Proceedings, 10th World Congress of Genetics Applied to Livestock Production

# Genomic prediction from whole genome sequence in livestock: the 1000 bull genomes project

**B.J. Hayes[1,2,3], I.M. MacLeod[3,4], H.D. Daetwyler[1,2,3], P.J. Bowman[2,3], A.J. Chamberlian[2,3], C.J. Vander Jagt[2,3], A. Capitan[5], H. Pausch[6], P. Stothard[7], X. Liao[7], C. Schrooten[8] , E. Mullaart[8], R. Fries[6], B. Guldbrandtsen[9], M.S. Lund[9], D.A. Boichard[5], R.F. Veerkamp[10], C.P. VanTassell[11], B. Gredler[12], T. Druet[13], A. Bagnato[14], J. Vilkki[15], D.J. deKoning[16], E. Santus[17], and M.E. Goddard[2,3,4].**

[1]Latrobe University, Melbourne, Australia, [2]Department of Environment and Primary Industries, Victoria, Australia, [3]Dairy Futures Co-operative Research Centre, Bundoora, Victoria, Australia, [4]University of Melbourne, [5]INRA UMR1313 Animal Genetics and Integrative Biology, [6]Chair of Animal Breeding, Technische Universität München, [7]University of Alberta, Edmonton, [8]CRV, 6800 AL, Arnhem, Netherland,[9]Aarhus University, Aarhus, Denmark, [10]Wageningen UR Livestock Research, [11]USDA-ARS-BFGL, [12]Qualitas AG, [13]Unit of Animal Genomics, GIGA , University of Liege, [14]Università degli Studi di Milano, [15]MTT Agrifood Research Finland, [16]Swedish University of Agricultural Sciences, Uppsala, Sweden, [17]ANARB, Italy

**ABSTRACT:** Advantages of using whole genome sequence data to predict genomic estimated breeding values (GEBV) might include better persistence of accuracy of GEBV across generations and more accurate GEBV across breeds. The 1000 bull genomes project provides a database of whole genome sequenced key ancestor bulls, that can be used for imputing sequence variant genotypes into reference sets for genomic prediction that are genotyped with SNP arrays. Run 3.0 of the 1000 bull genomes project included 429 sequences from 15 different breeds, with 31.8 million variants detected. Challenges with using this data in genomic predictions include the very large number of variants, accuracy of imputing sequence variants, and choice of method when deriving the prediction equation. Here we describe a new method that addresses at least some of these challenges, BayesRC, that takes advantage of biological information in genomic prediction from sequence data. In a dairy data set, predictions using BayesRC and imputed sequence data from the 1000 bull genomes were 2% more accurate than from an 800K data set, and we could demonstrate the method was able to identify causal mutations in some cases. Further improvements will come from more accurate imputation of sequence variant genotypes and improved biological information.

Keywords: Genomic prediction, whole genome sequence, biological information

## Introduction

Genomic prediction of breeding values is increasingly widely used in livestock breeding programs for dairy, beef, sheep, chickens and pigs (eg. Wiggans et al. 2011, Saatchi et al. 2012, Daetwyler et al. 2012, Hawken et al. 2014, Wellman et al. 2013). To date, most of these predictions have been based on arrays of approximately 50,000 SNP. The advantage of using whole genome sequence data for genomic prediction over such SNP arrays arises from the fact that with the sequence data, the actual casual mutations responsible for trait variation are now in the data set. This could have at least three benefits:

1) Better persistence of accuracy of genomic predictions over generations, and less erosion of accuracy in genomic predictions for individuals that are less related to the reference set. It has become clear that much of the accuracy of current genomic breeding values, based on 50,000 DNA markers, in fact derives from prediction of the effect of large chromosome segments that segregate within fairly closely related animals (eg. Habier et al. 2010). As a result, the accuracy of the prediction equation will rapidly decay over generations as large chromosome segments break up due to recombination. If the causal mutations were identified using the sequence data, and the prediction equation was based on their effects, accuracy would persist over many more generations, and in more distantly related animals. Macleod et al. (2014) and Meuwissen and Goddard (2010) could clearly demonstrate this in simulated sequence data.

