Black Sea Journal of Engineering and Science
doi: 10.34248/bsengineering.1387431

Open Access Journal
e-ISSN: 2619 – 8991

Research Article
Volume 7 - Issue 2: 155-159 / March 2024

A COMPARATIVE EVALUATION OF THE OUTLIER DETECTION METHODS

Melis ÇELİK GÜNEY[1]*, Gökhan Tamer KAYAALP[1]

[1]Çukurova University, Faculty of Agriculture, Department of Animal Science, 01330, Adana, Türkiye

Abstract: In data mining, in order to calculate descriptive statistics and other statistical model parameters correctly, outliers should be identified and excluded from the data set before starting data analysis. This paper studied and compared the performance of model-based, density-based, clustering-based, angle-based, and isolation-based outlier detection methods used in data mining. ROC and AUC curves were used to compare the performances of outlier detection methods. A data set with a standard normal distribution and fit a logistic regression was simulated. To compare the methods, the data was modified by randomly adding 30 outliers to the data set. The iForest algorithm was found to have higher predictive power than Mahalanobis, LOF, k-means, and ABOD. In addition, outliers were found in a real data set with the iForest algorithm and deleted from the data set. Then, the data sets with outliers and without outliers were compared. The results showed that the model without outliers has a higher predictive ability.

Keywords: Outlier, LOF, iForest, ROC curve, Data mining

***Corresponding author:** Çukurova University, Faculty of Agriculture, Department of Animal Science, 01330, Adana, Türkiye
E mail: celikm@cu.edu.tr (M. ÇELİK GÜNEY)
Melis ÇELİK GÜNEY https://orcid.org/0000-0002-6825-6884
Gökhan Tamer KAYAALP https://orcid.org/0000-0003-2193-848X

Received: November 22, 2023
Accepted: January 08, 2024
Published: March 15, 2024

Cite as: Çelik Güney M, Kayaalp GT. 2024. A comparative evaluation of the outlier detection methods. BSJ Eng Sci, 7(2): 155-159.

1. Introduction

The technique of simulating human intelligence with algorithms to create a new computer that can do the work that humans can do is defined as artificial intelligence (AI) (Bharadiya, 2023). Machine learning (ML) is a collection of algorithms that computers use to generate and refine predictions or behaviors based on data (Molnar, 2019). Logistic regression analysis, one of the machine learning methods, is used frequently in many fields.

In multi-category or ordinal scales, logistic regression forecasts the value of the dependent variable examines the connection between dependent and independent variables, and makes classification (Mertler and Vannatta, 2005).

Outliers are observations that are significantly different from other observations (Cebeci et al., 2022). These values may be due to incorrect entry of records, fraudulent behavior, humans or instruments error or a natural deviation in the population (Hodge and Austin, 2004). The frequent occurrence of outliers has increased interest in outlier detection methods in data mining.

Outlier detection methods are classified as univariate or multivariate; parametric, semi-parametric, and nonparametric; supervised, semi-supervised, and unsupervised. It is also classified as density-based, clustering-based, distance-based, and depth-based outlier detection methods (Ben-Gal, 2005; Gogoi et al., 2011; Yucel Altay, 2014; Cebeci, 2020; Cebeci et al., 2022). Statistical-based, deviation-based, and subspace-based outlier detection methods can also be added to this classification (Xu et al., 2018). There is also an isolation-based outlier detection method that has been actively used recently and has high performance (Liu et al., 2008). The aim of the study is to compare the performance of model-based method Mahalabonis distance, density-based method LOF, clustering-based method k-means, angle-based method ABOD, and isolation-based outlier detection method iForest used in data mining.

2. Materials and Methods

2.1. Dataset

The data set was generated a sample of size 3000 from a standard normal distribution using the R package. In this data set, there are 3 independent (X_1, X_2, X_3) and 1 dependent (Y) variable. The independent variables consist of continuous variables and the dependent variable consists of a binary variable containing 0 and 1 values. In order to compare the outlier detection methods, a total of 30 outliers, 10 in each independent variable, were randomly added to the data set and the data were modified.

