ARTICLE · Posted as a preprint on aiXiv (aixiv.260822.000005). Not submitted to or accepted by a journal.

# A fusion-junction vaccine in extraskeletal myxoid chondrosarcoma: what can be established today, and the capabilities that would change it

**Tristan D. McRae**

*Independent researcher, unaffiliated. Correspondence: trimcrae@gmail.com ORCID: 0000-0002-1823-1451*

Version of 2026-08-24 · typeset preview · built from 9aaf4dd29, tree not clean at build time

**ABSTRACT**

**Background.** Extraskeletal myxoid chondrosarcoma (EMC) is a rare soft-tissue sarcoma defined by rearrangement of *NR4A3*, most often to *EWSR1*. The fusion junction encodes a peptide sequence absent from either parent protein, and because the fusion is the truncal driver it is present in every tumour cell and cannot be lost without loss of the driver. The sponsors of an individualised neoantigen therapy have announced a positive phase 3 result in resected melanoma [9], which makes the platform question timely for other tumours. **Purpose.** This paper neither predicts that an EMC vaccine will work nor argues that it will not. It reports what current instruments and access establish about the target, separates limits of the tumour from limits of method or access, and records for each movable limit what would move it. **Methods.** Junctions were derived at the transcript level from Ensembl exon structure, so the acceptor exon is retained whole including its 5' untranslated region. Class I binding was predicted with MHCflurry 2.1.4, models release 2.2.0 [2], on a ten-allele panel for the junction screen and a 34-allele panel for the coverage scan, calling a peptide strong at a presentation percentile of 0.5 or below; class II binding with MHCnuggets [12] on 23 class II alleles across DR, DP and DQ at 100 and 1000 nM. Coverage is the union carrier frequency of the presenting alleles over Allele Frequency Net Database records [1]; the sampling model a pooled binomial would require does not hold, so no confidence interval is placed on it and three other readings are reported in its place: the exact within-locus form of the coverage expression, distribution-free Fréchet bounds on the between-locus dependence the formula assumes away, and the empirical spread of the same quantity recomputed inside each source population. Novelty was assessed by exact-match search against the UniProt reviewed human proteome including isoforms, with the unreviewed entries of the same reference proteome searched separately and reported separately. Clinical figures come from a curated EMC registry. No wet-laboratory data were generated. **Results.** Of 27 declared exon pairs, 5 are in frame, yielding 174 junction-spanning peptides and 11 distinct predicted binders of which 4 are strong; there is no pan-EMC epitope. Predicted coverage is a property of the screen as much as of the junction: the commonly reported *EWSR1* exon 7 to *NR4A3* exon 3 junction covers 8.5% on ten alleles, presented on HLA-B*15:01 alone, and 12.3% on 34, where the same lead peptide is also strong on HLA-A*30:02; pooling every in-frame junction gives 27.4% and 30.4% on those two panels. None is a ceiling. Coverage against the acceptance threshold is computed here as a continuous function rather than sampled: it is a step function, every step is one peptide-allele call, and the four steps below the conventional cut all fall inside a window 0.0844 percentile units wide that closes before the cut is reached. Across the span of cuts this field routinely uses it runs from 0% at a conservative 0.2 through 30.4% at 0.5 to 72.6% at 2.0, so the reported figure is one point on a curve rather than an estimate with a tolerance. It also moves with the predictor: an independently trained class I model over the same peptides and panel returns 8 presenting alleles and 45.2% at its own conventional cut, sharing only three alleles with the four above. In place of the confidence interval withdrawn from earlier versions, the pooled figure is bounded three ways: the exact within-locus form raises it by 0.33 percentage points, so it cannot be too high for that reason; Fréchet bounds place it in [17.5%, 29.9%] under any linkage disequilibrium whatever; and recomputed inside each of the 112 source populations that measured the whole presenting panel it has a median of 24.5% and a range of 0% to 66.0%. That last is an order of magnitude wider than the other two, and wider than the panel and the predictor axes, though not than the threshold. Dropping Hardy-Weinberg and between-locus independence together bounds the same figure in [9.2%, 31.1%] while assuming nothing about genotypes or haplotypes. 170 of 174 peptides are absent from the reviewed human proteome including isoforms and all 4 strong binders survive; the 4 that do not occur in an *NR4A3* isoform, belong to the four aspartate-seam junctions, and cost one predicted binder. A near-self search at one to two substitutions places the lead peptide one residue from DMPCVQAQY in that isoform and two from a paralogue peptide, neither difference at an anchor, against a chance expectation of 0.02; no binder in the screen has a near-self neighbour differing only at anchors. On a class II panel widened from three DRB1 alleles to 23 across DR, DP and DQ, 44 peptide-allele pairs bind and one is strong, on DRB1*14:01, so combined CD8 and CD4 coverage is computable for the first time and is 1.8%. The candidate construct is a 15-residue synthetic long peptide carrying both arms. The four out-of-frame junctions, screened here for the first time, give read-through tracts of 9 to 31 residues before a premature stop, and their 10 strong calls are tighter than any in-frame call — the only peptides in this work surviving a conservative cut come from junctions that cannot encode the driver. **What bounds each conclusion.** Ten limits are enumerated in Section 3 and graded there as bounded by the disease, by current instruments, or by access. Two of them, the cold-tumour limit and the physical-exclusion limit, are re-graded here against EMC immune evidence that postdates the previous version and does not choose between them. Each movable one is paired with the advance that would move it and the observation that would show it had arrived; no date is offered for any of them. **Interpretation.** The most defensible present statement is neither that the route is viable nor that it is closed, but that it is instrument-limited in identifiable ways, and that several of the numbers a reader would take as bounding it are bounding the screen instead. Two findings are offered as results. Seam-proximal peptides of four of the five in-frame junctions reproduce a sequence in a normal *NR4A3* isoform, which withdraws a predicted binder and is a defect in the novelty filter that will recur at any breakpoint whose seam reconstructs an isoform boundary. And the coverage figures this route has been graded on move with the panel and the threshold by more than the distance between them. The paper also observes, of this programme's own route ledger and not of the field, that several priming-directed classes were excluded there for want of antigen supply while a vaccine is an antigen supply, so the combination was never graded here as a unit; the standing objection to the vaccine is not the one that observation answers. Predicted binding is a screen and not evidence of presentation, immunogenicity or benefit, and nothing here supports use of any agent outside a clinical trial.

**Keywords** extraskeletal myxoid chondrosarcoma; EWSR1::NR4A3; fusion-junction neoantigen; cancer vaccine; HLA population coverage; MHC binding prediction; rare sarcoma

## 1. A standing-state report rather than a verdict

A verdict of "unpromising" delivered against a target whose presentation has never been measured records the state of the measuring apparatus, in a form that reads as a statement about the tumour and that nobody revisits when the apparatus improves. This paper separates the three kinds of limit such a verdict conflates. Some are properties of this disease and this junction (a quiet genome, a myxoid matrix proposed to exclude lymphocytes, an incidence below one per million per year) and will not move. Some are properties of today's instruments, chiefly a sequence-based predictor standing in for a measurement nobody has taken. Some are limits of access rather than of knowledge: no reachable patient material, no measured presentation, no manufacturing route at this incidence. Section 3 records, for each movable limit, the advance that would move it and the evidence that would count as that advance arriving, with no date attached; Section 6 records what would look like each arriving without being it.

The immediate occasion is external. The sponsors of an individualised neoantigen therapy given with pembrolizumab have announced that a randomised phase 3 trial met its primary endpoint of recurrence-free survival in resected stage IIB to IV melanoma [9]; that announcement is a company press release, no effect size was disclosed in it, and the peer-reviewed evidence in this setting remains the phase 2b trial [3]. The result does not transfer to EMC, and Section 4 sets out the axis on which the transfer fails. What it shows is that the manufacturing and delivery apparatus for an individualised RNA vaccine exists as a clinical reality rather than as a proposal.

## 2. Results: the target and its current evidence base

### 2.1 Disease and fusion

EMC accounts for roughly 1 to 3% of soft-tissue sarcomas, with an estimated incidence well under one per million per year [7]. It is defined by rearrangement of *NR4A3* on chromosome 9q22. *EWSR1*::*NR4A3* is the commonest fusion, reported in approximately 62 to 75% of cases [7] and in 79% of a molecularly confirmed series of 58 cases [10]; variant partners include *TAF15*, *TCF12*, *TFG* and *FUS* [7]. The genome is otherwise quiet, and the fusion is the truncal driver.

Two consequences follow, and they point in opposite directions. Because the fusion is truncal and clonal, it is present in every tumour cell and cannot be subclonally lost, so a T-cell response directed at the junction cannot be escaped by antigen loss in the way that a response against a passenger mutation can. Because the genome is quiet, the junction is close to the only tumour-exclusive antigen the disease offers, so if the junction fails there is no second candidate to fall back on. The same feature that makes the target durable makes the portfolio of targets shallow.

### 2.2 Junction structure and predicted binding

Junctions were derived from the spliced transcripts rather than from an assumed breakpoint, so the acceptor exon is retained whole with its 5' untranslated region, as a fusion transcript retains it. This matters: a superseded model that concatenated coding sequences discarded that retained region and selected a junction disjoint from the one the transcript model produces. All figures below are from the transcript model.

Of 27 declared exon pairs, 5 are in frame: *EWSR1* exons 7, 9, 10, 12 and 13 joined to *NR4A3* exon 3. The remaining 22 are graded out as non-coding acceptor (9), out of frame (4) or not producing the seam (9). The in-frame set yields 174 distinct junction-spanning peptides, screened with MHCflurry 2.1.4, models release 2.2.0 [2], over ten alleles — HLA-A*01:01, A*02:01, A*03:01, A*11:01, A*24:02, B*07:02, B*08:01, B*15:01, B*35:01 and B*44:02 — calling a peptide strong at a presentation percentile of 0.5 or below and weak at 2.0 or below. That screen returns 11 distinct predicted binders, 4 of them strong. The lead candidate at the commonly reported *EWSR1* exon 7 junction is NMPCVQAQY on HLA-B*15:01, at a presentation percentile of 0.37 and a predicted affinity of 73.4 nM. The class call is made on the percentile, not the affinity, and the two do not agree in rank order across this set, so affinities appear below only beside the percentile that classified them.

The derivation is stated as a procedure rather than described, because Section B5 reports a filter failure and a reader cannot locate a failure in prose. For one donor exon *d* and acceptor exon *a*:

1. Build the chimeric mRNA as the donor transcript through the 3' end of exon *d*, then acceptor exon *a* **whole** — its 5' untranslated region included, as a fusion transcript retains it — and everything 3' of it.
2. Translate from the donor's own initiator codon, so the reading frame is the donor's throughout.
3. Let $j_0$ be the index of the first residue not encoded wholly by donor nucleotides. If the donor cut leaves a partial codon, the residue at $j_0$ is completed by acceptor nucleotides and belongs to neither parent; if the cut is codon-aligned there is no such residue.
4. Verify the seam rather than assume it: the chimeric N-terminus must match the donor protein residue for residue up to $j_0$, the C-terminus must be the acceptor's own, and acceptor residue 1 must appear immediately 3' of the seam codon. A junction failing any of the three is an error, not a result.
5. Enumerate every 8- to 11-mer containing $j_0$ — including those that begin at it, which a straddle test requiring a residue on each side would drop.
6. **The novelty filter, and the step Section B5 is about:** keep a peptide if it is not a substring of wild-type *EWSR1* and not a substring of wild-type *NR4A3*. That is a two-protein test. It does not consult the rest of the proteome, and it does not consult other isoforms of either parent — which is precisely how four peptides that occur in a normal *NR4A3* isoform passed it.

Step 6 is the whole of the defect: the filter is correct for what it tests and the set it tests against is too small by everything except two canonical sequences. Figure 1 shows the seam this produces, and the out-of-frame case beside it. Table 4 gives every pair the window produces with the grade the screen assigned it, so which 22 drop out is checkable rather than counted.

