Ministério da Educação
Universidade Tecnológica Federal do Paraná
Pró-Reitoria de Pesquisa e Pós-Graduação

Relatório Final de Atividades

ITNGS: Aplicativo para análises iniciais de sequenciamentos de nova geração vinculado ao projeto Reconhecimento de padrões em sequências genômicas: um estudo de caso utilizando redes complexas

Juliana Costa Silva
Voluntária
Tecnologia em análise e desenvolvimento de sistemas
Data de ingresso no programa: 10/2013
Orientador: Prof. Dr. Fabrício Martins Lopes
Co-Orientador: Dr. Douglas Silva Domingues

Área do Conhecimento: 1.03.03.00-6 – Metodologia e Técnicas da Computação.

CAMPUS CORNÉLIO PROCÓPIO, 2014.

JULIANA COSTA SILVA
FABRÍCIO MARTINS LOPES
DOUGLAS SILVA DOMINGUES

ITNGS: Aplicativo para análises iniciais de sequenciamentos de nova geração

Relatório Técnico do Programa de Iniciação Tecnológica da Universidade Tecnológica Federal do Paraná.

CAMPUS CORNÉLIO PROCÓPIO, 2014.

Sumário

INTRODUÇÃO

As células e a transmissão de características através da herança genética são alvos de estudo a muito tempo. Em 1860, o monge Gregor Mendel [11] realizou uma pesquisa sobre herança genética em ervilhas. Em 1953, a estrutura química da molécula de DNA foi descrita [32].

A partir destas descobertas muitos estudos foram desenvolvidos acerca do código genético. Em 1977, foi desenvolvido o método Sanger [28], que tornou possível determinar em que ordem os nucleotídeos (A, T, C, G) aparecem em um fragmento de DNA. O processo de descobrir em que ordem os nucleotídeos se apresentam é chamado sequenciamento.

Com o tempo, o interesse em estudos genéticos aumentou e foi necessário encontrar formas de produzir um maior volume de sequências. Uma das formas de automatizar o método de sequenciamento Sanger foi proposta em 1986, quando uma empresa chamada Applied Biosystems [5] começou a fabricar máquinas de sequenciamento de DNA automatizadas. Atualmente, existem várias máquinas automatizadas de sequenciamento, que utilizam diferentes metodologias para se obter a sequência de DNA de diversos organismos. Os aparelhos que automatizam as técnicas de sequenciamento são chamados sequenciadores.

O princípio utilizado para realizar um sequenciamento é o processo de incorporação de nucleotídeos. Para que esta reação ocorra artificialmente, o DNA é fragmentado e são colocados no sequenciador um vetor de sequências, como DNA de bactérias (plasmídeo) ou adaptadores (pequenos fragmentos de DNA com sequência conhecida).

O DNA de bactéria ou adaptadores inseridos não representam o organismo que se esta sequenciando, e sim uma necessidade biológica para que o sequenciamento aconteça. Durante o sequenciamento, o DNA bactéria e/ou adaptadores inseridos passam pela incorporação de nucleotídeos, onde cada nucleotídeo incorporado é registrado pelo sequenciador.

O registro dos nucleotídeos pode apresentar erros, por isso o sequenciador também registra o grau de confiança para cada nucleotídeo registrado, chamado nota de qualidade. A nota de qualidade representa o quanto se pode ter certeza de que o nucleotídeo registrado é de fato o nucleotídeo incorporado.

O DNA de bactérias, adaptadores, e os nucleotídeos com nota de qualidade baixa são chamados artefatos de sequenciamento.

Após o sequenciamento é possível visualizar em arquivos de texto as sequências de DNA. O sequenciador registra todos os nucleotídeos, inclusive os de bactérias ou adaptadores. É necessário separar os artefatos das sequências, antes de utilizar os dados gerados pelo sequenciador.

Este trabalho tem como objetivo o desenvolvimento de uma ferramenta que remova artefatos de sequências de DNA, com interface gráfica e utilizando uma nova abordagem de remoção, com o objetivo de facilitar e melhorar este processo.

OBJETIVOS E JUSTIFICATIVA

Depois da criação dos primeiros sequenciadores em 1986, muitos outros sequenciadores foram criados. Fabricantes comercias de sequenciadores introduziram no mercado aparelhos que conseguem produzir muito mais dados do que os primeiros aparelhos fabricados baseados na metodologia Sanger [21].

Os sequenciadores que produzem uma grande quantidade de dados em um tempo relativamente menor, são chamados sequenciadores de nova geração. A ferramenta desenvolvida neste trabalho tem como objetivo remover os artefatos de sequências de DNA obtidas através do sequenciamento feito em aparelhos de nova geração, inicialmente apenas dos sequenciadores Sanger [28], Illumina Solexa [3], e Roche 454 [20].

Existem algumas ferramentas que executam a remoção/ mascaramento de artefatos para dados de nova geração, como a ferramenta Cross-match [14] que mascara com um 'X' sequências iguais as de vetores e/ou adaptadores, ou ferramentas do Kit Fastx [13], que é um conjunto de ferramentas para conversão e tratamento de dados de sequenciamento. Estas ferramentas são de difícil utilização para usuários que não conhecem linguagens utilizadas para acesso via linha de comando, já que não possuem interface gráfica. Isto faz com que o proprietário de dados genômicos não consiga tratá-los, já que a maioria dos usuários deste tipo de dados são biólogos, geneticistas ou pesquisadores de áreas correlacionadas.

A criação de uma ferramenta que possibilite realizar a remoção de artefatos e escolher como será feita esta remoção com interface gráfica é o principal objetivo deste trabalho.

A dificuldade em utilizar ferramentas computacionais que não possuem interface gráfica, faz com que recursos deixem de ser utilizados. Um bom exemplo desta situação em bioinformática é a ferramenta chamada BLAST *(Basic Alignment Search Tool)* [2], com um grande número de citações em artigos no período de 2000 e 2013, é uma das ferramentas mais utilizadas quando existe a necessidade de buscar informações sobre sequências genômicas e transcritas. Esta ferramenta possui duas formas de uso: é possível utiliza-la via web (com restrições relativas ao tamanho da consulta enviada) e localmente via linha de comando (com restrições de hardware).

Apesar das restrições a versão web é a mais utilizada, já que os usuários (em sua maioria biólogos e pesquisadores de áreas correlacionadas) encontram muitas dificuldades para realizar a instalação e a execução da ferramenta de forma local, que não possui implementação com interface gráfica.

