10 Aplicações da bioinformática na agricultura

Adhemar Zerlotini Neto
Antonio Nhani Jr.
Fábio Danilo Vieira
Leandro Carrijo Cintra
Maurício de Alvarenga Mudadu
Paula Regina Kuser Falcão
Poliana Fernanda Giachetto

1 Introdução

A biotecnologia tem sido fundamental para o avanço observado na Agropecuária nos últimos 30 anos. A bioinformática, área multidisciplinar responsável pela análise do grande volume de dados resultantes das tecnologias genômicas, foi imprescindível nesse avanço. Com o advento das chamadas tecnologias de sequenciamento de nova geração, passou a ser produzido um volume extraordinariamente grande de dados genômicos que precisavam ser analisados. Na era da transformação digital, a capacidade de geração de dados biológicos cada vez mais rápida, com valores mais acessíveis e em maior volume produz uma vasta quantidade de dados, o *Big Data*. Esse grande e crescente volume de dados exige soluções em pelo menos três âmbitos: infraestrutura escalável, gerenciamento dos dados e uso inteligente desses dados.

A bioinformática utiliza ferramentas computacionais para responder a perguntas biológicas complexas e contribuir com resultados inovadores. O tema envolve o uso de uma infraestrutura de computação de alto desempenho e ferramentas para organizar, analisar, integrar, processar, simular e armazenar grandes volumes de dados derivados de experimentos *in vivo* e *in vitro*. Um desafio da bioinformática é integrar os dados heterogêneos gerados

pelas ciências "ômicas" (tanto entre si como com os dados gerados pelas ciências "tradicionais"), permitindo descobertas que vão além das possíveis em cada uma das disciplinas individualmente. Várias novas camadas de "ômicas", como análises de genomas, metabolomas, transcriptomas, interactomas, tornaram-se importantes para os avanços das pesquisas. A integração de toda essa informação permite fazer descobertas e melhorar o conhecimento dos sistemas biológicos.

Acesso a alta capacidade de armazenamento e processamento, com poderosos algoritmos de indexação, assim como aplicações com aprendizado de máquina, é indispensável para a execução de atividades de bioinformática. Mais importante, uma equipe capacitada e em constante atualização para auxiliar no planejamento dos processos de geração dos dados, na análise de dados e na extração/obtenção de novos conhecimentos a partir do *Big Data* é o que vai propiciar à Embrapa ser um ator relevante nessa área do conhecimento.

Nesse contexto, em 2011, foi criado o Laboratório Multiusuário de Bioinformática (LMB) da Embrapa, com o propósito de dar suporte em bioinformática aos projetos de PD&I alinhados com os objetivos estratégicos da Embrapa. Desde sua criação, o LMB já atendeu uma ampla carteira de projetos, dentro de três diretrizes de atuação:

- **Acesso ao parque computacional** com infraestrutura de alto desempenho;
- **Consultoria na análise de dados biológicos** que requerem computação de alto desempenho, seja pelo volume de dados, seja pela complexidade das análises;
- **Treinamentos** visando multiplicar competências através de cursos e outras ações de capacitação.

O LMB tem atuado em projetos de pesquisa da Embrapa e de instituições parceiras, envolvendo mais de 20 culturas e criações estudadas em mais de 50 projetos de pesquisa. Uma peculiaridade importante é que, em bioinformática, cada projeto é único, e a equipe do LMB trabalha para atender essas demandas. Sua atuação em bioinformática baseia-se nas áreas de: análise da expressão gênica, montagem e análise de genomas, identificação de marcadores moleculares, análise de transcriptomas e metagenomas, estudos de evolução, modelagem de sistemas biológicos, predição de estruturas proteicas e interação molecular, interação ou inibição de moléculas, entre outras atividades.

1.1 Infraestrutura computacional do LMB para suporte a projetos de bioinformática aplicada à agropecuária

Os projetos em bioinformática exigem uma infraestrutura computacional diferenciada, sendo muito difícil ou mesmo impossível a execução da maioria deles apenas com o uso de equipamentos computacionais comuns. Pode-se

compreender o motivo de tais requisitos quando se considera a complexidade computacional dos algoritmos executados e o volume de dados biológicos analisados.

O objetivo desta sessão é apresentar a infraestrutura computacional utilizada para o armazenamento e o processamento do grande volume de dados produzidos pelos projetos de pesquisa em biotecnologia da Embrapa e de suas instituições parceiras. Essa infraestrutura está focada na disponibilização de capacidade de processamento e memória e de armazenamento de grandes volumes de dados.

Para fazer frente aos diversos algoritmos com alta complexidade computacional presentes na bioinformática, é padrão a utilização de *cluster*s computacionais para o processamento de dados. Para aqueles que estão menos familiarizados com a área de computação de alto desempenho, um *cluster* é formado por um conjunto de computadores ligados em rede com um nó de coordenação central, sendo utilizados juntos na solução dos problemas computacionais. A principal vantagem de um *cluster* é proporcionar o poder de computação de dezenas, centenas e, em alguns casos extremos, milhares de nós de processamento de uma forma transparente para o usuário, ou seja, sem que ele necessite interagir e disparar análises de dados em cada uma das máquinas individualmente. As tarefas (*jobs*) a serem executadas no sistema são disparadas a partir de um nó de gerenciamento, permanecendo em uma ou mais filas de execução e sendo automaticamente enviadas para um nó de processamento adequado, quando possível.

Com o advento da computação em multicores, cada nó de processamento nos *cluster*s modernos tem algumas dezenas de núcleos; em algumas situações excepcionais, cada nó pode chegar a centenas de núcleos de processamento. Surge, então, uma questão muito importante para o processamento em bioinformática: quanto de memória deverá ter cada nó de processamento? Para a resposta, deve-se considerar que a quantidade de memória é diretamente proporcional ao número de núcleos no nó de processamento; além disso, deve-se considerar que essa proporção tem aumentado com o desenvolvimento de novas técnicas de investigação biológicas, que geram quantidades de dados cada vez mais significativas. Sendo assim, até há pouco tempo, o recomendado era que houvesse em cada nó de processamento 8 Gb de memória RAM para cada núcleo disponível. Com o significativo aumento no volume de geração dos dados biológicos, essa quantidade sofreu uma atualização, e as novas plataformas de processamento destinadas a atividades de bioinformática estão sendo desenvolvidas com 16 Gb de memória RAM para cada núcleo disponível no nó.

Outra questão relevante nas plataformas de processamento de dados biológicos está relacionada com o armazenamento e a preservação dos dados. Basicamente, o gargalo mais significativo a ser superado aqui diz respeito à quantidade de dados a ser armazenada. A velocidade de acesso a esses dados

não impacta significativamente no desempenho das plataformas, pois, no geral, as ferramentas e os programas executados para a realização das análises irão carregar os dados para a memória e executar os procedimentos por um tempo significativo. Um atraso na carga inicial não modifica demasiadamente o tempo total de execução da tarefa. No entanto, uma restrição na capacidade de armazenamento do ambiente computacional terá uma ampla gama de ocorrências negativas. Não se pode trabalhar com diversos projetos ao mesmo tempo, pois comumente eles demandam algumas centenas de gigabytes, podendo, para alguns projetos excepcionais, chegar a algumas dezenas de terabytes para o armazenamento dos dados brutos. Durante as análises, é necessário armazenar dados intermediários, que podem ser de até uma ordem de grandeza dos dados originais. Sendo assim, atualmente as plataformas para processamento de dados biológicos utilizam corriqueiramente sistemas de armazenamento com capacidade de alguns petabytes.

O ambiente de processamento disponível hoje possui um *cluster* com um nó de controle (*head node*) e 14 (quatorze) nós de processamento. Destes, 13 (treze) têm 64 núcleos e 512 Gb de memória RAM cada um. Há também um nó especial, utilizado para a execução de tarefas que exigem grande quantidade de memória. Esse nó possui 2 Tb de memória RAM e 160 núcleos para processamento. No total, o *cluster* disponibiliza 992 núcleos. Para o gerenciamento das tarefas no *cluster*, utiliza-se um sistema gestor de filas, desenvolvido inicialmente pela *Sun Microsystems,* conhecido como *SGE – Sun Grid Engine.* Especificamente para a bioinformática, uma plataforma computacional em *cluster* é bastante útil, pois, em geral, os problemas da área envolvem múltiplos conjuntos de dados (*datasets*) processados em *pipelines* constituídos por múltiplos estágios, sendo fácil a paralelização do processamento em máquinas separadas no ambiente. Problemas com tais características são ideais para a execução em *cluster*s.