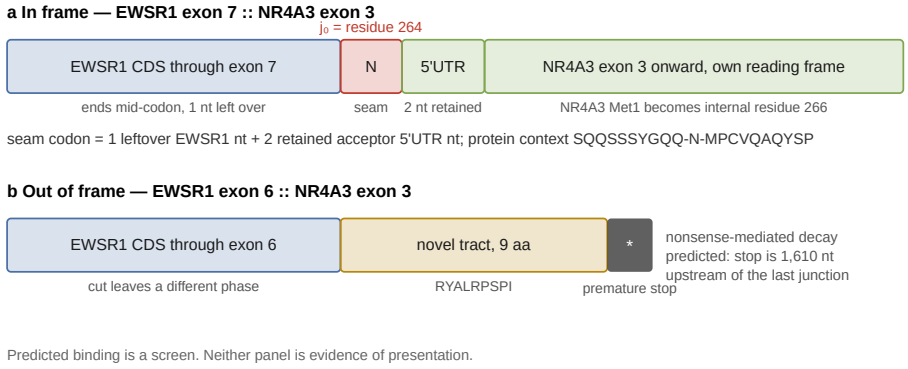


**Figure 1. The seam codon is built from leftover donor nucleotides and the acceptor exon's retained 5' untranslated region, and its identity decides whether the isoform collision occurs.** (a) The *EWSR1* exon 7 to *NR4A3* exon 3 junction. Donor coding sequence ends mid-codon with one nucleotide left over; two retained acceptor 5' untranslated nucleotides complete it, giving the seam residue at $j_0 = 264$, after which *NR4A3* resumes in its own frame with its methionine 1 as internal residue 266. Four of the five in-frame junctions place aspartate at this position rather than asparagine, and that is the difference Section B5 is about. (b) An out-of-frame junction reads the same acceptor exon in a shifted register: 9 novel residues, then a premature stop 1,610 nucleotides upstream of the chimera's last exon-exon junction, which is the canonical nonsense-mediated-decay configuration. Predicted binding is a screen; neither panel is evidence of presentation.

**Table 4. Every candidate exon pair, and why each is or is not carried forward.** All 27 pairs the declared window produces, graded by the screen of §2.2: 5 in frame, 4 out of frame, 9 acceptor exon is non-coding, 9 no seam produced at this pair. *Donor phase* is the number of coding nucleotides of the donor's last codon left over at the cut, so phase 0 is a codon-aligned break and phases 1 and 2 build the seam codon from retained acceptor nucleotides. An *In frame* cell reading n/a is a pair with no chimeric reading frame for the question to be about, not a pair that failed it. These are a combinatorial window of declared exon pairs, not observed breakpoints, and the window is an assumption of this work rather than a finding of it (§2.2).

**Table 4. Every candidate exon pair, and why each is or is not carried forward**

| *EWSR1* exon | *NR4A3* exon | Donor phase | In frame | Grade |
|---|---|---|---|---|
| 7 | 3 | 1 | yes | in frame; peptides emitted |
| 9 | 3 | 1 | yes | in frame; peptides emitted |
| 10 | 3 | 1 | yes | in frame; peptides emitted |
| 12 | 3 | 1 | yes | in frame; peptides emitted |
| 13 | 3 | 1 | yes | in frame; peptides emitted |
| 6 | 3 | 2 | no | out of frame; read-through tract, then a premature stop |
| 8 | 3 | 2 | no | out of frame; read-through tract, then a premature stop |
| 11 | 3 | 0 | no | out of frame; read-through tract, then a premature stop |
| 14 | 3 | 2 | no | out of frame; read-through tract, then a premature stop |
| 6 | 2 | 2 | n/a | acceptor exon is non-coding; no chimeric ORF |
| 7 | 2 | 1 | n/a | acceptor exon is non-coding; no chimeric ORF |
| 8 | 2 | 2 | n/a | acceptor exon is non-coding; no chimeric ORF |
| 9 | 2 | 1 | n/a | acceptor exon is non-coding; no chimeric ORF |
| 10 | 2 | 1 | n/a | acceptor exon is non-coding; no chimeric ORF |
| 11 | 2 | 0 | n/a | acceptor exon is non-coding; no chimeric ORF |
| 12 | 2 | 1 | n/a | acceptor exon is non-coding; no chimeric ORF |
| 13 | 2 | 1 | n/a | acceptor exon is non-coding; no chimeric ORF |
| 14 | 2 | 2 | n/a | acceptor exon is non-coding; no chimeric ORF |
| 6 | 4 | 2 | n/a | no seam produced at this pair |
| 7 | 4 | 1 | n/a | no seam produced at this pair |
| 8 | 4 | 2 | n/a | no seam produced at this pair |
| 9 | 4 | 1 | n/a | no seam produced at this pair |
| 10 | 4 | 1 | n/a | no seam produced at this pair |
| 11 | 4 | 0 | n/a | no seam produced at this pair |
| 12 | 4 | 1 | n/a | no seam produced at this pair |
| 13 | 4 | 1 | n/a | no seam produced at this pair |
| 14 | 4 | 2 | n/a | no seam produced at this pair |

That ten-allele panel is the instrument behind every binder figure here, and a wider one changes them: a 34-allele screen of the same peptides at the same threshold returns five strong peptide-allele calls rather than four, because NMPCVQAQY is also strong on HLA-A*30:02. Section 2.3 reports both panels.

**The out-of-frame junctions.** Twenty-two of the 27 pairs carry no peptides above, and four of those are out of frame — *EWSR1* exons 6, 8, 11 and 14 joined to *NR4A3* exon 3. A frameshifted junction reads the acceptor exon in a novel register, so every residue after the seam is non-self until a stop codon, and in other tumours that class is the richest antigen source available rather than the poorest. Here it is not. The four read-through tracts are 9, 9, 31 and 9 residues long before a premature stop, yielding 97 distinct 8- to 11-mers in total; and three of the four — those from *EWSR1* exons 6, 8 and 14 — converge on the identical eight-residue core YALRPSPI, differing only in the seam residue, because a frameshift into the same acceptor exon reads the same nucleotides in the same shifted register. The antigen supply from this class is therefore one short peptide and one 31-residue tract, not four independent ones. All four premature stops lie between 1,542 and 1,610 nucleotides upstream of the chimera's last exon-exon junction, which is the canonical configuration for nonsense-mediated decay by the 50-nucleotide rule — a positional criterion computed from the transcript model, not a decay measurement, and reported as such.

Screened against the same 34-allele panel at the same cut, those 97 peptides return 10 strong peptide-allele calls across 8 alleles, all of them from the single 31-residue tract, and the three 9-residue tracts return none. Those calls are tighter than anything the in-frame junctions produce. The strongest is LPLQVPVM on HLA-B*51:01 at a presentation percentile of 0.1039, against a best in-frame call of 0.3736; four of the ten fall below 0.2755. Section 2.3 reports that a conservative 0.2 cut leaves the in-frame screen with no presenting allele at all — and the peptides that survive that cut anywhere in this work come exclusively from a junction that cannot produce the driver. The antigen-poor part of this locus is the part that makes the disease. Two things bound this paragraph and are stated rather than implied: the 27 pairs are a combinatorial window of nine donor exons against three acceptor exons and not a set of observed breakpoints, and a frameshifted *EWSR1*::*NR4A3* cannot encode the chimeric transcription factor that defines the disease. So these peptides are not offered as EMC targets. What the screen establishes is the size of the antigen supply a frameshift at this locus would provide if one were ever observed, which is a property of the locus and answers the question rather than deferring it.

There is no pan-EMC epitope: the most widely shared candidate appears in 4 of the 5 junctions and is a weak binder, three of the five junctions return no strong binder at all, and every strong binder is specific to its breakpoint. One of the 11 predicted binders, DMPCVQAQY on HLA-B*35:01, is withdrawn on the proteome search reported under B5 below, leaving 10; all 4 strong binders survive that search.

### 2.3 Population coverage

Coverage here means the union carrier frequency of the presenting alleles. Writing $A$ for the set of alleles that present at least one junction peptide at the acceptance threshold, and $f_a$ for the pooled frequency of allele $a$ over Allele Frequency Net Database records [1]:

$$C(A) = 1 - \prod_{a \in A} (1 - f_a)^2$$

Each factor $(1 - f_a)^2$ is the probability that an individual carries neither copy of allele $a$ under Hardy-Weinberg, and the product treats loci as independent. Every coverage figure in the running text uses $f_a$ pooled across all AFND populations; Table 1 applies the same formula with $f_a$ restricted to one sub-region's records, which is the only difference between its cells and the global row; $C$ is therefore the fraction carrying at least one presenting allele. Because $A$ is itself a function of the acceptance threshold $t$ — an allele enters $A$ at the percentile of its best junction peptide — coverage is a function of $t$, and it is a right-continuous step function whose jumps are exactly the distinct peptide-allele percentiles. Section 2.3 computes it. Three properties of that quantity govern how the figures below should be read: the panel, the threshold and the predictor. All three are properties of the screen rather than of the tumour. A fourth, below them, is a property of the population the figure averages over, and a fifth is the uncertainty the pooled figure carries and does not print.

**It moves with the panel.** On the ten-allele screen the *EWSR1* exon 7 junction is presented on HLA-B*15:01 alone and covers 8.5%; on 34 alleles the same lead peptide is also strong on HLA-A*30:02 and the junction covers 12.3%. Pooling every in-frame junction gives 27.4% on ten alleles and 30.4% on 34. The four figures are one computation at two panel widths and two junction scopes, not four findings, and the 27.4% and 30.4% pair differs by exactly one allele. Extending the panel further can only raise them, so none of them is a ceiling and this paper does not call any of them one.

**It moves with the threshold further than it moves with anything else, and the whole function is computed here rather than sampled at three points.** Coverage against the acceptance threshold is a step function whose every step is a single peptide-allele call. Below the conventional cut it has four. HLA-B*15:01 enters at presentation percentile 0.3736 on NMPCVQAQY, taking coverage from zero to 8.5%; HLA-A*30:02 at 0.4033 on the same peptide, reaching 12.3%; HLA-A*01:01 at 0.4061 on RGDMPCVQAQY, reaching 23.2%; HLA-B*07:02 at 0.4580 on MPPPLRGDM, reaching 30.4%. Four steps and five strong calls, because the fifth — QQNMPCVQAQY on HLA-B*15:01 at 0.4986 — adds a second peptide to an allele already presenting and moves nothing. Sampling at round numbers, as earlier versions of this paper did — to 0.45 leaves three alleles and 23.2%, to 0.40 leaves one and 8.5%, to 0.37 leaves none and 0.0% — understates how compressed the function is. Every step falls inside a window 0.0844 units wide; the function is then flat from 0.4580 all the way to the cut; and it reaches zero 0.1264 below it. The headline figure is flat at the threshold not because the threshold is well placed but because it sits past the last step, with a cliff a tenth of a percentile unit beneath.

**Above the cut it does not plateau either, and that is the more useful half.** Extending the same screen to a percentile of 5 adds twenty-four further alleles in twenty-four further steps, none of them large and none of them absent: 8 presenting alleles and 47.8% at a cut of 1.0, 16 and 72.6% at 2.0, 28 and 90.2% at 5.0. The paper's own weak-binder cut is 2.0. So over the span of cuts this field routinely writes down — 0.2 as a conservative tier, 0.5 as strong, 2.0 as weak — predicted coverage of this junction runs from 0% to 72.6%, and the reported 30.4% is one arbitrary point inside that range rather than an estimate with a tolerance.

**And it moves with the predictor, which is a third axis and was untested until now.** Every figure above comes from one software suite, so the sensitivity reported here could be a property of that suite rather than of the junction. The same 174 peptides and the same 34 alleles were therefore re-screened with MHCnuggets [12], an independently trained class I predictor whose native output is a predicted IC50 rather than a presentation percentile. The two scales are not rescaled onto one axis — that would manufacture the agreement it reports — so each predictor is swept over its own conventional cut and the shapes are compared. Every allele on the panel returned a score, so no cell of the comparison is an absent reading. Two things follow. The threshold sensitivity is not an artifact of one suite: the second predictor's curve is equally steep, running from no presenting allele at 10 nM through 8 alleles and 45.2% at its conventional 500 nM to 32 alleles and 94.0% at 10,000 nM. But the two predictors do not agree on which alleles present: at each one's own conventional cut they share three — HLA-A*30:02, HLA-B*07:02 and HLA-B*15:01 — while HLA-A*01:01 is MHCflurry's alone and five more are MHCnuggets' alone. So the choice of predictor moves the headline figure from 30.4% to 45.2% on identical inputs, and the intersection of the two is three alleles of a union of nine. The lead allele HLA-B*15:01 is in the intersection, which is the one place in this paper where two independent instruments say the same thing.

**None of this is an argument for a looser cut.** A threshold chosen because it raises coverage would be the same defect this paper exists to name, arriving from the other side; the 28 alleles at a percentile of 5 are not a better answer than the 4 at 0.5, they are a demonstration that the question "what fraction of patients could this reach?" has no answer until somebody defends a cut, and nobody has. Figure 2 plots the function on a log axis, which is the axis its shape needs: the four steps span 0.0844 percentile units and a linear axis collapses them into the left margin. The whole curve is deposited as `coverage-threshold-curve.json`, so a reader who would set the cut elsewhere can read off what it gives instead of taking three points on trust.

