Decifrando o genoma em grande escala

Sylvia Morais de Sousa[1]
Andréa Almeida Carneiro[2]
Newton Portilho Carneiro[3]

Resumo - A determinação das funções gênicas tem demandado um grande avanço das ciências genômicas, cujas tecnologias concentram-se, principalmente, na geração e no estudo de uma grande quantidade de dados. O ponto de apoio para o entendimento da função gênica e da estrutura do genoma tem sido o sequenciamento de genomas completos e do genoma expresso em grande escala. Mapas físicos e genéticos têm sido integrados com informações genômicas e de expressão, resultando em bancos de dados públicos altamente informativos para diferentes espécies animais e vegetais. Tais informações auxiliam em vários aspectos a análise de expressão gênica, a determinação dos efeitos de processamento de éxons e do número de cópias gênicas e cromossômicas, culminando na determinação das funções biológicas e do mecanismo de ação de vários genes. São descritos o surgimento de novas tecnologias e a evolução de algumas inovações já existentes, voltadas para a identificação de funções gênicas.

Palavras-chave: Sequenciamento de DNA. Genômica funcional. Macroarranjo de DNA. Mutagênese. Fenotipagem.

INTRODUÇÃO

A melhoria e a redução no custo das tecnologias de genotipagem e fenotipagem, associadas a estudos de genoma em grande escala, rapidamente estão-se tornando a abordagem preferida para dissecar a genética de caracteres complexos. O sequenciamento de ácido desoxirribonucleico (DNA) teve dois grandes momentos tecnológicos na história que trouxeram um enorme avanço para a ciência. Um deles ocorreu em 1977 com o aparecimento do método de Sanger. Desde o surgimento dessa técnica, houve modificações que ajudaram a aperfeiçoar o processo. Uma delas foi o uso de dioxinucleotídeos marcados com fluorescência, capazes de ser visualizados por laser; a segunda foi o uso de capilares em substituição a placas, os quais auxiliaram na automatização do processo de carregamento de amostra no gel e número de amostras feitas por dia. O segundo grande momento do sequenciamento ocorreu próximo a 2005, com o surgimento do sequenciamento por síntese. Esse processo, apesar de bastante inovador (um aumento de cerca de cem vezes em comparação com o obtido pelo método Sanger), não anulou o primeiro, pelo fato de fazer leituras curtas (cerca de 200 pb)[4]. Dentre os métodos novos de sequenciamento encontram-se as plataformas 454 da Life Science, Illumina® e SOLiD™ da Applied Biosystems™[5]. Essas novas tecnologias deram oportunidades de estudos mais complexos de organismos poliploides, com grande quantidade de sequências repetitivas, genomas comparativos, identificação de polimorfismos e genes diferencialmente expressos (que atualmente são muito estudados usando microchips de DNA também conhecidos como microarranjos). Os microarranjos, estudo de expressão gênica em grande es-

[1]*Bióloga, D.Sc., Pesq. Embrapa Milho e Sorgo, Caixa Postal 151, CEP 35701-970 Sete Lagoas-MG. Correio eletrônico: smsousa@cnpms.embrapa.br*

[2]*Bióloga, D.Sc., Pesq. Embrapa Milho e Sorgo, Caixa Postal 151, CEP 35701-970 Sete Lagoas-MG. Correio eletrônico: andreac@cnpms.embrapa.br*

[3]*Biólogo, D.Sc., Pesq. Embrapa Milho e Sorgo, Caixa Postal 151, CEP 35701-970 Sete Lagoas-MG. Correio eletrônico: newtonc@cnpms.embrapa.br*

[4]*pb – Pares de bases.*

[5]*Disponíveis respectivamente nos sites: http://www.454.com; http://www.illumina.com e http://www3.appliedbiosystems.com*

cala, têm três grandes grupos no mercado: Affymetrix, NimbleGen e a Agilent[6].

A função de um dado gene não necessariamente representa o que está descrito no banco de dados. Muitas vezes é necessário verificar sua função bioquímica em um contexto biológico. Um dos processos mais bem-aceitos é a modificação da expressão do gene em sistemas heterólogos. Isso pode ser inicialmente feito por meio da identificação de mutantes obtidos por mutagênese química como é o caso da tecnologia *targenting induced local lesions in genomes* (*Tilling*) e *transposon* ou transgenia, utilizando construções gênicas que se baseiam em RNA de interferência – RNA *interference* (RNAi). Os processos de relacionar gene e função têm sido feitos em grande escala por companhias que têm otimizado desde a identificação de genes de interesse (por estudos genômicos), passando pela montagem de cassetes gênicos, transformação de plantas modelos (como tabaco, Arabidopsis, arroz e milho), até a análise fenotípica.

Tecnologias têm direcionado o sequenciamento de um genoma completo custar abaixo de mil dólares, contudo vários outros aspectos estão envolvidos no entendimento da relação gene e função. O conhecimento mais aprofundado dessa função gênica e a sua inter-relação (também conhecida como metabolômica), tanto em nível micro (célula) como macro (planta), têm como objetivo final o entendimento e o desenvolvimento de plantas cada vez mais produtivas e adaptadas às mais diversas condições de cultivo.

SEQUENCIAMENTO EM GRANDE ESCALA

Em 1977, foram publicados dois artigos metodológicos para a determinação rápida de sequências de DNA (SANGER et al., 1977; SANGER; NICKLEN; COULSON, 1977), que iriam transformar a biologia como um todo, fornecendo uma ferramenta poderosa para decifrar genes completos, que, mais tarde, seriam a base para o sequenciamento de genomas completos. O método sofreu uma série de melhorias, estabelecendo-se como o único de sequenciamento de DNA usado nos 30 anos seguintes. Com a meta de decifrar o genoma humano, houve um aumento, sem precedentes, na escala do sequenciamento de DNA, levando ao desenvolvimento de sequenciadores automáticos por capilaridade. A automação de laboratórios e a paralelização de processos resultaram em centros de sequenciamento com centenas de instrumentos. No entanto, mesmo com dois genomas humanos sequenciados e de outras tantas espécies, o desejo por uma tecnologia mais eficiente e barata continuou impulsionando as pesquisas na área (SCHUSTER, 2008).

