CRANFIELD UNIVERSITY

Matthew G Goulden

Analysis of mecA status of MRSA Staphylococcal Chromosome Cassette from Next Generation Sequence data

Cranfield Health
Applied Bioinformatics

MSc Thesis
Academic Year: 2010 - 2011

Supervisors: Lee Larcombe, T Clark & Conrad Bessant
March 2012-03-30

CRANFIELD UNIVERSITY

Cranfield Health
Applied Bioinformatics

MSc

Academic Year 2010 - 2011

Matthew G Goulden

Analysis of MRSA Staphylococcal Chromosome Cassette mecA status from Next Generation Sequence data

Supervisors: Lee Larcombe, TClark & Conrad Bessant

March 2012

This thesis is submitted in partial fulfilment of the requirements for the degree of MSc Applied Bioinformatics

ABSTRACT

NGS sequencing libraries prepared on an Illumina NGS platform for 10 isolates of *Staphylococcus aureus* were analysed. After extensive pre-processing to address library quality issues, for each isolate the status of the Staphylococcal Chromosome Cassette, and its mecA gene specifying resistance to meticillin, was determined. All mecA-positive isolates encoded canonical mecA. None encoded the new variant mecA identified in strain LGA251.

Keywords:

core genome, accessory genome, drug resistance,

ACKNOWLEDGEMENTS

I thank my supervisors at Cranfield, Lee Larcombe and Conrad Bessant. I thank Taane Clark at LSHTM for making available the *S.aureus* sequence libraries. I thank the BBSRC for financial contributions during this work.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The simple acronym ‘MRSA’, Meticillin-resistant *Staphylococcus aureus,* is unusual within the bacteriological world in having achieved widespread public recognition. Within the UK, no bacterial strain has attracted more mentions in the UK parliament (Hansard), nor been mentioned more frequently in the context of human health by British Newspapers. More globally as a subject of a Google search, ‘MRSA’ cedes only to tuberculosis (Google trends, run 19 Jan 2012), the world’s number one bacteriological killer. And as a subject of genomic sequencing, the NCBI genome project dedicated to *S aureus* includes over 200 genome sequencing projects, with 30 completed, 82 at an advanced contig stage and a further 86 Short Read Archive projects (NCBI_DataBase:154).

Despite the fact that MRSA has been characterised more extensively than any other bacterial species, this thesis presents the analysis using Illumina NGS data of a further 10 isolates of MRSA.

Section 1.1 introduces *S.aureus* and its remarkable genomic plasticity which, by conferring resistance to a broad range of antibiotics, ensures MRSA’s persistence as a virulent pathogen and its place in the microbiological limelight. Section 1.2 introduces the technological platform used to generate the data, Illumina’s NGS platform.

1.1 MRSA

1.1.1 *Staphylococcus aureus*

Staphylococcus aureus is among many bacterial species which can contribute to a normal human microflora with little ill effect. A gram-positive coccus, it typically colonises damp folds (perineal, inguinal, axillae) and also the anterior nares (Kuehnert et al., 2006, von Eiff et al., 2001). The majority of the human population are intermittent carriers, playing but transient host to distinct strains as they are encountered, up to one third are persistent carriers, and a minority appear never to become carriers (*Ibid.*).

Despite its typically benign symbiosis with its human host, *Staphylococcus aureus* is an opportunistic pathogen which causes significant human disease. Damage to carrier's skin, whether through accident, needle/inoculation or surgical intervention, allows bacteria to enter the tissues to cause either local disease (pimples, boils and abscesses), or systemic diseases including pneumonia, septicaemia and toxic shock syndrome. Such systemic diseases were typically fatal in the pre-antibiotic era (Kuehnert et al., 2006).

1.1.2 The antibiotic era

The introduction of systemic penicillin antibiotics in the 1940s lead within a few years to the identification of penicillin-resistant *S.aureus strains*; these isolates expressed an enzyme, β-lactamase, which by hydrolysing the common active center of penicillins rendered them inactive. A decade after the introduction of penicillin, the spread of this resistance had rendered these formerly potent therapeutics largely ineffective against *S.aureus* (de Lencastre et al., 2007).

The search for β-lactamase-resistant alternatives resulted in the development of the semi-synthetic penicillins including meticillin; within just two years of their introduction into clinical practice, meticillin-resistant clinical isolates were identified (*Ibid*.). These isolates expressed an alternative to the native Staphylococcal penicillin binding protein, which by virtue of its low affinity for penicillins could act as an effective surrogate even in the presence of formerly therapeutic levels of semi-synthetic penicillins.

In the decades that followed, drug-resistant *S.aureus* isolates became globally disseminated, initially within hospital environments as HA-MRSA, acquired additional drug-resistance functions & latterly spread into the community as CA-MRSA. Over a 60-year period of antibiotic use *S.aureus* has evolved to become what one author has described as "the spector of a totally resistant bacterial pathogen” (*Ibid.*) & remains the most notorious ‘poster-bug’ of antibiotic resistance.

1.1.3 Virulence & the accessory genome.

Microbial genomes can be described as consisting of core and accessory genomes. The core genome consists of those genetic functions essential for basic cell survival e.g. those required for RNA and DNA synthesis and replication, metabolism and energy systems, and for maintenance and replication of the organism’s physical entity. The accessory genome consists of those functions which confer on the organism adaptability to specific environmental niches, for example antibiotic resistance, or expression of virulence factors (Lindsay and Holden, 2004). The core genome of *S.aureus* has been estimated at ~75% of any specific *S. aureus* strain (*Ibid.*). This core is well conserved, with 98-100% identity at the amino acid level (*Ibid.*). A highly diverse collection of structural elements - plasmids, transposons, genomic islands, chromosome cassettes, bacteriophages, pathogenicity islands - comprise the accessory genome making up the remaining ~25% of any *S.aureus* genome. It is the combinatorial permutation of elements in this accessory genome which determine the pathogenicity and antibiotic resistance of an *S.aureus* isolate. The elements of this toolkit for “bacterial evolution in quantum leaps” are introduced in order of ascending size with the exception of the diversely-sized plasmids, which are introduced first.

1.1.3.1 Plasmids.

Many plasmids have been characterised from *S aureus*. (Markowitz et al., 2012). The Integrated Microbial Genomes resource currently lists 57 plasmids ranging in size between 1,288 and 57 889bp, and encoding between 1 and 58 open reading frames (*Ibid*.). These plasmids may exist as independently replicating episomal forms or as integrated plasmids concomitantly replicated with the genome. Many of these plasmids specify antibiotic resistance genes. The spread of a plasmid-specified penicillinase operon, the *bla* operon, was the cause of the early failure of penicillin therapy against *S. aureus;* the majority of

S.aureus isolates from healthy individuals now carry this marker (Oliveira et al., 2002). Other plasmids specify resistance to a wide range of antibiotics including streptomycin, tetracycline, erythromycin, aminoglycosides, chloramphenicol, trimethoprim, macrolides and vancomycin (Malachowa and DeLeo, 2010). Plasmids may also specify virulence factors including toxins such as Scalded Skin Syndrome-associated exfoliative toxin (*eta*) & food-poisoning-associated enterotoxins (*Ibid*.). Plasmids may also carry other elements of the accessory genome.

1.1.3.2 Insertion Sequences

Insertion sequences are abundant in most bacterial genomes and can insert at random into any locus, including other elements of the accessory genome. IS specify only the functions required for their transposition, implying a limited ability to contribute to adaptation. However, their insertion can change expression or function of adjacent genes of the core or accessory genome. Regulation of the meticillin resistance operon in many MRSA strains is disrupted as a result of the insertion of IS431or IS1272 into the operon’s regulatory genes (Malachowa and DeLeo, 2010). Run-on transcription from IS promoters can also drive adjacent gene expression, as in composite transposons composed of paired IS flanking intervening accessory sequences. The composite transposon Tn4001 specifies an operon conferring resistance to the aminoglycosides gentamicin, tobramycin and kanamycin. Expression of this operon is driven from promoters within the flanking IS256 (Byrne et al., 1989)**.**

1.1.3.3 Transposons

Transposons characteristically carry genes in addition to those required for their transposition. In Staphylococcal transposons, these genes frequently specify additional antibiotic resistance functions, including resistance to penicillins, tetracyclins, chloramphenicol, macrolides and vancomycin (Malachowa and DeLeo, 2010). Such transposons can integrate randomly into the core genome, or other elements of the accessory genome.

1.1.3.4 Pathogenicity Islands

S.aureus pathogenicity islands are a class of helper-phage-mobilised elements which integrate site-specifically into one of 6 conserved *att* sites in the core genome (*Ibid.*). The Pathogenicity Island Database currently lists 33 confirmed and candidate SaPIs (Yoon et al., 2007). In addition to a conserved core region specifying phage mobilization-related functions including an integrase, helicase, and terminase (Novick, 2003), pathogenicity islands generally specify virulence factors such as food-poisoning-associated enterotoxins or the toxin associated with toxic shock syndrome (Malachowa and DeLeo, 2010). Staphylococcal pathogenicity islands range in size between 15kbp and 27kbp (Novick, 2003).

1.1.3.5 Genomic Islands

Much of the literature uses ‘genomic island’ as an umbrella term to encompass all genetic elements most probably acquired by horizontal gene transfer. Such terminology sets ‘genomic islands’ as synonymous with ‘the accessory genome’ as used here. It has been shown by analysis of more than 600 complete bacterial genomes that virulence functions disproportionately associate with such genomic islands (Ho Sui et al., 2009). The Genomic Islands of *Staphylococcus aureus* are vestigial mobile genetic elements, originally acquired by horizontal gene transfer but thought now to be statically integrated into conserved core genome positions owing to a defective transposase gene (Malachowa and DeLeo, 2010). Three types of SaGI have been identified, *v*Saα, *v*Saβ and *v*Sa*γ,* each of which exists in multiple allelic forms and all of which variously specify virulence factors including exfoliative toxins, pore-forming leucocidins & haemolysins, superantigen enterotoxins and pro-inflammatory lipoprotein-like proteins (*Ibid*.). The staphylococcal genomic islands range in size between 20kbp to 46kbp (Gill et al., 2005).

1.1.3.6 Staphylococcal Cassette Chromosomes

A diverse class of mobile genetic elements unique to Staphylococcus, the SCC were initially united by sharing three common features; (1) they integrate into a conserved core genome *att* site at the 3’ end of a ribosomal methyl transferase gene (often referred to as ‘orfX’) near the origin of replication; this integration generates characteristic direct and inverted repeats at each end; (2) they include two site-specific recombinase genes termed ccrA and ccrB involved in element mobility; and (3) they include a mecA operon (IWG-SCC, 2009). The SSC was first characterised as an ‘additional DNA’ element present in a meticillin-resistant strain, N315, but absent from meticillin sensitive strains (Ito et al., 1999). The mecA operon was shown to specify an alternative low affinity penicillin binding protein, PBP2a, which conferred meticillin resistance on its host (*Ibid.*). The SSC was also shown to accommodate other elements of the accessory genome, including a transposon, Tn554, specifying resistance to erythromycin and spectinomycin, and an IS431-flanked insertion of a plasmid, pUB110, specifying resistance to bleomycin and tobramycin (*Ibid*.). Since that first characterisation, many allelic variants of all SSC regions have been identified, with eleven types of SSCmec currently recognised by the International Working Group on SSCmec classification, including pseudoSSCmec elements integrating into the same site but carrying no mec operon (IWG-SCC, 2009). SSC range in size between 24k and 51kbp (*Ibid.*) (Li et al., 2011)

1.1.3.7 Bacteriophages

Bacteriophage are a primary mechanism of horizontal gene transfer between *S. aureus* strains, and mobilise pathogenicity islands at very high frequency (Goerke et al., 2009) (Novick, 2003). Lysogenic bacteriophage integration typically occurs at conserved core genome sites and can result in gene inactivation. The integration site of group Sa6 bacteriophages, present in ~10% of *S. aureus* strains, results in the inactivation of a core virulence gene, glycerol ester hydrolase (*geh* (Lee and Buranen, 1989)). Similarly, the integration of the most frequently detected bacteriophages, of group Sa3, results the inactivation of another core virulence gene, α-hemolysin (*hlb,* (Coleman et al., 1989). Any resultant diminution in virulence is likely to be offset by expression of additional virulence determinants carried by the integrated prophage (*Ibid*.). More than 80 integrative bacteriophages have been characterized for *S. aureus*, and these specify various toxins including superantigens implicated in food poisoning (*sep*, *sek*, *sea*, *seq* (Malachowa and DeLeo, 2010)), inhibiting neutrophil activation, chemotaxis, phagocytosis and bacteriocidal activity (*chip*, *scn*, *sak Ibid.*), causing epidermal exfoliation of the host (*eta,* scalded skin syndrome) or simply lysing host neutrophils, macrophage and monocytes (PVL, Panton Valentine Leucocidin, *Ibid)*

1.1.4 Classification of MRSA

Phenotypic and molecular typing of *S.aureus* isolates is used to guide therapy of infected patients, management of carriers, and also as a tool in the epidemiological surveillance of emerging variants.

1.1.4.1 Phenotypic Testing

The British Society for Antimicrobial Chemotherapy has established standard tests for phenotypic measurements of antibiotic sensitivity (Andrews and Howe, 2011). These tests measure the growth inhibition of a sub-confluent bacterial lawn on agar plates supplemented with antimicrobial-impregnated discs (*Ibid.*). Standardised thresholds of growth inhibition are used to classify strains into sensitive, intermediate and resistant categories (*Ibid.*). Results are frequently confirmed using additional often molecular tests.

1.1.4.2 Pulse Field Gel Electrophoresis

PFGE is one of the most discriminatory tests for epidemiological studies of pathogenic organisms, for which it continues to be described as a ‘Gold Standard’ (Strommenger et al., 2006). An extension of agarose gel electrophoresis, it supplements the standard electrophoretic drive voltage with two others at 120 degrees to either side. By regularly cycling the applied drive voltage through these three directions PFGE can resolve DNA molecules up to 2M bases. Rigorous comparisons with sequence-based classification methods have confirmed its discriminatory power (Hallin et al., 2007). PFGE is however a technically-demanding, labour-intensive approach which is subject to some interpretation subjectivity (*Ibid.*). Despite these challenges the development of standardised protocols has allowed the establishment of a single pan-Europe approach for MRSA typing (Murchan et al., 2003)

1.1.4.3 Spa Typing

Spa-typing is a single locus molecular typing technique targeting a polymorphic region within the Staphylococcus core genome, protein A gene (*spa)*. Flanked by well conserved sequences, the polymorphic *spa* X-region consists of a variable number of 24bp repeats which can be characterised by sequencing or PCR/VNTR approaches (Frenay et al., 1996). It is less labour-intensive than PFGE, reproducibly returns data suitable for inter-laboratory comparison, and has proved useful for studies of both local and global epidemiology (Hallin et al., 2009). As a single-locus assay its discriminatory power is however limited; it is considered best applied together with additional typing methods (*Ibid*.).

1.1.4.4 Multi Locus Sequence Typing

MLST is a molecular typing technique targeting polymorphisms in multiple house-keeping genes within the core genome (Maiden et al., 1998). Application to *S. aureus* using sequencing to target 7 polymorphic genes and has been shown to be highly discriminatory, capable unambiguously of assigning *S.aureus* isolates to known strain types & to generate data suitable for inter-laboratory comparison (Enright et al., 2000). Its dependence on a sequencing platform may limit its accessibility for some laboratories, and some maintain its discriminatory power may limit its epidemiological applicability (Pourcel et al., 2009).

1.1.4.5 Multi Locus Tandem Repeat Typing

Various Multi-Locus Tandem Repeat techniques have been developed, all of which address the limited discriminatory power of single-locus Spa-typing while aiming to retain its rapid, PCR-based, reproducible generation of data suitable for inter-laboratory comparison (Holmes et al., 2010). A three-technique comparison between multi-locus VNTR fingerprinting (MLVF), which targets

Tandem Repeats in the coding regions of 7 core genome genes, multi-locus VNTR analysis (MLVA), which preferentially targets Tandem Repeats in non-coding regions of 8 core genome genes, and PFGE, showed that PFGE remained the most discriminatory of the three techniques (*Ibid.*). Both Tandem Repeat methods were capable of resolving isolates with the same PFGE profile, but interpretation of MLVF data was reported to be subjective (*Ibid.*).

1.1.4.6 mecA targeted molecular characterisation

None of the techniques described above specifically confirms the presence of the mecA operon. Many PCR assays have been developed which target SSC, its mecA operon, and the junction between SSC and the sequences flanking its insertion site in orfX. Such assays leverage the specificity achieved by primer selection with the sensitivity of PCR potentially to achieve direct assay of clinical samples containing unresolved Staphylococci. However, the diversity within the accessory genome presents such targeted assays, and indeed any molecular assay directly targeting elements of the accessory genome, with the molecular equivalent of a moving target, simply because the accessory genome is not static. 'Failures' of such targeted assays are legion, and highlight the need for alternative molecular characterisation.

The application of one such assay, which included a comprehensive set of primers targeting the junction between *orfX* and diverse SSC elements, failed to identify fully 7% of phenotypically-validated MRSA isolates (Huletsky et al., 2004). Targeted sequencing of the SSC from these MRSA isolates identified them as new sequence variants of this element, which the existing assay was unable to amplify (*Ibid.*).

The product of *mecA*, the protein PBP2a, can specifically be targeted in cell lysates prepared from cultures of isolated suspected MRSA using a slide agglutination assay including latex beads conjugated with a monoclonal antibody directed against PBP2a (Bowers et al., 2003). As for other molecular assays targeting elements of the accessory genome, this assay can return false negatives when presented with previously uncharacterised diversity within the accessory genome.

In one application of this assay, an isolate phenotypically confirmed as MRSA gave negative results (Garcia-Alvarez et al., 2011). This isolate also returned negative data from two PCR assays targeting mecA and the SSC/*orfX* junction (*Ibid.*).

In a similar report, two phenotypically confirmed MRSA isolates isolates returned positive PBP2a results from two assays directly detecting the protein, but negative from another (Shore et al., 2011). These isolates also returned negative results from PCR assays targeting the mecA gene (*Ibid.*).

1.1.4.7 When all else fails : sequence it

The authors of both these reports resolved the perplexing findings from their closed assays using an open assay system, Whole Genome Sequencing, to determine the genetic basis of meticillin resistance (Garcia-Alvarez et al., 2011, Shore et al., 2011). The isolates were shown to have highly divergent SSC structures accommodating a divergent *mecA* gene (*Ibid*.).

1.2 Whole Genome Sequencing

Whole Genome Sequencing refers to the complete determination of the primary nucleotide sequence of a microbial genome. In the decade since the Human Genome Project highlighted the prodigious expense of capillary based DNA sequencers, many alternative platforms capable of delivering such data at a fraction of the cost have been developed (Mardis, 2011). As these platforms become increasingly accessible, the WGS analytical route is rapidly becoming a new gold standard for microbial characterisation. Recent notable applications of this approach to outbreaks of enterohemorrhagic *E. coli* O104 and *Klebsiella* have lead to predictions of its use for ‘routine diagnosis in quasi real time’ in the foreseeable future (Mellmann et al., 2011, Kupferschmidt, 2011). As of early 2012, the Joint Genomes Institute-curated Genomes Online database (GOLD) lists over 3000 microbial genomes as completed and over 9000 ongoing and targeted projects (Pagani et al., 2012). The technologies that have made this possible are introduced in this section.

1.2.1 DNA sequencing

Watson and Crick's classic understatement - "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material" (Watson and Crick, 1953) - unknowingly alluded to what subsequently became the biochemical underpinnings of the 'sequencing by synthesis' approach used by most commercialised DNA sequencing platforms.

1.2.1.1 The Sanger Method

For the first 3 decades of DNA sequencing the Sanger dideoxy method dominated DNA sequencing. An outline of this method serves well not only to introduce the paradigm shift of the more recently developed NGS platforms, but also to highlight the common biochemical underpinnings of all these methods.

A sequencing project using the Sanger method has four clearly-defined phases (Sanger et al., 1977, Shendure and Ji, 2008). Briefly, phase 1 fragments the target DNA and provides for its amplification, originally by cloning and growth in bacterial systems, latterly often by PCR; phase 2 consists of using these amplified DNAs as templates for primer-directed 'cycle sequencing', in which PCR-like cycling with a single primer in the presence of fluorescently-labelled dideoxy chain-terminators generates stochastically–terminated, end-labelled extension products between 20 and 1500nucleotides in length; phase 3 consists of the physical separation of these products by high resolution capillary electrophoresis, with the sequence of the template DNA being determined by detection of the fluorescently-labelled termination products as they near the end of the capillaries. Refinement of the Sanger technique over 30 years culminated in read lengths approaching 1000 nucleotides and raw base accuracies as high as 99.999%. Multi-capillary cassettes allowed limited parallel processing of up to 384 samples (*Ibid*). In the final phase, downstream of sequence acquisition, read assembly and sequence finishing was time-consuming, with even the best software packages only optionally making use of base call quality values (measures of the uncertainty attached to each base call) provided in a separate file – for example, Phrap (Green, 1999) and Consed (Gordon et al., 1998). Legacy tools such as Blast, which makes no use whatsoever of base call quality values, were typically used to map reads and contigs.

1.2.1.2 The 'NGS Method'

The 'massively parallel' NGS platforms introduced since 2005 have much in common with the Sanger outline above. All variously fragment their target DNA, and most also provide for its amplification (step1). Conceptually, step 2 is also shared; with the sole exception of ABI's SOLiD platform, which uses a 'sequencing by ligation' paradigm, all rely on variously detecting the incorporation of a correctly base-paired nucleotide into the newly synthesized complement to a template strand. All, however, significantly diverge from Sanger, and achieve their eponymous massively parallel processing, by abandoning the post-synthesis CE-based physical resolution of termination products, and instead use physical resolution of templates prior to *in situ* sequencing on various physical support arrays (*Ibid*). This shift to what is termed 'cyclic array sequencing', in which hundreds of thousands to millions and soon billions of arrayed templates are simultaneously sequenced on various support structures, produces data on a scale orders of magnitude greater than that of Sanger, with up to 50Gb per run for some platforms (Metzker, 2010).

1.2.2 Data formats and analysis tools : 'all change, please'

As Di Lampedusa says in The Leopard, 'Everything must change to stay the same'. Three key differences between data from Sanger machines and that from the new platforms necessitated the development of new data analysis approaches – the sheer scale of the data, the length of the reads, and the error-rate of the base calls.

NGS platforms produce data on an unparalleled scale, with single machines run by a single operator capable of producing as much data in 24 hours as several hundred Sanger machines. The scale of this data alone would preclude use of legacy software such as blast to map reads (as do the characteristics of the search algorithm it uses). This reality has necessitated the creation of more efficient tools to handle read alignment/mapping.

NGS read-lengths are platform-characteristic and in general are far shorter than Sanger reads. The market-leader, Illumina, can generate read lengths of 100 or 150, depending on the exact platform used (Illumina, 2012); the ABI SOLiD can generate read lengths of just 75.(SOLiD_homepage, 2012); the Roche 454 has recently demonstrated longer reads with mode length of 700 (Karow, 2011); and the LifeTech IonTorrent has produced mean read lengths of 240 (Jenkins, 2011). Although each platform is capable of producing these read lengths, many datasets are deliberately shorter, because longer runs cost more and for many applications greater length will not necessarily be correspondingly beneficial. The preponderance of short reads, which are less likely to align uniquely to a reference sequence, is likely to persist. Legacy tools deal poorly with such short reads.

NGS platform base call error-rates far exceed those of the Sanger platform, which after 30 years development approached accuracy of 99.999% (Shendure and Ji, 2008). After 5 years development, the error-rate of even the most accurate NGS platform was more than 2 orders of magnitude greater than that of Sanger (Nowrousian, 2010), and in some cases exceeds the SNP rate in the human genome. There is consequently less certainty about base calls, and a correspondingly greater need to maintain and update measures of that uncertainty throughout all phases of data acquisition and processing.

1.2.3 NGS data format : fastq

The fastq data format (.fq, .fastq) was originally developed at The Sanger to maintain in a single file both the bases called for a sequence read and the associated base call quality scores, typically referred to as ‘Phred’ scores after the eponymous program originally introduced for capillary sequencing data (Ewing et al., 1998). Only recently formally defined (Cock et al., 2010), the fastq format has become the accepted standard for NGS data. Figure 1 illustrates the format with an entry from the data analysed here.

Figure 1. Fastq data format

```
@HWUSI-EAS1501_0027_FC70J05AAXX:3:1:12206:2249#TGACCA/2
AATATTATTTTGATATAAATATCAAAATAATATTAGATCGGAAGAGCGTCGTGTAGGGAAAG
AGGGTAGATCTCGGTGGTCGCCGTGTCCTTTAAAAAAGAGTCG
+HWUSI-EAS1501_0027_FC70J05AAXX:3:1:12206:2249#TGACCA/2
hhhhghhghhhcfghhghhghgcchffQ]cWZTZW^^^^^f]fc_^Y^[[^\_W[^``BBBBBBBBBBBB
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
```

An example sequence entry from a fastq file. The sequence entry consists of four lines, the second and fourth of which have wrapped on account of their length. The first and third lines, starting @[...] and +[...], specify colon-separated identifiers for the machine, flow cell lane (3) and tile (1), and the x-(12206) and y-(2249) coordinates of the sequenced cluster. #TGACCA indicates the index sequence used for sample identification in the multiplexed lane, and the trailing '/2' indicates this is read_2 from the sequenced cluster. The second line specifies the sequence read, and the fourth the quality calls. In this early Illumina data, the probability, p, that the call is wrong is integer transformed using the expression $-10\log_{10}$ (p/1-p), and the resulting integers represented by +64 offset ASCII characters.

1.2.4 NGS data aligners

Legacy aligners like blast, Basic Local Alignment Tool, are inappropriate for NGS short reads. Its seed-extension algorithm is optimised to identify distant homologs to query sequences by local alignment extension from initially-identified conserved 'seed' motifs (Altschul et al., 1990). This is not what NGS alignment requires. It is also slow. NGS alignment requires a rapid algorithm capable of identifying exact or nearly-exact matches, without excluding sequence variants that may represent either base call errors or SNPs. Two groups of aligners have been developed to serve this purpose, each built around a different algorithmic core (Li and Homer, 2010).

The first group of aligners are evolutions of the blast algorithm, using variations on hashed seed-extension. Blast by default used a seed of 11 consecutive matches. Discontinuous seeds (the ‘spaced-seed’ approach) improve sensitivity and are used by ELAND, ZOOM, SOAP and others (*Ibid.*).Use of multiple spaced seeds can all but eliminate match-extension, allowing accelerated performance for some aligners including ELAND. Allowing gaps within seeds and/or using multiple seeds to accelerate alignment are used by SHRiMP, RazerS, SSAHA2 and BLAT. Others make improvements to the speed of match extension, including Novoalign, CLC workbench and SHRiMP (*Ibid*.). The hash seed-extension aligners require more memory, run slower, but are more sensitive than the second group.

The second, smaller, group of aligners are built around suffix trees, data structures that allow rapid string matching by representing all string suffixes, and a rapid-access index data structure, the Burrows-Wheeler Transform. These aligners include Bowtie, BWA, SOAP2 & BWA-SW (*Ibid.*). Relative to the first group, these aligners require less memory, run faster, but are less sensitive, and generally do not handle indels well (*Ibid.*).

Factors that influence alignment performance include input data quality, existence of read pairs, and provision of base call quality. Input data quality must be maintained e.g. by removal of sequence contaminants such as adapter sequences and low-quality sequences. ‘Read pair’ refers to paired sequences derived from both ends of defined-length genomic fragments. For ‘paired-end’ reads, these fragments are relatively short, typically 150-450 nucleotides, and not greater than 700. ‘Mate-pair’ reads derive from larger genomic fragments (1kb-20kb or more) which are re-circularised about a common adapter from which bi-directional sequencing is subsequently primed. Read pair entries help resolve repeat structures, which can otherwise confound aligners. Providing base quality scores has been shown to improve alignment performance, but only MAQ, Novoalign and Bowtie can use base quality scores (Ibid).

Selection of an aligner from the more than 50 available depends on a number of factors including the expected divergence of the reads from the reference & the length of the reads. Less divergent sequences can take advantage of the faster suffix tree algorithms. In this group only BWA deals adequately with indels, or can be used with the longer reads from Roche454 and IonTorrent. For more divergent sequences, aligners from the hashing group are indicated. In this group BLAT, SHRiMP2 and Mosaik can handle the longer reads.

All aligners are highly dependent on the quality of the input data.

1.2.5 NGS alignment format : SAM/BAM

Legacy aligners such as blast output in an intuitive but inflexible data format poorly suited to representing multiple read alignments. The ‘Sequence Alignment/Map’ (SAM) format was originally developed for the 1000 Genomes Project and has become the accepted output format for NGS aligners (Li et al., 2009). SAM has a compressed companion format, Binary Alignment/Map (BAM), to improve performance, and an associated ‘SAMtools’ software package for manipulations in SAM/BAM format (SAMtools_home_page). Figure 2 illustrates the format with entries from the data analysed here.

Figure 2. SAM/BAM format

```
HWUSI-EAS1501_0027_FC70J05AAXX:3:1:12206:2249#TGACCA	77	*	0	0*
*	0	0
AATATTATTTTGATATAAATCTCAAAAAAATATTAGATCGGAAGAGCACACGTCTGAAATCCC
ATCACTGACCAATCTCGTCNGCCCTCTTCTTTTTTGTAAAAA
ffffffffdcccfcffffBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB	AS:i:0
HWUSI-EAS1501_0027_FC70J05AAXX:3:1:12206:2249#TGACCA	141	*	0	0*
*	0	0
AATATTATTTTGATATAAATATCAAAATAATATTAGATCGGAAGAGCGTCGTGTAGGGAAAG
AGGGTAGATCTCGGTGGTCGCCGTGTCCTTTAAAAAAGAGTCG
hhhhghhghhhcfghhghhghgcchffQ]cWZTZW^^^^^f]fc_^Y^[[^\_W[^``BBBBBBBBBBBB
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB	AS:i:0
```

An entry from a SAM file, using the same sequence cluster as figure 1. The format has a single line per entry, with 11 tab-separated fields per line, as detailed in Li et al., 2009. The underlining is not part of the format but indicates entry contamination with indexed adaptor (read_1, above) and universal adapter (read_2, below), an early indication of data quality issues with libraries used here. (file generated by Taane Clark by aligning on-QCed, non-sanitized, unfiltered data with aligner Smalt against MRSA252 reference sequence).

1.2.6 NGS assemblers

For sequence reads which do not align to reference, or in the absence of a reference, sequence assembly tools are necessary. Legacy tools such as the Human Genome project workhorse assembler Phrap, which requires high quality sequence read lengths of 500-1000, are unsuited to NGS data (Miller et al., 2010). NGS assemblers need to work efficiently with large volumes of the shorter, lower quality reads these platforms generate. Two groups of sequence assemblers have been developed, each based around a distinct algorithmic core (*Ibid.*).

The Overlap/Layout/Consensus (OLS) methods manipulate reads intact and establish links between them through use of exhaustive pair-wise mapping of their overlaps. The path through these links establishes a consensus sequence

(*Ibid*.). Manipulating the totality of all reads intact, OLS assemblers are computationally expensive, and most frequently applied to Sanger & other long read datasets. Their memory requirements preclude their application to read numbers above ~1million. The OLC assemblers Newbler and the Celera Assembler are used for 454 data; Edena and Shorty have been applied to short reads from Illumina and SOLiD platforms (Ibid.).

The de Bruijn graph (DBG) assemblers subdivide each read into defined-length kmers and represent each read as a linked path through kmer nodes. Re-occurrence of previously-encountered kmers within subsequent reads defines new links and causes path divergence/graph expansion; this kmer ‘re-use’ collapses sequence complexity and generates a consensus as a ‘by-product’ of graph assembly without the exhaustive pair-wise comparisons of the OLC assemblers. The DBG assemblers are less computationally-demanding than their OLS counterparts. Euler and Velvet are well established DBG assemblers (Ibid).

Factors that influence assembler performance include sequence variants, input data quality and sequence repeats. Sequence variants resulting from sequence error or low quality sequence inputs cause graph expansion and miss-assembly. Sequence inputs must be sanitised of contaminant sequences and of low quality sequence regions. Repeat sequences within target sequences confound assemblers. Effective resolution of repeats requires paired-end and/or mate-pair reads that span the repeat, and others which have just ‘one foot’ in the repeat by having one end within the repeat and the other in unique flanking sequence (Ibid).

Selection of aligner from the more than 20 available depends on a number of factors including the read length and number of reads. The OLC class assemblers work well with longer reads, but are precluded with high data volumes. The DBG assemblers work well with short reads, with Velvet frequently being used for bacterial assembly. All are dependent on data quality.

1.2.7 Before we begin : data QC

In any system, omitting data Quality Control prior to data analysis is likely to confirm one variant or another of the old adage ‘Rubbish In, Rubbish Out’. The scale of NGs data presents an opportunity to confirm this adage on an unparalleled scale.

The massively parallel nature of NGS data generation means that systematic data QC across entire datasets must be performed computationally. There are a number of NGS QC packages available, of which the FastX tookit from the HannonLab (Gordon, 2009), which is also available through the Galaxy site (Goecks et al., 2010), and the FastQC package from The Babraham Institute (Andrews, 2011a) are the most widely used owing to their complementarity.

The FastQC package is an entirely analytic package determining 11 NGS QC measures as detailed in Table 1.

