Genomika
od drożdży do człowieka

Termin „genomika" stworzył sir Thomas Roderick w 1986 roku, niejako na wyrost i przypadkiem, podczas poszukiwania dobrego tytułu dla nowego czasopisma. Impulsem do założenia „Genomics" był przewidywany, gwałtowny napływ prac związanych z planowanym już Projektem Sekwencjonowania Genomu Człowieka (ang. HGP). Pierwotnie nowe pismo miało się skupić na zagadnieniach związanych z genomem człowieka. W następnym dwudziestoleciu określenie to zrobiło oszałamiającą karierę, choć na wyniki analizy genomu człowieka przyszło jeszcze trochę poczekać.

MAREK ZAGULSKI

Genomika, a cóż to za nauka?

Genomika jest definiowana jako nauka dążąca do ustalenie sekwencji nukleotydowej, organizacji i funkcji wszystkich genów. Powstała dzięki „mariażowi" biologii molekularnej i biologii komórki z klasyczną genetyką, wspomaganemu przez informatykę.

Genomika narodziła się w 1992 roku, gdy opublikowano pierwszą sekwencję chromosomu 1. drożdży (*S. cerevisiae*). Pomimo że ustalono wtedy sekwencję zaledwie jednego z 16 chromosomów, a nie całego genomu. Postęp prac wymógł podział genomiki na poddyscypliny.

Analiza organizacji chromosomu, identyfikacja genów i sekwencji regulatorowych spowodowały powstanie **genomiki strukturalnej**.

Termin **genomika strukturalna** jest także stosowany od 1999 roku przez biologów zajmujących się ustalaniem struktury białek. Jako że biolodzy zajmujący się analizą struktury białek opierają się na danych uzyskanych przez naukowców badających genomy, można śmiało powiedzieć, że mamy do czynienia z przypadkiem pożarcia matki przez własne dzieci. Bardziej właściwą nazwą dla globalnej analizy proteomów wydaje się być coraz powszechniej używane określenie „proteomika strukturalna".

Badania prowadzone w celu identyfikacji funkcji nowo odkrytych genów zapoczątkowały **genomikę funkcjonalną**.

Ostatecznie porównywanie wyników uzyskanych przez genomiki strukturalną i funkcjonalną, dla różnych organizmów, dało początek **genomice porównawczej**.

Badanie około 6000 genów drożdży (*S. cerevisiae*), odkrytych w trakcie badania ich genomu, na długi czas zdominowało genomikę strukturalną. Następująca po nich, dzięki doskonale rozwiniętej genetyce

Genom oznacza kompletny zestaw genów zawartych w chromosomach danego organizmu. Sam termin został po raz pierwszy użyty już ponad 80 lat temu przez Hansa Winklera i powstał dzięki elizji słów **GEN**es i chromos**OME**s.

drożdży, fala doniesień o eksperymentalnie potwierdzonej funkcji nowo odkrytych genów stała się głównym źródłem wiedzy dla naukowców zajmujących się genomikami funkcjonalną i porównawczą innych gatunków, w tym również człowieka.

Gwałtowny napływ informacji o sekwencji nukleotydowej i organizacji genomów różnych organizmów zmienił oblicze biologii molekularnej i szybko zainicjował powstanie genomiki systemów, proteomiki, metabolomiki, bioinformatyki etc., a także całkowicie zmienił systematykę i medycynę. Trudno już sobie wyobrazić pracę genetyka czy hodowcy bez kompletnej informacji o genomie badanego organizmu. Genomika na pewno nie mogłaby istnieć bez komputerów, sieci informatycznej i baz danych. Dlatego opisując jej początki, nie można zignorować postępu w informatyzacji nauki. Początkowo wydawało się, że rola komputerów będzie ograniczona do zbierania i wymiany terabajtów informacji (np. sekwencji nukleotydowych i dedukowanych sekwencji aminokwasowych białek). Powstawały serwery udostępniające i klasyfikujące dane pozyskane czy to w czasie realizacji dużych projektów badawczych, czy też jako pojedyncze sekwencje deponowane przez niezależnych naukowców. Jednak ich dostępność była początkowo ograniczona przez konieczność logowania się na serwerach celem uruchomienia oprogramowania (np. GCG) lub długotrwałego i skomplikowanego ściągania danych (FTP-GeneBank). Brak rozwiniętej sieci i zaawansowanych protokołów sieciowych spowodował, że raczkujący internet wykorzystywano jedynie do wymiany korespondencji za pomocą poczty elektronicznej oraz przesyłania uzyskanych sekwencji nukleotydowych, w postaci plików tekstowych, do centralnego serwera projektu. Wyniki niezbędne do zweryfikowania jakości uzyskanych sekwencji przesyłane były zwykłą pocztą (!). Pierwsze edycje kompletnych sekwencji chromosomów drożdży dostępne były na dyskach CD. Dopiero powstanie i upowszechnienie HTML i WWW znakomicie poprawiło dostępność i jakość baz danych. Szybko okazało się, że aby mogły być one użyteczne, informacja deponowana w postaci sekwencji nukleotydowej wymaga przynajmniej wstępnego opracowania i klasyfikacji. Dlatego powstało oprogramowanie umożliwiające, poprzez WWW, analizę sekwencji na wyspecjalizowanych serwerach, a stąd był już tylko krok do powstania portali łączących wielorakie bazy danych sekwencji nukleotydowych kwasów nukleinowych i aminokwasowych białek, krystalograficznych i literaturowych (takich jak SGD, NCBI). Początkowo dane przesyłane i udostępniane przez serwery były pozyskiwane za pomocą procedur eksperymentalnych, szybko jednak okazało się, że gromadzona informacja staje się pokusą do spekulacji i porównań.

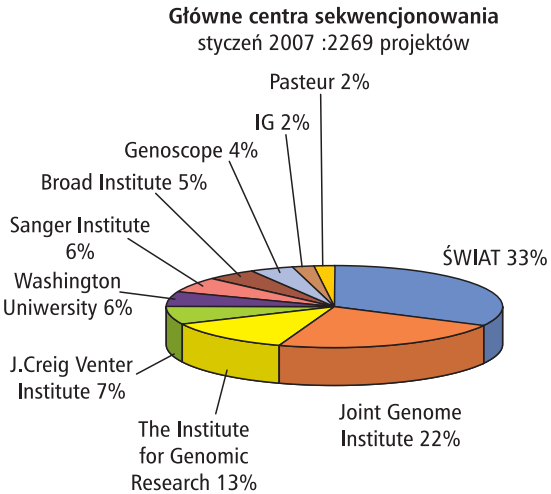
„Posiadanie" końcówki „-omics" stało się niesłychanie modne nie tylko w naukach biologicznych, zajmujących się analizą olbrzymiej ilości danych, ale i w całym świecie nauki. Dodanie końcówki „-omics" do uprzednio istniejącej nazwy nauki uznano za nobilitację sugerującą bardziej globalne i systemowe podejście do problemu. Przykładami tego typu pojęć są „metagenomika" i „paleogenomika".