Centenas de instrumentos, com base em sequenciamento capilar de 96 amostras, foram substituídos por poucos aparelhos capazes de fazer sequenciamento de milhões de pares de bases, paralelamente, em uma única corrida. Além disso, essa nova geração de sequenciadores utiliza fragmentos que não são sujeitos ao sistema convencional de clonagem em vetores de *Escherichia coli*.

Os primeiros sinais de que o mercado de sequenciamento poderia ser revolucionado apareceram em 2005, com a publicação da tecnologia sequenciamento por síntese, desenvolvida pela 454 Life Sciences (MARGULIES et al., 2005) e pelo protocolo multiplex de colônia de polimerase do laboratório de George Church da Escola de Medicina de Harvard, EUA (SHENDURE et al., 2005). Ambos os grupos usaram a estratégia de reduzir o volume necessário de reação, enquanto aumentavam dramaticamente o número de reações de sequenciamento por corrida. A estratégia consistia em colocar em uma matriz centenas de milhares de moldes em uma placa do tipo picotítulo (PTP) ou finas camadas de agarose, para que estas sequências pudessem ser analisadas em paralelo. Tais modificações culminaram em um aumento gigantesco de informações, quando comparadas com as 96 sequências obtidas pelo método Sanger em capilar (SCHUSTER, 2008).

O pirosequenciamento é um método para determinar a ordem dos nucleotídeos do DNA, com base na detecção da liberação de pirofosfato no ato da incorporação dos nucleotídeos, ao invés da terminação em cadeia com dideoxinucleotídeos, daí o nome sequenciamento por síntese. Nesse método, a detecção da atividade da DNA polimerase utiliza outra enzima quimioluminescente, a luciferase. O DNA molde é imobilizado e a solução contendo os nucleotídeos é adicionada e removida após cada reação, sendo que a luz é produzida apenas quando a solução de nucleotídeos complementa a primeira base sem par da fita molde (RONAGHI, 2001). A sequência dos sinais quimioluminescentes permite a determinação da sequência dos respectivos nucleotídeos complementares à fita molde (Fig 1).

O sistema de sequenciamento de DNA paralelo da 454 Life Sciences é cem vezes mais rápido do que o método de sequenciamento padrão e é capaz de sequenciar mais que 200 mil fragmentos, em 4 horas de corrida (MARGULIES et al., 2005). Contudo, esse aumento na velocidade de sequenciamento veio associado a uma redução no comprimento da leitura, sendo sequenciados em média fragmentos de, aproximadamente, 100 pb de comprimento (MARGULIES et al., 2005). Novas versões do sistema Genome Sequencer FLX e da química Titanium (HARKINS; JARVIE, 2007) aumentaram o comprimento médio da leitura para um pouco mais de 200 bases, com promessas de atingir leituras de até 1.000 pb. A principal vantagem de leituras mais longas é a facilidade na montagem e na organização das informações,

[6]*Disponíveis respectivamente nos sites: http://www.affymetrix.com; http://www.nimblegen.com e http://www.agilent.com*

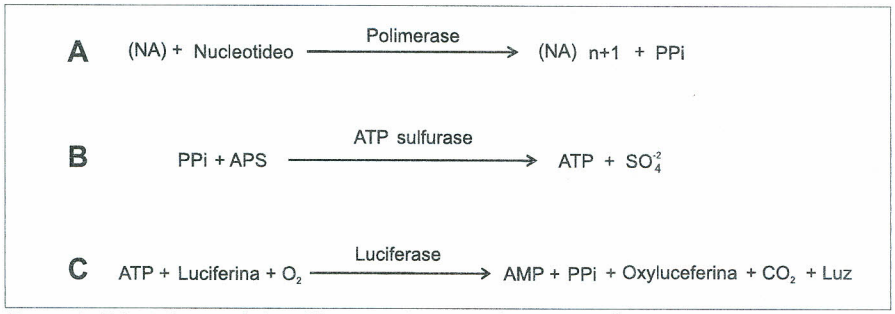

Figura 1- Princípio geral dos diferentes sistemas de reações de pirosequenciamento

FONTE: Ronaghi (2001).

NOTA: A - A polimerase catalisa a incorporação de nucleotídeos na cadeia de DNA; B - Como resultado da incorporação, a molecular pirofosfato (PPI) é liberada e subsequentemente convertida a adenosina-5'-trifosfato (ATP) pela ATP sulfurase; C - A luz é produzida em uma reação da luciferase durante a qual a molécula luciferina é liberada.

principalmente considerando a montagem *de novo* de genomas.

O método 454 reduz a dimensão e a complexidade de cada etapa do protocolo convencional de pirosequenciamento, aumentando a produtividade de informações em uma escala genômica. Ao invés de desenhar um novo par de *primers,* para a amplificação por reação em cadeia da polimerase – *polymerase chain reaction* (PCR) de cada fragmento genômico, o 454 divide o genoma em milhões de fragmentos e liga dois *primers* adaptadores universais (A e B). A ligação coesiva é não específica e os fragmentos que não incorporaram ambos os *primers* são removidos, usando esferas especiais que são revestidas com os *primers*. Em seguida, ocorre a reação de amplificação por PCR, também conhecida como emulsão de PCR (DRESSMAN et al., 2003), na qual o *primer* B e o fragmento molde estão livres na solução, enquanto o *primer* A é embebido nas esferas. As esferas de sefarose, cobertas com milhões de *primer* A, são misturadas com a solução de DNA molde, *primer* B e solução. As esferas são adicionadas em excesso, de tal forma que apenas uma molécula de DNA modelo e uma esfera acabam em cada gota d'água.

Esse passo é crítico, uma vez que cada fragmento de DNA tem as mesmas sequências de adaptadores e *primers* de PCR. Após a reação de PCR, cada esfera tem milhões de cópias da mesma sequência de DNA. Essas esferas são, então, colocadas em placas que cabem exatamente uma esfera por poço, contendo um total de 400 mil poços. O processo restante é o pirosequenciamento com enzima e substrato, sendo colocados na placa PTP, trinucleotídeos ATGC adicionados sequencialmente, cuja fluorescência é capturada com a câmera *charge coupled device* (CCD).

