UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MATHEUS ANDRADE RODRIGUES

EXPLORANDO TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE NA ANÁLISE DE DNA DE GRUPOS JUDAICOS E DE OUTROS GRUPOS ÉTNICOS: UMA COMPARAÇÃO ENTRE PCA, T-SNE E UMAP.

CAMPINA GRANDE - PB

2023

MATHEUS ANDRADE RODRIGUES

EXPLORANDO TÉCNICAS DE REDUÇÃO DE DIMENSIONALIDADE NA ANÁLISE DE DNA DE GRUPOS JUDAICOS E DE OUTROS GRUPOS ÉTNICOS: UMA COMPARAÇÃO ENTRE PCA, T-SNE E UMAP.

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador : Tiago Lima Massoni

CAMPINA GRANDE - PB

2023

MATHEUS ANDRADE RODRIGUES

Explorando técnicas de redução de dimensionalidade na análise de DNA de grupos judaicos e de outros grupos étnicos: Uma comparação entre PCA, t-SNE e UMAP.

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Tiago Lima Massoni
Orientador – UASC/CEEI/UFCG

Patrícia Duarte de Lima Machado
Examinador – UASC/CEEI/UFCG

Francisco Vilar Brasileiro
Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 28 de Junho de 2023.

CAMPINA GRANDE - PB

RESUMO

Aplicamos PCA, t-SNE e UMAP nos datasets de calculadoras de interpretação genética com dados de grupos étnicos judaicos, de vários vizinhos não-judeus e de etnias correlacionadas, utilizando o software R. Realizamos uma comparação visual entre os resultados gerados e utilizamos o microbenchmark para verificar o tempo de execução dos métodos. O t-SNE e o UMAP são eficientes para trabalharmos com âmbitos locais da visualização, enquanto o PCA é adequado quando o número de amostras é reduzido. t-SNE e UMAP são capazes de formar agrupamentos que não veríamos somente utilizando o PCA. Apesar disso, são mais lentos que o PCA, e as visualizações geradas por eles mudam ao executar o algoritmo novamente.

EXPLORING DIMENSIONALITY REDUCTION TECHNIQUES IN DNA ANALYSIS OF JEWISH AND OTHER ETHNIC GROUPS: A COMPARISON OF PCA, T-SNE, AND UMAP.

ABSTRACT

We applied PCA, t-SNE, and UMAP to datasets from genetic interpretation calculators containing data of Jewish ethnic groups, various non-Jewish neighbors, and correlated ethnicities, using the R software. We conducted a visual comparison of the generated results and used microbenchmarking to measure the execution time of the methods. t-SNE and UMAP are efficient for working with local aspects of visualization, while PCA is suitable when the number of samples is small. t-SNE and UMAP are capable of forming clusters that would not be seen using PCA alone. However, they are slower than PCA, and the visualizations generated by them change when the algorithm is run again.

.

Explorando técnicas de redução de dimensionalidade na análise de DNA de grupos judaicos e de outros grupos étnicos: Uma comparação entre PCA, t-SNE e UMAP.

Matheus Andrade Rodrigues
matheus.andrade.rodrigues@ccc.ufcg.edu.br
Unidade Acadêmica de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba

Tiago Lima Massoni
massoni@dsc.ufcg.edu.br
Unidade Acadêmica de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba

RESUMO

Aplicamos PCA, t-SNE e UMAP nos datasets de calculadoras de interpretação genética com dados de grupos étnicos judaicos, de vários vizinhos não-judeus e de etnias correlacionadas, utilizando o software R. Realizamos uma comparação visual entre os resultados gerados e utilizamos o microbenchmark para verificar o tempo de execução dos métodos. O t-SNE e o UMAP são eficientes para trabalharmos com âmbitos locais da visualização, enquanto o PCA é adequado quando o número de amostras é reduzido. t-SNE e UMAP são capazes de formar agrupamentos que não veríamos somente utilizando o PCA. Apesar disso, são mais lentos que o PCA, e as visualizações geradas por eles mudam ao executar o algoritmo novamente.

Keywords

PCA, t-SNE, UMAP, DNA, redução de dimensionalidade, visualização de dados, etnias, judeus, Oriente Médio.

REPOSITÓRIO

Repositório no GitHub contendo datasets e scripts utilizados no R: https://github.com/matheusyhwh/ethnic-data-visualization

AVISO LEGAL

O presente trabalho não é patrocinado por nenhum grupo, empresa ou pessoa, e não possui finalidade comercial ou política. Não é o objetivo do trabalho servir de apoio para quaisquer dos lados no conflito israelo-palestino ou algum outro conflito étnico, cultural ou social que envolva os povos citados aqui. Os autores também reafirmam que a classificação étnica é demasiadamente complexa, e vai muito além do âmbito genético - chegando, inclusive, aos âmbitos religiosos e linguísticos.

1. INTRODUÇÃO

Representar graficamente um vetor com duas variáveis é simples e pode ser feito com um básico gráfico bidimensional. Um dado com três variáveis pode ser representado com um gráfico tridimensional. Mas quando passamos disso, não há meios físicos de mapear essas informações perfeitamente nos eixos de um gráfico. Por exemplo: caso queiramos analisar a relação entre peso e altura de indivíduos, poderíamos utilizar o eixo x para representar os valores de peso e o eixo y para representar as alturas. Caso ainda queiramos adicionar uma terceira variável que represente a ingestão calórica diária, poderíamos criar um eixo z e utilizá-lo para representar este dado. Contudo, caso desejemos analisar uma quarta variável - como a quantidade de horas de atividade física semanal - não podemos adicionar um quarto eixo e utilizá-lo para visualizar esta relação. Nestes casos, faz-se uso de outros métodos para contornar esta limitação - cores, tamanhos e formatos dos pontos exibidos para representar outras características. Mas conforme temos dados de dimensões maiores, esta tarefa torna-se cada vez mais complexa e não-intuitiva. Muitos problemas de aprendizagem de máquina possuem datasets com dimensão na casa dos milhares. Isso pode deixar o treinamento dos dados extremamente custoso e impossibilitar a obtenção de bons resultados. A este desafio a literatura técnica deu o nome de Problema da Dimensionalidade, ou Maldição da Dimensionalidade - termo cunhado por Richard E. Bellman em 1957 em seu livro Dynamic Programming.[1]

Para resolver este problema, o ideal seria utilizar dados com menos variáveis. Entretanto, esta não será uma opção válida no mundo real, onde encontraremos frequentemente enormes conjuntos de dados de altíssimo número de variáveis nos campos de pesquisa em data mining e machine learning, entre outros. Quanto maior o número de variáveis, maior será o poder de processamento necessário para realizar a análise destes dados e também a quantidade de dados de treinamento necessários para criar modelos úteis [2].

Para contornar esta situação, utilizam-se técnicas de redução de dimensionalidade. Assim poderemos representar informações de dimensão elevada num gráfico bidimensional. Conforme novos métodos de redução de dimensionalidade surgem, faz-se útil a comparação entre eles, para averiguarmos quais melhor se enquadram nos escopos dos trabalhos. Um método muito utilizado para a análise visual dos dados é o PCA, porém mais recentemente outros métodos - como o t-SNE e o UMAP - surgem para servir de alternativa aos pesquisadores que desejam representar graficamente dados de alta dimensão.

Aplicamos os três métodos acima em datasets de resultados de análises genéticas com o DNA autossômico de comunidades judaicas e de diversos vizinhos não-judeus, e verificamos se os mapas gerados por estas visualizações refletiam adequadamente a proximidade entre grupos étnicos relacionados historicamente e/ou geograficamente. Após a análise visual dos mapas gerados por estes três métodos, posteriormente utilizamos o pacote microbenchmark para mensurar o tempo de execução de

cada um dos métodos, assim nos permitindo avaliar e destacar as diferenças de desempenho entre eles.

As visualizações geradas pelos métodos utilizando DNAs de grupos étnicos nos permitem compreender melhor a diversidade genética entre diferentes populações. Essa diversidade genética reflete a história evolutiva da humanidade, incluindo migrações e misturas étnicas. O estudo do DNA de grupos étnicos contribui para a compreensão da história evolutiva da humanidade, e pode fornecer indicativos que auxiliem no entendimento de êxodos antigos, misturas genéticas e adaptações a diferentes ambientes e culturas. Isso nos ajuda a reconstruir as rotas migratórias, entender a diversidade genética atual e explorar a relação entre genética e cultura humana.

Para fins de estudo de caso, o foco se deu nas etnias judaicas, que foram comparadas com as populações não-judaicas das localidades onde residiram. Ex: ashkenazim e centro/leste-europeus, mizrahim e comunidades árabes, persas, etc. As populações utilizadas foram os grupos judaicos devido ao histórico de isolamento genético que estas comunidades possuem em relação aos grupos não-judaicos. O software utilizado para implementação dos scripts foi o R.

2. BACKGROUND

2.1 Etnias judaicas

Os judeus são um grupo etno-religioso cuja origem remonta ao antigo Israel. Após as guerras judaico-romanas, os judeus foram expulsos de suas áreas urbanas e impedidos de visitar seu centro religioso, Jerusalém, sob pena de morte. Além disso, a cidade de Jerusalém, foi renomeada para Élia Capitolina, e a província foi renomeada como Siria Palestina [3][4]. Inicialmente, as principais comunidades judaicas estavam concentradas na Terra de Israel e na Babilônia, devido a uma Diáspora anterior, ocasionada pelas deportações dos judeus pelas tropas babilônicas de Nabucodonosor, que ocorreram no sexto século a.C. [5]. Após esta nova Diáspora pelo Império Romano, as populações judaicas concentraram-se principalmente no Norte da Terra de Israel (Galiléia), Ásia Menor, Itália e Mesopotâmia [6]. No decorrer da Idade Média, foram-se formando grupos judaicos com culinária, vestimentas, idiomas e liturgias distintas em diversos pontos da Europa, África e Ásia, e não mais o centro da produção cultural e religiosa dos judeus era a Terra de Israel. Os principais grupos atualmente são os judeus Ashkenazim, Mizrahim e Sephardim. Os Ashkenazim habitavam a região do centro e do leste europeu; os Mizrahim habitavam no Oriente Médio, Ásia Central e Norte da África; e os sephardim habitavam em regiões ao redor do Mediterrâneo: Península Ibérica (Espanha e Portugal), Império Otomano e Norte da África. Ainda há comunidades que distinguem-se destas citadas e preferem apresentar-se como grupos próprios, como por exemplo os Italkim (Itália), Romaniotes (Grécia), Kavkazim (Região do Cáucaso), e os Teimanim (Yemen).

Há certa variação quanto estas classificações, portanto, em certas análises poderemos ver Teimanim incluídos no grupo dos Mizrahim, ou os Parsim (Irã e Afeganistão) ou os Bukharim (Usbequistão e Tajiquistão) não incluídos nos Mizrahim, por exemplo. Por vezes, os grupos eram divididos apenas entre Ashkenazim e Sephardim (fusionando Mizrahim com Sephardim). Atualmente, todas estas comunidades possuem grande parte de seus membros concentrada no moderno Estado de Israel, que concentra cerca de 45% de toda a população judaica do mundo [7].

2.2 Análises de DNA e os SNPs

O DNA (sigla para “Ácido Desoxirribonucleico”) contém as informações genéticas dos indivíduos. É composto por duas tiras espiralizadas chamadas de dupla hélice. Essas duas tiras são ligadas por nucleotídeos: pares de bases nitrogenadas identificadas pelas letras A (adenina), T (timina), C (citosina) e G (guanina). São as unidades que compõem o DNA [8].

Para ver os resultados genéticos, o indivíduo deve encomendar uma análise de DNA em alguma empresa do ramo. Após isso, receberá um kit com ferramental para coletar uma amostra biológica. Como a amostra requerida normalmente é a saliva, o indivíduo receberá cotonetes para friccionar contra a parte interna da bochecha, ou um tubo para depositar alguns mililitros de saliva, por exemplo. Após ter realizado a coleta, o indivíduo envia a amostra para a empresa, que fará a extração do DNA presente naquela porção. Após este processo, a empresa fará o sequenciamento dos dados num arquivo chamado popularmente de Raw DNA Data (Dado de DNA cru).