Table 1. FastQC metrics

FastQC metric	Depicts per library
1. Basic statistics	Number of reads, read length, sequence %GC
2. Per base sequence quality	Boxplots of Phred/Q scores across read length
3. Per sequence quality scores	Phred/Q score distribution
4. Per base sequence content	Average base calls across read length (4 bases)
5. Per base GC content	Average base calls across read length (GC only)
6. Per sequence GC content	GC distribution over all sequences
7. Per base N content	N content across read length
8. Sequence length distribution	Distribution of all read lengths
9. Sequence duplication levels	Sequence duplication in bins 1..9, 10+
10. Over-represented sequences	Sequence duplications comprising > 0.1% of library
11. kmer content	Frequency of over-represented 5-mers across read length

FastX includes a limited set of analytical tools together with a comprehensive set of processing tools as detailed at http://hannonlab.cshl.edu/fastx_toolkit.

1.3 Aims & objectives

The overall objective of this thesis is to determine mecA status of the Staphylococcal Chromosome Cassette for 10 MRSA isolates for which NGS data generated on an Illumina GAIIx machine have been made available.

The intermediate steps of a corresponding analytical path include:

- Data quality appraisal; the overall quality of the isolate sequence libraries will be determined using the ‘FastQC’ toolset from the Brabraham Institute.

- Data pre-processing; based on the data quality metrics generated from the FastQC analysis, the data libraries will be pre-processed to address any QC issues, after which the FastQC analysis will be repeated to confirm the efficacy of pre-processing. Re-iterative cycles may necessary to identify useful pre-processing parameters.

- Assembly; pre-processed libraries will be assembled using the De Bruijn graph program Velvet, with assembly optimisation using the weighted mean contig length, ‘N50’, as an optimisation criterion.

- Comparison to reference sequences; the contigs from Velvet will be compared with MRSA reference sequences using the Mummer toolset. Initial comparisons will use MRSA252, representative of the epidemic eMRSA-16 lineage typical of UK HA-MRSA. Subsequent comparisons may use further MRSA sequences as necessary to identify the SCC sequence representation of the isolates under study.

- The mecA status of each SCC will be determined.

2 MATERIALS AND METHODS

2.1 Raw data quality appraisal.

Fastq-formatted raw sequencing data for 10 isolates of *Staphylococcus aureus* were made available at the LSHTM. They were generated on an Illumina Genome Analyzer IIx using the paired read protocol for obtaining sequence from both ends of each cluster, so two files were returned for each isolate library. The protocol used to generate the data was requested but not made available.

2.1.1 FastQC - data analysis

FastQC v0.9.4 was obtained from the Babraham Bioinformatics group at the BBSRC Babraham Institute. The originators of this comprehensive package describe it as "a quality control application for high throughput sequence data. It reads in sequence data in a variety of formats and can either provide an interactive application to review the results of several different QC checks, or create an HTML based report which can be integrated into a pipeline" (Andrews, 2011a). The raw fastq-formatted data were analysed using FastQC according to the online manual (*Ibid.*). Rather than running 20 individual fastq files through FastQC, the number of individual FastQC runs was reduced to the number of isolates by interleaving into a single file the two paired-end read files for each isolate using the Velvet perlscript shufflesequences_fastq.pl (Zerbino and Birney, 2008).

2.1.2 Data pre-processing

Data were processed to remove the sequence contaminants identified by FastQC. Preprocessing initially used the FastX toolkit of command-line NGS data-processing functions (Gordon, 2009). Calls were automated using a shell script sequentially to call FastX_clipper, to clip adapter sequences, FastX_trimmer, to remove low-quality sequence regions, and FastX_quality_filter, to remove entirely those sequences still falling below a quality threshold. External files were used to provide parameters to the fastX function calls (sample names, adapter sequences, trim length definitions and filter parameters). After each FastX function, a perl script call (syncReads.pl) was used to synchronise the processed sequences, after which a call to FastQC was made to determine the impact of the individual pre-processing steps. This strategy was terminated when the FastX_toolkit was found not accurately to process the longer reads (105nucleotides) present in these libraries. The FastX-toolkit use was abandoned in favour of cutadapt (Martin, 2011), which was used both to remove adapter sequences and to trim low quality regions of sequence. After cutadapt processing the read files were again compiled into a single file as above & re-analysed using FastQC. The wrapper shell script which coordinated pre-processing ('postFastQC1_v10.sh') and ancillary scripts it calls are described in 4.4.4Appendix A.

2.1.3 FastX_collapser: confirmation of data duplication.

Fastx_collapser, a program in the FastX package which collapses input FastQ or FastA files into a non-redundant FastA output file, concomitantly reporting the input and output read metrics was used to confirm sequence duplication levels.

2.2 De novo assembly with Velvet

The Debruijn graph-based assembler Velvet was used to assemble the processed libraries (Zerbino and Birney, 2008). Input libraries were subsetted as necessary to limit sequence inputs using the script described in appendixC. Assembly was optimised using Velvet Optimiser, a perl script systematically applying values within a user-specified range and returning those yielding an optimal assembly (Gladman, 2011).

2.3 Mapping assemblies to MRSA 252 with bl2seq

Assembled contigs returned by velvet were mapped against reference genomes using the blast suite program bl2seq called from the wrapper script compare_library.pl (Stothard).

2.4 SCC-specific contig mapping with mummer

The ‘mummer’ family of suffix tree sequence comparison programs was used extensively to visualise the mappings between contigs assembled by Velvet and the MRSA252 reference sequence (Kurtz et al., 2004). Initial dna:dna comparisons used the core mummer program, together with the ancillary program mummerplot, to construct ‘dotter’ type visualisations of contig mappings. NUCmer was used to generate graphical mappings against the reference sequence. Visualisations complete with genome annotations extracted from gff3 files were constructed using the programs NUCmer and PROTmer together with the ancillary programs show-coords, mapview and xfig (*Ibid*.). The tblastn-like sensitivity of PROTmer was used both as a primary search tool and also to validate and confirm mappings made with either blastn or blastp bl2seq.

3 RESULTS

3.1 Raw data analysis

3.1.1 FastQC analysis

Analysis of the raw fastq-formatted data files using the FastQC package from the Babraham Bioinformatics Group was used to obtain a quality control overview of the sequences.

The 11 output metrics generated by FastQC provide an indication of library and sequence read quality. Although the terms in which the metrics are expressed (‘Pass’, ‘Warn’ and ‘Fail’) suggest delineations into definitive sequence categories, the package documentation makes clear that these terms are but succinct substitutes for ‘entirely normal’, ‘slightly abnormal’ and ‘very unusual’ and that the output needs to be considered in the context of expectations for a given library (Andrews, 2011a). Table 2. FastQC analysis presents a summary of the FastQC output. In the description of this output that follows, the same three libraries are generally used to illustrate the FastQC metrics; from the left of the quality-sorted table 1, library 352 is used as an ‘entirely normal’ exemplar; from the right of this table, libraries 3016 and 374 are used as ‘very unusual’ exemplars. The metrics are discussed in the descending row order of table 1.

Metric 1 is presented in Table 3. These data show that the number of sequence reads per library ranges from 1 to 10 million, and that the reported %GC spans 32-41% (expected range 33-34%; (Holden et al., 2004). FastQC reported all libraries to be ‘entirely normal’ by this measure (table 1).

Metric 3, is presented in Figure 3. Per Sequence Quality Scores. This metric determines the average Qscore for each read and plots the distribution of this average Qscore for all sequences in the library. It allows identification of libraries requiring Qscore-based read filtering to remove low quality reads entirely before further analysis. Most libraries required no filtering, including library 352 included as an exemplar in figure 1. However the plots for libraries

3016 and 374 include leading spikes and secondary peaks indicative of low quality reads which would be better removed. FastQC considered all libraries to be ‘entirely normal’ by this measure (table 1).

Metric 7 (not shown), ‘Per base n content’, depicts ambiguous calls against read position FastQC considered all libraries to be ‘entirely normal’ by this measure (table 1).

Metric 8 (not shown), ‘Sequence Length Distribution’, plots the distribution of read sequence lengths. All sequences in these libraries were 105nucleotides. FastQC considered all libraries to be ‘entirely normal’ by this measure (table 1).

Metric 2 is illustrated in Figure 4. Per Base Sequence Quality Scores. This metric uses boxplots to depict sequence Q score against read length. It allows identification of libraries that would benefit from read-length trimming to remove low quality read regions before further analysis. The libraries 3016 and 374 included in figure 2 are among such libraries. The abrupt quality inflexion point in the 374 data is untypical, and may reflect issues during library preparation rather than issues with the sequencing process, since all libraries were sequenced within a single lane of the Illumina machine. Alternatively local anomalies within the sequencing lane (dust, lint) may have impacted these reads. FastQC reported 6 libraries to be ‘entirely normal’ and 4 to be ‘very unusual by this measure (table 1). The ‘very unusual’ libraries beyond those considered in the figure are illustrated after pre-processing in 3.2.2

Metric 6 is illustrated in Figure 5. Per Sequence GC content compares the distribution of GC over all sequences against a Normal distribution having the same mean and standard deviation as the library. Deviations from N are indicative of contaminants within the library. All libraries but 352 and 2999 showed evidence of contamination by this measure (not shown). FastQC considered 6 libraries to be ‘very unusual’, and 4 to be ‘slightly abnormal’ by this measure (table 1).

Metric 5 is illustrated in Figure 6. Per Base GC Content is much like the previous one but focuses solely on GC frequency. In a random library this frequency would be expected to be unchanged against read position. This metric allows identification of biased libraries. FastQC considered 5 libraries to be ‘very unusual’, and 5 to be ‘slightly abnormal’ by this measure (table 1).

Metric 10 (not shown), ‘Over-represented sequences‘, tabulates repetitive reads comprising more than 0.1% of total reads and uses database searching to identify them. All libraries were found to include between 1 and 7 variants of the TruSeq Adaptor sequences comprising between 0.26 and >4% of total reads.

Metric 9 is illustrated in Figure 7, Figure 8 and is summarised in Table 4. This metric depicts the extent to which reads in the library are unique. Libraries generated from highly diverse source sequences optimally consist of large numbers of individually unique reads. For the current libraries, generated from a source sequence of limited diversity, some degree of repetition is inevitable. The figures show that this is true for all libraries to some extent. Table 3 shows a clear relationship between the total number of library reads and the resulting sequence duplication levels (from 30% duplication for library 1723 with ~1 million reads, to ~80% duplication for those with > 8 million reads). This outcome suggests that generating more than ~1 million clusters per sample on the sequencing chip is not useful for samples of this size, and merely results in the generation of redundant information. Such overrepresentation of individual reads may reflect over-amplification of sequences prior to bridge amplification, or excess input onto the sequencing lane. The plots for some libraries (notably 535, 303, 352) in addition include a characteristic 10+ tail indicating a higher level of read duplication for a subset of contaminant sequences. FastQC reported one library to be ‘slightly abnormal’ and the remaining 9 to be ‘very unusual’ by this measure (table 1).

Metric 4 is illustrated in Figure 9. Per Base Sequence Content. This metric depicts base frequency against read position. In a random library all four base frequencies would be expected to be unchanged against read position. This metric allows identification of biased libraries. Libraries 3016 and 374 share a common primer-proximal profile within which the sequence of the TruSeq Adapter used for library construction can be discerned. FastQC considered all libraries to be ‘very unusual’ by this measure (table 1).

Metric 11 is illustrated in Figure 10. Sequence kmer Content depicts against read length the relative enrichment of those length 5 kmers which appear more frequently than expected within a sequence of the read’s base composition. It is useful for identifying library contaminants. All libraries showed evidence of contaminating sequences. The FastQC metric-sorted output presented in table 1 ranked library 352 most highly. Even for this most highly ranked library the peaks at the beginning of the Kmer content plot (figure 8) indicate adapter contamination; from the colour-coded peaks it is possible to assemble nucleotides 2-13 of the TruSeq adapter used in library construction (pink ATCGG, green GGAAG, yellow AGAGC, black GAGCA, in which the underline indicates alignment with the previous kmer). Of the total of ~6 x 10^8 enriched kmer tags identified in all 10 libraries, 48% derived from the TruSeq adapters used for library generation. Cumulatively these data were sufficient to allow complete reassembly of the first 14nucleotides of the indexed TruSeq TruSeq adaptor from all but 2 libraries (Table 5. Read-Depth Summary, TruSeq Adapter-Derived Kmers). A scatterplot of total library reads *versus* TruSeq Adapter-derived kmers reads (Figure 11. Library *vs* Adapter kmer Read-Depth) showed that for most libraries the number of kmer reads was linearly related to the library read depth; these libraries accumulated adapter kmer reads at ~21% per library read. With the exception of libraries 352 and 303, which respectively accumulated adapter kmers at ~6% and ~13%, other libraries had still greater adapter kmer accumulation rates (3016; 31%: 374; 49%). These data indicate a significant residual adapter presence in all libraries.

3.1.2 Analysis of adapter sequence position

FastQC analysis indicated residual adapter sequence contamination of library reads for all libraries. A shell script detailed in Appendix B was used to count adapter sequences in all libraries. With the single exception of Universal primer in library 2999, the majority of adapter sequences were at the start of reads as shown in Table 6. Counts for indexed adapter sequences exceeded those for the universal adapter sequence. For the position-in-read-independent counts, this excess ranged 2.5 to 13.5 fold across the 10 libraries. For the 5'-aligned counts, this excess ranged 2.4 to 30 fold.

Isolate_	352_	1940_	454_	492_	2999_	535_	1723_	374_	303_	3016_	total of
QC metric run index	**615**	**619**	**616**	**617**	**620**	**624**	**618**	**622**	**623**	**621**	30
Basic Statistics (1)	P	P	P	P	P	P	P	P	P	P	30
Per sequence quality scores (3)	P	P	P	P	P	P	P	P	P	P	30
Per base N content (7)	P	P	P	P	P	P	P	P	P	P	30
Sequence Length Distribution (8)	P	P	P	P	P	P	P	P	P	P	30
Per base sequence quality (2)	P	P	P	P	P	P	F	F	F	F	18
Per sequence GC content (6)	P	W	W	W	W	F	F	F	F	F	7
Per base GC content (5)	W	W	W	W	F	F	W	F	F	F	5
Overrepresented sequences (10)	W	W	F	F	W	F	F	W	F	F	4
Sequence Duplication Levels (9)	F	F	F	F	F	F	W	F	F	F	1
Per base sequence content (4)	F	F	F	F	F	F	F	F	F	F	0
Kmer Content (11)	F	F	F	F	F	F	F	F	F	F	0
total of 33	20	18	17	17	17	15	14	13	12	12	

Table 2. FastQC analysis, raw data (summary)

The output of FastQC has been quality-sorted to place the highest scoring library on the left and lowest on the right. The quality sort assigned integers to each of the three output terms (Pass = 3, Warn=1, Fail=0), summed these figures both across columns and down rows, and then sorted on the totals shown. The numbers in parentheses after each QC metric title indicate the order of metric presentation by FastQC.

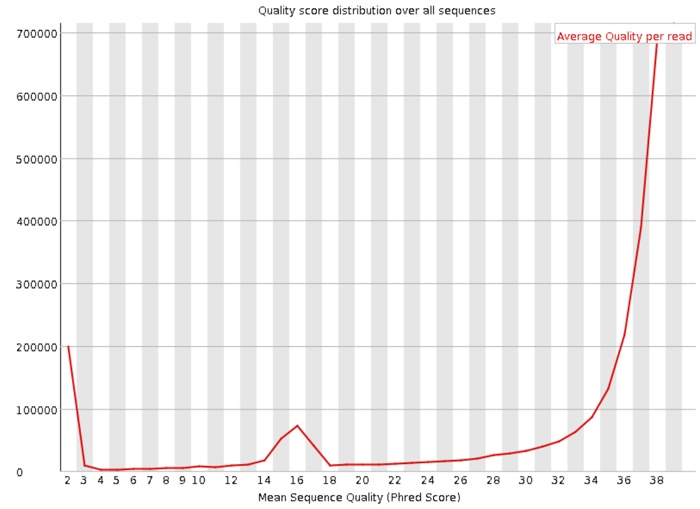
Isolate_ runorder	352_ 615	1940_ 619	454_ 616	492_ 617	2999_ 620	535_ 624	1723_ 618	374_ 622	303_ 623	3016_ 621
FastQC version	0.9.4	0.9.4	0.9.4	0.9.4	0.9.4	0.9.4	0.9.4	0.9.4	0.9.4	0.9.4
Call	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass
File type	Conventional base calls	=	=	=	=	=	=	=	=	=
Encoding	Illumina 1.5	=	=	=	=	=	=	=	=	=
Total Reads	9,918,842	3,768,246	10,040,306	5,109,018	2,194,204	8,121,702	992,154	5,720,698	8,985,616	3,085,900
Read length	105	105	105	105	105	105	105	105	105	105
%GC	32	33	33	33	33	36	33	41	35	37

Table 3. Summary of Basic Statistics

‘=’ indicates entry identical to that to the left.

Isolate_ runorder	352_ 615	1940_ 619	454_ 616	492_ 617	2999_ 620	535_ 624	1723_ 618	374_ 622	303_ 623	3016_ 621
Sequence duplication	~80%	~55%	~82%	~63%	~52%	~82%	~30%	~68%	~80%	~63%

Table 4. Sequence Duplication Levels



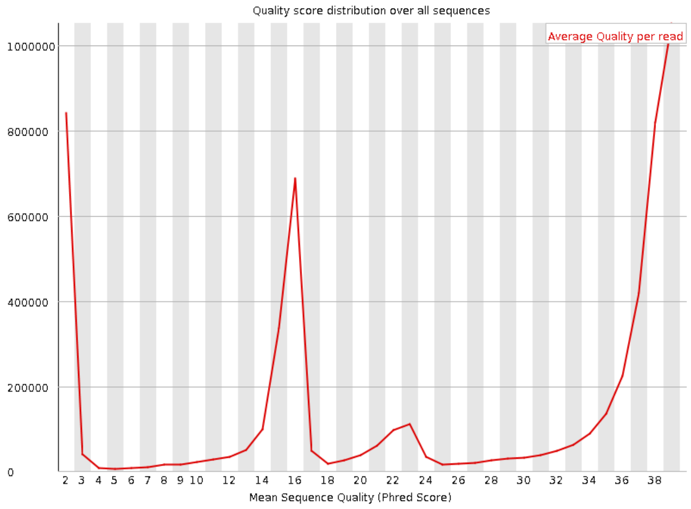
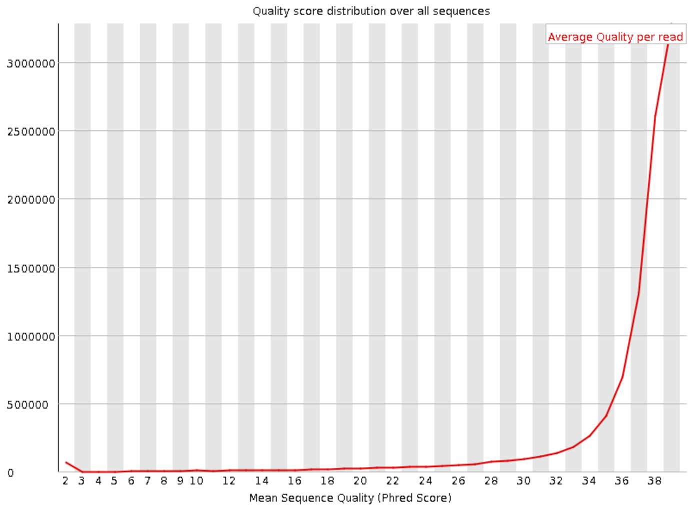
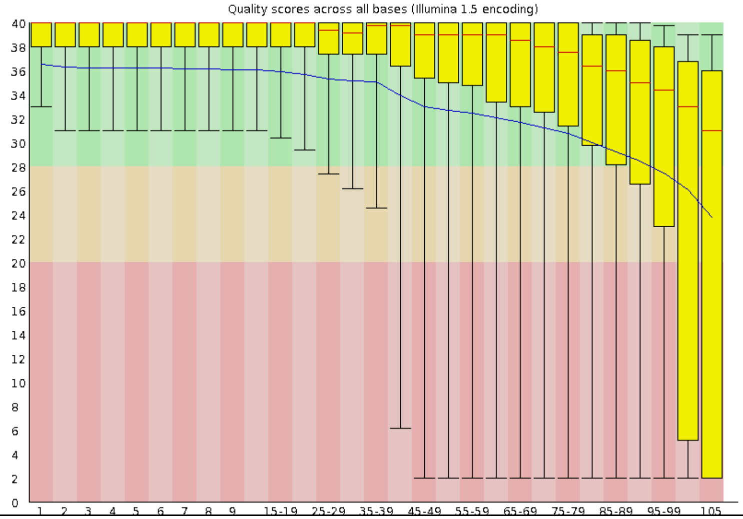


Figure 3. Per Sequence Quality Scores

Depiction of the read average Qscore distribution over all sequences. Library 352 (lower right) approaches the ideal, with the majority of reads having high Qscores. Libraries 3016 (top left), 374 (top right) have in addition secondary peaks indicating a subset of low quality sequence reads.





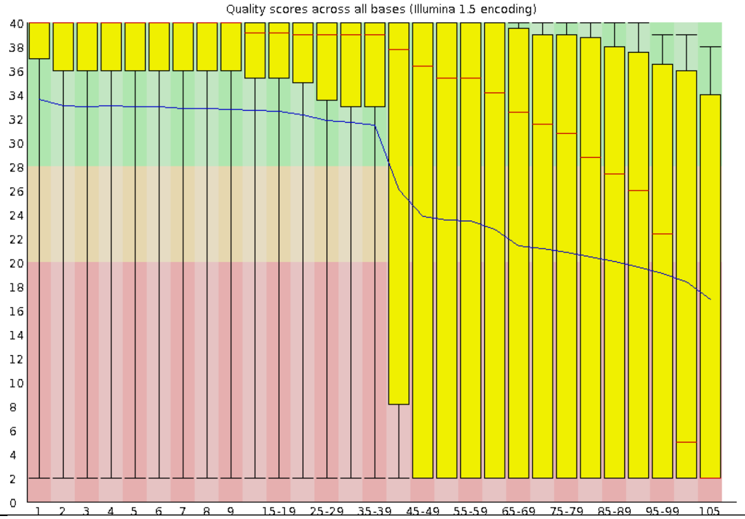
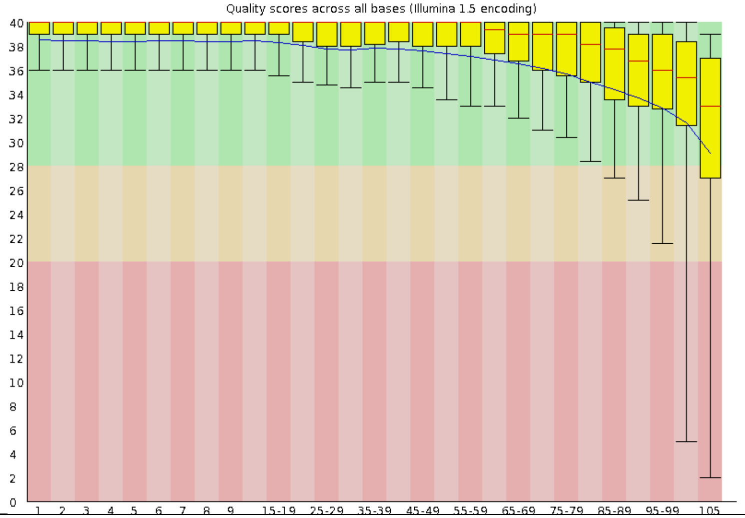
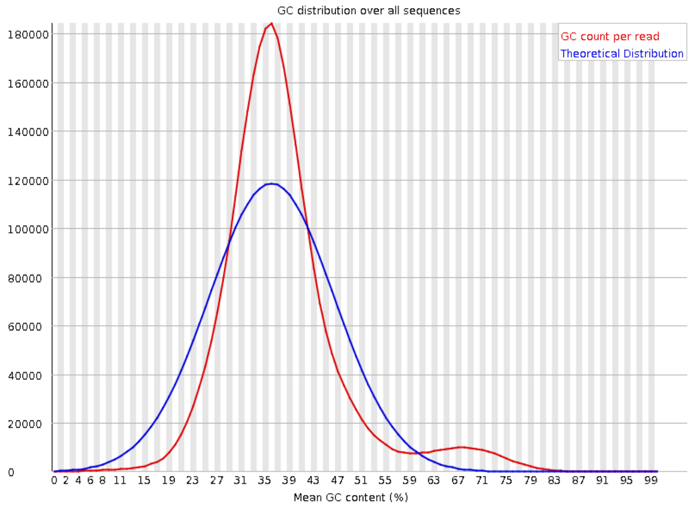


Figure 4. Per Base Sequence Quality Scores

Depiction of sequence quality against read length; whiskers represent the lower and upper deciles, the limits of the box the lower and upper quartiles, the red line the median, and the blue line the mean of the Q score. Libraries 3016 (top left), 374 (top right) drop below both the Q28 (green) and Q22 (beige) quality thresholds on longer reads. Library 352 (lower right) exceeds quality thresholds at all read lengths.





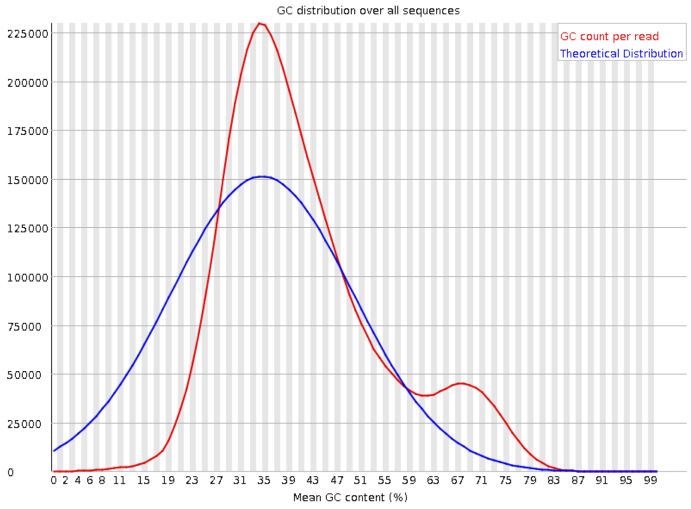
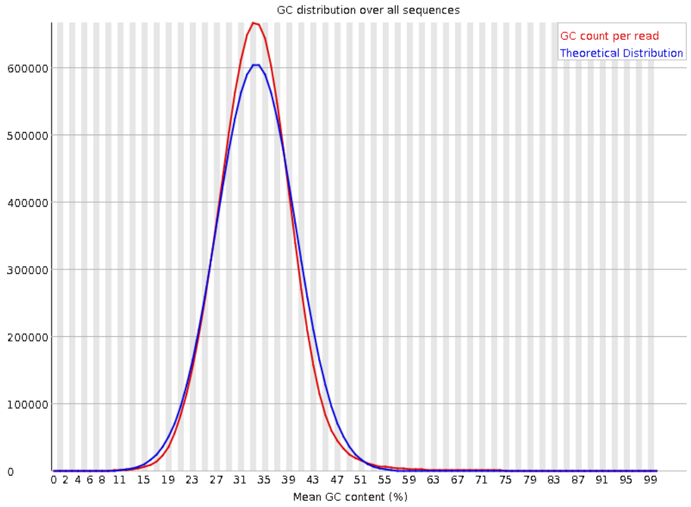
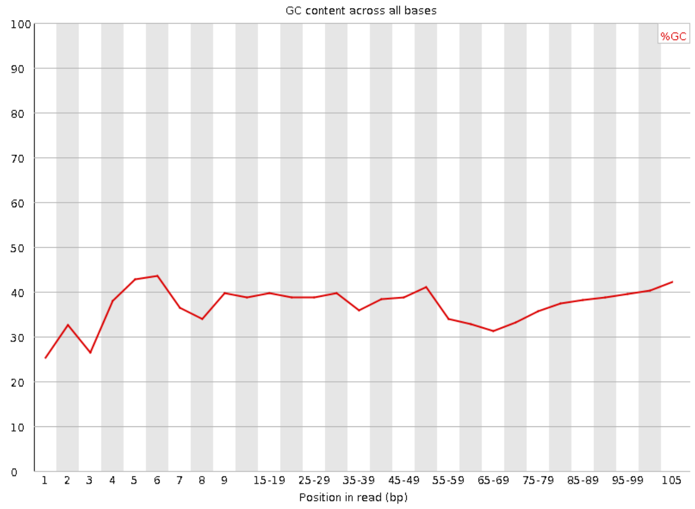


Figure 5. Per Sequence GC content

Depiction of observed GC distribution over all reads relative to a Normal distribution with the same mean and standard deviation as the library. Quality random libraries superimpose well on N (library 352, lower right). Deviations from N are indicative of contaminants within the library. Library 3016 (top left) & 374 (top right) depart from N, suggesting these libraries include contaminants..





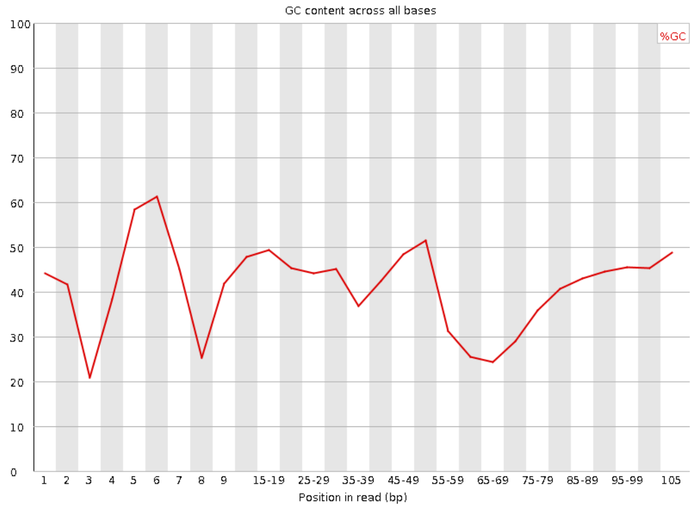
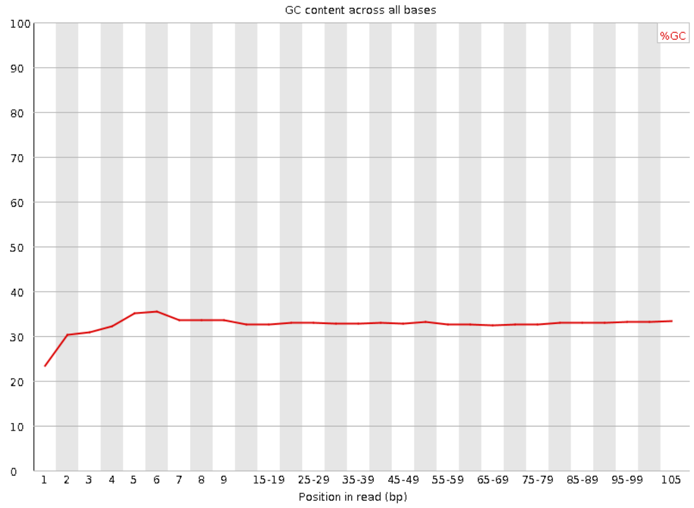
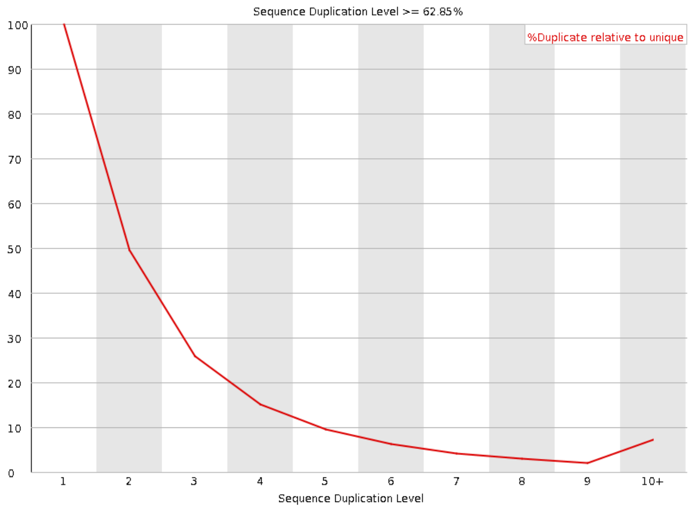


Figure 6. Per Base GC Content

Depiction of GC frequency against read position. Library 352 (lower right) has the expected flat profile with the sole exception of primer-proximal bases. Libraries 3016 (top left), 374 (top right) depart from this profile, indicating that these libraries are not random.





