Auditing bacterial dark-gene screens for superimposed open reading frame artefacts: a multi-layer analysis of Rv2438A in *Mycobacterium tuberculosis*

Christophe Guyeux[a,*]

[a]Université Marie et Louis Pasteur, CNRS, FEMTO-ST institute, F-90000 Belfort, France

*Corresponding author: christophe.guyeux@univ-fcomte.fr

Abstract

Essentiality and knockdown-vulnerability screens can promote spurious bacterial open reading frames when those frames overlap essential genes, because such a frame inherits its neighbour's signals undiluted and therefore satisfies the screen's criteria better than a genuine small gene. We present a multi-layer audit that tests this failure mode across genome annotation, transposon mutagenesis, CRISPR interference, homology, transcript mapping, proteomics, and population variation. We apply it to Rv2438A, a 92-codon conserved hypothetical open reading frame of *Mycobacterium tuberculosis* ranked first by our own dark-gene target screen. Rv2438A is superimposed on the essential NAD synthetase locus *nadE*: 44% lies within its coding sequence on the opposite strand, and the remainder covers its promoter and transcription start site. Consequently, three of five Himar1 sites lie within *nadE*, no CRISPRi guide can target Rv2438A without binding *nadE*, and the cross-species hit maps to the same *nadE* start junction. Rv2438A lacks its own transcription start site and is absent from every proteomic dataset that detects *nadE*. A genome-wide scan identifies six short, overlapping, uncharacterised loci among 3,907 annotated genes, but only Rv2438A combines overlap and essentiality with non-detection across all proteomic datasets; rare genome-wide, it ranked first among screen hits. We provide an implementable audit workflow and a codon-position control, but measure the control's sensitivity as only two of five genes with attested protein, limiting it to confirmatory use. Overlap coordinates and neighbour-specific experimental resolution should therefore be reported before bacterial dark genes are prioritised.

Keywords: *Mycobacterium tuberculosis*, genome annotation, spurious open reading frames, overlapping genes, gene essentiality, transposon sequencing, CRISPR interference

1 Introduction

Roughly one quarter of the *Mycobacterium tuberculosis* proteome remains functionally uncharacterised, an obvious reservoir of drug targets: a gene that escaped characterisation has, by construction, escaped the medicinal chemistry that followed. The standard way to mine it combines two genome-wide resources: saturating transposon mutagenesis partitions genes into essential and non-essential [1], and CRISPRi knockdown replaces that binary call with a continuous vulnerability index, identifying genes whose *partial* inhibition already imposes a large

fitness cost [2]. Ranking dark genes by the conjunction of the two, preferring small proteins for tractability, is a reasonable, widely used heuristic.

We use "dark gene" throughout for a coding sequence carrying no experimentally supported function — annotated "hypothetical" or "conserved hypothetical", without characterised domain, biochemical activity or phenotype. The term is borrowed from the "dark proteome", which names the same set at the protein level and is current in bacterial-pathogen genomics [3]; it is not, we grant, a standardised term of bacterial genetics, and we adopt it only because the alternatives name the annotation rather than the object: "hypothetical protein" presupposes the protein whose existence is precisely what is at issue here.

This work reports that the heuristic has a failure mode whose mechanism is structural rather than accidental, and that, in the one instance we can follow end to end, it outranked the true positives. We are careful about what that claim is. Transposon and CRISPRi screens are discovery tools: they nominate candidates, and a candidate that overlaps another gene ordinarily prompts locus-specific genetic validation before anything is concluded. Nothing here disputes that. What we show is that a particular class of candidate — a putative reading frame superimposed on an essential gene — passes every filter such a screen applies, does so *better* than a genuine small gene, and therefore arrives at the top of the ranking rather than in its body, where the cost of the eventual validation is highest. The failure is a demonstrated interpretive trap, established on one locus followed from promotion to refutation; it is not a documented distortion of published target prioritisation, and Section 2.1 states what is and is not known on that point.

That spurious open reading frames pervade bacterial annotation is well known, and antiparallel overlapping ORFs specifically have been examined directly. Analysing 929 prokaryotic genomes, Moshensky and Alexeevski [4] showed that reading frames opposite annotated genes are far more frequent than chance predicts, that their translations are generally not under negative selection, and concluded that essential genes are unlikely to be encoded by such frames. Their demonstrative case is sobering: 282 antiparallel ORFs opposite the highly conserved *dnaK* gene had been annotated as glutamate dehydrogenases and admitted to Pfam as a third family of that enzyme. Separately, Choe et al. [5] dissected mechanistic causes of false-positive and false-negative essentiality calls in *Escherichia coli* transposon data, attributing many false positives to nucleoid-associated protein occupancy rather than to biological essentiality.

We take that conclusion as our starting point rather than our result, and ask a different question: what happens to such a locus inside a modern target-discovery pipeline? Our contribution is threefold. First, the previous analyses rest on comparative sequence evidence, chiefly K_a/K_s and homolog distributions; we show the experimental layers target screens actually rely on — transposon essentiality, CRISPRi vulnerability, proteomic detection, transcription-start mapping — are contaminated too, and identify the mechanism in each case. Second, we show the artefact is not merely tolerated by such screens but can be actively promoted by them: a spurious ORF overlapping an essential gene satisfies the dark, small, essential and vulnerable criteria better than a genuine small gene does, inheriting its neighbour's signals undiluted. This is a statement about the selection criteria, which can be checked on their definition, rather than an estimate of how often the situation arises in practice. Third, the overlap threshold used previously was 180 bp; our locus overlaps by 123 bp and would not have been captured, so the concern extends

to shorter overlaps than examined.

We develop the case on Rv2438A, a 92-codon "conserved hypothetical" ORF that our own pipeline ranked as its best candidate: essential by transposon sequencing [1], robustly vulnerable by CRISPRi [2], conserved across 145,209 isolates, domainless but with a divergent cross-species hit in *Mycobacterium decipiens* [6, 7], and immediately next to *nadE*, the glutamine-dependent NAD synthetase, itself a validated antitubercular target [8–11]. None of this requires Rv2438A to encode a protein: all of it follows from a single architectural fact — the ORF is superimposed on the *nadE* locus, half inside its coding sequence, half over its promoter and transcription start site. We establish this with five independent lines of evidence, three formulated as quantitative predictions before the corresponding data were examined and then verified, plus one weaker indication whose limits we quantify.

In practice, this yields a codon-position control whose measured sensitivity (2 of 5 on genes with attested protein) confines it to a confirmatory role rather than a general diagnostic, a genome-wide screen showing the artefact to be rare, six candidate loci out of 3,907 and therefore easy to exclude once one thinks to look, and a screen-design recommendation: because the artefact is rare in the genome but concentrated at the top of the ranking, its prevalence among screen *hits* is far higher than among *genes*, so target pipelines should report overlap fraction as a first-class annotation for any short gene.

2 Materials and Methods

2.1 Origin of the candidate and timing of the audit

Because the argument of this paper concerns what a screen promotes, the provenance of the ranking matters as much as the audit itself, and we state it in full.

