Universidade Federal do Pará

Instituto de Ciências Biológicas

Programa de Pós-graduação em Biotecnologia

AVALIAÇÃO DO VIÉS GC EM PLATAFORMAS DE SEQUENCIAMENTO DE NOVA GERAÇÃO

Kenny da Costa Pinheiro

Belém - Pará
Março de 2015

Universidade Federal do Pará
Instituto de Ciências Biológicas
Programa de Pós-graduação em Biotecnologia

AVALIAÇÃO DO VIÉS GC EM PLATAFORMAS DE SEQUENCIAMENTO DE NOVA GERAÇÃO

Kenny da Costa Pinheiro

Plano de defesa submetido ao Programa de Pós Graduação em Biotecnologia da UFPA para obtenção do grau de Mestre em Biotecnologia

Orientador: Dr. Rommel Thiago Jucá Ramos

Belém – Pará
Março de 2015

Dados Internacionais de Catalogação-na-Publicação (CIP)
Sistema de Bibliotecas da UFPA

Pinheiro, Kenny da Costa, 1980-
Avaliação do viés GC em plataformas de sequenciamento de nova geração / Kenny da Costa Pinheiro. - 2015.

Orientador: Rommel Thiago Jucá Rammos.
Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Ciências Biológicas, Programa de Pós-Graduação em Biotecnologia, Belém, 2015.

1. Bioinformática. 2. Corynebacterium pseudotuberculosis. 3. Genoma. I. Título.

CDD 22. ed. 570.285

AGRADECIMENTOS

À Universidade Federal do Pará e ao Programa de Pós Graduação em Biotecnologia pela oportunidade e aprendizado.

À FAPESPA pela bolsa de pesquisa.

Ao professor Dr. Arthur Silva pela oportunidade oferecida e confiança depositada em meu trabalho, ao fazer parte da excelente equipe de profissionais que formam o LPDNA.

Ao professor Dr. Rommel Ramos, por ter sido meu orientador, pela confiança e ajuda que me deu durante o mestrado visando sempre meu crescimento pessoal e profissional, e por ter sido meu grande amigo nesta jornada.

Aos amigos Allan Veras e Pablo Caracciolo, pela paciência e compreensão em repassar seus conhecimentos de bioinformática, me auxiliando nos momentos difíceis, compartilhando laços de amizade e contribuindo de forma grandiosa para a conclusão deste trabalho.

A todos os amigos do laboratório que depositaram sua confiança em mim e que compartilharam seus conhecimentos possibilitando que eu amadurecesse profissionalmente e me permitindo valorizar o trabalho em equipe.

Ao meu pai que sempre me ajudou a superar os desafios que surgiram dando incentivo e forças para continuar. Por ter sido um grande pai me ensinando valores corretos, justos e honrados para alcançar minhas metas.

À minha mãe um agradecimento especial, sem a qual eu não estaria aqui agora, por tudo que me ensinou e por ter me dado a oportunidade de realizar meus sonhos

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

NGS	Sequenciadores de próxima geração (“Next Generation Sequencing”)
PGM	Sequenciador produzido pela Life Technologies (“Personal Genome Machine”)
ddNTP	didesoxinucleotídeos
INDEL	Erro de sequenciamento onde pode ocorrer uma inserção ou deleção de uma base (“Insertion and Deletion”)
DNA	Ácido desoxirribonucléico
SMRT	Sequenciamento de molécula única de DNA (“Single Molecule Real Time”)
CSV	Formato tabular para planilhas (“Comma-separated Values”)
SRA	Arquivo de leituras sequenciadas depositados no NCBI (“Sequence Read Archive”)
BWA	Algoritmo para alinhamento de sequências (“Burrows wheeler align”)

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

O surgimento das plataformas de sequenciamento de nova geração (NGS) proporcionou o aumento do volume de dados produzidos, tornando possível a obtenção de genomas completos. Apesar das vantagens alcançadas com estas plataformas, são observadas regiões de elevada ou baixa cobertura, em relação à média, associadas diretamente ao conteúdo GC. Este viés GC pode afetar análises genômicas e dificultar a montagem de genomas através da abordagem *de novo*, além de afetar as análises baseadas em referência. Além do que, as maneiras de avaliar o viés GC deve ser adequada para dados com diferentes perfis de relação/associação entre GC e cobertura, tais como linear e quadrático.

Desta forma, este trabalho propõe o uso do Coeficiente de Correlação de Pearson (*r*) para analisar a correlação entre conteúdo GC e Cobertura, permitindo identificar a intensidade da correlação linear e detectar associações não-lineares, além de identificar a relação entre viés GC e as plataformas de sequenciamento. Os sinais positivos e negativos de *r* também permitem inferir relações diretamente proporcionais e inversamente proporcionais respectivamente. Utilizou-se dados da espécie *Corynebacterium pseudotuberculosis*, conhecido por serem genomas clonais obtidas através de diferentes tecnologias de sequenciamento para identificar se há relação do viés GC com as plataformas utilizadas.
Palavras-chave: Viés GC, Pearson, Quadrático, Correlação.

ABSTRACT

The emergence of high throughput sequencing (HTS) platforms increased the amount of data making feasible to obtaining complete genomes. Despite the advantages and the throughput produced by these platforms, the high or low genomic coverage in the regions of the genome can be related to GC content. This GC bias may affect genomic analyzes and the genomic/transcriptomic analysis based on *de novo* and reference approach. In addition, the ways to evaluate the GC bias should be fit to data with different profiles of the GC vs coverage relationship, such as linear and quadratic.

Thus, this work proposes the use of Pearson's Correlation Coefficient (*r*) to analyze the correlation between GC content and coverage, allowing to identify the strength of linear correlation and detect nonlinear associations, beyond identify a relationship between GC bias and sequencing platforms. The positive and negative signs of *r* also allow us to infer directly and inversely proportional relationships, respectively. To evaluate the bias, we used the data of *Corynebacterium pseudotuberculosis* obtained from different sequencing technologies to identify if the CG bias is related to used platforms.

Keywords: GC Bias, Pearson, quadratic, Correlation.

1. INTRODUÇÃO

Com o advento dos sequenciadores de alto rendimento ou sequenciadores de próxima geração (NGS) em 2005, foram obtidas muitas vantagens no sequenciamento de genomas quando se compara com a metodologia de Sanger (1977), dentre estas pode-se citar: a grande quantidade de dados gerados, menor custo e tempo de corrida mais curto. O que propiciou o aumento da quantidade de projetos de sequenciamento de genomas completos (Shendure & Ji, 2008). (Figura 1)

As primeiras plataformas NGS a serem comercializadas foram 454 FLX Roche, Illumina e SOLiD System (Carvalho & Silva, 2010). Todas estas têm características específicas quanto aos métodos de sequenciamento (Liu et al., 2012). Além disto, há um grande desafio quanto ao processamento e manipulação dos dados produzidos, principalmente quanto à montagem de genomas, onde se observa regiões não representadas (lacunas), que podem estar relacionadas a limitações computacionais, mas também pode tratar-se de regiões pouco representadas no sequenciamento em função do conteúdo GC de cada organismo *(*Chen et al., 2*013;* Ross et al., 2013).

Estudos anteriores demonstram que muitas plataformas NGS, tais como 454 e Ion Torrent PGM possuem erros específicos de sequenciamento, tais como inserções e deleções (INDEL) decorrentes de regiões homopoliméricas *(*Zeng et al., 2013). Outras plataformas, como a Illumina, possuem erros de substituições (Shendure & Ji, 2008). Estes erros de sequenciamento dificultam o processo de montagem, *de novo* e por referência, e reduzem a acurácia dos alinhamentos das leituras, além de causar erros no processo de montagem de genomas e transcriptomas *(Wir*awan et al., 2014).

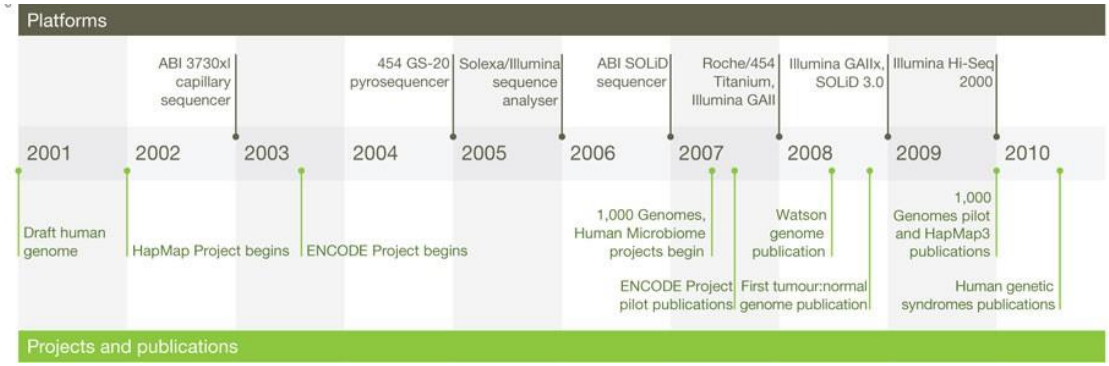

Gráfico mostrando ano de lançamento das plataformas mais utilizadas e seus respectivos rendimentos, além de apresentar o ano de inicialização de alguns dos projetos e publicações mais importantes (Adaptado de Mardis,2011)

Figura 1. Linha do tempo mostrando ano de lançamento das principais plataformas NGS

1.1 Sequenciadores de Primeira Geração

Frederick Sanger desenvolveu o método de sequenciamento por terminação de cadeia em 1977 (Sanger *et al*., 1977). Sanger revolucionou a metodologia de sequenciar a molécula de DNA ao introduzir didesoxinucleotídeos (ddNTP) marcados que têm como principal característica a ausência do radical hidroxila na terminação 3' do carbono da pentose. Desta forma a adição de um ddNTP específico (A, T, G ou C) à cadeia recém-formada interrompe a extensão da mesma, sendo que no final deste processo obteremos polinucleotídeos de diferentes tamanhos e consequentemente distintos pesos moleculares por conta da adição randômica de didesoxinucleotídeos. Esta diferença de tamanho permite que estas moléculas sejam separadas por eletroforese e posteriormente detectadas por um sequenciador (Sanger *et al*., 1977)

As técnicas utilizadas anteriormente eram árduas e aplicavam excessiva radioatividade, diferente do proposto por Sanger, sendo reconhecido como a tecnologia da primeira geração. A Applied Biosystems foi a responsável por introduzir o primeiro sequenciador automático (1987): ABI Prism 3700, capaz de produzir 500 kilobases (kb)

por dia além de contar com leituras que podiam alcançar um comprimento de 600 bases. O atual modelo chama-se ABI 3730xl e realiza um rendimento de 2.88 MB por dia, e gera leituras de até 900 bases. Esta tecnologia foi utilizada no projeto genoma humano como uma das principais ferramentas, apesar do alto custo e baixo rendimento quando comparada às novas tecnologias que surgiram a partir de 2005 (Liu et al., 2012).

1.2 Sequenciadores de Segunda Geração

Com o surgimento da plataforma 454 pela empresa Roche em meados de 2005, deu-se início a uma segunda geração de sequenciadores, conhecidos como NGS (*Next Generation Sequencing*). Posteriormente surgiram outras plataformas como a GA Illumina lançada pela Solexa e SOLiD da Life Tecnologies *(He*nson et al., 2012). Entre as principais características que diferenciam os sequenciadores NGS do método de Sanger pode-se citar o baixo custo, redução do tempo de sequenciamento e o alto rendimento (Schlebusch & Illing, 2012). Entretanto, as primeiras leituras produzidas por estes equipamentos eram muito curtas trazendo grandes desafios ao processo de montagem de genomas, devido a dificuldade em representar as diversas regiões repetitivas presentes no genoma, além de haver a necessidade de estruturas computacionais mais robustas, para o processamento do grande volume de dados gerados, e de algoritmos eficientes *(Mi*ller et al., 2010; Schlebusch & Illing, 2012).

1.2.1 454 Roche GS-FLX

Esta plataforma utiliza o sequenciamento baseado na síntese de uma nova molécula a partir da molécula molde também conhecido como Pirosequenciamento, onde a adição de um nucleotídeo à cadeia recém-formada usando DNA polimerase, causa a liberação de um pirofosfato que será convertido para ATP pela enzima ATP sulfurilase. Este ATP será utilizado na reação de oxidação da luciferina pela enzima luciferase (Figura 2) para a

produção de um sinal de luz que posteriormente será captado por uma câmera CCD ("Charge-coupled Device") (Kaur *et al*., 2013). Na sua primeira versão, a plataforma 454 produzia leituras com tamanho entre 100 – 150 pb (pares de bases) com rendimento de 20 Mb por corrida. Em 2008, uma atualização foi lançada e nomeada como 454 GS FLX Titanium produzindo leituras de aproximadamente 700 pb de comprimento com 99.9% de acurácia (Liu et al., 2012). A versão atual deste equipamento já produz leituras com comprimento superior a 1000 pb, rendimento de 700 Mb e trabalha com as bibliotecas genômicas fragments e paired-end (www.454.com). Dentre todos os NGS, o 454 detinha o maior comprimento de leitura e esta característica favoreceu a sua utilização para montagem *de novo* e estudos metagenômicos *(Z*hang et al., 2011). Entretanto esta plataforma apresenta uma baixa acurácia para representar regiões homopoliméricas, havendo muitos erros associados com inserções e deleções (Schlebusch & Illing, 2*012; Z*hang et al., 2011).

