Sekwencjonowanie genomów i rozwój biotechnologii

Zbigniew Przybecki, Magdalena E. Pawełkowicz, Rafał Wóycicki

Katedra Genetyki Hodowli i Biotechnologii Roślin,
Wydział Ogrodnictwa i Architektury Krajobrazu,
Szkoła Główna Gospodarstwa Wiejskiego, Warszawa

Genome sequencing and progress in biotechnology

Summary

As a result of the Human Genome Project and announcement of the project: Archon X Prize for Genomics "genome for 1000$", the sequencing technology made huge progress. The purpose was to develop radically new technology that would dramatically reduce the time and cost of sequencing and resequencing eukaryotic genomes (or large region), and transcriptomes. This progress has influenced many areas also widely understood biotechnology. Owing to comparative genomic and sequenced genomes, new opportunities have occurred such as acceleration of genomic polymorphism identification, genes and genome structure prediction. It has also stimulated the development of assignation of functional effect of SNP, CNP or INDEL polymorphism and enabled functional and structural annotation of genome, establishment of a net of genome interaction (netoms) and netom's controlling. Genetic and physical mapping of genome elements methods undergo considerable alterations, thus they become critical in association methodology of genome and their function. The analysis of metabolic network and its elements is basic in planning activities in new biotechnology. Here we present the achievements in cucumber which were possible due to sequencing the genome of this species by our team.

Key words:
whole genome sequencing (WGS), next-generation DNA sequencing (NGS), genome-wide association study (GWAS), expression QTL (eQTL), synthetic biology, genome engineering, netom.

Adres do korespondencji

Zbigniew Przybecki,
Katedra Genetyki Hodowli
i Biotechnologii Roślin,
Wydział Ogrodnictwa
i Architektury Krajobrazu,
Szkoła Główna
Gospodarstwa Wiejskiego,
ul. Nowoursynowska 159,
02-776 Warszawa;
e-mail:
zbigniew_przybecki@sggw.pl

biotechnologia
4 (91) 9–23 2010

1. Wstęp

Xiaoi i wsp. (1) uważają, że od 2001 r., czyli zakończenia sekwencjonowania genomu człowieka rozpoczęły się wielkie zmiany nie tylko w badaniach biologicznych. Objęły one wiele sfer, począwszy od molekularnych po genomiczne, poprzez genetykę i odwrotną genetykę, oraz stopniowe wpływanie na styl życia, szczególnie w ochronie zdrowia. Ich zdaniem sytuacja dojrzała do tego, aby zdefiniować na nowo pojęcie genomu. Uważają także, że genom może być rozpatrywany na poziomie, molekularnym, komórkowym, osobniczym i gatunkowym, co będzie rodziło różne implikacje w biotechnologii i diagnostyce. Proponują definiować **genom** jako **„pełen skład materiału genetycznego posiadanego przez wewnątrzkomórkowe parazyty, komórkę, czy organizm**. Jednocześnie dla mikroorganizmów wprowadzono pojęcie metagenom, obejmujące wartości genomów grup mikroorganizmów zasiedlających określone siedliska, np. mikroorganizmów glebowych, mikroorganizmów sfery korzeniowej danego gatunku rośliny, metagenom mikroorganizmów zasiedlających przewód pokarmowy lub jego części itp.

Genom pełni podstawową rolę w funkcjonowaniu organizmu, i dlatego znajomość jego pełnej sekwencji wpływa w decydujący sposób na rozwój wielu dziedzin nauki i biogospodarki jak: biologia systemów (integruje wiedzę omiczną), biologia syntetyczna (projektowanie i tworzenie sztucznych systemów biologicznych opartych na wiedzy z biologii systemów), czy biotechnologia.

Według Zespołu ds. Biogospodarki przy Ministrze Nauki i Szkolnictwa Wyższego (2007 r.); http://www.nauka.gov.pl/mn/_gAllery/33/69/33696/20071224_Raport_Biogospodarka.pdf biotechnologia definiowana jest jako **interdyscyplinarna dziedzina nauki i techniki, która zajmuje się zmianą materii żywej i nieożywionej poprzez zastosowanie metod naukowych i technologii z wykorzystaniem organizmów żywych, ich części bądź pochodzących od nich produktów i ich modeli z użyciem technik DNA/RNA, białek i innych cząsteczek, komórek, kultur komórkowych i inżynierii komórkowej, genowych i wektorów RNA, bioinformatyki, nanobiotechnologii, w celu tworzenia wiedzy, dóbr i usług"**. W definicji tej, jak się wydaje, mieści się także biologia syntetyczna.

2. Przedmiot badań biotechnologii

Z przytoczonej definicji wynika, że organizmy lub ich części, są głównymi obiektami zainteresowań biotechnologii. Wiemy także, że charakteryzują się one szeregiem właściwości (cech), będących przedmiotem zainteresowania biotechnologa, w jego dążeniu do uzyskania określonego celu utylitarnego, lub naukowego.

Każda właściwość organizmu jest wynikiem działania określonej sieci metabolicznej, złożonej ze szlaków metabolicznych i sterowanej określonym zbiorem genów i ich produktów. Zbiór ten, dla większej precyzji przekazu można określić jako

netom (ang. *network* – sieć), co powoduje wyodrębnienie terminu **netomika**. Netomy charakteryzują się różnym stopniem złożoności, wobec tego mogą składać się z netomów różnych rzędów, określanych również modułami. Złożoność netomu zależna będzie od kompleksowości cechy, którą tworzy. Można uznać wobec tego, że organizm działa dzięki funkcjonowaniu skomplikowanego, dobrze zorganizowanego netomu organizmalnego, złożonego z olbrzymiej liczby netomów niższych rzędów. Najwięcej sieci metabolicznych opracowano dla prostych organizmów, głównie bakterii i grzybów (2-4). Natomiast przykładem złożonego netomu u roślin jest sieć metabolizmu podstawowego (primary) u *Arabidopsis thaliana* (5) skonstruowana w wyniku analizy wykorzystującej modelową sieć metaboliczną AraGEM. Do rekonstrukcji sieci wykorzystano około 10 tys. elementów, w tym unikatowych dla kompartmentów komórkowych. Model metabolizmu obejmuje m.in. fotorespirację, fotosyntezę zróżnicowane dla komórek fotosyntetyzujących i niefotosyntetyzujących procesy redoks. Autorzy podkreslają rolę sekwencjonowania genomu w tych pracach. Złożonym netomem (uwzględniającym również interaktom) może być sieć procesu nowotworowego raka mózgu u człowieka glioblastoma (GBM) (6). Opracowując nowe podejście określane jako „automatyczna analiza sieciowa" (pakiet NetBox) oparte na analizie topologicznej, wyodrębniono 10 modułów tej sieci. Wśród nich wyróżniono 2 moduły stałe (core) czyli występujące we wszystkich odmianach (zawierające m.in. szlaki sygnalne) tego nowotworu i zmienne – u jednych pacjentów występujące u innych nie (specyficzne osobniczo). W tej analizie wykorzystano mutacje, CNP zunifikowaną sieć interakcji molekularnych (oddziaływania białko-białko i ścieżki przesyłu sygnału). Tą metodą rozbudowywany będzie netom GBM, w tym również dla dalszych wariantów tego nowotworu.

