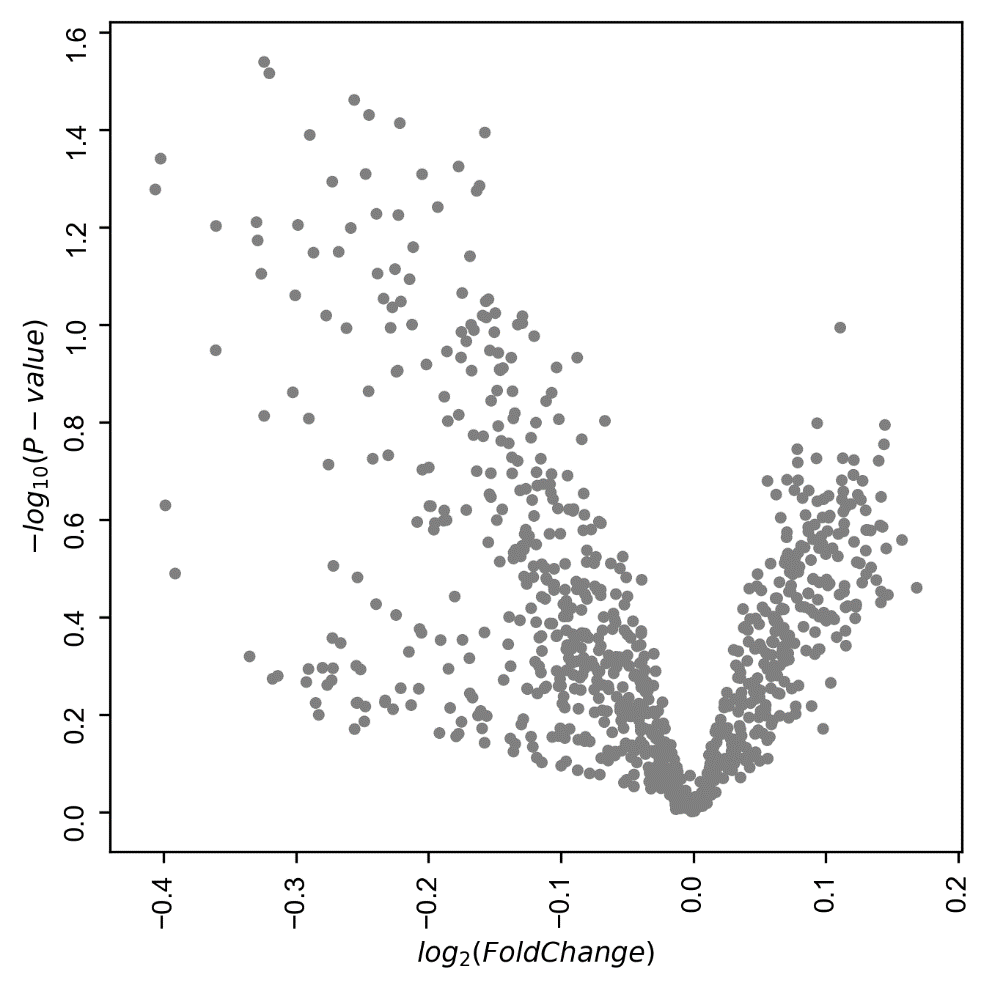
1 Introduction and project areaPredicting the functional impact of genetic mutations is key to develop personalized medicine strate-gies for treating complex diseases with a strong genetic component (i.e. Cancer, Type II Diabetes)[1]. Mutations predisposing an individual with a to certain diseases can be identified through experi-mental and observational genome wide association studies (GWAS). However, the exact mechanismby which the mutation causes the disease is often unknown. Moreover, finding causal mutationswithin the very large number of silent mutations is challenging. This is especially true for single nu-cleotide polymorphisms (SNPs), which are very common single nucleotide substitutions ubiquitouslydistributed throughout the whole genome. It is thought that about half of all SNPs are located innon-coding regions of the genome [1], indirectly interfering with protein expression by modulatingregulatory mechanisms such as transcriptional activity or epigenetic marks. To gain insight into therole of these SNPs, many research groups have developed bioinformatics tools aimed to characterizefunctional SNPs at the genome-scale. Although these efforts had long been limited to the study ofSNPs located in coding regions [2], the increasing availability of large experimental datasets like theEncyclopedia of DNA elements (ENCODE, [3]) have allowed researchers to address the problem ofnon-coding variant prediction.With large datasets come new computational methods to solve these problems. In particular, su-pervised machine learning models efficiently extract meaningful information from large datasetscontaining many different features [4]. This is particularly useful for solving complex classificationproblems where the final decision depends on many different parameters. Some of these algo-rithms are used to build models which predict the functional effects of sequence variants, a problemwhere integrating a wide range of different data types and formats such as contextual chromatininformation is crucial. For example, Ritchie et al. [2] used a Random Forest classifier to identifyfunctionally relevant variants in non-coding regions. In the present study, a Deep Learning modelbased on the convolutional neural network architecture [5] was used to solve the variant predictionproblem. The functional ’chromatin features’ predicted by the model include transcription factorbinding affinity, histone marks and DNase I-hypersensitive sites. The model is trained on sequencesannotated for these features and learns associations between them. This results in a model capableof simultaneously predicting the functional effects of non coding variants on 919 different chromatinfeatures.2 Model and data2.1 DeepSEA model descriptionThe input data to the DeepSEA is generally a set of 1000-bp sequences taken from anywhere inthe genome with coordinates in the GRCh37 reference genome format [6]. Each input 1000-bpsequence is one hot-encoded to represent the categorical variables (nucleotides) as binary vectors.The model itself is a classifier and is composed of three parts. The first part extracts featuresby successively applying convolutional filters to the input. The second part integrates