The candidate came from a target screen run over the re-annotated *Mycobacterium tuberculosis* gene set of our own atlas, in which each of the 3,907 H37Rv coding sequences matched to a curated record carries a functional verdict. Genes were retained when they were uncharacterised, short enough to be tractable, called essential by transposon sequencing [1], and robustly vulnerable by CRISPRi [2] — robustly meaning that the confidence interval of the vulnerability index excludes zero, which for Rv2438A it did (-4.56, CI $[-7.2, -1.9]$). A fifth criterion applied at that time, absence of a detectable homologue outside the complex, was withdrawn the same day when a relaxed `tblastn` recovered the divergent *M. decipiens* hit; that hit was itself shown to be an artefact later, and the sequence of these two corrections is reported in the Results.

The ranking was produced on 5 July 2026 and Rv2438A came first: among the genes retained by these four criteria, it was judged the strongest individual candidate by comparative review against the same criteria, not selected by a single numeric score. The rule that makes this first place independently verifiable is stated in full, with explicit thresholds, in Section 3.8: of 3,907 annotated genes, Rv2438A is the only one that combines all five boolean criteria defined there, a result stable over 49 threshold combinations (Supplementary Text S2). The audit reported here was run between 5 and 31 July 2026, that is *after* the ranking and on the strength of it: the project was opened to characterise the candidate as a drug target, not to refute it, and the overlap with *nadE* was found in the course of that characterisation. No screen output was

recomputed afterwards to produce the ranking retrospectively, and the audit did not select its case from among several. This ordering is what makes the case informative and also what limits it: we report one locus, followed prospectively from promotion to refutation, not a survey of screens.

We are equally explicit about the reach of that promotion. Rv2438A has never been proposed or pursued as an antitubercular target in the published literature: as of 8 September 2026 the locus tag returns no hit in PubMed titles or abstracts, and a single hit in the body of the open-access full-text literature, where it appears as one line of a supplementary table of "conserved hypotheticals" [12]. The failure documented here is therefore a demonstrated interpretive trap, reproduced end to end on a real pipeline, and not an established distortion of published target prioritisation. What generalises is the mechanism, not the incident.

2.2 The audit workflow

The audit is a fixed sequence of tests, each asking of one experimental layer whether its signal could have been produced by a neighbouring gene rather than by the candidate (Figure 1). It opens with a single gate — does the candidate overlap an annotated neighbour, and over what fraction of its own length — because a candidate that overlaps nothing needs no audit, and because every subsequent test is a test of that overlap's consequences. The five layers that follow are mutually independent: transposon insertion sites, CRISPRi guide placement, the coordinates of a cross-species hit, transcription-start mapping and proteomic detection each fail in a different way, and no two of them share a data source. A sixth, the codon-position test on population variation, is confirmatory only, for the reason quantified in the Recommendations. The remainder of this section specifies each layer; the Results apply them to Rv2438A in the same order.

2.3 Genome coordinates and overlap

Coordinates were taken from the H37Rv reference annotation (NC_000962.3, GFF3). Overlap between a gene and its immediate neighbours was the length of the intersection of their coordinate intervals, as a fraction of the shorter gene. Reading frames were verified by direct translation of the reference sequence (genetic code 11), checking for absence of internal stop codons and presence of a terminal one.

Himar1 inserts at TA dinucleotides, so the informative resolution of a transposon essentiality call is the number of TA sites a gene contains. TA sites were enumerated on the reference sequence and split into those inside the overlap with the neighbouring gene, where an insertion is lethal through the neighbour, and those specific to the gene under test.

2.4 Population variation and the codon-position test

Variant data for this test were queried directly from the laboratory TBannotator database (274,680 MTBC genomes), which stores variants in SPDI notation. The 145,209-isolate figure used above for the conservation statistic instead comes from a locally mirrored, independently curated subset of the same collection, so the two counts are not directly comparable. *SPDI coordinates are zero-based whereas GFF3 coordinates are one-based*; the conversion is applied before

