Mobile Genetic Elements

ISSN: (Print) 2159-256X (Online) Journal homepage: http://www.tandfonline.com/loi/kmge20

Proteomics technique opens new frontiers in mobilome research

Andrew D. Davidson, David A. Matthews & Kevin Maringer

To cite this article: Andrew D. Davidson, David A. Matthews & Kevin Maringer (2017) Proteomics technique opens new frontiers in mobilome research, Mobile Genetic Elements, 7:4, 1-9, DOI: 10.1080/2159256X.2017.1362494

To link to this article: http://dx.doi.org/10.1080/2159256X.2017.1362494

Accepted author version posted online: 01 Aug 2017.
Published online: 01 Aug 2017.

MOBILE GENETIC ELEMENTS
2017, VOL. 7, NO. 4, e1362494 (9 pages)
https://doi.org/10.1080/2159256X.2017.1362494

Taylor & Francis
Taylor & Francis Group

COMMENTARY

OPEN ACCESS Check for updates

Proteomics technique opens new frontiers in mobilome research

Andrew D. Davidson[a,†], David A. Matthews[a,†], and Kevin Maringer [b]

[a]School of Cellular and Molecular Medicine, University of Bristol, Bristol, UK; [b]Department of Microbial Sciences, Faculty of Health and Medical Sciences, University of Surrey, Guildford, UK

ABSTRACT

A large proportion of the genome of most eukaryotic organisms consists of highly repetitive mobile genetic elements. The sum of these elements is called the "mobilome," which in eukaryotes is made up mostly of transposons. Transposable elements contribute to disease, evolution, and normal physiology by mediating genetic rearrangement, and through the "domestication" of transposon proteins for cellular functions. Although 'omics studies of mobilome genomes and transcriptomes are common, technical challenges have hampered high-throughput global proteomics analyses of transposons. In a recent paper, we overcame these technical hurdles using a technique called "proteomics informed by transcriptomics" (PIT), and thus published the first unbiased global mobilome-derived proteome for any organism (using cell lines derived from the mosquito *Aedes aegypti*). In this commentary, we describe our methods in more detail, and summarise our major findings. We also use new genome sequencing data to show that, in many cases, the specific genomic element expressing a given protein can be identified using PIT. This proteomic technique therefore represents an important technological advance that will open new avenues of research into the role that proteins derived from transposons and other repetitive and sequence diverse genetic elements, such as endogenous retroviruses, play in health and disease.

ARTICLE HISTORY
Received 18 July 2017
Revised 25 July 2017
Accepted 28 July 2017

KEYWORDS
Aedes aegypti; endogenous retrovirus; LTR retrotransposon; mobilome; proteomics informed by transcriptomics (PIT); repetitive element; transposon proteomics

Mobile genetic elements are DNA sequences that can move within and between genomes. In eukaryotes, transposons make up the majority of such elements, comprising between 5% (yeast; *Saccharomyces cerevisiae*) and 77% (frog; *Rana esulenta*) of an organism's genome.[1] The sum of an organism's transposable elements is referred to as its mobilome. We recently reported the first high-throughput global profiling of an organism's mobilome-derived proteome.[2] In this commentary, we provide a more focused description of our transposon proteomics method, and discuss which aspects of transposon biology are best studied proteomically. While our emphasis here is on transposons, our technique is equally useful for studying endogenous retroviruses and other repetitive and/or sequence-diverse elements that are not fully represented in reference genome databases.

Why study the transposon proteome?

The fact that transposable elements constitute such a large proportion of most eukaryotic genomes makes their study important for fully understanding an organism's biology. The most widely known activity of transposons is their ability to transpose and insert themselves into new positions within the genome. class I elements replicate via a "copy and paste" mechanism in which an RNA transcript derived from the genomic transposon sequence acts as a template for cDNA (cDNA) production by a transposon-encoded reverse transcriptase.[3-5] This cDNA copy integrates elsewhere in the genome through the action of a transposon-encoded integrase to create new copies of the element.[4,5] class II elements do not replicate *via* an RNA intermediate.[3,6] Instead, "cut and paste" DNA transposons use transposase enzymes to excise and insert themselves elsewhere within the genome, with copies generated through DNA repair mechanisms, and during S phase if the donor, but not the acceptor, site has been replicated before transposition.[3,6] Non-RNA-mediated "copy and paste" transposition mechanisms also exist.[3,6] Transposons express several

CONTACT Kevin Maringer k.maringer@surrey.ac.uk Department of Microbial Sciences, Faculty of Health and Medical Sciences, University of Surrey, Guildford, Surrey, GU2 7XH, UK.
[†]These authors contributed equally to this work.