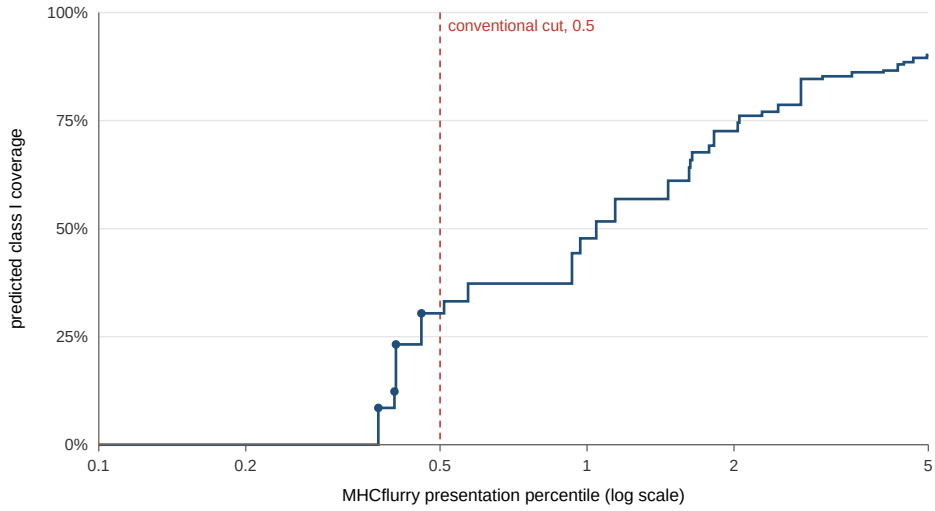


**Figure 2. Predicted class I coverage against the acceptance threshold, log axis.** The step function over the 34-allele panel, each step one peptide-allele call, drawn to the largest threshold the predictions can speak to. Filled circles mark the four steps below the conventional cut, which span 0.0844 percentile units; the dashed line is that cut, drawn as one annotated vertical rather than as the plot's endpoint. The axis is logarithmic because the function's shape occupies two decades and a linear axis collapses every step the argument rests on into the left margin. Coverage is the union carrier frequency of the presenting alleles on pooled AFND frequencies.

**It is an average over populations that do not resemble one another.** The pooled figure is a single number laid over a distribution whose ends are more than forty-fold apart, and a reader planning anything for a particular place needs the end they are standing at rather than the mean. Predicted coverage of any in-frame junction runs from 1.4% in Melanesia to 60.4% in Northern Europe, and of the public *EWSR1* exon 7 junction from 0.8% in Northern Africa to 16.1% in Northern Europe. So the statement that no panel screened here reaches half of patients is true of the pooled global frame and false in Northern Europe, where three alleles reach three fifths. This is not a limit of the screen — it is what HLA frequency is — but it means the pooled figure should not be quoted alone, and the table below is given so that it need not be.

**Table 1. Predicted class I coverage by UN M49 sub-region, ten-allele screen.** Coverage is the union carrier frequency of the presenting alleles over Allele Frequency Net Database records [1], computed within each sub-region rather than pooled. The pooled global figures are the last row. *Public junction* is the *EWSR1* exon 7 to *NR4A3* exon 3 junction, presented on HLA-B*15:01 alone; *any junction* pools every in-frame junction, presented on 3 alleles. No confidence interval is given, for the reason in §2.3; *n* is the largest per-allele sample the sub-region contributes and is what a reader should weigh a cell against. A dash is an allele absent from that sub-region's records, not a coverage of zero.

**Table 1. Predicted class I coverage by UN M49 sub-region, ten-allele screen**

| Sub-region | Public junction | Any junction | Presenting alleles | *n* |
|---|---|---|---|---|
| Northern Europe | 16.1% | 60.4% | HLA-A*01:01, HLA-B*07:02, HLA-B*15:01 | 1598 |
| Western Europe | 8.2% | 41.6% | HLA-A*01:01, HLA-B*07:02, HLA-B*15:01 | 4045 |
| Eastern Europe | 7.1% | 37.0% | HLA-A*01:01, HLA-B*07:02, HLA-B*15:01 | 2456 |
| Southern Europe | 5.9% | 32.1% | HLA-A*01:01, HLA-B*07:02, HLA-B*15:01 | 5982 |
| Northern America | 8.0% | 31.6% | HLA-A*01:01, HLA-B*07:02, HLA-B*15:01 | 5478 |
| Western Asia | 2.6% | 27.8% | HLA-A*01:01, HLA-B*07:02, HLA-B*15:01 | 2543 |
| Southern Asia | 4.0% | 25.1% | HLA-A*01:01, HLA-B*07:02, HLA-B*15:01 | 3580 |
| Australia and New Zealand | 5.0% | 23.9% | HLA-A*01:01, HLA-B*07:02, HLA-B*15:01 | 702 |
| Northern Africa | 0.8% | 21.2% | HLA-A*01:01, HLA-B*07:02, HLA-B*15:01 | 1537 |
| Latin America and the Caribbean | 7.8% | 21.1% | HLA-A*01:01, HLA-B*07:02, HLA-B*15:01 | 17277 |
| Sub-Saharan Africa | 2.0% | 20.4% | HLA-A*01:01, HLA-B*07:02, HLA-B*15:01 | 3066 |
| Eastern Asia | 15.4% | 20.0% | HLA-A*01:01, HLA-B*07:02, HLA-B*15:01 | 11585 |
| South-eastern Asia | 3.7% | 14.0% | HLA-A*01:01, HLA-B*07:02, HLA-B*15:01 | 3556 |
| Polynesia | — | 1.9% | HLA-B*07:02 | 450 |
| Melanesia | 0.9% | 1.4% | HLA-B*07:02, HLA-B*15:01 | 1269 |
| Micronesia | — | — | — | 129 |
| **Pooled global** | **8.5%** | **27.4%** | HLA-A*01:01, HLA-B*07:02, HLA-B*15:01 | — |

**What the withdrawn interval was, and what stands in its place.** Earlier versions printed Wilson 95% intervals about half a percentage point wide on every figure above. They are withdrawn. They were computed by pooling every reference population into one binomial, and the same records show HLA-B*15:01 ranging from 0 to 0.40 in frequency between those populations, so the single-urn model the interval assumes is refuted by its own input — and the interval is an order of magnitude narrower than the threshold sensitivity above. The regional figures in Table 1 are point values for the same reason, and

the reason is stronger there: the sample behind a sub-region's cell is smaller and less homogeneous than the pooled one, not more.

Withdrawing a wrong interval and printing a bare point estimate is not an improvement on printing a wrong one, and this section previously stopped there. Three quantities are given instead. Each answers a question the Wilson interval was standing in for, none of them is a confidence interval, and all three are deposited in `coverage-uncertainty.json`.

*The same-locus approximation, and its direction.* The formula above multiplies $(1 - f_a)^2$ over alleles, taking HLA-B*07:02 and HLA-B*15:01 to be independent events when a diploid genome carries at most two alleles at HLA-B and the two compete for the same pair of slots. Grouping by locus is exact under the same Hardy-Weinberg assumption: the probability of carrying neither of a locus's presenting alleles is $(1 - \Sigma f_a)^2$, not $\prod (1 - f_a)^2$. Since $(1 - f_1 - f_2) < (1 - f_1)(1 - f_2)$, the published form is a lower bound and not an approximation of unknown sign. Recomputed exactly, the pooled ten-allele figure is 27.7% against the 27.4% printed above: 0.33 percentage points, in the direction that means no coverage figure in this paper is too high for this reason. The correction is small here only because the two same-locus frequencies are small; it grows with the panel, and a screen reporting many alleles at one locus under the published form would understate by considerably more.

*Dependence between loci, bounded rather than assumed away.* What remains is independence between HLA-A and HLA-B, which linkage disequilibrium violates and which nothing in this work measures. It does not have to be measured to be bounded. Pooled per-locus carrier frequencies are 12.4% at A and 17.5% at B, and for a union of events with those marginals the Fréchet bounds place coverage in [17.5%, 29.9%] under *any* dependence structure whatever, exactly and distribution-free, assuming nothing about haplotypes. Positive disequilibrium, presenting alleles clustering in the same people, pushes coverage toward the lower end; negative disequilibrium toward the upper. So the independence assumption is worth at most 12.4 percentage points here, and the reader who distrusts it can take the bound instead of the estimate.

*Hardy-Weinberg is the assumption underneath that one, and it is bounded the same way.* Reading a carrier frequency off an allele frequency through $(1 - p)^2$ assumes Hardy-Weinberg at the locus, and nothing here tests it: AFND publishes allele frequencies rather than genotype counts, so it cannot be tested from this source. It can be bounded from it. An allele frequency is a fraction of chromosomes, and every way of arranging those copies into people lies between two extremes — every copy paired in a homozygote, giving a carrier frequency equal to $p$, and no homozygotes at all, giving $\min(1, 2p)$. Hardy-Weinberg's $2p - p^2$ sits inside that interval for every $p$, near its top when $p$ is small. Dropping Hardy-Weinberg and between-locus independence together, coverage of the ten-allele screen lies in [9.2%, 31.1%], and the 27.7% above sits inside it. That interval assumes nothing about genotypes or haplotypes at all, and it is the widest statement this paper can make that is still certainly true.

*The spread that matters is between populations, and it is far wider than either.* Recomputing $C$ inside each AFND study population from that population's own frequencies, restricted to the 112 populations of at least 50 individuals that measured all three presenting alleles, gives a median of 24.5%, an interquartile range of 9.6% to 35.6%, and a full range of 0% to 66.0%. The pooled 27.4% sits near the median of a distribution whose middle half spans 26 percentage points and whose extremes are 0% and 66.0%. That is an order of magnitude wider than either reading above it, wider than the panel and predictor axes of this same figure, and narrower only than the threshold axis. Taken instead over every population with any measurement, scoring an unmeasured allele as absent, the median falls to 3.1% across 482 populations. That is a floor rather than an estimate, because absence from AFND is a reporting gap and not a measured zero. The distance between those two readings is a statement about the database, not about HLA, and it is given here so that the 112-population figure is not read as a survey of the world.

None of the three is the quantity a confidence interval would give, and none is offered as one. What is still absent is any interval on the *screen* — on whether the peptide-allele calls that define $A$ are correct. That is limit B2, which no rearrangement of population statistics can reach.

Because an individualised platform selects against the patient's own genotype rather than a public epitope, the relevant figure is the pooled-junction one. Selecting per patient moves it from 8.5% to 30.4% on the panels screened here, and removes neither the panel dependence nor the threshold dependence.

#### 2.4 Limits of the current evidence

Predicted binding is a screen. It is not evidence that any peptide is presented on an EMC tumour cell, not a measure of the density at which it would be presented, and not evidence of T-cell recognition. No EMC immunopeptidomic dataset is known to the author. Sequence-level novelty against the proteome has now been tested and is reported under B5, which excludes one failure mode and leaves presentation, immunogenicity and cross-reactivity untouched. The machinery a presented peptide would require has since been read at the transcript level in EMC tumour tissue and is not lost (B8), which establishes a precondition rather than a presentation. These gaps are enumerated as limits B2 and B3 rather than treated as resolved.

## 3. The standing state, limit by limit

Ten limits are enumerated, and the second column is the one that matters. Disease-bounded limits are properties of this tumour and are not expected to move; instrument-bounded limits move when methods move; access-bounded limits are properties of this programme's circumstances rather than of anyone's knowledge.