Um método para melhorar a eficiência de perfis de transcrição com base no 454 é ancorar esses fragmentos sequenciados a sítios únicos próximos à região 3' das sequências expressas para reduzir o número de leituras necessário para identificar mRNAs individuais e maximizar a distinção de polimorfismos entre transcritos relacionados. A região 3'- não traduzida (UTR) é rica em polimorfismos de base única, que distinguem os transcritos (BHATTRAMAKKI et al., 2002; VROH BI et al., 2006). A especificidade da leitura da sequência 3'-UTR permite a efetiva anotação de cada mRNAs, sem a montagem completa dos cDNAs. Essa estratégia, por exemplo, foi adotada para analisar ovários de milho mutante e selvagem, utilizando uma estratégia multiplex que resultou em perfis de expressão quantitativos, em que se podem distinguir membros próximos de famílias gênicas (EVELAND; MCCARTY; KOCH, 2008).

O comprimento dos fragmentos sequenciados com a tecnologia 454 não é uma preocupação para a análise de transcriptomas, uma vez que os genes expressos são menores do que os genomas e apresentam menos DNA repetitivo. A técnica de captura por microdissecção laser (LCM) pode ser utilizada para isolar transcritos que se acumulam em determinados tipos de células, reduzindo, assim, a complexidade e o tamanho do transcriptoma alvo (SCHNABLE; HOCHHOLDINGER; NAKAZONO, 2004). Outra grande vantagem da tecnologia 454 é não envolver a clonagem dos fragmentos e a construção de bibliotecas, que são etapas caras e demoradas, além de possibilitar o sequenciamento de diferentes amostras de cDNA simultaneamente, aumentando a recuperação de transcritos altamente especializados e raros.

Emrich et al. (2007) relataram o sequenciamento de cDNA extraído de células do meristema apical do broto, utilizando a abordagem LCM-454. Uma única corrida de sequenciamento 454 foi capaz de gerar mais que 25 mil sequências genômicas de milho, sendo anotados quase 400 transcritos de genes órfãos (FU et al., 2005). No entanto, esses transcritos órfãos, detectados por LCM-454, foram validados experimentalmente, e a maioria deles não foi detectada em outros tecidos, incluindo espigas imaturas ricas em tecido meristemático.

Outra plataforma de sequenciamento da Illumina®, denominada Solexa e mais recentemente Genome Analyzer, também utiliza o sequenciamento por síntese de fragmentos aleatórios, ligados coesivamente aos adaptadores e processados em multiplex com reações, utilizando moléculas únicas. No entanto, durante a síntese, essa plataforma incorpora trinucleotídeos fluorescentes com a extremidade 3'-OH –inativada, enquanto no 454 são incorporados quatro nucleotídeos por fileira. O Genome Analyzer incorpora os nucleotídeos, um de cada vez, tirando fotos da intensidade das quatro cores fluorescentes

a cada ciclo, seguido pela ativação química do 3'-OH. Todos os quatro nucleotídeos são adicionados simultaneamente e cada fluorescência é marcada com uma cor diferente. Em vez de uma placa PTP, o sequenciamento no Genome Analyzer ocorre em fluxo microfluido das células, que são cobertos por dois *primers* diferentes ligados quimicamente. A hibridação dos *primers* com o DNA genômico com pontas coesivas é preparada de maneira similar ao descrito para o 454. Quando os fragmentos se ligam ao interior da célula, as duas pontas do fragmento ligam-se, formando uma ponte. Após a formação dessa ponte, uma polimerase isotérmica amplifica a sequência, usando os mesmos *primers* genéricos e os nucleotídeos não-marcados. O resultado da desnaturação é um fluxo de células cobertas com uma única fita molde de DNA, pronta para a incorporação de nucleotídeos, imagem fluorescente e ativação dos 3'-OH.

A terceira plataforma, denominada Sequencing by Oligo Ligation and Detection (SOLiD™) da Applied Biosystems, é semelhante às demais pelo fato de ser uma plataforma multiplex e massiva, que não necessita da clonagem dos fragmentos e amplifica o DNA fragmentado, que é ligado a dois *primers*, em uma única emulsão de PCR com esferas cobertas de *primers*.

Esse instrumento tornou-se comercial em outubro de 2007, utilizando um processo único de sequenciamento catalisado pela DNA ligase. Cada corrida do SOLiD™ requer cinco dias e produz de 3 a 4 milhões de etiquetas com uma média de 25 a 35 pb. A plataforma SOLiD™ tem um único protocolo a leitura de cada nucleotídeo, duas vezes sucessivamente em reações com dinucleotídeos, como um mecanismo de verificação de erros. A reação de sequenciamento é repetida cinco vezes em cada modelo com quatro cores fluorescentes em sondas de oito pares de base. Cada uma das cinco corridas de sequenciamento por ligação é precedida por uma hibridação com *primers* de sequenciamento universais. Esse *primer* é específico para os *primers* ligados à esfera, usado durante a amplificação de PCR, mas os cinco *primers* hibridizam a diferentes fragmentos que se sobrepõem na extremidade 5' do *primer* ligado à esfera. A extremidade 5' do *primer* 1 é o último nucleotídeo a se ligar no *primer* da esfera (posição n), enquanto a extremidade 5' do *primer* 2 está na posição n-1 e assim por diante.

O primeiro passo junta a ligação da extremidade 5' do *primer* universal a um dos quatro oligonucleotídeos fluorescentes da sonda octâmera, que constitui de três nucleotídeos aleatórios (64 combinações), um dinucleotídeo (que combina com as cores fluorescentes) e um trinucleotídeo universal idêntico em todos os octâmeros. Cada passo adiciona 256 (4 de 64) octâmeros e o sinal fluorescente indica o dímero sequenciado. Os octâmeros são ligados ao *primer* universal e, em preparação para o próximo ciclo, um passo de clivagem remove o trinucleotídeo universal e a etiqueta 5' fluorescente. O processo é repetido até o final do fragmento de DNA, resultando em dois de cada cinco nucleotídeos sequenciados. Para sequenciar o restante do fragmento de DNA, a reação começa novamente com o *primer* de sequenciamento universal 2. Após as cinco interações desse processo, cada nucleotídeo na sequência foi lido duas vezes. Uma vez que existem apenas quatro cores fluorescentes, o processo deve ser repetido quatro vezes, a fim de cobrir todos os possíveis dímeros de sequenciamento.

