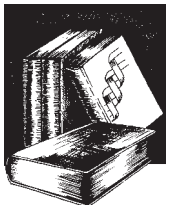


PRACE PRZEGLĄDOWE

Sekwencjonowanie porównawcze genomów: generowanie markerów genetycznych typu INDEL i SNP

Piotr A. Ziółkowski[1], Danuta Babula-Skowrońska[2], Małgorzata Kaczmarek[2], Agata Cieśla[2], Jan Sadowski[1,2]

[1]Zakład Biotechnologii, Instytut Biologii Molekularnej i Biotechnologii, Wydział Biologii, Uniwersytet im. Adama Mickiewicza, Poznań
[2]Pracownia Genomiki Funkcjonalnej, Instytut Genetyki Roślin, Polska Akademia Nauk, Poznań

Comparative sequencing of genomes: generating of INDEL and SNP genetic markers

Summary

Comparison of genome sequences has become an important approach to identify and understand biological significance of the variations and fluxes that occur through a genome. The main subject of the work concentrates on identification of indels and SNPs in large genomes and their potential application in biotechnology. Importantly, fine elaboration of genome structure and sequence polymorphism that results from resequencing promises to benefit breeding, biotechnology and medical research. The article also describes how the data extracted from comparative studies of genomes depends on phylogenetic distances of the species involved.

Key words:
comparative sequencing, resequencing, genome structure, DNA polymorphism, genetic markers, indel, SNP, biotechnology.

Adres do korespondencji

Jan Sadowski,
Zakład Biotechnologii,
Instytut Biologii Molekularnej
i Biotechnologii,
Wydział Biologii,
Uniwersytet
im. Adama Mickiewicza,
ul Umultowska 89,
61-614 Poznań;
Pracownia Genomiki Funkcjonalnej,
Instytut Genetyki Roślin,
Polska Akademia Nauk,
ul. Strzeszyńska 34,
60-479 Poznań;
e-mail: jsad@amu.edu.pl

1. Wstęp

1.1. Analiza porównawcza sekwencji DNA

Porównywanie sekwencji nukleotydowych jest obecnie w biologii jednym z najbardziej rozpowszechnionych i podstawowych podejść badawczych. Wraz ze wzrostem liczby różnorodnych sekwencji DNA, dostępnych w odpowiednich bazach danych, można kompletować wstępne informacje o strukturze i przewidywanej funkcji sekwencji nas interesujących. Narzędziem najpowszechniej wykorzystywanym w analizie porównawczej, jest BLAST (ang. *Basic Local Alignment Search Tool*) (1). Umożliwia on porównanie i ocenę podobieństwa sekwencji o bardzo różnej długości: od kilku do wielu milionów nukleotydów. Najbardziej złożoną i zarazem najważniejszą analizą w tym zakresie jest analiza porównawcza sekwencji całych genomów między badanymi osobnikami, liniami, odmianami, gatunkami itd. Porównanie takie dostarcza bardzo szczegółowych informacji na temat podobieństwa genetycznego, pochodzenia i ewolucji badanych organizmów. Coraz szerzej stosowane resekwencjonowanie różnych genotypów umożliwia efektywną identyfikację SNP (ang. *Single Nucleotide Polymorphism*), rearanżacji chromosomowych, mapowanie breakpointów, detekcję rzadko występujących wariantów sekwencyjnych, etc. Te całogenomowe informacje są podstawą budowania technologii genomowej opartej na różnego typu mikromacierzach DNA. Jednym z ważniejszych obecnie poszukiwań w dziedzinie genetyki, hodowli i biotechnologii jest ustalenie możliwie licznych różnic w sekwencji badanych genotypów w celu zidentyfikowania i opracowania efektywnych sekwencji znacznikowych (markerowych). Sekwencje takie są niezbędne w rozwoju efektywnej diagnostyki medycznej, w genetyce, oraz hodowli i biotechnologii.

1.2. Sekwencjonowanie szerokoprzepustowe i resekwencjonowanie genomów

Obecnie podstawą określania sekwencji całogenomowych są nowe technologie szerokoprzepływowego sekwencjonowania, takie jak Roche-454 Life Sciences (www.roche-applied-science.com), ABI-SOLiD (www.appliedbiosystems.com) i Solexa-Illumina (www.illumina.com). Technologie te są nadal doskonalone, przez co stają się coraz wydajniejsze, dokładniejsze i ponadto coraz tańsze. Umożliwiają już obecnie, nawet stosunkowo małemu zespołowi, przeprowadzenie kompletnej analizy genomowej, włączając w to koszt samego sekwencjonowania wykonanego usługowo przez wyspecjalizowane firmy.

Wspomniane, nowe strategie sekwencjonowania stoją obecnie u podstaw wielkoskalowych projektów porównawczego sekwencjonowania genomu różnych osobników lub linii w obrębie gatunku, jak i spokrewnionych gatunków. W pierwszym rzędzie analizą taką objęto porównanie genomów prokariotów, np. bakterii choro-

botwórczych (2), także genomów pojedynczych komórek; zakres i znaczenie tego ostatniego podejścia sprawiły, że powstała kolejna, nowa gałąź genomiki nazwana genomiką pojedynczych komórek. Liczne są już doniesienia o resekwencjonowaniu różnych linii gatunków bakterii. Ciekawym spostrzeżeniem w tych badaniach był fakt, że linie należące do tego samego gatunku wykazują znaczące różnice w sekwencji, w tym tak istotne, jak obecność grupy genów, które występują w genomie tylko jednej linii (nie wykryto ich w innych liniach tego gatunku). Obserwacja ta dotyczy wszystkich sekwencjonowanych dotychczas linii badanych gatunków i doprowadziła do powstania koncepcji pangenomu bakteryjnego (3). U postaw tej koncepcji stoi interesująca hipoteza, w której zakłada się, że wykryte dodatkowe sekwencje genomu mogą pełnić funkcje nie będące podstawowymi, natomiast mogą dostarczać nowych cech (w tym przewagę selekcyjną) przystosowujących do nowych niszy, oporność antybiotykową czy kolonizowanie nowych gospodarzy.

W przypadku porównań genomów organizmów eukariotycznych przeprowadzono szereg porównań mniej lub bardziej spokrewnionych gatunków, a także resekwencjonowanie różnych osobników i linii tego samego gatunku. Zakres tych badań był jednak z oczywistych powodów (wielkość genomów) znacznie skromniejszy.