2) Higher accuracy of genomic predictions. One issue with the current SNP arrays (eg. 50K, 800K for cattle) is that the SNP have been selected to have a high minor allele frequency. This means that the SNP arrays are less likely to have SNP in linkage disequilibrium with casual mutations where one of the alleles is at low frequency in the population. If this variation from rare alleles could be captured with the whole genome sequence data, and exploited in genomic predictions, accuracy of genomic breeding value may be able to be improved in the order of 2-30%, depending on trait (Druet et al. 2014). It should be pointed out that in populations with small effective population size, such as Holsteins, the gain in accuracy from using sequence data is likely to be limited, as the current 50K SNP arrays capture a large proportion of the genetic variance. For example for milk production, Haile-Mariam et al. (2013), and Jensen et al. (2012), the additive genetic variance captured by the 50K SNP was 80-90%. However, for some traits like fertility and survival, the genetic variance captured by SNP arrays is only ~ 60%, perhaps because there are more rare variants affecting these traits (Haile-Mariam et al. 2013). For these traits gains in accuracy from using sequence data might be higher.

3) More accurate genomic predictions across breeds. For a breed like Holstein dairy cattle, large reference populations have been assembled, leading to high accuracies of genomic predictions. For other breeds, assembling such large populations is challenging, so using genomic information across breeds would be appealing. However, the accuracy of genomic predictions across

breeds with the Bovine 50K array is close to zero, and this improves slightly when the 800K array is used (Erbe et al. 2012). With the whole genome sequence data, at least the causative mutations which do segregate across breeds could be captured and this information used in multi-breed genomic predictions. A multi-breed reference population would be required to achieve this. A multi-breed reference would also have the benefit from the fact that linkage disequilibrium across breeds is lower than that within breeds, so that causative mutations could be mapped more precisely (eg Raven et al. 2014). So multi-breed reference combined with sequence data should be the best approach to achieve the potential benefits from sequence data stated above.

In fact it becomes clear from the above that making best use of the sequence information in genomic prediction is actually a QTL mapping problem – the aim of both becomes to identify causative mutations.

In this paper, we review the whole genome sequence data that is available, in cattle at least, from the 1000 bull genomes project, discuss challenges with using such data in genomic predictions and how they can be overcome. We then present some preliminary results for accuracy of genomic predictions using sequence data.

## The 1000 bull genomes data set

A large reference set of animals with phenotypes and genotypes is required to estimate the genomic prediction equation. These reference sets must be large (10,000s) to achieve accurate genomic predictions, given the typical architecture of complex traits (large numbers of mutations of small effects). Such large numbers of animals are unlikely to be sequenced. An alternative strategy is to sequence key ancestors of the population, then impute the genotypes for the sequence variants into much larger reference sets with phenotypes and SNP array genotypes. The 1000 Bull Genomes Project is building this database of sequenced key ancestor bulls for the bovine research community (Daetwyler et al 2014). Run3.0 of the project included 429 full genome sequences of key ancestors from 15 breeds, sequenced at an average of 10.5 fold coverage, Table 1. There were 31.8 million filtered sequence variants detected in the sequences, including 29.1 million SNP and 1.7 million insertion-deletions. Various quality control steps were applied to determine both the accuracy of calling variants in the sequence data, and the accuracy of calling genotypes at the variants. This included assessing the rate of opposing homozygote genotypes for sire-son pairs, and assessing the agreement between sequence genotypes and genotypes from an 800K SNP array in the sequenced bulls. This was high, at 98.8%. The variants were annotated into different functional classes, including intergenic and intragenic, synonymous and non-synonymous, and other classes, Table 2. There was a considerable difference in allele frequency spectrum between the sequence variants, and the variants on the 800K Bovine HD Array, with many more sequence variants at low frequency. Missense (non-synonymous) variants were at more extreme allele frequencies than synonymous variants.

An interesting alternative approach to the strategy adopted by the 1000 bull genomes project (sequencing key ancestors at moderate fold coverage) would be to use the same sequencing effort to sequence a very large number of individuals at very low fold coverage. This strategy has been very successful in inbred plants like rice (Huang et al. 2010). However simulation suggests in outbred species (with heterozygote genotypes) a lower limit of 6 fold coverage is necessary to achieve accurate imputation of sequence variant genotypes into animals genotyped with SNP arrays (Druet et al. 2014).

## Challenges with using sequencing data in genomic prediction

There are at least three challenges for dealing with whole genome sequence data for genomic predictions; the very large number of variants, achieving accurate imputation of sequence variant genotypes into animals genotyped with SNP arrays, and choosing a method to estimate the genomic prediction equation that makes best use of the sequence data.

