Norges miljø- og biovitenskapelige universitet

**Masteroppgave 2022 30 stp**
Fakultet for kjemi, bioteknologi og matvitenskap

# Genomanalyse av *Salmonella enterica* subsp. *enterica* serovar Typhimurium i Norge

Comparative genomics of *Salmonella enterica* subsp. *enterica* serovar Typhimurium in Norway

Andrea Obstfelder
Kjemi og bioteknologi

# Forord

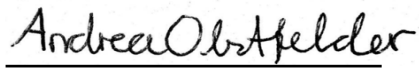
Denne masteroppgaven ble utført ved Norges miljø- og biovitenskapelige universitet ved fakultetet for kjemi, bioteknologi og matvitenskap i samarbeid med Veterinærinstituttet på Ås. Oppgaven er skrevet våren 2022 og har et omfang på 30 studiepoeng.

Jeg vil rette en stor takk til mine eksterne veiledere Karin Lagesen og Camilla Sekse, begge seniorforskere på Veterinærinstituttet, for muligheten til å skrive oppgaven hos dem. De har engasjert seg i oppgaven min, og hjulpet meg med tett oppfølging og god veiledning. I tillegg vil jeg takke Jeevan Karloss Antony-Samy, bioinformatiker ved Veterinærinstituttet, for timer og sene kvelder som har blitt brukt for å hjelpe meg i mitt arbeid.

Lars Snipen, førstemaneuensis ved biostatistikkgruppen ved KBM, har vært min interne veileder. Jeg vil takk ham for god undervisning som har inspirert meg til å velge en masteroppgave innenfor bioinformatiske analyser. I tillegg vil jeg takke for gode tilbakemeldinger og veiledning på slutten av skriveprosessen.

Avslutningsvis ønsker jeg å takke medstudent Oda Marie Bjørgum Karlsen for å ha vært en god venn og samarbeidspartner i ulike prosjekter gjennom mine fem år på studiet. I tillegg er jeg takknemlig for støtte og motiverende ord fra familie og venner under den siste innspurten.

Norges miljø- og biovitenskapelige universitet
*Fakultet for kjemi, bioteknologi og matvitenskap*
*Ås, 16.05.2022*



Andrea Obstfelder

# Sammendrag

*Salmonella enterica* subspecies *enterica* serovar Typhimurium, omtalt som *S.* Typhimurium i denne masteroppgaven, er en av de vanligste patogene serovariantene av *Salmonella* spp. *S.* Typhimurium har blitt isolert hos et bredt spekter av verter fra villfaunaen som er antatt ansvarlig for flere av de humane smittetilfellene av *Salmonella* i Norge. Hovedmålet med oppgaven var å teste hypotesen om at ville dyr er en potensiell smittekilde for *S.* Typhimurium blant produksjonsdyr i Norge.

Det ble i dette prosjektet plukket ut totalt 127 *S.* Typhimurium isolater tatt mellom 2001 og 2021 fra ville dyr, produksjonsdyr og sport- og familiedyr fra laboratoriebeholdningen hos Veterinærinstituttet. Ved hjelp av helgenomsekvensering (WGS) ble informasjon som fantes i genomsekvensen til *S.* Typhimurium isolatene undersøkt. WGS av patogener er et viktig verktøy for å forbedre forståelsen av smittsomme sykdommer som sprer seg mellom verter. Begrensninger i sekvenseringsteknologien gjør at DNAet kuttes opp i kortere fragmenter, og kun korte fragmenter opp til 300 bp sekvenseres, også kalt reads. Ved hjelp av bioinformatiske verktøy ble store mengder rådata fra sekvenseringen kvalitetsjekket før det opprinnelige genomet ble forsøkt rekonstruert ved hjelp av de novo SPAdes assembly.

For videre analyser ble ” Multilocus Sequence Typing” (MLST) brukt for bestemmelse av allele-variasjoner basert på 7 husholdningsgener, og hvert bakterieisolat fikk tildelt en sekvenstype der ST19 og ST34 var mest vanlige. Deretter ble evolusjonær likhet mellom isolatene innenfor samme ST vurdert ved SNP analyse. For fylogenetiske analyser ble det observert klustre med en evolusjonær likhet mellom ville dyr og produksjonsdyr som indikerer smitteoverføring. Basert på resultatene fra det brukte datasettet var det i flere tilfeller sterke indikasjoner på at ville dyr kan være smittekilden for *S.* Typhimurium blant produksjonsdyr i Norge.

# Abstract

*Salmonella enterica* subspecies *enterica* serovar Typhimurium, referred to as *S.* Typhimurium in this master`s thesis, is one of the most common pathogenic serovariants of *Salmonella* spp. *S.* Typhimurium has been isolated in a wide range of hosts from the wild fauna. These hosts are believed to be responsible for several of the human cases of *Salmonella* infections in Norway. The main goal of this thesis is to test the hypothesis that wild animals are a potential source of infection from *S.* Typhimurium among production animals in Norway.

In this project a totalt of 127 *S.* Typhimurium isolates taken between 2001 and 2021 from wild animals, production animals and sport- and family animals were selected from the laboratory inventory at the Norwegian Veterinary Institute. The information present in the genome sequences of the *S.* Typhimurium isolates were examined using whole genome sequencing (WGS). WGS of pathogens is an important tool to improve the understanding of how contagious diseases spread between hosts. Due to limitations in the sequencing technology, the DNA are fragmented and only up to 300 bp are sequenced, this is then called reads. Bioinformatic tools were used for quality check a large number of the raw data from the sequencing before the original genom was attempted reconstructed using de novo SPAdes assembly.

" Multilocus Sequence Typing" (MLST) were used in further analysis to determine allele-variations based on 7 housekeeping genes, and each bacterial isolate were assigned one sequence type where ST19 and ST34 were the most common. Then the evolutionary similarity between isolates within the same ST was assessed by SNP analysis. Using phylogenetic analysis we were able to observe clusters with evolutionary similarity between wild animals and production animals, which indicates that transmission of *S.* Typhimurium had taken place. Based on the results from the applied dataset we found several of these cases, which gives a strong indication that wild animals can be the source of infection for *S.* Typhimurium among production animals in Norway.

# Innholdsfortegnelse

# Ordforklaringer

*Forkortelser*

| | |
|---|---|
| cgMLST | Core genome MLST |
| DBG | De Bruijn-graf |
| DNA | Deoksyribonukleinsyre |
| MLST | Multilocus Sequencing Typing |
| MSIS | Meldingssystem for smittsomme sykdommer |
| NGS | Neste generasjons sekvensering |
| PCR | Polymerase-chain-reaction |
| PJS | Prøvejournalsystemet |
| SBS | Sequencing by synthesis |
| SNP | Single nucleotide polymorphisms |
| WGS | Whole genome sequencing |

*Engelske begreper*

| | |
|---|---|
| Assembly | Betyr montering og brukes i bioinformatikken om overlappende reads som blir koblet sammen til en eller flere sammenhengende sekvenser. Prosessen på norsk kalles «å assemblere». |
| Contig | En sammenhengende sekvens av assemblerte reads. |
| Input og Output | Input er dataen som skal analyseres for å oppnå en output fil med resultater. |
| Reads | I neste generasjons sekvensering referer en reads til DNA-sekvensen til ett fragment (en liten del av DNA). |
| Workflow | En workflow er direkte oversatt en arbeidsflyt. En workflow er i bioinformatikken et sett med sammensatte verktøy for å behandle rå sekvenseringsdata. |

# Innledning

## 1.1 *Salmonella* spp.

*Enterobacteriaceae* er en stor familie med gramnegative, fakultativt anaerobe og stavformede tarmbakterier som er utbredt i naturen (Bøvre, 2021). *Salmonella*, sammen med flere kjente bakterieslekter som blant annet *Escherichia*, *Shigella, Enterobacter* og *Yersinia* tilhører denne familien (Nelseon & Greene, 2022). *Salmonella* fikk navnet sitt etter Daniel E. Salmon som var den første til å isolere *Salmonella choleraesuis*, senere kalt *Salmonella enterica*, fra svinetarm (Monte & Sellera, 2020). I dag er *Salmonella* spp. delt inn i to arter som omfatter mer enn 2500 serovarianter funnet over hele verden.

*Salmonella* er en patogen bakterie som forårsaker infeksjonen salmonellose. De fleste serovariantene kan gi sykdom hos mange arter, inkludert mennesker. Symptomer som kan forekomme er feber, diare og magesmerter, og disse går som oftest over av seg selv. I noen tilfeller kan smitten føre til at infeksjonen blir mer alvorlig, og i verste fall føre til død. Salmonellose hos mennesker er en type A-sykdom som ved påvisning er meldepliktig til MSIS (Folkehelseinstituttet, 2019).

Ikke alle mennesker og dyr som er smittet med bakterien får symptomer. Disse er friske smittebærere, og kan skille salmonellabakterien ut med avføringen. *Salmonella* er en zoonotisk bakterie, som betyr at den kan overføres ved direkte kontakt med avføring eller indirekte mellom dyr og mennesker. Smitten fra dyr til mennesker kommer i hovedsak via mat, der rå kjøttprodukter og egg fra produksjonsdyr er vanlige kilder. Gjennom ulike overvåkingsprogram pågår det i Norge en aktiv overvåking av smitte hos produksjonsdyr for å begrense spredning. Veterinærinstituttet bistår med planlegging, analyser, bearbeiding av data og rapportering for denne overvåkingen. Dersom funn av *Salmonella* oppdages varsles Mattilsynet som iverksetter umiddelbare tiltak for å hindre smittespredningen (Veterinærinstituttet, 2022).

I motsetning til produksjonsdyr har ikke mennesker, villedyr (med unntak av villsvin), og sport- og familiedyr en aktiv smittesporing av salmonellose. Siden testingen utføres i hovedsak ved sykdomstilfeller er symptomfrie friske smittebærere blant disse gruppene vanskelig å oppdage. En passiv overvåking vil muligens føre til en underrapportering av antall smittede.

### 1.1.1 Typingsmetoder for *Salmonella*

Prosessen med å skille bakterier basert på deres fenotypiske og genotypiske forskjeller er kjent som «typing». Fenotypiske teknikker oppdager egenskaper uttrykt av mikroorganismen. Dette er egenskaper som form, størrelse, farge, og antigener som kan måles uten en referanse til genomet (Ferrari et al., 2017). Fenotypiske egenskaper til en bakterie kan være miljøpåvirket som gir en feil indikasjon på genotypen. Bakterier som er genetisk umulig å skille, kan dermed se ulike ut basert på fenotypiske egenskaper.

Genotypiske metoder innebærer å undersøke bakteriens DNA-sekvens ved hjelp av bioteknologiske analyser og sammenligning mot en referansesekvens. Samtidig som metoden er rask, gir den ofte bedre presisjon enn fenotypiske metoder. En vanlig typingsmetode for *Salmonella* er serotyping. Dette er tradisjonelt utført som en fenotypisk metode basert på gjenkjennelse av strukturen på overflateproteiner (Ferrari et al., 2017). I de fleste tilfeller vil det være forskjeller i DNA sekvensen som utgjør forskjeller i serotype og dermed kan genotypisk serotyping for *Salmonella* i de fleste tilfeller gi det samme resultatet som ved fenotypisk serotyping.

#### 1.1.1.1 Kauffman-White-Klassifisering

Typing basert på fenotypiske egenskaper har i mange år vært den tradisjonelle metoden brukt for karakterisering av *Salmonella* serovarianter. Kauffman-White-klassifisering (KWL) er et system som klassifiserer slekten *Salmonella* i serovarianter, basert på overflategener. Først bestemmes O-antigenet som er den ytterste delen av bakteriens overflate laget av lipopolysakkarider (Granum, 2017). Deretter bestemmes H-antigenet som kjennetegnes ved forskjellig proteininnhold i flagellen. Fasevariasjon som gir fenotypiske endringer brukes for å uttrykke to forskjellige H-antigener, H1 og H2. Monofasiske varianter av *Salmonella* har kun en variant av bakteriens flagell og kan dermed kun uttrykke ett H-antigen. Disse er ofte resistente mot flere typer antibiotika (Jarp, 2018).

Hvert O- og H antigen har et unikt kodenummer som brukes i kombinasjon for å klassifisere salmonellabakterien inn i serovarianter (Ibrahim & Morin, 2018). KWL-skjemaet deler *Salmonella* spp. inn i to arter: *Salmonella enterica* og *Salmonella bongori*. *Salmonella bongori* består av færre enn ti serovarianter som alle isoleres fra kaldblodige dyr (Granum, 2017). *Salmonella enterica* deles inn i seks underarter (subspecies) som igjen deles inn i mer enn 2500 ulike serovarianter (Kauffmann & Edwards, 1952):

- Subspecies I: *enterica*
- Subspecies II: *salamae*
- Subspecies IIIa: *arizona*
- Subspecies IIIb: *diarizonae*
- Subspecies IV: *houtenae*
- Subspecies V: *indica*

De fleste serovariantene er å finne i subspecies I, *enterica*. Denne underarten inneholder alle serovariantene, med noen unntak, som har blitt isolert fra menneske og andre varmblodige dyr. Blant disse serovariantene finner vi *Salmonella enterica* subspecies *enterica* serovar Enteritidis og *Salmonella enterica* subsp. *enterica* serovar Typhimurium, som de mest vanlige i Norge.

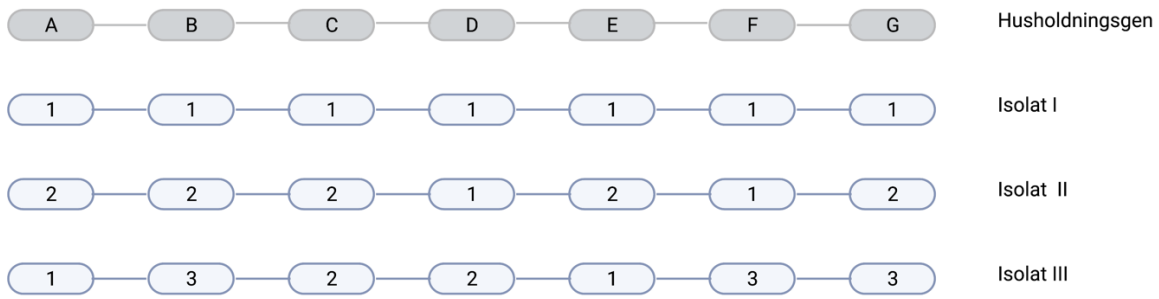
#### 1.1.1.2 Multilocus Sequence Typing (MLST)

Serotyping basert på fenotypiske egenskaper er en tidkrevende metode og krever et godt trent personell for å tolke resultatene korrekt. De siste årene har molekylære metoder gradvis erstattet fenotypisk analyser for typing av bakteriestammer. Molekylære typingsmetoder brukes hovedsakelig for å finne kilden i smitteutbrudd, men kan også brukes til å identifisere antibiotikaresistensgener og/ eller virulensgener mikroorganismer innehar. Ett eksempel på en kjent molekylær typingsmetode er Multi Locus- Sequencing Typing (MLST).

MLST er et eksempel på en molekylær typingsteknikk som har sekvenstype (ST) som resultat, og effektivt beskriver en bakteriepopulasjon for en organisme. Tradisjonelt involverer metoden PCR amplifikasjon av husholdningsgener ved hjelp av spesifikke primere. Dette er en tidkrevende prosedyre som er i ferd med å bli erstattet av MLST basert på helgenomsekvenserings-data. MLST er en gunstig tilnærming når nært beslektede bakteriearter studeres siden teknikken hjelper til med å oppdage variasjoner på nukleotidnivå (Maiden, 2006).

Teknikken baser seg på en rekke valgte husholdningsgener (vanligvis syv) som koder for de mest primære oppgavene. Genene som brukes skal prinsipielt finnes i alle isolater til den aktuelle arten. Hver bakterieart har sitt eget sett av husholdningsgener. For *Salmonella* er disse: *aro*C, *dna*N, *hem*D, *his*D, *pur*E, *suc*A og *thr*A (Leekitcharoenphon et al., 2012). Gensekvensen til de syv genene sammenlignes med DNA sekvensen fra reads eller det

assemblerte genomet (Larsen et al., 2012). På denne måten er det mulig å se varianter i et lokus der kun en base er forskjellig. Dersom nukleotidsekvensen er ulik anses den å være en ny allel og tildeles et unikt allelenummer, vist i figur 1 (Yoshida et al., 2016). Ved å tildele hver allel en merkelapp med et nummer kan en se på kombinasjoner av tall som er oppnådd på tvers av flere gener.




*Figur 1 Multilocus Sequencing Typing basert på de syv husholdningsgenene: aroC, dnaN, hemD, hisD, purE, sucA og thrA. Figuren er laget ved bruk av biorender.*

Kombinasjonene av merkelappene fra figur 1 lager en profil som tilordnes en sekvenstype for hvert isolat (Feil et al., 2004). Et eksempel på tildeling av ST basert på alleleprofilene er vist under:

| | | |
|---|---|---|
| Isolat I | 1-1-1-1-1-1-1 | Sekvenstype 1 (ST-1) |
| Isolat II | 2-2-2-1-2-1-2 | Sekvenstype (ST-66) |
| Isolat III | 1-3-2-2-1-3-3 | Sekvenstype (ST-98) |

Metoden er den mest brukte for typing av patogene bakterier. En har i tidligere studier sett at ulike *Salmonella* serovarianter er assosiert med ulike sekvenstyper. For eksempel er sekvenstypene ST19, ST313 og ST34 ofte assosiert med *S.* Typhimurium. I studien til Achtman i 2012 ble det funnet monofasiske varianter av *S.* Typhimurium i ST34 (Achtman et al., 2012)

Ulempen er begrensningen ved å kun se på syv gener, som fører til at en ikke får kartlagt forskjellene i resten av genomet. I takt med at teknologien har forbedret seg har MLST blitt utvidet til flere gener. cgMLST (Core genome MLST) er et eksempel på en utvidelse av MLST-konseptet som baserer seg på eksistensen av et sett med kjerne-gener (Enright et al., 2000). Dette er gener som deles av alle genomer innenfor arten.

### 1.1.2 *S.* Typhimurium

«*Salmonella enterica* subspecies *enterica* serovar Typhimurium» omtales for enkelhets skyld som «*S.* Typhimurium» i denne oppgaven. *S.* Typhimurium forårsaker 10-20% av smittetilfellene for *Salmonella* blant mennesker i Norge, og er en av de vanligste patogene serovariantene for *Salmonella* spp. Tidligere studier har isolert *S.* Typhimurium hos et bredt spekter av verter fra villfaunaen hos blant annet ville fugler og piggsvin som er ansvarlige for flere av smittetilfellene i Norge (MacDonald et al., 2018). Ville dyr kan gjennom forurensing av miljøet fungere som effektive spredere til blant annet produksjonsdyr (Refsum et al., 2002).

Produksjonsdyr som blant annet storfe, svin og fjørfe kan bli smittet av ville dyr som har levd i samme omgivelser, men også faktorer som fôr, transport og slakting kan bidra til smitte. Etter slakting kan bakterien komme fra blant annet produksjon, oppbevaring og tilberedning. Bakterien kan overleve flere måneder på overflater og formeres under visse vilkår (Veterinærinstituttet, 2022). På grunn av sin evne til å spre seg i produksjonskjeden og i miljøet i hvilket som helst næringsmiddel kan smittesporingen for *S.* Typhimurium være komplisert.

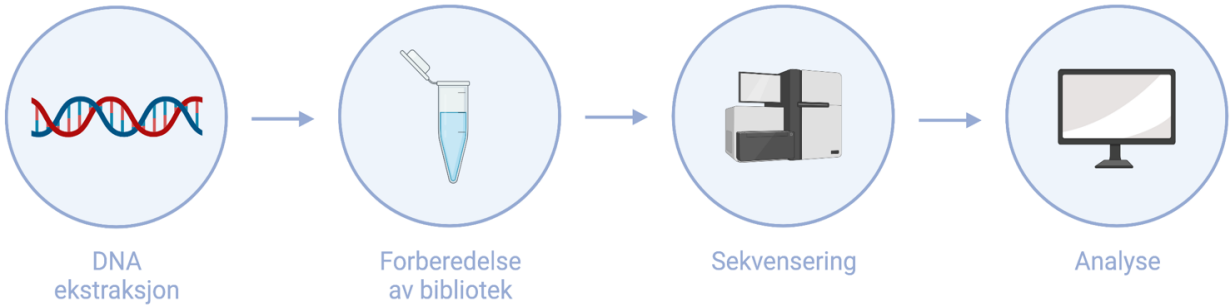
## 1.2 DNA sekvensering

Rekkefølgen av nukleinsyrer i polynukleotidkjeder inneholder informasjon om arvelige og biokjemiske egenskaper til levende liv. Evnen til å se på slike sekvenser er avgjørende for biologisk forskning.

Historien om DNA begynner i 1953 da Watson og Crick oppdaget strukturen til DNA. I 1964 utførte Richard Holley sekvensering av tRNA som det første forsøket på å sekvensere nukleinsyren (Heather & Chain, 2016). Biokjemikeren Fredrik Sanger utviklet i 1977 den første metoden som kunne lese baserekkefølgen til DNA-molekylet, herav navnet Sanger sekvensering. Denne førstegenerasjonsmetoden baserer seg på dideoksyribonukleotider (ddATP, ddCTP, ddTTP og ddGTP) i DNA polymerase reaksjon. Dideoksyribonukleotider mangler en hydroksylgruppe (3`OH) som avbryter kjedeforlengelsen og stopper reaksjonen. DNA-tråder av ulik lengde blir deretter analysert med gelelektroforese for å få frem rekkefølgen av basene i sekvensen (Sanger et al., 1977).

Forskere fra hele verden har i løpet av det siste halve århundre brukt mye tid og ressurser på å utvikle og forbedre teknologier for å lette sekvenseringen av DNA- og RNA-molekyler. Fra å tidligere bare kunne sekvensere korte oligonukleotider, har vi i dag teknologien til å sekvensere millioner av fragmenter i parallell med dypsekvensering. Denne høykapasitetssekvensering er mer effektiv, samtidig som kostnaden for sekvenseringen er lavere. En dominerende dypsekvenseringsteknikk som brukes i dag er Neste generasjonssekvensering (NGS), også kalt «high-throughput sequencing». NGS er i likhet med Sanger sekvensering en nukleotidsekvenseringsteknikk. Begge teknikkene bruker fluoriserende merking for å identifisere nukleotider som ved hjelp av DNA-polymerase legges til en etter en på en voksende DNA-templatstreng (Illumina, 2022b).

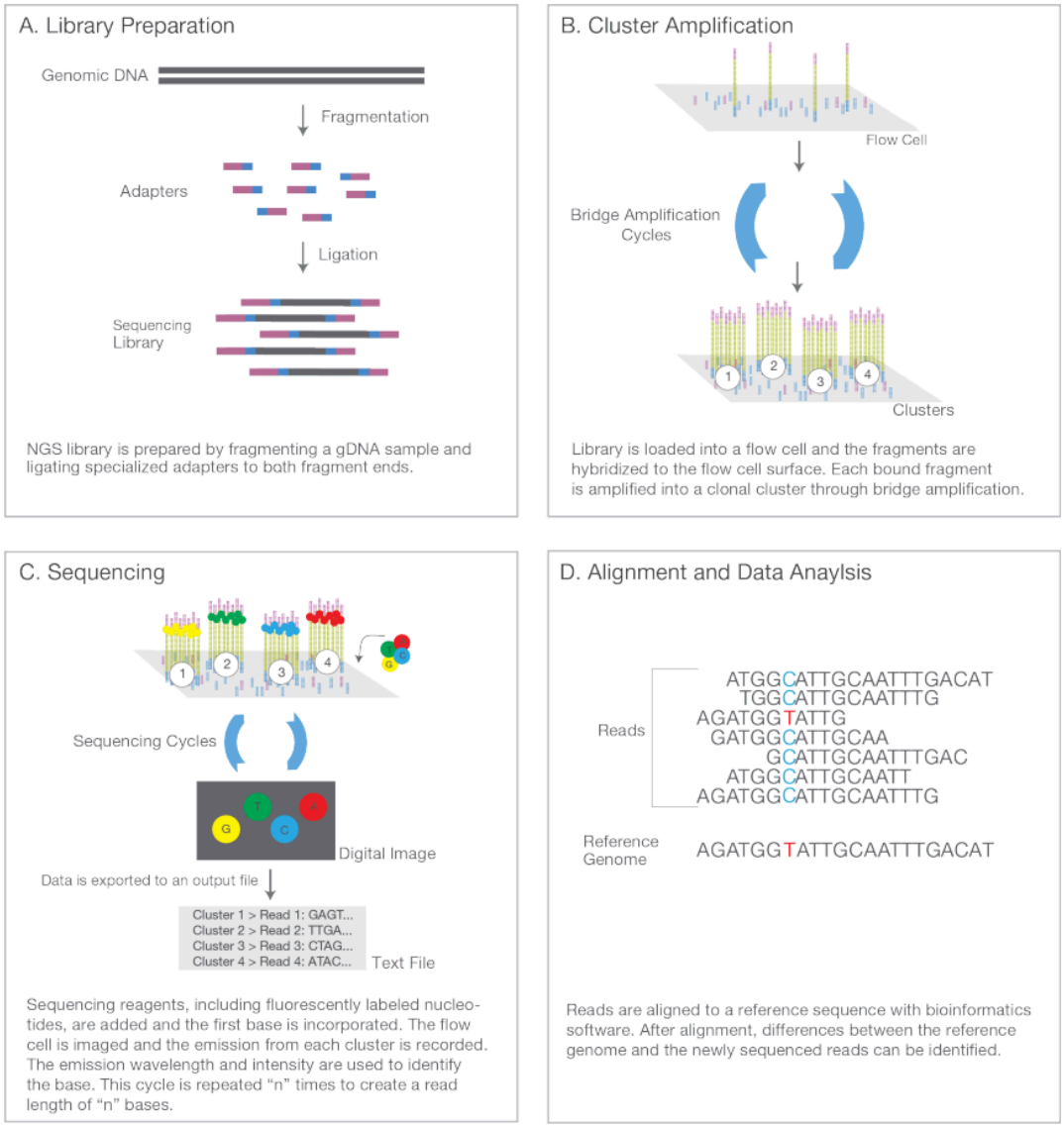
Sanger sekvensering sekvenserer kun et enkelt DNA-fragment av gangen, mens NGS sekvenserer millioner av fragmenter parallelt, noe som gir en raskere behandlingstid. Det er flere tilgjengelige plattformer for NGS metoder på markedet, der Illumina er blant de mest populære. Sekvenseringsteknikken baserer seg på hovedstegene illustrert i figur 2. Prosessen innebærer å fragmentere ekstrahert DNA i flere deler, legge til adaptere, sekvensere bibliotekene og sette dem sammen for å rekonstruere en genomisk sekvens (Illumina, 2022a).




*Figur 2 Hovedstegene for Neste generasjonssekvensering (NGS). Ekstrahert DNA blir fragmentert, adaptere legges til, biblioteket sekvenseres og genomet rekonstrueres. Figuren er laget ved bruk av biorender.*

### 1.2.1 Sequence by synthesis (SBS)

Den mest dominerende NGS-teknologien er Illumina sekvensering som bruker «sequencing by synthesis» (SBS) for å kartlegge baserekkefølgen. Metoden består av fire grunnleggende trinn: forberedelse av prøven, «cluster generation», sekvensering og dataanalyse illustrert i figur 3.




*Figur 3 Metoden for Neste generasjonssekvensering (NGS) som bruker «sequencing by synthesis» for å kartlegge baserekkefølgen. Inkluderer fire steg: Bibliotekforberedelse (A), «Cluster generation» (B), Sekvensering (C) og dataanalyse (D). Figuren er hentet fra (Illumina, 2022b).*

Prosessen begynner med at renset DNA fragmenteres og spesialiserte adaptere legges til endene av fragmentene, som danner et genomisk bibliotek. Adaptere inneholder segmenter som fungerer som referansepunkter under amplifikasjonen og sekvenseringen (A). Det modifiserte DNAet appliseres på en «flowcell» hvor amplifikasjonen og sekvenseringen finner sted. Flowcellen er en glassplate med brønner som består av oligonukleotider som er komplementære med adaptereregionene på fragmenttrådene. Etter at fragmentene har festet seg starter klusteringen. Dette trinnet resulterer i millioner av kopier av enkelttrådet DNA fra broamplifikasjon (B). Nukleotider med fluorescerende merker binder seg til DNA templattråden, som indikerer hvilket nukleotid som er tilsatt. Ved sekvensering av par gjentas prosessen for den omvendte tråden (C). Ved sekvensering fra begge ender av

genomfragmentet produseres paired end-reads (R1 og R2). R1 er reads som har blitt sekvensert i den første runden av illumina sekvenseringsmaskinen, mens R2 er reads som ble sekvensert i andre runde. Etter sekvenseringen kan instrumentprogramvaren identifisere nukleotidene (D) (Illumina, 2022b).

