UNIVERSIDADE FEDERAL DE VIÇOSA
CAMPUS FLORESTAL

Vinícius Cauê Furlan Roberto

Map Reduce Cluster and Annotation Tool for Rapid Analysis
Um ambiente para anotação e enriquecimento de dados biológicos

FLORESTAL - MINAS GERAIS
2019

Vinícius Cauê Furlan Roberto

Map Reduce Cluster and Annotation Tool for Rapid Analysis
Um ambiente para anotação e enriquecimento de dados biológicos

Monografia, apresentada ao Curso de Ciência da Computação da Universidade Federal de Viçosa como requisito para obtenção do título de bacharel em Ciência da Computação.

Orientador: Eduardo Martin Tarazona Santos

FLORESTAL - MINAS GERAIS
2019

Vinícius Cauê Furlan Roberto

Map Reduce Cluster and Annotation tool for Rapid Analysis
Um ambiente para anotação e enriquecimento de dados biológicos

Monografia, apresentada ao Curso de Ciência da Computação da Universidade Federal de Viçosa como requisito para obtenção do título de bacharel em Ciência da Computação.

Eduardo Martin Tarazona Santos

Avaliador 1

Avaliador 2

FLORESTAL - MINAS GERAIS,

DEDICATÓRIA

Primeiro eu gostaria de dedicar e agradecer ao universo, que apesar de vasto e de se comportar de forma estocástica me colocou nos lugares certos, na hora exata.
Agradeço também aos meus pais (Luís Carlos e Sueli Furlan), minha mulher (Marla Mendes de Aquino), minha avó (Maria Aparecida) e meus irmãos (Lucas Furlan e Nicolly Furlan) por todo apoio, carinho, dedicação e paciência durante todos estes anos de graduação. Aos meus orientadores Eduardo Martin Tarazona Santos e Thiago Peixoto Leal, em primeiro lugar pelo conhecimento agregado durante todo este tempo que tenho trabalhado no laboratório de Diversidade Genética Humana da UFMG. Em segundo lugar por confiarem à mim tarefas importantes, me dando assim oportunidades de mostrar todo conhecimento que adquiri durante este tempo. Além disso, agradeço ao pessoal que trabalha comigo no laboratório (Camila, Carol, Ricardo e Victor) pelas conversas, discussões, dicas e bons momentos proporcionados. Agradeço a todos os meus professosres da graduação (José Augusto Nacif, Daniel Mendes Barbosa, Fabrício Aguiar Silva, Thais Regina de Moura Braga Silva e Glaucia Braga e Silva) pela paciência, ensinamentos e todo o conhecimento agregado a partir das aulas e puxões de orelha. Por fim eu gostaria de agradecer a Universidade Federal de Viçosa e a Universidade Federal de Minas Gerais por proporcionarem educação pública gratuita de qualidade e oportunidades a qualquer um disposto a aprender e ensinar.

RESUMO

O desenvolvimento de técnicas de sequenciamento de nova geração trouxeram um grande volume de dados biológicos e com isso, novos desafios computacionais relativos ao processamento desses dados. Dentro deste contexto o Laboratório de Diversidade Genética Humana desenvolveu uma ferramenta que agrupa informação de 11 bancos de dados públicos de informações biológicas. Apesar de ter obtido bons resultados nessa integração, a forma como a ferramenta foi implementada não era a mais eficiente, o que levou a um gasto de recurso computacional desnecessário e demorado no retorno da análise. Neste trabalho nós realizamos a reengenharia da ferramenta num ambiente paralelo e distribuído através de um cluster de computadores, utilizando o framework Apache Hadoop, com o objetivo de apoiar o processamento deste grande volume de dados. Os primeiros testes mostraram que a nova ferramenta foi bem sucedida, gastando consideravelmente menos memória e tempo nos testes mais pesados. Além disso, uma nova ferramenta de anotação de SNPs baseada neste novo ambiente é proposta.

Palavras-chave:
Cluster; Hadoop; Paralelo; Distribuido; Big Data; Anotação; SNPs; Bioinformática;

ABSTRACT

The development of new generation sequencing techniques has brought a large volume of biological data and new computational challenges related to it. Within this context the Human Genetic Diversity Laboratory has developed a tool that gathers information from 11 public biological databases. Despite having been successful in this integration, the way the tool was implemented was not the most efficient, which led to an unnecessary expense consuming computational resource and time in the return of the analysis. In this work we re-engineered the tool in a distributed and parallel environment through a computer cluster, using the Apache Hadoop framework, with the objective of supporting the processing of this large data volume. Early tests showed that the new tool was successful, spending considerably less memory and time on the heavier tests. In addition, a new SNP annotation tool is also proposed in the context of this new environment.

Keywords:
Cluster; Hadoop; Parallel; Distributed; Big Data; Annotation; SNPs; Bioinformatics;

SUMÁRIO

1 Introdução **8**
1.1 Motivação 8
1.2 O crescimento do volume de dados de origem biológica 9
1.3 Mutações genéticas 11
1.4 SNPs e anotação de variantes genéticas 13
1.5 Objetivo 14
1.5.1 Objetivo geral 14
1.5.2 Objetivos específicos 14

2 Metodologia **15**
2.1 Multi Agent System for SNP Annotation 15
2.2 MapReduce Cluster 19
2.3 Annotation Tool for Rapid Analysis 22

3 Resultados **26**
3.1 Multi Agent System for SNP Annotation 26
3.2 Cluster Annotation Tool 27
3.3 Comparação entre performance e anotação das ferramentas 29
3.3.1 Performance 29
3.3.2 Anotação de variantes genéticas 30

4 Conclusão **33**

5 Bibliografia **34**

1 Introdução

1.1 Motivação

Sistemas voltados para análises bioinformáticas tem se tornado cada vez mais complexos, exigindo cada vez mais poder computacional e conhecimento por parte de quem está operando e desenvolvendo. Na era do big data a bioinformática passa pelos mesmos problemas encontrados em qualquer outra área que trabalhe com dados de grandes dimensões, no qual existem mais dados sendo gerados por intervalo de tempo do que análises sendo concluídas sobre esses dados (Trelles, O., Prins, P., Snir, M. et al., 2011).

Ferramentas bioinformáticas podem tornar-se obsoletas ou não conseguir lidar mais com o volume de dados fornecidos pelos usuários, o que exige atualizações e modificações para resolver os problemas. Uma abordagem comumente utilizada para resolver a inviabilidade do uso de uma metodologia devido ao crescimento dos dados fornecidos pelo usuário é afrouxar as restrições do modelo como acontece em programas bastantes conhecidos por bioinformatas como o shapeit2 (Reich et al., 2012), shapeit3 (Delaneau Marchini et al., 2016) e admixture (Alexander, Novembre et al., 2009), tornando o modelo mais propenso a erros, porém retornando resultados em tempo viável.

O Laboratório de Diversidade Genética Humana (LDGH) da UFMG trabalha para criar técnicas e ferramentas capazes de contornar estes problemas e responder questões relacionadas a genética de populações humanas, como miscigenação e ancestralidade. Na tentativa de melhorar a performance da principal ferramenta de estudos de variantes genéticas desenvolvida no laboratório (MASSA) foi proposto a reengenharia dos bancos de dados e sua implementação. Para isso propomos a utilização do framework Apache Hadoop (White, 2012) e de um ecossistema para operações com grandes conjuntos de dados de forma distribuída, de forma a manter os bancos de dados utilizados pela aplicação sempre atualizados.

1.2 O crescimento do volume de dados de origem biológica

O desenvolvimento e uso massivo de novas tecnologias, fez com que enormes quantidades de dados passassem a ser gerados. Nos anos de 2006 a 2010, por exemplo, o volume de dados computacionais gerados cresceu de 166 Exabytes para 988 Exabytes (Gantz, John and Reinsel, 2012). Na área da genética o mesmo aconteceu. O fim do Projeto Genoma Humano (International Human Genome Sequencing Consortium, 2003) culminou num aumento dos investimentos e pesquisas em áreas relacionadas a genética e biologia molecular. Com isso observou-se um aumento exponencial da quantidade de dados gerados a partir do sequenciamento de dados biológicos, diminuindo assim o custo do sequenciamento por megabase por consequência de um genoma como um todo (Mardis, 2011). No início do projeto genoma humano o custo para o sequenciamento de 1 Megabase era de aproximadamente $100 dólares (Figura 1). Hoje em dia é possível sequenciar está mesma megabase por apenas $0,01 dólar (Behjatti Tarpey, 2015). O Projeto genoma humano levou 10 anos para ser concluído e foi investido um montante total de $100 bilhões de dólares para que um genoma humano fosse sequenciado por inteiro. Hoje em dia, esse mesmo genoma pode ser sequenciado por "apenas" $1000 dólares e levaria apenas um dia para essa tarefa ser concluída (Behjatti Tarpey, 2015).

Em contrapartida à queda dos preços de sequenciamento, nos anos seguintes foi observado um aumento no número de genomas sequenciados e consequentemente do volume de dados gerados a partir deste tipo de análise (Figura 2).

Estima-se que até 2025 os dados gerados a partir de sequenciamento genético somem ao todo mais de 1 Exabyte. Com o aumento repentino do volume de dados, a bioinformática mostrou-se uma importante aliada nas análises e pesquisas feitas nas áreas relacionadas. Tanto ao garantir que estes dados sejam processados da melhor forma, quanto para encontrar a melhor forma de armazenar tamanho volume de informações. Novas técnicas, abordagens e ferramentas surgem a todo momento com o objetivo de organizar e enriquecer dados de origem biológica. Softwares como as ferramentas de anotação de variantes genéticas. Cujo objetivo é agregar informações valiosas sobre variantes, genes, doenças, entre outras informações de grande valor para o biologia molecular, genética e outras áreas relacionadas.

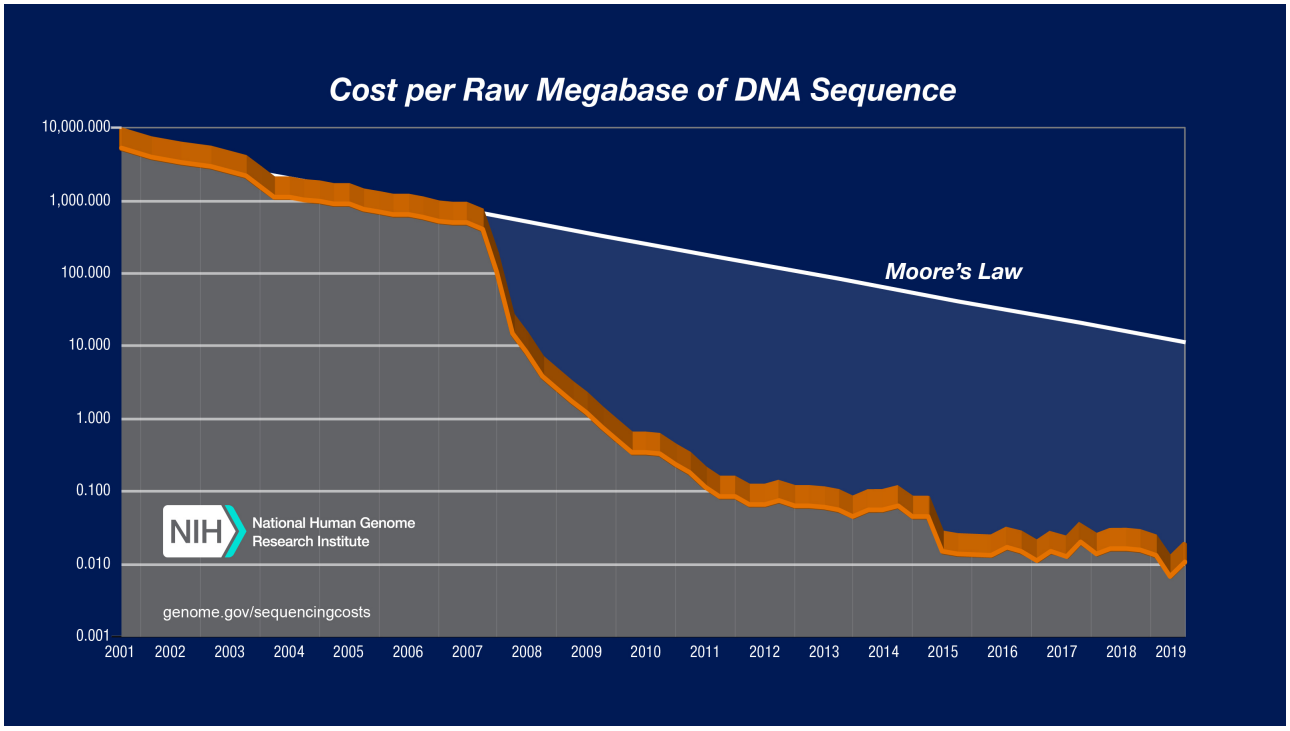

Figura 1 – Custo (em dólares) do sequenciamento por megabase. Retirado de: www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

