Relatório: Fração de linfócitos - Métodos para estimar a presença de células B e T

Jean Resende

2023-06-21

Fração de linfócitos

Contextualização

A progressão do tumor e o sucesso das terapias anticancerígenas são influenciadas pela composição e a densidade das células imunes no microambiente tumoral. As técnicas recomendadas com a finalidade de estudar tais células são: citometria de fluxo, coloração imuno-histoquímica ou sequenciamento de célula única. Geralmente, bases de dados de sequenciamento de RNA (RNA-seq). Sendo asssim, torna-se necessário a aplicação de métodos computacionais para estimar a composição de células imunes a partir dos dados RNA-Seq.

Há vários métodos computacionais propostos recentemente que prometem estimar a fração de células imunes em dados de sequenciamento do tipo RNA-Seq. Porém, diferentes métodos utilizam diferentes cálculos, trilham caminhos diferentes e consequentemente geram resultados que contradizem ou reforçam os resultados de outro método. Alguns métodos de quantificação de células imunes olham para a matriz de expressão gênica, e a partir de genes marcadores e/ou deconvulação estimam a composição celular. Outros métodos como MiXCR e TRUST4 (dentre outros) olham para o dado bruto do sequenciamento, ou seja, não olham a matriz de expressão gênica, mas sim as leituras que se alinham a região do transcriptoma referente ao receptor de células B/T por exemplo.

Nosso trabalho é direcionado às células B e T a partir de dados brutos de RNA-Seq. Isso faz com que utilizemos um método que acesse as leituras brutas do sequenciamento e então estime a composição dessas células. (1) Mas será que os métodos que utilizam genes marcadores ou os que são baseados em deconvolução estimam a presença de células B e T assim como os métodos TRUST4/MiXCR que olham diretamente para o dado bruto? (2) Quais métodos podemos eutilizar em nosso trabalho, reforçando a presença de células B e T?

Metodologia

Fiz uma revisão de literatura sobre os principais métodos de quantificação de linfócitos aplicados na imuno-oncologia. O artigo Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology foi o trabalho que me debrucei mais, pois os autores avaliaram os principais métodos de quantificação de tipos celulares da imuno-oncologia. Os métodos que eles avaliaram foram: CIBERSORT, EPIC, MCP-counter, quanTIseq, TIMER e xCell. Os métodos que eles recomendaram foram: EPIC, MCP-counter, xCell e quanTIseq (https://academic.oup.com/view-large/137497314).

EPIC e **quanTIseq** utilizam regressão de mínimos quadrados restrito para inferir frações a partir de uma matriz de assinatura e da expressão gênica em massa. Já o método **CIBERSORT** utiliza *v-Support Vector Regression* também sob uma matriz de assinatura. Já o método **MCP-counter** utiliza a expressão de genes marcadores em amostras heterogêneas, quantificando cada tipo de célula de forma independente. O **xCell**

também utiliza expressão de genes marcadores, no entanto aplica um teste estatístico de enriquecimento. Tanto o quanTIseq quanto o EPIC geram pontuações relativas à quantidade total de células sequenciadas.

A extração dos TCR e BCR por meio do TRUST4 como já vimos antes foi bem sucedida. Identificamos TCR e BCR nas amostras e começamos a visualizar um padrão relacionando a contagens desses receptores e o perfil esteroidal. Sendo assim, a primeira análise que fiz, foi utilizar os métodos indicados por Sturm et al. (2019) nos dados públicos de ACC (ou seja, a matriz de expressão gênica disponível no TCGA). Porém, o foco do estudo é direcionado para as células B e T, então tentei retirar dos métodos as células que eram diferentes de B e T, assim como fizemos com o CIBERSORT. Tais métodos utilizam métricas diferentes, ou seja, não posso aplicar a mesma lógica que apliquei no CIBERSORT.

Resultados

A Figura 1 mostra o resultado desta primeira análise. Porém, não consegui retirar as células diferentes de B e T para os métodos EPIC e quanTIseq, pois são métodos que geram uma fração (variando de 0 a 1) então teria que alterar na matriz de referência, mas esses dois métodos utilizam outros objetos de referência além da matriz. No entanto, consegui remover as células diferentes de B e T para os métodos MCP-counter e xCell, pois esses métodos não geram frações, mas sim pontuações e por causa disso, posso considerar apenas as células de interesse.

