Identifying mixed *Mycobacterium tuberculosis* infection and laboratory cross-contamination during Mycobacterial sequencing programs

Authors

David H Wyllie [1, 2, 3] *, Esther Robinson[4], Tim Peto[1,3], Derrick W Crook[1,3], Adebisi Ajileye[4], Priti Rathod[4], Rosemarie Allen[4], Lisa Jarrett[4], E Grace Smith[4], A Sarah Walker[1,3]

Affiliations

[1] Nuffield Department of Medicine, John Radcliffe Hospital, Headley Way, Oxford OX3 9DU, UK

[2] Public Health England Academic Collaborating Centre, John Radcliffe Hospital, Headley Way, Oxford OX3 9DU, UK

[3] The National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford

[4] Public Health England National Mycobacteriology Laboratory North and Central, Heartlands Hospital, Birmingham B9 5SS

Keywords

Mycobacterium tuberculosis, next generation sequencing, quality control, mixed infection, cross-contamination

ABSTRACT

Introduction: Detecting laboratory cross-contamination and mixed tuberculosis infection are important goals of clinical Mycobacteriology laboratories.

Objectives: To develop a method detecting mixtures of different *M. tuberculosis* lineages in laboratories performing Mycobacterial next generation sequencing (NGS).

Setting: Public Health England National Mycobacteriology Laboratory Birmingham, which performs Illumina sequencing on DNA extracted from positive Mycobacterial Growth Indicator tubes.

Methods: We analysed 4,156 samples yielding *M. tuberculosis* from 663 MiSeq runs, obtained during development and production use of a diagnostic process using NGS. Counts of the most common (major) variant, and all other variants (non-major variants) were determined from reads mapping to positions defining *M. tuberculosis* lineages. Expected variation was estimated during process development.

Results: For each sample we determined the non-major variant proportions at 55 sets of lineage defining positions. The non-major variant proportion in the two most mixed lineage defining sets (F2 metric) was compared with that in the 47 least mixed lineage defining sets (F47 metric). Three patterns were observed: (i) not mixed by either metric, (ii) high F47 metric suggesting mixtures of multiple lineages, and (iii) samples compatible with mixtures of two lineages, detected by differential F2 metric elevation relative to F47. Pattern (ii) was observed in batches, with similar patterns in the H37Rv control present in each run, and is likely to reflect cross-contamination. During production, the proportions of samples in each pattern were 97%, 2.8%, and 0.001%, respectively.

Conclusion: The F2 and F47 metrics described could be used for laboratory process control in laboratories sequencing *M. tuberculosis*.

(249 words)

INTRODUCTION

Mycobacterium tuberculosis is an organism which has co-evolved with humans during the early migrations of modern man, diverging from a common *M. tuberculosis* ancestor about 75,000 years ago (1). Distinct lineages, corresponding to evolution occurring during these early migrations, are readily identified by next generation sequencing (NGS), with each lineage characterised by ancient single nucleotide variants (SNV) which define deep branches in the *M. tuberculosis* phylogeny (1, 2).

Multiple *M. tuberculosis* lineages: Infection by multiple lineages of TB is well described, and has been detected by observing mixed results on fractional sequencing (e.g. spoligotyping, MIRU-VNTR) and validated by characterisation of multiple individual picks from solid media (3). Multi-lineage infection is characterised by isolates differing by many hundreds of SNVs, in which respect it differs from the increasingly recognised and more common in-host microevolution (4). Reported rates of mixed infection vary markedly as reviewed (3, 5), with between 10 and 30% reported in areas of current (5-7) or historical (8, 9) high prevalence. Much lower rates are reported in low incidence countries (3), although systematic under-detection is likely to occur due to both to the limited representation of bacteria in single sputum samples of pulmonary disease, and to the decrease in diversity occurring during differential strain growth in broth culture (10).

Implications of mixed infection: Mixed infection, assessed by either MIRU-VNTR polymorphisms (7) or by heterogeneity in drug susceptibility testing (11), is independently associated with reduced drug treatment response, so there are compelling clinical reasons to try to identify it. There are also important technical implications of isolating mixed TB strains from a culture. Firstly, such a finding may reflect cross-contamination within the laboratory (3). Secondly, mixed infection complicates the interpretation of drug resistance tests, whether phenotypic or genotypic, as one or other co-infecting strain may dominate the results from these tests. Thirdly, it complicates the understanding of relatedness when techniques such as SNV distance computation are applied, as these generally assume a single sequence is present when base-calling (5, 12-14), marking mixed sites as uncertain. Maximal likelihood tree drawing algorithms assume such 'uncertain' sites contain no information and impute a single nucleotide at each of such positions, an approach which may be inappropriate in the presence of mixtures.

Increasingly, NGS based species and resistance determination is becoming routine in Mycobacteriology laboratories, and has been deployed in reference laboratories in England (15, 16). As part of the quality control and accreditation of the routine process now operating in these laboratories, we describe an approach to identifying mixed samples using Illumina next generation

sequencing data, illustrating its use by studying over 4,000 consecutive positive cultures from a single reference laboratory.

METHODS

Isolation of DNA from Mycobacteria and sequencing

This study includes all mycobacteria processed between 01/06/2015 and 30/12/2017 in the Public Health England Midlands and North reference laboratory, whose catchment is approximately 15 million people, or about one third of England. Clinical specimens were decontaminated and inoculated into Mycobacterial Growth Indicator Tubes (MGIT) tubes. During the process (Fig. 1) positive MGIT tubes were batched when they became available, either following growth in the local laboratory, or following receipt from another laboratory. Positive control samples (H37Rv strain) were also grown in MGIT cultures. Batches of positive samples were extracted using a manual process, exactly as described in the Supplementary Methods in (16). Illumina sequencing libraries were prepared using Nextera XT chemistry from equal amounts of 12 (20.04.15-01.08.15) or 16 (01.08.15-30.12.17) Mycobacterial DNA extracts (16), using manual steps (Table 1). Positive control DNA (H37Rv, obtained from ATCC) was included as one of the 12 or 16 extracts in all libraries, either from a contemporaneously extracted broth culture, or from stored DNA. Libraries were loaded into an Illumina MiSeq instrument and sequenced (16).

Bioinformatic processing

The routine bioinformatics pipeline deployed by Public Health England has been previously described (15). Briefly, reads were first processed using the Mykrobe predictor tool, which identifies *Mycobacterium tuberculosis* using species specific k-mers (17). Specimens identified as containing *M. tuberculosis* were further processed (16), mapped to the H37Rv v2 genome (NC_000962.2) (18), as described (16), and vcf files generated using Samtools mPileup, with additional basecalling using GATK VariantAnnotator v2.1. A consensus base is called from high-quality bases provided one base accounts from >90% of the pileup; otherwise, the base is recorded as uncertain ('N') (15). In this analysis, high quality base counts (identified by the BaseCounts VCF tag) were extracted and summarised using code available at https://github.com/davidhwyllie/VCFMIX.

Nucleotides identifying lineage

Coll *et al* (2) described the *M. tuberculosis* phylogeny and identified 62 sets of nucleotide positions defining the deep branches of the *Mycobacterium tuberculosis* lineage. At each position within a nucleotide set, in one particular clade, one nucleotide is uniquely present (i.e. is not present in any

other of the known clades). These sets contain a median 108 nucleotide positions (range 1 – 898). In this analysis, we considered 55 branches, excluding branches 1.2, 3.1, 3.1.2, 4.1.2, 4.3.4.2.1, 4.6 and 4.7 because they contain fewer than 20 positions, making estimates of minor variation in these positions less reliable than estimates in other branches. We identified lineage using consensus basecalling in these 55 branches. If the signature SNV of a branch was called as uncertain, we called only to the level of the branch deeper (i.e. closer to the root) than the uncertain call. If more than one different lineage defining variant was called, or we could not call any lineage defining positions, we reported the samples as 'lineage not defined'.

Estimation of the minor variant frequency within a set

The minor variant frequency at a set of bases can be due to sequencing error, mapping error and/or *bona fide* inter-lineage mixtures (Supplementary Figure 1, panels A and B). Minor variant frequencies were determined as follows: if there are n bases in a lineage defining set, we count the high quality depths d at each base, e.g. if n=3 and d_1 = 30, d_2= 70, d_3= 100, then the total depth $D = \sum_{i=1}^{n} d_i$ = 200. For each position, we also identify the most common base; the minor depth m is the total depth minus the most common base depth. If the minor depths are m_1 = 3, m_2= 7, m_3= 10, then total minor depth $M = \sum_{i=1}^{n} m_i$ = 20; we estimate the minor allele fraction p in the set as M/D = 0.1.

F2 and F47 metrics

If sequences from two different *M. tuberculosis* lineages are mixed together, then the sets which uniquely define these lineages will be mixed (depicted in Supplementary Figure 1C); there will be a minimum of two and maximum of eight sets affected (e.g. a Lineage 5/7 mixture will mix two sets of lineage defining nucleotides, a 2.1/4.2.1 will mix five sets, and a 4.1.1.1 / 3.1.2.1 mixture will mix 8 sets). Only if more than two samples are mixed will more than 8 sets be mixed. In this work, we describe two metrics reflecting mixing. Having computed the minor allele frequency estimates, p_1, p_2, .. p_{55} we can sort these in descending order, identifying the sets with the highest and lowest minor allele frequencies. We then estimate the minor variant frequency across the nucleotides in the top two (F2 metric) and lowest 47 sets (F47 metric) as described above. The underpinning assumptions are that mixtures of biological origin are most likely to occur between two lineages, and therefore F2 is the most sensitive metric for identifying these. Since between two and 8 sets will be mixed in such genuine co-infections, then the lowest 47 (55 - 8) sets will not be mixed, and thus the F47 metric is more sensitive for identifying laboratory contamination involving more than two samples.

