Validation through a comparison of physical examination and DNA test results: OLFML3 case study

Abstract

Objective: To investigate the usefulness of predicting the presence of a genetic mutation in a population based on data arising from physical examinations and its use as a validator of grading schemes.

Methods: We employ an inference algorithm adjusted for large scale application and apply it to the particular case of Olfactomedin Like 3 (OLFML3) gene in the Border Collie, where essentially independent results of DNA test and physical exam exist, to find out to what extent the two predictions agree and highlight systematic discrepancies.

Results: Large scale inference over a population of 243,752 Border Collie dogs is combined with statistical data mining techniques on available DNA test and physical exam results corresponding to the OLFML3 gene and physical gonioscopy. Aside from validating the inference technique, the results are two fold: (1) sources of possible errors are identified indicating potential revisions of the physical examination grading scheme, and (2) the inference is found to approach close values within a short distance (1 or 2 steps away from a tested

_____

animal) based on physical examinations alone, supporting the usefulness of predictions based solely on the results of a physical exam.

Conclusion: A systematic comparison of the inference results based on a physical examination and on DNA tests highlights any problems with the grading scheme used for the physical examinations. The proximity of physical examination and DNA based inference suggests that the method could be employed to select the most informative subjects to explore when a method of inheritance of a condition is suspected but only a physical examination is possible.

Introduction

Reduction in expense, increase in speed and the ability to work at large scale without needing to satisfy stringent approvals is fuelling a desire to use *in silico* methods in the fields of genetic research [Lang et al. (2018)]. Indeed, approaches range from pedigree analysis for tracking of genetic inheritance to automatic predictions of deleterious missense variants. We hypothesise that given a condition with a discernible physical manifestation, large scale inference combined with statistical data mining techniques can be used to validate the results of a DNA test, physical examination and the inference algorithm itself.

Previously, a pipeline of computational approaches has been used to identify highly deleterious missense single nucleotide polymorphisms (SNPs) computationally, allowing for systematic exploration of numerous genetic loci for particular traits (e.g. processing DNA extracted from a population of 69 os-

triches to explore physiological and reproductive traits [Al-Shuhaib et al. (2018)]).
Such a technique can be used to influence a breeding program to potentially
increase desired outcomes such as yield (e.g. by focusing away from the highly
deleterious non synonymous SNPs (nsSNPs) found by investigating the effects
of amino acid substitutions of the *prl* gene on 223 members of the population
of Japanese quail, thus increasing egg production [Al-Shuhaib et al. (2019)]).
Existing datasets can also be exploited for comprehensive exploration of SNPs:
the database of single nucleotide polymorphisms [Sherry et al. (2001)] facili-
tated an exploration of the most deleterious nsSNPs of the bovine LEP gene
[Al-Shuhaib (2019)].

To avoid the need for animal / human testing, *in silico* methods have also
been used in toxicology, such as for the assessment of cosmetic substances.
These approaches include quantitative structure–activity relationship models
which predict the activities of new chemicals and read-across which allows par-
allels to extend existing results to new substances [Gellatly & Sewell (2019)].

A valuable contribution of computational approaches allows haplotypes to
be inferred from observed genotype data combined with pedigree structure. A
number of computational approaches are summarized in [Gao et al. (2009)].
While these approaches are known to have limitations – in particular, they
are not guaranteed to produce the true configuration [Lindholm et al. (2004)]
– they have been applied successfully on an increasing scale from relatively
small family investigations [Burdick et al. (2006)] to the large scale analyses
of plants [Velazco et al. (2019)].

Our investigation focuses on conditions with a suspected simple single-gene
inheritance which are discernible by a physical examination and have a (can-
didate) associated DNA test. In the case of complete penetrance, the results
of a physical exam and the DNA test should align, and, assuming no other
common causes of the condition, provide a validation of both types of test.

However, the number of sample points with both DNA and physical results is often very small, and the physical exam's inability to distinguish carrier (with alleles gG or Gg) from normal (alleles gg) status (for recessive inheritance) or carrier (gG or Gg) from affected (GG) status (for dominant inheritance), along with any errors of either evaluation type, further complicates analysis.