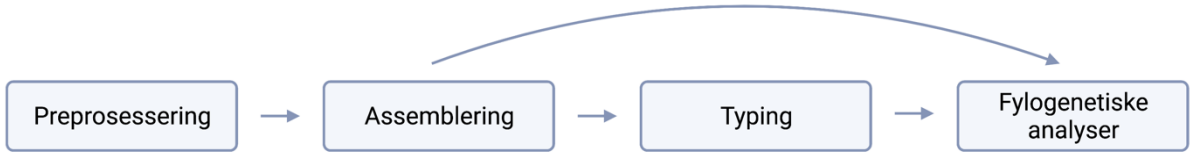
### 1.3 Bioinformatiske analyser

Grunnet begrensninger i NGS-teknologien er det ikke mulig å sekvensere genomet fra ende til ende. Sekvensene som kommer ut fra sekvenseringsmaskinen kalles reads. Reads fra NGS-maskiner er korte DNA fragmenter som til sammen potensielt dekker hele genomet. Lengden på readsene er avhengig av hva slags sekvenseringteknologi og flowcelle som er blitt brukt.

Reads fra sekvenseringen må settes sammen igjen for å rekonstruere genomet. Rådataen med de korte sekvensene samles i to FASTQ-filer (R1 og R2) som må bearbeides bioinformatisk for å kunne tolkes. En FASTQ-fil er en tekstfil som inneholder blant annet bokstavrekkefølge og en tilsvarende kvalitetsverdi kodet som ASCII-tegn, som indikerer kvaliteten på det avleste nukleotidet (Johnson et al., 2021).

Prosessen fra sekvensdata til et resultat med biologisk betydning innebærer flere steg. Det er i dag flere tilgjengelige verktøy som gir mange valgmuligheter. Valg av ulike innstillinger og programmer kan være avgjørende for endelig resultat, og det er derfor viktig å sette seg inn i valgmulighetene og velge det som passer best for formålet.

Stegene brukt i denne oppgaven for analyse av rådatafilene frem til illustrering av genomet er vist i figur 4. Stegene innebærer forbehandling av rådatafilene fra sekvenseringen (preprosessering), rekonstruering av det opprinnelige genomet (assemblering), identifisering av serovarianter (typing) og til slutt en fylogenetisk analyse. For fylogenetiske analyser benyttes data fra assembleringen som input og en kan dermed gå direkte fra assembleringen til fylogeni.



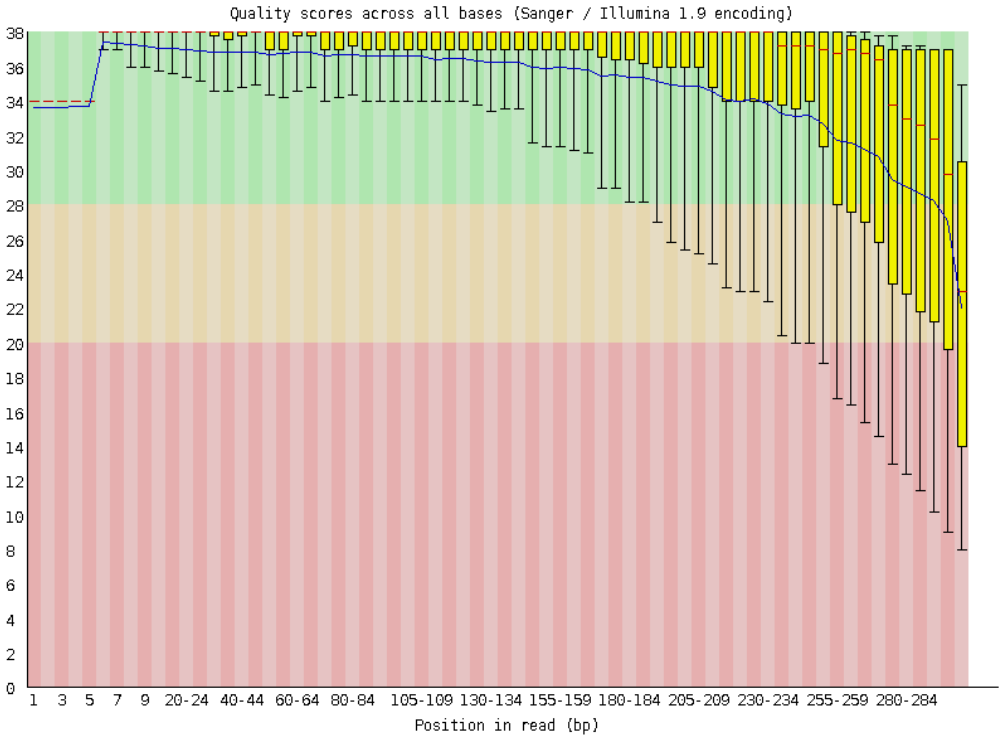

*Figur 4 Stegene for bioinformatiske analyser utført etter sekvensering. Figuren er laget ved bruk av biorender.*

### 1.3.1 Kvalitetskontroll og preprosessering av rådata

Det første trinnet under preprosessering er å identifisere og sjekke reads av dårlig kvalitet som kan ha oppstått fra feil under forberedelse av biblioteket og sekvenseringsfeil. Det er mye arbeid å sjekke kvaliteten på flere millioner reads manuelt. Kvalitetsverdiene, også kalt PHRED score, til rådataen plottes derfor ofte inn et Score Quality plot. Figur 5 er et eksempel på en slik plot som viser kvalitetsverdier. Hver stolpe er et boxplot som favner 50% av verdiene. X-aksen viser posisjonen i reads og y-aksen viser kvaliteten målt i PHRED score. PHRED score regnes ut ved å bruke formel 1 (Illumina, 2011).

$$PHRED\ score\ =\ -10 * log_{10}(P_{err}) \qquad (1)$$

Dersom PHRED score er på 20 eller høyere betyr det at 99% kvalifiserer seg som nøyaktig og 1% sjanse for feil. I figur 5 viser den blå streken gjennomsnitt for PHRED score og den røde streken viser median. Befinner den blå streken seg i den grønne regionen indikeres en god kvalitet. Y-aksen er delt inn i tre deler etter svært god kvalitet (grønn), rimelig god kvalitet (oransje) og dårlig kvalitet (rød). Baseparene med dårlig kvalitet tyder på sekvenseringsfeil, og må bearbeides før ytterligere analyser. Før assembleringen fjernes sekvensene med lav kvalitet og dermed den tilhørende ASCII-koden.




*Figur 5 Eksempel på en Score Quality plot hentet fra FastQC for isolatet 2021-04-15832 R1. X-aksen viser posisjonen i reads og y-aksen viser kvaliteten målt i PHRED score.*

I figuren 5 blir gjennomsnittet på kvalitetscoren relativt dårligere for de siste baseparene. R2 er ofte av dårligere kvalitet enn R1 som i stor grad skyldes at den stegvise syntetiseringen blir gradvis dårligere, blant annet fordi kjemikaliene har vært lengre på maskinen. Nukleotider med fluorescerende merker blir mer utydelig ved avlesning, og dermed hvilket nukleotid som har blitt syntetisert.

Før Illumina sekvenseringen festes adaptersekvenser til DNA fragmentene. Disse sekvensene er tekniske og blir ofte fjernet igjen av Illumina maskinen etter sekvenseringen. Det kan likevel finnes rester av adapterne til stede i readsene som forurenser dataen og gir dårlig kvalitet. Trimmomatic er et raskt verktøy som blir brukt til å trimme basepar med dårlig kvalitet og fjerne adaptere i Illumina (FASTQ) dataen. Ved å fjerne adaptere før en assembly forhindres falske overlappinger som gir feilaktige contigs/transkripsjoner under assembleringen (Bolger, Lohse, & Usadel, 2014).

### 1.3.2 Assemblering

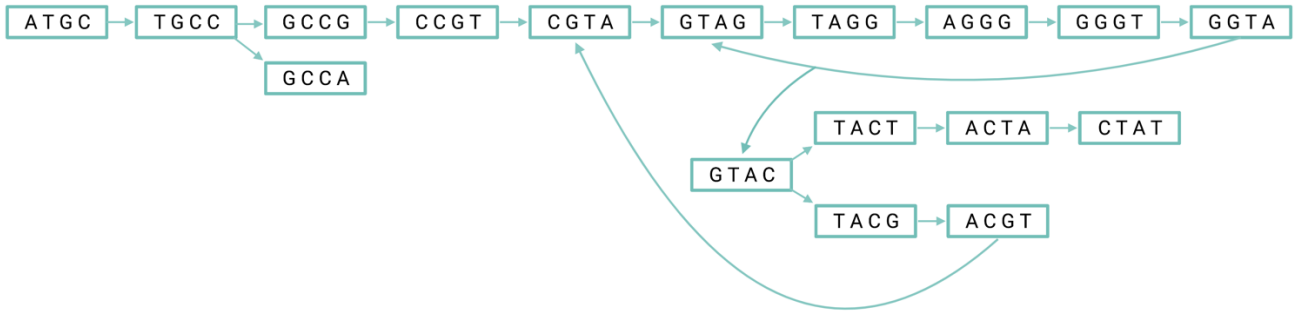
Etter DNA sekvensering er genomet delt opp i kortere sekvenser, kalt reads, samplet fra forskjellige steder i genomet. Rekonstruering av genomet er et av de første trinnene ved analysering av organismens genom. Det er viktig med en vellykket assemblering for identifisering av gener i senere analysemetoder.

Ved assemblering justeres og kombineres reads fra DNA sekvensen til lengre sekvenser kalt contigs. Contigs blir deretter satt sammen for å danne Scaffolds, som består av sekvenser adskilt med mellomrom, kalt «gaps». Dette kan gjøres både med og uten ett referansegenom (Lischer & Shimizu, 2017). Et referansegenom er et tidligere assemblert genom som brukes som en guide for å sette sammen det sekvenserte genomet. Fordelen med en referansebasert assemblering er at den er enklere å bruke og justeringen av reads gjøres raskt med et genom å ta utgangspunkt i. Samtidig som denne metoden krever en passende sekvens som kan brukes som referanse, begrenser den også mulighet til å oppdage nye sekvenser som ikke finnes i referansegenomet. De novo assemblering som er mest brukt for bakterier i dag kan oppdage nye sekvensrekkefølger og alleler uten hjelp av et referansegenom. Genomsekvensen blir gjenskapt gjennom overlappende sekvenserte reads som settes sammen til contigs (Lischer & Shimizu, 2017).

Det finnes flere tilgjengelige verktøy for assemblering som kan gi varierende resultater. Forskjellige sekvenseringsteknologier produserer ulike data, og det er derfor viktig å velge best tilpasset verktøy basert på sekvenseringsplattformen som er bitt brukt. I denne oppgaven ble NGS med Illumina brukt som gir korte reads. To kjente NGS assembly er Overlap/Layout/Consensus (OLC) og De Brujin Graph (DBG). OLC avhenger av en overlappende graf og er mer egnet for lange reads med lav coverage, som Sanger sekvensering. DBG-algoritmen bruker en form for K-mer graf og er i motsetning til OLC mer egnet for korte reads med høy coverage (Li et al., 2011).

#### *1.3.2.1* De Brujin graf (DBG)

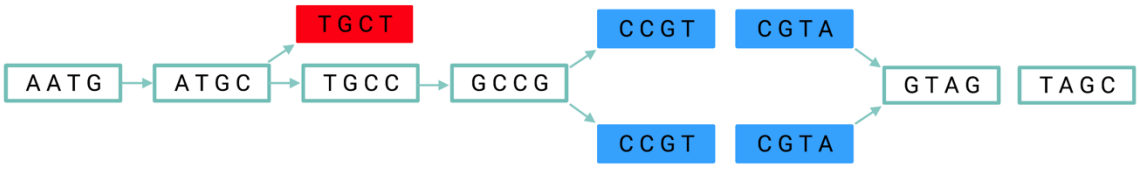
En vanlig de novo assembly metode er de Brujin graf (DBG) som bruker reads delt opp i k-merer for å konstruere en graf. DBG deler readsene opp i k-1-merer. Hver unike kmer med k-lengde produserer en node. To noder kobles sammen hvis k-meren har en k-1 overlapp. Det vil si at dersom de siste k-1-tegnene (suffiks) i k-mer representert av A, er like de første k-1-tegnene (prefiks) i k-mer representert av B vil A og B bindes sammen (Maduranga, 2020). Et eksempel på de Brujin graf er illustrert i figur 6.




*Figur 6 Eksempel på en de Brujin graf. Figuren er inspirert fra https://en.wikipedia.org/wiki/File:Example_1seq.pdf og laget ved bruk av biorender.*

For et perfekt genom kan readsene lage en sti der hver og en av nodene er representert en gang. Stien kalles Eulersk sti, og skal rekonstruere den opprinnelige genomsekvensen (Bankevich et al., 2012). Feil forårsaket av sekvenseringsprosessen eller forurensninger i prøven skaper utfordringer. Med optimalisering kan grafen ved hjelp av forenklinger og feilfjerninger bidra til å inkludere flere av nodene.

Assemblering av korte reads (fra Illumina) har høy nøyaktighet, men har begrenset evne til å assemble repeterende sekvenser. Repeterte regioner som er spred utover genomsekvensen vil kunne slå seg sammen til større sammenhengende sekvenser som ikke vil ha unike DNA-sekvenser på noen av sidene. Dette gjør assemblering av repeterte regioner som er lengre enn readlengde umulig (Baptista et al., 2018). I en DBG vil repeterende sekvenser føre til at to forskjellige stier starter og slutter ved de samme nodene, kalt bobler (vist med blått i figur 7). For å finne ut hvilken sti som skal beholdes og hvilken som skal fjernes kalkuleres Kmer coverage for hver sti. Forgreninger kalt tip er en annen feil som kan forekomme (vist med rødt i figur 7). Dette skjer dersom det er feil på slutten av en read, slik at nodene er frakoblet grafen i en av endene (Leggett et al., 2013). Slike feil skaper forvirring når en skal velge sti for å rekonstruere den opprinnelige genomsekvensen, og må fjernes fra grafen. Selv etter forenklingene vil det nesten alltid være regioner i grafen vi ikke kan inkludere i et større contig, og disse vil da danne sine egne små contiger (Li et al., 2011). Forventet størrelser på contiger fra assembleringen avhenger mye av størrelsen på genomet og typen sekvenseringsdata.




*Figur 7 Eksempel på bobler (merket med blått) og tips (merket med rødt) i en de Brujin graf. Figuren er laget ved bruk av biorender.*

Valg av riktig assembleringsverkøy avhenger av hva brukeren er interessert i å få ut av genom assembleringen, samt tilgjengelig budsjett. Et vanlig brukt assembleringsverktøy for korte reads som baserer seg på DBG algoritmen er SPAdes. Spades går gjennom 4 steg: konstruksjon av graf, justeringer og beregninger i forhold til avstand mellom parede k-merer i read-par, konstruksjon av graf med hensyn til read-par og konstruksjon av contigs (Bankevich et al., 2012).

#### 1.3.2.2 Kvalitetsjekk av assembly

Det er mange utfordringer knyttet til assemblering av genomet. En av de største utfordringene er håndtering av feillesninger. Avlesningene etter sekvenseringen vil aldri vært helt nøyaktig. For å håndtere dette kan den samme delen av DNA sekvenseres flere ganger. Antall ganger en

seksjon ble sekvensert kalles «coverage». Coverage av et datasett er det totale antallet baser i den leste dataen delt på det totale antallet basepar i det sekvenserte genomet. Verdien kan beregnes ved hjelp av formel (2) og blir ofte brukt som en faktor i kvalitetssjekk og eventuell fjerning av baser. Det er vanlig med mellom 30 og 50 X coverage for å få en vellykket assembly. Jo høyere coverage en har jo flere korte reads av DNA-sekvensen.

$$Coverage = \frac{Antall\ reads * Read\ lengde}{Total\ genomstørrelse} \quad (2)$$

QUAST (Quality Assessment Tool) er et verktøy for kvalitetsvurdering av assembly. Verktøyet kan vurdere sammenstillinger både med og uten et referansegenom (Gurevich et al., 2013). Verktøyet genererer en rapport med statistiske verdier og beregninger. Noen av disse verdiene er N50, antall contigs og total lengde på assembly. N50 verdien beregnes ved å sortere alle contigs etter lengde. Sekvenslengden til den korteste contigen ved 50% av den totale genomlengden gir N50 verdi.

### 1.3.3 Typing av helgenomsekvenserte data

DNA-baserte typingsmetoder er viktig for utbruddsoppklaring, smittesporing og identifisering av smittekilde som kan redusere smittepress. Typing av helgenomsekvenserte *Salmonella* isolater kan redusere salmonellose og øke mattryggheten. *Salmonella* in Silico Typing Resource (SISTR) er en bioinformatisk plattform som ble utviklet for å forutsi nøyaktig serovariant for assemblerte helgenomsekvensammenstillinger. Plattformen har flere sekvensbaserte typingsanalyser integrert som finner serovarianten. I denne oppgaven fant SISTR serovarianten ved bestemmelse av antigen, cgMLST-gen alleler og MASH som estimerer genom- og metagenomavstanden. Ved bestemmelse av antigen finner SISTR H1 og H2/flagell antigen og O/somatisk antigen. Den antigene profilen blir brukt til å identifisere serotype referert til Kauffman-white referanse katalog. cgMLST baserer seg på 330 kjernegener (core genes) identifisert gjennom en komparativ genomisk analyse basert på et sett med assemblerte genomer med best kvalitet i datasettet. BLAST blir brukt for å sammenligne DNA-sekvensen med data fra databaser (Yoshida et al., 2016). MASH identifiserer serovarianten ved å estimere genomavstanden ved bruk av MinHash-algoritmen. Jo likere to genomer er, jo flere MinHashes vil sannsynligvis matche (Ondov et al., 2016).

SISTR er en molekylær typingsmetode som baserer seg på genotypiske egenskaper for å finne serovarianten. Metodene mangler den nødvendige oppløsningen for å identifisere utbrudd forårsaket av nært beslektede bakterievarianter. En måte å utnytte WGS-data på er identifisering av SNP-er som varierer mellom isolater. SNP er en forkortelse for «Single-nucleotide polymorphism» som representerer en forskjell i en enkelt base i DNAet:


SNP-er er svært informative og er i stand til å avsløre evolusjonshistorier til homogene grupper, samt oppdage utbrudd (Pearce et al., 2018). Tilnærmingen kan være referansefri eller referansebasert, der sistnevnte er mest brukt. Ved bruk av referanse justeres sekvenserte reads til et nært beslektet referansegenom for å identifisere SNP-er. DNA-sekvensene som er felles mellom de sekvenserte isolatene og referansegenomet kan deretter analyseres basert på forskjellige SNP-er.

### *1.3.4* Fylogenetiske analyser

For å finne smitteoverføringen mellom ville dyr og produksjonsdyr må en finne ut hvordan *S.* Typhimurium bakterien i de ulike isolatene er relatert. Fylogenetiske analyser basert på helgenomsekvensering ser på den evolusjonære utviklingen og kartlegger slektskap mellom organismene (Dutta, 2021). Et fylogenetisk tre er en evolusjonær rekonstruksjon som viser den genetiske nærheten mellom organismene. Konstruksjonen bygges opp via klustering, som grupperer dataen i delmengder kalt klynger koblet sammen med grener. Lengden på grenene viser den genetiske avstanden mellom individene og en kan på denne måten se hvor lenge det er siden artene hadde en felles stamfar. Tidligere var dataene i fylogenetiske analyser basert på anatomiske likheter og forskjeller mellom artene. Trærne baserer seg nå i hovedsak på genetiske og molekylære data (Scott & Baum, 2016).

## 1.4 Målsetting for oppgaven

Hovedmålet med oppgaven er å teste hypotesen om at ville dyr er en potensiell smittekilde for *S.* Typhimurium blant produksjonsdyr i Norge. For å svare på hypotesen skal helgenomsekvenser (WGS) av *S.* Typhimurium fra ville dyr, produksjonsdyr og sport- og familiedyr bearbeides, analyseres og tolkes ved bruk av bioinformatiske analyser. Deretter skal en vurdere om datasettet som er tilgjengelig i oppgaven tillater å svare på om isolatene fra produksjonsdyr og ville dyr er like nok til å vurdere om de er fra samme kilde.

*Hovedmål*

- Er ville dyr en potensiell smittekilde av *S.* Typhimurium til produksjonsdyr i Norge?

*Delmål*

- Tillater datasettet som er tilgjengelig i oppgaven å svare på hovedmålet?
- Er isolatene fra produksjonsdyr og ville dyr like nok til å vurdere om de er fra samme kilde?
- Hvordan må eventuelt datasettet se ut for å svare på problemstillingen?

# Materialer og metode

## 1.5 Isolering, dyrking, DNA ekstraksjon og sekvensering

*Utført av Veterinærinstiuttet.*

Isolering, dyrking og DNA ekstraksjon ble utført av Veterinærinstituttet i henhold til en brukerveiledning for ekstraksjon av bakterielt genomisk DNA til dypsekvensering. Brukerveiledningen er utformet av Veterinærinstituttet for bruk i videre helgenomsekvensering (WGS) som krever DNA av god kvalitet.

Ekstrahert DNA ble sekvensert på Illumina Miseq og/eller NextSeq sekvenseringsmaskin. Miseq brukes rutinemessig på Veterinærinstiuttet for WGS av bakterier. Fordelen med Miseq er at den gir read-sekvenser med lengde på ca. 300 bp, mens NextSeq gir reads med kortere lengde på 150 bp. For isolatene i datasettet brukt i oppgaven var 109 av isolatene sekvensert med Illumina Miseq og 18 av isolatene med Illumina NextSeq. Isolatene sekvensert med Illumina Miseq var samlet fra ulike kjøringer, mens isolatene sekvensert med Illumina Nextseq var fra samme kjøring.

### 1.5.1 IRIDA og Galaxy

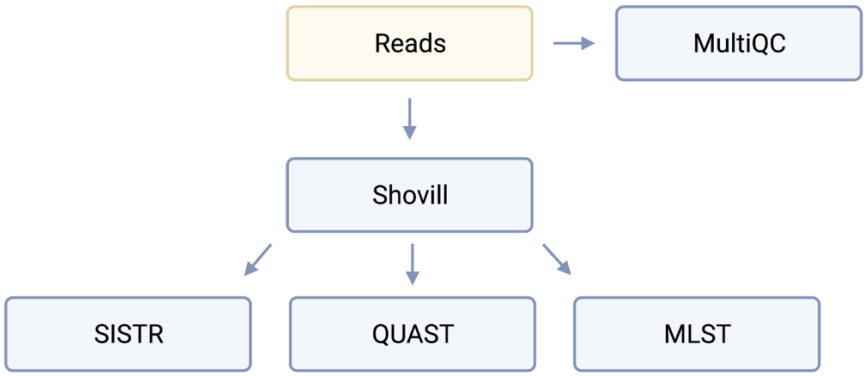
IRIDA er en bioinformatisk og analytisk web-plattform utviklet for utbruddsundersøkelser av infeksjonssykdommer ved bruk av WGS data (Matthews et al., 2018). Plattformen ble brukt til opplasting, lagring og administrering av *S.* Typhimurium isolatene. For analysene benyttet IRIDA seg av Galaxy.

Isolatene lagret i IRIDA ble delt med Galaxy-delen. Galaxy er en plattform hvor en tar i bruk bioinformatiske verktøy for å analysere filene fra sekvenseringen lagret i for eksempel IRIDA. Plattformen brukes til å designe egne «workflows» som ble brukt under preprosessering, assemblering og typing av sekvensdataene.

## 1.6 Galaxy workflow 1: Preprossesering, Assemblering og Typing

En forenklet utgave av workflowen som ble brukt for preprosessering, assemblering og typing er illustrert i figur 8. For en mer detaljert workflow designet i Galaxy med oversikt over inputs og outputs se vedlegg figur 19.

FastQC ble brukt til å utføre kvalitetskontroll på råsekvensdataene der resultatene ved hjelp av MultiQC ble samlet i en oversiktlig rapport. Deretter ble readsene satt sammen ved assemblering med bruk av Shovill som bruker Trimmomatic for å fjerne adaptere på paired end Illumina fastq reads og SPAdes som de novo assembler. Kvaliteten på de assemblerte genomene ble vurdert ved bruk av QUAST. Deretter ble SISTR brukt for serovarprediksjon og MLST for tildeling av sekvenstype. Alle verktøyene ble brukt med standard parametere, der det ikke er spesifisert noe annet.




*Figur 8 Forenklet versjon av workflow designet i Galaxy ved preprosessering, assemblering og typing av S. Typhimurium isolatene. Figuren er laget ved bruk av biorender.*

### 1.6.1 Kvalitetssjekk med MultiQC

For å sjekke kvaliteten på råsekvensdataen ble verktøyet MultiQC (Version 1.11) brukt. Dette gir en samlet rapport over FastQC profilene til isolatene. FastQC bruker rådata fra R1 og R2 FASTQ-filer per isolat som input. Sammendragsgrafer og tabeller fra alle isolatene ble ved hjelp av MultiQC brukt for å samle resultatene i en HTML-fil. HTML-filen inneholdt følgende resultater: unike reads, dupliserte reads, sekvens reads lengde og GC% innhold.

Basert på data fra MultiQC ble Qual bases (%) og coverage regnet. Qual bases (%), coverage og GC (%) ble brukt til å vurdere kvaliteten til rådataen. Coverage ble regnet ut ved bruk av formel (2), der den totale genomstørrelse på referansegenomet *Salmonella enterica* subsp. *enterica* serovar Typhimurium str. LT2 (4857450 bp) ble brukt for alle isolatene.

Qual bases (%) ble regnet ut ved å bruke formel 2.

$$Qual\ bases\ (\%) = \frac{Unique\ reads}{Duplicate\ reads} * 100 \qquad (2)$$

### 1.6.2 Assemblering og preprosessering med Shovill

Shovill (Version 1.1.0) ble brukt for assemblering av *Salmonella* sekvensdataene fra Illumina paired-end reads. Shovill er en pipeline som bruker et assembleringsprogram med Trimmomatic kombinert. Trimmomatic ble brukt for å fjerne adaptere og reads av dårlig kvalitet før assembleringen. SPAdes (St. Petersburg genome assembler) ble valgt som de novo assembler.

### 1.6.3 Kvalitetssjekk av assembly med QUAST

QUAST (Version 5.0.2) ble brukt til å vurdere kvaliteten etter assembleringen. Contigs-filen fra assembleringen ble brukt som input. Referansegenomet brukt var *Salmonella enterica* subsp. *enterica* serovar Typhimurium str. LT2, complete genome (NC_003197.2) hentet fra NCBI. QUAST produserte en rapport der N50, antall contigs og lengde på assembly ble plukket ut for å vurder kvaliteten.

### 1.6.4 SISTR

SISTR (Version 1.1.1) brukte contigs fra assembleringen med Shovill som input. Serovarianter ble identifisert fra helgenomsekvens assembly ved bestemmelse av antigen og cgMLST-genalleler ved bruk av BLAST. SISTR brukte i tillegg Mash for en rask estimering av genom- og metagenomavstanden. SISTR produserte en output fil med serovar-prediksjonen og resultater for in silico typingen. Filen inneholdt serovar-prediksjon basert på antigen, cgMLST- og Mash-resultater.

### 1.6.5 MLST

MLST (Version 2.16.1) brukte contigs fra assembleringen med Shovill som input. Inputfilen ble skannet mot PubMLST typing-skjemaer, som for *Salmonella* baserer seg på gensekvensen til de syv genene *aro*C, *dna*N, *hem*D, *his*D, *pur*E, *suc*A og *thr*A. En tabulatordelt datafil ble produsert med ST (sekvenstype) og allele-ID-ene. Datafilen ble lastet opp og visualisert ved hjelp av trevisualiseringsprogrammet Grapetree (Version 0.1.8) (Zhou et al., 2018).

