Pangenom-studier av genus Enterococcus

Pangenome studies of genus Enterococcus

Hans Petter Brodal

UNIVERSITETET FOR MILJØ- OG BIOVITENSKAP

Institutt for Kjemi, Bioteknologi og Matvitenskap

Masteroppgave 60 stp. 2011

Forord

Denne masteroppgaven ble utført ved Institutt for Kjemi, Bioteknologi og Matvitenskap (IKBM) ved Universitetet for Miljø- og Biovitenskap (UMB). Arbeidet med oppgaven har hovedsaklig foregått i tidsrommet høsten 2010 og våren 2011, i tillegg til forberedende arbeider våren 2010. Det har vært en lærerik prosess som har gitt nye, og forhåpentligvis nyttige erfaringer.

Hovedveileder for denne oppgaven har vært førsteamanuensis Lars Snipen fra Biostatistikk-gruppa ved IKBM, og jeg vil takke for god veiledning og nyttige innspill underveis med arbeidet. Jeg føler jeg har fått den hjelpen jeg har bedt om og at han har tatt seg god tid til meg. Min andre veileder, forsker Dag Anders Brede ved IKBM har bidratt med interessant og nyttig informasjon om de studerte bakteriene. Etterhvert som oppgaven tok mer og mer form, ble det klart at arbeidet endte opp et annet sted enn det som først var tenkt. Dette førte til at rollen til Dag Anders som veileder ble mindre enn først antatt, men han bidro med nyttige tolkninger som ga resultatene en biologisk mening i tillegg til bare tall og bokstaver.

I tillegg vil jeg takke mora mi, Guro, og faren min, Peter, for god hjelp med korrekturlesing og konstruktive innspill til oppgaven.

Ås, mai 2011

Hans Petter Brodal

Sammendrag

Sekvenseringsteknologien er i dag på et stadium som gjør at utfordringene ikke lenger ligger i selve sekvenseringen av genomer, men i det å utnytte og behandle all informasjon fra genomene. Fokuset og tilnærmingen til denne oppgaven er metoder der hele genomer, i stedet for bare utvalgte deler, ble brukt for å sammenligne stammer og arter av bakterier. Oppgaven har altså vært et helgenom-studie der stammer fra utvalgte *Enterococcus*-arter er blitt studert.

Helgenom-studiet har i hovedsak vært fokusert på *Enterococcus faecalis* og *Enterococcus faecium*. I tillegg ble det tatt med noen få stammer fra to andre arter, *Enterococcus casseliflavus* og *Enterococcus gallinarum*. Stammene innad i artene, og artene i forhold til hverandre, ble sammenlignet i forhold til genetisk likhet/ulikhet, og det ble generert pangenom-trær, estimert pangenom- og kjernegenom-størrelse og estimert en parameter kalt genomic fluidity. Såvidt jeg vet, er det ikke tidligere blitt gjort noen lignende studie der disse artene sammenlignes på denne måten.

Såkalte pangenom-trær er diagrammer som baserer seg på relative avstander, eventuelt grad av ulikhet, mellom genomer. Vi observerte at alle trærne ga tydelig inndeling etter art, og at de ulike trærne viste ulik grad av likhet mellom stammene. Stort sett virket det som om stammene innenfor *E. faecalis* ligner mer på hverandre enn stammene innenfor *E. faecium*.

Pangenomet til en art består av alle de forskjellige genfamiliene innenfor arten, og er interessant fordi det kan si noe om den genetiske diversiteten som arten innehar, og dermed artens evne til å tilpasse seg ulike miljøer. Kjernegenomet er alle genfamiliene som finnes i alle stammer innenfor en art, og disse genfamiliene sier også noe om diversiteten ved å gi et mål på hva som er likt for alle stammene innenfor en art. Binomiske mixture modeller ble brukt for å estimere pangenom og kjernegenom for *E. faecalis* og *E. faecium*. *E. faecium* fikk estimert pangenomet til å være større enn for *E. faecalis*. Parametereren genomic fluidity, som er ment som et annet mål på diversiteten innenfor en art, ble også estimert til å være høyere for *E. faecium* enn for *E. faecalis*.

Det ble funnet en del genfamilier blant kjernegenene som kun fantes innenfor arten. For å finne ut mer om disse unike kjernegenene, ble de klassifisert etter hvilke funksjonelle COG-grupper (Cluster of Orthologous Groups) de tilhørte. Deretter ble det utført Fisher's exact test for å undersøke om noen av gruppene var over- eller underrepresentert blant de unike kjernegenene. Her så vi at noen av de funksjonelle gruppene var overrepresentert blant de unike kjernegenene, og disse kan kanskje forklare noen artskjennetegn.

Abstract

Sequencing-technology is now at a stage where the challenges no longer lies in the sequencing of genomes, but rather in utilizing and processing all of the data. The main focus this thesis is the study of whole genomes; a so called whole-genome study. The methods that are used focuses on entire genomes, rather than just selected parts, to compare strains and species of bacteria.

Two different *Enterococci* species, *Enterococcus faecalis* and *Enterococcus faecium* have been the main area of interest. In addition, a few genomes from two other species, *Enterococcus casseliflavus* and *Enterococcus gallinarum* were also included. The strains within the species, and also the species in relation to each other, were compared in terms of genetic similarity and difference. Pangenome trees were generated, pangenome and core-genome sizes were estimated, and a statistic called genomic fluidity was also estimated.

Pangenome trees are diagrams that are based on the relative difference between genomes. The distances between genomes reflects the amount of gene families they have in common, and not in common. The pangenome trees that were generated showed a clear separation due to different species. It also seemed that the largest distances between genomes were found within *E. faecium*, and that this species contains more diversity between strains than *E. faecalis*.

The pangenome of a species consists of all the different gene families found within that species. This can give an idea of the genetic diversity that the species possesses, and thus the species' ability to adapt to different environments. The core genome consists of all gene families found in all strains within a species, and the size of the core genome can provides a measure of expected similarity within a species. Binomial mixture models were used to estimate the pangenome and core-genome sizes of *E. faecalis* and *E. faecium*. We found the pangenome size to be larger for *E. faecium* than *E. faecalis*, but the core genome sizes were quite similar. Genomic fluidity, a proposed alternative measure of diversity within a species was also estimated, and the statistic showed lower amount of shared genes within *E. faecium* than *E. faecalis*.

Some of the gene families among the core genes only existed within the species. To find out more about these unique core genes, they were classified according to which functional COG (Cluster of Orthologous Groups) group they belong. Fisher's exact test was utilized to examine whether any of the groups were significantly over- or under-represented among the unique core genes, compared to all of the core genes. We found that some of the functional groups were over-represented among the unique core genes of *E. faecalis*, and these genes can maybe explain some characteristics of the species.

Innholdsliste

Kapittel 1

Innledning

Fokuset og tilnærmingen til denne oppgaven er komparative helgenom-studier, som betyr at det er brukt metoder som benytter hele genomer for å sammenligne bakterier.

1.1 Komparativ genomikk

Genomikk handler om studien av, eventuelt læren om, genomer. Komparativ genomikk er en retning innen genomikk der man studerer genomsekvenser fra flere organismer for å avdekke funksjonelt og genetisk slektskap. Det brukes i vår moderne tid automatiserte metoder for å sekvensere organismers genomer og bioinformatiske verktøy for å behandle disse, og behovet for bioinformatikk har økt etterhvert som sekvenseringsteknologien har utviklet seg. Begrensningen ligger ikke lenger i selve sekvenseringen, men i det å behandle store mengder sekvensdata.

Ulike organismer har en del gener som ligner mer eller mindre på hverandre, og som i mange tilfeller kan sies å være samme gen, særlig med tanke på funksjon. Man sier da at disse ulike variantene av et gen er medlemmer i samme genfamilie. Disse er sentrale innen komparativ genomikk, da man studerer hvilke genfamilier organismer har felles, og hvilke genfamilier de ikke deler. Når genomer sammenlignes for å finne ut hvor like eller ulike de er, er det dette man ofte baserer seg på. Det er også denne tilnærmingen som er blitt brukt i arbeidet med denne oppgaven.

Gener som er viktige, for eksempel husholdningsgener, er ofte godt konserverte og finnes som regel hos de aller fleste av organismene man antar er nært beslektet. Dette er gener man kan tenke seg har blitt nedarvet fra en felles stamfar, og kan kalles for homologe gener. Gener som kun finnes hos et fåtall i en beslektet gruppe, kan være et reultat av horisontal overføring fra et annet genom. Dette gjelder i all hovedsak for prokaryote organismer (bakterier). Bakterier, nærmere bestemt *Enterococcus sp.*, vil være hovedfokus for dette masterprosjektet.

Mange biologiske forskningsområder kan dra nytte av og bygge på informasjon fra komparative genomanalyser. Det genomiske materialet til ulike arter, eller stammer innenfor en art, blir

sammenlignet, og målet er å finne ut hvordan organismer er beslektet, hvilke egenskaper de deler og eventuelt ikke deler, og hvorfor. Muligens vil det også være mulig å si noe om den evolusjonære historien, og et mulig felles opphav for organismene man studerer.

Komparativ genomikk forutsetter at man har tilgjengelig DNA-sekvenser (genomsekvenser), og før gode sekvenseringsteknikker ble utviklet, var dette et tidkrevende, dyrt og lite effektivt arbeid. Man kunne altså ikke ta utgangspunkt i så store deler av et genom for å sammenligne organismer genetisk. Når man mot slutten av 1800-tallet såvidt begynte å klassifisere bakterier, ble dette i all hovedsak gjort ved å bruke fenotypiske markører, som for eksempel utseendet eller andre observerbare egenskaper. Man begynte etterhvert også å klassifisere bakterier på bakgrunn av de kjemiske komponentene i cellen, deriblant andel GC-nukleotider, komponenter i celleveggen(Gram-positive/-negative, annet), cytokromer med mer. Generelt sett har framskritt innen klassifisering og genomikk naturlig nok fulgt den tekniske utviklingen innen mikrobiologi og bioteknologi (Schleifer 2009) .

De første DNA-sekvensene som ble sekvensert var fra et virus, bakteriofagen φX174. Dette ble gjort av Fred Sanger og hans team på 1970-tallet, og dette ble gjort ved hjelp av DNA polymerase I fra *Escherichia coli* og DNA polymerase fra en bakteriofag, T4. De utarbeidet metoder og teknikker for sekvensering, genom-mapping, lagring av data og bioinformatiske analyser (Sanger & Coulson 1975).

1.2 16S rRNA - ett gen som markør

16S rRNA er en komponent i 30S rRNA, som igjen er en del av ribosomet til prokaryoter. Genet som koder for 16S rRNA er en viktig og godt konservert genetisk komponent, og dette betyr at genet ikke vil være veldig forskjellig fra art til art, og særlig lite forskjellig innenfor en art. Innen komparativ genomikk har 16S rRNA blitt brukt lenge, og det er flere grunner til dette. 16S rRNA er konstant med tanke på funksjon, samtidig som det er tilstede i alle organismer (Coenye & Vandamme 2003). I tillegg er sekvensene av en slik lengde at de er enkle å sekvensere direkte, ved hjelp av enzymet revers transkriptase (Woese 1987).

I 1977 ble det ved hjelp av delvis komplette, og etterhvert komplette sekvenser med 16S RNA-gener slått fast at archaeabakterier må klassifiseres som et eget rike blant prokaryoter, og det oppstod dermed et klart skille mellom bakterier og archaebakterier (Schleifer 2009). Bruken av

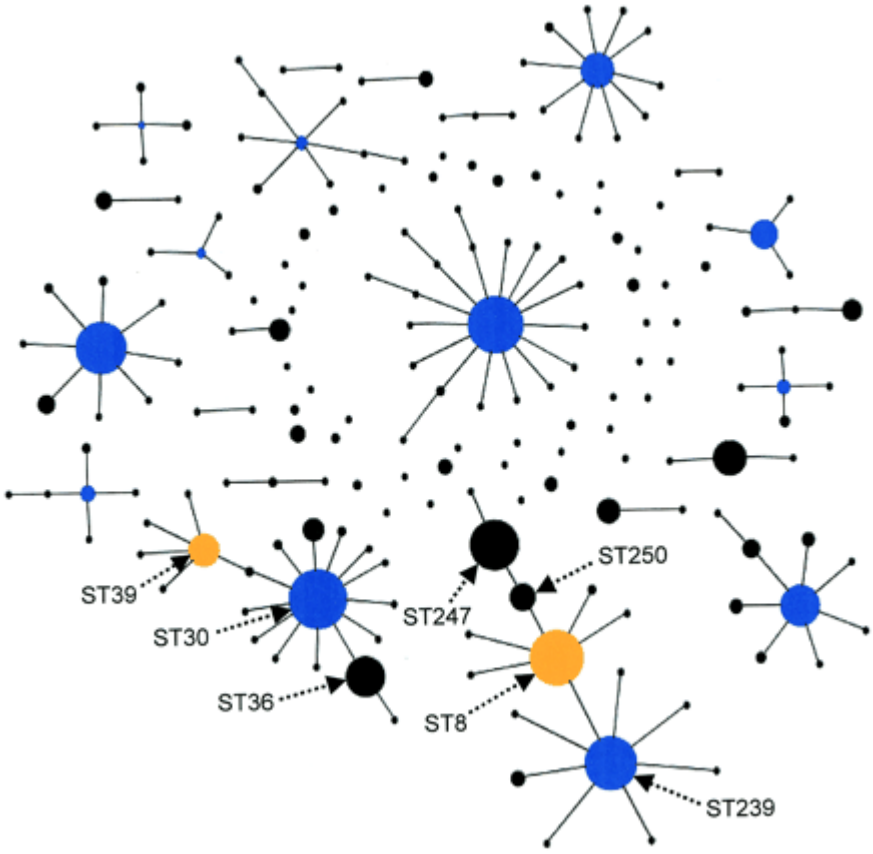
16S rRNA-gensekvenser var et gjennombrudd innen klassifisering, og førte til at man kunne klassifisere bakterier ved bruk av fylogeni.

16S RNA-gener er godt konservert, men kan være tilstede i flere varianter i et genom, og en variasjon på en til to prosent er ikke uvanlig. Dermed vil 16S RNA-markører best kunne benyttes for å skille, og finne relative avstander mellom ulike arter. Siden forskjellene er små, kan det være problematisk å skille mellom arter som er nært beslektet, og særlig innenfor samme art kan forskjellene være så små at denne markøren ikke vil kunne avdekke særlige forskjeller(Coenye & Vandamme 2003).

1.3 Multi Locus Sequence Typing(MLST)

Multi Locus Sequence Typing er en metode som ved å måle variasjonen i DNA-sekvensene til en bestemt gruppe gener, kan brukes til å finne relativt slektskap mellom organismer og muligens beregne den evolusjonære historien til organismene. Ofte brukes delsekvenser av sju-åtte husholdningsgener, og delsekvensene som brukes er av en slik lengde at de kan sekvenseres direkte med automatisert sekvenseringsutstyr, da dette bidrar til å sikre mer nøyaktige sekvenser. Det er et definert sett med gener som brukes for hver art, og det kan variere hvilke husholdningsgener, og hvor mange, som benyttes for ulike arter. Disse genene (allelene) utgjør en såkalt allel-profil for en art. Det finnes en rekke gensekvenser for hvert MLST-allel, og disse er åpent tilgjengelig i databaser på internett (Maiden et al. 1998; Urwin & Maiden 2003).
For eksempel fra www.mlst.net (MLST).

Resultatet av en MLST-analyse kan framstilles som et dendrogram (trediagram) basert på parvise forskjeller mellom allelprofilen til de forskjellige organismene, men det er også vanlig å fremstille resultatene grafisk ved hjelp av et radialdiagram (figur 1).



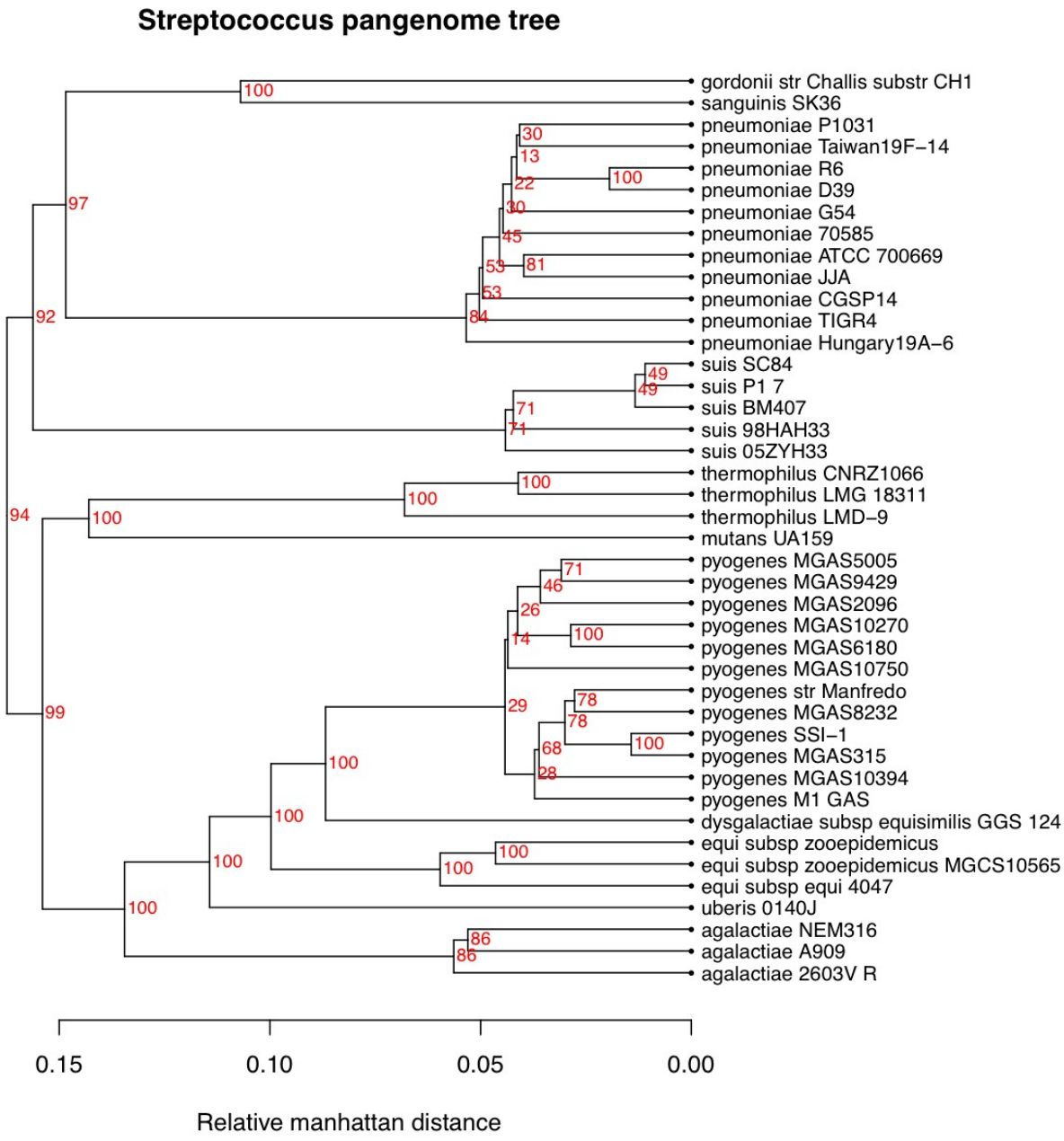
Figur 1: Eksempel på radial-diagram som output fra eBURST. Figuen er hentet fra en studie av arten Staphylococcus aureus (Feil et al. 2004). *Diagrammet viser clustre (sirkler) av stammer, og størrelsen på sirklene gjenspeiler antall stammer der avstanden mellom genomene er null. Nivået utenfor er stammer der avstanden er én. Dersom det brukes sju husholdningsgener kan avstanden være 0, 1,...,7 der avstanden er antall MLST-alleler som er ulikt mellom genomene.*

MLST er en automatisert metode som kombinerer mulighetene og fordelene innenfor high throughput sekvenseringsteknologi og bioinformatikk. Analyser kan lett reproduseres, da sekvensdata kan utveksles mellom laboratorier. Metoden gir som regel klare resultater, og vil i de fleste tilfeller kunne skille mellom stammer innenfor en bakterieart. Men siden MLST er basert på forskjeller mellom husholdningsgener, som er godt konserverte gener, vil man kunne miste evnen til å skille mellom organismer som er nært beslektet fordi forskjellene er små. (Enright et al. 2000)

1.4 Helgenom-studier

Sekvensenseringsteknologien har kommet langt fra den spede begynnelsen, og man er nå i stand

til å effektivt sekvensere hele genomer (Mardis 2008). Dette gjør igjen at man er kan utnytte hele genomer når man sammenligner organismer for å forhåpentligvis danne seg et mer riktig bilde av slektskapet mellom organismene. Et genom er definert som alle genene til en organisme.