The discovery and subsequent verification of a specific genetic mutation can prove to be an expensive process, with cost of the subsequent deployment of a resulting DNA test in a population limited by the means of the individuals or health organizations employing it. Lack of testing and misunderstandings based on miscomprehension of the DNA test's scope, for example when other causes of the condition exist, could be detrimental to subsequent generations. A basic understanding of simple genetics and a suspected mode of inheritance can be used to make offspring $\rightarrow$ parent inferences – for example, to produce affected progeny, both parents must be at least carriers in the case of recessive inheritance – however large scale programming methods need to be employed to make longer distance inferences over a large population. Such algorithms can also account for degrees of error, incomplete penetrance or variable expression.

To explore our hypothesis, we require a dataset consisting of the following:

1. A large population with known pedigree information.
2. A single gene mutation present in the population with a suspected (or confirmed) simple mode of inheritance.
3. Results of a physical examination for some animals within the population.
4. Results of a DNA test for some animals within the population.

To this end, the recently discovered olfactomedin like 3 (OLFML3) mutation indicating a predisposition to the development of narrow angled glaucoma in Border Collie dogs [Pugh et al. (2019)] was chosen. Initial whole genome sequencing of six affected animals indicated the involvement of chromosome

17, and further studies of the region with the lowest Bonferroni corrected significance value SNP uncovered a missense variant ($c.590G > A$) in the OLFML3 gene with the predicted change of arginine at position 197 to glutamine (p.R197Q). The inheritance was suspected to be autosomal recessive and prior to the publication of this manuscript, this was further supported by a larger study [Oliver et al. (2020)] of 209 Border Collie dogs. While the smaller study of 6 dogs manually utilised pedigree information, the larger did not have access to pedigree information.

The physical exam consists of a gonioscopy performed by an ophthalmic specialist assessing any drainage angle changes. The condition has been studied in the Border Collie since 2008, when a number of instances of primary glaucoma presented in the breed, with the main screening tool employed in the UK being a gonioscopic examination. The grading scheme used in the gonioscopic evaluation has undergone a number of changes during this time.

The main objective of our work is to demonstrate that a large scale inference algorithm combined with statistical data mining techniques can be used to validate the results of a DNA test, physical examination and the inference algorithm itself. Moreover, the inference algorithm can also be used to explore the utility of a physical examination grading scheme or to highlight any obvious errors in either the DNA test or physical examination results. As a by product, the inference algorithm could be employed to suggest subjects to consider for investigation in the quest for a disease's DNA test.

Materials and Methods

A convergence adjusted form of the blocking Gibbs algorithm [Jensen et al. (1995)] is used to make inferences over a population of 243,752 dogs based on available DNA and physical exam results yielding two separate probability distributions

(over affected-carrier-normal status) for each animal in the population by exploiting available parentage links. For the test case, an animal DNA OLFML3 tested carrier with a clear (i.e. normal or carrier) physical examination result will yield a $(0.0, 1.0, 0.0)$ DNA probability distribution, while (assuming no other information) the exam result probability distribution will contain non zero values in both the normal and carrier fields (such that these sum to 1.0). An animal without test result(s) will have its probability distributions inferred.[1]

The inferred probability distributions can be compared by statistical methods either directly, or they can be filtered by dates of birth or (more usefully) their distance to the nearest tested animal – it is expected that as the distance away from a tested animal increases the probability distributions will approach a background probability and thus will not be particularly informative since the data contains animals with dates of birth ranging from 1894 to 2019 (with over 92% of the dogs having a date of birth listed).

Gonioscopic examination

The classification of the observed result has undergone numerous changes in the UK, starting with a 0-5 scale (Table 1) of which G0-G2 are considered clear and G3-G5 affected, reverting to a simple clear / affected distinction, through to a new scheme supported by the British Veterinary Association (BVA) assigning one of four grades (Table 2) with only 0 being clear. UK results outside the dates listed and results from other countries are only reported as "clear" or "affected", with the distinction between the two left to the individual ophthalmologists.

[1] Note that the probability distribution is not fixed for tested animals and inference can adjust the values based on other inputs – for example, if two parents are DNA tested affected while an offspring is DNA tested normal, all three animals will have their probabilities adjusted.

