UNIVERSIDADE FEDERAL DO PARANÁ

Análise de *SNP's* do genoma da estirpe HM053 de *Azospirillum brasilense*, utilizando sequenciamento de nova geração.

Curitiba, Dezembro de 2013

VINICIUS DE SARAIVA CHAGAS

Análise de *SNP's* do genoma da estirpe HM053 de *Azospirillum brasilense*, utilizando sequenciamento de nova geração.

Monografia apresentada ao Departamento de Genética, Setor de Ciências Biológicas da Universidade Federal do Paraná, como requisito para obtenção do título de bacharel em Ciências Biológicas.

Orientadora: Profª Dra. Roseli Wassem
Co-orientador: Dr. Joseph A. Medeiros Evaristo

Curitiba, Dezembro de 2013.

AGRADECIMENTOS

Primeiramente agradeço ao “Tonho” (papai), “Iaia” (mamãe), “Juba” (maninha), “Biel” meu irmão de nascimento separado e as famílias Chagas e Saraiva. Todos que me deram o primeiro empurrão na vida e continuam me ajudando até hoje, sem vocês não existiria base para o que sou hoje.

Agradeço a Professora Drª. Roseli Wassem e o Dr. Joseph Evaristo pela orientação, ensinamentos e principalmente pela paciência em me ensinar como pensar ciência e escrevê-la.

Porém tão importante quanto este trabalho para a conclusão de uma etapa e início de outra, estão todas as pessoas que me fizeram chegar até aqui e me ajudaram a ser quem eu sou.

Portanto muito obrigado a todo o pessoal da Biologia de 2009 que me deram uma força nos estudos e que me fizeram rir e chorar. Um beijo especial para a “Xarcelli” (maninha), “Mà”(maninha), “Fran”(protetora), “Gonza”(brother), “Line”(agitada), “Vandrão”(crânio) e o “Chito”(lindão), minha família de sala. É galera nos estranhamos de vez em sempre mas também rimos de tudo isso ai, e eu só tenho que agradecer e muito a todos vocês.

Agradecimento especial a todas as famílias que me adotaram nesses cinco anos, com muito carinho e amor, que resolveram levar este “caiçara” para seus lares.

Muito obrigado aos Kosmalas: Myrthes, Maurício e “Gê”. Sempre serei agradecido por tudo que fizeram, e por terem me adotado de verdade.

Agradeço também aos “poderosos” Accioly: “Brunão” (irmão), a médica (quase doutora) Marina que me fez crescer, ensinou muito e estará sempre comigo, a “Cris” e “Caro” por terem permitido um intruso na casa de vocês e me acolherem sempre tão bem, a “Bia” por permitir o filho dela andar com um elemento igual a mim e a “Isa” porque ela é uma fofa!

E a mais recente família que me aceitou como agregado, agradeço ao Professor Luís “Zão” Fávaro por toda sua amizade e de toda sua “prole”: “Wan” (poderosa), “Burda” e “Di” (amorecos), “Bibis” (parceira), Anderson (novato). E

os agregados Silvia (guria), “Weg” (poderoso), e “Claudjenha” (paciente) muito obrigado pela força e carinho nos últimos momentos desta etapa.

Um gigantesco agradecimento a todos vocês e todos os outros que não tenho espaço pra citar, mas me ajudaram a chegar aqui.

RESUMO

Azospirillum brasilense é uma protobactéria que possui hábito de vida livre ou associado com raízes de plantas, sendo responsável por inúmeros benefícios às plantas hospedeiras auxiliando, por exemplo, no desenvolvimento de um sistema radicular mais robusto, absorção de água e de minerais e favorecendo o crescimento vegetal. A principal característica desta bactéria é a de utilizar o dinitrogênio atmosférico (N_2), que é convertido em amônia (NH_3) através da ação do complexo nitrogenase. Todo este processo de fixação é regulado em diferentes níveis, tanto transcricionais quanto pós-traducionais, uma vez que o processo de fixação de nitrogênio possui elevado gasto energético. Machado (1988) produziu mutantes de *A. brasilense* entre os quais está a linhagem HM053, que fixa nitrogênio mesmo na presença de nitrogênio fixado e também excreta amônio. Já Hauer (2012) identificou uma mutação no gene codificador de glutamina sintetase (GS), *glnA,* na estirpe HM053. A enzima GS é uma das principais enzimas responsáveis pela assimilação de amônio e esta mutação é a provável causa dos fenótipos observados. Para avaliar se esta é a única mutação presente no seu genoma, o objetivo deste trabalho foi analisar os dados de sequenciamentos do genoma, fornecidos por dois sequenciadores de nova geração (Ion Proton™ System e Applied Biosystems (SOLiD) Sequencing). Os três sequenciamentos apresentaram características diferentes, e por consequência foram tratados separadamente, apresentando assim resultados distintos. Como resultado do tratamento de qualidade dos *reads* foram selecionados 99,82% dos dados de Ion Proton 7, 99,87% de Ion Proton 6 e 98,97% do SOLiD. Do volume de dados resultante foram alinhados com o genoma referência 92,05% dos *reads* de Ion Proton 7, 92,17% de Ion Proton 6 e 54,15% do SOLiD. Em cada sequenciamento foi possível detectar a seguinte quantidade de SNP's: 37 para Ion Proton 7 e 20 para Ion Proton 6 e SOLiD. A partir do estudo foi possível confirmar a mutação em GS com duas tecnologias diferentes de sequenciadores, corroborando o estudo de HAUER (2012), outras 16 SNP's presentes nos dados de ambas tecnologias de sequenciadores necessitam de estudos posteriores para saber seu relacionamento com a fixação biológica de nitrogênio.

Palavras-chave: *Azospirilum brasilense,* HM053, GS, bioinformática, SNPs.

ABSTRACT

Azospirillum brasilense is a protobactéria that can be found having a free lifestyle or associated with plants roots, being responsible for numerous benefits to the host plants aiding, for example, developing a more robust root system, absorption of water and minerals and promoting plant growth. The main feature of this bacteria is to use atmospheric dinitrogen (N_2), which is converted into ammonia (NH_3) through the action of nitrogenase complex. This whole process of fixing is set to different transcriptional and post-translational levels, since the process of nitrogen fixation have high energy expenditure. Machado (1988) produced mutants from *A. brasilense* among which are the HM053 strain which fixed nitrogen even in the presence of fixed nitrogen, and excretion of ammonium. Hauer (2012) discovered a mutation in the glutamine synthetase (GS) gene *glnA*, the strain HM053. The GS enzyme is a key enzyme responsible for the assimilation of ammonium and this mutation is the likely cause of the observed phenotypes. To assess whether this is the only mutation present in their genome, the objective of this study was to analyze data from the genome sequencing by two next-generation sequencing (Ion Proton™ System e Applied Biosystems (SOLiD) Sequencing). The three sequencing methods had different features, and consequently were treated separately and so presenting different results. As a result for the quality of the treatment, reads were selected, being, 99,82% data of Ion Torrent 7, 99,87% of Ion Torrent 6 and 98,97% of SOLiD. The resulting volume of data were aligned with the reference genome with a match of 92.05% in the reads from the Ion Torrent 7, 92.17% from Ion Torrent 6 and 54.15% of match in the SOLiD. In each sequencing was possible to detect the following amounts of SNP's: 37 for Ion Torrent 7 and 20 to Ion Torrent 6 and SOLiD. From the study it was confirmed the mutation in GS with two different technologies sequencers corroborating the study of HAUER (2012), other 16 SNPs present in the data of both technologies sequencers require further studies to know his relationship with the biological nitrogen fixation.

KEYWORDS: *Azospirilum brasilense,* HM053, GS, bioinformatics, SNP’s

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

O gênero *Azospirillum* pertence à subclasse α de protobactérias, possuindo hábito de vida livre ou associado com raízes de plantas. Este gênero pode ser encontrado aderido na superfície das raízes e/ou livre na rizosfera de gramíneas com importância econômica como milho, arroz e trigo. Estas bactérias são caracterizadas por serem gram negativas e diazotróficas, ou seja, possuem a capacidade de se desenvolverem utilizando o dinitrogênio atmosférico (N_2), que é convertido em amônia NH_3 (STEENHOUDT; VANDERLEYDEN, 2000).

Estudos sobre a estrutura genômica do genêro de *Azospirillum* indicam a presença de cinco a sete megareplicons que variam de 0,65 a 2,6 Mpb. Alguns replicons possuem genes que expressam RNA ribossomal, indicando a existência de múltiplos cromossomos no genoma (MARTIN-DIDONET *et al*., 2000). Ainda com base em estudos genéticos, foi proposta a existência de sete espécies de *Azospirillum*: *A. brasilense*, *A. lipoferum* (TARRAND *et al.,* 1978), *A. halopraeferans* (REINHOLD *et al.,* 1987), *A. amazonense* (MAGALHÃES *et al.,* 1983)*, A. irakense* (KHAMMAS *et al.,* 1989). Porém os principais estudos genéticos e bioquímicos, especialmente com relação a fixação biológica de nitrogênio, foram realizados com *A. brasilense* (HUERGO *et al.*, 2008).

Morfologicamente *A. brasilense* apresenta um padrão misto de flagelos, onde um flagelo polar é sintetizado durante seu desenvolvimento em meio líquido e tem função primária natatória e os flagelos laterais adicionais são induzidos durante o crescimento em meio sólido sendo responsáveis pelo povoamento bacteriano em superfícies sólidas. Quando o ambiente apresenta condições desfavoráveis, a espécie se converte a forma de cisto e desenvolve uma camada protetora de polissacarídeos, obtendo energia a partir do acúmulo de grânulos de poli-hidroxialcanoatos (STEENHOUDT; VANDERLEYDEN, 2000).

Dentre os benefícios que este microorganismo pode dar as plantas estão: a fixação de nitrogênio, o desenvolvimento de um sistema radicular mais robusto, produção de fitormônios, aumento na absorção de água e de minerais do solo, levando ao crescimento mais rápido da planta (HUNGRIA, 2011).