Para o armazenamento de dados, estão disponíveis: um *storage SGI Infinite* com capacidade para 150 Tb em uma configuração com RAID 6; e um *storage IBM DS3412* com capacidade para armazenar 51 Tb em uma configuração com RAID 5. Além do armazenamento principal, é vital que haja uma política de backup que garanta a segurança dos dados na plataforma. Em virtude do volume de informações recebido e gerado constantemente, a metodologia com o melhor custo-benefício para o backup envolve o uso de fitas LTO. Atualmente, há no ambiente uma biblioteca de fitas com capacidade para 44 unidades LTO6. Como cada fita LTO6 propicia, em média, o armazenamento de 6,25 Tb de dados, a biblioteca tem a capacidade de tratar até 275 Tb de backup on-line.

Esse tipo de infraestrutura computacional é indispensável para a execução das análises dos dados de projetos de pesquisa em bioinformática na agricultura.

2 Aplicações

2.1 A bioinformática e a cadeia produtiva do tambaqui

O primeiro objetivo estratégico da Embrapa é "desenvolver conhecimentos e tecnologias para o adequado manejo e aproveitamento sustentável dos biomas brasileiros". A Embrapa, historicamente, sempre se preocupou com o desenvolvimento regional, atuando em linhas de frente em que o risco científico ou econômico eram fatores desestimuladores para a iniciativa privada. Em outras palavras, enfrentando problemas agropecuários que empresas privadas do setor avaliavam ser economicamente inviáveis. Esse papel insubstituível da Embrapa garantiu, para citar um único exemplo, o aproveitamento do bioma Cerrado para a agricultura, levando desenvolvimento e riqueza à região. O Brasil é um país continental com desigualdades socioeconômicas entre suas regiões geográficas. A região Norte é rica em recursos naturais, mas suas cadeias produtivas ainda carecem de desenvolvimento e inovação. Encontra-se aí uma cadeia produtiva de pescado cuja produção anual de peixes nativos é de 290 mil toneladas, segundo o Anuário da Piscicultura 2019, sendo o tambaqui (*Colossoma macropomum*) o principal produto. Visando promover o desenvolvimento dessa importante cadeia produtiva, entre outros objetivos igualmente relevantes, a Embrapa, através do projeto BRS Aqua[1], identificou pontos críticos para o incremento da produção que, se adequadamente resolvidos, aumentariam a competitividade e a sustentabilidade da cadeia produtiva do tambaqui.

Um dos pontos críticos identificado pela Embrapa na cadeia produtiva do tambaqui[2] foi a ocorrência de cruzamentos entre matrizes aparentadas. Muitos piscicultores não sabem, mas a simples escolha das matrizes para o cruzamento pode, se mal feita, reduzir de 10% a 30% o peso final dos peixes. Isto é, usando a mesma quantidade de ração na alimentação, o produtor poderia perder até 30% de conversão alimentar. Na literatura científica, esse fenômeno é conhecido como depressão endogâmica, e poucos produtores de peixes sabem de sua existência. Só para dimensionar o tamanho do problema, veja-se o caso dos peixes nativos que são muito apreciados na região Norte. Como mencionado, em 2019, a produção foi de 290 mil toneladas, assumindo uma estimativa conservadora, pois o cruzamento de matrizes aparentadas pode ter impactado negativamente a produção em pelo menos

1 O projeto BRS Aqua tem financiamento do Fundo Tecnológico do BNDES/Funtec, da Secretaria de Pesca e Aquicultura (SAP) do Ministério da Agricultura, Pecuária e Abastecimento (MAPA), do CNPq, da FAPDF e Embrapa. Nesta parte do projeto BRS Aqua atuaram, principalmente, as Unidades: Embrapa Recursos Genéticos e Biotecnologia, Embrapa Pesca e Aquicultura e Embrapa Informática Agropecuária.

2 Esse ponto crítico ocorre em todas as cadeias produtivas de peixe em que não há como identificar o parentesco entre as matrizes.

30 mil toneladas. Diz-se "pelo menos" porque, além da depressão endogâmica, o cruzamento entre matrizes aparentadas acarreta mais um fenômeno danoso, conhecido cientificamente como "alelos fatais". Numa população qualquer, alelos fatais são raros; porém, quando ocorrem em homozigose, prejudicam o desenvolvimento do embrião. Ou seja, esses alelos causam deformações nos embriões ou abortam o seu desenvolvimento quando herdados tanto do pai quanto da mãe. Daí a recomendação de se evitarem casamentos consanguíneos. Se esses alelos são raros na população como um todo, dentro de famílias portadoras desses alelos a ocorrência da homozigose é significativamente mais frequente, chegando até 25%. Isto é, num cruzamento consanguíneo, até 25% dos embriões podem ser perdidos ou apresentar defeitos congênitos. Tanto a depressão endogâmica quanto os alelos fatais são problemas críticos na cadeia produtiva dos peixes, mas que os produtores ou simplesmente ignoram ou não têm condições técnicas de evitar tais cruzamentos pela dificuldade de aferir o grau de parentesco entre as matrizes.

Além da depressão endogâmica e dos alelos fatais, outro ponto crítico é a existência de híbridos férteis no plantel de matrizes. Nas aulas de Biologia, aprende-se que, quando duas espécies diferentes cruzam, o resultado é um animal infértil. Infelizmente, no caso dos peixes, isso nem sempre é verdade. Por exemplo, o tambaqui pode cruzar com o pacu (*Piaractus mesopotamicus*), e o híbrido é um animal fértil. Ocorre, entretanto, que muitos produtores realizam o cruzamento de tambaqui com pacu porque os híbridos ganham mais peso que os animais puros e o sabor da carne não é significativamente afetado. Na literatura, esse fenômeno é conhecido como "vigor híbrido", e é bastante usado na produção de grãos, por exemplo. O problema ocorre quando híbridos, em vez de ir para o abate, são erroneamente escolhidos para compor o plantel de matrizes. Embora essa escolha possa parecer improvável num primeiro momento, ela ocorre porque a seleção muitas vezes se baseia nas características externas, e, devido ao vigor híbrido, não é raro que um híbrido seja erroneamente selecionado por apresentar maior peso, por exemplo. Nesse caso, como os híbridos são férteis, o erro da escolha só será descoberto no momento do cruzamento, quando o produtor observará a segregação natural que acarreta muita variabilidade nas características de interesse econômico, como o peso ao abate. Produtores que vendem alevinos para engorda podem ter sua credibilidade afetada por vender animais de baixa qualidade, pois a variabilidade da segregação afeta bastante a engorda.

Uma vez identificados esses problemas, os pesquisadores da Embrapa desenvolveram dois chips de DNA que resolvem tais questões de forma inovadora, eficiente e de baixo custo. Nesses chips, marcadores moleculares, conhecidos como polimorfismos de um único nucleotídeo, ou simplesmente SNPs, podem fornecer informações suficientes para determinar o grau de parentesco e pureza de espécie. No caso de parentesco, os marcadores devem

possuir bastante variabilidade na população estudada. Matematicamente, isso equivale a exigir que o *Minor Allele Frequency* (MAF) seja próximo a 0,5. O princípio é exatamente o mesmo de um teste de paternidade, só que nessa aplicação busca-se identificar qualquer grau de parentesco para evitar cruzamentos consanguíneos, diminuindo a depressão endogâmica e minimizando a ocorrência de alelos fatais. O desafio científico é justamente escolher os tais marcadores moleculares SNPs. No caso do tambaqui, por exemplo, a falta de um genoma de referência publicamente disponível foi um primeiro obstáculo a ser vencido. Isso levou a Embrapa a realizar, internamente, um Projeto Genoma do Tambaqui, e o LMB foi responsável pela montagem desse Genoma Tambaqui. São aproximadamente 1,3 bilhão de nucleotídeos divididos em 27 cromossomos (ou grupos de ligação). De posse do genoma, o próximo passo foi selecionar uma subpopulação representativa da população de tambaquis e sequenciar o DNA do pool dessa subpopulação. O resultado desse sequenciamento foi mapeado no genoma de referência, e finalmente foi realizada a descoberta de SNPs. Mesmo com a exigência de uma cobertura mínima de 150X, foram identificados mais de 2 milhões de SNPs (Ianella et al., 2019). Apesar do número significativo, apenas um subconjunto satisfaz inúmeras exigências. A tarefa de selecionar 96 SNPs para compor o chip de parentesco levou em conta o MAF, o espaçamento dentro dos cromossomos, a anotação funcional e, finalmente, a ausência de variações genômicas nas regiões flanqueadoras do SNP candidato. Como se pode notar, o trabalho de bioinformática foi bastante intenso para realizar todas essas tarefas, o que justifica a necessidade de uma infraestrutura como a do LMB. Após a fase de validação dos SNPs em uma população diferente daquela usada na fase anterior, os SNPs validados foram incorporados ao chip, que se mostrou extremamente eficiente na determinação do grau de parentesco e já está sendo usado na cadeia produtiva do tambaqui. Ou seja, o produtor já possui uma ferramenta inovadora para eliminar a depressão endogâmica e os alelos fatais, evitando assim prejuízos silenciosos causados pelos cruzamentos consanguíneos.