From Run 3.0 of the 1000 bull genomes project, 31.8 million variants were detected. While such large numbers of variants could be used in genome wide association studies (GWAS), because the analysis is highly parallelizable, for genomic prediction methods this presents a significant challenge. One option is to use biological information to prioritise or filter variants. This biological information comes in two forms, sites in the genome where variants are more likely to have an effect on any trait, for example coding regions or regulatory regions, and gene sets in which mutations are more likely to affect specific traits, for example genes expressed in mammary gland for milk production in dairy cattle. For the first type of information, analysis of genome annotations for enrichment of GWAS hits (eg. significant SNP) has identified coding regions, particularly missense mutationsand regions upstream and downstream of genes as enriched for trait associated variants in both humans and cattle (Kindt et al. 2013, Koufariotis et al. 2014). More recently, in human studies, regulatory regions (which are often in these upstream and downstream regions), have been shown to be enriched for GWAS hits for disease traits (Maurano et al 2012). For the second type of information, “Atlases” of bovine gene expression provide a means of identifying potentially important gene sets. The atlases define , genes which are differential expressed across tissues, using either digital gene expression tags or RNA Sequence data (Harhay et al. 2010, Chamberlain this proceedings respectively). The utility of both types of information for genomic prediction is demonstrated later in this paper.

Once a subset of sequence variants has been identified for further analysis, the next step towards genomic prediction is imputation of these genotypes into reference populations, typically already genotyped with SNP arrays. A major challenge here is accurate imputation of rare variants. To demonstrate this, we used cross

validation within the 1000 bull genomes data set to assess the accuracy of imputation. Randomly chosen subsets of sequenced Holstein animals had their sequence variant genotypes reduced to the 777K genotypes on the Bovine High density SNP array (there were 625,000 of these that were polymorphic and passed quality control). Beagle4.0 (Browning and Browning 2014) was used to impute full sequence variant genotypes in these animals (using all other sequenced animals as a reference). While the accuracy was reasonable for variants with minor allele frequency greater than 5%, below this, accuracy of imputation rapidly declined, Figure 1. This is in agreement with other studies on the same data (eg. Binsbergen this proceedings). Use of pedigree information only improved accuracy of imputing rare variants slightly (Figure 1), but other imputation programs may perform better with rare variants., Alternative methods are being developed to better utilise multiple breed information in imputation (Bouwman and Veerkamp this proceedings, Daetwyler et al. 2014). Improving the imputation accuracy of sequence variant genotypes, particularly rare variants, is necessary before the full advantage of sequence data for genomic predictions can be realized. Imputation of rare variants is likely to be important, because for example non-synonymous variants tend to have low minor allele frequencies (Figure 1).

The final challenge to be addressed is choice of statistical method to derive the genomic prediction equation that will make best use of the sequence data. Best linear unbiased prediction methods (BLUP as described by Meuwissen et al. 2001, or it's mathematical equivalent, GBLUP, Habier et al. 2007), are unlikely to make the best use of sequence data for two reasons. The first is that the prior used in these methods assumes all variants have an effect. One of the main attractions of using sequence data is to identify the causal variants, and remove all other variants from the prediction equation. The BLUP methods cannot arrive at this solution, as every variant will have a predicted non-zero effect derived from a single normal distribution. Another problem with BLUP methodologies is that the severe shrinkage imposed means that the effect of a causative mutation is rarely captured by a single variant, rather the effect is split across several or many SNP (eg. Verbyla et al. 2009). So we require methods that allow a proportion of the variants to have zero effect, and preferably a variable degree of shrinkage, such that moderate to large effects of causal mutations are not smeared across multiple variants that are in moderate LD with the QTL. Methods which meet both these criteria include BayesB (Meuwissen et al. 2001), and BayesR (Erbe et al. 2012). Both BayesB and BayesR have been tested on simulated sequence data (Clark et al. 2011 and Macleod et al. 2014), and lead to both higher genomic prediction accuracy and greater persistence of accuracy over generations than BLUP methods in those simulations. A drawback of the Bayesian methods however is that they are typically implemented using Gibbs sampling, or Metropolis Hastings algorithms in the case of BayesB, and such sampling algorithms are costly in computing time. Alternatives to sampling such as expectation-maximisation algorithms have been proposed for both BayesB (Shepherd et al. 2010), and BayesR (Wang et al these proceedings). These revised Bayesian methods can decrease computer processing time by up to 10 fold.