Regression modelling

Because of high leverage by a small number of observations, we used quantile regression to estimate the relationship between median values of log-transformed non-callable base numbers and log-transformed F47, using the quantreg R package (R 3.3.1).

Ethical framework

Only anonymised data was used in this work; ethical approval is not required.

Data availability

The data analysed is available at https://ora.ox.ac.uk/objects/uuid: 5e4ec1f8-e212-47db-8910-161a303a0757.

RESULTS

Samples studied

4,156 samples were included since they (i) were identified using MyKrobe (17) as belonging to the *M. tuberculosis* complex, and (ii) had at least 0.5 x 10^6 read pairs mapped to the H37Rv reference genome, a criterion reflecting successful DNA extraction and sequencing. These sequences were highly diverse, originating from six branches of lineage 1 (n=320), five branches from lineage 2 (n=278), five branches of lineage 3 (n=1,010), thirty lineage 4 branches (n=2,266). 106 samples were from *M. bovis* or *africanum*, and 176 did not have their lineages defined.

The laboratory processes operated under three different phases: in the first, *development*, laboratory processes were being actively refined; in the second, *pre-production*, laboratory processes were fixed and controlled by standard operating procedures, with version controlled changes. The third *production* state was similar to the second, except that the process had received ISO15189 accreditation.

Variation in lineage defining positions in H37Rv controls

The F2 mixture metric reflects the estimated mixture in the two most mixed lineage defining sets; in the H37Rv controls, this follows a distribution skewed to the right (Figure 2A). The MiSeq runs with H37Rv controls with F2 mixture metrics in the top 5% (Fig. 2A) are temporally clustered (red lines on Fig. 2B, C) with a number of examples in the Development phase. Among clinical (non-control) samples, variation in F2 metric is explained in part by MiSeq run (Kruskal-Wallis test, $p < 10^{-16}$), and a strong correlation exists between the F2 metric in H37Rv controls and that in clinical samples on the same plate ($\rho = 0.61$, 95% CI 0.56-0.61, Spearman's Rank Correlation). That is, in plates with elevated F2 metrics in the H37Rv control, the clinical samples are more likely to have elevated F2, as is evident visually (e.g. Fig 2B and C, around run 2301).

Different patterns of mixtures were observed during Development

We ordered specimens first by the order of the plates analysed and the order in which the bioinformatics processing was completed, which is the order that an automated quality control monitoring system would encounter output. During the Development phase (Fig. 3), blocks of samples derived from runs with elevated mixtures in the H37Rv control are seen (red bars in Fig. 3A), coincident with clear increases in both F2 and F47 metrics (Fig. 3B, C), reflecting elevations in mixed bases across most or all lineage defining positions (Fig. 3D). These blocks of samples typically span multiple MiSeq runs (Fig 3A,B). In addition to the blocks of samples with elevated F2 and F47 metrics, we also observed small numbers of single samples with elevated F2, but not F47, metrics (Fig. 3B, Arrow). This latter pattern is expected in cases of inter-lineage mixtures of only two

183 samples (Supp. Fig 1, and Methods). These patterns were also seen in the subsequent phases
184 (Supplementary Figures S2-5, yellow dots).

185 ## Mixtures of multiple lineages are common

186 Based on the pattern observed in the Development phase, we categorised samples as having one of
187 (i) neither F2 nor F47 raised, (ii) raised F47, or (iii) raised F2 without raised F47. We defined a raised
188 F2 and F47 as more than 10x and 5x the respective median metric during Development in all control
189 and clinical samples, cutoffs which correspond to 4.7% (F2) and 0.2% (F47) minor variant frequencies
190 across the relevant lineage-defining sets, respectively. In the Pre-production and Production phases,
191 F2 and F47 values below these thresholds (reflecting unmixed samples) were observed in 97.5% of
192 the samples studied, raised F47 and F2 values (reflecting a mixture of multiple samples) in 2.5%,
193 while six samples (0.001%) had raised F2 but normal F47 values (Table 2).

194 ## Isolated F2 metric elevation is rare

195 Isolated elevation of F2 is expected if bacteria from two different lineages are mixed. In Fig. 4, we
196 show the minor variant frequencies from all samples from the six individuals with raised F2, but
197 normal F47, metrics. In one case, patient 3, two technical repeats of the same sample (#2) showed
198 the same pattern, as did a separate sample taken contemporaneously. In other cases (patients 1, 5,
199 6) the mixed pattern was only observed in one out of two positive samples taken on the same day,
200 and in two cases (2, 3) only a single sample was positive. Thus, between 1 and 6 samples of the
201 4,156 studied may truly reflect mixed co-infections.

202 ## Impact of inter-lineage variation on basecalling

203 One obvious question is whether very low level cross-contamination impacts the consensus
204 sequence which can discerned from the pileup. As cross-contamination increases, at some point
205 minor variant frequencies in some parts of the genome will start to rise above the 10% cutoff
206 specified by the basecalling algorithm. The numbers of uncalled bases will then rise; this
207 relationship can be observed in Figure 5: the number of uncallable bases rises rapidly when F47
208 exceeds the cutoff value, but only slowly below it. Below the cutoff value of 4.7% (red line in Fig. 5),
209 which is 10 times the median (black line in Fig. 5), the number of uncallable bases increased by 1.25
210 fold (95% CI 1.22, 1.28) for every 10-fold increase in F47; above the cutoff, the corresponding
211 increase was 9.24 (95% CI 5.5, 11.2).

212

DISCUSSION

In this work, we describe methods for monitoring the presence of mixtures of *Mycobacterium tuberculosis* of different lineages. It assumes that multiple lineages and sublineages of *M. tuberculosis* are being sequenced contemporaneously; this is the case in our setting, and is also true globally (19, 20). Using single nucleotide variants, each of which uniquely defines a branch in the phylogenetic tree of *M. tuberculosis*, we can show two patterns of mixtures. The first, which occurred in about 2.5% of samples during the pre-production and production phases of our project, is indicative of multiple samples being mixed together, since mixtures are seen in most or all of the lineage defining branches. This occurred in batches, was characterised by cross-contamination at levels of less than 1%, and can be monitored by a metric we term F47. This pattern likely reflects process failures. The strength of the F47 metric is that the depth analysed is very high, as about 5,000 nucleotides typically contribute across the lineage-defining sets included in it. If there is a sequencing depth of 50-100 at each of these, the effective sequencing depth analysed is of the order of 25,000 – 50,000, making detection of minor variation at sub- 1% levels readily feasible with high statistical confidence.

Such low level cross-contamination, as observed during our production process and illustrated in Supplementary Figures 2-4, is likely to have minimal influence on inference drawn from the sequence, unless highly sensitive assays for heteroresistance are required. A sensitive metric such as F47 will, however, allow early detection of emerging problems and allow review of process, as part of continuous quality improvement.

A second class of mixture, which we found was rarely detected in this setting, is compatible with co-infection with two organisms of differing lineage within the patient. This kind of mixture is clinically relevant (7, 11), and may be under-detected using the laboratory process we describe here, since culture based amplification can reduce diversity in the sample inoculated (10). Its frequency may rise if direct-from-sample sequencing is employed, or if samples from high-endemicity areas are studied, but here we identified only one probable case of such mixtures, and five other possible cases. Confirmatory approaches are available: microbiological techniques conducted separately on multiple picks from the same samples have been used as validation(3); a limitation of this study is that we could not undertake such work as only multiply sub-cultured stored isolates exist for historical samples. Techniques for reconstructing the contributing sequences also have been described in detail (21-23), and we did not study them here. Another limitation is that we were not able to consider six of the lineage defining sets in Coll *et al* because they covered <20 nucleotide positions and therefore we considered that they did not contain sufficient information to be used in

F2 or F47 metrics. A consequence is that our method would not identify mixtures of samples if they only involved mixtures in these excluded branches.

The clinical of use of bacterial genome sequencing is rising (16, 17, 24), and given the importance of *M. tuberculosis* and the complexity of treatment, *M. tuberculosis* has been one of the first organisms tackled (15). The processes followed involve multiple steps at which the opportunity for cross-contamination exists. The availability of tools monitoring critical aspects of laboratory process is required for accreditation under ISO15189, and the F2 and F47 metrics described here will contribute to this.