Wciąż rosnącą listę mniej lub bardziej wydarzonych „dzieci" genomiki można znaleźć na stronie http://www.genomic-glossaries.com/content/omes.asp. Większość z nich ma z kilkunastoletnią „babcią" wspólną jedynie końcówkę nazwy. Ludzie składający mozolnie sekwencje pierwszych genów mogą być dumni i nieco zagubieni, nie rozpoznając części trochę „hałaśliwego potomstwa" genomiki.

Fabryki sekwencji

W 1995 r. opublikowano pierwszy kompletny genom (bakterii *Haemophilus influenzae*) o wielkości 1,830 MB (milion par zasad), a kilka miesięcy później 0,580 MB genomu *Mycoplasma genitalium*, obwieszczając światu sukces, który był raczej medialny i miał osłabić znaczenie sekwencjonowania genomu drożdży (projekt YGSP). Nie umniejszając znaczenia ustalenia pierwszej, kompletnej sekwencji nukleotydowej geno-

mu bakterii, należy pamiętać, że w tym samym czasie w bazach danych konsorcjum realizującego YGSP zdeponowano ponad 6 MB sekwencji, z czego 2,85 MB sekwencji 6 chromosomów już opublikowano. Niekwestionowaną zasługą zespołu Craiga Ventera realizującego ten projekt było wykazanie wyższości – w przypadku sekwencjonowania genomów – systemu shot-gun połączonego z odpowiednim opracowaniem wyników *in silico* nad metodą sekwencjonowania uporządkowanego. Konsekwencją sukcesu Craiga Ventera było zwycięstwo wyspecjalizowanych centrów sekwencjonowania nad projektami rozproszonymi.



Rys. 1. Procentowy udział w ustalaniu sekwencji genomów na podstawie: http://www.genomesonline.org/gold_statistics.htm#aname

Powstanie fabryk sekwencji (Rys. 1) doprowadziło do gwałtownego napływu danych. Istniejące i powstające centra sekwencjonowania (JGI, TIGR, Genoscope), aby wykorzystać swój wciąż rosnący potencjał, zaczęły same zabiegać o fundusze na następne, coraz bardziej „ambitne" projekty. Obecnie większość sekwencjonowania DNA odbywa się w dziewięciu największych centrach. Ponad 60% sekwencji genomów dostarczają centra zlokalizowane w USA. Sekwencjonowanie w centrach, prowadzone przez „technologów sekwencjonowania", jak każda produkcja masowa zaowocowało niestety obniżeniem jakości otrzymywanych sekwencji. Mnożą się błędy!

Projekt Sekwencjonowania Genomu Człowieka

Oczywiście ze względu na „jedynie słuszny antropomorfizm", a także wielkość projektu, celem największego wyścigu stało się zsekwencjonowanie genomu człowieka. Wyścig był sprawą polityczną i prestiżową. Jego efekty, udostępniane w prasie, były cokolwiek żenujące. Przez lata, co kilka miesięcy ogłaszano „przełom" w postaci zmapowania lub zsekwencjonowania tego czy innego chromosomu człowieka lub nawet całego genomu.

Rozpoczęty w 1990 roku, przez amerykański Department of Energy (DOE) we współpracy z NIH, Human Genome Project (HGP) oficjalnie został zakończony w 2004 roku. W międzyczasie, 15 lutego 2001 roku konsorcjum, starając się uprzedzić spodziewaną publikację konkurencyjnej grupy Craiga Ventera, opublikowało w „Nature" „Initial sequencing and analysis of the human genome". Dzień później ukazała się w „Science" publikacja Craiga Ventera z Celera Genomics, dotycząca ustalenia konsensusu sekwencji euchromatyny. Grupa Ventera uzyskała 27 mln odczytów. Było to o 30% za mało, by ustalić wiarygodny konsensus sekwencji. Aby osiągnąć zadowalającą wiarygodność, posłużono się publicznie dostępnymi danymi z... Human Genome Project! Mimo tego, wątpliwego etycznie posunięcia, w publikacji Ventera nie podano pełnej sekwencji poszczególnych chromosomów, a tylko zbiór tysięcy przypisanych do nich fragmentów.

Poczynając od 2000 roku, ekipy związane z HGP publikowały kolejne, pełne sekwencje pojedynczych chromosomów. Podanie 18 maja 2006 roku pełnej sekwencji chromosomu 1 wraz z informacją o zakończeniu projektu wywołało pewną konsternację. Ale to właśnie 18.05.2006 r. należy uznać za moment zakończenia projektu.

Końcowa jakość danych uzyskanych w trakcie HGP pozostawia wiele do życzenia. W przypadku ostatniego, zsekwencjonowanego chromosomu nie odczytano – z powodów technologicznych – około 1,3 mln par zasad euchromatyny. Jeśli dodamy do tego brakujące 13,3 mln par zasad heterochromatyny, to okaże się, że około 6% całkowitej sekwencji chromosomu 1 człowieka wciąż nie jest poznana.

Odnośnie do pozostałych chromosomów sytuacja jest często znacznie gorsza. Reasumując, pełna sekwencja genomu człowieka nadal nie jest ustalona i nie wiadomo, czy ją poznamy.

Genomika – kalendarium

Genomika jest nierozerwalnie związana z pojawieniem się komputerów i internetu. Większość rezultatów projektów sekwencjonowania genomów, baz danych sumujących wyniki analiz wykonanych przez małe zespoły badaczy i konsorcja oraz „eksperymentów bioinformatycznych" jest dostępna tylko w formie elektronicznej. Dlatego w prezentowanym poniżej kalendarium powstanie i upowszechnienie komputera PC oraz internetu jest równie ważne jak praca Sangera opisująca metodę sekwencjonowania DNA, która w znacznie udoskonalonej i częściowo zautomatyzowanej postaci jest stosowana do dziś.