Figure 1: Fração e pontuação de linfócitos gerados pelos principais métodos sugeridos por Sturm et al., (2019). As colunas representam as amostras enquanto que as linhas representam anotações das amotras (primeiro bloco) e fração de linfócitos para os métodos EPIC e quanTIseq e pontuação de linfócitos para MCP-counter e xCell (segundo bloco).

O xCell mostrou um perfil interessante, mais parecido com as contagens que encontramos com o TRUST4 (Figura 2). Apesar disso, o xCell indicou uma presença de células T e B que o TRUST4 não encontrou. O xCell separa algumas soubcategorias dessas células, isso o TRUST4 não faz. O TRUST4 encontrou mais células T em low steroid, e o xCell encontrou mais células CD8 Tcm, CD8 T e CD4 naive T. Mas o xCell apontou para uma presença maior de células CD4 Tcm para high steroid e uma presença parecida entre low steroid e high steroid (não fiz teste estatístico, são hipóteses com base na visualização).

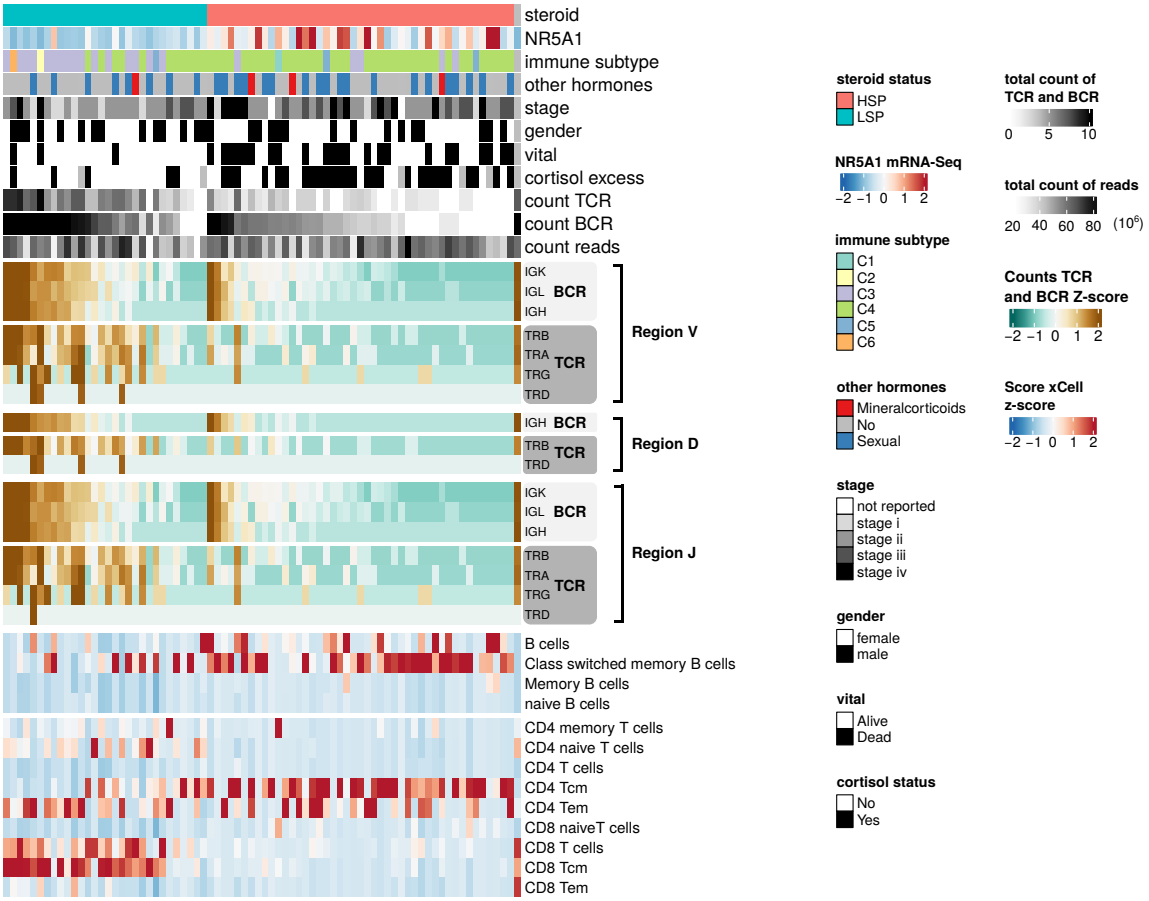

Figure 2: Contagens de TCR e BCR com o TRUST4 e pontuação de linfócitos com xCell. As colunas representam as amostras enquanto que as linhas representam anotações das amotras (primeiro bloco) e contagem de TCR e BCR com TRUST4 e pontuação de linfócitos com xCell (segundo bloco).

O xCell foi o método que mais se aproximou da extração de TCR e BCR com o TRUST4, comparado com os outros métodos (EPIC, quanTIseq e MCP-couter). Prefiro utilizá-lo como o método de quantificação de tipo celular baseado na matriz de expressão gênica em conjunto com o TRUST4 que olha para os dados brutos. Outro ponto que reforça esta escolha é que muitos trabalhos tem utilizado o xCell, inclusive o trabalho que o João publicou em 2021.

Expressão de TCR e BCR

Contextualização

