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 genomic 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.

### Feature Highlights

- A software tool that integrates and reports overlap and proximity in genomic information from any user-defined datasets including those from next generation sequencing applications.
- Considerably fast and user friendly and is easy to use even with limited computing resources.
- 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.
- Maintains a very low memory footprint and can run with ease on a personal computer.
- 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.
- Reports “High-impact” SNPs altering the start codons, stop codons, splice sites or transcription start sites.
- Output includes graphical summaries and data tables.

### Implementation

- Data input includes information on the genomic features to be annotated
- Reference annotation information as a GFF3 file
- Reference Sequence as a FASTA file
- Identifies positional overlap of input experimental features and reference genome annotation.
- High impact feature calls include single nucleotide features (SNPs, methylation sites etc.) overlapping with the first base of annotation “gene” feature (Transcription Start Site)
- SNPs overlapping with splice sites (two nucleotide positions on either side of the exon-intron splice boundary)
- The coding SNPs are categorized into “NONSENSE”, “MISSENSE” and “SILENT” SNPs.
- Generates proximity output by calculating the number of nucleotides from a gene annotation as the closest distance to the nearest gene.

### 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.

### Output Files and Tables

- Classificaion and counts of coding (CDS) SNPs based on impacts upon protein translation.
- List of reference annotation elements that do not overlap with input features. "XXX" is for example, gene.
- Counts of input features by overlap with each element of the reference annotation.
- Detailed summary of annotation integration.
- Includes summary of feature input data and report regions.
- Graphical summaries of specificity of feature input data and report regions.
- Graphical summaries of specific inclusion of feature input data.
- Detailed results and counts of overlap data (SNP) between input features.
- Detailed results and counts of overlap data (SNP) between input features.

### 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 genomic structural elements of reference genes as labeled. Smaller pie chart inset summarized proportions of coding SNPs according to their predicted translational impacts.

### 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).