2. Identyfikacja indeli w genomie *Arabidopsis thaliana*

2.1. Identyfikacja polimorfizmów typu indel

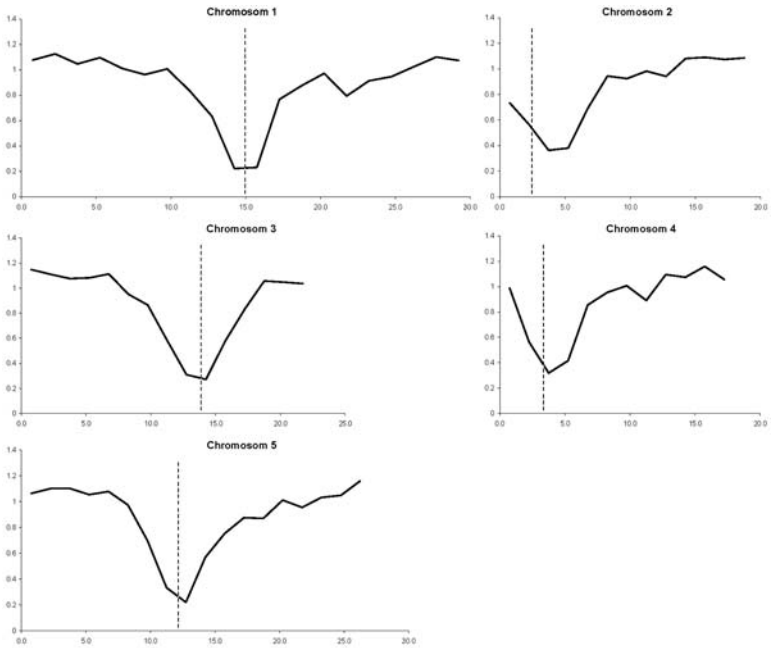
Polimorfizmy typu **in**sercja-lub-**del**ecja (indel) są najczęstszymi, obok SNP, polimorfizmami występującymi w genomach eukariotycznych. Są one identyfikowane najczęściej poprzez porównanie sekwencji nukleotydowej dwóch linii lub odmian: wykryta w jednej linii insercja jest jednocześnie delecją w drugiej linii; bez porównania większej liczby linii i analizy mechanizmu pochodzenia, nie można w sposób jednoznaczny określić, czy mamy do czynienia z insercją, czy też delecją. Wielkość mutacji indel waha się w bardzo szerokim przedziale od 1 pz do nawet kilku milionów pz; dlatego też z reguły dzieli się je na tzw. małe indele (ang. *small indels*) o wielkości od 1 pz (taka mutacja bywa też często definiowana jako SNP) do 100 pz, oraz duże indele (ang. *large indels*) $\geqslant$100 pz (4). Przeprowadzona przez nas szacunkowa analiza samych tylko dużych indeli różniących dwie linie *A. thaliana* (Col i L*er*) sugeruje obecność ok. 8500 polimorfizmów tego typu (5); jest natomiast pewne, że małe indele są o kilka rzędów wielkości częstsze; dane te wskazują na potencjał jaki przedstawiają polimorfizmy typu indel w projektach hodowlanych (6).

W przeprowadzonej przez nas analizie polimorfizmów dużych indeli w genomie *A. thaliana* wykorzystaliśmy dostępność sekwencji genomowej dla dwóch linii – Col i L*er*. Pierwsza z nich była przedmiotem projektu sekwencjonowania genomu *A. thaliana* oparta na mapie fizycznej zbudowanej za pomocą klonów BAC (7). Druga

natomiast, to wynik sekwencjonowania przeprowadzonego z wykorzystaniem techniki „shotgun" przez Cereon Genomics (obecnie część Monsanto Co.)(4). Obie linie są stosunkowo blisko ze sobą spokrewnione, a czas ich ewolucyjnego rozejścia został oszacowany na ok. 200 tys. lat (5). Sekwencja L*er* była dostępna w postaci 81 306 kontigów o średniej długości ok. 1,3 kpz.

W pierwszej fazie analizy kontigi L*er* zostały przyrównane do sekwencji Col reprezentowanej przez pięć tzw. pseudochromosomów (odtworzona sekwencja całego chromosomu z kilkoma przerwami, głównie w okolicach centromerowych) z wykorzystaniem zestawu skryptów specjalnie do tego celu przygotowanych. Łącznie przypisano w ten sposób blisko 70% kontigów, pokrywając w ten sposób 58,4% genomu Col. Kontigi, które zostały odrzucone, składały się niemal wyłącznie z sekwencji powtarzających się, bądź mogły być nieprawidłowo złożone.

Przy tej okazji należy zwrócić uwagę na fakt, że przy odtwarzaniu sekwencji chromosomowej zawsze pojawiają się regiony problematyczne, dla których nie można ustalić naturalnego porządku sekwencji nukleotydowej. Dotyczy to w szczególności regionów bogatych w sekwencje powtarzające się, zarówno o charakterze satelitarnym (np. powtórzenia centromerowe), jak i transpozonów (skupiska elemen-



Rys. 1. Przypisywanie kontigów L*er* do chromosomów Col. Oś X – reprezentuje pozycję na chromosomie Col mierzoną w ramkach o długości 1,5 Mpz; oś Y – określa pokrycie danego regionu kontigmi L*er*. Przykładowo, jeśli Y uzyskuje wartość równą 1, oznacza to, że dany region o długości 1,5 Mpz jest pokryty kontigami w 2/3 długości. Pionowe, przerywane linie pokazują przybliżoną pozycję centromerów (5).

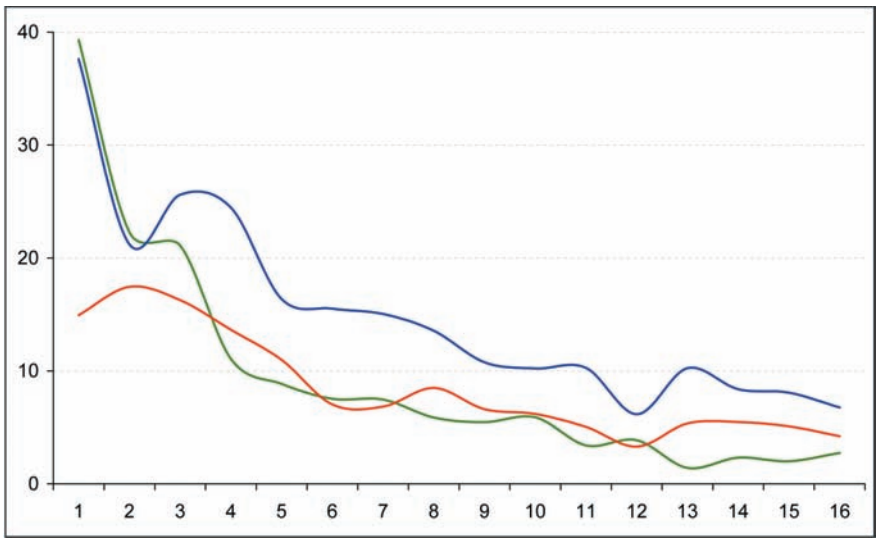
tów ruchomych w regionach przycentromerowych oraz heterochromatynowych). Ogromna liczba sekwencji powtarzających się o wysokiej homogeniczności uniemożliwia w tych regionach definitywne przypisanie sekwencji, przez co wszystkie dostępne obecnie algorytmy składające kontigi, jak i budujące alignmenty pomijają te obszary tworząc puste przerwy. Problem z odtwarzaniem takich regionów uwidocznił się także podczas naszych badań porównawczych – regiony przycentromerowe wykazywały wielokrotny spadek liczby przypisanych regionów (rys. 1).

