Myndigheten för samhällsskydd och beredskap

Universell detektion av biologiska högriskorganismer

FORSKNING

MSB:s kontaktpersoner:
Sara Brunnberg, 010-240 40 87
Susanna Ekströmer, 010-240 4352

Publikationsnummer MSB 895-juni 2015

Förord

Denna rapport utgör slutredovisning för projektet "Universell detektion av högriskorganismer" som finansierats av MSBs öppna utlysning av forskningsmedel 2010. Projektet har letts av Statens Veterinärmedicinska Anstalt, SVA under tiden Januari 2011 till April 2014. Projektet har haft ett fokus på att utnyttja den nya generationens sekvenseringsteknologi för att söka efter DNA sekvenser som indikerar att högpatogena organismer finns närvarande i prov som även innehåller en komplex bakgrund av andra organismer. Under projekttiden har sekvenseringsteknologin utvecklats mycket snabbt. Metoderna för informationsbehandling ligger emellertid hela tiden steget efter. Detta projekt har bidragit till att förbättra möjligheterna till snabbare och säkrare användning av sekvensering i diagnostiska sammanhang.

Innehållsförteckning

Sammanfattning

Detta projekt har utvecklat metoder och verktyg för att förbättra förmågan att kunna dra konklusiva slutsatser vid analys av prover med okänt innehåll som kan innehålla smittämnen associerade med allvarlig sjukdom. Projektet har använt sig av den nya generationens DNA sekvenseringsteknologi och ett stort fokus har lagts vid att utveckla och utvärdera dataanalysmetoder för att ge bästa möjliga beslutsunderlag. Speciellt har fokus lagts på att kunna skilja på högriskorganismer och de betydligt mindre farliga nära släktingar som finns spridda i vår omgivning och som kan skapa stora beslutssvårigheter vid bedömningen av resultaten. Detta har gjorts genom att genomsignaturer identifierats som med stor tillförlitlighet signalerar att alvarliga smittämnen finns närvarande. Projekt har skapat verktyg som ökat vår förmåga att, vid en oklar hotbild med misstanke om biologiskt hot, kunna göra en tillförlitlig förutsättningslös analys av prover som ställer en öppen fråga "vad finns i provet", snarare än en serie specifika frågor "finns det detta i provet?, finns det detta i provet?, finns det detta i provet?, etc... Vidare har fokus lagts vid att skapa en så komplett lösning som möjligt, dvs, från provtagning, till sekvensering, till dataanalysen, till tolkningen, till rapporteringen och framförallt visualiseringen av resultaten.

1. Bakgrund

1.1 Sekvenseringsteknologins utveckling och användning

Sekvenseringsteknologin (avläser informationen i organismers arvsmassa, det vill säga information lagrat i DNA trådar) har varit ett av dom områden som utvecklats mest inom livsvetenskaperna den senaste tiden. Det har berört nästan alla forskningsfält som har anknytning till biologi/medicin och är nu även på väg att börja användas i rena diagnostiska tillämpningar.
Det började med att det storskaliga sekvenseringsprojektet HUGO som under 1990 talet hade som mål att sekvensera, det vill säga avläsa informationen i DNA trådarna för hela det 3 miljarder baspar stora humana genomet (DNA är uppbyggt av "baspar", som är en minsta informationsenhet likt en bokstav i en text; Ett genom [jeno:´m] är den totala mängden DNA som bär på alla organismens gener). Mycket automatiseringar utvecklades i HUGO men fortfarande skapades varje enskild sekvensinformationsavläsning var för sig. Den stora tekniska revolutionen kom i mitten av första årtiondet på 2000 talet när metoder utvecklades för att samtidigt skapa sekvensavläsningar från ett stort antal slumpmässiga regioner i det studerade genomet. Detta kallas för parallellsekvensering eller nya generationens sekvenseringsteknologi (NGS). Grundprincipen i NGS är fortfarande liknande men tekniken har utvecklats vidare så att större datamängder genereras av bättre kvalitet och varje enskild sekvensavläsning har blivit längre, vilket har många fördelar vid dataanalysstegen.

Parallellsekvenseringsmaskinerna fanns från början enbart i form av mycket stora instrument som lämpade sig för specialiserade laboratorier som bara gjorde sekvenseringar. Dock har det den senaste tiden kommit ett antal mindre maskiner som visserligen inte ger lika mycket data vid varje körning, men som är mycket mer lämpade för normalstora diagnostiska och/eller forskningsinriktade laboratorier.