O nitrogênio é um dos elementos fundamentais para o desenvolvimento

das plantas, porém é um fator limitante na produção agrícola, sendo a suplementação de nitrogênio no solo por fertilizantes químicos uma das soluções para auxiliar no desenvolvimento das plantas. Entretanto, estes métodos vêm ocasionando problemas ambientais como: aumento de emissões de óxido de nitrogênio na atmosfera, acidificação de solos e eutrofização de águas (DIXON; KHAN, 2004). Em contrapartida, o processo de fixação biológica fornece nitrogênio menos propenso a lixiviação e volatilização, uma vez que ele é utilizado *in situ*, tornando o processo biológico uma alternativa barata, limpa e sustentável para o fornecimento de nitrogênio na agricultura comercial (HUERGO *et al*., 2008). Por este motivo *A. brasilense* tem sido utilizado como inoculante em culturas agrícolas, pois pode converter nitrogênio atmosférico em amônia em ambiente com baixos níveis de nitrogênio, através da ação do complexo nitrogenase, que catalisa a seguinte reação:

$$N_2 + 10H^+ + 8e^- + 16Mg.ATP \xrightarrow{\text{Nitrogenase}} 2NH_3 + H_2 + 16Mg.ADP + 16Pi$$

Figura 1: Reação de redução de dinitrogênio em amônia pela nitrogenase

Esta enzima possui dois componentes protéicos, a dinitrogenase redutase, ou proteina Fe (NifH), transportadora ativa (dependente de ATP) de elétrons para a proteina Fe-Mo (NifDK), onde se encontra o sítio de redução de N_2 (STEENHOUDT; VANDERLEYDEN, 2000; DIXON; KHAN, 2004).

Pelo fato da redução de N_2 ser um processo de alto gasto energético e a nitrogenase ser uma enzima instável na presença de oxigênio, o processo de fixação biológica de nitrogênio possui uma rígida e complexa regulação, em todas as etapas, em resposta principalmente à presença ou ausência de oxigênio e amônio (STEENHOUDT; VANDERLEYDEN, 2000).

Em protobactérias os genes *nif*, como os *nifHDK,* codificadores da nitrogenase, são regulados pela proteína ativadora de transcrição NifA em conjunto com a RNA-polimerase σ^N. A proteína NifA faz parte da família EBP (enhancer-binding protein), envolvida na ativação da transcrição gênica em resposta a estímulos celulares, sendo estes, baixos níveis de oxigênio e amônio (DIXON; KHAN, 2004). NifA permanece inativa após expressa, sendo a proteína da família P_{II}, GlnB, necessária para sua ativação em condições de

fixação de nitrogênio (ARAUJO *et al.,* 2004; ARSENE; KAMINSKI; ELMERICH, 1996). Nessa condição, a nitrogenase expressa é ativa. Entretanto, quando os níveis de nitrogênio fixado ou de oxigênio aumentam, ela é inativada rapidamente. A regulação pós traducional da nitrogenase é decorrente da ação da DraT (dinitrogenase redutase ADP-ribosiltransferase) que ADP-ribosila a porção Fe da nitrogenase, inativando-a, quando existe alta concentração de NH_4^+ no ambiente. Esta modificação é revertida pela DraG (dinitrogenase redutase glicohidrólise ativadora) que remove a ADP-ribose do componente protéico da enzima, reativando a nitrogenase e a produção de NH_3 (DIXON; KHAN, 2004).

Em *A. brasilense* as proteínas que compõem o sistema regulatório de metabolismo do nitrogênio são: Glutamina sintetase (GS, codificada pelo gene *glnA*), uridililtransferase (GlnD, codificada pelo gene *glnD*), adenililatransferase (GlnE, codificada pelo gene *glnE*), proteínas transdutoras de sinal da família P_{II}, GlnB e GlnZ (codificadas pelos genes *glnB* e *glnZ*), o canal protéico de amônia AmtB (codificado pelo gene *amtB*), os dois componentes do sistema regulatório NtrB-NtrC, além de NtrY-NtrX (codificados respectivamente pelos genes *ntrB, ntrC, ntrY* e *ntrX*). O amônio disponibilizado após a fixação de nitrogênio é assimilado em *A. brasilense* através de duas vias: uma envolvendo a ação sequencial da GS e a glutamato sintase (GOGAT) e outra com a glutamato desidrogenase (GDH) sendo a primeira a principal rota de assimilação de amônio (revisado por HUERGO *et al.*, 2008).

A GS é uma enzima chave no metabolismo de nitrogênio, uma vez que esta reação constitui a única maneira de sintetizar Gln, que é um doador de nitrogênio na biossíntese de diversos compostos (revisado por ANTONIUK, 2007). A enzima é constituída de um dodecamêro composto de duas matrizes hexagonais sobrepostas, onde cada monômero pode ser modificado por adenililação, independentemente, sendo o número de monômeros modificados determinante do nível de atividade da enzima. De acordo com FORCHHAMMER (2007) a GS é considerada uma enzima ancestral devido à sua ampla distribuição e história filogenética.

A GS assimila amônio convertendo Glu em Gln, para então a GOGAT transferir o grupo amino de Gln para 2-oxoglutarato (2OG), formando duas

moléculas de Glu. Na segunda via de assimilação ocorre o processo de formação de Glu pela GDH a partir de 2OG e amônio. Depois da produção de Glu e Gln, são sintetizados outros compostos derivados de nitrogênio por transferência secundária. Assim, a presença de 2OG intracelular, o percussor de assimilação de amônia, indica baixos níveis de nitrogênio fixado, enquanto a presença de Gln, produto da reação, indica que os níveis de nitrogênio fixado encontram-se suficientes (Figura 2) (LEIGH; DODSWORTH, 2007).

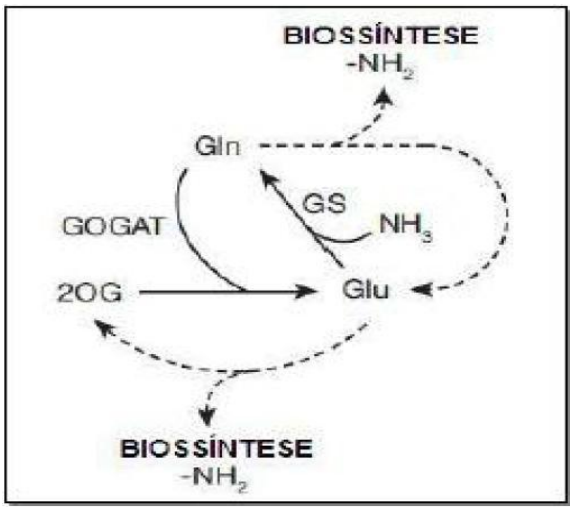

Figura 2: Vias GS/GOGAT de assimilação de amônia. Linhas sólidas: reações que levam a formação de Glu e Gln a partir de 2OG (2-oxoglutarato) e NH_3. Linhas pontilhadas: representação da transferência do grupamento amino para as vias biossintéticas (adaptado de LEIGH; DODSWORTH, 2007).

Porém estas vias só são ativas quando a GS não se encontra adenililada, pois quando adenililada, a GS permanece inativa. O processo de adenililação da GS é dependente dos componentes do sistema Ntr, respondendo diretamente a proteína GlnB. A proteína GlnB pode estar em seu estado natural ou uridililada, dependendo da disponibilidade de nitrogênio intracelular. Quando uridililada interage com GlnE que desadenilila GS. A proteina GlnD também pode agir como reguladora de GS, uma vez que ela catalisa a adição de uma molécula UMP a partir de uma molécula UTP para um resíduo de Tyr de P_{II} e também faz a remoção de UMP via hidrólise.

O produto e o substrato da reação de assimilação catalisada pela GS, 2OG e a Gln, são também reguladores da atividade da própria GS, pois sinalizam a disponibilidade de nitrogênio e o estado energético da célula. A

proporção de Gln e 2OG regula a atividade de GlnD, estimulando a uridililação, das proteínas GlnB e GlnZ quando há pouca Gln e a atividade de remoção do uridil em excesso de Gln. Dessa forma, GlnB encontra-se uridililada quando há pouco nitrogênio fixado disponível, e estimula a desadenililação da GS pela GlnE, ativando-a. A atividade elevada de GS leva a um aumento de Gln intracelular, fazendo com que a GlnD remova o grupamento uridil de GlnB, e assim, estimulando GlnE para a inativação da GS (Figura 3) (LEIGH; DODSWORTH, 2007).

A regulação da expressão de GS em proteobactérias, além de ser a nível pós-traducional, também ocorre durante a transcrição. Em enterobactérias o gene *glnA* está contido no operon *glnAntrBC*, onde os genes *ntrB e ntrC* codificam duas proteínas de um sistema de dois componentes. Enquanto a NtrB trata-se de uma proteína sensora histidina quinase, a NtrC é uma proteína reguladora de resposta que se liga ao DNA e regula sua transcrição (MERRICK.; EDWARDS, 1995).

Em *A. brasilense* o gene codificador de Gln sintetase faz parte do operon bicistronico *glnBA* que é regulado por três tipos de promotores: dois localizados anteriormente ao gene *glnB* (*glnBp*σ^{70} e *glnBp2*σ^{N}); e um promotor não identificado na região intergênica de *glnBA*. Em ambiente onde existe uma limitação de amônio, os genes *glnB* e *glnA* são co-transcritos pelo promotor dependente de σ^{N} e ativado pela proteína NtrC-P (HUERGO *et al*., 2008). Nestas condições o operon *glnBA* é fortemente transcrito a partir do promotor *glnBp2*, que é dependente do fator σ^{N} da RNA polimerase e de NtrC (ZAMAROCZY; PAQUELIN; ELMERICH, 1993). Por outro lado, em condições de excesso de amônio o promotor do operon *glnBA* é *glnBp1*, que é dependente de σ^{70} e pouco expresso nessas condições. O gene *glnA* é também transcrito de forma independente de *glnB* a partir de seu promotor ainda não conhecido (ZAMAROCZY; PAQUELIN; ELMERICH, 1993). Além disso, o promotor *glnBp1* é sobreposto a um sítio de ligação de NtrC, que ao ser ocupado por NtrC-P, favorece a ação do promotor *glnBp2* em condições limitantes de amônio. (BOZOUKLIAN; ELMERICH, 1986).