1953 – opisanie struktury DNA,
1977 – opracowanie chemicznej metody sekwencjonowania DNA,
1977 – opracowanie enzymatycznej metody sekwencjonowania DNA, wykorzystującej dideoksynukleotydy,
1977 – ustalenie sekwencji pierwszego genu
1981/1983 – komputer osobisty Apple II/IBM PC/XT,
1982 – powstanie GenBanku,
1982 – zsekwencjonowano pierwszy genom – bakteriofag ΦX174,
1986 – powstaje termin „genomika",
1987 – pierwszy automatyczny sekwenator firmy Applied Biosystems,
1989 – wprowadzono HTTP & HTML Tim Breners-Lee, CERN,
1991/1992 – internet – upowszechnienie – Gopher/WWW,
1992 – zsekwencjonowany chromosom III *S. cerevisiae*,
1992–1996 – YGSP projekt drożdżowy – początek prawdziwej genomiki,
1995 – odczytany genom *H. influenze*,
1997 – bioinformatyka – pierwsza opublikowana praca, w której wszystkie prezentowane wyniki otrzymano przez analizę komputerową,
18.05.2006 – zsekwencjonowany genom *Homo sapiens* 2001, 2003,
2005 – The genome Sequencer 20™ firmy 454 Life Sciences – sekwenator paralelny,
2006 – sekwenatory genomowe (Solexa, Solid).

YGSP

☐ Sekwencjonowanie i analiza funkcjonalna genomu drożdży

W publikacji przeglądowej „Sequencing and functional analysis of the yeast genome" zawarto podsumowanie projektu sekwencjonowania genomu drożdży, ostateczne rezultaty oraz przegląd podstawowych, dostępnych po zakończeniu projektu metod analizy funkcjonalnej. Przewagą konsorcjum drożdżowego nad zespołem Craiga Ventera było to, że sekwencjonowaniem zajęli się ludzie doskonale znający klasyczną genetykę, fizjologię i cytologię. **Dla nich poznanie sekwencji nukleotydowej genomu drożdży nie było celem samym w sobie, a tylko etapem niezbędnym do kontynuacji dotychczasowej pracy. Poznanie genomu drożdży oraz olbrzymia ilość informacji wynikająca z intensywnych badań nad wszystkimi aspektami życia drożdży dały możliwość natychmiastowej ewaluacji informacji uzyskanych w trakcie sekwencjonowania. W rezultacie to drożdże stały się organizmem modelowym i to nie tylko w badaniach genomicznych. Stały się również przykładem właściwie wykorzystanego potencjału genomiki.** Dane eksperymentalne dotyczące funkcji nowo odkrytych genów

do dziś są podstawą bioinformatycznej anotacji genów innych organizmów, dla których ze względu na trudności metodyczne przeprowadzenie takich analiz nie jest łatwe lub wręcz jest niemożliwe z różnych przyczyn. Sukces wynikający z równoczesnego zaangażowania setek laboratoriów w fazę postsekwencyjną YGSP nie został już właściwie nigdy powtórzony. Podstawową i najważniejszą różnicą pomiędzy YGSP a HGP jest stopień wykorzystania wyników uzyskanych w trakcie projektu sekwencjonowania. Naukowcy sekwencjonujący genom drożdżowy, dzięki jego stosunkowo prostej organizacji (np. prawie brak intronów), łatwo identyfikowali nawet te geny, których wcześniej nie znano. Wystarczyło za pomocą prostego algorytmu znaleźć w sekwencji otwartą ramkę odczytu (sekwencję rozpoczynającą się kodonem AUG, począwszy od którego kolejne trójki nukleotydów oznaczają kolejne aminokwasy, a kończącą się kodonem stop) i już można było przystąpić do klonowania i określania funkcji potencjalnego genu.

Poznanie sekwencji nukleotydowej genomu nie jest równoznaczne z poznaniem sekwencji nukleotydowych zlokalizowanych w nim genów. Poznanie sekwencji nukleotydowej genu nie oznacza, że znamy jego funkcję. Czasem możemy domyślać się, jaką rolę pełni. Jednak są to jedynie przypuszczenia, które trzeba potwierdzić eksperymentalnie. Podstawowym sposobem badania funkcji genów jest ich precyzyjne uszkadzanie lub nawet usuwanie z genomu. Usunięcie lub poważne uszkodzenie genu istotnego dla funkcjonowania komórki spowoduje jej śmierć. W przypadku innych genów lub mniejszych uszkodzeń efektem może być zmiana fenotypu. To informacja określająca funkcję badanego genu. Z przyczyn oczywistych tego typu badania na komórkach ludzkich nie mogą być przeprowadzane w przypadku wielu genów, np. związanych z rozwojem i różnicowaniem się naszego organizmu. Z drugiej strony są to geny kluczowe nie tylko dla zrozumienia embriogenezy, ale również wielu chorób genetycznych, również chorób nowotworowych. Dlatego tak ważne są badania prowadzone na organizmach modelowych takich jak drożdże (*S. cerevisiae*). Ponieważ jednak istnieją ogromne różnice pomiędzy drożdżami a człowiekiem, ważne jest, aby wyniki uzyskane dla tego prostego organizmu eukariotycznego potwierdzać na innych, bliższych człowiekowi organizmach modelowych, np. myszach.

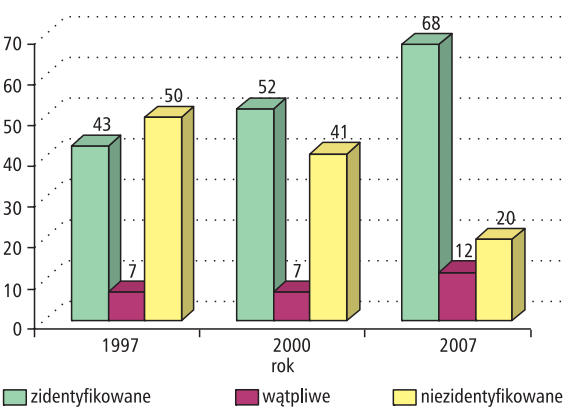
Jedynie słuszny antropomorfizm doprowadził do uznania służebnej roli genomiki organizmów modelowych jako generatora danych dotyczących funkcji i regulacji genów. Uzyskane informacje miały być wykorzystane poprzez analizę porównawczą przy opisaniu genomu ludzkiego. Oczywiście użyteczność wyników uzyskanych w „mokrych" eksperymentach przeprowadzonych na organizmach modelowych zależy od jakości algorytmów, oprogramowania i sprzętu komputerowego, dostępnego w trakcie realizacji projektów.