ACKNOWLEDGEMENTS

This study is supported by the Health Innovation Challenge Fund (a parallel funding partnership between the Wellcome Trust [WT098615/Z/12/Z] and the Department of Health [grant HICF-T5-358]) and NIHR Oxford Biomedical Research Centre. Professors Derrick Crook, Tim Peto and Sarah Walker are affiliated to the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Healthcare Associated Infections and Antimicrobial Resistance at University of Oxford in partnership with Public Health England. Professor Crook is based at University of Oxford. Professor Tim Peto is an NIHR Senior Investigator. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health or Public Health England. The sponsors of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

FIGURE LEGENDS

Figure 1 Laboratory and Bioinformatic processing

Figure 2 Mixtures in H37Rv controls

Histogram showing the F2 metric, which reflects the mixture in the two most mixed lineage associated sets, in H37Rv control DNA (A). Median F2 metric among clinical samples other than H37Rv is shown in (B); red lines indicate that the F2 mixture metric in H37Rv controls is raised (as shown in A). (C) shows F2 metric for each *M. tuberculosis* sequence from a clinical sample.

Figure 3 Mixture metrics in the Development phase

Samples arranged first by the order of the MiSeq runs (depicted as solid gray blocks, in A), and the order bioinformatics processing completed. Only samples yielding *M. tuberculosis* are shown, which is why some blocks in A are longer than others. If the H37Rv control samples had increased F2 statistics, a red bar is shown above each sample in A. We depicted the F2 (B) and F47 metrics (C), as well as the estimated mixture F in each of the 58 lineage defining sets (D). The arrow illustrates a sample with elevated F2, but low F47 metric.

Figure 4 Consistency of isolated F2 elevation in individuals

Six individuals with elevated F2, but not F47, statistics were identified during the pre-production and production phases. The observed minor variant proportion for all deep branches analysed are shown in a heatmap. For example, patient #4 had two samples taken in December 2015; sample 2 was analysed twice (sequencing ids 0d9d5,9276f) and sample 3 once. A similar pattern of minor variation is seen in all three samples.

Figure 5 F47 metric and the number of callable bases

The relationship between the F47 metric and the number of uncallable bases. The red line corresponds to the cutoff used to define F47 as being elevated.

Supplementary Figures

Supplementary Figure 1

Illustrates the computation of variation between samples of two different lineages, Lineage 1 (black text) and Lineage 2 (blue text) (A). When a mixture of these samples is present, and mapped to a reference sequence, a major base and minor base(s) are present in the pileup (B). Variation may be due to either sequencing error (underlined) or to lineage associated variation; the non-major variant frequency included both classes of variation. Lineage defining sites, as defined by Coll *et al* (2), mark branches of the phylogenetic tree. If a lineage 4.1.1.1 *M. tuberculosis* is mixed with a lineage 3.1.2.1 *M. tuberculosis*, eight sets of lineage defining sites will be mixed (red boxes).

298 *Supplementary Figures 2-5*

299 These illustrate mixture patterns observed during the Production stage. The layout is similar to
300 Figure 3; samples arranged first by the order of the MiSeq runs (depicted as solid gray blocks, in A),
301 and the order bioinformatics processing completed. Only samples yielding *M. tuberculosis* samples
302 are shown, which is why some blocks in A are longer than others. If the H37Rv control samples had
303 increase F2 statistics, a red bar is shown above each sample in A. We depicted the F2 metric (B) and
304 F47 metrics (C), as well as the estimated mixture F in each of the 58 lineage defining sets (D).

305

306

TABLES

Table 1 Samples analysed

Development Stage	Total sequences (samples and controls)	Date range	Number of individuals providing samples	MiSeq run range	Number of clinical samples	Number of MiSeq Runs
Development	938	20.04.2015-15.12.2015	630	101 .. 291	776	154
Pre-Production	1167	01.04.2016-06.12.2016	753	1152 ..1522	919	163
Production	2191	07.12.16-30.12.2017	1481	1523 .. 2307	1794	346

Table 2 Detection of mixtures in clinical samples

Development Stage	Neither F2 nor F47 raised	F2 raised, but F47 normal	F47 raised
Pre-Production (n=919)	900 (98%)	5 (0.003%)	14 (1.1%)
Production (n=1794)	1741 (97%)	1 (0.001%)	52 (2.8%)

REFERENCES

1. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, Yeboah-Manu D, Bothamley G, Mei J, Wei L, Bentley S, Harris SR, Niemann S, Diel R, Aseffa A, Gao Q, Young D, Gagneux S. 2013. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet 45:1176-82.
2. Coll F, McNerney R, Guerra-Assuncao JA, Glynn JR, Perdigao J, Viveiros M, Portugal I, Pain A, Martin N, Clark TG. 2014. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. Nat Commun 5:4812.
3. Cohen T, van Helden PD, Wilson D, Colijn C, McLaughlin MM, Abubakar I, Warren RM. 2012. Mixed-strain mycobacterium tuberculosis infections and the implications for tuberculosis treatment and control. Clin Microbiol Rev 25:708-19.
4. Lieberman TD, Wilson D, Misra R, Xiong LL, Moodley P, Cohen T, Kishony R. 2016. Genomic diversity in autopsy samples reveals within-host dissemination of HIV-associated Mycobacterium tuberculosis. Nat Med 22:1470-1474.
5. Bryant JM, Harris SR, Parkhill J, Dawson R, Diacon AH, van Helden P, Pym A, Mahayiddin AA, Chuchottaworn C, Sanne IM, Louw C, Boeree MJ, Hoelscher M, McHugh TD, Bateson AL, Hunt RD, Mwaigwisya S, Wright L, Gillespie SH, Bentley SD. 2013. Whole-genome sequencing to establish relapse or re-infection with Mycobacterium tuberculosis: a retrospective observational study. Lancet Respir Med 1:786-92.
6. Wang X, Liu H, Wei J, Wu X, Yu Q, Zhao X, Lyu J, Lou Y, Wan K. 2015. An investigation on the population structure of mixed infections of Mycobacterium tuberculosis in Inner Mongolia, China. Tuberculosis (Edinb) 95:695-700.
7. Cohen T, Chindelevitch L, Misra R, Kempner ME, Galea J, Moodley P, Wilson D. 2016. Within-Host Heterogeneity of Mycobacterium tuberculosis Infection Is Associated With Poor Early Treatment Response: A Prospective Cohort Study. J Infect Dis 213:1796-9.
8. Kay GL, Sergeant MJ, Zhou Z, Chan JZ, Millard A, Quick J, Szikossy I, Pap I, Spigelman M, Loman NJ, Achtman M, Donoghue HD, Pallen MJ. 2015. Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. Nat Commun 6:6717.
9. Chan JZ, Sergeant MJ, Lee OY, Minnikin DE, Besra GS, Pap I, Spigelman M, Donoghue HD, Pallen MJ. 2013. Metagenomic analysis of tuberculosis in a mummy. N Engl J Med 369:289-90.
10. Plazzotta G, Cohen T, Colijn C. 2015. Magnitude and sources of bias in the detection of mixed strain M. tuberculosis infection. J Theor Biol 368:67-73.
11. Zetola NM, Modongo C, Moonan PK, Ncube R, Matlhagela K, Sepako E, Collman RG, Bisson GP. 2014. Clinical outcomes among persons with pulmonary tuberculosis caused by Mycobacterium tuberculosis isolates with phenotypic heterogeneity in results of drug-susceptibility tests. J Infect Dis 209:1754-63.
12. Walker TM, Lalor MK, Broda A, Ortega LS, Morgan M, Parker L, Churchill S, Bennett K, Golubchik T, Giess AP, Del Ojo Elias C, Jeffery KJ, Bowler I, Laurenson IF, Barrett A, Drobniewski F, McCarthy ND, Anderson LF, Abubakar I, Thomas HL, Monk P, Smith EG, Walker AS, Crook DW, Peto TEA, Conlon CP. 2014. Assessment of Mycobacterium tuberculosis transmission in Oxfordshire, UK, 2007-12, with whole pathogen genome sequences: an observational study. Lancet Respir Med 2:285-292.
13. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TE. 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis 13:137-46.
14. Guerra-Assuncao JA, Houben RM, Crampin AC, Mzembe T, Mallard K, Coll F, Khan P, Banda L, Chiwaya A, Pereira RP, McNerney R, Harris D, Parkhill J, Clark TG, Glynn JR. 2015. Recurrence due to relapse or reinfection with Mycobacterium tuberculosis: a whole-genome sequencing

approach in a large, population-based cohort with a high HIV infection prevalence and active follow-up. J Infect Dis 211:1154-63.

15. Quan TP, Bawa Z, Foster D, Walker T, Del Ojo Elias C, Rathod P, Iqbal Z, Bradley P, Mowbray J, Walker AS, Crook DW, Wyllie DH, Peto TEA, Smith EG. 2017. Evaluation of whole genome sequencing for Mycobacterial species identification and drug susceptibility testing in a clinical setting: a large-scale prospective assessment of performance against line-probe assays and phenotyping. J Clin Microbiol doi:10.1128/jcm.01480-17.

16. Pankhurst LJ, Del Ojo Elias C, Votintseva AA, Walker TM, Cole K, Davies J, Fermont JM, Gascoyne-Binzi DM, Kohl TA, Kong C, Lemaitre N, Niemann S, Paul J, Rogers TR, Roycroft E, Smith EG, Supply P, Tang P, Wilcox MH, Wordsworth S, Wyllie D, Xu L, Crook DW. 2016. Rapid, comprehensive, and affordable mycobacterial diagnosis with whole-genome sequencing: a prospective study. Lancet Respir Med 4:49-58.

17. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, Earle S, Pankhurst LJ, Anson L, de Cesare M, Piazza P, Votintseva AA, Golubchik T, Wilson DJ, Wyllie DH, Diel R, Niemann S, Feuerriegel S, Kohl TA, Ismail N, Omar SV, Smith EG, Buck D, McVean G, Walker AS, Peto TE, Crook DW, Iqbal Z. 2015. Rapid antibiotic-resistance predictions from genome sequence data for Staphylococcus aureus and Mycobacterium tuberculosis. Nat Commun 6:10063.

18. Lunter G, Goodson M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21:936-9.

19. Stucki D, Brites D, Jeljeli L, Coscolla M, Liu Q, Trauner A, Fenner L, Rutaihwa L, Borrell S, Luo T, Gao Q, Kato-Maeda M, Ballif M, Egger M, Macedo R, Mardassi H, Moreno M, Tudo Vilanova G, Fyfe J, Globan M, Thomas J, Jamieson F, Guthrie JL, Asante-Poku A, Yeboah-Manu D, Wampande E, Ssengooba W, Joloba M, Henry Boom W, Basu I, Bower J, Saraiva M, Vaconcellos SEG, Suffys P, Koch A, Wilkinson R, Gail-Bekker L, Malla B, Ley SD, Beck HP, de Jong BC, Toit K, Sanchez-Padilla E, Bonnet M, Gil-Brusola A, Frank M, Penlap Beng VN, Eisenach K, Alani I, Wangui Ndung'u P, et al. 2016. Mycobacterium tuberculosis lineage 4 comprises globally distributed and geographically restricted sublineages. Nat Genet 48:1535-1543.

20. Manson AL, Cohen KA, Abeel T, Desjardins CA, Armstrong DT, Barry CE, 3rd, Brand J, Chapman SB, Cho SN, Gabrielian A, Gomez J, Jodals AM, Joloba M, Jureen P, Lee JS, Malinga L, Maiga M, Nordenberg D, Noroc E, Romancenco E, Salazar A, Ssengooba W, Velayati AA, Winglee K, Zalutskaya A, Via LE, Cassell GH, Dorman SE, Ellner J, Farnia P, Galagan JE, Rosenthal A, Crudu V, Homorodean D, Hsueh PR, Narayanan S, Pym AS, Skrahina A, Swaminathan S, Van der Walt M, Alland D, Bishai WR, Cohen T, Hoffner S, Birren BW, Earl AM. 2017. Genomic analysis of globally diverse Mycobacterium tuberculosis strains provides insights into the emergence and spread of multidrug resistance. Nat Genet 49:395-402.

21. Eyre DW, Cule ML, Griffiths D, Crook DW, Peto TE, Walker AS, Wilson DJ. 2013. Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in Clostridium difficile transmission. PLoS Comput Biol 9:e1003059.

22. Pulido-Tamayo S, Sanchez-Rodriguez A, Swings T, Van den Bergh B, Dubey A, Steenackers H, Michiels J, Fostier J, Marchal K. 2015. Frequency-based haplotype reconstruction from deep sequencing data of bacterial populations. Nucleic Acids Res 43:e105.

23. Gan M, Liu Q, Yang C, Gao Q, Luo T. 2016. Deep Whole-Genome Sequencing to Detect Mixed Infection of Mycobacterium tuberculosis. PLoS One 11:e0159029.

24. De Silva D, Peters J, Cole K, Cole MJ, Cresswell F, Dean G, Dave J, Thomas DR, Foster K, Waldram A, Wilson DJ, Didelot X, Grad YH, Crook DW, Peto TE, Walker AS, Paul J, Eyre DW. 2016. Whole-genome sequencing to determine transmission of Neisseria gonorrhoeae: an observational study. Lancet Infect Dis 16:1295-1303.

FIGURE 1

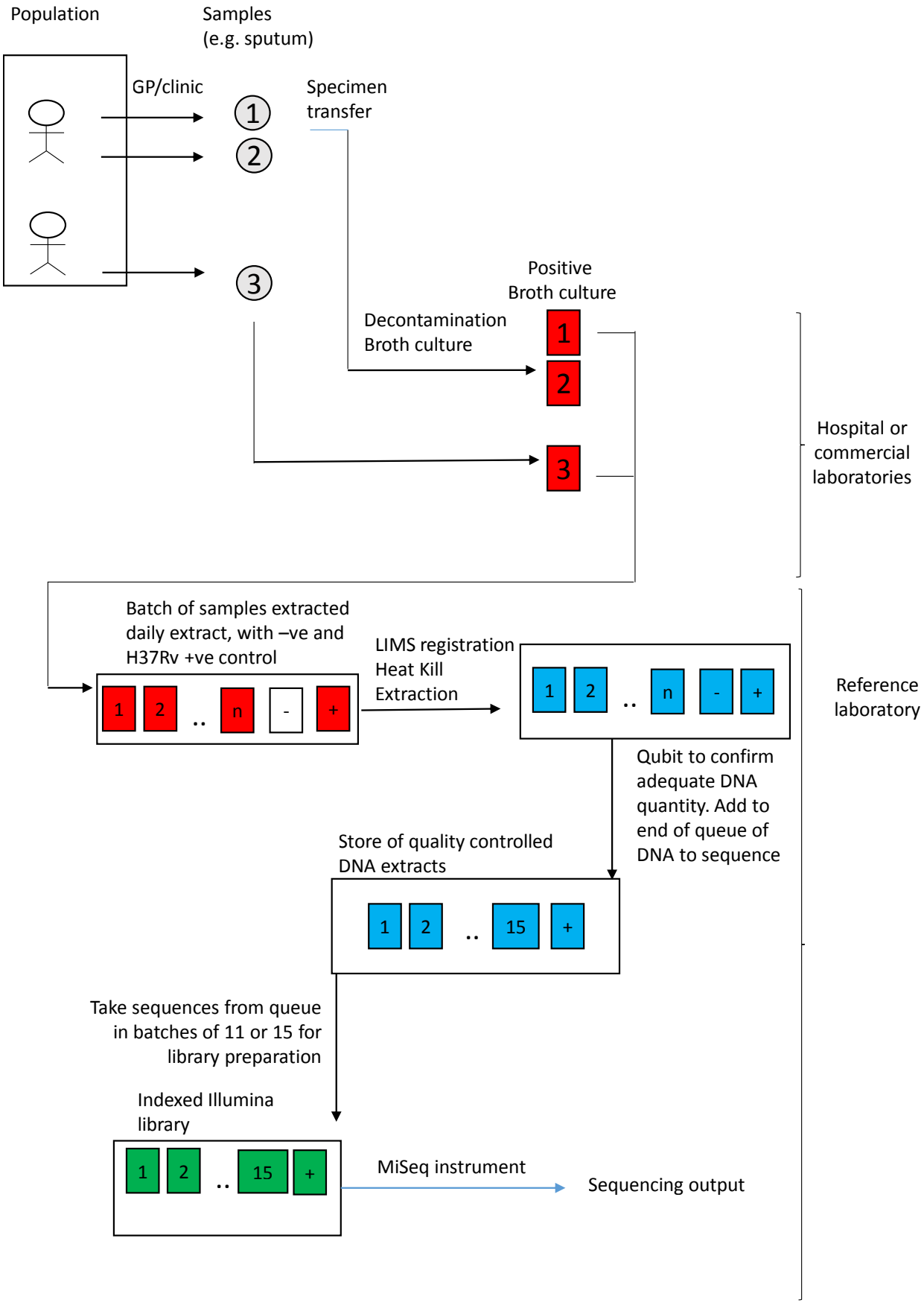

Population
Samples
(e.g. sputum)
GP/clinic
Specimen
transfer
1
2
3
Positive
Broth culture
Decontamination
Broth culture
1
2
3
Hospital or
commercial
laboratories
Batch of samples extracted
daily extract, with –ve and
H37Rv +ve control
1
2
..
n
-
+
LIMS registration
Heat Kill
Extraction
1
2
..
n
-
+
Reference
laboratory
Qubit to confirm
adequate DNA
quantity. Add to
end of queue of
DNA to sequence
Store of quality controlled
DNA extracts
1
2
..
15
+
Take sequences from queue
in batches of 11 or 15 for
library preparation
Indexed Illumina
library
1
2
..
15
+
MiSeq instrument
Sequencing output