Następnym etapem analizy było wyszukanie przerw w alignmentach o długości 100-20 000 pz. Górna granica indeli została przyjęta arbitralnie, z uwagi na fakt, że największe transpozony u rzodkiewnika mają wielkość ok. 15 kpz. W ten sposób zidentyfikowano 6636 mutacji indel. W wielu przypadkach, ze względu na niewielką długość kontigów L*er*, nie było możliwe określenie obu granic mutacji, co jest warunkiem koniecznym do obliczenia długości indela oraz scharakteryzowania mechanizmu odpowiadającego za jego powstanie. Dlatego do dalszej analizy wybrano tylko 2201 mutacji, które posiadały zdefiniowane oba końce i były insercjami w Col (bądź delecjami w L*er*).

2.2. Rozkład mutacji indel i SNP wzdłuż chromosomów

Polimorfizmy, aby mogły być efektywnymi markerami genetycznymi, muszą rozkładać się równomiernie wzdłuż chromosomów tak, by każdy region chromosomowy był odpowiednio wysycony. W idealnej sytuacji rozkład taki przybierałby postać prostej równoległej do osi X, reprezentującej pozycję na chromosomie. Niestety, w naturze rozkład miejsc polimorficznych znacząco odbiega od tego schematu, przy czym niektóre regiony chromosomowe mogą zawierać od kilkunastu do kilkuset więcej polimorfizmów, aniżeli średnia chromosomowa. Z reguły największe dysproporcje dotyczą obszarów przycentromerowych. W regionach tych, ze względu na względnie niskie wysycenie sekwencjami kodującymi, ograniczoną rekombinację oraz wysoki poziom metylacji DNA, dochodzi do akumulacji sekwencji powtarzających się. Ich modyfikacja czy eliminacja nie wpływa negatywnie na funkcjonowanie organizmu, dlatego właśnie tempo zmian i poziom polimorfizmu dla tych regionów jest największy.

Z przeprowadzonej przez nas analizy wynika, że rozkład częstości poszczególnych typów indeli wzdłuż chromosomów nie jest jednakowy (rys. 2). Liczba indeli związanych z aktywnością transpozonów wzrasta stopniowo wraz z przesuwaniem się od telomeru w kierunku centromeru i przechodzi we wzrost wykładniczy ok. 2 Mpz przed centromerem. Ten typ mutacji jest silnie negatywnie skorelowany z częstością występowania genów, stąd właśnie bierze się ich akumulacja w regionach przycentromerowych (8). Z kolei rozkład częstości mutacji spowodowanych rekombinacją niewyrównaną jest odmienny: częstość indeli w tej kategorii wzrasta stopniowo w kierunku centromeru, osiągając maksimum ok. 1,5 Mpz przed sa-



Rys. 2. Częstotliwość występowania mutacji indel (oś Y) mierzona względem uśrednionego ramienia chromosomowego *A. thaliana*, w sekcjach o długości 1 Mpz (oś X) poczynając od regionu (pery)centromerowego w kierunku telomerów. W kolorze czerwonym – indele powstałe na drodze rekombinacji niewyrównanej; w kolorze zielonym – indele będące rezultatem aktywności transpozonów; w kolorze niebieskim – indele utworzone w wyniku rekombinacji nieuprawnionej (5).

mym centromerem, po czym stopniowo obniża się. Jest to spowodowane, jak się wydaje, powiązaniem rekombinacji niewyrównanej z ogólną rekombinacją homologiczną, która jest silnie inhibowana w okolicach centromerów (9). Przeprowadzone przez nas analizy statystyczne wykazały istotną korelację tych dwóch wartości. Rozkład częstości indeli wywołanych działaniem rekombinacji nieuprawnionej jest natomiast bimodalny stanowiąc swego rodzaju stan pośredni pomiędzy poprzednimi rozkładami: uzyskuje on dwa maksima, jedno położone ok. 2,5 Mpz przed centromerem, oraz drugie, formujące wyraźny wierzchołek przycentromerowy (rys. 2).

Chociaż uśrednione analizy rozkładu mutacji wzdłuż ramion chromosomowych nie były, jak dotąd, analizowane dla polimorfizmów SNP, możemy jednak przyrównać dystrybucję liczby SNP (10) i indeli (5) wzdłuż całych chromosomów *A. thaliana*. Okazuje się, że oba typy polimorfizmów wykazują bardzo podobne rozkłady. Dotyczy to nie tylko regionów przycentromerowych, ale również ramion chromosomowych, na których pojawiają się dodatkowe wierzchołki zwiększonej częstości występowania indeli i SNP. Na tej podstawie można wnioskować, że obszary o zwiększonym poziomie polimorfizmu są w znacznym stopniu tożsame dla obu typów mutacji. Z tych też względów polimorfizmy SNP i indel mogą być wykorzystywane zamiennie w pracach hodowlanych i innych działaniach aplikacyjnych.

2.3. Analiza molekularnych mechanizmów powstawania mutacji indel

W przeciwieństwie do mutacji SNP, które są stosunkowo homogenne pod względem mechanizmów powstawania (czytaj poniżej), mutacje indel mogą zachodzić na wiele sposobów. W przeprowadzonych przez nas badaniach polimorfizmów typu indel w genomie *A. thaliana* najczęściej wykrywane były trzy mechanizmy: rekombinacja niewyrównana, wstawienie/wycięcie transpozonu oraz rekombinacja nieuprawniona (tab. 1) (5).

Tabela 1

Liczba i wielkość mutacji indel zidentyfikowanych poprzez porównanie sekwencji genomowych linii Col i L*er* (5)

	Mechanizm powstania				Razem
	nieznany	rekombinacja niewyrównana	rekombinacja nieuprawniona	wstawienie/wycięcie transpozonu	
liczba	151	569	980	526	2201
udział procentowy	6,9%	25,9%	44,5%	23,9%	
mediana wielkości	1531	2924	215	1314	682

Rekombinacja niewyrównana zachodzi w sytuacji, gdy parują ze sobą dwa homologiczne regiony znajdujące się na tym samym chromosomie, z reguły w niewielkiej odległości od siebie (<10 kpz). W wyniku rekombinacji dochodzi do duplikacji, bądź delecji, odcinka chromosomu znajdującego się pomiędzy regionami homologii. W naszych badaniach mutacje generowane przez ten mechanizm były identyfikowane na drodze wyszukiwania odcinków homologii flankujących obszar insercji, bądź delecji (podobieństwo sekwencji >95%, długość regionów homologicznych ⩾50 pz). Porównanie całogenomowej sekwencji linii Col i L*er* pozwoliło na identyfikację 569 mutacji tego typu, przy czym mediana długości mutacji wynosiła 2924 pz. Warto podkreślić, że ten mechanizm odgrywa decydującą rolę w ewolucji genów i sekwencji kodujących (tab. 2) (5).