A remoção de artefatos de sequências pode ser realizada em alguns sites, nos quais é possível enviar as sequências, escolher como a remoção será feita e acessar o resultado após o processamento. Como EGassembler, PRINSEQ e Galaxy.

Os sites citados a pouco tem como objetivo realizar a remoção de artefatos de sequências de nova geração e outras análises. As técnicas utilizadas se diferem em alguns pontos e são semelhantes ou iguais em outros. A dificuldade de utilização aparece quando os arquivos são maiores que 1024 MB, o que é comum quando se trata de sequências de nova geração. Enviar estes arquivos para os servidores das ferramentas se torna uma tarefa difícil e nem sempre bem sucedida, até por que em geral o tamanho de arquivos é limitado a 2048 MB.

A partir desta observação, a criação de uma ferramenta que integra funcionalidades de remoção de artefatos para sequências de nova geração, torna a execução local e com interface gráfica, é a ideia explorada neste trabalho.

Este trabalho apresenta uma ferramenta com interface gráfica, para tratamento de sequências de DNA obtidas através de sequenciadores de nova geração.

REVISÃO BIBLIOGRÁFICA

Ácido Desoxirribonucleico: DNA. O ácido desoxirribonucleico (DNA) é um polímero composto por unidades de nucleotídeos (desoxirribonucleotídeos) também chamados de bases. Os nucleotídeos que formam as sequências de DNA dos seres vivos são adenina, representada pela letra A, guanina, representada pela letra G, citosina representada pela letra C e timina representada pela letra T. Estes nucleotídeos formam uma fita de dupla hélice [32].

Os nucleotídeos fazem ligações entre si de forma especifica, A só faz ligação com T, G só faz ligação com C. Assim ao observarmos um lado de uma fita de DNA é possível saber o que contém na outra [17].

Sequências de nova geração. Ao observar o DNA, suas quatro bases, e a quantidade de cada uma, não é possível obter muitas informações. A maior preocupação ao observar as sequências é na disposição e frequência destes quatro nucleotídeos [1]. As metodologias para se observar esta disposição foram evoluindo com a necessidade de estudos mais apurados acerca das sequências genéticas.

Em abril de 2003 o projeto genoma humano terminou sequenciamento e mapeamento todo o genoma humano, o custo para realizar o sequenciamento de um genoma era alto, com o interesse da comunidade cientifica em estudos genéticos e o empenho para o projeto genoma humano, as técnicas de sequenciamento evoluíram muito, e em pouco tempo, entre 1990 e 2003 (início e fim do projeto Genoma Humano) o custo de sequenciamento foi reduzido de forma drástica. Neste período foram criados aparelhos mais rápidos e técnicas de custo reduzido, para sequenciamento [22].

O sequenciamento de nova geração pode ser feito de duas formas, *paired-end* ou *single-end*. Enquanto na técnica *single-end* os nucleotídeos são incorporados do começo ao fim de um fragmento, na técnica *paired-end* duas cópias do mesmo fragmento recebem a incorporação de nucleotídeos, partindo de extremidades opostas. Como apresentado na Figura 1.

O material genético utilizado para sequenciamento pode ser obtido através da inserção de fragmentos de DNA em bactérias, que replicam o DNA inserido durante o processo de replicação, esta forma de obter cópias do material genético é chamada BACs *(Bacterial artificial chromosome)*.

O sequenciamento de DNA pode ser de genomas inteiros, de pedaços de genomas clonados em BACs, ou de cDNA com mRNA oriundos de diferentes tecidos ou condições. As metodologias trazem informações diferentes, e idealmente para um projeto genoma usa-se as duas [12]. Como resultado do sequenciamento são gerados dados em pedaços, já que os aparelhos que realizam esta tarefa possuem uma limitação de tamanho de leitura. Este DNA é quebrado em pequenos fragmentos, os fragmentos recebem adaptadores (pequenas sequências de DNA) em suas extremidades como os cortes não são exatos e o material foi replicado, existem áreas que se repetem, ou seja, são sequenciadas duas vezes [4].

Todo este processo faz com que alguns tratamentos sejam necessários antes de fazer qualquer afirmação acerca de dados de um sequenciamento. Tanto a bactéria quanto

