UNIWERSYTET IM. ADAMA MICKIEWICZA W POZNANIU

Wydział Biologii

Laboratorium Technik Biologii Molekularnej

Rozprawa Doktorska

Opracowanie zestawu polimorficznych markerów STR do genotypowania ludzi oraz jego wdrożenie do analiz pokrewieństwa w dalszych relacjach rodzinnych

Wojciech Łuczak

Rozprawa doktorska wykonana pod kierunkiem:

prof. UAM dr hab. Mirosławy Dabert

Laboratorium Technik Biologii Molekularnej, Wydział Biologii

Uniwersytet im. Adama Mickiewicza w Poznaniu

Doktorat wdrożeniowy zrealizowany pod kierunkiem opiekuna pomocniczego z ramienia Laboratorium Diagnostyki Molekularnej GenMed sp.j.:

dr hab. Katarzyny Świerczyńskiej

Poznań, 2024

Składam serdeczne podziękowania:

Pani prof. UAM dr hab. Mirosławie Dabert

Za danie mi możliwości realizacji moich badań, wsparcie, przekazaną wiedzę

i ogromne zaangażowanie.

Praca z Panią Profesor była dla mnie wielką przyjemnością.

Dziękuję moim wspólnikom: Katarzynie Świerczyńskiej i Jakubowi Świerczyńskiemu za wsparcie i wiarę w moje możliwości.

Dziękuję mojemu współpracownikowi Dawidowi Leciejowi za nieoceniony wkład w realizację projektu.

Dziękuję moim współpracownikom z Laboratorium Diagnostyki Molekularnej GenMed i studentom Wydziału Biologii UAM: Jagodzie Pazdioara, Michałowi Radaszewskiemu, Łukaszowi Skibińskiemu, Pauli Paczesnej, Oliwii Różańskiej, Jakubowi Śnieciowi, Monice Langner, Natalii Olszewskiej, Szymonowi Błaszczykowi, Bognie Juskowiak za pomoc w realizacji prac badawczych i wdrożeniu metody Kinfinder do codziennej pracy Laboratorium.

Dziękuję mojej partnerce Barbarze Urbańskiej za pomoc, cierpliwość i wyrozumiałość.

Dziękuję moim rodzicom: Urszuli Łuczak oraz Stefanowi Łuczakowi za bezgraniczne wsparcie, motywację i pomoc na każdym etapie mojej edukacji.

Finansowanie

Niniejsza praca powstała przy finansowym udziale:

1. Ministerstwa Nauki i Szkolnictwa Wyższego w ramach programu “Doktorat Wdrożeniowy” edycja IV.
2. Przedsiębiorstwa Laboratorium Diagnostyki Molekularnej Genmed sp.j. J. Świerczyński.

Streszczenie

Analiza długości sekwencji mikrosatelitarnych, czyli krótkich powtórzeń tandemowych (ang. STR – Short Tandem Repeats), jest jedną z najważniejszych metod stosowanych w genetyce sądowej, w tym w badaniach biologicznego pokrewieństwa. Jej powszechność stosowania wynika z wysokiej informatywności, czułości oraz możliwości badania zdegradowanego DNA. Dodatkowo, metoda ta charakteryzuje się niskim kosztem jednostkowym analizy i krótkim czasem realizacji badań. Rutynowa analiza loci STR w laboratoriach genetyczno-sądowych opiera się na technice multipleks-PCR, wykorzystującej fluorescencyjnie znakowane startery, oraz na elektroforezie kapilarnej, która umożliwia określenie długości zamplifikowanych alleli. W badaniach kryminalistycznych i analizach pokrewieństwa najczęściej stosuje się technologię pozwalającą na jednoczesną analizę dwudziestu loci STR z systemu CODIS (ang. Combined DNA Index System) oraz dodatkowych loci STR, takich jak SE33, PENTA D i PENTA E, przy użyciu komercyjnych zestawów odczynników. Analiza dwudziestu loci STR systemu CODIS jest zazwyczaj wystarczająca w stosunkowo prostych sprawach dotyczących badania biologicznego ojcostwa lub macierzyństwa, jednak okazuje się być niewystarczająco informatywna w przypadku badania pokrewieństwa w relacjach pokrewieństwa drugiego i trzeciego stopnia.

Celem niniejszego doktoratu wdrożeniowego było opracowanie metody Kinfinder do badania biologicznego pokrewieństwa w dalszych relacjach rodzinnych. Metoda miała zostać wdrożona w Laboratorium Diagnostyki Molekularnej GenMed, w związku z czym musiała opierać się na technologii reakcji multipleks-PCR oraz elektroforezy kapilarnej. Takie podejście pozwoliłoby na stosowanie metody przez firmę bez konieczności ponoszenia dodatkowych kosztów inwestycyjnych w nową aparaturę, przestrzeń laboratoryjną oraz w szkolenia pracowników.

Projekt opracowania metody Kinfinder obejmował analizę bioinformatyczną genomu ludzkiego, w celu identyfikacji 150-200 najbardziej polimorficznych loci STR, następnie zaprojektowanie starterów do reakcji PCR do amplifikacji każdego z wybranych loci i przeprowadzenie laboratoryjnych badań przesiewowych loci STR pozwalających na oszacowanie heterozygotyczności loci w populacji polskiej. W kolejnym etapie wybrano 50

najbardziej polimorficznych loci i opracowano dwie reakcje multipleks-PCR do genotypowania ludzkiego DNA w zakresie wybranych loci STR. Po zaprojektowaniu i walidacji metody, dysponując możliwością analizy 50 loci w dwóch reakcjach multipleks-PCR, przeprowadzono badania populacyjne mające na celu wyznaczenie częstości występowania alleli każdego locus w populacji polskiej. Badania opisane w niniejszej pracy objęły także sekwencjonowanie loci w celu powiązania długości sekwencji tandemowo powtórzonej z długością amplikonów, a także objęły stworzenie drabiny allelicznej oraz odczynnika do kalibracji analizatora genetycznego. Badania populacyjne potwierdziły bardzo wysoką polimorficzność opracowanych 50 loci STR. Ich średnia heterozygotyczność wynosiła 88,07% co w porównaniu ze średnią heterozygotycznością loci CODIS wynoszącą w polskiej populacji 78,95% jest wyraźnie wyższą wartością. Najbardziej polimorficzne loci w zestawie Kinfinder (D8A26, D15L495, D13S742) cechowały się heterozygotycznością odpowiednio 93,45%, 93,53%, 93,9% co odpowiada najbardziej polimorficznemu locus STR wykorzystywanego w genetyce sądowej (SE33), którego heterozygotyczność w populacji polskiej wynosi od 93,4% do 95,4% w zależności od źródeł literaturowych.

Wyniki badań wskazują, że wykorzystane dane wstępne oraz opracowana od podstaw metoda szacowania polimorfizmu loci STR pozwoliły na weryfikację heterozygotyczności wybranych loci w populacji polskiej. Zgodnie z założeniami, potwierdzono, że heterozygotyczność części loci STR jest wyższa niż ta zgłoszona w bazach projektu 1000 Genomes oraz w bazach STRCatalog i WebSTR, co uwidoczniło ograniczenia technologii sekwencjonowania krótkich odczytów zastosowanej w projekcie 1000 Genomes w analizie polimorfizmu sekwencji mikrosatelitarnych w genomie ludzkim.

Wyniki analiz statystycznych badań pokrewieństwa w różnych relacjach rodzinnych potwierdziły wysoką informatywność metody Kinfinder i jej przydatność zwłaszcza w badaniach pokrewieństwa drugiego i trzeciego stopnia. Zgodnie z założeniami doktoratu wdrożeniowego, metoda Kinfinder została włączona do standardowych procedur Laboratorium Diagnostyki Molekularnej GenMed i została już zastosowana w ponad trzydziestu badaniach pokrewieństwa w tym w sprawach sądowych.

Abstract

The analysis of microsatellite sequences, also known as Short Tandem Repeats (STR), is one of the most important methods used in forensic genetics, including in biological kinship testing. Its widespread use results from its high informativeness, sensitivity, and the ability to analyze degraded DNA. Additionally, this method is characterized by low per-sample cost and a short turnaround time for analysis. Routine STR locus analysis in forensic genetics laboratories relies on the multiplex-PCR technique, which uses fluorescently labeled primers, and capillary electrophoresis, which enables the determination of the lengths of amplified alleles. In forensic investigations and kinship analysis, technology allowing for the simultaneous analysis of twenty CODIS STR loci (Combined DNA Index System) and additional STR loci, such as SE33, PENTA D, and PENTA E, using commercial reagent kits is commonly employed. The analysis of twenty CODIS STR loci is sufficient for relatively simple cases of biological paternity or maternity testing, but it proves to be insufficiently informative for kinship testing in second and third-degree familial relationships.

The aim of this implementation PhD was to develop the KinFinder method for investigating biological kinship in more distant familial relationships. The method was intended to be implemented in the GenMed Molecular Diagnostics Laboratory, and therefore, it needed to be based on the technology of multiplex-PCR and capillary electrophoresis. This approach would allow the company to use the method without incurring additional investment costs for new equipment, laboratory space, or employee training.

The development of the Kinfinder method involved a bioinformatic analysis of the human genome to identify 150-200 of the most polymorphic STR loci, followed by the design of PCR primers for the amplification of each selected locus, and laboratory screening of STR loci to estimate the heterozygosity of these loci in the Polish population. In the next phase, 50 of the most polymorphic loci were selected, and two multiplex-PCR reactions were developed for genotyping human DNA at the selected STR loci. After the design and validation of the method, with the ability to analyze 50 loci in two multiplex-PCR reactions, population studies were conducted to determine the allele frequencies of each locus in the Polish population.

The research described in this dissertation also included sequencing of the loci to link the length of the tandem repeat sequence to the length of the amplicons, as well as the creation of an allelic ladder and a reagent for calibrating the genetic analyzer. Population studies confirmed the very high polymorphism of the 50 developed STR loci. Their average heterozygosity was 88.07%, which is significantly higher compared to the average heterozygosity of CODIS loci in the Polish population, which is 78.95%. The most polymorphic loci in the KinFinder set (D8A26, D15L495, D13S742) had heterozygosity rates of 93.45%, 93.53%, and 93.9%, respectively, comparable to the most polymorphic STR locus used in forensic genetics (SE33), whose heterozygosity in the Polish population ranges from 93.4% to 95.4%, depending on the literature reports.

The results of the study show that the preliminary data used, as well as the newly developed method for estimating STR locus polymorphism, allowed for the verification of the heterozygosity of selected loci in the Polish population. As predicted, it was confirmed that the heterozygosity of some STR loci is higher than that reported in the 1000 Genomes project database and the STRCatalog and WebSTR databases, highlighting the limitations of the short-read sequencing technology used in the 1000 Genomes project for analyzing microsatellite sequence polymorphism in the human genome.

The statistical analysis of kinship tests in various familial relationships confirmed the high informativeness of the KinFinder method and its usefulness, particularly in second and third-degree kinship testing. In line with the objectives of this implementation PhD, the KinFinder method has been incorporated into the standard procedures of the GenMed Molecular Diagnostics Laboratory and has already been applied in over thirty kinship tests, including court cases.

Spis treści

1. Wstęp

1.1. Loci mikrosatelitarne genomu człowieka

Mikrosatelity, czyli proste powtórzenia sekwencji (ang. SSRs - Simple Sequence Repeats) nazywane też krótkimi powtórzeniami tandemowymi (ang. STRs - Short Tandem Repeats), to sekwencje DNA zawierające krótkie motywy ułożonych tandemowo od jednego do sześciu nukleotydów. Ten typ sekwencji jest charakterystyczny dla genomów organizmów eukariotycznych (Li i in., 2002; Zane i in., 2002). W obszarach mikrosatelitarnych motywy powtarzają się od kilku do kilkuset razy, przy czym liczba powtórzeń jest wysoce zmienna (Selkoe i Toonen, 2006). W ludzkim genomie liczba sekwencji STR przekracza milion i zajmują one ok. 3% jego długości (Sawaya i in., 2013). Najczęściej występującymi mikrosatelitami w genomie ludzkim są te z powtórzeniami dinukleotydowymi. Sekwencje z powtórzeniami mono- i tetranukleotydowymi są rzadsze, natomiast powtórzenia tri-, penta-, oraz heksaukleotydowe są najmniej liczne (Ellegren, 2004).

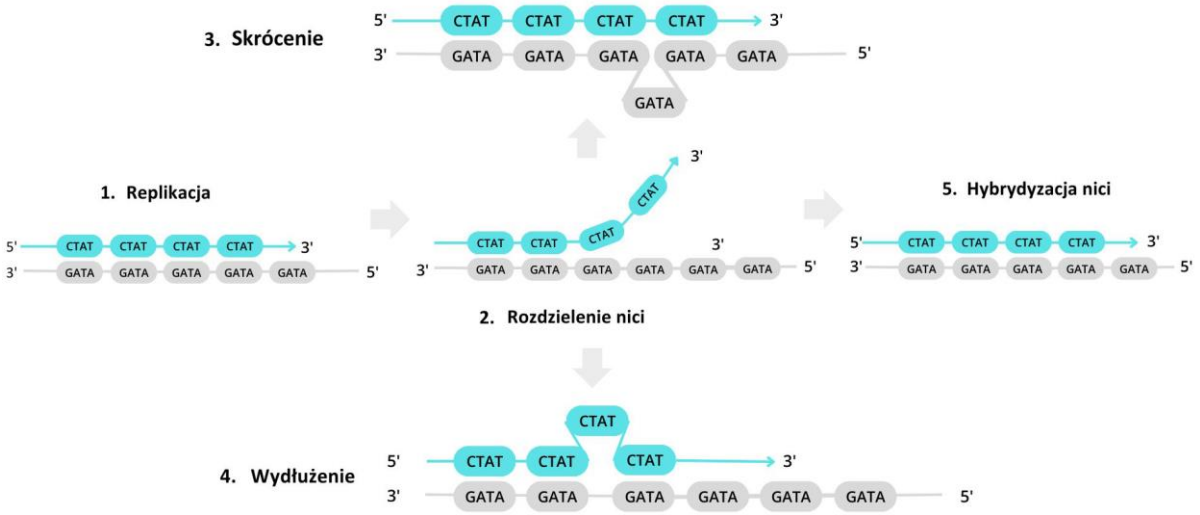
Ze względu na biologiczne właściwości sekwencji mikrosatelitarnych, takie jak dziedziczenie mendlowskie, wysoka polimorficzność, wysokie tempo mutacji oraz obecność dużej liczby tych sekwencji w genomie, są one wykorzystywane w wielu obszarach biologii, m. in. w genetyce sądowej do identyfikacji osób, w ustalaniu pokrewieństwa (np. Zhang i in., 2020; Wilkening i in., 2006) oraz w analizach kryminalistycznych (np. Nwawuba Stanley i in., 2020), w genetyce populacyjnej do badania różnorodności genetycznej, struktury populacji oraz procesów ewolucyjnych (np. Schlötterer, 2000), w hodowli roślin i zwierząt w celu identyfikacji cech pożądanych i monitorowania linii genetycznych (Xu i in., 2017). Ponadto analiza sekwencji mikrosatelitarnych znajduje zastosowanie w genetyce konserwatorskiej do badania genetyki zagrożonych gatunków, co pomaga w planowaniu strategii ochrony i zarządzania populacjami (Abdul-Muneer, 2014), w biologii ewolucyjnej w badaniach mechanizmów ewolucyjnych oraz zjawisk takich jak dryf genetyczny i dobór naturalny (Takezaki i Nei, 2009), a także w badaniach antropologicznych i archeologicznych gdzie są stosowane do analizy różnorodności genetycznej wśród dawnych populacji ludzkich, co pomaga w badaniach nad pochodzeniem i migracją ludzi (Slatkini i Racimo, 2016). Analiza loci STR znajduje również zastosowanie w diagnostyce nowotworów (Boland i Goel, 2010),

monitorowaniu terapii oraz w diagnozowaniu wrodzonych chorób genetycznych takich jak ataksje rdzeniowo-móżdżkowe czy pląsawica Huntingtona (Brouwer i in., 2009).

Termin “satelitarny DNA” został po raz pierwszy użyty w 1961 r do opisania frakcji DNA wykazującej inną gęstość w trakcie wirowania DNA w gradiencie chlorku cezu, przez co wyraźnie odróżniającej się od głównej frakcji DNA (Kit., 1961). Termin „mikrosatelita” został wprowadzony znacznie później, w 1989 roku do opisania w genomie człowieka sekwencji tandemowo powtórzonych zawierających motyw TG (Litt i Luty, 1989). Autorzy publikacji jednocześnie wykazali znaczny polimorfizm locus zawierającego ten motyw w obrębie badanej grupy osób. Pierwsza sekwencja fragmentu ludzkiego genomu, zawierająca powtórzenia tandemowe, została opisana wcześniej, w 1984 r. (Weller et al 1984 r.). Sekwencja zawierała powtórzony tandemowo motyw GGAT znajdujący się w genie mioglobiny. Funkcjonalny ludzki gen mioglobiny ma długość ok. 10,4 tys. par zasad (pz) i posiada strukturę podobną do hemoglobiny, z trzema eksonami i dwoma intronami oraz długimi regionami niekodującymi. Sekwencja STR zawierająca ok. 165 powtórzeń motywu GGAT położona jest w sekwencji regulatorowej 1100-1750 pz powyżej sekwencji kodującej genu. Postęp w technikach biologii molekularnej, takich jak PCR i sekwencjonowanie, znacząco przyspieszył poznawanie genomu ludzkiego oraz odkrywanie licznych sekwencji mikrosatelitarnych w różnych jego obszarach.

Sekwencje mikrosatelitarne mutują z częstością o kilka rzędów wielkości wyższą (10^{-2}-10^{-6}) w porównaniu do mutacji punktowych w innych rejonach genomu (Hodel i in., 2016). W przeciwieństwie do mutacji punktowych, które wpływają jedynie na pojedynczy nukleotyd, mutacje mikrosatelitarne prowadzą do zysku lub utraty całej jednostki powtórzonej tandemowo, a w niektórych przypadkach dwóch lub więcej powtórzeń. Tempo mutacji loci mikrosatelitarnych zależy od sekwencji i długości motywu powtórzonego oraz liczby jednostek powtórzonych (Gymrek i in., 2017). Zmiany długości w DNA mikrosatelitarnym są zazwyczaj uznawane za wynik tzw. poślizgów replikacyjnych — to znaczy przejściowego rozłączenia replikujących się nici DNA, a następnie ich niewłaściwego ponownego połączenia (Schlötterer i Tautz, 1992). Kiedy nowa nić przyłączy się w niewłaściwej pozycji, kontynuacja replikacji prowadzi do wstawienia lub usunięcia jednostek powtórzeń w porównaniu do nici

matrycowej (Ryc. 1).

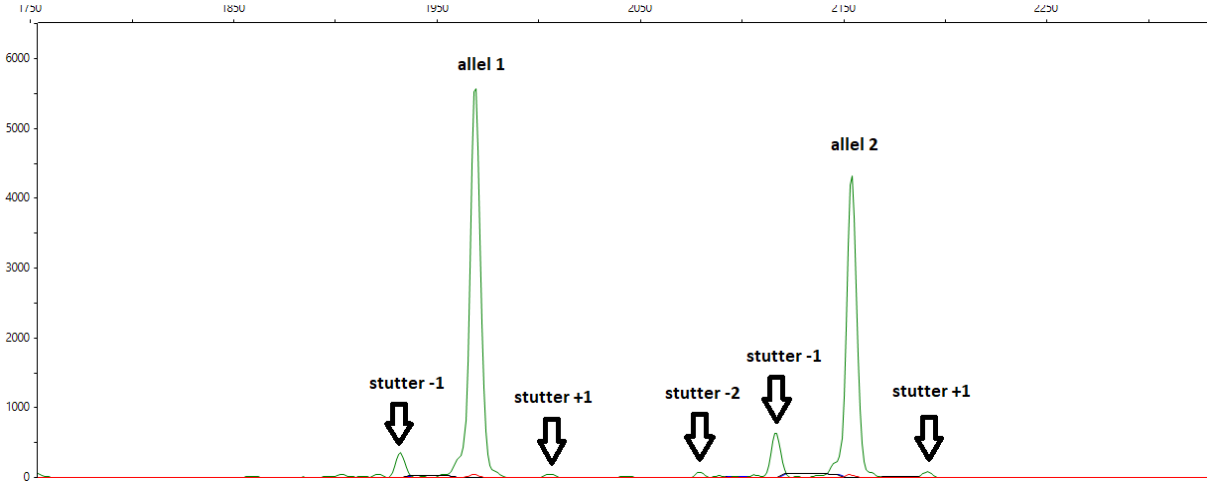


Ryc. 1. Mechanizm powstawania mutacji sekwencji STR. Podczas replikacji sekwencje DNA są połączone, a polimeraza DNA rozpoznaje i wiąże się z końcem 3' nici DNA, co pozwala jej na syntezę identycznej komplementarnej kopii (1). Czasami nici ulegają dysocjacji, kompleks polimerazy DNA odpada, a replikacja zatrzymuje się (2). Gdy dwie nici DNA ponownie się łączą, zazwyczaj dopasowują się poprawnie (5). Jednak czasami, z powodu powtarzających się sekwencji (GATA), nici mogą ulec błędnemu dopasowaniu, co prowadzi do skrócenia potomnej nici o jedno powtórzenie tandemowe (3) lub do wydłużenia nici potomnej o jedno powtórzenie tandemowe (4).

Większość pierwotnych mutacji tego typu jest naprawiana przez system naprawy błędnie sparowanych zasad, a jedynie mała część, która nie została naprawiona, staje się rzeczywistymi zdarzeniami mutacji mikrosatelitarnych. Eksperymenty in vitro z wykorzystaniem oczyszczonych enzymów eukariotycznych lub prokariotycznych potwierdzają, że polimeraza DNA jest jedyną aktywnością enzymatyczną potrzebną do powstania zjawiska poślizgu (Schlötterer i Tautz, 1992). Wysokie tempo mutacji mikrosatelit koreluje pozytywnie z polimorficznością tych sekwencji w populacji co jest wykorzystywane w badaniach pokrewieństwa oraz analizie kryminalistycznej.

Poślizg replikacyjny występuje również in vitro podczas amplifikacji sekwencji mikrosatelitarnych metodą PCR (Ryc. 2). Charakterystyczną cechą takich amplifikacji jest obecność amplikonów nazywanych „stutter" – czyli produktów ubocznych, które różnią się długością od głównego produktu o wielokrotność długości jednostki powtórzonej (Hauge i Litt, 1993: Murray i in., 1993). Badania ilościowe pokazują, że częstość poślizgów polimerazy

Taq rośnie wraz z liczbą jednostek powtórzonych i jest odwrotnie skorelowana z długością jednostki powtórzonej. Wiele osób zajmujących się mikrosatelitami zaobserwowało amplikony typu stutter podczas amplifikacji loci STR; markery powtórzeń tetranukleotydowych generują zazwyczaj mniej produktów stutter niż powtórzenia dinukleotydowe, a zwłaszcza mononukleotydowe. Amplikony typu stutter zazwyczaj pojawiają się jako produkty krótsze od amplifikowanego allelu, co sugeruje, że tempo mutacji polegającej na skróceniu sekwencji przez polimerazę Taq jest znacznie wyższe niż tempo mutacji prowadzących do jej wydłużenia (Shinde i in., 2003).



Ryc. 2. Elektroforegram przedstawiający amplifikację locus D13S317 zawierającego powtórzenia tetranukleotydowe. Strzałkami oznaczono amplikony typu stutter będące niespecyficznymi produktami reakcji PCR. Stutter -2 jest amplikonem skróconym o dwie jednostki powtórzone, stutter -1 amplikonem skróconym o jedną jednostkę, natomiast stutter +1 jest amplikonem zawierającym dodatkową jednostkę sekwencji powtórzonej.

1.2 Zastosowania analizy loci STR w badaniach biologicznego pokrewieństwa

Ze względu na wysoki polimorfizm sekwencji mikrosatelitarnych, łatwość analizy długości sekwencji metodą PCR oraz stosunkowo niski koszt badania, analiza loci STR jest metodą pierwszego wyboru w genetyce kryminalistycznej oraz rutynowych testach pokrewieństwa. Dominującą technologią analizy alleli w loci STR, zarówno w badaniach kryminalistycznych jak i badaniach pokrewieństwa, jest analiza dwudziestu tzw. loci CODIS (ang. COmbined DNA Index System) oraz kilku innych markerów, w tym: SE33, PENTA D, PENTA E (np. Shrivastava i in., 2021) przy użyciu komercyjnie dostępnych zestawów reagentów takich jak zestaw

GlobalFiler™ PCR Amplification Kit firmy Applied Biosystems lub zestaw PowerPlex® Fusion 6C System firmy Promega. Zestawy te wykorzystują reakcję multipleks-PCR i elektroforezę kapilarną do analizy długości amplifikowanych produktów DNA.

Analiza długości amplifikowanych fragmentów DNA zawartych w loci STR pozwala na precyzyjne określenie liczby tandemowych powtórzeń w polimorficznej sekwencji mikrosatelitarnej, a tym samym na ustalenie profilu genetycznego badanej osoby, który można później porównać z profilem DNA izolowanym z materiału biologicznego lub z profilem genetycznym innej osoby w celu ustalenia pokrewieństwa lub tożsamości. Analiza dwudziestu loci STR systemu CODIS jest wystarczająca w stosunkowo prostych sprawach dotyczących badania biologicznego ojcostwa lub macierzyństwa, jednak może być niewystarczająca w przypadku badania pokrewieństwa w dalszych relacjach rodzinnych (Pedroza Matute i Iyavoo, 2024). Trudność w uzyskaniu pewnych i jednoznacznych wyników dotyczących biologicznego pokrewieństwa w dalszych relacjach rodzinnych, innych niż rodzic-dziecko przy użyciu analizy loci CODIS wynika z niewystarczającej liczby analizowanych loci oraz ich ograniczonej polimorficzności (Tamura i in., 2015).

Podobnie, niewystarczająca informatywność analizy loci CODIS może stanowić wyzwanie w przypadkach identyfikacji ofiar katastrof masowych, totalitaryzmów, zbrodni wojennych oraz osób zaginionych (Bradford i in. 2011). W takich przypadkach profile genetyczne osób zmarłych lub zaginionych porównuje się z profilami dostępnych krewnych lub potencjalnych krewnych, których dane znajdują się w bazach. Gdy analiza obejmuje zbyt małą liczbę loci, istnieje ryzyko, że profile poszukiwanych osób mogą przypadkowo pasować do innych profili z bazy, co poważnie komplikuje identyfikację i stanowi istotne ograniczenie badania (Jabłońska-Milczarek i Frankowski, 2020).

Dzięki powszechności występowania sekwencji mikrosatelitarnych w ludzkim genomie i dostępności baz całogenomowych, możliwa jest identyfikacja i praktyczne wykorzystanie nowych, wysoce polimorficznych loci mikrosatelitarnych w badaniach pokrewieństwa, co stwarza możliwość poprawy możliwości analitycznych w trudnych przypadkach. W wyniku prowadzenia projektu 1000 Genomes (1000 Genomes Project Consortium, 2012) opublikowano dane na temat tysięcy wcześniej nieopisanych loci STR (Willems i in., 2014).

Identyfikując tego rodzaju markery genetyczne należy mieć na uwadze, że powinny spełniać określone kryteria dotyczące struktury samej sekwencji powtórzonej, sekwencji flankujących, a także powinny cechować się wysoką zmiennością genetyczną w obrębie badanej populacji.

Markery mikrosatelitarne o dużym znaczeniu dla genetyki sądowej to przede wszystkim sekwencje z powtarzającym się motywem składającym się z 3-5 nukleotydów. Krótszy motyw dinukleotydowy wiąże się z ryzykiem częstszego występowania amplikonów typu stutter (Ellegren, 2004). Pojawienie się takich dodatkowych pików utrudnia interpretację wyników, zwłaszcza w przypadku niezrównoważonych mieszanin DNA, gdy może być problematyczne ustalenie, czy dany amplikon DNA pochodzi od rzeczywistego allelu danej osoby, czy też jest amplikonem typu stutter allelu innej osoby (Brookes i in., 2012). Inną istotną cechą, którą powinien charakteryzować się informatywny locus w genetyce kryminalistycznej jest odpowiednia struktura sekwencji flankujących, czyli sekwencji otaczających ten locus. Nie powinna ona zawierać powtórzeń mono- i dinukleotydowych, powinna być sekwencją konserwatywną, tzn. w obrębie sekwencji flankujących locus STR nie powinno być częstych mutacji punktowych, w tym insercji/delecji, które utrudniałyby zaprojektowanie specyficznych, silnie hybrydyzujących starterów do reakcji PCR. Dodatkowo, ważna jest unikalność sekwencji flankujących w genomie ludzkim. Obecność sekwencji powtarzających się wielokrotnie w genomie zmniejsza szansę na udane zaprojektowanie odpowiednich starterów, szczególnie w reakcjach multipleks-PCR. Ponadto, lokalizacja markera STR w obrębie dłuższej sekwencji powtórzonej może prowadzić do powstawania niespecyficznych produktów reakcji PCR, co powoduje, że docelowy marker DNA amplifikuje się mniej efektywnie. Inną cechą wartościowego markera STR jest jego zwartość. Przykładowo marker o stosunkowo wąskim zakresie długości alleli oferuje trzy rodzaje zalet w porównaniu do markerów z szerszym zakresem długości amplikonów: ułatwia multipleksowanie reakcji PCR (Zeng i in., 2015), zmniejsza ryzyko wystąpienia nierównowagi heterozygotycznej pików elektroforetycznych spowodowanej bardziej efektywną amplifikacją krótszych alleli w reakcji PCR oraz pozwala uzyskać produkt PCR nawet ze śladów DNA wykazujących oznaki degradacji DNA (Westen i in., 2013). Dodatkowo, przy opracowywaniu nowych markerów STR należy uwzględnić wiele czynników, takich jak, niska częstość mutacji w jednostce powtórzonej i jej sekwencjach flankujących oraz wysoka heterozygotyczność locus.

1.3 Alternatywne metody analizy DNA człowieka w genetyce sądowej

Skomplikowane przypadki badań biologicznego pokrewieństwa obejmujące m.in. analizę pokrewieństwa w relacjach 2-go i 3-go stopnia, pokrewieństwa z historią kazirodztwa lub katastrofy masowe sprawiają, że obecnie dostępne komercyjne zestawy odczynników do analizy loci STR nie gwarantują uzyskania rozstrzygającego wyniku badań (Zhang i in., 2022). Gdy standardowe metody zawodzą, istnieje potrzeba wprowadzenia nowych, bardziej zaawansowanych technologii opartych np. na sekwencjonowaniu nowej generacji (NGS) (Ballard i in. 2020), analizie chromosomów płci (analiza STR Y, X) (Jobling, i in., 1997) lub analizach polimorfizmu pojedynczych nukleotydów (SNP) z wykorzystaniem mikromacierzy (de Vries i in., 2022). Wspomniane metody, pomimo swoich zalet, mają także wady, takie jak: wysokie koszty analizy (NGS, mikromacierze SNP), wysokie koszty wdrożenia nowej metody badań laboratoryjnych (NGS, mikromacierze SNP) oraz niemożność wykorzystania metody w badaniach określonych relacji rodzinnych (analizy chromosomów Y, X). Pomimo wysokiej wartości informacyjnej tych metod, ze względu na powyższe czynniki, techniki te nie zastąpiły analizy długości alleli STR z wykorzystaniem elektroforezy kapilarnej w rutynowych badaniach pokrewieństwa.

W związku z powyższym, kluczowe znaczenie ma opracowanie metody analizy dużej liczby wysoce polimorficznych autosomalnych loci STR, która opiera się na szeroko stosowanej w laboratoriach genetyczno-sądowych technologii multipleks-PCR oraz elektroforezie kapilarnej. Jest to szczególnie istotne w badaniach pokrewieństwa w bardziej złożonych relacjach rodzinnych, w których dostępne komercyjne zestawy odczynników do analizy autosomalnych loci STR nie gwarantują jednoznacznych wyników (Tamura i in., 2015). Dlatego rozwój technologii, która umożliwiłaby pokonanie tych trudności bez potrzeby wprowadzania nowych technologii analitycznych w laboratoriach genetyczno-sądowych, pozostaje przedmiotem zainteresowania specjalistów w tej dziedzinie (Xu i in., 2022). Dzięki obecności dużej liczby wysoce polimorficznych loci STR w genomie człowieka oraz wykorzystaniu technologii sekwencjonowania całogenomowego w badaniach populacyjnych, opracowanie nowej wysoce informatywnej metody analizy loci mikrosatelitarnych w celu badania biologicznego pokrewieństwa jest w pełni realne.

2. Cel pracy

Celem niniejszego doktoratu wdrożeniowego było opracowanie nowej metody do badania biologicznego pokrewieństwa w dalszych relacjach rodzinnych. Ta nowa metoda o nazwie Kinfinder miała zostać wdrożona w Laboratorium Diagnostyki Molekularnej GenMed, w związku z czym musiała opierać się na technologii reakcji multipleks-PCR oraz elektroforezy kapilarnej. Takie podejście pozwoliłoby na stosowanie metody przez firmę bez konieczności ponoszenia dodatkowych kosztów inwestycyjnych w nową aparaturę, przestrzeń laboratoryjną oraz w szkolenia pracowników.

Aby osiągnąć główny cel pracy należało zrealizować następujące cele cząstkowe:

- zidentyfikowanie w genomie człowieka najbardziej polimorficznych loci STR;
- przeprowadzenie badań przesiewowych w celu oszacowania heterozygotyczności wytypowanych loci STR oraz rozpiętości ich alleli w populacji polskiej;
- zaprojektowanie dwóch reakcji multipleks-PCR umożliwiających równoczesną amplifikację 2 x 25 loci STR w jednym procesie analitycznym;
- opracowanie drabiny allelicznej;
- opracowanie odczynnika do kalibracji spektralnej analizatora genetycznego;
- ewaluację metody Kinfinder przez jej porównanie z wiodącym zestawem komercyjnym;
- wdrożenie metody do rutynowej pracy Laboratorium Diagnostyki Molekularnej GenMed w badaniach pokrewieństwa w różnych relacjach rodzinnych.

3. Materiały i metody

3.1 Izolaty DNA

W badaniach wykorzystano 200 izolatów DNA pobranych od zanonimizowanych klientów Laboratorium Diagnostyki Molekularnej GenMed, którzy wcześniej wyrazili pisemną zgodę na wykorzystanie ich materiału genetycznego do celów badawczych. Dodatkowo, badania prowadzono na izolatach DNA pozyskiwanych w trakcie realizacji badań od trzech osób, ochotników biorących udział w pracach laboratoryjnych, którzy również wyrazili pisemną zgodę na udział w badaniu. Izolację DNA prowadzono za pomocą zestawu Swab (A&A Biotechnology, Gdańsk, Polska) zgodnie z protokołem producenta zestawu.

3.2 Startery do reakcji PCR

Startery do reakcji PCR zostały zsyntetyzowane przez firmę Sigma Aldrich. W przypadku bezpośredniego znakowania amplikonów, każdy ze starterów forward znakowany był jednym z czterech barwników fluorescencyjnych: 6-FAM, HEX, TAMRA lub ROX. Startery reverse nie były znakowane. Wszystkie startery zostały oczyszczone przez producenta metodą HPLC.

3.3 Identyfikacja nowych polimorficznych loci STR z danych genomowych

Do identyfikacji loci STR w genomie człowieka wykorzystano bazę danych STRCatalog, dostępną pod adresem http://strcat.teamerlich.org (Willems i in., 2014) oraz bazę WebSTR, dostępną pod adresem http://webstr.ucsd.edu (Lundström i in., 2023). Sekwencje nukleotydowe loci STR były analizowane w programie Genome Data Viewer, wykorzystując referencyjny genom człowieka (https://www.ncbi.nlm.nih.gov/gdv/browser/genome/), dostępnym w serwisie internetowym National Center for Biotechnology Information (NCBI). Kryteriami wyboru loci w bazach danych STRCatalog i WebSTR były: heterozygotyczność loci STR (>80% dla populacji światowej); rozpiętość locus STR, czyli różnica długości pomiędzy najkrótszym i najdłuższym allelem locus STR wynosząca do 100 pz; podobny rozkład częstości alleli w populacjach kontynentalnych; długość motywu powtórzonego tandemowo w locus STR wynosząca 3-5 nukleotydów; brak sekwencji zawierających powtórzenia mononukleotydowe dłuższe niż 7 pz oraz powtórzenia dinukleotydowe dłuższe niż 20 pz w regionach bezpośrednio flankujących locus STR, brak sekwencji wielokrotnie powtórzonych w

genomie człowieka w sekwencjach flankujących locus STR.

3.4 Identyfikacja polimorficznych loci STR z danych literaturowych

Do poszukiwania dodatkowych loci STR spełniających kryteria zawarte w pkt. 3.3 zostały wykorzystane baza Pubmed (https://pubmed.ncbi.nlm.nih.gov/) oraz wyszukiwarka Google.pl. Dane wyszukiwano kierując do baz zapytania zawierające następujące frazy kluczowe: “highly polymorphic STR loci”, “most heterozygous STR loci” “heterozygosity of STR loci”, “highly hetorozygous microsatellite locus" i pokrewne.

3.5 Projektowanie starterów do reakcji PCR i multipleks-PCR

Do projektowania starterów do reakcji singlepleks-PCR i multipleks-PCR wykorzystano ogólnodostępne programy i narzędzia bioinformatyczne: program Primer3 (https://primer3.ut.ee/), primerBLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) oraz BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch).

Przy projektowaniu starterów przyjęto następujące kryteria: długość starterów 18-30 bp, procent nukleotydów G i C na poziomie 25-50%, temperatura topnienia Tm = 60°C (+/- 5°C). Do projektowania starterów w programie BLAST wykorzystano bazę danych RefSeq Representative Genomes. Sekwencja każdego startera zaproponowanego przez wyżej wymienione programy do projektowania starterów została sprawdzona przez przyrównanie do genomu referencyjnego GRCh38.p14 w serwisie internetowym 1000 Genomes Browser (https://www.ensembl.org/Homo_sapiens/Info/Index) pod kątem występowania potencjalnych polimorfizmów SNP oraz insercji/delecji w miejscach przyłączania starterów. Progiem akceptacji częstości występowania polimorfizmów w sekwencji starterów była wartość 0,0002 (0,02%). Weryfikacja potencjalnych niepożądanych oddziaływań pomiędzy starterami w reakcji multipleks-PCR została przeprowadzona z użyciem programów MultiPLX (Kaplinski i Remm, 2015), Autodimer (Vallone i Butler, 2015) oraz Multiple Primer Analyzer (https://www.thermofisher.com/pl/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/multiple-primer-analyzer.html). Startery, które w reakcji PCR generowały artefakty lub nie pozwalały na skuteczną amplifikację, były dodatkowo modyfikowane

poprzez przesunięcie startera bliżej/dalej od docelowej sekwencji STR oraz skracanie/wydłużanie sekwencji na końcu 3' danego startera w celu zwiększenia jego specyficzności bez zmiany długości amplikonu. Zmienione startery ponownie sprawdzano za pomocą narzędzi BLAST, MultiPLX, Autodimer oraz Multiple Primer Analyzer jak opisano powyżej.