Já o chip de DNA para determinação de pureza exigiu análises mais complexas. Isso porque foi necessário incluir na análise mais duas espécies que cruzam com o tambaqui e produzem híbridos férteis, a saber, o pacu e a caranha (*Piaractus brachypomus*). Como nenhuma dessas espécies possui genoma de referência, foi necessário usar o genoma do tambaqui como referência. Esse procedimento não é trivial porque, além das variações intraespécies, há também as variações interespécies (tambaqui x pacu / tambaqui x caranha), o que aumenta o grau de complexidade das análises. Até mesmo na fase de mapeamento dos *reads* no genoma de referência, a exigência de similaridade teve de ser reduzida devido às diferenças interespecíficas. Diferentemente dos SNPs de parentesco, os SNPs para aferição de pureza de espécie devem estar

"fixados", isto é, não apresentar variação na espécie, ou seja, MAF = 0. Um exemplo pode ajudar a entender melhor o problema. Se em uma determinada posição no genoma tem-se um nucleotídeo "A" fixado no tambaqui, e nessa mesma posição tem-se o nucleotídeo "C" fixado no pacu, então essa posição do genoma é uma séria candidata a compor o chip de pureza de espécie, pois, num teste de DNA, um resultado "A" significaria "tambaqui" e um "C", pacu. E para reduzir ainda mais os custos, buscou-se por marcadores genômicos capazes de separar simultaneamente o tambaqui das outras duas espécies. No exemplo anterior, isso significaria que a caranha também tivesse um "C" fixado naquela mesma posição genômica[3]. Dessa forma, com um único chip de DNA é possível avaliar a pureza do tambaqui em relação às duas principais espécies que produzem híbridos[4]. Mais uma vez, usando frequência alélica, espaçamento físico no genoma e anotação funcional, foram selecionados 96 SNPs para compor o chip, e depois da fase de validação em populações independentes, os SNPs validados foram incorporados ao chip de aferição de pureza. Com essa ferramenta genômica pode-se eliminar todos os híbridos que erroneamente tenham sido escolhidos para compor o plantel de matrizes.

Estudos de impacto econômico realizados pela Embrapa, supondo uma produção média de 150 mil toneladas de tambaqui, preveem ganhos adicionais entre R$ 9 milhões e R$ 28 milhões para os produtores[5]. Cada análise de amostras para pureza e parentesco, atualmente, custa R$ 60,00. Para um produtor com 100 matrizes, isso equivaleria a um investimento de R$ 12 mil. Como cada matriz tem uma vida útil de três anos, esse valor é amortizado por igual período. O investimento é insignificante quando comparado ao retorno. Tanto é assim que essas duas tecnologias, batizadas de TambaPlus[6], já foram adotadas por produtores de cinco estados: Mato Grosso, Tocantins, Roraima, Amazonas e Rondônia. E mais de 1.500 testes já foram realizados. A importância do TambaPlus é tal que a tecnologia foi selecionada para compor um seleto grupo de tecnologias que foram destaque no 47º Aniversário da Embrapa[7].

3 Os SNPs são marcadores bialélicos, o que viabiliza separar uma espécie de outras duas simultaneamente. Há SNPs trialélicos, porém são raríssimos, e, portanto, não é factível produzir um único chip de genotipagem que separe as três espécies duas a duas simultaneamente.

4 Pelo que já foi exposto, esse chip de pureza não separa pacu de caranha.

5 Notícia fornecida em vídeo conferência intitulada TambaPlus®: Ferramentas genômicas para análise e gestão de matrizes de tambaqui destinadas à produção de alevinos, disponível na plataforma Agrotins: https://agrotins.to.gov.br/programacao/tambaplus-ferramentas-genomicas-para-analise-e-gestao-de-matrizes-de-tambaqui-destinadas-a-produca.html. Acesso em: 23 jun 2020.

6 Disponível em: https://www.embrapa.br/busca-de-noticias/-/noticia/46203188/ferramentas-genomicas-ajudarao-a-evitar-cruzamentos-consanguineos-entre-matrizes-de-tambaqui

7 Disponível em: https://www.embrapa.br/47-anos/solucoes-tecnologicas-em-destaque?link=47-anos

As pesquisas na cadeia produtiva do tambaqui prosseguirão. Ainda há muito espaço para aprimorar a produção de peixes. Em qualquer programa de melhoramento genético há, grosso modo, duas fases principais, a saber, a fase de Seleção e a de Cruzamento. O tambaqui ainda está numa etapa anterior, conhecida como pré-melhoramento. Nesse início, a preocupação principal foi evitar cruzamento consanguíneo e a presença de híbridos no plantel de matrizes.

2.2 Bioinformática no desenvolvimento de vacinas: vacinologia reversa

Na produção animal, a utilização de vacinas é uma alternativa efetiva e de menor custo para prevenção ou redução da severidade de doenças que afetam os rebanhos. A vacinação contribui para a manutenção da saúde e do bem-estar animal, para o aumento da eficiência na produção de alimentos e para a redução da transmissão de zoonoses. Comparadas a outras formas de controle, como o uso de antibióticos e pesticidas, as vacinas apresentam vantagens, como a não contaminação do meio ambiente e dos produtos de origem animal (carne, leite e ovos).

Seguindo a metodologia convencional de desenvolvimento de vacinas, o patógeno é cultivado *in vitro* no laboratório e utilizado em sua forma atenuada (na qual perde a habilidade de causar a doença) ou morta para elicitar uma resposta imune protetora no hospedeiro. Alternativamente, componentes purificados do patógeno também podem ser utilizados como antígenos, nas chamadas vacinas de subunidades (Rappuoli; Covacci, 2003).

Embora as vacinas obtidas da forma convencional figurem entre as invenções mais importantes da humanidade, constituindo uma ferramenta poderosa no combate aos agentes biológicos causadores de doenças, nem todos os patógenos podem ser cultivados *in vitro* e utilizados no desenvolvimento de vacinas, na sua forma convencional. Além disso, os métodos convencionais são bastante demorados, podendo ser necessário de cinco a 15 anos para a obtenção de uma vacina eficaz (Vernikos, 2008).

A vacinologia reversa, metodologia publicada pela primeira vez por Rappuoli (2000), surgiu como uma estratégia alternativa para a descoberta de antígenos protetores para o desenvolvimento de vacinas que parte da análise do genoma do patógeno alvo. Viabilizada em função do sequenciamento genético em larga escala, juntamente com o desenvolvimento de ferramentas de bioinformática, a vacinologia reversa utiliza ferramentas de predição *in silico* para a identificação de alvos (antígenos) para o desenvolvimento de vacinas. Por meio dessas ferramentas, genomas, transcriptomas e proteomas são examinados *in silico*, proteínas preditas são selecionadas com base em atributos desejáveis – que podem induzir uma resposta imune capaz de proteger contra uma determinada doença, e os alvos são então identificados.

A partir deles, diferentes tipos de vacinas podem ser delineados e desenvolvidos dentro de um intervalo de um a dois anos.

Vacinas comerciais obtidas por meio dessa metodologia já são realidade. Uma vacina desenvolvida contra a doença meningocócica invasiva, causada pela bactéria *Neisseria meningitidis* sorogrupo B, foi liberada para uso na Europa em 2014 (Andrews; Pollard, 2014). Nessa vacina, a resposta imune é desencadeada por epítopos – sequências específicas de resíduos de aminoácidos presentes no antígeno que participam diretamente da interação com anticorpos, que foram identificados por meio de ferramentas de bioinformática. Os epítopos têm sido considerados particularmente interessantes no desenvolvimento de vacinas, uma vez que tem sido demonstrado que vacinas compostas por esses peptídeos são capazes de otimizar ou mesmo exceder o potencial de proteção induzido pela proteína nativa cognata (Kao; Hodges, 2009). Em contraste com as vacinas vivas atenuadas, uma vacina contendo um epítopo sintético não é capaz de reverter a virulência de um patógeno (Palatnik-De-Sousa et al., 2018). Ainda, vacinas baseadas em epítopos são mais específicas, não induzindo respostas imunes indesejáveis, são capazes de gerar imunidade de longa duração e são mais baratas do que as vacinas convencionais (Ahmad et al., 2016).