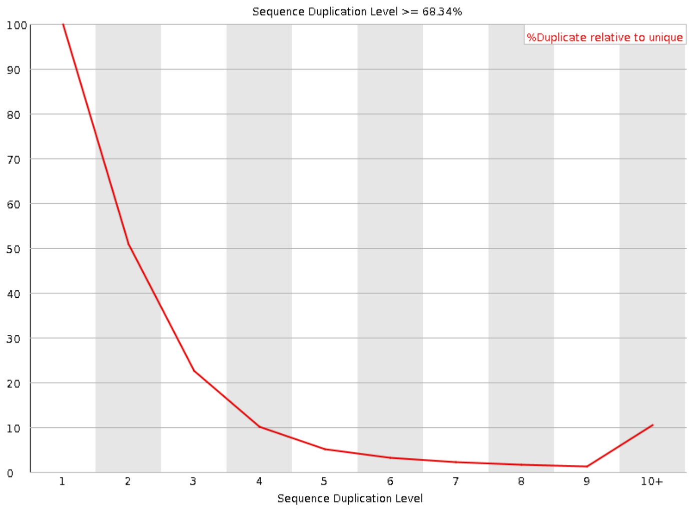
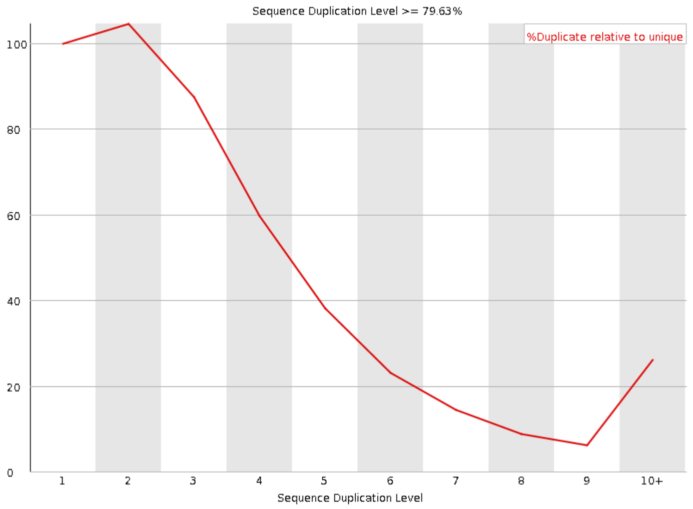
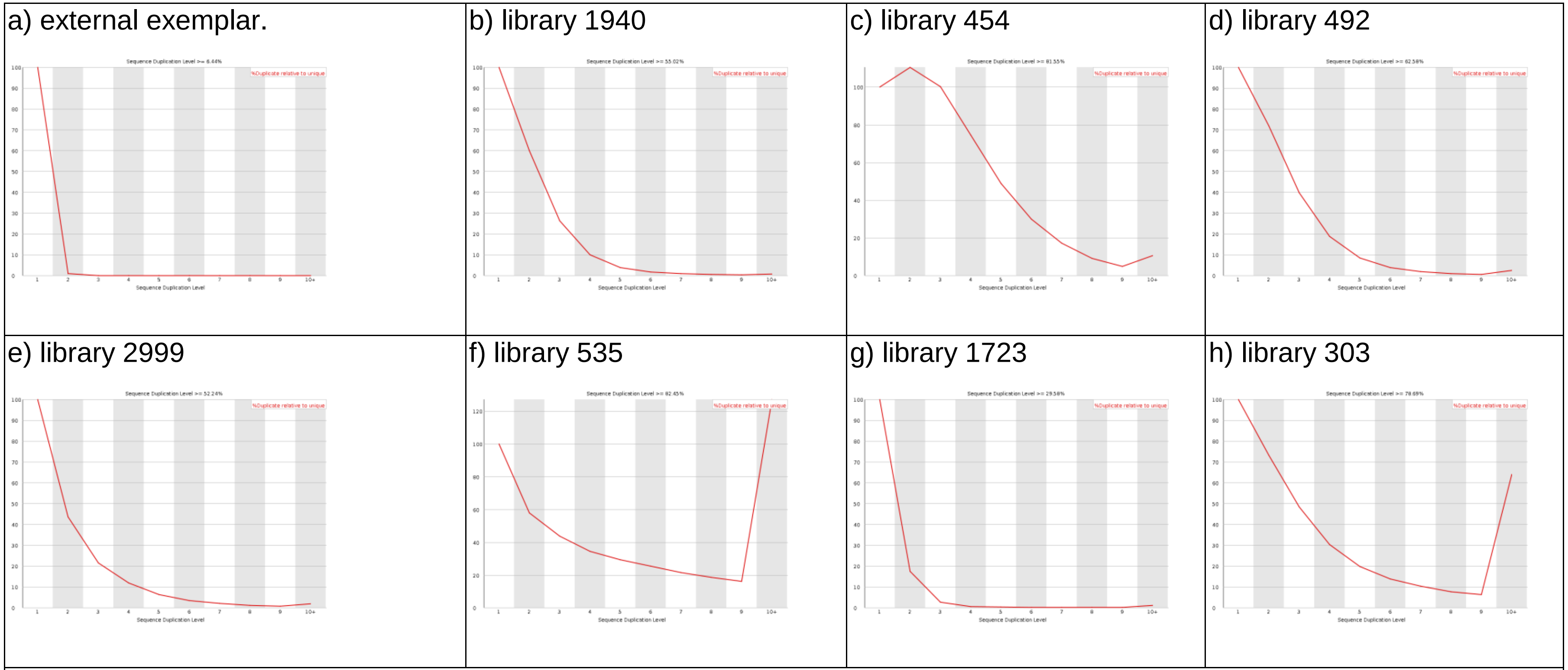


Figure 7. Sequence Duplication (exemplars)

The number of reads duplicated between 2 and >10 times is shown on a scale relative to the number of unique reads, which is set to 100%. Libraries 3016 (above left) and 374 (above right) have a lesser read duplication level than library 352 (right). The upturned ‘tail’ on each profile reflects a higher degree of duplication of a subset of reads.





a) external exemplar.

b) library 1940

c) library 454

d) library 492

e) library 2999

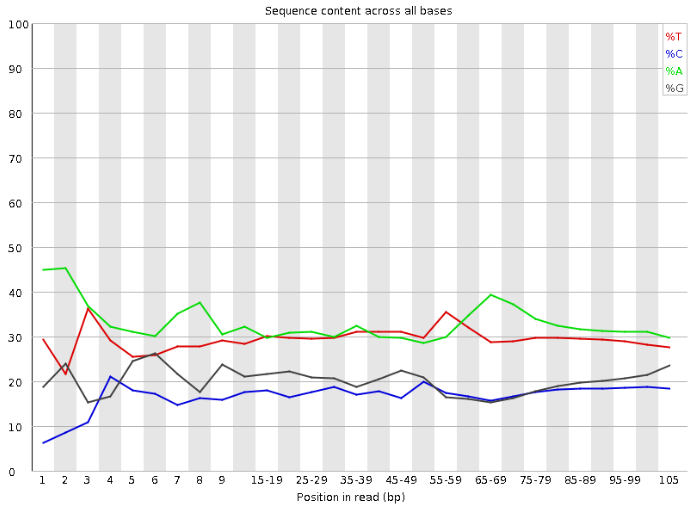
f) library 535

g) library 1723

h) library 303

Figure 8. Sequence Duplication (external exemplar)

Comparison with an external exemplar (Simon 2011c) highlights the extent of sequence duplication within the MRSA libraries, which are presented in the rank order established in table 1. Even those libraries deviating least from the ideal profile (e.g. 1723, 2999, 1940, 492) have substantial sequence duplication. Libraries 535 and 303 have in addition 10+ tails indicating a sequence subset with a still higher rate of read duplication.



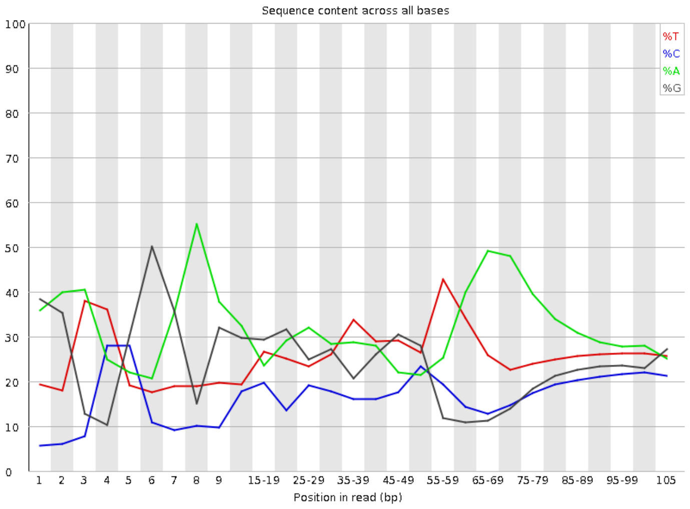
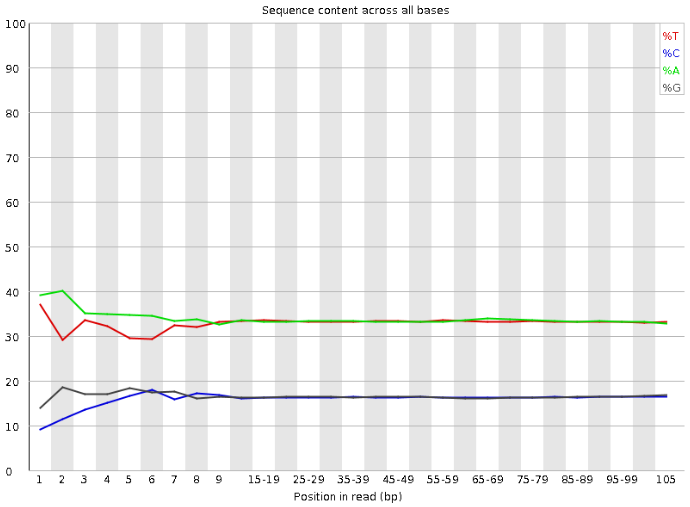
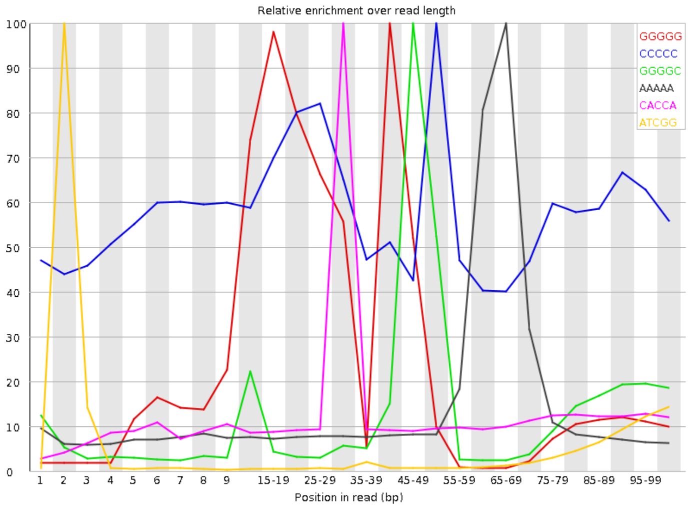


Figure 9. Per Base Sequence Content

Depiction of sequence content for all four bases against read position. Library 352 (lower right) has the expected flat profile with the sole exception of primer-proximal bases. Libraries 3016 (top left), 374 (top right) depart from this profile, suggesting these libraries are not random.





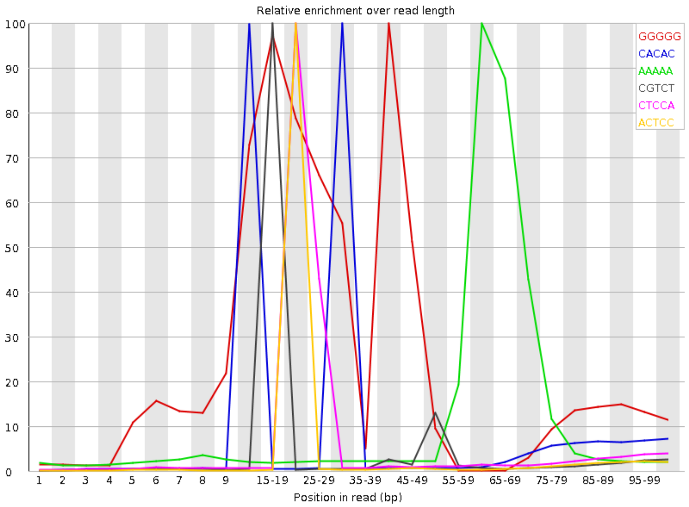
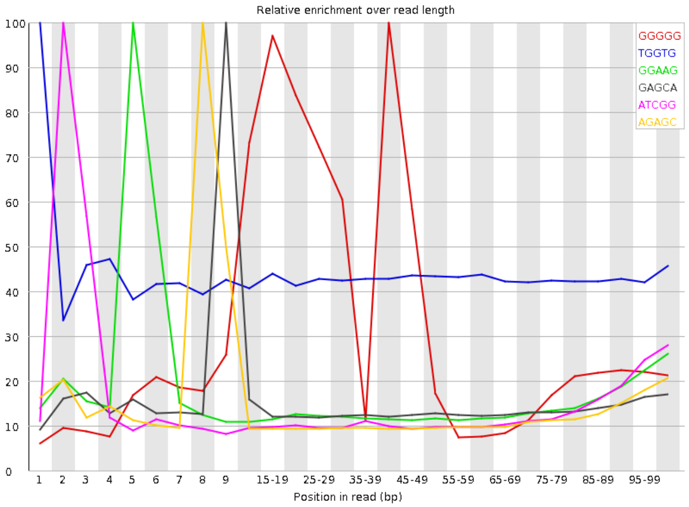


Figure 10. Sequence kmer Content

The relative enrichment of length 5 Kmers is depicted against read length for libraries 3016 (above), 374 (above right) and 352 (right). The offset peaks are characteristic of library adapter contamination.



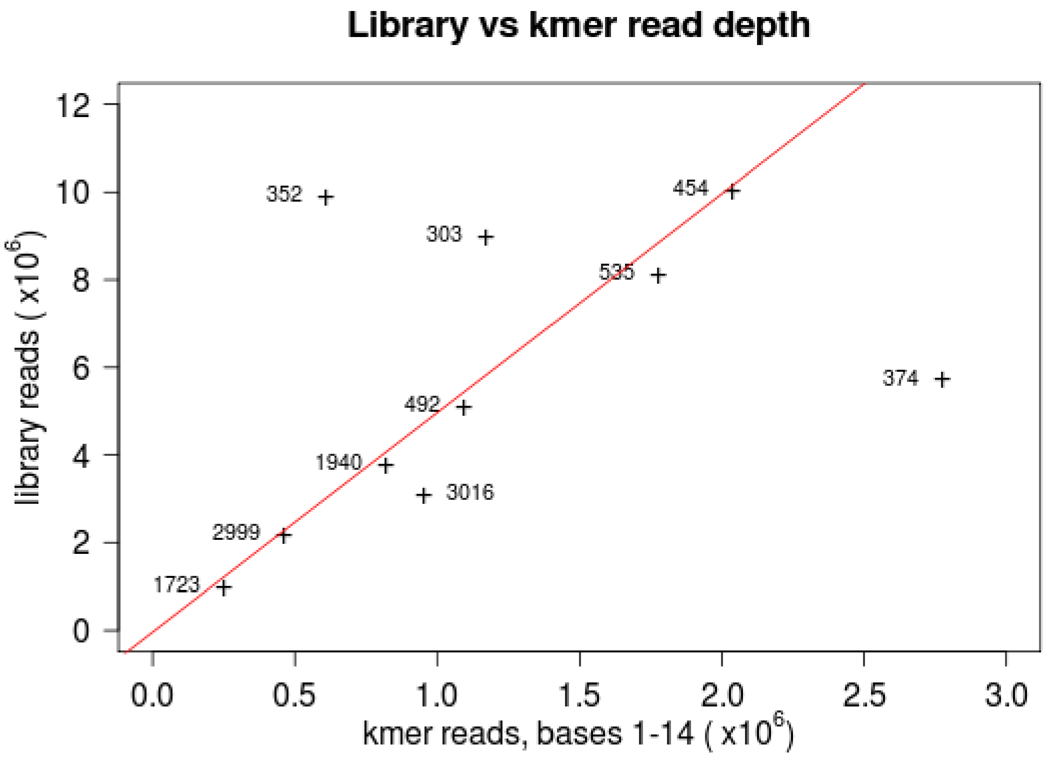
Kmer sequence	Adapter bases	352_ 615	1940_ 619	454_ 616	492_ 617	2999_ 620	535_ 624	1723_ 618	374_ 622	303_ 623	3016_ 621
GATCG	**1 - 5**	3.97E+5	4.16E+5	9.47E+5	5.39E+5	2.30E+5	9.66E+5	1.54E+5	2.06E+6		6.90E+5
ATCGG	**2 - 6**	5.30E+5	4.47E+5	1.04E+6	5.88E+5	2.47E+5	1.07E+6	1.62E+5	2.18E+6		7.14E+5
TCGGA	**3 - 7**	4.05E+5	3.92E+5	9.03E+5	5.11E+5	2.13E+5	9.34E+5	1.46E+5	2.05E+6		6.56E+5
CGGAA	**4 - 8**	4.55E+5	4.01E+5	9.27E+5	5.21E+5	2.14E+5	9.53E+5	1.48E+5	2.08E+6		6.50E+5
GGAAG	**5 - 10**	6.07E+5	4.60E+5	1.08E+6	6.09E+5	2.38E+5	1.14E+6	1.63E+5	2.42E+6		7.19E+5
GAAGA	**6 - 11**		8.19E+5	2.04E+6	1.09E+6	4.63E+5	1.78E+6	2.51E+5	2.51E+6		9.53E+5
AAGAG	**7 - 12**		5.95E+5	1.36E+6	7.52E+5	2.89E+5	1.23E+6	2.03E+5	2.78E+6		7.47E+5
AGAGC	**8 - 13**	4.81E+5	3.72E+5	8.62E+5	4.78E+5	1.84E+5	8.77E+5	1.38E+5	2.01E+6		5.55E+5
GAGCA	**9 - 14**	5.51E+5	3.22E+5	8.03E+5	4.29E+5	1.65E+5	8.46E+5	1.11E+5	1.57E+6	1.17E+6	4.55E+5
TruSeq Adapter seq:		GATCGGAAGAGCACACGTCTGAACTCCAGTCACnnnnnnATCTCGTATGCCGTCTTCTGCTTG									

Table 5. Read-Depth Summary, TruSeq Adapter-Derived Kmers

For each library, the read depth for the length5 kmers in positions 1-9 is presented. The sequence of the TruSeq indexed adapter is shown at the bottom of the table. With few exceptions (libraries 352 & 303) the first 14 nucleotides of the adapter (underlined) could readily be reconstructed from the FastQC Kmer Content analysis. The ‘nnnnnn’ within the TruSeq adapter strand indicates the position of the variable indexing tags.

Figure 11. Library *vs* Adapter kmer Read-Depth

The maximum read depth for each TruSeq-derived kmer is plotted against total read depth for each library. With the exception of three libraries (352, 303 & 374) the kmer read depths linearly reflect the library read depths (linear fit line for these libraries in red).



Isolate_Runorder	352_615	1940_619	454_616	492_617	2999_620	535_624	1723_618	374_622	303_623	3016_621
Universal	9,023	25,567	35,761	24,527	4,190	21,892	12,563	332,566	220,188	15,620
- **5' of read**	6,799	18,470	24,767	15,908	938	14,804	8,329	315,038	205,544	5,151
- **5’ (%)**	75.4	72.2	69.3	64.9	22.4	67.6	66.3	94.7	93.3	33.0
Indexed	38,768	84,732	195,790	100,770	16,262	294,672	31,384	880,027	552,361	162,567
- **5' of read**	37,165	79,343	188,461	94,335	14,221	289,314	28,308	866,550	537,093	156,029
- **5’ (%)**	95.9	93.6	96.3	93.6	87.4	98.2	90.2	98.5	97.2	96.0
Total library reads:	9,918,842	3,768,246	10,040,306	5,109,018	2,194,204	8,121,702	992,154	5,720,698	8,985,616	3,085,900
Total Adapter reads:	47,791	110,299	231,551	125,297	20,452	316,564	43,947	1,212,593	772,549	178,187
As %s	0.48	2.93	2.31	2.45	0.93	3.90	4.43	21.20	8.60	5.77

Table 6. Adapter Read Position Summary

The universal and the indexed TruSeq adapter sequences were counted for each library. Shown for each adapter is the total number of reads including the adapter, the number having the adapter aligned with the start of the read, and the corresponding percentage of the total. Also shown are the total library reads (from table 2) and the percentage of that total which the adapter-containing reads comprise.

3.2 Library pre-processing

On the basis of the FastQC analysis, libraries were pre-processed to remove sequence contaminants. Pre-processing using programs from the FastX package was unsuccessful; in particular the clipper program was found to clip input sequences at positions entirely unrelated to the adapter string provided on the command line (not shown). Email contact with the package author (Assaf Gordon (gordon@cshl.edu) indicated that use of these programs for 105nuc read length libraries was inappropriate ("The clipper program is quite old (3 years old now), and was not designed to properly handle long reads (when it was developed, 36nt was a norm)."). No documentation of this limitation could be identified

3.2.1 Adapter removal with cutadapt

Reiterative cycles of library pre-processing using cutadapt (Martin, 2011) followed by FastQC analysis were used to identify suitable pre-processing parameters. After 17 iterations a parameter set which yielded acceptable outputs for 7 of the 10 libraries were identified. The final parameters used are detailed in Appendix A (Pre-processing shell script).

3.2.2 FastQC analysis

Table 7. FastQC analysis, pre-processed (summary) shows the outcome of the final pre-processing steps. From a library perspective, and without exception, the sum quality scores depicted in the columns of the table showed all to be improved by the final pre-processing conditions (median improvement +6.5). Improvement was least (+1) for library SA352, used in many figures above as an 'entirely normal' exemplar, and greatest (+8, +9) for the two libraries used as 'extremely unusual' exemplars, SA3016 & SA374.

From a metric perspective, four remained static (at 30: 'Basic Stats', 'Per Sequence Quality Score' & 'Per Base N Content', and at zero, 'Per Base Sequence Content'), one diminished ('Sequence Length Distribution'), and the remainder all improved. Modest improvements (+1, +2) to ‘Per Base GC Content’ and ‘Per Base Sequence Content’, intermediate improvements (+12, +13) to ‘Per Base sequence Quality’ and ‘Per Sequence GC’, and substantial improvements (+26, +24) to ‘Over-Represented Sequences’ and ‘Kmer Content’ metrics usefully reflect the libraries’ sequence quality issues.

Metric 1 for pre-processed data is presented in Table 8. Summary of Basic Statistics (raw & pre-processed). These data show that pre-processing removed between 73 thousand and 2.5 million, or between 1.8 and 44% of library reads. The reported GC figures for the pre-processed libraries converge with the expected value for MRSA reference sequences (Holden et al., 2004). FastQC reported all libraries to be ‘entirely normal’ by this measure (table 1p).

Metric 3 is presented in Figure 12. Per Sequence Quality Scores (pre-processed). Pre-processing has removed the profile irregularities indicating sequence contaminants. FastQC reported all libraries to be ‘entirely normal’ by this measure (table 1p).

Metric 7 (not shown), ‘Per base n content’, depicts ambiguous calls against read position FastQC considered all libraries to be ‘entirely normal’ by this measure (table 1p), a call unchanged by the pre-processing.

Metric 8 (not shown), ‘Sequence Length Distribution’, is the sole metric for which diminished scores resulted from pre-processing; all libraries were reported as length 11-105, which FastQC considered ‘slightly abnormal’. This outcome reflects the minimum sequence length parameter selected for cutadapt during pre-processing.

Metric 2 is presented in Figure 13. Per Base Sequence Quality Score (pre-processed), and improved from 18 to 30. All libraries exceed quality thresholds at all read lengths.

Metric 6 is presented in Figure 14. Per Sequence GC Content (pre-processed), and improved from 7 to 20. The departures from an N distribution exhibited by the raw libraries have largely been removed by pre-processing. Residual kurtosis indicates some persistent sequence contamination.

Metric 5 is presented in Figure 15. Per Base GC Content (pre-processed), and improved modestly from 5 to 6. The deviations from the expected profile exhibited by the raw libraries have largely been removed.

Metric 10, ‘Over-represented sequences‘ (not illustrated), improved from 4 to 30, and is the metric improved most by the pre-processing. This improvement reflected the removal of adapter sequences from the libraries.

Metric 9 is presented in Figure 16. Sequence Duplication (pre-processed, 1) and Figure 17. Sequence Duplication (pre-processed, 2). This metric was improved but modestly, from 1 to 3, by pre-processing. The figures show sequence duplication profiles essentially unchanged by removal of adapter sequences from the libraries.Table 9. Sequence Duplication Levels (raw & pre-processed). The modest effect of adapter removal on overall sequence duplication levels is confirmed by Table 9. Sequence Duplication Levels (raw & pre-processed).

Metric 4 is presented in Figure 18. Per Base Sequence Content (pre-processed). Overall this metric remained unchanged by pre-processing despite the improvements visible in the figure. This outcome reflects the persistence of non-random base distribution, which is visible at the start of the library consensus reads in the figure.

Metric 11 is presented in Figure 19. Kmer Content (pre-processed). Pre-processing improved this metric from 0 to 24. For most (n=7) libraries, no kmer content was reported by FastQC after pre-processing; for three libraries (SA303, SA352 & SA535), enriched kmer content persisted. Examination of the kmers reported showed that with a single exception they did not derive from the Illumina adaptors used in library construction. The origin of these library contaminants is not known.

3.2.3 Sequence duplication (confirmation)

The FastQC analysis after pre-processing indicated high sequence duplication within all libraries. This persistent duplication was independently confirmed using the ‘collapser’ program from the FastX package. Table 10 shows that pre-processing had little effect on overall library redundancy. Both before and after pre-processing, sequence reiteration constituted between 10 and 50% of the library reads (median raw 24.9, pre-processed 23.7).

Isolate	SA1940	SA454	SA492	SA374	SA1723	SA352	SA2999	SA3016	SA303	SA535	score
_run index	_619	_616	_617	_622	_618	_615	_620	_621	_623	_624	of 30
Basic Statistics (1)	P	P	P	P	P	P	P	P	P	P	30
Per seq quality scores (3)	P	P	P	P	P	P	P	P	P	P	30
Per base N content (7)	P	P	P	P	P	P	P	P	P	P	30
Seq Length Distrib (8)	W	W	W	W	W	W	W	W	W	W	10
Per base seq quality (2)	P	P	P	P	P	P	P	P	P	P	30
Per seq GC content (6)	P	P	P	W	W	W	P	W	P	W	20
Per base GC content (5)	W	W	W	W	W	F	W	F	F	F	6
Overrepresented seq (10)	P	P	P	P	P	P	P	P	P	P	30
Seq Duplication Levels (9)	F	F	F	W	W	W	F	F	F	F	3
Per base seq content (4)	F	F	F	F	F	F	F	F	F	F	0
Kmer Content (11)	P	P	P	P	P	P	W	P	W	w	24
total of 33	23	23	23	22	22	21	21	20	20	18	

Table 7. FastQC analysis, pre-processed (summary)

The output of FastQC was assigned quality integers sorted horizontally to place the highest-scoring library on the left, as in Table 2. FastQC analysis, raw data (summary). The columns of the table have been sorted as in table1. The rows of the table however are as in Table 1 to aid comparison with the raw data & highlight those metrics changed by the pre-processing. (The quality sort assigned integers to each of the three output terms (Pass = 3, Warn=1, Fail=0), summed these figures both across columns and down rows, and then sorted on the column totals shown). The numbers in parentheses after each QC metric title indicate the order of metric presentation by FastQC.

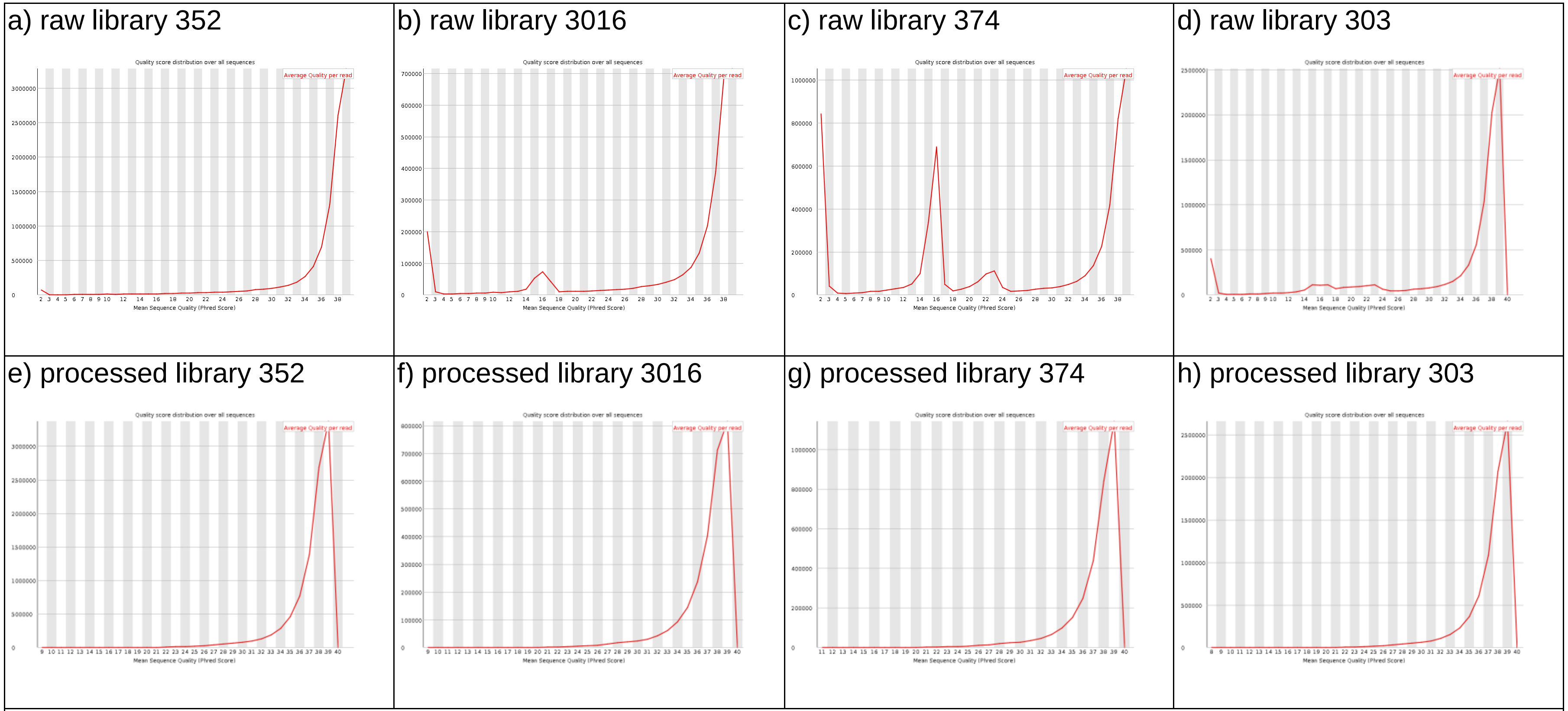
Isolate_ runorder	**1940_ 619**	**454_ 616**	**492_ 617**	**374_ 622**	**1723_ 618**	**352_ 615**	**2999_ 620**	**3016_ 621**	**303_ 623**	**535_ 624**
FastQC version	0.9.4	0.9.4	0.9.4	0.9.4	0.9.4	0.9.4	0.9.4	0.9.4	0.9.4	0.9.4
Call	pass	pass	pass	pass	pass	pass	pass	pass	pass	pass
File type	Conventional base calls	=	=	=	=	=	=	=	=	=
Encoding	Illumina 1.5	=	=	=	=	=	=	=	=	=
Total Reads(raw)	3,768,246	10,040,306	5,109,018	5,720,698	992,154	9,918,842	2,194,204	3,085,900	8,985,616	8,121,702
Total Reads(proc)	3,528,834	9,452,484	4,826,166	3,201,430	851,668	9,742,788	2,121,134	2,647,486	7,651,184	7,343,372
Removed Reads	239,412	587,822	282,852	2,519,268	140,486	176,054	73,070	438,414	1,334,432	778,330
“ as %	6.4	5.9	5.5	44	14.2	1.8	3.3	14.2	14.9	9.6
Read length	10-105	10-105	10-105	10-105	10-105	10-105	10-105	10-105	10-105	10-105
%GC, raw, proc	33, 31	33, 32	33, 32	41, 32	33, 31	32, 32	33, 33	37, 34	35, 32	36, 34

Table 8. Summary of Basic Statistics (raw & pre-processed)

‘=’ indicates entry identical to that to the left.

Isolate_ runorder	**1940_ 619**	**454_ 616**	**492_ 617**	**374_ 622**	**1723_ 618**	**352_ 615**	**2999_ 620**	**3016_ 621**	**303_ 623**	**535_ 624**
Raw	~55%	~82%	~63%	~68%	~30%	~80%	~52%	~63%	~80%	~82%
Pre-processed	~50%	~79%	~59%	~47%	~23%	~78%	~50%	~56%	~74%	~80%

Table 9. Sequence Duplication Levels (raw & pre-processed).



a) raw library 352

b) raw library 3016

c) raw library 374

d) raw library 303

e) processed library 352

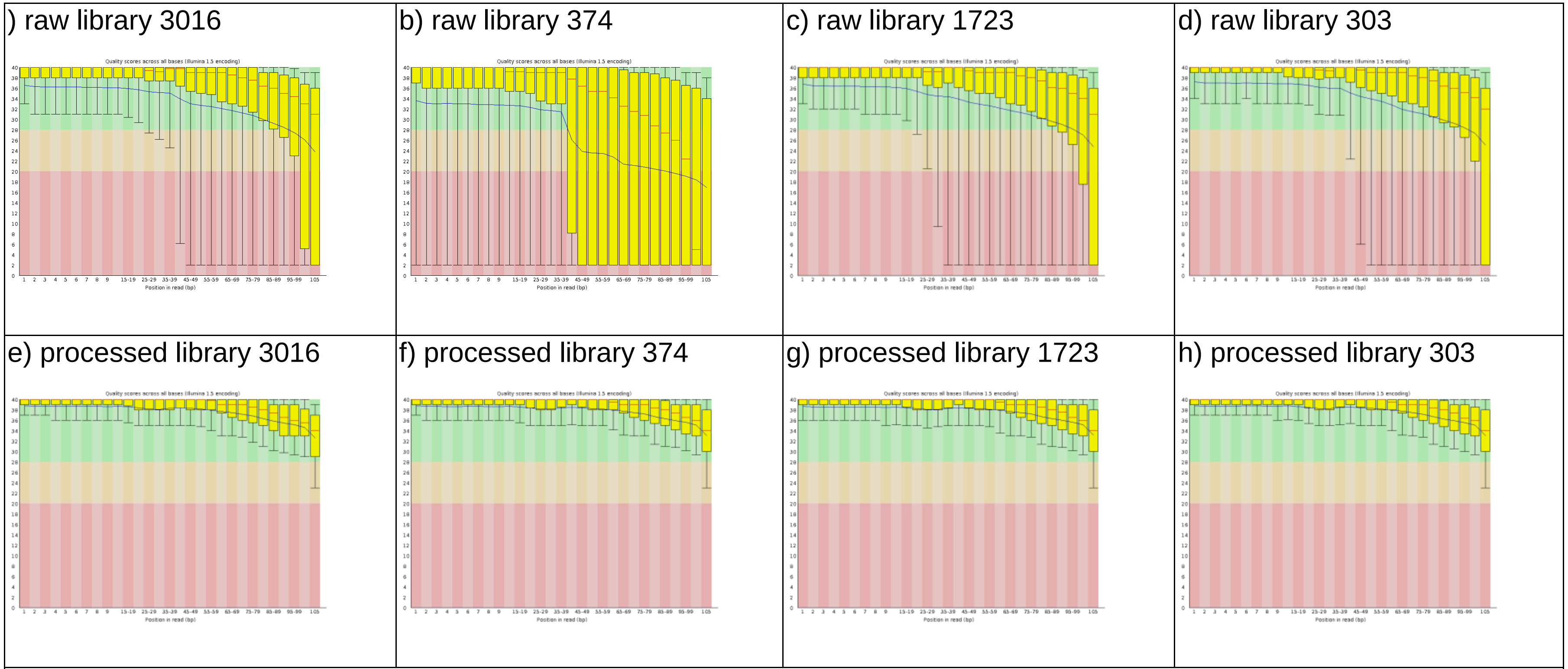
f) processed library 3016

g) processed library 374

h) processed library 303

Figure 12. Per Sequence Quality Scores (pre-processed)

Depiction of the read average Qscore distribution over all sequences for libraries 352, 3016, 374 and 303 before (top) and after (bottom) pre-processing.



