CHALMERS | GÖTEBORGS UNIVERSITET

Analys av gener och arter i metagenomikdata

Examensarbete för kandidatexamen i matematik vid Göteborgs universitet

Emma Eriksson
Sofia Lebens

Institutionen för matematiska vetenskaper
Chalmers tekniska högskola
Göteborgs universitet
Göteborg 2017

Analys av gener och arter i metagenomikdata

Examensarbete för kandidatexamen i matematisk statistik vid Göteborgs universitet

Emma Eriksson Sofia Lebens

Handledare: Tobias Österlund
Examinator: Marina Axelson-Fisk

Institutionen för matematiska vetenskaper
Chalmers tekniska högskola
Göteborgs universitet
Göteborg 2017

Populärvetenskaplig presentation

I miljön som vi lever i finns en mängd icke-naturliga ämnen som kommer från användning av diverse produkter i vår vardag. Substansernas effekt på människor och andra organismer är ofta inte fullständigt kartlagd och det är därför viktigt att utveckla metoder för att kunna undersöka vilken effekt dessa ämnen har, både ur hälso- och miljöperspektiv. Ett ämne som flitigt har använts i exempelvis tandkräm för att förhindra bakterietillväxt är triclosan. Det har visats att ämnet inte enbart har bakteridödande effekt utan kan även utgöra en hälsorisk vilket gjort att ämnet nu har förbjudits i en mängd olika produkter. Trots att konsumtionen av triclosan nu kraftigt har reducerats återfinns ämnet i naturen där det kan påverka både djur- och växtliv. Till följd av den bakteriedödande effekten hos triclosan påverkar ämnet mikroorganismer i exempelvis havsmiljö. Detta kan ge effekter i form av förändringar i den biologiska mångfalden vilket i sin tur kan påverka hela ekosystem och därmed få stora konsekvenser i miljön.

I det här projektet utvecklades en statistisk metod för att studera prov av havsvatten som hämtats från den svenska västkusten. Havsvatten innehåller en stor mängd olika mikroorganismer i komplexa sammansättningar och för att studera hur triclosan påverkar dessa samhällen av mikroorganismer behandlades proven med olika mängd av ämnet. Studiens ena syfte var att försöka hitta samband mellan artsammansättning och funktionalitet i proven och från dessa dra slutsatser om associationer mellan arter och gener. För att studera dessa samband undersöktes hur förekomsten av gener och arter påverkas av triclosan. Den data som ligger till grund för studien består av den kvantifierade mängden av arter och gener i de olika proven. Denna typ av data, som erhållits genom att extrahera och kartlägga allt genetiskt material i prov hämtade direkt från miljön, benämns som metagenomikdata. Den stora mängd information som återfinns i metagenomikdata kan nu i större utsträckning än tidigare undersökas till följd av nya sekvenseringsmetoder vilket gör att mikrobsamhällen kan studeras på en nivå som tidigare varit omöjlig.

Studiens andra syfte var att undersöka själva metoden som tillämpades för att bedöma hur väl den lämpar sig för att analysera här typen av data. Metoden baseras på att identifiera grupper av arter som vissa funktioner är associerade med. De konstellationer som bildas på detta sätt består av arter och gener som påverkas på samma sätt av triclosan. Möjliga asso-

ciationer mellan en grupp arter och en funktionalitet kan därmed erhållas men det går inte att med säkerhet fastställa vilka gener och därmed funktioner som finns i en viss grupp av arter. Resultaten visar på vilka gener som kan vara sammankopplade med funktioner som gör att vissa organismer klarar av att hantera triclosan medan andra gener visar på det motsatta. Gener som kodar för proteiner involverade i bakteriers immunförsvar observerades bland arter som överlever i triclosan och dessa gener kan därför antas kunna bidra till att arterna har förmåga att etablera sig i höga koncentrationer av triclosan. I grupper av arter som inte har förmåga att överleva i triclosan hittades gener som kodar för DNA-överföring mellan organismer. Denna funktion skulle därför kunna bidra till att arter som har dessa gener inte lyckas etablera sig i triclosan.

De samband mellan gener och arter som observerats i denna studie kan utgöra en viktig utgångspunkt för fortsatta studier av hur triclosan och andra substanser påverkar mikroorganismer i den naturliga miljön och kan bidra till förståelse om tidigare okända relationer. Vidare ger studien en indikation på att en metod som bygger på associationer mellan gener och arter i prov av denna typ kan användas för att hitta komplexa samband i samhällen av mikroorganismer. Genom att som i denna utforskande studie kombinera olika dataset i den statistiska analysen kan nya sätt att analysera data av olika slag utvecklas och den här typen av metoder för att extrahera information från olika dataset har stor potential att kunna användas inom många olika områden.

Sammanfattning

I takt med att nya DNA-sekvenseringsmetoder utvecklas kan den stora mängd information som återfinns i metagenomikdata i större utsträckning än tidigare undersökas vilket öppnar upp för nya möjligheter att studera mikrobsamhällen. I det här projektet har vi undersökt hur relationer och samband mellan arter och gener kan hittas baserat enbart på information om deras respektive förekomst i två separata dataset. De aktuella dataseten innehåller förekomst av gener och arter från 16 olika mikrobsamhällen bestående av havsvatten som behandlats med olika koncentrationer av den antibakteriella substansen triclosan. Gener och arter som svarar på samma sätt vid ändringar i koncentrationen av triclosan grupperades och dessa konstellationer analyserades vidare. Metoden för att skapa konstellationerna bygger i ett första steg på hierarkisk klustring av gener respektive arter baserat på korrelationen mellan förekomsten i proverna. I nästa steg i analysen identifierades till varje kluster av arter gener med stark korrelation till arterna och till varje kluster av gener identifierades arter med stark korrelation till generna. De bildade konstellationerna av arter och gener visade på hög stabilitet vid variationer av olika parametrar i metoden och var i stort sett oberoende av om klustringen baserades på arter eller gener. Konstellationerna hade även stor homogenitet i avseende på arter och genfunktionalitet vilket vi tolkar som att sannolikheten är stor att verkliga samband kan identifieras. Sanna associationer mellan gener och arter kan med säkerhet inte konstateras med denna metod men flera intressanta mönster och samband i konstellationerna observerades och bör undersökas vidare. Exempelvis hittades gener som kodar för DNA-överföring mellan bakterier i konstellationer som inte överlever i triclosan medan gener associerade med bakteriers immunförsvar återfanns i konstellationer som har förmåga att etablera sig i triclosan. Analyser där data för både gener och arter kombineras utgör ett område som inte är väl studerat men där det troligen finns mycket ny information att extrahera. Statistiska analyser av associationer mellan gener och arter kan bidra till ökad förståelse för tidigare okända samband mellan dessa samt ge upphov till nya idéer och hypoteser värda att testas ytterligare.

Abstract

As new extreme high throughput DNA sequencing methods continue to develop the large amounts of information that they give rise to in terms of metagenomic data opens the way for completely novel approaches to the study of microbial ecosystems. In this project we have investigated how relationships and connections between microbial genes and species that carry them can be found based entirely on their appearance in two separate datasets accumulated from the same samples using metagenomic analysis. The dataset used in the study consists of abundance of genes and species derived from DNA isolated from microbial communities in biofilms formed in seawater treated with different concentrations of the antimicrobial agent triclosan. Genes and species that responded in the same way to changes in the concentration of triclosan were grouped together for further analysis. The method used for creating the different constellations consisted of a first step where genes and species were clustered based on their abundance in the samples. In the next step genes with strong correlations to each cluster of species and species with strong correlations to each cluster of genes were identified. These constellations based on species and genes were robust appearing not to vary with variations in the parameters of the analysis and not to be dependent on whether the clustering was based on associations of genes or vice versa. The constellations were also homogeneous with respect to species and gene functionality (the same genes clustering with the same species) which we interpret as meaning that the likelihood of a tangible connection between them being identified is high. Clearly, concrete conclusions regarding the species and the genes they carry cannot me made using the methods we present here, but several interesting patterns have emerged that would bare further scrutiny. For example, genes involved with the horizontal transfer of DNA between species do not appear to survive in triclosan whereas genes associated with the bacterial immune system were highly associated with bacteria that were able to establish themselves in the presence of triclosan. Few analyses have been done in which information about the species present in a studied niche or ecosystem and the genes that they collectively contain are combined and there is much new information to be derived from such studies. Statistical approaches to the analysis of species and their collective genome has the potential to give new insights into previously unknown associations and to develop hypotheses that can be further tested experimentally.

Innehåll

1 Inledning **7**
- 1.1 Metagenomik . . . 7
- 1.2 Triclosan . . . 8
- 1.3 Syfte . . . 9

2 Metod och implementering **10**
- 2.1 Förbehandling och inledande analys av data . . . 10
  - 2.1.1 Data . . . 10
  - 2.1.2 Filtrering av data . . . 10
  - 2.1.3 Normalisering av data . . . 10
  - 2.1.4 Inledande analys av data . . . 11
- 2.2 Del 1 och 2 - Undersökning av gener och arter i de individuella dataseten . . . 11
- 2.3 Del 3 - Korrelation mellan gener och arter . . . 12
  - 2.3.1 Framtagande av korrelationsmatris . . . 12
  - 2.3.2 Analys av korrelationsmatris . . . 13
    - 2.3.2.1 Inledande analys av korrelation mellan gener och arter . . . 13
    - 2.3.2.2 Klustring av gener/arter . . . 13
      - 2.3.2.2.1 Hierarkisk klustring . . . 13
      - 2.3.2.2.2 Identifiering av kluster . . . 13
      - 2.3.2.2.3 Analys av klustringsresultat . . . 14

3 Resultat **15**
- 3.1 Inledande analys av data . . . 15
- 3.2 Analys del 1 - Undersökning av geners variation med koncentrationen triclosan 16
- 3.3 Analys del 2 - Undersökning av arters variation med koncentrationen triclosan 18
- 3.4 Analys del 3 - Korrelation mellan gener och arter . . . 18
  - 3.4.1 Inledande analys av korrelation mellan gener och arter . . . 18
  - 3.4.2 Konstellationer av arter och gener skapade baserat på klustring av arter 20
  - 3.4.3 Konstellationer av arter och gener skapade baserat på klustring av gener 22

4 Diskussion **26**

Referenser **30**

Förord

Projektet genomfördes i en grupp av två utan uppdelning av särskilda ansvarsområden mellan personerna. Planeringen av projektet och samtliga beslut togs gemensamt och det praktiska arbetet med projektet utfördes till största delen tillsammans. Hela projektet utfördes i R och utveckling av programkod för de olika analyserna gjordes tillsammans. Vissa mindre praktiska delar genomfördes var för sig men alltid i nära samråd med den andra personen. Även viss inläsning på bakgrund och metoder gjordes individuellt. Metoder och erhållna resultat har under projektets gång kontinuerligt diskuterats inom gruppen och med handledaren för att driva projektet framåt. Detta har resulterat i många nya frågeställningar och idéer som undersökts vidare. I slutskedet av projektet lades mycket tid på att diskutera resultaten som erhållits för att dra slutsatser om hur metoder som de som använts i detta projekt kan tillämpas på metagenomikdata men också vilken problematik som uppkommer i och med komplexiteten i datan.
Rapporten har till stor del skrivits tillsammans. I arbetet med rapporten skrev Emma grunden till Introduktion och Metod och implementering och Sofia skrev grunden till Diskussion, men därefter bearbetades texten grundligt av båda personerna vilket resulterat i att ingen varit huvudansvarig för olika avsnitt. Övriga delar skrevs av båda personerna.
En loggbok där aktivitet samt varje persons spenderade tid dokumenterats har förts under projektet.

1 Inledning

1.1 Metagenomik

Metagenomik är ett relativt nytt område där genetiskt material från prover hämtade direkt från den naturliga miljön sekvenseras och studeras (*1*, *2*). Tidigare har studier utförts genom att enskilda stammar av mikroorganismer isolerats och odlats i laboratoriemiljö och därefter har DNA sekvenserats från arter som isolerats ur dessa. Då mikroorganismer i naturen bildar komplexa samhällen av en stor mängd olika arter (*3*) och då majoriteten av mikroorganismer dessutom är svåra att kultivera i laboratoriet (*4*) ger de traditionella metoderna inte svar på hur mikrobsamhällena egentligen ser ut och den mikrobiologiska biodiversiteten missas således.

Metagenomik baseras på att allt DNA från samtliga celler i ett prov först extraheras och därefter slumpmässigt klyvs vilket resulterar i ett stort antal mindre fragment. Fragmenten utgör ett slumpmässigt prov från metagenomet, det totala genomet som representerar den totala mängden DNA i provet. Fragmenten sekvenseras därefter för att kartlägga ordningen av nukleotiderna i sekvenserna. Snabbare och mer kostnadseffektiva sekvenseringsmetoder har på senare år utvecklats vilket resulterat i att mikrobiologiska samhällen kan undersökas i en mycket större skala än någonsin tidigare. De moderna metoderna möjliggör sekvensering av allt genetiskt material i ett prov och ger på så sätt en mer komplett bild av vilka arter som finns och från informationen om vilka gener som förekommer kan man dra slutsatser om metabolism och annan aktivitet i mikrobsamhället. Storskaliga projekt med syfte att identifiera exempelvis mikroorganismer associerade med olika hälso- och sjukdomstillstånd (Human Microbiome Project (*5*, *6*)) och mikrobsamhällen i olika miljöer på jorden (Earth Microbiome Project (*7*)) resulterar i miljarder DNA-fragment som nu är möjliga att sekvensera tack vare nya metoder. Användningsområden för metagenomik är omfattande och i det medicinska området har studier genomförts för att identifiera mikrobsamhällen associerade med exempelvis inflammatoriska tarmsjukdomar (*8*, *9*). Liknande studier har genomförts på prov från individer med typ 2-diabetes där både arter och de funktioner som kodas av genomen hos organismerna i samhällen associerade till sjukdomen identifierats (*10*, *11*). Metagenomik har även använts för att kartlägga tidigare okända virus (*12*) och gener (*13*), samt kan ge information om miljöförhållanden associerade med gener vars funktioner tidigare varit okända (*14*).

Två aspekter av mikrobsamhällen, nämligen vilka arter som finns där och vilka funktioner de bidrar med, kan undersökas med hjälp av metagenomikdata (*15*). Detta är möjligt till följd av att vissa DNA-fragment härstammar från kodande regioner (gener) av genomet som ger information om taxonomisk tillhörighet medan andra fragment härstammar från regioner som ger information om biologiska funktioner hos organismerna. Fragmenten jämförs mot en referenssekvens som innehåller redan identifierade regioner och generna kvantifieras genom att summera antalet fragment som matchar varje region. För att bestämma den taxonomiska tillhörigheten används som referens den gen som kodar för 16S ribosomalt RNA (rRNA) som utgör en del av den prokaryota ribosomen. Motsvarande gen som kodar för 18S rRNA finns i eukaryoter. Vissa regioner av 16S/18S rRNA-generna är starkt konserverade eftersom de är viktiga för att upprätthålla cellens funktion medan andra regioner skiljer sig mycket mellan olika arter vilket gör att de kan användas för att bestämma arttillhörighet. Databaser innehållande sekvenser för 16S/18S rRNA-generna i olika arter används för att tillskriva ett sekvenserat fragment till en viss art och databaser för proteinfamiljer används för att klassificera ett fragment till en viss funktion. En proteinfamilj är en grupp evolutionärt relaterade proteinsekvenser som kodar för gener som antas ha samma biologiska funktion.

