Ciência Rural, Santa Maria, v.40, n.3, p.735-744, mar, 2010
ISSN 0103-8478

Sequenciamento de DNA de nova geração e suas aplicações na genômica de plantas

Next generation DNA sequencing and its applications in plant genomics

Mayra Costa da Cruz Gallo de Carvalho[I] Danielle Cristina Gregorio da Silva[II]

- REVISÃO BIBLIOGRÁFICA -

RESUMO

As plataformas de sequenciamento de nova geração são uma alternativa poderosa para estudos de genômica estrutural e funcional. Na genômica de plantas, os trabalhos com as novas plataformas têm sido destinados ao sequenciamento de transcritos, ressequenciamento ou sequenciamento de novo de genomas plastidiais. Neste trabalho, são detalhadas as tecnologias das plataformas mais utilizadas atualmente, bem como é revisada a aplicação dessas tecnologias na genômica estrutural e funcional de plantas.

Palavras-chave: *leituras curtas, ressequenciamento, sequenciamento de novo, genômica estrutural de plantas, transcritômica de plantas, genômica funcional.*

ABSTRACT

The next-generation DNA sequencing technologies are a powerful alternative to studies in structural and functional genomics. In plant genomics studies, the work with these new platforms has been used for the sequencing of transcripts, re-sequencing, and the de novo sequencing of plastid genomes. This research details the technological principles of the next-generation DNA sequencing platforms most used and reviews its application in structural and functional plant genomics.

Key words*: short reads, res-sequencing, de novo sequencing, plants structural genomics, plants transcriptomics, functional genomics.*

INTRODUÇÃO

As novas tecnologias de sequenciamento, denominadas de tecnologias de sequenciamento de nova geração, começaram a ser comercializadas em 2005 e estão evoluindo rapidamente. Todas essas tecnologias promovem o sequenciamento de DNA em plataformas capazes de gerar informação sobre milhões de pares de bases em uma única corrida. Dentre as novas plataformas de sequenciamento, duas já possuem ampla utilização em todo o mundo: a plataforma 454 FLX da Roche e a Solexa da Illumina. Outros dois sistemas de sequenciamento que começam a ser utilizados são a plataforma da *Applied Biosystems*, denominada *SOLiD System*, e o *HeliscopeTrue Single Molecule Sequencing* (tSMS), da Helicos. Essas novas plataformas possuem como características comuns um poder de gerar informação muitas vezes maior que o sequenciamento de Sanger, com uma grande economia de tempo e custo por base para o sequenciamento. Essa maior eficiência advém do uso da clonagem *in vitro* e de sistemas de suporte sólido para as unidades de sequenciamento, não precisando mais do intensivo trabalho laboratorial de produção de clones bacterianos, da montagem das placas de sequenciamento e da separação dos fragmentos em géis. A clonagem *in vitro* em suporte sólido permite que milhares de leituras possam ser produzidas de uma só vez com a plataforma 454, Solexa ou SOLiD.

[I]Departamento de Ciências Biológicas, Universidade Estadual Paulista (UNESP), Avenida Dom Antonio, 2100, 19806-900, Assis, SP, Brasil. E-mail: mayra@assis.unesp.br. Autor para correspondência.
[II]Departamento de Biologia e Tecnologia, Universidade Estadual do Norte do Paraná (UENP), Campus Luiz Meneghel, Bandeirantes, PR, Brasil.

Recebido para publicação 23.03.09 Aprovado em 09.11.09

Pirosequenciamento e a tecnologia 454

O sistema 454 foi a primeira plataforma de sequenciamento de nova geração a ser comercializada. A plataforma 454 realiza o sequenciamento baseado em síntese, o pirosequenciamento (RONAGHI et al., 1998). A leitura da sequência nesse sistema é realizada a partir de uma combinação de reações enzimáticas que se inicia com a liberação de um pirofosfato, oriundo da adição de um desoxinucleotídeo à cadeia. Em seguida, esse pirofosfato é convertido para ATP, pela ATP sulfurilase, sendo este utilizado pela luciferase para oxidar a luciferina, produzindo um sinal de luz (Figura 1) capturado por uma câmera CCD (*charge-coupled device*) acoplada ao sistema.