The cost-to-move column's units are whose permission and whose material a row requires, rather than money or months, neither of which this programme can estimate. No time estimate is offered for any row, for the reason Section 3.1 gives.

| ID | Limit | Bounded by | Best available answer today | Cost to move | What it |
|---|---|---|---|---|---|
| B1 | Predicted class I coverage is low and panel-dependent | instrument, partly disease | 8.5% / 12.3% public junction, 27.4% / 30.4% pooled, on 10 / 34 alleles; 0.0% at a 0.37 cut | **computational** — needs nobody's permission | Wide defer meas prese |
| B2 | Presentation predicted, never measured | instrument and access | Presentation percentiles on 174 peptides, from the predictor named in Section 2.2 | **tissue + a proteomics facility** — the binding constraint on this route | Immu on El patie |
| B3 | Self-adjacency and central tolerance | disease | Lead peptide is 1 substitution (position 1, not an anchor) from DMPCVQAQY in an *NR4A3* isoform, against a chance expectation of 0.02; 0 of 11 binders has an anchor-only near-self neighbour | **a T-cell assay on matched donors** — needs material and a laboratory | T-cel assay speci HLA |
| B4 | One strong CD4 epitope, on one allele of 23 | instrument | 44 binders, 1 strong (SYGQQNMPCVQAQYS on DRB1*14:01, 66.1 nM) over 14 of 23 alleles; class II coverage 6.5%, combined CD8 and CD4 1.8% | **computational**, and now largely spent | A clas meas prese |
| B5 | Seam-proximal peptides of four junctions occur in an *NR4A3* isoform | instrument, one failure mode resolved | 170 of 174 novel proteome-wide; one binder withdrawn | **computational** — needs nobody's permission | Sequ answ react the fi beco awar |
| B6 | Immunologically cold microenvironment | disease, addressable in combination | Quieter than comparator sarcomas on both archival cohorts; two individual-patient reports point the other way | **spatial profiling on tissue** | A vac antig checl supp |
| B7 | Physical exclusion by the myxoid matrix | disease | Inferred from histology; the matrix pathway is now read in EMC and does not carry the analogy | **spatial profiling on tissue** | Vasc norm matri agen |
| B8 | EMC immune profiling is thin, not absent | access | Two archival cohorts read here; one published 12-tumour series, its spatial arm 2 specimens | **a larger series with spatial resolution** | Per-t profil that e infiltra class |
| B9 | Manufacturing economics at this incidence | access | Five enumerable in-frame junctions, not per-patient discovery | **not this programme's to move** | A pla mast vehic |
| B10 | Trial design below the randomisation threshold | disease | A 24-patient histology cohort has been run across nine centres | **a trial design decision, not a measurement** | Adap histol desig defer |

#### 3.1 The limit worth watching

The single most consequential of the movable limits is B2. Direct mass-spectrometry evidence that a fusion-junction peptide is presented at measurable abundance on a common allele, in any fusion-driven tumour, would convert the central question of this route from a prediction into an empirical one. The coverage instrument used here is disclosed as failing for exactly that reason: it computes an eligibility fraction for an epitope whose presentation is unestablished. No date is offered for that arrival, or for any other in this section. An earlier version of this paper carried a table of optimistic, expected and conservative years for four such capabilities; it is withdrawn, because a forecast a reader cannot check does not become checkable by being tabulated, and Section 6 does the work it was there for by saying what each arrival would and would not look like.

#### B1. Low predicted class I coverage, dependent on the screen

**Proposition.** On the panels and thresholds screened here, fewer than a third of patients carry an allele predicted to present any junction peptide — and that fraction is set by the screen at least as much as by the tumour.

**Evidence.** The 34-allele screen finds 4 presenting alleles and 30.4%; the ten-allele screen finds 3 and 27.4%. Both rest on five strong peptide-allele calls, one per presenting allele, and Section 2.3 gives their threshold sensitivity.

**A high-confidence tier, and it is empty.** The natural response to an undefended cut is to report a conservative tier beside it, and the conventional conservative cut for class I presentation is 0.2. At 0.2 this screen returns no presenting allele and no coverage at all: the least permissive call it makes is 0.3736, so every figure in this paper lives strictly between the conventional cut and a cut half its size. That is reported rather than suppressed, and it is the sharpest available statement of what these predictions are worth: they are not a set of confident calls with a permissive tail, they are entirely a permissive tail.

**What would clear or move it.** Three things, in ascending cost. Extending the panels beyond 10 and 34 class I alleles to the full validated set is a computational task and would raise the denominator. The acceptance threshold is a convention rather than a result — nothing here defends 0.5 against 0.4 or 0.6 — and settling what the cut should be for junction peptides would do more to this figure than any panel extension. Calibrating that cut against a benchmark of experimentally validated neoepitopes is the form that settling would take, and no such benchmark restricted to fusion-junction peptides is known to the

author; calibrating on point-mutation neoantigens instead would import an assumption about junction peptides that is the very thing in question. Measured immunopeptidomics (B2) could promote peptides the predictor ranks weakly or remove ones it ranks strongly, in a direction not knowable in advance.

**Residual.** Some fraction of patients will have no presented junction peptide and no second antigen to substitute. That fraction is a real and permanent exclusion from this approach, and it should be stated in any protocol rather than discovered at screening.

### B2. Predicted rather than measured presentation

**Proposition.** No junction peptide has been shown to be presented on the surface of an EMC cell.

**Evidence.** All binding figures in Section 2 come from the sequence-based predictor named there [2]. No EMC immunopeptidomic dataset is known to the author, and the repository contains none.

**What would clear it.** Mass-spectrometry immunopeptidomics on EMC tumour tissue or on one of the patient-derived EMC cell lines that have been established and characterised [6,14]. A single positive identification of a junction-spanning peptide in an EMC eluate would convert the central premise of this route from predicted to observed.

**What a negative would have to specify to mean anything, and this paper cannot supply it.** The sentence "a negative result would be close to decisive" appeared in an earlier version of this section and was doing no work, because a null eluate bounds nothing until three quantities are fixed in advance: a stated limit of detection in peptide copies per cell, so that "not presented" is separated from "below the instrument"; a stated number of independent specimens and their HLA types, since the lead peptide is presented on one allele and a specimen not carrying it cannot test the claim; and a positive control peptide of known abundance eluted in the same run, without which a null is indistinguishable from a failed elution. This paper names those three requirements and does not meet any of them — naming them is not the same as having them. Their absence is why B2 is graded as bounded by access rather than by the disease, and it is what makes this the highest-value single experiment in the ledger.

**Cost and owner.** Requires tissue and a proteomics facility. It is not computational and cannot be performed by this programme.

### B3. Self-adjacency and central tolerance

**Proposition.** The junction peptide is mostly self sequence with one or two novel residues at the seam, so the T-cell repertoire capable of recognising it may have been deleted or anergised.

**Evidence.** At the *EWSR1* exon 7 junction the seam carries a single novel codon formed from one leftover *EWSR1* nucleotide and two retained acceptor 5' untranslated nucleotides, after which *NR4A3* methionine 1 follows as an internal residue. The remainder of each peptide is parental sequence. Whether one altered residue is sufficient to break tolerance is not answerable by binding prediction.

**The near-self search, which this section previously listed as unbuilt.** Two computational filters were specified for this route and never built: a distance-to-self filter, and an analysis of whether the differences from self fall at anchor positions, which affect binding, or at positions contacting the T-cell receptor, which affect recognition. Both have now been run. Every one of the 11 predicted binders was searched against the reviewed human proteome including isoforms at a tolerance of one or two substitutions, with a per-peptide chance baseline from 50 residue shuffles of the same peptide re-searched identically.

Eight of the 11 have at least one human peptide within two substitutions, and every such neighbour lies in *EWSR1* itself, in an *NR4A3* isoform, or in the paralogue *NR4A1* — which is what a junction assembled from two self proteins should give, and is a check on the search as much as a result. Three findings matter. **The lead peptide is one substitution from self.** NMPCVQAQY differs at position 1 from DMPCVQAQY, which occurs in *NR4A3* isoform 3 — the same isoform Section B5 is about — and at positions 1 and 5 from EMPCIQAQY in an *NR4A1* isoform. The chance expectation is stated as a model rather than as a number: for a query $q$, draw $N = 50$ uniform random permutations of its own residues, search each against the same proteome at the same tolerance, and let $\mu(q)$ be the mean neighbour count over those draws. This holds length and amino-acid composition fixed and destroys only the order, so $\mu$ is the neighbour count attributable to $q$ being a string of those residues rather than to $q$ being that particular sequence. For NMPCVQAQY, $\mu = 0.02$ against an observed 2. **No p-value is quoted and none should be:** 50 draws bound a rate near zero loosely, the draws are not independent of the proteome's own composition, and the comparison this paper needs is an order of magnitude, not a tail probability. **Neither difference is at an anchor.** Position 1 and position 5 face outward or into the groove's middle rather than serving as the primary anchors at position 2 and the C-terminus, so a T cell raised against the neoepitope reads a surface that differs from the self peptide's at the positions it actually contacts — which is the configuration in which cross-recognition is most plausible and in which central tolerance is most likely to have acted. **And no binder in the screen has an anchor-only neighbour.** Zero of the 11 has a near-self peptide whose differences are confined to anchor positions, which would have been the worst case: an identical TCR-facing surface distinguished only by residues the T cell cannot see. MPPPLRGDM's five neighbours are one *EWSR1* peptide, MPPPLRGGP, counted once per isoform, differing at positions 8 and 9 — and against a chance expectation of 1.34 for that peptide, five is not distinguishable from chance.

**Two caveats carried with the result.** The anchor assignment is the general class I rule of position 2 and the C-terminus applied uniformly, not an allele-specific motif; HLA-A*01:01 in particular reads position 3 as a primary anchor, and the artifact marks every row so a reader can apply that. And sequence distance is not receptor distance: peptides three substitutions apart can present near-identical surfaces and peptides one apart can present different ones. This search excludes one more failure mode — a close self peptide nobody had looked for — and leaves the question of whether a repertoire exists exactly where it was.

**What would clear or narrow it.** Only a T-cell reactivity assay against the specific peptide-HLA complex answers the question. The search above changes what such an assay would be testing: the comparator is no longer a hypothetical self peptide but a named one, DMPCVQAQY, at a known single-residue distance, in a protein the tumour also expresses.

**Comparator.** This limit is where EMC differs most sharply from the melanoma setting, in which selected neoantigens frequently arise from point mutations in a repertoire that has not been tolerised against them and in a tumour already under immune pressure.

### B4. One strong CD4 epitope, on one allele of twenty-three

**Proposition.** An effective vaccine generally requires CD4 helper epitopes. The corrected junction supplies one predicted strong class II binder on a 23-allele panel, restricted to an allele carried by about one person in fifteen. That is a statement about the screen, and it is reported here as one.

**History, and why this section changed.** This arm was previously withheld because it sat on the superseded coordinate system, on a seam disjoint from the class I set; the builder was moved to the transcript model and the arm regenerated, so the two arms are certified to sit on the same seam. It was then reported as a negative on three DRB1 alleles — and Section 6.2 of that version named a wider DR, DP and DQ panel as the single most likely source of a change to it. That panel has now been run. The prediction was correct and the negative did not survive it.

**Result.** At the corrected *EWSR1* exon 7 junction, 15 candidate 15-mers were screened with MHCnuggets [12] against 23 class II alleles — 15 DRB1, DRB3*01:01, DRB4*01:01, DRB5*01:01, two DPB1 and three DQB1 — calling a peptide a binder below 1000 nM and strong below 100 nM. Every declared allele returned a score; none was unscreenable, so no cell of this panel is an absent reading. 44 peptide-allele pairs bind, spread over 14 of the 23 alleles, and one is strong: SYGQQNMPCVQAQYS on DRB1*14:01 at 66.1 nM. The three-allele subset the previous version reported reproduces exactly within the wider run — YSQQSSSYGQQNMPC at 261.6 nM and QYSQQSSSYGQQNMP at 438.9 nM, both on DRB1*07:01 — so the widening is additive and the earlier negative was a property of the panel, not of a different computation.

**What the positive does and does not bound, and it is less than it looks.** The 15 candidates are single-residue-offset windows over one seam sharing 14 of 15 residues, so the number of independent peptides tested is nearer one than fifteen, and the 44 binders are overwhelmingly the same short stretch of sequence seen on different alleles. One strong call on 23 alleles is a per-allele rate of 1 in 23, and the allele it lands on is not a common one. Class II prediction is substantially less accurate than class I and is not treated here as its equal: class II peptides are not length-restricted by a closed groove, the binding register is not fixed, and predictors for it are trained on far less measured data. The single strong call should be read as a weaker statement than any single class I call in this paper, and no class II prediction here has been calibrated against anything measured. What the result does establish is narrow and worth having: the junction is not devoid of predicted helper epitope, which is what three alleles could not distinguish from a junction that is.

**What would clear or move it.** Only measured class II presentation settles it; a wider panel now has much less headroom to move it than it did. A construct can also supply help from a heterologous source rather than from the junction itself, which is standard practice in peptide vaccine design and would make this limit a design constraint rather than a blocking one.