Na abordagem da vacinologia reversa, as sequências das proteínas de um organismo são analisadas utilizando-se programas de predição *in silico*. Essas proteínas, no entanto, são, em sua grande maioria, preditas a partir do sequenciamento de genomas e transcriptomas, por meio de ferramentas de bioinformática. Isso porque o sequenciamento genético em larga escala, possível graças às novas tecnologias que reduziram dramaticamente o custo de geração das sequências, do mesmo modo que aumentou exponencialmente o número de sequências geradas a partir de uma amostra, tem acumulado uma quantidade de dados genômicos e transcriptômicos sem precedentes. Por outro lado, um avanço tecnológico que permitisse o desenvolvimento de técnicas de sequenciamento de proteínas com elevada sensibilidade e em larga escala ainda não aconteceu. O avanço nas metodologias de obtenção de sequências expressas causou uma subsequente evolução nas metodologias de análise. Uma lista de programas pode ser acessada na página "*List of RNA-Seq bioinformatics tools*" (Wikipedia, 2020). Faremos a seguir uma breve descrição comentada da metodologia aplicada para obtenção de genes diferencialmente expressos na glândula salivar do carrapato bovino (Andreotti et al., 2018). Todas as ferramentas citadas são obtidas através de licença acadêmica ou de instituição de pesquisa governamental, ou possuem distribuição livre.

Com o objetivo de melhor compreender a interação parasita-hospedeiro e identificar possíveis genes e mecanismos envolvidos, um estudo iniciado em 2015, financiado pela Embrapa, gerou mais de 600 milhões de sequências a partir do sequenciamento do RNA (utilizando a metodologia de RNA-Seq)

de larvas, ninfas, glândula salivar, intestino e ovários do carrapato do boi, *Rhipicephalus* (*Boophilus*) *microplus* (Andreotti et al., 2018). Além da caracterização dos transcriptomas dos diferentes tecidos, por meio da montagem *de novo*, nosso grupo de pesquisa também identificou os genes diferencialmente expressos (GDE) entre carrapatos crescidos em bovinos resistentes (Nelore), bovinos susceptíveis (Holstein) e animais cruzados, com resistência intermediária ao parasita (Nelore x Holstein). A análise desse conjunto de dados, por meio de ferramentas que informam a função das proteínas preditas pelos GDE e as vias biológicas em que atuam, trouxe novas descobertas acerca da interação carrapato-bovino e apontou potenciais candidatos que podem ser utilizados como antígenos no desenvolvimento de vacinas para o controle do carrapato bovino (Giachetto et al., 2020).

O primeiro passo na análise de *RNA-Seq* é a verificação da qualidade das sequências geradas. Ferramentas como *FastX Toolkit* (FastX-GitHub, 2020) e *FastQC* (FastQC-GitHub, 2020) verificam vários parâmetros, dentre os quais destacamos:

- Qualidade média de bases e qualidade média por sequência. Para um bom resultado, a sequência deve ter um "Phred score" superior a 30.
- Conteúdo de GC (%GC). A porcentagem da presença das bases nucleotídicas Guanina e Citosina na sequência deve aproximar-se da distribuição normal, uma vez que o conteúdo em GC muito elevado dificulta a síntese e, muitas vezes, o agrupamento (contigagem) das sequências durante os processos de obtenção e montagem.
- Quantidade de bases indeterminadas (%N). Bases indeterminadas dificultam o processo de contigagem. Podem ocorrer no início do sequenciamento, onde existe uma saturação de reagentes; no final, pela diminuição da concentração de reagentes; ou em uma região com alta %GC, que dificulta a leitura da região pela polimerase.
- Presença de adaptadores. Adaptadores são sequências de nucleotídeos curtas, utilizadas para a preparação da biblioteca e o sequenciamento. Sua presença prejudica a contigagem, originando sequências quiméricas. Para eliminá-los, ferramentas como Trimmomatic (Bolger et al., 2014) e Trim Galore (TrimGalore-GitHub, 2020) são frequentemente utilizadas.

Como lidamos com um grande número de sequências, uma ótima ferramenta para agrupar e visualizar os dados obtidos na análise de qualidade (e mesmo passos posteriores) é o MultiQC (Ewels et al., 2016), que organiza os resultados obtidos em uma página web.

Com a qualidade das sequências verificada, passamos para a obtenção do transcriptoma, através da comparação sequência a sequência e da contigagem delas por similaridade. Várias ferramentas podem ser usadas nesse passo, citando, por exemplo, QUAST (Gurevich et al., 2013), que é recomendado

para a análise de metagenomas. A ferramenta de escolha para a análise deste trabalho foi o programa Trinity (Grabherr et al., 2011). Essa ferramenta é, na verdade, um *pipeline* que reúne, através de scripts desenvolvidos nas linguagens de programação Perl[8] e Python[9], várias ferramentas de análise para qualidade, contigagem de sequências e estatísticas para a identificação dos GDEs, tendo como diferencial a possibilidade de identificação de isoformas (o mesmo que transcritos) de um mesmo gene, oriundas do *splicing* alternativo. Diferentes tecidos podem expressar diferentes isoformas em diferentes quantidades. Identificar a isoforma expressa localmente possibilita melhor entendimento da expressão de um determinado gene em uma determinada via metabólica ou tecido.

Obtido o transcriptoma, o próximo passo é a verificação da qualidade da montagem. Uma primeira abordagem é o mapeamento das sequências utilizadas para a montagem de volta ao transcriptoma obtido. Em uma boa montagem, mais de 80% das sequências utilizadas mapeiam no transcriptoma. Uma segunda forma de avaliação consiste na identificação e na quantificação de sequências completas, através da análise de similaridade contra bancos de dados curados, como o SwissProt ou o TrEMBL (The UniProt Consortium, 2019), ou na busca de ortólogos presentes na mais próxima classificação do organismo estudado, neste caso, os artrópodos, utilizando o software BUSCO (Seppey et al., 2019).

Vários fatores influenciam o delineamento experimental de um ensaio de *RNA-Seq* para a identificação de GDEs:

- No preparo das amostras, desde a extração do *RNA* total até a obtenção de bibliotecas para sequenciamento, pode ocorrer o efeito de lote, em que são incluídos desde a utilização de diferentes soluções (feitas em dias diferentes) até quem as prepara (Conesa et al., 2016);
- A profundidade de sequenciamento (o número de sequências geradas), que influencia no número de sequências obtidas e, portanto, na quantificação do número de GDEs identificados (Conesa et al., 2016; Lamarre et al., 2018);
- O número de réplicas técnicas (quantas vezes uma mesma amostra é sequenciada), que influencia no poder estatístico para a detecção dos GDEs, sendo recomendadas não menos que três repetições (Conesa et al., 2016), embora um maior número (cerca de seis repetições) possa aumentar a representatividade das sequências do transcriptoma (Lamarre et al., 2018). É comumente aceito um ensaio com triplicatas, pois o aumento de réplicas implica no aumento do custo de ensaio;

8 Disponível em: https://www.perl.org

9 Disponível em: https://www.python.org

- A preparação de uma repetição biológica. Conesa et al. (2016) apontam que a variabilidade biológica é particular para cada ensaio, e apesar de difícil de controlar, é importante para um estudo que envolve populações, sugerindo ao menos que a amostra biológica seja feita em triplicata. Lamarre et al. (2018) apontam a detecção de até 20% de GDEs devido à variabilidade biológica, o que pode não justificar elevar os custos do ensaio.

A correlação entre as amostras utilizadas no ensaio é também uma medida importante da qualidade da montagem e das bibliotecas construídas. A análise de componentes principais permite visualizar correlações entre replicatas técnicas e biológicas, que devem, preferencialmente, formar agrupamentos não muito distantes. Uma discrepância entre amostras de um mesmo grupo pode indicar contaminação, mistura de amostras, erro de sequenciamento ou efeitos de lote, que devem ser considerados para o descarte da referida amostra. Importante também o fato de que sem uma triplicata técnica, uma duplicata biológica deverá ser descartada, prejudicando toda a análise.

Com um transcriptoma de boa qualidade, passamos à identificação das sequências diferencialmente expressas, os GDEs. Trinity incorpora diversas ferramentas estatísticas para esse fim. Neste caso, optamos pelo uso de RSEM (Li; Dewey, 2011), que estima a quantidade de cada transcrito realinhando as sequências de cada biblioteca (ou tratamento experimental) ao transcriptoma gerado – motivo da importância da qualidade e da relação entre as replicatas – e edgeR (Robinson et al., 2010), um pacote desenvolvido no programa estatístico R (R Core Team, 2020) e integrante do Projeto Bioconductor (Huber et al., 2015) para análise de dados biológicos, que realiza a comparação par a par das sequências geradas entre todas as amostras e identifica aquelas com expressão diferencial.