Visto que o número de leituras de TCR e BCR pode depender do total de leituras do sequenciamento, defini a expressão dos receptores sendo: *TCR|BCR = Mi/(Ni + Mi)* em que i corresponde a cada amostra, M é o número de leituras que mapeiam para um BCR/TCR específico e N é o snúmero de leituras que mapeiam para qualquer outra coisa no genoma. Esse cálculo é usado para esta finalidade - expressão de BCR e TCR específicos (Pineda et al., 2021, Yu et al., 2022, Selitsky et al., 2019).

Olhar somente as contagens de TCR e BCR sem considerar o tamanho da biblioteca (quantidade de leituras sequenciadas pelo RNA-Seq) pode ser algo perigoso, pois essas amostras não tem uma quantidade padronizada de tamanho de biblioteca. Sendo assim, uma amostra pode ter mais/menos contagens de TCR/BCR quando comparada a outra, e essa diferença ser devido a quantidade de leituras a mais/menos que esta amostra tem em relação a outra.

Metodologia

Apliquei a fórmula *TCR|BCR = Mi/(Ni + Mi)* sob as contagens de TCR e BCR resultantes do TRUST4. Para uma amostra i, inseri como valor Mi a contagem de um TCR/BCR. Para o valor de Ni inseri a quantidade de leituras da amostra i resultante do controle de qualidade feita pelo FastQC (um dos resultados desta ferramenta foi a quantidade de leituras sequenciadas).

Resultados

A Figura 3 mostra a expressão dos TCR e BCR com base no cálculo apresentado acima. É possível observar um perfil visual parecido com a contagem bruta (Figura 2). No entanto, algumas amostras, assim como algumas cadeias de receptores evidenciaram mais a expressão dos TCRs e BCRs.

A diferença entre as contagens e a expressão de TCR e BCR para LSP e HSP foi significativa de acordo com o teste de Wilcoxon (Figura 4). A Firua 4 mostra a comparação das contagens e normalização sob o perfil esteroidal. Em ambas comparações (LSP_contagens X HSP_ contagens (non-normalized); LSP_expression X HSP_expression (normalized)) o LSP apresentou um valor maior.

