Norges miljø- og biovitenskapelige universitet
Fakultet for veterinærmedisin og biovitenskap
Institutt for kjemi, bioteknologi og matvitenskap

Masteroppgave 2014
30 stp

Statistiske beregninger i forbindelse med DNA-basert identifikasjon ved massekatastrofer

Statistical calculations for DNA based disaster victim identification

Julie Johnsen Kjetså

FORORD

Denne masteroppgaven er utført ved Institutt for kjemi, bioteknologi og matvitenskap ved Norges miljø- og biovitenskapelige universitet i perioden oktober 2013 til mai 2014. Oppgaven er den avsluttende delen av master i teknologi (sivilingeniør) - kjemi og bioteknologi. Til tross for et noe dystert overordnet tema, har prosessen vært svært interessant og lærerik.

Jeg vil først og fremst takke min hovedveileder, professor Thore Egeland, for god rettledning og oppfølging under arbeidet. At du har delt din teoretiske og praktiske kompetanse på feltet har gjort dette til et særdeles spennende år. Tusen takk til stipendiat Daniel Kling for all hjelp, gode råd og raske tilbakemeldinger ved arbeidet i Familias.

Til slutt vil jeg takke min familie for støtte, gode ord og inspirasjon til å komme i mål.

Ås, 9. mai 2014

_____

Julie J. Kjetså

SAMMENDRAG

Identifisering av personer etter masseulykker er en gren av familiegenetikk innen forensisk vitenskap som benytter DNA-bevis til å indikere slektskap. Personer kan enten identifiseres ved hjelp av DNA-profil fra den antatt uidentifiserte, eller, som i denne oppgaven, basere seg på beregning av slektskap til antatte referansepersoner. Søsken vil kunne dele mellom null og to alleler, mens en forelder og barn deler ett, sett bort fra mutasjoner, og dette benyttes i beregningene. Hypotesen om at den savnede personen og den antatte referansefamilien er beslektet veies opp mot hypotesen om at de er ubeslektet, og presenteres som en «likelihood ratio».

I denne oppgaven er det undersøkt hvordan ulike scenario påvirker sikkerheten i identifiseringen av ukjente personer opp mot referansepersoner. Referansepersoner og antall markører som benyttes er variert for å studere hvordan resultatet påvirkes, og er gjort for å kunne anta hva som er hensiktsmessig å benytte for å kunne identifisere personer etter masseulykker. For de statistiske beregningene for hvor sannsynlige dataene gitt en hypotese er, er programmet Familias benyttet. Det er studert simulerte data for 100 personer og referansefamilier i hvert scenario. Generelt vil naturligvis et større antall markører og referansepersoner bedre nøyaktigheten av identifiseringen, men bruk av for mange vil kunne være overflødig. Resultatene i oppgaven bekreftet at bruk av begge foreldene som referansepersoner er de foretrukne slektningene til å identifisere savnede personer, hvor gjennomsnittlig LR-verdi for treff mellom foreldre og uidentifisert person med markørsystemet CODIS er $3.47 \cdot 10^6$ ganger større enn ved søk med bror som referanseperson. Flere markører vil også gi en høyere gjennomsnittlig LR-verdi, og dermed øke antall treff med informativ identifisering.

ABSTRACT

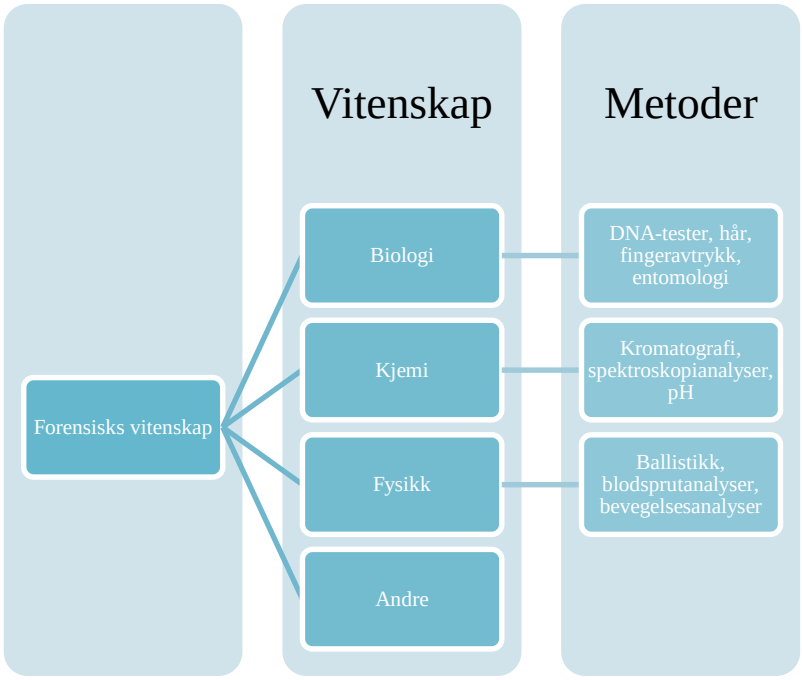
Disaster victim identification is a part of family genetics in forensic science, and to identify missing persons by kinship analysis, DNA evidence is used. Identification of missing persons can be performed either by a direct match with reference samples from personal items, or, as in this thesis, based on calculation of kinship between the missing person and assumed reference relatives. Siblings share between zero and two alleles IBD, while a parent and child will share one, apart from mutations. In the calculations, this and allele frequencies in the population is used. To indicate kinship, the hypothesis that the missing person is related to the reference members is compared to the hypothesis stating that the missing person is unrelated to the known reference members, and the results are presented as a likelihood ratio.

This thesis examined how different scenarios will affect the identification of missing persons by use of reference relatives. The accuracy of the identification has been studied by having a different number of reference relatives and genetic markers, and is done to be able to say what is appropriate so that the identification will be informative. To calculate the probabilities of the data given a hypothesis, the program Familias is used. For each scenario, simulated data for 100 missing persons and reference relatives are studied. In general, more reference relatives will result in a greater accuracy of the identification, but using too many might be unnecessary and too costly. The results of this thesis confirm that using both parents of the missing person are preferred as relatives to kinship analysis. The mean value of LR is $3.47 \cdot 10^6$ times greater for match with parents compared to sibling when using the CODIS genetic marker system. A higher number of genetic markers will achieve a greater mean value of LR, and in that way give a higher number of informative identifications.

Innhold

1 INNLEDNING

«Forensic science» er en samlebetegnelse på vitenskaper som samler og undersøker informasjon i rettslige sammenhenger, og stammer fra det latinske ordet «forēnsis», som betyr «in open court, public». Det finnes ikke et godt tilsvarende ord på norsk, ettersom rettsvitenskap ikke dekker alle de vitenskapene som anvendes rettslig, men den fornorskede versjonen av ordet, forensisk vitenskap, er nå å finne i norsk litteratur.

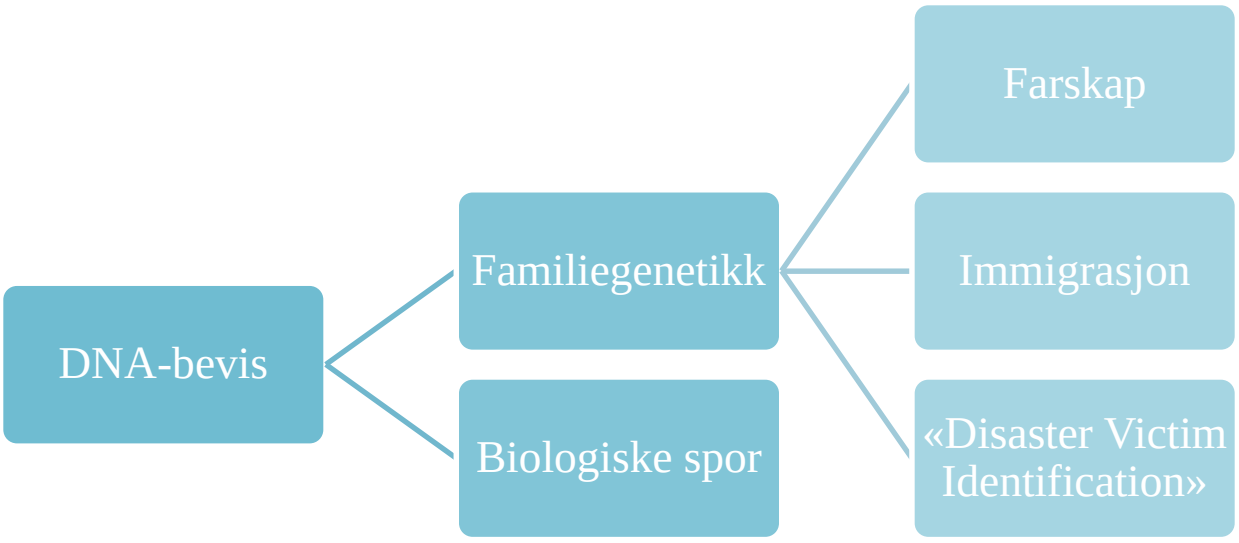


Figur 1.1: Vitenskaper og metoder anvendt i arbeid med forensisk vitenskap.

Innenfor forensisk vitenskap anvendes blant annet kjemi, biologi og fysikk til å gi vitenskapelige bevis i rettslig sammenheng. Figur 1.1 viser disse vitenskapene, og eksempler på hvilke metoder innenfor disse grenene som kan være formålstjenlige i denne sammenheng. Metodene under kjemi kan for eksempel brukes i toksikologi, hvor man kan være interessert i å finne ut om en person er forgiftet eller dopet, analyser i fysikk kan gi en indikasjon på hvordan en transportulykke kan ha forløpt, mens entomologi kan benyttes for å si noe om dødstidspunkt. En kombinasjon av de

relevante grenene vil forhåpentligvis være med på å danne et bilde av hvordan de faktiske forholdene har vært ved spørsmål i rettslig sammenheng.

DNA-bevis og -profiler kan i forensisk vitenskap benyttes i flere typer rettsgenetiske anvendelser, som i all hovedsak deles i to undergrupper; biologiske spor og familiegenetikk. Sporsaker er saker der det er mistanke om at det har skjedd noe kriminelt, som for eksempel voldtekt, drap, innbrudd og annet, mens familiegenetikk favner farskap, familiegjenforening ved immigrasjon og identifisering av personer ved store ulykker. I forbindelse med store ulykker kan noe kriminelt ligge bak, noe som er tilfellet i for eksempel terroraksjoner eller massakre, men til tross for mistanke om noe kriminelt, håndteres arbeidet ofte av familiegenetikk. Dette var for eksempel tilfellet under identifisering av personene som omkom på Utøya og i Regjeringskvartalet 22. juli 2011, der det rettsgenetiske arbeidet ble utført av avdeling for familiegenetikk på Folkehelseinstituttet i Oslo.



Figur 1.2: Forensiske anvendelser av DNA-bevis.

Siden DNA-teknologi i rettsgenetikk ble introdusert internasjonalt på midten av 1980-tallet, har teknologien blitt utviklet til å være et nyttig verktøy for flere bruksområder der identifisering er

involvert. På grunn av biologisk materiales tilgjengelighet og store variasjoner mellom individer, har utarbeidelse av denne teknologien hatt stor betydning i identifiseringsarbeid. For å skape DNA-profiler behøver man kun små mengder DNA, og profiler kan dannes på bakgrunn av DNA fra blod, spytt, sæd og andre celler. Utvikling av DNA-profiler ble første gang benyttet for å identifisere en gjerningsperson i forbindelse med en drapssak, men ble deretter et nyttig verktøy innen familiegenetikk, og har blitt mest brukt for å bestemme farskap. Etter å ha klart å identifisere levende personer med stor suksess, gikk veien videre til å bruke informasjonen i DNA til å identifisere omkomne. Situasjoner som krever en slik identifisering er krigsofre i massegraver, savnede soldater og savnede personer i masseulykker. Første gang DNA var den viktigste metoden for identifisering ved masseulykker, var ved Operafjell-ulykken på Svalbard i 1996 (Olaisen et al. 1997; Rognum 2010).

I forbindelse med store ulykker og katastrofer ønsker man å identifisere individene involvert, for å kartlegge dødsårsakene og for å kaste lys over årsaken til ulykken. Temaet i denne oppgaven er å studere hvordan dette kan gjøres når man har DNA-profiler fra de omkomne og antatte familiemedlemmer.

1.1 Oppgavens oppbygning

I innledningen av denne oppgaven, presenteres først «disaster victim identification» (DVI). Det finnes ingen fullgod norsk betegnelse, men det engelske uttrykket kan oversettes til identifisering ved eller etter massekatastrofer. Videre i oppgaven vil derfor den engelske forkortelsen «DVI» benyttes. Deretter presenteres et rammeverk for hvordan bevis tolkes ved hjelp av den frekventistiske metoden og den bayesianske tilnærmingen. Ettersom dette kan anses som grunnleggende verktøy for å tolke statistiske beregninger, er dette lagt til innledningen som bakgrunn for de videre beregningene i Avsnitt 2 hvor materialer og metoder beskrives. Videre gis det et kort motiverende eksempel for å illustrere tankegangen i DVI, og til slutt hensikten med oppgaven.

I Avsnitt 2.5 presenteres de statistiske beregningene som foretas i forbindelse med DVI, som anses som mer kompliserte og spesifikke beregninger for denne oppgaven. Beregningene gjøres på bakgrunn av teorien og antakelsene beskrevet i Avsnitt 2.1-2.4. På bakgrunn av tilnærmingene presentert i innledningen, vil disse beregningene kunne tolkes i resultat- og diskusjonsdelen. Ved siden av de statistiske beregningene i rettsgenetikk og styrkeberegning av disse, introduseres programmene brukt i oppgaven nærmere, før resultatene presenteres i den tredje delen. Avslutningsvis diskuteres resultatene og hva som kan ha innvirkning på dem, samt hva som kan gjøres videre, før oppgavens konklusjon fattes.

1.2 DNA-identifisering etter massekatastrofer («DVI»)

Tradisjonelt avhenger «DVI» av innsats fra både politi, tannleger og patologer som sammenligner *ante mortem*-prøver fra savnede personer, med *post mortem*-prøver fra avdøde. Det vil i denne oppgaven fokuseres på hvordan DNA-basert informasjon kan anvendes til identifisering av personer ved slike store ulykker der mange personer skal identifiseres, hvor beregningene baseres på DNA-profiler fra uidentifiserte personer og antatte familiemedlemmer.

Massekatastrofer kan hovedsakelig deles i to hovedgrupper; naturkatastrofer og menneskeskapte katastrofer. Naturkatastrofer omfavner jordskjelv, vulkanutbrudd, tsunamier, snøskred og orkaner, mens menneskeskapte involverer transportulykker, terroraksjoner, krig og politiske kriser (Prinz et al. 2007). Avhengig av hva slags type massekatastrofe, vil tilnærmingen i identifiseringsarbeidet være forskjellig. Dette kan avhenge av hvordan *post-morten*-prøvene er tatt, hva slags materiale som blir tatt og i hvilken forfatning avdøde er i, og har påvirkning på DNA-typingens suksessrate, og dermed på hvor sikker konklusjon man kan gi av resultatet. Helst ønsker man å kunne ta prøver fra blod eller bløtvev, da disse er rike på DNA, og derfor vil skape minst utfordringer i utforming av DNA-analyser. Utfordringer ved å få nok DNA-materiale i god nok kvalitet kan oppstå når de avdøde har vært utsatt for ekstremt høy temperatur, kjemiske ødeleggelser eller mikrobiell nedbrytning.

Når en fullstendig DNA-profil er utarbeidet fra den uidentifiserte personen eller levningene, må denne sammenlignes med enten DNA fra en savnet person eller ved hjelp av slektskapsanalyser hvor man har DNA fra antatte referansepersoner. Dersom man har klart å utarbeide en DNA-profil for den savnede personen basert på for eksempel personlige eiendeler som tannbørste, barberhøvel, undertøy, hårbørste eller andre prøver man tror inneholder den savnedes DNA, eller om personen allerede er å finne i en nasjonal DNA-database, vil det bli foretatt et direkte søk. Basert på hvor sannsynlig det er at den uidentifiserte og den savnede personen er samme person, beregnes det er «likelihood ratio», som vil beskrives ytterligere i avsnitt 2.2.2. Videre i oppgaven vil «likelihood ratio» bli omtalt ved forkortelsen «LR».

I denne oppgaven vil det ses nærmere på identifisering av personer der man ikke har disse surrogat-DNA-prøvene som nevnt over, men der identifiseringen vil basere seg på beregning av slektskap fra antatte referansepersoner. Personlige eiendeler benyttet til å lage referanseprøver kan være forurenset, for eksempel kan en annen enn den savnede ha brukt tannbørste, barberblader og lignende. Derfor vil DNA-profiler fra antatte referansepersoner kunne være en mer pålitelig kilde i identifiseringsberegninger. En DNA-profil fra slektninger av den savnede er kjent, og disse er i denne oppgaven simulert i Familias 3 (Mostad et al. 2013). I denne oppgaven benyttes Familias 3 med modulene «Blind search» og «DVI», og vil heretter omtales som Familias.

DNA-identifisering etter massekatastrofer kan deles inn i to kategorier, åpent eller lukket. En åpen massekatastrofe er en hendelse der man ikke har registre eller data tilgjengelig for gruppen mennesker involvert. Dette gjør det vanskelig å finne det sanne antallet ofre etter en katastrofe. Ved et lukket system er problemet å finne treff mellom et kjent antall avdøde og savnede personer i en identifisert gruppe. Et eksempel på dette kan være ved et flystyrt, hvor man har en kjent passasjerliste. Som regel vil det i en lukket katastrofe være lettere å få sammenlignbare *ante mortem*-prøver raskere (Interpol 2009).

1.3 Den frekventistiske metoden

Det finnes flere måter å tolke DNA-bevis på, og derfor også hvordan bevisene skal presenteres for retten på best mulig måte. Å beskrive statistiske konsepter som beskriver de tolkede dataene i retten på en forståelig måte for dommer og jury kan være vanskelig, og det finnes både fordeler og ulemper ved bruk av de forskjellige metodene. Det vil i denne oppgaven bli gitt vist hvordan to av disse metodene kan anvendes til tolkning av data i forhold til DVI.

En frekventistisk tilnærming i forensisk vitenskap er relatert til, men ikke identisk med den frekventistiske tilnærmingen i sannsynlighetsteori (Buckleton et al. 2005). I den frekventistiske metoden anvendt i forensisk vitenskap, finner man bevis mot en hypotese ved å vise at dataene er lite sannsynlige hvis hypotesen antas sann, og dermed støtter den alternative hypotesen. Det vil si at jo mindre sannsynlige dataene er under hypotesen, jo større sannsynlig er alternativet, og at et tilfeldig treff er usannsynlig. I denne oppgaven vil ikke nullhypotese og alternativ hypotese benyttes, men H_p og H_d som tradisjonelt står for de engelske ordene «prosecution» og «defense».

I det motiverende eksempelet, som forestilles i avsnitt 1.5, presenteres 3 savnede personer og 2 referansepersoner. For å teste om det første offeret V1 tilhører familie F2, må det formuleres to hypoteser, hvor den første er

$H_\mathrm{p}: Offeret, V1, tilhører\ familie\ F2.$

Deretter beregnes sannsynligheten for dataene dersom hypotesen er sann, hvor dataene i dette tilfellet er DNA fra den antatte familien, her kalt data. Sannsynligheten for dataene hvis nullhypotesen er sann, blir da

$$P(data|H_\mathrm{p})$$

På samme måte beregnes sannsynligheten for dataene gitt en annen hypotese, i dette tilfellet

$H_{\mathrm{d}}: Offeret, V1, er\ ikke\ beslektet\ til\ familie\ F2.$

LR er en sannsynlighetskvote for dataene gitt to hypoteser. I forbindelse med DVI, beregnes ratioen ut ifra hvor sannsynlig dataene er under hypotesen H_p i forhold til H_d:

$$LR = \frac{P(\mathrm{data}|H_{\mathrm{p}})}{P(\mathrm{data}|H_{\mathrm{d}})}$$

Dersom LR-verdien er høy, antas det at bevisene støtter at H_d er sann, og at H_d er usann, altså at dataene er mer sannsynlige gitt H_p enn H_d. Avhengig av hvor sannsynlige dataene er under en hypotese i forhold til en annen, vil man kunne si noe om hvor sterk støtte det er for hver av hypotesene.

En annen måte å bruke den frekventistiske metoden, er å se på sannsynligheten for å ekskludere en tilfeldig person. Dersom det motiverende eksempelet benyttes til å illustrere dette, vil V1 ikke ekskluderes fra å kunne tilhøre familie F2, men en sannsynlighet for at en tilfeldig person blir avvist beregnes. Ut fra dette vil det være usannsynlig at offeret er en tilfeldig person dersom sannsynligheten for å utelate tilfeldige personer fra familien er høy nok. Det blir ikke gått nærmere inn på sannsynligheten for å ekskludere en tilfeldig person i denne oppgaven (Buckleton et al. 2005).

I denne oppgaven benyttes den frekventistiske metoden til å fremlegge databevis i form av LR, noe som gir enklere beregninger enn ved den bayesianske metoden som omtales i Avsnitt 1.4. Fordelen ved bruk av den frekventistiske metoden er at det gir lettere beregninger, men kan være vanskeligere å tolke og fremlegge i retten enn ved bruk av den bayesianske metoden.

1.4 Den bayesianske metoden

Bayes teorem gir en matematisk regel for å kunne utnytte informasjon fra erfaring og observasjon til å finne et estimat, ved at den gir *a posteriori*-sannsynligheter til de gitte hypotesene (Buckleton et al. 2005).

Bayes teorem er utledet ved hjelp av lovene for sannsynlighet, og kan skrives med ord som

$$a\ posteriori - \text{sannsynlighet} = LR \cdot a\ priori - \text{sannsynlighet}$$

A priori-sannsynligheten defineres videre som

$$P(H_i) = \pi_i, i = 1, \dots, I$$

I det motiverende eksempelet, som presenteres i Avsnitt 1.4, vil *a priori*-sannsynligheten være sannsynligheten for hypotesen med at offeret V1 er beslektet til familien F2, mens LR vil være sannsynlighetsratioen for data gitt to hypoteser. LR vil da formuleres på samme måte som i den frekventistiske tilnærmingen.

$$P(data|H_i) = L_i$$

LR sier noe hvor mange ganger mer sannsynlig dataene er gitt en hypotese, i forhold til en annen. Ved å beskrive *a posteriori*-sannsynligheten i det første, motiverende eksempelet med ord, vil den være den være sannsynligheten for hypotesen om at offeret V1 er beslektet til familien F2 gitt DNA-bevisene. Bayes teorem vil dermed uttrykkes ved

$$P(H_i|data) = \frac{P(data|H_i) \cdot \pi_i}{\sum_j P\left(data\middle|H_j\right) \cdot \pi_j} = \frac{L_i \pi_i}{\sum L_j \, \pi_j}$$

hvor π er *a priori*-sannsynliheten. Dette uttrykket kan forenkles dersom *a priori*-sannsynligheten er lik for alle hypotesene, slik at $\pi_1 = \ldots = \pi_I$.

$$P(H_i|data) = \frac{L_i}{L_i + \cdots + L_I}$$

$$= \frac{LR_i}{L_1 + \cdots + LR_{I-1} + 1}$$

der $LR_i = L_i/L_I$. Dette er en generell versjon av Bayes teorem når man antar flat *a priori*-fordeling.

Anvendt konkret på det motiverende eksempelet, finner man

$$P(V_1 \in F_1|data) = \frac{\pi_1 P(data|V_1 \in F_1)}{\pi_1 P(data|V_1 \in F_1) + \pi_2 P(data|V_1 \in F_2) + \pi_3 P(data|V_1 \in F_3)}$$

Med like *a priori*-sannsynligheter, det vil si

$$\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$$

blir

$$P(V_1 \in F_1|data) = \frac{LR_1}{LR_1 + LR_2 + 1} \qquad \textbf{Formel 1.1}$$

Dersom LR_2 er tilnærmet lik null, og LR_1 er stor, vil *a posteriori*-sannsynligheten for at offeret V1 tilhører familie F1, være tilnærmet 1.

