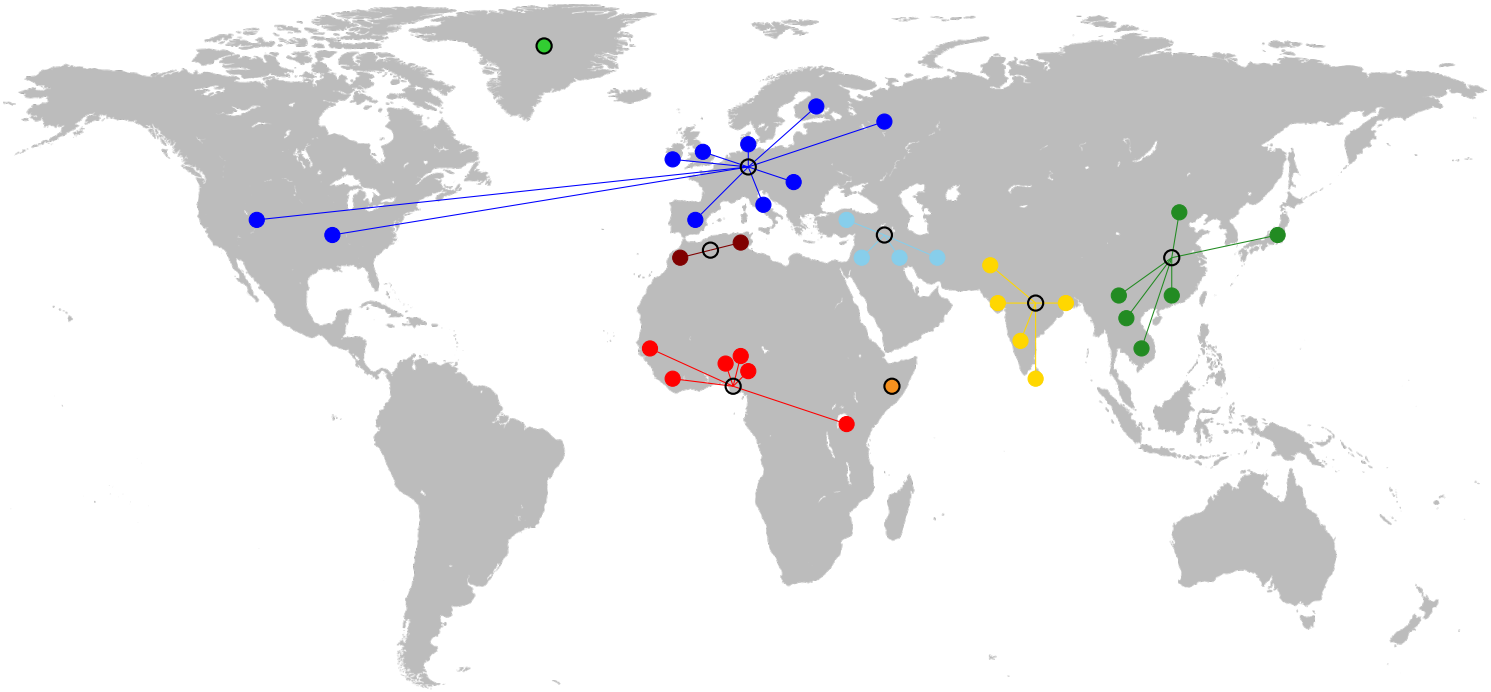
Prædiktion af Individers Afstamning ved Anvendelse af Ancestry Informative Markers

Speciale

Martin Nørskov

Efterår 2018 - Forår 2019



Aalborg Universitet

Institut for Matematiske Fag

AALBORG UNIVERSITET
STUDENTERRAPPORT

Institut for
Matematiske Fag
Matematik
Skjernvej 4A
9220 Aalborg Øst

Titel:

Prædiktion af Individers Afstamning
ved Anvendelse af
Ancestry Informative Markers

Tema:

Statistisk modellering i retsgenetik

Projektperiode:

Efterår 2018 - Forår 2019

Forfatter:

Martin Nørskov

Vejleder:

Torben Tvedebrink

Sidetal:

63

Afleveringsdato:

3. juni 2019

Synopsis:

I dette speciale undersøges individers afstamning ved brug af Ancestry Informative Markers (AIMs). Der opstilles et hypotesetest, som vurderer, om en specifik profil stammer fra en given population. I analysen er der brugt klassifikationstræer, random forest og lasso-regression til at udvælge en delmængde af markørerne, som bedst skelner mellem populationerne. Derudover er der undersøgt markørernes sensitivitet samt markørernes coverage og tilbøjelighed til at give et dropout. Analysen viste, at klassifikationstræer med en filtrering for de mindst sigende markører er de bedste modeller. Det førte til to modeller, CART20 og CART50, som har henholdsvis 21 og 7 markører. Med hensyn til markørernes sensitivitet er rs671, rs3811801, rs1800414, rs4821004, rs2125345 og rs2899826 de mest sensitive, og i forhold til markørernes dropout er rs1296819, rs12439433 og rs4833103 de mest tilbøjelige til at give et dropout.

Forord

Dette speciale er udarbejdet af en studerende på tredje og fjerde semester på kandidatuddannelsen i Matematik ved Aalborg Universitet.

Specialet er opbygget i kapitler og afsnit, hvor første kapitel begynder med en indledning. I andet kapitel introduceres teori, hvor et hypotesetest opstilles. I tredje kapitel laves en analyse af to datasæt med anvendelse af teorien. Analysen vil blandt andet finde en delmængde af markører, som bedst skelner mellem populationer samt undersøge markørernes sensitivitet og tilbøjelighed til at være et dropout.

Figurer, tabeller og ligninger er nummereret fortløbende efter kapitel og afsnit. Kilder er angivet efter forfatter og årstal. Hvis et helt afsnit er baseret på en eller flere kilder, er disse angivet i starten af afsnittet, hvorimod hvis kun en enkel paragraf er baseret på en eller flere kilder, er disse angivet i slutningen af paragraffen. En samlet litteraturliste med kilderne findes sidst i specialet. Programmet `R` er anvendt til statistiske analyser.

En stor tak skal gå til min vejleder Torben Tvedebrink for god vejledning igennem hele forløbet.

Aalborg Universitet, 3. juni 2019

Martin Nørskov

mnarsk14@student.aau.dk

Summary

The aim of this Master Thesis has been to investigate the ancestry of individuals using Ancestry Informative Markers (AIMs), which are markers containing genetic information about ancestry. This is of great interest as the ancestry of a perpetrator can be predicted, which will help the police in their investigations of criminal cases.

In the theory a hypothesis test is constructed, that assess whether a specific profile originates from a given population based on the genotype on the markers of the profile. To this both bi-allelic and tri-allelic markers will be used.

In the analysis two data set were analyzed using the hypothesis test from the theory. The first data set contains the genotype on 164 markers of 3.560 profiles. In the first part of the analysis the attempt was to select a subset of these markers, which could best distinguish between the populations. To this Classification- and Regressiontrees, Random Forest and Lasso regression were used. The analysis showed that the best models were CART20 and CART50, which had 21 and 7 markers. Furthermore, it was found that rs16891982 was the most remarkable marker as it had the highest weight in all of the subsets that it appeared in.

In the second part the sensitivity of the markers were analyzed in terms of accept or reject the populations in two cases. Simulated profiles were analyzed, where they were accepted or rejected of their own population and if the profiles were rejected or accepted of their population by a change on a marker, the marker was said to be sensitive, respectively. In the first case the analysis showed that rs671, rs3811801 and rs1800414 were the most sensitive markers, and in the second case it was rs4821004, rs2125345 and rs2899826.

In the third part a data set with six dilutions were examined in terms of the coverage of the markers and how likely they were to give a dropout. At first, coverage for chosen markers were examined, where markers with low, medium and high coverage were accessible. The first dilutions had a relative high coverage wheras the coverage seemed to decrease in the last dilutions. The marker rs735480 was among the markers with the highest coverage and was even one of the markers from CART20.

The heterozygous cases showed that a few cases were outside the interval on $[0,3;3]$ and there were considerably more cases outside the interval on $[0,67;1,67]$.

In order for a marker to be reliable the percentage of positive coverage for the markers should be around 50%, so positive and negative coverage are nearly the same size. An analysis showed that 31 markers exceeded a difference on 10 from 50%. The markers rs1760921 and rs2001907 were the most remarkable as they had the biggest differences and exceeded the difference most of the times.

Also, in order for a marker to be reliable the percentage of the major allel should be around 50% and

100% for heterozygous and homozygous cases, respectively. An analysis showed that 69 markers exceeded a difference on 10 more than one time. The marker rs9530435 was the most remarkable as it exceeded the difference most of the times with the highest average of the differences.

An analysis for the relationship between positive coverage and the major allel showed no clear tendency. However, there were a few cases where some of the markers exceeded the difference on 10 for both positive coverage and the major allel. This was in particular valid for rs9530435, which also was the most remarkable marker in the analysis for the major allel.

In terms of how likely the markers were to give a dropout two cases were examined. In the first case the requirement for giving a dropout was that coverage should be 100 or lower and the heterozygous cases should be outside the balance on $[0,3;3]$. The marker rs1296819 was the most likely to give a dropout, but also rs12439433 and rs4833103 gave many dropouts. In the second case a tightened change on the heterozygous balance was made as the balance was $[0,67;1,67]$. The result for this analysis showed the same tendency as the first heterozygous balance, but in general there were more dropouts in this case.

Finally, simulated profiles were examined, where the profiles markers were diluted with the frequencies founded in the dropout analysis. It was found that the percentage of the classification for the simulated diluted profiles was close to the percentage of the classification for the simulated profiles in most of the populations. Thus, in this case dilutions of the profiles markers do not give a substantial effect.

Indholdsfortegnelse

1 Indledning

Indvandringen i Danmark har været stigende de seneste 10 år, hvilket har medført et stigende antal af invandrer og efterkommere fra både vestlige og ikke-vestlige lande. I takt med den stigende indvandring er antallet af indvandrer og efterkommere, som er skyldige i kriminalitet, også steget. Kriminaliteten blandt de personer, som ikke er af dansk oprindelse, er dermed et stigende problem i Danmark. Der er derfor motivation for at udvikle værktøjer til at identificere de kriminelle indvandrer og efterkommere. [Danmarks Statistik, 2019]

Ancestry Informative Markers (AIMs) er markører, som indeholder genetisk information om individers afstamning. Med udgangspunkt i et individs DNA kan markørerne dermed informere, om individet er af dansk oprindelse eller stammer fra et andet folkeslag. Dette kan derfor være en hjælp i politiets arbejde, da markørerne kan anvendes til at identificere afstamningen af en potentiel gerningsmand via DNA fra et gerningssted, hvilket vil skærpe politiets eftersøgning af gerningsmanden. [Tvedebrink et al., 2018]

Dette speciale vil derfor undersøge de genetiske markører med henblik på at bidrage med informationen om individers afstamning.

1.1 Deoxyribonucleic Acid

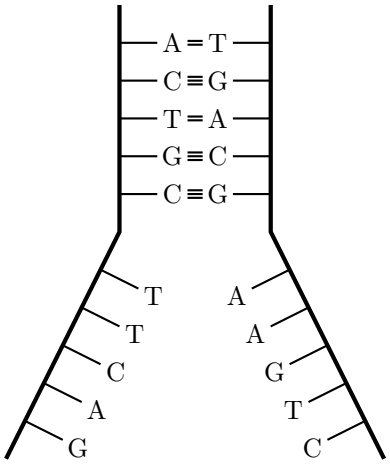
Dette afsnit er baseret på [Butler, 2010].

Den menneskelige krop består cirka af 100 trillioner celler, hvor hver celle har en kerne, som indeholder et kemisk stof kaldet *Deoxyribonucleic Acid* (*DNA*). Stoffet indeholder genetisk information, som anvendes til at reproducere cellen, hvor den komplette information fra hele DNA'et i en celle kaldes for *genom*. DNA er sammensat af nukleotid-enheder, som er lavet af tre ting: en nukleobase, et sukker og et fosfat. Basen indeholder variationen i hver nukleotid-enhed, hvorimod sukker og fosfat opbygger strukturen af DNA'et. Nukleobasen kan være en af de fire mulige nukleobaser: *Adenine* (A), *Cytosine* (C), *Guanine* (G) og *Thymine* (T). Mennesket har cirka 3 milliarder nukleotid-positioner i sit genetiske DNA, hvor disse fire baser kan sidde. Informationen i DNA'et er kodet som ordningen af baserne på nukleotid-positionerne.

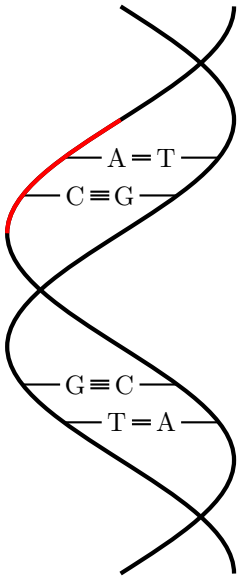
Et DNA er sammensat af to strenge, som er knyttet sammen under en proces kaldet *hybridization*. Individuelle nukleotider sammensættes med sin komplementære base via hydrogen-bindinger mellem baserne. Det vil sige, at basen A kun kan bindes sammen med basen T, mens basen C kun kan bindes sammen med basen G. Der er tre hydrogen-bindinger mellem C og G, mens der kun er to bindinger mellem A og T, så bindingen mellem C og G er lidt stærkere end bindingen mellem A og T. Hydrogen-bindingerne mellem baserne kan brydes ved opvarmning eller en kemisk behandling, hvilket er en proces kaldet *denaturation*. Dette vil separere de to DNA-strenge, jævnfør Figur 1.1 (a). Hvis DNA-strengene nedkøles, vil de finde sammen igen, så deres komplementære sekvens stemmer overens, hvilket er en anden proces kaldet *renaturation*. De to DNA-strenge med hydrogen-bindingerne mellem baserne former tilsammen en dobbelthelix, hvilket kan ses i Figur 1.1 (b).

En lang dobbelthelix af to DNA-strenge udgør tilsammen et *kromosom*, hvor genomet i en celle er

fordelt på 46 kromosomer. Kromosomerne sidder i par, så de 46 kromosomer danner 23 par af kromosomer. En specifik position på et kromosom kaldes et *locus* (pluralis: *loci*) eller en *markør*, som består af en sekvens af *alleller*, der kan være baserne A, C, G og T. Hvis et locus består af to alleller, kaldes det *bi-allelisk*, og hvis det består af tre alleller, kaldes det *tri-allelisk*. Hvis de to alleller på et bi-allelisk locus er ens, kaldes de for *homozygot*, og hvis de to alleller ikke er ens, kaldes de for *heterozygot*. Karakteriseringen af disse alleller på et locus kaldes en *genotype*. I Figur 1.1 (b) angiver det røde område et bi-allelisk locus, hvor genotypen er AC og dermed er heterozygot. En samling af genotypen fra flere loci udgør tilsammen en DNA-profil.



(a) To DNA-strenge, som bliver denatureret, så hydrogen-bindingerne mellem baserne brydes.



(b) To DNA-strenge udgør tilsammen en dobbelthelix. Det røde område på den ene streng angiver et bi-allel locus.

Figur 1.1: Denaturation og en dobbelthelix med to DNA-strenge.

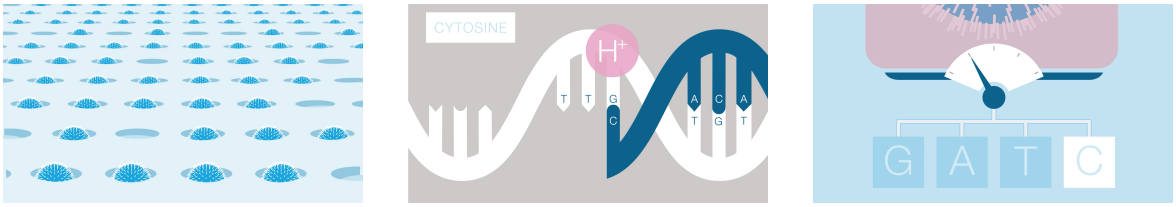
En variation på en enkel base på et locus mellem individer kaldes en *single nucleotide polymorphism* (*SNP*). Der findes millioner af SNP'er for hvert individ, så dermed kan SNP'er være nyttige i at adskille individer fra hinanden. Nogle af SNP'erne betegnes som *Ancestry Informative Markers* (*AIMs*), da variationerne i baserne på markørerne mellem individer kan separere individerne med hensyn til deres afstamning.

1.2 Ion AmpliSeq Targeted Sequencing Teknologien

Dette afsnit er baseret på [Scientific, 2019a], [Scientific, 2019b] og [Scientific, 2019c].

Teknologien, som er blevet anvendt til at udtrække data fra DNA-prøverne, kaldes *Ion AmpliSeq Targeted Sequencing*. Det er en metode, som sekventerer informationen fra DNA'et fra baserne A, C, G og T til digital information. Derudover anvender metoden også *Targeted Sequencing*, som er en metode til at målrette sin sekventering til specifikke områder af DNA'et. Teknologien er hurtig, mere simpel og mindre omkostningsfuld end andre sekventeringsmetoder.

Processen for sekventering begynder med en DNA-prøve, som denatureres og dermed adskilles i millioner af DNA-strenge. Hver af disse strenge knytter sig til en kugle, hvorefter strengen kopieres, så den dækker hele kuglen. Disse kugler med strenge tilføjes nu til en chip, som består af millioner af brønde, hvor hver kugle deponerer ned i hver sin brønd, jævnfør Figur 1.2 (a). Herefter fyldes chippen med en af de fire baser, hvor den enkelte base inkorporeres på en base i en DNA-streng, hvis de to baser komplimentært stemmer overens. Hver gang en base inkorporeres i en DNA-streng, frigives der en hydrogenion, jævnfør Figur 1.2 (b). Frigivelsen af en hydrogenion ændrer pH-værdien i brønden, hvor en måler i bunden af brønden måler denne ændring og omsætter den til en spænding. Denne spænding indikerer en af de fire baser, som derefter læses og gemmes, jævnfør Figur 1.2 (c). Chippen tømmes for basen og fyldes igen med en ny base hvert 15. sekund. Hvis der er to identiske baser ved siden af hinanden på DNA-strengen, inkorporeres to ens baser, spændingen fordobles og to identiske baser læses. Denne proces for sekventering foregår simultant i alle brøndene.



(a) En chip med brønde, hvori der er deponerede kugler med DNA-strenge.

(b) En inkorporering af basen C på basen G, hvilket frigører et hydrogenion.

(c) En brønd med en kugle, hvor måleren registrerer og læser basen C.

Figur 1.2: Processen med sekventering. [Scientific, 2019a]

Som sagt anvender sekventeringsmetoden også Targeted Sequencing, som målretter sin sekventering til specifikke områder af et DNA. Dette gøres ved hjælp af processen *Polymerase Chain Reaction* (*PCR*), som amplificerer et enkelt DNA til millioner af kopier på kort tid. Derudover er der også brug for *primer pairs*, som er designede par, der amplificerer de specifikke områder af DNA'et.

1.3 Introduktion af data

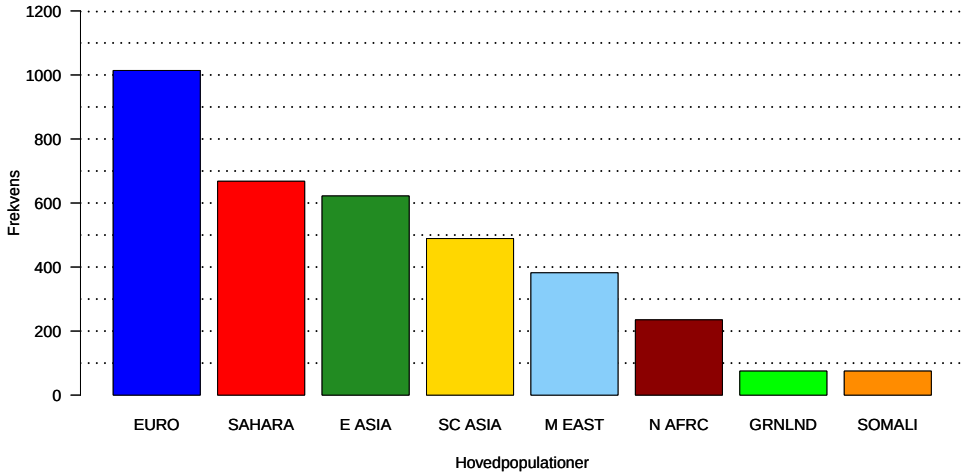
Sekventeringsmetoden fra afsnit 1.2 er brugt til at lave to datasæt, som vil blive introduceret i dette afsnit.

Det første datasæt består af genotypen på 164 markører fra 3.560 profiler. Profilerne kommer fra otte hovedpopulationer, som kan ses i Tabel 1.1 samt deres forkortelse og antallet af profiler. I projektet vil hovedpopulationerne blive refereret til med deres forkortelser.

Hovedpopulation	Forkortelse	Antal profiler
Europa	EURO	1.014
Sahara	SAHARA	668
Østasien	E ASIA	622
Syd- og Centralasien	SC ASIA	489
Mellemøsten	M EAST	382
Nordafrika	N AFRC	235
Grønland	GRNLND	75
Somalien	SOMALI	75

Tabel 1.1: De otte hovedpopulationer med deres forkortelse og antallet af profiler.

Antallet af profiler i stikprøverne fra de otte hovedpopulationer kan ses i Figur 1.3.



Figur 1.3: Antallet af profiler i stikprøverne fra de otte hovedpopulationer.

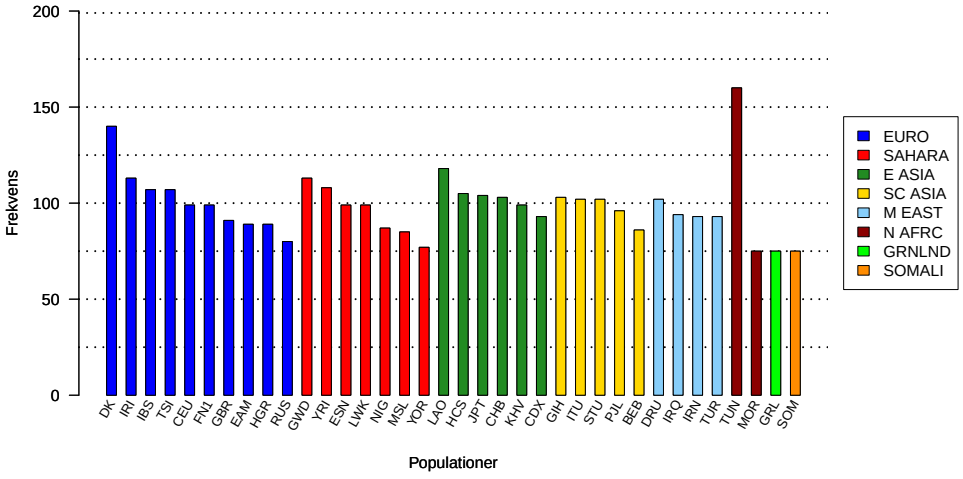
Det kan ses i Figur 1.3, at EURO har den klart største stikprøve med over 1.000 profiler, hvorefter de andre hovedpopulationer følger efter. GRNLND og SOMALI er de to hovedpopulationer, som har de mindste stikprøver med 75 profiler.

De otte hovedpopulationer kan inddeles i 36 mindre populationer, som er angivet i Tabel 1.2 samt deres forkortelser.

Hovedpopulation	Population	Forkortelse	Antal profiler
EURO	Danskere, Danmark	DK	140
EURO	Irer, Irland	IRI	113
EURO	Iberer, Spanien	IBS	107
EURO	Toscanere, Italien	TSI	107
EURO	Utah beboere, Nordvesteuropa	CEU	99
EURO	Finnere, Finland	FN1	99
EURO	Britere, England og Skotland	GBR	91
EURO	Europæisk amerikanere	EAM	89
EURO	Ungarere, Ungarn	HGR	89
EURO	Russere, Rusland	RUS	80
SAHARA	Gambianere, Western, Gambia	GWD	113
SAHARA	Yorubaere, Ibadan, Nigeria	YRI	108
SAHARA	Esan folket, Nigeria	ESN	99
SAHARA	Luhya folket, Webuye, Kenya	LWK	99
SAHARA	Nigerianere, Nigeria	NIG	87
SAHARA	Mende folket, Sierra Leone	MSL	85
SAHARA	Yorubaere, Benin, Nigeria	YOR	77
E ASIA	Laosianere, Laos	LAO	118
E ASIA	Syd Hankinesere, Kina	HCS	105
E ASIA	Japanere, Tokyo, Japan	JPT	104
E ASIA	Hankinesere, Beijing, Kina	CHB	103
E ASIA	Kinh folket, Ho Chi Minh City, Vietnam	KHV	99
E ASIA	Kinesere, Xishuangbanna, Kina	CDX	93
SC ASIA	Gujarati folket, Indien	GIH	103
SC ASIA	Teluguere, Indien	ITU	102
SC ASIA	Tamilere, Sri Lanka	STU	102
SC ASIA	Punjabiere, Lahore, Pakistan	PJL	96
SC ASIA	Bengaliere, Bangladesh	BEB	86
M EAST	Drusere, Israel	DRU	102
M EAST	Irakere, Irak	IRQ	94
M EAST	Iranere, Iran	IRN	93
M EAST	Tyrkere, Tyrkiet	TUR	93
N AFRC	Tunesere, Tunesien	TUN	160
N AFRC	Marokkanere, Marokko	MOR	75
GRNLND	Grønlændere, Grønland	GRL	75
SOMALI	Somaliere, Somalien	SOM	75

Tabel 1.2: De 36 populationer med deres hovedpopulation, forkortelse og antallet af profiler.

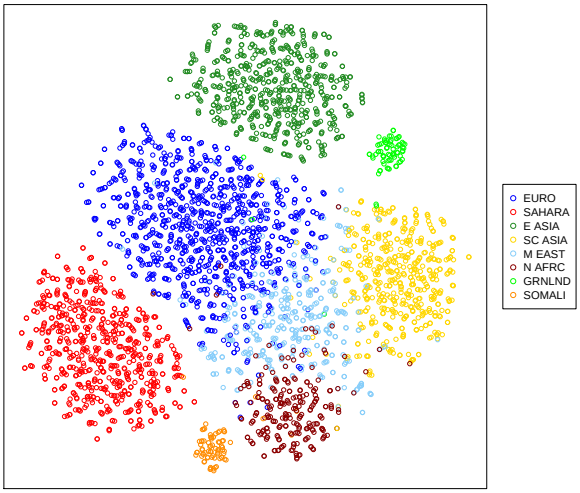
Antallet af profiler i stikprøverne fra de 36 populationer er illustreret i Figur 1.4.



Figur 1.4: Antallet af profiler i stikprøverne fra de 36 populationer.

De fleste stikprøver i Figur 1.4 er på omkring 100 profiler. Populationerne MOR, GRL og SOM har de mindste stikprøver på 75 profiler, hvorimod TUN har den største stikprøve med over 150 profiler.