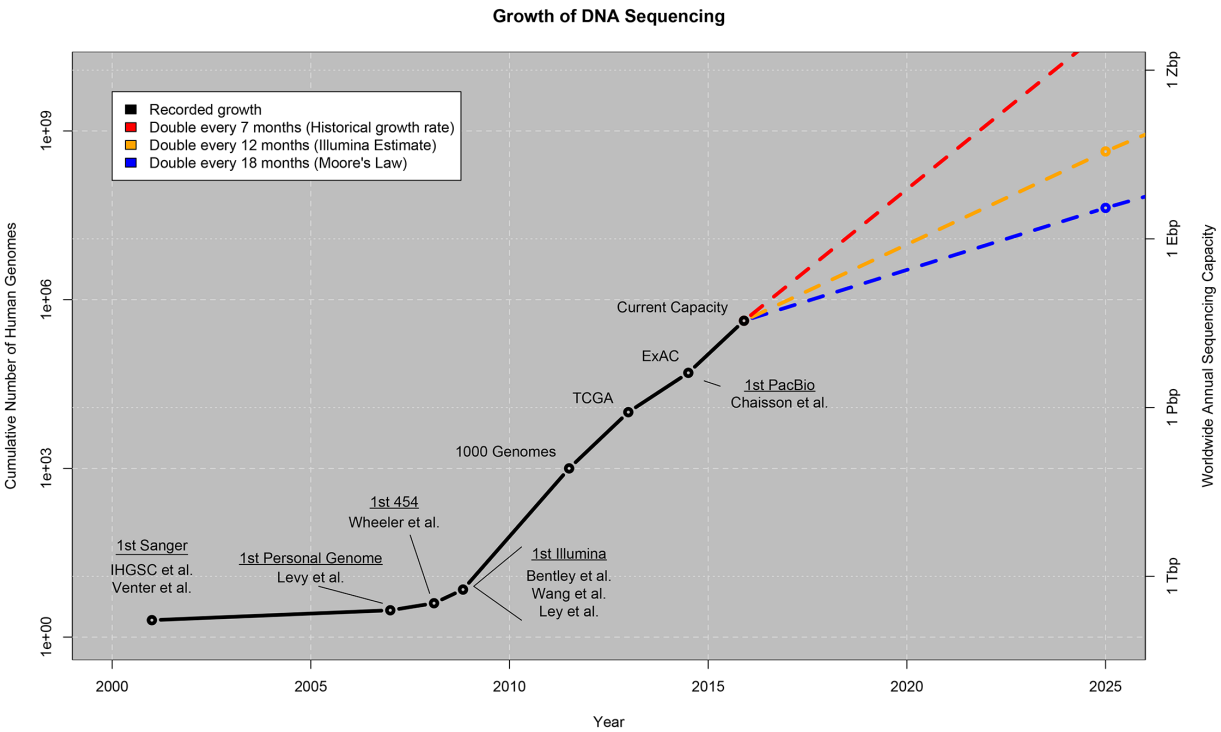

Figura 2 – Aumento do número e do volume de dados gerados a partir do sequenciamento de genomas. Retirado de: DOI:10.1371/journal.pbio.1002195

1.3 Mutações genéticas

Mutações podem ser definidas como mudanças nas sequências de nucleotídeos do material genético de um indivíduo. Mutações podem ser explicadas através do dogma central da biologia. O dogma central da biologia descreve o fluxo de informação dentro da célula. A informação está contida dentro de uma molécula de DNA, que por sua vez pode ser transcrita em RNA mensageiro que então é traduzido em um aminoácido através de outros processos celulares (Figura 3).

Outro possível caminho que informação pode tomar é o da replicação. No processo de replicação do DNA, podem ocorrer erros. Em sua maioria, os erros são rapidamente removidos e corrigidos por uma série de enzimas do sistema de reparo do DNA, porém alguns erros ainda podem persistir e estes são denominados mutações. As Mutações podem ocorrer de forma induzida, quando existe exposição do indivíduo à um agente mutagênico, ou de forma natural, ocorrendo durante a fase de replicação do DNA. Essas mutações podem ser classificadas de duas maneira: 1) Mutações sinônimas, sendo estas mutações pontuais onde a mudança de um nucleotídeo não afeta o aminoácido resultante no processo de transcrição e tradução. 2) Mutações não-sinônimas, são mutações onde a mudança do nucleotídeo pode acarretar na mudança do aminoácido produzido. Podendo resultar ou não na mudança da estrutura e/ou função de uma célula. Um exemplo bastante fácil de entender como uma mutação pode resultar em mudanças na estrutura de uma célula pode ser visto observando uma doença denominada Anemia Falciforme. Na anemia falciforme a mudança de um único nucleotídeo, resulta na troca do aminoácido glutamina para o aminoácido valina (Figura 4)

Causando a alteração da estrutura da hemoglobina e resultando assim na alteração da função da mesma. Portadores de anemia falciforme possuem dificuldades na oxigenação dos tecidos do corpo humano como um dos resultados dessa alteração causada pela mutação. Além disso, esta mudança pontual de nucleotídeo único é denominada SNP (Single Nucleotide Polymorphism ou Polimorfismo de nucleotídeo único). Estas variações devem ocorrer em no mínimo 1% de uma determinada população para ser classificada como um SNP.

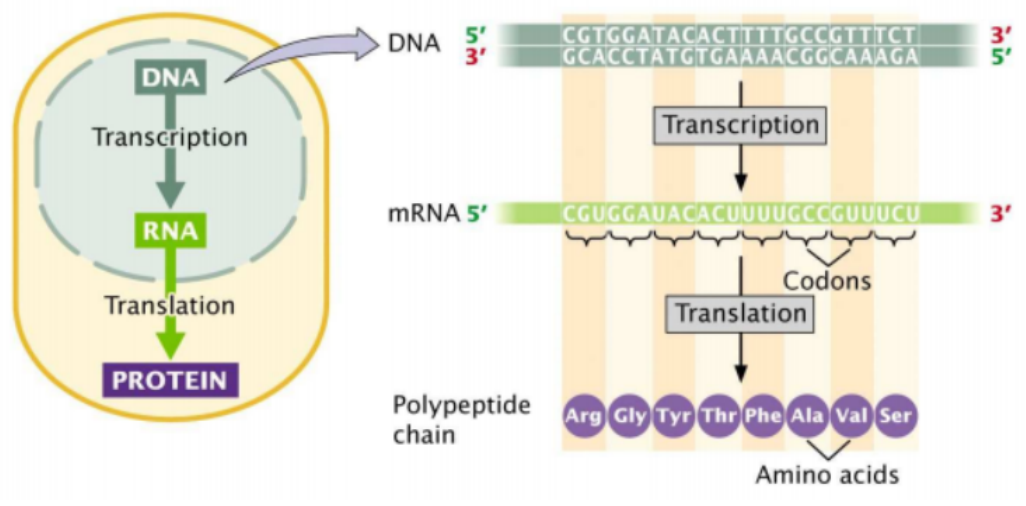

Figura 3 – Descrição do fluxo de informação dentro de uma célula. Retirado de: khanacademy.org

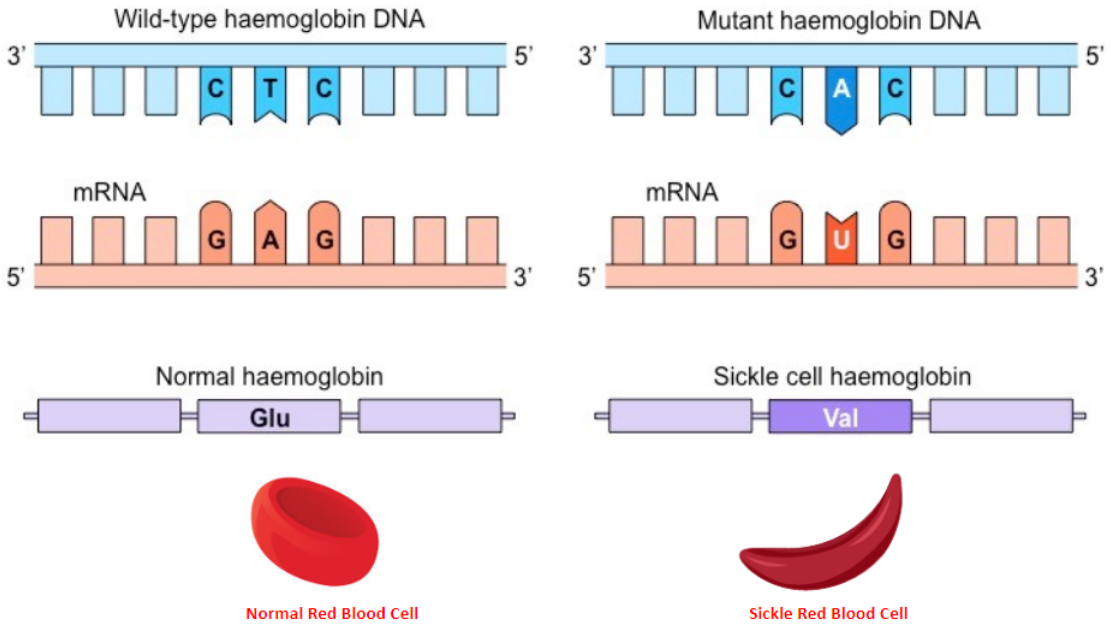

Figura 4 – Mutação causadora da anemia falciforme.