Działalność biotechnologa, czy hodowcy, polega bądź na przewidywaniu występowania określonej właściwości wynikającego z zestawu genów bądź też na dokonywaniu określonych zmian cech organizmu lub komórki. Wymaga to wiedzy o zależnościach w obrębie netomu i między netomami. Ingerencja w działanie netomów może mieć charakter trwały lub czasowy. Ten ostatni wywoływany jest czynnikami pozagenetycznymi (jednak sposób reakcji na nie jest zależny od stuktury genetycznej), natomiast trwałe zmiany w netomie są efektem zmian w części genomu kontrolującego netom.

Określone zmiany w netomach, można dokonywać w dwojaki sposób: 1) wprowadzając zmiany w sposób losowy – stosuje się w przypadku braku wiedzy na temat działania netomu (podłoża molekularnego cechy) i/lub brak możliwości wprowadzania celowych zmian z powodu braku odpowiednich metod; 2) wprowadzając zmiany w zdefiniowanym miejscu co gwarantuje zamierzony efekt i jest stosowane, kiedy wiadomo jak system (netom) działa i istnieją metody wprowadzania zmian. Tak zatem, w zależności od nagromadzonej wiedzy o netomie i dostępności metod ingerencji w jego działanie można uzyskiwać efekty ściśle zaplanowane, co przekłada się na skuteczność działania.

Wiedzę o netomach i ich modyfikacjach, zdobywa się w największym stopniu przez wielkoskalowe sekwencjonowania (WGS, ang. *Whole Genome Sequencing*) geno-

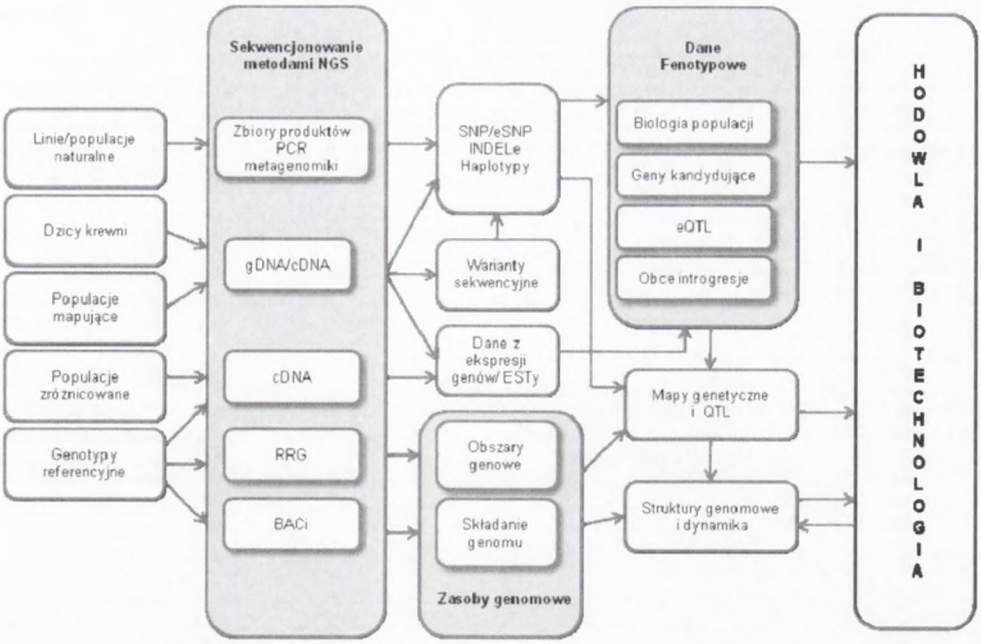
mów (referencyjnego i resekwencjonowanych) i transkryptomów. Inne sposoby (tab.), dają obraz wycinkowy i ich znaczenie będzie marginalizowane w tym sensie, że będą używane do weryfikacji w określonych sytuacjach. Postęp w technologii i obniżenie kosztów sekwencjonowania WGS, ogromnie przyspieszy postęp zastępując wiele istniejących metod i technik, lub modyfikując je poprzez włączenie WGS. Samo zsekwencjonowanie i składanie średniej wielkości genomu można zrealizować w warunkach średniej wielkości zespołu badawczego z budżetem około 3 mln zł (7) dobrze przygotowanego od strony bioinformatycznej, dysponującego komputerem o 48 GB RAM oraz komputerami mniejszej mocy. Wielkie konsorcja międzynarodowe nie są potrzebne do samego sekwencjonowania i składania, natomiast współpraca z różnymi specjalistami (fizjomika, genomika, metabolomika) jest niezbędna do pełnego wykorzystania uzyskanych danych.

3. Genomika porównawcza – użycie w hodowli i biotechnologii

Genomika porównawcza polega na charakteryzowaniu bioinformatycznym przez porównanie sekwencji całych genomów, lub dużych jego obszarów, oraz porównaniu sekwencji genomowych z innymi zbiorami sekwencji, np. transkryptomów, EST-ów i in. Powstała z chwilą zsekwencjonowania całych genomów i dużych zbiorów sekwencji oraz ich zgromadzenia w bazach danych (8). O jej rozwoju decyduje postęp w bioinformatyce oraz dostępność nowych metod sekwencjonowania szczególnie tzw. *Next-Generation-Sequencing* (NGS). Znaczenie NGS dla genetyki, hodowli roślin i biotechnologii (tab.; rys. 1) jest wielorakie, ale głównie polega na: dostarczaniu zasobów genomowych (sekwencji całych genomów, lub ich elementów, transkryptomów, proteomów i in.), dostarczaniu markerów i mapowaniu QTL-i, identyfikacji introgresji genów w wyniku oddalonych krzyżowań, ułatwieniu analizy ekspresji oraz wykorzystaniu możliwości mapowania asocjacyjnego. Na przykład sekwencjonowanie genomowego DNA obejmującego BACi (np. BES, ang. *Bac End Sequence*), redukuje reprezentacje genomu (RRG), lub sekwencjonowanie cDNA z genotypów referencyjnych używając technologii NGS może dostarczyć genomowych zasobów (sekwencje genomowe) takich jak EST, elementy genowe i obszary genomowe.

