Stable Feature Ranking with Logistic Regression Ensembles

Ronald J. Nowling and Scott J. Emrich
Computer Science & Engineering
University of Notre Dame
Notre Dame, IN 46656
Email: rnowling@nd.edu, semrich@nd.edu

***Abstract*—Beyond automated classification, supervised machine-learning models can be *interpreted* to find which features or combination of features distinguish sets of classes. Logistic Regression (LR) is an example of a model well-suited for human interpretation. Unfortunately, results from feature ranking with LR may not be reliable and reproducible for the same dataset. We demonstrate that stability and consistency can be achieved via ensembles ("LR ensembles"). As a specific example of the real-world utility of our associated framework, we apply LR ensembles to single-nucleotide polymorphisms (SNPs) associated with the recent speciation of the malaria vectors *Anopheles gambiae* and *Anopheles coluzzii* and compare with the more common univariate metric F_{ST}.**

I. Introduction

Supervised machine-learning (ML) models are most commonly used to "learn" patterns from labeled data and then use the learned patterns to predict the associated classes of unlabeled data. To extract patterns that distinguish input classes, ML models separate out predictive input features (variables) from the rest. Beyond being used just for automated classification, supervised ML models can be *interpreted* to find which features or combination of features distinguish the classes and lead to human understanding and insight.

Interpretable models have many applications in bioinformatics and are often used in genome-wide association studies (GWAS). In this context, features are engineered from observed variants and a predictive model is trained on individuals labeled by their characterized phenotypes. The model is then interpreted to find variants that may be associated with the differences.

Although single-SNP association tests [1] or univariate measures such as F_{ST} for population-wide differences [2], [3] are more common, interpretable machine learning models, particularly Logistic Regression (LR), have been applied successfully to identify important and relevant genetic variants [4]–[14].

Unfortunately, it has been independently observed that feature selection with Logistic Regression can be unreliable and unstable with highly-correlated features, which are expected in gene variant analysis. Toloşi, et al. [15] haved proposed the application of clustering to identify and collapse correlated features[1]. However, standard clustering methods such as *k*-means are ill-suited for categorical variables, which is one method by which to encode variants. Finally, specialized clustering methods such as *k*-modes [16] are not widely available in common machine-learning libraries, limiting their availability for the broader bioinformatics community.

Applying ideas from Random Forests [17], we introduce an alternative method for variant analysis we call "Logistic Regression Ensembles," which we expect will be broadly useful. Our framework provides mechanisms for averaging the feature weights across the models in an ensemble and determining the number of models needed to achieve stable rankings, which achieves stable, consistent, and reproducible results. Because this approach is compatible with standard Logistic Regression implementations, this framework can be used with the user's programming language and libraries of choice and automatically inherits the advantages of new versions as they are released. We provide a reference implementation using scikit-learn in Asaph, a variant analysis toolkit.

As a specific example of the real-world utility of our method, we apply LR ensembles to single-nucleotide polymorphisms (SNPs) associated with the recent speciation of the malaria vectors *Anopheles gambiae* and *Anopheles coluzzii* from [18]. Tools and methods for identifying sequence-based differences are valuable for molecularly characterizing such closely-related species and ultimately to help understand speciation in these model systems [19]. PCA analysis of samples from the two species has shown strong evidence for strong similarity within species and clear differences between species [2], but localizing key variants is ongoing work [18].

First, we compare rankings generated by pairs of single LR models trained using the common Stochastic Gradient Descent (SGD) method to show that the rankings of SNPs from the same dataset vary significantly. Then, we show that ensembles of 250 models can achieve agreement as high as 99.0%. Finally, we compare LR ensembles and the more common metric F_{ST} using Jaccard similarity and the coefficient of determination (r^2) between scores computed by the two methods.

II. Methods

A. Dataset

Analyses were performed using biallelic SNPs at ≈4.6 million positions from 149 *Anopheles gambiae* (BFM) and

[1]Standard distance metrics such as the Euclidean distance do not account for cases where labels are permuted but have perfect association.

Anopheles coluzzii (BFS) mosquitoes. Details on the sequencing and variant calling (including filtering) are given in [18].