O penúltimo passo é a anotação (ou identificação) de cada sequência diferencialmente expressa, através da análise de similaridade em bancos de dados de sequências nucleotídicas e proteicas, buscando homologia a sequências já conhecidas, e em bancos de dados de vias metabólicas que informam em qual (quais) delas o gene participa. Seguem uma análise manual de cada resultado, o embasamento bibliográfico buscando a importância de tal gene ao desenvolvimento no ciclo de vida do carrapato, e a seleção de possíveis alvos para a fabricação de vacinas.

A existência de vacinas comerciais disponíveis para o controle do carrapato bovino demonstrou que elas podem atuar de maneira efetiva no controle das infestações, reduzindo a aplicação de acaricidas. A adoção dessas vacinas, no entanto, tem sido limitada, principalmente, por não se mostrarem efetivas contra todos os estágios de vida do parasita, além de apresentarem baixa eficácia contra algumas cepas regionais do *R.* (*B.*) *microplus* (Andreotti,

2006). Resultados obtidos em teste conduzido pela Embrapa com um isolado regional do carrapato evidenciaram uma eficácia de 46,4% e 49,2%, respectivamente, para as vacinas TickGARD® e GavacTM (Andreotti, 2006). Assim, tendo como base o banco de dados descrito anteriormente, nossa equipe coordena hoje um estudo que prevê a identificação de epítopos imunogênicos candidatos ao desenvolvimento de vacinas contra o carrapato bovino, utilizando a metodologia da vacinologia reversa, a partir das proteínas preditas dos transcriptomas do parasita. Por meio da execução de um *pipeline* contendo uma série de ferramentas de análise, os genes candidatos a alvos para a produção de vacinas são analisados quanto à presença dos epítopos que podem interagir com o sistema imune do bovino para a produção de anticorpos, auxiliando no combate à infestação do carrapato.

A obtenção de uma vacina com alta eficácia, utilizada de forma integrada em estratégias de controle do carrapato bovino, deverá reduzir consideravelmente as infestações dos rebanhos e as implicações relacionadas ao uso de acaricidas, que incluem, além do custo e da contaminação ambiental, uma preocupação crescente da população com a segurança alimentar, o que tem levado, cada vez mais, ao consumo de alimentos livres de resíduos químicos, obtidos a partir de sistemas produtivos sustentáveis. Ainda, com a validação do *pipeline* que estamos propondo, o LMB poderá aplicar a metodologia de vacinologia reversa na identificação de alvos para o controle de outros problemas de interesse da agropecuária.

2.3 Ferramentas de bioinformática

Conforme preconizado pela agricultura digital, para serem transformadas em conhecimento útil, as informações geradas a partir de experimentos biológicos devem estar acessíveis e, se possível, disponibilizadas na Internet. Os bioinformatas e os biólogos computacionais lidam com esse cenário há mais de uma década, num ambiente com infraestrutura adequada como a que foi descrita anteriormente, e implementam bibliotecas de software, kits de ferramentas, plataformas e bancos de dados para obter sucesso nesse assunto.

No LMB da Embrapa, várias ferramentas de análise de dados são utilizadas, e tornou-se necessária uma busca por uma solução de integração dos dados. Os resultados das análises são armazenados criteriosamente em uma estrutura de diretórios e relatórios são gerados. Algumas ferramentas geram resultados em formato já disponível para a Internet ou, até mesmo, podem ser executadas diretamente on-line. Duas ferramentas em desenvolvimento têm contribuído muito para a integração dos dados gerados e a transformação desses dados em informação.

2.3.1 Machado: um framework de integração de dados genômicos

Iniciou-se, em 2017, um projeto para descoberta de proteínas candidatas para *pipelines* de construção de plantas transgênicas (Prado et al., 2014; Napier et al., 2019) resistentes a estresses abióticos denominado PlantAnnot – desenvolvimento de um sistema de bioinformática aplicado na descoberta de genes relacionados a estresses abióticos em plantas, focado no tema de mudanças climáticas. Para realização desse projeto, um grande volume de dados genômicos foi extraído de bancos de dados públicos. O conjunto de dados extraído corresponde a 53 genomas de plantas, totalizando mais de 1,8 milhão de genes e mais de 2,3 milhões de proteínas. Esses dados foram utilizados para realizar análises computacionais de forma a selecionar 72 mil proteínas de interesse para os *pipelines*. Um dos objetivos do projeto era o de armazenar e disponibilizar os dados e as análises realizadas.

Para solucionar esse problema de uma forma mais ampla, desenvolveu-se um software de código aberto chamado Machado, um framework de integração de dados genômicos escrito em Python[10] que permite aos grupos de pesquisa armazenar dados genômicos e que também oferece interfaces para navegação, buscas e visualização. O Machado utiliza a biblioteca BioPython (Cock et al., 2009) que suporta a grande maioria dos formatos de arquivos e programas utilizados na bioinformática. Além disso, o Python vem se consolidando como uma das principais linguagens de programação na área de ciências de dados (Millman; Aivazis, 2011), e o Machado pode também se beneficiar das ferramentas dessa área. Esse framework utiliza o esquema de banco de dados Chado e, portanto, deve ser bastante intuitivo para adoção ou execução em bancos de dados que já existem, pelos atuais desenvolvedores.

O esquema de banco de dados relacional biológico do GMOD, Generic Model Organism Database Project[11], conhecido como Chado (Mungall; Emmert, 2007), é uma das poucas iniciativas de código aberto que obteve relativo sucesso em adoção pela comunidade. Muitos softwares conseguem se conectar a ele, como o Gbrowse (Stein et al., 2002), o Jbrowse (Skinner et al., 2009) e o Apollo (Lee et al., 2013), que são importantes ferramentas para visualização e anotação de genomas. Existem algumas ferramentas para integração de dados que usam o Chado como esquema de banco de dados ou que conseguem extrair os dados desse esquema, porém elas foram desenvolvidas em linguagens de programação pouco utilizadas na bioinformática (Kalderimis et al., 2014; Spoor et al., 2019).

10 Disponível em: https://www.python.org

11 Disponível em: http://www.gmod.org

O Machado possui várias ferramentas de carregamento de dados para dados genômicos e para resultados de análises de softwares conhecidos no meio biológico (BLAST, InterproScan etc.) (Altschul et al., 1990; Quevillon et al., 2005), e sua interface web contém uma poderosa ferramenta de buscas que permite filtrar e ordenar os resultados de forma rápida.

No âmbito do projeto PlantAnnot, foi criada uma ferramenta, utilizando o Machado, denominada *Plant Co-expression Annotation Resource*, para armazenar e disponibilizar esses dados[12]. Essa ferramenta é uma implementação do Machado que serve como exemplo de sua utilidade para pesquisadores que necessitam armazenar e tornar acessível um grande volume de dados genômicos.

Para exemplificar, uma das utilidades do *Plant Co-expression Annotation Resource* é a de possibilitar a navegação pelo genoma de 53 espécies de plantas angiospermas, permitindo a visualização de detalhes sobre genes, proteínas e RNA por meio do navegador de genomas JBrowse. Outra utilidade dessa ferramenta é a de realizar buscas por palavras-chave e uso de filtros. Dessa forma, o usuário consegue realizar buscas simples por genes, proteínas e RNA, pelo uso de palavras de interesse. Mas também poderá agregar à busca filtros mais complexos, produzindo listas de resultados mais específicas, por exemplo um conjunto de proteínas sem função conhecida, candidatas para a criação de plantas transgênicas resistentes à estresses abióticos, como seca, calor, frio, entre outros.

O Machado pretende ser um framework objeto-relacional moderno, que usa os mais recentes módulos Python para produzir um programa de código aberto eficaz para pesquisa genômica, podendo ser um projeto envolvente para novos desenvolvedores, colaboradores e usuários. Para tanto, criamos uma conta corporativa para o LMB no GitHub, que acreditamos ser a primeira conta da Embrapa nessa plataforma[13]. Também foi criada uma versão demonstração do sistema[14].

A ferramenta Machado vai passar por fases de aperfeiçoamento para projetos em andamento na Embrapa, como o projeto "O Hologenoma de Nelore: Implicações na Qualidade de Carne e em Eficiência Alimentar", com foco em melhoramento genômico de bovinos, liderado pela Embrapa Pecuária Sudeste. Esse projeto pretende identificar mecanismos moleculares relacionados à maciez da carne, e, para isso, foram produzidos diversos conjuntos de dados que precisam ser integrados, como genomas, transcriptomas, proteomas, genotipagens, entre outros.