) raw library 3016

b) raw library 374

c) raw library 1723

d) raw library 303

e) processed library 3016

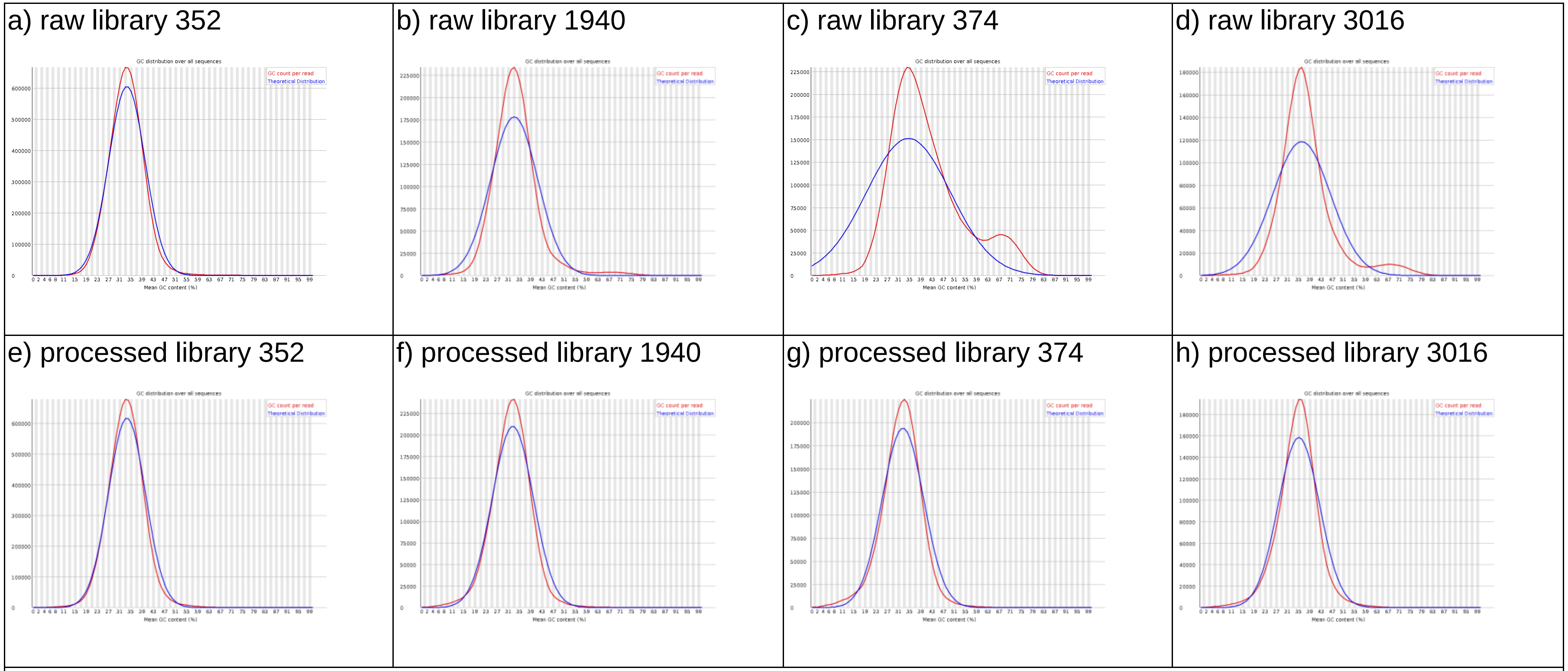
f) processed library 374

g) processed library 1723

h) processed library 303

Figure 13. Per Base Sequence Quality Score (pre-processed)

Comparison between those libraries rated as 'highly unusual' by FastQC when unprocessed (top row) and after pre-processing (bottom row).



a) raw library 352

b) raw library 1940

c) raw library 374

d) raw library 3016

e) processed library 352

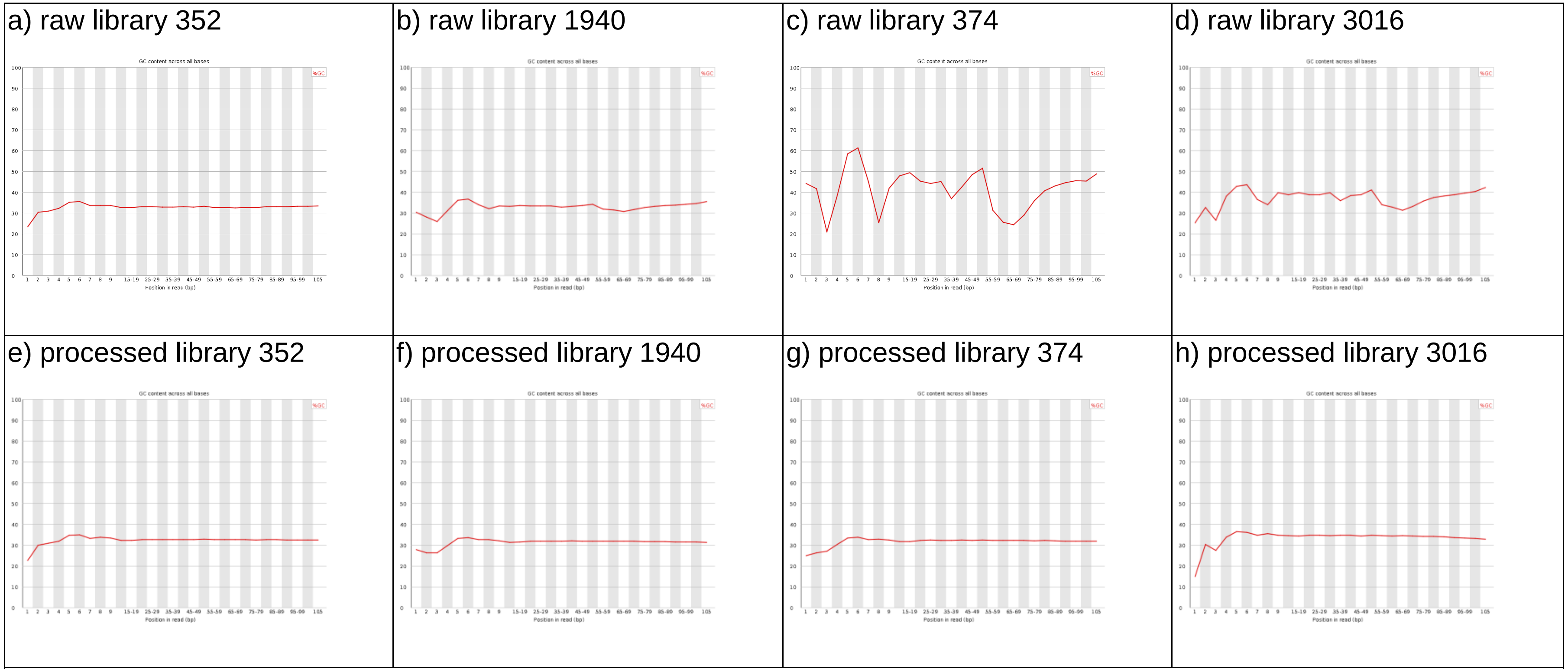
f) processed library 1940

g) processed library 374

h) processed library 3016

Figure 14. Per Sequence GC Content (pre-processed)

Depiction of observed GC distribution over all reads (red) relative to a Normal distribution with the same mean and standard deviation as the library (blue), before (top) and after (bottom) pre-processing for selected libraries. Quality random libraries superimpose on N. The persistent kurtosis indicates a low level of persistent library contamination. The reduction in skewness reflects removal of major library contaminants.



a) raw library 352

b) raw library 1940

c) raw library 374

d) raw library 3016

e) processed library 352

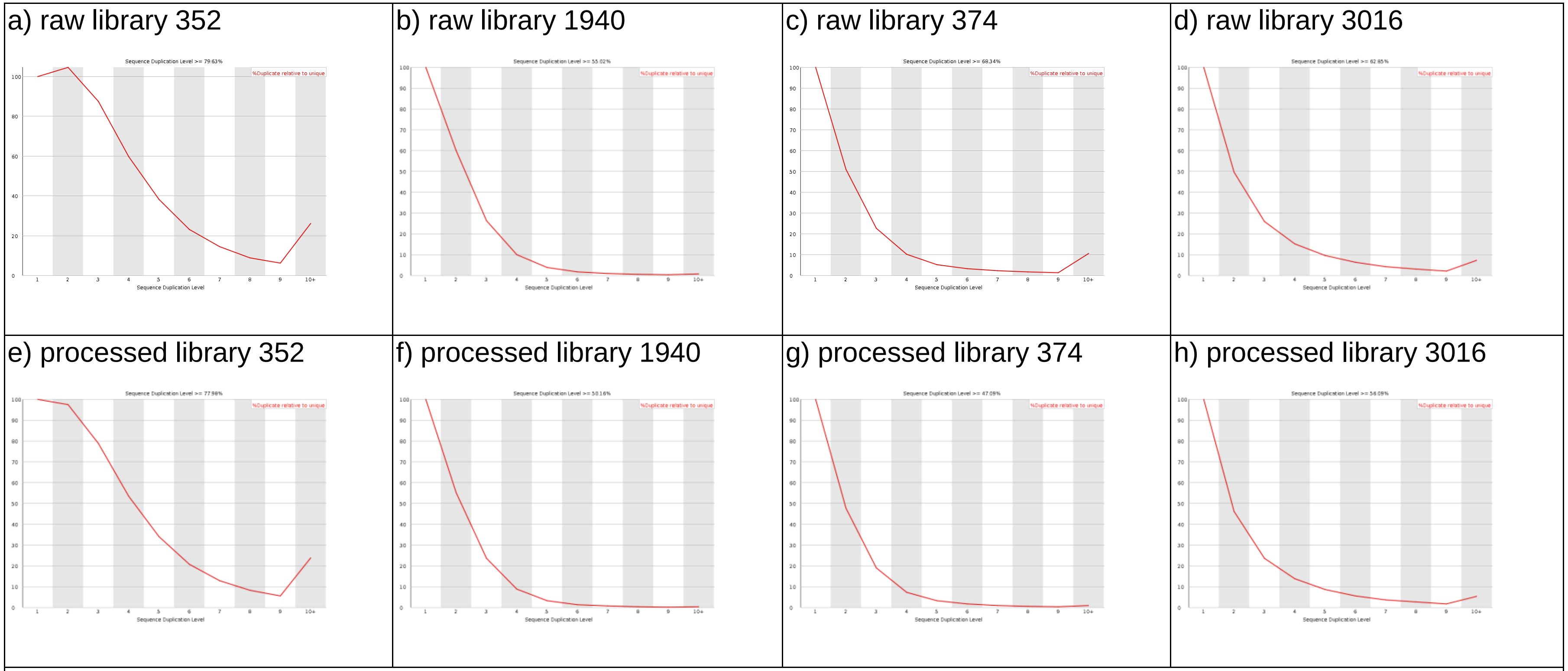
f) processed library 1940

g) processed library 374

h) processed library 3016

Figure 15. Per Base GC Content (pre-processed)

Depiction of GC frequency against read position before (top) and after (bottom) pre-processing for selected libraries. After pre-processing all libraries approach the profile of the top-ranked 352 library**.**



a) raw library 352

b) raw library 1940

c) raw library 374

d) raw library 3016

e) processed library 352

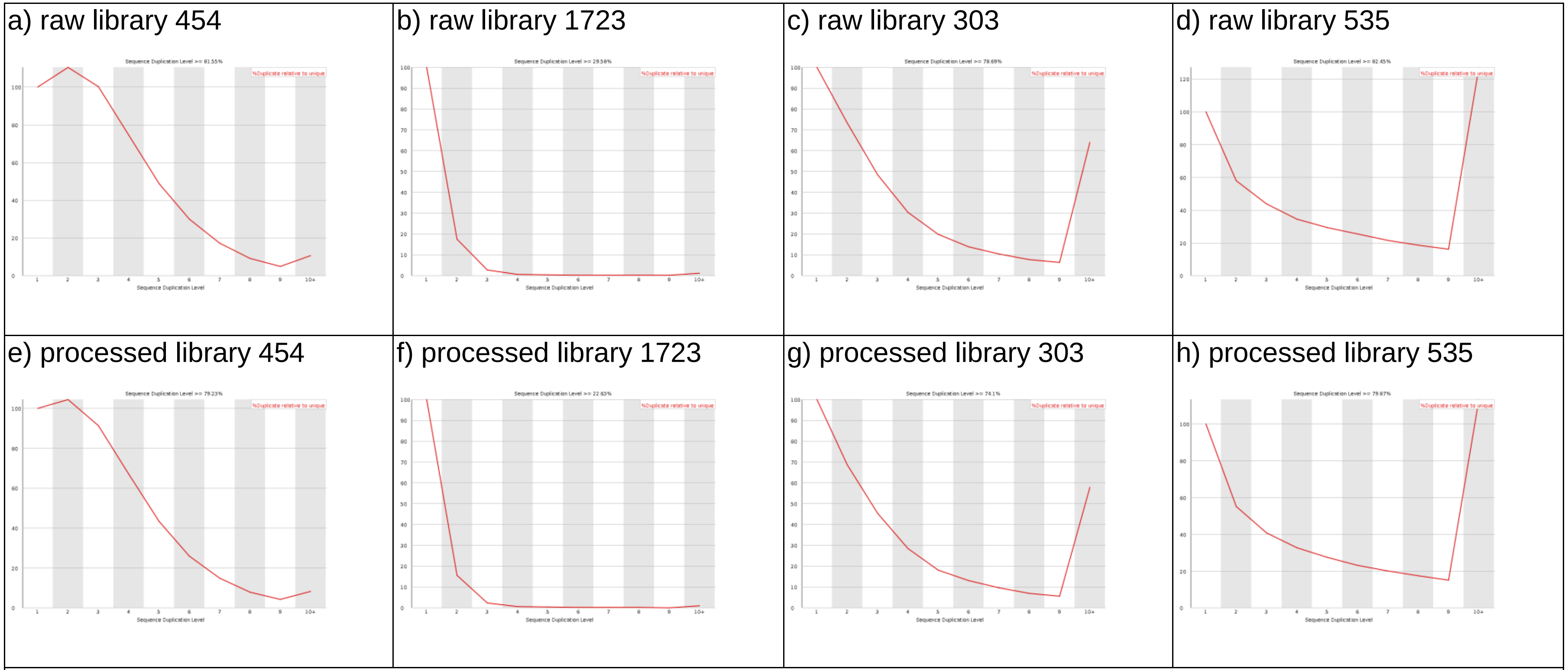
f) processed library 1940

g) processed library 374

h) processed library 3016

Figure 16. Sequence Duplication (pre-processed, 1)

The number of reads duplicated between 2 and >10 times is shown on a scale relative to the number of unique reads, which is set to 100% before (top) and after (bottom) pre-processing for selected libraries.



a) raw library 454

b) raw library 1723

c) raw library 303

d) raw library 535

e) processed library 454

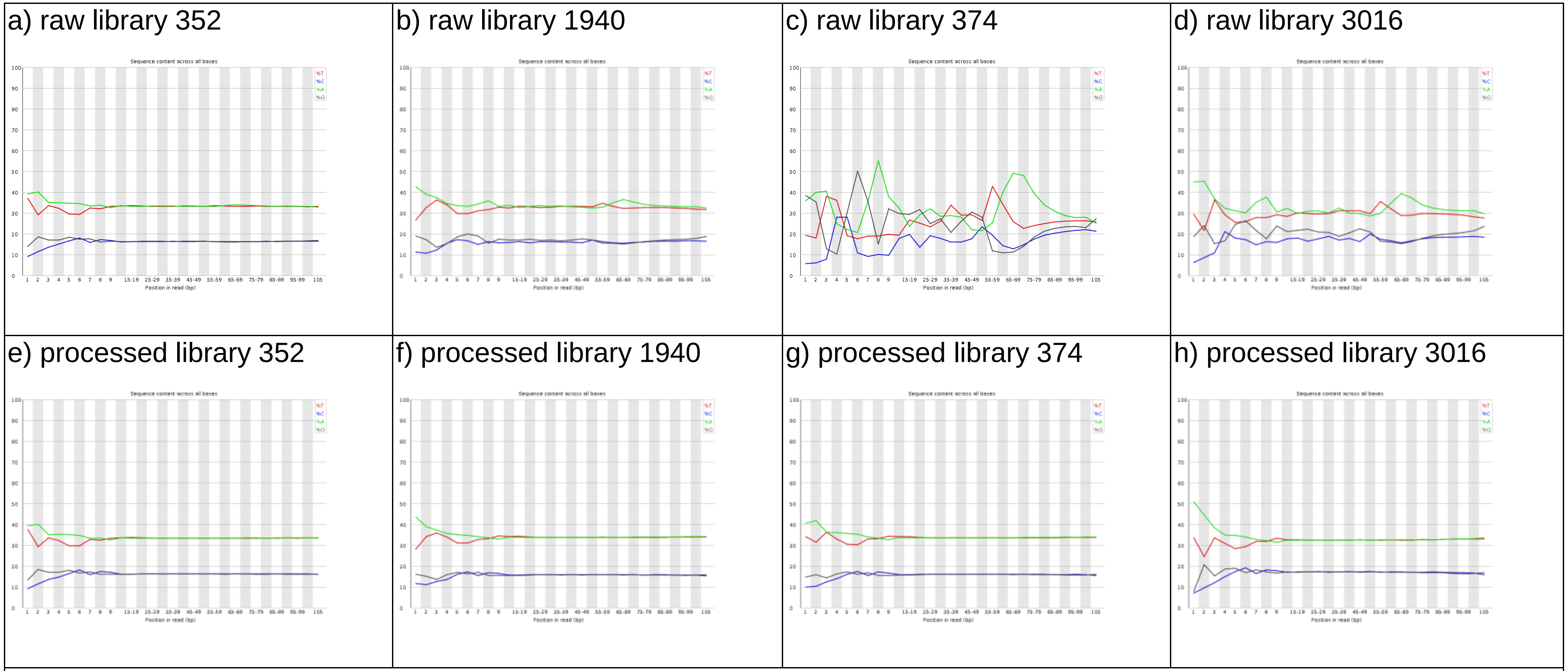
f) processed library 1723

g) processed library 303

h) processed library 535

Figure 17. Sequence Duplication (pre-processed, 2)

The number of reads duplicated between 2 and >10 times is shown on a scale relative to the number of unique reads, which is set to 100% before (top) and after (bottom) pre-processing for selected libraries.



a) raw library 352

b) raw library 1940

c) raw library 374

d) raw library 3016

e) processed library 352

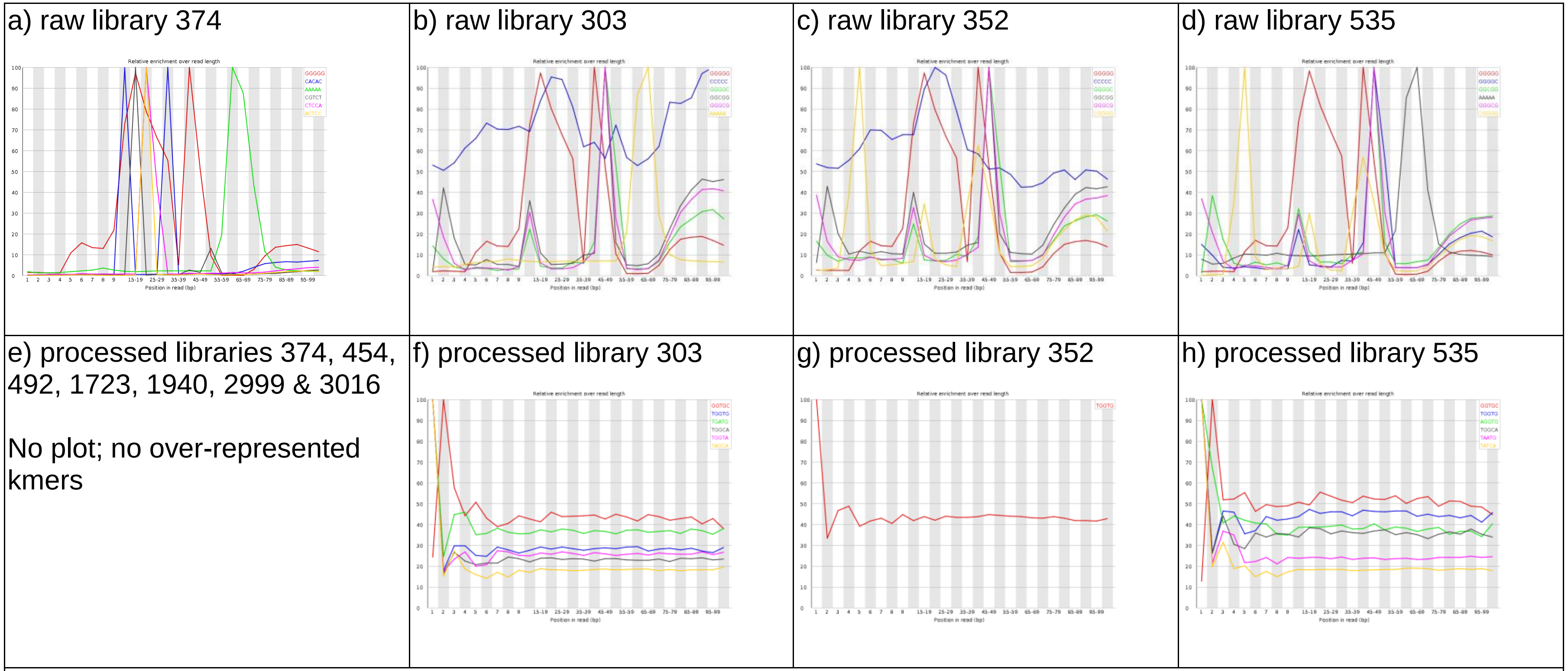
f) processed library 1940

g) processed library 374

h) processed library 3016

Figure 18. Per Base Sequence Content (pre-processed)

Depiction of sequence content for all four bases against read position before (top) and after (bottom) pre-processing for selected libraries.



a) raw library 374

b) raw library 303

c) raw library 352

d) raw library 535

e) processed libraries 374, 454, 492, 1723, 1940, 2999 & 3016

No plot; no over-represented kmers

f) processed library 303

g) processed library 352

h) processed library 535

Figure 19. Kmer Content (pre-processed)

Depiction of the relative enrichment of length 5 Kmers against read length before (top) and after (bottom) pre-processing. Most libraries (e) were freed of enriched kmers. Kmers from unknown sources other than adapter sequences persisted in three libraries, 303 (b, f), 352 (c, g) and 535 (d, h).

Isolate_ runorder	**1940_ 619**	**454_ 616**	**492_ 617**	**374_ 622**	**1723_ 618**	**352_ 615**	**2999_ 620**	**3016_ 621**	**303_ 623**	**535_ 624**
raw, total	3,768,246	10,040,306	5,109,018	5,720,698	992,154	9,918,842	2,194,204	3,085,900	8,985,616	8,121,702
raw, uniq	2,949,184	5,811,409	3,727,195	5,026,778	904,292	5,261,068	1,696,639	2,442,032	5,475,874	4,168,566
uniq as %	78.3	57.9	73.0	87.9	91.1	53.0	77.3	79.1	60.9	51.3
proc, total	3,528,834	9,452,484	4,826,166	3,201,430	851,668	9,742,788	2,121,134	2,647,486	7,651,184	7,343,372
proc, uniq	2,786,719	5,604,783	3,579,485	2,572,501	769,815	5,474,738	1,673,777	2,075,424	4,396,427	3,685,897
uniq as %	79.0	59.3	74.2	80.4	90.4	56.2	78.9	78.4	57.5	50.2

Table 10. FastX_collapser: sequence redundancy

Total raw and pre-processed libraries were input to FastX_collapser; shown are the total read numbers, unique read numbers, and the percent of total which unique reads constitute for each of raw and pre-processed libraries.

3.3 Assembly and analysis

The sanitized libraries were assembled de novo using the de Bruijn graph assembler Velvet (Zerbino and Birney, 2008), via the wrapper script Velvet Optimizer (Gladman, 2011). No library subsetting was done for initial assembly, which yielded output reports summarised in Table 11. Velvet assembly summary. The optimal overlap length (kmer) for all initial assemblies was found to be 69 nucleotides or more. The total coverage ranged between 8 (SA1723) and 82 (SA352), total contig number between 1317 and 120 and n50 (the optimisation criterion defined to Velvet Optimiser) between 3,523 and 78,679. The cumulative length of all assemblies ranged 94 to 102% of reference length, and with the exception of library SA1723 these figures changed little when contigs shorter than 1kb were excluded.

3.3.1 Assembly appraisal

The output of Velvet was appraised in three ways.

Velvet requires coverage sufficient to allow short, low-coverage contigs resulting from mis-assembly to be resolved from the longer, higher quality contigs generated by accurate assembly. The ability of each library to support this resolution was determined using test assemblies. Test assemblies were made using the kmer length identified by Velvet optimiser and a coverage-cutoff of zero, a value which prevents ‘rescue’ of reads mis-assembled into low-coverage contigs during Velvet assembly. Plots of contig coverage for each test assembly were used to visualise the resolution available to Velvet. Most libraries were found to provide adequate coverage for quality assembly (Figure 20), with even the intermediate coverage libraries SA3016 and SA2999 (Figure 21) providing adequate resolution between short, low-coverage, likely erroneous contigs and the remainder of the assembly. The library with lowest coverage, SA1723, provided lesser resolution for high quality assemblies (Figure 22).

The Velvet contigs were compared to MRSA252 across a core genome region of 180kb starting at 2.612 megabases, a region less subject to large scale divergence than those including elements of the accessory genome. The contigs for two isolates, SA352 and SA492, appeared virtually identical to MRSA252 (Figure 23). The contigs for other isolates diverged from MRSA252; those for isolates SA2999, SA3016, SA303 and SA535 suggested a common departure from MRSA252 (Figure 23), as did those for isolates SA374 and SA454 (Figure 24); those for isolates SA1940 and SA1723 were individually distinct from MRSA252, with that of SA1723 particularly fragmented, most probably on account of the limited coverage this library provided.

For isolates which diverged from MRSA252, blastn searches using the divergent single contig which spans ~30-40k in figures 23 & 24 identified more closely-related database entries. Isolates SA2999, SA3016, SA303 and SA535 appeared closely related to MSSA476 (Figure 25), a community-acquired meticillin-sensitive strain for which the complete genome is available (Holden et al., 2004). The divergent contigs from isolates SA374 and SA454 matched over their 21kb length a contig from a non-assembled WGS *S.aureus* project (Jones, 2012). In the absence of an assembled genome for this project, a preliminary comparison of the non-assembled contigs from all three isolates with MRSA252 shows that the three diverge similarly from the reference (Figure 26). Isolate SA1940 was found to be closely related to Mu50 (Figure 27), a vancomycin-resistant strain for which the complete genome is available (Cui et al., 2009). No comparable database entry was identified for SA1723, possibly on account of the highly fragmented assembly for this isolate.

3.3.2 Preliminary SCC mapping against MRSA252

The contigs from preliminary Velvet assembly were mapped in two ways to the full-length genome from the reference strain MRSA252.

The blastn variant bl2seq was used, with default parameters, to gain an insight into the proportion of contigs which mapped to the reference; contigs mapping to the reference ranged from 48.6 to 87.7 % (median 84.6%; Table 12).

The mummer family program NUCmer was used, with default settings, to generate independent contig:reference mappings. Mapview was then used to generate graphical outputs, yielding a first impression of the SCC status of the isolates. Isolates SA352 and SA492 appear to share most of the MRSA252 SSC structure, but the remaining isolates have only fragmentary coverage across this region and the upstream flanking region (Figure 28).

3.3.3 Targeted SCC mapping

The homologies identified during assembly appraisal were used to make more targeted alignments of each isolates contigs to appropriate reference sequences.

3.3.3.1 MRSA252 : SA352 & SA492

The SCC of MRSA252 was isolated as a reference sequence and compared using nucmer to the contigs for SA352 and SA492 (Figure 30). Contigs from both isolates aligned with the entire SCC, including the mec operon, Tn554 and the ccr operon. The contigs of both isolates which included the ccr operon and the downstream remainder of the SCC diverged slightly from the reference. Likewise, the SA352 contig spanning the mec operon, and that of SA492 spanning Tn554 also diverged. In addition, both isolates included contigs aligning discontinuously with the reference, consistent with either a subpopulation with deletions removing much of the SCC, or a consistent miss-assembly. The striking ‘stack’ of divergent and interrupted contig alignments beneath the IS431-flanked pUB110 plasmid insert in each isolate reflects nucmer’s rendering of two types of contig. The first are simply short contigs (~150-250 nuc) consisting of a sublength of variously divergent IS431-like elements. The second are longer contigs which include intact duplications of these divergent IS431 elements separated by additional sequences (not shown). Both result from the presence in the assemblies for these isolates of plasmids bearing duplications of IS431-like elements. When run with the maxmatch parameter, nucmer displays all match positions between query contigs and the reference, and joins these match positions with colour-coded contig-specific lines. This rendering can give a false impression that a contig maps discontinuously across the reference. Blast database searches with contigs aligning in this way in the nucmer output confirmed that both isolates harboured plasmids shown to confer broadened drug resistance phenotypes and to include IS431-like elements (McDougal et al., 2010).

3.3.3.2 MSSA476 : SA2999, SA3016, SA303 and SA535

The SCC of MSSA476 was isolated as a reference and compared using nucmer to the contigs for isolates SA3016, SA2999, SA303 and SA535 (Figure 31). Isolate SA3016 included sequences matching all regions of the MSSA476 SCC, with 5 of 6 contigs matching the reference perfectly across their length (not shown). Isolate SA2999 included contigs diverging slightly from MSSA476 upstream and running into the SCC, with sequences homologous to the first CDS in the hsd operon, hsdR, not reported. Isolate SA303 included contigs largely identical to SCC, but upstream sequences diverged somewhat. In addition hybrid, discontinuously matching contigs with both highly similar and significantly divergent regions were also present. The contigs of isolate SA535 were most different from the reference, with the contigs over much of the length of SCC diverging slightly. In addition, and as for SA303, discontinuously matching contigs with more significantly divergent sequences were also present.

3.3.3.3 Mu50 & S0385 : SA1940

The contigs of isolate SA1940 were found to be similar to strain Mu50 in the core genome regions (Figure 27). This high similarity was found not to extend into the SCC region. To identify potential SCC-specific contigs from the SA1940 library, the library was screened against the SCC sequences from *S aureus* strains MRSA252 (Holden et al., 2004), MSSA476 (*Ibid*), Mu50 (Cui et al., 2009) and LGA251 (Garcia-Alvarez et al., 2011) using promer. The compiled hit contigs were then used to blastp against the *Staph aureus* subset of genbank. In addition to previous matches this identified a newly-accessible full genome sequence for strain S0385 (Schijffelen et al., 2010), a livestock-associated meticillin resistant isolate from an endocarditis patient, against which the SA1940 contigs were better matched. Comparison of 8 contigs with the S0385 SCC (Figure 32) showed that the upstream half of the SCC was largely represented by contigs within the SA1940 assembly, including the mec operon. Many contigs matched discontinuously, however, suggesting rearrangement either between the isolates or during assembly. Analysis of one such contig, which contributed both the leading and trailing matches to the S0385 SCC depicted in figure 32, suggested mis-assembly (Figure 33). The contig consisted of a fusion between sequences matching those at the start and at the center of the S0385 SCC, with these sequences transposed end-to-end within the contig. Such mis-assembly is a characteristic assembly error in regions with repeats when library depth is inadequate to provide reads which both span and ‘step into’ repeats. However, although this region of the S0385 is rich with repeats (Figure 34), none could be identified which could readily explain the structure of contig 115. If nothing else this outcome highlights the limitations of the representations generated by nucmer, which serve merely to indicate that the sequences within the contig set for an isolate include some which align, possibly only in part, with those of the reference; in the absence of a single continuously matching contig across the entire SCC they do not support conclusions about the overall SCC structure of an isolate. Figure 32 also shows the mapping of these contigs against Mu50, which confirms that the similarity observed across the core genome between SA1940 and this strain does not extend into the SCC sequences of the accessory genome.