As três plataformas de análise genômica em larga escala são instrumentos em nanoescala, com sistema para análise de imagem e alto poder de bioinformática que usa supercomputadores (Quadro 1). Apenas a tecnologia SOLiD™ proporciona checagem de erros no processo de sequenciamento, enquanto Genome Analyzer e 454 garantem alta qualidade, mas não implementam medidas de controle de qualidade. Por outro lado, a plataforma SOLiD™ necessita de um sistema computacional massivo e recursos para estocar dados, por causa da complexidade da abordagem. Apesar de já terem alcançado grandes avanços, os métodos de sequenciamento em larga escala ainda têm alguns desafios, como a redução de custos, a diminuição nas taxas de erro de sequenciamento e a leitura de fragmentos mais longos. Novos sistemas que não necessitam de amplificação por PCR e que se iniciam a partir de uma única molécula de DNA estão em desenvolvimento, como a tecnologia Helicos Biosystem - True Single Molecule Sequencing e Complete Genomics, que propõe a visualização da exata localização onde o oligonucleotídeo fluorescente de pentâmero incorpora-se ao longo de uma única fita de DNA.

Essa nova geração de tecnologias de sequenciamento fornece uma velocidade e processividade na geração de sequências de DNA sem precedentes, permitindo um impressionante avanço científico e novas aplicações biológicas.

O objetivo de gerar grande quantidade de dados de sequência de organismos relacionados está direcionado com uma aplicação denominada ressequenciamento, que manipula os dados de sequência de diferentes modos que a montagem *de novo* de genoma. No ressequenciamento, a montagem é direcionada para a sequência-referência e requer menor cobertura (8 a 12 X), que a montagem do genoma *de novo* (25 a 70 X).

ANÁLISE DE EXPRESSÃO DIFERENCIAL EM GRANDE ESCALA

A tecnologia de microarranjos de DNA possibilita a avaliação simultânea da expressão de milhares de genes em diferentes tecidos de um determinado organismo, em diferentes estádios de desenvolvimento ou submetidos a condições de estresse. Os microarranjos são bastante utilizados em experimentos de genômica funcional, com diversas espécies animais e vegetais, sendo gradativamente incorporados em diferentes áreas da pesquisa zootécnica, como crescimento e metabolismo, resposta imune a doenças, reprodução e resposta a fatores de

QUADRO 1 - Comparação das três plataformas de sequenciamento em larga escala

Característica	Plataformas de sequenciamento		
	454	Genome Analyzer	SOLiD™
Química de sequenciamento	Pirosequenciamento	Polimerase com base em sequenciamento por síntese	Ligação com base em sequenciamento
Abordagem de amplificação	Emulsão PCR	Amplificação por ponte	Emulsão de PCR
Finais pareados/separação	Sim/3 kb	Sim/200 pb	Sim/3 kb
Mb/corrida	100 Mb	1.300 Mb	3000 Mb
Tempo/corrida (finais pareados)	7 horas	4 dias	5 dias
Comprimento de leitura	250 pb	32-40 pb	35 pb
Custo por corrida (total)	$8,439.00	$8,950.00	$17,447.00
Custo por Mb	$84.39	$5.97	$5.81

FONTE: Dados básicos: Mardis (2008).

NOTA: Mb - Megabases; kb - Quilobases; pb - Pares de bases; PCR - *Polymerase chain reaction*.

estresse não-infecciosos (restrição alimentar, exposição a elementos tóxicos e a outras condições ambientais desfavoráveis), além de melhoramento genético animal. Tais experimentos são consideravelmente caros e como consequência, são, em geral, conduzidos com tamanhos amostrais reduzidos. A realização de experimentos com microarranjos envolve uma série de procedimentos laboratoriais de alta complexidade, desde a coleta das amostras até a obtenção das imagens para análise, que frequentemente introduzem variações adicionais aos resultados (ROSA; ROCHA; FURLAN, 2007).

A Affymetrix, em Santa Clara, Califórnia, dominou o mercado por muitos anos, aplicando tecnologia de fotolitografia para a impressão de oligonucleotídeos em microarranjos de alta densidade. Seu *chip* é bastante usado, mas a dinâmica do mercado está mudando. Novos autores têm lançado arranjos de alta densidade em menor tempo e custo.

A Xeotron da Houston Techonology Center[7], e a NimbleGen, da Roche são duas companhias que produzem microarranjos com tecnologia digital, com base em uma série de pequenos espelhos, tornando a sua impressão mais rápida e com menor custo.

A Agilent Technologies, em Palo Alto, Califórnia, também usa um processo de síntese *in situ* com a impressão de oligos de 60 nucleotídeos, base por base, em lâminas de vidro especialmente preparadas. Cada oligo representa um gene e essa lâmina pode ser lida na maioria dos escâneres comerciais.

Além das lâminas de microarranjos no formato de 3 x 1 polegadas, algumas companhias estão desenvolvendo novos formatos de arranjos que levam a processos paralelos de múltiplas amostras, como é o caso da Illumina® em San Diego, Califórnia. A tecnologia de esferas da companhia está disponível em dois substratos distintos, o Sentrix Array Matrix (para até 96 amostras) e o Sentrix LD BeadChip (para até 8 amostras). Cada arranjo é fabricado para processar múltiplas amostras de cada vez e ambas suportam genotipagem de polimorfismo de nucleotídeo único – *single-nucleotide polymorphism* (SNP) e aplicações para análise de expressão gênica. Cada arranjo em cada substrato contém milhares de pequenos poços nos quais esferas são montadas de maneira aleatória. A companhia usa sondas de 50 nucleotídeos concatenadas com sequências conhecidas imobilizadas na superfície. Após a montagem das esferas, cada arranjo é decodificado, para determinar quais tipos de esfera contêm qual gene em cada um dos poços do substrato (GUNDERSON et al., 2004).