Em 2021 houve um aumento na demanda por kits de DNA para descoberta de ancestralidade de até 700% [9]. A empresa Genera, uma deste ramo a atuar no Brasil, declarou que suas vendas em 2019 aumentaram 14 vezes em relação a 2018, e em 2021 novamente aumentaram, em 20 vezes em relação a 2019 [10][11]. Várias empresas fazem a coleta do DNA e fornecem o resultado com as respectivas porcentagens étnicas do usuário. Esta é uma tendência que possivelmente continuará pelos próximos anos.

Todos os humanos possuem quase a mesma sequência de três bilhões de bases de DNA distribuídos pelos 23 pares de cromossomos - mais de 99% do DNA é idêntico. No entanto, existem variações em certas posições chamadas polimorfismos, que são responsáveis por diferenciar um indivíduo de outro [12]. O Raw DNA é um arquivo digital normalmente de extensão .csv, .txt ou .tsv, onde cada linha deste arquivo é referente a um SNP (Polimorfismo de nucleotídeo único). O Raw DNA conterá: 1-o código RSID (O identificador do SNP), 2-o número do cromossomo onde aquela informação se encontra, 3- a posição do cromossomo e 4- o resultado do genótipo com os nucleotídeos (dois alelos, que podem ser A, C, G, T)[13][14]. Na tabela 1 é possível ver como estas informações estão organizadas no arquivo.

Tabela 1 - exemplo de seção de um arquivo Raw DNA.

cod.RSID	cromossomo	posição	genótipo
rs4477212	1	82154	AA
rs3094315	1	752566	AC
rs3131972	1	752721	GG
rs12562034	1	768448	--
rs12124819	2	776546	CT

O arquivo Raw DNA fornecido pelas empresas em testes simples de DNA geralmente contém por volta de 600 mil a 700 mil linhas.

2.3 Calculadoras étnicas

Após obter o Raw DNA, o usuário pode realizar o upload deste arquivo em diversos sites que oferecem serviços de interpretação dos SNPs - desde serviços de saúde até calculadoras étnicas. Há no mercado atual um crescente número de empresas que possuem calculadoras privadas, onde o usuário deve comprar os kits para ter acesso ao serviço. Contudo, há outros sites de genética e genealogia que possuem calculadoras gratuitas. Um exemplo é o GEDmatch, sendo ele o maior site de calculadoras genéticas gratuitas. O GEDmatch é um site onde o usuário pode fazer upload de seu(s) arquivo(s) de DNA e ver seus resultados em diversas calculadoras étnicas. Em algumas calculadoras é possível ver o spreadsheet com a média dos resultados de várias etnias. A calculadora Eurogenes K13 possui 13 dimensões, por exemplo. A MDLP 22 possui vinte e duas dimensões.

No GEDmatch há tabelas que mostram quais são as pontuações médias para diversas etnias (Tabela 2). Cada coluna representa a porcentagem étnica identificada pela calculadora, e cada linha corresponde a uma etnia. O site foi criado em 2010, por Curtis Rogers e John Olson, e nele são oferecidas gratuitamente diversas calculadoras étnicas (além de outros serviços, como navegador de cromossomos, lista de pessoas com possível parentesco e possibilidade de montagem de árvore genealógica). Até Abril de 2021, o GEDmatch possuía em seu portfólio 37 calculadoras, que estão distribuídas em 7 projetos, a saber: MDLP Project, Eurogenes, Dodecad, HarappaWorld, EthioHelix, puntDNAL e GedrosiaDNA.

O presente estudo terá como dataset estas tabelas de interpretação dos SNPs, disponibilizadas pelo GEDmatch com as médias dos resultados étnicos de vários grupos. Três métodos de redução de dimensionalidade serão aplicados nas tabelas de duas calculadoras: Eurogenes K13 e MDLP (Magnus Ducatus Lituaniae Project) 22. As calculadoras foram escolhidas devido à diversidade de povos judaicos que possuem. A Eurogenes K13 é uma das mais utilizadas, e a MDLPK23b é a que possui maior dimensão dentre as calculadoras que disponibilizaram tabelas e possui mais de seiscentas etnias, mas a MDLP 22 foi preferida pois é menos extensa e também possui os grupos judaicos que a K23b possui.

Tabela 2: exemplo de seção do dataset da calculadora Eurogenes K13

label	North_Atlantic	Baltic	West_Med	West_Asian	East_Med	Red_Sea	South_Asian	East_Asian	Siberian	Amerindian	Oceanian	Northeast_African	Sub-Saharan
Abhkasian	1.64	4.62	9.81	54.30	22.78	1.84	1.91	1.48	1.11	0.17	0.33	0.01	0.01
Algerian	10.96	0.51	22.38	0.11	28.26	15.84	0.08	0.39	0.08	0.20	0.23	10.44	10.51
Algerian_Jewish	12.58	3.47	22.86	10.35	37.03	7.63	0.75	0.51	0.27	0.00	0.43	2.22	1.91
Armenian	2.79	0.85	13.02	38.87	34.79	5.23	3.41	0.43	0.03	0.05	0.44	0.04	0.05
Ashkenazi	15.32	10.16	18.44	10.29	39.61	5.98	01.04	0.79	0.97	0.46	0.29	1.15	0.50
Assyrian	1.41	1.84	11.50	32.63	39.64	7.78	3.76	0.71	0.20	0.19	0.15	0.19	0.02
Bedouin	1.25	1.19	11.06	14.43	38.15	21.40	1.85	0.25	0.69	0.27	0.21	6.33	2.91
Belorussian	29.37	49.40	7.59	3.87	4.98	1.44	0.97	0.18	1.32	0.18	0.38	0.20	0.13

3. METODOLOGIA

O trabalho utiliza a abordagem quanti-qualitativa (mista). Criamos um script R para aplicação de três métodos de redução de dimensionalidade (PCA, t-SNE e UMAP) nos datasets de interpretação dos SNPs de diversos grupos étnicos. A máquina utilizada foi um notebook com processador i5 5200U, 8GB de memória RAM DDR3L, e os algoritmos foram executados diversas vezes anteriormente para estudo dos autores, e somente uma vez para a captação dos screenshots presentes no apêndice. O software foi o R, versão 4.0.3.

A análise visual dos gráficos gerados pelos métodos é qualitativa. Etnias relacionadas geograficamente muitas vezes também o são geneticamente, portanto as visualizações geralmente irão posicionar estas etnias próximas umas das outras (por exemplo, portugueses e espanhóis). Caso etnias próximas geograficamente tenham ficado dispostas muito distantes entre si na visualização, verificamos o porquê disso ocorrer: por ser uma exceção (por exemplo, no caso das etnias judaicas, que geralmente são mais próximas de outras etnias judaicas do que seus vizinhos não-judeus), ou um erro de visualização. A análise com microbenchmark para avaliar o tempo de execução de cada um dos métodos é quantitativa.

Para a coleta de dados, fizemos o download dos datasets e aplicamos uma limpeza neles, removendo etnias que não estão no escopo da pesquisa. Para a realização das análises, utilizamos o R. Para ver os scripts criados e os datasets, o leitor pode acessar o repositório do GitHub do projeto, encontrado no link fornecido no início do trabalho. Para ver os datasets também é possível entrar no site do GEDmatch e pesquisar por Admixture (heritage), escolher o projeto público desejado, rodar a calculadora e ver a tabela (spreadsheet), ou ainda entrar em contato com os vários autores, colaboradores e pesquisadores que forneceram as calculadoras. Por exemplo, a calculadora Eurogenes foi criada em 2011 por David Wesolowski, que mantém o Blog Eurogenes [15], onde discute sobre populações antigas e mantém contato com estudantes e entusiastas do tema.

As amostras dos DNAs utilizadas para a formulação das tabelas são extraídas pelos colaboradores de artigos científicos publicados previamente e disponibilizados para pesquisa posterior [16] ou sites de arquivos genéticos para utilização de pesquisadores, como European Genome-phenome Archive [17] gnomAD: Genome Aggregation Database [18] e Allen Ancient DNA Resource (AADR) [19], entre outras fontes que disponibilizam material deste cunho aos pesquisadores.

Os métodos de redução de dimensionalidade possuem como finalidade realizar um *trade-off* onde possam representar o conjunto de dados num espaço de dimensão menor de forma que a acurácia não seja demasiadamente afetada com esse mapeamento. Eles possibilitam que seja possível a visualização de dados de alta dimensão em gráficos bidimensionais ou tridimensionais, por exemplo. Os métodos abaixo foram utilizados com a linguagem de programação e ambiente R.

3.1 PCA

O PCA (Análise de Componente Principal) é um método que verifica a importância das variáveis do dataset para a variância no eixo do gráfico. As variáveis de maior importância são aquelas que ocasionam maior distância entre pontos extremos do gráfico. A combinação linear que possui maior variação é chamada de Componente Principal 1. A que possui segunda maior variação é chamada de Componente principal 2, até Componente Principal n, onde n é o número de variáveis do dataset. O PCA obtém variáveis não-correlacionadas através das variáveis correlacionadas, que são chamadas de componentes principais (PCs) idealmente com perda mínima de informações [20].

Para a realização do PCA, deve-se padronizar as variáveis caso elas estejam fora de escala (variável de 0 a 1, enquanto todas as outras são de 0 a 50, por exemplo). Isso pode ser feito diminuindo o valor da variável pela média e dividindo pelo desvio padrão.

$$z = \frac{\text{value - mean}}{\text{standard deviation}}$$

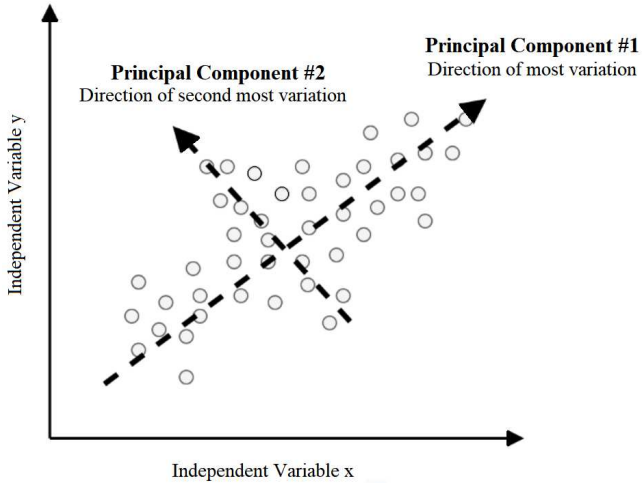
Calcula-se então a matriz de covariância. Esta matriz é simétrica e possui p linhas e colunas, sendo p o número de variáveis no dataset. A covariância de uma variável consigo mesma é sua variância (Cov (a, a) = Var (a)), portanto, na diagonal principal temos as variações de cada variável inicial. E como a covariância é comutativa (Cov (a, b) = Cov (b, a)), as porções triangulares superior e inferior da matriz são iguais.[11]

$$\begin{bmatrix} \text{Cov(x, x)} & \text{Cov(x, y)} & \text{Cov(x,z)} \\ \text{Cov(y, x)} & \text{Cov(y, y)} & \text{Cov(y,z)} \\ \text{Cov(z, x)} & \text{Cov(z, y)} & \text{Cov(z,z)} \end{bmatrix}$$

Exemplo de matriz de covariância com três variáveis

Isso resultará no seguinte cenário: Variáveis que possuem dependência entre si terão covariância positiva. Variáveis independentes possuirão covariância igual a zero. Variáveis inversamente dependentes possuirão valores negativos. Quanto maior a dependência entre variáveis, maior será a covariância.

A partir dos resultados da matriz de covariância, consegue-se os autovetores e autovalores para que sejam obtidos os componentes principais.



