# Documentation for BESC KB Genome Resequencing toolkit

Tatiana V Karpinets
tvkarpinets@ornl.gov
12/15/12

## Introduction

This kit was developed for BESC genome resequencing projects. The target for resequencing is usually the genome of a strain, which is further referred to as the mutant strain, or M strain, with a practically important phenotype evolved by an adaptation of a wild type strain. This phenotype is absent in the wild type strain, or WT strain, with a completely sequenced genome, which is used as the reference genome. The goal of the resequencing projects is to discover genomic modifications underlying the specific phenotype of the mutant. The toolkit we describe here was developed to analyze sequencing data produced for the resequenced mutant genomes in terms of their biological content and in terms of physiological effects on the mutant phenotype. By developing tools we aimed not just to re-call SNPs and indels from the resequencing data, but also to facilitate the biological interpretation of the mutations by their comprehensive annotation. In addition, we sought to predict molecular functions that likely lost, gained, or changed in the mutant. Below is a brief description of tools available in the kit.

## Resequencing tools

The tool kit consists of 6 tools that can be applied either as a pipeline (Figure 1) or independently. The pipeline starts with calling SNPs and indels from the text file with high confidence differences produced by 454 sequencing. An example of the text file is provided at http://cricket.ornl.gov/html/download/resequencing/454HCDiffs.txt. The tool, which is referred as **"SNP-indel caller"** finds changes in the genome of the M strain and their location given a file with high confidence 454-reads as the input. The output file of the tool has 5 columns that include the contig name in the reference genome, the stat position of the mutation in the reference genome (Reference_start), the end position of the mutation (Reference_end), the corresponding sequence (from Reference_start to Reference_end) in the reference genome, and the corresponding sequence in the mutant genome (Mutant_seq). For insertions, Reference_start and Reference_end are the same. For deletions, Mutant_seq value is equal "-". The described presentation of SNPs and small indels are typical for resequensing software and allows one to locate changed nucleotides in the wild type (reference) genome and to reproduce a new nucleotide sequence in the mutant strain for further analysis. An example of the output file is provided at http://cricket.ornl.gov/html/download/resequencing/454HCDiffs_dif.txt. The rest 5 tools in the toolkit annotate the revealed mutational changes in the M strain genome to understand the biological effects of each change on the mutant phenotype. They use output file of the "SNP-indel caller" as an input.

Tools named **"Mutant proteins – Fasta"** and **"Mutant proteins – CDD"** takes the file produced by the SNP-indel caller and the contig/contigs of the WD (reference) as inputs. The first tool generates a FASTA file of proteins for the M strain and the second tool annotates the proteins with protein family domains using the CDD pipeline.