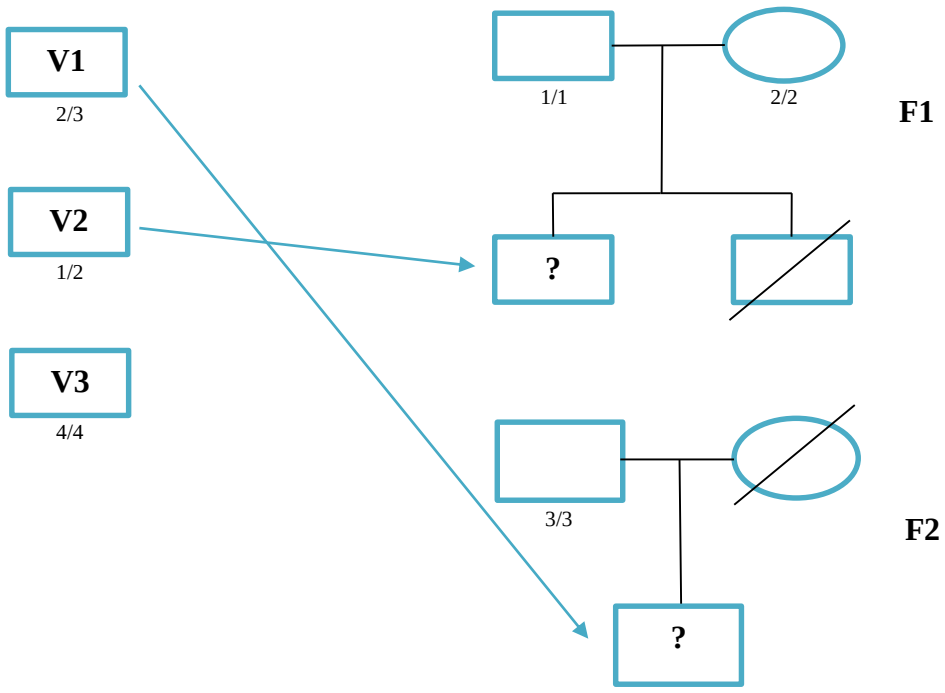
Felles for både den frekventistiske og den bayesisanske metoden er at det beregnes en LR-verdi, som beskrives nærmere i tilknytning til identifisering av personer ved massekatastrofer i Avsnitt 2.5.2. I den bayesianske metoden bestemmer man i tillegg en *a priori*-sannsynlighet, for å kunne

beregne en *a posteriori*-sannsynlighet. For å kunne bestemme en *a priori*-sannsynlighet vil man ta i bruk kjent informasjon, for eksempel anslått alder ut ifra obduksjon.

A priori-fordelingen er vanskelig å fastsette, og kreves for å beregne *a posteriori*-sannsynligheten. I programmet Familias, hvor de statistiske beregningene i denne oppgaven er gjort, vises, i tillegg til LR, *a posteriori*-sannsynlighetene for hver beregning. Disse er basert på en flat *a priori*-fordeling i DVI-anvendelsene, ettersom det er antatt at det ikke finnes noen erfaring eller informasjon som tilsier noe annet. Når man skal presentere data for medlemmer av retten uten statistisk bakgrunn, vil det kunne være enklere å presentere dataene ved hjelp av *a posteriori*-sannsynligheten som angir sannsynligheten for en hypotese gitt dataene, da det gir en lettere fortolkning. Man kan må den måten forklare *a posteriori*-sannsynligheten enten med prosent eller på odds-form, som gir en mer verbal forklaring på bevis og hypotesene. En innarbeiding av denne forhåndsinformasjonen vil ikke nødvendigvis være påkrevd da DNA-bevis står veldig sterkt, men kan være hendig ved dårlig DNA-kvalitet eller –kvantitet.

1.5 Motiverende eksempel

Hensikten med dette eksempelet er å enkelt illustrere prinsippene i arbeid med DVI. Når man jobber med DVI, ønsker man å plassere ofrene til riktig familie. I et lite eksempel kan man se på tre ofre og to referansefamilier, med kun en markør.



Figur 1.3: Illustrasjon av familietrærne som benyttes i det motiverende eksempelet.

Figur 1.3 illustrerer de 3 savnede personene og deres alleler for en markør, samt to referansefamilier. Pilene illustrerer en mulig løsning, der det første offeret, V1, tilhører familie F2, det andre offeret, V2, tilhører familie F1, mens offer nummer tre, V3, ikke har noen registrert familie, og at det derfor ikke finnes noe grunnlag for å kunne identifisere ham. Ved beregninger i Familias vil dette undersøkes nærmere. Programmet presenteres nærmere i Avsnitt 2.7.

I denne oppgaven benyttes den marginale tilnærmingen som er implementert i Familias. Der vil si at man tester et individ mot alle familiene. I dette eksempelet er det 3 uidentifiserte personer, og i tillegg er det en sannsynlighet for at en ukjent person er en del av hvert enkelt familietre. Dette gjør at *a priori*-sannsynlighet for hver beregning er 0.25.

I Familias vil det i hver beregning settes opp to hypoteser, hvor en av disse indikerer at den savnede personen tilhører en bestemt referansefamilie, mot hypotesen om at denne personen ikke tilhører dette familietreet. For å beregne LR for at den uidentifiserte personen V2 tilhører familien F1, vil følgende hypoteser settes opp:

H_p: *Den uidentifiserte personen V2 tilhører F1*

H_d: *Den uidentifiserte personen V2 tilhører ikke dette familietreet*

I dette eksempelet er allelfrekvensen for hver av de fire allelene 0.25, de forekommer altså like ofte i populasjonen. Dette motiverende eksempelet er gjort uten mutasjonsmodell, men innvirkning av mutasjonsmodell beskrives nærmere i Avsnitt 2.4. Beregningene gjort i Familias gjøres basert på fremgangsmåten som beskrives i Vedlegg 4. Resultatet av beregningene er som følger:

DVI module - Results

Project name is: SmallDviExample2

Family id	Unidentified person	Prior	Posterior	LR
F1	Reference pedig...	0.25	0.111111	1
F1	V1 (Son)	0.25	0	0
F1	V2 (Son)	0.25	0.888889	8
F1	V3 (Son)	0.25	0	0
F2	Reference pedig...	0.25	0.333333	1
F2	V1 (SonF2)	0.25	0.666667	2
F2	V2 (SonF2)	0.25	0	0
F2	V3 (SonF2)	0.25	0	0

Figur 1.4: Utskrift av resultatene i Familias.

I linje 4, hvor hypotesene H_p og H_d testes for den uidentifiserte V2 og familien F1, beregnes

$$LR = \frac{P(data|V_2 \text{ er i familie } F1)}{P(data|V_2 \text{ er ubeslektet})} = \frac{1}{2p_2p_3} = \frac{1}{2\left(\frac{1}{4}\right)\left(\frac{1}{4}\right)} = 8$$

LR-verdien angir dataene som 8 ganger mer sannsynlig under hypotesen om at V2 tilhører F1, i forhold til hypotesen om at V2 ikke tilhører dette familietreet. Et alternativ til LR-verdien er, som nevnt, *a posteriori*-sannsynligheten. I dette eksempelet er *a priori*-sannsynligheten lik for alle

familietrærne. *A posteriori*-sannsynligheten kan dermed beregnes som LR for alternativet dividert på summen av alle LR for samme familietre multiplisert med *a priori*-sannsynligheten. Ettersom det er en sannsynlighet for at en ukjent er del av familietreet, vil resultatene vise LR-verdi for en mer enn antall uidentifiserte personer. Basert på Formel 1.1, blir *a posteriori*-sannsynligheten

$$P(V2\ er\ i\ familie\ F1|data) = \frac{8}{1+0+8+0} = 0.889$$

1.6 Hensikten

Hensikten med denne oppgaven er å identifisere personer ved store ulykker ved hjelp av DNA-profiler fra de omkomne og antatte familiemedlemmer. Ettersom det ikke finnes noe grunnlag for å sette *a priori*-sannsynligheten, er den den samme for alle beregningene av familietrær. For enklere beregninger vil LR-verdien vil benyttes til å presentere resultatene. Dersom en omkommet person tilhører en av referansefamiliene, vil man anta at DNA-beviset er sterkt, og LR-verdien blir typisk relativt høy. Ulike faktorer kan påvirke hvor høy denne LR-verdien blir, altså hvor sikker man kan være på at denne personen tilhører en familie. Disse faktorene kan være antall markører, hvor mange personer som var i ulykken og som skal identifiseres, samt hvor mange referansepersoner man har DNA-profilen til i familien, og hvordan disse er beslektet.

For beregninger og til å simulere data i denne oppgaven, brukes Familias, og en introduksjon av programmet er å finne under avsnitt 2.4 i materiale og metoder, mens en fullstendig fremgangsmåte av simulering og bruk av Familias finnes i Vedlegg 2, 3 og 4. I oppgaven blir det brukt ulike eksempler for å illustrere både hvordan sannsynligheter beregnes ved å gå i dybden på enkelte tilfeller med få personer, og generelt i større saker med mange involverte og flere faktorer. Basert på disse ulike scenarioene, vil det ses nærmere på hvordan LR påvirkes, og hvor gode beregningene i programmet er til å indikere informativ identifisering ut ifra slektskapene.

2 Materiale og metoder. Teori

2.1 Genetiske markører

Likhet mellom foreldre og avkom både blant mennesker, dyr og planter har alltid vært åpenlyst, men alle individers genomer, med unntak av eneggede tvillinger, er unike, og kan derfor brukes som en unik personlig identifisering. Hver enkeltes genom er en kombinasjon av foreldrenes kromosomer, og kan på den måten anvendes til å indikere familierelasjoner, for eksempel i farskapssaker eller identifisering ved masseulykker.

Til tross for at det stadig utvikles nye metoder og maskiner for å øke hastigheten og minke kostnadene for å sekvensere hele genomet, er dette unødvendig arbeid ved identifisering av personer, da det størsteparten av genomet er likt blant mennesker. Det vil derfor være optimalt å finne områder i genomet som har høy variasjon og som skiller dermed individer fra hverandre, slik at sannsynligheten for at to tilfeldige personer har identiske DNA-profiler blir minst mulig. Loci brukt som markører for å utvikle en DNA-profiler består av områder som ikke er protein-kodende, men innehar egenskapen at de er polymorfe, noe som vil si at det finnes flere varianter av DNA-sekvensen i et locus (Buckleton et al. 2005).

Man finner størsteparten av DNA i en celle i cellekjernen hvor det er pakket i menneskets 46 kromosomer i det fleste cellene, og dette kalles kjerne-DNA, men noe komplementært DNA er også å finne i mitokondriene. DNA funnet her behandles annerledes i forensisk genetikk. Forskjellige individer har forskjellige mønster i DNA, og i det ikke-kodende delen av DNA, finner man områder hvor 2-4 nukleotider repeteres. Antall repetisjoner er forskjellig i hvert allel, og områdene er svært polymorfe. Slike variable «tandems repeats» som er flankert av samme gjenkjennelsessete for samme restrikssjonsenzym, som dermed vil gi fragmenter med ulik lengde mellom individer og kan separeres på en gel (Lesk 2012).

I denne oppgaven blir simulerte «short tandem repeats»-markører benyttet, ettersom profiler basert

på disse er den mest brukte metoden for genetisk identifisering. Allelfrekvensene og hvilke markører som er benyttet i hvert av systemene fremgår av Vedlegg 1. Det nevnes også andre alternativer, som kan brukes som supplerende analyser, men dette er ikke gjort videre i denne oppgaven.

2.2 STR

«Short tandem repeats» (STR) loci, består av repeterte segmenter med lengde på to til åtte baser. Vanligvis er den repeterte sekvensen 4-5 basepar lang, og finnes på mange ulike steder blant intron-områder i kromosomene, hvor de varierer i lengde og hvilken sekvens som repeteres.

Når man skal velge STR loci, spiller flere faktorer inn. Det er ønskelig med en stor variabilitet innen et locus, slik at man har en lav sannsynlighet for tilfeldig treff. De fleste systemene anvendt i dag gir en estimert sannsynlighet for et tilfeldig treff på mellom 1 av 10^{10} og 1 av 10^{20}, noe som avhenger av antall markører brukt og allelfrekvenser i relevant populasjon (NCBI, 05.03.14). Lengden på allelene bør være mellom 90-500 bp, da mindre alleler degraderes i mindre grad, og det er derfor mindre sannsynlig at de dropper ut. Dette gir typisk bedre presisjon i målingene enn ved høyere molekylær vekt (Buckleton et al. 2005). Antall alleler i et forensisk relevant STR-loci er vanligvis mellom 5 og 20 alleler.

STR-profiler er den mest brukte metoden ved genetisk indentifisering i de fleste sammenhenger, og ble for eksempel brukt i identifisering av ofrene både ved den etniske rensingen i Jugoslavia og angrepet 11. september 2001. Flesteparten av STR-markørene befinner seg på separate kromosomer, eller med en avstand på minst 25 Mb dersom de befinner seg på samme kromosom. Det er viktig med en viss avstand mellom markørene, ettersom to loci som ligger nært hverandre har en tendens til å nedarves sammen (Buckleton et al. 2005).

«Polymerase chain reaction» (PCR) brukes til å analysere STR, der primere designet for spesifikke

sekvenser på en av sidene av den repeterte sekvensen benyttes. Disse fluorescerende DNA-fragmentene med varierende lengde vil videre, avhengig av antall repeterte segmenter, separeres ved elektroforese og detekteres ved hjelp av et kamerasystem, som dermed kan benyttes til å lage en genetisk profil, unik for et individ (Fletcher et al. 2007).

De tidligste muliplexene baserte seg på kun noen få STR loci, noe som gjorde at sannsynligheten for tilfeldige treff ble høy sammenlignet med moderne standarder. Det vil i denne oppgaven bli benyttet to ulike systemer, som omtales i Avsnitt 2.2.1 og 2.2.2. De to systemet er valgt for å illustrere hvilken effekt antall markører som brukes ved identifisering av personer i Familias har.

2.2.1 «Combined DNA index system» (CODIS)

CODIS består av 13 STR markører, og inneholder i tillegg Amelogenin (AMEL) for å bestemme kjønn. Systemet var et resultat av standardisering i Canada og USA, og har siden 1997 blitt brukt av FBI i kriminalsaker. Blant de 13 markørene i CODIS finner man syv som også brukes i systemene som benyttes i Europa. Det finnes forskjellige allelfrekvenser for forskjellige populasjoner, men i denne oppgaven er data fra en kaukasisk populasjon benyttet. Allelfrekvensene er de samme som benyttet av Ge et al. (2011), og er fått på forespørsel fra PhD Andreas Tillmar ved «Rättsmedicinalverket», Linköping.

2.2.2 «Second-generation multiplex» (SGM)

SGM ble introdusert i 1995, og er et seks-locus STR system kombinert med Amelogenin kjønnstest. Ved å utvide det eldre systemet med to markører, ble sannsynligheten for tilfeldige treff senket, og SGM ble tatt i bruk under utviklingen av de nasjonale databasene i U.K. og New Zealand. For enkelhetens skyld, er amelogenin-kjønnstesten utelatt i forsøkene gjort i denne oppgaven, ettersom de resterende markørene også inngår i CODIS databasen. Kjønnstesten benyttes derfor ikke i denne oppgaven, men allelene er med i det innlastede datasettet for å indikere kjønn. SGM inneholder kun en markør mindre enn det som er utviklet av ENFSI, som er et

standardsystem utviklet for å kunne ha kunne møte utfordringene ved kriminalitet på tvers av landegrensene.

2.3 Andre genetiske markører

Foruten STR, kan andre markører benyttes til identifisering av personer, og disse presenteres kort her. I tilfeller der kvantitet og/eller kvaliteten på STR-markørene ikke er god nok til å indikere informativ identifisering, kan andre markører kunne gi mer informasjon.

2.3.1 SNP

«Single nucleotide polymorphisms» (SNP-er) er mutasjoner på et basepar som viser variasjon innen en populasjon. Blant det 3 milliarder store genomet, finnes det en slik mutasjon i gjennomsnitt per 300 bp, og det er derfor et mangfold av basepar som kan analyseres (Fletcher et al. 2007).

Ettersom SNP-er er mindre i størrelse enn STR, vil det være en fordel å ta i bruk SNP-er i identifsering av personer ved masseulykker hvor DNA-fragmentene i prøvene er sterkt nedbrutt. SNP-er kan bli amplifisert helt ned i lengder på under 60 bp, men for å oppnå en sikker konklusjon på nivå med et 13 STR loci-system trenger man et stort antall SNP-er. Etter at «high-throughput»-teknologi kom på markedet, har det blitt mer kostnadseffektivt og mulig å standardisere (Zietkiewicz et al. 2011).

Det er lav mutasjonsrate i SNP-er, noe som gjør de til stabile markører. For å kunne avgjøre hvilke SNP-er som er egnet til identifisering må man ta i betraktning tekniske og statistiske sider ved SNP-analyser, og samtidig se på frekvensen av den enkelte SNP i populasjonen.

2.3.2 Mitokondrielt DNA (mtDNA)

Til tross for at man finner mesteparten av DNAet i cellekjernen, finner man også komplementært DNA i mitokondriene. mtDNA er nedarvet ved en annen mekanisme, og behandles derfor annerledes i forensisk vitenskap. Ettersom spermmitokondiret ødelegges når det entrer det fertiliserte egget, vil DNAet være nedarvet kun fra mor. Det gjør det mulig å spore mors avstamming over mange generasjoner da det ikke forkommer noen rekombinasjon. mtDNA vil derfor ikke være unikt for hvert individ, ettersom alle slektninger med samme morslinje vil ha identisk mtDNA-sekvens. På grunn av høy mutasjonsrate og manglende reparasjonsmekanismer vil ikke mtDNA være likt mellom populasjoner (Fletcher et al. 2007). Det er derfor fordelaktig i situasjoner der det er få referansepersoner, og der disse er mer enn en generasjon unna personen som skal identifiseres, da informasjonen i mtDNA kan gi en forbindelse der man mangler data i slektstrær. Disse faktorene gjør at mtDNA brukes stadig i masseulykker.

En annen fordel ved å bruke mtDNA-typing fremfor kjerne-DNA, er det store antallet mtDNA i hver celle, som videre fører til større sjanse for å finne egnet templat-DNA. I arbeid med DVI kan tilstanden til personene involvert i ulykken, og dermed templat-DNA, være degradert grunnet for eksempel høy temperatur, kontaminering med for eksempel jord, forråtnelsesprosesser eller mikrobiologiske angrep, og valg av templat-DNA kan være avgjørende for å få et pålitelig resultat.

Gitt at man har utarbeidet en mtDNA-profil basert på bevismaterialet og dette gir en treff med referansematerialet, vil den statistiske tolkningen av hvor signifikant en treff er avhenge av frekvensen til den enkelte kombinasjonen av tilstøtende alleler som nedarves sammen, altså haplotypen. Dersom haplotypen er sjelden, vil sannsynligheten for at prøvene er nedarvet fra samme morslinje være høy (Zietkiewicz et al. 2011). Ved å benytte seg av et område med en kjent funksjon, antas det at dette området med høy sannsynlighet undergår større selektivt press enn områder med antatt ukjent funksjon, og derfor er mer polymorft (Buckleton et al. 2005).

2.3.3 Y-STR

Dersom bruk av standard, autosomal STR-markører ikke gir tilstrekkelig informasjon, kan Y-kromosomale STR (Y-STR) benyttes for å finne treff mellom mannlige slektninger (Prinz et al. 2007). Y-STR er svært polymorfe, og videreføres fra generasjon til generasjon med få forandringer, ettersom det i rundt 95% av Y-kromosomets lengde ikke foregår noen overkrysning mellom X og Y.

I identifiseringssaker hvor nære slektninger ikke er tilgjengelige eller begrenset, vil en LR-verdi basert på fjerne slektninger muligens ikke være tilstrekkelig til å kunne gi en informativ identifisering. I disse tilfellene kan markører som Y-STR, som avhenger av fars avstamming og er overført fra en generasjon til neste med lite mutasjoner og overkrysning, kunne brukes for å øke LR-verdien, på samme måte som mtDNA fra mors avstamming (Ge et al. 2011). Man kan på den måten for eksempel avkrefte at to menn er brødre.

2.4 Mutasjoner

En mutasjon er en forandring i DNA-sekvensen, som vanligvis er forårsaket av en feil under DNA-replikasjonen i meiosen. Mutasjonene er enten er enkel substitusjon av et nukleotid eller at et eller flere nukleotider fjernes eller legges til DNA-sekvensen. STR-locus er ofte utsatt for mutasjoner sammenlignet med kodende og ikke-kodende områder som ikke er repetitive i genomet, og å observere en mutasjon mellom en savnet person og referansepersoner vil kunne føre til at slektskap ikke indikeres. Et av kravene til STR som bruk av markør er at locus er, som tidligere nevnt, polymorfe, og dette har oppstått som følge av høy mutasjonsrate i områdene (Buckleton et al. 2005). Det er derfor viktig å ta hensyn til mutasjonene når slektskap undersøkes.

I denne oppgaven gjøres simuleringene og beregningene uten mutasjoner, for å kunne sammenligne med eksisterende rapporter på området og for enkelhets skyld, men vil ha innvirkning på beregninger med virkelige data. Uten mutasjoner vil en far antas å være ubeslektet dersom de

likevel deler 12 av 13 alleler, noe som er urimelig. Det er derfor designet ulike mutasjonsmodeller som gir forskjellige mutasjonsmatriser, og som også kan angis i ved beregninger i Familias.

I Familias finnes det fire ulike mutasjonsmodeller som vektlegger mutasjonsrate og størrelsen på mutasjonen forskjellig. En modell er stasjonær dersom allelfrekvensene ikke endres fra en generasjon til neste, og dermed påvirkes ikke beregningen dersom irrelevante personer legges til. Mutasjonsmodellene i Familias er

Modell 1. Uniform: Lik sannsynlighet for mutasjon for allel alleler, med mutasjonsrate R, og tar ikke hensyn til avstand mellom allelene.

Modell 2. Proporsjonal: Sannsynligheten for mutasjon er proporsjonal til frekvensen av allelet det muterer til og tar ikke hensyn til avstand mellom allelene. Stasjonær.

Modell 3. Synkende (equal): Tar hensyn til hvor stor avstand det er mellom allelene før og etter mutasjon, da alleler oftere muterer til nærliggende alleler.

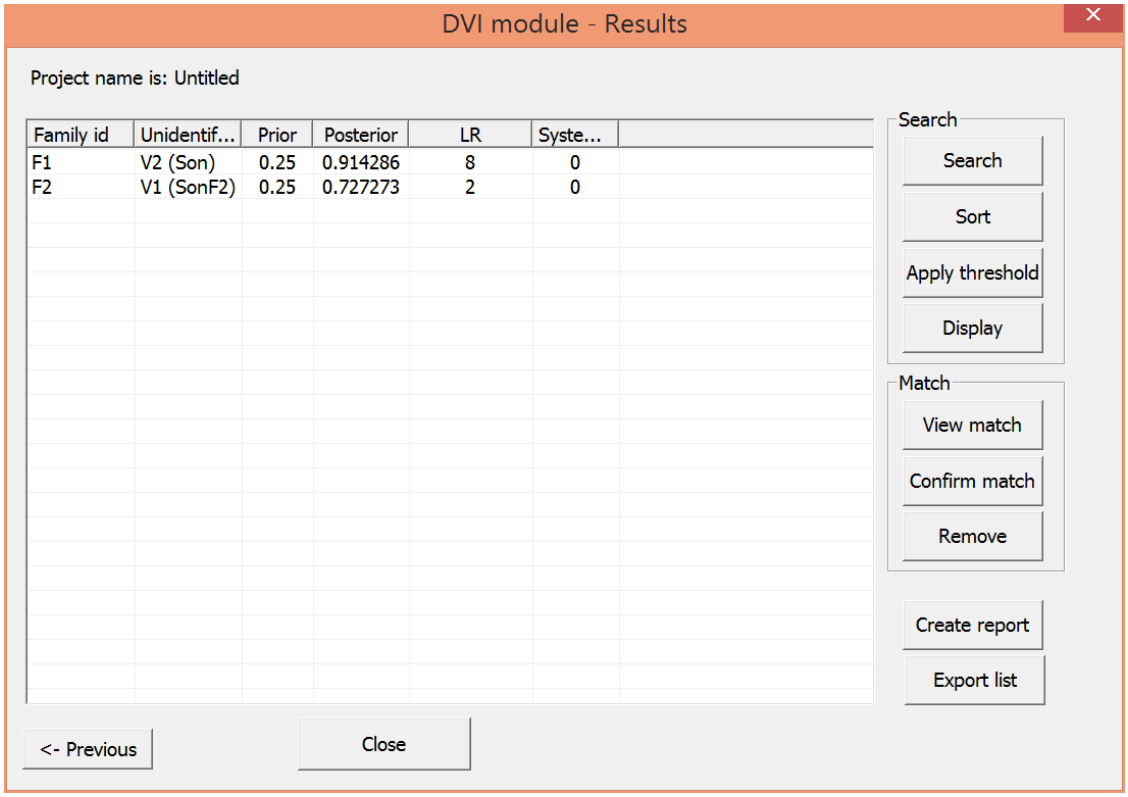
Modell 4. Synkende (stable): Samme som modell 3, men stasjonær.