**Consequence for the construct, and for the combined figure.** Two things follow directly, and both changed with the panel. The candidate construct regenerated at the corrected junction now carries the CD4 epitope as well as the two strong class I epitopes NMPCVQAQY and QQNMPCVQAQY on HLA-B*15:01, and its minimal synthetic long peptide is 15 residues carrying both arms, where the class I-only version was 11. And the combined CD8 and CD4 coverage figure can be computed at all for the first time: class II coverage is 6.5% on DRB1*14:01 alone, and the fraction of patients carrying at least one presenting allele of each class is 1.8%, the product of the two arms' coverages:

$$\mathbf{C_{I\wedge II} = C_I \times C_{II} = 0.2737 \times 0.0649 = 0.0178}$$

The class I term is the ten-allele screen's 27.4%, not the 34-allele screen's 30.4% quoted two paragraphs above: the instrument builds its class I allele set from the junction screen rather than from the broad-panel scan. On the 34-allele set the same product gives 2.0%. The product assumes independence between the class I and class II loci, the same assumption the within-class formula makes and no weaker. HLA-A, HLA-B and DRB1 sit on chromosome 6 in linkage disequilibrium, so it is an approximation whose direction is not known without haplotype frequencies. The magnitude is bounded even though the direction is not, and by the same argument §2.3 applies to the within-class formula. This figure is an intersection rather than a union, so the bounds invert: under any dependence structure whatever it lies in [0.0%, 6.5%], and the upper end is the useful half. A construct that needs a presenting allele of each class can never reach a larger fraction of patients than its scarcer arm alone, which here is the class II arm at 6.5%. That ceiling holds without any haplotype data, and it is 3.6 times the product rather than an order of magnitude above it. It is a carrier-frequency product and nothing more: it says what fraction of patients carry a presenting allele of each class, not what a class II response contributes. Class II help for a CD8 response is cognate rather than additive, and no model of that is offered here. That figure is smaller than either arm alone and is the eligibility fraction for a construct that needs both — and, like every coverage number in this paper, it is a point on the curves of Section 2.3 rather than an estimate with a tolerance.

**Status.** Positive, on one allele of 23, with the previous negative explained as a panel artifact rather than overturned by a different method.

### B5. Four peptides occur in an *NR4A3* isoform

**Proposition, as originally stated.** The junction peptides had been tested for novelty against two proteins rather than against the proteome, so their absence from normal human proteins was not established.

**Result.** All 174 distinct junction peptides were searched by exact substring against the UniProt reviewed human proteome, isoform sequences included. 170 are absent from every reviewed human protein. All 4 strong binders survive, including the *EWSR1* exon 7 lead NMPCVQAQY. The limit is largely cleared, and the sequence-level novelty premise of this route holds for the great majority of the peptide set.

**The four that do not, and which junctions they belong to.** DMPCVQAQ, DMPCVQAQY, DMPCVQAQYS and DMPCVQAQYSP all occur in Q92570-3, an isoform of *NR4A3* itself. One of them, DMPCVQAQY, is a predicted binder on HLA-B*35:01 at 369.1 nM. Those four peptides are not tumour-exclusive, and DMPCVQAQY is withdrawn as a candidate.

**The pattern is a mechanism, and it survives the transcript model.** A seam defined by exon boundaries is defined by whichever transcript declares them, and every figure in Section 2 comes from one canonical transcript per gene. The junction was rebuilt across all 99 protein-coding *EWSR1* transcripts against all 4 of *NR4A3*, at each of the five in-frame donor exons: 1,980 pairs, of which 970 emit an in-frame seam. The seam residue is not stable across them, taking nine distinct values — aspartate in 502 pairs, glycine in 251, asparagine in 118, serine in 59, no seam residue at all in 31, and four other residues in the remaining 9.

**The collision appears in exactly the 502 aspartate-seam pairs and in none of the other 468.** It is therefore not a property of the canonical annotation, and not a single case study: it is a property of the aspartate seam, reproduced independently in 502 transcript pairs, and its absence is equally determined. That is the mechanism this section had been asserting from one pair.

The pattern is not random across the junction set. All four collided peptides belong to the same four junctions, *EWSR1* exons 9, 10, 12 and 13, which are the four whose seam codon is aspartate on the canonical transcript; the *EWSR1* exon 7 junction has an asparagine seam codon and none of its peptides

collides. So four of the five in-frame junctions share a seam whose most seam-proximal peptides reproduce a normal *NR4A3* isoform sequence, and the one junction that is clean is the commonly reported public one that carries the lead binder. Within the affected junctions the collision is confined to peptides that begin at the seam residue, which are the peptides with the least donor content; those extending further into *EWSR1*, including the strong binder RGDMPCVQAQY, remain novel. The consequence for design is specific rather than general: for the aspartate-seam junctions, a construct should not rely on the seam-proximal window.

**The methodological finding.** The upstream novelty filter compares each candidate against the canonical parent proteins only, so an isoform that carries the seam sequence passes it unseen. That is a defect in the filter rather than in these particular junctions, and it will recur for any breakpoint whose seam residue reconstructs an isoform boundary. The proteome search reports this condition explicitly rather than silently discarding the hits.

**The repair, stated as a test rather than as an aspiration.** Saying the filter should be made isoform-aware names the defect without saying what to compute, so here is what to compute. Everything 3' of the seam residue is the acceptor's own sequence, retained whole, so a seam-proximal peptide is the seam residue followed by an acceptor stem. It reproduces a normal protein exactly when some isoform of the acceptor gene already places that same residue immediately 5' of that same stem, which is what an alternative first exon does. The residues for which that holds are a property of the acceptor gene alone; call them its collision alphabet. The test is then one character:

> a junction whose seam residue lies in the acceptor's collision alphabet has a seam-proximal window that is not tumour-exclusive.

The alphabet is computed once per acceptor from its isoform set and answers the question for every donor and every breakpoint without a further search, where the proteome scan it anticipates costs one pass over the proteome per peptide. Here the alphabet is a single residue, aspartate, contributed by Q92570-3.

**It is exact on every transcript pair this locus can produce.** Applied to the 970 in-frame pairs of the 1,980 above, the one-residue test agrees with the proteome search on all 970: 502 predicted collisions, all observed; 468 predicted clean, none colliding; no disagreement in either direction. That the two ends match is the point. A rule that flagged the aspartate seam would be unremarkable if it also flagged the glycine seam's 251 pairs or the asparagine seam's 118, and it flags neither.

**And it does not replace the search, which is the failure mode from the other side.** The test anticipates one condition: a seam-proximal peptide reconstructing an acceptor isoform boundary. It is silent about a peptide colliding with an unrelated protein, and silent about peptides extending far enough into the donor to become novel again — this locus's own strong binder RGDMPCVQAQY carries the same stem, begins three residues 5' of the seam, and survives. A filter that ran this instead of the proteome search would reintroduce the defect it was built to catch. It is a pre-screen that says where the search will find something, and it costs one comparison.

**What a clean result does not license.** A peptide absent from every reviewed human protein is not thereby safe. A T-cell receptor engages a peptide-MHC surface rather than a sequence, so a peptide differing from a self peptide at a position that does not contact the receptor can still be cross-recognised. Unreviewed sequences were searched separately, over 127,090 entries: of the 170 peptides absent from every reviewed protein, 12 occur in at least one unreviewed entry, and none of those 12 is a predicted binder. A hit among predicted-and-unreviewed entries is not evidence that a normal protein carries the peptide, and a miss there is not evidence of absence, so this withdraws no peptide and confirms none; what it establishes is where the sequence-novelty premise is weakest, and that is not where the candidates are. This test excludes one specific failure mode and leaves the others standing.

### B6. Immunologically cold microenvironment

**Proposition.** EMC has a low mutational burden and a sparse infiltrate, so there is little pre-existing antigen-specific response for an intervention to amplify. The mutational-burden half is not disputed here. The infiltrate half was an assumption when this section was first written and is now a measurement, reported at B8 and not restated here.

**What that measurement does to this proposition, and what it does not.** It supports the direction B6 asserts and does not license the word cold, because the comparator is other sarcomas rather than an immune-normal reference, and because a bulk average over whole tissue cannot tell a tumour with no T cells from one whose T cells are held outside it. That second reading is B7's, the two are not the same claim, and B7 states why the evidence now available does not choose between them.

**Evidence, and an observation about this programme's own grading.** This proposition is why this programme set most immune-modulating classes aside for this disease in its route ledger, a committed record of forty candidate modalities each carrying a verdict and a reason. Those reasons are quoted below from that ledger: they are this author's own prior notes, not positions taken in the literature. Innate agonists of the STING, TLR and RIG-I classes were excluded there because such agonism "supplies the danger signal and the priming context; it does not supply antigens, and a genome as quiet as EMC's is short of antigens rather than short of priming". In-situ vaccination was excluded because it "releases and adjuvants the antigens the tumour already has", so "releasing more of nothing does not help". Checkpoint inhibitors beyond PD-1, costimulatory agonists, adenosine-axis inhibitors and regulatory T-cell depletion were each excluded for acting on a response presumed absent.

Every one of those is an argument about antigen supply, and a vaccine is an antigen supply, so none of them argues against the class in a configuration where antigen is supplied exogenously. That is why the combination in Section 4 was never graded here as a unit.

The symmetry is not exact, and an earlier version of this paper asserted that it was. The vaccine was not parked for want of priming: its ledger entry is on the board rather than excluded, and its stated reason for being parked is immunogenicity, "which its own record states is not a computational question". The standing blocker against it names two things, that EMC is antigen-cold and that the fusion junction is a weak peptide-HLA, and the second is an antigen-side objection no priming-directed partner answers. The correct statement is the weaker one: a partner supplies what several exclusions called missing, and does not supply what the vaccine's own blocker calls missing.

**What would clear it.** Clinical evaluation of a vaccine together with an agent that supplies priming or releases inhibition. Section 4 sets out the specific backbone for which EMC evidence already exists.

### B7. Physical exclusion by the myxoid matrix

**Proposition.** EMC is immune-excluded by a dense chondroitin-sulfate gel, which is a physical barrier rather than a signalling programme.

**Evidence.** The myxoid matrix is the disease's defining histological compartment. The oncofetal chondroitin sulfate pathway is the mechanism proposed for it here by analogy: the glycosaminoglycan biosynthesis genes of that pathway are differentially expressed and correlated with immune response in placenta and colorectal cancer [8], which is the tissue setting that study examined. That pathway has since been read in the same two EMC cohorts and the reading does not carry the analogy across: the linker and core-protein modules are higher in EMC than in the comparator sarcomas on both platforms, the backbone-polymerisation and 4-O-sulfotransferase modules disagree between platforms, the sulfate-donor module is lower on both, and three sulfotransferase modules fall below the read's own readability floor. Values have one home, `research/modalities/emc-expression-panels.json`, read 2. A transcript level is a capacity to build a glycan and not the glycan, so the inference from those tissues to this one remains the author's. Transforming growth factor beta inhibition, which is the standard proposal for immune-excluded tumours, was set aside for EMC precisely on the ground that the exclusion here is physical rather than driven by a fibroblast programme.

**Why this limit is distinct from B6.** A cold tumour lacks a response. An excluded tumour may have a response that cannot reach the target. These require different interventions, and an intervention aimed at one does not address the other. A vaccine can raise the circulating frequency of junction-specific T cells without changing whether those cells can enter the tumour.

**Why the new EMC immune evidence does not choose between them.** Both instruments that speak to the question arrived in 2026 and neither sees what the other sees. The bulk cohort reading of B8 runs in the direction B6 predicts, and it is an average over whole tissue with no spatial resolution, so a cold tumour and an excluded one return the same value. The multiplex immunofluorescence arm of the twelve-tumour series runs in the direction B7 predicts, because a T cell that is present and suppressed is not a T cell that is absent, and it is two specimens compared within themselves and called proof of concept by its own authors [15]. The instrument with a cohort has no spatial resolution and the instrument with spatial resolution has no cohort. This paper therefore does not resolve the disagreement and does not treat either limit as retired by the other's evidence; what would resolve it is the observation B8 names, per-tumour spatial profiling on a series large enough to estimate a fraction.

**What would clear or mitigate it.** Vascular normalisation is the mechanism with the most direct EMC evidence, discussed in Section 4. Matrix-directed approaches, including addressing the oncofetal chondroitin sulfate modification itself, are registered in this programme as candidates and are not resolved.

### B8. Thin EMC immune-profiling data

**Proposition.** The characterisations of EMC as cold and as excluded were inferred from the disease's mutational burden, its histology and sarcoma-wide immunotherapy experience rather than from EMC-specific immune profiling. The proposition as originally stated, that no such profiling exists, no longer holds; what follows replaces it.

**Consequence.** Several exclusions above rested on an assumption nobody had measured in this disease, and an absent reading was never a reading of absence.