Zasoby genowe takie jak: obszary wysycone genami (ang. *gene spaces* – długie obszary bogate w geny, zjawisko powszechne u roślin – (9)) z poskładanych fragmentów genomu (ang. *genome assembly* – kontigi, superkontigi) mają bezpośredni wpływ na zrozumienie architektury genomu przydatnej dla genetyki hodowli i biotechnologii ważnych gospodarczo obiektów.

Innym zastosowaniem NGS jest genotypowanie form rodzicielskich populacji mapujących, lub spokrewnionych z innymi formami nie uprawnymi. Przyspiesza to konstruowanie markerów molekularnych, np. SSR-ów i SNP-ów, które są bardzo przydatne do konstrukcji map genetycznych, identyfikacji QTL-i i wykrywania fragmentów pochodzących z introgresji. Markery sprzężone z QTL-ami mogą być użyte



Rys. 1. Przegląd zastosowań sekwencjonowania NGS w genetyce i hodowli/biotechnologii wg (8). Opis w tekście.

w procesie selekcji wspomaganej markerami (MAS) do selekcji potomstwa zawierającego korzystne allele. Dla uzyskania markerów funkcjonalnych lub genowych (ang. *gene-based* – stanowiące fragment genu wykazującego sprzężenie z markerem), sekwencjonowanie NGS cDNA genotypów kontrastowych (zawierających cechy alleliczne) ze względu na interesujące cechy mogą być użyte do identyfikacji genów kandydackich determinujących cechę lub zasocjowanych z nią. Mapowanie ekspresji genów (jakość i poziom produktów ekspresji) kandydackich, razem z fenotypowaniem populacji segregujących, uzyskanych z genotypów kontrastowych dostarcza QTL-i ekspresyjnych (eQTL-i) i markerów zasocjowanych z tymi eQTL-ami (np. eSNP-y). Są one przydatne jako funkcjonalne (perfect) markery do MAS. Innym ważnym obszarem zastosowania NGS jest genetyka asocjacyjna i biologia populacji, gdzie z setek osobników można sekwencjonować zbiory amplikonów PCR tysięcy genów kandydatów. Dane sekwencyjne mogą zostać użyte do identyfikacji SNP-ów lub haplotypów genów lub genomów w genetyce asocjacyjnej i biologii populacji (8).

Tabela

Zmiany w metodyce niektórych badań z udziałem sekwencji genomu i WGS

Rodzaj analizy	Wykonanie bez WGS	Wykonanie z użyciem WGS	Efekt użycia WGS
izolacja genów	standardowe metody klonowania funkcjonalnego i pozycyjnego (*mape-based*)	identyfikacja funkcjonalna i pozycyjna (jeżeli genom jest na chromosomach) genów *in silico* w adnotowanym strukturalnie i funkcjonalnie genomie	wyszukanie *in silico* dowolnej sekwencji kandydackiej
genotypowanie – zmienność DNA genomowego	wszystkie metody wyszukiwania markerów molekularnych w tym subtrakcyjne	subtrakcja *in silico* (analiza porównawcza między genomami; między genomami a transkryptomami lub zbiorami EST i in.)	cały polimorfizm SNP i INDEL porównywanych genomów w ciągu kilku godzin w złożonej części genomu i zmapowany – jeżeli genom przypisany został do chromosomów
mapowanie genetyczne (dystans w jednostkach względnych wyrażających wielkość sprzężenia)	tradycyjne metody mapowania, z udziałem tradycyjnych markerów	morganowskie* (często określane jako rekombinacyjne) i asocjacyjne	zmapowany cały polimorfizm w złożonej części genomu
mapowanie fizyczne (dystans i położenie w jednostkach fizycznych)	standardowe mapowania cytogenetyczne; mapy klonalne	*in silico*	szczegółowe mapy fizyczne adnotowanego genomu i zasocjowanych cech
GWAS (ang. *Genome-Wide Association Study*)	mapowanie asocjacyjne przy użyciu tradycyjnych markerów i mikromacierzy	mapowanie asocjacyjne z użyciem genomu referencyjnego i całych lub częściowo resekwencjonowanych genomów populacji mapujących	przegląd całej zmienności genomu i określanie jej związków z cechami
mapowanie zbiorów genów sterujących ekspresją złożonych cech	standardowe metody (hybrydyzacja northerna, QT-RT-PCR; mikromacierze i in.) ilościowego i jakościowego wyznaczania ekspresji genów (transkryptom, proteom) i mapowanie asocjacyjne, lub morganowskie	ilościowa i jakościowa analiza porównawcza sekwencji całych genomów i transkryptomów, białek	eQTL (*expression QTLs*); eSNP (*expression SNPs*); kontrola ekspresji genów; tworzenie netomów

*mapowanie morganowskie – jest mapowaniem wg zasad podanych przez Morgana –jego twórcę, często nazywane jest rekombinacyjnym. Jednak mapowanie asocjacyjne jest również mapowaniem genetycznym i u jego podstaw również leży zjawisko rekombinacji genetycznej. Dlatego dla wyraźnego rozróżnienia tych mapowań pierwsze z nich określono tutaj jako morganowskie.

W tabeli pokazano w jaki sposób sekwencjonowanie genomów może zmieniać tradycyjne metody rozwiązywania problemów i jaki jest efekt takiej zmiany.

1. Izolowanie genów „wspomagane genomem" polega na wyszukaniu *in silico* w sekwencji adnotowanego genomu sekwencji homologicznych do sekwencji zidentyfikowanej funkcjonalnie (sekwencja warunkująca analogiczną funkcje w innym

gatunku, sekwencja uzyskana w wyniku odczytu sekwencji aminokwasów, sekwencja uzyskana z transkryptu i in.), lub homologicznych do markera(ów) sprzężonego z poszukiwanym genem i identyfikacji genów kandydackich w przewidzianej przez mapowanie odległości od sekwencji markerowej (odszukiwanie na podstawie mapy lub poszukiwanie pozycyjne). Jeżeli genom wcześniej nie był adnotowany, to trzeba zrobić adnotacje przynajmniej przeszukiwanego regionu. Jeżeli genom nie jest przeniesiony jeszcze na chromosomy, lub jeżeli kontig zawierający marker nie wchodzi w skład dostatecznie wielkiego metakontigu, poszukiwanie pozycyjne może się nie powieść.