Figur 2: Eksempel på pangenom-tre for Streptococcus sp. (Snipen & Ussery 2010).

Helgenom-studier omhandler å studere hvilke gener som finnes innen en art, kun hos én bestemt stamme eller i alle stammene som utgjør en art. Det er også av interesse å studere hvilke gener som utgjør forskjellene mellom stammer, og hva som kjennetegner disse. Basert på alle genene

prøver man å beregne relative forskjeller mellom stammene, og dette kan illustreres ved å konstruere såkalte pangenom-trær (figur 2). Dette er diagrammer som fremstiller hvor forskjellige genomene til ulike stammer er. Trærne viser relative avstander i form av Manhattan- eller Jaccard-avstander mellom genomene(Snipen & Ussery 2010). Ved beregningen av disse avstandene kan det legges vekt på ulike grupper gener avhengig av hvor ofte de opptrer i en populajon. På denne måten kan man fokusere undersøkelsen mot genfamilier mange organismer har til felles, eller man kan fokusere på sjeldne gener som kun finnes i få genomer. Det kan også legges lik vekt på alle genfamilier, enten de er vanlige eller sjeldne.

Selv om man undersøker hele genomer, er man ikke nødt til å ha komplette sekvenserte genomer. De fleste bakteriegenomener foreligger i forskjellige antall contigs, noe som betyr at genomet ikke består av en komplett sammensatt sekvens, men i deler man ikke har klart å sette sammen til én sekvens. Dette henger sammen med måten genomene blir sekvensert på. DNA-molekylet som utgjør et bakteriekromosom er for stort til å kunne sekvenseres direkte. Derfor brukes det spesifikke enzymer som kutter opp DNA-tråden i lesbare lengder. Når bitene av DNA-molekylet sekvenseres skjer dette i en helt tilfeldig rekkefølge, og sekvensen må etterpå settes sammen igjen på bakgrunn av overlappende sekvenslikhet. Den siste delen av sekvenseringsarbeidet, å få satt sammen en komplett genomsekvens er arbeidskrevende, og dermed kostbar, og belønningen per i dag veier ikke opp kostnadene i forhold til tid og penger.

Alle forskjellige gener man finner innenfor en art er en samling man kaller for et pangenom, mens samlingen av de genene som er felles for alle genomene i en art kalles kjernegenom. Størrelsen på pangenomet kan si noe om artens evne til å tilpasse seg forskjellige miljøer. En art som er representert i mange forskjellige miljøer vil sannsynligvis inneha større genetisk variasjon enn en art som kun finnes i ett eller få forskjellige miljøer. Det er dermed også forventet at pangenomet til en slik art er større enn hos den andre. En sentral del av denne oppgaven vil være å estimere pangenomet og kjernegenomet til *E. faecalis* og *E. faecium*.

Komparative helgenom-studier har tidligere blitt utført for blant annet *Escherichia coli* (Rasko et al. 2008) og *Streptococcus agalactiae* (Tettelin et al. 2005). Her ble det funnet at diversiteten innen disse bakterieartene var overraskende stor, og at helgenom-studier av bakterier vil være avgjørende for å studere årsaken til at enkelte stammer innenfor en art har bestemte egenskaper. Dette er blant annet viktig i sammenheng med utvikling av vaksiner mot patogene bakterier (Mora & Telford 2010).

1.5 Enterococcus

Enterococcus er en gruppering melkesyrebakterier, og består av bakterier som finnes naturlig i menneskers og dyrs fordøyelseskanaler, men også i andre miljøer som blant annet jord og vann. Enterococcer i jord og vann kan ofte gi en indikasjon på forurensing forårsaket av avføring. I tillegg til dette utgjør enterococcer ofte en del av de naturlige organismene i fermenterte matvarer (Lepage et al. 2006).

Navnet *Enterococcus* ble først brukt av Thiercelin i 1899, i forbindelse med å beskrive bakterier funnet i menneskelig avføring. De er klassifisert som Gram-positive, fakultativt anaerobe, katalase-negative kokker. De opptrer alene, i par eller i kortere kjeder. Fram til 1984 ble flere av enterococcene, både *E. faecalis* og *E. faecium*, kalt *Streptococcus faecalis* og *Streptococcus faecium* (Devriese et al. 1993; Solheim 2010) .

Fram til omtrent 1980 ble ikke enterococcer regnet for å være spesielt viktige patogene organismer, men dette bildet endret seg etter identifiseringen av multiresistente stammer. Det ble lenge kun innrapportert infeksjoner i pasienter med nedsatt immunforsvar, men senere har man funnet at bakteriene er langt mer viktige patogene organismer enn først antatt. I tillegg til å forårsake ulike infeksjoner, har mange enterococcer utviklet, ved evolusjon, eller anskaffet ved horisontal genoverføring, resistens mot en del antibiotika. Dette har ført til at enterococcus-infeksjoner har blitt vanskelige å behandle. Spesielt er det *E. faecalis* og *E. faecium* som står for de fleste infeksjoner forårsaket av enterococcer, der igjen de fleste, historisk sett, har blitt forårsaket av *E. faecalis*. Men etterhvert som resistens mot vancomycin og ampicillin har spredd seg, har dette forholdet forskjøvet seg mer mot *E. faecium*, da denne arten regnes som den dominerende blant vancomycin- og ampicillinresistente enterococcer (Levine 2006; Solheim 2010) .

I helgenom-studiet som er gjort i denne oppgaven, er det blitt brukt genomer fra fire forskjellige *Enterococcus*-arter: *E. faecalis, E. faecium, E. casseliflavus* og *E. gallinarum*.

1.6 Problemstilling

Innenfor *Enterococcus*-artene finnes det stammer som blant annet er kommensale, probiotiske eller patogene, det vil si det er en viss variasjon mellom stammene innenfor disse artene. Både *E. faecalis* og *E. faecium* har vist stor motstandsdyktighet mot antibiotika, og det er påvist resistens mot både vancomycin og ampicillin blant stammer innenfor disse artene, og flere andre enterococcer.

For å kunne si noe om hvor sannsynlig det er at en bakterie er patogen, er det relevant å vite noe om variasjonen man finner innenfor arten. Hovedfokuset for denne oppgaven har vært å kartlegge diversitet innenfor enterococcer, da spesielt for *E. faecalis* og *E. faecium*. Dette ble gjort ved å klustre gener i genfamilier og estimere pangenom-størrelse og kjernegenom-størrelse, både totalt og innenfor de to nevnte artene respektivt. Det var også noe fokus på å finne karakteristiske gener for de to artene, og undersøke om disse representerer spesielle funksjonelle grupper av gener. I tillegg ble det vurdert et alternativ til pangenom-størrelse, kalt genomic fluidity, for å beskrive diversiteten innen en gruppe bakterier.

Kapittel 2

Materialer og metoder

Alle analyser og beregninger tilknyttet arbeidet med denne masteroppgaven ble gjort i R, som er et åpent og gratis programvaremiljø utviklet for statistisk dataanalyse. R kan lastes ned fra: www.r-project.org (*The R Project for Statistical Computing*).

2.1 Sekvensdata - Innsamling og bearbeiding

For de fleste sekvenserte organismer, ligger sekvensene tilgjengelig i databaser på internett, og sekvensdata ble til å begynne med lastet ned fra National Center for Biotechnology Information (NCBI) manuelt ved å bruke "copy – paste"-metoden. De færreste av genomene er fullstendig sekvensert, og dette betyr at de fleste genomene som er brukt, foreligger i et varierende antall(tre-fire til flere hundre) contigs. Nettsidene til NCBI har ikke noen god løsning for å laste ned alle delsekvensene(contigs) til et genom på en rask måte, og dette viste seg dermed å være en strevsom, tidkrevende og lite effektiv metode. Innsamling av sekvensdata ble derfor løst ved at hovedveileder skaffet til veie sekvensdata på en mer egnet måte. Totalt ble det samlet inn sekvensdata for 50 genomer fra fire ulike *Enterococcus*-arter; 25 fra *E. faecalis*, 21 fra *E. faecium*, tre fra *E. casseliflavus* og ett fra *E. gallinarum*.

2.2 Genprediksjon

Det ble brukt Prodigal (Hyatt et al. 2010) for å predikere gener i de nedlastede genomene. Dette er et effektivt og gratis program, og ble brukt for å finne gener i alle genomene. Prodigal ble kjørt via et R-script *script_prodigal.R*. Resultatene fra Prodigal ble lagret i fasta-formaterte filer, der sekvensen til alle genene funnet i hvert genom ble samlet i én fil, altså én fil per genom. Deretter ble det telt opp hvor mange gener som ble funnet i hvert genom, og dette ble lagret i en oversiktstabell for genomene (tabell 1).

Etter å ha samlet inn sekvensdata og gjort genprediksjoner på disse, ble det laget en oversikt over genomene. Her ble det tatt med informasjon om antall basepar, antall contigs, andel GC-

nukleotider og antall gener funnet med Prodigal. Denne informasjonen ble hentet ut av dataene ved hjelp av et script skrevet i R, *script_deskriptiv_genom.R*. For å illustrere sammenhengen mellom ulike størrelser innen bakteriegenomer, ble det også laget plot av antall contigs mot antall gener funnet, antall basepar mot antall gener funnet og andel GC-nukleotider mot antall gener funnet. Prodigal finner også en del ufullstendige gener, kalt partielle gener. For å vise sammenhengen mellom antall contigs og antall partielt predikert gener, ble det også laget et plot for dette.

2.3 Genfamilier

2.3.1 BLASTing

Første steg i helgenom-analysen, er å BLASTe alle genomene mot hverandre, alle mot alle. Basic Local Alignment Search Tool, forkortet BLAST, er et verktøy som benytter en algoritme som finner de beste lokale sammenstillingene av to sekvenser. Vi ønsket å undersøke slektskap mellom genomene, og brukte derfor BLAST for å finne likheter mellom proteinsekvensene fra genomene. Mer nøyaktig lette vi etter genfamilier, som er gener i forskjellige organismer som er så like at vi sier de er samme gen, og dermed utgjør en genfamilie. Det ble brukt protein-BLAST, og alle gensekvensene funnet med Prodigal ble translaterert til proteinsekvenser ved hjelp av et R-script, *script_dna_til_protein.R*, før BLASTing.

For å ha kontroll på alle filnavn og navn på genomer, ble det opprettet en mapping-fil, *mapping_enterococcus.txt*, ved hjelp av R-scriptet *script_enterococcus_mapping.r*. Det ble brukt en del kildekode utviklet av hovedveileder i forbindelse med BLASTing, og denne forutsatte en slik mapping-fil. Fila består av en tabell med tre kolonner; genom-id (unik identifikator for hvert genom), filnavn og navn på organismen/genomet. Fila inneholder én rad for hvert genom, og brukes blant annet av R-scriptet *script_fastaprep.R* for å redigere beskrivelsen for hver delsekvens i filene med proteinsekvenser. Dette er nødvendig for å kunne BLASTe alle sekvensene mot hverandre, og samtidig ha god kontroll på resultatene ved å bruke entydige navn på alle sekvenser. Derfor ble beskrivelsen til hver delsekvens fra hvert genom gitt et nummer i tillegg til den unike identifikatoren for genomet sekvensen tilhører.

Selve BLASTingen ble gjort slik at alle genomene ble BLASTet mot hverandre. Dette betyr at det ble BLASTet begge veier, dvs at alle genomer ble brukt som både spørresekvens og database,

og alle genomene ble også BLASTet mot seg selv. BLASTingen ble gjort lokalt på en datamaskin veileder har tilgang til. Dette fordi det er en datamaskinkrevende analyse som tar en god del til, og som viste seg å være for krevende for min egen datamaskin.

Alle de lokale sammenstillingene som ble funnet av BLAST ble lagret i en resultatfil. Det ble opprettet én resultatfil for sammenstillingene av sekvenser fra ett genom mot et annet genom. Denne fila består av tolv kolonner, der hver rad inneholder blant annet navn på spørresekvens(query), treffsekvens(hit), lengde på sammenstilling, andel identitet, bit score og E-verdi. Etter å ha BLASTet 50 genomer, alle mot alle, var det til slutt generert 2500 (50x50) resultatfiler.

2.3.2 Clustring

Genfamilier er samlinger av gener som er så like at de kan sies å være samme gen, selv om de ikke er identisk like. For å behandle resultatene fra BLAST-søket, og bestemme genfamilene, ble det brukt kildekode utviklet av veileder (Snipen et al. 2009). Scriptet *script_panMatrix.r* leser inn BLAST-resultatene og kaller funksjoner som beregner genfamilier og som deretter oppretter en pan-matrise.

Det første steget i clustringen av genfamilier er at en funksjon, *preBlast*, setter opp en tabell (data frame) med sammenstilling av alle sekvenser mot seg selv, og lengde på sammenstillingen. Denne brukes for å kontrollere at alle sekvenser er med i BLASTingen.

Neste steg er beregning av genfamilier. Dette ble basert på 50-50 regelen, som er en regel som brukes for å bestemme om gener tilhører samme genfamilie. Den er definert ved at sekvenslikheten mellom to gener må være minst 50%, og samtidig må sammenstillingen dekke minst 50% av både treff- og spørresekvens.

Clustring betyr i prinsippet at det dannes forbindelser mellom gener som tilhører samme familie og som oppfyller 50-50-regelen. Hvert gen representerer en node i et trediagram, og dersom ett gen tilhører samme familie som et annet gen, dannes det en forbindelse mellom disse. Dersom et tredje gen er likt nok til å være i familie med ett av de andre genene i familien, blir også dette genet regnet som et medlem av genfamilien. Dette genet kan dermed sies å være i familie med et

gen det ikke direkte er likt nok til å være familie med, men blir det allikevel via genene mellom disse. Altså blir gener tilhørende samme familie fordi de begge er i familie med gener mellom dem, selv om de kanskje ikke ville vært i direkte familie med hverandre i følge 50-50 regelen. Denne måten å clustre genfamilier på kalles single-linkage.

Funksjonen som utfører clustringen av genfamilier heter *fiftyFifty*, og denne leverer en tabell (data frame), *gfam.frame*, med kolonner for sekvenstag og nummer på genfamilie. Alle sekvenstagene er unike, og inneholder identifikator for genom og nummer på sekvens fra genomet. Denne tabellen ble i neste runde brukt for å sette opp en pan-matrise.

Pan-matrisa består av en kolonne for hvert genom og en rad for hver genfamilie. Om en genfamilie er representert med minst ett medlem i et genom, markeres dette med "1" i matrisa. Dersom genfamilien ikke fins i genomet markeres dette med "0". Pan-matrisa brukt i denne studien ble satt opp av funksjonen *panmat*, og som brukte gfam.frame som grunnlag.

Etter å ha clustret BLAST-resultatene og opprettet en pan-matrise, ble det laget et stolpediagram over fordelingen av genfamilier. Det vil si en framstilling av hvor mange genfamiler som finnes i ett, to, ..., femti genomer. Det ble også laget slike stolpediagram for *E. faecalis* og *E. faecium*, respektivt.

2.3.3 Pangenom-trær

For å illustrere den relative likheten mellom genomene ble det laget pangenom-trær. Det ble laget trær der ulike grupper av genfamilier ble lagt vekt på, og disse grupperingene baseres her på hvor ofte genfamilier opptrer blant et utvalg genomer. Genfamilier som alltid er tilstede, kalles *core*-gener, genfamilier man finner i mange genomer kalles for *shell*-gener og genfamilier som kun opptrer i noen få genomer kalles *cloud*-gener. På bakgrunn av dette, ble det konstruert trær med forskjellig typer vekting; lik vekting (*flat*) for alle genfamilier, vekting som legger mer vekt på det som er likt hos de fleste genomene (*shell*) og vekting som legger vekt på genfamilier som finnes i få genomer (*cloud*). Kjernegener som vi finner i alle genomene, vil ikke påvirke et pangenom-tre, men dersom det bare er ett genom som mangler et gen, vil dette genet ha en effekt.

Relativ likhet illustreres ved å bruke relative manhattan-avstander, der avstanden mellom to genomer i og k er gitt ved:

$$D_{i,k} = (1/W) \sum_{j=1}^{n} w_j |M_{i,j} - M_{k,j}|$$

Der n er totalt antall genfamilier, w_j er vektingen til gen j, og W er summen av alle disse vektene. M representerer pan-matrisa, og $M_{i,j}$ betyr dermed rad (genfamilie) j og kolonne (genom) i.

For å illustrere hvor stabile grenene i trærne er, ble det gjort bootstrappinger der genfamilier blir re-samplet. Tallene som vises i diagrammene er prosentandelen av de re-samplede, og re-clustrede trærne hvor grenene har de samme "bladene" (nodene) (Snipen & Ussery 2010).

Selve konstruksjonen av pangenom-trærne ble gjort i R, ved hjelp av et script, *script_panTree.R,* og kildekode, *trees.R*, utviklet av Snipen & Ussery (2010).

2.4 Estimering av kjernegenom og pangenom

2.4.1 Kjernegenom og pangenom

Mange bakteriegenomer er blitt sekvensert og mange flere er underveis. Dette bidrar til at utfordringene innen komparativ genomikk stort sett er forbundet med komparativ pan-genomikk, eller helgenom-studier som det også kan kalles. I denne sammenhengen fokuseres det ofte på to grupper gener, og størrelsen på disse. En del av genene hos bakterier finnes hos alle stammene innenfor arten (kan også være tilstede i andre arter), og disse kalles for kjernegener, og utgjør det såkalte kjernegenomet. Dette er en samling av gener som er så godt konservert at de finnes i alle stammene av en art, og de kan tenkes å representere noe av det essensielle ved arten. Av kjernegenene er det igjen et utvalg som kun finnes innenfor arten, men som fortsatt finnes i alle stammene, og disse kalles unike kjernegener.

En annen viktig og interessant gruppe består av alle de forskjellige genfamiliene som finnes innenfor en populasjon av genomer, som regel en art. Denne samlingen gener kalles for et pangenom, og dette kan fortelle noe om artens evne til å tilpasse seg forskjellige miljøer. Et stort pangenom kan tyder på at bakterien har tilpasset seg mange miljøer, og dermed har en genetisk diversitet av en viss størrelse.

Å sekvensere absolutt alle eksisterende stammer av en art, er praktisk umulig, så derfor vil den reelle størrelsen på både kjerne- og pangenom forbli ukjente parametere. Men det er foreslått metoder for å kunne estimere disse størrelsene. Å estimere pangenom betyr i praksis å estimere hvor mange nye genfamiler som finnes i de genomene som ennå ikke er sekvensert. Altså må det estimeres hvor mye pangenom-størrelsen øker når man legger til nye genomer fra populasjonen. Hvor mange nye genfamilier kan man forvente å finne for hvert nye genom? Dette kan være problematisk å estimere, særlig i forbindelse med veldig lav sannsynlighet for at svært sjeldne gener dukker opp (Kislyuk et al. 2011).

2.4.2 Mixture model

Estimering av kjernegenom går ut på å estimere hvor mange genfamiler som finnes i absolutt alle genomer i populasjonen. Altså må man gå ut fra hvor mange gener man finner i alle sekvenserte genomer og hvor mange av disse man ikke finner etterhvert som nye genomer blir sekvensert.