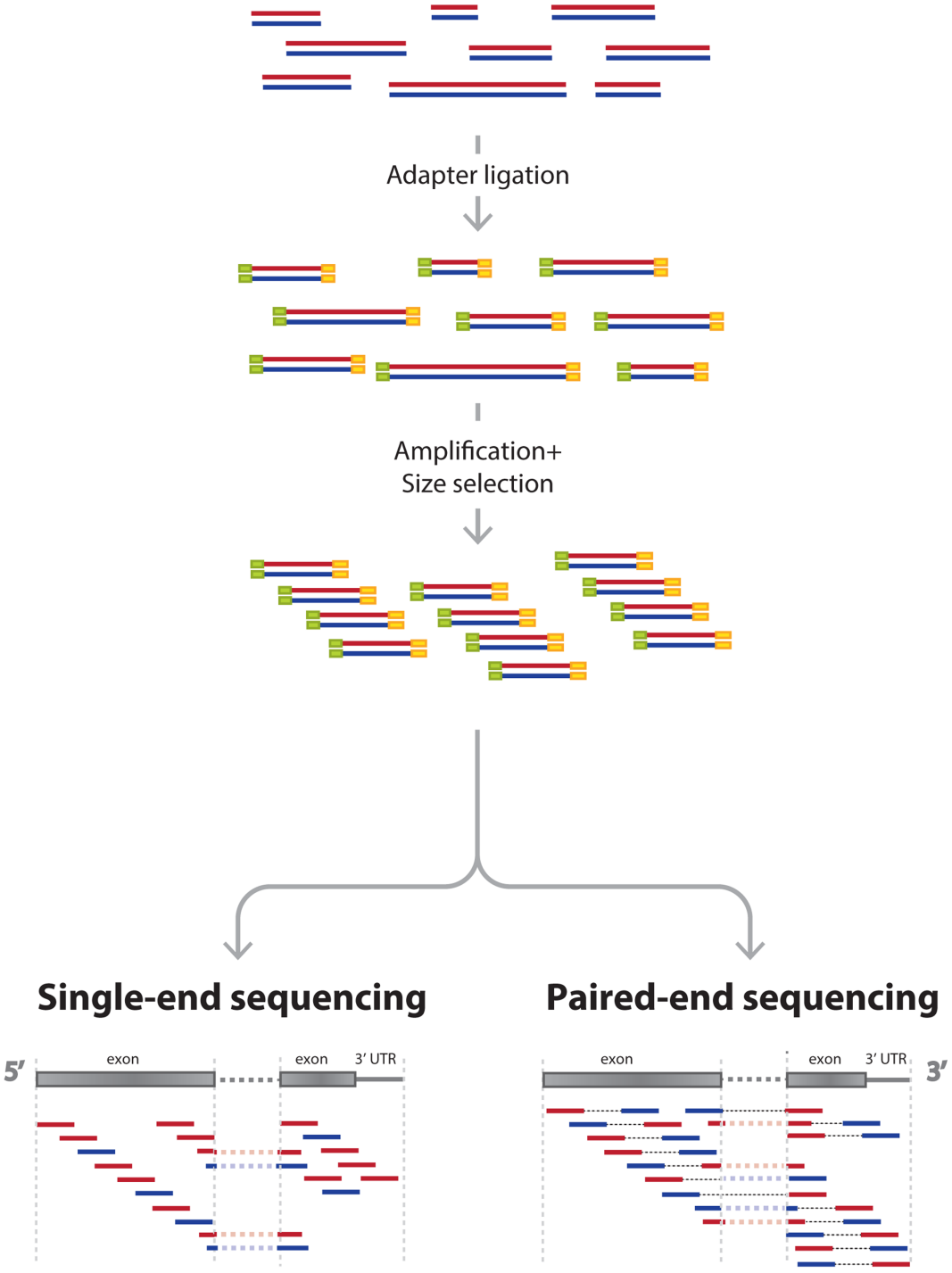

Figura 1: Adaptadores são ligados as extremidades das sequências **Fonte:** (ZHERNAKOVA et al., 2013)

os adaptadores são lidos em conjunto com as sequências, remover estes adaptadores ou contaminantes e as sequências com baixa qualidade é essencial para a obtenção de dados puros [22].

Os sequenciadores de nova geração, possibilitam realizar sequenciamento em um laboratório relativamente simples, com custos relativamente pequenos [21]. São considerados sequenciadores de nova geração os aparelhos que fazem sequenciamento em larga

escala, ou seja, que conseguem ler milhões de sequências em um tempo relativamente curto se comparado a outros aparelhos.

Estudos comparativos de sequenciadores de nova geração observaram que cada aparelho utiliza uma forma de reação química para obter sequências, e os arquivos de saída de cada aparelho devem ser tratados de forma especifica em alguns aspectos [22].

Em geral as saídas dos sequenciadores de nova geração geram vários tipos de arquivos, entre eles o FASTQ, ou arquivos que são facilmente convertidos para este formato, como os arquivos de extensão FASTA associados a arquivos de extensão QUAL.

Complexidade de sequências. Sequências com distribuição de nucleotídeos desequilibrada, em geral apresentam informações com pouco significado biológico [23]. A complexidade de qualquer sequência de caracteres pode ser calculada com a entropia de Sahnnon [29]. A entropia mostra o quanto de informação é necessária para reproduzir uma sequência genômica (ou qualquer conjunto de caracteres). A entropia de Shannon foi o método escolhido para verificar a complexidade das sequências, visto que a baixa complexidade pode indicar não só erros de sequenciamento mas regiões de sequências poli-A e poli-T, por exemplo. O valor de complexidade de qualquer sequência de caracteres pode ser calculado por:

$$H = -\sum_{i} p_i \, log_2 \, p_i \tag{1}$$

Utilizando-se de log na base 2, o valor de entropia é dado em bits. Esta técnica observa o quanto os 4 nucleotídeos possíveis se repetem, caso a entropia seja muito alta, indica uma sequência pouco complexa, ou seja, um nucleotídeo aparece muito na sequência, o que pode indicar um erro no sequenciamento. As sequências de baixa complexidade são excluídas, antes mesmo da análise de nota qualidade.

Arquivos FASTA. O formato FASTA utiliza o símbolo "$>$" para indicar que a linha traz o nome a outras informações sobre a sequência, nas linhas seguintes traz a sequência de nucleotídeos. Cada sequência representa uma leitura (*read*) do aparelho, quando se trata de dados ainda sem tratamento.

```
>SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
```

Figura 2: Exemplo de arquivo FASTA **Fonte:** (COCK et al., 2010)

O formato FASTA é muito utilizado para disponibilizar sequências sendo que a maior parte das sequências depositadas em bancos públicos, como genomas e parte deles estão em formato FASTA. Muitas ferramentas de análise de sequências exigem como arquivos de entrada sequências no formato FASTA.

***Scores* de qualidade PHRED e formato QUAL.** O software PHRED faz a leitura de arquivos de sequências de DNA, e atribui um valor de qualidade a cada base [10]. O cálculo PHRED foi criado em 1992 e melhorado em 1995. PHRED considera métricas de sequenciamento Sanger como resolução e formato de pico para cada base, relacionando estas informações a grandes tabelas de pesquisa multivariadas [9]. Esta metodologia provou ser muito precisa [27]. Devido a sua eficácia se tornou um padrão exigido em aparelhos comerciais de sequenciamento.

Os sequenciadores de nova geração se diferenciam pela química utilizada, mas o processo de geração da pontuação de qualidade é o mesmo de forma geral.

$$Q_{PHRED} = -10\, x\, log_{10}\,(P_e) \tag{2}$$

Onde P_e é a probabilidade de erro na codificação desta base, ou seja, quando Q é 30, para uma base por exemplo, isto é equivalente a 1 base com leitura incorreta a cada 1.000 vezes, a precisão de leitura é de 99,9%. Para um *score* PHRED de 20 pode-se afirmar que a probabilidade de uma leitura incorreta é de uma em 100, o que nos diz que a cada 100 bases 1 estará errada, com precisão de leitura a 99%.

Para armazenar estes valores PHRED pode apresentar outro formato de arquivo, o formato QUAL.

```
>SRR014849.1 EIXKN4201CFU84 length=93
18 10 5 3 2 1 1 1 1 1 1 1 1 1 1 22 37
31 22 16 11 6 1 26 34 30 11 33 26 30 21
33 26 25 36 32 16 36 32 16 36 32 20 6
24 33 25 30 25 2 24 36 32 15 35 31 17
36 32 20 6 25 29 20 30 25 4 32 26 32 23
32 26 30 24 33 26 35 31 14 28 27 30 22
28 24 27 17 32 23 28 28
```

Figura 3: Exemplo de arquivo QUAL **Fonte:** (COCK et al., 2010)

O arquivo QUAL é associado um arquivo FASTA, como o apresentado na Figura 2 o arquivo QUAL traz o símbolo > seguido do mesmo nome e informações do FASTA a ele relacionado, nas linhas seguintes apresenta o *score* de qualidade de cada nucleotídeo do FASTA como observado na Figura 3.