3.6 Modyfikacja starterów do przeszukiwania loci STR

Do badań przesiewowych zastosowano metodę M13-tailing, polegającą na wykorzystaniu dodatkowego, uniwersalnego startera (Boutin-Ganache i in., 2001). Do amplifikacji loci STR wykorzystano startery M13(-40) o sekwencji GTTTTCCCAGTCACGAC oraz M13(-47) o sekwencji CGCCAGGGTTTTCCCAGTCACGAC znakowane barwnikami fluorescencyjnymi 6-FAM, VIC oraz ROX oraz startery wskazane w tabeli nr 1.

3.7 Modyfikacje starterów w celu ograniczenia niepełnej adenylacji

Wybrane startery, które wykazywały niepełną adenylację końcowego produktu PCR, zostały zmodyfikowane metodą PIG-tail poprzez dodanie sekwencji GTTTCTT na końcu 5' starterów reverse (Brownstein i in., 1996).

3.8 Przygotowanie puli matrycowego DNA do testowania polimorfizmu loci STR

Aby przygotować matryce do testowania polimorfizmu nowych loci STR, 200 izolatów DNA podzielono na cztery grupy, z których utworzono cztery mieszaniny zawierające DNA pochodzące od 50 różnych osób w jednakowym stężeniu. Stężenie DNA w mieszaninach zostało ustalone na poziomie 2 ng/µl, z równym wkładem każdego izolatu (40 pg/µl DNA od każdej z 50 osób obecnych w mieszance DNA). W trzech izolatach zawierających DNA od jednej osoby stężenie DNA doprowadzono do wartości 1 ng/µl.

3.9 Testowanie polimorfizmu loci STR metodą PCR

Reakcję PCR przeprowadzono za pomocą zestawu Type-it Microsatellite PCR Kit (Qiagen, Niemcy) w końcowej objętości 10 µl zawierającej: 5 µl mieszaniny Type-it Multiplex PCR Master Mix, 0,2 µM specyficzny starter reverse, 0,075 µM specyficzny fuzyjny starter forward

z modyfikacją M13(-40), 0,125 µM fluorescencyjnie znakowany uniwersalny starter M13(-40) i 1 µl matrycowego DNA (2 ng mieszaniny przygotowanej jak w pkt. 3.6). Reakcję przeprowadzono przez 30 cykli według następującego profilu termicznego: początkowa aktywacja polimerazy: 95°C - 5 minut, denaturacja w 95°C - 30 sekund, przyłączanie starterów w 60°C - 90 sekund, wydłużanie w 72°C - 60 sekund, końcowe wydłużanie w 65°C - 60 minut. W reakcji użyto kontroli negatywnej (woda zamiast matrycowego DNA) oraz kontroli pozytywnych (matrycą były analizowane osobno izolaty DNA pochodzące od trzech różnych osób z populacji polskiej).

3.10 Amplifikacja alleli w multipleksowanej reakcji PCR

Amplifikację alleli 50 loci STR wybranych w badaniach przesiewowych prowadzono w dwóch reakcjach multipleks-PCR. W reakcji multipleks-PCR A amplifikowano loci: D04L885. D14L953, D06L106, D16L554, D01L267, D03L115, D2N43, D12L794, D02L114, D1S1656, D05L169, D14L699, D3A57, D8S1132, D3N61, D05L140, D13S742, D2L174, D17L432, D1N16, D15L495, D21L291, D14L785, D05L207, D01L569 oraz locus amelogeniny (Amel), w reakcji multipleks-PCR B amplifikowano loci: D09L159, D02L221, D02L142, D8A26, D05L113, D14L276, D17L255, D20L226, D08L110, D16L732, D01L217, D03L109, D07L144, D7S3048, D12S391, D01L228, D03L194, D07L101, D10L126, D3N54, D01L215, D12L908, D07L147, D12L630, D10S2325, D07L134.

Wyjściową mieszaninę starterów do reakcji multipleks-PCR przygotowano poprzez dodanie 1 µl roztworu każdego startera o stężeniu 10 µM, a następnie objętość całej mieszaniny uzupełniano do objętości 200 µl. Do każdej reakcji multipleks-PCR prowadzonej w objętości 10 µl dodawano 2 µl mieszaniny starterów. Wydajność reakcji multipleks-PCR optymalizowano tak modyfikując stężenie starterów dla każdego z amplifikowanych loci, aby osiągnąć sygnał fluorescencyjny wszystkich amplikonów w układzie heterozygotycznym w zakresie 1000-2000 RFU.

Reakcje multipleks-PCR przeprowadzono z wykorzystaniem zestawu HOT FIREPol® MultiPlex Mix (Solis Biodyne, Estonia) w końcowej objętości 10 µl zawierającej: 3 µl mieszaniny HOT FIREPol® MultiPlex Mix, 2 µl mieszaniny starterów (o stężeniu 0,05 µM każdy), 1 µl

matrycowego DNA o stężeniu 1ng/µl. Reakcję przeprowadzono przez 32 cykle według następującego profilu termicznego: początkowa aktywacja polimerazy: 95°C - 5 minut, denaturacja w 95°C - 30 sekund, przyłączanie starterów w 60°C - 5 minut, wydłużanie w 72°C - 65 minut, końcowe wydłużanie w 65°C - 60 minut. W reakcji użyto kontroli negatywnej (woda zamiast matrycowego DNA) oraz kontroli pozytywnych.

3.11 Sporządzenie odczynnika do kalibracji spektralnej analizatora genetycznego

Odczynnik do kalibracji spektralnej analizatora genetycznego został stworzony poprzez zmieszanie pięciu produktów reakcji PCR o różnej długości, z których każdy wyznakowany był jednym z pięciu barwników fluorescencyjnych: ATTO633, 6-FAM, HEX, TAMRA lub ROX (Tab. 1). Do amplifikacji fragmentów DNA użyto matrycy DNA w postaci plazmidu pGEM-3Z (Promega) oraz fluorescencyjnie znakowanych starterów forward oraz nieznakowanych reverse.

Tabela 1. Sekwencje starterów wykorzystanych do sporządzenia odczynnika do kalibracji spektralnej analizatora genetycznego.

Nazwa startera	Sekwencja od 5- do 3'	Długość amplikonu	Barwnik
pGEM3ZforATTO633	GGGCGAATTCGAGCTCGGTA	100	ATTO633
pGEM3ZrevATTO633	GATTACGCCAAGCTATTTAGGTGA		-
pGEM3ZforROX	GGGCGAATTCGAGCTCGGTA	120	ROX
pGEM3ZrevROX	CAGGAAACAGCTATGACCATGATTA		-
pGEM3ZforTAMRA	GGGCGAATTCGAGCTCGGTA	140	TAMRA
pGEM3ZrevTAMRA	GAGCGGATAACAATTTCACACAG		-
pGEM3ZforHEX	GGGCGAATTCGAGCTCGGTA	160	HEX
pGEM3ZrevHEX	CGTATGTTGTGTGGAATTGTGAG		-
pGEM3Zfor6FAM	GGGCGAATTCGAGCTCGGTA	180	6-FAM
pGEM3Zrev6FAM	TACACTTTATGCTTCCGGCTCG		-

Reakcję PCR przeprowadzono za pomocą zestawu HOT FIREPol® MultiPlex Mix (Solis Biodyne, Estonia) w końcowej objętości 20 µl zawierającej: 6 µl mieszaniny HOT FIREPol® MultiPlex Mix, 0,25 µM starter reverse, 0,25 µM starter forward, 2 µl matrycowego DNA o stężeniu 1ng/µl. Reakcję przeprowadzono przez 30 cykli według następującego profilu termicznego: początkowa aktywacja polimerazy: 95°C - 5 minut, denaturacja w 95°C - 30 sekund,

przyłączanie starterów w 60°C - 90 sekund, wydłużanie w 72°C - 60 sekund, końcowe wydłużanie w 65°C - 60 minut. Mieszanina poreakcyjna została oczyszczona z wykorzystaniem zestawu odczynników Clean-Up Concentrator (A&A Biotechnology, Polska) zgodnie z protokołem producenta.

Roztwory oczyszczonych wyznakowanych fluorescencyjnie amplikonów DNA rozdzielono na drodze elektroforezy kapilarnej, a następnie zostały one zmieszane ze sobą w takim stosunku objętościowym, aby sygnał fluorescencji każdego z pięciu amplikonów zawierał się w przedziale 1000 – 2000 RFU. Otrzymana mieszanina amplikonów została wykorzystana jako odczynnik do kalibracji spektralnej analizatora genetycznego ABI3130XL (Applied Biosystems, Stany Zjednoczone) zgodnie z zaleceniami producenta analizatora.

3.12 Analiza czułości reakcji multipleks-PCR

Czułość dwóch zaprojektowanych reakcji multipleks-PCR została przetestowana poprzez przeprowadzenie reakcji dla izolatu DNA zawartego w zestawie AmpFℓSTR™ (DNA Control 007) (Applied Biosystems) w stężeniach DNA: 1 ng/µl, 0,5 ng/µl, 0,25 ng/µl, 0,125 ng/µl, 0,1 ng/µl. Reakcje multipleks PCR prowadzono zgodnie z metodą opisaną w pkt. 3.8.

3.13 Elektroforeza kapilarna

Elektroforeza znakowanych fluorescencyjnie amplikonów była prowadzona przy użyciu analizatora genetycznego ABI 3130XL. Fragmenty były rozdzielane z wykorzystaniem polimeru POP-7 (Applied Biosystems) z użyciem kapilar o długości 36 cm w obecności standardu długości DNA Size Standard v2.0 GeneScan™ 600 LIZ™ (Applied Biosystems). Przed elektroforezą, 1 µl produktów PCR mieszano z 9,6 µl formamidu i 0,4 µl standardu długości DNA GeneScan™ 600 LIZ™ dye Size Standard v2.0 (Applied Biosystems) i denaturowano w 95°C przez 5 minut, a następnie chłodzono przez 10 min lub przechowywano w 4°C do czasu wykonania elektroforezy. Elektroforezę produktów uzyskanych w pojedynczej reakcji PCR prowadzono w następujących warunkach: czas iniekcji 20 sekund, napięcie iniekcji 1,5 kV; czas elektroforezy 1200 sekund, napięcie elektroforezy 15 kV. Elektroforezę produktów reakcji multipleks-PCR A i B prowadzono w następujących warunkach: czas iniekcji 18 sekund, napięcie iniekcji 1,2 kV; czas elektroforezy 1200 sekund, napięcie elektroforezy 15 kV.

Elektroforegramy były analizowane przy użyciu programów PeakScanner v1.0 (Applied Biosystems) oraz GeneMapper ID-X v1.5 (Applied Biosystems).

3.14 Szacowanie heterozygotyczności loci STR na podstawie puli DNA

Do obliczania częstości alleli na podstawie powierzchni poszczególnych pików elektroforetycznych użyto następującego wzoru: F1 = P1/PA, gdzie F1 = częstość allelu 1; P1 = powierzchnia allelu 1; PA = suma powierzchni wszystkich alleli w danym locus. Uzyskane częstości występowania alleli w populacji polskiej dla każdej z czterech mieszanin DNA uśredniano do finalnej wartości dodając do siebie częstości występowania allelu w każdej próbce pulowanego DNA i dzieląc tę wartość przez liczbę próbek. Heterozygotyczność każdego locus STR obliczano poprzez sumowanie kwadratów częstości występowania wszystkich alleli tego locus (homozygotyczność), a następnie odjęcie tej wartości od liczby 1 (heterozygotyczność).

3.15 Analiza statystyczna

Analizy statystyczne dotyczące zakresów wyników badań biologicznego pokrewieństwa przeprowadzono za wykorzystaniem programu KinBN v1.1.2 (Morimoto i in., 2020). Analizy wyników badań pokrewieństwa wykonano dla następujących relacji rodzinnych: rodzic-dziecko (pierwszy stopień pokrewieństwa), pełne rodzeństwo, wujek-siostrzeniec (drugi stopień pokrewieństwa) oraz kuzyn-kuzyn (trzeci stopień pokrewieństwa). Analizy biostatystyczne wykonano dla dwóch hipotez: hipotezy H1 - pokrewieństwo w badanej relacji rodzinnej oraz hipotezy H2 – brak pokrewieństwa pomiędzy badanymi osobami.

Symulacje przebiegały w następujący sposób: dla hipotezy H1 (obecność pokrewieństwa) program KinBN losował profile genetyczne hipotetycznych osób rodzicielskich (założycieli), zgodnie z częstością występowania alleli w populacji polskiej, a następnie symulował dziedziczenie alleli w sposób mendlowski zgodnie z określoną relacją rodzinną. Na tej podstawie program obliczał prawdopodobieństwo pokrewieństwa dla każdej pary osób. W przypadku hipotezy H2 (brak pokrewieństwa) program jedynie losował profile genetyczne dla dwóch niespokrewnionych, po czym obliczał prawdopodobieństwo pokrewieństwa tych osób w określonej relacji rodzinnej.

Obliczenia zakresów spodziewanych wyników badań pokrewieństwa wykonano dla standardowej metody analizy 21 loci STR (20 loci CODIS oraz locus SE33) wykorzystywanych w systemie Globalfiler (Thermofisher) oraz dla rozszerzonej metody analizy 69 loci STR, w tym 21 loci systemu Globalfiler oraz 50 loci metody Kinfinder (2 loci STR są wspólne dla metody Globalfiler i Kinfinder).

Dla każdej relacji biologicznego pokrewieństwa, dla obu metod analitycznych i dla każdej z obu hipotez (H1 i H2) wykonano po 10 000 analiz obliczeniowych (łącznie 16 000 analiz).

3.16 Nazewnictwo nowo scharakteryzowanych loci STR

Na potrzeby realizacji niniejszej pracy przyjęto następujący schemat nazewnictwa nowo scharakteryzowanych loci STR: loci były oznaczane według wzoru DxxLyyy, gdzie "xx" odnosi się do numeru chromosomu, a "yyy" do pierwszych trzech cyfr chromosomowej lokalizacji początku sekwencji tandemowo powtórzonej w referencyjnej wersji genomu GRCh37.p13. Nazw loci opisanych w literaturze naukowej i wykorzystanych w niniejszej pracy nie zmieniano.

3.17 Sekwencjonowanie metodą Sangera

Locus STR w układzie homozygotycznym było amplifikowane w reakcji PCR z użyciem nieznakowanych starterów w warunkach opisanych w pkt. 3.9. Mieszaniny poreakcyjne zawierające zamplifikowaną sekwencję tandemowo powtórzoną oczyszczano przy użyciu zestawu EPPiC Fast (A&A Biotechnology, Polska) zgodnie w protokołem załączonym przez producenta. Mieszanina reakcyjna do sekwencjonowania składała się z 3 µl oczyszczonej mieszaniny poreakcyjnej PCR, 1,5 µl nieznakowanego startera forward lub reverse o stężeniu 10 µM, 1,5 µl wody i 4 µl mieszaniny odczynników BigDye™ Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems). Produkty sekwencjonowania rozdzielono na analizatorze genetycznym ABI 3130XL, a wyniki analizowano w programach Chromas (Technelysium Pty Ltd) i FinchTV (Geospiza, Inc.). Sekwencjonowanie było prowadzone dla obu nici matrycowego DNA z użyciem starterów forward i reverse.

3.18 Konstrukcja drabin allelicznych

Drabiny alleliczne opracowano poprzez zmieszanie amplikonów alleli locus STR, które występują w populacji polskiej z częstością przekraczającą 1%. Mieszaniny poreakcyjne, zawierające zamplifikowane allele, były rozdzielane na drodze elektroforezy kapilarnej. Na podstawie intensywności sygnału fluorescencyjnego (RFU, ang. Relative Fluorescence Unit) określono proporcje, w jakich amplikony alleli w danym locus STR zostaną zmieszane, aby każdy allel był możliwie równomiernie reprezentowany w drabinie allelicznej. Tak uzyskana mieszanina amplikonów była następnie ponownie amplifikowana, tworząc drabinę alleliczną. Wszystkie reakcje PCR przeprowadzono w warunkach opisanych w pkt. 3.7.

4. Wyniki

4.1 Identyfikacja wysoce polimorficznych loci STR w genomie człowieka

W pierwszym etapie badań z bazy STRCatalog, agregującej dane dotyczące sekwencji mikrosatelitarnych z projektu 1000 Genomes wybrano 155 wysoce polimorficznych loci STR, które spełniały wszystkie kryteria opisane w pkt. 3.3. Dodatkowo na podstawie analizy danych literaturowych (Novroski i in., 2018; Shin i in., 2004) wybrano 28 loci STR spełniające te same kryteria wyboru. Lista wybranych loci STR zamieszczona jest w tabeli nr 1.

Tabela 2. Loci STR wyznaczone do szacowania heterozygotyczności. W tabeli przedstawiono nazwę locus, lokalizację chromosomową początku (5') i końca (3') sekwencji tandemowo powtórzonej dla wersji genomu referencyjnego GRCh37.p13, motyw sekwencji tandemowo powtórzonej, heterozygotyczność locus wskazaną w bazie STRCatalog lub w danych literaturowych, sekwencje starterów forward i reverse wykorzystanych do amplifikacji danego locus w badaniach przesiewowych. Liczba w nazwie locus występująca po literze "D" oznacza chromosom, na którym znajduje się wskazany locus.

Lp.	Locus	Początek	Koniec	Motyw	Het.	Starter forward (5'-3')	Starter reverse (5'-3')
1.	D01L154	15437756	15437793	AAC	0.859	GTTTTCCCAGTCACGACGCATTTCCTGACCCACCACT	GTTTCTTAATTAGCTGGGCATGGTGGT
2.	D01L202	202037376	202037425	AAT	0.870	GTTTTCCCAGTCACGACGGAAGTTGTAATGAGCTGAGATCGAG	GTTTCTTCACCATGCCTGGCTAATTGTT
3.	D01L215	215409389	215409441	AGAT	0.879	GTTTTCCCAGTCACGACCTATACATTTTTAAAGGCAGAAGAAG	ACTGTGCCACCTCCTCTCAT
4.	D01L217	217960093	217960155	AAG	0.875	GTTTTCCCAGTCACGACGATGCAGTGAAACCCATTAGGATA	GTTTCTTCTGGTGGAGATACAAACAAAGATGG
5.	D01L228	228654226	228654313	AAG	0.913	GTTTTCCCAGTCACGACAAGAGCCATGTGAGGTTCGTTTAC	CAAAACTTAGCCTAGCATGGTCAG
6.	D01L229	229467535	229467578	AAT	0.860	GTTTTCCCAGTCACGACCCCTTTCCCCATATCTCACCC	GAGCTGAGATCGCGCCATT
7.	D01L267	26704377	26704454	AAAG	0.889	GTTTTCCCAGTCACGACACTTTCCAGGCATTTGAAGAAGT	GTTTCTTTGTGAATTACTTTGGGAGGCTGT
8.	D01L284	28418676	28418710	AAT	0.839	GTTTTCCCAGTCACGACCAGGACCTGGGAGACGGA	GCTAGGATTACAGATGGGAGCC
9.	D01L569	56936756	56936835	AAAG	0.863	GTTTTCCCAGTCACGACGGGCAAATTATTTACAGGCACCA	GTTTCTTTCTTCGCTTCCCTTCTTCCC
10.	D01L679	6792427	6792507	AAAG	0.863	GTTTTCCCAGTCACGACGGATGACAGTGTGGGTTCTCT	GTTTCTTGCTACTCGGGAGGCTGA
11.	D01L801	80115411	80115486	AAAG	0.884	GTTTTCCCAGTCACGACATTGGAAGAAAGAGAAACCCATACA	GTTTCTTTCTCACCAGGGAAAATAAGTGCA
12.	D02L106	106337999	106338047	AAAG	0.896	GTTTTCCCAGTCACGACTCAGGAGTTCAAGACCAGCC	GCCACAATCCAGCCACTACT
13.	D02L114	114054526	114054608	AAG	0.900	GTTTTCCCAGTCACGACTCTCACATCTGCTCTGATAAATTCCA	TGATCCATTGTCCCAGCCAG
14.	D02L117	117592756	117592804	AGAT	0.853	GTTTTCCCAGTCACGACGGATACCCAAGTCCCTCAGTC	GTTTCTTTGGAATCATGAAATCGAAGCCATTT
15.	D02L133	133034774	133034843	AAGG	0.897	GTTTTCCCAGTCACGACGGGTGATCCTGGGCGACA	ATCATTTGCTTGCTTGCTTGC
16.	D02L142	142567661	142567756	AAAG	0.887	GTTTTCCCAGTCACGACCATTGAAATAATTAACCTCACATTTTC	GAAAGGTTCTCATGCCATTCGG
17	D02L174	174327360	174327425	AAAG	0.911	GTTTTCCCAGTCACGACAAAACAGTGTGCGGGGTGA	GTTTCTTAGCCAGAAAAGCATACCCGT

Lp.	Locus	Początek	Koniec	Motyw	Het.	Starter forward (5'-3')	Starter reverse (5'-3')
18	D02L190	190609717	190609755	ACT	0.916	GTTTTCCCAGTCACGACAGTT CAACATCAGCCTAGCCA	CCTCAGCCTCCCAAGTAGC
19.	D02L199	199865283	199865326	AAG	0.866	GTTTTCCCAGTCACGACAGCT GAGATCACACCACTGC	CACCATCTATGTTTGTGTAGGCA
20.	D02L212	212066367	212066433	AAAG	0.863	GTTTTCCCAGTCACGACTCTG AAGGGATTGAGGTTGCA	GTTTCTTTGGTAATTTGGTTTGTTG TACGCA
19.	D02L220	220182877	220182914	AAT	0.892	GTTTTCCCAGTCACGACATGA GCCACCATGCCCAT	GTTTCTTAGCCAAGATTGTACCACT GCT
20.	D02L221	221218006	221218073	AAAG	0.861	GTTTTCCCAGTCACGACGAAC CATTCTCATCCAGCAGG	GTTTCTTTTGTGAGCTATGATTATG CCATTG
21.	D02L227	227245834	227245891	AGAT	0.908	GTTTTCCCAGTCACGACTCCCC AGTCCGCACACA	GTTTCTTGGAATTTAGGTAAAGGT GTGATTCC
22.	D02L231	231153171	231153239	AAAG	0.883	GTTTTCCCAGTCACGACAGGC TGATCATTTGACTTTCTTTGT	GTTTCTTCACATGCTCTCACTCATA AGTGGA
23.	D02L232	232425114	232425158	AAT	0.918	GTTTTCCCAGTCACGACGAGC TGCTGGAACCTCTTC	GTTTCTTTTCAGCCTCCCAAAGTGC
24.	D02L234	234570270	234570328	AAAG	0.878	GTTTTCCCAGTCACGACAAAG TGCAAGGTTTGAAGCC	AGCTGTGGTTGGCGATCATT
25.	D02L504	50493104	50493138	AAT	0.851	GTTTTCCCAGTCACGACCGGT CACACAAAGATAGAGGT	AGCCAAGATCGTGCCACTG
26.	D02L570	570334	570394	AAG	0.905	GTTTTCCCAGTCACGACACCTT TTCATGTTTGCTCGCT	GTTTCTTTTGAGCCATCCTTGCATC CC
27.	D02L707	70788225	70788288	AAGG	0.895	GTTTTCCCAGTCACGACGGCG TAGGAGAAGGGAAAGTAAA	GTTTCTTTGAGGTGGGAGAATTGC TTGA
28.	D03L101	101684290	101684317	AAT	0.855	GTTTTCCCAGTCACGACTCTTT GTTCACTTGTCATGGGTG	CACAAAGTACCTAATGCCTGGC
29.	D03L109	109358	109420	AAG	0.869	GTTTTCCCAGTCACGACCTGA TGTCCCAGCACCCATA	TCACTGCAGCCTCAAACTCC
30.	D03L115	115658378	115658432	AAAG	0.894	GTTTTCCCAGTCACGACATCA GTATCAATGTCAAGGAGCT	GAGGCTGAAGTGGGAGGATC
31.	D03L142	14246925	14246965	AAT	0.903	GTTTTCCCAGTCACGACAGAT GAAACCAGGCGCAAAG	CTGGGCGAGAGAGTGAGACT
32.	D03L184	184209812	184209861	AAT	0.851	GTTTTCCCAGTCACGACACAA AAGCGAAACTCTGTCTCA	GTTTCTTAGAAAGGATATGGGTAA TGCTATTGA
33.	D03L194	1943917	1943976	AAGG	0.883	GTTTTCCCAGTCACGACCAAA GTTTGCAAGGGTAAAGAGTG	GTTTCTTGAAAACTGTAGCTACCG AAAGGTCTG
34.	D03L321	32165284	32165331	AGAT	0.852	GTTTTCCCAGTCACGACCCTT GACCCAGCCATGTGAA	AAGAAAGAAATGTGACTGGCCA
35.	D03L617	61778343	61778421	AAAG	0.854	GTTTTCCCAGTCACGACTACTT CTCCGTTGAGCCTACACTTA	GTTTCTTCACGTGCCTGTAGTCCTA GTTACTT
36.	D04L109	109332496	109332573	AAAG	0.889	GTTTTCCCAGTCACGACGCTG TACGTTCTAGCCAGTG	GTTTCTTTGTATTCTGCAACCTTTA CTAAACTT
37.	D04L152	152941396	152941477	AAAGG	0.831	GTTTTCCCAGTCACGACAGTC CCAGCTACTCAGGAGG	GTTTCTTTGGCTAAGACCCGTCCTA TGA
38.	D04L164	164955214	164955271	AGAT	0.864	GTTTTCCCAGTCACGACTGGA AGACAGTGTGGTGATTTCT	GTTTCTTGTGGTGTTTGGTTTTCTA TTCCCA
39.	D04L166	166279041	166279110	AAGGG	0.836	GTTTTCCCAGTCACGACCTCA GCCTCCCGAGTAGC	GTTTCTTGCTGGGCTTGATGAAGA ACC
40.	D04L180	180861419	180861451	AAC	0.841	GTTTTCCCAGTCACGACAGGA TGGGAGAGTTTATTAGAAGC	GTTTCTTACATACAACCCTCCCTCC CA
41.	D04L885	88549700	88549730	AAT	0.863	GTTTTCCCAGTCACGACTGAC CCATAGAAGTTCAAAAGA	GTTTCTTACCTAGTGACTTGAAACC TCA
42.	D04L888	888767	888805	AAAT	0.853	GTTTTCCCAGTCACGACGCAA CAAAGCGAGACTCTGTC	GTTTCTTCAGTGCATGAGTGTGGG AAC
43.	D05L105	105979402	105979440	AAT	0.846	GTTTTCCCAGTCACGACAGAA CAAGTAAGATCACCCAGAA	GCCAGGGGACAACAGTGAAA
44.	D05L113	113248117	113248176	AGAT	0.858	GTTTTCCCAGTCACGACTGTG GGTCTTTGTGATCATGT	AGGAGGGTTGTTAGGGAGGG
45.	D05L140	140726699	140726748	AAAG	0.902	GTTTTCCCAGTCACGACGAAA TTAAATGATGTTTGACTCCTTC	CAGCCTGGGAAACATAGCAATTGC
46.	D05L169	169085985	169086038	AAT	0.876	GTTTTCCCAGTCACGACCCTAT AATTAATCATTGCCAACACC	GTTTCTTCTTACATACCCAAGATGA CTCCACT
47.	D05L207	20789541	20789609	AAAG	0.884	GTTTTCCCAGTCACGACGATC TGAAACCCAGTAACCTCC	CTCTAGTGATTCTCCTGCCTAG
48.	D06L106	106214600	106214663	AGAT	0.864	GTTTTCCCAGTCACGACCCTAT AAGCAGGACATTTTTAACAC	TTGCACTCCAGCCTGGGAATTAG
49.	D06L144	144348361	144348432	AAAG	0.843	GTTTTCCCAGTCACGACATCC CAGCTACTCAGGAGGC	GTTTCTTATGGTGGGAGGTGGAG GTTA
50.	D06L160	160573676	160573745	AAAG	0.887	GTTTTCCCAGTCACGACCACC AGCAGACTTATACTACAGGA	GTTTCTTAGAATGGTGTGAACCTG GGAG
51.	D06L312	31264024	31264076	AAT	0.834	GTTTTCCCAGTCACGACTCTA AGGACACCATCAAGAAAGT	GTTTCTTTTCTAAAGGGGCTATGTC TCTTAAA
52.	D06L322	32269746	32269772	AAT	0.837	GTTTTCCCAGTCACGACCCTG TGAGGTAAGAATGGATTTT	GTTTCTTACAATTTTCATGTGCCAC AAAA
53.	D06L391	39130992	39131034	AAT	0.886	GTTTTCCCAGTCACGACTGGG AGATAGAGCAAGATTCTGT	GTTTCTTCTCCAGTAATACATTTTG GGGCA

Lp.	Locus	Początek	Koniec	Motyw	Het.	Starter forward (5'-3')	Starter reverse (5'-3')
54.	D06L402	40211270	40211324	AGAT	0.889	GTTTTCCCAGTCACGACTGTGGGACTTCTCAGCATTCA	GTTTCTTTGATATGTGAGTTAGCTGGCGA
55.	D06L477	47780805	47780859	AAC	0.858	GTTTTCCCAGTCACGACACCTCAGTACATGGAACCTTCA	GTTTCTTTGTGTCCACCTCTATCTGACG
56.	D06L788	78871278	78871326	AAT	0.855	GTTTTCCCAGTCACGACACCACTTTTGCCCTTCCAGT	GTTTCTTGACAGAGCAAGACTCCGTC
57.	D06L815	81581236	81581302	AGAT	0.847	GTTTTCCCAGTCACGACGGAATCAGCCCAAATGCCCA	GTTTCTTCATGGTGGGTGGGTGGATG
58.	D07L101	101470525	101470584	AAGAG	0.836	GTTTTCCCAGTCACGACGATCCCTTCTGCTGTTCCCC	AGGCTGATGTGGGAGGACT
59.	D07L121	121867745	121867781	AAT	0.882	GTTTTCCCAGTCACGACCATCCTTACATCATCCCCAATAGT	GTTTCTTCGCACACCTGTAGTCCCATA
60.	D07L122	122611130	122611194	AAGG	0.855	GTTTTCCCAGTCACGACTCTGCCTGCCATTCTACTCC	GTTTCTTGGCAAGAGCAGTAGCAAGATG
61.	D07L134	134201476	134201552	AAAAG	0.874	GTTTTCCCAGTCACGACAAGACGTCTTTTTCTTCAACATGTCTTTCC	GGTGATCCTGCATTCCTGAGA
62.	D07L138	138795366	138795412	AAT	0.915	GTTTTCCCAGTCACGACGGTGCCATTAAAAGCCTCTCG	GTTTTAGGGCTTCAGTAGGAGAGT
63.	D07L144	144893365	144893439	AAAAG	0.880	GTTTTCCCAGTCACGACGTTCATTTGTGGCATATTAATATCC	ATCAGCCTGGGCAACACCATAAAG
64.	D07L147	147834762	147834829	AAGG	0.874	GTTTTCCCAGTCACGACAGGTTGAGGCAGGAGAATAATTTA	GTTTCTTCTCTTCAACTACAGTGCATTTGC
65.	D07L323	32322098	32322123	AGAT	0.927	GTTTTCCCAGTCACGACCCAGCCTGGGCAACAAAGGCCAA	GACTACTTGTGAGATTGAAACTG
66.	D07L521	52117027	52117114	AAAAG	0.846	GTTTTCCCAGTCACGACGCAACATGACGAAACCCCAT	GTTTCTTTGCAGGTTGTGTCTTAGTACTCA
67.	D07L666	66614710	66614768	AGAT	0.890	GTTTTCCCAGTCACGACTGTGTTTTGCCTGTTAGAAATGT	AGAAAAGAAGGGAAATGTGCTTT
68.	D07L806	80619368	80619395	AAT	0.857	GTTTTCCCAGTCACGACGCCTGTAATCCCAGCACTTT	CCGATTTCATCACCCACCAGT
69.	D07L884	8843423	8843469	AAT	0.892	GTTTTCCCAGTCACGACCCTATGTAACAAACCTGCGCG	AGTATCAAGAGGTGGAGCTTTT
70.	D07L960	96057619	96057670	AGAT	0.860	GTTTTCCCAGTCACGACTGTCCAGCCTCTTTTCAGGA	GTTTCTTCCAAGGGGCAATAGGTAGGG
71.	D08L104	104599628	104599699	AAAG	0.882	GTTTTCCCAGTCACGACGTATTACCCTGATACCAAAACCAAA	TTCATCAGAAATATTGGTGCATAGTT
72.	D08L110	11069866	110169932	AAAG	0.892	GTTTTCCCAGTCACGACAGTCCTAGCTACTCCAGGGG	GTTTCTTAGCCTTCAATATGAGGTGGTCT
73.	D08L120	120874740	120874787	AAT	0.842	GTTTTCCCAGTCACGACGGTGGGATTACAGGCACGAG	GTTTCTTAGTTCACTGTTGTCCATGCT
74.	D08L120	120482221	120482298	AGAT	0.849	GTTTTCCCAGTCACGACACCCCATGACAGAAGTTTACCT	GTTTCTTTTGGTCACACAGTCAACTCTAA
75.	D08L135	135277346	135277386	AAAG	0.931	GTTTTCCCAGTCACGACTGCAGTGAGTCGAGATGGC	GTTTCTTCCACTAATACATCCCTTACACAAATTGA
76.	D08L375	37591980	37592035	AAGG	0.874	GTTTTCCCAGTCACGACCAGAGCTGAGATCGCGCC	GTTTCTTTGTTACCCAGCAGATACATTTGTG
77.	D08L836	83634636	83634683	AAT	0.849	GTTTTCCCAGTCACGACGCCAAAGTCTTGTTTTCAGGC	GTTTCTTTGAATCCACTTGCAGTACACTTTA
78.	D08L980	98003782	98003831	AAT	0.839	GTTTTCCCAGTCACGACCACAAATTAGCTGGGCGTGG	GTTTCTTCCAAAATCCACCAAACTCTAGGG
79.	D09L134	134707976	134708036	AAAG	0.913	GTTTTCCCAGTCACGACGGGTGACAGAGCAAGACTACA	GTTTCTTGCTGCTATTTCTTTTCCTCGGC
80.	D09L159	15944824	15944885	AAAG	0.862	GTTTTCCCAGTCACGACCTTTATTTCTAGGCAGAGCATGGTAG	GTTTCTTATTGTGCACATGTACCCTAAAACTT
81.	D09L182	18289134	18289190	AGAT	0.845	GTTTTCCCAGTCACGACTGGGTTCTCCAAAGAAATAGAACC	GTTTCTTAAGATGCCCAGCTACTTGGTTA
82.	D09L762	76286311	76286361	AGAT	0.856	GTTTTCCCAGTCACGACGATGGAGCAAAGAGACAGTTATTTC	GTTTCTTAGTATACCTGAAGAAAAAGGGGTTG
83.	D09L794	79452358	79452393	AAT	0.897	GTTTTCCCAGTCACGACAGGTGAAGCGATTGAGACCA	CAGCATGTCAGATACTTCCTGC
84.	D10L126	12644895	12644923	AAT	0.859	GTTTTCCCAGTCACGACGGACCCACAGTGACTCCCAT	GTTTCTTGCAACAGGGCGAGACTCT
85.	D10L300	30059574	30059640	AAG	0.872	GTTTTCCCAGTCACGACGCGGGCAGATCACTTGAGAT	CCTGGGTGAGAGAGCAAGAC
86.	D10L311	31112493	31112549	AAT	0.881	GTTTTCCCAGTCACGACTAGGAGCATGGTTTGAGCCC	GTTTCTTGCTCTGGGTGTGTGTGCA
87.	D10L616	61690083	61690113	AAAT	0.872	GTTTTCCCAGTCACGACGGGACTTGGGAAAGGGAGGA	GTTTCTTCAGGAGAATGGCGTGAACCC
88.	D10L677	67704607	67704690	AGAT	0.845	GTTTTCCCAGTCACGACAGGCTTGTTTCATAGACTCCAGT	GTTTCTTTCTCCTTTAGCTGTTACTTACCT
89.	D10L963	96378219	96378266	AAT	0.859	GTTTTCCCAGTCACGACCTGCCTGGGGAAATGGGTG	GTTTCTTTAGCCCGGGAGACAGTGAG
90.	D11L161	16147363	16147423	AAGG	0.841	GTTTTCCCAGTCACGACAGAATACAAACTCAAGCCCAAAGA	GTTTCTTTGTGGATTGATGCTTGGCCA

