TITLE

FEATnotator: A Tool for Integrated Annotation of Sequence Features and Variation,
Facilitating Interpretation in Genomics Experiments

AUTHORS

Ram Podicheti^{*a,b} and Keithanne Mockaitis^c

^a Center for Genomics and Bioinformatics, Indiana University, 1001 E Third Street,
Bloomington IN, USA 47405

mnrusimh@indiana.edu +1 812 856 1821

^b School of Informatics and Computing, Indiana University, 919 E Tenth Street,
Bloomington IN, USA 47408

^c Department of Biology, Indiana University, 915 E. Third Street, Bloomington IN, USA
47405

kmockait@indiana.edu

* Author for correspondence

ABSTRACT

As approaches are sought for more efficient and democratized uses of non-model and expanded model genomics references, ease of integration of genomic feature datasets is especially desirable in multidisciplinary research communities. Valuable genomic conclusions are often missed or slowed when researchers refer experimental results to a single reference sequence that lacks integrated pan-genomic and multi-experiment data in accessible formats. Association of genomic positional information, such as results from an expansive variety of next-generation sequencing experiments, with annotated reference features such as genes or predicted protein binding sites, provides the context essential for conclusions and ongoing research. When the experimental system includes polymorphic genomic inputs, rapid calculation of gene structural and protein translational effects of sequence variation from the reference can be invaluable. Here we present FEATnotator, a lightweight, fast and easy to use open source software program that

integrates and reports overlap and proximity in genomic information from any user-defined datasets including those from next generation sequencing applications. We illustrate use of the tool by summarizing whole genome sequence variation of a widely used natural isolate of *Arabidopsis thaliana* in the context of gene models of the reference accession. Previous discovery of a protein coding deletion influencing root development is replicated rapidly. Appropriate even in investigations of a single gene or genic regions such as QTL, comprehensive reports provided by FEATnotator better prepare researchers for interpretation of their experimental results. The tool is available for download at <http://featnotator.sourceforge.net>.

KEYWORDS:

SNP; Genome Features; Non-model; Genomic loci; Overlap

1 INTRODUCTION

Applications of next generation sequencing have made available a growing assemblage of assays for scientists across disciplines interested in probing genomic information (1-4). DNA re-sequencing (5, 6) followed by mapping of reads to a reference assembly enables detection of genomic variation in the form of single nucleotide polymorphisms (SNPs), multiple nucleotide polymorphisms (MNPs) including simple sequence repeats (SSRs), and insertions and deletions. RNA-Seq and transcript mapping similarly promotes discovery of RNA polymorphism as well as quantification of reference gene expression and genomic annotation of coding and noncoding transcribed regions (7-11). Chromatin-immunoprecipitation applied in ChIP-Seq and applications that include larger complex crosslinking prior to sequencing (*e.g.* chromatin conformation capture) allows genome-wide mapping of protein-DNA complexes or higher order structure in chromosomes (12-16). Best uses of each these applications rely upon contextualizing experimental results spatially among annotated genomic loci such as genes and genetic markers, and among results of other related sequencing experiments.