Commentary to: Maringer K, et al. Proteomics informed by transcriptomics for characterising active transposable elements and genome annotation in *Aedes aegypti*. BMC Genomics 2017;18:1-18; PMID:28049423; https://doi.org/10.1186/s12864-016-3432-5

proteins during transposition, including enzymes and structural proteins.[3-5] Some transposable elements do not encode their own proteins, hijacking the machinery of other elements instead; these include short interspersed nuclear elements (SINEs) and miniature inverted repeat transposable elements (MITEs).[3] These non-autonomous elements are not detectable proteomically and will not be discussed further. Individual transposons tend to lose their ability to transpose over time, both through host defense mechanisms and through the acquisition of inactivating mutations.[7-12]

Transposition is biologically interesting because the insertion of transposons into host gene coding sequences or regulatory elements can generate new phenotypes. Exons or entire genes may be copied, disrupted or shuffled, new introns created, epigenetic modifications altered, and gene expression modulated.[6,13,14] Large-scale chromosomal rearrangements also occur.[6] Transposon activity is therefore both a driver in the evolution of new functions,[6,13,14] and a contributing factor in diseases such as cancer and hereditary disorders.[6,13-17]

Defining the transpositionally active mobilome is challenging. Genomic studies only reveal whether a transposable element was recently active in general terms, evidenced by new genomic insertions in offspring compared to parents, or by insertion site variation between individuals or species in which elements have been active since the last common ancestor.[9] Transposition in specific cells or tissues under varying conditions however is difficult to capture. On the other hand, RNA sequencing can detect transposon RNA in individual samples, but also picks up RNA-mediated host defenses against mobile elements that are not indicative of transposition.[18-20] Reporter assays measuring the transposition of specific elements are useful for targeted studies, but do not provide a complete picture of the active mobilome and do not identify which genomic copies of an element are active. In contrast to these approaches, proteomics has the potential to provide a complete picture of mobilome activity by identifying all protein-producing transposons in a sample, many of which will be in the process of active transposition.

In addition to transposition-mediated effects, it has become evident that transposons can be "domesticated" and their genetic material co-opted for new cellular functions.[1,6,21] At least 50–100 plant and mammalian proteins are known to originate from transposons.[1] For example, transposase-derived genes contribute to V(D)J recombination during B- and T-cell receptor maturation, and the DNA-binding domains of several transcription factors and proteins involved in chromosome segregation also originate from transposases.[1,6] Meanwhile, proteins derived from the structural gag and env proteins of long-terminal repeat (LTR) retrotransposons (class I) and endogenous retroviruses have been linked to placental development, cell proliferation, apoptosis, and antiviral defenses.[1,21-23]

Transposon-derived cellular genes can be distinguished from non-domesticated transposable elements by their lack of functional transposition sequences, lack of inactivating mutations, evolution under purifying selection, and single-copy coding sequences that are maintained at orthologous loci across species.[1] Especially those with known functions should in principle be annotated in reference genomes. However, identifying domesticated transposons, particularly recently domesticated ones, can be challenging in genomes containing many related and recently active transposable elements.[6] Domesticated transposon-derived proteins have so far been identified either serendipitously in molecular studies of cellular and disease mechanisms, or through bioinformatic genome analyses that provide no evidence for protein production and often focus on just one type of transposon protein. Here too, unbiased proteomic experiments can help identify unrecognised cellular functions derived from mobile genetic elements by surveying the complete repertoire of transposons that demonstrably produce protein. Protein function may also be hinted at from protein expression dynamics in different contexts (e.g. cancerous *versus* non-cancerous cells).

Proteomics therefore has several advantages over genomics and transcriptomics in measuring global mobilome activity, and can make valuable contributions to all investigations into the numerous aspects of normal physiology and disease processes in which transposons and transposon-derived proteins play a role. The major limitation of proteomics is that it cannot definitively prove active transposition, even if all proteins from a single element are detected. On the other hand, detection of only a single protein from a given element may not discount active transposition, due to experimental limits of protein detection and

the potential contribution of proteins from other transposons to transposition. Nevertheless, proteomics provides a valuable springboard into mechanistic follow-on studies, and adds the capability of detecting transposition-independent protein expression from mobile genetic elements.

Why is defining the transposon proteome technically challenging?

In typical global proteomic workflows, protein isolated from an experimental sample is separated by gel electrophoresis, tryptically digested, and analyzed by

