Genetic Correlates of Morton's Toe

Emma Weeding

ABSTRACT

Introduction: Morton's toe—also known as Greek toe or *metatarsus atavicus*—is an anatomical variant where an individual's first toe (i.e. big toe) is shortened relative to their second toe. This trait is present in roughly 10-30% of humans, though its prevalence varies widely depending on the population studied. Morton's toe is believed to be genetic and has been observed in families across multiple generations, yet specific genetic correlates for Morton's toe have never been described in the literature. We developed two binary classification models to predict which individuals are likely to have Morton's toe based on SNPs, and then examined these models to identify SNPs which associate with Morton's toe based on predictive ability.

Methods: The genetic data of OpenSNP users who had declared a phenotype with respect to "Morton's Toe" and/or "Index Toe Longer than Big Toe" (n = 181 total; n = 88 with Morton's toe) were used in this work. As specific genetic correlates of Morton's toe have never been previously described, the genetic data was first filtered to include only SNPs within genes from the gene ontology set HP: Abnormality of Toe from MSigDB. Fisher's exact test was applied to each SNP (using an allelic model), and the 250 SNPs with the lowest p values were used as regressors to train logistic regression (LR) and random forest (RF) binary classification models to predict the presence of Morton's toe. All models were trained and assessed via 10-fold cross-validation, and this full training process was repeated five times for each model. Averaged results were used to generate final summary performance statistics and assess SNP importance.

Results: Overall, both models were successful at predicting the presence of Morton's toe based on SNPs. Compared to the LR model, the RF model demonstrated marginally higher overall performance (average AUC 0.84 ± 0.01 vs 0.81 ± 0.01, p = 0.002) with particularly improved specificity (0.81 ± 0.01 vs 0.72 ± 0.02, p = 0.0001) at the cost of sensitivity (0.66 ± 0.03 vs 0.72 ± 0.02, p = 0.006). SNPs most influential to the LR model were quite different from those influential to the RF model–generally, LR identified uncommon SNPs with a stronger individual effect on Morton's toe risk, whereas RF identified more common SNPs with a more complex relationship to Morton's toe risk.

Conclusion: While this work was significantly limited by the small sample size and lack of an independent test set, the performance of these SNP-based exploratory models certainly supports the view of Morton's toe as a genetic trait. However, its specific genetic underpinnings appear to be complex and heterogenous. Further studies would be needed to determine if these models generalize to individuals outside of OpenSNP.

INTRODUCTION

Morton's toe—also known as Greek toe or *metatarsus atavicus*—is a generally benign anatomical variant where an individual's first metatarsal is shortened relative to their second. Occurring in roughly 10-30% of the population globally, this trait is relatively common in humans though its reported prevalence varies widely depending on the population studied, ranging from 3% in Sweden to 90% in northern Japan [1-3]. Morton's toe is believed to be genetic and can be observed in families across several generations, though its mode of inheritance has been a subject of debate [4, 5]. The few modern studies on this subject have suggested a more complex pattern of inheritance rather than Mendelian, as was historically proposed [1-3, 5, 6]. Moreover, specific genetic correlates with Morton's toe have never been described in the literature.

In this work, we developed small exploratory models of Morton's toe presence versus absence based on single-nucleotide polymorphisms (SNPs). Specifically, logistic regression- and random forest-based binary classification models were developed using publicly available genetic data in individuals without or without Morton's toe. The parameters of these models were then examined to determine which SNPs were most important to successful prediction of the Morton's toe phenotype. By extension, these SNPs of interest might inform understanding of which genes or biological pathways drive the development of Morton's toe.

METHODS

Data Source and Study Population

All raw genetic data was obtained via OpenSNP [7]. Specifically, 23andMe files were downloaded for all OpenSNP users reporting a phenotype with respect to "Morton's Toe" and/or "Index Toe Longer than Big Toe" (n = 281). Users with indeterminate or discordant responses to

these phenotype questions were excluded (n = 24), as were any genetic data files not based on human reference genome GRCh37 (n = 66). Several users had uploaded multiple copies of their genetic data (n = 10), and these duplicate files were also removed from the dataset. The final study cohort consisted of 181 individuals either with (n = 88, 48.6% of total) or without (n = 93, 51.4% of total) Morton's toe.