With increasing numbers and types of experimental datasets anchored to genomic

positions in a given reference, many researchers aim to compare results from a variety of experiments, alternate annotations, etc. to approach comprehensive conclusions. One of the most common needs being addressed today is the annotation and summary of single nucleotide polymorphism (SNP) information, as more and more contrasting whole genome sequence (WGS) datasets become available for association genetics and population genomics studies. Several different feature annotation software tools are available, most of which are designed for annotating SNPs (17-24) or genomic coverage peaks (25-27). SNP annotation tools are generally expected to classify SNPs according to their positions with respect to gene structure, and further when found in coding regions, based on their impacts upon protein products. User constraints such as input format requirements, available reference annotations, database preprocessing requirements, software environment requirements, program installation, processing speed, memory use and utility of output formats can limit employment of software for many researchers. For example, most currently available tools require sequence variant information to be provided in VCF format exclusively. Tools such as SnpEff (18) and AnnTools (21) require the annotation data to be preprocessed, the later even necessitating loading into a MySQL database. Packages such as PopGenome (22) that offer powerful data analytical and graphical capabilities operate only within the R environment. Other tools (17, 19, 24) are restricted to uses of genomic information in a single species, most commonly human. Variant Annotation Tool (28), a cloud-enabled SNP annotation tool, provides a means to visualize the functional impact of variants on a gene, however it requires several dependency libraries to be installed and configured which can be challenging. PAVIS (26) is a peak annotation tool that is limited to genomes with annotation information available through the UCSC genome browser. Although TRAMS (23) appears more flexible with respect to data input formats, it is applicable to only prokaryotic genomes. Also offering flexibility in input formats, Snpdat (20) additionally eliminates the need for much data preprocessing. However, this tool loads completely all SNP data and annotation information into memory prior to processing, thus requiring a larger memory footprint and rendering the annotation process very slow. More consequentially, in annotation trials we noticed lack of fundamental biology knowledge in the program, as Snpdat appears to assume erroneously that transcription start site positions (TSS)

provided in genome annotation determine a protein's reading frame for translation. Tools optimized to rapidly detect or retrieve genomic feature overlaps, such as Bedtools (29) and Tabix (30), are available, however their outputs require extensive further processing to attain integrated biological annotation. Bedtools, for example, reports feature overlap with a gene, a corresponding transcript, an exon and a CDS as independent rows lacking relational information among them. Finally, nearly all currently available tools able to accomplish single nucleotide annotation are designed exclusively to utilize datasets of sequence polymorphism and therefore expect input of reference and alternative alleles. The nature of some research would benefit from more generalized annotations that note loci without prior assumption of sequence characteristics. For example, Gent et al. (31) showed recently that CHH methylation sites in maize exhibit a gradient that correlates with proximity to genes. The methylation sites are single nucleotide loci and yet do not represent genomic variation.

We developed FEATnotator, an epithet for Feature Annotator, to provide a method for genomic feature annotation that operates in a manner that is considerably fast and user-friendly, and further eschews most common burdens for data preprocessing. The FEATnotator workflow is illustrated schematically in Figure 1. The tool annotates and integrates any genomic positional information derived from next-generation sequencing datasets and other traditional and contemporary experimental sources. Since it handles positional information that extends over a single nucleotide position to any length, data inputs may include but are not restricted to SNPs, transcribed regions, *cis* regulatory elements, peak calls from RNA sequence coverage or ChIP-seq experiments, or even results from sequence alignments such as BLAST.

FEATnotator maintains a very low memory footprint by not requiring complete loading of input data and reference annotation data all at once. It can be run with ease on a personal computer. The tool accepts tab delimited text input in any format, with the simple requirement that separate columns designate reference sequence name (*e.g.*, scaffold, chromosome or transcript) and position, such as VCF. FEATnotator generates user-sortable feature tables based on the nature of the input provided, and additionally

compiles total findings into informative summary tables and optional graphs. It is written in Perl and does not require any special modules or relational database support. Currently R (<http://www.r-project.org>) is required for the plotting of tabulated summary information into graphs. As the summary tables are provided in simple text format, a user without R may comfortably generate custom plots within an external graphing program of choice. Prior knowledge of computer programming is not required to run FEATnotator.

In this report we demonstrate the use and illustrate select outputs of FEATnotator by associating a publicly available and widely used SNP dataset with reference gene models and traditional genetic markers of a model plant system. We analyze SNPs determined previously in the Wassilewskija, or Ws, accession against the Columbia, or Col, genome reference sequence of the model plant *Arabidopsis thaliana*.

2 METHODS AND IMPLEMENTATION (MATERIALS AND METHODS)

2.1 USER INPUTS

2.1.1 FEATURE INPUT

Genomic feature inputs for FEATnotator are tab delimited text files in which each row represents a single locus. Scaffold names, loci start and end positions should be available in separate columns in any order, adjacent to any data a user wishes to integrate alongside. The column positions for these three fields are provided as input to the program. If loci of interest are single nucleotide positions, such as DNA methylation sites or single nucleotide indels, then a file containing only the two columns of scaffold and position will suffice as input. Any additional notation present in separate column(s) of the input file will be preserved in the output file.

For SNP analyses, sequence variant information can be specified in the input file either in one column representing the alternate allele, or in two columns that represent the major and minor alleles. When major and minor alleles are specified, a variation from the reference base is considered the alternate allele, and this can exist in either column. In

situations where both major and minor alleles differ from the reference base, the major allele is treated as the alternate allele. FEATnotator also accepts VCF files for variant sequence input, and when using this format there is no need for the user to specify the column positions described above.

2.1.2 ANNOTATION INPUT

Genome annotation information is provided as one or more files in GFF3, GTF or Genbank formats. FEATnotator converts each of the files to GFF3 format and combines them into a single file, thus allowing multiple features to be combined and analyzed, regardless of their source. The annotation information may include features such as gene models, markers, BLAST alignments or other features mapped to the reference sequence. The FEATnotator algorithm requires the input data to be sorted by scaffold name and position, however the program spares the users from this sorting task by implicitly sorting the supplied input data before processing them. When annotations are supplied as a single GFF3 file and pre-sorted, the user can bypass the sorting process, thereby reducing processing time.

2.1.3 REFERENCE SEQUENCE

In uses of FEATnotator for SNP analyses, a reference sequence in FASTA format is provided for determination of sequence substitutions consequences. The sequence IDs in the reference sequence file should match with those in the reference annotation file. Example data input files are available on the sourceforge project website along with software documentation.

2.2 QUERY IMPLEMENTATION

FEATnotator identifies positional overlap of input experimental features and reference genome annotation. Figure 2 illustrates the algorithm describing in detail how the input features are compared against the reference annotation. Input features are queried sequentially against data that are maintained in a “buffer space” of ordered genomic annotation information. Portions of the genomic reference previously fully queried are removed from the buffer space as new portions are added. In this manner input features

are compared against positional windows smaller than the total reference, thus efficiently utilizing computational time and space.

2.3 ANALYSES

2.3.1 HIGH IMPACT FEATURES

Feature overlap single nucleotide position indicated by FEATnotator as transcription start site (TSS) derives from the first base of annotation “gene” feature. Overlap with transcript splice sites consider two nucleotide positions on either side of the exon-intron splice boundary, as this is the core of canonical sites used commonly in eukaryotic gene model recognition algorithms (32, 33). In the proximity output, the number of nucleotides from a gene annotation is calculated as the closest distance to the nearest gene, in either orientation (*i.e.*, last or first nucleotide of total gene model annotation).

2.3.2 SNP INTERPRETATION

For analyses of translational impact in cases of coding SNPs, the codon affected by a SNP in the CDS is determined with the help of the “phase” (8th) and the “strand” (7th) columns of the standard reference annotation GFF3 file. The program uses standard genetic code (34) to categorize the coding SNPs into “NONSENSE”, “MISSENSE” and “SILENT” SNPs.

2.3.3 INDEL ANALYSIS

The InDel Analysis option computes the Insertion and Deletion lengths (value is positive for Insertion, negative for Deletion) and annotates InDel position. If the InDel overlaps with a coding region, the program predicts whether the variation causes a protein translational frameshift based on the length of the InDel. Both SNP and InDel input data can coexist in the same file, such as in a VCF file.

2.4 USER OUTPUT OPTIONS

Genomic coordinates encompassing sequence less than the complete reference may be specified for FEATnotator generation of outputs. Graphical outputs of summaries may be

specified if R is installed. A list of parameter option commands is provided in the FEATnotator manual.

2.5 EXAMPLE

In this report we demonstrate selected outputs of FEATnotator by integrating SNP and indel data of *Arabidopsis thaliana* Ws-0 provided by Gan et al. (2011), and by analyzing SNP consequences. For the annotation we used gene models of TAIR10 (www.arabidopsis.org). In a subsequent run additional annotation proceeded with 2010 genetic marker data provided on the same site.

3 RESULTS

The output of FEATnotator comprises multiple annotation and statistical summary tables as well as optional graphical output depicting the summary information. The complete software, along with example inputs and outputs, can be downloaded from <http://featnotator.sourceforge.net>.

3.1 ANNOTATION TABLES

A full summary of output Tables provided by FEATnotator are listed in Table 1. The primary annotation table, output.FEATnote.consolidated.txt, contains every entry of the input data file extended by integrated annotation information, including feature annotation signatures, proximity calculation of feature to nearest gene and other user-specified information. The annotation signature includes each element of a reference annotation such as gene structural components (exon, intron, UTR), repeats, etc. found by FEATnotator to overlap with input data. A subset of coordinates from this table generated in our analysis of *Arabidopsis thaliana* Ws ecotype indel sequence variation is shown in Table 2. See legend for details.

In the consolidated table, each annotation element is represented as “Type:ID” and multiple matching annotation elements are separated by a space. For example, a feature input matching two overlapping genes in Arabidopsis, AT1G01770 and AT1G01780 will

present the annotation signature: “gene:AT1G01770 gene:AT1G01780”. If an element is not accompanied by an ID in the reference annotation file, the ID of the corresponding Parent element is used along with its index value. For example, the annotation signature for a data element matching the third exon of mRNA AT4G14290.1 will be listed as “gene:AT4G14290 mRNA:AT4G14290.1 exon:AT4G14290.1_3”. For SNP analysis, where the reference element is a protein coding reading frame (CDS), the annotation signature output will further include the reference codon, reference amino acid, alternate codon and alternate amino acid. The annotation information for the input features designated by a single nucleotide position within the coding genes can specify high impact overlapping positions such as overlapping translation start or stop codons, splice sites and transcription start sites. Analysis of intergenic features will present additional columns specifying the nearest gene to the input feature and the distance to it. Proximity to other features, such as genetic marker positions, can be calculated routinely by adding the coordinates of these to the reference annotation file.

An additional output table, output.FEATnote.detailed.txt, includes the detail of integrated annotation listed as separate rows. This allows input features overlapping multiple annotation elements to be listed multiple times, once for each annotation feature, and allows the user to filter the input according to overlaps of interest. For example, when opened in a spreadsheet application, the user can quickly filter and selectively display only SNPs that overlap gene coding regions if desired.

Additional tabulated outputs of FEATnotator are listed in Table 1 and include those that compile summary statistics used for graphical outputs.

3.2 EXPERIMENT SUMMARIES

Stats such as the total number of input features, number of features overlapping the intergenic regions and those overlapping the genes, their distribution across intronic, exonic, coding and UTR regions, start and stop codons, splice sites, transcription start sites etc. are included in output summary file output.FEATnote.summary.txt (Table 1). For SNP data, this summary also includes transition and transversion counts, number of

synonymous and non-synonymous SNPs, number of premature stop codons generated or stop codons lost.

3.3 GRAPHS

FEATnotator provides an accessory script that plots summary information graphically, using data derived from the output tables. The graphical output includes distribution of features across the length of a selected reference span using sliding windows, a summary of variation over gene components, and a summary of coding SNPs impacts.

3.3 PROCESSING

In the determination of feature overlap data without added SNP CDS interpretation, FEATnotator processing completes within a few minutes for standard eukaryotic genomic references. For the integration and analysis of Ws accession SNPs in *Arabidopsis thaliana* shown here, the program processed the input SNP data (35) at ~2600 SNPs per minute. Use of the entire dataset of 700,000 rows completed in approximately 4.5 hours.

4 DISCUSSION

Laboratory and computational experiments that generate results with genomic positional information are accumulating at rapid rates in both model and non-model organism research programs. Today many references further incorporate sequence variant information from alternative races and admixtures to expand and accelerate trait or disease associations beyond a species' original reference haplotype (36-39). For a single reference sequence, furthermore, boundaries of multiple gene models per locus continue to be generated with expansive transcriptome sequencing experiments (40, 41). Gene regulation at the level of transcription and alternative splicing of mRNA naturally varies among accessions/races and among differing expression contexts in the same organism (42), sometimes limiting usefulness of a single gene model annotation for interpretations of sequence variation.

The Ws line of *Arabidopsis thaliana*, whose sequence variation (35) is summarized here

in our example software run (Figure 3), originated in Belarus and contrasts to the reference line Col, of U.S. origin, in some of its growth, development and response traits. In addition to its use in studies of natural variation using WGS (35), sequence variation in this accession is broadly noteworthy because Ws is the genetic background of community mutant collections in use internationally for over two decades in forward and reverse genetic screens (43). Numerous laboratories specialized in molecular genetics and plant physiology, not necessarily whole reference genomics, continue to sort through differences between Ws and Col to discover and refine our understanding of individual gene functions and ecotype-sensitive systems biology (44-46). To illustrate some of the FEATnotator output, here we showed the rediscovery of a reported (47) genomic deletion in Ws (Table 2). This protein structural deletion enhances root growth vigor over that of the Col reference allele (47).

Model organism genome reference annotations and publicly available, broadly familiar experimental data are used here for software demonstration purposes. By design FEATnotator offers perhaps an even greater appeal for genomic studies in non-model organisms, and early or experimental draft genome assemblies, due to its lack of dependence upon species-specific databases or strictly limited formats, its low memory requirement, and most of all speed to a detailed genomic interpretation that might be followed up with direct molecular confirmation, genetic and phenotypic consequence testing.

Often geneticists and other researchers are focused upon a genomic region rather than an entire genomic reference. Without the need to alter input files, FEATnotator allows the user to select genomic coordinates for summaries specific to their region of interest. When genetic markers are included as features, for example, genes and sequence variants within quantitative trait loci (QTL) can be integrated using FEATnotator and reduced to a simple, user-friendly vertically oriented table. Generation of QTL position-customized reference regions allowed us to readily identify functional consequence of SNPs in newly annotated genes in the genome of *T. cacao*, leading to discovery of natural variation in a conserved trans-acting small interfering, or *tasi*-RNA interaction site (48). This SNP was

initially annotated as merely one of four synonymous coding SNPs, but progressive use of experimental feature inputs marked an overlapping target site for *TAS4*-si81, a small RNA that destabilizes MYB113 expression and eliminates downstream anthocyanin production for red color in cacao pods (45).

SNP annotation tools with reportedly high processing rates (18) require building of the necessary reference databases and loading these completely into memory, thus demanding computing resources with larger memory. Since FEATnotator does not load the input feature data or the genome annotation data completely into memory (see Section 2.2 and Figure 2), sizes of these inputs do not influence the processing rate. The reference sequence FASTA file will however be loaded into memory in the use of FEATnotator for SNP analysis. Over the runtime of Ws SNP annotation, the maximum memory occupied was only around 500 MB, allowing this experiment to be run on personal computer. Use of FEATnotator may effectively complement or replace the need for genome visualization software installations for a variety of applied studies, of particular value to researchers who lack continuously robust Internet access and/or computing resource support.

5 CONCLUSIONS

The tabulated outputs of FEATnotator integrate genomic positional information that is limited only by the user, such as polymorphism in a non-reference genetic background and loci with transcript coverage, into a research dataset such as reference gene models. Associated whole dataset summaries provide researchers with a simple and functionally meaningful survey of the genic locations of queried features and sequence variation. These are being used to summarize new sequence references, to quickly compare genomic datasets relative to one another, and to guide continuing research such as surveys of genomic selection and genetic marker development. Both overlap and proximity to genomic annotation are uncovered by FEATnotator, and details of protein translation consequence are indicated when the program is used for SNP analyses. Remarkable ease of use, low memory requirement and lack of dependence upon other

software installations or specialized data input formats make FEATnotator especially attractive to a diversity of researchers.

References

1. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005;437(7057):376-80.
2. Bentley DR. Whole-genome re-sequencing. *Current opinion in genetics & development*. 2006;16(6):545-52.
3. Mardis ER. Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*. 2008;9:387-402.
4. Shendure J, Ji H. Next-generation DNA sequencing. *Nature biotechnology*. 2008;26(10):1135-45.
5. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, et al. Whole-genome sequencing and variant discovery in *C. elegans*. *Nature methods*. 2008;5(2):183-8.
6. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*. 2008;456(7218):66-72.
7. Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, et al. Characterization of the piRNA complex from rat testes. *Science*. 2006;313(5785):363-7.
8. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*. 2008;133(3):523-36.
9. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*. 2008;5(7):621-8.
10. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*. 2008;320(5881):1344-9.
11. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*. 2008;453(7199):1239-43.

12. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. *Science*. 2007;316(5830):1497-502.
13. Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, et al. High-resolution profiling of histone methylations in the human genome. *Cell*. 2007;129(4):823-37.
14. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods*. 2007;4(8):651-7.
15. Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, et al. ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*. 2009;457(7231):854-8.
16. Sajjan SA, Hawkins RD. Methods for identifying higher-order chromatin structure. *Annual review of genomics and human genetics*. 2012;13:59-82.
17. Bai Y, Cavalcoli J. SNPAAMapper: An efficient genome-wide SNP variant analysis pipeline for next-generation sequencing data. *Bioinformatics*. 2013;9(17):870-2.
18. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*. 2012;6(2):80-92.
19. Dayem Ullah AZ, Lemoine NR, Chelala C. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic acids research*. 2012;40(Web Server issue):W65-70.
20. Doran AG, Creevey CJ. Snpdat: easy and rapid annotation of results from de novo snp discovery projects for model and non-model organisms. *BMC bioinformatics*. 2013;14:45.
21. Makarov V, O'Grady T, Cai G, Lihm J, Buxbaum JD, Yoon S. AnnTools: a comprehensive and versatile annotation toolkit for genomic variants. *Bioinformatics*. 2012;28(5):724-5.
22. Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Molecular biology and evolution*. 2014;31(7):1929-36.
23. Reumerman RA, Tucker NP, Herron PR, Hoskisson PA, Sangal V. Tool for rapid annotation of microbial SNPs (TRAMS): a simple program for rapid annotation of genomic variation in prokaryotes. *Antonie van Leeuwenhoek*. 2013;104(3):431-4.
24. Shen TH, Carlson CS, Tarczy-Hornoch P. SNPit: a federated data integration system for the purpose of functional SNP annotation. *Computer methods and programs in biomedicine*. 2009;95(2):181-9.
25. Blahnik KR, Dou L, O'Geen H, McPhillips T, Xu X, Cao AR, et al. Sole-Search: an integrated analysis program for peak detection and functional annotation using ChIP-seq data. *Nucleic acids research*. 2010;38(3):e13.
26. Huang W, Loganantharaj R, Schroeder B, Fargo D, Li L. PAVIS: a tool for Peak Annotation and Visualization. *Bioinformatics*. 2013;29(23):3097-9.
27. Zhu LJ, Gazin C, Lawson ND, Pages H, Lin SM, Lapointe DS, et al. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC bioinformatics*. 2010;11:237.

28. Habegger L, Balasubramanian S, Chen DZ, Khurana E, Sboner A, Harmanci A, et al. VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics*. 2012;28(17):2267-9.
29. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-2.
30. Li H. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*. 2011;27(5):718-9.
31. Gent JJ, Ellis NA, Guo L, Harkess AE, Yao Y, Zhang X, et al. CHH islands: de novo DNA methylation in near-gene chromatin regulation in maize. *Genome research*. 2013;23(4):628-37.
32. Gelfand MS, Mironov AA, Pevzner PA. Gene recognition via spliced sequence alignment. *Proceedings of the National Academy of Sciences of the United States of America*. 1996;93(17):9061-6.
33. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*. 2003;19 Suppl 2:ii215-25.
34. NCBI. NCBI Taxonomy Homepage [cited 2014 28 October]. Available from: <http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi?chapter=tgencodes-SG1>.
35. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, et al. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*. 2011;477(7365):419-23.
36. Rounsley SD, Last RL. Shotguns and SNPs: how fast and cheap sequencing is revolutionizing plant biology. *Plant Journal*. 2010;61(6):922-7.
37. Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, et al. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature biotechnology*. 2012;30(1):105-11.
38. Chia JM, Song C, Bradbury PJ, Costich D, de Leon N, Doebley J, et al. Maize HapMap2 identifies extant variation from a genome in flux. *Nature genetics*. 2012;44(7):803-7.
39. LaFave MC, Varshney GK, Vemulapalli M, Mullikin JC, Burgess SM. A Defined Zebrafish Line for High-Throughput Genetics and Genomics: NHGRI-1. *Genetics*. 2014;198(1):167-70.
40. Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, et al. Diversity and dynamics of the *Drosophila* transcriptome. *Nature*. 2014;512(7515):393-9.
41. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, et al. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science*. 2010;330(6012):1775-87.
42. Reddy AS, Marquez Y, Kalyna M, Barta A. Complexity of the alternative splicing landscape in plants. *The Plant cell*. 2013;25(10):3657-83.
43. Feldmann KA, Marks MD, Christianson ML, Quatrano RS. A Dwarf Mutant of *Arabidopsis* Generated by T-DNA Insertion Mutagenesis. *Science*. 1989;243(4896):1351-4.
44. Pacurar DI, Pacurar ML, Pacurar AM, Gutierrez L, Bellini C. A novel viable allele of *Arabidopsis* CULLIN1 identified in a screen for superroot2 suppressors by next generation sequencing-assisted mapping. *PloS one*. 2014;9(6):e100846.

45. Kalbina I, Strid A. Supplementary ultraviolet-B irradiation reveals differences in stress responses between *Arabidopsis thaliana* ecotypes. *Plant, cell & environment*. 2006;29(5):754-63.
46. Chan Z, Loescher W, Grumet R. Transcriptional variation in response to salt stress in commonly used *Arabidopsis thaliana* accessions. *Plant physiology and biochemistry : PPB / Societe francaise de physiologie vegetale*. 2013;73:189-201.
47. Beuchat J, Li S, Ragni L, Shindo C, Kohn MH, Hardtke CS. A hyperactive quantitative trait locus allele of *Arabidopsis* BRX contributes to natural variation in root growth vigor. *Proceedings of the National Academy of Sciences of the United States of America*. 2010;107(18):8475-80.
48. Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, Livingstone D, 3rd, Cornejo O, et al. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome biology*. 2013;14(6):r53.

LEGENDS FOR FIGURES AND TABLES

Figure 1. Schematic Overview of FEATnotator Software Workflow. Types of input data are indicated on the top left. Example output data are shown on the lower right side pointing out. From positional information in input feature files, FEATnotator identifies overlap with reference genome annotation. Outputs include the tabulated feature association with, or proximity to, the experimental input features. For SNP analysis, a reference sequence file is provided along with the SNP variation input data. Output in this case also includes analyzed translational impacts of the coding SNPs with details of the variation. FEATnotator offers options for the generation of graphical outputs illustrating the summary information. Graphs include those showing the density of features across the length of a reference scaffold using sliding windows, distribution over gene components and intergenic regions, and translational impacts of coding SNPs.

Figure 2. Schematic illustration of the FEATnotator Algorithm. FEATnotator identifies the positional overlap between experimental feature input data and annotation data, both sorted by chromosomal position. The solid lines indicate the workflow and the dashed lines indicate the direction of data flow.

Figure 3. Graphical summaries of SNPs in the Ws ecotype of *Arabidopsis thaliana* with respect to the Col reference. **A**, Distribution of total SNPs and predicted “high-impact” SNPs across the entire length of Chr1 binned into 10,000 bp overlapping windows with a step size of 1000 bp. **B**, Data as above graphed in the subset region from 24.1 MB to 24.2 MB of Chr1. **C**, Summary distributions of Ws SNPs with respect to TAIR10 intergenic and genic structural elements of reference genes as labeled. Smaller

pie chart inset summarized proportions of coding SNPs according to their predicted translational impacts.