Utfordringene ved estimering av den sanne pan-genom-størrelsen til en art, er knyttet til avhengighet mellom genomer og prediksjon av sjeldne gener. For å hanskes med avhengighet mellom genomer, og gener som opptrer i forskjellig grad, kan man benytte en binomisk mixture-model for å estimere størrelsen på pangenom og kjernegenom. Kjernegenomet, pangenomet og tettheten, eventuelt fordelingen, av genfamilier i utvalget brukes som utgangpunktet for å estimere pan- og kjernegenom for populasjonen utvalget representerer. I denne studien ble det benyttet samme fremgangsmåte, kildekode og script for bruk i R, som Snipen et al. (2009). Det vil i praksis si at en funksjon, *binomixestimate*, estimerte pangenom og kjernegenom ut fra pan-matrisa som ble satt opp tidligere. Grunnlaget for estimering av pangenom og kjernegenom ved bruk av binomiske mixture modeller er beskrevet i etterfølgende avsnitt:

Størrelsen på pangenomet til utvalget kalles n. Denne er gitt ved:

$$n=\sum_{g=1}^{G} y_g \quad (1)$$

der G er antall genomer i utvalget, og y_g er antall genfamilier som finnes i g genomer. Den virkelige størrelsen på pangenomet er da gitt ved

$$\eta=n+y_0 \quad (2)$$

der y_0 står for de genfamiliene som til nå er observert i 0 genomer, det vil si det ukjente antallet genfamilier som ennå ikke er observert. Å estimere pangenomet vil derfor i praksis si å estimere y_0. Det antas uavhengighet mellom genfamiliene, og kan dermed knytte y_0 til $y_1,...,y_g$, slik at $y = (y_0,y_2,..., y_g)$. y er dermed en multinomisk vektor. Det antas at den reelle pangenom-størrelsen er konstant, i alle fall på et gitt tidspunkt, og om vi antar uavhengighet mellom genfamiliene, er det en vektor med tilhørende multinomiske sannsynligheter $\theta=(\theta_0,...,\theta_G)$ knyttet til y. Disse sannsynlighetene gjenspeiler sannsynligheten for at et gen blir oppdaget i 0, 1, ..., G genomer. For y_0 gjelder $E(y_0)=\eta\theta_0$. Dette leder videre til at $E(n)=\eta(1-\theta_0)$, og kombinert gir disse uttrykkene følgende:

$$E(y_0)=E(n)\frac{\theta_0}{1-\theta_0} \quad (3)$$

Ved å bruke *n* som estimat for E(n), kan vi predikere y_0 om vi kan estimere θ_0. For å kunne finne dette estimatet, antas det en viss glatthet i den multinomiske sannsynlighetstettheten. Dette gjør det mulig å bruke en binomisk mixture model for å estimere θ_0. Det kan da tilpasses en kombinasjon av forskjellige binomiske tettheter, etter tettheten av genfamiler som finnes i 1,2,...,G genomer. Nærmere bestemt benyttes det binomiske fordelinger med ulik sannsynlighet for suksess, som i denne sammenhengen er at et gen (egentlig genfamilie) er tilstede i et genom tilfeldig trukket fra en populasjon. De forskjellige binomiske tetthetene som benyttes, kaller vi komponenter, der K er antall komponenter i modellen og π_k angir hvor mye vekt som tillegges komponent k:

$$\theta_g = \sum_{k=1}^{K} \pi_k f(g;\rho_k) \quad der\ g = 0,\ldots,G \qquad (4)$$

Det er naturlig å knytte en av komponentene i modellen til estimeringen av kjernegenom, og her antas det at sannsynligheten for å finne et kjernegen i et tilfeldig valgt genom er 1,0. Kjernegener er per definisjon alltid tilstede i alle genomer innenfor populasjonen. Den første komponenten i modellen gis dermed en deteksjonssannsynlighet $\rho_1 = 1{,}0$.

For å estimere de resterende parameterene er det benyttet en maximum-likelihood-funksjon:

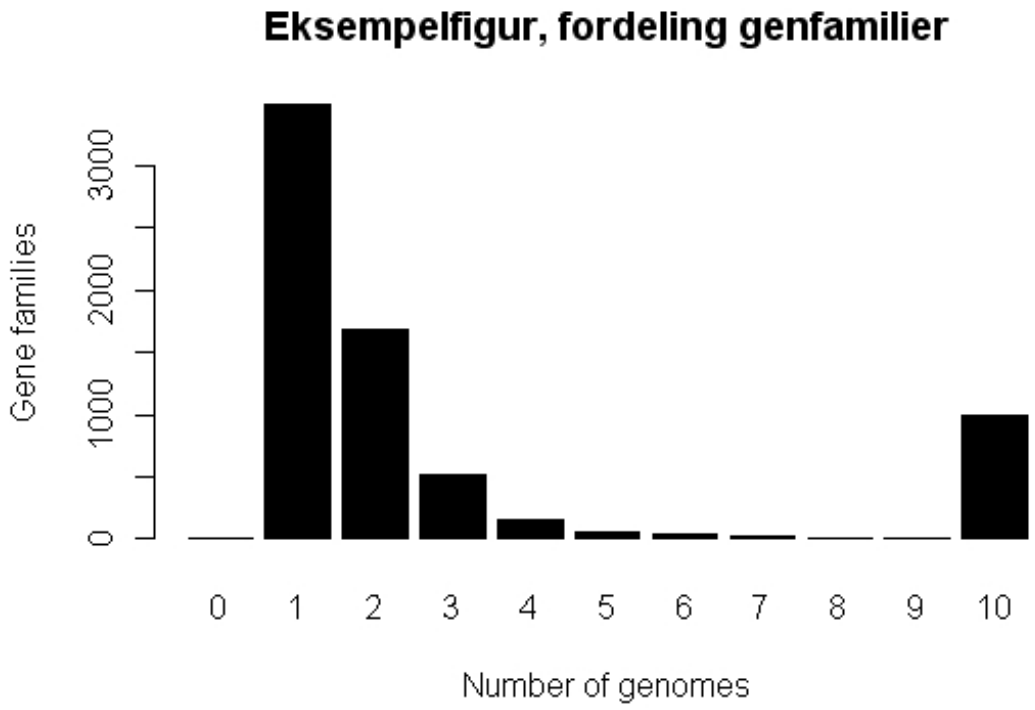
$$l(\pi,\rho|K) = \sum_{g=1}^{G} \log(\theta_g / (1-\theta_0)) + C \qquad (5)$$

Der $\theta_0, \ldots, \theta_G$ er avhengig av π og ρ, og C er en konstant uavhengig av disse parameterene. For et valg av K, estimeres π og ρ ved å maksimere for uttrykket over, som bare involverer($y_1,\ldots, y_G$). Estimatene vi får ut av dette brukes igjen i de tidligere uttrykkene for å estimere y_0.

Det siste steget er å avgjøre hvor kompleks modell som bør brukes. Fordi det brukes et maximum likelihood-kriterie i bunn, kan dette gjøres ved å benytte Bayesian Information Criterion (BIC). Dette for å sikre at man bruker en forklaringsmodell som er godt tilpasset data som samtidig sikrer at man ikke estimerer flere parametere enn nødvendig (Schwartz 1978). Det blir valgt et antall K komponenter til modellen som minimerer BIC:

$$BIC(K) = -2l(\pi, \rho | K) + (2K - 2)\log n \quad (6)$$

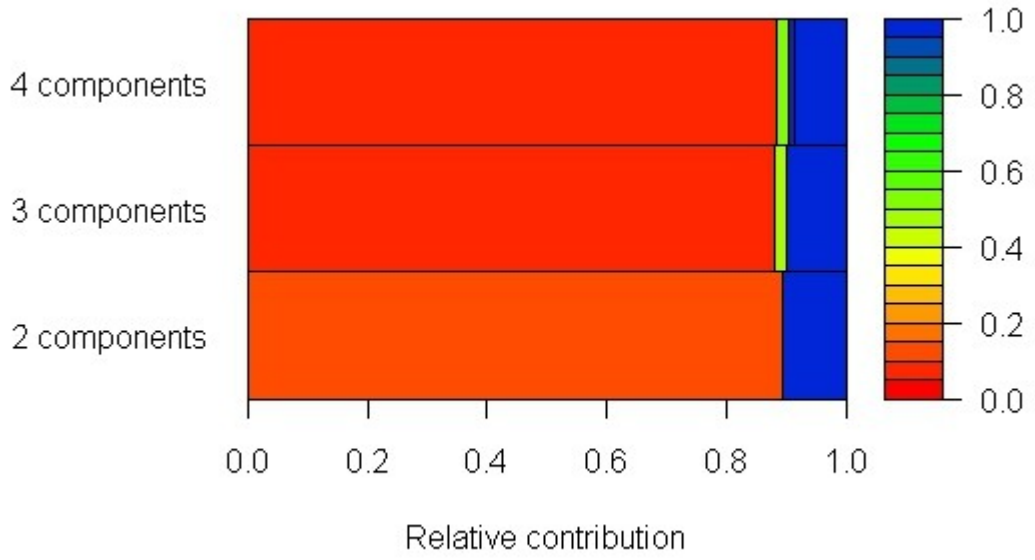
I likhet med Snipen et al. (2009) ble det også her brukt en bootstrap-metode for å si noe om usikkerheten til estimatene, som foreslått av (Kuhnert et al. 2008).



Figur 3: Fordelingen av genfamilier i genomer simulert med mixture model; hvor mange genfamilier som er tilstede i 1,2,...,10 simulerte genomer.

For å illustrere estimering av pangenom og kjernegenom ble det simulert et datasett med 10 genomer, mixture model med tre komponenter; deteksjonssannsynligheter på 0.10, 0.5, og 1.0 med mixture-proportions på henholdsvis 0.88, 0.02, og 0.1. Pangenom-størrelsen på utvalget ble satt til 10 000. Figur 3 viser hvor mange genfamilier som ble simulert i ett, to, ..., ti genomer. Ved estimering av pangenom og kjernegenom for genomene blir det tilpasset en mixture model etter disse dataene, det vil si at man prøver å finne binomiske tettheter som ligner på fordelingen til data i flere områder av diagrammet. Man prøver å predikere hvordan fortsettelsen av diagrammet i begge ender vil se ut på bakgrunn av de binomiske fordelingene som passer best med dataene. En tenkt søyle til venstre for søylen som representerer genfamiliene som fins i ett genom, vil her være et estimat av antall genfamilier som ennå ikke er observert, altså y_0.

Resultatene fra estimering ved mixture model fremstilles blant annet som en figur som viser forholdene mellom deteksjonssannsynlighet og mixture proportion (bidrag) i modellen. Det ble estimert modeller med 2 til 8 komponenter, og hovedfokuset har vært på de to artene *E. faecalis* og *E. faecium*. Under er en eksempelfigur (figur 4) basert på de simulerte(forklart over) genomene.



Figur 4: Eksempel på framstilling av resultater fra estimering av pangenom og kjernegenom med mixture model. De fargede komponentene for hver modell illustrerer deteksjonssannsynlighet(farge) og mixture proportion(bredde) for hver komponent i modellen.

En annen, og enkel metode for å si noe om et pangenom, er å beregne noe som kalles Chao's lower bound. Dette estimerer y_0 ut fra antall genfamilier som finns i hhv ett og to genomer, altså y_1 og y_2.

$$\hat{\eta} = n + y_1^2/(2y_2) \quad (7)$$

Dette estimatet er ment å gi en indikasjon på hvor stort pangenomet minst må være. Som navnet indikerer er dette et estimat for den nedre grensen til pangenomet. Det ble beregnet Chao's lower bound estimat for *E. faecalis* og *E. faecium*.

Det ble også laget et plot som viser den kumulative summen av forskjellige genfamilier etterhvert som det legges til genomer. Dette ble gjort ved å telle opp antall genfamilier tilstede i første genom, gå til neste genom og legge til de genfamiliene som ikke fantes i første genom, og så videre helt til alle genomene var telt opp. Plottet vil til slutt ende opp på det som er pangenom-størrelsen til utvalget. Genomene ble sortert etter art for å kunne se effekten av å gå fra en art til en annen. I denne sammenhengen er dette den mest interessante observasjonen å gjøre, da det ikke foreligger noen fornuftig rekkefølge å bruke for genomene innenfor hver art. Størrelsen som diagrammet ender på til slutt vil være pangenom-størrelsen for hele utvalget av genomer som er brukt i dette prosjektet, men det er ikke rimelig å anta at de genomene som utgjør utvalget er et representativt utvalg for *Enterococcus*, da det finnes en rekke flere *Enterococcus*-arter som ikke er med i denne helgenom-studien. Derfor er ikke pangenom-størrelsen for kun disse artene egnet til å si noe særlig om *Enterococcus* som gruppe.

2.5 Genomic fluidity

Som et alternativ til å estimere pangenom-størrelser, har det blitt lansert å bruke en parameter som kalles genomic fluidity, som skal kunne fungere som et slags mål på genomisk diversitet innenfor en gruppe genomer. Genomic fluidity angir hvor mye ulikhet det sannsynligvis er mellom to genomer innenfor et utvalg eller populasjon av genomer og baserer seg på den gjennomsnittlige forskjellen mellom genomer fra samme populasjon (Kislyuk et al. 2011).

Dersom man antar at genomer trekkes tilfeldig fra en populasjon, og at de utgjør et representativt utvalg, kreves det få genomer for å gi et godt estimat av genomic fluidity.
Selve parameteren er definert ved:

$$\Phi = \frac{2}{N(N-1)} \sum_{k,l=1}^{N} \left(\frac{U_k + U_l}{M_k + M_l} \right) \quad (8)$$

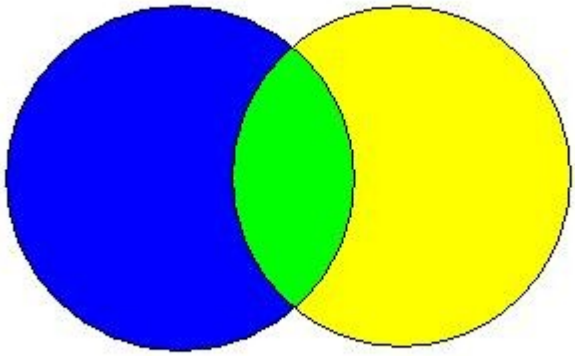
Der N er antall genomer, enten totalt i populasjonen eller i utvalget. Når det baseres på et utvalg blir dette som et estimat for populasjonen å regne. U_k og U_l er genfamilier som er unike respektivt for genom k og l, mens M_k og M_l er det totale antallet genfamilier respektivt i genom k og l. Φ angir den forventede ulikheten mellom to genomer fra samme populasjon. Uttrykket ble implementert i et R-script, *script_genomic_fluidity.R*, og dette ble brukt for å estimere genomic fluidity for henholdsvis *E. faecalis* og *E. faecium*. Det ble også estimert genomic fluidity for

artene sammen, for å se hvordan dette skiller seg fra artenes genomic fluidity, respektivt.

For å si noe om usikkerheten til estimatene ble varians estimert etter jackknife-prinsippet. Det vil si at det ble beregnet genomic fluidity når ett av genomene ikke er med. Alle genomene ble utelatt ett og ett, og det ble beregnet en genomic fluidity der hver av dem var utelatt. For hver av disse ble forskjellen til den opprinnelige genomic fluidity beregnet, og variansen ble estimert ved:

$$(\hat{\sigma})^2 = \hat{var}[\hat{\phi}] = \frac{(N-1)}{N}\sum_i (\hat{\phi}_i - \hat{\phi})^2 \quad (9)$$

Der $\hat{\phi}_i$ er estimert genomic fluidity når genom i er utelatt, $\hat{\phi}$ er estimert genomic fluidity. Dette er altså et estimat basert på leave-one-out statistikk.



Figur 5: De to sirklene illustrerer genomer. Det grønne området representerer det som er felles i begge genomene, mens gult og blått representerer det som er unikt i hvert genom i forhold til det andre.

Figur 5 illustrerer det som beregnes for genomic fluidity. I formelen (8) er det en størrelse U og M for hvert genom. Hver av sirklene i figuren representerer et genom, der det grønne området inneholder de genfamiliene som finnes i begge genomene, mens det gule og det blå området inneholder de genfamiliene som kun finnes i det respektive genom, og ikke i det andre. For hvert av genomene tilsvarer U antallet gener i det gule eller det blå området respektivt, mens M er antallet gener i det grønne feltet i tillegg til antall gener i det gule eller det blå feltet. Genomic fluidity er det gjennomsnittlige forholdet mellom summen av det to genomer ikke har til felles,

altså 'blå' + 'gul' og summen av alle gener i de to genomene, som tilsvarer (blå + grønn) + (gul + grønn). φ er forholdet (gul + blå) / ((gul + grønn) + (blå + grønn)), og genomic fluidity, Φ, er gjennomsnittet av φ for alle tilfeldige genompar i populasjonen.

2.6 COG-klassifisering

For å finne ut mer om kjernegenomet og de unike genene, ble det gjort et lokalt BLAST-søk mot COG-databasen for å finne hvilke funksjonelle grupper genene innenfor kjernegenomet og de unike *E.faecalis*-genene tilhører. COG står for Cluster of Orthologous Groups, og med ortologer menes det homologe gener, som antas å ha oppstått ved at et felles opphav (organisme) har utviklet seg i flere retninger. Resultatet er to litt forskjellige utgaver av samme gen. Genene er altså ikke helt like, men de har samme funksjon i hver sin respektive organisme (Fang et al. 2010).

COG-databasen ble satt opp som BLAST-database. Det ble laget én fasta-formatert fil med sekvensene til de unike kjernegenene til *E. faecalis,* og én fasta-formatert fil med alle kjerne-genene til *E. faecalis.* Gensekvensene ble hentet ut fra v583-stammen, og de ble BLASTet mot COG-databasen for å finne ut hvilke funksjonelle grupper genene tilhører, både kjernegener og unike kjernegener. Det ble satt en øvre E-verdi-grense på $1x10^{-5}$, og de beste treffene for hver kombinasjon av gen og funksjonell gruppe ble tatt med videre. Deretter ble det telt opp antall gener innenfor hver gruppe, og her bør det legges til at enkelte gener falt under flere grupper, som regel to men flere forekom også. Disse genene ble dermed telt opp like mange ganger som antall funksjonelle grupper de tilhører.

Det er ønskelig å finne ut om noen funksjonelle grupper er over- eller underrepresenterte blant de unike genene til *E. faecalis*, og derfor ble det benyttet en test kalt Fisher's exact test. Den beregner sannsynligheten for avvik fra nullhypotesen på en eksakt måte, siden testen ikke krever noen antakelser om fordeling av dataene. Testen brukes typisk for å avdekke signifikante forskjeller mellom forhold i to grupper (Routledge 2005). I dette tilfellet er det altså snakk om å finne forskjellige forhold i fordelingen av alle kjernegener og unike kjernegener innenfor funksjonelle grupper.

Fisher-testen gir en p-verdi, som her er sannsynligheten for våre data gitt at det er samme fordeling av genfamilier innen en funkjsonell gruppe for både kjernegener og unike kjernegener.

Testnivået ble satt til 5%, altså må p-verdier være under 0,05 for at man kan si at forskjellen er signifikant. I tillegg gir testen et forholdstall, som er et maximum-likelihood estimat, kalt odds ratio. Dette tallet viser hvilken vei forholdet er forskjøvet. Er odds raten lavere enn 1, er den funksjonelle gruppa underrepresentert blant de unike genene, mens den er overrepresentert blant de unike genene dersom odds ratio er større enn 1. Det ble gjort én test for hver funksjonell gruppe, der det ble testet om antall genfamilier tilhørende gruppa er ulikt fordelt blant kjernegener og unike kjernegener.

Etter å ha undersøkt kjernegener og unike kjernegener hos *E. faecalis* i forhold til COG-klassifisering, ble den samme undersøkelsen gjennomført for *E. faecium.* Stammen det ble tatt utgangspunkt i, var DO-stammen med prosjekt ID 30627.