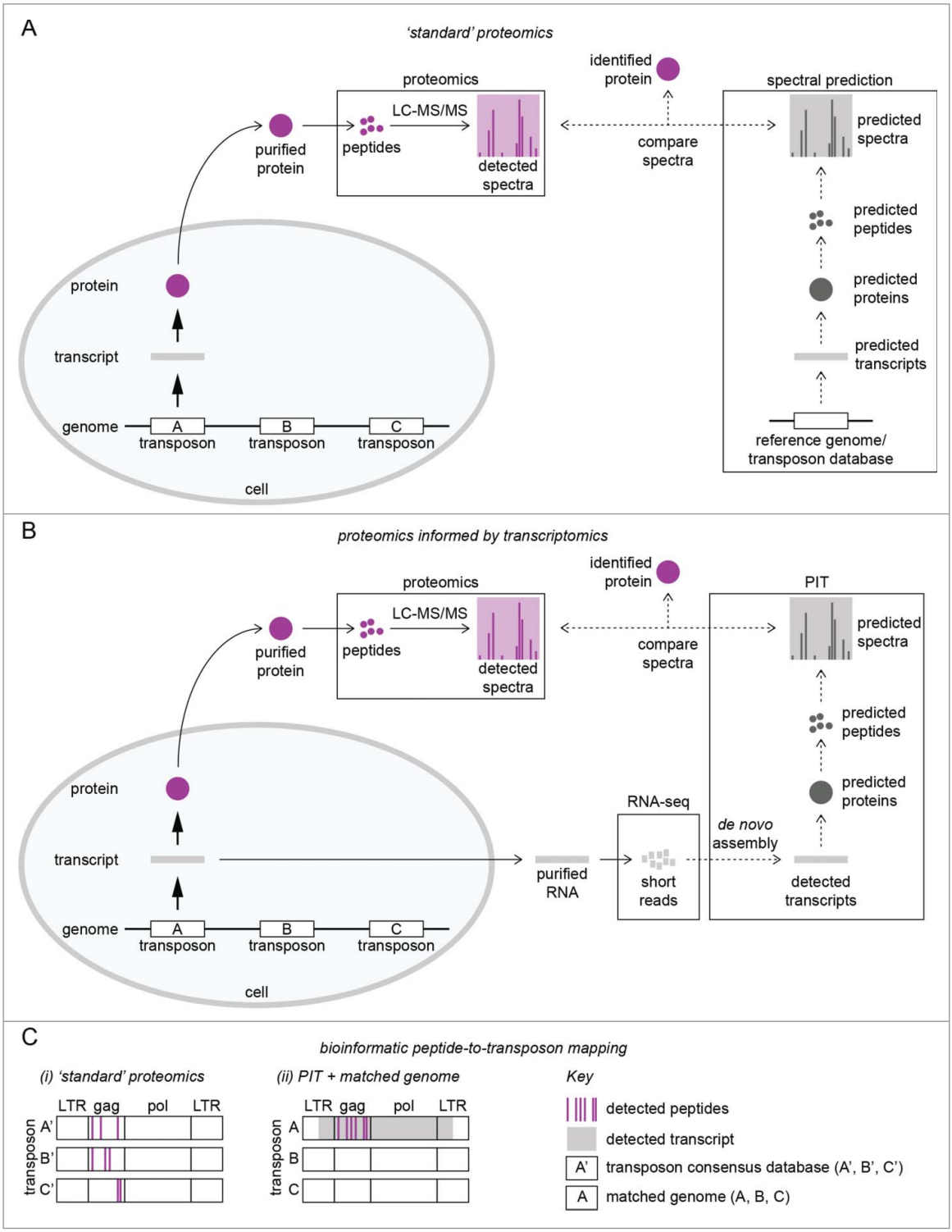

Figure 1. (For figure legend, see page 4.)

liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) to produce a set of spectra that correspond to the detected peptides (Fig. 1A). Peptides, and ultimately proteins, are identified by comparing these spectra to spectra bioinformatically predicted from protein annotation in reference genomes (Fig. 1A). From obtaining good DNA sequence coverage of transposons present in the genome, to bioinformatically relating detected peptides back to individual transposable elements, there are several hurdles that make mobilome proteomics challenging technically.

(1) Coverage of highly repetitive elements is frequently incomplete in genomes sequenced using Sanger and Illumina platforms, because short reads often do not span the full length of large transposons.[24] (2) High quality genome annotation of transposable elements is often lacking, partly because their highly repetitive and sequence-diverse nature complicates their identification, and partly because automating transposon annotation is difficult.[24-26] (3) To facilitate gene annotation, repetitive sequences are purposefully masked in genome assemblies,[27] meaning reference genomes cannot be used for predicting transposon proteins, peptides, and spectra in proteomic workflows. (4) Dedicated repetitive element databases such as Repbase (girinst.org)[28] and Tefam (tefam.biochem.vt.edu) do exist, but mostly list consensus sequences of phylogenetically related elements.[28,29] Individual transposons may diverge considerably from this consensus.[29] (5) Reference genomes may not accurately reflect the mobilome of a given experimental sample, because transposon sequences and insertion sites can vary substantially between populations, individuals, and tissues.[9,30,31] (6) Large copy numbers (up to one million copies for the most common transposon family (Alu) in humans)[9,32] make bioinformatically assigning detected peptides to a specific genomic copy of an element virtually impossible (Fig. 1Ci; but see later).

