We first created features for both the training and the test variants. The set of features were grouped into three:

1) Deepbind scores: We used the set of 972 deepbind scores (http://tools.genes.toronto.edu/deepbind/) and scanned the sequences in the similar manner as the PWM scan. The scores are obtained by using deep learning technique, and the scores provide information on sequence specificities of DNA binding proteins.

2) UniProbe k-mer scores: We scored each of the 150 nucleotide sequences with the k-mer E-scores from the UniProbe (http://the_brain.bwh.harvard.edu/uniprobe/). Specifically, we used 320 kmer sets corresponding to *** human and *** mouse transcription factors. For a given TF and oligonucleotide pair, we summed the E-scores of the all the k-mers , counted in a 1nucleotide sliding window fashion and by taking taking into account both strands, within the 150 nucleotide oligonucleotide. As a result we obtained a UniProbe k-mer score set of size 320 for both alleles of the GVs.

3) ENCODE chromatin scores: We utilized all the ENCODE histone modification, TF binding ChIP-seq, DNase-seq and RRBS-seq data available in GM128787 from the ENCODE portal (https://www.encodeproject.org/). Specifically, we overlapped the coordinates of the 150 nucleotide oligos with the ENCODE-derived peak sets. For the 10 histone datasets we used the original ENCODE-reported peak boundaries. For DNase-seq and TF ChIP-seq, we considered a 151 bps window centered around the ENCODE reported summit of the peak. As a cumulative TF ChIP-seq score, we counted the number of times each 150 nucleotide sequence overlaps with a refined TF ChIP-seq peak described as above. This processing resulted in a total of 13 features that we refer to as ENCODE chromatin scores.

We used random forest to build all the predictive models. Specifically, for continuous responses (log2FC), we used random forest regression and tuned the mtry parameter with 5-fold cross validation by minimizing the mean squared error. For binary responses (regulatory Hit, emVarHit), we tuned the mtry parameter of random forest by maximizing the area under the ROC curve with cross-validation. For all the models, the numbers of trees were set as 750.

Par1: To predict the log2FC responses, we considered random forest models over the features described for their respective allele (for the ENCODE chromatin scores case, we used the same features for both models). To predict the Regulatory Hits, with the first two data types we considered the maximum score between the two alleles. We considered random forest models with the max. score feature, the ENCODE chromatin scores and the predicted log2FC gene expression for both the reference and alternative alleles.

Part2:

No	LogSkew.Comb		emVar_Hit	
	Features	Optimal mtry	Features	Optimal mtry
1	PWM scan + Deepbind	150	PWM scan + ENCODE chromatin scores	150

2	Predicted expression change from Part1 + ENCODE chromatin scores	12	Predicted expression change from Part1 + Deepbind + UniProbe k-mer scores + ENCODE chromatin scores	700
3	Predicted expression change from Part1 + Deepbind	900	Predicted expression change from Part1 + PWM scan	600