Zapotrzebowanie na podobne analizy zadecydowało o rozwoju bioinformatyki w jej klasycznej postaci. Pogląd o służebnej roli projektów genomowych organizmów modelowych prezentowali przede wszystkim Amerykanie, o dziwo także członkowie Yeast Community. Można się zastanawiać, czy ten konformizm (koncepcja organizmów modelowych) był podyktowany pragmatyką związaną z pozyskiwaniem funduszy, czy też autentyczną wiarą, że wyniki analizy funkcjonalnej, uzyskane dla drożdży, będzie można w pełni odnieść do organizmu człowieka. Faktem jest, że nawet w przypadku podstawowej analizy sekwencji wyższych eukariontów poszukiwania i anotacji genów, bez odnoszenia się do wyników uzyskanych dla organizmów modelowych niewiele ona daje. Nawet jeśli uda się, co bardzo wątpliwe, za pomocą prostego porównania uzyskanej sekwencji z dostępnymi w bazach danych sekwencjami nukleotydowymi cDNA, zlokalizować gen składający się z intronów i z eksonów, to określenie jego funkcji jedynie w oparciu o dane uzyskane dla organizmów mo-

delowych nie jest już takie proste, a często prowadzi do fałszywych wniosków.

Podstawowa analiza genomów wyższych eukariontów opiera się na wynikach uzyskanych przez analizę sekwencji zdeponowanych w bazach danych. Sekwencja kodująca (np. polipeptyd), często nie dłuższa niż 2000 par zasad (bp), u wyższych eukariontów często bywa pofragmentowana i przez to ukryta w kilkudziesięciu tysiącach liter DNA. Geny kodujące pre mRNA, zawierające od kilku do kilkudziesięciu małych eksonów, są niezwykle trudne do wykrycia, co nie znaczy, że jest to niemożliwe. Przykładem może być gen LAG1hs, którego 7 eksonów jest rozrzuconych na przestrzeniu 25 kb. Gen został zlokalizowany w genomie dzięki porównaniu sekwencji genomu z sekwencją jego cDNA. Z kolei cDNA LAGhs zidentyfikowano dzięki jego istotnemu podobieństwu do sekwencji drożdżowego genu LAG1sc. Ostatecznym potwierdzeniem jego funkcji był przeprowadzony eksperyment pokazujący komplementację fenotypu delecji LAG1sc przez nadeksprymowany LAGhs. Doświadczenie polegało na wprowadzeniu do szczepu drożdży (*S. cerevisiae*) pozbawionego genu LAG1sc jego ludzkiego odpowiednika (LAG1hs) w takiej postaci, że można było uzyskać wysoką produkcję kodowanego przez ten gen białka. Jeśli wytwarzanie przez drożdże ludzkiego białka znosi efekt delecji określonego genu (w tym wypadku LAG1sc), oznacza to, że ludzkie białko może pełnić (w drożdżach) funkcję drożdżowego białka, którego brak (delecja genu) powoduje wykrywalną zmianę fenotypu. Ludzkie białko może funkcjonalnie zastąpić białko drożdżowe.

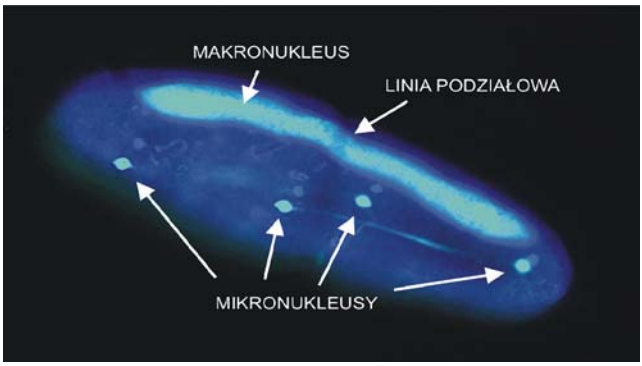
Tuż przed zakończeniem realizacji YGSP, w styczniu 1996, wystartował EUROFAN (European network for functional analysis of yeast genes discovered by systematic DNA sequencing) projekt europejski, mający na celu dokonanie analizy funkcjonalnej 1000 genów odkrytych w trakcie trwania projektu sekwencjonowania genomu drożdży. W projekcie tym ewaluacja uzyskanych sekwencji nukleotydowych była nie tylko najbardziej wszechstronna, ale i natychmiastowa. Fragmenty sekwencji odczytywane w poszczególnych laboratoriach niezwłocznie poddawane były analizie informatycznej. Najciekawsze znalezione geny stawały się przedmiotem badań genetycznych i biochemicznych. Jednocześnie trwały gorączkowe prace nad adaptacją metod analizy genów do badań na dużą skalę oraz nad wprowadzaniem i testowaniem nowych, wydajniejszych i precyzyjniejszych metod. Doprowadziło to do powstania odwrotnej genetyki (ang. reverse genetics). Historia dowodzi, że konsorcjum drożdżowe było kuźnią metod dla genomik funkcjonalnej oraz porównawczej. Warto zauważyć, że rezultaty analiz masowych są najczęściej traktowane tylko jako punkt wyjścia lub uzupełnienie wyników analizy pojedynczego genu lub szlaku metabolicznego, wykonywanych przez niezależnych badaczy czy zespoły badawcze. Metody identyfikacji funkcji genów, oparte na fantastycznie rozwiniętej genetyce klasycznej i biochemii były na bieżąco modyfikowane. Dzięki mrówczej pracy genetyków badających drożdże od 1997 roku około 1500 genom została przypisana funkcja (Rys. 2).



Rys. 2. Postępy w identyfikacji funkcji produktów genów S. cerevisiae (lata 1997, 2000 i 2007 w procentach); źródło: SGD

Projekt sekwencjonowania genomu pantofelka

Pantofelek (*Paramecium tetraurelia*) jest przedstawicielem typu orzęsków (*Cilliate*). Unikalną dla tej rodziny cechą jest posiadanie złożonego aparatu jądrowego, funkcjonującego w bardzo skomplikowany sposób. Posiadanie dwóch typów jąder: somatycznego (makronukleus) i aktywnego wyłącznie w procesach płciowych (mikronukleus), zostało wymuszone poprzez funkcjonowanie orzęsków jako olbrzymich jednokomórkowych organizmów.