I Figur 1.5 ses et forsøg på at skelne mellem informationen i datasættet, hvor genotypen på profilernes markører er anvendt i funktionen `Rtsne`, som bedst muligt bevarer parvise afstande fra $\mathbb{R}^n$ til $\mathbb{R}^2$ ved brug af en ikke-lineær transformation. Den transformerer altså strukturen i høj-dimensionel data til en lav-dimensionel struktur. [Maaten and Hinton, 2008] [Donaldson, 2016]



Figur 1.5: Visualisering af genotypen på profilernes markører med et tSNE-plot.

Ved at observere Figur 1.5 kunne det kunne tyde på, at profilernes markører indeholder genetisk information om profilernes afstamning. Hovedpopulationerne EURO, SC ASIA, M EAST og N AFRC ligger tæt op ad hinanden, så det tages i mente, at det kan være svært at skelne mellem disse hovedpopulationer. Derimod distancerer SAHARA, E ASIA, GRNLND og SOMALI sig mere fra de andre hovedpopulationer.

Det andet datasæt er et fortyndingsdatasæt, som består af fem profiler, der er blevet genotypet på 165 markører i seks fortyndinger og i to duplikater. Dette giver ialt 9.900 observerede markører.

I datasættet er antallet af de læste baser for hver observation angivet som *reads*, for eksempel er antallet af læste A-baser angivet som *A-reads*. For troværdigheden af genotypen på markørerne bør reads for én base være over 100, som i projektet vil blive referet til som *allelgrænse*. Summen af alle reads for de fire baser er angivet som *coverage*. Sekventeringen af en DNA-streng er foregået både forlæns og baglæns, så antallet af forlæns læste baser og antallet af baglæns læste baser tilsammen udgør coverage. Antallet af forlæns og baglæns læste baser er henholdsvis angivet som *positiv coverage* og *negativ coverage*. Procentandelen af positiv coverage er angivet som *procent positiv coverage*, hvilket bør ligge på omkring 50%, så positiv og negativ coverage er lige store. Jo tættere procentandelen for positiv coverage er på 50%, jo mere valid er den pågældende markør. Procentandelen for den mest læste base for en observation er angivet som *procent dominerende allel*. For homozygote og heterozygote alleller bør denne procentandel ligge på henholdsvis omkring 100% og 50%, da der er én base for homozygote alleller og to baser for heterozygote alleller.

2 Hypotesetest for population

I dette kapitel udledes teori, som anvender genotypen på profilernes markører til at afgøre, om profilerne stammer fra givne populationer. Der vil både være fokus på bi-alleliske og tri-alleliske markører i henholdsvis afsnit 2.1 og afsnit 2.2.

2.1 Hypotesetest med bi-alleliske markører

Dette afsnit er baseret på [Tvedebrink et al., 2018].

I dette afsnit opstilles et hypotesetest, som vurderer på baggrund af bi-alleliske markører, om en specifik profil stammer fra en given population. På grund af bi-alleliske markører ses der kun på frekvensen af det første allel, kaldet *allel 1*, som er det første i leksikografisk rækkefølge. For eksempel vil basen C være allel 1 på en markør, hvor der er alleller med baserne C og T. Der tages altså udgangspunkt i, om frekvensen af profilens allel 1 på en specifik markør stemmer overens med frekvensen af populationens allel 1. Der fremstilles en såkaldt *z-score*, som vurderer hypotesen.

Antag, at I bi-alleliske markører er genotypet i J forskellige populationer for n_j profiler. Frekvensen af allel 1 for en population er angivet som $x_{ij} \in \{0, 1, \ldots, 2n_j\}$ for markør $i \in \{1, 2, \ldots, I\}$ i population $j \in \{1, 2, \ldots, J\}$. Det vil sige, at $\mathbf{x}_j = \begin{bmatrix} x_{1j} & x_{2j} & \cdots & x_{Ij} \end{bmatrix}$ er en vektor, som angiver frekvensen af allel 1 på forskellige markører i population j. En enkel specifik profil er angivet som $\mathbf{x}_0 = \begin{bmatrix} x_{01} & x_{02} & \cdots & x_{0I} \end{bmatrix}$, som er vektoren over alle profilens markører, hvor indgangene $x_{0i} \in \{0, 1, 2\}$ er frekvensen af allel 1 på markør i. Givet en population j ønskes det at sammenligne to hypoteser om $\mathbf{x}_0$:

$$\begin{aligned} H_0 &: \mathbf{x}_0 \text{ stammer fra population } j \\ H_1 &: \mathbf{x}_0 \text{ stammer ikke fra population } j. \end{aligned} \tag{2.1}$$

For at sammenligne de to hypoteser opstilles en likelihood ratio test, hvor likelihooden for hver af $\mathbf{x}_0$ og $\mathbf{x}_j$ bliver sammenlignet. Denne test vil så angive, om det er mere sandsynligt, at $\mathbf{x}_0$ og $\mathbf{x}_j$ stammer fra den samme population, eller om de stammer fra to forskellige populationer. Antag, at markørerne og allellerne inden for en population hver især er indbyrdes uafhængige. Dette medfører, at X_{0i} og X_{ij} er binomialfordelt, så under nulhypotesen er $X_{0i} \sim \mathrm{B}(2, p_{ij})$ og $X_{ij} \sim \mathrm{B}(2n_j, p_{ij})$, hvor p_{ij} er sandsynligheden for allel 1 på markør i i population j. Endvidere haves det, at $X_{+i} = X_{0i} + X_{ij} \sim \mathrm{B}(2(n_j + 1), p_{ij})$. På grund af bi-alleliske markører er $q_{ij} = 1 - p_{ij}$ sandsynligheden for allel 2 på markør i i population j. Hele essensen af hypotesetestet er at finde ud af, om $\mathbf{x}_0$ og $\mathbf{x}_j$ stammer fra den samme population, så det ønskes at betinge med x_{+i} i likelihood ratioen. Dermed vises det med den betingede sandsynlighed, at fordelingen af $X_{0i} = x_{0i} \mid X_{+i} = x_{+i}$ er hyper-geometrisk

$$
\begin{aligned}
P\left(X_{0i}=x_{0i}\mid X_{+i}=x_{+i}\right) &= \frac{P\left(X_{0i}=x_{0i},X_{+i}=x_{+i}\right)}{P\left(X_{+i}=x_{+i}\right)} \\
&= \frac{P\left(X_{0i}=x_{0i},X_{ij}=x_{+i}-x_{0i}\right)}{P\left(X_{+i}=x_{+i}\right)} \\
&= \frac{P\left(X_{0i}=x_{0i}\right)P\left(X_{ij}=x_{+i}-x_{0i}\right)}{P\left(X_{+i}=x_{+i}\right)} \\
&= \frac{\binom{2}{x_{0i}}p_{ij}^{x_{0i}}q_{ij}^{2-x_{0i}}\binom{2n_j}{x_{+i}-x_{0i}}p_{ij}^{x_{+i}-x_{0i}}q_{ij}^{2n_j-x_{+i}+x_{0i}}}{\binom{2\left(n_j+1\right)}{x_{+i}}p_{ij}^{x_{+i}}q_{ij}^{2(n_j+1)-x_{+i}}} \\
&= \frac{\binom{2}{x_{0i}}\binom{2n_j}{x_{+i}-x_{0i}}}{\binom{2\left(n_j+1\right)}{x_{+i}}}. \qquad (2.2)
\end{aligned}
$$

Den betingede sandsynlighed for $x_{0i}=0$ er givet ved

$$
\begin{aligned}
P\left(X_{0i}=0\mid X_{+i}=x_{+i}\right) &= \frac{\binom{2}{0}\binom{2n_j}{x_{+i}-0}}{\binom{2\left(n_j+1\right)}{x_{+i}}} \\
&= \frac{2n_j!}{x_{+i}!\left(2n_j-x_{+i}\right)!}\frac{x_{+i}!\left(2\left(n_j+1\right)-x_{+i}\right)!}{2\left(n_j+1\right)!} \\
&= \frac{2n_j!}{\left(2n_j-x_{+i}\right)!}\frac{\left(2\left(n_j+1\right)-x_{+i}\right)!}{2\left(n_j+1\right)\left(2\left(n_j+1\right)-1\right)\left(2\left(n_j+1\right)-2\right)!} \\
&= \frac{1}{\left(2n_j-x_{+i}\right)!}\frac{\left(2\left(n_j+1\right)-x_{+i}\right)!}{2\left(n_j+1\right)\left(2\left(n_j+1\right)-1\right)} \\
&= \frac{1}{\left(2n_j-x_{+i}\right)!}\frac{\left(2\left(n_j+1\right)-x_{+i}\right)\left(2\left(n_j+1\right)-x_{+i}-1\right)\left(2\left(n_j+1\right)-x_{+i}-2\right)!}{2\left(n_j+1\right)\left(2\left(n_j+1\right)-1\right)} \\
&= \frac{\left(2\left(n_j+1\right)-x_{+i}\right)\left(2n_j+1-x_{+i}\right)}{2\left(n_j+1\right)\left(2n_j+1\right)}.
\end{aligned}
$$

På tilsvarende måde findes den betingede sandsynlighed for $x_{0i}\in\{1,2\}$.

$$
P\left(X_{0i}=1\mid X_{+i}=x_{+i}\right) = 2\ \frac{\left(2\left(n_j+1\right)-x_{+i}\right)x_{+i}}{2\left(n_j+1\right)\left(2n_j+1\right)}
$$

$$
P\left(X_{0i}=2\mid X_{+i}=x_{+i}\right) = \frac{x_{+i}\left(x_{+i}-1\right)}{2\left(n_j+1\right)\left(2n_j+1\right)}.
$$

I dette tilfælde kan likelihood ratioen med de to forskellige hypoteser skrives som

$$
\begin{aligned}
Q\left(x_{0i}, x_{+i}\right) &= \frac{q\left(x_{0i}, x_{+i} \mid H_0\right)}{q\left(x_{0i}, x_{+i} \mid H_1\right)} \\
&= \frac{\left(\hat{p}_{+i}\right)^{x_{+i}} \left(1-\hat{p}_{+i}\right)^{2(n_j+1)-x_{+i}}}{\left(\hat{p}_{0i}\right)^{x_{0i}} \left(1-\hat{p}_{0i}\right)^{2-x_{0i}} \left(\hat{p}_{ij}\right)^{x_{+i}-x_{0i}} \left(1-\hat{p}_{ij}\right)^{2n_j-x_{+i}+x_{0i}}} \\
&= \frac{\left(\frac{x_{+i}}{2\left(n_j+1\right)}\right)^{x_{+i}} \left(1-\frac{x_{+i}}{2\left(n_j+1\right)}\right)^{2(n_j+1)-x_{+i}}}{\left(\frac{x_{0i}}{2}\right)^{x_{0i}} \left(1-\frac{x_{0i}}{2}\right)^{2-x_{0i}} \left(\frac{x_{+i}-x_{0i}}{2n_j}\right)^{x_{+i}-x_{0i}} \left(1-\frac{x_{+i}-x_{0i}}{2n_j}\right)^{2n_j-x_{+i}+x_{0i}}},
\end{aligned}
$$

hvor udtrykket $q(x_{0i}, x_{+i} \mid H_0)$ repræsenterer tilfældet, hvor x_{0i} og x_{ij} stammer fra den samme population, og estimatet af sandsynligheden for allel 1 på markør i i population j er $\hat{p}_{+i}$. Udtrykket $q(x_{0i}, x_{+i} \mid H_1)$ repræsenterer derimod tilfældet, hvor x_{0i} og x_{ij} ikke stammer fra den samme population. De første to faktorer angiver x_{0i}, hvor estimatet er $\hat{p}_{0i}$, og de sidste to faktorer angiver $x_{ij} = x_{+i} - x_{0i}$, hvor estimatet er $\hat{p}_{ij}$. Hvis der betinges med den sufficiente statistik x_{+i}, fører dette til Fishers exact test, hvor likelihood ratioen betinget med x_{+i} bliver

$$
Q\left(x_{0i}, x_{+i}\right) = Q\left(x_{0i} \mid x_{+i}\right) = \frac{q\left(x_{0i} \mid x_{+i}, H_0\right)}{q\left(x_{0i} \mid x_{+i}, H_1\right)}. \tag{2.3}
$$

Udtrykket $q(x_{0i} \mid x_{+i}, H_0)$ i (2.3) bliver nu blot en konstant, når middelværdien og variansen tages af den, da den ikke afhænger af x_{0i}. Tages $-\log(Q(x_{0i} \mid x_{+i}))$ kan bidraget fra $q(x_{0i} \mid x_{+i}, H_1)$ skrives som

$$
\begin{aligned}
\log\left(q\left(x_{0i} \mid x_{+i}, H_1\right)\right) &= \log\left(\left(\frac{x_{0i}}{2}\right)^{x_{0i}}\right) + \log\left(\left(1-\frac{x_{0i}}{2}\right)^{2-x_{0i}}\right) \\
&\quad + \log\left(\left(\frac{x_{+i}-x_{0i}}{2n_j}\right)^{x_{+i}-x_{0i}}\right) + \log\left(\left(1-\frac{x_{+i}-x_{0i}}{2n_j}\right)^{2n_j-x_{+i}+x_{0i}}\right) \\
&= x_{0i}\log\left(x_{0i}\right) - x_{0i}\log\left(2\right) + \left(2-x_{0i}\right)\log\left(2-x_{0i}\right) \\
&\quad - \left(2-x_{0i}\right)\log\left(2\right) + \left(x_{+i}-x_{0i}\right)\log\left(x_{+i}-x_{0i}\right) - \left(x_{+i}-x_{0i}\right)\log\left(2n_j\right) \\
&\quad + \left(2n_j-x_{+i}+x_{0i}\right)\log\left(2n_j-x_{+i}+x_{0i}\right) - \left(2n_j-x_{+i}+x_{0i}\right)\log\left(2n_j\right) \\
&= x_{0i}\log\left(x_{0i}\right) - x_{0i}\log\left(2\right) + \left(2-x_{0i}\right)\log\left(2-x_{0i}\right) \\
&\quad - \left(2-x_{0i}\right)\log\left(2\right) + \left(x_{+i}-x_{0i}\right)\log\left(x_{+i}-x_{0i}\right) - 2n_j\log\left(2n_j\right) \\
&\quad + \left(2n_j-x_{+i}+x_{0i}\right)\log\left(2n_j-x_{+i}+x_{0i}\right).
\end{aligned}
$$

Idet $x_{0i} \in \{0, 1, 2\}$, er den stokastiske del af $-\log Q(x_{0i} \mid x_{+i})$

$$
\left(x_{+i}-x_{0i}\right)\log\left(x_{+i}-x_{0i}\right) + \left(2n_j-x_{+i}+x_{0i}\right)\log\left(2n_j-x_{+i}+x_{0i}\right) - 2\log\left(2\right)\mathbb{I}\left(x_{0i}=1\right), \tag{2.4}
$$

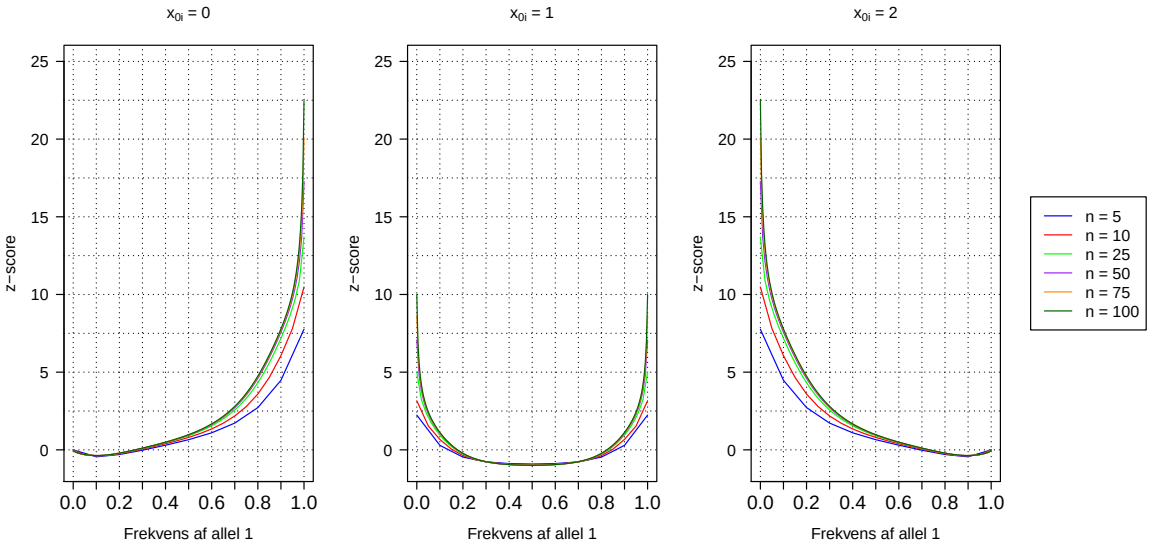
hvor det bruges, at $0 \cdot \log(0) = 0$, og $\mathbb{I}(\cdot)$ er en indikatorfunktion. For at standardisere $-\log Q(x_{0i} \mid x_{+i})$ tages middelværdien og variansen af den, så den standardiserede negative log-likelihood ratio kan skrives som

$$z_{ij} = \frac{-\log Q\left(x_{0i} \mid x_{+i}\right) + \mathbb{E}_{x_{0i}}\left[\log Q\left(x_{0i} \mid x_{+i}\right)\right]}{\sqrt{\mathbb{V}_{x_{0i}}\left[\log Q\left(x_{0i} \mid x_{+i}\right)\right]}}. \tag{2.5}$$

Denne værdi kaldes z_{ij}*-scoren* for markør i for population j, og den angiver, hvor angiveligt det er, at x_{0i} stammer fra den samme population som x_{ij}. Som det tidligere er antaget, er der uafhængighed hver især mellem markørerne og mellem allellerne. Det vil sige, at z_{ij}-scoren kan generaliseres til at gælde for alle allellerne. Dette gøres ved at summe alle ledene sammen i tælleren og nævneren af (2.5), hvilket giver

$$z_j = \frac{\sum_{i=1}^{I}\left(-\log Q\left(x_{0i} \mid x_{+i}\right) + \mathbb{E}_{x_{0i}}\left[\log Q\left(x_{0i} \mid x_{+i}\right)\right]\right)}{\sqrt{\sum_{i=1}^{I}\mathbb{V}_{x_{0i}}\left[\log Q\left(x_{0i} \mid x_{+i}\right)\right]}}. \tag{2.6}$$

Ligeledes kaldes denne værdi z_j*-scoren* for population j, og den angiver, hvor angiveligt det er, at $\mathbf{x}_0$ stammer fra den samme population som $\mathbf{x}_j$. Figur 2.1 viser tre illustrationer af z-scoren for forskellige frekvenser af allel 1 og forskellige valg af n.



Figur 2.1: z-scoren for forskellige frekvenser af allel 1 og forskellige valg af populationens størrelse.

Den første illustration repræsenterer situationen, hvor $x_{0i} = 0$. Jo højere frekvens af allel 1 giver en højere z-score, og det bliver derfor mere antageligt, at $\mathbf{x}_0$ ikke stammer fra $\mathbf{x}_j$. Den anden illustration angiver situationen, hvor $x_{0i} = 1$. En lav og høj frekvens af allel 1 giver en høj z-score. Her er det altså mest antageligt, at $\mathbf{x}_0$ stammer fra $\mathbf{x}_j$, hvis $\mathbf{x}_j$ har en frekvens af allel 1 på omkring 0,5. Den tredje illustration repræsenterer situationen, hvor $x_{0i} = 2$, hvilket er det modsatte scenarie af den første illustration.

Fordelingen af z_j-scoren under nulhypotesen kan approksimeres ved standard normalfordelingen jævnfør den centrale grænseværdisætning. Der kan dog forekomme svagheder i markørerne i form af frafald på grund af antallet eller kvaliteten af dem, eller nogle af markørerne kan være fixed i en population. Med udtrykket fixed menes der, at nogle af markørerne kan være fast bestemt for en population. Det vil sige, at frekvensen af allel 1 på en bestemt markør er ens for alle i populationen. Disse svagheder kan forårsage afvigelser fra normaliteten, men for at undgå dette i analysen er der brug for numeriske metoder, når p-værdien udregnes for test-statistikken.

2.2 Hypotesetest med tri-alleliske markører

Dette afsnit er baseret på [Tvedebrink et al., 2018].

I dette afsnit opstilles en hypotesetest, som følger samme fremgangsmåde som i afsnit 2.1, men i stedet for bi-alleliske markører indgår der tri-alleliske markører. På grund af, at markørerne er tri-alleliske, fokuseres der nu på frekvensen af både det første og andet allel, kaldet *allel 1* og *allel 2*.

Antag derfor nu, at I tri-alleliske markører er genotypet i J forskellige populationer for n_j profiler. Frekvensen af allel 1, 2 og 3 for en population er dermed angivet som

$$x_{ji} = \left\{ \begin{bmatrix} x_{jiA} & x_{jiC} & x_{jiT} \end{bmatrix}^\top \;\middle|\; \{x_{jiA}, x_{jiC}, x_{jiT}\} \in \{0, 1, \ldots, 2n_j\}\,, \sum_{k \in \{A,C,T\}} x_{jik} = 2n_j \right\} \tag{2.7}$$

for markør $i \in \{1, 2, \ldots, I\}$ for population $j \in \{1, 2, \ldots, J\}$. Det giver, at $\mathbf{x}_j = \begin{bmatrix} \mathbf{x}_{j1} & \mathbf{x}_{j1} & \cdots & \mathbf{x}_{jI} \end{bmatrix}^\top$ er en vektor med vektorer, som angiver frekvensen af allel 1, 2 og 3 på forskellige markører i population j. En enkel specifik profil er angivet som $\mathbf{x}_0 = \begin{bmatrix} \mathbf{x}_{01} & \mathbf{x}_{02} & \cdots & \mathbf{x}_{0I} \end{bmatrix}^\top$, som er vektoren over alle profilens markører, hvor indgangene $\mathbf{x}_{0i} = \begin{bmatrix} x_{0iA} & x_{0iC} & x_{0iT} \end{bmatrix}^\top$ igen er vektorer, som angiver antallet af hvert allel, hvor $\{x_{0iA}, x_{0iC}, x_{0iT}\} \in \{0, 1, 2\}$ og $\sum_{k \in \{A,C,T\}} x_{0ik} = 2$. Antag igen, at markørerne og allellerne inden for en population hver især er indbyrdes uafhængige. Dette medfører, at $\mathbf{X}_{0i}$ og $\mathbf{X}_{ji}$ er multinomialfordelt, så under nulhypotesen er $\mathbf{X}_{0i} \sim \text{mult}(2, \mathbf{p}_{ji})$ og $\mathbf{X}_{ji} \sim \text{mult}(2n_j, \mathbf{p}_{ji})$, hvor $\mathbf{p}_{ji} = \begin{bmatrix} p_{jiA} & p_{jiC} & p_{jiT} \end{bmatrix}$ er sandsynligheden for allel A, C og T for population j på markør i. Endvidere haves det, at $\mathbf{X}_{+i} = \mathbf{X}_{0i} + \mathbf{X}_{ji} \sim \text{mult}(2(n_j + 1), \mathbf{p}_{ji})$. Det ønskes igen at betinge med $\mathbf{x}_{+i}$ i likehood ratioen, og den betingede sandsynlighed bliver

$$\begin{aligned}
P\left(\mathbf{X}_{0i} = \mathbf{x}_{0i} \mid \mathbf{X}_{+i} = \mathbf{x}_{+i}\right) &= \frac{P\left(\mathbf{X}_{0i} = \mathbf{x}_{0i}\right) P\left(\mathbf{X}_{ji} = \mathbf{x}_{+i} - \mathbf{x}_{0i}\right)}{P\left(\mathbf{X}_{+i} = \mathbf{x}_{+i}\right)} \\
&= \frac{\binom{2}{\mathbf{x}_{0i}} \prod\limits_{k \in \{A,C,T\}} p_{jik}^{x_{0ik}} \binom{2n_j}{\mathbf{x}_{+i} - \mathbf{x}_{0i}} \prod\limits_{k \in \{A,C,T\}} p_{jik}^{x_{+ik} - x_{0ik}}}{\binom{2\left(n_j + 1\right)}{\mathbf{x}_{+i}} \prod\limits_{k \in \{A,C,T\}} p_{jik}^{x_{+ik}}} \\
&= \frac{\binom{2}{\mathbf{x}_{0i}} \binom{2n_j}{\mathbf{x}_{+i} - \mathbf{x}_{0i}}}{\binom{2\left(n_j + 1\right)}{\mathbf{x}_{+i}}} \\
&= \frac{2!\; 2n_j!}{\prod\limits_{k \in \{A,C,T\}} x_{0ik}!\left(x_{+ik} - x_{0ik}\right)!} \frac{\prod\limits_{k \in \{A,C,T\}} x_{+ik}!}{2\left(n_j + 1\right)!} \\
&= \frac{\prod\limits_{k \in \{A,C,T\}} \binom{x_{+ik}}{x_{0ik}}}{\binom{2\left(n_j + 1\right)}{2}}.
\end{aligned}$$

Den betingede sandsynlighed for $\mathbf{x}_{0i} = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^\top$ er givet ved

$$\begin{aligned} P\left(\mathbf{X}_{0i} = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^\top \;\middle|\; \mathbf{X}_{+i} = \mathbf{x}_{+i}\right) &= \frac{\binom{x_{+iA}}{1}\binom{x_{+iC}}{1}\binom{x_{+iT}}{0}}{\binom{2\,(n_j+1)}{2}} \\ &= \frac{x_{+iA}!}{(x_{+iA}-1)!}\frac{x_{+iC}!}{(x_{+iC}-1)!}\frac{2!\,(2\,(n_j+1)-2)!}{2\,(n_j+1)!} \\ &= \frac{x_{+iA}\,(x_{+iA}-1)!}{(x_{+iA}-1)!}\frac{x_{+iC}\,(x_{+iC}-1)!}{(x_{+iC}-1)!} \\ &\quad \frac{2!\,(2\,(n_j+1)-2)!}{2\,(n_j+1)\,(2\,(n_j+1)-1)\,(2\,(n_j+1)-2)!} \\ &= 2\;\frac{x_{+iA}\;x_{+iC}}{2\,(n_j+1)\,(2n_j+1)}. \end{aligned}$$

På tilsvarende måde findes den betingede sandsynlighed for de resterende $\mathbf{x}_{0i}$.

$$P\left(\mathbf{X}_{0i} = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}^\top \;\middle|\; \mathbf{X}_{+i} = \mathbf{x}_{+i}\right) = 2\;\frac{x_{+iA}\;x_{+iT}}{2\,(n_j+1)\,(2n_j+1)}$$

$$P\left(\mathbf{X}_{0i} = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}^\top \;\middle|\; \mathbf{X}_{+i} = \mathbf{x}_{+i}\right) = 2\;\frac{x_{+iC}\;x_{+iT}}{2\,(n_j+1)\,(2n_j+1)}$$

$$P\left(\mathbf{X}_{0i} = \begin{bmatrix} 2 & 0 & 0 \end{bmatrix}^\top \;\middle|\; \mathbf{X}_{+i} = \mathbf{x}_{+i}\right) = \frac{x_{+iA}\,(x_{+iA}-1)}{2\,(n_j+1)\,(2n_j+1)}$$

$$P\left(\mathbf{X}_{0i} = \begin{bmatrix} 0 & 2 & 0 \end{bmatrix}^\top \;\middle|\; \mathbf{X}_{+i} = \mathbf{x}_{+i}\right) = \frac{x_{+iC}\,(x_{+iC}-1)}{2\,(n_j+1)\,(2n_j+1)}$$

$$P\left(\mathbf{X}_{0i} = \begin{bmatrix} 0 & 0 & 2 \end{bmatrix}^\top \;\middle|\; \mathbf{X}_{+i} = \mathbf{x}_{+i}\right) = \frac{x_{+iT}\,(x_{+iT}-1)}{2\,(n_j+1)\,(2n_j+1)}.$$