Hver av modellene beskriver en mutasjonsmatrise som angir sannsynligheten m_{ij} for en mutasjon fra allel *i* til allel *j*. Modell 1 beskriver den enkleste mutasjonsmatrisen for et system med tre alleler:

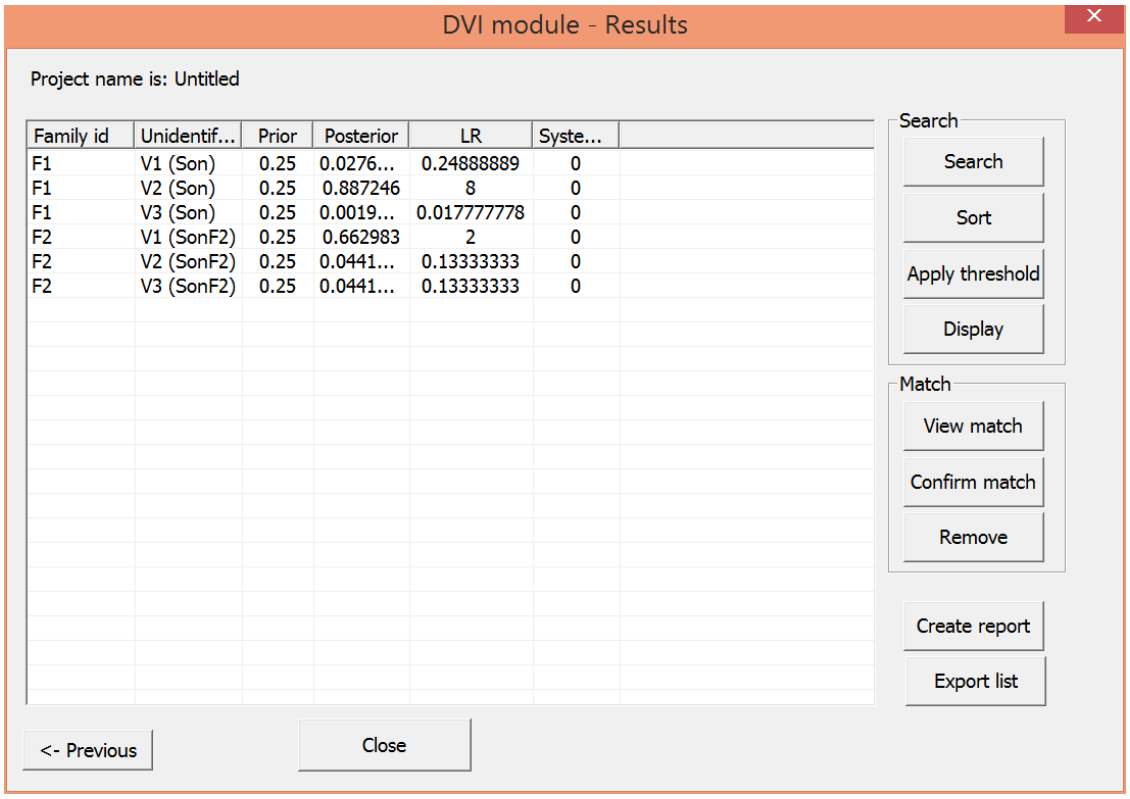
$$M = \begin{bmatrix} 1-R & \frac{R}{N-1} & \frac{R}{N-1} \\ \frac{R}{N-1} & 1-R & \frac{R}{N-1} \\ \frac{R}{N-1} & \frac{R}{N-1} & 1-R \end{bmatrix}$$

hvor R er mutasjonsraten og N=3 antall alleler. Sannsynligheten for at allel 1 forblir allel 1, blir derfor $1-R$, mens sannsynligheten for at allel 1 muterer til allel 2, altså m_{12}, angis i første rad, andre kolonne som $\frac{R}{N-1}$ (Berggreen 2013).

For å illustrere mutasjonsmatrisenes påvirkning på resultatet i en DVI-situasjon, benyttes det motiverende eksempelet fra Avsnitt 1.5. Resultatet av søket vises igjen i Figur 2.1, mens samme søk med angitt mutasjonsmodell 2 og mutasjonsrate 0.1 vises i Figur 2.2.



Figur 2.1: Resultat av motiverende eksempel uten mutasjonsmodell.



Figur 2.2: Resultat av motiverende eksempel med mutasjonsmodell 2.

Av resultatene etter søk i DVI-modulen, kan det ses at ved bruk av mutasjonsmodell vil ingen av treffene gi en LR-verdi på 0, noe som er tilfellet uten mutasjonsmodell. I et tilfelle hvor for eksempel far og sønn deler 12 av 13 alleler, vil derfor LR-verdien kunne være høy nok til å indikere slektskap.

I Figur 2.2 er det også verdt å legge merke til at *a posteriori*-sannsynlighetene ikke summerer seg til 1, men til 0.917. Dette betyr at det er en sannsynlighet på 0.083 for at en ukjent person tilhører familietreet F1 gitt dataene. Med en slik marginal metode, kan man prinsipielt få et scenario der det er mer enn 50% sannsynlighet for at en person tilhører F1, og samtidig mer enn 50% sannsynlighet for at den samme personen tilhører F2. Dette kan virke inkonsistent, og diskuteres videre i Avsnitt 4.10.

2.5 Statistiske beregninger i forensisk genetikk

Innledningsvis ble to måter å tolke DNA-bevis på presentert, den frekventistiske og den bayesianske metoden, der det også ble vist noen eksempler på beregninger. Videre i denne oppgaven blir antakelser presisert, og det gis grundigere eksempler på beregninger av slektskap i forbindelse med DNA-basert identifisering. Ettersom resultatene i denne oppgaven presentes av LR-verdien, vil hovedfokuset i de statistiske beregningene basere seg på dette.

2.5.1 Hardy-Weinberg lov

Hardy og Weinberg utarbeidet, uavhengig av hverandre, en beregning av genotypefrekvenser fra allelfrekvensene. De viste dermed at en likevekt i genotypefrekvenser vil oppstå etter en generasjon med tilfeldig parring, hvor populasjonen er ubegrenset og uten forstyrrende krefter som seleksjon, migrasjon eller mutasjon. De forventede frekvensene for genotypene 1/1, 1/2 og 2/2 er henholdsvis

$$p^2, 2pq \; og \; q^2$$

hvor

$$p^2 + 2pq + q^2 = 1$$

i et locus med kun to alleler. Tilsvarende uttrykk gjelder for locus med flere alleler. Det er uavhengighet for en markør, og i tillegg antas det uavhengighet mellom markørene. Dersom betingelsene for denne ideelle populasjonen opprettholdes, vil genotypefrekvensene holdes konstante over flere generasjoner. Til tross for at betingelsene ikke er sann for alle populasjoner, vil det være mulig å modellere avvik. Modellen kan dermed benyttes for å kunne beregne sannsynligheter for alle genotyper, gitt allelfrekvensene i populasjonen. I denne oppgaven benyttes det kaukasiske datasettet med allelfrekvenser for å beregne genotypesannsynlighetene (Buckleton et al. 2005; Fletcher et al. 2007).

2.5.2 «Likelihood-ratio» (LR)

«Likelihood ratio» har norske oversettelser som sannsynlighetskvote og sannsynlighetsbrøk, men forkortelsen LR brukes, som tidligere nevnt, i denne oppgaven. LR er et forholdstall som beregnes basert på sannsynligheten for DNA-bevis under to hypoteser. De to hypotesene som sammenlignes i identifisering av savnede personer (MP) vil være:

H_p: *MP tilhører familietreet som er formulert*
H_d: *MP er ubeslektet til de kjente referansepersonene i familietreet som er formulert*

For å sammenligne disse to hypotesene, utformes det generelle uttrykket

$$LR = \frac{P(G_{MP}, G_P | H_\mathrm{p})}{P(G_{MP}, G_P | H_\mathrm{d})}$$

Hvor sannsynligheten for at DNA-profilen til den savnede personen, G_{MP}, og referanseperson(ene), G_P, er beslektet, delt på sannsynligheten for de samme profilene dersom uidentifisert person og referanseperson(ene) er ubeslektet. Dette betyr at man ikke får noen indikasjon på hvilken hypotese som er sann dersom LR er 1. Dersom LR er større enn 1, er dataene mer sannsynlige under H_p, enn H_d, og mindre 1 vil angi dataene som mer sannsynlige under H_d enn H_p. For eksempel vil en LR på 1000 si at dataene er 1000 ganger med sannsynlige under hypotesen at MP tilhører familetreet enn at personen er ubeslektet.

For å identifisere en person ønsker man en høy LR, men det kan være vanskelig å sette en grense for hva som er høyt nok til at man kan være «sikre» i rettsgenetiske anvendelser. Dette vil diskuteres nærmere i Avsnitt 4.1. Ge et al. (2011) beskriver tilsvarende simuleringer og beregninger som i denne oppgaven, og markerer log LR over 6 som informativ identifisering i sine resultater. Buckleton et al. (2005) karakteriserer en tilsvarende LR på over 1,000,000 som «Extremely Strong» støtte for H_1, men en terskel for LR behøver nødvendigvis ikke være så høy for å karakteriseres som informativ identifikasjon.

2.5.3 Familietreanalyser med foreldre

Ved nedarving, vil en person få et allel fra hver forelder, med utgangspunkt i Mendels prinsipper. Hvert individ har to alleler, der et allel er en kopi av et korresponderende allel i individets far, mens det andre allelet er en kopi av en korresponderende allelet i individets mor. I dag vet man at alleler ikke nedarves helt uavhengig, da nærliggende locus ofte blir nedarvet sammen. Det blir derfor valgt locus som ligger fysisk langt fra hverandre, gjerne på forskjellige kromosomer, slik at man kan benytte seg av antakelsene i HW ved valg av alleler brukt som markører. Basert på dette vil man kunne ta sikre beslutninger dersom man har DNA-prøver fra både mor og far. Avhengig av foreldrenes alleler, finnes det mellom en og fire forskjellige kombinasjoner i hvert locus for deres avkom.

Dersom man har en uidentifisert person med genotype G_B, og personens antatte foreldre, hvor mor har genotypen G_M og far har genotypen G_F, vil evaluering av dette scenarioet kreve følgende formulerte hypoteser:

H_P: *De antatte foreldene er de sanne foreldrene til den uidentifiserte personen*
H_d: *Et ukjent foreldrepar er de sanne foreldrene til den uidentifiserte personen*

Basert på dette, vil LR bestemmes som

$$LR = \frac{P(data|H_P)}{P(data|H_D)}$$

$$= \frac{P(G_B, G_{M,} G_F|H_P)}{P(G_B, G_{M,} G_F|H_D)}$$

$$= \frac{P(G_B|G_M, G_F, H_P)}{P(G_B|G_M, G_F, H_D)} \cdot \frac{P(G_{M,}G_F|H_P)}{P(G_{M,}G_F|H_D)}$$

Det antas videre at den samlede sannsynligheten for å observere foreldrenes genotyper er uavhengig av hypotesene H_P og H_d, og det siste leddet i ligningen vil derfor bli 1. LR vil derfor beregnes som følger:

$$LR = \frac{P(G_B|G_M, G_F, H_P)}{P(G_B|G_M, G_F, H_D)} \qquad \textbf{Formel 2.1}$$

Sannsynligheten for den uidentifiserte personens alleler under hypotesen om at personene i familietreet er ubeslektet, vil være lik sannsynligheten for å observere disse allelene i populasjonen, og ved å forenkle Formel 2.1 vil utregningene baseres på ligningen i Formel 2.2 (Fung & Hu 2008).

$$LR = \frac{P(G_B|G_M, G_F, H_P)}{P(G_B|H_D)} \qquad \textbf{Formel 2.2}$$

Basert på Formel 2.2 kan LR for at den uidentifiserte personen er avkom av de antatte foreldrene gitt deres genotyper beregnes, og dette er gjort ved ulike scenario. En slik beregning for et scenario der mor har allelene A/C, far B/D og den uidentifiserte personen A/B blir vises her:

$$LR = \frac{\frac{1}{2} \cdot \frac{1}{2}}{2p_A p_B} = \frac{1}{8p_A p_B}$$

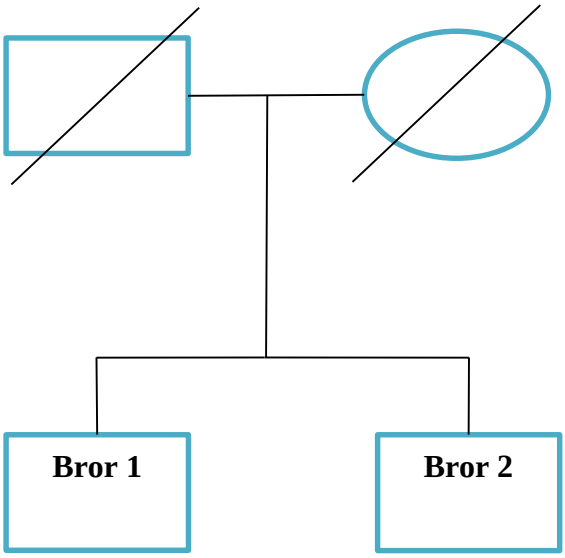
Utregning for de ulike scenarioene er å finne i Vedlegg 7, mens scenario og tilhørende beregning av LR er gitt i Tabell 2.1 LR for et allelsystem der frekvensen av allel A er 0.49, B er 0.01, C er 0.49 og D er 0.01.

Tabell 2.1: LR-verdi ved ulike genotyper for foreldre og uidentifisert person, samt validisering i Familias.

G_F	G_M	G_B	LR	Familias
A/A	A/A	A/A (1.1)	$\frac{1}{p_A^2}$	4.165
	A/B	A/A (1.2)	$\frac{1}{2p_A^2}$	2.082
		A/B (1.3)	$\frac{1}{4p_Ap_B}$	51.02
	B/B	A/B (1.4)	$\frac{1}{2p_Ap_B}$	102.04
	B/C	A/B (1.5)	$\frac{1}{4p_Ap_B}$	51.02
A/B	A/B	A/A (2.1)	$\frac{1}{4p_A^2}$	1.041
		A/B (2.2)	$\frac{1}{4p_Ap_B}$	51.02
	A/C	A/A (2.3)	$\frac{1}{4p_A^2}$	1.041
		A/B (2.4)	$\frac{1}{8p_Ap_B}$	25.51
	C/D	A/C (2.5)	$\frac{1}{8p_Ap_C}$	0.521

2.5.4 «Identity by descent» (IBD) ved søsken

«Identity by descent», som kan formuleres på norsk ved identisk ved nedarving, er et konsept som ble introdusert i 1940 av Cotterman, men som har blitt revidert av Malecot, Li og Sacks og Jacquard. Konseptet går ut på at to alleler er IBD hvis de er like fordi de er kopier av det samme allelet fra en felles stamfar (Buckleton et al. 2005). Dette vil derfor ha stor innvirkning på beregninger av slektskap mellom to personer. Ved beregninger av LR for søsken, kan man betinge med tanke på om de deler 0, 1 eller 2 alleler IBD. En slik teoretisk beregning vil bli gjøres nedenfor.



Figur 2.3: Illustrasjon av et familietre med foreldre og to brødre.

Konseptet med IBD mellom to brødre illustreres i Figur 2.3. Her står *I* for andel alleler som deles IBD, og hvor brødre, som nevnt, kan dele 0, 1 eller 2 alleler IBD. *I* er for brødre binomisk fordelt med parameterne p=0.5 og n=2. På den måten blir P(I = 0) = P(I = 2) = 0.25, mens P(I = 1) = 0.5 (Vigeland et al. 2012).

Tabell 2.2: Sannsynlighet for genotypepar gitt antall alleler delt IBD.

Genotype	***I* = 0**	***I* = 1**	***I* = 2**
A/A, A/A	p_A^4	p_A^3	p_A^2
A/A, A/B	$2P_A^3P_B$	$p_A^2p_B$	0
A/A, B/B	$p_A^2p_B^2$	0	0
A/B, A/B	$4p_A^2p_B^2$	$p_Ap_B(p_A+p_B)$	$2p_Ap_B$
A/B, A/C	$4p_A^2p_Bp_C$	$p_Ap_Bp_C$	0
A/B, B/C	$4p_Ap_B^2p_C$	$p_Ap_Bp_C$	0
A/B, C/C	$2p_Ap_Bp_C^2$	0	0

Basert på sannsynlighetene for genotypepar gitt antall alleler delt IBD av Tabell 2.2, er det mulig å finne LR for at personene er brødre, i forhold til at referansepersonen er en ubeslektet. Et konkret eksempel på denne utregningen vil bli gitt her, hvor LR for at personene er brødre gitt at begge har allelene AA beregnes. De resterende utregningene er å finne i Vedlegg 8. Det antas at foreldrenes genotype er ukjent, og følgende hypoteser settes opp:

H_p: *Den uidentifiserte personen er bror av person B med allelene A/A*

H_d: *Den uidentifiserte er ubeslektet til person B med allelene A/A*

Sannsynligheten for genotypene gitt hypotesen om at personene er brødre, er lik sannsynligheten for å observere allelene, gitt antall alleler de deler IBD, multiplisert med sannsynligheten for antall alleler delt IBD:

$$\begin{aligned} &P(A/A, A/A|Brødre) \\ &= P(A/A, A/A|I=0)P(I=0) \\ &+ P(A/A, A/A|I=1)P(I=1) \\ &+ P(A/A, A/A|I=2)P(I=2) \end{aligned}$$

Ettersom sannsynligheten for å observere at 0, 1 og 2 alleler deles IDB er henholdsvis ¼, ½ og ¼, vil LR-verdien være

$$= P(A/A, A/A|I = 0) \cdot \frac{1}{4} + P(A/A, A/A|I = 1) \cdot \frac{1}{2} + P(A/A, A/A|I = 2) \cdot \frac{1}{4}$$

$$= p_A^4 \cdot \frac{1}{4} + p_A^3 \cdot \frac{1}{2} + p_A^2 \cdot \frac{1}{4} \qquad \textbf{Formel 2.3}$$

Sannsynligheten for å genotypene, gitt at personene er ubeslektet, vil være sannsynligheten for å observere de fire allelene i en populasjon, altså

$$P(A/A, A/A|Ubeslektet) = p_A^4 \qquad \textbf{Formel 2.4}$$

Basert på Formel 2.3 og 2.4 kan LR-verdien beregnes:

$$LR = \frac{P(A/A, A/A|Brødre)}{P(A/A, A/A|Ubeslektet)} = \frac{p_A^4 \frac{1}{4} + p_A^3 \frac{1}{2} + p_A^2 \frac{1}{4}}{p_A^4}$$

$$= \frac{\frac{1}{4} p_A^2 + \frac{1}{2} p_A + \frac{1}{4}}{p_A^2} \qquad \textbf{Formel 2.5}$$

I et enkelt eksempel med kun tre alleler, antas det at allelfrekvensene for A, B og C er som følger: p_A=0.01, p_B=0.49 og p_C=0.5. Frekvensene benyttes i Formel 2.5 for å finne

$$LR = \frac{\frac{1}{4} \cdot 0.01^2 + \frac{1}{2} \cdot 0.01 + \frac{1}{4}}{0.01^2} = 2550.25$$

Under disse betingelsene, vil dataene være 2550,25 ganger mer sannsynlige dersom personene er brødre i forhold til at de er ubeslektet. LR er utregnet for alle scenario, og sjekket mot Familias, hvor eksempelet med samme allelfrekvenser og familietre ble gjennomført.

Tabell 2.3. Beregning av LR for at to personer er brødre gitt genotypedata.

Savnet	**Bror**	**LR**	**Familias**
A/A	A/A (1.1.)	$\frac{\frac{1}{4}p_A^2 + \frac{1}{2}p_A + \frac{1}{4}}{p_A^2}$	2550.25
	A/B (1.2)	$\frac{\frac{1}{4}p_A + \frac{1}{4}}{p_A}$	25.25
	B/B (1.3)	$\frac{1}{4}$	0.25
A/B	A/B (2.1)	$\frac{p_A p_B + \frac{1}{2}(p_A + p_B) + \frac{1}{2}}{4p_A p_B}$	38.5153
	A/C (2.2)	$\frac{p_A + \frac{1}{2}}{4p_A}$	12.75
	B/C (2.3)	$\frac{p_B + \frac{1}{2}}{4p_B}$	0.5051
	C/C (2.4)	$\frac{1}{4}$	0.25

Tabell 2.3 viser at det er stor spredning i LR-verdiene, selv der allelene for individene er like hverandre. Der begge individene har allelene AA, blir LR-verdien på 2550.25, noe som er betydelig høyere enn når begge individene har AB, noe som gir en verdi på 38.52. Dette viser at selv ved et lite eksempel kan LR-verdien bli nokså høy dersom alleler er sjeldne. Dersom flere markører benyttes, vil man kunne forvente høyere LR-verdi hvis personene er beslektet. Spesielt høye LR-verdier for hver markør, vil forventes i de tilfellene der allelene i flere markører er sjeldne.

Tabell 2.4: Sannsynligheten for at to brødre deler X antall alleler.