SNP Selection

As above, no specific genetic correlates with Morton's toe have previously been described in the literature, and applying a GWAS approach to this cohort was not possible given the small sample size and suspected complexity of the phenotype. To narrow the analysis to a pool of potentially biologically relevant SNPs, the OpenSNP data was filtered to only include SNPs within genes from a gene ontology term for human toe abnormalities, HP:0001780 [8-10]. This reduced the total number of SNPs in the OpenSNP files from 1,258,255 to 27,037. Via OpenCRAVAT [11], the Fisher's exact test was then used to calculate an allelic p-value for each SNP with respect to the Morton's toe phenotype. Finally, the 250 SNPs with the lowest p-values (not necessarily meeting any particular threshold) were selected for use as regressors in all models. This number of SNPs was determined empirically based on model performance during cross-validation–in all cases, using a smaller number of SNPs resulted in a lower area under the curve (AUC), whereas using an increased number of SNPs beyond 250 did not significantly improve AUC.

Data Preprocessing and Model Training

For each SNP, a value of 2, 1, or 0 was assigned depending on whether a given individual was homozygous for the SNP, heterozygous for the SNP, or had the reference genotype,

respectively. These values were standardized to a mean of 0 and standard deviation of 1 for each SNP to facilitate later interpretation of model coefficients.

Subsequently, binary classification models to predict the presence or absence of Morton's toe based on the 250 SNPs were developed using either logistic regression (with elastic net regularization) or random forest algorithms. In both cases, models were trained and assessed using 10-fold cross-validation. Cross-validation performance was also used to inform hyperparameter tuning. Finally, each full training cycle of 10 training-validation folds was performed five times for each modeling approach in order to generate averaged performance summary statistics as well as to assess consistent SNP importance in each model. Summary statistics for the different modeling approaches were compared using Student's t-test.

Performance Metrics and SNP Importance

For both models, AUC, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated. For logistic regression, model coefficients were extracted for the cross-validation fold with the highest validation AUC. SNPs with a higher coefficient magnitude were considered to be more important to the model's predictive ability, and vice versa. For random forest, mean decrease in impurity was extracted for each SNP in a similar fashion, and SNPs with a higher mean decrease in impurity were considered to be more important to the model's predictive ability.

RESULTS

SNP Selection

A list of the top 250 SNPs with the lowest allelic p-values is included in **Supplementary Table 1** (attached as .csv). As expected, no individual SNP met the threshold for statistical significance after adjustment for multiple comparisons via the Benjamini-Hochberg procedure

using a false discovery rate of 5%. Of the 250 SNPs, the overwhelming majority (n = 234) were intron variants. Only three were coding variants: rs4782300 (missense, *ZNF469*), rs6520618 (synonymous, *BCOR*), and rs2066844 (missense, *NOD2*).

Model Performance

Overall performance of the logistic regression and random forest models is shown in **Figure 1**. The logistic regression model demonstrated all-around balanced performance with an average AUC of 0.81 ± 0.01 (mean ± standard deviation), sensitivity of 0.72 ± 0.02, specificity of 0.72 ± 0.02, PPV of 0.70 ± 0.02, and NPV of 0.77 ± 0.01. Overall performance of the random forest model was surprisingly similar, though slightly better with respect to average AUC, which was 0.84 ± 0.01 ($p = 0.0015$ compared to logistic regression). The random forest average sensitivity, specificity, PPV, and NPV were 0.66 ± 0.03 ($p = 0.0059$), 0.81 ± 0.01 ($p = 0.0001$), 0.77 ± 0.02 ($p = 0.0006$), and 0.73 ± 0.02 ($p = 0.0039$) respectively.

SNP Importance

The top 10 most important SNPs (as defined in the Methods section) for each model are shown in **Table 1** and **Table 2**. All SNPs shown are intron variants. SNPs highly influential to the logistic regression model were of notably lower global allelic frequency (average 0.20, ranging from 0.01 to 0.36) compared to those important to the random forest model (average 0.35, ranging from 0.20 to 0.70). Interestingly, no top 10 SNPs were common to both models, though at least one SNP within *DMD* was in the top 10 SNPs for each (one SNP for logistic regression, three different SNPs for random forest). Two SNPs in *EVC* were highly important to the random forest model, but not the logistic regression model.