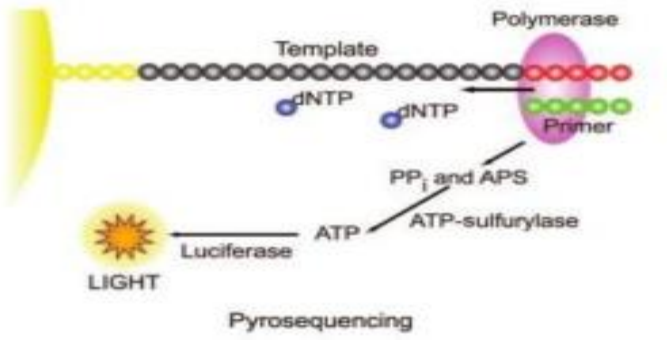

Figura 2. Pirosequenciamento na plataforma 454

Química de sequenciamento utilizando pirofosfatos liberados a partir da adição de nucleotídeos à cadeia em extensão (Adaptado de Kaur *et al*., 2013).

1.2.2 Illumina GA

A plataforma Illumina realiza o sequenciamento através da síntese de uma nova molécula usando DNA polimerase e nucleotídeos terminadores marcados com diferentes fluoróforos, muito semelhante ao que ocorre na técnica de Sanger. O diferencial desta metodologia de sequenciamento consiste na utilização de uma plataforma sólida de vidro

onde ocorre a amplificação das amostras por PCR (Henson *et al*., 2012).

O lançamento da Genome Analyzer em 2006 pela Solexa que logo em seguida (2007) foi comprada pela Illumina (Liu et al., 2012), foi um marco na revolução dos sequenciadores de nova geração. Produzindo inicialmente leituras de 35 pb e com um rendimento de 1Gb *(He*nson et al., 2012), utilizando a tecnologia de sequenciamento por síntese. Em pouco tempo a plataforma recebeu diversas atualizações e melhorias. Em 2010, foi lançado o HiSeq 2000 (http://www.illumina.com) que teve seu rendimento inicial em torno de 200 GB, passando em seguida para 600 GB por corrida, com leituras de 100 pb. O HiSeq 2000 utiliza bibliotecas fragments e paired-ends, e o principal erro desta plataforma são as substituições *(He*nson et al., 2012).Comparado às outras plataformas de segunda geração, a HiSeq 2000 é a mais barata contendo uma taxa de erro menor que 2%, tornando esta tecnologia amplamente utilizada em estudos de transcriptoma e montagens *de novo. (He*nson et al., *2012;* Liu et al., 2012).

Atualmente, a HiSeq 2500 conta com um rendimento de até 1000 Gb para leituras de tamanho 2x125 pb (http://www.illumina.com).

1.2.3 ABI SOLiD

A plataforma SOLiD (*support oligonucleotide ligation detection*) foi lançada e comercializada em 2007 pela Applied Biosystems (Life Technologies). Utiliza uma abordagem de sequenciamento por ligação catalisada pela enzima DNA ligase. As primeiras versões deste equipamento produziam leituras de 25-35 pb de comprimento com um rendimento de até 4 Gb. Atualmente, a versão SOLiD 5500 xl system produz até 250 Gb (Kaur & Malik, 2013). O tamanho das leituras varia de acordo com a biblioteca utilizada: Mate-paired com 2x60 pb, Paired-end com 75x35 pb e Fragment com 75 pb (http://www.appliedbiosystems.com). ABI SOLiD utiliza um sistema de codificação com 16 possíveis combinações de 2 nucleotídeos em quatro possibilidades de cores através da implementação de ligações químicas de sondas marcadas com *di-base.* Desta forma, cada par de nucleotídeos representa uma cor, indicada pelos números de 0 a 3 (Figura 3), o que

permite diferenciar polimorfismos verdadeiros de erros de sequenciamento pois cada base é interrogada até duas vezes, além de permitir uma identificação mais precisa de mutações pontuais como inserções e deleções (Shendure & Ji, 2*008; Z*hang et al., 2011). O principal erro para esta plataforma é a substituição de bases (Salmela, 2010). Na figura 3 pode-se notar um diagrama exemplificando o código de cores onde as bases no eixo vertical identificam a primeira base e as bases no eixo horizontal simbolizam a segunda base. Se a di-base identificada for por exemplo G-T (guanina e timina) a cor será verde e o código numérico será 1.

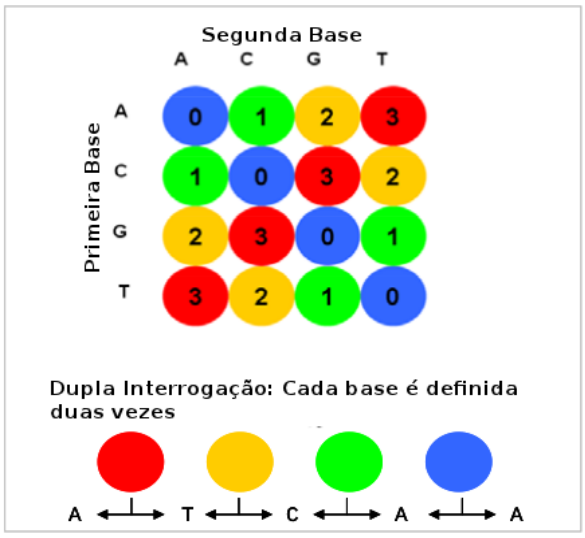

Figura 3. Código de cores para plataforma SOLiD

4 cores utilizadas para detectar as duas bases, em 16 possíveis combinações, e um exemplo de dupla interrogação usando o código de cores (Adaptado de Applied, 2011c).

Como arquivos de saída após o sequenciamento, a plataforma produz um arquivo com extensão csfasta contendo as bases sequenciadas no formato do código numérico que pode variar de 0 a 3 (Figura 4 - 1) e outro arquivo com os valores de qualidade Phred.

(Figura 4 - 2).

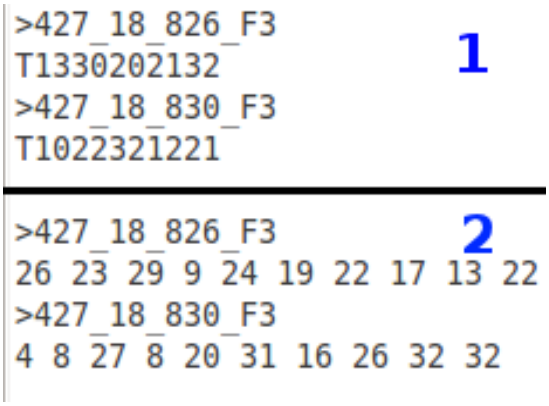

Figura 4. Exemplo de arquivo .csfasta (1) e arquivo .qual (2)

Bases em código de cores(1) e qualidades Phred (2) de cada base do arquivo .csfasta.

1.3 Sequenciadores de Terceira Geração

Novos métodos de decodificação de DNA foram lançados após as plataformas de segunda geração, utilizando abordagens como *Single Molecule Real Time* – (SMRT*)* que não requer amplificação, o que possibilita uma alta acurácia dos dados e leituras potencialmente longas quando comparadas à geração anterior *(Z*hang et al., 2011).

1.3.1 Ion Torrent Personal Genome Machine (PGM)

A plataforma Ion Torrent além de fazer parte da terceira geração, inaugurou a era dos sequenciadores pós-luz ao utilizar um semicondutor como sistema de detecção de bases. O método de sequenciamento baseia-se na detecção de íons hidrogênio que são liberados durante o processo de polimerização do DNA *(He*nson et al., 2012). Atualmente são oferecidos 3 chips para sequenciamento: Ion 314™ Chip v2 com rendimento de até 100 Mb, Ion 316™ Chip v2 com rendimento de até 1Gb e Ion 318™ Chip v2 com rendimento de até 2Gb. O tamanho das leituras pode alcançar até 400 pb, exibindo o comprimento médio em torno de 200 pb (http://www.lifetechnologies.com). O principal tipo de erro desta plataforma são as inserções e deleções (INDEL) pois, em regiões homopoliméricas, não existe linearidade entre a intensidade do fluxo de íons hidrogênios detectados e o número de nucleotídeos incorporados, fazendo com que erros na determinação do tamanho de tais regiões sejam frequentes *(*Zeng et al., 2013).

1.3.2 PacBio (Pacific Biosciences)

Em 2010, a Pacific Biosciences lançou a plataforma de terceira geração baseada na tecnologia SMRT, onde não há necessidade das etapas de amplificação por PCR, assim

como todas as demais plataformas de 3º geração (Kaur & Malik, 2013). A abordagem de sequenciamento utilizada, identifica diferentes nucleotídeos marcados com distintas cores através dos fosfatos. Durante o processo de síntese, o sinal de fluorescência é detectado assim que um fosfato é liberado na reação de incorporação do nucleotídeo à fita de DNA *(Z*hang et al., 2011).

O PacBio RS System produz um rendimento de 35-45 Mb com uma média de tamanho de leitura em torno de 1500 pb *(He*nson et al., 2012). Atualmente, a versão PacBio RS II produz um throughput maior que 35 Mb com leituras de tamanho médio entre 4.200-8.500 pb (http://www.pacificbiosciences.com). Apesar disto, há uma elevada taxa de erro das bases (13 – 15%) nos dados *(He*nson et al., 2012).

1.4 Erros de Sequenciamento

As tecnologias da nova geração introduzem erros nos dados sequenciados (inserções, deleções e substituições), dificultando análises baseadas em genomas de referência em análises genômicas e transcriptômicas (Tabela 1).

Além destes erros, o sequenciamento de regiões preferenciais como resultado das químicas de sequenciamento pode resultar no viés GC, que influencia a avaliação dos dados produzidos por estas plataformas *(*Ross et al., 2013). Algumas métricas já foram descritas para avaliar este viés, baseado na declividade da reta observada no gráfico de dispersão para a relação GC/cobertura. Uma janela deslizante, de tamanho igual ao tamanho de leitura para uma respectiva plataforma, é utilizada para coletar os dados de conteúdo GC e cobertura ao longo de todo o genoma do organismo. A janela desliza pelo genoma, sendo que o tamanho do passo da janela é igual ao tamanho da janela para que não haja sobreposição. Em cada passo a janela deve coletar o valor de conteúdo GC e cobertura para determinada região este procedimento repete-se até chegar ao final das bases do genoma. Ao final deste processo teremos vários valores de GC e cobertura que podem ser avaliados através de um gráfico de dispersão, onde no eixo X teremos os respectivos valores de GC capturados em cada janela e no eixo Y o valor de cobertura

associado a cada janela. A reta de regressão linear é ajustada ao pontos do gráfico de dispersão e o grau de viés GC definido como o ângulo de declividade (Inclinação) que a reta forma com o eixo X (Chen *et al*., 2013). Este tipo de análise só pode ser mensurada quando o comportamento do gráfico de dispersão é linear.

Contudo, abordagens capazes de identificar associações não-lineares como o coeficiente de Pearson podem ser importantes por ser adequarem a diferentes perfis dos dados (Asuero *et al*., 2006).

Tabela 1. Plataformas NGS e suas principais características e erros associados.

Plataforma		Tamanho da leitura	Rendimento	Biblioteca Genômica	Tipo de erro
454		>1000	700 Mb	Fragments / Paired-end	Indel
Illumina HiSeq 2500		2x125	1 Tb	Fragments / Paired-end	Substituição
SOLiD		2x60 bp (Mate-paired) 75x35 bp (Paired-end) 75 bp (Fragment)	>20Gb/dia	Fragments / Paired-end / Mate-paired	Substituição
ION Torrent PGM	Chip 314 v2	35–400 bases (Média: 200 bases)	60 – 100 Mb	Fragments / Paired-end / Mate-paired	Indel
	Chip 316 v2		600 Mb - 1Gb	Fragments / Paired-end / Mate-paired	Indel
	Chip 318 v2		1.2 – 2 Gb	Fragments / Paired-end / Mate-paired	Indel

1.4.1 Viés GC.

O conteúdo GC médio de um organismo pode ser calculado somando-se o total de Guaninas e Citosinas e dividindo este valor pelo número total de bases do genoma. Em seguida multiplica-se o valor por 100 para que o resultado possa ser mensurado em porcentagem (Fórmula 1).

$$\frac{G+C}{A+T+G+C} \times 100$$

Fórmula 1. Cálculo do conteúdo GC médio (%).