Tabela 2

Mutacje indel wykryte w genach i sekwencjach kodujących (5)

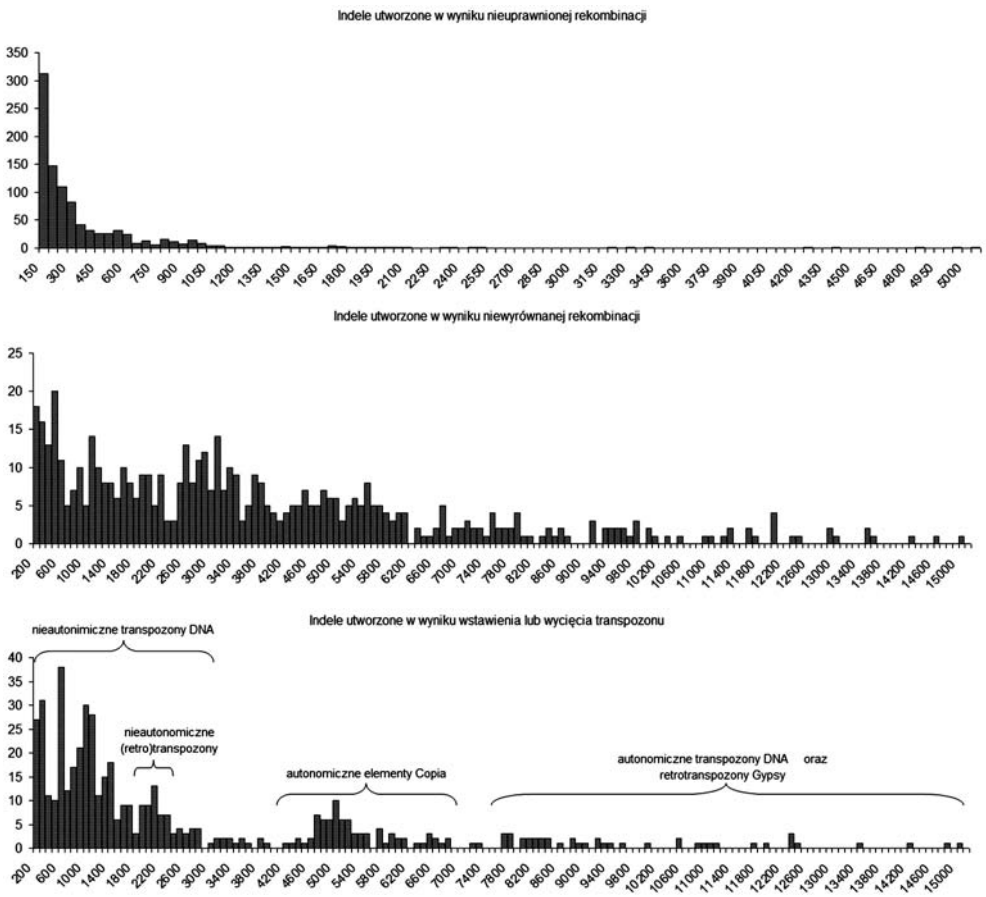
		Rekombinacja niewyrównana	Wstawienie/wycięcie transpozonu	Rekombinacja nieuprawniona	Razem
w genach	liczba	392	63	400	894
	%	43,8	7,0	44,7	
w sekwencjach kodujących	liczba	375	32	194	631
	%	59,4	5,1	30,7	

Wstawienie, bądź wycięcie transpozonów, jest zjawiskiem bardzo częstym w genomach eukariotycznych, tak zwierzęcych, jak i roślinnych. Chociaż transpozony często są inaktywowane poprzez mechanizmy epigenetyczne, takie jak metylacja DNA, czy też liczne delecje generowane przez rekombinację nieuprawnioną oraz niewyrównaną, pewna liczba transpozonów stale pozostaje aktywna i poprzez integrację z genomem gospodarza powoduje powstawanie mutacji indel. W analizie polimorfizmu różnicującego linie Col i L*er* zastosowaliśmy podejście, które opiera się na wykorzystaniu programu RepeatMasker (A.F.A. Smit, R. Hubley i P. Green, RepeatMasker Open-3.0. 1996 do 2004; http://www.repeatmasker.org) do wyszukiwania sekwencji powtarzających się: insercje znajdujące się w jednej, bądź drugiej linii były testowane pod kątem obecności sekwencji transpozonów. Na tej podstawie można było jednoznacznie określić pojawienie się mutacji indel jako wynik wstawienia lub wycięcia transpozonu, przy czym dla retrotranspozonów zawsze mamy do czynienia z wstawieniem nowego elementu. Łącznie zidentyfikowaliśmy 526 przypadków indeli, które powstały w wyniku aktywności transpozonów, a mediana ich długości wynosiła 1314 pz. Na podstawie dalszej analizy wykazano, że udział indeli związanych z aktywnością transpozonów w ewolucji genów i sekwencji kodujących jest bardzo mały w porównaniu ze znaczeniem dwóch pozostałych mechanizmów generujących indele (tab. 2).

Rekombinacja nieuprawniona, określana również jako niehomologiczne łączenie końców (NHEJ, ang. *Non-Homologous End Joining*), polega na delecji, rzadziej duplikacji, fragmentu DNA znajdującego się pomiędzy dwoma dwuniciowymi pęknięciami DNA. Ze względu na fakt, że DNA pęka na obu niciach, nie jest możliwe zastosowanie jednego z mechanizmów naprawczych opartych na homologii sekwencji nukleotydowej. Dotyczy to w szczególności niedzielących się komórek somatycznych, w których chromosomy homologiczne nie są sparowane i przez to nie mogą być wykorzystane jako matryca do naprawy DNA. Jednak pęknięcia nici są przeważnie przesunięte względem siebie (tzn. powstają wystające końce), w związku z tym podczas łączenia końców powstają charakterystyczne krótkie powtórzenia, które znakują granice mutacji (tzw. mikrohomologie). Warto podkreślić fakt, że ten mechanizm naprawczy powszechnie występuje u kręgowców i roślin wyższych w miejsce naprawy opartej na sekwencjach homologicznych, która przeważa u bakterii i niższych eukariontów (np. u drożdży *Saccharomyces cerevisiae*) (11). Poprzez porównanie linii Col i L*er* zidentyfikowaliśmy łącznie 980 przypadków polimorfizmów indel, które były wynikiem działania tego właśnie mechanizmu. Stanowi to blisko 50% wszystkich wykrytych przez nas mutacji (tab. 1). Jednocześnie indele generowane przez rekombinację nieuprawnioną były najmniejsze, z medianą wynoszącą zaledwie 215 pz. Interesujące jest, że stosunkowo duża liczba mutacji indel odnalezionych w genach i sekwencjach kodujących była wynikiem zajścia właśnie tego mechanizmu – świadczy to o dużym znaczeniu nieuprawnionej rekombinacji w ewolucji genów. Należy zaznaczyć, że przeprowadzone przez nas badania obejmowały wyłącznie polimorfizm tzw. dużych indeli (o wielkości ⩾100 pz); w innych badaniach