Denna typ av genbaserad metagenomikdata kan användas för att undersöka skillnader i relativ gen- och artförekomst under olika experimentella förhållanden som exempelvis olika sjukdomstillstånd eller vid tillsatser av antibakteriella substanser. En statistisk analys genomförs för att identifiera gener och arter vars relativa förekomst ändras mellan mikrobsamhällena i proven. På grund av storleken och komplexiteten hos sekvenserade metagenom

är den statistiska analysen dock komplicerad. Metagenomikdata är ofta högdimensionell eftersom förekomsten av flera tusen gener testas samtidigt. Dessutom innebär varje test en risk att felaktigt förkasta nollhypotesen om att förekomsten inte ändras mellan olika prov och därför krävs korrigering för multipla test för att kontrollera typ-I fel och i och med det öka styrkan på testet för att detektera verkliga skillnader mellan prov (*16*). Trots att kostnaden för DNA-sekvensering gått ned med utvecklingen av nya metoder är det fortfarande dyrt att sekvensera den stora mängd DNA som finns i ett metagenomikprov och därför är antalet biologiska replikat i allmänhet litet i den här typen av studier. Biologisk och teknisk variation påverkar också i hög grad metagenomikdata. Biologisk variation hänrör från naturliga skillnader i genförekomst mellan mikrobsamhällen medan teknisk variation uppkommer vid den experimentella bearbetningen av prover. Källor till teknisk variation kan vara exempelvis processen för förbehandling av prover (*17*) samt sekvenseringsfel (*18*). DNA-fragment kan också felaktigt matchas till en viss kodande region i en referenssekvens (*19*). Det sistnämnda kan delvis vara ett resultat av att databaserna som används för att tillskriva ett sekvenserat fragment till en viss gen enbart innehåller gener som tidigare identifierats. Det är viktigt att i detta sammanhang poängtera att hela genom tidigare endast sekvenserats för en liten andel av det totala antalet mikroorganismer (*20*) och därför saknas en stor del av informationen som krävs för att korrekt identifiera samtliga gener i ett metagenom. För att kunna utföra en statistisk analys av metagenomikdata krävs att metoderna som används kan hantera komplexiteten i datan. Många metoder har utvecklats med syfte att identifiera gener som ändras mellan prov från olika mikrobsamhällen. I en nyligen publicerad studie baserad på data innehållande förekomst av gener i två typer av metagenom jämfördes en mängd olika metoder för detta ändamål (*21*). Metoderna presterade olika på olika dataset och stora skillnader i prestation observerades generellt till följd av exempelvis antalet prov.

I detta projekt kommer metagenomikdata från samhällen av mikroorganismer som härstammar från Gullmarsfjorden utanför Lysekil att undersökas med hjälp av statistiska metoder. Den experimentella delen av projektet genomfördes vid Sven Lovén Center för marin infrastruktur (för experimentella detaljer se (*22*)). Havsvatten pumpades kontinuerligt in i 16 stycken 20 liter stora akvarium till vilka den kemiska substansen triclosan tillsattes i olika mängd. Mikrobsamhällen som bildats på glaset i akvarierna togs om hand efter 18 dagar och varje prov sekvenserades med Illumina-metoden. DNA-fragmenten klassificerades därefter taxonomiskt med 16S/18S rRNA och för att identifiera biologiska funktioner användes TIGRFAM-databasen (*23*) som utnyttjar en dold Markovmodell för klassificering av proteinsekvenser.

1.2 Triclosan

Triclosan (5-chloro-2-(2,4-dichlorophenoxy)phenol) är en organisk förening som verkar både på bakterier och svamp. Substansen är vanligt förekommande i tvål, tandkräm, kosmetika och schampo. Även i leksaker, kläder och skor finns triclosan för att förhindra bakteritillväxt. Triclosan verkar genom att binda till och inhibera ett enzym som är essentiellt för fettsyrasyntesen i bakterier och kloroplaster (*24*). Bristen på fettsyror påverkar stabiliteten på cellmembranet som är avgörande för cellens överlevnad. En stor del av ämnet hamnar i naturen eftersom vattenreningsverk inte kan avlägsna all triclosan (*25*) och därför har ämnet detekterats i höga halter i sjöar, hav och vattendrag över hela världen (*26–29*). Triclosan är giftigt för vattenlevande organismer och studier visar att mikroalger är de känsligaste organismerna (*30–33*). På den svenska västkusten har koncentrationer på upp till 0,55 nM uppmätts (*34*). I Sverige har triclosan hittats både i human plasma och bröstmjölk (*35*, *36*). Användandet av triclosan och andra substanser som verkar som biocider kan också ge upphov till att bakterier utvecklar korsresistens mot antimikrobiella läkemedel (*37*). Studier visar också att triclosan kan påskynda cancertillväxt (*38*, *39*) och att ämnet har hormonstörande effekt (*40*). Sedan 2016 är triclosan förbjuden att användas i ett stort antal produkter inom EU (*41*).

1.3 Syfte

Syftet med studien är att studera samband mellan gener och arter i metagenomikdata och till grund för studien ligger kvantifierad förekomst av gener och arter i prov av havsvatten som behandlats med olika koncentration av triclosan. Datan kommer inledningsvis att analyseras för att identifiera gener respektive arter vars förekomst påverkas av triclosan. Vidare är syftet att studera vilka förändringar i biologisk funktion mellan olika metagenom som är associerade med förändringar i artsammansättningen mellan prover. Statistiska studier på metagenomikdata utförs ofta genom att undersöka hur förekomsten av antingen gener (*21*) eller arter (*42*) skiljer sig åt mellan prov från olika miljöer, men denna analys syftar till att undersöka hur korrelationen mellan gener och arter i proven kan analyseras. Genom att studera både gener och arter tillsammans kan en ökad förståelse för tidigare okända samband mellan dessa erhållas. Hela genom hos enbart en liten andel av det totala antalet mikroorganismer har färdigställts vilket innebär att den här typen av statistiska analyser kan bidra till observationer om samband mellan gener och arter som kan öppna upp för tillämpningar inom en lång rad olika områden.

2 Metod och implementering

2.1 Förbehandling och inledande analys av data

2.1.1 Data

Studien utfördes på två olika dataset med förekomst av gener respektive arter. Förekomsten av en gen eller art anger antalet DNA-fragment som matchar respektive gen eller art i ett visst prov. Varje dataset består av en matris där raderna representerar gener/arter och kolumnerna representerar de 16 prov som behandlats med olika koncentrationer av triclosan. Totalt innehåller datan förekomst av 3676 gener och 6186 arter. Både prokaryoter (bakterier) och eukaryoter finns i artdatan. I denna rapport används härefter ordet gen för att representera något element identifierat i TIGRFAM-databasen, vilket kan vara en genfamilj eller en enskild gen. Tabell 1 visar antal prov som behandlats med olika koncentrationer av triclosan. Notera att de angivna koncentrationerna genom hela denna rapport avser den mängd triclosan som adderades till de olika akvariumen under experimentet och den ursprungliga koncentrationen triclosan i havsvattnet är därför inte inräknad i dessa.
Analysen av de två dataseten genomfördes i R (*43*) och beskrivs nedan.

Tabell 1: Antal prov som behandlats med olika koncentrationer av triclosan.

Triclosan konc (nM)	Antal prov
0	4
0,316	1
1	1
3,16	3
10	1
31,6	1
100	1
316	3
1000	1

2.1.2 Filtrering av data

Dataseten filtrerades inledningsvis genom att ta bort arter och gener vars totala förekomst över samtliga prov var mindre än 10. Denna filtrering gjordes för att reducera antalet gener och arter med mycket låg förekomst vilket kan härstamma från exempelvis felaktig klassificering av DNA-fragment. Artdatan innehåller förekomst av mitokondrier och kloroplaster eftersom 16S/18S rRNA-genen som används för taxonomisk klassificering även finns i dessa och de togs därför bort, tillsammans med en grupp som klassificerats som okända. Dataseten som användes i den följande analysen innehåller 3498 gener och 1963 arter.

2.1.3 Normalisering av data

Normalisering av den här typen av data är viktigt eftersom det reducerar tekniska skillnader mellan prov så att dessa har minimal påverkan på de statistiska resultaten. Provet som behandlats med 31,6 nM triclosan är djupare sekvenserat än de övriga proven och förekomsten av gener och arter i detta prov är därför generellt högre. Normalisering gör att denna och liknande tekniska skillnader inte kommer att spela in i den statistiska analysen. Normalisering av datan i denna studie utfördes genom att dividera förekomsten av varje gen och art med den totala förekomsten av gener respektive arter i det aktuella provet, om inget annat anges nedan. Denna typ av normalisering resulterar i relativ förekomst som motsvarar proportionen av varje gen respektive art av den totala förekomsten i ett prov. Summan av den totala förekomsten i varje prov är därför 1. Då ordet förekomst härefter används i denna rapport menas relativ förekomst efter normalisering.

2.1.4 Inledande analys av data

Principalkomponentanalys genomfördes för att hitta mönster i datan och visualisera generella skillnader mellan proven. Genom att beskriva alla gener respektive arter i form av linjärkombinationer kan dimensionen på datan minimeras för att underlätta analys och visualisering. De p generna eller arterna transformeras till M nya variabler, där $M < p$. De nya variablerna är linjärkombinationer av de ursprungliga variablerna (gener eller arter) och kan skrivas som

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j, \quad m = 1, ..., M \tag{1}$$

för konstanterna $\phi_{1m}, \phi_{2m}, ..., \phi_{pm}$. Den första principalkomponenten definieras som den linjärkombination av variabler för vilken variansen är som störst. Denna linjärkombination beskriver riktningen i vilken datan varierar som mest. Den andra principalkomponenten definieras därefter som den linjärkombination av variabler som har störst varians bland de linjärkombinationer som är okorrelerade med den första, och beskriver därför en riktning som är ortogonal mot den första. Efterföljande principalkomponenter definieras på motsvarande sätt. Analysen utfördes med `prcomp` i `stats`-paketet i R på båda dataseten innehållande förekomst av gener respektive arter. Datan normaliserades inte men skalades och centrerades före analysen.

2.2 Del 1 och 2 - Undersökning av gener och arter i de individuella dataseten

För att inledningsvis få en bild av hur triclosan påverkar gener plottades förekomsten av varje gen i proven som inte behandlats med triclosan (kontroll) mot samma gens förekomst i prov som behandlats med fyra olika koncentrationer (0,316, 3,16, 31,6 och 316 nM) av triclosan. Fyra kontrollprov fanns i dataseten men eftersom ett av dessa visar ett något avvikande mönster i förekomsten av gener och arter plottades medelvärdet av förekomsten i de tre andra kontrollproven. För koncentrationerna 0,316 och 31,6 nM fanns enbart ett prov av varje, men för proven som behandlats med 3,16 och 316 nM triclosan fanns tre replikat och medelvärdet av förekomsten i dessa användes i plottarna. Därefter gjordes samma plottar för förekomsten av arter. Datan normaliserades inledningsvis genom att dividera förekomsten av varje gen och art med den totala förekomsten av gener och arter i respektive prov.

Efter denna parvisa jämförelse genomfördes en statistisk analys i `edgeR` (*44*) i R för att identifiera gener och arter som ändras signifikant med triclosan-koncentrationen. I en studie med syfte att jämföra olika metoder för att identifiera gener som ändras signifikant mellan metagenom var `edgeR` en av de metoder som presterade överlag bäst (*21*). Förekomsten av en gen eller art i prov är diskret och kan antas vara Poisson-fördelad, men till följd av biologisk variation mellan replikat är variansen större än väntevärdet. `edgeR` modellerar därför datan som överspridd (overdispersed) med hjälp av en negativ binomialfördelning. En generalized linear model (GLM) (*45*) baserad på en negativ binomialfördelning användes för att testa om förekomsten varje gen respektive art ändras signifikant med koncentrationen triclosan. GLM kan ses som en förlängning av klassiska linjära regressionsmodeller och används för icke-normalfördelad data. Modellen specificerar en viss sannolikhetsfördelning med hjälp av fördelningens förhållande mellan väntevärde och varians. För en negativ binomialfördelning definieras väntevärdet för en gen/art i i prov j som $\mathbb{E}[Y_{ij}] = \mu_i$ och variansen som $\text{var}[Y_{ij}] = \mu_i + \phi_i \mu_i^2$ där ϕ_i är så kallad dispersion som inkluderar alla typer av variation mellan replikat, både sådan som uppkommer till följd av tekniska och biologiska skillnader. Linkfunktionen i GLM-modellen för denna fördelning definieras som logaritmen av väntevärdet av förekomsten av en gen eller art. För att ta hänsyn till att datan innehåller teknisk variation mellan prov måste även en normaliseringfaktor inkluderas i modellen i form av $\log(N_j)$. En log-linjär modell kan därefter anpassas för varje gen respektive art i enligt

$$\log(\mathbb{E}[Y_{ij}]) = \alpha_i + \beta_i x_j + \phi_i + \log(N_j) \tag{2}$$

där x_j beskriver koncentrationen (log) av triclosan som prov j behandlats med och β_i är regressionskoefficienten för gen/art i. För att testa om förekomsten av genen eller arten ändras mellan prov som behandlats med olika koncentration triclosan testas nollhypotesen att $\beta_i = 0$. I denna analys normaliserades datan med trimmed mean of M-values (TMM) (*46*) som visats vara en metod som konsekvent presterar bra på denna typ av dataset (*47*). TMM utvecklades som ett alternativ till metoden som använder den totala förekomsten i varje prov som normaliseringsfaktor. TMM utgår från en normaliseringsfaktor som anger produkten mellan den totala förekomsten i provet och en skalningsfaktor. Skalningsfaktorn baseras på varje par av prov och beräknas genom att använda log-fold change (förhållandet mellan förekomsten i de två proven) och absolut intensitet från vilka de mest extrema värdena tas bort (trimming).

I fallet med multipla test då flera hypoteser testas samtidigt måste p-värdena korrigeras för antalet hypotestest som utförs för att kontrollera fel av typ I och därmed minimera antalet falska förkastanden. I denna analys gjordes ett hypotestest per gen respektive art där nollhypotesen är att genen eller arten inte ändras med koncentrationen triclosan. Korrigerade p-värden baserade på Benjamini-Hochberg-metoden (*48*) kontrollerar false discovery rate (FDR) och beräknades med `edgeR`. FDR är den förväntade andelen felaktiga förkastanden (false positives) bland de gener/arter för vilka nollhypotesen förkastats. Samtliga metoder för att korrigera p-värden kräver att de m p-värdena initialt ordnas enligt $p_{(1)} \leq ... \leq p_{(m)}$. För Benjamini-Hochberg-metoden multipliceras därefter varje p-värde $p_{(i)}$ med $a_i = m/i$ där $i = 1, ..., m$ så att $p'_{(i)} = a_i p_{(i)}$. Om denna multiplikation gör att p-värdena inte längre följer samma ordning som tidigare minskas det största värdet i varje par där ordningen ändrats, enligt $\widetilde{p}_{(i)} = \max_{j=i,...,m} p'_{(i)}$. Därefter definieras de korrigerade p-värdena som $\widetilde{p}_{(i)} = \min(\widetilde{p}_{(i)}, 1)$ för alla i. Genom att de korrigerade p-värdena är större än de ursprungliga förkastas ett mindre antal hypoteser för en förbestämd signifikansnivå och resulterar i ett mindre antal signifikanta gener/arter. I denna studie användes en signifikansnivå på 5 % vilket också motsvarar den FDR som de korrigerade p-värdena förväntas generera.

2.3 Del 3 - Korrelation mellan gener och arter

2.3.1 Framtagande av korrelationsmatris

En korrelationskoefficient ger ett mått på hur starkt relaterade två variabler är till varandra. Pearsons korrelationskoefficienter för samtliga kombinationer av förekomst av gener och arter i proven beräknades och sparades i en matris där arter återfinns som rader och gener som kolumner. Datan normaliserades först genom att dividera förekomsten av varje gen och art med den totala förekomsten av gener och arter i respektive prov. En hög korrelation mellan en gen och en art indikerar att förekomsten av genen och arten varierar på liknande sätt med förändringar i koncentrationen av triclosan. För två variabler där X och Y kan representeras av en gen respektive en art i populationen beräknas Pearsons korrelationskoefficient ρ enligt

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{3}$$

där

$$\mathrm{cov}(X,Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \tag{4}$$

σ_X och σ_Y är standardavvikelsen för X respektive Y, och μ_X och μ_Y är deras väntevärden. För att beräkna korrelationskoefficienten baserad på ett stickprov uppskattas kovariansen och variansen utifrån provet. Korrelationskoefficienten r för genen $\{x_1, ..., x_n\}$ och arten $\{y_1, ..., y_n\}$ i n prov beräknas då enligt

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{5}$$

För varje korrelationskoefficient genomfördes ett hypotestest som testar avsaknaden av korrelation med nollhypotesen $H_0 : \rho = 0$ och den tvåsidiga alternativa hypotesen $H_1 : \rho \neq 0$. En teststatistika baserad på korrelationskoefficienten r för stickprovet beräknades enligt

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{6}$$

som under nollhypotesen följer en t-fördelning med $n-2$ frihetsgrader, där n är antalet prov, 16 i detta fall. p-värden för samtliga korrelationskoefficienter mellan gener och arter beräknades, och korrigerades därefter med Benjamini-Hochberg-metoden (*48*) som beskrivits ovan. Dessa p-värden korrigerades per rad i korrelationsmatrisen.

