Supplementary Methods

An overview of the workflow for scheme creation, validation and threshold calibration is summarized in Additional file 2: Fig S3.

Creation of cgMLST scheme

Core loci were defined using the cgMLST target definer v1.5 implemented in the SeqSphere+ software (client v7.7.5). The finished genome of the *Mycobacterium abscessus* subsp. *abscessus* (Mab_A) ATCC19977 type strain was used as seed genome. For the penetration set, all genome assemblies up to chromosome or complete genome level were included (n=67) as well as 31 draft genomes (Additional file 2: Fig S1 and Additional file 2: Fig S2). An initial quality screen with Ridom SeqSphere+ (v7.7.5) software identified two outliers (NZ_CP065033.1 and NZ_CP022230.1) with more than 1,000 missing/bad cgMLST targets, which were therefore removed. The final penetration set thus comprised 96 genomes (Additional file 1: TableS1). All publicly available plasmids (n=17 on July 2nd, 2021; CU458745.1, NC_021278.1, NZ_AP022622.1, NZ_CP065264.1 NZ_CP065184.1, NZ_CP063329.1, NZ_CP065285.1, NZ_CP063325.1, NZ_CP063321.1, NZ_CP065281.1, NZ_CP065277.1, NZ_CP065270.1, NZ_CP065267.1, NZ_CP022234.1, NZ_CP050977.1, NZ_ATFQ01000044.1, NC_020994.1) were used to exclude plasmid-borne sequences from the scheme.

As first step of the scheme creation process, all protein-coding genes were extracted from the seed genome (n=4,920) and filtered according to the default SeqSphere+ Seed Genome Filters: (i) minimum Length Filter (requires >=50 bases), Start Codon Filter (requires start codon at beginning of the gene), Stop Codon Filter (requires single stop codon at end of gene), Homologous Gene Filter (requires no multiple copies of gene with BLAST overlap>=100bp, identity>=90.0%), Gene Overlap Filter (requires no overlap with other genes >4 bases) and Exclude Sequences Filter (requires no BLAST hit with overlap ≥ 100bp and identity ≥ 90.0% with sequences defined to be excluded (e.g. plasmids)). From the initial set of 4920 loci, 30 were discarded because multiple copies were found in the seed genome (n=28) or because

they were located on plasmids (n=2; exclude sequence filter) (Additional file 1: Table S3). Targets were further filtered using the 96 genomes from the penetration set according to the default SeqSphere+ Query Genome Filters: (i) Blast search (requires BLAST hit with overlap=100%, identity ≥ 90% in every query genome; BLAST v2.2.12 with options: word size=11, mismatch penalty=-1, Match reward=1, Gap open costs=5, Gap extension costs=2) and (ii) Stop Codon Percentage Filter (requires single stop codon at end of gene in >80% penetration genomes). In total, 1,986 loci were discarded because they overlapped in the seed genome (n=338), were not found in each of the query genomes (n=1,646) or because more than 20% of the query genomes had an invalid/missing stop codon for this locus (n=2). The remaining hard defined core genome that makes up the final cgMLST scheme thus consisted of 2,904 loci (Additional file 1: Table S3).

Technical validation of cgMLST scheme

To assess the robustness and reproducibility of the scheme, we compared cgMLST results obtained for genomes generated from the same isolate with different assembly approaches. This technical validation was performed using a diverse set of 30 Mab strains (10 strains of each Mab subspecies) for which both publicly available WGS read sets (Illumina fastQ files) and assemblies (fastA files) were available (Additional file 1: Table S1 and Additional file 2: Figure S1). Assemblies from the short read sets were made using two popular *de novo* (SPAdes and skesa) assemblers and one mapping assembler (BWA-MEM), using different assembly pipelines (SeqSphere$^+$ and shovill) and using different read preprocessing steps (default trimming or prior read error correction turned off or on) (see Methods and Additional file 1: Table S4). Next to the average percentage of “good” cgMLST targets (i.e. those for which an allele number was assigned), we also recorded the processing time to get an idea about the total time required for cgMLST analysis (Additional file 1: Table S4).