1.4 SNPs e anotação de variantes genéticas

SNPs (Single Nucleotide Polymorphisms) são variações pontuais no genoma que muitas vezes podem fornecer informações relevantes sobre o funcionamento de uma determinada região do mesmo. Um SNP pode fornecer informações como a interação com fármacos e fenótipos, regulação de alguns genes até servir como biomarcadores para pesquisas envolvendo o genoma humano. A localização desses biomarcadores pode ser extremamente importante em termos de previsão de significado funcional, mapeamento genético e genética de populações (Shen, Carlson Tarczy-Hornoch, 2009). Com o grande número de SNPs no genoma humano há a necessidade de priorizar cada um desses SNPs de acordo com seu efeito potencial, a fim de agilizar a genotipagem e análises (Capriotti et al., 2012). O processo de anotação de variantes genéticas se baseia em agregar o máximo de informações relacionadas a um conjunto de polimorfismos ou genes. Anotar um grande número de variantes é um processo difícil e complexo que requer métodos computacionais sofisticados para lidar muitas vezes com um grande volume de dados. Atualmente existem 3 formas principais de anotação e enriquecimento de variantes, cada uma delas com um contexto e objetivos próprios: (i) anotação baseada em genes, (ii) anotação baseada em gene knowledge e (iii) anotação funcional. Na anotação baseada em genes as informações de um gene conhecido são usadas como referência para indicar se a variante observada reside em um gene ou próximo a ele e se tem o potencial de interromper a sequência de proteínas, modificar a estrutura da proteína e sua função. A anotação baseada em genes é baseada no fato de que mutações não-sinônimas (aquelas que a alteração do nucleotídeo no DNA é repassada ao mRNA que, posteriormente, acarretará numa modificação do aminoácido a ser incorporado na proteína) alteram a sequência de uma proteína (M. J. Li Wang, 2015). A anotação baseada em gene knowledge é feita com base nas informações do atributo do gene, função da proteína e seu metabolismo. Nesse tipo de anotação é dada mais ênfase à variação genética que interrompe o domínio da função da proteína, a interação proteína-proteína e a via biológica. A anotação funcional identifica principalmente a variante com base nas informações sobre se os locais variantes estão na região funcional que abriga sinais genéticos ou epigenéticos conhecidos. A função das variantes não codificantes é extensa em termos da região genômica afetada e envolve quase todos os processos de regulação de genes do nível transcricional ao pós-traducional (Sauna Kimchi-Sarfaty, 2011).

1.5 Objetivo

1.5.1 Objetivo geral

Este trabalho tem como objetivo testar a performance da atual ferramenta de anotação de variantes genéticas do Laboratório de Diversidade Genética Humana (LDGH) - Multi Agent System for SNP Annotation (MASSA) - e, caso necessário implementar novas abordagens a fim de resolver problemas de performance relativos à ferramenta alem disso este trabalho visa implementar uma política de atualização para os bancos de dados utilizados pela aplicação.

1.5.2 Objetivos específicos

1. Avaliar se o uso de multi-agentes no emprego do paralelismo é adequado para o problema abordado.
2. Testar e implementar o uso do ecossistema hadoop para operações com grandes conjuntos de dados.
3. Implementar uma ferramenta de anotação de variantes genéticas equivalente ao MASSA porém que contorne os problemas da ferramenta.

2 Metodologia

2.1 Multi Agent System for SNP Annotation

A ferramenta de anotação de SNPs MASSA (Multi Agent System for SNP Annotation) foi implementada na linguagem Java utilizando o framework JADE (Bellifemine, Poggi, Rimassa, 1999), um framework orientado a agentes que seguem o padrão FIPA (Foundation For Intelligent, Physical Agents). O framework foi utilizado com o intuito de paralelizar as consultas feitas pelo MASSA de forma a acelerar o processo de anotação das variantes. A aplicação possui três tipos abstratos de agentes (Figura 2.1), sendo estes:O (i) agente de interface, (ii) agente coordenador e (iii) agente de banco de dados (Figura 5). O agente de interface gerencia as atividades de entrada e saída de dados. O agente coordenador é responsável por coordenar o processo de anotação de SNPs e genes, delegando tarefas aos agentes DB e combinando os resultados. O agente DB encapsula o acesso aos bancos de dados, retornando destes atributos relacionados a genes e SNPs (Souza, 2014). Cada um dos agentes implementados possui como objetivo representar uma abstração para uma camada de paralelização de forma que cada um dos agentes realize tarefas de forma independente, aumentando a velocidade na qual as anotações são feitas.

Figura 5 - Arquitetura do MASSA (Multi Agent System for SNP Annotation)

O MASSA realiza os testes de enriquecimento para os seguintes atributos biológicos: Doenças (OMIM e PharmGKB), vias metabólicas (PharmGKB e Reactome), fármacos (PharmGKB), processo biológico, função molecular e componente celular (GO), banda citogenética (OMIM) e família gênica (HGNC). Esses bancos de dados públicos são de grande interesse à área da genética de populações, farmacogenômica e epidemiologia genética (Figura 6). A seguir apresentaremos uma breve descrição de cada um dos bancos

MASSA realiza os testes de enriquecimento para os seguintes atributos biológicos: Doenças (OMIM e PharmGKB), vias metabólicas (PharmGKB e Reactome), fármacos (PharmGKB), processo biológico, função molecular e componente celular (GO), banda citogenética (OMIM) e família gênica (HGNC). Esses bancos de dados públicos são de grande interesse à área da genética de populações, farmacogenômica e epidemiologia genética. A seguir apresentaremos uma breve descrição de cada um dos bancos.

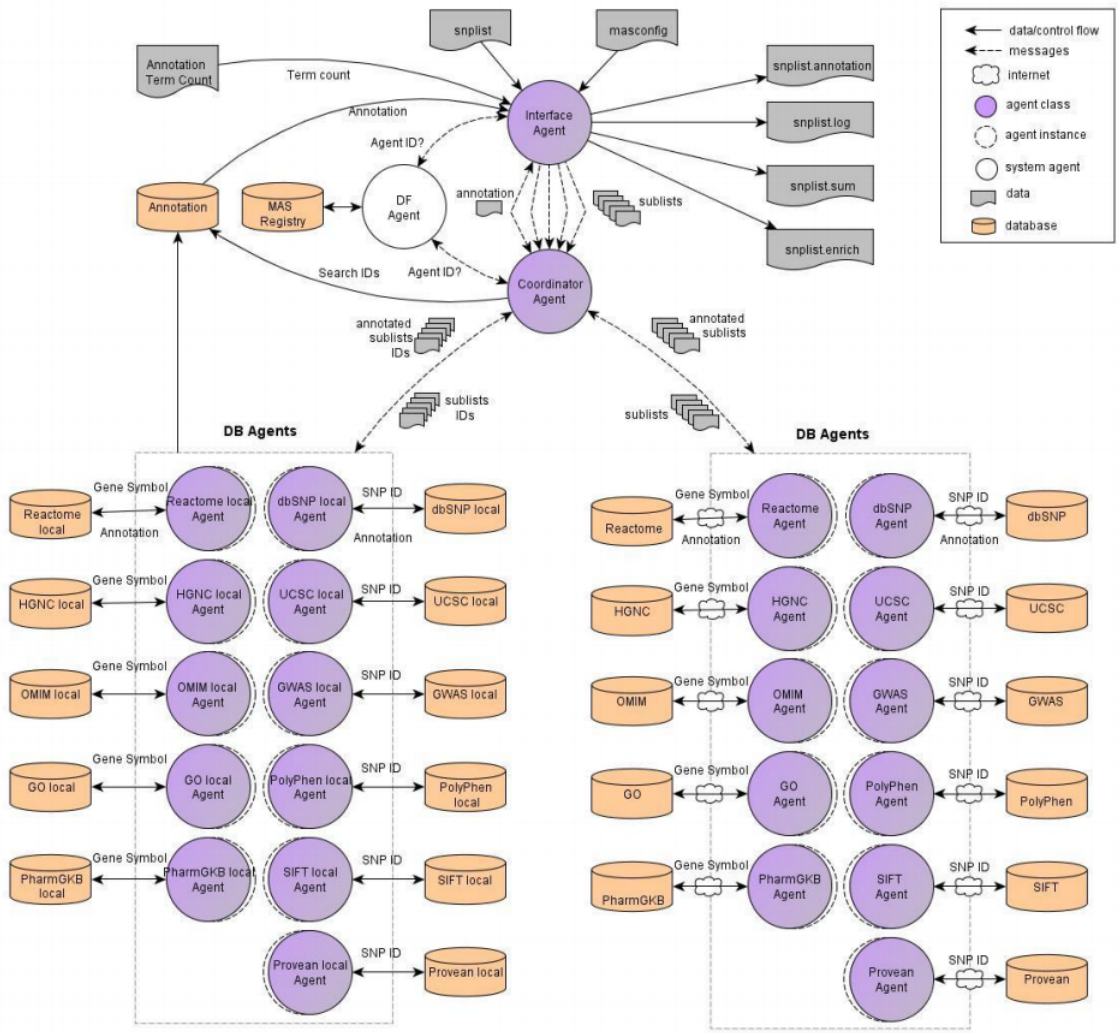

Figura 5 – Arquitetura da aplicação MASSA

1. dbSNP (Sherry et al., 2001) - dbSNP é um dos principais repositórios de variantes genéticas humanas existentes, em especial dos polimorfismos de nucleotídeo único (SNPs) e, em menor escala, de pequenas inserções, deleções e

2. UCSC (Karolchik et al., 2014) - A plataforma UCSC Genome Browser permite a visualização de diferentes níveis de anotação genômica em diversos organismos.

3. Gene Ontology (Ashburner et al., 2000) - A base de dados do Gene Ontology é atualmente a maior base de dados atual sobre genes e suas funções. As informações disponíveis são legíveis por humanos e por máquinas de forma que a base de dados seja consultada tanto para análises robustas quanto para pesquisas manuais.

4. PharmGKB (Whirl-Carrilo et al., 2012) - É uma ferramenta voltada para a área da farmacogenômica que abrange informações incluindo diretrizes clínicas e rótulos de medica-

mentos, associações gene-medicamento potencialmente acionáveis e relações genótipo-fenótipo. O PharmGKB coleta, organiza e divulga conhecimento sobre o impacto da variação genética humana nas respostas aos medicamentos.

5. OMIM ("Online Mendelian Inheritance in Man, OMIM®," 2014) - É um catálogo de genes humanos, distúrbios e características genéticas. Continuamente atualizado, com foco na relação molecular entre variação genética e expressão fenotípica. As informações em cada entrada do OMIM são citadas e a referência completa é fornecida. O OMIM é curado pelo McKusick-Nathans Institute of Genetic Medicine, The Johns Hopkins University School of Medicine.

6. Reactome (Croft et al., 2014; Joshi-Tope et al., 2003 Matthews et al., 2007, 2009, Vastrik et al., 2007) - É um banco de dados sobre vias metabólicas humanas e suas reações. As informações no banco de dados são de autoria de biólogos especialistas, que são inseridas e mantidas pela equipe de curadores e equipe editorial da Reactome.

7. HGNC (Gray et al., 2013) - O Comitê de Nomenclatura Gênica da HUGO (Human Genome Organisation) é responsável por definir e padronizar a nomenclatura de genes. A partir deste comitê são definidos os nomes e símbolos oficiais dos genes humanos. Na comunicação científica é necessário respeitar os princípios de clareza, precisão, comunicabilidade e consistência. Para isso, existe um comitê de nomenclatura genética, cuja função é assegurar que cada gene humano tenha um nome e símbolo únicos que sejam usados consistentemente na literatura científica. Apesar dos esforços, ainda encontramos textos onde o autor se refere a um gene usando um símbolo obsoleto, ou não faz a distinção adequada entre o gene e a proteína, prejudicando a compreensão por parte do leitor (Splendore, 2005) além de trazer enormes dificuldades para técnicas de análises de dados que utilizam robôs para realizar a coleta dos dados.

8. NHGRI GWAS Catalog (Welter et al., 2014) - O Gwas Catalog foi fundado pelo NHGRI (National Human Genome Research Institute) em 2008, em resposta ao rápido aumento no número de estudos de associação genômica (GWAS) publicados. Um GWAS tem como objetivo verificar se existe associação estatística entre o fenótipo estudado (podendo ser características como altura e peso ou doenças como diabetes, obesidade ou câncer) e milhões de variantes genotipadas ao longo do genoma. O GWAS Catalog fornece um

banco de dados consistente, pesquisável, visualizável e disponível gratuitamente sobre associações entre fenótipos e SNPs publicados.