B. Logistic Regression Ensembles

Assume we have N samples with V bialleic positions. Each position has a reference allele and an alternative allele, and at each position, each sample has one of three genotypes (homozygous reference, homozygous alternate, or heterogyzous).

Here, we consider two ways of encoding the variants as a feature matrix $\mathbf{X}$ with dimensions $N \times P$. With the *genotype categories* feature-encoding scheme, we represent each genotype for each position as three categorical variables, and there are $P = 3V$ features. If sample i has the homozygous reference genotype at position k, then we set $\mathbf{X}_{i,3k+1} = 1$. If sample i has the homozygous alternate genotype at position k, then we set $\mathbf{X}_{i,3k+2} = 1$. If sample i has the heterozygous genotype at position k, then we set $\mathbf{X}_{i,3k+3} = 1$. If the genotype of sample i is unknown at position k, then we do nothing. For each variant v, we define a multiset $S_v = \{3k+1, 3k+2, 3k+3\}$ of corresponding feature indices.

With the *allele counts* feature-encoding scheme, we record the number of times each allele occurs for sample i at position k, and there are $P = 2V$ features. If sample i has the homozygous reference genotype at position k, then we set $\mathbf{X}_{i,2k+1} = 2$. If sample i has the homozygous alternate genotype at position k, then we set $\mathbf{X}_{i,2k+2} = 2$. If sample i has the heterozygous genotype at position k, then we set $\mathbf{X}_{i,2k+1} = 1$ and $\mathbf{X}_{i,2k+2} = 1$. If the genotype of sample i is unknown at position k, then we set $\mathbf{X}_{i,2k+1} = 0$ and $\mathbf{X}_{i,2k+2} = 0$. For each variant l, we define a multiset $S_l = \{2k+1, 2k+2\}$ of corresponding feature indices.

From the samples' population labels, we define a N-length vector $\mathbf{y}$ of class labels. We then form an ensemble of M Logistic Regression models with the form:

$$P(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + \exp(-\beta_m \cdot \mathbf{x} + \beta_{0,m})} \tag{1}$$

where y_i is the class label and $\mathbf{x}_i$ is the feature vector for a single sample i and β_m is the P-length weight vector and $\beta_{0,m}$ is the intercept for model m.

We train each model separately using Stochastic Gradient Descent and an L_2 penalty. (For the experiments in this paper, we performed 20 epochs of training for each model.) When bagging is employed, N samples are sampled with replacement from the original set to obtain a new feature matrix $\mathbf{X}^m$ and vector $\mathbf{y}^m$ of class labels for each model m.

After training the models, we then employ the weight vectors to rank the SNPs. The weight vector β_m of each model m is transformed by taking the absolute values of the elements and normalizing the vector to get $\hat{\beta}_m$ (Equation 2). From the normalized weight vectors, we then calculate an weight w_k for each feature k by averaging the k^{th} elements of each normalized vector $\hat{\beta_m}$ for each model m (Equation 3). From the normalized feature weights, we then calculate the weight v_l for each SNP l by taking the average of the feature weights w_k for $k \in S_l$ (Equation 4). We then sort the variants in descending order by their weights v_l for $l = 1, 2, \ldots, P$.

$$\hat{\beta}_m = \frac{|\beta_m|}{\|\beta_m\|} \tag{2}$$

$$w_k = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}_{m,k} \tag{3}$$

$$v_l = \frac{1}{\|S_l\|} \sum_{k \in S_l} w_k \tag{4}$$

A single Logistic Regression model is a just special case where $M = 1$.

C. Ranking SNPs with F_{ST}

To rank the SNPs, we first calculated the the F_{ST} score for each position using VCFTools [20]. We filtered out all positions with invalid (nan) scores. Then, we sorted the SNPs in descending order by their F_{ST} scores.

D. Metrics for Comparing Rankings and Scores

Given two pairs of SNP rankings, we selected the top k SNPs from each input ranking to get two sets, A and B, of SNPs. We then calculated the Jaccard Similarity as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{5}$$

where $0 \leq J(A, B) \leq 1$.