The Asian rice (*Oryza sativa*) data obtained by Zhao et al. (2011) were used. From this data set, 282 observations were included in the analysis. From this data set, seed length (X_1), seed width (X_2), and seed volume (X_3) were used as independent variables and leaf pubescence (Y) was used as the dependent variable. The independent variables consist of continuous variables and the dependent variable is a categorical variable (no pubescence: 0, pubescence: 1). The simulated and real data set split of training (70%) and test (30%) set.

2.2. Logistic Regression

Regression analysis is used to determine the relationship between dependent and independent variables. In this analysis, the dependent variable consists of continuous data. If the dependent variable is a categorical or ordinal, logistic regression analysis is needed. In logistic regression analysis, independent variables can be discrete, continuous or a mixture of these variables.

In binary logistic regression, one of the logistic regression model types, the dependent variable is two categories. The logistic regression model (Equation 1) is as follows.

$$In(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \qquad (1)$$

where π is the likelihood of the event, β_0 is the constant term, β_k is the kth regression coefficient, X is the independent variable, and Y is the dependent variable (Juarto, 2023).

2.3. Outlier Detection Methods

2.3.1. Mahalanobis distance

Statistical methods (model-based methods) for outlier detection are divided into parametric methods and nonparametric methods (Han et al., 2012). In parametric methods, multivariate outlier detection using Mahalanobis distance (Equation 2) is performed as follows (Rousseeuw and Van Zomeren, 1990).

$$M = \sqrt{(x-\mu)\Sigma^{-1}(x-\mu)'} \qquad (2)$$

Where M is Mahalanobis distance, x is vector of variables $x = (x_{1,}x_{2,...,}x_k)$, $\mu = (\mu_{1,}\mu_{2,...,}\mu_k)$ is vector of mean values, Σ is the covariance matrix (Leys et al., 2017).

If the squared Mahalanobis distances of each observation is greater than the quantile (e.g. 0.975 quantile) of the χ^2 distribution, these observations are outliers (Rousseeuw and Van Zomeren, 1990; Prykhodko et al., 2018; Cebeci, 2020).

The Mahalanobis distances were examined using the “Moutlier” function of the “chemometrics” package in R (Filzmoser and Varmuza, 2017).

2.3.2. Local outlier factor (LOF):

Proximity-based outlier detection methods are divided into density-based and distance-based outlier detection methods (Han et al., 2012). One of the density-based outlier detection algorithms is the LOF. This algorithm is based on ⍰ nearest neighbors (Breunig et al., 2000).

The distance between an object O and its nearest ⍰ neighbors is the k-distance of O, denoted by $\mathrm{dist}_k(O)$, and the ⍰-distance neighborhood of an object O (Equation 3) is as follows.

$$N_k(O) = \{O' | O' \in D, dist(O,O') \le dist_k(O)\} \qquad (3)$$

The reachable distance is defined as the maximum of k-distance of O' and the distance between O and O', and its formula (Equation 4) is:

$$reachdist_k(O \leftarrow O') = max\,\{dist_k(O), dist(O,O')\} \qquad (4)$$

The local reachability density of object O (Equation 5) can be written as follows (Hofmann and Klinkenberg, 2014).

$$ldist_k(O) = \frac{\|N_k(O)\|}{\sum O'\epsilon N_k(O)\; reachdist_{\boldsymbol{k}}(O \leftarrow O')} \qquad (5)$$

The local outlier factor (LOF) of O can be written as follows (Equation 6).

$$LOF_k(O) = \frac{\sum O' \in N_k(O)\;.\;ldist_{\boldsymbol{k}}(O')}{\|N_k(O)\|\;.\;ldist_k(O)} \qquad (6)$$

Observations that have a substantially lower density than their neighbors are identified as outliers (Cebeci, 2020).

The local outlier factors were examined using the “lof” function of the “Rlof” package in R (Hu et al., 2015).

2.3.3. k-means algorithm

Clustering-based outlier detection methods are divided into hierarchical and non-hierarchical clustering methods. In the k-means, which is one of the outlier detection methods with non-hierarchical clustering, the aim is to divide n objects into k number of clusters and to minimise the similarity between clusters and maximise the similarity within clusters (Yadav and Sharma, 2013; Deb and Dey, 2017).

