PRACE PRZEGLĄDOWE

Sekwencjonowanie największego chromosomu *Paramecium tetraurelia* – przykład opracowania metody

Jacek K. Nowak

Instytut Biochemii i Biofizyki, Polska Akademia Nauk, Warszawa

Sequencing of the largest chromosome of *Paramecium tetraurelia*

Summary

The largest, megabase, somatic chromosome of *P. tetraurelia* was isolated and sequenced in order to explore its organization and gene content. The AT-rich chromosome is compact with very small introns, short intergenic regions and a coding density of at least 74%, higher than that reported for budding yeast or any other free-living eukaryote. Homology to known proteins could be detected only for 57% of the 464 potential protein coding genes. Subsequently, the megabase chromosome sequence was used during the whole genome sequencing project as a reference to evaluate sequence assembly and gene annotation accuracy. In a pilot project of the global analysis of *P. tetraurelia* gene expression during autogamy, DNA microarrays were used. Statistical data analysis allowed the identification of four clusters of co-expressed genes. Screening for silencing phenotypes of 15 autogamy specific genes revealed that 4 genes were essential during vegetative growth and 3 others were essential for successful sexual process.

Key words:
ciliates, *Paramecium*, sequencing, DNA microarrays, genome rearrangements.

Adres do korespondencji

Jacek K. Nowak,
Instytut Biochemii
i Biofizyki,
Polska Akademia Nauk,
ul. Pawińskiego 5A,
02-106 Warszawa.

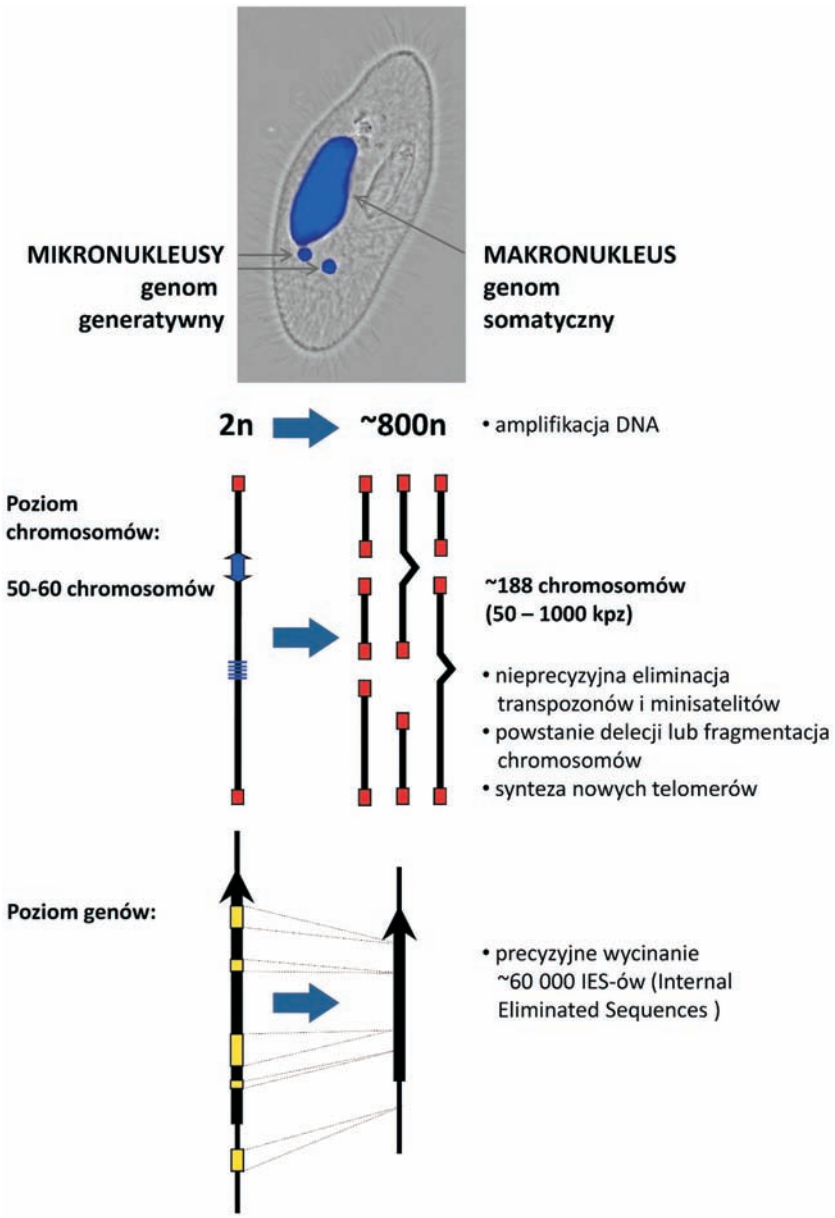
biotechnologia
4 (91) 76–90 2010

1. *Paramecium tetraurelia* jako organizm modelowy

Paramecium tetraurelia, znany pod nazwą pantofelek, jest występującym powszechnie wolno żyjącym jednokomórkowym organizmem eukariotycznym należącym do orzęsków. Pantofelek jest jednym z laboratoryjnych organizmów modelowych po-

zwalających badać różnorodne zjawiska biologiczne. Wyniki uzyskane dla konkretnego modelu mogą zostać w pewnym zakresie uogólnione dla innych organizmów zwiększając naszą wiedzę o świecie żywym. Ze względu na dużą ewolucyjną odległość od innych organizmów modelowych stosowanych w laboratoriach, wykorzystanie *Paramecium* pozwala na nowe podejście do badań podstawowych procesów wspólnych dla komórek eukariotycznych. Ponadto pantofelek charakteryzuje się zestawem cech, które czynią go użytecznym dla badań naukowych. Wystarczy wymienić łatwość hodowli, niepatogenność, poznany niedawno genom oraz szereg zaawansowanych technik biologii molekularnej dostępnych w badaniach, włączając w to interferencję RNA (RNAi), aby uzasadnić dlaczego *P. tetraurelia* jest ciekawym i posiadającym duży potencjał organizmem modelowym (1). *Paramecium* oraz inny orzęsek – *Tetrahymena thermophila*, przyczyniły się do poznania ważnych zjawisk. Pierwszy wyjątek od uniwersalności kodu genetycznego (poza genomem mitochondrialnym) (2,3), katalityczna aktywność RNA, sekwencje telomerów i telomeraza (4), translacyjna kontrola splicingu u eukariontów (5) – to przykłady odkryć, które zawdzięczamy badaniom orzęsków. Ze względu na duży rozmiar komórki (~120 μm), złożoną budowę cytoszkieletu oraz możliwość precyzyjnej obserwacji ruchu komórek, pantofelek jest wykorzystywany do badań obecnych również w organizmie człowieka rzęsek i ciałek podstawowych, których nieprawidłowości odpowiedzialne są za wiele ludzkich chorób.