**What has since been read, in EMC tumour tissue.** Two archival bulk expression cohorts, 6 and 10 EMC tumours against 29 and 6 comparator sarcomas, were scored against lineage, effector and antigen-presentation modules and against published hallmark infiltration proxies. The interferon-γ effector axis, the myeloid compartment and class II antigen presentation are lower in EMC than in the comparator sarcomas on both platforms; the cytotoxic-lymphocyte and T-cell modules are lower on the larger cohort; and two hallmark proxies, the interferon-γ response and the lymphocyte-weighted allograft-rejection set, are lower on both. The class I apparatus is the exception. β2-microglobulin and the class I heavy chains sit near the top of the expressed range on the larger platform, and HLA-A, HLA-C and TAP1 read flat against the comparator rather than lost. Values and their contrasts have one home, `research/modalities/emc-expression-panels.json`, reads 8 and 19.

**What points the other way, in individual patients.** Twelve archival EMC tumours profiled by whole-transcriptome targeted sequencing carry more B-cell infiltration in the low-risk half of the series than in the high-risk half, and proof-of-concept multiplex immunofluorescence on two of the specimens reports exhausted CD3+CD8+PD1+ T cells and FOXP3+ regulatory T cells in the high-risk one; the authors state that every immunofluorescence comparison is a within-specimen region-of-interest analysis, that their prognostic model overfits, and that the work is exploratory and hypothesis-generating [15]. A single reported case profiled as immune-enriched despite a low mutational burden is graded in Section 4 [16]. Neither is a cohort estimate, and a quiet distribution with occasional infiltrated members is consistent with both those observations and with the paragraph above. What the twelve-tumour series adds beyond direction is that the variation it reports is within EMC rather than between EMC and something else, which is the axis a patient-selection argument would have to run on.

**What none of it settles.** The comparator arm is other sarcomas, which are not an immune-normal reference, so lower than that comparator is not the same statement as cold. Bulk expression cannot locate a transcript in a cell, and therefore cannot separate a cold tumour from the excluded one B7 describes. Sixteen tumours cannot estimate what fraction of EMC is infiltrated, which is the quantity a patient-selection argument would need. The class I reading is transcript abundance in a cohort, not protein on a cell surface in a patient, and class I loss would independently disable every antigen-directed route including this one. And the twelve-tumour series has been read here only as its published abstract, because the publisher returned an access error to this work's retrieval runner, so its infiltration method, its external validation cohorts and its per-specimen values are unexamined.

**What would still clear it.** Per-tumour profiling with spatial resolution on a series large enough to estimate that fraction, and class I assessed as protein.

**Cost and owner.** The cohort reading is done. It ran on a continuous-integration runner over two public archival series, with no rental, no accelerator and no purchased compute. What remains requires tissue, and it is still the cheapest tissue-based item in the ledger and the one that most efficiently informs the others.

### B9. Manufacturing economics at this incidence

**Proposition.** An individualised vaccine requires per-patient sequencing, design and manufacture, and EMC's incidence is well under one per million per year.

**Consequence.** No independent programme can supply the manufacturing, and the economics that support an individualised product in melanoma do not obviously extend to a disease with this incidence.

**What would mitigate it.** Two features of this target reduce the requirement relative to a general individualised product. First, the antigen is not discovered per patient by tumour sequencing but is determined by which exon pair the patient carries, of which five are in frame; the design space is therefore small, enumerable in advance, and shared across patients with the same breakpoint. Second,

because the fusion is truncal, the antigen does not need re-selection over the disease course. A small fixed panel of breakpoint-specific constructs, allocated by a diagnostic assay, is closer to a stratified product than to a bespoke one. Whether that is commercially tractable is a question for a platform holder and not one this analysis can answer.

#### B10. Trial design below the randomisation threshold

**Proposition.** At this incidence a conventional randomised trial of a vaccine addition is not feasible.

**Evidence and mitigation.** A histology-specific EMC cohort within a sarcoma master protocol has already been executed and reported for a different regimen, enrolling 24 patients across nine centres in three countries over four years [5]. That shows the vehicle exists, and it gives an accrual rate of about six unselected patients a year across nine centres.

**And that rate does not survive this paper's own eligibility filter.** A junction-vaccine cohort cannot enrol unselected EMC: it needs *EWSR1*::*NR4A3*, which is 62 to 75% of cases [7], and then an HLA type that presents a junction peptide, which is 8.5 to 30.4% of those (B1). Applying both fractions to that trial's own accrual leaves roughly 0.3 to 1.4 eligible patients a year, so a cohort of the size already run would take on the order of two decades, and a powered comparison considerably longer. That is what makes B10 a limit of the disease rather than of study design: the vehicle exists and the patients to put in it do not arrive fast enough. The endpoint question is separate — response-based endpoints behave poorly in indolent tumours, and the existing EMC cohort reported progression-free survival at a fixed time point.

## 4. An ungraded combination

The limits above do not resolve independently. B6 and B7 are properties of the tumour that a vaccine cannot address, and B1 and B3 are properties of the antigen that a checkpoint inhibitor cannot address. The relevant question is therefore not whether a junction vaccine works in EMC, which is the question that was asked and answered negatively, but whether a junction vaccine adds anything to a backbone that already has EMC-specific activity.

Such a backbone exists. A phase 2 histology-specific EMC cohort within the IMMUNOSARC II master protocol evaluated sunitinib with nivolumab in adults with advanced, progressing, centrally confirmed EMC across nine centres in Spain, Italy and the United Kingdom. Of 23 evaluable patients, 16 were progression-free at 6 months, median progression-free survival was 13.2 months (95% CI 5.7 to 20.7), and there were 2 partial responses [5]. This is a conference abstract and has not been peer-reviewed; the full publication is not available, and no results are posted for the registration. It is reported here as the most direct EMC-specific evidence available for an immunotherapy-containing regimen, and not as evidence of efficacy. The preceding mixed-histology phase Ib/II study of the same combination is published [4].

Three features make this backbone the natural context for the vaccine question. The antiangiogenic component addresses B7 by a mechanism — vascular normalisation reduces the physical and vascular barriers to lymphocyte entry — and antiangiogenic tyrosine kinase inhibitors carry the most consistent prospective signal in this disease: pazopanib gave an objective response in 4 of 22 evaluable patients with a median progression-free survival of about 19 months [11], and a sunitinib series reported activity in translocated EMC [13]. The checkpoint component addresses the release arm of B6. Neither addresses antigen supply, which is what the vaccine would contribute.

**One reported single-agent checkpoint response, and what it changes.** The checkpoint half of that backbone has never been given alone in a reported EMC cohort, and one patient has been reported outside one: a man with *EWSR1*::*NR4A3* EMC, microsatellite stable at a tumour mutational burden of 0.67 mut/Mb and with *TSC2* loss, whose microenvironment a commercial assay profiled as immune-enriched and fibrotic with high PD-L1, and who received single-agent pembrolizumab with a near-complete response reported over 15 cycles [16]. One patient in an industry-authored conference abstract that has not been peer reviewed supports no response rate, and it is not read here as a promotion of the checkpoint arm. What it does is name the selection axis: the markers conventionally used to choose a patient for checkpoint inhibition and the markers describing that patient's microenvironment pointed in opposite directions in the same tumour, which is the heterogeneity B8's cohort reading permits and cannot measure. That is a question for B8's observation to settle, not a case for adding a third agent.

This is also the architecture of the melanoma programme behind the recent announcement, where the individualised vaccine is given with pembrolizumab and never alone [3,9]. The transfer to EMC fails on antigen depth — melanoma supplies a pool of private neoantigens, EMC one junction — and not on architecture.

**What this section is claiming, and what it is not.** The proposition is that adding a breakpoint-matched junction construct to a checkpoint and antiangiogenic backbone gives a combination in which each component covers a limit the others do not, and that it has been graded as a unit nowhere — including in this programme's own ledger, which is where B6's observation comes from. It is not a trial proposal. No population, comparator, endpoint, effect size or sample size is specified here; a single-arm addition to a backbone whose own six-month progression-free rate is 16 of 23 could not attribute an effect to the vaccine; and B10's arithmetic says the eligible patients do not arrive at a workable rate. Whether the combination merits evaluation at all turns on B2. Nothing here recommends that any patient receive any of these agents outside a clinical trial.

## 5. Present work and pending arrivals

Four items stood on this list in the previous version, all needing nothing but public data. Two have since been done and are reported above rather than promised here: the distance-to-self and anchor-versus-contact-position filters, which are now the near-self search of B3, and the extension of the class II panel from three DR alleles to 23 across DR, DP and DQ, which is now the positive result of B4. The class I panels remain at 10 and 34 alleles against the full validated set, which would raise the denominator of every coverage figure. A defended acceptance threshold for junction peptides is still absent, and Section 2.3 now shows exactly how much rests on it. The isoform-aware novelty filter identified in B5 is still unbuilt. None of the remainder requires permission, material or funding.

Everything else waits. Measured presentation (B2) would change the character of this route rather than its score, because the screen in Section 2 would stop being a stand-in and become a hypothesis with a calibrating dataset. B8 is no longer an absence: two archival EMC expression series have been read here and a twelve-tumour series has been published, and what remains of that limit is spatial resolution on a series large enough to estimate an infiltrated fraction, together with class I assessed as protein rather than as transcript. It is still the cheapest tissue-gated item and the one that informs the most others, because infiltrate location and HLA class I status bear on every antigen-directed route in this disease. Access to a patient-derived line would additionally make B2 executable rather than merely specified; these are separate arrivals and neither implies the other. B6 and B7 are properties of the tumour that no computational advance mitigates.

## 6. Distinguishing a real advance from an apparent one

Each capability this route waits on has a plausible near-miss that the literature would report in language resembling a hit.

**A result on a different fusion is not a result on this one.** A presentation or immunogenicity study on another fusion-driven tumour raises the prior that fusion junctions in general can be seen by T cells. It does not establish that this junction, on these alleles, is presented at usable abundance. The discriminating question to ask of any such report is whether its evidence concerns presentation, abundance, or only a different fusion.

**Robotic execution is not material access.** A cloud laboratory with per-experiment pricing supplies robots and generic reagents, not an EMC line or organoid, and without that none of the tissue-gated items in Section 3 becomes runnable.

**A better predictor is not a measurement.** An improved class I or class II model would change the numbers in Section 2 and leave B2 where it is. The limit there is that nobody has looked, not that the looking-glass is imprecise.

**A larger allele panel raises a figure without grounding it.** Extending the panels can only move coverage upward, and a higher predicted-presentation figure is not evidence that any patient's tumour presents anything.

### 6.1 Order of the remaining questions

No dates are offered and none of the steps below is scheduled; what follows is an ordering, and the ordering is the content. Each step is cheaper than the one after it and each can close the route, so running them out of order spends the expensive ones on a question a cheap one would have settled.

1. **Defend the acceptance threshold, or record that it cannot be defended.** Computational, needs nothing but public data. Calibrate the cut against experimentally validated epitopes, and if no validated set restricted to fusion-junction peptides exists, that absence is itself the finding and Section 2.3's curve is the only honest report of coverage. Until this is done every figure in B1 is a point on a curve.
2. **Measure whether the lead peptides bind their alleles at all.** Four peptide-allele pairs carry every class I figure in this paper; Table 3 ranks them by what each costs if it fails. Synthesise them and test binding in a cell-based stabilisation assay, first pair first.

**Table 3. The four peptide-allele pairs every class I figure rests on, ranked by what each costs if it fails.** *Coverage at risk* is a leave-one-out difference — the pooled figure with all four presenting alleles minus the figure without this one — and not the allele's marginal contribution as it enters Section 2.3's curve, which is order-dependent and larger. *Percentile* is the call that put the allele in the set. Testing binding for these four is the only step in Section 6.1 that needs neither an EMC specimen nor a proteomics facility.

**Table 3. The four peptide-allele pairs every class I figure rests on, ranked by what each costs if it fails**

| Order | Peptide | Allele | Percentile | Coverage at risk |
|---|---|---|---|---|
| 1 | RGDMPCVQAQY | HLA-A*01:01 | 0.4061 | 9.9 pp |
| 2 | MPPPLRGDM | HLA-B*07:02 | 0.458 | 7.2 pp |
| 3 | NMPCVQAQY | HLA-B*15:01 | 0.3736 | 6.5 pp |
| 4 | NMPCVQAQY | HLA-A*30:02 | 0.4033 | 3.0 pp |

3. **Look for the peptide on a tumour.** Immunopeptidomics on EMC tissue or a patient-derived line, with the limit of detection, the specimen count and their HLA types, and the abundance-matched positive control all fixed before the run, per B2. This is the decisive step and the first that requires material this programme cannot obtain.
4. **Only if step 3 is positive, ask whether a T cell sees it.** Recognition of the peptide-HLA complex by T cells from HLA-matched donors. A negative here is the one that B3 predicts and that nothing computational can rule out.
5. **Only then is there anything to build.** The construct in Section 2 is an output of a screen and is not a candidate for administration; it exists so that steps 2 to 4 have something specific to test.