To directly evaluate the agreement of the scoring of SNPs, we computed the coefficient of determination (r^2). Given two pairs of scores for a set of SNPs, we generated a pair of column vectors. We then used SciPy's 'linregress' function [21] to compute the coefficient of determination (r^2).

E. Asaph: an Open-Source Toolkit for Variant Analysis

Our method (see Figure 1) was evaluated using Asaph, an open-source toolkit for variant analysis. Asaph was implemented in Python using Numpy / Scipy [22], Matlotlib [23], and Scikit Learn [24] and is available at https://github.com/rnowling/asaph under the Apache Public License v2.

III. Results

A. Disagreement in Top-Ranked SNPs from Two LR Models

Genomes often have on the order of millions of variants. With such large data sizes, approximate, stochastic, optimization methods such as Stochastic Gradient Descent (SGD) are often used to train the Logistic Regression (LR) models. LR models trained with SGD on the same data, however, can produce vastly different weight distributions, resulting in significant disagreement in the rankings. This is especially true when the intrinsic dimensionality of the problem is much lower than the number of features.

We demonstrate this inconsistency. For each of the two feature-encoding schemes (genotype categories and allele counts), we trained two LR models. We then ranked the SNPs using each model's feature weights as described in

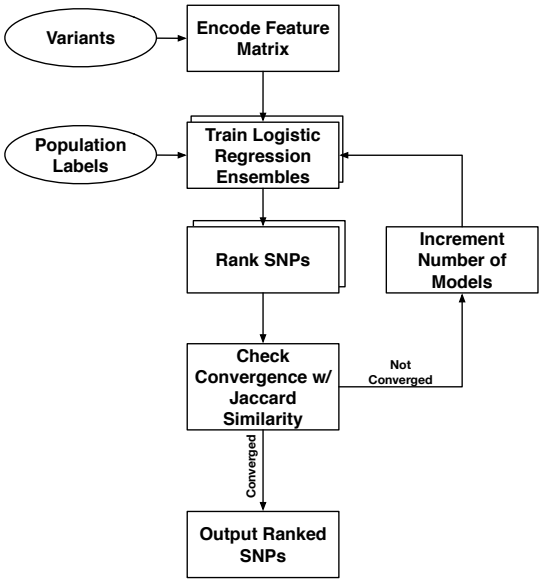


Fig. 1: Logistic Regression Ensembles workflow

Section II-B. We selected the top 0.01% (466) of the SNPs ranked by each model and used Jaccard Similarity to compare the rankings of the four models. As measured by Jaccard Similarity, the SNPs ranked by the two allele count models agreed on only 80.7% of SNPs, while two genotype category models agreed on as few as 38.8% of SNPs. (See Table I.) Such low-levels of agreement suggest that rankings generated from a single LR model are inconsistent.

TABLE I: Comparison of ranking of SNPs by two pairs (one pair for each feature-encoding scheme) of LR models

Threshold (SNPs)	Jaccard Similarity	
	Allele Counts	Genotype Categories
0.01% (466)	80.7%	38.8%
0.1% (4,662)	83.8%	63.0%
1% (46,620)	79.2%	62.9%
10% (466,202)	76.6%	63.8%

B. Achieving Agreement with an Ensemble of LR Models

Leo Breiman, the inventor of Decision Trees, identified similar instabilities when he observed that small changes in the training sets resulted in large changes in resulting Decision Tree and Linear Regression models [17]. In part to overcome those instabilities, Breiman suggested using an ensemble, leading to Random Forests [25]. We propose a similar approach. In particular, we propose training an ensemble of LR models and using an average of their weights to rank variants. In doing so, we reduce variance and should enable consistent, reproducible rankings across models.