2.3.2 Analys av korrelationsmatris

2.3.2.1 Inledande analys av korrelation mellan gener och arter

Bland de gener och arter som identifierats som sådana vars förekomst ändras signifikant med koncentrationen triclosan beräknades antalet arter som varje gen är starkt korrelerad med och antalet gener som varje art är starkt korrelerad med. Korrelationskoefficienter över 0,95 undersöktes. En korrelationskoefficient på 0,95 motsvarar ett korrigerat p-värde på $3 \cdot 10^{-6}$. En gen starkt korrelerad med ett stort antal arter kan antas vara en vanligt förekommande gen bland arter som ändras på samma sätt som genen i triclosan-gradienten. För de gener och arter till vilka flest starka korrelationer hittades undersöktes hur dessa varierar med koncentrationen triclosan.

2.3.2.2 Klustring av gener/arter

2.3.2.2.1 Hierarkisk klustring

Baserat på matrisen av korrelationskoefficienter för korrelationen mellan samtliga gener och arter användes hierarkisk klustring för att hitta grupper av arter respektive gener. Arterna representeras av rader i korrelationsmatrien och sådana som uppvisar liknande korrelation till samtliga gener (kolumner) i matrisen grupperades. Grupper av gener med liknande korrelation till samtliga arter bildades på motsvarande sätt genom att klustra kolumner med liknande korrelationsmönster. Hierarkisk klustring bygger på att varje observation (art eller gen) initialt tillhör ett separat kluster, varefter de observationer för vilka avståndet är kortast grupperas. Därefter identifieras det kortaste avståndet mellan två ytterligare observationer eller mellan det befintliga klustret och en observation. Processen fortgår fram till att alla observationer tillhör ett kluster. Hierarkisk klustring genererar ett dendrogram som kan liknas vid ett uppochnedvänt träd med grenar som representeras av kluster i vilka bladen utgör de enskilda observationerna och där längden på grenarna representerar hur olika klustren är. Euklidiskt avstånd mellan observationer tillämpades eftersom detta ger ett mått på hur lika korrelationsmönstren för gener alternativt arter är. För att uppskatta avståndet mellan två kluster användes det längsta avståndet mellan observationerna i klustren. Klustringen genomfördes med `hclust` i `stats`-paketet i R.

2.3.2.2.2 Identifiering av kluster

För att identifiera enskilda kluster i ett dendrogram kan olika metoder tillämpas. Ofta klipps dendrogrammet på en konstant höjd alternativt klipps för att generera ett förutbestämt antal kluster. Dessa metoder är inte optimala för att identifiera samtliga kluster korrekt eftersom det är svårt att bestämma lämpliga värden på höjden eller antal kluster, särskilt om klustringen resulterar i komplicerade dendrogram. För att identifiera kluster av gener respektive arter i denna studie användes `cutreeDynamic` i `dynamicTreeCut`-paketet (*49*). Denna metod tillämpar en dynamisk process baserad på en analys av hur grenarna ser ut och har visat sig kunna identifiera biologiskt relevanta genkluster (*49*). Processen börjar i botten av dendrogrammet

och i varje förgrening bedöms de två underliggande grenarna utifrån kriterier baserade på deras struktur. Kriterierna baseras bland annat på avståndet mellan förgreningen i fråga och förgreningarna längst ned i det eventuella klustret, avståndet från toppen av dendrogrammet till observationerna längst ned i klustret, samt det totala antalet observationer i klustret. Om båda grenarna uppfyller kraven för att utgöra enskilda kluster klipps dendrogrammet i förgreningen, och om inte fortsätter processen uppåt i dendrogrammet till nästa förgrening som bedöms på samma sätt. Observationer som inte verkar tillhöra något kluster kan tillskrivas det närmaste klustret genom en process liknande partitioning around medoids (PAM), men denna funktion användes inte vilket leder till att vissa observationer kan lämnas utan något kluster (`pamStage`=FALSE). Olika värden på parametern som kontrollerar hur små klustren kan bli (`deepSplit`) testades och för dendrogrammen som genererades för gener respektive arter gav detta inga större variationer i de identifierade klustren. För samtliga analyser användes därför deepSplit=2.

2.3.2.2.3 Analys av klustringsresultat

De identifierade klustren innehållande gener respektive arter analyserades i två steg. Nedan följer en beskrivning av analysen av klustren innehållande arter. Motsvarande analys gjordes för de identifierade klustren av gener. I det första steget identifierades vilka arter som återfinns i respektive kluster. Arter i samma kluster har liknande korrelation till samtliga gener i datasetet, och dessa arter varierar därför med triclosan-koncentrationen på liknande sätt. För att bestämma hur arterna påverkas av triclosan beräknades korrelationskoefficienter (Pearson) mellan förekomsten av varje art och triclosan-koncentrationen. I varje kluster togs därefter arter som ändras signifikant med triclosan-koncentrationen ut. Arter och gener som ändras signifikant identifierades i den tidigare genomförda `edgeR`-analysen (sektion 2.2). I nästa steg i analysen undersöktes varje kluster av signifikanta arter separat. Gener som ändras signifikant med koncentrationen triclosan och som är starkt korrelerade till arterna i respektive kluster identifierades därefter.

För att en gen ska anses vara starkt korrelerad med arterna i ett kluster sattes ett villkor på att summan av samtliga korrelationer mellan den genen och samtliga arter i klustret skulle överstiga (antal arter i klustret) $\times\, x$, där x kan ses som en genomsnittlig korrelation mellan genen i fråga och arterna i klustret. Två värden på villkoret x, 0,80 och 0,85, användes för att studera hur olika grader av korrelation påverkar generna som är sammanlänkade med arterna i ett kluster. En korrelationskoefficient på 0,80 motsvarar ett korrigerat p-värde på $2 \cdot 10^{-2}$ och 0,85 motsvarar ett korrigerat p-värde på $8 \cdot 10^{-3}$. Det mer restriktiva fallet där ett högt genomsnittligt värde på korrelationen mellan en gen och arterna i klustret krävs ger ett mindre antal gener som uppfyller villkoret och kan ge en indikation på vilka gener som kan antas vara starkast sammankopplade med arterna i klustret. För att visualisera klustren och generna som är starkt korrelerade med dessa användes `igraph`-paketet (*50*). Denna typ av graf ger även information om gener som är starkt korrelerade med arter i flera kluster.

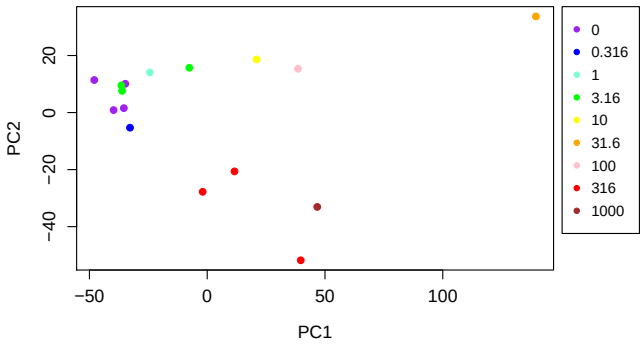
För varje identifierat kluster resulterar analysen i en lista av signifikanta arter som utgör klustret och en lista av signifikanta gener starkt korrelerade med dessa arter. De gener som uppfyllde villkoret på en genomsnittlig korrelation över 0,85 undersöktes vidare genom att identifiera deras funktionalitet. Vissa konstellationer bestående av artkluster och gener som arterna korrelerar starkt till valdes ut för en mer detaljerad analys där potentiella samband mellan generna och arterna utforskades.

I den motsvarande analysen av kluster av gener identifierades de gener i varje kluster som ändras signifikant med koncentrationen triclosan samt deras funktion. Signifikanta arter som är starkt korrelerade med generna i respektive kluster undersöktes med samma villkor på den genomsnittliga korrelationen som för klustren av arter. Klustren visualiserades genom att rita upp arter starkt korrelerade med generna i varje kluster. Arter korrelerade med en genomsnittlig korrelation över 0,85 med generna i varje kluster identifierades och utvalda konstellationer bestående av genkluster och arter starkt korrelerade till dessa undersöktes.

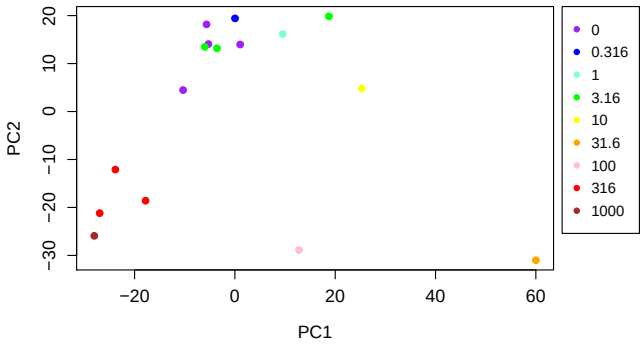
3 Resultat

3.1 Inledande analys av data

Principalkomponentanalys utfördes på båda dataseten för att visualisera proven som behandlats med olika koncentration av triclosan. Majoriteten av variansen beskrivs av de första principalkomponenterna och dessa kan därför användas för att visualisera datan. De två första principalkomponenterna för geners förekomst i proven beskriver 83 % av variansen i datan. I Figur 1a visas de två första principalkomponenterna för varje prov baserat på förekomsten av gener i proven. Figuren visar att proven som behandlats med de låga koncentrationerna av triclosan återfinns nära varandra medan proven med de högsta koncentrationerna bildar en separat grupp, men dock inte lika tät. Provet som behandlats med 31,6 nM triclosan befinner sig långt från de övriga proven vilket tyder på att det skiljer sig från de andra. För de koncentrationer för vilka replikat finns kan en en viss variation observeras mellan dessa. Störst variation mellan replikat observeras för proven som behandlats med 316 nM triclosan. Sammanfattningsvis kan det noteras att den andra principalkomponenten separerar proven i två grupper baserade på om de behandlats med någon av de två högsta koncentrationerna av triclosan eller någon av de lägre koncentrationerna. Den första principalkomponenten däremot separerar proven inom de två grupperna, dvs prov som har mer lika koncentration.



(a)



(b)

Figur 1: Visulisering av de två första principalkomponenterna för prov med olika koncentration triclosan baserat på förekomsten av (a) gener och (b) arter i proven. Prov som behandlats med olika koncentration av triclosan representeras av olika färg.

Motsvarande principalkomponenter baserade på förekomsten av arter i proven visas i Figur

1b. De två första principalkomponenterna för arters förekomst i proven förklarar endast 44 % av variansen. Trots detta tycks den andra principalkomponenten även här separera proven i två distinkta grupper, men här återfinns de prov som behandlats med de fyra högsta koncentrationerna av triclosan i den ena gruppen. Den första principalkomponenten separerar även i detta fall proven inom respektive grupp som identifierats av den andra komponenten. Provet som behandlats med 31,6 nM triclosan är även här mycket olikt de övriga proven.

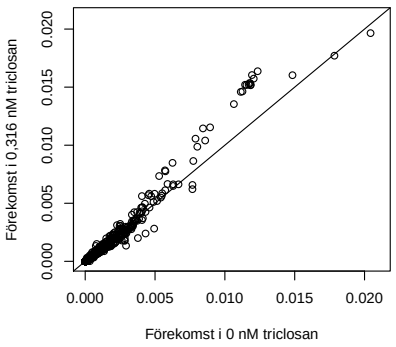
3.2 Analys del 1 - Undersökning av geners variation med koncentrationen triclosan

För att undersöka hur förekomsten av varje gen varierar med triclosan-koncentrationen plottades först medelvärdet av förekomsten i prov som inte behandlats med triclosan (kontroll) mot förekomsten i fyra prov som behandlats med olika koncentration av triclosan. För de koncentrationer för vilka replikat finns plottades medelvärden. Figur 2 visar de fyra plottarna med triclosan-koncentrationerna 0,316, 3,16, 31,6 och 316 nM. Punkter som ligger under den 45-gradiga linjen representerar gener vars förekomst minskar med ökad triclosan-koncentration och punkter ovanför linjen är gener som ökar med ökad triclosan-koncentration.

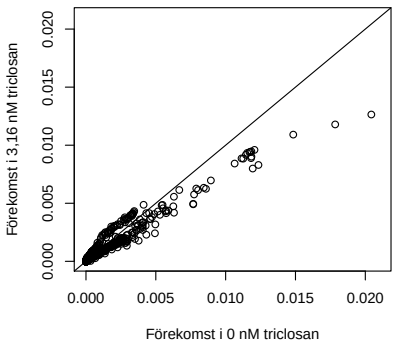
Vid de två lägsta koncentrationerna av triclosan syns inga stora avvikelser från linjen, men vid koncentrationen 31,6 nM syns tydliga trender. Vid denna koncentration har ett stort antal geners förekomst minskat jämfört med förekomsten i proven som inte behandlats med triclosan medan ett stort antal andra geners förekomst har ökat. Bland de gener som ökat i förekomst är majoriteten gener som kodar för diverse proteiner som transporterar ämnen in och ut ur cellen. Proteiner kan bidra till att arter som innehåller dessa klarar av triclosan genom att effektivt transportera ut ämnet innan det hinner skada cellen. Vid ännu högre koncentration av triclosan minskar dock förekomsten av dessa gener men många av dem är fortfarande högre än i proven som inte behandlats med triclosan.

Att en kraftig ökning i förekomst av vissa gener observeras vid en koncentration på 31,6 nM och att förekomsten av dessa gener vid högre koncentration generellt är lägre indikerar att upp till en viss koncentration triclosan finns det arter som kan hantera ämnet och som ökar i förekomst till följd av att många andra arter dör ut. Förekomsten av de gener som finns i arterna som ökar ökar därmed också. Vid högre koncentrationer av triclosan har inte alla dessa arter förmåga att överleva och förekomsten av generna minskar därför. Det kan också noteras att flera gener vars förekomst var mycket låg i provet som behandlats med 31,6 nM triclosan har ökat i förekomst i proven med 316 nM triclosan. Dessa gener kodar för proteiner som också exporterar ämnen ut ur cellen men även enzym involverade i fotosyntesen. Generna kan finnas i arter som tål triclosan men vars tillväxt vid de lägre koncentrationerna hämmats av många andra arter som frodats men när de arterna minskar i förekomst kan de här arterna börja växa igen.