Diminuindo o valor do denominador na fórmula 1, podemos ter tamanhos diferentes de janelas que podem ser utilizadas para coletar valores de conteúdo GC ao longo de todo o genoma. Tais janelas deslizantes são úteis para avaliar regiões onde o conteúdo GC está acima da média ou abaixo da média. Quando o conteúdo GC está muito acima da média (regiões de alto conteúdo GC) ou quando o conteúdo GC está muito abaixo da média (regiões de baixo conteúdo GC), costuma-se observar em dados NGS uma cobertura muito baixa ou muito elevada. Tal fenômeno é conceituado como Viés GC e trata-se de um viés de cobertura (Figura 5). A cobertura simboliza o total de bases sequenciadas (leituras obtidas através do sequenciamento) divididas pelo tamanho esperado do genoma.

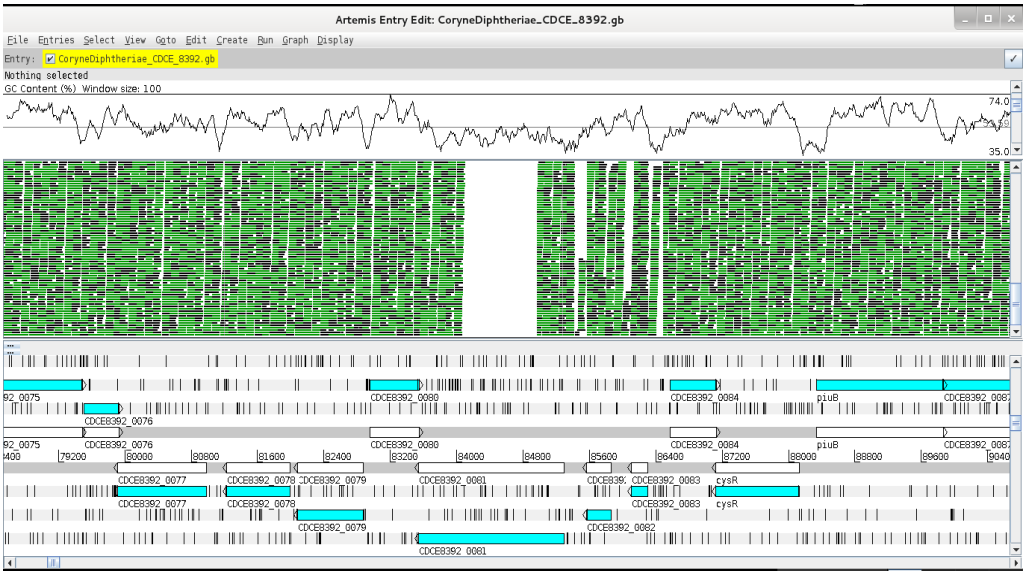

Figura 5. Software artemis demonstrando um genoma bacteriano

Parte superior da imagem apresenta um gráfico em linha ilustrando a variação do conteúdo GC de acordo com um tamanho de janela igual a 100 pb. A linha horizontal que passa pelo gráfico representa conteúdo GC

Este viés tem sido observado em regiões de genomas com alto e baixo conteúdo GC (de acordo com um tamanho de janela deslizante pré-determinado), produzindo uma tendência ao erro em função da baixa ou alta cobertura (viés GC).

Um dos principais mecanismos pelo qual o viés de cobertura pode ser introduzido, ocorre nas etapas de amplificação por PCR, comum à todas as plataformas NGS. Por esta razão, trabalhos abordando este viés GC foram produzidos a fim de identificar e quantificar como os diferentes graus deste viés afetam o processo de montagem. Apesar disto, a sua influência na montagem é pouco discutida, mesmo sendo notório que a falta de cobertura irá contribuir para a formação de gaps *(*Aird et al., 2011; Benjamini and Speed, *2012;* Chen et al., *2013;* Ross et al., 2013).

Uma das formas de se avaliar o conteúdo GC é baseada na declividade da reta de regressão linear a partir do gráfico de dispersão (GC *versus* cobertura) *(*Chen et al., 2013). Chen *et al* utilizaram 5 espécies de bactérias sequenciadas apenas com a plataforma Illumina, o que não permitiu avaliar a influência dos diferentes métodos de sequenciamento quanto ao viés GC. Ainda neste estudo, observou-se o comportamento linear dos gráficos de dispersão na maioria das amostras, entretanto este tipo de análise não pode ser aplicada quando observamos um comportamento não-linear no gráfico de dispersão.

Desta forma, este trabalho propõe o uso do coeficiente de correlação de Pearson para verificar diferentes padrões de associação (linear ou não-linear) entre as variáveis conteúdo GC e Cobertura. Para tanto, utilizou-se dados da espécie *Corynebacterium pseudotuberculosis*, conhecido por serem genomas bem sintênicos e conservados *(So*ares et al., 2013), obtidas através de diferentes tecnologias de sequenciamento: SOLiD system, 454, Ion Torrent PGM (com e sem a enzima Hi-Q) e Illumina, para identificar a relação do viés GC com o método de sequenciamento utilizado.

1.5 Coeficiente de Pearson

O Coeficiente de Correlação de Pearson (Fórmula 1) também conhecido como Coeficiente de Correlação Produto-Momento (*r*) foi desenvolvido por Pearson em 1896 ao pesquisar os trabalhos anteriores de Galton (1888). Este Coeficiente adimensional mede a intensidade da correlação linear sendo que a correlação pode ser definida como o grau de associação entre 2 variáveis e pode ser apresentada através de um gráfico de dispersão *(A*suero et al., 2006).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{Onde} \quad \bar{x} = \frac{\sum_{i=1}^{n} x}{n} \quad \text{e} \quad \bar{y} = \frac{\sum_{i=1}^{n} y}{n}$$

Fórmula 2. Equação para cálculo do Coeficiente de Pearson.

O valor de *r* pode variar de -1 até +1, onde o sinal é responsável por caracterizar associações inversamente proporcionais (-) e diretamente proporcionais (+) respectivamente. Valores próximos a 1 indicam uma forte correlação linear e valores próximo a zero demonstram uma fraca associação. Os valores intermediários são classificados de acordo com a tabela 2 (Taylor, 1990).

Caso ocorresse uma situação hipotética onde o valor de *r* seria igual ou muito próximo a 1 (raramente ocorre na natureza), significaria respectivamente uma associação linear perfeita ou quase perfeita entre as 2 variáveis, indicando que conforme a medida da variável *x* aumenta, a de *y* aumenta na mesma proporção (diretamente proporcional) ou indicando que enquanto o valor de *x* aumenta, o valor de *y* diminui na mesma proporção (inversamente proporcional) *(A*suero et al., 2006; Taylor, 1990).

Tabela 2. Valores de *r* e suas interpretações.

Valores de *r*	Interpretação
0.90 to 1.00	Correlaçao Muito Alta
0.70 to 0.89	Alta Correlação
0.50 to 0.69	Correlação Moderada
0.30 to 0.49	Baixa Correlação

Outra característica importante acerca do coeficiente, é que o *r* é incapaz de inferir associações com interferência ou sem interferência, sendo que para determinar relações de causa e efeito entre 2 variáveis são necessárias outras análises *(A*suero et al., 2006).

1.5.1 Gráficos de Dispersão.

Os gráficos de dispersão são muito informativos e fazem parte de uma etapa extremamente importante na análise de dados estatísticos e determinação de correlação entre 2 variáveis analisadas. A distribuição dos pontos em um gráfico de dispersão juntamente com os valores de *r* permitem inferir a intensidade da associação linear e até mesmo detectar ausência de correlação linear em casos onde o valor de *r* corresponde a zero (Figura 6).

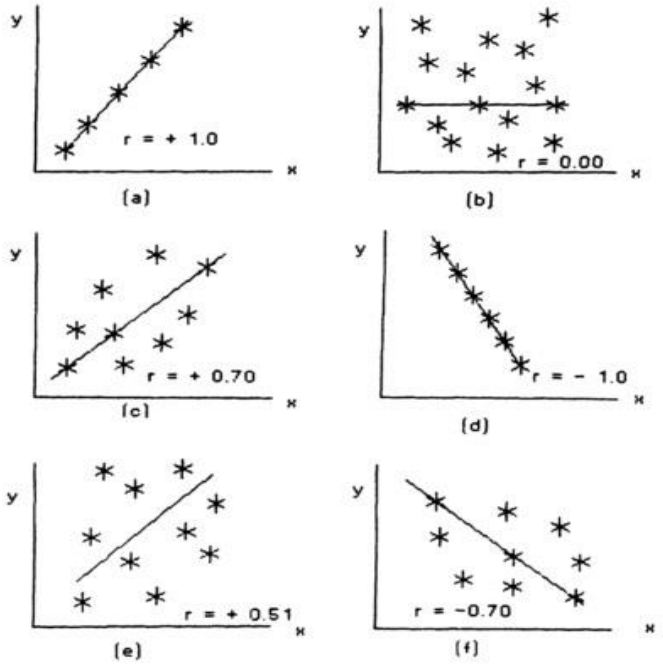

Figura 6. Gráficos de dispersão com diferentes intensidades de associação linear

Gráficos de dispersão demonstrando associação linear perfeita (a) onde os pontos se ajustam a uma reta e o valor de r é máximo e igual a + 1 (diretamente proporcional) e na imagem (d) uma associação linear perfeita porém com valor de r negativo demonstrando uma associação indiretamente proporcional. Em (b) pode-se observar uma associação não-linear. Em (c), (e) e (f) exemplificam casos de associações com alta correlação (Adaptado de Taylor., 1990).

Existe entretanto, mais um caso particular que deve ser analisado: onde o gráfico de dispersão apresenta pontos dispersos na forma de uma parábola (quadrático). Nesta situação específica, não existe correlação linear e o valor de *r* é zero ou muito próximo a zero. Isto ocorre quando a correlação é polinomial e quadrática ($y=x^2$) sendo que o valor de *r* não serve para mensurar este tipo de associação quadrática, porém pode-se utilizar a análise do gráfico de dispersão juntamente com o valor de *r* para detectar este tipo específico de associação não-linear, quando existir, o que pode ser utilizado para a avaliação do viés GC *(A*suero et al., 2006) (Figura 7).

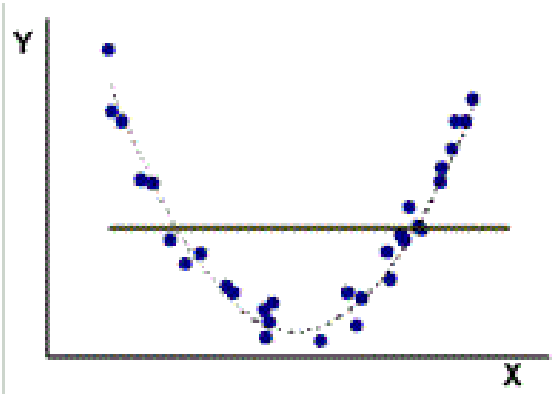

Figura 7. Gráfico de dispersão apresentando correlação não-linear

Gráfico de dispersão demonstrando regressão polinomial com associação quadrática (Adaptado de http://stat2.med.up.pt/cursop/print_script.php3?capitulo=regressao&numero=3&titulo=Correla%E7%E3o%20e%20regress%E3o%20linear%20simples).

2. OBJETIVOS

2.1 Objetivo Geral

Avaliar a relação do viés GC com as químicas de sequenciamento das plataformas NGS.

2.2 Objetivos Específicos

- Alinhar as leituras oriundas das plataformas ION Torrent PGM, SOLiD, Illumina e 454 utilizando os softwares de alinhamento: CLC Genomics Workbench, Bowtie 2 e TMAP;
- Avaliar o viés GC para os dados do organismo modelo *Corynebacterium pseudotuberculosis* utilizando métricas já descritas.
- Propor uma nova métrica de estatística descritiva para avaliação do viés GC quando o gráfico de dispersão não apresentar perfil linear.
- Avaliar a nova métrica utilizando os dados de estudos anteriores a respeito do viés GC.

3. MATERIAIS E MÉTODOS

3.1 Avaliação estatística

O software *Statistical computing environment* R (www.R-project.org) foi utilizado para a construção dos gráficos de dispersão e cálculo do coeficiente de correlação de Pearson (*r*) obtidos a partir das métricas calculadas para 2 variáveis: conteúdo GC e cobertura. Para tanto, a tabela de saída padrão do software Picard foi convertida para o formato ".csv". Em seguida, importou-se a tabela para dentro do ambiente R utilizando a linha de comando abaixo, que define que os valores da tabela serão salvos dentro de uma variável nomeada como "variavel":

variavel <- read.table("tabela.csv",header=T,sep=",",dec=".")