X	0	1	2
P(X=x)	$^1/_4$	$^1/_2$	$^1/_4$

Brødre behøver nødvendigvis ikke dele noen alleler, ettersom de i ¼ tilfeller ikke deler noen alleler IBD. Forventningsverdien til antall alleler delt IBD i en markør blir basert på Tabell 2.4

$$E(X) = 0\frac{1}{4} \cdot 1\frac{1}{2} \cdot 2\frac{1}{4} = 1$$

hvor forventningsverdien til systemet blir

$$S = X_1 + \cdots + X_n$$
$$E(S) = N \cdot 1 = N$$

Videre vil

$$E(X^2) = 0^2\frac{1}{4} + 1^2\frac{1}{2} + 2^2\frac{1}{4} = \frac{3}{2}$$

med variansen til antall alleler og systemet

$$Var(X) = \frac{3}{2} - 1^2 = \frac{1}{2}$$

$$Var(S) = N \cdot \frac{1}{2}$$

Standardavviket til et system vil derfor bli

$$SD(S) = \sqrt{\frac{1}{2}} \cdot \sqrt{N}$$

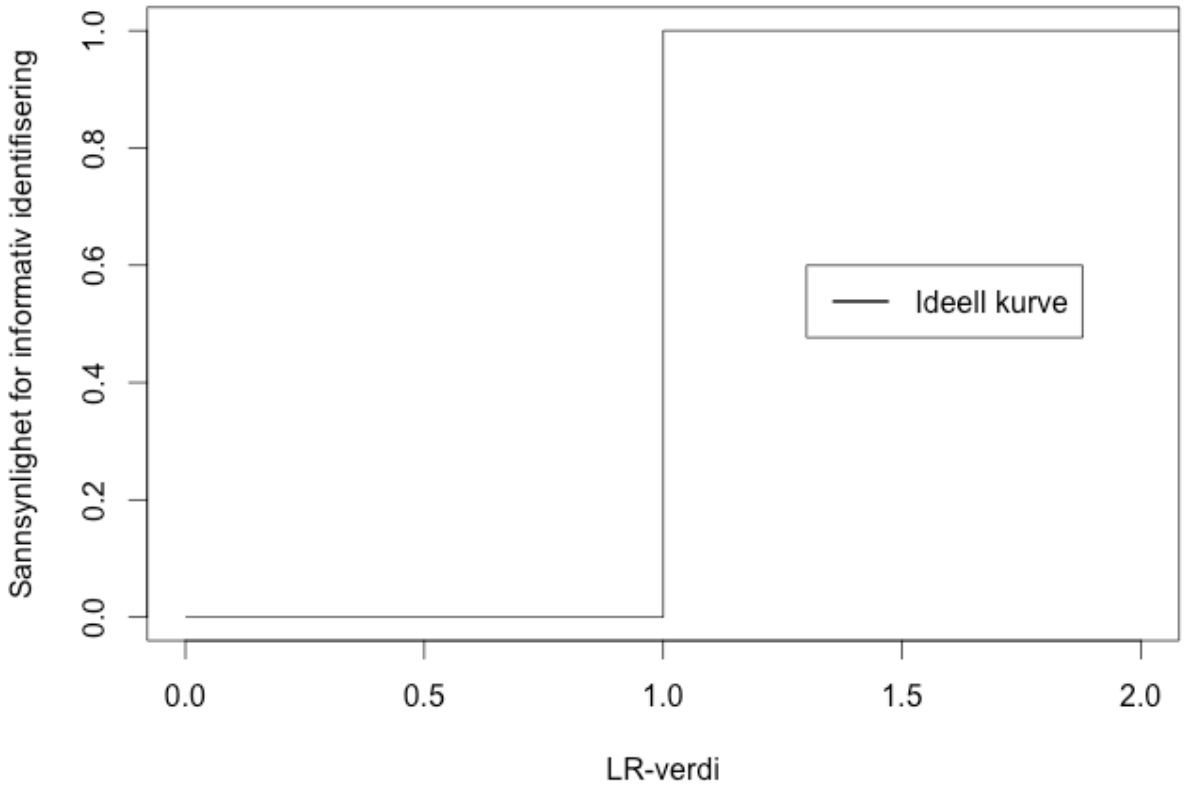
Forventningen er at søsken deler 1 allel per markør, hvor standardavviket i et system med 13 alleler blir 2.55. På grunn av variasjon i IBD mellom søskenpar, vil det kunne være forventet at det er større spredning i LR-verdiene dersom man kun har et helsøsken som referanseperson, enn om man skulle ha begge foreldrene hvor den savnede deler ett allel med begge foreldrene. Der man har

tilgang til begge foreldrene som referansepersoner, skal det være mulig å finne igjen begge allelene til den savnede blant allelene i foreldrenes DNA-profil, sett bort fra mutasjoner.

2.6 Styrkeberegning

En styrkeberegning angir sannsynligheten for å forkaste nullhypotesen for ulike verdier av den ukjente parameteren θ. Styrken opp mot et spesifikt alternativ er beregnet som sannsynligheten for at testen vil avvise nullhypotesen når dette spesifikke alternativet er sant. I gode eksperimenter må man påse at styrken vil bli rimelig høy for å oppdage rimelige avvik fra nullhypotesen (Løvås 2011). Når styrken øker, vil sjansene for at en type II-feil vil inntreffe bli mindre. Sannsynligheten for en type II-feil kalles en falsk negative rate, som også kalles sensitivitet.

For ulike verdier av den ukjent parameteren θ vil styrkeberegningen angi sannsynligheten for å forkaste nullhypotesen i klassisk styrkeberegning. I forensiske anvendelser er det ingen parameter eller signifikans, og det er typisk to alternativer, noe som gjør styrkeberegninger anvendes annerledes. De to hypotesene er at den uidentifiserte personen tilhører et bestemt familietre (nullhypotesen, H_p), eller at det det er en annen ukjent som tilhører familietreet (alternativ hypotese, H_d). For å få en informativ identifisering, må dataene være mye mer sannsynlige under nullhypotesen enn den alternative hypotesen, og det settes derfor en terskel for hvor høy LR skal være for å kunne bestemme at den uidentifiserte personen tilhører et familietre.



Figur 2.4: Ideell kurve i en styrkefunksjon med LR-verdi for treff og sannsynlighet for informativ identifisering.

Figur 2.4 viser en ideell kurve for styrkeberegninger, hvor alle treff med sanne familietrær får en LR-verdi over 1, mens alle usanne treff får LR-verdi under 1. Kurven vil i realiteten ikke bli slik, og det er derfor viktig å sette en høy nok terskel for LR til å utelate falske positive. I denne oppgaven er terskelen satt 5000, med en log LR-verdi på omtrent 3.7, men hvor denne terskelen bør settes diskuteres videre i avsnitt 4.1.

Hvor sikker kan man være på å få et riktig resultat? I Familias kan dette avhenge av blant annet antall markører, antall DNA-profiler blant referansepersonene og hvor mange personer som skal identifiseres. Ved å gjøre simuleringer i programmet vil man kjenne en fasit, som gjør det mulig å

teste sannsynligheten for å avvise nullhypotesen H_p, når man samtidig vet at denne hypotesen er sann. Dersom man kjører programmet ved ulike scenario, vil det være mulig å se innvirkningen på styrken fra hver av dem. Dette kan gjøre det mulig å gi en bekreftelse på om plattformen har potensiale til å bli brukt («proof of concept»), og hva som kan være med på å gi et mer riktig resultat av testene.

Tabell 2.5: Ordforklaring av variablene i Tabell 2.6.

Antall ofre	**Antall markører**
Lavt = -1	Få = -1
Høyt = 1	Mange = 1

Tabell 2.6: Scenario som testes for å se på styrken av beregningene.

Seed (eks.)	**Ofre, x_1**	**Markører, x_2**	**Feilrate, y**
17	-1	-1	…
17	1	-1	…
17	-1	1	…
17	1	1	…

For å få data til styrkeberegningene gjøres det simuleringer i Familias, som deretter behandles i Excel. I Excel kan man sortere ut data, slik at man kun får for eksempel savnede personers alleler i en fil, som kan lagres som tekstfil og leses inn i Familias. På samme måte kan man sortere familiene, der hver families alleler lagres som en fil, som kan leses inn i under «referance families» i Familias, for deretter å indikere familieforhold. Ettersom simuleringene vil gi oss fasiten, vil man kunne sjekke om programmet har klart å finne de sanne relasjonene mellom savnede personer og referansepersonene.

Seed i simuleringene kan varieres, men ved å angi seed kan man gjenta forsøket. DNA-profilene blir da de samme ved hver simulering med samme seed. Ved å variere antall referansepersoner og markører vil man kunne se hva som påvirker feilraten.

Etter å ha gjort slektskapsberegninger i Familias og ha fått en LR-verdi for hvert av treffene mellom uidentifisert person og antatte referansefamilier, er det derfor interessant å se hvor gode antagelser om identifisering beregningene gir. Det finnes da to ulike typer feil hypotesetestingen kan gi, disse er angitt av Tabell 2.7.

Tabell 2.7: Klassifisering av type feil og riktige avgjørelser i hypotesetesting.

	Tilhører en familien		
Klassifisert til riktig familie?		Ja	Nei
	Ja	Sann positiv, n_{11}	Falsk positiv, n_{12}
	Nei	Falsk negativ, n_{21}	Sann negativ, n_{22}

Basert på antall falske negative og falske positive, kan en feilrate beregnes.

$$Feilrate = \frac{n_{12} + n_{21}}{N}$$

hvor falske positive er den mest alvorlige typen feilen, og det kan derfor være hensiktsmessig å vekte feilraten.

$$Vektet\ feilrate = \alpha\frac{2n_{12}}{N} + (1-\alpha)\frac{2n_{21}}{N}$$

hvor α=0.5 er vanlig feilrate, mens for eksempel α=0.1 er vektet, hvor man legger større vekt på den mer alvorlige feilen å klassifisere en person til feil familie, når den egentlig tilhører en annen.

For å se på styrken til programmet i denne oppgaven, testes det hvor godt de statistiske beregningene i programmet er i stand til å indikere informativ identifisering, altså beholde H_0 når denne er sann. Dersom H_p er sann, men ikke har en LR-verdi over den gitte LR-terskelen, vil dette tilsvare en falsk negativ-feil. Det betyr at den uidentifiserte personen i virkeligheten tilhører familietreet, men har ikke høy nok LR-verdi, og at H_p dermed forkastes. Den mer alvorlige feilen med å klassifisere en uidentifisert person som en del av et familietre når sannheten av at en annen

ukjent person tilhører familietreet, vil dermed være en type 2-feil, angitt av antallet n_{21}. Da vil det antas en informativ identifisering, når den i virkeligheten er usann, altså er falsk positiv.

2.7 Familias 3

Den første versjonen av Familias ble utgitt i 1995 etter et samarbeid mellom Norsk Regnesentral, ved Thore Egeland og Petter Mostad, og Rettmedisinsk Institutt, ved Bjørnar Olaisen og Margurethe Stenersen. Programmet er en gratis «software» som kan lastes ned fra nettsiden familias.name, og kommer stadig i nye versjoner, hvor utviklingen nå er tatt over av Daniel Kling ved Folkehelseinstituttet og doktergradstipendiat ved NMBU.

Familias benyttes til å finne sannsynligheter og LR, som er beskrevet nærmere i avsnittene 2.5.2, 2.5.3 og 2.5.4, basert på DNA-profiler hvor personene er kjente, mens deres familierelasjon er ukjent. I Familias finnes det nå også en ny modul som gir muligheten til å finne slektskap mellom ukjente personer og referansefamilier eller –personer, som kan være et nyttig verktøy i identifisering av personer ved store ulykker (DVI). Identifiseringen baserer seg på alternative familietrær for personer med kjente DNA-profiler, og benytter informasjon fra DNA-observasjoner fra relevant populasjon for å beregne hvilke familietrær som er mest sannsynlig, og hvor sannsynlig det er i forhold til andre (Egeland & Mostad 2010).

For å kunne gjøre sannsynlighetsberegninger trenger man en database med informasjon om allelfrekvensene til de markørene som finnes i personenes DNA-profil. Denne databasen kan også inneholde informasjon om mutasjonsjonsrater i populasjonen. Deretter må man, uavhengig av modul, definere personene (også de man ikke har DNA-profil til, for å få skape fullstendige familietrær) og deres DNA-informasjon, og videre hvilke familietrær som skal testes mot hypotesen om at de ikke er beslektet.

2.7.1 DVI

Familias har vært brukt mest i forbindelse med slektskapssaker, men i Familias 3 finner man et par nye moduler. Den ene av disse er DVI-modulen, hvor programmet kan brukes til å identifisere mange personer ved hjelp av deres referansefamilier. For å kunne gjøre dette, må man ha DNA-profiler til de savnede personene, altså ukjent identitet, samt DNA-profilene til en eller flere referansepersoner med kjent identitet eller personlige eiendeler som potensielt kan gi et direkte treff. Basert på DNA-profilene til de savnede personene og definerte familitrær med minst en kjent DNA-profil vil programmet beregne LR-verdi og *a posteriori*-sannsynlighet for hver enkelt savnet person og hver familie. I noen tilfeller må referansepersoner uten kjent DNA-profil tillegges familietreet for å få det fullstendig for å kunne søke, for eksempel må man legge inn mor og far uten data dersom man kun har et søsken som referanseperson med DNA-informasjon. Hvordan DVI-beregninger i Familias er utført i denne oppgaven er å finne i Vedlegg 4.

Basert på de forenklede beregningene i avsnitt 2.2.3 vil programmet beregne LR-verdi for slektskapet, som i denne oppgaven dreier seg om relasjonene foreldre-barn og brødre. Når LR for et locus er funnet, kan man ved bruk av produktregelen beregne en samlet LR-verdi over flere uavhengige loci, ved å multiplisere de respektive LR-verdiene.

2.7.2 Blindsøk («Blind search»)

Foruten DVI-modulen i Familias3, er også «Blind Search» en ny modul som kan være behendig i identifikasjonsberegninger ved store ulykker. I denne delen av programmet behøver man ikke definere noen slektstrær før man søker, men programmet søker selv etter en relasjon mellom DNA-profilene som er lagt inn. Det er mulig å søke etter relasjonene foreldre-barn, søsken, halv-søsken, kusiner eller tremenning, og man kan på den måten finne ut om noen av ofrene er beslektet. Det er også mulig å finne trioer, for eksempel om to av personene er foreldre til en tredje. Dersom flere DNA-prøver stammer fra samme person vil det også være mulig å gjøre et søk i denne modulen, altså et søke etter et direkte treff – en situasjon som kan oppstå dersom det er nødvendig å

identifisere flere levninger av samme offer. Resultatet av et Blindsøk gir en LR-verdi for den enkelte relasjonen mellom to personer i datasettet som er lastet inn. En fremgangsmåte for bruk av Blindsøk i Familias3 er å finne i Vedlegg 5.

2.8 R

R er et gratis programmeringsspråk og -miljø for statistisk databehandling og grafikk, i utgangspunktet utviklet av Robert Gentleman og Ross Ihaka i 1993, men siden 1997 har en stadig økende gruppe personer bidratt til å skrive programmet. Programmet kan lastes ned gratis fra http://cran.uib.no/. R gir et bredt utvalg av statistiske og grafiske teknikker, og ettersom R er en «open source», kan brukere utvide mulighetene i programmet ved å lage nye pakker. En del pakker finnes allerede i installasjonen i R, og disse er benyttet for plotting i oppgaven (The R Project 2014).

Grafikken i denne oppgaven skrevet i R, består av Kernel tetthetsestimering av log LR-verdiene, og skriptene er å finne i Vedlegg 9. I tetthetsfordelingen brukes x-verdiene log LR-verdi som input, og figurene viser sannsynligheten for hvordan log LR-verdiene er fordelt for hvert av scenarioene, hvor de fordeles over en jevn kurve, i motsetning til et histogram. «Default»-verdier på «bandwidth» og standardavvik er brukt i plottingen. «Bandwidth» er derfor «nrd0», som er basert på «Silvermans tommelfingerregel». Det er brukt en «Gaussian Kernel tetthetsestimator», Formel 2.6.

$$\widehat{h_{rot}} = 1.06\, min\left\{\hat{\sigma}\frac{R}{1.34}\right\} n^{-1/5} \qquad \textbf{Formel 2.6}$$

hvor $\hat{\sigma}$ står for det minste standardavviket, R er det interkvartile spekteret og n er prøveutvalget (The R Project 2013).

For å beregne gjennomsnitt, standardavvik, varians og persentiler i avsnitt 3.1, benyttes pakken

RcmdrPlugin.NMBU. R Commander er en pakke beregnet på og brukt i flere kurs ved NMBU, og er vedlikeholdt hovedsakelig av Kristian Hovde Liland (Liland 2014).

3 Resultater

3.1 DVI-eksempler i Familias

Eksemplene på bruk av modulen DVI i Familias er basert på simulerte data. Dataene er også simulert i Familias, og allelene og allelfrekvensene fra databasene CODIS og SGM er benyttet i testene. Alle fremgangsmåter i Familias og etterarbeid av resultatene for videre bruk fremgår av Vedlegg 2, 3 og 4. Ved å ha simulerte data har man fasiten på hvilken savnet person som tilhører hvilket familietre, og det vil på den måten være mulig se hvor godt programmet klarer å finne en høy LR-verdi der det er forventet en høy LR-verdi. Ved å benytte seg av forskjellige referansepersoner og antall markører vil det forventes forskjell i gjennomsnittlig LR-verdi, noe som kan indikere hvilke scenario som vil gi informativ identifisering. Det vil i hvert tilfelle bli gitt en seed, noe som gjør det mulig å gjenskape forsøkene, men filene brukt her vil også være tilgjengelig fra forfatteren. Simuleringene er gjort med antakelsene om ingen mutasjon og at hvert locus er uavhengig, og det ses bort ifra eventuelle genotypingsfeil.

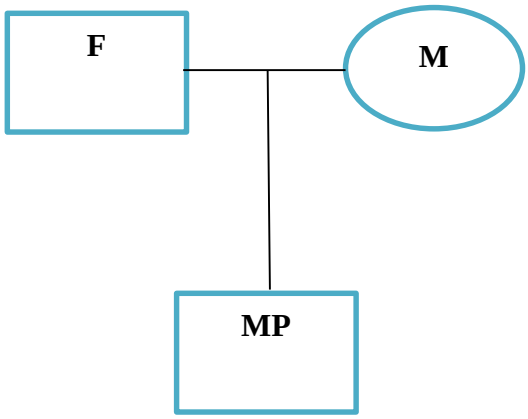
3.1.1 To foreldre

I dette første eksempelet i denne delen av oppgaven benyttes CODIS-systemets 13 markører for en kaukasisk populasjon, og det er simulert data for 100 familietrær basert på dette. Et utdrag av resultatene av simuleringene i Familias er vist i Tabell 3.1, hvor siste og første markør vises, samt de to første og det siste familietreet. I tillegg viser simuleringene «Amel» som indikerer hvilket kjønn personene er, slik at man slipper å legge til denne informasjonen når personene importeres til Familias.

Tabell 3.1: Utdrag av resultatet av simulering av foreldre og uidentifisert persons genotype i Familias.

Family	Sample	CSF1PO 1	CSF1PO 2	…	D21S11 1	D21S11 2
1	Mother	10	11	…	30	31.2
1	Father	10	11	…	30	28
1	MP	11	11	…	31.2	30
2	Mother	10	10	…	28	28
2	Father	11	10	…	31.2	28
2	MP	10	10	…	28	31.2
…	…	…	…	…	…	…
100	Mother	11	12	…	30	32.2
100	Father	12	10	…	29	32.2
100	MP	11	12		32.2	32.2

Seed i disse simuleringene i Familias er satt til 500. For å identifisere disse personene finnes det to referansepersoner i familietreet, nemlig mor og far til den uidentifiserte. Foreldreparenes genotyper er simulert sammen med genotypene til de uidentifiserte personene, men importeres separat til Familias i familietreanalysene. Hvordan filene er bearbeidet i Excel for å få genotypedataene i separate filer som igjen kan lastes inn i Familias, går frem av Vedlegg 3. Av utdraget av simuleringene kan det ses at de savnede personenes alleler er mulig å finne igjen blant foreldrenes alleler.



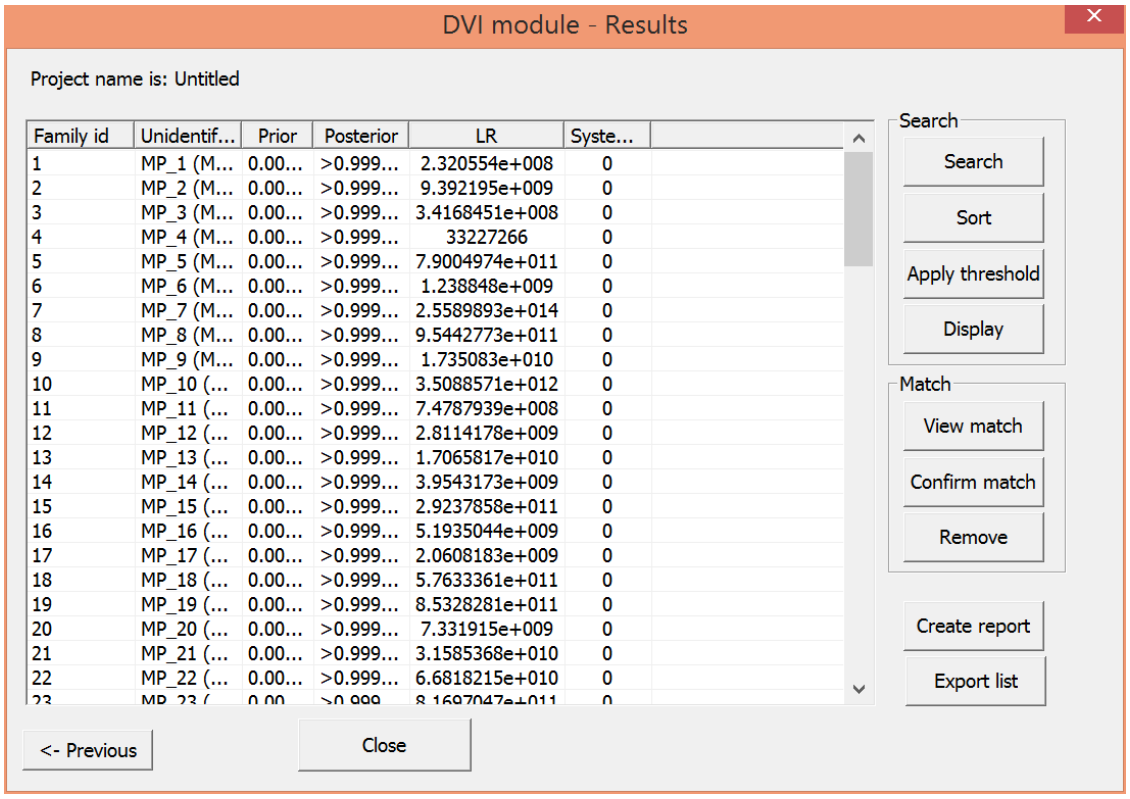
Figur 3.1. Illustrert familietre for H_p.

Det er først beregnet en LR-verdi for hver enkelt markør, som baseres på metodikken beskrevet i Avsnitt 2.5.3 med familietre-analyser. LR-verdien for hvert enkelt familietre beregnes ved å multiplisere LR for hver markør med hverandre, ved å beregne hvor sannsynlig DNA-profilene er under de to hypotesene:

H_p: *Den uidentifiserte personen er barn av foreldrene M og F*

H_d: *Levningene er er fra en ukjent person ubeslektet til M og F*

I det andre eksempelet er forsøket gjentatt med de samme referansepersonene, men systemet SGM, som kun inneholder 6 av de samme markørene benyttes.



Figur 3.2 Utdrag av resultatene fra Familias med foreldre og CODIS.

LR-verdi for den forbindelsen mellom hver uidentifisert person og tilhørende referansefamilie som i simuleringene stammer fra samme familietre, altså sann positive, er brukt i beregningene her. I scenarioet med foreldre sammenfaller dette med de slektskapene som gir de høyeste LR-verdiene. Disse LR-verdiene er lagret i en egen fil, og videre beregninger er foretatt i R Commander, og resultatene av dette er å finne i Tabell 3.2. For sammenligninger med relevante artikler og bedre visuell visning, er det videre valgt å bruke 10-logaritmen av LR-verdiene i beregningene. I tillegg til gjennomsnitt og standardavvik for hvert av scenarioene, er det valgt å presentere varians og persentiler for sammenligning med relevant litteratur.

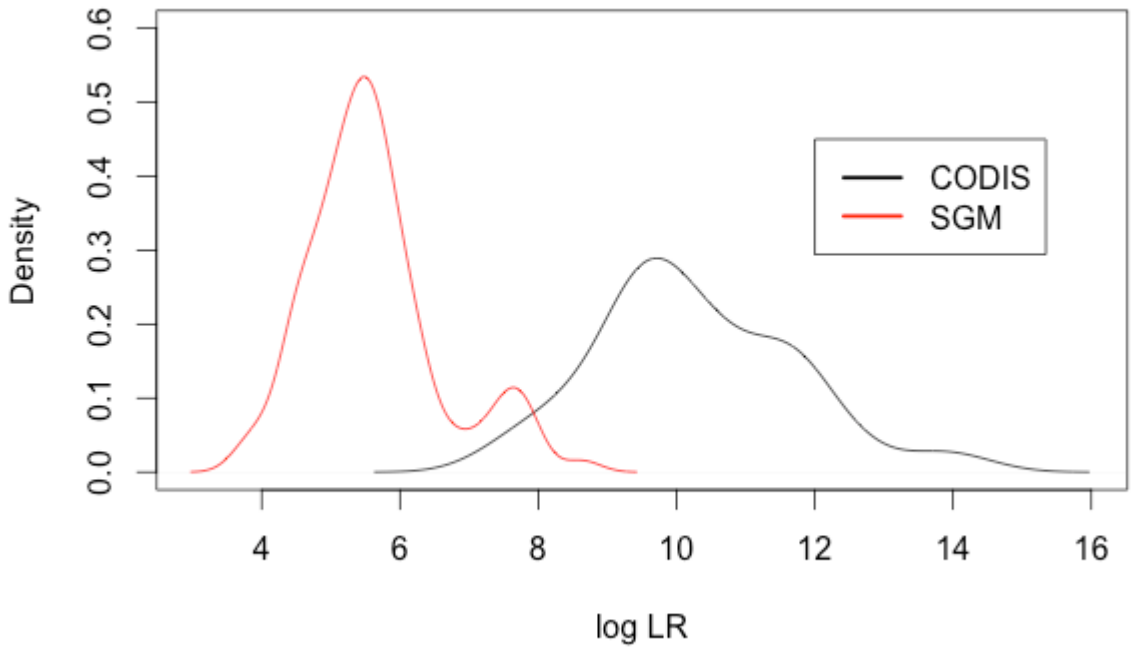
Tabell 3.2: Gjennomsnitt, standardavvik, varians, 5-, 1- og 0.1-persentiler for Log_{10}-verdiene hvor to foreldre er referansepersoner.

System	Gjennom-snitt	Varians	Standard-avvik	5%	1%	0.1%	Simuler-inger
CODIS	10.28	2.12	1.46	7.94	7.36	7.19	100
SGM	5.61	0.93	0.97	4.38	3.81	3.76	100

Gjennomsnittlig log LR-verdi for systemet CODIS på 10.28 er en del høyere enn for systemet SGM, og bruk av flere markører vil derfor indikere større sjanse for en informativ identifisering. I alle tilfellene er det treff mellom den uidentifiserte personen og den riktige familien som gir den høyeste LR-verdi, hvor ingen andre treff kom over LR-grensen på 1 som er satt for visning av resultater i Familias. Det finnes derfor ingen søk mellom uidentifisert person og familietrær med foreldre som referansepersoner med et falskt positivt treff.

Ved hjelp av en Welch t-test i R-Commander (utskrift R-Commander Vedlegg 11), er det testet om det er forskjell på gjennomsnittet av log LR-verdiene for CODIS og SGM ved konfidensnivå 0.99. Dette gir en t-verdi på 26.17, med frihetsgradene 171.95 og p-verdi mindre enn $2.2 \cdot 10^{16}$. Ettersom det antas at den sanne forskjellen i gjennomsnitt er større enn 0, beregnes det at LR-verdien for systemet CODIS er gjennomsnittlig 46 773 ganger høyere enn SGM ved bruk av foreldre som referansepersoner. Standardavviket til systemet CODIS er en del høyere enn SGM, hvor også ratioen mellom standardavvik og gjennomsnitt er høyere for CODIS. Det vil si at log LR-verdiene for SGM er mer sentrert rundt gjennomsnittet, med færre avvikende resultater enn ved beregning av log LR-verdi med systemet CODIS. SGM er derfor mer konsistent.

Et Kernel tetthetsplot er gitt i Figur 3.3 for å gi en visuell sammenligning av log av LR-verdiene for systemene CODIS og SGM.



Figur 3.3: Log LR for testing mot foreldre ved systemene CODIS og SGM.

Når antall loci i systemet øker, skifter LR-fordelingen til høyre for de sanne referansepersonene, noe som indikerer sterkere støtte til det antatte familietreet. Alle LR-verdiene er høyere enn 1, og det er derfor mer sannsynlig at den uidentifiserte personen tilhører familietreet enn en ukjent person i alle tilfellene. Av figuren kan det også ses at bredden på fordelingen er høyere for familietrær med et større antall markører. Det stemmer overens med beregnet standardavvik for CODIS og SGM, hvor CODIS har større standardavvik. Det viser dermed at det blir et større spekter log LR-verdier når flere markører benyttes i identifiseringen. Dersom man antar en terskel for log LR på 3.7, vil hele arealet av kurven med CODIS klart være til høyre for terskelen. Størsteparten av arealet under SGM-kurven ser ut til å være til høyre for LR-terskelen, men det er ikke like tydelig.

3.1.2 Én bror

For å se på innvirkningen av hvilke referansepersoner som benyttes i DNA-basert identifisering har på LR-verdiene, benyttes et eksempel hvor familietreet består av en bror med kjent genotype,

mens mor og far er ukjente. Det undersøkes om en av de uidentifiserte personene er medlem av familietreet som savner en bror. På samme måte som i eksemplene i avsnitt 3.1.1, er genotypedata for både den uidentifiserte personen og broren i familietreet simulert i Familias, basert på allelfrekvensene i databasen CODIS og uten mutasjonsmodell. 100 simuleringer er foretatt, slik at det finnes 100 uidentifiserte personer og 100 antatte brødre av de uidentifiserte, og simuleringene er gjort med seed 501.

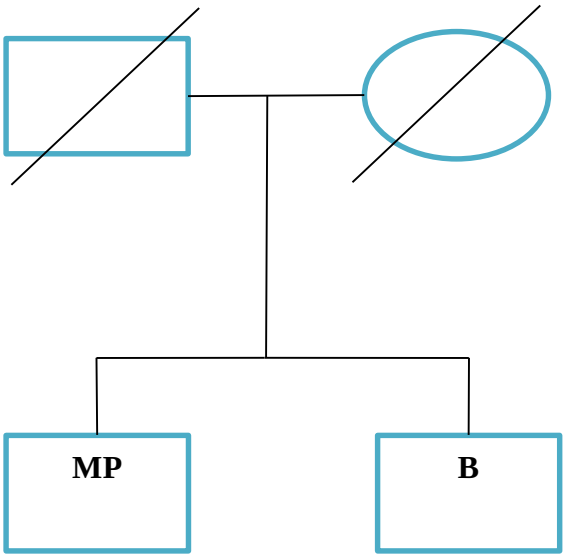
Tabell 3.3: Utdrag av resultatet av simulering av brødres genotype i Familias.

Family	**Sample**	**CSF1PO 1**	**CSF1PO 2**	**…**	**D21S11 1**	**D21S11 2**
1	Brother	12	12	…	28	32.2
1	MP	12	12	…	28	29
2	Brother	13	11	…	28	32.2
2	MP	13	11	…	29	32.2
…	…	…	…	…	…	…
100	Brother	11	11	…	31.2	29
100	MP	11	12	…	30	29

For å få simulerte data for en bror, må også data for foreldre simuleres. Disse utelates i de videre beregningene i Familias. Av utdraget av simuleringene illustreres det at brødre deler forskjellig antall alleler IBD.

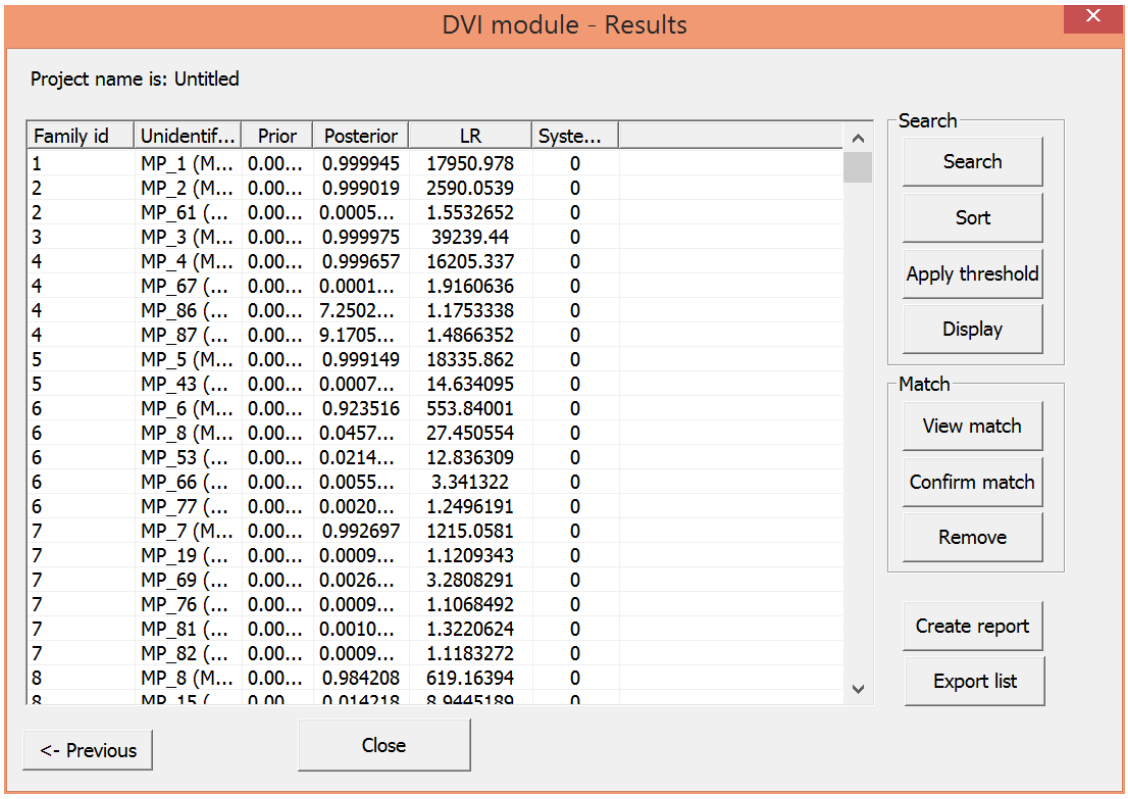
Hypotesene for hver uidentifisert persons forhold til hvert familietre som benyttes for å beregne LR-verdien, er utformet som følger:

H_P: *Den uidentifiserte, MP, er bror av personen B med DNA − profil* G_B
H_d: *Den uidentifiserte, MP, er en ukjent person ubeslektet til person B*



Figur 3.4 Familietreet som blir undersøkt i dette avsnittet hvor kun genotypene til savnet person (MP) og bror (B) er kjent.

Basert på metodikken i avsnitt 2.5.4 hvor IBD introduseres og fremgangsmåte i Familias i Vedlegg 4, beregnes en LR-verdi for hvert av de potensielle familietrærne. Et utdrag av resultatene i Familias kan ses av Figur 3.5. LR-verdiene for treff mellom familie og uidentifisert person som er simulert sammen er utplukket manuelt, og de sanne positive lagres i en egen excelfil for videre beregninger. Det samme er gjort med høyeste LR-verdi for treff mellom familie og uidentifisert person, uavhengig av om de er simulert i samme familietre, altså sanne positive og falske positive med høyest LR-verdi.



Figur 3.5: Utdrag av resultatene i Familias, hvor søk med bror som referanseperson er benyttet.

LR-verdiene for de familietrærne der den uidentifiserte personen er testet mot den referansefamilien den er simulert fra, er benyttet til å beregne gjennomsnitt og varians som finnes i Tabell 3.4. Som tidligere er LR-verdi for hver markør multiplisert for å finne LR-verdien for hele systemet.

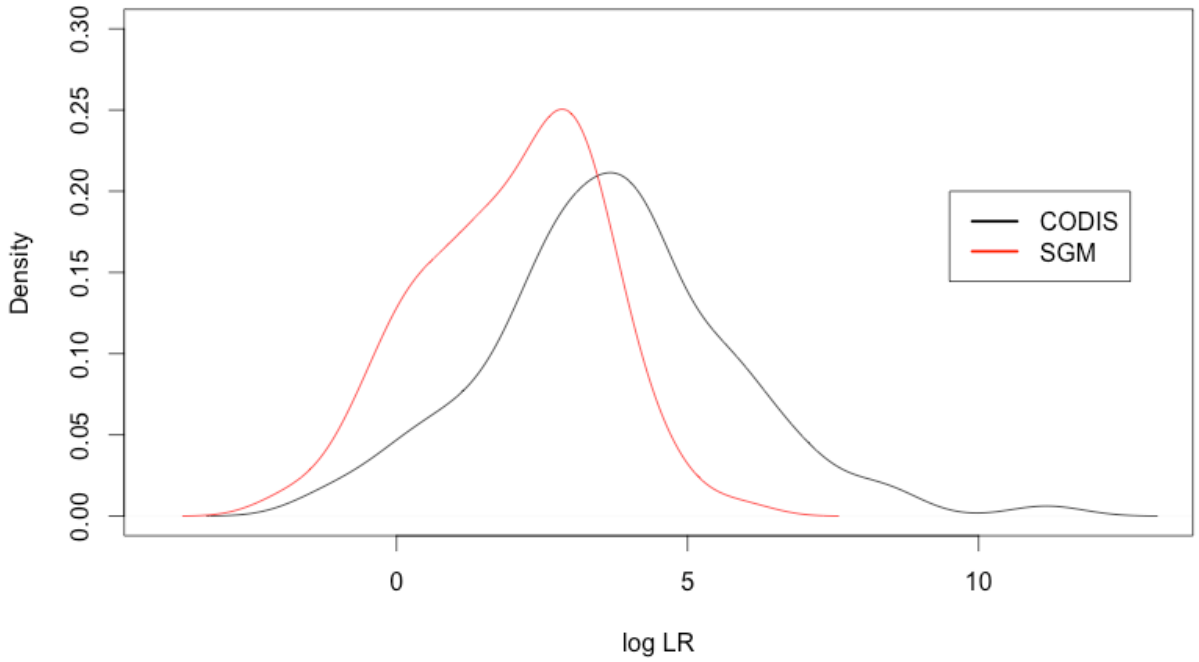
Tabell 3.4: Gjennomsnitt, standardavvik, varians, 5-, 1- og 0.1-persentiler for Log_{10}-verdiene hvor én bror er referanseperson.

System	Gjennom-snitt	Varians	Standard-avvik	5%	1%	0.1%	Simuler-inger
CODIS	3.68	4.59	2.14	0.07	-1.18	-1.36	100
SGM	1.98	2.40		-0.07	-1.77	-2.02	100

Gjennomsnittlig log LR-verdi for familietreet ved bruk av systemet CODIS med 13 markører (3.6845) er en del høyere enn ved SGM med 6 markører (1.9880), og dermed vil bruk av flere markører som ventet gi sikrere DNA-identifisering.

Ved hjelp av en Welch t-test i R-Commander (utskrift R-Commander Vedlegg 11), er det testet om det er forskjell på gjennomsnittet av log LR-verdiene for CODIS og SGM ved konfidensnivå 0.99. Dette gir en t-verdi på 6.42, med frihetsgradene 180.25 og p-verdi mindre enn 5.91e-10. Det er altså mindre forskjell på gjennomsnittlig log LR-verdi for systemene CODIS og SGM ved bruk av brødre enn foreldre. Ettersom det antas at den sanne forskjellen i gjennomsnitt er større enn 0, beregnes det at LR-verdien for systemet CODIS er gjennomsnittlig 50 ganger høyere enn SGM ved bruk av bror som referanseperson. Standardavviket til systemet CODIS er en del høyere enn SGM, hvor også ratioen mellom standardavvik og gjennomsnitt er høyere for CODIS. Det vil si at log LR-verdiene for SGM er mer sentrert rundt gjennomsnittet, med færre avvikende resultater enn ved beregning av log LR-verdi med systemet CODIS. SGM er derfor, også ved bruk av bror som referanseperson, mer konsistent.

Log LR for begge systemene er også plottet i R ved Kernel tetthetsfordeling for å gi en visuell sammenligning av testene, se Figur 3.6.

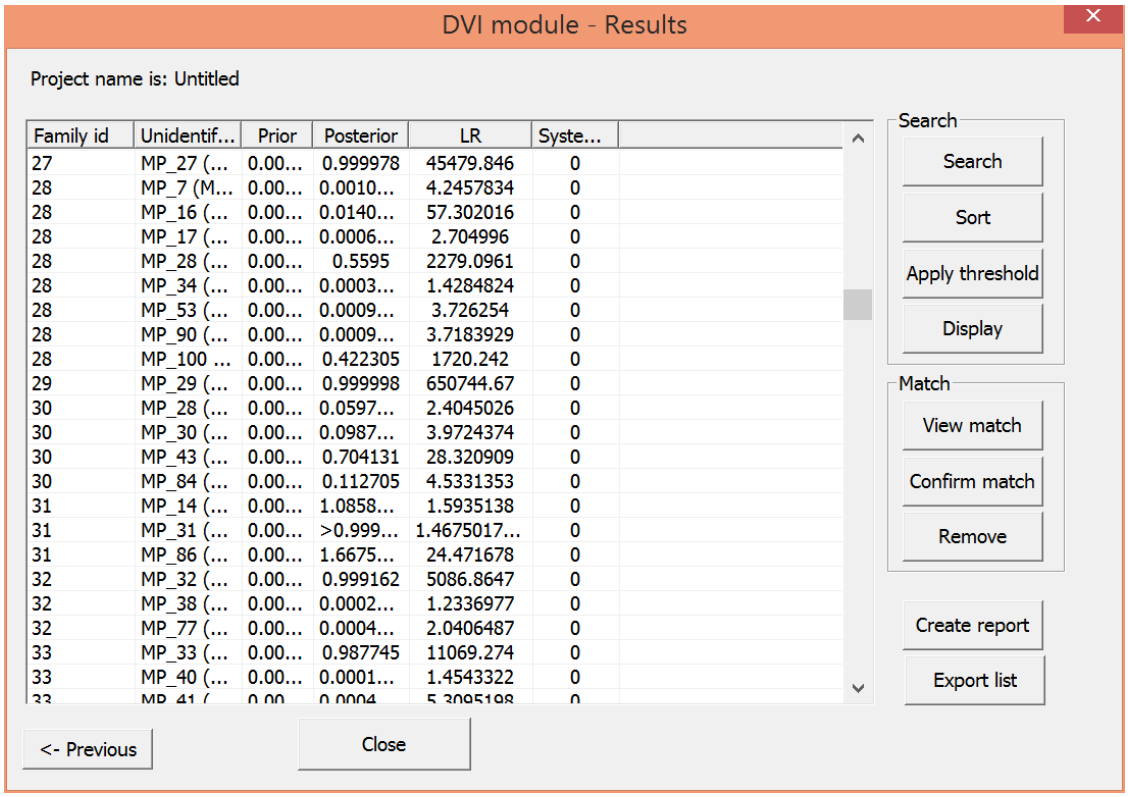


Figur 3.6: Log LR ved bruk av bror som referanseperson med systemene CODIS og SGM.

Det antas at log LR=3.7, eller 5000 er satt som grense. Arealet til høyre for denne grensen i Figur 3.6 ser ut til å være omtrent halvparten av arealet under CODIS-kurven, mens størsteparten av arealet under SGM-kurven befinner seg til venstre for denne grensen. Som i forsøket med foreldre, viser det seg at LR-verdiene er høyere ved bruk av flere markører i søk hvor DNA-profilen til en bror er kjent. Fordelingen av LR-verdiene hvor familietreanalysene inneholder brødre gir større spredning, sammenlignet med der begge foreldrenes DNA-profil er kjent. Fordelingen strekker seg over et større område, og ikke alle LR-verdiene ligger over 1 (log 0). Arealet under kurven til venstre for null med CODIS har et mindre areal enn tilsvarende areal under kurven med SGM. Det indikerer at i denne regionen til venstre for null i Figur 3.6 vil hypotesetestingen støtte hypotesen Hd, at den uidentifiserte, MP, er en ukjent person ubeslektet med person B.

I en del tilfeller blant simuleringene av familietrærne som inneholder profilene til en bror, får et annet familietre høyere LR-verdier enn det simulerte og sanne familietreet, altså et tilfeldig treff. Ettersom familie og savnet person er nummerert ut ifra plassering i simuleringene, vil det sanne

familietreet tilsvare familie 1 til MP1, familie 2 til MP2, og så videre til familie 100 til MP100. I noen tilfeller får en savnet person treff med feil familie, og høyere LR-verdi enn den sanne savnede personen som tilhører familietreet. For eksempel gir hypotesen med familie 30 høyere LR-verdi med person 43, enn person 30, som er den uidentifiserte personen som er simulert sammen med familie 30, som illustrert i utdraget av resultatene i Familias vist i Figur 3.7.



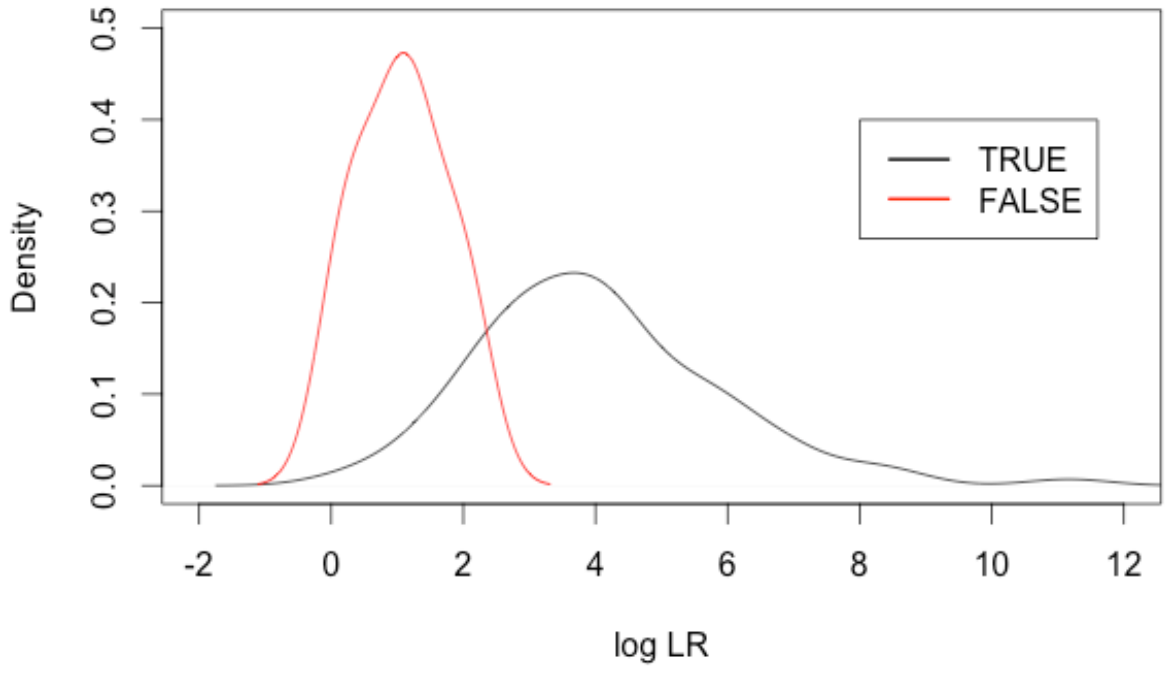
Figur 3.7: Utdrag av resultatene i Familias med bror som referanseperson som illustrerer at et treff mellom en referansefamilie og usann uidentifisert person får høyest LR-verdi i noen tilfeller.

Gjennomsnitt, varians, standardavvik og persentiler for disse tilfeldige treffene med feil uidentifisert person er gitt i Tabell 3.5. Det er i disse tilfellene større sannsynlighet for at en tilfeldig uidentifisert person er den savnede i et annet familietre enn i den sanne uidentifiserte personen.

Tabell 3.5: Gjennsomsnitt, varians, 5-, 1- og 0.1-persentiler for Log_{10}-verdiene der et annet familietre har fått høyere LR enn den riktige referansefamilien.

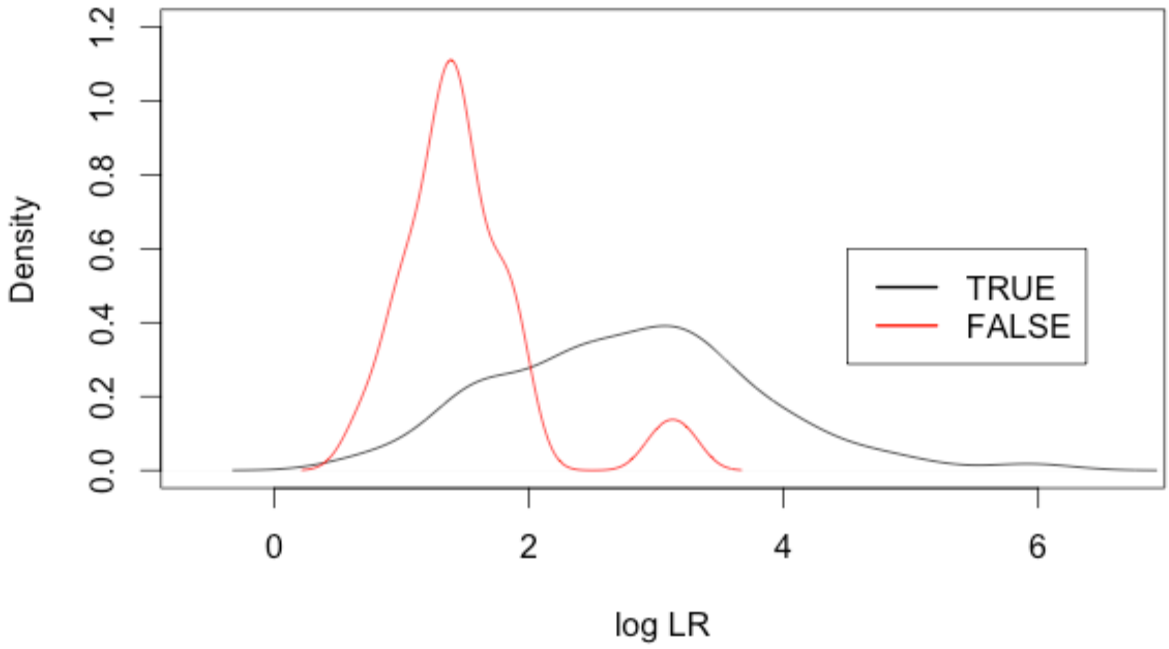
System	Gjennomsnitt	Varians	5%	1%	0.1%	Antall
CODIS	1.071	0.4907	0.0856	0.0856	2.117	9
SGM	1.489	0.29527	0.7916	0.65319	3.230224	31

For plotting av Figur 3.8 og 3.9 er de høyeste LR-verdiene for hvert familie plukket ut manuelt, uavhengig av om dette er det sanne familietreet eller ikke. Disse verdiene er plottet for å gi et visuelt inntrykk av fordelingen for henholdsvis CODIS og SGM, og gruppert, hvor «True» står for LR-verdiene i de tilfellene der det sanne familietreet gir høyeste LR, mens «False» står for LR-verdiene i de tilfellene der en annen uidentifisert person i søket ga større sannsynlighet for dataene H_p, enn den sanne uidentifiserte personen.



Figur 3.8: Høyeste log LR for hvert familietre med bror som referanseperson ved systemet CODIS, fordelt på om det er det sanne familietreet eller ikke.

Fordelingen av de høyeste sanne LR-verdiene er bredere og høyere enn fordelingen av de falske LR-verdiene, men de falske LR-verdiene ligger innenfor samme spekter som de sanne LR-verdiene. Det er også en del flere LR-verdier for de sanne LR-verdiene. Ettersom LR-grensen for informativ identifisering i denne oppgaven er satt til 5000, som altså tilsvarer log LR-grense på tilnærmet 3.7 gir ingen av de tilfeldige treffene i dette eksempelet høy nok LR-verdi til feilaktig plassering av en uidentifisert person i et familitre. Det kan tyde på at de falske LR-verdiene er mer konsistente.



Figur 3.9: Høyeste log LR for hvert familietre med bror som referanseperson ved systemet SGM, fordelt på om det er det sanne familietreet eller ikke.

Fordelingen av de høyeste LR-verdiene ved bruk av en bror som referanseperson og systemet SGM viser samme trender som ved systemet CODIS, men gir flere LR-verdier hvor urelaterte personer ser ut til å være beslektet. Disse falske positive gir en liten topp nærmere terskelen for LR rundt 3.7, men gir gjennomsnittlig lavere LR-verdier enn de sanne positive.

Dersom fordelingene av beslektede og falske ubeslektede overlapper, noe som er tilfellet ved bruk av bror som referanseperson, vil det finnes et område blant LR-verdiene som kan gi falske positive

og falske negative indikasjoner. I forsøkene i denne oppgaven ga ingen av de falske ubeslektede høy nok LR-verdi til å overstige terskelen for LR på 5000. De ubeslektede ville derfor ikke, med denne terskelen, blitt brukt til å anta et slektskap mellom de uidentifiserte og referansepersonene. Det er derimot ikke en fullstendig seperasjon av de to fordelingene med sann/usann i verken søk med SGM eller CODIS. Det er noe man ønsker for å redusere sjansen for å feilaktig identifisere individer som beslektet eller ubeslektet, noe som i disse tilfellene vil medføre at en andel individer vil karakteriseres som ubeslektet når de egentlig er beslektet med denne terskelen.

3.1.3 Sammenligning av eksempler med to foreldre og én bror

For å sammenligne om det er forskjell i gjennomsnittlig LR-verdi ved bruk av henholdsvis foreldre og bror som referanseperson, er det foretatt en Welch-t-test for hvert av systemene CODIS og SGM. For systemet CODIS, testes hypotesen om at gjennomsnittlig log LR-verdi for foreldre er større enn gjennomsnittlig log LR-verdi for bror. Det gir en t-verdi på 25.45, og p-verdi mindre enn 2.2e-16. Hypotesen om at forskjellen er større enn null støttes på konfidensnivå 0.99. For de simulerte dataene i oppgaven, er gjennomsnittlig LR-verdi for foreldre $3.47 \cdot 10^6$ ganger større enn gjennomsnittlig LR-verdi for brødre ved bruk av markørene i CODIS.

Samme test er gjennomført for systemet SGM, og utskriften fra R-Commander kan ses i Vedlegg 11. Testen gir en t-verdi på 19.85 og p-verdi mindre enn 2.2e-16. Dette støtter også at differansen mellom gjennomsnittlig LR-verdi ved bruk av henholdsvis foreldre og bror, med samme system, er positiv og større enn null. For de simulerte dataene i oppgaven ved systemet SGM, er gjennomsnittlig LR-verdi for foreldre er 4168 ganger større enn gjennomsnittlig LR-verdi ved bruk av bror som referanseperson.

Det er størst spredning i LR-verdiene i søk med systemet CODIS og bror som referanseperson, hvor standardavviket er størst. Søk med systemet SGM og foreldre som referanseperson gir lavest standardavvik, og dermed minst spredning i LR-verdiene. Det forholder seg likt dersom man ser på variasjonskoeffisienten, som gir et normalisert forholdstall mellom standardavvik i forhold til

gjennomsnittet.

3.2 Blindsøk i Familias

For å teste blindsøk-modulen i Familias, er de fem første simulerte familietrærne fra avsnitt 3.2 og 3.3 benyttet. I tillegg er DNA-profilen til MP_2 fra søk med søsken lagt til to ganger, for å teste direkte treff-funksjonen som ikke har vært prøvd tidligere. En forklaring av enkelte personer er gitt av Tabell 3.6 for å lettere kunne sammenligne med tidligere presenterte resultater.

Tabell 3.6: Forklaring av enkelte personer benyttet i blindsøket, for sammenligning med tidligere resultater.

Person i blindsøk	Tilsvarer person i tidligere søk
18	MP_2 i avsnitt 3.1.2
26	MP_2 i avsnitt 3.1.2
1	MP_1 i avsnitt 3.1.1
2	Simulert mor til MP_1, avsnitt 3.1.1
3	Simulert far til MP_1, avsnitt 3.1.1
16	MP_1 i avsnitt 3.1.2
17	Simulert bror til MP_1, avsnitt 3.1.2

Fremgangsmåte for bruk av blindsøk i Familias fremgår av Vedlegg 5. Et utdrag av resultatene er vist her i Tabell 3.7, mens de resterende er å finne i Vedlegg 6.

Tabell 3.7: Utdrag av resultatene av blindsøket.

Person 1	Person 2	Forhold	Log LR-verdi
18	26	Direkte treff	18.88
18	26	Søsken	12.47
18	26	Foreldre-barn	8.51
1	2	Foreldre-barn	3.29
1	3	Foreldre-barn	2.96
16	17	Søsken	4.25
16	17	Foreldre-barn	4.25

Ettersom person 18 og 26 i søket innehar samme DNA-profil, vil søket, som forventet, gi en høy LR-verdi. Ettersom det i dette tilfellet kan antas at det har vært god nok kvalitet på DNAet til å kunne utforme en fullstendig DNA-profil, vil hvor høy LR-verdien blir, kun avhenge av allelfrekvensene i populasjonen. En høy LR-verdi for et direkte treff, vil også resultere i høy LR-verdi for søk etter søsken og foreldre-barn.

Ved søk etter relasjonen foreldre-barn mellom person 1 og 2, og 1 og 3 kan det ses at søket ved blindsøk ikke gir høy nok LR-verdi til å sikre informativ identifisering. Når foreldrene søkes opp mot en savnet person uten søk med også den andre forelderen, er det vanskelig å få høy nok LR til å gi informativ identifisering. Det skal i Familias også være mulig å søke etter trioer, for å gjøre et samlet søk med begge foreldrene, og dermed få tilsvarende verdier som i treffene i Avsnitt 3.1.1, men dette var ikke ferdig implementer i tide for denne oppgaven.

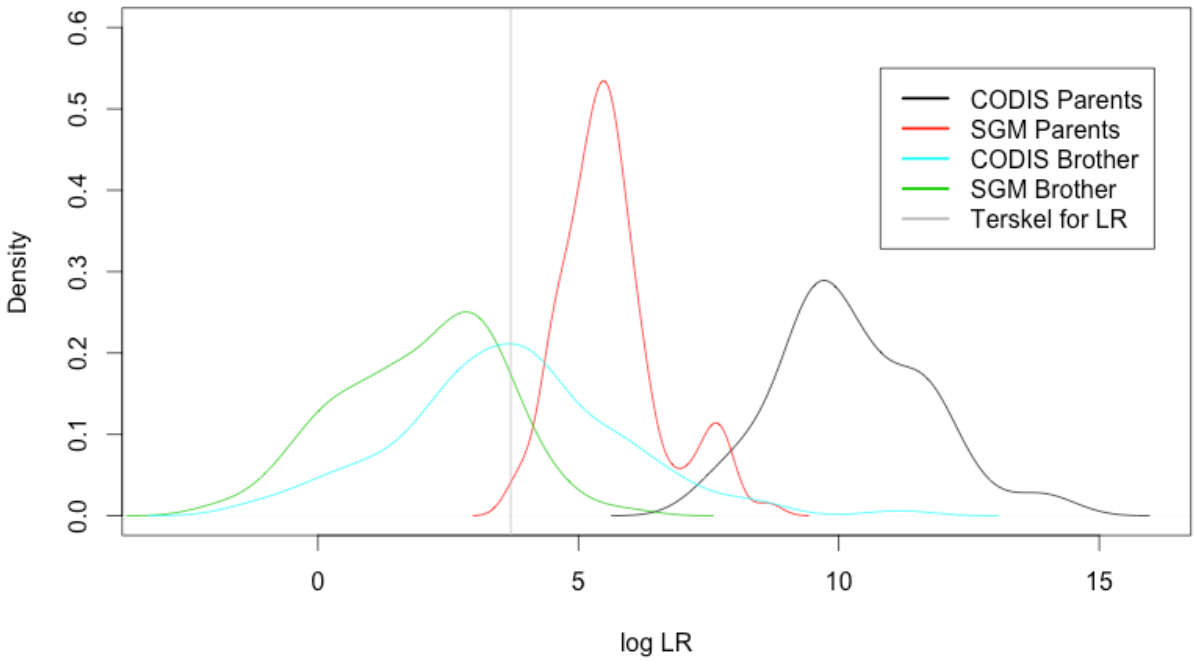
I søk mellom person 16 og 17, er LR-verdien den samme i søk etter relasjonen søsken, som i søket i avsnitt 3.1.2, som tilsvarer «Family ID» 2 og MP_2, da det er basert på de samme beregningene. Verdt å legge merke til av resultatene i Tabell 3.7, er at søket gir forholdene søsken og foreldre-barn samme LR-verdi. At den er akkurat den samme i dette tilfellet er tilfeldig basert på allelfrekvensene, men ved å se på de resultatene fra søkene med de andre uidentifiserte personene hvor en bror er simulert som referanseperson, kan det ses at det er vanskelig ut fra blindsøket å

skille mellom forholdene søsken og foreldre-barn. I disse tilfellene vil ytterligere informasjon være nødvendig for å kunne vite med sikkerhet relasjonen mellom disse, for eksempel ved å anslå alder ut ifra obduksjon, eller ved bruk av mtDNA for å finne morslinjen, da søsken vil ha likt mtDNA, mens far og barn vil ha ulikt.

3.3 Styrkeberegning

Som beskrevet i avsnitt 2.6, er det foretatt en styrkeberegning for å se hvor godt beregningene gir støtte til H_p når denne er sann. Basert på de samme dataene som er brukt i forsøkene i avsnittene 3.1.1 og 3.1.2, er også her LR-verdiene log-transformert og plottet ved hjelp av Kernel tetthetsfordeling, og dette vises i Figur 3.10. Terskelen for LR er fortsatt 5000, og log-verdien av denne er markert i grått. Skriptet for hvordan dette er gjort i R, er å finne i Vedlegg 9.

Ettersom det ikke finnes noen LR-verdier som overgikk LR-terskelen blant de usanne familietrærne, er disse utelatt i de videre beregningene. Det finnes derfor ingen falske positive blant beregningene og hypotesetestingen gjort med simuleringsdataene.



Figur 3.10: LR-verdiene for de sanne slektstrærne ved hvert scenario, med terskel for log LR≈3.7.

Tetthetsfordelingene av LR-verdiene ved de ulike scenarioene i Figur 3.10 viser at i alle LR-verdiene for simuleringer med foreldre og de 13 CODIS-markørene, som eneste scenario, befinner seg over terskelen for LR. De fleste LR-verdiene ved bruk av foreldre, men med færre markører i systemet SGM, gjør at sannsynlighetsspekteret også går noe under terskelen for LR, mens omtrent halvparten av LR-verdiene ved simulering av DNA-profiler for helsøsken basert på systemet CODIS sannsynligvis vil indikere informativ identifisering. Flesteparten av LR-verdiene vil ut ifra beregningene i denne oppgaven ikke ha høy nok LR-verdi til å indikere slektskap i tilfeller med en bror som referanseperson og kun de 6 markørene fra SGM, hvor kun 12 treff ga en LR-verdi over terskelen for LR.

Videre ses det på sannsynligheten for å beholde H_p der den er sann og terskelen for LR, altså styrken ved de ulike scenarioene. Som beskrevet i avsnitt 2.6, er styrkeberegninger i forensisk vitenskap annerledes enn klassisk styrkeberegning. Basert på LR-verdiene ved de ulike scenarioene er

sannsynligheten for at de er over LR-terskelen estimert, med tilhørende standardavvik og 95% konfidensintervall. Sannsynlighetene er formulert som følger:

$$1: P = P(LR \geq 5000 | H_0, CODIS, Foreldre)$$
$$2: P = P(LR \geq 5000 | H_0, SGM, Foreldre)$$
$$3: P = P(LR \geq 5000 | H_0, CODIS, Bror)$$
$$4: P = P(LR \geq 5000 | H_0, SGM, Bror)$$

Utregningene av estimatet p̂, standardavvik og 95% konfidensintervall er å finne i Vedlegg 10, mens resultatene finnes i Tabell 3.8.

Tabell 3.8: Estimert sannsynlighet for å ikke forkaste H_0 ved de ulike scenarioene, med tilsvarende standardavvik og konfidensintervall.

Scenario	p̂	SE(p̂)	95% KI
1	1	≈0	≈1
2	1	≈0	≈1
3	0.5	0.05	[0.402,0.598]
4	0.12	0.032	[0.057,0.183]

Basert på simuleringene av foreldre, estimeres det at alle sanne familietrær vil få en LR-verdi høyere enn terskelen for LR, både ved CODIS og SGM. Ved bruk av bror som referanseperson, er derimot verdiene lavere, og dermed færre tilfeller som vil karakteriseres som informativ identifikasjon. Halvparten av simuleringene ga en LR-verdi over 5000, og det er estimert med 95% sannsynlighet at ved bruk av bror som referanseperson, vil mellom 40.2% og 59.8% av de sanne familietrærne få en LR-verdi over denne terskelen. Kun en andel på 0.12 av simuleringene med en bror som referanseperson og systemet SGM ga en LR-verdi over terskelen ved de sanne familietrærne. Dermed estimeres det at i 8,8% til 15,2% av tilfellene med bror som referanseperson

og systemet SGM vil man beholde H_p og kunne indikere sann identifisering, basert på de 100 simuleringene.

4 DISKUSJON

I denne oppgaven er formålet å se på de statistiske beregningene knyttet til identifisering av personer ved store ulykker ved hjelp av DNA-profiler fra den ukjent og antatte familiemedlemmer. Å velge de mest informative slektskapscenarioene vil potensielt føre til sann identifisering, og redusere kostnader ved å minimere overflødig testing. En teoretisk bakgrunn er gitt, mens beregningene er utført i Familias.

4.1 Terskel for LR

Før en skal foreta en DNA-basert identifisering må det settes en statistisk terskel. I artikkelen av Ge et al. (2011) anses log LR-verdier på over 6 som indikasjon på informativ identifikasjon, mens under identifiseringsarbeidet etter 9/11 krevde en LR-verdi over 10,000,000,000, altså log LR=10 (Gonzales et al. 2006).

I denne oppgaven er terskel på LR satt til 5000 (log-verdi≈3.7), noe som er den LR-terskelen som brukes i praksis i for eksempel Ungarn. Det kan diskuteres om denne er høy nok, eller om det bør kreves forskjellige LR-terskler for forskjellige indentifiseringssakene for å opprettholde forutsatte grenser for spesifisitet og sensitivitet, da det er forskjell mellom DNA-basert identifisering i farskapstesting, familiesøk, immigrasjonssaker og DVI. Det er ønskelig å sette en så høy LR-terskel som mulig for å unngå at falske familietrær inkluderes, men samtidig ikke så høy at sanne familietrær utelates. Det er derfor viktig å finne en balanse, og LR-verdien kan variere mellom ulike anvendelser, som farskapstesting, immigrasjonssaker og DVI.

Terskelen på 5000 som benyttes i denne oppgaven, gjør at alle treff hvor foreldre benyttes som referansefamilie indikerer informativ indikasjon. For søsken derimot, er ikke LR-verdien for alle sanne treff over terskelen for LR, men ekskluderer alle falske treff. Dersom en lavere terskel skulle bli satt for å inkludere flere sanne familietrær, vil det kunne gå på bekostning av å ekskludere de

falske. Terskelen benyttet i denne oppgaven ser derfor ut til å være en rimelig verdi, da den utelater alle falske positive som er den alvorligste typen feil.

4.2 Hensiktsmessig valg av referansepersoner

Bakgrunnen for beregning av LR for at to personer er søsken er beskrevet i Avsnitt 2.5.4, og Tabell 2.3 viser at det er stor variasjon av LR avhengig av hvor mange alleler som deles IBD mellom personene. For hver av markørene, vil LR-verdien blir lavere enn 1 dersom ingen alleler deles, og høyere dersom begge deles, da forventningsverdien er at de deler ett allel (Buckleton et al. 2005). Det er uvanlig at et søskenpar ikke deler noen alleler IBD over flere locus, men det kan forekomme, og en LR for et tilfeldig treff med et ubeslektet individ kan dermed bli høyere. Dersom profilen fra en savnet person har flere loci hvor ingen alleler deles med et søsken, vil LR bli lavere enn 1, og den alternative hypotesen om at individet ikke er et søsken av den savnede støttes.

Beregningene er Familias viser stor variasjon i LR-verdiene til søskenparene simulert og beregnet i Familias. Gjennomsnittet på 4836,15 er lavere enn terskelen for LR som er satt som grense for informativ identifisering, og i en del av tilfellene er det et tilfeldig treff med andre referansefamilier som gir en høyere LR. Den store variasjonen blant LR-verdiene ved helsøsken-scenarioet skyldes den vide IBD-distribusjonen.

Simuleringene og beregningene gjort i artikkelen av Ge et. al. (2011) er gjort med samme betingelser som i denne oppgaven. Tabell 1 i artikkelen presenterer deres resultater, og viser at bruk av helsøsken gir den laveste gjennomsnittlige LR-verdien blant nær familie (altså ikke besteforeldre, onkler, søskenbarn eller halvsøsken). Bruk av begge foreldrene som referansepersoner gir blant de høyeste gjennomsnittlige LR-verdiene, hvor kun bruk av 3 barn og ektefelle, samt 4 barn gir høyere LR-verdi (Ge et al. 2011). Mens gjennomsnittlig LR-verdi for treff mellom uidentifisert person og foreldre er $3.47 \cdot 10^6$ ganger større enn gjennomsnittlig LR-verdi for treff mellom uidentifisert person og bror med markørsystemet CODIS, er tilsvarende forhold i beregningene av Ge et al. $6.68 \cdot 10^6$. Gjennomsnittlig LR og varians i denne oppgaven er

noe høyere enn i artikkelen, noe som kan være tilfeldig og uavhengig av programmet, ettersom det er gjort færre simuleringer i denne oppgaven enn i artikkelen.

Dersom man har flere søsken enn ett, vil det være større sjanse for å finne ut fra hvilke alleler den uidentifiserte har sitt opphav, selv om man ikke kan finne ut hvilke alleler som kommer fra mor og hvilke som kommer fra far. Når alle de fire allelene er observert blant de referansesøsknene, vet man hvilke fire alleler som kommer fra foreldrene, som sammen med allelfrekvensen vil kunne gi høyere LR-verdier.

Det er viktig å bruke de mest informative referansepersonene, og antall slektninger kan variere. Å utarbeide en DNA-profil for alle vil ofte være overflødig, da det vil medføre større kostnadsbruk uten å nødvendigvis øke LR-verdien nevneverdig. Bruk av kun en referanseperson vil ofte ikke være tilstrekkelig, da det ikke vil medføre store nok LR-verdier til å kunne indikere slektskap. Det er intuitivt fordelaktig å ha minst mulig genetisk avhengighet mellom referansepersonene, for å bevare mest mulig informasjon. Dette gjør at DNA-profiler fra foreldre gir høyere LR-verdi fremfor søsken, da foreldre som oftest ikke er beslektet. Ge et al. (2011) hevder at det trengs syv helsøsken i tillegg til en av foreldrene for å kunne oppnå lignende LR-verdier som to foreldre gir.

4.3 Hensiktsmessig antall markører

I resultatene i Avsnitt 3.1.1 er det beregnet at simuleringene i oppgaven gir 46 773 ganger høyere gjennomsnittlig LR-verdi ved bruk av CODIS i forhold til SGM. Forsøkene utført i denne oppgaven viser at SGM-databasen på generelt basis ikke inneholder nok markører til å gi høye nok LR-verdier til å skille mellom nært beslektede og ubeslektede individer. Lav LR-verdi i både scenarioet der foreldre benyttes i slektsskapsanalyser og ved bruk av et søsken, i dette tilfellet en bror, viser at flere familietrær ekskluderes der det i virkeligheten er sant, noe som indikerer at et større antall markører bør benyttes for å øke LR-verdien.

Ved bruk av CODIS-systemets 13 markører viser slektskapsanalyser med bruk av foreldre som referansepersoner at alle familietrærne får høy nok LR-verdi til å kunne indikere slektskap, mens gjennomsnittet ved bruk av brødre ikke ga høy nok LR-verdi til å overstige terskelen og indikere slektskap, og dermed bli ekskludert. For videre tester vil det kunne være interessant å finne ut om et enda større antall markører vil være hensiktsmessig for å kunne heve gjennomsnittlig LR-verdi for slektskapsanalyser et søsken er involvert, eller om dette også vil gjøre at flere falske familietrær vil inkluderes.

Til tross for at et større antall markører vil øke styrken på identifiseringen, kan det være vanskelig å gjennomføre i praksis, da kvaliteten og kvantiteten på DNA-materialet funnet i levningene etter de uidentifiserte personene kan være begrenset. Er det nok DNA tilgjengelig, kan flere markører benyttes, men det er ikke tilfellet i mange DVI-saker (Ge et al. 2011).

4.4 Blindsøk

Et blindsøk etter relasjoner mellom de uidentifiserte er et nyttig verktøy i tilfeller der det antas at enkelte av de uidentifiserte personene er beslektet. Dersom man ikke vet med sikkerhet hvilken relasjon man skal søke etter, kan det likevel være vanskelig å skille mellom de ulike forholdene mellom de uidentifiserte basert på LR-verdien, dersom man ikke har noen annen informasjon tilgjengelig. At det indikeres slektskap kan likevel benyttes i videre søk. Dersom man i tillegg til de uidentifiserte personene med antatt slektskap har referansepersoner, kan et søk med antatte familietrær være oppklarende.

I noen DVI-saker vil man møte på utfordringen å ha flere DNA-profiler enn savnede personer, som følge av at det finnes flere levninger fra samme person. I disse tilfellene vil et blindsøk kunne avgjøre hvilke DNA-profiler som tilhører samme person, og slå disse sammen før videre søk. Det vil derfor være hensiktsmessig å foreta et blindsøk før man søker etter slektskap gjennom referansepersoner.

4.5 Antall simuleringer

På grunn av tidsbegrensinger som følge av tidkrevende etterarbeid av simuleringene i Excel, er det kun foretatt 100 simuleringer for hvert scenario. Forsøket kan med fordel gjentas med flere simuleringer for å få flere DNA-profiler og mer pålitelig resultater. Særlig gjør dette seg gjeldene ved simulering av DNA-profiler som benyttes i familietrær med større spekter på LR-verdiene, som i tilfellet med helsøsken i denne oppgaven. Da vil potensielt et større antall simuleringer kunne øke tettheten av datapunktene og dermed jevne ut kurvene, noe som kan gjøre en direkte visuell sammenligning av testene mer pålitelig.

Til tross for et færre antall simuleringer og familietrær, viser forsøkene med de ulike identifikasjonsscenarioene samme trender blant gjennomsnittlig log LR-verdi og tilhørende varians som beskrevet i artikkelen av Ge et al. (2011), hvor antall simuleringer er mellom 10 000 og 1 000 000. Ved å sammenligne gjennomsnittlig log LR-verdi ved bruk av foreldre med de samme betingelsene i oppgaven og artikkelen i en Welch t-test, vil det ikke være grunnlag til å forkaste hypotesen om at log LR-gjennomsnittene er forskjellige på et 95%-konfidensnivå.

4.6 Andre kombinasjoner av referanseslektninger

Dersom nære slektninger ikke er tilgjengelige som referansepersoner, kan fjernere slektninger måtte tas i bruk i identifiseringsarbeidet. Dersom det genetiske slektskapet blir fjernere, vil antall falske positive treff øke, ettersom det deles færre alleler IBD. I disse tilfellene kan det være informative å ta i bruk mtDNA og/eller Y-STR, ettersom det kan spore slektskap over flere generasjoner på henholdsvis mors og fars avstamming. Bruk av disse markørene i tillegg til STR-locus vil kunne øke LR-verdien.

4.7 Enegget tvilling og nære slektninger

Dersom begge personene et enegget tvillingpar er utsatt for samme ulykke, vil det være vanskelig å kunne identifisere dem. De vil da ha den samme DNA-profilen, sett bort ifra eventuelle mutasjoner. I forhold til å bruke referansepersoner fra familien vil det være mulig å identifisere dem som et par, men ikke kunne skille dem som to individer. Dette gjør seg også gjeldende ved bruk av blindsøk, da programmet vil anta personene som samme person.

Et lignende problem vil oppstå der det er flere personer som er beslektet. For eksempel vil en onkel og nevø kunne gi høy LR-verdi med de samme referansepersonene, og en høy nok LR-verdi til å indikere slektskap, ettersom de har en del av de samme genotypene som for eksempel far/bror. Dersom man sidestiller disse to i forhold til hverandre kan forskjellen i LR-verdi muligens ikke bli høy nok til å kunne avgjøre med sikkerhet hvem som er hvem. Et slikt scenario vil kunne kreve flere referansepersoner, der en eller flere kun er beslektet til den ene, og ved å benytte seg av mtDNA eller Y-STR på motsatt side av familien til eventuell onkel.

4.8 Identifiseringspraksis i Norge

Identifiseringsgruppen i Norge er lagt til Kripos, og ble opprettet ved kgl. resolusjon av 25.4.1975. Ved siden av rettsgenetiker, består gruppen av administrativt personell, kriminalteknikere, rettspatologer og rettsodontologer, som skal søke å samle opplysninger om den savnede og de opplysningene som er fremkommet om den avdøde under sakkyndige likundersøkelser. Under identifiseringsarbeidet skal gruppen sikre bevis som kan være av betydning for å fastslå årsaks- og ansvarsforhold ved katastrofen, som kan benyttes videre ved en eventuell rettergang dersom det er mistanke om noe kriminelt (Politiet.no 2013). Retten kan oppnevne uavhengige sakkyndige til å bistå under granskning, som skal fungere som rettens rådgiver. Den sakkyndige utformer en sakkyndigrapport som formulerer en konklusjon, og må i noen tilfeller bistå retten.

Den sakkyndige bør presentere sine observasjoner og vurderinger uten å ta parti for eller mot noen

personer i saken, og slik at de blir best mulig forståelig for juristene og rettens legfolk (Rognum 2010). Ved fremlegging av resultater etter identifisering av personer ved massekatastrofer, ønsker man også at resultatene skal være forståelig for pårørende. Dersom resultatene presenteres som en LR-verdi, som er mest vanlig i arbeid med DVI, vil den være vanskeligere å tolke enn dersom en *a posteriori*-sannsynlighet beregnes. Vanligvis presenteres resultatene av identifiseringen som en LR-verdi, ettersom det er vanskelig å finne korrekte *a priori*-sannsynligheter for beregning av *a-posteriori*-sannsynligheter. I Norge er det ikke utarbeidet noen standard for hvilke ord som skal benyttes for å beskrive resultatene av LR, men en nyansert verbalisering, tilsvarende tabell 2.3 på side 40 i Buckleton et. al. (2005), kan med fordel utarbeides for slik at det finnes en felles terminologi (Egeland 2009).

4.9 Treff med eksisterende database

I Familias er det som nevnt, i tillegg til DVI-mode også mulig å foreta et blindsøk som tester sannsynligheten for at to personer i det innlastede datasettet er samme person eller i familiær relasjon til hverandre. Maguire et. al. (2014) har tatt for seg hvordan et slikt søk kan brukes til å finne treff mellom prøver og den eksisterende nasjonale DNA-databasen (NDNAD) og hvordan dette kan hjelpe arbeidet i kriminalsaker, men påpeker også ulike etiske utfordringer et slik søk kan medføre, og hvordan praksisen på dette området er i ulike land. I artikkelen fokuseres det på hvordan familiære søk kan gjennomføres, og ikke et direkte treff som man vanligvis er ute etter ved å søke opp mot databasen som inneholder DNA-profilene til kjente lovbrytere.

Søk etter familier baserer seg på at beslektede individer har større sannsynlighet for å dele alleler i et locus enn ubeslektede, og at foreldre og barn deler alleler på en spesifikk måte, og dette kan benyttes til å løse de vanskeligste sakene, ved å tilføre ny informasjon til fastlåste saker, og på den måten være et verdifullt verktøy for rettssystemet. Sammenlignet med DVI, vil de sammen beregningene av DNA-bevis foretas i et familiesøk for å identifisere nære biologiske slektninger.

Det er ulik praksis rundt bruk av familiesøk i forskjellige land. UK har den største DNA-databasen

i verden per innbygger, og denne har blitt et viktig verktøy i etterforskning og avdekking av kriminalsaker, hvor familiesøk har blitt bruk i rundt 210 saker siden det ble introdusert i 2002. Familiesøk har også blitt foretatt i Nederland og USA, men reglene for når slike søk er tillatt varierer mellom landene og statene.

Problematikken rundt emnet dreier seg om personvernet rundt både personer allerede i databasen NDNAD, samt mulige slektninger av personene i databasen, som ellers ikke ville kommet i politiets søkelys. Samtidig vil denne søkemetoden kunne virker forsterkende på synspunkter eller fordommer om den påståtte utbredelsen av kriminalitet innenfor enkelte familier. Et databasesøk vil kunne avsløre slektninger og finne en genetisk link mellom to personer, som tidligere ikke er oppdaget, altså avsløre nye familieforhold, eller i motsatt fall, avkrefte en genetisk link man trodde var der (Maguire et al. 2014).

Ved å foreta et blindsøk, som i Familias, vil det være mulig å finne ut om ulike DNA-profiler kommer fra samme person, noe som ikke vil være utenkelig ved masseulykker med store personskader. Dette vil hjelpe arbeidet med å finne rett antall personer i ulykken dersom man har å gjøre med en ulykke der man ikke har en fullstendig liste over ofrene, samt å få identifisert alle kroppsdeler. Om det finnes et slektskap mellom to eller flere av personene i ulykken, vil det være mulig å finne en link før man i det hele tatt ser på familietresannsynlighetene, og på den måten gjøre at man forventer å finne høyere LR-verdier.

En del av de samme betenkelighetene det nevnes ved å bruke et familiesøk til å finne lovbryter, kan også overføres til å bruke NDNAD til søk opp mot savnede personer. Familiesøk kan også være til hjelp for å unngå en mer kostnad prosedyre enn nødvendig, ved å benytte seg av allerede eksisterende DNA-profiler, fremfor å utarbeide nye. Det vil derfor kunne være et sterkt verktøy i bekjempelsen av kriminalitet, men må veies opp i forhold til personvern.

4.10 Videre arbeid

I innledningseksempelet ble det påpekt at det benyttes den marginale tilnærmingen som er implementert i Familias. For å unngå situasjoner hvor en referansefamilie kan få en *a posteriori*-sannsynlighet over 50% for flere uidentifiserte personer, kan en simultan metode benyttes. Dette er ikke analysert i denne oppgaven, men kan være aktuelt å se nærmere på for å eventuelt forbedre resultatene for hele søket. På den måten kan de «sikre» treffene, for eksempel de tilfellene som gir en LR-verdi over LR-terskelen fjernes fra søk med de andre referansefamiliene. Dette vil ta lenger tid, men kan gi forbedret resultat.

I innledningen ble også *a priori*-sannsynligheten nevnt, som i beregninger i denne oppgaven er den samme for alle individene, og det har blitt valgt å se på LR-verdien. Dersom programmet benyttes på virkelige data, vil det kunne være interessant å se på innvirkningen der kjent informasjon benyttes for å unngå å miste informasjon i arbeidet og presentasjonen. For eksempel vil det etter en obduksjon i enkelte tilfeller være mulig å si noe om den uidentifiserte er et barn eller en voksen, noe som kan benyttes i beregningene. DVI-modulen i Familias beregner *a priori*-sannsynligheten basert på antall savnede personer, og det er derfor vanskelig å inkludere eventuell annen informasjon. Dette gjelder også i et lukket problem med et kjent antall uidentifiserte personer, hvor størrelsen og beregningene tar utgangspunkt i at også en ukjent, uidentifisert person kan være slektning av referansepersonene. For et lukket system, hvor man vet hvilke uidentifiserte personer som er inkludert, vil man derfor ikke ha behov for å inkludere denne ukjente når man skal angi *a priori*-sannsynligheten.

I denne oppgaven er det benyttes to typer familietrær, og for videre å se på hvor godt programmet Familias presterer vil det være interessant å se på også enda flere typer familietrær. Det er varierende grad av informasjon fra DNA-profiler til referansepersoner, men for bedre se hvilke kombinasjoner av nært og fjernere beslektede personer som kan gi et riktig resultat vil kunne være til hjelp for å unngå unødvendig stort ressursbruk, og dermed være kostnadsbesparende.

Ved bruk av blindsøk-modulen i denne oppgaven avsluttet programmet ved søk etter å danne trioer

blant de uidentifiserte. Dette er i utgangspunktet en funksjon som skal fungere i Familias, men som under arbeid med oppgaven ikke var ferdig implementer i programmet. Dersom søket hadde blitt gjennomført med de samme personene som nevnt i Avsnitt 3.3, ville man kunne forvente trioer med høy LR-verdi mellom de personene som er simulert i samme familietre, tilsvarende LR-verdiene i resultatene i Avsnitt 3.1.1.

Ettersom simuleringene og beregningene i denne oppgaven ga varierende LR-verdier med bruk av bror som referanseperson, vil det kunne være interessant å se på i hvor stor grad bruk av Y-STR eller mtDNA til tillegg til STR-analyse vil kunne påvirke LR-verdien. Muligens kombinasjonen i en del tilfeller kunne øke LR-verdien såpass at man kan konkludere med slektskap mellom dem.

For videre testing av Familias, bør virkelige data tas i bruk, der blant annet mutasjoner vil kunne ha innvirkning på LR-verdien.

5 KONKLUSJON

Hvilke referansepersoner som benyttes til identifisering av savnede personer, viste seg å ha størst utslag for resultatene i denne oppgaven, da alle LR-verdiene med foreldre som referansepersoner ble høyere enn terskelen for LR, uavhengig av system. Det er viktig å velge de mest informative individene i et familietre for å redusere kostnader i identifiseringsarbeid, og jo mindre avhengighet det er mellom referansepersonene, jo høyere LR-verdi vil man oppnå. Det er kommet frem til numeriske mål som viser betydning av å øke antall markører eller bruke mer informative referansepersoner, der gjennomsnittlig LR-verdi for sanne positive treff ved foreldre og markørsystemet CODIS er klart høyest. Dersom kvalitet og kvantitet tillater det, er det ønskelig å benytte seg av flere markører. Beregningene i Familias viser at programmet gir lignende LR-verdier som programmer benyttet i annen relevant litteratur. Det er viktig å sette en terskel for LR slik at falske positive treff utelates, nemlig at en uidentifiserte personer antas å være beslektet til en familie de i virkeligheten ikke er beslektet til ikke indikeres som beslektet. Samtidig er det ønskelig få flest mulige treff mellom den uidentifiserte personen og den sanne referansefamilien med høyere LR-verdi enn terskelen. En LR-verdi på 5000 utelater alle falske positive treff av simuleringene i oppgaven.

REFERANSER

Berggreen, M. (2013). *Familiegenetikk og mutasjoner: betydningen av modellvalg*. Ås: NMBU, Institutt for kjemi og bioteknologi. 51 s.

Buckleton, J., Triggs, C. M. & Simon, J. W. (2005). *Forensic DNA evidence Interpretation*. Boca Raton: FL: CRC Press. 534 s.

Egeland, T. (2009). Statistisk vekting av DNA-funn i straffesaker. *Tidsskrift for Strafferett*, 2: 190-204.

Egeland, T. & Mostad, P. (2010). Manual Familias. 52. Tilgjengelig fra: familias.name/manual.pdf.

Fletcher, H., Hickey, I. & Winter, P. (2007). *Genetics*. 3 utg. New York, Abingdon: Taylor & Francis Group. 379 s.

Fung, W. K. & Hu, Y.-Q. (2008). Parentage testing. I: *Statistical DNA Forensics: Theory, Methods and Computation*, s. 262. Chichester, Great Britain: John Wiley & Sons LTD.

Ge, J., Budowle, B. & Chakraborty, R. (2011). Choosing Relatives for DNA Identification of Missing Persons. *Journal of Forensic Sciences*, 56: 6.

Gonzales, A. R., Schofield, R. B. & Schmitt, G. R. (2006). Lessons Learned From 9/11: DNA Identification in Mass Fatality Incidents: Office of Justice Programs. 68 s.

Interpol. (2009). *DVI guide*. Tilgjengelig fra: http://www.interpol.int/INTERPOL-expertise/Forensics/DVI-Pages/DVI-guide (lest 21.04.2014).

Lesk, A. M. (2012). *Introduction to genomics*. 2. utg. New York, U.S.: Oxford University Press. 397 s.

Liland, K. H. (2014). *NMBU/UMB R repository*: The R Project. Tilgjengelig fra: http://repository.umb.no/R/ (lest 04.04.2014).

Løvås, G. (2011). *Statistikk for universiteter og høgskoler*. 2. utg. Oslo: Universitetsforlaget 2004. 489 s.

Maguire, C. N., McCallum, L. A., Storey, C. & Whitaker, J. P. (2014). Familial Searching: A specialist forensic DNA profiling service utilising the National DNA Database to identify unknown offenders via their relatives - The UK experience. *Forensic Science International: Genetics*, 8: 1-9.

Mostad, P., Egeland, T. & Kling, D. (2013). *Familias 3*. 3 utg.: Norwegian Institute of Public Health.

Olaisen, B., Stenersen, M. & Mevåg, B. (1997). Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster. *Nature Genetics*, 15: 402-405.

Politiet.no. (2013). *Kripos, Identifiseringsarbeid*. Tilgjengelig fra: https://www.politi.no/kripos/identifiseringsarbeid/ (lest 30.04.2014).

Prinz, M., Carracedo, A., Mayr, W. R., Morling, N., Parsons, T. J., Sajantila, A., Scheithauer, R., Schmitter, H. & Schneider, P. M. (2007). DNA Commision of the International Society for Forensic Genetics (ISFG): Recommendations regarding the role of forensic genetics for disaster victim identification (DVI). *Forensic Science International: Genetics*, 1 (1): 3-12.

Rognum, T. O. (red.). (2010). *Lærebok i rettsmedisin*. 2. utg. Oslo: Gyldendahl Norsk Forlag AS. 491 s.

The R Project. (2013). *Bandwidth Selectors for Kernel Density Estimation*: The R Project. Tilgjengelig fra: http://stat.ethz.ch/R-manual/R-patched/library/stats/html/bandwidth.html (lest 04.04.2014).

The R Project. (2014). *What is R?*: The R Project. Tilgjengelig fra: http://www.r-project.org/ (lest 04.04.2014).

Vigeland, M. D., Selmer, K. K. & Egeland, T. (2012). Statistical methods in genetics. I: Veierød, M. B., Lydersen, S. & Laake, P. (red.) *Medical Statistics*, s. 38: Gyldendal Forlag.

Zietkiewicz, E., Witt, M., Daca, P., Zebracka-Gala, J., Goniewicz, M., Jarzac, B. & Witt, M. (2011). Current genetic methodologies in the identification of disaster victims and in forensic analysis. Tilgjengelig fra: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3265735/ (lest 31.03.2014).

VEDLEGG 1: MARKØRER OG ALLELFREKVENSER FOR SYSTEMET CODIS OG SGM

Tabell 1: Markørene og deres plassering på kromosomene benyttet ved henholdsvis CODIS og SGM.

Locus	Plassering på kromosomet	CODIS	SGM
CSF1PO	5q33.3-34	X	
FGA	4q	X	X
TH01	11p15.5	X	X
TPOX	2p13	X	
VWA	12p	X	X
D3S1358	3	X	
D5S818	5q21-31	X	
D7S820	7	X	
D8S1179	8	X	X
D13S317	13	X	
D16S539	16	X	
D18S51	18	X	X
D21S11	21	X	X

*Ved siden av markørene som fremgår av tabell 1 er egentlig AMEL en del av systemet. I beregningene i oppgaven fremkommer det hvilket kjønn personene er, men er ikke en del av beregningene. SGM er derfor benyttes for å illustrere innvirkning av antall markører, og AMEL er for enkelhets skyld utelatt.

CSF1PO

8	0.00496995
9	0.0115899
10	0.216888
11	0.301317
12	0.360926
13	0.096029
14	0.00827992

FGA

18	0.02649
19	0.05298

20	0.12748
21	0.18543
21.2	0.00497
22	0.21854
22.2	0.01159
23	0.13411
23.2	0.00331
24	0.13576
24.2	0.00166
25	0.07119
26	0.02318
27	0.00331

TH01

5	0.00165997
6	0.231785
7	0.190396
8	0.0844383
9	0.114238
9.3	0.367543
10	0.00827983
11	0.00165997

TPOX

5	0.00165997
6	0.00165997
8	0.534759
9	0.119208
10	0.0562889
11	0.243375
12	0.0413892
13	0.00165997

VWA

13	0.00165998
14	0.0943691
15	0.110929
16	0.200328
17	0.281457
18	0.200328
19	0.104299
20	0.00496995
21	0.00165998

D3S1358

11	0.00165998
14	0.102649
15	0.261587
16	0.253307
17	0.215228
18	0.152318
19	0.0115899
20	0.00165998

D5S818

7	0.00165998
8	0.00330997
9	0.0496695
10	0.0513195
11	0.360926
12	0.384106
13	0.140729
14	0.00661993
15	0.00165998

D7S820

7	0.01821
8	0.15066
8.1	0.00166
9	0.17715
10	0.24338
11	0.20695
12	0.16556
13	0.03477
14	0.00166

D8S1179

8	0.01159
9	0.00331
10	0.10099
11	0.08278
12	0.18543
13	0.30464
14	0.16556
15	0.11424
16	0.03146

D13S317

8	0.112582
9	0.0745015
10	0.051321
11	0.339407
12	0.248345
13	0.124172
14	0.048011
15	0.00166003

D16S539

8	0.01821
9	0.11258
10	0.05629
11	0.32119
12	0.32616
13	0.1457
14	0.01987

D18S51

10	0.00827988
11	0.0165598
12	0.127479
13	0.132449
14	0.137419
14.2	0.00165998
15	0.158938
16	0.139069
17	0.125829
18	0.0761588
19	0.0380796
20	0.0215198
21	0.00827988
22	0.00827988

D21S11

25.2	0.00165997
27	0.0264895
28	0.158937
29	0.195356
29.2	0.00330993
30	0.278144
30.2	0.0281494
31	0.0827783
31.2	0.099338
32	0.00661987
32.2	0.0844383
33	0.00165997
33.2	0.0264895
34.2	0.0049699
35	0.00165997

VEDLEGG 2: FREMGANGSMÅTE FOR SIMULERINGER I FAMILIAS

1. Programfilen Familias 3 lastes ned gratis fra familias.no. Åpne programmet.
2. Finn «Advanced» under rullegardinen «File», og huk av «Save genotype data».
3. Trykk på General DNA data (Ctrl+G). Her legges alleler og allelfrekvenser inn. Dersom man har et eksisterende DNA-system på egen fil kan den importeres ved hjelp av import. Filene CODIS og SGM er lastet inn i eksemplene i denne oppgaven, informasjon om disse systemene er å finne i vedlegg 1. Her legges også eventuell informasjon om mutasjonsmodellene inn, noe som ikke er benyttet i denne oppgaven. Importer eller legg inn allelinformasjonen manuelt, og trykk på «save».
4. Trykk på «Persons» (Ctrl+E). Det er i denne oppgaven foretatt to simuleringer, der henholdsvis DNA-databasen CODIS og SGM er benyttet. Fyll inn navn, hvilken database som skal benyttes og kjønn. I simuleringene brukt i denne oppgaven vil det si:

Name	Database	Gender
MP (missing person)	CODIS/SGM	Male
Mother	CODIS/SGM	Female
Father	CODIS/SGM	Male
Brother	CODIS/SGM	Male

   Når alle personene er definert og lagt til, lukk vinduet.
5. Under «Pedigrees» (Ctrl+P), legg til minst to pedigreer. Finn «Add» under «Actions», og definer det første pedigreet. Under «Add relations» vil «Mother» angis som forelder til «MP» og legges til, deretter «Father» som forelder til «MP». Trykk «Close» når første pedigree er ferdig definert. Trykk «Add» nok en gang for å definere pedigree nummer 2. Her utvides det samme pedigreet, hvor «Mother» og «Father» er foreldre til «MP», og «Mother» og «Father» angis også som foreldre til «Brother». Det er ikke mulig å definere direkte at «MP» og «Brother» er brødre, uten å legge inn personer som er deres foreldre. Trykk «add» for å legge til det andre pedigreet.
6. Under «Actions» finnes også knappen «Simulate». Dobbeltklikk på de personene som det skal simuleres en DNA-profil til, slik at disse havner under kolonnen «Will be genotyped». Angi antall

simuleringer, og velg enten «random seed» eller skriv inn ønsket seed. Det er i denne oppgaven gjort 100 simuleringer, og benyttet seed 500 og 501 for henholdsvis simuleringer av genotypene til foreldre og bror. Trykk «simulate».

7. Lagre resultatene. Dersom genotypdataene skal benyttes videre i Familias, må de bearbeides i Excel/Notepad, som er beskrevet i Vedlegg 3.

VEDLEGG 3: BEARBEIDING AV GENOTYPEDATA FRA FAMILIAS I EXCEL/NOTEPAD

For å kunne laste DNA-informasjonen inn i Familias igjen, og for lettere å kunne definere familietrær i DVI-mode må filene sorteres i ulike filer og enkelte kolonner fjernes.

1. Det lagres to filer som følge av simuleringene i Familias, genotypedata og rawdata. Åpne genotypedata i Excel.
2. For å lagre genotypene til de savnede personene, filtrer først under Sample-kolonnen, slik at kun MP-radene vises. For å få de data for MP i familietreet, der det er simulerte data for to foreldre, filtreres også i kolonnen «True Pedigree», i tilfellet i denne oppgaven ble det «Ped 1». Kopier deretter alle rader som samsvarer med kolonnene «Ped 1» og «MP» inn i et nytt dokument. Lagre det nye dokumentet som «MP_parents.txt», Tekst (tabulatordelt).
3. I det nye dokumentet, gi nytt navn til den første savnede personen, kall denne MP_1. Marker ruta, og dra nedover kolonnen, slik at den savnede personene får navn MP_X, der X er i stigende rekkefølge.
4. Slett kolonnen «True Pedigree» for alle rader, og lagre filen.
5. Gå tilbake til genotypedataene i det opprinnelige dokumentet. Filtrer nå på det samme familietreet, altså fortsatt de som tilhører «Ped 1». Fjern MP, og huk av både «Mother» og «Father» i kolonnen «Sample». Kopier alle rader som samsvarer med kolonnene «Ped 1» og «Mother» og «Father» inn i et nytt dokument. Lagre det nye som dokumentet som «Parents.txt», tekst (tabulatordelt).
6. I det nye dokumentet, gi kolonnen «True Pedigree» nytt navn som «Family». Indiker hvilke personer som tilhører samme familie ved å nummerere de i stigende rekkefølge, slik at første «Mother» og «Father» er «Family» 1, andre «Family» to, osv, tilsvarende antall «MP».
7. I «Sample»-kolonnen, sett inn klammeparentes rundt navnet på familiemedlemmet, for eksempel [Mother] og [Father] for alle personene, og lagre filen. Dette gjøres for at familietrær automatisk lages når filen importeres til Familias.
8. I arbeidet med denne oppgaven ble alle alleler som innehold punktum automatisk omgjort til dato, for eksempel ble 9.3 omgjort til 09.mar. Dette ble lagret videre i de nye filene, og må derfor

endres før de importeres til Familias. For å gjøre dette, åpne de nye tekstfilene i for eksempel Notisblokk. Trykk rediger, og deretter erstatt. For eksempelet med 9.3, søk etter 09.mar og erstatt med 9.3, og trykk erstatt alle.

Samme fremgangsmåte ble også benyttet for brødre. En annen måte å effektivisere tidsbruken er ved å lage familietrær uten å bruke klammeparentes, er ved først å lage kun et familietre i Familias. Lagre deretter filen, og åpne programfilen i et redigeringsprogram (for eksempel Notisblokk) hvor man kan kopiere delen om omtaler pedigree til alle familiene, som deretter åpnes i Familias igjen.

VEDLEGG 4: FREMGANGSMÅTE FOR DVI-MODUL I FAMILIAS 3

1. Åpne Familias 3.exe.
2. Trykk på «General DNA data» for å legge inn markører og allelfrekvenser. For å manuelt legge inn ny markør, trykk på «Add» og legg til navn og frekvens for hvert allel, og trykk «Save». Gjør dette for hver markør. Markørsystemet må tilsvare systemet benyttet i DNA-data for personene som importeres, som i denne oppgaven er Codis, også benyttet i simuleringene. I denne oppgaven ble en eksisterende database med markører og allelfrekvenser benyttet ved å trykke på import, og finne den lagrede filen på maskinen. Denne filen må være lagret som en tekstfil. Under «General DNA data» legges også mutasjonsmodeller inn, ved å trykke på «Mutations», om samme modell skal benyttes for alle markører. Mutasjoner er ikke benyttet i denne oppgaven, og er derfor utelatt. Trykk deretter på «Close» for å lukke vinduet «General DNA data».
3. Under «Tools», og deretter «DVI module», trykk på «Add unidentified data» (Ctrl+L). Legg inn antall personer ved å gi personen et navn og angi kjønn, og deretter trykk «Add». Dobbeltklikk på personen for å legge til DNA-data for personen. Bla i rullegardinen for å finne riktig system og allel. Trykk deretter på «Close». Som beskrevet i Vedlegg 2 og 3, er en tekstfil med DNA-informasjonen til de uidentifiserte personene lagd, og kan dermed leses inn ved «Import». På den måten defineres alle de savnede personene ut ifra navnet i den første kolonnen i tekstfilen, og tilhørende alleler og kjønn angitt av de resterende.
4. Trykk «Next» for å komme til vinduet «Add reference families». For å legge til et nytt familietre, trykk på «Add». Legg deretter inn personene i familietreet under ruta «Persons», ved å angi navn og kjønn, også de man ikke har DNA-profil til, men som man trenger for å lage et fullstendig familietre. For eksempel må begge foreldrene legges inn om man skal lage et familietre med to brødre. Dobbeltklikk på personen for å angi allelinformasjonen for hvert system i rullegardinen. Når DNA-informasjonen er lagt til, trykk «Close». Under «Pedigrees» ligger allerede et «Reference pedigree» som er tomt. Dette benyttes for utregning av LR, og er tilsvarer hypotesen om at ingen av de uidentifiserte personene tilhører denne familien. Trykk på «Add» for å legge til et nytt familietre som det skal beregnes LR for. Velg foreldre og barn i rullegardinen for å danne et fullstendig familietre. Trykk «Close» når alle er lagt til familietreet. I arbeid med denne

oppgaven var allerede en tekstfil med DNA-informasjon til referansepersonene og familitrær for disse tillagd. Istedenfor å legge inn hver enkelt person, importeres dermed denne filen ved å trykke på «Data only», og finne den aktuelle filen for innlasting.

5. Trykk deretter «Next» for å komme til søkevinduet. Trykk så på «Search», og sett en «Match limit» for LR. I denne oppgaven ble denne grensen satt til 1 ved søk etter treff med begge foreldrene som referansepersoner, mens søk med bror som referanseperson krevde en LR på 0.01 for at de «riktige» treffene skulle vises i resultatene. Lagre en enkel rapport av resultatene.

For videre beregninger er de høyeste LR-verdiene for hvert «riktige» treff mellom uidentifisert person og referansepersoner manuelt plukket ut som lagres i en egen fil for beregninger i R. Det lagres også en fil med de treffene for hver familie som ga høyest LR-verdi, der treff med «feil» uidentifisert person ga høyere LR-verdi enn den «riktige».

VEDLEGG 5: FREMGANGSMÅTE FOR BLINDSØK I FAMILIAS 3

For å teste blindsøk-modulen i Familias, lagres de fem første familietrærne fra simuleringene med foreldre, de fem første med brødre i tillegg til at uidentifisert person nummer 2 fra brødre i en egen fil. Systemet CODIS er benyttet for alle personene. Kolonnene «True Pedigree» fjernes, og i kolonnen «Samples» nummereres personene fra 1-26 i stigende rekkefølge.

1. Importer eller legg til allelfrekvenser som beskrevet under punkt 2 i Vedlegg 4.
2. Importer eller legg til uidentifiserte personer som beskrevet under punkt 3 i Vedlegg 4.
3. Trykk på «Blind Search» for å få opp søkevinduet. Trykk på «New Search».
4. Marker de familieforholdene som skal undersøkes. I denne oppgaven ble forholdene «Parents», «Siblings» og «Direct match» undersøkt og skalert mot ubeslektet, men en treffgrense på 1.
5. Søk, og lagre rapporten.

VEDLEGG 6: RESULTAT AV BLINDSØK – LR-VERDIER

| Match list (Divided into matches)

Person 1	Person 2	Relationship	LR
18	26	Direct-match	7.67867e+018
18	26	Siblings	2.93819e+012
18	26	Parent-Child	3.30075e+008
13	14	Siblings	254778
13	14	Parent-Child	174344
20	21	Siblings	39239.4
20	21	Parent-Child	25284.3
24	25	Siblings	18335.9
16	17	Siblings	17951
16	17	Parent-Child	17950.2
4	6	Parent-Child	16848.1
22	23	Siblings	16205.3
22	23	Parent-Child	14495.3
7	9	Parent-Child	10217
4	5	Parent-Child	4932.38
4	6	Siblings	4431.56
13	15	Parent-Child	4388.17
18	19	Siblings	2590.05
19	26	Siblings	2590.05
1	2	Parent-Child	1940.12
1	3	Parent-Child	915.563
7	9	Siblings	828.524
1	2	Siblings	686.362
10	12	Parent-Child	606.901
7	8	Parent-Child	509.25
13	15	Siblings	395.824
10	11	Parent-Child	237.114
4	5	Siblings	144.355
1	6	Siblings	118.317
10	11	Siblings	53.8482
7	8	Siblings	49.4782
10	12	Siblings	38.8892
1	3	Siblings	28.771
15	22	Siblings	9.75407
10	20	Siblings	9.3403
11	21	Siblings	6.57775
10	21	Siblings	5.9688
11	15	Siblings	1.98903
3	6	Siblings	1.83257
1	11	Siblings	1.81763
2	4	Siblings	1.58744
2	6	Siblings	1.28239
1	20	Siblings	1.13239
1	4	Siblings	1.06983
9	21	Siblings	1.03523

VEDLEGG 7: UTREGNING AV LR FOR FORELDRE

Utregningene under er gjort basert på følgende formel:

$$LR = \frac{P(G_B|G_M, G_F, H_P)}{P(G_B|H_D)}$$

Hvor G_B er genotypen til den uidentifiserte personen («body»), G_M er genotypen til mor, G_F er genotypen til far, H_P er hypotesen om at personene er beslektet og H_D er hypotesen før betinging.

1.1

$$LR = \frac{P(A/A|A/A, A/A, Foreldre)}{P(A/A|A/A, A/A, Ubeslektet)} = \frac{1 \cdot 1}{p_A^2} = \frac{1}{p_A^2}$$

1.2

$$LR = \frac{P(A/A|A/B, A/A, Foreldre)}{P(A/A|A/B, A/A, Ubeslektet)} = \frac{\frac{1}{2} \cdot 1}{p_A^2} = \frac{1}{2p_A^2}$$

1.3

$$LR = \frac{P(A/B|A/B, A/A, Foreldre)}{P(A/B|A/B, A/A, Ubeslektet)} = \frac{1 \cdot \frac{1}{2}}{2p_A p_B} = \frac{1}{4p_A p_B}$$

1.4

$$LR = \frac{P(A/B|B/B, A/A, Foreldre)}{P(A/B|B/B, A/A, Ubeslektet)} = \frac{1 \cdot 1}{2p_A p_B} = \frac{1}{2p_A p_B}$$

1.5

$$LR = \frac{P(A/B|B/C, A/A, Foreldre)}{P(A/B|B/C, A/A, Ubeslektet)} = \frac{\frac{1}{2} \cdot 1}{2p_A p_B} = \frac{1}{4p_A p_B}$$

2.1

$$LR = \frac{P(A/A|A/B, A/B, Foreldre)}{P(A/A|A/B, A/B, Ubeslektet)} = \frac{\frac{1}{2} \cdot \frac{1}{2}}{p_A^2} = \frac{1}{4p_A^2}$$

2.2

$$LR = \frac{P(A/B|A/B, A/B, Foreldre)}{P(A/B|A/B, A/B, Ubeslektet)} = \frac{2 \cdot \left(\frac{1}{2} \cdot \frac{1}{2}\right)}{2p_A p_B} = \frac{1}{4p_A p_B}$$

2.3

$$LR = \frac{P(A/A|A/C, A/B, Foreldre)}{P(A/A|A/C, A/B, Ubeslektet)} = \frac{\frac{1}{2} \cdot \frac{1}{2}}{p_A^2} = \frac{1}{4p_A^2}$$

2.4

$$LR = \frac{P(A/B|A/C, A/B, Foreldre)}{P(A/B|A/C, A/B, Ubeslektet)} = \frac{\frac{1}{2} \cdot \frac{1}{2}}{2p_A p_B} = \frac{1}{8p_A p_B}$$

2.5

$$LR = \frac{P(A/C|C/D, A/B, Foreldre)}{P(A/C|C/D, A/B, Ubeslektet)} = \frac{\frac{1}{2} \cdot \frac{1}{2}}{2p_A p_C} = \frac{1}{8p_A p_c}$$

VEDLEGG 8: UTREGNING LR BRØDRE MED IBD VS. UBESLEKTET

1.1

$$LR = \frac{P(A/A, A/A|Brødre)}{P(A/A, A/A|Ubeslektet)} = \frac{p_A^4 \frac{1}{4} + p_A^3 \frac{1}{2} + p_A^2 \frac{1}{4}}{p_A^4} = \frac{\frac{1}{4}p_A^2 + \frac{1}{2}p_A + \frac{1}{4}}{p_A^2}$$

1.2

$$LR = \frac{P(A/A, A/B|Brødre)}{P(A/A, A/B|Ubeslektet)} = \frac{p_A^2 \times 2p_A p_B \times \frac{1}{4} + p_A^2 p_B \times \frac{1}{2}}{2p_A^2 p_B}$$

$$= \frac{2p_A^3 p_B \times \frac{1}{4} + p_A^2 p_B \frac{1}{2}}{2p_A^3 p_B}$$

$$= \frac{\frac{1}{4}(p_A + 1)}{p_A}$$

1.3

$$LR = \frac{P(A/A, B/B|Brødre)}{P(A/A, B/B|Ubeslektet)} = \frac{\frac{1}{4}p_A^2 p_B^2}{p_A^2 p_B^2} = \frac{1}{4}$$

2.1

$$LR = \frac{P(A/B, A/B|Brødre)}{P(A/B, A/B|Ubeslektet)}$$

$$= \frac{\frac{1}{4}2p_A p_B \times 2p_A p_B + \frac{1}{2}p_A p_B(p_A + p_B) + \frac{1}{4} \times 2p_A p_B}{4p_A^2 p_B^2}$$

$$= \frac{p_A^2 p_B^2 + \frac{1}{2}p_A p_B(p_A + p_B) + \frac{1}{2}p_A p_B}{4p_A^2 p_B^2}$$

$$= \frac{p_A p_B + \frac{1}{2}(p_a + p_B) + \frac{1}{2}}{4 p_A p_B}$$

2.2

$$LR = \frac{P(A/B, A/C|Brødre)}{P(A/B, A/C|Ubeslektet)} = \frac{2p_A p_B \times 2p_A p_C \times \frac{1}{4} + p_A p_B p_C \times \frac{1}{2}}{2p_A p_B \times 2p_A p_C}$$

$$= \frac{p_A^2 p_B p_C + \frac{1}{2} p_A p_B p_C}{4 p_A^2 p_B p_C}$$

$$= \frac{p_A + \frac{1}{2}}{4 p_A}$$

2.3

$$LR = \frac{P(A/B, B/C|Brødre)}{P(A/B, B/C|Ubeslektet)} = \frac{2p_A p_B \times 2p_B p_C \times \frac{1}{4} + p_A p_B p_C \times \frac{1}{2}}{2p_A p_B \times 2p_B p_C}$$

$$= \frac{p_A p_B^2 p_C + \frac{1}{2} p_A p_B p_C}{4 p_A p_B^2 p_C}$$

$$= \frac{p_B + \frac{1}{2}}{4 p_B}$$

2.4

$$LR = \frac{P(A/B, C/C|Brødre)}{P(A/B, C/C|Ubeslektet)} = \frac{2 \times p_A p_B \times p_C^2 \times \frac{1}{4}}{2 \times p_A p_B \times p_C^2} = \frac{1}{4}$$

VEDLEGG 9: R-SKRIPT FOR PLOTTING AV TETTHETER

Figur 3.3: Log LR for testing mot foreldre ved systemene CODIS og SGM**.:**

```
parents_LR <- read.csv2("parents_LR.csv")

plot(density(parents_LR$log_LR_CODIS), xlim=c(3, 16), ylim=c(0.00, 0.60),
    main="Log LR for foreldre ved systemene CODIS og SGM", xlab="log LR")
lines(density(parents_LR$log_LR_SGM), col=2)
legend(12,0.45, c("CODIS", "SGM"), lty=c(1,1),
      lwd=c(2.5,2.5), col=c("black", "red"))
```

_____

Figur 3.6: Log LR ved bruk av bror som referanseperson med systemene CODIS og SGM.:

```
brother_LR <- read.csv2("brother_LR.csv")

plot(density(brother_LR$log_LR_CODIS), xlim=c(-4, 13), ylim=c(0.00, 0.30),
    main="Log LR for broedre ved systemene CODIS og SGM", xlab="log LR")
lines(density(parents_LR$log_LR_SGM), col=2)
legend(9.5,0.20, c("CODIS", "SGM"), lty=c(1,1),
      lwd=c(2.5,2.5), col=c("black", "red"))
```

_____

Figur 3.8: Høyeste log LR for hvert familietre med bror som referanseperson ved systemet CODIS, fordelt på om det er det sanne familietreet eller ikke.:

```
brother_codis <- read.csv2("lr_brother_codis.csv")

plot(density(brother_codis$log_LR_CODIS), xlim=c(-2, 13), ylim=c(0.00, 0.5),
    main="Hoyeste log LR for sosken ved systemet CODIS", xlab="log LR")
```

```
lines(density(brother_codis$log_LR_FALSE[1:9]), col=2)
legend(9,0.35, c("TRUE", "FALSE"), lty=c(1,1),
      lwd=c(2,2), col=c("black", "red"))
```

_____

Figur 3.9: Høyeste log LR for hvert familietre med bror som referanseperson ved systemet SGM, fordelt på om det er det sanne familietreet eller ikke.:

```
brother_sgm <- read.csv2("lr_sgm_brother.csv")

plot(density(brother_sgm$log_LR), xlim=c(-0.6, 6.7), ylim=c(0.00, 1.2),
    main="Hoyeste log LR for sosken ved systemet SGM", xlab="log LR")
lines(density(brother_sgm$log_LR_false[1:31]), col=2)
legend(4.5,0.6, c("TRUE", "FALSE"), lty=c(1,1),
      lwd=c(2,2), col=c("black", "red"))
```

_____

Figur 3.10: LR-verdiene for de sanne slektstrærne ved hvert scenario, med terskel for log LR≈3.7.:

```
allLR <- read.csv2("allLR.csv")

plot(density(allLR$log_LR_CODIS_parents), xlim=c(-3, 16), ylim=c(0.00, 0.60),
    main="Alle LR fordelts for scenario og antall markorer med terskel for LR", xlab="log LR")
lines(density(allLR$log_LR_SGM_parents), col=2)
lines(density(allLR$log_LR_SGM_brother), col=3.2)
lines(density(allLR$log_LR_CODIS_brother), col=5)
abline(v=3.7, col=8)
legend(10.8,0.55, c("CODIS Parents", "SGM Parents", "CODIS Brother", "SGM Brother", "Terskel for
LR"),
      lty=c(1,1), lwd=c(2.5,2.5), col=c("black", "red", 5, 3.2, 8))
```

VEDLEGG 10: UTREGNING AV STANDARDFEIL OG KONFIDENSINTERVALL FOR ANDELEN LR-VERDIER OVER TERSKELEN I DE FORSKJELLIGE SCENARIOENE

1: $\boldsymbol{P = P(LR \geq 5000 | H_0, CODIS, Foreldre)}$

$$\hat{p} = \frac{X}{n} = \frac{100}{100} = 1$$

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{1(1-1)}{100}} \approx 0$$

$$KI = \left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) \approx 1$$

2: $\boldsymbol{P = P(LR \geq 5000 | H_0, SGM, Foreldre)}$

$$\hat{p} = \frac{X}{n} = \frac{100}{100} = 1$$

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{1(1-1)}{100}} \approx 0$$

$$KI = \left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) \approx 1$$

3: $\boldsymbol{P = P(LR \geq 5000 | H_0, CODIS, Bror)}$

$$\hat{p} = \frac{X}{n} = \frac{50}{100} = 0.5$$

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.5(1-0.5)}{100}} = 0.05$$

$$KI = \left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

$$= \left(0.5 - 1.96\sqrt{\frac{0.5(1-0.5)}{100}}, 0.5 + 1.96\sqrt{\frac{0.5(1-0.5)}{100}} \right) = (0.402, 0.598)$$

4: $\boldsymbol{P = P(LR \geq 5000 | H_0, SGM, Bror)}$

$$\hat{p} = \frac{X}{n} = \frac{12}{100} = 0.12$$

$$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.12(1-0.12)}{100}} = 0.032$$

$$KI = \left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

$$= \left(0.12 - 1.96\sqrt{\frac{0.12(1-0.12)}{100}}, 0.12 + 1.96\sqrt{\frac{0.12(1-0.12)}{100}} \right) = (0.057, 0.183)$$

VEDLEGG 11: UTSKRIFT R COMMANDER, WELCH T-TEST

H_0: Gjennomsnittlig log LR-verdi for foreldre med CODIS er lik gjennomsnittlig LR-verdi for foreldre med SGM.

H_1: Gjennomsnittlig log LR-verdi er større for foreldre med CODIS enn foreldre med SGM.

```
        Welch Two Sample t-test

data:  two samples
t = 26.7124, df = 171.952, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 4.255822      Inf
sample estimates:
mean of x mean of y
  10.2764    5.6104
```

H_0: Gjennomsnittlig log LR-verdi for brødre med CODIS er lik gjennomsnittlig LR-verdi for brødre med SGM.

H_1: Gjennomsnittlig log LR-verdi er større for brødre med CODIS enn brødre med SGM.

```
        Welch Two Sample t-test

data:  two samples
t = 6.4184, df = 180.248, p-value = 5.905e-10
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 1.07609     Inf
sample estimates:
mean of x mean of y
   3.6845    1.9880
```

H_0: Gjennomsnittlig log LR-verdi for foreldre med SGM er lik gjennomsnittlig LR-verdi for brødre med SGM.

H_1: Gjennomsnittlig log LR-verdi er større for foreldre enn brødre.

```
        Welch Two Sample t-test

data:  two samples
t = 19.8544, df = 165.867, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 3.193822      Inf
sample estimates:
mean of x mean of y
   5.6104    1.9880
```

H_0: Gjennomsnittlig log LR-verdi for foreldre med CODIS er lik gjennomsnittlig LR-verdi for brødre med CODIS.

H_1: Gjennomsnittlig log LR-verdi er større for foreldre enn brødre.

```
        Welch Two Sample t-test

data:  two samples
t = 26.7124, df = 171.952, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 4.255822      Inf
sample estimates:
mean of x mean of y
  10.2764    5.6104
```

H_0: Gjennomsnittlig log LR-verdi for foreldre i oppgaven er det samme som gjennomsnittlig LR-verdi for foreldre beregnet av Ge et al.

H_1: Gjennomsnittlig log LR-verdi er ulik i oppgaven og artikkelen.

```
        Welch Two Sample t-test

data:  two samples
t = -0.1637, df = 99.019, p-value = 0.8703
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.3136091  0.2658091
sample estimates:
mean of x mean of y
  10.2561   10.2800
```

Norges miljø- og
biovitenskapelige
universitet

Postboks 5003
NO-1432 Ås
67 23 00 00
www.nmbu.no