Podobnie jak u innych orzęsków, w cytoplazmie komórki *P. tetraurelia* znajdują się dwa rodzaje jąder: dwa małe, diploidalne, nieaktywne transkrypcyjnie jądra generatywne – mikronukleusy oraz duże, wysoce poliploidalne (~800n), odpowiedzialne za ekspresję genów jądro somatyczne – makronukleus. Podczas każdego procesu płciowego (koniugacji – zachodzącej pomiędzy komórkami dwóch typów płciowych lub autogamii – samozapłodnienia) stary makronukleus ulega degradacji, a nowy makronukleus jest odtwarzany z mikronukleusa. W czasie powstawania nowego makronukleusa genom generatywny podlega złożonym, powtarzalnym, precyzyjnie regulowanym procesom, w wyniku których część materiału genetycznego jest eliminowana (rys. 1) (6). Rearanżacja materiału genetycznego polega na nieprecyzyjnym usunięciu elementów wielokopijnych, m.in. transpozonów i minisatelitów oraz precyzyjnym wycięciu jednokopijnych, niekodujących sekwencji (IES, ang. *Internal Eliminated Sequence*), które można nazwać „DNA intronami" (7). Precyzyjne usunięcie IES-ów obecnych zarówno w obrębie sekwencji kodujących jak i niekodujących, jest niezbędne do odtworzenia funkcjonalnych ramek odczytu w makronukleusie. Jakiekolwiek poważne zaburzenia w ich usuwaniu prowadzą do śmierci komórek po zakończeniu procesu płciowego. Poza orzęskami swój genom na dużą skalę rearanżuje podczas różnicowania się linii komórkowych m.in. *Ascaris* (nicień), *Cyclops* (skorupiak) oraz śluzice i minogi (kręgowce) (8).



Rys. 1. Rozwój makronukleusa u *P. tetraurelia*. Na zdjęciu przedstawiono komórkę pantofelka z nałożonymi wybarwionymi fluorescencyjnie mikronukleusami i makronukleusem. Na schemacie przedstawiono zachodzące podczas powstawania makronukleusa procesy prowadzące do powstania genomu somatycznego: replikację DNA, usunięcie elementów wielokopijnych, m.in. transpozonów i minisatelitów, prowadzące do powstania heterogennej populacji około 200 chromosomów somatycznych, co jest powiązane z syntezą *de novo* telomerów na końcach powstałych fragmentów. Ponadto precyzyjnie wycinanych jest około 60 000 krótkich (26-882 pz), jednokopijnych, niekodujących IES-ów (7); (zdjęcie autora, schemat na podstawie (6)).

2. Pilotażowy projekt sekwencjonowania genomu *Paramecium tetraurelia*

Znajomość sekwencji DNA badanego organizmu, jak się wydaje, jest w tej chwili absolutnie niezbędna do prowadzenia zaawansowanych badań naukowych. Wraz ze spadkiem kosztów i czasu sekwencjonowania stało się możliwe poznanie sekwencji nukleotydowej genomów wszystkich organizmów modelowych, a nawet sekwencjonowanie bardzo dużych genomów roślin istotnych ekonomicznie (ziemniak, pszenica). W latach dziewięćdziesiątych ubiegłego wieku sytuacja była jednak znacznie mniej korzystna. Naukowcy wykorzystujący w swoich badaniach *Paramecium tetraurelia* jako organizm modelowy, pomimo że prowadzili zaawansowane badania różnorodnych procesów komórkowych, stanowili zbyt małą grupę, by uzyskać fundusze wystarczające do poznania sekwencji całego genomu. Stworzono zatem konsorcjum laboratoriów francuskich, niemieckich i polskich, którego celem było poznanie całego genomu pantofelka. W 2002 r. francuski Narodowy Ośrodek Badań Naukowych oficjalnie powołał GDRE (fr., *Groupement de Recherches Européen*) „Paramecium Genomics", koordynowane przez Jeana Cohena z Centre de Génétique Moléculaire w Gif-sur-Yvette we Francji.

W przypadku *Paramecium* w jednej komórce znajdują się dwa różniące się od siebie genomy. Należało zatem podjąć decyzję – który genom sekwencjonować? Mikro- czy makronuklearny? Zdecydowano o podjęciu się zbadania w pierwszej kolejności genomu makronuklearnego, którego wielkość szacowano na 75-200 Mpz. Za tą decyzją przemawiało szereg faktów: zawartość DNA makronuklearnego w komórce jest około 200 razy większa niż mikronuklearnego, makronukleus jest wolny od niezwykle trudnych do sekwencjonowania sekwencji repetytywnych, ponadto, genomem odpowiedzialnym za ekspresję genów jest właśnie genom somatyczny.

Pierwszym etapem prowadzącym do poznania genomu *Paramecium* było sekwencjonowanie ponad 3000 końców plazmidów pochodzących ze zindeksowanej biblioteki genomowej. Wynikiem było 1,5 Mpz sekwencji stanowiącej ok. 1-1,5% całego genomu (9,10). Zidentyfikowano 722 potencjalnych otwartych ramek odczytu poprzez porównanie z bazami danych oraz 119 potencjalnych nowych, nieznanych wcześniej genów. Sekwencje zdeponowano w bazach danych w listopadzie 2000 r. Istotnym wnioskiem z pilotażowego projektu była wysoka zawartość sekwencji kodującej w DNA pantofelka, szacowana wtedy na ponad 68%. Do jej zwiększenia przyczynia się także długość intronów oszacowana na 18-35 nt. Potwierdziło to, że dalsze sekwencjonowanie losowych fragmentów, przy dość niewielkich kosztach, jest dobrą metodą poznawania nowych genów *Paramecium*.