Arquivos FASTQ. O formato FASTQ passou a ser muito utilizado por ser um formato que facilita o acesso a informação, ele combina dados numéricos e alfabéticos em um único arquivo, facilitando assim a extração de informações [8]. Diferentemente do arquivo FASTA associado ao QUAL, é possível obter a sequência e seu *score* de qualidade em um único arquivo, facilitando a organização e o uso deste.

O arquivo FASTQ contém algumas informações sobre uma sequência:

- O símbolo "@" seguido de identificação do *read* (pode ser numérica alfabética, ou alfanumérica);

- A sequência de nucleotídeos lida;
- O símbolo de soma "+", em alguns casos seguido do nome da sequência;
- Notas de qualidade;

Exemplo da composição de um arquivo FASTQ é apresentado na figura 4:

```
@UNC13-SN749:143:C0CA2ACXX:6:1101:4320:1971 2:N:0:TGCTGT
TTAAATCACAAGCACTGAATTAAGAAAAGAATCAAGAATGAGAAAATCCATCTTGTTTCATCCAGAAAA
+
@CCFFFBFHHDFGIIGGIJIIIJGGGIJGGHIGIIIIEIGIIHIIIJGE@BFEDFGIIJIJGGIJCFDH
```

Figura 4: Exemplo de FASTQ **Fonte:** Autoria própria

A última linha traz as notas de qualidade. As notas mostram o quão precisa foi a leitura daquela letra naquela posição da sequência, ou seja, notas baixas simbolizam sequências que tendem ao erro. Os valores são alfabéticos devido a nota de qualidade, que varia entre -5 e 93, e permite sua representação utilizando 1 caractere, já que a maior parcela dos valores possíveis são representados numericamente por 2 dígitos.

Cada letra representa o valor da nota para o nucleotídeo acima dela na mesma posição, esta nota é obtida através da conversão do caractere para código ASCII *(American Standard Code for Information Interchange)*.

A tabela ASCII foi criada em 1963 com o objetivo de padronizar a codificação entre computadores de todos os fabricantes [24]. ASCII apresenta cada símbolo alfabético como um valor numérico padrão. Por este motivo esta codificação foi escolhida para representar valores de qualidade. A partir do valor de referência na tabela ASCII é retirado 33 ou 64 dependendo do sequenciador que gerou os dados, o resultado deste cálculo é o valor da a nota de qualidade do nucleotídeo [8].

Variação da nota de qualidade. Em 2004 a Solexa Inc. apresentou uma nova versão de FASTQ, na qual a nota de qualidade é calculada como na equação (2):

$$QSolexa = -10\,x\,log_{10}(\frac{P_e}{1\,-\,P_e}) \tag{3}$$

A Solexa, Inc., foi adquirida pela Illumina, Inc., em 2006. A variante fastq-solexa apresentada na Tabela 1 continuou sendo utilizada. Entretanto a pipeline Genome Analyser a partir da versão 1.3 passou a utilizar notas de qualidade PHRED, a Illumina resolveu não utilizar o formato fastq-sanger apresentado na Tabela 1, criaram o terceiro tipo de FASTQ diferente dos arquivos fastq-solexa, este novo formato codifica a nota PHRED com um a partir do código 64 da tabela ASCII e pode atingir *scores* de 0 a 62. Embora atualmente, para os sequenciadores Illumina 1.8 e posteriores, em dados de Illumina são esperados *scores* de 0-40, o formato fastq-illumina como apresentado na Tabela 1 ainda deve ser considerado.

A conversão da nota de qualidade é realizada de forma diferente para cada arquivo, como observado na Tabela 1.

Tabela 1: As três variantes descritas FASTQ, com colunas: descrição, o nome do formato usado em projetos OBF *Open Bioinformatics Foundation*, a gama de ASCII, valor a compensar da codificação ASCII, o tipo de índice de qualidade codificado e a gama possível de pontuação **Fonte:** (COCK et al., 2010)

Description, OFB name	ASCII characters		Quality score	
	Range	Offset	Type	Range
Sanger standard fastq-sanger	33-126	33	PHRED	0 to 93
Solexa/ early Illumina fastq-solexa	59-126	64	Solexa	-5 to 62
Illumina 1.3 + fastq-illumina	64-126	64	PHRED	0 to 62

Alinhamento de sequências. Alinhamento de sequências pode ser definido como a comparação entre duas ou mais sequências, por meio de uma série de caracteres na mesma ordem. Como indicado na Figura 5, o esquema de um alinhamento busca regiões similares, onde podem aparecer mais ou menos caracteres e também caracteres diferentes. Cada técnica de alinhamento trata estas variações de uma forma.

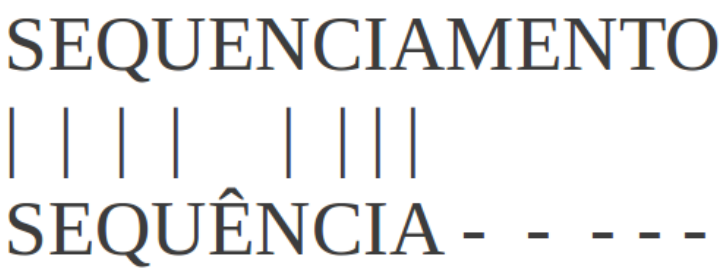

Figura 5: Componentes de alinhamento de sequências **Fonte:** (MAIA; OLIVEIRA, 2011)

Neste trabalho, o processo de alinhamento tem como objetivo encontrar regiões onde ficaram fragmentos de vetor, ou adaptadores (sequências que não pertencem ao organismo que foi sequenciado). Existem dois tipos de alinhamento, alinhamento global e alinhamento local.

O alinhamento local (escolhido para este trabalho) compara e pontua as igualdades penalizando as diferenças, e faz isso por região, não necessariamente alinha uma sequência inteira contra a outra, mas regiões de uma sequência que sejam iguais (ou muito parecidas) com a outra. Os dois alinhamentos utilizam uma matriz de pontuação para identificar os melhores alinhamentos, a forma de pontuar analisar a matriz é um dos motivos de diferença entre o resultado de cada método [19]. A ferramenta de alinhamento escolhida para fazer o alinhamento foi o MegaBLAST.

