Legends for Supplementary Figures

Supplementary Figure 1. A schematic outlining of the strategy used for proteomic analysis of *Anopheles gambiae*. Protein lysates from larvae, pupae and adult mosquito tissues (head, salivary gland, malpighian tubules, ovary, midgut, viscera and male accessory gland and testis) were subjected to trypsin digestion. LC-MS/MS analysis was carried out using a high resolution Fourier transform mass spectrometer. The mass spectrometry-derived data were searched against a six-frame translated protein database, splice variant protein database and protein database of *Anopheles gambiae* (AgamP3.6 build). Bioinformatics analysis was carried out to further characterize the peptides uniquely identified from the genome search. Selected novel and corrected gene structures were further validated by RT-PCR and sequencing.

Supplementary Figure 2. A) Correlation of counts of number of peptides per protein with the number of proteins in each category. **B)** Distribution of peptide coverage per protein from nine different tissues.

Supplementary Figure 3. Coverage of sequence by mass spectrometry data. A) AGAP009323-PA was identified with 100% sequence coverage including an N-terminal acetylated peptide. B) Extensive sequence coverage of AGAP003153 (VATA_ANOGA) gene is shown that is based on 69 unique peptides that validated each of its five exons and four junctional peptides that confirmed all exon-exon junctions.

Supplementary Figure 4. Change in the structure of the AGAP010657 gene based on GSSPs. A) Seven peptides mapped to intergenic and intragenic region of the AGAP010657 gene. The SNAP prediction model supported a C-terminal exon extension of AGAP010657 and the presence of novel exon in the intronic region of gene AGAP010657. B) The MS/MS spectra of two genome search specific peptides, DYFGFLEVVAR and DVLIFHSDEFK, are shown.

Supplementary Figure 5. Joining of two genes. A) A single gene structure was identified in place of two separate predicted genes using 11 intergenic peptides mapped between the two Ensembl annotated genes, AGAP011872 and AGAP011873. B) Representative MS/MS spectra for peptides SDVLISGLGGLGVEVAK and HNIALIVADTR supporting the novel gene are shown.

Supplementary Figure 6. Identification of novel genes using peptides identified in intergenic regions. A) Thirteen peptides were identified in the intergenic region on chromosome 2L where intron of VectorBase gene model AGAP007548-RB is annotated on the opposite strand. Presence of a novel gene is also predicted by SNAP prediction program. B) Representative MS/MS spectra for peptides THFYNATQSVGENVTQWY and FGSNLEAILLDK supporting novel gene are shown.

Supplementary Figure 7. Change in the gene structure using peptide in the intron of the annotated gene. A) Fifteen peptides were identified in 3 different introns of the gene AGAP010021 which suggest change in the gene structure. These intronic peptides and other intergenic peptides support longer gene model supported the SNAP and Fgenesh models. B)

Representative MS/MS spectra for peptides LCRPVCANDQSCLNNER and ASCSCFSGMVPSPTAK supporting change in gene structure are shown.

Supplementary Figure 8. Confirmation of splice sites from peptides spanning exon-exon junctions. Peptides identified in protein database search but not detected in tblastn search were further investigated. In the case of the AGAP011026 gene, four such exon spanning peptides were mapped to exon1-exon2, exon4-exon5, exon5-exon6 and exon6-exon7 junctions. Exons are marked in alternate black and blue letters to distinguish the junctions.

Supplementary Figure 9. Confirmation of splice sites from peptides spanning exon-exon junctions. The AGAP002350 gene (Q3B723_ANOGA) encoding troponin-T exists in 5 different isoforms. Five splice junction peptides are shown thap map all 5 splice variants.

Supplementary Figure 10. Identification of a novel splice form for the AGAP003656 gene. A peptide, DCSDGEDEICEAQR, was identified in the MS/MS ion search against a synthetic database of potential splice isoforms which indicates a novel splice event occurring between exons 3 and 6 of the annotated gene. The MS/MS spectrum, which led to identification of the novel exon-exon junction, DCSDGEDEICEAQR, is shown.

Supplementary Data. 1) Details of the contribution of proteomic data to VectorBase. 2) Selected MS/MS spectra of genome search specific peptides. 3) Details of sequences submitted to Genbank along with a list of the sequences, primers and *Anopheles gambiae* mosquito organs used for the analysis.

Legends for Supplementary Tables

Supplementary Table 1. The list of peptides identified from organ-wise proteomic analysis. The peptides are designated as JHU_Ag_XXXX where JHU refers to Johns Hopkins University and Ag refers to *Anopheles gambiae*. The peptide data was searched against protein database of *Anopheles gambiae* (genebuild AgamP3.6). The table shows the JHU peptide ID, which denotes the peptide identifier displayed in the VectorBase genome browser. Genome coordinates are listed by tblastn analysis of peptide data against *Anopheles gambiae* genome sequence. Peptides can be viewed as separate tracks on VectorBase genome browser using the URL http://funcgen.VectorBase.org/gdav/das as DAS server and JHU_Ag_v2 as data source.