O PCA calcula então os pesos que cada variável possuirá nos componentes principais. O vetor com os pesos que compõem aquele Componente Principal é chamado de autovetor. Os Componentes Principais são novas variáveis construídas através dessas misturas das variáveis iniciais. Os Componentes Principais são construídos de forma que não possuam correlação entre si. A maioria das informações dentro das variáveis iniciais é comprimida no primeiro componente. Um dataset de n dimensões fornecerá n Componentes Principais, e o PCA tenta colocar o máximo de informações possíveis no primeiro componente, o máximo de informações restantes no segundo e assim por diante. Portanto, os componentes 1 e 2 serão os mais adequados para que o mapa étnico seja montado. Os componentes principais são menos interpretáveis e não têm nenhum significado real, uma vez que são construídos como combinações lineares das variáveis iniciais.pense nos componentes principais como novos eixos que fornecem o melhor ângulo para ver e avaliar os dados, de modo que as diferenças entre as observações sejam mais visíveis.é a linha que maximiza a variância (a média das distâncias quadradas dos pontos projetados (pontos vermelhos) até a origem). O segundo componente principal é calculado da mesma maneira, com a condição de que não esteja correlacionado com (ou seja, perpendicular a) o primeiro componente principal e que seja responsável pela próxima variância mais alta. Autovetores e autovalores sempre vêm em pares, de modo que cada autovetor tem um autovalor. E seu número é igual ao número de dimensões dos dados. Por exemplo, para um conjunto de dados tridimensional, existem 3 variáveis, portanto, existem 3 autovetores com 3 autovalores correspondentes.os autovetores da matriz de covariância são na verdade as direções dos eixos onde há mais variância (mais informações) e que chamamos de Componentes Principais. E os autovalores são simplesmente os coeficientes anexados aos autovetores, que fornecem a quantidade de variância carregada em cada componente principal. Classificando seus autovetores em ordem de seus autovalores, do

mais alto para o mais baixo, obtemos os componentes principais em ordem de significância [21].

3.2 t-SNE

Assim como o PCA, o t-SNE (t-distributed stochastic neighbor embedding) é um método de redução de dimensionalidade e visualização de dados de alta dimensão. Foi desenvolvido por Laurens van der Maaten e Geoffrey Hinton em 2008 [22].

Primeiro, a similaridade entre todos os pares de pontos é calculada. Normalmente, é utilizada uma medida de similaridade, como a distância euclidiana, para determinar a proximidade entre os pontos no espaço de alta dimensão. Essas similaridades são então convertidas em probabilidades de similaridade, onde pontos mais próximos têm probabilidades maiores. O próximo passo é construir uma matriz de probabilidades de similaridade conjunta, que representa as similaridades entre todos os pares de pontos. Essa matriz é simétrica e captura a similaridade entre os pontos tanto no espaço de alta dimensão quanto no espaço de menor dimensão. A matriz é construída usando um kernel de similaridade (normalmente um kernel Gaussiano) aplicado às probabilidades de similaridade calculadas anteriormente.

O t-SNE visa minimizar a divergência entre as distribuições de similaridade no espaço de alta dimensão e no espaço de menor dimensão. Para isso, utiliza-se um processo iterativo de otimização para encontrar uma configuração de pontos no espaço de menor dimensão que seja consistente com as probabilidades de similaridade conjunta. Isso é alcançado minimizando uma função de custo que representa a divergência entre as distribuições. Durante o processo de otimização, é utilizado um método chamado gradiente descendente estocástico para ajustar a posição dos pontos no espaço de menor dimensão. O gradiente é calculado com base na função de custo e é usado para atualizar as posições dos pontos iterativamente. O processo continua até que uma condição de parada seja atingida, como um número máximo de iterações ou uma convergência satisfatória. Após a conclusão do processo de otimização, a configuração final dos pontos no espaço de menor dimensão é obtida. Essa configuração pode ser visualizada para analisar a estrutura e os agrupamentos presentes nos dados de alta dimensão [23][24][25].

O t-SNE é uma técnica estocástica e não determinística, o que significa que os resultados podem variar a cada execução. Além disso, o t-SNE é uma abordagem não linear e não preserva a distância entre os pontos originais (ver fig. 3 e 4).

3.3 UMAP

O primeiro passo é construir uma estrutura de vizinhança para capturar as relações de proximidade entre os pontos de dados. Isso envolve calcular as distâncias ou similaridades entre os pontos no espaço de alta dimensão e identificar os vizinhos mais próximos para cada ponto. A estrutura de vizinhança é construída de forma a capturar tanto as relações locais quanto as globais entre os pontos. Em seguida, o UMAP utiliza um grafo de vizinhança ponderado, no qual os pontos de dados são representados como nós e as conexões entre eles são representadas por arestas com pesos. A otimização do grafo de vizinhança é realizada usando técnicas de otimização, como descida de gradiente estocástico, para ajustar os pesos das arestas de forma a preservar as relações de vizinhança dos pontos originais no espaço de menor dimensão.

Após a otimização do grafo de vizinhança, um gráfico de distâncias é criado. Nesse gráfico, cada ponto de dados é representado como um nó e as distâncias entre eles são representadas como arestas ponderadas. O gráfico de distâncias captura a conectividade global dos dados e é usado para preservar a estrutura global dos agrupamentos durante a projeção. Finalmente, o algoritmo UMAP realiza a projeção dos dados de alta dimensão para um espaço de menor dimensão. Isso envolve a minimização de uma função de custo que leva em consideração as relações de vizinhança preservadas no espaço de menor dimensão. A projeção é realizada usando técnicas de otimização não linear, como descida de gradiente estocástico, para encontrar uma configuração de pontos que minimize a função de custo [26].

4. RESULTADOS

4.1 Pré-processamento dos dados

O pré-processamento consiste em uma série de manipulações dos dados com o objetivo de melhorar a qualidade do dataset, facilitar seu manuseio e melhorar as visualizações geradas a partir dele. Foram feitas modificações nos dois conjuntos de dados utilizados: o da calculadora MDLP 22 e o da calculadora Eurogenes K13.

Os datasets foram convertidos de .txt para .tsv para serem melhor manipulados. Cada linha representa uma etnia, e nas colunas estão as porcentagens da interpretação do DNA pela calculadora para aquele grupo étnico. Foram removidas etnias que estão fora do escopo do estudo, como grupos étnicos ameríndios, oceânicos e siberianos. Esses grupos não possuem relações genéticas próximas ou não mantiveram contato cultural substancial com os grupos étnicos judaicos ao longo da história. As porcentagens, anteriormente separadas por vírgulas, foram alteradas para pontos decimais. Os arquivos foram então convertidos para o formato .csv para uso no software R.

Ao excluir amostras irrelevantes para a pesquisa, a visualização dos dados tornou-se mais clara e o tamanho do arquivo diminuiu significativamente. No caso do Eurogenes, o tamanho do arquivo reduziu de 4,8KB para 1,9KB, representando uma redução de 69,81%. Para o MDLP, o tamanho do arquivo .csv diminuiu de 33,3KB para 14,4KB, indicando uma redução de 56,76%.

O dataset MDLP 22 possui dados genéticos de 275 etnias, com vinte e duas variáveis - a saber: Pigmeu, Oeste-asiático, Norte-europeu-mesolítico, Indo-tibetano, mesoamericano, artico-amerindio, amerindio da América do Sul,Indiano, siberiano do norte, mediterrânico atlântico neolítico, samoiedos, indo-iraniano, Leste-siberiano, nordeste-europeu, sul africano, norte ameríndio, subsaariano, sudeste asiatico, oriente próximo, melanésio, paleo-siberio e austronésio. A calculadora MDLP pode ser utilizada para encontrar ancestralidades variadas, incluindo oceânicas e ameríndias e também pode servir de análise de DNA arcaico. Das duzentas e setenta e cinco etnias, vinte e duas possuem tradições orais com relatos de ancestralidade direta israelita: Dois grupos de judeus Ashkenazim, Judeus da Argélia, Judeus do Azerbaijão, dois grupos de Judeus da Etiópia, Judeus antigos da França, Judeus da Geórgia, Judeus da Índia, Judeus do Irã, dois grupos de judeus do Iraque, Judeus da Itália, Judeus curdos, Judeus da Líbia, Judeus do Marrocos, Judeus da Romênia, Judeus da Síria, Judeus tat, Judeus da Tunísia, Judeus do Uzbequistão, e Judeus do Iêmen, além dos Samaritanos. Após a limpeza dos dados, o número de etnias foi diminuído para 117.

O dataset Eurogenes possui dados genéticos de duzentas e quatro etnias, divididos em treze variáveis: Atlântico Norte, Báltico, Oeste Mediterrâneo, Oeste Asiático, Leste Mediterrâneo, Mar Vermelho, Sul da Ásia, Leste Asiático, Sibérico, Amerindio, Oceânico, Nordeste Africano e Africano Subsaariano. A calculadora é particularmente adequada para tratar de ancestralidade européia e médio-oriental. Das duzentas e quatro etnias, onze possuem em suas tradições orais relatos de ancestralidade direta israelita: Judeus da Argélia, Ashkenazim, Judeus da Geórgia, Judeus do Irã, Judeus da Itália, Judeus curdos, Judeus da Líbia, Samaritanos, Judeus sepharadim, Judeus da Tunísia e Judeus do Iêmen, além de também incluir os samaritanos. Após a limpeza dos dados, o número de etnias foi diminuído para 58.

Nas figuras 1 e 2 é possível ver o PCA da tabela Eurogenes, antes e depois do pré-processamento, respectivamente.

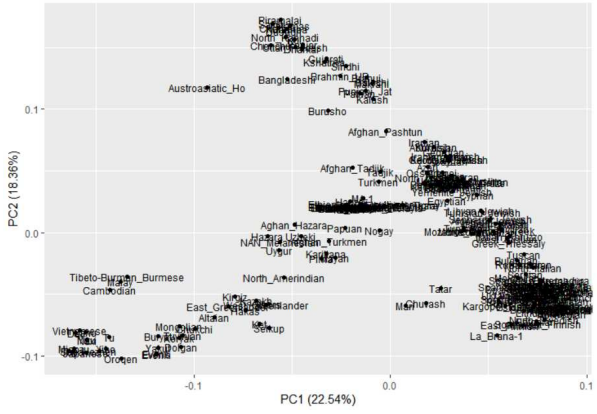


Figura 1: PCA do dataset sem pré-processamento dos resultados da calculadora Eurogenes K13. Como muitas etnias estão presentes, a visualização dos rótulos é afetada

Antes da remoção das etnias irrelevantes para o estudo, o PCA mostrou-se demasiadamente complexo e confuso. Entretanto, o t-SNE e o UMAP já apresentaram resultados inteligíveis.

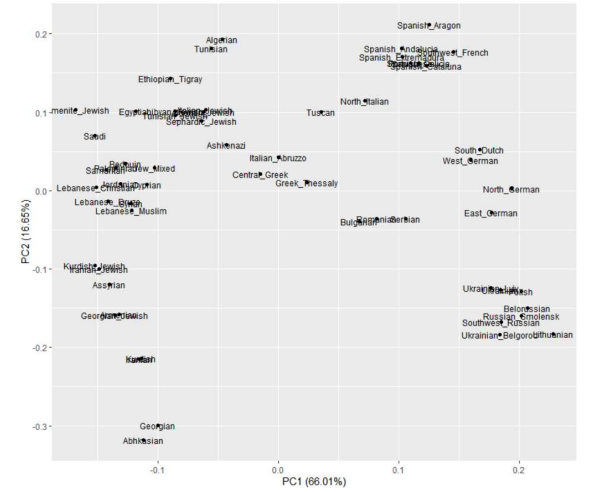


Figura 2: PCA do dataset após pré-processamento dos resultados da calculadora Eurogenes K13.