Durante o processo de construção dos gráficos de dispersão, optou-se por padronizar todos os gráficos seguindo o seguinte método: a média da cobertura normalizada possui valor igual a 1, sendo que valores superiores significam cobertura elevada, e valores inferiores indicam uma baixa cobertura em relação à média da cobertura normalizada. Os pontos dos gráficos de dispersão que apresentaram cobertura acima da média foram marcados de vermelho, e uma reta paralela ao eixo Y (Cobertura Normalizada) foi traçada em todos os gráficos indicando o conteúdo GC médio de cada organismo específico. Pontos que foram marcados na cor azul indicam as coordenadas de GC e Cobertura Normalizada que ficaram abaixo da média.

3.2 Validação do Coeficiente de Pearson

Com o objetivo de validar os resultados do Coeficiente de Correlação de Pearson e gráficos de dispersão utilizou-se os dados analisados por Chen *et al*, (conforme tabela 3),

que avaliaram o viés GC com base na declividade da reta de regressão linear (Inclinação). Assim, foram analisadas as espécies *Pseudomonas fluorescens*, *Shewanella amazonensis*, *Escherichia coli*, *Staphylococcus aureus* e *Mycobacterium tuberculosis*, sequenciadas na plataforma Illumina, e cujas bibliotecas foram obtidas através do "Sequence Read Archive database (SRA)" assim como os genomas completos dos organismos utilizados no estudo de Chen *et al,* 2013. Algumas linhagens foram sequenciadas mais de uma vez, porém em tempos distintos, sendo que o identificador SRA pode ser utilizado para diferenciar estes sequenciamentos de linhagens idênticas da mesma espécie.

Tabela 3. Amostras de 14 bibliotecas genômicas analisadas no estudo de Chen et al.

Espécie	Identificador	Identificador (SRA)	Conteúdo GC Médio (%)	GC Bias (Inclinação)
Pseudomonas fluorescens Pf0-1	NC_007492.2	DRR001171	60.5	-1.96
Shewanella amazonensis SB2B	NC_008700.1	SRR090701	53.6	3.41
Escherichia coli K-12 MG1655	NC_000913.2	SRR001666	50.8	-0.07
Escherichia coli K-12 MG1655	NC_000913.2	SRR350605	50.8	-1.9
Escherichia coli K-12 MG1655	NC_000913.2	SRR398955	50.8	-1.55
Escherichia coli K-12 MG1655	NC_000913.2	SRR402738	50.8	-2.6
Staphylococcus aureus USA 300	NC_010079.1	SRR022866	32.8	-5.3
Staphylococcus aureus USA 300	NC_010079.1	SRR022867	32.8	-4.49
Staphylococcus aureus USA 300	NC_010079.1	SRR022868	32.8	-5.05
Staphylococcus aureus MRSA252	NC_002952.2	SRR342227	32.8	4.13
Mycobacterium tuberculosis H37Rv	NC_000962.2	SRR099031	65.6	-1.1
Mycobacterium tuberculosis H37Rv	NC_000962.2	SRR017680	65.6	-5.24
Mycobacterium tuberculosis H37Rv	NC_000962.2	SRR023440	65.6	-8.86
Mycobacterium tuberculosis H37Rv	NC_000962.2	SRR023441	65.6	-8.96

3.3 Alinhamento dos dados

Os dados brutos do sequenciamento em formato SRA das 14 bibliotecas utilizadas por Chen *et al* foram convertidos para o formato fastq. Então, as leituras foram mapeadas contra seus respectivos genomas de referência, cujos números de acesso são apresentados na tabela 3, utilizando o software CLC Genomics Workbench 7 com os seguintes parâmetros: *mismatch cost* igual a 2, *insertion* e *deletion cost* igual a 3, *lenght fraction* de 0,7 e *similarity fraction* igual a 0,7. Além do software Bowtie 2 (Langmead & Salzberg, 2012) com o valor 32 para o parâmetro -L (*seed*), 1 para -N (número de *mismatches* permitidos).

Para as leituras provenientes de 10 sequenciamentos de *Corynebacterium pseudotuberculosis*, em formato fastq (dados de Illumina e Ion Torrent PGM) e color space (dados de Solid), foram realizados mapeamentos contra os seus genomas completos, disponíveis no NCBI (Tabela 4). O alinhamento foi realizado através do software CLC Genomics Workbench 7 e Bowtie 2, com os mesmos parâmetros apresentados acima. Apenas as bibliotecas de *Corynebacterium pseudotuberculosis 31*, sequenciada na plataforma Ion Torrent PGM com e sem a nova Ion Hi-Q™ Sequencing Chemistry, teve seu alinhamento realizado através do software Tmap (http://mendel.iontorrent.com/ion-docs/) e CLC Genomics Workbench, devido ao tamanho diferenciado das leituras.

Tabela 4. Amostras de *Corynebacterium pseudotuberculosis* avaliadas.

NGS	Espécie	Identificação	Conteúdo GC médio (%)
SOLiD	*C. pseudotuberculosis* 258	NC_017945.1	52.1
	C. pseudotuberculosis 267	NC_017462.1	52.2
	C. pseudotuberculosis |19	NC_017303.1	52.2
	C. pseudotuberculosis 31	NC_017730.1	52.2
Illumina	*C. pseudotuberculosis* 1/06-A	NC_017308.1	52.2
	C. pseudotuberculosis 3/99-5	NC_016781.1	52.2
	C. pseudotuberculosis 42/02-A	NC_017306.1	52.2
Ion Torrent	*C. pseudotuberculosis* 31	NC_017730.1	52.2
HiQ	*C. pseudotuberculosis* 31	NC_017730.1	52.2
454	*C. pseudotuberculosis* 1002	NC_017300.1	52.2

Para o TMAP, utilizou-se o parâmetro mapall, *--max-seed-band* de 18, *--num-threads* com 2, penalização por *mismatch* (-M) com 3 e *-v stage1 map1*, para mapear sequências curtas contra o genoma de referência com o algoritmo BWA *short-read* (Li & Durbin, 2009), map2 para mapear sequências longas contra o genoma de referência utilizando BWA *long-read* (Li & Durbin, 2010) e map3 que se trata de uma simplificação do algoritmo SSAHA *long-read (*Ning et al., 2001) para mapear sequências longas.

3.4 Processamento dos alinhamentos pelo Picard

Os resultados dos alinhamentos foram processados por módulos do software Picard (http://picard.sourceforge.net). Primeiramente, o alinhamento em formato SAM foi ordenado pelo módulo SortSam, e em seguida o módulo CollectGCBiasMetrics foi utilizado para se obter as métricas baseadas no conteúdo GC (AT dropout e GC dropout) com base em uma janela de 100 pares de bases, para as quais foram calculados os valores de cobertura normalizados.

O módulo CollectGCBiasMetrics requer que os arquivos SAM de input estejam devidamente ordenados. Como arquivo de saída tem-se uma tabela com 5 colunas. A primeira coluna tem como cabeçalho a nomenclatura “GC” e fornece o valor do conteúdo GC coletado de acordo com um tamanho de janela pré-determinado. A segunda coluna tem o identificador “*WINDOWS*” e informa o número de janelas encontradas para o respectivo valor de conteúdo GC. A terceira coluna, “*READ_STARTS*”, mostra o número de leituras que iniciaram exatamente naquela janela com determinado conteúdo GC. A quarta coluna, definida como “*MEAN_BASE_QUALITY*”, identifica a qualidade média das bases que caíram na respectiva janela de conteúdo GC identificado. A quinta coluna fornece o valor da cobertura média normalizada de todas as janelas com determinado conteúdo GC. A sexta coluna, “*ERROR_BAR_WIDTH*”, mostra a taxa de erro para o cálculo da cobertura normalizada, sendo que o valor mostrado nesta janela indica o quanto a cobertura normalizada pode variar para mais ou para menos do seu respectivo valor. Você pode ver um exemplo desta tabela abaixo (Tabela 5).

Tabela 5. Dados coletados pelo software picard.

GC	WINDOWS	READ_STARTS	MEAN_BASE_QUALITY	NORMALIZED_COVERAGE	ERROR_BAR_WIDTH
16	7	0	0	0	0
17	12	0	0	0	0
18	22	0	0	0	0
19	21	1	0	0.011423	0.011423
20	36	14	0	0.093284	0.024931
21	66	51	22	0.185356	0.025955
22	64	63	26	0.236125	0.029749
23	105	94	29	0.214743	0.022149
24	136	186	26	0.328062	0.024055
25	170	251	25	0.354165	0.022355
26	263	544	26	0.496163	0.021273
27	288	567	27	0.47225	0.019833
28	421	812	27	0.462653	0.016236
29	596	1442	27	0.580364	0.015283
30	808	1812	27	0.537933	0.012637
31	1224	3088	28	0.60517	0.01089
32	1623	4104	28	0.606555	0.009468
33	2140	6105	29	0.684311	0.008758
34	2863	9365	28	0.784635	0.008108
35	4080	13115	28	0.771062	0.006733
36	5405	18271	29	0.810864	0.005999
37	7222	25881	28	0.859617	0.005343
38	9680	35480	28	0.879204	0.004668
39	13414	50782	29	0.908098	0.00403
40	17703	70388	29	0.953747	0.003595

A métrica de AT-dropout foi utilizada para medir o quanto abaixo da cobertura estão as regiões com conteúdo GC abaixo de 50% em relação a cobertura média observada. O valor de AT-dropout representa o percentual de leituras que foram mapeadas em regiões com conteúdo GC acima de 50%, e GC-dropout mostra o percentual de leituras que mapearam em regiões com conteúdo GC abaixo de 50%. Ambas foram obtidas através da opção "SUMMARY_OUTPUT" que localiza-se dentro do módulo CollectGCBiasMetrics.(http://picard.sourceforge.net/picard-metric-definitions.shtml#GcBiasSummaryMetrics).

4. RESULTADOS E DISCUSSÃO

A influência do conteúdo GC a cerca da baixa representatividade de certas regiões em estudos genômicos, utilizando plataforma de nova geração (NGS), já vem sendo discutido (Ross et al., 2013). Assim, neste trabalho avaliou-se a relação entre conteúdo GC e cobertura em 8 genomas de *Corynebacterium pseudotuberculosis*, cujas sequências completas estão depositados no banco de dados do NCBI. Por se tratar de uma bactéria com alta conservação gênica *(*Soares et al., 2013), tornou possível a associação do conteúdo GC / Cobertura genômica com as plataformas NGS.

Como alternativa à metodologia de medir o grau de viés GC com base na declividade da reta de regressão linear no gráfico de dispersão, entre conteúdo GC e Cobertura (Chen et al., 2013) propõem-se utilizar o valor numérico do Coeficiente de Correlação Linear de Pearson (*r*), cujos valores são adimensionais variando entre -1 e +1, para medir a intensidade da associação linear e observar associações não-lineares entre as variáveis GC e cobertura, quando existirem. Além disto, *r* permite inferir se as variáveis analisadas são diretamente proporcionais (+) ou inversamente proporcionais (-) *(*Asuero et al., 2006).

4.1 Validação dos dados utilizando amostras de Chen *et al*

A fim de validar o coeficiente proposto, aplicou-se a metodologia baseada no valor de *r* para a lista de bactérias apresentadas no trabalho de Chen *et al* (Tabela 6), para comparar o viés GC positivo e negativo com os valores positivos e negativos de *r*. Assim, 13 amostras tiveram concordância quanto a positividade e negatividade entre o GC Bias Slope e os sinais de *r* (Tabela 6), e apenas *Escherichia coli* K-12 MG1655 (GenBank: NC_000913.2), figura 5A, teve seu GC Bias Slope igual a -0.07 (negativo) e *r* igual a 0.02 (CLC) e 0.05 (Bowtie). Com exceção desta amostra que apresentou um viés nulo, as demais de *Escherichia coli* apresentaram um viés negativo com alta correlação linear (Figura 8).

Tabela 6. 14 Bibliotecas genômicas e seus respectivos valores de *r* para cada alinhador utilizado.