Rys. 3. Pantofelek (*Paramecium tetraurelia*) (fot. dr Jacek K. Nowak)

Skomplikowane zależności pomiędzy makro- i mikronukleusami zostały „wypracowane" w toku trwającej 2 mld lat, niezależnej ewolucji. Sugeruje się, że powielenie do 800n informacji genetycznej w gigantycznym makronukleusie jest sposobem na zapewnienie wystarczającej ilości RNA niezbędnego do funkcjonowania olbrzymiej i niezwykle złożonej komórki.

Gigantyczny, aktywny transkrypcyjnie makronukleus (Rys. 3) nie jest najlepszym miejscem przechowywania DNA. Świadczy o tym degradacja chromosomów makronukleusa, postępująca wraz z kolejnymi podziałami wegetatywnymi. Dlatego funkcją mikronukleusa, aktywnego jedynie w procesach seksualnych i w czasie podziału komórki, jest przechowywanie materiału genetycznego w niezmienionej postaci. Proces dojrzewania nowego makronukleusa, polegający na amplifikacji do 800n i fragmentacji chromosomów oraz na dodawaniu do ich końców telomerów i usuwaniu sekwencji IES (ang. Internal Eliminated Sequences), występuje tylko u orzęsków. Istnienie takiego procesu wraz z innymi unikalnymi dla tych pierwotniaków cechami funkcjonowania materiału genetycznego, oraz duży dystans ewolucyjny dzielący je od organizmów, których genomy zsekwencjonowano wcześniej, pozwalały przypuszczać, że badając genom pantofelka, poznamy bardzo wiele genów o nieznanych funkcjach lub odkryjemy nowe funkcje białek podobnych do wcześniej odkrytych. Analiza informatyczna fragmentu DNA pantofelka o długości około 1,5 MB (około 1% całego genomu) wykazała obecność co najmniej 722 genów, w tym 119 potencjalnych genów kodujących białka nieposiadające odpowiedników w bazach danych. Sugerowało to bardzo dużą ilość genów zawartych w jego genomie. Uzyskana we wstępnych badaniach informacja dotycząca łatwo identyfikowalnych, krótkich (18 35 nukleotydowych) intronów sugerowała, że anotacja sekwencji, w czasie realizacji projektu sekwencjonowania genomu, będzie stosunkowo łatwa.

W publikacji „Paramecium genome survey: a pilot project" przedstawiono narodziny nowego projektu. Można jednoznacznie stwierdzić, że w przypadku pantofelka YGSP było modelem określającym sposób działania. Początkowo, wobec ograniczonej wiedzy, każdy wielki projekt sekwencjonowania genomu był wyprawą w nieznane. W momencie zakończenia YGSP niemal 60% sekwencji odkrytych genów kodowało białka o niezidentyfikowanej funkcji. Wynik ten był i tak zaskakujący, zważywszy na prominentną pozycję drożdży (*S. cerevisiae*) jako najintensywniej badanego w tamtych czasach organizmu eukariotycznego. Nawiasem mówiąc, po latach wytrwałego sekwencjonowania geno-

mów, znalezienie organizmu zwierającego dużą liczbę nieznanych genów było niesłychanie trudne, z wyżej omówionych powodów pantofelek dawał taką nadzieję.

Tak jak siłą napędową projektu YGSP była olbrzymia społeczność drożdżowa, tak niewielka grupa naukowców zajmujących się pantofelkiem stanowiła piętę achillesową projektu sekwencjonowania jego genomu. Niestety, mimo iż był jednym z pierwszych organizmów jednokomórkowych, odkrytych (XVIII wiek) przez obserwację mikroskopową, to laboratoriów zajmujących się pantofelkiem są jedynie dziesiątki, a naukowców setki. Od tamtego czasu badano go bardzo intensywnie, dokonując szeregu istotnych odkryć. Protozoolodzy nie mieli szczęścia w pozyskiwaniu funduszy na projekty sekwencjonowania. Ta mała, ale naukowo bardzo aktywna, międzynarodowa społeczność zabiegała o fundusze już od lat dziewięćdziesiątych. Jej członkowie zdawali sobie sprawę z niemożności kontynuacji badań na światowym poziomie, bez ustalenia pełnej sekwencji genomu. Rozległa wiedza o fizjologii oraz o procesach rearanżacji materiału genetycznego pantofelka była gwarancją właściwej oceny uzyskanych w projekcie sekwencjonowania wyników. Dobrze przygotowane techniki transformacji oraz wyciszania ekspresji badanych genów poprzez RNAi gwarantowały skuteczną analizę funkcjonalną. W 2001 roku w bazach danych istniało tylko kilkadziesiąt sekwencji DNA tego organizmu. Przedstawiony projekt pilotażowy polegał na sekwencjonowaniu obu końców 1800 przypadkowo wybranych fragmentów DNA z biblioteki genów zawierającej fragmenty chromosomalnego DNA pantofelka. Jest to najprostsza i najbardziej ekonomiczna metoda pozyskiwania informacji o strukturze genomu. Ogromnym sukcesem projektu pilotażowego, opartego na poznaniu sekwencji około 1% genomu, była identyfikacja ponad 700 sekwencji kodujących. Sekwencje te zidentyfikowano przez porównanie uzyskanych sekwencji ze znanymi sekwencjami innych organizmów, zdeponowanymi w bazach danych. Uzyskane wyniki, sugerujące bardzo duże upakowanie genów w genomie pantofelka, dowodziły, że projekt sekwencjonowania jego genomu jest sensowny i powinien wnieść wiele nowych informacji o strukturze i funkcjonowaniu genomu prokariotycznego.

Pomimo postępu technologicznego i związanego z nim obniżenia kosztów sekwencjonowania (bardziej wydajne, automatyczne sekwenatory, użycie robotów do izolacji DNA), koszty oznaczenia sekwencji całego genomu nadal są wysokie. Jednak wstępne fazy takiego projektu mogły być realizowane w stosunkowo małym laboratorium, np. w Pracowni Sekwencjonowania DNA IBB PAN, dysponującej kilkoma starego typu, automatycznymi sekwenatorami. Uzyskane, również przez polskich naukowców, wyniki nie zostały zignorowane przez francuskich decydentów. Dzięki temu koszt ostatniej fazy sekwencjonowania genomu pantofelka poniósł rząd francuski. Zrealizowano ją w jednym z dużych centrów sekwencjonowania (Genoscope).

Dane dotyczące gęstości kodowania wraz z informacją o potencjalnej wielkości genomu sugerowały, że pantofelek posiada aż 30 000 genów. Dla porównania w genomie człowieka jest ich około 27 000.