Likelihood ratioen med de to forskellige hypoteser kan skrives som

$$\begin{aligned} Q\left(\mathbf{x}_{0i}, \mathbf{x}_{+i}\right) &= \frac{q\left(\mathbf{x}_{0i}, \mathbf{x}_{+i} \mid H_0\right)}{q\left(\mathbf{x}_{0i}, \mathbf{x}_{+i} \mid H_1\right)} \\ &= \frac{\left(\frac{x_{+iA}}{2\,(n_j+1)}\right)^{x_{+iA}}\left(\frac{x_{+iC}}{2\,(n_j+1)}\right)^{x_{+iC}}\left(\frac{x_{+iT}}{2\,(n_j+1)}\right)^{x_{+iT}}}{\left(\frac{x_{0iA}}{2}\right)^{x_{0iA}}\left(\frac{x_{0iC}}{2}\right)^{x_{0iC}}\left(\frac{x_{0iT}}{2}\right)^{x_{0iT}}\left(\frac{x_{+iA}-x_{0iA}}{2n_j}\right)^{x_{+iA}-x_{0iA}}\left(\frac{x_{+iC}-x_{0iC}}{2n_j}\right)^{x_{+iC}-x_{0iC}}\left(\frac{x_{+iT}-x_{0iT}}{2n_j}\right)^{x_{+iT}-x_{0iT}}}. \end{aligned}$$

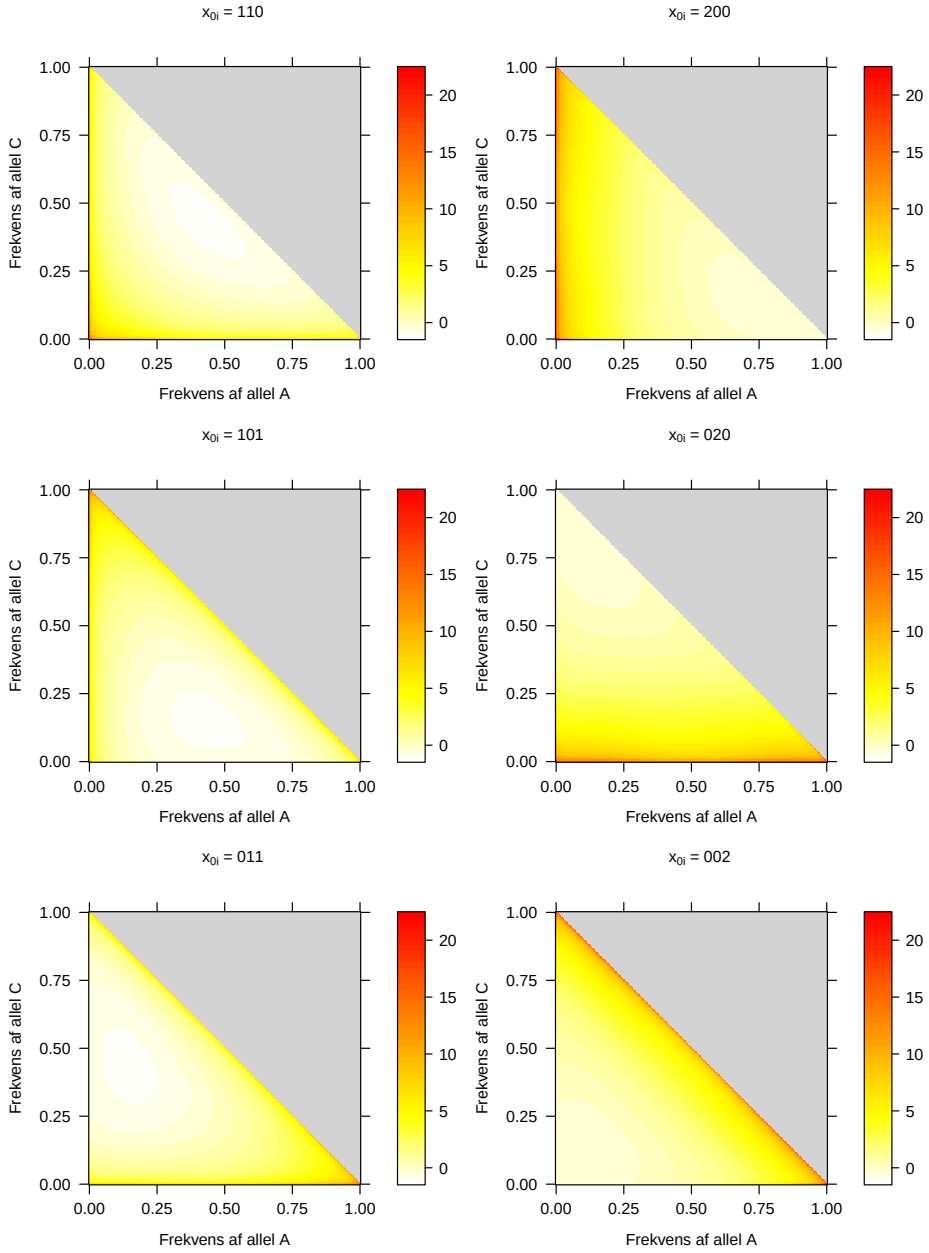
Dermed kan den standardiserede negative log likelihood ratio skrives som

$$\mathbf{z}_{ij} = \frac{-\log Q\left(\mathbf{x}_{0i} \mid \mathbf{x}_{+i}\right) + \mathbb{E}_{\mathbf{x}_{0i}}\left[\log Q\left(\mathbf{x}_{0i} \mid \mathbf{x}_{+i}\right)\right]}{\sqrt{\mathbb{V}_{\mathbf{x}_{0i}}\left[\log Q\left(\mathbf{x}_{0i} \mid \mathbf{x}_{+i}\right)\right]}}, \tag{2.8}$$

og på den generaliserede form, hvor summen tages over alle allellerne, skrives den som

$$\mathbf{z}_j = \frac{\sum_{i=1}^{I}\Big(-\log Q\left(\mathbf{x}_{0i} \mid \mathbf{x}_{+i}\right) + \mathbb{E}_{\mathbf{x}_{0i}}\left[\log Q\left(\mathbf{x}_{0i} \mid \mathbf{x}_{+i}\right)\right]\Big)}{\sqrt{\sum_{i=1}^{I}\mathbb{V}_{\mathbf{x}_{0i}}\left[\log Q\left(\mathbf{x}_{0i} \mid \mathbf{x}_{+i}\right)\right]}}. \tag{2.9}$$

Figur 2.2 består af seks illustrationer, som viser intensiteten af z-scoren for $\mathbf{x}_{0i}$. Overskriften på hver illustration angiver, om $\mathbf{x}_{0i} = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^\top$, $\mathbf{x}_{0i} = \begin{bmatrix} 2 & 0 & 0 \end{bmatrix}^\top$ osv.



Figur 2.2: z-scoren for forskellige frekvenser af allel 1, 2 og 3 samt en populationsstørrelse på 100 profiler. Overskriften på hver illustration angiver værdien af $\mathbf{x}_{0i}$, og det grå område angiver NA-værdier.

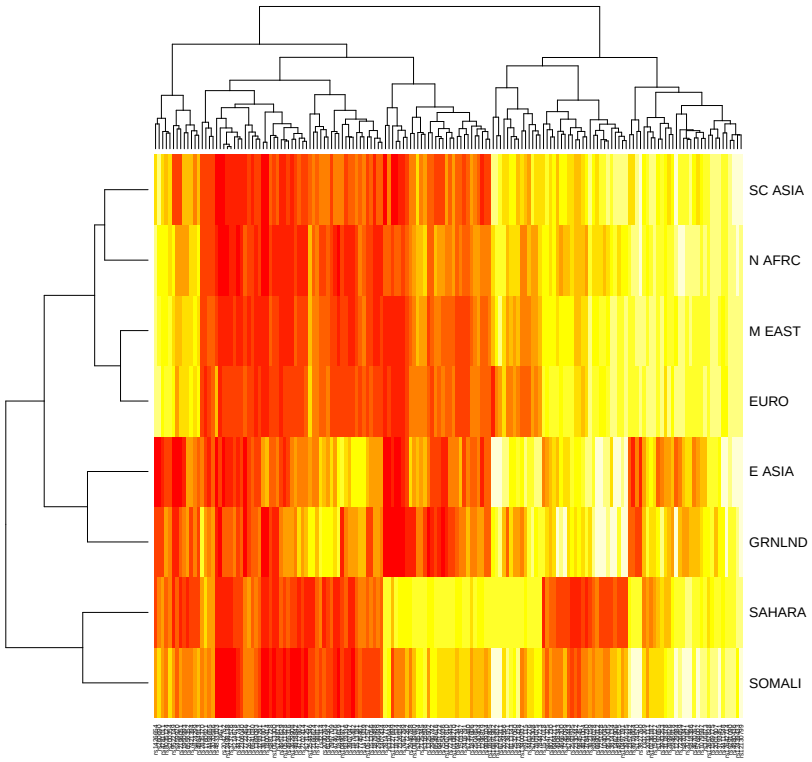
Det kan ses i Figur 2.2, at der fremkommer den samme tendens på alle seks illustrationer. Intensiteten er højest i det område, hvor frekvensen er højest af det allel, som er 0 i vektoren $\mathbf{x}_{0i}$. For eksempel i den første illustration er $\mathbf{x}_{0i} = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}^\top$. Her er $x_{0iT} = 0$, så intensiteten er højest, der hvor frekvensen af T-allellet er højest. Jævnfør $p_{jiA}^{x_{0iA}} + p_{jiC}^{x_{0iC}} + p_{jiT}^{x_{0iT}} = 1$, bliver frekvensen af T-allellet højere, jo lavere frekvensen af A-allellet og C-allellet er. En højere frekvens af T-allellet giver dermed en højere intensitet og en højere z-score.

3 Analyse

I dette kapitel laves en analyse af datasættene ved anvendelse af teorien fra kapitel 2. I afsnit 3.1 forsøges det ved anvendelse af matematiske modeller at udvælge en delmængde af markørerne, som bedst skelner mellem hovedpopulationerne. I afsnit 3.2 analyseres markørernes sensitivitet med hensyn til at acceptere og forkaste hovedpopulationerne. Dette er gjort for simulerede profiler, som er accepteret eller forkastet af deres egen hovedpopulation, men henholdsvis bliver forkastet eller accepteret efter en ændring på den pågældende markør. I afsnit 3.3 undersøges fortyndingsdatasættet i forhold til markørernes coverage og tilbøjelighed til at være et dropout.

3.1 Udvælgelse af relevante markører

Fra det første datasæt er der hentet frekvensen af de observerede allel 1 fra hovedpopulationerne, altså x_{ij} for markør i for hovedpopulation j. Denne x_{ij} er divideret med $2n_j$, som er antallet af de mulige allel 1 for hovedpopulation j. Dette giver frekvensen over, hvor mange allel 1 for de forskellige markører der er i hovedpopulationerne, som er vist i Figur 3.1.



Figur 3.1: Heatmap over frekvensen af de observerede allel 1 for hver markør i de forskellige hovedpopulationer. De lyse/gule og mørke/røde farver indikerer henholdsvis lave og høje frekvenser. Der er brugt *complete* som linkage og den euklidiske norm som distance.

I Figur 3.1 ses det tydeligt, at der er en forskel mellem frekvenserne, både inden for markørerne og hovedpopulationerne. Det ses også, at der er delmængder af markørerne, som har høje frekvenser for delmængder af hovedpopulationerne. Et eksempel på dette kunne være det røde område øverst til venstre,

som viser mange høje frekvenser for mange af de første markører og hovedpopulationerne SC ASIA, N AFRC, M EAST og EURO. Det kunne derfor være interessant at finde forskellige delmængder af markørerne, som kunne være nyttige til at skelne mellem forskellige grupper af hovedpopulationerne. I dette afsnit vil dette blive undersøgt.

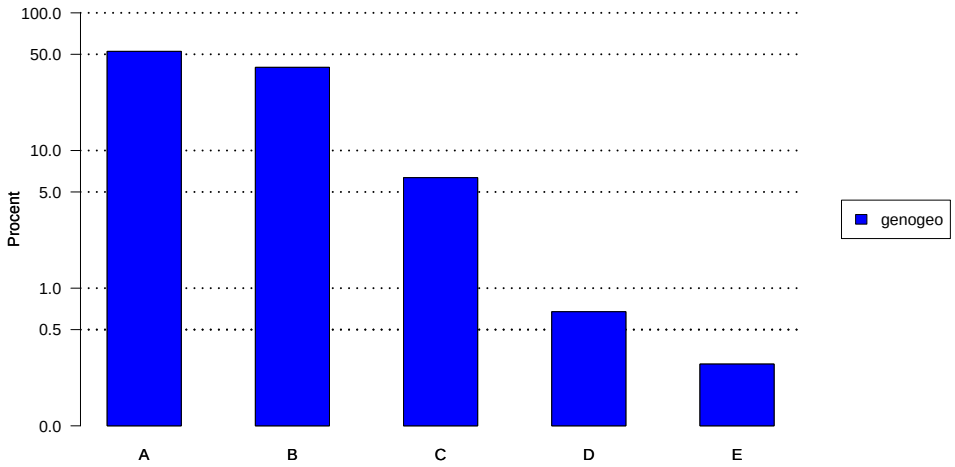
3.1.1 Udvælgelse med genogeo

Der startes med en anvendelse af funktionen `genogeo`, som udregner z_j-scoren for hver hovedpopulation i forhold til $\mathbf{x}_0$ og bekræfter med hensyn til de to hypoteser (2.1), hvilke hovedpopulationer $\mathbf{x}_0$ kan stamme fra. Dette er gjort for alle 3.560 profiler, hvor udfaldet er klassificeret i fem grupper. Beskrivelsen af gruppens profiler er beskrevet i Tabel 3.1.

Gruppe	Beskrivelse
A	Profilen er kun accepteret af den sande hovedpopulation.
B	Profilen er accepteret af flere hovedpopulationer inklusiv den sande hovedpopulation.
C	Profilen er ikke accepteret af nogen hovedpopulation.
D	Profilen er accepteret af én hovedpopulation, som ikke er den sande hovedpopulation.
E	Profilen er accepteret af flere hovedpopulationer eksklusiv den sande hovedpopulation.

Tabel 3.1: Klassificeringen af grupperne.

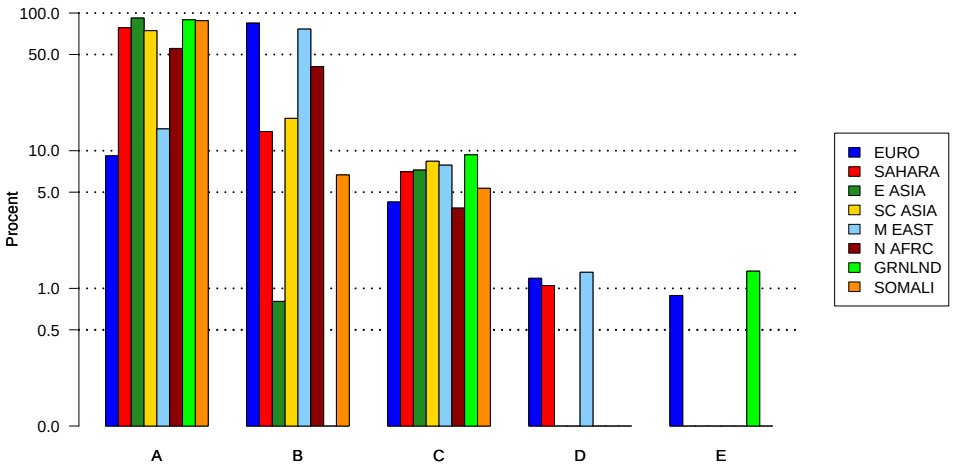
Procentandelen af klassificeringen fra udfaldet af `genogeo` kan ses i Figur 3.2.



Figur 3.2: Procentandel på logaritmisk skala af klassificeringen for `genogeo`.

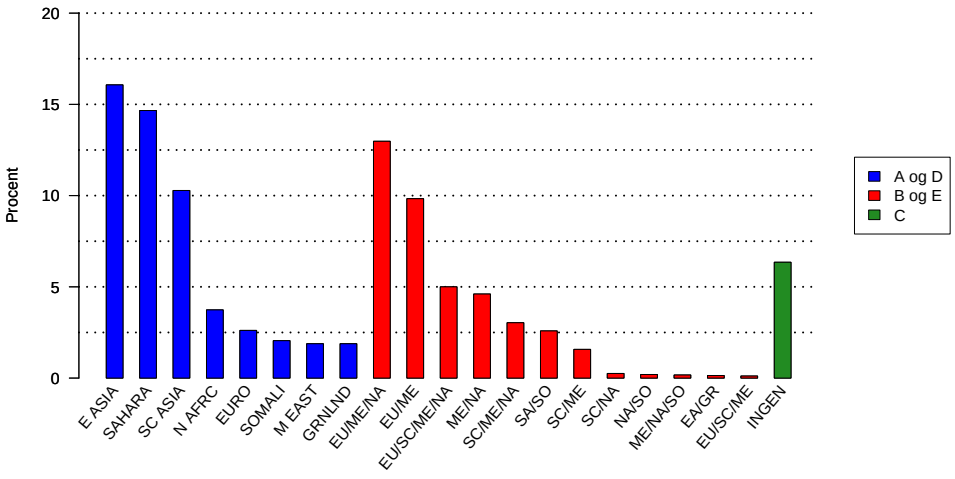
Det ses i Figur 3.2, at lidt over 50% af profilerne er blevet accepteret kun af den sande hovedpopulation, men samtidig er der stadig omkring 40% af profilerne, som bliver accepteret af flere hovedpopulationer inklusiv den sande hovedpopulation. Omkring 6% af profilerne bliver ikke accepteret af nogen hovedpopulation, mens de sidste to grupper udgør mindre end 1%. Det ønskes at forøge andelen af gruppe A og reducere andelen af gruppe B, så der er flere profiler, som kun bliver accepteret af den sande hovedpopulation. Samtidig skal andelen af gruppe C, D og E holdes nede på et minimum, så det mindst mulige antal af profilerne bliver misklassificeret. Dette bliver uddybet i afsnit 3.1.2.

Det undersøges nærmere, om der er nogen hovedpopulationer, som skiller sig ud fra de andre hovedpopulationer med hensyn til klassificeringen. Dette er illustreret i Figur 3.3, hvor klassificeringen inden for hver hovedpopulation bliver betragtet.



Figur 3.3: Procentandel på logaritmisk skala af klassificeringen inden for hver hovedpopulation.

Det fremgår af Figur 3.3, at EURO, M EAST og N AFRC har en meget lille andel af gruppe A, og en høj andel af gruppe B i forhold til de andre hovedpopulationer. Dette underbygger altså observationen fra Figur 1.5, at det kan være svært at skelne mellem disse hovedpopulationer. Alle hovedpopulationerne har andelen af gruppe C nede under 10%, mens for de få hovedpopulationer, som har en andel i gruppe D og E, er andelen omkring 1%. For at undersøge de tre omtalte hovedpopulationer nærmere er procentandelen af delmængderne fra udfaldet af `genogeo` fundet, hvilket er illustreret i Figur 3.4.



Figur 3.4: Procentandel af delmængderne fra udfaldet af `genogeo`. Navnet for delmængderne med flere end én hovedpopulation er forkortet til de to første bogstaver af hovedpopulationens forkortelse.

Det ses i Figur 3.4, at udfaldet af `genogeo` kan inddeles i 21 delmængder. De første otte blå delmængder er udfaldet, hvor `genogeo` kun har accepteret én hovedpopulation for profilen, altså gruppe A og D sammenlagt. De næste 12 røde delmængder er udfaldet, hvor `genogeo` har accepteret flere end én hovedpopulation, altså gruppe B og E sammenlagt. Her vises navnet på delmængden som de to første bogstaver af hovedpopulationens forkortelse. Den sidste grønne delmængde er den tomme mængde, hvor `genogeo` ikke har accepteret nogen hovedpopulation, altså gruppe C. Det ses også, at de to første af de røde delmængder, EU/ME/NA og EU/ME, har de største procentandele med udgangspunkt i de røde delmængder, hvilket underbygger observationerne fra Figur 1.5 og Figur 3.3.

3.1.2 Udvælgelse af markører med CART

Det ønskes som sagt at forøge andelen af gruppe A og reducere andelen af gruppe B fra forrige afsnit, hvor klassifikationsmetoden CART (Classification- and Regressiontrees) vil blive anvendt i et forsøg på at løse dette. Metoden vil blive anvendt på de profiler, som er accepteret af flere end én hovedpopulation, og den anvender kun de accepterede hovedpopulationer, som blev udvalgt med `genogeo`. Hvis en profil først er blevet forkastet for at stamme fra en vis hovedpopulation med `genogeo`, tages denne hovedpopulation altså ikke i betragtning igen. Metoden vil finde en delmængde af markørerne, hvor den mener, at denne delmængde bedst kan skelne mellem de accepterede hovedpopulationer. Denne delmængde af markører vil så blive anvendt i `genogeo` til at udregne en ny z_j-score for hver hovedpopulation i forhold til $\mathbf{x}_0$. Dette kan medføre, at en eller flere af de accepterede hovedpopulationer fra `genogeo` nu bliver afvist. Det kan dog også ske, at metoden afviser den sande hovedpopulation eller afviser alle de accepterede hovedpopulationer, men metoden anvendes til trods for dette.

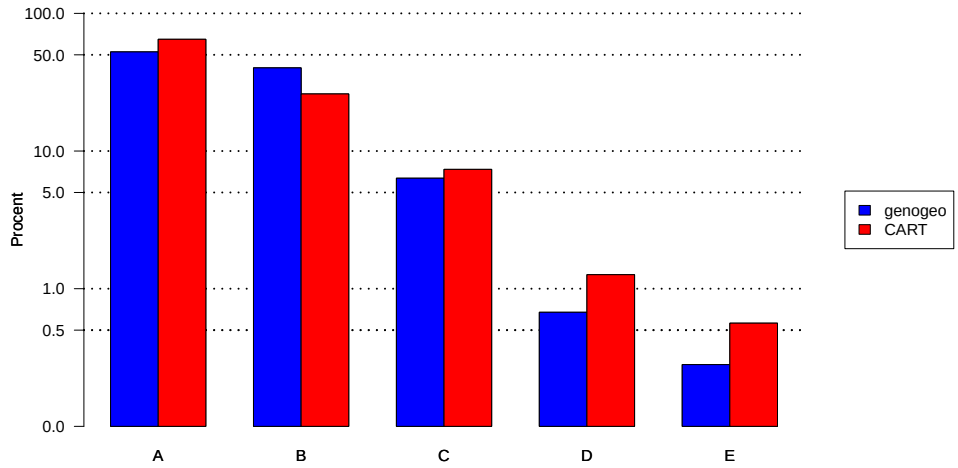
Efter at have anvendt CART på de profiler, som blev accepteret af flere hovedpopulationer, er metoden anvendt yderligere to gange på de profiler, som fortsat blev accepteret af flere hovedpopulationer. Efter tredje iteration med metoden var der ingen af profilerne, som fortsat blev klassificeret til en anden gruppe. For hver ny iteration blev CART anvendt med den nye delmængde af de accepterede hovedpopulationer fra forrige iteration. I Tabel 3.2 vises et eksempel med fire af profilerne fra datasættet, hvor det er angivet, hvilke hovedpopulationer `genogeo` og de tre iterationer med CART accepterer.

Profil	Hovedpopulation	genogeo	CART iteration 1	CART iteration 2	CART iteration 3
HG03006	SC ASIA	SC ASIA	-	-	-
NA11919	EURO	EURO, M EAST, N AFRC	EURO, M EAST	EURO	-
BT940	M EAST	EURO, M EAST	(ingen)	-	-
JK4912	E ASIA	(ingen)	-	-	-

Tabel 3.2: De accepterede hovedpopulationer fra `genogeo` og de tre iterationer med CART for fire af profilerne. Tegnet "-" angiver ingen anvendelse af den pågældende iteration med CART.

I Tabel 3.2 ses det, at profil HG03006 stammer fra SC ASIA og bliver også accepteret af SC ASIA med `genogeo`, så profilen bliver klassificeret i gruppe A og bliver ikke anvendt i analysen med CART. Profil NA11919 stammer fra EURO, men profilen bliver accepteret af EURO, M EAST og N AFRC med `genogeo`. Profilen bliver efter første iteration med CART accepteret af EURO og M EAST og efter anden iteration kun accepteret af EURO, så profilen bliver også klassificeret i gruppe A. Profil BT940 stammer

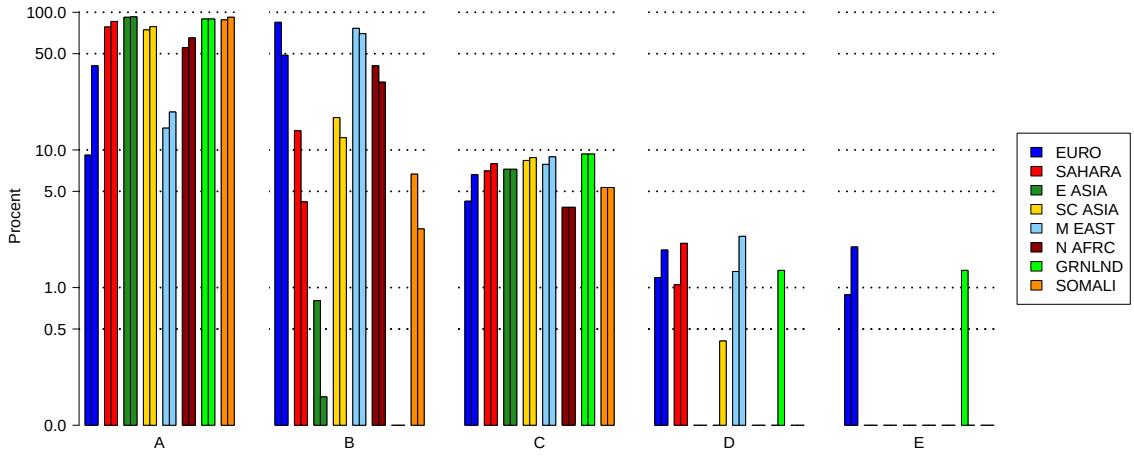
fra M EAST, men profilen bliver accepteret af EURO og M EAST med `genogeo`. Profilen bliver efter første iteration med CART ikke accepteret af nogen hovedpopulation og bliver så klassificeret i gruppe C. Profil JK4912 stammer fra E ASIA, men profilen bliver ikke accepteret af nogen hovedpopulation med `genogeo` og bliver så også klassificeret i gruppe C. En sammenligning af procentandelen af klassificeringen for `genogeo` og tredje iteration med CART, som blot kaldes *CART*, kan ses nedenfor i Figur 3.5.



