brcaindica
Collaborative Compendium of *BRCA1* and *BRCA2* genetic variants in India

USER MANUAL

About BRCAIndica

The BRCAIndica compendium is a powerful resource that offers researchers and clinicians with over 10,000 *BRCA1* and *BRCA2* genetic variants classified as per the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) guidelines for the accurate interpretation of variant pathogenicity.

BRCAIndica is accessible at:
https://clingen.igib.res.in/brcaindica/index.php

Homepage

The homepage offers an overview of the resource, along with the breakdown of variant numbers of each gene in the database. The top right section contains links to this user manual, along with the paper containing details about ACMG annotations by Richards et. al. for easy reference. The section also contains a 'Contact Us' button - clinicians and researchers who wish to collaborate by submitting their variants can click on it to participate in the BRCAIndica Collaborative Compendium program.

Next, the Search section offers various example formats in which the database can be queried, including variant ID, nucleotide change, amino acid change and rsID formats. Clicking on any of the example formats opens up the Results section, offering some basic details about the variant. It also contains a "More Information" button, which leads the user to a new

page with detailed annotations of the variant. These include:

1. **Information on Variants:** This section includes more detailed information describing the variant, including its gene, chromosomal location, links to rsID (dbSNP), nucleotide and amino acid changes, reference and alt alleles, as well as information about the function, and clinical significance of the variant based on ClinVar.

2. **Allele Frequencies:** An allele is an alternative form of a given gene which is located at the same genetic locus on the chromosome. An individual could carry one or more allelic forms of a gene. Allele frequencies represent the frequency of the occurrence of a particular allele in a particular population. It is calculated as the number of times an allele has been observed in the population i.e. allele count to the number of the total alleles present at that locus i.e. allele number.

   BRCAIndica offers allele frequencies for the following populations:
   (a) **Genome Aggregation Database (GnomAD):** It is a global population dataset bearing whole genome sequencing information from 9 major populations, including African/African-American (AFR), Amish (AMI), Latino/Admixed American (AMR), Ashkenazi Jewish (ASJ), East Asian (EAS), Finnish (FIN), Non-Finnish European (NFE), South Asian (SAS), and Other (OTH) populations.
   (b) **1000 Genome project (1000Genomes):** It is another global population dataset database is composed of whole genome sequence dataset of 2,504 individuals from 5 major population i.e. Europe (EUR), South Asia (SAS), Africa (AFR), East Asia

(EAS), and the Americas (AMR) comprising of 26 sub-populations.

(c) **Esp6500 (Esp6500siv2):** This dataset is composed of a whole exome sequence dataset of 6,503 individuals from two populations: 2,203 individuals of African-American and 4,300 individuals of European-American population.

(d) **IndiGen:** The IndiGen data contains whole genome sequencing data from 1029 self-declared healthy Indian individuals from 27 states, representing diverse ethic subgroups.

The Allele Frequencies section links to this manual for greater clarity on the section.

3. **Computational Predictions:** This section bears pathogenicity predictions regarding the variant through three benchmarked computational prediction tools:

   (a) **SIFT (Sorting Intolerant From Tolerant):** Predictions are based on protein sequence homology as well as the physical properties of amino acids. The scores indicate:
   D - Deleterious (sift<=0.05)
   T - Tolerated (sift>0.05)

   (b) **Polyphen2 (Polymorphism Phenotyping v2):** Predicts the effect of amino acid substitutions on the structure and function of the protein. The scores indicate:
   D - Probably damaging (>=0.957)
   P - Possibly Damaging (0.453<=pp2_hdiv<=0.956)
   B - Benign (pp2_hdiv<=0.452)

(c) **CADD (Combined Annotation Dependent Depletion):** It offers scores ("scaled C-scores") obtained from trained linear kernel support vector machines using multiple annotations correlated with allelic diversity, regulatory effects experimentally measured, pathogenicity of both coding and non-coding variants, and high rank pathogenic variant in human genome sequence. The scores indicate:
>15 : predicts a pathogenic effect of the variant
<15: predicts a benign effect of the variant

The Computational Predictions section also links to this manual for greater clarity.

4. **ACMG/AMP Classification:** This section bears the ACMG/AMP attributes, as well as the final classification assigned to the variant and the date the variant was last updated. The section links to the ACMG paper by Richards et. al. so that it can act as a ready reference for greater clarity on each attribute.

5. **Reference:** This section offers a link to the relevant paper or database that is associated with the given variant.