Um alinhamento sem lacunas pode ser definido como duas sequências idênticas en-

tre si, nos caracteres e na ordem em que se encontram. Chama-se de *match* (igual/ compatível) dois nucleotídeos idênticos encontrados em um alinhamento, são chamados de *mismatch* (incompatibilidade) nucleotídeos diferentes, e chama-se *gap* (lacuna/ abertura/ brecha) nenhum nucleotídeo encontrado. A Figura 5 apresenta os elementos de um alinhamento.

Para cada *match*, *mismatch* e *gap* ocorridos durante o alinhamento é dada uma pontuação. A maior pontuação é dada para *matches*.Como a pontuação de *mismatch* e *gap* é negativa elas são chamadas penalidades.

MegaBLAST. O MegaBLAST [31] é otimizado para o alinhamento de sequências que possuem pequenas diferenças decorridas de erros de sequenciamento ou de variações naturais [6]. MegaBLAST utiliza um algoritmo de alinhamento conhecido como guloso. Algorítimos gulosos se caracterizam por partirem de problemas grandes, e reduzi-los a problemas menores.

Para alinhamento de sequências o MegaBLAST busca semelhanças, assim parte da diferença entre as duas sequências, onde um alinhamento é avaliado pela contagem de nucleotídeos que não se alinham. Por isso, para alinhar sequências que se diferem somente por erros de sequenciamento, ou outras influências o algorítimo guloso pode ser uma forma mais rápida e tão eficaz quanto as outras metodologias [33]. Neste contexto o MegaBLAST foi escolhido como o alinhador de sequências para este trabalho.

O MegaBLAST gera um relatório que contem informações sobre os alinhamentos (se encontrados). Como o MegaBLAST é acessado em segundo plano, o tipo de relatório fixo escolhido foi o relatório padrão do BLAST. Este modelo de relatório traz uma grande quantidade de informações, inclusive sobre o banco de dados utilizado, como fatores K e Lambda de [16] que são utilizados para avaliar a relevância de cada alinhamento.

MATERIAIS E MÉTODOS

O tratamento inicial de sequências é realizado com dados sequenciados e sem nenhum tratamento, chamados geralmente de dados brutos. Antes da utilização dos dados, é necessário remover a maior quantidade possível de erros de sequenciamento e contaminações. O objetivo é que fiquem apenas dados de interesse, e então seja possível realizar tentativas de reordenar a sequência (montagem), de forma que sua disposição fique o mais próximo possível de como ela está ordenada no organismo.

O tratamento de sequências é feito em duas etapas: 1) Remoção de sequências de baixa qualidade; e 2) Limpeza de sequências dos adaptadores ou vetor. Durante a remoção de sequências de baixa qualidade também são removidas sequências de baixa complexidade, a análise de complexidade influi diretamente no aceite da sequência para a segunda etapa do processamento.

Neste capítulo será apresentada a metodologia utilizada para cada etapa, e a motivação de cada método escolhido.

Remoção de sequências de baixa qualidade e complexidade. Este trabalho é desenvolvido inicialmente para tratar arquivos de saída dos aparelhos Sanger [28], Illumina Solexa [3], e Roche 454 [20], que geram como saída arquivos no formato FASTQ ou arquivos facilmente conversíveis para este formato.

FASTQ é o tipo de arquivo de entrada para a limpeza de qualidade. Eliminar pares de base pares de base (pb) que não foram detectados corretamente pelo aparelho é o objetivo desta etapa do trabalho. Para que sejam eliminadas as sequências com baixa qualidade o usuário deverá escolher a nota de corte adequada a situação para delimitar o menor valor de nota aceitável.

Existem três tipos de arquivo FASTQ, descritos na Tabela 1, nota-se que o diferencial entre a nota de qualidade destes arquivos é a posição da tabela ASCII escolhida para iniciar a pontuação e o *offset*. A ferramenta apresentada neste trabalho permite o tratamento de sequências *paired-end* e *single-end*, sendo possível inserir dois arquivos de entrada para a remoção de sequências de baixa qualidade. Quando a opção de sequenciamento escolhida é single-end o arquivo 2 (se selecionado) é ignorado, e ocorre apenas o processamento do arquivo 1.

Antes da remoção de sequências com baixa qualidade, são necessárias algumas validações. A leitura do arquivo é realizada em bloco, 4 linhas por vez, devido a estrutura de arquivo FASTQ apresentada anteriormente. São observadas as regras de estrutura básica para este tipo de arquivo, as linhas do bloco devem conter respectivamente as seguintes características, 1) Descrição; 2) Apenas caracteres que contemplem uma sequência de DNA; 3) Como primeiro caractere o símbolo de adição, e 4) Apenas caracteres que representem nota entre o mínimo e máximo descrito para o sequenciador selecionado. Se o arquivo possuir as características descritas o processamento segue.

Somente sequências com complexidade maior ou igual a escolhida pelo usuário passam pela remoção de sequências de baixa qualidade. Sequências com complexidade muito alta (próximo de 1.0) tendem a ter muitas repetições e por isso (dependendo do objetivo do estudo) possuem significado biológico pouco relevante, como existe uma grande diversidade de informações a se obter de dados de sequenciamento a complexidade pode ser parametrizada de acordo com o objetivo do estudo.

O limite de início e fim da sequência pode ser definido pelo usuário, por isso, é importante observar que início é posição inicial de corte + 1 e fim é o tamanho da sequência - posição final de corte.

Como as notas de qualidade são fornecidas para cada base, é necessário avaliar em que circunstância deve-se excluir uma sequência de baixa qualidade. A metodologia utilizada para avaliar a qualidade das sequências foi a janela deslizante, também utilizada para outras análises de sequências de DNA e RNA [7] [26]. Visto que as extremidades das sequências geralmente são de baixa qualidade [7], automaticamente são criadas duas janelas com tamanho de aproximadamente 10% do comprimento da sequência, as janelas têm como ponto de partida o início e o fim da sequência como mostra a Figura 6.