9. PolyPhen2 (Adzhubel et al., 2010) - O Polyphen 2 é uma ferramenta que tem como objetivo prever o possível impacto de uma substituição de aminoácidos na estrutura e função de uma proteína humana. Esta previsão é baseada em várias características que compreendem a sequência, informações filogenéticas e estruturais que caracterizam a substituição.

10. Provean/SIFT (Choi et al., 2012), SIFT (Kumar, Henikoff, 2009) - O SIFT é uma ferramenta bioinformática que tem o objetivo de prever se uma substituição de aminoácidos afeta a função da proteína com base na homologia de sequência e nas propriedades físicas dos aminoácidos. O SIFT pode ser aplicado a polimorfismos não-sinônimos que ocorrem naturalmente e outros tipos de mutações.

A lista de SNPs indicada no arquivo de configuração é lida pelo agente de Interface, checada em busca de inconsistências (identificadores diferentes do previsto) e dividida pela taxa de paralelização (pR) em N sub listas (onde N = pR), as quais são enviadas separadamente para o agente coordenador em N mensagens. O controle de cada uma das listas é feito através de um identificador de pesquisa e o coordenador também é informado sobre o modelo de anotação, se local ou remoto, e se há opção de anotação simples. Após receber as sublistas de SNPs, o agente coordenador as envia aos agentes DB para anotação. Há dois passos sequenciais a serem seguidos pelo Coordenador para gerenciar o processo de anotação: a) a anotação de informações sobre polimorfismos, e b) a anotação de informações sobre genes.

No passo (a) cada sublista é enviada aos agentes DB que disponibilizam as informações sobre polimorfismos, como dbSNP, UCSC, GWAS Catalog, PolyPhen2, Provean e SIFT. Dentre os atributos retornados neste primeiro momento está o nome do gene onde o SNP se encontra. Com isso, o Coordenador segue à próxima fase, passo (b), onde filtra a lista de genes e a envia aos demais agentes DB: OMIM, GO, PGKB,HGNC e Reactome. Entretanto, o passo b) só é executado na anotação completa, na anotação rápida o Coordenador requisita apenas a anotação do agente dbSNP.

Porém através do relato dos usuários e de análises feitas através de algoritmos escritos por mim, foi identificado perda de desempenho na aplicação. De forma que, para entradas muito grandes, é observado um aumento considerável do tempo de resposta da aplicação. Utilizando recursos disponíveis no próprio sistema operacional do servidor do laboratório, eu coletei dados

da aplicação enquanto a mesma era executada para atestar esta perda de desempenho. Para que fosse possível manter a leitura do número de threads e recursos totais que a aplicação utilizava, foi utilizado um script em python cujo objetivo é monitorar algumas variáveis do arquivo contido no caminho '/proc/¡pID¿' (Caminho no qual o sistema Linux mapeia todos os processos em execução, assim como os recursos utilizados pelos mesmos) e guardar essas informações em vetores, de modo que ao fim do processo fosse possível plotar gráficos relacionados à execução do processo em questão. Para os testes foi utilizado o servidor do Laboratório de Diversidade Genética Humana (LDGH) (Dell PowerEdge R810, Processador Intel(R) Xeon(R) CPU E7-4820 @ 2.00GHz (64 núcleos), 128 GB DDR3 de RAM, 17 TB de HD mecânico, SIstema operacional CentOS 6.8 x86_x64 e utilizando o SGBD MySQL 5.2). Foram utilizados 6 arquivos como entrada com 100.000, 200.000, 500.000, 750.000, 1.500.000 e 2.500.000 SNPs respectivamente. A fim de medir o desempenho da aplicação em diferente cenários variamos o parâmetro (pR) com 18,26,36,50. A primeira análise realizada teve como objetivo medir o desempenho da aplicação, observando o consumo de memória RAM a partir de diferentes tamanhos de entrada, utilizando os parâmetros pré estabelecidos pelo autor. Foram utilizados os modos de anotação remoto e local para esta análise. A segunda análise mediu o desempenho da aplicação, observando o tempo de execução a partir de diferentes tamanhos de entrada, utilizando os parâmetros pré estabelecidos pelo autor. Foram utilizados os modos de anotação remoto e local para esta análise. Os resultados que foram obtidos através das duas análises são a média resultante dos valores obtidos para memória e tempo, de 3 execuções, para cada um dos tamanhos de entrada.

2.2 MapReduce Cluster

O MapReduce Cluster é um conjunto de ferramentas implementadas no intuito de apoiar operações com grandes conjuntos de dados. Um cluster de computadores é um conjunto de máquinas onde dois ou mais computadores operam de forma sincronizada e paralela. Pela forma como os computadores trabalham em conjunto, um cluster pode ser considerado como um único sistema. O Cluster proposto e testado neste trabalho foi implementado com base no framework Apache Hadoop. O Hadoop é um framework de código aberto escrito na linguagem Java, cuja principal finalidade é permitir o armazenamento e processamento de grandes

volumes de dados de forma paralela e distribuída. O Hadoop é composto por 4 ferramentas base: (i) O Hadoop Common, um conjunto de bibliotecas e utilidades básicas necessárias para a execução das tarefas em um cluster; (ii) O Hadoop Distributed File System (HDFS), o sistema de arquivos distribuído implementado pela ferramenta, que permite que os dados sejam distribuídos e processados através dos nós do cluster; (iii) Hadoop Yarn, uma plataforma responsável pelo gerenciamento dos recursos do cluster e agendamento de tarefas; (iv) Hadoop MapReduce, a implementação do modelo de programação MapReduce para processamento de dados em larga escala. O MapReduce é um modelo de programação baseado no paradigma de programação funcional. A principal ideia por trás do MapReduce é mapear (função map) um conjunto de dados em uma coleção, de forma que estes dados sejam mapeados e identificados através de tuplas ¡chave, valor¿ que então são operados de forma paralela. A partir disso todas as tuplas são reduzidas (Função reduce) com a mesma chave produzida na saída final do processamento. Esta abordagem adota o princípio de abstrair toda a complexidade da paralelização de uma aplicação usando apenas as funções Map e Reduce, que por serem funções puras (funções que não produzem efeito colateral no código) são facilmente paralelizáveis. O MapReduce traz para o Hadoop um modo mais simples de escrever e realizar operações em grandes conjuntos de dados de forma paralela, porém, consultas complexas sobre uma quantidade grande de arquivos pode se tornar uma tarefa bastante difícil de ser realizada no HDFS. Buscado uma maneira de simplificar as operações onde fossem necessárias partes do conteúdo de arquivos contidos no HDFS e para facilitar a utilização de algumas funções de MapReduce foi implementado o data warehouse Apache Hive. O Hive é um data warehouse implementado sob o ecossistema Apache Hadoop com o objetivo de simplificar a leitura, escrita e o gerenciamento de grandes datasets contidos no HDFS usando a linguagem Hive Query Language (HQL). O HQL é uma linguagem SQL-Like que implicitamente é convertida em funções MapReduce pela ferramenta toda vez que uma consulta é realizada. Em bancos de dados relacionais tradicionais, quando uma tabela é criada ela precisar ter um esquema bem definido e com isso essa tabela impõe seu esquema aos dados quando estes são carregados. Os dados são verificados pelo SGBD (Sistema Gerenciador de Banco de Dados) no momento em que vão sendo inseridos na tabela (schema on write). Em comparação, o Hive não verifica os dados no esquema da tabela no momento da gravação, ao invés disso, a ferramenta executa verificações quando os mesmo são lidos (schema on read) (White, 2012). O Hive trabalha com a arquitetura schema on read pois sua principal função não é realizar consultas habituais e trazer resultados baseados nelas. O Hive é um data

warehouse que busca uma forma de realizar operações sobre grandes arquivos e/ou conjuntos de dados distribuídos de forma paralela e utilizando uma linguagem simples que é convertida em funções de MapReduce. Ao trabalhar com o modo schema on read é possível ganhar flexibilidade e desempenho nas operações, uma vez que grandes volumes de dados possuem várias fontes e formatos diferentes além de consultas não serem a principal finalidade desta ferramenta. Com isso, diferente de bancos de dados relacionais tradicionais, o Hive não oferece recursos de acesso aleatório aos dados do HDFS. Com recursos de gravação e interatividade limitados pelo Hadoop e pelo MapReduce, o Hive como relatado anteriormente é destinado à execução de transformações e operações em lote além de grandes consultas analíticas. No intuito de suprir demandas por consultas de baixa latência sobre os dados contidos em alguns arquivos do HDFS, foi implementado um banco de dados não-relacional de baixa latência para operar em conjunto com o ecossistema Hadoop. O banco de dados não relacional escolhido para esta tarefa foi o MongoDB. Por ser orientado a documentos JSON, o mongoDB permite que os bancos de dados possam ser modelados de forma mais natural. Onde os dados podem ser aninhados em hierarquias complexas e ainda assim serem indexáveis e consequentemente fáceis de recuperar. Ao invés do conceito de normalização dos dados que utilizam chaves estrangeiras e relações presentes no modelo relacional, o MongoDB preza pela desnormalização dos dados (redundância), de forma que operações envolvendo joins entre tabelas sejam desencorajadas. As 3 ferramentas descritas formam o ecossistema do MapReduce Cluster. Através de funções de MapReduce é possível operar sobre grandes conjuntos de dados e a partir destas operações é possível popular bancos de dados no MongoDB para análises posteriores. O Ecossistema proposto se comunica através de funções MapReduce emitidas por um usuário ou através de consultas feitas diretamente ao MongoDB, o caminho de dados pode ser visto na figura 6.

O cluster Hadoop implementado no Laboratório de Diversidade Genética Humana conta com 2 nós de dados (Datanodes) e 1 nó mestre (Namenode). O nó principal (Namenode) está implementado no servidor principal do laboratório, um servidor modelo Dell PowerEdge R810, 128 GB de memória RAM, 64 Cores (Intel(R) Xeon(R) CPU E7-4820 @ 2.0 Ghz) e 3 TB de HD. Os 2 datanodes do cluster foram implementados em computadores com as seguintes configurações: (i) Dell PowerEdge Mini Tower Server T130 (processador Intel Xeon E3-1220 v6 3.0GHz e 8GB de memória RAM e 1.5 TB de HD); (ii) Iomega StorCenter px12-400r, Intel Core i3 (Quad Core 3.3GHz, 4GB DDR3 e 12TB distribuídos em 4 HDs SATA 3.5). Ao todo, o cluster implementado possui 3 nós, 100 GB de RAM, 72 processadores e 8.75TB de memória