## 1.7 Galaxy workflow 2: SNP analyse

For SNP analysen ble verktøyene illustrert i figur 9 brukt. Kjerne-genom ble identifisert ved hjelp av ParSNP, hvorpå Harvesttools ble brukt for å formatere outputet til fasta. Rekombinante områder ble identifisert med Gubbins. Deretter ble et Maximum Likelihood-tre generert med IQTREE som visualisert i iTOL. Snp-dists produserte en avstandsmatrise og ble

brukt for å beregne antall SNP-er mellom alle sekvenser i FASTA-filen fra Gubbins multippel sammenstillingen. I beskrivelsene av verktøyene under er kun outputet brukt i analysen beskrevet. Workflow designet i Galaxy med oversikt over alle inputs og outputs fra analysen er lagt ved som vedlegg i figur. 20.


*Figur 9 Forenklet versjon av workflow designet i Galaxy for SNP analyse av S. Typhimurium isolatene. Figuren er laget ved bruk av biorender.*

### 1.7.1 ParSNP og Harvesttools

ParSNP (Version 1.2) brukte ferdig assemblert ST19 og ST34 separat som inputs. For ST19 SNP analyse ble *Salmonella enterica* subsp. *enterica* serovar Typhimurium str. LT2, complete genome (NC_003197.2) brukt som referansegenom. For ST34 SNP analysen ble *Salmonella enterica* subsp. *enterica* serovar Typhimurium WW01 chromosome, complete genome (CP022168.1) brukt som referansegenom. ParSNP utførte en kjernegenom alignment som produserte en "Gingr formatted binary archive" fil sammen med flere outputs som ikke ble brukt i analysen. Gingr filen ble konvertert til en fasta-fil med Harvesttools (Version 1.2.0) som ble brukt som input i Gubbins.

### 1.7.2 Gubbins

Gubbins (Version 0.1.0) brukte «multi-FASTA whole genome alignment» outputet fra Harvesttools til å se etter rekombinasjon i genomet. Ved en homolog rekombinasjon bytter en *Salmonella*-bakterie ut biter av genomet sitt med omtrent samme region i en annen *Salmonella*-bakterie. Siden disse endringene ikke er forårsaket av vertikal arv brukes Gubbins til å detektere rekombinasjon i genomet så ikke regionene brukes for å finne SNP-er (Croucher et al., 2014). Gubbins produserte flere outputs, inkludert "FASTA format alignment of filtered polymorphic sites", som ble brukt som input for IQTREE og Snp-dists

### 1.7.3 Snp-dists

Snp-dists (Version 0.8.2) brukte "FASTA format alignment of filtered polymorphic sites" produsert fra Gubbins som input. Verktøyet beregner SNP-avstanden mellom hvert

sekvenspar i en multippel sekvenssammenstilling (av like lange sekvenser) og outputer de parvise SNP-avstandene som matrise (Seemann, 2019), med en rad/kolonne for hvert genom. Matrisen ble brukt for å se hvor mange SNP-er som skilte to genomer.

### 1.7.4 IQTREE

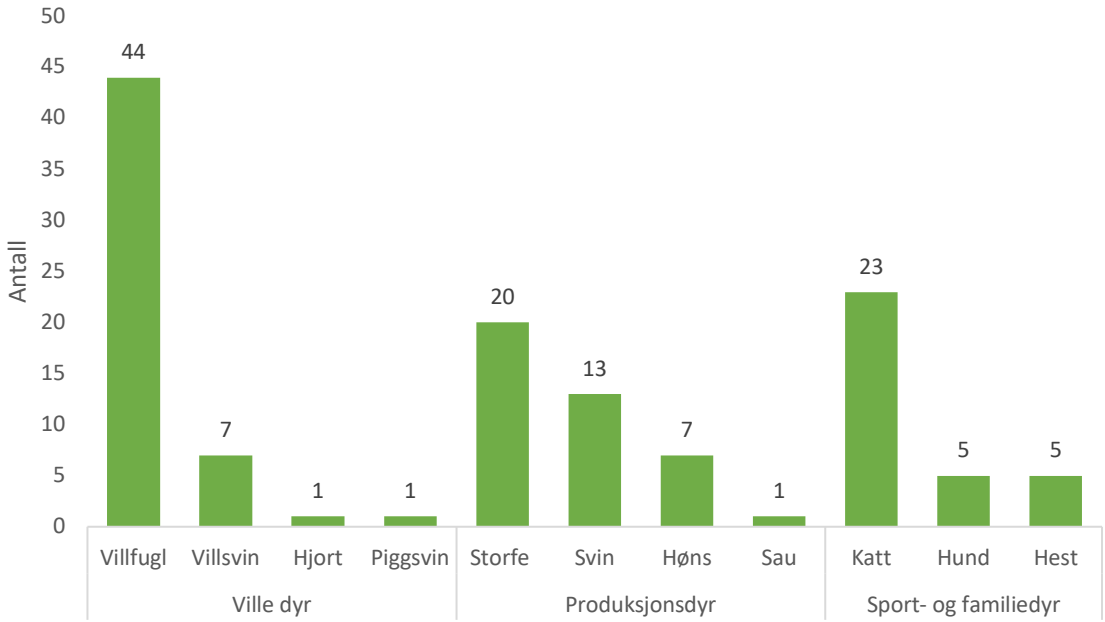
IQTree (Version 1.5.5.3) brukte "FASTA format alignment of filtered polymorphic sites" produsert fra Gubbins som input og rekonstruerer et evolusjonstre. Etter kjøringen produserte IQ-TREE flere outputs, inkludert «Maximum Likelihood Distance Matrix». iTOL (Interactive Tree Of Life) v5 ble brukt som programverktøy for å visualisere og manipulere et fylogenetiske trær basert på SNP forskjellene (Letunic & Bork, 2021) fra Maximum Likelihood distansematrisen.

# Resultater

## 1.8 Datasett

Datasettet i oppgaven bestod av *S.* Typhimurium isolater valgt ut fra konfidensielle dokumenter tilgjengelig hos Veterinærinstituttet. Utgangspunktet var *Salmonella* spp. isolater som allerede var helgenomsekvensert. For å inkluderes i denne oppgaven ble det stilt krav om at isolatene var fra ville dyr, produksjonsdyr og sport- og familiedyr i Norge, og at serovarianten var *S.* Typhimurium basert på serotyping gjort på bakteriologi laboratoriet ved Veterinærinstituttet. Kravene ble brukt til å plukke ut isolater fra prøvejournalsystemet (PJS) til Veterinærinstituttet. Isolater som var eldre enn 2006 ble slått opp i et dokument over historiske *Salmonella* isolater fra 1951 til 2006 for å finne tilstrekkelig med informasjon for å vurdere om isolatene oppfylte kravene for å bli inkludert i datasettet. Isolatene som ikke oppfylte disse kravene, samt isolater fra ringtester, ble ikke inkludert.

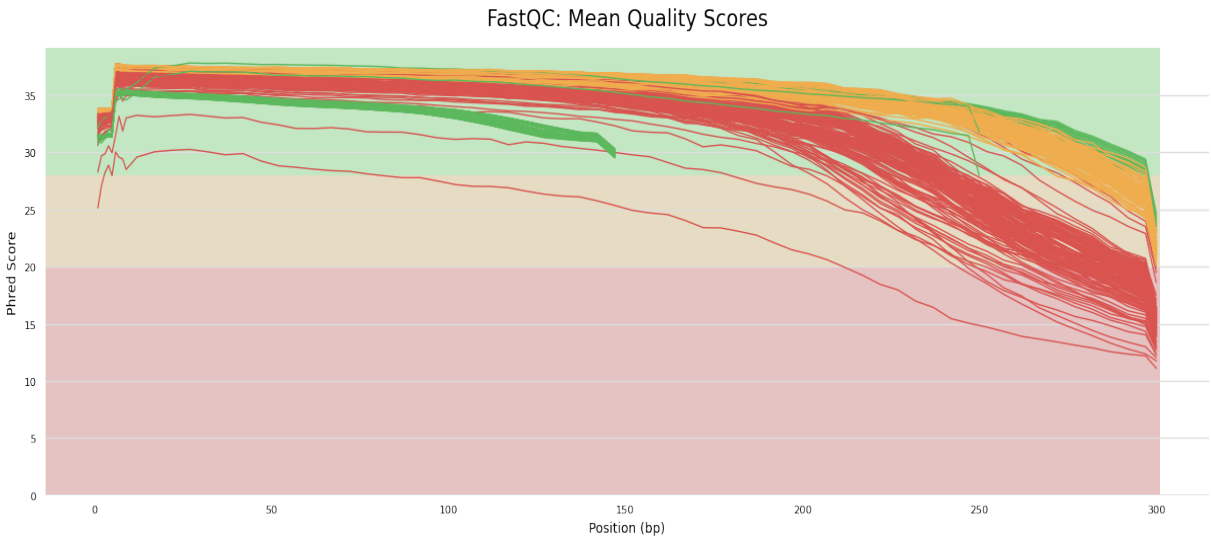
Sekvensdataene i datasettet bestod av 127 isolater som var samlet i paired end R1 og R2 fastq-filer. 109 av isolatene var sekvensert med Illumina NextSeq og hadde en sekvenslengde på 300 bp, mens 18 isolater var sekvensert med Illumina MiSeq og hadde en sekvenslengde på 150 bp. Isolatene i datasettet ble samlet inn mellom 2001 og 2021 fra ville dyr, produksjonsdyr og sport- og familiedyr i Norge. Blant de 127 *S.* Typhimurium isolatene var 53 isolert fra ville dyr, 41 isolert fra produksjonsdyr og 33 isolert fra sport- og familiedyr. Arter og tilhørende antall isolater er presentert i figur 10. Isolater fra 2020 utgjorde store deler av datasettet, der flere av isolatene var isolert fra katt. De resterende isolatene var jevnt fordelt over årene fra 2001 til 2021. Tabell 7 i vedlegg inneholder informasjon om alle isolatene inkludert i oppgaven, samt tilhørende art og et PJS nummer som inkluderer året prøvene ble samlet inn og bakteriene isolert.




*Figur 10 Oversikt over antall ville dyr, produksjonsdyr og sport- og familiedyr med artsopprinnelse for de 127 isolatene brukt i oppgaven.*

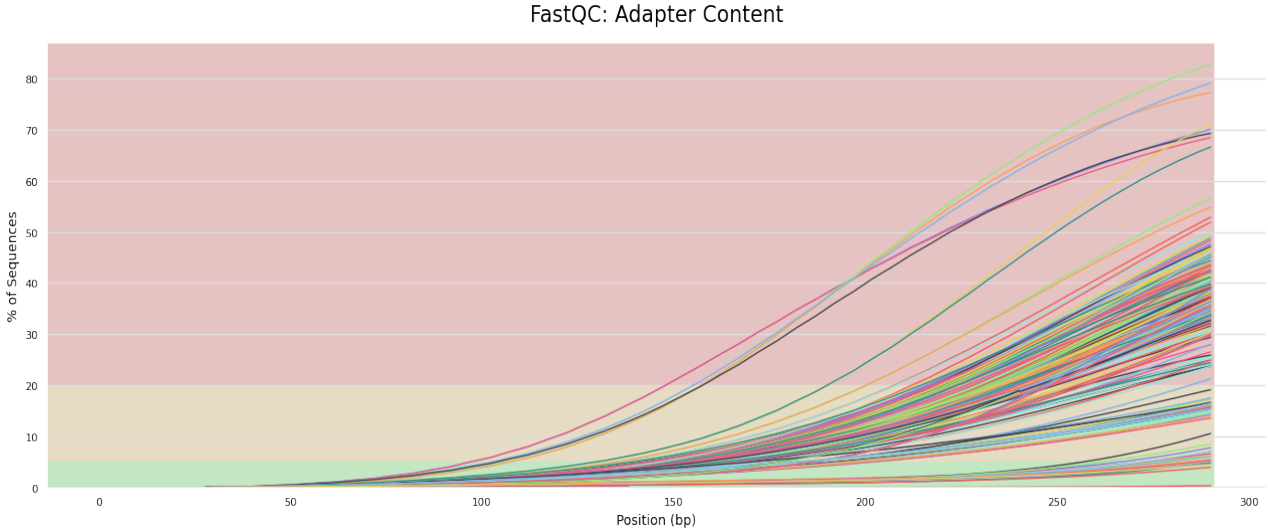
## 1.9 Kvalitetskontroll av rådata og assembly

Rådatafilene fra sekvenseringen ble kvalitetsjekket ved hjelp av verktøyet MultiQC, og kvaliteten på assembly ble vurdert ved hjelp av QUAST. Kvalitetsscore for rådataen plottet i en Score Quality plot fra MultiQC-rapporten er illustrert i figur 11, der y-aksen viser kvalitetscoren (PHRED score) og x-aksen viser posisjonen i reads. For å gjøre det mulig å plotte flere prøver i samme graf viser plottet gjennomsnittlig kvalitetsscore. Plottet inkluderer reads med ulik lengde basert på sekvenseringsteknologien og flowcelle som ble brukt. De korte grønne linjene (R1 og R2) i figur 11 stopper ved 150 bp og er sekvensert med Illumina NextSeq. Isolater sekvensert med Illumina MiSeq viser reads med lengde på 300 bp. For reads med lengde på 300 bp viser grønne og oransje linjer reads som ble sekvensert i første runde (R1), og røde linjer viser, med unntak, reads sekvensert i andre runde (R2).




*Figur 11 Score Quality plot fra MultiQC rapporten. Figuren viser kvalitetsverdier for R1 og R2 rådata fra Illumina Miseq (300 bp) og Nextseq (150 bp). Posisjonen i reads er plottet langs x-aksen og kvaliteten målt i PHRED score langs y-aksen. Y-aksen er delt inn i tre deler etter svært god kvalitet (grønn), rimelig god kvalitet (oransje) og dårlig kvalitet (rød).*

Adapterinnholdet i readsene (R1 og R2) fra rådataen er plottet i figur 12, der x-aksen viser posisjonen (bp) og y-aksen viser mengde adaptersekvenser i %. Figuren viser kun isolater med ≥ 0,1% adapterkontaminering. Hver farget linje representerer en adapter som er oppdaget i isolatet, og kan derfor inneholde flere linjer per isolat. For y-aksen indikerer en høy kurve en større andel adaptere, mens x-aksen viser hvor tidlig adapterinnholdet oppdages i readsen.




*Figur 12 Adapterinnholdet i reads fra assembleringen. X-aksen viser posisjonen (bp) og y-aksen viser mengde adaptersekvenser i %. Hver farget linje representerer en adapter som er oppdaget i isolatet. Y-aksen er delt inn i tre deler etter svært god kvalitet (grønn), rimelig god kvalitet (oransje) og dårlig kvalitet (rød).*

Gjennomsnittlig kvalitetsverdier for rådata fra MultiQC og assembly fra QUAST er presentert i tabell 1. Kvaliteten ble sjekket for alle de 127 isolatene i datasettet. Dette inkluderte 109 isolater med en sekvenslengde på 300 bp sekvensert med Illumina NextSeq og 18 isolater med sekvenslengde på 150 bp sekvensert med Illumina MiSeq. For MultiQC var det ønsket et GC-innhold på rundt 52%, som er gjennomsnittlig verdi for *S.* Typhimurium (Papanikolaou et al., 2009). For QUAST ble det satt et krav på at N50 skulle være (>15000) og antall contigs (<100). Basert på de nevnte grensene var det ingen av de 18 isolatene som ble sekvensert med Illumina NextSeq (150 bp) som oppfylte kravene, og ble dermed fjernet fra datasettet. For isolatene sekvensert med Illumina MiSeq (300 bp) var det 5 isolater som ikke oppfylte kravene og ble fjernet fra datasettet. 104 isolater sekvensert med Illumina Miseq (300 bp) viste gode resultater i kvalitetskontrollen og kunne brukes i ytterligere analyser (tabell 1). Tabell 8 i vedlegg presenterer en samlet rapport med kvalitetsverdier for alle isolatene fra MultiQC og QUAST der isolater fjernet fra datasettet er markert.

*Tabell 1 Gjennomsnittsverdier av resultatene fra kvalitetskontroll av rådata og assembly ved bruk av hhv. MultiQC og QUAST. Sekvenslengde (bp), qual bases (%), GC (%) og coverage er hentet fra MultiQC rapporten og N50 og antall contigs er hentet fra QUAST rapporten.*

| | MultiQC | | | | QUAST | |
|---|---|---|---|---|---|---|
| Antall isolater | Sekvenslengde (bp) | Qual bases % | GC % | Coverage | N50 | Antall contigs |
| 18 | 149 | 64 | 48 | 70 | 6603 | 1478 |
| 5 | 300 | 84 | 51 | 69 | 22022 | 10977 |
| 104 | 300 | 84 | 51 | 67 | 326988 | 50 |

** Kvalitetsverdier som ikke oppfylte kravene for god nok kvalitet er merket med rødt, og tilhørende isolater ble fjernet fra datasettet (totalt 23).*

### 1.10 Serovarprediksjon

SISTR ble brukt for å sjekke at de gjenværende 104 isolatene i datasettet var serovariant *S.* Typhimurium. Alle isolatene, med ett unntak, ble karakterisert som subspecies *enterica*. Isolatet 2018-01-3742-8 manglet alle cgMLST330 loci. En nøyaktig serovariant fra antigenene kunne derfor ikke bestemmes. For de resterende isolatene ble O-antigen faktor (1,4, [5] og 12) og H1 antigen (i) karakterisert. For H2 antigen ble (1,2) karakterisert for 102 isolater, mens for isolatene 2018-01-1983-10 og 2021-01-1675-2 ble ingen H2 antigen funnet.

SISTR kombinerte resultater fra karakterisering basert på serovar_antigen, serovar_cgMLST og serovar_Mash for å gi et endelig resultat (tabell 2). Det endelige resultater viste serovariant *S.* Typhimurium for alle 104 isolater.

*Tabell 2 Salmonella serovariant, samt O-antigen og H-antigen fra SISTR analysen for totalt 104 isolater. For 2 av isolatene ble ingen H2 antigen funnet.*

| Antall isolater | O - gruppe | Serovariant | O-antigen | H1 antigen | H2 antigen |
|---|---|---|---|---|---|
| 102 | *enterica* | Typhimurium | 1,4,5,12 | i | 1,2 |
| 2 | *enterica* | Typhimurium | 1,4,5,12 | i | - |

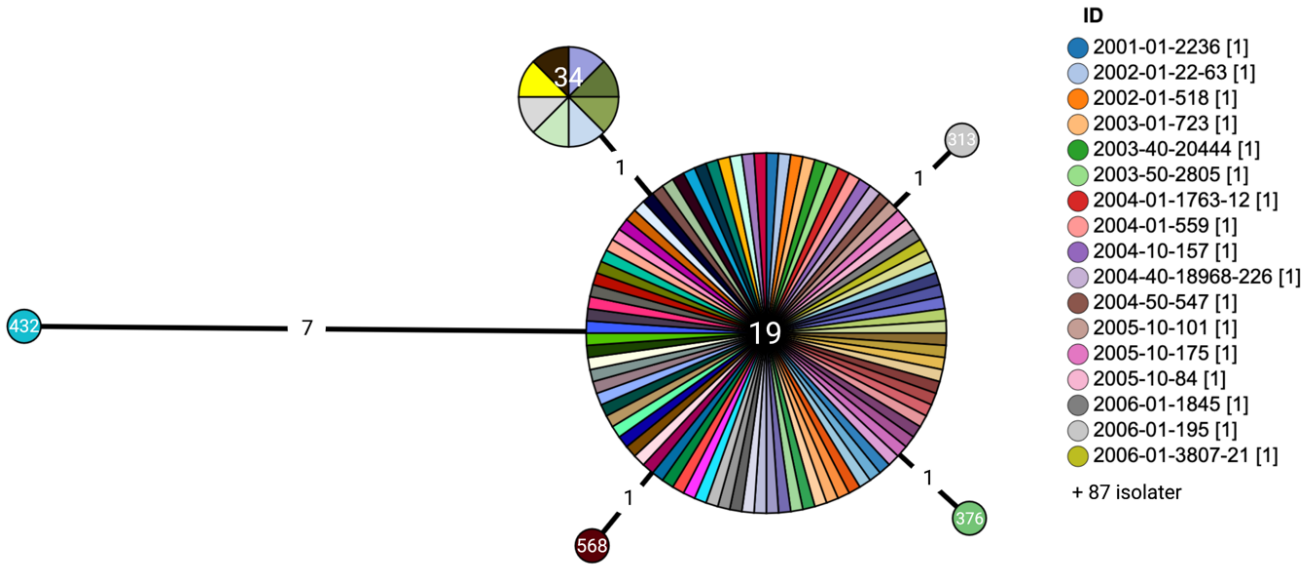
### 1.11 *S.* Typhimurium MLST sekvenstype

MLST tildelte de 104 isolatene fra *S.* Typhimurium til 6 sekvenstyper (ST). 92 isolater tilhørte ST19, 8 isolater tilhørte ST34 og ett isolat tilhørte hhv. ST313, ST376, ST432 og ST 568. De 7 husholdningsgenene *aro*C, *dna*N, *hem*D, *his*D, *pur*E, *suc*A, *thr*A funnet for sekvenstypene med tilhørende allelenummer, samt antall isolater knyttet til sekvenstypen er presentert i tabell 3. I tabell 3 er allelene som skiller seg ut fra ST19 profilen merket rødt. ST34, ST313, ST376 og ST568 har en avstand på 1, som betyr at det er kun en allele som skiller seg fra ST19. ST432 skiller seg ut med en avstand på 7 alleler, som er merket med grått i tabellen.

*Tabell 3 S.Typhimurium stammene sortert etter sekvenstype, samt den tilknyttende allele-ID-ene. Alleler som skiller seg ut fra ST19 profilen for ST 34, ST313, ST376, ST432 og St568 er merket med rødt som gir en indikasjon på avstanden mellom STene.*

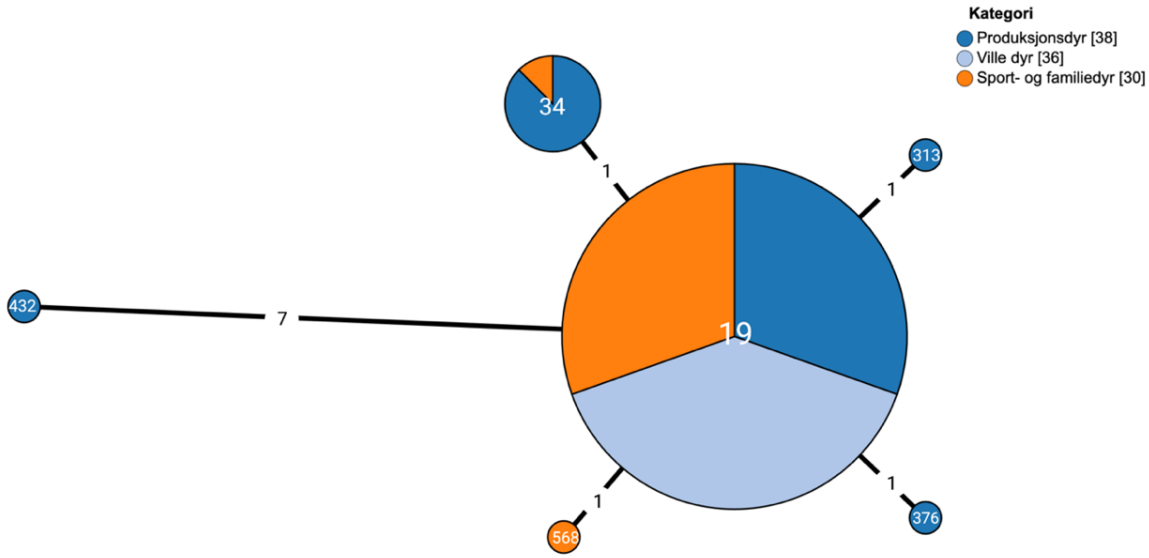
| Sekvenstype (ST) | *aro*C | *dna*N | *hem*D | *his*D | *pur*E | *suc*A | *thr*A | Antall isolater |
|---|---|---|---|---|---|---|---|---|
| 19 | 10 | 7 | 12 | 9 | 5 | 9 | 2 | 92 |
| 34 | 10 | 19 | 12 | 9 | 5 | 9 | 2 | 8 |
| 313 | 10 | 7 | 12 | 9 | 112 | 9 | 2 | 1 |
| 376 | 10 | 7 | 12 | 9 | 5 | 121 | 2 | 1 |
| 568 | 183 | 7 | 12 | 9 | 5 | 9 | 2 | 1 |
| 432 | 88 | 26 | 30 | 55 | 21 | 68 | 80 | 1 |

Trevisualiseringsprogrammet Grapetree visualiserte et tre basert på klustering på basis av allele-tall fra MLST analysen er illustrert i figur 13. Treet viser totalt 104 isolater, representert med hver sin farge, fordelt på 6 klynger basert på sekvenstype.




*Figur 13 Grapetree basert på klustering på basis av allele-tall for S.Typhimurium isolatene fra MLST analysen. Treet viser totalt 104 isolater fordelt på 6 klynger basert på ST.*

Grapetree basert på klustering på basis av allele-tall fra MLST analysen fordelt på produksjonsdyr, ville dyr og sport- og familiedyr er illustrert i figur 14. Treet viser totalt 104 isolater fordelt på sekvenstype og type dyr.




*Figur 14 Grapetree basert på klustering på basis av allele-tall for S.Typhimurium isolatene fra MLST analysen. Treet viser totalt 104 isolater fordelt på 6 klynger basert på ST. Andelen av isolatene som er produksjonsdyr er merket med mørkeblå, ville dyr er merket med lys blå og sport- og familiedyr merket med oransje.*

### 1.12 Analyse av SNP forskjeller

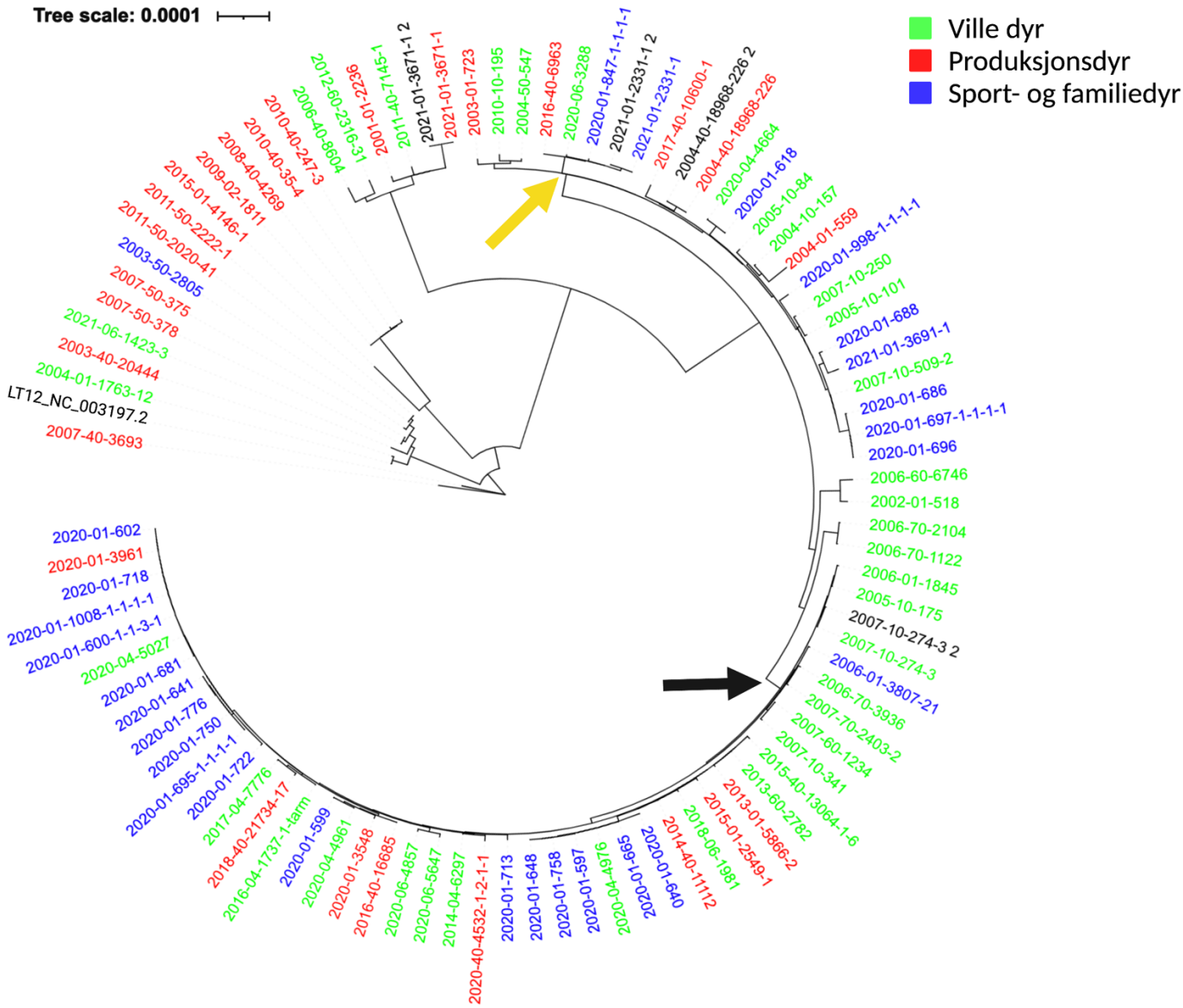
ST19 og ST34 som inneholdt flest isolater ble analysert separat basert på hele genomet. Resultatene fra SNP analysen ble illustrert i et fylogenetisk tre generert ved bruk av iTOL. Under SNP analysen ble noen av isolatene ikke inkluder grunnet tekniske årsaker, og disse er dermed ikke inkludert i de fylogenetiske trærne. Dette gjelder fire isolater for ST19: 2002-01-22-63, 2006-40-8604, 2012-60-2316-31, 2018-40-12050-5 og ett isolat for ST34: For ST34: 2007-01-3478.