4.2 Aplicação dos métodos nos conjuntos de dados

O script que pode ser encontrado em nosso repositório GitHub gera o PCA de um .csv que esteja devidamente rotulado. Por exemplo: no caso do dataset Eurogenes, lemos o .csv e o atribuímos para a variável `seuDataset.` Depois, definimos que os rótulos serão a primeira coluna do dataset. Eles serão necessários pois precisamos ver os nomes das etnias, e não somente os pontos no gráfico. Em seguida, atualizamos o `seuDataset` excluindo a primeira coluna dele (usada para os rótulos) ao selecionar apenas a partir da segunda coluna até *n*, onde *n* é o número da última coluna. No caso do Eurogenes, o n será 14 e no MDLP 22, será 23.

É indispensável a instalação de todos os pacotes antes do uso. Caso ainda não possua os pacotes necessários (ggplot2, ggfortify, Rtsne, umap), basta ir em Pacotes > Instalar Pacotes e procurar por eles e realizar a instalação, e em seguida rodar os códigos disponibilizados no repositório.

Chamamos o pacote ggortify, e realizamos a análise dos componentes principais utilizando o prcomp. É possível omitir o scale. = TRUE caso você não queira realizar a normalização dos dados. Após isso, plotamos o gráfico com base na lista de valores retornada por prcomp.

O pacote prcomp realiza a análise de Componentes Principais nos dados fornecidos. O valor retornado é usado para a construção do gráfico através do pacote ggplot2 ou da função genérica autoplot. O gráfico terá duas dimensões. Os outros parâmetros fazem com que seja utilizado o rótulo das amostras ao invés do ponto, e o tamanho da fonte do rótulo é setado. Por fim, o software apresentará a tela com os rótulos dispostos com suas respectivas distâncias.

Finalizados os PCAs, iniciamos as análises das visualizações com o t-SNE. O pacote Rtsne realiza a plotagem do t-SNE no dataset passado como parâmetro. Caso ocorram valores duplicados ou duas ou mais amostras de uma mesma etnia, pode-se opcionalmente removê-las sentando check_duplicates para TRUE. Em termos práticos, a perplexidade diz respeito ao modo de agrupamentos: quanto menor for a perplexidade, mais ilhas com menor quantidade de componentes haverão; e quanto maior, mais dispersa a representação será. Além da perplexidade, podemos realizar diversas outras alterações, como o número máximo de iterações, e o valor de theta, que diz respeito ao trade-off entre velocidade e acurácia, sendo 0 o valor do t-SNE original, com maior acurácia, e 1 o valor para maior rapidez e menos acurácia (sendo o valor default 0.5). Utilizamos todos os valores default, e testamos uma perplexidade alta e uma baixa.

Por fim, o último método foi o UMAP: Chamamos o pacote umap, utilizamos o método default Naive, preparamos o dataframe e plotamos, semelhantemente ao realizado nos métodos anteriores.

Após a obtenção de todas as visualizações geradas pelos métodos, utilizamos o microbenchmark para verificar o tempo consumido por cada um deles.

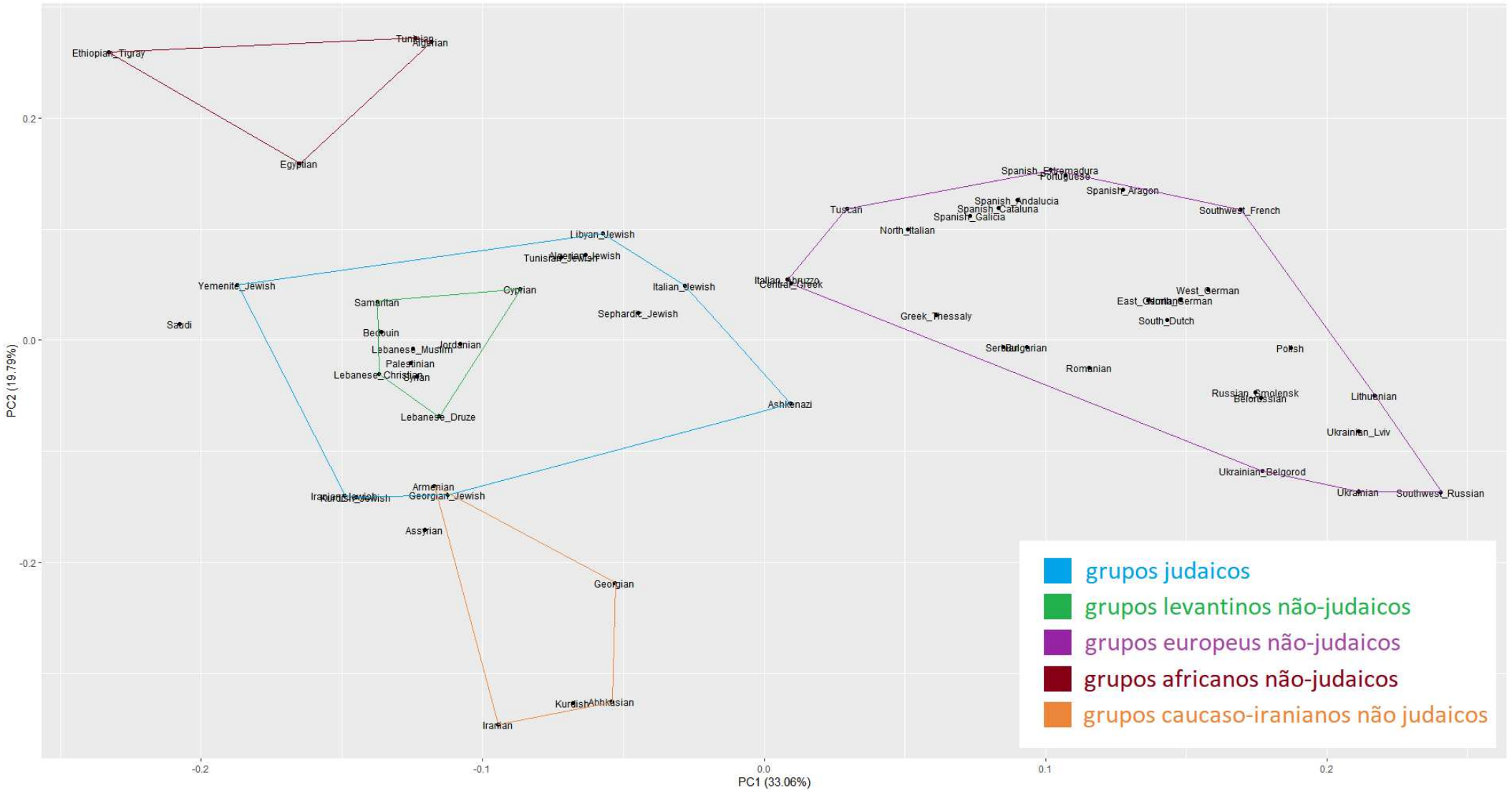


Imagem 1: PCA dos resultados da calculadora Eurogenes K13.

4.3 Análise visual dos gráficos gerados

Para a realização do PCA no dataset Eurogenes (imagem 1), utilizou-se os dois primeiros componentes principais: O Componente Principal 1 representou 33.06% e o Componente Principal 2 representou 19.79%. O método foi capaz de estabelecer divisões que condizem com a realidade geográfica das etnias presentes na calculadora. Note que linhas limítrofes foram adicionadas para auxílio visual.

O PCA separou todos os grupos europeus não-judaicos na metade direita do gráfico, e todos os grupos judaicos, levantinos, africanos e cáucaso-iranianos na metade esquerda. Vemos que, dentre as etnias africanas mostradas, os etíopes são os mais afastados dos levantinos, enquanto que os norte-africanos são mais próximos, especialmente os egípcios. Isso faz sentido, tendo em vista que a Etiópia é mais longe que o Egito em relação ao Levante. Ao focar na região dos grupos da Europa, vê-se que o PCA espelha a geografia do continente, como foi discutido por Johnson et al [27]. Os ibéricos (Portugueses e Espanhóis) estão na extremidade superior e os russos e ucranianos se encontram na outra extremidade. Alemães, franceses, poloneses e neerlandeses se encontram entre as extremidades, o que condiz com sua posição mais ao centro da Europa. Já os grupos balcânicos, gregos e italianos se encontram posicionados mais próximos das etnias do Oriente Médio, o que condiz com a posição que possuem mais ao Leste do Mar Mediterrâneo, compartilhando esta área com países como Turquia, Líbano e Egito.

É possível ver a propriedade de linearidade do PCA através do posicionamento das misturas judaicas na figura 4. Se criarmos uma nova amostra judaica através da média dos resultados das amostras Yemen_Jewish e Lybian_Jewish, veremos que essa nova amostra será equidistante destes elementos. O mesmo não ocorre nos outros métodos testados, onde as distâncias são exageradas em relação ao PCA e não seguem essa característica.

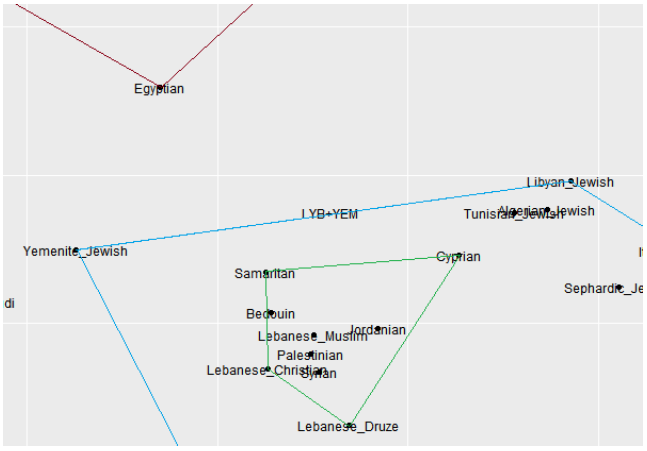


Figura 3: A nova amostra (LYB+YEM) feita da média dos dados de Yemenite_Jewish e Lybian_Jewish fica no meio do caminho entre as duas.

No PCA do dataset da calculadora MDLP (imagem 2 no apêndice), o Componente Principal 1 representou 23.54% e o Componente principal 2 representou 14.94% da variância. As informações tornaram-se demasiadamente confusas e várias etnias de regiões muito distintas estavam em regiões próximas, dificultando a visualização.

Podemos ver que no t-SNE, os resultados são bem diferentes. Enquanto o PCA mostra maior similaridade entre quaisquer dois pontos conforme forem mais próximos no gráfico e menor conforme estiverem afastados, o t-SNE se preocupa em criar ilhas populacionais de etnias semelhantes. Portanto, ele é excelente para criar clusters e grupos, mas pode não ser indicado para verificar a relação entre os grupos mais distantes, especialmente se o hiperparâmetro da Perplexidade for baixo. Rodamos o t-SNE no Eurogenes K13 com a perplexidade 19 e no MDLP 22 com perplexidade 38 (as perplexidades máximas permitidas para o número de amostras de cada dataset), e obtivemos o resultado exposto nas imagens 3 e 4, respectivamente (cf. apêndice). É válido notar que, caso haja a tentativa de reprodução posterior, ainda que utilizando o mesmo código, o gráfico será diferente, já que o t-SNE sempre gerará uma nova visualização ao ser executado novamente.

Nota-se no Eurogenes K13 que todas as etnias européias foram dispostas na metade direita do gráfico e as outras ficaram à esquerda, do mesmo modo que ocorreu no PCA. Entretanto, rodando outras vezes, as etnias européias ficaram em posições diferentes, por exemplo nos quadrantes superiores e as outras nos inferiores. O que se manteve constante é que enquanto as etnias européias estavam dispostas em uma seção, as outras estavam na outra metade.

É interessante lembrar que no t-SNE a mistura LYB+YEM não será equidistante das amostras que foram usadas para a compor. Neste gráfico gerado, a mistura se encontrou muito mais próxima de Lybian_Jewish.