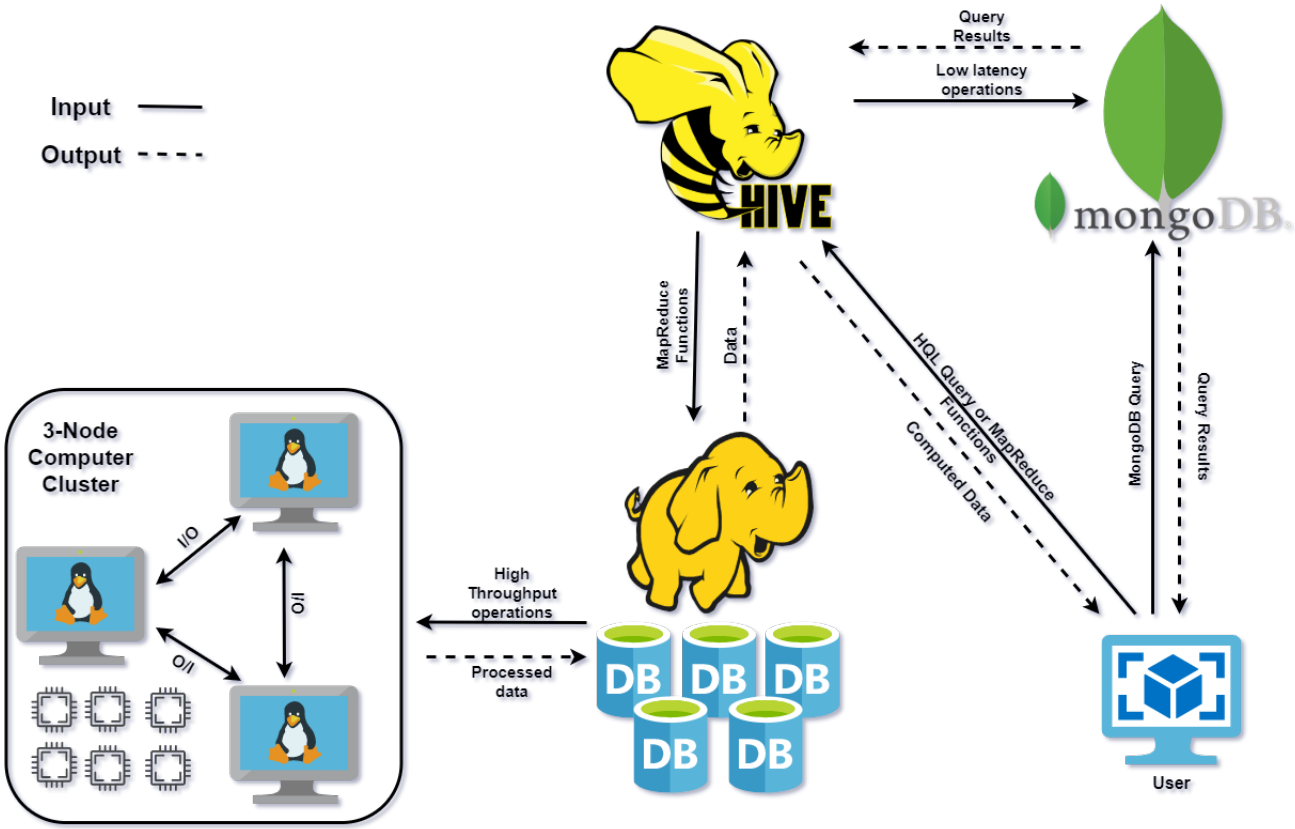

Figura 6 – Caminho de dados e comunicação das ferramentas do ecossistema MapReduce Cluster

distribuídos em 5 HDS.

2.3 Annotation Tool for Rapid Analysis

A ferramenta de anotação de variantes genéticas aqui proposta foi criada com base na principal idéia trazida pelo MASSA (Souza, 2014); A união de diversos bancos de dados genéticos públicos com objetivo de agregar e enriquecer informações sobre variantes genéticas humanas. Uma ferramenta de anotação de variantes genéticas pode ser vista como uma aplicação movida estritamente à consultas feitas sobre um banco de dados. Dessa forma, encontrar uma modelagem eficiente e que trouxesse informações relevantes para os pesquisadores para os bancos de dados foi a tarefa mais importante executada durante o desenvolvimento da ferramenta. Utilizando o apoio do MapReduce Cluster foi possível retirar informações contidas nos arquivos dos bancos de dados que estavam no HDFS. Através de funções MapReduce foi realizada a conversão dos arquivos disponibilizado pelos bancos de dados públicos (dos formatos .vcf e

.gff) para o formato JSON (formato utilizado pelo MongoDB), filtrar informações consideradas relevantes desses dados e então inseri-los em coleções do MongoDB, através de um driver disponibilizado pelo próprio MongoDB e que possui como objetivo realizar a comunicação entre o Hive e o MongoDB, de forma que as saídas das consultas feitas através do Hive sejam direcionadas e inseridas em coleções do MongoDB. Com o objetivo de montar um banco de dados para anotação de variantes genéticas baseado nos bancos de dados do MASSA (Multi Agent System for SNP Annotation), o primeiro passo foi elaborar, utilizando o processo descrito anteriormente, um modo de trazer informações relevantes para anotação de SNPs do principal banco de dados utilizado pelo MASSA, o dbsnp. Na figura 7 é possível observar o esquema de como ocorreram as consultas e comunicações entre as ferramentas do cluster assim como um exemplo de um documento originado a partir das consultas e já inserido no MongoDB em sua respectiva coleção de documentos.

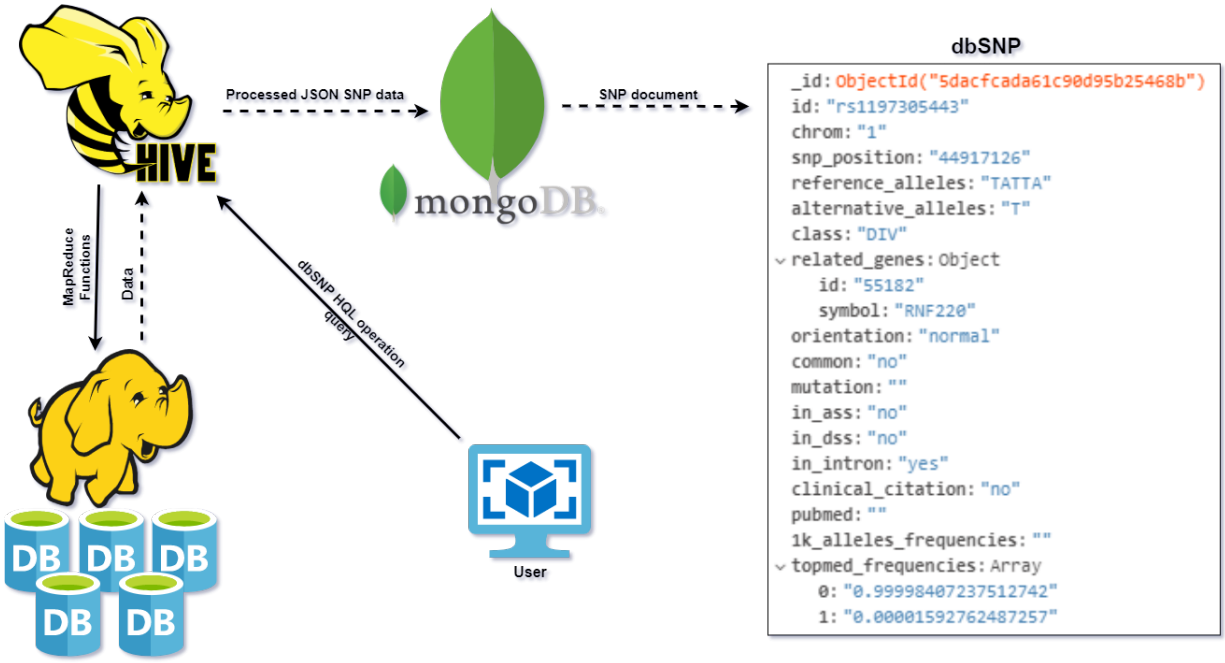

Figura 7 – Processo de inserção de informações contidas nos arquivos do dbsnp na coleção de SNPs do MongoDB.

No passo seguinte, optamos por criar uma única coleção de genes, contendo informações dos três bancos de dados que contém informações sobre genes: HGNC (Gray et al., 2013), Gene Ontology (Ashburner et al., 2000) e UCSC (Karolchik et al., 2014). A partir de consultas utilizando HQL e joins, foi possível unir os 3 bancos de dados através de atributos que ambos tinham em comum. Então essa consulta foi filtrada e apenas atributos considerados relevantes (como o nome do gene, nomes sinônimos para o mesmo gene, posição no genoma, etc..) fo-

ram inseridas na coleção de genes do MongoDB. Um esquema deste processo bem como um exemplo dos documentos de genes resultante pode ser visto na figura 8.

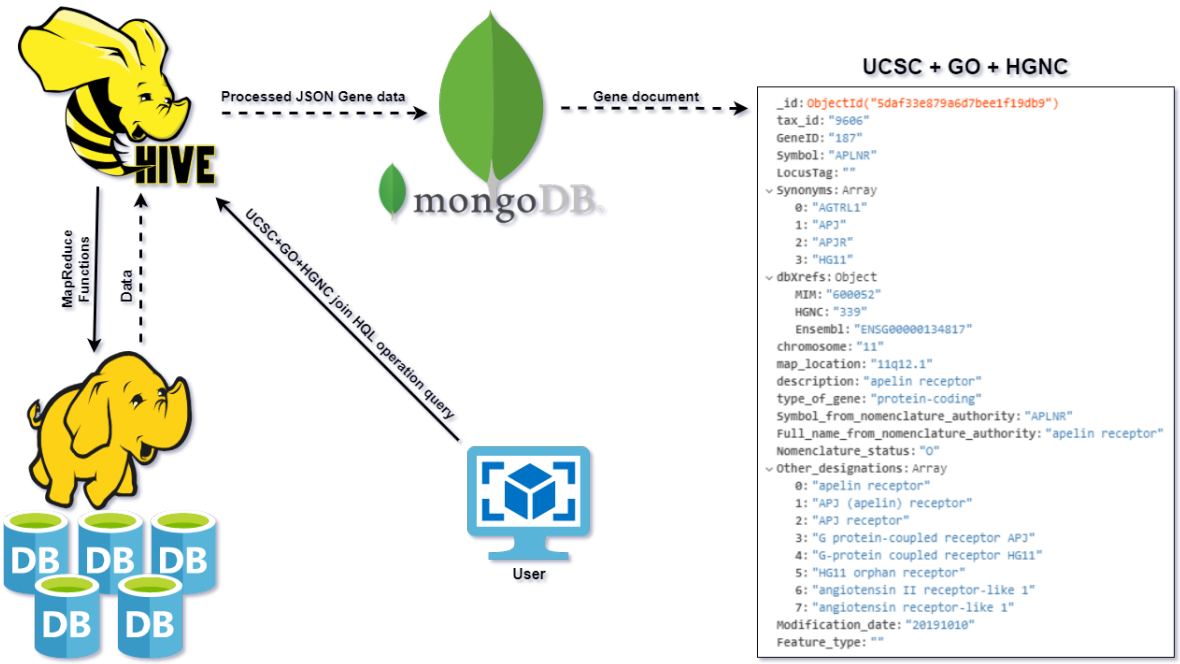

Figura 8 – Processo de inserção de informações contidas nos arquivos do UCSC, GO e HGNC na coleção de genes do MongoDB.

O passo seguinte consistiu em integrar os bancos de dados relacionados à atributos médicos (fenótipos e doenças) e farmacológicos à coleção de SNPs. O processo foi o mesmo utilizado nos passos anteriores e consistiu em agregar informações de fenótipos dos bancos de dados do Gwas Catalog e do ClinVar, além de informações sobre interações entre fármacos e SNPs contidos no PharmGKB. Resultando em documentos que podem como os que estão exemplificados na figura 9.