FIGURE 2

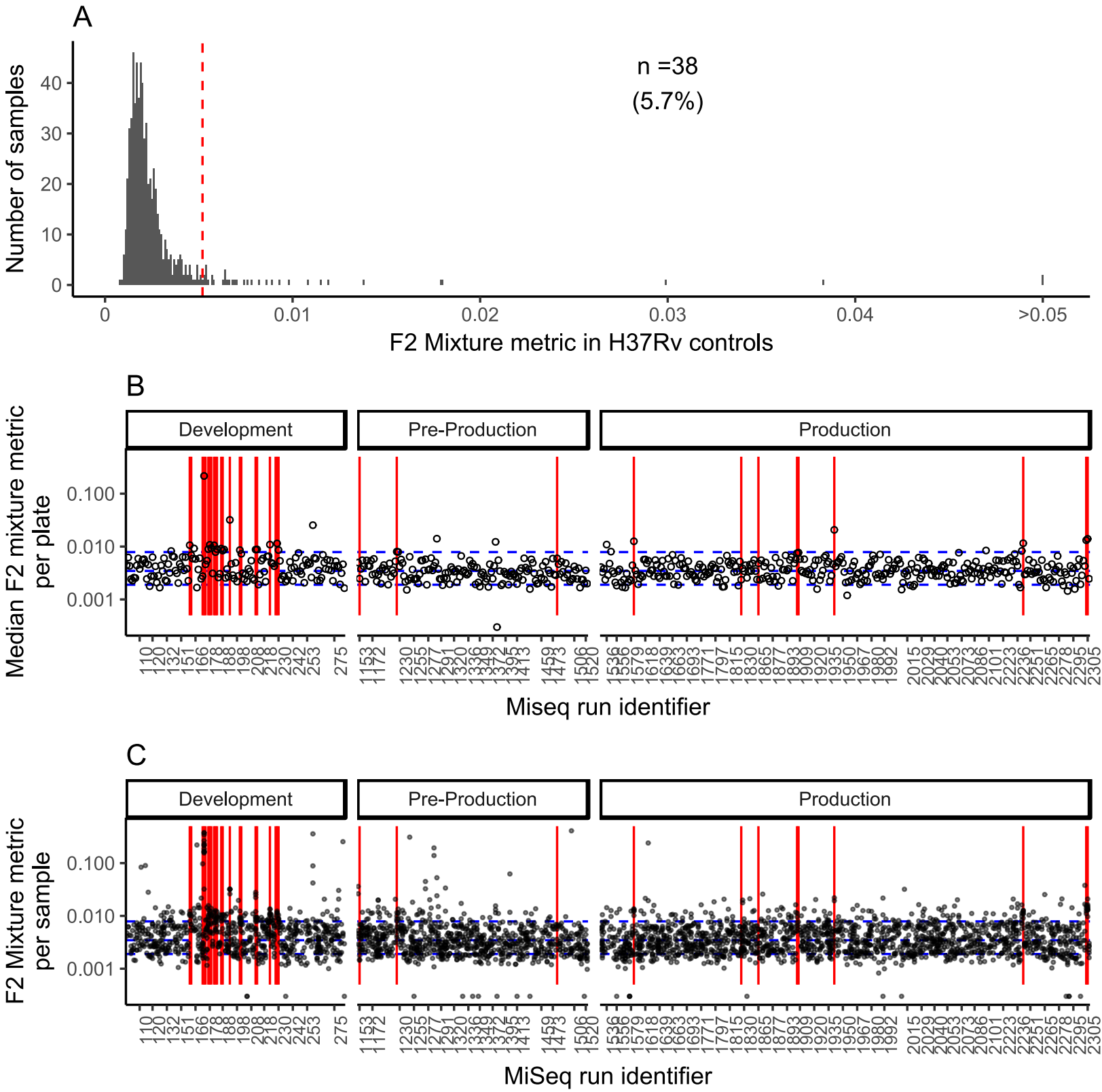

A
n =38
(5.7%)
Number of samples
F2 Mixture metric in H37Rv controls
B
Development
Pre-Production
Production
Median F2 mixture metric per plate
Miseq run identifier
C
F2 Mixture metric per sample
MiSeq run identifier

Figure 3

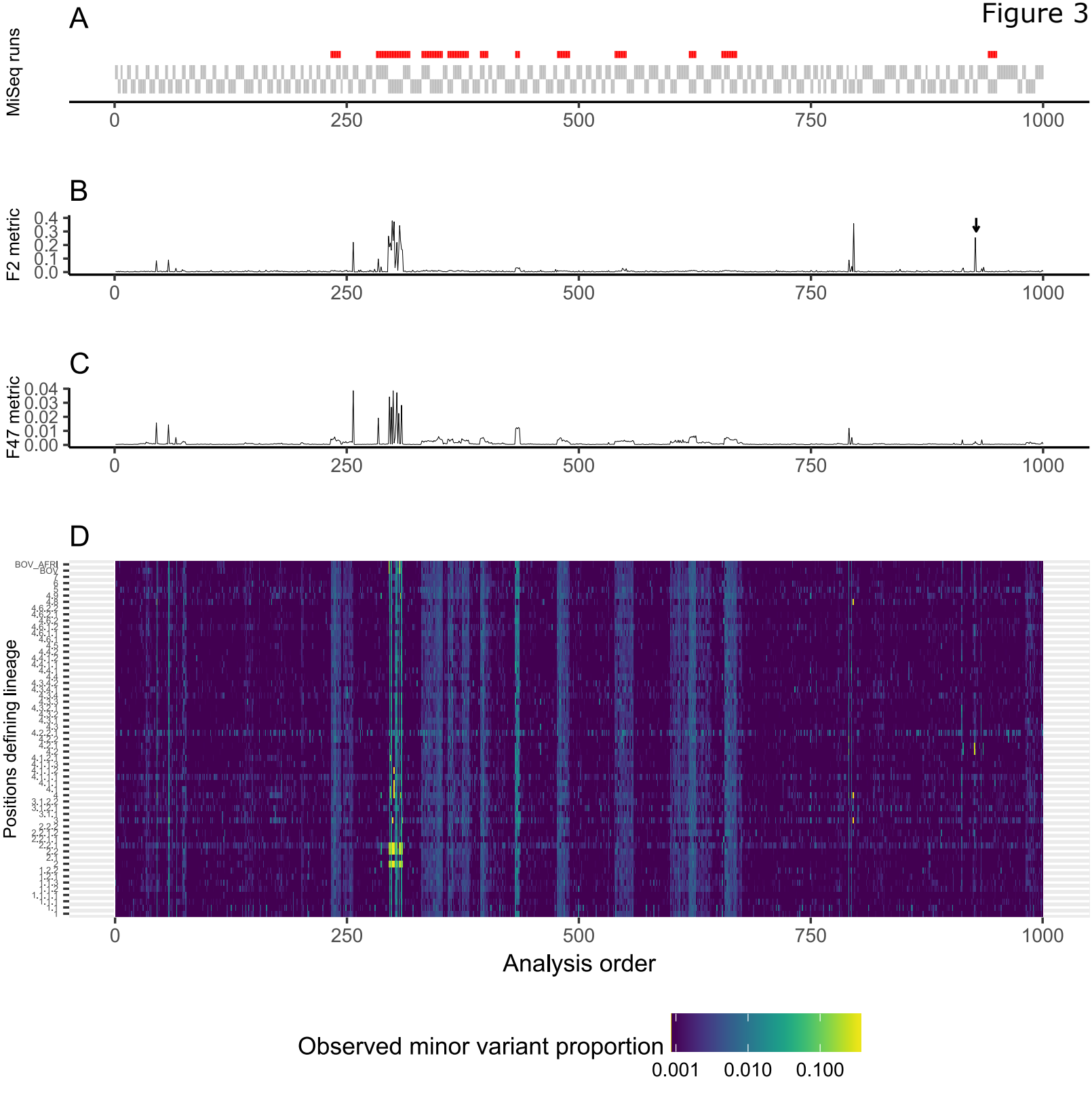

A
MiSeq runs
0
250
500
750
1000
B
F2 metric
0.4
0.3
0.2
0.1
0.0
C
F47 metric
0.04
0.03
0.02
0.01
0.00
D
Positions defining lineage
Analysis order
Observed minor variant proportion
0.001
0.010
0.100