Powszechność zjawiska duplikacji genomów

Znaczenie duplikacji materiału genetycznego jako jednej z podstawowych, o ile nie najważniejszych, sił napędowych ewolucji postulowano już w latach 70. (hipoteza Ohno). Wyniki uzyskane w trakcie YGSP przemawiały za jej potwierdzeniem. Analiza duplikacji zaobserwowanych u *S. cerevisiae* genów zaowocowała hipotezą, że genom tych drożdży powstał w wyniku duplikacji genomu protoplasty – WGD (ang. whole genome duplication). Strukturę genomu *S. cerevisiae* opisuje się jako zdegenerowany genom tetraploidalny. Słowo „zdegenerowany" odnosi się do następujących po duplikacji serii delecji, prowadzących do powstania 16 chromosomów z wyjściowych sześciu oraz zachowania tylko 13% z powielonych

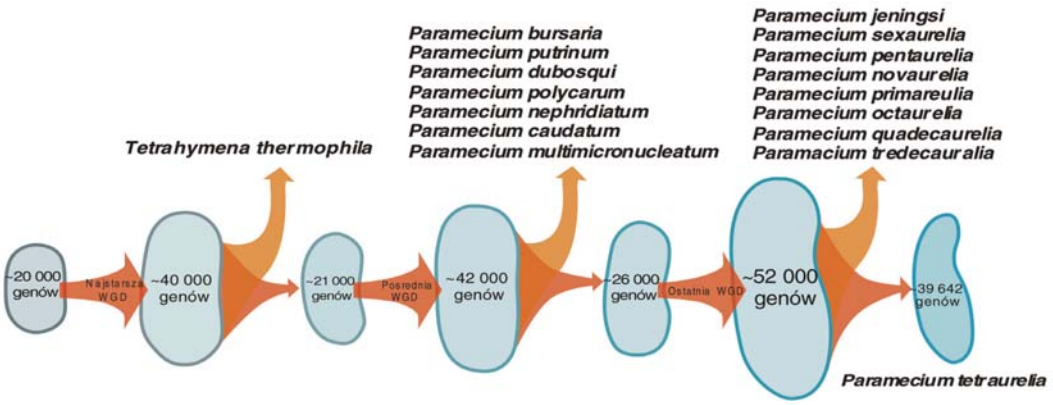
genów. Hipoteza znalazła potwierdzenie w 2004 roku po porównaniu sekwencji i struktury genomów *S. cerevisiae* i innych drożdży – *Kluyveromyces waltii*. Zdaniem autorów poliploidyzacja jako wstępny etap ewolucji prowadzi do niestabilności genomu. Konsekwencją są następujące po duplikacji mutacje, delecje i rearanżacja struktury chromosomów, prowadzące do odzyskania ich stabilności. Nie bez znaczenia dla procesów redukcji ilości materiału genetycznego, w przypadku organizmów jednokomórkowych (R strategów), jest także presja selekcyjna, prowadząca do „pozbycia się zbędnych kosztów". Precyzyjna analiza porównawcza sekwencji czterech gatunków drożdży, należących do *Hemiascomycetes* oraz *S. cerevisiae*, wykazała zaskakującą różnorodność i złożoność mechanizmów specjacji. Poza WGD i następującymi procesami intensywnej utraty genów, dużą rolę w formowaniu genomów odgrywają powtórne duplikacje genów lub całych segmentów chromosomów. Przewaga określonego typu zmian w czasie ewolucji poszczególnych gatunków zdeterminowana jest najczęściej adaptacją do niszy środowiskowej.

Zaskakująco wysoka (w porównaniu z przewidywaniami opartymi na wynikach wcześniejszych prac) liczba zidentyfikowanych genów pantofelka (39 642) wynika z kilku rund duplikacji oraz wysokiego poziomu utrzymanych genów po duplikacji. Trzy rundy duplikacji doprowadziły do sytuacji, w której tylko 32% genów nie posiada paraloga (**paralogi** – geny powstałe w wyniku duplikacji pierwotnego genu). Proces redukcji ilości materiału genetycznego, aczkolwiek ciągły i względnie łatwy do wykrycia, wydaje się działać u pantofelka mniej efektywnie niż w innych organizmach. Może się wydawać, że przyjęta przez orzęski strategia multiplikacji jądra wegetatywnego przy zredukowanej aktywności jądra generatywnego prowadzi do zmniejszonej presji na procesy redukujące genom.

Ustalono, że doszło do co najmniej trzech rund duplikacji, a następnie utraty części genów. Odkrycia dokonane w trakcie projektu sekwencjonowania całego genomu pantofelka stanowią nową wartość. Wcześniej hipoteza WGD opierała się na wynikach uzyskanych za pomocą genomiki porównawczej. W przypadku genezy *S. cerevisiae* na ostateczne dowody potwierdzające rolę WGD w ewolucji tego gatunku trzeba było czekać osiem lat. Analiza sekwencji genomu rzodkiewnika (*Arabidopsis thaliana*) wykazała, że najwyraźniejszym efektem co najmniej trzech, zaszłych w ciągu ostatnich 350 mln lat, rund duplikacji, było zwiększenie stopnia komplikacji systemów regulacyjnych poprzez niemal podwojenie ilości czynników sygnałowych i transkrypcyjnych.

W przeciwieństwie do innych genomów, dla których opisano WGD (drożdże, ryby, rośliny), procesy redukcji wielkości genomu po WGD u pantofelka zostały zahamowane lub znacznie spowolnione. Pozwoliło to na względnie łatwą identyfikację paralogów pochodzących z różnych rund duplikacji. Przeprowadzona szczegółowa analiza profilu utraty genów doprowadziła do zaskakujących wniosków. Niższy poziom utraty paralogów genów będących częścią kompleksów metabolicznych czy białkowych nie jest niczym dziwnym. Działa tu mechanizm konserwujący, wymuszony przez konieczność utrzymania właściwego stosunku stechiometrycznego produktów ekspresji genów. Zaskakująca jest natomiast, stwierdzona u pantofelka, wysoka konserwacja zduplikowanych genów.