Sekvensering används till att analysera informationen som finns lagrad i DNA, men det finns ett stort antal varianter på detta. I ett klassiskt genomprojekt görs sekvensering av ett stort antal slumpmässiga regioner i genomet och sedan pusslas informationen ihop så att en så bra bild som möjligt av hela arvsmassan skapas. Normalt producerar man så många sekvensavläsningar så att varje del av genomet i genomsnitt beskrivs av ett större antal oberoende sekvensavläsningar.

Man kan även sekvensera andra typer av nukleinsyraprover. Det är till exempel mycket vanligt att sekvensera RNA istället för DNA. RNA liknar DNA och skapas då en gen aktiveras och ju mer aktiv en gen är desto mer RNA bildas från den. Man kan således mäta genernas aktivitetsgrad. En annan variant är att sekvensera ett prov som innehåller en blandning av många olika organismer. Det brukar kallas för metagenomsekvensering. Metagenomsekvensering kan till exempel användas för att förstå hur den normala bakteriefloran i människor ser ut och hur den påverkar vår hälsa. Mikroorganismer finns överallt på vår kropp och i mag-tarmkanalen och det finns storskaliga projekt som försöker förstå betydelsen av detta (Human Microbiome Project, HMP). Metagenomsekvensering kan även användas för att identifiera sjukdomsframkallande mikroorganismer i bakgrunden av de normala "snälla" mikroorganismerna. Detta projekt handlar om att utveckla metoder för att förbättrar vår förmåga att upptäcka sekvenser som signalerar att det finns farliga mikroorganismer i ett prov vid metagenomsekvenseringar.

1.2 Den bioinformatiska utmaningen

I samband med att sekvenseringsmaskinerna blir fler, effektivare och sekvenseringarna billigare öppnas många nya möjligheter men det skapas också en stor utmaning i att hantera den stora informationsmängden. En del av problemet är att mycket stora mängder data skapas. Analys av de stora datamängderna kan vara mycket krävande och lagring av data är mycket kostsamt. Det pågår en debatt om vad som egentligen behöver sparas och hur länge man bör spara informationen. Det finns ett samarbete mellan Europa (EBI), USA (NCBI) och Japan (DDBJ) för att hålla ett öppet arkiv för sekvensdata. Det handlar dels om färdigbearbetade sekvenser och dels om de underliggande sekvensavläsningarna (rådata från sekvenseringsmaskinerna). Databasen för färdigbehandlade sekvenser innehåller just nu information om cirka 850 miljarder baspar (maj 2014) och den dubbleras vartannat år. Databasen för rådata har nyligen passerat 1 biljard baspar (en miljon miljarder). Dubblingstiden är här 19 månader (källa http://www.ebi.ac.uk/ena/about/statistics).

Sekvenseringen är inte längre bara ett forskningsverktyg utan är även på väg in i diagnostikverksamheten. En annan utmaning i detta sammanhang är att skapa jämförbara och standardiserade analysmetoder samt skapa förutsättningar att snabbt dela data med varandra. Det handlar både om tekniska lösningar men också om att skapa juridiska förutsättningar och om att få laboratorier att aktivt vilja dela med sig av sina data. Stora problem är även kopplade till hur metadata skall hanteras (kringinformation om vad som sekvenserades). Denna information kan vara mycket känslig men utan metadata är sekvensinformationen oftast av begränsat värde.

2. Projektbeskrivning

2.1 Provbehandling

För sekvensering behövs ett framrenat DNA prov. Det är i stort sett samma typ av provberedning som används vid PCR analys (den vanligaste molekylära diagnostikmetoden, som påvisar förekomst av en utvald kort DNA region). Sekvenseringar har dock som regel lite högre krav på DNA preparationens kvalitet. Allt DNA renas fram ur provet, vanligtvis med hjälp av specifik inbindning till en kiselmatris. Eftersom detta steg i sekvenseringsprocessen har optimerats under många år i samband med optimeringen av PCR diagnostik har detta projekt enbart utnyttjat tillgängliga metoder och inte utvecklat nya provberedningsprotokoll.