Supplementary Table 2. A complete list of proteins identified from *Anopheles gambiae*.

Supplementary Table 3. Classification of 2,682 genome specific search peptides (GSSPs). A) The list of 1,873 genome search specific peptides (GSSPs) mapping to intergenic regions of VectorBase genes B) The list of 489 genome search specific peptides mapping to intronic regions of VectorBase genes C) The list of 43 genome search specific peptides mapping to exonic regions of VectorBase genes D) The list of 104 genome search specific peptides mapping to exon-intron junctions of VectorBase genes E) The list of 59 genome search specific peptides mapping to gene boundaries of the VectorBase genes; and, F) The list of 114 genome search specific peptides mapping to UTRs of VectorBase genes

Supplementary Table 4. Confirmation of splice sites with peptides spanning exon-exon junctions. The list contains a total of 4,106 unique peptides corresponded to exon-exon junctions of VectorBase transcripts.

Supplementary Table 5. Confirmation of novel splice sites with peptides spanning hypothetical exon-exon junctions. The list consists of 83 unique peptides that mapped to novel exon-exon junctions of VectorBase transcripts.

Supplementary Table 6. The list of 616 N-terminally acetylated unique peptides with their corresponding protein entries.

Supplementary Table 7. The list of peptides that directed the RT-PCR validation and sequencing of novel coding regions which are absent in AgmaP3.6 genebuild.

Legends for Supplementary Figures

Supplementary Figure 1. A schematic outlining of the strategy used for proteomic analysis of *Anopheles gambiae*. Protein lysates from larvae, pupae and adult mosquito tissues (head, salivary gland, malpighian tubules, ovary, midgut, viscera and male accessory gland and testis) were subjected to trypsin digestion. LC-MS/MS analysis was carried out using a high resolution Fourier transform mass spectrometer. The mass spectrometry-derived data were searched against a six-frame translated protein database, splice variant protein database and protein database of *Anopheles gambiae* (AgamP3.6 build). Bioinformatics analysis was carried out to further characterize the peptides uniquely identified from the genome search. Selected novel and corrected gene structures were further validated by RT-PCR and sequencing.

Supplementary Figure 2. A) Correlation of counts of number of peptides per protein with the number of proteins in each category. **B)** Distribution of peptide coverage per protein from nine different tissues.

Supplementary Figure 3. Coverage of sequence by mass spectrometry data. A) AGAP009323-PA was identified with 100% sequence coverage including an N-terminal acetylated peptide. B) Extensive sequence coverage of AGAP003153 (VATA_ANOGA) gene is shown that is based on 69 unique peptides that validated each of its five exons and four junctional peptides that confirmed all exon-exon junctions.

Supplementary Figure 4. Change in the structure of the AGAP010657 gene based on GSSPs. A) Seven peptides mapped to intergenic and intragenic region of the AGAP010657 gene. The SNAP prediction model supported a C-terminal exon extension of AGAP010657 and the presence of novel exon in the intronic region of gene AGAP010657. B) The MS/MS spectra of two genome search specific peptides, DYFGFLEVVAR and DVLIFHSDEFK, are shown.

Supplementary Figure 5. Joining of two genes. A) A single gene structure was identified in place of two separate predicted genes using 11 intergenic peptides mapped between the two Ensembl annotated genes, AGAP011872 and AGAP011873. B) Representative MS/MS spectra for peptides SDVLISGLGGLGVEVAK and HNIALIVADTR supporting the novel gene are shown.

Supplementary Figure 6. Identification of novel genes using peptides identified in intergenic regions. A) Thirteen peptides were identified in the intergenic region on chromosome 2L where intron of VectorBase gene model AGAP007548-RB is annotated on the opposite strand. Presence of a novel gene is also predicted by SNAP prediction program. B) Representative MS/MS spectra for peptides THFYNATQSVGENVTQWY and FGSNLEAILLDK supporting novel gene are shown.

Supplementary Figure 7. Change in the gene structure using peptide in the intron of the annotated gene. A) Fifteen peptides were identified in 3 different introns of the gene AGAP010021 which suggest change in the gene structure. These intronic peptides and other intergenic peptides support longer gene model supported the SNAP and Fgenesh models. B)

Representative MS/MS spectra for peptides LCRPVCANDQSCLNNER and ASCSCFSGMVPSPTAK supporting change in gene structure are shown.

Supplementary Figure 8. Confirmation of splice sites from peptides spanning exon-exon junctions. Peptides identified in protein database search but not detected in tblastn search were further investigated. In the case of the AGAP011026 gene, four such exon spanning peptides were mapped to exon1-exon2, exon4-exon5, exon5-exon6 and exon6-exon7 junctions. Exons are marked in alternate black and blue letters to distinguish the junctions.