Lp.	Locus	Początek	Koniec	Motyw	Het.	Starter forward (5'-3')	Starter reverse (5'-3')
91.	D11L499	49957927	49958012	AAAAT	0.857	GTTTTCCCAGTCACGACGAGGCTGAGACACGAGAATCT	GTTTCTTGTATGGTGGTGGGCAACTGT
92.	D11L932	93285609	93285652	AAT	0.881	GTTTTCCCAGTCACGACTTCAAACTCCCAGCCTCAGG	ATCCCAGCTACTCAGGAGGC
93.	D12L115	115894994	11589507	AAAAG	0.908	GTTTTCCCAGTCACGACATTAGCTGGATGTGGTGGCG	GTTTCTTGTGGTCAGACTTCCTATTGTTCCT
94.	D12L130	130925947	130925994	AAT	0.871	GTTTTCCCAGTCACGACGGGTGACAGAGCGAGACT	TTCCTTCTTTCTCGTCCTTGTAA
95.	D12L131	13174142	13174180	AAC	0.870	GTTTTCCCAGTCACGACCGGAGCTTGCAGTGTG	GTTTCTTACCAAGCCCATGTAAGGAGC
96.	D12L434	43428051	43428122	AAAG	0.868	GTTTTCCCAGTCACGACAGAGGCAGGAGGATAGCTTGA	GTTTCTTTGTGAGCCTGTGTATGTTCATCA
97.	D12L441	44101186	44101246	AAAG	0.893	GTTTTCCCAGTCACGACGGCAACACATCATTGGGCAT	TGTTTGTTTGCTTATACCAACAGG
98.	D12L616	6166647	6166693	AAT	0.873	GTTTTCCCAGTCACGACCCCACGTAGCTCCCAAGTAG	GAGGGCAGGAGTTCGAGAC
99.	D12L630	63076807	63076866	AAT	0.877	GTTTTCCCAGTCACGACAGACTACAGGCAGGTACCACT	ACAAGATGTGGTACGTGGTCA
100.	D12L794	79458707	79458779	AAAG	0.862	GTTTTCCCAGTCACGACTCTCTTATTCCCTCAGCCCCA	GTTTCTTCAGCGTGCAAACCCTTAGAG
101.	D12L908	90823354	90823397	AGAT	0.870	GTTTTCCCAGTCACGACGCACATCTAGTAAACACTACCTAGC	GTTTCTTTGCTACCCTCTCTTTCTGAAACA
102.	D13L385	38541503	38541569	AAGG	0.936	GTTTTCCCAGTCACGACTAATCTTCGATTTCTGCTGTAACAA	AGGATTATTCTGGATTATCGGGG
103.	D13L459	45923472	45923533	AAAG	0.853	GTTTTCCCAGTCACGACTTCTTACAACTTTCTTCGGTGTAGG	GTTTCTTTCTGTCCATTACAGATCATGAAAAA
104.	D13L516	51688743	51688784	AGAT	0.834	GTTTTCCCAGTCACGACTTCTCATTCCTCCCTGCACC	GTTTCTTACTCCTCATGTTGTACGTTAGAGT
105.	D13L576	57688832	57688897	AAGG	0.853	GTTTTCCCAGTCACGACTGGGGAACAGAGTGAGACT	GTTTCTTGGTGCTAAACATCATTAATCATCAGA
106.	D14L194	19452207	19452256	AAT	0.850	GTTTTCCCAGTCACGACGAGGTCAAGGCTGCAGTGAG	ATGATCTGTTCTCTCTCTGGAAGC
107.	D14L276	27688436	27688493	AAAG	0.881	GTTTTCCCAGTCACGACTGTTCAGTTCTTTGTTCTCCATTT	GTTTCTTTAGCTGGACTTGGTGGCG
108.	D14L364	36424694	36424723	AAT	0.857	GTTTTCCCAGTCACGACGAGCCAAGATCACGCCACT	GTTTCTTCACACACATGAACTCAACCCAC
109.	D14L372	37229537	37229597	AAGG	0.861	GTTTTCCCAGTCACGACTCGCTCCTGGGATGTGCC	GTTTCTTCAGTATGGAGCTGGCAGGAC
110.	D14L409	40925722	40925767	AAT	0.919	GTTTTCCCAGTCACGACAGCTACTTGTGGGACCGAGG	TTGGACTCACTGTAGGCTCA
111.	D14L699	69965306	69965379	AAGAG	0.858	GTTTTCCCAGTCACGACGCTTCACTTTCCTCAGATGGC	GAAATTAGCTGGGTGTGGTGG
112.	D14L781	78104645	78104678	AAT	0.838	GTTTTCCCAGTCACGACCTGCAGCCTTGACCTCCTG	GTTTCTTCCAGGAGTTCAAGACCAGCC
113.	D14L785	78588181	78588247	AAAG	0.840	GTTTTCCCAGTCACGACGAGAGTAAGTACATCCTGCTGAG	GTTTCTTAGCAAGAATGAAGGCGGAAAG
114.	D14L953	95326741	95326782	AAAG	0.878	GTTTTCCCAGTCACGACTGCTATTTCTACTGTGGTTTGGT	GTTTCTTTCAGGAGTTCGAGGTTGCAG
115.	D15L258	25895801	25895861	AGAGG	0.802	GTTTTCCCAGTCACGACGAATAGAACAAAAAGGTGGAGGAAG	GTTTCTTTAATGCAGGTGGTATAATTCAGTCC
116.	D15L296	29670028	29670068	AAT	0.861	GTTTTCCCAGTCACGACCAGTGAGCCAAGATCGTACCA	GTTTCTTCCTCCTTTGTTCCAGCCAAC
117.	D15L495	49591907	49591992	AAAG	0.900	GTTTTCCCAGTCACGACTAGTTTTCATGGGATTGTATGGTTT	GTTTCTTGCACGTATTATAAGACTGACCACAT
118.	D15L752	75225761	75225807	AGAT	0.901	GTTTTCCCAGTCACGACGATCACTTGAGGCCAGGAGT	TGCACTTTCAGTATTCAGTTAAGGC
119.	D16L104	10466424	10466477	AAAG	0.869	GTTTTCCCAGTCACGACACAAGAAAAGAAAAGGGGCTCC	GTTTCTTTGAAATCATGCCACTGTACTCCA
120.	D16L547	5475067	5475123	AGAT-CTTT	0.869	GTTTTCCCAGTCACGACTGTTGGGGTTTGGTGTATGAA	GTGAGCCAAGATCGCACC
121.	D16L554	55495723	55495777	AAAG	0.906	GTTTTCCCAGTCACGACTGAATACTTGTAAAACACCTGGCA	GTTTCTTACAGAGTGAGACCTTCCAAGAG
122.	D16L732	73202111	73202178	AAAG	0.872	GTTTTCCCAGTCACGACCCCCATAAATGCCAGCCTTTTA	GTTTCTTAGTCTGCATTTCTACAACAAAACA
123.	D17L172	17264525	17264552	AGAT	0.833	GTTTTCCCAGTCACGACACCTACTTAACCTTATTCTACCAAC	GTTTCTTCAAGCTTGCCAAACTCTTCA
124.	D17L255	25565619	25565687	AAAG	0.835	GTTTTCCCAGTCACGACCTGAGCAATAGAGTGAAACTGCCT	GTTTCTTGGGGAAGAAGTAATGACAATGAGTA
125.	D17L388	38838897	38838962	AAAG	0.844	GTTTTCCCAGTCACGACAGGAAGTGACATCAAGAACAATAG	GTTTCTTCTGAGTGACAGAGTGAGACTCCAT
126.	D17L432	43294887	43294947	AAGG	0.864	GTTTTCCCAGTCACGACGCAAGAGATCTTTCCTTTCTTTCC	GTTTCTTAACAGAGCAAGACTCCATCTCG

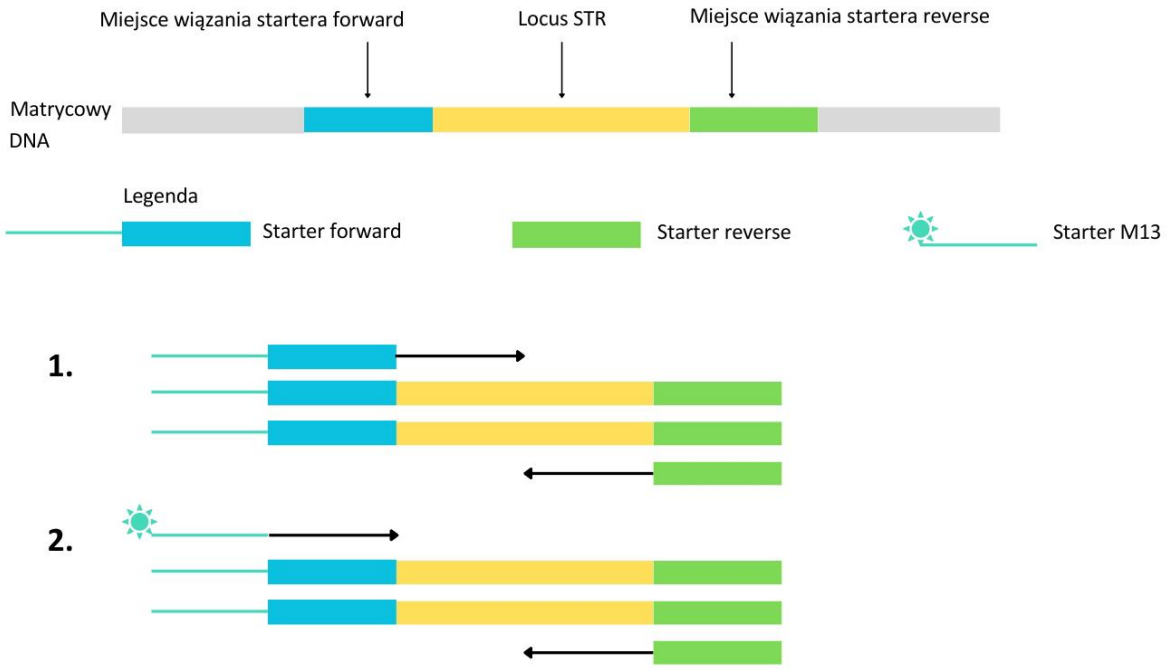
Lp.	Locus	Początek	Koniec	Motyw	Het.	Starter forward (5'-3')	Starter reverse (5'-3')
127.	D18L119	11940262	11940333	AAAG	0.869	GTGAAACCCCATCTCTGCTA	GCCTTGTGTCTATTAAACTCTACTA
128.	D18L4991	49913945	49914013	AAGG	0.906	GTTTTCCCAGTCACGACTAAATTACAGGAGACCAGAGAGCG	GTTTCTTTTCATTGCAGCAGAATATTTGAAAGAG
129.	D18L4995	49954178	49954264	AAGG	0.838	GTTTTCCCAGTCACGACCAGGGTCCAAGGCCAAGTT	GTTTCTTCAGCTGCGTGAGATAATACTTTGA
130.	D18L689	68913552	68913599	AAT	0.871	GTTTTCCCAGTCACGACACCATTGACTCACTGCCCAA	GGCTGAGGTGAAGGATTGCT
131.	D19L240	2408989	2409030	AAT	0.857	GTTTTCCCAGTCACGACGTGGGAGGCTCCATCTCAA	GTTTCTTACAGAATCAGAGAATAATGCAGACC
132.	D19L298	29876730	29876796	AAAG	0.878	GTTTTCCCAGTCACGACAGACTTCACTTCCTCCAAGATCAG	GTTTCTTCCACAGAGCAAGATCTCATCTAAA
133.	D19L762	7628330	7628391	AAT	0.859	GTTTTCCCAGTCACGACAGTCACAGTTGCTTCACTTGTTT	GTTTCTTTGGGTGACAGAGTGGGATCT
134.	D19L901	9011984	9012055	AAAAT	0.811	GTTTTCCCAGTCACGACCTGGTGTGCAGTAAGACTATGTTTC	GTTTCTTCTCTTCATTGGTCTCACGTCTG
135.	D20L110	60909810	60909839	AGGGCG	0.903	GTTTTCCCAGTCACGACCCGCCAATGTCACACCGG	GTTTCTTCACTCGCTCGTCCCAGAC
136.	D20L193	19302989	19303073	AAGG	0.873	GTTTTCCCAGTCACGACGCTTTGTCTATGTAAGGTGCTCC	GTTTCTTTTGGAAATTCTGTTAGGACCATGTC
137.	D20L226	22677741	22677821	AAGAG	0.903	GTTTTCCCAGTCACGACCTGTGAAAACCTGAGGCCCT	GCAACGTAGGAGAGAGAGGG
138.	D21L110	11077556	11077590	AAAT	0.868	GTTTTCCCAGTCACGACATGCCTTGCTTCAAAGGTCAG	GTTTCTTAGCTGCATTAGCCCCTAACA
139.	D21L291	29133704	29133778	AAAG	0.856	GTTTTCCCAGTCACGACGGTGAAGGCACTCAAGAGATATTTA	GTTTCTTGAAAGAGGGTAAGAGACAGAAGAGC
140.	D21L317	31706806	31706875	AGAT	0.854	GTTTTCCCAGTCACGACTGAATAGTGTTTGGCCTCTCTCA	GTTTCTTTGGTTTCCTTGAGTTTCCAGGT
141.	D22L375	37566130	37566172	ATCC	0.889	GTTTTCCCAGTCACGACTATGCTGCCTGGTATGTGCC	CACACCCCATCTCTTCTGCA
142.	D22L505	50519439	50519487	AAAG	0.897	GTTTTCCCAGTCACGACAGCCTAGAATGTAACTCACGCA	TGGGGTTTCACTATGTTGGCC
143.	D3S2406			[TATC]a [TGTC]b [CGTC]c [CATC]d	0.986	CGCCAGGGTTTTCCCAGTCACGACGTGGGGCAGTTTGAGTCTGACC	GAGCCACATGGAGAGGCTTTAG
144.	D1N10	230905363	230905429	[TATC]a-ATC-[TATC]b	0.937	TAACGGGAATTGACCAGGTAGGC	GGTAGAGATGGAAGAAAATCCCCCATA
145.	D07L5211	52117027	52117114	AAAAG	0.9	GTTTTCCCAGTCACGACATTGAGAATTCAATACTTTGATAC	GTACTCATTCAAGTGAATAATTCATATG
146.	D19L298	29876730	29876796	AAAG	0.9	GTTTTCCCAGTCACGACAGACTTCACTTCCTCCAAGATCAG	GCACACCAGGCTAAGAAAGTTGG
147.	D12L131	13174142	13174180	AAC	0.9	GTTTTCCCAGTCACGACTTGTTCAAGATTTTAAGTGTGG	CTGCGGATGCTGCCCCACGA
148.	D03L101	101684290	101684317	AAT	0.8	GTTTTCCCAGTCACGACCTTGTCATGGGTGATAAGTA	CACAAAGTACCTAATGCCTGGC
149.	D05L105	105979402	105979440	AAT	0.8	GTTTTCCCAGTCACGACACTCCTAGAACAAGTAAGATCACC	GCCAGGGGACAACAGTGAAA
150.	D21S205	39819523	39819659	GATA	0.848	GTTTTCCCAGTCACGACGCCATTACCATATGAGTTAGTC	CAAATCTTGAACTCATCCTC
151.	D02L234	234570270	234570328	AAAG	0.9	GTTTTCCCAGTCACGACAAAGTGCAAGGTTTGAAGCC	AGCTGTGGTTGGCGATCATT
152.	D15L752	75225761	75225807	AGAT	0,901	GTTTTCCCAGTCACGACGATCACTTGAGGCCAGGAGT	TGCACTTTCAGTATTCAGTTAAGGC
153.	D01L679	6792427	6792507	AAAG	0.9	GTTTTCCCAGTCACGACGGATGACAGTGTGGGTTCTCT	GCTACTCGGGAGGCTGA
154.	D04L164	164955214	164955271	AGAT	0.9	GTTTTCCCAGTCACGACTGGAAGACAGTGTGGTGATTTCT	GTGGTGTTTGGTTTTCTATTCCCA
155.	D06L815	81581236	81581302	AGAT	0.8	GTTTTCCCAGTCACGACCAATCAGTAATGAAAATGTTATC	CATGGTGGGTGGGTGGATG
156.	D18S386	68127490	68127652	AAAG	0.952	GTTTTCCCAGTCACGACTGAGTCAGGAGAATCACTTGGAAC	CTCTTCCATGAAGTAGCTAAGCAG
157.	D15S822	27390752	27390811	AGAT	0.9	CGCCAGGGTTTTCCCAGTCACGACAGTCAACAGTCTCAGAGACC	GTACATGATGAGCTGCTTCTC
158.	D8A29	138636482	138636536	AAAGG	0.922	CGCCAGGGTTTTCCCAGTCACGACGAGCATGGAGTATCTGTGAAGC	GGGAGAGAGGGCACTTAGAGT
159.	D11N29	129963384	129963453	GAGAA	0.911	CGCCAGGGTTTTCCCAGTCACGACCCAGTTCAAGGTTAGCCAGG	CCTACATGGAACATGTTATGAC
160.	D2S1360	17310718	17310801	[AGAT]a [AGAC]b [AGAT]c	0.936	CGCCAGGGTTTTCCCAGTCACGACTAGAGTCAATGTATTATGAAACCTG	GATTGATGGGGTCTTTGTTCAAGG

Lp.	Locus	Początek	Koniec	Motyw	Het.	Starter forward (5'-3')	Starter reverse (5'-3')
161.	D4A38	92191665	92191716	[TAGA]a [TAGT]b [TACA]c [TAGA]d [TACA]e [TAGA]f [TACA]g [TGGA]f [TACA]g [TAGA]h	0.947	CGCCAGGGTTTTCCCAGTCAC GACCCTGTTTGCTGTTACATA GGCC	AGTCAAAAGTCTCCAATAGAACCT TTC
162.	D3A65	32165284	32165331	AGAT	0.8	CGCCAGGGTTTTCCCAGTCAC GACGCCCTCGATTCTCAAAAT GC	CACTACACAAGCATAGTCAG
163.	D1N21	80115411	80115486	AAAG	0.9	CGCCAGGGTTTTCCCAGTCAC GACGCATACAAACTTTTGCTT GAATG	TTGCTGTGTAGAATTTCTATCTTT
164.	D12N15	44101186	44101246	AAAG	0.8	CGCCAGGGTTTTCCCAGTCAC GACGTAGGCAACACATCATTG GGCATTG	GAAATTAAGTCACATCACCAGAC
165.	D15N26	95280196	95280281	AGAT	0.8	CGCCAGGGTTTTCCCAGTCAC GACGACATAAAGATACAGAG ACAGAC	CACTCACTGAACAGGCGAAAAGC
166.	D12N3	115894994	115895070	AAAAG	0.8	CGCCAGGGTTTTCCCAGTCAC GACGTGAGACTCCATCTGAAA GAGAGAA	GCAAGGAGTCAAGACTTCCAG
167.	D4N70	120055406	120055461	AGAT	0.9	CGCCAGGGTTTTCCCAGTCAC GACCACATAAATAGATGCAAC CTCTACC	GTTGCTGAGTGACAACTCAG
168.	D5N72	173748145	173748190	AGAT	0.8	CGCCAGGGTTTTCCCAGTCAC GACCCTAGATGAGACACAAA GGGATTG	GGGATTTCTTGACCTCCATAGCC
169.	D11N52	132008458	132008526	AGAT	0.7	CGCCAGGGTTTTCCCAGTCAC GACCTTGGTGTGGCATAGGA ATTTGTC	CTGTCCTTGGGCATCAGCTCC
170.	D11S236	19259601	19259719	[TATC]a [TGTC]b [TATC]c [TCCA]d [CCCA]e [TCCA]f	0.907	CGCCAGGGTTTTCCCAGTCAC GACGAGGTGCAAGAATGTCT GTCTGTC	GTTTCTACAGTGCATTGTATACG
171.	D6N71	132397638	132397707	[GAAA]a -A- [GAAA]b -AA- [GAAA]c	0.989	CGCCAGGGTTTTCCCAGTCAC GACGTGAATGCACACTATATG ATGG	GAACAGGCAGGATGTAATGCAAC
172.	D14N56	41415668	41415710	AGAT	0.9	CGCCAGGGTTTTCCCAGTCAC GACGATAAACACAGGAATAA AGCTAG	CTTAAGAAAATTTCTGTCCCAAG
173.	D2N43	42072419	42072494	AGAT	0.9	CGCCAGGGTTTTCCCAGTCAC GACGTTTTGAGACTCGGACTC CCAAG	GAGCTGAAATGCACTGTGTATTAG
174.	D8S1132	107328920	107329002	TCTA	0.867	GGCTAGGAAAGGTTAGTGGC	TATTGCTCGAAAGAGAGAGGG
175.	D3N61	89050930	89051003	AAAG	0.9	CGCCAGGGTTTTCCCAGTCAC GACCATAGGATTTGGCAATG GATCATG	GTAAGCCTTTATTGACTACACTG
176.	D13S742	25282949	25283113	AAAG	0,891	GTTTTCCCAGTCACGACGGGC TAGGAATGGAAATAGGTT	GGGCTAGGAATGGAAATAGGTTG TAC
177.	D1N16	218178446	218178512	AAGAG	0.9	CGCCAGGGTTTTCCCAGTCAC GACCAATTATGTTCAAGAGGG AGGAAATG	CCTTCTTTCTCTACCTTGAAGAC
178.	D8A26	106405526	106405617	[TTCC]a [CTTT]b- C- [CTTT]c	0.9491	CGCCAGGGTTTTCCCAGTCAC GACATAGCTGAAGCATGTGC	TTTGTCTTTGCATTATAC
179.	D7S3038	21266718	21266793	[TATC]a [TACC]b [CACC]c	0.9270	GTTTTCCCAGTCACGACCGCC AGGGTTTTCCCAGTCACGACC TGGAGCTGCATAGTGTCCTTT	GAAAATCATCCTGTGTGCTTTCCCC
180.	D04L164	164955214	164955271	AGAT	0.9	GTTTTCCCAGTCACGACTGGA AGACAGTGTGGTGATTTCT	GTGGTGTTTGGTTTTCTATTCCCA
181.	D19L298	29876730	29876796	AAAG	0.878	GTTTTCCCAGTCACGACAGAC TTCACTTCCTCCAAGATCAG	GTTTCTTCCACAGAGCAAGATCTC ATCTAAA

Lp.	Locus	Początek	Koniec	Motyw	Het.	Starter forward (5'-3')	Starter reverse (5'-3')
182.	D3N54	70651745	70651840	[GAAA]a -A- [GAAA]b	0.9137	GTTTTCCCAGTCACGACCTCC AATAAGCAGAAAGAG	GAGAATATGCATCTGTACAT
183.	D10S2325	12793050	12793125	TCTTA	0.8380	GTTTTCCCAGTCACGACGGCC AGTCAGCTAAAGGA	CACGAAAGAAGCCTTCTGAAGCC

4.2 Testowanie zaprojektowanych starterów do amplifikacji loci STR z wykorzystaniem metody M13-tailing

Dla wszystkich loci STR wskazanych w tabeli nr 1 zaprojektowano i zamówiono startery do reakcji PCR. Do startera forward na końcu 5' dołączono sekwencję jednego z dwóch starterów M13(-40) lub M13(-47), umożliwiającą wykorzystanie pośredniej metody znakowania produktów PCR jak w pkt. 3.6 (Ryc. 3).



Ryc. 3 Schemat reakcji PCR z wykorzystaniem metody M13-tailing. W pierwszym cyklu reakcji locus STR jest amplifikowane z wykorzystaniem starterów forward wydłużonego o sekwencję startera M13 i reverse (1). W kolejnych cyklach reakcji jedna z nici matrycowego DNA jest syntetyzowana z użyciem fluorescencyjnie znakowanego startera M13. Dzięki użyciu fluorescencyjnie znakowanego startera M13 produkty reakcji mogą być analizowane na drodze elektroforezy kapilarnej.

Sekwencje starterów opracowywano z wykorzystaniem programów komputerowych wskazanych w pkt. 3.5. Specyficzność i wydajność działania zaprojektowanych starterów w

reakcji PCR weryfikowano poprzez amplifikację danego locus z wykorzystaniem izolatów DNA od pojedynczych osób, metodą M13-tailing (pkt. 3.6) i analizę elektroforegramów mieszaniny poreakcyjnej (pkt. 3.13). Startery kwalifikowano do dalszych prac, jeżeli wydajność reakcji PCR była wysoka i analizowane piki elektroforetyczne w układzie heterozygotycznym miały wysokość minimum 2000 RFU dla zastosowanego analizatora genetycznego ABI3130XL, a także jeśli w elektroforegramie nie znajdowały się dodatkowe, niepożądane produkty reakcji (artefakty). Dodatkowym kryterium akceptacji starterów były niskie piki elektroforetyczne amplikonów typu stutter nie mogące przekraczać 20% wysokości prawidłowego piku elektroforetycznego odpowiadającego rzeczywistemu allelowi badanego locus STR. Startery reverse, dla których w elektroforegramie obserwowano rozdwojone piki elektroforetyczne modyfikowano dodając do 5' końca startera sekwencję PIG-tail GTTTCTT promującą końcową adenylację produktów reakcji PCR (pkt. 3.7). W przypadku uzyskania zbyt niskiej wydajności reakcji PCR lub zaobserwowania artefaktów startery projektowano od nowa i poddawano je kolejnej weryfikacji poprzez wykonanie reakcji PCR i ponowną analizę elektroforegramów. W przypadku niemożności uzyskania pożądanych wyników amplifikacji po trzykrotnej próbie zmiany sekwencji starterów taki locus eliminowano z dalszych prac badawczych. Do kolejnego etapu badań przesiewowych ze 183 wstępnie wybranych loci zakwalifikowano 172 loci STR.

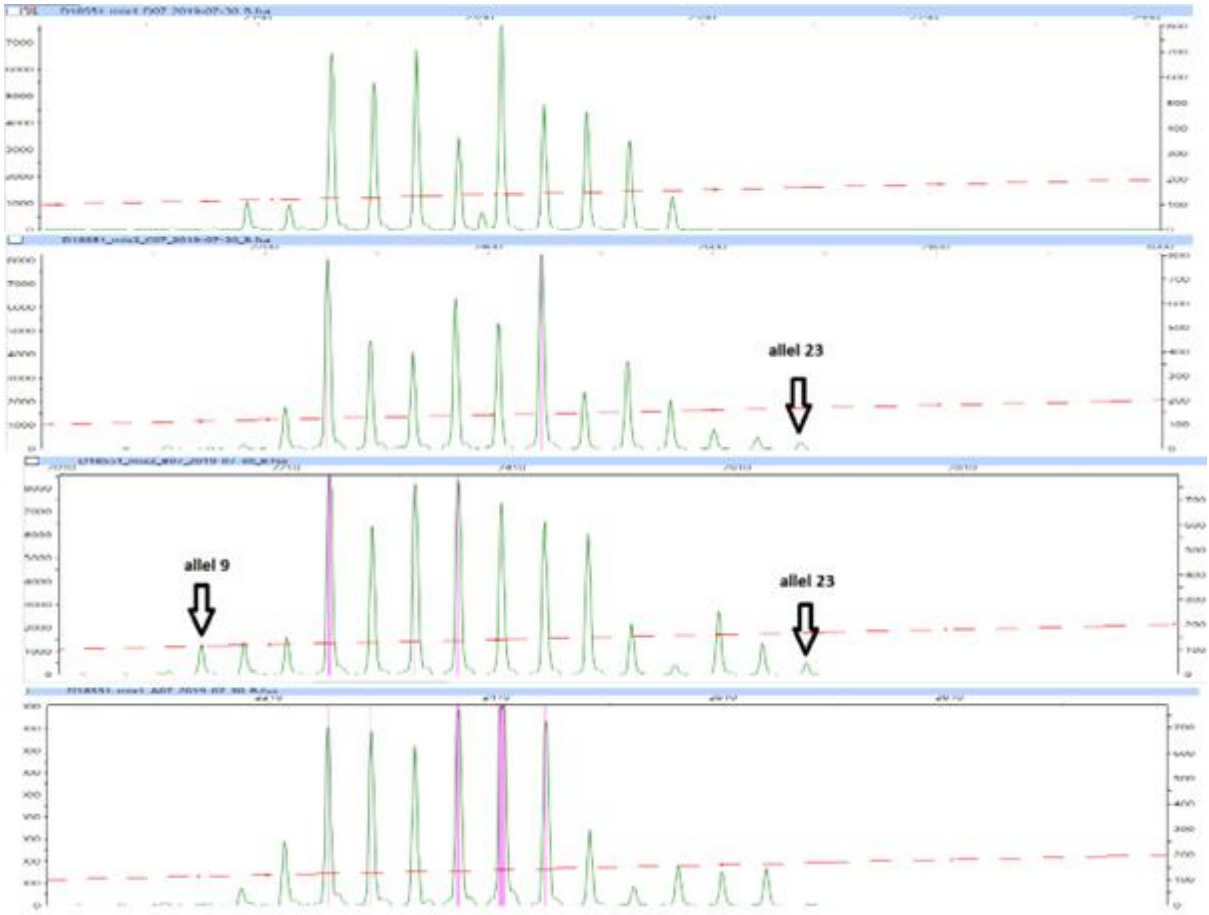
4.3 Szacowanie heterozygotyczności loci STR w populacji polskiej

Wyselekcjonowane 172 loci STR zostały poddane badaniu przesiewowemu w celu weryfikacji ich polimorfizmu w populacji polskiej oraz określenia rozpiętości locus, czyli zakresu długości alleli. Do badań przesiewowych wykorzystano metodę M13-tailing oraz wcześniej opracowane startery.

Szacowanie heterozygotyczności loci STR przeprowadzono wykonując po cztery reakcje PCR dla każdego analizowanego locus STR. Jako matrycy do każdej reakcji PCR użyto mieszaniny DNA pochodzącej od 50 losowo wybranych osób z populacji polskiej, innych dla każdej z czterech reakcji PCR (pkt. 3.8). W ten sposób w czterech reakcjach PCR zamplifikowano łącznie allele wszystkich badanych loci STR znajdujące się w genomach 200 badanych osób. Na podstawie analizy elektroforegramów szacowano częstość występowania alleli w populacji polskiej oraz heterozygotyczność badanych loci STR metodą opisaną w pkt. 3.14. W celu oceny

dokładności metody wykonano badania przesiewowe dla 13 loci CODIS, dla których częstości występowania alleli w populacji polskiej oraz heterozygotyczność są znane.

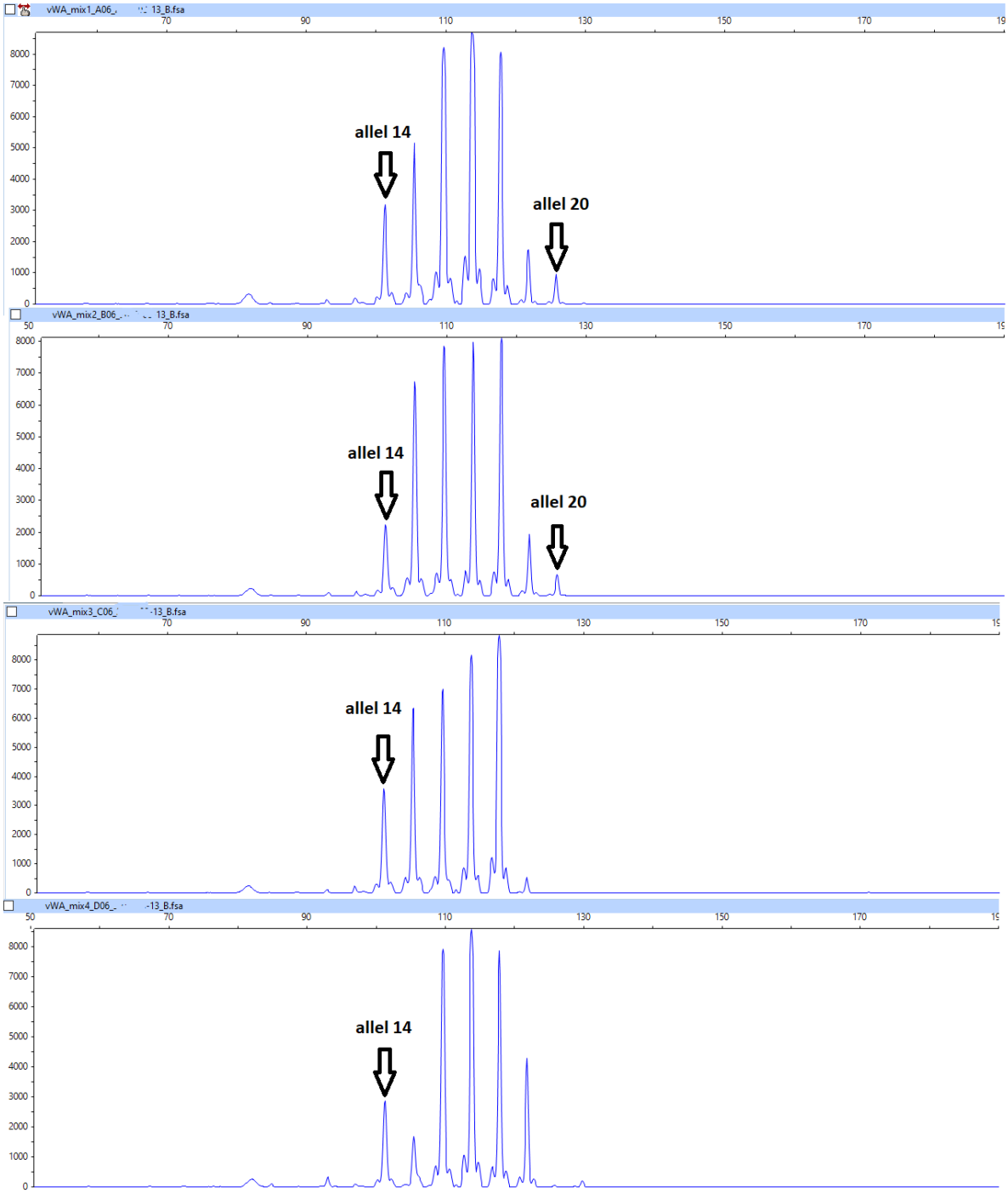
Wartości heterozygotyczności loci CODIS obliczone w ramach badań przesiewowych przeprowadzonych w niniejszej pracy doktorskiej były wysoce zgodne z wartościami heterozygotyczności tych loci w populacji polskiej znanymi z danych literaturowych (Ossowski i in., 2017). Przykładowe wyniki badań przesiewowych oraz obliczenia częstości występowania alleli dla dwóch loci CODIS D18S51 oraz vWA przedstawiono poniżej, odpowiednio na ryc. 4 i w tab. 3 oraz na ryc. 5 i w tab. 4.



Ryc. 4. Badania przesiewowe locus D18S51. Rysunek zawiera cztery elektroforegramy z amplifikacją locus D18S51 dla czterech mieszanin pulowanego DNA pochodzącego od 50 niespokrewnionych ze sobą osób każda. Rozkład pików elektroforetycznych oraz pole ich powierzchni w założeniu metody ma odpowiadać ich rozkładowi w populacji polskiej. Na elektroforegramach wskazany jest najkrótszy zidentyfikowany w puli DNA pochodzącego od 200 osób allel 9 występujący w populacji polskiej z częstością 0,1% oraz najdłuższy zidentyfikowany w dwóch elektroforegramach allel 23 występujący w populacji polskiej z częstością 0,2%.

Tabela 3. Wyniki badań przesiewowych locus D18S51. W tabeli w kolejnych kolumnach wskazano allele locus D18S51, pola powierzchni pików elektroforetycznych odpowiadających poszczególnym allelom w ujęciu procentowym w stosunku do sumy pól powierzchni wszystkich pików elektroforetycznych w elektroforegramie w każdej z czterech mieszanin DNA (Mieszanina DNA nr 1-4), średnią wartość stosunku pola powierzchni piku do pól powierzchni wszystkich pików w elektroforegramie dla czterech mieszanin DNA, znaną częstość występowania alleli w populacji polskiej wg Ossowski i in. 2017. W dolnym wierszu wskazano wartości heterozygotyczności dla locus D18S51 obliczone na podstawie badań przesiewowych przeprowadzonych w niniejszej rozprawie oraz wskazane w publikacji Ossowski i in., 2017.

Allel locus D18S51	**Mieszanina DNA nr 1**	**Mieszanina DNA nr 2**	**Mieszanina DNA nr 3**	**Mieszanina DNA nr 4**	**średnia 1-4.**	**Częstość występowania allelu**
9	-	--	2,19	-	0,55	0,1
10	2,26	-	2,44	1,28	1,49	0,7
11	2,04	2,63	2,75	4,60	3,00	1,9
12	14,2	15,06	14,88	12,95	14,27	9,20
13	11,93	10,50	11,10	12,60	11,53	9,20
14	14,54	10,03	14,22	11,62	12,60	15,15
15	8,49	14,19	14,52	14,22	12,86	17,85
15.2	1,46	-	-	-	0,36	-
16	16,45	11,00	12,82	14,50	13,56	18,45
17	10,14	16,06	11,41	13,30	12,73	12,00
18	8,50	5,03	10,61	5,49	7,40	6,650
19	7,25	7,80	3,80	1,38	5,06	3,90
20	2,7	4,30	0,71	2,88	2,65	2,40
21	-	1,75	4,74	2,43	2,23	1,30
22	-	1,05	2,33	2,73	1,52	0,80
23	-	0,62	0,87	-	0,37	0,20
Heterozygotyczność					88,7%	88,83



Ryc. 5. Badania przesiewowe locus vWA. Rysunek zawiera cztery elektroforegramy z amplifikacją locus vWA dla czterech mieszanin pulowanego DNA pochodzącego od 50 niespokrewnionych ze sobą osób każda. Rozkład pików elektroforetycznych oraz pole ich powierzchni w założeniu metody ma odpowiadać ich rozkładowi w populacji polskiej. Na elektroforegramach wskazany jest najkrótszy zidentyfikowany w puli DNA pochodzącego od 200 osób allel 14 występujący w populacji polskiej z częstością 0,1% oraz najdłuższy zidentyfikowany w dwóch elektroforegramach allel 20 występujący w populacji polskiej z częstością 1,9%. W badaniach przesiewowych nie zidentyfikowano allelu 13 występującego w populacji polskiej z częstością 0,5%.

Tabela 4. Wyniki badań przesiewowych locus vWA. W tabeli w kolejnych kolumnach wskazano allele locus vWA, wielkości pików elektroforetycznych odpowiadających poszczególnym allelom w ujęciu procentowym w stosunku do sumy wielkości wszystkich pików elektroforetycznych w elektroforegramie w każdej z czterech mieszanin DNA (Mieszanina DNA nr 1-4), średnią wartość stosunku wielkości piku do wielkości wszystkich pików w elektroforegramie dla czterech mieszanin DNA, częstość występowania alleli locus D18S51 w populacji polskiej wg. Ossowski i in. 2017. W dolnym wierszu wskazano wartości heterozygotyczności dla locus vWA obliczone na podstawie badań przesiewowych oraz wskazane w publikacji (Ossowski i in., 2017).

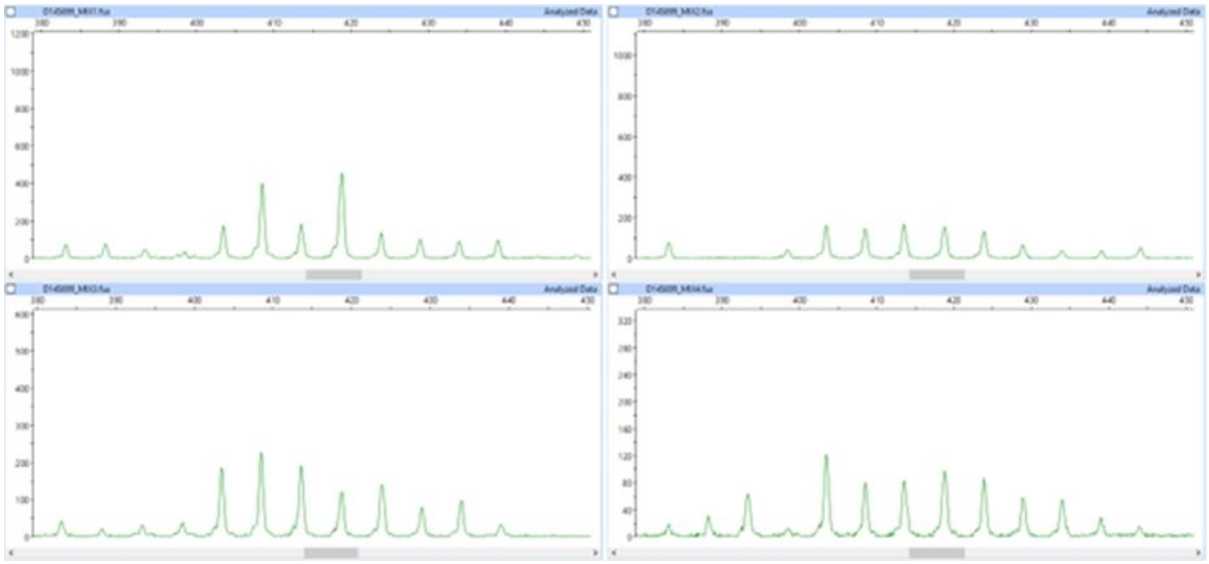
Allel locus vWA	Mieszanina DNA nr 1	Mieszanina DNA nr 2	Mieszanina DNA nr 3	Mieszanina DNA nr 4	średnia 1-4.	Częstość występowania allelu
13	-	-	-	-	-	0,5
14	8,87	6,3	10,4	8,66	8,56	9,00
15	14,29	17,98	17,40	5,04	13,68	10,60
16	22,81	22,13	20,35	23,88	22,29	17,90
17	24,08	23,42	24,66	25,82	24,50	30,20
18	22,42	22,80	25,67	23,69	23,64	22,80
19	4,86	5,46	1,56	12,9	6,19	7,00
20	2,67	1,89	-	-	1,14	1,90
21	-	-	-	-	-	-
22	-	-	-	-	-	-
Heterozygotyczność					80,36%	80,50%

Średnia różnica w heterozygotyczności wyznaczonej dla 13 loci CODIS metodą opracowaną w niniejszej pracy a heterozygotycznością tych loci w populacji polskiej znaną z literatury (Ossowski i in. 2017) dla wszystkich 13 analizowanych loci CODIS wyniosła 2,57%. Największą różnicę wyników zaobserwowano w przypadku locus D5S818 wynoszącą 7,64%, zaś najmniejszą w przypadku loci D18S51 oraz vWA: odpowiednio 0,13% oraz 0,17%. Dane dotyczące heterozygotyczności 13 loci CODIS obliczonej na podstawie badań przesiewowych oraz rzeczywistej heterozygotyczności tych loci zostały zestawione w tabeli nr 5.