3. Sekwencjonowanie największego somatycznego chromosomu *P. tetraurelia*

3.1. Izolacja chromosomu i przygotowanie biblioteki genów

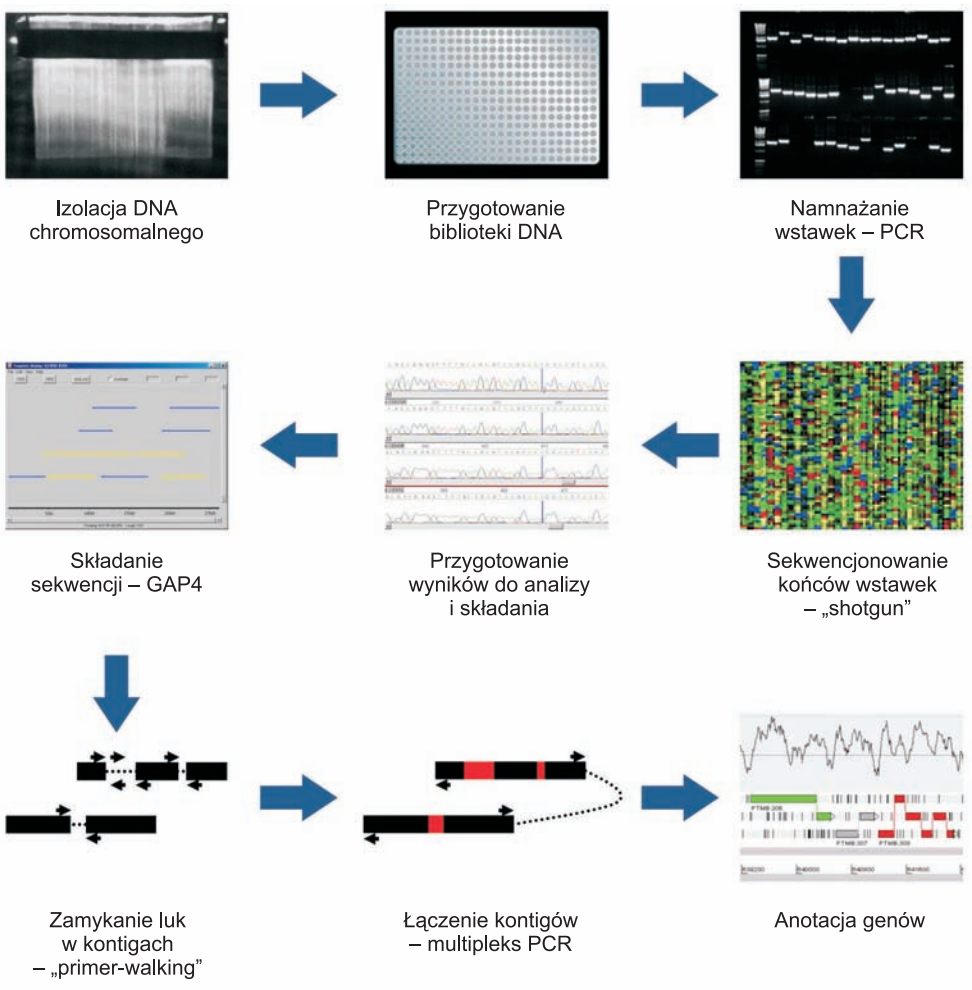
Dobre wyniki projektu pilotażowego skłoniły nas do podjęcia kolejnego kroku w kierunku poznania genomu *Paramecium*. Logicznym etapem było ustalenie sekwencji nukleotydowej jednego całego chromosomu, co pozwoliłoby na poznanie struktury chromosomów i zawartości genów. Pozwoliłoby zdobyć doświadczenie oraz oszacować koszty i potencjalne problemy, z jakimi przyjdzie się zmierzyć podczas projektu genomowego. Izolacja pojedynczego chromosomu nie była jednak łatwym zadaniem. Podczas powstawania makronukleusa chromosomy są fragmentowane i powstająca populacja około 200. chromosomów somatycznych jest wysoce heterogenna: chromosomy mogą występować w całości lub w kilku częściach, notuje się obecność wewnętrznych delecji. Podczas rozdzielania chromosomów metodą elektroforezy CHEF DNA migruje w postaci „rozmazu", trudno wyróżnić chromosomy. Jedynym fragmentem widocznym jako pojedynczy prążek, co pozwoliło na jego wyizolowanie, był fragment o wielkości ok. 1 Mpz odpowiadający największemu chromosomowi. Ze względu na postępującą wraz z kolejnymi podziałami wegetatywnymi fragmentację chromosomów, konieczne było wykorzystanie DNA z młodych komórek, które wykonały jedynie kilka podziałów po przejściu autogamii. Następnie przygotowano losową bibliotekę fragmentów restrykcyjnych o wielkości 1-3 kpz, którą poddano sekwencjonowaniu i składaniu wykorzystując strategię *shotgun*.

3.2. Ustalenie sekwencji nukleotydowej chromosomu

Pierwszy etap obejmował wykonanie około 12 600 reakcji sekwencjonowania końców wstawek namnożonych w reakcji PCR. Około 5000 odczytów zostało odrzuconych jako zanieczyszczenie pochodzące z innych chromosomów. Pozostałe 7600 odczytów złożono przy wykorzystaniu programu Gap4 (*Staden Package*) (11). W wyniku składania sekwencji uzyskano aż 180 kontigów od 3,5 do 85 kpz o sumarycznej długości 940 kpz, bliskiej spodziewanemu milionowi par zasad. Średnio sekwencja została pokryta odczytami czterokrotnie, jednak etap *shotgun* nie był wystarczający do ustalenia pełnej i nieprzerwanej sekwencji chromosomu.