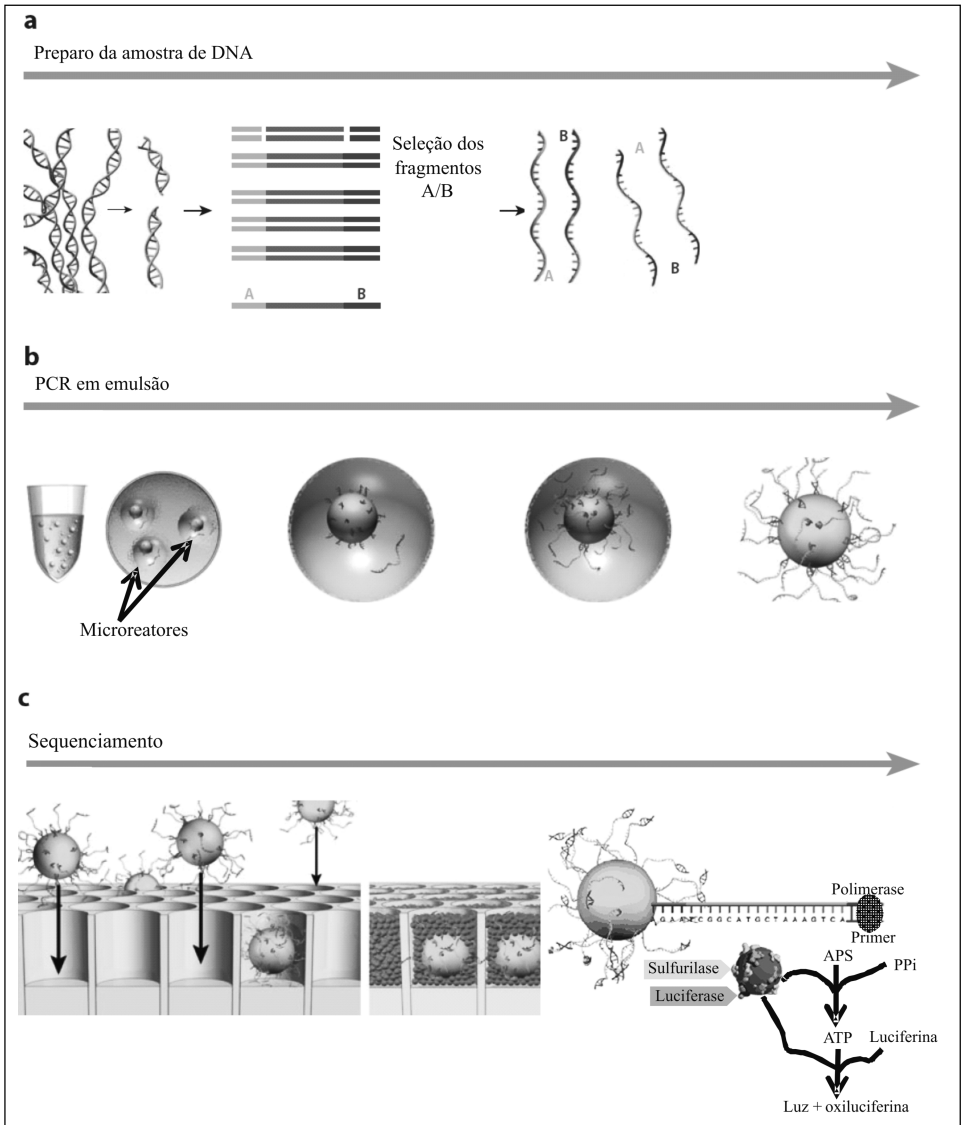

Figura 1 - Resumo ilustrativo do sequenciamento na plataforma 454. O sequenciamento é dividido em três etapas: (a) preparo da amostra, (b) PCR em emulsão e (c) Sequenciamento. (a) O DNA é fragmentado aleatoriamente e ligado a adaptadores A e B em suas extremidades. Os fragmentos A/B são selecionados para o sequenciamento. (b) Os fragmentos são ligados às microesferas magnéticas por meio do pareamento com sequências curtas complementares presentes na superfície da microesfera. Apenas um único tipo de fragmento se liga a uma determinada microesfera. As microesferas são capturadas individualmente em gotículas oleosas onde a PCR em emulsão ocorre. Milhares de cópias do fragmento alvo são produzidas nessa fase. (c) As microesferas ligadas às sequências alvo fita simples são capturadas individualmente em poços no suporte de sequenciamento. São fornecidos os reagentes para a reação de pirosequenciamento, e o sinal de luz emitido é identificado a cada base incorporada, em cada poço de sequenciamento.

O sistema requer que o DNA seja mecanicamente fragmentado em sequências de 300 – 800pb, transformado em fragmentos abruptos fosforilados e ligado a adaptadores de sequência específica (Figura 1). A biblioteca de DNA da amostra é ligada a adaptadores A e B nas extremidades 3' e 5' dos fragmentos, respectivamente, os quais são utilizados nas etapas posteriores de isolamento dos fragmentos (A-B) e amplificação e nas reações de sequenciamento. O adaptador B possui biotina ligada à extremidade 5', o que permite o isolamento dos fragmentos ligados ao adaptador A na extremidade 3' e adaptador B na extremidade 5' na amostra. Somente os fragmentos A-B são eluídos na reação de purificação e são especificamente ligados às microesferas que carregam várias cópias da sequência complementar exata ao adaptador B de um único fragmento (MARGULIES et al., 2005). O outro adaptador é utilizado no anelamento do *primer* que inicia a reação de sequenciamento. As microesferas ligadas aos fragmentos únicos de fita simples são então emulsionadas em uma mistura de água e óleo com reagentes de PCR para amplificação clonal do fragmento fita simples em cerca de 1 milhão de cópias. Na PCR em emulsão, o óleo em solução aquosa forma micelas, nas quais as microesferas são capturadas. Cada micela funcionará como um microrreator, produzindo muitas cópias idênticas de um mesmo fragmento isoladamente em um microssuporte (DRESSMAN ET AL., 2003).

Após a PCR de emulsão, as microesferas ligadas aos fragmentos de fita simples são depositadas em poços distintos em uma placa de sílica onde os reagentes para o sequenciamento são distribuídos. As reações de sequenciamento ocorrem em cada poço, para um único tipo de fragmento ligado à microesfera, não havendo, portanto, competição por reagentes com outros fragmentos da biblioteca. A placa de sequenciamento é dividida em 1,6 milhões de poços com diâmetro suficiente para alojar uma única microesfera (Figura 1).

A placa de sequenciamento é inserida junto ao sistema óptico de leitura no equipamento. Os reagentes e as soluções de sequenciamento são então distribuídos por toda a placa a cada ciclo para obtenção do sequenciamento paralelo dos 1,6 milhões de poços. O sequenciamento é realizado em ciclos, e a cada ciclo um tipo determinado de nucleotídeo é adicionado à reação. Se o nucleotídeo adicionado for incorporado à sequência em síntese, um sinal de luz é emitido, sendo a intensidade desse sinal um reflexo do número de nucleotídeos desse tipo específico que foram sucessivamente incorporados na molécula. Como o nucleotídeo que é adicionado a cada ciclo é conhecido, o sinal de luz emitido pode ser diretamente utilizado como informação de sequência (RONAGHI, 2001).

Os fragmentos sequenciados nessa plataforma passam por sistemas de análise de qualidade em que sequências distintas oriundas do sequenciamento de uma única microesfera são eliminadas, bem como as leituras em que a sequência inicial TCGA (quatro primeiros nucleotídeos dos adaptadores) não aparece. As leituras produzidas possuem geralmente cerca de 250pb, o que representa um comprimento de leitura muito menor que o produzido pelo sistema de Sanger (~700pb). A Roche divulgou recentemente o lançamento da série *Titanium* de pirosequenciamento, em que leituras maiores que 400pb são conseguidas. Esse aprimoramento das leituras advém de otimizações nas reações químicas do pirosequenciamento, as quais reduzem o ruído de fundo e aumentam o número de leituras por corrida, e do novo desenho do suporte de sequenciamento (*PicoTiterPlate*), o qual agregou duas mudanças principais: o uso de uma estrutura metálica, permitindo leituras mais acuradas, e esferas ainda menores, aumentando, tanto o tamanho das leituras, quanto o número de leituras por corrida (ROCHE, 2008). O maior tamanho das leituras e a grande capacidade de gerar informação tornam o processo de montagem mais fácil num projeto de sequenciamento *de novo* (sequenciamento de genomas desconhecidos) e permite trabalhar com coberturas genômicas mais amplas, favorecendo o processo de montagem. Além disso, pelo fato de não envolver clonagem bacteriana, a representação do genoma é bem mais fiel, de forma que sequências difíceis de clonar e manter em bibliotecas genômicas podem ser acessadas.