Ao todo existem 8 bancos de dados integrados em uma única coleção de SNPs no mongoDB, de forma que, consultas possam trazer o máximo de informações possíveis sobre um SNP em uma única requisição ao banco de dados. Os dados oferecidos pelas ferramentas Provean 2 e Sift ainda não foram integradas ao banco de dados pelo fato de serem necessários que alguns passos sejam executados antes, para que os dados de anotação sejam fornecidos pelas ferramentas. Ainda não encontramos um modo de integrá-las ao cluster de modo que as mesmas não resultem em um atraso em outras outras operações. Uma alternativa para as ferramentas ainda está sendo discutida.

dbsnp


```
_id: ObjectId("5dacfcada61c90d95b25468b")
id: "rs1197305443"
chrom: "1"
snp_position: "44917126"
reference_alleles: "TATTA"
alternative_alleles: "T"
class: "DIV"
related_genes: Object
    synonyms: "C1orf164"
    dbxrefs: Object
        MIM: "616136"
        HGNC: "25552"
        Ensembl: "ENSG00000187147"
    map_location: "1p34.1"
    description: "ring finger protein 220"
    gene_type: "protein-coding"
    Symbol_from_nomenclature_authority: "RNF220"
    full_nomenclature_name: "ring finger protein 220"
    other_designations: Array
        0: "E3 ubiquitin-protein ligase RNF220"
        1: "RING-type E3 ubiquitin transferase RNF220"
    feature_type: ""
orientation: "normal"
common: "no"
mutation: ""
in_ass: "no"
in_dss: "no"
in_intron: "yes"
clinical_citation: "no"
pubmed: ""
1k_alleles_frequencies: ""
topmed_frequencies: Array
    0: "0.99998407237512742"
    1: "0.00001592762487257"
```


GWAS Catalog


```
gwas_catalog: Object
    phenotype: "Breast cancer"
    risk_allele: "rs11076805-C"
    allele_frequency: "0.75"
    pubmed: "www.ncbi.nlm.nih.gov/pubmed/29058716"
    p_value: "1E-8"
    or_or_beta: 1.0869565
    region: "16p13.3"
    mapped_gene: "ADCY9"
    intergenic: 0
    context: "intron_variant"
```


Clinvar


```
clinvar: Object
    disease_name: "HEMOGLOBIN_G_(MAKASSAR)"
    clinical_significance: "other"
    hgvs_expression: "NC_000011.9:g.5248232T>G"
    dbXrefs: Array
    sequence_ontology: Object
    RS: "rs334"
```


PharmGKB


```
pharmGKB: Object
    phenotype_interaction: Object
        related_chemical: "ABT-751 (PA166104276)"
        related_phenotype: "metabolism/PK"
        significance: "no"
        notes: "p-value>0.05"
        sentence: "Allele G is not associated with ABT-751 pharmacokinetic parameters whe..."
        alleles: "G"
    drug_interaction: Object
        related_chemical: "lamotrigine (PA450164)"
        related_phenotype: "metabolism/PK"
        significance: "no"
        notes: "The concentration was measured as lamotrigine trough concentration / d..."
        sentence: "Genotype AA is not associated with increased concentrations of lamotri..."
        alleles: "AA"
```


Figura 9 – Documentos contidos na coleção de SNPs resultantes agregação de informações sobre fenótipos e fármacos.

3 Resultados

3.1 Multi Agent System for SNP Annotation

Ao realizar testes de desempenho na ferramenta de anotação MASSA, foi possível observar através dos gráficos obtidos que a ferramenta possui problemas tanto no modo como a memória RAM é utilizada pela aplicação quanto pela demora ao retornar resultados referentes à anotação das variantes. No gráfico exibido na figura 10 é possível observar como o uso de memória RAM escala de forma gradual ao variarmos o número de SNPs que são passados como entrada para a ferramenta.

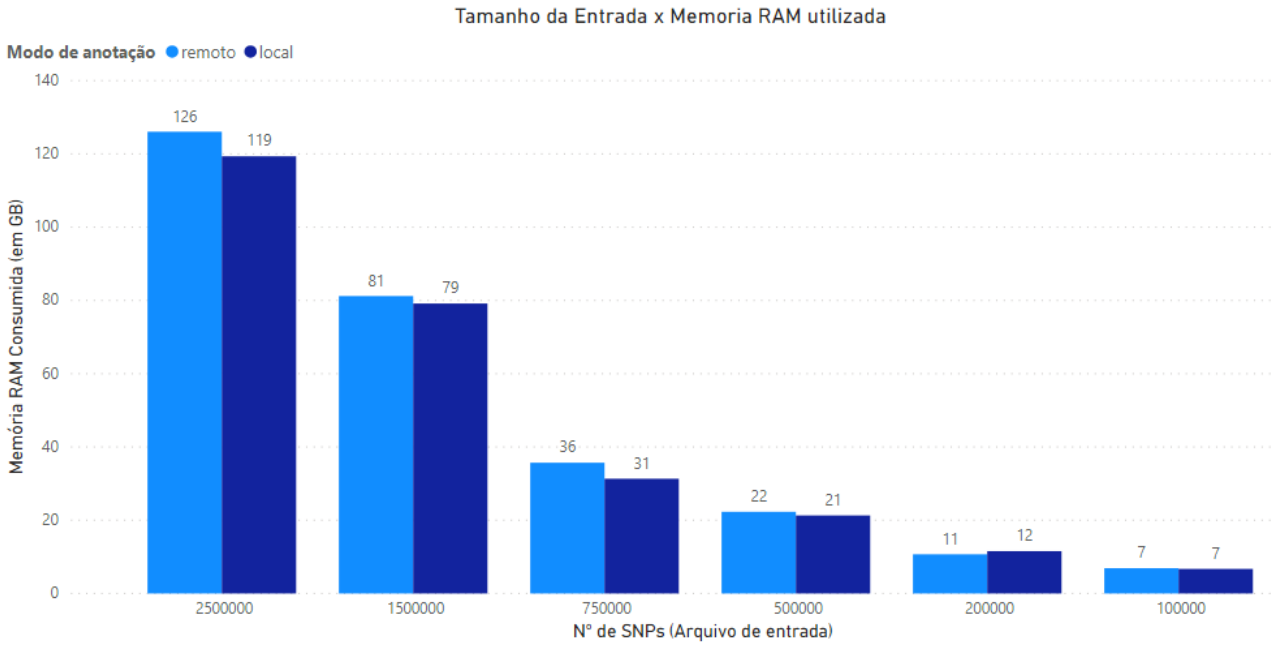

Figura 10 – Utilização de memória RAM consumida em relação à quantidade de SNPs para anotação remota (Azul claro) e anotação local (Azul escuro). Na execução envolvendo 2.5M SNPs da anotação remota aplicação não pode executar completamente. nas três execuções foram lançados exceções do tipo "OutOfMemory" quando a aplicação atingia 126 GB de RAM.

No gráfico mostrado na figura 11 pode-se observar como o tempo de execução da aplicação aumenta de forma exponencial ao se variar o número de SNPs que são passados como entrada para o MASSA.

Após observarmos todos os gráficos obtidos através dos testes é possível concluir que a aplicação sofre uma grande perda de desempenho quando entradas muito grandes são fornecidas

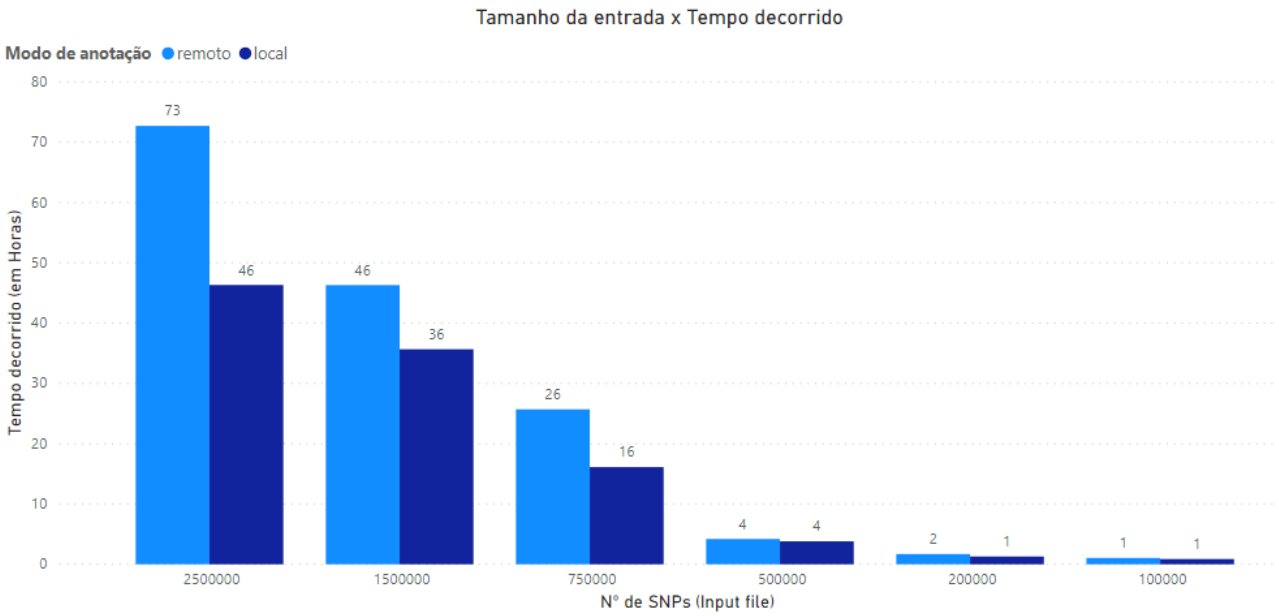

Figura 11 – Tempo de execução em relação ao tamanho da entrada para a anotação remota (Azul claro) e anotação local (Azul escuro).

e o modo de anotação completo é utilizado. Essa perda de desempenho se dá, principalmente, por por três motivos: (i) A desatualização dos dados presentes nos bancos força o usuário a utilizar a anotação remota; (ii) O modelo utilizado não extrai o máximo do desempenho de um banco de dados relacional e (iii) que o paralelismo utilizado não é adequado a aplicação.

3.2 Cluster Annotation Tool

Os testes de desempenho realizados no MASSA também foram realizados na ferramenta proposta neste trabalho. Através do gráfico da figura 12 é possível notar um grande ganho de desempenho em relação ao consumo de memória de RAM na execução da ferramenta Cluster Annotation Tool.

Em relação ao tempo de execução da ferramenta é possível notar também um grande ganho de desempenho. Ao invés de horas uma anotação de 2.5 milhões de SNPs pode ser feita em minutos. Na figura 13 é possível observar o gráfico relativo ao tempo decorrido durante a anotação de variantes em relação ao tamanho da entrada.

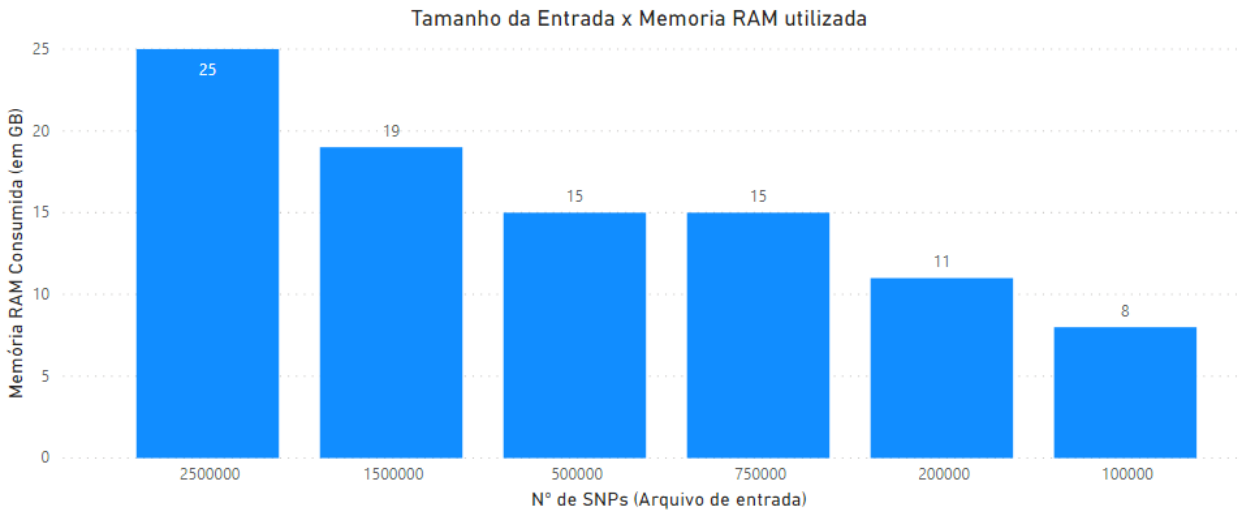

Figura 12 – Consumo de memória RAM em relação ao tamanho da entrada da ferramenta Cluster Annotation Tool.