2. Genotypowanie tradycyjnie oparte jest na wyszukiwaniu markerów standardowymi metodami takim jak RFLP, RAPD, FELP, subtrakcja, hybrydyzacja różnicowa, metody mikroukładowe jak DArT i SNP-y bez sekwencji genomu i in. Każda z tych metod wyławia tylko niewielką część zmienności genomu, przy często dużych nakładach. Natomiast poprzez subtrakcję *in silico* zsekwencjonowanych genomów można wyłowić całą zmienność (SNP-y, INDEL-e, CNP-y, rearanżacje) w złożonej części genomu. Wobec tego wielkość wyszukanej zmienności, zależy od stopnia złożenia i ułożenia na chromosomach genomu. Jeżeli złożona część genomu jest przypisana odpowiednim miejscom na chromosomach to tym samym wyszukane zmiany są zmapowane. Takie genotypowanie *in silico* w porównaniu z laboratoryjnym („mokrym") jest tanie i krótkie.

3. Zaangażowanie WGS w mapowaniach genetycznych, będzie rosło wraz ze spadkiem ceny i wzrostem szybkości sekwencjonowań NGS. W tej chwili najbliższy „ideału" jest zapowiadany system HANS. Znacznie przebija nawet kryteria konkursu „genom za 1000 USD", w tym cenowe ponad 10x. http://nabsys.com/. W najbardziej zaawansowanej formie będzie to podanie pełnych sekwencji wszystkich genotypów biorących udział w mapowaniu, jak również stworzenie pełnej mapy genetycznej SNP-ów i INDEL-i oraz sprzężonych z nimi cech z wytypowaniem genów kandydackich dla tych cech (w tym obszarów QTL-i), w jednym doświadczeniu, w ciągu kilku godzin. Z finansowego punktu widzenia już obecnie możliwe jest resekwencjonowanie populacji mapujących, np technologią Illuminy, przez pojedyncze zespoły badawcze przy większych projektach. W takim przypadku do mapowania nie ma potrzeby zamiany SNP-ów INDEL-i i innych rearanżacji w standartowe markery, np. typu SCAR.

W przypadku braku resekwencjonowania populacji mapującej, konieczne jest wyszukanie zmiany w genomie, projektowanie różnego typu markerów standartowych (SCAR-y, SSR-y i in.) i ich mapowanie.

4. Mapowanie fizyczne w zsekwencjonowanym genomie polega na wyszukaniu *in silico* lokalizowanej w genomie sekwencji i nadaniu jej adresu, który może zawierać szereg danych. W zależności od stopnia złożenia genomu, położenie sekwencji może być podane z dokładnością do nukleotydu na kontigach, superkontigach, metakontigach i chromosomach. Adnotacja genomu pozwala na zmapowanie fizyczne wszystkich genów w genomie.

5. GWAS (ang. Genome-Wide Association Study) jest analizą funkcjonalną polegającą na wykorzystaniu zmienności genomowej do szukania funkcji genów lub regionów genomu (QTL-e) w tworzeniu fenotypu poprzez mapowanie głównie asocjacyjne. Liczba cech zasocjowanych w eksperymencie z określonymi genami lub obszarami genomu zależy od wielkości zmienności znalezionej w DNA genomowym (szczególnie tzw. funkcjonalnej, czyli zlokalizowanej w genach) i liczby polimorficznych cech w populacji mapującej. Jeżeli można sobie pozwolić na resekwencjonowanie populacji mapującej, to teoretycznie można zidentyfikować cały polimorfizm DNA tej populacji. W takim przypadku jako populacją mapującą najlepiej jest dysponować kolekcjami odległych linii, co daje dużą zmienność i dostateczną nierównowagę cech allelicznych umożliwiającą asocjację zmienności DNA i fenotypowej.

6. Mapowanie eQTL-i przedstawiono szczegółowiej na rysunku 2 i opisano w tekście.

4. Wielkoskalowa analiza funkcjonalna

Podstawą działań biotechnologicznych jest wiedza o efektach działania netomów i regulacji ich funkcjonowania. Konstrukcja sieci metabolicznych najprostszych nawet cech czy struktur, wymaga przypisania olbrzymiej liczbie elementów jeszcze większej liczby funkcji i interakcji między tymi elementami. Dlatego potrzebne są metody wielkoskalowej analizy funkcji. Do pewnego stopnia możliwe jest to metodami analizy mikroukładowej. Jednak do istotnego zwiększenia skali analiz funkcjonalnych potrzebne są dane sekwencyjne wielkich zbiorów sekwencji, głównie genomicznych i transkryptomicznych.

Do zdobycia wiedzy funktomicznej konieczna jest ponadto zmienność genetyczna (polimorfizm w DNA genomowym) i umiejętność znalezienia efektu funkcjonalnego tegoż polimorfizmu DNA (GWAS, ang. *Genome-Wide Association Study* – analiza eQTL).

4.1. Efekt funkcjonalny polimorfizmu DNA

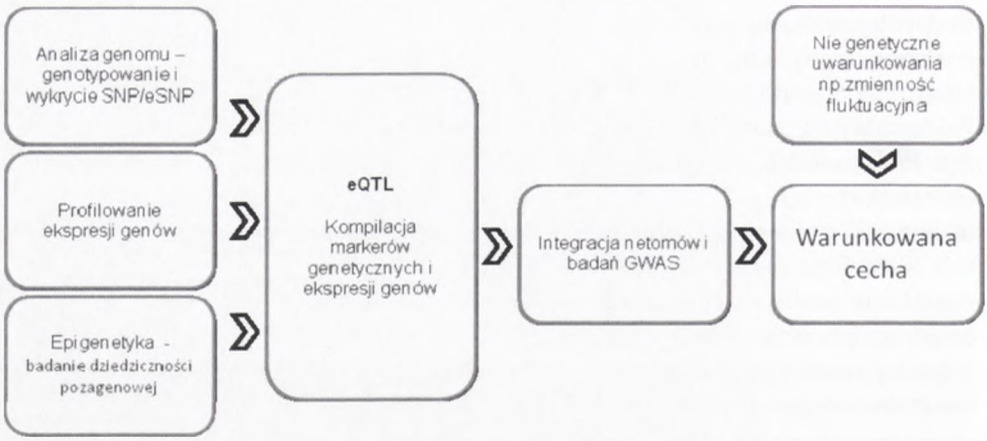
Jednoczesna analiza reguł funkcjonowania wielu sieci wymaga dwojakiego rodzaju działań. Jedne to mapowania w tym genetyczne QTL-i, najlepiej wysokorozdzielcze i identyfikujące funkcjonalnie QTL-e (10). Natomiast drugie są analizami typu GWAS i obrazowania regulacji ilościowej ekspresji genów – eQTl-i (11-15).