Where k is no. of the cluster, D is a dataset containing n objects, the steps of the algorithm are given below (Kaya and Koymen, 2008; Han et al., 2012; Bertizlioglu and Ozgonenel, 2012).

1. Initially, to determine the cluster center, n objects are randomly selected from D to form the number k of clusters.
2. The average of each object is calculated and the center points are determined.
3. Based on the mean values, each object is grouped with the closest center point.
4. The new mean value of each data item is calculated.
5. Step 2 and 3 are repeated until k does not change.

In order to determine outliers with hierarchical clustering, after the 3rd step, the distances of each object are calculated by summing the squares of the deviations from the center of the cluster to which each object belongs and taking the square root. The objects with the maximum distance are considered outliers (Cebeci, 2020).

The distances of each object were examined using the “kmeans” function of the “Stats” package in R.

2.3.4. Angle-based outlier detection (ABOD)

In Angle-based outlier detection method, the variances of the angles between the difference vectors of the data objects are taken into account (Kriegel et al., 2008).

In dataset D, when one point $\vec{\mathrm{A}} \in \mathrm{D}$ ($\vec{\mathrm{A}} = (\mathrm{A}_1, \mathrm{A}_2, \ldots \mathrm{A}_\mathrm{n})$) and two other points $\vec{\mathrm{B}}, \vec{\mathrm{C}} \in \mathrm{D}$ and $\vec{\mathrm{B}}, \vec{\mathrm{C}} \neq \mathbf{A}$, the Angle-Based Outlier Factor (ABOF) is calculated by Equation 7.

$$ABOF(\vec{A}) = Var_{\vec{B},\vec{C}\in D}(\frac{(\overline{AB},\overline{AC})}{\|\overline{AB}\|^2.\|\overline{AC}\|^2} \qquad (7)$$

where $\overline{\mathrm{AB}}$ is the difference vectors $(\vec{\mathrm{B}} - \vec{\mathrm{A}})$, $\|\overline{\mathrm{AB}}\|$ is the Euclidean distance between $\vec{\mathrm{A}}$ and $\vec{\mathrm{B}}$, $\mathrm{ABOF}(\vec{\mathrm{A}})$ is the variance over the angles between the difference vectors of $\vec{A}$ to all pairs of points in D weighted by the distance of the points (Kriegel et al., 2008). The ABOF values found by equation 7 are ranked and those that are smaller than the ABOF values of other observations are called outliers. The ABOFs were examined using the "abod" function of the "abodOutlier" package in R (Jimenez, 2015).

2.3.5. Isolation forest (iForest)

The iForest algorithm, which is one of the isolated-based outlier detection methods, used to calculate the anomaly score of a data point is based on the observation that the isolation tree (iTrees) structure is equivalent to binary search trees (BST). The anomaly score of a data point is calculated by Equations 8 and 9 (Liu et al., 2008; Negi, 2020).

$$c(m) = \begin{cases} 2\mathrm{H}(\mathrm{m}-1) - \dfrac{2(\mathrm{m}-1)}{\mathrm{n}}, & m > 2 \\ 1, & m = 2 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

$$\mathrm{s}(\mathrm{x},\mathrm{m}) = 2^{-\frac{\mathrm{E}(\mathrm{h}(\mathrm{x}))}{\mathrm{c}(\mathrm{m})}} \tag{9}$$

where s is anomaly score, h(x) is the path length of an x observation, E(h(x)) is the average of h(x) of the iTrees set, c(m) is average length of unsuccessful search in BST, H is a harmonic number, and n is the no. of external nodes.

The determination of whether or not the observations are outliers is based on the anomaly score. An observation is considered an outlier if its anomaly score is near 1; if it is near 0.5, it is not. The value is normal value if the anomaly score is much less than 0.5 (Liu et al., 2008).

The anomaly scores were examined using the "iForest" function of the "isofor" package in R (Graves and Drozdov, 2019).

2.4. Comparison of the Performance of Methods

In this paper, the receiver operating characteristic (ROC) curve is used for measuring the performance of the outlier detection methods. The confusion matrix used when plotting the ROC curve is given in Table 1 (Sharma et al., 2022).