DETERMINAÇÃO DE FUNÇÃO GÊNICA EM GRANDE ESCALA

Genética reversa é um processo de descoberta de genes que ocorre, conforme o próprio nome indica, de forma oposta ao processo de genética clássica. As genéticas clássica e reversa são parecidas, pelo fato de investigadores tipicamente deduzirem a função de um gene por meio de um efeito de mudança no fenótipo. Por outro lado, o contraste dos dois processos está no fato de que a genética clássica procura indivíduos raros com fenótipos não usuais, buscando, então, o gene ou alelo responsável pela característica fenotípica. A localização de um gene associado com tal fenótipo é o ponto final da investigação.

O avanço nas técnicas de sequenciamento de DNA resultou em vários genomas completamente sequenciados e em uma infinidade de sequências gênicas disponíveis. A abundância de tais informações estimulou a genética reversa, onde, com base nas sequências dos genes, procura-

[7]*Disponível no site: http://www.houstontech.org*

se entender a influência delas no fenótipo, descobrindo sua função biológica. Assim, uma série de estratégias pode ser utilizada para auxiliar na determinação da função de genes.

Mutagênese por silenciamento gênico

O RNA de interferência – RNA *interference* (RNAi) é um processo no qual a expressão de um gene específico é inibido por RNA senso e antisenso. Baseia-se na capacidade de sequências de dupla fita reconhecerem e degradarem sequências que sejam complementares a esses (LEWIN, 2004). O RNAi foi primeiramente descrito em *Caenorhabditis elegans*, quando introduzida uma fita dupla de RNA e observou-se o silenciamento da expressão do gene (FIRE et al., 1998; KUTTENKEULER; BOUTROS, 2004). A primeira classe a participar do RNAi é o *double stranded* RNA (dsRNA), que é formado pela complementariedade de bases de duas fitas simples de RNA e automaticamente reconhecidos por um complexo enzimático – RNAses tipo III, específicas para RNAs dupla fita (DICER). Esse primeiro complexo tem atividade RNAse III e digere o dsRNA em fragmentos de 21 a 25 pb. Esses pequenos fragmentos são reconhecidos por um segundo complexo enzimático que se acopla a regiões homólogas desses fragmentos de 25 pb no mRNA alvo (que nesse caso será o próprio genoma do *potyvirus*), degradando-o e impossibilitando que o vírus produza as enzimas necessárias a sua multiplicação. Para que a fita de dsRNA ocorra nesse processo, uma região conservada do genoma do vírus (cerca de 400 pb) é colocada duas vezes na construção gênica sendo que uma delas é invertida em relação à outra. Quando a construção é transcrita, ocasiona a formação de uma dupla fita de RNA (Fig. 2).

O silenciamento, com base em RNAi, é uma excelente estratégia para genética reversa (WATERHOUSE; GRAHAM; WANG, 1998). O RNAi tem sido usado como uma ferramenta poderosa para silenciar genes e analisar a perda de função, quando alelos não mutantes não estão disponíveis (PATTANAYAK et al., 2005). Processos de análise de função gênica em grande escala, usando o RNAi, têm sido usados em processos como TraitMill™[8].

Mutagênese por *transposons* e agrobactéria

Os *transposons* são elementos móveis que podem translocar de uma região do genoma para outra (HAYES, 2003). *Transposons* são sequências de DNA que podem inserir em uma nova localidade do genoma sem ter relação com a região inserida (LEWIN, 2004). A mutagênese

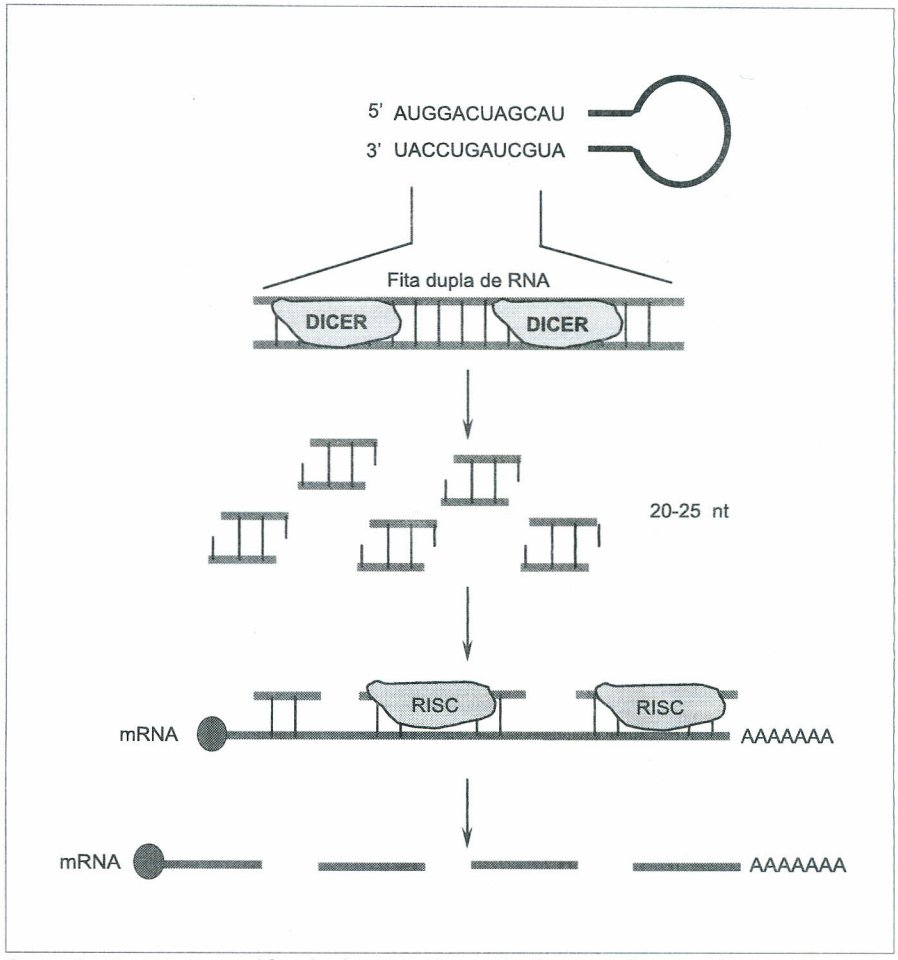

Figura 2 - Esquema simplificado da degradação do dsRNA pelos complexos enzimáticos

NOTA: O fragmento de interesse é inserido duas vezes invertido no vetor, para que ocorra durante a transcrição a formação do grampo de RNA. Esse complexo é reconhecido pela enzima – RNAses tipo III, específicas para RNAs dupla fita (DICER) que fragmenta dessa dupla fita de RNA em fragmentos de 25 bases. Esse produto é então reconhecido pelo complexo silenciador induzido por RNA – RNA *induced silencing complex* – (RISC) que reconhece RNAs produzidos na celular que são homólogos a esse RNA. Esse processo impede que os RNAs provenientes da célula venham a ser traduzidos ocasionando plantas mutantes para o fenótipo específico. A grande vantagem desse sistema é de ser praticamente independente do número de cópias do gene de interesse, já que o alvo é o RNA proveniente dessas cópias.