Analiza GWAS polega na identyfikacji polimorfizmu (głównie SNP-ów) w skali genomowej i jego przypisaniu (asocjacji) określonym cechom (13). Jeśli dla danego gatunku istnieje platforma SNP-ów (mikroukład SNP), wtedy wystarczy analiza z mikromacierzy SNP-ów osobników populacji mapującej, względnie innej populacji dostatecznie licznej i z dostatecznie dużą nierównowagą cech allelicznych. Jeżeli natomiast brak takiej platformy, wtedy pierwszą część (znalezienie polimorfizmu) moż-

na łatwiej wykonać dysponując zsekwencjonowanym genomem. Najlepiej jednak dokonać zresekwencjonowania wszystkich genomów użytych do szukania polimorfizmu. Ten sposób wyszukania polimorfizmu ilustruje przykład opisany w rozdziale 7 dotyczącym genomiki ogórka. Znaleziony polimorfizm SNP-ów (szczególnie cenny jest funkcjonalny), ewentualnie też INDEL-i jest wykorzystany do przypisania mu określonych cech w klasycznym mapowaniu morganowskim lub asocjacyjnym. Do mapowania polimorfizm SNP-ów lub INDEL-i wymaga przekonwertowania (*in silico*) w polimorfizm markerów sekwencyjnych np. SCAR-ów. Po weryfikacji markerów z większej liczby resekwencjonowań genomów reprezentujących polimorfizm gatunku można użyć je do konstrukcji platformy zmienności służącej do identyfikacji polimorfizmów w populacjach i w mapowaniach. Jeśli natomiast platforma nie jest tworzona, wygenerowane markery można użyć do zwykłego mapowania SCAR-ów, morganowskiego lub asocjacyjnego, które może być również wydajne jeżeli do izolacji DNA i składania mieszanin reakcyjnych PCR użyje się robotów.

W najbliższym czasie jest zapowiadany skokowy rozwój metodyki sekwencjonowania genomów. Wyrazem tego postępu jest metoda HANS (*Hybrydyzation Assisted Nanopore Sequencing* (16) – http://nabsys.com/), której twórcy zapowiadają koszt sekwencji wielkości genomu człowieka < 100 USD, czas sekwencjonowania i składania genomu – 1 godzina, w urządzeniu wielkością porównywalną do cyklera PCR. Pozwoli to na znaczne uproszczenie procedur mapowania i polegać będzie na zsekwencjonowaniu wszystkich osobników populacji mapującej, ich analizie porównawczej i zmapowaniu genetycznym i fizycznym polimorfizmu DNA i cech fenotypu. W ten sposób w jednym postępowaniu do genomu przypisane zostają także określone efekty fenotypowe (funkcjonalny efekt polimorfizmu DNA). Byłoby to w szczególności pomocne przy wyszukiwaniu zbiorów genów regulujących nazywanych eQTL-ami (ang. expression QTLs, wg (13).

Określenie eQTL związane jest z eSNP-em. SNP jest nazywany eSNP-em, wtedy gdy związany jest z regulacją genu (jednego lub więcej). Inaczej mówiąc eSNP jest SNP-em znajdującym się w genie wchodzącym w skład eQTL-a, czyli zbioru genów regulujących ekspresję (poziom produktu) określonego genu. Jest więc SNP-em funkcjonalym. Rolę regulacyjną określonemu genowi (przynależność do eQTL-a) w regulacji poziomu ekspresji innego genu (produktu – najczęściej transkryptu, lub białka) przypisuje się na podstawie danych mapowania (asocjacyjne) mówiących, czy istnieje związek (sprzężenie) między zmiennością genu z eQTL-a (eSNP-y w genie regulatorze) a poziomami produktu badanego genu w ontogenezie osobnika(ów). Powiedzmy, że ekspresja genu A jest regulowana przez eQTL, który jest zbiorem genów B, C, D,....., z których każdy ma formy alleliczne spowodowane występowaniem w nich eSNP-ów. Żeby wykazać regulacyjną funkcję dowolnego genu tego eQTL-a dla genu A, trzeba wykazać istnienie sprzężenia między allelami tego genu eQTL-a a poziomem produktu genu A w różnych tkankach i różnych momentach rozwoju. Na rysunku 2 przedstawiono schemat tworzenia eQTL-i i ich wykorzystania.



Rys. 2. Mapowanie eQTL-i. Mapowanie ekspresyjnych QTL-i (eQTL) zaczyna się pomiarem ekspresji genu w komórce lub tkance z wielu osobników. Informacja ta jest podstawą badań efektów polimorfizmu DNA (wszystkich typów) na ekspresję poszczególnych genów. Inne czynniki, które mogą zmieniać transkrypcję, takie jak epigenetyczna metylacja CpG, mogą być również mapowane. Analiza sieciowa opiera się na mocnych korelacjach istniejących między transkryptami i pozwala identyfikować moduły genowe, które pośredniczą w tworzeniu złożonych funkcji. Ta informacja może następnie służyć do interpretacji genetycznych asocjacji i informacji mapowych z badań złożonych cech lub chorób, wg (13).

Możliwość określenia wielkości eQTL-a (liczba wykrytych genów regulujących ekspresję genu) regulującego działanie genu w całym organizmie, lub jego części zależy od liczby wykrytych SNP-ów i zmian epigenetycznych z jednej strony oraz liczby przebadanych transkryptomów z drugiej. Potrzebne dane transkryptomiczne uzyskuje się z analiz *Real Time* RT-PCR, analiz mikromacierzowych i sekwencjonowania całych transkryptomów. Najlepsze i najwięcej danych ilościowych i jakościowych uzyskuje się z tych ostatnich analiz (cena sekwencjonowania wynosi kilka tysięcy USD za transkryptom, w zależności od pokrycia) (17,18). Asocjacja tych dwóch zbiorów danych polega głównie na badaniu sprzężeń między ich elementami i pozwala ustalić, który SNP jest eSNP-em, danego genu, a następnie pogrupować je w eQTL-e. Jeśli położenia na mapie eQTL i QTL dowolnej cechy pokrywają się to może oznaczać, że eQTL steruje również ekspresją cechy warunkowanej przez jej QTL (13); (19,20). Wyznaczanie eQTL-i regulujących określoną cechę umożliwia budowanie sieci metabolicznej dla niej. Dostępność do systematycznie gromadzonych eQTL-i pozwala na wgląd w podstawy biologiczne powiązań uzyskanych w badaniach GWAS i pomaga ustalić obraz netomu badanej cechy.