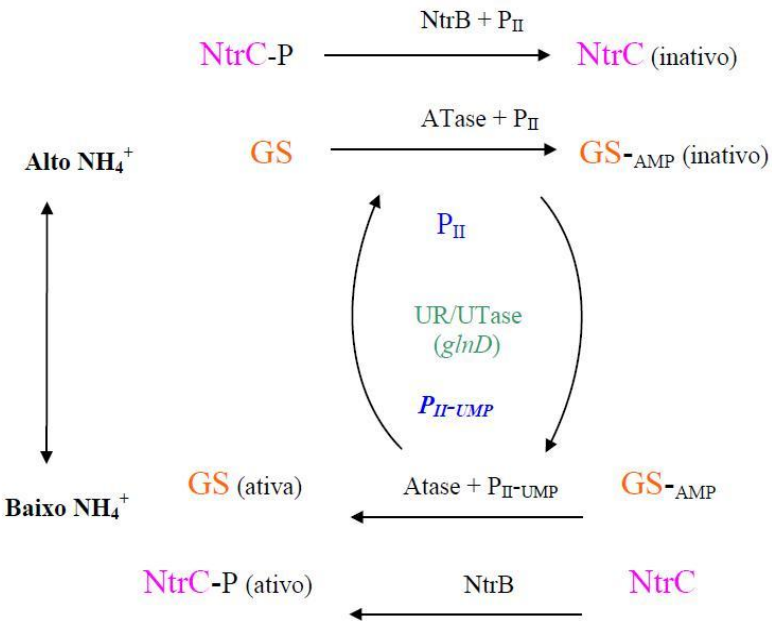

Figura 3: Modelo esquemático da regulação das atividades da Gln sintetase e sistema Ntr em resposta à disponibilidade de nitrogênio fixado no meio. A Utase (GlnD) catalisa a uridililação e desuridililação de P_{II} (GlnB e GlnZ). A NtrB catalisa a fosforilação e desfosforilação da NtrC (MERRICK & EDWARDS, 1995).

O fato do processo de fixação de nitrogênio, junto com suas etapas de assimilação e incorporação do nitrogênio fixado, serem altamente reguladas a nível transcricional, pós-transcricional e pós-traducional, é consequência do alto custo energético do processo (DIXON; KHAN, 2004). Apesar da forte regulação, estirpes capazes de fixar nitrogênio na presença de amônia já foram isoladas e são denominados de Nif^C, ou seja, Nif constitutivas (DOMMELEN *et al*., 2003). O isolamento de tais estirpes é de grande interesse uma vez que estas têm grande potencial para serem utilizadas em formulações de inoculantes de uso agrícola já que potencialmente, o nitrogênio fixado constitutivamente será liberado para consumo da planta. Para isso a maior parte destas linhagens é selecionada utilizando compostos análogos ao amônio, como MA (metilamônia) e EDA (etilenodiamina) que são tóxicos quando metabolizados pela GS. Consequentemente, os mutantes selecionados apresentam absorção deficiente de MA e/ou uma GS não funcional (HUERGO *et al*, 2008).

DOMMELEN *et al* (2003) caracterizaram dois mutantes derivados da estirpe Sp7 de *A. brasilense* (7028 e 7029), com baixa atividade de GS e

capazes de excretar altas quantidades de amônio. O mutante 7028 apresentava 18% da atividade de GS, comparada a estirpe selvagem e possuía capacidade de fixar nitrogênio constitutivamente. Esta estirpe apresenta uma mutação que leva à troca de uma Arg na posição 322 da GS por uma cisteína. Já o mutante 7029 não tem a capacidade de fixar nitrogênio, apresenta uma troca de Asp por asparagina na posição 52 da GS e possui menos de 1,5% da atividade de GS, comparada à estirpe selvagem. Neste mesmo estudo, um plasmídeo contendo o gene *glnA* de A*. brasilense* fez com que o mutante 7028 recuperasse a regulação da atividade da nitrogenase por amônio, e 7029 se tornasse capaz de fixar nitrogênio.

Outros mutantes excretores de amônio foram obtidos por MACHADO (1988) utilizando a estirpe Sp7 de *A. brasilense* como origem. Além desse fenótipo, todos os mutantes fixam nitrogênio constitutivamente e têm alterações na atividade e/ou regulação da GS. Acredita-se que a habilidade de excretar amônio seja decorrente do acúmulo de maior quantidade de amônio, já que este não é assimilado devido às alterações na GS (SRIVASTAVA; TRIPATHI, 2006). Uma das estirpes obtidas, denominada HM053, apresenta ainda outras características, tais como: alterações no transporte de amônio e baixa atividade da GS; o não crescimento em meio com baixa concentração de amônio ou apenas com nitrato; alterações no perfil de adenililação da GS; capacidade de excretar altos níveis de amônio (MACHADO, 1988). Apesar das tentativas de identificar a causa desses fenótipos, não foi possível identificar quais genes estão mutados nessa estirpe (VITORINO, 2000). O sequenciamento do gene *glnA* dessa estirpe, realizada por HAUER (2012) permitiu a detecção de uma mutação que leva à mudança do resíduo de Pro na posição 347, para Leu na GS. Esses resíduos estão próximos a Arg 339 e a Arg 359, as quais são importantes como sítios de ligação a ATP e Glu, porém se postula que estas alterações não sejam a única explicação para o fenótipo de HM053. A identificação de outras mutações poderá contribuir para a compreensão dos mecanismos que regulam a fixação e assimilação de nitrogênio em A. brasilense, assim, o presente trabalho visa analisar o sequenciamento da linhagem mutante HM053, na busca de outras possíveis mutações.

3. OBJETIVOS

3.1 Objetivo Geral

Analisar os dados de sequenciamentos do genoma da linhagem mutante HM053, fornecidos por dois sequenciadores de nova geração (Ion Proton™ System e Applied Biosystems (SOLiD) Sequencing), confrontados com o genoma da estirpe selvagem *Azospirillum brasilense* FP2.

3.2 Objetivos específicos

- Encontrar parâmetros, no CLC genome ManWorkbench, de análise de polimorfismos nucleotídicos para dois tipos diferentes de sequenciadores.
- Analisar as variações nucleotídicas encontradas.
- Comparar as variações nucleotídicas identificadas em todos os sequenciamentos

4. MATERIAL E MÉTODOS

4.1 Estratégia geral

Para a busca de SNP's (*Single Nucleotide Polymorphism*) em HM053, utilizou-se o software CLC Workbench (versão 5.5.1) para a manipulação de dados previamente disponíveis: o genoma da estirpe FP2 de *Azospirillum brasilense* e três sequenciamentos do mutante HM053. O genoma da estirpe FP2 de *A. brasilense,* disponibilizado pelo banco de dados do grupo de pós-graduação em bioinformática da Universidade Federal do Paraná, foi utilizado como referência para identificação de SNP's. O núcleo de Fixação de Nitrogênio da UFPR disponibilizou os sequenciamentos do mutante HM053, sendo dois sequenciamentos gerados pelo sequenciador de nova geração Ion Proton™ System (denominadas Ion Proton 6 e Ion Proton 7), e um sequenciamento gerado pelo sequenciador SOLiD Sequencing, denominado SOLiD.

Os dois sequenciamentos do Ion Proton™ System (Ion Proton 6 e Ion Proton 7) contavam com 221.597 *reads* e 1.927.885 *reads* respectivamente, enquanto o sequenciamento do Applied Biosystems (SOLiD) Sequencing, SOLiD apresentou 911.118 *reads*.

Para tornar o mais confiável possível as detecções de SNP's estes dados passaram por diversas análises, com parâmetros específicos para as necessidades de cada sequenciamento. Primeiramente ocorreu a triagem de qualidade dos dados com a remoção de *reads* duplicados, e posteriormente os *reads* foram aparados de acordo com sua qualidade e tamanho. Na segunda etapa os *reads* foram mapeados com o genoma referência, para depois ocorrer a detecção de possíveis SNP's a partir da ferramenta de detecção de variações com base na qualidade (www.clcsupport.com, 03/06/2013).

4.2 Triagem dos dados

Os arquivos gerados pelo SOLiD Sequencing, passaram por um processo de refinamento, utilizando o SOLiD™ Accuracy Enhancement Tool (SAET). Este refinamento é utilizado para correção de erros na pré-montagem dos

reads, utilizando a alta cobertura e os dados de qualidade dos *reads* para diminuir a taxa de erro por um fator de 5 (Applied Biosystems SOLiD™ 3 Plus System, 2010).

Os *reads* de cada sequenciamento foram aparados a partir de parâmetros de probabilidade baseado na qualidade das bases, onde se é imposto um limite de probabilidade (*Limit*) variando de 0 a 1. Assim, o primeiro passo no processo do software é converter o índice de qualidade (Q) de cada base para uma probabilidade de p_{erro}: $p_{erro} = 10^{\frac{Q}{-10}}$, então bases com baixa qualidade terão valores de p_{erro} maior e bases de alta qualidade terão um p_{erro} menor. Em seguida cada base tem um novo valor calculado: *Limit* - p_{erro}, este valor irá ser negativo ou nulo (zero) para bases cuja probabilidade de erro é igual ou maior ao valor de limite, e positivo para bases que tem probabilidade de erro menor que o limite. Então o software quando realiza a leitura de cada base do *read* seleciona aquela que possuir o primeiro valor positivo e todas as bases em sequência até que exista uma que possua um valor igual a somatória de todas as bases anteriores pertencentes a sequência, tudo antes e depois desta região será cortado. O *read* será completamente removido se não existir nenhuma base com valor positivo.

Para comprovar a qualidade dos dados que seriam utilizados para a detecção de SNP's, foi realizada uma avaliação de qualidade em cada sequenciamento através da ferramenta repórter de qualidade (QC), que mostra diversos dados: qualidade dos *reads* e qualidade de bases. Foi pertinente para este trabalho a confirmação de que os *reads* que seriam utilizados na etapa de mapeamento teriam em sua maior parte uma qualidade média equivalente ou superior a Phred 20 que representa 1 erro em 100 bases (www.clcsupport.com, 03/06/2013).