Podwojenie ilości białka, produktu genu już wyrażanego na wysokim poziomie, często prowadzi do istotnych zmian fenotypowych. Ich utrwalenie może spowodować problemy z cyklem komórkowym i rozmnażaniem, a w konsekwencji wywołać rozdział populacji, pierwszy etap specjacji. Po pewnym czasie zduplikowane geny mogą ulec zróżnicowaniu, zwłaszcza jeśli zajdą w nich mutacje zwiększające poziom ekspresji lub aktywność kodowanego białka. Następnym etapem jest zwykle utrata lub zmiana funkcji jednego ze zduplikowanych genów. Mutacje zmieniające funkcje



Rys. 4. WGD a specjacja

genów lub ich poziom ekspresji mają prawdopodobnie większe znaczenie w utrwalaniu powstałych na skutek duplikacji fenotypów niż w tworzeniu nowych. Proponowany model specjacji znajduje potwierdzenie w analizie filogenetycznej, łączącej ostatnie WGD z powstaniem grupy bliźniaczych gatunków *Paramecium aurelia*.

Kres genomiki strukturalnej?

Zsekwencjonowanie całego genomu pantofelka było wielkim sukcesem zawiązanego w 2000 roku, międzynarodowego konsorcjum. Sukces ten jest efektem konsekwencji i profesjonalizmu w działaniu małej społeczności badaczy pantofelka. Kolejne etapy projektu prowadziły od mało kosztownego sekwencjonowania biblioteki stosunkowo małych fragmentów genomu, poprzez sekwencjonowanie pojedynczego chromosomu do sekwencjonowania całego genomu. Równolegle z postępami w projekcie sekwencjonowania rozwijano molekularne techniki analizy funkcjonalnej pantofelka. Biorąc pod uwagę uzyskane efekty, organizacja i realizacja tego projektu może być przykładem dla każdej małej społeczności naukowej. Współdziałanie trzech grup: fizjologów i cytologów, laboratoriów sekwencjonowania oraz bioinformatyków doprowadziło do maksymalnego wykorzystania uzyskanych wyników. Jest całkiem możliwe, że projekt sekwencjonowania genomu pantofelka może być jednym z ostatnich istotnych dla genomiki strukturalnej. O jego powodzeniu zadecydowało nie tylko znalezienie 10 000 nowych genów, ale także udokumentowanie kolejnych rund duplikacji, potwierdzających podstawowe znaczenie tego procesu jako głównego mechanizmu ewolucji na poziomie molekularnym.

Paradoksalnie genomika strukturalna pomimo, a może właśnie z powodu, tak wielu podjętych projektów sekwencjonowania ma się ku schyłkowi. Coraz trudniej znaleźć organizm, którego nietypowa organizacja genomu lub posiadanie nieznanych wcześniej genów wystarczy do opublikowania wyników badań w poważnym czasopiśmie. Bazy danych nadal będą rosły w miarę kończenia rozpoczętych projektów, ale czas łowców genów powoli przemija. Nowe projekty sekwencjonowania mogą z góry zakładać odkrycie nie więcej niż kilku procent nieopisanych dotąd genów. Po zsekwencjonowaniu przedstawicieli wszystkich ważniejszych grup taksonomicznych nowe geny można znaleźć teraz tylko w bardzo egzotycznych organizmach lub poprzez metagenomikę, odczytując wszystkie łańcuchy DNA obecne w organizmach symbiotycznych, których wyizolowanie w warunkach laboratoryjnych jest praktycznie niemożliwe, lub sekwencjonując genomy gatunków tworzących całe nisze ekologiczne. Oczywiście nadal będzie po-

pyt na sekwencje genomów organizmów mających znaczenie ekonomiczne, celem usprawnienia ich modyfikacji, ale możemy to traktować już nie jako wielką przygodę odkrywców białych plam, ale jako żmudne rysowanie precyzyjnych planów i map poprzez hodowców. Odczytanie kompletnych sekwencji genomów jest częścią opisania świata i naszym dziedzictwem – to tu znajdziemy informację o tym, jak sobie ze światem radzić, jak odpowiadać na jego niszczący wpływ, jak optymalnie pobierać energię i potrzebne do życia składniki, wreszcie jak reagować na inne osobniki tego samego czy innego gatunku. To wszystko, aby osiągnąć sukces, to znaczy reprodukować się z maksymalną wydajnością i zapewnić przetrwanie oraz rozród swego potomstwa. Genom, a właściwie zawarta w nim informacja, jest z punktu widzenia gatunku najważniejsza. Nasza cywilizacja całkiem niedawno „odkryła", że najcenniejsza jest informacja. Niesłychanie ważny staje się sposób jej uporządkowania, zapewnienie autoryzowanego dostępu, bezpieczne przechowywanie i kopiowanie, a także zabezpieczenie przed informacją szkodliwą oraz przed destrukcyjnymi czynnikami zewnętrznymi. Bardzo istotne okazały się procedury precyzyjnego kopiowania oraz modyfikowania informacji bez zakłóceń. Procesy te od miliardów lat są nieustannie testowane w naturze. Dodatkowo w przypadku zapisu genetycznego żywych, rosnących i ewoluujących organizmów mamy do czynienia z informacją podlegającą nieustannym zmianom. Przypomina to funkcjonowanie szalonego programu wciąż uszkadzanego i naprawianego, a czasem nienaprawianego, ale dzięki szczególnemu mechanizmowi (selekcji) ciągle udoskonalanego.

Odczytana przez nas sekwencja nukleotydowa genomu, z którego jesteśmy tacy dumni, najbardziej przypomina fotografię zmontowaną z kilkudziesięciu klatek różniących się drobnymi szczegółami. Senny koszmar informatyka. Ludzie rozpoczynający sekwencjonowanie genomów przypominają tych, którzy od czasów wielkich odkryć geograficznych niemal do współczesności odkrywali białe plamy na mapie świata. Wówczas technologia dopiero się rodziła i trzeba było na bieżąco rozwiązywać problemy. Jednak od momentu powstania genomiki do dziś, pomimo gwałtownego rozwoju technologii, jej postęp jest uzależniony od kombinacji efektów pracy wielkich zespołów oraz mrówczej pracy pojedynczych naukowców, spędzających życie na analizie funkcji pojedynczego genu. Ewaluacja i kompilacja wyników pracy setek tysięcy ludzi z całego świata jest możliwa tylko dzięki istnieniu systemu wymiany, deponowania i przechowywania informacji naukowej, będącego częścią internetu.

Sekwencjonowanie genomów stało się sprawą nie tyle naukowców, co inżynierów i organizatorów nauki. Sekwencjonowanie genomów stało się działalnością gospodarczą. Coraz częściej to właśnie firmy związane z medycyną czy biotechnologią dostarczają funduszy na kolejne projekty, odciążając budżet nauki. Przy obecnym potencjale sekwencjonowania uzyskanie pełnej sekwencji kolejnego organizmu jest właściwie tylko kwestią pozyskania odpowiednich środków. Podstawowym pytaniem, które należy zadać, planując nowy projekt, jest to, czy istnieje grupa eksperymentatorów, którzy właściwie wykorzystają wysiłek włożony w odczytanie kolejnych miliardów zasad DNA.

