**Mass Spectrometry (for the main text)**

Two independent tumors of each tumor type (MDA-MB-231 or LM2) were analyzed. The ECM-enriched samples from each tumor were subsequently analyzed by mass spectrometry as previously described (Naba et al., 2012). Briefly, ECM-enriched fractions were solubilized in urea, disulfide bonds reduced and alkylated, and proteins digested with PNGaseF, Lys-C, and trypsin. Solutions that began cloudy upon initial reconstitution were clear after overnight digestion.

Aliquots of each digest (150 ng) were analyzed with an automated nano LC-MS/MS system, consisting of an Agilent 1200 nano-LC system (Agilent Technologies, Wilmington, DE) coupled to an Orbitrap Elite Fourier transform mass spectrometer (Thermo Fisher Scientific, San Jose, CA) equipped a nanoflow ionization source (James A. Hill Instrument Services, Arlington, MA). Peptides were eluted from a 10 cm column (Picofrit 75 um ID, New Objectives) packed in-house with reversed phase resin that was either Microsorb C-18 3um particles, 300A pores (Agilent Technologies, )for CuGS, ErSyd, and SmG venoms or ReproSil-Pur C18-AQ 3um particles, 120A pores (Dr. Maisch, Ammerbuch Germany) for TiC venom. The 90 min acetonitrile/0.1% formic acid gradient was run at a flow rate of 200 nl/min to yield ~20 sec peak widths. Solvent A was 0.1% formic acid and solvent B was 90% acetonitrile / 0.1% formic acid. The elution portion of the LC gradient was 3-6% solvent B in 1 min, 6-36% in 40 min, 36-56% in 10 min, 56-90% in 5 min and held at 90% solvent B for 9 min. Data-dependent LC-MS/MS spectra were acquired in ~3 sec cycles; each cycle was of the following form: one full Orbitrap MS scan at 60,000 resolution followed by 12 MS/MS scans in the orbitrap at 15,000 resolution using an isolation width of 3.5 m/z. The top 5 most abundant precursor ions were each sequentially subjected to CID, HCD, and ETD dissociation. Dynamic exclusion was enabled with a mass width of +/- 10 ppm, a repeat count of 1, and an exclusion duration of 12 sec. Charge state screening was enabled along with monoisotopic precursor selection and non-peptide monoisotopic recognition to prevent triggering of MS/MS on precursor ions with unassigned charge or a charge state of 1. For CID, the normalized collision energy was set to 30 with an activation Q of 0.25 and activation time of 10 ms. For HCD, the normalized collision energy was set to 35. For ETD, fluoranthene was used as the ETD reagent with an anion AGC target of 500,000 ions, supplemental activation was enabled, and the reaction time was dependent on the precursor charge state (precursor charge state - reaction time in msec: +2-100, +3-66.7, +4-50, +5-40, +6-33.3, etc). All MS/MS spectra were collected with an AGC target ion setting of 100,000 ions. The instrument control software does not currently allow for separate AGC targets for each dissociation mode. Optimal AGC targets would be closer to 30,000 ions for CID, HCD; and 200,000 ions for ETD. All mass spectra associated with this paper may be downloaded from ftp://ftp.broadinstitute.org/distribution/proteomics/public\_datasets/Undheim\_journalName\_2013.

Mass spectra were interpreted with Spectrum Mill. MS/MS spectra were searched against a UniProt database containing human (78,369 entries) sequences (including isoforms and excluding fragments) downloaded from the UniProt web site on June 30, 2010 with a set of common laboratory contaminant proteins (73 entries) appended.

7 centipede 454 RNA-Seq libraries (SaH\_RL1, SmG\_RL2, ErCut\_RL3, CwT\_RL4, SxPM\_RL5, TlC\_RL7, ?RL8 ), described earlier and in table X were assembled with Trinity, concatenated, and translated into 6-frames of protein sequence for use in searching with MS/MS spectra. The sequence data as raw reads, assembled transcripts, and translated ORFs is available at the following URL: where are we going to put this? How polished do you think the protein database should be?

Peptides identified with a false discovery rate < 1.6% were assembled into identified proteins, and annotated as being ECM-derived or not (described in Naba et al., 2012). MS/MS spectra searches allowed for carbamidomethylation of cysteines and possible carbamylation of N-termini as fixed/mix modifications. Allowed variable modifications were oxidized methionine, deamidation of asparagine, pyro-glutamic acid modification at N-terminal glutamine, pyro-carbamidomethyl modification at N-terminal cysteine, and hydroxylation of proline with a precursor MH+ shift range of -18 to 97 Da. Hydroxyproline was only observed in the proteins known to have it (collagens and proteins containing collagen domains, emilins, etc.) and only within the expected GXPG sequence motifs. Additional detailed information can be found in Supplemental Materials and Methods. Upon acceptance of this manuscript, the original mass spectra associated with this manuscript may be downloaded from ftp://ftp.broadinstitute.org/distribution/proteomics/public\_datasets/Morgenstern\_journalName\_2014.