A média de qualidade da janela é calculada, caso a média esteja abaixo do valor mínimo definido a janela avança uma posição em relação ao centro como mostra a Figura 6. As janelas são independentes e param quando encontram a primeira média acima ou igual ao valor mínimo definido. Quando as duas janelas encontram o trecho de qualidade ideal é verificada a distância entre a nova posição de início e fim da sequência de boa qualidade. Se a distância representa mais de 50% do tamanho original da sequência, o trecho que tem qualidade ideal é escrito em dois arquivos de saída, nos formatos FASTA e FASTQ. A média de qualidade da sequência registrada em um arquivo de extensão ".txt", que serve de base para a geração de um gráfico que relaciona a média de qualidade e a quantidade de sequências que obtiveram esta média.

Para sequências *paired-end* é realizada uma comparação, onde, se as duas sequências

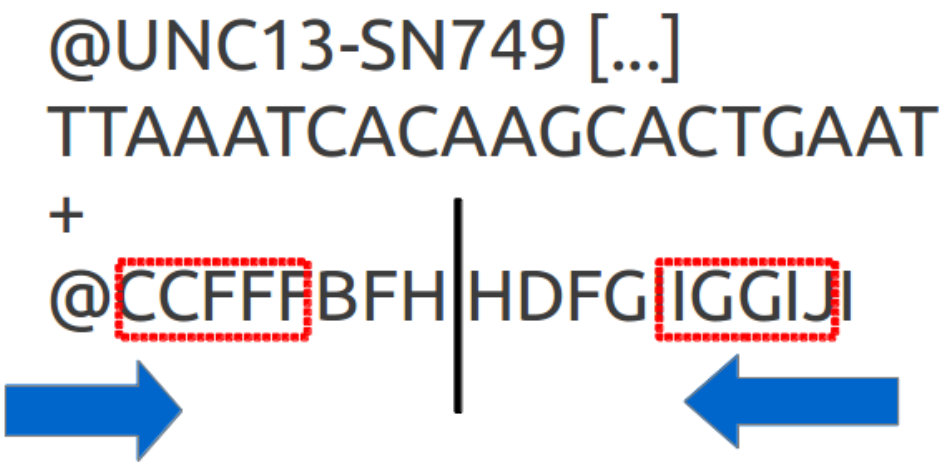

Figura 6: Esquema de funcionamento da janela deslizante. As janelas são independentes, o que permite que uma pare e a outra continue em movimento. **Fonte:** Autoria própria.

possuem qualidade e complexidade aceitável estas são escritas em seus respectivos arquivos de saída. Caso apenas uma das sequências obtenha qualidade e complexidade aceitável ela será escrita em um arquivo single (este arquivo contem o nome do arquivo 1 + "_single" e extensão FASTQ e FASTA). As sequências aprovadas na fase de remoção por qualidade são utilizadas para a remoção de vetor e/ou adaptadores.

Remoção do Vetor ou adaptadores. Para a remoção de artefatos a ferramenta de alinhamento escolhida foi o MegaBLAST. Esta ferramenta do pacote de aplicativos BLAST+, está disponível para download em ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/ [6], e é uma dependência para a fase de remoção de vetores e adaptadores deste trabalho em sistemas operacionais Linux. Para usuários de sistemas operacionais Windows os executáveis megablast.exe e formatdb.exe acompanham a ferramenta.

A execução do MegaBLAST é feita via linha de comando, a aplicação aguarda a execução do comando enviado para o sistema, que por sua vez retorna se obteve sucesso ou não durante a execução.

Para que o alinhamento seja realizado é necessário informar alguns parâmetros, como a sequência que será alinhada e contra qual banco de dados será alinhada. O banco de dados utilizado será sempre a sequência de vetor/ adaptadores que queremos remover (artefatos da sequência), estas sequências podem ser escolhidas pelo usuário, entretanto caso o usuário não tenha as sequências de vetor adaptadores utilizados no sequenciamento, é possível utilizar o arquivo FASTA UniVec [18], que contem vetores e adaptadores geralmente utilizados, este arquivo é disponibilizado pelo NCBI *(National Center for Biotechnology Information)* e acompanha o executável desta ferramenta.

Ao comparar sequências com objetivo de encontrar alinhamento entre elas, chamaremos de 'consulta' as sequências onde se busca um padrão ou informação, e chamaremos de 'banco de dados' as sequências que você acredita que estão contidas na sua consulta. As sequências que foram aprovadas na limpeza de qualidade serão utilizadas como consulta, como banco de dados serão utilizadas as sequências de vetor/ adaptadores.

Antes da execução do MegaBLAST é necessário criar o banco de dados. Para que ele seja reconhecido pelo MegaBLAST é necessário formatar as sequências nele contidas

em formato de banco de dados BLAST, a obtenção deste banco formatado é possível através do executável *formatdb*. Este executável também faz parte do kit de ferramentas do BLAST+, e exige como entrada um arquivo no formato FASTA, o nome do banco que se deseja criar e o tipo de banco (nucleotídeo ou proteína) [6]. A criação deste banco de dados que é feita via linha de comando pelo ITNGS, sem a necessidade de comando do usuário.

Para sequências escolhidas pelo usuário, o nome do banco sempre será *database*, assim ao escolher um arquivo FASTA ele será copiado para o local de execução desta ferramenta, e renomeado para *database.fasta*. Após a execução do MegaBLAST serão avaliadas quais sequências possuem alta similaridade com vetor/ adaptadores e estas devem ser descartadas. As sequências que tiverem pouca similaridade devem ser mantidas. Esta avaliação esta diretamente ligada aos *scores* de alinhamento gerados pelo MegaBLAST.

Após o processamento do MegaBLAST, é realizada a leitura do relatório gerado e extraído o melhor alinhamento de cada sequência. De cada alinhamento ficam registrados apenas o *score* e a identificação da sequência. Para encontrar um *score* ideal de corte que represente sequências com pouca ou nenhuma contaminação, os *scores* são convertidos em uma distribuição normal, com grau de liberdade 0,5 e considerados apenas os *scores* abaixo ou igual ao valor t gerado desta distribuição, este valor t representará o *score* de limiar para corte.

Para registrar as informações de *score* e nome das sequências é utilizado um HashMap [25], onde a chave é o nome da sequência, e o valor relacionado a cada chave é o *score*. Para selecionar as sequências que possuem *score* $\leq$ t, ou que não tiveram alinhamento com vetor/ adaptadores é feita a leitura do arquivo de saída da primeira etapa (remoção de sequências de baixa qualidade) e verificado se cada uma das sequências estão ou não no HashMap, caso a sequência esteja no HashMap ainda é necessário verificar se o *score* de alinhamento a ela atribuído é $\leq$ t. Ao satisfazer as duas condições a sequência é escrita em arquivos com extensão FASTA e FASTQ e nomeado com o nome do arquivo original acrescido de "_clean". As sequências que não aparecem no alinhamento são diretamente escritas nos arquivos, pois o fato de não ocorrer um alinhamento leva a crer que não possuem vetor/ adaptadores em sua composição.