När väl rent DNA finns framtaget fragmenteras det, det vill säga slumpmässiga brottpunkter i DNA kedjan introduceras så att det skapas en uppsättning ungefär lika långa DNA fragment som representerar slumpvis utvalda områden av genomet. Fragmenteringen görs normalt mekaniskt (behandling med kraftiga ultraljudvågor) eller enzymatiskt (med hjälp av speciella proteiner som klyver sönder DNAt). Fragmenten kopplas sedan samman med korta "DNA adaptrar" av känd sekvens, en på varje sida. Eftersom DNA adaptrarnas sekvenser är kända kan man förstärka DNAt med PCR så att tillräckliga mängder skapas för att sekvenseringsreaktionens signal skall kunna detekteras. Förstärkningen sker med fragmenten inneslutna i små vattendroppar i en oljeemulsion eller fast förankrade på en yta. Detta hindrar molekylerna från att blanda sig med varandra och ger möjlighet att avläsa signalerna som små punkter. Det är fortfarande mycket svårt att detektera signalen från en enskild molekyl, även om de första maskinerna som gör detta redan finns och detta tros bli standard i framtidens maskiner.

För att minska mängden sekvensdata som behöver produceras kan man anrika det fragmenterade DNAt för de regioner man är mest intresserad av. Det kan man göra med hjälp av en uppsättning med förankrade DNA bitar som fiskar ut fragment med liknande sekvens ur bakgrundsblandningen. Denna strategi har används framförallt när utvalda delar av det mänskliga genomet skall sekvenseras. I detta projekt analyseras metagenomprover som kan ha hög komplexitet och vi ville därför även utvärdera om denna typ av anrikning skulle kunna vara användbar. Vi gjorde därför ett pilotförsök där förankrade DNA bitar designade mot specifika områden som är förknippad med bakteriers sjukdomsframkallande förmåga användes. Utfallet av dessa försök visade på ett flertal problem. Känsligheten att detektera låga koncentrationer av ett smittämne var begränsad. Det krävdes lång tid för DNA bitarna att hitta sina mål vilket leder till långa analystider. Kostnaderna per provanalys blev höga. I och med att sekvenseringsmaskinernas kapacitet hela tiden ökar kraftigt framstod det därför som en bättre lösning att producera mer sekvensdata och

istället använda datorbaserade metoder för att filtrera bort bakgrunden. Denna strategi medför dessutom mindre risk för att man missar någonting som man saknar i sin uppsättning utfisknings-DNA bitar, det vill säga man får ett mer universellt tillvägagångssätt. En stor tonvikt i detta projekt lades därför vid dataanalyssteget, vilket utgör en stor utmaning.

2.2 Att skilja på högpatogena organismer och deras nära släktingar

Detta projekt fokuserar speciellt på en grupp organismer som klassificeras som "högpatogena", det vill säga ger allvarlig sjukdom i människor. Många av dessa smittämnen är zoonoser, vilket innebär att människor kan få smittan från djur. Mikrobiella organismer delas in i biologiska skyddsnivåer,1-4. Vi har i detta projekt jobbat med organismer som klassas i skyddsnivå 3 och 4. Både virus och bakterier finns representerade. En förteckning över högpatogena organismer som ofta refereras till är USAs "Select agent" lista (http://www.selectagents.gov/). Listan inkluderar virus såsom smittkoppor, olika blödarfebervirus, SARS mm. Bakteriella smittämnen inkluderar mjältbrandsbakterien, pestbakterien, harpestbakterien mm.

Arvsmassan i de högpatogena bakteriella smittämnena har i de flesta fallen mycket låg variabilitet mellan olika stammar. Samtidigt finns det närbesläktade bakterier som inte alls är högpatogena, men som är nästan identisk med den högpatogena bakterien i stora delar av sin arvsmassa. I PCR baserad diagnostik har man försökt välja ut regioner som sällan förekommer i närbesläktade arter. När tvetydiga resultat uppkommer hamnar beslutsfattarna i svåra situationer. Man vill i största möjliga mån undvika att ge falska larm om högpatogena organismer samtidigt som man inte vill låta bli att varna om det föreligger en reell risk. Vid metagenomsekvensering är det inte ovanligt att det dyker upp sekvenser som är mycket lika högpatogener. Man måste då avgöra om det är sannolikt att dom kommer från en närbesläktad art eller om det faktiskt är troligt att man påvisat en högpatogen bakterie. Projektet har fokuserat mycket på att skapa metoder och referensinformation som gör att man får så bra förutsättningar som möjligt att avgöra detta.