Steps 1 and 2 would together decide whether the rest is worth commissioning, and neither needs an EMC specimen. That is the whole of what this paper can say about what to do next.

### 6.2 Conditions for revision

Four statements in this paper are falsifiable by an observation a reader could go and make, and they are listed so a future reader can check them rather than re-derive them. Sequence-level novelty for 170 of 174 peptides would be overturned by a proteome release adding a protein that contains one of the four strong binders. Every coverage figure would move upward if a wider panel added a presenting allele and downward if measured presentation removed one of the strong calls, and all of them go to zero if the acceptance threshold is set below 0.37. The class II statement said, in the version this paper supersedes, that it would be changed by a wider DR, DP and DQ panel — named there as the single most likely source of a change to it. That panel has since been run and it did change it, from no strong epitope on three alleles to one on 23; the statement in B4 is now the wider panel's, and what would change it again is measured class II presentation rather than more prediction. And the argument in Section 4 would be closed altogether by an EMC series showing HLA class I loss as protein; the transcript-level reading in B8 does not show it, and a protein-level series that did would end every antigen-directed route in this disease including this one.

## 7. Limitations

All binding figures are predictions from sequence-based models and no EMC-specific validation of either model exists. They predict peptide-MHC affinity alone: proteasomal cleavage and TAP transport are not modelled here, so a strong call bounds what could be presented rather than naming a peptide that is. Coverage is computed by multiplying non-carrier frequencies across alleles, which assumes independence both between loci and between alleles at the same locus. The same-locus case arises here — two alleles pooled are both HLA-B — and handling it correctly moves the pooled figure by about 0.3 percentage points. Cross-locus haplotype linkage disequilibrium is not modelled and its effect is not estimated; bounding it would need haplotype-frequency data rather than the allele frequencies used here. The population-to-region mapping is an approximation, regional figures are point values on samples as

small as 579 individuals, and the frequencies are pooled across populations whose allele frequencies differ by more than the figures being reported. A fourth dependence sits underneath all of these and was not measured until now: every peptide in Section 2 is derived from one canonical transcript per gene, and across the 970 in-frame transcript pairs tested under B5 no junction peptide is common to all of them. The seam residue itself takes nine values, so the peptide identities move with it. The counts are more stable than the identities — each pair yields 30, 34 or 38 novel peptides and no other value — but a reader should take the specific sequences named here, NMPCVQAQY included, as conditioned on the canonical annotation rather than as properties of the locus. Which transcript a patient's fusion actually uses is not decidable from annotation and is not decided here. The binder counts and every coverage figure derived from them depend on an acceptance threshold this paper does not defend, and no multiplicity correction is applied anywhere: 174 peptides were screened against 10 and then 34 alleles at two thresholds, with no decoy control and no null expectation, so the calls that pass are reported as what the screen returned rather than as an enrichment over chance. The near-self search of B3 is the one analysis here that carries a null, and it carries one because a count of near-self neighbours is meaningless without the number a peptide of that length and composition finds by chance; the binding screen still has none, and a shuffle null for predicted binding would need a defended threshold to be a null of anything.

The IMMUNOSARC II EMC cohort result is a conference abstract, single-arm and not peer reviewed, and it evaluates a combination whose component with the larger independent EMC evidence base is the tyrosine kinase inhibitor rather than the checkpoint inhibitor; it is cited to establish that a vehicle and a backbone exist, not to attribute activity to the immune arm. The single-agent checkpoint response cited beside it is one patient in an industry-authored conference abstract and bounds nothing. EMC's characterisation as immune-excluded rests on inference; its characterisation as immunologically quiet rests on the two archival cohorts read here against comparator sarcomas, which is a contrast and not an absolute, and on one twelve-tumour series read here only as an abstract. Neither instrument separates a cold tumour from an excluded one, which is limit B8 and the reason B6 and B7 remain graded apart. The class II panel is 23 alleles across DR, DP and DQ, its one strong call lands on a single uncommon allele, and class II prediction is less accurate than class I, so B4's positive bounds less than a class I call of the same nominal strength. No claim is made that any peptide is presented, that any construct would be immunogenic, that any combination would be safe or effective, or that any of this is ready for clinical use. No wet-laboratory work was performed, and the measurements this characterisation most needs require work this programme cannot carry out.

## 8. Methods and reproducibility

Every figure in Sections 2 and 3 is generated by a script in `research/modalities/` and is committed as a JSON artifact beside it: `fusion_breakpoints.py` for the junction set and predicted binders, `hla_coverage.py` for population coverage, `coverage_scan.py` for the coverage-versus-allele-count curve, `coverage_threshold_curve.py` for the coverage-versus-threshold curve of Section 2.3, `junction_proteome_novelty.py` for the proteome search of Section B5, `junction_frameshift_peptides.py` for the out-of-frame screen of Section 2.2, `junction_selfsimilarity.py` for the near-self search of Section B3, `predictor_concordance.py` for the second-predictor check, `coverage_uncertainty.py` for the within-locus form, the Fréchet bounds and the between-population spread of Section 2.3, `novelty_seam_test.py` for the one-residue pre-screen of Section B5 and its validation, `patient_neoepitopes.py` and `patient_cd4_epitopes.py` for the per-patient shortlisters, and `vaccine_construct.py` for the candidate construct and its minimal synthetic long peptide. The corresponding artifacts are `fusion-breakpoint-neoantigens.json`, `hla-coverage.json`, `coverage-curve.json`, `coverage-threshold-curve.json`, `epitope-allele-matrix.json`, `epitope-allele-loose-matrix.json`, `junction-proteome-novelty.json`, `junction-frameshift-peptides.json`, `junction-selfsimilarity.json`, `predictor-concordance.json`, `coverage-uncertainty.json`, `novelty-seam-test.json`, `patient-cd4-demo.json` and `vaccine-construct.json`. The two tables in this paper are generated from those artifacts by `vaccine_path_tables.py` and a build fails if a cell and its source diverge. The predictor versions those artifacts record are MHCflurry 2.1.4 with models release 2.2.0 for class I and MHCnuggets 2.4.1 for class II. MHCnuggets publishes no models-release identifier separate from its package version — it ships its trained weights inside the distribution — so the version is the whole of the provenance available for that arm, and the artifact records why rather than leaving the field blank.

**Table 2. The 34-allele class I panel.** Listed so the coverage scan can be reproduced and the choice of panel assessed rather than taken on trust. It is a common-allele HLA-A and HLA-B panel spanning global diversity, and it carries no HLA-C and no class II allele — so every figure computed on it is a partial coverage of the class I repertoire, and §2.3's statement that extending the panel can only raise the figures applies to HLA-C before it applies to anything else.

**Table 2. The 34-allele class I panel**

| Locus | Alleles |
|---|---|
| HLA-A (16) | HLA-A*01:01, HLA-A*02:01, HLA-A*02:03, HLA-A*02:06, HLA-A*03:01, HLA-A*11:01, HLA-A*23:01, HLA-A*24:02, HLA-A*26:01, HLA-A*30:01, HLA-A*30:02, HLA-A*31:01, HLA-A*32:01, HLA-A*33:01, HLA-A*68:01, HLA-A*68:02 |
| HLA-B (18) | HLA-B*07:02, HLA-B*08:01, HLA-B*13:02, HLA-B*14:02, HLA-B*15:01, HLA-B*18:01, HLA-B*27:05, HLA-B*35:01, HLA-B*38:01, HLA-B*40:01, HLA-B*40:02, HLA-B*44:02, HLA-B*44:03, HLA-B*46:01, HLA-B*51:01, HLA-B*53:01, HLA-B*57:01, HLA-B*58:01 |

**These are not regenerated on every commit, and this paper does not claim they are.** The workflow that runs them is dispatched by hand, its steps do not fail the run when a generator fails, and it writes its outputs to a separate cache branch from which they are copied in. So the artifacts in the repository are the record of a run that happened, verifiable by their embedded timestamps and input hashes, rather than the output of continuous re-execution. An earlier version of this section said they were regenerated in continuous integration; that was not true of any branch a reader would fetch.

Clinical figures are quoted from the curated EMC registry at `research/data/emc-clinical-registry.json`, which carries a structured citation entry for every source and a retrieval note for those whose identification required one.

## 9. Declarations

**Use of AI tools.** A large language model (Claude, Anthropic) was used throughout this work: to write the analysis code, to run the screening pipelines, to draft and revise this manuscript, and to conduct the internal adversarial review of earlier drafts whose findings this version incorporates. The model versions used over the span of the work are the ones the repository's commit record names, and are not restated here. No quantitative result was generated by a language model directly: every figure in Sections 2 and 3 is produced by the code named in Section 8 and is reproducible from the committed artifacts, and the clinical figures are transcribed from the publications cited for them. Every literature identifier in Section 10 was checked against a retrieved bibliographic record held in this repository, and any identifier that could not be so anchored was removed rather than retained. The author takes full responsibility for all content, including for the correctness of the code and for the interpretation of the results.

**Data and code availability.** All code and all artifacts underlying every figure are in the public repository that accompanies this manuscript, under the paths given in Section 8. No restricted data were used.

**Competing interests.** The author declares no financial competing interests: he holds no position, equity, consultancy or patent relating to any gene, sequence, peptide or agent named here. One non-financial interest is declared: the author is a survivor of extraskeletal myxoid chondrosarcoma, the disease this work addresses.

**Funding.** This work received no external funding and was self-funded by the author. No funder had any role in the analyses, the interpretation of the results, or the decision to publish.

**Ethics.** No human subjects, human material or animals were involved. No patient data were used; the clinical figures quoted are published or publicly presented aggregate results.

**Not clinical guidance.** Nothing in this manuscript is medical advice, and nothing in it is evidence that any agent or combination is safe or effective in extraskeletal myxoid chondrosarcoma. No peptide reported here has been synthesised, formulated, or tested in any cell, tissue or animal by anyone, and none may be administered to any person. The agents named in Section 4 are discussed as a research hypothesis and their use outside a clinical trial is not supported by anything in this paper.

## 10. References

Every identifier below is transcribed from a bibliographic record — either one held in this repository's literature and registry files, or one retrieved from Europe PMC or Crossref by the verification workflow that accompanies this work. None is written from recollection.