#### 1.12.1 ST19

ST19-gruppen omfattet 92 isolater fra *S.* Typhimurium hos villedyr, produksjonsdyr og sport- og familiedyr i Norge. Av disse ble 88 av isolatene fra SNP analysen illustrert i et fylogenetisk tre ved bruk av iTOL v6 (figur 15). Innenfor ST19-gruppen ble det oppdaget SNP-er med avstander mellom isolatene som strekker seg fra 0 til 1044 SNP-er, med median på 176 SNP-er og gjennomsnittlig 378 SNP-er. SNP distanse matrisen for hele ST19 var for stor til å inkluderes som vedlegg i rapporten, og kan gis ved forespørsel.

I figur 15 er de to største klusterne merket med en svart pil (kluster 1) og en gul pil (kluster 2). Isolater som ikke ble definert som villedyr, produksjonsdyr og sport- og familiedyr i figur 15, og derfor merket svart, ble dupliserte under SNP analysen. «LT12_ NC_003197.2» er referansegenomet *Salmonella enterica* subsp. *enterica* serovar Typhimurium str. LT2, complete genome (NC_003197.2).

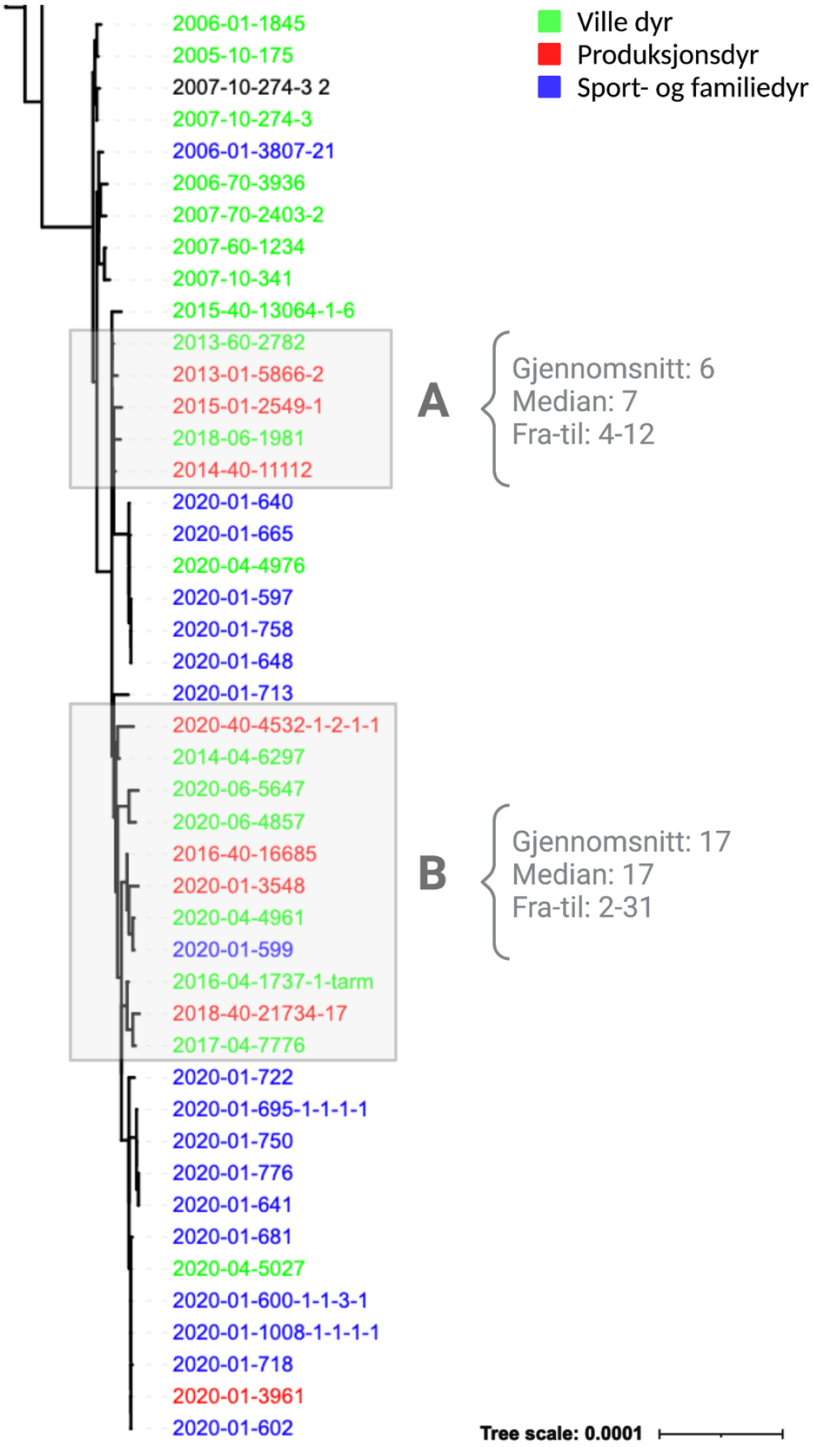
| Gjennomsnitt | Median | Fra-til |
|---|---|---|
| 378 | 176 | 0-1044 |




*Figur 15 Fylogenetisk tre basert på SNP analysen for S.Typhimurium ST19. Treet viser SNP forskjellen mellom S. Typhimurium isolatene der ville dyr er merket grønt, produksjonsdyr rødt og sport- og familiedyr blått. Svart pil peker på kluster 1 og gul pil på kluster 2.*

Figur 16 viser en forstørret figur av kluster 1 som består av 44 isolater fra 18 ville dyr, 8 produksjonsdyr og 18 sport- og familiedyr. Isolatene i klusteret har avstander som strekker seg fra 0 til 43 SNP forskjeller, med en median på 25 og et gjennomsnitt på 24 SNP Avstanden mellom isolatene innad i klusteret og referansen brukt i analysen var ca. 800-900 SNP-er. SNP distanse matrise for ST19 kluster 1 er presentert i vedlegg tabell 9.

| Gjennomsnitt | Median | Fra-til |
|---|---|---|
| 24 | 25 | 0-43 |




*Figur 16 Kluster 1 forstørret fra figur 14. Klusteret består av 44 isolater fra hhv. ville dyr, produksjonsdyr og sport- og familiedyr. Figuren er delt inn i gruppe A og B som inneholder isolater med små SNP forskjeller med utgangspunkt i isolatene samlet fra produksjonsdyr.*

Gruppe A i figur 16 inkluderer tre isolater samlet fra produksjonsdyr og to isolater fra ville dyr. Isolatene i Gruppe A strekker seg fra 4-12 SNP forskjeller, med en median på 7 og et gjennomsnitt på 6 SNP. SNP distanse matrisen for Gruppe A er presentert i tabell 4. Isolater fra produksjonsdyr er merket rødt og isolater fra ville dyr er merket grønt.

*Tabell 4 SNP distanse matrise for gruppe A i ST19 kluster 1. Tabellen presenterer tre isolater samlet fra produksjonsdyr (merket med rødt) og to isolater samlet fra ville dyr (merket med grønt).*

| Gruppe A | 2013-01-5866-2 | 2013-60-2782 | 2014-40-11112 | 2015-01-2549-1 | 2018-06-1981 |
|---|---|---|---|---|---|
| 2013-01-5866-2 | 0 | 5 | 7 | 12 | 11 |
| 2013-60-2782 | 5 | 0 | 4 | 9 | 8 |
| 2014-40-11112 | 7 | 4 | 0 | 7 | 6 |
| 2015-01-2549-1 | 12 | 9 | 7 | 0 | 11 |
| 2018-06-1981 | 11 | 8 | 6 | 11 | 0 |

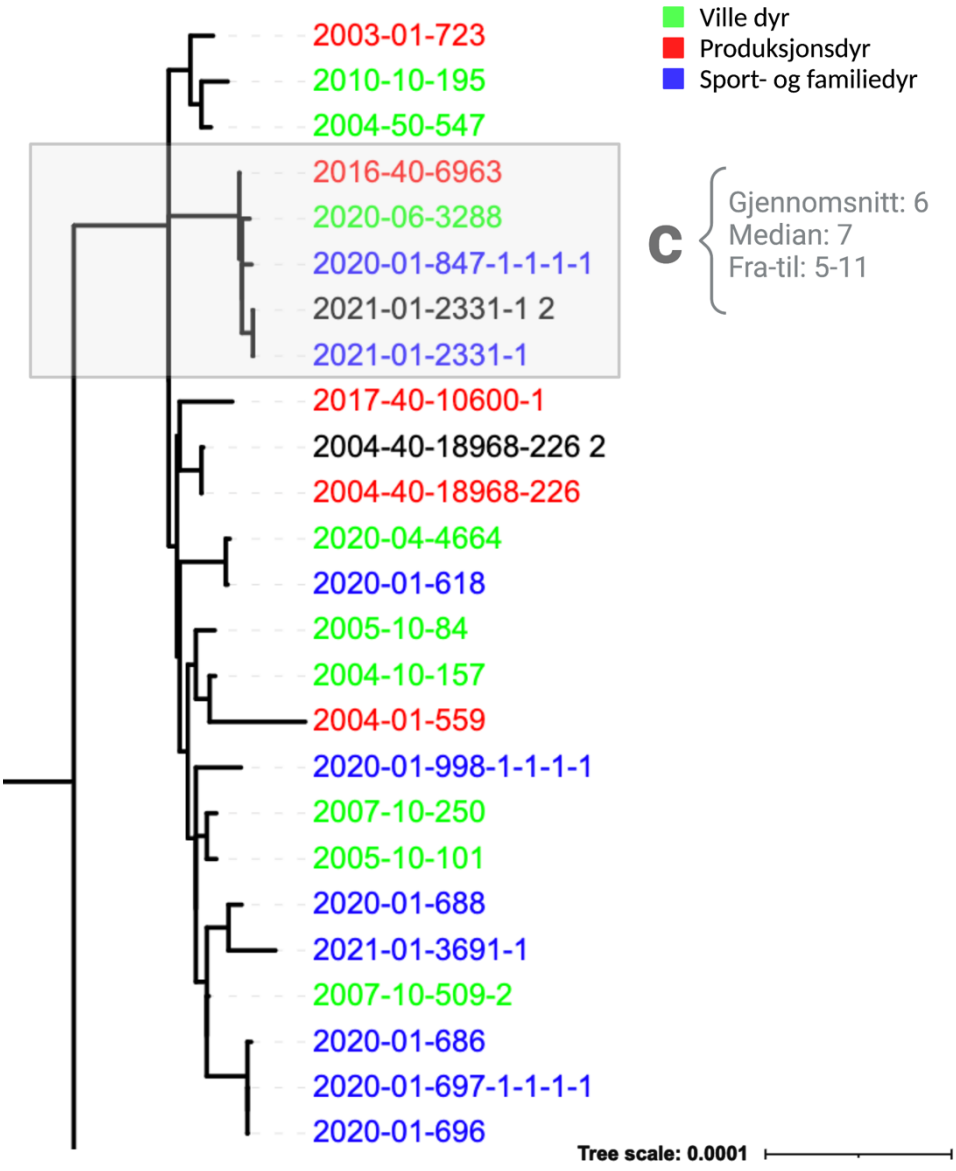
Gruppe B i figur 16 inkluderer fire isolater samlet fra produksjonsdyr, seks isolater fra ville dyr og ett isolat fra sport- og familiedyr. Isolatene i Gruppe B strekker seg fra 2-31 SNP forskjeller, med en median på 17 og et gjennomsnitt på 17 SNP. SNP distanse matrisen for Gruppe B er presentert i tabell 5. Isolater fra produksjonsdyr er merket rødt, isolater fra ville dyr er merket grønt og isolater fra sport- og familiedyr er merket blått.

*Tabell 5 SNP distanse matrise for gruppe B i ST19 kluster 1. Tabellen presenterer fire isolater samlet fra produksjonsdyr (merket med rødt), to isolater samlet fra ville dyr (merket med grønt) og ett isolat samlet fra sport- og familiedyr (merket blått)*

| Gruppe B | 2014-04-6297 | 2016-04-1737-1-tarm | 2016-40-16685 | 2017-04-7776 | 2018-40-21734-17 | 2020-01-599 | 2020-04-4961 | 2020-01-3548 | 2020-06-4857 | 2020-06-5647 | 2020-40-4532-1-2-1-1 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2016-40-16685 | 9 | 9 | 0 | 14 | 17 | 5 | 5 | 8 | 20 | 22 | 20 |
| 2014-04-6297 | 0 | 10 | 9 | 15 | 18 | 14 | 14 | 17 | 15 | 17 | 13 |
| 2016-04-1737-1-tarm | 10 | 0 | 9 | 7 | 10 | 14 | 14 | 17 | 21 | 23 | 21 |
| 2020-01-599 | 14 | 14 | 5 | 19 | 22 | 0 | 2 | 11 | 25 | 27 | 25 |
| 2020-04-4961 | 14 | 14 | 5 | 19 | 22 | 2 | 0 | 11 | 25 | 27 | 25 |
| 2017-04-7776 | 15 | 7 | 14 | 0 | 7 | 19 | 19 | 22 | 26 | 28 | 26 |
| 2020-01-3548 | 17 | 17 | 8 | 22 | 25 | 11 | 11 | 0 | 28 | 30 | 28 |
| 2018-40-21734-17 | 18 | 10 | 17 | 7 | 0 | 22 | 22 | 25 | 29 | 31 | 29 |
| 2020-06-4857 | 15 | 21 | 20 | 26 | 29 | 25 | 25 | 28 | 0 | 12 | 26 |
| 2020-40-4532-1-2-1-1 | 13 | 21 | 20 | 26 | 29 | 25 | 25 | 28 | 26 | 28 | 0 |
| 2020-06-5647 | 17 | 23 | 22 | 28 | 31 | 27 | 27 | 30 | 12 | 0 | 28 |

Figur 17 viser en forstørret figur av kluster 2. Klusteret består av 23 isolater fra 9 ville dyr, 5 produksjonsdyr og 9 sport- og familiedyr. Isolatene i klusteret har avstander som strekker seg fra 1 til 83 SNP forskjeller, med median på 51 SNP og gjennomsnittlig 50 SNP. Avstanden mellom isolatene innad i klusteret og referansen brukt i analysen var ca. 800-900 SNP-er. SNP distanse matrisen for ST19 kluster 2 er presentert i vedlegg tabell 10.

| Gjennomsnitt | Median | Fra-til |
|---|---|---|
| 50 | 51 | 1-83 |




*Figur 17 Kluster 2 forstørret fra figur 14. Klusteret består av 23 isolater fra hhv. ville dyr, produksjonsdyr og sport- og familiedyr. Figuren er delt inn i gruppe C som inneholder isolater med små SNP forskjeller med utgangspunkt i isolatet samlet fra produksjonsdyr.*

Gruppe C i figur 17 inkluderer ett isolat samlet fra produksjonsdyr, ett isolat fra ville dyr og to isolater fra sport- og familiedyr. Isolatene i Gruppe C strekker seg fra 5-11 SNP forskjeller, med en median på 7 og et gjennomsnitt på 6 SNP. SNP distanse matrisen for Gruppe C er presentert i tabell 6, der isolatet fra produksjonsdyr er merket rødt, isolatet fra ville dyr er merket grønt og isolater fra sport- og familiedyr er merket blått.

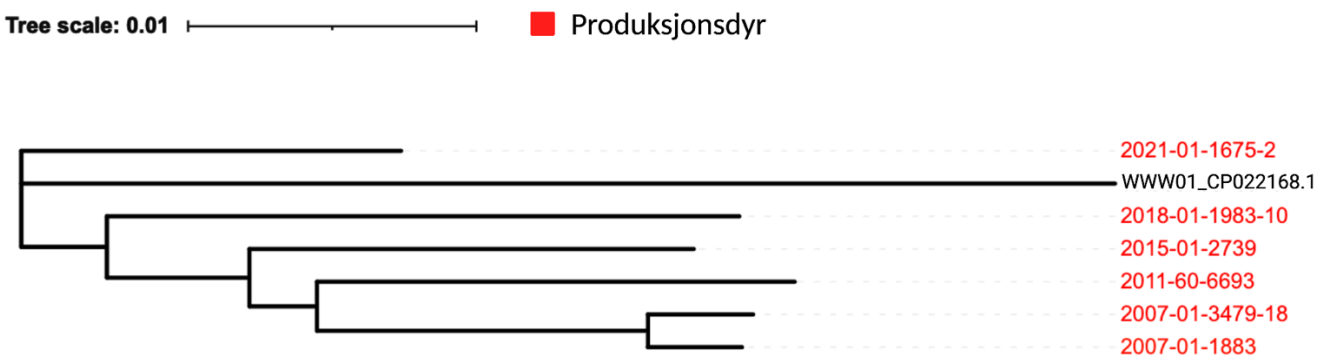
*Tabell 6 SNP distanse matrise for gruppe C i ST19 kluster 2. Tabellen presenterer ett isolat samlet fra produksjonsdyr (merket med rødt), ett isolat samlet fra ville dyr (merket med grønt) og to isolater samlet fra sport- og familiedyr (merket blått)*

| Gruppe C | 2016-40-6963 | 2020-06-3288 | 2021-01-2331-1 | 2020-01-847-1-1-1-1 |
|---|---|---|---|---|
| 2021-01-2331-1 | 7 | 10 | 0 | 11 |
| 2020-06-3288 | 5 | 0 | 10 | 9 |
| 2020-01-847-1-1-1-1 | 6 | 9 | 11 | 0 |
| 2016-40-6963 | 0 | 5 | 7 | 6 |

### 1.12.2 ST34

ST34-gruppen omfattet åtte isolater fra *S.* Typhimurium hos villedyr, produksjonsdyr og sport- og familiedyr i Norge. Av disse ble seks av isolatene fra SNP analysen illustrert i et fylogenetisk tre ved bruk av iTOL v6 (figur 18). Innenfor ST34 gruppen ble det oppdaget SNP-er med avstander mellom isolatene som strekker seg fra 12 til 103 SNP-er, med median på 64 SNP-er og gjennomsnittlig 60 SNP-er. «WW01_ CP022168.1» er referansegenomet *Salmonella enterica* subsp. *enterica* serovar Typhimurium WW01 chromosome, complete genome (CP022168.1). Avstanden mellom isolatene innad i klusteret og referansen brukt i analysen var ca. 100 SNP-er. SNP distanse matrisen for ST34 er presentert i tabell 11 i vedlegget.

| Gjennomsnitt | Median | Fra-til |
|---|---|---|
| 60 | 64 | 12-103 |




*Figur 18 Fylogenetisk tre basert på SNP analysen for S.Typhimurium ST34. Treet viser SNP forskjellen mellom S. Typhimurium isolatene der produksjonsdyr er merket med rødt.*

# Diskusjon

Prossesen fra reads til fylogenetiske sammenstillinger inneholder flere steg med flere valgmuligheter. I alle steg skal bioinformatiske verktøy velges og tilhørende parametere settes. Hva som velges kan være med på å påvirke det endelige resultatet.

### 1.13 Datasett

Datasettet bestod opprinnelig av 127 *S.* Typhimurium isolater jevnt fordelt fra villedyr, produksjonsdyr og sport- og familiedyr (figur 10). Flere av isolatene i datasettet ble samlet fra 2020 som inkluderte mange ville dyr isolater fra katt. De resterende isolatene var jevnt fordelt på årene mellom 2001 og 2021 (vedlegg tabell 7). Definisjonen av et ideelt datasett avhenger av hensikten i en studie. Ved analyser av genomer forbundet med smitteoverføring er det viktig at datasettet er mangfoldig og med god kvalitet. Flere av isolatene samlet fra produksjonsdyr var ikke like nok isolatene samlet fra villedyr, som gjorde det vanskelige å vurdere i hvor stor grad ville dyr smitter produksjonsdyr.

Gjennom bioteknologiske metoder som kvalitetssjekk av reads og assembly er det ikke usannsynlig at isolater ekskluderes fra datasettet. Å fjerne isolater vil dermed ha en større påvirkning for et mindre datasett. På den andre siden er et lite datasett lettere håndterbart som gjør analysene mindre tidkrevende.

Isolatene i datasettet ble sekvensert med NGS av hele genomet som gir en dypere kunnskap om bakterien. NGS går gjennom flere steg som kan ha innvirkning på kvaliteten til read sekvensene. Det er blant annet avgjørende med en god forberedelse for en vellykket sekvensering. WGS gir høy oppløselighet som gjør identifisering av sammenhenger mellom *S.* Typhimurium lettere, og er essensielt for en bedre overvåking under identifisering av sannsynlig smittekilde. Tidligere ble det brukt metoder med lavere oppløselighet (blant annet serotyping). Dersom isolatene som ble karakterisert som *S.* Typhimurium hadde en epidemiologisk link (f.eks dyr fra samme besetning), kunne man anta at de var en del av ett utbrudd/hadde samme opprinnelse.

## 1.14 Bioinformatiske analyser i Galaxy

Etter sekvenseringen ble plattformen Galaxy brukt for alle analysemetodene utført i oppgaven. Workflowene brukt i analysestegene ble designet i Galaxy og tilpasset for analyser av *S.* Typhimurium. Det brukes ofte ikke standardiserte workflows ved analyse av patogene bakterier. Dette fordi patogene bakterier omfatter mange forskjellige arter med ulike behov, der det avhenger fra art til art hva som trengs/ønskes. Ikke alle arter har like utfyllende databaser som *S.* Typhimurium, som kan gjøre noen analyser vanskeligere. Workflowene utvikles dermed i henhold til nødvendigheter, men også ressurser som er tilgjengelige.

I likhet med Galaxy bruker kommandolinjen bioinformatiske analyser for å analysere filer med reads fra sekvenseringen. Bruk av kommandolinjen krever betydelig mer kompetanse for å bygge workflows sammenlignet med Galaxy. En fordel med kommandolinjen er at en har tilgang til alle opsjoner som et program må ha, mens for Galaxy har en kun tilgang til opsjonene som er definert tilgjengelig for deg av noen andre. Dette førte til at for noen av analysene ble det produsert outputs med format som ikke kunne brukes som input i neste analyse, som førte til at workflowen ikke fungerte. Flere av analysene i Galaxy måtte derfor utføres separat som var mer tidkrevende. I tillegg ble det erfart på grunn av systemet som Galaxy var satt opp i fungerte analysene dårlig for datasettet i oppgaven. Dette førte til en del feilmeldinger under analysene siden Galaxy ikke håndterte antallet isolater som ble kjørt samtidig. Isolatene måtte derfor deles opp i kolleksjoner som gjorde det uoversiktlig og dermed vanskelig å finne og sortere resultatene. Hvor mange isolater som kunne inkluderes i en kolleksjon var avhengig av underliggende datakapasitet og om det interne systemet var i bruk av noen andre.

## 1.15 Kvalitetskontroll av reads fra sekvenseringen

Kvaliteten på readsene fra sekvenseringen ble vurdert i en samlet MultiQC rapport med statistiske kvalitetsverdier på tvers av isolatene. Det er viktig å luke ut reads som viser mangler før assembleringen da de kan forstyrre og føre til at sammensetning av det komplette genomet blir av lavere kvalitet. På den andre siden er det viktig å få med så mye informasjon fra readsene som mulig for å rekonstruere genomet. Det er derfor viktig å vurdere kvaliteten nøye slik at ikke reads med viktig informasjon forkastes som fører til en ufullstendig assemblering.  Dette kan føre til at contigs blir delt opp slik at de ikke kan brukes i senere fylogenetiske analyser.

MultiQC rapporten inneholdt kvalitetskontroll av isolater med readlengde på 150 bp og 300 bp. Det var en tydelig sammenheng mellom redusert kvalitet basert på Phred score (figur 11) og økende adapter innhold (figur 12) mot slutten av avlesningen. Det var forventet med dårligere kvalitet for de siste baseparene. Dette fordi den stegvise syntetiseringen gradvis blir dårligere og adaptere forurenser dataen. Kvalitetsverdiene for R2 var derfor dårligere sammenlignet med R1.

Reads med 150 bp lengde hadde et GC-innhold på 48% og kun 64% ble betegnet som kvalifiserte baser som skyldtes dupliserte sekvenser. Disse verdiene skilte seg ut fra resten av readsene med 300 bp lengde som hadde gjennomsnittlig et GC-innhold på 51% og 84% kvalifiserte baser. Dette skyldtes antagelig at det var kortere reads som alle inneholdt adaptere, som ga en mindre prosentandel kvalifiserte baser. For *Salmonella* genomet var det ønskelig med et GC-innhold på rundt 52% som er gjennomsnittlig for *S.* Typhimurium, som gjorde at GC innholdet til isolatene på 150 bp var lavere enn ønsket.

Trimmomatic ble brukt for å fjerne adaptere og filtrere sekvenser av dårlig kvalitet. Ideelt sett burde kvaliteten og antall gjenværende reads blitt kvalitetsjekket med FastQC etter preprosesseringen. Opptelling av antall reads etter trimmingen kan være en god indikator på hvordan preprosesseringen gikk. Siden trimmomatic og assembleringen ble gjort i en ferdig oppsatt workflow med Shovill var ikke dette mulig. En stolte derfor på at readsene var av god nok kvalitet etter trimmingen.

### 1.16 Kvalitetskontroll av assembly

Reads fra sekvensering ble forsøkt rekonstruert med de novo assembleringsverktøyet SPAdes. For å se hvor vellykket assembleringen var ble QUAST brukt for å sammenligne sekvensene med et eksisterende fullstendig genom fra samme stamme (*S.* Typhimurium). Contigs-filene fra assembleringen ble evaluert basert på genomsekvensen til et referansegenom. Etter at genomet ble forsøkt rekonstruert var det ønskelig med så få contigs som mulig. Ideelt sett bør et perfekt assembly for ett isolat gi en enkel contig som dekker hele genomet. For korte reads er ikke dette oppnåelig grunnet repeterende sekvenser, men en lavest mulig antall contigs er ønsket. Quast-rapporten viste 23 isolater som skilte seg ut basert på antall contigs og N50 verdier. 5 av isolatene med readlengde på 300 bp hadde gjennomsnittlig 10977 contigs, mens alle isolatene med readlengde på 150 bp hadde gjennomsnittlig 1478 contigs som er høyt sammenlignet med 50 contigs som gjaldt for de resterende isolatene. Kortere reads har en

tendens til å få mer contigs etter assembleringen. Basert på erfaring gjelder dette rundt 10-15 contigs flere og er dermed ikke grunnlaget til 1478 contigs som oppdaget her.

N50 verdien for de isolatene som skilte seg ut basert på antall contigs var 6630 (readlengde 150 bp) og 22022 (readlengde 300 bp). Dette er sammenlignet med N50 verdien på 326988 for de resterende isolatene svært lavt. N50 verdien er sekvenslengden til den korteste contigen ved 50% av den totale genomlengden. Ideelt sett burde verdien vært likt antall basepar totalt i genomet. Lave N50 verdier kan tyde på mer repeterende genomer, lav kvalitet og korte reads. N50 verdier i seg selv er ofte ikke nok til å evaluere hvor suksessfull assembleringen var og ble derfor brukt kombinert med antall contigs for å vurdere kvaliteten. Det tillates med en viss variasjon mellom kvalitetsverdiene, men siden de 23 isolatene viste stor variasjon mellom N50 verdi og antall contigs i forhold til de andre isolatene ble de fjernet fra datasettet før videre fylogenetiske analyser.