Figur 3.5: Procentandel på logaritmisk skala af klassificeringen for `genogeo` og CART.

Det ses i Figur 3.5, at anvendelsen med CART har forøget andelen af gruppe A og reduceret andelen af gruppe B som ønsket. Andelen af gruppe A er efter tredje iteration med CART over 60%, og andelen af gruppe B er under 30%. Anvendelsen har dog også givet en lille forøgelse i gruppe C, D og E, hvor gruppe C er kommet op på omkring 7%, og gruppe D og E er samlet set nede under 2%.

Hovedpopulationerne hver især bliver igen undersøgt efter anvendelsen med CART. En sammenligning for `genogeo` og tredje iteration med CART er vist i Figur 3.6.

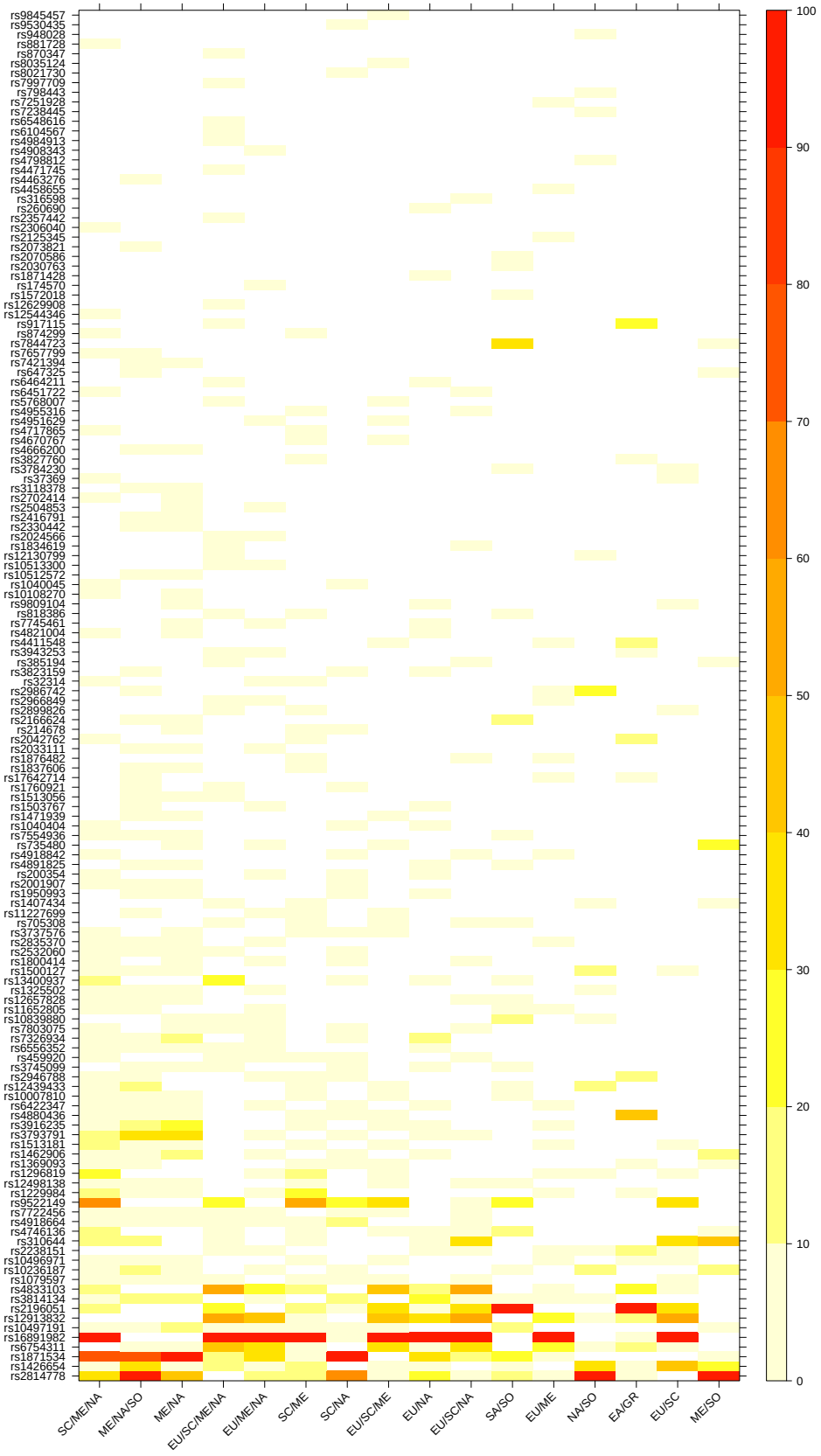


Figur 3.6: Procentandel på logaritmisk skala af klassificeringen inden for hver hovedpopulation. For hver gruppe og hovedpopulation indikerer den første farvede søjle `genogeo`, og den anden farvede søjle indikerer CART.

Det fremgår af Figur 3.6, at anvendelsen med CART har haft en stor betydning for EURO, hvor andelen af gruppe A og B nu er på omkring 40% og 50%. Med hensyn til de andre hovedpopulationer har de også fået små forbedringer.

Alle markørerne, som er taget i betragtning i analysen med CART, har en vægt, som siger noget om, hvor vigtig den er i analysen. I CART er der anvendt 142 markører, hvor vægten for hver markør for hver delmængde af hovedpopulationer er divideret med den største vægt fra markørerne inden for den pågældende delmængde af hovedpopulationer. Omregnet til procent giver dette et indblik i, hvor stor en indflydelse markørerne har i forhold til den vigtigste markør i delmængden, hvilket er illustreret i Figur 3.7. Bemærk, at fire nye delmængder af hovedpopulationer er tilføjet: EU/NA, EU/SC/NA, EU/SC og ME/SO. Disse delmængder er skabt ud fra de 12 delmængder af hovedpopulationer, da nogle profiler kunne forkaste enkelte hovedpopulationer og kun stå tilbage med en af de fire delmængder. Det vil sige, at der nu er 16 delmængder af hovedpopulationer.

Det fremgår af Figur 3.7, at rs2814778 optræder i 14 delmængder og dermed er den markør, som er mest hyppig. Derudover optræder de næste ni markører fra rs2814778 i mellem 10 og 13 delmængder, hvor specielt rs16891982 skiller sig ud ved at have den største vægt for flere af delmængderne. Det bemærkes også, at rs1871534, rs12913832 og rs2196051 har store vægte for flere delmængder.

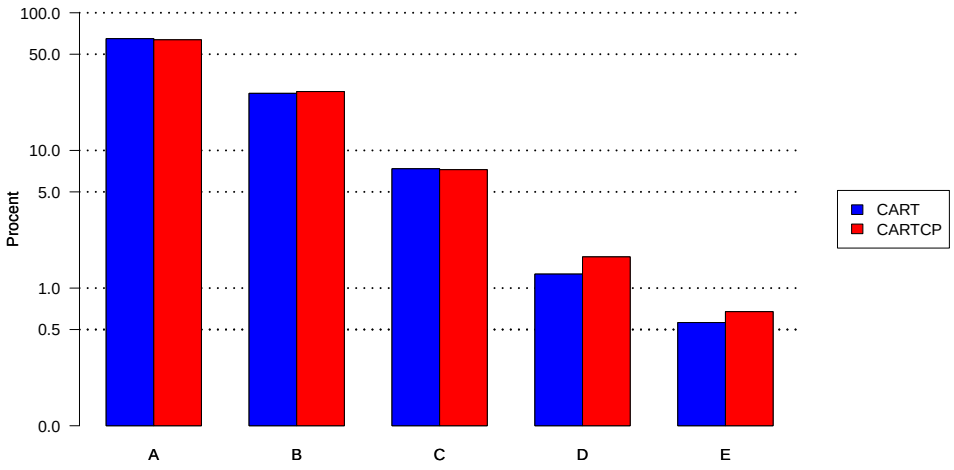


Figur 3.7: Relative vægte af markørerne fra CART for de forskellige delmængder af hovedpopulationer. Markørerne og delmængderne af hovedpopulationer er sorteret efter hyppighed. Farven hvid indikerer, at markøren ikke er anvendt i den pågældende delmængde. Navnet for delmængderne er forkortet til de to første bogstaver af hovedpopulationens forkortelse.

3.1.3 Udvælgelse af markører med CP-værdi

Jævnfør Figur 3.1 er der mange markører, som har en lav frekvens af allel 1, og jævnfør Figur 3.7 er der mange markører, som har små vægte i analysen med CART. Når der analyseres på, hvilke hovedpopulationer, der bliver accepteret, kan disse markører blot være støj i analysen og dermed være med til at forringe resultatet. Dette vil nu blive taget i betragtning i et forsøg på at forbedre resultatet fra forrige afsnit. I analysen med CART er der markører, som er betydelig mere relevante end andre markører. Disse relevante markører tænkes at være vigtigere i analysen i forhold til de andre markører, der ses som den såkaldte støj. I udvælgelsen af disse relevante markører anvendes der en såkaldt CP-værdi, som er en værdi, der afskaffer flere af de mindst vigtige markører i analysen med CART, jo højere værdien er.

Denne analyse følger den samme proces, som blev vist i Tabel 3.2. CP-værdien er fundet for hver af delmængderne i Figur 3.4, som havde flere end én hovedpopulation. I Figur 3.8 ses en sammenligning af procentandelen af klassificeringen for CART og CART med anvendelse af CP-værdi, som kaldes *CARTCP*.



Figur 3.8: Procentandel på logaritmisk skala af klassificeringen for CART og CARTCP.

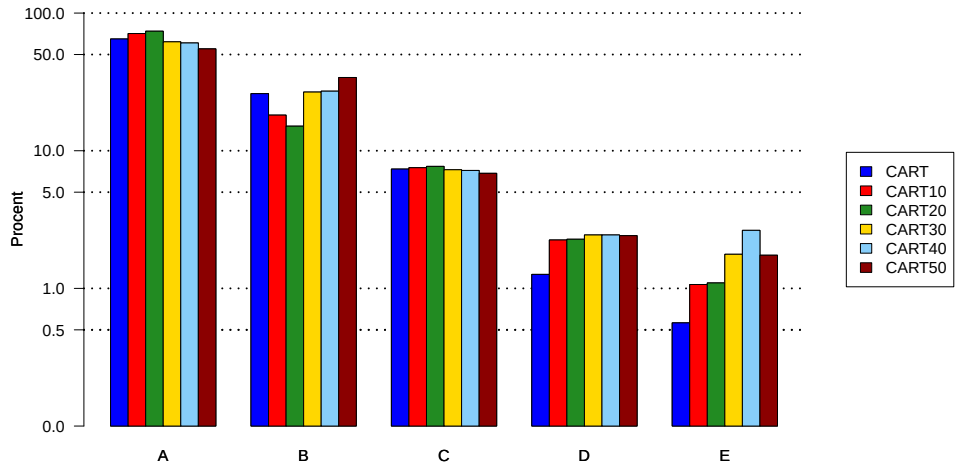
Det ses i Figur 3.8, at der ikke er den store forskel mellem CART og CARTCP.

Analysen med hovedpopulationerne er igen lavet med anvendelse af en CP-værdi, men grundet den lille forskel er den udeladt her.

3.1.4 Udvælgelse af markører med grænser

På trods af den meget lille forbedring af resultatet med anvendelse af CP-værdi fortsættes der med at udvælge markører på andre måder. Som sagt har alle markørerne fra analysen med CART en vægt, som siger noget om, hvor vigtig den er i analysen. I dette afsnit fokuseres der på, hvad der sker, hvis der bliver sat grænser på, hvor stor vægten skal være for en markør for at forblive i analysen. Jævnfør Figur 3.7 er disse grænser sat til 10%, 20%, 30%, 40% og 50%, da vægten for de fleste markører ligger under 50%.

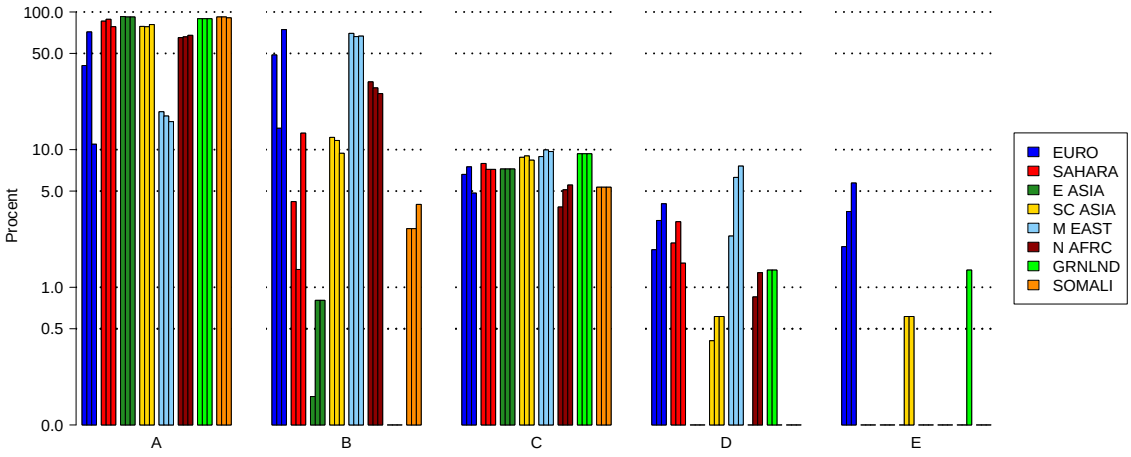
Denne analyse følger igen den samme proces, som blev vist i Tabel 3.2. I Figur 3.9 ses en sammenligning af procentandelen af klassificeringen for CART og de fem grænser med CART, som kaldes *CART10*, *CART20*, *CART30*, *CART40* og *CART50*.



Figur 3.9: Procentandel på logaritmisk skala af klassificeringen for CART, CART10, CART20, CART30, CART40 og CART50.

Det ses i Figur 3.9, at CART20 er den model, som har den største andel i gruppe A, den mindste andel i gruppe B og en lille forøgelse i de sidste tre grupper. Jævnfør Tabel 5.1 anvender modellen kun 21 markører, så det vælges at gå videre med denne model. CART50 er den model, som har den største andel i gruppe B, og jævnfør Tabel 5.1 anvender den kun syv markører. Det vælges også at gå videre med denne model, da den kun anvender syv markører, men selvom den har en stor andel i gruppe B, accepterer den stadig den sande hovedpopulation.

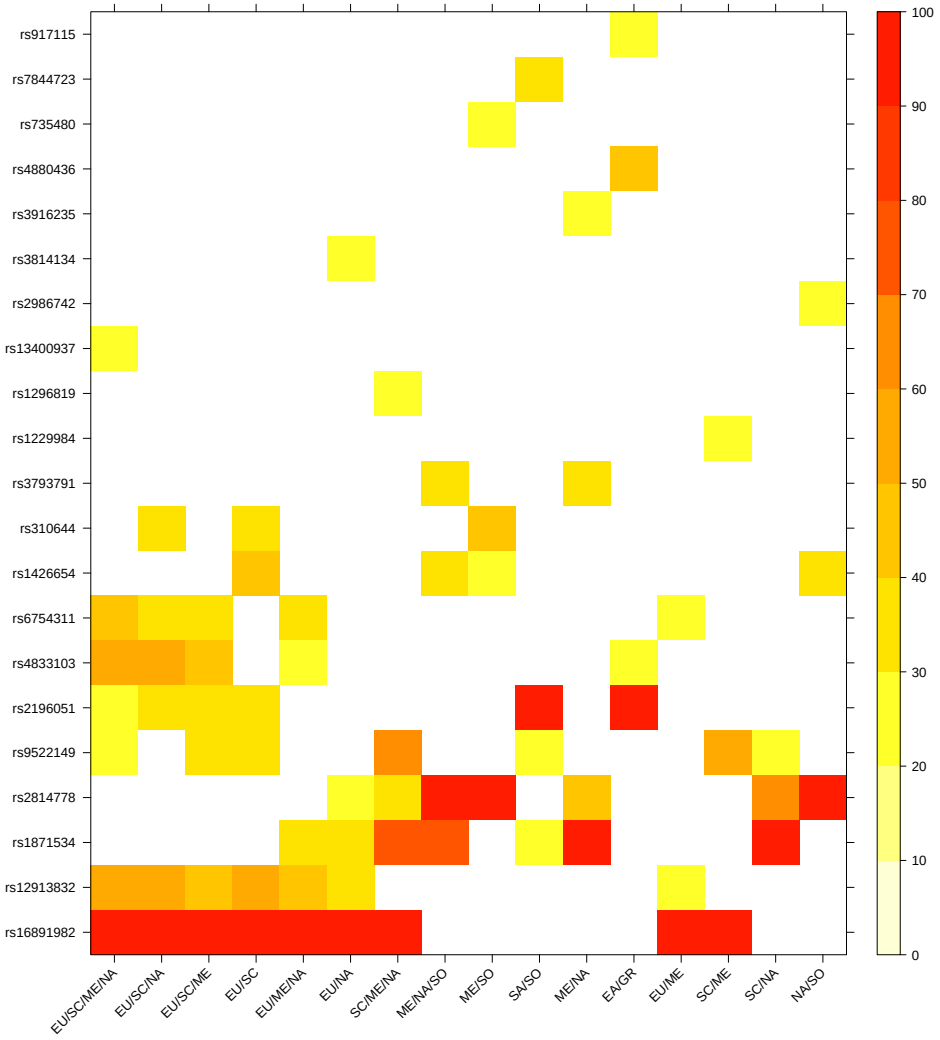
I forhold til hovedpopulationerne ses en sammenligning af CART, CART20 og CART50 i Figur 3.10.



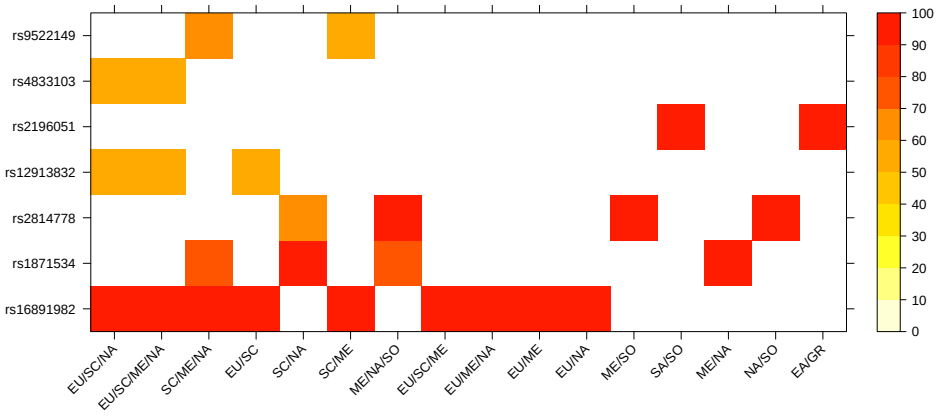
Figur 3.10: Procentandel på logaritmisk skala af klassificeringen inden for hver hovedpopulation for CART, CART20 og CART50.

I Figur 3.10 ses det, at M EAST stadig har en lille andel i gruppe A og en stor andel i gruppe B, uanset hvilken model der bliver taget i betragtning.

I CART20 og CART50 er der som sagt anvendt henholdsvis 21 og 7 markører. Figur 3.11 og Figur 3.12 viser de relative vægte for disse markører.



Figur 3.11: Relative vægte af markørerne fra CART20 for de forskellige delmængder af hovedpopulationer. Markørerne og delmængderne af hovedpopulationer er sorteret efter hyppighed. Farven hvid indikerer, at markøren ikke er anvendt i den pågældende delmængde. Navnet for delmængderne er forkortet til de to første bogstaver af hovedpopulationens forkortelse.



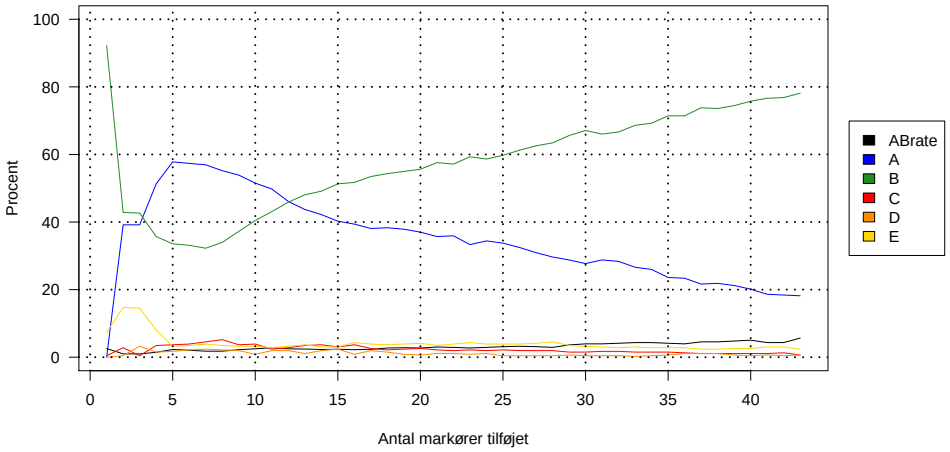
Figur 3.12: Relative vægte af markørerne fra CART50 for de forskellige delmængder af hovedpopulationer. Markørerne og delmængderne af hovedpopulationer er sorteret efter hyppighed. Farven hvid indikerer, at markøren ikke er anvendt i den pågældende delmængde. Navnet for delmængderne er forkortet til de to første bogstaver af hovedpopulationens forkortelse.

Det fremgår tydeligt i Figur 3.11 og Figur 3.12, at rs16891982 er den klart vigtigste markør, da den optræder i flest delmængder og har den største vægt for alle de delmængder, som den optræder i. Derudover er det også værd at bemærke rs1871534 og rs2814778 i Figur 3.12, som hver især optræder i fire delmængder med store vægte. Alle markørerne i Figur 3.12 er også de vigtigste markører i Figur 3.11 per konstruktionen af dem, hvor nogle af dem optræder i flere delmængder i Figur 3.11 med mindre vægte end 50.

3.1.5 Udvælgelse af markører manuelt

I et forsøg på at forbedre klassificeringen af grupperne er der lavet en manuel udvælgelse af markørerne, hvor der tages udgangspunkt i markørerne fra CART. Fra udfaldet af `genogeo` er der som sagt 21 delmængder med forskellige kombinationer af hovedpopulationerne, jævnfør Figur 3.4. For hver af delmængderne med flere end én hovedpopulation er der brugt CART for at finde den bedste delmængde af markører, som bedst skelner mellem hovedpopulationerne i den pågældende delmængde. Disse delmængder af markører er så blevet brugt i `genogeo` for de profiler, som blev accepteret af de pågældende hovedpopulationer. Dette er altså den samme metode til udvægelse af markørerne, som blev brugt i afsnit 3.1.2, men forskellen er nu, at `genogeo` kun får én markør af gangen til at finde en delmængde af markører. Tilføjelsen af markørerne sker i samme rækkefølge som deres vægt, så den vigtigste markør bliver tilføjet først.

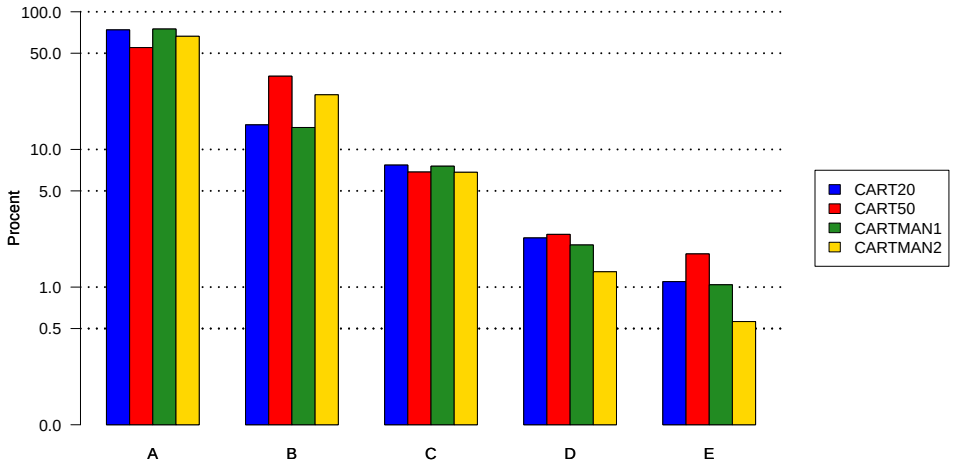
I Figur 3.13 vises et eksempel med delmængden EU/ME/NA, hvor procentandelen er angivet for de fem grupper efter tilføjelse af hver markør. Delmængden har 43 markører fra CART og er også den delmængde, som har den største procentandel af profiler, jævnfør Figur 3.4. En rate, som kaldes *ABrate*, er også blevet tilføjet, som angiver en rate for summen af andelen fra gruppe A og B divideret med summen af andelen fra gruppe C, D og E.



Figur 3.13: Procentandelen for de fem grupper samt ABrate efter tilføjelse af hver markør for EU/ME/NA.

Figur 3.13 viser, at delmængden med de fem vigtigste markører, som bedst skelner mellem EURO, M EAST og N AFRC, er den delmængde af markører, hvormed der er flest profiler, som bliver klassificeret i gruppe A. Derefter bliver procentandelen for gruppe A mindre og mindre, jo flere markører der bliver tilføjet. Procentandelen for gruppe B er tilnærmelsesvis en spejling af gruppe A. Dette stemmer godt overens med den støj, som blev nævnt i starten af afsnit 3.1.3, hvor `genogeo` accepterer mange hovedpopulationer, når der er mange markører med i analysen.

Der er valgt to modeller med det udgangspunkt at tilføje markørerne enkeltvist. Modellerne, som kaldes *CARTMAN1* og *CARTMAN2*, bruger henholdsvis den delmængde af markører, hvor procentandelen for gruppe A og ABrate. De to modeller er anvendt på de profiler, som blev accepteret af flere end én hovedpopulation fra `genogeo`. Resultatet af dette kan ses i Figur 3.14, hvor der bliver sammenlignet med CART20 og CART50.



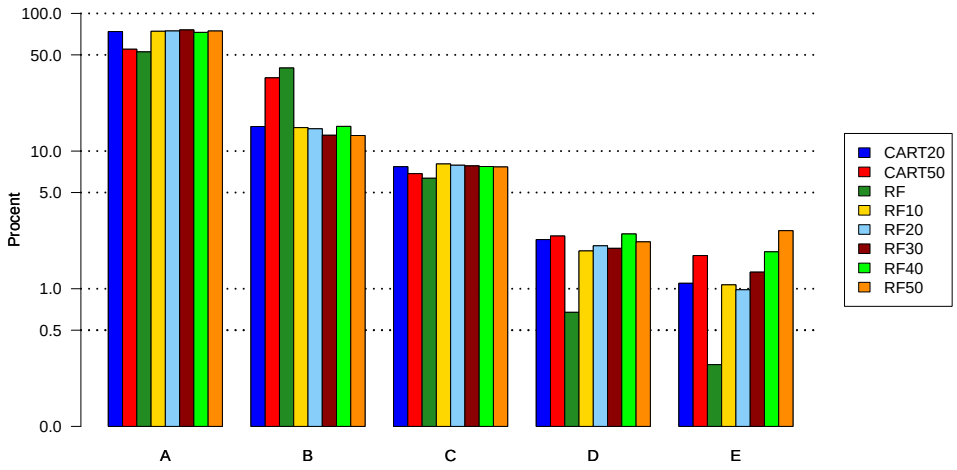
Figur 3.14: Procentandel på logaritmisk skala af klassificeringen for CART20, CART50, CARTMAN1 og CARTMAN2.