4.3 Mapeamento dos reads contra o genoma referência

Para que o *read* possa parear com determinada área do genoma referência, é necessário que este atinja diversos requisitos pré-estabelecidos pelo software na ferramenta de mapeamento. Entre estes requisitos estão o custo para a aceitação das variações (bases diferentes, inserção e deleção) pareadas com determinada região, além de requisitos mínimos de tamanho

(número de nucleotídeos) e similaridade das bases do *read* com a região pareada (www.clcsupport.com, 03/06/2013).

4.4 Detecção de variação com base na qualidade

A detecção de variação com base na qualidade trata-se de uma ferramenta que busca variações encontradas após a etapa de mapeamento, desde que elas atinjam os requisitos mínimos pré-estabelecidos. Estes requisitos consistem nas variáveis do algoritmo de qualidade de vizinhança desenvolvido por ALTSHULER (*et al* 2000), onde se estabelece um número nucleotídeos vizinhos da base variante (raio), esta vizinhança deve possuir em cada nucleotídeo uma qualidade mínima pré-estabelecida, além de uma qualidade mínima necessária do próprio nucleotídeo variante analisado. A ferramenta também permite estabelecer qual a cobertura mínima de *reads* que a variação deve apresentar para ser considerada e a frequência que aparece nestes *reads* de cobertura (www.clcsupport.com, 03/06/2013).

5. RESULTADOS

O início de triagem dos dados consistiu na remoção de possíveis *reads* duplicados provenientes do processo de sequenciamento. Inicialmente os dois sequenciamentos do Ion Proton (Ion Proton 6 e Ion Proton 7) computavam 221.597 e 1.927.885 *reads* respectivamente, e o sequenciamento do SOLiD, 911.118 de *reads*. Porém após esse processo de triagem dos *reads* duplicados, se constatou que o sequenciamento Ion Proton 6 contava com 20,49% de sequencias duplicadas, o Ion Proton 7 com 23,26% e enquanto o SOLiD apresentou uma porcentagem de 14,66% de sequencias duplicadas (Tabela 1).

Após o processo de remoção de *reads* duplicados os sequenciamentos passaram por tratamentos com alguns parâmetros exclusivos, diferenciando o tratamento das amostras entre os sequenciadores Ion Proton e SOLiD, buscando a melhoria dos dados para a detecção de possíveis SNP's. Nesta etapa foram aparadas as bases dos *reads* com baixa qualidade, o sequenciamento SOLiD teve como parâmetro de limite (*Limit*) o valor de 0,05 o que seleciona sequências de bases do *read* com uma qualidade média superior a Phred 15. Já os sequenciamentos Ion Proton 6 e Ion Proton 7 tiveram este limite mais restritivo, para garantir o uso apenas de regiões com alta confiabilidade. Nestes dados o limite utilizado foi o de 0,11, selecionando sequências do *read* com média de qualidade igual ou superior a de Phred 20 (Tabela 1). Uma vez que a maioria dos *reads* possuía alguma sequência de bases com qualidade média que satisfez os parâmetros impostos acima, a porcentagem de *reads* aparados foi de aproximadamente 99% nos três sequenciamentos, porém com o corte das sequências de bases com baixa qualidade houve uma redistribuição do tamanho dos *reads* observado nas figuras 4 à 6, sendo que aproximadamente 1% dos *reads* foram totalmente removidos. A maioria dos *reads* produzidos nas corridas Ion Proton 6 (Figura 4) e Ion Proton 7 (Figura 5) tiveram qualidade média entre Phred 23 e 25 enquanto os produzidos pelo SOLiD possuía um valor de Phred entre 25 e 32 (Figura 6).

Tabela 01: Resultados das triagens dos sequenciamentos.

Sequenciamento	**Nº inicial de *reads***	**% *reads* duplicados**	**Nº final de reads**	**Tamanho médio dos reads (pb)**
Ion Proton 7	1.927.885	23,26	1.476.677	64,6
Ion Proton 6	221.597	20,49	175.950	63,9
SOLiD	911.118	14,66	769.508	24,6

Nº inicial de *reads* = número de *reads* fornecidos inicialmente; % *reads* duplicados = porcentagem de *reads* idênticos no dado original; Nº final de *reads* = número de *reads* resultantes após triagem; Tamanho médio dos *reads* = tamanho (número de nucleotídeos) médio dos *reads* após o processo de triagem

Após a triagem o tamanho médio dos *reads* obtidos nas corridas Ion Proton 6 e Ion Proton 7 apresentaram valores similares, sendo 63,9 e 64,6, respectivamente, enquanto SOLiD apresentou o valor de 24,6 (Tabela 1). A distribuição do tamanho dos *reads* antes e depois da fase de triagem podem ser observados nas figuras 4 a 6. Em todos os casos, houve uma redução dos tamanhos com o processo de seleção, sendo que nas corridas Ion Proton 6 (Figura 4) e Ion Proton 7 (Figura 5) a maioria dos *reads* possuía 100 a 125 pb antes da triagem, após o processo de triagem a maior parte passou a possuir um tamanho entre 50 e 65pb. Por outro lado, a maioria dos *reads* produzidos pelo SOLiD possuía 50pb, que é o tamanho máximo produzido por este seqüenciador, resultando em *reads* de tamanhos menores após seleção por qualidade.

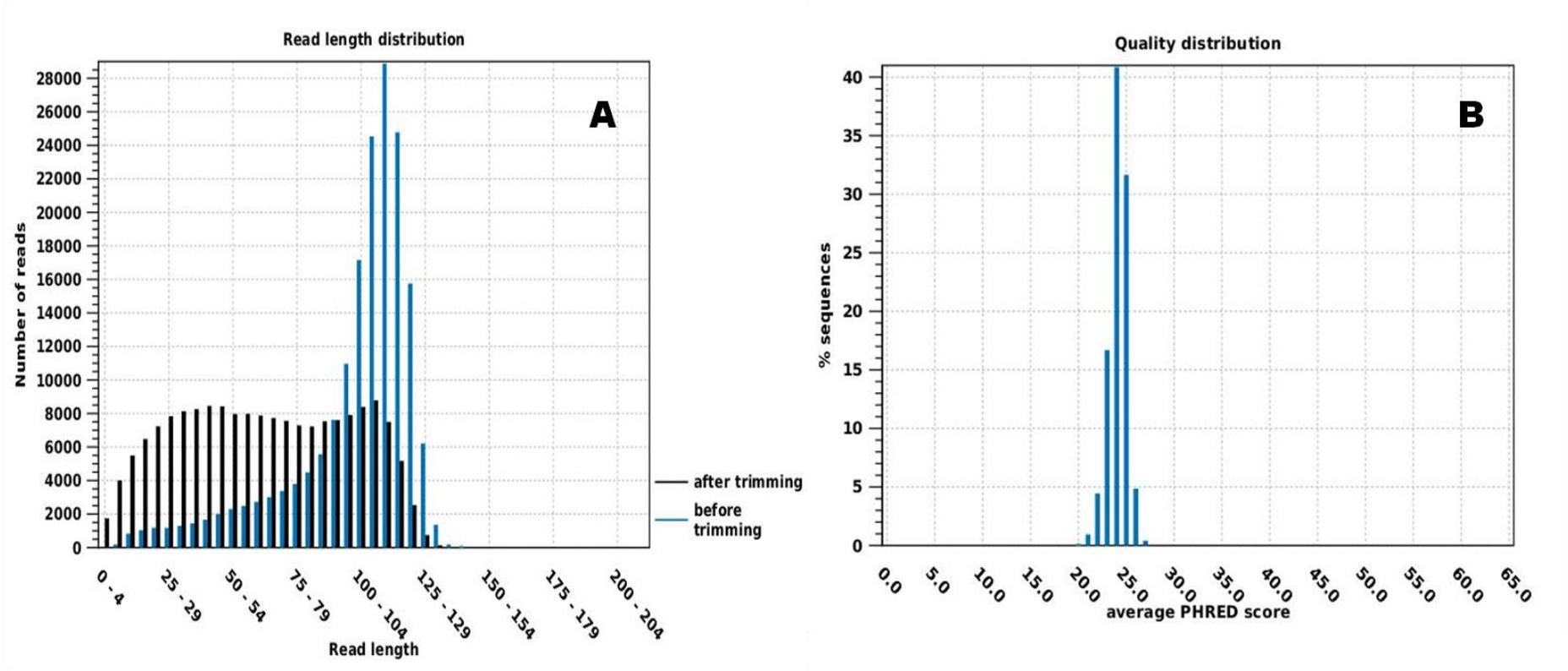

Figura 04: Gráfico das características de qualidade dos dados obtidos na corrida Ion Proton 6. (A) Distribuição dos tamanhos dos reads antes da triagem (azul) e depois da triagem (preto). Eixo X, tamanho do *read* e eixo Y, quantidade de *reads*. (B) Distribuição da qualidade de *reads* após o processo de triagem. Eixo X, qualidade média em Phred e eixo Y, porcentagem (%) dos *reads*.