For each assembly (method), more than 98% good cgMLST targets were identified. However, the number of good targets was significantly lower for the mapping approach compared to *de novo* assembly approaches (Additional file 1: Table S4) ($p<0.05$). Total processing time per

sample ranged between 4.4 and 18.4 minutes, although we have to note that this is dependent on your hardware configuration (Additional file 1: Table S4). The actual time to determine the cgMLST profile from the assembly in SeqSphere+ (i.e. searching for all 2,904 cgMLST loci and translation of the sequences into allele numbers) was on average 4.5 minutes per sample (Additional file 1: Table S4). Once cgMLST profiles were calculated and stored in the SeqSphere+ database, the time required to compare cgMLST profiles (i.e. calculate distances) or create minimum spanning trees was neglectable (<1 min for 210 profiles).

We calculated the cgMLST distance as the number of cgMLST loci with allele differences between the different assemblies, not considering missing alleles in the pairwise comparisons. If different assemblies of the same isolate were analyzed, there were no cgMLST loci with a different allele number (cgMLST distance = 0) between the assemblies generated with Shovill and SeqSphere^{+} (Additional file 2: Fig S4).

For 16 out of 30 strains, one up to six allele differences were found between publicly available assemblies (i.e. those downloaded from the public repository RefSeq) and those generated as part of this study in Shovill or SeqSphere^{+} (Additional file 2: Fig S4). For 25 strains, 1 up to 19 cgMLST targets with different allele numbers were found when the mapping approach was compared with de novo assembly approaches (Additional file 2: Fig S4).