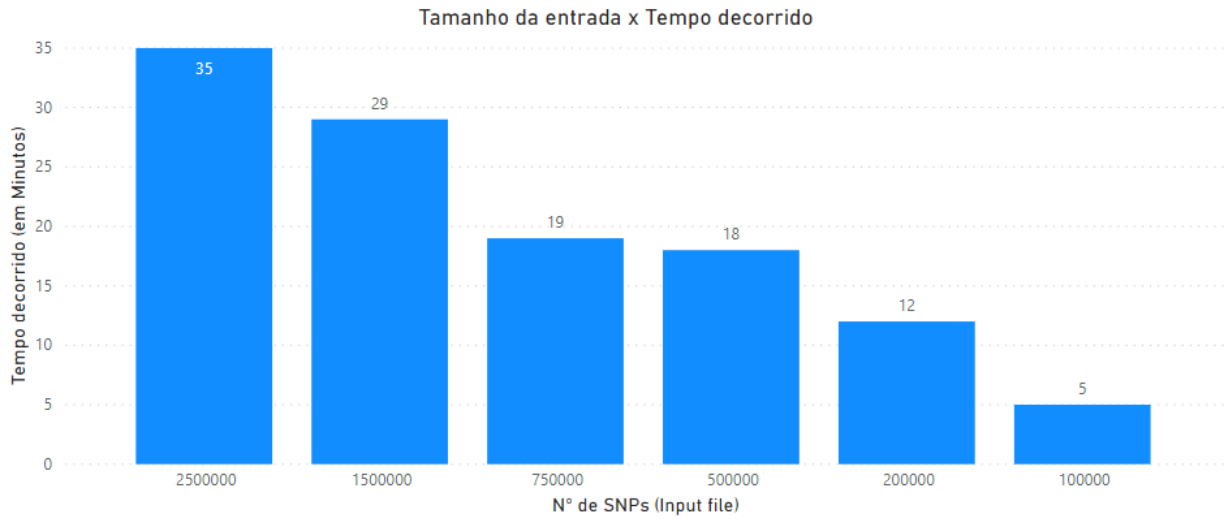

Figura 13 – Tempo de execução da ferramenta Cluster Annotation Tool em relação ao tamanho da entrada.

3.3 Comparação entre performance e anotação das ferramentas

3.3.1 Performance

Comparando os resultados obtidos através das anotações feitas da nova ferramenta com o MASSA é possível notar um grande ganho de desempenho. As anotações acontecem de forma muito mais rápida, de modo que, com apenas uma requisição ao mongoDB é possível trazer dados relevantes de 8 bancos de dados. Além disso, o uso de memória RAM por parte do cluster está limitado por contêineres gerenciados pelo YARN, evitando assim que qualquer overflow de memória possa acontecer. Na figura 12 é possível observar um gráfico comparando o consumo de memória RAM entre o MASSA e a ferramenta de anotação Cluster Annotation Tool.

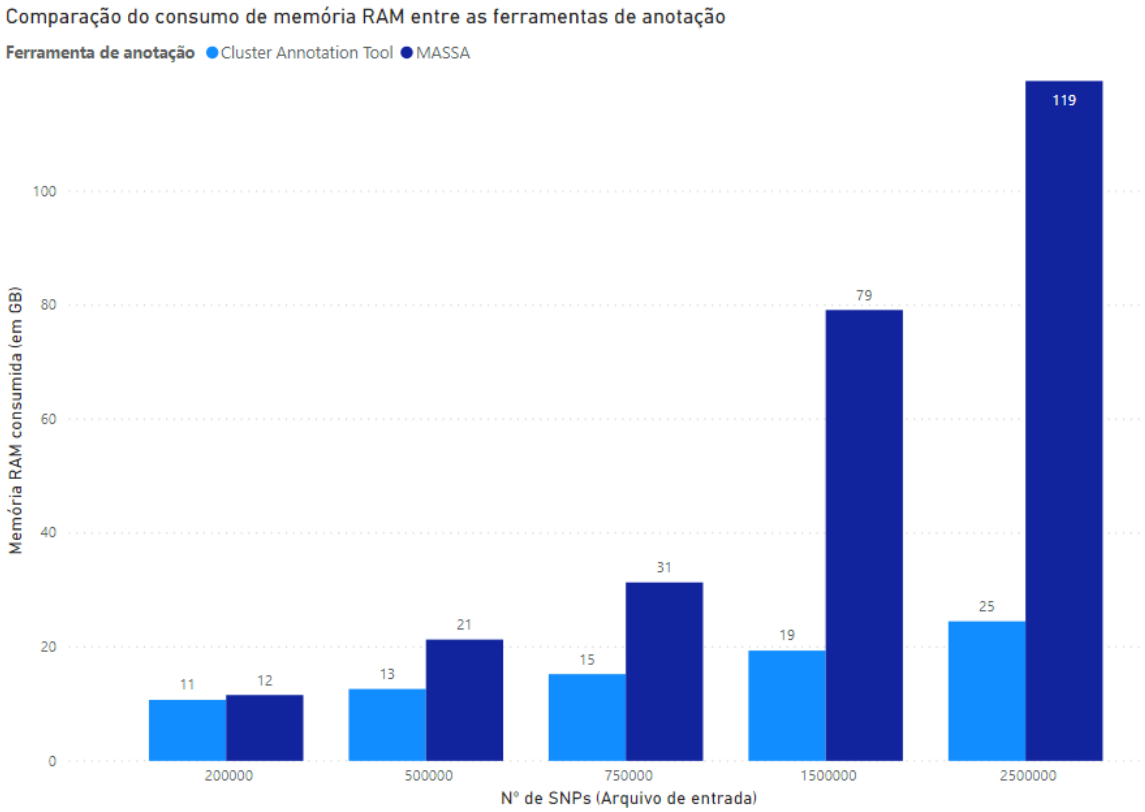

Figura 14 – Comparação do consumo de memória RAM durante a execução das duas ferramentas de anotação de variantes genéticas.

Quando comparamos o desempenho em relação ao tempo de resposta da aplicação a ferramenta Cluster Annotation Tool possui um grande ganho de desempenho quando compa-

rado ao MASSA, principalmente pelo fato de grandes operações (como consultas envolvendo 2.5 milhões de SNPs) ocorrerem de forma distribuída no cluster. Na figura 13 é possível observar um gráfico comparando o tempo de execução entre o MASSA e a ferramenta Cluster Annotation Tool.

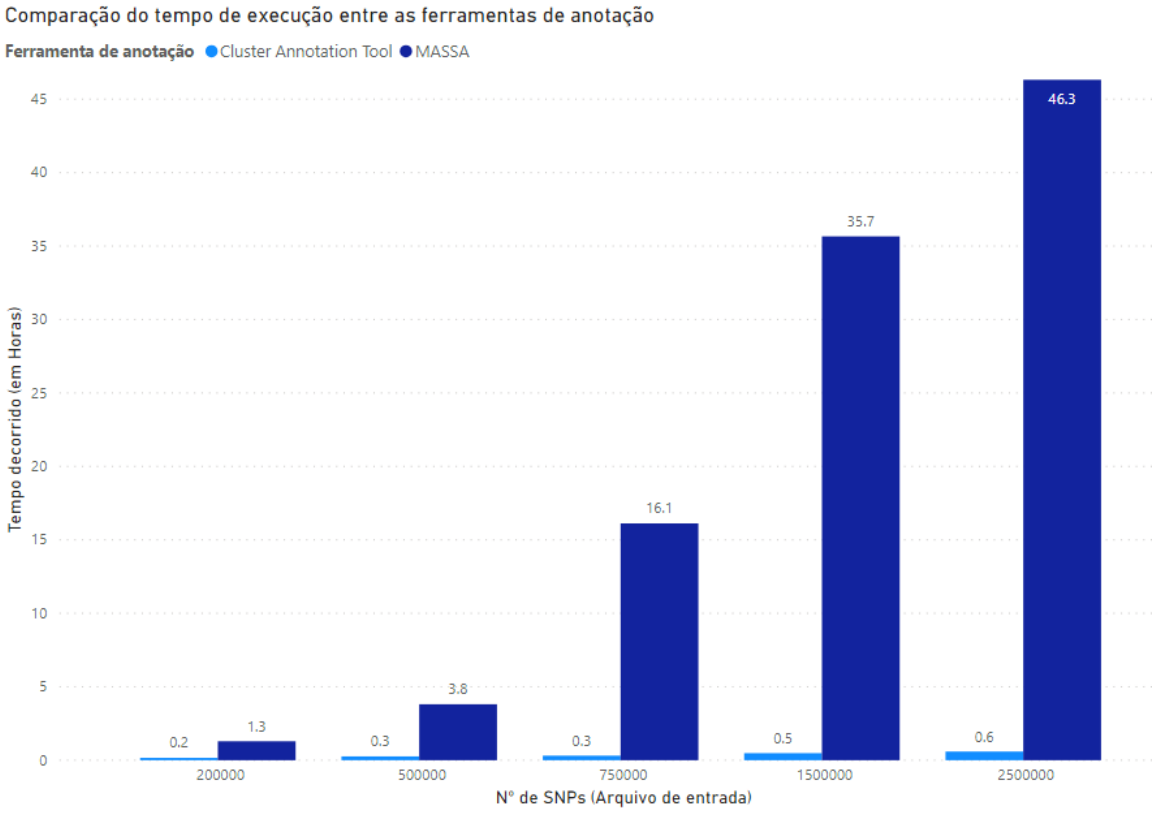

Figura 15 – Comparação do tempo de execução entre as ferramentas de anotação MASSA e Cluster Annotation Tool para diferentes tamanhos de entrada.

3.3.2 Anotação de variantes genéticas

Outro relevante fator à ser comparado é o resultado das anotações. O MASSA tem como objetivo trazer informações de até 66 colunas de 11 bancos de dados para agregar informações sobre SNPs. Porém, por possuir grandes problemas de performance, seu modo de anotação completo é inviável. Com isso a ferramenta só é utilizável para grandes conjuntos de dados no modo de anotação simples. No modo de anotação simples são retornadas 17 atributos sobre SNPs que provém apenas do dbsnp, estes atributos são: 1) O ID do SNP; 2) O tipo do polimorfismo; 3) O gene relacionado; 4) O ID do gene relacionado; 5) A região de transcrição; 6)

A Numeração do nucleotídeo; 7) O Cromossomo à qual o SNP pertence; 8) A posição do snp no cromossomo; 9) O Alelo ancestral; 10) Orientação da fita de dna; 11) A versão do genoma; 12) A posição Inicial do SNP; 13) A Posição final do SNP; 14) o ID do mRNA; 15) a versão do mRNA; 16) Outros alelos relacionados; 17) A frequência do alelo em populações continentais. Na figura 14 é possível observar um exemplo de parte do arquivo de anotação de SNPs produzido pela ferramenta MASSA.