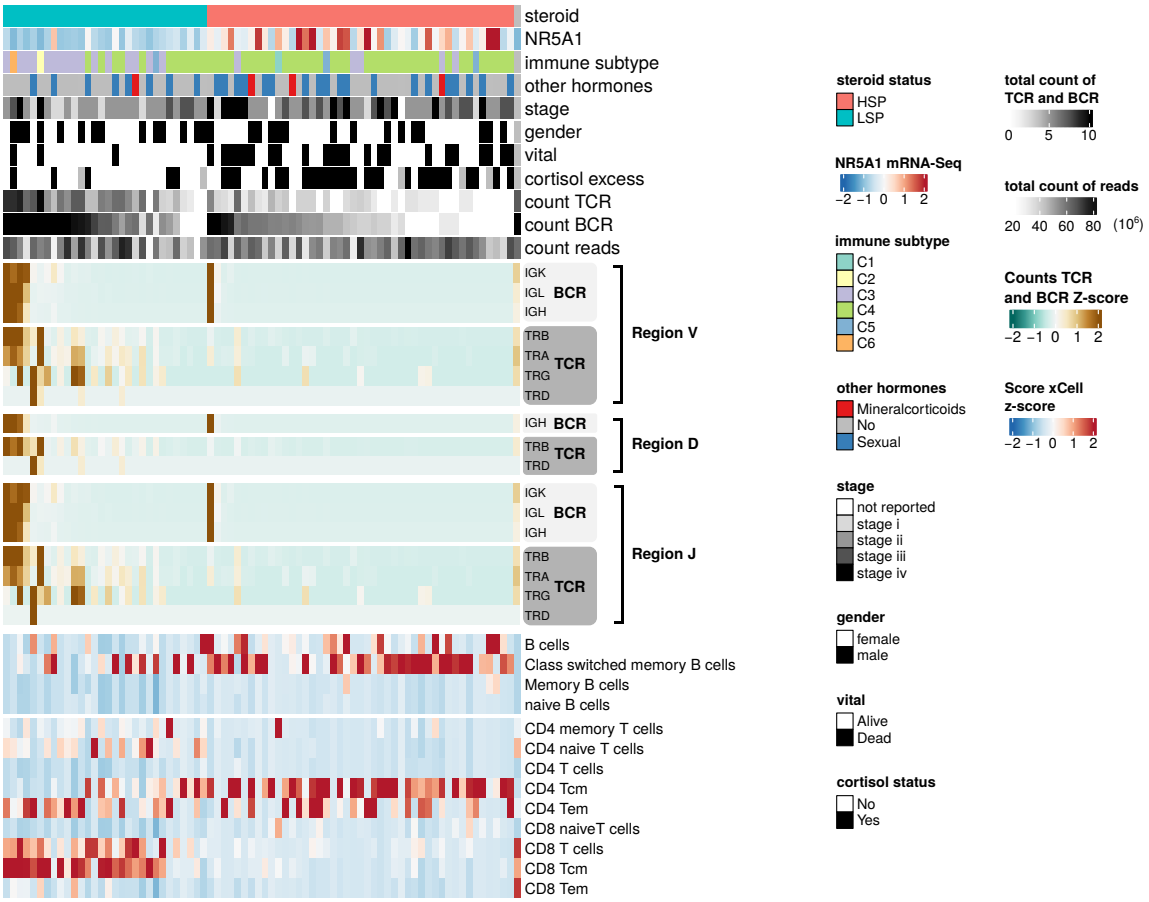

Figure 3: Expressão de TCR e BCR. As colunas representam as amostras enquanto que as linhas representam anotações das amotras (primeiro bloco) e expressão de TCR e BCR sob a fórmula TCR|BCR = Mi/(Ni + Mi), em que i representa uma amostra, M a contagem de um TCR/BCR e N a quantidade de leituras.