3.3.3.4 IS-160 : SA374 & SA454

Contigs of isolates SA374 and SA454 were found (Figure 26) to be similar across the core genome to those of an unassembled WGS project for *S.aureus* strain IS-160 (Jones, 2012). The SCC-specific contigs of this project were identified using promer comparisons against the SCC of 12 *S. aureus* strains typifying the currently recognised SCC types (IWG-SCC, 2009); this identified the SCC of IS-160 as being highly similar to that of strain CA05 across regions specifying the mecA operon, but divergent thereafter (Figure 35a). Comparison of the contigs from SA374 and SA454 with CA05 SCC confirmed a similar profile for SA374 (Figure 35b) and indicated that SA454 was more closely similar still to SA05 (Figure 35c). Ordered concatenation of the SCC-specific contigs from IS-160 enabled comparisons with those from SA374 (Figure 35d) and SA454 (Figure 35e), which confirmed that the mecA region of both was less divergent than the remainder of the sequences aligning with the IS-160 SCC.

3.3.3.5 WIS : SA1723

The fragmented contig structure of isolate SA1723 most probably contributed to the failure of assembly validation alignments against a 180kb core genome region to identify consistent database entries giving confidence in this assembly. Across the shorter region of the SCC, contigs in the SA1723 assembly were identified using promer as above (3.3.3.4), which identified the SCC of strain WIS_JCSC3624 (Ito et al., 2004), for which no complete genome sequence is available, as being most similar to that of SA1723 across regions including the mecA operon (Figure 36). As for isolates SA352 and SA492, the SA1723 library included multiple sequences aligning to the IS431 elements of the SCC; these derived from plasmids carried by isolate SA1723 and which were previously described as conferring extended drug resistance phenotypes (McDougal et al., 2010).

3.3.4 mecA alignment

The mecA gene products of the reference sequences used here - CA05, IS-160, MRSA252, Mu50, JCSC3624_WIS, s0385 and LGA251 - were compared using Clustal (Figure 37). All but LGA251 proved highly related. The mecA peptide of strains MRSA252, CA05, IS-160 and Mu50 were identical. Those of s0385 and JCSC3624_WIS differed from this group by a single Serine to Arginine substitution at position 225. As has previously been reported (Garcia-Alvarez et al., 2011, Shore et al., 2011) LGA251 diverged from this otherwise common mecA sequence.

The mecA gene products from the reference sequences MRSA252, s0385 and LGA251 were compared with those from the isolates examined here using Clustal (Figure 38). As for the reference sequences above, all proved highly related with LGA251 as an outlier. The S225R substitution proved the sole point of divergence between the isolate mecA peptides. At this position isolates SA1723, SA1940 and SA374 cluster with the s0385 Arginine-specifying reference. Isolates SA492, SA454 and SA352 cluster with the Serine-specifying MRSA252 reference. The divergent LGA251 mecA specifies Threonine at this position and is otherwise divergent as previously reported (Garcia-Alvarez et al., 2011, Shore et al., 2011).

Isolate	Contigs (all)					Contigs > 1k				Velvet parameters		
	number	n50	max length	cumulative length	% of ref length	number	% of total	cumulative length	% of ref length	kmer	Coverage (Ck)	covCut
SA352	120	78,679	205,267	2,972,781	102.4	67	55.8	2,954,890	101.8	77	82	4.84
SA454	144	43,578	156,723	2,742,187	94.5	107	74.3	2,728,596	94	79	79	20.9
SA1940	163	36,460	88,198	2,872,374	99	133	81.6	2,860,912	98.6	77	29	6.46
SA303	168	39,846	228,825	2,756,535	95	112	66.7	2,742,254	94.5	77	62	5.27
SA535	173	37,799	231,421	2,740,314	94.4	116	67.1	2,723,186	93.8	77	64	0.93
SA374	185	31,655	145,270	2,812,581	96.9	145	78.4	2,799,947	96.5	77	27	8.07
SA492	220	33,966	134,774	2,974,228	102.5	148	67.3	2,952,644	101.7	73	45	0.29
SA3016	256	27,774	106,020	2,827,503	97.4	162	63.3	2,799,688	96.5	69	25	7.37
SA2999	309	23,388	81,262	2,854,771	98.4	199	64.4	2,823,502	97.3	69	21	1.87
SA1723	1317	3,523	15,618	2,817,869	97.1	821	62.3	2,558,150	88.1	69	8	3.73

Table 11. Velvet assembly summary

The outcome of optimising the assembly of isolates using Velvet Optimiser. Complete library read sets were provided to Velvet Optimiser (no library subsetting was used). Maximal n50 was defined as the optimisation criterion for Velvet Optimiser. Coverage is expressed in kmer coverage terms, Ck, defined as Ck=C (L – k +1)/L where C is nucleotide coverage, L is read length & k is kmer length.

Isolate	CN number	CN bl2seq hits to reference	Hits as % of CN number
SA352	120	105	87.5
SA454	144	122	84.7
SA1940	163	142	87.1
SA303	168	142	84.5
SA535	173	147	85
SA374	185	152	82.2
SA492	220	193	87.7
SA3016	256	214	83.6
SA2999	309	247	79.9
SA1723	1317	640	48.6

Table 12. Blastn hits to reference

The number of contigs from Velvet asssmblies, and their percentage of total contigs numbers, mapping to the reference sequence by blastn.

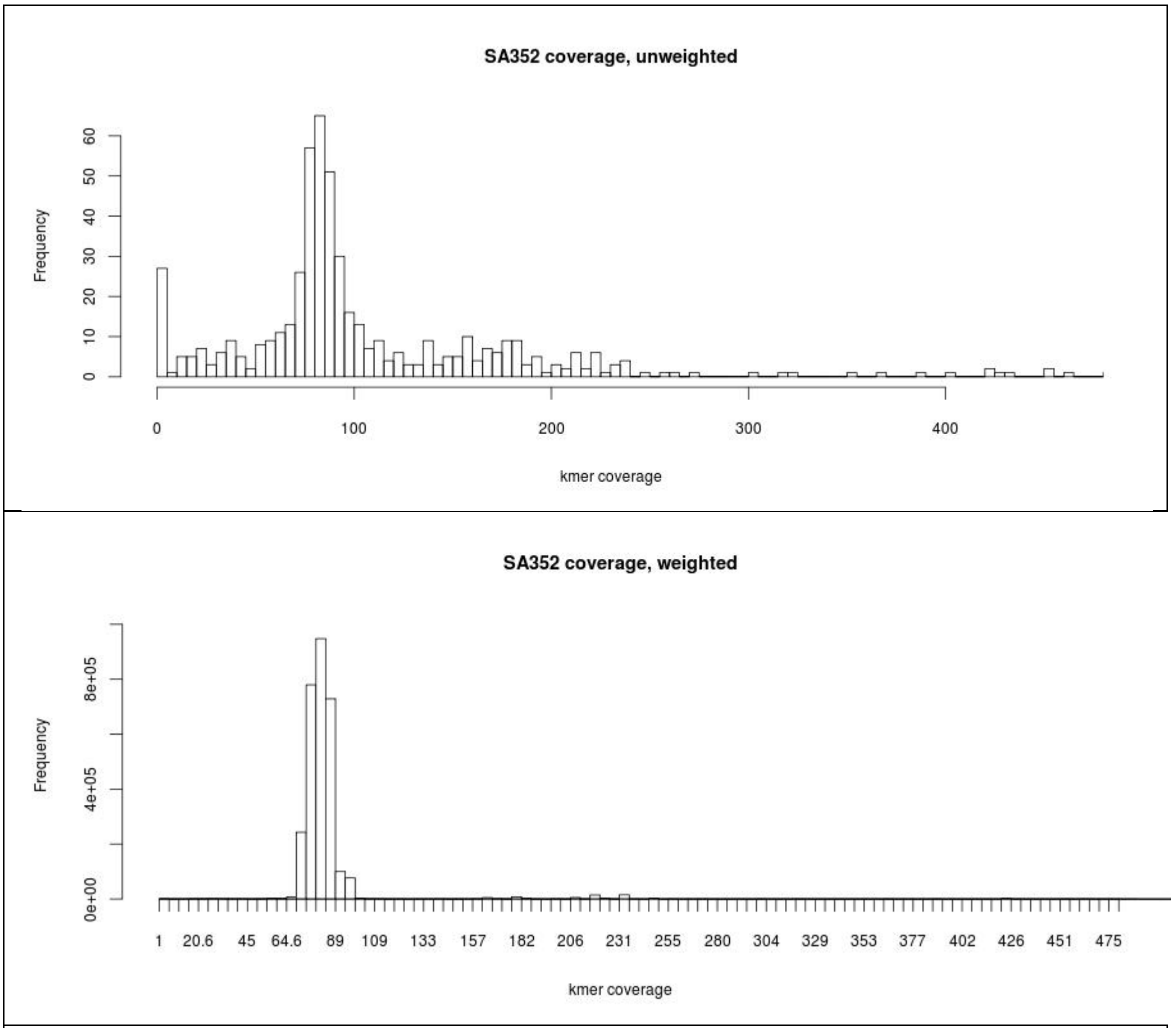


Figure 20 Assembly appraisal – high coverage libraries

Contig coverage distribution for SA352, an exemplar high coverage library, unweighted (top), and weighted against contig length (bottom).

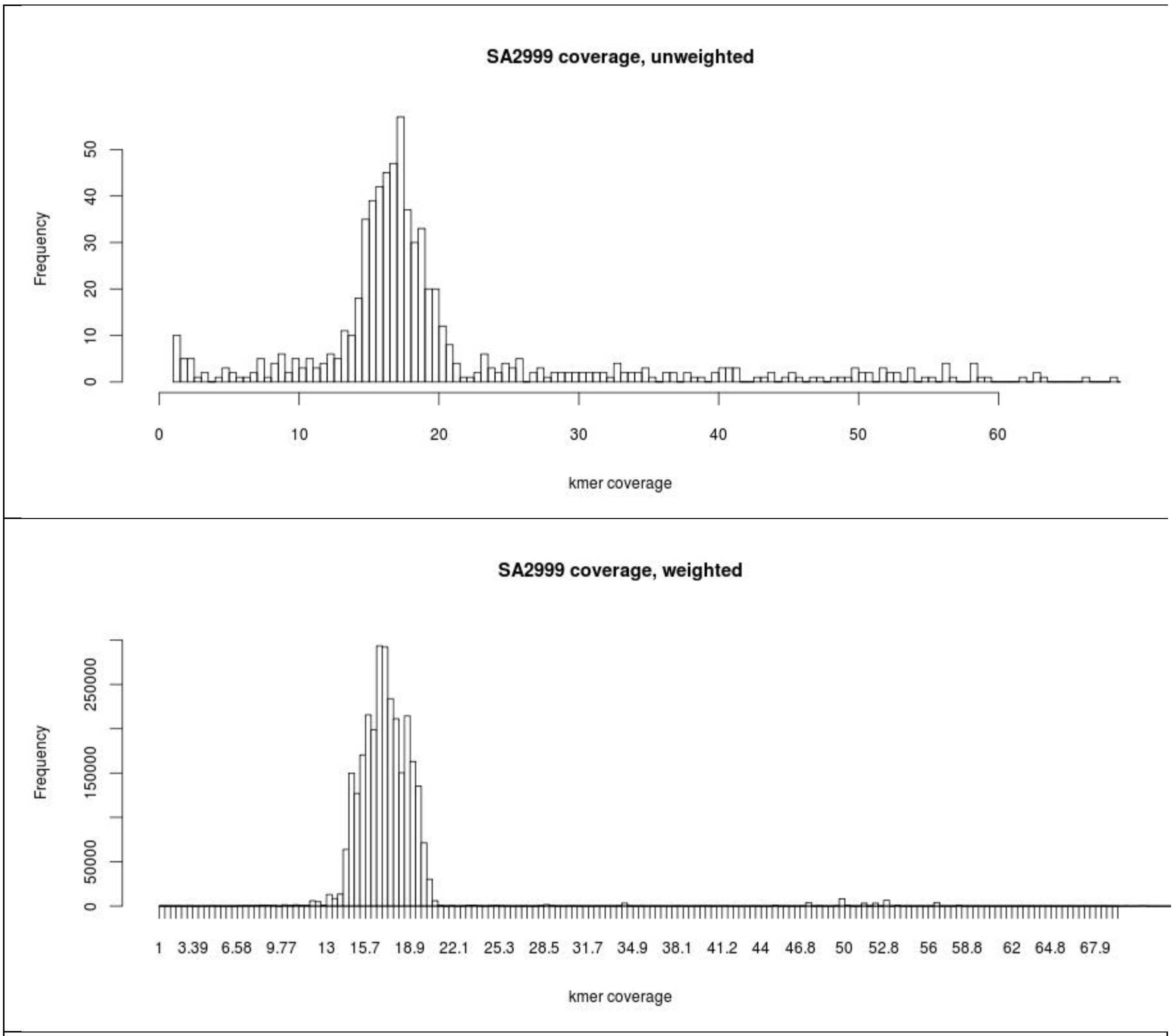


Figure 21. Assembly appraisal – intermediate coverage libraries

Contig coverage distribution for SA2999, exemplar intermediate coverage library, unweighted (top), and weighted against contig length (bottom).

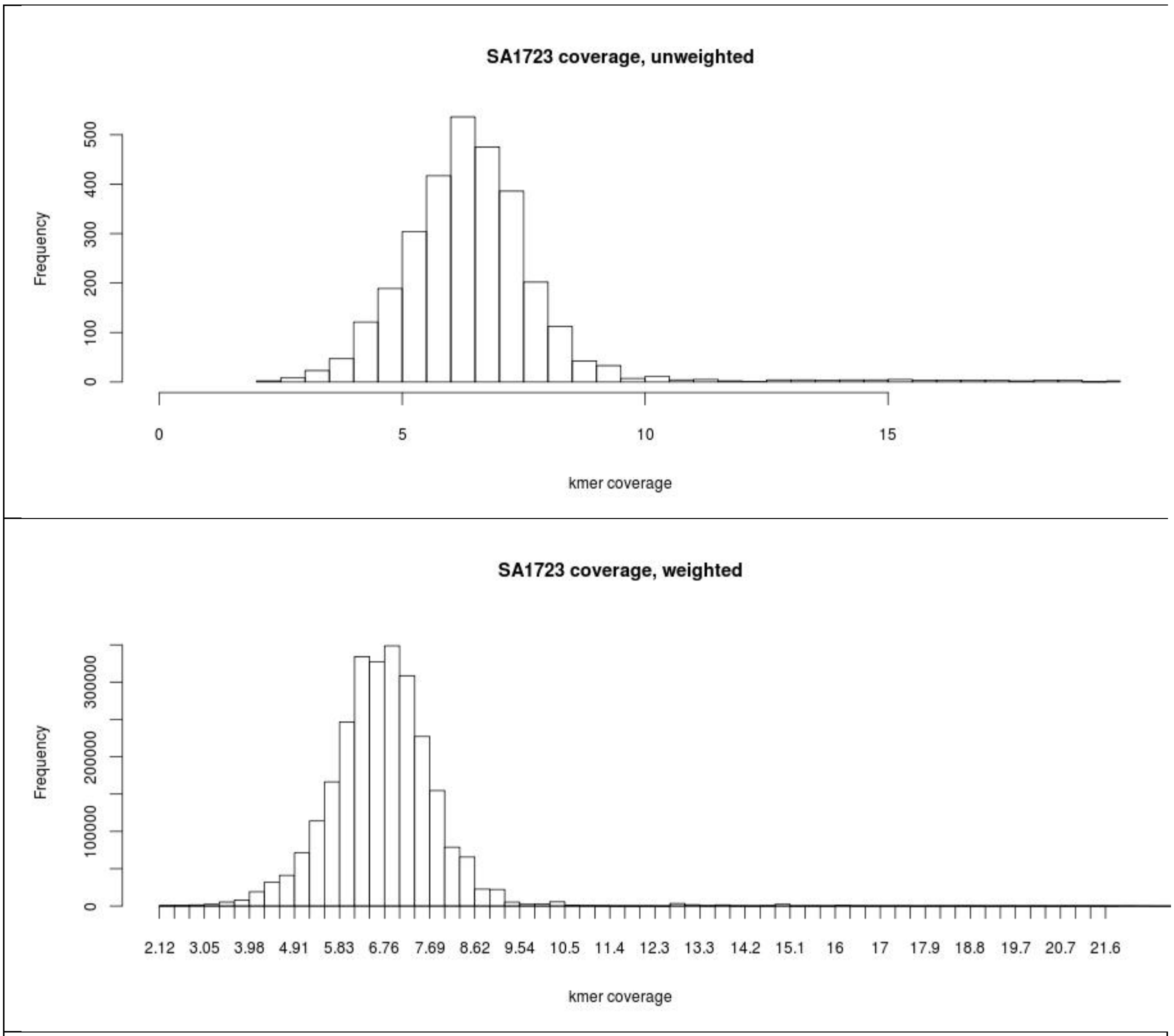


Figure 22. Assembly appraisal – low coverage library

Contig coverage distribution for SA1723, unweighted (top), and weighted against contig length (bottom).

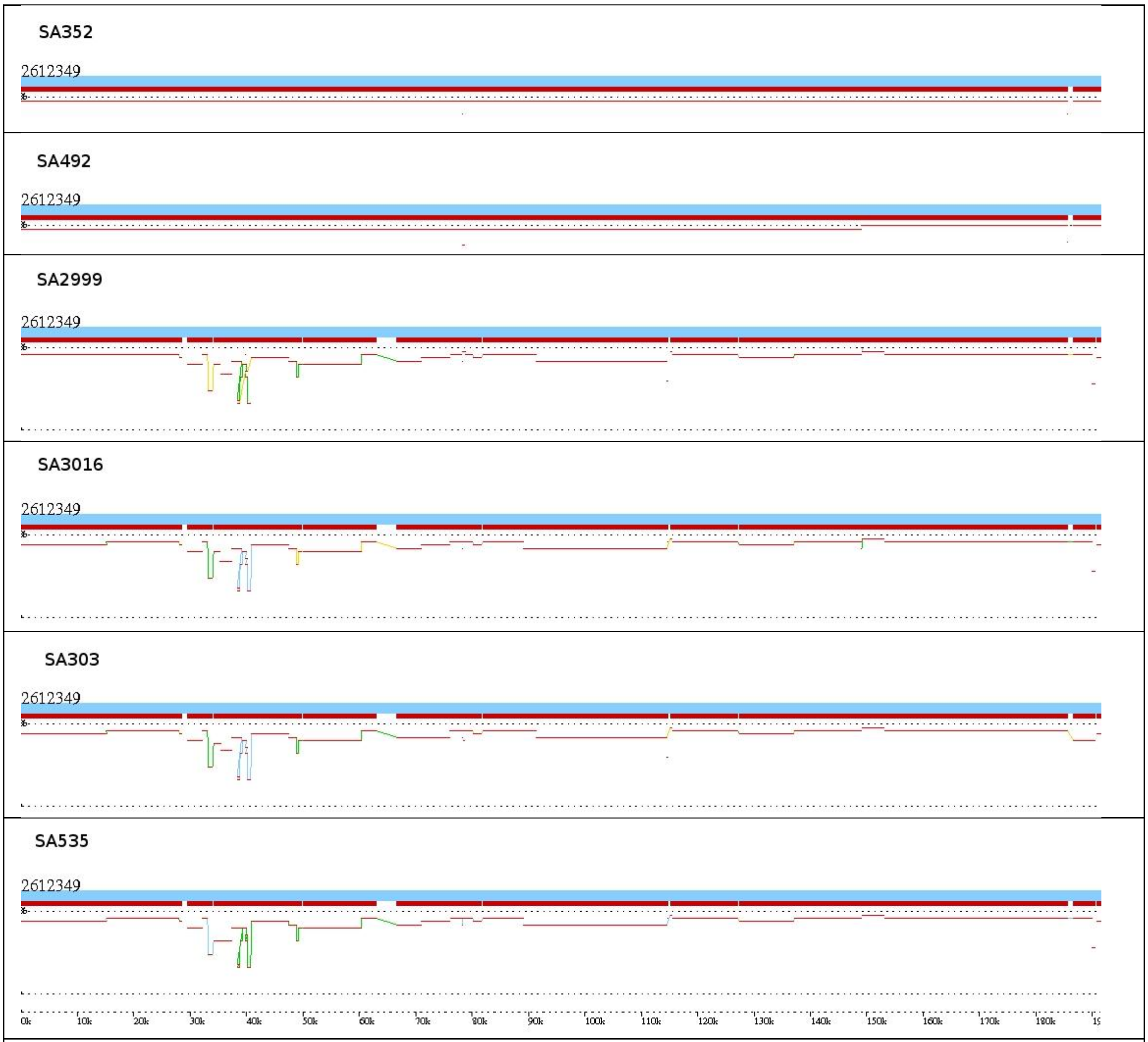


Figure 23. Assembly appraisal – core genome

Comparison of isolates with MRSA252 across 180kb of core genome starting at 2.612 megabases. SA352 and SA492 appear closely related to the reference. SA2999, SA3016, SA303 and SA535 are closely-related and apparently have a common divergence from the reference.

.

Bold elements represent the reference (blue) and the regions of reference matched by contigs (red). The lower, thinner elements represent the individual contigs, with distance below the reference representing sequence divergence. Elements of contigs matching discontinuously are connected by lines coloured to distinguish individual contigs. .

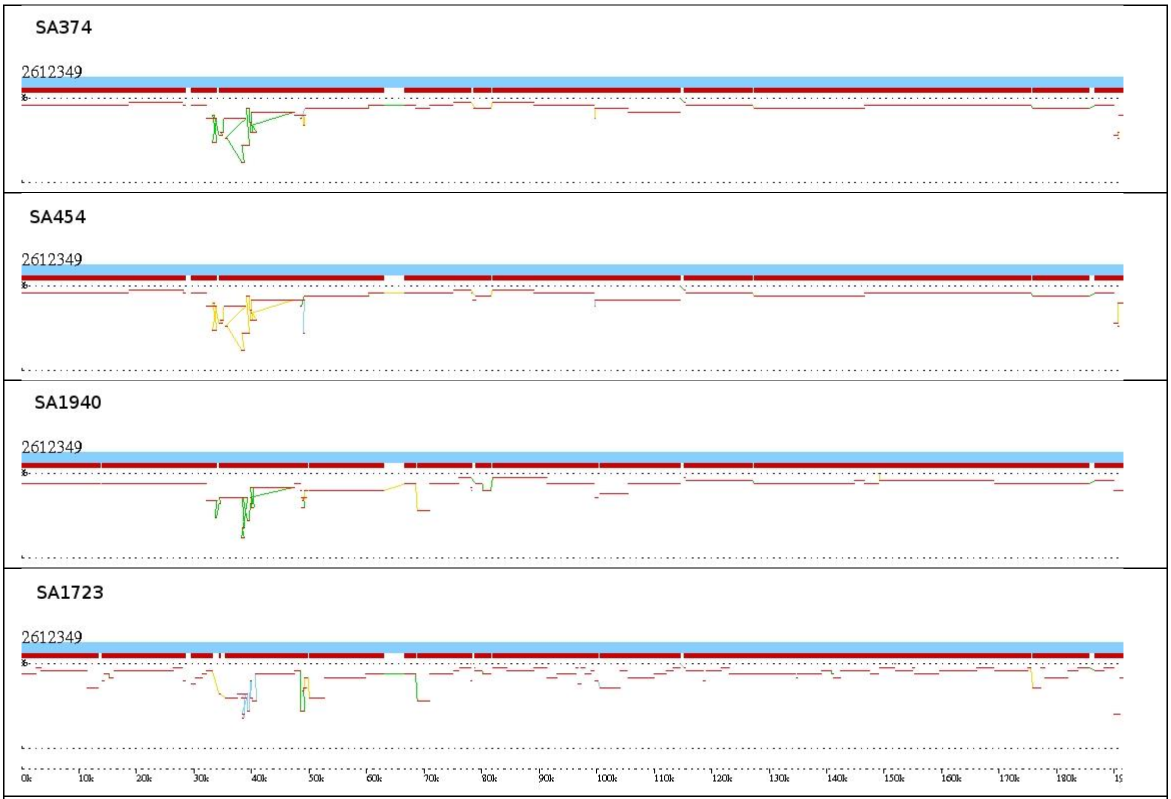


Figure 24. Assembly appraisal - core genome

Comparison of isolates with MRSA252 across 180kb of core genome starting at 2.612 megabases. SA374 and SA454 are closely-related and apparently have a common divergence from the reference. SA1940 and SA1723 have distinct divergence form the reference.

Bold elements represent the reference (blue) and the regions of reference matched by contigs (red). The lower, thinner elements represent the individual contigs, with distance below the reference representing sequence divergence. Elements of contigs matching discontinuously are connected by lines coloured to distinguish individual contigs.

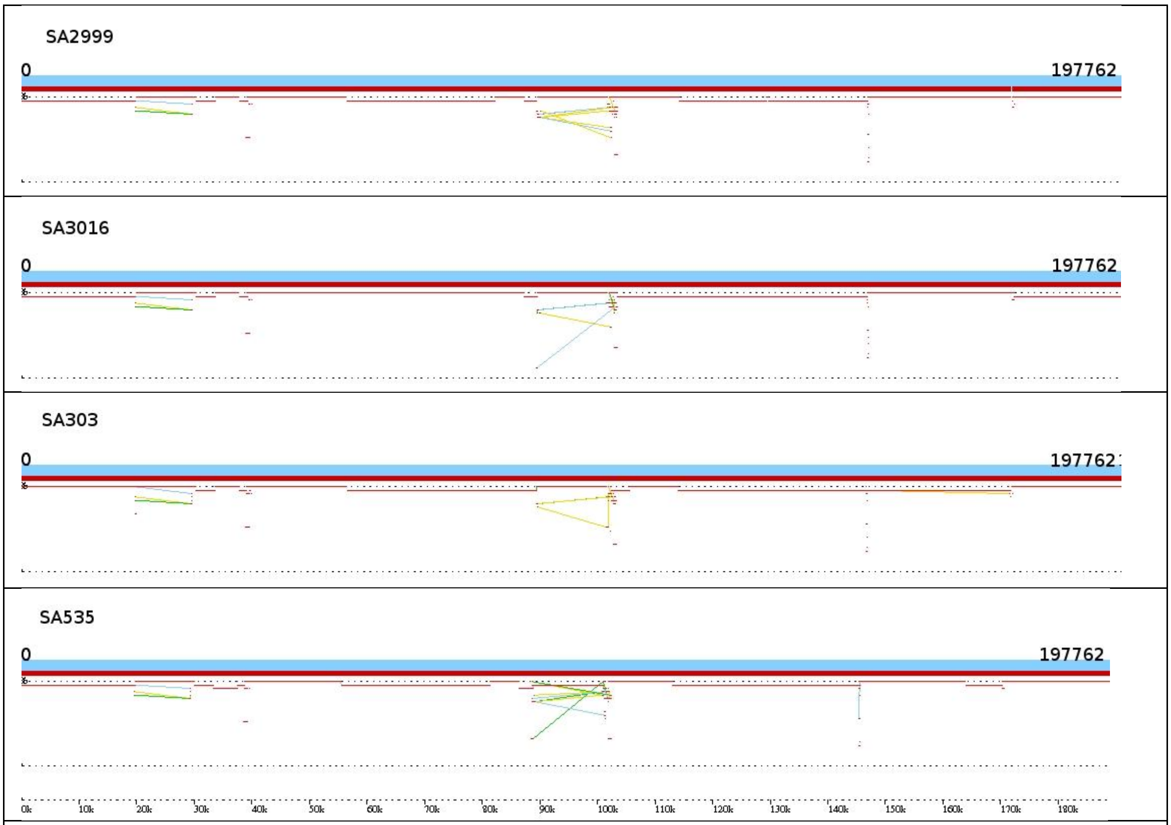


Figure 25. Assembly appraisal – MSSA476 core genome

Comparison of selected isolates with MSSA476. The 197kb comparison sequence starting at 2.530 megabases is homologous to that used in figures 25 & 26 but includes insertions cumulatively adding 17kb relative to MRSA252. The assemblies for SA2999, SA3016, SA303 and SA535 are all highly similar to MSSA476.

Bold elements represent the reference (blue) and the regions of reference matched by contigs (red). The lower, thinner elements represent the individual contigs, with distance below the reference representing sequence divergence. Elements of contigs matching discontinuously are connected by lines coloured to distinguish individual contigs.

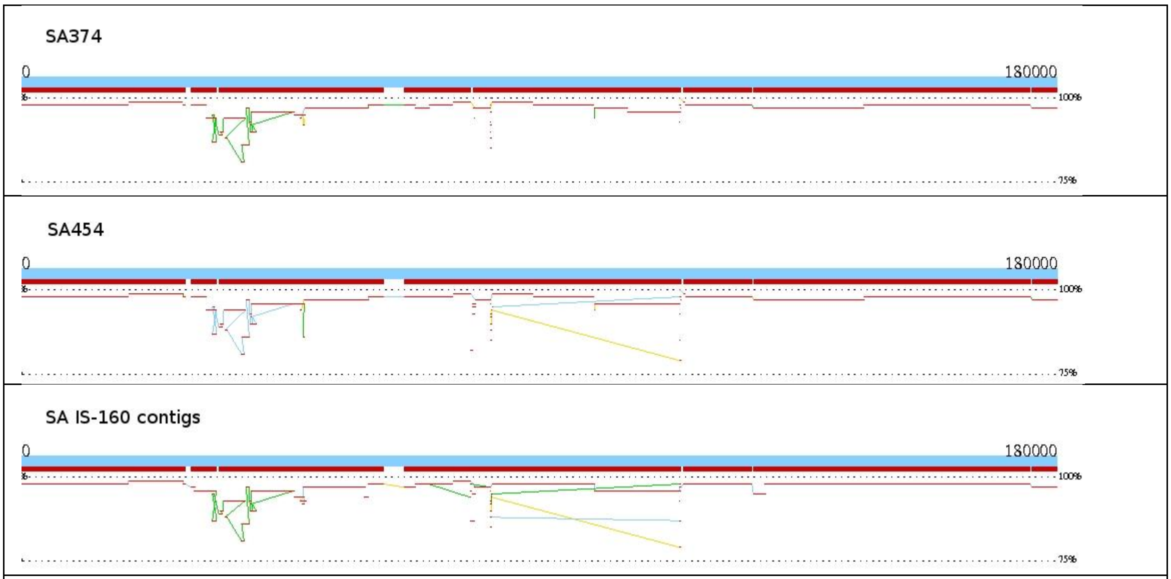


Figure 26. Assembly appraisal - IS-160 core genome

Comparison of isolates SA374, SA454 and the contigs of isolate IS-160 with MRSA252 across 180kb of core genome starting at 2.612 megabases. Isolates SA374 and SA454 and IS-160 are highly similar in their divergence pattern.

Bold elements represent the reference (blue) and the regions of reference matched by contigs (red). The lower, thinner elements represent the individual contigs, with distance below the reference representing sequence divergence. Elements of contigs matching discontinuously are connected by lines coloured to distinguish individual contigs.

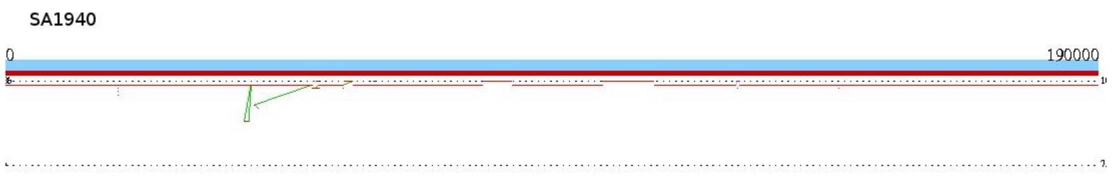


Figure 27. Assembly appraisal - Mu50 core genome

Comparison of isolate SA1940 with Mu50. The 195kb Mu50 sequence homologous to that used in figures 25 & 26 has been clipped at 190kb. SA1940 is highly similar to Mu50.

Bold elements represent the reference (blue) and the regions of reference matched by contigs (red). The lower, thinner elements represent the individual contigs, with distance below the reference representing sequence divergence. Elements of contigs matching discontinuously are connected by lines coloured to distinguish individual contigs.