Table 1. The seven standard tabulated file outputs of FEATnotator. Descriptions of content of each file output are listed. User may specify prefix other than the default prefix “output.”

Table 2. FEATnotator output indicating a protein coding genomic deletion in Ws relative to the Col reference. Shown are a subset of FEATnote.consolidated table outputs showing coordinates that correspond to the *Arabidopsis thaliana* gene encoding *BREVIS RADIX* (*BRX*), previously reported by Beuchat *et al.* to harbor a deletion that eliminates seven amino acids in the Ws protein [44]. Integrated were indel sequence variation features provided by Gan *et al.* [32]. FEATnotator output indicates the Ws deletion start site position (shaded in grey, *row 4*) along with display of nucleotide length (column 3) and sequence (column 4) difference between the two accessions. Further detail of sequence variation in and around *BRX* is shown in rows above and below, revealing additional deletions (*row 1*, 3 nt 802 upstream of the gene, *row 3*, 2 nts in an upstream intron; and *row 7*, 34 nts in a downstream intron) and insertions (*row 2*, 2 nts in an upstream intron; *rows 5 and 6*, 7 and 28 nts in a downstream intron; and *row 8*, 6 nts in the 3’UTR).

Podicheti and Mockaitis, **Figure 1.**

Podicheti and Mockaitis, **Figure 2.**

A**SNP Distribution Across Chr1 – Windowsize 10000 – Stepsize 1000** **B****SNP Distribution Across Chr1:24.1MB..24.2MB – Windowsize 10000 – Stepsize 1000**

C

Distribution of SNPs in Genome Annotation

Table 1.

Output File Name	Description
output.FEATnote.consolidated.txt	Primary annotation table. Contains every entry from feature input file extended by the annotation information. Includes feature Annotation Signature, proximity to nearest gene, and other user-specified information.
output.FEATnote.detailed.txt	Detailed annotation table. Lists each overlapping annotated feature as a separate row. Suitable for filtering dataset based on overlaps with features of interest.
output.FEATnote.summary.txt	Detailed summary of annotation integration.
output.FEATnote.genic.txt	Summary classification of feature input content with respect to genic and intergenic elements of the annotation.
output.FEATnote.XXX.txt	Counts of input features by overlap with each element of the reference annotation. XXX is for example, <i>gene</i> .
output.FEATnote.NOoverlap.XXX.txt	List of reference annotation elements that do not overlap with input features. XXX is for example, <i>gene</i> .
output.FEATnote.SNPsense.txt	Classification and counts of coding (CDS) SNPs based on impacts upon protein translation.

Table 2.

Scaffold	Pos	L	Reference	Consensus	Annotation Signature	Inter Genic	Intron	CDS	UTR	TSS	Nearest	Distance
Chr1	11444688	-3	TTT	-	NA	YES	NO	NO	NO	NO	AT1G31880	802
Chr1	11445854	2	-	TT	mRNA:AT1G31880.1[intron]	NO	YES	NO	NO	NO	NA	NA
Chr1	11448744	-2	TG	-	mRNA:AT1G31880.1[intron]	NO	YES	NO	NO	NO	NA	NA
Chr1	11449673	-21	GGACCAGGAAGTTATC ATGGA	-	mRNA:AT1G31880.1 CDS:AT1G31880.1_1	NO	NO	YES	NO	NO	NA	NA
Chr1	11449896	7	-	GGCAAGA	mRNA:AT1G31880.1[intron]	NO	YES	NO	NO	NO	NA	NA
Chr1	11449898	28	-	TTAGACATTTGAA ACTCGATAATCGTA	mRNA:AT1G31880.1[intron]	NO	YES	NO	NO	NO	NA	NA
Chr1	11449900	-34	AATCATATACATTTTGA AAGTTGGTAATCATACC	-	mRNA:AT1G31880.1[intron]	NO	YES	NO	NO	NO	NA	NA
Chr1	11450073	6	-	TTTCTC	mRNA:AT1G31880.1 exon:AT1G31880.1_1	NO	NO	NO	YES	NO	NA	NA