Espécie	Identificador (SRA)	Alinhadores	*r*	AT Dropout	GC Dropout	Leituras Mapeadas	Total de Leituras
Pseudomonas fluorescens Pf0-1	DRR001171	CLC	-0.86	0	3.06	90.89%	10.129.958
		Bowtie	-0.83	0	2.84	92.05%	
Shewanella amazonensis SB2B	SRR090701	CLC	0.95	7.04	0.42	93.89%	5.860.354
		Bowtie	0.95	7.04	0.41	92.22%	
Escherichia coli K-12 MG1655	SRR001666	CLC	0.02	0.40	1.25	98.33%	14.095.336
		Bowtie	0.05	0.41	1.18	98.67%	
Escherichia coli K-12 MG1655	SRR350605	CLC	-0.78	0	5.32	70.25%	103.989.664
		Bowtie	-0.78	0	5.06	69.00%	
Escherichia coli K-12 MG1655	SRR398955	CLC	-0.72	0	4.53	92.18%	84.098.170
		Bowtie	-0.72	0	4.48	92.22%	
Escherichia coli K-12 MG1655	SRR402738	CLC	-0.97	0	7.16	98.64%	40.062.962
		Bowtie	-0.97	0	7.03	99.16%	
Staphylococcus aureus USA300	SRR022866	CLC	-0.38	13.68	0.29	73.77%	25.551.716
		Bowtie	-0.35	13.36	0.28	69.94%	
Staphylococcus aureus USA300	SRR022867	CLC	-0.16	13.39	0.25	83.03%	3.816.486
		Bowtie	-0.16	13.28	0.24	82.25%	
Staphylococcus aureus USA300	SRR022868	CLC	-0.13	12.23	0.28	78.10%	31.125.794
		Bowtie	-0.28	11.95	0.26	77.11%	
Staphylococcus aureus MRSA252	SRR342227	CLC	0.84	8.19	0	89.81%	32.087.596
		Bowtie	0.84	8.51	0	86.57%	
Mycobacterium tuberculosis	SRR099031	CLC	-0.89	0	2.66	91.70%	43.236.170

H37Rv		Bowtie	-0.88	0	2.59	90.99%	
Mycobacterium tuberculosis H37Rv	SRR017680	CLC	-0.96	0	12.47	72.06%	11.611.065
		Bowtie	-0.97	0	12.35	71.10%	
Mycobacterium tuberculosis H37Rv	SRR023440	CLC	-0.94	0	22.34	57.45%	43.429.770
		Bowtie	-0.93	0	21.87	55.80%	
Mycobacterium tuberculosis H37Rv	SRR023441	CLC	-0.97	0	23.30	56,33%	47.165.936
		Bowtie	-0.97	0	22.51	49.10%	

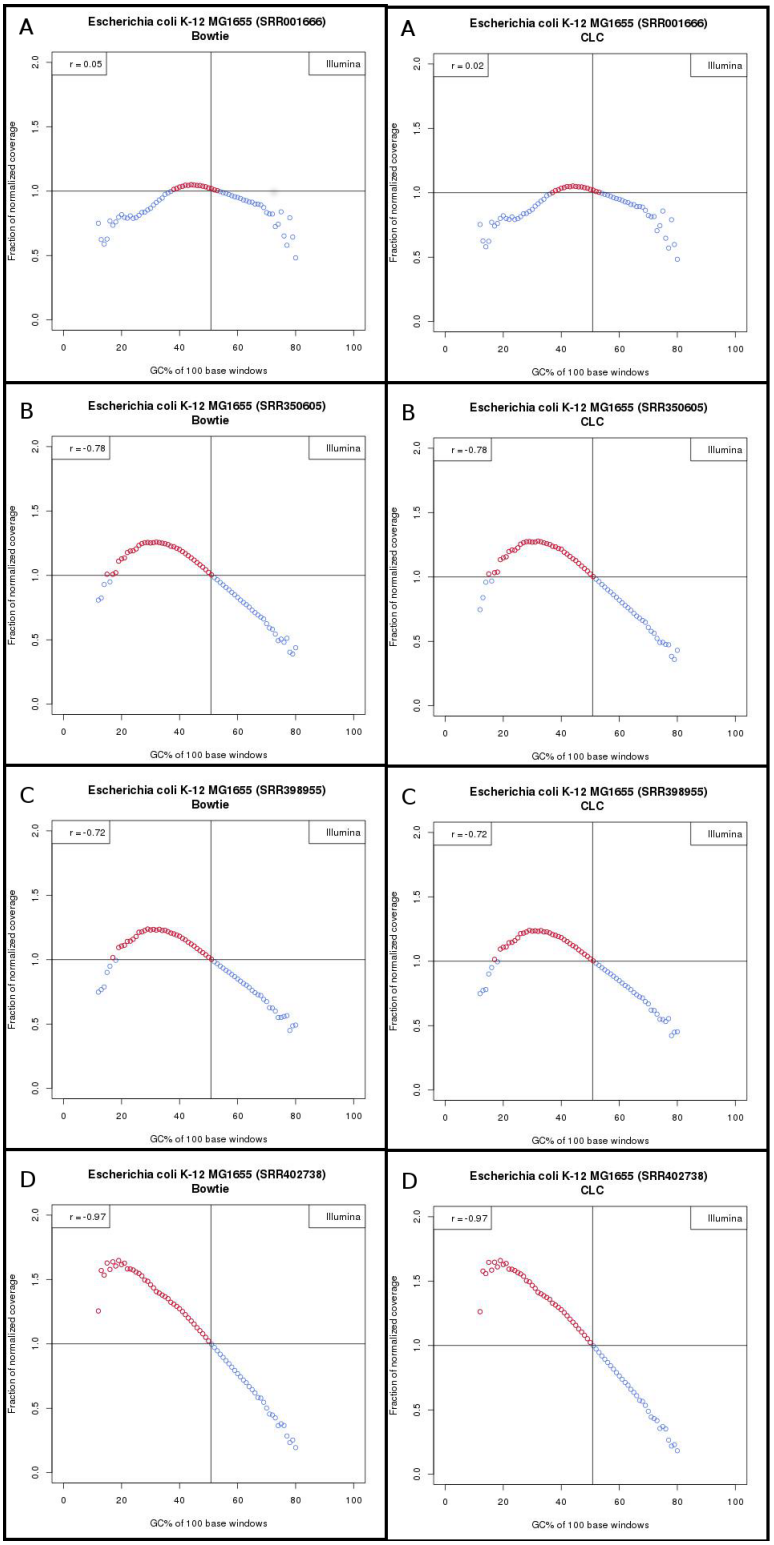

Figura 8. Gráficos de dispersão para *Escherichia coli*

O comportamento dos gráficos para *Escherichia coli* mostram alta correlação linear (valores de r no canto superior esquerdo da figura) , com exceção da figura 1A que mostrou correlação próxima a zero. Os pontos em vermelho no gráfico apresentam valores de GC e Cobertura acima da cobertura média normalizada representada graficamente pela reta horizontal que passa pelo valor 1 no eixo y. Os pontos em azul mostram valores abaixo da cobertura média normalizada. A reta vertical que corta o gráfico, passa pelo conteúdo GC médio de cada organismo e portanto é específica para cada organismo.

Avaliando as amostras do trabalho de Chen *et al*. que foram sequenciadas na plataforma Illumina, observa-se para *Pseudomonas fluorescens* e *Shewanella amazonensis* valores de *r* negativo e positivo, respectivamente. Assim, o comportamento dos gráficos de dispersão são diferenciados (Figura 9), demonstrando que o sinal de *r* pode inferir relações diretamente proporcionais (+) ou inversamente proporcionais (-) entre as variáveis GC / Cobertura.

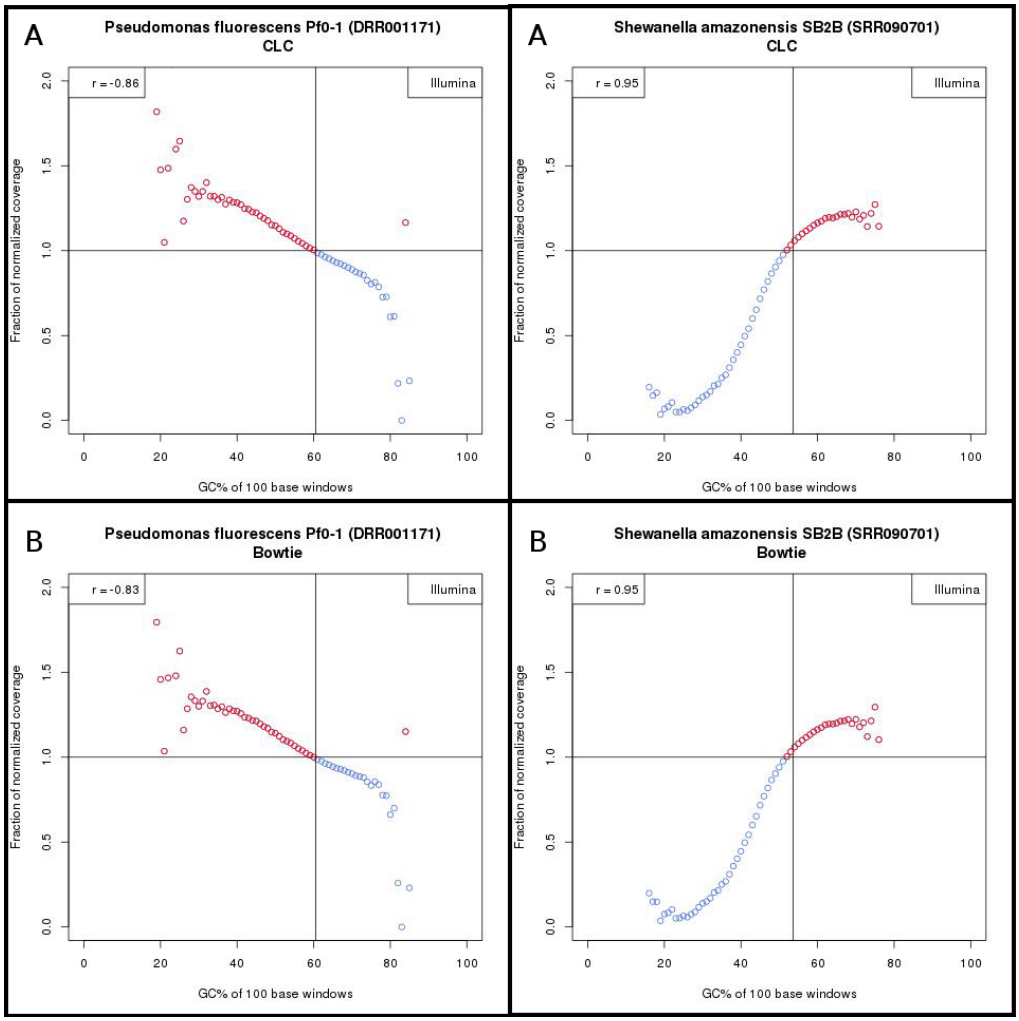

Figura 9. Gráficos de dispersão para *Pseudomonas fluorescens* e *Shewanella amazonensis*

Pseudomonas fluorescens apresentando viés negativo e *Shewanella amazonensis* com viés positivo. Sinal negativo de r para *Pseudomonas fluorescens* e sinal positivo para *Shewanella amazonensis* concordantes quanto ao Viés negativo e Viés positivo destes organismos. Os pontos em vermelho no gráfico apresentam valores de GC e Cobertura acima da cobertura média normalizada representada graficamente pela reta horizontal que passa pelo valor 1 no eixo y. Pontos em azul mostram valores abaixo da cobertura média normalizada. A reta vertical que corta o gráfico, passa pelo conteúdo GC médio de cada organismo e portanto é específica para cada organismo

Notou-se também que a mesma linhagem de *Mycobacterium tuberculosis* apresentou ângulos de declividade, em relação ao eixo *x*, visualmente distintos nos gráficos de dispersão para cada sequenciamento (Figura 10), além dos sinais de *r* serem negativamente concordantes com os sinais de declividade (*Slope*) (Tabela 6). Os quatro sequenciamentos de *M. tuberculosis* consistem em uma biblioteca de fragmentos e as demais pareadas, com tamanhos de leituras e distância de insertos muito aproximados (Chen et al., 2013), mostrando que não há relação do tipo de biblioteca com o viés GC. Em contrapartida, as três amostras de *Staphylococcus aureus* USA 300, apesar de também representarem diferentes sequenciamentos, resultaram em gráficos de dispersão e valores de *r* muito próximos (Figura 11A-C).