Table 1. Confusion matrix

	Actual Positive	Actual Negative
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative

When plotting the ROC curve, the true positive rate (TPR) should be available in addition to the false positive rate (FPR). TPR and FPR are formulated in Equations 10 and 11 (Omar and Nassif, 2023).

$$\mathrm{TPR} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{10}$$

$$\mathrm{FPR} = \frac{\text{False Positive}}{\text{False Positive} + \ \text{True Negative}} \tag{11}$$

The Area Under the Curve (AUC) is defined as the area under the ROC curve and shows the percentage of correct classification of positive and negative results. A larger AUC value means better performance (Auslander et al., 2011). A lower FPR and a higher TPR are desired because of admirable predictive prowess (Hou et al., 2023).

3. Results and Discussion

In simulated data, the AUCs of the different methods were compared in Figure 1. The AUC of the iForest model was found to be higher than the Mahalanobis, LOF, k-means, and ABOD models. This demonstrated that the iForest model has higher predictive power compared to the other models. The AUC of the LOF model was found the second highest AUC. The iForest has higher predictive power than the LOF (Gao et al.,, 2019; Gnat, 2020; Negi, 2020; Kiruthika and Sowmyarani, 2020; Vijayakumar et al., 2020).

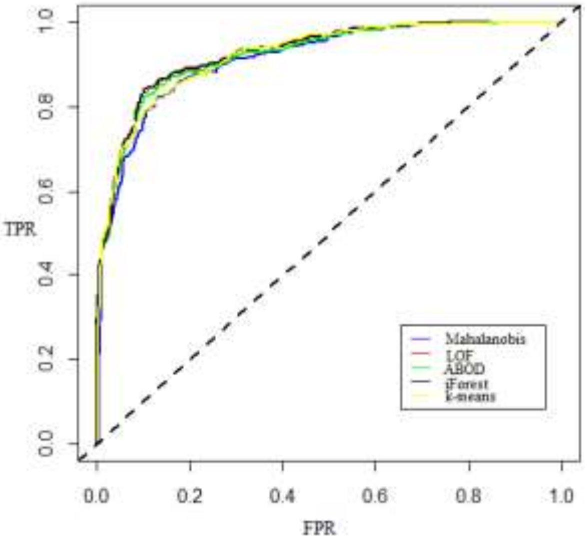

Figure 1. The ROC curve.

Since the iForest algorithm had higher predictive power than the other models, outliers were found using iForest in the real data. According to the outlier scores found by Equations 8 and 9, seven observations were identified as outliers and deleted from the data set. Then, the logistic models were developed for the data set with and without outliers and the ROC curves of these models were plotted in Figure 2.

In Figure 2, The AUC of without outliers and with outliers logistic regression models were 0.8526 and 0.7841. The AUC of the without outliers model was found to be higher than the outlier model. In other words, predictive modeling and classification are not reliable in the data set with outliers. Nurunnabi and West (2012) compared outlier and without outlier data sets and reported that the results of the outlier data set were not reliable in logistic regression analysis. Osborne and Amy (2004) reported that outliers significantly affect the analysis.

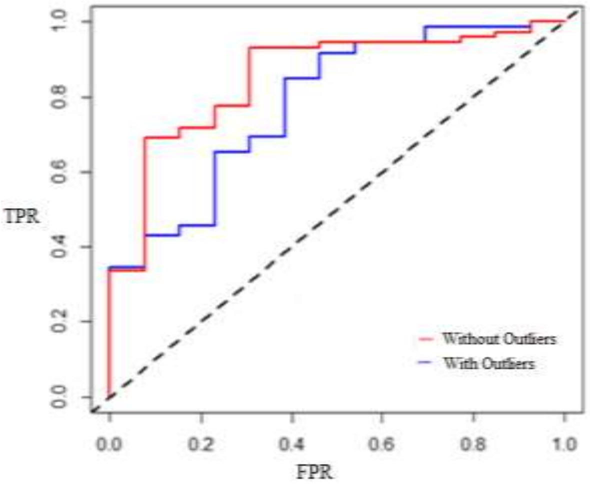

Figure 2. ROC curve of the data set with and without outliers.

4. Conclusion

In this study, the performances of Mahalanobis, LOF, k-means, ABOD, and iForest methods were compared. iForest algorithm was found to have a higher predictive power compared to the other methods. It is also concluded that outliers in logistic regression analysis affect the model considerably.