(1)Polymorphism Id	(2)Polymorphism Type				(3)Gene Symbol	(4)Gene ID	(5)Transcript Region		(6)Nucleotide Numbering coding DNA			(7)Chromosome	(8)Chromosome Position		(9)Ancestral Allele
(10)Orientation	(11)Assembly Build Version				(12)Assembly Coord Start	(13)Assembly Coord End			(14)mRNA accession			(15)mRNA version	(16)Alleles	(17)Frequency	
rs9310888	null	null	null	null	g.29286762	3	29286762	null	null	g37	null	null	NC_000003	11	G,A	0.299913,0.700087
rs1517634	null	null	null	null	g.224183485	2	224183485	null	null	g37	null	null	NC_000002	11	G,A	0.42557,0.57443
rs10497705	null	null	null	null	g.190492014	2	190492014	null	null	g37	null	null	NC_000002	11	N,T,C	0.000149701,0.437425,0.562425
rs10498255	null	CAB39	51719	intron	c.-43-12444	2	231612230	null	null	37_3	818216471	818216471	NM_001130850	1	T,C	0.75886,0.24114
rs798887	null	null	null	null	g.54793188	19	54793188	A	null	g37	null	null	NC_000019	9	G,A	0.338315,0.661685
rs9325872	null	null	null	null	g.20480271	8	20480271	G	null	g37	null	null	NC_000008	10	G,A	0.288755,0.711245
rs10519979	null	null	null	null	g.149634951	4	149634951	G	null	g37	null	null	NC_000004	11	G,A	0.451865,0.548135
rs10508349	null	null	null	null	g.8298964	10	8298964	null	null	g37	null	null	NC_000010	10	G,A	0.814768,0.185232
rs868179	null	null	null	null	n.351+292	2	177549497	null	null	g37	null	null	XR_108435	1	G,A	0.600227,0.399773
rs10488172	null	EXOC4	60412	intron	c.1514+20282	7	133335176	null	null	37_3	713680181	713680181	NM_021807	3	G,T	0.220853,0.779147

Figura 16 – Exemplo de um arquivo de anotação resultante da ferramenta MASSA.

Na figura acima é possível observar que além da limitação imposta pelo modo de anotação simples existem também muitos SNPs anotados que possuem campos com atributo NULL (ou Nulo), isto devido a desatualização dos bancos de dados utilizados pela ferramenta, fazendo com que muitos SNPs ou informações não pudessem ser encontradas. o Cluster Annotation Tool foi pensado de modo que este problema não ocorra. Separar as idéias propostas pelo Dr. Giordano no MASSA em ferramentas distintas, foi um passo importante para reestruturar o funcionamento da ferramenta. Para que o banco de dados de SNPs esteja sempre atualizados, foram desenvolvidos scripts por mim que são executados periodicamente, e possuem o objetivo de comparar a data da última atualização no FTP dos bancos de dados e quando necessário baixá-los e realizar a atualização dos mesmos no HDFS. Na anotação fornecida pela ferramenta Cluster Annotation Tool são agregados informações de mais 20 atributos além dos 17 atributos trazidos pelas anotações do MASSA. Destes 37 atributos, 15 são informações sobre SNP, 12 são informações sobre fenótipos e interações do SNP com fármacos, 6 são informações sobre genes e suas funções e 4 relacionados a vias metabólicas à qual o mesmo pode estar relacionado. Na figura 15 é possível observar um exemplo do arquivo de anotação produzido pela ferramenta Cluster Annotation Tool. Os dados anotados foram SNPs relacionados a anemia falciforme, citada anteriormente.

Ao comparar o resultado relativo a anotação de variantes produzido pelas duas ferramentas é possível concluir que a ferramenta Cluster Annotation Tool além de possuir um desempenho superior e anotar as mesmas informações básicas que o MASSA, agrega mais informações que abrangem uma variedade maior de bancos de dados.

```
#rs334
## Other related gene Information

rsid    tax_id  GeneID  Symbol  LocusTag        Synonyms        chromosome      map_location    description     type_of_
rs334   9606    3040    HBA1            ['ECYT7', 'HBA-T2', 'HBH']      16      16p13.3 hemoglobin subunit alpha 2
rs334   9606    3040    HBA2            ['ECYT7', 'HBA-T2', 'HBH']      16      16p13.3 hemoglobin subunit alpha 2
```

```
type_of_gene    Symbol_from_nomenclature_authority      Full_name_from_nomenclature_authority   Nomenclature_status     Other
 2      protein-coding  HBA2    hemoglobin subunit alpha 2      O       ['hemoglobin subunit alpha', 'alpha globin', 'alpha-2
 2      protein-coding  HBA2    hemoglobin subunit alpha 2      O       ['hemoglobin subunit alpha', 'alpha globin', 'alpha-2
```

```
#related Pharmaco/Phenotype
## CLINVAR

rsid    reference_allele        alternative_alleles     disease_name    clinical_significance   other   hgvs_expression dbXrefs sequence_ontology
rs334   T       A       HEMOGLOBIN_G_(MAKASSAR) other   NC_000011.9:g.5248232T>A        { "id" : "SO:0001483", "url" : "http://www.sequenceontology.org/browser/current_release/term/SO:0001483" }
```

Figura 17 – Exemplo de anotação da variante genética relacionada à anemia falciforme produzida pela ferramenta Cluster Annotation Tool

4 Conclusão

Através de análises realizadas por mim na Ferramenta de anotação de variantes MASSA ficou claro quanto a perda de desempenho que a ferramenta sofre quando grandes conjuntos de dados são passados como entrada. Baseado neste problema foi proposto neste trabalho a implementação de um cluster de computadores utilizando o framework Apache Hadoop para apoiar operações com grandes volumes de dados. Com base neste cluster e em algoritmos MapReduce foi proposta uma nova ferramenta de anotação de dados biológicos baseado no MASSA, que se mostrou bastante eficiente quando comparada ao mesmo, tanto no consumo de memória RAM quanto no tempo de resposta da aplicação. Além disso, a ferramenta Cluster Annotation Tool é capaz de anotar até 37 atributos relativos a um SNP. 20 atributos a mais além dos atributos que o MASSA também anota. Uma política de atualização para os bancos de dados, também foi implementado. De forma a manter os bancos de dados utilizados para anotação de variantes sempre atualizados. O ambiente de computação paralela proposto aqui abre um novo leque de possibilidades para os bioinformatas do Laboratório de Diversidade Genética Humana, de forma que grandes volumes de dados possam ser tratados em horas ao invés de dias. Por outro lado para tirar proveito deste tipo de ferramenta requer conhecimento sobre programação e também sobre modelo MapReduce.

5 Bibliografia

A global reference for human genetic variation, The 1000 Genomes Project Consortium, Nature 526, 68-74 (01 October 2015) doi:10.1038/nature15393.

Alexander, D. H., et al. “Fast Model-Based Estimation of Ancestry in Unrelated Individuals.” Genome Research, vol. 19, no. 9, 2009, pp. 1655–1664., doi:10.1101/gr.094052.109.

Amdahl, Gene M. “Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities.” Proceedings of the April 18-20, 1967, Spring Joint Computer Conference on - AFIPS ’67 (Spring), 1967, doi:10.1145/1465482.1465560.

Araújo, Gilderlanio S., et al. “Integrating, Summarizing and Visualizing GWAS-Hits and Human Diversity with DANCE (Disease-ANCEstry Networks).” Bioinformatics, vol. 32, no. 8, 2015, pp. 1247–1249., doi:10.1093/bioinformatics/btv708.

Barabási, Albert-László, et al. “Network Medicine: a Network-Based Approach to Human Disease.” Nature Reviews Genetics, vol. 12, no. 1, 2010, pp. 56–68., doi:10.1038/nrg2918. “Concept 15 DNA and Proteins Are Key Molecules of the Cell Nucleus.” Friedrich Miescher :: DNA from the Beginning, www.dnaftb.org/15/bio.html.

“DNA Sequencing Costs: Data.” Genome.gov, www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data.

Dahm, Ralf. “Discovering DNA: Friedrich Miescher and the Early Years of Nucleic Acid Research.” Human Genetics, vol. 122, no. 6, 2007, pp. 565–581., doi:10.1007/s00439-007-0433-0. Dean, Jeffrey, and Sanjay Ghemawat. “MapReduce.” Communications of the ACM, vol. 51, no. 1, 2008, p. 107., doi:10.1145/1327452.1327492.

Hogeweg, Paulien. “The Roots of Bioinformatics in Theoretical Biology.” PLoS Computational Biology, vol. 7, no. 3, 2011, doi:10.1371/journal.pcbi.1002021.

"Human Genome Project FAQ." Genome.gov, www.genome.gov/human-genome-project/Completion-FAQ. "Initial Sequencing and Analysis of the Human Genome." Nature, vol. 409, no. 6822, 2001, pp. 860–921., doi:10.1038/35057062.

"Initial Sequencing and Analysis of the Human Genome." Nature, vol. 409, no. 6822, 2001, pp. 860–921., doi:10.1038/35057062.

"International Human Genome Sequencing Consortium Publishes Sequence and Analysis of the Human Genome." Genome.gov, www.genome.gov/10002192/2001-release-first-analysis-of-human-genome.

Karp, Gerald. Cell and Molecular Biology: Concepts and Experiments. John Wiley, 2008.

Li, Mulin Jun, and Junwen Wang. "Current Trend of Annotating Single Nucleotide Variation in Humans – A Case Study on SNVrap." Methods, vol. 79-80, 2015, pp. 32–40., doi:10.1016/j.ymeth.2014.10.003.

Mardis, Elaine R. "A Decade's Perspective on DNA Sequencing Technology." Nature, vol. 470, no. 7333, 2011, pp. 198–203., doi:10.1038/nature09796.

Metzker, Michael L. "Sequencing Technologies — the next Generation." Nature Reviews Genetics, vol. 11, no. 1, 2009, pp. 31–46., doi:10.1038/nrg2626.

O'connell, Jared, et al. "Haplotype Estimation for Biobank-Scale Data Sets." Nature Genetics, vol. 48, no. 7, 2016, pp. 817–820., doi:10.1038/ng.3583.

Person. "State of MongoDB March, 2010: MongoDB Blog." MongoDB, MongoDB, 8 Mar. 2010, www.mongodb.com/blog/post/state-of-mongodb-march-2010.

Sauna, Zuben E., and Chava Kimchi-Sarfaty. "Understanding the Contribution of Synonymous Mutations to Human Disease." Nature Reviews Genetics, vol. 12, no. 10, 2011, pp. 683–691., doi:10.1038/nrg3051.

Shen, Terry H., et al. “SNPit: A Federated Data Integration System for the Purpose of Functional SNP Annotation.” Computer Methods and Programs in Biomedicine, vol. 95, no. 2, 2009, pp. 181–189., doi:10.1016/j.cmpb.2009.02.010.

Thusoo, Ashish, et al. “Hive - a Petabyte Scale Data Warehouse Using Hadoop.” 2010 IEEE 26th International Conference on Data Engineering (ICDE 2010), 2010, doi:10.1109/icde.2010.5447738.

Trelles, Oswaldo, et al. “Big Data, but Are We Ready?” Nature Reviews Genetics, vol. 12, no. 3, 2011, pp. 224–224., doi:10.1038/nrg2857-c1.

Wan, Yue, et al. “Landscape and Variation of RNA Secondary Structure across the Human Transcriptome.” Nature, vol. 505, no. 7485, 2014, pp. 706–709., doi:10.1038/nature12946.

White, Tom. Hadoop: the Definitive Guide. O’Reilly, 2012. Wu, C. H. “The Protein Information Resource.” Nucleic Acids Research, vol. 31, no. 1, 2003, pp. 345–347., doi:10.1093/nar/gkg040.

“A Map of Human Genome Variation from Population-Scale Sequencing.” Nature, vol. 467, no. 7319, 2010, pp. 1061–1073., doi:10.1038/nature09534.

“The Most Popular Database for Modern Apps.” MongoDB, www.mongodb.com/.