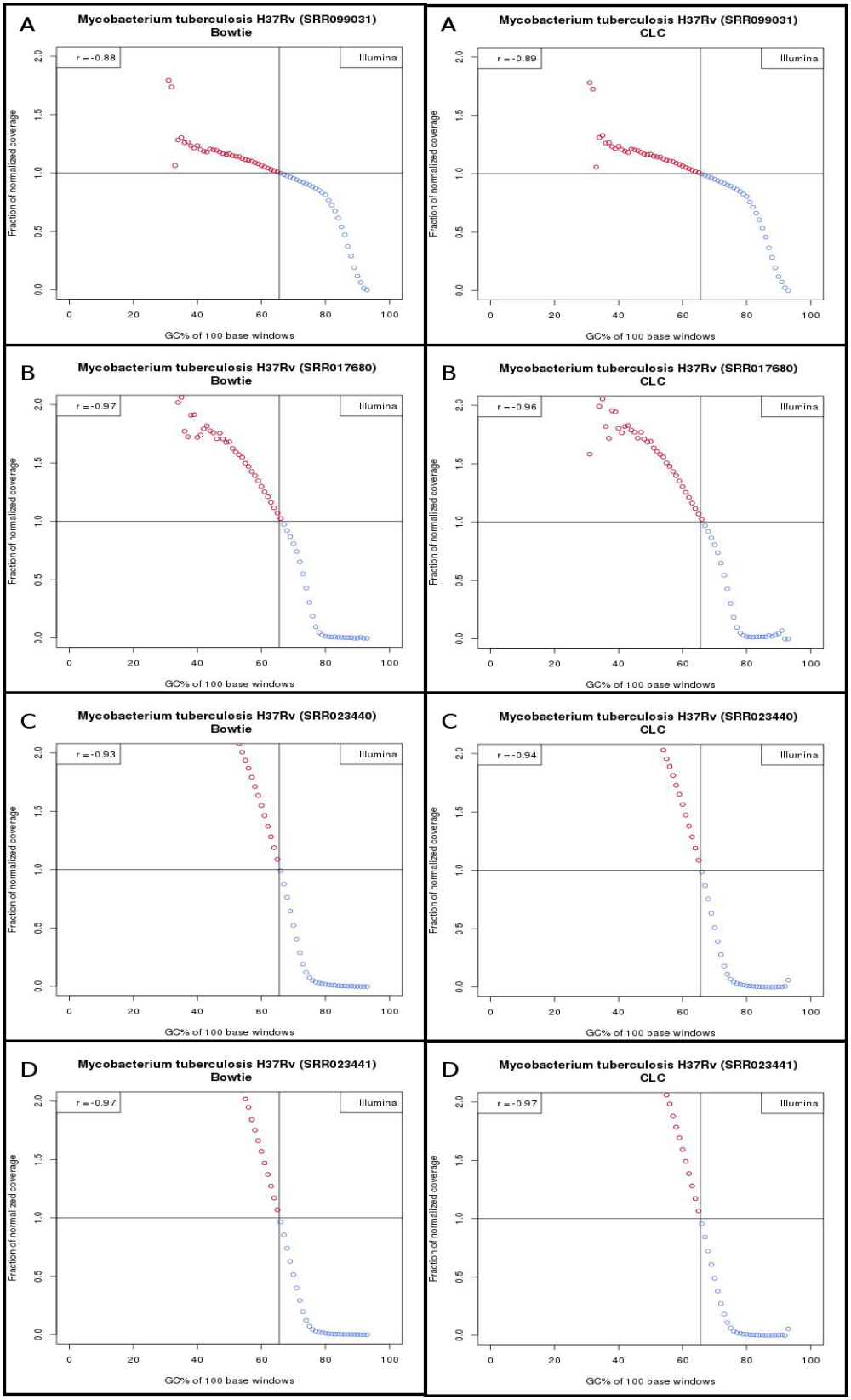

Figura 10. Gráficos de dispersão para *Mycobacterium tuberculosis*

Pode-se observar diferentes padrões de declividades no gráfico de dispersão para a mesma linhagem de *Mycobacterium tuberculosis*, onde a maior inclinação foi observada para a C e D, o que corrobora com as informações obtidas por Chen et al.. Os pontos em vermelho no gráfico apresentam valores de GC e Cobertura acima da cobertura média normalizada representada graficamente pela reta horizontal que passa pelo valor 1 no eixo y. Pontos em azul mostram valores abaixo da cobertura média normalizada. A reta vertical que corta o gráfico, passa pelo conteúdo GC médio de cada organismo e portanto é específica para cada organismo**.**

Com as três amostras de *Staphylococcus aureus* USA 300, os gráficos e valores de *r* mostram um padrão de associação diferenciado quando comparados a linhagem *Staphylococcus aureus* MRSA252, que apresentou um forte viés positivo, evidenciando que o viés GC também pode estar associado às diferentes linhagens da mesma espécie (Figura 11), apesar dos percentuais de conteúdo GC das 4 amostras de *Staphylococcus aureus* serem idênticos (Tabela 3).

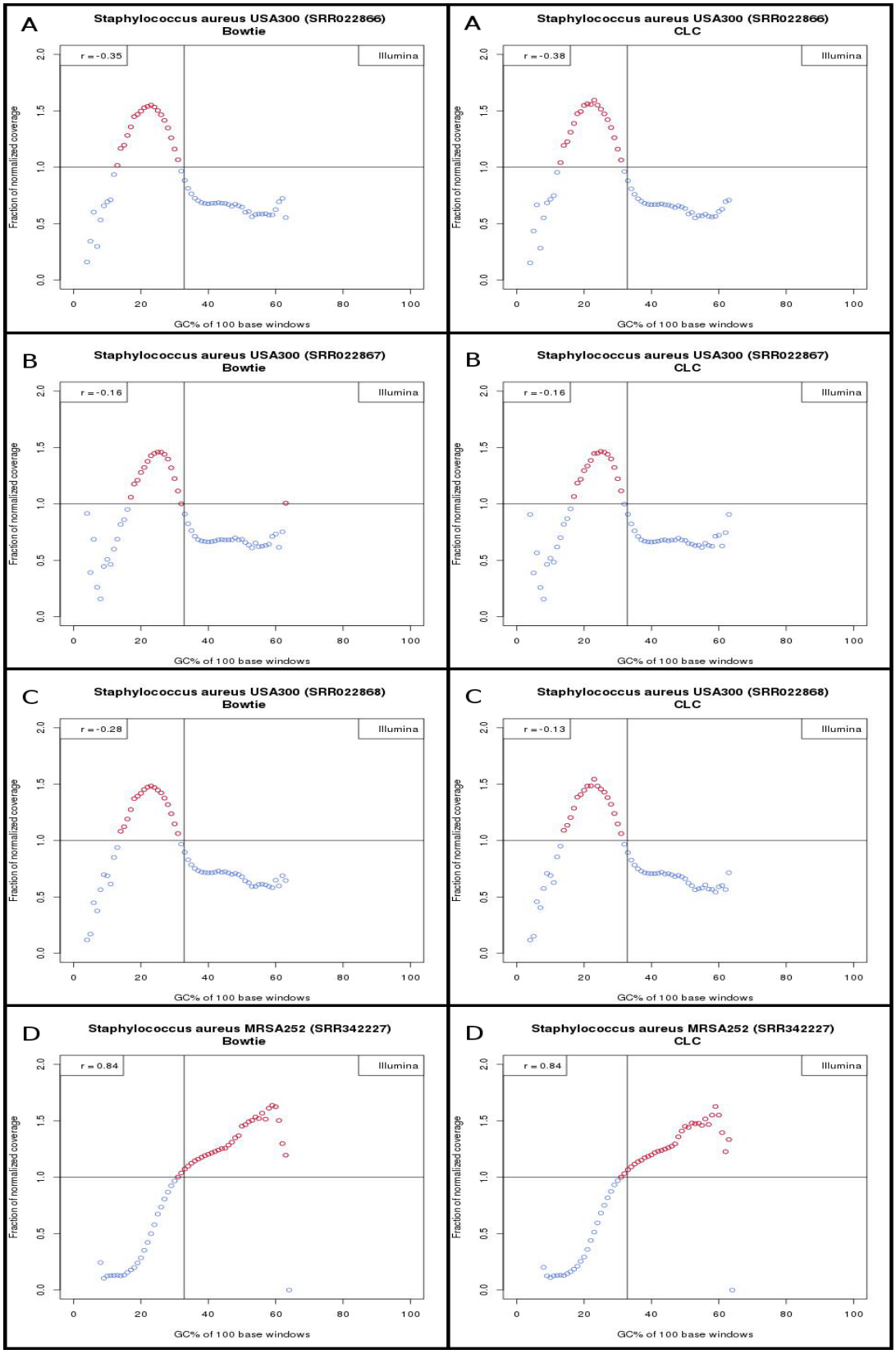

Figura 11. Gráfico de dispersão para *Staphylococcus aureus*

S. aureus apresentou padrões de viés GC diferentes entre linhagens distintas, quando se compara a linhagem MRSA252 (D) com as amostras de USA300 (A, B e C). Os pontos em vermelho no gráfico apresentam valores de GC e Cobertura acima da cobertura média normalizada representada graficamente pela reta horizontal que passa pelo valor 1 no eixo y. Pontos em azul mostram valores abaixo da cobertura média normalizada. A reta vertical que corta o gráfico, passa pelo conteúdo GC médio de cada organismo e portanto é específica para cada organismo.

Para todas as 14 amostras avaliadas por Chen *et al*, não houve diferenças significativas no valor do *r* e no gráfico de dispersão entre os diferentes alinhadores: CLC Genomics Workbench e Bowtie2. O que demonstra que o software de alinhamento não interfere na associação GC/Cobertura Normalizada. (Tabela 6)

4.2 Avaliação dos dados de *Corynebacterium pseudotuberculosis*

Os gráficos de dispersão para as variáveis conteúdo GC e cobertura demonstraram um mesmo padrão gráfico em forma de parábola na associação para amostras de *C. pseudotuberculosis* sequenciadas na plataforma SOLiD (Figura 12), e um padrão diferenciado entre as amostras de *C. pseudotuberculosis* obtidas a partir do sequenciador Illumina (Figura 13), apesar da espécie apresentar conservação gênica *(*Soares et al., 2013).

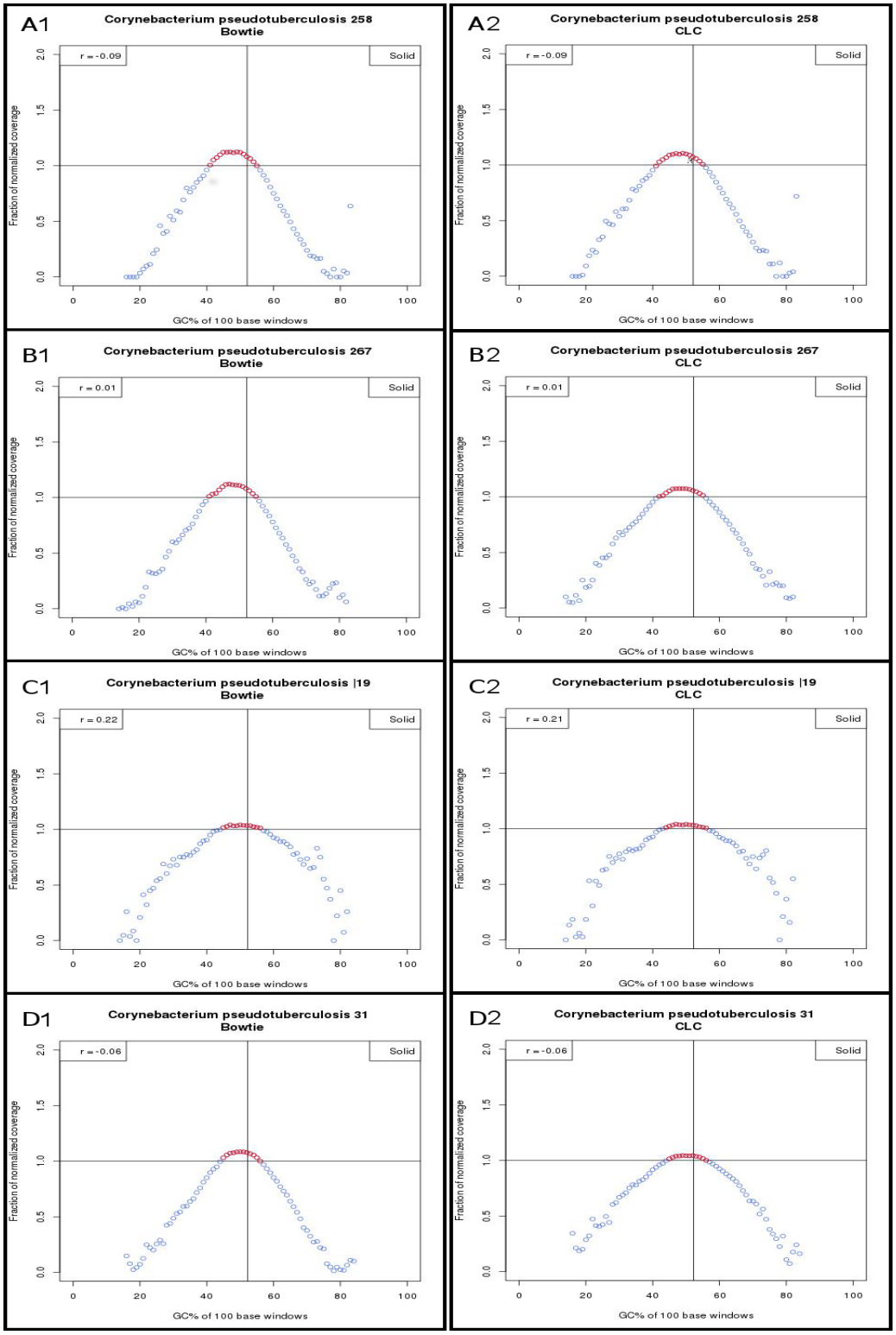

Figura 12. Gráficos de dispersão para amostras de *Corynebacterium pseudotuberculosis* sequenciadas na plataforma SOLiD

Comparação entre os gráficos de dispersão obtidos a partir dos softwares de mapeamento Bowtie 2 (A1, B1, C1 e D1) e CLC Genomic Workbench (A2, B2, C2 e D2), onde não se observou diferença significativa entre os dados obtidos por estes alinhadores. Os pontos em vermelho no gráfico apresentam valores de GC e Cobertura acima da cobertura média normalizada representada graficamente pela reta horizontal que passa pelo valor 1 no eixo y. Os Pontos em azul mostram valores abaixo da cobertura média normalizada. A reta vertical que corta o gráfico, passa pelo conteúdo GC médio de cada organismo e portanto é específica para cada organismo.