Esta ferramenta foi desenvolvida em linguagem JAVA [15]. Na próxima seção serão apresentados mais detalhes sobre a linguagem e o modelo de desenvolvimento.

RESULTADOS E DISCUSSÕES

A ferramenta desenvolvida foi nomeada ITNGS, uma abreviação para *Initial Treatment of Next Generation Sequences*. Com o objetivo de averiguar a eficácia da metodologia utilizada para a remoção de artefatos de sequências oriundas de sequenciadores de nova geração, foram realizados testes com sequências de bancos públicos, onde são disponibilizados dados brutos (*raw data*) de sequenciadores de nova geração utilizados em estudos, o banco de dados do NCBI que armazena estas sequências é chamado SRA e disponibilizado em: http://www.ncbi.nlm.nih.gov/sra. É possível buscar dados brutos refiando a busca por sequenciador, projeto, organismo entre outros.

Os dados selecionados foram o estudo SRR096789, oriundo de Papaver bracteatum sequenciado em um aparelho 454 GS FLX Titanium com a metodologia *single-end*, a escolha dos dados se justifica no tamanho do conjunto e no aparelho utilizado no se-

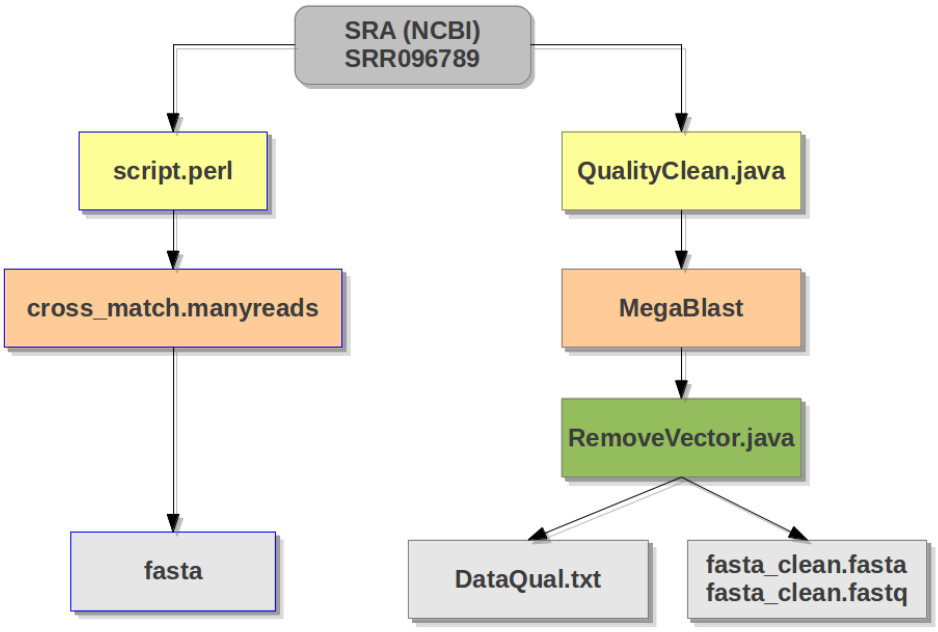

Figura 7: Os dados do banco SRR096789, passaram pela limpeza de qualidade e alinhamento com vetor/ adaptadores com duas metodologias deferentes, para possibilitar a comparação dos resultados. **Fonte:**Autoria própria.

quenciamento, já que dados com mais que 1024MB tornaram a execução dos testes de comparação demorados e inviáveis, pois a ferramenta CrossMatch exige processador e RAM acima dos disponibilizados para o estudo.

Esses dados passaram pela limpeza na ferramenta desenvolvida neste trabalho, foram analisados os resultados de cada fase (qualidade e vetor/ adaptadores). Os resultados foram comparados com o script clean_solexa.pl em linguagem PERL disponibilizado em http://www.lge.ibi.unicamp.br/cursobioinfo2013/, o script analisa a complexidade e qualidade de arquivos FASTQ, para a remoção de sequências de baixa qualidade o script calcula a média de qualidade de cada read, se a média e a complexidade estiverem abaixo do parametrizado os reads inteiros são excluídos. Para a comparação da remoção de vetor/adaptadores foi utilizado o software CrossMatch que mascara as sequências identificadas como vetor/ adaptadores com o caractere 'X'.

Os parâmetros de qualidade utilizados foram os apresentados a seguir:

- Mínimo de média de qualidade: 20;
- Complexidade máxima: 0.6.

Os parâmetros de alinhamento utilizados foram os apresentados a seguir:

- MegaBLAST: padrão da ferramenta;
- CrossMatch: padrão da ferramenta, e *Score* mínimo: 20.

Os testes foram realizados com o banco de SRA SRR096789 que contém sequências single-end geradas pelo sequenciador 454 GS FLX Titanium. Os resultados dos testes

foram registrados de duas formas: pela contagem de *reads* e pela contagem de nucleotídeos, como apresentado nas Tabelas 2 e 3:

Tabela 2: Quantidade de reads de entrada e quantidade de reads excluídos a cada fase da limpeza. **Fonte:** Autoria própria

Fase	ITNGS	Script + CrossMatch
Entrada	595.176	595.176
Baixa qualidade	5.620	243.713
Vetor/ adaptadores	0	3.752

A Tabela 2 confirma que utilizar a média como parâmetro de seleção pode acarretar em um alto índice de descarte de sequências de boa qualidade, como o software desenvolvido faz o descarte de nucleotídeos a partir das extremidades da sequência avaliando a média de qualidade por porção (janela) a quantidade de reads excluídos é bem menor.

Ao comparar as quantidades de nucleotídeos excluídos na Tabela 3 nota-se que a ferramenta desenvolvida excluiu poucos *reads* e muitos nucleotídeos, ou seja, temos reads menores, porem de boa qualidade. Se comparado a quantidade de *reads* inteiros excluídos pela outra técnica (script), é possível afirmar que dados que poderiam ser aproveitados tenham sido excluídos. É importante salientar que o *score* de qualidade é utilizado até mesmo para a montagem e mapeamento de sequências [30].