Kapittel 3

Resultater

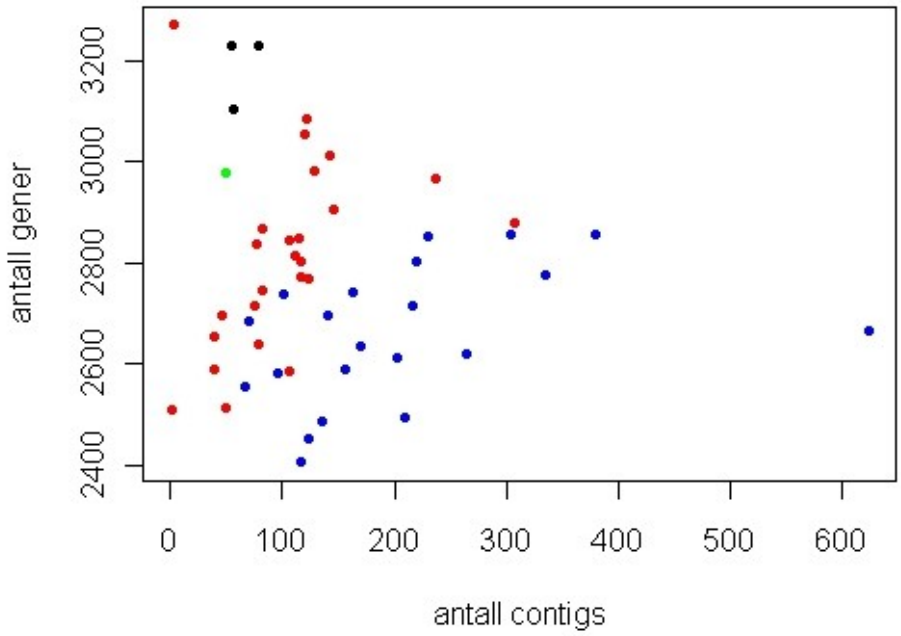
3.1 Deskriptiv statistikk

Tabell 1 inneholder deskriptiv statistikk for alle genomene som er brukt i denne oppgaven. Det som er tatt med i tabellen er antall contigs genomet består av, antall basepar for å gi et mål på størrelsen til genomet, andel GC-nukleotider og antall gener predikert ved bruk av Prodigal.

Tabell 1: Deskriptiv statistikk(antall contigs, antall basepar(bp), andel GC-nukleotider(gc) og antall predikerte gener(gener)) for 50 Enterococcus-genomer; 25 E. faecalis, 21 E. faecium, 3 E. casseliflavus og 1 E. gallinarum.

genom	contigs	bp	gc	gener
E_casseliflavus_EC10	54	3423271	0.425	3228
E_casseliflavus_EC20	57	3392503	0.428	3104
E_casseliflavus_EC30	79	3415886	0.425	3228
E_faecalis_ARO1DG	39	2821090	0.376	2655
E_faecalis_ATCC29200	123	2936062	0.375	2770
E_faecalis_ATCC4200	83	3008853	0.374	2867
E_faecalis_CH188	120	3158914	0.371	3056
E_faecalis_D6	46	2886935	0.375	2695
E_faecalis_DS5	128	3128175	0.371	2980
E_faecalis_E1Sol	75	2853152	0.375	2715
E_faecalis_Fly1	106	2790650	0.374	2584
E_faecalis_HH22	307	3049823	0.372	2879
E_faecalis_HIP11704	143	3129968	0.371	3012
E_faecalis_JH1	112	2994535	0.373	2816
E_faecalis_Merz96	106	3037892	0.376	2843
E_faecalis_OG1RF	1	2739626	0.378	2510
E_faecalis_R712	116	2900639	0.373	2771
E_faecalis_S613	145	3042102	0.373	2905
E_faecalis_T1	82	2905630	0.375	2744
E_faecalis_T11	49	2729090	0.377	2515
E_faecalis_T2	121	3204818	0.370	3085
E_faecalis_T3	40	2783550	0.376	2591
E_faecalis_T8	115	2985410	0.373	2850
E_faecalis_TUSoD-Ef11	79	2800312	0.376	2639
E_faecalis_TX0104	237	3106827	0.373	2965
E_faecalis_TX1322	116	2929603	0.373	2804
E_faecalis_V583	4	3359975	0.374	3270
E_faecalis_X98	77	2909530	0.374	2836
E_faecium_1141733	101	2865115	0.381	2740
E_faecium_1230933	304	2951888	0.378	2855
E_faecium_1231408	379	2889449	0.377	2855
E_faecium_1231410	230	2943814	0.377	2853
E_faecium_1231501	140	2799376	0.379	2698
E_faecium_1231502	220	2926115	0.377	2804
E_faecium_C68	170	2726073	0.378	2636
E_faecium_Com12	67	2685403	0.381	2555
E_faecium_Com15	70	2771456	0.382	2684
E_faecium_D344SRF	203	2636262	0.379	2612
E_faecium_DO	623	2830149	0.378	2665
E_faecium_DO	163	2848381	0.379	2742
E_faecium_E1039	124	2503231	0.380	2452
E_faecium_E1071	96	2700771	0.379	2583
E_faecium_E1162	136	2635043	0.380	2488
E_faecium_E1636	210	2609495	0.378	2494
E_faecium_E1679	335	2874725	0.377	2777
E_faecium_E980	117	2447102	0.381	2405
E_faecium_TC_6	264	2705326	0.377	2621
E_faecium_TX1330	156	2721168	0.381	2591
E_faecium_U0317	217	2823309	0.377	2717
E_gallinarum_EG2	49	3134430	0.406	2979

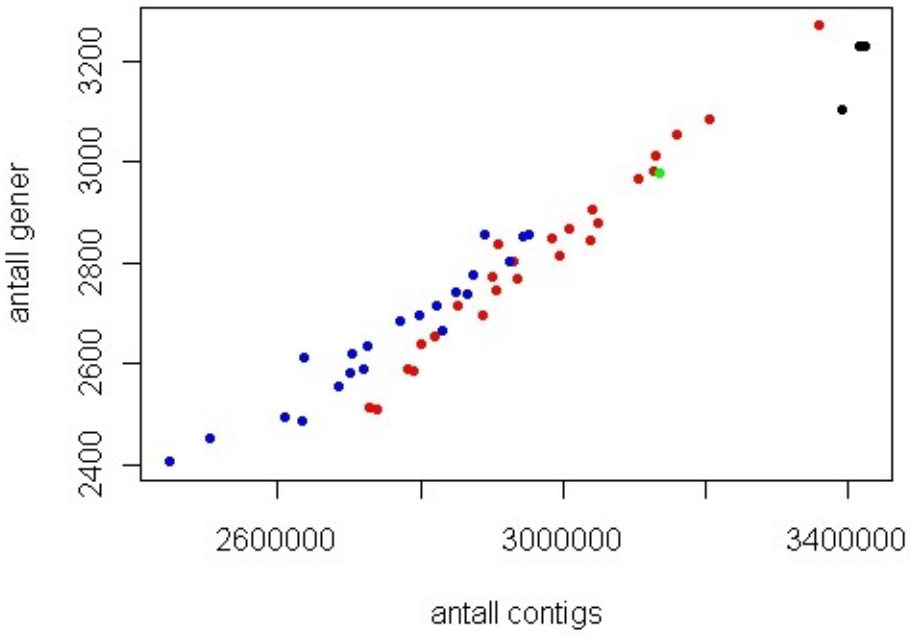
For å illustrere sammenhenger mellom ulike størrelser ble det generert plott.



Figur 6: Sammenhengen mellom antall contigs genomet består av, og antall gener funnet av Prodigal. Svart representerer E. casseliflavus, rød E. faecalis, blå E. faecium og grønn E. gallinarum. Utliggeren helt til høyre er E. faecium DO med prosjekt ID PID30627

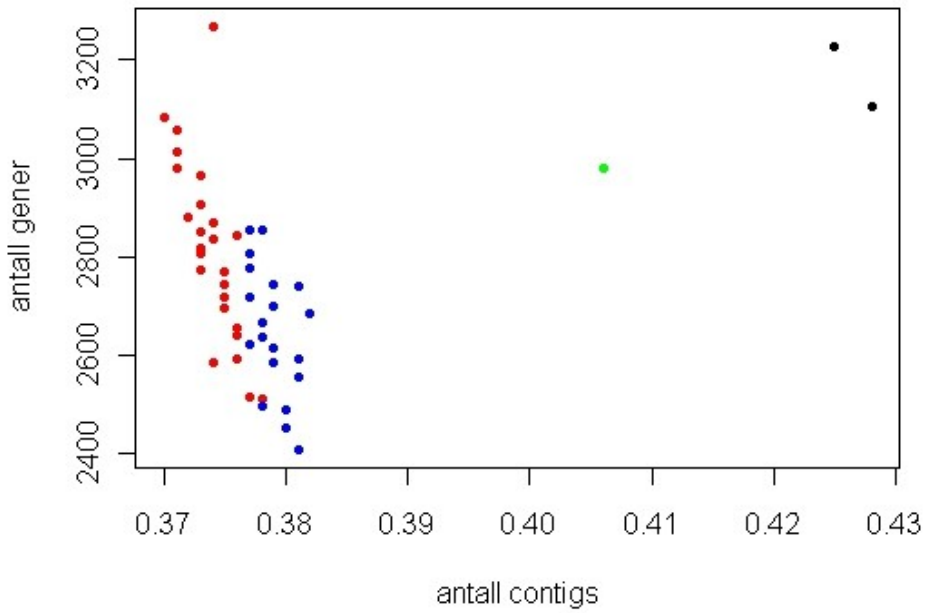
Figur 6 viser sammenhengen mellom antall contigs og antall predikerte gener. Men dataene er kanskje ikke varierte nok, i forhold til antall contigs, for å kunne si noe sikkert om hvor mange gener Prodigal finner. Genomene med flest gener består av færrest contigs, men det er også genomer som består av få contigs og likevel har få predikerte gener. Det eneste som ikke observeres er genomer med svært mange contigs og svært mange gener. Plottet tar ikke hensyn til at de ulike genomene varierer i størrelse (antall basepar).

Observerer et genom, *E. faecium* DO PID30627 (prosjekt ID), med svært mange contigs i forhold til de andre genomene. Allikevel skiller ikke dette genomet seg ut i forhold til antall gener som er predikert, sammenlignet med resten av stammene innen arten.



Figur 7: Sammenhengen mellom antall gener funnet med Prodigal og antall basepar i genomet. Rød er E. faecalis, blå E. faecium, svart E. casseliflavus og grønn E. gallinarum.

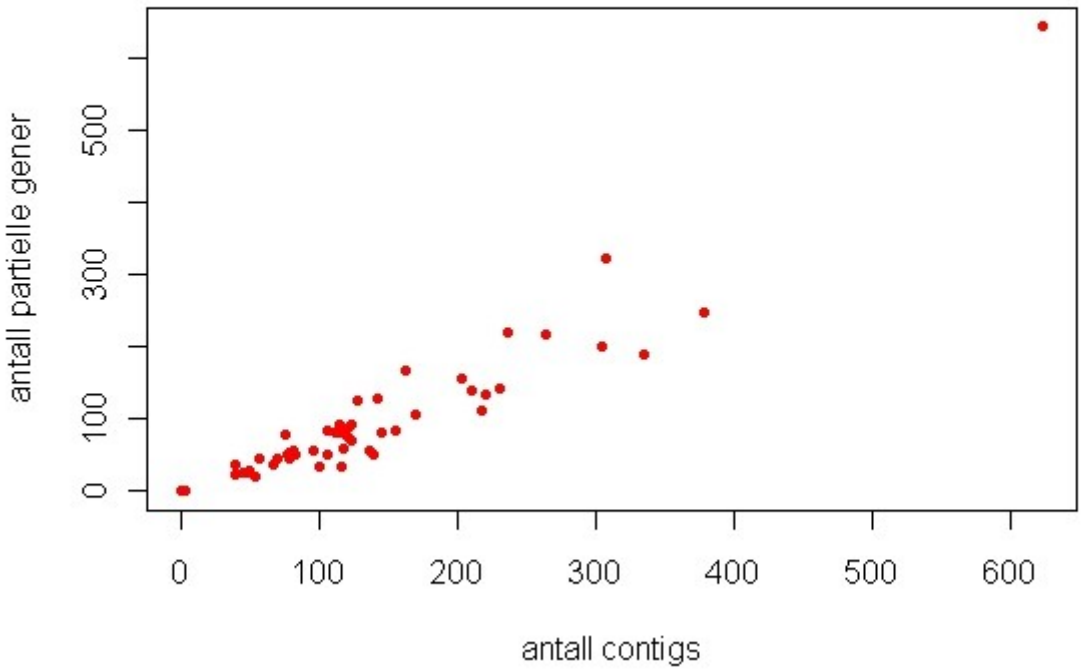
Figur 7 viser en klar sammenheng mellom antall basepar, altså den fysiske størrelsen på genomet og antall gener som er predikert av Prodigal. Bakteriegenomer er konsentrerte, og store genomer vil stort sett bestå av flere gener enn små genomer. Vi observerer også en viss ulik spredning i størrelse blant genomene. Rødt representerer *E. faecalis*, blått *E. faecium*, svart *E. casseliflavus* og grønt *E. gallinarum*.



Figur 8: Sammenhengen mellom andel GC-nukleotider i genomet, og antall gener funnet med Prodigal. De tre, faktisk fire, utliggerne er de tre E. casseliflavus (svart) og det ene E. gallinarum (grønn). To av E. casseliflavus er svært like i denne sammenhengen, og vises som én prikk. Rødt representerer E. faecalis og blått E. faecium.

Ut fra figur 8 er det ikke lett å si om det er noen sammenheng mellom antall gener og andel GC-nukleotider i genomene. Men det er midlertidig lett å se at de fleste genomene har en andel GC-nukleotider mellom 0.38 og 0.37. Ut fra tabell 1 kan vi se at det er *E. gallinarum* og de tre *E. casseliflavus*-genomene som er utliggerne langt til høyre. To av genomene til *E. casseliflavus* har lik GC-andel (0.425) og likt antall gener, og disse to genomene vises kun som én prikk i plottet.

Vi observerer også at det er en viss forskjell mellom artene med tanke på andel GC-nukleotider, og dette ser vi ut fra fargene de ulike artene er representert med i plottet: *E. faecalis* er rød, *E. faecium* er blå, *E. casseliflavus* er svart og *E. gallinarum* er grønn.

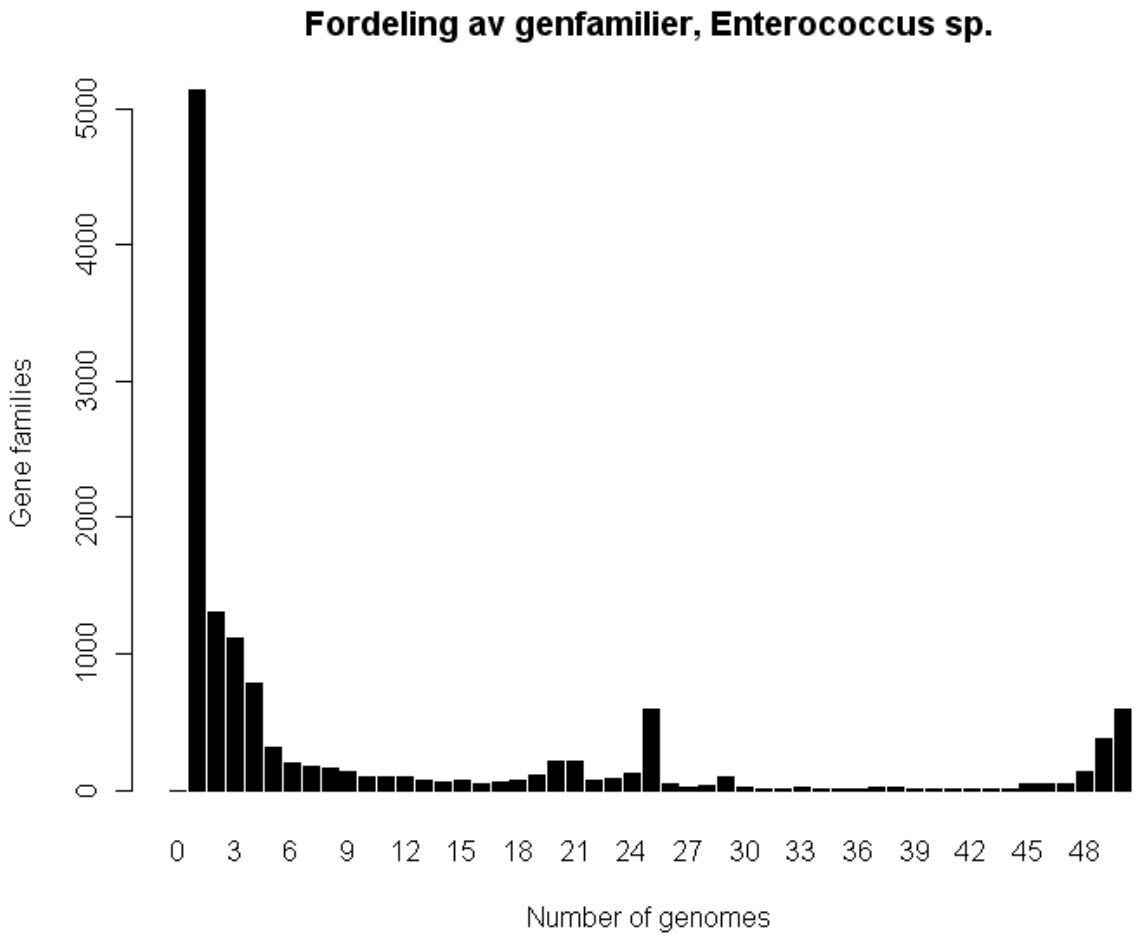


Figur 9: Sammenhengen mellom antall contigs og antall partielle gener predikert av Prodigal. Den tidligere nevnte utliggeren, E. faecium DO PID30627 ser vi her igjen lengst oppe til høyre.

Siden genomene i hovedsak foreligger som contigs, altså delsekvenser, blir det predikert en del ufullstendige gener, også kalt partielle gener. Som figur 9 viser, er det en klar sammenheng mellom antall contigs og antall partielle gener. Dette er forsåvidt forventet, da man regner med å finne flere ufullstendige gener jo mer genomet er oppdelt.

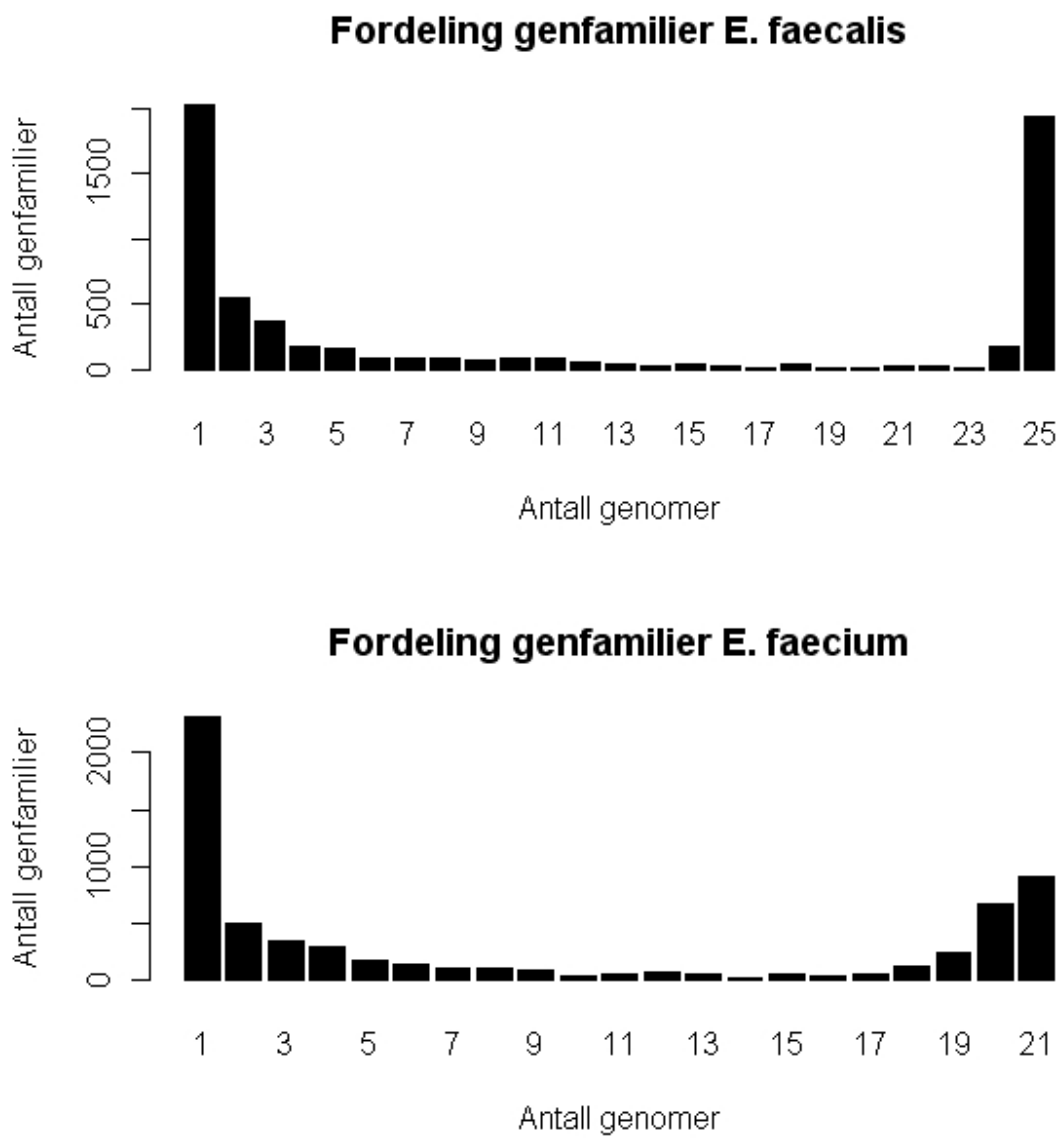
Observerer at genomet(*E. faecium* DO PID30627) som består av svært mange contigs, også er det genomet der det predikeres flest partielle gener. Denne observasjonen avviker ikke i forhold til tendensen blant resten av observasjonene. Men likevel kan dette sies å være en noe ekstrem observasjon.