DISCUSSION

Both the logistic regression and random forest models were generally successful at predicting the presence of Morton's toe based on a set of 250 SNPs. Overall performance of the two modeling approaches was similar, with an average AUC in the low to mid-80s for both LR and RF. The behavior of the models was quite different in other respects, however. The RF model was significantly more specific and less sensitive than the LR model. Moreover, there was no overlap between the top 10 most influential SNPs (based on coefficient magnitude or decrease in impurity) for each model. The LR model appeared to identify individual SNPs which are relatively uncommon, but, when present, significantly increase the likelihood that an individual has Morton's toe. Conversely, the RF model found a surprising amount of importance in SNPs with a higher global allelic frequency (up to 0.70 in the top 10 SNPs). We suspect that the RF model is better able to identify higher-level SNP interactions–that is, cases in which a given SNP has a marginal effect on the Morton's toe phenotype when present in isolation, but a strong effect when combined with other SNPs. Generalized linear models such as logistic regression can struggle to identify these types of relationships. It is thus both biologically and computationally reasonable for the two modeling approaches to have identified different SNPs as being important in different ways.

Because of these varied results, it is difficult to immediately appreciate a unified biological pathway to explain why and how Morton's toe develops in certain individuals. By design (i.e. filtering by the gene ontology term for toe abnormalities), all SNPs assessed in this work have some direct or indirect relevance to toe development in humans. The notable genes *DMD* and *EVC* identified above are each associated with diseases involving musculoskeletal dystrophy or dysplasia–specifically, muscular dystrophy and Ellis-Van Creveld syndrome,

respectively–but their pathophysiologic influence on first and second toe length is not clear [12, 13]. Nonetheless, it seems clear that Morton's toe is a complex and likely highly heterogenous phenotype.

Major limitations of this work include the small sample size relative to the complexity of the phenotype, as well as the lack of a truly independent test cohort. As it was empirically determined that a set of 250 SNPs resulted in optimal model performance, without an independent test set, we cannot be certain that the high performance of the models is due to appropriate identification and weighting of biologically meaningful SNPs, or whether these models have simply been overfitted to this OpenSNP population. The lack of adequate ethnicity data in OpenSNP is another significant limitation for similar reasons. As the prevalence of Morton's toe varies significantly by population, it is likewise unclear if the performance of these two models is being driven by SNPs directly related to toe length or by confounding factors related to ethnicity. Stratifying the analysis by ethnicity could elucidate which SNPs or sets of SNPs are truly important to this phenotype in different populations, as well as overall. This manner of future work is clearly needed to determine whether truly generalizable genetic models for Morton's toe could be developed.

REFERENCES

1. Romanus, T., *Heredity of a long second toe.* Hereditas, 1949. **35**: p. 651-652.
2. Kaplan, A., *Genetics of Relative Toe Lengths.* Acta Genet Med Gemellol (Roma), 1964. **13**: p. 295-304.
3. Vounotrypidis, P. and P. Noutsou, *The Greek Foot: Is it a Myth or Reality? An Epidemiological Study in Greece and Connections to Past and Modern Global History.* Rheumatology, 2015. **54**: p. i182-i183.
4. Beers, C. and L. Clark, *Tumors and short-toe—a dihybrid pedigree: a family history showing the inheritance of hemangioma and metatarsus atavicus.* J Hered, 1942. **33**: p. 366-368.
5. Aigbogun, E.O., et al., *Morton's Toe: Prevalence and Inheritance Pattern among Nigerians.* Int J Appl Basic Med Res, 2019. **9**(2): p. 89-94.
6. Morton, D., *Inheritance of a long second toe.* J Hered, 1952. **43**: p. 49-50.
7. Greshake, B., et al., *openSNP--a crowdsourced web resource for personal genomics.* PLoS One, 2014. **9**(3): p. e89204.
8. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
9. Liberzon, A., et al., *The Molecular Signatures Database (MSigDB) hallmark gene set collection.* Cell Syst, 2015. **1**(6): p. 417-425.
10. Köhler, S., et al., *The Human Phenotype Ontology in 2021.* Nucleic Acids Res, 2021. **49**(D1): p. D1207-D1217.

11. Pagel, K.A., et al., *Integrated Informatics Analysis of Cancer-Related Variants.* JCO Clin Cancer Inform, 2020. **4**: p. 310-317.

12. Flanigan, K.M., et al., *Rapid direct sequence analysis of the dystrophin gene.* Am J Hum Genet, 2003. **72**(4): p. 931-9.

13. Ruiz-Perez, V.L., et al., *Mutations in a new gene in Ellis-van Creveld syndrome and Weyers acrodental dysostosis.* Nat Genet, 2000. **24**(3): p. 283-6.