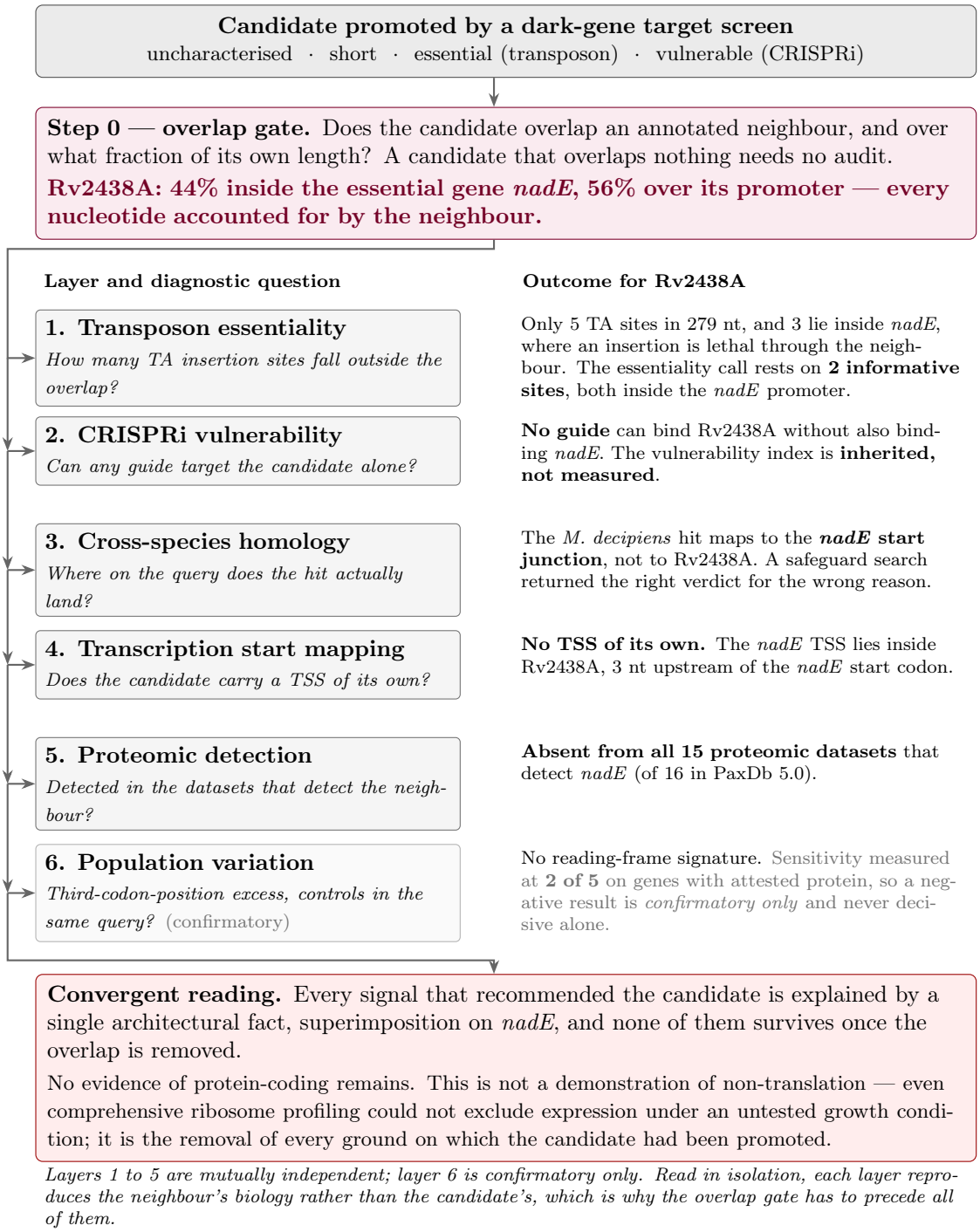


Figure 1: The audit workflow, and its application to Rv2438A. A candidate promoted by a dark-gene target screen enters at the top. Step 0 asks the only question that conditions the rest: does the candidate overlap an annotated neighbour? Layers 1 to 5 then test, independently of one another, whether each experimental signal that supported the candidate can be attributed to that neighbour; layer 6 is a confirmatory control whose measured sensitivity forbids reading a negative result as evidence of non-coding. The right-hand column reports the outcome obtained for Rv2438A at each layer. The workflow is not specific to *Mycobacterium tuberculosis*: it requires only a compact genome, gene prediction permissive enough to call short reading frames on both strands, and a screen that rewards essentiality in short uncharacterised genes.

any frame arithmetic. Codon position was computed as $((p + 1 - \text{start}) \bmod 3) + 1$ on the plus strand and $((\text{end} - (p + 1)) \bmod 3) + 1$ on the minus strand, where p is the SPDI position.

A translated coding sequence under purifying selection accumulates substitutions preferentially at the third codon position, largely synonymous there. The principle is not new: three-base periodicity and third-position composition bias are classical coding statistics microbial gene finders have used for decades [13, 14]. What we use here is not a new statistic but its application in the opposite direction — as an *a posteriori* control on a candidate a screen has already promoted, not as a predictor. The test compares the observed distribution across the three positions to the uniform expectation by a χ^2 test with two degrees of freedom. The test is only interpretable with positive controls: neighbouring, unambiguously coding genes must be included in the same query, and in our hands they also caught an off-by-one frame error that would otherwise have inverted the conclusion (Recommendations, below). The expected third-position fraction depends on the frequency filter applied upstream, since rare variants are largely sequencing noise diluting the signal: genuine coding sequences give approximately 0.53 over all variants and 0.75 over variants carried by at least twenty isolates, against 0.33 for an unread sequence under either filter. These are properties of the catalogue, not universal constants (Supplementary Text S1), so expected rates are always taken from controls run in the same query. The implementation is a reusable module (Data availability).

2.5 Homology and its localisation

Cross-species homology was assessed by `tblastn` of the query protein against assembled non-tuberculous mycobacterial genomes at an E-value threshold of 10^{-3}. Alignment coordinates were retained — the point on which our earlier pipeline had failed: a low-coverage hit cannot be interpreted without knowing *which* part of the query it covers. Where a hit was suspected to reflect a neighbouring gene, that gene's protein was blasted against the same target and the two sets of high-scoring pairs compared.

2.6 CRISPRi guide mapping

The individual guide sequences underlying the published vulnerability index are not part of the summary statistic and were obtained separately, from the library deposited by the original authors [2] rather than re-derived. Each guide assigned to a locus was located on the reference genome by exact string matching on both strands, its position assigned to the annotated gene whose coordinates contain it. A guide with more than one genomic match, or a match outside its assigned gene, is flagged.

2.7 Expression evidence

Proteomic detection was taken from PaxDb 5.0 (16 datasets); transcription start sites and N-terminal peptides for *Mycobacterium tuberculosis* from the supplementary tables of [15]. Coverage of each resource was quantified before use, and the interpretive weight of a negative result fixed in advance: the N-terminomic dataset covers 208 of roughly 4,000 proteins, so absence from it is uninformative and not counted as evidence, whereas the TSS map covers 3,038 genes, so a

dedicated TSS would have been strong counter-evidence.

All 3,978 annotated genes were paired with their curated annotation records (3,907 matched) and scored on five criteria: uncharacterised verdict, length at most 600 nucleotides, overlap of at least 25% with an immediate neighbour, essentiality, and absence from all proteomic datasets.

2.8 What the annotation record claims

Evidence qualifiers were read directly from the RefSeq GFF3 annotation of NC_000962.3. The `experiment=` and `inference=` qualifiers are carried by the CDS feature, not the gene feature; reading them from the latter returns no evidence for any gene in the genome — which has the appearance of a result. A locus was called unsupported when it carried neither. Gene identifiers came from the `Dbxref=GeneID:` field of the gene feature; these integers are assigned by the archive in increasing order, so contiguous blocks separated by large discontinuities correspond to successive deposition rounds. We did not fix block boundaries in advance but cut the sorted identifiers wherever a gap exceeded a hundred times the median gap between consecutive values, and verified no gap fell in the intermediate range that would make the partition arbitrary. Curated free-text comments were read from the reference resource for this genome. Comparisons of the unsupported fraction between groups of genes were matched on length throughout, since short genes are less studied for reasons unrelated to the question; where a group is compared with the rest of the genome, the null came from drawing, for each member, a control from the same length class, over 20,000 replicates with a fixed seed.

3 Results

3.1 Rv2438A is superimposed on the *nadE* locus

Rv2438A spans 2,736,709–2,736,987 on the plus strand (279 nt, 92 codons). *nadE* (Rv2438c) spans 2,734,792–2,736,831 on the minus strand, so its start codon lies at 2,736,831. The two features overlap by 123 nt, that is 44% of Rv2438A, covering the first 41 codons of *nadE*. The remaining 156 nt of Rv2438A lie between +1 and +156 upstream of the *nadE* start codon, in its promoter region (Figure 2, Table 1). Every nucleotide of Rv2438A is therefore accounted for by *nadE*: 44% as coding sequence, 56% as promoter.

3.2 The essentiality call rests on two informative insertion sites

Rv2438A contains only five TA dinucleotides over its 279 nt. Three lie inside *nadE* — itself essential, CRISPRi vulnerability −5.54 (CI [−5.75, −5.33]) — where an insertion is lethal through disruption of that gene. Only two TA sites, at 2,736,884 and 2,736,954, are specific to Rv2438A, both in the *nadE* promoter. The "essential" call therefore rests on two informative sites, in a region independently expected to be lethal if disrupted.