3.2 Genfamilier



Figur 10: Fordeling av genfamilier. Diagrammet viser hvor mange genfamilier som har medlemmer i ett, to, ..., femti genomer.

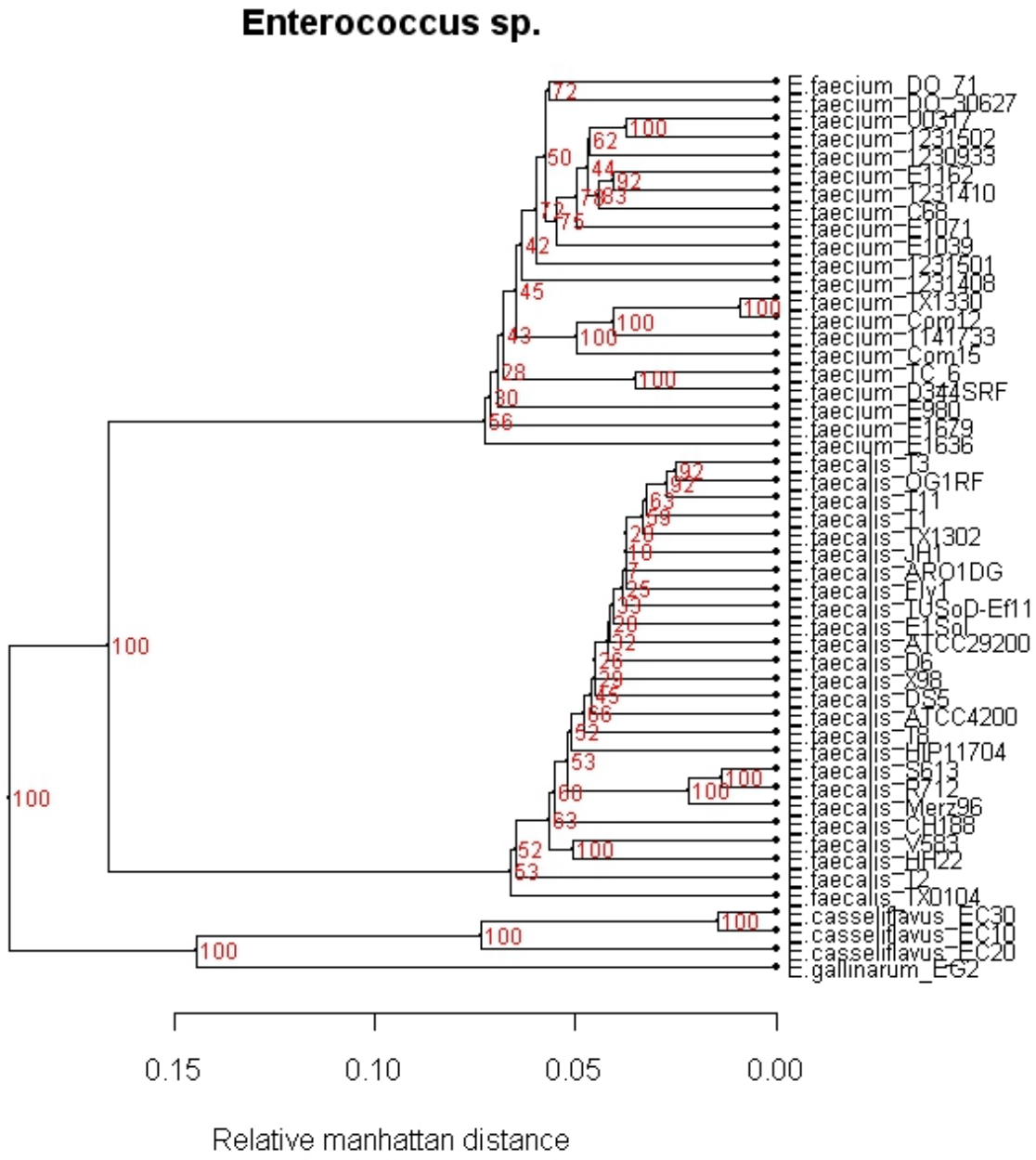
Figur 10 over viser fordelingen av genfamilier som er representert i ett, to, ..., femti genomer. I diagrammet observeres det en liten topp for genfamilier som finnes i 25 stammer. Dette skyldes blant annet de unike kjernegenene til *E. faecalis*. En topp ved 21 svarer til de unike kjernegenene til *E. faecium*. Vi ser også et par topper helt i enden, som skyldes genfamilier som finnes i nesten alle, og alle stammene. Dette er høyt konserverte gener, og man kan anta at de fleste står for viktige funksjoner i organismene. De kan tenkes å ha fulgt med organismen helt fra den oppstod, og en del av disse genfamiliene som finnes i alle stammene utgjør kjernegenomet til alle *Enterococcer.*



Figur 11: Fordelingen av genfamilier innenfor artene E. faecalis og E. faecium. Stolpehøyden tilsvarer hvor mange genfamiler som er representert i det korresponderende antall genomer.

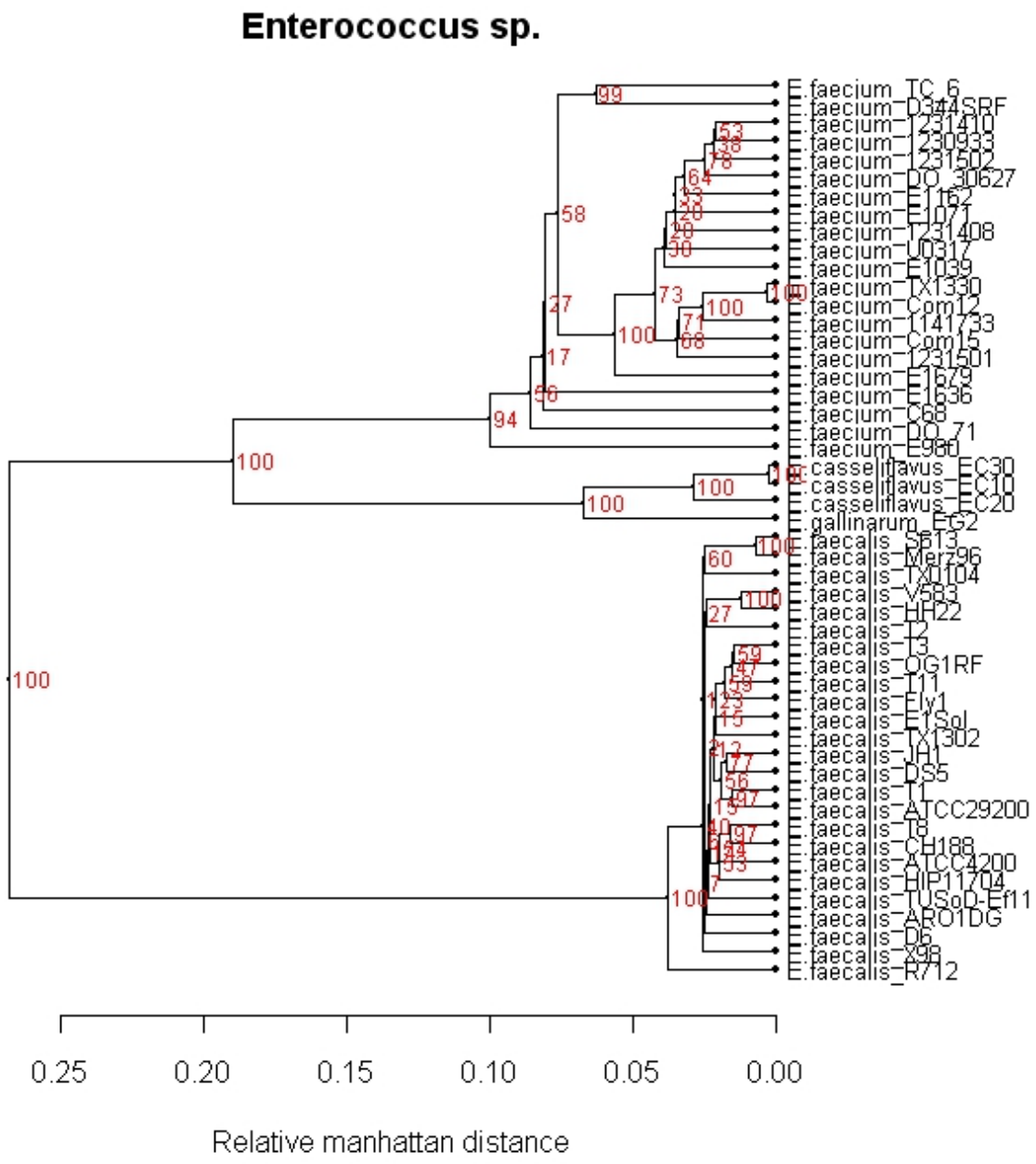
Stolpediagrammene i figur 11 viser fordelingen av genfamilier i henholdvis *E. faecalis* og *E. faecium*. Pangenomet til utvalget av begge artene ser relativt likt ut, men kjernegenomet til *E. faecium* virker å være mindre enn for *E. faecalis*. Riktignok er det hos *E. faecium* et høyere antall genfamiler som finnes i nesten alle stammene, og det kan tenkes at mange contigs og dermed flere partielle gener er skyld i at kjernegenomet ser ut til å være mindre.

3.3 Pangenom-trær



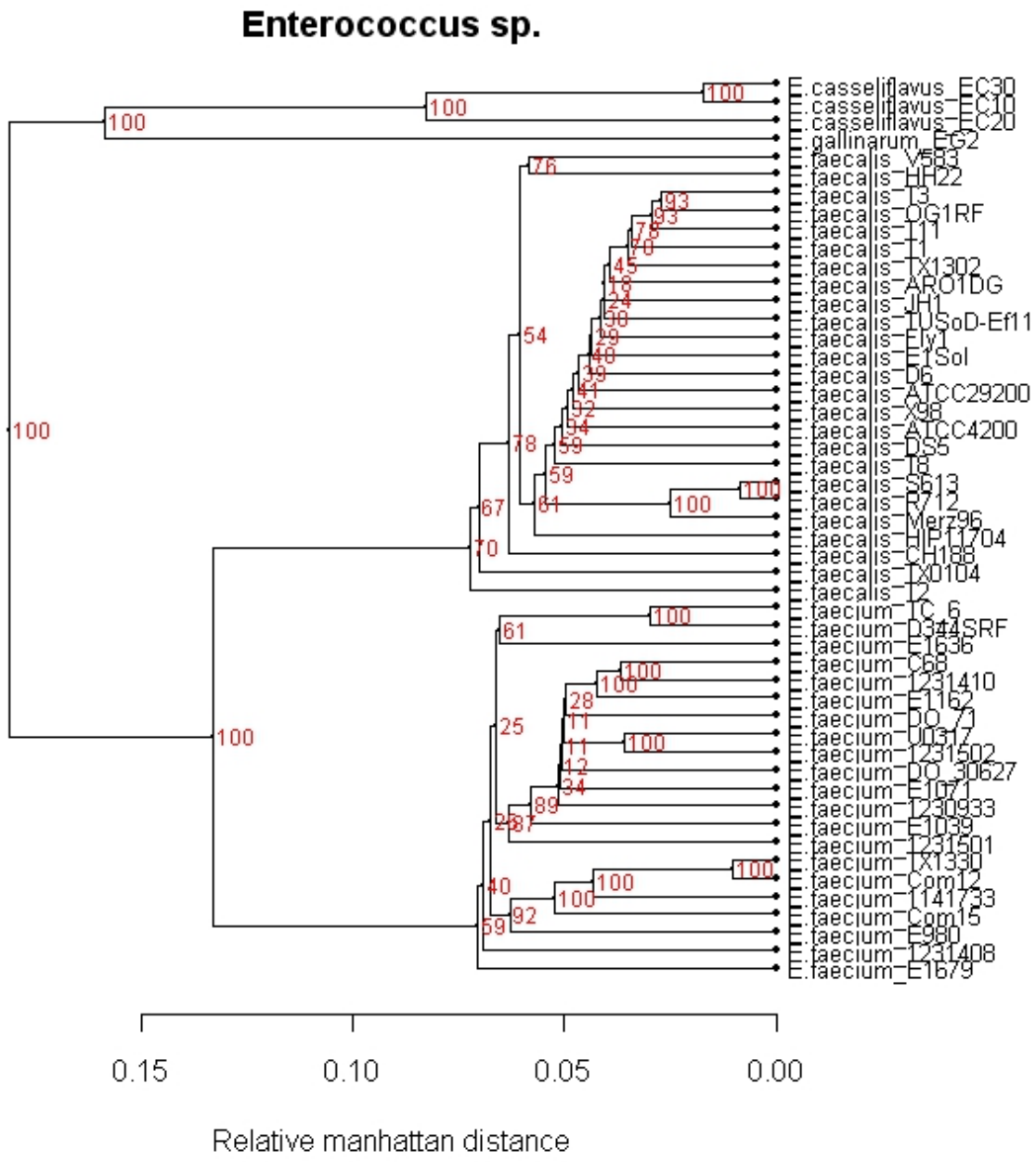
Figur 12: Pangenom-tre for fire Enterococcus-arter generert med flat vekting.

Pangenom-tre i figur 12 er generert med flat vekting. Dette betyr at det legges like mye vekt på alle genfamilier, uansett om de opptrer sjeldent eller ofte. Vi observerer i figuren at artene skiller seg klart fra hverandre. Treet viser også at det er en varierende grad av likhet mellom genomer innenfor hver art.



Figur 13: Pangenom-tre for fire Enterococcus-arter generert med shell-vekting.

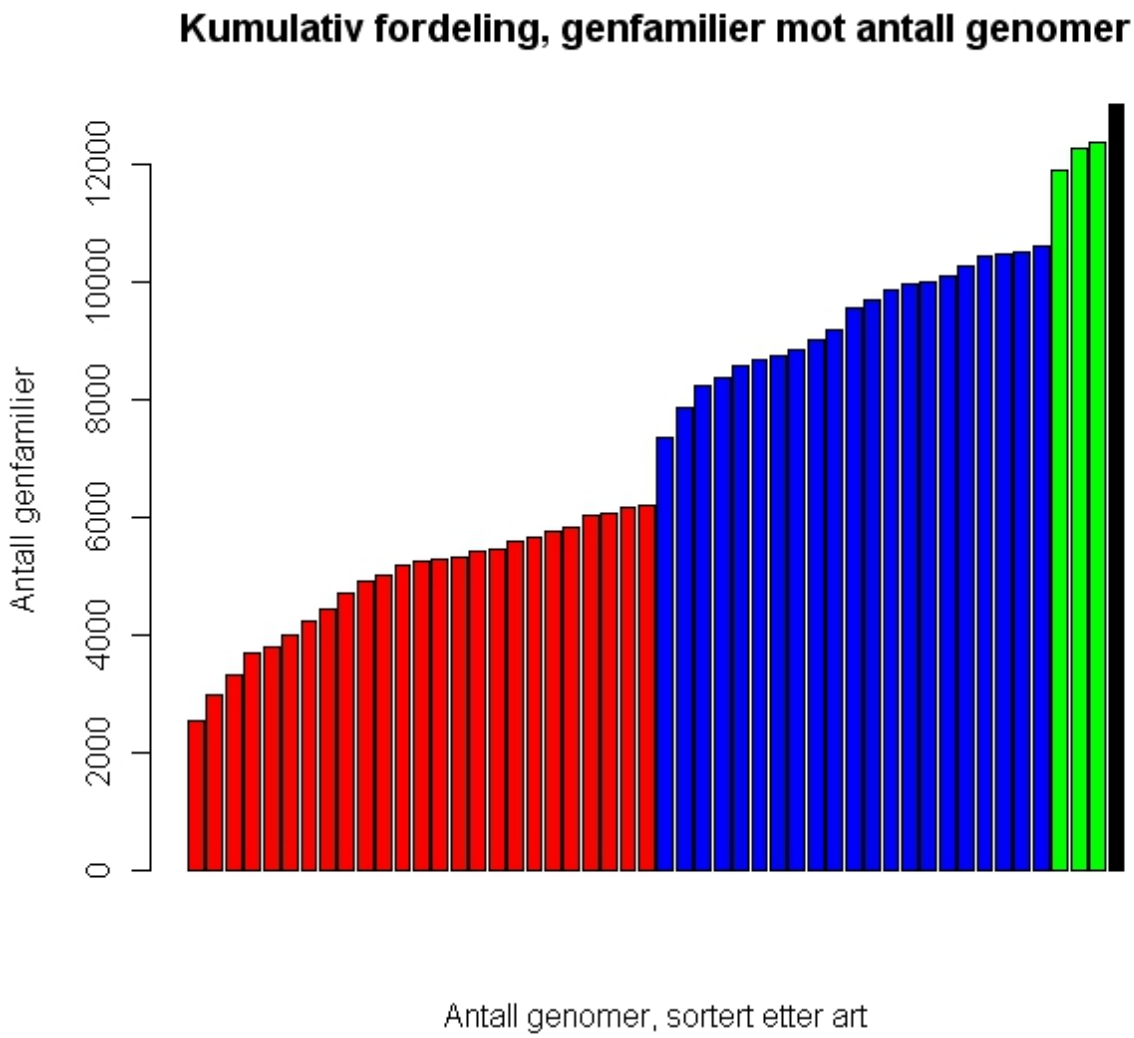
Pangenom-treet over (figur 13) er generert med *shell*-vekting. Det betyr at genfamiliene som er vanlige, og forventet å finne i de fleste genomene, vektlegges mer enn genfamilier som er sjeldne. Det oppstår en tydelig inndeling etter art, og det observeres ulik grad av likhet innen de forskjellige artene. Det er også tydelig at ulikheten er mindre innenfor *E. faecalis*, enn innenfor *E. faecium*. Dette tyder på at stammene innenfor *E. faecalis* muligens deler flere genfamilier og at de har mer til felles.



Figur 14: Pangenom-tre for fire Enterococcus-arter genenerert med cloud-vekting.

Pangenom-treet over (figur 14) er generert med vektlegging av sjeldne gener, såkalt *cloud*-vekting. Dette treet illustrerer ulikheten mellom genomene når de mer sjeldne genene blir vektlagt. Cloud-gener er gener som opptrer sjeldent, og det er rimelig at det observeres større avstander mellom genomene når det vektlegges det som kun er felles for ett eller få genomer. Selv om treet er generert med cloud-vekting, dannes det fortsatt en klar inndeling etter art. Dette tyder på at sjeldne gener også er artsbundet. Observerer også tre par genomer der forskjellen er svært liten.

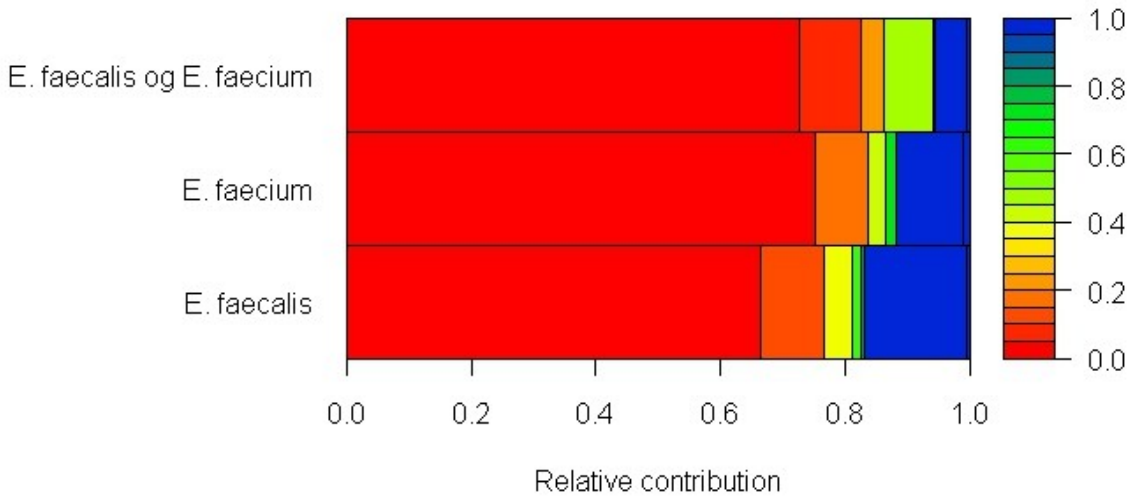
3.4 Kjernegenom og pangenom



Figur 15: Kumulativ fordeling av genfamilier etterhvert som nye genomer blir lagt til. E. faecalis er rød, E. faecium er blå, E. casseliflavus er grønn og E. gallinarum er svart.

Diagrammet i figur 15 viser hvordan det totale antallet forskjellige genfamilier øker etterhvert som antall genomer øker. De ulike artene er angitt med farge, som gjør det lettere å se hvordan antallet forskjellige genfamilier øker ved å legge til en ny art. Stolpene representerer altså ikke antall genfamilier i enkelte stammer, men det kumulative antallet forskjellige genfamilier etterhvert som flere genomer inkluderes.

3.4.1 Mixture model estimering av pangenom og kjernegenom



Figur 16: Mixture model plot som illustrerer den beste (minimum BIC) binomiske mixture modellen for hver art, i tillegg til de to artene kombinert. Beste modell for E. faecalis har sju komponenter, beste modell for E. faecium har seks komponenter og beste modell for kombinasjonen har sju komponenter. Fargen på elementene angir deteksjonssannsynlighet, mens bredden represesenterer bidrag(mixture proportion) til modellen.

Figur 16 illustrerer den beste modellen for hver av de to *Enterococcus*-artene, i tillegg til en modell for artene sammen. Bredden på de horisontale elementene representerer mixing proportion av deteksjonssannsynlighet som er angitt med farge. De røde elementene svarer til områder av pangenomet med lav deteksjonssannsynlighet, altså gener som forekommer sjelden. De områdene i den blå enden av skalaen representerer konserverte gener, som er felles for de fleste genomene.

Observerer at når artene kombineres, blir elementet som svarer til gener med høy deteksjonssannsynlighet mindre, og området som tilsvarer lav deteksjonssannsynlighet desto større. Når arter blandes, vil det naturlig være mindre som er felles for alle.

De blå elementene er mindre for *E. faecium* enn for *E. faecalis*, og dette indikerer at genomene

innen *E. faecalis* har mer til felles med hverandre enn genomene innen *E. faecium*, og at variasjonen mellom genomene innen *E. faecium* er større.

3.4.2 E. faecalis, mixture model

Tabell 2: Resultater fra mixture model estimering av pan- og kjernegenom, E. faecalis. Den beste(minimum BIC) mixture-modellen er den med sju komponenter.

	E. faecalis			
	Core.size	Pan.size	BIC	LogLikelihood
2 components	1926	6238	53885.23	-26929.52
3 components	1929	7030	30863.26	-15409.80
4 components	1903	7717	26447.33	-13193.09
5 components	1821	9461	25772.21	-12846.80
6 components	166	16269	25457.42	-12680.68
7 components	74	12529	25432.89	-12659.68
8 components	0	12543	25451.24	-12660.12
Sample	1930	6210	NA	NA

Observerer i tabell 2 at når man går fra modell med fem komponenter til en med seks komponenter, blir kjernegenomet plutselig mye mindre. Dette skyldes at det eneste komponentet med høy (1,0) deteksjonssannsynlighet blir delt opp i to komponenter, der det nye komponentet med deteksjonssannsynlighet 1,0 blir svært lite. Dermed estimeres kjernegenomet til å bli mye mindre.

Tabell 3: Resultat av bootstrap-estimering av pan- og kjernegenom, 100 bootstrappinger, E. faecalis

		E. faecalis		
mean core	var core	mean pan	var pan	antall komponenter
1890.86	2719.112	7702.83	12411.01	4
1236.99	576810.192	8969.32	986901.90	5
210.73	137626.300	11015.14	1812342.24	6
64.64	23797.667	11484.25	2300499.91	7

Det ble beregnet varians til både kjernegenom og pangenom-størrelse, og ser tydelig i tabell 3 at variansen øker og minker med de estimerte pangenom- og kjernegenom-størrelsene. For den beste modellen (sju komponenter) for *E. faecalis* er det estimerte standardavviket 1516,7 for pangenomet, og 154,3 for kjernegenomet.

Tabell 4: Estimert pangenom og kjernegenom ved bruk av ti E. faecalis genomer.

	E. faecalis, 10 genomer			
	Core.size	Pan.size	BIC	LogLikelihood
2 components	2091	5208	18561.29	-9267.90
3 components	2086	6873	15278.89	-7618.20
4 components	2064	7652	15122.82	-7531.66
5 components	411	7850	15127.74	-7525.62
6 components	1373	7927	15144.96	-7525.74
Sample	2090	4910	NA	NA

Tabell 4 inneholder resultater fra estimering av pangenom og kjernegenom ved å kun bruke ti genomer fra *E. faecalis*. Den beste modellen ble nå den med 4 komponenter, og ga estimater på 2064 og 7652 for henholdsvis kjernegenom og pangenom.

3.4.3 E. faecium, mixture model

Tabell 5: Resultater fra mixture model estimering av pan- og kjernegenom, E. faecium. Den beste(minimum BIC) mixture-modellen for E. faecium er den med seks komponenter.

	E. faecium			
	Core.size	Pan.size	BIC	LogLikelihood
2 components	904	6407	86089.56	-43031.63
3 components	836	6859	35156.02	-17556.09
4 components	623	8943	30095.60	-15017.12
5 components	352	12549	29070.02	-14495.56
6 components	182	16369	28947.20	-14425.39
7 components	192	16730	28964.39	-14425.23
8 components	0	16222	28978.40	-14423.46
Sample	908	6404	NA	NA

I tabell 5 ser vi resultatene fra estimering av mixture models for *E. faecium*. Observerer at BIC minimeres for modellen med 6 komponenter, og dette er per definisjon den beste modellen. Vi ser ikke her den samme dramatiske utviklingen for kjernegenomet etterhvert som modellen blir mer kompleks, og dette kan ha en sammenheng med at kjernegenomet for utvalget også er vesentlig mindre enn for *E. faecalis*.

Tabell 6: Resultater fra bootstrap-metode for estimering av pan- og kjernegenom, 100 bootstrappinger, E. faecium

E. faecium				
mean core	var core	mean pan	var pan	antall komponenter
618.05	2194.331	8936.08	105853.2	4
352.35	3395.078	12780	550169.3	5
152	14314.424	15688.58	3181571.8	6
86.74	13908.336	15705.84	5581762.3	7

Ser ut fra resultatene i tabell 6 at variansen til de bootstrappede-estimtaene for *E. faecium* gir et estimert standardavvik på 1783,7 for pangenomet og 119,6 for kjernegenomet.

Tabell 7: Resultater fra estimering av pangenom og kjernegenom ved å bruke ti genomer fra E. faecium.

	E. faecium, 10 genomer			
	Core.size	Pan.size	BIC	LogLikelihood
2 components	1523	4946	25611.36	-12792.94
3 components	1357	5339	19096.66	-9527.10
4 components	817	7686	18168.26	-9054.41
5 components	0	8377	18156.01	-9039.79
6 components	0	8399	18172.97	-9039.78
Sample	1525	4876	NA	NA

Tabell 7 viser resultatene fra estimering av pangenom og kjernegenom ved å bruke et utvalg på ti genomer fra *E. faecium*. Den beste modellen ut fra BIC-kriteriet er den med fem komponenter. Dette gir estimater på 0 og 8377 for henholdsvis pangenom og kjernegenom.

Det ble beregnet Chao's lower bound estimater for begge artene, og dette ga følgende:
E. faecalis: 9913
E. faecium: 11873

Dette betyr at pangenomene til henholdsvis *E. faecalis* og *E. faecium* forventes å være minst så store som de estimerte Chao's lower bound.

3.5 Genomic Fluidity

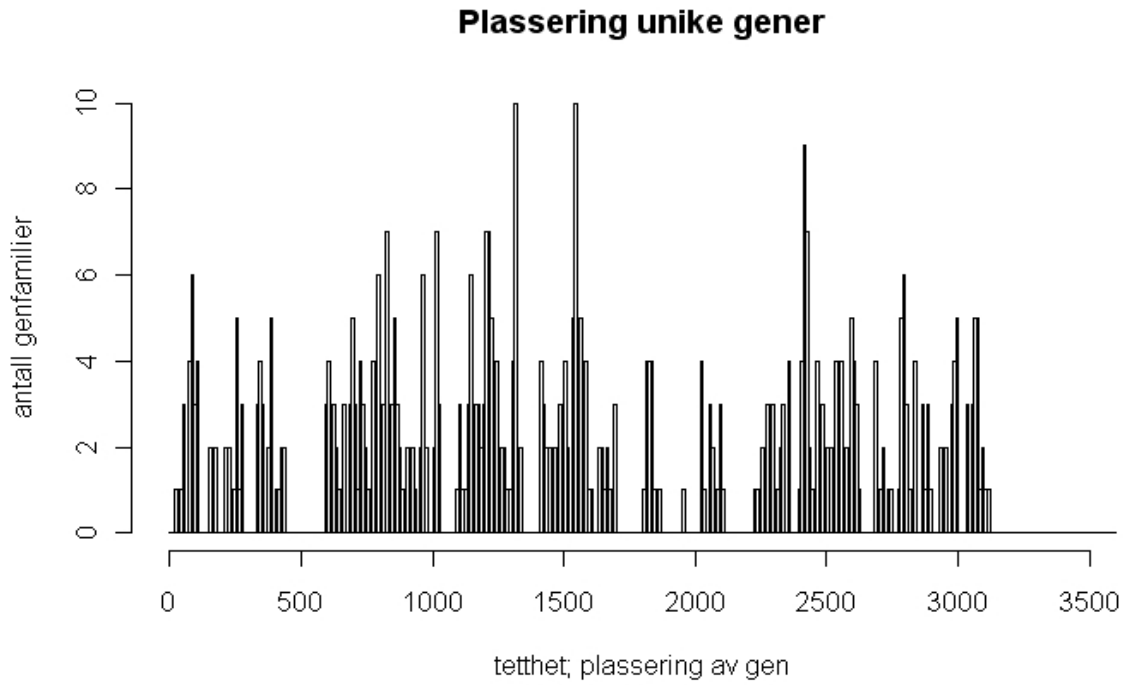
For utvalget av *E. faecalis*- og *E. faecium*-genomer ble det beregnet genomic fluidity og variansen til dette estimatet ble beregnet som beskrevet under metoder.

Tabell 8: Genomic fluidity med varians og standardavvik for E. faecalis, E. faecium og begge sammen.

Art	Genomic fluidity	Var	Standardavvik
E. faecalis	0,299	0,004	0,064
E. faecium	0,423	0,01	0,1
E. faecalis og E. faecium	0,679	0,011	0,104

Tabell 8 viser resultatene fra estimering av genomic fluidity for *E. faecalis*, *E. faecium* og begge artene sammen. Ser at innen *E. faecalis* er det forventet at genomene har en andel ulikhet på 0,299, mens det innenfor *E. faecium* er forventet at forskjellen mellom genomene er på 0,423. For en populasjon med både *E. faecalis* og *E. faecium*, vil den forventede andelen forskjell mellom genomer være på 0,679.

3.6 Unike gener



Figur 17: Tettheten av unike gener langs kromosomet til E. faecalis v583. Plasseringen tilsvarer gensekvensens nummer i genomet.

Det ble laget et histogram (figur 17) som viser tettheten av gener innenfor intervaller av genomet til *E. faecalis*. Dette illusterer hvor på kromosomet til v583-stammen de unike *E. faecalis*-genene er plassert i forhold til andre gener. Det er altså kun brukt gener som finnes i alle faecalis-stammene, og som ikke er tilstede i noen andre arter.

Observerer at det er enkelte områder langs kromosomet hvor vi ikke finner noen unike kjernegener. Disse samsvarer med definerte mobile genetiske elementer(se vedlegg)

Siden det ikke finnes noen komplette kromosomer fra *E. faecium* eller de andre artene, ble dette kun gjort for *E. faecalis*.

Tabell 9: Tabell med statistikk over genomene innenfor de ulike artene. Antall gener og basepar er gjennomsnittet for arten, basert på utvalget genomer.

Art	Antall stammer	Gener kun i art	Antall gener	Antall basepar
E. faecalis	25	515	2814	2967726
E. faecium	21	179	2658	2756841
E. casseliflavus	3	464	3187	3410553
E. gallinarum	1	650	2979	3134430

Tabell 9 over viser antall unike kjernegener, gjennomsnittlig antall gener i genomene og gjennomsnittlig størrelse på genomene, i form av antall basepar, for hver art. Ser umiddelbart at antallet unike kjernegener innen *E. faecalis* er høyt. Man skulle kanskje forvente å finne færre genfamilier som finnes i alle stammene av en art og i ingen andre arter, når man har mange stammer representert. Altså, flere stammer gjør det kanskje mindre sannsynlig å finne mange genfamilier som finnes i alle stammene, men som ikke finnes hos noen andre arter.

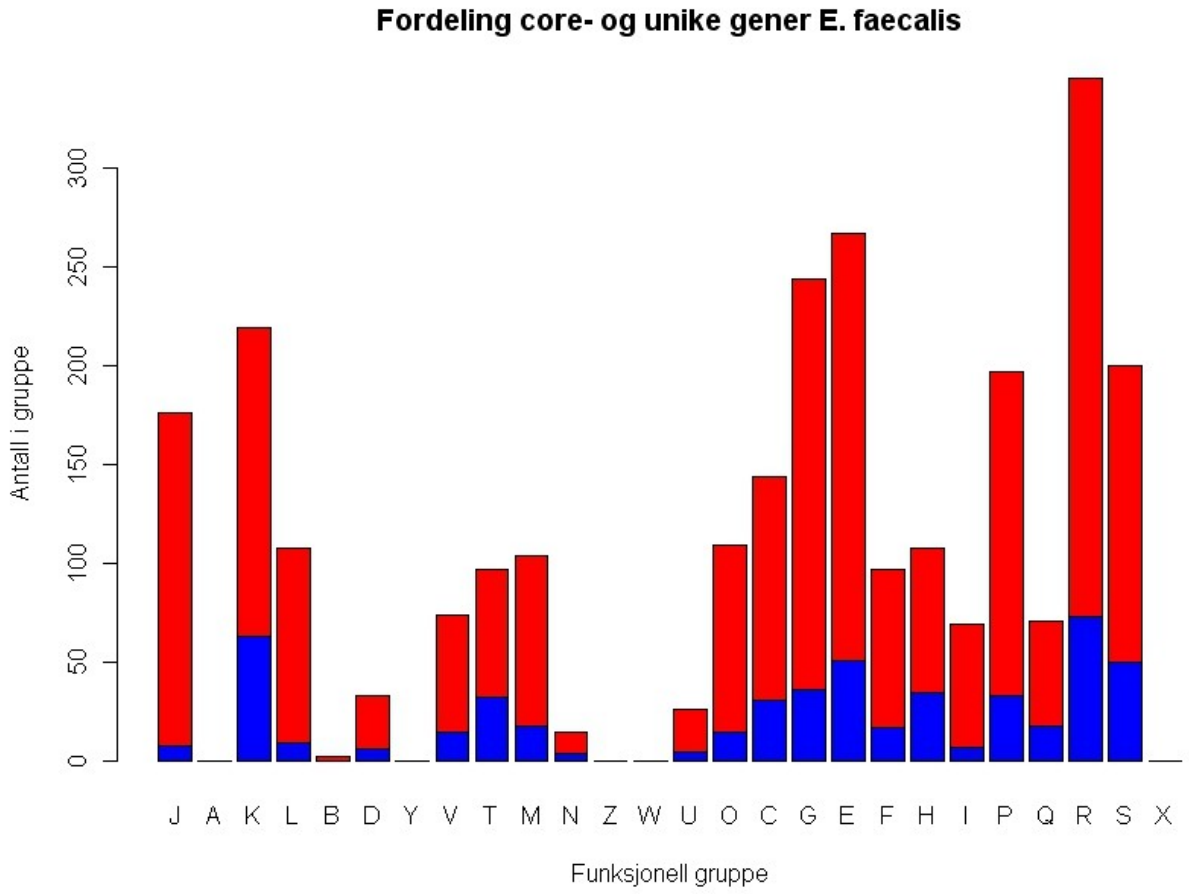
Det tilsynelatende mest overraskende er å finne så mange genfamilier som finnes i alle faecalis-stammene og i ingen andre arter, sammenlignet med det tilsvarende antallet hos *E. faecium*. Det er blitt brukt 25 *E. faecalis*-stammer og 21 *E. faecium*-stammer, og det ser ut til å være en forskjell som ikke kan forklares ut fra antall genomer.

Antallet unike genfamilier for henholdsvis *E. casseliflavus* og *E. gallinarum* er rimeligere med tanke på at det er brukt henholdsvis tre og ett genom.

E. casseliflavus ser ut til å ha et gjennomsnittlig større genom enn de andre artene artene, noe vi ser både av antall gener og antall basepar. Denne arten er kun representert ved tre stammer, så det kan være problematisk å generalisere for *E. casseliflavus* kun på bakgrunn av de tre genomene.

3.7 COG-klassifisering

3.7.1 E. faecalis



Figur 18: Fordelingen av genfamiler innen COG-grupper. Rødt representerer antallet blant alle kjernegener, mens blått representerer antallet blant de unike kjernegenene.

Figur 18 viser fordelingen av genfamilier fra *E. faecalis* i de forskjellige funksjonelle gruppene. Det røde området er antall kjernegener, mens den blå andelen er antall unike kjernegener som tilhører den funksjonelle gruppa.

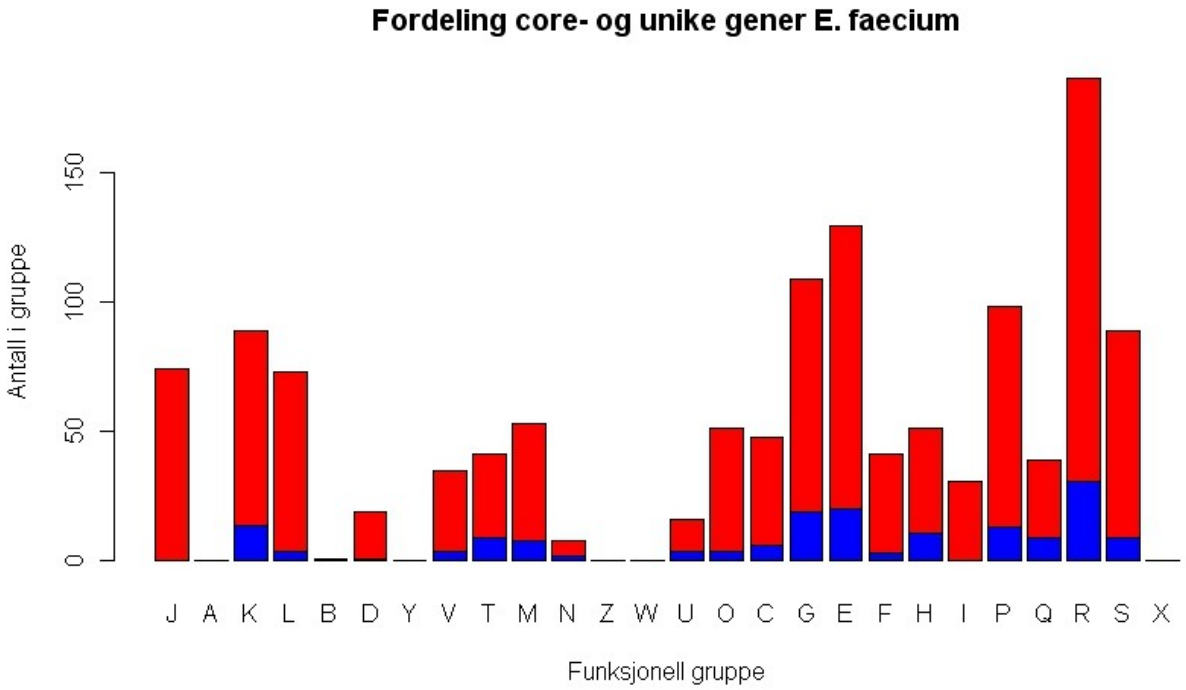
Det ble gjort en Fisher-test for å se om fordelingen var ulik blant de unike kjernegenene i forhold til alle kjernegenene. Resultatene fra denne testen er presentert i tabell 8 under:

Tabell 10: Signifikant forskjellig fordeling av genfamilier blant unike kjernegener mot alle kjernegener, funnet ved Fisher's exact test.