These specific challenges are exacerbated by the generally poor assembly and annotation quality of many genome sequences, and the large and diverse array of bioinformatic tools used to identify repetitive elements, which complicate comparisons between genomes.[24] Performing proteomics on endogenous retroviruses and other sequence-diverse non-annotated genetic elements poses similar challenges.

"Proteomics informed by transcriptomics" captures the total mobilome-derived proteome, and identifies specific elements expressing protein

We recently performed the first high-throughput global proteomic analysis of an organism's transposon proteome in a cell line derived from the mosquito *Aedes aegypti*.[2] Several previous studies had proteomically analyzed a subset of protein spots excised after 2D gel electrophoresis, but had focused on only a limited selection of transposon proteins (e.g., transposase).[33-40] The method we used, "proteomics informed by transcriptomics" (PIT),[41,42] solves the aforementioned problems afflicting mobilome proteomics by circumventing the requirement for genome annotation and instead identifying peptides based on matched RNA-Seq data (Fig. 1B). In PIT, the experimental sample is split; protein is extracted from one part and processed for LC-MS/MS as usual, while RNA is isolated from the rest and used for RNA-Seq. RNA sequencing reads are assembled into transcripts *de novo* (without the use of a reference genome) using one of several bioinformatic transcriptome assembly programmes, and translated *in silico* to predict proteins, peptides, and spectra that are ultimately used to

Figure 1. (see previous page) PIT is a superior proteomic method for detecting proteins derived from mobile genetic elements. (A) 'Standard' workflow for global proteomic analysis. Purified proteins are separated by gel electrophoresis, tryptically digested into <20-residue peptides, and analyzed by LC-MS/MS. Detected peptides are identified from their respective spectra using predicted proteins, peptides, and spectra ultimately derived from genome annotation and/or transposons reference databases. (B) PIT workflow. Experimental samples are split into protein (processed as described in A) and RNA; the latter is subjected to RNA sequencing. Short RNA-Seq reads are bioinformatically assembled into transcripts *de novo* (without referring to a reference genome),[49] and proteins, peptides, and spectra are predicted from these transcripts to produce a bespoke reference database for identifying peptides. In A & B, solid arrows indicate 'wet' experiments; dashed arrows indicate 'dry' (bioinformatic) analyses. (C) Bioinformatic identification of transposable elements expressing protein, in this case 'transposon A' (LTR retrotransposon used for illustration purposes). (i) Short peptides detected by LC-MS/MS often match multiple transposons sharing short stretches of perfect amino acid conservation, and individual elements deviate from the database consensus, complicating the assignment of peptides to their correct transposon. (ii) In PIT, detected peptides perfectly match their experimentally verified transcripts, allowing protein-producing mobile genetic elements to be accurately identified. When combined with a perfectly matched genome sequence, the precise genomic location of protein-producing transposons can also be determined (see Fig. 2).

determine which proteins were detected by LC-MS/MS (Fig. 1B).[41,42] The result is a bespoke reference database exquisitely matched to the proteome of the experimental sample, which is limited only by RNA sequencing depth.[42] PIT therefore solves the combined problems of incomplete repetitive element sequence coverage, identification, and annotation in genomes, as well as the potentially poor fit of experimental data to reference databases.

In our study, we identified transposon proteins by BLASTing the *in silico* translation of detected peptide-associated transcripts against the Tefam and Repbase reference databases.[2] Using the full-length amino acid sequence is important, because the short (<20 residue) peptides detected by LC-MS/MS could map to multiple transposons, while increased sequence coverage allows specific elements to be detected confidently (Fig. 1C). Although nucleotide BLAST could in theory be performed instead, protein BLAST is preferable because it reduces divergence from the consensus by excluding synonymous sequence differences. In this way, we identified a total of 136 transposon proteins in our sample with high confidence.[2] It is important to tailor the thresholds for transposon protein identification to each species and reference database, as we observed differences in a side-by-side comparison of the Tefam and Repbase databases.[2] Only 15 of the 136 identified transposon proteins closely matched the *Ae. aegypti* transposon reference database,[2] confirming the aforementioned technical challenges to transposon proteomics posed by incomplete transposon identification, the inclusion of only consensus sequences in databases, and potential differences between a given experimental system and the reference genome.