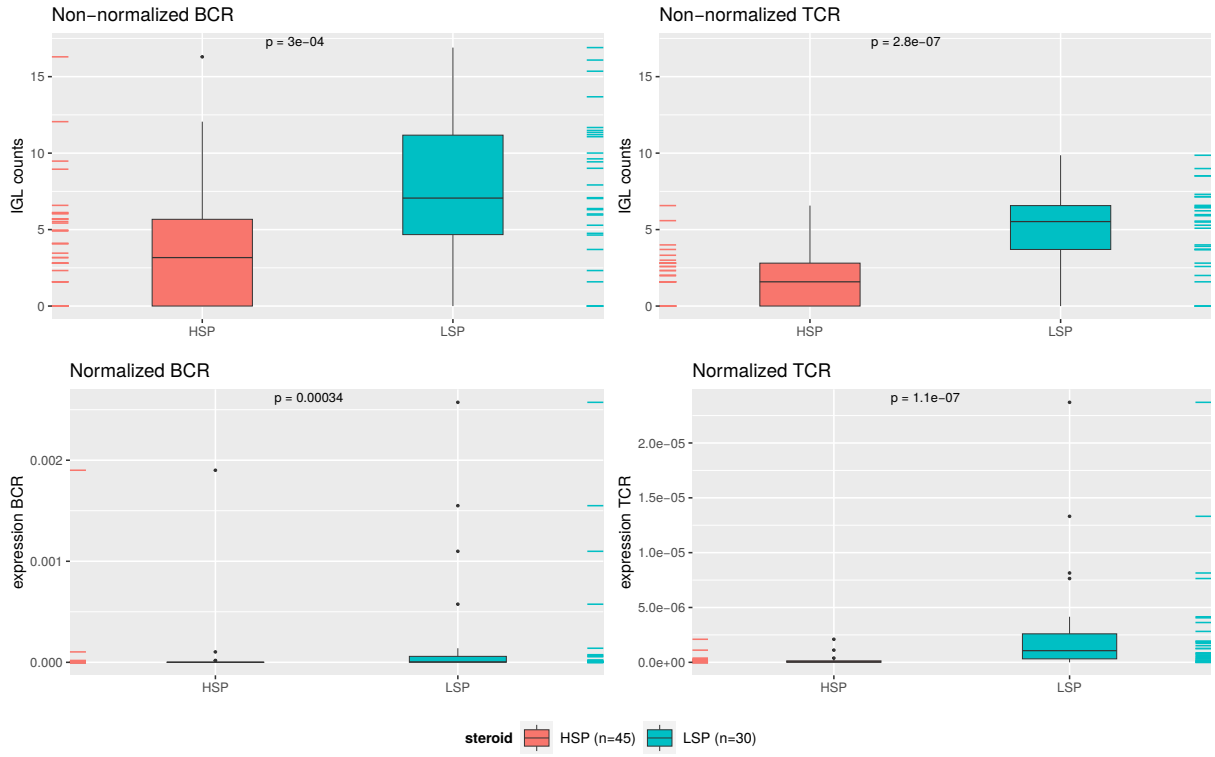

Figure 4: Comparação de contagens e expressões de TCR e BCR sob o perfil esteroidal. Teste de Wilcoxon foi utilizado para comparação entre os grupos.

Immunarch

Contextualização

Immunarch é um pacote R com funções específicas para análise de repertórios de receptores de células T e células B. Ele foi desenvolvido pela ImmunoMind (startup da UC Berkeley SkyDeck). Seu diferencial é a possibilidade de trabalhar com diferentes formatos resultantes dos softwares popualares de análise e pós-análise de TCR e BCR. O Immunarch possui funções para cálculos de métricas no contexto de TCR e BCR como sobreposição de repertório, estimativa de uso de genes, diversidade, rastreamento de clonótipos além de fornecer funções de plots com parâmetros de funções de gráficos que podem ser ajustadas conforme a escolha do usuário.

Na etapa em que o projeto está é interessante a exploração de ferramentas e métricas usadas em análises de TCR e BCR. O immunarch tem chamado atenção, devido as métricas que ele calcula envolvendo e as funções de plots pré-configuradas. Nesta ideia comecei a executar algumas funções e plotar alguns gráficos.

Metodologia

Executei as principais funções do pacote immunarch com métricas recomendadas. Gerei alguns gráficos mostrando a diferença entre o número de clonótipos únicos, distribuição do comprimento da região CDR3, sobreposição de clonotipos e diversidade.

Resultados

O intuito deste momento não é olhar afundo a parte biológica, mas mostrar o que o immunarch faz. Nos últimos dias desde a última reunião tenho me dedicado nas metodologias de artigos que analisaram o TCR e BCR em dados de RNA-Seq.