Following the approach described in Section II-B, we trained two pairs of ensembles for each of the two feature-encoding schemes (genotype categories and allele counts) for a range of ensemble sizes. We selected the top 0.01% (466) of the SNPs ranked by each ensemble and used Jaccard Similarity to compare the rankings between each pair. For both feature-encoding schemes, agreement plateaued around 250 models. (See Figure 2.) As measured by Jaccard Similarity, the SNPs ranked by the allele count ensembles agreed on 99.0% SNPs versus the 80.7% agreement achieved for single LR models. Similarly, the genotype category ensembles agreed on 95.0% of SNPs versus 38.8% agreement achieved for single LR models. The ensemble approach is thus able to compensate for the variance by adequately sampling and then averaging the various possible weight distributions.

C. Accounting for Uncertainty in Population Frequencies with Bagging

Once again employing ideas from Random Forests, we employ bagging, or bootstrap aggregation, to sample over different possible population genotype frequencies [17], [26]. Bagging alone introduces additional variance in the models, which could cause inconsistency in the rankings, but is accounted for by an ensemble approach. Using the same assessment described in Section III-B and with 250 models, the allele counts-encoded (left) and genotype categories-encoded (right) ensembles achieved overlaps of 97.0% and 95.0% for the top 0.01% (466) ranked SNPs, respectively (see Figure 2).

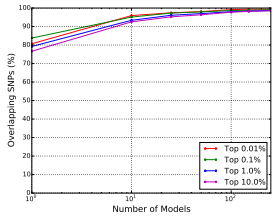
D. Comparison of SNPs as Ranked by F_{ST} and LR Ensembles

Fixation index (F_{ST}) is an univariate method commonly used to rank genetic variations between populations. Since SNPs identified via F_{ST} have received the most attention and analysis in the *Anopheles* research community, we compared rankings obtained via Logistic Regression (LR) Ensembles those obtained with F_{ST}. SNPs from the data set described in Section II-A ranked with LR Ensembles (with and without bagging) of 250 models using the categories and counts feature-encoding schemes as described in Section II-B and F_{ST} as described in Section II-C.

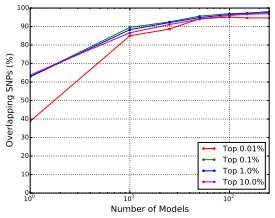
Agreement was measured in two ways: correlations between scores and, as above, Jaccard similarity. Scatter plots of scores for each SNP from the four LR Ensembles versus their F_{ST} scores are shown in Figure 3. Linear Regression models were trained for each ensemble vs F_{ST}, from which coefficients of determination (r^2) were computed (see Table II). Reasonably high r^2 values were achieved using the allele counts feature-encoding scheme, regardless of whether bagging was used ($r^2 = 0.812$) or not ($r^2 = 0.889$). However, bagging led to a significant improvement in r^2 (from 0.643 to 0.850) when using the genotype categories feature-encoding scheme.

As the top 0.01% and 0.1% of SNPs are likely to be the most interesting variants, we used Jaccard Similarity to compare agreement of the top-ranked SNPs of the LR ensembles with F_{ST}. The comparisons using Jaccard Similarities of the top-ranked SNPs tell a more nuanced story (see Table III) than the comparison of scores for all SNPs. For both the allele counts and genotype categories feature-encoded LR ensembles, bagging improved agreement of the top-ranked SNPs vs F_{ST} over not using bagging. The allele counts feature-encoded LR ensemble achieved agreement as high as

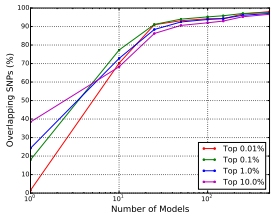
Fig. 2: Jaccard Similarities of Top-Ranked SNPs from Pairs of LR Ensemble Models



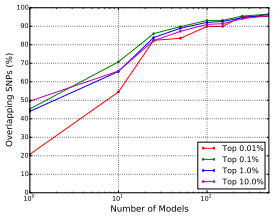
(a) Allele Counts



(b) Genotype Categories



(c) Allele Counts (Bagging)



(d) Genotype Categories (Bagging)

95.1% and 91.9% with F_{ST} for the top 0.01% (466) and 0.1% (4,662) SNPs, while the genotype categories feature-encoded LR ensemble had lower levels of agreement of 86.8% and 79.2%, respectively. When expanded out to the top 1% (46,620) and 10% (466,202) SNPs, the genotype categories feature-encoded LR ensemble achieved agreements as high as 98.7% and 99.7%, while the allele counts feature-encoded LR ensemble showed reduced agreement at 79.6% and 77.4%.