En årsak til at noen av isolatene viste dårlig kvalitet etter assemblering kan være at Trimmomatic ikke fikk fjernet alt av adaptere og sekvenser med dårlig kvalitet før assembleringen. På en annen side kan det også bety at trimmingen har fjernet reads med viktig informasjon før assemblering av genomet. Av de 23 isolatene som ved hjelp av QUAST ble vurdert til å ha for dårlig kvalitet til å inkluderes i videre analyser var 19 isolater sekvensert med Nextseq som ga reads med lengde på 150 bp. Sekvensering med NextSeq brukes ikke rutinemessig på Veterinærinstituttet. Siden Veterinærinstituttet har mindre erfaringer med denne metoden, var det vanskelig å tolke hva som var grunnen til den lave kvaliteten.

### 1.17 Sammenligning av SISTR resultater med rapportert serovariant

*S.* Typhimurium isolater kan deles inn i serovarianter i henhold til Kauffman-White-klassifiseringen basert på deres flagellære (H) og somatiske antigener eller ved bruk av genombaserte serotypingsmetoder. Tabell 7 i vedlegg viser en oversikt over datasettet brukt i oppgaven med rapporterte serovarianter gjort fenotypisk av laboratoriepersonell på Veterinærinstituttet. Karakterisering av serovariant fra SISTR ble brukt for å sjekke samsvar med serotypen som ble identifisert fenotypisk på laboratoriet.

SISTR predikerer serovar fra arten *Salmonella enterica* basert på både antigen og cgMLST-metoden som beskrevet i (1.3.3). For alle isolatene i oppgaven samsvarte rapporterte

serovarianter fra SISTR med fenotypisk serotyping, og bekrefter at alle isolatene er av arten *Salmonella enterica* med serovar *S*. Typhimurium. Det er ikke alle serovarianter det er like godt samsvar mellom fenotypisk analyse og SISTR som for *S*. Typhimurium. Men etter hvert som flere genomer legges til NCBI *Salmonella*-databasen, vil SISTR utvide nøyaktigheten enda mer i verifisering av *Salmonella*-serovarianter.

Ved karakterisering av serovarianten ble O-antigen faktor (1,4, [5] og 12) H1 antigen (i) og H2 antigen (1,2) karakterisert for alle *Salmonella* isolatene, med unntak av to. Disse funnene har i tidligere studier blitt karakterisert som *S*. Typhimurium og dette samsvarer med serovarianten foreslått av SISTR (Grimont & Weill, 2007). For to av isolatene klarte ikke SISTR å finne H2 antigener, som kan indikere at disse var monofasiske. Ved å slå opp isolatene i PJS til Veterinærinstituttet kunne en se at de monofasiske funnene samsvarte med resultater fra serotypingen som ble identifisert fenotypisk på laboratoriet.

### 1.18 Tildeling av ST med MLST

MLST tildelte en allele profil som bestemte ST for *S*. Typhimurium isolatene. Tre sekvenstyper som ofte er assosiert med *S*. Typhimurium er ST19, ST313 og ST34 (Achtman et al., 2012). Resultatene fra MLST inkluderte alle de nevnte STene, i tillegg ble ST376, ST432 og ST568 identifisert fra ett isolat hver. ST19 inkluderte absolutt flest isolater, og er også den sekvenstypen som ser ut til å være mest vanlig andre steder (Alikhan et al., 2018). Et interessant funn var at de monofasiske variantene oppdaget i SISTR ble tildelt ST34, som samsvarer med studien til Achtman (Achtman et al., 2012). ST34-gruppen inneholdt også isolater som ikke var monofasiske. ST432 skilte seg ut med en avstand på 7 alleler fra de andre STene. Den genetiske endringen viser at evolusjonstiden fra ST432 sannsynligvis er lengre enn for de resterende isolatene.

MLST er en fin metode for å få en rask oversikt over populasjonsstrukturen til *S*. Typhimurium. En ulempe med denne metoden er at den ikke ga den fine oppløsningen som var nødvendig for uttbruddsanalyser.

### 1.19 SNP analyse for ST19 og ST34

SNP analyser har en høy oppløsning og ble brukt for å vurdere en mulig smitteoverføring mellom produksjonsdyr og ville dyr. Noen av isolatene inkludert i datasettet for SNP ble ikke analysert av parSNP. Det ble ikke oppdaget noen årsak til dette etter at fasta-filen fra assembly og kvalitetsrapporten fra QUAST ble kontrollsjekket. For å se nærmere på dette kunne SNP analysen blitt gjort flere ganger for å se om de samme isolatene ble utelukket. Grunnet tidsbegrensinger ble ikke dette undersøkt nærmere og isolatene som ikke var med i outputet etter parSNP ble dermed ikke inkludert i SNP analysen.

Det ble observert to klustre (kluster 1 og 2) i SNP analysen for ST19 (figur 15). Flere av isolatene samlet fra produksjonsdyr falt utenfor klusterne og ble dermed ikke inkludert for å vurdere smitteoverføring mellom produksjonsdyr og ville dyr. Kluster 1 bestod av 8 isolater samlet fra produksjonsdyr. Siden isolatene inkludert i klusteret hadde små SNP-forskjeller mellom seg, var det vanskelig å dele isolatene inn i subkluster. Kluster 1 ble derfor delt inn i to grupper (A og B) som inkluderte isolatene samlet fra produksjonsdyr og tilhørende isolater med små SNP forskjeller.

Gruppe A i kluster 1 inkluderte 3 isolater fra produksjonsdyr og 2 isolater fra ville dyr med få (4-12 SNP) forskjeller mellom seg (figur 17). Alle isolatene samlet fra produksjonsdyr hadde en tett kobling til ett isolat fra ville dyr, som kan være en indikasjon på at det har skjedd en smitteoverføring fra villedyr til produksjonsdyr. Isolatene fra 2013 og 2014 hadde færre SNP forskjeller mellom seg, sammenlignet med isolatet fra 2015 (tabell 4). Dette kan tyde på at genomet har utviklet seg/forandret seg med tiden, ved for eksempel miljøpåvirkninger og vertstilpasninger. Siden isolatene i klusteret fra 2013 inkluderte både et produksjonsdyr og et vilt dyr var det vanskelig å vurdere om produksjonsdyr ble smittet av ville dyr eller omvendt i gruppe A.

For å finne en sannsynlig smittekilde kan en se på isolatene utenfor gruppe A i kluster 1 fra de tidligere årene. Alle isolatene fra senere år, med unntak av ett, var isolert fra ville dyr. Dette kan være en indikasjon på at smitten av *S.* Typhimurium var etablert blant de ville dyrene først, og deretter smittet produksjonsdyr. På en annen side består datasettet av et utvalg av isolater, og isolater fra produksjonsdyr fra de senere årene kan dermed ikke ha blitt fanget opp, og en kan med datasettet ikke utelukke at produksjonsdyrene er smittekilden.

Gruppe B fra kluster 1 inkluderte 4 isolater fra produksjonsdyr, 6 isolater fra ville dyr og ett isolat fra sport- og familiedyr. SNP distansen innad i gruppe B var større enn for Gruppe A. I likhet med isolater i Gruppe A hadde flere av produksjonsdyrene flere SNP forskjeller til isolater fra ville dyr sammenlignet med et annet isolat fra produksjonsdyr. SNP forskjeller i gruppe A og B i ST19 viste dermed en sterk indikasjon på at det har skjedd en smitteoverføring av *S.* Typhimurium blant produksjonsdyr og ville dyr.

For ST19 var det ett isolat fra produksjonsdyr som ikke ble inkludert i en gruppe (figur 17) grunnet større SNP forskjeller til andre produksjonsdyr og ville dyr. Isolatet hadde på en annen side få og i noen tilfeller ingen SNP-forskjeller til flere katteisolater fra sport- og familiedyr. Alle isolatene, inkludert produksjonsdyret, var fra 2020, og en kan dermed anta at det har skjedd ett smitteutbrudd blant katter. Katter beveger seg ute over større områder, og kan på den måte ha kommet seg inn i områder i nærheten av produksjonsdyret som kan være årsak til overføring av smitte.

Selv om kluster 2 bestod av en klynge med nærere relaterte isolater var det fremdeles større forskjeller innad i klusteret enn for kluster 1. For ett av isolatene samlet fra produksjonsdyr ble det observert små SNP forskjeller til tre andre isolater som inkluderte ett isolat fra ville dyr (Gruppe C figur 18). Isolatet fra produksjonsdyret hadde minst SNP forskjeller til et annet isolat fra ville dyr. Dette funnet kan støtte at ville dyr og produksjonsdyr har smittet hverandre, men for isolat C er produksjonsdyrisolatet samlet fra tidligere år enn det fra ville dyr. Dette motsier hypotesen at ville dyr smitter produksjonsdyr med *S.* Typhimurium. Siden dette kun inkluderer funn for to isolater, kan resultatene i gruppe C ikke slå fast hvor smitten kommer fra.

I motsetning til Gruppe A og B inkluderer gruppe C to isolater fra sport- og familiedyr som har lite SNP forskjeller til isolatet fra produksjonsdyr. Siden kluster 2 er lite og består av isolater fordelt på flere år, vil hvert år være dårlig representert. Det er vanskelig å vurdere eventuelle smitteutbrudd for isolater i ett lite kluster fordelt på 17 år fordi genomet kan utvikle seg over tid. Dataen fra ST19 gjør at vi ikke kan utelukke at det er en forbindelse mellom smitte av produksjonsdyr og ville dyr, og resultatene (spesielt for gruppe A) gir en sterk indikasjon på at det er ville dyr som har smittet produksjonsdyr.

ST34 bestod av 8 isolater kun samlet fra produksjonsdyr (figur 19). Isolatene hadde en større varians i SNP forskjeller enn det vi så for ST19 (figur 17). Basert på SNP analysen av ST34 var det dermed ikke mulig å definere ett utbrudd, eller vurdere smitten mellom produksjonsdyr og ville dyr.

SNP analysen gjort i Galaxy produserte ikke output filer som viste hvor mye av genomet som ble dekket. Dette er viktig informasjon for å vite hvor mye en kan stole på resultatene. Det ble derfor i tillegg gjort en ny SNP analyse ved bruk av workflowen ALPPACA (version 1.0.0) (https://github.com/NorwegianVeterinaryInstitute/ALPPACA/) utført av Veterinærinstituttet, der ParSNP produserte en output som viste hvor mye av genomet som ble dekket. Rapporten viste at for ST19 (Vedlegg tabell 12) ble gjennomsnittlig 93 % og for ST34 (Vedlegg tabell 13) ble 88 % av genomene dekket under analysen. Hvis deler av genomet ikke er blitt brukt i sammenligningen kan det føre til at analysen fjerner flere forskjeller som fører til mindre SNP-forskjeller. For å kontrollere dette kunne SNP analysen blitt gjort på nytt kun med isolatene innad i klusteret for å se om SNP forskjellene ble mindre når større del av genomet ble dekket.

# Konklusjon

Gjennom arbeidet i denne masteroppgaven ble det ved hjelp av bioinformatiske analyser undersøkt om ville dyr er en potensiell smittekilde av *S.* Typhimurium til produksjonsdyr i Norge. For isolatene fra produksjonsdyr som var samlet i ett kluster med isolater fra ville dyr ble det observert en likhet mellom ville dyr og produksjonsdyr som indikerer smitteoverføring. Basert på resultatene fra det anvendte datasettet var det flere steder sterke indikasjoner på at ville dyr er en potensiell smittekilde av *S.* Typhimurium til produksjonsdyr i Norge.

Et best mulig datasett bør inneholde flere produksjonsdyr isolater som blir inkludert i samme klustre som isolater fra ville dyr. Datasettet burde også inneholde isolater representert av både produksjonsdyr og villedyr samlet opp over flere år. Flere av isolatene samlet fra produksjonsdyr var ikke like nok andre isolater fra ville dyr til å vurdere om de var fra samme kilde. Siden datasettet inkludert i klusterne bestod av få produksjonsdyrisolater fra tidligere år var det vanskeligere å vurdere om det var ville dyr eller produksjonsdyr som var kilden til smitten eller om både produksjonsdyr og ville dyr er et mulig reservoar. På bakgrunn av dette kunne ikke resultatene si noe om at ville dyr er den vanligste smittekilden av *S.* Typhimurium til produksjonsdyr i Norge, men resultatene viste at smitte fra ville dyr mest sannsynlig har skjedd.

# Referanser


Achtman, M., Wain, J., Weill, F.-X., Nair, S., Zhou, Z., Sangal, V., Krauland, M. G., Hale, J. L., Harbottle, H., Uesbeck, A., et al. (2012). Multilocus Sequence Typing as a Replacement for Serotyping in Salmonella enterica. *PLOS Pathogens*, 8 (6): e1002776. doi: 10.1371/journal.ppat.1002776.

Alikhan, N. F., Zhou, Z., Sergeant, M. J. & Achtman, M. (2018). A genomic overview of the population structure of Salmonella. *PLoS Genet*, 14 (4): e1007261. doi: 10.1371/journal.pgen.1007261.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*, 19 (5): 455-77. doi: 10.1089/cmb.2012.0021.

Baptista, R. P., Reis-Cunha, J. L., DeBarry, J. D., Chiari, E., Kissinger, J. C., Bartholomeu, D. C. & Macedo, A. M. (2018). Assembly of highly repetitive genomes using short reads: the genome of discrete typing unit III Trypanosoma cruzi strain 231. *Microbial genomics*, 4 (4): e000156. doi: 10.1099/mgen.0.000156.

Bøvre, K. (2021). *Enterobacteriaceae*. Tilgjengelig fra: https://sml.snl.no/Enterobacteriaceae (lest 09.02.2022).

Croucher, N. J., Page, A. J., Connor, T. R., Delaney, A. J., Keane, J. A., Bentley, S. D., Parkhill, J. & Harris, S. R. (2014). Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research*, 43 (3): e15-e15. doi: 10.1093/nar/gku1196.

Dutta, S. (2021). *What is Phylogenetic Analysis?* News medical life sciences. Tilgjengelig fra: https://www.news-medical.net/health/What-is-Phylogenetic-Analysis.aspx (lest 18.03.18).

Enright, M. C., Day, N. P., Davies, C. E., Peacock, S. J. & Spratt, B. G. (2000). Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of Staphylococcus aureus. *J Clin Microbiol*, 38 (3): 1008-15. doi: 10.1128/jcm.38.3.1008-1015.2000.

Feil, E. J., Li, B. C., Aanensen, D. M., Hanage, W. P. & Spratt, B. G. (2004). eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *J Bacteriol*, 186 (5): 1518-30. doi: 10.1128/jb.186.5.1518-1530.2004.

Ferrari, R. G., Panzenhagen, P. H. N. & Conte-Junior, C. A. (2017). Phenotypic and Genotypic Eligible Methods for Salmonella Typhimurium Source Tracking. *Frontiers in microbiology*, 8: 2587-2587. doi: 10.3389/fmicb.2017.02587.

Folkehelseinstituttet. (2019). *Salmonellose*. Tilgjengelig fra: https://www.fhi.no/nettpub/smittevernveilederen/sykdommer-a-a/salmonellose---veileder-for-helsepe/#meldings-og-varslingsplikt.

Granum, E. P. (2017). *Matforgiftning*. Oslo: Cappelen Damm.

Grimont, P. & Weill, F.-X. (2007). Antigenic Formulae of the Salmonella serovars, (9th ed.) Paris: WHO Collaborating Centre for Reference and Research on Salmonella. *Institute Pasteur.*: 1-166.

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29 (8): 1072-5. doi: 10.1093/bioinformatics/btt086.

Heather, J. M. & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, 107 (1): 1-8. doi: 10.1016/j.ygeno.2015.11.003.

Ibrahim, G. M. & Morin, P. M. (2018). Salmonella Serotyping Using Whole Genome Sequencing. *Frontiers in Microbiology*, 9. doi: 10.3389/fmicb.2018.02993.
Illumina. (2011). Quality Scores for Next-Generation Sequencing *Technical Note: Sequencing*.
Illumina. (2022a). *A beginner`s guide to NGS*. Tilgjengelig fra: https://www.illumina.com/science/technology/next-generation-sequencing/beginners.html (lest 24.02.2022).
Illumina. (2022b). An introduction to Next-Generation
Sequencing Technology.
Jarp, J. (2018). *Dette vet vi om Salmonella Typhimurium-bakterien*. Tilgjengelig fra: https://www.vetinst.no/nyheter/dette-vet-vi-om-salmonella-typhimurium-bakterien (lest 23.02.2022).
Johnson, Z. J., Krutkin, D. D., Bohutskyi, P. & Kalyuzhnaya, M. G. (2021). Chapter Eight - Metals and methylotrophy: Via global gene expression studies. I: Cotruvo, J. A. (red.) b. 650 *Methods in Enzymology*, s. 185-213: Academic Press.
Kauffmann, F. & Edwards, P. R. (1952). Classification and Nomenclature of Enterobacteriaceae. *International Journal of Systematic and Evolutionary Microbiology*, 2 (1): 2-8. doi: https://doi.org/10.1099/0096266X-2-1-2.
Larsen, M. V., Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R. L., Jelsbak, L., Sicheritz-Pontén, T., Ussery, D. W., Aarestrup, F. M., et al. (2012). Multilocus sequence typing of total-genome-sequenced bacteria. *J Clin Microbiol*, 50 (4): 1355-61. doi: 10.1128/jcm.06094-11.
Leekitcharoenphon, P., Lukjancenko, O., Friis, C., Aarestrup, F. M. & Ussery, D. W. (2012). Genomic variation in Salmonella enterica core genes for epidemiological typing. *BMC genomics*, 13: 88-88. doi: 10.1186/1471-2164-13-88.
Leggett, R., Ramirez-Gonzalez, R., Verweij, W., Kawashima, C., Iqbal, Z., Jones, J., Caccamo, M. & Maclean, D. (2013). Identifying and Classifying Trait Linked Polymorphisms in Non-Reference Species by Walking Coloured de Bruijn Graphs. *PloS one*, 8: e60058. doi: 10.1371/journal.pone.0060058.
Letunic, I. & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, 49 (W1): W293-W296. doi: 10.1093/nar/gkab301.
Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., et al. (2011). Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics*, 11 (1): 25-37. doi: 10.1093/bfgp/elr035.
Lischer, H. E. L. & Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, 18 (1): 474. doi: 10.1186/s12859-017-1911-6.
MacDonald, E., White, R., Mexia, R., Bruun, T., Kapperud, G., Brandal, L. T., Lange, H., Nygård, K. & Vold, L. (2018). The role of domestic reservoirs in domestically acquired Salmonella infections in Norway: epidemiology of salmonellosis, 2000-2015, and results of a national prospective case-control study, 2010-2012. *Epidemiology and infection*, 147: e43-e43. doi: 10.1017/S0950268818002911.
Maduranga, U. (2020). *Genome Assembly using de Bruijn Graphs*. Tilgjengelig fra: https://towardsdatascience.com/genome-assembly-using-de-bruijn-graphs-69570efcc270 (lest 04.03.22).
Maiden, M. C. (2006). Multilocus sequence typing of bacteria. *Annu Rev Microbiol*, 60: 561-88. doi: 10.1146/annurev.micro.59.030804.121325.

Matthews, T. C., Bristow, F. R., Griffiths, E. J., Petkau, A., Adam, J., Dooley, D., Kruczkiewicz, P., Curatcha, J., Cabral, J., Fornika, D., et al. (2018). The Integrated Rapid Infectious Disease Analysis (IRIDA) Platform. *bioRxiv*: 381830. doi: 10.1101/381830.

Monte, D. F. M. & Sellera, F. P. (2020). Salmonella. *Emerging Infectious Diseases*, 26 (12): 2955-2955. doi: 10.3201/eid2612.ET2612.

Nelseon, G. E. & Greene, M. H. (2022). Enterobacteriaceae. In Mandell, Douglas and Bennett's principles and practice of infectious diseases.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S. & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*, 17 (1): 132. doi: 10.1186/s13059-016-0997-x.

Papanikolaou, N., Trachana, K., Theodosiou, T., Promponas, V. J. & Iliopoulos, I. (2009). Gene socialization: gene order, GC content and gene silencing in Salmonella. *BMC Genomics*, 10: 597. doi: 10.1186/1471-2164-10-597.

Pearce, M. E., Alikhan, N.-F., Dallman, T. J., Zhou, Z., Grant, K. & Maiden, M. C. J. (2018). Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak. *International journal of food microbiology*, 274: 1-11. doi: 10.1016/j.ijfoodmicro.2018.02.023.

Refsum, T., Heir, E., Kapperud, G., Vardund, T. & Holstad, G. (2002). Molecular epidemiology of Salmonella enterica serovar typhimurium isolates determined by pulsed-field gel electrophoresis: comparison of isolates from avian wildlife, domestic animals, and the environment in Norway. *Appl Environ Microbiol*, 68 (11): 5600-6. doi: 10.1128/aem.68.11.5600-5606.2002.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74 (12): 5463-7. doi: 10.1073/pnas.74.12.5463.

Scott, A. D. & Baum, D. A. (2016). Phylogenetic Tree. I: Kliman, R. M. (red.) *Encyclopedia of Evolutionary Biology*, s. 270-276. Oxford: Academic Press.

Seemann, T. (2019). *snp-dists*. Tilgjengelig fra: https://github.com/tseemann/snp-dists.

Veterinærinstituttet. (2022). *Salmonella*. Tilgjengelig fra: https://www.vetinst.no/sykdom-og-agens/salmonella (lest 29.04.22).

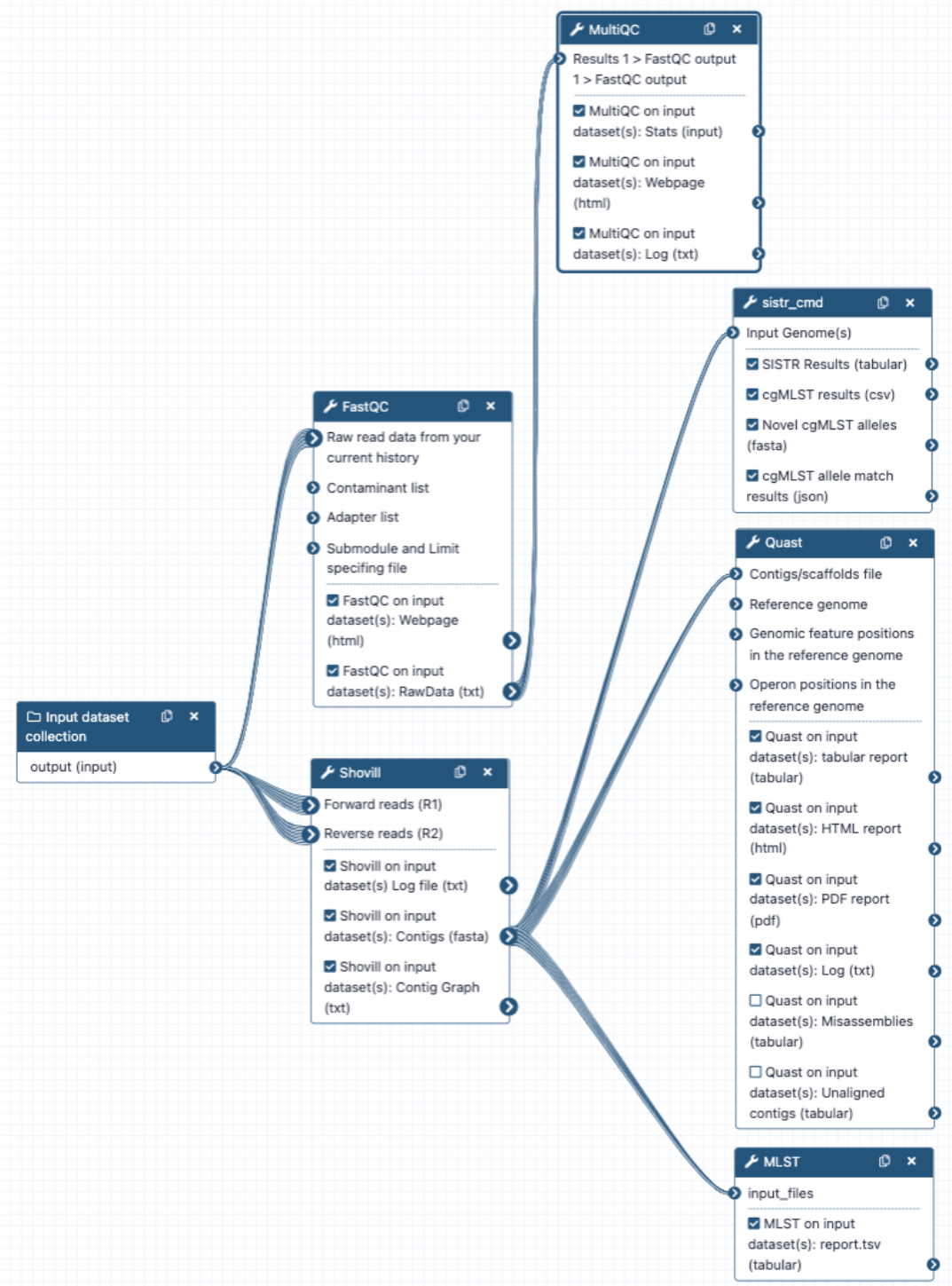
Yoshida, C. E., Kruczkiewicz, P., Laing, C. R., Lingohr, E. J., Gannon, V. P., Nash, J. H. & Taboada, E. N. (2016). The Salmonella In Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft Salmonella Genome Assemblies. *PLoS One*, 11 (1): e0147101. doi: 10.1371/journal.pone.0147101.

Zhou, Z., Alikhan, N.-F., Sergeant, M. J., Luhmann, N., Vaz, C., Francisco, A. P., Carriço, J. A. & Achtman, M. (2018). GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome research*, 28 (9): 1395-1404. doi: 10.1101/gr.232397.117.

# Vedlegg

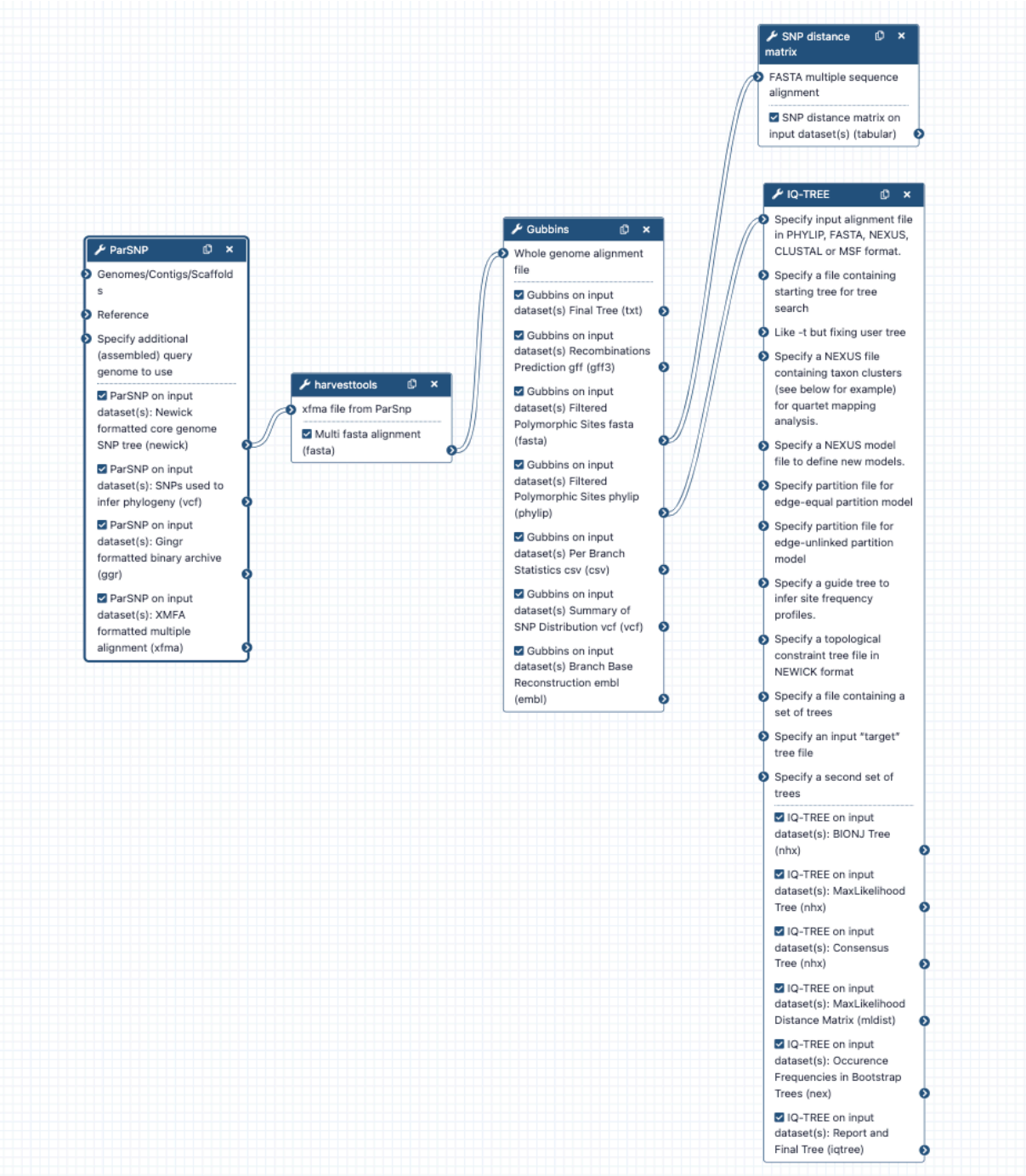
**Galaxy workflow 1: Preprossesering, Assemblering og Typing**

**Figur 19:** Detaljert workflow designet i Galaxy brukt for preprosessering, assemblering og typing med oversikt over inputs og outputs. Workflowen inkluderer verktøyene FastQC, MultiQC, Shovill, QUAST, SISTR og MLST.