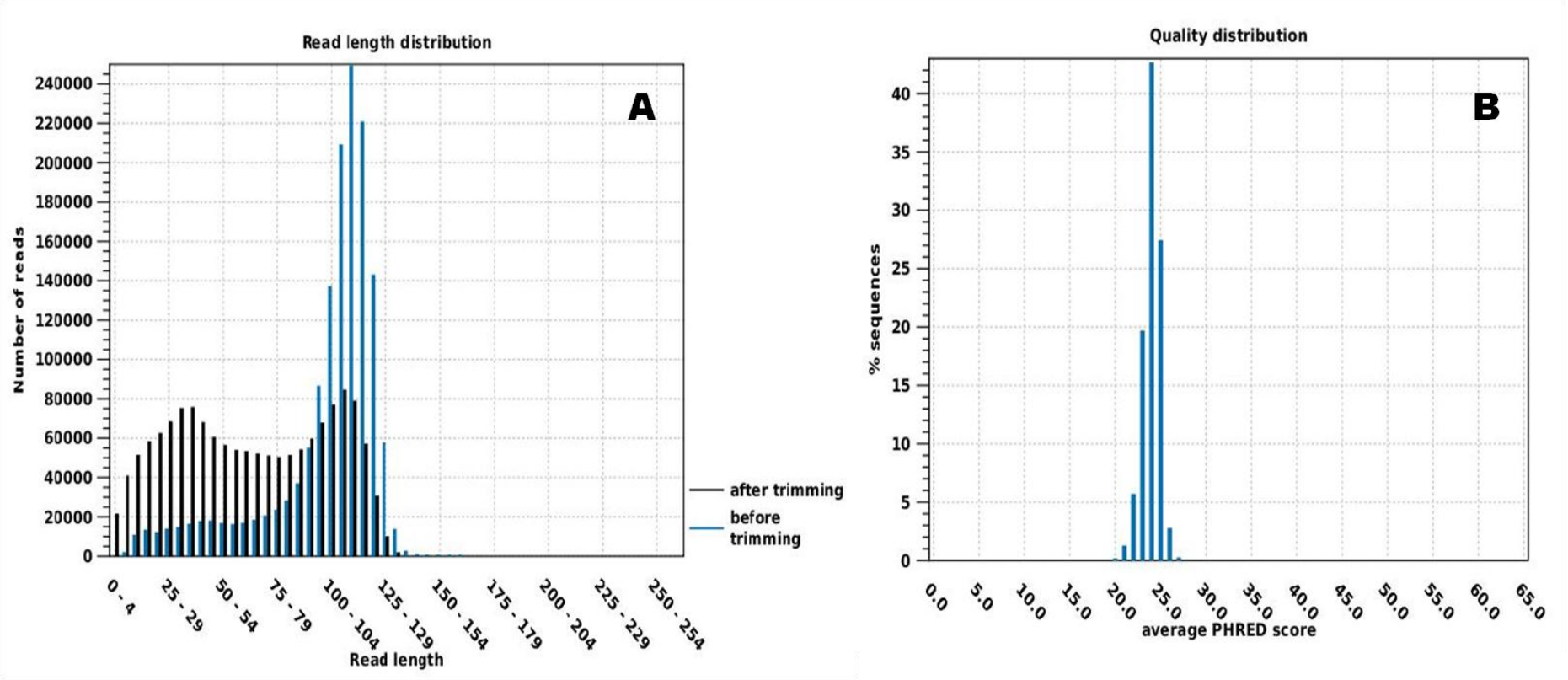

Figura 05: Gráfico das características de qualidade dos dados obtidos na corrida Ion Proton 7. (A) Distribuição dos tamanhos dos *reads* antes da triagem (azul) e depois da triagem (preto). Eixo X, tamanho do *read* e eixo Y, quantidade de *reads*. (B) Distribuição da qualidade de *reads* após o processo de triagem. Eixo X, qualidade média em Phred e eixo Y, porcentagem (%) dos *reads*.

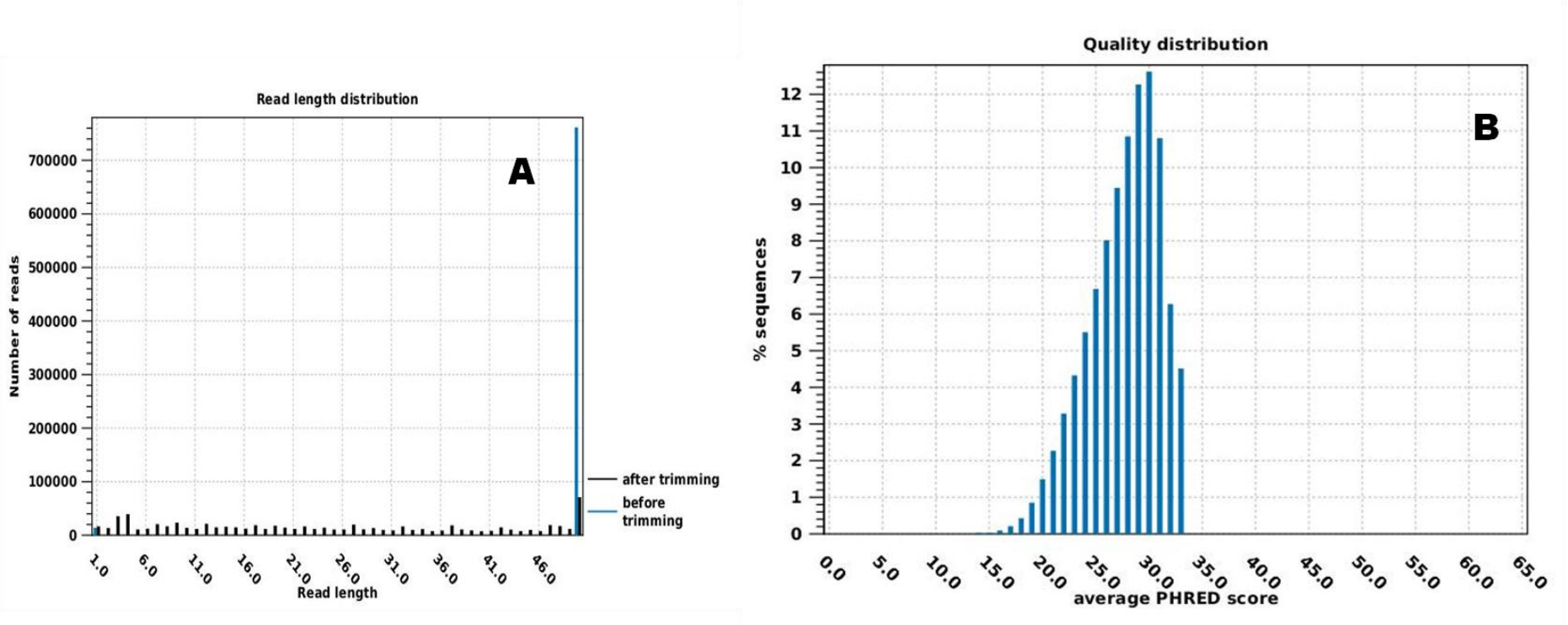

Figura 06: Gráfico das características de qualidade dos dados obtidos na corrida SOLiD. (A) Distribuição dos tamanhos dos reads antes da triagem (azul) e depois da triagem (preto). Eixo X, tamanho do *read* e eixo Y, quantidade de *reads*. (B) Distribuição da qualidade de *reads* após o processo de triagem. Eixo X, qualidade média em Phred e eixo Y, porcentagem (%) dos *reads*.

Após o processo de seleção, todos os *reads* foram mapeados contra o genoma referência da linhagem selvagem, FP2, de *A. brsilense*. Nesta etapa todos os sequenciamentos receberam parâmetros que impõem altos custos para ocorrer qualquer variação (troca, deleção e inserção) entre a leitura e o genoma referência, sendo o maior valor de penalidade que o software permite de 3. Visando a maior credibilidade na identificação de SNP's, foi imposto um valor de 3 para variação de nucleotídeo e de 2 para inserções e deleções. Além disso, foi imposta também uma porcentagem mínima de similaridade e extensão de alinhamento com o genoma de referência. Para as leituras produzidas pelo Ion Proton foram impostos os limites de 90% de similaridade e também de extensão do alinhamento. Já o sequenciamento proveniente do SOLiD, por possuir tamanho médio menor, teve seus parâmetros de extensão de alinhamento e similaridade iguais a 0,8 (Tabela 2).

Tabela 02: Parâmetros de Mapeamento.

Parâmetros	Ion Proton 6 e 7	SOLiD
Custo de variação	3	3
Custo de inserção/deleção	2	2
Extensão do alinhamento	0,9	0,8
Similaridade do alinhamento	0,9	0,8

Custo de variação = custo para que ocorra o alinhamento de um *read* que apresenta variação; Custo de inserção/deleção = custo para que ocorra o alinhamento de um *read* que apresenta inserção e/ou deleção; Extensão do alinhamento = valor mínimo de bases do read compatíveis com o genoma referência; Similaridade do alinhamento = valor mínimo de bases do read similares com o genoma referência

A sequência consenso de todas as leituras mapeadas em cada região é comparada com o genoma referência, dando uma prévia de todas as variações apresentadas. Assim, um bom alinhamento dos *reads* é crucial para a detecção de SNP's verdadeiros (KUMAR *et al*, 2012). Os sequenciamentos Ion Proton 6 e Ion Proton 7 apresentaram porcentagens de *reads* alinhados de 92,17% e 92,05% respectivamente. Entretanto, houve uma grande diferença da cobertura destes seqüenciamentos, visto que ambos possuíam número de reads totais

muito diferentes. Portanto, Ion Proton 6 apresentou uma média de cobertura de 1,45 *reads* alinhados e Ion Proton 7 possui uma média de cobertura de 12,22 *reads* alinhados. Já o sequenciamento SOLiD obteve uma porcentagem de 54,15% de alinhamento dos *reads* e com uma média de cobertura de 4,04 *reads* alinhados, o que pode explicar o tamanho médio de *reads* (35,69 pb) do sequenciamento na etapa de mapeamento, pois apesar do tamanho médio na etapa de triagem ter sido de 24,6pb os *reads* que satisfizeram os parâmetros de alinhamento possuíam um tamanho médio maior (Tabela 3).

Tabela 3: Dados dos Mapeamentos

Sequenciamento	**% *reads* alinhados**	**Tamanho médio**	**Cobertura média de *reads* alinhados**	**Cobertura máxima**
Ion Proton 6	92,17%	62,88	1,45	18
Iont Proton 7	92,05%	63,34	12,22	129
SOLiD	54,15%	35,69	4,04	147

% *reads* mapeados = porcentagem de *reads* que foram alinhados; Tamanho médio = tamanho médio dos *reads* que foram alinhados; Cobertura média de *reads* alinhados = média da cobertura de *reads* alinhados em todo o mapeamento; Cobertura máxima = número máximo de *reads* alinhados em uma região.