5. Inżynieria genomowa

Dysponowanie wiedzą o zależnościach występujących w funkcjonowaniu całych genomów otwiera możliwości nadawania nowych właściwości organizmom na drodze inżynierii genomowej. Inżynieria genomowa jest definiowana jako obszerne i zamierzone modyfikacje systemu replikującego się (genomów?), mające prowadzić do realizacji określonych celów (21) (poznawczych lub gospodarczych). Inaczej mówiąc jest to inżynieria genetyczna na skalę genomową.

Podstawą jej istnienia jest zsekwencjonowany genom i możliwości sekwencjonowań po rearanżacjach genomowych (22). Synteza *de novo* genomu bakteryjnego z zaplanowanymi licznymi rearanżacjami była możliwa tylko dzięki znajomości genomu (23). Wraz z dużą skalą syntezy *de novo* DNA tradycyjne techniki ingerowania, jak proste modyfikacje genów, klonowanie, mutageneza, zastąpione zostają przez nowoczesne projektowanie, automatyczną syntezę i charakterystykę (21). Przykładami prac inżynieryjnych w skali genomowej oprócz wspomnianej pracy (23) mogą być prace obejmujące duże obszary genomów (przekraczające często 100 kb, manipulacja, która wymagała użycia metod opartych na rekombinacjach *in vivo*) dostosowujące genomy do specyficznych celów jak: przeniesienie większości genomu archea do innego prokariotycznego genomu (24); wprowadzenie dużej liczby zaplanowanych zmian w jednym genomie (25); delecja wielu dużych segmentów genomu *E. coli* w celu eliminacji niestabilnych fragmentów (26); opracowanie systemu translacji do testowania nowych funkcji, bez wprowadzania genów do komórek (27); rozłożenie genomu faga T7 w rekonfigurowalne moduły (28).

6. Biologia syntetyczna

Biologię syntetyczną można określić, jako połączenie biologii molekularnej i inżynierii do **kreowania** i doskonalenia urządzeń (ang. *devices*) genetycznych i małych molekuł, które przez te urządzenia są konstruowane (29,30). Jest to zatem tworzenie systemów (szlaków metabolicznych i sieci – netomów) możliwych do programowania i przewidywania w działaniu. Biologia syntetyczna wymaga współdziałania badaczy z wielu dziedzin: biologii molekularnej, chemii, chemii fizycznej, biofizyki, inżynierii, modelowania matematycznego, bioinformatyki i innych. Nie jest to klasyczna inżynieria genetyczna, gdzie rozwiązanie problemu polega na skoncentrowaniu się na jednym lub kilku genach (29). Więcej szczegółów na temat metod biologii syntetycznej zob. m.in. (21,29-32). Urządzenie takie może być wprowadzone do organizmu i uruchomione tylko, wtedy gdy zostanie w najdrobniejszych szczegółach zapisane w genomie. Oznacza to możliwość całkowitego przewidywania skutków jego wprowadzenia i kontrolę nad nim. Wprowadzany układ przebudowuje metabolizm. Skonstruowano układy, które mogą wykonywać obliczenia (33). Może to wymagać znacznej i bardzo precyzyjnej przebudowy, a niekiedy budowy od

nowa genomu – określanej jako „inżynieria genomowa" (21-23). Opracowano szybką zautomatyzowaną metodę *in vivo* kierowanej ewolucji szlaków metabolicznych (25). Nazwano ją multipleksową automatyczną inżynierią genomową MAGE (ang. *Multiplex Automated Genome Engineering*). Zdolność tworzenia takich urządzeń powoduje całkowicie nowe podejście do wielu aplikacji np. w bioremediacji, produkcji energii, czy medycynie (29).

Biologię syntetyczną można uznać za najbardziej zaawansowany technologicznie dział biotechnologii, o olbrzymim, ale trudnym do przewidzenia wpływie w przyszłości na nasze życie.

7. Genomika ogórka po zsekwencjonowaniu genomu

Genom ogórka został zsekwencjonowany dla trzech różnych genotypów przez trzy grupy badawcze: Konsorcjum Azjatyckie – zsekwencjonowało ogórek azjatycki, odmianę długoowocową „Chinese Long" linia 9930 (34), i udostępniło sekwencję w NCBI); Polskie **Konsorcjum Sekwencjonowania Genomu Jądrowego Ogórka** (http://csgenome.sggw.pl/) zsekwencjonowało wysoce wsobną (>20 pokoleń wsobnych) jednopienną linię B10 odmiany „Borszczagowski" (35,36) – sekwencja znajduje się od września 2009 r. w NCBI, ale jest niedostępna. Żeńska linia Gy14 została zsekwencjonowana przez zespół Y. Weng'a w University of Wisconsin (37).

Katedra Genetyki Hodowli i Biotechnologii Roślin SGGW ma półwieczne tradycje intensywnych badań nad ogórkiem, i w tym czasie osiągnięto wiele liczących się w Polsce i na świecie wyników, zarówno naukowych, jak i wdrożeniowych (7). Przy szybkim postępie w metodyce badań molekularnych relatywnie efektywność naszych badań nad tym obiektem malała. Motywami, które skłoniły nas do zsekwencjonowania genomu ogórka *de novo* były: możliwość radykalnego zwiększenia efektywności wszystkich naszych badań związanych z ogórkiem spowodowana faktem posiadania sekwencji genomu własnego obieku badań, i postęp w metodyce sekwencjonowania genomów czyniący takie sekwencjonowanie dostępnym przy naszych możliwościach.