wskazuje się jednakże na znaczną liczbę stosunkowo krótkich mutacji (często <100 pz) na tej drodze wywołanych (12,13). W konkluzji można zatem stwierdzić, że nieuprawniona rekombinacja jest najważniejszym mechanizmem odpowiedzialnym za powstawanie mutacji typu indel w genomach organizmów eukariotycznych.

Warto zwrócić uwagę na fakt, że wielkość mutacji indel w dużym stopniu zależy od mechanizmu, który odpowiada za jej powstanie. Różnice w tym względzie są bardzo widoczne, co przedstawiono na rysunku 3 (5). Rozkład wielkości mutacji generowanych przez rekombinację niewyrównaną jest stosunkowo duży; choć największa liczba przypadków w tej kategorii ma długość ~500 pz, to wysoka częstość mutacji jest obserwowana nawet do 6 kpz. Wynik ten wskazuje na szeroki zakres długości indeli generowanych w wyniku działania rekombinacji niewyrównanej.



Rys. 3. Histogram liczby indeli (oś Y) w zależności od długości wstawionego/wyciętego fragmentu DNA (oś X). W przypadku indeli powstałych w wyniku rekombinacji nieuprawnionej zastosowano wielkość przedziału 150 pz, natomiast dla pozostałych dwóch kategorii polimorfizmów zastosowano przedziały o wielkości 300 pz (2).

Z kolei mutacje powstałe na skutek aktywności sekwencji powtarzających się charakteryzują się wielomodalnym rozkładem, gdy analizuje się ich wielkość. Pojawiają się tutaj wierzchołki w przedziale długości od 150 do 2600 pz, silny wierzchołek przy mutacjach ~5 kpz, a także wierzchołki przy ~8 i ~10 pz. Odpowiadają one odpowiednio indelom powstałym poprzez włączenie/wycięcie 1) nieautonomicznych transpozonów, 2) retroelementów z rodziny copia, oraz 3) autonomicznych transpozonów DNA i retrotranspozonów z rodziny gypsy (rys. 3).

Natomiast rozkład częstotliwości występowania mutacji związanych z rekombinacją nieuprawnioną wykazuje charakterystyczny przebieg hiperboidalny: liczebność w tej kategorii jest największa przy długości 100 pz, od której rozpoczęty był pomiar, i gwałtownie spada wraz ze wzrostem długości indeli. Ponieważ krótkie mutacje, jak się wydaje, są najlepsze w kontekście ich zastosowania jako markery genetyczne, to właśnie polimorfizmy generowane przez rekombinację nieuprawnioną mają największe znaczenie w selekcji.

2.4. Możliwości zastosowania polimorfizmów indel w procesie hodowlanym i selekcji nowych odmian

Polimorfizmy indel bez wątpienia mogą być zastosowane do tworzenia wydajnych systemów markerowych, analogicznie do polimorfizmów SNP. Jednakże sama identyfikacja polimorfizmów indel jest trudniejsza, aniżeli identyfikacja SNP. Wynika to z faktu, że większe mutacje wymagają zsekwencjonowania dłuższych fragmentów genomowego DNA, takich, które pokryłyby całą insercję/delecję. W prezentowanej przez nas analizie stosowane były dane pochodzące z projektów opartych na sekwencjonowaniu metodą „BAC-by-BAC" oraz „shotgun", które dostarczyły stosunkowo długich sekwencji. Rozwijające się dynamicznie w ostatnim czasie podejścia wysoko przepustowe każą zadać pytanie, na ile identyfikacja polimorfizmów indel może być przeprowadzona przy użyciu właśnie tych metod.

Dotychczas u rzodkiewnika pospolitego stosowano analizę sekwencji opartą na hybrydyzacji do mikromacierzy DNA pokrywających cały genom (tzw. mikromacierze tillingowe). Clark i in. (10) wykorzystał tę technologię do poznania polimorfizmu 20. różnych linii *A. thaliana* identyfikując ponad milion polimorfizmów SNP. Autorzy stwierdzili również, że ok. 4% jest niezachowanych pomiędzy poszczególnymi liniami, przy czym nie udało się jednoznacznie stwierdzić, czy było to rezultatem dużych delecji danego regionu, czy też znacznym stopniem ich dywergencji. W konsekwencji trzeba stwierdzić, że przytoczona technika hybrydyzacji do mikromacierzy nie jest przydatna w identyfikacji mutacji indel większych, aniżeli kilka par zasad (10). Z kolei technologie oferowane przez takie firmy, jak Illumina, Roche, czy Applied Biosystems generują wciąż stosunkowo krótkie sekwencje, co ogranicza ich wykorzystanie do identyfikacji polimorfizmów indel (13). Nie ulega jednak wątpliwości, że wraz ze wzrostem długości odczytów oraz dalszego obniżania kosztów

nowoczesne technologie sekwencjonowania wysoko przepustowego zostaną zaprzęgnięte również i do tego celu. Należy jednak zauważyć, że jedną sprawą jest identyfikacja mutacji indel, inną natomiast wykorzystanie już wykrytych mutacji do przeglądania polimorfizmu występującego w populacji pomiędzy osobnikami: ta druga kwestia jest o wiele prostsza metodycznie. Pierwsze próby konstrukcji specjalnych mikromacierzy umożliwiających mapowanie polimorfizmów indel zostały już przeprowadzone i zakończyły się sukcesem (14,15).

3. Identyfikacja SNP

3.1. Mechanizmy powstawania

SNP, czyli polimorfizm pojedynczego nukleotydu, jest najczęściej obserwowanym typem polimorfizmu DNA. Duża liczba i wysoka stabilność części SNP w genomach czyni z nich doskonałe markery genetyczne (16). Przyczyną pojawiania się SNP są mutacje punktowe, jednak nie każda z nich jest uznawana za SNP. Aby mutacja mogła być sklasyfikowana jako SNP, oba warianty (allele) *locus* muszą być reprezentowane z częstością $>1\%$ w populacji.