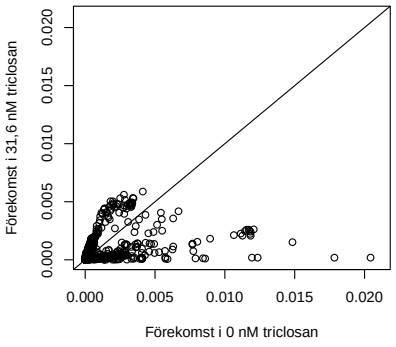
Signifikansen mellan proverna som behandlats med olika koncentration av triclosan testades genom att anpassa en GLM-modell till datan för varje gen respektive art. Korrigerade p-värden (FDR) beräknades och för en signifikansnivå på 5 % kunde 923 gener klassificeras som signfikanta vilket innebär att deras förekomst ändras signifikant med koncentrationen triclosan. Både gener vars förekomst ökar och minskar med ökad koncentration triclosan identifieras med denna metod. Om förekomsten av en gen ökar monotont med ökad koncentration av triclosan identifieras den som ökande, medan om förekomsten av en gen minskar monotont identifieras den som minskande. Det är dock viktigt att poängtera att gener vars förekomst ökar vid låga koncentrationer av triclosan jämfört med kontrollen men därefter minskar eventuellt inte identifieras som signifikanta i denna analys. Ett exempel på en gen som inte är signifikant till följd av detta är genen som kodar för det enzym som triclosan antas inhibera (enoyl-acyl carrier protein reductase; TIGR03151). I Figur 3 visas genens förekomst mot koncentrationen triclosan (log+1). Förekomsten av denna gen ökar inledningsvis med ökad koncentration av triclosan men efter 100 nM minskar dess förekomst. Detta resultat visar på att arter som innehåller denna gen inte klarar av höga koncentrationer av triclosan. Bland



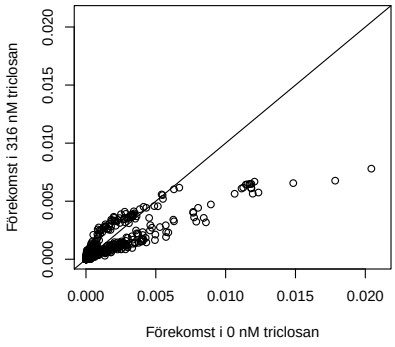
(a) 0,316 nM triclosan



(b) 3,16 nM triclosan



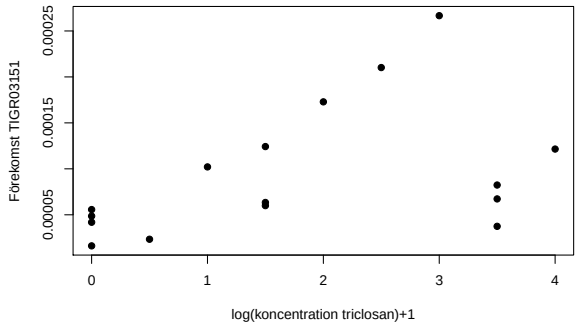
(c) 31,6 nM triclosan



(d) 316 nM triclosan

Figur 2: Scatter plots som visar hur förekomsten av gener ändras med ökad koncentration av triclosan jämfört med prov som inte behandlats med triclosan (på x-axeln). Medelvärdet av förekomsten plottades för koncentrationer där replikat finns.

de arter som har högst korrelation till denna gen, och därmed varierar på liknande sätt med triclosan, finns enbart bakterier som tillhör klassen Flavobacteria (fylum[1] Bacteroidetes).

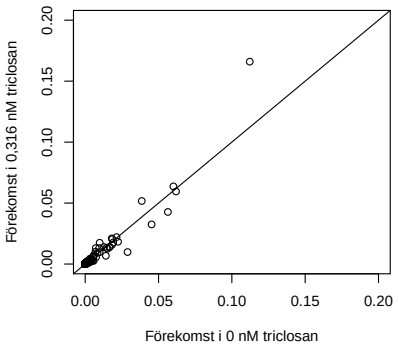


Figur 3: Förekomst av genen som kodar för enoyl-acyl carrier protein reductase (TIGR03151) plottad mot koncentrationen triclosan.

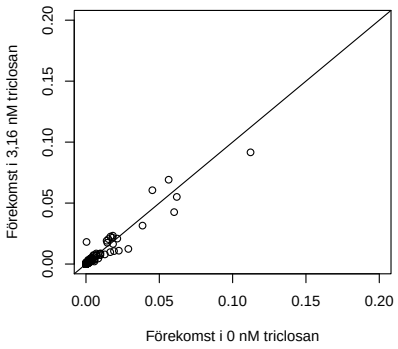
[1]fylum är en taxonomisk rang som i sin tur är indelad i klasser

3.3 Analys del 2 - Undersökning av arters variation med koncentrationen triclosan

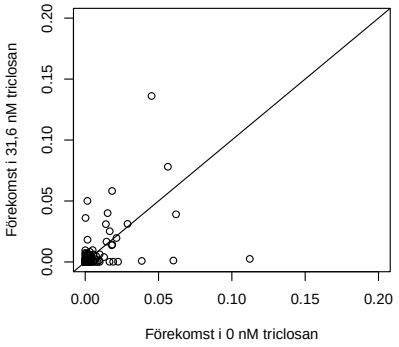
På motsvarande sätt som för gener plottades förekomsten av varje art i kontrollproven mot förekomsten i fyra prov som behandlats med olika koncentration av triclosan (Figur 4). Inzoomade plottar finns i Figur 11 i Appendix. Liknande trender som för generna kan ses. Vid 31,6 nM triclosan har förekomsten av ett fåtal arter ökat kraftigt medan ett stort antal arter ökat men inte fullt så mycket, jämfört med proven som inte behandlats med triclosan. Vid den högre koncentrationen triclosan minskar förkomsten för många arter, på liknande sätt som observerats för gener. I analysen för att hitta signifikanta arter var det korrigerade p-värdet (FDR) under 5 % för 658 arter och dessa antas därför variera signifikant med koncentrationen triclosan.



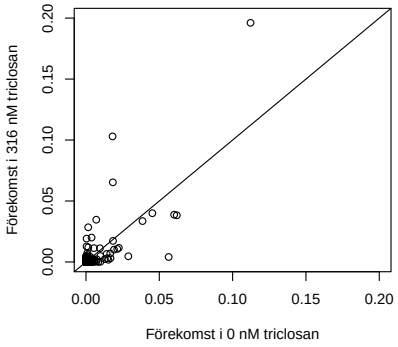
(a) 0,316 nM triclosan



(b) 3,16 nM triclosan



(c) 31,6 nM triclosan



(d) 316 nM triclosan

Figur 4: Scatter plots som visar hur förekomsten av arter ändras med ökad koncentration av triclosan jämfört med prov som inte behandlats med triclosan (på x-axeln). Medelvärdet av förekomsten plottades för koncentrationer där replikat finns.

3.4 Analys del 3 - Korrelation mellan gener och arter

3.4.1 Inledande analys av korrelation mellan gener och arter

Samband mellan förekomst av gener och arter i prov där koncentrationen triclosan varierer undersöktes genom att korrelationskoefficienter för samtliga kombinationer av gener och arter beräknades. Bland de gener och arter vars förekomst ändras signifikant med koncentrationen triclosan undersöktes först vilka gener som är starkt korrelerade med störst antal arter. För att anses vara starkt korrelerad sattes ett villkor på en korrelationskoefficient över 0,95.

Tabell 2 visar de 10 gener som är starkt korrelerade med störst antal arter, samt respektive gens funktion. Som mest korrelerar en gen till 25 olika arter. Förekomsten av samtliga gener i denna lista ökar med ökad koncentration av triclosan. Tabell 3 visar de 10 arter som är starkt korrelerade ($r > 0,95$) med störst antal gener. Hela 43 gener är starkt korrelerade med den art som korrelerar till flest gener. Tre av arterna i listan är eukaryoter (Diatomea; kiselalger) medan resten är bakterier som tillhör fylum Proteobacteria (klassen Gammaproteobacteria). Förekomsten av samtliga arter i listan ökar med ökad triclosan-koncentration.

Tabell 2: Gener starkt korrelerade med störst antal arter.

Gen	Funktion	Antal arter
TIGR02922	TIGR02922: TIGR02922 family protein	25
TIGR03758	conj_TIGR03758: integrating conjugative element protein, PFL_4701 family	21
TIGR01752	flav_long: flavodoxin	17
TIGR02205	septum_zipA: cell division protein ZipA	16
TIGR01753	flav_short: flavodoxin	16
TIGR02548	casB_cse2: CRISPR type I-E/ECOLI-associated protein CasB/Cse2	14
TIGR02443	TIGR02443: conserved hypothetical protein	14
TIGR01109	Na_pump_decarbB: sodium ion-translocating decarboxylase, beta subunit	13
TIGR00013	taut: 4-oxalocrotonate tautomerase family enzyme	13
TIGR03907	QH_beta: quinohemoprotein amine dehydrogenase, beta subunit	13

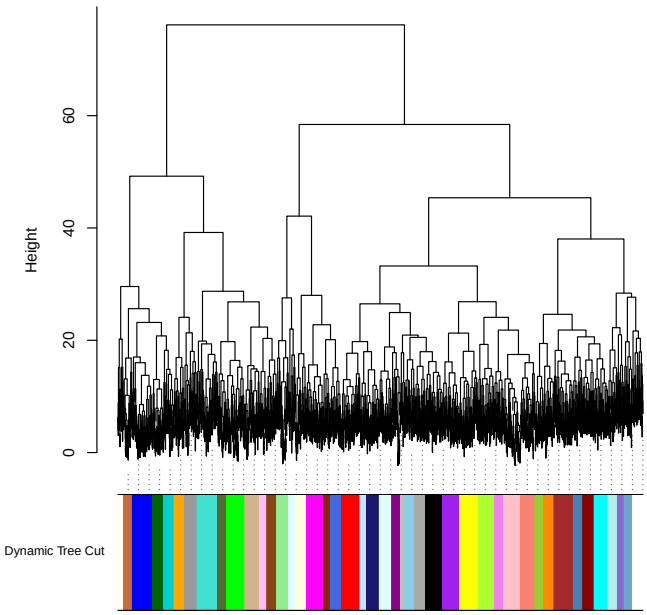
Tabell 3: Arter starkt korrelerade med störst antal gener.

Art	Antal gener
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oceanospirillaceae, Oceanobacter, Unclassified Oceanobacter	43
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Eionea, Eionea nigra	36
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oceanospirillaceae, Oceanobacter, Thalassolituus sp. H61	33
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Unclassified Alteromonadaceae,	31
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Entomoneis	29
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Candidatus Endobugula, Candidatus Endobugula glebosa	24
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Alteromonas, Unclassified Alteromonas	24
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Placoneis	23
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Unclassified Alteromonadales, ,	23
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Gomphonema	17

3.4.2 Konstellationer av arter och gener skapade baserat på klustring av arter

I syfte att kunna studera mönster och samband mellan arter och gener i de två dataseten skapades konstellationer[2] av arter och gener baserat endast på dess korrelationsmönster i triclosan-gradienten. Konstellationer av arter och gener genererade baserat på klustring av arter skapades i två steg. I första steget skapades kluster av arter baserat på hur de korrelerar med samtliga gener i datasetet. I andra steget formades grupper av gener tillhörande varje artkluster, där gener identifierades efter kriteriet att de korrelerar starkt med samtliga arter i respektive kluster.

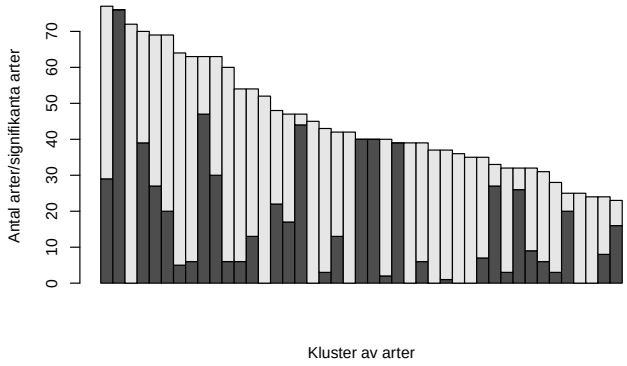
I det första steget utfördes hierarkisk klustring i syfte att forma grupper av arter med liknande korrelation till samtliga gener i datasetet. Arter förekommer som rader i korrelationsmatrisen innehållande samtliga kombinationer av korrelationskoefficienter mellan gener och arter. Klustringsalgoritmen identifierar rader med kortast avstånd, vilket i det här fallet innebär mest liknande korrelationskoefficienter till generna, och skapar grupper av dessa. Arter som återfinns i samma kluster uppvisar således samma typ av korrelation till samtliga gener (kolumner) i datasetet. Resultatet från hierarkisk klustring av arter visas i Figur 5 i form av ett dendrogram. De 43 kluster av arter som identifierades representeras av olika färg i dendogrammet. Därefter sorterades icke signifikanta arter bort ur klustren. Fyra kluster innehåller enbart signifikanta arter medan tio kluster inte innehåller några signifikanta arter alls. Figur 6 visar fördelningen av det totala antalet arter i varje kluster och antalet arter som ändras signifikant med koncentrationen av triclosan.



Figur 5: Klustring av arter baserad på korrelation till gener visualiserad i form av ett dendrogram. Färgblocken representerar kluster identifierade med `dynamicTreeCut`.

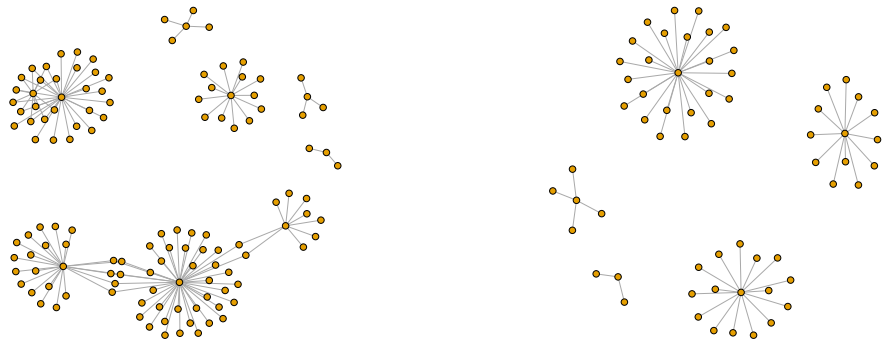
I det andra steget i skapandet av art-gen-konstellationerna identifierades gener som ändras signifikant med triclosan och som är starkt korrelerade till de signifikanta arterna i respektive kluster. Generna identifierades genom att använda ett villkor på en genomsnittlig korrelation till arterna i klustret. Två olika värden på villkoret för den genomsnittliga korrelationen undersöktes för att se vilken effekt detta har på vilka gener som kan identifieras. Figur 7 visar grafer som illustrerar gener som är korrelerade till arterna i varje kluster med en genomsnittlig korrelation på över 0,80 respektive 0,85. Generna länkas till respektive kluster

[2]en konstellation definieras här som ett kluster av arter eller gener samt de gener eller arter som är starkt korrelerade till dessa



Figur 6: Histogram som visar fördelningen av det totala antalet arter och antalet signifikanta arter i varje kluster. Den totala höjden av varje stapel representerar det totala antalet arter i ett kluster och den svarta stapeln representerar andelen signifikanta arter.

med en edge. Med villkoret 0,80 innehåller nio kluster arter som är starkt korrelerade med minst en gen, men för det högre villkoret minskar antalet gener som är starkt korrelerade till arterna och enbart fem kluster innehåller arter som är starkt korrelerade till några gener. I fallet med det lägre villkoret är vissa gener starkt korrelerade till flera olika kluster (Figur 7a) men dessa försvinner då villkoret höjs.



(a) (b)

Figur 7: Grafer som visualiserar gener som är starkt korrelerade till kluster av arter med krav på en genomsnittlig korrelation på minst (a) 0,80 och (b) 0,85 till arterna i ett kluster. Figuren i (a) visar 9 kluster och (b) 5 kluster av arter.

Bland de skapade konstellationerna, bestående av artkluster och tillhörande gener, valdes tre ut för vidare analys. I denna analys undersöktes gener med en genomsnittlig korrelation över 0,85 till arterna i respektive kluster. Konstellationerna valdes ut enligt något eller några av följande tre kriterier: 1) hög korrelationskoefficient mellan förekomst av arter och koncentration triclosan (pga att korrelationsrelationen är transitiv har även de tillhörande generna hög korrelation mellan förekomst och koncentration triclosan), 2) stor homogenitet av arter i klustret, eller 3) stor homogenitet bland de gener som korrelerar starkt med arterna. Arter och gener i de tre utvalda konstellationerna anges i Tabell 4-9 i Appendix.

Den första konstellationen innehåller sex stycken signifikanta arter, varav tre är prokaryoter

(två som tillhör fylum Lentisphaerae och en Proteobacteria) samt tre eukaryoter (två som tillhör fylum Metazoa och en Chromaleovata). 24 gener identifierades som starkt korrelerande till samtliga signifikanta arter i detta kluster. Både arter och gener har en svag negativ korrelation med triclosan och ett medeltal av korrelationskoefficienten med triclosan ligger på -0,20. Gruppen av gener i denna konstellation är mycket homogen och kodar för funktioner som är associerade med horisontell DNA-överföring. Gener med denna funktionalitet förekommer mycket sparsamt i de övriga konstellationerna och inte i någon konstellation där arter och gener har hög korrelation med triclosan. Det faktum att många gener av den här typen har en negativ korrelation med triclosan indikerar att de inte är vanligt förekommande hos arter som lyckas etablera sig i höga koncentrationer av triclosan. Det ligger nära till hands att spekulera i att arter som klarar att etablera sig i höga koncentrationer av triclosan har energikrävande strategier för att kunna hantera triclosan, som exempelvis blockering, nedbrytning och transport ut ur cellen. Organismerna skulle då kunna spara energi genom att göra sig av med plasmider då horisontell överföring av DNA inte är nödvändig för överlevnad.