Dzięki sekwencjonowaniu metodą przedłużania starterów (*primer-walking*) udało sie połączyć kontigi, dla których istniały wstawki, których końce znajdowały się w dwóch różnych kontigach. Pozostałe 24 przerwy zostały połączone metodą „multiplex PCR" (12). W metodzie tej projektuje się startery do PCR skierowane „na zewnątrz" każdego fragmentu. Następnie przeprowadza się reakcję PCR wykorzystując kilka, kilkanaście starterów jednocześnie oraz genomowe DNA jako matrycę.

Uzyskane produkty PCR następnie sekwencjonowano. W drugiej fazie projektu wykonano łącznie dodatkowe 1400 reakcji sekwencjonowania, co pozwoliło uzyskać nieprzerwaną sekwencję całego chromosomu. Blisko pięciokrotne pokrycie sekwencji gwarantowało bardzo niski statystyczny błąd sekwencjonowania wynoszący jedynie 1:700 000. Największy somatyczny chromosom *P. tetraurelia,* jak się okazało, był bogaty w zasady A i T (72%), co stanowi jedną z najwyższych wartości obserwowanych u eukariontów. Kolejną zaobserwowaną cechą DNA pantofelka okazała się niższa od oczekiwanej zawartość dinukleotydów CpG, która może świadczyć o intensywnej metylacji DNA i TpA – związana najprawdopodobniej z wykorzystaniem tego dinukleotydu przy rearanżacjach genomu podczas powstawania makronukleusa.



Rys. 2. Etapy sekwencjonowania chromosomu.

3.3. Potwierdzenie złożenia sekwencji

Złożenie odczytów w jedną sekwencję zostało potwierdzone na dwa sposoby. Po pierwsze, wykonano trawienie chromosomu enzymem BsiWI i porównano wyniki z konceptualnym (*in silico*) trawieniem sekwencji chromosomu uzyskując zgodność wielkości wszystkich fragmentów. Ponadto, cały chromosom został pokryty produktami reakcji PCR o długości 5-22 kpz „na zakładki". Produkty zostały następnie strawione różnymi enzymami, a uzyskane wzory restrykcyjne były zgodne z przewidywaniami. Należało jeszcze rozstrzygnąć, czy udało się poznać kompletną sekwencję chromosomu – od jednego telomeru do drugiego. Metoda konstrukcji banku genów wykluczała obecność w bibliotece DNA fragmentów zawierających powtórzenia telomerowe, jednak wykorzystanie wstępnych wyników z projektu sekwencjonowania całego genomu *Paramecium* pozwoliło odnaleźć telomery w pobliżu końców uzyskanej sekwencji. Ponadto zaobserwowano obecność 126-nukleotydowych powtórzeń przypominających motywy WD-40 białka G, które były wcześniej spotykane w pobliżu końców różnych chromosomów makronuklearnych (13).

3.4. Poszukiwanie genów – anotacja sekwencji nukleotydowej

Samo odczytanie kolejnych liter sekwencji nukleotydowej może dostarczyć pewnych danych do analizy *in silico*, jednak nie jest wystarczające, by uzyskany wynik mógł być użyteczny dla badaczy wykonujących eksperymenty *in vivo*. Kolejnym istotnym etapem było zatem poszukiwanie genów w obrębie ustalonej sekwencji. Ze względu na fakt, że w tym czasie nie były znane żadne genomy blisko spokrewnionych ewolucyjnie organizmów, w tym ani jednego przedstawiciela orzęsków, zadanie to było niezwykle trudne. Konieczne stało się opracowanie zupełnie nowych procedur prowadzących do anotacji chromosomu. Narzędzia bazowały m.in. na: porównaniu z bazą znanych białek (programem Blastx) (14), analizie zawartości nukleotydów G i C (jest ona dwukrotnie wyższa w rejonach kodujących), analizie częstości wykorzystania kodonów, algorytmie wyszukującym potencjalne introny (bazującym na znanych intronach), szukaniu paralogów w pierwszych sekwencjach z projektu sekwencjonowania całego genomu *Paramecium*. Ponadto wykorzystano program GlimmerM (15) służący do przewidywania genów *ab initio* na podstawie modelu sekwencji kodującej wygenerowanego na bazie znanych genów *Paramecium*. Program Artemis (16) pozwolił analizować wszystkie wymienione dane jednocześnie i decydować o strukturze genów. Końcowym etapem było potwierdzenie poprawności i poprawienie identyfikacji genów poprzez przyrównanie sekwencji potencjalnych białek do kilku najbardziej podobnych białek, poszukiwanie domen białkowych w bazach CDD i InterPro, poszukiwanie peptydów sygnałowych oraz przewidywanie helis transbłonowych i pęczków helis (*coiled-coil*).

Rys. 3. Anotacja genów. Podczas anotacji wykorzystywano program Artemis (16) pozwalający na jednoczesną wizualizację wykresów zawartości GC (**A**), wykresów wykorzystania najczęstszych kodonów na obu niciach DNA (**B**), wyników uzyskanych m. in. dzięki programom Blastx, GlimmerM, wyszukiwaniu intronów (**C**). (**D**) Przykładowy fragment chromosomu ze znalezionymi genami na obu niciach DNA. Uwagę zwracają krótkie odcinki międzygenowe.

Największy somatyczny chromosom *P. tetraurelia* zawiera niezwykle gęsto ułożone geny (średnia długość odcinków pomiędzy genami – 202 pz) z bardzo krótkimi intronami (średnia długość 25 pz). Zawartość sekwencji kodującej wynosi ponad 74%, co odpowiada najwyższej znanej gęstości genów dla wolno żyjących organizmów eukariotycznych. Dla 57% ze 464 zidentyfikowanych genów przypisano prawdopodobne funkcje przez porównanie sekwencji ich potencjalnych produktów do znanych sekwencji białkowych. Pozostałe zidentyfikowane geny kodują białka niewykazujące podobieństwa do żadnych znanych dotychczas białek.