Det kan ses i Figur 3.14, at CARTMAN1 ligner meget CART20, men jævnfør Tabel 5.1 har CARTMAN1 brugt 79 markører i analysen til forskel fra CART20, som kun har brugt 21 markører. Med hensyn til CARTMAN2 ligger den lige midt imellem CART20 og CART50, men den har brugt 107 markører til analysen, så det vælges derfor at beholde CART20 og CART50 som værende de bedste modeller.

3.1.6 Udvælgelse af markører med Random Forest

En udvælgelse af markørerne er også foretaget med klassifikationsmetoden Random Forest, hvor fremgangsmåden er den samme som for CART. Random Forest giver også en delmængde af markørerne med vægte, som bedst skelner mellem de delmængder af hovedpopulationer, som blev vist i Figur 3.4. Metoden med at sætte grænser for vægtene er også anvendt i dette afsnit, hvor grænserne er de samme som i afsnit 3.1.4.

I Figur 3.15 ses en sammenligning af CART20, CART50 og de seks modeller med Random Forest, som kaldes *RF*, *RF10*, *RF20*, *RF30*, *RF40* og *RF50*.



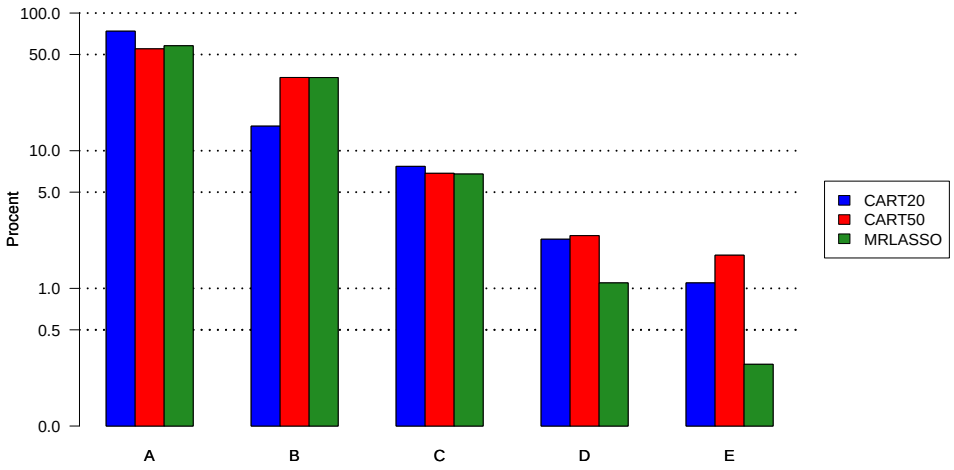
Figur 3.15: Procentandel på logaritmisk skala af klassificeringen for CART20, CART50, RF, RF10, RF20, RF30, RF40 og RF50.

Det fremgår umiddelbart af Figur 3.15, at RF kunne være en god erstatning for CART50, men jævnfør Tabel 5.1 anvender RF alle 164 markører. Desuden ligner de øvrige modeller med Random Forest meget CART20, men jævnfør Tabel 5.1 igen anvender alle modellerne med Random Forest flere markører end de 21 markører, som CART20 anvender. Det vælges derfor igen at gå videre med CART20 og CART50.

Analysen med hovedpopulationerne er igen lavet med anvendelse af Random Forest, men grundet den manglende forbedring er den udeladt her.

3.1.7 Udvælgelse af markører med Lasso

Et sidste forsøg på at finde en delmængde af markørerne er foretaget med en multinomial regression, Lasso, hvor fremgangsmåden igen er den samme som for CART. Figur 3.16 viser en sammenligning af CART20, CART50 og analysen med multinomial regression, som kaldes MRLASSO.



Figur 3.16: Procentandel på logaritmisk skala af klassificeringen for CART20, CART50 og MRLASSO.

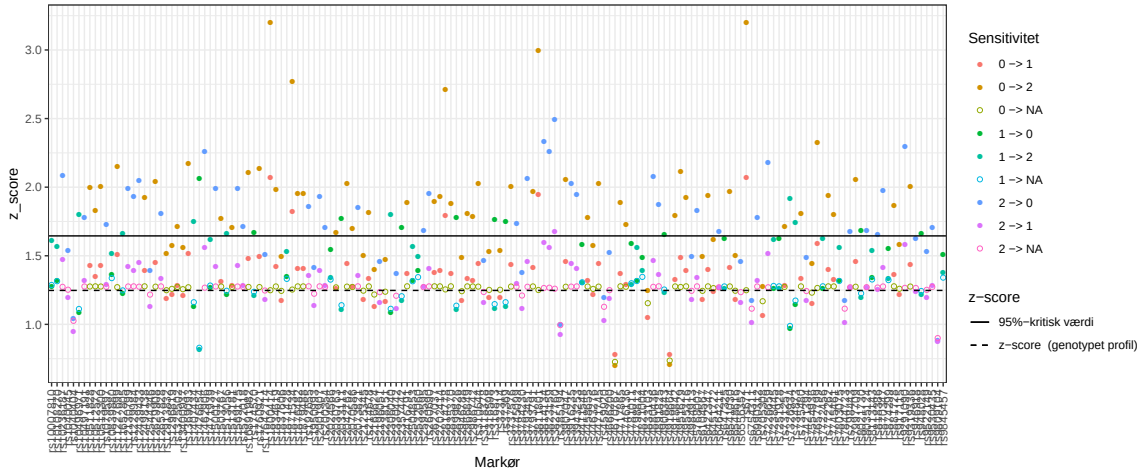
Figur 3.16 viser umiddelbart også, at analysen med MRLASSO kunne være en god erstatning for CART50, men jævnfør Tabel 5.1 anvender MRLASSO 152 markører, så CART20 og CART50 anses fortsat som værende de bedste modeller.

Analysen med hovedpopulationerne er igen lavet med MRLASSO, men grundet den manglende forbedring er den udeladt her.

En samlet oversigt over klassificeringen for alle modellerne kan ses i Figur 5.1, og en oversigt over modellernes anvendelse af markørerne kan ses i Figur 5.2.

3.2 Sensitivitet af markører

Genotypen på en specifik markør for en profil kan omsættes til digital information ved at se på frekvensen af allel 1. For eksempel kan en profil have genotypen AA, AG eller GG på en markør, som vil blive omsat til henholdsvis 2, 1 eller 0. I visse tilfælde kan der opstå fejl ved klassificeringen af profilens genotype på markørerne, hvilket medfører afvigelser fra den korrekte z-score. Figur 3.17 viser en simuleret profil fra Europa, hvor der enkeltvist er blevet ændret på frekvensen af allel 1 for alle markører.



Figur 3.17: En simuleret profil fra Europa med enkeltvise ændringer på frekvensen af allel 1 for alle markører. En ændring til en NA-værdi betyder, at den pågældende markør er udeladt. Den sorte linje angiver 95%-kvartilen af standard normalfordelingen (1,64), og den stiplede linje angiver profilens z-score for at stamme fra Europa.

Det fremgår af Figur 3.17, at ændringer på den simulerede profil fra Europa i værste tilfælde kan medføre, at profilen bliver forkastet for at stamme fra Europa. Det fremgår også, at ændringerne $0 \to 2$ og $2 \to 0$ ofte resulterer i en højere z-score end 1,64, så profilen bliver forkastet for at stamme fra Europa. Dette stemmer godt overens med Figur 2.1, da illustration 1 og 3 i Figur 2.1 er hinandens spejling omkring 0,5. Det vil sige, at ændringerne $0 \to 2$ og $2 \to 0$ ofte resulterer i en stor differens på z-scoren og en højere z-score end 1,64. Ændringerne $0 \to 1$, $1 \to 0$, $1 \to 2$ og $2 \to 1$ i Figur 3.17 resulterer ofte i en lavere z-score end 1,64, hvilket også stemmer godt overens med Figur 2.1, da z-scoren ofte ikke ændrer sig betydeligt for disse ændringer og forbliver under en z-score på 1,64. Ændringerne til en NA-værdi i Figur 3.17 ligger ofte tæt på den korrekte z-score, hvilket stemmer godt overens med (2.6), da markørernes bidrag i ligningen blot bliver udeladt.

I dette afsnit vil det derfor blive undersøgt, hvilke markører der er særlige sensitive i forhold til at acceptere eller forkaste en hovedpopulation.

3.2.1 Udvælgelse af særlige sensitive markører

Der bliver først undersøgt de tilfælde, hvor en profil er blevet accepteret af sin egen hovedpopulation. I analysen er der simuleret 1.000 profiler inden for hver hovedpopulation. Profilerne har fået ændret én indgang i deres $\mathbf{x}_0 = \begin{bmatrix} x_{01} & x_{02} & \dots & x_{0I} \end{bmatrix}$ jævnfør Tabel 3.3, hvorefter en ny z-score er blevet udregnet for profilen i forhold til hovedpopulationen.

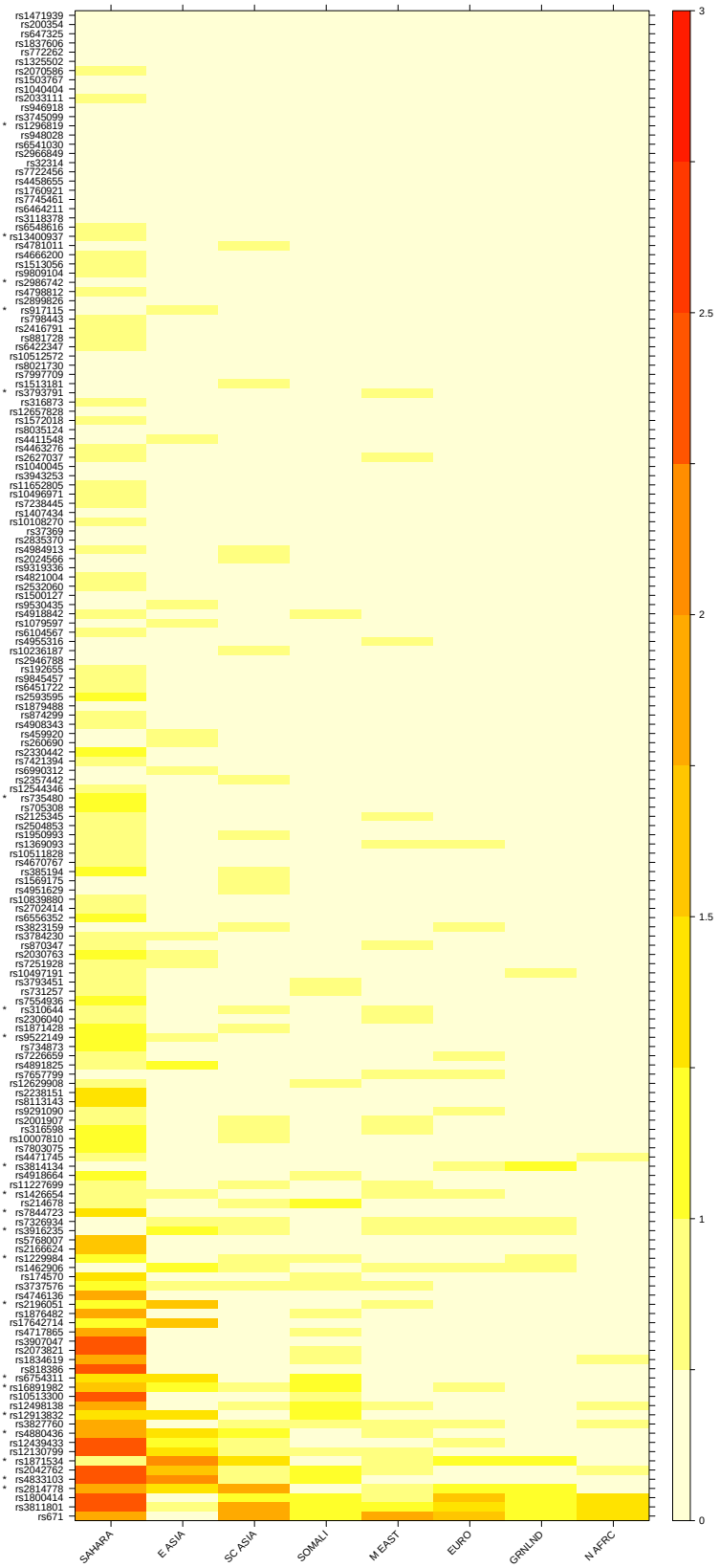
Oprindelig x_{0i}	Ny x_{0i}
0	1 eller NA
1	0, 2 eller NA
2	1 eller NA

Tabel 3.3: Mulige ændringer af den oprindelige x_{0i}.

Det vil sige, at hvis profilen for eksempel har $x_{0i} = 0$ for markør $i = 1, 2, \dots, I$, bliver denne ene markør ændret til $x_{0i} = 1$ eller $x_{0i} = NA$ samtidig med, at profilen beholder informationen fra alle de andre markører. En ændring til en NA-værdi betyder, at den pågældende markør er udeladt.

Situationerne, hvor $x_{0i} = 0$ ændres til $x_{0i} = 2$ eller $x_{0i} = 2$ ændres til $x_{0i} = 0$, undersøges ikke, da det er meget usandsynligt, at en markør for en profil ændres fra 0 til 2 eller omvendt. For eksempel ville det for ændringen $2 \rightarrow 0$ med A som allel 1 betyde, at kun få A-reads skulle genotypes, mens mange G-reads skulle genotypes og dermed medføre, at $x_0 = 0$. Disse ændringer er yderst usandsynlige, hvorimod ændringerne i Tabel 3.3 er mere realistiske.

Hvis ændringen på en markør har givet en z-score over 95%-kvartilen af standard normalfordelingen (1,64), bliver denne markør betragtet som værende sensitiv. I Figur 3.18 vises procentandelen for hyppigheden af de sensitive markører.

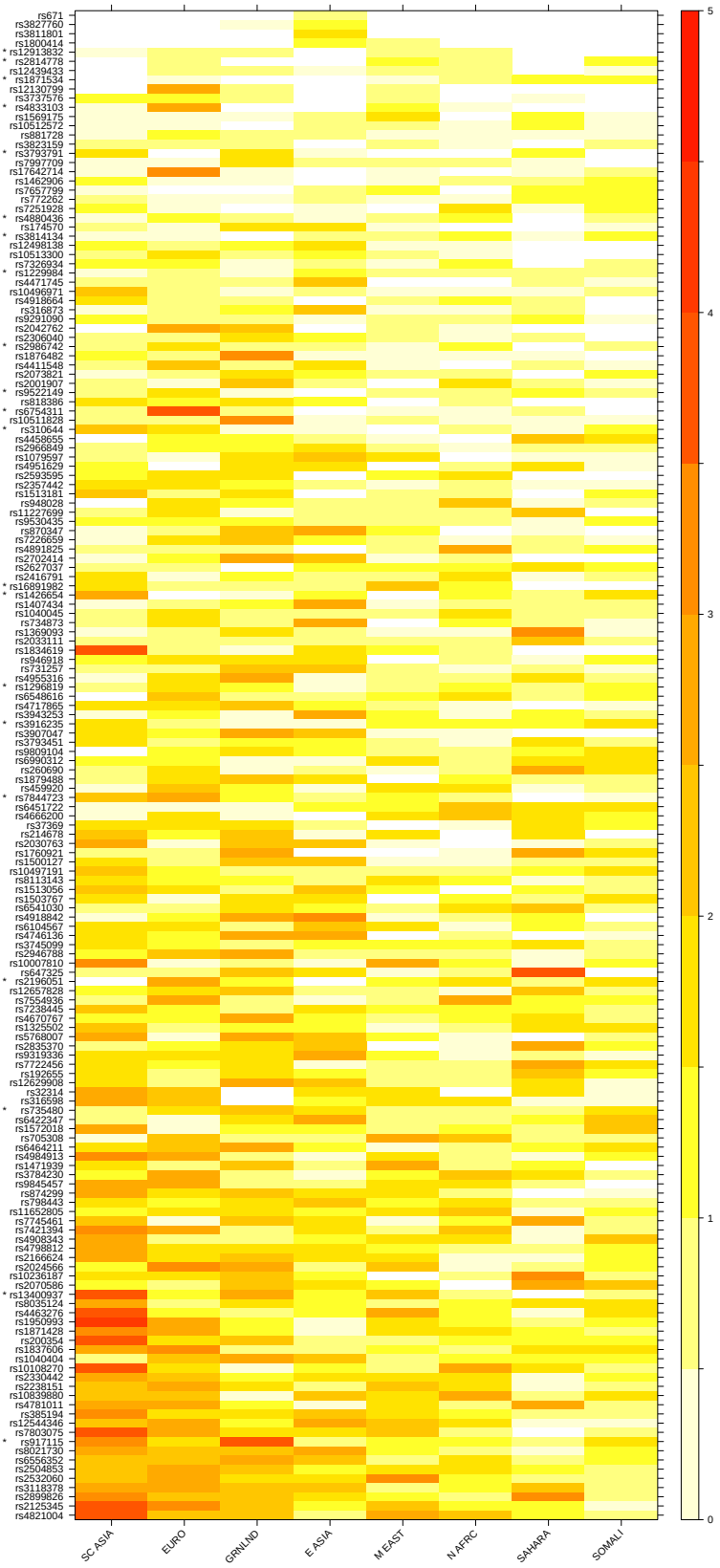


Figur 3.18: Procentandelen for hyppigheden af de sensitive markører for simulerede profiler, som er accepteret af deres egen hovedpopulation, men er blevet forkastet efter en ændring på den pågældende markør. Markørerne og hovedpopulationerne er sorteret efter hyppighed. Markørerne, som er markeret med en stjerne, blev anvendt i CART20.

Det fremgår af Figur 3.18, at alle markørerne er sensitive, men der er nogle få af de første markører fra bunden, som er mere sensitive end andre. Dette gælder specielt rs671, rs3811801 og rs1800414, da de har de største hyppigheder for at være sensitive for flere hovedpopulationer. Det fremgår også tydeligt af Figur 3.18, at der er nogle hovedpopulationer, som er mere sensitive end andre. Her skiller SAHARA sig mest ud ved at have ti markører, som har en procentandel på over 2%. Det vil altså sige, at SAHARA er mest tilbøjelig til at forkaste profiler efter en ændring på en markør, selvom profilerne i første omgang var accepteret af SAHARA. Det bemærkes også, at de fleste af de markører, som blev anvendt i CART20, ligger nede blandt de markører, som har de største hyppigheder til at være sensitive.

Der vil nu blive undersøgt tilfældene, hvor en profil er blevet forkastet af sin egen hovedpopulation. Denne analyse følger samme fremgangsmåde som den tidligere analyse, men en markør bliver nu betragtet som værende sensitiv, hvis ændringen på markøren giver en z-score under 95%-kvartilen af standard normalfordelingen (1,64). Resultatet af dette er vist i Figur 3.19.

Det kan ses i Figur 3.19, at nogle markører er mere sensitive end andre, hvor de mest sensitive markører er rs4821004, rs2125345 og rs2899826. Der er også tilfælde, hvor procentandelen for nogle markører og hovedpopulationer er lig nul. Dette er specielt gældende for rs671, rs3827760, rs3811801 og rs1800414, da de har en procentandel på nul for 6-7 hovedpopulationer. Det bemærkes også, at markørerne, som blev anvendt i CART20, ikke er koncentreret nede blandt de markører med de højeste procentandele på samme måde som i Figur 3.18, men de er derimod spredt mere ud blandt alle de andre markører.



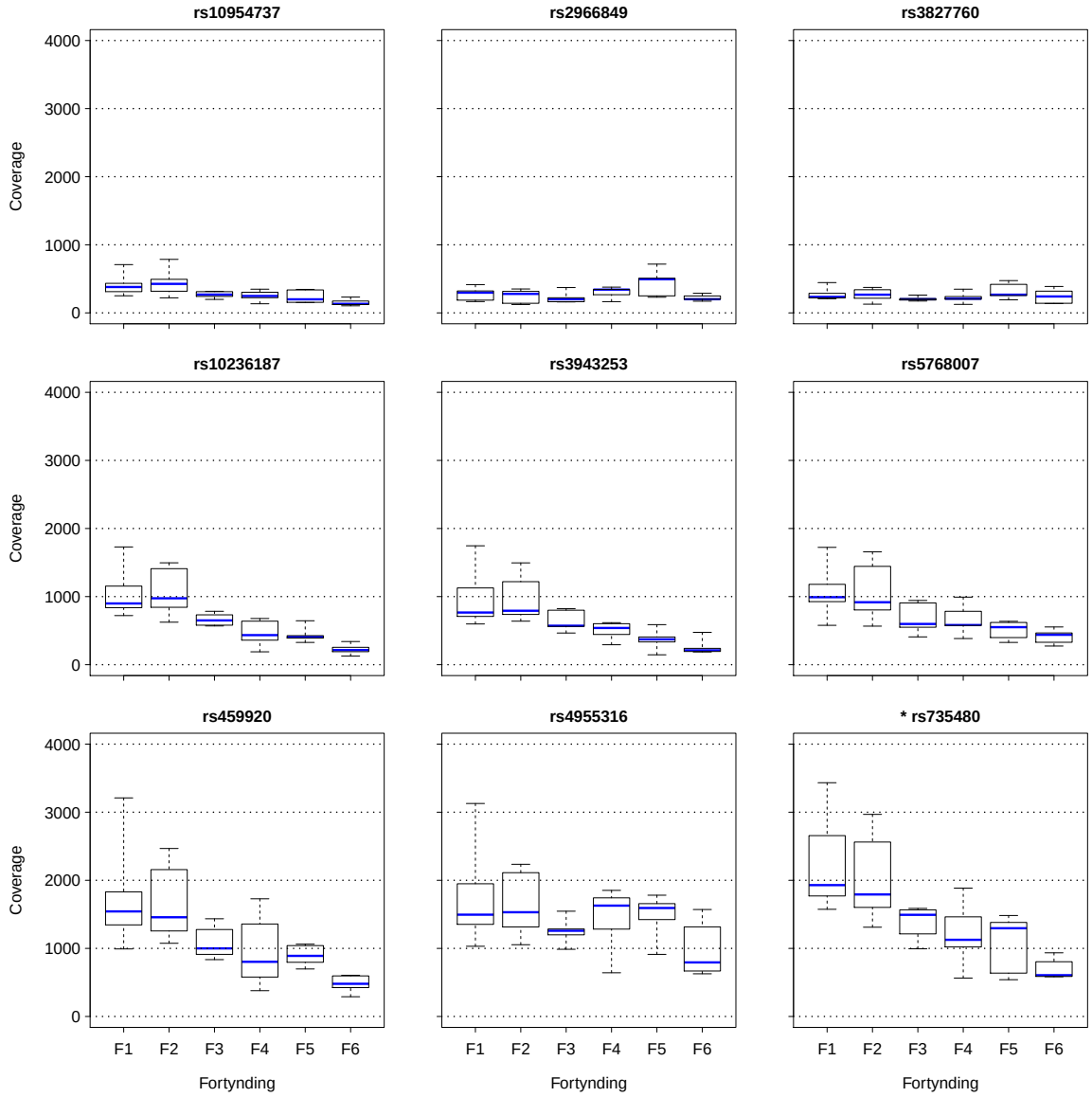
Figur 3.19: Procentandelen for hyppigheden af de sensitive markører for simulerede profiler, som er forkastet af deres egen hovedpopulation, men er blevet accepteret efter en ændring på den pågældende markør. Markørerne og hovedpopulationerne er sorteret efter hyppighed. Markørerne, som er markeret med en stjerne, blev anvendt i CART20.

3.3 Coverage for markører

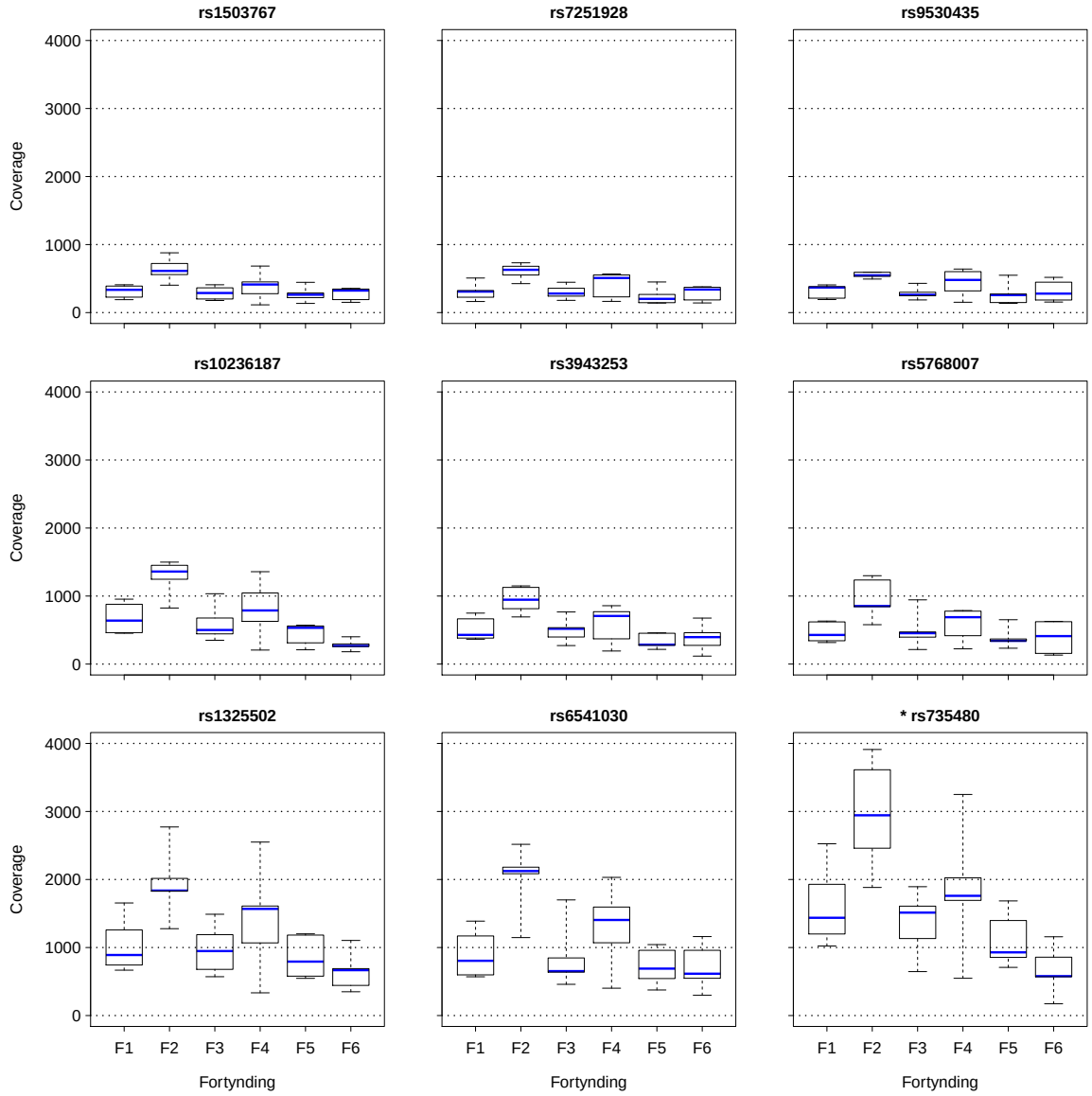
Genotypen på markørerne er baseret på markørernes coverage, altså hvor mange gange der er blevet læst de forskellige baser på markørerne. I fortyndingsdatasættet er der lavet et forsøg med fortynding af markørernes coverage. En fortynding af coverage kan medføre, at troværdigheden for genotypen af baserne på markørerne mindskes. I værste tilfælde kan en markør give et dropout i analysen grundet nogle betingelser, som bliver beskrevet i afsnit 3.3.6. Derudover er der i fortyndingsdatasættet for hver observation også angivet positiv og negativ coverage samt procentandelen for det dominerende allel. Hvis markørernes coverage opfører sig usædvanlig i disse tilfælde, vil det også mindske troværdigheden for dem. Dette afsnit vil derfor omhandle en analyse af markørernes coverage og tilbøjelighed til at give et dropout.