**Galaxy workflow 2: SNP analyse**

**Figur 20:** Detaljer workflow designet i Galaxy brukt for SNP analyse med oversikt over inputs og outputs. Workflowen inkluderer ParSNP, Harvesttools, Gubbins, IQ-TREE og SNP distance matrix.



## Oversikt over datasettet brukt i oppgaven

**Tabell 7:** Oversikt over datasettet brukt i oppgaven. Tabellen viser en oversikt over ID nummer, kategori, art, serovariant og opprinnelse.

| ID | Kategori | Art | Serovariant | Opprinnelse |
|---|---|---|---|---|
| 2001-01-2236 | Produksjonsdyr | Storfe | S. Typhimurium | Norsk |
| 2002-01-22-63 | Produksjonsdyr | Storfe | S. Typhimurium | Norsk |
| 2002-01-518 | Ville dyr | Villfugl | S. Typhimurium | Norsk |
| 2003-01-723 | Produksjonsdyr | Gris | S. Typhimurium | Norsk |
| 2003-40-20444 | Produksjonsdyr | Høns | S. Typhimurium | Norsk |
| 2003-50-2805 | Kjæledyr | Hest | S. Typhimurium | Norsk |
| 2004-01-1763-12 | Ville dyr | Piggsvin | S. Typhimurium | Norsk |
| 2004-01-559 | Produksjonsdyr | Gris | S. Typhimurium | Norsk |
| 2004-10-157 | Ville dyr | Villfugl (Grønnfink) | S. Typhimurium | Norsk |
| 2004-40-18968-226 | Produksjonsdyr | Gris | S. Typhimurium | Norsk |
| 2004-50-547 | Ville dyr | Villfugl (Dompap) | S. Typhimurium | Norsk |
| 2005-10-101 | Ville dyr | Villfugl | S. Typhimurium | Norsk |
| 2005-10-175 | Ville dyr | Villfugl | S. Typhimurium | Norsk |
| 2005-10-84 | Ville dyr | Villfugl | S. Typhimurium | Norsk |
| 2006-01-1845 | Ville dyr | Villfugl (Moskusand) | S. Typhimurium | Norsk |
| 2006-01-195 | Produksjonsdyr | Høns | S. Typhimurium | Norsk |
| 2006-01-3807-21 | Kjæledyr | Hest | S. Typhimurium | Norsk |
| 2006-40-8604 | Ville dyr | Villfugl | S. Typhimurium | Norsk |
| 2006-50-3052 | Produksjonsdyr | Storfe | S. Typhimurium | Norsk |
| 2006-60-6746 | Ville dyr | Villfugl | S. Typhimurium | Norsk |
| 2006-70-1122 | Ville dyr | Villfugl (Grønnfink) | S. Typhimurium | Norsk |
| 2006-70-2104 | Ville dyr | Villfugl (Grønnfink) | S. Typhimurium | Norsk |
| 2006-70-3936 | Ville dyr | Villfugl | S. Typhimurium | Norsk |
| 2007-01-1883 | Produksjonsdyr | Svin | S. Typhimurium | Norsk |
| 2007-01-3478 | Produksjonsdyr | Svin | S. Typhimurium | Norsk |
| 2007-01-3479-18 | Produksjonsdyr | Svin | S. Typhimurium | Norsk |
| 2007-10-250 | Ville dyr | Villfugl (Gråsisik) | S. Typhimurium | Norsk |
| 2007-10-274-3 | Ville dyr | Villfugl (Gråsisik) | S. Typhimurium | Norsk |
| 2007-10-341 | Ville dyr | Villfugl (Gråsisik) | S. Typhimurium | Norsk |
| 2007-10-509-2 | Ville dyr | Villfugl (Gråsisik) | S. Typhimurium | Norsk |
| 2007-40-3693 | Produksjonsdyr | Storfe | S. Typhimurium | Norsk |
| 2007-50-375 | Produksjonsdyr | Storfe | S. Typhimurium | Norsk |
| 2007-50-378 | Produksjonsdyr | Storfe | S. Typhimurium | Norsk |
| 2007-60-1234 | Ville dyr | Villfugl (Gråsisik) | S. Typhimurium | Norsk |
| 2007-70-2403-2 | Ville dyr | Villfugl (Gråsisik) | S. Typhimurium | Norsk |
| 2008-40-4269 | Produksjonsdyr | Storfe | S. Typhimurium | Norsk |
| 2009-02-1811 | Produksjonsdyr | Storfe | S. Typhimurium | Norsk |
| 2010-10-195 | Ville dyr | Villfugl (Grønnfink) | S. Typhimurium | Norsk |
| 2010-40-247-3 | Produksjonsdyr | Svin | S. Typhimurium | Norsk |
| 2010-40-35-4 | Produksjonsdyr | Svin | S. Typhimurium | Norsk |
| 2010-80-25-1 | Ville dyr | Villfugl (Dompap) | S. Typhimurium | Norsk |
| 2011-04-7074-1 | Ville dyr | Villfugl (Gråsisik) | S. Typhimurium | Norsk |
| 2011-04-9056-1 | Ville dyr | Villfugl (Duefamilien) | S. Typhimurium | Norsk |
| 2011-40-7145-1 | Ville dyr | Villfugl (Tamgås) | S. Typhimurium | Norsk |
| 2011-50-2020-41 | Produksjonsdyr | Svin | S. Typhimurium | Norsk |
| 2011-50-2222-1 | Produksjonsdyr | Storfe | S. Typhimurium | Norsk |
| 2011-60-6693 | Produksjonsdyr | Høns | S. Typhimurium | Norsk |
| 2011-70-2808-2 | Ville dyr | Villfugl (Gråsisik) | S. Typhimurium | Norsk |
| 2012-01-6145-1 | Kjæledyr | Hest | S. Typhimurium | Norsk |
| 2012-04-7212-1 | Ville dyr | Villfugl (Tamdue) | S. Typhimurium | Norsk |
| 2012-60-2316-31 | Ville dyr | Villfugl (Gråmåke) | S. Typhimurium | Norsk |
| 2013-01-5866-2 | Produksjonsdyr | Sau | S. Typhimurium | Norsk |
| 2013-40-7492-9 | Produksjonsdyr | Svin | S. Typhimurium | Norsk |
| 2013-60-2782 | Ville dyr | Villfugl (Grønnsisik) | S. Typhimurium | Norsk |
| 2014-04-2493-1 | Ville dyr | Villfugl (Kjøttmeis) | S. Typhimurium | Norsk |
| 2014-04-6297 | Ville dyr | Villfugl (Dompap) | S. Typhimurium | Norsk |
| 2014-40-11112 | Produksjonsdyr | Høns | S. Typhimurium | Norsk |
| 2014-40-15837 | Produksjonsdyr | Storfe | S. Typhimurium | Norsk |
| 2015-01-2549-1 | Produksjonsdyr | Storfe | S. Typhimurium | Norsk |
| 2015-01-2739 | Produksjonsdyr | Storfe | S. Typhimurium | Norsk |

# Kvalitetsverdier fra MultiQC og QUAST

**Tabell 8:** En samlet rapport med kvalitetsverdier for alle isolatene fra MultiQC og QUAST. Isolater fjernet fra datasettet etter kvalitetskontrollen er markert.

| | MultiQC | | | | | | | | QUAST | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sample | Unique Reads | Duplicate Reads | Total reads | Qual bases % | Sequence read lenght | %GC | Genome size, average | Coverage | N50 (>15000) | Number of contigs (<500) | Total assembly lenght |
| 2001-01-2236_R1 | 444520 | 100922 | 545442 | 81,49720777 | 301 | 51 | | | | | |
| 2001-01-2236_R2 | 449719 | 95723 | 545442 | 82,45037969 | 301 | 51 | 4857450 | 68 | 341585 | 54 | 4928457 |
| 2002-01-22-63_R1 | 604179 | 169313 | 773492 | 78,11056869 | 301 | 51 | | | | | |
| 2002-01-22-63_R2 | 615288 | 158204 | 773492 | 79,54678264 | 301 | 51 | 4857450 | 96 | 223684 | 58 | 4972202 |
| 2002-01-518_R1 | 350901 | 58261 | 409162 | 85,76089666 | 301 | 51 | | | | | |
| 2002-01-518_R2 | 356372 | 52790 | 409162 | 87,09801986 | 301 | 51 | 4857450 | 51 | 364381 | 56 | 4811800 |
| 2003-01-723_R1 | 368135 | 59834 | 427969 | 86,01908082 | 301 | 51 | | | | | |
| 2003-01-723_R2 | 371780 | 56189 | 427969 | 86,87077802 | 301 | 51 | 4857450 | 53 | 376582 | 49 | 4767935 |
| 2003-40-20444_R1 | 385729 | 66272 | 452001 | 85,33808554 | 301 | 51 | | | | | |
| 2003-40-20444_R2 | 390030 | 61971 | 452001 | 86,28963210 | 301 | 52 | 4857450 | 56 | 207707 | 42 | 4730513 |
| 2003-50-2805_R1 | 343141 | 49457 | 392598 | 87,40263578 | 301 | 51 | | | | | |
| 2003-50-2805_R2 | 377536 | 15062 | 392598 | 96,16350567 | 301 | 51 | 4857450 | 49 | 198691 | 66 | 4994495 |
| 2004-01-1763-12_R1 | 365888 | 58527 | 424415 | 86,20995959 | 301 | 51 | | | | | |
| 2004-01-1763-12_R2 | 366155 | 58260 | 424415 | 86,27286971 | 301 | 51 | 4857450 | 53 | 266893 | 43 | 4731054 |
| 2004-01-559_R1 | 396294 | 66262 | 462556 | 85,67481559 | 301 | 49 | | | | | |
| 2004-01-559_R2 | 399116 | 63440 | 462556 | 86,28490388 | 301 | 50 | 4857450 | 57 | 182236 | 73 | 4767046 |
| 2004-10-157_R1 | 497036 | 108899 | 605935 | 82,02794029 | 301 | 51 | | | | | |
| 2004-10-157_R2 | 502493 | 103442 | 605935 | 82,92853194 | 301 | 51 | 4857450 | 75 | 341452 | 53 | 4766934 |
| 2004-40-18968-226_R1 | 445245 | 85763 | 531008 | 83,84901922 | 301 | 51 | | | | | |
| 2004-40-18968-226_R2 | 481700 | 49308 | 531008 | 90,71426419 | 301 | 52 | 4857450 | 66 | 376603 | 50 | 4767849 |
| 2004-50-547_R1 | 452890 | 89413 | 542303 | 83,51235379 | 301 | 51 | | | | | |
| 2004-50-547_R2 | 458706 | 83597 | 542303 | 84,58481698 | 301 | 51 | 4857450 | 67 | 376579 | 46 | 4808954 |
| 2005-10-101_R1 | 284307 | 36002 | 320309 | 88,76022840 | 301 | 51 | | | | | |
| 2005-10-101_R2 | 286273 | 34036 | 320309 | 89,37401072 | 301 | 51 | 4857450 | 40 | 303402 | 49 | 4809008 |
| 2005-10-175_R1 | 326269 | 46719 | 372988 | 87,47439596 | 301 | 51 | | | | | |
| 2005-10-175_R2 | 330101 | 42887 | 372988 | 88,50177486 | 301 | 51 | 4857450 | 46 | 376329 | 51 | 4807529 |
| 2005-10-84_R1 | 318571 | 44450 | 363021 | 87,75552929 | 301 | 51 | | | | | |
| 2005-10-84_R2 | 320641 | 42380 | 363021 | 88,32574424 | 301 | 51 | 4857450 | 45 | 376603 | 46 | 4766767 |
| 2006-01-1845_R1 | 329847 | 50048 | 379895 | 86,82583345 | 301 | 51 | | | | | |
| 2006-01-1845_R2 | 333842 | 46053 | 379895 | 87,87743982 | 301 | 51 | 4857450 | 47 | 322477 | 50 | 4806034 |
| 2006-01-195_R1 | 415904 | 87909 | 503813 | 82,55126406 | 301 | 51 | | | | | |
| 2006-01-195_R2 | 419409 | 84404 | 503813 | 83,24695869 | 301 | 51 | 4857450 | 62 | 377380 | 50 | 4948752 |
| 2006-01-3807-21_R1 | 324644 | 47287 | 371931 | 87,28608263 | 301 | 51 | | | | | |
| 2006-01-3807-21_R2 | 327371 | 44560 | 371931 | 88,01928315 | 301 | 51 | 4857450 | 46 | 341394 | 50 | 4806320 |
| 2006-40-8604_R1 | 333834 | 64022 | 397856 | 83,90824821 | 301 | 51 | | | | | |
| 2006-40-8604_R2 | 337425 | 60431 | 397856 | 84,81083608 | 301 | 51 | 4857450 | 49 | 324913 | 55 | 4927809 |
| 2006-60-6746_R1 | 418232 | 88293 | 506525 | 82,56887617 | 301 | 51 | | | | | |
| 2006-60-6746_R2 | 422294 | 84231 | 506525 | 83,37081092 | 301 | 51 | 4857450 | 63 | 376580 | 50 | 4811253 |
| 2006-70-1122_R1 | 444632 | 90466 | 535098 | 83,09356417 | 301 | 51 | | | | | |
| 2006-70-1122_R2 | 448670 | 86428 | 535098 | 83,84819229 | 301 | 51 | 4857450 | 66 | 377148 | 49 | 4769605 |
| 2006-70-2104_R1 | 387262 | 68269 | 455531 | 85,01331413 | 301 | 51 | | | | | |
| 2006-70-2104_R2 | 390709 | 64822 | 455531 | 85,77001346 | 301 | 51 | 4857450 | 56 | 377148 | 50 | 4769298 |
| 2006-70-3936_R1 | 330071 | 49046 | 379117 | 87,06309662 | 301 | 51 | | | | | |
| 2006-70-3936_R2 | 333376 | 45741 | 379117 | 87,93485916 | 301 | 51 | 4857450 | 47 | 376329 | 47 | 4805386 |
| 2007-01-1883_R1 | 342156 | 50362 | 392518 | 87,16950560 | 301 | 51 | 4857450 | | | | |
| 2007-01-1883_R2 | 347619 | 44899 | 392518 | 88,56128891 | 301 | 51 | 4857450 | 49 | 301490 | 42 | 4803379 |
| 2007-01-3478_R1 | 294101 | 36736 | 330837 | 88,89604246 | 301 | 51 | 4857450 | | | | |
| 2007-01-3478_R2 | 294904 | 35933 | 330837 | 89,13876017 | 301 | 51 | 4857450 | 41 | 262876 | 47 | 4797170 |
| 2007-01-3479-18_R1 | 255861 | 29218 | 285079 | 89,75091115 | 301 | 51 | | | | | |
| 2007-01-3479-18_R2 | 257955 | 27124 | 285079 | 90,48544439 | 301 | 51 | 4857450 | 35 | 301490 | 50 | 4801365 |
| 2007-10-250_R1 | 455471 | 92932 | 548403 | 83,05406790 | 301 | 51 | 4857450 | | | | |
| 2007-10-250_R2 | 460835 | 87568 | 548403 | 84,03218071 | 301 | 51 | 4857450 | 68 | 376603 | 49 | 4808961 |
| 2007-10-274-3_R1 | 387893 | 65054 | 452947 | 85,63761323 | 301 | 51 | 4857450 | | | | |
| 2007-10-274-3_R2 | 394912 | 58035 | 452947 | 87,18724266 | 301 | 52 | 4857450 | 56 | 376329 | 49 | 4806694 |
| 2007-10-341_R1 | 342019 | 52604 | 394623 | 86,66980891 | 301 | 51 | | | | | |
| 2007-10-341_R2 | 345672 | 48951 | 394623 | 87,59550254 | 301 | 51 | 4857450 | 49 | 376329 | 52 | 4805759 |
| 2007-10-509-2_R1 | 320239 | 43837 | 364076 | 87,95938211 | 301 | 51 | | | | | |
| 2007-10-509-2_R2 | 322636 | 41440 | 364076 | 88,61776113 | 301 | 52 | 4857450 | 45 | 376603 | 49 | 4808935 |
| 2007-40-3693_R1 | 287817 | 55448 | 343265 | 83,84688215 | 301 | 51 | 4857450 | | | | |
| 2007-40-3693_R2 | 288308 | 54957 | 343265 | 83,98992032 | 301 | 51 | 4857450 | 43 | 279635 | 54 | 4907266 |
| 2007-50-375_R1 | 344180 | 52310 | 396490 | 86,80672905 | 301 | 51 | 4857450 | | | | |
| 2007-50-375_R2 | 346940 | 49550 | 396490 | 87,5028374 | 301 | 51 | 4857450 | 49 | 266893 | 39 | 4757073 |
| 2007-50-378_R1 | 373980 | 61804 | 435784 | 85,81774457 | 301 | 51 | 4857450 | | | | |
| 2007-50-378_R2 | 376481 | 59303 | 435784 | 86,39165275 | 301 | 51 | 4857450 | 54 | 266893 | 40 | 4758232 |
| 2007-60-1234_R1 | 340226 | 51057 | 391283 | 86,95138813 | 301 | 51 | 4857450 | | | | |
| 2007-60-1234_R2 | 341732 | 49551 | 391283 | 87,33627579 | 301 | 51 | 4857450 | 48 | 325504 | 54 | 4807562 |
| 2007-70-2403-2_R1 | 484346 | 103626 | 587972 | 82,37569136 | 301 | 51 | | | | | |
| 2007-70-2403-2_R2 | 484865 | 103107 | 587972 | 82,46396087 | 301 | 52 | 4857450 | 73 | 376329 | 50 | 4806589 |
| 2008-40-4269_R1 | 555896 | 146969 | 702865 | 79,09001017 | 301 | 51 | 4857450 | | | | |
| 2008-40-4269_R2 | 559770 | 143095 | 702865 | 79,64118287 | 301 | 51 | 4857450 | 87 | 226123 | 46 | 4862016 |
| 2009-02-1811_R1 | 429302 | 92615 | 521917 | 82,25484129 | 301 | 51 | 4857450 | | | | |
| 2009-02-1811_R2 | 434050 | 87867 | 521917 | 83,16456448 | 301 | 51 | 4857450 | 65 | 259667 | 41 | 4861712 |
| 2010-10-195_R1 | 436576 | 84583 | 521159 | 83,77021216 | 301 | 51 | 4857450 | | | | |
| 2010-10-195_R2 | 442021 | 79138 | 521159 | 84,81499888 | 301 | 52 | 4857450 | 65 | 341432 | 51 | 4808504 |
| 2010-40-247-3_R1 | 378656 | 76891 | 455547 | 83,12117081 | 301 | 51 | 4857450 | | | | |
| 2010-40-247-3_R2 | 381563 | 73984 | 455547 | 83,75930475 | 301 | 51 | 4857450 | 56 | 225886 | 50 | 4967407 |
| 2010-40-35-4_R1 | 591356 | 165885 | 757241 | 78,09349996 | 301 | 51 | 4857450 | | | | |
| 2010-40-35-4_R2 | 599774 | 157467 | 757241 | 79,20516718 | 301 | 51 | 4857450 | 94 | 226123 | 48 | 4974102 |
| 2010-80-25-1_R1 | 767883 | 656290 | 1424173 | 53,91781757 | 149 | 46 | 4857450 | | | | |
| 2010-80-25-1_R2 | 772495 | 651678 | 1424173 | 54,24165463 | 149 | 46 | 4857450 | 87 | 2336 | 2028 | 3533834 |
| 2011-04-7074-1_R1 | 607958 | 271895 | 879853 | 69,09767882 | 149 | 48 | | | | | |
| 2011-04-7074-1_R2 | 602540 | 277313 | 879853 | 68,48189413 | 149 | 48 | 4857450 | 54 | 4155 | 1625 | 4169382 |
| 2011-04-9056-1_R1 | 708937 | 492522 | 1201459 | 59,00634146 | 149 | 47 | 4857450 | | | | |
| 2011-04-9056-1_R2 | 715337 | 486122 | 1201459 | 59,53902713 | 149 | 47 | 4857450 | 74 | 2559 | 2017 | 3804448 |
| 2011-40-7145-1_R1 | 326959 | 52424 | 379383 | 86,18177409 | 301 | 51 | | | | | |
| 2011-40-7145-1_R2 | 326913 | 52470 | 379383 | 86,16964914 | 301 | 51 | 4857450 | 47 | 341591 | 53 | 4926601 |
| 2011-50-2020-41_R1 | 424673 | 81090 | 505763 | 83,96679868 | 301 | 51 | 4857450 | | | | |
| 2011-50-2020-41_R2 | 428985 | 76778 | 505763 | 84,81937192 | 301 | 51 | 4857450 | 63 | 219818 | 37 | 4760441 |
| 2011-50-2222-1_R1 | 421195 | 78729 | 499924 | 84,25180627 | 301 | 51 | | | | | |
| 2011-50-2222-1_R2 | 428823 | 71101 | 499924 | 85,7776382 | 301 | 51 | 4857450 | 62 | 275352 | 42 | 4799904 |
| 2011-60-6693_R1 | 380615 | 72735 | 453350 | 83,95610455 | 301 | 51 | | | | | |
| 2011-60-6693_R2 | 381796 | 71554 | 453350 | 84,21660968 | 301 | 52 | 4857450 | 56 | 316044 | 38 | 4800776 |
| 2011-70-2808-2_R1 | 837228 | 532865 | 1370093 | 61,10738468 | 149 | 48 | | | | | |
| 2011-70-2808-2_R2 | 846759 | 523334 | 1370093 | 61,80303089 | 149 | 48 | 4857450 | 84 | 4987 | 1488 | 4334445 |
| 2012-01-6145-1_R1 | 562022 | 345947 | 907969 | 61,89880932 | 149 | 48 | 4857450 | | | | |
| 2012-01-6145-1_R2 | 557744 | 350225 | 907969 | 61,42764786 | 149 | 48 | 4857450 | 56 | 5092 | 1470 | 4419711 |
| 2012-04-7212-1_R1 | 712433 | 319552 | 1031985 | 69,03520885 | 149 | 49 | 4857450 | | | | |
| 2012-04-7212-1_R2 | 716918 | 315067 | 1031985 | 69,46980819 | 149 | 49 | 4857450 | 63 | 7390 | 1088 | 4572623 |
| 2012-60-2316-31_R1 | 336635 | 106467 | 443102 | 75,97234948 | 301 | 52 | 4857450 | | | | |
| 2012-60-2316-31_R2 | 341665 | 101437 | 443102 | 77,10752829 | 301 | 52 | 4857450 | 55 | 324913 | 49 | 4925648 |
| 2013-01-5866-2_R1 | 448426 | 92203 | 540629 | 82,94523601 | 301 | 51 | | | | | |
| 2013-01-5866-2_R2 | 457559 | 83070 | 540629 | 84,63456455 | 301 | 52 | 4857450 | 67 | 376326 | 48 | 4805948 |
| 2013-40-7492-9_R1 | 892359 | 435834 | 1328193 | 67,18594361 | 149 | 50 | 4857450 | | | | |
| 2013-40-7492-9_R2 | 903051 | 425142 | 1328193 | 67,9909471 | 149 | 50 | 4857450 | 81 | 18854 | 547 | 4845397 |
| 2013-60-2782_R1 | 557564 | 147831 | 705395 | 79,0428058 | 301 | 51 | 4857450 | | | | |
| 2013-60-2782_R2 | 567051 | 138344 | 705395 | 80,38772603 | 301 | 51 | 4857450 | 87 | 338753 | 49 | 4805167 |
| 2014-04-2493-1_R1 | 1157157 | 743450 | 1900607 | 60,88354931 | 149 | 50 | 4857450 | | | | |
| 2014-04-2493-1_R2 | 1161999 | 738608 | 1900607 | 61,13831002 | 149 | 50 | 4857450 | 117 | 23323 | 436 | 4758382 |
| 2014-04-6297_R1 | 475879 | 108140 | 584019 | 81,48347913 | 301 | 51 | 4857450 | | | | |
| 2014-04-6297_R2 | 489241 | 94778 | 584019 | 83,7714184 | 301 | 52 | 4857450 | 72 | 303403 | 52 | 4806861 |
| 2014-40-11112_R1 | 607159 | 175389 | 782548 | 77,58744512 | 301 | 51 | 4857450 | | | | |
| 2014-40-11112_R2 | 622195 | 160353 | 782548 | 79,5088608 | 301 | 51 | 4857450 | 97 | 376326 | 52 | 4806612 |
| 2014-40-15837_R1 | 539565 | 144541 | 684106 | 78,87154915 | 301 | 51 | 4857450 | | | | |
| 2014-40-15837_R2 | 550430 | 133676 | 684106 | 80,45975331 | 301 | 51 | 4857450 | 85 | 313285 | 55 | 5049427 |
| 2015-01-2549-1_R1 | 504861 | 120749 | 625610 | 80,69899778 | 301 | 51 | 4857450 | | | | |
| 2015-01-2549-1_R2 | 513211 | 112399 | 625610 | 82,03369511 | 301 | 52 | 4857450 | 78 | 376326 | 51 | 4807372 |
| 2015-01-2739_R1 | 424796 | 80969 | 505765 | 83,99078623 | 301 | 51 | 4857450 | | | | |
| 2015-01-2739_R2 | 429715 | 76050 | 505765 | 84,96337232 | 301 | 51 | 4857450 | 63 | 316044 | 57 | 5006421 |
| 2015-01-4146-1_R1 | 508600 | 112988 | 621588 | 81,82268641 | 301 | 51 | 4857450 | | | | |
| 2015-01-4146-1_R2 | 524135 | 97453 | 621588 | 84,32193028 | 301 | 51 | 4857450 | 77 | 225886 | 53 | 4947918 |
| 2015-04-8860-1_R1 | 735610 | 556227 | 1291837 | 56,94294249 | 149 | 46 | 4857450 | | | | |
| 2015-04-8860-1_R2 | 748492 | 543345 | 1291837 | 57,94012712 | 149 | 47 | 4857450 | 79 | 2558 | 1975 | 3710760 |
| 2015-40-13064-1-6_R1 | 526091 | 130252 | 656343 | 80,15488853 | 301 | 51 | 4857450 | | | | |
| 2015-40-13064-1-6_R2 | 533694 | 122649 | 656343 | 81,31327675 | 301 | 51 | 4857450 | 81 | 376329 | 53 | 4806460 |
| 2016-02-351-2_R1 | 788896 | 531453 | 1320349 | 59,7490512 | 149 | 47 | 4857450 | | | | |
| 2016-02-351-2_R2 | 795633 | 524716 | 1320349 | 60,25929508 | 149 | 47 | 4857450 | 81 | 3150 | 1843 | 4069577 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 2016-04-1737-1-tarm_R1 | 404598 | 73654 | 478252 | 84,59933257 | 301 | 51 | 4857450 | | | | |
| 2016-04-1737-1-tarm_R2 | 404414 | 73838 | 478252 | 84,56085913 | 301 | 51 | 4857450 | 59 | 376326 | 59 | 4806584 |
| 2016-04-17630-5_R1 | 692290 | 342118 | 1034408 | 66,9262032 | 149 | 48 | 4857450 | | | | |
| 2016-04-17630-5_R2 | 695954 | 338454 | 1034408 | 67,28041546 | 149 | 48 | 4857450 | 63 | 5152 | 1474 | 4396371 |
| 2016-04-3959-1_R1 | 636595 | 320237 | 956832 | 66,53153323 | 149 | 48 | 4857450 | | | | |
| 2016-04-3959-1_R2 | 642593 | 314239 | 956832 | 67,15839353 | 149 | 48 | 4857450 | 59 | 4157 | 1616 | 4158491 |
| 2016-04-4931-1_R1 | 687903 | 490254 | 1178157 | 58,38805864 | 149 | 48 | 4857450 | | | | |
| 2016-04-4931-1_R2 | 695871 | 482286 | 1178157 | 59,06436918 | 149 | 48 | 4857450 | 72 | 3856 | 1700 | 4156388 |
| 2016-40-16685_R1 | 513108 | 124246 | 637354 | 80,50596686 | 301 | 51 | 4857450 | | | | |
| 2016-40-16685_R2 | 535639 | 101715 | 637354 | 84,04105097 | 301 | 51 | 4857450 | 79 | 376326 | 52 | 4806889 |
| 2016-40-6963_R1 | 489203 | 115584 | 604787 | 80,88847809 | 301 | 51 | 4857450 | | | | |
| 2016-40-6963_R2 | 503645 | 101142 | 604787 | 83,27642625 | 301 | 52 | 4857450 | 75 | 325071 | 46 | 4809099 |
| 2017-04-21988-1_R1 | 683544 | 284013 | 967557 | 70,64638052 | 149 | 49 | 4857450 | | | | |
| 2017-04-21988-1_R2 | 678996 | 288561 | 967557 | 70,1763307 | 149 | 49 | 4857450 | 59 | 7165 | 1149 | 4665938 |
| 2017-04-7776_R1 | 538906 | 132936 | 671842 | 80,21320489 | 301 | 51 | 4857450 | | | | |
| 2017-04-7776_R2 | 546737 | 125105 | 671842 | 81,37880633 | 301 | 51 | 4857450 | 83 | 341394 | 49 | 4806360 |
| 2017-40-10600-1_R1 | 475568 | 104893 | 580461 | 81,92936304 | 301 | 51 | 4857450 | | | | |
| 2017-40-10600-1_R2 | 484048 | 96413 | 580461 | 83,39027084 | 301 | 51 | 4857450 | 72 | 376602 | 46 | 4804970 |
| 2017-60-1003-1_R1 | 700490 | 326547 | 1027037 | 68,20494296 | 149 | 50 | 4857450 | | | | |
| 2017-60-1003-1_R2 | 706200 | 320837 | 1027037 | 68,76091124 | 149 | 50 | 4857450 | 63 | 14297 | 680 | 4681325 |
| 2018-01-1983-10_R1 | 572600 | 170221 | 742821 | 77,08451969 | 301 | 51 | 4857450 | | | | |
| 2018-01-1983-10_R2 | 583191 | 159630 | 742821 | 78,5103006 | 301 | 52 | 4857450 | 92 | 277466 | 53 | 4957420 |
| 2018-01-3742-8_R1 | 476769 | 114367 | 591136 | 80,65301386 | 301 | 51 | 4857450 | | | | |
| 2018-01-3742-8_R2 | 484703 | 106433 | 591136 | 81,99517539 | 301 | 52 | 4857450 | 73 | 277466 | 53 | 4955869 |
| 2018-06-1981_R1 | 419763 | 78749 | 498512 | 84,20318869 | 301 | 51 | 4857450 | | | | |
| 2018-06-1981_R2 | 426595 | 71917 | 498512 | 85,57366723 | 301 | 52 | 4857450 | 62 | 376326 | 52 | 4807697 |
| 2018-40-12050-5_R1 | 391919 | 77730 | 469649 | 83,44934196 | 301 | 51 | 4857450 | | | | |
| 2018-40-12050-5_R2 | 398310 | 71339 | 469649 | 84,81014545 | 301 | 51 | 4857450 | 58 | 259480 | 57 | 4843665 |
| 2018-40-21734-17_R1 | 565615 | 144181 | 709796 | 79,68698048 | 301 | 51 | 4857450 | | | | |
| 2018-40-21734-17_R2 | 561482 | 148314 | 709796 | 79,10470051 | 301 | 51 | 4857450 | 88 | 376327 | 50 | 4806788 |
| 2019-01-1106-1-1-1-1-1_R1 | 363934 | 55841 | 419775 | 86,69739742 | 301 | 50 | 4857450 | | | | |
| 2019-01-1106-1-1-1-1-1_R2 | 365971 | 53804 | 419775 | 87,18265738 | 301 | 50 | 4857450 | 52 | 4564 | 1222 | 4203552 |
| 2020-01-1008-1-1-1-1_R1 | 533121 | 122635 | 655756 | 81,29868427 | 301 | 51 | 4857450 | | | | |
| 2020-01-1008-1-1-1-1_R2 | 545937 | 109819 | 655756 | 83,25306974 | 301 | 51 | 4857450 | 81 | 325064 | 49 | 4806674 |
| 2020-01-1354_R1 | 419372 | 93967 | 513339 | 81,69494233 | 301 | 51 | 4857450 | | | | |
| 2020-01-1354_R2 | 425150 | 88189 | 513339 | 82,82051432 | 301 | 51 | 4857450 | 64 | 61 | 222665 | 4981381 |
| 2020-01-3548_R1 | 408551 | 76064 | 484615 | 84,30424151 | 301 | 51 | 4857450 | | | | |
| 2020-01-3548_R2 | 418930 | 65685 | 484615 | 86,44594162 | 301 | 52 | 4857450 | 60 | 363573 | 51 | 4807662 |
| 2020-01-597_R1 | 572514 | 151063 | 723577 | 79,12274713 | 301 | 51 | 4857450 | | | | |
| 2020-01-597_R2 | 589858 | 133719 | 723577 | 81,51972769 | 301 | 52 | 4857450 | 90 | 376326 | 53 | 4807338 |
| 2020-01-599_R1 | 421362 | 80575 | 501937 | 83,94718859 | 301 | 51 | 4857450 | | | | |
| 2020-01-599_R2 | 437335 | 64602 | 501937 | 87,12946047 | 301 | 52 | 4857450 | 62 | 376326 | 51 | 4803446 |
| 2020-01-600-1-1-3-1_R1 | 562436 | 135396 | 697832 | 80,59762235 | 301 | 51 | 4857450 | | | | |
| 2020-01-600-1-1-3-1_R2 | 585017 | 112815 | 697832 | 83,83350147 | 301 | 51 | 4857450 | 86 | 325064 | 50 | 4807028 |
| 2020-01-602_R1 | 558932 | 142744 | 701676 | 79,65670765 | 301 | 51 | 4857450 | | | | |
| 2020-01-602_R2 | 572458 | 129218 | 701676 | 81,58437798 | 301 | 51 | 4857450 | 87 | 376326 | 50 | 4806354 |
| 2020-01-618_R1 | 467070 | 99192 | 566262 | 82,48302023 | 301 | 51 | 4857450 | | | | |
| 2020-01-618_R2 | 478134 | 88128 | 566262 | 84,43688611 | 301 | 52 | 4857450 | 70 | 376606 | 47 | 4807914 |
| 2020-01-640_R1 | 434069 | 83595 | 517664 | 83,85149441 | 301 | 51 | 4857450 | | | | |
| 2020-01-640_R2 | 443459 | 74205 | 517664 | 85,66541231 | 301 | 52 | 4857450 | 64 | 376326 | 52 | 4805860 |
| 2020-01-641_R1 | 343043 | 49852 | 392895 | 87,3116227 | 301 | 51 | 4857450 | | | | |
| 2020-01-641_R2 | 345378 | 47517 | 392895 | 87,90592907 | 301 | 52 | 4857450 | 49 | 305377 | 50 | 4805298 |
| 2020-01-648_R1 | 445693 | 87698 | 533391 | 83,55840275 | 301 | 51 | 4857450 | | | | |
| 2020-01-648_R2 | 456929 | 76462 | 533391 | 85,66492498 | 301 | 52 | 4857450 | 66 | 376326 | 54 | 4805973 |
| 2020-01-653_R1 | 379634 | 64784 | 444418 | 85,42273265 | 301 | 51 | 4857450 | | | | |
| 2020-01-653_R2 | 388392 | 56026 | 444418 | 87,39339991 | 301 | 52 | 4857450 | 55 | 376581 | 45 | 4807490 |
| 2020-01-665_R1 | 368149 | 61761 | 429910 | 85,6339699 | 301 | 52 | 4857450 | | | | |
| 2020-01-665_R2 | 378929 | 50981 | 429910 | 88,14147147 | 301 | 52 | 4857450 | 53 | 376326 | 49 | 4806643 |
| 2020-01-681_R1 | 462537 | 95741 | 558278 | 82,85065863 | 301 | 51 | 4857450 | | | | |
| 2020-01-681_R2 | 472266 | 86012 | 558278 | 84,5933388 | 301 | 52 | 4857450 | 69 | 376326 | 48 | 4806242 |
| 2020-01-686_R1 | 448501 | 92631 | 541132 | 82,88199552 | 301 | 52 | 4857450 | | | | |
| 2020-01-686_R2 | 460899 | 80233 | 541132 | 85,17311857 | 301 | 52 | 4857450 | 67 | 376603 | 47 | 4808350 |
| 2020-01-688_R1 | 416038 | 76517 | 492555 | 84,46528814 | 301 | 51 | 4857450 | | | | |
| 2020-01-688_R2 | 422132 | 70423 | 492555 | 85,70251038 | 301 | 52 | 4857450 | 61 | 248954 | 52 | 4807289 |
| 2020-01-695-1-1-1-1_R1 | 577283 | 147356 | 724639 | 79,66490901 | 301 | 51 | 4857450 | | | | |
| 2020-01-695-1-1-1-1_R2 | 593456 | 131183 | 724639 | 81,89677895 | 301 | 51 | 4857450 | 90 | 325064 | 50 | 4807756 |
| 2020-01-696_R1 | 280581 | 34391 | 314972 | 89,08125167 | 301 | 52 | 4857450 | | | | |
| 2020-01-696_R2 | 285730 | 29242 | 314972 | 90,71600015 | 301 | 52 | 4857450 | 39 | 325072 | 48 | 4808764 |
| 2020-01-697-1-1-1-1_R1 | 530838 | 124004 | 654842 | 81,06352372 | 301 | 51 | 4857450 | | | | |
| 2020-01-697-1-1-1-1_R2 | 543415 | 111427 | 654842 | 82,98413969 | 301 | 52 | 4857450 | 81 | 325072 | 45 | 4809015 |
| 2020-01-713_R1 | 442462 | 87889 | 530351 | 83,42814476 | 301 | 51 | 4857450 | | | | |
| 2020-01-713_R2 | 452262 | 78089 | 530351 | 85,27597761 | 301 | 52 | 4857450 | 66 | 303403 | 51 | 4803619 |
| 2020-01-718_R1 | 544826 | 135436 | 680262 | 80,09061215 | 301 | 51 | 4857450 | | | | |
| 2020-01-718_R2 | 556428 | 123834 | 680262 | 81,79613149 | 301 | 52 | 4857450 | 84 | 376326 | 47 | 4805164 |
| 2020-01-722_R1 | 553604 | 139042 | 692646 | 79,92596507 | 301 | 51 | 4857450 | | | | |
| 2020-01-722_R2 | 563679 | 128967 | 692646 | 81,38053205 | 301 | 51 | 4857450 | 86 | 303403 | 50 | 4805733 |
| 2020-01-750_R1 | 379700 | 62663 | 442363 | 85,83448435 | 301 | 51 | 4857450 | | | | |
| 2020-01-750_R2 | 382697 | 59666 | 442363 | 86,51198224 | 301 | 52 | 4857450 | 55 | 324559 | 49 | 4805916 |
| 2020-01-756_R1 | 392179 | 54950 | 447129 | 87,71048176 | 301 | 51 | 4857450 | | | | |
| 2020-01-756_R2 | 393013 | 54116 | 447129 | 87,89700511 | 301 | 52 | 4857450 | 55 | 341394 | 419 | 5339510 |
| 2020-01-758_R1 | 488878 | 106064 | 594942 | 82,17237983 | 301 | 51 | 4857450 | | | | |
| 2020-01-758_R2 | 499813 | 95129 | 594942 | 84,01037412 | 301 | 52 | 4857450 | 74 | 376326 | 52 | 4806517 |
| 2020-01-776_R1 | 539229 | 131233 | 670462 | 80,42648204 | 301 | 51 | 4857450 | | | | |
| 2020-01-776_R2 | 550216 | 120246 | 670462 | 82,0652028 | 301 | 51 | 4857450 | 83 | 338753 | 53 | 4805954 |
| 2020-01-847-1-1-1-1_R1 | 555366 | 132707 | 688073 | 80,71323828 | 301 | 51 | 4857450 | | | | |
| 2020-01-847-1-1-1-1_R2 | 575540 | 112533 | 688073 | 83,64519462 | 301 | 52 | 4857450 | 85 | 325071 | 44 | 4767992 |
| 2020-01-998-1-1-1-1_R1 | 538549 | 125463 | 664012 | 81,10531135 | 301 | 51 | 4857450 | | | | |
| 2020-01-998-1-1-1-1_R2 | 549750 | 114262 | 664012 | 82,79217845 | 301 | 51 | 4857450 | 82 | 325073 | 49 | 4808955 |
| 2020-04-4664_R1 | 740546 | 253776 | 994322 | 74,47748315 | 301 | 51 | 4857450 | | | | |
| 2020-04-4664_R2 | 763340 | 230982 | 994322 | 76,76989949 | 301 | 52 | 4857450 | 123 | 376606 | 48 | 4768002 |
| 2020-04-4961_R1 | 489000 | 109833 | 598833 | 81,65882642 | 301 | 51 | 4857450 | | | | |
| 2020-04-4961_R2 | 498923 | 99910 | 598833 | 83,31588273 | 301 | 52 | 4857450 | 74 | 338753 | 52 | 4805925 |
| 2020-04-4976_R1 | 513268 | 117615 | 630883 | 81,35708206 | 301 | 51 | 4857450 | | | | |
| 2020-04-4976_R2 | 525945 | 104938 | 630883 | 83,36648792 | 301 | 52 | 4857450 | 78 | 376326 | 52 | 4806816 |
| 2020-04-5027_R1 | 559911 | 143502 | 703413 | 79,59918284 | 301 | 51 | 4857450 | | | | |
| 2020-04-5027_R2 | 573954 | 129459 | 703413 | 81,59559178 | 301 | 51 | 4857450 | 87 | 376326 | 51 | 4807142 |
| 2020-04-5069-1_R1 | 718573 | 500781 | 1219354 | 58,93063048 | 149 | 47 | 4857450 | | | | |
| 2020-04-5069-1_R2 | 724139 | 495215 | 1219354 | 59,38710169 | 149 | 47 | 4857450 | 75 | 3071 | 1843 | 3965129 |
| 2020-04-5964-1_R1 | 577236 | 240787 | 818023 | 70,56476407 | 149 | 48 | 4857450 | | | | |
| 2020-04-5964-1_R2 | 571244 | 246779 | 818023 | 69,83226633 | 149 | 48 | 4857450 | 50 | 3749 | 1692 | 4113449 |
| 2020-06-3288_R1 | 438264 | 83873 | 522137 | 83,93659135 | 301 | 51 | 4857450 | | | | |
| 2020-06-3288_R2 | 447893 | 74244 | 522137 | 85,78074337 | 301 | 51 | 4857450 | 65 | 376603 | 43 | 4767567 |
| 2020-06-4857_R1 | 431200 | 82259 | 513459 | 83,9794414 | 301 | 51 | 4857450 | | | | |
| 2020-06-4857_R2 | 438830 | 74629 | 513459 | 85,46544125 | 301 | 52 | 4857450 | 64 | 341394 | 55 | 4808326 |
| 2020-06-5647_R1 | 419243 | 80566 | 499809 | 83,88064241 | 301 | 51 | 4857450 | | | | |
| 2020-06-5647_R2 | 427211 | 72598 | 499809 | 85,47485139 | 301 | 52 | 4857450 | 62 | 376326 | 51 | 4765785 |
| 2020-06-6905_R1 | 491621 | 95716 | 587337 | 83,70339345 | 301 | 50 | 4857450 | | | | |
| 2020-06-6905_R2 | 496074 | 91263 | 587337 | 84,46156125 | 301 | 52 | 4857450 | 73 | 35742 | 249 | 4801304 |
| 2020-40-4532-1-2-1-1_R1 | 528923 | 117602 | 646525 | 81,81013882 | 301 | 51 | 4857450 | | | | |
| 2020-40-4532-1-2-1-1_R2 | 544757 | 101768 | 646525 | 84,25923205 | 301 | 51 | 4857450 | 80 | 325064 | 44 | 4964540 |
| 2020-40-5378-1_R1 | 643204 | 357998 | 1001202 | 64,2431797 | 149 | 47 | 4857450 | | | | |
| 2020-40-5378-1_R2 | 649907 | 351295 | 1001202 | 64,91267496 | 149 | 47 | 4857450 | 61 | 3007 | 1928 | 4031966 |
| 2021-01-1675-2_R1 | 389568 | 61213 | 450781 | 86,42067878 | 301 | 51 | 4857450 | | | | |
| 2021-01-1675-2_R2 | 396626 | 54155 | 450781 | 87,98640582 | 301 | 51 | 4857450 | 56 | 282781 | 51 | 4927381 |
| 2021-01-2331-1_R1 | 421784 | 71780 | 493564 | 85,45679993 | 301 | 51 | 4857450 | | | | |
| 2021-01-2331-1_R2 | 426104 | 67460 | 493564 | 86,33206636 | 301 | 52 | 4857450 | 61 | 325071 | 45 | 4808937 |
| 2021-01-3671-1_R1 | 543886 | 162602 | 706488 | 76,984464 | 301 | 52 | 4857450 | | | | |
| 2021-01-3671-1_R2 | 541973 | 164515 | 706488 | 76,71368799 | 301 | 52 | 4857450 | 88 | 272152 | 54 | 4927921 |
| 2021-01-3691-1_R1 | 643068 | 171449 | 814517 | 78,95083835 | 301 | 51 | 4857450 | | | | |
| 2021-01-3691-1_R2 | 629239 | 185278 | 814517 | 77,25302234 | 301 | 51 | 4857450 | 101 | 222895 | 62 | 4765778 |
| 2021-06-12535-3_R1 | 480182 | 107678 | 587860 | 81,68305379 | 301 | 51 | 4857450 | | | | |
| 2021-06-12535-3_R2 | 480988 | 106872 | 587860 | 81,82016126 | 301 | 51 | 4857450 | 73 | 5898 | 1323 | 4417297 |
| 2021-06-1423-3_R1 | 542513 | 142632 | 685145 | 79,1822169 | 301 | 52 | 4857450 | | | | |
| 2021-06-1423-3_R2 | 552935 | 132210 | 685145 | 80,70335476 | 301 | 52 | 4857450 | 85 | 266894 | 37 | 4730519 |
| 2006-50-3052_R1 | 264037 | 30519 | 294556 | 89,63898206 | 301 | 51 | 4857450 | | | | |
| 2006-50-3052_R2 | 265221 | 29335 | 294556 | 90,04094298 | 301 | 51 | 4857450 | 37 | 197553 | 65 | 4847416 |
| 2020-01-3961_R1 | 624165 | 175276 | 799441 | 78,07518003 | 301 | 51 | 4857450 | | | | |
| 2020-01-3961_R2 | 661720 | 137721 | 799441 | 82,77283752 | 301 | 52 | 4857450 | 99 | 376326 | 50 | 4806380 |