2.3 Genomsignaturer

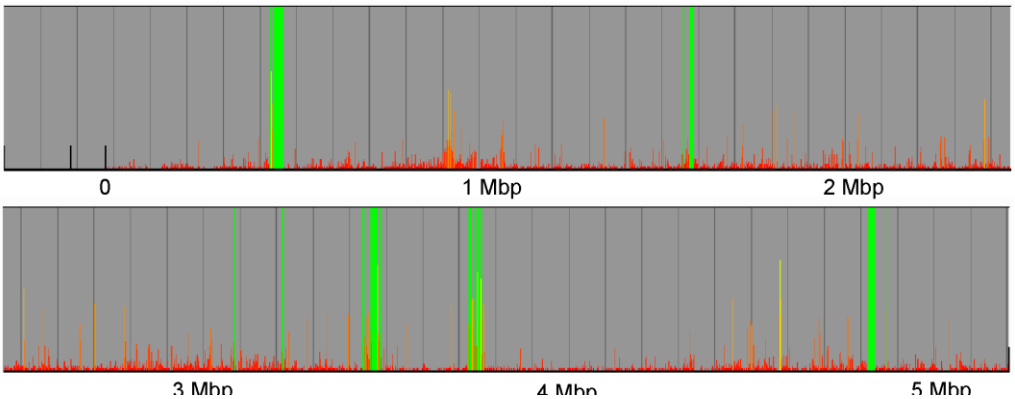
Om man hittar sekvenser som har likheter med högpatogena organismer i ett metagenomprov behövs verktyg och referensinformation för att avgöra om det är troligt att det är en nära besläktad bakterie eller om det finns anledning att se det som ett riktigt larm. För att få bästa möjliga beslutsunderlag har vi i detta projekt utvecklat ett koncept med "genomiska signaturer" som gör att man direkt, utan fördjupad analys, kan göra en tolkning huruvida en sekvens som liknar en specifik region i en högpatogen organisms genom skall tas på största alvar eller om den är vanligt förekommande i nära besläktade bakterier.

Den genomiska signaturen räknas ut genom att alla i analysen ingående genomsekvenser fragmenteras upp och datorjämförelser görs så att ett värde på likheten för varje genomisk region (fragment) fås mot alla i analysen ingående genom. Det blir således en allt mot alla jämförelse som skapar en stor databas med information om hur samtliga delar av samtliga genom är relaterade till alla andra analyserade genom. Databasen kan sedan utforskas grafiskt eller i tabellform för att identifiera genetiska markörer och signaturer för specifika grupper av bakterier (till exempel en högpatogen bakterie).

Signaturerna kan även användas för att skapa nya eller validera befintliga PCR analysmetoder. I vårat arbete har vi använt signaturanalys för att skapa en ny PCR metod riktad mot mjältbrandsbakterien samt utvärderat tidigare beskrivna PCR metoder. Mer information finns länkad till under fördjupad läsning.

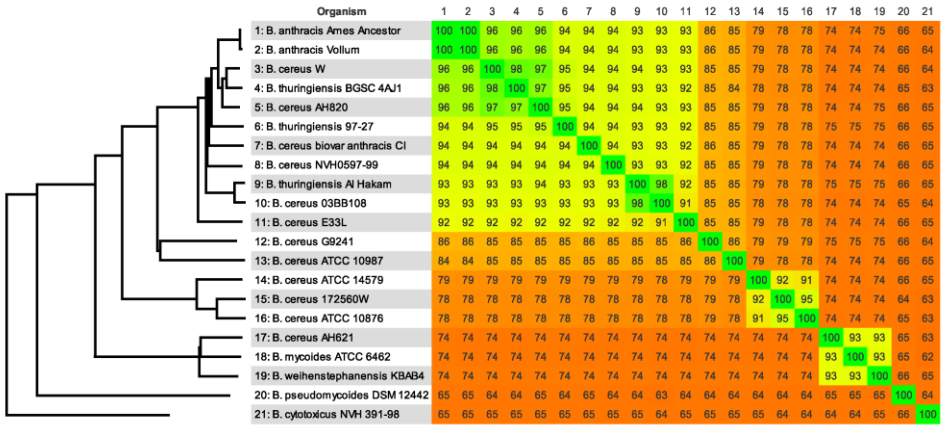
För att lätt kunna skapa de genomiska signaturerna har vi utvecklat ett datorprogram med grafiskt användargränssnitt, Gegenees. Programmets huvudfunktion är signaturanalysen men utöver detta kan programmet även organisera genomsekvenser, interagera med de internationella sekvensdatabaserna samt utvärdera PCR specificitet. Hur många genom som kan analyseras samtidigt beror på datorstorleken men utan att ha tillgång till superdatorer kan man jämföra ett hundratal genomsekvenser. När jämförelse väl är gjord finns ett det verktyg för att definiera och utforska genomiska signaturer, både grafiskt och i tabellform. Mer information om detta datorprogram finns länkad till under fördjupad läsning.