Sposób w jaki sekwencja genomu wpłynęła na postęp w badaniach nad ogórkiem ilustrują niektóre wyniki: 1) oszacowanie udziału różnego typu sekwencji w genomie, np. sekwencje mikrosatelitarne, roślinne elementy powtórzone, i in.; 2) adnotacja strukturalna (wyszukanie w genomie struktur genopodobnych, czyli potencjalnie mogących być genami – CDS-y z niektórymi charakterystykami, np. liczba CDS-ów w genomie, średnia liczba egzonów i intronów w genie, średnia długość egzonu i intronu i in. oraz funkcjonalna (przypisanie znalezionym CDS-om – a właściwie produktom ich ekspresji, wszystkich możliwych funkcji) całego genomu; 3) znalezienie sekwencji kandydatów dwóch genów płci; 4) lokalizacja na mapie fizycznej chromosomów czterech genów płci w tym gen M/m dotąd nie zmapowany, nawet genetycznie i z których trzy znajdują się w stosunkowo małym obszarze, co

sugeruje istnienie w genomie obszarów (klastrów) o większym zagęszczeniu czynników rozwoju płci i kwiatów. Sugeruje to możliwość rozwoju chromosomów płci; 5) wykazano istnienie zmienności proporcji w promotorach genów trzech elementów CRE (DRE, ABRE i ERE) w genomach pięciu gatunków, zależnie od warunków siedliska w których się kształtowały. Sugeruje to Darwinowską selekcję w odniesieniu do tych elementów (36,38); 6) wykonano analizę porównawczą genomów B10 i 9930, w której znaleziono 540 tys. SNP-ów; 7) zintegrowano z genomem ponad 28 000 (z ~34 000) klonów biblioteki BAC, co oznacza praktycznie wyszukiwanie w bibliotece dowolnej sekwencji bez konieczności przeglądu biblioteki przez hybrydyzację lub PCR.

W wyniku zsekwencjonowania genomu, powstała platforma do wyszukiwania genów i markerów istotnych cech hodowlanych; uzyskano źródło danych dla genomiki porównawczej (genom referencyjny do porównania z resekwencjonowanymi liniami, zsekwencjonowanymi transkryptomami i innymi zbiorami sekwencji) służącej badaniom podstawowym jak i aplikacyjnym w hodowli molekularnej; uzyskano dokładną lokalizację interesujących genów dzięki mapowaniu *in silico*; możliwe jest szybkie wyszukiwanie różnic sekwencyjnych, określanie ich funkcji, wyszukiwanie genów głównych dla danej cechy na podstawie położenia markerów w ścieżkach metabolicznych i/lub regulacji ekspresji genów; analiza funkcji genów cech złożonych, może być przyśpieszona poprzez amplifikacje genów głównych poprzez PCR i np. transformację; istnieje możliwość zastąpienia „mokrych" metod laboratoryjnych odwrotnej genetyki metodami *in silico* (resekwencjonowanie i genomika, transktyptomika, metabolomika, netomika porównawcza). Tak odbywa się np. izolacja genów i znalezienie klonów BAC je zawierających. W tej sytuacji trudno mówić o izolacji genów, jest to raczej ich odszukiwanie w sekwencji genomu; od resekwencjonowania genomu do wyszukania interesujących nas genów może minąć zaledwie godzina przy wykorzystaniu ścieżki przetwarzania danych złożonej z programów do składania genomu, adnotacji strukturalnej, funkcjonalnej, tworzenia ścieżek metabolicznych, regulacji ekspresji genów, genomiki, transkryptomiki porównawczej itp.

Objaśnienia skrótów:

BES-BAC, ang. *Ends Sequence* – zsekwencjonowane końce klonów BAC

CDS, ang. *Coding DNA Seguence* – część kodująca genu

CNP, ang. *Copy Number Polymorphism* – polimorfizm liczby kopii sekwencji

eQTL, ang. *expresion QTL* – zbiór genów sterujących ekspresją genu

eSNP – ang. *expresion SNP* – SNP z genu ulegającego ekspresji

GWAS, ang. *Genome-Wide Association Study* – asocjacja zmienności genomowej i cech

HANS, ang. *Hybrydyzation Assisted Nanopore Sequencing* – sekwencjonowanie przez nanopory wspomagane hybrydyzacją

INDEL, ang. *Insertion/Deletion Polymorphism* – polimorfizm insercji i delecji nukleotydów

MAGE, ang. *Multiplex Automated Genome Engineering* – automatyczna *in vivo* kierowana ewolucja szlaków metabolicznych

MAS, ang. *Marker Assisted Selection* – selekcja wspomagana markerami
NGS, ang. *Next-Generation DNA Sequencing* – metody sekwencjonowania nowej generacji
SNP, ang. *Single Nucleotide Polymorphism* – polimorfizm pojedynczego nukleotydu
WGS, ang. *Whole Genome Sequencing* – sekwencjonowanie całego genomu jednocześnie

Literatura

1. Xiao L. Saldivar J.-S., Zhou C., Chen C., Zhang J., Sirois P., Li K., (2009), Mol. Biotechnol., 41, 152-156.
2. Andersen M. R., Nielsen M. L., Nielsen J., (2008), Molecular Systems Biology, 4, 1-13.
3. Feist A. M., Palsson B. Ø., (2008) Nature Biotechnology, 26 (6), 659-668.
4. Feist A. M., Palsson B. O., (2010) Current Opinion in Microbiology, 13, 344-349.
5. Gomes de Oliveira Dal'Molin C., Quek L.-E., Palfreyman R. W., Brumbley S. M., Nielsen L. K., (2010), Plant Physiology, 152, 579-589.
6. Cerami E., Demir E., Schultz N., Taylor B. S., Sander C., (2010), PLoS ONE, 5(2), e8918.
7. Przybecki Z., Wóycicki R., Malepszy S., (2009), Post. Biol. Kom., 36., suppl. 25, 19-31.
8. Varshney R. K., Nayak S. N., May G. D., Jackson S. A., (2009), Trends in Biotechnol., 27, 522-530.
9. Varshney R. K., Graner A., Sorrells M. E., (2005), Trends Plant Sci., 10, 621-630.
10. Masojć P., Lebiecka K., Milczarski P., Wiśniewska M., Łań A., Owsianicki R., (2009), Euphytica, 170, 123-129.
11. Barrett J. C., Hansoul S., Nicolae D. L., Cho J. H., Duerr R. H., Rioux J. D., Brant S. R., Silverberg M. S., Taylor K. D., Barmada M. M., Bitton A., Dassopoulos T., Datta L. W., Green T., Griffiths A. M., Kistner E. O., Murtha M. T., Regueiro M. D., Rotter J. I., Schumm L. P., Steinhart A. H., Targan S. R., Xavier R. J., (The NIDDK IBD Genetics Consortium), Libioulle C., Sandor C., Lathrop M., Belaiche J., Dewit O., Gut I., Heath S., Laukens D., Mni M., Rutgeerts P., van Gossum A., Zelenika D., Franchimont D., Hugot J.-P., de Vos M., Vermeire S., Louis E., (The Belgian-French IBD Consortium), The Wellcome Trust Case Control Consortium, Cardon L.R., Anderson C. A., Drummond H., Nimmo E., Ahmad T., Prescott N. J., Onnie C. M., Fisher S. A., Marchini J., Ghori J., Bumpstead S., Gwilliam R., Tremelling M., Deloukas P., Mansfield J., Jewell D., Satsangi J., Mathew C. G., Parkes M., Georges M., Daly M. J., (2008), Nature Genet., 40, 955-962.
12. Hirschhorn J. N., (2009), N. Engel. J. Med., 360, 1699-1701.
13. Cookson W., Liang L., Abecasis G., Moffatt M., Lathrop M., (2009), Nature Rev., 10, 184-194.
14. Emilsson V., Thorleifsson G., Zhang B., Leonardson A. S., Zink F., Zhu J., Carlson S., Helgason A., Walters G. B., Gunnarsdottir S., Mouy M., Steinthorsdottir V., Eiriksdottir G. H., Bjornsdottir G., Reynisdottir I., Gudbjartsson D., Helgadottir A., Jonasdottir A., Jonasdottir A., Styrkarsdottir U., Gretarsdottir S., Magnusson K. P., Stefansson H., Fossdal R., Kristjansson K., Gislason H. G., Stefansson T., Leifsson B. G., Thorsteinsdottir U., Lamb J. R., Gulcher J. R., Reitman M. L., Kong A., Schadt E. E., Stefansson K., (2008), Nature, 452, 423-428.
15. Schadt E. E., Monks S. A., Drake T. A., Lusisk A. J., Chek N., Colinayok V., Ruff T. G., Milligan S. B., Lamb J. R., Cavet G., Linsley P. S., Mao M., Stoughton R. B., Friend S. H., (2003), Nature, 422, 297-302.
16. HANS NABsys. http://nabsys.com/
17. Wang Z., Gerstein M., Snyder M., (2009), Nature Rev. Genet., 10, 57-63.
18. Yassoura M., Kaplana T., Fraser H. B., Levin J. Z., Pfiffner J., Adiconis X., Schroth G., Luo S., Khrebtukova I., Gnirke A., Nusbaum C., Thompson D.-A., Friedman N., Regev A., (2009), PNAS, 106, 3264-3269.
19. Yin Z., Meng F., Song H., Wang X., Xu X., Yu D., (2010), Plant Physiology, 152, 1625-1637.
20. Chen X., Hackett C. A., Niks R. E., Hedley P. E., Booth C., et al., (2010), PLoS ONE, 5(1), e8598.
21. Carr P. A., Church G. M., (2009), Nature Biotechnol., 27, 12, 1151-1163.
22. Dietz S., Panke S., (2010), BioEssays, 32, 356-362.