			E. faecalis
odds.ratio	p.verdi	FG	beskrivelse
0.185	5.49e-09	J	Translation, ribosomal structure and biogenesis
1.764	4.87e-04	K	Transcription
0.366	1.78e-03	L	Replication, recombination and repair
2.106	1.48e-03	T	Signal transduction mechanisms
2.056	1.11e-03	H	Coenzyme transport and metabolism
0.461	4.58e-02	I	Lipid transport and metabolism
1.421	4.14e-02	S	Function unknown

Tabell 10 inneholder de COG-gruppene der det var signifikant forskjellig fordeling blant unike kjernegener og alle kjernegener. P-verdien angir sannsynligheten for at den ulike fordelingen er observert ved en ren tilfeldighet. For alle p-verdier under 0.05, eller 5 %, ble forskjellen sagt å være signifikant. Der odds-ratio er større enn 1 er gruppa overrepresentert blant de unike kjernegenen, mens den er underrepresentert om odds-ratio er mindre enn 1. Ser at fire av gruppene er overrepresentert blant de unike kjernegenene, mens tre av gruppene er underrepresentert.

3.6.1 E. facium



Figur 19: Fordelingen av genfamiler innen COG-grupper. Rødt representerer antallet blant alle kjernegener, mens blått representerer antallet blant de unike kjernegenene.

Figur 19 viser fordelingen av unike kjernegener (blå) og alle kjernegener (rød) blant funksjonelle grupper for *E. faecium*.

Tabell 11: Signifikant forskjellig fordeling av genfamilier blant unike kjernegener mot alle kjernegener, funnet ved Fisher's exact test.

			E. faecium
odds.ratio	p.verdi	FG	beskrivelse
0.000	2.75e-05	J	Translation, ribosomal structure and biogenesis
0.362	4.90e-02	L	Replication, recombination and repair
0.000	2.70e-02	I	Lipid transport and metabolism

I tabell 11 ser vi at for to av gruppene er odds-ratio er 0, og dette kommer av at de aktuelle COG-gruppene ikke er representert blant de unike genene mens det finnes genfamilier blant de øvrige kjernegenene som tilhører disse COG-gruppene. For *E. faecium* er det ingen av de unike kjernegenene som er overrepresentert blant noen av de funksjonelle gruppene

Kapittel 4

Diskusjon

4.1 Sekvensdata - deskriptiv statistikk

Som man kan se i tabell 1, er det en del umiddelbare forskjeller mellom genomene, særlig med tanke på antall contigs. Dette kan gjenspeile hvor langt man har kommet med arbeidet med å sette sammen de aktuelle genomene til mer komplette sekvenser. Problemer som kan oppstå når det benyttes genomer som foreligger i svært mange contigs, er at det blir vanskeligere å predikere gener. Som et resultat vil man predikere en del partielle gener fordi det forekommer at bare deler av gener er med i delsekvensene som tilsammen utgjør genomet.

Plot over antall contigs mot antall gener (figur 6) funnet viser en vag sammenheng, som i alle fall sier noe om at man kan forvente å finne færre gener i et genom, når genomet er delt opp i mange contigs. Men selv om et genom består av få contigs, betyr ikke dette nødvendigvis at man finner flere gener. Det kan muligens sies at antall contigs setter en slags begrensning på hvor mange gener man vil kunne predikere, og at mange contigs begrenser evnen til å predikere gener.

Vi observerte at de fleste genomene består av mindre enn 250 contigs. Det er alltid ønskelig at genomene man jobber med er så komplette som mulig, da genomer i form av mange contigs kan gi problemer i sammenheng med prediksjon av gener. Plot over antall contigs mot antall partielle gener predikert (figur 9) viser at det predikeres flere partielle, eller delvise, gener når genomer består av mange contigs. For genomer som består av mange contigs er det ofte flere contigs som er små, og dermed vil antallet partielt predikerte gener øke.

Når det plottes antall basepar mot antall predikerte gener (figur 7) observeres det også en klar sammenheng. Vi ser at jo større et genom er, desto flere gener består det av. Bakteriegenomer er konsentrerte og består av lite DNA som ikke er kodende, og dette er dermed en sammenheng man forventer å se.

Når det gjelder plottet (figur 8) over andel GC-nukleotider og antall gener, viste ikke dette noen

spesielt klar sammenheng. Kodende DNA har ofte litt høyere andel GC-nukleotider, men hos bakterier består DNA-molekylet stort sett bare av kodende DNA. Ulik GC-andel kan kanskje heller tenkes å være karakteristisk for art. Selv om *E. casseliflavus* her bare er representert med 3 genomer, ser det ut til *E. casseliflavus* har en høyere andel GC-nukleotider enn de andre og dette er sannsynligvis noe som er karakteristisk for arten,

4.2 Genfamilier

Etter clustring og oppretting av panmatrise, ble det gjort en oppsummering av hvor mange genfamilier som finnes i én, to, ..., alle femti stammer, og tilsvarende for artene *E. faecalis* og *E. faecium*. Basert på dette ble det generert noen stolpediagram som viser tettheten av genfamilier for henholdsvis alle enterococcer som er brukt (figur 10), *E. faecalis* og *E. faecium* hver for seg (figur 11). I forbindelse med stolpediagrammene bør det nevnes at blant genfamiliene som bare finnes i én organisme kan man regne med noe støy som skyldes at genprediksjonsverktøyet Prodigal finner gener som ikke egentlig er gener, og som dermed ikke vil ligne på reelle gener i andre genomer. I tillegg finnes det også en del partielle gener som skyldes contigs. Om Prodigal finner to deler av samme gen, vil genet i praksis bli telt opp to ganger. Om de partielle genene er korte, vil de sannsynligvis ikke oppfylle 50-50-regelen, og dermed ikke bli gjenkjent som medlemmer av en genfamile. Da blir de, feilaktig, definert som en egen genfamilie.

De reelle genfamiliene som bare finnes i én stamme, kalles ofte for ORFans, en slags ordlek basert på ORF og "orphans", altså menes det foreldreløse gener. De er "forlatte" gener i den forstand at de ikke har noen åpenbare slektninger(homologer) innenfor arten, og ser ikke ut til å stamme fra et felles opphav.

Ut fra diagrammene i figur 11 kan det virke som om det er forskjell på de to artenes kjernegenom. Det er omtrent dobbelt så mange genfamilier som er representert i alle *E. faecalis*-genomene sammenlignet med *E. faecium*. I tillegg er *E. faecalis* representert med flere genomer enn *E. faecium*, og tatt i betraktning at det normalt forventes at kjernegenom blir mindre når antall genomer øker, ser dette ut til å være noe som karakteriserer forskjellen på disse to artene. Det kan være interessant å se på antall contigs som genomene til de to artene består av. Det kan kanskje se ut som om *E. faecium*-genomene generelt består av flere contigs enn *E. faecalis* (figur 6). Partielle gener kan muligens være en grunn til at færre av genene til *E. faecium*-genomene faller innenfor samme genfamilier.

4.3 Pangenom-trær

Det ble generert tre forskjellige pangenom-trær, basert på tre forskjellige måter å vekte genfamiliene. Det første treet (figur 12) har såkalt flat vekting, som betyr at det blir lagt like mye vekt på alle genfamilier. Ved å studere treet med flat vekting, gir det et inntrykk av hvor stor avstand, eller relativ ulikhet, det er mellom genomene når alle genene blir vektlagt likt. Dermed vil dette treet gi det mest oversiktlige inntrykket av hvor like alle genomene er i forhold til hverandre. Treet viser både varierende grad av likhet innenfor artene, samtidig som det gir en indikasjon på hvor like artene er hverandre.

Det andre treet (figur 13) ble generert med såkalt shell-vekting. Dette betyr at genfamiliene som forekommer ofte blant genomene blir lagt mer vekt på enn de sjeldne genfamiliene. Her ser vi at det er ulik grad av variasjon mellom genomene innenfor de ulike artene. Det kan se ut til at det er mindre variasjon mellom genomene innen *E. faecalis* enn innen *E. faecium*. Altså har stammene innen *E. faecalis* mer til felles med hverandre enn stammene tilhørende de andre artene. Dette kan igjen gi en indikasjon på at kjernegenomet til *E. faecalis* er større enn hos de andre artene, noe som stemmer godt overens med at det er observert langt flere kjernegener og unike kjernegener for *E. faecalis* enn for de andre artene, antall genomer tatt i betraktning.

Det tredje og siste treet (figur 14) som ble generert er basert på cloud-vekting. Med cloud-gener menes gener som forekommer sjeldent i en populasjon. Her ble naturlig nok avstandene mellom genomene større i gjennomsnitt, men allikevel er det genomer som også her ser ut til å ligge svært nærme hverandre. Selv om treet er basert på vekting av genfamilier som deles av få genomer innen artene, observerte vi allikevel en tydelig artsinndeling. Ved å legge vekt på sjeldne genfamilier, skulle man tro at det ikke ville oppstå en så tydelig inndeling etter art. Det virker som at en del av de sjeldne genfamiliene er artspesifikke.

Den lille avstanden vi observerte for enkelte stammer i cloud-treet kan kanskje skyldes at disse stammene deler en del spesielle egenskaper, som igjen kan ha oppstått i forbindelse med tilpasning til spesielle miljøer eller at stammene på en eller annen måte har blitt utsatt for mye av den samme påvirkningen fra miljø, eller fra andre organismer. Det kan også tenkes at dette er svært nære slektinger som nylig har skilt lag fra et felles opphav.

Det bør nevnes at databasen alle genomene er lastet ned fra, i sjeldne tilfeller inneholder to eller flere sekvenseringsprosjekter for samme organisme. Det kan også hende at genomer blir lagt inn i databasen med feil stammenavn, artsnavn og lignende. Dette bør helst ikke forekomme, men skjer allikevel. I tilfelle noen genomer er så like at man skulle tro de egentlig kommer fra samme stamme, kan det altså hende at de faktisk gjør det.

Før pangenom-trærne (presentert under resultater) ble generert, ble det først generert noen pangenom-trær som virket ganske urimelige, da det ikke oppstod noen inndeling etter art i noen av dem. Det er jo forventet at stammene innenfor hver art ligner mest på sin egen art, men dette var foreløpig ikke tilfelle. Alle genomene ble lastet ned og BLASTet på nytt, men dette ga ikke forskjellige resultater fra første gang. Det ble også forsøkt å lage nye trær basert på en pan-matrise der partielle gener ikke var med. Dette ga ingen spesielt annerledes resultater, og man kan altså ikke anta at de partielle genene har særlig stor innvirkning på pangenom-trærne.

Det ble observert en grov inndeling, men denne skilte tilsynelatende ikke mellom arter. Det som riktignok var interessant med oppdelingen som ble observert var antallet genomer i hver hovedforgrening; 21, 3, 1 og 25. Når det tas i betraktning at det er brukt 25 *E. faecalis*-, 21 *E. faecium*-, 3 *E. casseliflavus*- og ett *E. gallinarum*-genom syntes dette å stemme for bra til å være tilfeldig. Dermed virket det meget sannsynlig at det var en feil med hvordan navnene på genomene ble håndtert.

Kildekoden ble gjennomgått, og det viste seg å være en bug i en funksjon som beregner pangenom-trærne. Problemet ble oppdaget og løst ved å sortere kolonnene i pan-matrisa i samme rekkefølge som genomene i mapping-fila. Dette ga i grunn like trær, men nå med riktige navn på de riktige stedene.

4.4 Kjernegenom og pangenom

Størrelsen på både kjernegenom og pangenom til de to artene *E. faecalis* og *E. faecium* ble estimert ved å bruke mixture model-metoden beskrevet av (Snipen et al. 2009). Det ble estimert modeller med 2 til 8 komponenter. For hvert antall komponenter ble det altså estimert størrelse på kjernegenom og pangenom ved å estimere deteksjonssannsynligheter og andel bidrag fra disse til modellen. For å velge antall komponenter som bør brukes, brukes et kriterium kalt Bayesian Information Criterion (BIC) (Schwartz 1978). Dette går ut på å velge et antall

komponenter som minimerer BIC, og for *E. faecalis* og *E. faecium* faller valget på å bruke henholdsvis 7 og 6 komponenter. Dette gir estimerte kjernegenom-størrelser på henholdsvis 74 og 182, og estimerte pangenom-størrelser på henholdsvis 12529 og 16369. Jo flere komponenter som benyttes, jo likere ser artene ut til å bli i forhold til kjernegenom, mens *E. faecium* fortsatt ser ut til å ha et større pangenom enn *E. faecalis*. Modellene med færre komponenter antyder en større forskjell på artene i forhold til pangenom og kjernegenom.

Det ble også brukt en bootstrap-metode for å estimere pangenom og kjernegenom, hvor målet var å få fram usikkerheten til estimatene. Ut fra de bootstrappede verdiene ble det estimert varians som et mål på usikkerheten. Vi ser her at når estimatet for pangenom øker, øker også variansen, og estimatene kan sies å bli mer usikre. Vanligvis vil estimater for parametere bli mer usikre jo flere parametere som estimeres fra de samme dataene.

Ved å studere utvalget i figur 11 ser det ut til at kjernegenomet tilhørende *E. faecium* er omtrent halvparten så stort som for *E. faecalis*, mens pangenomet ser ut til å være størst for *E. faecium*. Dette kan tyde på at det er større diversitet mellom genomene innen *E. faecium*, enn hos *E. faecalis*. Selv om genomene til *E. faecalis* i gjennomsnitt er større enn *E. faecium*, ser det ut til at de har et betydelig større kjernegenom, og dermed mer til felles med hverandre. Denne forskjellen er også illustrert ved mixture model plot (figur 16); de røde elementene representerer pangenom, mens de blå representerer kjernegenom. De røde elementene er større og de blå mindre hos *E. faecium* enn *E. faecalis*. Forskjellen kommer også fram i tabell 2 og 5, der vi også ser at forskjellen blir mindre etterhvert som flere komponenter benyttes i mixture modellen, det vil si forskjellen på artene utlignes etterhvert som det benyttes flere komponenter. Muligens kan det være et problem med å estimere mange parametere ut fra datautvalget som foreligger.

I figuren (figur 16) som illustrerer mixture modellene observerte vi at det er (minst) to blå elementer, altså elementer som representerer gener med høy deteksjonssannsynlighet, i hver modell. Elementet som blir liggende helt til høyre blir svært lite, og dette skyldes en meget rigid restriksjon i form av deteksjonssannsynlighet for kjernegener på 1,0. Denne sannsynligheten svarer til at kjernegenene er definert som de genfamiliene som finnes i absolutt alle genomer innen en populasjon. Etterhvert som flere genomer blir sekvensert, øker sjansen for at et kjernegen ikke blir oppdaget i et genom. Dette fører til at når det estimeres modeller med såpass mange komponenter, splittes området med høy deteksjonssannsynlighet opp i flere elementer fordi mange av genene vil kunne havne i en gruppe med meget høy deteksjonssannsynlighet,

men som allikevel ikke er 1,0.

For å illustrere hva som skjer dersom man velger en deteksjonssannsynlighet som er svært nær 1,0 , ble det estimert mixture modeller med deteksjonssannsynlighet for kjernegener på 0,999. Dette førte til at den optimale mixture modellen for *E. faecalis* består av seks komponenter og at estimatet for kjernegenomet ble vesentlig større, 1049 mot 74 fra den opprinnelige modellen. For *E. faecium* var fortsatt mixture model med seks komponenter den optimale, med kjernegenom-størrelse på 223. Her ble det ikke opprettet noe nytt komponent med høyere deteksjonssannsynlighet enn 0,999 , og dette kan kanskje forklares med at det er større variasjon innen *E. faecium* og at kjernegenomet er mindre.

Hvilken deteksjonssannsynlighet som skal brukes for kjernegener, bør begrunnes med hva man egentlig definerer som kjernegener, og hvor mye slingsringsmonn man vil tillate. En streng restriksjon for kjernegener i mixture modeller i tillegg til en streng definisjon av genfamilier vil kunne føre til en underestimering av kjernegenom, mens man kan risikere det motsatte om restriksjonene er for åpne. Det har vist seg at andelen variasjon innen forekjellige arter kan være ganske ulik, og kan hende finnes det ikke et enkelt svar som gjelder for alle arter, men at man heller må prøve å tilpasse restriksjonene etter de organismene man jobber med. Dette gir muligens en lite konsekvent bruk av restriksjoner, men naturen er igjen heller ikke alltid så konsekvent som man skulle ønske.

Tabell 4 og 7 inneholder resultater fra estimering av pangenom- og kjernegenom-størrelse der utvalget av genomer ble redusert til ti genomer for hver art. Dette ble gjort for å illustrere hvordan estimatene påvirkes av antall genomer som ligger til grunn for mixture modell-estimeringen. Pangenom-estimatene på 7652 og 8377 for henholdsvis *E. faecalis* og *E. faecium* er vesentlig lavere enn når det benyttes alle genomene fra hver art. Dette indikerer at jo flere genomer som inkluderes, jo høyere blir estimatene for pangenom-størrelse. Dette tyder på at den reelle pangenom-størrelsen for en bakterieart sannsynligvis er meget stor, og at det kreves et betydelig antall sekvenserte stammer for å få til et godt estimat for pangenomet til en bakterieart (Tettelin et al. 2005).

Det ble også beregnet noe som kalles Chao's lower bound. Dette er et estimat på den nedre grensa for pangenom-størrelsen. Estimatet sier ikke noe om hvor stort pangenomet kan være, men hvor stort det minst må være. Dette estimatet tar kun utgangspunkt i hvor mange

genfamilier som finnes i ett genom og hvor mange genfamilier som finnes i to genomer blant utvalget. Estimater på 9913 og 11873 for henholdsvis *E. faecalis* og *E. faecium* tyder også på at pangenomet til *E. faecium* er større enn pangenomet til *E. faecalis*.

4.5 Genomic Fluidity

Genomic fluidity er ment å skulle gi et mål på diversiteten innenfor en populasjon eller gruppe genomer, og er ment som et alternativ til pangenom-størrelse. Det ble beregnet genomic fluidity for *E. faecalis* og *E. faecium* og for disse to artene sammen (tabell 8). Estimatet for *E. faecium* er større enn for *E. faecalis*, noe som igjen tyder på at det er større variasjon mellom genomene innen denne arten. Dette stemmer til en viss grad overens med at estimatene for pangenom-størrelsen til *E.faecium* også er større enn for *E. faecalis*, mens det motsatte er tilfelle for kjernegenom-estimatene. Større pangenom og mindre kjernegenom vil i praksis si at det er større forskjeller på genomene, da de tilsammen innehar et større genomisk materiale mens det samtidig er mindre som er felles for alle genomene i populasjonen. Men det bør merkes at dette gjelder for mixture modellene med få komponenter. Forskjellen på pan- og kjernegenom utlignes etterhvert som det brukes mer kompliserte modeller. Genomic fluidity og de binomiske mixture model estimatene indikerer det samme: Det er større diversitet mellom stammene innen *E. faecium* enn mellom stammene *E. faecalis*.