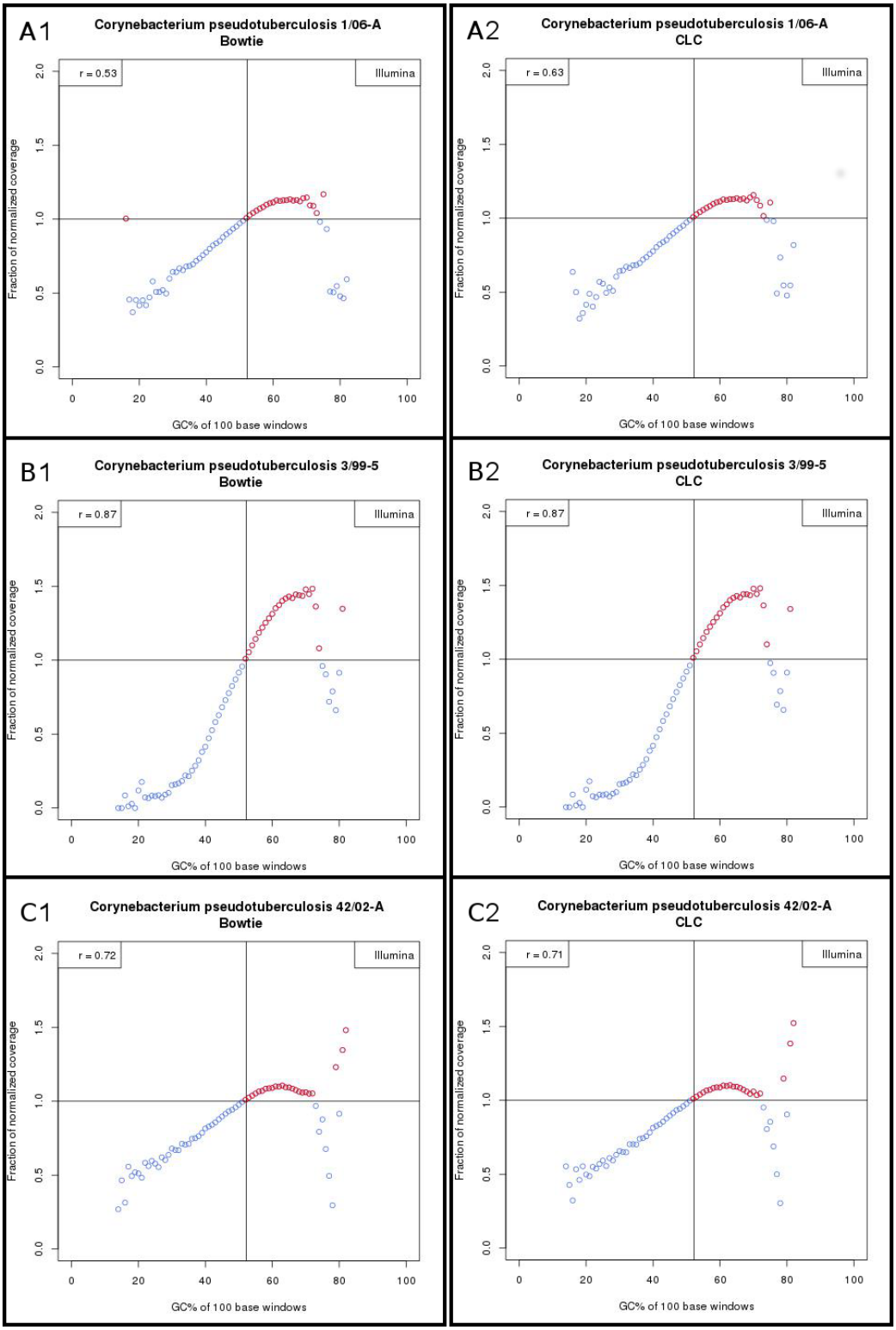

Figura 13. Gráficos de dispersão para amostras de *Corynebacterium pseudotuberculosis* sequenciadas na plataforma Illumina

Comparação entre os gráficos de dispersão obtidos a partir dos softwares de mapeamento Bowtie 2 (A1, B1 e C1) e CLC Genomic Workbench (A2, B2 e C2), onde não se observou diferença significativa entre os resultados dos distintos alinhadores. Os pontos em vermelho no gráfico apresentam valores de GC e Cobertura acima da cobertura média normalizada representada graficamente pela reta horizontal que passa pelo valor 1 no eixo y. Pontos em azul mostram valores abaixo da cobertura média normalizada. A reta vertical que corta o gráfico, passa pelo conteúdo GC médio de cada organismo e portanto é específica para cada organismo.

A avaliação das amostras pelo coeficiente de Pearson permitiu observar um viés positivo com alta correlação linear para os dados de Illumina (Figura 13), para SOLiD um viés diferenciado (Figura 12) com uma correlação não-linear (valores de *r* próximo de 0), e gráficos de dispersão em forma de parábola demonstrando uma associação quadrática ($y=x^2$, onde *y* é a cobertura normalizada e *x* o conteúdo GC). Esta diferença pode ser resultado da química do sequenciamento, visto que a plataforma SOLiD é a única a utilizar a enzima DNA ligase no sequenciamento (Schlebusch & Illing, 2012).

Para as linhagens de *C. pseudotuberculosis*, não se observou diferenças significativas no valor do *r* e no gráfico de dispersão entre os diferentes alinhadores: CLC Genomic Workbench, Bowtie2 e Tmap, o que demonstra que o software de alinhamento não interfere na associação GC/Cobertura Normalizada (Tabela 7).

A fim de validar a possível relação do viés GC ao sequenciador, analisou-se a linhagem *C. pseudotuberculosis* 31 pertencente ao biovar *equi* e sequenciada na plataformas SOLiD e Ion Torrent PGM. Segundo os valores de *r* (Tabela 7), os dados de Ion Torrent PGM apresentaram uma associação linear de baixa a moderada (-0,49 – -0,52) para a linhagem 31 (Figura 14 – A1 e B1), enquanto que os dados de SOLiD para esta mesma linhagem apresentaram uma relação próxima a zero (-0,06 – -0,09) (Figura 12 – D1-2). Quanto ao gráfico de dispersão, observou-se um padrão quadrático (forma de parábola) para os dados de SOLiD (Figura 12), e moderadamente linear para Ion Torrent PGM (Figura 14 – A1 e B1).

O sequenciamento da *C. pseudotuberculosis* 31 também foi realizado utilizando a nova química de sequenciamento Ion Hi-Q™. Com estes dados, observou-se a redução dos valores de *r* (-0.25) quando comparados à plataforma Ion Torrent PGM, evidenciando a redução do viés GC para a categoria "Pouca ou Nenhuma Correlação" tanto no alinhamento realizado no CLC Genomics Workbench quanto no alinhamento realizado

pelo software Tmap (Figura 14.A2;Figura 14.B2). Esta redução pode ser avaliada ao se comparar os gráficos de dispersão, onde o sequenciamento utilizando a enzima Ion Hi-Q™ mostrou um padrão próximo a uma reta paralela ao eixo das abscissas (Figura 14.A2-B2).

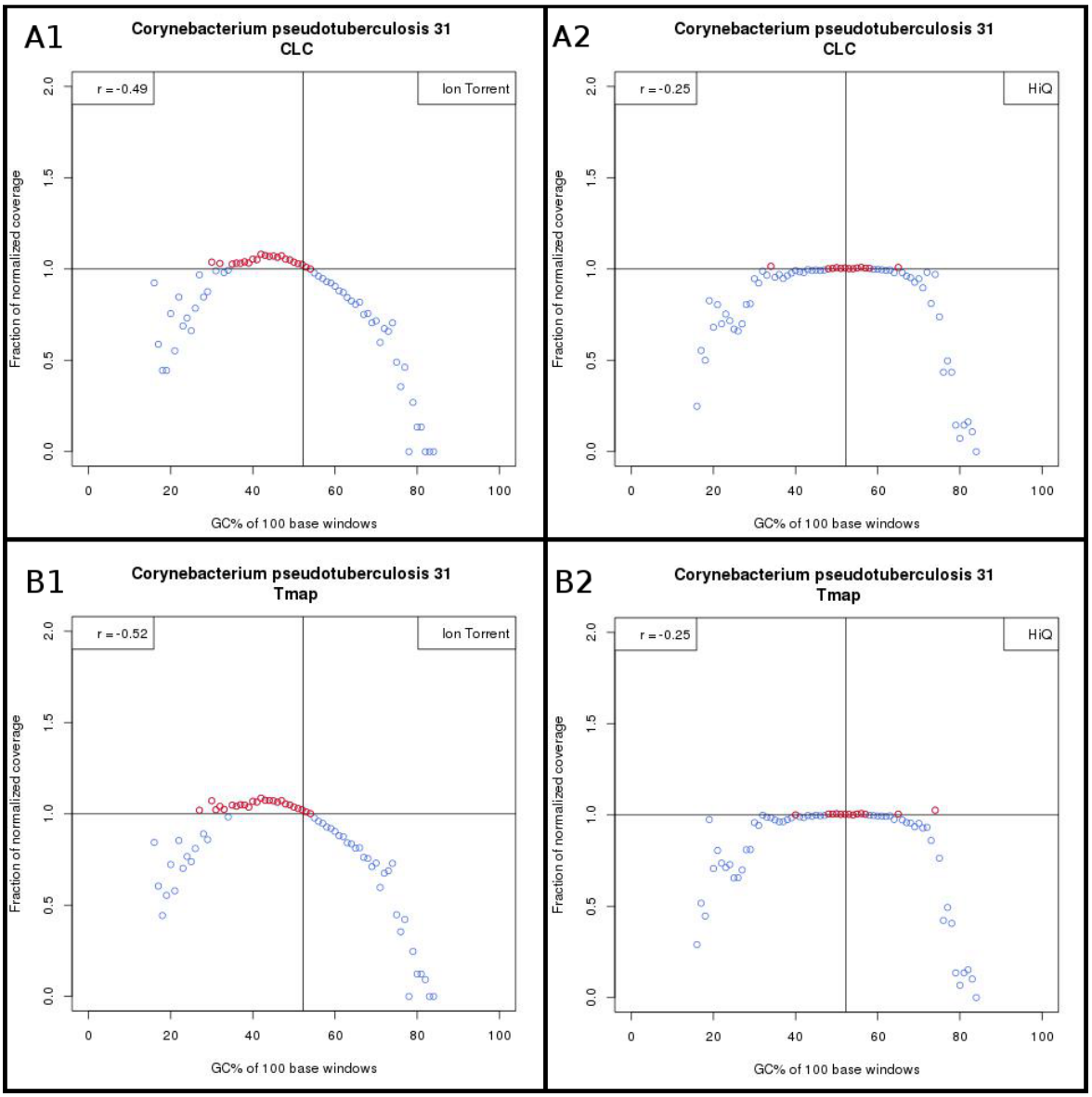

Figura 14. Gráfico de Dispersão para *C. pseudotuberculosis* 31

Comparação entre a química de sequenciamento Ion Torrent PGM (A1 e B1) e a nova Ion Hi-Q™ Química de Sequenciamento (A2 e B2). Os pontos em vermelho no gráfico apresentam valores de GC e Cobertura acima da cobertura média normalizada representada graficamente pela reta horizontal que passa pelo valor 1 no eixo y. Pontos em azul mostram valores abaixo da cobertura média normalizada. A reta vertical que corta o gráfico, passa pelo conteúdo GC médio de cada organismo e portanto é específica para cada organismo.

As linhagens *C. pseudotuberculosis* 258 e 31 foram as únicas a apresentarem associação negativa de acordo com os valores de Pearson (Tabela 7) para todas as plataformas, apesar de diferentes intensidades de correlação linear.

Em *C. pseudotuberculosis 1002*, cujo genoma foi sequenciado na plataforma 454, observou-se uma Baixa Correlação para a plataforma 454 (Figura 15). Os gráficos de dispersão apresentaram padrões de pontos distintos das outras plataformas e uma tendência a linearidade. Os valores de *r* para 454 foram de 0.42(Bowtie) e 0.47(CLC).