3.3.1 Coverage for udvalgte markører

I fortyndingsdatasættet er der angivet coverage for samtlige markører. Coverage for udvalgte markører er angivet som boksplots i Figur 3.20 og Figur 3.21 for henholdsvis duplikat A og B. De tre øverste, midterste og nederste illustrationer i figurene repræsenterer markører med henholdsvis lav, middel og høj coverage.



Figur 3.20: Coverage for udvalgte markører for duplikat A. Markørerne, som er markeret med en stjerne, blev anvendt i CART20.

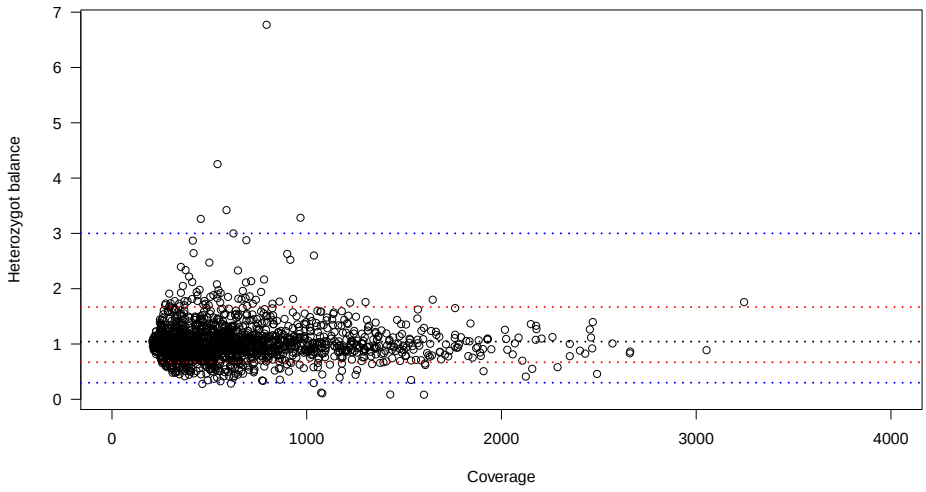


Figur 3.21: Coverage for udvalgte markører for duplikat B. Markørerne, som er markeret med en stjerne, blev anvendt i CART20.

Det ses i Figur 3.20 og Figur 3.21, at der kan være markører med lav, middel og høj coverage, som tydeligt er forskellige fra hinanden. Tendensen for de fleste illustrationer er, at de første fortyndinger har relativ høj coverage, hvorimod coverage bliver lavere i de sidste fortyndinger. Bemærk, at enkelte markører går igen i de to figurer, for eksempel rs735480 repræsenterer markørerne med høj coverage og er tilmed også den eneste markør i Figur 3.20 og Figur 3.21, som blev anvendt i CART20.

3.3.2 Heterozygote alleller

I fortyndingsdatasættet er der 1.539 observationer, hvor der er set heterozygote alleller på en profil. Forholdet imellem reads for de to læste baser på allellerne kaldes *heterozygot balance*. For at to alleller bliver betragtet som heterozygote, skal den heterozygote balance for dem være inden for et interval, som ofte bliver sat til $[0,3;3]$. Det vil sige, at forholdet imellem reads for baserne skal ligge inden for dette interval. En skærpet heterozygot balance på $[0,67;1,67]$ bruges også, hvor styrken for dét at være heterozygot er stærkere. Coverage og den heterozygote balance for de heterozygote alleller er angivet i Figur 3.22.

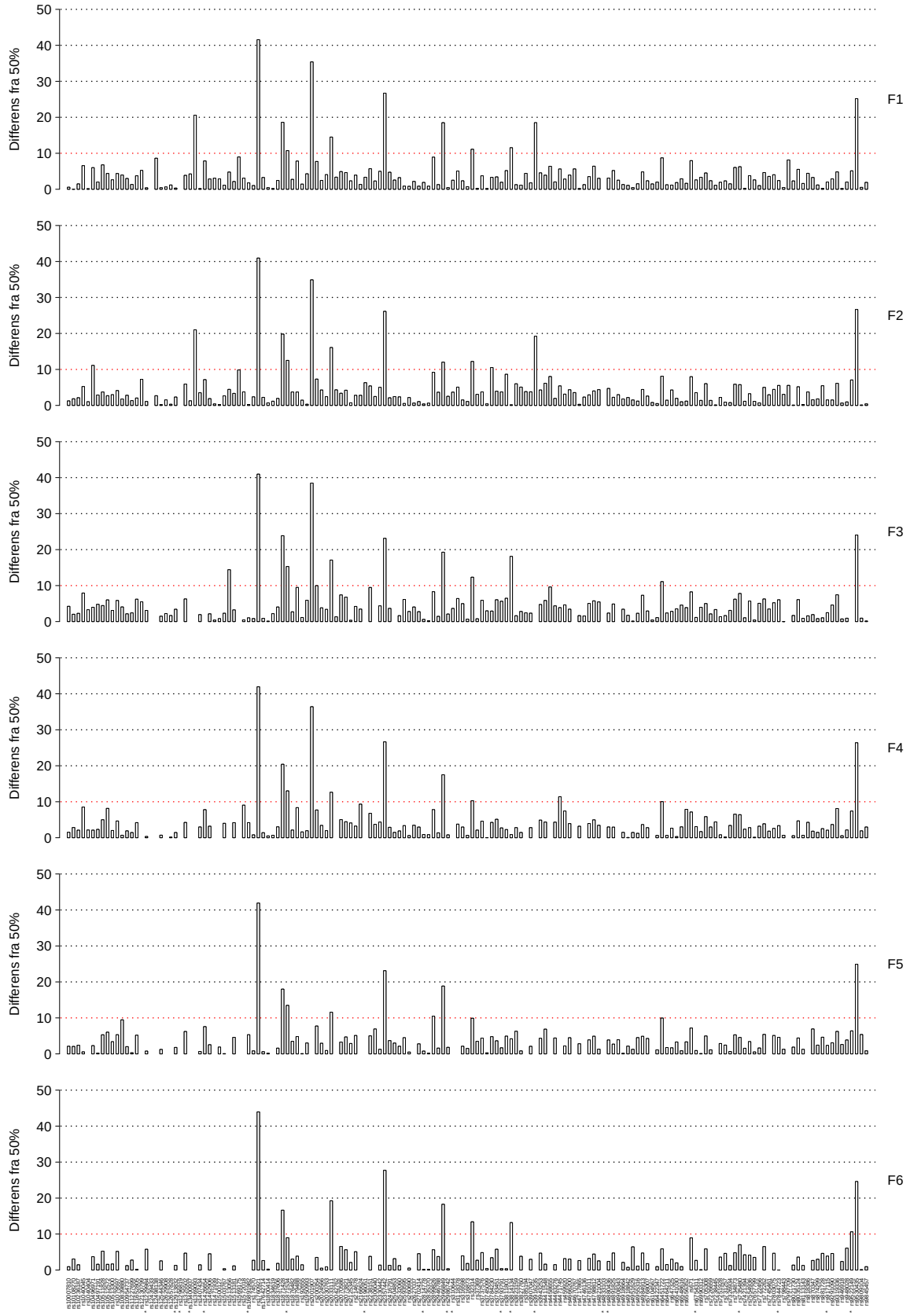


Figur 3.22: Coverage og den heterozygote balance for de heterozygote alleller. Den sorte stiplede linje angiver gennemsnittet for heterozygot balancen for observationerne, og de blå og røde stiplede linjer angiver henholdsvis de to heterozygot balancer på $[0,3;3]$ og $[0,67;1,67]$.

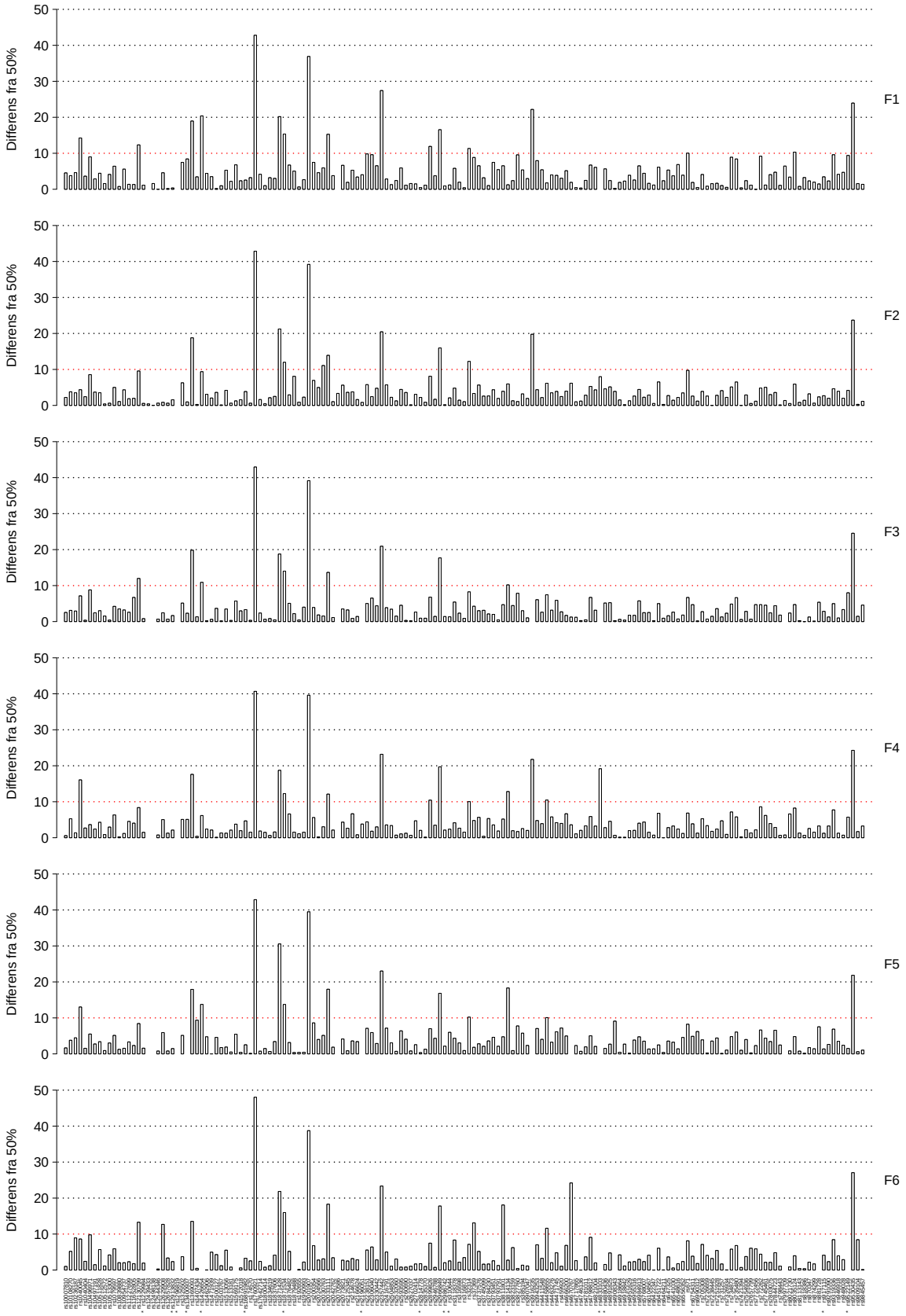
Det ses i Figur 3.22, at der er enkelte observationer, som ligger uden for heterozygot balancen på $[0,3;3]$, mens der er betydelig flere observationer, som ligger uden for balancen på $[0,67;1,67]$. Gennemsnittet af den heterozygote balance for alle observationerne er 1,04.

3.3.3 Positiv og negativ coverage

I fortyndingsdatasættet er der også angivet positiv og negativ coverage. For en højere troværdighed af markørerne bør procentandelen for positiv coverage ligge på omkring 50%, så positiv og negativ coverage er omtrent lige store. I Figur 3.23 og Figur 3.24 vises gennemsnittet af markørernes differens fra 50% for positiv coverage for henholdsvis duplikat A og B efter en filtrering med en allelgrænse på 100 og en heterozygot balance på $[0,3;3]$.



Figur 3.23: Gennemsnit af markørernes differens fra 50% for positiv coverage for hver fortynding fra duplikat A. Den røde stiplede linje angiver en differens på 10. Markørerne, som er markeret med en stjerne, blev anvendt i CART20.



Figur 3.24: Gennemsnit af markørernes differens fra 50% for positiv coverage for hver fortynding fra duplikat B. Den røde stiplede linje angiver en differens på 10. Markørerne, som er markeret med en stjerne, blev anvendt i CART20.

Det ses i Figur 3.23 og Figur 3.24, at gennemsnittet for de fleste af markørernes differens fra 50% ligger under 10. I Tabel 3.4 er der angivet de markører, som overskrider differensen på 10 med gennemsnittet af deres differens fra 50% samt antallet af gange, hvor markøren overskrider differensen på 10.

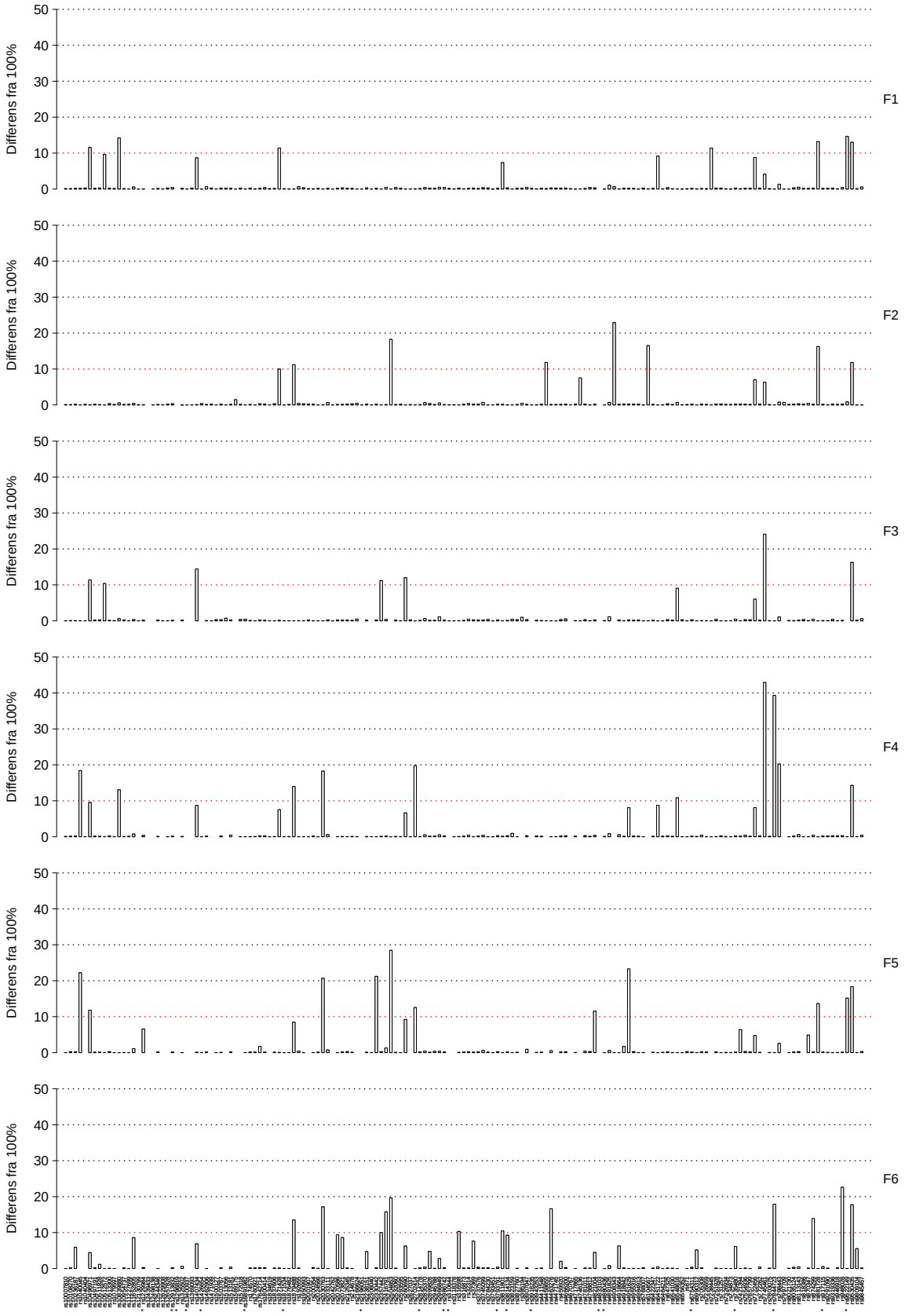
	Markør	Gennemsnit for differens	Antal overskridelser
	rs1760921	42,62	12
	rs2001907	37,82	10
	rs9530435	24,76	12
	rs2357442	24,31	12
	rs1871428	20,72	12
*	rs3916235	20,28	5
	rs1369093	18,53	8
	rs2966849	17,41	12
	rs2033111	15,19	12
*	rs4833103	13,56	1
*	rs1871534	13,10	11
	rs32314	10,70	9
*	rs1426654	10,10	3
	rs12130799	9,10	3
	rs2899826	8,50	3
	rs1040045	8,38	3
*	rs3814134	8,27	6
	rs671	8,09	1
	rs4458655	7,95	3
	rs6451722	7,23	2
	rs4670767	6,61	1
*	rs9522149	6,54	1
	rs10496971	6,21	1
	rs3811801	5,91	1
	rs8035124	5,62	1
	rs4471745	5,50	1
	rs1513056	5,03	1
	rs37369	4,04	1
	rs12629908	3,99	1
	rs3784230	3,79	1
	rs2030763	3,62	1

Tabel 3.4: Markører, som overskrider differensen på 10 for positiv coverage med gennemsnittet af deres differens fra 50% samt antallet af gange, hvor markøren overskrider differensen på 10. Markørerne, som er markeret med en stjerne, blev anvendt i CART20.

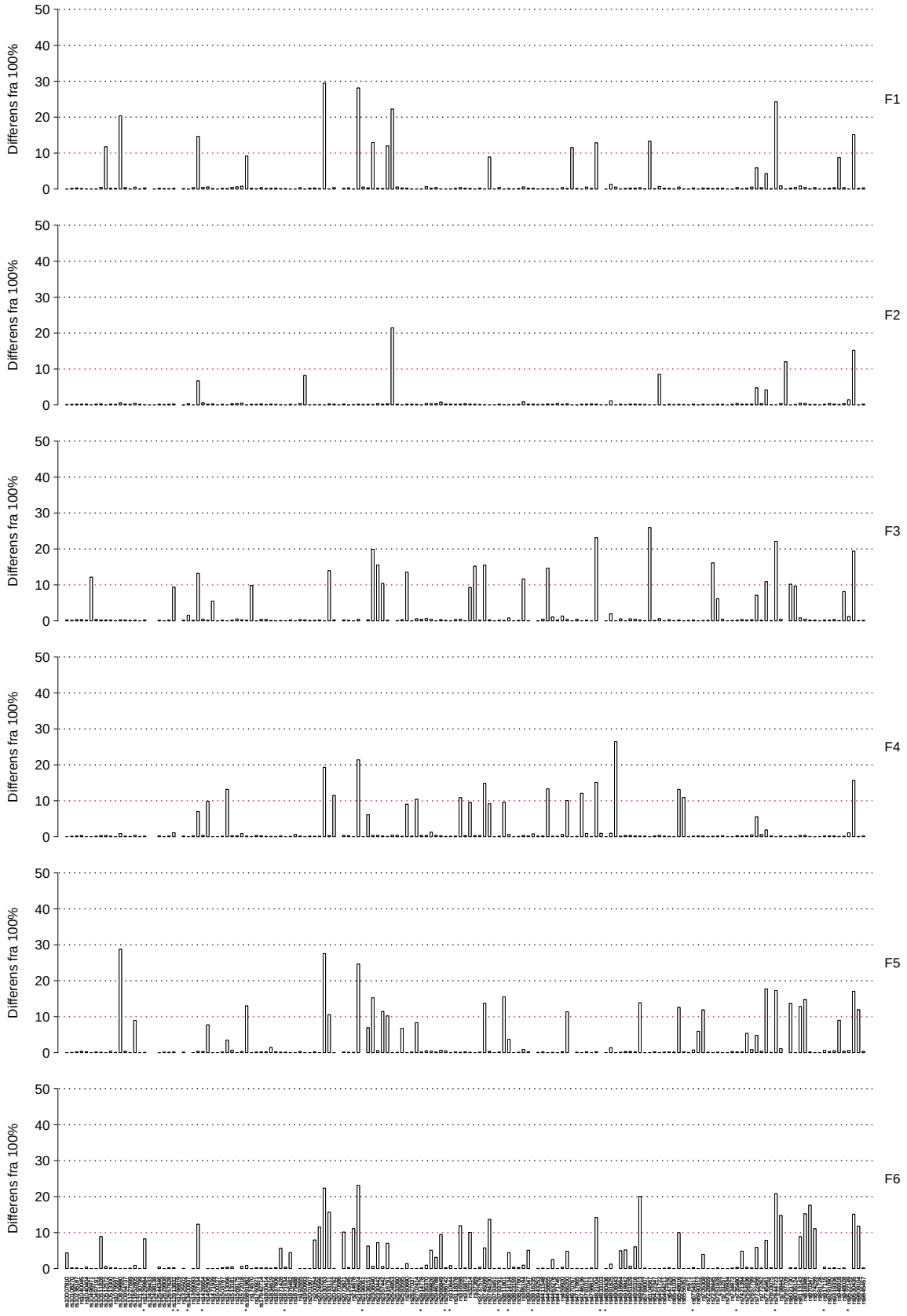
Som det fremgår i Tabel 3.4, er der 31 markører, som overskrider differensen på 10. Markørerne rs1760921 og 20011907 er dem med de største differenser, da de har et gennemsnit for deres differens på over 30. Ti af markørerne overskrider mere end halvdelen af gangene, hvor seks af disse markører overskrider i alle fortyndinger. Desuden er der også nogle markører fra CART20, som er med i tabellen, hvor rs3916235 og rs1871534 skiller sig mest ud. Markøren rs1871534 overskrider 11 gange, men den har dog kun et gennemsnit på 13,10, hvorimod rs3916235, som har et gennemsnit på 20,28, overskrider fem gange.

3.3.4 Det dominerende allel

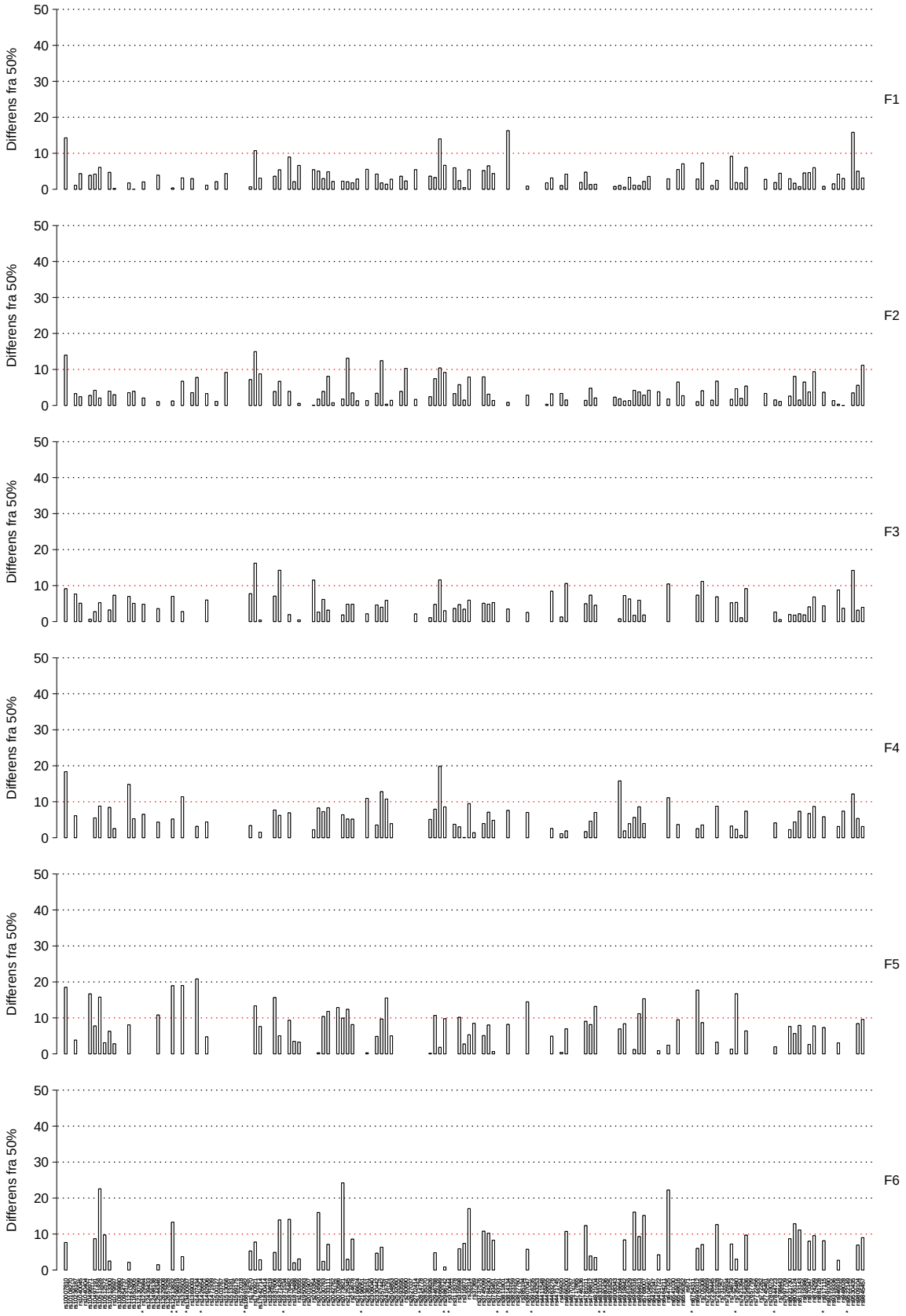
I fortyndingsdatasættet er der også oplyst procentandelen for det dominerende allel på profilernes markører. For en højere troværdighed af markørerne bør procentandelen for det dominerende allel for homozygote og heterozygote alleller ligge på henholdsvis 100% og 50%. I Figur 3.25 og Figur 3.26 ses gennemsnittet af homozygote allellers differens fra 100% på markørerne for duplikat A og B, hvorimod gennemsnittet af heterozygote allellers differens fra 50% på markørerne for duplikat A og B ses i Figur 3.27 og Figur 3.28. Analysen er lavet med en allelgrænse på 100 og en heterozygot balance på $[0,3;3]$.