Figure 1. Representative ROC curves for the logistic regression (**left**) and random forest (**right**) models. ROC curves for each validation fold are indicated by faint blue lines, whereas the darker blue lines and shaded gray areas represent the average ± standard deviation ROC curves across all validation folds.

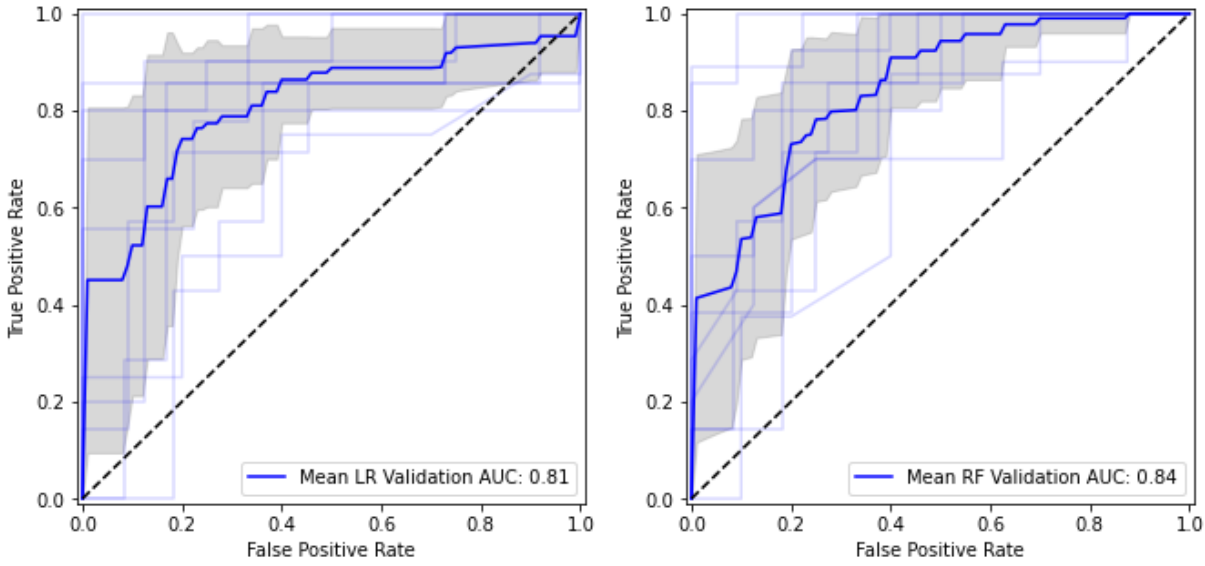

Table 1. Top 10 most influential SNPs for the logistic regression model based on model coefficient magnitude. Abbreviations: AF, allele frequency.

Ref. SNP ID	Location	Ref. Base	Alt. Base	Gene	Global AF
rs12852580	chrX:107830607	G	A	*MID2*	0.04
rs77685903	chr3:189754938	G	A	*TP63*	0.10
rs6631392	chrX:31781620	G	A	*DMD*	0.26
rs148991318	chr7:255297	C	T	*FAM20C*	0.16
rs5936560	chrX:70478804	A	T	*DLG3*	0.18
rs75108301	chr12:879458	G	T	*WNK1*	0.01
rs35263707	chr3:30642605	G	A	*TGFBR2*	0.36
rs9928399	chr16:55431047	T	C	*MMP2*	0.29
rs12935162	chr16:49530783	G	A	*ZNF423*	0.21
rs10448080	chr8:28747274	T	C	*EXTL3*	0.35

Table 2. Top 10 most influential SNPs for the random forest model based on mean decrease in impurity. Abbreviations: AF, allele frequency.

Ref. SNP ID	Location	Ref. Base	Alt. Base	Gene	Global AF
rs2291155	chr4:5733576	C	T	*EVC*	0.21
rs2468240	chr12:88050815	G	A	*CEP290*	0.50
rs4829222	chrX:31439635	A	C	*DMD*	0.44
rs2185385	chr1:210757467	A	G	*KCNH1*	0.23
rs13103693	chr4:5734090	C	T	*EVC*	0.20
rs6790272	chr3:179571641	A	C	*ACTL6A*	0.43
rs2094147	chrX:33097459	C	T	*DMD*	0.43
rs11966489	chr6:157092895	T	C	*ARID1B*	0.21
rs11127298	chr2:1966454	T	G	*MYT1L*	0.20
rs1321398	chrX:32872222	A	G	*DMD*	0.70