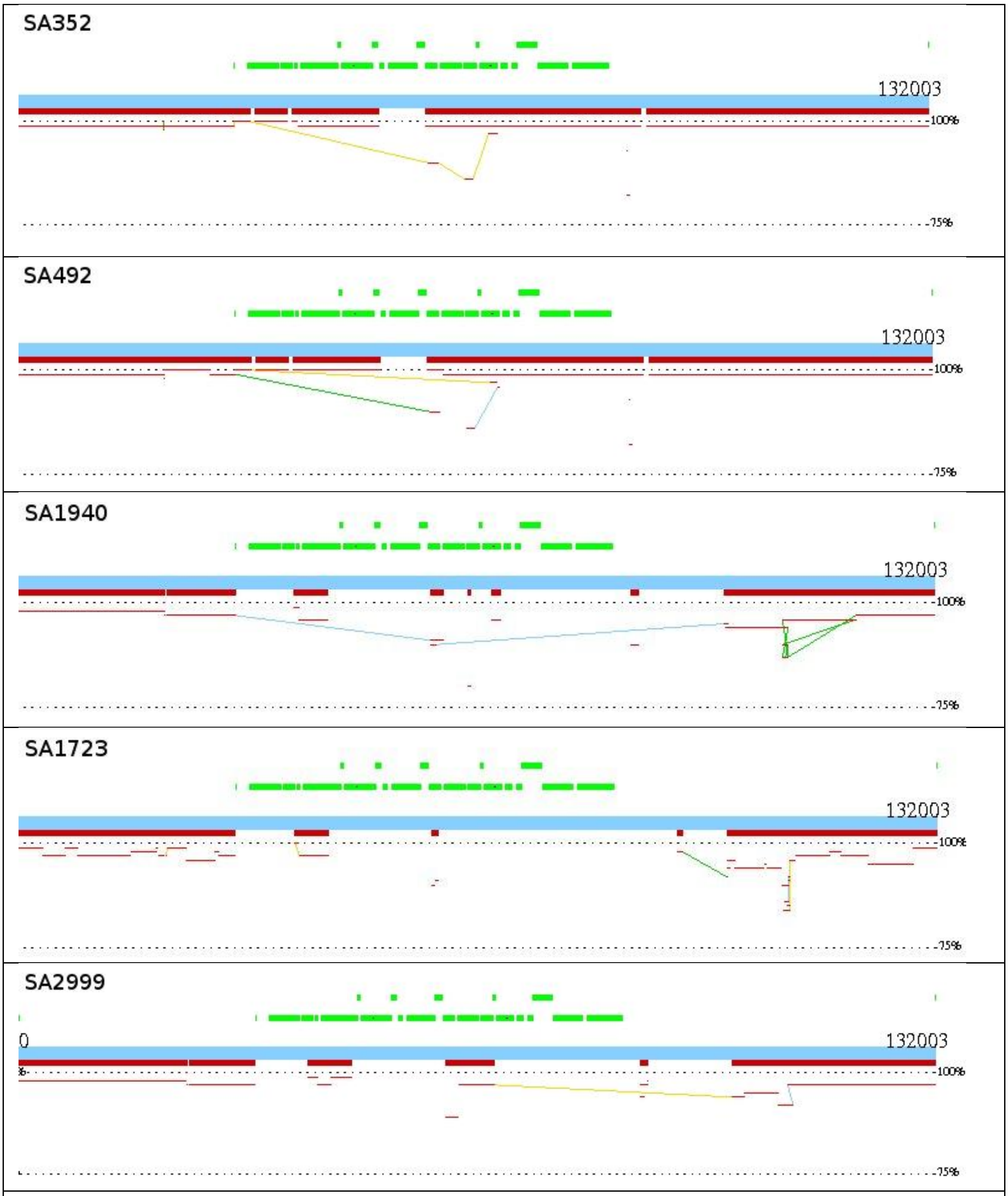


Figure 28. Preliminary SCC mapping with NUCmer

Preliminary alignments across the SCC region of MRSA252 made using NUCmer with default parameters. Bold elements represent the reference (blue), the genes of the SCC (green), and the reference regions matched by contigs (red). The lower, thinner elements represent the individual contigs, with distance below the reference representing sequence divergence. Elements of contigs matching discontinuously are connected by lines coloured to distinguish individual contigs.

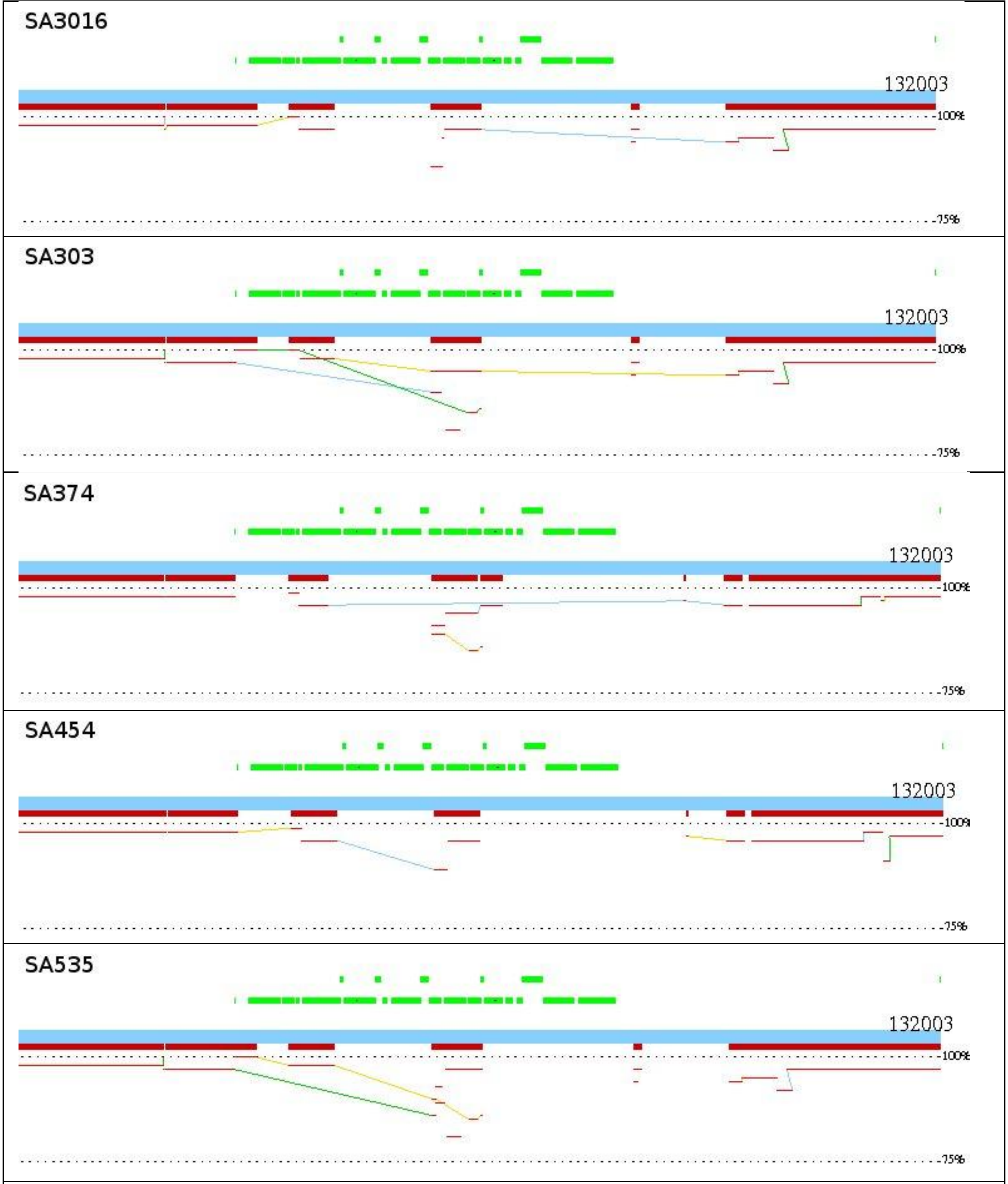


Figure 29. Preliminary SCC mapping with NUCmer

Preliminary alignments across the SCC region of MRSA252. Bold elements represent the reference (blue), the genes of the SCC (green), and the reference regions matched by contigs (red). The lower, thinner elements represent the individual contigs, with distance below the reference representing sequence divergence. Elements of contigs matching discontinuously are connected by lines coloured to distinguish individual contigs..

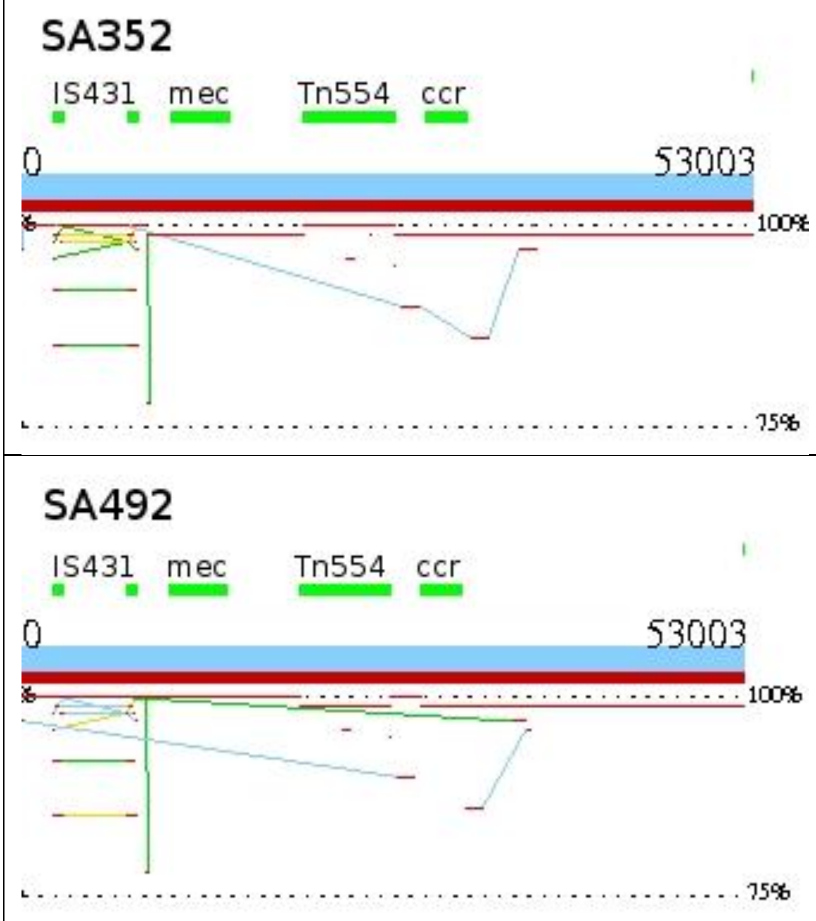


Figure 30. Mapping SCC against MRSA252

Comparison between the MRSA252 SCC and the contigs of isolates SA352 (top) and SA492 (bottom), made using nucmer with the maxmatch parameter.

In green are indicated selected elements of the SCC including the IS431 elements which flank the pUB110 plasmid insert, the mec operon (mecA, mecRI, mecI), Tn554, and the cassette recombinases (ccrB,ccrA).

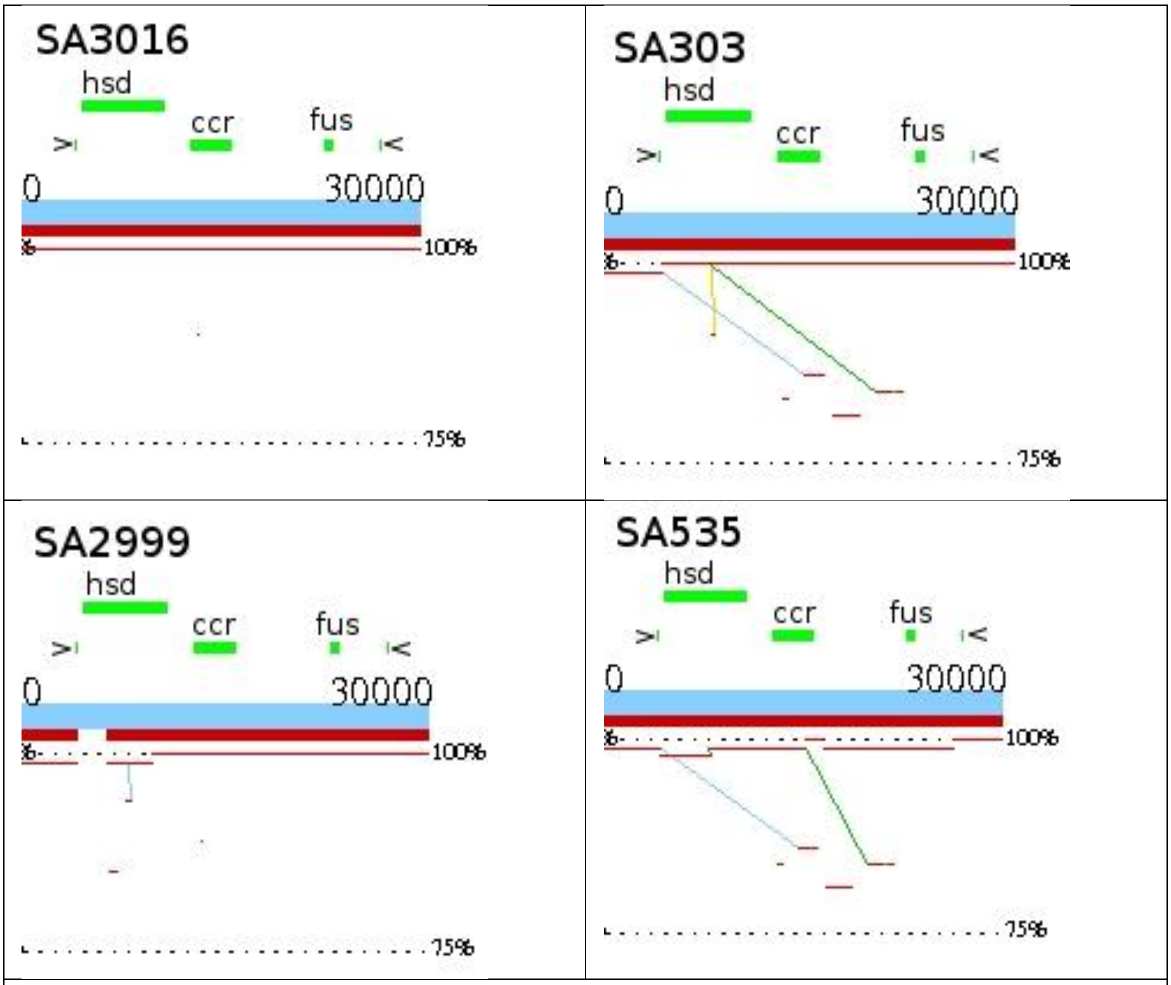


Figure 31. Mapping SCC against MSSA476

Comparison between the MSSA476 SCC and the contigs of isolates SA2999, SA3016, SA535 & SA303, made using nucmer with the maxmatch parameter... The bounds of the SCC are indicated by the arrowheads. In green are indicated selected elements of the SCC including the restriction/modification hsd operon (hsdR, hsdS, hsdM), the recombinase operon (ccrB, ccrA).& the fusidic acid resistance homolog (fus).

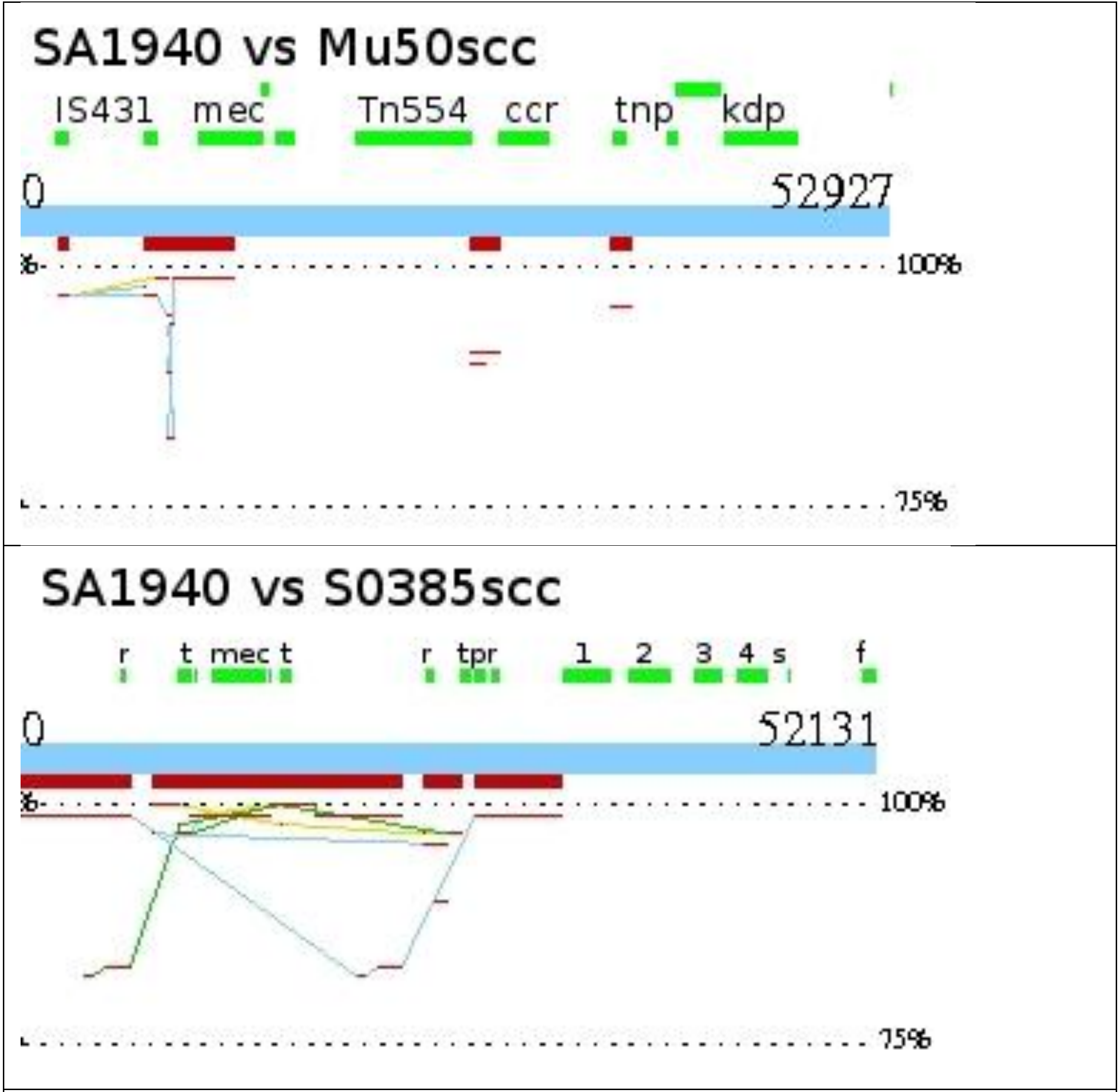


Figure 32. Mapping SSC against Mu50 and S0385

Comparison between the contigs of isolate SA1940 and the SCC of isolates Mu50 and S0385, made using nucmer with the maxmatch parameter.

In green are indicated selected elements of the SCC including (top) the IS 431 which flank the pUB110 insert, the mec operon (mecA, mecI, MecR), Tn554, the restriction/modification hsd operon (hsdR, hsdS, hsdM), the recombinase operon (ccrB, ccrA).& the operons tnp, kdp and (bottom) cassette recombinases (r), transposons (t), topoisomerase (p), the mec operon (mecA, mecI, MecR) and additional junk region genes for which SA1940 has no contig counterparts.

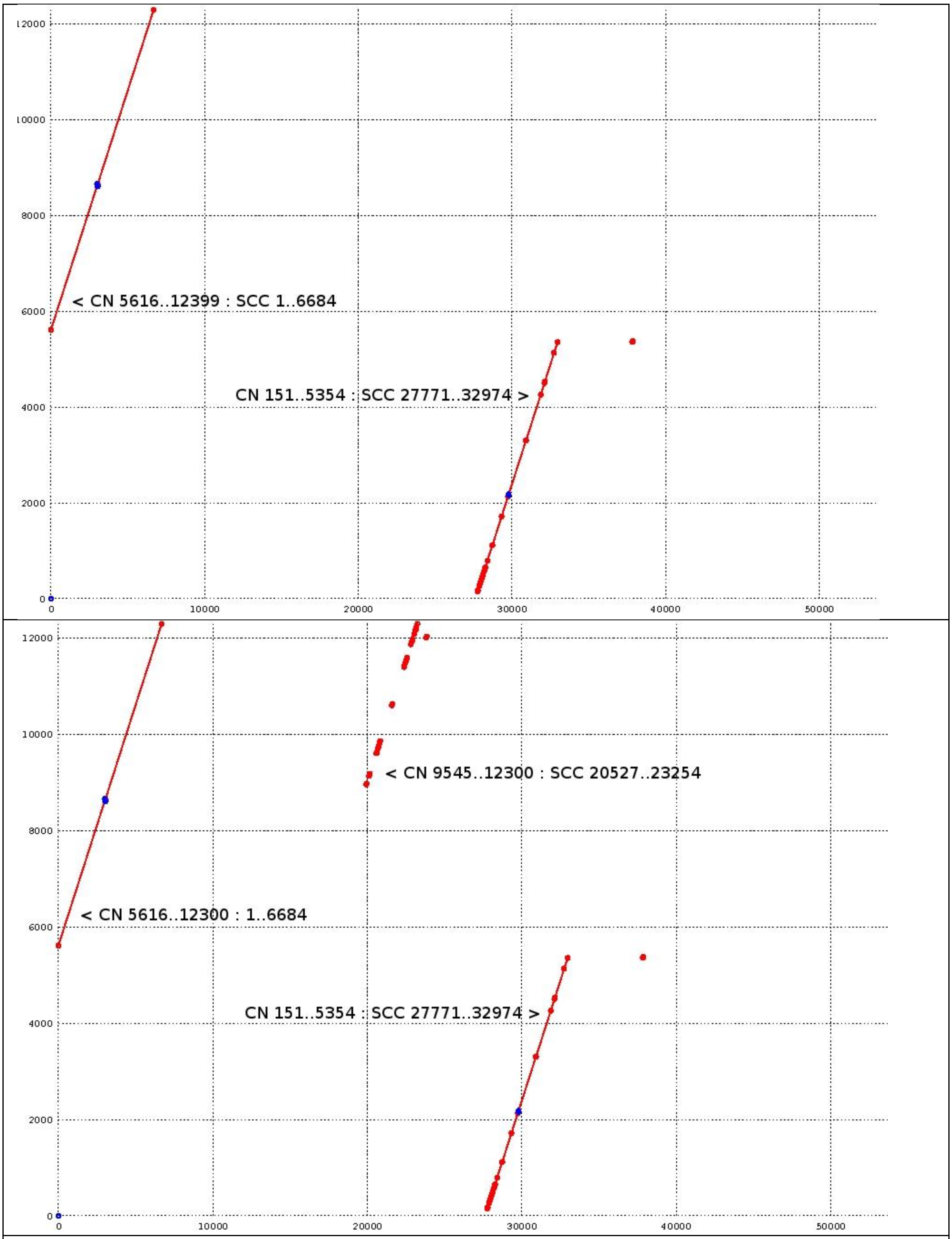


Figure 33. SA1940 CN115: potential mis-assembly

Comparison between CN115 (y-axis) and S0385 SCC (x-axis) made using mummer with parameters ‘unique matches’ (top) and maxmatch (bottom). The contig (CN) bases matching SCC bases are indicated.

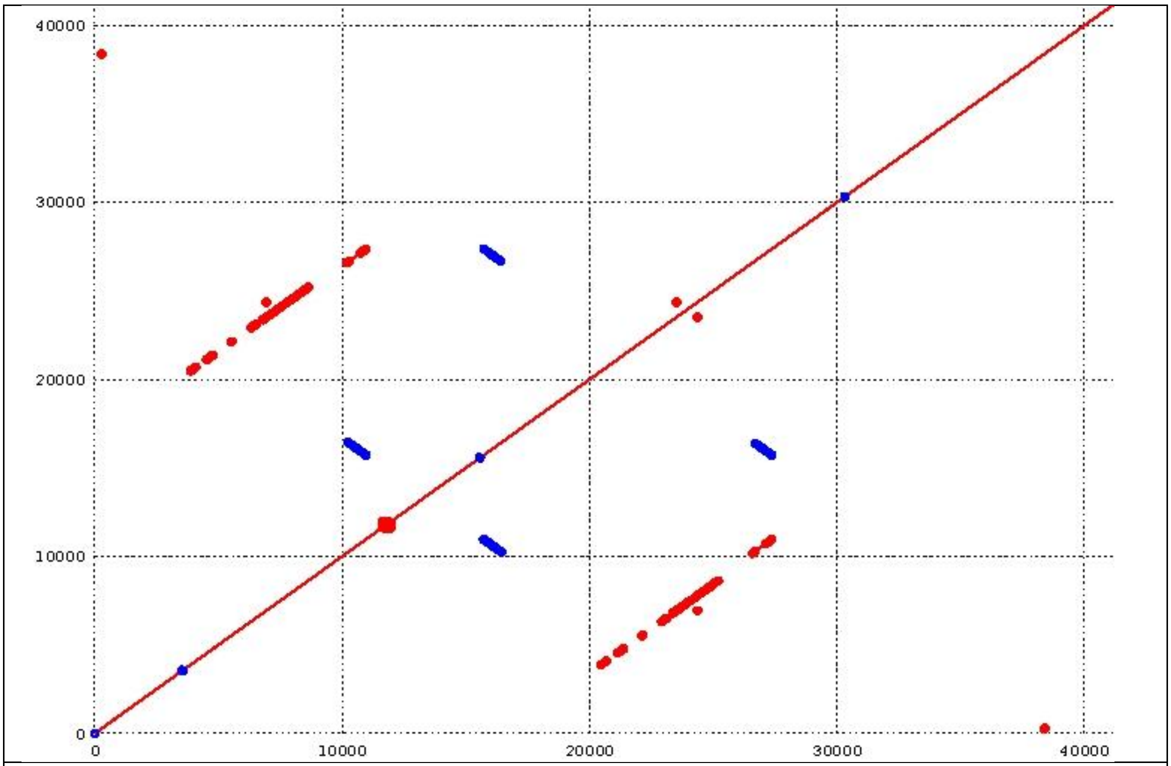


Figure 34. Repeat structure in S0385 SCC

Self-comparison of the S0385 SCC between ~500 nucleotides upstream of the SCC and 40k within the SCC. The structure of CN115 (above) suggested a common region of ~250 nucleotides which resulted in the sequences at the start of the SCC being fused to those upstream of SCC position 32974. However no such repeat is evident.

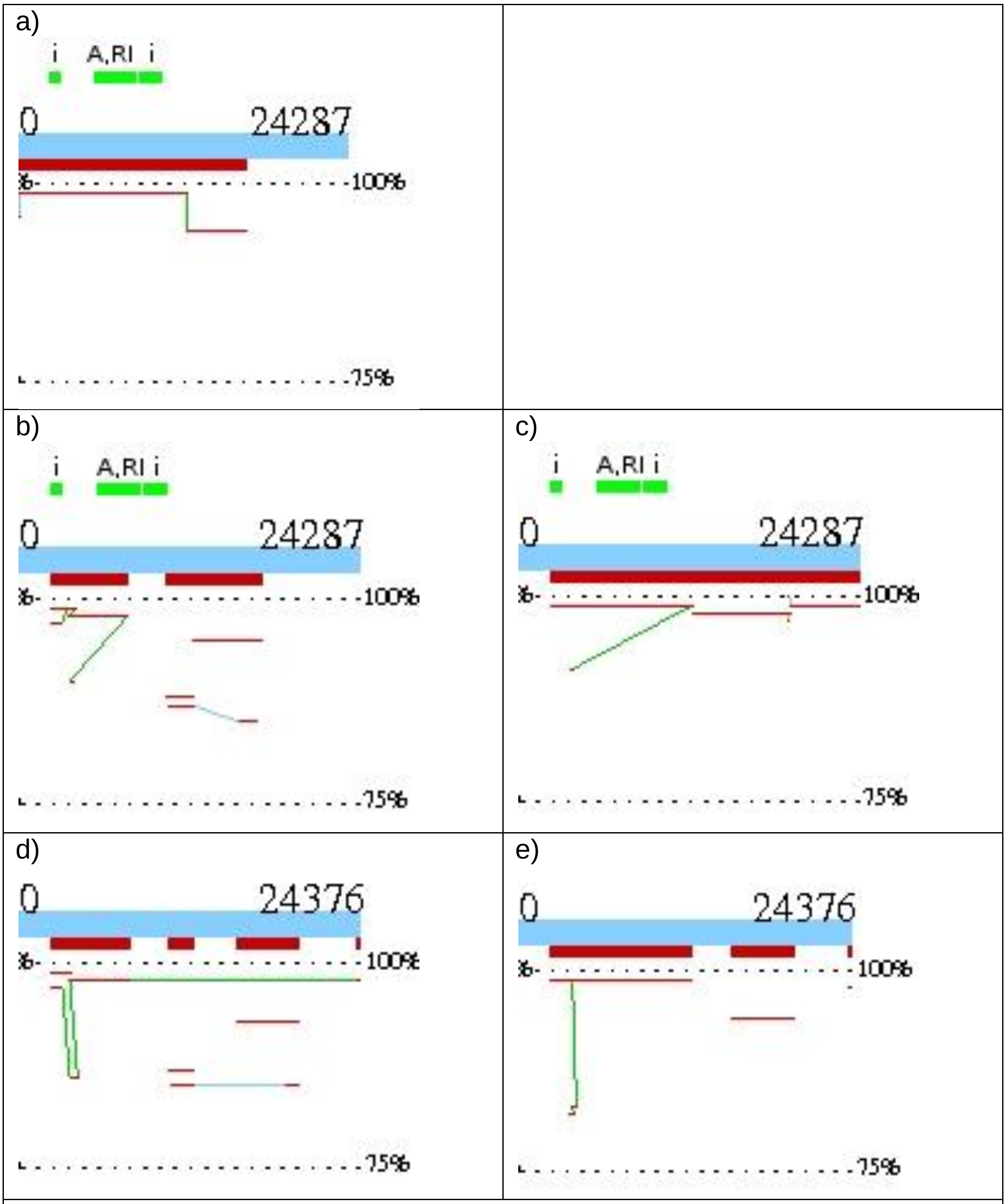


Figure 35. Mapping SCC against CA05 & IS160

Comparison between the SCC of strain CA05 and the contigs of unassembled strain IS-160 (a), isolate SA374 (middle left) and isolate SA454 (middle right) made using nucmer with the maxmatch parameter. In green are indicated selected elements of the SCC including (for all diagrams) the single IS431 element (i), the truncated mec operon (A, RI) and the single IS1272 element (i).

Figures d & e compare the assembled IS-160 contigs with those from isolate SA374 (left) and isolate SA454 (right).

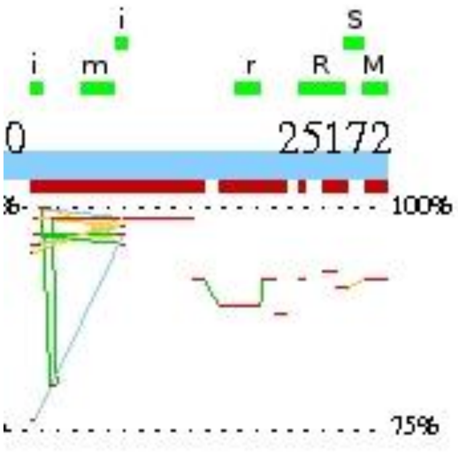
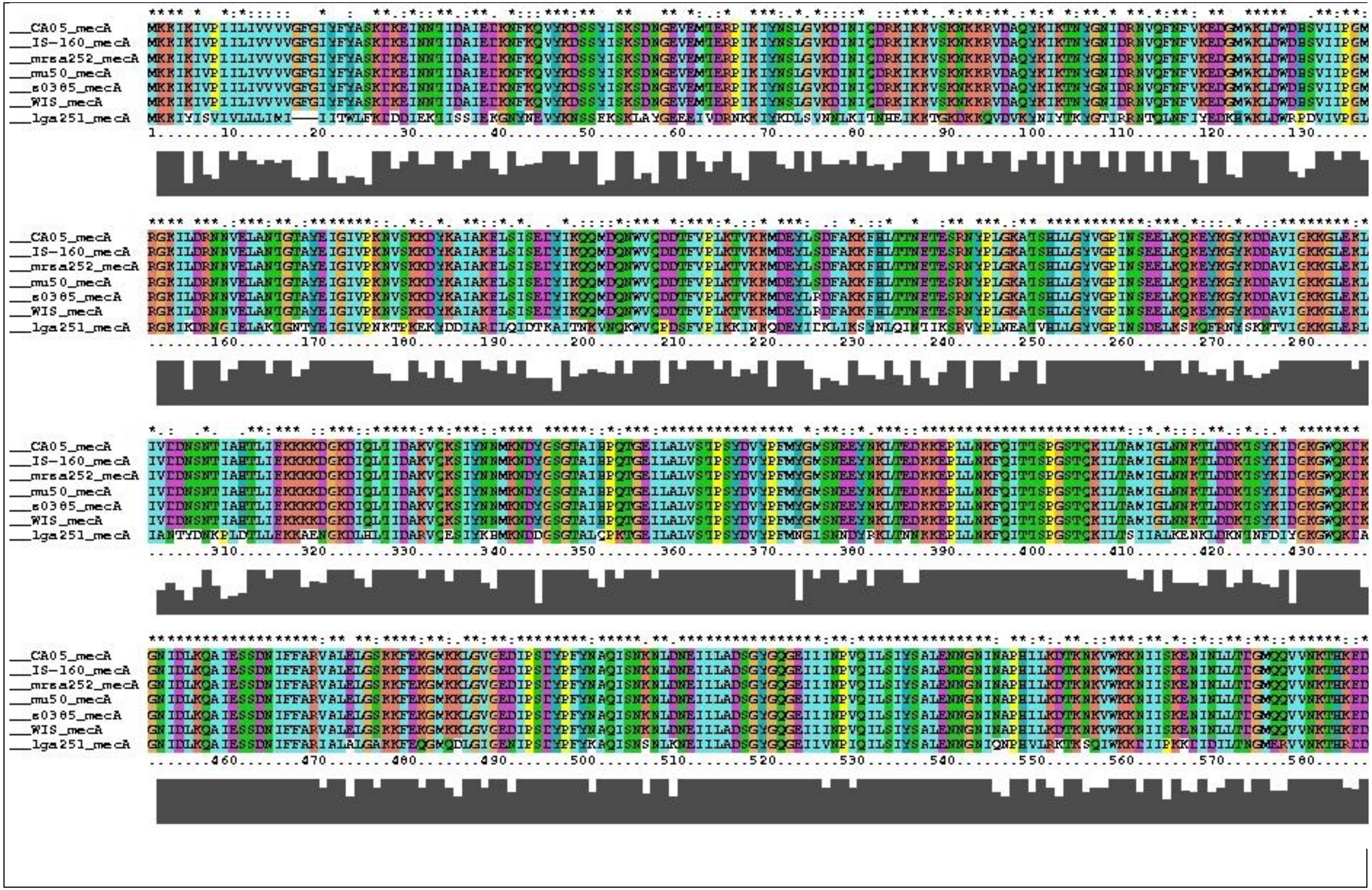
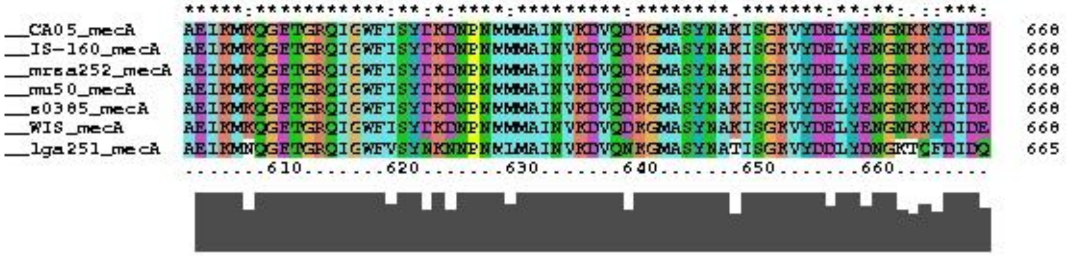


Figure 36. Mapping SCC against WIS_JCSC3624

Comparison between the SCC of strain WIS_JCSC3624 and the fragmentary contigs of isolate SA1723 made using nucmer with the maxmatch parameter.