[8]*Disponível no site: http://www.cropdesign.com*

com base em *transposons*, tem sido usada com sucesso para identificar genes essenciais (HAYES, 2003). Métodos com base em *transposon* têm sido usados em Arabidospsis, milho e outras espécies (STEMPLE, 2004). Um problema da mutagênese por inserção de *transposons* é o grande número de indivíduos necessários para fazer a caracterização fenotípica e identificar a mutação em um gene específico (GILCHRIST; HAUGHN, 2005; TILL et al., 2003).

O segmento do plasmídeo Ti de *Agrobacterium tumefaciens*, conhecido como T-DNA, carrega genes de transformação de plantas e pode ser utilizado para mutagênese insercional. Essa mutagênese é usada para produzir *knockouts* principalmente em Arabidopsis (ALONSO et al., 2003). Esse processo também tem sido utilizado em arroz e milho, porém em uma escala menor, apesar da cobertura estar aumentando (HENIKOFF; TILL; COMAI, 2004). Ao contrário de outros sistemas de identificação de genes, o mecanismo preciso da integração do T-DNA no genoma da planta ainda é desconhecido. Como outras técnicas de supressão de RNA, a mutagênese insercional é limitada pelo hospedeiro e seu alcance é limitado pelos tipos de alelos (MCCALLUM et al., 2000).

Mutagênese por EMS - *Tilling*

A tecnologia *targeting induced local lesions in genomes* (*Tilling*) é um processo de genética reversa desenvolvido por Colbert et al. (2001). Nos processos de mutagênese por *transposons* e agrobactéria é teoricamente possível identificar a região onde foi inserido o fragmento, já que a sequência deste é conhecida. Quanto ao método, baseia-se na identificação de regiões do DNA em que não existe esse fragmento conhecido inserido no local. Em razão de a mutagênese química causar grande densidade de mutações, virtualmente todos os genes podem ser atingidos por esse método. Além disso, trata-se de um processo independente de transgênicos. O método tem como base a capacidade de uma enzima em detectar o desparelhamento de fitas de DNA mutante e normal, quando aneladas. As plantas são tratadas com etilmetanosulfanato (EMS), para gerar uma população de mutações de ponto aleatórias. Por seletividade, o DNA é amplificado com *primers* marcados com fluorescência e os heteroduplexes são formados pelo não-pareamento de algumas bases entre o DNA selvagem e o mutante. Os heteroduplexes são incubados com a endonuclease de planta CEL I (endo-1,4-β-glucanase) que cliva o heteroduplex em sítios não-pareados, resultando em produtos que são visualizados nos sequenciadores automáticos de DNA. Após a análise do DNA de plantas individuais a partir do DNA do *pool*, é possível identificar a planta com a mutação (Fig. 3). Um centro de grande escala de *Tilling* para Arabidopsis está sendo montado no Instituto do Câncer Fred Hutchinton em Seattle-USA (PROWEB PROJECT, 2009). O usuário solicita e recebe sementes de Arabidopsis, contendo mutações no gene de interesse.

FENOTIPAGEM EM GRANDE ESCALA

Uma plataforma de fenotipagem em grande escala para milho e arroz, denominada TraitMill™, tem sido desenvolvida na Bélgica pela CropDesign. Plataformas semelhantes têm sido descritas por outras companhias ao redor do mundo. O princípio dessa plataforma engloba a caracterização de um grande número de plantas digitalmente, utilizando ferramentas de bioinformática, sistema de engenharia de genes, transformação em grande escala e um sistema automatizado

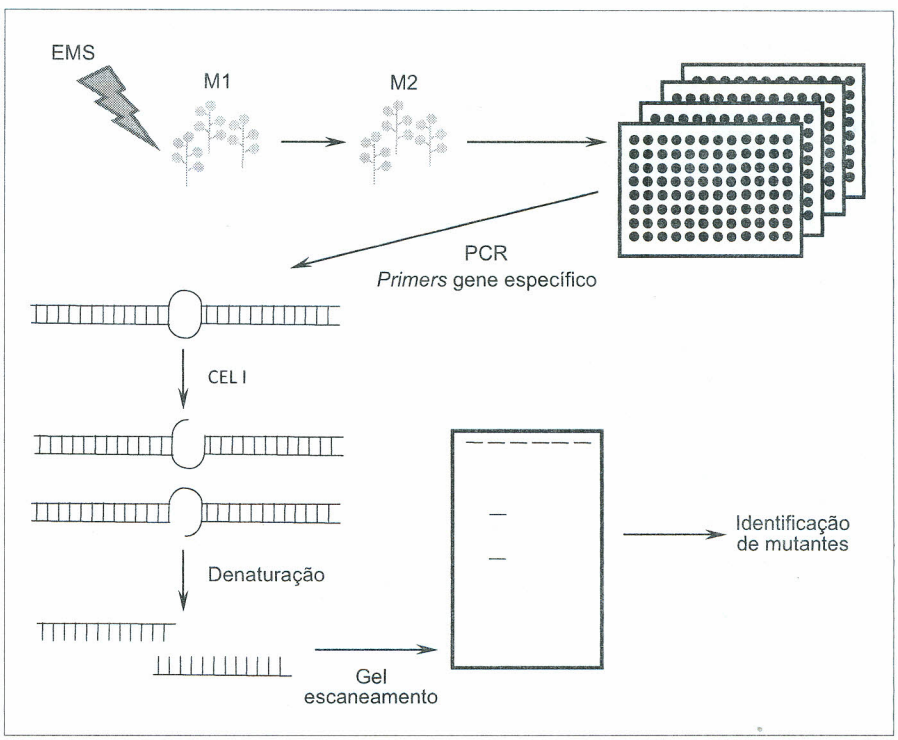

Figura 3 - *Tilling* em alta escala

FONTE: Colbert et al. (2001).