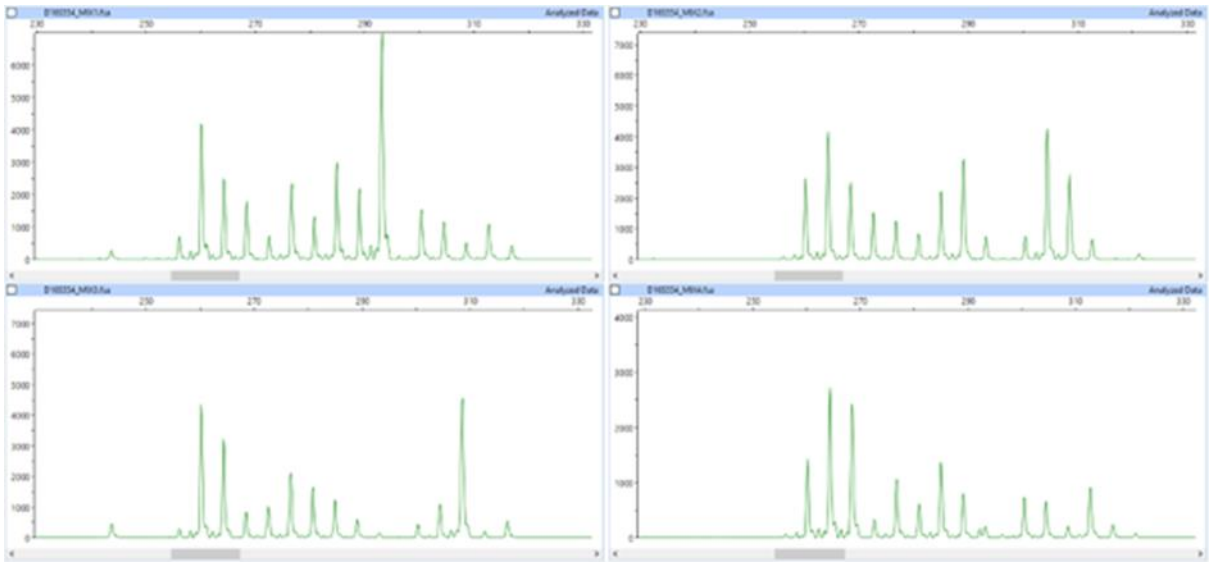
Tabela 5. Heterozygotyczność 13 podstawowych loci CODIS (CSF1PO, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D21S11, FGA), obliczona na podstawie częstości występowania alleli w populacji polskiej (Ossowski et al., 2017) - Het., w porównaniu do heterozygotyczności oszacowanej przy zastosowaniu metody badań loci STR opisywanej w niniejszej pracy (Het. N=200).

Locus CODIS	Het.	Het. N=200
CSF1PO	0,725	0,7546
TH01	0,7533	0,7169
TPOX	0,6167	0,6367
vWA	0,8050	0,8033
D3S1358	0,805	0,7828
D5S818	0,6817	0,7581
D7S820	0,7767	0,8055
D8S1179	0,79	0,8289
D13S317	0,79	0,7995
D16S539	0,7333	0,7755
D18S51	0,8870	0,8883
D21S11	0,8383	0,8197
FGA	0,845	0,854

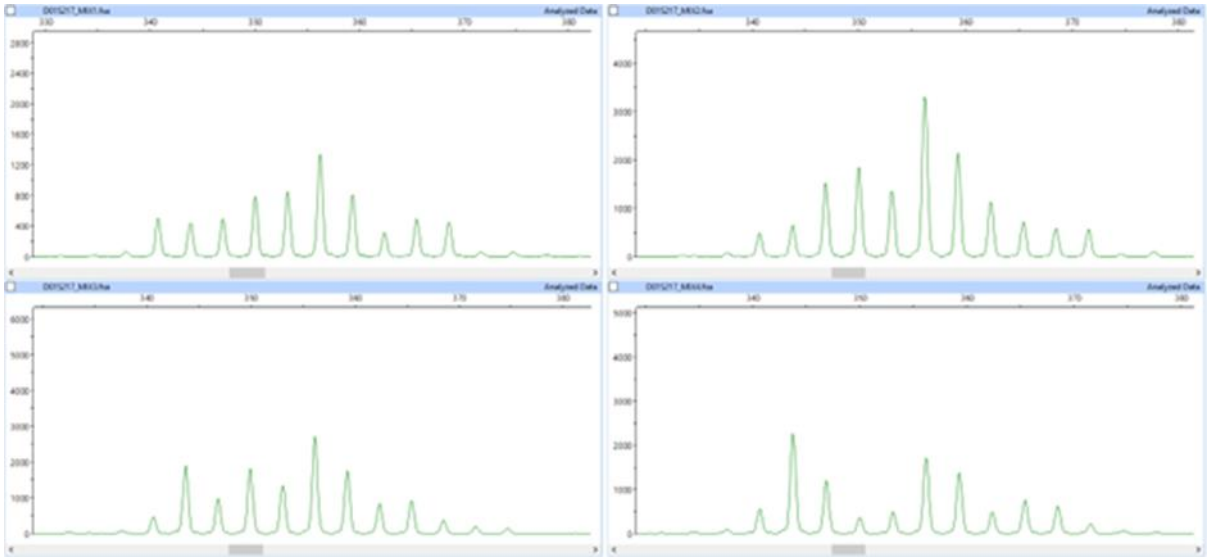
Wykorzystując opracowaną w niniejszej pracy metodę szacowania heterozygotyczności przeprowadzono badania przesiewowe dla wszystkich wyselekcjonowanych 172 loci STR. Elektroforegramy z badań przesiewowych dla dziesięciu wybranych nowo scharakteryzowanych loci STR (D14L699, D16L554, D01L217, D02L142, D03L109, D05L140, D14L276, D07L134, D01L228, D05L169) zostały przedstawione na rycinach 6-15. Wartości częstości występowania alleli dla tych loci w populacji polskiej wyznaczone na podstawie badań przesiewowych zostały wskazane w tabeli nr 6.



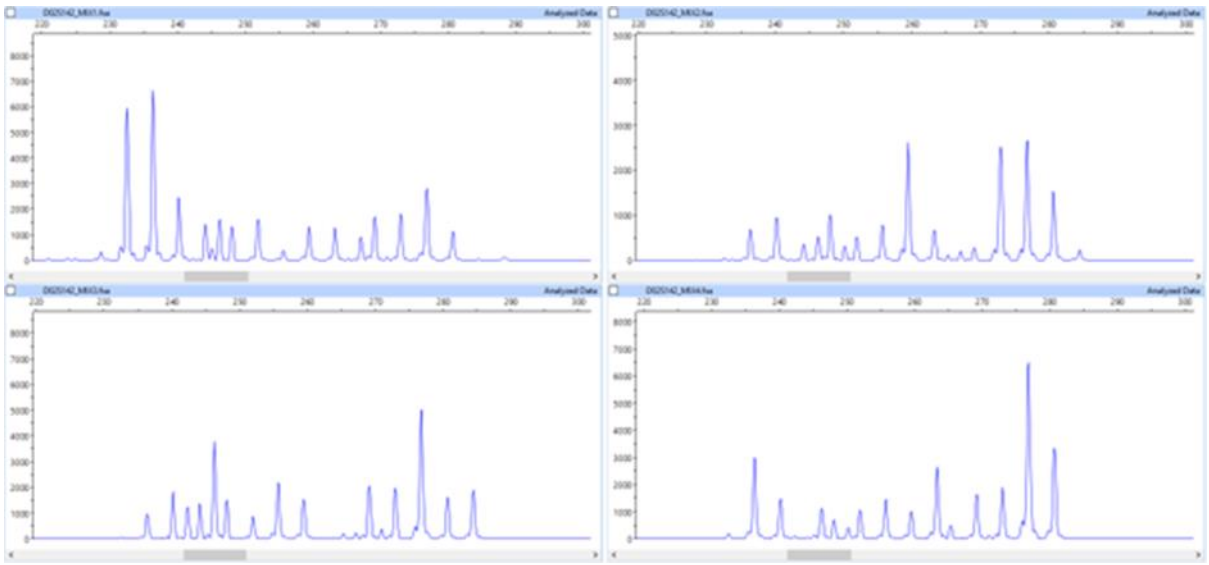
Ryc. 6 Locus D14L699. W obrębie locus zidentyfikowano 13 alleli. Rozpiętość locus (różnica długości pomiędzy najdłuższym, a najkrótszym zidentyfikowanym allelem) wyniosła 62 pz.



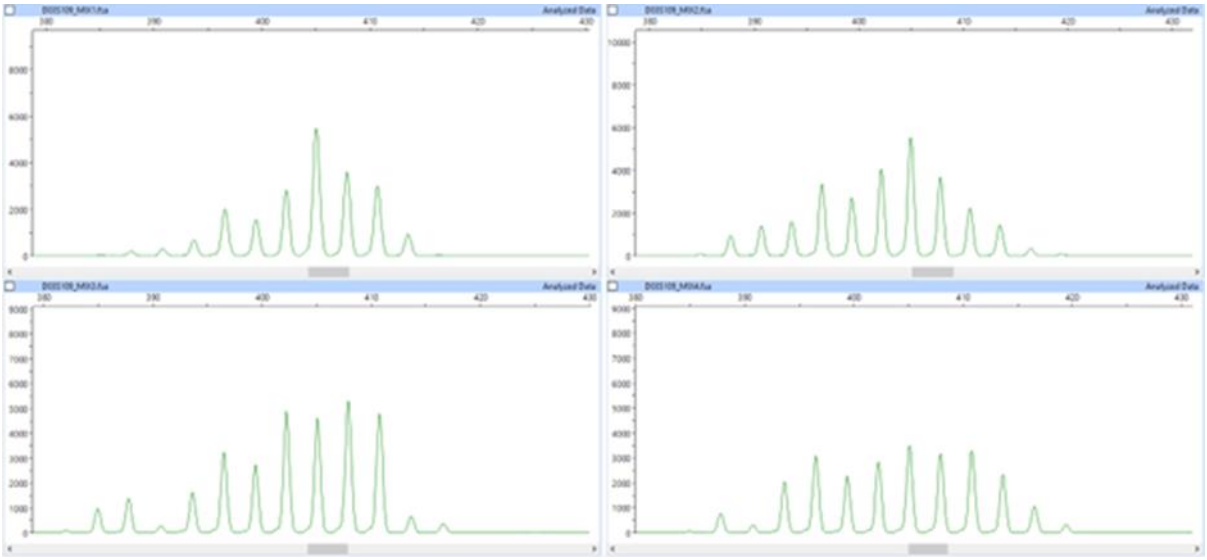
Ryc. 7. Locus D16L554. W obrębie locus zidentyfikowano 17 alleli. Rozpiętość locus wyniosła 78 pz.



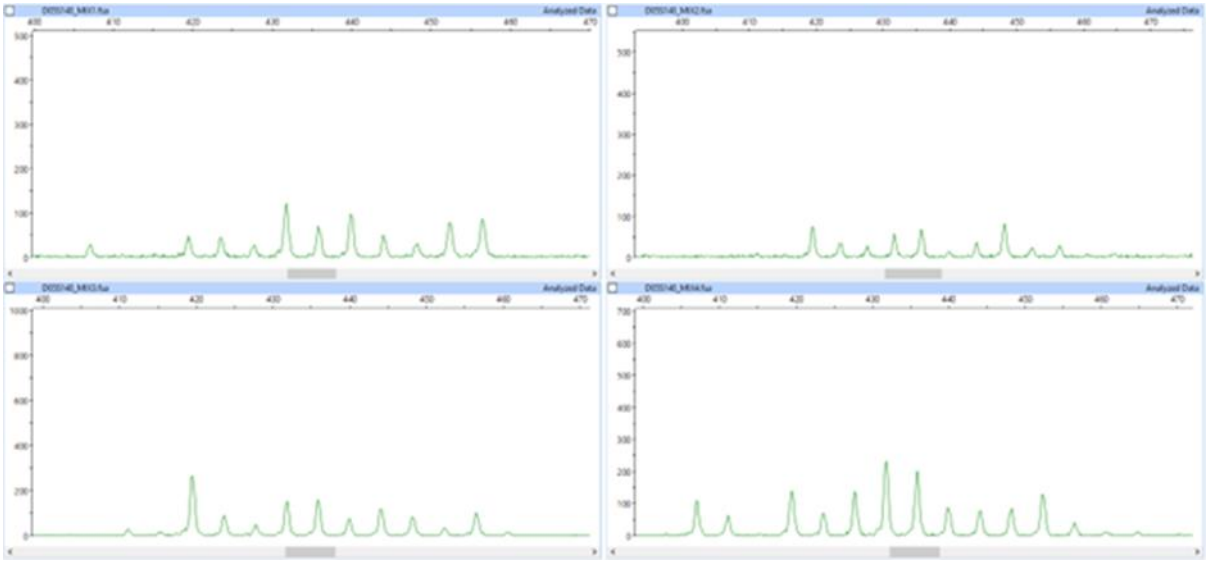
R**yc. 8** Locus D01L217. W obrębie locus zidentyfikowano 15 alleli, rozpiętość locus wyniosła 41 pz.



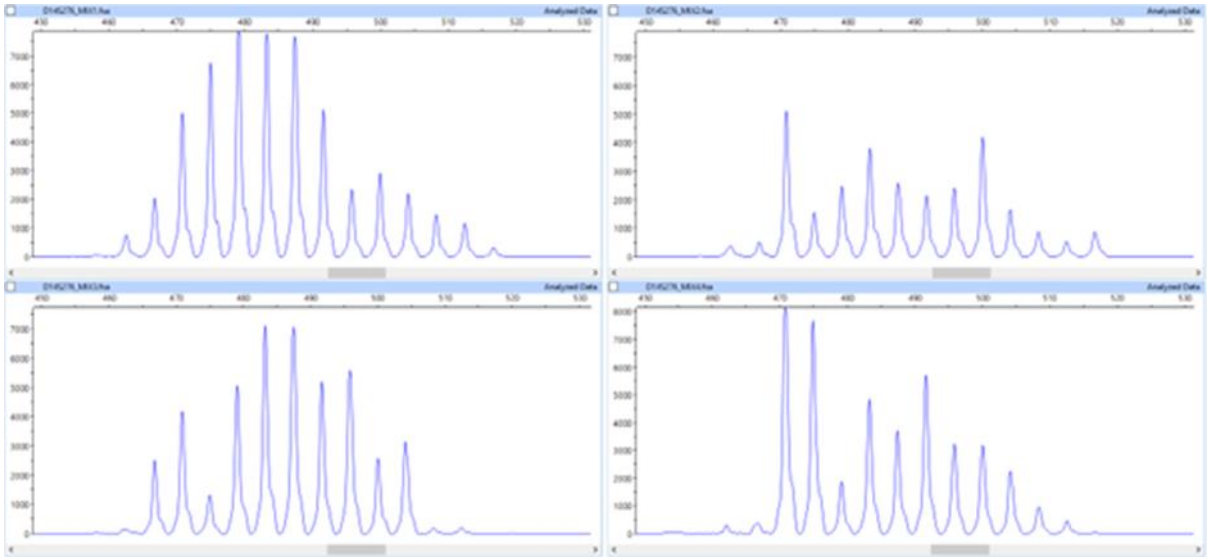
Ryc. 9. Locus D02L142. W obrębie locus zidentyfikowano 20 alleli, rozpiętość locus wyniosła 60 pz.



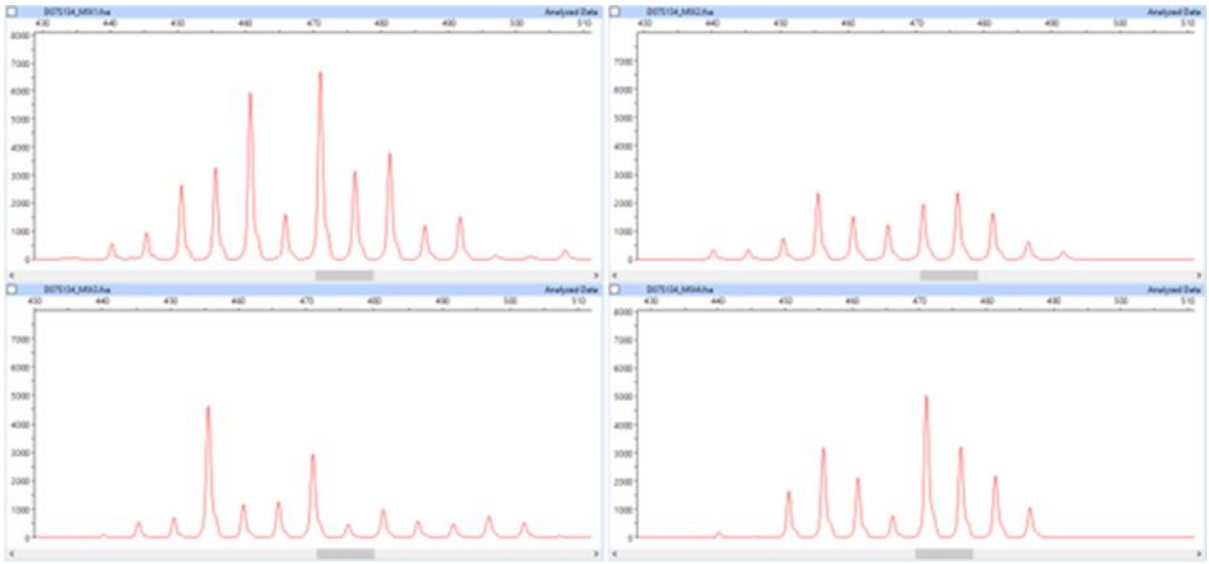
Ryc. 10 Locus D03L109. W obrębie locus zidentyfikowano 13 alleli, rozpiętość locus wyniosła 38 pz.



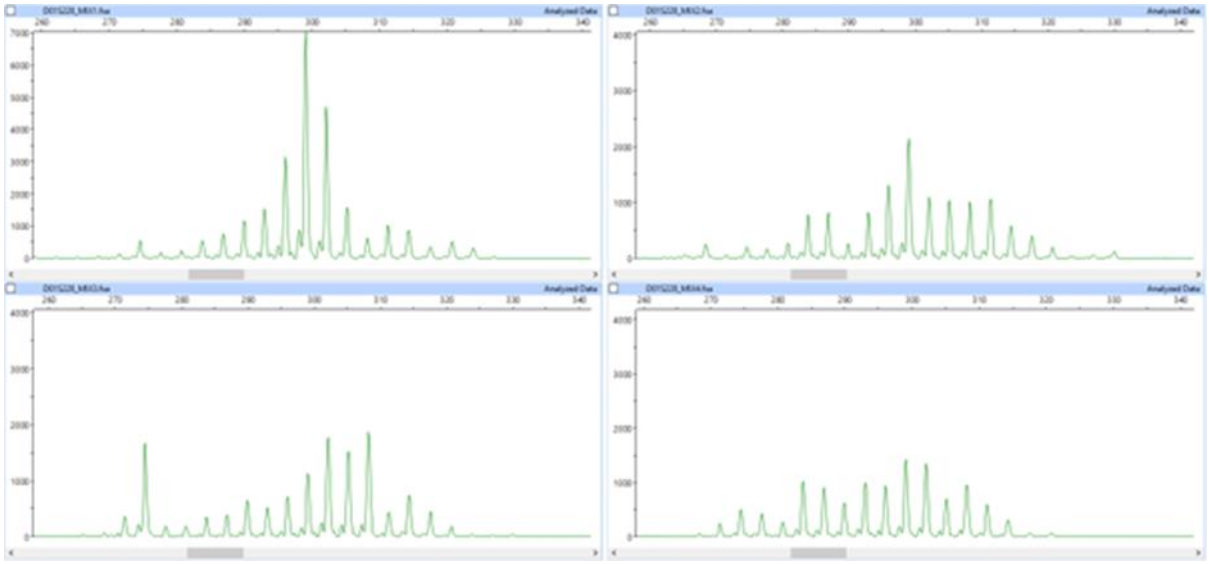
Ryc. 11 Locus D05L140. W obrębie locus zidentyfikowano 14 alleli, rozpiętość locus wyniosła 58 pz.



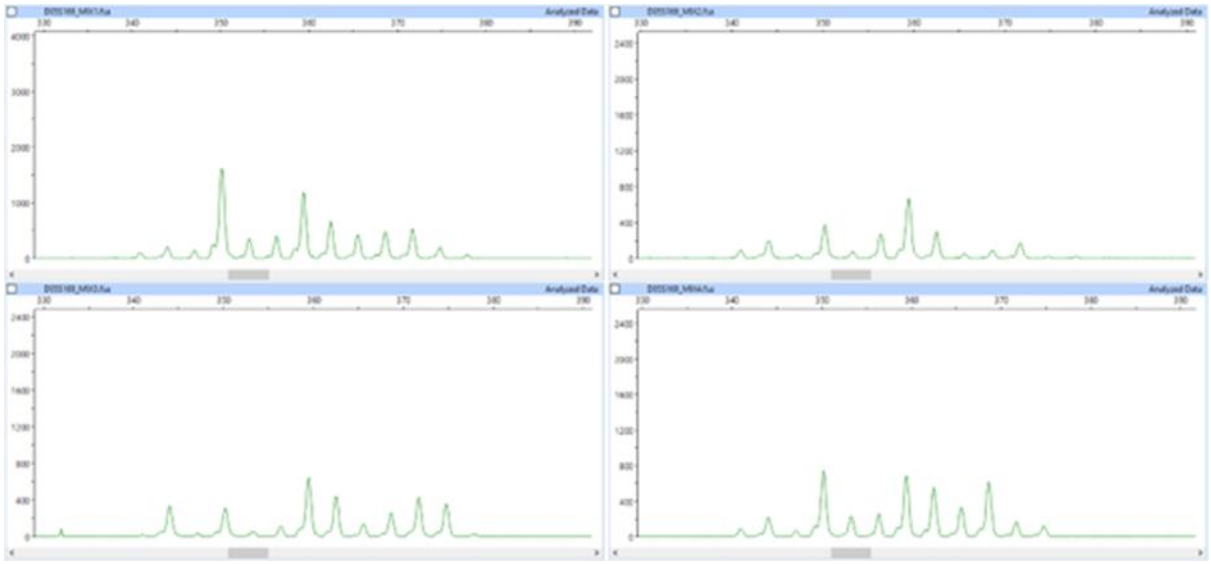
Ryc. 12 Locus D14L276. W obrębie locus zidentyfikowano 14 alleli, rozpiętość locus wyniosła 55 pz.



Ryc. 13 Locus D07L134. W obrębie locus zidentyfikowano 13 alleli, rozpiętość locus wyniosła 94 pz.



Ryc. 14 Locus D01L228. W obrębie locus zidentyfikowano 19 alleli, rozpiętość locus wyniosła 56 pz.



Ryc. 15 Locus D05L169 W obrębie locus zidentyfikowano 14 alleli, rozpiętość locus wyniosła 46 pz.

Tabela 6. Częstości alleli w przykładowych 10 loci STR, które zostały wybrane do zaprojektowania dwóch reakcji multipleks-PCR. Częstości występowania alleli w populacji polskiej zostały obliczone metodą opracowaną w niniejszej pracy doktorskiej (pkt. 3.14). Allel jest definiowany przez długość amplikonu w pz. DNA nr 1-4 oznacza częstość występowania danego allelu w puli 50 testowanych osób.

Locus	Allel	DNA nr 1	DNA nr 2	DNA nr 3	DNA nr 4	Średnia 1-4
D14L699	383	0	0,0638	0,0348	0,0000	0,0332
	388	0,0373	0,0000	0,0139	0,0369	0,0220
	393	0,0262	0,0000	0,0239	0,0906	0,0352
	398	0,0157	0,0378	0,0338	0,0000	0,0218
	403	0,0891	0,1438	0,1493	0,1762	0,1396
	408	0,2096	0,1373	0,1801	0,1074	0,1586
	414	0,0982	0,1600	0,1632	0,1158	0,1343
	419	0,2534	0,1524	0,1065	0,1443	0,1642
	424	0,0747	0,1265	0,1174	0,1225	0,1103
	429	0,0570	0,0616	0,0657	0,0822	0,0666
	434	0,0537	0,0346	0,0836	0,0839	0,0639
	439	0,0511	0,0292	0,0279	0,0403	0,0371
	444	0,0000	0,0530	0,0000	0,0000	0,0132
D16L554	247	0,1569	0,1406	0,1617	0,1273	0,1466
	251	0,0757	0,1837	0,1481	0,1964	0,1510
	255	0,0513	0,0965	0,0428	0,1816	0,0930
	259	0,0155	0,0344	0,0206	0,0255	0,0240
	263	0,0603	0,0483	0,1367	0,0664	0,0779
	267	0,0495	0,0333	0,0958	0,0417	0,0551
	272	0,0728	0,1007	0,0701	0,1179	0,0904
	276	0,0697	0,0529	0,0608	0,0654	0,0622
	280	0,2649	0,0712	0,0172	0,0334	0,0967
	287	0,0672	0,0405	0,0288	0,0266	0,0408
	291	0,0634	0,1168	0,0328	0,0411	0,0635

Locus	Allel	DNA nr 1	DNA nr 2	DNA nr 3	DNA nr 4	Średnia 1-4
	295	0,0308	0,0535	0,1697	0,0249	0,0697
	299	0,0222	0,0277	0,0149	0,0520	0,0292
D01L217	338	0,0105	0,0000	0,0057	0,0088	0,0062
	341	0,0741	0,0330	0,0330	0,0537	0,0484
	344	0,0643	0,0437	0,1377	0,2198	0,1164
	347	0,0729	0,1041	0,0710	0,1153	0,0908
	350	0,1182	0,1282	0,1348	0,0344	0,1039
	353	0,1265	0,0932	0,1004	0,0483	0,0921
	356	0,2015	0,2314	0,2014	0,1683	0,2007
	359	0,1228	0,1484	0,1306	0,1357	0,1344
	362	0,0463	0,0773	0,0621	0,0486	0,0586
	365	0,0750	0,0494	0,0699	0,0759	0,0676
	368	0,0000	0,0000	0,0000	0,0618	0,0154
	369	0,0692	0,0414	0,0277	0,0000	0,0346
	372	0,0088	0,0390	0,0145	0,0231	0,0213
	375	0,0100	0,0035	0,0114	0,0063	0,0078
	378	0,0000	0,0076	0,0000	0,0000	0,0019
D02L142	233	0,1847	0,0000	0,0000	0,0000	0,0462
	236	0,2049	0,0426	0,0344	0,1107	0,0981
	240	0,0780	0,0601	0,0601	0,0540	0,0631
	242	0,0000	0,0000	0,0400	0,0000	0,0100
	244	0,0399	0,0196	0,0440	0,0000	0,0259
	246	0,0459	0,0321	0,1344	0,0421	0,0636
	248	0,0365	0,0622	0,0513	0,0237	0,0434
	250	0,0000	0,0182	0,0000	0,0144	0,0082
	252	0,0504	0,0293	0,0303	0,0372	0,0368
	256	0,0114	0,0493	0,0783	0,0530	0,0480

Locus	Allel	DNA nr 1	DNA nr 2	DNA nr 3	DNA nr 4	Średnia 1-4
	259	0,0421	0,1678	0,0564	0,0000	0,0666
	260	0,0000	0,0000	0,0000	0,0380	0,0095
	263	0,0400	0,0415	0,0000	0,0968	0,0446
	265	0,0000	0,0000	0,0000	0,0166	0,0041
	267	0,0266	0,0116	0,0000	0,0000	0,0095
	269	0,0529	0,0158	0,0734	0,0614	0,0509
	273	0,0575	0,1678	0,0705	0,0705	0,0916
	277	0,0924	0,1804	0,1951	0,2529	0,1802
	281	0,0368	0,1017	0,0601	0,1288	0,0819
	284	0,0000	0,0000	0,0716	0,0000	0,0179
D03L109	385	0,0000	0,0000	0,0459	0,0000	0,0115
	387	0,0000	0,0000	0,0645	0,0184	0,0207
	388	0,0140	0,0364	0,0000	0,0000	0,0126
	389	0,0000	0,0000	0,0000	0,0690	0,0172
	390	0,0171	0,0424	0,0126	0,0000	0,0180
	393	0,0521	0,0734	0,0598	0,0230	0,0521
	396	0,1414	0,1490	0,1334	0,0483	0,1180
	398	0,0000	0,0000	0,0000	0,1552	0,0388
	399	0,0966	0,1099	0,0781	0,0000	0,0711
	401	0,0000	0,0000	0,0000	0,1011	0,0253
	402	0,1440	0,1287	0,1327	0,0000	0,1013
	404	0,0000	0,0000	0,0000	0,0874	0,0218
	405	0,1985	0,1697	0,1134	0,0000	0,1204
	407	0,1671	0,1275	0,1756	0,1172	0,1469
	409	0,0000	0,0000	0,0000	0,1713	0,0428
	410	0,1198	0,0775	0,1419	0,0000	0,0848
	412	0,0000	0,0000	0,0000	0,1264	0,0316
	413	0,0495	0,0554	0,0251	0,0000	0,0325

Locus	Allel	DNA nr 1	DNA nr 2	DNA nr 3	DNA nr 4	Średnia 1-4
	415	0,0000	0,0000	0,0000	0,0483	0,0121
	416	0,0000	0,0212	0,0171	0,0000	0,0096
	418	0,0000	0,0000	0,0000	0,0345	0,0086
	419	0,0000	0,0088	0,0000	0,0000	0,0022
D05L140	304	0,0266	0,0555	0,1029	0,0099	0,0487
	308	0,0161	0,0231	0,0588	0,0293	0,0318
	317	0,0512	0,0565	0,0927	0,1515	0,0880
	321	0,0710	0,0578	0,0745	0,0613	0,0662
	325	0,0576	0,0625	0,0966	0,0854	0,0755
	329	0,2088	0,1674	0,1505	0,1108	0,1594
	333	0,1229	0,1236	0,1299	0,0970	0,1184
	337	0,0731	0,0808	0,0446	0,0561	0,0636
	341	0,0892	0,1196	0,0000	0,0774	0,0715
	342	0,0000	0,0000	0,0830	0,0000	0,0207
	346	0,0532	0,0555	0,0389	0,1789	0,0816
	350	0,1138	0,0808	0,0753	0,0648	0,0837
	354	0,1039	0,1038	0,0206	0,0433	0,0679
	358	0,0126	0,0131	0,0316	0,0345	0,0229
D14L276	463	0,0144	0,0158	0,0000	0,0000	0,0076
	467	0,0380	0,0167	0,0542	0,0000	0,0272
	471	0,0905	0,1676	0,0919	0,2085	0,1396
	475	0,1210	0,0522	0,0314	0,1713	0,0940
	479	0,1558	0,0834	0,1140	0,0439	0,0993
	483	0,1493	0,1302	0,1640	0,1148	0,1396
	487	0,1423	0,0885	0,1625	0,0871	0,1201
	492	0,0933	0,0765	0,1201	0,1356	0,1064
	496	0,0438	0,0847	0,1298	0,0766	0,0837

Locus	Allel	DNA nr 1	DNA nr 2	DNA nr 3	DNA nr 4	Średnia 1-4
	500	0,0543	0,1459	0,0589	0,0754	0,0836
	504	0,0415	0,0558	0,0733	0,0535	0,0560
	508	0,0280	0,0307	0,0000	0,0225	0,0203
	512	0,0221	0,0190	0,0000	0,0107	0,0130
	517	0,0059	0,0328	0,0000	0,0000	0,0097
D07L134	440	0,0164	0,0224	0,0000	0,0000	0,0097
	445	0,0289	0,0234	0,0353	0,0000	0,0219
	450	0,0791	0,0511	0,0000	0,0000	0,0326
	451	0,0000	0,0000	0,0418	0,0797	0,0304
	456	0,0989	0,1688	0,3027	0,1626	0,1832
	461	0,1864	0,1140	0,0753	0,1077	0,1208
	466	0,0505	0,0889	0,0797	0,0379	0,0642
	471	0,2207	0,1496	0,1956	0,2650	0,2077
	476	0,1030	0,1805	0,0327	0,1683	0,1211
	481	0,1255	0,1278	0,0684	0,1197	0,1103
	486	0,0402	0,0506	0,0422	0,0592	0,0481
	492	0,0505	0,0229	0,0340	0,0000	0,0268
	497	0,0000	0,0000	0,0540	0,0000	0,0135
	502	0,0000	0,0000	0,0383	0,0000	0,0096
D01L228	244	0,0000	0,0127	0,0000	0,0000	0,0032
	247	0,0061	0,0096	0,0194	0,0216	0,0142
	250	0,0288	0,0400	0,0881	0,0284	0,0463
	253	0,0126	0,0160	0,0176	0,0335	0,0199
	256	0,0042	0,0180	0,0113	0,0176	0,0128
	259	0,0173	0,0434	0,0450	0,0696	0,0438
	262	0,0354	0,0560	0,0297	0,0724	0,0484
	265	0,0385	0,0147	0,0730	0,0377	0,0409

Locus	Allel	DNA nr 1	DNA nr 2	DNA nr 3	DNA nr 4	Średnia 1-4
	268	0,0523	0,0447	0,0425	0,0759	0,0539
	271	0,1112	0,0998	0,0644	0,1283	0,1009
	274	0,2509	0,1570	0,0820	0,1120	0,1505
	277	0,1783	0,0847	0,1470	0,0966	0,1266
	280	0,0754	0,0785	0,1019	0,0611	0,0792
	283	0,0386	0,0970	0,1158	0,0946	0,0865
	286	0,0500	0,0841	0,0335	0,0974	0,0662
	289	0,0407	0,0632	0,0687	0,0252	0,0494
	292	0,0303	0,0511	0,0428	0,0236	0,0369
	295	0,0184	0,0296	0,0174	0,0046	0,0175
	299	0,0109	0,0000	0,0000	0,0000	0,0027
D05L169	341	0,0149	0,0349	0,0000	0,0201	0,0175
	344	0,0337	0,0921	0,1167	0,0541	0,0741
	347	0,0178	0,0142	0,0076	0,0129	0,0131
	350	0,2671	0,1673	0,1082	0,1899	0,1831
	353	0,0519	0,0300	0,0173	0,0531	0,0381
	356	0,0571	0,1079	0,0314	0,0572	0,0634
	359	0,1904	0,2997	0,2088	0,1735	0,2181
	362	0,1020	0,0000	0,0000	0,1349	0,0592
	363	0,0000	0,1188	0,1380	0,0000	0,0642
	366	0,0661	0,0213	0,0414	0,0805	0,0523
	369	0,0767	0,0371	0,0784	0,1559	0,0870
	372	0,0838	0,0768	0,1388	0,0380	0,0844
	375	0,0297	0,0000	0,1134	0,0299	0,0432
	378	0,0088	0,0000	0,0000	0,0000	0,0022

Badania przesiewowe wykazały heterozygotyczność badanych loci w zakresie 0,611-0,939. Na podstawie tych danych, z puli 172 badanych loci wybrano 50 najbardziej polimorficznych do

opracowania metody Kinfinder. Lista loci oraz ich szacowana heterozygotyczność zostały zamieszczone w tabeli nr 6.

Szacowana heterozygotyczność 50 najbardziej polimorficznych loci wytypowanych do opracowania metody KinFinder zawierała się w przedziale 0,8105-0,9390. Najniższą heterozygotycznością w tej grupie cechowały się loci D12L908 (0,8105), D07L147 (0,8112) oraz D10L126 (0,8261), najwyższym zaś wskaźnikiem heterozygotyczności charakteryzowały się loci D13L742 (0,9390), D15L495 (0,9353) oraz D8A26 (0,9345). Średnia heterozygotyczność dla wszystkich 50 wytypowanych loci oszacowana na podstawie badań przesiewowych wyniosła 0,8762.

Tabela 7. Heterozygotyczność 50 loci wytypowanych do zestawu Kinfinder wskazana w bazie STRCatalog/WebSTR lub w danych literaturowych (Het. STRCat.) dla populacji światowej oraz heterozygotyczność obliczona na podstawie przeprowadzonych w niniejszej pracy doktorskiej badań przesiewowych dla populacji polskiej (Het N=200) dla grupy 200 osób.

Locus	Het. STRCat.	Het N=200	Locus	Het. STRcat.	Het N=200
D12L908	0,870	0,818	**D03L109**	0,869	0,917
D07L147	0,847	0,830	**D01L217**	0,875	0,888
D10L126	0,859	0,843	**D07L134**	0,874	0,872
D17L255	0,835	0,873	**D14L785**	0,84	0,869
D16L732	0,872	0,859	**D02L221**	0,861	0,863
D21L291	0,856	0,851	**D14L699**	0,858	0,885
D04L885	0,863	0,844	**D02L114**	0,9	0,870
D17L432	0,864	0,920	**D05L113**	0,858	0,899
D12L794	0,862	0,854	**D14L276**	0,881	0,898
D12L630	0,877	0,855	**D01L267**	0,889	0,850
D8S1132	0,867	0,833	**D1N16**	0,9	0,874
D06L106	0,864	0,890	**D03L115**	0,894	0,904
D07L101	0,836	0,865	**D05L140**	0,902	0,911
D10S2325	0,838	0,856	**D02L142**	0,887	0,917
D05L169	0,876	0,881	**D3N54**	0,9137	0,875
D20L226	0,903	0,863	**D16L554**	0,906	0,905
D2N43	0,9	0,833	**D12S391**	0,874	0,862
D7S3048	0,927	0,894	**D02L174**	0,911	0,922
D05L207	0,884	0,887	**D3A57**	0,8	0,934
D07L144	0,880	0,891	**D3N61**	0,9	0,900
D01L215	0,879	0,854	**D01L569**	0,863	0,901
D14L953	0,878	0,851	**D8A26**	0,9491	0,909
D09L159	0,862	0,866	**D15L495**	0,900	0,905
D03L194	0,883	0,861	**D13S742**	0,891	0,911
D08L110	0,892	0,832	**D1S1656**	0,874	0,880

4.4 Opracowanie dwóch multipleksowanych reakcji PCR do jednoczesnej amplifikacji 50 loci STR

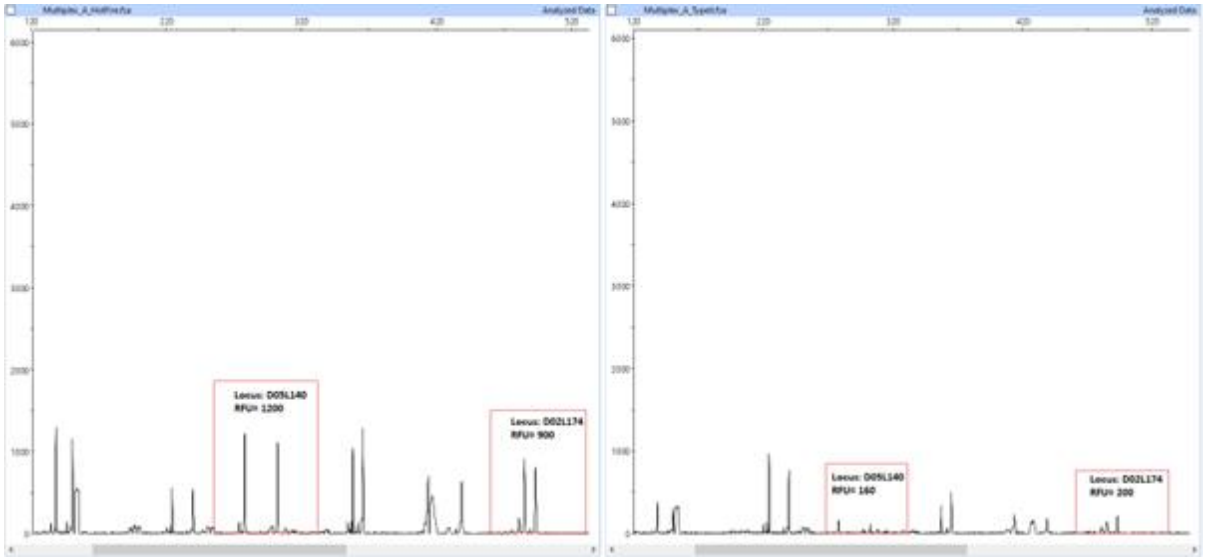
Ze 172 loci STR, które poddano badaniom przesiewowym wybrano 50 loci o najwyższym poziomie heterozygotyczności do opracowania dwóch reakcji multipleks-PCR (multipleks A i multipleks B), umożliwiających analizę długości alleli w próbkach ludzkiego DNA. W przypadku 21 loci sekwencje starterów do reakcji multipleks-PCR były takie same lub miały minimalnie zmodyfikowaną sekwencję nukleotydową, z usuniętym sztucznym starterem M13 na końcu 5', jak te zastosowane w badaniach przesiewowych (tabela nr 2). Dla pozostałych 29 loci startery zostały znacząco zmienione lub zaprojektowane na nowo z uwzględnieniem wcześniej określonych zakresów długości amplikonów, aby zapewnić skuteczne multipleksowanie reakcji i uniknąć możliwości nakładania się alleli różnych loci STR wyznakowanych tym samym barwnikiem fluorescencyjnym. Dodatkowo w celu umożliwienia identyfikacji płci badanej osoby zaprojektowano startery do amplifikacji locus amelogeniny (AMEL) i włączono je do reakcji multipleks-PCR A. Startery forward wykorzystane w reakcjach multipleks PCR były znakowane na końcu 5' jednym z następujących barwników fluorescencyjnych: 6-FAM, HEX, ROX lub TAMRA (pkt. 3.8), startery reverse nie były znakowane fluorescencyjnie.