Table 1 UK gonioscopy grading scheme: 2007-2008

Grade	Gonioscopic findings
G0	Wide. Open no apparent thickening of ligament.
G1	Normal. Some thick fibres.
G2	Normal. Some thick fibres, some sheets.
G3	Narrowed. Abnormal pectinate ligament or pectinate ligament not seen.
G4	Narrow. Pectinate ligament not seen.
G5	Very narrow. Pectinate ligament not seen, angle looks closed.

Table 2 UK gonioscopy grading scheme: 2017-

Grade	Gonioscopic findings
0	Normal iridocorneal angle (ICA) with no/minimal (0%-1%) pectinate ligament abnormality (PLA).
1	1-25% of ICA affected by PLA.
2	26-75% of ICA affected by PLA
3	>76% of ICA affected, and / or severe narrowing of ICA

Data

The dataset contains information on 243,752 Border Collie dogs (June 2019) collected via the online interface of the freely available Anadune Border Collie Database [Preiss, 2019-06-19]. Table 3 presents information contained in the dataset at the time of evaluation: the total number of dogs, numbers of tested dogs, the number of dogs with a minimum of complete 3 generation pedigrees. The number of dogs at a given number of steps away from a tested animal is also included, where the number of steps corresponds to the (shortest path) Dijkstra distance [Dijkstra (1959)]: i.e. animals at step 0 are directly tested themselves, step 1 have a directly tested parent or offspring, while siblings are treated as 2 steps away as are grandparents and grandchildren etc. The counts for gonioscopy affected include dogs that have developed primary glaucoma but were not examined gonioscopicaly prior to this (this accounts for 6 animals).

Table 3 Dataset statistics: breaking down numbers of directly tested dogs and the numbers of dogs at a given number of steps from a tested animal

	OLFML3			gonioscopy	
	normal	carrier	affected	clear	affected
Total dogs	———————		243,752	———————	
Dogs with 3+ generations	———————		197,840	———————	
Tested dogs	873	335	35	3,753	158
	70.23%	26.95%	2.82%	95.96%	4.04%
	———	1,243	———	— 3,911	—
Tested dog 1 step away	———	10,249	———	— 38,246	—
Tested dog 2 steps away	———	23,869	———	— 48,222	—
Tested dog 3 steps away	———	39,736	———	— 49,659	—

Inference algorithm

The size of the population examined makes full exact computation intractable. Since the pedigree structure of the data can be represented as a Bayesian network, we base our inference algorithm on blocking Gibbs [Jensen et al. (1995)], a variant of a Markov chain Monte Carlo method which integrates exact local calculations with Gibbs sampling. The performance gain of blocking Gibbs over plain Gibbs is due to simultaneous sampling of sets (blocks) of variables (rather than sampling of individual variables), which reduces the dependence between the blocks and so increases the mixing rate (i.e. increasing the rate of convergence).

Results and discussion

It is possible to find an average of the difference between the DNA and physical examination generated probability distributions for each animal. An average distance, defined in equation where D denotes the population, p the probability distribution generated from DNA tests, q that from physical examinations, thus $p(a)(d)$ being used as shorthand for probability of affected status for dog d based on DNA test inference, can be used to provide a normalized (between

0 and 1) measure of similarity of the two distributions. A score close to 0 indicates that further statistical exploration of the inferred data is warranted.

$$\frac{1}{|D|} \sum_{d \in D} \frac{1}{2} (|p(a)(d) - q(a)(d)| + |p(c)(d) - q(c)(d)| + |p(n)(d) - q(n)(d)|) \quad (1)$$

As only 660 (14.7% of either DNA or physically tested) dogs have both observed sets of results, with only 14 affected animals, a better picture is provided by comparing the outcomes (affected, carrier or normal) corresponding to the highest inferred probability for each animal over the population. For example, for an untested animal which has the two (normal, carrier, affected) inferred probability distributions of $(0.879, 0.117, 0.004)$ and $(0.731, 0.263, 0.006)$ physical examination and DNA test respectively, the comparison will be of the two, highest scoring, normal results. Exploring the proportion of the number of animals with identical highest results generated by the two test methods out of the total number of animals yields the Jaccard index and gives a measure of similarity between two samples.