12 Disponível em: https://www.machado.cnptia.embrapa.br/plantannot

13 Disponível em: https://github.com/lmb-embrapa

14 Disponível em: https://www.machado.cnptia.embrapa.br/demo_machado

2.3.2 BDPFG: sistema web para recuperação de informação de pedigree, fenótipos e genótipos

O desenvolvimento de tecnologias de genotipagem em larga escala de marcadores moleculares do tipo *Single Nucleotide Polymorphisms* (SNP) – para estimar o perfil genômico de animais – permitiu tanto o desenvolvimento de estudos de associação genótipo-fenótipo em escala genômica (do inglês *genome-wide association studies* – GWAS) quanto a introdução da tecnologia de seleção genômica em programas de melhoramento genético. As tecnologias atuais para geração de dados moleculares são capazes de realizar a genotipagem de dezenas a centenas de milhares de marcadores SNP, em um único ensaio para cada indivíduo, com enorme velocidade e automação (Caetano, 2009).

Por outro lado, essa conjuntura implica na necessidade de armazenamento de um enorme volume de dados, não somente de genótipos, mas também de fenótipos e pedigree de um número cada vez maior de animais. Dessa forma, realizar o armazenamento adequado e a extração de conhecimento útil a partir dessa quantidade de dados torna-se um grande desafio. Dado o volume de dados armazenado, uma questão importante a se considerar no desenvolvimento de uma solução computacional é a adequabilidade da modelagem do banco de dados à aplicação desejada, pois esta terá impacto direto nos tempos de consulta e escrita em sistemas gerenciadores de bancos de dados relacionais (SGBD), onde essa informação estará armazenada.

Diante disso, com o objetivo de fornecer uma solução que fosse eficiente tanto no armazenamento quanto na integração e na consulta desse alto volume de dados, o sistema Banco de Dados de Pedigree, Fenótipos e Genótipos (BDPFG) foi desenvolvido. O objetivo desse sistema é integrar dados enviados, de vários formatos, para que se possam analisá-los nos softwares de avaliação genética/genômica. O BDPFG foi inicialmente desenvolvido utilizando um diagrama de dados proposto por Higa e Oliveira (2015). Esse diagrama foi redesenhado de forma que possibilitasse a implementação do tipo *JavaScript Object Notation* (JSON). Com a implementação dos tipos JSON e texto em algumas tabelas, foi possível o uso da abordagem *Not Only SQL*[15] (NoSQL) para armazenar parte dos dados, agilizando consultas que necessitariam realizar junções (*joins*) com outras tabelas.

Para o desenvolvimento do sistema, foram escolhidos componentes de tecnologia de informação dentro da filosofia do uso de software livre. O sistema gerenciador de banco de dados escolhido foi o PostgreSQL[16], por ser um SGBD confiável, amplamente utilizado no mercado. Como software

15 Disponível em: http://nosql-database.org

16 Disponível em: https://www.postgresql.org

para controle de versão, foi utilizado o GitLab[17], hospedado na Embrapa. A linguagem de programação escolhida foi Java[18] e seus componentes da tecnologia *Java Enterprise Edition* (Java EE).

Dentre as tecnologias Java EE disponíveis e utilizadas pelo BDPFG destaca-se, entre outras, a estrutura *Java Server Faces* (JSF). A arquitetura do framework JSF emprega o modelo MVC (*Model, View, Controller*), que faz a separação entre as camadas de apresentação e de aplicação. O servidor de aplicação escolhido para abrigar o sistema BDPFG foi o WildFly[19].

O projeto de desenvolvimento do sistema utilizou alguns conceitos do Scrum, que é um framework ágil para a realização de projetos complexos. O Scrum reúne atividades de monitoramento e *feedback*, em geral, por meio de reuniões rápidas e diárias com toda a equipe, procurando identificar e corrigir quaisquer deficiências no processo de desenvolvimento. Além disso, o método Scrum baseia-se em fundamentos como: equipes pequenas, requisitos desconhecidos e iterações curtas, estas denominadas de *sprints* (Schwaber, 2004).

O sistema BDPFG possui muitos recursos implementados e está em processo de homologação pelos usuários. Por meio de sua interface web é possível realizar consultas e importações de dados fenotípicos, genotípicos e de pedigree de diversas espécies de animais. Ao acessá-lo, a página de login será exibida (Figura 1):

Entre suas funcionalidades, destaca-se a visualização dos dados de animais (Figura 2). Nessa tela, o usuário encontra diversas informações sobre o indivíduo, tais como: código identificador do indivíduo, nome original, pai,

Figura 1. Tela de login do sistema BDPFG[20].

17 Disponível em: https://gitlab.com

18 Disponível em: https://www.oracle.com/br/java/

19 Disponível em: http://wildfly.org/downloads/

20 Disponível em: http://www.bdpfg.cnptia.embrapa.br/

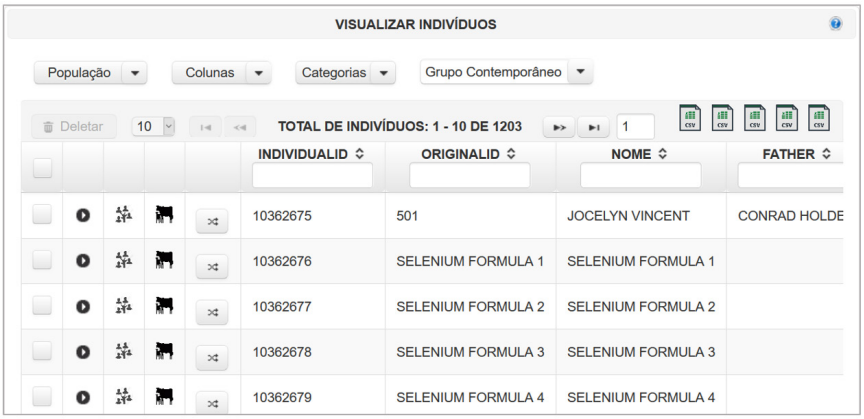

Figura 2. Tela mostrando indivíduos cadastrados no sistema[21].

mãe, data de inserção na população, população e outras informações contidas nas variáveis JSON relativas ao tipo do indivíduo (gado de corte, ave, etc.). Contudo, cabe ressaltar que as variáveis dos fenótipos relacionados às espécies consideradas pelo sistema devem ser previamente registradas, sendo importadas do Sistema de Experimentos da Embrapa – SIEXP (Apolinário et al., 2016), onde foram definidas para a espécie com a qual o usuário trabalhará no seu grupo de usuários (ex: bovinos, suínos, etc.).

É possível também importar dados de arquivos com colunas separadas por tabulações (TSV). Esses arquivos precisam seguir um formato padronizado. Depois de importar os dados, é possível visualizar o pedigree de um animal listado na página de visualização de animais. A janela de pedigree pode ser expandida para facilitar a visualização dos animais e dos seus antepassados.

O banco de dados disponibiliza vários filtros para que o usuário possa conferir os dados que foram carregados e, então, exportar para o formato dos softwares de avaliação. Geralmente, os dados são exportados em formato tabular, para serem analisados no programa R, já que são extensas tabelas com medições de características dos animais. Também é possível exportar os dados desses animais (fenótipos, pedigree) para arquivos no formato CSV e manipulá-los no Excel. Os filtros existentes permitem consultas por população, categoria, nome do animal, nome do pai, nome da mãe. Uma outra ferramenta, talvez a mais importante do sistema, é a de identificação de animais duplicados, possibilitando ao usuário realizar a associação de animais duplicados em um animal apenas.

O sistema BDPFG faz parte de uma solução computacional proposta em outros projetos Embrapa (MaxiDep e MaxiPlat). Esses projetos buscaram aglutinar esforços para estruturação de uma solução computacional (da qual

21 Disponível em: http://www.bdpfg.cnptia.embrapa.br/

o BDPFG é um dos componentes) para suporte à rotina de avaliação genética de programas de melhoramento genético de gado de corte, no escopo do programa Embrapa-Geneplus. Tal esforço compreendeu tanto o desenvolvimento de ativos para suporte à organização dos dados utilizados nas avaliações genéticas (sistema BDPFG) quanto o desenvolvimento de uma solução nacional para a resolução de modelos genético-estatísticos (software brBlup). Dessa forma, o sistema BDPFG faz o "meio de campo" na organização dos dados para que o software brBlup os utilize na geração de modelos genético-estatísticos.

Uma comparação com a busca em outros softwares com interface web desenvolvidos pela Embrapa Informática Agropecuária (Vieira, 2012a, 2012b), com funcionalidade de armazenamento de genótipos e fenótipos e que contemplam consultas básicas a dados moleculares (SNPs), mostra que uma consulta simples em cerca de 800 animais e 700 mil marcadores SNP demorava, pelo menos, uma hora para ser processada nesses outros softwares desenvolvidos. Uma consulta semelhante realizada no banco BDPFG leva menos de um minuto, pois a utilização de campos dos tipos JSON e texto nas tabelas retira parte da normalização necessária do modelo tradicional, agilizando as pesquisas.