Importantly, we also validated PIT's ability to make biologically relevant observations about mobile genetic elements.[2] For example, non-LTR retrotransposons (class I) encode 2 ORFs, with ORF1 often truncated and not transcribed.[5] This was reflected in our PIT data, with fewer proteins detected for ORF1 than ORF2 for non-LTR retrotransposons.[2] Another interesting finding was the overabundance of proteins detected from LTR retrotransposons compared to other elements,[2] despite the fact that non-LTR retrotransposons are more abundant in *Ae. aegypti*.[43] Although this result must be interpreted with caution, as our proof-of-principle study included just one data point from a cell line that may not reflect the *in vivo* situation, our results are in agreement with the enrichment of LTR retrotransposon-derived small RNAs, known to correlate with transposon activity,[44] in the related insect *Drosophila melanogaster*.[45] Since LTR retrotransposons specifically are implicated in antiviral defenses in *Ae. aegypti*,[22,23] we postulated that this mosquito may differentially allow LTR retrotransposons to remain active while suppressing other elements. If this result is corroborated, investigating the mechanisms by which the organism achieves differential transposon silencing, and copes with the potential deleterious consequences of heightened LTR retrotransposon activity, would be highly interesting. Although a discordance between genomic abundance and transposition activity has previously been observed in genomic studies,[30] we are the first to describe this at the protein level,[2] which may reflect not only transposition but also other (possibly cellular) functions of transposon proteins.

After publishing our study, the genome for the *Ae. aegypti* cell line we used (Aag2) was sequenced and made available at vectorbase.org.[46,47] We wanted to test whether combining our PIT data with a matched genome sequence would allow us to pinpoint precisely which genomic copies of an element express protein. We therefore BLASTed (blast.ncbi.nlm.nih.gov) the full experimentally determined sequence of our 17 detected transposon transcripts that were associated with at least 2 peptides against the Aag2 cell genome (with repeats unmasked). In principle, each detected RNA transcript sequence should match the genomic DNA sequence at the locus from which it derives with 100% sequence identity across the full transcript length (100% query coverage). In practice, many of the thousands of genomic copies of a transposable element may be almost identical to each other and the transcript. Furthermore, sequencing errors and differences between our Aag2 cell clone and the published reference sequence may reduce the observed sequence identity. For our purposes, we considered transcripts exhibiting at least 99% nucleic acid sequence identity over 99% query coverage to be an "exact match." Using these criteria, we were able to identify the exact genomic transposon sequence expressing protein for 5 elements (Fig. 2A). By cross-referencing the Aag2 contig containing the protein-expressing transposon with the *Ae. aegypti* reference genome (Liverpool strain version L3,[43] vectorbase.org), and a physical chromosome map for *Ae. aegypti*,[48] we were also able to identify the physical chromosomal location of the identified elements (Fig. 2B).

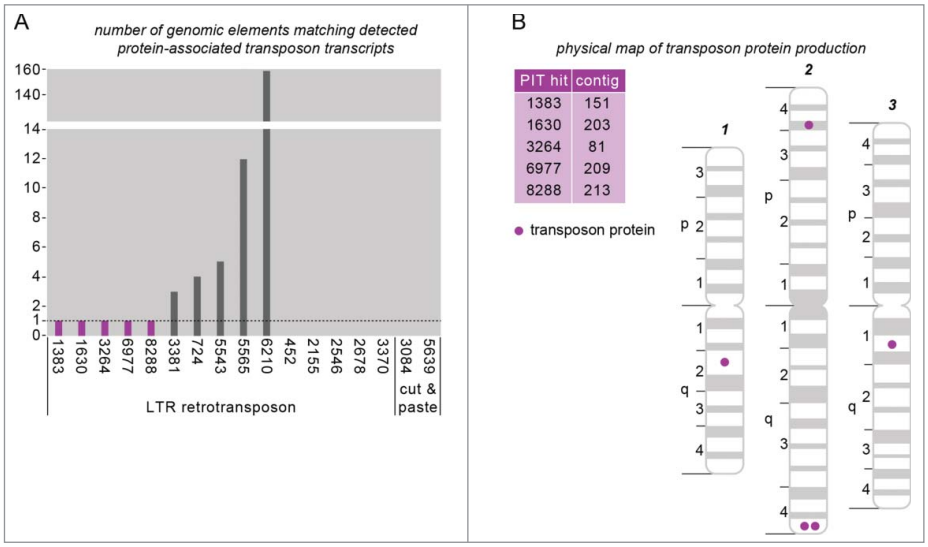

Figure 2. PIT can identify the specific unique genomic copy of an element that is expressing protein. (A) Number of genomic transposon sequences that precisely match PIT transcripts associated with 2 or more peptides. Transcripts with only one genomic match (purple bars) can be mapped to their exact source. Numbers (x axis) correspond to PIT hit IDs from our associated paper.[2] (B) Physical chromosome map for *Ae. aegypti*[48]; protein-expressing transposons for which the genomic source could be identified are annotated.

However, it is not always possible to map protein-expressing transposons in this way. For example, 5 transposon transcripts matched multiple almost identical genomic transposon sequences and could thus not be accurately located to a single source (Fig. 2A). Identical insertions contained within larger repeat regions are also expected to complicate this kind of analysis. Finally, several transcripts had no close match in the reference genome (Fig. 2A), either due to incomplete sequence coverage of repetitive elements,[24] or because these elements differ between our clone of the cell line and the sequenced clone. Due to mobilome divergence, it was not possible to accurately directly map protein-expressing transposons using the main *Ae. aegypti* reference genome (Liverpool strain version L3,[43] vectorbase.org; data not shown), highlighting the need for perfectly matched genome, transcriptome, and proteome data for this kind of analysis.

