IMPACT
CASE STUDY

Variant Effect Predictor

e!Ensembl®

THE IMPACT OF STANDARDISATION

A genome is an organism's complete set of DNA, including all of its genes, and holds the key to greater understanding of an organism's development, how and why we differ and what makes us susceptible to diseases. Many organisms including humans have now been fully sequenced and reference data sets are held in Ensembl's databases.

We can start to understand what is happening in individual genomes by comparing them with reference genomes, and combining that information with knowledge from all other fields of molecular biology.

Ensembl's Variant Effect Predictor (VEP) is a powerful open software tool that can analyse most types of variation data. It uses the extensive annotation in Ensembl to provide detailed functional predictions and annotation on the effects of variants.

FUNDING

Ensembl receives funding from the Wellcome Trust with additional funding for project specific components from the BBSRC, EC, NIH, CTTV and EMBL.

wellcometrust

EMBL-EBI CTTV NIH BBSRC European Commission

IMPACT

VEP is deployed in many critical areas of research such as cancer and rare diseases, where strong links have been established between changes in the genome and disease development. VEP also supports conversion of variant data into the format most familiar to clinicians (HGVS codes) allowing the knowledge gained to be directly applied.

VEP is often adopted above other similar tools due to its high performance, the stability of funding, extensive user support, and it's open licence does not restrict users in distributing their results. Below are some findings from interviews with three diverse users of the VEP tool demonstrating the research it supports.

USER IMPACTS

illumina® Genomics england

The 100,000 Genomes project will sequence 100,000 whole genomes from NHS patients by 2017. The project is focussing on patients with cancer and patients with a rare disease and their families. It is hoped that the project's legacy will be a service ready for adoption by the NHS, new medicines, treatments and diagnostics, and a country which hosts the world's leading genomic companies. *(Genomics England 2015)*

Illumina is a leading developer, manufacturer, and marketer of life science tools and integrated systems for large-scale analysis of genetic variation and function. It is the sequence provider for the world-leading 100,000 Genomes Project. VEP is part of the annotation engine that is used to deliver annotated genomes for this project, and also for Illumina's VariantStudio software product.

"The Illumina VariantStudio data analysis software application enables researchers to quickly identify and classify disease-relevant variants, and then communicate significant findings in a structured report. VariantStudio talks to an annotation tool which has VEP at its core. VEP was selected by Illumina because it was more robust and more production-ready than other annotation tools. Because of VEP's superior quality and accuracy its users are able to catch some edge cases where annotation would be otherwise incorrectly handled."

Elliott Margulies, Illumina

GENE VARIANTS

88 MILLION

The human genome is made of 3.2 billion bases of DNA which code for approximately 20,000 protein coding genes. Scientists from around the world catalogued 88 million variants.

RARE VARIANTS

64 MILLION

The 100,00 Genomes Project considered that the majority of variants, 64 million were considered rare in frequency, occuring in only 1% or less of the population.

RARE DISEASE

6-7%

Rare diseases are individually very uncommon but in total they affect a surprisingly large number of people, between 6 and 7 percent of the UK population.

PROCESSING POWER

The VEP software tool can process more than **4 million** genetic variants per hour.

IMPACT
CASE STUDY

Variant Effect Predictor

USERS IMPACTS Cont.

As part of the **Deciphering Development Disorders study,** a collaboration between the Wellcome Trust Sanger Institute and NHS regional genetic services to understand a range of development disorders in children, just under 14,000 families had their DNA sequenced. The VEP tool was a central part of the analysis.

‘we achieved a diagnostic yield of 27% among 1133 previously investigated yet undiagnosed children … In families with developmentally normal parents, whole exome sequencing of the child and both parents resulted in a 10-fold reduction in the number of potential causal variants that needed clinical evaluation … Most diagnostic variants identified in known genes were novel and not present in current databases of known disease variation.” (Wright et all 2015)

“The benefit of VEP is in annotating and predicting the likely consequences of variants identified in the study, allowing us to identify disease-causing variants much more efficiently and effectively. This is important for the DDD team, clinicians and families. About a third of families will likely receive a diagnosis. Only a small number will be treatable, but the information is valuable for counselling and helping them to make informed choices about having further children, based on whether the variant is likely to be inherited or spontaneous”.

Dr Caroline Wright, Wellcome Trust Sanger Institute

The Daniel MacArthur Lab and the **ExAC** project are prominent users of the VEP tool. The lab is jointly based at Massachusetts Hospital and the Broad Institute. VEP is central to three major projects in the lab including the Exome Aggregation Consortium (ExAC). ExAC is an international coalition of investigators with a focus on data from exome sequencing and variant discovery on the regions of the genome that encode proteins, known collectively as the exome. It is by far the largest single aggregation of coding variants in the world and a key comparison data set for childhood-onset Mendelian diseases.

“VEP is central to the ExAC project which is building a large reference database of human genetic variation. This is using exome sequencing to understand variation in human genes, and uses VEP to predict functional variation. By the end of 2015 we expect to have aggregated data from around 100,000 individuals and identified over 15 million genetic variants. Between launching in October 2014 and June 2015 EXAC has had over 1.5 million page views. This represents more than 80,000 unique users over 8 months.”

We went with VEP for three main reasons:

- The quality and completeness of annotations: VEP leverages the Ensembl gene set. On manual inspection of results there were inaccuracies in the other tools and VEP was correct.
- The software is beautifully documented, which makes it easy to expand and build plugins.
- VEP integrates seamlessly with Ensembl. We often need to pull in other types of data and can do this smoothly with VEP”.

Daniel MacArthur, Massachusetts Hospital/Broad Institute and lead analyst ExAC project.

PROCESS POWER

83%

From a survey of over 2,500 Ensembl users, 83% reported a high or very high benefit of the data and tools supplied.

DIRECT IMPACT

1 in 3

In the developed world, cancer will affect one in three people at some stage in their life. Without research we condeming tomorrow’s generation to today’s treatments.

RESEARCH EFFICIENCY

46%

MORE EFFICIENT

From a survey of over 4,000 EMBL-EBI users, the efficency value placed on our services represents a direct worth of between £5,382 to £26,000 per respondant with the overall average that EMBL-EBI services allowed users to be 46% more efficient in their work.

Charles Beagrie All interviews and subsequent economic analysis were undertaken by Charles Beagrie Ltd. on behalf of EMBL-EBI.