A partir do mapeamento foi possível observar que existiam muitas regiões com baixa ou nenhuma cobertura, o que dificulta a etapa de detecção de SNP's. Em função dessa heterogeneidade de cobertura em todos os sequenciamentos foi necessário estabelecer critérios mínimos para a identificação de SNP's, sendo estes diferentes para cada conjunto de dados. A partir do estudo de HAUER (2012) obteve-se a localização da mutação no gene *glnA* em HM053 (3.744.916 no genoma referência), e esta mutação foi utilizada como uma variação controle para o planejamento dos parâmetros individuais de cada seqüenciamento. Sendo assim, na etapa de detecção das variações de nucleotídeos, os parâmetros foram estabelecidos de forma que melhor se

adaptassem ao volume, tamanho, alinhamento e qualidade dos *reads* de cada sequenciamento visando a detecção da mutação em *glnA* (Tabela 4).

Tabela 04: Parâmetros de detecção de SNP's.

PARAMETROS	Ion Proton 6	Ion Proton 7	SOLiD
Raio de vizinhança	10	10	5
Máximo de variações	2	2	2
Qualidade mínima da vizinhança (Phred)	20	20	20
Qualidade mínima central (Phred)	20	20	15
Cobertura mínima	4	16	5
Mínimo de leituras concordantes	100,00%	100,00%	100,00%

Raio de vizinhança = número de nucleotídeos analisados vizinhos ao que apresenta variação; Máximo de variações = Número máximo de variações nucleotídicas permitidas; Qualidade mínima da vizinhança (Phred) = qualidade (Phred) mínima exigida em todos os nucleotídeos no raio de vizinhança do nucleotídeo analisado; Qualidade mínima central = qualidade mínima exigida no nucleotídeo que se encontra variando; Cobertura mínima = mínimo de *reads* alinhados que apresentem a variação; Mínimo de leituras concordantes = frequência mínima da variação em todos os *reads* de cobertura

Em função das características de cada conjunto de dados e dos parâmetros para detecção de SNP's, cada sequenciamento apresentou resultados distintos. Ion Proton 6 (Tabela 5) apresentou 20 variações com uma faixa de cobertura de 4 à 7 *reads*, e uma média de qualidade de aproximadamente Phred 25. Já Ion Proton 7 apresentou 37 variações com uma faixa de cobertura de 16 à 31 reads, e uma média de qualidade superior a Phred 25 (Tabela 6). E por último, o sequenciamento SOLiD apresentou 20 variações com uma cobertura de 5 à 26 reads, com uma qualidade média acima de Phred 25 (Tabela 7).

Quando todas as tabelas foram comparadas para encontrar alterações em comum não foi encontrada nenhuma variação (troca, inserção e deleção)

compartilhada entre os três seqüenciamentos, dentro dos parâmetros utilizados. Porém foram encontradas variações compartilhadas entre o sequenciamento Ion Proton 7 e SOLiD, e entre os sequenciamentos Ion Proton 6 e Ion Proton 7 (Tabela 8).

Uma vez que apenas duas variações foram compartilhadas entre Ion Proton 7 e SOLiD, as variações detectadas pelo Ion Proton 7 foram verificadas no mapeamento do sequenciamento SOLiD, para avaliar sua presença ou ausência. Foi possível observar 16 SNP's compartilhadas entre os dois sequenciamentos, porém a baixa cobertura de *reads* alinhados nos dados gerados pelo SOLiD foi a principal causa da não detecção (Tabela 9).

Para a complementação das informações foram selecionadas seis variações para a confecção de *primers* (Tabela 9). Porém até a conclusão do presente trabalho a empresa responsável pela síntese dos *primers* ainda se encontrava com a entrega em atraso.

Tabela 05: Variações nucleotídicas em Ion Proton 6.

Posição no genoma referência	Tipo da variação	Base referência	Número de variações	Alelo variante	Cobertura de *reads*	Média de qualidade Phred	Variação do aminoácido
6.440	SNV	C	1	A	7	26.4	Ala → Asp
121.400	SNV	C	1	A	4	26.0	Região intergênica
215.526	SNV	T	1	G	6	25.0	Gln → His
734.434	SNV	T	1	C	6	23.8	L
1.377.181	SNV	G	1	T	4	24.8	Gly → Val
1.712.110	SNV	C	1	G	7	24.0	Região intergênica
2.413.655	SNV	G	1	A	4	25.3	Gly → Ser
2.413.665	SNV	G	1	A	4	24.8	Arg → Gln
2.413.676	SNV	G	1	A	4	24.3	Gly → Ser
2.940.730	SNV	T	1	G	4	25.5	Glu → Ala
3.010.934	InDel	G	1	-	4	16.8	FS
3.611.671	InDel	G	1	-	4	22.5	Região intergênica
3.744.916	SNV	C	1	T	4	25.3	Pro → Leu
3.783.234	SNV	C	1	A	4	25.3	L
3.783.240	SNV	C	1	T	5	27.2	L
3.963.895	SNV	G	1	C	4	22.5	L
5.192.638	SNV	A	1	G	4	25.3	Ile→ Val
5.847.350	SNV	T	1	C	4	24.5	L
5.869.520	SNV	C	1	A	4	26.0	Gly → Cys
6.051.986	SNV	G	1	T	4	26.0	Gly → Val

Posição no genoma referência = Posição do nucleotídeo no genoma referência FP2; Tipo da variação = Variação simples de nucleotídeo (SNV), Inserção/Deleção (InDel), Múltipla variação de nucleotídeo (MNV); Alelo referência = Base encontrada na posição do genoma referência; Número de variações = Quantas variações ocorreram nessa base; Alelo variante = Base variante gerada pelo consenso de *reads*; Cobertura de *reads* = Quantidade de *reads* alinhados que apresentam a variação; Média de qualidade = Média de qualidade da base variante, entre os *reads* alinhados; Variação do Aminoácido = Alteração de aminoácido conseqüência da variação: Frameshift (FS), Quando existe uma lacuna de nucleotídeo na região, no genoma referência (L)e região entre genes (Região intergênica).

Tabela 06: Variações nucleotídicas em Ion Proton 7.

Posição no genoma referência	Tipo da variação	Base referência	Número de variações	Alelo variante	Cobertura de *reads*	Média de qualidade Phred	Variação do aminoácido
6.440	SNV	C	1	A	22	25.9	Ala → Asp
122.579	SNV	A	1	G	28	24.6	L
505.734	InDel	0	1	A	25	26.2	Região intergênica
505.735	InDel	A	1	0	26	24.6	Região intergênica
911.747	SNV	C	1	A	19	26.0	Gly → Val
1.270.732	SNV	C	1	A	20	25.8	Asp → Tyr
1.441.983	SNV	G	1	T	16	26.1	Pro → Thr
1.506.977	SNV	C	1	A	22	26.0	Gly → Val
2.024.828	InDel	T	1	0	17	19.3	FS
2.087.859	SNV	G	1	A	24	25.8	Gly → Asp
2.087.868	SNV	G	1	T	26	26.0	Gly → Val
2.304.873	SNV	C	1	T	21	25.4	Arg → Gln
2.576.045	InDel	G	1	0	16	20.2	FS
2.710.923	SNV	C	1	G	18	23.1	Pro → Ala
2.842.308	SNV	C	1	A	17	26.5	L
2.874.175	SNV	G	1	C	16	24.3	L
3.082.649	SNV	G	1	T	17	25.7	Pro → Thr
3.082.651	SNV	G	1	T	17	26.3	Pro → His
3.143.861	InDel	C	1	0	16	24.0	FS
3.244.153	SNV	T	1	A	18	25.4	Glu → Val
3.380.970	SNV	C	1	A	24	24.2	L
3.428.103	SNV	A	1	C	20	24.2	Asp → Ala
3.472.497	SNV	C	1	A	20	26.1	Gly → Val
3.702.929	MNV	AC	1	GT	20	26.0	His → Met/ His → Tyr
3.744.916	SNV	C	1	T	16	24.4	Pro → Leu

Posição no genoma referência	Tipo da variação	Base referência	Número de variações	Alelo variante	Cobertura de *reads*	Média de qualidade Phred	Variação do aminoácido
3.783.234	SNV	C	1	A	16	25.0	L
3.783.240	SNV	C	1	T	19	25.5	L
4.339.334	SNV	G	1	T	25	25.5	L
4.656.008	SNV	A	1	G	19	25.7	Região intergênica
4.772.490	SNV	A	1	C	29	23.0	Ser → Arg
4.772.501	SNV	C	1	G	22	25.5	Asp → Glu
5.242.677	SNV	C	1	A	31	25.7	Ser → Tyr
6.018.707	SNV	A	1	C	16	24.5	Gln
6.158.969	SNV	C	1	G	17	25.0	Ser → Thr
6.367.074	SNV	G	1	A	30	25.5	Gly → Ser
6.427.010	SNV	G	1	A	17	25.0	Região intergênica
6.589.980	InDel	0	1	A	17	25.8	Thr

Posição do nucleotídeo no genoma referência = Posição do nucleotídeo no genoma referência FP2; Tipo de variação nucleotídica = Variação simples de nucleotídeo (SNV), Inserção/Deleção (InDel), Multipla variação de nucleotídeo (MNV); Alelo referência = Base encontrada na posição do genoma referência; Número de variações = Quantas variações ocorreram nessa base; Alelo variante = Base variante gerada pelo consenso de *reads*; Cobertura de *reads* = Quantidade de *reads* alinhados que Apresentam a variação; Média de qualidade = Média de qualidade entre os *reads* de cobertura; Variação do Aminoácido = Troca de aminoácido com a variação. Sendo que frameshift (FS), Quando existe uma lacuna de nucleotídeo na região, no genoma referência (L) e região entre genes (Região intergênica).

Tabela 07: Variações nucleotídicas SOLiD.