**Results from genomic prediction with sequence data, with and without biological knowledge**

The accuracy of genomic prediction using imputed sequence data was assessed using a reference population of 16,214 Holstein and Jersey cows and bulls (3049, 770 Holstein and Jersey bulls, and 8478, 3917 Holstein and Jersey cows, respectively). The phenotypes were daughter trait deviations for bulls, and trait deviations for cows, for milk volume, fat kg and protein kg. Phenotypes were weighted according to Garrick et al. (2009).

The validation population was 873 Dutch Red Holsteins, with phenotypes represented by de-regressed Interbull proofs on the Australian scale.

Genotypes were either imputed 632,003 SNP genotypes from the Illumina Bovine HD array, or imputed 1,674,245 sequence variant genotypes, where the variants were within genes or +/- 2kb from gene start and stop positions (a crude filtering of variants on biological information). Beagle 4.0 was used for imputation and one of all pairs of variants in complete linkage disequilibrium ($r^2$>0.99) were pruned out. .

Prediction equations were derived using either BayesR (Erbe et al. 2012), modified to use weights (Kemper et al. this proceedings), or a new method (BayesRC) that incorporates biological information, (MacLeod et al. this proceedings). BayesR assumes variant effects come from one of four normal distributions, one with zero variance (so zero effect), and the other three with increasing variance. BayesRC is a modified version of BayesR (Erbe et al 2012, Kemper et al 2014). The key modification in BayesRC requires that all variants used in the analysis are first allocated to one of several "classes" based on prior knowledge (for example non-synonymous mutations versus synonymous mutations), and the frequency with which SNP effects come from one of the four normal distributions is estimated separately for each class of SNP. This allows, for instance, the possibility that non-synonymous SNPs have a non-zero effect more often than synonymous SNPs (details in MacLeod et al. these proceedings). Biological information, used to define classes in BayesRC, was from two sources

1) Genome annotations. The three classes in BayesRC were then non-synonymous variants, (Table 1), variants in upstream and downstream regions and annotated microRNAs (Table 1), or 800K (variants on the BovineHD array and not in the two classes above. This analysis was called **Seq_BayesRC**,

2) Micro-array experiments to identify a set of genes in mammary gland that were found to play a role in milk synthesis, (Vander Jagt 2012). The classes here were 1) non synonymous coding variants in differentially expressed genes in the micro-array experiments 2) all other variants in genes differentially expressed in lactation experiments, and 3) all other sequence variants. This method was called **Lact_ BayesRC**.

We assessed the accuracy of prediction of BayesR using either the 800K genotypes ("**800K**"), using the imputed subset of sequence data ("**Seq BayesR**") and **Seq_BayesRC** and **Lact_ BayesRC**. The proxy for accuracy was the correlation of genomic predicted EBV and de-regressed proof in the 876 Red Holstein bulls.

The largest increase in accuracy for all traits was as a result of including cows in the reference population (~ 5%), Figure 2. Using sequence variants gave a 2% increase in accuracy over the 800K genotypes, averaged over traits. The BayesRC method did not give higher accuracy of genomic prediction in this data set, but identification of a relevant gene set could be further improved by including information from RNAseq atlases (Chamberlain et al. these proceedings).

The BayesRC method, did appear to lead to more precise mapping of QTL effects (MacLeod et al these proceedings), which is critical for persistence of accuracy of genomic predictions over generations. To assess the potential of BayesRC to identify causative mutations, we first investigated whether or not two previously identified mutations with effects on milk production traits were identified by the method. These were a mutation in the DGAT1 gene with an effect on fat%, fat kg, and milk volume (Grisart et al. 2002), and a mutation in the promoter of the PAEP gene (previously called Beta Lactoglobulin) with effects on protein percentage (Braunschweig and Leeb 2006). Both mutations were more clearly identified in Lact_ BayesRC, much more clearly than when BayesR was used, Figures 2 and 3 in MacLeod et al this proceedings. This is encouraging, especially since there are four 800K SNP in high linkage disequilibrium ($r^2$~0.75) with the DGAT1 causative mutation, but these SNP received a lower posterior probability in the BayesRC analysis, indicating that causative mutations can be identified even when there are other variants in relatively high linkage disequilibrium.