# SNP distanse matrise for ST19 kluster 1

**Tabell 9:** SNP distanse matrise for ST19 kluster 1 produsert i Galaxy. Matrisen er delt inn i to deler grunnet mangel på plass.

| snp-dists 0.8.2 | 2013-01-5866- | 2013-60-278 | 2014-04-629 | 2014-40-1111 | 2015-01-2549- | 2016-04-1737-1-tarr | 2016-40-1668 | 2017-04-777 | 2018-40-21734-1 | 2007-10-274- | 2020-01-597 | 2020-01-60 | 2020-01-64 | 2020-01-641 | 2020-01-648 | 2020-01-665 | 2020-01-681 | 2020-01-718 | 2020-01-722 | 2020-01-750 | 2020-01-758 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2013-01-5866-2 | 0 | 5 | 11 | 7 | 12 | 16 | 15 | 21 | 23 | 24 | 18 | 19 | 17 | 22 | 18 | 19 | 20 | 20 | 21 | 21 | 18 |
| 2013-60-2782 | 5 | 0 | 8 | 4 | 9 | 13 | 12 | 18 | 20 | 21 | 15 | 16 | 14 | 19 | 15 | 16 | 17 | 17 | 18 | 18 | 15 |
| 2014-04-6297 | 11 | 8 | 0 | 10 | 15 | 11 | 10 | 16 | 18 | 27 | 19 | 14 | 20 | 15 | 19 | 22 | 15 | 13 | 16 | 16 | 19 |
| 2014-40-11112 | 7 | 4 | 10 | 0 | 7 | 15 | 14 | 20 | 22 | 23 | 13 | 18 | 12 | 21 | 13 | 14 | 19 | 19 | 20 | 20 | 13 |
| 2015-01-2549-1 | 12 | 9 | 15 | 7 | 0 | 20 | 19 | 25 | 27 | 28 | 18 | 23 | 17 | 26 | 18 | 19 | 24 | 24 | 25 | 25 | 18 |
| 2016-04-1737-1-tarm | 16 | 13 | 11 | 15 | 20 | 0 | 9 | 7 | 9 | 32 | 26 | 13 | 25 | 16 | 26 | 27 | 14 | 14 | 15 | 15 | 26 |
| 2016-40-16685 | 15 | 12 | 10 | 14 | 19 | 9 | 0 | 14 | 16 | 31 | 25 | 12 | 24 | 15 | 25 | 26 | 13 | 13 | 14 | 14 | 25 |
| 2017-04-7776 | 21 | 18 | 16 | 20 | 25 | 7 | 14 | 0 | 6 | 37 | 31 | 18 | 30 | 21 | 31 | 32 | 19 | 19 | 20 | 20 | 31 |
| 2018-40-21734-17 | 23 | 20 | 18 | 22 | 27 | 9 | 16 | 6 | 0 | 39 | 33 | 20 | 32 | 23 | 33 | 34 | 21 | 21 | 22 | 22 | 33 |
| 2007-10-274-3 | 24 | 21 | 27 | 23 | 28 | 32 | 31 | 37 | 39 | 0 | 34 | 35 | 33 | 38 | 34 | 35 | 36 | 36 | 37 | 37 | 34 |
| 2020-01-597 | 18 | 15 | 19 | 13 | 18 | 26 | 25 | 31 | 33 | 34 | 0 | 29 | 1 | 30 | 0 | 3 | 30 | 28 | 31 | 31 | 0 |
| 2020-01-602 | 19 | 16 | 14 | 18 | 23 | 13 | 12 | 18 | 20 | 35 | 29 | 0 | 28 | 7 | 29 | 30 | 1 | 1 | 6 | 6 | 29 |
| 2020-01-640 | 17 | 14 | 20 | 12 | 17 | 25 | 24 | 30 | 32 | 33 | 1 | 28 | 0 | 31 | 1 | 2 | 29 | 29 | 30 | 30 | 1 |
| 2020-01-641 | 22 | 19 | 15 | 21 | 26 | 16 | 15 | 21 | 23 | 38 | 30 | 7 | 31 | 0 | 30 | 33 | 8 | 6 | 9 | 1 | 30 |
| 2020-01-648 | 18 | 15 | 19 | 13 | 18 | 26 | 25 | 31 | 33 | 34 | 0 | 29 | 1 | 30 | 0 | 3 | 30 | 28 | 31 | 31 | 0 |
| 2020-01-665 | 19 | 16 | 22 | 14 | 19 | 27 | 26 | 32 | 34 | 35 | 3 | 30 | 2 | 33 | 3 | 0 | 31 | 31 | 32 | 32 | 3 |
| 2020-01-681 | 20 | 17 | 15 | 19 | 24 | 14 | 13 | 19 | 21 | 36 | 30 | 1 | 29 | 8 | 30 | 31 | 0 | 2 | 7 | 7 | 30 |
| 2020-01-718 | 20 | 17 | 13 | 19 | 24 | 14 | 13 | 19 | 21 | 36 | 28 | 1 | 29 | 6 | 28 | 31 | 2 | 0 | 7 | 7 | 28 |
| 2020-01-722 | 21 | 18 | 16 | 20 | 25 | 15 | 14 | 20 | 22 | 37 | 31 | 6 | 30 | 9 | 31 | 32 | 7 | 7 | 0 | 8 | 31 |
| 2020-01-750 | 21 | 18 | 16 | 20 | 25 | 15 | 14 | 20 | 22 | 37 | 31 | 6 | 30 | 1 | 31 | 32 | 7 | 7 | 8 | 0 | 31 |
| 2020-01-758 | 18 | 15 | 19 | 13 | 18 | 26 | 25 | 31 | 33 | 34 | 0 | 29 | 1 | 30 | 0 | 3 | 30 | 28 | 31 | 31 | 0 |
| 2020-01-776 | 22 | 19 | 15 | 21 | 26 | 16 | 15 | 21 | 23 | 38 | 30 | 7 | 31 | 0 | 30 | 33 | 8 | 6 | 9 | 1 | 30 |
| 2020-04-4961 | 20 | 17 | 15 | 19 | 24 | 14 | 5 | 19 | 21 | 36 | 30 | 17 | 29 | 20 | 30 | 31 | 18 | 18 | 19 | 19 | 30 |
| 2020-04-4976 | 17 | 14 | 20 | 12 | 17 | 25 | 24 | 30 | 32 | 33 | 1 | 28 | 0 | 31 | 1 | 2 | 29 | 29 | 30 | 30 | 1 |
| 2020-04-5027 | 19 | 16 | 14 | 18 | 23 | 13 | 12 | 18 | 20 | 35 | 29 | 0 | 28 | 7 | 29 | 30 | 1 | 1 | 6 | 6 | 29 |
| 2018-06-1981 | 11 | 8 | 12 | 6 | 11 | 19 | 18 | 24 | 26 | 27 | 15 | 22 | 16 | 23 | 15 | 18 | 23 | 21 | 24 | 24 | 15 |
| 2020-01-3548 | 23 | 20 | 18 | 22 | 27 | 17 | 8 | 22 | 24 | 39 | 33 | 20 | 32 | 23 | 33 | 34 | 21 | 21 | 22 | 22 | 33 |
| 2020-06-4857 | 21 | 18 | 14 | 20 | 25 | 21 | 20 | 26 | 28 | 37 | 31 | 24 | 30 | 27 | 31 | 32 | 25 | 25 | 26 | 26 | 31 |
| 2020-06-5647 | 23 | 20 | 16 | 22 | 27 | 23 | 22 | 28 | 30 | 39 | 33 | 26 | 32 | 29 | 33 | 34 | 27 | 27 | 28 | 28 | 33 |
| 2020-01-3961 | 19 | 16 | 14 | 18 | 23 | 13 | 12 | 18 | 20 | 35 | 29 | 0 | 28 | 7 | 29 | 30 | 1 | 1 | 6 | 6 | 29 |
| 2005-10-175 | 24 | 21 | 25 | 23 | 28 | 32 | 31 | 37 | 39 | 4 | 32 | 35 | 33 | 36 | 32 | 35 | 36 | 34 | 37 | 37 | 32 |
| 2006-01-1845 | 26 | 23 | 27 | 25 | 30 | 34 | 33 | 39 | 41 | 6 | 34 | 37 | 35 | 38 | 34 | 37 | 38 | 36 | 39 | 39 | 34 |
| 2006-01-3807-21 | 23 | 20 | 24 | 22 | 27 | 31 | 30 | 36 | 38 | 13 | 31 | 34 | 32 | 35 | 31 | 34 | 35 | 33 | 36 | 36 | 31 |
| 2006-70-3936 | 28 | 25 | 31 | 27 | 32 | 36 | 35 | 41 | 43 | 18 | 38 | 39 | 37 | 42 | 38 | 39 | 40 | 40 | 41 | 41 | 38 |
| 2007-10-341 | 28 | 25 | 31 | 27 | 32 | 36 | 35 | 41 | 43 | 18 | 38 | 39 | 37 | 42 | 38 | 39 | 40 | 40 | 41 | 41 | 38 |
| 2007-60-1234 | 25 | 22 | 28 | 24 | 29 | 33 | 32 | 38 | 40 | 15 | 35 | 36 | 34 | 39 | 35 | 36 | 37 | 37 | 38 | 38 | 35 |
| 2007-70-2403-2 | 27 | 24 | 30 | 26 | 31 | 35 | 34 | 40 | 42 | 17 | 37 | 38 | 36 | 41 | 37 | 38 | 39 | 39 | 40 | 40 | 37 |
| 2020-01-1008-1-1-1-1 | 19 | 16 | 14 | 18 | 23 | 13 | 12 | 18 | 20 | 35 | 29 | 0 | 28 | 7 | 29 | 30 | 1 | 1 | 6 | 6 | 29 |
| 2020-01-600-1-1-3-1 | 19 | 16 | 14 | 18 | 23 | 13 | 12 | 18 | 20 | 35 | 29 | 0 | 28 | 7 | 29 | 30 | 1 | 1 | 6 | 6 | 29 |
| 2020-01-695-1-1-1-1 | 21 | 18 | 16 | 20 | 25 | 15 | 14 | 20 | 22 | 37 | 31 | 6 | 30 | 1 | 31 | 32 | 7 | 7 | 8 | 0 | 31 |
| 2020-40-4532-1-2-1-1 | 22 | 19 | 13 | 21 | 26 | 22 | 21 | 27 | 29 | 38 | 30 | 25 | 31 | 26 | 30 | 33 | 26 | 24 | 27 | 27 | 30 |