**Reduction/alkylation *and digestion***

1ug venom was obtained from n animals. Cysteines were reduced and alkylate using (EU/DM fill in details to yield ethanolyl cys). After solubilizing 500ng venom in 100 mM ammonium bicarbonate pH 8.0 digestion was done using either trypsin or Glu-C (Sequencing Grade, Promega, Madison, WI), at a ratio of 1:50 enzyme:substrate, with vortexing at 37°C for 4 hours. Solutions were desalted using Ziptips? David did the digestions at the Broad Institute, please edit Karl’s placeholder sentences.

***Mass Spectrometry***

Aliquots of each digest (150 ng) were analyzed with an automated nano LC-MS/MS system, consisting of an Agilent 1200 nano-LC system (Agilent Technologies, Wilmington, DE) coupled to an Orbitrap Elite Fourier transform mass spectrometer (Thermo Fisher Scientific, San Jose, CA) equipped a nanoflow ionization source (James A. Hill Instrument Services, Arlington, MA). Peptides were eluted from a 10 cm column (Picofrit 75 um ID, New Objectives) packed in-house with reversed phase resin that was either Microsorb C-18 3um particles, 300A pores (Agilent Technologies, )for CuGS, ErSyd, and SmG venoms or ReproSil-Pur C18-AQ 3um particles, 120A pores (Dr. Maisch, Ammerbuch Germany) for TiC venom. The 90 min acetonitrile/0.1% formic acid gradient was run at a flow rate of 200 nl/min to yield ~20 sec peak widths. Solvent A was 0.1% formic acid and solvent B was 90% acetonitrile / 0.1% formic acid. The elution portion of the LC gradient was 3-6% solvent B in 1 min, 6-36% in 40 min, 36-56% in 10 min, 56-90% in 5 min and held at 90% solvent B for 9 min. Data-dependent LC-MS/MS spectra were acquired in ~3 sec cycles; each cycle was of the following form: one full Orbitrap MS scan at 60,000 resolution followed by 12 MS/MS scans in the orbitrap at 15,000 resolution using an isolation width of 3.5 m/z. The top 5 most abundant precursor ions were each sequentially subjected to CID, HCD, and ETD dissociation. Dynamic exclusion was enabled with a mass width of +/- 10 ppm, a repeat count of 1, and an exclusion duration of 12 sec. Charge state screening was enabled along with monoisotopic precursor selection and non-peptide monoisotopic recognition to prevent triggering of MS/MS on precursor ions with unassigned charge or a charge state of 1. For CID, the normalized collision energy was set to 30 with an activation Q of 0.25 and activation time of 10 ms. For HCD, the normalized collision energy was set to 35. For ETD, fluoranthene was used as the ETD reagent with an anion AGC target of 500,000 ions, supplemental activation was enabled, and the reaction time was dependent on the precursor charge state (precursor charge state - reaction time in msec: +2-100, +3-66.7, +4-50, +5-40, +6-33.3, etc). All MS/MS spectra were collected with an AGC target ion setting of 100,000 ions. The instrument control software does not currently allow for separate AGC targets for each dissociation mode. Optimal AGC targets would be closer to 30,000 ions for CID, HCD; and 200,000 ions for ETD. All mass spectra associated with this paper may be downloaded from ftp://ftp.broadinstitute.org/distribution/proteomics/public\_datasets/Undheim\_journalName\_2013.

***Protein Sequence Database used for searching MS/MS spectra***

7 centipede 454 RNA-Seq libraries (SaH\_RL1, SmG\_RL2, ErCut\_RL3, CwT\_RL4, SxPM\_RL5, TlC\_RL7, ?RL8 ), described earlier and in table X were assembled with MIRA concatenated, and translated into protein sequence for use in searching with MS/MS spectra. The 6-frame translation retained as many ORFs from each DNA entry that met the conditions of being longer than 8 amino acid residues and not-shorter-than the next longest ORF from that DNA entry by a factor of 0.8. The sequence data as raw reads, assembled transcripts, and translated ORFs is available at the following URL: where are we going to put this? How polished do you think the protein database should be?

Eivind, I used the starting FASTA file that you described in a 12/15/2012 email as:

The last file is the compiled fasta file with all the centipede 454 data I have, including cuticle from ErC (the ancestral organ) and one on a PNG Scolopendra.

We can include a table with characteristics of the libraries including a count of the transcripts. Since I count the transcripts according to the prefix of the identifiereach library is broken down by subtype which you clarified in a 1/2/2013 email as:

The \_c, rep\_c, and \_s prefixes are the type of contigs. As far as I've understood it. \_s means it is a singlet "contigs", \_c means it's a "proper" contig, and \_rep\_c is a contig where repeats of reads were used to construct it. I avoid relying on singlets due to the error rate of 454, although since the sequencing depth is not great they may well be low abundance transcripts that I suppose could be verified with proteomic data if that's investigated "deeper".