23. Gibson D. G., Benders G. A., Andrews-Pfannkoch C., Denisova E. A., Baden-Tillson H., Zaveri J., Stockwell T. B., Brownley A., Thomas D. W., Algire M. A., Merryman C., Young L., Noskov V. N., Glass J. I., Venter J. C., Hutchison III C. A., Smith H. O., (2008), Science, 319, 1215-1220.
24. Itaya M., Tsuge K., Koizumi M., Fujita K., (2005), Proc. Natl. Acad. Sci. USA, 102, 15971-15976.
25. Wang H. H., Isaacs F. J., Carr P. A., Sun Z. Z., Xu G., Forest C. R., Church G. M., (2009), Nature, 460, 894-898.
26. Posfai G., Plunkett III G., Feher T., Frisch D., Keil G. M., Umenhoffer K., Kolisnychenko V., Stahl B., Sharma S. S., de Arruda M., Burland V., Harcum S. W., Blattner F. R., (2006), Science, 312, 1044--1046.
27. Shimizu Y., Kanamori T., Ueda T., (2005), Methods, 36, 299-304.
28. Chan L. Y., Kosuri S., Endy D., (2005), Mol. Syst. Biol., 1, 2005.0018.
29. Purnick P. E. M., Weiss R., (2009), Nature Rev. Mol. Cell Biol., 10, 410-422.
30. Khalil A. S., Collins J. J., (2010), Nature Rev. Genet., 11, 367-379.
31. Tyo K. E. J., Kocharin K., Nielsen J., (2010), Curr. Opinion in Microbiol., 13, 1-8.
32. McArthur IV G. H., Fong S. S., (2010), J. of Biomed. and Biotech., Article ID 459760.
33. Friedland A. E., Lu T. K., Wang X., Shi D., Church G., Collins J. J., (2009), Science, (May 29), 324(5931): 1199-1202.doi:10.1126/science. 1172005.
34. Huang S., Li R., Zhang Z., Li L., Gu X., Fan W., Lucas W. J., Wang X., Xie B., Ni P., Ren Y., Zhu H., Li J., Lin K., Jin W., Fei Z., Li G., Staub J., Kilian A., van der Vossen E. A G., Wu Y., Guo J., He J., Jia Z., Ren Y., Tian G., Lu Y., Ruan J., Qian W., Wang M., Huang Q., Li B., Xuan Z., Cao J., Asan, Wu Z., Zhang J., Cai Q., Bai Y., Zhao B., Han Y., Li Y., Li X., Wang S., Shi Q., Liu S., Cho W. K., Kim J.-Y., Xu Y., Heller-Uszynska K., Miao H., Cheng Z., Zhang S., Wu J., Yang Y., Kang H., Li M., Liang H., Ren X., Shi Z., Wen M., Jian M., Yang H., Zhang G., Yang Z., Chen R., Liu S., Li J., Ma L., Liu H., Zhou Y., Zhao J., Fang X., Li G., Fang L., Li Y., Liu D., Zheng H., Zhang Y., Qin N., Li Z., Yang G., Yang S., Bolund L., Kristiansen K., Zheng H., Li S., Zhang X., Yang H., Wang J., Sun R., Zhang B., Jiang S., Wang J., Du Y., Li S., (2009), Nat. Genet., published online, 1 November 2009; doi:10.1038/ng.475.
35. Wóycicki R., Witkowicz J., Pawełkowicz M., Siedlecka E., Gutman W., Pląder W., Seroczyńska A., Śmiech M., Niemirowicz-Szczytt K., Bartoszewski G., Gawroński P., Dąbrowska J., Karpiński S., Borodovsky M., Lomsadze A., Malepszy S., Przybecki Z. (w przygotowaniu).
36. Miller J., Koren S., Walenz B., Sutton G., Knight J., Jarvie T., Kodira C., Affourtit J., Harkins T., (2009), Poster Santa Fe.
37. Gawronski P., Dabrowska J., Woycicki R., Bartoszewski G., Przybecki Z., Malepszy S., Karpinski S., (2010), SEB Annual Main Meeting 2010 Clarion Congress Hotel, Prague, Czech Republic 30th June – 3rd July 2010.