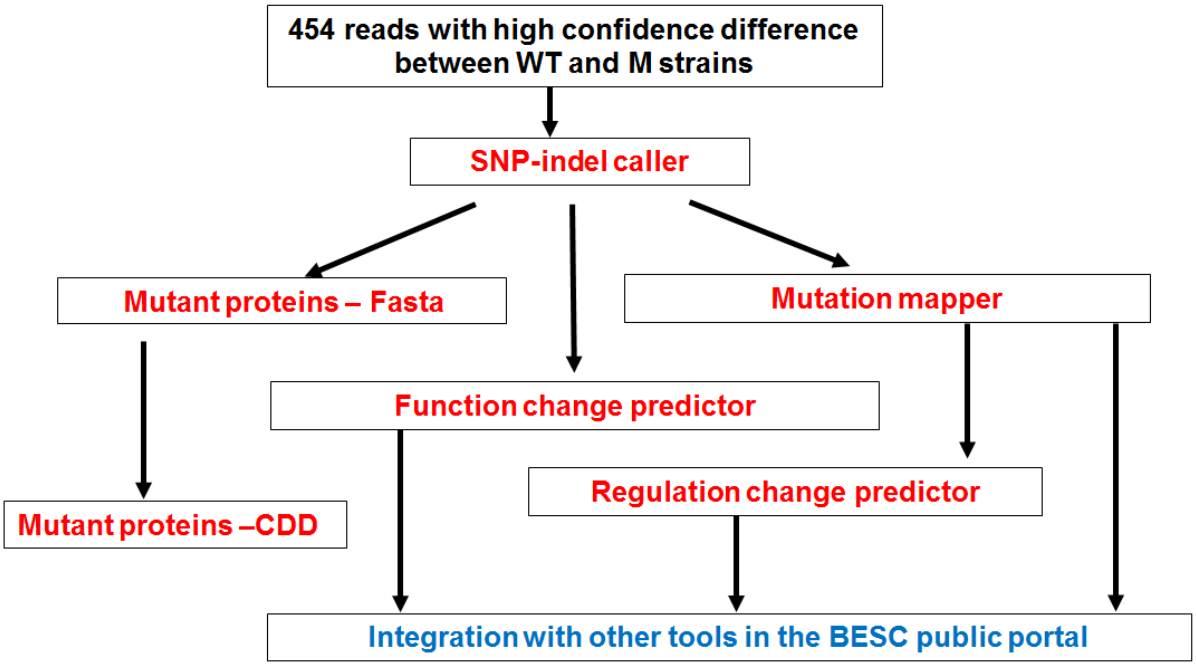


Figure 1. Bacterial Genome resequencing pipeline to predict mutations and their biological functions in the mutant (M) strain from 454 sequence reads using the genome of a wild type (WT) strain as the reference. The outputs produced by each tool (red color) are the following. *SNP-indel caller*: a list of changes in the M strain genome and their location. *Mutant proteins – Fasta*: Protein fasta file for the M strain. *Mutation mapper*: List of mutations, their type and position in the genome, mapping to gene and intergenic regions. *Function change predictor*: List of genes with probable loss, change, or gain of function. *Mutant proteins –CDD*: CDD annotation of the M strain proteins. *Regulation change predictor*: List of genes with potential changes in their regulation.

**"Mutation Mapper".** The tool "Mutation Mapper" annotates all identified changes in the genomic sequence (contigs) of the M strain. The output file gives a comprehensive annotation for each mutational change in the genome, including a location of the change within an intergenic region or gene, its position in the genome, the type of change (synonymous or non-synonymous, multiple changes, insertions, or deletions), and amino acids and codons in the M and WT strains.

**"Function Change Predictor".** The "Function change predictor" evaluates potential changes in the protein function, namely a loss or a gain of a protein family domain, or a Pfam domain. The predictions are implemented using two different approaches, and each approach produces a separate output table. The tool requires two input files. One input file is the output of the "SNP-indel caller". The other input is the feature tables (also called gbk-files) of the contig/contigs of the WD (reference) strain. An example of the second input is available at . http://cricket.ornl.gov/html/download/resequencing/Contig.gbk

The first approach finds genes that have putatively lost or gained their Pfam domain/domains because of a change in the start/stop codon or indels. The most confident loss or gain of the protein function often takes place because of insertions, deletions in the coding sequence of genes leading to a premature stop codon or because of substitutions in its start/stop. These changes can lead to truncated genes with no functional domains or to fused genes. To identify this set of genes, our script re-calls each coding sequence of the mutant from 454 reads and annotates it by protein family domains Then the domains are compared with Pfam domains of the related coding sequence in the WT strain to find those that are lost or gained. The identified lists of genes and their lost/gained Pfam domains are produced as output of the approach.

Our second approach targets for the analysis only non- synonymous single nucleotide substitutions that occur within a Pfam domain of the protein. These mutations also more likely affect protein function and are comprised of a great number of mutations in coding sequences. We propose that a part of the protein sequence aligned with the Pfam model is the most important for the protein function, and substitutions of the amino acids within this part more likely lead to changes in the gene function. These changes, however, are not necessarily deleterious for the protein. They may also confer a gain of the protein function or its change. Many oncogenes in cancer cells, for example, increase activity because of amino acid substitutions that change a non-conserved allele to one that is more conserved in protein homologs (Lee, et al., 2009). Thus, a change of a common amino acid within a Pfam domain to an uncommon one, is likely deleterious for the domain, but a change in the opposite direction, from an uncommon to a common amino acid, may lead to a gain or change of the protein function. This concept has been recently used to calculate bi-directional SIFT score. We have employed the concept to calculate the Pfam score of the non-synonymous amino acid substitution within a Pfam domain. The Pfam score is calculated from the PSSM scores of the Pfam model, taking into consideration the position of the substitution and the amino acid in the mutant and in the reference sequence. The resulted Pfam score of the mutation can be positive, which indicates a potential gain of function, or negative, which points to a potential loss of the protein function.