Mutacje prowadzące do SNP są spowodowane spontanicznymi błędami podczas replikacji chromosomów: dzieje się tak, gdy mimo działania mechanizmów naprawy, polimeraza DNA włącza błędny nukleotyd. Gdy cząsteczka z błędnym nukleotydem ulegnie replikacji, to jedna z helis potomnych będzie posiadała prawidłową sekwencję wyjściową, a druga sekwencję zmutowaną. Czynnikami wywołującymi mutacje jednonukleotydowe są również mutageny chemiczne i fizyczne. Mutageny chemiczne mogą występować w formie analogów zasad, czynników deaminujących, alkilujących i interkalujących, co może powodować błędy w parowaniu nukleotydów czy delecję. Najpowszechniejszymi mutagenami fizycznymi są promieniowanie nadfioletowe (UV), promieniowanie jonizujące i ciepło.

Teoretycznie, dla każdego *locus* istnieje możliwość występowania w populacji czterech form allelicznych (z uwagi na obecność czterech typów nukleotydów), jednak w rzeczywistości SNP występują zazwyczaj jako formy bialleliczne (17,65). Jest to związane z większą częstotliwością występowania tranzycji (zamiana puryny na purynę A-G, G-A lub pirymidyny na pirymidynę C-T, T-C) niż transwersji (zamiana puryny na pirymidynę A-C, A-T, G-C, G-T i pirymidyny na purynę C-A, C-G, T-A, T-G). Dzieje się tak prawdopodobnie na skutek spontanicznej deaminacji 5-metylocytozyny do tymidyny w przypadku dwunukleotydów CG (18).

SNP można podzielić na kilka typów ze względu na miejsce ich występowania w genomie. Znaczna większość SNP zlokalizowana jest w rejonach międzygenowych i stanowi tzw. mutacje ciche. Nie wpływają one na funkcjonowanie organizmu, jednak w przypadku genomu ludzkiego, w niektórych przypadkach ich obecność może

korespondować z ryzykiem zachorowania na pewne choroby. SNP zlokalizowane w regionach regulatorowych genów mogą prowadzić do zmiany poziomu ekspresji danego genu lub spowodować pojawienie się białka w tkance, w której wcześniej nie występowało.

W przypadku pojawienia się SNP w genie, ich efekt może być różny:

1. Może spowodować mutację synonimiczną, w przypadku gdy nowy kodon reprezentuje ten sam aminokwas co kodon wyjściowy. Powstające białko będzie identyczne z tym kodowanym przez gen niezmutowany.

2. Może spowodować zmianę niesynonimiczną i wówczas powstałe białko będzie posiadać jeden zmieniony aminokwas, co może mieć wpływ na jego funkcję.

3. W wyniku mutacji może dojść do zamiany kodonu kodującego białko na kodon STOP. Zachodzi wówczas mutacja typu nonsens, która powoduje przedwczesną terminację translacji, a powstające krótsze białko w większości przypadków będzie niefunkcjonalne.

4. Może dojść również do sytuacji odwrotnej niż w omawianym przypadku i kodon STOP zostanie zastąpiony przez kodon odpowiadający jakiemuś aminokwasowi. Powstające dłuższe białko również może nie spełniać swych fizjologicznych funkcji.

Częstość występowania SNP w genomach jest bardzo różna w zależności od gatunku. W genomie człowieka 1 SNP przypada na ok. 300 nukleotydów (19). W przypadku genomów roślinnych, jednym z bardziej polimorficznych jest genom kukurydzy, w którym 1 SNP przypada na 60–104 pz (20). U jęczmienia, 1 SNP występuje średnio co 200 pz (21), w genomie soi co 237 pz (22), u rzodkiewnika co 336 pz (23), a u pszenicy co 540 pz (24).

3.2. Metody identyfikacji SNP w kolekcjach linii genetycznych i hodowlanych

W ostatnich latach wzrasta zainteresowanie wykorzystaniem markerów typu SNP w pracach badawczych i hodowlanych dotyczących modelowych i uprawnych gatunków roślin (25). Występowanie SNP z wysoką częstością w genomach osobników danej populacji oraz fakt, że potencjalnie każdy może stać się użytkowym markerem, to tylko nieliczne z cech, które zadecydowały o ich popularności. Jej wyznacznikiem mogą być miliony zidentyfikowanych SNP w genomie człowieka, z których około 1 mln może być analizowanych jednocześnie (w jednym eksperymencie) dzięki zastosowaniu technik mikromacierzowych (26). Dostępność tak dużej liczby markerów SNP pozwala na skanowanie całego genomu z dużą dokładnością w celu poszukiwania markerów sprzężonych z cechą ilościową (27).

W porównaniu ze stopniem zaawansowania prac badawczych dotyczących człowieka lub innych kręgowców, badania z wykorzystaniem markerów SNP do analizy genomów roślinnych są jeszcze w fazie początkowej. Identyfikacja markerów typu SNP na drodze analizy sekwencji EST, bądź dostępnych w bazach danych (np. NCBI

EST; http://www.ncbi.nlm.nih.gov/dbEST) lub generowanych na potrzeby tego podejścia była wykorzystywana w pracach dotyczących zarówno roślin modelowych takich jak *A. thaliana* (23), jak i uprawnych: kukurydzy (28), jęczmienia (29) czy pomidora (30).

Technika mikromacierzy, kojarzona zwykle z analizą poziomu ekspresji genów, może również służyć do poszukiwania nowych SNP. W tym przypadku porównuje się wyniki uzyskane z hybrydyzacji cDNA lub DNA pochodzących z różnych osobników. Takie podejście było wykorzystywane w odniesieniu do *A. thaliana* (22,31,32), ryżu (33), jęczmienia (34), kukurydzy (35), a nawet w układzie heterologicznym – pomiędzy soją i fasolnikiem chińskim (36).

W odróżnieniu od dwóch poprzednich metod, obciążonych fałszywie pozytywną identyfikacją SNP (15-50%), powtórne sekwencjonowanie amplifikowanych fragmentów genów pozwala na wysoce wiarygodną identyfikację SNP [szczegóły w (25)]. Podejście takie było wykorzystywane we wspomnianych już badaniach genomu człowieka (26), a w odniesieniu do roślin uprawnych – genomu kukurydzy (zanalizowano kilka tysięcy genów (37,38); Panzea (http://www.panzea.org) i soi (ponad 4 tys. genów (39). Poza tymi gatunkami, prace z wykorzystaniem powtórnego sekwencjonowania były prowadzone w przypadku *A. thaliana* (40), ryżu (41), pomidora (42), buraka cukrowego (43) i jęczmienia (44).