***Protein identification and quantitation***

All MS data was interpreted using the Spectrum Mill software package v4.1 beta (Agilent Technologies, Santa Clara, CA). Similar MS/MS spectra acquired on the same precursor m/z within +/- 30 sec were merged. MS/MS spectra with precursor charge >11 and poor quality MS/MS spectra, which failed the quality filter by not having a sequence tag length > 0 (i.e., minimum of two masses separated by the in-chain mass of an amino acid) were excluded from searching. MS/MS spectra were searched against the database described above that contains 6-frame translation of de novo assembled transcripts from RNA-Seq reads. A set of common laboratory contaminant proteins was NOT appended but probably should be. Searches were conducted separately for spectra from each dissociation method. Scoring parameters were ESI Orbitrap (CID), ESI Orbitrap-ETD, or ESI Orbitrap-HCD respectively. CID and HCD spectra were allowed +/- 20 ppm product mass tolerance and 40% minimum matched peak intensity. ETD spectra were allowed +/- 100 ppm product mass tolerance (due to an apparent instrument acquisition software bug affecting mass calibration of high mass ETD product ions) and 30% minimum matched peak intensity. All spectra were allowed +/- 20 ppm precursor mass tolerance. Ethanolyl cysteine was a fixed modification. Allowed variable modifications were oxidized methionine, deamidation of asparagine, and pyro-glutamic acid modification at N-terminal glutamine with a precursor MH+ shift range of -18 to 64 Da. Tryptic digests were searched with trypsin enzyme specificity with a maximum of 3 missed cleavages. Glu-C digests were searched with Glu-C with cleavage specificity of post D or E with a maximum of 6 missed cleavages.

Identities interpreted for individual spectra were automatically designated as confidently assigned using the Spectrum Mill autovalidation module to apply target-decoy based false-discovery rate (FDR) scoring threshold criteria via a two-step auto threshold strategy at the spectral and protein levels. First, peptide mode was set to allow automatic variable range precursor mass filtering with score thresholds optimized to yield at spectral level FDR of <1.2% for each dissociation method and each precursor charge state in each LC-MS/MS run. Second, protein mode was applied to further filter all the peptide-level validated spectra combined from both LC-MS/MS runs (trypsin and Glu-C digests) derived from a single venom sample using a minimum protein score of 20 and a maximum protein-level FDR of zero. Since the maximum peptide score is 25, the protein level step filters the results so that each identified protein is comprised of multiple peptides unless a single excellent scoring peptide was the sole match. The above criteria yielded false discovery rates of <1.0% for each sample at the peptide-spectrum match level and <1.2 % at the distinct peptide level as estimated by target-decoy-based searches using reversed sequences. In calculating scores at the protein level and reporting the identified proteins, redundancy is addressed in the following manner: the protein score is the sum of the scores of distinct peptides. A distinct peptide is the single highest scoring instance of a peptide detected through an MS/MS spectrum. MS/MS spectra for a particular peptide may have been recorded multiple times, (i.e. as different precursor charge states, different dissociation methods, modified by deamidation at Asn or oxidation of Met) but are still counted as a single distinct peptide. When a peptide sequence >8 residues long is contained in multiple protein entries in the sequence database, the proteins are grouped together and the highest scoring one and its accession number are reported. In some cases when the protein sequences are grouped in this manner there are distinct peptides which uniquely represent a lower scoring member of the group (isoforms, family members, and different species). Each of these instances spawns a subgroup and multiple subgroups are reported and counted towards the total number of proteins and in Table S2 they are given related protein subgroup numbers (e.g., Latent-transforming growth factor beta-binding protein 2(LTBP2), the human and murine forms and are listed as subgroup members 781.1 and 781.2 respectively).

Relative abundances of proteins were determined using extracted ion chromatograms (XIC’s) for each peptide precursor ion in the intervening high resolution FT-MS scans of the LC-MS/MS runs. An individual protein’s abundance was calculated as the sum of the ion current measured for all peptide precursor ions with MS/MS spectra confidently assigned to that protein. The peak area for the XIC of each precursor ion subjected to MS/MS was calculated automatically by the Spectrum Mill software in the intervening high-resolution MS1 scans of the LC-MS/MS runs using narrow windows around each individual member of the isotope cluster. Peak widths in both the time and *m*/*z* domains were dynamically determined based on MS scan resolution, precursor charge and *m*/*z,* subject to quality metrics on the relative distribution of the peaks in the isotope cluster vs theoretical. Although the determined protein ratios are generally reliable to within a factor of 2-fold of the actual ratio, numerous experimental factors contribute to variability in the determined abundance for a protein. These factors may include incomplete digestion of the protein; widely varying response of individual peptides due to inherent variability in ionization efficiency as well as interference/suppression by other components eluting at the same time as the peptide of interest, differences in instrument sensitivity over the mass range analyzed, and inadequate sampling of the chromatographic peak between MS/MS scans.

***Functional Annotation of Protein sequences***

All of the protein translated transcript sequences with peptide spectrum matches described above were annotated using BLAST2GO. Provide details.

Not sure how much of this was done by Karl then re-done in part or in whole by Eivind.