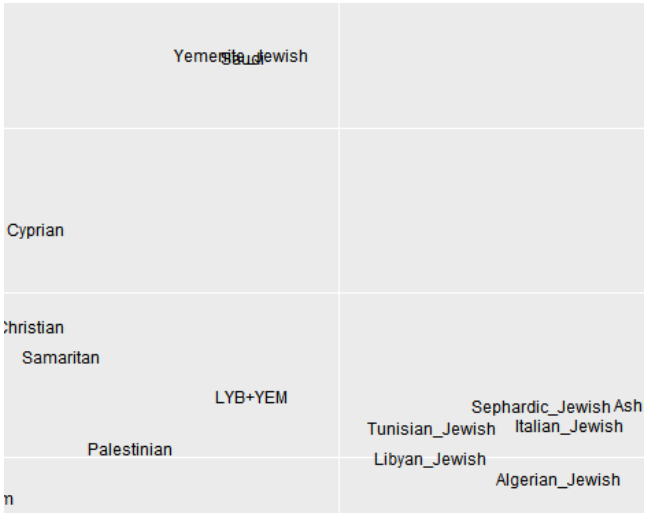


Figura 4: LYB+YEM não equidistante de Lybian_Jewish e Yemenite_Jewish no t-SNE

Assim como no PCA, vemos que os grupos judaicos se concentraram principalmente em duas regiões: houve um agrupamento mais ocidental, como os ashkenazim, italkim, sephardim e judeus da Tunísia, Argélia e Líbia, foram alocados entre os levantinos e os gregos e italianos. E um outro agrupamento mais oriental, dos judeus da Geórgia, judeus do Irã e os judeus curdos, que se posicionou entre os levantinos e suas populações locais não-judias. Os judeus do Iêmen foram alocados

separadamente, porém ainda próximos dos levantinos.

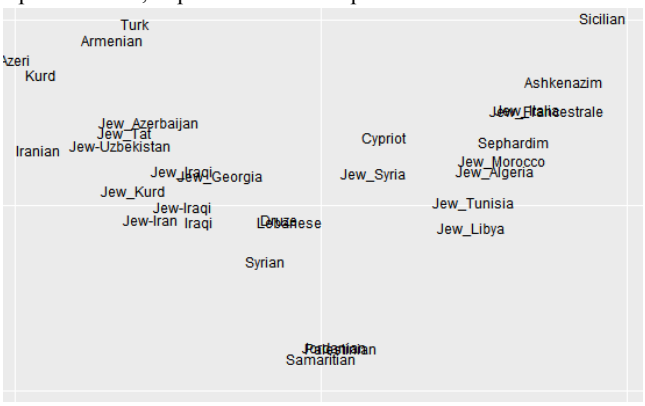


Figura 5: Dois principais agrupamentos judaicos que apareceram em várias visualizações produzidas pelo t-SNE

No t-SNE do MDLP, também vemos uma configuração semelhante na disposição das etnias judaicas, que ficaram concentradas em dois principais grupos, assim como no Eurogenes. Os judeus do Iêmen ficaram afastados dos dois clusters judaicos na maioria das vezes, mas ainda foram alocados relativamente próximos aos levantinos. Já os judeus da Índia e os judeus da Etiópia sempre foram posicionados muitíssimo afastados dos clusters judaicos, e ficaram mais próximos das etnias locais não-judaicas.

Ao diminuir o hiperparâmetro da perplexidade (imagens 5 e 6), notamos que o t-SNE gera gráficos com clusters mais definidos, porém as relações entre grupos menos próximos foi muito afetada. Por exemplo, ao exagerar nos clusters, podemos ver claramente a grande proximidade entre alemães e holandeses, entretanto não vemos mais a relação entre alemães e poloneses, como víamos no PCA. A perplexidade realiza o equilíbrio dos aspectos locais e globais da visualização.[28] Um gráfico gerado com perplexidade baixa pode ser interessante caso queiramos estudar clusters populacionais, e pode trazer esclarecimentos valiosos aliando-o com outros gráficos.

O UMAP também trouxe um gráfico interessante. Ele criou trilhas populacionais tanto no Eurogenes (imagem 7. cf. apêndice) quanto no MDLP (imagem 8, cf. apêndice). No Eurogenes, houve uma clusterização mais intensa dos leste-europeus, que ficaram consideravelmente separados dos outros grupos étnicos. Após isso, temos a "trilha" étnica, onde numa ponta vemos os europeus do centro e do oeste, passando pelos gregos e italianos, depois pelos grupos judaicos ocidentais, pelos levantinos e norte-africanos, chegando aos cluster judaico oriental e findando nos grupos cáucaso-iranianos. Ele formou uma linha que partiu da Europa, passou pelo Norte da África e Oriente Médio e encerrou no Cáucaso. Já no MDLP, duas trilhas se formaram: a primeira partindo do leste europeu, indo aos balcãs e ao centro europeu, e findando no norte da Itália e na península ibérica. A segunda trilha inicia-se com os italianos do sul e gregos, passando pelos grupos judaicos ocidentais, levantinos, judaicos oriental, região do Cáucaso e Irã, passa pelo centro da ásia e encerra na índia. A grosso modo, formou-se uma trilha majoritariamente europeia e a outra, majoritariamente asiática.

O UMAP se assemelha ao t-SNE por não ser linear, logo, uma mistura entre duas etnias não será posicionada exatamente ao meio delas. E se assemelha ao PCA por sempre fornecer o mesmo gráfico ao executar.

4.4 Microbenchmark

Aplicamos o pacote Microbenchmark para avaliar os três métodos. Segundo a documentação do R, O Microbenchmark é um pacote de funções de temporização precisas, e "fornece infraestrutura para medir e comparar com precisão o tempo de execução de expressões R" [29][30].

O pacote Microbenchmark é uma biblioteca que permite medir e comparar o desempenho de pequenas porções de código em R. Ele é amplamente utilizado para analisar o tempo de execução de diferentes operações ou funções em um ambiente controlado. A sua principal função é fornecer uma maneira fácil e precisa de medir o tempo que uma determinada operação ou função leva para ser executada. Isso é útil para identificar partes do código que podem estar consumindo muito tempo de processamento e necessitem de otimização. Ao usar o Microbenchmark , escrevemos o conjunto de expressões ou funções que desejamos medir e executá-las várias vezes para obter uma estimativa confiável do tempo de execução. O pacote lida com a precisão das medições, descartando possíveis flutuações causadas por interrupções do sistema operacional ou outros fatores externos. Além disso, permite comparar diferentes versões de um código para determinar qual implementação é mais eficiente em termos de tempo de execução. Isso é útil quando você está otimizando seu código e deseja verificar se as mudanças resultaram em melhorias reais no desempenho. O pacote é uma ferramenta valiosa para analisar o desempenho de código em R, ajudando os desenvolvedores a identificar gargalos de desempenho e aprimorar a eficiência de seus programas.

Iniciamos chamando os pacotes necessários para a análise. Depois, chamamos o microbenchmark, passando como parâmetro os diversos códigos aos quais desejamos comparar - quatro situações foram testadas em cada um dos datasets: PCA, t-SNE com perplexidade alta (34 no MDLP, 18 no Eurogenes), t-SNE com perplexidade baixa (3), UMAP. Para ver a lista de valores coletados, como (média mediana, menor valor, maior valor, etc), basta chamar o objeto mbm. Para ver o gráfico de violino, basta realizar o autoplot.

Após isso, obtemos o resultado da imagem 9. O PCA verificou-se o mais rápido dos métodos, com média de 146.2211 ms no PCA e 149.9398 no MDLP. Não houve diferença considerável nos tempos do PCA dos dois datasets apesar da diferença de tamanho. Já nos outros métodos é possível notar uma discrepância entre os resultados do MDLP e Eurogenes. Por exemplo, o UMAP no MDLP obteve média de 820.1477 ms, enquanto no Eurogenes a média foi de 541.0432 segundos - 279 ms mais rápido no Eurogenes.

No t-SNE também houve discrepância entre o tempo para aplicar o método entre os datasets MDLP e Eurogenes, mas não houve diferença significativa no t-SNE com perplexidade baixa ou alta em um mesmo dataset: 462.8076 e 458.4846 no MDLP e 260.8859 e 268.1207 no Eurogenes (4ms a mais e 8 ms a menos ao aumentar a perplexidade, respectivamente). Embora tais diferenças de tempo possam parecer insignificantes e desprezíveis à primeira vista, em trabalhos que utilizem datasets maiores, essas discrepâncias podem se tornar expressivas e significar minutos ou horas.

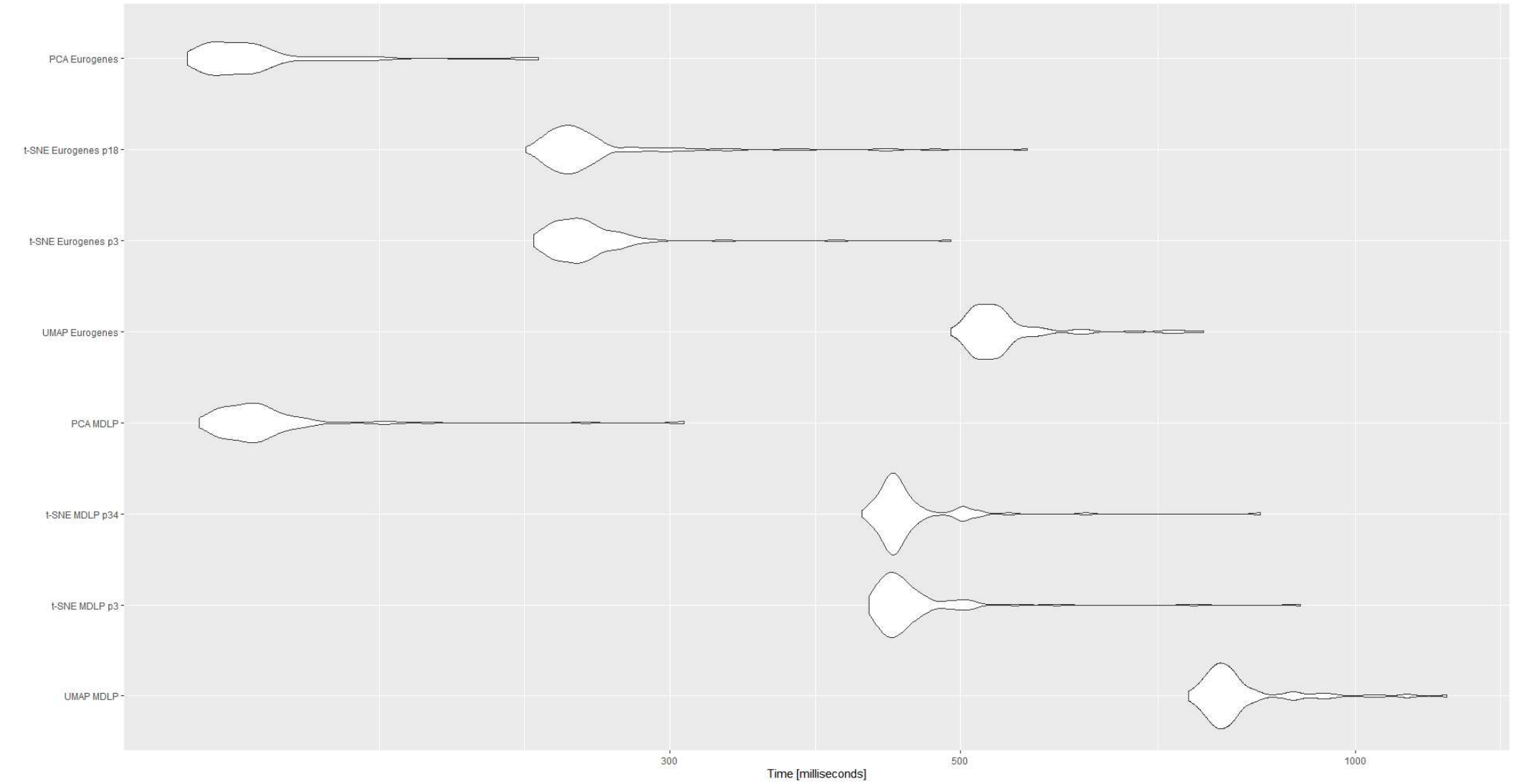


Imagem 9: Gráfico de violino dos resultados do microbenchmark

5. CONCLUSÃO

Aplicamos os três métodos de redução de dimensionalidade em dois datasets pré-processados. Após isso, utilizamos um pacote de análise de desempenho para averiguar como se comportavam os métodos. O PCA é o método mais rápido, preserva até mesmo as distâncias entre etnias menos próximas, e apresentou uma visualização satisfatória no primeiro dataset. Entretanto, no segundo dataset (que possui mais amostras), a visualização tornou-se confusa e não separou as etnias a nível continental. Aplicamos duas vezes o t-SNE - a primeira com perplexidade alta, e a segunda com perplexidade três. Conforme baixamos a perplexidade, percebemos que o método focava em aspectos locais e não nos globais, formando ilhas populacionais, sendo interessante para estudar proximidade alta entre grupos. Por fim, utilizamos UMAP default, que criou visualizações com aspecto de trilhas, respeitando as separações a nível continental.