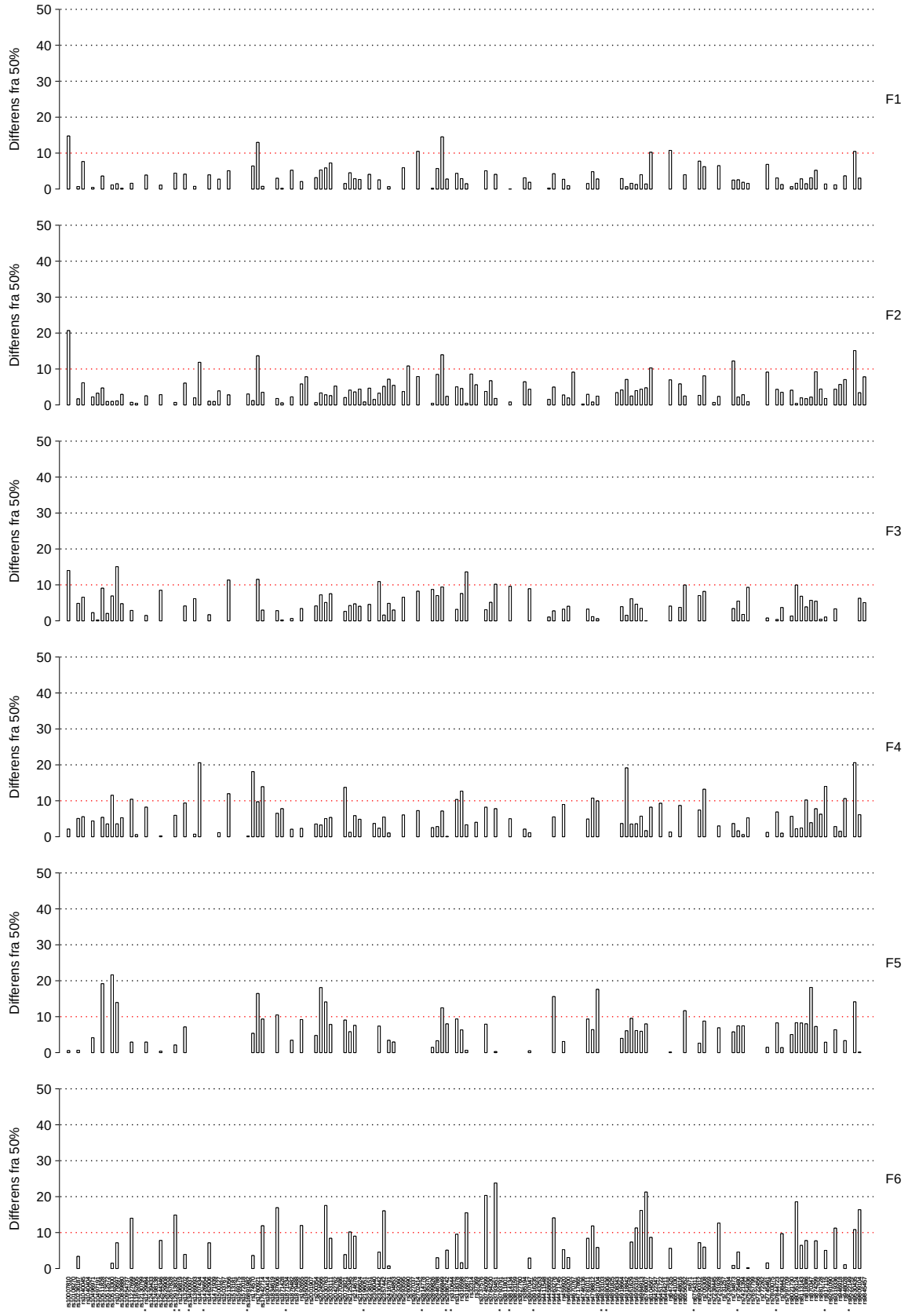
Figur 3.25: Gennemsnit af homozygote allellers differens fra 100% for det dominerende allel på markørerne for hver fortynding fra duplikat A. Den røde stiplede linje angiver en differens på 10. Markørerne, som er markeret med en stjerne, blev anvendt i CART20.



Figur 3.26: Gennemsnit af homozygote allellers differens fra 100% for det dominerende allel på markørerne for hver fortynding fra duplikat B. Den røde stiplede linje angiver en differens på 10. Markørerne, som er markeret med en stjerne, blev anvendt i CART20.



Figur 3.27: Gennemsnit af heterozygote allellers differens fra 50% for det dominerende allel på markørerne for hver fortynding fra duplikat A. Den røde stiplede linje angiver en differens på 10. Markørerne, som er markeret med en stjerne, blev anvendt i CART20.



Figur 3.28: Gennemsnit af heterozygote allellers differens fra 50% for det dominerende allel på markørerne for hver fortynding fra duplikat B. Den røde stiplede linje angiver en differens på 10. Markørerne, som er markeret med en stjerne, blev anvendt i CART20.

Det ses i Figur 3.25, Figur 3.26, Figur 3.27 og Figur 3.28, at differensen fra 100% og 50% for de fleste markører ligger under 10. Der er kun en enkelt gang, hvor to markører kommer over 30, nærmere bestemt rs7745461 og rs7844723 i Figur 3.25. I Tabel 3.5 er der angivet de markører, som overskrider differensen på 10 mere end én gang med gennemsnittet af deres allellers differens fra 100%/50% og antallet af gange, hvor markørerne overskrider differensen på 10 mere end én gang.

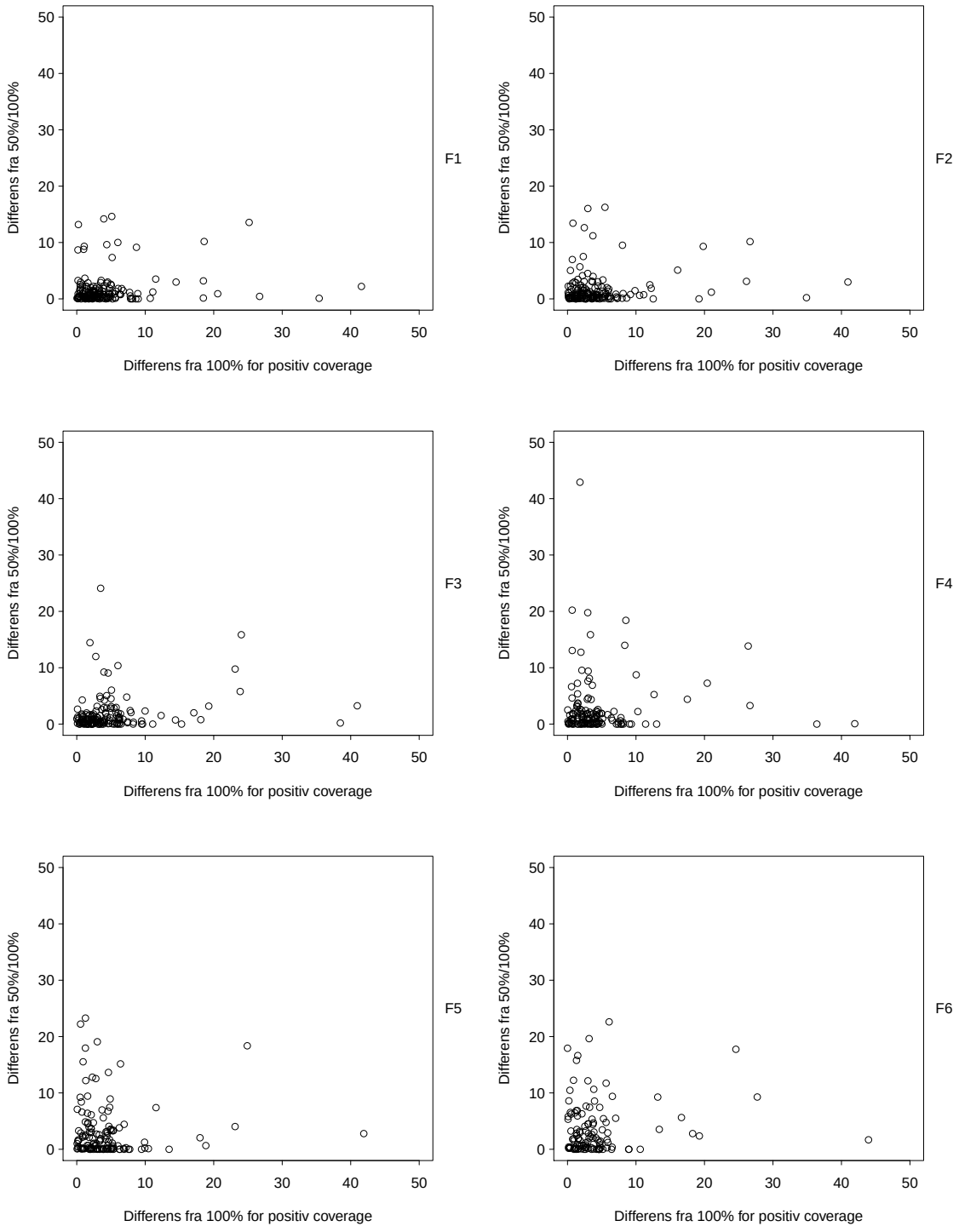
		Homozygot		Heterozygot		Samlet	
	Markør	Gennemsnit	Overskridelser	Gennemsnit	Overskridelser	Gennemsnit	Overskridelser
	rs9530435	15,75	12	12,97	8	14,36	20
	rs1760921			12,73	8	12,73	8
	rs7745461	12,42	4			12,42	4
	rs10007810			12,18	7	12,18	7
*	rs7844723	11,81	6			11,81	6
	rs2966849			11,51	7	11,51	7
	rs2504853	11,05	5			11,05	5
	rs1407434	7,76	4	12,83	3	10,30	7
	rs2030763	12,89	7	6,95	3	9,92	10
	rs2166624	9,88	4			9,88	4
	rs10511828			9,31	3	9,31	3
	rs10839880	7,20	4			7,20	4
	rs1837606			7,02	3	7,02	3
	rs6104567	6,20	3	7,53	2	6,86	5
	rs1325502			6,79	2	6,79	2
*	rs12913832			6,74	3	6,74	3
	rs647325			6,64	4	6,64	4
	rs7251928			6,55	2	6,55	2
	rs5768007			6,53	3	6,53	3
	rs4821004	6,91	5	5,91	2	6,41	7
	rs260690	4,88	2	7,80	2	6,34	4
	rs4908343	6,32	2			6,32	2
	rs8035124			6,28	2	6,28	2
	rs3793451			6,05	2	6,05	2
	rs10513300			6,05	2	6,05	2
	rs2125345			5,87	3	5,87	3
	rs11227699			5,82	3	5,82	3
	rs3745099	4,36	3	7,20	2	5,78	5
	rs2357442	3,70	4	7,53	3	5,61	7
	rs32314	2,59	1	8,52	1	5,56	2
	rs17642714			5,55	2	5,55	2
	rs4798812			5,48	2	5,48	2
	rs1079597			5,28	2	5,28	2
	rs2702414	4,32	3	6,15	1	5,23	4
	rs2033111	3,60	3	6,86	1	5,23	4
	rs4955316			5,07	2	5,07	2
	rs2306040	4,92	3			4,92	3
	rs4984913	2,95	2	6,61	2	4,78	4
	rs6548616	4,77	3			4,77	3
	rs1513056	2,01	1	7,45	2	4,73	3
	rs3784230	2,78	1	6,45	1	4,62	2
	rs10496971	5,07	4	4,15	1	4,61	5
	rs705308	1,48	1	7,68	2	4,58	3
	rs1871428	2,95	1	6,02	2	4,48	3
	rs4458655	4,48	3			4,48	3
	rs2416791	4,00	3	4,70	2	4,35	5
	rs1879488	4,34	3			4,34	3
	rs2330442	3,85	2	4,68	1	4,27	3
	rs316598	2,88	3	5,63	2	4,26	5
	rs316873			4,21	2	4,21	2
	rs818386	3,24	2	5,10	1	4,17	3
	rs2073821	1,74	1	6,60	2	4,17	3
	rs9809104	2,49	2	5,81	1	4,15	3
	rs4463276	1,79	1	6,31	2	4,05	3
	rs798443	3,98	2			3,98	2
	rs881728	3,92	3			3,92	3
	rs870347	1,99	1	5,71	1	3,85	2
	rs1040045	3,83	2			3,83	2
	rs2024566	1,04	1	6,46	2	3,75	3
	rs6556352	1,06	1	6,30	1	3,68	2
	rs3811801	3,61	2			3,61	2
	rs948028	2,78	1	4,41	1	3,60	2
	rs8113143	2,21	1	4,95	1	3,58	2
*	rs9522149	3,49	2			3,49	2
	rs4666200	2,40	2	4,58	2	3,49	4
	rs7238445	2,79	2			2,79	2
	rs10512572	2,74	2			2,74	2
	rs874299	2,25	2			2,25	2
	rs8021730	2,06	2			2,06	2

Tabel 3.5: Markører, som overskrider differensen på 10 mere end én gang for det dominerende allel med gennemsnittet af deres allellers differens fra 100%/50%. Derudover antallet af gange, hvor markørerne overskrider differensen på 10. Til sidst er der en samlet oversigt, som indeholder informationen fra homozygot og heterozygot. Markørerne, som er markeret med en stjerne, blev anvendt i CART20.

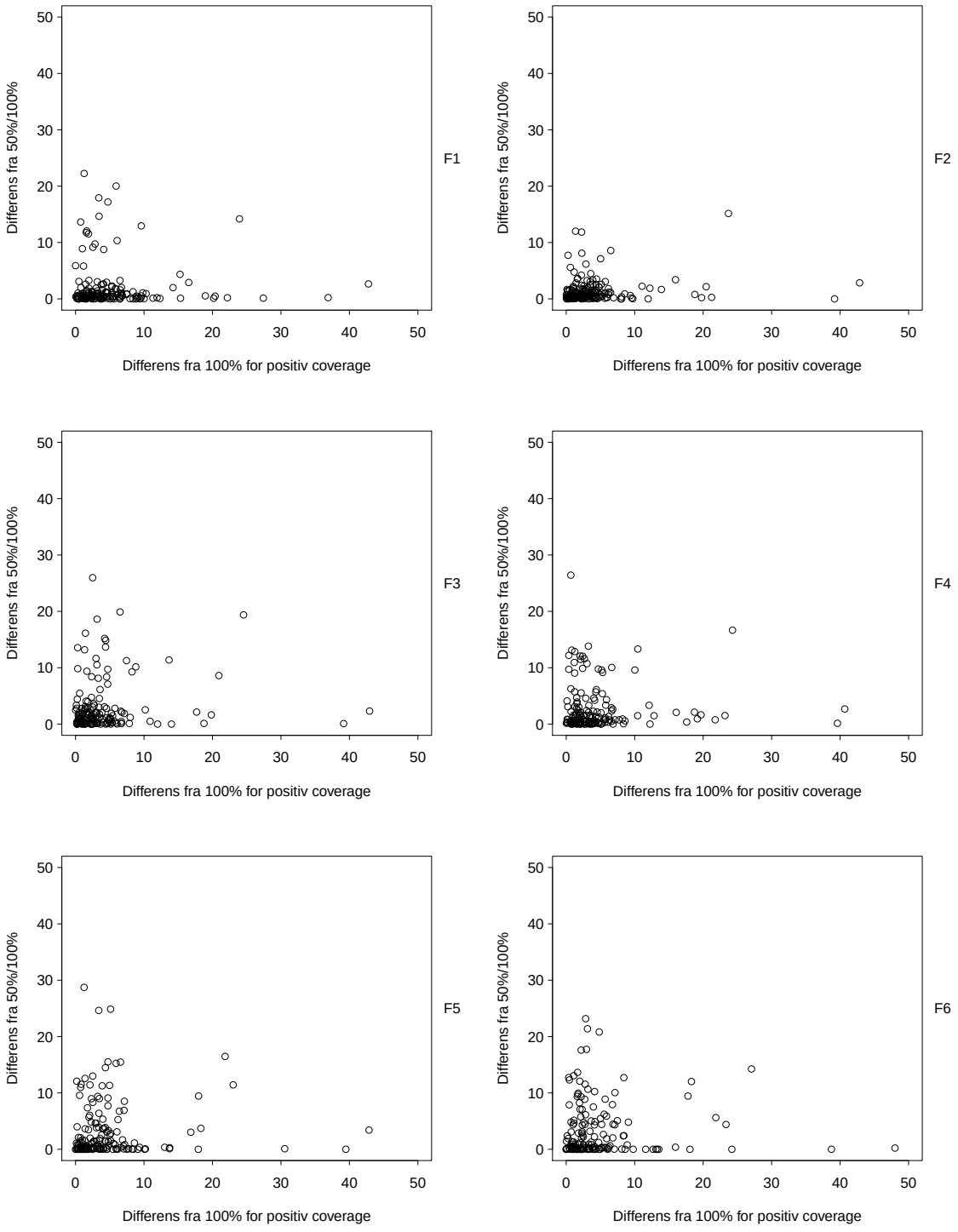
Det kan ses i Tabel 3.5, at der er 69 markører, som overskrider differensen på 10 mere end én gang. Markør rs9530435 er den, som skiller sig mest ud, da den overskrider differensen på 10 hele 20 gange, men den har dog kun et gennemsnit på 14,36. De fleste af de resterende markører overskrider kun få gange med et lavt gennemsnit. Desuden er der tre markører fra CART20, som også optræder i tabellen, hvor rs7844723 har det højeste gennemsnit på 11,81 med seks overskridelser.

3.3.5 Positiv coverage og det dominerende allel

Informationen om positiv coverage og det dominerende allel er undersøgt yderligere for at undersøge, om der er en sammenhæng mellem dem. Resultatet af dette er vist i Figur 3.29 og Figur 3.30 for henholdsvis duplikat A og B.



Figur 3.29: Sammenhæng mellem differens fra 100% for positiv coverage og differens fra 50%/100% for det dominerende allel for duplikat A.



Figur 3.30: Sammenhæng mellem differens fra 100% for positiv coverage og differens fra 50%/100% for det dominerende allel for duplikat B.

Det kan ses i Figur 3.29 og Figur 3.30, at der ikke er en klar sammenhæng mellem positiv coverage og det dominerende allel. Det, at profilen har en stor differens for positiv coverage, er derfor ikke ensbetydende med, at profilen har en stor differens for det dominerende allel. Der er dog nogle tilfælde, hvor enkelte markører overskrider differensen på 10 for både positiv coverage og det dominerende allel, hvilket kan ses i Tabel 3.6.

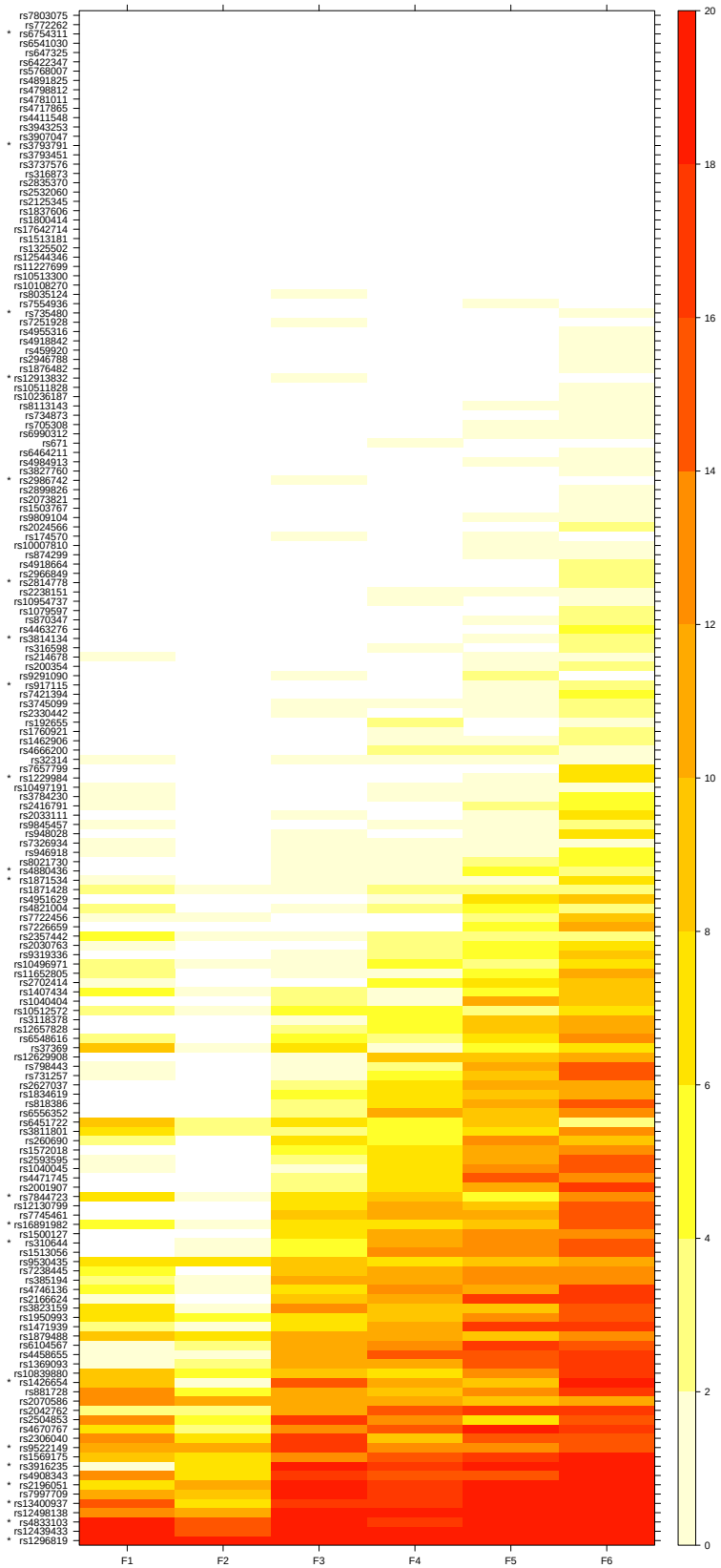
	Positiv coverage		Det dominerende allel	
Markør	Gennemsnit	Overskridelser	Gennemsnit	Overskridelser
rs9530435	24,76	12	15,47	12
rs2033111	15,97	2	11,70	2
rs2357442	23,02	1	11,44	1
rs1871428	18,60	1	10,18	1
rs4458655	10,46	1	13,32	1

Tabel 3.6: Markører, som overskrider differensen på 10 både for positiv coverage og det dominerende allel. For positiv coverage er der listet gennemsnittet af deres differens fra 50% samt antallet af overskridelser, og for det dominerende allel er der listet gennemsnittet af deres differens fra 50% og 100% for henholdsvis heterozygot og homozygot samt antallet af overskridelser.

Som det kan ses i Tabel 3.6, er der enkelte markører, som overskrider differensen på 10 både for positiv coverage og det dominerende allel. Dette gælder specielt rs9530435, som overskrider samtlige gange.

3.3.6 Udvælgelse af dropout markører

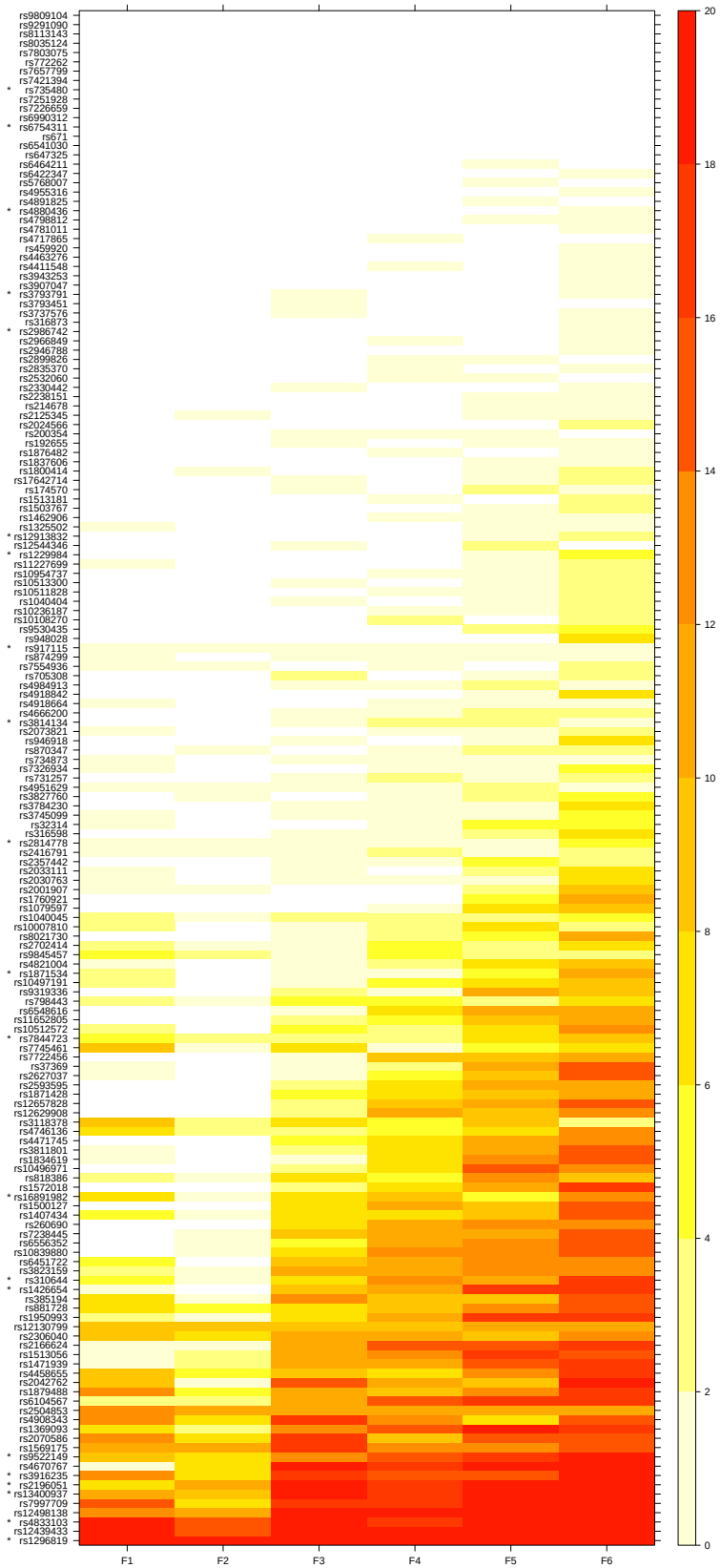
I fortyndingsdatasættet findes fem profiler, som ud fra allellernes reads er blevet genotypet på 165 markører i seks fortyndinger og i to duplikater. Hvis reads for alle baser hver især er 100 eller derunder for en markør, bliver markøren betragtet som to dropouts, da den har to alleller. I tilfælde af heterozygote alleller, hvor reads for baserne hver især er over 100, men heterozygot balancen for dem ligger uden for $[0,3;3]$, bliver markøren set som ét dropout. Hvis de heterozygote alleller kun har ét allel, hvor reads for én base er over 100, bliver markøren også set som ét dropout. Antallet af dropouts er til sidst summeret sammen for hver fortynding (F1-F6) og for hver markør, hvilket kan ses i Figur 3.31.