Kompatybilność starterów w reakcji multipleks-PCR sprawdzano in silico (pkt. 3.5), a także empirycznie poprzez wykonanie reakcji multipleks-PCR i ocenę specyficzności i wydajności reakcji poprzez analizę elektroforegramów. W przypadku uzyskania niskiej wydajności reakcji dla któregokolwiek z loci zwiększano stężenie starterów dla tego locus w mieszaninie reakcyjnej, a w przypadku braku poprawy wydajności projektowano startery od nowa. Modyfikacji lub zmiany sekwencji starterów dokonano także w przypadkach, gdy w reakcji multipleks-PCR pojawiały się artefakty lub rozdwojone piki elektroforetyczne, wskazujące na niepełną adenylację końcowego produktu reakcji. Takie problemy zaobserwowano dla pięciu loci: D01L569, D02L174, D03L194, D14L276 i D14L785. W tych przypadkach startery reverse zostały wydłużone na końcu 5' o sekwencję GTTTCTT (tzw. PIG-tail), co skutecznie zwiększyło wydajność końcowej adenylacji amplikonów i wyeliminowało problem generowania rozdwojonych pików elektroforetycznych dla tych loci.

Obie reakcje multipleks-PCR zostały zaprojektowane w taki sposób, aby pomiędzy amplikonami poszczególnych loci, wyznakowanych tym samym barwnikiem fluorescencyjnym w tej samej reakcji multipleks-PCR istniały przerwy od około 10 do około 20 pz. Te przerwy zostały zaprojektowane na wypadek istnienia w populacji niezidentyfikowanych rzadkich alleli, krótszych lub dłuższych niż najkrótszy lub najdłuższy allel zidentyfikowany w badaniach przesiewowych przeprowadzonych na pulach DNA. Przerwy były również istotne w kontekście ewentualnych drobnych zmian w sekwencjach starterów, jeśli wymagałyby one np. wydłużenia na końcu 5' o siedmionukleotydową sekwencję PIG-tail GTTTCTT wspomagającą końcową adenylację amplifikowanych fragmentów DNA (pkt. 3.7).

Ze względu na planowaną komercjalizację zestawu, w obu reakcjach multipleks-PCR zarezerwowano miejsce dla jednego z loci systemu CODIS: D1S1656 w multipleksie A i D12S391 w multipleksie B. Włączenie tych dwóch loci, po jednym w każdym multipleksie, miało na celu umożliwienie kontroli obiegu badanych próbek w laboratorium. Dzięki temu w kolejnych badaniach możliwe było porównanie profilu genetycznego badanej osoby w zakresie tych loci z profilem uzyskanym przy użyciu komercyjnych zestawów odczynników, obejmujących loci CODIS. Umożliwiło to weryfikację, czy DNA amplifikowane w różnych reakcjach multipleks-PCR pochodziło od tej samej osoby, co pozwalało wykluczyć ewentualne przypadkowe pomylenie próbek w trakcie procesu laboratoryjnego.

Analiza elektroforegramów reakcji multipleks-PCR z użyciem różnych komercyjnych zestawów odczynników do amplifikacji DNA (Type-it Microsatellite PCR Kit firmy Qiagen oraz HOT FIREPol® MultiPlex Mix firmy Solis Biodyne) wykazała, że zestaw Type-it Microsatellite PCR Kit nie amplifikował wydajnie dwóch loci STR (D05L140 i D02L17) nawet pomimo wielokrotnych prób modyfikacji warunków reakcji zalecanych przez producenta zestawu oraz zwiększenia stężenia starterów do amplifikacji tych loci w mieszaninie reakcyjnej. Dopiero użycie zestawu HOT FIREPol® MultiPlex Mix pozwoliło na uzyskanie wystarczająco efektywnej amplifikacji tych loci (Ryc. 16). Dalsze badania prowadzono z użyciem zestawu odczynników HOT FIREPol® MultiPlex Mix.



Ryc. 16 Elektroforegramy przedstawiające skuteczność amplifikacji loci STR w reakcji multipleks-PCR przy użyciu zestawu HOT FIREPol® MultiPlex Mix (lewa sekcja) oraz Type-it Microsatellite PCR Kit (prawa sekcja). Wzrost wydajności amplifikacji był najbardziej widoczny dla loci D05L140 i D02L174 (zaznaczonych na czerwono), gdzie wartość RFU wzrosła 7,5-krotnie z 160 RFU (zestaw Type-it Microsatellite PCR Kit (Qiagen) do 1200 RFU (zestaw HOT FIREPol® MultiPlex Mix (Solis Biodyne) w przypadku locus D05L140 oraz 4,5-krotnie w przypadku locus D02L174, z 200 RFU do 900 RFU.

Kolejnym etapem optymalizacji parametrów reakcji multipleks-PCR było dostosowanie stężeń starterów w celu zbalansowania intensywności sygnałów fluorescencyjnych amplikonów. Część par starterów pozostała na standardowym poziomie, co oznacza, że dodawano je w stężeniu 0,05 µM. Największe modyfikacje wprowadzono dla starterów amplifikujących loci D10S2325, D14L699, D02L174 i D05L140, w przypadku których stężenia starterów zwiększono sześciokrotnie oraz dla locus D17L432, w przypadku którego stężenie starterów zwiększono siedmiokrotnie do stężenia 0,35 µM. Zmiany stężeń starterów dla wszystkich loci STR amplifikowanych w reakcjach multipleks PCR A i B zostały przedstawione w tabeli nr 8.

Tabela 8. Modyfikacje stężeń starterów dla każdego locus wchodzącego w skład multipleksów A i B. „1x" oznacza stężenie każdego z pary starterów danego locus (forward i reverse) w mieszaninie reakcyjnej równe 0,05 µM. Wartość liczbowa w rubryce stężenie starterów oznacza mnożnik stężenia wyjściowego każdej pary starterów dla wskazanego locus.

Multipleks A			**Multipleks B**		
Locus	**Stężenie starterów**	**Barwnik fluorescencyjny**	**Locus**	**Stężenie starterów**	**Barwnik fluorescencyjny**
D04L885	1x	6-FAM	**D09L159**	3x	6-FAM
D14L953	1,5x	6-FAM	**D02L221**	2,5x	6-FAM
D06L106	1,5x	6-FAM	**D02L142**	3x	6-FAM
D16L554	1,5x	6-FAM	**D8A26**	3x	6-FAM
D01L267	1,5x	6-FAM	**D05L113**	1,5x	6-FAM
D03L115	1x	6-FAM	**D14L276**	3,5x	6-FAM
D2N43	1x	6-FAM	**D17L255**	1x	HEX
D12L794	2x	HEX	**D20L226**	1x	HEX
D02L114	2x	HEX	**D08L110**	5,5x	HEX
D1S1656	4x	HEX	**D16L732**	1,5x	HEX
D05L169	3x	HEX	**D01L217**	2x	HEX
D14L699	6x	HEX	**D03L109**	2x	HEX
D3A57	4x	HEX	**D07L144**	5x	HEX
D8S1132	4x	TAMRA	**D7S3048**	2,5x	TAMRA
D3N61	2x	TAMRA	**D12S391**	2x	TAMRA
D05L140	6x	TAMRA	**D03L194**	4x	TAMRA
D13S742	3x	TAMRA	**D07L101**	4x	TAMRA
D02L174	6x	TAMRA	**D10L126**	3,5x	TAMRA
D17L432	7x	TAMRA	**D3N54**	2x	ROX
Amel	3x	ROX	**D01L215**	1,5x	ROX
D1N16	2x	ROX	**D12L908**	1x	ROX
D15L495	2x	ROX	**D07L147**	2x	ROX
D21L291	3x	ROX	**D12L630**	4x	ROX
D14L785	2x	ROX	**D10S2325**	6x	ROX
D05L207	2x	ROX	**D07L134**	3x	ROX
D01L569	3x	ROX			

Po wprowadzeniu wszystkich powyższych zmian uzyskano finalne warunki dla obu reakcji multipleks-PCR A i B, które umożliwiają jednoczesną analizę 50 loci STR oraz markera płci – amelogeniny. Długości poszczególnych amplikonów w zaprojektowanych multipleksach mieszczą się w zakresach 66-534 pz (tabele nr 9-10).

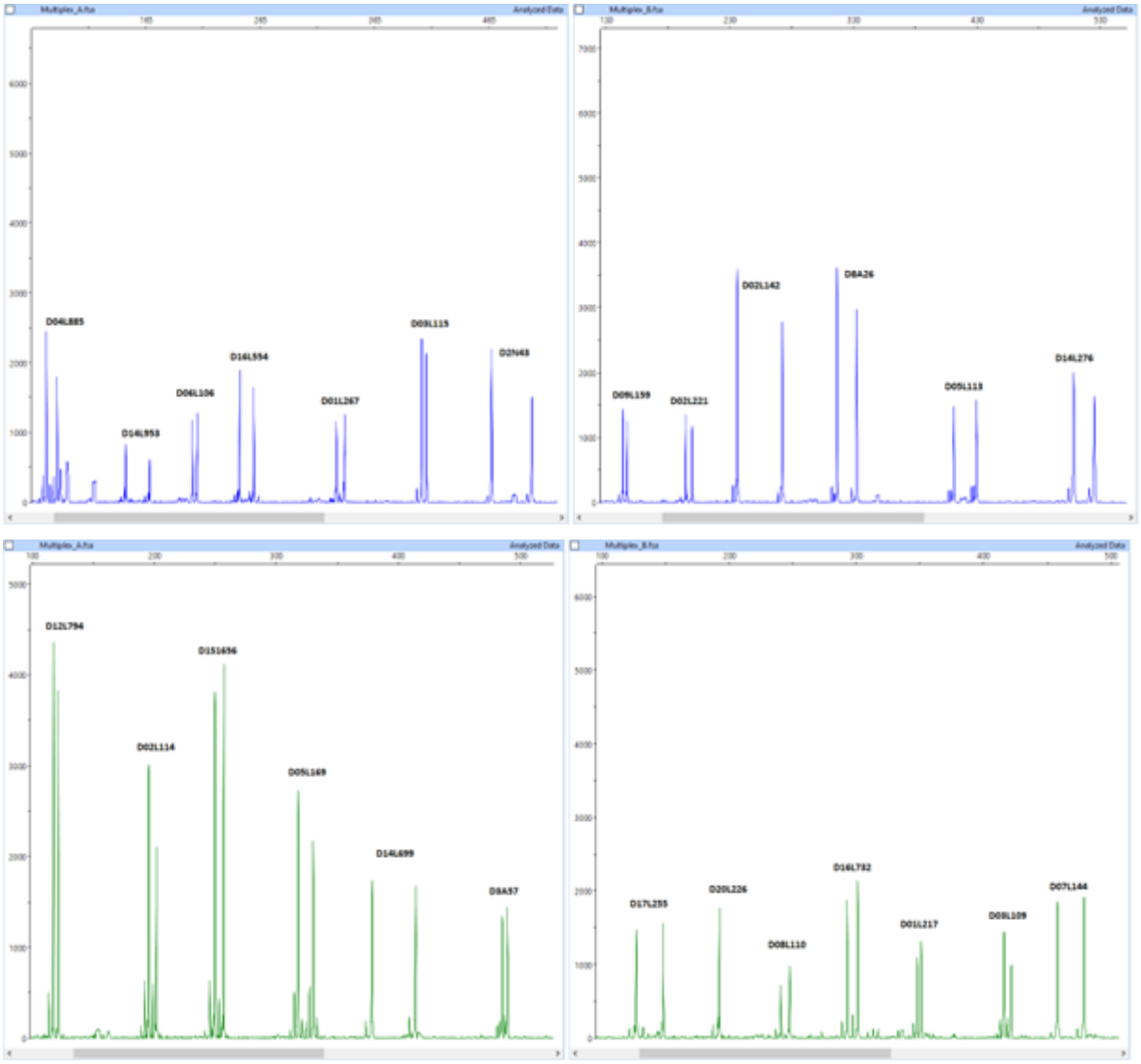
Tabela 9. Długości motywu powtórzonego w locus STR oraz zakres długości amplikonów w reakcji multipleks-PCR A.

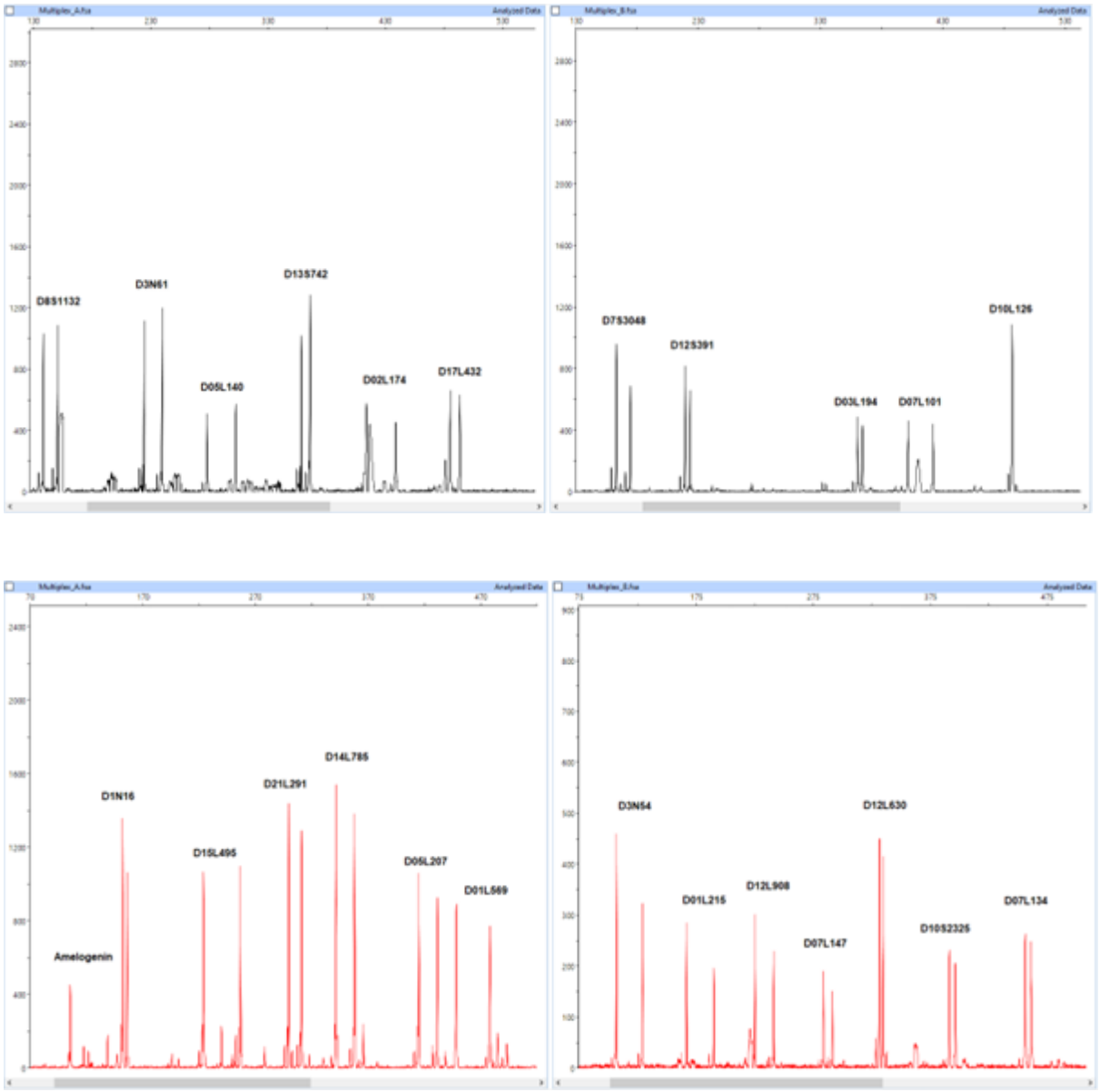
Locus	Długość motywu sekwencji STR	Zakres długości amplikonów (pz)	Kanał detekcji
D04L885	3	66,93 - 90,44	6-FAM
D14L953	4	139,2 - 176,41	6-FAM
D06L106	4	191,51 - 222,13	6-FAM
D16L554	4	242,82 - 299,43	6-FAM
D01L267	4	313,9 - 361,97	6-FAM
D03L115	4	383,56 - 426,6	6-FAM
D2N43	4	440,0 - 487,18	6-FAM
D12L794	4	101,86 - 140,77	HEX
D02L114	3	152,06 - 217,16	HEX
D1S1656	4	217,37 - 264,88	HEX
D05L169	3	308,39 - 345,22	HEX
D14L699	5	383,1 - 454,1	HEX
D3A57	4	464,1 - 505,68	HEX
D8S1132	4	135,83 - 172,02	TAMRA
D3N61	4	204,58 - 255,39	TAMRA
D05L140	4	268,12 – 322,21	TAMRA
D13S742	4	344,02 - 390,09	TAMRA
D2L174	4	407,74 - 448,05	TAMRA
D17L432	4	469,49 - 506,12	TAMRA
Amel	-	105 - 111	ROX
D1N16	4	141,07 - 180,74	ROX
D15L495	4	206,15 - 268,48	ROX
D21L291	4	284,49 - 324,46	ROX
D14L785	4	336,97 - 382,48	ROX
D05L207	4	394,94 - 436,14	ROX
D01L569	4	447,41 - 506,9	ROX

Tabela 10. Długości motywu powtórzonego w locus STR oraz zakres długości amplikonów w reakcji multipleks-PCR B.

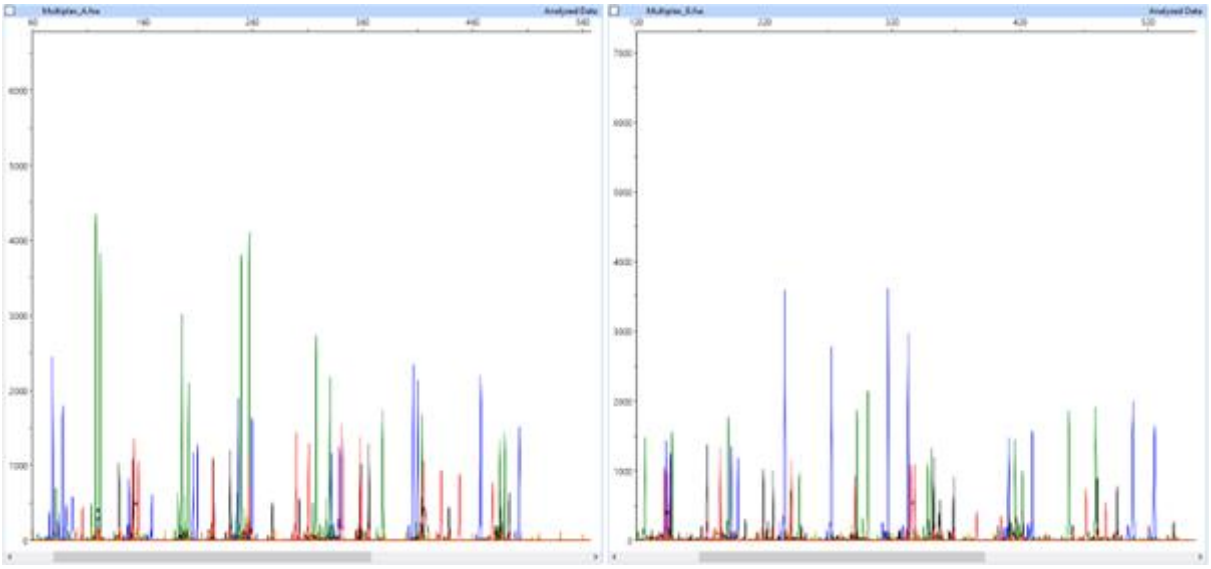
Locus	Długość motywu sekwencji STR	Zakres długości amplikonów w pz	Kanał detekcji
D09L159	4	122,05 - 154.03	6-FAM
D02L221	4	165,43 - 202,61	6-FAM
D02L142	4	228,67 - 288,31	6-FAM
D8A26	4	300,38 - 357,44	6-FAM
D05L113	4	389,29 - 441,77	6-FAM
D14L276	4	462,7 – 533,11	6-FAM
D17L255	4	105,92 - 144,21	HEX
D20L226	5	154,14 - 207,58	HEX
D08L110	4	221,91 - 267,11	HEX
D16L732	4	276,28- 321,65	HEX
D01L217	3	341,55 - 384,75	HEX
D03L109	3	396,61- 439,12	HEX
D07L144	5	447,6 - 515,31	HEX
D7S3048	4	148,18 - 187,91	TAMRA
D12S391	4	204,56 - 256,17	TAMRA
D01L228	3	273,93 - 324,95	TAMRA
D03L194	4	333,74 - 381,06	TAMRA
D07L101	5	410,74 - 464,99	TAMRA
D10L126	3	474,56 - 502,18	TAMRA
D3N54	4	102,92 - 147,43	ROX
D01L215	4	162,12 - 197.52	ROX
D12L908	4	221,07 - 249,2	ROX
D07L147	4	260,19 - 298,85	ROX
D12L630	3	319,1 - 346,82	ROX
D10S2325	5	361,07 - 415,32	ROX
D07L134	5	440,21 - 533,26	ROX

W każdym z czterech kanałów detekcji każdej z obu reakcji multipleks-PCR analizowanych jest od 5 do 7 loci STR (ryc.17).





Ryc. 17 Elektroforegramy dwóch reakcji multipleks-PCR z rozdziałem na kanały detekcji dla amplikonów wyznakowanych 6-FAM (niebieski), HEX (zielony), TAMRA (czarny) oraz ROX (czerwony). Lewe segmenty przedstawiają elektroforegramy multipleksu A, zaś prawe multipleksu B.



Ryc. 18. Elektroforegramy reakcji multipleks-PCR A (lewy segment) i multipleks-PCR B (prawy segment) w czterech kanałach detekcji.

Opracowaną metodę analizy biologicznego pokrewieństwa Kinfinder przetestowano na wszystkich badanych izolatach DNA, pochodzących od niespokrewnionych ze sobą osób z populacji polskiej. Badania populacyjne nie wykazały żadnych nieprawidłowości działaniu metody Kinfinder. W oparciu o te wyniki wyznaczono częstości alleli badanych loci w populacji polskiej, a także obliczono heterozygotyczność loci STR (Tabela 11).

Tabela 11. Porównanie szacowanej heterozygotyczności loci STR w badaniach przesiewowych (Het. Badania przesiewowe) oraz heterozygotyczności oznaczonej na podstawie badań populacyjnych (Het. Badania populacyjne). Różnica pomiędzy wartościami heterozygotyczności uzyskanymi w dwóch różnych badaniach jest wyrażona w procentach (%)

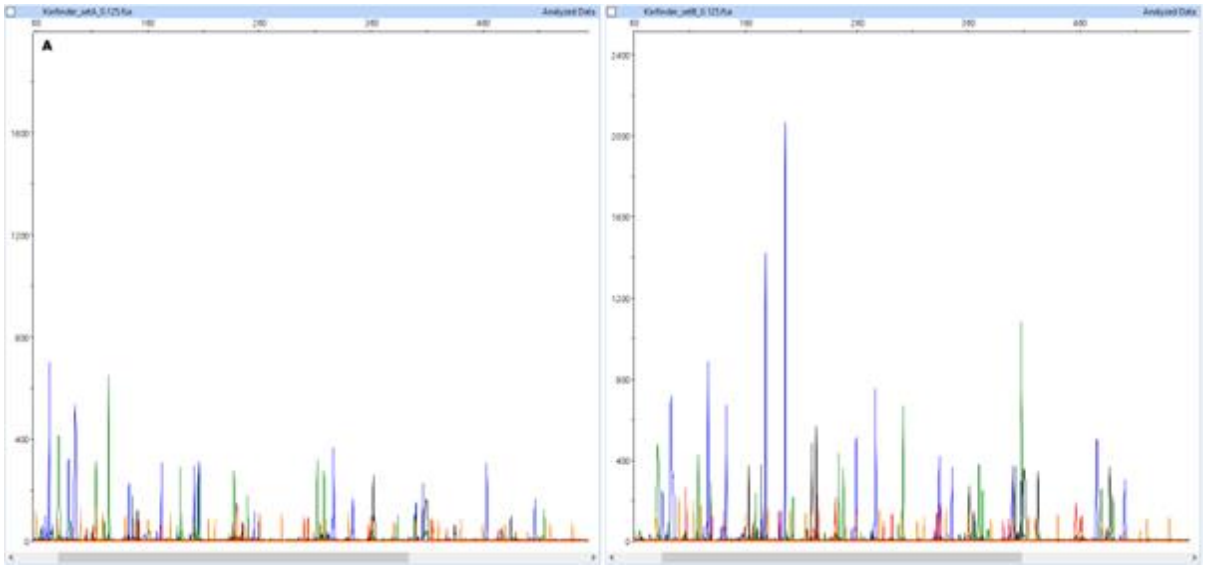
Locus	Różnica	Het. Badania przesiewowe	Het. Badania populacyjne	Locus	Różnica	Het. Badania przesiewowe	Het. Badania populacyjne
D12L908	0,75	0,818	0,8105	**D03L109**	3,22	0,917	0,8848
D07L147	1,88	0,83	0,8112	**D01L217**	0,25	0,888	0,8855
D10L126	1,69	0,843	0,8261	**D07L134**	1,36	0,872	0,8856
D17L255	3,97	0,873	0,8333	**D14L785**	1,73	0,869	0,8863
D16L732	1,73	0,859	0,8401	**D02L221**	2,5	0,863	0,888
D21L291	0,93	0,851	0,8417	**D14L699**	0,98	0,885	0,8948
D04L885	0,2	0,844	0,842	**D02L114**	2,5	0,87	0,895
D17L432	7,7	0,92	0,843	**D05L113**	0,3	0,899	0,896
D12L794	0,81	0,854	0,8459	**D14L276**	0,06	0,898	0,8986

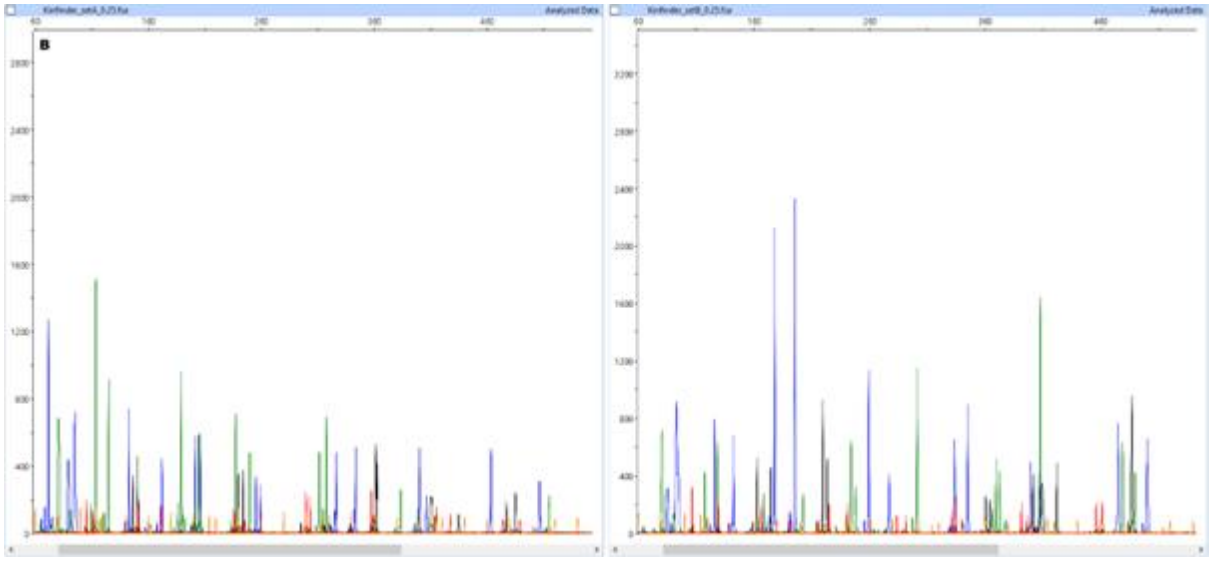
D12L630	0,32	0,855	0,8518	**D01L267**	4,3	0,85	0,8993
D8S1132	2,2	0,833	0,855	**D1N16**	2,8	0,874	0,902
D06L106	3,18	0,89	0,8582	**D03L115**	0,02	0,904	0,9042
D07L101	0,28	0,865	0,8622	**D05L140**	0,59	0,911	0,9051
D10S2325	0,81	0,856	0,8641	**D02L142**	1,17	0,917	0,9053
D05L169	1,19	0,881	0,8691	**D3N54**	3,64	0,875	0,9114
D20L226	0,71	0,863	0,8701	**D16L554**	1,11	0,905	0,9161
D2N43	3,92	0,833	0,8722	**D12S391**	1,9	0,862	0,881
D7S3048	2,03	0,894	0,8737	**D02L174**	0,4	0,922	0,918
D05L207	1,3	0,887	0,8745	**D3A57**	1,1	0,934	0,923
D07L144	1,53	0,891	0,8757	**D3N61**	2,71	0,9	0,9271
D01L215	2,24	0,854	0,8764	**D01L569**	2,67	0,901	0,9277
D14L953	2,85	0,851	0,8795	**D8A26**	2,55	0,909	0,9345
D09L159	1,49	0,866	0,8809	**D15L495**	3,03	0,905	0,9353
D03L194	2,2	0,861	0,883	**D13S742**	2,8	0,911	0,939 0
D08L110	2,62	0,832	0,8582	**D1S1656**	1,52	0,88	0,8952

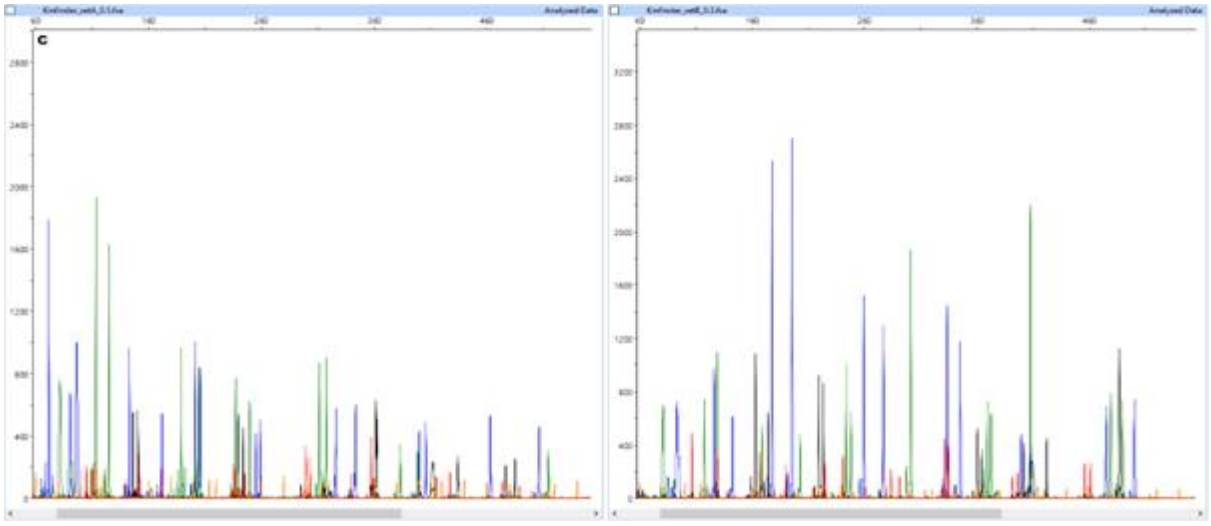
Średnia heterozygotyczność 50 loci STR oznaczona na podstawie badań populacyjnych wyniosła 0,8807 (88,07%) przy średniej hetrozygotyczności loci wyznaczonej na podstawie badań przesiewowych wynoszącej 0,8762 (87,62%). Średnia różnica w heterozygotyczności locus oznaczonej w badaniach populacyjnych i przesiewowych dla 50 badanych loci STR wyniosła 1,875%.

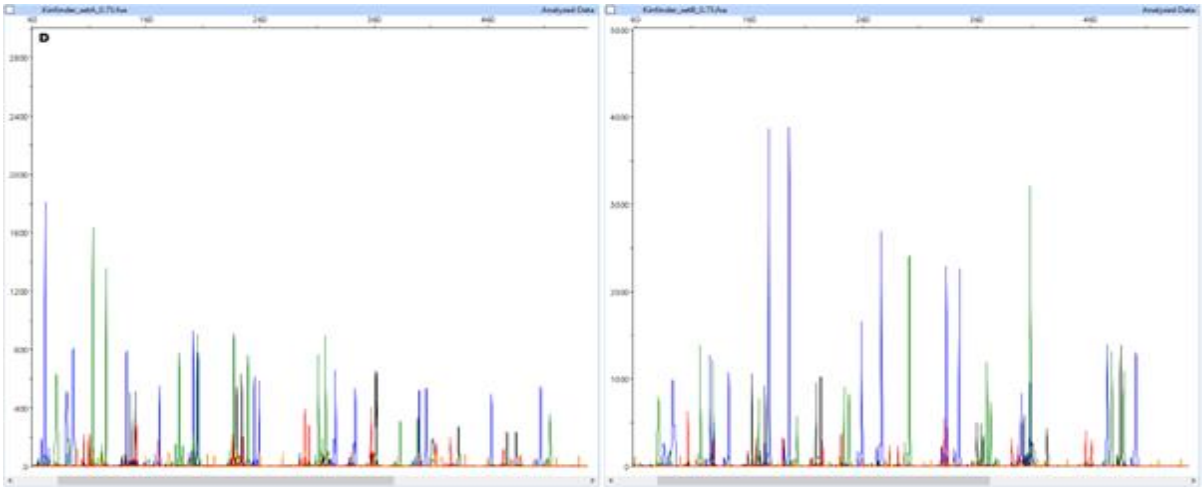
4.5 Analiza czułości metody Kinfinder

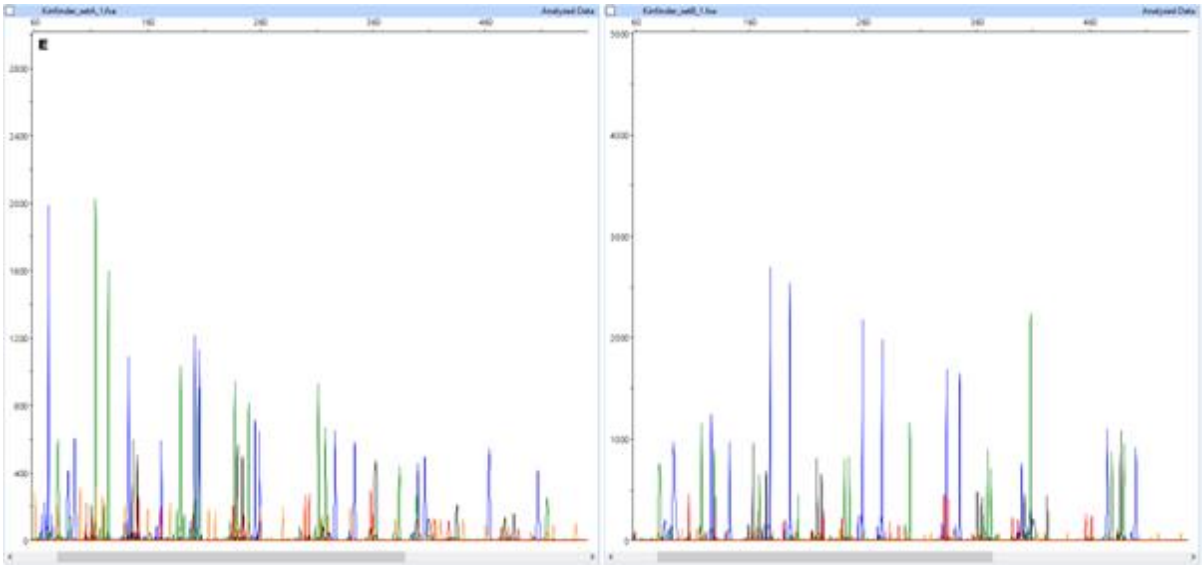
Opracowaną metodę Kinfinder do genotypowania ludzkiego DNA w zakresie 50 wysoce polimorficznych loci STR poddano testom czułości, stosując w reakcjach multipleks-PCR A i B różne ilości DNA człowieka: od 0,125 ng do 2 ng (pkt. 3.10) (Ryc. 19 A-I). Testy wykazały, że pełen profil genetyczny w obu multipleksach uzyskano już przy amplifikacji próbki zawierającej 0,125 ng ludzkiego DNA, co odpowiada ilości DNA znajdującej się w około 20 komórkach somatycznych (ryc. 19).

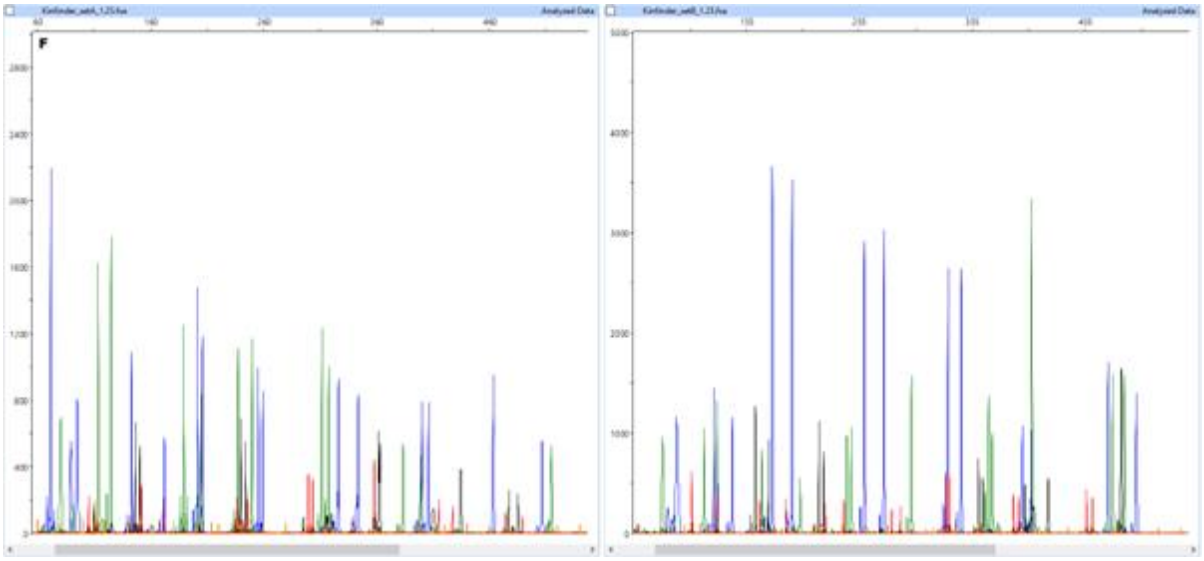

A

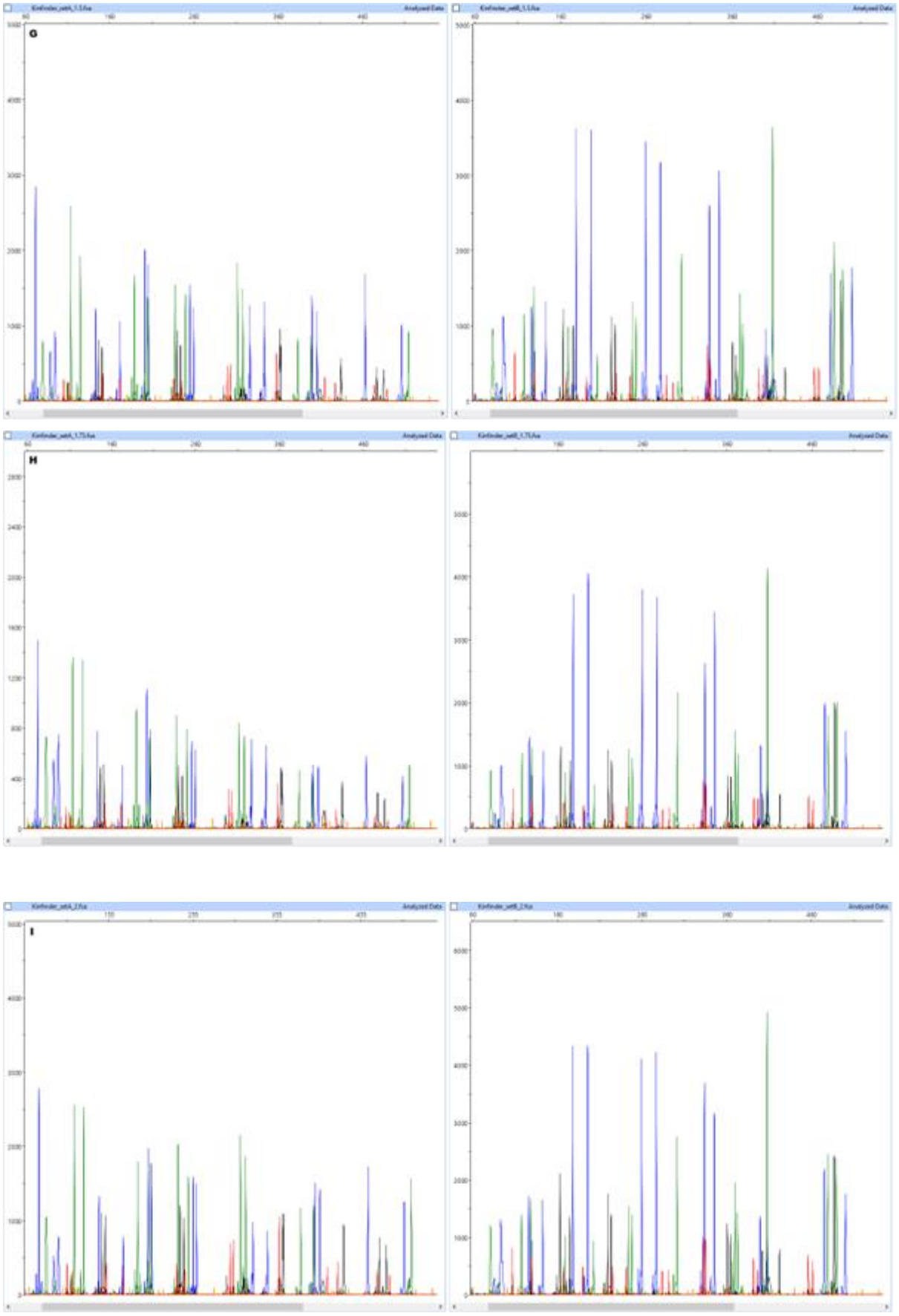

B


C


D


E

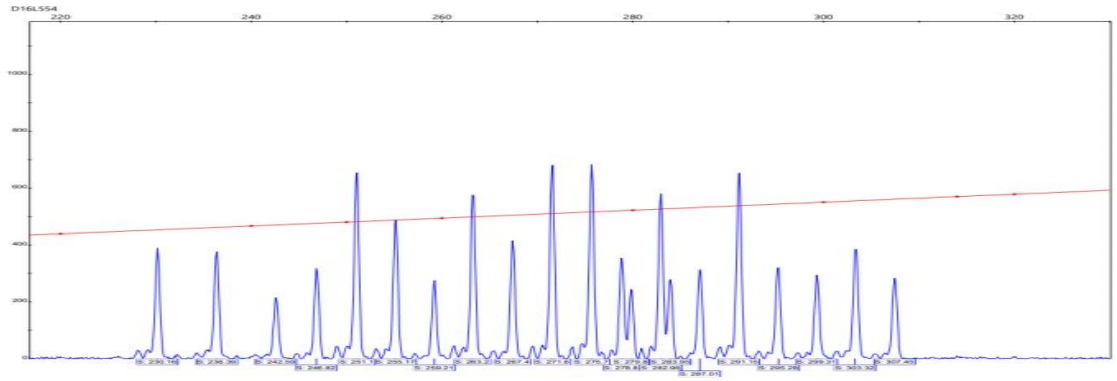

F


G
H
I

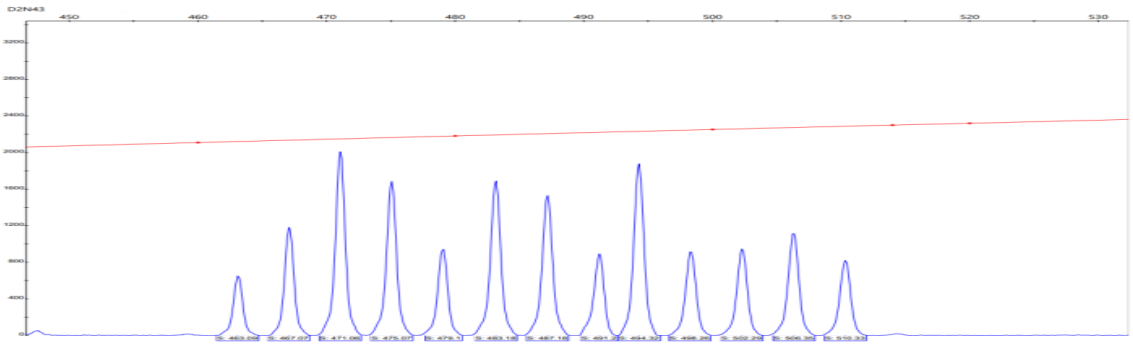
Ryc. 19. Analiza czułości obu reakcji multipleks-PCR (multipleks A - lewy segment, multipleks B - prawy segment). Elektroforegramy przedstawiają wynik amplifikacji DNA dla próbek zawierających izolat DNA Control DNA 007 (Thermofisher) w następującej ilości: A- 0.125ng; B- 0.25ng; C- 0.5ng; D- 0.75ng; E- 1ng; F- 1.25ng; G- 1.5ng; H- 1.75ng; I- 2ng.

4.6 Opracowanie drabiny allelicznej

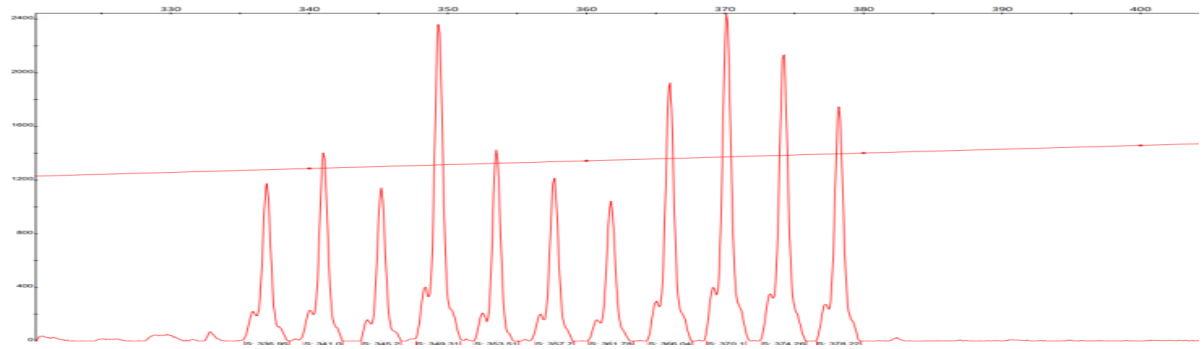
Drabina alleliczna jest elementem zestawu odczynników służącym prawidłowej identyfikacji alleli bez względu na warunki elektroforezy kapilarnej i typu analizatora genetycznego, na którym przeprowadzane są analizy profilowania genetycznego badanych osób. Drabiny alleliczne poszczególnych loci STR w zestawie Kinfinder zostały skonstruowane w sposób opisany w pkt. 3.16. Poprawność skonstruowania drabiny była weryfikowana poprzez analizę elektroforegramów. W szczególności weryfikowano czy w drabinie allelicznej obecne są wszystkie amplifikowane allele, czy wysokość pików elektroforetycznych odpowiadających poszczególnym allelom jest wyrównana oraz czy w drabinie allelicznej nie znajdują się artefakty reakcji PCR mogące utrudnić interpretację elektroforegramów. Drabiny alleliczne wybranych loci STR wraz z opisem przedstawione są na rycinach 20-24.



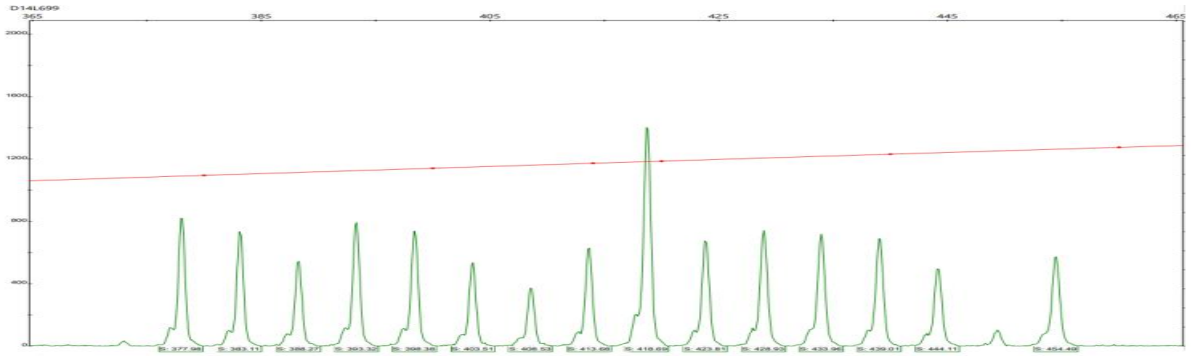
Ryc. 20. Drabina alleliczna locus D16L554. W drabinie allelicznej znajdują się amplikony 21 alleli.



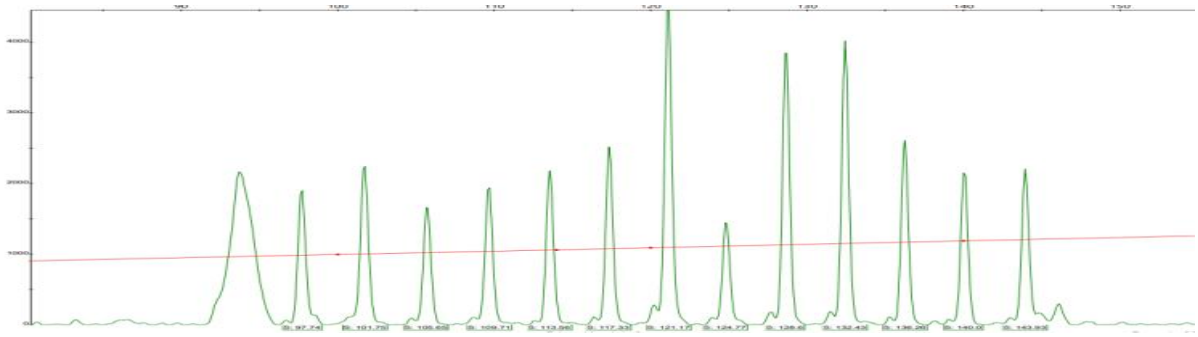
Ryc. 21. Drabina alleliczna locus D2N43. W drabinie allelicznej znajdują się amplikony 13 alleli.



Ryc. 22. Drabina alleliczna locus D14L785. W drabinie allelicznej znajdują się amplikony 11 alleli. Wybrzuszenia po lewej stronie pików elektroforetycznych są efektem niepełnej adenylacji produktów reakcji PCR.



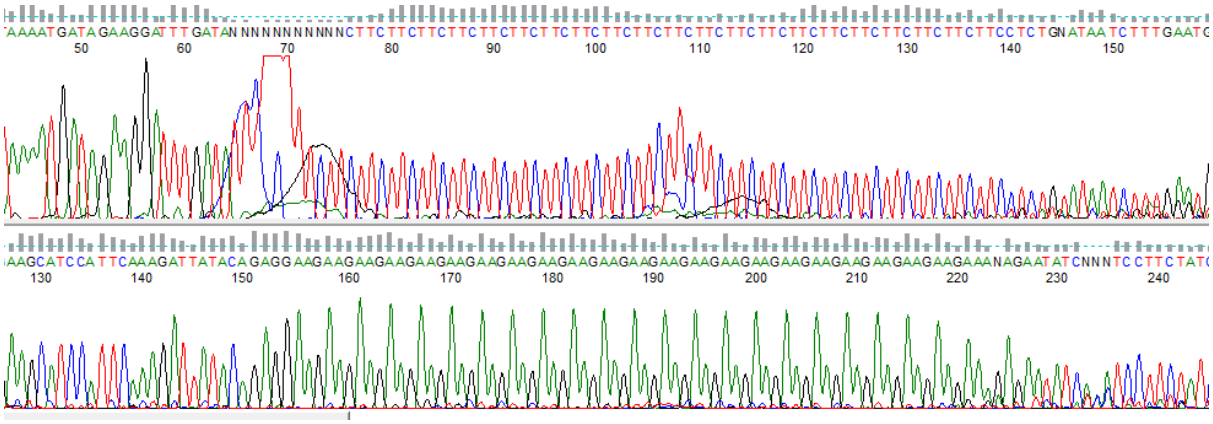
Ryc. 23. Drabina alleliczna locus D14L699. W drabinie allelicznej znajdują się amplikony 15 alleli.



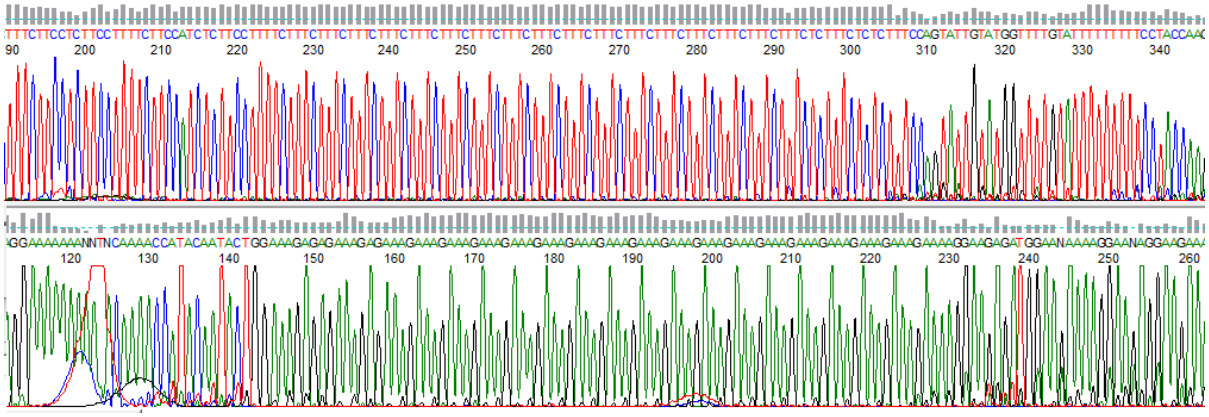
Ryc. 24. Drabina alleliczna dla locus D12L794. W drabinie allelicznej znajdują się amplikony 13 alleli. Po lewej stronie widoczny jest artefakt - pik elektroforetyczny o innej morfologii niż produkty reakcji PCR. Artefakt jest prawdopodobnie efektem niedostatecznego oczyszczenia fluorescencyjnie znakowanego startera forward przez dostawcę startera do reakcji PCR.

4.7 Sekwencje nowo opracowanych loci STR

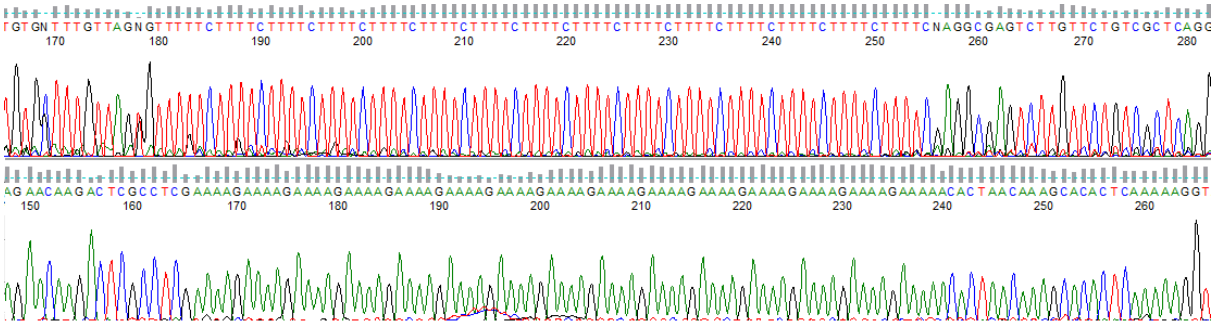
W celu weryfikacji poprawności amplifikacji właściwych loci STR, a także w celu identyfikacji liczby jednostek tandemowo powtórzonych dla amplikonu o określonej długości, produkty PCR sekwencjonowano metodą Sangera. Do sekwencjonowania wybrano amplikony pochodzące od osób homozygotycznych dla analizowanego locus STR. Sekwencjonowanie loci STR przeprowadzono dla obu nici DNA używając do sekwencjonowania nieznakowanych starterów forward i reverse (pkt. 3.17) Na rycinach nr 25-27 przedstawiono przykłady elektroforegramów reakcji sekwencjonowania trzech loci STR: D01L217, D01L569 oraz D07L144, zawierających odpowiednio powtórzenia tri-, tetra- oraz pentanuklotydowe.



Ryc. 25. Elektroforegramy przedstawiające sekwencję allelu locus D01L217 składającą się z 23 powtórzeń tandemowych motywu trinukleotydowego AAG/TTC. Obie reakcje sekwencjonowania wykonano dla tego samego izolatu DNA. Górny elektroforegram przedstawia reakcję sekwencjonowania z użyciem startera forward, dolny z użyciem startera reverse.

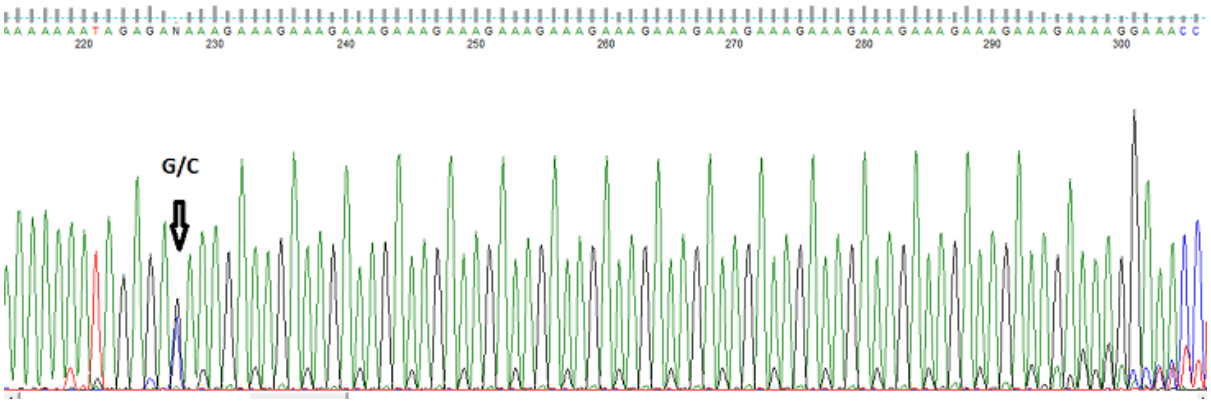


Ryc. 26. Elektroforegramy przedstawiające sekwencję allelu locus D01L569 składającą się z 17 powtórzeń motywu trinukleotydowego AAAG/TTTC. Obie reakcje sekwencjonowania wykonano dla tego samego izolatu DNA. Górny elektroforegram przedstawia reakcję sekwencjonowania z użyciem startera forward, dolny z użyciem startera reverse.

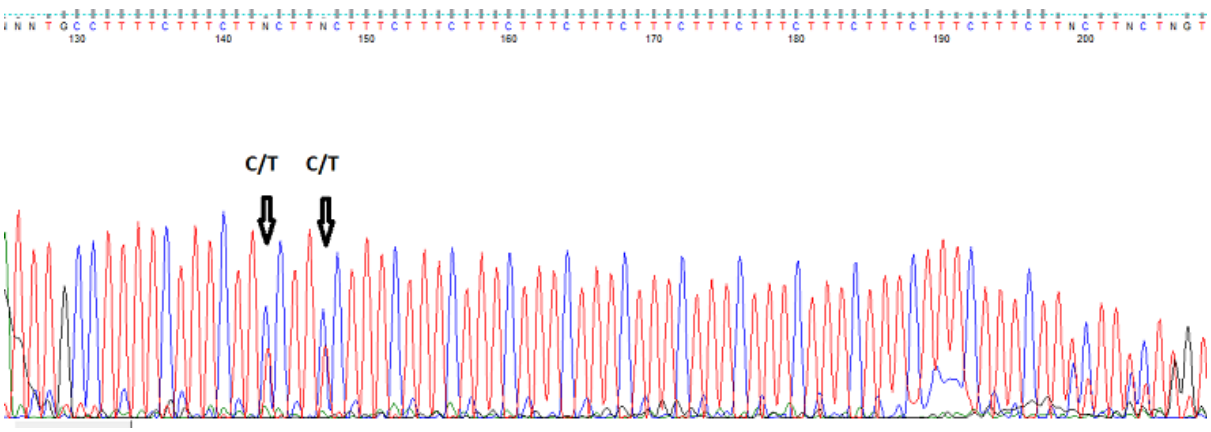


Ryc. 27. Elektroforegramy przedstawiające sekwencję allelu locus D07L144 składającą się z 14 powtórzeń tandemowych motywu pentanukleotydowego AAAAG/TTTTC. Obie reakcje sekwencjonowania wykonano dla tego samego izolatu DNA. Górny elektroforegram przedstawia reakcję sekwencjonowania z użyciem startera forward, dolny z użyciem startera reverse.

W przypadku pięciu loci STR (D12L794, D01L267, D13S742, D03L109 oraz D14L699) w sekwencji wykryto polimorfizm typu SNP (ryc. 28, 29). Analiza sekwencji tych pięciu loci w serwisie Genome Data Viewer NCBI potwierdziła występowanie zidentyfikowanych wcześniej polimorfizmów SNP w sekwencjach genomowych.



Ryc. 28. Elektroforegram przedstawiający sekwencję tandemowo powtórzoną locus D12L794. W obrębie sekwencji tandemowo powtórzonej zidentyfikowano polimorfizm typu SNP G/C.



Ryc. 29. Elektroforegram przedstawiający sekwencję tandemowo powtórzoną locus D01L267. W obrębie sekwencji tandemowo powtórzonej zidentyfikowano zmianę dwóch jednostek motywu powtórzonego TTTC na TTCC.

4.8 Analiza informatywności metody Kinfinder

Biologiczne pokrewieństwo między badanymi osobami analizowano poprzez obliczenie ilorazu wiarygodności (LR, Likelihood Ratio). LR pozwala określić, które z dwóch założeń – pokrewieństwo lub jego brak – jest bardziej prawdopodobne na podstawie wprowadzonych do programu danych w postaci profili genetycznych badanych osób oraz na podstawie częstości występowania alleli analizowanych loci w populacji polskiej.

Obliczenia LR przeprowadzono dla 21 loci systemu Globalfiler, standardowo wykorzystywanego w badaniach pokrewieństwa przez Laboratorium Diagnostyki Molekularnej GenMed oraz dla 69 loci obejmujących system Globalfiler (21 loci) i Kinfinder

(50 loci), z czego dwa loci są wspólne dla obu metod.

Analizy biostatystyczne dotyczące biologicznego pokrewieństwa oraz symulacje tych analiz przeprowadzono za pomocą programu KinBN v1.1.2 (Morimoto i in., 2020), jak opisano w pkt. 3.15. Symulacje wyników pokrewieństwa wykonano dla następujących relacji: rodzic-dziecko (pierwszy stopień pokrewieństwa), pełne rodzeństwo, wujek-siostrzeniec (drugi stopień pokrewieństwa) oraz kuzyn-kuzyn (trzeci stopień pokrewieństwa) (tabele nr 12-15). Celem tych symulacji było określenie informatywności i przewidywanie spodziewanych wyników badań pokrewieństwa w różnych relacjach rodzinnych.

Wszystkie symulacje wykonano dla zestawu 21 loci STR (20 loci CODIS oraz locus SE33) wykorzystywanych w komercyjnym systemie Globalfiler oraz dla zestawu 69 loci STR, w tym 21 loci systemu Globalfiler oraz 48 nowych loci STR systemu Kinfinder. Dla każdej relacji rodzinnej i dla każdej hipotezy (H1 i H2) wykonano po 10 000 symulacji.

Tabela 12. Symulowane wyniki badań biologicznego pokrewieństwa w relacji rodzic-dziecko dla osób spokrewnionych w tej relacji (H1) oraz osób niespokrewnionych (H2) otrzymywane dla metody analizy 21 loci STR (Globalfiler) i 69 loci STR (Globalfiler + Kinfinder). "min." i "maks." oznaczają minimalny i maksymalny wynik LR osiągnięty w 10 tysiącach symulacji. Progi wyników 5%, 25%, 75%, 95% wskazują najwyższy wynik LR uzyskany we wskazanej grupie procentowej najniższych wyników LR. Mediana jest wartością środkową otrzymanych wyników.

Analiza biologicznego pokrewieństwa **relacji rodzic - dziecko** vs osoby niespokrewnione					
Globalfiler (21 loci STR)			Globalfiler + Kinfinder (69 loci STR)		
Próg wyników	Stosunek prawdopodo-bieństw LR (H1)	Stosunek prawdopodo-bieństw LR (H2)	Próg wyników	Stosunek prawdopodo-bieństw LR (H1)	Stosunek prawdopodo-bieństw LR (H2)
min.	431	0	min.	$1{,}91 \times 10^{24}$	0
5%	$1{,}52 \times 10^{5}$	0	5,00%	$2{,}71 \times 10^{26}$	0
25%	$2{,}79 \times 10^{6}$	0	25,00%	$1{,}61 \times 10^{29}$	0
mediana	$2{,}90 \times 10^{7}$	0	mediana	$7{,}64 \times 10^{30}$	0
75%	$2{,}55 \times 10^{8}$	0	75,00%	$3{,}93 \times 10^{32}$	0
95%	$1{,}01 \times 10^{10}$	$1{,}17 \times 10^{-20}$	95,00%	$1{,}39 \times 10^{35}$	0
maks.	$1{,}48 \times 10^{12}$	$2{,}78 \times 10^{-5}$	maks.	$1{,}55 \times 10^{40}$	0

Tabela 13. Symulowane wyniki badań biologicznego pokrewieństwa w relacji pełne rodzeństwo dla osób spokrewnionych w tej relacji (H1) oraz osób niespokrewnionych (H2) otrzymywane dla metody

analizy 21 loci STR (Globalfiler) i 69 loci STR (Globalfiler + Kinfinder). “min.” i “maks.” oznaczają minimalny i maksymalny wynik LR osiągnięty w 10 tysiącach symulacji. Progi wyników 5%, 25%, 75%, 95% wskazują najwyższy wynik LR uzyskany we wskazanej grupie procentowej najniższych wyników LR. Mediana jest wartością środkową otrzymanych wyników.

Analiza biologicznego pokrewieństwa w relacji pełne rodzeństwo vs osoby niespokrewnione					
Globalfiler (21 loci STR)			Globalfiler + nowo opracowana metoda (69 loci STR		
Próg wyników	Stosunek prawdopodo-bieństw LR (H1)	Stosunek prawdopodo-bieństw LR (H2)	Próg wyników	Stosunek prawdopodo-bieństw LR (H1)	Stosunek prawdopodo-bieństw LR (H2)
min.	$4{,}35\text{x}10^{-3}$	$1{,}49\text{x}10^{-10}$	min.	$1{,}37\text{x}10^{8}$	0
5,00%	74,6	$2{,}75\text{x}10^{-8}$	5,00%	$4{,}8\text{x}10^{17}$	0
25,00%	$2{,}32\text{x}10^{4}$	$8{,}79\text{x}10^{-7}$	25,00%	$1{,}93\text{x}10^{22}$	$6{,}61\text{x}10^{-22}$
mediana	$2{,}01\text{x}10^{6}$	$1{,}24\text{x}10^{-5}$	mediana	$8{,}96\text{x}10^{25}$	$2{,}50\text{x}10^{-20}$
75,00%	$1{,}37\text{x}10^{8}$	$1{,}85\text{x}10^{-4}$	75,00%	$1{,}64\text{x}10^{28}$	$2{,}08\text{x}10^{-17}$
95,00%	$7{,}42\text{x}10^{10}$	$1{,}30\text{x}10^{-2}$	95,00%	$1{,}33\text{x}10^{33}$	$5{,}79\text{x}10^{-14}$
maks.	$8{,}64\text{x}10^{14}$	$1{,}47\text{x}10^{3}$	maks.	$4{,}89\text{x}10^{40}$	$1{,}10\text{x}10^{-11}$

Tabela 14. Symulowane wyniki badań biologicznego pokrewieństwa w relacji pokrewieństwa drugiego stopnia dla osób spokrewnionych w tej relacji (H1) oraz osób niespokrewnionych (H2) otrzymywane dla metody analizy 21 loci STR (Globalfiler) i 69 loci STR (Globalfiler + Kinfinder). “min.” i “maks.” oznaczają minimalny i maksymalny wynik LR osiągnięty w 10 tysiącach symulacji. Progi wyników 5%, 25%, 75%, 95% wskazują najwyższy wynik LR uzyskany we wskazanej grupie procentowej najniższych wyników LR. Mediana jest wartością środkową otrzymanych wyników.

Analiza biologicznego pokrewieństwa **relacji drugiego stopnia** vs osoby niespokrewnione					
Globalfiler (21 loci STR)			Globalfiler + nowo opracowana metoda (69 loci STR)		
Próg wyników	Stosunek prawdopodo-bieństw LR (H1)	Stosunek prawdopodo-bieństw LR (H2)	Próg wyników	Stosunek prawdopodo-bieństw LR (H1)	Stosunek prawdopodo-bieństw LR (H2)
min.	$1{,}05\text{x}10^{-2}$	$4{,}07\text{x}10^{-5}$	min.	1,1	$3{,}37\text{x}10^{-13}$
5,00%	0,6	$6{,}28\text{x}10^{-4}$	5,00%	536	$1{,}34\text{x}10^{-10}$
25,00%	7,35	$6{,}23\text{x}10^{-3}$	25,00%	$3{,}30\text{x}10^{5}$	$2{,}14\text{x}10^{-8}$
mediana	71,4	$3{,}18\text{x}10^{-2}$	mediana	$1{,}13\text{x}10^{7}$	$1.9\text{x}10^{-7}$
75,00%	391	0,2	75,00%	$2{,}16\text{x}10^{9}$	$6{,}51\text{x}10^{-6}$
95,00%	$9{,}02\text{x}10^{3}$	2,75	95,00%	$9{,}91\text{x}10^{11}$	$2{,}02\text{x}10^{-3}$
maks.	$4{,}88\text{x}10^{5}$	124	maks.	$1{,}52\text{x}10^{14}$	11,4

Tabela 15. Symulowane wyniki badań biologicznego pokrewieństwa w relacji pokrewieństwa trzeciego

stopnia dla osób spokrewnionych w tej relacji (H1) oraz osób niespokrewnionych (H2) otrzymywane dla metody analizy 21 loci STR (Globalfiler) i 69 loci STR (Globalfiler + Kinfinder). “min.” i “maks.” oznaczają minimalny i maksymalny wynik LR osiągnięty w 10 tysiącach symulacji. Progi wyników 5%, 25%, 75%, 95% wskazują najwyższy wynik LR uzyskany we wskazanej grupie procentowej najniższych wyników LR. Mediana jest wartością środkową otrzymanych wyników.

Analiza biologicznego pokrewieństwa relacji trzeciego stopnia vs osoby niespokrewnione					
Globalfiler (21 loci STR)			Globalfiler + nowo opracowana metoda (69 loci STR)		
Próg wyników	Stosunek prawdopodo-bieństw LR (H1)	Stosunek prawdopodo-bieństw (H2)	Próg wyników	Stosunek prawdopodo-bieństw LR (H1)	Stosunek prawdopodo-bieństw LR (H2)
min.	$4{,}55\times10^{-2}$	$1{,}60\times10^{-2}$	min.	$4{,}05\times10^{-2}$	$1{,}05\times10^{-4}$
5,00%	0,29	$5{,}20\times10^{-2}$	5,00%	0,46	$6{,}35\times10^{-4}$
25,00%	1,03	0,16	25,00%	12,4	$4{,}48\times10^{-3}$
mediana	2,93	0,34	mediana	158	$2{,}2\times10^{-2}$
75,00%	9,4	0,78	75,00%	$1{,}11\times10^{3}$	0,15
95,00%	61,5	3,2	95,00%	$2{,}85\times10^{4}$	1,93
maks.	$8{,}28\times10^{3}$	94,8	maks.	$2{,}72\times10^{5}$	33,1

4.9 Wdrożenie metody Kinfinder do bieżącej pracy laboratorium genetycznego

Nowo opracowana metoda analizy 50 loci STR w dwóch reakcjach multipleks-PCR została nazwana metodą Kinfinder i została wdrożona do rutynowej pracy Laboratorium Diagnostyki Molekularnej GenMed. Do dnia 30.09.2024 r. metoda Kinfinder została wykorzystana w ponad trzydziestu badaniach biologicznego pokrewieństwa realizowanych zarówno na zlecenie sądów powszechnych jak i klientów indywidualnych. Zdecydowana większość badań dotyczyła analizy biologicznego pokrewieństwa w relacji drugiego stopnia (25% wspólnego autosomalnego DNA u badanych osób, np. wuj-bratanek, dziadek - wnuk, przyrodnia siostra, przyrodni brat), ale wykorzystano tę metodę również w bardziej skomplikowanych i nietypowych sprawach, z których trzy wybrane sprawy opisane są poniżej.

4.9.1 Sprawa nr 1 - disomia jednorodzicielska

Laboratorium Diagnostyki Molekularnej GenMed otrzymało zlecenie przeprowadzenia

standardowego testu ojcostwa na potrzeby sprawy sądowej dotyczącej ustalenia ojcostwa. Otrzymano wymazy z policzków od trzech osób zaangażowanych w badanie (dziecko, matka, domniemany ojciec) i przeprowadzono standardową analizę 21 autosomalnych loci STR przy użyciu zestawu GlobalFiler. Analiza profili genetycznych badanych osób ujawniła trzy niezgodności genetyczne pomiędzy badanym mężczyzną a dzieckiem, które stanowiły podstawę do wykluczenia biologicznego ojcostwa mężczyzny względem badanego dziecka. Bardziej szczegółowa analiza profili genetycznych wykazała, że te wszystkie trzy niezgodności wystąpiły w układach TPOX, D2S441 i D2S1338, które znajdują się na chromosomie 2.

Te trzy niezgodności mogły być spowodowane albo brakiem biologicznego pokrewieństwa między domniemanym ojcem a dzieckiem, albo rzadkim zjawiskiem genetycznym znanym jako disomia jednorodzicielska (ang. uniparental disomy), w tym przypadku dotycząca chromosomu 2. Disomia jednorodzicielska to zjawisko, w którym para homologicznych chromosomów pochodzi od jednego rodzica. Aby dokładniej zbadać i rozwiązać tę sprawę, przeprowadzono szerszą analizę genetyczną angażując do badań zespoły naukowców z Instytutu Ekspertyz Sądowych z Krakowa, Zakładu Medycyny Sądowej Pomorskiego Uniwersytetu Medycznego ze Szczecina oraz Uniwersytetu Jagiellońskiego z Krakowa. Analiza obejmowała łącznie 53 loci STR, w tym sześć nowo scharakteryzowanych w niniejszej pracy loci STR chromosomu 2, a także 32 markery insercji-delecji i 94 polimorfizmy pojedynczego nukleotydu (SNP).

Badanie wykazało, że wszystkie obserwowane niezgodności w markerach DNA między domniemanym ojcem a dzieckiem były ograniczone do chromosomu 2. Ostatecznie doprowadziło to do potwierdzenia biologicznego ojcostwa, ujawniając jednocześnie obecność disomii jednorodzicielskiej matczynej. Wyniki tego badania zostały opublikowane w czasopiśmie naukowym „Genes” (Doniec i in., 2021).

4.9.2 Sprawa nr 2 - analiza pokrewieństwa w sprawie spadkowej

Laboratorium Diagnostyki Molekularnej GenMed otrzymało od jednego z wielkopolskich sądów rejonowych postanowienie o zasięgnięciu opinii biegło-sądowej w sprawie ustalenia biologicznego pokrewieństwa dwóch osób: osoby A (mężczyzna) i osoby B (kobieta), która była uważana za jego ciotkę. Celem było ustalenie ojcostwa zmarłego mężczyzny (osoba C) wobec mężczyzny A w sposób pośredni poprzez analizę DNA jego żyjącej siostry (osoba B) i ustalenie czy badane osoby A i B są spokrewnione w relacji bratanek-ciotka. Początkowo przeprowadzono standardowe badanie pokrewieństwa poprzez analizę 21 loci STR przy zastosowaniu komercyjnego zestawu GlobalFiler. Analiza wykazała wynik nierozstrzygający - LR=1,2.

Aby uzyskać bardziej jednoznaczne wyniki badań wykorzystano metodę Kinfinder i tym samym analiza została rozszerzona do 68 loci STR. Rozszerzona analiza wykazała prawdopodobieństwo pokrewieństwa LR=0,79 dla relacji ciotka-bratanek co również jest wynikiem nierozstrzygającym. Jednocześnie analiza porównawcza profili genetycznych badanych osób A i C wykazała wysokie prawdopodobieństwo biologicznego pokrewieństwa (LR=207,1) dla relacji pokrewieństwa trzeciego stopnia. Wyniki tego badania mogły wskazywać, że zmarły mężczyzna i badana kobieta nie byli pełnym rodzeństwem, lecz rodzeństwem przyrodnim (mieli wspólnego jednego rodzica). Ze względu na charakter relacji rodzinnej nie mogły zostać wykorzystane analizy loci STR chromosomów X i Y.

Realizując postanowienie sądu należało kierować się przepisami polskiego prawa, które w postępowaniu cywilnym nakazują biegłemu podjęcie czynności jedynie w zakresie określonych w postanowieniu sądu i nieprzekazywanie w sporządzonej opinii żadnych innych informacji odnośnie przebiegu realizacji ekspertyzy, o których biegły powziął informacje w trakcie realizacji opinii, a o które nie pytały strony postępowania. Mając na uwadze te okoliczności przekazano sędziemu prowadzącemu sprawę wątpliwości co do wyników analizy DNA i wydając opinię przekazano, że wynik jest niejednoznaczny i jednocześnie zasugerowano potrzebę pozyskania materiału biologicznego od zmarłego mężczyzny (osoba B) na drodze ekshumacji zwłok. Uzyskanie profilu genetycznego zmarłego pozwoliłoby na uzyskanie jednoznacznego wyniku świadczącego o pokrewieństwa badanych osób (A i B) lub jego braku bez konieczności badania DNA żyjącej siostry zmarłego. Sąd przychylił się do sugestii i w

kolejnym postanowieniu zarządził wykonanie ekshumacji zwłok i pobranie materiału biologicznego od zmarłego celem ustalenia biologicznego ojcostwa względem osoby A. Czynności związane z ekshumacja zwłok zostały wykonane i z materiału biologicznego w postaci fragmentów skóry pochodzącej od zmarłego wyizolowano DNA i porównano profile genetyczne obu osób (A i B) w zakresie 69 loci STR, czyli z użyciem zestawu komercyjnego Globalfiler rozszerzonego o loci w zestawie Kinfinder.

Analiza porównawcza profili genetycznych w rozszerzonym zakresie wykazała stosunek prawdopodobieństw LR wynoszący $1{,}38x10^{31}$ (1,38 razy dziesięć trzydziestej pierwszej potęgi), co ostatecznie potwierdziło ojcostwo zmarłego mężczyzny wobec osoby A.

Dodatkowo, już wyłącznie dla potrzeb sprawdzenia przydatności metody Kinfinder w badaniach biologicznego pokrewieństwa przeprowadzono analizę porównawczą profili genetycznych zmarłego mężczyzny (osoba B) i badanej kobiety (osoby C), biorąc pod uwagę dwie hipotezy: pełne rodzeństwo i rodzeństwo przyrodnie. Analiza biostatystyczna przeprowadzona za pomocą programu KinBN potwierdziła biologiczne pokrewieństwo (LR = $4{,}34x10^{6}$) dla relacji rodzeństwa przyrodniego. Z kolei analiza porównawcza dla relacji pełnego rodzeństwa wykazała wynik LR = $1{,}66x10^{-7}$, co ostatecznie potwierdziło wcześniejsze przypuszczenie, że zmarły mężczyzna i badana kobieta byli rodzeństwem przyrodnim (mieli jednego wspólnego rodzica).

4.9.3 Sprawa nr 3 - rzadka mutacja w badaniu ojcostwa

Laboratorium Diagnostyki Molekularnej GenMed otrzymało od klienta indywidualnego standardowe zlecenie wykonania badania biologicznego ojcostwa. Klient przesłał do laboratorium wymazy z policzka pobrane od siebie oraz swojego domniemanego dziecka, z których następnie wyizolowano DNA i oznaczono profile genetyczne badanych osób w standardowym zakresie 21 loci STR system Globalfiler. Analiza wykazała zgodność genetyczną w zakresie 20 loci STR i jedną niezgodność w locus SE33. Niezgodność w zakresie pojedynczego locus jest najczęściej spowodowana mutacją locus STR i zwykle nie stanowi podstawy do wykluczenia ojcostwa, a jedynie obniża końcowy wynik prawdopodobieństwa pokrewieństwa, jednakże w tym przypadku charakter potencjalnej mutacji obniżył łączny

wynik parametru szansy ojcostwa (PI) do wartości 0,04 co wskazywało na brak pokrewieństwa pomiędzy badanymi osobami. Profil genetyczny w zakresie locus SE33 dla ojca został oznaczony jako układ homozygotyczny 33,33, zaś dla dziecka również jako homozygotyczny, ale z allelami 25.2,25.2. Częstość mutacji ojcowskiej dla tego locus wyznaczona na podstawie 51610 analiz pokrewieństwa i wskazana w raporcie AABB (American Association of Blood Banks) z 2004 r. wynosi 0,62% (www.aabb.org/sa/facilities/Pages/relationshipreports.aspx). Parametr ten dotyczy jednak mutacji prowadzących do skrócenia lub wydłużenia dziedziczonego allelu o jeden motyw powtórzenia tandemowego. Mutacje prowadzące do utraty lub zyskania dwóch lub więcej jednostek motywu tandemowo powtórzonego są dużo rzadsze. W opisywanym przypadku mogło dojść do utraty dużego fragmentu sekwencji tandemowo powtórzonej zawierającego ponad 7 powtórzeń tetranukleotydowych sekwencji co jest zjawiskiem ekstremalnie rzadkim. Model matematyczny programu DNAStat wersja 2.1 (Berent, 2010), który jest rutynowo stosowany w Laboratorium Diagnostyki Molekularnej GenMed zakładał, że taka mutacja cechuje się częstością niższą niż 1/1 000 000, co obniżyło końcowy wynik analizy 21 loci STR do LR=0,04 i wskazywało na to, że bardziej prawdopodobny jest brak pokrewieństwa pomiędzy badanymi osobami niż ich pokrewieństwo.

Z uwagi na fakt, że zjawisko mutacji o takim charakterze występuje ekstremalnie rzadko, wzięto pod uwagę okoliczność, że otrzymany wynik analizy może nie być efektem zjawiska mutacji, a rezultatem wystąpienia zjawiska odziedziczenia przez badane dziecku allelu zerowego (ang. null allele). Allel zerowy to allel, który jest obecny w genomie, ale nie ulega amplifikacji w reakcji PCR i tym samym nie jest obserwowany w elektroforegramie (Kline i in., 2011). Najczęstszą przyczyną zaistnienia takiego zjawiska jest mutacja punktowa w miejscu wiązania startera, która może powodować brak amplifikacji tego allelu. Jeśli dana osoba jest heterozygotyczna i ma mutację w miejscu wiązania startera dla jednego z alleli, wówczas w badaniu zostanie zgenotypowana jako homozygota. W opisywanym przypadku zarówno domniemany ojciec jak i syn były osobami homozygotycznymi w obrębie locus SE33. Biorąc pod uwagę, że locus SE33 jest silnie polimorficzny i jego homozygotyczność w populacji polskiej wynosi ok. 6,3%, obecność u dwóch badanych niespokrewnionych osób układu homozygotycznego (o ile hipoteza o braku pokrewieństwa pomiędzy badanymi osobami byłaby prawdziwa) jest mało prawdopodobna (0,4%). Ta okoliczność uprawdopodobniała hipotezę, że brak wspólnego allelu u badanego mężczyzny i dziecka w obrębie locus SE33

może być wynikiem odziedziczenia allelu zerowego, a nie mutacji.

Jednym z rozwiązań mogącym rozstrzygnąć kwestię domniemanego pokrewieństwa pomiędzy badanymi osobami mogłaby być analiza loci STR chromosomu Y, jednakże nawet uzyskanie pełnej zgodności haplotypu Y u badanych mężczyzn nie uprawniałoby do stwierdzenia, że badany mężczyzna jest ojcem dziecka z prawdopodobieństwem graniczącym z pewnością. Zwykle profil genetyczny chromosomu Y jest wspólny dla wszystkich mężczyzn linii ojcowskiej wobec czego nie jest możliwe rozróżnienia, który z mężczyzn w linii ojcowskiej jest ojcem dziecka.

Mając na uwadze powyższe okoliczności zdecydowano się wykorzystać metodę Kinfinder do analizy dodatkowych autosomalnych loci STR. Wynik badań ojcostwa uzyskany z uwzględnieniem prawdopodobieństwa rzadkiej mutacji wyniósł LR = $19x10^{18}$ co jednoznacznie potwierdziło ojcostwo badanego mężczyzny.

5. Dyskusja

5.1 Kryteria wyboru loci STR do badań pokrewieństwa

Pierwszym etapem prac realizowanym w ramach doktoratu wdrożeniowego był wybór nowych loci STR do wykorzystania w metodzie Kinfinder. W niniejszej pracy wykorzystano dane dotyczące zmienności sekwencji mikrosatelitarnych w ludzkim genomie, pochodzące z projektu „1000 Genomes" zagregowane w bazach danych STRCatalog oraz WebSTR. Informacje zebrane w ramach tego projektu oferują ogromny potencjał dla badań nad genetyką populacyjną (Okazaki i in. 2020). Na etapie badań obejmujących selekcję najbardziej polimorficznych loci STR genomu ludzkiego zidentyfikowano kilka problemów, które mogą stanowić przeszkody w poszukiwaniu nowych wysoce polimorficznych loci STR w bazie danych projektu 1000 Genomes. Jednym z nich było ograniczenie technologiczne związane z dostępnymi w bazie danymi, które były pozyskiwane metodą sekwencjonowania następnej generacji (NGS) opartą na sekwencjonowaniu krótkich fragmentów DNA (100 pz). Metoda ta ogranicza możliwość identyfikacji potencjalnie polimorficznych, ale jednocześnie długich loci STR (Tang i in., 2017; Frontanilla i in., 2022). Powodem takiego stanu rzeczy jest wybór metody sekwencjonowania i charakter sekwencji tandemowo powtórzonych. Żeby móc poprawnie zsekwencjonować locus STR koniecznym jest otrzymanie odczytów zawierających pełną sekwencję tandemowo powtórzoną wraz z sekwencjami flankującymi sekwencjonowany locus STR (Rajan-Babu i in., 2021). W przypadkach, w których uda się zsekwencjonować w pojedynczych odczytach jedynie sekwencję flankującą oraz część sekwencji tandemowo powtórzonej nie jest możliwe określenie, jak długi fragment sekwencji tandemowo powtórzonej został zamplifikowany i zsekwencjonowany. Zatem metoda sekwencjonowania krótkich odczytów umożliwia jedynie identyfikację alleli locus STR, w których długość fragmentu zawierającego powtórzenia tandemowe jest krótsza niż 100 pz.

W praktyce nawet dla krótszych loci STR, zawierających się w przedziale długości 50-80 pz, przy głębokości sekwencjonowania 30x zastosowanej w projekcie 1000 Genomes, okazało się problematyczne uzyskanie odczytu zawierającego pełną sekwencję obu alleli locus STR.

Baza STRCatalog agregująca dane uzyskane w Projekcie "1000 Genomes" dotyczące loci STR

zawiera sekwencje 1092 genomów osób z populacji światowej (1000 Genomes Project Consortium, 2012), jednakże z powodu wykorzystania technologii sekwencjonowania krótkich odczytów dane dotyczące długości loci STR badanych osób są niepełne (Gymrek i in., 2016; Read i in., 2023). Dla większości loci STR spełniających kryteria wskazane w pkt. 2.1 niniejszej pracy bazy STRCatalog oraz WebSTR zawierały jedynie dane dotyczące ok. 100-200 alleli i rzadko przekraczały liczbę 300 alleli. Powyższe ograniczenia zastosowanej technologii sekwencjonowania krótkich odczytów powodowały, że w bazie STRCatalog nie zostały uwzględnione loci zawierające sekwencję tandemowo powtórzoną dłuższą niż 100 pz w tym m.in. najdłuższe z loci CODIS takie jak: locus FGA z zakresem długości: 13-51 powtórzeń tetranuklotydowych, locus D21S11 z zakresem 24-38 powtórzeń tetranuklotydowych, a także locus SE33 z zakresem 4-37 powtórzeń tetranukleotydowych.

Dodatkowym wyzwaniem związanym z technologią sekwencjonowania krótkich odczytów jest nadreprezentacja krótkich alleli w bazie danych 1000 Genomes w porównaniu z dłuższymi allelami tego samego locus. Wynika to z faktu, że rezultaty sekwencjonowania fragmentów zawierających wyłącznie motyw powtórzony są odrzucane na etapie składania genomów.

Powyższe ograniczenia spowodowały, że wybór potencjalnych loci STR był ograniczony, jednakże mimo tego udało się zidentyfikować w grupie stosunkowo krótkich loci STR takie, które wykazywały założone w niniejszej pracy doktorskiej parametry pozwalające na wykorzystanie ich do opracowania metody badania pokrewieństwa w dalszych relacjach rodzinnych. Należy mieć również na uwadze, że krótkie, ale wysoce polimorficzne loci STR są bardziej przydatne w badaniach pokrewieństwa z wykorzystaniem metody multipleks-PCR niż loci z długimi allelami o takiej samej heterozygotyczności, ponieważ krótkie sekwencje tandemowo powtórzone generują niższe artefakty typu stutter (Leclair i in. 2004), a także cechują się średnio niższym współczynnikiem mutacji (Steely i in. 2022: Verbiest i in., 2023) co znacząco ułatwia zarówno analizę wyników badań jak i późniejsze wnioskowanie w sprawach biologicznego pokrewieństwa.

W ramach prac prowadzonych w trakcie realizacji rozprawy doktorskiej, dane zagregowane w bazie STRCatalog zostały wykorzystane do wstępnej selekcji 155 loci STR wykazujących odpowiednie parametry predestynujące te loci do wykorzystania w opracowaniu nowej

metody badania pokrewieństwa. Jednym z podstawowych kryteriów wyboru loci STR na tym etapie była długość motywu tandemowo powtórzonego, wynosząca od 3 do 5 par zasad.

Kolejnym kryterium wyboru loci STR była maksymalna heterozygotyczność danego locus w populacji światowej. Do dalszych analiz wybierano jedynie loci o heterozygotyczności przekraczającej 80% - 85% w populacji światowej. Ważnym czynnikiem była również podobna dystrybucja alleli danego locus w różnych populacjach kontynentalnych. Przyjęcie tego dodatkowego kryterium wynikało z faktu, że zgodnie z założeniami niniejszej pracy doktorskiej, zaplanowane badania zostaną wykonane na DNA pochodzącym od osób z populacji polskiej. W przypadku niektórych loci mogło się okazać, że mimo wysokiej heterozygotyczności w populacji polskiej, w innych populacjach zmienność genetyczna tego samego locus byłaby niższa. Przyjęcie kryterium podobnego rozkładu alleli w populacjach kontynentalnych zwiększało prawdopodobieństwo, że loci będą miały podobną wartość informacyjną także w innych populacjach krajowych.

Trzecim kryterium wyboru loci STR był brak sekwencji zawierającej powtórzenia mono- i dinukleotydowe w sekwencjach flankujących locus STR. Obecność takich sekwencji w amplikonie powodowałaby powstawanie artefaktów typu stutter (Fazekas i in., 2010), co jest zjawiskiem bardzo niekorzystnym przy interpretacji elektroforegramów reakcji PCR. Zasadniczo odrzucano z dalszych badań loci, które zawierały sekwencję z powtórzeniami mononukleotydowymi dłuższą niż 7 pz lub sekwencję z powtórzeniami dinukleotydowymi dłuższą niż 20 pz, jednakże w przypadku niektórych dobrze rokujących wysoce polimorficznych loci STR (np. D02L234) zdecydowano się na empiryczne sprawdzenie czy sekwencja mono- lub dinukleotydowa nieco dłuższa niż w założonym kryterium będzie rzeczywiście generowała artefakty w takiej liczbie, która utrudniałaby interpretację otrzymywanych wyników. Po analizie wyników wobec dużej reprezentacji innych wysoce polimorficznych loci STR, wybrano jedynie loci nie sprzężone z sekwencjami zawierającymi powtórzenia mono- i dinukleotydowe.

Czwartym kryterium wyboru loci STR był brak obecności sekwencji wielokrotnie powtórzonych w genomie w sekwencjach flankujących locus STR. Obecność takich sekwencji mogłaby powodować powstawanie artefaktów w reakcji PCR (Hommelsheim i in., 2014),

a także obniżenie wydajności amplifikacji poprzez zużywanie starterów oraz trifosforanów deoksyrybonukleotydów do amplifikacji produktów niespecyficznych (Ruiz-Villalba i in., 2017).

Ostatnim istotnym kryterium wyboru locus była jego rozpiętość, czyli różnica długości sekwencji pomiędzy najkrótszym i najdłuższym allelem danego locus występującym w populacji. Pierwotnie założono, że kryterium granicznym będzie rozpiętość locus mniejsza niż 100 pz, jednak w trakcie postępów badań przesiewowych zmieniono to kryterium do 80 pz. Akceptacja loci STR o dużej rozpiętości byłaby niekorzystna ze względu na trudność w multipleksowaniu reakcji PCR. Biorąc pod uwagę wykorzystanie czterech kanałów fluorescencji do detekcji amplikonów oraz założenie, że długość amplikonów wszystkich loci powinna zawierać się w przedziale 60-550 pz wykorzystanie loci STR o małej rozpiętości pozwalało na umieszczenie większej liczby loci w dwóch opracowywanych reakcjach multipleks-PCR. Przyjęcie wymienionych kryteriów pozwoliło na wybranie łącznie 183 loci STR genomu ludzkiego spełniających wszystkie wymienione kryteria i przeprowadzenie badań przesiewowych mających na celu określenie heterozygotyczności oraz zakresu długości alleli tych loci. Przykładowa analiza sekwencji dwóch loci STR: D02L234 oraz D3A57 oraz wyniki amplifikacji tych loci wraz z komentarzem przedstawione są na rycinach 30 i 31.

Ryc. 30. Analiza sekwencji nukleotydowej (A) oraz wyniki amplifikacji locus D02L234 dla DNA pochodzącego od pojedynczej osoby (B) oraz dla mieszaniny DNA pochodzącego od 50 osób (C). Analiza sekwencji nukleotydowej w programie Genome Data Viewer (NCBI) wskazuje obecność dwóch sekwencji tandemowo powtórzonych: sekwencję podkreśloną zieloną linią zawierającą 14 powtórzeń motywu teranukleotydowego AAAG oraz sekwencję podkreśloną czerwoną linią zawierającą ciąg 14 adenin (14 A). Niebieskie linie oraz prostokąty znajdujące się w polu pod sekwencją nukleotydową oznaczają występowanie insercji/delecji, natomiast czerwone prostokąty oznaczają miejsca występowania polimorfizmów SNP. Optymalnym sposobem amplifikacji tego locus byłoby zaprojektowanie startera w sekwencji znajdującej bezpośrednio przy locus STR z powtórzeniami tetranukleotydowymi omijając amplifikację ciągu mononukleotydowego (14 A) jednak ze względu na występowanie w tej sekwencji szeregu polimorfizmów zaprojektowanie startera w tym miejscu jest niewskazane. Zaprojektowanie startera obejmującego amplifikację sekwencji 14 A powoduje generowanie w reakcji PCR dużej liczby artefaktów stutter, co utrudnia analizę elektroforegramów. Ze względu na powyższe okoliczności locus D02L234 odrzucono z dalszych prac zmierzających do opracowania reakcji multipleks-PCR.

Ryc. 31. Analiza sekwencji nukleotydowej (A) oraz wyniki amplifikacji locus D3A57 dla DNA pochodzącego od pojedynczej osoby (B) oraz dla czterech mieszanin DNA pochodzącego od 50 osób nałożonych na siebie (C). Analiza sekwencji nukleotydowej w programie Genome Data Viewer (NCBI) wskazuje obecność dwóch sekwencji tandemowo powtórzonych sprzężonych ze sobą: sekwencji zawierającej powtórzenia tetranukleotydowe GAAA (czerwone podkreślenie) oraz GGAA (niebieskie podkreślenie) oraz sekwencji zawierającej powtórzenia dinukleotydowe GA (zielone podkreślenie). Sekwencja zawierająca powtórzenia dinukleotydowe GA również wykazuje polimorfizm długości, a także zawiera wtrącone jedno powtórzenie motywu AA co redukuje liczbę amplikonów typu stutter generowaną poprzez amplifikację tej sekwencji (B). Analiza elektroforegramów z badań przesiewowych wskazuje na wysoką polimorficzność locus D3A57.

5.2 Badania przesiewowe wybranych loci STR genomu ludzkiego

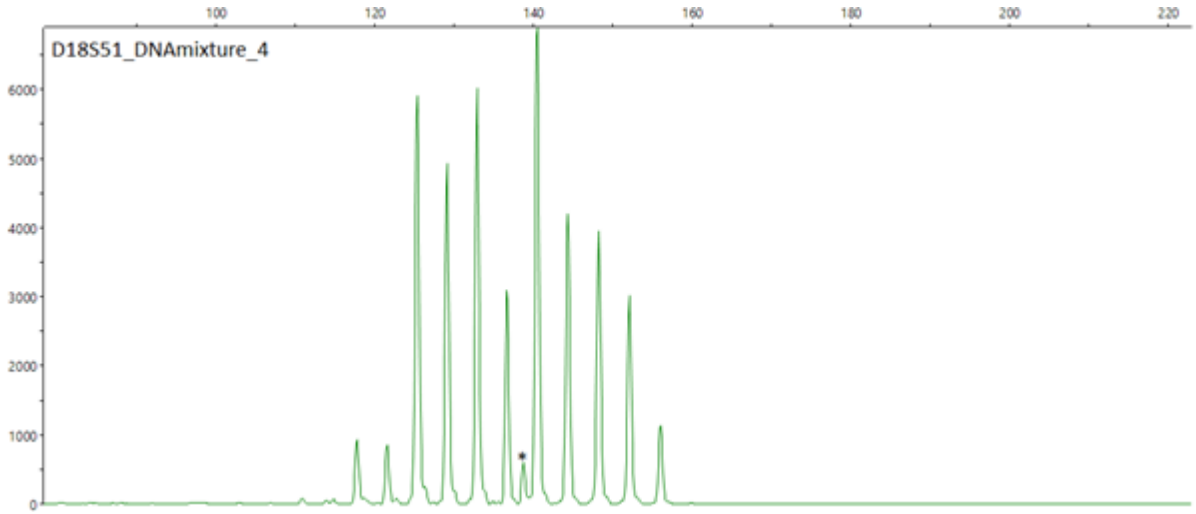
Omówione wcześniej ograniczenia w danych z sekwencjonowania NGS w bazie 1000 Genomes spowodowały, że koniecznym etapem prac stała się weryfikacja polimorficzności oraz zakresu długości alleli wybranych 183 loci STR poprzez realizację badań laboratoryjnych.

Na etapie planowania prac badawczych rozważano dwa sposoby weryfikacji heterozygotyczności wyselekcjonowanych loci STR w populacji polskiej, które pozwoliłyby wybrać najbardziej przydatne loci do badań pokrewieństwa. Metodologicznie oczywistym rozwiązaniem byłoby wykonanie genotypowania reprezentatywnej grupy osób z populacji polskiej w zakresie 183 badanych loci poprzez wykorzystanie pojedynczych reakcji PCR, a następnie określenie częstości występowania alleli analizowanych loci w populacji polskiej oraz wyznaczenie heterozygotyczności loci. Niestety takie podejście wiązałoby się z koniecznością wykonania tysięcy reakcji PCR i rozdziałów elektroforetycznych co znacząco podniosłoby koszty realizacji projektu. Dlatego w niniejszej rozprawie doktorskiej zastosowano innowacyjne podejście do oszacowania polimorfizmu wytypowanych loci poprzez wykonanie analizy dla każdego z 183 badanych STR na czterech pulach izolatów DNA pochodzących od pięćdziesięciu osób z populacji polskiej każda, które zostały zmieszane w równych stężeniach. Częstości występowania alleli w populacji polskiej miały zostać określone na podstawie analizy ilościowej i jakościowej amplikonów. Zgodnie z założeniami tej metody, stosunek pola powierzchni piku elektroforetycznego danego allelu do sumy powierzchni wszystkich pików elektroforetycznych miałby odzwierciedlać częstość występowania allelu w badanej grupie pięćdziesięciu osób. Takie podejście umożliwiło znaczące ograniczenie liczby reakcji PCR i rozdziałów elektroforetycznych, co przyczyniło się do oszczędności czasu i kosztów analiz. Ponieważ proponowana metoda badań przesiewowych była nowa dla wiedzy zweryfikowano jej poprawność i przydatność dla celów rozprawy doktorskiej wykonując badania przesiewowe na spulowanych izolatach DNA dla 13 loci CODIS, powszechnie stosowanych w rutynowej genetyce sądowej (Hares, 2015), dla których częstość występowania alleli w populacji polskiej jest znana (Wróbel i in., 2019; Ossowski i in., 2017).

Badanie to wykazało, że średnia różnica między heterozygotycznością określoną w badaniach przesiewowych metodą zaproponowaną w niniejszej rozprawie doktorskiej a heterozygotycznością wyznaczoną na podstawie analizy pojedynczych osób, podaną

w literaturze naukowej (Ossowski i in., 2017), wyniosła zaledwie 2,57%. Różnicę tę uznano za nieistotną dla tego etapu badań, ponieważ badania przesiewowe na tym etapie doktoratu wdrożeniowego miały na celu jedynie rozróżnienie loci o wysokiej polimorficzności (z heterozygotycznością >80%) od tych o mniejszej polimorficzności (heterozygotyczność 60% -80%). Dokładny rozkład alleli w populacji oraz heterozygotyczność badanych loci zostały wyznaczone na późniejszym etapie badań po opracowaniu reakcji multipleks-PCR.

Zastosowane podejście, polegające na jednoczesnej amplifikacji alleli występujących u 50 osób okazało się eksperymentalnie bardzo wydajne i czułe, o czym świadczy wykrycie bardzo rzadkiego allelu 15.2 w locus D18S51 (ryc. 32), co dowiodło, że analiza pulowanego DNA od 50 osób pozwala na identyfikację pojedynczych alleli w próbce.



Ryc. 32. Elektroforegram przedstawia zamplifikowane allele locus D18S51. Matrycę DNA do reakcji PCR stanowiła mieszanina DNA pochodzącego od 50 osób z populacji polskiej. Rzadki allel 15.2 został oznaczony symbolem „*”.

Analiza pulowanego DNA pochodzącego od większej liczby osób mogłaby utrudnić identyfikację rzadkich alleli ze względu na obniżenie wysokości piku elektroforetycznego dla takiego allelu i możliwość pomylenia go z pikiem typu stutter lub innym artefaktem reakcji.

Badania przesiewowe zrealizowane w niniejszej pracy przy zastosowaniu nowatorskiej metody szacowania heterozygotyczności wykazały, że analiza nawet niekompletnych danych

uzyskanych z projektu 1000 Genomes może prowadzić do identyfikacji niezwykle cennych w kontekście badań pokrewieństwa loci STR, które są dużo bardziej informatywne niż loci STR systemu CODIS. Dodatkowo, zastosowana w pracy szybka i oszczędna metoda pośredniego znakowania produktu PCR poprzez M13-tailing okazała się wysoce efektywna zarówno pod względem wydajności amplifikacji, jak i oszczędności reagentów. Zdaniem autora rozprawy doktorskiej opracowane podejście może być z powodzeniem stosowane w badaniach przesiewowych loci STR innych organizmów eukariotycznych.

Wyniki pracy pokazują, że zastosowane dane wstępne i opracowana de novo metoda szacowania polimorfizmu loci STR, umożliwiły zweryfikowanie heterozygotyczności wytypowanych loci w populacji polskiej. Zgodnie z wcześniejszymi założeniami potwierdzono, że obserwowana heterozygotyczność części loci STR jest wyższa niż raportowana w bazie projektu 1000 Genomes oraz w bazach STRCatalog i WebSTR, co potwierdziło ograniczenia zastosowanej w projekcie 1000 Genomes technologii sekwencjonowania krótkich odczytów w analizie polimorfizmu sekwencji mikrosatelitarnych genomu ludzkiego.

5.3 Opracowany zestaw Kinfinder

Zdecydowanie najtrudniejszym zadaniem pracy doktorskiej było zaprojektowanie dwóch reakcji multipleks-PCR umożliwiających jednoczesną analizę 25 loci STR w każdej z nich. Projektowanie reakcji multipleks-PCR stanowi wyzwanie, ponieważ konieczne jest jednoczesne spełnienie wielu wymagań (Markoulatos i in., 2002; Elnifro i in., 2000). Kluczowym elementem reakcji multipleks-PCR jest opracowanie starterów PCR, które zapewnią efektywną amplifikację alleli wszystkich analizowanych w tej reakcji loci STR. Aby zagwarantować wydajną amplifikację, startery muszą zostać zaprojektowane tak, aby nie hybrydyzowały z innymi starterami w tej samej reakcji. (Xie i in., 2022). Biorąc pod uwagę, że w opracowywanych reakcjach znajdowała się mieszanina 50 (multipleks B) lub 52 oligonukleotydów (multipleks A), to liczba potencjalnych oddziaływań starter-starter w reakcji multipleks-PCR A wyniosła 4950, zaś dla reakcji multipleks B liczba ta wyniosła 5051. W praktyce manualne sprawdzenie niepożądanych oddziaływań dla każdej pary starterów w reakcji przy takiej liczbie amplifikowanych produktów jest niemożliwe i koniecznym staje się wykorzystanie specjalistycznego oprogramowania (Ganschow i in., 2019; Guo i in., 2021). W

trakcie prac badawczych opisanych w niniejszej pracy, do identyfikacji potencjalnych niepożądanych oddziaływań pomiędzy starterami wykorzystano programy MultiPLX oraz Autodimer (pkt. 3.5). Wykorzystanie tych programów umożliwiło identyfikację potencjalnych niepożądanych oddziaływań pomiędzy projektowanymi starterami co znacznie usprawniło prace przy projektowaniu reakcji multipleks-PCR.

Kolejnym istotnym warunkiem jest zastosowanie starterów o zbliżonej temperaturze przyłączania (annealingu) do matrycy (Sint i in., 2012). Przy projektowaniu reakcji multipleks-PCR przyjęto zgodnie z zaleceniami innych autorów (Xu i in., 2012; Sint i in., 2012), że najodpowiedniejszą będzie temperatura równa 60 st.C. przy tolerancji +/- 5st.C. W niektórych przypadkach do 5' końca starterów dodawano siedmionukleotydową sekwencję PIG-tail: GTTTCTT (Brownstein i in., 1996). Wydłużenie starterów o tę sekwencję nie wpływało negatywnie na wydajność reakcji.

Dodatkowym utrudnieniem jest konieczność takiego zaprojektowania starterów, aby amplikony dla poszczególnych loci miały odpowiedni zakres długości, a produkty amplifikacji alleli z jednego locus nie nakładały się na allele innego locus w tym samym kanale fluorescencyjnym detekcji (Schilz i in., 2004). Ze względu na spodziewaną konieczność modyfikacji sekwencji niektórych starterów oraz na możliwość wystąpienia w populacji niezidentyfikowanych w badaniach przesiewowych rzadkich alleli wykraczających długością poza oznaczony zakres długości locus zdecydowano się na zaprojektowanie przerw pomiędzy amplikonami alleli amplifikowanych loci o długości 10-20 pz, które okazały się wystarczające, co wykazały analizy prowadzone w wyniku wdrożenia metody Kinfinder w Laboratorium Diagnostyki Molekularnej GenMed.

Kolejnym istotnym aspektem branym pod uwagę przy projektowaniu reakcji multipleks-PCR była identyfikacja polimorfizmów typu SNP w sekwencjach wiązania starterów do matrycy mogących powodować zjawisko alleli zerowych (ang. null alleles) (Kline i in., 2011). Zjawisko występowania alleli zerowych było już wielokrotnie obserwowane w komercyjnych odczynnikach do identyfikacji osobniczej takich jak AmpFlSTR Identifiler kit (Applied Biosystems) (Dauber i in., 2008; Mizuno i in., 2008) lub PowerPlex® 21 System (Promega) (Yao i in., 2018). Często obserwowane zjawisko alleli zerowych w locus D8S1179 w populacji

filipińskiej doprowadziło firmę Applied Biosystems do zmiany starterów dla tego locus w zestawie do identyfikacji osobniczej AmpFlSTR Profiler Plus PCR amplification kit (Leibelt i in., 2003). Ze względu na powszechność występowania polimorfizmów SNP w sekwencjach flankujących loci STR (ryc. 31) zaprojektowanie starterów w miejscach nie zawierających polimorfizmów SNP było niemożliwe. Żeby zminimalizować możliwość wystąpienia zjawiska alleli zerowych w opracowywanych reakcjach multipleks-PCR projektowano startery w miejscach sekwencji flankujących niezawierających polimorfizmów SNP lub o częstości występowania SNP w populacji światowej nie wyższej niż 0,0002 (pkt. 3.5).

Konieczność spełnienia wszystkich tych wymagań jednocześnie znacznie komplikowała proces projektowania starterów, a zmiana każdego startera powodowała konieczność ponownej analizy wszystkich parametrów reakcji. W trakcie projektowania reakcji multipleks-PCR wielokrotnie zmieniano sekwencje starterów, zmieniano zakresy długości amplikonów loci STR, a nawet wymieniono dwa loci pomiędzy multipleksami A i B. Wydajność reakcji multipleks-PCR nie była równa dla wszystkich loci wobec czego konieczna okazała się zmiana stężeń starterów (tabela 8).

5.4 Sekwencje opracowanych loci STR

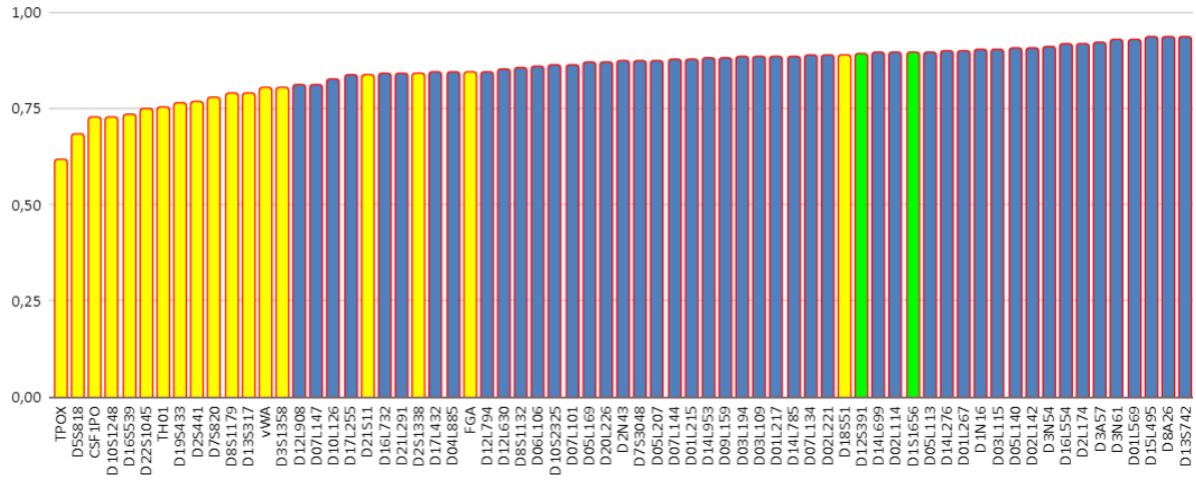
Znajomość sekwencji alleli i liczby powtórzeń tandemowych były niezbędne do przyjęcia nomenklatury dla wszystkich alleli. W przypadku alleli typowych, tzn. różniących się długością motywu powtórzonego, do nazwania alleli wystarczyło poznanie sekwencji jednego z alleli. W przypadku pięciu sekwencjonowanych loci (D12L794, D01L267, D13S742, D03L109 oraz D14L699) zidentyfikowano polimorfizm sekwencji nukleotydowej przy braku polimorfizmu długości (ryc. 28 i 29).

Polimorfizm sekwencji nukleotydowej alleli o tej samej długości jest zjawiskiem znanym i dotyczy również loci CODIS (Zhang i in., 2018; Dai i in., 2019). Polimorfizmów SNP nie prowadzących do zmiany długości sekwencji nie da się zaobserwować wykorzystując standardową analizę długości amplifikowanych fragmentów, jednak może być ona obserwowana przy zastosowaniu metody sekwencjonowania NGS (Novroski i.in. 2019), która powoli znajduje zastosowanie w niektórych badaniach z zakresu genetyki sądowej (Børsting i Morling, 2015; Alvarez-Cubero i in., 2017). Ewentualne wykorzystanie w przyszłości

technologii NGS w standardowych badaniach polimorfizmu loci STR zwiększyłoby jeszcze bardziej informatywność analizy tych samych loci w badaniach pokrewieństwa.

5.5 Heterozygotyczność opracowanych loci STR w populacji polskiej

Częstość występowania alleli 50 wysoce polimorficznych loci STR w populacji polskiej analizowano przy wykorzystaniu reakcji multipleks-PCR A i B. Badania populacyjne potwierdziły wysoką heterozygotyczność badanych loci STR (tabela nr 16). Najwyższą zmiennością w populacji polskiej spośród wybranych loci cechowały się loci D8A26, D15L495, D13S742 o heterozygotyczności odpowiednio: 0,9345, 0,9353, 0,939. Wyniki wskazują, że te loci są zbliżone do heterozygotyczności locus SE33, uznawanego za najbardziej polimorficzne locus ludzkiego genomu wykorzystywane w genetyce sądowej (Santos i in., 2023; Bhinder i in., 2018) i cechujące się w populacji polskiej heterozygotycznością od 93,4% do 95,4% (Wojtkiewicz i in., 2016; Jacewicz i in. 2008; Wróbel i in. 2019). Badania populacyjne wykazały zdecydowanie wyższą heterozygotyczność 50 analizowanych loci STR wchodzących w skład systemu Kinfinder w populacji polskiej (średnia heterozygotyczność = 88,07%) w porównaniu do heterozygotyczności loci CODIS (średnia heterozygotyczność = 78,95%) (Wróbel i in., 2019). Heterozygotyczność 50 loci STR opisanych w niniejszej pracy w porównaniu do loci CODIS przedstawiono na ryc. 33.



Ryc. 33. Heterozygotyczność loci STR systemu CODIS (żółty) oraz loci STR zawartych w zestawie Kinfinder (niebieski). Dwa najbardziej polimorficzne loci CODIS: D12S391 D1S1656 (zielony) zostały również włączone do metody Kinfinder w celu kontroli przebiegu procesu laboratoryjnego badania próbek DNA oraz ograniczenia możliwości pomyłki próbek przez personel laboratoryjny poprzez

możliwość porównania profili genetycznych w obrębie tych loci z profilem genetycznym uzyskanym z użyciem komercyjnego zestawu do identyfikacji osobniczej.