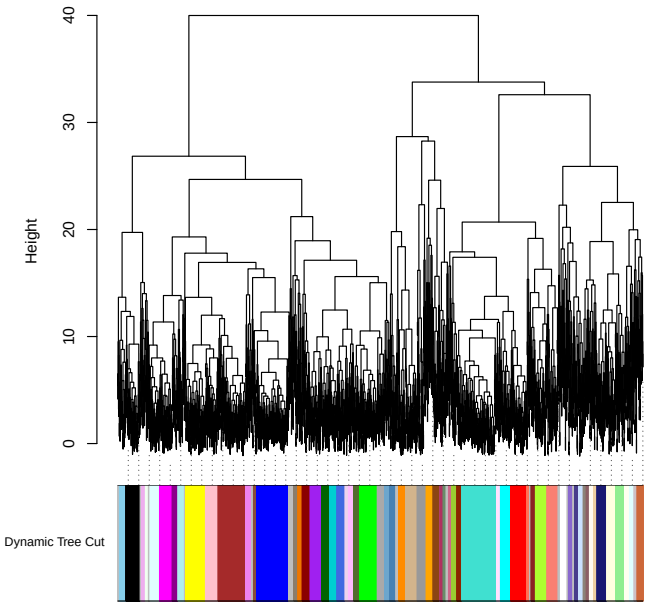
Den andra konstellationen utgörs av en homogen grupp av 40 signifikanta arter och 15 tillhörande gener. De flesta bakterier i gruppen tillhör klassen Gammaproteobacteria (fylum Proteobacteria), och utöver dessa ingår fem eukaryoter. Värt att notera här är att samtliga dessa arter återfinns bland de tio arter som är korrelerade med störst antal gener (Tabell 3). Av de 15 gener som identifierades som starkt korrelerade till arterna i klustret är sex stycken associerade med CRISPR-systemet (clustered regularly interspaced short palindromic repeats). Dessa gener kodar för proteiner involverade i bakteriers immunförsvar som exempelvis känner igen och bryter ner inkommande DNA. Generna och arterna som hör till denna konstellation har höga korrelationskoefficienter till triclosan, där CRISPR-generna ligger högst på 0,98-0,99.

Den tredje konstellationen består av 26 arter och 12 gener. Artklustret, som är något spretigare i sin sammansättning än i de två tidigare konstellationerna, innehåller en art som tillhör fylum Actinobacteria (klassen Actinobacteria), 3 Bacteroidetes (klassen Flavobacteria), 10 Proteobacteria (klassen Gammaproteobacteria) samt 12 eukaryoter som tillhör 5 olika fyla. Även i denna gengrupp återfinns CRISPR-gener och flera gener som kodar för transport över membran. Gener och arter i konstellationen har stark korrelation till triclosan. Då även dessa CRISPR-gener har mycket stark korrelation till triclosan går det att spekulera i att förekomsten av dessa gener kan användas som markör för att de associerade arterna lyckas etablera sig i höga koncentrationer av triclosan.

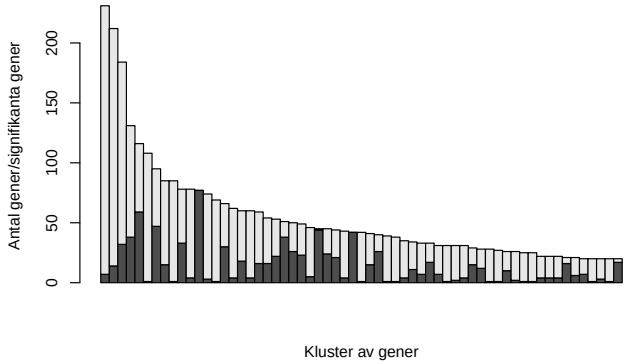
3.4.3 Konstellationer av arter och gener skapade baserat på klustring av gener

Skapandet av konstellationer av arter och gener innehållande kluster av gener med tillhörande starkt korrelerade arter genomfördes på motsvarande sätt. Först skapades kluster av gener som uppvisar liknande korrelationsmönster med samtliga arter i datasetet. I Figur 8 illustreras ett dendrogram med de 61 identifierade klustren av gener representerade i olika färg. I dessa 61 genkluster sorterades därefter icke signifikanta gener bort. Då 13 kluster helt saknade signifikanta gener erhölls totalt 48 kluster av gener. Figur 9 visar fördelningen av det totala antalet gener i varje kluster och antalet gener som ändras signifikant med ändringar i koncentrationen av triclosan. Figur 10 visar grafer som illustrerar signifikanta arter som är starkt korrelerade till de signifikanta generna i varje kluster för en genomsnittlig korrelation på över 0,80 respektive 0,85.

Bland de bildade konstellationerna av genkluster med tillhörande arter med en genomsnittlig korrelation över 0,85 till generna, valdes tre ut för vidare analys enligt samma kriterier som tidigare. Två av konstellationerna valdes för att de signifikanta generna i dessa hade en hög korrelationskoefficient till triclosan och det tredje valdes ut på grund av homogeniteten i genfunktionerna. I Tabell 10-15 i Appendix anges generna och arterna i de tre utvalda konstellationerna.



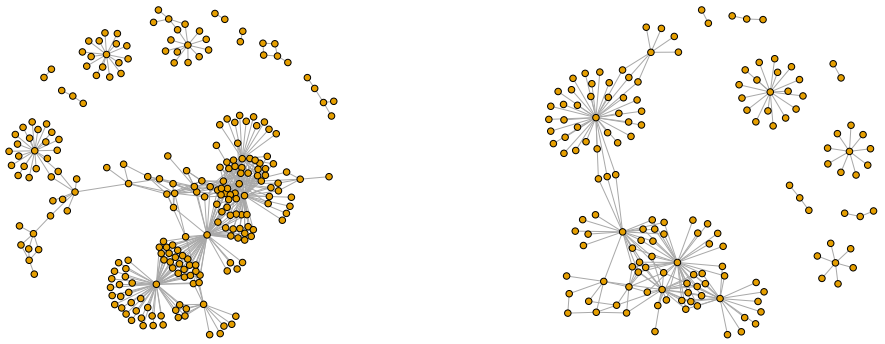
Figur 8: Klustring av gener baserad på korrelation till arter visualiserad i form av ett dendrogram. Färgblocken representerar kluster identifierade med `dynamicTreeCut`.



Figur 9: Histogram som visar fördelningen av det totala antalet gener och antalet signifikanta gener i varje kluster. Den totala höjden av varje stapel representerar det totala antalet gener i ett kluster och den svarta stapeln representerar andelen signifikanta gener.

Den första konstellationen innehåller 47 gener vars funktioner till största delen är olika typer av transport över membran. Tre av generna kodar för proteiner involverade i typ II-sekretionssystemet som transporterar olika proteiner över cellmembranet ut i periplasman hos gramnegativa bakterier. Vidare förekommer gener som kodar för enzymer som transporterar metaboliter in i cellen. De 36 arter som är starkt korrelerade till detta genkluster bildar en mycket homogen grupp bestående av arter som enbart tillhör klassen Gammaproteobacteria (fylum Proteobacteria), samt eukaryoter av klassen Diatomea. Flera av dessa arter återfinns i den lista av arter som har störst antal starka korrelationer till generna i datasetet (Tabell 3). Generna och arterna i den här konstellationen har höga korrelationskoefficienter till triclosan (0,71-0,89).

Den andra konstellationen består av 18 gener, som samtliga kodar för någon form av horison-



(a) (b)

Figur 10: Grafer som visualiserar arter som är starkt korrelerade till kluster av gener med krav på en genomsnittlig korrelation på minst (a) 0,80 och (b) 0,85 till generna i ett kluster. Figuren i (a) visar 32 kluster och (b) 21 kluster av gener.

tell DNA-överföring, samt 9 tillhörande arter. Gener och arter i denna konstellation har en svag negativ korrelation till triclosan och generna sitter i huvudsak på plasmider. De 9 arter som är starkt korrelerade till den här gruppen av gener är prokaryoter som tillhör tre olika fyla (Cyanobacteria, Lentisphaerae och Proteobacteria) samt eukaryoter som tillhör fylum Metazoa och Alveolata. Den här konstellationen är mycket lik den första av de konstellationer som beskrivs i sektion 3.4.2. Den konstellationen bestod av ett artkluster innehållande sex stycken arter varav alla även ingår i denna konstellations artgrupp. Även gengrupperna visar på stor likhet och de 18 gener som ingår i denna konstellation är en delmängd av de 24 gener som finns i den andra konstellationen.

Denna kombination av gener och arter visar alltså på särskilt stor symmetri i och med att samma grupperingar av gener och arter observeras oberoende av om det är gener eller arter som klustras i det första steget i analysen. Denna symmetri beror på att samtliga korrelationskoefficienter i den delmatris som utgörs av dessa arter och gener måste vara relativt lika. Den observerade symmetrin innebär att varken gener eller arter i denna kombination förekommer i någon av de övriga gen-art-konstellationerna vilket i denna analys innebär att dessa gener endast är relaterade till precis de här arterna och vice versa. Då dessa arter har en negativ korrelation till triclosan innebär det att de har svårt att etablera sig i höga koncentrationer av triclosan vilket även skulle kunna innebära att den här genfunktionen möjligen saknas hos arter som klarar det.

Den tredje konstellationen innehåller 23 gener och 19 arter som har en hög korrelation till triclosan (0,80-0,98). Några gener i denna konstellation är associerade med lipoproteiner men det som är mest noterbart är den höga andelen CRISPR-gener. Gruppen av arter som korrelerar starkt till generna i denna grupp består av 19 arter som tillhör klassen Gammaproteobacteria (fylum Proteobacteria) och 4 eukaryoter. Den här konstellationen uppvisar stor symmetri med den tredje konstellationen som diskuterades i sektion 3.4.2, som också hade en hög andel CRISPR-gener.

Nedan följer en sammanfattning av observationer i de analyserade gen-art-konstellationerna i 3.4.2 och 3.4.3.

1. Endast ett litet antal gener i konstellationerna kodar för basala funktioner som exempelvis central metabolism och DNA-replikation. De flesta gener kodar för mer specialiserade funktioner som är specifika för vissa arter. Orsaken till detta kan vara att basala gener förekommer hos så många arter att förekomsten av dessa inte ändras signifikant i triclosan-gradienten.

2. En vanlig genfunktion i flera av konstellationerna är olika typer av transport över membran. Vissa konstellationer innehåller till största delen gener som kodar för transportrelaterade proteiner.

3. I de konstellationer av gener och arter som har starkast korrelation till triclosan förekommer CRISPR-gener associerade med bakteriers immunförsvar.

4. Det förekommer en viss symmetri mellan gen- och artklustren vilket innebär att generna och arterna i dessa konstellationer inte ingår i några andra konstellationer.

5. Plasmid-DNA och gener som kodar för proteiner involverade i horisontell överföring av DNA förekommer i huvudsak endast i en konstellation. Denna konstellation har en hög grad av symmetri och korrelationen med triclosan är negativ.

6. Grupperna av arter som förekommer i konstellationerna är mycket homogena, och detta oberoende av om konstellationerna skapas genom att klustra gener eller arter.

4 Diskussion

Syftet med analysen var att studera eventuella samband mellan gener och arter i metagenomikdata bestående av förekomst av gener och arter i mikrobsamhällen som härstammar från havsvatten som behandlats med olika koncentrationer av den antibakteriella substansen triclosan. En statistisk metod baserad på korrelation mellan förekomsten av gener och arter i de olika mikrobsamhällena utvecklades. Konstellationer av gener och arter som varierar signifikant och svarar på samma sätt i triclosan-gradienten bildades och dessa analyserades i syfte att identifiera genfunktionaliteter kopplade till specifika grupper av arter.

Då det genetiska materialet som ligger till grund för analysen kommer från prov extraherade från miljön finns ett antal variabler som kan påverka resultaten och därmed de slutsatser som bygger på dessa. Havsvatten innehåller en viss mängd triclosan och det bör noteras att den koncentrationen av triclosan som hänvisas till i denna rapport avser mängden som proven behandlats med. Kontinuerliga mätningar under experimentets fortgång visar också att en viss variation i koncentrationen finns till följd av det kontinuerliga inflödet av havsvatten till varje akvarium vilket gör att en konstant koncentration av triclosan är svår att upprätthålla (*22*). Dessa faktorer bidrar till att den uppmätta koncentrationen i de prover som behandlats med de två längsta koncentrationerna av triclosan överlappar med kontrollproverna till vilka ingen triclosan adderats. Variationen i koncentration kan därför resultera i svårigheter att bedöma om de observerade likheterna i art- och gensammansättning mellan dessa prover beror på att mikrobsamhällena inte påverkas av låga halter av triclosan eller om koncentrationen i själva verket är mer eller mindre densamma i proven. I de prov som behandlats med högre koncentration triclosan noterades dock inget överlapp mellan de uppmätta koncentrationerna.

Ytterligare en variabel som bör tas i beaktande när prov hämtas direkt från miljön är den biologiska variationen som naturligt påverkar förekomsten av gener och arter i vad som kan antas vara identiska mikrobsamhällen. Detta tillsammans med teknisk variation i form av felkällor exempelvis relaterade till bearbetningen av proven och DNA-sevenseringen samt bristen på replikat för flera av koncentrationerna gör att säkerheten i resultaten kan diskuteras. De analyser som genomfördes i denna studie för att hitta trender och mönster i datan baserades dock på samtliga prover och det kan antas att de variationer som eventuellt ger mindre avvikelser mellan proven inte påverkar resultaten markant. För att identifiera gener och arter vars förekomst ändras signifikant med triclosan-koncentrationen användes en regressionsmodell med samtliga prov och beräkningarna av korrelationskoefficienter baserades på förekomsten av gener respektive arter i samtliga prov. Det bör dock påpekas att det är viktigt att följa upp studier som denna med ytterligare undersökningar innehållande exempelvis fler replikat för att bekräfta resultaten och kunna dra mer solida slutsatser.

Det centrala i det här projektet är utvecklandet av en metod för att skapa konstellationer av gener och arter utifrån de två dataseten med information om hur förekomst av dessa varierar i triclosan-gradienten. Den relation som ligger till grund för vilka kombinationer av arter och gener som bildas är arternas och genernas korrelation till varandra i triclosan-gradienten. Gener och arter som svarar på samma sätt vid ändringar i triclosan-koncentrationen är starkt korrelerade till varandra. I det första steget i analysen användes hierarkisk klustring för att bilda kluster av arter med liknande korrelation till samtliga gener i datasetet. På samma sätt skapades kluster av gener med liknande korrelation till samtliga arter. Genom att basera klustringen på korrelationsmönstret till samtliga gener eller arter i datan togs all tillgänglig information med i beräkningarna, vilket antas ge stabila kluster av gener respektive arter. Redan i detta steg, då dessa grupper av gener respektive arter undersöks separat, fås indikationer på att metoden som används resulterar i grupper där intressanta samband eventuellt kan påvisas. Grupperna av arter är överlag mycket homogena. I vissa av dessa förekommer endast arter av en taxonomisk klass medan andra innehåller arter från ett fåtal olika fyla. Eventuella samband mellan olika taxonomiska fyla/klasser som observerades i samma grupper av arter studerades inte vidare men detta kan vara av intresse för fortsatta studier. I grupperna av gener observerades spännande samband när det gäller genernas funktionalitet.

I vissa grupper förekommer exempelvis gener som i det nästan uteslutande är relaterade till transport över membran och i andra grupper är samtliga gener lokaliserade på plasmider. Vissa gener förekommer frekvent och i hög andel i flera kluster, som exempelvis CRISPR-gener och gener involverade i fettsyntes.