We therefore provide proof-of-principle that PIT can not only characterize the global profile of the mobilome-derived proteome, but also that detected transposon proteins can be matched to their precise genomic source. It should be noted however that, overall, our proteomic approach is likely facilitated by the fact that mosquitoes encode a large diversity of mobile genetic elements, each with a relatively low copy number compared to mammals.[24] Using PIT (and other approaches) to characterize the mobilome-derived proteome may be more challenging in humans and (almost all) other placental mammals, where the major active protein-producing transposable element is the highly abundant non-LTR retrotransposon L1.[24]

Concluding remarks and future prospects

Our PIT pipeline allows interrogation of the mobilome-derived proteome in a global and unbiased way for the first time, opening up exciting new opportunities for defining the total contribution of transposon-derived proteins to cellular function, as well as for characterizing transposon activity in different contexts. Importantly, global transposon proteome profiling will allow the field to move away from targeted studies and the serendipitous discovery of transposon protein functions in health and disease, and toward holistic experiments that give a complete picture of the positive and negative impacts of the mobilome on its host organism. Combining transcriptomic and proteomic data with matched genomic information provides a powerful toolkit for dissecting the contribution of individual transposons, out of the thousands of genomic copies of an element, to the overall global activity of the mobilome. Inherently, our methods are equally valuable for studying endogenous retroviruses and other repetitive and/or divergent genetic elements that may or may not be accurately represented in reference genomes. The tools we have developed

therefore open exciting new avenues of research into the dynamic role these mobile DNA sequences play in cellular function, disease, and the evolution of new phenotypes, while also capturing their changing activity during invasion and eventual silencing, inactivation, and domestication in new hosts.

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

Acknowledgments

The authors thank Catriona Macfarlane for constructive comments on the manuscript.

Funding

The work associated with this commentary was funded by: Wellcome Trust fellowship 096062 and a seed grant from the Faculty of Health and Medical Sciences, University of Surrey, to KM; Medical Research Council (UK) grant G0801973 to ADD; BBSRC grant BB/M02542X/1 to ADD and DAM; BBSRC grants BB/M020118/1, BB/L018438/1, and BB/K016075/1 to DAM.

ORCID

Kevin Maringer http://orcid.org/0000-0003-0977-8807

References

[1] Alzohairy AM, Gyulai G, Jansen RK, Bahieldin A. Transposable elements domesticated and neofunctionalized by eukaryotic genomes. Plasmid. 2013;69:1-15. doi:10.1016/j.plasmid.2012.08.001. PMID:22960324

[2] Maringer K, Yousuf A, Heesom KJ, Fan J, Lee D, Fernandez-Sesma A, Bessant C, Matthews DA, Davidson AD. Proteomics informed by transcriptomics for characterising active transposable elements and genome annotation in Aedes aegypti. BMC Genomics. 2017;18:1-18. doi:10.1186/s12864-016-3432-5. PMID:28049423

[3] Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973-82. doi:10.1038/nrg2165. PMID:17984973

[4] Nefedova L, Kim A. Mechanisms of LTR-Retroelement transposition: lessons from drosophila melanogaster. Viruses. 2017;9:81. doi:10.3390/v9040081

[5] Han JS. Non-long terminal repeat (non-LTR) retrotransposons: mechanisms, recent developments, and unanswered questions. Mob DNA. 2010;1:15. doi:10.1186/1759-8753-1-15. PMID:20462415

[6] Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 2007;41:331-68. doi:10.1146/annurev.genet.40.110405.090448. PMID:18076328

[7] Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ. An epigenetic role for maternally inherited piRNAs in transposon silencing. Science. 2008;322:1387-92. doi:10.1126/science.1165171. PMID:19039138

[8] Khurana JS, Wang J, Xu J, Koppetsch BS, Thomson TC, Nowosielska A, Li C, Zamore PD, Weng Z, Theurkauf WE. Adaptation to P element transposon invasion in drosophila melanogaster. Cell. 2011;147:1551-63. doi:10.1016/j.cell.2011.11.042. PMID:22196730

[9] Huang CRL, Burns KH, Boeke JD. Active transposition in genomes. Annu Rev Genet. 2012;46:651-75. doi:10.1146/annurev-genet-110711-155616. PMID:23145912

[10] Lohe AR, Moriyama EN, Lidholm DA, Hartl DL. Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. Mol Biol Evol. 1995;12:62-72. doi:10.1093/oxfordjournals.molbev.a040191. PMID:7877497

[11] Selker EU, Cambareri EB, Jensen BC, Haack KR. Rearrangement of duplicated DNA in specialized cells of Neurospora. Cell. 1987;51:741-52. doi:10.1016/0092-8674(87)90097-3. PMID:2960455