Todos os métodos forneceram visualizações interessantes para os dados trabalhados. No entanto, ao comparar os métodos de redução de dimensionalidade aplicados, é possível inferir que, em datasets menores, o PCA realiza uma divisão equilibrada e corresponde à realidade geográfica e cultural dos povos estudados. Todavia, sua confiabilidade e intuitividade diminuem drasticamente quando trabalhamos com dados de dimensões maiores ou datasets com um número maior de amostras. No dataset Eurogenes, a visualização gerada pelo PCA foi de fácil entendimento e replicou a disposição geográfica dos povos, entretanto, no caso da visualização gerada com os dados da calculadora MDLP, o PCA aproximou etnias de origens continentais distintas e criou uma visualização confusa.

O t-SNE e o UMAP são mais indicados para esses casos, sendo que, quando comparados ao PCA, estes podem exigir um alto custo computacional se os dados forem muito grandes. Além disso, se a perplexidade escolhida for muito baixa, o método do t-SNE realizará muitas clusterizações e a visualização pode ser afetada no contexto geral. Como o PCA é linear, ele é adequado para a análise do cenário global. No entanto, quando se trata de representar relacionamentos complexos em datasets maiores, o t-SNE e o UMAP tendem a ser mais eficazes. O t-SNE formou clusters menores e mais numerosos que o UMAP.

A utilização em conjunto dos três métodos na análise poderá enriquecer as pesquisas que necessitem de redução de dimensionalidade, e espera-se que o presente trabalho contribua para uma compreensão mais abrangente acerca da visualização de dados genéticos e sirva de auxílio na seleção do(s) método(s) mais apropriado(s) para o desenvolvimento de outras pesquisas, e que possa servir de base para trabalhos futuros sobre o assunto - que possam inclusive incluir outros métodos (como LDA e MDS), ou outros datasets (por exemplo, haplogrupos Y ou mtDNA, ao invés do DNA autossômico).

Tabela 3: comparativo entre os três métodos

	PCA	**t-SNE**	**UMAP**
Linearidade	linear	não-linear	não-linear
Estabilidade nas visualizações geradas	determinístico (sempre gera mesma visualização)	não-determinístico/estocástico (mesma entrada gera resultados distintos)	não-determinístico/estocástico (mesma entrada gera resultados distintos)
Comportamento em dataset de muitas amostras	Confuso em muitas amostras, muito bom com poucas	Comportou-se bem com muitas amostras	Comportou-se bem com muitas amostras
Customizabilidade	Baixa	Alta (theta, perplexidade, etc)	Alta (method)
Tempo de execução no MDLP	menor que 200 milissegundos	Maior que 200 milissegundos	Maior que 200 milissegundos
Forma tradicional	Disperso	Disperso se perplexidade for alta; forma de ilhas populacionais quando perplexidade é baixa	Tendeu a criação de trilhas

6. REFERENCIAS

[1] BELLMAN, R. Dynamic Programming. Rand Corporation research study, Princeton University Press, 1957.

[2] BuiltIn. What is the curse of dimensionality? https://builtin.com/data-science/curse-dimensionality

[3] EUSÉBIO DE CESARÉIA, História Eclesiástica IV, capítulo 6. Traduzido por Wolfgang FIscher, Novo Século, 2002.

[4] ELIZABETH SPELLER. Following Hadrian: A Second-Century Journey Through the Roman Empire, at Google Books, Oxford University Press, 2004, p. 218

[5] The Jewish Encyclopedia: A Descriptive Record of the History, Religion, Literature, and Customs of the Jewish People from the Earliest Times to the Present Day. Funk & Wagnalls company, 1906.

[6] GOODMAN, M. SCHWARTZ, S. The Oxford Handbook of Jewish Studies. Oxford University Press, 2005. p80

[7] The Jewish Agency. Jewish Population Rises to 15.2 million Worldwide. https://www.jewishagency.org/jewish-population-5782/#:~:text=The%20percentage%20of%20Jews%20living,by%20the%20Pew%20Research%20Center

[8] Revista Fleury, Edição. 33. Como funciona o exame de DNA. https://www.fleury.com.br/noticias/como-funciona-o-exame-de-dna-revista-fleury-ed-33

[9] Veja Ciência. Testes genéticos que mapeiam origem dos ancestrais estão em alta no Brasil. https://veja.abril.com.br/ciencia/testes-geneticos-que-mapeiam-origem-dos-ancestrais-estao-em-alta-no-brasil

[10] Saúde Digital News. Genera bate meta de receita e quer testar mais 10 milhões de pessoas até 2028. https://saudedigitalnews.com.br/28/04/2022/genera-bate-meta-de-receita-e-quer-testar-mais-10-milhoes-de-pessoas-ate-2028/

[11] Genera. A crescente demanda por testes de ancestralidade no Brasil. https://www.genera.com.br/blog/a-crescente-demanda-por-testes-de-ancestralidade-no-brasil/

[12] SNPedia. Single Nucleotide Polymorphism. https://www.snpedia.com/index.php/Single_Nucleotide_Polymorphism

[13] Behold Genealogy. Comparing Raw Data from 5 DNA Testing Companies. https://www.beholdgenealogy.com/blog/?p=2700

[14] tellmeGen. O que é uma variante genética comum ou SNP? https://www.tellmegen.com/pt/?_ga=2.212280855.2020948951.1686773886-1146115295.1686593884

[15] Eurogenes Blog: Focusing on ancient population genomics. https://eurogenes.blogspot.com/

[16] BRACE, S. et al. Genomes from a medieval mass burial show Ashkenazi-associated hereditary diseases pre-date the 12th century. Current Biology Volume 32, Issue 20.

[17] European Genome-phenome Archive: https://ega-archive.org/

[18] gnomAD: Genome Aggregation Database. https://gnomad.broadinstitute.org/

[19] Harvard: Allen Ancient DNA Resource (AADR): https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data

[20] Byun, J., Han, Y., Gorlov, I. P., Busam, J. A., Seldin, M. F., & Amos, C. I. (2017). Ancestry inference using principal component analysis and spatial analysis: a distance-based analysis to account for population substructure. BMC genomics, 18(1), 789.

[21] BuiltIn. A Step-by-Step Explanation of Principal Component Analysis (PCA) https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[22] van der Maaten, L.J.P. t-Distributed Stochastic Neighbor Embedding. https://lvdmaaten.github.io/tsne/

[23] L.J.P. van der Maaten and G.E. Hinton. Visualizing Non-Metric Similarities in Multiple Maps. Machine Learning 87(1):33-55, 2012.

[24] L.J.P. van der Maaten. Learning a Parametric Embedding by Preserving Local Structure. In Proceedings of the Twelfth International Conference on Artificial Intelligence & Statistics (AI-STATS), JMLR W&CP 5:384-391, 2009.

[25] L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research 9(Nov):2579-2605, 2008.

[26] Dimensionality Reduction for Data Visualization: PCA vs TSNE vs UMAP vs LDA. https://towardsdatascience.com/dimensionality-reduction-for-data-visualization-pca-vs-tsne-vs-umap-be4aa7b1cb29

[27] Novembre, J., Johnson, T., Bryc, K. et al. Genes mirror geography within Europe. Nature 456, 98–101 (2008).

[28] Distill. How to Use t-SNE Effectively. https://distill.pub/2016/misread-tsne/#:~:text=A%20second%20feature%20of%20t,effect%20on%20the%20resulting%20pictures

[29] RDocumentation. Microbenchmark 1.4.10. https://www.rdocumentation.org/packages/microbenchmark/versions/1.4.10/topics/microbenchmark

[30] CRAN. Package microbenchmark. https://CRAN.R-project.org/package=microbenchmark