Sensitivity to the annotated *nadE* start

A curated ancestral annotation places the *nadE* start 11 codons downstream (2,736,798 against 2,736,831), which would move four of the five TA sites onto Rv2438A instead of two; we retain

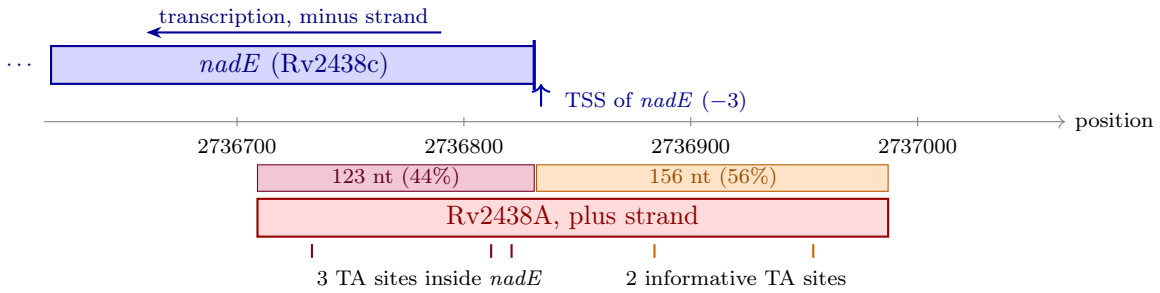


Figure 2: Architecture of the Rv2438A locus. The annotated open reading frame (red, plus strand) overlaps the essential *nadE* coding sequence (blue, minus strand) over 123 nt, that is 44% of its length (purple). Its remaining 156 nt (orange) lie upstream of the *nadE* start codon and carry the promoter, including the mapped transcription start site three nucleotides upstream of that start. Of the five *Himar1* TA insertion sites of Rv2438A, three fall inside *nadE* and only two are specific to it, both within the *nadE* promoter and flanking its initiation site.

the reference-annotation figures because the mapped transcription start site fits it far better (Supplementary Text S1.5). The TA-site argument is the only line of evidence sensitive to this choice — the codon-position, proteomic, homology and transcription-start results all concern the non-overlapping region, on which both annotations agree.

3.3 No guide RNA can target Rv2438A without targeting *nadE*

The CRISPRi signal admits the same architectural explanation, geometric rather than statistical. Guides in this library target the non-template strand at every available protospacer-adjacent motif [2], distributed over the whole length of Rv2438A — but that length is entirely accounted for by *nadE*, 44% within its coding sequence and 56% over its promoter and transcription start site, so whichever guide is considered, the dCas9 complex binds *nadE* sequence. Confirmed directly: of the 28 guides the library assigns to Rv2438A, located by exact match, six fall inside the *nadE* coding sequence and the remaining 22 in its promoter, none outside *nadE* (Supplementary Text S3).

The two zones act by different routes and predict a specific signature: within the coding sequence dCas9 impedes elongation partially, efficiency depending on which strand is bound; over the promoter, where *nadE*'s mapped start site lies, it occludes initiation strand-independently, the more effective silencing mode. An indirect knockdown through this mixture should be both attenuated and noisier than a direct one. Both hold: *nadE* measured directly gives -5.54 (CI $[-5.75, -5.33]$, width 0.42) against Rv2438A's -4.56 (CI $[-7.19, -1.90]$, width 5.29) — 18% weaker, an interval thirteen times wider, the profile of an indirect measurement. Bosch et al. [2] do control for neighbour effects, but for polar silencing of co-directional operonic genes downstream of the binding site; that control does not cover an antiparallel frame superimposed on the same DNA, where the affected gene is neither downstream nor on the same strand.

3.4 The cross-species homologue is the *nadE* start junction

The *M. decipiens* hit that had previously downgraded Rv2438A from "MTBC-specific" covers query residues 25–53 and touches neither the N-terminus nor the last 39 residues. Before ex-

amining its location we formulated the prediction that, under the superimposition hypothesis, it should map onto the *nadE* start junction. It does. In *M. decipiens* the hit lies at 3,563,961–3,564,047, while *nadE* occupies 3,561,972–3,564,011 with its start codon at 3,564,011 (alignment spanning the full query, 680 aligned positions, 91.6% identity, $E = 0$). Thus 58.6% of the hit lies inside *nadE* and 41.4% upstream of it, the same proportion as in H37Rv (17 of 29 residues and 12 of 29).

An independent confirmation follows. *M. decipiens* carries a second *nadE*-like region near 3,565,500–3,565,660, and the second high-scoring pair of Rv2438A falls there too. Wherever *nadE* has a hit, Rv2438A has one at the same place, and nowhere else. The apparent homology is the shadow of the neighbour's: because *nadE* is essential, its nucleotide sequence is conserved between species, and so is its translation in any superimposed frame.

3.5 Rv2438A has no transcription start site; *nadE*'s lies inside it

A single transcription start site is mapped in the window 2,736,500–2,737,200: position 2,736,834 on the minus strand, assigned to *nadE* at position −3, a near-leaderless transcript, the majority initiation mode in mycobacteria [15]. That site lies inside Rv2438A, whose non-overlapping part begins at 2,736,832. The only TSS the map associates with Rv2438A itself lies 1,215 nt upstream, inside *nadE* on the opposite strand — a 5' untranslated region of that length is implausible in a bacterium, and the association is an artefact of the "downstream same-strand gene" column. Rv2438A therefore has no evidence of autonomous transcription, and the two informative TA sites of the essentiality call flank the initiation site of an essential transcript.

3.6 No protein is detected where *nadE* is

The audit's fifth layer asks not whether the candidate is detected, which for a 92-codon product would prove little, but whether it is detected in the datasets that detect its neighbour — a paired comparison in which the two sequences share every sample, every digestion and every instrument. Across the sixteen datasets of PaxDb 5.0, *nadE* is detected in fifteen and Rv2438A in none.

The limit of this observation is real and we do not overstate it: a short, possibly disordered protein can be produced and still escape mass spectrometry, so absence here is not proof of absence in the cell. Its weight comes from the pairing rather than from the absence alone. It is also the reason the audit does not rest on this layer: the argument is carried by the four preceding ones, each of which identifies a mechanism by which a signal attributed to Rv2438A is in fact produced by *nadE*.

3.7 Substitutions show no reading-frame signature

The codon-position test uses the unfiltered variant catalogue, not the frequency-thresholded count quoted above: at a sampling depth of 274,680 genomes almost every position eventually varies in some isolate, and the 121 variants of the non-overlapping region are overwhelmingly singletons or doubletons. That is deliberate, since the statistic of interest is the distribution across codon positions, which noise dilutes but does not bias, and restricting to frequent variants would leave too few observations. Over the non-overlapping region of Rv2438A, substitutions

distribute across codon positions as 37/45/39, 32.2% at the third position, indistinguishable from the one-third expected without a reading frame ($\chi^2 = 0.86$, 2 d.f., $p \approx 0.65$). The two flanking control genes, analysed in the same query, behave as expected of translated sequences: *nadE* gives 74.6% and *proB* 75.3% at the third position among variants carried by at least twenty isolates (Table 2). Applied to the 121 non-overlapping variants, that rate predicts roughly 65 third-position variants against 39 observed: the test excludes a coding signature of the magnitude carried by the immediate neighbours, though not a weak one.

Applied to every gene fully contained in the sequenced windows with protein directly attested by mass spectrometry, the test recovers only 2 of 5 as coding — a likelihood ratio near 1.7 for the non-coding hypothesis, weak, specific but insensitive. The failure is not gene length but variant density, which tracks purifying constraint rather than translation as such (Supplementary Text S1, validated against three independent controls there). Below 0.55 variants per nucleotide the test recovers 71% of coding genes; above it, 34%, and only one essential gene in three falls below that threshold. Rv2438A, at 0.63, sits inside the range the test cannot resolve, why this line of evidence is ranked last below. No false positive occurred in any of these controls: the statistic never called a locus coding that lacked protein evidence.