3.5. Wnioski

W wyniku tego projektu sekwencjonowania poznano organizację i strukturę oraz ustalono blisko milion par zasad (1 Mb) sekwencji nukleotydowej największego somatycznego chromosomu *P. tetraurelia* (17).

Krótkie odcinki międzygenowe sugerowały istnienie grup współregulowanych genów kodujących białka zaangażowane w ten sam proces lub szlak metaboliczny. Takich grup jednak nie znaleziono. Na chromosomie znajduje się jedynie 6 par podobnych genów ułożonych obok siebie. Niewielki poziom identyczności aminokwasowej (25-56%) sugeruje, że może być to śladem dość starej ewolucyjnie duplikacji genów. Pośród białek kodowanych na chromosomie znaleziono wiele białek z motywem MORN, kinaz, metalofosfoesteraz, co zgodne jest z hipotezą, według której pantofelek ma znacząco rozbudowane szlaki przewodzenia sygnałów w komórce, co pozwala mu odpowiadać na zmieniające się warunki otoczenia. Ponadto w genomie znajduje się niezwykle dużo genów kanałów potasowych, które związane są z ruchem rzęsek. Najbardziej niespodziewany wynik dała ekstrapolacja liczby genów dla całego, szacowanego na 75 Mpz genomu. Przy założeniu, że zawartość sekwencji kodującej jest równie wysoka na innych chromosomach, łączna liczba genów *Paramecium* mogła być większa niż 30 000, co stawiałoby pantofelka pośród genomów z najwyższą liczbą genów.

4. Sekwencjonowanie całego genomu

Dane uzyskane w wyniku kolejnych projektów sekwencjonowania: pilotażowego (10) oraz projektu sekwencjonowania największego somatycznego chromosomu *P. tetraurelia* (17) okazały się podstawowe do uzyskania finansowania projektu sekwencjonowania całego genomu *Paramecium*. Konsorcjum GDRE „*Paramecium* Genomics" miało teraz argumenty pozwalające udowodnić, że zadanie to jest warte realizacji. Ostatecznie projekt przeprowadziło i sfinansowało Francuskie Narodowe Centrum Sekwencjonowania „Genoscope".

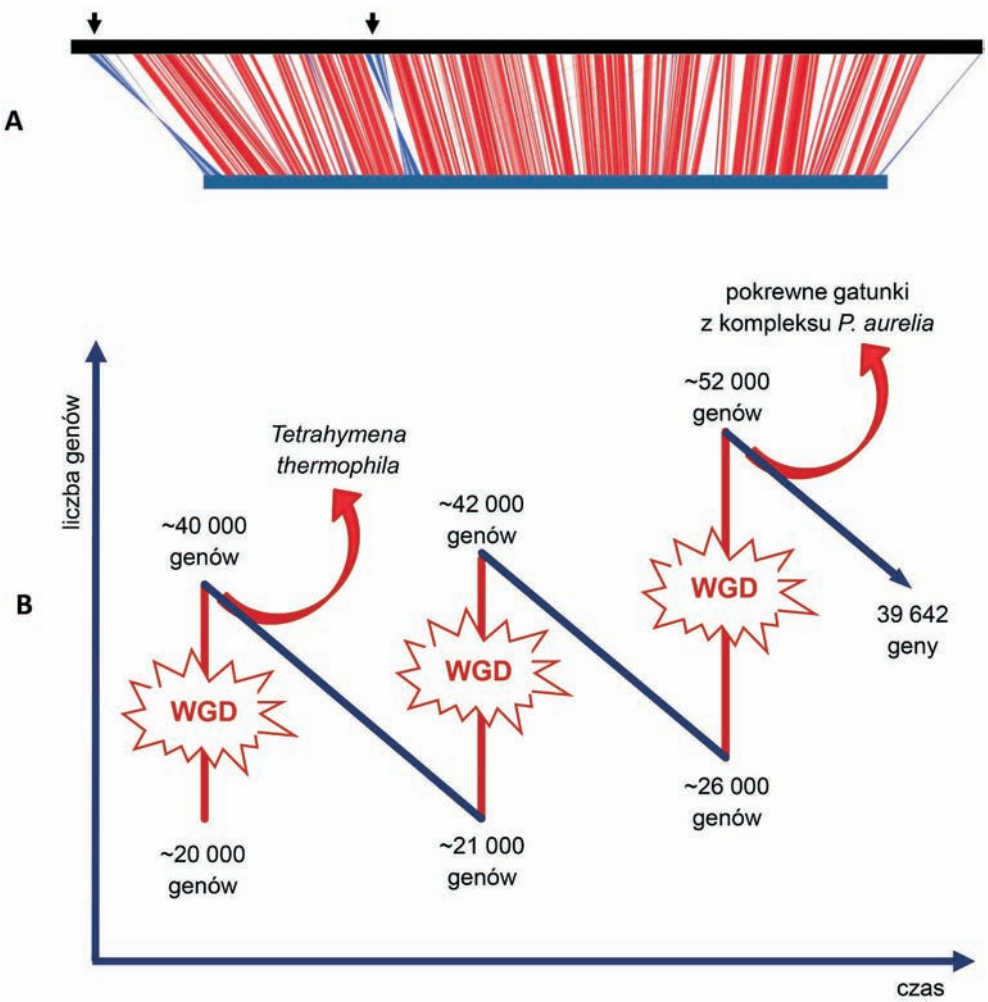
Znalezione w projekcie sekwencjonowania największego chromosomu geny posłużyły jako zestaw wyjściowy do stworzenia modeli genów *Paramecium* dla oprogramowania służącego do automatycznego poszukiwania genów. Ponadto sekwencja nukleotydowa tego chromosomu, będąca najdłuższym znanym fragmentem genomu pantofelka, została wykorzystana jako sekwencja referencyjna do sprawdzania poprawności złożenia sekwencji całego genomu, natomiast ekspercka anotacja genów kodowane na tym chromosomie została wykorzystana do oceny poprawności wyszukiwania genów na skalę genomową.