Using the inferred probabilities provides a larger sample while allowing the hypothesis that an inference mechanism based on pedigrees can be used reliably in the absence of a DNA test to be tested. The results can be further broken down by:

1. Distance to nearest tested animal.
2. Result: affected, carrier or normal.

Although the effect of actual observations will be propagated through pedigrees, it is necessarily going to be larger closer to a tested animal – the DNA test for OLFML3 became available in 2018, therefore a very small number (if any) animals born before 2000 would be directly tested meaning the impact

Table 4 Population metrics (average distance and Jaccard index) broken down by distance to tested animal

Distance	0	≤ 1	≤ 2	≤ 3	≤ 4	≤ 5
Jaccard	0.691	0.734	0.789	0.830	0.880	0.908
Avg distance	0.391	0.246	0.202	0.183	0.171	0.169

of tested dogs on dogs born in 1900 would be negligible, unlike the effect of the same on their parents and offspring. Separating by distance to the nearest tested animal allows both an investigation of the performance of the inference algorithm to a certain distance, as well as the correlation between physical exam and the DNA test.

Lastly, Spearman's correlation coefficient can be used to assess how well the relationship between two variables can be described using a monotonic function. It ranks the possible outcomes and examines the similarity of the ranking. A direct comparison of each animal's probability distribution turned into ranks would only be based on three data points ($n = 3$) and although these could be averaged over the entire population such a small value of n would not be sufficient for significant results. Instead, the probability distributions are converted into ($n =$) 11 bins $(0.0, 0.1, \ldots, 1.0)$ whose frequencies are gathered over the population. The ranks of these bins (based on DNA and physical examination tests) are then compared. Specifically, the rank of frequency of assigning 0.0 to the carrier status based on the DNA test inference is compared to the rank of frequency of assigning same based on physical examination based inference.

The results of the average distance and Jaccard index broken down by distance to tested animal can be found in Table 4.[2] Table 5 presents Spearman's correlation coefficient broken down by result and distance from tested animal.

[2] The maximum distance is 19, however, the effect of tests reduces greatly after approximately 3 steps.

Table 5 Spearman's coefficient (n = 11): each row represents the results for a given distance to tested animal

| Distance	Num dogs	Normal	Carrier	Affected
0	660	-0.459	-0.179	0.857
≤ 1	9,967	0.482	0.673	0.866
≤ 2	32,980	0.555	0.745	0.943
≤ 3	71,377	0.609	0.773	0.943

Table 4 demonstrates the power of the inference algorithm: since a gonioscopic examination can only assign a clear or an affected result, the average distance is expected to be high for directly tested, distance 0, animals (0.391) but the two probabilities quickly tend towards similar values as the distance away from the tested animals increases. The significant drop of the average from directly tested to animals those just one step away shows that within a short distance, the inference algorithm based on physical examination input quickly approaches the probabilities generated from DNA test results.

Examining distance 2 carrier probabilities based on DNA and physical examination based inference plotted against each other (Figure 1) indicates the presence of a correlation (the line of best fit is shown). However, the Figure also contains a large number of outliers, particularly around 0.5 and 1.0 (this is expected as for some animals DNA inference will be almost exact due to the parents' status). The outliers are further confirmed by the boxplots shown in Figure 2: unequal whiskers with a non centered mean and a small number of outliers for DNA based inference and a large number of outliers for physical examination based inference indicate that a comparison based on an assumption of bivariate normally distributed datasets, such as the Pearson coefficient, would not be appropriate.