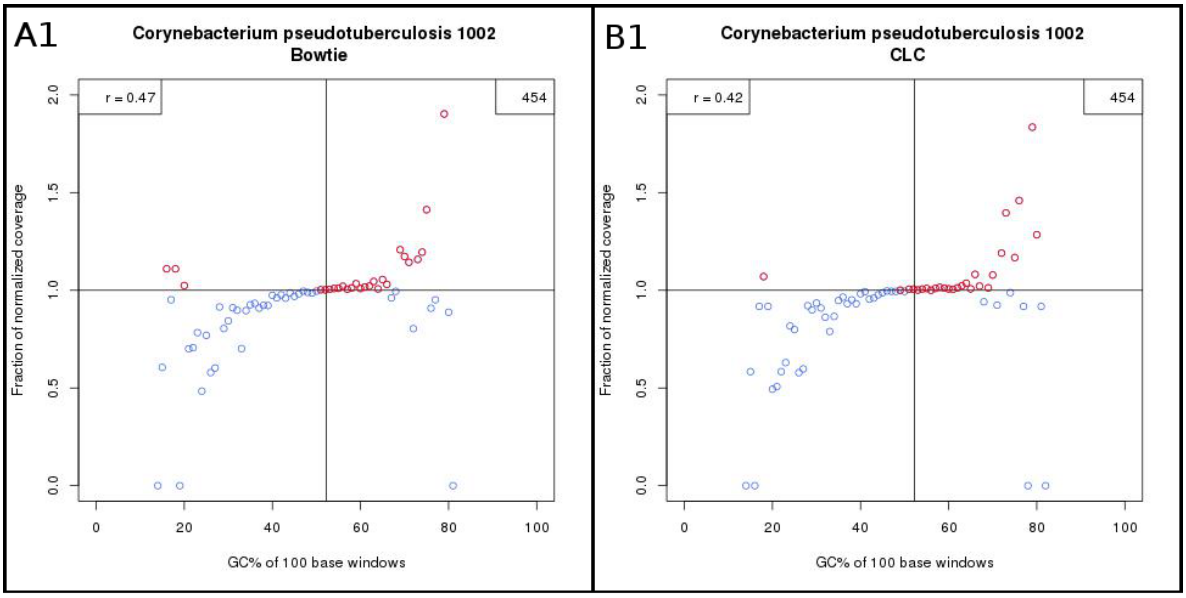

Figura 15. Gráfico de dispersão para *C. pseudotuberculosis* 1002 sequenciada na plataforma 454

A linhagem 1002 sequenciada na plataforma 454 apresenta o r como baixa correlação, e observa-se uma tendência a elevação da linearidade. Os pontos em vermelho no gráfico apresentam valores de GC e Cobertura acima da cobertura média normalizada representada graficamente pela reta horizontal que passa pelo valor 1 no eixo y. Pontos em azul mostram valores abaixo da cobertura média normalizada. A reta vertical que corta o gráfico, passa pelo conteúdo GC médio de cada organismo e portanto é específica para cada organismo.

A avaliação da métrica de AT-dropout e GC-dropout em *C. pseudotuberculosis* para os dados das plataformas Illumina e SOLiD revelaram padrões opostos. Enquanto os valores de AT-dropout ficaram próximos de 0 (zero) para SOLiD, para Illumina variaram de 3% à 9%. Quanto ao valor de GC-dropout para a plataforma SOLiD, houve variação de 1.38% até 5.37%, enquanto que para Illumina, GC-dropout manteve-se próximo de 0 (zero) (Tabela 7).

Segundo os valores de *r*, a intensidade da associação linear foi mais evidente em dados de Illumina, seguida por Ion Torrent PGM, 454, Hi-Q e SOLiD, este último apresentando uma associação não-linear quadrática (Figura 16).

Tabela 7. Amostras de C. pseudotuberculosis e seus respectivos valores de *r*

NGS	Espécie	Alinhadores	*r*	AT Dropout (%)	GC Dropout (%)	Leituras alinhadas (%)	Total de Leituras
SOLID	*Corynebacterium pseudotuberculosis* 258	CLC	-0.09	0.44	4.41	95.89%	10.061.554
		Bowtie	-0.09	0.44	5.37	54.12%	
	Corynebacterium pseudotuberculosis 267	CLC	0.01	0.43	3.13	80.00%	29.103.069
		Bowtie	0.01	0.45	4.89	64.11%	
	Corynebacterium pseudotuberculosis |19	CLC	0.21	0.42	1.42	87,10%	7.320.762
		Bowtie	0.22	0.57	1.38	61.52%	
	Corynebacterium pseudotuberculosis 31	CLC	-0.06	0.63	1.59	88.86%	183.034.704
		Bowtie	-0.06	1.07	3.07	43.64%	
Illumina	*Corynebacterium pseudotuberculosis* 1/06-A	CLC	0.63	3.70	0.25	95.77%	8.717.549
		Bowtie	0.53	3.71	0.25	95.51%	
	Corynebacterium pseudotuberculosis 3/99-5	CLC	0.87	9.75	0.79	94,96%	14.431.252
		Bowtie	0.87	9.77	0.80	94.40%	
	Corynebacterium pseudotuberculosis 42/02-A	CLC	0.71	3.20	0.16	96.83%	8.719.556
		Bowtie	0.72	3.19	0.17	99.57%	
ION	*Corynebacterium pseudotuberculosis* 31	CLC	-0.49	0.02	2.53	79.64%	3.566.141
		Tmap	-0.52	0.01	2.56	87.24%	
HiQ	*Corynebacterium pseudotuberculosis* 31	CLC	-0.25	0.23	0.08	89.49%	5.898.526
		Tmap	-0.25	0.16	0.11	95.85%	
454	*Corynebacterium pseudotuberculosis* 1002	CLC	0.42	0.50	0.03	91.54%	397.147
		Bowtie	0.47	0.73	0.03	88.29%	

5. CONCLUSÃO

A análise do viés GC através do coeficiente de Pearson mostrou-se eficaz e mais adequada na avaliação dos gráficos de dispersão na presença de associações lineares e não-lineares entre conteúdo GC e cobertura normalizada, tanto para os dados simulados por Chen *et al., 2013* quanto para os dados de *Corynebacterium pseudotuberculosis* obtidos através do sequenciamento pelas distintas plataformas NGS, o que possibilitou avaliar os fatores que podem afetar a decodificação do DNA e sua influência no viés GC.

A avaliação das amostras de *Corynebacterium pseudotuberculosis* nas plataformas Illumina, SOLiD, 454, Ion Torrent PGM com e sem a enzima Hi-Q, permitiu definir a correlação de linearidade do viés GC associado as tecnologias de sequenciamento (Figura 13), principalmente no tocante às suas químicas, considerando a redução do viés observado para o sequenciamento com a química de sequenciamento Hi-Q para Ion Torrent PGM.

Identificou-se um viés diferenciado para a plataforma SOLiD onde a associação não-linear e quadrática foi confirmada pelos gráficos de dispersão e valores de *r*. Este viés quadrático pode ser uma influência da química de sequenciamento diferenciada desta tecnologia que é o único NGS a utilizar a enzima DNA ligase ao invés da polimerase.

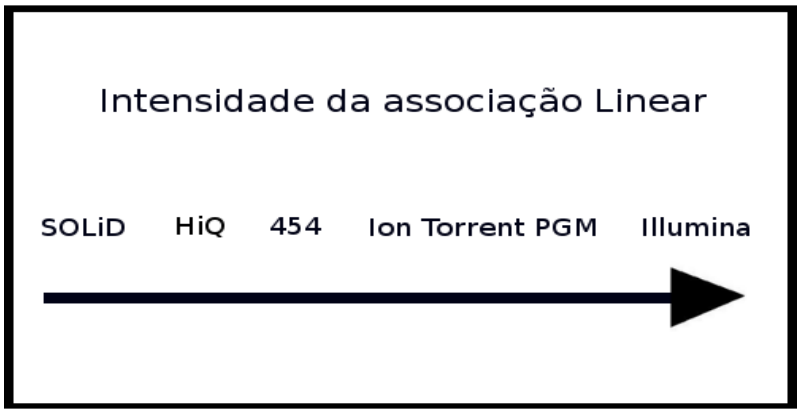

Figura 16. Relação da intensidade da correlação linear associado às plataformas de sequenciamento.

Observa-se o sentido crescente da força de correlação linear, iniciando pela plataforma SOLiD (associação não linear), seguindo para Hi-Q, 454, Ion Torrent PGM e Illumina, com associações lineares de diferentes intensidades.

A metodologia utilizada neste trabalho pode ser aplicada para outras plataformas e organismos a fim de identificar presença ou ausência de viés GC. Mesmo na ausência deste viés, o coeficiente de Pearson junto ao gráfico de dispersão são aplicáveis, pois os valores de Pearson para um viés nulo ou quase inexistente seriam muito próximos a zero. Além disto, observaria-se um gráfico de dispersão no qual os pontos se aproximariam de uma reta paralela ao eixo das abscissas, não havendo comportamento quadrático em qualquer gráfico analisado sem viés.

O viés GC pode ser resultado de diversos fatores: amplificação da biblioteca por PCR em análises genômicas como transcriptômicas (Aird et al., 2011; Korf, 2013), química de sequenciamento, além do conteúdo GC de cada organismo que influencia no grau e intensidade deste viés. No presente estudo, a associação do viés GC às diferentes plataformas foi confirmado, tendo como modelo *Corynebacterium pseudotuberculosis*. Esta informação pode auxiliar a definição dos parâmetros de softwares para análises baseadas em cobertura de sequenciamento, como montagem de genomas (por referência ou abordagem *de novo*) e análises de transcriptômicas.

6. REFERÊNCIAS BIBLIOGRÁFICAS

Aird D, Ross M G, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe D B, Nusbaum C, and Gnirke A (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome biology, 12(2), R18. BioMed Central Ltd.

Asuero a. G, Sayago a., and González a. G (2006). The Correlation Coefficient: An Overview. Critical Reviews in Analytical Chemistry, 36(1), 41–59.

Benjamini Y, and Speed T P (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic acids research, 40(10), e72.

Carvalho M da C G de, and Silva D da (2010). Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas. Ciência Rural, 735–744.

Chen Y-C, Liu T, Yu C-H, Chiang T-Y, and Hwang C-C (2013). Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PloS one, 8(4), e62856.

Henson J, Tischler G, and Ning Z (2012). Next-generation sequencing and large genome assemblies. Pharmacogenomics, 13(8), 901–915.

Kaur R, and Malik C (2013). Next Generation Sequencing: A REVOLUTION IN GENE SEQUENCING, 2(4), 1–20.

Korf I (2013). Genomics: the state of the art in RNA-seq analysis. Nature methods, 10(12), 1165–6. Nature Publishing Group.

Langmead B, and Salzberg S L (2012). Fast gapped-read alignment with Bowtie 2. Nature methods, 9(4), 357–9.

Li H, and Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England), 25(14), 1754–60.

Li H, and Durbin R (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England), 26(5), 589–95.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, and Law M (2012). Comparison of next-generation sequencing systems. Journal of biomedicine & biotechnology, 2012, 251364.

Miller J, Koren S, and Sutton G (2010). Assembly algorithms for next-generation sequencing data. Genomics, 95(6), 315–327.

Ning Z, Cox A J, and Mullikin J C (2001). SSAHA: A Fast Search Method for Large DNA Databases, 1725–1729.

Ross M G, Russ C, Costello M, Hollinger A, Lennon N J, Hegarty R, Nusbaum C, and Jaffe D B (2013). Characterizing and measuring bias in sequence data. Genome biology, 14(5), R51. BioMed Central Ltd.

Salmela L (2010). Correction of sequencing errors in a mixed set of reads. Bioinformatics (Oxford, England), 26(10), 1284–90.

Schlebusch S, and Illing N (2012). Next generation shotgun sequencing and the challenges of de novo genome assembly. South African Journal of Science, 108(11/12), 1–8.

Shendure J, and Ji H (2008). Next-generation DNA sequencing. Nature biotechnology, 26(10), 1135–45.

Soares S C, Silva A, Trost E, Blom J, Ramos R, Carneiro A, Ali A, Santos A R, Pinto A C, Diniz C, Barbosa E G V, Dorella F a, Aburjaile F, Rocha F S, Nascimento K K F, Guimarães L C, Almeida S, Hassan S S, Bakhtiar S M, Pereira U P, Abreu V a C, Schneider M P C, Miyoshi A, Tauch A, and Azevedo V (2013). The pan-genome of the animal pathogen Corynebacterium pseudotuberculosis reveals differences in genome plasticity between the biovar ovis and equi strains. PloS one, 8(1), e53818.

Taylor R (1990). Interpretation of the correlation coefficient: a basic review. Journal of diagnostic medical sonography, 35–39.

Thompson J F, and Steinmann K E (2010). Single molecule sequencing with a HeliScope genetic analysis system. (F M Ausubel, R Brent, R E Kingston, D D Moore, J G Seidman, J A Smith, and K Struhl Eds)Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.], Chapter 7, Unit7.10. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Wirawan A, Harris R S, Liu Y, Schmidt B, and Schröder J (2014). HECTOR: a parallel multistage homopolymer spectrum based error corrector for 454 sequencing data. BMC bioinformatics, 15, 131.

Zeng F, Jiang R, and Chen T (2013). PyroHMMsnp: an SNP caller for Ion Torrent and 454 sequencing data. Nucleic acids research, 41(13), e136.

Zhang J, Chiodini R, Badr A, and Zhang G (2011). The impact of next-generation sequencing on genomics. Journal of Genetics and Genomics, 38(3), 95–109.