3 Considerações finais

As pesquisas relatadas neste capítulo estão em andamento e prosseguirão para outras etapas. Na pesquisa com o tambaqui, com o avanço da produção no futuro próximo, poder-se-á dar início ao melhoramento genético propriamente dito, a exemplo do que já ocorre no exterior. As ferramentas genômicas apresentadas neste capítulo poderão evoluir para ajudar na fase de seleção das matrizes, com o objetivo de melhorar alguma característica de interesse econômico, por exemplo o peso ao abate. Na cadeia produtiva de carne bovina, a seleção genômica já é uma realidade, e os resultados são excelentes. O mesmo pode ocorrer com a cadeia produtiva de peixes. Com a crescente importância da proteína de peixes no cardápio mundial, talvez a região amazônica possa se tornar, em breve, uma grande produtora e, quem sabe, até exportadora de peixes nativos. Há ainda muito caminho a percorrer, mas a Embrapa já deu uma contribuição significativa indicando e abrindo o caminho, e a bioinformática desempenha um papel fundamental.

A validação de uma metodologia que inclui a identificação de antígenos por meio de um *pipeline* de vacinologia reversa e a obtenção de uma vacina multiepítopos está em andamento na Embrapa, com a participação do LMB, e tem como objetivo o controle do carrapato bovino. A infestação de rebanhos bovinos por esse parasita é considerada, hoje, um dos problemas mais

importantes na pecuária em termos econômicos, atingindo todos os países de clima tropical e subtropical. Só no Brasil, as perdas anuais devidas à infestação pelo carrapato são da ordem de US$3,24 bilhões (Grisi et al., 2014). A obtenção de uma vacina eficaz certamente irá contribuir para o controle do parasita, reduzindo as aplicações de acaricidas e o prejuízo ambiental e econômico decorrentes dessa prática. Ainda, uma vez validada, são várias as aplicações possíveis da metodologia, incluindo a identificação de alvos para o controle de outros problemas de interesse da agropecuária envolvendo sanidade e bem-estar animal.

A ferramenta Machado vai atender outros projetos em andamento na Embrapa. Já existe programação para seu uso no projeto Genômica Aplicada à Otimização de Programas de Melhoramento Genético de Espécies Forrageiras Tropicais, liderado pela Embrapa Cerrados, com foco em melhoramento de plantas forrageiras. Nesse projeto, está previsto o sequenciamento de genomas de referência para seis espécies forrageiras tropicais, com a caracterização de conjuntos amplos de variantes genômicas, e espera-se usar o Machado como base para a implementação de um portal de acesso aos dados genômicos gerados.

O banco de dados BDPFG está sendo estruturado de forma a permitir sua utilização em outras coleções de dados, com algumas alterações específicas para cada projeto.

Como mostrado nas pesquisas aqui relatadas, a bioinformática tornou-se fundamental e será ainda mais importante nas agendas de inovação em direção à transformação digital da agricultura. A existência de estruturas multiusuários para atender projetos de pesquisa que não possuem a estrutura necessária para análises complexas é fundamental, possibilitando, ainda, melhor uso de recursos. Com a dependência da bioinformática da disponibilidade de uma equipe especialista e de infraestrutura adequada, o gerenciamento da estrutura que atende os projetos de pesquisa deve estar com a atenção voltada para manter ambos os aspectos atualizados.

4 Referências

AHMAD, T. A.; EWEIDA, A. E.; SHEWEITA, S. A. B-cell epitope mapping for the design of vaccines and effective diagnostics. **Trials in Vaccinology**, v. 5, p. 71-83, 2016. DOI: 10.1016/j.trivac.2016.04.003.

ALTSCHUL, S. F.; GISH, W.; MILLER, W.; MYERS, E. W.; LIPMAN, D. J. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215, n. 3, p. 403-410, Oct 1990. DOI: 10.1016/S0022-2836(05)80360-2.

ANDREOTTI, R. Performance of two Bm86 antigen vaccine formulation against tick using crossbreed bovines in stall test. **Revista Brasileira de Parasitologia Veterinária**, v.15, p. 97-100, 2006.

ANDREOTTI, R.; GIACHETTO, P. F.; CUNHA, R. C. Advances in tick vaccinology in Brazil: from gene expression to immunoprotection. **Frontiers in Biosciences**, v. 10, p. 127-42, Jan 2018. DOI: 10.2741/s504.

ANDREWS, S. M.; POLLARD, A. J. A vaccine against serogroup B Neisseria meningitidis: dealing with uncertainty. **The Lancet Infectious Diseases**, v. 14, n. 5, p. 426-434, May 2014. DOI: 10.1016/s1473-3099(13)70341-4.

APOLINÁRIO, D. R. de F.; QUEIROS, L. R.; VACARI, I.; CRUZ, S. A. B. da. **SIExp – Sistema de Informação de Experimentos da Embrapa**. Versão v. 1.7.6. Campinas: Embrapa Informática Agropecuária, 2016.

BOLGER, A. M.; LOHSE, M.; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114-2120, Aug 2014. DOI: 10.1093/bioinformatics/btu170.

CAETANO, A. R. Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. **Revista Brasileira de Zootecnia**, v. 38, p. 64-71, 2009. Número especial. DOI: 10.1590/s1516-35982009001300008.

COCK, P. J. A.; ANTAO, T.; CHANG, J. T.; CHAPMAN, B. A.; COX, C. J.; DALKE, A.; FRIEDBERG, I.; HAMELRYCK, T.; KAUFF, F.; WILCZYNSKI, B.; DE HOON, M. J. Biopython: freely available Python tools for computational molecular biology and bioinformatics. **Bioinformatics**, v. 25, n. 11, p. 1422-1423, June 2009. DOI: 10.1093/bioinformatics/btp163.

CONESA, A.; MADRIGAL, P.; TARAZONA, S.; GOMEZ-CABRERO, D.; CERVERA, A.; MCPHERSON, A.; SZCZEŚNIAK, M. W.; GAFFNEY, D. J.; ELO, L. L.; ZHANG, X.; MORTAZAVI, A. A survey of best practices for RNA-seq data analysis. **Genome Biology**, v. 17, article number 13, 2016. DOI: 10.1186/s13059-016-0881-8.

EWELS, P.; MAGNUSSON, M.; LUNDIN, S.; KÄLLER, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. **Bioinformatics**, v. 32, n. 19, p. 3047-3048, Oct 2016. DOI: 10.1093/bioinformatics/btw354.

FastQC-GitHub. Disponível em: https://github.com/s-andrews/FastQC/releases. Acesso em: 7 maio 2020.

FastX-Github. Disponível em: https://github.com/agordon/fastx_toolkit. Acesso em: 7 maio 2020.

GIACHETTO, P. F.; CUNHA, R. C.; NHANI JUNIOR, A.; GARCIA, M. V.; FERRO, J. A.; ANDREOTTI, R. Gene expression in the salivary gland of Rhipicephalus (Boophilus) microplus fed on tick-susceptible and tick-resistant hosts. **Frontiers in Cellular and Infection Microbiology**, v. 9, p. 477, Jan 2020. DOI: 10.3389/fcimb.2019.00477.

GRABHERR, M. G.; HAAS, B. J.; YASSOUR, M.; LEVIN, J. Z.; THOMPSON, D. A.; AMIT, I.; ADICONIS, X.; FAN, L.; RAYCHOWDHURY, R.; ZENG, Q.; CHEN, Z.; MAUCELI, E.; HACOHEN, N.; GNIRKE, A.; RHIND, N.; DI PALMA, F.; BIRREN, B. W.; NUSBAUM, C.; LINDBLAD-TOH, K.; FRIEDMAN, N.; REGEV, A. Full-length transcriptome assembly from RNA-seq data without a reference genome. **Nature Biotechnology**, v. 29, n. 7, p. 644-652, 2011. DOI: 10.1038/nbt.1883.

GRISI, L.; LEITE, R. C.; MARTINS, J. R. de S.; BARROS, A. T. M. de; ANDREOTTI, R.; CANÇADO, P. H. D.; LEÓN, A. A. P. de; PEREIRA, J. B.; VILLELA, H. S. Reassessment of the potential economic impact of cattle parasites in Brazil. **Revista Brasileira de Parasitologia Veterinária**, v. 23, n. 2, p. 150-156, Apr/June 2014. DOI: 10.1590/S1984-29612014042.

GUREVICH, A.; SAVELIEV, V.; VYAHHI, N.; TESLER, G. QUAST: quality assessment tool for genome assemblies. **Bioinformatics**, v. 29, n. 8, p. 1072-1075, Apr 2013. DOI: 10.1093/bioinformatics/btt086.