NOTA: Sementes são mutagenizadas em concentrações de 20, 25 e 30 mM de etilmetanosulfonado (EMS). Plantas M1 são colocadas em bandejas e sementes plantadas em vasos para a geração M2, onde cada M2 deriva de uma planta M1 diferente. Os DNAs das plantas M2 são preparados e submetidos a reação em cadeia da polimerase (PCR) usando *primers* específicos. A reação é submetida a tratamento com endo-1,4-β-glucanase (CEL I), limpeza e eletroforese e escaneamento.

de alta resolução para avaliação do fenótipo. Esse processo tem a capacidade de manipular centenas de genes e avaliar o efeito na planta por meio de um sistema automático de avaliação fenotípica global. Parte desse programa investiga genes, vias metabólicas e elementos regulatórios que tenham papéis importantes no crescimento e no desenvolvimento da planta. Os genes são geralmente testados sob o controle de promotores constitutivos e órgãos específicos. Nesse processo, são gerados cerca de 50 mil transformantes independentes por ano. As plantas são cultivadas em casa de vegetação, onde são conduzidas por meio de esteiras automáticas para a captura de imagens de vários ângulos, durante o seu ciclo vegetativo. São processadas mais de 30 mil fotos de plantas por dia. Para assegurar a correta análise dos dados, cada planta leva uma etiqueta que contém um *chip* que faz com que todas as informações sejam registradas automaticamente, quando a planta passa pelo sistema de imagem. Algorítmos são usados para análise digital e extração de vários parâmetros relacionados com o crescimento, produção e tolerância a estresses. As sementes são colhidas para cada planta e as amostras são etiquetadas com código de barra. Um robô analisa o peso das sementes enquanto um sistema de análise de imagem faz sua contagem e determina o seu tamanho (Fig. 4).

CONSIDERAÇÕES FINAIS

Muitas metodologias têm sido desenvolvidas para auxiliar na compreensão das funções gênicas. Um grande número de sequências depositadas nos bancos de dados teve suas funções desvendadas por meio de experimentos de avaliação da expressão e localização de mRNAs e proteínas, assim como na reconstituição de mutantes com fenótipos específicos. O sequenciamento tornou-se um elemento-chave na compreensão da funcionalidade do gene. A recente introdução de instrumentos capazes de produzir milhões de sequências em uma única corrida está rapidamente mudando o cenário da genética, aumentando a capacidade de fornecer respostas com velocidades inimagináveis. Essas tecnologias permitirão o sequenciamento total de genomas a preços cada vez mais acessíveis e com maior rapidez.

Nesta revisão, foram abordadas as mais recentes tecnologias de sequenciamento e de análise de expressão gênica. As técnicas de sequenciamento têm permitido a comparação de vários organismos e/ou situações de forma mais acessível e frequente. As companhias que elaboram microarranjos de DNA têm utilizado dados de sequenciamento, mapeamento e função gênica para montagem de arranjos cada vez mais específicos e direcionados. Os processos de análise gênica, tanto individual quanto em larga escala, fornecerão acesso a novas descobertas sem precedentes em todas as áreas da Biologia, incluindo a agropecuária, nas quais existe um grande interesse na identificação de genes envolvidos com resistência a doenças, tolerância a estresses, aumento da qualidade nutricional, biorremediação de ambientes, dentre outras características de valor econômico, social e ambiental.

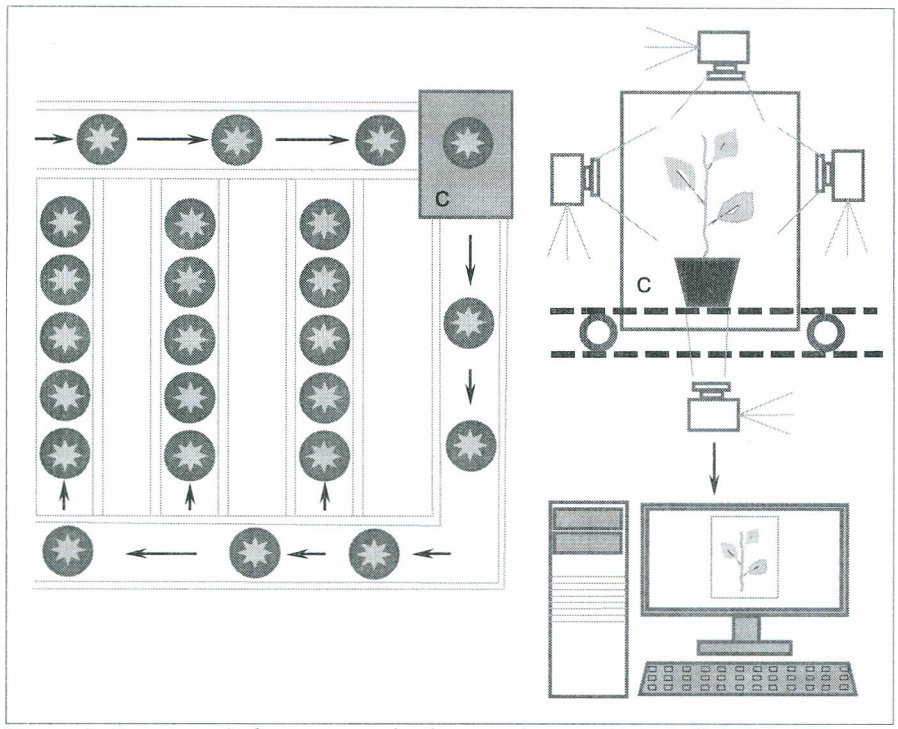

Figura 4 - Descrição da fenotipagem de plantas pelo sistema TraitMill

NOTA: Plantas são dispostas na casa de vegetação em esteiras que se movem em tempo programado. Cada uma das plantas é fotografada digitalmente de vários ângulos e as imagens são processadas em programas de computador que permitem grande número de análises, incluindo crescimento, cor e formato de folhas e caule, densidade de raiz, dentre outras características.