Author Contributions

The percentage of the author(s) contributions is presented below. All authors reviewed and approved the final version of the manuscript.

	M.Ç.G.	G.T.K.
C	50	50
D	50	50
S		100
DCP	100	
DAI	100	
L	100	
W	80	20
CR	20	80
SR	100	

C=Concept, D= design, S= supervision, DCP= data collection and/or processing, DAI= data analysis and/or interpretation, L= literature search, W= writing, CR= critical review, SR= submission and revision.

Conflict of Interest

The authors declared that there is no conflict of interest.

Ethical Consideration

Ethics committee approval was not required for this study because of there was no study on animals or humans. The authors confirm that the ethical policies of the journal, as noted on the journal's author guidelines page, have been adhered to.

Acknowledgements

We gratefully thank to Prof. Dr. Zeynel CEBECİ at the Çukurova University for his contributions in this study.

We would like to thank Cukurova University Scientific Research Coordinatorship for supporting this study with project number FDK-2018 10287.

It was produced from the thesis titled “Comparative Examination of Outlier Detection Methods in Binary Logistics Regression Analysis” at Cukurova University Thesis no: 794371. https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp.

References

Auslander B, Gupta KM, Aha DW. 2011. A comparative evaluation of anomaly detection algorithms for maritime video surveillance. Proceedings of the Society of Photographic Instrumentation Engineers Conference, June 15-17, Orlando, US, Vol. 8019, pp: 27-40.

Bharadiya JP. 2023. A comparative study of business intelligence and artificial intelligence with big data analytics. American J Artific Intel, 7(1): 24-30.

Ben-Gal I. 2005. Outlier detection. In Data Mining and Knowledge Discovery Handbook, Springer, Boston, US, pp: 288.

Bertizlioglu IN, Ozgonenel O. 2012. Blackout detection using k-means clustering method. ELECO'2012 Electrical and Electronics Engineering Symposium, November 29-December 1, Bursa, Turkiye.

Breunig MM, Kriegel HP, Ng RT, Sander J. 2000. LOF: Identifying Density-Based Local Outliers. In ACM Sigmod Record, 29(2): 93-104.

Cebeci Z. 2020. Data preprocessing with R in data science. Nobel Academic Publishing, Ankara, Türkiye, opp: 552.

Cebeci Z, Cebeci C, Tahtali Y, Bayyurt L. 2022. Two novel outlier detection approaches based on unsupervised possibilistic and fuzzy clustering. PeerJ Comp Sci, 8: e1060.

Deb AB, Dey L. 2017. Outlier detection and removal algorithm in k-means and hierarchical clustering. World J Comp Appl Technol, 5(2): 24-29.

Filzmoser P, Varmuza K. 2017. Chemometrics: Multivariate Statistical Analysis in Chemometrics. URL: https://CRAN.R-project.org/package=chemometrics. (accessed date: February 10, 2023).

Gao R, Zhang T, Sun S, Liu Z. 2019. Research and improvement of isolation forest in detection of local anomaly points. J Physics: Conf Series, 1237(5): 1-6.

Gnat S. 2020. Testing the effectiveness of outlier detecting methods in property classification. Real Estate Manag Valuat, 28(4): 81-92.

Gogoi P, Bhattacharyya D, Borah B, Kalita JK. 2011. A survey of outlier detection methods in network anomaly identification. Comput J, 54(4): 570-588.

Graves E, Drozdov I. 2019. Zelazny7/isofor: Isolation forest anomaly detection. URL: https://github.com/Zelazny7/isofor. (accessed date: February 01, 2023).

Han J, Pei J, Pei J. 2012. Data mining: concepts and techniques, Third Edition. Morgan Kaufmann Publishers Elsevier, US, pp: 744.

Hou S, Gao J, Wang C. 2023. Order acceptance choice modeling of crowd-sourced delivery services: a systematic comparative study. URL: https://www.techrxiv.org/doi/full/10.36227/techrxiv.24139491.v1 (accessed date: February 23, 2023).

Hodge V, Austin J. 2004. A survey of outlier detection methodologies. Artific Intel Rev, 22(2): 85-126.