Apêndice

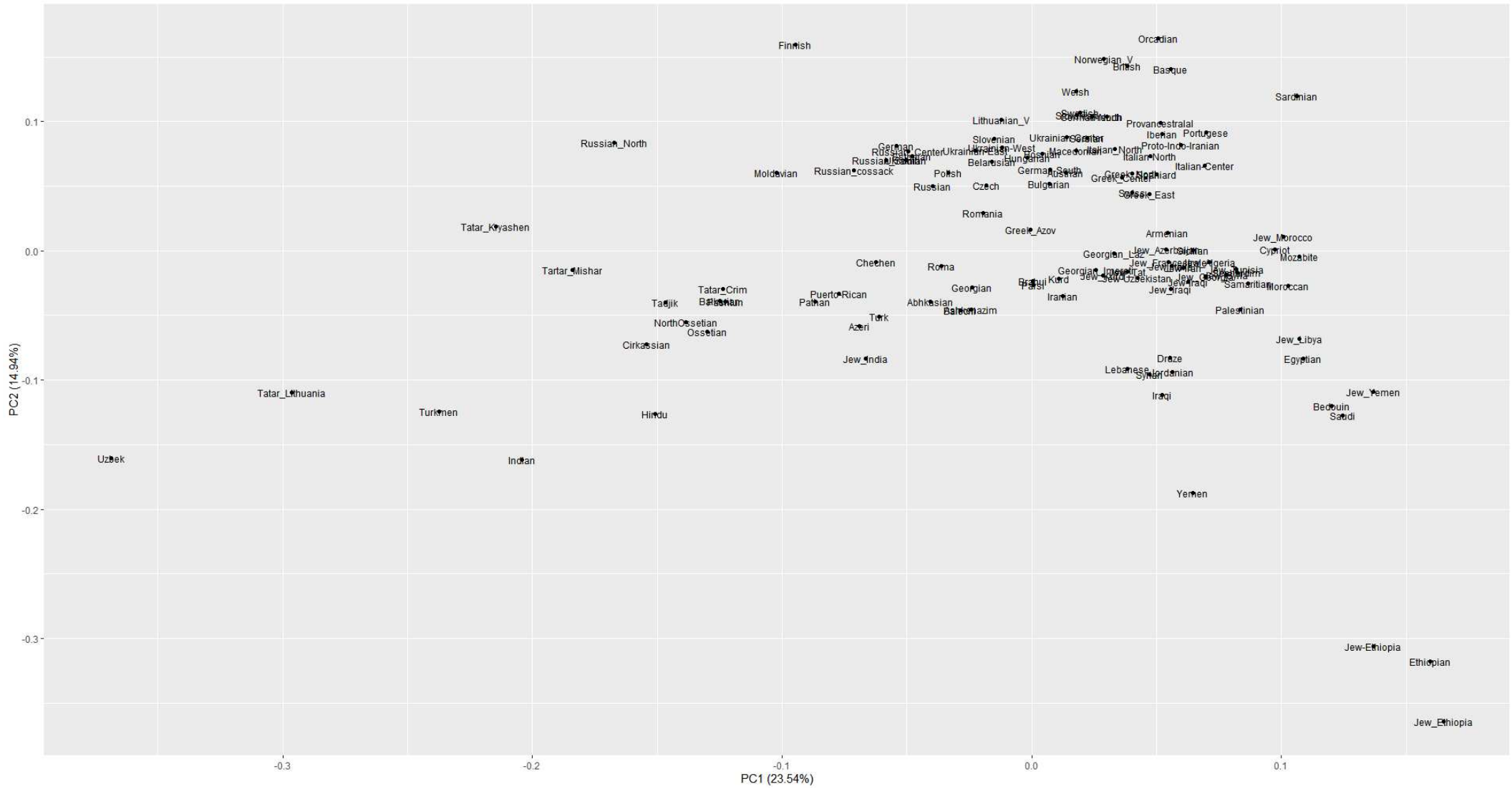


Imagem 2: PCA dos resultados da calculadora MDLP 22

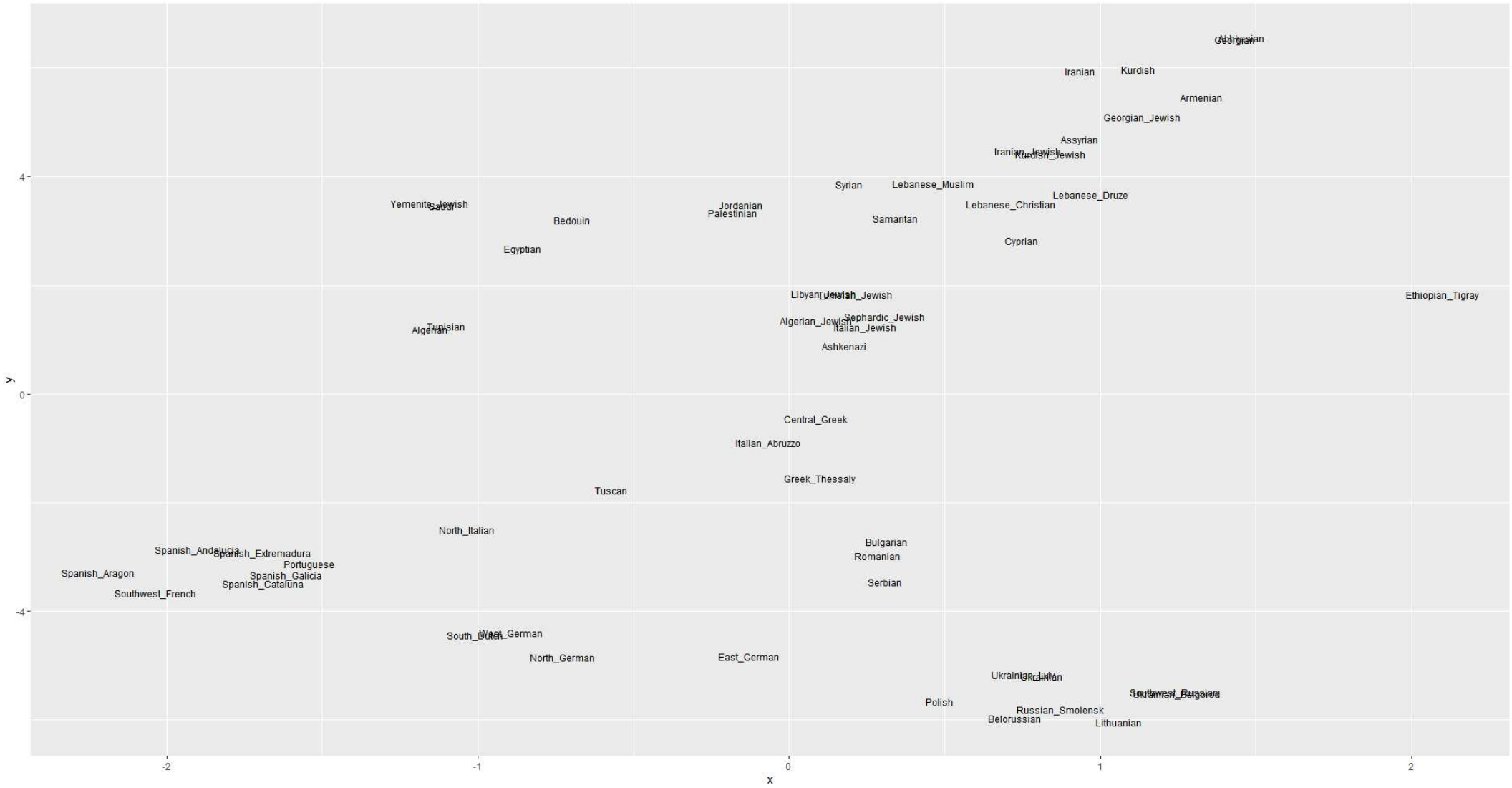


Imagem 3: t-SNE dos resultados da calculadora Eurogenes com perplexidade alta

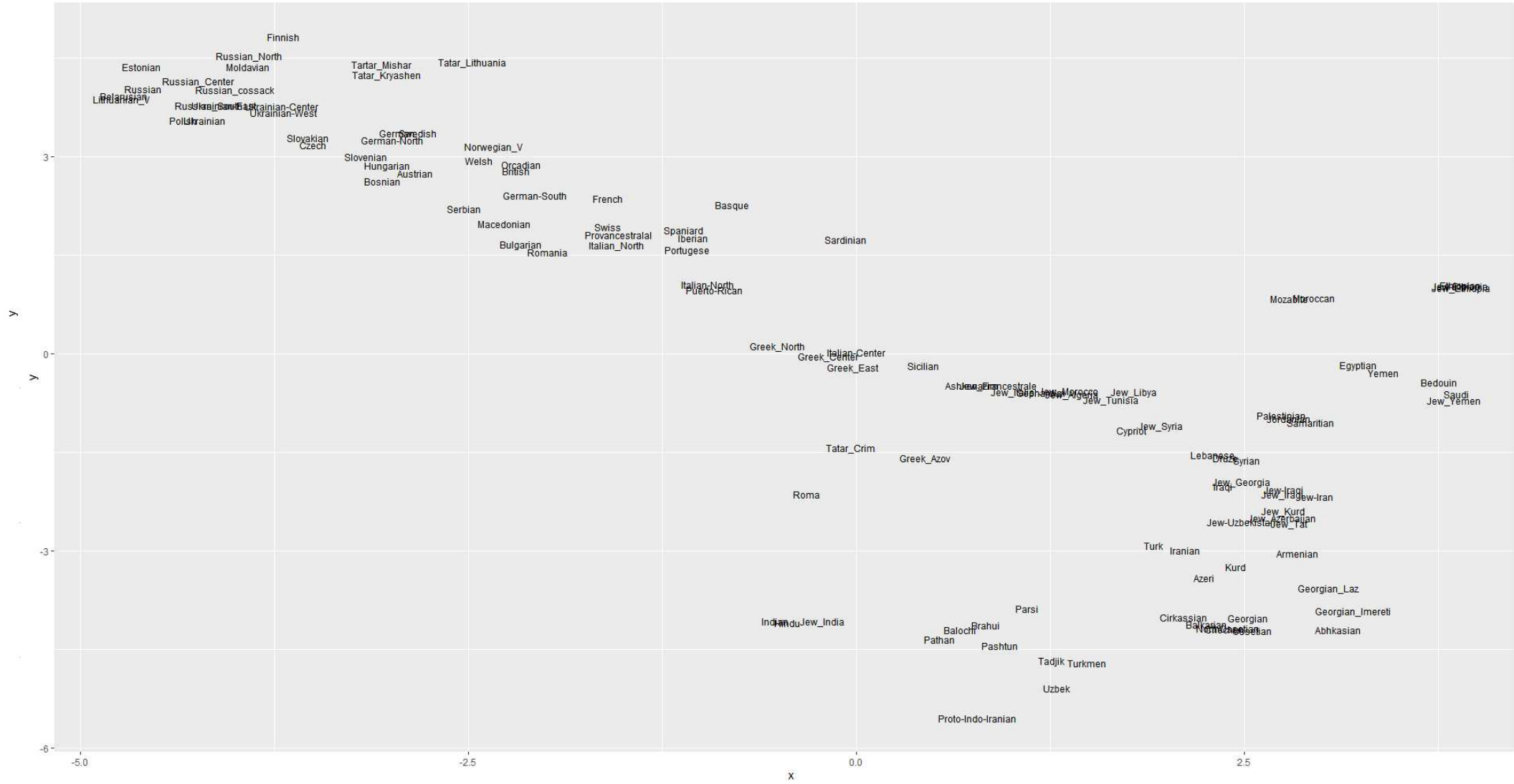


Imagem 4: t-SNE dos resultados da calculadora MDLP 22 com perplexidade alta

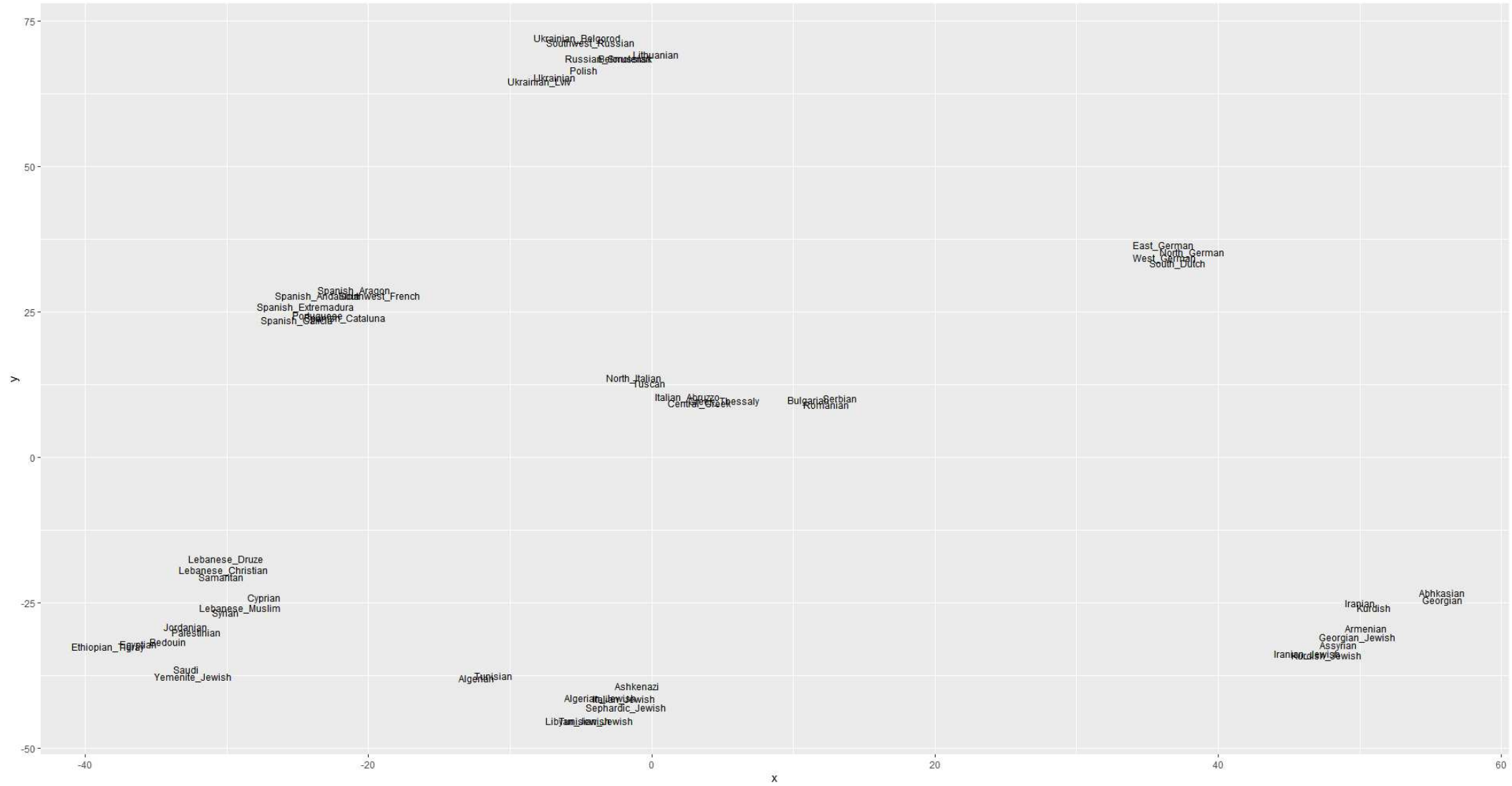


Imagem 5: t-SNE dos resultados da calculadora Eurogenes com perplexidade baixa. Relações distantes não são preservadas, mas relações próximas são evidenciadas.