In green are indicated selected elements of the SCC including (i) the two IS431 elements, (m) the truncated mec operon, (r) the novel ccrC recombinase which characterises JCSC3624, and the three genes of the hsd operon, R, S & M.

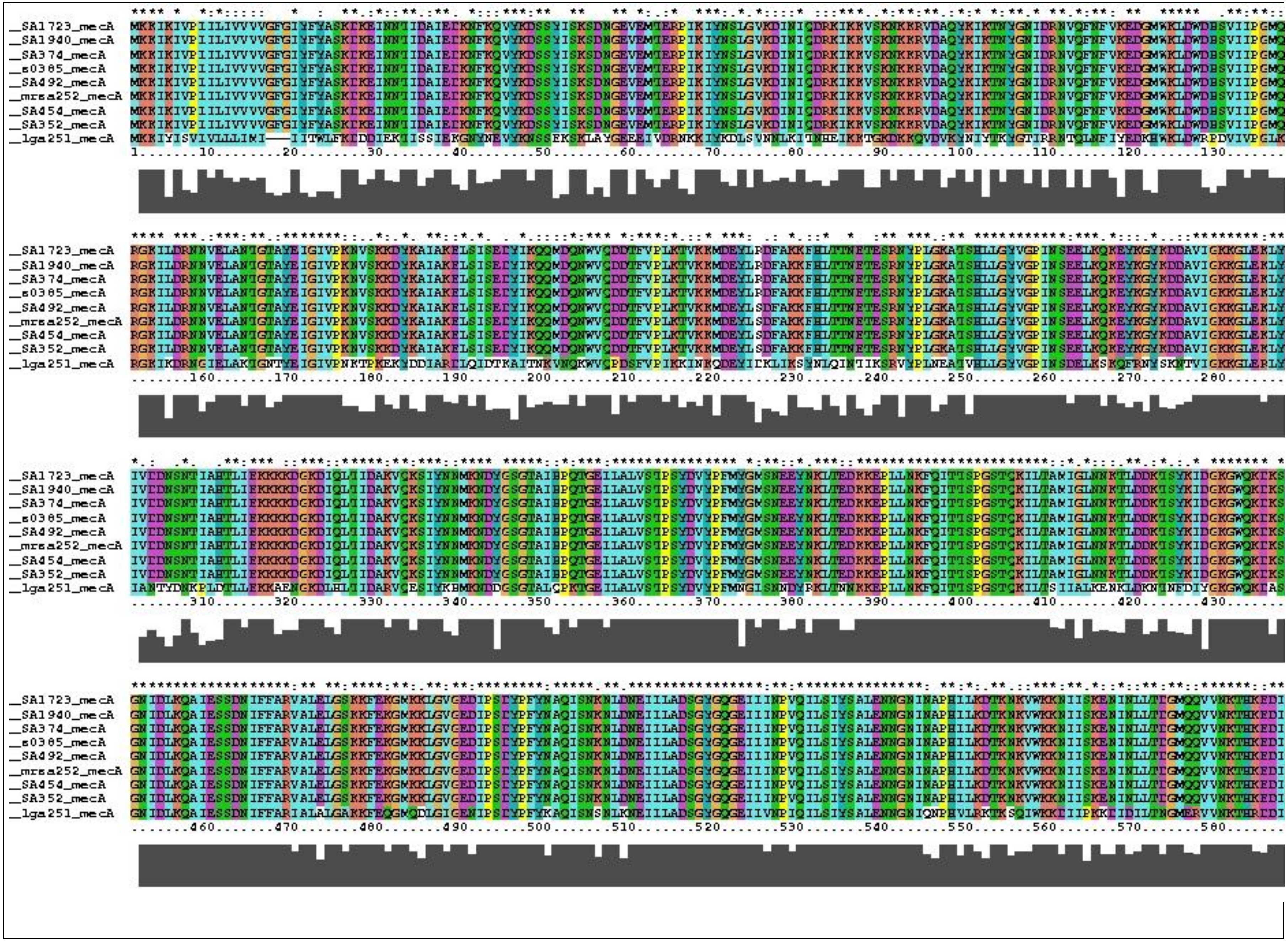

__CA05_mecA
__IS-160_mecA
__mrsa252_mecA
__mu50_mecA
__s0385_mecA
__WIS_mecA
__lga251_mecA



```
                  *****:***********:**:*:****:*********:********.*******:**:**:. :***:
___CA05_mecA      AEIKMKQGETGRQIGWFISYDKDNPNKMMAINVKDVQDKGMASYNAKISGKVYDELYENGNKKYDIDE   668
___IS-160_mecA    AEIKMKQGETGRQIGWFISYDKDNPNKMMAINVKDVQDKGMASYNAKISGKVYDELYENGNKKYDIDE   668
___mrsa252_mecA   AEIKMKQGETGRQIGWFISYDKDNPNKMMAINVKDVQDKGMASYNAKISGKVYDELYENGNKKYDIDE   668
___mu50_mecA      AEIKMKQGETGRQIGWFISYDKDNPNKMMAINVKDVQDKGMASYNAKISGKVYDELYENGNKKYDIDE   668
___s0385_mecA     AEIKMKQGETGRQIGWFISYDKDNPNKMMAINVKDVQDKGMASYNAKISGKVYDELYENGNKKYDIDE   668
___WIS_mecA       AEIKMKQGETGRQIGWFISYDKDNPNKMMAINVKDVQDKGMASYNAKISGKVYDELYENGNKKYDIDE   668
___lga251_mecA    AEIKMNQGETGRQIGWFVSYNKNNPNKIMAINVKDVQNKGMASYNATISGKVYDDLYDNGKTCFDIDQ   665
                  .......610.......620.......630.......640.......650.......660........
```

Figure 37. Reference mecA alignment

Clustal alignment of mecA peptides from reference strains (starts on previous page)

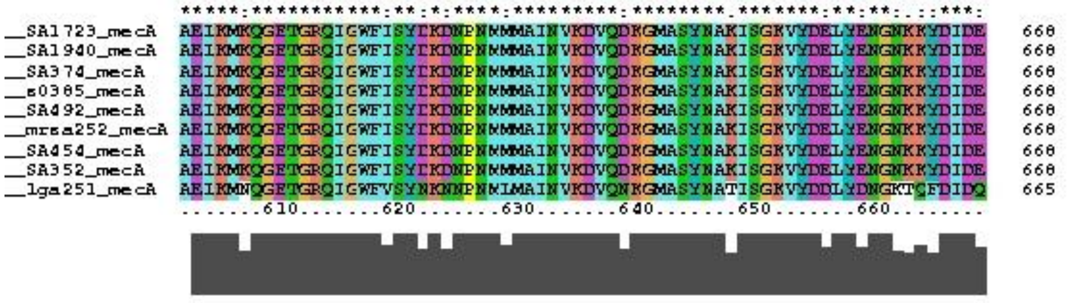


```
                **** *  *:*:::::   *  :  **.:*::**.:**. *:::***:**  **    ** *:.:*  ***:.*.*::::*  :::***..*:**:**.:*:* *:**.* **.*:**: **  *****  .**:**::
__SA1723_mecA   MKKIKIVPIILIVVVVGFGIYFYASKDKEINNTIDAIEDKNFKQVYKDSSYISKSDNGEVEMTERPIKIYNSLGVKDINIQDRKIKKVSKNKKRVDAQYKIKTNYGNIDRNVQFNFVKEDGMWKLDWDHSVIIPGMQ
__SA1940_mecA   MKKIKIVPIILIVVVVGFGIYFYASKDKEINNTIDAIEDKNFKQVYKDSSYISKSDNGEVEMTERPIKIYNSLGVKDINIQDRKIKKVSKNKKRVDAQYKIKTNYGNIDRNVQFNFVKEDGMWKLDWDHSVIIPGMQ
__SA374_mecA    MKKIKIVPIILIVVVVGFGIYFYASKDKEINNTIDAIEDKNFKQVYKDSSYISKSDNGEVEMTERPIKIYNSLGVKDINIQDRKIKKVSKNKKRVDAQYKIKTNYGNIDRNVQFNFVKEDGMWKLDWDHSVIIPGMQ
__s0385_mecA    MKKIKIVPIILIVVVVGFGIYFYASKDKEINNTIDAIEDKNFKQVYKDSSYISKSDNGEVEMTERPIKIYNSLGVKDINIQDRKIKKVSKNKKRVDAQYKIKTNYGNIDRNVQFNFVKEDGMWKLDWDHSVIIPGMQ
__SA492_mecA    MKKIKIVPIILIVVVVGFGIYFYASKDKEINNTIDAIEDKNFKQVYKDSSYISKSDNGEVEMTERPIKIYNSLGVKDINIQDRKIKKVSKNKKRVDAQYKIKTNYGNIDRNVQFNFVKEDGMWKLDWDHSVIIPGMQ
__mrsa252_mecA  MKKIKIVPIILIVVVVGFGIYFYASKDKEINNTIDAIEDKNFKQVYKDSSYISKSDNGEVEMTERPIKIYNSLGVKDINIQDRKIKKVSKNKKRVDAQYKIKTNYGNIDRNVQFNFVKEDGMWKLDWDHSVIIPGMQ
__SA454_mecA    MKKIKIVPIILIVVVVGFGIYFYASKDKEINNTIDAIEDKNFKQVYKDSSYISKSDNGEVEMTERPIKIYNSLGVKDINIQDRKIKKVSKNKKRVDAQYKIKTNYGNIDRNVQFNFVKEDGMWKLDWDHSVIIPGMQ
__SA352_mecA    MKKIKIVPIILIVVVVGFGIYFYASKDKEINNTIDAIEDKNFKQVYKDSSYISKSDNGEVEMTERPIKIYNSLGVKDINIQDRKIKKVSKNKKRVDAQYKIKTNYGNIDRNVQFNFVKEDGMWKLDWDHSVIIPGMQ
__lga251_mecA   MKKIYISVIVLLLIMI---IITWLFKDDDIEKTISSIEKGNYNEVYKNSSEKSKLAYGEEEIVDRNKKIYKDLSVNNLKITNHEIKKTGKDKKQVDVKYNIYTKYGTIRRNTQLNFIYEDKHWKLDWRPDVIVPGLK
                1.......10........20........30........40........50........60........70........80........90........100.......110.......120.......130.......

                **** ***.:***:**.:*******:...*:.*. **::*.*.  . *.::::*:*** *:***:*.::* ***: .: *.::*  *  :** ***.:** ************:***.*:::.*...::********:**
__SA1723_mecA   RGKILDRNNVELANTGTAYEIGIVPKNVSKKDYKAIAKELSISEDYIKQQMDQNWVQDDTFVPLKTVKKMDEYLPDFAKKFHLTTNETESRNYPLGKATSHLLGYVGPINSEELKQKEYKGYKDDAVIGKKGLEKLY
__SA1940_mecA   RGKILDRNNVELANTGTAYEIGIVPKNVSKKDYKAIAKELSISEDYIKQQMDQNWVQDDTFVPLKTVKKMDEYLPDFAKKFHLTTNETESRNYPLGKATSHLLGYVGPINSEELKQKEYKGYKDDAVIGKKGLEKLY
__SA374_mecA    RGKILDRNNVELANTGTAYEIGIVPKNVSKKDYKAIAKELSISEDYIKQQMDQNWVQDDTFVPLKTVKKMDEYLPDFAKKFHLTTNETESRNYPLGKATSHLLGYVGPINSEELKQKEYKGYKDDAVIGKKGLEKLY
__s0385_mecA    RGKILDRNNVELANTGTAYEIGIVPKNVSKKDYKAIAKELSISEDYIKQQMDQNWVQDDTFVPLKTVKKMDEYLPDFAKKFHLTTNETESRNYPLGKATSHLLGYVGPINSEELKQKEYKGYKDDAVIGKKGLEKLY
__SA492_mecA    RGKILDRNNVELANTGTAYEIGIVPKNVSKKDYKAIAKELSISEDYIKQQMDQNWVQDDTFVPLKTVKKMDEYLSDFAKKFHLTTNETESRNYPLGKATSHLLGYVGPINSEELKQKEYKGYKDDAVIGKKGLEKLY
__mrsa252_mecA  RGKILDRNNVELANTGTAYEIGIVPKNVSKKDYKAIAKELSISEDYIKQQMDQNWVQDDTFVPLKTVKKMDEYLSDFAKKFHLTTNETESRNYPLGKATSHLLGYVGPINSEELKQKEYKGYKDDAVIGKKGLEKLY
__SA454_mecA    RGKILDRNNVELANTGTAYEIGIVPKNVSKKDYKAIAKELSISEDYIKQQMDQNWVQDDTFVPLKTVKKMDEYLSDFAKKFHLTTNETESRNYPLGKATSHLLGYVGPINSEELKQKEYKGYKDDAVIGKKGLEKLY
__SA352_mecA    RGKILDRNNVELANTGTAYEIGIVPKNVSKKDYKAIAKELSISEDYIKQQMDQNWVQDDTFVPLKTVKKMDEYLSDFAKKFHLTTNETESRNYPLGKATSHLLGYVGPINSEELKQKEYKGYKDDAVIGKKGLEKLY
__lga251_mecA   RGKIKDRNGIELAKTGNTYEIGIVPNKTPKEKYDDIARDLQIDTKAITNKVNQKWVQPDSFVPIKKINKQDEYIDKLIKSYNLQINTIKSRVYPLNEATVHLLGYVGPINSDELKSKQFRNYSKNTVIGKKGLERLY
                .......160.......170.......180.......190.......200.......210.......220.......230.......240.......250.......260.......270.......280.......

                *.:   .*.   .**:***  ::***::*****:**:***::****  *****::*:***:**************** *:**::*.***::*************************::*.*::::.**.:*.:.* ******* *
__SA1723_mecA   IVDDNSNTIAHTLIEKKKKDGKDIQLTIDAKVQKSIYNNMKNDYGSGTAIHPQTGEILALVSTPSYDVYPFMYGMSNEEYNKLTEDKKEPLLNKFQITTSPGSTQKILTAMIGLNNKTLDDKTSYKIDGKGWQKDKS
__SA1940_mecA   IVDDNSNTIAHTLIEKKKKDGKDIQLTIDAKVQKSIYNNMKNDYGSGTAIHPQTGEILALVSTPSYDVYPFMYGMSNEEYNKLTEDKKEPLLNKFQITTSPGSTQKILTAMIGLNNKTLDDKTSYKIDGKGWQKDKS
__SA374_mecA    IVDDNSNTIAHTLIEKKKKDGKDIQLTIDAKVQKSIYNNMKNDYGSGTAIHPQTGEILALVSTPSYDVYPFMYGMSNEEYNKLTEDKKEPLLNKFQITTSPGSTQKILTAMIGLNNKTLDDKTSYKIDGKGWQKDKS
__s0385_mecA    IVDDNSNTIAHTLIEKKKKDGKDIQLTIDAKVQKSIYNNMKNDYGSGTAIHPQTGEILALVSTPSYDVYPFMYGMSNEEYNKLTEDKKEPLLNKFQITTSPGSTQKILTAMIGLNNKTLDDKTSYKIDGKGWQKDKS
__SA492_mecA    IVDDNSNTIAHTLIEKKKKDGKDIQLTIDAKVQKSIYNNMKNDYGSGTAIHPQTGEILALVSTPSYDVYPFMYGMSNEEYNKLTEDKKEPLLNKFQITTSPGSTQKILTAMIGLNNKTLDDKTSYKIDGKGWQKDKS
__mrsa252_mecA  IVDDNSNTIAHTLIEKKKKDGKDIQLTIDAKVQKSIYNNMKNDYGSGTAIHPQTGEILALVSTPSYDVYPFMYGMSNEEYNKLTEDKKEPLLNKFQITTSPGSTQKILTAMIGLNNKTLDDKTSYKIDGKGWQKDKS
__SA454_mecA    IVDDNSNTIAHTLIEKKKKDGKDIQLTIDAKVQKSIYNNMKNDYGSGTAIHPQTGEILALVSTPSYDVYPFMYGMSNEEYNKLTEDKKEPLLNKFQITTSPGSTQKILTAMIGLNNKTLDDKTSYKIDGKGWQKDKS
__SA352_mecA    IVDDNSNTIAHTLIEKKKKDGKDIQLTIDAKVQKSIYNNMKNDYGSGTAIHPQTGEILALVSTPSYDVYPFMYGMSNEEYNKLTEDKKEPLLNKFQITTSPGSTQKILTAMIGLNNKTLDDKTSYKIDGKGWQKDKS
__lga251_mecA   IANTYDNKPLDTLLEKKAENGKDLHLTIDARVQESIYKHMKNDDGSGTALQPKTGEILALVSTPSYDVYPFMNGLSNNDYRKLTNNKKEPLLNKFQITTSPGSTQKILTSIIALKENKLDDKTNFDIYGKGWQKDAS
                .......310.......320.......330.......340.......350.......360.......370.......380.......390.......400.......410.......420.......430.......

                ********************:** **:****:**:.**:**:*********:*****.**.*******************:**:****************: **:*:.**.::***:**.*::*::**:**::******::**
__SA1723_mecA   GNIDLKQAIESSDNIFFARVALELGSKKFEKGMKKLGVGEDIPSDYPFYNAQISNKNLDNEILLADSGYGQGEIIINPVQILSIYSALENNGNINAPHILKDTKNKVWKKNIISKENINLLTDGMQQVVNKTHKEDI
__SA1940_mecA   GNIDLKQAIESSDNIFFARVALELGSKKFEKGMKKLGVGEDIPSDYPFYNAQISNKNLDNEILLADSGYGQGEIIINPVQILSIYSALENNGNINAPHILKDTKNKVWKKNIISKENINLLTDGMQQVVNKTHKEDI
__SA374_mecA    GNIDLKQAIESSDNIFFARVALELGSKKFEKGMKKLGVGEDIPSDYPFYNAQISNKNLDNEILLADSGYGQGEIIINPVQILSIYSALENNGNINAPHILKDTKNKVWKKNIISKENINLLTDGMQQVVNKTHKEDI
__s0385_mecA    GNIDLKQAIESSDNIFFARVALELGSKKFEKGMKKLGVGEDIPSDYPFYNAQISNKNLDNEILLADSGYGQGEIIINPVQILSIYSALENNGNINAPHILKDTKNKVWKKNIISKENINLLTDGMQQVVNKTHKEDI
__SA492_mecA    GNIDLKQAIESSDNIFFARVALELGSKKFEKGMKKLGVGEDIPSDYPFYNAQISNKNLDNEILLADSGYGQGEIIINPVQILSIYSALENNGNINAPHILKDTKNKVWKKNIISKENINLLTDGMQQVVNKTHKEDI
__mrsa252_mecA  GNIDLKQAIESSDNIFFARVALELGSKKFEKGMKKLGVGEDIPSDYPFYNAQISNKNLDNEILLADSGYGQGEIIINPVQILSIYSALENNGNINAPHILKDTKNKVWKKNIISKENINLLTDGMQQVVNKTHKEDI
__SA454_mecA    GNIDLKQAIESSDNIFFARVALELGSKKFEKGMKKLGVGEDIPSDYPFYNAQISNKNLDNEILLADSGYGQGEIIINPVQILSIYSALENNGNINAPHILKDTKNKVWKKNIISKENINLLTDGMQQVVNKTHKEDI
__SA352_mecA    GNIDLKQAIESSDNIFFARVALELGSKKFEKGMKKLGVGEDIPSDYPFYNAQISNKNLDNEILLADSGYGQGEIIINPVQILSIYSALENNGNINAPHILKDTKNKVWKKNIISKENINLLTDGMQQVVNKTHKEDI
__lga251_mecA   GNIDLKQAIESSDNIFFARIALALGAKKFEQGMQDLGIGENIPSDYPFYKAQISNSNLKNEILLADSGYGQGEIIVNPIQILSIYSALENNGNIQNPHVLRKTKSQIWKKDIIPKKDIDILTNGMERVVNKTHREDI
                .......460.......470.......480.......490.......500.......510.......520.......530.......540.......550.......560.......570.......580.......
```



```
               *****:***********:**:*:****:*********:********.*******:**:**:. :***:
__SA1723_mecA  AEIKMKQGETGRQIGWFISYDKDNPNKMMAINVKDVQDKGMASYNAKISGKVYDELYENGNKKYDIDE   668
__SA1940_mecA  AEIKMKQGETGRQIGWFISYDKDNPNKMMAINVKDVQDKGMASYNAKISGKVYDELYENGNKKYDIDE   668
__SA374_mecA   AEIKMKQGETGRQIGWFISYDKDNPNKMMAINVKDVQDKGMASYNAKISGKVYDELYENGNKKYDIDE   668
__s0385_mecA   AEIKMKQGETGRQIGWFISYDKDNPNKMMAINVKDVQDKGMASYNAKISGKVYDELYENGNKKYDIDE   668
__SA492_mecA   AEIKMKQGETGRQIGWFISYDKDNPNKMMAINVKDVQDKGMASYNAKISGKVYDELYENGNKKYDIDE   668
__mrsa252_mecA AEIKMKQGETGRQIGWFISYDKDNPNKMMAINVKDVQDKGMASYNAKISGKVYDELYENGNKKYDIDE   668
__SA454_mecA   AEIKMKQGETGRQIGWFISYDKDNPNKMMAINVKDVQDKGMASYNAKISGKVYDELYENGNKKYDIDE   668
__SA352_mecA   AEIKMKQGETGRQIGWFISYDKDNPNKMMAINVKDVQDKGMASYNAKISGKVYDELYENGNKKYDIDE   668
__lga251_mecA  AEIKMNQGETGRQIGWFVSYNKNNPNKIMAINVKDVQNKGMASYNATISGKVYDDLYDNGKTCFDIDQ   665
               .......610.......620.......630.......640.......650.......660........
```

Figure 38. Isolate MecA alignment

Clustal alignment of the selected reference strain mecA peptides with those from the isolates analysed here (figure starts on previous page)

4 DISCUSSION

Illumina sequence libraries for 10 isolates of Staphylococcus aureus were made available by Taane Clark of the London School of Hygiene and Tropical Medicine. The libraries consisted of paired-end reads, so each consisted of two fastq files, one for each read. The source of the libraries and the reason for the samples being of interest were requested but not made available.

All steps of NGS data analysis have been shown to be reliant on the provision of high quality sequence reads (Kircher et al., 2011). For this reason data analysis was preceeded by data QC and pre-processing.

4.1 Data quality control

Sequence libraries made available were quality-appraised using the FastQC package from The Babraham Institute (Andrews, 2011a). The libraries were found to have significant quality issues, including indications of contaminant sequences, clear evidence of persistent adapter sequences, and a high degree of sequence reiteration.

When summarised in a scheme assigning and summing quality scores, fewer than half the libraries scored more than half the total possible (table2). Basic library statistics showed that the number of reads in each library ranged between 0.9M to 10M, that the read length in all libraries was consistent at 105nucleotides (table3). Some libraries had GC content significantly different from that expected of *S aureus* (table3).

Plots of read Qscores over all sequences for each library indicated inconsistent library quality, with some libraries having subsidiary peaks of lower-quality reads (figure 3). Plots of sequence quality against read length (figure 4) likewise indicated inconsistent library quality, with four libraries flagged by FastQC as ‘extremely unusual’.

Plots of observed GC distribution over all reads relative to a Normal distribution with the same mean and standard deviation as the library (figure 5), an indicator of library sequence contamination, suggested that all but one library had sequence contamination; FastQC flagged 4 libraries as ‘slightly abnormal’, and 5 as ‘extremely unusual’. A single library was flagged as ‘entirely normal’.

Plots of nucleotide frequencies against read position for quality random libraries are essentially flat. Plots of either GC frequency (figure 6) or of all four base frequencies (figure 9) strongly suggested the libraries were not random. On the basis of the GC frequency plots, FastQC flagged 5 libraries as ‘slightly abnormal’, and 5 as ‘extremely unusual’. On the basis of the four base frequency plots, all libraries were flagged by FastQC as ‘extremely unusual’

Plots of sequence duplication levels (figures 7, 8 and table 4) indicated that all libraries had a high degree of sequence duplication, with FastQC estimating sequence duplication levels between 30 and 82 % (median 65). FastQC flagged all libraries as ‘extremely unusual’ on this metric.

Two measures of adapter sequence contamination highlighted adapter presence in the libraries. Tabulation of over-represented sequences exceeding a threshold of 0.1% of total reads identified Illumina TruSeq adapter sequences as comprising between 0.26 and >4% of total reads. A second measure, a depiction against read length of the relative enrichment of length 5 kmers which appear more frequently than expected within a sequence of the read’s base composition, confirmed the presence of adapters in the libraries (figure 8). Approximately 50% of all over-represented kmers could be shown to derive from the Illumina TruSeq adapters. Working solely from the over-represented kmers, it proved possible to derive from all libraries but 2 the first 14 nucleotides of the TruSeq adapter (table 4).

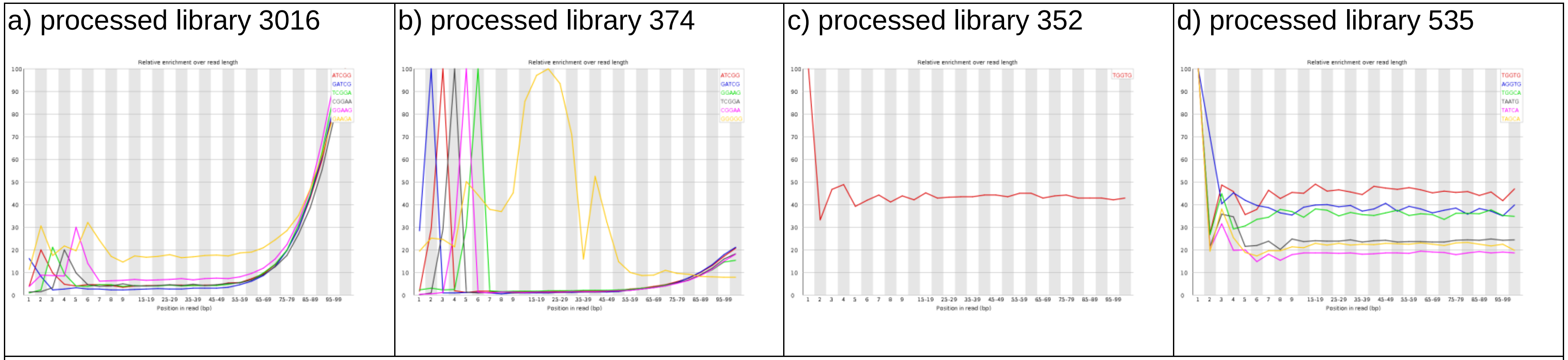
Analysis of the rate at which libraries accumulated adapter reads relative to library read depth indicated that most accumulated adapter reads at a constant rate of 0.21 per library read (figure 9). Two libraries had fewer adapter reads, SA352 and SA303 (6% and 13%), and two had more, SA3016 and SA374 (~ 31% and ~49%).

4.2 Data pre-processing

FastQC analysis indicated adapter contamination of the libraries. To guide adapter clipping approaches, the position of adapters within library reads was determined using a simple shell script calling grep (AppendixB). With the exception of the universal adapter in libraries SA2999 and SA303, most adapters were at the start of library reads (table 6): Universal adapter counts at read starts comprised 65 - 95% (median 71%) of total universal adapter counts. Indexed adapter counts at read starts comprised 87 - 99% (median 96%) of total indexed adapter counts. A variable excess of indexed adapter sequences over universal adapter sequences was identified. For instances at the start of reads, this excess ranged 2.6 - 30.3 fold (median 5.7 fold); the excess for position-independent adapter counts ranged 2.5 - 13.5 fold (median 4 fold). These figures suggest inconsistently optimised paired-end sequencing reactions.

Initial data sanitation approaches used programs from the FastX package of short reads pre-processing tools (Gordon, 2009). It was found that the tool designed for adapter clipping, fastx_clipper, processed library reads at positions entirely unrelated to the adapter sequence provided on the command line (not shown). Direct contact with the package author indicated that the package, originally written at a time when reads were just 36nucleotides, was poorly suited to the 105 nucleotide read length of the libraries. No documentation of this limitation could be found.

Data sanitization was achieved using a custom shell script, ultimately as described in appendixA. Suitable parameters were identified through reiterative cycles of adapter removal and reanalysis using FastQC. The outcome of some intermediate pre-processing parameters high-lighted the variable quality of the libraries. Figure 39. Intermediate data sanitization).


a) processed library 3016

b) processed library 374

c) processed library 352

d) processed library 535

Figure 39. Intermediate data sanitization

Inconsistent library outcomes after intermediate pre-processing. Some libraries (a:SA3016, also SA1940, SA1723, SA492); maintained high levels of adapter-derived kmers at the end of reads but much reduced kmers elsewhere; library SA374 (b) maintained high levels of adapter-derived kmers at the start of reads, some at the end, and an elevated GGGGG contaminant (yellow); library SA352 (c) maintained a single kmer TGGTG not derived from the adapters; others (d: SA535, and SA303) maintained multiple kmers not derived from adapters.

Processing details: using a minimum overlap of 4, a 10% permitted mismatch rate, and allowing wildcards to count as adapter matches, universal and sample-specific indexed adapters were trimmed twice each from the reads.

Re-analysis using FastQC after pre-processing showed that all but one data-quality metric was improved by the processing applied Table 8. Summary of Basic Statistics (raw & pre-processed). The ‘sequence length distribution’ metric universally returned warnings as a result of the amended library read length after adapter trimming. Overall, pre-processing resulted in the removal of between 1.8 and 44% of library reads, or between 73 thousand and 2.5 million reads. The contaminating adapter sequences reported by the ‘over-represented sequences’ metric were entirely removed from all libraries. For 7 libraries, pre-processing entirely removed the corresponding over-represented kmers, but for three libraries, residual over-represented kmers which did not derive from the Illumina adapter persisted. The origin of the persistent contaminants from which these kmers derived was not identified. For all libraries, the impact of pre-processing on sequence duplication was minimal (Table 9). Use of the FastX ‘collapser’ function independently confirmed a persistent high level of sequence duplication within all libraries (Table 10. FastX_collapser: sequence redundancy). This duplication in part reflects the high sequencing depth of the libraries; relative to the length of the MRSA252 genome standard (2902216 nuc), most libraries have more reads than reference bases, so some duplication is inevitable.

4.3 De novo assembly and assembly appraisal

Sanitized library reads were assembled de novo using the de Bruijn graph assembler Velvet, with optimisation coordinated by the wrapper script VelvetOptimiser (Zerbino and Birney, 2008, Gladman, 2011). De novo assembly rather than reference-based mapping was used in part because of the well characterised (see chapter 1) variability of elements of the accessory genome including the Staphylococcal Cassette Chromosome. Scripts for library subsetting to limit the number of reads were prepared and used in independent optimisations of Velvet, but little difference in the outcome was observed (not shown). Communication with the package author indicated that the higher input coverage from non-subsetted libraries would not degrade Velvet performance (D Zerbino, pers comm.).

Contigs from Velvet were validated first by appraising the ability of each library to enable Velvet adequately to distinguish between the low-coverage shorter contigs which result from miss-assembly and the longer higher-coverage contigs which accurate assembly produces. This metric is largely determined by the sequence coverage or read depth each library provides. Most libraries were found to have coverage adequate to allow Velvet clearly to resolve the two classes of contig (Figure 20, Figure 21). The coverage from one library, SA1723, was lower and less suited to resolution of the two classes (Figure 22).

Contigs from Velvet were appraised by alignment to core genome regions of MRSA252, which are less subject to the compositional fluidity typical of the regions hosting elements of the accessory genome. The rationale for this comparison to relatively static core genome sequences was that accurate assembly by Velvet should result in contigs which match these regions; whereas consistent lack of such matches would drawn into question the accuracy of the assembly overall. Just two isolates proved closely to match the MRSA252 reference sequence, SA352 and SA492 (Figure 23) over a 180kb core genome region starting at 2.6megabases. The remaining isolates diverged from the reference sequence, initially confounding conclusions about the quality of the velvet assemblies. Nonetheless, some isolate groupings could be identified on the basis of distinctive patterns of divergence from the reference; four isolates, SA2999, SA3016, SA303 and SA535 could be seen to be closely-related (Figure 23), as could the two isolates SA474 and SA454 (Figure 24). The remaining two isolates SA1940 and SA1723 were differently divergent from the reference (Figure 24). The lower coverage of the SA1723 library resulted in a highly fragmented alignment, diminishing confidence in this assembly.