Genom za 1000 dolarów

W latach 1990–2005 koszt ustalenia jednej zasady sekwencji spadł tysiąckrotnie z 10 dol. do jednego centa. Ufundowany na początku 2004 roku grant NIH w wysokości 70 mln dol., na obniżenie kosztów technik sekwencjonowania stawia za cel przede wszystkim uzyskanie sekwencji genomu człowieka za 100 000 dol., a docelowo za 1000 dol. Za osiągnięcie tego ostatniego celu ufundowane zostały jeszcze dwie nagrody: 0,5 mln dol. J. Craig Venter Science Foundation i 5,20 mln dol. X Prize

Foundation. Kilkanaście firm przystąpiło do technologicznego wyścigu. Główną stawką nie jest jednak pozyskanie grantu czy nagrody za obniżenie kosztów projektów naukowych. Walka trwa o pozyskanie olbrzymiego rynku przyszłości, rynku indywidualnej diagnostyki i analizy genetycznej. Obecnie znanych jest około 1600 genów ludzkich, w których opisano mutacje skorelowane z chorobami. Ilość SNP (ang. single nucleotide polymorphism) w genomie ludzkim jest obecnie szacowana na 1/10 000 par zasad. Wygląda na to, że to sekwencjonowanie DNA, a nie testy oparte na analizie fragmentów będzie metodą diagnostyki przyszłości. Do tego celu zbudowano sekwenatory klasy genomowej: 454 Life Sciences sprzedawany jako Genome Sequencer 20 firmy Roche, Solexa firmy Illumina oraz Solid firmy Applied Biosystems. Genome Sequencer 20 (454), pomimo nazwy, nie jest tak naprawdę sekwenatorem genomowym – w jednym eksperymencie jest w stanie odczytać najwyżej 100 mln par zasad przy koszcie porównywalnym z klasycznymi metodami. Bez wątpienia robi to szybciej, ale uzyskane sekwencje są krótsze i gorszej jakości. W projektach sekwencjonowania de novo, gdzie ostateczna jakość uzyskanych sekwencji jest ważna, powinien on być stosowany raczej w połączeniu z klasycznymi metodami. Wprowadzona w końcu 2006 roku maszyna Solexa firmy Illumina jest w stanie dostarczyć w pojedynczym eksperymencie kilka miliardów par zasad przy koszcie odczynników na poziomie 4000 dol. Niestety generowane odcinki sekwencji są zbyt krótkie (25 nukleotydów), aby pokusić się o składanie całych genomów. Podobne ograniczenie dotyczy także sekwenatora Solid firmy Applied Biosystems. Nie jest to jednak żaden problem dla producentów. Maszyny te zostały skonstruowane przede wszystkim z myślą o indywidualnej diagnostyce medycznej i są przeznaczone do resekwencjonowania i to w dodatku do resekwencjonowania genomu człowieka. Wydaje się wysoce prawdopodobne, że w ciągu najbliższych lat każdy, tak jak James Watson, będzie dysponował sekwencją własnego genomu wraz z analizą mutacji, przyczyn potencjalnych chorób i będzie to kosztowało o wiele mniej niż 1 mln dolarów.

dr hab. **MAREK ZAGULSKI**
GENOMED

PIŚMIENNICTWO

- Hieter P., Boguski M. S., *Functional Genomics: It's All How You Read It*. Science: 601–602, 24 October 1997.
- Poinar H. N. et al., *Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA*. Science: 392–394, 20 January 2006.
- Fleischmann R. D. et al., *Whole-genome Random Sequencing and Assembly of Haemophilus Influenzae Rd*. Science 269: 496–512, 28 July 1995.
- Fraser C. M. et al., *The Minimal Gene Complement of Mycoplasma genitalium*. Science 270: 397–404, 20 October 1995.
- International Human Genome Sequencing Consortium, *Initial sequencing and analysis of the human genome*. Nature 409: 860–921, 15 February 2001.
- Venter C. J. et al., *The Sequence of the Human genome*. Science 291: 1304–1351, 16 February 2001.
- Botstein D., Cherry J. M., *Molecular linguistics: Extracting information from gene and protein sequences*. PNAS 94: 5506–5507, May 1997.
- Watson J. D., Crick F. H. C., *Molecular structure of nucleic acids. A structure for deoxiribose nucleic acid*. Nature 171: 737–738, 1953.
- Maxam M. A., Gilbert W. A., *New Method for Sequencing DNA*. Proc. Natl. Acad. Sci. U. S. A. 74: 560–564, 1977.
- Sanger F., Nicklen S., Coulson A. R., *DNA sequencing with chain-terminating inhibitors*. Proc. Natl. Acad. Sci. U. S. A. 74: 5463–5467, 1977.
- Mushegian A. R., Bassett D. E. Jr., Boguski M. S., Bork P., Koonin E. V., *Positionally cloned human disease genes: Patterns of evolutionary conservation and functional motifs*. Proc. Natl. Acad. Sci. U. S. A. 94: 5831–5836, 1997.
- Ohno S. 1970. *Evolution by gene duplication*. Springer-Verlag, NY za: E. Pennisi, *Molecular Evolution, Genome Duplications: The Stuff of Evolution?* Science 294: 2458–2460, 21 December 2001.
- Wolfe K. H., Shields D. C., *Molecular evidence for an ancient duplication of the entire yeast genome*, Nature (letters to Nature) 387: 708–713, 12 June 1997.
- Kellis M. et al., *Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae*. Nature 428: 617–624, 8 April 2004.
- Dujon B. et al., *Genome evolution in yeasts*. Nature 430: 35–44, 1 July 2004.
- Maere S. et al., *Modeling gene and genome duplications in eukaryotes*. Proc Natl. Acad. Sci. U. S. A. 102 (15): 5454–5459, 12 April 2005.
- Venter J. C. et al., *Environmental genome shotgun sequencing of the Sargasso Sea*. Science 305: 66–74, 2 April 2004.
- Service R. F., *GENE SEQUENCING; The race for the $1000 genome*. Science 311: 1544–1546, 17 March 2006.
- James Watson's genome sequenced Published online: news@nature.com, 1 June 2007; http://www.nature.com/news/2007/070528/full/070528-10.html.