Posição no genoma referência	Tipo da variação	Base referência	Número de variações	Alelo variante	Cobertura de *reads*	Média de qualidade Phred	Variação do aminoácido
17.427	SNV	G	1	A	9	28.1	L
17.430	SNV	C	1	G	9	31.9	L
17.452	SNV	C	1	T	7	28.9	Região intergênica
52.404	SNV	C	1	T	6	28.5	Região Intergênica
476.110	SNV	G	1	T	10	31.2	Gly → Val
1.255.992	SNV	A	1	G	5	31.4	Asp → Gly
1.329.790	SNV	A	1	T	12	28.8	Phe → Tyr
1.329.792	SNV	C	1	T	12	29.9	Região Intergênica
1.329.820	SNV	C	1	A	26	31.5	Gly → Val
1.655.550	InDel	Y	1	0	7	28.1	FS
1.671.423	SNV	C	1	A	6	28.2	Ser → Tyr
1.722.557	InDel	Y	1	0	10	27.5	Região Intergênica
1.722.567	SNV	C	1	T	14	31.1	Região intergênica
2.842.217	SNV	G	1	C	5	32.8	Pro → Ala
2.842.220	SNV	G	1	T	5	30.2	Pro → Thr
4.572.481	SNV	T	1	C	10	30.7	L
4.656.008	SNV	A	1	G	5	26.8	Região intergênica
4.772.501	SNV	C	1	G	6	32.2	Asp → Glu
4.858.880	SNV	C	1	T	5	30.6	Ser → Asn
5.099.235	SNV	A	1	G	6	30.8	Thr → Ala

Posição do nucleotídeo no genoma referência = Posição do nucleotídeo no genoma referência FP2; Tipo de variação nucleotídica = Variação simples de nucleotídeo (SNV), Inserção/Deleção (InDel), Multipla variação de nucleotídeo (MNV); Alelo referência = Base encontrada na posição do genoma referência; Número de variações = Quantas variações ocorreram nessa base; Alelo variante = Base variante gerada pelo consenso de *reads*; Cobertura de *reads* = Quantidade de *reads* alinhados que Apresentam a variação; Média de qualidade = Média de qualidade entre os *reads* de cobertura; Variação do Aminoácido = Troca de aminoácido com a variação. Sendo que frameshift (FS), Quando existe uma lacuna de nucleotídeo na região, no genoma referência (L) e região entre genes (Região intergênica).

Tabela 08: Variações compartilhadas entre os diferentes sequenciamentos

Posição no genoma referência	**Variação no Ion Proton 7**	**Variação no SOLiD**	**Região**	**Variação do Aminoácido**
4.656.008	A – G	A – G	Região não codificadora	Região Intergênica
4.772.501	C – G	C – G	Proteína conservada sem função	Asp → Glu
Posição no genoma referência	**Variação no Proton 7**	**Variação no Proton 6**	**Região**	**Variação do Aminoácido**
6.440	C - A	C - A	*COG0243: Anaerobic dehydrogenases,*	Ala → Asp
3.744.916	C – T	C – T	*glutamine synthetase*	Pro → Leu
3.783.234	C – A	C – A	*conserved hypothetical protein*	L
3.783.240	C – T	C – T	*conserved hypothetical protein*	L

Posição no genoma referência = posição do nucleotídeo no genoma referência FP2; Variação no Ion Proton 7 = variação nucleotídica apresentada no Ion Proton 7; Variação no SOLiD = variação nucleotídica apresentada no SOLiD; Variação no Ion Proton 6 = variação nucleotídica apresentada no Ion Proton 6; Região = Gene que a região representa; Variação do Aminoácido = Troca de aminoácido com a variação. Sendo que frameshift (FS), Quando existe uma lacuna de nucleotídeo na região, no genoma referência (L) e região entre genes (Região intergênica).

Tabela 9: Tabela de observação de cobertura

Posição no genoma referência	**Variação no Ion Proton 7**	**Variação no Solid**	**Região**	**Número de cobertura de *reads* pelo Mapeamento SOLiD**
6.440	C – A		*COG0243: Anaerobic dehydrogenases, typically*	0
122.579	A – G		*hypothetical protein FrCN3DRAFT_6881*	0
911.747	C – A	C – A	*exported protein of unknown function; putative*	1
1.270.732	C – A		*hypothetical protein HMPREF1318_0759*	0
1.441.983	G – T	G – T	*putative Peroxide-responsive repressor (PerR-like),*	1
1.506.977 *	C – A		*GntR family transcriptional regulator*	0
2.087.859	G – A		*2-oxo/hydroxy acid reductase*	0
2.087.868	G – T		*2-oxo/hydroxy acid reductase*	0
2.304.873	C – T	-	*putative transcriptional regulator (ArsR family) with*	0
2.710.923*	C – G	-	*amino acid adenylation*	0
2.842.308	C – A	-	*phosphoadenosine phosphosulfate reductase*	0
2.874.175	G – C	G – C	*putative dipeptide/oligopeptide/nickel ABC*	1
3.082.649	G – T	-	*conserved exported protein of unknown function*	0
3.082.651	G – T	-	*conserved exported protein of unknown function*	0
3.244.153	T – A	-	*putative Heat shock protein*	0
3.380.970	C – A	-	*Região não codificadora*	0
3.428.103	A – C	A – C	*chloride peroxidase*	2
3.472.497	C – A	C – A	*cobT gene product*	3
3.744.916	C – T	C – T	*glutamine synthetase*	3
3.783.234	C – A	C – A	*conserved hypothetical protein*	2
3.783.240	C – T	C – T	*conserved hypothetical protein*	1
4.339.334	G – T	G – T	*iron complex outermembrane receptor protein*	1
4.656.008*	A – G	A – G	*Região não codificadora*	1
4.772.490	A – C	A – C	*conserved protein of unknown function (DUF520)*	4
4.772.501*	C – G	C – G	*conserved protein of unknown function (DUF520)*	6

Posição no genoma referência	**Variação no Ion Proton 7**	**Variação no Solid**	**Região**	**Número de cobertura de *reads* pelo Mapeamento SOLiD**
5.242.677	C – A	-	*cell division protein FtsZ*	0
6.018.707*	A – C	A – C	*enoyl-CoA hydratase*	4
6.158.969	C – G	C – G	*FlbD*	1
6.367.074*	G – A	G – A	*branched-chain amino acid ABC*	4
6.427.010	G – A	G – A	*Região não codificadora*	2

Posição no genoma referência = posição do nucleotídeo no genoma referência FP2; Variação no Ion Proton 7 = variação nucleotídica apresentada no Ion Proton 7; Variação no SOLiD = variação nucleotídica apresentada no SOLiD, quando não ocorreu variação (-); Região = Gene que a região representa; Número de cobertura de *reads* pelo Mapeamento SOLiD = Número de *reads* alinhados na região da variação pelo mapeamento SOLiD

* variação selecionada para confecção de *primers*

6. DISCUSSÃO

A detecção de *Single Nucleotide Polymorphism* (SNP), ou, polimorfismo de nucleotídeo individual ou único, no genoma de HM053 pode tornar possível a identificação das variações que poderiam explicar o fenótipo da estirpe mutante HM053. Para analisar estes possíveis SNPs foi realizada uma análise dos dados produzidos por dois tipos de sequenciadores de nova-geração (SOLiD e Ion Proton), estes dados foram comparados com o genoma referência da estirpe FP2 de *A. brasilense*, utilizando o software CLC genome MainWorkbench.

O estudo de HAUER (2012) demonstrou a presença de uma mutação na GS de HM053 através do método de sequenciamento de Sanger, este mesmo estudo demonstrou qual seria a localização desta mutação no gene *glnA*, e a partir desta informação foi possível saber qual é a localização da base no genoma utilizado como referência da estirpe FP2, e se esta mutação foi encontrada nos três sequenciamentos de HM053.

O dado analisado do sequenciamento SOLiD não selecionou a mutação no gene *glnA* pela ferramenta de detecção de variações, este fato ocorreu pela baixa cobertura de *reads* na região porém com a observação manual no mapeamento foi possível observar a presença desta SNP com 3 *reads* de cobertura. Enquanto os dados dos sequenciamentos Ion Proton 6 e Ion Proton 7 apresentaram a detecção da mutação, sendo que em Ion Proton 6 a cobertura foi de 4 *reads* e a média de qualidade dos *reads* em Phred de 25.3, enquanto Ion Proton 7 apresentou uma cobertura de 16 *reads* com media de qualidade destes de 24.4.

A capacidade de fixar nitrogênio constitutivamente (Nif^C) e de excretar amônio (MACHADO 1988) em HM053, acredita-se ser decorrente do acúmulo de amônio intracelular que não é assimilado, assim a mutação presente em GS seria uma das causas do fenótipo da estirpe, uma vez que a glutamina sintetase é a principal enzima responsável pela assimilação de amônio. Porém apenas esta mutação pode não ser a única explicação para o fenótipo da estirpe mutante, existindo a possibilidade de que outros genes tenham sofrido mutação. VITORINO (2001) teorizou que os genes mais cotados para explicar

o fenótipo de excreção de amônio seriam *glnD*, *glnE*, *glnB* e *glnZ*, uma vez que todos estão relacionados com a regulação da atividade de GS, mas em seus estudos não foram encontradas mutações em *glnB* e nem em *glnZ*.

O trabalho realizado também não detectou mutações nestes genes, KUMAR *et al* (2012) explica que a detecção de SNP's utilizando somente análise computacional de dados provenientes de NGS (*next-generation sequencing*) possuem diversas dificuldades, como as encontradas neste trabalho, entre elas um volume de dados baixo aonde nenhum sequenciamento ultrapassou uma quantidade superior a 2.000.000 de *reads*. No entanto os parâmetros seletivos de qualidade e mapeamento como os apresentados neste trabalho, garantiram que todas as variações detectadas possuíssem um bom grau de confiabilidade, sendo 20 variações detectadas em Ion Proton 6 e SOLiD, e 37 variações detectadas em Ion Proton 7. A presença de regiões com baixa cobertura criou uma dificuldade para que todos os sequenciamentos apresentassem SNP's em comum com base na ferramenta de detecção, sendo 2 variações compartilhadas entre SOLiD e Ion Proton 7, e 4 compartilhadas entre Ion Proton 6 e Ion Proton 7.