Czynnikami ograniczającymi postęp w identyfikacji markerów typu SNP były do niedawna czasochłonność i koszty sekwencjonowania. Tradycyjnie stosowana metoda Sangera w połączeniu z kapilarną elektroforezą generowała wprawdzie dłuższe sekwencje, bo >800 pz, ale przy wyższych kosztach w przeliczeniu na parę zasad. Obecnie, dostępne są wcześniej wymienione trzy systemy sekwencjonowania, należące do nowej generacji. Wszystkie trzy były dotychczas stosowane głównie do powtórnego sekwencjonowania całkowicie zsekwencjonowanych genomów należących do gatunków takich jak *C. elegans* (45), mikroorganizmy morskie (46) i *A. thaliana* (13). W odniesieniu do gatunków o mniej scharakteryzowanych genomach, pojawiły się pierwsze prace dotyczące identyfikacji SNP w genomach eukaliptusa (47) i kukurydzy (48,49).

Ograniczenia metodyczne nie są jedynymi jakie stoją na drodze masowego wykorzystania SNP w analizie genomów roślin uprawnych. Większość z tych roślin to organizmy poliploidalne, zarówno allopoliploidalne takie jak rzepak (*Brassica napus*), bawełna (*Gossypium hirsutum*), tytoń (*Nicotiana tabacum*), czy pszenica (*Triticum aestivum*), jak i autopoliploidalne np. ziemniak (*Solanum tuberosum*) lub trzcina cukrowa (*Saccharum officinarum*). Identyfikacja SNP w genomach tych gatunków może być utrudniona ze względu na konieczność rozróżnienia pomiędzy markerami różnicującymi w obrębie genomów (markery przydatne) a genomami (markery nieprzydatne) (25). Przykładowo, zastosowanie metody powtórnego sekwencjonowania amplifikowanych fragmentów genowych nie sprawdziłoby się w przypadku heksaploidalnego genomu pszenicy *Triticum aestivum* L. Analiza sekwencji uzyskanych z wszystkich trzech genomów byłaby bardzo utrudniona ze względu na występujące

między nimi różnice typu indel. Rozwiązaniem stało się zastosowanie genomowo-specyficznych starterów do amplifikacji sekwencji genowych (50,51). Pomimo wymaganych wysokich nakładów pracy i zaawansowanych narzędzi bioinformatycznych, prace nad gatunkami poliploidalnymi są prowadzone z wykorzystaniem zarówno analizy ogromnej liczby sekwencji EST (52,53), powtórnego sekwencjonowania amplifikowanych fragmentów sekwencji genowych (54,55), technik mikromacierzy (24) i sekwencjonowania typu Solexa (56).

Warto też wspomnieć, że identyfikacja i opracowanie markerów SNP są prowadzone również przez firmy hodowlane; Dupont/Pionier przeprowadził sekwencjonowanie prawie 10 tys. genów w 500 liniach kukurydzy (57), a TraitGenetics wykorzystał metodę powtórnego sekwencjonowania, oraz stosował mikromacierze GoldenGate multiplex do analizy od 5 do 10 tys. genów w genomach takich roślin uprawnych jak pomidor, kukurydza, papryka i gatunków z rodzaju *Brassica* (25).

3.3. Wykorzystanie SNP w pracach genetycznych, hodowlanych i biotechnologii

Ze względu na powszechność, ewolucyjną stabilność oraz rozwój szybkich, wydajnych i stosunkowo niedrogich metod wykrywania polimorfizmów, SNP stały się atrakcyjnym systemem markerowym w badaniach genomów, analizie sprzężeń oraz ukierunkowanej selekcji pożądanych cech użytkowych za pomocą techniki MAS (ang. *Marker-Assisted Selection*) (58). Wraz z insercjami/delecjami stanowią podstawę większości różnic występujących między allelami, umożliwiając identyfikację markerów o silnych sprzężeniach z *loci* dla cech użytkowych (59,60). Wykorzystując dostępne informacje o genach kandydatach u roślin modelowych takich jak *A. thaliana* i ryż, SNP poszerzyły naszą wiedzę o genetycznych podstawach dziedziczenia ważnych cech odpowiedzialnych za wzrost, rozwój oraz obronę przed szkodnikami. Wiedza ta pozwoliła udoskonalić zespół cech u roślin uprawnych (61). Jednym z pierwszych podejść w pracach hodowlanych stało się opracowanie map genetycznych o wysokim nasyceniu markerów m.in. SNP. Obecnie dostępnych jest kilka map obejmujących tysiące markerów SNP m.in. dla kukurydzy i winorośli (szczepy Cabernet Sauvignon i Pinot Noir) (62,63). U kukurydzy wykazano wysoki stopień polimorfizmu w SNP (1/80 pz) i indelach (1/240 pz), co pozwoliło zmapować 164 z 311 *loci* (64). Dostępność takich map pozwala na określenie pokrewieństwa genetycznego między badanymi genotypami przy użyciu wybranego zestawu kilku czy kilkunastu markerów SNP.

Na podstawie ostatnich obliczeń wskazuje się, że SNP występują w genomie człowieka co najmniej raz na 300 nukleotydów. W genomie tym występuje zatem co najmniej 10 mln SNP. Ponad 4 mln SNP zostało zidentyfikowanych, a informacje o nich są publicznie dostępne dzięki pracom konsorcjum TSC i specjalistycznym platformom. Dla większości z tych 4 mln SNP ewentualne asocjacje funkcjonalne pozostają nieznane. Jednak kompilacja publicznie dostępnych SNP przez NCBI dała

w rezultacie zestaw SNP określanych jako specyficzne markery, które nazwano referencyjnymi (ang. *reference* SNP, w skrócie rsSNP). Ponad 2,6 mln SNP zostało uznanych jako rsSNP. W ostatnich pracach wskazuje się także, że 10 mln SNP wspólnych dla populacji ludzkiej nie dziedziczy się niezależnie od innych SNP; wydaje się raczej, że sąsiadujące SNP tworzą zestawy związane z odpowiednimi allelami genów; taki blok SNP nazywa się haplotypem. Cechą haplotypów jest ich przekazywanie z pokolenia na pokolenie bez rekombinacji. Obecne zawansowanie tej analizy sprawia, że haplotyp obejmujący z reguły wiele SNP, może być skutecznie identyfikowany w oparciu na zaledwie kilku wybranych SNP. Znaczącą jest także obserwacja, że sekwencje chromosomowe dwóch losowo wybranych osobników populacji ludzkiej są identyczne w 99,9% i że w obrębie różniących je sekwencji (0,1%) aż 80% stanowią SNP.