Genomas pequenos, como os de bactérias e de alguns eucariotos, podem ser facilmente montados usando a plataforma 454. Com relação às demais tecnologias de sequenciamento da segunda geração, a plataforma 454 é a que produz as maiores leituras e por isso tem sido mais utilizada, inclusive para o sequenciamento de genomas eucariotos (WICKER et al., 2009). Outra limitação importante da plataforma 454 é a baixa eficiência na determinação de homopolímeros. Como a intensidade do sinal de fluorescência relaciona-se ao número de vezes que um determinado nucleotídeo foi incorporado à sequência, a determinação precisa de sequências em que um único nucleotídeo é repetido mais de três vezes torna-se imprecisa. O custo do sequenciamento com essa plataforma é superior ao custo das plataformas Solexa e SOLiD (Tabela 1), mas, nos casos em que a produção de leituras maiores é necessária, a plataforma 454 deve ser a melhor opção.

Tabela 1 - Resumo das principais características técnicas das plataformas 454 GS-FLX, Solexa e SOLiD e laboratórios no Brasil que já adquiriram essas novas plataformas. A duração da corrida inclui o tempo para o preparo, a leitura e o processamento das amostras; o custo da corrida e o valor do equipamento são fornecidos na capacidade máxima do equipamento.

Plataforma	----------Corrida----------			----------Custo----------		Acurácia (%)	Laboratório**
	Informação (Gb)	Duração (dias)	Reads (pb)	Equipamento (U$)	Base (U$)		
GS-FLX *Titanium*	0,5	3 a 4	Até 400	531.500	10.000	99,5	- LNCC - IQ-USP
Genome analyzer (Solexa)	3	5	25-35	430.000	6.250	98,5	- Nenhum - Fiocruz
SOLiD *System*	25	4-12	35-50	599.000	10.000	99	- Instituto Ludwig - UFPA

*Valores cotados em janeiro de 2009.
**Pesquisa realizada em janeiro de 2009.

Plataforma Solexa

O sequenciamento na plataforma Solexa, assim como o sequenciamento de Sanger, é realizado por síntese usando DNA polimerase e nucleotídeos terminadores marcados com diferentes fluoróforos. A inovação dessa plataforma consiste na clonagem *in vitro* dos fragmentos em uma plataforma sólida de vidro, processo também conhecido como PCR de fase sólida (FEDURCO et al., 2006; TURCATTI et al., 2008). A superfície de clonagem (*flow cells*) é dividida em oito linhas que podem ser utilizadas para o sequenciamento de até oito bibliotecas. Em cada linha, adaptadores são fixados à superfície pela extremidade 5', deixando a extremidade 3' livre para servir na iniciação da reação de sequenciamento dos fragmentos imobilizados no suporte por hibridização (Figura 2).

Os fragmentos de DNA da amostra são também ligados aos adaptadores em ambas as extremidades, o que permite sua fixação ao suporte de sequenciamento por hibridização a um dos adaptadores fixados (Figura 2). No primeiro ciclo de amplificação, nucleotídeos não marcados são fornecidos para que haja a síntese da segunda fita do fragmento imobilizado no suporte. A alta densidade de adaptadores no suporte facilita a hibridização do adaptador livre dos fragmentos imobilizados a sua sequência complementar fixa perto do clone inicial durante o ciclo de anelamento. Após o ciclo de anelamento, o fragmento forma uma estrutura em "ponte" na superfície de sequenciamento e a extensão ocorre, formando a fita complementar também em "ponte". No ciclo de desnaturação, as fitas são separadas e linearizadas. Esses ciclos são repetidos 35 vezes e assim as cerca de mil cópias geradas de cada fragmento nessa PCR de fase sólida permanecem próximas umas das outras, formando um *cluster* de sequenciamento. Etapas de desnaturação são necessárias para a separação dos duplex formados e, nos próximos ciclos de amplificação, nucleotídeos terminadores marcados são fornecidos para as reações de sequenciamento que ocorrem dentro de cada cluster. A alta densidade dos clusters de sequenciamento possibilita que o sinal de fluorescência gerado com a incorporação de cada um dos nucleotídeos terminadores tenha uma intensidade suficiente para garantir sua detecção exata. Até 50 milhões de *clusters* podem ser produzidos por linha, correspondendo a uma representação satisfatória da biblioteca. Após a incorporação de cada nucleotídeo no fragmento em síntese, a leitura do sinal de fluorescência é realizada. Em seguida, ocorre uma etapa de lavagem para remoção dos reagentes excedentes e remoção do terminal 3' bloqueado e do fluoróforo do nucleotídeo incorporado no ciclo anterior para que a reação de sequenciamento prossiga. A leitura das bases é feita pela análise sequencial das imagens capturadas em cada ciclo de sequenciamento. Em geral, leituras de 25-35 bases são obtidas de cada *cluster* (SHENDURE & JI, 2008).