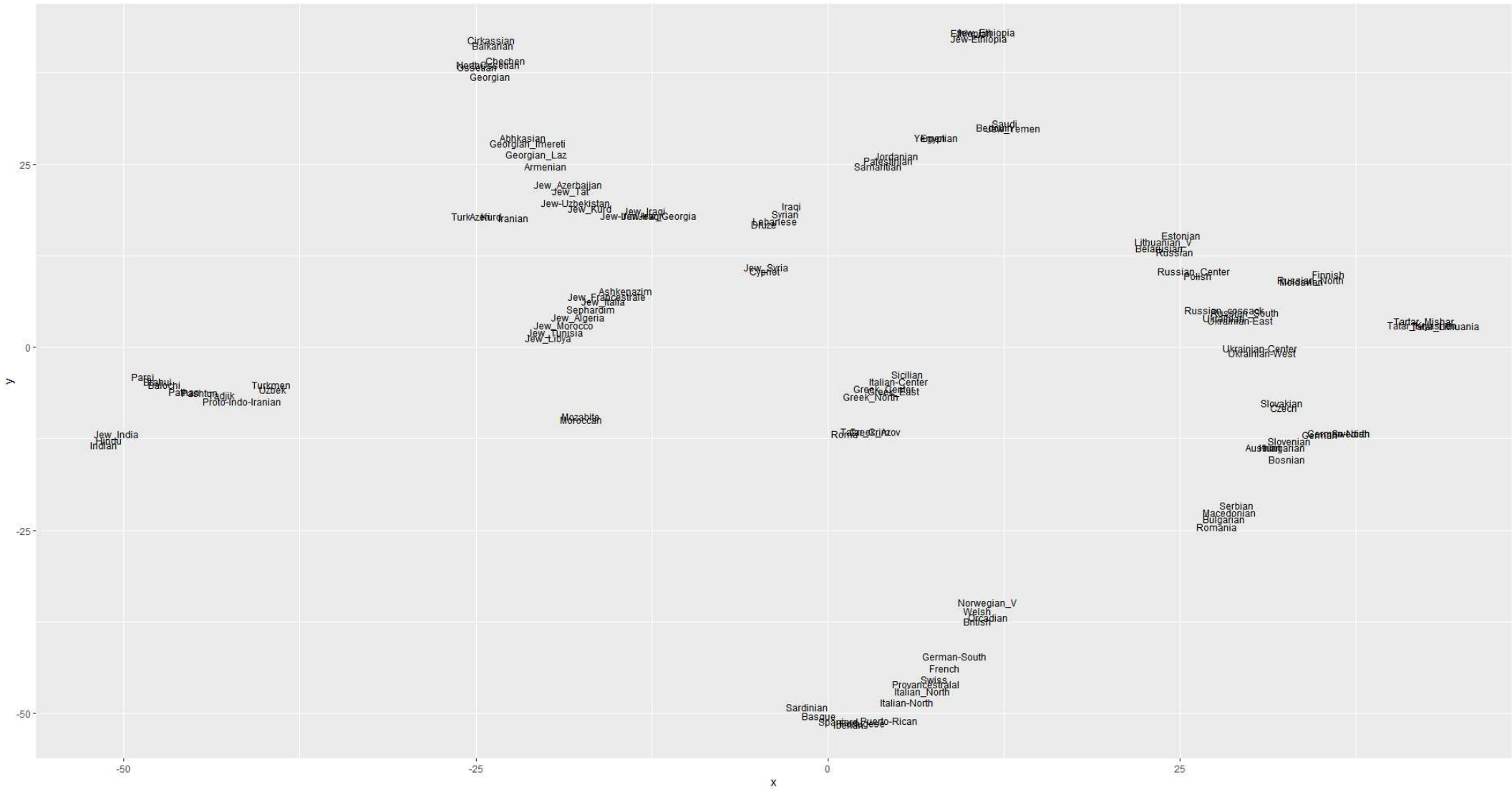


Imagem 6: t-SNE dos resultados da calculadora MDLP 22 com perplexidade baixa.

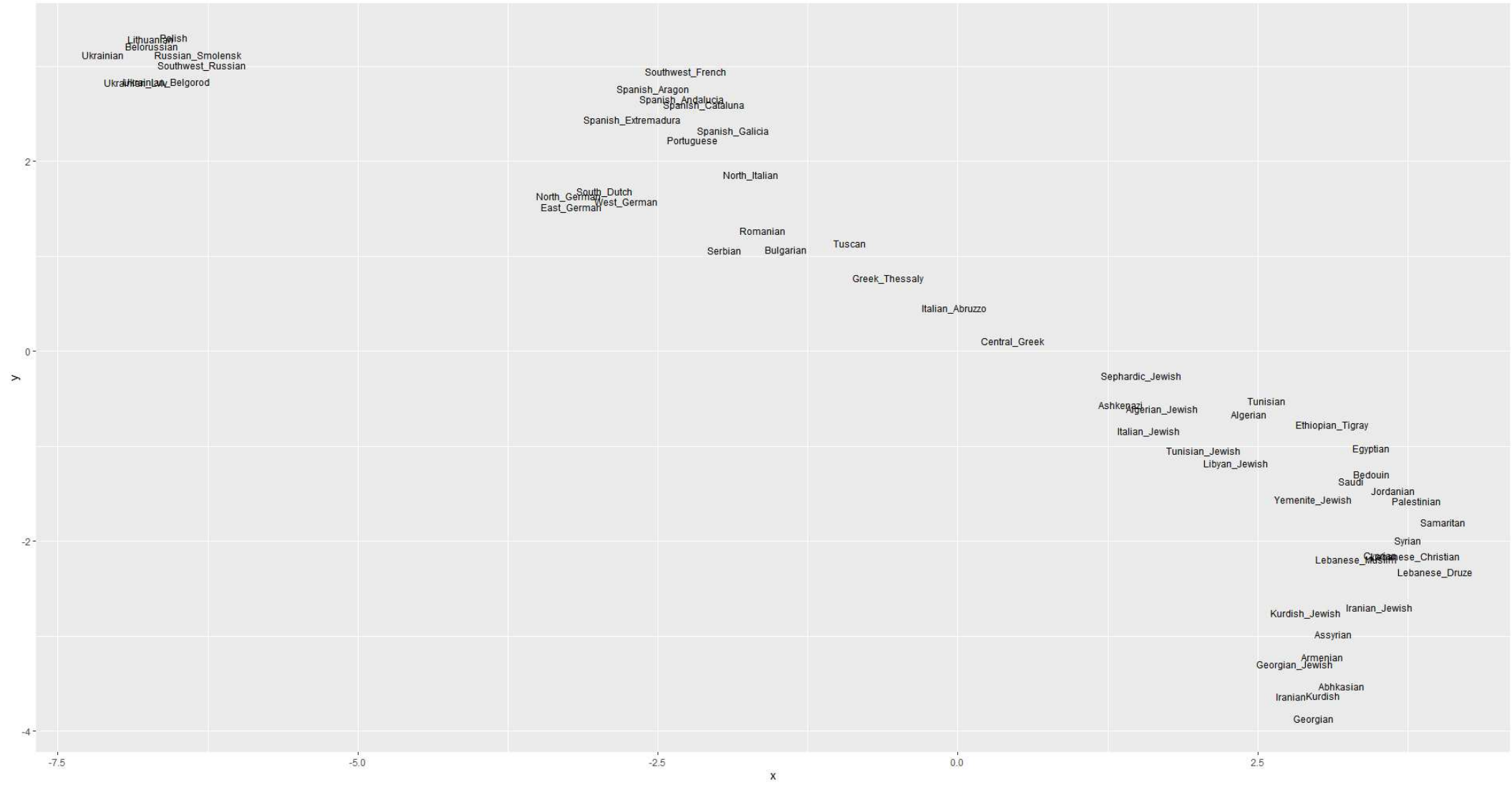


Imagem 7: UMAP dos resultados da calculadora Eurogenes K13

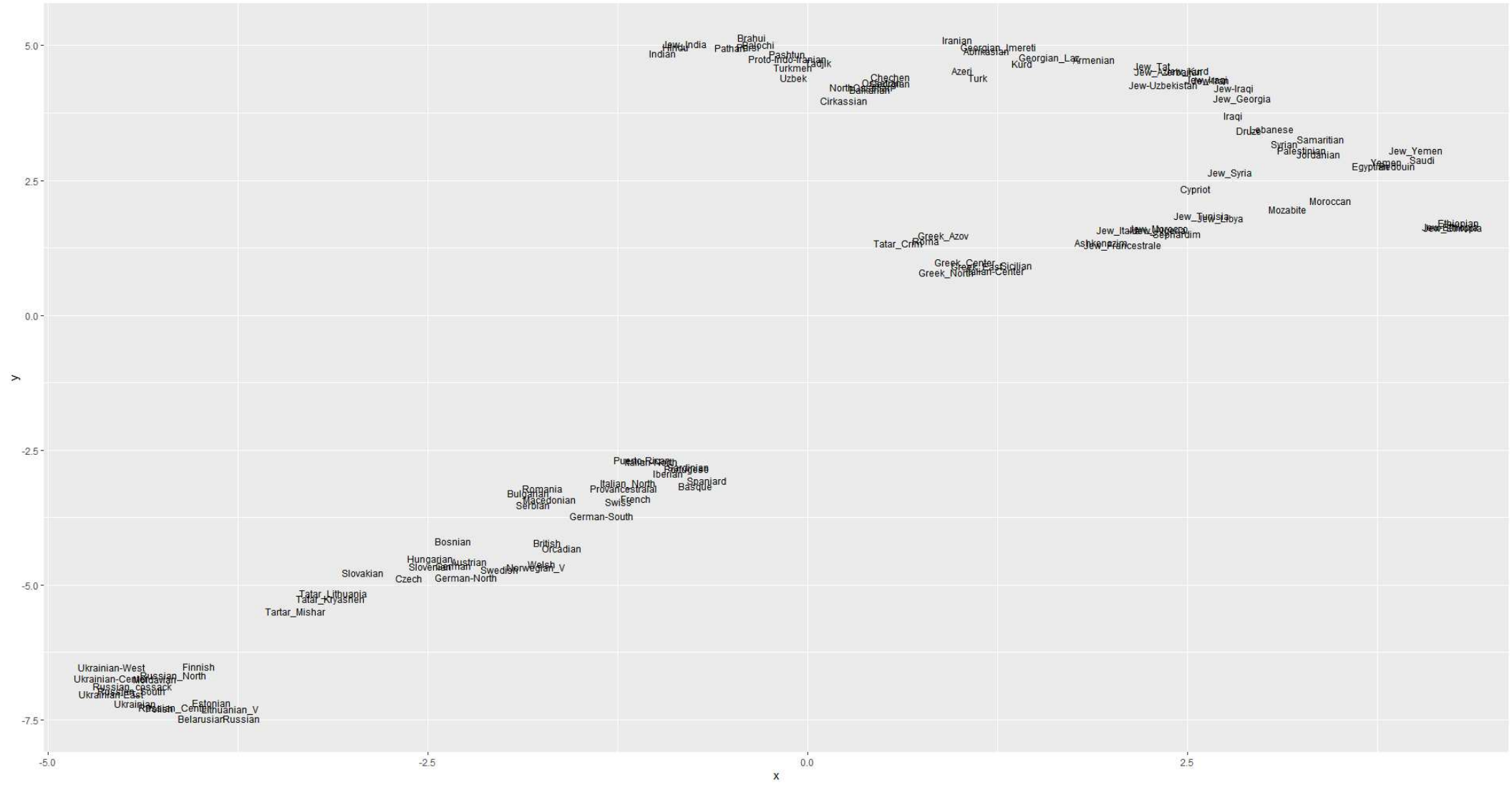


Imagem 8: UMAP dos resultados da calculadora MDLP 22