Supplementary Figures

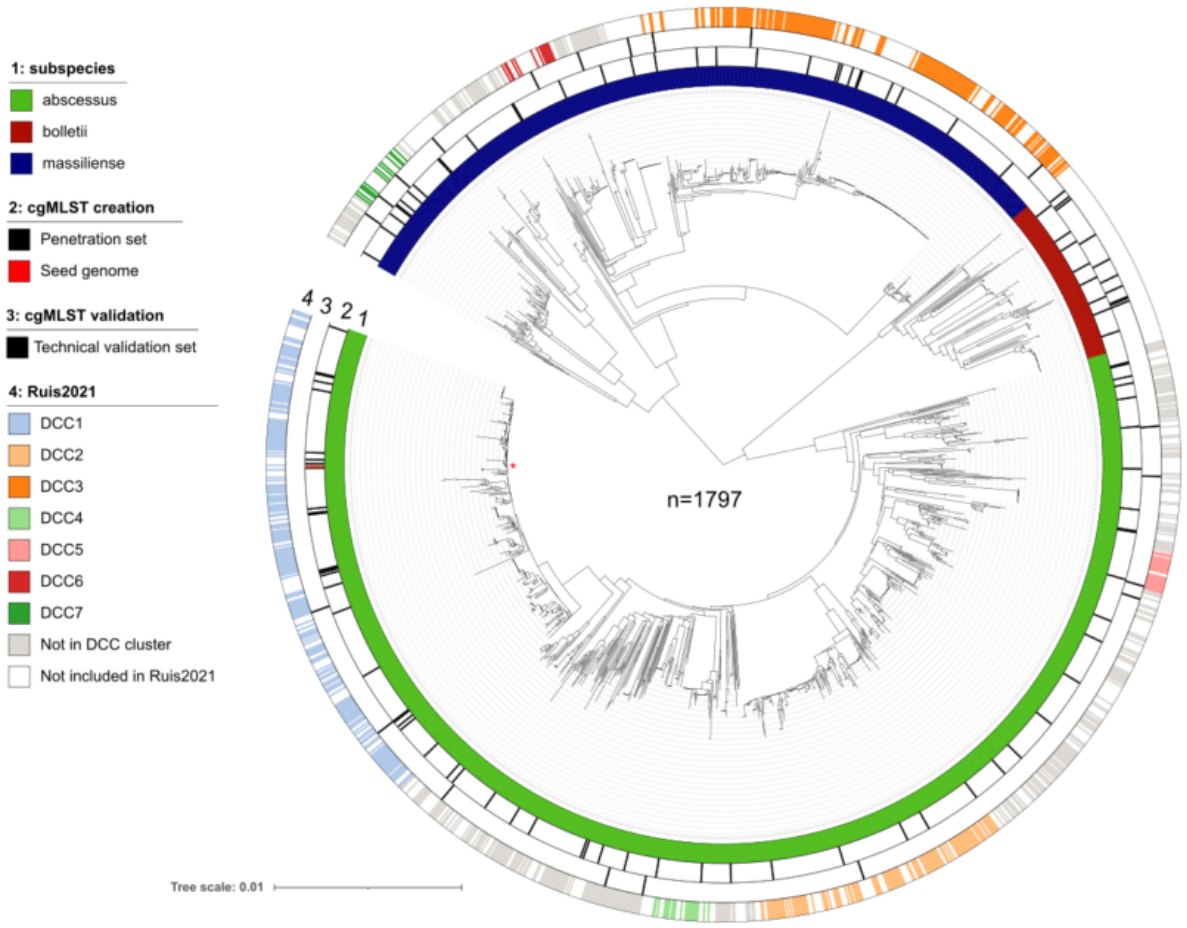

Fig S1*: *Mash distance-based neighbor-joining tree (NJ) of 1,797 Mycobacterium abscessus isolates (evaluation set). 1. Subspecies classification based on position in NJ tree. 2. Genomes used for cgMLST scheme creation: penetration set (n=96) and seed genome. 3. Genomes used for technical validation of the cgMLST scheme (n=30). 4. DCC classification as reported by Ruis and coworkers[1]. NJ tree was created with Mashtree [2] and visualized using iTOL[3].

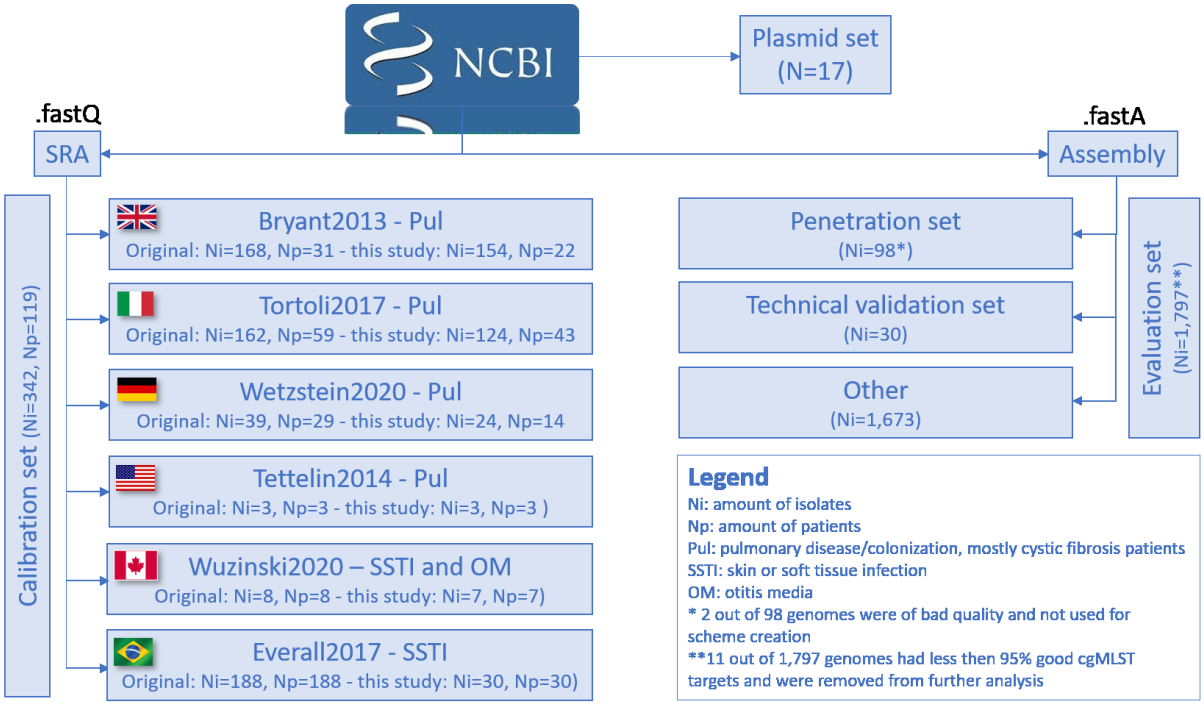

Fig S2: Overview of datasets used for cgMLST scheme creation, validation and threshold calibration.