Fig. 4

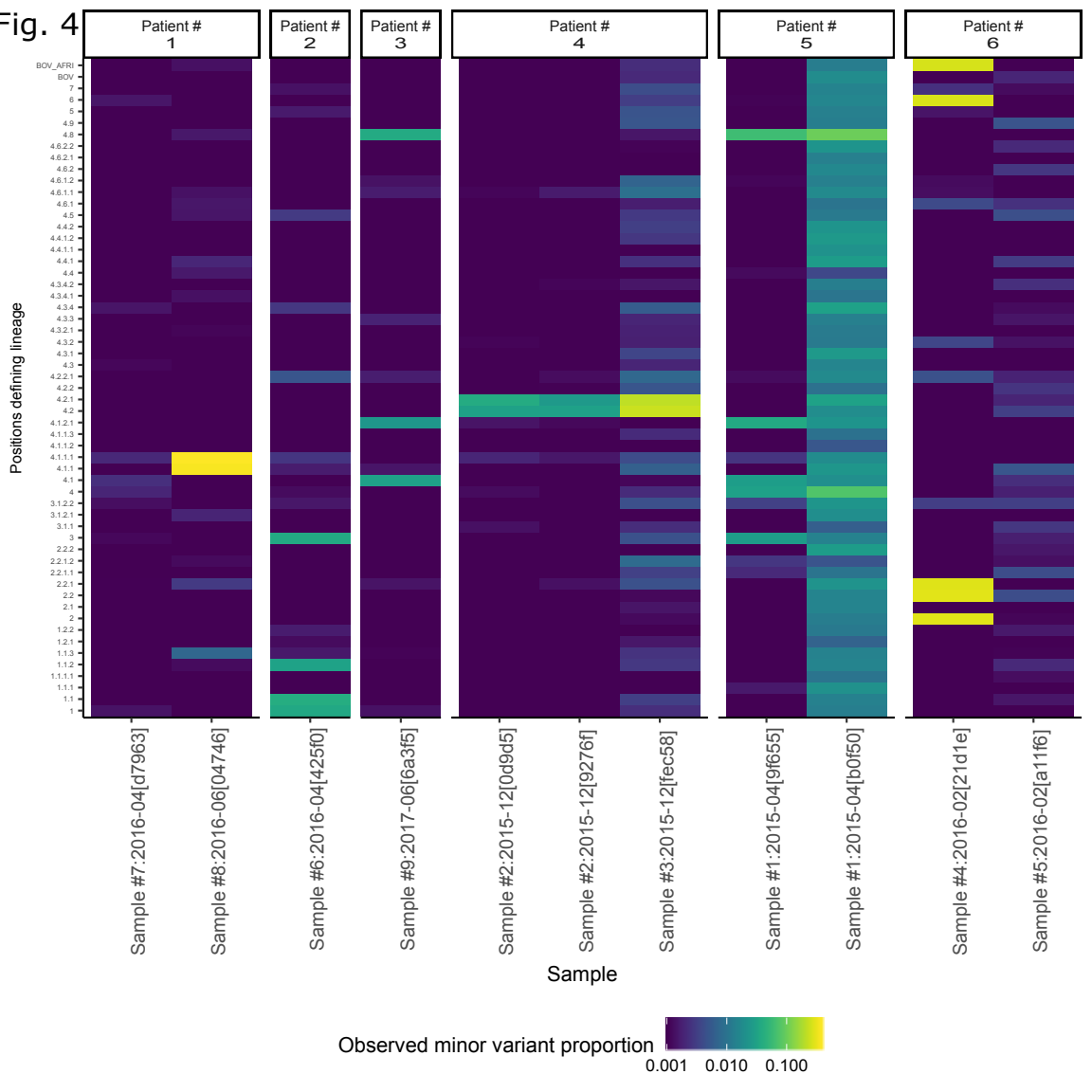
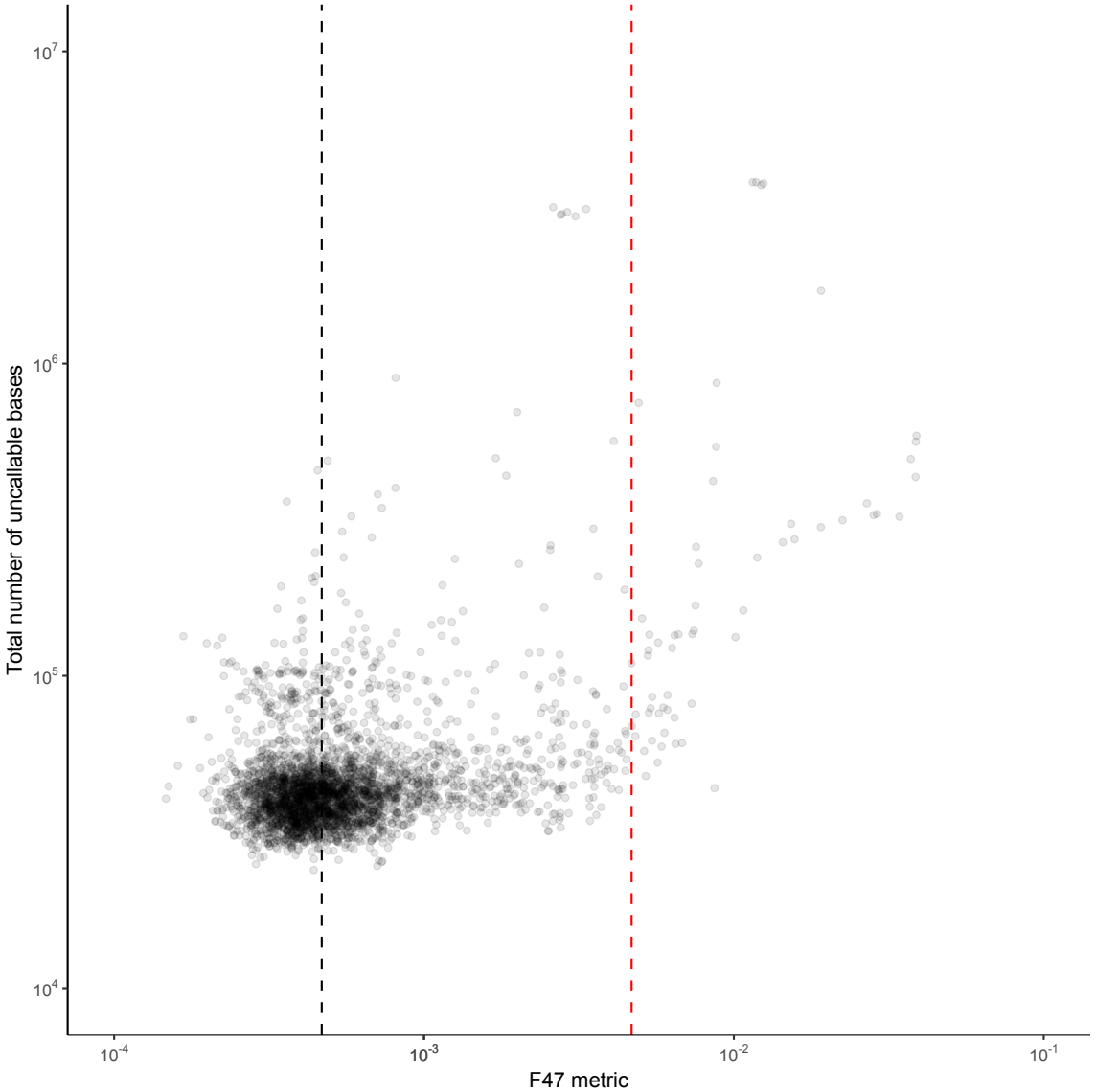

Patient # 1
Patient # 2
Patient # 3
Patient # 4
Patient # 5
Patient # 6
Positions defining lineage
BOV_AFRI
BOV
7
6
5
4.9
4.8
4.6.2.2
4.6.2.1
4.6.2
4.6.1.2
4.6.1.1
4.6.1
4.5
4.4.2
4.4.1.2
4.4.1.1
4.4.1
4.4
4.3.4.2
4.3.4.1
4.3.4
4.3.3
4.3.2.1
4.3.2
4.3.1
4.3
4.2.2.1
4.2.2
4.2.1
4.2
4.1.2.1
4.1.1.3
4.1.1.2
4.1.1.1
4.1.1
4.1
4
3.1.2.2
3.1.2.1
3.1.1
3
2.2.2
2.2.1.2
2.2.1.1
2.2.1
2.2
2.1
2
1.2.2
1.2.1
1.1.3
1.1.2
1.1.1.1
1.1.1
1.1
1
Sample #7:2016-04[d7963]
Sample #8:2016-06[04746]
Sample #6:2016-04[425f0]
Sample #9:2017-06[6a3f5]
Sample #2:2015-12[0d9d5]
Sample #2:2015-12[9276f]
Sample #3:2015-12[fec58]
Sample #1:2015-04[9f655]
Sample #1:2015-04[b0f50]
Sample #4:2016-02[21d1e]
Sample #5:2016-02[a11f6]
Sample
Observed minor variant proportion
0.001
0.010
0.100

Figure 5


Total number of uncallable bases
10^7
10^6
10^5
10^4
10^{-4}
10^{-3}
10^{-2}
10^{-1}
F47 metric

A

ACATACGTACGTACGTACGT
ACGTACGT**T**C**A**TACGTACGT

Sequence of
lineage 1
lineage 2

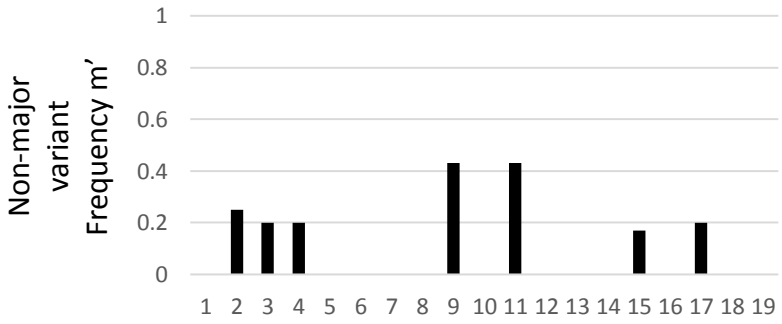
ACATACGTACGT
GTACGTACGTACGT
CGTACGTACTTACGT
GTACGTACGTACGT
ACGTACGT**T**C**A**TACGTACGT
AAGTACGT**T**C**A**TACGTCCG
CGGACGT**T**C**A**TACGTACGT