| snp-dists 0.8.2 | 2020-01-776 | 2020-04-496 | 2020-04-497 | 2020-04-5027 | 2018-06-1981 | 2020-01-3548 | 2020-06-4857 | 2020-06-5647 | 2020-01-3961 | 2005-10-175 | 2006-01-184 | 2006-01-38 | 2006-70-39 | 2007-10-341 | 2007-60-123 | 2007-70-240 | 2020-01-100 | 2020-01-600 | 2020-01-695 | 2020-40-4532-1-2-1-1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2013-01-5866-2 | 22 | 20 | 17 | 19 | 11 | 23 | 21 | 23 | 19 | 24 | 26 | 23 | 28 | 28 | 25 | 27 | 19 | 19 | 21 | 22 |
| 2013-60-2782 | 19 | 17 | 14 | 16 | 8 | 20 | 18 | 20 | 16 | 21 | 23 | 20 | 25 | 25 | 22 | 24 | 16 | 16 | 18 | 19 |
| 2014-04-6297 | 15 | 15 | 20 | 14 | 12 | 18 | 14 | 16 | 14 | 25 | 27 | 24 | 31 | 31 | 28 | 30 | 14 | 14 | 16 | 13 |
| 2014-40-11112 | 21 | 19 | 12 | 18 | 6 | 22 | 20 | 22 | 18 | 23 | 25 | 22 | 27 | 27 | 24 | 26 | 18 | 18 | 20 | 21 |
| 2015-01-2549-1 | 26 | 24 | 17 | 23 | 11 | 27 | 25 | 27 | 23 | 28 | 30 | 27 | 32 | 32 | 29 | 31 | 23 | 23 | 25 | 26 |
| 2016-04-1737-1-tarm | 16 | 14 | 25 | 13 | 19 | 17 | 21 | 23 | 13 | 32 | 34 | 31 | 36 | 36 | 33 | 35 | 13 | 13 | 15 | 22 |
| 2016-40-16685 | 15 | 5 | 24 | 12 | 18 | 8 | 20 | 22 | 12 | 31 | 33 | 30 | 35 | 35 | 32 | 34 | 12 | 12 | 14 | 21 |
| 2017-04-7776 | 21 | 19 | 30 | 18 | 24 | 22 | 26 | 28 | 18 | 37 | 39 | 36 | 41 | 41 | 38 | 40 | 18 | 18 | 20 | 27 |
| 2018-40-21734-17 | 23 | 21 | 32 | 20 | 26 | 24 | 28 | 30 | 20 | 39 | 41 | 38 | 43 | 43 | 40 | 42 | 20 | 20 | 22 | 29 |
| 2007-10-274-3 | 38 | 36 | 33 | 35 | 27 | 39 | 37 | 39 | 35 | 4 | 6 | 13 | 18 | 18 | 15 | 17 | 35 | 35 | 37 | 38 |
| 2020-01-597 | 30 | 30 | 1 | 29 | 15 | 33 | 31 | 33 | 29 | 32 | 34 | 31 | 38 | 38 | 35 | 37 | 29 | 29 | 31 | 30 |
| 2020-01-602 | 7 | 17 | 28 | 0 | 22 | 20 | 24 | 26 | 0 | 35 | 37 | 34 | 39 | 39 | 36 | 38 | 0 | 0 | 6 | 25 |
| 2020-01-640 | 31 | 29 | 0 | 28 | 16 | 32 | 30 | 32 | 28 | 33 | 35 | 32 | 37 | 37 | 34 | 36 | 28 | 28 | 30 | 31 |
| 2020-01-641 | 0 | 20 | 31 | 7 | 23 | 23 | 27 | 29 | 7 | 36 | 38 | 35 | 42 | 42 | 39 | 41 | 7 | 7 | 1 | 26 |
| 2020-01-648 | 30 | 30 | 1 | 29 | 15 | 33 | 31 | 33 | 29 | 32 | 34 | 31 | 38 | 38 | 35 | 37 | 29 | 29 | 31 | 30 |
| 2020-01-665 | 33 | 31 | 2 | 30 | 18 | 34 | 32 | 34 | 30 | 35 | 37 | 34 | 39 | 39 | 36 | 38 | 30 | 30 | 32 | 33 |
| 2020-01-681 | 8 | 18 | 29 | 1 | 23 | 21 | 25 | 27 | 1 | 36 | 38 | 35 | 40 | 40 | 37 | 39 | 1 | 1 | 7 | 26 |
| 2020-01-718 | 6 | 18 | 29 | 1 | 21 | 21 | 25 | 27 | 1 | 34 | 36 | 33 | 40 | 40 | 37 | 39 | 1 | 1 | 7 | 24 |
| 2020-01-722 | 9 | 19 | 30 | 6 | 24 | 22 | 26 | 28 | 6 | 37 | 39 | 36 | 41 | 41 | 38 | 40 | 6 | 6 | 8 | 27 |
| 2020-01-750 | 1 | 19 | 30 | 6 | 24 | 22 | 26 | 28 | 6 | 37 | 39 | 36 | 41 | 41 | 38 | 40 | 6 | 6 | 0 | 27 |
| 2020-01-758 | 30 | 30 | 1 | 29 | 15 | 33 | 31 | 33 | 29 | 32 | 34 | 31 | 38 | 38 | 35 | 37 | 29 | 29 | 31 | 30 |
| 2020-01-776 | 0 | 20 | 31 | 7 | 23 | 23 | 27 | 29 | 7 | 36 | 38 | 35 | 42 | 42 | 39 | 41 | 7 | 7 | 1 | 26 |
| 2020-04-4961 | 20 | 0 | 29 | 17 | 23 | 11 | 25 | 27 | 17 | 36 | 38 | 35 | 40 | 40 | 37 | 39 | 17 | 17 | 19 | 26 |
| 2020-04-4976 | 31 | 29 | 0 | 28 | 16 | 32 | 30 | 32 | 28 | 33 | 35 | 32 | 37 | 37 | 34 | 36 | 28 | 28 | 30 | 31 |
| 2020-04-5027 | 7 | 17 | 28 | 0 | 22 | 20 | 24 | 26 | 0 | 35 | 37 | 34 | 39 | 39 | 36 | 38 | 0 | 0 | 6 | 25 |
| 2018-06-1981 | 23 | 23 | 16 | 22 | 0 | 26 | 24 | 26 | 22 | 25 | 27 | 24 | 31 | 31 | 28 | 30 | 22 | 22 | 24 | 23 |
| 2020-01-3548 | 23 | 11 | 32 | 20 | 26 | 0 | 28 | 30 | 20 | 39 | 41 | 38 | 43 | 43 | 40 | 42 | 20 | 20 | 22 | 29 |
| 2020-06-4857 | 27 | 25 | 30 | 24 | 24 | 28 | 0 | 12 | 24 | 37 | 39 | 36 | 41 | 41 | 38 | 40 | 24 | 24 | 26 | 25 |
| 2020-06-5647 | 29 | 27 | 32 | 26 | 26 | 30 | 12 | 0 | 26 | 39 | 41 | 38 | 43 | 43 | 40 | 42 | 26 | 26 | 28 | 27 |
| 2020-01-3961 | 7 | 17 | 28 | 0 | 22 | 20 | 24 | 26 | 0 | 35 | 37 | 34 | 39 | 39 | 36 | 38 | 0 | 0 | 6 | 25 |
| 2005-10-175 | 36 | 36 | 33 | 35 | 25 | 39 | 37 | 39 | 35 | 0 | 4 | 11 | 18 | 18 | 15 | 17 | 35 | 35 | 37 | 36 |
| 2006-01-1845 | 38 | 38 | 35 | 37 | 27 | 41 | 39 | 41 | 37 | 4 | 0 | 13 | 20 | 20 | 17 | 19 | 37 | 37 | 39 | 38 |
| 2006-01-3807-21 | 35 | 35 | 32 | 34 | 24 | 38 | 36 | 38 | 34 | 11 | 13 | 0 | 11 | 11 | 8 | 10 | 34 | 34 | 36 | 35 |
| 2006-70-3936 | 42 | 40 | 37 | 39 | 31 | 43 | 41 | 43 | 39 | 18 | 20 | 11 | 0 | 16 | 13 | 11 | 39 | 39 | 41 | 42 |
| 2007-10-341 | 42 | 40 | 37 | 39 | 31 | 43 | 41 | 43 | 39 | 18 | 20 | 11 | 16 | 0 | 5 | 15 | 39 | 39 | 41 | 42 |
| 2007-60-1234 | 39 | 37 | 34 | 36 | 28 | 40 | 38 | 40 | 36 | 15 | 17 | 8 | 13 | 5 | 0 | 12 | 36 | 36 | 38 | 39 |
| 2007-70-2403-2 | 41 | 39 | 36 | 38 | 30 | 42 | 40 | 42 | 38 | 17 | 19 | 10 | 11 | 15 | 12 | 0 | 38 | 38 | 40 | 41 |
| 2020-01-1008-1-1-1-1 | 7 | 17 | 28 | 0 | 22 | 20 | 24 | 26 | 0 | 35 | 37 | 34 | 39 | 39 | 36 | 38 | 0 | 0 | 6 | 25 |
| 2020-01-600-1-1-3-1 | 7 | 17 | 28 | 0 | 22 | 20 | 24 | 26 | 0 | 35 | 37 | 34 | 39 | 39 | 36 | 38 | 0 | 0 | 6 | 25 |
| 2020-01-695-1-1-1-1 | 1 | 19 | 30 | 6 | 24 | 22 | 26 | 28 | 6 | 37 | 39 | 36 | 41 | 41 | 38 | 40 | 6 | 6 | 0 | 27 |
| 2020-40-4532-1-2-1-1 | 26 | 26 | 31 | 25 | 23 | 29 | 25 | 27 | 25 | 36 | 38 | 35 | 42 | 42 | 39 | 41 | 25 | 25 | 27 | 0 |

## SNP distanse matrise for ST19 kluster 2

**Tabell 10:** SNP distanse matrise for ST19 kluster 2 produsert i Galaxy.

| snp-dists 0.8.2 | 2017-40-10600-1 | 2004-40-18968-226 | 2016-40-6963 | 2020-06-3288 | 2021-01-2331-1 | 2021-01-3691-1 | 2020-01-618 | 2020-01-686 | 2020-01-688 | 2020-01-696 | 2020-04-4664 | 2003-01-723 | 2004-01-559 | 2004-10-157 | 2004-50-547 | 2005-10-101 | 2005-10-84 | 2007-10-250 | 2007-10-509-2 | 2010-10-195 | 2020-01-697-1-1-1-1 | 2020-01-847-1-1-1-1 | 2020-01-998-1-1-1-1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2017-40-10600-1** | 0 | 40 | 71 | 76 | 78 | 81 | 57 | 68 | 63 | 66 | 58 | 57 | 97 | 49 | 56 | 49 | 49 | 49 | 45 | 65 | 66 | 77 | 63 |
| **2004-40-18968-226** | 40 | 0 | 55 | 60 | 62 | 65 | 41 | 52 | 47 | 50 | 42 | 41 | 81 | 33 | 40 | 33 | 33 | 33 | 29 | 49 | 50 | 61 | 47 |
| **2016-40-6963** | 71 | 55 | 0 | 5 | 7 | 92 | 68 | 79 | 74 | 77 | 69 | 60 | 108 | 60 | 59 | 60 | 60 | 60 | 56 | 68 | 77 | 6 | 74 |
| **2020-06-3288** | 76 | 60 | 5 | 0 | 10 | 97 | 73 | 84 | 79 | 82 | 74 | 65 | 113 | 65 | 64 | 65 | 65 | 65 | 61 | 73 | 82 | 9 | 79 |
| **2021-01-2331-1** | 78 | 62 | 7 | 10 | 0 | 99 | 75 | 86 | 81 | 84 | 76 | 67 | 115 | 67 | 66 | 67 | 67 | 67 | 63 | 75 | 84 | 11 | 81 |
| **2021-01-3691-1** | 81 | 65 | 92 | 97 | 99 | 0 | 76 | 61 | 32 | 59 | 77 | 78 | 108 | 60 | 77 | 52 | 60 | 52 | 38 | 86 | 59 | 98 | 66 |
| **2020-01-618** | 57 | 41 | 68 | 73 | 75 | 76 | 0 | 63 | 58 | 61 | 3 | 54 | 92 | 44 | 53 | 44 | 44 | 44 | 40 | 62 | 61 | 74 | 58 |
| **2020-01-686** | 68 | 52 | 79 | 84 | 86 | 61 | 63 | 0 | 43 | 2 | 64 | 65 | 95 | 47 | 64 | 39 | 47 | 39 | 25 | 73 | 2 | 85 | 53 |
| **2020-01-688** | 63 | 47 | 74 | 79 | 81 | 32 | 58 | 43 | 0 | 41 | 59 | 60 | 90 | 42 | 59 | 34 | 42 | 34 | 20 | 68 | 41 | 80 | 48 |
| **2020-01-696** | 66 | 50 | 77 | 82 | 84 | 59 | 61 | 2 | 41 | 0 | 62 | 63 | 93 | 45 | 62 | 37 | 45 | 37 | 23 | 71 | 0 | 83 | 51 |
| **2020-04-4664** | 58 | 42 | 69 | 74 | 76 | 77 | 3 | 64 | 59 | 62 | 0 | 55 | 93 | 45 | 54 | 45 | 45 | 45 | 41 | 63 | 62 | 75 | 59 |
| **2003-01-723** | 57 | 41 | 60 | 65 | 67 | 78 | 54 | 65 | 60 | 63 | 55 | 0 | 94 | 46 | 23 | 46 | 46 | 46 | 42 | 32 | 63 | 66 | 60 |
| **2004-01-559** | 97 | 81 | 108 | 113 | 115 | 108 | 92 | 95 | 90 | 93 | 93 | 94 | 0 | 54 | 93 | 76 | 68 | 76 | 72 | 102 | 93 | 114 | 90 |
| **2004-10-157** | 49 | 33 | 60 | 65 | 67 | 60 | 44 | 47 | 42 | 45 | 45 | 46 | 54 | 0 | 45 | 28 | 20 | 28 | 24 | 54 | 45 | 66 | 42 |
| **2004-50-547** | 56 | 40 | 59 | 64 | 66 | 77 | 53 | 64 | 59 | 62 | 54 | 23 | 93 | 45 | 0 | 45 | 45 | 45 | 41 | 19 | 62 | 65 | 59 |
| **2005-10-101** | 49 | 33 | 60 | 65 | 67 | 52 | 44 | 39 | 34 | 37 | 45 | 46 | 76 | 28 | 45 | 0 | 28 | 10 | 16 | 54 | 37 | 66 | 34 |
| **2005-10-84** | 49 | 33 | 60 | 65 | 67 | 60 | 44 | 47 | 42 | 45 | 45 | 46 | 68 | 20 | 45 | 28 | 0 | 28 | 24 | 54 | 45 | 66 | 42 |
| **2007-10-250** | 49 | 33 | 60 | 65 | 67 | 52 | 44 | 39 | 34 | 37 | 45 | 46 | 76 | 28 | 45 | 10 | 28 | 0 | 16 | 54 | 37 | 66 | 34 |
| **2007-10-509-2** | 45 | 29 | 56 | 61 | 63 | 38 | 40 | 25 | 20 | 23 | 41 | 42 | 72 | 24 | 41 | 16 | 24 | 16 | 0 | 50 | 23 | 62 | 30 |
| **2010-10-195** | 65 | 49 | 68 | 73 | 75 | 86 | 62 | 73 | 68 | 71 | 63 | 32 | 102 | 54 | 19 | 54 | 54 | 54 | 50 | 0 | 71 | 74 | 68 |
| **2020-01-697-1-1-1-1** | 66 | 50 | 77 | 82 | 84 | 59 | 61 | 2 | 41 | 0 | 62 | 63 | 93 | 45 | 62 | 37 | 45 | 37 | 23 | 71 | 0 | 83 | 51 |
| **2020-01-847-1-1-1-1** | 77 | 61 | 6 | 9 | 11 | 98 | 74 | 85 | 80 | 83 | 75 | 66 | 114 | 66 | 65 | 66 | 66 | 66 | 62 | 74 | 83 | 0 | 80 |
| **2020-01-998-1-1-1-1** | 63 | 47 | 74 | 79 | 81 | 66 | 58 | 53 | 48 | 51 | 59 | 60 | 90 | 42 | 59 | 34 | 42 | 34 | 30 | 68 | 51 | 80 | 0 |

## SNP distanse matrise for ST34

**Tabell 11:** SNP distanse matrise for ST34 produsert i Galaxy.

| | dataset_0848a134-2 | 2011-60-6693 | 2015-01-2739 | 2021-01-1675-2 | 2018-01-1983-1 | 2007-01-1883 | 2007-01-3479-18 |
|---|---|---|---|---|---|---|---|
| **dataset_0848a13** | 0 | 103 | 94 | 80 | 99 | 97 | 99 |
| **2011-60-6693** | 103 | 0 | 59 | 65 | 75 | 52 | 50 |
| **2015-01-2739** | 94 | 59 | 0 | 62 | 68 | 47 | 55 |
| **2021-01-1675-2** | 80 | 65 | 62 | 0 | 64 | 65 | 61 |
| **2018-01-1983-10** | 99 | 75 | 68 | 64 | 0 | 69 | 73 |
| **2007-01-1883** | 97 | 52 | 47 | 65 | 69 | 0 | 12 |
| **2007-01-3479-18** | 99 | 50 | 55 | 61 | 73 | 12 | 0 |

# ALPACCA Core Genome Report ST19

**Tabell 12:** Resultater fra ParSNP ved bruk av ALPACCA utført av Veterinærinstituttet. Tabellen viser oversikt over hvor mye av genomet som ble dekket for alle ST19 isolater.

## ParSNP

*ParSNP average coverage report. The "Type" column refers to which of the samples were used as a reference in the analysis. The "Percent Coverage" column represents the % coverage of each genome that was used to generate the core genome alignment. The green bar is relative to 100%. The total coverage is the average across all samples. Samples with expected high similarity (e.g. outbreak or same ST) are expected to have a high percent coverage.*

Code

| Sequence | Sample | Type | Percent Coverage |
|---|---|---|---|
| Sequence 1 | 2016-04-1737-1-tarm.fasta.ref | Reference | 93.8 |
| Sequence 2 | 2001-01-2236.fasta | Sample | 91.5 |
| Sequence 3 | 2002-01-22-63.fasta | Sample | 90.5 |
| Sequence 4 | 2002-01-518.fasta | Sample | 93.6 |
| Sequence 5 | 2003-01-723.fasta | Sample | 94.5 |
| Sequence 6 | 2003-40-20444.fasta | Sample | 95.2 |
| Sequence 7 | 2003-50-2805.fasta | Sample | 90.1 |
| Sequence 8 | 2004-01-1763-12.fasta | Sample | 95.2 |
| Sequence 9 | 2004-01-559.fasta | Sample | 94.5 |
| Sequence 10 | 2004-10-157.fasta | Sample | 94.5 |
| Sequence 11 | 2004-40-18968-226.fasta | Sample | 94.5 |
| Sequence 12 | 2004-50-547.fasta | Sample | 93.7 |
| Sequence 13 | 2005-10-101.fasta | Sample | 93.7 |
| Sequence 14 | 2005-10-175.fasta | Sample | 93.7 |
| Sequence 15 | 2005-10-84.fasta | Sample | 94.5 |
| Sequence 16 | 2006-01-1845.fasta | Sample | 93.7 |
| Sequence 17 | 2006-01-3807-21.fasta | Sample | 93.8 |
| Sequence 18 | 2006-40-8604.fasta | Sample | 91.5 |
| Sequence 19 | 2006-60-6746.fasta | Sample | 93.7 |
| Sequence 20 | 2006-70-1122.fasta | Sample | 94.5 |
| Sequence 21 | 2006-70-2104.fasta | Sample | 94.5 |
| Sequence 22 | 2006-70-3936.fasta | Sample | 93.8 |
| Sequence 23 | 2007-10-250.fasta | Sample | 93.7 |
| Sequence 24 | 2007-10-274-3.fasta | Sample | 93.7 |
| Sequence 25 | 2007-10-341.fasta | Sample | 93.8 |
| Sequence 26 | 2007-10-509-2.fasta | Sample | 93.7 |
| Sequence 27 | 2007-40-3693.fasta | Sample | 91.8 |
| Sequence 28 | 2007-50-375.fasta | Sample | 94.7 |
| Sequence 29 | 2007-50-378.fasta | Sample | 94.7 |
| Sequence 30 | 2007-60-1234.fasta | Sample | 93.7 |
| Sequence 31 | 2007-70-2403-2.fasta | Sample | 93.7 |
| Sequence 32 | 2008-40-4269.fasta | Sample | 92.7 |
| Sequence 33 | 2009-02-1811.fasta | Sample | 92.7 |
| Sequence 34 | 2010-10-195.fasta | Sample | 93.7 |
| Sequence 35 | 2010-40-247-3.fasta | Sample | 90.7 |
| Sequence 36 | 2010-40-35-4.fasta | Sample | 90.6 |

| Sequence | Sample | Type | Percent Coverage |
|---|---|---|---|
| Sequence 37 | 2011-40-7145-1.fasta | Sample | 91.5 |
| Sequence 38 | 2011-50-2020-41.fasta | Sample | 94.7 |
| Sequence 39 | 2011-50-2222-1.fasta | Sample | 93.9 |
| Sequence 40 | 2012-60-2316-31.fasta | Sample | 91.6 |
| Sequence 41 | 2013-01-5866-2.fasta | Sample | 93.7 |
| Sequence 42 | 2013-60-2782.fasta | Sample | 93.7 |
| Sequence 43 | 2014-04-6297.fasta | Sample | 93.8 |
| Sequence 44 | 2014-40-11112.fasta | Sample | 93.7 |
| Sequence 45 | 2015-01-2549-1.fasta | Sample | 93.7 |
| Sequence 46 | 2015-01-4146-1.fasta | Sample | 91.0 |
| Sequence 47 | 2015-40-13064-1-6.fasta | Sample | 93.8 |
| Sequence 48 | 2016-04-1737-1-tarm.fasta | Sample | 93.8 |
| Sequence 49 | 2016-40-16685.fasta | Sample | 93.7 |
| Sequence 50 | 2016-40-6963.fasta | Sample | 93.7 |
| Sequence 51 | 2017-04-7776.fasta | Sample | 93.7 |
| Sequence 52 | 2017-40-10600-1.fasta | Sample | 93.8 |
| Sequence 53 | 2018-06-1981.fasta | Sample | 93.7 |
| Sequence 54 | 2018-40-12050-5.fasta | Sample | 93.0 |
| Sequence 55 | 2018-40-21734-17.fasta | Sample | 93.8 |
| Sequence 56 | 2020-01-1008-1-1-1-1.fasta | Sample | 93.8 |
| Sequence 57 | 2020-01-3548.fasta | Sample | 93.7 |
| Sequence 58 | 2020-01-3961.fasta | Sample | 93.7 |
| Sequence 59 | 2020-01-597.fasta | Sample | 93.7 |
| Sequence 60 | 2020-01-599.fasta | Sample | 93.8 |
| Sequence 61 | 2020-01-600-1-1-3-1.fasta | Sample | 93.8 |
| Sequence 62 | 2020-01-602.fasta | Sample | 93.8 |
| Sequence 63 | 2020-01-618.fasta | Sample | 93.7 |
| Sequence 64 | 2020-01-640.fasta | Sample | 93.7 |
| Sequence 65 | 2020-01-641.fasta | Sample | 93.7 |
| Sequence 66 | 2020-01-648.fasta | Sample | 93.7 |

| Sequence | Sample | Type | Percent Coverage |
|---|---|---|---|
| Sequence 67 | 2020-01-665.fasta | Sample | 93.8 |
| Sequence 68 | 2020-01-681.fasta | Sample | 93.7 |
| Sequence 69 | 2020-01-686.fasta | Sample | 93.7 |
| Sequence 70 | 2020-01-688.fasta | Sample | 93.7 |
| Sequence 71 | 2020-01-695-1-1-1-1.fasta | Sample | 93.7 |
| Sequence 72 | 2020-01-696.fasta | Sample | 93.7 |
| Sequence 73 | 2020-01-697-1-1-1-1.fasta | Sample | 93.7 |
| Sequence 74 | 2020-01-713.fasta | Sample | 93.8 |
| Sequence 75 | 2020-01-718.fasta | Sample | 93.7 |
| Sequence 76 | 2020-01-722.fasta | Sample | 93.7 |
| Sequence 77 | 2020-01-750.fasta | Sample | 93.7 |
| Sequence 78 | 2020-01-758.fasta | Sample | 93.7 |
| Sequence 79 | 2020-01-776.fasta | Sample | 93.7 |
| Sequence 80 | 2020-01-847-1-1-1-1.fasta | Sample | 94.6 |
| Sequence 81 | 2020-01-998-1-1-1-1.fasta | Sample | 93.7 |
| Sequence 82 | 2020-04-4664.fasta | Sample | 94.5 |
| Sequence 83 | 2020-04-4961.fasta | Sample | 93.7 |
| Sequence 84 | 2020-04-4976.fasta | Sample | 93.7 |
| Sequence 85 | 2020-04-5027.fasta | Sample | 93.7 |
| Sequence 86 | 2020-06-3288.fasta | Sample | 94.6 |
| Sequence 87 | 2020-06-4857.fasta | Sample | 93.7 |
| Sequence 88 | 2020-06-5647.fasta | Sample | 94.5 |
| Sequence 89 | 2020-40-4532-1-2-1-1.fasta | Sample | 90.8 |
| Sequence 90 | 2021-01-2331-1.fasta | Sample | 93.7 |
| Sequence 91 | 2021-01-3671-1.fasta | Sample | 91.5 |
| Sequence 92 | 2021-01-3691-1.fasta | Sample | 94.5 |
| Sequence 93 | 2021-06-1423-3.fasta | Sample | 95.3 |
| **Total coverage** | | | **93.0** |

# ALPACCA Core Genome Report ST34

**Tabell 13:** Resultater fra ParSNP ved bruk av ALPACCA utført av Veterinærinstituttet. Tabellen viser oversikt over hvor mye av genomet som ble dekket for alle ST34 isolater.

## ParSNP

*ParSNP average coverage report. The "Type" column refers to which of the samples were used as a reference in the analysis. The "Percent Coverage" column represents the % coverage of each genome that was used to generate the core genome alignment. The green bar is relative to 100%. The total coverage is the average across all samples. Samples with expected high similarity (e.g. outbreak or same ST) are expected to have a high percent coverage.*

Code

| Sequence | Sample | Type | Percent Coverage |
|---|---|---|---|
| Sequence 1 | 2011-60-6693.fasta.ref | Reference | 89.6 |
| Sequence 2 | 2007-01-1883.fasta | Sample | 89.6 |
| Sequence 3 | 2007-01-3478.fasta | Sample | 89.6 |
| Sequence 4 | 2007-01-3479-18.fasta | Sample | 89.5 |
| Sequence 5 | 2011-60-6693.fasta | Sample | 89.6 |
| Sequence 6 | 2015-01-2739.fasta | Sample | 85.9 |
| Sequence 7 | 2018-01-1983-10.fasta | Sample | 86.8 |
| Sequence 8 | 2021-01-1675-2.fasta | Sample | 87.3 |
| **Total coverage** | | | **88.0** |


**Norges miljø- og biovitenskapelige universitet**
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway