



Universiteit
Leiden
The Netherlands

Development of forensic genomics research toolkits by the use of Massively Parallel Sequencing

Gaag, K.J. van der

Citation

Gaag, K. J. van der. (2019, November 27). *Development of forensic genomics research toolkits by the use of Massively Parallel Sequencing*. Retrieved from <https://hdl.handle.net/1887/80956>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/80956>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/80956> holds various files of this Leiden University dissertation.

Author: Gaag, K.J. van der

Title: Development of forensic genomics research toolkits by the use of Massively Parallel Sequencing

Issue Date: 2019-11-27

Nederlandse samenvatting

Dit proefschrift richt zich op de ontwikkeling, toepassing en validatie van nieuwe forensische methoden gebruik makend van nieuwe beschikbare technieken bekend als 'Massively Parallel Sequencing' (afgekort MPS, ook bekend onder de naam 'Next Generation Sequencing').

In **hoofdstuk I** wordt de achtergrond beschreven van MPS zelf en de verschillende methoden die verder in het proefschrift worden besproken.

De ontwikkeling en toepassing van forensisch DNA onderzoek heeft zich snel ontwikkeld sinds de ontdekking van zeer variabele 'DNA vingerafdrukken' door Jeffreys et al. (1985).

De laatste ca. twintig jaar heeft de analyse van 'Short Tandem Repeats' (STRs) een grote rol gespeeld in identificatie m.b.v. DNA onderzoek. STRs zijn stukjes DNA die een herhaalde sequentie bevatten van 2-6 nucleotiden. Het aantal herhalingen van deze sequentie kan variëren tussen personen en door het analyseren van meerdere van deze STRs kan een vrijwel uniek DNA profiel worden vastgesteld. Een voorbeeld van een stukje DNA sequentie met daarin een STR wordt hieronder weergegeven.

Figuur 1, DNA sequentie die een AGAT-herhaling bevat

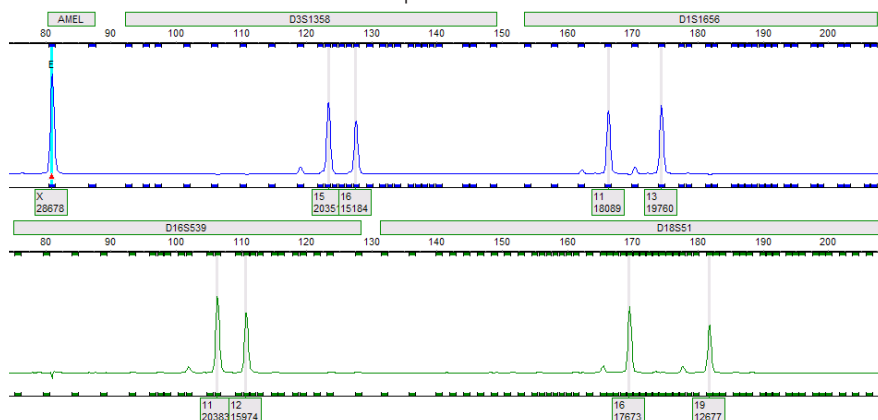
ATCCGA AGAT AGAT AGAT AGAT AGAT AGAT AGAT AGAT TGGACCAG

Conventionele DNA profielen

Capillaire elektroforese (CE) is de conventionele methode voor het analyseren van STRs. Voor analyse d.m.v. CE worden fragmenten die een STR bevatten eerst vermeerderd en gelabeld m.b.v. de zogenaamde polymerase chain reaction (PCR). Vervolgens worden de fragmenten op lengte van elkaar gescheiden en vergeleken met een ladder van fragmenten met een bekende lengte om zo het aantal herhalingen van elke STR vast te stellen in een DNA profiel.

Hoewel CE analyse van STRs een vrij eenvoudige techniek is, heeft het ook zijn limieten. Tijdens de amplificatie van STR-fragmenten ontstaan zgn. stutter-artefacten die vooral de interpretatie van een ongebalanceerd gemengd DNA profiel ingewikkelder kunnen maken. Daarnaast zijn er, om een voldoende uniek DNA profiel te verkrijgen, suboptimale condities nodig voor de analyse van gedegrademd DNA. Figuur 2 geeft een voorbeeld van een (deel van een) DNA profiel.

Figuur 2, Voorbeeld van een DNA profiel



Deel van een DNA profiel van een (enkele) persoon. De pieken zijn weergegeven voor vijf markers die gescheiden zijn bij lengte (in hetzelfde kleurkanaal) of fluorescent label.

Nieuwe technologieën: Massively Parallel Sequencing

Recent zijn er nieuwe DNA analyse technieken ontwikkeld die parallel de sequentie kunnen bepalen van miljoenen DNA moleculen. Terwijl deze MPS-technieken al toegepast worden in het medische DNA onderzoek en andere moleculaire onderzoeksgebieden zijn ze nog relatief nieuw in het forensische werkveld.

In principe zou MPS een aantal beperkingen van CE moeten kunnen voorkomen. Enkele beperkingen in CE STR analyse worden echter veroorzaakt door het type DNA marker en niet door de gebruikte techniek. Aangezien het type data, maar vooral de hoeveelheid data voor MPS anders is dan bij CE, is het cruciaal dat er gespecialiseerde forensische software ontwikkeld wordt om goed met deze data om te gaan en MPS toepassing in zaakwerk te implementeren.

In dit proefschrift wordt de toepassing van MPS in humaan forensisch DNA onderzoek verkend en wordt niet enkel het lab-gerelateerde praktische aspect onderzocht, maar ook de ontwikkeling van gespecialiseerde forensische software voor analyse van MPS data. Daarnaast verkennen we nieuwe niet-STR DNA markers als alternatief voor de huidige analyse om de toepassing van DNA analyse in de nabije toekomst nog verder uit te breiden.

Massively Parallel Sequencing vs Capillaire Elektroforese

De verwachte perspectieven van MPS in forensisch DNA onderzoek zijn gebaseerd op de volgende verschillen tussen CE en MPS.

- Terwijl CE enkel fragmentlengte analyseert, typeert MPS de exacte sequentie.

Sequentie variatie (bovenop lengtevariatie) verhoogt de statistische kracht van een STR. Daarnaast kan sequentie-variatie helpen om stutter artefacten en echte allelen van elkaar te onderscheiden.

- *De detectierange van CE is beperkt terwijl MPS een vrijwel ongelimiteerde detectierange heeft.* Voor CE zorgt te hoog signaal voor extra artefacten (in andere kleuren) terwijl erg laag signaal niet van ruis onderscheiden kan worden. Voor MPS kan het aantal uitgelezen sequenties vrijwel oneindig worden verhoogd zonder dat er artefacten ontstaan bij andere STRs (al stijgen de kosten mee met het aantal sequenties).
- *De capaciteit van CE om meerdere markers tegelijk te analyseren is beperkt terwijl MPS in principe duizenden markers tegelijk kan analyseren.* Aangezien er momenteel niet meer dan zes fluorescente labels gebruikt kunnen worden met CE, wordt het gewenste aantal markers in een reactie verkregen door de lengte van verschillende markers niet te laten overlappen. Voor MPS kunnen alle markers in dezelfde fragment-range ontworpen worden omdat ze onderscheiden worden op sequentie in plaats van kleur en lengte.

Massively Parallel Sequencing van Short Tandem Repeats

Het meeste werk dat beschreven wordt in dit proefschrift richt zich op MPS analyse van STRs. Aangezien forensische DNA databanken zijn opgebouwd met STR data is het logisch om te beginnen met de analyse van deze zelfde markers zodat er nog altijd zoekingen uitgevoerd kunnen worden in de DNA databank.

Basale analyse van MPS STR data:TSSV

Terwijl STRs de gebruikelijke markers zijn voor forensisch DNA onderzoek, worden ze niet heel veel gebruikt in andere DNA onderzoeksgebieden. De initieel beschikbare software pakketten voor MPS data analyse konden dan ook slecht om gaan met herhalende sequenties. In **hoofdstuk 2** van dit proefschrift wordt de (open-source) software TSSV beschreven. Dit was een van de eerste software pakketten vooral gericht op STR analyse. Door het gebruik van twee anker sequenties (aan beiden kanten van de STR) worden markers herkend en vervolgens wordt de waargenomen variatie tussen deze anker sequenties samengevat en worden herhaalde sequenties afgekort. Naast de analyse van STRs bleek deze software ook goed toepasbaar voor de analyse van errors in sequentie-data.

Voor sequentie data is ook sequentie nomenclatuur nodig

Allelen van CE worden benoemd als een nummer wat het aantal herhalingen in

de STR weergeeft. Aangezien het bepalen van de exacte sequentie extra variatie aan toont, is er een nomenclatuur nodig die ook deze variatie kan beschrijven. In **hoofdstuk 3** beschrijven we de eerste aanbevelingen voor een forensische STR sequentie nomenclatuur met het doel om alle relevante sequentie informatie weer te geven in een zo kort en leesbaar mogelijke naam.

Validatie van een (prototype) commerciële methode voor STR sequentie-analyse

In samenwerking met het bedrijf Promega© werd een prototype van een commerciële methode, ontwikkeld voor sequentie-analyse van STRs, getest. Met deze methode werd het DNA van 17 STRs en een marker voor geslachtstypering vermenigvuldigd in een reactie en werd, na verdere voorbereiding, de sequentie bepaald met behulp van de Miseq™ sequencer. Hoofdstuk 4 is een gedetailleerde beschrijving van de analyse van de prestaties van deze methode en de waargenomen sequentie-variantie van elke STR.

Om te berekenen hoe uniek een DNA profiel is, moeten eerst de frequenties vastgesteld worden van elke sequentie-variant in de populatie. Wanneer deze frequenties bekend zijn kan er een statistische kansberekening gebruikt worden om in te schatten hoe groot de kans is dat een willekeurig persoon een bepaald DNA profiel heeft. De STR-sequenties van 297 DNA-monsters van personen uit een Europese, Aziatische en Afrikaanse populatie werden bepaald. Zo werd inzicht verkregen in de extra sequentie-variantie t.o.v. de CE lengte-variantie van de STRs en de frequenties van deze sequentie-varianten. Voor de meeste STRs werden substantieel meer sequentie-varianten waargenomen dan CE lengte varianten.

Bij de analyse van STRs resulteert de PCR in het vormen van zogenaamde stutter-artefacten die veroorzaakt worden door een 'slippend' enzym (Polymerase). STR stutter bemoeilijkt de interpretatie van ongebalanceerde mengsels aangezien pieken van de lager bijdragende donor niet altijd onderscheiden kunnen worden van stutter. De extra sequentie-informatie levert meer inzicht in deze stutter. Het bleek dat de hoeveelheid stutter bepaald wordt door de lengte van de langste ononderbroken herhaling in de STR. Deze informatie gaf nieuwe mogelijkheden voor de interpretatie van gemengde profielen op basis van sequentie-data.

Correctie van stutter in een monster m.b.v. bioinformatica tools

In **hoofdstuk 5** is de bioinformatica software FDS Tools beschreven als de eerste software die werkelijk PCR- en sequentie-artefacten kan corrigeren in MPS data. Op basis van een training dataset wordt systematische ruis vastgesteld en deze informatie wordt vervolgens gebruikt om in een onbekend monster de ruis te corrigeren zodat

de basislijn van de data substantieel wordt gereduceerd.

De software bevat ook tools om kwaliteitscontrole uit te voeren op de referentie data en om analysegrenzen vast te stellen na het uitvoeren van de correctie. Met behulp van deze vastgestelde grenzen kan data van onbekende monsters worden gefilterd en wordt de data weergegeven in een interactieve format.

Analyse van ongebalanceerde mengsels toonde aan dat allelen van de laag bijdragende donor vaker werden teruggevonden na correctie terwijl deze zonder correctie niet herkend konden worden.

Alternatieve forensische markers voor STRs

Ook met de verbeterde analyses m.b.v. bioinformatica tools blijft het duidelijk dat STRs niet voor alle toepassingen ideale forensische markers zijn. Terwijl de hoge mate van variatie helpt om een zo uniek mogelijk profiel te verkrijgen met een beperkt aantal markers zorgen stutter artefacten en lengte-variatie voor beperkingen in de interpretatie van mengsels en de analyse van afgebroken DNA. In **hoofdstuk 6** zijn microhaplotypes geselecteerd uit publiek beschikbare genoom data die vier of meer SNPs (DNA verschillen op een enkele positie) bevatten in een fragment van 70 baseparen. Deze markers kunnen bijna net zo variabel zijn als STRs zonder het nadeel van stutter artefacten en variatie in lengte.

De studie toonde aan dat het merendeel van deze geselecteerde fragmenten voortkwamen uit verkeerd gerapporteerde variatie in de publiek beschikbare genoom data maar er werd toch een subset van 16 microhaplotypes geselecteerd met een discriminerend vermogen dat ongeveer overeen komt met negen STRs. Naast het hoge discriminerende vermogen lijkt de data ook informatie te geven over de biogeografische afkomst van een persoon. Het was namelijk mogelijk om aan de hand van de data de drie geteste populaties vrijwel compleet van elkaar te scheiden.

Conclusies en toekomstperspectieven

In **hoofdstuk 7** wordt het geheel van de uitgevoerde studies besproken in combinatie met de toekomstperspectieven. MPS is nu klaar om toegepast te worden in forensisch zaken en wordt ook al in enkele zaken toegepast. Terwijl MPS nu nog een relatief dure methode is, is het niet onmogelijk dat dit in de nabije toekomst verandert en dat CE geleidelijk vervangen kan worden door MPS wanneer het in een gestroomlijnde geautomatiseerde manier toegepast kan worden. MPS lijkt een veelbelovende tool voor de analyse van mengsels, die mogelijk is door het sequencen van STRs, maar ook door het sequencen van andere markers zoals microhaplotypes. Er zijn veel andere toepassingen voor MPS die niet (of beperkt) mogelijk zijn met CE, vooral door het hogere aantal markers dat in een keer geanalyseerd kan worden. Momenteel

zijn er al toepassingen voor voorspelling van biogeografische herkomst en een aantal uiterlijke kenmerken op basis van SNP. Voor een meer gedetailleerde voorspelling van biogeografische herkomst, maar ook voor veel andere uiterlijke karakteristieken moet nog meer basale kennis verzameld worden voor het toegepast kan worden in zaken. Bij verschillende instituten worden momenteel studies uitgevoerd om deze basale kennis uit te breiden.