Supplementary Figure 9. Confirmation of splice sites from peptides spanning exon-exon junctions. The AGAP002350 gene (Q3B723_ANOGA) encoding troponin-T exists in 5 different isoforms. Five splice junction peptides are shown thap map all 5 splice variants.

Supplementary Figure 10. Identification of a novel splice form for the AGAP003656 gene. A peptide, DCSDGEDEICEAQR, was identified in the MS/MS ion search against a synthetic database of potential splice isoforms which indicates a novel splice event occurring between exons 3 and 6 of the annotated gene. The MS/MS spectrum, which led to identification of the novel exon-exon junction, DCSDGEDEICEAQR, is shown.

Supplementary Data. 1) Details of the contribution of proteomic data to VectorBase. 2) Selected MS/MS spectra of genome search specific peptides. 3) Details of sequences submitted to Genbank along with a list of the sequences, primers and *Anopheles gambiae* mosquito organs used for the analysis.

Legends for Supplementary Tables

Supplementary Table 1. The list of peptides identified from organ-wise proteomic analysis. The peptides are designated as JHU_Ag_XXXX where JHU refers to Johns Hopkins University and Ag refers to *Anopheles gambiae*. The peptide data was searched against protein database of *Anopheles gambiae* (genebuild AgamP3.6). The table shows the JHU peptide ID, which denotes the peptide identifier displayed in the VectorBase genome browser. Genome coordinates are listed by tblastn analysis of peptide data against *Anopheles gambiae* genome sequence. Peptides can be viewed as separate tracks on VectorBase genome browser using the URL http://funcgen.VectorBase.org/gdav/das as DAS server and JHU_Ag_v2 as data source.

Supplementary Table 2. A complete list of proteins identified from *Anopheles gambiae*.

Supplementary Table 3. Classification of 2,682 genome specific search peptides (GSSPs). A) The list of 1,873 genome search specific peptides (GSSPs) mapping to intergenic regions of VectorBase genes B) The list of 489 genome search specific peptides mapping to intronic regions of VectorBase genes C) The list of 43 genome search specific peptides mapping to exonic regions of VectorBase genes D) The list of 104 genome search specific peptides mapping to exon-intron junctions of VectorBase genes E) The list of 59 genome search specific peptides mapping to gene boundaries of the VectorBase genes; and, F) The list of 114 genome search specific peptides mapping to UTRs of VectorBase genes

Supplementary Table 4. Confirmation of splice sites with peptides spanning exon-exon junctions. The list contains a total of 4,106 unique peptides corresponded to exon-exon junctions of VectorBase transcripts.

Supplementary Table 5. Confirmation of novel splice sites with peptides spanning hypothetical exon-exon junctions. The list consists of 83 unique peptides that mapped to novel exon-exon junctions of VectorBase transcripts.

Supplementary Table 6. The list of 616 N-terminally acetylated unique peptides with their corresponding protein entries.

Supplementary Table 7. The list of peptides that directed the RT-PCR validation and sequencing of novel coding regions which are absent in AgmaP3.6 genebuild.

Legends for Supplementary Figures

Supplementary Figure 1. A schematic outlining of the strategy used for proteomic analysis of *Anopheles gambiae*. Protein lysates from larvae, pupae and adult mosquito tissues (head, salivary gland, malpighian tubules, ovary, midgut, viscera and male accessory gland and testis) were subjected to trypsin digestion. LC-MS/MS analysis was carried out using a high resolution Fourier transform mass spectrometer. The mass spectrometry-derived data were searched against a six-frame translated protein database, splice variant protein database and protein database of

Anopheles gambiae (AgamP3.6 build). Bioinformatics analysis was carried out to further characterize the peptides uniquely identified from the genome search. Selected novel and corrected gene structures were further validated by RT-PCR and sequencing.

Supplementary Figure 2. A) Correlation of counts of number of peptides per protein with the number of proteins in each category. **B)** Distribution of peptide coverage per protein from nine different tissues.

Supplementary Figure 3. Coverage of sequence by mass spectrometry data. A) AGAP009323-PA was identified with 100% sequence coverage including an N-terminal acetylated peptide. B) Extensive sequence coverage of AGAP003153 (VATA_ANOGA) gene is shown that is based on 69 unique peptides that validated each of its five exons and four junctional peptides that confirmed all exon-exon junctions.

Supplementary Figure 4. Change in the structure of the AGAP010657 gene based on GSSPs. A) Seven peptides mapped to intergenic and intragenic region of the AGAP010657 gene. The SNAP prediction model supported a C-terminal exon extension of AGAP010657 and the presence of novel exon in the intronic region of gene AGAP010657. B) The MS/MS spectra of two genome search specific peptides, DYFGFLEVVAR and DVLIFHSDEFK, are shown.

