SNPHunter Input/Output Overview:

SNPHunter, version 2.21:

SNPHunter searches an annotated, *de novo* assembly alignment file for high quality single nucleotide polymorphisms (SNPs) by using a series of adjustable search parameters and custom filters to exclude suspect assembly regions (e.g. a region with polymorphism due to an alternative splice site). As of version 2.10, these filters are generated (as a text file) separately by the new AssemblyFilter script in order to streamline and speedup SNPHunter execution. The primary SNPHunter/AssemblyFilter input is an assembly alignment file such as that generated by the AceParser script from assembly Ace formatted output (i.e. Phred/Phrap output). This assembly alignment file can contain embedded annotation, allowing for frame orientation (see AceParser doc). SNPHunter also requires (as of version 2.00) quality score input from a padded quality score file (i.e. padded with zero's to match aligned assembly sequence lengths), also generated by AceParser. Requiring quality score input allows generation of SNP p-values and consensus sequence quality statistics, as well as simple filtering of assembly nucleotides by quality score during the SNP calling process. As of version 2.21, SNPHunter automatically reads and writes to a condensed quality score file format in order to cut down these potentially huge file sizes by as much as 50% (see also AceParser docs). Additional, optional inputs are formatted consensus ORFs (generated through PipeMeta) to provide an alternate source of framing information and a tab-delimited population key file for generating population specific output columns. Output is one row per SNP and consists of tab-delimited columns loosely divided into basic Consensus statistics, framing information, general SNP information, population specific SNP information, modified sequence output, and annotation. There are several customizable search and filter parameters, including filters for homopolymer runs, alternative splicing/clustered polymorphism, nucleotide depth, and presence of annotation that are available using AssemblyFilter script, and a read trimming filter and quality score filter available directly from within SNPHunter. Additional basic SNP search parameters are explained below.

Input: description of all possible input files for SNPHunter. Assembly alignment and padded quality score files are required by SNPHunter, all others files are optional.

- Assembly file: Tab-delimited, annotated, *de novo* assembly alignment file. A file containing the padded (i.e. aligned with hyphens for spacing) assembly consensus Contig sequences with optional integrated Uniprot annotation, followed by their respective aligned ESTs. The annotation (if available) consists of the following fields: Blast available/unavailable, Blast Contig frame, number of gaps, Blast alignment range, query start, query end, best Blast subject name (not always Uniprot), best Blast bitscore, Uniprot primary name (i.e. Uniprot ID), Uniprot description, species, taxonomy, GO terms, Pfam, KEGG KO (i.e. KEGG ID), KEGG levels (i.e. hierarchy IDs and descriptions). The file can be created from an Ace file using the AceParser script, which is run from PipeMeta or from the command line (see AceParser help for details).
    - E.g. single ass file entry:

Contig_24

GAAGCTATAGTATTTTGCAATTTTAACCAATTATACAAAACGGATCCAAAAAT
TGTTGCAACATGGGTTAAGATTTTaAAAAAAGTACCAAACAGTGTTCTCTGGCTTTTAA
GTTTTCCAGTTGCAGGGGAACGAAATTTACAAAAATACGTTCGAAGTTTAGGAATATC
G 1 1 0 162 1 162 HEC00418_1 86.3

B4HCC1_DROPE SubName: Full=GL13214; Drosophila persimilis [Fruit fly] Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Ephydroidea; Drosophilidae; Drosophila; Sophophora.

GO:0005488,F:binding,IEA:InterPro. K09667 01100: Metabolism --> 01107: Glycan Biosynthesis and Metabolism --> 00512: O-Glycan biosynthesis --> EC:2.4.1.41: polypeptide N-acetylglucosaminyltransferase, OGT;

==========

008052_2922_1845_ECKAT2S12HEUN7

GAAGCTATAGTATTTTGCAATTTTAACCAATTATACAAAACGGATCCAAAAAT
TGTTGCAACATGGGTTAAGATTTTAAAAAAAGTACCAAACAGTGTTCTCTGGCTTTTAA
GTTTTCCAGTTGCAGGGGAACG----------

010218_3127_1413_ECKAT2S13HWT8N

----------GGTTAAGATTTT-AAAA
AAGTACCAAACAGTGTTCTCTGGCTTTTAAGTTTTCCAGTTGCAGGGGAACGAAATTTA
CAAAAATACGTTCGAAGTTTAGGAATATCG

****

- Padded quality score file: Required file containing the padded assembly EST quality scores. Used to exclude low quality (below minimum threshold) nucleotides from SNP calling process and during SNP quality statistical analysis. The padded quality score file can be generated by the Aceparser script from the quality score file used in the assembly process. Basically, the padded quality score file is a very large file of 'aligned' quality scores (relative to their corresponding padded ESTs) padded with zeroes instead of hyphens. As of SNPHunter version 2.21 and AceParser version 1.33, quality score files are written/read using a condensed format to help limit file sizes.

- Filter file: Optional tab-delimited, two-column file containing Contig names and a sequences of ones and zeros denoting filtered nucleotides and non-filtered nucleotides respectively. As of SNPHunter version 2.10, filtering of the assembly is performed by a separate script called AssemblyFilter in order to improve run-time, provide more accurate statistics (with regard to the effects of filtering on the assembly), and streamline SNPHunter code.

- ORF file: Optional tab-delimited protein sequence file used to help predict codon

positions for SNPs. Created through Pipemeta using the 'Find ORFs' menu option in 'Post Assembly Menu -> SNP File Menu'. Pipemeta creates the file by running the Translator script for all six frames of the Contig sequences and then keeping the largest ORF (or ORFs in the case of ties for largest) for each Contig. The resulting ORF sequences contain coordinates in the titles linking back to the contigs, as well as frame and length information used to calculate SNP positions by SNPHunter.

- E.g. single ORF file entry:

  Contig_10087_Frame:-3_Pos1:0-131_Frame:3_Pos1:120-U248_Len:43
  EWLRQSEEQKPTYLRRANAQKKTIRLGTVNTHIFDQHTASLT*
  MLEMRYVDRKCACLLCQAVLFSSAHLHVEGTLVSVPQTVSAIP

- Population Key file: a simple tab-delimited optional file used to determine population origin of tagged ESTs in the ass file. The file should consist of two columns (tab-delimited), one row per population. The first column should contain the population names and the second the corresponding population EST tag. The assembly Est/read sequences should be tagged by attaching a short, alpha-numeric ID string to the ends of the titles (e.g. >SeqName_A).

  - E.g. pop key file:

    AdultpopK _K

    AdultpopM _M

    AdultpopP _P

    Larvalpop _L

Filters and search parameters: SNPHunter uses several search parameters to call SNPs, as well as several optional filters (also with parameters), read from a filter file created by the AssemblyFilter script, to screen out suspect or non-desirable regions within an assembly. Most of the filter options described here are no longer available from SNPHunter (they are run prior to SNPHunter using AssemblyFilter), but have been kept in this overview for user convenience.

Search Parameters (SNPHunter):

- *Minimum Minority Allele Number*: Minimum number of instances of the minority allele(s) present at the potential SNP position necessary in order to call the SNP. Default value is 2.

- *Minimum SNP Depth*: Minimum number of EST nucleotides at the potential SNP position (i.e. nucleotide depth) required in order for a SNP to be called. Default value

is 4.

- *Minimum Quality Score*: Minimum quality score in order for an assembly EST/read nucleotide to be considered during the SNP calling process. Default value is 14 ($P < 0.05$).

- *Alpha*: Parameter to determine allowable error probability at SNP P-value field. Higher SNP P-values will cause individual allele P-values to be calculated (see below for details). Default parameter value is 0.01.

Filters (SNPHunter):

- *EST/Read End Filter*: Filters out a given number of nucleotides from the ends of all assembly ESTs/reads. Adds additional stringency to the alternative splicing filter by preventing these end nucleotides (which are often ignored during the assembly process) from counting towards SNP detection. The adjustable parameter determines the number of nucleotides to filter from each EST/read end (default value is 2).

  E.g. SNPHunter –i infile.ass –o outfile.txt –e 2

Filters (AssemblyFilter):

- *Blast Filter*: Filters out assembly alignment regions that are outside the range of their Blast alignments (or that don't have Blast alignments). This is useful when only framed SNPs are desired. This filter affects the output for all 'adjusted' descriptive statistics.

  E.g. AssemblyFilter –i infile.ass –o outfile.txt –b

- *Minimum Nucleotide Depth Filter:* Filters out assembly alignment regions that have EST nucleotide depths below a minimum value. The adjustable parameter is the minimum nucleotide depth number (no default value). This filter affects the output for all 'adjusted' descriptive statistics.

  E.g. AssemblyFilter –i infile.ass –o outfile.txt –n 10

- *Homopolymer Run Filter*: Filters out assembly alignment regions containing homopolymer runs above a minimum run value. Homopolymer runs are strings of repeat nucleotides (e.g. ATGGGGGGGGCC) and are known to have higher rates of sequencing errors in 454 pyrosequencing reads. The adjustable parameter is the minimum homopolymer run number (default value is 5). This filter affects the output for all 'adjusted' descriptive statistics.

  E.g. AssemblyFilter –i infile.ass –o outfile.txt –p 10

- *Alternative Splicing Filter*: Uses a pattern searching algorithm to filter out polymorphic assembly regions suspected of occurring due to alternative splicing, assembly error, or

other non-SNP mechanisms common within *de novo* transcriptome assemblies. The adjustable filter parameters are seed size and cluster spacing. Seed size is the number of polymorphisms clustered together required to trigger the filter (default setting is 2). Cluster spacing is the space allowed between clusters in order to be considered part of the excluded region (no default setting, but recommended setting is 4). This filter affects the output for all 'adjusted' descriptive statistics.

E.g. AssemblyFilter –i infile.ass –o outfile.txt –a 4,2

Output: description of SNPHunter tab-delimited output fields. Exact output varies depending on options specified, so not all fields may be present. Each row corresponds to a single SNP or to an 'empty' result for Contigs lacking a SNP call (unless the 'minimum output' option is specified at runtime).

- *Contig Name*: Consensus Contig sequence name (e.g. Contig_100). One or more rows (one row per SNP or 'empty' result).
- *Number of SNPs in Contig*: Total number of SNPs found within the Contig assembly alignment matching all search parameters and after all filters and quality controls.
- *Contig Padded Length*: Length of the Contig including all internal alignment spacing (padding).
- *Contig Trim Length*: Length of the Contig without any alignment spacing. Used to calculate raw statistics.
- *Nucleotide Count*: Total number of Contig nucleotides with one or more EST/read nucleotides above the quality score cut-off, excluding those in filtered regions. Basically, this number represents the remaining pool of candidate consensus nucleotides used to search for SNPs by SNPHunter after all quality control and filtering.
- *Flag: Blast Info*: Binary flag denoting availability of Blast annotation (1=Blast available, 0=Blast unavailable).
- *Orientation, from Blast*: Contig orientation (i.e. sense or antisense strand) in the original sequence file and ace file, according to best available Blast alignment. **NOTE: Assembly (.ass) file sequences are modified to always be in the sense orientation when Blast information is available.**
- *Codon Position, from Blast*: SNP codon position according to best available Blast alignment. Codon positions are calculated from best Blast frame information and Contig length.
- *Flag: SNP within Blast Range*: Binary flag showing whether or not SNP is within the

Blast alignment range (i.e. between query start and end coordinates). Codon position estimates for SNPs within the Blast range have a higher degree of certainty.

- *Peptide String, from Blast*: Corresponding amino acids of the Blast predicted codons generated from the SNP. Amino acids appear in the peptide string in order of highest allele frequency to lowest. The 'X' character represents codons that could not be translated due to non-SNP related degeneracy (e.g. 'yT[AT]', where [AT] is the SNP with 'A' and 'T' alleles and 'y' is another polymorphic site in the contig ['A' and 'C'] representing a rejected potential SNP).

- *Syn/Nonsyn, from Blast*: Field value can be Synonymous, Nonsynonymous, Degenerate, or n/a. A SNP is considered Degenerate when the Blast predicted codon has additional nonsynonomous polymorphism at other positions.

- *Orientation, from Largest ORF*: Contig orientation (i.e. sense or antisense strand) in the original sequence file and ace file, according largest ORF. Largest ORF orientation does not modify consensus sequence output like Blast orientation information does (i.e. output consensus sequences are only oriented relative to Blast data). It is possible for two or more ORFs to tie for largest, in which case two or more orientations will be output, delimited by semi-colon. **Optional field**.

- *Codon Position, from Largest ORF*: SNP codon position according to largest ORF detected in the Contig. Only available if an ORF file was created and incorporated in the SNP analysis (see above under Input). Codon positions are calculated from the ORF sequence frame information and Contig length. It is possible for two or more ORFs to tie for largest, in which case two or more codon positions will be, delimited by semi-colon (e.g. 1;2). **Optional field.**

- *Flag: SNP within ORF Range*: Binary flag showing whether or not SNP is within the ORF range (i.e. between query start and end coordinates). Assuming the largest detectable ORF is actually the real, full length reading frame, codon position calls outside the ORF are meaningless. However, in many cases the actual reading frame length will be longer (or sometimes shorter) than the largest detectable ORF, in which case a codon position call occurring outside of the ORF range may be valid. **Optional Field.**

- *Peptide String, from Largest ORF*: Corresponding amino acids of the largest ORF predicted codons generated from the SNP. Amino acids appear in the peptide string in order of highest allele frequency to lowest. The 'X' character represents codons that could not be translated due to non-SNP related degeneracy (e.g. 'yT[AT]', where [AT] is the SNP with 'A' and 'T' alleles and 'y' is another polymorphic site in the contig ['A' and 'C'] representing a rejected potential SNP). Multiple peptide string from different frames (from a tie for largest ORF) are delimited by semi-colon. **Optional Field.**

- *Syn/Nonsyn, from Largest ORF*: Field value can be Synonymous, Nonsynonymous,

Degenerate, or n/a. A call of Nonsynonymous occurs when any predicted largest ORF produces an amino acid change (in cases of a tie for largest ORF). A call of Synonymous occurs when all predicted ORFS fail to produce an amino acid change. A SNP is considered Degenerate when one or more predicted largest ORF produces a codon that fails to translate due to additional polymorphism at other codon positions. **Optional Field.**

- *Flag: Codon Position Match*: Binary flag showing whether or not the fields *Codon position from Blast* and *Codon position from largest ORF* are in agreement with each other. In cases where more than one 'largest' ORF is present, agreement occurs if any of the ORF codon positions match the estimated Blast codon position. **Optional Field.**

- *Gaps*: Number of Gaps present *in best Blast alignment*. Best blast alignment is used to calculate codon position; therefore any gaps in the alignment throw the estimated codon position call into serious question. In such cases, having the codon position calculated from the largest ORF match the Blast codon position can help validate the codon position call.

- *Trim Blast Start*: Best Blast alignment query start position. Coordinate refers to the position on the trim (unpadded) Contig where alignment with the best subject begins.

- *Trim Blast End*: Best Blast alignment query end position. Coordinate refers to the position on the trim (unpadded) Contig where alignment with the best subject ends.

- *Trim Blast Length*: Length of Blast alignment between Contig and best subject.

- *Longest ORF Length*: Length of the longest predicted ORF nucleotide sequence(s). Useful when comparing best Blast results to longest predicted ORF results (i.e. when the results contradict each other).

- *Average Coverage Depth, Raw*: Average depth of EST/read coverage in the Contig's raw assembly alignment (i.e. before quality filtering). Raw average coverage depth is calculated by taking the number of EST/read nucleotides at each nucleotide position on the Contig and dividing by the *Trim Contig Length*. Nucleotides at padded sites within the aligned assembly (i.e. hyphens in the consensus Contig) are not considered.

- *Average Coverage Depth, Adjusted*: Average depth of EST/read coverage in the Contig's adjusted assembly alignment (i.e. after filters). Adjusted average coverage depth is calculated by taking the EST/read nucleotides above the minimum quality score threshold at each non-filtered Contig position and dividing by the *Nucleotide Count*. Nucleotides at padded sites within the aligned assembly (i.e. hyphens in the consensus Contig) are not considered.

- *Average Quality Score, Raw (SD)*: Average quality score of the Contig assembly

alignment calculated before quality filtering, and its standard deviation in parentheses. Raw Contig average quality score is calculated from the summed average scores at each non-pad nucleotide position divided by the *Trim Contig Length*.

- *Average Quality Score, Adjusted (SD)*: Average quality score of the adjusted Contig assembly alignment (i.e. after filtering), and its standard deviation (SD) in parentheses. Adjusted Contig average quality score is calculated from the summed average scores above the minimum quality threshold at each non-pad and non-filtered nucleotide position, divided by the *Nucleotide Count*.

- *SNP Name (Depth/Relative Frequency)*: Allele names and minority allele nucleotide depths and relative frequencies (in parentheses). Alleles appear in order of relative abundance from major allele on the left, to lowest frequency on the right. Each minority allele is followed by its corresponding nucleotide depth at the SNP position and its relative frequency in parentheses (e.g. GT(33/0.33)C(5/.05)).

- *Nucleotide Depth*: Depth of EST coverage at the position of the SNP. Nucleotides below the minimum quality score threshold are not included.

- *Average SNP Quality, Adjusted*: **Adjusted** average quality score at the SNP position.

- *Average SNP Quality SD, Adjusted:* **Adjusted** average quality score standard deviation at the SNP position.

- *SNP P-value*: Probability of at least one or more base calls being an error at the SNP position.

- *Allele P-value*: Probability of *all* allele-specific base calls being errors at the SNP position, listed separately by allele. Probabilities are listed (separated by semi-colons) in order of highest to lowest allele frequency.

- *Padded Location*: Position of the SNP on the aligned (with internal spacing) consensus Contig.

- *Trim Location*: Position of the SNP on the trim (without internal spacing) consensus Contig.

- *Population Fields*: Each population specified in the population key file gets five additional fields. **Optional output.**

    - *Population Specific Average Coverage Depth, Adjusted*: Average **adjusted** coverage depth for the Contig alignment, but specific to the given population.

    - *Population Specific Contig Length, Adjusted*: Number of **adjusted** Contig nucleotides with representative population specific EST/read nucleotides.

- *Population Specific Average Quality Score, Adjusted*: Average **adjusted** quality score for the Contig alignment, but specific to the given population.

- *Population SNP Name (Depth/Relative Frequency)*: Population specific alleles in order of frequency. Minority alleles are followed by population specific allele depth and relative frequencies in parentheses.

- *Population Depth*: Total number of EST nucleotides at the SNP position belonging to the population, excluding those below the minimum quality score cut-off.

- *Flag: Population SNP*: Binary flag denoting the presence or absence of a population specific SNP (i.e. a SNP present within the population). Useful for finding SNPs present within one population, but not another.

- *Trim Contig Sequence*: Contig sequence without internal assembly alignment spacing. The sequence is modified in several ways from the original in the Ace and Contig file. If Blast annotation is available it has been oriented to be sense strand. The sequence is reverse complimented if the Blast query vs. subject orientation is -/+ or just reversed if the Blast is -/-. The Contig Frame is also adjusted so that the sequence starts in position one (i.e. one or two nucleotides are snipped off from a sequence end). Regions excluded from the SNP search algorithm by the SNPHunter filters (e.g. alternative splicing filter) are represented by lowercase nucleotides, as are polymorphic sites that do not meet the minimum SNP parameters. Uppercase standard nucleotides and uppercase degenerate nucleotides are included Contig regions (added to Nucleotide Count) and SNPs that made it through the SNP calling process, respectively.

- *Padded Contig Sequence*: Contig sequence with internal assembly alignment spacing still intact. The sequence is modified in several ways from the original in the Ace and Contig file. If Blast annotation is available it has been oriented to be sense strand. The sequence is reverse complimented if the Blast query vs. subject orientation is -/+ or just reversed if the Blast is -/-. The Contig Frame is also adjusted so that the sequence starts in position one (i.e. one or two spaces are added to a sequence end). Regions excluded from the SNP search algorithm by the SNPHunter filters (e.g. alternative splicing filter) are represented by lowercase nucleotides, as are polymorphic sites that do not meet the minimum SNP parameters. Uppercase standard nucleotides and uppercase degenerate nucleotides are for included Contig regions (added to Nucleotide Count) and SNPs that made it through the SNP calling process, respectively. The padded Contig sequence is useful for comparison to the Contig assembly alignment (e.g. it can be cut and paste directly into the assembly alignment in BioEdit from Pipemeta).

- *Best Blast Subject*: Subject name of the top hit from the Blast analysis with the best bitscore. Not necessarily the same Blast subject as the one used to derive the Uniprot annotation if a multiple subject field Blast table was used during the Ass file creation

process (see AceParser docs)! In such a case the subject is often from a Blast alignment versus a reference sequence set from a closely related species not found in Uniprot. All Blast data (e.g. codon positions) used in the various Blast related fields are derived from the Best Blast Subject.

- *Best Blast Bitscore*: Bitscore from the best blast subject alignment.
- AnnotationBlastBitscore: Bitscore from the best annotation Blast (i.e. Uniprot/Uniref).
- *Annotation Subject*: Subject name (actually the primary subject name) of the top hit from the Uniprot Blast analysis, regardless of whether or not the Uniprot subject is the best blast alignment for the contig. Allows best available Uniprot annotation for the Contig to be included in the SNP table.
- *Description*: Uniprot annotation description of the Contig, derived from Blast analysis of Contig versus the Uniprot database.
- *Species*: Species binomial name derived from Uniprot. Common name is included in brackets if available.
- *Taxonomy*: taxonomic classification of Contig derived from Uniprot. Useful for sorting SNPs by various taxonomic divisions.
- *GO Terms*: Gene Ontology terms. The Gene Ontology project is a major bioinformatics initiative with the aim of standardizing the representation of gene and gene product attributes across species and databases. The GO project has developed three structured controlled vocabularies (ontologies) that describe gene products in terms of their associated biological processes, cellular components and molecular functions in a species-independent manner. For more information, visit http://www.geneontology.org/GO.doc.shtml . Each of the three GO categories can have up to one entry per row, separated by semi-colons and denoted by its representative letter (i.e. p=biological processes, c=cellular components, and f=molecular functions). Each entry consists of a GO term ID, definition/description, and evidence source (e.g GO:0033180, C:proton-transporting V-type ATPase, V1 domain, IEA:InterPro.). Each of the three possible entries can be thought of as the tip of a three branched annotation tree, where the tips are the most specific available terms in an annotation hierarchy that starts at very general descriptions near the trunk and gets more specific near the tips.
- *KEGG KO*: KEGG (Kyoto Encyclopedia of Genes and Genomes) orthology object ID. KEGG is an integrated database resource consisting of 16 linked databases, broadly categorized into systems information, genomic information, and chemical information. KEGG has been widely used as a reference knowledge base for biological interpretation of large-scale datasets generated by sequencing and other high-throughput experimental technologies. For more information on KEGG, visit http://www.genome.jp/kegg/kegg1.html . KEGG KOs represent an instance of the molecular interaction/reaction network, which is stored as a collection of pathway maps (graphical diagrams) in the PATHWAY database. Reflecting the map resolution, KEGG

PATHWAY is organized in a hierarchy. The KEGG KO allows the sequence annotation information pathway to be identified in the KEGG hierarchy.

- *KEGG Levels*: The Contig's associated KEGG pathway/ontology through the molecular interaction/reaction network (identified by the KEGG KO). Represented as a hierarchy of annotations, from most general at the top level, to most specific at the fourth (bottom) level. Some Contigs will have more than one pathway through the hierarchy (separated by semi-colon), representing multiple functions or simply multiple perspectives on the gene (e.g. molecular vs. disease entry). Each level node in the pathway contains an ID followed by a description. Levels are connected by an arrow symbol.
    - *E*.g. KEGG levels:

      01140: Cellular Processes --> 01151: Transport and Catabolism --> 04144: Endocytosis --> EC:2.7.11.16: G protein-coupled receptor kinase; 01140: Cellular Processes --> 01145: Immune System --> 04062: Chemokine signaling pathway --> EC:2.7.11.16: G protein-coupled receptor kinase;

Contact Information: please feel free to contact for questions/comments and for reporting bugs/issues.

Author: J. Cristobal Vera

Email: jcv128@psu.edu

Phone: 814-863-5927