TABLE II: Coefficients of Determination (r^2) of Scores from LR Ensembles vs F_{ST}

	Allele Counts	**Genotype Categories**
No Bagging	0.812	0.643
Bagging	0.889	0.850

TABLE III: Comparison of rankings of SNPs by LR Ensembles and F_{ST}

Threshold (SNPs)	**Allele Counts**		**Genotype Categories**	
	Bagging	**No Bagging**	**Bagging**	**No Bagging**
0.01% (466)	95.1%	94.2%	86.8%	69.7%
0.1% (4,662)	91.9%	91.7%	79.2%	65.6%
1% (46,620)	79.6%	79.6%	98.8%	98.7%
10% (466,202)	77.4%	76.7%	99.8%	99.7%

IV. Discussion and Conclusion

We show that feature weights from Logistic Regression models trained with Stochastic Gradient Descent (SGD) can vary significantly between instances trained on the same SNP data. In Section III-A, we compared the top 0.01% (466) ranked SNPs from two models and found that they can agree on as few as 38.8% of SNPs and, in the best case, may only agree on as many as 80.7% of SNPs.

To remedy this problem, we proposed Logistic Regression Ensembles. With our approach, the weights from the models in the ensemble are averaged and then used to rank the features. In Section III-B we demonstrated that, by using a pair of ensembles with 250 models each, agreement, as measured by Jaccard Similarity, of up to 99.0% of the top 0.01% (466) of SNPs ranked can be achieved. Even with the additional variance introduced by using bagging to account for sampling bias, we were able to demonstrate agreement up to 97.0% (see Section III-C).

In Section III-D, we applied LR ensembles to single-nucleotide polymorphisms (SNPs) associated with the recent speciation of the malaria vectors *Anopheles gambiae* and *Anopheles coluzzii* from [18] and validated LR Ensembles against the popular univariate method F_{ST}. With the allele counts encoding scheme, LR ensembles achieved a coefficient of determination (r^2) of up to 0.889, indicating significant agreement. Using Jaccard Similarity, LR Ensembles achieved an agreement of up to 95.1%.

LR has been used successfully in several GWAS studies. If scientists desire to use LR instead of F_{ST} or other methods, our LR ensembles method enables them to do so while resolving issues with consistency and stability. These LR Ensembles are implemented and available in Asaph, an open-source toolkit for exploring machine-learning approaches to ranking SNPs in incipient species of insects.

Acknowledgment

The authors would like to thank Nora Besansky, Michael Fontaine, Becca Love, and Aaron Steele for thoughtful discussions that provided the motivation for this effort.

References

[1] D. J. Balding, "A tutorial on statistical methods for population association studies," *Nat Rev Genet*, vol. 7, no. 10, pp. 781–791, Oct 2006.

[2] M. C. Fontaine, J. B. Pease *et al.*, "Extensive introgression in a malaria vector species complex revealed by phylogenomics," *Science*, vol. 347, no. 6217, 2015.

[3] D. Neafsey, M. Lawniczak, and D. Park, "SNP genotyping defines complex gene-flow boundaries among African malaria vector mosquitoes," *Science*, vol. 2984, 2010.

[4] Q. He and D.-Y. Lin, "A variable selection method for genome-wide association studies," *Bioinformatics*, vol. 27, no. 1, p. 1, 2011.

[5] "Finding predictive gene groups from microarray data," *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 106 – 131, 2004.

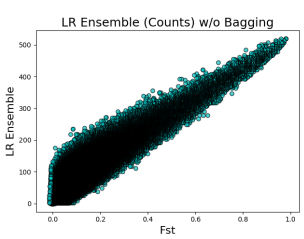
[6] C. J. Hoggart, J. C. Whittaker, M. De Iorio, and D. J. Balding, "Simultaneous analysis of all snps in genome-wide and re-sequencing association studies," *PLOS Genetics*, vol. 4, 07 2008.