Uzyskana sekwencja somatycznego genomu *P. tetraurelia* obejmuje około 72 Mpz. Ze względu na fakt, że, podobnie jak dla większości realizowanych obecnie projektów sekwencjonowania, nie wykonywano pracochłonnego zamykania luk po-

zostałych w sekwencji po fazie *shotgun*, genom ma kształt 188 skafoldów (odpowiadających w większości całym chromosomom) zawierających statystycznie jedną lukę o długości 660 pz na 120 kpz (18). Cała sekwencja genomu wraz z rosnącą liczbą danych genetycznych i literaturowych jest dostępna w bazie danych ParameciumDB (http://paramecium.cgm.cnrs-gif.fr) (19). Obecnie prowadzone są prace nad uzyskaniem sekwencji genomu mikronuklearnego *P. tetraurelia*, która znacząco zwiększy wiedzę na temat eliminowanych sekwencji DNA, trwa też interesujące, z ewolucyjnego punktu widzenia, sekwencjonowanie innego blisko spokrewnionego gatunku *P. biaurelia*. Potwierdzając nasze przypuszczenia bazujące na wynikach z projektu sekwencjonowania największego chromosomu, w genomie *Paramecium* odnaleziono blisko 40 000 genów, co stanowi większą liczbę, niż liczba genów w genomie człowieka. Na podstawie dokładnej analizy sekwencji DNA wykazano, że genom *P. tetraurelia* uległ w toku ewolucji co najmniej trzem kolejnym duplikacjom całego genomu (WGD, ang. *Whole-Genome Duplication*), co tłumaczy tak dużą liczbę genów oraz występowanie wielu grup paralogów (rys. 4A). *Paramecium* znalazło się tym samym pośród organizmów, które służą do zaawansowanych badań dotyczących ewolucji będącej rezultatem duplikacji genomu: losu zduplikowanych genów, uzyskiwaniu nowych funkcji przez geny.

4.1. Duplikacje genomu w ewolucji organizmów

Duplikacja genomu jest rzadkim zjawiskiem, które jednak było już obserwowane u eukariontów, m.in. u *A. thaliana* czy *S. cerevisiae*. Postulowano, że duplikacje mogą być przyczyną najważniejszych zmian ewolucyjnych, gdyż podwojenie liczby genów może być źródłem nowych funkcji pozwalających na lepszą adaptację do warunków środowiska (20). Zgodnie z oczekiwaniami, na podstawie analizy trzech kolejnych duplikacji w genomie *Paramecium* wykazano, że dla większości zduplikowanych genów jedna kopia jest tracona po pewnym czasie, wynoszącym niekiedy miliony lat (rys. 4B). Jednak liczba genów obecnych w kilku kopiach jest znacznie większa niż u innych organizmów. Można to wytłumaczyć dualizmem jądrowym komórki *P. tetraurelia*, a w szczególności obecnością silnie poliploidalnego makronukleusa, w którym liczba kopii chromosomów jest regulowana. Ponadto, jak wcześniej postulowano, duplikacje genomu prowadzą do specjacji. Udowodniono, że rozdzielenie *Paramecium* od *Tetrahymena* nastąpiło po najstarszej WGD, a powstanie 15. siostrzanych gatunków *Paramecium* jest wynikiem ostatniej duplikacji. Kolejnym, nieoczekiwanym wnioskiem było wykazanie, że dwie kopie zduplikowanego genu są zachowywane w genomie nie w wyniku nabycia nowych funkcji przez te geny, lecz gdy poziom ekspresji danych genów ma znaczący wpływ na równowagę w komórce. Tak zatem, obie kopie genów zduplikowanych są zachowywane gdy wymaga tego utrzymanie równowagi pomiędzy innymi białkami, np. z tego samego kompleksu białkowego lub szlaku metabolicznego. Udało się także wykazać związek pomiędzy



Rys. 4. Duplikacje całego genomu *P. tetraurelia*. (**A**) Uzyskane za pomocą Artemis Comparison Tool (21) porównanie największego chromosomu (górna linia) oraz jego kopii z ostatniej duplikacji całego genomu (WGD) (linia dolna). Odcinki o dużej homologii sekwencji DNA są połączone liniami. Linie krzyżujące się (zaznaczone strzałkami) oznaczają odcinki o zmienionej orientacji – inwersje. (**B**) Na schemacie przedstawiono prawdopodobną ewolucję liczby genów *P. tetraurelia* w czasie. W okresach pomiędzy duplikacjami geny ulegały stopniowej pseudogenizacji i były sukcesywnie tracone. Specjacja będąca wynikiem duplikacji genomu została zaznaczona strzałkami – po najstarszej duplikacji oddzieliła się linia prowadząca do *T. thermophila*, po ostatniej powstały gatunki kompleksu *P. aurelia*. Pośrednia duplikacja jest trudna do umiejscowienia w czasie ewolucji.

poziomem ekspresji genów a zachowywaniem zduplikowanych genów w genomie – geny wyrażane na wysokim poziomie są rzadziej tracone.

5. Od sekwencji do funkcji – analiza funkcjonalna nowo odkrytych genów

Poznanie genomu *P. tetraurelia* pozwoliło na nowe, bardziej globalne podejście do badań. Jednym z bardziej interesujących tematów podejmowanych z wykorzystaniem tego organizmu są badania regulacji rearanżacji genomowych zachodzących podczas procesów płciowych. Celem kolejnego projektu realizowanego dzięki współpracy pomiędzy laboratoriami polskimi i francuskimi była identyfikacja nowych genów zaangażowanych w to zjawisko, gdyż w chwili obecnej znanych jest jedynie kilka czynników białkowych związanych z rearanżacjami genomu.