Reads from
lineage 1
lineage 2

T (bold) lineage defining variant

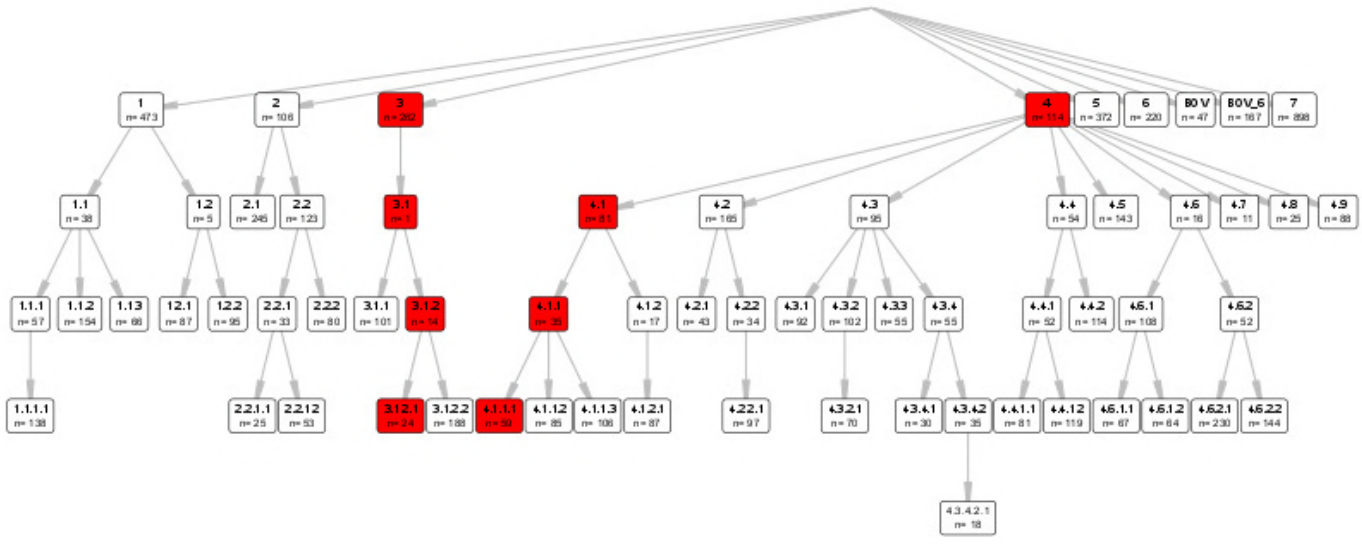
T (underlined) Variation due to error

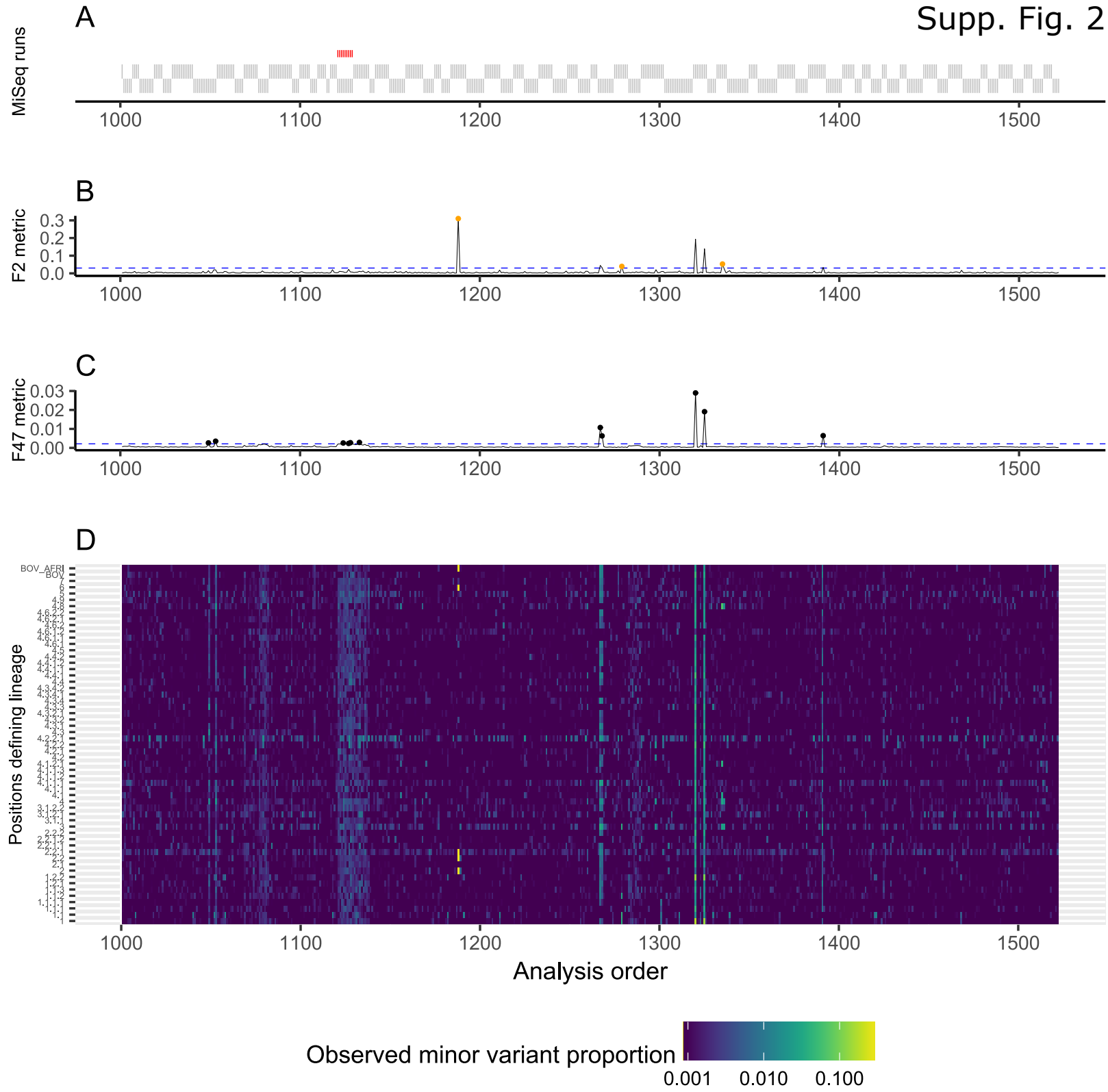
B

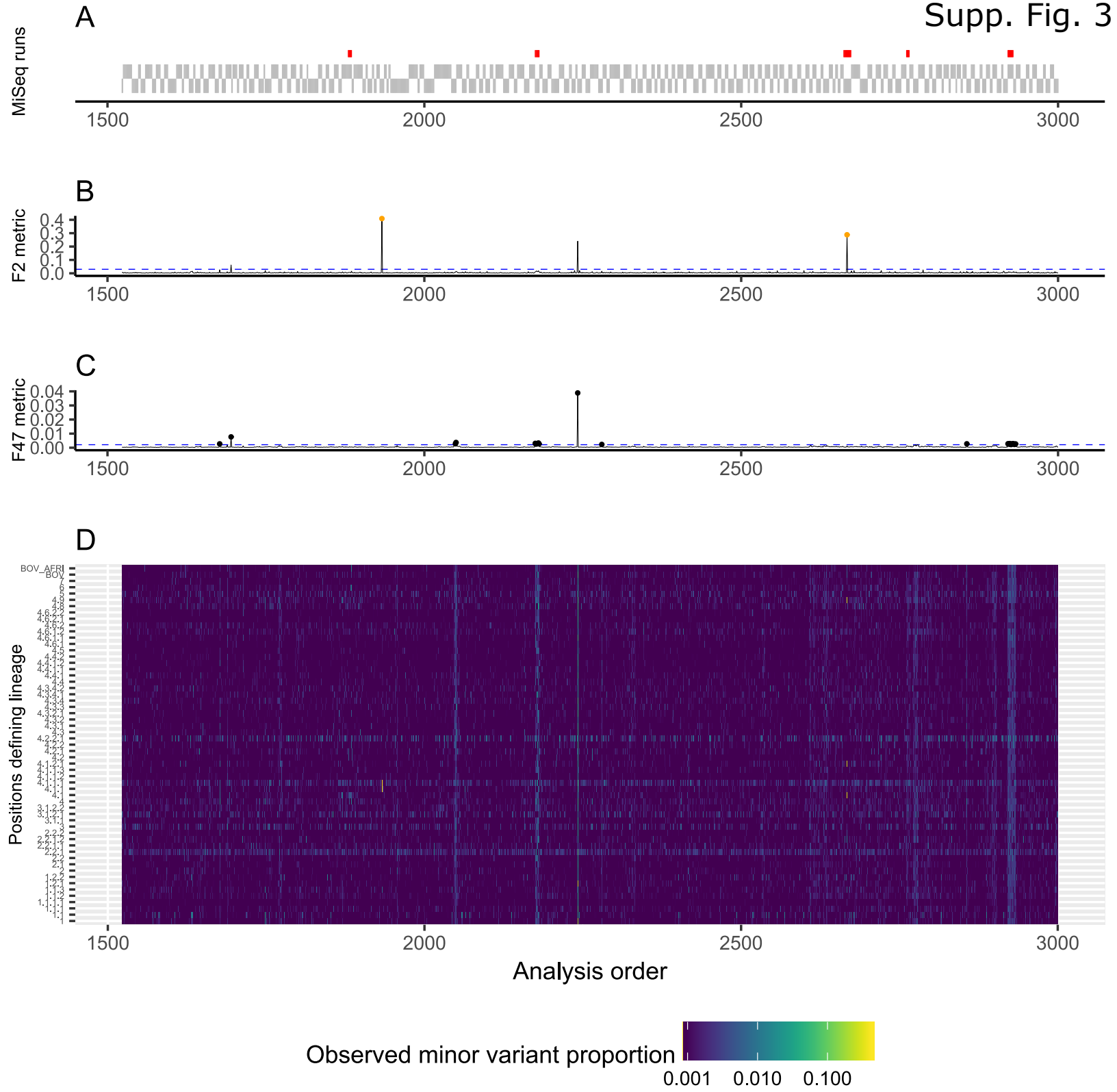


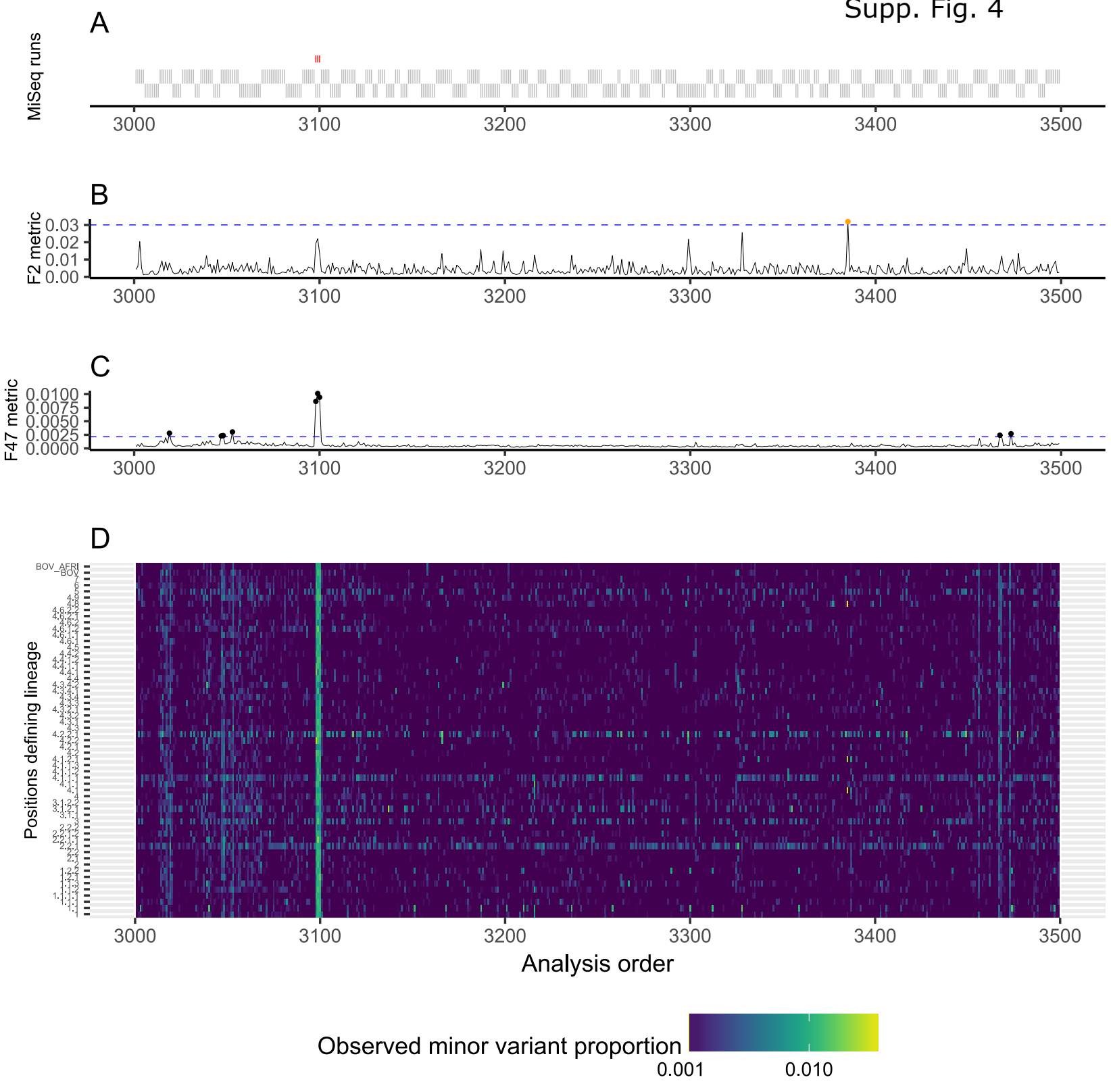
Non-major variant frequency m'

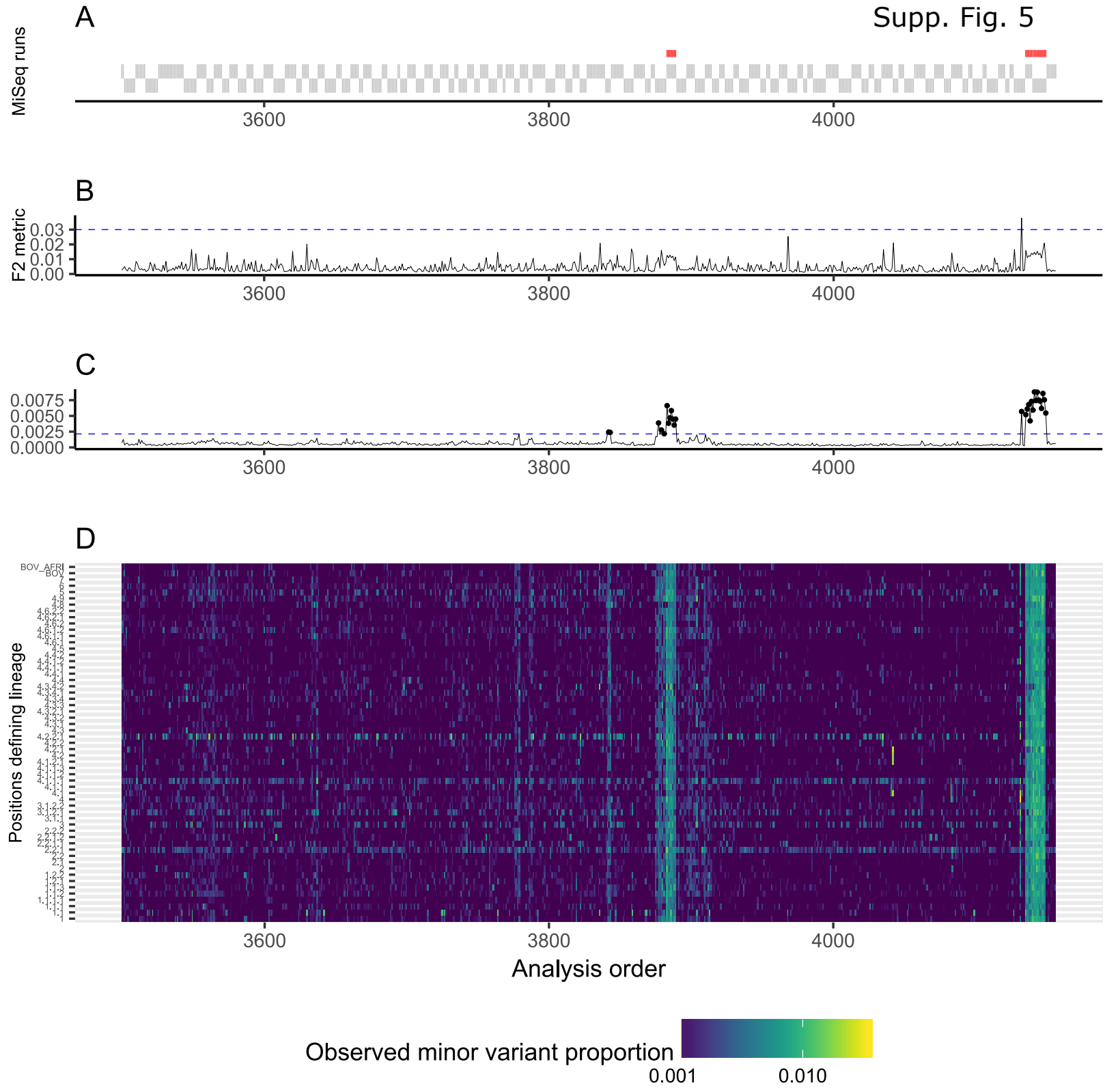
C




Supp. Fig. 2
A
MiSeq runs
1000
1100
1200
1300
1400
1500
B
F2 metric
0.3
0.2
0.1
0.0
C
F47 metric
0.03
0.02
0.01
0.00
D
Positions defining lineage
BOV_AFRI
BOV
Analysis order
Observed minor variant proportion
0.001
0.010
0.100


A
Supp. Fig. 3
MiSeq runs
1500
2000
2500
3000
B
F2 metric
0.4
0.3
0.2
0.1
0.0
C
F47 metric
0.04
0.03
0.02
0.01
0.00
D
Positions defining lineage
Analysis order
Observed minor variant proportion
0.001
0.010
0.100


A
MiSeq runs
3000
3100
3200
3300
3400
3500
B
F2 metric
0.03
0.02
0.01
0.00
C
F47 metric
0.0100
0.0075
0.0050
0.0025
0.0000
D
Positions defining lineage
Analysis order
Observed minor variant proportion
0.001
0.010


A
Supp. Fig. 5
MiSeq runs
3600
3800
4000
B
F2 metric
0.03
0.02
0.01
0.00
3600
3800
4000
C
0.0075
0.0050
0.0025
0.0000
3600
3800
4000
D
Positions defining lineage
BOV_AFR1
BOV
3600
3800
4000
Analysis order
Observed minor variant proportion
0.001
0.010