Ett exempel på en högpatogen signatur visas i figur 1. Denna figur visar en grafisk representation av mjältbrandsbakteriens signatur. de gröna signalerna utgör mycket specifika områden som, om man finner dem i sin metagenomsekvensering, signalerar att det är stor sannolikhet att man har en riktig mjältbrandsbakterie i provet.



Figur 1. Mjältbrandsbakteriens genomsignatur.

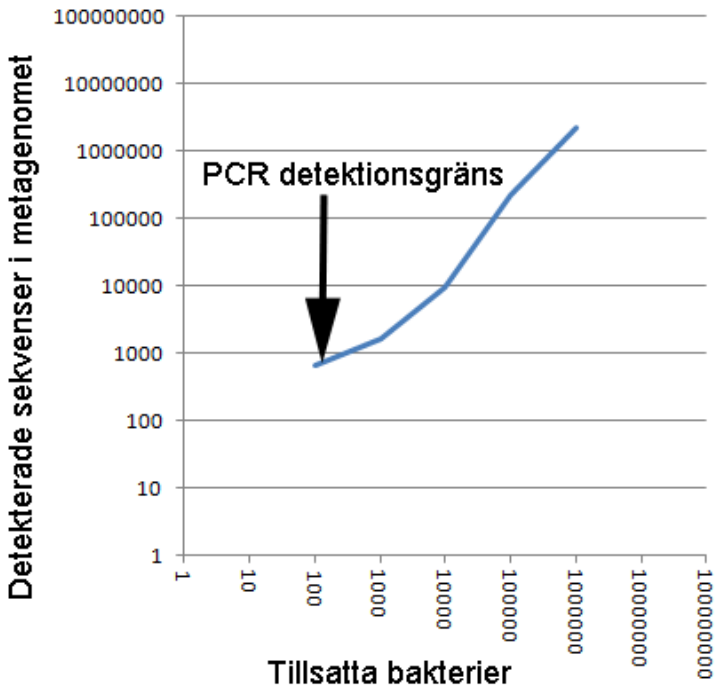
Datorprogrammet mäter även genomisk medellikhet vilket är ett snabbt sätt att gruppera nya genomsekvenser. Medellikheten kan visas i en tabell där färgkodningen visar graden av likhet (Figur 2).



Figur 2. Medellikheten mellan olika genomsekvenser.

2.4 Detektionsgränsen vid metagenomsekvensering

Metagenomsekvensering har visat sig vara ett mycket värdefullt verktyg för att identifiera infektiösa mikroorganismer när man inte vetat vad man letat efter. Till exempel identifierades det "nya" viruset Schmallenberg med hjälp av metagenomsekvensering. Ett inneboende problem med metagenomsekvensering är dock att desto mer bakgrunds DNA som finns desto mer måste man sekvensera för att få samma känslighet i analysen. Det har visat sig fungera väl med metagenomanalyser av vätskor som till exempel blod och cellodlingsmedium. I detta projekt har vi gjort en direkt jämförelse mellan detektionsgränsen för PCR och metagenomsekvensering i dricksvatten och funnit att den var minst lika bra för sekvenseringen (Figur 3). Detta resultat krävde dock att en signaturanalys gjordes för att filtrera bort sekvenser från bakgrundsfloran som liknade smittämnets.