The non-overlapping part of Rv2438A is nonetheless the least variable region of the neighbourhood (25.6 variants per kb carried by at least twenty isolates, against 100.8 and 154.9 for the adjacent intergenic region and *nadE*) — real constraint, one of the gene's original arguments, but also exactly what an essential gene's promoter is expected to show without any coding constraint.

A translated essential gene cannot tolerate premature stop codons, yet the non-overlapping part of Rv2438A accumulates four among 98 substitutions (4.1%) in six viable isolates, against none among 1,043 in *nadE* and one among 570 in *proB* (odds ratio 68.6, $p = 4.9 \times 10^{-5}$). The comparison confounds coding constraint with the constraint of the locus itself, though: a shifted-frame control removes most of the signal and places Rv2438A at an unremarkable 80th percentile among sixty real genes (Supplementary Text S1.6). What survives, independent of any control, is that six viable isolates carry a premature stop where *nadE* carries none: a weak converging signal, not a demonstration.

3.8 The screening criterion promotes the artefact

A spurious ORF overlapping an essential gene is dark, since there is nothing to annotate; small, a predicted fragment; essential, since insertions destroy the neighbour or its promoter; and vulnerable, since knockdown silences the neighbour. It satisfies the four criteria of a dark-target screen better than a genuine small gene, inheriting the neighbour's signals undiluted.

We tested how often this occurs. Of 3,907 annotated genes, six are uncharacterised, at most 600 nt long, and overlap a neighbour by at least 25%; exactly one also combines essentiality with absence from all proteomic datasets, and it is Rv2438A (Table 3). The artefact is therefore rare among genes but, in our screen, it was ranked first. This asymmetry is the practical point: rarity in the genome does not translate into rarity among hits. The codon-position test was applied to the other five candidates and their neighbours as well; none falls in the variant-density window where the test is informative (Supplementary Text S1.7), which in hindsight is why Rv2438A and *nadE*, at 0.50–0.51 variants per nucleotide, were the pair on which it could decide anything.

4 Discussion

4.1 Relation to previous work

That an antiparallel overlapping ORF is unlikely to be a genuine essential gene is not our claim. Moshensky and Alexeevski [4] reached it across 929 prokaryotic genomes using stop-codon avoidance and K_a/K_s; their *dnaK* example, 282 such frames entered as a family of glutamate dehydrogenases in Pfam, shows how far the consequences propagate. The mechanistic dissection of transposon essentiality artefacts is likewise established, by Choe et al. [5] in *E. coli*. Applying their approach here (Supplementary Text S2) shows the antiparallel frame itself to be ordinary, not rare: 5,458 frames of at least Rv2438A's length oppose the 3,906 coding sequences of H37Rv, essentially what composition alone predicts (ratio 0.97), with no counterpart of the *dnaK* Pfam family in this genome.

What we add is what happens downstream: those analyses work on comparative sequence evidence and conclude about the reading frame's reality; they do not examine the experimental layers a target pipeline actually consumes. We show each layer is contaminated by an identifiable mechanism — essentiality via shared TA sites, promoter disruption via the remaining two, conservation via promoter constraint, cross-species homology via the neighbour's own conservation read in the superimposed frame, apparent transcription via the neighbour's start site lying inside the ORF, knockdown vulnerability via guides that cannot bind anywhere but the neighbour — and quantify a selection bias not previously stated: because the artefact scores well on every criterion, its frequency among hits is unrelated to its frequency among genes. One locus in 3,907 carried the complete profile and was ranked first, independent of where the screen's two thresholds sit, over 49 combinations tested (Supplementary Text S2).

In this species the failure mode we describe has not yet produced a published error: of the six short, uncharacterised genes overlapping an antiparallel neighbour by a fifth of their length or more, only one carries any publication (Supplementary Text S2). That does not make the problem theoretical: the annotation has already reached three reference proteomes and a clone repository, and Moshensky and Alexeevski [4] traced an entire Pfam family to antiparallel frames elsewhere — the mechanism reaches databases well before it reaches articles.

A second, independent reason compounds that bias: the criteria such screens combine are proxies for gene length more than for anything else. Between coding sequences shorter than 150 nt and longer than 1200 nt, the proportion called dark rises 38-fold, domainless 17-fold, proteomically silent 9-fold, while essentiality runs the other way as supporting TA sites collapse from a median of 25 to four. A screen conjoining these criteria applies nearly the same filter four times, and an essentiality call on a short gene rests, in most cases, on a handful of observations before overlap even arises — Rv2438A, three of five sites inside *nadE*, illustrates a general situation, not a special one. A length threshold instead of an overlap threshold retrieves the same locus independently (Supplementary Text S2): of 293 coding sequences under 300 nt, exactly one is dark, proteomically undetected, essential, and overlapping a neighbour by at least a quarter of its length — Rv2438A.

Two boundaries are worth stating. Moshensky and Alexeevski's overlap threshold, 180 bp, would not have retrieved Rv2438A's 123 bp, so shorter overlaps deserve the same scrutiny; a

reader who already accepts their conclusion should read this work as being about screen design, not about the reality of one reading frame. The two studies differ in null model, not in fact: their reported excess of antiparallel frames disappears once codon usage is held constant, as we do (Supplementary Text S2) — the cause they themselves propose, stop-codon avoidance as compositional, not selected.

4.2 Which lines of evidence carry the argument

The six observations are not equally strong, and a list invites averaging them. We rank them on two criteria: whether they survive a change in the annotated *nadE* start, and how much they would move a reader who did not already accept the conclusion.

The architecture is the foundation, and it is invariant. Under the reference annotation the split is 44%/56% coding sequence/promoter; under the ancestral annotation, 32%/68%. The proportions move, the total does not: every nucleotide of the locus belongs to *nadE* either way. It proves nothing about translation alone, but it is what makes the others interpretable, and no annotation choice threatens it.

Two observations carry most of the argument: both took the form of predictions, and neither rests on a contestable parameter. The *M. decipiens* hit was predicted to fall on the *nadE* start junction, and does, in both species and identical proportion, with the secondary high-scoring pairs coinciding too. The CRISPRi argument is geometric, not statistical: of the 28 guides actually assigned to Rv2438A, none lands outside *nadE*, and the predicted signature of an indirect measurement, attenuated and noisier, is what the published index shows.

Two further observations are strong but bounded: the absence of a dedicated transcription start site is direct evidence, tempered by 76% map coverage, so an absence is suggestive, not conclusive; the proteomic contrast, zero detections against fifteen for the neighbour, is a sharp comparison, tempered by the possibility that a short protein escapes mass spectrometry.

Two observations carry little weight. The TA-site argument is the one line sensitive to the annotation question above, moving from two informative sites to four. The codon-position test, sensitivity 2 of 5 and likelihood ratio near 1.7, is consistent with the conclusion and produced no false positive, but cannot establish anything alone — and its own applicability condition is not met here: variant density 0.63 places it where the test misses two thirds of genes with attested protein, so we treat it as corroboration, not evidence.

Read this way, the case does not depend on its weakest parts: removing the last two still leaves the architecture, two verified predictions and two bounded observations — the argument we would defend.