[7] C. Kooperberg, M. LeBlanc, and V. Obenchain, "Risk prediction using genome-wide association studies," *Genetic Epidemiology*, vol. 34, no. 7.

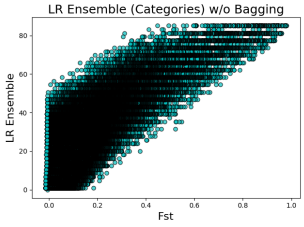
[8] J. Zhu and T. Hastie, "Classification of gene microarrays by penalized logistic regression," *Biostatistics*, vol. 5, no. 3, p. 427, 2004.

[9] J. O. Ogutu, T. Schulz-Streeck, and H.-P. Piepho, "Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions," *BMC Proceedings*, vol. 6, no. 2, p. S10, 2012.

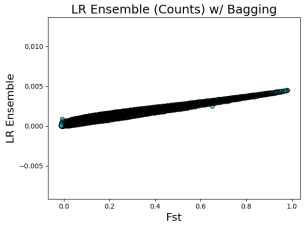
Fig. 3: Scatter plots of F_{ST} vs scores from LR Ensembles with counts and categories feature-encoding schemes, with and without bagging



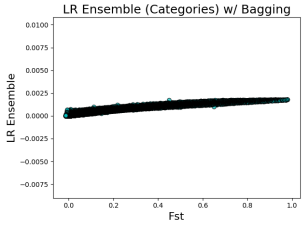
(a) Allele Counts (No Bagging)



(b) Genotype Categories (No Bagging)



(c) Allele Counts (Bagging)



(d) Genotype Categories (Bagging)

[10] C. Riedelsheimer, F. Technow, and A. E. Melchinger, "Comparison of whole-genome prediction models for traits with contrasting genetic architecture in a diversity panel of maize inbred lines," *BMC Genomics*, vol. 13, no. 1, p. 452, 2012.

[11] P. Waldmann, G. Mszros, B. Gredler, C. Frst, and J. Slkner, "Evaluation of the lasso and the elastic net in genome-wide association studies," *Frontiers in Genetics*, vol. 4, p. 270, 2013.

[12] G. Abraham, A. Kowalczyk, J. Zobel, and M. Inouye, "Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease," *Genetic Epidemiology*, vol. 37, no. 2.

[13] L. Shen and E. C. Tan, "Dimension reduction-based penalized logistic regression for cancer classification using microarray data," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 2, no. 2, pp. 166–175, Apr. 2005.

[14] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2.

[15] L. Toloi and T. Lengauer, "Classification with correlated features: unreliability of feature ranking and solutions," *Bioinformatics*, vol. 27, no. 14, p. 1986, 2011.

[16] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Mining and Knowledge Discovery*, vol. 2, no. 3, pp. 283–304, 1998.

[17] L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[18] A. Miles, N. J. Harding *et al.*, "Natural diversity of the malaria vector anopheles gambiae," 2016.

[19] A. P. Michel, W. M. Guelbeogo *et al.*, "Molecular differentiation between chromosomally defined incipient species of *Anopheles funestus*." *Insect Molecular Biology*, vol. 14, no. 4, pp. 375–87, 2005.

[20] P. Danecek, A. Auton *et al.*, "The variant call format and vcftools," *Bioinformatics*, vol. 27, no. 15, p. 2156, 2011.

[21] E. Jones, T. Oliphant, P. Peterson *et al.*, "SciPy: Open source scientific tools for Python," 2001–, [Online; accessed 2017-09-30]. [Online]. Available: http://www.scipy.org/

[22] S. v. d. Walt, S. C. Colbert, and G. Varoquaux, "The numpy array: A structure for efficient numerical computation," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, 2011.

[23] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing In Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[24] F. Pedregosa, G. Varoquaux *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 5, pp. 1–35, 1999.

[26] D. Gianola, K. A. Weigel, N. Krmer, A. Stella, and C.-C. Schn, "Enhancing genome-enabled prediction by bagging genomic blup," *PLOS ONE*, vol. 9, 04 2014.