[12] Goyon C, Faugeron G. Targeted transformation of Ascobolus immersus and de novo methylation of the resulting duplicated DNA sequences. Mol Cell Biol. 1989;9:2818-27. doi:10.1128/MCB.9.7.2818. PMID:2674671

[13] Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008;9:397-405. doi:10.1038/nrg2337. PMID:18368054

[14] Kidwell M, Lisch D. Transposable elements and host genome evolution. Trends Ecol Evol. 2000;15:95-9. doi:10.1016/S0169-5347(99)01817-0. PMID:10675923

[15] Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, Vogelstein B, Nakamura Y. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. Cancer Res. 1992;52:643-5. PMID:1310068

[16] Scott E, Devine S. The role of Somatic L1 Retrotransposition in human cancers. Viruses. 2017;9:131. doi:10.3390/v9060131

[17] Chen JM, Cooper DN. A Mechanistic link between L1 retrotransposition and chromothripsis. Hum Mutation. 2016;37:329-9. doi:10.1002/humu.22870. PMID:26950403

[18] Bronkhorst AW, Miesen P, van Rij RP. Small RNAs tackle large viruses: RNA interference-based antiviral defense against DNA viruses in insects. Fly. 2013;7:216-23. doi:10.4161/fly.25708. PMID:23974177

[19] Ribeiro JMC, Arcà B, Lombardo F, Calvo E, Phan VM, Chandra PK, Wikel SK. An annotated catalogue of salivary gland transcripts in the adult female mosquito, Aedes aegypti. BMC Genomics. 2007;8:6. doi:10.1186/1471-2164-8-6. PMID:17204158

[20] Castellano L, Rizzi E, Krell J, Di Cristina M, Galizi R, Mori A, Tam J, De Bellis G, Stebbing J, Crisanti A, et al. The germline of the malaria mosquito produces

abundant miRNAs, endo-siRNAs, piRNAs and 29-nt small RNAs. BMC Genomics. 2015;16:100. doi:10.1186/s12864-015-1257-2. PMID:25766668

[21] Naville M, Warren IA, Haftek-Terreau Z, Chalopin D, Brunet F, Levin P, Galiana D, Volff JN. Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. Clin Microbiol Infect. 2016;22:312-23. doi:10.1016/j.cmi.2016.02.001. PMID:26899828.

[22] Goic B, Vodovar N, Mondotte JA, Monot C, Frangeul L, Blanc H, Gausson V, Vera-Otarola J, Cristofari G, Saleh M-C. RNA-mediated interference and reverse transcription control the persistence of RNA viruses in the insect model Drosophila. Nat Immunol. 2013;14:396-403. doi:10.1038/ni.2542. PMID:23435119

[23] Goic B, Stapleford KA, Frangeul L, Doucet AJ, Gausson V, Blanc H, Schemmel-Jofre N, Cristofari G, Lambrechts L, Vignuzzi M, et al. Virus-derived DNA drives mosquito vector tolerance to arboviral infection. Nat Commun. 2016;7:12410. doi:10.1038/ncomms12410. PMID:27580708

[24] Sotero-Caio CG, Platt RN, Suh A, Ray DA. Evolution and diversity of transposable elements in vertebrate genomes. Genome Biol Evol. 2017;9:161-77. doi:10.1093/gbe/evw264. PMID:28158585

[25] Platt RN, Blanco-Berdugo L, Ray DA. Accurate transposable element annotation is vital when analyzing new genome assemblies. Genome Biol Evol. 2016;8:403-10. PMID:26802115. doi:10.1093/gbe/evw009

[26] Hoen DR, Hickey G, Bourque G, Casacuberta J, Cordaux R, Feschotte C, Fiston-Lavier A-S, Hua-Van A, Hubley R, Kapusta A, et al. A call for benchmarking transposable element annotation methods. Mob DNA. 2015;6:1-9. doi:10.1186/s13100-015-0044-6

[27] Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13:329-42. doi:10.1038/nrg3174. PMID:22510764

[28] Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015; 6(11):1-6

[29] de Parseval N, Heidmann T. Human endogenous retroviruses: from infectious elements to human genes. Cytogenet Genome Res. 2005;110:318-32. doi:10.1159/000084964. PMID:16093684

[30] Reiss D, Mager DL. Stochastic epigenetic silencing of retrotransposons: does stability come with age? Gene. 2007;390:130-5. doi:10.1016/j.gene.2006.07.032. PMID:16987613

[31] Chen X-G, Jiang X, Gu J, Xu M, Wu Y, Deng Y, Zhang C, Bonizzoni M, Dermauw W, Vontas J, et al. Genome sequence of the Asian Tiger mosquito, Aedes albopictus, reveals insights into its biology, genetics, and evolution. Proc Natl Acad Sci. 2015;112:E5907-15. doi:10.1073/pnas.1516410112. PMID:26483478