## Conclusion

Genomic prediction with whole genome sequence data is now possible for cattle. The 1000 bull genomes project provides a database of 31 million variant genotypes in key ancestor bulls that can be imputed into reference populations genotyped with SNP arrays, and genomic prediction methods that can deal with such large data sets are under development. One such method, BayesRC, takes advantage of biological information in genomic prediction from sequence data. In a dairy data set, predictions using BayesR or BayesRC and imputed sequence data from the 1000 bull genomes were 2% more accurate than from an 800K data set, and we could demonstrate BayesRC was able to identify causal mutations in some cases. Further improvements in accuracy of genomic prediction from sequence data, and particularly persistence of accuracy of these predictions across generations, will come from more accurate imputation of sequence variant genotypes, larger data sets, and improved biological information on sites and gene sets that are more likely to harbor mutations affecting quantitative traits.

## Literature Cited


Braunschweig, M.H., Leeb, T. (2006). J. Dairy Sci. 89:4414-9

Browning, B.L., Browning, S.R. (2013). Genetics 194:459-71

Clark, S.A., Hickey, J.M., van der Werf, J.H. (2011). Genet. Sel. Evol. 17:43:18

Daetwyler, H.D., Capitan, A., Pausch, H. et al. (2014). Nat. Gen. Accepted.

Daetwyler, H.D., Swan, A.A., van der Werf, J.H., Hayes, B.J. (2012). Genet. Sel. Evol. 44:33

Druet, T., Macleod, I.M., Hayes, B.J. (2014). Heredity 112:39-47

Erbe, M. Hayes, B.J., Matukumalli, L.K., et al. (2012) J. Dairy Sci. 95: 4114-29

Garrick, D.J., Taylor, J.F., Fernando, R.L. (2009). Genet Sel Evol. 31:41:55

Grisart, B., Coppieters, W., Farnir, F. (2002). Genome Res. 12:222-31

Habier, D. et al. (2010). Genet. Sel. Evol. 42:5

Habier, D. et al. (2011) BMC Bioinformatics 12:186

Habier, D., Fernando, R.L., Dekkers, J.C. (2007). Genetics 177:2389-97

Habier, D., Tetens, J., Seefried, F.R. (2010). Genet. Sel. Evol. 19:42:5

Haile-Mariam, M., Nieuwhof, G.J., Beard, K.T. et al. (2013). J. Anim. Breed. Genet. 130:20-31.

Harhay, G.P., Smith, T.P., Alexander, L.J. et al. (2010). 11:R102

Hawken, R.J. (2014). Proc. Plant Anim. Genome XXII. W520

Huang, X., Wei, X., Sang, T. et al. (2010). Nat Genet. 42:961-7

Jensen, J., Su, G., Madsen, P. (2012). BMC Genet. 13: 44

Kindt, A.S., Navarro, P., Semple, C.A., Haley, C.S. (2013). BMC Genomics 14:10

Koufariotis, R. Chen, Y.P., Bolormaa, S., Hayes, B.J. (2014). BMC Genomics. Submitted.

Macleod, I.M., Hayes, B.J., Goddard, M.E. (2014). Genetics

Maurano, M.T., Humbert, R., Rynes, E. (2012). Science 337:1190-1195

Meuwissen, T.H., Hayes, B.J., Goddard, M.E. (2001). Genetics 157:1819-29

Saatchi, M., Schnabel, R.D., Rolf, M.M. et al. (2012). Genet. Sel. Evol. 44:38

Shepherd, R.K., Meuwissen, T.H., Woolliams, J.A. (2010). BMC Bioinformatics. 11:529

Vander Jagt, C.J. (2012). PhD Thesis. University of Melbourne, Australia.