HIGA, R. H.; OLIVEIRA, G. B. **Banco de Dados de Genótipos e Fenótipos (BDGF) para suporte a estudos de associação genômica ampla e seleção genômica em programas de melhoramento animal**. Campinas: Embrapa Informática Agropecuária, 2015. 30 p. (Embrapa Informática Agropecuária. Documentos, 133). Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/138127/1/Doc133.pdf. Acesso em: 7 maio 2020.

HUBER, W.; CAREY, V. J.; GENTLEMAN, R.; ANDERS, S.; CARLSON, M.; CARVALHO, B. S.; BRAVO, H. C.; DAVIS, S.; GATTO L.; GIRKE, T.; GOTTARDO, R.; HAHNE, F.; HANSEN, KD.; IRIZARRY, R. A.; LAWRENCE, M.; LOVE, M. I.; MACDONALD, J.; OBENCHAIN, V.; OLE'S; A. K.; PAG'ES, H.; REYES, A.; SHANNON, P.; SMYTH, G.K; TENENBAUM, D.; WALDRON, L.; MORGAN, M. Orchestrating high-throughput genomic analysis with Bioconductor. **Nature Methods**, v. 12, n. 2, p. 115-121, Jan 2015. DOI: 10.1038/nmeth.3252.

IANELLA, P.; YAMAGISHI, M. E. B.; VARELA, E. S.; VILLELA, L. C. V.; PAIVA, S. R.; CAETANO, A. R. Tambaqui (Colossoma macropomum) single nucleotide polymorphism discovery by reduced representation library deep sequencing. In: AQUACULTURE, 2019, New Orleans. **Abstracts**. [S.l.: s.n.], 2019. p. 491. Disponível em: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/208846/1/CNPASA-2019-Aqua2.pdf. Acesso em: 7 maio 2020.

KALDERIMIS, A.; LYNE, R.; BUTANO, D.; CONTRINO, S.; LYNE, M.; HEIMBACH, J.; HU, F.; SMITH, R.; ŠTĚPÁN, R.; SULLIVAN, J.; MICKLEM, G. InterMine: extensive web services for modern biology. **Nucleic Acids Research**, v. 42, n. W1, p. W468-W472, July 2014. DOI: 10.1093/nar/gku301.

KAO, D. J.; HODGES, R. S. Advantages of a synthetic peptide immunogen over a protein immunogen in the development of an anti-pilus vaccine for Pseudomonas aeruginosa. **Chemical Biology & Drug Design**, v. 74, p. 33-42, 2009. DOI: 10.1111/j.1747-0285.2009.00825.x.

LAMARRE, S.; FRASSE, P.; ZOUINE, M.; LABOURDETTE, D.; SAINDERICHIN, E.; HU, G.; LE BERRE-ANTON, V.; BOUZAYEN, M.; MAZA, E. Optimization of an RNA-Seq differential gene expression analysis depending on biological replicate number and library size. **Frontiers in Plant Science**, v. 9, article 108, Feb 2018. DOI: 10.3389/fpls.2018.00108.

LEE, E.; HELT, G. A.; REESE, J. T.; MUNOZ-TORRES, M. C.; CHILDERS, C. P.; BUELS, R. M.; STEIN, L.; HOLMES, I.H; ELSIK, C.G.; LEWIS, S.E. Web Apollo: a web-based genomic annotation editing platform. **Genome Biology**, v. 14, n. 8, article number R93, Aug 2013. DOI: 10.1186/gb-2013-14-8-r93.

LI, B.; DEWEY, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. **BMC Bioinformatics**, v. 12, n. 1, article number 323, Aug 2011. DOI: 10.1186/1471-2105-12-323.

MILLMAN, K. J.; AIVAZIS, M. Python for scientists and engineers. **Computing in Science & Engineering**, v. 13, n. 2, p. 9-12, Mar 2011. DOI: 10.1109/MCSE.2011.36.

MUNGALL, C. J.; EMMERT, D. B. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. **Bioinformatics**, v. 23, n. 13, p. i337-i346, July 2007. DOI: 10.1093/bioinformatics/btm189.

NAPIER, J. A.; HASLAM, R. P.; TSALAVOUTA, M.; SAYANOVA, O. The challenges of delivering genetically modified crops with nutritional enhancement traits. **Nature Plants**, v. 5, n. 6, p. 563-567, June 2019. DOI: 10.1038/s41477-019-0430-z.

PALATNIK-DE-SOUSA, C. B.; SOARES, I. da S.; ROSA, D. S. Epitope discovery and synthetic vaccine design. **Frontiers in Immunology**, v. 9, p. 826, 2018. DOI: 10.3389/978-2-88945-522-5.

PRADO, J. R.; SEGERS, G.; VOELKER, T.; CARSON, D.; DOBERT, R.; PHILLIPS, J.; COOK, K.; CORNEJO, C.; MONKEN, J.; GRAPES, L.; REYNOLDS, T.; MARTINO-CATT, S. Genetically engineered crops: from idea to product. **Annual Reviews of Plant Biology**, v. 65, n. 1, p. 769-790, Apr 2014. DOI: 10.1146/annurev-arplant-050213-040039.

QUEVILLON, E.; SILVENTOINEN, V.; PILLAI, S.; HARTE, N.; MULDER, N.; APWEILER, R.; LOPEZ, R. InterProScan: protein domains identifier. **Nucleic Acids Research**, v. 33, p. W116-W120, July 2005. Issue suppl_2. DOI: 10.1093/nar/gki442.

R CORE TEAM. **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2020. Disponível em: https://www.R-project.org. Acesso em: 7 maio 2020.

RAPPUOLI, R. Reverse vaccinology. **Current Opinion in Microbiology**, v. 3, n. 5, p. 445-450, Oct 2000. DOI: 10.1016/s1369-5274(00)00119-3.

RAPPUOLI, R.; COVACCI, A. Reverse vaccinology and genomics. **Science**, v. 302, n. 5645, p. 602, Oct 2003. DOI: 10.1126/science.1092329.

ROBINSON, M. D.; MCCARTHY, D. J.; SMYTH, G. K. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics**, v. 26, n. 1, p. 139-140, Jan 2010. DOI: 10.1093/bioinformatics/btp616.

SCHWABER, K. **Agile project management with scrum**. United States: Microsoft Press, 2004. 163 p.

SEPPEY, M.; MANNI, M.; ZDOBNOV, E. M. BUSCO: assessing genome assembly and annotation completeness. In: KOLLMAR, M. (ed.). Gene prediction. New York: Humana, 2019. p. 227-245. (Methods in molecular biology, v. 1962). DOI: 10.1007/978-1-4939-9173-0_14.

SKINNER, M. E.; UZILOV, A. V.; STEIN, L. D.; MUNGALL, C. J.; HOLMES, I. H. JBrowse: a next-generation genome browser. **Genome Research**, v. 19, n. 9, p. 1630-1638, 2009. DOI: 10.1101/gr.094607.109.

SPOOR, S.; CHENG, C. H.; SANDERSON, L. A.; CONDON, B.; ALMSAEED, A.; CHEN, M.; BRETAUDEAU, A.; RASCHE, H.; JUNG, S.; MAIN, D.; BETT, K.; STATON, M.; WEGRZYN, J. L.; FELTUS, F. A.; FICKLIN, S. P. Tripal v3: an ontology-based toolkit for construction of FAIR biological community databases. **Database**, v. 2019, 2019. DOI: 10.1093/database/baz077.

STEIN, L. D.; MUNGALL, C.; SHU, S.; CAUDY, M.; MANGONE M.; DAY, A.; NICKERSON, E.; STAJICH, J. E; HARRIS, T. W.; ARVA, A.; LEWIS, S. The generic genome browser: a building block for a model organism system database. **Genome Research**, n. 516, p. 1599-1610, 2002. DOI: 10.1101/gr.403602.

THE UNIPROT CONSORTIUM. UniProt: a worldwide hub of protein knowledge. **Nucleic Acids Research**, v. 47, p. D506-D515, Jan 2019. Issue D1. DOI: 10.1093/nar/gky1049.

TRIMGALORE-GITHUB. Disponível em: https://github.com/FelixKrueger/TrimGalore. Acesso em: 7 maio 2020.

WIKIPEDIA. **List of RNA-Seq bioinformatics tools**. 2020. Disponível em: https://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools. Acesso em: 07 maio 2020.

VERNIKOS, G. S. Genome watch: overtake in reverse gear. **Nature Reviews Microbiology**, v. 6, n. 5, p. 334-335, 2008. DOI: 10.1038/nrmicro1898.

VIEIRA, F. D. **Sistema Bife de Qualidade**. Versão 1.6. Campinas: Embrapa Informática Agropecuária, 2012a. 1 CD-ROM.

VIEIRA, F. D. **Sistema Suínos**. Versão 1.1. Campinas: Embrapa Informática Agropecuária, 2012b. 1 CD-ROM.