Badania populacyjne potwierdziły przydatność i dokładność nowo opracowanej metody badań przesiewowych w szybkiej estymacji heterozygotyczności loci mikrosateliatarnych. Średnia różnica w heterozygotyczności oszacowanej w badaniach przesiewowych i heterozygotyczności wyznaczonej na podstawie wyników badań populacynych wyniosła 1,875%. Tak niska różnica w uzyskanych wynikach jest akceptowalna i jest mniejsza niż pomiędzy wartościami heterozygotyczności niektórych loci wskazanymi w różnych badaniach populacyjnych dla populacji polskiej dla tego samego locus; przykładowo heterozygotyczność locus TPOX w publikacji (Wróbel i in., 2019) wynosi 0,579, natomiast w publikacji (Jacewicz i in., 2007) wynosi 0,604, co oznacza różnicę 2,5%. Wartości heterozygotyczności 50 loci STR wskazane w bazie STRCatalog, oszacowane na podstawie badań przesiewowych oraz wyznaczone na podstawie badań populacyjnych zestawione są w tabeli 16.

Tabela 16. Heterozygotyczność 50 loci STR wchodzących w skład metody Kinfinder. Het. 1 oznacza heterozygotyczność wykazaną w bazach danych STRCatalog/WebSTR dla populacji światowej lub w danych literaturowych dla populacji krajowych, Het.2 oznacza heterozygotyczność wykazaną w badaniach przesiewowych dla populacji polskiej, Het.3 oznacza heterozygotyczność loci wykazaną w badaniach populacyjnych dla populacji polskiej.

Locus	Het. 1	Het. 2	Het. 3	Locus	Het. 1	Het. 2	Het. 3
D12L908	0,87	0,818	0,8105	D03L109	0,869	0,917	0,8848
D07L147	0,847	0,83	0,8112	D01L217	0,875	0,888	0,8855
D10L126	0,859	0,843	0,8261	D07L134	0,874	0,872	0,8856
D17L255	0,835	0,873	0,8333	D14L785	0,84	0,869	0,8863
D16L732	0,872	0,859	0,8401	D02L221	0,861	0,863	0,888
D21L291	0,856	0,851	0,8417	D14L699	0,858	0,885	0,8948
D04L885	0,863	0,844	0,842	D02L114	0,9	0,87	0,895
D17L432	0,864	0,92	0,843	D05L113	0,858	0,899	0,896
D12L794	0,862	0,854	0,8459	D14L276	0,881	0,898	0,8986
D12L630	0,877	0,855	0,8518	D01L267	0,889	0,85	0,8993
D8S1132	0,867	0,833	0,855	D1N16	0,9	0,874	0,902
D06L106	0,864	0,89	0,8582	D03L115	0,894	0,904	0,9042
D07L101	0,836	0,865	0,8622	D05L140	0,902	0,911	0,9051
D10S2325	0,838	0,856	0,8641	D02L142	0,887	0,917	0,9053
D05L169	0,876	0,881	0,8691	D3N54	0,9137	0,875	0,9114
D20L226	0,903	0,863	0,8701	D16L554	0,906	0,905	0,9161
D2N43	0,9	0,833	0,8722	D12S391	0,874	0,862	0,881
D7S3048	0,927	0,894	0,8737	D02L174	0,911	0,922	0,918
D05L207	0,884	0,887	0,8745	D3A57	0,8	0,934	0,923
D07L144	0,88	0,891	0,8757	D3N61	0,9	0,9	0,9271
D01L215	0,879	0,854	0,8764	D01L569	0,863	0,901	0,9277
D14L953	0,878	0,851	0,8795	D8A26	0,9491	0,909	0,9345
D09L159	0,862	0,866	0,8809	D15L495	0,9	0,905	0,9353
D03L194	0,883	0,861	0,883	D13S742	0,891	0,911	0,939 0
D08L110	0,892	0,832	0,8582	D1S1656	0,874	0,88	0,8952

5.6 Symulacje badań pokrewieństwa z wykorzystaniem metody Kinfinder

W celu oceny przydatności metody Kinfinder w badaniach pokrewieństwa wykonano komputerowe symulacje badania pokrewieństwa dla czterech relacji rodzinnych; rodzic - dziecko, pełne rodzeństwo, relacja drugiego stopnia oraz relacja trzeciego stopnia. W badaniach porównano informatywność metody wykorzystującej system Globalfiler rutynowo stosowany w badaniach pokrewieństwa w laboratoriach genetyczno-sądowych oraz połączenie systemu Globalfiler z metodą Kinfinder umożliwiającą łączną analizę 69 loci STR. Symulacje wykonano z wykorzystaniem programu komputerowego KinBN.

Symulacje wykazały, że standardowa analiza 21 loci STR (Globalfiler) jest wystarczająca do zastosowania w badaniach ojcostwa (mediana LR = $2{,}90 \times 10^{7}$) oraz w badaniach pokrewieństwa w relacji pełne rodzeństwo (mediana LR = $2{,}01 \times 10^{6}$) co pokrywa się z danymi literaturowymi (Inoue i in., 2016). Dane literaturowe wskazują również, że analiza 21 loci STR jest niewystarczająca w badaniach pokrewieństwa w relacji drugiego i trzeciego stopnia (Tamura i in., 2015). Wyniki symulacji analiz pokrewieństwa przedstawione w niniejszej pracy potwierdzają ten pogląd, mediany spodziewanych wyników badań pokrewieństwa dla relacji drugiego i trzeciego stopnia pokrewieństwa uzyskiwane przy zastosowaniu analizy 21 loci STR (Globalfiler) wynoszą odpowiednio: 71,4 oraz 2,93. Co jest również istotne analiza 21 loci w badaniach pokrewieństwa drugiego i trzeciego stopnia jest obarczona ryzykiem uzyskania wyników fałszywie pozytywnych. Wyniki analiz statystycznych wykazały, że w badaniu pokrewieństwa drugiego stopnia przy analizie 21 loci STR (Globalfiler) możliwe jest otrzymanie dla osób spokrewnionych wyniku LR = $1{,}05 \times 10^{-2}$ co z dużym prawdopodobieństwem wskazuje na brak pokrewieństwa pomiędzy badanymi osobami. Z kolei w przypadku badania osób niespokrewnionych możliwe jest otrzymanie wyniku LR= $1{,}24 \times 10^{2}$ co z dużym prawdopodobieństwem wskazuje na pokrewieństwo tych osób. W przypadku rozszerzonej analizy 69 loci STR (Globalfiler + Kinfinder) problem ten nie występuje. Połączenie metod Kinfinder i Globalfiler wykazało doskonałą przydatność w badaniach drugiego stopnia i dobrą przydatność w badaniach pokrewieństwa w relacji trzeciego stopnia (mediany LR odpowiednio: $1{,}13 \times 10^{7}$ i 158).

Metoda Kinfinder może być również przydatna przy identyfikacji ofiar katastrof masowych, ofiar wojen i totalitaryzmów, a także osób zaginionych. W takich przypadkach profil

genetyczny osoby zmarłej lub zaginionej porównywany jest z bazą profili genetycznych krewnych osób zaginionych (Prinz i in. 2007). Bazy profili DNA krewnych nie zawsze obejmują profile genetyczne rodziców lub dzieci osób zaginionych, dlatego w niektórych przypadkach konieczne jest przeprowadzenie analizy dalszych stopni pokrewieństwa. W takich badaniach kluczowa jest wysoka informatywność analizy genetycznej (Hartman i in., 2011; Prinz i in., 2007). Wysoka informatywność metody Kinfinder w analizie pokrewieństwa drugiego stopnia (mediana LR = $1{,}13 \times 10^{7}$) sugeruje, że może ona być przydatna nawet przy porównywaniu profili genetycznych w dużych bazach danych zawierających tysiące profili.

5.5 Wdrożenie metody Kinfinder do rutynowej pracy laboratorium genetycznego.

Metoda Kinfinder, jest obecnie rutynowo stosowana w Laboratorium Diagnostyki Molekularnej GenMed zarówno w badaniach świadczonych na rzecz klientów indywidualnych, jak i w sprawach sądowych w postępowaniu spadkowym oraz w sprawach o ustalenie ojcostwa. Zastosowanie metody Kinfinder umożliwiło uzyskanie jednoznacznych wyników potwierdzających lub wykluczających biologiczne pokrewieństwo w przypadkach, w których dostępne komercyjnie rozwiązania okazywały się niewystarczające. Do momentu złożenia pracy doktorskiej metoda została zastosowana w Laboratorium Diagnostyki Molekularne GenMed ponad trzydzieści razy, każdorazowo pozwalając na uzyskanie rozstrzygającego wyniku biologicznego pokrewieństwa. Obecnie z jej stosowania przeszkolono trzech genetyków sądowych pracujących w laboratorium. Wdrożenie metody przebiegło sprawnie, bez potrzeby zakupu nowego sprzętu, korzystając wyłącznie z istniejącej aparatury laboratorium. Obecnie badania z wykorzystaniem tej metody i realizowane w Laboratorium Diagnostyki Molekularne GenMed obejmujące analizę 72 markerów DNA (69 loci STR oraz loci amelogeniny, DYS391, i Y-indel) oferowane są przez podmioty zewnętrzne będące partnerami firmy, w tym przez lidera badań diagnostycznych na polskim rynku spółkę Diagnostyka S.A (link) oraz ich podmiot zależny Zdrowegeny.pl sp. z o.o. (link). Kolejnym krokiem w rozwoju opracowanej metody będzie produkcja zestawu odczynników Kinfinder i ich sprzedaż do wykorzystania w zewnętrznych laboratoriach.

6. Wnioski

1. Baza danych projektu 1000 Genomes zawiera istotne informacje odnośnie do polimorficznych loci STR, jednakże dane są niepełne. Baza zawiera pełne informacje na temat krótkich loci STR, długie sekwencje są nieobecne lub dane są tym silniej zniekształcone im locus zawiera dłuższe allele.

2. Baza WebSTR (https://webstr.ucsd.edu/) jest doskonałym narzędziem do wyszukiwania potencjalnie polimorficznych sekwencji STR w genomie ludzkim.

3. Genom człowieka zawiera dużą liczbę polimorficznych loci STR, jednak elementy ich sekwencji flankujących, takie jak obecność: sekwencji wielokrotnie powtórzonych w genomie, sprzężonych sekwencji zawierających powtórzenia mono- i dinukleotydowe oraz sekwencji wykazujących wysoki polimorfizm w populacji utrudniający zaprojektowanie starterów może utrudniać lub nawet uniemożliwiać wykorzystanie tych loci w badaniach pokrewieństwa.

4. Opracowana w tej rozprawie doktorskiej, nowatorska metoda badań przesiewowych umożliwia szybką selekcję polimorficznych loci STR zawierających powtórzenia tri- tetra- i pentanukleotydowe. Ze względu na uniwersalne właściwości sekwencji STR, metoda ta może być wykorzystana do testowania polimorfizmu loci STR u innych organizmów w trakcie opracowywania analogicznych metod badania pokrewieństwa zwierząt hodowlanych, identyfikacji odmian roślin i in.

5. Opracowanie reakcji multipleks PCR jest trudnym wyzwaniem rosnącym wraz z liczbą produktów generowanych w reakcji, jednak stworzenie reakcji do jednoczesnej analizy 25 loci STR jest możliwe.

6. Średnia heterozygotyczność nowo scharakteryzowanych loci STR wykorzystywanych w metodzie Kinfinder w populacji polskiej wynosi 88,07% i jest znacząco wyższa niż heterozygotyczność loci CODIS wynosząca 78,95%.

7. Sekwencjonowanie alleli STR sugeruje, że przynajmniej część z nich oprócz polimorfizmu długości sekwencji wykazuje także polimorfizm sekwencji nukleotydowej.

8. Nowo opracowana metoda analizy pokrewieństwa Kinfinder wykorzystująca analizę 50 wysoce polimorficznych loci STR cechuje się bardzo wysoką informatywnością w badaniach

pokrewieństwa.

9. Metoda Kinfinder wykorzystuje analizę autosomalnych loci STR, a przez to nie ma ograniczeń związanych z analizą chromosomów płci, wobec czego można ją wykorzystać w badaniach pokrewieństwa w dowolnej relacji rodzinnej.

10. Wdrożenie metody Kinfinder w laboratoriach genetycznych jest bezproblemowe, gdyż metoda wykorzystuje powszechnie stosowaną w laboratoriach genetyczno-sądowych technologię reakcji PCR w multipleksie w połączeniu z elektroforezą kapilarną.

7. Bibliografia

1000 Genomes Project Consortium; Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56-65. doi: 10.1038/nature11632. PMID: 23128226; PMCID: PMC3498066.

Abdul-Muneer PM. Application of microsatellite markers in conservation genetics and fisheries management: recent advances in population structure analysis and conservation strategies. Genet Res Int. 2014;2014:691759. doi: 10.1155/2014/691759. Epub 2014 Apr 7. PMID: 24808959; PMCID: PMC3997932.

Alvarez-Cubero MJ, Saiz M, Martínez-García B, Sayalero SM, Entrala C, Lorente JA, Martinez-Gonzalez LJ. Next generation sequencing: an application in forensic sciences? Ann Hum Biol. 2017 Nov;44(7):581-592. doi: 10.1080/03014460.2017.1375155. Epub 2017 Sep 26. PMID: 28948844.

Ballard D, Winkler-Galicki J, Wesoły J. Massive parallel sequencing in forensics: advantages, issues, technicalities, and prospects. Int J Legal Med. 2020 Jul;134(4):1291-1303. doi: 10.1007/s00414-020-02294-0. Epub 2020 May 25. PMID: 32451905; PMCID: PMC7295846.

Berent J. DNAStat, version 2.1--a computer program for processing genetic profile databases and biostatistical calculations. Arch Med Sadowej Kryminol. 2010 Apr-Sep;60(2-3):118-26. English, Polish. PMID: 21516944.

Bhinder MA, Zahoor MY, Sadia H, Qasim M, Perveen R, Anjum GM, Iqbal M, Ullah N, Shehzad W, Tariq M, Waryah AM. SE33 locus as a reliable genetic marker for forensic DNA analysis systems. Turk J Med Sci. 2018 Jun 14;48(3):611-614. doi: 10.3906/sag-1801-21. PMID: 29916220.

Blouin AG, Askar M. Chimerism analysis for clinicians: a review of the literature and worldwide practices. Bone Marrow Transplant. 2022 Mar;57(3):347-359. doi: 10.1038/s41409-022-01579-9. Epub 2022 Jan 26. PMID: 35082369; PMCID: PMC9446524.

Boland CR, Goel A. Microsatellite instability in colorectal cancer. Gastroenterology. 2010 Jun;138(6):2073-2087.e3. doi: 10.1053/j.gastro.2009.12.064. PMID: 20420947; PMCID: PMC3037515.

Boutin-Ganache I, Raposo M, Raymond M, Deschepper CF. M13-tailed primers improve the readability and usability of microsatellite analyses performed with two different allele-sizing methods. Biotechniques. 2001 Jul;31(1):24-6, 28. PMID: 11464515.

Børsting C, Morling N. Next generation sequencing and its applications in forensic genetics. Forensic Sci Int Genet. 2015 Sep;18:78-89. doi: 10.1016/j.fsigen.2015.02.002. Epub 2015 Feb 14. PMID: 25704953.

Bradford L, Heal J, Anderson J, Faragher N, Duval K, Lalonde S. Disaster victim investigation recommendations from two simulated mass disaster scenarios utilized for user acceptance testing CODIS 6.0. Forensic Sci Int Genet. 2011 Aug;5(4):291-6. doi: 10.1016/j.fsigen.2010.05.005. Epub 2010 Jul 9. PMID: 20620126.

Brookes C, Bright JA, Harbison S, Buckleton J. Characterising stutter in forensic STR multiplexes. Forensic Sci Int Genet. 2012 Jan;6(1):58-63. doi: 10.1016/j.fsigen.2011.02.001. Epub 2011 Mar 8. PMID: 21388903.

Brouwer JR, Willemsen R, Oostra BA. Microsatellite repeat instability and neurological disease. Bioessays. 2009 Jan;31(1):71-83. doi: 10.1002/bies.080122. PMID: 19154005; PMCID: PMC4321794.

Brownstein MJ, Carpten JD, Smith JR. Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. Biotechniques. 1996 Jun;20(6):1004-6, 1008-10. doi: 10.2144/96206st01. PMID: 8780871.

Dai W, Pan Y, Sun X, Wu R, Li L, Yang D. High polymorphism detected by massively parallel sequencing of autosomal STRs using old blood samples from a Chinese Han population. Sci Rep. 2019 Dec 12;9(1):18959. doi: 10.1038/s41598-019-55282-9. PMID: 31831766; PMCID: PMC6908607.

Doniec A, Łuczak W, Wróbel M, Januła M, Ossowski A, Grzmil P, Kupiec T. Confirmation of Paternity despite Three Genetic Incompatibilities at Chromosome 2. Genes (Basel). 2021 Jan 4;12(1):62. doi: 10.3390/genes12010062. PMID: 33406744; PMCID: PMC7824413.

Dauber E.M, Glock B, Mayr W.R. Two examples of null alleles at the D19S433 locus due to the same 4bp deletion in the presumptive primer binding site of the AmpFlSTR Identifiler kit. Forensic Science International: Genetics Supplement Series. 2008; 1(1): 107-108.

de Vries JH, Kling D, Vidaki A, Arp P, Kalamara V, Verbiest MMPJ, Piniewska-Róg D, Parsons TJ, Uitterlinden AG, Kayser M. Impact of SNP microarray analysis of compromised DNA on kinship classification success in the context of investigative genetic genealogy. Forensic Sci Int Genet. 2022 Jan;56:102625. doi: 10.1016/j.fsigen.2021.102625. Epub 2021 Nov 1. PMID: 34753062.

Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004 Jun;5(6):435-45. doi: 10.1038/nrg1348. PMID: 15153996.

Elnifro EM, Ashshi AM, Cooper RJ, Klapper PE. Multiplex PCR: optimization and application in diagnostic virology. Clin Microbiol Rev. 2000 Oct;13(4):559-70. doi: 10.1128/CMR.13.4.559. PMID: 11023957; PMCID: PMC88949.

Fazekas A, Steeves R, Newmaster S. Improving sequencing quality from PCR products containing long mononucleotide repeats. Biotechniques. 2010 Apr;48(4):277-85. doi: 10.2144/000113369. PMID: 20569204.

Frontanilla TS, Valle-Silva G, Ayala J, Mendes-Junior CT. Open-Access Worldwide Population STR Database Constructed Using High-Coverage Massively Parallel Sequencing Data Obtained from the 1000 Genomes Project. Genes (Basel). 2022 Nov 24;13(12):2205. doi: 10.3390/genes13122205. PMID: 36553472; PMCID: PMC9778533.

Ganschow S, Silvery J, Tiemann C. Development of a multiplex forensic identity panel for massively parallel sequencing and its systematic optimization using design of experiments. Forensic Sci Int Genet. 2019 Mar;39:32-43. doi: 10.1016/j.fsigen.2018.11.023. Epub 2018 Nov 30. PMID: 30529891.

Guo J, Starr D, Guo H. Classification and review of free PCR primer design software. Bioinformatics. 2021 Apr 1;36(22-23):5263-5268. doi: 10.1093/bioinformatics/btaa910. PMID: 33104196.

Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL, Pritchard JK, Sharp AJ, Erlich Y. Abundant contribution of short tandem repeats to gene expression variation in humans. Nat Genet. 2016 Jan;48(1):22-9. doi: 10.1038/ng.3461. Epub 2015 Dec 7. PMID: 26642241; PMCID: PMC4909355.

Gymrek M, Willems T, Reich D, Erlich Y. Interpreting short tandem repeat variations in humans using mutational constraint. Nat Genet. 2017 Oct;49(10):1495-1501. doi: 10.1038/ng.3952. Epub 2017 Sep 11. PMID: 28892063; PMCID: PMC5679271.

Hares DR. Selection and implementation of expanded CODIS core loci in the United States. Forensic Sci Int Genet. 2015 Jul;17:33-34. doi: 10.1016/j.fsigen.2015.03.006. Epub 2015 Mar 12. PMID: 25797140.

Hartman D, Drummer O, Eckhoff C, Scheffer JW, Stringer P. The contribution of DNA to the disaster victim identification (DVI) effort. Forensic Sci Int. 2011 Feb 25;205(1-3):52-8. doi: 10.1016/j.forsciint.2010.09.024. Epub 2010 Nov 23. PMID: 21106312.

Hauge XY, Litt M. A study of the origin of 'shadow bands' seen when typing dinucleotide repeat polymorphisms by the PCR. Hum Mol Genet. 1993 Apr;2(4):411-5. doi:

10.1093/hmg/2.4.411. PMID: 8504301.

Hile SE, Eckert KA. Positive correlation between DNA polymerase alpha-primase pausing and mutagenesis within polypyrimidine/polypurine microsatellite sequences. J Mol Biol. 2004 Jan 16;335(3):745-59. doi: 10.1016/j.jmb.2003.10.075. PMID: 14687571.

Hodel RG, Segovia-Salcedo MC, Landis JB, Crowl AA, Sun M, Liu X, Gitzendanner MA, Douglas NA, Germain-Aubrey CC, Chen S, Soltis DE, Soltis PS. The report of my death was an exaggeration: A review for researchers using microsatellites in the 21st century. Appl Plant Sci. 2016 Jun 16;4(6):apps.1600025. doi: 10.3732/apps.1600025. PMID: 27347456; PMCID: PMC4915923.

Hommelsheim CM, Frantzeskakis L, Huang M, Ülker B. PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications. Sci Rep. 2014 May 23;4:5052. doi: 10.1038/srep05052. PMID: 24852006; PMCID: PMC4031481.

Inoue H, Manabe S, Fujii K, Iwashima Y, Miyama S, Tanaka A, Saitoh H, Iwase H, Tamaki K, Sekiguchi K. Sibling assessment based on likelihood ratio and total number of shared alleles using 21 short tandem repeat loci included in the GlobalFiler™ kit. Leg Med (Tokyo). 2016 Mar;19:122-6. doi: 10.1016/j.legalmed.2015.07.008. Epub 2015 Jul 21. PMID: 26254055.

Jabłońska-Milczarek M, Frankowski A. Latest Interpol reports on disaster victim identification and the process of implementing international DVI standards in Poland. Arch Med Sadowej Kryminol. 2020;70(2-3):163-180. English. doi: 10.5114/amsik.2020.104493. PMID: 33853286.

Jacewicz R, Jedrzejczyk M, Berent J. The most efficient STR loci in forensic genetics in population of central Poland. Forensic Science International: Genetics Supplement. 2008;1(1):340-342, doi: 10.1016/j.fsigss.2007.10.056.

Jobling MA, Pandya A, Tyler-Smith C. The Y chromosome in forensic analysis and paternity testing. Int J Legal Med. 1997;110(3):118-24. doi: 10.1007/s004140050050. PMID: 9228562.

Kaplinski L, Remm M. MultiPLX: automatic grouping and evaluation of PCR primers. Methods Mol Biol. 2015;1275:127-42. doi: 10.1007/978-1-4939-2365-6_9. PMID: 25697656.

KIT S. Equilibrium sedimentation in density gradients of DNA preparations from animal tissues. J Mol Biol. 1961 Dec;3:711-6. doi: 10.1016/s0022-2836(61)80075-2. PMID: 14456492.

Kline MC, Hill CR, Decker AE, Butler JM. STR sequence analysis for characterizing normal, variant, and null alleles. Forensic Sci Int Genet. 2011 Aug;5(4):329-32. doi: 10.1016/j.fsigen.2010.09.005. Epub 2010 Oct 6. PMID: 20932816.

Leclair B, Frégeau CJ, Bowen KL, Fourney RM. Systematic analysis of stutter percentages and allele peak height and peak area ratios at heterozygous STR loci for forensic casework and database samples. J Forensic Sci. 2004 Sep;49(5):968-80. PMID: 15461097.

Leibelt C, Budowle B, Collins P, Daoudi Y, Moretti T, Nunn G, Reeder D, Roby R. Identification of a D8S1179 primer binding site mutation and the validation of a primer designed to recover null alleles. Forensic Sci Int. 2003 May 5;133(3):220-7. doi: 10.1016/s0379-0738(03)00035-5. PMID: 12787655.

Li YC, Korol AB, Fahima T, Beiles A, Nevo E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol Ecol.* 2002;11(12):2453–2465. doi: 10.1046/j.1365-294X.2002.01643.x.

Litt M, Luty JA. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet. 1989 Mar;44(3):397-401. PMID: 2563634; PMCID: PMC1715430.

Lundström OS, Adriaan Verbiest M, Xia F, Jam HZ, Zlobec I, Anisimova M, Gymrek M. WebSTR: A Population-wide Database of Short Tandem Repeat Variation in Humans. J Mol Biol. 2023 Oct 15;435(20):168260. doi: 10.1016/j.jmb.2023.168260. Epub 2023 Sep 7. PMID: 37678708.

Markoulatos P, Siafakas N, Moncany M. Multiplex polymerase chain reaction: a practical approach. J Clin Lab Anal. 2002;16(1):47-51. doi: 10.1002/jcla.2058. PMID: 11835531; PMCID: PMC6808141.

Mizuno N, Kitayama T, Fujii K, Nakahara H, Yoshida K, Sekiguchi K, Yonezawa N, Nakano M, Kasai K. A D19S433 primer binding site mutation and the frequency in Japanese of the silent allele it causes. J Forensic Sci. 2008 Sep;53(5):1068-73. doi: 10.1111/j.1556-4029.2008.00806.x. Epub 2008 Jul 11. PMID: 18636979.

Morimoto C, Tsujii H, Manabe S, Fujimoto S, Hirai E, Hamano Y, Tamaki K. Development of a software for kinship analysis considering linkage and mutation based on a Bayesian network. Forensic Sci Int Genet. 2020 Jul;47:102279. doi: 10.1016/j.fsigen.2020.102279. Epub 2020 Mar 19. PMID: 32289730.

Murray V, Monchawin C, England PR. The determination of the sequences present in the shadow bands of a dinucleotide repeat PCR. Nucleic Acids Res. 1993 May 25;21(10):2395-8. doi: 10.1093/nar/21.10.2395. PMID: 8506134; PMCID: PMC309538.

Novroski NMM, Wendt FR, Woerner AE, Bus MM, Coble M, Budowle B. Expanding beyond the current core STR loci: An exploration of 73 STR markers with increased diversity for enhanced DNA mixture deconvolution. Forensic Sci Int Genet. 2019 Jan;38:121-129. doi:

10.1016/j.fsigen.2018.10.013. Epub 2018 Oct 29. PMID: 30396008.

Nwawuba Stanley U, Mohammed Khadija A, Bukola AT, Omusi Precious I, Ayevbuomwan Davidson E. Forensic DNA Profiling: Autosomal Short Tandem Repeat as a Prominent Marker in Crime Investigation. Malays J Med Sci. 2020 Jul;27(4):22-35. doi: 10.21315/mjms2020.27.4.3. Epub 2020 Aug 19. PMID: 32863743; PMCID: PMC7444828.

Okazaki A, Yamazaki S, Inoue I, Ott J. Population genetics: past, present, and future. Hum Genet. 2021 Feb;140(2):231-240. doi: 10.1007/s00439-020-02208-5. Epub 2020 Jul 18. PMID: 32683493; PMCID: PMC7368598.

Ossowski A, Diepenbroek M, Szargut M, Zielińska G, Jędrzejczyk M, Berent J, Jacewicz R. Population analysis and forensic evaluation of 21 autosomal loci included in GlobalFiler™ PCR Kit in Poland. Forensic Sci Int Genet. 2017 Jul;29:e38-e39. doi: 10.1016/j.fsigen.2017.05.003. Epub 2017 May 10. PMID: 28522272.

Pedroza Matute S, Iyavoo S. Applications and Performance of Precision ID GlobalFiler NGS STR, Identity, and Ancestry Panels in Forensic Genetics. Genes (Basel). 2024 Aug 28;15(9):1133. doi: 10.3390/genes15091133. PMID: 39336724; PMCID: PMC11431077.

Prinz M, Carracedo A, Mayr WR, Morling N, Parsons TJ, Sajantila A, Scheithauer R, Schmitter H, Schneider PM; International Society for Forensic Genetics. DNA Commission of the International Society for Forensic Genetics (ISFG): recommendations regarding the role of forensic genetics for disaster victim identification (DVI). Forensic Sci Int Genet. 2007 Mar;1(1):3-12. doi: 10.1016/j.fsigen.2006.10.003. Epub 2006 Nov 28. PMID: 19083722.

Rajan-Babu IS, Peng JJ, Chiu R, Li C, Mohajeri A, Dolzhenko E, Eberle MA, Birol I, Friedman JM. Genome-wide sequencing as a first-tier screening test for short tandem repeat expansions. Genome Med. 2021 Aug 9;13(1):126. doi: 10.1186/s13073-021-00932-9. Erratum in: Genome Med. 2021 Sep 13;13(1):151. doi: 10.1186/s13073-021-00961-4. PMID: 34372915; PMCID: PMC8351082.

Read JL, Davies KC, Thompson GC, Delatycki MB, Lockhart PJ. Challenges facing repeat expansion identification, characterisation, and the pathway to discovery. Emerg Top Life Sci. 2023 Dec 14;7(3):339-348. doi: 10.1042/ETLS20230019. PMID: 37888797; PMCID: PMC10754332.

Ruiz-Villalba A, van Pelt-Verkuil E, Gunst QD, Ruijter JM, van den Hoff MJ. Amplification of nonspecific products in quantitative polymerase chain reactions (qPCR). Biomol Detect Quantif. 2017 Nov 1;14:7-18. doi: 10.1016/j.bdq.2017.10.001. PMID: 29255685; PMCID: PMC5727009.

Santos NBPD, de Paula Filho MFF, Silva AMDS, Teló EP, Junior JBDN, de Queiroz Balbino V, Takenami IO, Cansanção IF. Allele Frequencies and Forensic Data of 25 STR Markers for Individuals in Northeast Brazil. Genes (Basel). 2023 May 29;14(6):1185. doi: 10.3390/genes14061185. PMID: 37372365; PMCID: PMC10298256.

Sawaya S, Bagshaw A, Buschiazzo E, Kumar P, Chowdhury S, Black MA, Gemmell N. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. PLoS One. 2013;8(2):e54710. doi: 10.1371/journal.pone.0054710. Epub 2013 Feb 6. PMID: 23405090; PMCID: PMC3566118.

Schilz F, Hummel S, Herrmann B. Design of a multiplex PCR for genotyping 16 short tandem repeats in degraded DNA samples. Anthropol Anz. 2004 Dec;62(4):369-78. PMID: 15648845.

Shin CH, Jang P, Hong KM, Paik MK. Allele frequencies of 10 STR loci in Koreans. Forensic Sci Int. 2004 Feb 10;140(1):133-5. doi: 10.1016/j.forsciint.2003.11.027. PMID: 15013178.

Shinde D, Lai Y, Sun F, Arnheim N. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. Nucleic Acids Res. 2003 Feb 1;31(3):974-80. doi: 10.1093/nar/gkg178. PMID: 12560493; PMCID: PMC149199.

Schlötterer C, Tautz D. Slippage synthesis of simple sequence DNA. Nucleic Acids Res. 1992 Jan 25;20(2):211-5. doi: 10.1093/nar/20.2.211. PMID: 1741246; PMCID: PMC310356.

Schlötterer C. Evolutionary dynamics of microsatellite DNA. Chromosoma. 2000 Sep;109(6):365-71. doi: 10.1007/s004120000089. Erratum in: Chromosoma 2001 Feb;109(8):571. PMID: 11072791.

Shinde D, Lai Y, Sun F, Arnheim N. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. Nucleic Acids Res. 2003 Feb 1;31(3):974-80. doi: 10.1093/nar/gkg178. PMID: 12560493; PMCID: PMC149199.

Shrivastava P, Jain T, Kumawat RK. Direct PCR amplification from saliva sample using non-direct multiplex STR kits for forensic DNA typing. Sci Rep. 2021 Mar 29;11(1):7112. doi: 10.1038/s41598-021-86633-0. PMID: 33782478; PMCID: PMC8007628.

Sint D, Raso L, Traugott M. Advances in multiplex PCR: balancing primer efficiencies and improving detection success. Methods Ecol Evol. 2012 Oct;3(5):898-905. doi: 10.1111/j.2041-210X.2012.00215.x. PMID: 23549328; PMCID: PMC3573865.

Slatkin M, Racimo F. Ancient DNA and human history. Proc Natl Acad Sci U S A. 2016 Jun 7;113(23):6380-7. doi: 10.1073/pnas.1524306113. Epub 2016 Jun 6. PMID: 27274045; PMCID: PMC4988579.

Steely CJ, Watkins WS, Baird L, Jorde LB. The mutational dynamics of short tandem repeats in large, multigenerational families. Genome Biol. 2022 Dec 12;23(1):253. doi: 10.1186/s13059-022-02818-4. PMID: 36510265; PMCID: PMC9743774.

Takezaki N, Nei M. Genomic drift and evolution of microsatellite DNAs in human populations. Mol Biol Evol. 2009 Aug;26(8):1835-40. doi: 10.1093/molbev/msp091. Epub 2009 Apr 30. PMID: 19406937.

Tamura T, Osawa M, Ochiai E, Suzuki T, Nakamura T. Evaluation of advanced multiplex short tandem repeat systems in pairwise kinship analysis. Leg Med (Tokyo). 2015 Sep;17(5):320-5. doi: 10.1016/j.legalmed.2015.03.005. Epub 2015 Apr 1. PMID: 25851967.

Tang H, Kirkness EF, Lippert C, Biggs WH, Fabani M, Guzman E, Ramakrishnan S, Lavrenko V, Kakaradov B, Hou C, Hicks B, Heckerman D, Och FJ, Caskey CT, Venter JC, Telenti A. Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. Am J Hum Genet. 2017 Nov 2;101(5):700-715. doi: 10.1016/j.ajhg.2017.09.013. PMID: 29100084; PMCID: PMC5673627.

Vallone PM, Butler JM. AutoDimer: a screening tool for primer-dimer and hairpin structures. Biotechniques. 2004 Aug;37(2):226-31. doi: 10.2144/04372ST03. PMID: 15335214.

Verbiest M, Maksimov M, Jin Y, Anisimova M, Gymrek M, Bilgin Sonay T. Mutation and selection processes regulating short tandem repeats give rise to genetic and phenotypic diversity across species. J Evol Biol. 2023 Feb;36(2):321-336. doi: 10.1111/jeb.14106. Epub 2022 Oct 26. PMID: 36289560; PMCID: PMC9990875.

Westen AA, van der Gaag KJ, de Knijff P, Sijen T. Improved analysis of long STR amplicons from degraded single source and mixed DNA. Int J Legal Med. 2013 Jul;127(4):741-7. doi: 10.1007/s00414-012-0816-1. Epub 2013 Jan 10. PMID: 23306520.

Wilkening S, Chen B, Hemminki K, Försti A. STR markers for kinship analysis. Hum Biol. 2006 Feb;78(1):1-8. doi: 10.1353/hub.2006.0030. PMID: 16900878.

Willems T, Gymrek M, Highnam G; 1000 Genomes Project Consortium; Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res. 2014 Nov;24(11):1894-904. doi: 10.1101/gr.177774.114. Epub 2014 Aug 18. PMID: 25135957; PMCID: PMC4216929.

Wojtkiewicz R, Markiewicz B, Jędrzejczyk M, Jacewicz R. Polymorphism of 12 STR loci included in the Investigator HD-plex kit in Polish population of Lodz region. Arch Med Sadowej Kryminol. 2016;66(1):13-22. doi: 10.5114/amsik.2016.62331. PMID: 28155985.

Wróbel M, Parys-Proszek A, Marcińska M, Ba G, Sekuła A, Kowalczyk M, Januła M, Doniec A, Kupiec T. Analysis of the frequency of occurrence in the polish population of alleles of 21 genetic markers in the Globalfiler kit. *Probl. Forensic Sci.* 2019, *117*, 49–61.

Xie NG, Wang MX, Song P, Mao S, Wang Y, Yang Y, Luo J, Ren S, Zhang DY. Designing highly multiplex PCR primer sets with Simulated Annealing Design using Dimer Likelihood Estimation (SADDLE). Nat Commun. 2022 Apr 11;13(1):1881. doi: 10.1038/s41467-022-29500-4. PMID: 35410464; PMCID: PMC9001684.

Xu L, Haasl RJ, Sun J, Zhou Y, Bickhart DM, Li J, Song J, Sonstegard TS, Van Tassell CP, Lewin HA, Liu GE. Systematic Profiling of Short Tandem Repeats in the Cattle Genome. Genome Biol Evol. 2017 Jan 1;9(1):20-31. doi: 10.1093/gbe/evw256. PMID: 28172841; PMCID: PMC5381564.

Xu Q, Wang Z, Kong Q, Wang X, Huang A, Li C, Liu X. Improving the system power of complex kinship analysis by combining multiple systems. Forensic Sci Int Genet. 2022 Sep;60:102741. doi: 10.1016/j.fsigen.2022.102741. Epub 2022 Jun 18. PMID: 35780597.

Xu W, Zhai Z, Huang K, Zhang N, Yuan Y, Shang Y, Luo Y. A novel universal primer-multiplex-PCR method with sequencing gel electrophoresis analysis. PLoS One. 2012;7(1):e22900. doi: 10.1371/journal.pone.0022900. Epub 2012 Jan 17. PMID: 22272223; PMCID: PMC3260127.

Yao Y, Yang Q, Shao C, Liu B, Zhou Y, Xu H, Zhou Y, Tang Q, Xie J. Null alleles and sequence variations at primer binding sites of STR loci within multiplex typing systems. Leg Med (Tokyo). 2018 Jan;30:10-13. doi: 10.1016/j.legalmed.2017.10.007. Epub 2017 Oct 28. PMID: 29125964.

Zeng X, King JL, Stoljarova M, Warshauer DH, LaRue BL, Sajantila A, Patel J, Storts DR, Budowle B. High sensitivity multiplex short tandem repeat loci analyses with massively parallel sequencing. Forensic Sci Int Genet. 2015 May;16:38-47. doi: 10.1016/j.fsigen.2014.11.022. Epub 2014 Dec 3. PMID: 25528025.

Zhang S, Niu Y, Bian Y, Dong R, Liu X, Bao Y, Jin C, Zheng H, Li C. Sequence investigation of 34 forensic autosomal STRs with massively parallel sequencing. Sci Rep. 2018 May 1;8(1):6810. doi: 10.1038/s41598-018-24495-9. PMID: 29717145; PMCID: PMC5931506.

Zhang Q, Wang X, Cheng P, Yang S, Li W, Zhou Z, Wang S. Complex kinship analysis with a combination of STRs, SNPs, and indels. Forensic Sci Int Genet. 2022 Nov;61:102749. doi: 10.1016/j.fsigen.2022.102749. Epub 2022 Jul 20. PMID: 35939875.

Zhang Q, Zhou Z, Wang L, Quan C, Liu Q, Tang Z, Liu L, Liu Y, Wang S. Pairwise kinship testing with a combination of STR and SNP loci. Forensic Sci Int Genet. 2020 May;46:102265. doi:

10.1016/j.fsigen.2020.102265. Epub 2020 Feb 25. PMID: 32145445.