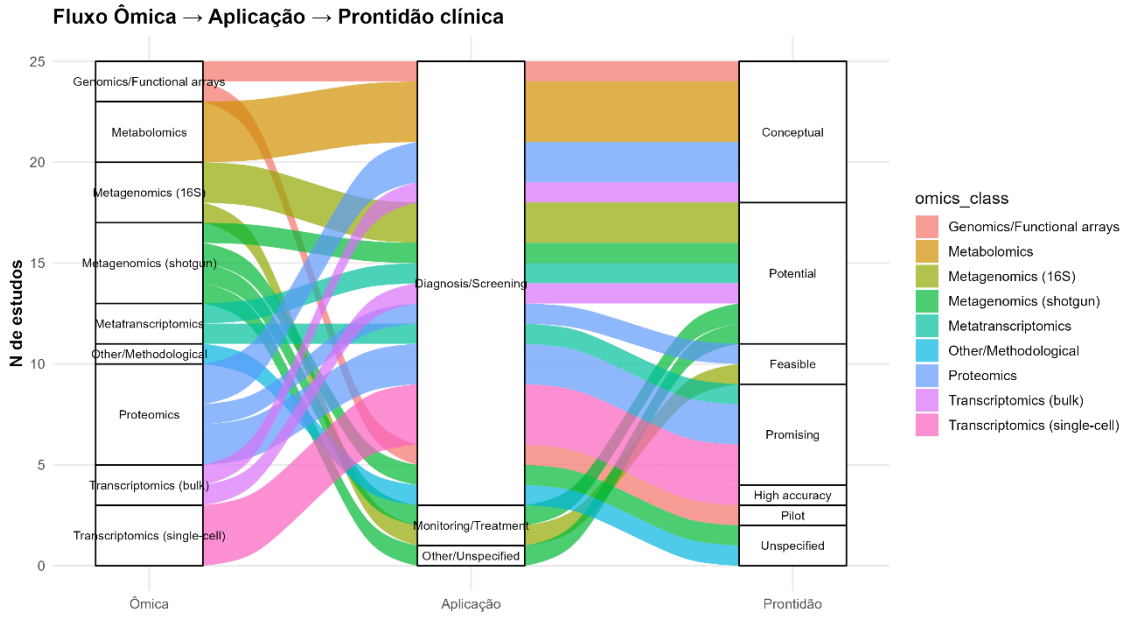

Figura 5. Fluxo alluvial conectando classes ômicas, aplicações clínicas e estágio de prontidão. Observa-se uma predominância de diagnóstico/triagem, fortemente sustentada por proteômica, metagenômica (shotgun e 16S) e transcriptômica de célula única. Aplicações de monitoramento/tratamento e prognóstico/risco aparecem em menor escala, enquanto iniciativas de descoberta de alvos terapêuticos são pontuais. No eixo de prontidão, predominam os estágios conceitual/potencial, seguidos por subconjuntos promissores/viáveis, com poucos exemplos de alta acurácia ou piloto, notadamente em *single-cell RNA-seq* com *deep learning* e metagenômica associada a *machine learning*. O panorama evidencia o avanço gradual da área, com plataformas tradicionais consolidando o diagnóstico e tecnologias emergentes delineando caminhos para personalização e validação clínica ampliada.