Tabela 3: Quantidade de nucleotídeos de entrada e quantidade de nucleotídeos excluídos a cada fase da limpeza. **Fonte:** Autoria própria

Fase	ITNGS	Script + CrossMatch
Entrada	321.393.431	321.393.431
Baixa qualidade	72.658.733	134.401.747
Vetor/ adaptadores	0	128.385

A comparação entre os resultados de alinhamento deixa clara a eficácia da ferramenta CrossMatch em relação ao MegaBLAST e também reafirma a importância do tratamento das extremidades de sequências, já que vetores/ adaptadores em geral estão nas extremidades da sequência. Algumas ferramentas já desenvolvidas como Lucy [7] utilizam o *score* como fator de influência para o alinhamento com vetores. O software desenvolvido está disponível no Google Code.

CONCLUSÃO

Este trabalho apresentou uma forma de implementação para tratamento de dados brutos de sequenciamento. A remoção de sequências de baixa qualidade e de contaminações nesta implementação foi desenvolvida buscando facilitar e otimizar as técnicas já existentes. A implementação foi avaliada comparando os resultados de técnicas anteriormente descritas. Observou-se que técnicas de remoção de sequências com baixa qualidade tendem a ser muito agressivas se considerando a sequência com um todo, e que ainda é necessário buscar formas de identificar contaminantes que consigam obter uma boa identificação com pouca capacidade computacional.

Os objetivos deste trabalho foram alcançados parcialmente, visto que todas as ferramentas de alinhamento testadas não tiveram sucesso em computadores de uso doméstico, com grande quantidade de dados.

As dificuldades encontradas no desenvolvimento foram resumidamente necessidades computacionais. As ferramentas existentes em geral precisam de mais de 10 GB de RAM *(Random Access Memory)*, ainda assim um arquivo de bancos de dados público como o SRA leva mais de 72 horas para ser analisado pelo CrossMatch, por exemplo. O acesso a computadores com essa característica também torna a validação mais difícil.

A evolução de técnicas de sequenciamento traz cada vez mais desafios computacionais. Para trabalhos futuros, o objetivo é a busca de uma melhor identificação de artefatos com a linguagem JAVA, ou a pesquisa de metodologias que juntas tratem sequências de baixa qualidade e contaminantes para sequenciadores de nova geração.

Referências Bibliográficas

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Biologia molecular da célula*. Artmed, 2010.

[2] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[3] Simon Bennett. Solexa ltd. *Pharmacogenomics*, 5(4):433–438, 2004.

[4] Jan Berka, Zhoutao Chen, Michael Egholm, Brian C Godwin, Stephen Kyle Hutchison, John Harris Leamon, Gary James Sarkis, and Jan Fredrik Simons. Paired end sequencing, October 13 2009. US Patent 7,601,499.

[5] Applied Biosystems. Applied biosystems, 2013.

[6] Christiam Camacho, Thomas Madden, Ning Ma, Tao Taoa, Richa Agarwala, and Aleksandr Morgulis. BLAST Command Line Applications User Manual, July 2013.

[7] Hui-Hsien Chou and Michael H Holmes. Dna sequence quality trimming and vector removal. *Bioinformatics*, 17(12):1093–1104, 2001.

[8] Peter JA Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. The sanger fastq file format for sequences with quality scores, and the solexa/illumina fastq variants. *Nucleic acids research*, 38(6):1767–1771, 2010.

[9] Brent Ewing and Phil Green. Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome research*, 8(3):186–194, 1998.

[10] Brent Ewing, LaDeana Hillier, Michael C Wendl, and Phil Green. Base-calling of automated sequencer traces usingphred. i. accuracy assessment. *Genome research*, 8(3):175–194, 1998.

[11] Newton FREIRE-MAIA. Gregor mendel–vida e obra. *TA Queiroz, São Paulo*, 1995.

[12] C. Gibas, P. Jambeck, C. de Amorin Machado, and Milarepa Ltda. *Desenvolvendo Bioinformática: ferramentas de software para aplicações em biologia*. Campus, 2001.

[13] A Gordon and GJ Hannon. Fastx-toolkit. *FASTQ/A short-reads pre-processing tools (unpublished) http://hannonlab. cshl. edu/fastx_toolkit*, 2010.

[14] Phil Green. Documentation for phrap and cross-match. *University of Washington, Seattle*, 1999.

[15] Oracle Company Information. Java, 2014.

[16] Samuel Karlin and Stephen F Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences*, 87(6):2264–2268, 1990.

[17] G. Karp. *Biologia celular e molecular*. MANOLE, 2005.

[18] PA Kitts, TL Madden, H Sicotte, L Black, and JA Ostell. Univec database. *Available from: ncbi. nlm. nih. gov/VecScreen/UniVec. html.[Links]*, 2011.

[19] I. Korf, M. Yandell, and J. Bedell. *BLAST*. BLAST (Electronic resource). O'Reilly Media, Incorporated, 2003.

[20] Mark D'Ascenzo Lindsay Freeberg. Gs flx sequencing, 2014.

[21] Elaine R Mardis. Next-generation sequencing platforms. *Annual Review of Analytical Chemistry*, (0), 2013.

[22] Michael L Metzker. Sequencing technologies—the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

[23] Aleksandr Morgulis, E Michael Gertz, Alejandro a Schäffer, and Richa Agarwala. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *Journal of computational biology : a journal of computational molecular cell biology*, 13(5):1028–40, June 2006.

[24] Computer History Museum. Internet history, 2004.

[25] Oracle and its affiliates. Hashmap api, 2014.

[26] V. Proutski and E. Holmes. SWAN: sliding window analysis of nucleotide sequence variability. *Bioinformatics*, 14(5):467–468, June 1998.

[27] Peter Richterich. Estimation of errors in "raw" dna sequences: a validation study. *Genome research*, 8(3):251–259, 1998.

[28] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, December 1977.

[29] Claude E Shannon. A note on the concept of entropy. *Bell System Tech. J*, 27:379–423, 1948.

[30] Andrew D Smith, Zhenyu Xuan, and Michael Q Zhang. Using quality scores and longer reads improves accuracy of Solexa read mapping. *BMC bioinformatics*, 9(1):128, January 2008.

[31] Tao Tao. Standalone blast setup for unix. 2010.

[32] James D Watson, Francis HC Crick, et al. Molecular structure of nucleic acids. *Nature*, 171(4356):737–738, 1953.

[33] Z Zhang, S Schwartz, L Wagner, and W Miller. A greedy algorithm for aligning DNA sequences. *Journal of computational biology : a journal of computational molecular cell biology*, 7(1-2):203–14, January 2004.