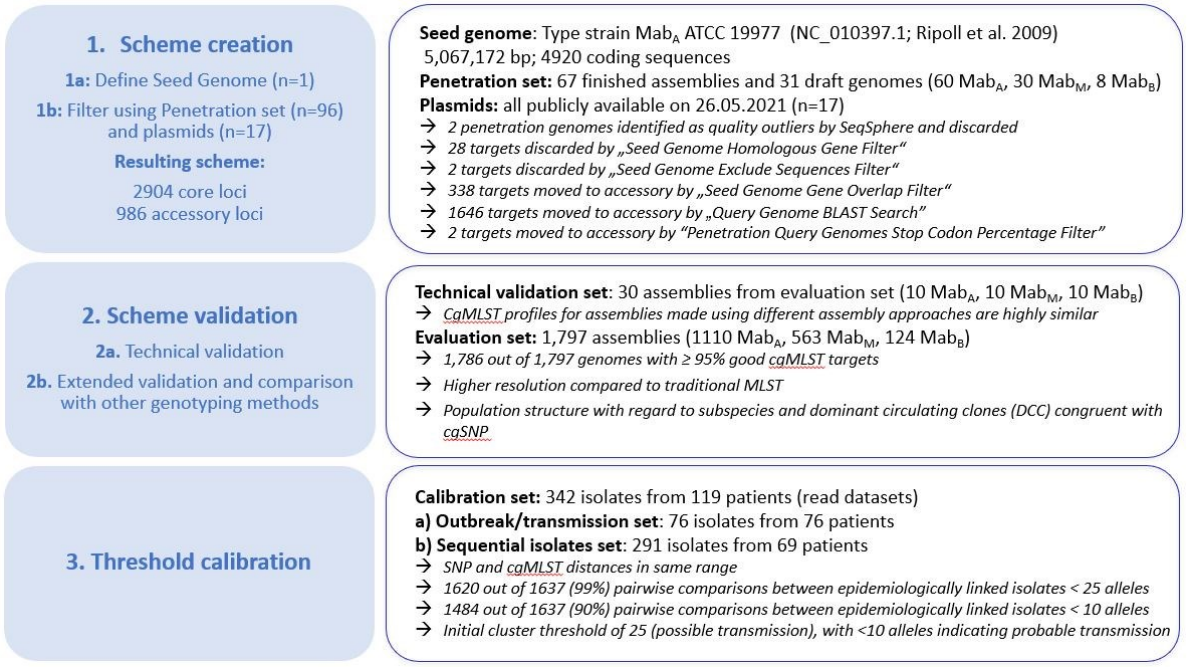

Fig S3: Workflow used for cgMLST scheme creation, validation and threshold calibration.