1. Gonzalez-Galarza FF, et al. Allele Frequency Net Database (AFND) 2020 update. *Nucleic Acids Research*
2. doi:10.1093/nar/gkz1029. Accessed via the MIT-licensed `slowkow/allelefrequencies` mirror.
3. O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides. *Cell Systems* 2020. doi:10.1016/j.cels.2020.09.001. The version run here is 2.1.4 with models release 2.2.0.
4. Weber JS, Carlino MS, Khattak A, Meniawy T, Ansstas G, Taylor MH, et al. Individualised neoantigen therapy mRNA-4157 (V940) plus pembrolizumab versus pembrolizumab monotherapy in resected melanoma (KEYNOTE-942): a randomised, phase 2b study. *Lancet* 2024;403(10427):632-644. doi:10.1016/S0140-6736(23)02268-7. PMID 38246194.
5. Martin-Broto J, Hindi N, Grignani G, Martinez-Trufero J, Redondo A, Valverde C, et al. Nivolumab and sunitinib combination in advanced soft tissue sarcomas: a multicenter, single-arm, phase Ib/II trial. *Journal for ImmunoTherapy of Cancer* 2020. doi:10.1136/jitc-2020-001561. PMID 33203665.
6. Hindi N, Palmerini E, Carrasco-Garcia I, Gonzalez-Billalabeitia E, Valverde C, Strauss SJ, et al. Phase II of sunitinib plus nivolumab in extraskeletal myxoid chondrosarcoma: results from the GEIS, ISG and UCL IMMUNOSARC II study. *Journal of Clinical Oncology* 2025;43(16_suppl):11513. doi:10.1200/JCO.2025.43.16_suppl.11513. **Conference abstract, single-arm, not peer reviewed**; registration NCT03277924, for which no results are posted and no full publication is available.
7. Iwata S, Noguchi R, Osaki J, Adachi Y, Shiota Y, Osaki S, et al. Establishment and characterization of NCC-EMC1-C1: a novel patient-derived cell line of extraskeletal myxoid chondrosarcoma. *Human Cell* 2025;38(4):122. doi:10.1007/s13577-025-01250-7. PMID 40580361.
8. Remiszewski P, Falkowski S, Szumera-Ciećkiewicz A, Spałek MJ, Rutkowski P, Czarnecka AM. From pathogenesis to the patient's bedside: a comprehensive review of extraskeletal myxoid chondrosarcoma. *Journal of Cancer Research and Clinical Oncology* 2025;151(11):283. doi:10.1007/s00432-025-06316-5. PMID 41055792.
9. Wu ZY, He YQ, Wang TM, Yang DW, Li DH, Deng CM, et al. Glycogenes in oncofetal chondroitin sulfate biosynthesis are differently expressed and correlated with immune response in placenta and colorectal cancer. *Frontiers in Cell and Developmental Biology* 2021;9:763875. doi:10.3389/fcell.2021.763875. PMID 34966741. The tissues studied are placenta and colorectal cancer, not sarcoma.
10. Merck and Moderna announce that the phase 3 INTerpath-001 trial of intismeran autogene plus pembrolizumab met its endpoints of recurrence-free survival and distant metastasis-free survival in completely resected stage IIB-IV melanoma. Company press release, 19 August 2026. https://www.merck.com/news/merck-and-moderna-announce-phase-3-interpath-001-trial-of-intismeran-autogene-plus-keytruda-met-endpoints-of-recurrence-free-survival-in-completely-resected-stage-iib-iv-melanoma/ (accessed 22 August 2026; retrieval record `literature/interpath-001-announcement-2026-08-22/`). **This is an announcement, not a publication**: no effect size was disclosed in it, none is quoted here, it carries no digital object identifier, and it is not indexed in any bibliographic database this work can query. It is cited only for the fact that the announcement was made.
11. Huang SC, Lee JC, Hsu YC, Tsai JW, Kao YC, Hsieh TH, et al. Extraskeletal myxoid chondrosarcomas: the uncommon clinicopathologic manifestations and significance of TAF15::NR4A3 fusion. *Modern Pathology* 2023;36(7):100161. doi:10.1016/j.modpat.2023.100161. PMID 36948401. This is the 58-case molecularly confirmed series.
12. Stacchiotti S, Ferrari S, Redondo A, Hindi N, Palmerini E, Vaz Salgado MA, et al. Pazopanib for treatment of advanced extraskeletal myxoid chondrosarcoma: a multicentre, single-arm, phase 2 trial. *The Lancet Oncology* 2019. doi:10.1016/S1470-2045(19)30319-5. PMID 31331701. Registration NCT02066285; 26 patients started treatment and 22 were evaluable for the primary endpoint.
13. Shao XM, Bhattacharya R, Huang J, Sivakumar IKA, Tokheim C, Zheng L, et al. High-throughput prediction of MHC class I and II neoantigens with MHCnuggets. *Cancer Immunology Research* 2020. doi:10.1158/2326-6066.CIR-19-0464. PMID 31871119. This is the class II predictor used for the CD4 arm of Section B4. The artifact records the tool and the version run, 2.4.1; MHCnuggets publishes no models release separate from its package version, and the artifact says so in that field rather than leaving it empty. *Superseded, retained: earlier versions of this entry and of Section 8 stated that the artifact recorded neither a version nor a models release, which was true of them and is no longer true of this one.*
14. Stacchiotti S, Pantaleo MA, Astolfi A, Dagrada GP, Negri T, Dei Tos AP, et al. Activity of sunitinib in extraskeletal myxoid chondrosarcoma. *European Journal of Cancer* 2014. doi:10.1016/j.ejca.2014.03.013. PMID 24703573. Retrospective series of 10 patients.
15. Bangerter JL, Harnisch KJ, Chen Y, Hagedorn C, Planas-Paz L, Pauli C. Establishment, characterization and functional testing of two novel ex vivo extraskeletal myxoid chondrosarcoma (EMC) cell models. *Human Cell* 2023;36(1):446-455. doi:10.1007/s13577-022-00818-x. PMID 36316541.
16. Chaiboonchoe A, Chanthercrob J, Sakamula R, Likhityungyuen T, Muangsomboon S, Chantharasamee J, et al. Prognostic biomarkers for enhanced risk stratification in extraskeletal myxoid chondrosarcoma: a retrospective cohort study. *PeerJ* 2026;14. doi:10.7717/peerj.21497. PMID 42465974. Twelve archival formalin-fixed EMC tumours by whole-transcriptome targeted sequencing, with multiplex immunofluorescence on two of them; the authors report internal validation indicating that their prognostic model overfits, state that every immunofluorescence comparison is a within-specimen region-of-interest analysis, and describe the work as exploratory and hypothesis-generating. **Only the abstract has been read for this manuscript**: the publisher returned HTTP 403 to the retrieval runner used here, so the infiltration method, the external validation cohorts and the per-specimen values are unexamined. Retrieval record `research/literature/emc-fourth-cohort-publication-2026-08-24.json`.
17. Galitskiy E, Livingston AJ, Plair T, Tarasov A, Putintsev V, Novokrestchenova A, et al. P35-1 Effective immune-based treatment of extraskeletal myxoid chondrosarcoma guided by next-generation gene profiling. *Annals of Oncology* 2025;36:S1732-S1733. doi:10.1016/j.annonc.2025.08.310. **Conference abstract, one patient, not peer reviewed**; fifteen of the seventeen authors are employees of the company whose assay is the subject, and the abstract text held in this repository has that company's own publication page as its only source. Registry entry `galitskiy2025emcpembrolizumab` in `research/data/emc-clinical-registry.json`.


**Data sources cited as resources rather than as publications.** Transcript structures are Ensembl records for *EWSR1* and *NR4A3*, retrieved and cached as committed inputs. The proteome searched in Section B5 is UniProtKB reference proteome UP000005640, reviewed entries with isoforms included, 42,547 sequences and 24,513,032 residues at retrieval, fetched from the UniProt REST stream; the isoform that carries the four colliding peptides is Q92570-3.

## Appendix A. Superseded figures

The following values were reported before the 2026-08-07 coordinate-system correction and must not be quoted. They are retained so that earlier drafts and derived documents can be identified.

| Quantity | Superseded value | Current value |
|---|---|---|
| In-frame junctions | 7 | 5 |
| Distinct predicted binders | 26 | 11 |
| *EWSR1* exon 7 junction, presenting alleles | A*11:01 and B*08:01 | B*15:01 on 10 alleles; B*15:01 and A*30:02 on 34 |
| *EWSR1* exon 7 junction coverage | 29.7% | 8.5% on 10 alleles, 12.3% on 34 |
| Any-strong-binder coverage | 58.0% | 27.4% on 10 alleles, 30.4% on 34 |
| Regional range | 36% to 79% | 1.4% to 60.4% |
| Broad-panel presenting alleles | 20 of 34 | 4 of 34 |
| Broad-panel coverage ceiling | 84.5% | 30.4%, and not a ceiling |
| Combined CD8 and CD4 coverage | 16.5% | 1.8% |
| Candidate minimal synthetic long peptide | 27 residues, both arms | 15 residues, both arms |
| Class II predicted binders at the *EWSR1* exon 7 junction | 9, of which 4 strong | 44 on 23 alleles, of which 1 strong |

The superseded values arose from a model that concatenated coding sequences and thereby discarded the acceptor exon's retained 5' untranslated region. The corrected junction set is disjoint from the superseded one, so the earlier peptide identifiers do not appear in the current artifacts.

## Appendix B. Statements withdrawn by the first adversarial review of this manuscript

This manuscript was reviewed on 2026-08-22 by five independent blind readers against a pinned commit, each with a different lens, none able to see the others. The statements below stood in the version they read and do not stand now. They are recorded because a reader who met an earlier copy, or a derived document that quoted one, needs to be able to identify what changed and why — and because a correction that leaves no trace is indistinguishable from a claim that was never made.

| Withdrawn statement | Why it does not stand |
|---|---|
| "presented on HLA-B*15:01 alone", stated without naming a panel | True of the ten-allele screen only. The 34-allele screen finds the same lead peptide strong on HLA-A*30:02, and the junction's coverage is then 12.3% rather than 8.5%. |
| The 30.4% figure described as a "ceiling" | Sections 5 and 6 of the same manuscript said extending the panel could only raise it. It is a union over four alleles at one threshold, and moving that threshold to 0.37 takes it to zero. |
| "It does not attain 50% at any panel size" — withdrawn | No search over panel sizes was performed; the panel is fixed at 34 and the scanned variable is the number of presenting alleles. In Northern Europe two alleles reach 52.8%. |
| Wilson 95% confidence intervals on every coverage figure | They were computed by pooling every reference population into one binomial. The same records show the frequency of one pooled allele ranging from 0 to 0.40 between those populations, so the model the interval assumes is refuted by its own input. |
| "the combined CD8 and CD4 figure is null rather than unreported. That is a result" — withdrawn | The instrument's class II branch never evaluates when no allele qualifies, so the empty field records "not computed". The instrument's own note calls the class II coverage it would compute a floor that untested alleles could only raise. |
| "The one substantive reasoning correction this work offers …", and the symmetry it asserted | The quoted exclusions are this author's own route-ledger notes, not positions in the literature, and the vaccine's own entry in that ledger is parked on immunogenicity rather than on absent priming. The claim is narrowed in B6 and the Abstract accordingly. |
| The dated capability bands of the former Section 3.1 (2027H2 / 2029 / 2032) | Withdrawn entire. They were self-described as the weakest material in the paper, and a forecast a reader cannot check does not become checkable by being tabulated. |
| "regenerated in continuous integration", of the artifacts in Section 8 | The workflow is dispatched by hand, its steps do not fail the run when a generator fails, and it writes to a cache branch. Section 8 now says what is actually true. |
| Reference 8 quoted as "correlated with disease outcome" | The study's title names immune response in placenta and colorectal cancer, which is the tissue setting it examined. |
| References 10 and 11 as "[citation to verify]" | Both were resolvable in this repository at no cost and are now written out. The one reference that genuinely has no record — the class II predictor — says so in its own entry instead. |
| "24 patients across nine centres … gives a realistic accrual rate" | Unfiltered by this paper's own eligibility criteria. Applying them gives roughly 0.3 to 1.4 eligible patients a year. |

## Appendix C. Statements withdrawn after the first adversarial review

The review recorded in Appendix B was not the last correction. The statements below stood after it and do not stand now, and they are recorded on the same principle: a correction that leaves no trace is indistinguishable from a claim that was never made.

| Withdrawn statement | Why it does not stand |
|---|---|
| B8, "rather than from published EMC-specific immune profiling" | Such profiling was published on 2026-07-13, four weeks before this manuscript's date, on the twelve-tumour cohort described in B8's own revision. |
| B8, HLA class I expression status "is not known for this disease" | The antigen-presentation precondition — β2-microglobulin, the class I heavy chains, the peptide transporters, tapasin, NLRC5, the immunoproteasome subunits and ERAP1 — had been read on both EMC array platforms on 2026-08-09, in this author's own artifacts, before the sentence was written. |
| B6 and B8 read together, that whether EMC is immunologically quiet was unmeasured here | It is now measured in the two archival cohorts, and the direction is that EMC is quieter than the comparator sarcomas on the interferon axis, the myeloid compartment and class II presentation. That is a cohort contrast against other sarcomas and not a statement that the disease is cold; the distinction is stated in B8. |
| Section 1, "no published EMC immune profiling" listed among the limits of access | The same statement as B8's, in a second place, and false for the same reason. It is the one-of-a-pair defect: correcting B8 alone would have left the claim standing where a reader meets it first. |
| Section 7, "The characterisations of EMC as cold and as immune-excluded rest on inference rather than on published EMC-specific immune profiling" | Third home of the same claim. The exclusion half still rests on inference; the quiet half no longer does. |
| Section 7, "The class II panel is three DR alleles with no DP or DQ, so its negative bounds a narrow question" | Superseded by the panel widening reported in B4 and in Appendix A, and stale in this section since that widening. Section 7 was contradicting B4 inside the same document: the panel is 23 alleles and the arm is not negative. This was not found by any adversarial round; it is a wording that no numeric guard reads. |
| B7, "No EMC-specific expression evidence for the pathway is cited, because none is known to the author" | The chondroitin-sulfate biosynthetic and sulfation machinery had been read on both EMC array platforms in this author's own artifacts before the sentence was written. The reading is reported in B7 and does not support the analogy the sentence was protecting, which is a different correction from the one the sentence claimed to make. |
| The B6 and B7 rows of Section 3, cost to move "not movable by any computation" | True of each limit and false of the evidence bearing on it: a computation over public archival series moved what is known about both. Those cells now name what the remaining move costs, which is spatial profiling on tissue. |