4.3 What is and is not established

We have not proved that Rv2438A is untranslated: a short protein can be genuinely produced and still escape mass spectrometry, the project's original argument and still admissible. What we have established is that a single architectural fact accounts for every observation that supported the gene, and that the alternative hypothesis made three quantitative predictions, verified: the *M. decipiens* hit, the transcription start site, and the CRISPRi index profile. A formal demonstration would require ribosome profiling: interrogating Smith et al. [16]'s genome-wide Ribo-RET map,

no footprint falls on the plus strand within Rv2438A — consistent with the rest of the case but inconclusive, since the assay recovers an initiation site for only 29% of annotated coding sequences and *nadE* itself has none detected (Supplementary Text S3). We are equally explicit about evidence declined for want of power: the N-terminomic dataset of [15] covers too few proteins (208 of roughly 4,000) to weigh an absence, and contains neither gene. No observation of the dossier remains uninstructed.

A further limitation is specific to the expression evidence and deserves to be stated on its own. The proteomic and transcript resources interrogated here sample a finite set of growth conditions, and none was designed to capture a locus expressed only under stress. Absence from PaxDb and from the transcription-start map therefore excludes constitutive expression under the conditions assayed; it does not exclude expression under a condition nobody has assayed, whether hypoxia, nutrient starvation, acid or oxidative stress, or intracellular residence. A dedicated transcript or proteomic survey under such conditions would be the direct test, and we do not treat the question as closed. It would, however, have to contend with the architecture: a transcript specific to Rv2438A would have to initiate independently of the *nadE* start site mapped three nucleotides upstream of the *nadE* start codon, and any protein produced would be encoded, over 44% of its length, on the strand opposite an essential gene.

4.4 What such an annotation leaves in the reference databases

The consequences do not stay local. Rv2438A's UniProt entry (Q79FD9) is unreviewed, scored 1 of 5 and recorded as "4: Predicted", yet carries twelve cross-references across ten databases, including a downloadable AlphaFold DB structure and a clone-repository identifier (full inventory in Supplementary Text S3). None of these resources asserts the protein exists, and UniProt says the opposite in two fields — but those fields do not travel: what travels is a sequence, a structure file, a clone identifier, each usable without consulting what it rests on, regardless of a pLDDT of 54. This is an inventory on one locus, not a frequency estimate, since every annotated reading frame receives such entries; it complements, on a case inspectable in full, the Pfam family Moshensky and Alexeevski [4] traced to antiparallel frames opposite *dnaK*.

The annotation has also travelled between genomes: in *M. bovis* AF2122/97 the architecture is reproduced to the nucleotide — same overlap, same upstream length — and the UniRef90 cluster holds three equally unreviewed, equally "Predicted" entries, one per reference proteome. This is not independent evidence: annotation in these genomes derives from the reference by automatic orthology, and reading it as cross-species conservation would repeat the *M. decipiens* error. One feature does carry weight: UniRef50 contains the same four members, from the same three species and no other, exactly where *nadE* ceases to be conserved closely enough for the opposite phase to stay free of stop codons (Supplementary Text S3) — a gradient governed by sequence divergence alone.

4.5 Where such an annotation comes from

The record itself answers part of the question, in the open. The curated comment on Rv2438A justifies "conserved" by three FASTA similarities at $E = 1.5$, the number of chance matches of that quality, to proteins from three unrelated domains of life; it opens, in its own words, with

"showing few similarity". Neighbouring records show this is not formulaic: curators recorded genuine homology or overlap where they found one, for *nadE*, Rv0810c and one scan candidate, but never for Rv2438A, whose 44% overlap with an essential gene appears nowhere in the entry (Supplementary Text S3). Entry order points the same way — *nadE* from this genome's original submission [17], Rv2438A from the later batch adding the 82 coding sequences of its re-annotation [18] — establishing at least that the host gene was there first (Supplementary Text S3).

Beyond this locus, why is Rv2438A annotated at all? Microbial gene finders build species-specific models by self-training on long, unambiguous open reading frames, then extrapolate to short candidates [13, 14]; their authors are explicit that many genes they miss "appear to be hypothetical proteins whose existence is only supported by the predictions of other programs" [13] — a description that fits a spurious ORF as well as a missed one. Mycobacteria compound the difficulty: nearly a quarter of their transcripts are leaderless, an ATG or GTG at the mRNA 5' end being necessary and sufficient for initiation [15], so the features that normally discriminate a genuine start are weak or absent [19]. A short GTG-initiated frame opposite an essential gene is close to the worst case for prediction, and such an annotation should be treated as a hypothesis, particularly once a screen has selected it.

4.6 Why the artefact is invisible to the usual controls

Each signal was, alone, reasonable: essentiality came from a saturating screen, conservation was measured on 145,209 isolates, the cross-species hit came from a deliberately sensitive safeguard search. None of these controls examines the coordinates of the feature relative to its neighbours: overlap is not part of the standard annotation payload, and in its absence every downstream layer silently inherits the neighbour's biology.

The cross-species hit deserves particular attention because it was itself introduced as a corrective. A strict `tblastn` threshold had called Rv2438A MTBC-specific; a relaxed search found the *M. decipiens* hit and correctly withdrew that claim. The correction was right, and it was right for the wrong reason: the hit was never a homologue. A safeguard can produce the correct verdict through an artefact, and remain an artefact.

4.7 Recommendations

Four measures follow, all cheap. First, report overlap fraction and the neighbour's identity as a first-class annotation for genes under 400 nt; this single field would have flagged Rv2438A on day one. Second, for transposon-based essentiality on a short gene, report TA sites falling outside any overlap: five sites of which three are shared is not the same evidence as fifty. Third, retain alignment coordinates in any homology screen and keep sub-threshold hits — both are free in the aligner's output; re-running our own orthology screen this way shows 43% of the genes it had called complex-specific do have a weak hit elsewhere, Rv2438A among them, mostly genuine but divergent homologues rather than superimposed frames (Supplementary Text S3). Fourth, read the evidence qualifiers RefSeq already writes in the annotation file every screen parses: *nadE* cites mass spectrometry, Rv2438A, three kilobases away, cites nothing — a systematic contrast, short overlapping frames carrying markedly less claimed evidence than short frames generally,

matched by length ($p = 0.0013$, Supplementary Text S3; a simpler flag, the locus-tag letter suffix, showed the same pattern but did not survive the same control).

A fifth, methodological, recommendation follows from our own error: the codon-position test should run with positive controls of *comparable length*. Ours were four to seven times longer than the locus under test, which flattered the contrast and revealed an off-by-one frame error that would otherwise have inverted the conclusion for *nadE* itself. Measured against genes with attested protein the test recovers only 2 of 5, so a negative outcome is weak, never decisive, and usable only within its applicability condition: not above roughly 0.55 variants per nucleotide in a catalogue of this depth, a threshold to recalibrate against sampling depth, not reuse as a constant. Only about a third of essential genes meet it, and very few others — not this locus — which is why we rank the test last among our lines of evidence; the released code implements the condition, so a negative verdict above threshold returns as uninformative rather than as absence of coding.