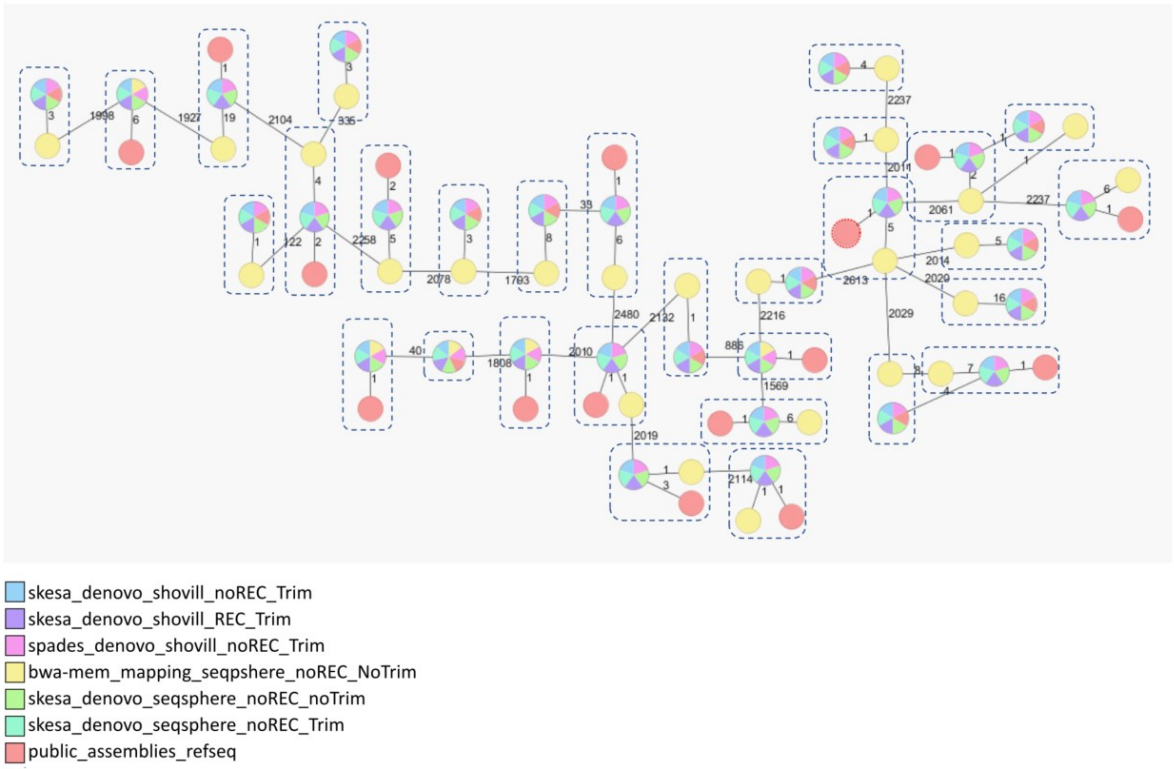

Fig S4: CgMLST-based minimum spanning tree of 30 isolates included in the technical validation set. For each isolate, cgMLST analysis was performed on seven assemblies generated using different assemblers and/or preprocessing steps (Details in Methods and Additional file 2: Table S4). REC: read error correction. The number on the branches represents the amount of cgMLST allele differences (distance) between assemblies. Assemblies belonging to the same isolate are grouped with dashed lines. Each node represents an assembly from a particular isolate. Assemblies without any allelic differences are collapsed into one node. For visualization purposes, the branch length is not proportional to the distance. The minimum spanning tree was created in SeqSphere^{+} and modified using inkscape.

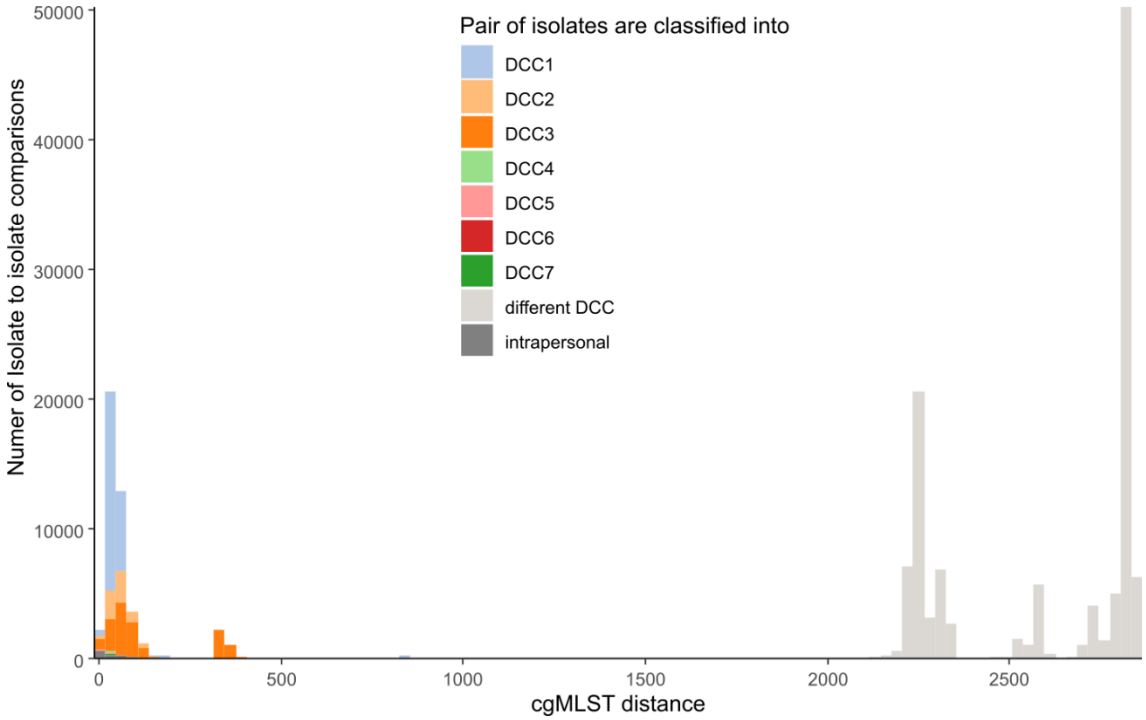

Fig S5: Histogram including all pairwise comparisons between 571 isolates previously classified into one of seven dominant circulating clone (DCC) complexes (Ruis et. al, 2021).