Supplementary Figure 5. Joining of two genes. A) A single gene structure was identified in place of two separate predicted genes using 11 intergenic peptides mapped between the two Ensembl annotated genes, AGAP011872 and AGAP011873. B) Representative MS/MS spectra

for peptides SDVLISGLGGLGVEVAK and HNIALIVADTR supporting the novel gene are shown.

Supplementary Figure 6. Identification of novel genes using peptides identified in intergenic regions. A) Thirteen peptides were identified in the intergenic region on chromosome 2L where intron of VectorBase gene model AGAP007548-RB is annotated on the opposite strand. Presence of a novel gene is also predicted by SNAP prediction program. B) Representative MS/MS spectra for peptides THFYNATQSVGENVTQWY and FGSNLEAILLDK supporting novel gene are shown.

Supplementary Figure 7. Change in the gene structure using peptide in the intron of the annotated gene. A) Fifteen peptides were identified in 3 different introns of the gene AGAP010021 which suggest change in the gene structure. These intronic peptides and other intergenic peptides support longer gene model supported the SNAP and Fgenesh models. B) Representative MS/MS spectra for peptides LCRPVCANDQSCLNNER and ASCSCFSGMVPSPTAK supporting change in gene structure are shown.

Supplementary Figure 8. Confirmation of splice sites from peptides spanning exon-exon junctions. Peptides identified in protein database search but not detected in tblastn search were further investigated. In the case of the AGAP011026 gene, four such exon spanning peptides were mapped to exon1-exon2, exon4-exon5, exon5-exon6 and exon6-exon7 junctions. Exons are marked in alternate black and blue letters to distinguish the junctions.

Supplementary Figure 9. Confirmation of splice sites from peptides spanning exon-exon junctions. The AGAP002350 gene (Q3B723_ANOGA) encoding troponin-T exists in 5 different isoforms. Five splice junction peptides are shown thap map all 5 splice variants.

Supplementary Figure 10. Identification of a novel splice form for the AGAP003656 gene. A peptide, DCSDGEDEICEAQR, was identified in the MS/MS ion search against a synthetic database of potential splice isoforms which indicates a novel splice event occurring between exons 3 and 6 of the annotated gene. The MS/MS spectrum, which led to identification of the novel exon-exon junction, DCSDGEDEICEAQR, is shown.

Supplementary Data. 1) Details of the contribution of proteomic data to VectorBase. 2) Selected MS/MS spectra of genome search specific peptides. 3) Details of sequences submitted to Genbank along with a list of the sequences, primers and *Anopheles gambiae* mosquito organs used for the analysis.

Legends for Supplementary Tables

Supplementary Table 1. The list of peptides identified from organ-wise proteomic analysis. The peptides are designated as JHU_Ag_XXXX where JHU refers to Johns Hopkins University and Ag refers to *Anopheles gambiae*. The peptide data was searched against protein database of *Anopheles gambiae* (genebuild AgamP3.6). The table shows the JHU peptide ID, which denotes the peptide identifier displayed in the VectorBase genome browser. Genome coordinates are listed by tblastn analysis of peptide data against *Anopheles gambiae* genome sequence. Peptides can be

viewed as separate tracks on VectorBase genome browser using the URL http://funcgen.VectorBase.org/gdav/das as DAS server and JHU_Ag_v2 as data source.

Supplementary Table 2. A complete list of proteins identified from *Anopheles gambiae*.

Supplementary Table 3. Classification of 2,682 genome specific search peptides (GSSPs). A) The list of 1,873 genome search specific peptides (GSSPs) mapping to intergenic regions of VectorBase genes B) The list of 489 genome search specific peptides mapping to intronic regions of VectorBase genes C) The list of 43 genome search specific peptides mapping to exonic regions of VectorBase genes D) The list of 104 genome search specific peptides mapping to exon-intron junctions of VectorBase genes E) The list of 59 genome search specific peptides mapping to gene boundaries of the VectorBase genes; and, F) The list of 114 genome search specific peptides mapping to UTRs of VectorBase genes

Supplementary Table 4. Confirmation of splice sites with peptides spanning exon-exon junctions. The list contains a total of 4,106 unique peptides corresponded to exon-exon junctions of VectorBase transcripts.

Supplementary Table 5. Confirmation of novel splice sites with peptides spanning hypothetical exon-exon junctions. The list consists of 83 unique peptides that mapped to novel exon-exon junctions of VectorBase transcripts.

Supplementary Table 6. The list of 616 N-terminally acetylated unique peptides with their corresponding protein entries.

Supplementary Table 7. The list of peptides that directed the RT-PCR validation and sequencing of novel coding regions which are absent in AgmaP3.6 genebuild.