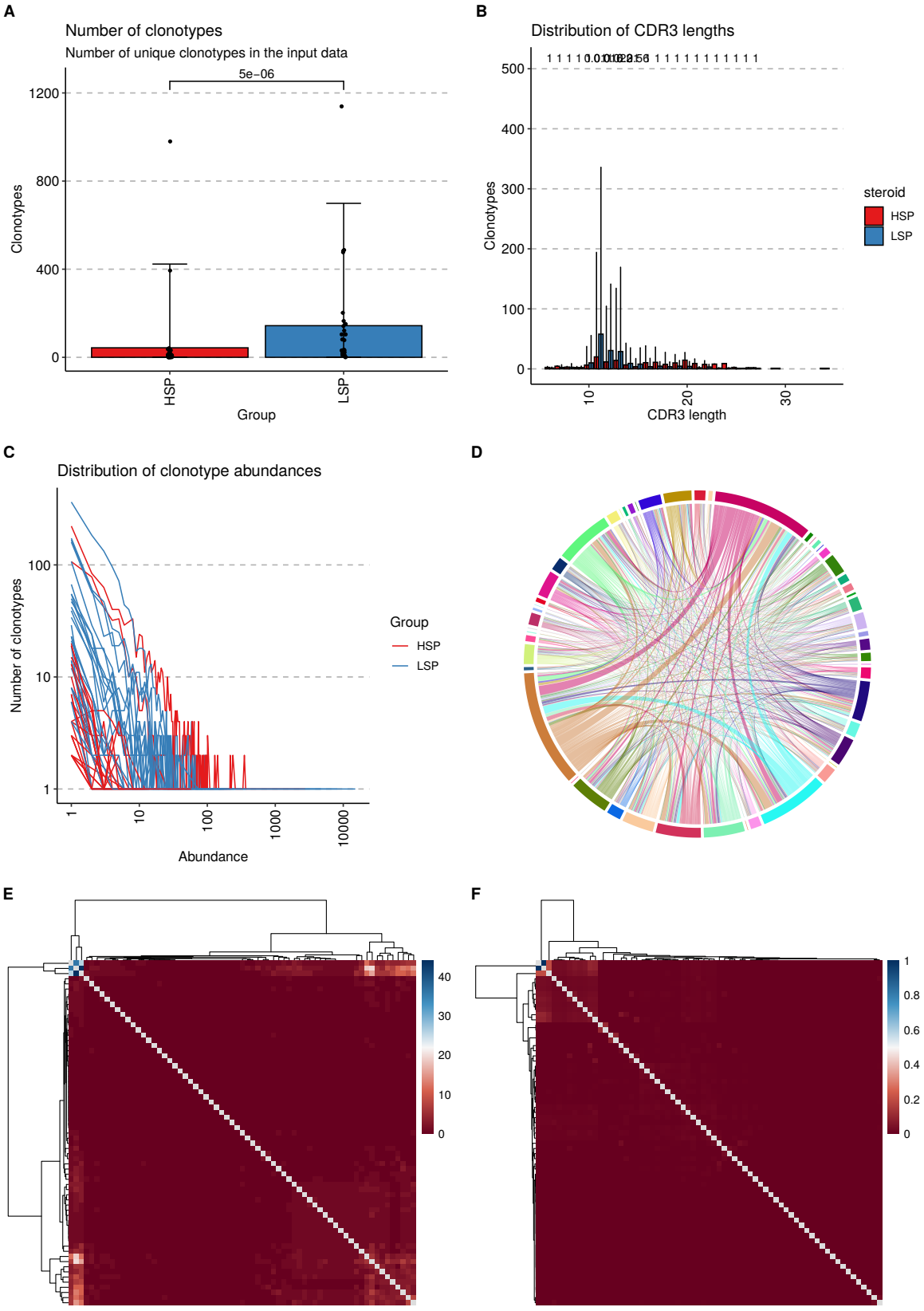

Figure 5: Alguns exemplos de resultados do immunarch. **A**. Número de clonótipos. **B**. Comprimento da região CDR3. **C** Distribuição do número de clonótipos e abundância. **D** Sobreposição dee repertórios. **E**. Sopreposição de repertórios com método 'público'. **F**. Sobreposição de repertórios com método Jaccard.