None of this is specific to *Mycobacterium tuberculosis*: it requires only a compact genome, gene prediction permissive enough to call short ORFs on both strands, and a screen rewarding essentiality in short uncharacterised genes — conditions common across bacterial pathogens, where the diagnostic transfers unchanged.

5 Conclusion

Rv2438A satisfies every criterion of a promising antitubercular target, and no evidence that it encodes a protein survives the audit: it is superimposed on the essential gene *nadE*, and every experimental signal that recommended it as a target traces to that single architectural fact. We do not claim to have proved it untranslated — a short protein can be produced and still escape mass spectrometry, and even comprehensive ribosome profiling could not exclude expression under an untested growth condition — but no ground remains on which it was promoted. Because such a locus satisfies the criteria of a dark-target screen better than a genuine small gene, it is not merely tolerated by these pipelines but promoted by them: our own screen ranked it first among 3,907 genes, while a genome-wide scan finds it to be the only locus in *Mycobacterium tuberculosis* with the complete profile.

Table 1: Architecture of the *nadE* locus in H37Rv (NC_000962.3). Rv2438A overlaps the essential *nadE* coding sequence on the opposite strand, and its remaining portion carries the *nadE* promoter and transcription start site.

Locus	Gene	Start	End	Strand	Length (nt)
Rv2437	,	2734376	2734795	+	420
Rv2438c	*nadE*	2734792	**2736831**	−	2040
Rv2438A	,	**2736709**	2736987	+	**279**
Rv2439c	*proB*	2737117	2738247	−	1131
Overlap Rv2438A / *nadE*: **123 nt = 44% of Rv2438A** (first 41 codons of *nadE*)					
Remainder: 156 nt at +1 to +156 upstream of the *nadE* start codon					
nadE transcription start site: **2736834** (−3), inside Rv2438A					

Table 2: Codon-position distribution of substitutions. The two flanking genes, analysed in the same query as positive controls, show the third-position excess expected of translated sequences; the non-overlapping part of Rv2438A does not.

Sequence	Pos. 1	Pos. 2	Pos. 3	% at pos. 3	χ^2 (2 d.f.)
nadE (control), all variants	377	310	**788**	**53.4**	272.5
variants ≥ 20 isolates	30	22	**153**	**74.6**	157.8
proB (control), all variants	233	177	**490**	**54.4**	185.7
variants ≥ 20 isolates	26	14	**122**	**75.3**	129.8
Rv2438A, non-overlapping, all variants	37	45	39	**32.2**	**0.86** ($p \approx 0.65$)

Table 3: Genome-wide screen for superimposed ORFs among 3,907 annotated genes. Six loci are uncharacterised, at most 600 nt and overlap a neighbour by at least 25%; only Rv2438A also combines essentiality with absence from all 16 proteomic datasets.

Locus	Length (nt)	Overlap (nt)	Overlap (%)	Essential	MS-detected	Neighbour
Rv2438A	279	123	**44**	**yes**	**no**	Rv2438c
Rv1434	138	52	38	no	no	Rv1435c
Rv2662	273	95	35	no	no	Rv2661c
Rv3861	327	102	31	no	yes	Rv3862c
Rv1374c	459	140	31	no	yes	RVnc0035
Rv0609A	228	58	25	no	no	Rv0609

Data and code availability

Twenty-two standard-library scripts are released with this article: an implementation of the codon-position test, which encapsulates the coordinate-system conversion, keeps the phase reference distinct from the queried window, refuses to build a query containing fewer than two regions so that positive controls cannot be omitted, and returns a negative verdict as uninformative when the variant density of the locus exceeds the applicability threshold; the genome-wide screen for superimposed open reading frames; the construction of the reference population with its available power; a declarative cross-check that recomputes every coordinate and insertion-site figure reported here from the reference genome and fails on any divergence; and five scripts measuring the applicability of the codon-position control, which respectively sweep the frequency threshold with its statistical power at each step, test density against length on sixty randomly drawn CDS under a fixed seed, separate a calling artefact from a constraint effect by frequency stratification and by overlap with the repeat mask, compare variant density between length-matched essential and non-essential genes, evaluate a frequency-trend variant of the test that we report as unsuccessful, measure the base rate of each screening criterion by gene-length class, census the antiparallel reading frames of the genome against a synonymous-codon null, cross Pfam domains with overlap and strand, reproduce the inter-genome architecture and UniRef cluster composition, translate every catalogued substitution in the frame to count premature stop codons against coding and frame-shifted controls, sweep the screen's two thresholds to test the stability of its count, cross antiparallel overlap with the publication record at matched gene length, and place the frame-contrast of a locus in the distribution of sixty randomly drawn coding sequences. Four further scripts read the ninth column of the RefSeq annotation, discarded

by the other parsers: a reader for its experiment and inference qualifiers and for gene-identifier deposition order; the script applying it to test whether short overlapping loci carry less claimed evidence than length-matched controls; a reader that locates every CRISPRi guide assigned to a locus on the reference genome by exact match and reports whether it falls inside that locus alone; and the script applying it to Rv2438A. A twenty-second script applies the codon-position test to the five other scan candidates and their neighbours (Supplementary Text S1.7). All are deposited, with a README documenting the input formats and the measured limits of the codon-position control, at https://doi.org/10.5281/zenodo.21718307. Genome annotation is from NC_000962.3; population variation was queried from a curated MTBC variant database; proteomic detection from PaxDb 5.0; transcription start sites and N-terminal peptides from the supplementary tables of [15]; CRISPRi guide sequences from the library deposited by Bosch et al. [2].

CRediT authorship contribution statement

Christophe Guyeux: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing: original draft, Writing: review and editing.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The author declares no competing interests.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author used OpenAI Codex to assist with literature searching, code development, verification, and language editing, and Claude Code (Anthropic, Claude Sonnet 5) to assist with the present revision, namely the measurement of prior art reported in the Discussion, the reorganisation of the Results along the hierarchy of evidence, and the drafting of Figure 1. The author subsequently reviewed and edited the content, independently verified the reported results against the source data, and takes full responsibility for the content of the publication.

References

[1] M.A. DeJesus, E.R. Gerrick, W. Xu, S.W. Park, J.E. Long, C.C. Boutte, E.J. Rubin, D. Schnappinger, S. Ehrt, S.M. Fortune, C.M. Sassetti, and T.R. Ioerger. Comprehensive Essentiality Analysis of the Mycobacterium tuberculosis Genome via Saturating Transposon Mutagenesis. *mBio*, 8(1), 2017. doi: 10.1128/mBio.02133-16. verified 2026-07-30 via NCBI E-utilities.

[2] B. Bosch, M.A. DeJesus, N.C. Poulton, W. Zhang, C.A. Engelhart, A. Zaveri, S. Lavalette, N. Ruecker, C. Trujillo, J.B. Wallach, S. Li, S. Ehrt, B.T. Chait, D. Schnappinger, and J.M. Rock. Genome-wide gene expression tuning reveals diverse vulnerabilities of M. tuberculosis. *Cell*, 184(17):4579–4592.e24, 2021. doi: 10.1016/j.cell.2021.06.033. verified 2026-07-30 via NCBI E-utilities.

