Specific Aims. The overarching goal of this proposal is to dramatically improve the accuracy, breadth, and depth of automated eukaryotic genome annotation by introducing new types of next-generation sequencing (NGS) data into the gene-annotation process. For maximal impact, this project will leverage MAKER, a widely adopted, easy-to-use genome annotation pipeline. MAKER was developed as part of an NHGRI funded project entitled *Software for the creation and quality control of genome annotations*, (R01HG004694) and an NSF pilot program grant for better annotation of plant genomes (IOS-1126998). MAKER already identifies DNA repeats, performs EST, RNA-seq and protein alignments, predicts genes with both evidence-informed and *ab initio* methods, and automatically consolidates this information into gene annotations with evidence-based quality scores. Here we propose to create the first fully NGS-enabled genome-annotation annotation pipeline. This will be accomplished by expanding MAKER to support the many different types of NGS data, and by constructing a probabilistic evidence synthesis engine built upon the Mackey lab's ENIGMA code-base. This ENIGMA engine will allow MAKER-*NGS* to employ diverse NGS data-types in the gene-annotation process, including population resequencing data, species-comparative gene annotations, predicted non-coding RNA genes and pseudogenes, modENCODE-like RNA pol II ChIP-seq profiles, 5' and 3' UTR tags, and MS-based proteomics data. The outputs of MAKER-*NGS* will be Sequence Ontology-compliant GFF3 and GVF files that can be directly imported into any GMOD genome database and analyzed with any GMOD compliant tool.

Aim 1. Use population re-sequencing data to improve genome annotation. Our goal here is to make MAKER the first genome annotation pipeline capable of using re-sequencing data. Recent work has shown that the average human genome harbors some 200 rare nonsense alleles. Re-sequencing data is now commonly available for many novel genomes at the time of annotation. MAKER-*NGS* will use these data to distinguish sequencing errors and rare nonsense alleles from true species-wide pseudogenes. The Yandell lab will extend MAKER to enable it to use re-sequencing data to inform the annotation process, allowing underlying gene predictors and sequence alignments to consider alternative alleles at candidate loci; Dr. Mackey's lab will concomitantly extend ENIGMA to identify and classify these variant-induced alternative annotations and true species-wide pseudogenes. These annotations together with their associated evidence will be included in MAKER-*NGS*'s GFF and GVF files. These output files will also enable the development of 3rd party tools that will use MAKER's outputs enriched with re-sequencing data for population genetic analyses (see letter of collaboration from Dr. Shapiro).

Aim 2. Exploit RNA-seq and species-comparative evidence for annotation of alternative splicing. We will enable the Enigma evidence synthesis engine to use RNA-seq alignments, in the context of pre-existing gene predictions and annotations to further revise and enhance MAKER's ability to identify and annotate alternatively spliced transcript isoforms. The ENIGMA engine will do so by exploring the wealth of alternative splicing events identifiable by RNA-seq alignments, together with other sources of alternative isoform evidence, including high-scoring, yet otherwise suboptimal, *ab initio* gene predictions, and evidence of evolutionarily-conserved alternatively spliced exons seen in comparative data. This will not only improve MAKER-*NGS*'s overall accuracy, it will also provide users with novel and highly desirable functionality for the identification of new transcript isoforms using their RNA-seq data in the context of pre-existing annotations.

Aim 3. Use the complete gamut of NGS data for gene-finder training. MAKER employs a bootstrap training procedure, whereby output from initial runs can be used to further refine the training of its underlying gene prediction methods, e.g. SNAP, Augustus, GeneMARK, FgenesH. NGS data such as RNA-seq, re-sequencing, and other evidence obviously have great potential to improve the gene-finder training process. Despite this fact, no tools or procedures currently exist to train gene-finders using the full gamut of NGS data-types. Dr. Mackey's lab will develop an ENIGMA-based tool that will mine these data from MAKER-*NGS* outputs and make them available for training purposes. This tool will also employ MAKER's new non-coding RNA gene predictions to compartmentalize candidate protein-coding segments of the genome for still better performance. This ENIGMA-based tool will operate on the augmented MAKER GFF and GVF files produced by the tools proposed in Aim 1 & 2 to produce synergistic improvements to MAKER's overall performance.

Collectively, these aims will create the first genome-annotation pipeline capable of exploiting the full potential of NGS data. The pipeline's NGS data-containing GFF and GVF outputs will also provide a rich new resource for downstream analyses and 3rd party tool development. Additional strengths of this proposal include the fact that it partners a new investigator with an established one, leveraging the strengths of both for rapid, coordinated software development that will employ best practices for software design and data sharing. Impact is assured, as this work will directly address the needs of the large and international MAKER user community.