The following algorithm is implemented in Perl to calculate the mutation Pfam score. First we identify Pfam domains in WT proteins with non-synonymous mutations using RPS-BLAST (Altschul, et al., 1997). Only proteins with mutations within Pfam domains are considered. An alignment of the Pfam model with the WT protein was used to identify the location of the mutation. If the mutation aligned with a gap in the model, then scoring was not available. If the mutation aligned with an amino acid in the model, then the score for the WT protein (WT_AA_Score) was equal to the PSSM score for the same position and the amino acid found in the WT protein. Similar to the PSSM score, WT_AA_Score was positive or negative integers in the range from -1000 to +1000. Positive scores indicate that the given amino acid substitution occurs more frequently in the alignment than expected by chance, while negative scores indicate that the substitution occurs less frequently than expected. Large positive scores often indicate critical functional residues, which may be active site residues or residues required for other intermolecular interactions. Thus, the WT_AA_Score indicates the frequency of the WT amino acid versus the one expected by chance. The score for the mutant protein (Mutant_AA_Score) was also equal to the PSSM score, but it referred to the amino acid found in the mutant protein at the same position. The Mutant_AA_Score, therefore, indicates the frequency of the mutant

amino acid versus the one expected by chance. The Pfam score for the non-synonymous mutation was calculated as the difference between the Mutant_AA_Score and the WT_AA_Score. Thus, a high positive value of the Pfam score indicates that the amino acid found in the mutant strain is more typical for the Pfam domain then the amino acid found in the WT strain at the same position. This change likely enhances the protein function (Lee, et al., 2009). The high negative score indicates a potential deleterious effect of the mutation, because the mutation takes place at a conserved position within the Pfam domain. The output file of this approach includes the list of proteins with non-synonymous SNPs and their scores.

The predicted effect of mutations on a protein family domain is potential and must be validated experimentally. First, the effect may be not as simple as just a loss or gain of function. One example of a complex effect of mutations on protein function is AdhE (Cthe_0423), bifunctional acetaldehyde-CoA/alcohol dehydrogenase, which is a crucial enzyme in the fermentation to ethanol in *Clostridium thermocellum*. Our tool predicted 2 mutaitons in this enzyme in an ethanol tolerant strain of *C. thermocellum*. Both mutations were within the iron-containing alcohol dehydrogenase (Fe-ADH) domain; although one mutation was predicted as deleterious, and the other one as beneficial. Thus, a loss of the domain was expected from the predictions. Our further experimental study (Brown, 2011) revealed that the enzyme indeed lost its NADH-dependent activity with simultaneous acquisition of NADPH-dependent activity. Thus, changes in the enzyme function were more complex than expected. Second, the tool takes into consideration only mutations that effect protein family domains. Mutations in other regions of the coding sequence, however, also can be important and lead to changes in protein function.

**Regulation change predictor:** This tool produces the list of genes with mutational changes, either SNPs or indels, in the upstream intergenic regions, given WT_M_Differences.txt and the contig/contigs of the WD (reference) as inputs. The output table (tab delimited) of the tool includes sequences of the intergenic regions for both the mutant and reference, and their differences and characteristics. A Perl script was used to reproduce coding and intergenic regions of the mutant from the identified mutations (SNP-indel caller output) and to compare each intergenic region with the same intergenic regions in the wild type strain. For each changed promoter the script finds the downstream gene/genes, reasons for the change (truncation, mutation), the location of the new region in case of truncation, the location of the mutation if the reason for the change is mutation, the sequence of the promoter in the mutant, the sequence of the promoter in the reference genome, and the difference in the length of the promoters.

**Integrative analysis of the identified mutations using genome annotations and other tools in the portal**

To reveal the biological meaning of the identified mutations, the annotations can be further integrated into the portal environment at the level of specific genes, intergenic regions, or protein family domains using the appropriate tools described above and genome annotations collected for BESC-targeted bacterial organisms. The pathway tools software (Caspi, et al., 2010) and metabolic reconstructions in BESC BeoCyc can be used to map mutations to transcriptional units and metabolic pathways and to overlay them with the genome and pathway overviews.

## References


Lee W, Zhang Y, Mukhyala K, Lazarus RA, Zhang Z: **Bi-directional SIFT predicts a subset of activating mutations**. *PLoS One* 2009, **4**(12):e8311.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

Brown SD, Guss AM, Karpinets TV, Parks JM, Smolin N, Yang S, Land ML, Klingeman DM, Bhandiwad A, Rodriguez M, Jr. *et al*: **Mutant alcohol dehydrogenase leads to improved ethanol tolerance in Clostridium thermocellum**. *Proc Natl Acad Sci U S A* 2011, **108**(33):13752-13757.

Caspi R, Altman T, Dale JM, Dreher K, Fulcher CA, Gilham F, Kaipa P, Karthikeyan AS, Kothari A, Krummenacker M *et al*: **The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases**. *Nucleic Acids Res* 2010, **38**(Database issue):D473-479.