[3] S. Lata and D. Heinz. ECLIPSE: exploring the dark proteome of ESKAPE pathogens through the sequence similarity network of the Protein Universe Atlas. *Bioinformatics*, 42 (8):btag613, 2026. doi: 10.1093/bioinformatics/btag613. verified 2026-09-08 via CrossRef API.

[4] Denis Moshensky and Andrei Alexeevski. Long antiparallel open reading frames are unlikely to be encoding essential proteins in prokaryotic genomes. *bioRxiv*, 2019. doi: 10.1101/724807. preprint, posted 8 August 2019; no journal publication found in Crossref (relation field empty, type still posted-content) or PubMed (author-pair search returns one unrelated 2023 paper, PMID 37072330, different topic) as of 2026-08-03; metadata read directly from the deposited PDF.

[5] D. Choe, U. Kim, S. Hwang, S.W. Seo, D. Kim, S. Cho, B. Palsson, and B.K. Cho. Revealing Causes for False-Positive and False-Negative Calling of Gene Essentiality in Escherichia coli Using Transposon Insertion Sequencing. *mSystems*, 8(1):e0089622, 2023. doi: 10.1128/msystems.00896-22. verified 2026-07-31 via NCBI E-utilities.

[6] B.A. Brown-Elliott, P.J. Simmer, A. Trovato, E.P. Hyle, S. Droz, S.P. Buckwalter, E. Boroni, J.A. Branda, E. Iana, A. Mariottini, J. Nelson, A. Matteelli, N.C. Toney, C. Scarparo, T.J.B. de Man, R. Vasireddy, R.T. Gandhi, N.L. Wengenack, D.M. Cirillo, R.J. Wallace, and E. Tortoli. Mycobacterium decipiens sp. nov., a new species closely related to the Mycobacterium tuberculosis complex. *International journal of systematic and evolutionary microbiology*, 68(11):3557–3562, 2018. doi: 10.1099/ijsem.0.003031. verified 2026-07-30 via NCBI E-utilities.

[7] G. Sapriel and R. Brosch. Shared Pathogenomic Patterns Characterize a New Phylotype, Revealing Transition toward Host-Adaptation Long before Speciation of Mycobacterium tuberculosis. *Genome biology and evolution*, 11(8):2420–2438, 2019. doi: 10.1093/gbe/evz162. verified 2026-07-30 via NCBI E-utilities.

[8] R. Cantoni, M. Branzoni, M. Labò, M. Rizzi, and G. Riccardi. The MTCY428.08 gene of Mycobacterium tuberculosis codes for NAD+ synthetase. *Journal of bacteriology*, 180

(12):3218–21, 1998. doi: 10.1128/JB.180.12.3218-3221.1998. verified 2026-07-30 via NCBI E-utilities.

[9] H.I. Boshoff, X. Xu, K. Tahlan, C.S. Dowd, K. Pethe, L.R. Camacho, T.H. Park, C.S. Yun, D. Schnappinger, S. Ehrt, K.J. Williams, and C.E. Barry, 3rd. Biosynthesis and recycling of nicotinamide cofactors in mycobacterium tuberculosis. An essential role for NAD in nonreplicating bacilli. *The Journal of biological chemistry*, 283(28):19329–41, 2008. doi: 10.1074/jbc.M800694200. verified 2026-07-30 via NCBI E-utilities.

[10] I.A. Rodionova, B.M. Schuster, K.M. Guinn, L. Sorci, D.A. Scott, X. Li, I. Kheterpal, C. Shoen, M. Cynamon, C. Locher, E.J. Rubin, and A.L. Osterman. Metabolic and bactericidal effects of targeted suppression of NadD and NadE enzymes in mycobacteria. *mBio*, 5(1), 2014. doi: 10.1128/mBio.00747-13. verified 2026-07-30 via NCBI E-utilities.

[11] R. Sharma, T.E. Hartman, T. Beites, J.H. Kim, H. Eoh, C.A. Engelhart, L. Zhu, D.J. Wilson, C.C. Aldrich, S. Ehrt, K.Y. Rhee, and D. Schnappinger. Metabolically distinct roles of NAD synthetase and NAD kinase define the essentiality of NAD and NADP in Mycobacterium tuberculosis. *mBio*, 14(4):e0034023, 2023. doi: 10.1128/mbio.00340-23. verified 2026-07-30 via NCBI E-utilities.

[12] S. Talwar, M. Pandey, C. Sharma, R. Kutum, J. Lum, D. Carbajo, R. Goel, M. Poidinger, D. Dash, A. Singhal, and A.K. Pandey. Role of VapBC12 Toxin-Antitoxin Locus in Cholesterol-Induced Mycobacterial Persistence. *mSystems*, 5(6):e00855–20, 2020. doi: 10.1128/msystems.00855-20. verified 2026-09-08 via CrossRef API and Europe PMC full text.

[13] A.L. Delcher, D. Harmon, S. Kasif, O. White, and S.L. Salzberg. Improved microbial gene identification with GLIMMER. *Nucleic acids research*, 27(23):4636–41, 1999. doi: 10.1093/nar/27.23.4636. verified 2026-07-31 via NCBI E-utilities.

[14] J. Besemer, A. Lomsadze, and M. Borodovsky. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic acids research*, 29(12):2607–18, 2001. doi: 10.1093/nar/29.12.2607. verified 2026-07-31 via NCBI E-utilities.

[15] S.S. Shell, J. Wang, P. Lapierre, M. Mir, M.R. Chase, M.M. Pyle, R. Gawande, R. Ahmad, D.A. Sarracino, T.R. Ioerger, S.M. Fortune, K.M. Derbyshire, J.T. Wade, and T.A. Gray. Leaderless Transcripts and Small Proteins Are Common Features of the Mycobacterial Translational Landscape. *PLoS genetics*, 11(11):e1005641, 2015. doi: 10.1371/journal.pgen.1005641. verified 2026-07-30 via NCBI E-utilities.

[16] Carol Smith, Jill G. Canestrari, Archer J. Wang, Matthew M. Champion, Keith M. Derbyshire, Todd A. Gray, and Joseph T. Wade. Pervasive translation in *Mycobacterium tuberculosis*. *eLife*, 11:e73980, 2022. doi: 10.7554/eLife.73980.

[17] S.T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S.V. Gordon, K. Eiglmeier, S. Gas, C.E. Barry, 3rd, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd,

T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M.A. Quail, M.A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J.E. Sulston, K. Taylor, S. Whitehead, and B.G. Barrell. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685): 537–44, 1998. doi: 10.1038/31159. verified 2026-08-01 via NCBI E-utilities.

[18] J.C. Camus, M.J. Pryor, C. Médigue, and S.T. Cole. Re-annotation of the genome sequence of Mycobacterium tuberculosis H37Rv. *Microbiology (Reading, England)*, 148(Pt 10):2967–2973, 2002. doi: 10.1099/00221287-148-10-2967. verified 2026-08-01 via NCBI E-utilities.

[19] A. Lomsadze, K. Gemayel, S. Tang, and M. Borodovsky. Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome research*, 28(7):1079–1089, 2018. doi: 10.1101/gr.230615.117. verified 2026-07-31 via NCBI E-utilities.