Para sanar a não detecção em regiões com baixa cobertura, foi realizada uma comparação manual, aonde as SNP's detectadas em Ion Proton 7 foram observadas manualmente no mapeamento do sequenciamento SOLiD, afim de confirmar ou não sua presença. Sendo possível observar 16 variações não detectadas presentes em SOLiD, que eram compartilhas pelo sequenciamento Ion Proton 7. Este resultado indica a confirmação de SNP's compartilhadas em sequenciamentos com diferentes tecnologias, e somado a isto algumas destas SNP's estão presentes em regiões de genes importantes, que poderiam estar relacionados com o fenótipo do mutante, como: *putative Peroxide-responsive repressor (PerR-like)* e *iron complex outermembrane receptor protein*, mas não foram encontradas mutações nas regiões de regulação de GS. Portanto se faz necessário um aprofundamento da ligação das proteínas, que apresentam SNP's, com o fenótipo apresentado na estirpe mutante HM053.

7. CONCLUSÕES

A análise dos dados dos sequenciamentos da estirpe mutante HM053 de *Azospirillum brasilense*, utilizando o software CLC genome ManWorkbench, confirmou a presença da mutação na enzima glutamina sintetase (GS) descrita por HAUER (2012) em dados provenientes de dois tipos de sequenciadores de nova geração (SOLiD e Ion Proton), a presença desta mutação em todas as análises corrobora sua existência, confirmando a importância da mutação para o fenótipo presente em HM053 de excreção de amônio e fixação constitutiva de nitrogênio, visto a importância da enzima na biossíntese do amônio intracelular no processo de fixação de nitrogênio.

Além disso, as análises dos dados demonstraram a existência de outras possíveis SNP's compartilhadas pelos dois tipos de sequenciadores, que utilizam diferentes tecnologias de sequenciamento, sendo que algumas destas SNP's carecem de estudos mais aprofundados, pois também podem ter ligação com o fenótipo da estirpe mutante, tais como: *putative Peroxide-responsive repressor (PerR-like)* e *iron complex outermembrane receptor protein* .

Portanto a utilização de ferramentas de bioinformática como o CLC genome ManWorkbench demonstrou grande potencial para análise de SNP's utilizando dados provenientes de sequenciadores com tecnologias distintas.

8. REFERÊNCIAS BIBLIOGRÁFICAS

ALTSHULER, D.; POLLARA, V. J.; COWLES, C. R.; ETTEN, W. J. V.; BALDWIN, J., LINTON.; L.; LANDER, E. S. An snp map of the human genome generated by reduced representation shotgun sequencing. **Nature**, v.407, p.513-516, 2000.

ANTONYUK, L.P. Glutamine Synthetase of the Rhizobacterium *Azospirillum brasilense*: Specific Features of Catalysis and Regulation. **Applied Biochemistry and Microbiology**., v.43, no3, p.244-249, 2007.

ARAUJO, M. S.; BAURA, V. A.; SOUZA, E.M.; BENELLI, E.M. RIGO, L. U.; STEFFENS, M. B. R.; PEDROSA, F. O.; CHUBATSU, L. S. In vitro uridylation of the *Azospirillum brasilense* N-signal transducing GlnZ protein. **Protein Expression and Purification** V.33, p. 19-24, 2004.

ARSENE, F.; KAMISNKI, P. A.; ELMERICH, C. Modulation of NifA activity by PII in *Azospirillum brasilense*: evidence for a regulatory role of NifA N-teminal domain. **Journal of Bacteriology**. v. 178, p. 4830-4838, 1996.

BOZOUKLIAN, H. ; ELMERICH, C. Nucleotide sequence of the *Azospirillum brasilense* Sp7 glutamine synthetasse structural gene. **Biochimie**., v.68, p.1181-1187, 1986.

DE MEL, V. S. J.; KAMBEROV, E. S.; MARTIN, P. D.; ZHANG, J.; NINFA, A. J.; EDWARDS, B. F. P. Preliminary X-ray diffraction analysis of crystals of the PII protein from *Escherichia coli*. **J. Mol. Biol**., v.243, p. 796-798, 1994.

DIXON, R.; KAHN, D. Genetic regulation of biological nitrogen fixation**. Nature Reviews Microbiology**., v.2, p.621-631, 2004.

DOMMELEN, A. V.; KEIJERS, V.; WOLLEBRANTS, A.; VANDERLEYDEN, J. Phenotypic Changes Resulting from Distinct Point Mutations in *the Azospirillum brasilense glnA* Gene, Encoding Glutamine Synthetase. **Applied and Environmental Microbiology**, v.69, nº9, p.5699-5701, 2003.

DRAZEN, JM.; YANDAVA, CN.; DUBE, L.; et al. Pharmacogenetic association between ALOX5 promoter genotype and the response to anti-asthma treatment. **Nat Genet**, v. 22, p.168-170, 1999.

FORCHHAMMER, K. Glutamine signaling in bacteria. **Frontiers in Bioscience**, v.12, p.358-370. 2007.

GANAL, MW.; ALTMANN, T.; RODER, MS. SNP identification in crop plants. **Plant Biol**., v. 12(2), p. 211-7, 2009.

HAUER, Vanessa. **SEQUENCIAMENTO DO GENE glnA DAS ESTIRPES MUTANTES HM14, HM26, HM053 E HM210 DE *Azospirillum brasilense***. Curitiba, 2012. Monografia (Ciêcnais Biológicas) - Setor de Ciências Biológicas, Universidade Federal do Paraná.

HUERGO, LF.; MONTEIRO, RA.; BOATTO, AC.; et al. Regulation of Nitrogen Fixation in *Azospirillum Brasilense*. ***Azospirillum* sp.: cell physiology, plant interactions and agronomic research in Argentina**, 2008.

HUNGRIA, M. Inoculação com *Azospirillum brasiliense*: inovação em rendimento a baixo custo**. Embrapa Soja** (Londrina), no325, 2011. 36p.

KHAMMAS, K. M.; AGERON, E.; GRIMONT, P. A. D.; KAISER, P*. Azospirillum irakense* sp. nov., a nitrogen-fixing bacterium associated with rice roots and rhizosphere soil. **Research Microbiology**. v. 140, p. 679-694, 1989.

KRAWEZAK, M, REISS, J, COOPER, DN, The mutational spectrum of single base-pair substitutions in mRNA splice junctions of human genes: causes and consequences. **Hum-Genet**, v. 90, p. 41-54, 1992.

KUMAR, s.; BANKS, T.; CLOUTIER S. SNP Discovery throught Next-Generation Sequencing and Its Applications. **International Journal of Plant Genomics**. v.2012, article ID 831460, 2012.

LEIGH, J. A.; DODSWORTH, J. A. Nitrogen Regulation in Bacteria and Archaea. **Annual Review of Microbiology**, v.61, p.349-377, 2007.

LIAO, P.; LEE, KH. From SNPs to functional polymorphism: The insight into biotechnology applications. **Biochemical Engineering Journal**. v.49, p. 149-158, 2010.

MACHADO, H. B**. Isolamento e caracterização de mutantes de *Azospirillum brasilense* constitutivos para fixação de nitrogênio**. 128f. Dissertação (Mestrado em Ciências-Bioquímica) – Departamento de Bioquímica, Universidade Federal do Paraná. Curitiba, 1988.

MAGALHAES, F. M. M.; BALDANI, J. I.; SOUTO, S. M.; KUYKENDALL, J. DOBEREINER, J. A new acid-tolerant *Azospirillum* species. **Anais Acadêmia Brasileira de Ciências**. v. 55, p. 471-430, 1983.

Manual online do **CLC Main Workbench**, acessado em, 03/06/2013 (www.clcsupport.com).

MARTIN-DIDONET, C. C. G.; CHUBATSU, L. S.; SOUZA, E. M.; KLEINA, M.; REGO, F. G. M.; RIGO, L. U.; GEOFFREY YATES, M.; PEDROSA, F. O. Genome Structure of the Genus Azospirillum. **Journal of Bacteriology**, v. 182, no14, p. 4113–4116. 2000.

MERRICK, M. J.; EDWARDS, R. A. Nitrogen Control in Bacteria. **Microbiological Reviews**, v.59, nº 4, p.604-622, 1995.

REINHOLD, B.; HUREK, T.; FENDRIK, I.; POT, B.; GILLIS, M.; KERSTERS, K.; THIELEMANS, S.; DE LEY, J. *Azospirillum halopraeferens* sp. nov., a nitrogen-fixing organism associated with roots of kallar grass (*Leptchloa fusca*) **International Journal of Systematic Bacteriology**. v. 37, p. 43-51, 1987.

SOLiD™ 3 Plus to SOLiD™ 4 System User Documentation Changes. **Applied Biosystems SOLiD™ System**, 2010.

SRIVASTAVA, A.; TRIPATHI, A. K. Adenosine Diphosphate Ribosylation of Dinitroge-nase Reductase and Adenylylation of Glutamine Synthetase Control Ammonia Excretion in Ethylenediamine-Resistent Mutants of Azospirillum brasilense Sp7. **Current Microbiology**, v.53, p.317-323. 2006.

STEENHOUDT, O.; VANDERLEYDEN, J. Azospirillum, a free-living nitrogen-fixing bacterium closely associated with grasses: genetic, biochemical and ecological aspects. **FEMS Microbiology Reviews**. v. 24, no4, p. 487-506, 2000.

TARRAND, J.J.; KRIEG, lipoferum group with description of a new genus, *Azospirillum* gen. Nov. and two species, *Azospirillum lipoferum* (Beijerinck) comb. nov. and *Azospirillum brasilense* sp. nov. **Canadian Journal of Microbiology**. v. 24, p. 967-980, 1978.

ZAMAROCZY, de M.; PAQUELIN, A.; ELMERICH, C. Functional Organization of the glnB-glnA Cluster of *Azospirillum brasilense*. **Journal of Bacteriology**, v. 175, nº9, p. 2507-2515, 1993.

VITORINO, J. C.; STEFFENS B. R.; MACHADO, H. B.; YATES, G.; SOUZA, E. M.; PEDROSA, F. O. Potential roles for the *glnB* and *ntrXY* genes in *Azospirillum brasilense*. **FEMS Microbiology Letters**, 201, 199-204, 2001.