Det er naturlig å anta at bakterier isolert fra det samme miljøet vil ligne mer på hverandre enn bakterier isolert fra forskjellige miljøer, men i denne sammenhengen har vi ingen informasjon om hvor hver enkelt stamme er isolert fra. Om det benyttes få genomer fra en populasjon, kan det virke noe dristig å hevde at genomic fluidity kan estimeres nøyaktig fra et lite utvalg (Kislyuk et al. 2011), uten noen antakelser om at genomene som ligger til grunn er et representativt utvalg av arten.

Genomic fluidity er et relativt primitivt mål, som tilbyr et enkelt mål på diversiteten innenfor en gruppe bakterier. Det kan diskuteres om man med dette henter ut all den informasjonen man kan om populasjonen, da det hevdes at man ikke trenger særlig mange genomer for å få et godt estimat. Genomic fluidity bidrar til å danne et bilde av hvor åpent genomet til organismen er, men det sier ikke noe særlig om størrelsen på populasjonen slik et pangenom kan gjøre. Begrepet er lansert som et alternativ i forbindelse med at det hevdes at pangenom-størrelse ofte er et høyst usikkert estimat, og at estimater for pangenom-størrelser vokser når antallet genomer bak

estimatet blir flere (Kislyuk et al. 2011; Snipen et al. 2009). Det vil kanskje være rimelig å anta at også genomic fluidity vil kunne øke noe dersom flere genomer benyttes i estimeringen, da flere genfamlier kommer inn og øker andelen som er ulikt mellom tilfeldige genomer.

Selve parameteren minner svært mye om såkalte Jaccard-avstander, som ikke kan sies å være noe nytt begrep innen statistikk. Det som skiller estimeringen av genomic fluidity fra Jaccard-avstander er at når det beregnes andel ulikhet for to genomer blir alle genfamilier i hvert genom telt opp og lagt sammen som det totale antallet genfamilier for de to genomene. Dermed blir genfamiliene som er felles for genomene, telt opp for hvert genom. For å beregne en tilsvarende Jaccard-avstand mellom de samme genomene, ville det heller blitt telt opp alle forskjellige genfamilier som tilsammen fins i to og to genomer. Altså en pangenom-størrelse for par av genomer. Gjennomsnittlige Jaccard-avstander ville sannsynligvis gitt et ganske likt mål på diversitet innen en gruppe bakterier, i forhold til genomic fluidity.

4.6 Unike genfamilier

Det første som ble gjort for å nærmere undersøke de unike kjernegenene, var å finne plasseringen til disse, og se om dette kunne gi noen interessante sammenhenger. Kun genomet til *E. faecalis v583* er komplett, og denne stammen ble derfor brukt som utgangspunkt for å finne plasseringen til de unike *E. faecalis*-genene.

Ved å se på sekvenstagen til alle genene til v583-stammen, så vi at de unike genfamiliene er spredt utover hele kromosomet, i alle fall i forhold til de andre genene i v583-genomet. Dette støttes også av et histogram (figur 17) som viser tettheten av unike gener langs kromosomet, og vi ser her at det er noen bestemte samlinger av gener mens det også er områder hvor vi ikke finner unike gener. Disse områdene uten unike gener korresponderer med definerte mobile genetiske elementer i v583-stammen, som tyder på at histogrammet viser konserverte genklustere.

Det ble funnet et høyt antall unike gener innen *E. faecalis* sammenlignet med hvor mange som ble funnet i *E. faecium*, som det på grunn av antall genomer er naturlig å sammenligne med. Det er rimelig å forvente at man finner færre unike kjernegener i en stor gruppe genomer enn i en liten gruppe. Jo flere genomer, desto mindre sjanse for at en genfamilie har minst et medlem i hvert genom.

Forholdet mellom antallet unike gener og antall genomer stemmer bedre for *E. faecium*, i forhold til antallet genomer og antall unike kjernegener funnet for de to artene med få genomer; henholdsvis *E. casseliflavus* og *E. gallinarum*. Her har vi få genomer, og naturligvis flere unike kjernegener i utvalget.

4.7 COG-klassifisering

For å finne ut mer om de unike kjernegenene innenfor artene *E. faecalis* og *E. faecium*, ble det altså gjort et BLAST-søk mot COG-databasen med både de unike kjernegenene og alle kjernegenene til *E. faecalis* og *E. faecium*. For å avgjøre om et treff var godt nok, ble det brukt en øvre E-verdi- grense på 1×10^{-5}. Dette førte til at ikke alle genfamilier fikk gode nok treff mot COG-databasen. Om grensa var for streng, eller om genene som ikke ga treff er tvilsomme vites ikke, men ved å eksperimentere med forskjellige øvre E-verdier, fant vi at det krevde en meget høy(E-verdi >>1) E-verdi-grense for at alle genene skulle gi treff.

Det ble generert stolpediagrammer som viser fordelingen av genfamilier innen de ulike COG-gruppene. Den blå delen av stolpene representerer andelen unike genfamilier for den enkelte gruppe, mens det røde er antallet genfamilier i kjernegenomet som tilhører den aktuelle funksjonelle gruppa.

For å undersøke forholdet mellom unike kjernegener og alle kjernegener med tanke på fordeling innen funksjonelle grupper, ble det utført en fisher-test for å se om fordelingen av genfamilier innen unike kjernegener og kjernegener er forskjellig. En forskjell i fordelingen kan kanskje gi en indikasjon på hvilke typer gener som står for noe av det som er ulikt mellom artene, og da særlig de COG-gruppene som er overrepresentert blant de unike kjernegenene i forhold til alle kjernegenene.

Tre funksjonelle grupper viste seg å være underrepresentert blant de unike kjernegenene til både *E. faecalis* og *E. faecium*. Det dreier seg om gener for translasjon og ribosomstrukur (J), replikasjon, rekombinasjon og reparasjon (L) og til dels lipidtransport og -metabolisme(I). Disse gruppene består av konserverte gener som er viktige for mange organismer, og det er dermed lite sannsynlig at disse genfamiliene kun finnes innenfor en art. To av disse gruppene, var ikke representert blant de unike kjernegene til *E. faecium* i det hele tatt. Eventuelt kan det være at det

finnes unike gener innenfor disse gruppene, men at disse ikke ga gode nok BLAST-treff mot COG-databasen.

Tatt i betraktning at kjernegenomet ser ut til å være større for *E. faecalis* enn for *E. faecium*, er det ikke overraskende at det her ble funnet flere unike kjernegener. Det viste seg også å være en signifikant forskjellig fordeling av genfamilier innen flere COG-grupper når fordelingen blant de unike kjernegenene og alle kjernegenene ble sammenlignet.
Hos *E. faecalis* var det fire COG-grupper som viste seg å være signifikant overrepresentert blant de unike kjernegenene:

- Transkripsjonsgener (K)
- Gener for signaltransduksjons-mekanismer (T)
- Gener for koenzymtransport og -metabolisme (H)
- Gener uten kjent funksjon (S).

Gruppene for transkripsjonsgener (K) og gener for signaltransduksjon-mekanismer (T) kan inneholde transkripsjonsfaktorer og regulatoriske gener som er spesifikke for art eller stamme. Også gener for koenzymtransport og -metabolisme (H) vil kunne variere mellom arter. Genene uten kjent funksjon utgjør muligens de genene som er mest interressante i forhold til å beskrive det som gjør at *E. faecalis* skiller seg fra andre nært beslektede arter. Det vil kunne være av interresse å studere disse genene mer, for å finne ut om de faktisk er aktive gener, og om de har noen interessante funksjoner.

4.8 Konklusjon

Selv om artene som er brukt i denne oppgaven antas å være nært beslektede, viser det seg å være vesentlige forskjeller mellom dem. Dette er særlig illustrert ved pangenom-treet basert på cloud-vekting. Dette treet vektlegger sjeldne gener innenfor en art, og man skulle kanskje ikke forvente at det ville vise en så tydelig artsinndeling som vi her har sett. Selv om genfamilier er sjeldne innenfor en art, ser de ut til å ha en tilknytning til arten.

Pangenom-treet med shell-vekting antyder at diversiteten i de to artene er forskjellig, og at det er forskjeller mellom artenes kjernegenom. Avstandene mellom genomene innen *E. faecium* er større, eller minst like store, som mellom genomene innen *E. faecalis*. Forskjellen på kjernegenomet ser egentlig ikke ut til å være så stor når det estimeres komplekse mixture modeller, selv om utvalget gir et ganske annet bilde av artene. Når vi videre sammenligner pangenom-størrelser virker også artene her å skille seg fra hverandte. Pangenom-estimatet til *E. faecium* er større enn estimatet for *E. faecalis*, og dette tyder på at diversiteteten er større mellom stammene innen *E. faecium*.

Genomic fluidity gir et estimat på forventet andel ulike gener mellom to genomer fra samme populasjon. Estimert genomic fluidity for de to artene bidrar ytterligere til å kunne si at diversiteten innen *E. faecium* er større enn hos *E. faecalis*, da det er en viss forskjell på estimatene.

Ved å undersøke kjernegenene, kom det fram at den delen av kjernegenene som er unike var av ganske forskjellig størrelse for de to artene i forhold til gjennomsnittlig antall gener og antall genomer. *E. faecalis* har et langt høyere antall unike kjernegener, som naturlig nok henger sammen med et større kjernegenom. For å finne ut mer om de unike kjernegenene, ble de klassifisert etter COG-inndelingen. Her viste det seg at blant de unike kjernegenene til *E. faecalis*, viste gruppa for gener uten kjent funksjon en signifikant større andel blant de unike kjernegenene sammenlignet med andelen blant alle kjernegenen.

Alle undersøkelsene tatt i betraktning, kan det konkluderes med at det er en viss ulik grad av diversitet innen henholdsvis *E. faecalis* og *E. faecium* og at artene skiller seg klart fra hverandre når hele genomer sammenlignes. Helgenom-studier spiller helt klart en viktig rolle innen komparativ genomikk, da dette tilbyr bedre og mer nøyaktig kartlegging av bakteriearter, som

igjen vil kunne bidra til å styrke artsbegrepet og artsinndelingen blant bakterier. I tillegg viser pangenom-størrelse og genomic fluidity at den genetiske variasjonen innenfor bakteriearter kan være stor. Kunnskapen om pangenomet til *E. faecalis* og *E. faecium* vil være viktig i forbindelse med studier av blant annet antibiotikaresistens og patogenitet blant disse artene.

4.9 Mulig videre arbeid

Denne oppgaven har ikke fokusert mye på selve beregningen av genfamilier, selv om genfamilier er et viktig begrep innen helgenom-studier. Selve definisjonen av hva som skal kunne kalles genfamilier og hvor like gener må være for å tilhøre samme genfamilie kan ha mye å si for undersøkelsene og metodene som også er brukt i denne oppgaven. Derfor bør det vies mer tid til å studere hvordan genfamilier bør defineres og beregnes.

For å følge opp arbeidet som er gjort i denne oppgaven, kan man ta utgangspunkt i COG-klassifisering, der det kom fram at funksjonen til en del godt representerte kjernegener er ukjent. Mange av disse genene er unike for art, og det er naturlig å tro at disse genene kan stå for noe av det som er særegent ved en art. Derfor vil det være interessant å studere disse genfamiliene nærmere, og prøve å finne ut om de muligens kan forklare interessante egenskaper som fins innen arten.

I tillegg vil man muligens studere egenskaper som kun opptrer i enkelte stammer i en populasjon, som for eksempel patogenitet eller resistens mot antibiotika. Dersom man ønsker å studere hva som forårsaker en slik egenskap, kan man muligens benytte en statistisk metode for variabelutvelgelse for å prøve å finne forklaringer på egenskapen man er interessert i. Et mulig valg av metode er logisk regresjon (Ruczinski et al. 2003). Dette er en regresjonsmetode der det benyttes logiske operatorer i stedet for aritmetiske, for eksempel 'og', 'eller', 'ikke' med fler, og mulige kombinasjoner av disse. Når det er snakk om gener gir dette mening, da det avgjørende er om et gen er tilstede eller ikke. En annen mulighet er å benytte en partial least square-metode for variabelutvelgelse.

Å benytte statistisk regresjon for å finne årsaken til interessante egenskaper krever mer kunnskap om bakteriene. Det ble vurdert om dette skulle være en del av oppgaven, men av ulike årsaker ble det ikke noe av.

Bibliografi

Coenye, T. & Vandamme, P. (2003). Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiology Letters*, 228 (1): 45-49.

Devriese, L. A., Pot, B. & Collins, M. D. (1993). Phenotypic identification of the genus Enterococcus and differentiation of phylogenetically distinct enterococcal species and species groups. *Journal of Applied Microbiology*, 75 (5): 399-408.

Enright, M. C., Day, N. P. J., Davies, C. E., Peacock, S. J. & Spratt, B. G. (2000). Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of Staphylococcus aureus. *Journal of Clinical Microbiology*, 38 (3): 1008-1015.

Fang, G., Bhardwaj, N., Robilotto, R. & Gerstein, M. B. (2010). Getting Started in Gene Orthology and Functional Analysis. *PLoS Comput Biol*, 6 (3): e1000703.

Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P. & Spratt, B. G. (2004). eBURST: Inferring Patterns of Evolutionary Descent among Clusters of Related Bacterial Genotypes from Multilocus Sequence Typing Data. *J. Bacteriol.*, 186 (5): 1518-1530.

Hyatt, D., Chen, G.-L., LoCascio, P., Land, M., Larimer, F. & Hauser, L. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11 (1): 119.

Kislyuk, A. O., Haegeman, B., Bergman, N. H. & Weitz, J. S. (2011). Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC genomics*, 12: 32.

Kuhnert, R., Del Rio Vilas, V. J., Gallagher, J. & Böhning, D. (2008). A Bagging-Based Correction for the Mixture Model Estimator of Population Size. *Biometrical Journal*, 50 (6): 993-1005.

Lepage, E., Brinster, S., Caron, C., Ducroix-Crepy, C., Rigottier-Gois, L., Dunny, G., Hennequet-Antier, C. & Serror, P. (2006). Comparative Genomic Hybridization Analysis of Enterococcus faecalis: Identification of Genes Absent from Food Strains. *J. Bacteriol.*, 188 (19): 6858-6868.

Levine, D. P. (2006). Vancomycin: A History. *Clinical Infectious Diseases*, 42 (Supplement 1): S5-S12.

Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., et al. (1998). Multilocus sequence typing: A portable

approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95 (6): 3140-3145.

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24 (3): 133-141.

MLST. *Multi Locus Sequence Typing homepage*. Tilgjengelig fra: www.mlst.net.

Mora, M. & Telford, J. (2010). Genome-based approaches to vaccine development. *Journal of Molecular Medicine*, 88 (2): 143-147.

NCBI. *National Center for Biotechnology Information*. Tilgjengelig fra: http://www.ncbi.nlm.nih.gov/.

The R Project for Statistical Computing. Tilgjengelig fra: http://www.r-project.org/.

Rasko, D. A., Rosovitz, M. J., Myers, G. S., Mongodin, E. F., Fricke, W. F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N. R., Chaudhuri, R., et al. (2008). The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *Journal of bacteriology*, 190 (20): 6881-93.

Routledge, R. (2005). *Fisher's Exact Test*. Encyclopedia of Biostatistics: John Wiley & Sons, Ltd.

Ruczinski, I., Kooperberg, C. & LeBlanc, M. (2003). Logic Regression. *Journal of Computational and Graphical Statistics*, 12 (3): 475-511.

Sanger, F. & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94 (3): 441-446.

Schleifer, K. H. (2009). Classification of Bacteria and Archaea: past, present and future. *Systematic and applied microbiology*, 32 (8): 533-42.

Schwartz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6 (2): 4.

Snipen, L., Almoy, T. & Ussery, D. W. (2009). Microbial comparative pan-genomics using binomial mixture models. *BMC genomics*, 10: 385.

Snipen, L. & Ussery, D. W. (2010). Standard operating procedure for computing pangenome trees. *Standards in genomic sciences*, 2 (1): 135-41.

Solheim, M. (2010). *A Study of Mechanisms Involved in The Pathogenicity of Enterococcus faecalis by DNA Microarrays*. Philosophiae Doctor Thesis. Ås: Norwegian University of Life Sciences (UMB), Department of Chemistry, Biotechnology and Food Science.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., et al. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of*

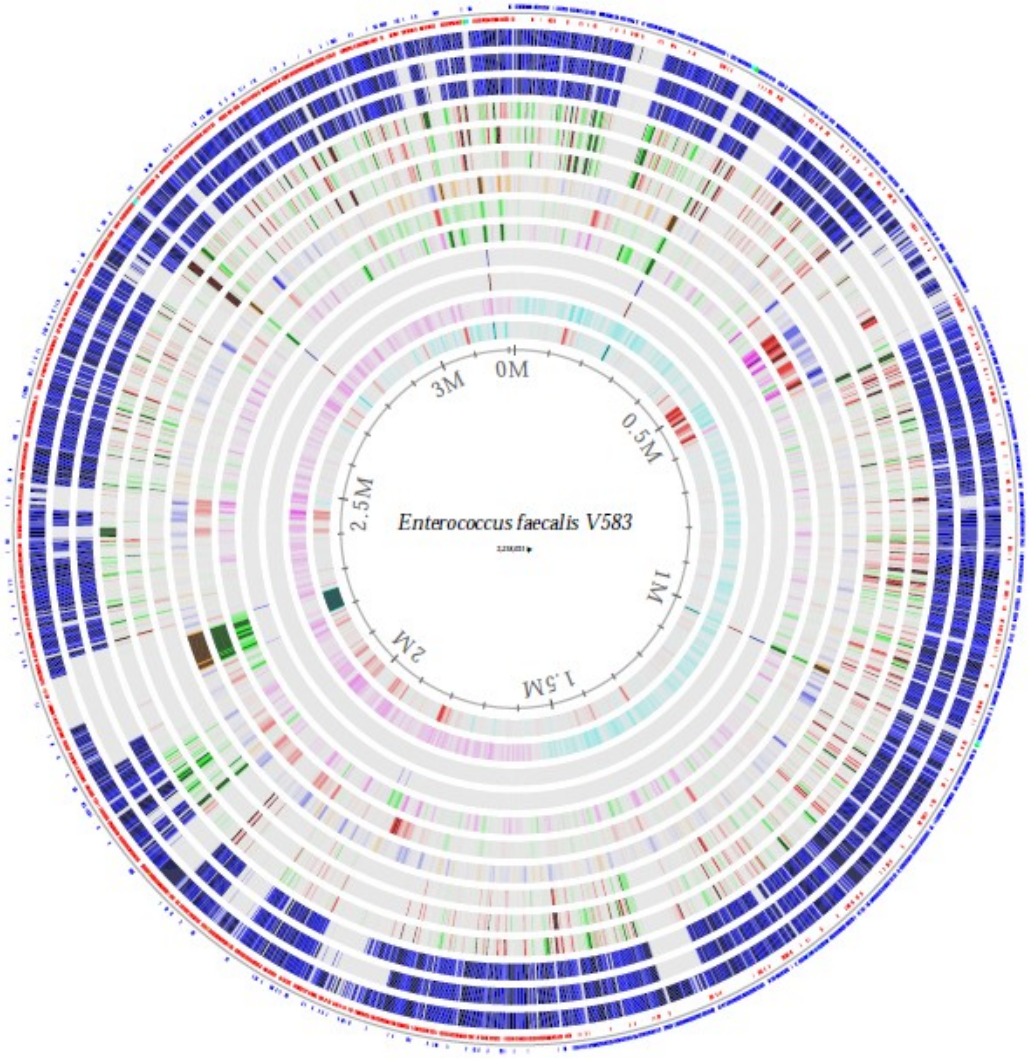
America, 102 (39): 13950-5.

Urwin, R. & Maiden, M. C. J. (2003). Multi-locus sequence typing: a tool for global epidemiology. *Trends in Microbiology*, 11 (10): 479-487.

Vebø, H. C., Solheim, M., Snipen, L., Nes, I. F. & Brede, D. A. (2010). Comparative Genomic Analysis of Pathogenic and Probiotic *Enterococcus faecalis* Isolates, and Their Transcriptional Responses to Growth in Human Urine. *PLoS ONE*, 5 (8): e12489.

Woese, C. R. (1987). BACTERIAL EVOLUTION. *Microbiological Reviews*, 51 (2): 221-271.

Vedlegg



Figur 20: Illustrasjon av kromosomet til E. faecalis v583 (Vebø, 2010).