Hofmann M, Klinkenberg R. 2014. RapidMiner: Data mining use cases and business analytics applications. CRC Press, New York, US, pp: 528.

Hu Y, Murray W, Australia YS. 2015. Rlof: R parallel implementation of local outlier factor (LOF). URL: https://CRAN.R-project.org/package=Rlof (accessed date: January 12, 2023).

Jimenez J. 2015. abodOutlier: angle-based outlier detection. URL: https://CRAN.R-project.org/package=abodOutlier (accessed date: January 12, 2023).

Juarto B. 2023. Breast Cancer classification using outlier detection and variance inflation factor. Eng Math Comp Sci J, 5(1): 17-23.

Kaya H, Koymen K. 2008. Data mining concept and application areas. Fırat Univ Doğu Araşt Derg, 6(2): 159-164.

Kiruthika S, Sowmyarani CN. 2020. Credit card fraud detection using machine learning and deployment of model in public cloud as a web service. Int J Recent Technol Eng, 9(2): 548-552.

Kriegel HP, Schubert M, Zimek A. 2008. Angle-based outlier detection in high-dimensional data. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, Las Vegas, US, pp: 444-452.

Leys C, Klein O, Dominicy Y, Ley C. 2017. Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. J Exp Soc Psychol, 74: 150-156.

Liu FT, Ting KM, Zhou ZH. 2008. Isolation forest. Eighth IEEE International Conference on Data Mining, December 15-19, Pisa, Italy, pp: 413-422.

Mertler CA, Vannatta RA. 2005. Advanced and multivariate statistical methods: practical application and interpretation, 3rd edition. Glendale, Pyrczak Publishing, Los Angeles, US, pp: 234.

Molnar C. 2019. Interpretable machine learning: a guide for making black box models explainable. URL: https://christophm.github.io/interpretable-ml-book/ (accessed date: September 20, 2023).

Negi SS. 2020. Early prediction of credit card fraud detection using isolation forest tree and local outlier factor machine learning algorithms. A Project Report of Capstone Project-2. Galgotias University, Uttar Pradesh, India, Act No: 14.

Nurunnabi A, West G. 2012. Outlier detection in logistic regression: A quest for reliable knowledge from predictive modeling and classification. IEEE 12th international conference on data mining workshops, December 10, pp: 643-652.

Omar AAC, Nassif AB. 2023. Lung cancer prediction using machine learning based feature selection: a comparative study. Advances in Science and Engineering Technology International Conferences (ASET), February 20-23, pp: 1-6.

Osborne JW, Amy O. 2004. The power of outliers (and why researchers should always check for them). Pract Asses Res Eval, 9(6): 1-12.

Prykhodko S, Prykhodko N, Makarova L, Pukhalevych S. 2018. Application of the squared mahalanobis distance for detecting outliers in multivariate non-Gaussian data. 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET), February 20-24, Lviv-Slavske, Ukraine, pp: 962-965.

Rousseeuw PJ, Van Zomeren BC. 1990. Unmasking multivariate outliers and leverage points. J American Stat Assoc, 85(411): 633-639.

Sharma DK, Chatterjee M, Kaur G, Vavilala S. 2022. Deep learning applications for disease diagnosis. Academic Press, Cambridge, US, pp: 31-51.

Vijayakumar V, Divya NS, Sarojini P, Sonika K. 2020. Isolation forest and local outlier factor for credit card fraud detection system. Int J Eng Adv Technol, 9(4): 261-265.

Xu X, Liu H, Li L, Yao M. 2018. A comparison of outlier detection techniques for high-dimensional data. Int J Comput Intel Syst, 11(1): 652-662.

Yadav J. Sharma M. 2013. A review of k-mean algorithm. Int J Eng Trends Technol, 4(7): 2972-2976.

Yucel Altay S. 2014. Using of spatio-temporal data mining for trajectory outlier detection and interpretation in health care services. MS Thesis, Atatürk University, Graduate School of Natural and Applied Sciences, Erzurum, Türkiye, pp: 25-32.

Zhao K, Tung CW, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J, McClung AM, Bustamante CD, McCouch SR. 2011. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nature Commun, 2(1): 467.