all of theinformation extracted in the previous one through a fully connected layer. The final part uses thesigmoid activation function to compute an output probability vector of length 919, which correspondsto the number of chromatin features predicted. To predict the effect of a single nucleotide variant(SNP), one can feed the model with the SNP information in the right format containing: the positionon the chromosome, the reference allele and the alternative allele. The model considers this inputas two 1000-bp sequences each centered around the reference and alternative allele and computestwo output probability vectors for each. To determine if a chromatin feature is enriched in the SNP,Log2 fold changes between the reference and alternative output probability vectors are computed.2

BIO-463 -Genomics and Bioinformatics2.2 Data, training and testing procedureThe data used to train the model is taken from the ENCODE and Roadmap Epigenomic projects.It consists of a set 919 chromatin profiles in narrowPeak (BED4+6) format: 104 histone marks(HM) profiles, 125 Dnase Hypersentitive Sites (DHS) profiles and 690 Transcription factor (TF)profiles. Each file contains called peaks based on pooled and normalized data. HM and TF profileswere obtained from Histone Chip-seq (TF Chip-seq) experiments and DHS from DNase-seq. Qualitycontrols such as the number biological replicates, the number of fragments per replicate, evaluationof antibody quality and input controls were all performed before data release. The training input forthe DeepSEA model is a set of n 1000-bp sequences each paired with a (1x919) vector of booleanvalues: ’1’ if the chromatin feature is enriched in the input sequence and ’0’ otherwise. The weightsof this model are calculated by minimizing an objective function (negative log likelihood) usingstochastic gradient descent. The output of each sequence sample is a (1x919) vector of normalizedprobabilities for all chromatin features. A higher probability for a given chromatin feature in theoutput vector indicates that this chromatin feature is more likely to be enriched in the input sequence.As with any classifier, the decision to classify a sample as positive or negative for a given featuredepends on a predefined threshold.3 Methods and Results3.1 Evaluating the DeepSEA modelThe aim of this section is to evaluate the performance of the DeepSEA model. For this task, wefirst generated a test dataset containing a total of 5000 1000-bp long DNA sequences randomlylocated on chromosome 9 spanning coordinates 30,000,924 - 38,000,661. Note that because weare only interested in the genomic location along the chromosome, the strand orientation is notimportant. Given that chromosome 9 was not used for training the model in the original paper,it should provide an unbiased evaluation of the model’s performance. For each sample in the testdataset, we calculated a label vector for the 919 chromatin features using the same method as theauthors: a feature was labeled one if at least half of the central 200-bp in the 1000-bp sequence iscontained in in the peak region of the corresponding chromatin profile and 0 otherwise.- what metric is used to evaluate mode performance ? (describe ROC curves and AUC scores)



E