[32] Szmulewicz MN, Novick GE, Herrera RJ. Effects of Alu insertions on gene function. Electrophoresis. 1998;19:1260-4. doi:10.1002/elps.1150190806. PMID:9694261

[33] Torres AR, Rodrigues EP, Batista JS, Gomes DF, Hungria M. Proteomic analysis of Soybean [Glycine max (L.) Merrill] roots inoculated with Bradyrhizobium japonicum Strain CPAC 15. Proteomics Insights. 2013;6:7-11. PMID:25288888

[34] Kleiner M, Young JC, Shah M, VerBerkmoes NC, Dubilier N. Metaproteomics reveals abundant transposase expression in mutualistic endosymbionts. MBio. 2013;4:e00223-13. doi:10.1128/mBio.00223-13. PMID:23781067

[35] Marondedze C, Thomas LA. Apple hypanthium firmness: new insights from comparative proteomics. Appl Biochem Biotechnol. 2012;168:306-26. doi:10.1007/s12010-012-9774-9. PMID:22733236

[36] Ding C, You J, Wang S, Liu Z, Li G, Wang Q, Ding Y. A proteomic approach to analyze nitrogen- and cytokinin-responsive proteins in rice roots. Mol Biol Rep. 2012;39:1617-26. doi:10.1007/s11033-011-0901-4. PMID:21607616

[37] Anderson DC, Campbell EL, Meeks JC. A soluble 3D LC/MS/MS proteome of the filamentous cyanobacterium Nostoc punctiforme. J Proteome Res. 2006;5:3096-104. doi:10.1021/pr060272m. PMID:17081061

[38] Li W, Gao Y, Xu H, Zhang Y, Wang J. A proteomic analysis of seed development in Brassica campestri L. PLoS One. 2012;7:e50290. doi:10.1371/journal.pone.0050290. PMID:23189193

[39] Goodchild A, Raftery M, Saunders NFW, Guilhaus M, Cavicchioli R. Biology of the cold adapted archaeon, Methanococcoides burtonii determined by proteomics using liquid chromatography-tandem mass spectrometry. J Proteome Res. 2004;3:1164-76. doi:10.1021/pr0498988. PMID:15595725

[40] Guo B, Chen Y, Zhang G, Xing J, Hu Z, Feng W, Yao Y, Peng H, Du J, Zhang Y, et al. Comparative proteomic analysis of embryos between a maize hybrid and its parental lines during early stages of seed germination. PLoS One. 2013;8:e65867. doi:10.1371/journal.pone.0065867. PMID:23776561

[41] Evans VC, Barker G, Heesom KJ, Fan J, Bessant C, Matthews DA. De novo derivation of proteomes from transcriptomes for transcript and protein identification. Nat Methods. 2012;9:1207-11. doi:10.1038/nmeth.2227. PMID:23142869

[42] Fan J, Saha S, Barker G, Heesom KJ, Ghali F, Jones AR, Matthews DA, Bessant C. Galaxy integrated Omics: Web-based standards-compliant workflows for proteomics informed by transcriptomics. Mol Cell Proteomics. 2015;14:3087-93. doi:10.1074/mcp.O115.048777. PMID:26269333

[43] Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, et al. Genome sequence of Aedes aegypti, a major arbovirus vector. Science. 2007;316:1718-23. doi:10.1126/science.1138878

[44] van Rij RP, Berezikov E. Small RNAs and the control of transposons and viruses in Drosophila. Trends Microbiol. 2009;17:163-71. doi:10.1016/j.tim.2009.01.003. PMID:19299135

[45] Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, Xu J, Kittler ELW, Zapp ML, Weng Z, et al. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. Science 2008;320:1077-81. doi:10.1126/science.1157396. PMID:18403677

[46] Zach W, Dolan PT, Kunitomi M, Tassetto M, Seetin MG, Oh S, Heiner C, Paxinos E, Andino R. Long-read assembly of the Aedes aegypti genome reveals the nature of heritable adaptive immunity sequences. bioRxiv. 2017; 127498

[47] Giraldo-Calderón GI, Emrich SJ, Maccallum RM, Maslen G, Dialynas E, Topalis P, Ho N, Gesing S, VectorBase Consortium, Madey G, et al. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic Acids Res. 2015;43:D707-13. doi:10.1093/nar/gku1117. PMID:25510499

[48] Timoshevskiy VA, Kinney NA, Debruyn BS, Mao C, Tu Z, Severson DW, Sharakhov IV, Sharakhova MV. Genomic composition and evolution of Aedes aegypti chromosomes revealed by the analysis of physically mapped supercontigs. BMC Biol. 2014;12:27. doi:10.1186/1741-7007-12-27. PMID:24731704

[49] Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644-52. doi:10.1038/nbt.1883