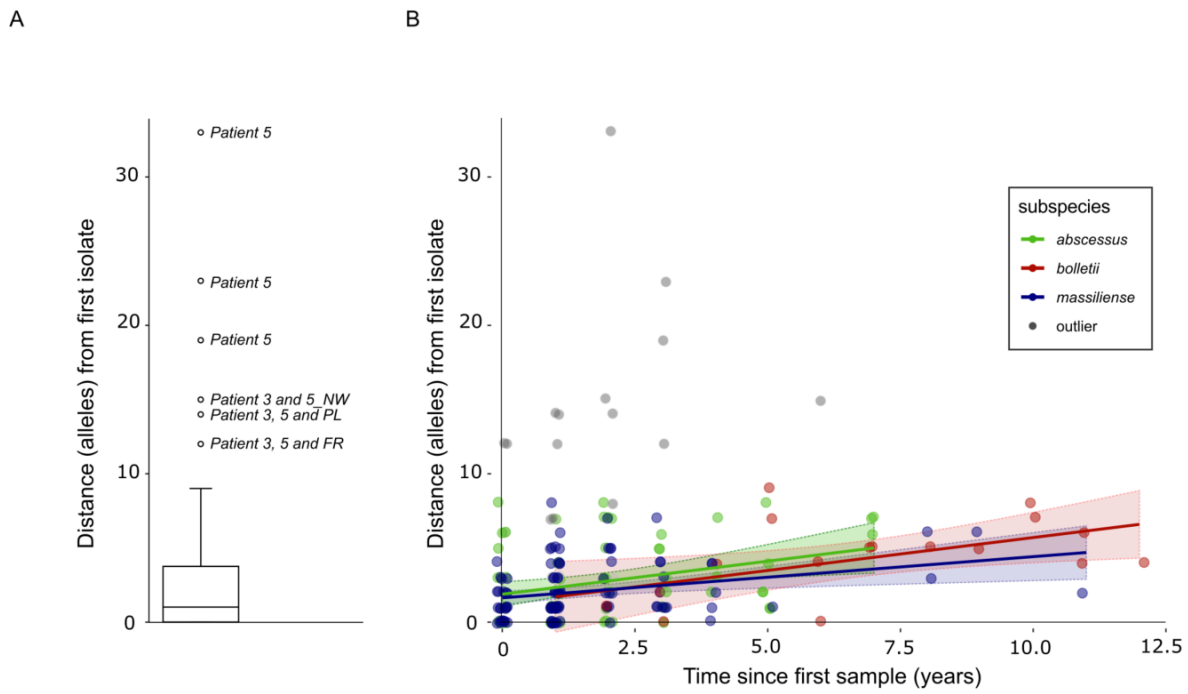

Fig S6: Within-patient diversity. Pairwise distances are calculated as the amount of allelic differences (cgMLST analysis) between an isolate (collected at x years after the first available isolate) and the first available isolate. Data is plotted for 43 patients with sequential *Mycobacterium abscessus* subsp. *abscessus* (MabA) isolates, 18 patients with sequential subsp. *massiliense* (Mab_M) *isolates* (n=18), and 8 patients with sequential Mab subsp. *bolletii* (MabB) isolates included in three studies[4–6]. A. Boxplot to identify outliers with high genetic distance compared to first isolate. These outliers might be the result of mixed infections with closely related isolates or hypermutator strains. B. Correlation of genetic divergence with the duration of infection (scatter plot with jitter function). Linear regressions were fitted by the least-squares method and did not include outliers (i.e. patient 3, 5, PL, FR and 5_NW). The slope of the linear fit was used to estimate evolutionary rate. The estimated evolutionary rate for Mab (including all subspecies but excluding outliers) was 0.38 alleles/genomes/year (95% CI 0.26 – 0.50, t Stat = 6.2, df = 189, p = 2.96E-09, R^2 = 0.170; regression line not shown) with 0.45 (95% CI 0.13-0.76, t Stat = 2.8, df = 68, p = 6.32E-03, R^2 = 0.105) for Mab_A, 0.28 alleles/genome/year (95% CI 0.09-0.46, t Stat = 2.9, df = 101, p = 3.99E-03 R^2 = 0.079) for Mab_M, and 0.44 (95% CI 0.08-0.81, t Stat = 2.6, df = 16, p = 1.99E-02 and R^2 = 0.294) for Mab_B.