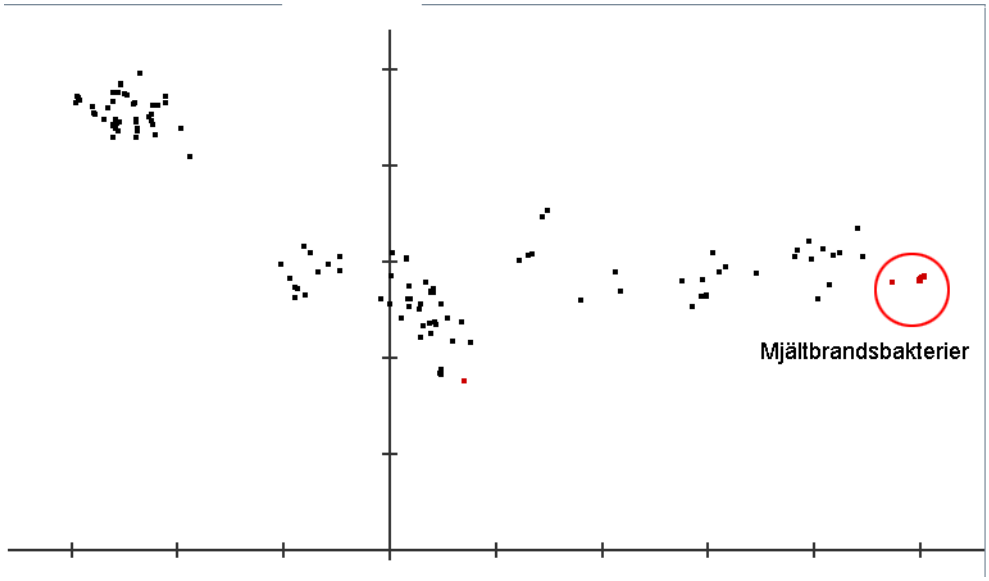
Figur 3. Jämförelse av detektionsgräns för metagenomsekvensering och PCR i 1.5 liter dricksvatten

"Smutsigare" prover som avföring, jordprover och dylikt är också möjligt att analysera men det krävs relativt höga kontaminationsnivåer för att det skall vara användbart utan att producera mycket stora mängder sekvensdata. Alltjämt kan detta vara mycket motiverat att, vid en oklar situation, genomföra för att göra den första identifieringen av vad man letar efter. När ett sjukdomsagens väl är definierat kan specifika PCR reaktioner utformas.

2.5 Gruppering med referensdata

Dataprogrammet som vi utvecklat kan användas för att gruppera genomsekvenser genom att räkna ut deras medellikheter. Detta är ett ganska grovt mått på likhet då variationerna kan vara stora då man jämför olika delar av samma genom. För att få en bättre placering av nya sekvenser (rena genomsekvenser eller metagenomsekvenser) gentemot de redan existerande referensdatabassekvenserna har vi även utvecklat en metod där vi jämför hela konserveringsprofiler snarare än bara ett medellikhetsvärde. Detta ger en mycket bättre och mer tillförlitlig separering av genomsekvenserna.

Metoden börjar med att ett "pan-genom" skapas. Det kan beskrivas som en sekvens som innehåller information om alla de gener som finns i hela populationen som utgör bakteriearten. Utifrån detta pan-genom görs en signaturanalys och sedan väljs regioner av pan-genomet ut som är associerade med stor sekvensvariation. Dessa utvalda regioner används för att göra konserveringsprofiler som sedan kan jämföras med avancerade multivariata dataanalysmetoder. Detta öppnar möjligheter för att mycket snabbt och med hög noggrannhet placera in en ny sekvens i referensdatasamlingen och även identifiera biomarkörer som definierar dessa grupperingar.



Figur 4 visualisering av en ny genomsekvens tillsammans med referensdata.

3. Fördjupad läsning

Websidor:

HUGO projektet:
http://www.hugo-international.org/

HMP projektet :
http://commonfund.nih.gov/hmp/index

Schmallenbergviruset:
http://news.sciencemag.org/plants-animals/2012/01/new-animal-virus-takes-northern-europe-surprise

Vetenskaplig litteratur:

Ågren J, Hamidjaja RA, Hansen T, Ruuls R, Thierry S, Vigre H, Janse I, Sundström A**,** Segerman B, Koene M, Löfström C, Van Rotterdam B, Derzelle S
In silico and in vitro evaluation of PCR-based assays for the detection of Bacillus anthracis chromosomal signature sequences.
Virulence. 2013 Nov 15;4(8):671-85
https://www.landesbioscience.com/journals/virulence/article/26288/?nocache=1222199637

Ågren J, Sundström A, Håfström T, Segerman B.
Gegenees: fragmented alignment of multiple genomes for determining phylogenomic distances and genetic signatures unique for specified target groups.
PLoS One. 2012;7(6):e39107.
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0039107

Agren J, Finn M, Bengtsson B, **Segerman B.**
Microevolution during an Anthrax Outbreak Leading to Clonal Heterogeneity and Penicillin Resistance
PLoS One. 2014 Feb 13;9(2)
http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0089112

Myndigheten för samhällsskydd och beredskap
651 81 Karlstad Tel 0771-240 240 www.msb.se
Publ.nr MSB 895-juni 2015