Gener som är vanligt förekommande hos alla arter, exempelvis sådana som kodar för proteiner involverade i central metabolism och DNA-replikation, förekommer mycket sällan i de grupper som bildades. En trolig förklaring till det är att den relativa förekomsten av vanligt förekommande gener inte ändras signifikant med koncentrationen triclosan eftersom de finns hos arter som både ökar och minskar i triclosan-gradienten. Vi tolkar det som att de flesta gener som identifieras i grupperna troligen är mer specialiserade och inte förekommer hos många olika arter. Vidare kan gener och arter vars förekomst initialt ökar med ökad koncentration av triclosan och som sedan minskar vid högre koncentrationer vanligtvis inte identifieras som signifikanta i analysen och därför observeras dessa inte heller i några av de bildade konstellationerna. Ett exempel på ett sådant fall är den gen som kodar för enzymet som triclosan inhiberar. Förekomsten av denna gen ökar upp till en koncentration på 100 nM men därefter minskar förekomsten kraftigt och därför identifieras den inte som en gen som ändras signifikant med triclosan. De gener och arter vars förekomst ökar eller minskar monotont med ökad koncentration av triclosan är därmed enklast att identifiera som signifikanta eftersom en GLM-modell som letar efter linjära samband anpassades i denna analys.

I ett andra steg i analysen formades konstellationer av gener och arter som har hög korrelation till varandra i triclosan-gradienten. Till grupperna av arter identifierades de gener som har en stark korrelation till arterna i respektive kluster och på samma sätt bildades konstellationer genom att utgå ifrån kluster av gener och därefter identifiera arter som korrelerar starkt till dessa. I urvalet av arter och gener sorterades dessa ut efter kriteriet på hur starkt de korrelerar till arterna respektive generna i klustret. Vid ett lägre villkor associerar flera arter/gener med ett antal olika kluster medan vid ett högre villkor är generna/arterna associerade med betydligt färre kluster. Vid ett mycket högt villkor finns inga associationer mellan arter och gener i konstellationerna utan de individuella arterna och generna förekommer endast i en konstellation. Det hade varit bekymmersamt om det motsatta hade observerats. Till följd av transitiviteten i korrelationsrelationen skulle det innebära att om en art korrelerar starkt med två olika genkluster måste samtliga arter som korrelerar starkt med något av dessa två även korrelera starkt med det andra, dvs de två klustrena skulle då vara mycket lika.

Flera av de konstellationer som bildats uppför sig symmetriskt i det avseende att samma kombinationer mellan grupper av gener och grupper av arter hittas oberoende av om gener eller arter klustras i det första steget. Detta betyder att de individuella arterna och generna inte förekommer i flera olika konstellationer. Symmetrin innebär även att de arter som väljs ut till ett genkluster har mycket lika korrelationskoefficienter till samtliga gener i klustret och vice versa. Hade arterna haft varierande koefficienter hade de antagligen hamnat i olika grupper när klustringen utförts med avseende på dessa. Det faktum att det endast verkar finnas en association mellan en viss grupp arter och en viss grupp gener vid ett högre villkor och att konstellationerna uppför sig symmetriskt indikerar att arter och gener som korrelerar starkt till varandra formar ”slutna” system, dvs system där arter och gener relateras till varandra i högst en konstellation (1:1 förhållande).

En annan intressant observation är att de konstellationer av arter och gener som bildas förefaller vara mycket stabila. Många olika parametrar i metoden testades under projektets gång. Exempelvis klipptes dendogrammen på olika sätt, arter och gener som inte korrelerar starkt till några andra gener och arter samt icke-signifikanta gener och arter filtrerades bort i olika skeden av processen. Kombinationerna av gener och arter i de resulterande konstellationerna förändrades inte märkvärt och inte heller de trender och mönster som observerades. Mycket talar för att den metod som utvecklats för att skapa konstellationer av gener och arter för att hitta samband verkar resultera i stabila grupper som skulle kunna innehålla sanna associationer vilka kan vara värda att undersöka vidare.

Sammanfattningsvis kan vi ställa oss följande frågor: Vilka säkra slutsater kan vi dra av de samband vi ser? Vad kan vi säga om generna och arterna som ingår i en viss konstellation? Går det att säga något om hurvida det föreligger en sann association eller är den enda säkra slutsaten vi kan dra angående gener och arter i en viss konstellation att de uppför sig på ett liknande sätt i den aktuella triclosan-gradienten?

Vi kan med säkerhet säga att de arter som är associerade med ett visst genkluster i en konstellation inte behöver innehålla en enda av generna i klustret. Ett exempel på det såg vi i en av konstellationerna som innehöll CRIPSR-gener. Bland de arter som ingick i konstellationen förekom det eukaryota arter och dessa innehåller inte CRISPR-gener. Vidare är det så att arter som inte ingår i en konstellation mycket väl kan ha en sann association till någon eller några av generna i densamma. Ett möjligt scenario för detta skulle kunna representeras av en gen som finns i en art som inte lyckas etablera sig i de högsta koncentrationerna av triclosan, men samtidigt finns genen även i en annan art som istället tar över i de högre koncentrationerna. Då kan det se ut som att båda arterna har en låg korrelation med genen och med triclosan, medan genen har en hög korrelation med triclosan.

Vi har inte kunnat visa att det går att dra några absolut säkra slutsatser om vilka gener och arter som har sanna associationer men trots det innehåller de resulterande konstellationerna, framtagna endast baserat på korrelation, intressanta samband och mönster som kan vara värda att studera ytterligare. Att konstellationerna uppträder symmetriskt, dvs att vi formar samma grupper oavsett om vi börjar med att klustra gener eller arter, indikerar att det kan finnas stabila relationer som kan tyda på sanna associationer. Grader av symmetri varierade bland konstellationerna. Mest symmetrisk var den konstellation med gener som kodar för horisontell DNA-överföring där nästan helt identiska grupper bildades. Även den konstellation där en hög andel av generna bestod av CRISPR-gener uppvisade stor symmetri. Vi skulle kunna tolka den här symmetrin som att associationerna är extra starka i dessa grupper. Det är också i just dessa grupper som vi har hittat det vi tolkar som de mest intressanta sambanden. De CRISPR-gener som finns i dessa konstellationer har mycket höga korrelationskoefficienter med triclosan vilket innebär att arterna som förekommer i samma konstellationer klarar att etablera sig i höga koncentrationer av triclosan. Eventuellt skulle CRISPR-genernas förekomst kunna användas som markör för konstellationer vars arter lyckas etablera sig i triclosan. Både när det gäller horisontell överföring av DNA och transport över membran ser vi att dessa genfunktionaliteter ändras över triclosan-gradienten. När det gäller överföring av DNA minskar den funktionaliteten i triclosan-gradienten vilket innebär att arterna som associeras med den funktionaliteten skulle ha svårt att etablera sig i höga koncentrationer. Man skulle kunna spekulera vidare i att arter som lyckas etablera sig i höga koncentrationer av triclosan har gjort sig av med plasmid-DNA i syfte att spara energi. Det kan spekuleras i om detta i så fall saktar ner evolutionen så att resistens mot triclosan tar längre tid för arterna att utveckla.

Flera aspekter relaterade till detta projekt har inte undersökts men skulle vara intressant att studera närmare. I studien undersöktes exempelvis nästan enbart trender i grupper av gener och arter som ökar med ökad koncentration av triclosan. Eventuellt skulle andra intressanta samband kunna hittas då kluster av gener och arter som minskar med ökad triclosan-koncentration undersöks. Vidare analyserades endast relativt stora kluster då det var ett mycket litet antal gener och arter associerade med mindre kluster. Riktigt små kluster och förhållandet och samband mellan arter och gener i dessa skulle vara intressant analysera. Även gener och arter som inte klassificeras till något kluster, till följd av att deras korrelationsmönster till generna/arterna skiljer sig mycket åt från de andra, skulle kunna analyseras närmare. Metoder för att beräkna sannolikheter för sanna associationer mellan data i olika dataset skulle också kunna studeras.

Slutligen kan vi säga att den framtagna metoden för att koppla ihop information från två dataset, med separat data om arter och geners förekomst i en triclosan-gradient, resulterade i vad vi uppfattar som stabila konstellationer av arter och gener. Det vi kunde hitta i konstellationerna var genfunktionaliteter som ändrades signifikant med triclosan, kopplade

till artgrupper som ändrades på samma sätt. Det här projektet kan därmed ses som en utforskande studie där metoder för att kombinera information från olika dataset utvecklas och evalueras. I takt med att nya sekvenseringsmetoder kontinuerligt utvecklas produceras allt större mängder data och därmed finns ett behov av metoder som kan analysera denna så effektivt som möjligt. I syfte att extrahera så mycket information som möjligt från datan kan därför metoder för att undersöka olika kombinationer av dataset undersökas vilket skulle kunna öppna upp för upptäckter om olika typer av samband som tidigare varit okända.

Referenser

(*1*) Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., och Goodman, R. M., (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology 5*, R245–R249.

(*2*) Rondon, M. R., August, P. R., Bettermann, A. D., Brady, S. F., Grossman, T. H., Liles, M. R., Loiacono, K. A., Lynch, B. A., MacNeil, I. A., Minor, C., m. fl. (2000). Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Applied and environmental microbiology 66*, 2541–2547.

(*3*) Roesch, L. F., Fulthorpe, R. R., Riva, A., Casella, G., Hadwin, A. K., Kent, A. D., Daroub, S. H., Camargo, F. A., Farmerie, W. G., och Triplett, E. W., (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME journal 1*, 283–290.

(*4*) Schloss, P. D., och Handelsman, J., (2005). Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome biology 6*, 229.

(*5*) Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C., Knight, R., och Gordon, J. I., (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature 449*, 804.

(*6*) Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., Bonazzi, V., McEwen, J. E., Wetterstrand, K. A., Deal, C., m. fl. (2009). The NIH human microbiome project. *Genome research 19*, 2317–2323.

(*7*) Gilbert, J. A., Jansson, J. K., och Knight, R., (2014). The Earth Microbiome project: successes and aspirations. *BMC biology 12*, 69.

(*8*) Manichanh, C., Rigottier-Gois, L., Bonnaud, E., Gloux, K., Pelletier, E., Frangeul, L., Nalin, R., Jarrin, C., Chardon, P., Marteau, P., m. fl. (2006). Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach. *Gut 55*, 205–211.

(*9*) Morgan, X. C., Tickle, T. L., Sokol, H., Gevers, D., Devaney, K. L., Ward, D. V., Reyes, J. A., Shah, S. A., LeLeiko, N., Snapper, S. B., m. fl. (2012). Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome biology 13*, R79.

(*10*) Karlsson, F. H., Tremaroli, V., Nookaew, I., Bergström, G., Behre, C. J., Fagerberg, B., Nielsen, J., och Bäckhed, F., (2013). Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature 498*, 99–103.

(*11*) Qin, J., Li, Y., Cai, Z., Li, S., Zhu, J., Zhang, F., Liang, S., Zhang, W., Guan, Y., Shen, D., m. fl. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature 490*, 55–60.

(*12*) Tang, P., och Chiu, C., (2010). Metagenomics for the discovery of novel human viruses. *Future microbiology 5*, 177–189.

(*13*) Nacke, H., Engelhaupt, M., Brady, S., Fischer, C., Tautzt, J., och Daniel, R., (2012). Identification and characterization of novel cellulolytic and hemicellulolytic genes and enzymes derived from German grassland soil metagenomes. *Biotechnology letters 34*, 663–675.

(*14*) Buttigieg, P. L., Hankeln, W., Kostadinov, I., Kottmann, R., Yilmaz, P., Duhaime, M. B., och Glöckner, F. O., (2013). Ecogenomic perspectives on domains of unknown function: correlation-based exploration of marine metagenomes. *PLoS One 8*, e50869.

(*15*) Sharpton, T. J., (2014). An introduction to the analysis of shotgun metagenomic data. *Frontiers in plant science 5*, 209.

(*16*) Dudoit, S., Shaffer, J. P., och Boldrick, J. C., (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 71–103.

(*17*) Morgan, J. L., Darling, A. E., och Eisen, J. A., (2010). Metagenomic sequencing of an in vitro-simulated microbial community. *PloS one 5*, e10209.

(*18*) Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., och Gu, Y., (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics 13*, 341.

(*19*) Wooley, J. C., och Ye, Y., (2010). Metagenomics: facts and artifacts, and computational challenges. *Journal of computer science and technology 25*, 71–81.

(*20*) Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., m. fl. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & integrative genomics 15*, 141–161.

(*21*) Jonsson, V., Österlund, T., Nerman, O., och Kristiansson, E., (2016). Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. *BMC genomics 17*, 78.

(*22*) Eriksson, K. M., Johansson, C. H., Fihlman, V., Grehn, A., Sanli, K., Andersson, M. X., Blanck, H., Arrhenius, Å., Sircar, T., och Backhaus, T., (2015). Long-term effects of the antibacterial agent triclosan on marine periphyton communities. *Environmental Toxicology and Chemistry 34*, 2067–2077.

(*23*) Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K., och Beck, E., (2013). TIGRFAMs and genome properties in 2013. *Nucleic acids research 41*, D387–D395.

(*24*) Escalada, M. G., Harwood, J., Maillard, J.-Y., och Ochs, D., (2005). Triclosan inhibition of fatty acid synthesis and its effect on growth of Escherichia coli and Pseudomonas aeruginosa. *Journal of Antimicrobial Chemotherapy 55*, 879–882.

(*25*) Thompson, A., Griffin, P., Stuetz, R., och Cartmell, E., (2005). The fate and removal of triclosan during wastewater treatment. *Water environment research 77*, 63–67.

(*26*) Kolpin, D. W., Furlong, E. T., Meyer, M. T., Thurman, E. M., Zaugg, S. D., Barber, L. B., och Buxton, H. T., (2002). Pharmaceuticals, hormones, and other organic wastewater contaminants in US streams, 1999- 2000: A national reconnaissance. *Environmental science & technology 36*, 1202–1211.

(*27*) Lindström, A., Buerge, I. J., Poiger, T., Bergqvist, P.-A., Müller, M. D., och Buser, H.-R., (2002). Occurrence and environmental behavior of the bactericide triclosan and its methyl derivative in surface waters and in wastewater. *Environmental science & technology 36*, 2322–2329.

(*28*) Xie, Z., Ebinghaus, R., Flöser, G., Caba, A., och Ruck, W., (2008). Occurrence and distribution of triclosan in the German Bight (North Sea). *Environmental Pollution 156*, 1190–1195.

(*29*) Fair, P. A., Lee, H.-B., Adams, J., Darling, C., Pacepavicius, G., Alaee, M., Bossart, G. D., Henry, N., och Muir, D., (2009). Occurrence of triclosan in plasma of wild Atlantic bottlenose dolphins (Tursiops truncatus) and in their environment. *Environmental Pollution 157*, 2248–2254.

(*30*) Orvos, D. R., Versteeg, D. J., Inauen, J., Capdevielle, M., Rothenstein, A., och Cunningham, V., (2002). Aquatic toxicity of triclosan. *Environmental toxicology and chemistry 21*, 1338–1349.

(*31*) Tatarazako, N., Ishibashi, H., Teshima, K., Kishi, K., och Arizono, K., (2003). Effects of triclosan on various aquatic organisms. *Environmental sciences: an international journal of environmental physiology and toxicology 11*, 133–140.

(*32*) Ishibashi, H., Matsumura, N., Hirano, M., Matsuoka, M., Shiratsuchi, H., Ishibashi, Y., Takao, Y., och Arizono, K., (2004). Effects of triclosan on the early life stages and reproduction of medaka Oryzias latipes and induction of hepatic vitellogenin. *Aquatic Toxicology 67*, 167–179.

(*33*) Franz, S., Altenburger, R., Heilmeier, H., och Schmitt-Jansen, M., (2008). What contributes to the sensitivity of microalgae to triclosan? *Aquatic Toxicology 90*, 102–108.

(*34*) Remberger, M., Sternbeck, J., och Strömberg, K., *Screening av triclosan och vissa bromerade fenoliska ämnen i Sverige*; IVL Svenska Miljöinstitutet AB: 2002.

(*35*) Adolfsson-Erici, M., Pettersson, M., Parkkonen, J., och Sturve, J., (2002). Triclosan, a commonly used bactericide found in human milk and in the aquatic environment in Sweden. *Chemosphere 46*, 1485–1489.

(*36*) Allmyr, M., Adolfsson-Erici, M., McLachlan, M. S., och Sandborgh-Englund, G., (2006). Triclosan in plasma and milk from Swedish nursing mothers and their exposure via personal care products. *Science of the Total Environment 372*, 87–93.

(*37*) Wesgate, R., Grasha, P., och Maillard, J.-Y., (2016). Use of a predictive protocol to measure the antimicrobial resistance risks associated with biocidal product usage. *American journal of infection control 44*, 458–464.

(*38*) Yueh, M.-F., Taniguchi, K., Chen, S., Evans, R. M., Hammock, B. D., Karin, M., och Tukey, R. H., (2014). The commonly used antimicrobial additive triclosan is a liver tumor promoter. *Proceedings of the National Academy of Sciences 111*, 17200–17205.

(*39*) Lee, H.-R., Hwang, K.-A., Nam, K.-H., Kim, H.-C., och Choi, K.-C., (2014). Progression of breast cancer cells was enhanced by endocrine-disrupting chemicals, triclosan and octylphenol, via an estrogen receptor-dependent signaling pathway in cellular and mouse xenograft models. *Chemical research in toxicology 27*, 834–842.

(*40*) Witorsch, R. J., (2014). Critical analysis of endocrine disruptive activity of triclosan and its relevance to human exposure through the use of personal care products. *Critical reviews in toxicology 44*, 535–555.

(*41*) European Commission, Commission Implementing Decision (EU) 2016/110 of 27 January 2016 not approving triclosan as an existing active substance for use in biocidal products for product-type 1.,
`http://data.europa.eu/eli/dec_impl/2016/110/oj`, 2016.

(*42*) Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y.-H., Tu, Q., Xie, J., Van Nostrand, J. D., He, Z., och Yang, Y., (2011). Reproducibility and quantitation of amplicon sequencing-based detection. *The ISME journal 5*, 1303–1313.

(*43*) R Core Team, R: A Language and Environment for Statistical Computing.; R Foundation for Statistical Computing, Vienna, Austria, 2016.

(*44*) Robinson, M. D., McCarthy, D. J., och Smyth, G. K., (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics 26*, 139–140.

(*45*) McCullagh, P., och Nelder, J., *Generalized Linear Models*, 2. utg.; Chapman och Hall/CRC: London, England, 1989.

(*46*) Robinson, M. D., och Oshlack, A., (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome biology 11*, R25.

(*47*) Wallroth, M., (2016). Normalization of metagenomic data - A comprehensive evaluation of existing methods. *Master thesis Chalmers University of Technology, University of Gothenburg.*

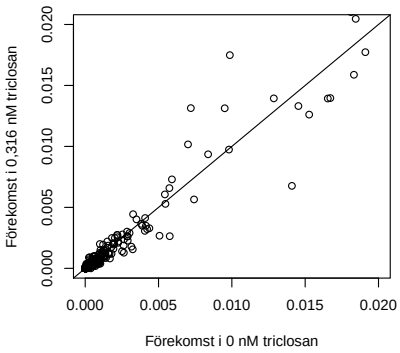
(*48*) Benjamini, Y., och Hochberg, Y., (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289–300.

(*49*) Langfelder, P., Zhang, B., och Horvath, S., (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics 24*, 719–720.

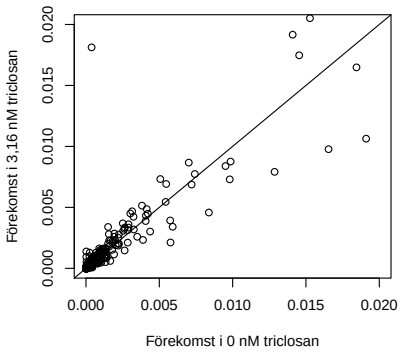
(*50*) Csardi, G., och Nepusz, T., (2006). The igraph software package for complex network research. *InterJournal Complex Systems*, 1695.

Appendix

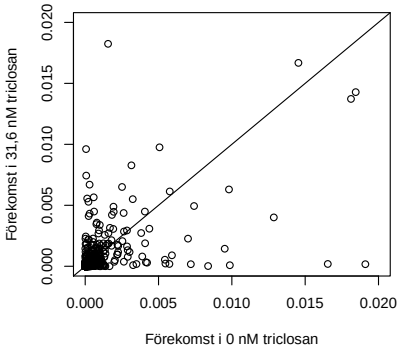
Undersökning av arters variation med koncentrationen triclosan



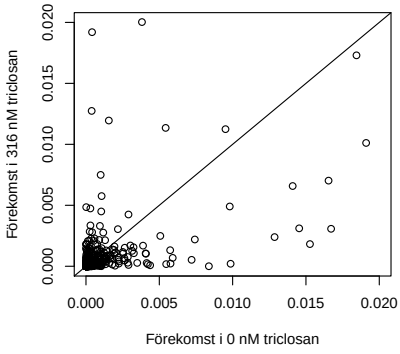
(a) 0,316 nM triclosan



(b) 3,16 nM triclosan



(c) 31,6 nM triclosan



(d) 316 nM triclosan

Figur 11: Scatter plots som visar hur förekomsten av arter ändras med ökad koncentration av triclosan jämfört med prov som inte behandlats med triclosan (på x-axeln). Medelvärdet av förekomsten plottades för koncentrationer där replikat finns. Plottarna är inzoomade jämfört med Figur 4.

Konstellationer av arter och gener skapade baserat på klustring av arter

Tabell 4: Arter i konstellation 1.

Arter
Bacteria, Lentisphaerae, Lentisphaeria, Lentisphaerales, Lentisphaeraceae, Lentisphaera, Unclassified Lentisphaera
Bacteria, Lentisphaerae, Lentisphaeria, Lentisphaerales, Lentisphaeraceae, Unclassified Lentisphaeraceae,
Bacteria, Proteobacteria, Alphaproteobacteria, Rhizobiales, Rhizobiaceae, Rhizobium, Unclassified Rhizobium
Eukaryota, Chromalveolata, Hacrobia, Centrohelida, Acanthocystidae, Pterocystis, Chlamydaster
Eukaryota, Metazoa, Mollusca, Bivalvia, Mytiloida, Unclassified Mytiloida,
Eukaryota, Metazoa, Platyhelminthes, Turbellaria, Rhabdocoela, Unclassified Rhabdocoela,

Tabell 5: Gener i konstellation 1.

Gen	Funktion	Antal i genfamilj	Korr med triclosan
TIGR02768	TraA_Ti: Ti-type conjugative transfer relaxase TraA	751	-0.22
TIGR02760	TraI_TIGR: conjugative transfer relaxase protein TraI	1960	-0.23
TIGR02767	TraG-Ti: Ti-type conjugative transfer system protein TraG	623	-0.27
TIGR02739	TraF: type-F conjugative transfer system pilin assembly protein TraF	256	-0.14
TIGR02738	TrbB: type-F conjugative transfer system pilin assembly thiol-disulfide isomerase TrbB	153	-0.15
TIGR00590	pcna: proliferating cell nuclear antigen (pcna)	259	-0.16
TIGR01971	NuoI: NADH-quinone oxidoreductase, chain I	121	-0.27
TIGR01030	rpmH_bact: ribosomal protein L34	44	-0.10
TIGR03738	PRTRC_C: PRTRC system protein C	66	-0.28
TIGR02920	acc_sec_Y2: accessory Sec system translocase SecY2	395	-0.14
TIGR02791	VirB5: P-type DNA transfer protein VirB5	220	-0.14
TIGR02781	VirB9: P-type conjugative transfer protein VirB9	248	-0.16
TIGR02686	relax_trwC: conjugative relaxase domain	286	-0.21
TIGR04332	gamma_Glu_sys: poly-gamma-glutamate system protein	307	-0.19
TIGR03741	PRTRC_E: PRTRC system protein E	106	-0.22
TIGR03743	SXT_TraD: conjugative coupling factor TraD, SXT/TOL subfamily	636	0.07
TIGR03744	traC_PFL_4706: conjugative transfer ATPase, PFL_4706 family	892	-0.09
TIGR02759	TraD_Ftype: type IV conjugative transfer system coupling protein TraD	566	-0.16
TIGR02743	TraW: type-F conjugative transfer system protein TraW	202	-0.15
TIGR02740	TraF-like: TraF-like protein	276	-0.15
TIGR02746	TraC-F-type: type-IV secretion system protein TraC	800	-0.34

Gen	Funktion	Antal i genfamilj	Korr med triclosan
TIGR02747	TraV: type IV conjugative transfer system protein TraV	145	-0.14
TIGR01912	TatC-Arch: Sec-independent protein translocase TatC	234	-0.27
TIGR02775	TrbG_Ti: P-type conjugative transfer protein TrbG	206	-0.16

Tabell 6: Arter i konstellation 2.

Arter
Bacteria, Bacteroidetes, Flavobacteria, Flavobacteriales, Flavobacteriaceae, Kordia, Bacteroidetes bacterium SOVto10
Bacteria, Bacteroidetes, Flavobacteria, Flavobacteriales, Flavobacteriaceae, Kordia, Unclassified Kordia
Bacteria, Firmicutes, Bacilli, Bacillales, Paenibacillaceae, Ammoniphilus, Unclassified Ammoniphilus
Bacteria, Fusobacteria, Fusobacteria, Fusobacteriales, Leptotrichiaceae, Unclassified Leptotrichiaceae,
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Aestuariibacter, Aestuariibacter salexigens
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Aestuariibacter, Alteromonas sp. S89
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Aestuariibacter, bacterium 1H105
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Alteromonas, Alteromonas genovensis
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, Glaciecola pallidula
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, Glaciecola sp. S577
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, gas vacuolate str. 206
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Porticoccus, gamma proteobacterium IMCC2136
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Saccharophagus, Saccharophagus degradans 2-40
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Simiduia, Simiduia agarivorans
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Simiduia, Simiduia areninigrae
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Simiduia, Simiduia sp. KLE1111
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Teredinibacter, Teredinibacter turnerae T7901
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Unclassified Alteromonadaceae,
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Colwelliaceae, Colwellia, Colwellia psychrerythraea 34H
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Colwelliaceae, Colwellia, Colwellia sp. STAB404
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Colwelliaceae, Thalassomonas, Thalassomonas sp. Za6a-12
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Shewanellaceae, Shewanella, Shewanella putrefaciens

Arter
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Unclassified Alteromonadales, ,
Bacteria, Proteobacteria, Gammaproteobacteria, Enterobacteriales, Enterobacteriaceae, Escherichia-Shigella, Unclassified Escherichia-Shigella
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Hahellaceae, Hahella, Hahella chejuensis KCTC 2396
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oceanospirillaceae, Amphritea, Amphritea japonica
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oceanospirillaceae, Amphritea, Unclassified Amphritea
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oceanospirillaceae, Oceanobacter, Thalassolituus sp. H61
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oceanospirillaceae, Oceanobacter, Unclassified Oceanobacter
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oceanospirillaceae, Thalassolituus, Thalassolituus sp. IMCC1883
Bacteria, Proteobacteria, Gammaproteobacteria, Order Incertae Sedis, Family Incertae Sedis, Arenicella, Stilbonema sp. associated bacterium 8_II_6
Bacteria, Proteobacteria, Gammaproteobacteria, Pasteurellales, Pasteurellaceae, Aggregatibacter, Unclassified Aggregatibacter
Bacteria, Proteobacteria, Gammaproteobacteria, Pseudomonadales, Unclassified Pseudomonadales, ,
Bacteria, Proteobacteria, Gammaproteobacteria, Vibrionales, Unclassified Vibrionales, ,
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Eolimna
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Rhopalodia
Eukaryota, Rhizaria, Cercozoa, Granofilosea, Massisteria, Massisteria marina, Unclassified Massisteria marina
Eukaryota, Rhizaria, Cercozoa, Granofilosea, Unclassified Granofilosea, ,
Eukaryota, Unikonta, Amoebozoa, Discosea, Flabellinia, Vannellida, Platyamoeba
Eukaryota, Unikonta, Choanomonada, Craspedida, Salpingoeca, Salpingoeca tuba, Unclassified Salpingoeca tuba

Tabell 7: Gener i konstellation 2.

Gen	Funktion	Antal i genfamilj	Korr med triclosan
TIGR03703	plsB: glycerol-3-phosphate O-acyltransferase	799	0.93
TIGR03822	AblA_like_2: lysine-2,3-aminomutase-related protein	321	0.92
TIGR00238	TIGR00238: KamA family protein	331	0.91
TIGR03820	lys_2_3_AblA: lysine-2,3-aminomutase	417	0.90
TIGR03821	AblA_like_1: lysine-2,3-aminomutase-like protein	321	0.92
TIGR02317	prpB: methylisocitrate lyase	285	0.94
TIGR01527	arch_NMN_Atrans: nicotinamide-nucleotide adenylyltransferase	167	0.95
TIGR03890	nif11_cupin: nif11 domain/cupin domain protein	171	0.98
TIGR01873	cas_CT1978: CRISPR-associated endoribonuclease Cas2, subtype I-E/ECOLI	87	0.96
TIGR01907	casE_Cse3: CRISPR-associated protein Cas6/Cse3/CasE, subtype I-E/ECOLI	208	0.97
TIGR04211	SH3_and_anchor: SH3 domain protein	198	0.92

Gen	Funktion	Antal i genfamilj	Korr med triclosan
TIGR02103	pullul_strch: alpha-1,6-glucosidases, pullulanase-type	900	0.93
TIGR02816	pfaB_fam: PfaB family protein	534	0.95
TIGR01868	casD_Cas5e: CRISPR-associated protein Cas5/CasD, subtype I-E/ECOLI	230	0.96
TIGR01869	casC_Cse4: CRISPR-associated protein Cas7/Cse4/CasC, subtype I-E/ECOLI	325	0.97

Tabell 8: Arter i konstellation 3.

Arter
Bacteria, Actinobacteria, Actinobacteria, Micrococcales, Microbacteriaceae, Microbacterium, Unclassified Microbacterium
Bacteria, Bacteroidetes, Flavobacteria, Flavobacteriales, Flavobacteriaceae, Flavobacterium, Flavobacterium kamogawaensis
Bacteria, Bacteroidetes, Flavobacteria, Flavobacteriales, Flavobacteriaceae, Flavobacterium, Flavobacterium sp. GSW-R14
Bacteria, Bacteroidetes, Flavobacteria, Flavobacteriales, Flavobacteriaceae, Winogradskyella, Bacteroidetes bacterium E42
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Aestuariibacter, Aestuariibacter litoralis
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, Glaciecola nitratireducens
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, Glaciecola nitratireducens FR1064
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Haliea, Haliea mediterranea
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Porticoccus, Porticoccus litoralis
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Colwelliaceae, Colwellia, Colwellia demingiae
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oceanospirillaceae, Amphritea, Amphritea sp. MEBiC05461T
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oceanospirillaceae, Oceanospirillum, Unclassified Oceanospirillum
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oleiphilaceae, Oleiphilus, Oleiphilus messinensis
Bacteria, Proteobacteria, Gammaproteobacteria, Thiotrichales, Family Incertae Sedis, Caedibacter, Caedibacter taeniospiralis
Eukaryota, Alveolata, Ciliophora, Intramacronucleata, Conthreep, Phyllopharyngea, Suctoria
Eukaryota, Alveolata, Ciliophora, Intramacronucleata, Spirotrichea, Euplotia, Euplotes
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Craticula
Eukaryota, Chromalveolata, Stramenopiles, Dictyochophyceae, Pedinellales, Dictyochophyceae, Pteridomonas
Eukaryota, Excavata, Euglenozoa, Kinetoplastea, Metakinetoplastina, Neobodonida, Neobodo
Eukaryota, Excavata, Euglenozoa, Kinetoplastea, Metakinetoplastina, Neobodonida, Unclassified Neobodonida
Eukaryota, Excavata, Euglenozoa, Kinetoplastea, Metakinetoplastina, Trypanosomatida, Trypanosoma
Eukaryota, Excavata, Euglenozoa, Kinetoplastea, Metakinetoplastina, Trypanosomatida, Unclassified Trypanosomatida
Eukaryota, Excavata, Euglenozoa, Kinetoplastea, Metakinetoplastina, Unclassified Metakinetoplastina,

Arter
Eukaryota, Rhizaria, Cercozoa, 7-2.3, Gymnophrys sp. COHH 17, Unclassified Gymnophrys sp. COHH 17,
Eukaryota, Unikonta, Amoebozoa, Discosea, Flabellinia, Dactylopodida, Unclassified Dactylopodida
Eukaryota, Unikonta, Amoebozoa, Discosea, Flabellinia, Vannellida, Unclassified Vannellida

Tabell 9: Gener i konstellation 3.

Gen	Funktion	Antal i genfamilj	Korr med triclosan
TIGR02548	casB_cse2: CRISPR type I-E/ECOLI-associated protein CasB/Cse2	159	0.95
TIGR03907	QH_beta: quinohemoprotein amine dehydrogenase, beta subunit	338	0.88
TIGR01503	MthylAspMut_E: methylaspartate mutase, E subunit	480	0.89
TIGR00816	tdt: C4-dicarboxylate transporter/malic acid transport protein	320	0.91
TIGR02922	TIGR02922: TIGR02922 family protein	67	0.96
TIGR02910	sulfite_red_A: sulfite reductase, subunit A	334	0.86
TIGR02680	TIGR02680: TIGR02680 family protein	1356	0.89
TIGR03908	QH_alpha: quinohemoprotein amine dehydrogenase, alpha subunit	510	0.88
TIGR03758	conj_TIGR03758: integrating conjugative element protein, PFL_4701 family	76	0.95
TIGR03295	frhA: coenzyme F420 hydrogenase, subunit alpha	412	0.84
TIGR00142	hycI: hydrogenase maturation peptidase HycI	146	0.85
TIGR03993	hydrog_HybE: [NiFe] hydrogenase assembly chaperone, HybE family	143	0.87

Konstellationer av arter och gener skapade baserat på klustring av gener

Tabell 10: Gener i konstellation 1.

Gen	Funktion	Antal i genfamilj	Korr med triclosan
TIGR00540	TPR_hemY_coli: heme biosynthesis-associated TPR protein	385	0.83
TIGR01848	PHA_reg_PhaR: polyhydroxyalkanoate synthesis repressor PhaR	107	0.80
TIGR03010	sulf_tusC_dsrF: sulfur relay protein TusC/DsrF	116	0.81
TIGR03012	sulf_tusD_dsrE: sulfur relay protein TusD/DsrE	127	0.78
TIGR01935	NOT-MenG: RraA family	150	0.88
TIGR01434	glu_cys_ligase: glutamate–cysteine ligase	512	0.90
TIGR03822	AblA_like_2: lysine-2,3-aminomutase-related protein	321	0.92
TIGR02449	TIGR02449: TIGR02449 family protein	65	0.68
TIGR02443	TIGR02443: conserved hypothetical protein	59	0.83
TIGR00148	TIGR00148: UbiD family decarboxylase	438	0.81
TIGR02451	anti_sig_ChrR: anti-sigma factor, putative, ChrR family	216	0.91
TIGR02998	RraA_entero: regulator of ribonuclease activity A	161	0.90
TIGR01709	typeII_sec_gspL: type II secretion system protein L	389	0.87
TIGR01345	malate_syn_G: malate synthase G	721	0.91
TIGR01344	malate_syn_A: malate synthase A	511	0.91
TIGR02518	EutH_ACDH: acetaldehyde dehydrogenase (acetylating)	488	0.78
TIGR00238	TIGR00238: KamA family protein	331	0.91
TIGR03138	QueF: queuine synthase	275	0.84
TIGR03136	malonate_biotin: Na+-transporting malonate decarboxylase, carboxybiotin decarboxylase subunit	399	0.71
TIGR01752	flav_long: flavodoxin	167	0.74
TIGR01753	flav_short: flavodoxin	140	0.82
TIGR03820	lys_2_3_AblA: lysine-2,3-aminomutase	417	0.90
TIGR03821	AblA_like_1: lysine-2,3-aminomutase-like protein	321	0.92
TIGR01711	gspJ: type II secretion system protein J	192	0.88
TIGR01713	typeII_sec_gspC: type II secretion system protein C	259	0.82
TIGR00481	TIGR00481: Raf kinase inhibitor-like protein, YbhB/YbcL family	142	0.70
TIGR00484	EF-G: translation elongation factor G	691	0.61
TIGR00312	cbiD: cobalamin biosynthesis protein CbiD	347	0.82
TIGR01300	CPA3_mnhG_phaG: monovalent cation/proton antiporter, MnhG/PhaG subunit	97	0.77
TIGR01834	PHA_synth_III_E: poly(R)-hydroxyalkanoic acid synthase, class III, PhaE subunit	324	0.62
TIGR02334	prpF: probable AcnD-accessory protein PrpF	390	0.89
TIGR00013	taut: 4-oxalocrotonate tautomerase family enzyme	64	0.69

Gen	Funktion	Antal i genfamilj	Korr med triclosan
TIGR01109	Na_pump_decarbB: sodium ion-translocating decarboxylase, beta subunit	354	0.72
TIGR02205	septum_zipA: cell division protein ZipA	284	0.83
TIGR02804	ExbD_2: TonB system transport protein ExbD	121	0.84
TIGR00190	thiC: thiamine biosynthesis protein ThiC	423	0.74
TIGR01298	RNaseT: ribonuclease T	200	0.82
TIGR00452	TIGR00452: tRNA (mo5U34)-methyltransferase	314	0.73
TIGR01195	oadG_fam: sodium pump decarboxylases, gamma subunit	84	0.77
TIGR01594	holin_lambda: phage holin, lambda family	107	0.80
TIGR03503	TIGR03503: TIGR03503 family protein	374	0.64
TIGR00230	sfsA: sugar fermentation stimulation protein	234	0.83
TIGR02364	dha_pts: dihydroxyacetone kinase, phospho-transfer subunit	125	0.64
TIGR01450	recC: exodeoxyribonuclease V, gamma subunit	1063	0.91
TIGR00942	2a6301s05: multicomponent Na+:H+ antiporter	144	0.76
TIGR00941	2a6301s03: multicomponent Na+:H+ antiporter, MnhC subunit	104	0.76
TIGR01707	gspI: type II secretion system protein I	101	0.87

Tabell 11: Arter i konstellation 1.

Arter
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Alteromonas, Alteromonas macleodii
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Alteromonas, Alteromonas sp. SN2
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Alteromonas, Unclassified Alteromonas
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Candidatus Endobugula, Candidatus Endobugula glebosa
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Eionea, Eionea nigra
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, Alteromonas macleodii
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, Glaciecola pallidula
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, Glaciecola punicea
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, Glaciecola punicea DSM 14233 = ACAM 611
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, Glaciecola siphonariae
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, Pseudoalteromonas atlantica T6c
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, Unclassified Glaciecola
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, bacterium QM22
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Saccharophagus, Saccharophagus degradans 2-40

Arter
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Simiduia, Simiduia sp. KLE1111
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Unclassified Alteromonadaceae,
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Unclassified Alteromonadales, ,
Bacteria, Proteobacteria, Gammaproteobacteria, Enterobacteriales, Enterobacteriaceae, Cronobacter, Cronobacter turicensis z3032
Bacteria, Proteobacteria, Gammaproteobacteria, Enterobacteriales, Enterobacteriaceae, Enterobacter, Unclassified Enterobacter
Bacteria, Proteobacteria, Gammaproteobacteria, Enterobacteriales, Enterobacteriaceae, Erwinia, Erwinia tasmaniensis
Bacteria, Proteobacteria, Gammaproteobacteria, Enterobacteriales, Enterobacteriaceae, Escherichia-Shigella, Unclassified Escherichia-Shigella
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oceanospirillaceae, Oceanobacter, Thalassolituus sp. H61
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oceanospirillaceae, Oceanobacter, Unclassified Oceanobacter
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oceanospirillaceae, Thalassolituus, Thalassolituus sp. IMCC1883
Bacteria, Proteobacteria, Gammaproteobacteria, Pasteurellales, Pasteurellaceae, Aggregatibacter, Unclassified Aggregatibacter
Bacteria, Proteobacteria, Gammaproteobacteria, Pasteurellales, Pasteurellaceae, Unclassified Pasteurellaceae,
Bacteria, Proteobacteria, Gammaproteobacteria, Pseudomonadales, Pseudomonadaceae, Cellvibrio, Cellvibrio japonicus Ueda107
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Amphiprora
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Encyonema
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Entomoneis
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Gomphonema
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Lemnicula
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Placoneis
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Rhopalodia
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Surirella
Eukaryota, Unikonta, Amoebozoa, Discosea, Flabellinia, Dactylopodida, Vexillifera

Tabell 12: Gener i konstellation 2.

Gen	Funktion	Antal i genfamilj	Korr med triclosan
TIGR01448	recD_rel: helicase, RecD/TraA family	720	-0.05
TIGR02768	TraA_Ti: Ti-type conjugative transfer relaxase TraA	751	-0.22
TIGR02760	TraI_TIGR: conjugative transfer relaxase protein TraI	1960	-0.23
TIGR02739	TraF: type-F conjugative transfer system pilin assembly protein TraF	256	-0.14

Gen	Funktion	Antal i genfamilj	Korr med triclosan
TIGR02738	TrbB: type-F conjugative transfer system pilin assembly thiol-disulfide isomerase TrbB	153	-0.15
TIGR00590	pcna: proliferating cell nuclear antigen (pcna)	259	-0.16
TIGR01030	rpmH_bact: ribosomal protein L34	44	-0.10
TIGR02791	VirB5: P-type DNA transfer protein VirB5	220	-0.14
TIGR02781	VirB9: P-type conjugative transfer protein VirB9	248	-0.16
TIGR02686	relax_trwC: conjugative relaxase domain	286	-0.21
TIGR03743	SXT_TraD: conjugative coupling factor TraD, SXT/TOL subfamily	636	0.07
TIGR03744	traC_PFL_4706: conjugative transfer ATPase, PFL_4706 family	892	-0.09
TIGR03674	fen_arch: flap structure-specific endonuclease	338	-0.18
TIGR02759	TraD_Ftype: type IV conjugative transfer system coupling protein TraD	566	-0.16
TIGR02743	TraW: type-F conjugative transfer system protein TraW	202	-0.15
TIGR02740	TraF-like: TraF-like protein	276	-0.15
TIGR02747	TraV: type IV conjugative transfer system protein TraV	145	-0.14
TIGR02775	TrbG_Ti: P-type conjugative transfer protein TrbG	206	-0.16

Tabell 13: Arter i konstellation 2.

Arter
Bacteria, Cyanobacteria, Cyanobacteria, SubsectionIII, FamilyI, Oscillatoria, Phormidium sp. MBIC10210
Bacteria, Lentisphaerae, Lentisphaeria, Lentisphaerales, Lentisphaeraceae, Lentisphaera, Unclassified Lentisphaera
Bacteria, Lentisphaerae, Lentisphaeria, Lentisphaerales, Lentisphaeraceae, Unclassified Lentisphaeraceae,
Bacteria, Proteobacteria, Alphaproteobacteria, Rhizobiales, Rhizobiaceae, Rhizobium, Unclassified Rhizobium
Eukaryota, Alveolata, Ciliophora, Intramacronucleata, Conthreep, Phyllopharyngea, Ephelota
Eukaryota, Alveolata, Protalveolata, Syndiniales, Syndiniales, Syndiniales Group I, Syndiniales
Eukaryota, Chromalveolata, Hacrobia, Centrohelida, Acanthocystidae, Pterocystis, Chlamydaster
Eukaryota, Metazoa, Mollusca, Bivalvia, Mytiloida, Unclassified Mytiloida,
Eukaryota, Metazoa, Platyhelminthes, Turbellaria, Rhabdocoela, Unclassified Rhabdocoela,

Tabell 14: Gener i konstellation 3.

Gen	Funktion	Antal i genfamilj	Korr med triclosan
TIGR03703	plsB: glycerol-3-phosphate O-acyltransferase	799	0.93
TIGR01435	glu_cys_lig_rel: glutamate–cysteine ligase/gamma-glutamylcysteine synthetase	719	0.84
TIGR01346	isocit_lyase: isocitrate lyase	527	0.95
TIGR01190	ccmB: heme exporter protein CcmB	211	0.82
TIGR00254	GGDEF: diguanylate cyclase (GGDEF) domain	168	0.89
TIGR02319	CPEP_Pphonmut: carboxyvinyl-carboxyphosphonate phosphorylmutase	294	0.85
TIGR02317	prpB: methylisocitrate lyase	285	0.94

Gen	Funktion	Antal i genfamilj	Korr med triclosan
TIGR01527	arch_NMN_Atrans: nicotinamide-nucleotide adenylyltransferase	167	0.95
TIGR02798	ligK_PcmE: 4-carboxy-4-hydroxy-2-oxoadipate aldolase/oxaloacetate decarboxylase	222	0.83
TIGR03890	nif11_cupin: nif11 domain/cupin domain protein	171	0.98
TIGR01705	MTA/SAH-nuc-hyp: putative 5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase	212	0.86
TIGR02320	PEP_mutase: phosphoenolpyruvate phosphomutase	284	0.84
TIGR02321	Pphn_pyruv_hyd: phosphonopyruvate hydrolase	290	0.88
TIGR02911	sulfite_red_B: sulfite reductase, subunit B	261	0.85
TIGR02547	casA_cse1: CRISPR type I-E/ECOLI-associated protein CasA/Cse1	504	0.96
TIGR01873	cas_CT1978: CRISPR-associated endoribonuclease Cas2, subtype I-E/ECOLI	87	0.96
TIGR01907	casE_Cse3: CRISPR-associated protein Cas6/Cse3/CasE, subtype I-E/ECOLI	208	0.97
TIGR04211	SH3_and_anchor: SH3 domain protein	198	0.92
TIGR02103	pullul_strch: alpha-1,6-glucosidases, pullulanase-type	900	0.93
TIGR02816	pfaB_fam: PfaB family protein	534	0.95
TIGR01868	casD_Cas5e: CRISPR-associated protein Cas5/CasD, subtype I-E/ECOLI	230	0.96
TIGR01869	casC_Cse4: CRISPR-associated protein Cas7/Cse4/CasC, subtype I-E/ECOLI	325	0.97
TIGR03948	butyr_acet_CoA: butyryl-CoA:acetate CoA-transferase	445	0.91

Tabell 15: Arter i konstellation 3.

Arter
Bacteria, Bacteroidetes, Flavobacteria, Flavobacteriales, Flavobacteriaceae, Kordia, Unclassified Kordia
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Aestuariibacter, bacterium 1H105
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Candidatus Endobugula, Candidatus Endobugula glebosa
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Dasania, marine gamma proteobacterium HTCC2143
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, Alteromonas macleodii
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, Glaciecola pallidula
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, Glaciecola siphonariae
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Glaciecola, Unclassified Glaciecola
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Melitea, Alteromonadaceae bacterium HB10001
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Saccharophagus, Saccharophagus degradans 2-40

Arter
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Simiduia, Simiduia sp. KLE1111
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Alteromonadaceae, Unclassified Alteromonadaceae,
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Colwelliaceae, Colwellia, Colwellia psychrerythraea 34H
Bacteria, Proteobacteria, Gammaproteobacteria, Alteromonadales, Unclassified Alteromonadales, ,
Bacteria, Proteobacteria, Gammaproteobacteria, Enterobacteriales, Enterobacteriaceae, Escherichia-Shigella, Unclassified Escherichia-Shigella
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oceanospirillaceae, Oceanobacter, Thalassolituus sp. H61
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oceanospirillaceae, Oceanobacter, Unclassified Oceanobacter
Bacteria, Proteobacteria, Gammaproteobacteria, Oceanospirillales, Oceanospirillaceae, Thalassolituus, Thalassolituus sp. IMCC1883
Bacteria, Proteobacteria, Gammaproteobacteria, Unclassified Gammaproteobacteria, , ,
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Eolimna
Eukaryota, Chromalveolata, Stramenopiles, Diatomea, Bacillariophytina, Bacillariophyceae, Rhopalodia
Eukaryota, Rhizaria, Cercozoa, Granofilosea, Massisteria, Massisteria marina, Unclassified Massisteria marina
Eukaryota, Rhizaria, Cercozoa, Granofilosea, Unclassified Granofilosea, ,