5.1. Rearanżacje genomu u *Paramecium tetraurelia*

Niezwykle duża skala oraz powtarzalność procesu rearanżacji DNA (rys. 1) czynią orzęski doskonałym organizmem modelowym do badania tego procesu. Odtworzenie funkcjonalnego makronukleusa jest kontrolowane epigenetycznie poprzez kilka klas niekodujących RNA, włączając w to krótkie scnRNA (skaningowe RNA) (22), podobne do kontrolujących transpozony w komórkach zwierzęcych piRNA oraz długie transkrypty (23), będące przykładem badanych obecnie intensywnie niegenowych transkryptów (24). Badanie ściśle regulowanych rearanżacji genomowych może przyczynić się do poszerzenia wiedzy na temat czynników zapewniających stabilność genomu, które odpowiedzialne są m.in. za różnicowanie się komórek. Najbardziej istotnym ze względu na znaczenie medyczne przykładem tego procesu jest rekombinacja V(D)J prowadząca do powstania regionów zmiennych genów kodujących immunoglobuliny i receptory powierzchniowe limfocytów B i T podczas dojrzewania układu odpornościowego kręgowców (25).

5.2. Poszukiwanie genów biorących udział w mejozie i rearanżacjach genomowych kodowanych na największym somatycznym chromosomie *P. tetraurelia*

Duża liczba nieznanych genów na nowo poznanym chromosomie wymusza wykorzystanie globalnej strategii poszukiwania funkcji poszczególnych genów. Dzięki współpracy z laboratoriami francuskimi mieliśmy szansę wykorzystania technologii mikromacierzy do badania profilu ekspresji wszystkich genów kodowanych na największym chromosomie *P. tetraurelia* (Nowak i wsp., dane nie publikowane). Do eksperymentu wykorzystano RNA wyizolowane z komórek wegetatywnych oraz komórek w różnych fazach procesu płciowego – autogamii. Na podstawie analizy statystycznej otrzymanych wyników wyodrębniono cztery grupy genów o podobnych profilach ekspresji: 1) geny z maksimum ekspresji na początku autogamii, 2) geny z maksimum ekspresji w późniejszej fazie autogamii, 3) geny podlegające represji,

4) geny podlegające stopniowej aktywacji. W celu weryfikacji danych wykonano sprawdzenie poziomu ekspresji wybranych genów w różnych fazach autogamii za pomocą metody Northern potwierdzając zasadność wyodrębnienia wymienionych grup genów. Uzyskane wyniki były dla nas punktem wyjścia do poszukiwania genów potencjalnie zaangażowanych w procesy związane z mejozą oraz w trakcie powstawania nowego makronukleusa w autogamii. W tym celu wykonano analizę fenotypów wyciszania dla 15. genów o maksimum ekspresji na początku oraz w późniejszej fazie autogamii. W eksperymentach wykorzystano technikę wyciszania genów na drodze interferencji RNA (RNAi) indukowanej poprzez karmienie komórek *Paramecium* bakteriami produkującymi dwuniciowe RNA homologiczne do badanego genu (26). Wykazano, że ekspresja czterech z badanych genów jest niezbędna w trakcie wzrostu wegetatywnego, a wyrażenie trzech innych jest konieczne do pomyślnego procesu autogamii. Tylko jeden z tych ostatnich genów, kodujący potencjalną helikazę RNA z motywem DExH, jest zaangażowany w rearanżacje genomowe. Dalsze badania dotyczące szczegółowej funkcji tego genu są obecnie w toku.

6. Podsumowanie

Znaczny wysiłek włożony w poznanie całych genomów znajduje uzasadnienie w efektach. Ustalenie sekwencji genomów, jak się wydaje, jest zasadnicze dla dalszego rozwoju szeroko rozumianej biologii molekularnej. Poznanie wszystkich genów genomu pozwala prowadzić badania nawet, wtedy gdy funkcje wszystkich genów nie są znane. Możliwe jest analizowanie sekwencji regulatorowych prowadząc do lepszego opisania funkcjonowania komórek na poziomie molekularnym. Jest to również prosta droga (i o dziwo, dość tania) do wyodrębnienia genów istotnych ze względów medycznych lub biotechnologicznych, które można wybrać z poznanego „katalogu" znacznie łatwiej, niż poszukiwać w nieznanym genomie. Sekwencjonowanie otwiera drogę do analiz biologicznych na dużą skalę, pozwalających analizować funkcjonowanie komórki czy organizmu jako całości – jak np. poznanie ekspresji genów z wykorzystaniem mikromacierzy lub sekwencjonowania transkryptów (RNA-seq).

Dla badacza wykorzystującego sekwencje DNA do badań laboratoryjnych zasadnicza jest jakość sekwencji i jakość jej anotacji. Błędy w sekwencji DNA, które łatwo znaleźć w sekwencjach zdeponowanych w bazach danych, mogą doprowadzić do poważnych problemów w prowadzonych badaniach, podobnie jak nieprawidłowy lub niekompletny opis sekwencji może doprowadzić do niewłaściwych wniosków. Wydaje mi się, że w przypadku genomu *Paramecium* udało się utrzymać odpowiednie standardy, a sekwencja i jej anotacja są nieustanie poprawiane i aktualizowane w bazie danych ParameciumDB. Poznanie sekwencji chromosomu, a później całego genomu, pozwoliło już – i pozwoli w najbliższej przyszłości – dokonać szeregu odkryć, zapoczątkowując erę „postgenomową" w badaniach z wykorzystaniem

Paramecium. Możliwe stało się na przykład odnalezienie w bardzo krótkim czasie od opublikowania sekwencji genomu poszukiwanego od wielu lat białka odpowiedzialnego za wprowadzanie pęknięć DNA podczas rearanżacji genomowych (27).

Słowniczek terminów i skrótów

nt – nukleotyd.
kpz – tysiąc par zasad.
Mpz – milion par zasad.
Strategia shotgun – strategia sekwencjonowania polegająca na losowej fragmentacji sekwencjonowanego fragmentu, sekwencjonowaniu końców powstałych fragmentów, a następnie złożeniu sekwencji z wykorzystaniem zachodzących na siebie sekwencji.
Kontig – nieprzerwana sekwencja DNA złożonych z wielu pojedynczych, zachodzących na siebie odczytów sekwencjonowania.
Skafold – ang. scaffold – „rusztowanie"; połączenie kontigów w większą całość; w odróżnieniu od kontigów, może zawierać luki w sekwencji nukleotydowej.
Anotacja – polega na oznaczeniu w sekwencji DNA informacji przydatnych do dalszych badań, jak np. jednostek strukturalnych (geny, promotory, introny, sekwencje repetytywne, niekodujące RNA, etc.) oraz informacji dotyczących samej sekwencji nukleotydowej (znane mutacje, polimorfizmy sekwencji, potencjalne błędy sekwencjonowania).
Anotacja genów *ab initio* – metoda identyfikacji genów *in silico* oparta m.in. na ukrytych modelach Markowa (HMM) – używają one statystycznej informacji na temat miejsc splicingu, wykorzystania kodonów, długości eksonów i intronów, etc., opartej na znanym katalogu genów danego organizmu.