Plataforma SOLiD (*Sequencing by Oligonuclotide Ligation and Detection*)

No sistema SOLiD (MCKERNAN et al. 2006), diferentemente dos demais processos, a reação de sequenciamento é catalisada por uma DNA ligase, e não uma polimerase. O DNA alvo é mecanicamente fragmentado em um sonicador em fragmentos de 60-90pb, para as bibliotecas de *tags* únicas, ou 1-10Kb, para as bibliotecas de *tags* duplas (*mate-pair*). Os fragmentos de 60-90pb são diretamente ligados a adaptadores universais (P1 e P2) em ambas as extremidades. Já nas bibliotecas *mate-pair*, a

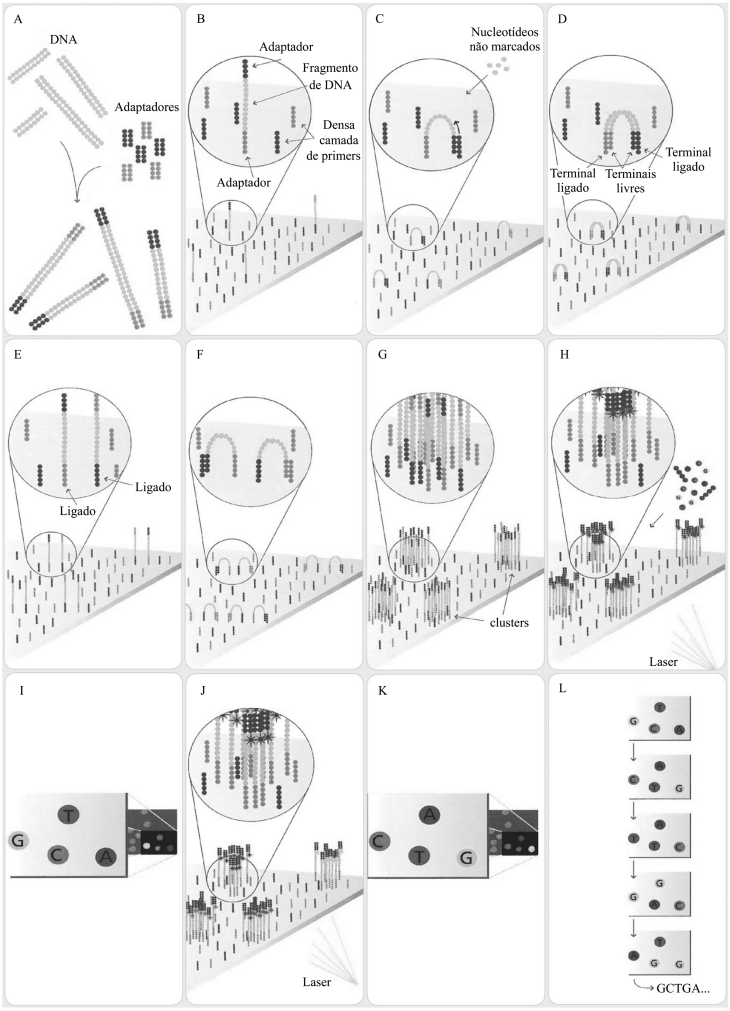

Figura 2 - Representação esquemática do princípio tecnológico da plataforma Illumina. O DNA é fragmentado aleatoriamente e ligado a adaptadores A e B em ambas as extremidades (A). As moléculas de DNA fita simples são aderidas por afinidade ao suporte sólido onde estão também aderidos em alta densidade oligonucleotídeos complementares aos adaptadores A e B (B). Durante a etapa de anelamento (C), no primeiro ciclo de amplificação da PCR em fase sólida, o adaptador da extremidade livre da molécula aderida ao suporte encontra seu oligonucleotídeo complementar no suporte, formando uma estrutura em ponte. Uma vez fornecidos os reagentes necessários, a PCR é iniciada utilizando a extremidade 3' livre do oligonucleotídeo como *primer* (C e D). Na etapa de desnaturação (E), a "ponte" é desfeita mediante elevação de temperatura. Repete-se a etapa de anelamento (F), formando novas estruturas em ponte e iniciando um novo ciclo de amplificação. Após uma série desses ciclos, serão obtidos *clusters* de moléculas idênticas ligadas ao suporte (G). Com a incorporação de nucleotídeos terminadores marcados e excitação a laser (H), é gerado sinal, o qual é captado por dispositivo de leitura e interpretado como um dos quatro possíveis nucleotídeos componentes da cadeia (I). O processo de incorporação de nucleotídeo marcado, excitação e leitura é repetido para cada nucleotídeo componente da sequência (J, K). A leitura é feita de forma sequencial, o que permite a montagem da sequência completa de cada *cluster* (L).

fragmentação resulta na produção de um contínuo de fragmentos de 1 a 10Kb, que são visualizados em gel para seleção da faixa de tamanho de interesse. Uma vez selecionados, os fragmentos são ligados aos mesmos adaptadores P1 e P2, mas são circularizados e clivados com uma enzima de restrição que reconhece seu sítio no adaptador e cliva adiante, liberando fragmentos formados por: 27 bases de uma região, mais a sequência dos adaptadores e mais 27 bases adicionais de outra região que está separada da primeira pela distância utilizada no intervalo de seleção dos fragmentos.

O adaptador P1 é utilizado no anelamento do *primer* da PCR de emulsão. A amplificação da biblioteca na plataforma SOLiD permite, da mesma forma que na plataforma 454, a ligação dos fragmentos por hibridização com sequências complementares aos adaptadores fixos a microesferas metálicas que são capturadas nas micelas da PCR de emulsão (Figura 3). As bibliotecas resultantes contêm milhões de moléculas únicas representando a sequência alvo inteira. As esferas são ligadas covalentemente a uma lâmina de vidro com uma substância desenvolvida pela *Applied Biosystems* que leva a ligação covalente das microesferas. Em cada corrida, são utilizadas duas lâminas ou *chips*, cada um com capacidade atual para 100 mil microesferas. Um aspecto interessante desse equipamento é que, embora as microesferas sejam aleatoriamente distribuídas sobre o *chip*, cada chip pode ser dividido em oito áreas, as quais podem então ser utilizadas na análise de oito bibliotecas diferentes. Alternativamente, é possível adquirir o sistema de código de barras da empresa que possibilita a identificação das diferentes

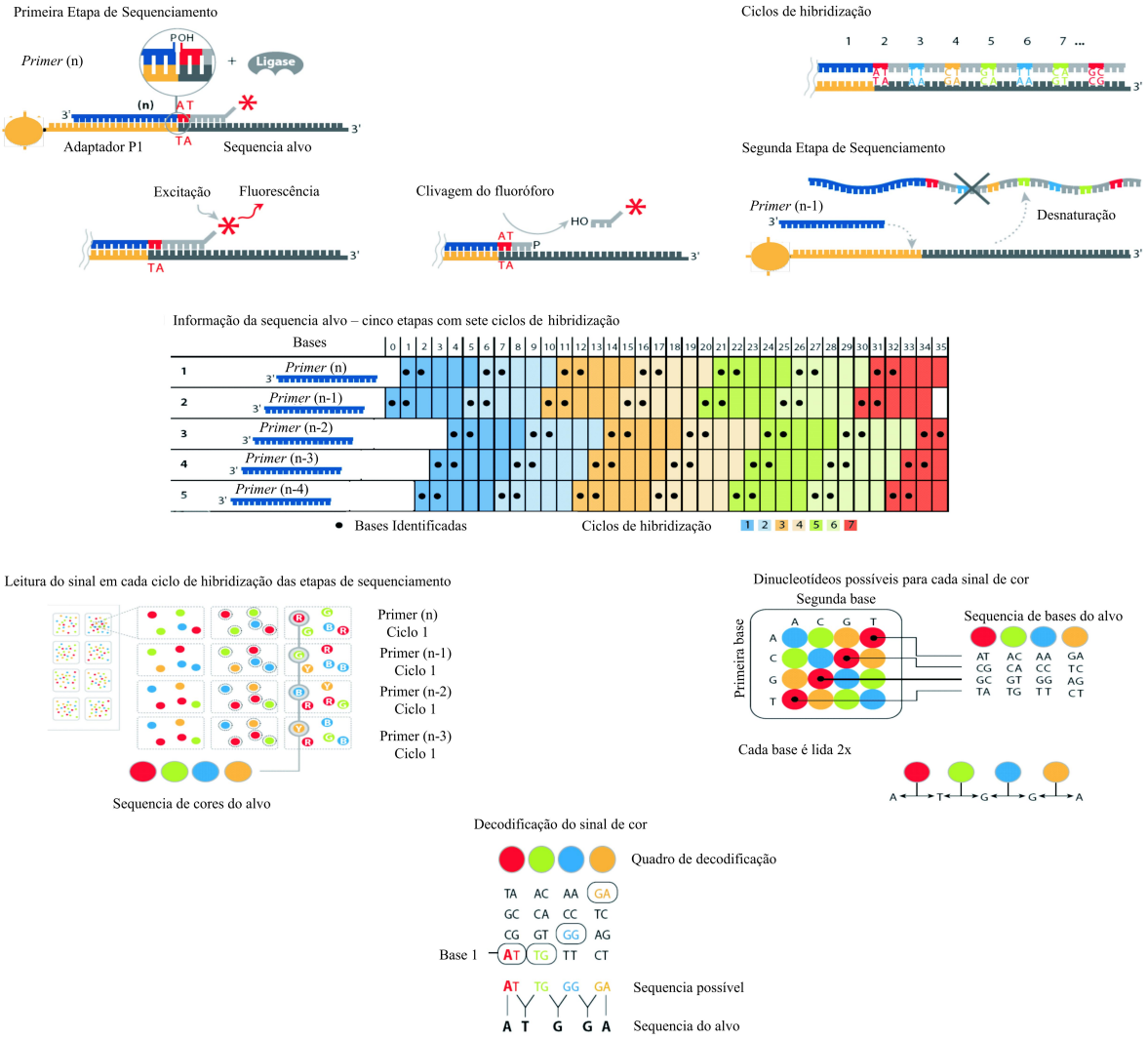

Figura 3 - Plataforma SOLiD. Os fragmentos de DNA são gerados e ligados ao adaptador P1 que se liga especificamente a uma microesfera. O sequenciamento ocorre por hibridização de sondas fluorescentes com o alvo em cinco etapas distintas. Na primeira etapa, o *primer* (n) é utilizado, liberando as primeiras bases da sequência alvo para hibridização com a sonda. Uma das sondas do *pool* encontrará similaridade ao alvo ligando-se a ele. O sinal de fluorescência é lido, e as três últimas bases da sonda, incluindo o fluoróforo, são removidas. Inicia-se o segundo ciclo de hibridização e assim sucessivamente, até que o alvo seja todo coberto (35pb). A sequência fita dupla é desnaturada, e uma nova etapa de sequenciamento é iniciada com o *primer* (n-1). Os ciclos de hibridização são repetidos, fornecendo informação de outras bases da sequência alvo. Novas etapas de sequenciamento com os *primers* (n-2), (n-3), e (n-4) são realizadas para que toda a sequência alvo seja determinada. Todas as combinações possíveis de dinucleotídeos são marcadas nas sondas com apenas quatro fluoróforos. Assim, duas leituras são necessárias de cada base para que a sequência do dinucleotídeo da sonda seja resolvida. Esse processo inicia-se com a identificação da primeira base do alvo na segunda etapa de sequenciamento (*primer* n-1), que libera para hibridização com a sonda uma base já conhecida, a última base do adaptador.

amostras distribuídas sobre um mesmo *chip* não dividido com base em cinco nucleotídeos específicos (código de barras) que são adicionados ao adaptador P2. Atualmente, são disponíveis 20 códigos distintos, os quais, se usados em *chips* divididos, possibilitam a análise simultânea de 320 amostras.

No analisador SOLiD, os moldes ligados às esferas são combinados aos *primers* universais de sequenciamento, a enzimas ligase e a sondas (1024 sondas). O sequenciamento é dividido em etapas distintas pelo uso do *primer* universal que tem n bases na primeira etapa, n-1 bases na segunda etapa, e assim

sucessivamente até a quinta etapa em que o *primer* possui n-4 bases. São também utilizadas, nas reações de sequenciamento, sondas curtas (oito bases) randômicas marcadas com um entre quatro fluoróforos possíveis em função do tipo de dinucleotídeo que apresentam na sua extremidade 3' (Figura 3).

As únicas bases seletivas da sonda são a primeira e a segunda; a terceira, a quarta e a quinta base são degeneradas em todas as combinações possíveis. As bases 6, 7 e 8 são inosinas que carregam o fluoróforo marcador. Na primeira etapa, é adicionado o *primer* universal completo que se anela exatamente na extremidade do adaptador P1. A sonda que for complementar à sequência alvo dentro do *pool* de sondas se hibridizará com a sequência molde e será ligada ao *primer* universal, pela ação da ligase. A fluorescência da sonda ligada é detectada, e o fluoróforo é clivado, deixando um grupo 5' fosfato disponível para reações adicionais. No próximo ciclo, adicionam-se novamente as sondas e a ligase para a leitura das próximas bases seletivas. Esses ciclos se repetem até que toda a sequência seja coberta. Na etapa seguinte, o fragmento é desnaturado e adiciona-se o segundo *primer* universal com n-1 bases, liberando desde a última base do adaptador para o sequenciamento. Novamente todos os ciclos com as sondas são realizados e esse processo é repetido produzindo uma leitura de 35 pb nas bibliotecas de *tags* únicas ou de 50pb nas bibliotecas *mate-pair*. As cinco etapas de sequenciamento são necessárias porque a cada ciclo de hibridização da sonda apenas a sequência do *primer* universal mais os grupos de dinucleotídeos marcadores das sondas que hibridizaram são conhecidos. Para descobrir o restante da sequência alvo, são necessárias, portanto, outras quatro etapas de sequenciamento usando o *primer* universal para o adaptador com uma base a menos no seu terminal 5' a cada etapa (n-1, n-2, n-3 e n-4). Assim, quando o *primer* n-1 for usado, por exemplo, a primeira sonda a se hibridizar fornecerá informação sobre a última base da sequência do adaptador e uma segunda informação da primeira base da amostra e assim por diante. Esse complexo processo ocorre sucessivamente, proporcionando dupla leitura para cada base e, como consequência, reduzindo muito a chance de erros de sequenciamento (Figura 3).

Como cada sinal de fluorescência especifica um dinucleotídeo e não uma única base, a decodificação dos sinais de leitura é feita combinando-se os dados (Figura 3). As bases do adaptador P1 são conhecidas, o que permite a identificação correta da primeira base do fragmento durante a segunda etapa de sequenciamento, quando se utiliza o *primer* com n-1 bases. Os demais sinais de fluorescência são especificados pela única combinação possível de cores que inclui a base conhecida. Esse sistema de leitura é muito eficiente na detecção de polimorfismos (SNPs), os quais são facilmente confundidos com erros de sequenciamento em outras plataformas. Na plataforma SOLiD, a presença de um SNP resulta sempre em uma das três alterações previstas de dois sinais de leitura, enquanto as demais seis alterações possíveis representam erros de sequenciamento com alteração de um único sinal. As leituras produzidas com o SOLiD apresentam acurácia muito superior às demais técnicas, sendo perfeitamente adequadas à identificação de polimorfismos genômicos reais.

As leituras curtas produzidas pela plataforma SOLiD foram utilizadas no sequenciamento de novo somente no caso de bactérias e com a produção de bibliotecas *mate-pair* (DURFEE et al., 2008). No entanto, a alta eficiência e sensibilidade da plataforma, aliadas à possibilidade de analisar 320 amostras distintas em uma única corrida, tornaram a plataforma SOLiD destinada principalmente aos estudos de transcritômica (CLOONAN et al., 2008; PASSALACQUA et al., 2009; TANG et al., 2009).

Genomas de plantas e as novas plataformas de sequenciamento

Para espécies vegetais com genomas desconhecidos, a utilização das novas plataformas de sequenciamento é ainda limitada. O tamanho das leituras produzidas é incompatível com a montagem dos genomas nucleares gigantescos e altamente repetitivos das plantas. Os poucos trabalhos realizados têm sido destinados ao sequenciamento de transcritos, ressequenciamento e sequenciamento *de novo* de genomas plastidiais, os quais são menores (~150Kb) e contêm pouca quantidade de DNA repetitivo.

A plataforma 454 é a mais utilizada em plantas. O primeiro trabalho que utilizou essa plataforma teve como objetivo avaliar o seu potencial para a análise de genomas repetitivos, comparando resultados do sequenciamento convencional com os obtidos para quatro clones BACs de cevada (WICKER et al., 2006). Os resultados mostraram que a plataforma 454 é capaz de gerar a mesma quantidade de informação obtida com o sequenciamento de Sanger com alta qualidade. No entanto, as leituras curtas produziram de seis a nove vezes mais "gaps", principalmente devido a erros de sequenciamento observados em regiões de homopolímeros. Para as regiões repetitivas, as leituras curtas só apresentaram problemas de montagem quando presentes em múltiplas cópias em um único clone. Parte do genoma da cevada (~10% do genoma

haploide) foi também sequenciada em uma única corrida utilizando leituras ainda menores produzidas com a plataforma Solexa (WICKER et al., 2008). Muito importante nesse trabalho foi o desenvolvimento de um índice matemático para predição e exclusão de regiões repetitivas. A aplicação desse índice possibilitou a montagem de 5.500Mb do genoma da cevada e a identificação de regiões desconhecidas.

O problema de *gaps* em regiões de homopolímeros foi relatado também no sequenciamento dos genomas plastidiais de *Nandina* e *Platanus* com as plataformas 454 e Solexa (MOORE et al., 2006). Apesar disso, uma redução considerável de custo (~$4.500 por genoma) e tempo (~2 semanas para finalização dos dois genomas) foi conseguida utilizando o sistema GS20 de pirosequenciamento para sequenciar 99,75% de tais genomas.

Um dos trabalhos que melhor ilustra o potencial das novas plataformas de sequenciamento foi o conduzido com o eucalipto (NOVAES et al., 2008), espécie para a qual pouca informação genômica é disponível. Nesse trabalho, 148,4Mb de ESTs de vários genótipos e tecidos foram sequenciados com a plataforma 454, gerando um número maior de genes do que o gerado por sequenciamento convencional, 23.742 SNPs altamente confiáveis e, muito importante, o enriquecimento de 37 vezes nas sequências de ESTs disponíveis para o eucalipto nos bancos de dados públicos. O sequenciamento de ESTs representa certamente uma estratégia de sucesso para obtenção de informação genômica das plantas a partir das novas plataformas de sequenciamento. Ele reduz os problemas de montagem associados às leituras curtas (EMRICH et al., 2007) e pode ser ainda mais informativo, uma vez que mais informação é produzida e sequências de baixa expressão são também amostradas (não há efeito de clonagem bacteriana). Além disso, o custo do sequenciamento de ESTs é menor com as novas plataformas de sequenciamento, uma vez que não depende da construção de bibliotecas de cDNA para cada um dos tecidos amostrados. Além do banco de ESTs do eucalipto, os bancos de milho (EMRICH et al., 2007), ***Medicago*** *sp* (CHEUNG et al., 2006) e arabidopsis (WEBER et al., 2007) também estão sendo enriquecidos com o pirosequenciamento.

O potencial de identificação de novos genes com os novos sistemas de sequenciamento é especialmente importante quando se deseja conhecer genes funcionais em tipos celulares restritos. Nesses casos, utilizando os métodos convencionais de sequenciamento, um grande número de bibliotecas deve ser construído e muitos ESTs devem ser sequenciados para maximizar a chance de encontrar os genes de interesse. Já com o sequenciamento livre de clonagem bacteriana, genes de células específicas podem ser facilmente identificados, como realizado para as células meristemáticas apicais do milho, que compõem apenas uma porção do ápice da planta. Um total de 400 novos genes foi identificado em uma única corrida na plataforma 454, sendo a maior parte deles genes especificamente expressos nesse tecido (EMRICH et al., 2007).

O ressequenciamento genômico em plantas é também muito informativo para estudos de polimorfismo. Indivíduos variantes de arabidopsis foram sequenciados utilizando a plataforma Solexa para buscar variações genotípicas (OSSOWSKI et al., 2008). A montagem das leituras curtas da plataforma Solexa foi auxiliada pela informação genômica disponível para arabidopsis, e o sequenciamento de pequena cobertura (11 vezes em uma única corrida) utilizado nesse trabalho foi suficiente para detecção de deleções, duplicações e de SNPs com uma especificidade de 99%. A plataforma 454 foi também utilizada com sucesso na identificação de SNPs em transcritos das células meristemáticas apicais de milho. Em uma única corrida, foi possível identificar cerca de cinco mil SNPs válidos entre os 2.472 genes identificados (BARBAZUK et al., 2007).

As novas plataformas de sequenciamento, em um futuro bem próximo, revolucionarão o conhecimento sobre o genoma das plantas, principalmente no que concerne ao estudo de variantes alélicas, SNPs, ao desenvolvimento de marcadores para seleção assistida e à clonagem baseada em mapeamento, representando uma importante ferramenta no melhoramento vegetal. Essa revolução será possível com o avanço tecnológico das próprias plataformas de sequenciamento, produzindo leituras maiores, avanço das ferramentas de análise de leituras curtas e montagem de sequências.

O que parece estar cada vez mais evidente é, na verdade, o potencial de uso imediato dessas tecnologias na genômica funcional com o estudo de transcritomas que vão desde organismos completos até células individuais (ANDREAS et al., 2007; TANG et al. 2009). Todas as plataformas de sequenciamento da segunda geração podem ser utilizadas no sequenciamento de transcritomas ou RNA-seq. A maior parte dos estudos de transcritômica em plantas é realizada utilizando os microarranjos de DNA, os quais, além de depender de um conhecimento genômico prévio e de serem influenciados pelo elevado ruído de fundo, possuem ainda uma faixa de detecção de expressão limitada quando comparada às novas plataformas de sequenciamento (~100 vezes *versus* 9.000 vezes)

(MARIONI et al., 2008; WANG et al., 2009). O grande sucesso das novas tecnologias na transcritômica se deve também ao fato de que estas possibilitam a superação de uma das maiores limitações dos projetos ESTs – a brusca redução no número de sequências novas amostradas com o aumento na quantidade de informação sequenciada. No estudo do transcritoma de plântulas de arabidopsis, ANDREAS et al. (2007) identificaram 16.000 novos ESTs ainda não caracterizados no dbESTs dos quais pelo menos 60 representam genes ainda não anotados, conferindo maior confiabilidade aos dados principalmente com relação à quantificação dos níveis de expressão gênica, os quais dependem muito do efeito de amostragem.

Um exemplo interessante do efeito de amostragem é apresentado no estudo do transcritoma de ***S. cerevisae*** com a plataforma Solexa (NAGALAKSHMI et al., 2008). Um total de 66 íntrons previamente identificados foram encontrados entre as sequências expressas na levedura, alguns dos quais foram tão expressos quanto seus éxons adjacentes.

CONCLUSÕES

As novas plataformas de sequenciamento apresentam a grande vantagem de permitir um sequenciamento altamente representativo de genomas e/ou transcritomas em um único passo, o que é extremamente relevante, em razão da grande redução de custo alcançada com essas metodologias. Seu emprego tem revolucionado a transcritômica com a geração de dados altamente reprodutíveis e informativos e com precisão na quantificação de transcritos. Em função do problema da montagem das leituras curtas produzidas por essas tecnologias, seu uso na genômica de plantas tem sido direcionado para o sequenciamento dos genomas plastidiais, sequências expressas, clones de interesse, ressequenciamento e detecção de variantes genotípicas. Uma combinação de alguma dessas tecnologias à tecnologia de Sanger poderia associar o baixo custo e a alta representatividade da primeira à facilidade de montagem do genoma da segunda, facilitando seu emprego no sequenciamento genômico de plantas.

REFERÊNCIAS

ANDREAS, P.M. et al. Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. **Plant Physiology**, v.144, p.32-42, 2007. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1913805/?tool=pubmed>. Acesso em: 5 jun. 2009. doi: 10.1104/pp.107.096677.

BARBAZUK, W.B. et al. SNP discovery via 454 transcriptome sequencing. **The plant journal**, v.51, p.910-918, 2007. Disponível em: <http://www.ncbi.nlm.nih.gov/ pmc/articles/PMC2169515/?tool=pubmed>. Acesso em: 5 jun. 2009. doi: 10.1111/j.1365-313X.2007.03193.x.

CHEUNG, F. et al. Sequencing ***Medicago trunculata*** expressed sequenced tags using 454 Life Sciences technology. **BMC genomics**, v.7, p.272, 2006. Disponível em: <http://www.biomedcentral.com/1471-2164/7/272>. Acesso em: 5 jun. 2009. doi:10.1186/1471-2164-7-272.

CLOONAN, N. et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. **Nature methods**, v.5, n.7, p.613-619, 2008. Disponível em: <http://www.nature.com/nmeth/journal/v5/n7/abs/nmeth.1223.html>. Acesso em: 5 jun. 2009. doi:10.1038/nmeth.1223.