Literatura

1. Altschul S. F., Gish W., Miller W., et al., (1990), J Mol Biol., 215, 403-410.
2. Pallen M. J., Wren B. W., (2007), Nature., 449, 835-842.
3. Tettelin H., Riley D., Cattuto C., et al., (2008), Curr Opin Microbiol., 11, 472-477.
4. Jander G., Norris S. R., Rounsley S. D., et al., (2002), Plant Physiol., 129, 440-450.
5. Ziolkowski P. A., Koczyk G., Galganski L., et al., (2009), Nucleic Acids Res., 37, 3189-3201.
6. Nagano A. J., Fukazawa M., Hayashi M., et al., (2008), Plant J., 56, 1058-1065.
7. AGI, (2000), Nature, 408, 796-815.
8. Wright S. I., Agrawal N., Bureau T. E., (2003), Genome Res., 13, 1897-1903.
9. Copenhaver G. P., Nickel K., Kuromori T., et al., (1999), Science, 286, 2468-2474.
10. Clark R. M., Schweikert G., Toomajian C., et al., (2007), Science, 317, 338-342.
11. Guirouilh-Barbat J., Huck S., Bertrand P., et al., (2004), Mol Cell, 14, 611-623.
12. Zhang Z., Gerstein M., (2003), Nucleic Acids Res., 31, 5338-5348.
13. Ossowski S., Schneeberger K., Clark R. M., et al., (2008), Genome Res., 18, 2024-2033.
14. Belo A., Beatty M. K., Hondred D., et al., (2010), Theor Appl Genet., 120, 355-367.
15. Salathia N., Lee H. N., Sangster T. A., et al., (2007), Plant J., 51, 727-737.
16. Lijavetzky D., Cabezas J. A., Ibanez A., et al., (2007), BMC Genomics., 8, 424.
17. Khlestkina E. K., Salina E. A., (2006), Genetika, 42, 725-36.
18. Vignal A., Milan D., SanCristobal M., et al., (2002), Genet Sel Evol., 34, 275-305.
19. Bondi C. O., Rodriguez G., Gould G. G., et al., (2008), Neuropsychopharmacology, 33, 320-331.
20. Ching A., Caldwell K. S., Jung M., et al., (2002), BMC Genet., 3, 19.
21. Rostoks N., Mudie S., Cardle L., et al., (2005), Mol Genet Genomics, 274, 515-527.
22. Borevitz J. O., Liang D., Plouffe D., et al., (2003), Genome Res., 13, 513-523.
23. Schmid K. J., Sorensen T. R., Stracke R., et al., (2003), Genome Res., 13, 1250-1257.
24. Somers D. J., Kirkpatrick R., Moniwa M., et al., (2003), Genome, 46, 431-437.
25. Ganal M. W., Altmann T., Roder M. S., (2009), Curr Opin Plant Biol., 12, 211-217.
26. Frazer K. A., Ballinger D. G., Cox D. R., et al., (2007), Nature, 449, 851-861.
27. McCarthy M. I., Abecasis G. R., Cardon L. R., et al., (2008), Nat Rev Genet., 9, 356-369.
28. Batley J., Barker G., O'Sullivan H., et al., (2003), Plant Physiol., 132, 84-91.
29. Kota R., Rudd S., Facius A., et al., (2003), Mol Genet Genomics, 270, 24-33.
30. Yamamoto N., Tsugane T., Watanabe M., et al., (2005), Gene, 356, 127-134.
31. Borevitz J. O., Hazen S. P., Michael T. P., et al., (2007), Proc Natl Acad Sci U S A, 104, 12057-12062.
32. Singer T., Fan Y., Chang H. S., et al., (2006), PLoS Genet., 2, e144.
33. Kumar R., Qiu J., Joshi T., et al., (2007), PLoS One, 2, e284.

34. Rostoks N., Borevitz J. O., Hedley P. E., et al., (2005), Genome Biol., 6, R54.
35. Kirst M., Caldo R., Casati P., et al., (2006), Plant Biotechnol J., 4, 489-498.
36. Das S., Bhat P. R., Sudhakar C., et al., (2008), BMC Genomics, 9, 107.
37. Wright S. I., Bi I. V., Schroeder S. G., et al., (2005), Science, 308, 1310-1314.
38. Yamasaki M., Tenaillon M. I., Bi I. V., et al., (2005), Plant Cell, 17, 2859-2872.
39. Choi I. Y., Hyten D. L., Matukumalli L. K., et al., (2007), Genetics, 176, 685-696.
40. Nordborg M., Hu T. T., Ishino Y., et al., (2005), PLoS Biol., 3, e196.
41. Nasu S., Suzuki J., Ohta R., et al., (2002), DNA Res., 9, 163-171.
42. van Deynze A., Stoffel K., Buell C. R., et al., (2007), BMC Genomics, 8, 465.
43. Schneider K., Kulosa D., Soerensen T. R., et al., (2007), Theor Appl Genet., 115, 601-615.
44. Kota R., Varshney R. K., Prasad M., et al., (2008), Funct Integr Genomics, 8, 223-233.
45. Hillier L. W., Marth G. T., Quinlan A. R., et al., (2008), Nat Methods, 5, 183-188.
46. Goldberg S. M., Johnson J., Busam D., et al., (2006), Proc Natl Acad Sci U S A, 103, 11240-11245.
47. Novaes E., Drost D. R., Farmerie W. G., et al., (2008), BMC Genomics, 9, 312.
48. Barbazuk W. B., Emrich S. J., Chen H. D., et al., (2007), Plant J., 51, 910-918.
49. van Orsouw N. J., Hogers R. C., Janssen A., et al., (2007), PLoS One., 2, e1172.
50. Chang S. H., Wang K. S., Chao S. J., et al., (2009), J Hazard Mater., 166, 1127-1133.
51. Ravel C., Praud S., Murigneux A., et al., (2006), Genome. 49, 1131-1139.
52. Cordeiro G. M., Eliott F., McIntyre C. L., et al., (2006), Theor Appl Genet., 113, 331-343.
53. Tang J., Vosman B., Voorrips R. E., et al., (2006), BMC Bioinformatics, 7, 438.
54. Li L., Paulo M. J., Strahwald J., et al., (2008), Theor Appl Genet., 116, 1167-1181.
55. Simko I., Haynes K. G., Jones R. W., (2006), Genetics., 173, 2237-2245.
56. Trick M., Long Y., Meng J., et al., (2009), Plant Biotechnol J., 7, 334-346.
57. Belo A., Zheng P., Luck S., et al., (2008), Mol Genet Genomics, 279, 1-10.
58. Ye S., Dhillon S., Ke X. Y., et al., (2001), Nucleic Acids Research, 29, art. no.-e88.
59. Cho R. J., Mindrinos M., Richards D. R., et al., (1999), Nat Genet., 23, 203-207.
60. Brookes A. J., (1999), Gene, 234, 177-186.
61. Gutterson N., Zhang J. Z., (2004), Curr Opin Plant Biol., 7, 226-230.
62. Jones E., Chu W. C., Ayele M., et al., (2009), Molecular Breeding, 24, 165-176.
63. Troggio M., Malacarne G., Coppola G., et al., (2007), Genetics. 176, 2637-2650.
64. Bhatrramakki D., Ching A., Dolan M., Tingey S., Rafalski A., (2000), Maize Genet. Coop. Newslett., 74, 54.
65. Jehan T., Lakhanpalu S., (2006), Indian Journal of Biotechnology, 5, 435-459.