From the contigs of each isolate not closely matching the MRSA252 reference sequence a highly-divergent contig was selected as a query subject for blastn searches for more closely-related database entries. The rationale for selecting a highly divergent contig was that such contigs would be more selective ‘handles’ for identifying related database entries than more conserved contigs. The potential flipside - that these contigs potentially represent divergent assembly by

Velvet rather than genuinely divergent sequences - could be addressed by determining whether their matches could be extended into adjoining sequences. With the exception of SA1723, clear database counterparts for each of the isolates not closely matching the MRSA252 reference could be identified, indicating that the divergent contigs represented accurate assembly of genuinely divergent sequences rather than assembly artefacts.

The previously-identified group of four isolates, SA2999, SA3016, SA303 and SA535 were highly related to MSSA476, a community–acquired meticillin-sensitive isolate (Figure 25; (Holden et al., 2004). The core genome region of MSSA476 used for the comparison is homologous to that of MRSA252 used in comparisons with SA352 and SA492 but includes insertions cumulatively adding 17kb to its length (*Ibid*). Relative to MSSA476, the four related isolates had two insertion points matched by both continuous contigs, which include the entirety of the insertion, and discontinuous contigs, which include the sequences flanking the insertion points but lack the insertions themselves (Figure 25). The positions of these insertions, at 20-30kb and 90-100kb, do not correspond to those between MRSA252 and MSSA476 and may indicate either divergence of this core genome region or a local Velvet miss-assembly.

The previously-identified group of two isolates, SA374 and SA454, was highly related to an unassembled WGS project for *S. aureus* subsp aureus IS-160 from the J Craig Venter Institute (Jones, 2012). In the absence of a complete assembly for this genome, nucmer plots of the contigs from SA372, SA454 and IS-160 against MRSA252, which displayed highly similar patterns of divergence, were used to confirm the similarity between these three isolates exhibited (Figure 26).

Isolate SA1940 was found to be highly related to Mu50, a vancomycin-resistant MRSA strain (Figure 27). No comparable reference was identified for the highly fragmented contig assembly for SA1723.

4.4 Assembly analysis

4.4.1 Preliminary alignment to MRSA252

Notwithstanding the sequence homologies noted while appraising the isolate assemblies, the contigs for all isolates were aligned with MRSA252 using the default nucmer setting to gain an early view of the SCC status of each isolate relative to this representative of UK endemic MRSA. Consistent with the above homologies, isolates SA352 and SA492, which were highly homologous to MRSA252 across the core genome, were also the most closely-related to MRSA252 across the genome region including the SCC, most of which both isolates maintained (Figure 28, top two figures). Other isolates lacked sequences aligning to much of the MRSA252 SCC, and also to adjoining downstream sequences (Figure 28, Figure 29).

4.4.2 SCC-specific alignments

Alignments of isolate contigs against the SCC from specific reference strains were used to identify the SCC status of each isolate; appropriate reference strains were identified either during assembly appraisal or from searching the *S.aureus* subset of NCBI.

The contigs of four isolates - SA3016, SA2999, SA303 and SA535 - were found to align to the SCC of MSSA476, a community-acquired meticillin-sensitive strain (Holden et al., 2004). The genome of MSSA476 includes a limited SCC which lacks the mecA operon but which includes the cassette recombinase genes ccrB & ccrA, a restriction-modification operon (*hsdR*, *hsdS*, *hsdM*), and a gene specifying a fusidic acid resistance homolog (*fus*). Isolate SA3016 included contigs precisely and continuously matching the full length of the MSSA SCC, but those for the other isolates diverged somewhat; those of SA2999 included no sequence corresponding to the first gene of the hsd operon, *hsdM*, and also diverged over the second, *hsdS*; over the cassette recombinase and the *fus* gene however it matched precisely the MSSA sequence. Those of isolate SA303 similarly matched the MSSA SCC, but showed clear evidence that sequences immediately upstream were divergent, perhaps indicating a more recent SCC acquisition by the host strain. Isolate SA535 homology was more disjointed, but included contigs corresponding to all elements of the MSAA SCC including the mec operon. The contigs of isolates SA303 and SA535 both included ‘hybrid’ assemblies, contigs which align to two otherwise-unconnected genomic locations, with one perfect alignment and one, generally shorter, divergent alignment. Although time constraints prevented further exploration of these contigs, they were in general more frequent in assemblies with higher coverage, and may consequently reflect a degree of excess coverage-induced mis-assembly by Velvet.

The contigs assembled for isolates SA352 and SA492 demonstrated a certain consistency across the Staphylococcal genome. Across a region of core genome used to appraise assemblies (Figure 23) both isolates proved identical to the UK epidemic MRSA reference strain MRSA252. Across the SCC of this strain the contigs of both isolates included sequence counterparts to all genes including the mecA operon (Figure 30). In addition, nucmer analysis depicted multiple variously divergent contigs aligning with the IS431 elements flanking the pUB110 plasmid insert of the MRSA252 SCC. These were found not to derive from the core SCC sequence in these isolates, but instead to derive from the sequences of plasmids harboured by both isolates (not shown). These plasmids include IS431-like elements which align with those in the MRSA252 SCC (McDougal et al., 2010). Originally reported in the context of the broadening drug resistance profile of MRSA strains causing invasive disease, these plasmids specify multiple drug resistance markers (Ibid.).

The contigs assembled for isolate SA1940 demonstrated the dynamic nature of the Staphylococcal genome. They aligned well with regions of the Mu50 core genome used to appraise assemblies (Figure 27) but poorly across the SCC of this strain (Figure 32 top). Database searches using the SA1940 contigs identified strain S0385 as more closely related to isolate SA1940; however even against the SCC of this strain the contigs of SA1940 matched but discontinuously, with one contig apparently indicating the deletion of much of the central portion of the SCC (Figure 32 lower). The contig providing sequences matching across the mecA operon however was continuous. Analysis of the discontinuously-matching contig suggested mis-assembly.

The contigs assembled for isolates SA374 and SA454 likewise demonstrated the variability of the staphylococcal genome. Database searches identified the unassembled IS-160 strain, which proved to be highly similar to both strains across a region of the core genome used for assembly appraisal (Figure 26). Identification of SCC-specific IS-160 contigs by comparison to 12 previously-characterised SCC sequences concomitantly revealed sequence identity between IS-160 and the SCC of strain CA05 across the region specifying the insertion elements and mec operon (Figure 35, top). Comparisons with these two reference sequences confirmed that both SA374 and SA454 included contigs specifying mecA (Figure 35 b, c) and that SA454 was more similar to CA05 than was SA374.

4.4.3 MecA status & Summary

Comparison of the mecA peptide of the references used to map the isolates analysed here showed that they were highly related, with the previously-identified LGA251 outlier (Figure 37). Extension of the comparison to include the mecA-encoding isolates showed that they included canonical mecA. None expressed the new variant mecA characteristic of LGA251 (Figure 38). The objectives of this thesis have been achieved.

4.4.4 Perspectives and future work

Library quality issues, as determined using Babraham's FastQC package (Andrews, 2011a), together with the earlier FastX package (Gordon, 2009), presented a significant learning opportunity in this work. .

Some aspects of the libraries were excellent. FastQC's 'Per Base N Content' metric indicated that the ambiguous 'N' base call was essentially absent from all libraries. Likewise the 'Sequence Length Distribution' metric indicated a high degree of consistency, with all reads at 105 nucleotides.

Other aspects were typical of the Illumina platform. The gradual fall in base call certainty with read position is platform characteristic, with error rates at the ends of reads as high as 10%. These error rates occur in part as a result of de-synchronisation between the individual molecules comprising a sequence cluster. The abrupt drop in quality in the SA374 library as reads passed 40 was untypical (Figure 4). Likewise the ‘Per Sequence Quality’ metric, which alternatively depicts the mean sequence quality, showed most libraries had a preponderance of high-quality calls with a largely insignificant number of low-quality calls. Some libraries, however, had broad (SA1723) or localised (SA3016, SA374, Figure 3, also SA535, SA303) deviations from this pattern indicating the presence of low-quality base calls.

Other FastQC metrics indicated more systematic issues. The Compositional metrics (Per Sequence GC content, Per Base GC Content and Per Base Sequence Content) indicated that the libraries were not random. While this may appear trivial, it is not; the Illumina base-calling model takes two run-specific parameters from cycle 2 and cumulatively from the first 20; even a cursory examination of a base frequency profile for this region for any library shows it to be far from random (Figure 9). It is not known whether any internal read standard was used in the lane to correct for these deviations.

The Contaminant Identification metrics (Over-represented Sequences, Sequence kmer Content, together with Per Sequence GC content) consistently indicated the presence of sequence contaminants within all libraries.

Lastly, two independent indicators of Read Duplication within the libraries indicated a high degree of reiteration; from the FastQC package, the extent of this duplication was estimated at 66% (median; range 30 - 82%; Table 4), and was determined by the FastX Toolkit Collapser to be 25% across all libraries (median; range 9 - 49%; Table 10). The difference between these figures reflects their differing origins; that from FastX Collapser is an accurate record of the number of sequences removed as the redundancy is collapsed from the totality of each read file. By contrast, that from FastQC is an estimate derived from tallying the instances of the first 200,000 unique reads across the remainder of each read file, using only the first 50bases of each read (Andrews, 2011b). The source of this redundancy is not clear; potential sources include either excess amplification after size-selection or simply an excess material loading onto the sequencing run. The protocol used to generate and sequence the libraries was requested but not made available. Sequencing libraries to such depth wastes resources by consuming reagents but returning no corresponding informational benefit, since repeat reads add little to the sum total information extracted from the library. In addition, such reiterative libraries cannot be used to determine features such as SNPs and INDELs without first discarding the redundant reads. By the determinations of FastX Collapser, such repeats are equivalent to squandering between 10 and 50% of the consumables resource used to sequences these libraries.

Constructing Illumina libraries requires coordination between the intended read length and the size-selected fragment length (Linnarsson, 2010). The present libraries were processed with 105 read cycles. Smalt mapping of entirely unsanitized reads to MRSA252 (from which Figure 2 derived) indicated a median insert size of 140 nucleotides across all libraries (range: 121 - 183). Such read length/fragment length ratios allow insert ‘read-through’, which generate reads including adapter sequences at the distal end (e.g. Figure 39a, b). Given the increased base call uncertainty and error rate which characterise the distal end of reads, removal of such adapters is non-trivial and is consequently time-consuming. Potential solutions during library construction include imposing a lower size-selection boundary with a great margin over the intended read length. Post-hoc solutions include read-merging, originally developed by the Pääbo lab for study of ancient DNAs (Briggs et al., 2010), or error-correction approaches. Read-merging capitalises on the overlap between the paired end reads from short inserts to return higher-quality sequence information for this overlap (Kircher et al., 2011, Magoc and Salzberg, 2011). More specifically, since a sequence at the distal end of one read is necessarily positioned at a shorter, and therefore less error-prone, read length on the other, read-merging permits identification & removal of ‘read-through’ adapter sequences with greater confidence. Still greater confidence can be acquired by implementing error-correction approaches such as Quake (Kelley et al., 2010) after read-merging. Applicable to projects with coverage greater than 15, Quake accumulates run-specific kmer graphs and, by weighting each kmer by its constituent base call qualities, can resolve ‘trusted’ high-coverage, high-quality kmers from single or other low-coverage kmers arising as a result of sequencing errors. Such base call errors are automatically corrected using the high quality call from the ‘trusted’ kmers. Quake preprocessing of datasets before assembly even by an advanced assembler such as Velvet, which implements error-correction in its de Bruijn graph assembly algorithm, results in higher-quality assemblies (*Ibid.*). Similarly, such preprocessing before aligning reads to reference sequences results in more SNP alignments and greater SNP discovery (*Ibid*.).

Future projects of this type with comparable datasets would better make use of a data sanitization workflow in the order below.

1. Read merging using e.g. FLASH (Magoc and Salzberg, 2011)
2. Error correction using e.g. Quake (Kelley et al., 2010)
3. DataQC using e.g. FastQC (Andrews, 2011a)
4. Data sanitisation, followed by repeat data QC using FastQC.

REFERENCES

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

ANDREWS, J. M. & HOWE, R. A. 2011. BSAC standardized disc susceptibility testing method (version 10). *J Antimicrob Chemother,* 66**,** 2726-57.

ANDREWS, S. 2011a. *FastQC, a quality control tool for high throughput sequence data* [Online]. Available: http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/.

ANDREWS, S. 2011b. Interpreting the duplicate sequence plot in FastQC Available from: http://proteo.me.uk/2011/05/interpreting-the-duplicate-sequence-plot-in-fastqc/ 2012].

BOWERS, K. M., WREN, M. W. & SHETTY, N. P. 2003. Screening for methicillin resistance in Staphylococcus aureus and coagulase-negative staphylococci: an evaluation of three selective media and Mastalex-MRSA latex agglutination. *Br J Biomed Sci,* 60**,** 71-4.

BRIGGS, A. W., STENZEL, U., MEYER, M., KRAUSE, J., KIRCHER, M. & PAABO, S. 2010. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res,* 38**,** e87.

BYRNE, M. E., ROUCH, D. A. & SKURRAY, R. A. 1989. Nucleotide sequence analysis of IS256 from the Staphylococcus aureus gentamicin-tobramycin-kanamycin-resistance transposon Tn4001. *Gene,* 81**,** 361-7.

COCK, P. J., FIELDS, C. J., GOTO, N., HEUER, M. L. & RICE, P. M. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res,* 38**,** 1767-71.

COLEMAN, D. C., SULLIVAN, D. J., RUSSELL, R. J., ARBUTHNOTT, J. P., CAREY, B. F. & POMEROY, H. M. 1989. Staphylococcus aureus bacteriophages mediating the simultaneous lysogenic conversion of beta-lysin, staphylokinase and enterotoxin A: molecular mechanism of triple conversion. *J Gen Microbiol,* 135**,** 1679-97.

CUI, L., NEOH, H. M., SHOJI, M. & HIRAMATSU, K. 2009. Contribution of vraSR and graSR point mutations to vancomycin resistance in vancomycin-intermediate Staphylococcus aureus. *Antimicrob Agents Chemother,* 53**,** 1231-4.

DE LENCASTRE, H., OLIVEIRA, D. & TOMASZ, A. 2007. Antibiotic resistant Staphylococcus aureus: a paradigm of adaptive power. *Curr Opin Microbiol,* 10**,** 428-35.

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**,** 1008-15.

EWING, B., HILLIER, L., WENDL, M. C. & GREEN, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res,* 8**,** 175-85.

FRENAY, H. M., BUNSCHOTEN, A. E., SCHOULS, L. M., VAN LEEUWEN, W. J., VANDENBROUCKE-GRAULS, C. M., VERHOEF, J. & MOOI, F. R. 1996. Molecular typing of methicillin-resistant Staphylococcus aureus on the basis of protein A gene polymorphism. *Eur J Clin Microbiol Infect Dis,* 15**,** 60-4.

GARCIA-ALVAREZ, L., HOLDEN, M. T., LINDSAY, H., WEBB, C. R., BROWN, D. F., CURRAN, M. D., WALPOLE, E., BROOKS, K., PICKARD, D. J., TEALE, C., PARKHILL, J., BENTLEY, S. D., EDWARDS, G. F., GIRVAN, E. K., KEARNS, A. M., PICHON, B., HILL, R. L., LARSEN, A. R., SKOV, R. L., PEACOCK, S. J., MASKELL, D. J. & HOLMES, M. A. 2011. Meticillin-resistant Staphylococcus aureus with a novel mecA homologue in human and bovine populations in the UK and Denmark: a descriptive study. *Lancet Infect Dis,* 11**,** 595-603.

GILL, S. R., FOUTS, D. E., ARCHER, G. L., MONGODIN, E. F., DEBOY, R. T., RAVEL, J., PAULSEN, I. T., KOLONAY, J. F., BRINKAC, L., BEANAN, M., DODSON, R. J., DAUGHERTY, S. C., MADUPU, R., ANGIUOLI, S. V., DURKIN, A. S., HAFT, D. H., VAMATHEVAN, J., KHOURI, H., UTTERBACK, T., LEE, C., DIMITROV, G., JIANG, L., QIN, H., WEIDMAN, J., TRAN, K., KANG, K., HANCE, I. R., NELSON, K. E. & FRASER, C. M. 2005. Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant Staphylococcus aureus strain and a biofilm-producing methicillin-resistant Staphylococcus epidermidis strain. *J Bacteriol,* 187**,** 2426-38.

GLADMAN, S. S., TORSTEN. . 2011. *VelvetOptimiser; wrapper software scanning parameters of Velvet to produce an optimal assembly.* [Online]. Available: http://www.bioinformatics.net.au/software.velvetoptimiser.shtml.

GOECKS, J., NEKRUTENKO, A. & TAYLOR, J. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol,* 11**,** R86.

GOERKE, C., PANTUCEK, R., HOLTFRETER, S., SCHULTE, B., ZINK, M., GRUMANN, D., BROKER, B. M., DOSKAR, J. & WOLZ, C. 2009. Diversity of prophages in dominant Staphylococcus aureus clonal lineages. *J Bacteriol,* 191**,** 3462-8.

GORDON, A. 2009. *FASTX-Toolkit; FASTQ/A short-reads pre-processing tools* [Online]. Available: http://hannonlab.cshl.edu/fastx_toolkit/ 2011].

GORDON, D., ABAJIAN, C. & GREEN, P. 1998. Consed: a graphical tool for sequence finishing. *Genome Res,* 8**,** 195-202.

GREEN, P. 1999. *DOCUMENTATION FOR PHRAP AND CROSS_MATCH (VERSION 0.990319)* [Online]. Available: http://www.phrap.org/phredphrap/phrap.html 2012].

HALLIN, M., DEPLANO, A., DENIS, O., DE MENDONCA, R., DE RYCK, R. & STRUELENS, M. J. 2007. Validation of pulsed-field gel electrophoresis and spa typing for long-term, nationwide epidemiological surveillance studies of Staphylococcus aureus infections. *J Clin Microbiol,* 45**,** 127-33.

HALLIN, M., FRIEDRICH, A. W. & STRUELENS, M. J. 2009. spa typing for epidemiological surveillance of Staphylococcus aureus. *Methods Mol Biol,* 551**,** 189-202.

HO SUI, S. J., FEDYNAK, A., HSIAO, W. W., LANGILLE, M. G. & BRINKMAN, F. S. 2009. The association of virulence factors with genomic islands. *PLoS One,* 4**,** e8094.

HOLDEN, M. T., FEIL, E. J., LINDSAY, J. A., PEACOCK, S. J., DAY, N. P., ENRIGHT, M. C., FOSTER, T. J., MOORE, C. E., HURST, L., ATKIN, R., BARRON, A., BASON, N., BENTLEY, S. D., CHILLINGWORTH, C., CHILLINGWORTH, T., CHURCHER, C., CLARK, L., CORTON, C., CRONIN, A., DOGGETT, J., DOWD, L., FELTWELL, T., HANCE, Z., HARRIS, B., HAUSER, H., HOLROYD, S., JAGELS, K., JAMES, K. D., LENNARD, N., LINE, A., MAYES, R., MOULE, S., MUNGALL, K., ORMOND, D., QUAIL, M. A., RABBINOWITSCH, E., RUTHERFORD, K., SANDERS, M., SHARP, S., SIMMONDS, M., STEVENS, K., WHITEHEAD, S., BARRELL, B. G., SPRATT, B. G. & PARKHILL, J. 2004. Complete genomes of two clinical Staphylococcus aureus strains: evidence for the rapid evolution of virulence and drug resistance. *Proc Natl Acad Sci U S A,* 101**,** 9786-91.

HOLMES, A., EDWARDS, G. F., GIRVAN, E. K., HANNANT, W., DANIAL, J., FITZGERALD, J. R. & TEMPLETON, K. E. 2010. Comparison of two multilocus variable-number tandem-repeat methods and pulsed-field gel electrophoresis for differentiating highly clonal methicillin-resistant Staphylococcus aureus isolates. *J Clin Microbiol,* 48**,** 3600-7.

HULETSKY, A., GIROUX, R., ROSSBACH, V., GAGNON, M., VAILLANCOURT, M., BERNIER, M., GAGNON, F., TRUCHON, K., BASTIEN, M., PICARD, F. J., VAN BELKUM, A., OUELLETTE, M., ROY, P. H. & BERGERON, M. G. 2004. New real-time PCR assay for rapid detection of methicillin-resistant Staphylococcus aureus directly from specimens containing a mixture of staphylococci. *J Clin Microbiol,* 42**,** 1875-84.

ILLUMINA. 2012. *Illumina support faqs: questions & answers* [Online]. Available: http://www.illumina.com/support/faqs.ilmn.

ITO, T., KATAYAMA, Y. & HIRAMATSU, K. 1999. Cloning and nucleotide sequence determination of the entire mec DNA of pre-methicillin-resistant Staphylococcus aureus N315. *Antimicrob Agents Chemother,* 43**,** 1449-58.

ITO, T., MA, X. X., TAKEUCHI, F., OKUMA, K., YUZAWA, H. & HIRAMATSU, K. 2004. Novel type V staphylococcal cassette chromosome mec driven by a novel cassette chromosome recombinase, ccrC. *Antimicrob Agents Chemother,* 48**,** 2637-51.

IWG-SCC 2009. Classification of staphylococcal cassette chromosome mec (SCCmec): guidelines for reporting novel SCCmec elements. *Antimicrob Agents Chemother,* 53**,** 4961-7.

JENKINS, D. 2011. *Ion Torrent Releases New 664.13 MB Run* [Online]. Available: http://www.edgebio.com/blog/?p=364.

JONES, M., DURKIN,A.S., HUANG,X.-Z., KIM,M., MCGANN,P., MISHRA,P., NIKOLICH,M., SINGH,I. AND PETERSON,S 2012. Staphylococcus aureus subsp. aureus IS-160, whole genome shotgun sequencing project: AICI01000001-AICI01000221.

KAROW, J. 2011. *As 454 Preps Launch of Sanger-Length Reads, Early Customers Highlight Utility for De Novo Assembly* [Online]. Available: http://www.genomeweb.com/sequencing/454-preps-launch-sanger-length-reads-early-customers-highlight-utility-de-novo-a 2012].

KELLEY, D. R., SCHATZ, M. C. & SALZBERG, S. L. 2010. Quake: quality-aware detection and correction of sequencing errors. *Genome Biol,* 11**,** R116.

KIRCHER, M., HEYN, P. & KELSO, J. 2011. Addressing challenges in the production and analysis of illumina sequencing data. *BMC Genomics,* 12**,** 382.

KUEHNERT, M. J., KRUSZON-MORAN, D., HILL, H. A., MCQUILLAN, G., MCALLISTER, S. K., FOSHEIM, G., MCDOUGAL, L. K., CHAITRAM, J., JENSEN, B., FRIDKIN, S. K., KILLGORE, G. & TENOVER, F. C. 2006. Prevalence of Staphylococcus aureus nasal colonization in the United States, 2001-2002. *J Infect Dis,* 193**,** 172-9.

KUPFERSCHMIDT, K. 2011. Epidemiology. Outbreak detectives embrace the genome era. *Science,* 333**,** 1818-9.

KURTZ, S., PHILLIPPY, A., DELCHER, A. L., SMOOT, M., SHUMWAY, M., ANTONESCU, C. & SALZBERG, S. L. 2004. Versatile and open software for comparing large genomes. *Genome Biol,* 5**,** R12.

LEE, C. Y. & BURANEN, S. L. 1989. Extent of the DNA sequence required in integration of staphylococcal bacteriophage L54a. *J Bacteriol,* 171**,** 1652-7.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LI, H. & HOMER, N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform,* 11**,** 473-83.

LI, S., SKOV, R. L., HAN, X., LARSEN, A. R., LARSEN, J., SORUM, M., WULF, M., VOSS, A., HIRAMATSU, K. & ITO, T. 2011. Novel types of staphylococcal cassette chromosome mec elements identified in clonal complex 398 methicillin-resistant Staphylococcus aureus strains. *Antimicrob Agents Chemother,* 55**,** 3046-50.

LINDSAY, J. A. & HOLDEN, M. T. 2004. Staphylococcus aureus: superbug, super genome? *Trends Microbiol,* 12**,** 378-85.

LINNARSSON, S. 2010. Recent advances in DNA sequencing methods - general principles of sample preparation. *Exp Cell Res,* 316**,** 1339-43.

MAGOC, T. & SALZBERG, S. L. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics,* 27**,** 2957-63.

MAIDEN, M. C., BYGRAVES, J. A., FEIL, E., MORELLI, G., RUSSELL, J. E., URWIN, R., ZHANG, Q., ZHOU, J., ZURTH, K., CAUGANT, D. A., FEAVERS, I. M., ACHTMAN, M. & SPRATT, B. G. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A,* 95**,** 3140-5.

MALACHOWA, N. & DELEO, F. R. 2010. Mobile genetic elements of Staphylococcus aureus. *Cell Mol Life Sci,* 67**,** 3057-71.

MARDIS, E. R. 2011. A decade's perspective on DNA sequencing technology. *Nature,* 470**,** 198-203.

MARKOWITZ, V. M., CHEN, I. M., PALANIAPPAN, K., CHU, K., SZETO, E., GRECHKIN, Y., RATNER, A., JACOB, B., HUANG, J., WILLIAMS, P., HUNTEMANN, M., ANDERSON, I., MAVROMATIS, K., IVANOVA, N. N. & KYRPIDES, N. C. 2012. IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res,* 40**,** D115-22.

MARTIN, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal,* 17.

MCDOUGAL, L. K., FOSHEIM, G. E., NICHOLSON, A., BULENS, S. N., LIMBAGO, B. M., SHEARER, J. E., SUMMERS, A. O. & PATEL, J. B. 2010. Emergence of resistance among USA300 methicillin-resistant Staphylococcus aureus isolates causing invasive disease in the United States. *Antimicrob Agents Chemother,* 54**,** 3804-11.

MELLMANN, A., HARMSEN, D., CUMMINGS, C. A., ZENTZ, E. B., LEOPOLD, S. R., RICO, A., PRIOR, K., SZCZEPANOWSKI, R., JI, Y., ZHANG, W., MCLAUGHLIN, S. F., HENKHAUS, J. K., LEOPOLD, B., BIELASZEWSKA, M., PRAGER, R., BRZOSKA, P. M., MOORE, R. L., GUENTHER, S., ROTHBERG, J. M. & KARCH, H. 2011. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One,* 6**,** e22751.

METZKER, M. L. 2010. Sequencing technologies - the next generation. *Nat Rev Genet,* 11**,** 31-46.

MILLER, J. R., KOREN, S. & SUTTON, G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics,* 95**,** 315-27.

MURCHAN, S., KAUFMANN, M. E., DEPLANO, A., DE RYCK, R., STRUELENS, M., ZINN, C. E., FUSSING, V., SALMENLINNA, S., VUOPIO-VARKILA, J., EL SOLH, N., CUNY, C., WITTE, W., TASSIOS, P. T., LEGAKIS, N., VAN LEEUWEN, W., VAN BELKUM, A., VINDEL, A., LACONCHA, I., GARAIZAR, J., HAEGGMAN, S., OLSSON-LILJEQUIST, B., RANSJO, U., COOMBES, G. & COOKSON, B. 2003. Harmonization of pulsed-field gel electrophoresis protocols for epidemiological typing of strains of methicillin-resistant Staphylococcus aureus: a single approach developed by consensus in 10 European laboratories and its application for tracing the spread of related strains. *J Clin Microbiol,* 41**,** 1574-85.

NCBI_DATABASE:154. Available: http://www.ncbi.nlm.nih.gov/genome/154.

NOVICK, R. P. 2003. Mobile genetic elements and bacterial toxinoses: the superantigen-encoding pathogenicity islands of Staphylococcus aureus. *Plasmid,* 49**,** 93-105.

NOWROUSIAN, M. 2010. Next-generation sequencing techniques for eukaryotic microorganisms: sequencing-based solutions to biological problems. *Eukaryot Cell,* 9**,** 1300-10.

OLIVEIRA, D. C., TOMASZ, A. & DE LENCASTRE, H. 2002. Secrets of success of a human pathogen: molecular evolution of pandemic clones of meticillin-resistant Staphylococcus aureus. *Lancet Infect Dis,* 2**,** 180-9.

PAGANI, I., LIOLIOS, K., JANSSON, J., CHEN, I. M., SMIRNOVA, T., NOSRAT, B., MARKOWITZ, V. M. & KYRPIDES, N. C. 2012. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res,* 40**,** D571-9.

POURCEL, C., HORMIGOS, K., ONTENIENTE, L., SAKWINSKA, O., DEURENBERG, R. H. & VERGNAUD, G. 2009. Improved multiple-locus variable-number tandem-repeat assay for Staphylococcus aureus genotyping, providing a highly informative technique together with strong phylogenetic value. *J Clin Microbiol,* 47**,** 3121-8.

SAMTOOLS_HOME_PAGE. *SAMtools* [Online]. Available: http://samtools.sourceforge.net/.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A,* 74**,** 5463-7.

SCHIJFFELEN, M. J., BOEL, C. H., VAN STRIJP, J. A. & FLUIT, A. C. 2010. Whole genome analysis of a livestock-associated methicillin-resistant Staphylococcus aureus ST398 isolate from a case of human endocarditis. *BMC Genomics,* 11**,** 376.

SHENDURE, J. & JI, H. 2008. Next-generation DNA sequencing. *Nat Biotechnol,* 26**,** 1135-45.

SHORE, A. C., DEASY, E. C., SLICKERS, P., BRENNAN, G., O'CONNELL, B., MONECKE, S., EHRICHT, R. & COLEMAN, D. C. 2011. Detection of staphylococcal cassette chromosome mec type XI carrying highly divergent mecA, mecI, mecR1, blaZ, and ccr genes in human clinical isolates of clonal complex 130 methicillin-resistant Staphylococcus aureus. *Antimicrob Agents Chemother,* 55**,** 3765-73.

SOLID_HOMEPAGE. 2012. *ABI SOLiD homepage* [Online]. Available: http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems.html.

STOTHARD, P. *compare_library.pl; a script accepting two multi-fasta files and running bl2seq for all entry permutations.* [Online]. Available: http://www.auburn.edu/~santosr/scripts/compare_library.prl [Accessed 12 Feb 2012 2012].

STROMMENGER, B., KETTLITZ, C., WENIGER, T., HARMSEN, D., FRIEDRICH, A. W. & WITTE, W. 2006. Assignment of Staphylococcus isolates to groups by spa typing, SmaI macrorestriction analysis, and multilocus sequence typing. *J Clin Microbiol,* 44**,** 2533-40.

VON EIFF, C., BECKER, K., MACHKA, K., STAMMER, H. & PETERS, G. 2001. Nasal carriage as a source of Staphylococcus aureus bacteremia. Study Group. *N Engl J Med,* 344**,** 11-6.

WATSON, J. D. & CRICK, F. H. 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. J.D. Watson and F.H.C. Crick. Published in Nature, number 4356 April 25, 1953. *Nature,* 248**,** 765.

YOON, S. H., PARK, Y. K., LEE, S., CHOI, D., OH, T. K., HUR, C. G. & KIM, J. F. 2007. Towards pathogenomics: a web-based resource for pathogenicity islands. *Nucleic Acids Res,* 35**,** D395-400.

ZERBINO, D. R. 2010. Using the Velvet de novo assembler for short-read sequencing technologies. *Curr Protoc Bioinformatics,* Chapter 11**,** Unit 11 5.

ZERBINO, D. R. & BIRNEY, E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res,* 18**,** 821-9.

APPENDICES

Appendix A (Pre-processing shell script)

Pre-processing the 10 MRSA libraries was automated using a custom shell script, postFastQC1_v10.sh (cd). The script remains sample and adapter-agnostic by reading sample names and adapter sequences from external files (samplesnames.txt, adapterseqs.txt; cd). The script coordinates calls to the core pre-processing program cutadapt, which removes contaminant adapter sequences and low sequence quality regions (Martin, 2011). In addition, the script makes calls to ancillary scripts allowing re-analysis of the pre-processed data by FastQC to be automated. Lastly, the script self-documents its output by calling custom shell library functions which write into each output directory files (‘data_trail.README'; examples on cd) which document directory content origins.

The first ancillary script (syncReads_wrapper_v5.pl; cd) synchronises the reads in the cutadapt-processed files. Loss of synchronisation between paired end read files occurs when pre-processing results in the loss of one read of a pair. Removal of the resultant ‘other’ read is necessary for most assemblers, which disallow such singleton reads within paired-end read libraries. Synchronisation output directories are self-documented as described above and also by the sync script writing a ‘sync.metrics’ file to disk (cd).

The second ancillary script, shuffleSequences_fastq.pl, is provided with the Velvet installation and ‘interleaves’ the reads of each pair into a single file. Not only is this required for subsequent assembled by Velvet, but also usefully reduces by 50% the number of files for submission to the last program called, FastQC. ‘Shuffled’ output directories are self-documented as described above, as is the FastQC output directory. .

Appendix B (Read-through sequence analysis)

The position of adapter sequence in library reads was examined using a shell script, readthruseqanal_v3.sh (cd). As an early shell script learning opportunity this script has a somewhat idiosyncratic structure using arrays where a hash would now be implemented. The script was made sample-agnostic by pulling sample names from an external file, samplenames.txt. A second external file, readthrusequences.txt, specifies the generic universal and sample-specific indexed adapter sequences then thought likely to be found at the 3' ('read-through') of reads. The content of a third file, sampleindexing.txt, serves as an adapter linking sample names to adapter sequences. The script indexes through the sample list, using grep to count the lines on which adapter sequence occur both in total and at the start of each read. By returning line counts of each instance, grep accurately counts the 5' instances but understates the total number of adapter inserts, since lines with multiple instances will be recorded as a single instance.

Appendix C (Subsetting for Velvet)

Velvet is sensitive to the number of reads provided, and assembly accuracy suffers when excess reads are input (Zerbino, 2010). The shell script subset4velvet_v1.sh (cd) was written to subset libraries according to the apparent sequence coverage relative to the MRSA 252 reference sequence. The script is generic, pulling required parameters from the external files samplenames.txt and numberofbins.txt, and using calls to sed to subset the input files.