DURFEE, T. et al. The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. **Journal of bacteriology**, v.190, n.7, p. 2597-2606, 2008. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2293198/?tool= pubmed>. Acesso em: 5 jun. 2009. doi: 10.1128/JB.01695-07.

EMRICH, S.J. et al. Gene discovery and annotation using LCM-454 transcriptome sequencing. **Genome research**, v.17, n.1, p.69-73, 2007. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1716268/?tool=pubmed>. Acesso em: 5 jun. 2009. doi: 10.1101/gr.5145806.

FEDURCO, M. et al. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. **Nucleic acids research**, v.34, n.3, p.e22, 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1363783/?tool= pubmed>. Acesso em: 5 jun. 2009. doi: 10.1093/nar/gnj023.

PASSALACQUA, K.D. et al. Structure and complexity of a bacterial Transcriptome. **Journal of bacteriology**, v.191, n.10, p.3203-3211, 2009. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2687165/?tool=pubmed>. Acesso em: 5 jun. 2009. doi:10.1128/JB.00122-09.

MARIONI, J.C. et al. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. **Genome research**, v. 18, n.9, p.1509-1517, 2008. Disponível am: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2527709/?tool=pubmed>. Acesso em: 5 jun. 2009. doi: 10.1101/gr.079558.108.

MARGULIES, M. et al. Genome sequencing in open microfabricated high density picoliter reactors. **Nature**, v.437, n. 7057, p. 376-380, 2005. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1464427/>. Acesso em: 5 jun. 2009. doi: 10.1038/nature03959.

MCKERNAN, K. et al. Reagents, methods, and libraries for bead-based sequencing. **US patent application 20080003571**, 2006.

MOORE, M.J. et al. Rapid and accurate pyrosequencing of angiosperm plastid genomes. **BMC Plant Biology**, v.6, n.1, p.17, 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1564139/?tool=pubmed>. Acesso em: 5 jun. 2009. doi: 10.1186/1471-2229-6-17.

NAGALAKSHMI, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. **Science**, v. 320, p.1344-1349, 2008. Disponível em: <http:// www.sciencemag.org/cgi/content/full/320/5881/1344>. Acesso em: 5 jun. 2009. doi: 10.1126/science.1158441.

NOVAES, E. et al. High-throughput gene and SNP discovery in ***Eucalyptus grandis***, an uncharacterized genome. **BMC genomics, v.**9, p.312, 2008. Disponível em: <http:// www.ncbi.nlm.nih.gov/pmc/articles/PMC2483731/ ?tool=pubmed>. Acesso em: 5 jun. 2009. doi: 10.1186/1471-2164-9-312.

OSSOWSKI, S. et al. Sequencing of natural strains of ***Arabidopsis thaliana*** with short reads. **Genome research**, v.18, p.2024–2033, 2008. Disponível em: <http://www.ncbi.nlm.nih.gov/ pmc/articles/PMC2593571/?tool=pubmed>. Acesso em: 5 jun. 2009. doi: 10.1101/gr.080200.108.

ROCHE 454 SEQUENCING. System features for GS FLX Titatnium series. (November 24, 2008). Disponível em: <http:/ /www.454.com/products-solutions/system-features.asp>. Acesso em: 5 jun. 2009.

RONAGHI, M. et al. A sequencing method based on real-time pyrophosphate. **Science, v.**281, p.363–365, 1998. Disponível em: <http://www.sciencemag.org/cgi/content/full/ 281/5375/ 363>. Acesso em: 5 jun. 2009. doi: 10.1126/ science.281.5375.363.

RONAGHI, M. Pyrosequencing sheds light on DNA sequencing. **Genome research**, v.11, p.3-11, 2001. Disponível em: <http:/ /genome.cshlp.org/content/11/1/3.long>. Acesso em: 5 jun. 2009. doi: 10.1101/gr.150601.

SHENDURE, J.; JI, H. Next-generation DNA sequencing. **Nature biotechnology**, v.26, n.10, p.1135-1145, 2008. Disponível em: <http://www.nature.com/nbt/journal/v26/n10/ abs/ nbt1486.html>. Acesso em: 5 jun. 2009. doi:10.1038/ nbt1486.

TANG, F. et al. mRNA-Seq whole-transcriptome analysis of a single cell. **Nature methods**, v.6, p.377-382, 2009. Disponível em: <http://www.nature.com/nmeth/journal/v6/n5/ abs/ nmeth.1315.html>. Acesso em: 5 jun. 2009. doi:10.1038/ nmeth.1315.

TURCATTI, G. et al. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. **Nucleic acids research**, v.36, e25, 2008. Disponível em: <http://www.ncbi.nlm.nih.gov/ pmc/articles/ PMC2275100/?tool=pubmed>. Acesso em: 5 jun. 2009. doi: 10.1093/nar/gkn021.

WANG, Z. et al. RNA-seq: A revolutionary tool for transcriptomics. **Nature**, v. 10, p.57-63, 2009. Disponível em: <http://www.nature.com/nrg/journal/v10/n1/abs/ nrg2484.html>. Acesso em: 5 jun. 2009. doi:10.1038/nrg2484.

WEBER, A.P. et al. Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. **Plant physiology**, v.144, n.1, p.32-42, 2007. Disponível em: <http:// www.ncbi.nlm.nih.gov/pmc/articles/PMC1913805/ ?tool=pubmed>. Acesso em: 5 jun. 2009. doi: 10.1104/ pp.107.096677.

WICKER, T. et al. 454 sequencing put to the test using the complex genome of barley. **BMC genomics, v.**7, p.275, 2006. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/ PMC1633745/?tool=pubmed>. Acesso em: 5 jun. 2009. doi: 10.1186/1471-2164-7-275.

WICKER, T. et al. Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. **BMC genomics**, v.9, p.518, 2008. Disponível em: <http://www.ncbi.nlm.nih.gov/ pmc/articles/ PMC2584661/?tool=pubmed>. Acesso em: 5 jun. 2009. doi: 10.1186/1471-2164-9-518.

WICKER, T. et al. A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. **Plant journal**, v.59, n.5, p.712-22, 2009. Disponível em: <http:// www3.interscience.wiley.com/cgi-bin/fulltext/ 122381633/PDFSTART>. Acesso em: 5 jun. 2009. doi: 10.1111/j.1365-313X.2009.03911.x.