Figur 3.31: Antallet af dropouts for hver fortynding og for hver markør med en heterozygot balance på $[0,3;3]$. Farven hvid indikerer, at den pågældende markør ikke har nogen dropouts i den pågældende fortynding. Markørerne, som er markeret med en stjerne, blev anvendt i CART20.

Det fremgår tydeligt i Figur 3.31, at nogle markører har flere dropouts end andre, og mange af markørerne har kun få eller ingen dropouts. Markør rs1296819 er den eneste markør, som har 20 dropouts for hver fortynding, hvilket også er det maksimale antal dropouts for en fortynding. Derudover har rs12439433 og rs4833103 også 14-20 dropouts for hver fortynding. Det fremgår også tydeligt, at de fleste markører har flere dropouts for hver fortynding på nær F2-fortyndingen, som er den med mindst dropout. Dette stemmer godt overens med Figur 3.20 og Figur 3.21, da coverage for markørerne bliver mindre for hver fortynding. Det bemærkes også, at flere af de markører, som blev anvendt i CART20, ligger nede blandt de markører, som har flest dropouts.

En analyse er også blevet lavet med heterozygot balancen på $[0,67; 1,67]$. Resultatet af dette kan ses i Figur 3.32.

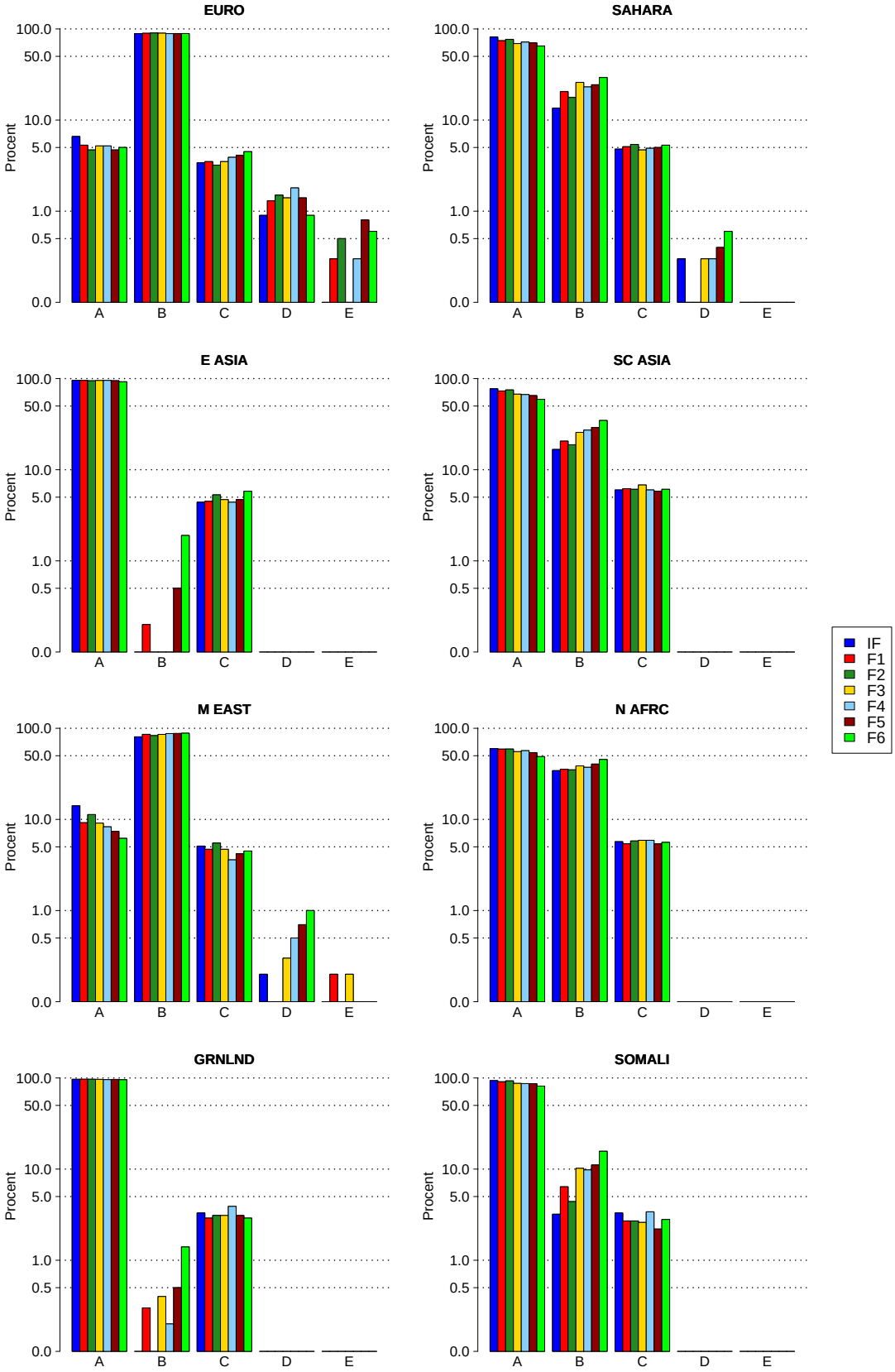


Figur 3.32: Antallet af dropouts for hver fortynding og for hver markør med en heterozygot balance på $[0,67; 1,67]$. Farven hvid indikerer, at den pågældende markør ikke har nogen dropouts i den pågældende fortynding. Markørerne, som er markeret med en stjerne, blev anvendt i CART20.

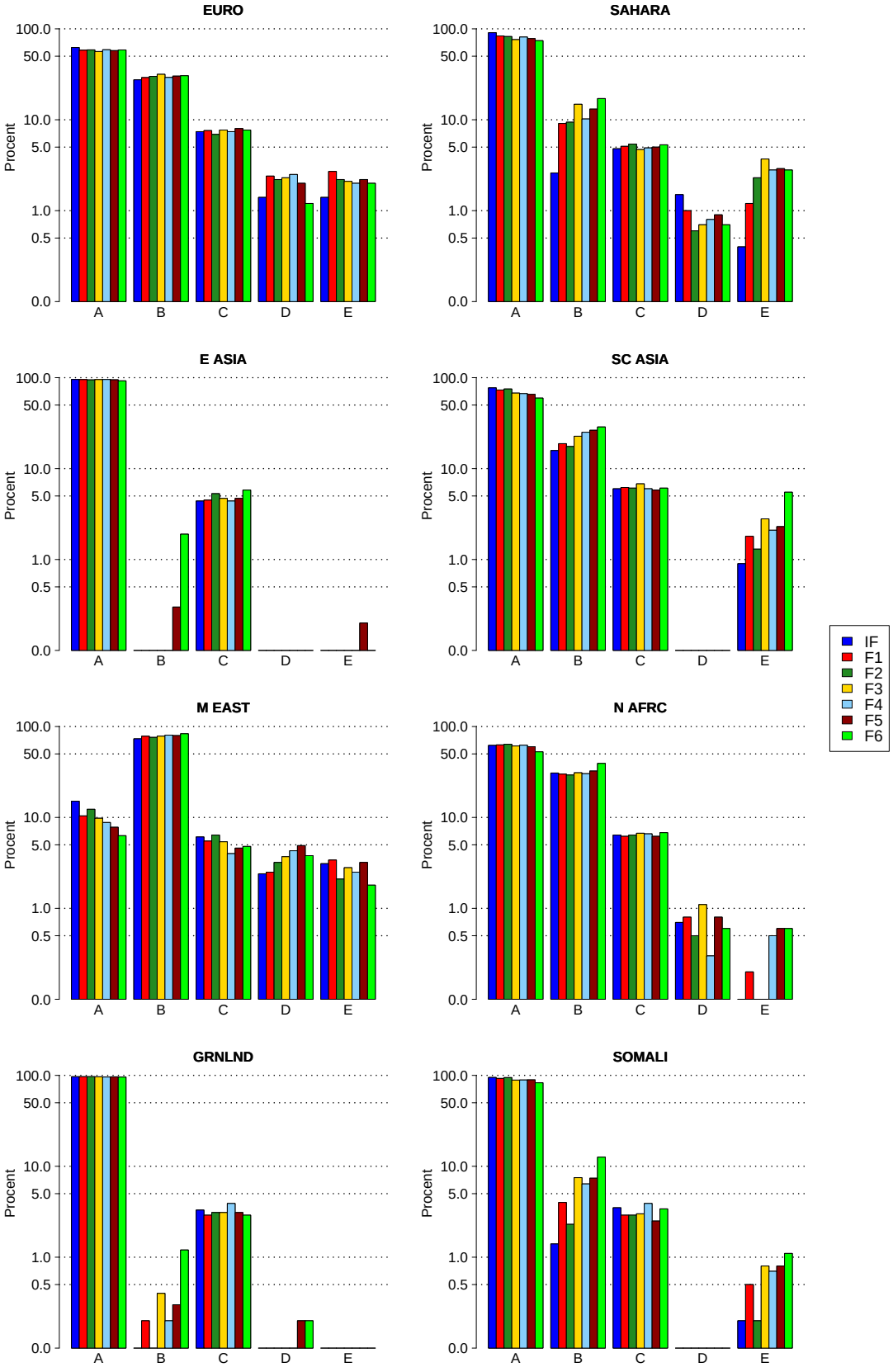
Det kan ses i Figur 3.32, at antallet af dropouts følger den samme tendens som fra Figur 3.31 dog med lidt flere dropouts i hver fortynding. Dette stemmer godt overens med Figur 3.22, da der var betydelig flere observationer, hvor deres heterozygot balance lå uden for $[0,67; 1,67]$, hvilket har givet flere dropouts. Den samme tendens giver dog, at konklusionen er den samme.

3.3.7 Fortyndede profiler

Der er lavet en analyse, som undersøger om fortyndede profiler har den samme klassificering af grupper i forhold til profiler, der ikke er blevet fortyndet. Der er simuleret 1.000 profiler inden for hver hovedpopulation, hvorefter profilernes markører er blevet fortyndet med hyppighederne fra Figur 3.31 og Figur 3.32. Til sidst er funktionen `genogeo` anvendt for at se, hvilke hovedpopulationer der bliver accepteret. I Figur 3.33 og Figur 3.34 er der henholdsvis anvendt alle markørerne og markørerne fra CART20, og de viser klassificeringen for de simulerede profiler, kaldet *IF*, da de ikke er fortyndet, samt de simulerede profiler, som er fortyndet med de seks fortyndinger, kaldet *F1*, *F2*, *F3*, *F4*, *F5* og *F6*.



Figur 3.33: Procentandel på logaritmisk skala af klassificeringen for de simulerede profiler samt de simulerede profiler, som er fortyndet med de seks fortyndinger. I analysen er der anvendt alle markørerne.



Figur 3.34: Procentandel på logaritmisk skala af klassificeringen for de simulerede profiler samt de simulerede profiler, som er fortyndet med de seks fortyndinger. I analysen er der anvendt markørerne fra CART20.

Det kan ses både i Figur 3.33 og Figur 3.34, at procentandelen af klassificeringen for de simulerede profiler, som er fortyndet, næsten er lige så stor som procentandelen af klassificeringen for de simulerede profiler i de fleste hovedpopulationer. SAHARA og SC ASIA er de hovedpopulationer, som skiller sig mest ud og har de største reduceringer og forøgeleser i henholdsvis gruppe A og B. Det er umiddelbart dem, som er mest tilbøjelig til at acceptere flere hovedpopulationer inklusiv den sande hovedpopulation efter deres profiler er blevet fortyndet.

4 Diskussion og konklusion

Formålet med dette speciale har været at undersøge individers afstamning ved brug af AIMs.

I den første del af analysen var formålet at finde en delmængde af markørerne, som bedst kunne skelne mellem hovedpopulationerne. Analysen viste, at funktionen `genogeo` blandt andet havde svært ved at skelne mellem 12 forskellige delmængder af hovedpopulationerne. Derudover viste det sig også, at forskellige delmængder af markørerne var bedst til at skelne mellem forskellige delmængder af hovedpopulationerne. De bedste modeller til at skelne mellem delmængderne af hovedpopulationer var CART20 og CART50, som bruger markørerne, der er angivet i Figur 3.11 og Figur 3.12. CART20 var en af de bedste modeller med hensyn til kun at acceptere den sande hovedpopulation, den havde en relativ høj procentandel i at acceptere flere hovedpopulationer inklusiv den sande hovedpopulation og brugte kun 21 markører. Med hensyn til CART50 var den en af de bedste til at acceptere flere hovedpopulationer inklusiv den sande hovedpopulation og brugte kun syv markører, hvilket gjorde den til den model, som brugte færrest markører. Markør rs16891982 var den klart vigtigste, da den optrådte i flest delmængder af hovedpopulationer samt havde den største vægt for alle de delmængder, som den optrådte i.

Grunden til at vælge to modeller til at være de bedste modeller opstod i dilemmaet om enten at vælge en model, som var bedst i kun at acceptere den sande hovedpopulation, eller vælge en model, som var bedst i at acceptere flere hovedpopulationer inklusiv den sande hovedpopulation. Det bedste scenarie er at stå tilbage kun med den sande hovedpopulation som værende accepteret. Ulempen ved at fortsætte med en profil, som er accepteret af flere hovedpopulationer inklusiv den sande hovedpopulation, er, at profilen kan ende med kun at blive accepteret af en hovedpopulation, som ikke er den sande hovedpopulation, eller overhovedet ikke at være accepteret af nogen hovedpopulation. Derfor blev CART50 også valgt som en af de bedste modeller, da den accepterer flere hovedpopulationer inklusiv den sande hovedpopulation.

I den anden del af analysen var hensigten at undersøge sensitiviteten af markørerne. I tilfældet, hvor en profil er accepteret af den sande hovedpopulation, men blev forkastet efter en ændring på en markør, var alle markørerne sensitive, men specielt rs671, rs3811801 og rs1800414 var mest sensitiv. Derudover viste det sig også, at SAHARA var mest tilbøjelig til at forkaste sine accepterede profiler efter en ændring på en markør. Markørerne fra CART20 var nede omkring de markører, som havde de største hyppigheder til at være sensitive. Det er altså meget informativt, hvis der for en profil, som er accepteret af den sande hovedpopulation, ændres på en af de markører, som bedst skelner mellem hovedpopulationerne.

I det andet tilfælde, hvor en profil først er forkastet af den sande hovedpopulation, men blev accepteret efter en ændring på en markør, viste det sig, at rs4821004, rs2125345 og rs2899826 var de mest sensitive markører. Derudover var SC ASIA mest tilbøjelig til at acceptere sine forkastede profiler efter en ændring på en markør. I dette tilfælde var markørerne fra CART20 mere spredt ud blandt alle de andre markører. Det er altså mindre informativt, hvis der for en profil, som er forkastet af den sande hovedpopulation, ændres på en af de markører, som bedst skelner mellem hovedpopulationerne.

I den tredje og sidste del af analysen var der fokus på at undersøge fortyndingsdatasættet med hensyn til markørernes coverage og tilbøjelighed til at give et dropout. Coverage for udvalgte markører viste, at der kan være markører med lav, middel og høj coverage. Tendensen for de fleste tilfælde var, at de første fortyndinger havde en relativ høj coverage, hvorimod coverage blev lavere i de sidste fortyndinger. Markør rs735480 var blandt dem med mest coverage samtidig med, at den også optrådte i CART20.

Analysen for de heterozygote tilfælde viste, at enkelte observationer lå uden for heterozygot balancen på $[0,3;3]$, og der var betydelig flere observationer, som lå uden for balancen på $[0,67;1,67]$.

Med hensyn til positiv og negativ coverage var der 31 markører, som overskred differensen på 10 fra 50%. Markør rs1760921 og rs20011907 var dem med de største differenser, da de havde et gennemsnit for deres differens på over 30. Derudover var der ti af markørerne, som overskred mere end halvdelen af gangene, hvor seks af disse markører overskred i alle fortyndingerne. De to markører, rs3916235 og rs1871534, var dem fra CART20, som skildte sig mest ud.

En analyse for det dominerende allel viste, at lidt under halvdelen af markørerne overskred differensen på 10 mere end én gang. Markør rs9530435 skildte sig mest ud ved at overskride differensen på 10 hele 20 gange med det højeste gennemsnit.

Det blev yderligere undersøgt, om der var en sammenhæng mellem positiv coverage og det dominerende allel, men der blev ikke fundet en klar tendens. Der var dog tilfælde, hvor enkelte markører overskred differensen på 10 for både positiv coverage og det dominerende allel, hvilket specielt var gældende for rs9530435, som også skildte sig mest ud i analysen for det dominerende allel.

I analysen med markørernes dropout med en heterozygot balance på $[0,3;3]$ viste det sig, at de fleste markører var tilbøjelige til at give et dropout og havde flere dropouts for hver fortynding. Dette var mest gældende for rs1296819, som var den eneste markør, der havde det maksimale antal dropouts i alle fortyndingerne. Derudover havde rs12439433 og rs4833103 også relativ mange dropouts. I det andet tilfælde med en heterozygot balance på $[0,67;1,67]$ kom der generelt lidt flere dropouts i alle fortyndingerne, men tendensen var den samme, så dermed forblev konklusionen også den samme.

Til sidst blev der undersøgt de simulerede profiler, hvor profilernes markører var fortyndet med hyppighederne fra analysen med dropout. Det viste sig, at procentandelen af klassificeringen for de simulerede profiler, som var fortyndet, næsten var lige så store som procentandelen af klassificeringen for de simulerede profiler i de fleste hovedpopulationer. SAHARA og SC ASIA havde de største reduceringer og forøgelser i henholdsvis gruppe A og B. Fortyndinger i profilernes markører har altså kun en lille negativ effekt med henyn til at acceptere flere hovedpopulationer inklusiv den sande hovedpopulation, selvom mange af markørerne fra CART20 lå nede blandt de markører, som havde flest dropouts.

På tværs af de forskellige analyser har der ikke været nogen markører, som har skildt sig særligt ud i mere end én analyse på nær i de få tilfælde, der er blevet nævnt.

Litteratur

Butler, J. M. (2010). *Fundamentals of Forensic DNA Typing.* Elsevier.

Donaldson, J. (2016). *T-Distributed Stochastic Neighbor Embedding for R (t-SNE).* https://cran.r-project.org/web/packages/tsne/tsne.pdf

Maaten, L. og Hinton, G. (2008). *Visualizing Data Using t-SNE.* Journal of Machine Learning Research 9. http://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf

Scientific, T. F. (2019a). *Sequencing Education.* https://www.thermofisher.com/dk/en/home/life-science/sequencing/sequencing-education.html.

Scientific, T. F. (2019b). *Targeted Sequencing.* https://www.thermofisher.com/dk/en/home/life-science/sequencing/dna-sequencing/targeted-sequencing.html.

Scientific, T. F. (2019c). *PCR Basics.* https://www.thermofisher.com/dk/en/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/pcr-education/pcr-reagents-enzymes/pcr-basics.html.

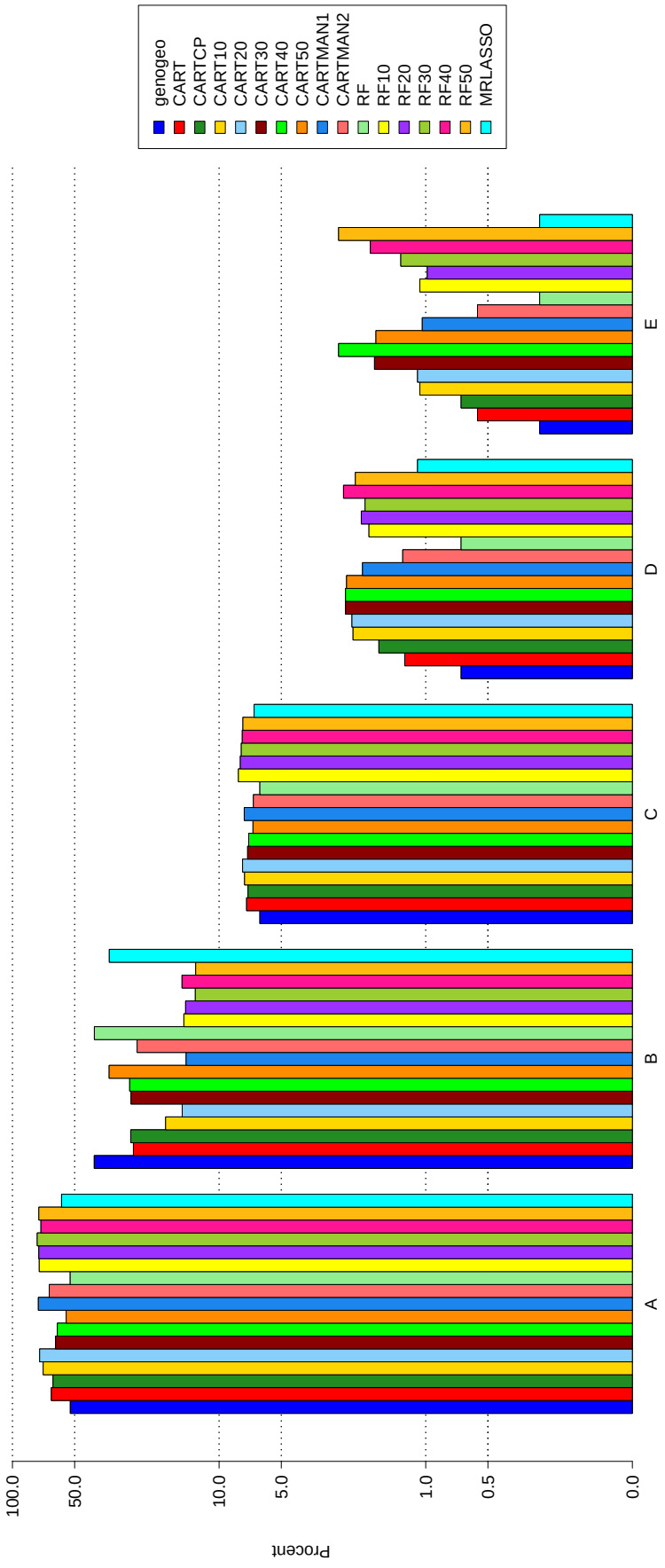
Danmarks Statistik (2019). *Statistikbanken.* https://www.statistikbanken.dk/statbank5a/default.asp?w=1920.

Tvedebrink, T. og Eriksen, P. S. og Mogensen, H. S. og Morling, N. (2018). *Weight of the Evidence of Genetic Investigations of Ancestry Informative Markers.* Theoretical Population Biology 120, side 1-10.

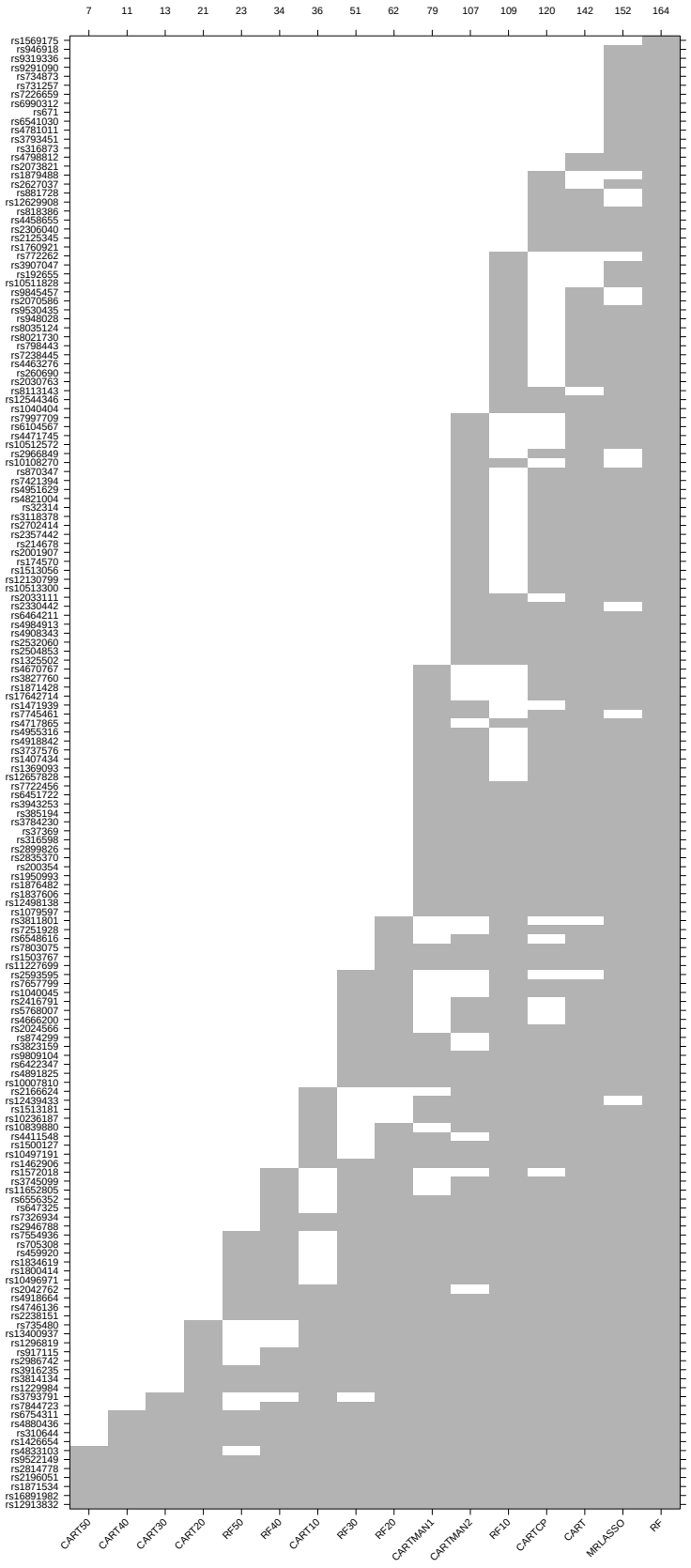
5 Bilag

Model	Antal markører anvendt
CART	142
CARTCP	120
CART10	36
CART20	21
CART30	13
CART40	11
CART50	7
CARTMAN1	79
CARTMAN2	107
RF	164
RF10	109
RF20	62
RF30	51
RF40	34
RF50	23
MRLASSO	152

Tabel 5.1: Antallet af markører, som de forskellige modeller anvender.



Figur 5.1: Procentandel på logaritmisk skala af klassificeringen for alle modeller.



Figur 5.2: Modellernes anvendelse af de forskellige markører. En anvendelse af den pågældende markør er markeret med grå. Modellerne er sorteret efter antallet af anvendte markører, og markørerne er sorteret efter hyppighed. Antallet af anvendte markører er angivet øverst.