Osoby zaangażowane w nie publikowane badania opisane w rozdz. 5.2. Nowak J. K., Gromadka R., Juszczuk M., Jerka-Dziadosz M., Maliszewska K., Mucchielli M.-H., Gout J.-F., Arnaiz O., Agier N., Tang T., Aggerbeck L., Cohen J., Delacroix H., Sperling L., Herbert C., Zagulski M., Bétermier M.

Wyniki omówione w pracy były finansowane w części z grantów Ministerstwa Nauki i Szkolnictwa Wyższego: 3 PO4A 006 25 i N303 075 32/2520.

Literatura

1. Beisson J., Bétermier M., Bré M.-H., Cohen J., Duharcourt S., Duret L., Kung C., Malinsky S., Meyer E., Preer J. R. Jr, Sperling L., (2010), Cold Spring Harb Protoc, 2010doi:10.1101/pdb.emo140
2. Caron F., Meyer E., (1985), Nature, 314(6007), 185-188.
3. Preer J. R. Jr., Preer L. B., Rudman B. M., Barnett A. J., (1985), Nature, 314(6007), 188-190.
4. Greider C. W., Blackburn E. H., (1985), Cell, 43(2 Pt 1), 405-413.
5. Jaillon O., Bouhouche K., Gout J., Aury J., Noel B., Saudemont B., Nowacki M., Serrano V., Porcel B., Ségurens B., Le Mouël A., Lepère G., Schächter V., Bétermier M., Cohen J., Wincker P., Sperling L., Duret L., Meyer E., (2008), Nature, 451 (7176), 359-362.
6. Duharcourt S., Lepère G., Meyer E., (2009), Trends Genet., 25(8), 344-350.
7. Gratias A., Betermier M., (2001), Biochimie, 83, 1009-1022.
8. Yao M. C., Duharcourt S., Chalker D. L., (2002), *Mobile DNA II*, Eds. Craig N. L., Craigie R., Gellert M., Lambowitz A. M., ASM Press, Washington DC.
9. Dessen P., Zagulski M., Gromadka R., Plattner H., Kissmehl R., Meyer E., Betermier M., Schultz J. E., Linder J. U., Pearlman R. E., Kung C., Forney J., Satir B. H., van Houten J. L., Keller A. M., Froissard M., Sperling L., Cohen J., (2001), Trends Genet., 17(6), 306-308.
10. Sperling L., Dessen P., Zagulski M., Pearlman R. E., Migdalski A., Gromadka R., Froissard M., Keller A. M., Cohen J., (2002), Eukaryot. Cell., 3, 341-352.

11. Staden R., (1996), Mol. Biotechnol., 5, 233-241.
12. Tettelin H., Radune D., Kasif S., Khouri H., Salzberg S., (1999), Genomics, 62, 500-507.
13. Forney J., Rodkey K., (1992), Nucleic Acids Res., 20, 5397-5402.
14. Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J., (1990), J. Mol. Biol., 215, 403-410.
15. Salzberg S., Pertea M., Delcher A., Gardner M., Tettelin H., (1999), Genomics, 59, 24-31.
16. Rutherford K., Parkhill J., Crook J., Horsnell T., Rice P., Rajandream M. A., Barrell B., Bioinformatics, (2000), 16, 944-945.
17. Zagulski M., Nowak J. K., Le Mouel A., Nowacki M., Migdalski A., Gromadka R., Noel B., Blanc I., Dessen P., Wincker P., Keller A. M., Cohen J., Meyer E., Sperling L., (2004), Curr. Biol., 15, 1397-1404.
18. Aury J. M., Jaillon O., Duret L., Noel B., Jubin C., Porcel B. M., Ségurens B., Daubin V., Anthouard V., Aiach N., Arnaiz O., Billaut A., Beisson J., Blanc I., Bouhouche K., Câmara F., Duharcourt S., Guigo R., Gogendeau D., Katinka M., Keller A.-M., Kissmehl R., Klotz C., Koll F., Le Mouël A., Lepère G., Malinsky S., Nowacki M., Nowak J. K., Plattner H., Poulain J., Ruiz F., Serrano V., Zagulski M., Dessen P., Bétermier M., Weissenbach J., Scarpelli C., Schächter V., Sperling L., Meyer E., Cohen J., Wincker P., (2006), Nature, 444 (7116), 171-178.
19. Arnaiz O., Cain S., Cohen J., Sperling L., (2007), Nucleic Acids Res., 35 (Database issue), D439-444.
20. Ohno S., (1970), *Evolution by Gene Duplication*, Ed. Allen & Unwin, London.
21. Carver T. J., Rutherford K. M., Berriman M., Rajandream M. A., Barrell B. G., Parkhill J., (2005), 21, 3422-3423.
22. Lepère G., Nowacki M., Serrano V., Gout J.-F., Guglielmi G., Duharcourt S., Meyer E., (2009), Nucleic Acids Res., 37(3), 903-915.
23. Lepère G., Bétermier M., Meyer E., Duharcourt S., (2008), Genes Dev., 22(11), 1501-1512.
24. Ponting C. P., Oliver P. L., Reik W., (2009), Cell, 136, 629-641.
25. Dudley D. D., Chaudhuri J., Bassing C. H., Alt F. W., (2005), Adv. Immunol., 86, 43-112.
26. Galvani A., Sperling L., (2001), Nucleic Acids Res., 29, 4387-4394.
27. Baudry C., Malinsky S., Restituito M., Kapusta A., Rosa S., Meyer E., Bétermier M., (2009), Genes Dev., 23, 2478-2483.