Verbyla, K.L. et al. (2009). Genet. Res. 91: 307-11

Wellmann, R., Preuss, S., Tholen, E. et al. (2013). Genet. Sel. Evol. 45:28

Wiggans, G.R., Vanraden, P.M., Cooper TA. (2011). J. Dairy Sci. 94:3202-11

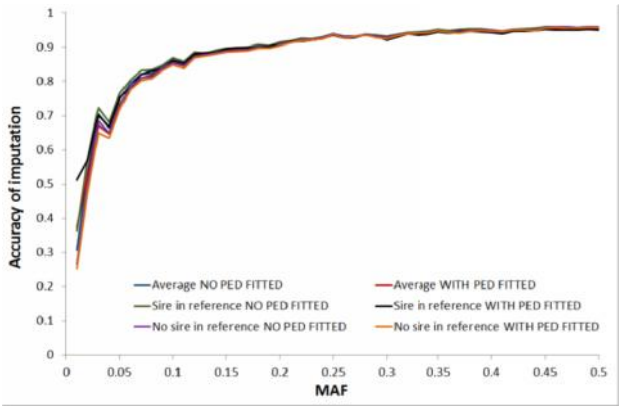


**Figure 1. Accuracy of imputing sequence variant genotypes by minor allele frequency (MAF), for Chromosome 26. The accuracy of imputing genotypes was assessed by cross validation, where sets of 25 Holstein animals were removed from the sequence dataset, and the genotypes for these animals were reduced to the genotypes that were on the Bovine HD array, then imputed all the sequence variants using Beagle4. The correlation between the real sequence variants and imputed variants were then calculated. Pedigree information was either included in the imputation (with ped fitted) or not included (no ped fitted).**

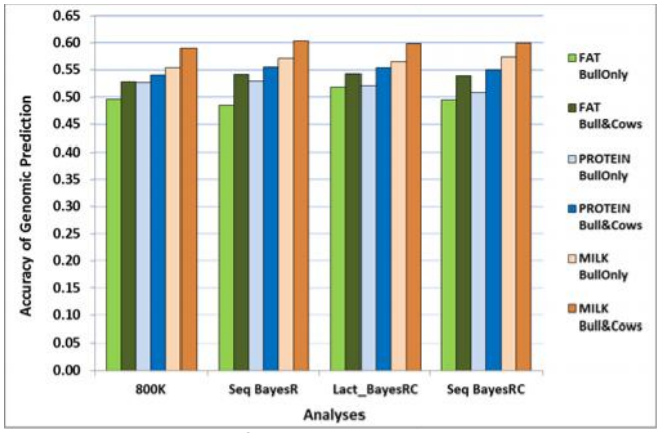


**Figure 2. Accuracy of genomic prediction in the validation group of Red Holstein (Dutch origin) using either Australian Bulls only (Holstein and Jersey) or Australian Bulls and Cows (Holstein and Jersey) reference sets. The proxy for accuracy was the correlation between de-regressed MACE proofs and genomic predictions.**

**Table 1**. Breeds and number per breed sequenced in Run 3.0 of 1000 bull genomes project.

| Breed | Number sequenced |
|---|---|
| Holstein | 122 |
| Jersey | 26 |
| Simmental | 87 |
| Angus | 54 |
| Swedish Reds | 16 |
| Piedmontese | 2 |
| Limousin | 25 |
| Hereford | 1 |
| Guelph Composite | 9 |
| Finnish Aryshire | 17 |
| Charolais | 8 |
| Brown Swiss | 43 |
| Belgian Blue | 10 |
| Beef Booster | 8 |
| All | 429 |

**Table 2A. Annotation of SNP from Run 3.0 of the 1000 bull genomes project.**

| Annotation | Number |
|---|---|
| intergenic_variant | 19277503 |
| intron_variant | 7587343 |
| upstream_gene_variant | 1007825 |
| downstream_gene_variant | 879304 |
| missense_variant | 119236 |
| synonymous_variant | 109805 |
| 3_prime_UTR_variant | 68408 |
| splice_region_variant | 22713 |
| 5_prime_UTR_variant | 12800 |
| stop_gained | 3083 |
| splice_donor_variant | 2391 |
| non_coding_exon_variant | 9894 |
| splice_acceptor_variant | 1629 |
| initiator_codon_variant | 235 |
| stop_lost | 105 |
| coding_sequence_variant | 184 |
| stop_retained_variant | 74 |
| mature_miRNA_variant | 273 |
| nc_transcript_variant | 70 |
| Total | 29102875 |

**Table 2B. Annotation of SNP from Run 3.0 of the 1000 bull genomes project.**

| Annotation | Number |
|---|---|
| Intergenic | 1135727 |
| intron_variant | 460164 |
| upstream_gene_variant | 64332 |
| downstream_gene_variant | 57775 |
| 3_prime_UTR_variant | 4823 |
| frameshift_variant | 1585 |
| splice_region_variant | 1207 |
| inframe_deletion | 998 |
| 5_prime_UTR_variant | 834 |
| inframe_insertion | 349 |
| splice_acceptor_variant | 160 |
| splice_donor_variant | 120 |
| missense_variant | 88 |
| non_coding_exon_variant | 406 |
| coding_sequence_variant | 59 |
| nc_transcript_variant | 19 |
| stop_gained | 8 |
| mature_miRNA_variant | 45 |
| stop_lost | 1 |
| Total | 1728700 |