REFERÊNCIAS

ALONSO, J.M. et al. Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. **Science**, v. 301, n. 5633, p.653-657, Aug. 2003.

BHATTRAMAKKI, D. et al. Insertion-deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers. **Plant Molecular Biology**, v. 48, n. 516, p. 539-547, Mar. 2002.

COLBERT, T. et al. High-throughput screening for induced point mutations. **Plant Physiology**, v. 126, p. 480-484, June 2001.

EMRICH, S.J. et al. Gene discovery and annotation using LCM-454 transcriptome sequencing. **Genome Research**, v. 17, n. 1, p. 69-73, Jan. 2007.

EVELAND, A.L.; MCCARTY, D.R.; KOCH, K.E. Transcript profiling by 3'- untranslated

region sequencing resolves expression of gene families. **Plant Physiology**, v. 146, p. 32-44, Jan. 2008.

FIRE, A. et al. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. **Nature**, v. 391, n. 6669, p. 806-811, Feb. 1998.

FU, Y. et al. Quality assessment of maize assembled genomic islands (MAGIs) and large-scale experimental verification of predicted genes. **Proceedings the National Academy of Sciences**, v. 102, n. 34, p. 12282-12287, Aug. 2005.

GILCHRIST, E.J.; HAUGHN, G.W. Tilling without a plough: a new method with applications for reverse genetics. **Current Opinion in Plant Biology**, v. 8, n.2, p. 211-215, 2005

GUNDERSON, K.L. et al. Decoding randomly ordered DNA arrays. **Genome Research**, v. 14, n.5, p. 870-877, May 2004.

HARKINS, T.; JARVIE, T. Metagenomics analysis using the genome sequencer FLX system. **Nature Methods**, v.4, n.6, p. 533, June 2007.

HAYES F. Transposon-based strategies for microbial functional genomics and proteomics. **Annual Review of Genetics**, v.37, p. 3-29, Dec. 2003.

HENIKOFF, S.; TILL, B.J.; COMAI, L. Tilling: traditional mutagenesis meets functional genomics. **Plant Physiology**, v. 135, p. 630-636, June 2004.

KUTTENKEULER, D.; BOUTROS, M. Genome-wide RNAi as a route to gene function in *Drosophila*. **Briefings in Functional Genomics and Proteomics**, v. 3, n. 2, p. 168-176, Aug. 2004.

LEWIN, B. **Genes VII.** New Jersey: Pearson Prentice Hall, 2004. 1027p.

MARDIS, E.R. The impact of next-generation sequencing technology on genetics. **Trends in Genetics**, v. 24, n.3, p. 133-141, Mar. 2008.

MARGULIES, M. et al. Genome sequencing in microfabricated high-density picolitre reactors. **Nature**, v. 437, n. 7057, p. 376-380, Sept. 2005.

MCCALLUM C.M. et al. Targeted screening for induced mutations. **Nature Biotechnology**, v. 18, n. 4, p.455-457, Apr. 2000.

PATTANAYAK, D. et al. Small but mighty RNA-mediated interference in plants. **Indian Journal of Experimental Biology**, v. 43, n. 1, p. 7-24, Jan. 2005.

PROWEB PROJECT. [S.l.: 2009]. Disponível em: <http://www.proweb.org>. Acesso em: 19 nov. 2009.

RONAGHI, M. Pyrosequencing sheds light on DNA sequencing. **Genome Research**, v. 11, p. 3-11, 2001.

ROSA, G.J. de M.; ROCHA, L.B. da; FURLAN, L.R. Estudos de expressão gênica utilizando-se *microarrays*: delineamento, análise, e aplicações na pesquisa zootécnica. **Revista Brasileira de Zootecnia**, Viçosa, MG, v.36, p. 185-209, 2007. Suplemento especial.

SANGER, F. et al. Nucleotide sequence of bacteriophage x 174 DNA. **Nature**, v. 265, n. 5596, p. 687-695, Feb. 1977.

_____; NICKLEN, S.; COULSON, A.R. DNA sequencing with chain-terminating inhibitors. **Proceedings of the National Academy of Sciences**, v. 74, n. 12, p. 5463-5467, Dec. 1977.

SCHNABLE, P.S.; HOCHHOLDINGER, F.; NAKAZONO, M. Global expression profiling applied to plant development. **Current Opinion in Plant Biology**, v. 7, p. 50-56, 2004.

SCHUSTER, S.C. Next-generation sequencing transforms today's biology. **Nature Methods**, v. 5, n.1, p. 16-18, Jan.2008.

SHENDURE, J. et al. Accurate multiplex polony sequencing of an evolved bacterial genome. **Science**, v. 309, n. 5741, p. 1728-1732, Sept. 2005.

STEMPLE, D.L. Tilling – a high-throughput harvest for functional genomics. **Nature Reviews Genetics**, v. 5, p. 145-150, Feb. 2004.

TILL, B.J. et al. Large-scale discovery of induced point mutations with high-throughput Tilling. **Genome Research**, v. 13, n. 3, p. 524–530, Mar. 2003.

VROH BI et al. Single nucleotide polymorphisms and insertion-deletions for genetic markers and anchoring the maize fingerprint contig physical map. **Crop Science**, v. 46, n. 1, p. 12-21, Jan./Feb. 2006.

WATERHOUSE, P.M.; GRAHAM, M.W.; WANG, M. B. Virus resistance and gene silencing in plants can be induced by simultaneous expression of sense and antisense RNA. **Proceedings of the National Academy of Science**, v. 95, p. 13959-13964, Nov. 1998.

BIBLIOGRAFIA CONSULTADA

BAURLER, I.; LAUX, T. Apical meristems: the plant's fountain of youth. **Bioessays**, v. 25, n. 10, p.961-970, Sept. 2003.

DRESSMAN, D. et al. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. **Proceedings of the National Academy of Sciences**, v. 100, n. 15, p. 8817-8822, July 2003.

GUYOMARC'H, S. et al. Regulation of meristem activity by chromatin remodelling. **Trends in Plant Science**, v. 10, p. 332-338, 2005.

WEBER, A.P.M. et al. Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. **Plant Physiology**, v. 144, n.1, p.32-42, May 2007.

186 mm x 50 mm