Jaccard index shows a clearly rising, strong, correlation between the two probability distributions [Real (1999)], particularly given the large number of points over which it is found (the number of dogs at a given distance away from
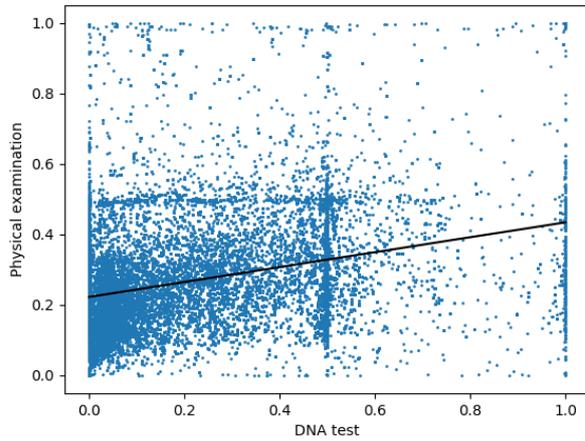
Fig. 1 Carrier probabilities inferred from DNA test against those from physical examination for distance 2 animals
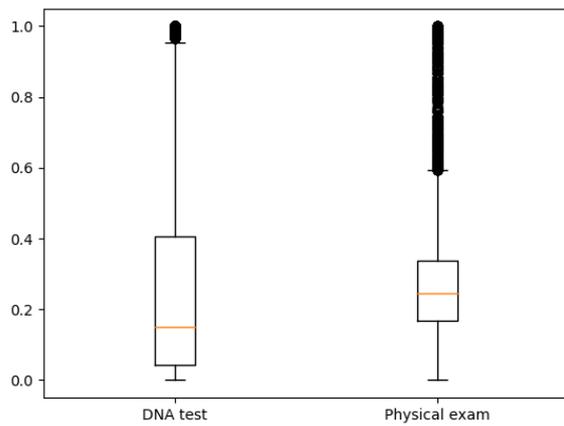
Fig. 2 Boxplot of carrier probabilities inferred from DNA test against those from physical examination for distance 2 animals

tested dogs can be seen in Table 5). Table 5 also shows a strong correlation (based on Spearman's coefficient) from distance 1 with significance levels of 0.1 for normal, 0.05 for carrier and 0.01 for affected (i.e. we can say with 99% certainty that the inferred affected probabilities at distance 1 did not occur by chance). There is also an obvious change from the observed (distance 0)

animals' Spearman correlation, where the negative value indicates that the normal DNA probabilities are generally lower than the physical examination based probabilities: this is due to the fact that a clear result on a physical examination will be distributed between a normal and a carrier with a strong bias towards normal due to the low number of affected animals, while DNA tested carriers will have probabilities (close to) 1.0. As the distance away from tested animals rises, this discrepancy reduces, supporting the hypothesis that the inference algorithm is capable of being employed with physical test results to yield results comparable to those generated from DNA test results.

The higher correlation at a small distance away from tested animals also suggests that the approach is capable of addressing omissions in the dataset: the data used in this work is based on voluntarily submitted results. Due to the stigma attached to the mutation being known to exist in a breeding line, it is possible that the database contains fewer affected (and probably carrier) animals than are present in the population. However, the distributions contained are feasible: assuming independence and random, OLFML3 uninformed, selection of partners (which would have been the case when DNA status was unknown) and denoting $P(\text{affected}) = x, P(\text{carrier}) = p, P(\text{normal}) = q(= 1 - x - p)$, then in a subsequent generation $P(carrier) = xp + 2xq + p^2/2 + pq$ as to produce a carrier, its parents must be either affected and carrier ($xp/2$), affected and normal (xq), both carriers ($p^2/2$), or one carrier other normal ($pq/2$). Substituting in for $P(normal)$ and ignoring x^2 (as it is small) gives $P(affected) = \frac{p^2}{4(1-p)}$. Substituting in 27% carriers yields 2% affected, i.e. 31 affected animals, which is comparable to the current value of 35.[3] Note that the independence assumption is unlikely to be valid, as some stud dogs will

---

[3] One testing laboratory, Animal Genetics UK, provided their tested numbers at June 2019, 753 normal (69%), 293 carrier (27%) and 40 affected (4%), showing that the number of affected animals in the dataset is slightly lower than true values.

be more popular than others and therefore it's possible for a mutated gene to be wider spread if a popular sire happens to be a carrier.

For completeness, animals which have both a physical gonioscopic examination result and a DNA test result are discussed. There are 660 such which show agreement in 618 clear and 14 affected cases. The disagreeing 28 cases consist of 9 dogs who were DNA tested affected but passed their physical gonioscopic exam (out of 23 dogs with an affected DNA result and a physical examination), 10 (out of 188) dogs DNA tested carrier but found affected during their physical gonioscopic exam and 9 (out of 449) dogs DNA tested normal who were found to be physically affected. The unexpected affected gonioscopic result is consistent with difficulty with the applied grading system as mentioned by [Oliver et al. (2017)]: for example, the 9 cases of disagreement for the normal DNA test contain 6 animals that were tested in the UK since the BVA grading scheme described in Table 2 was introduced with 5 of the animals scoring BVA grade 1 (the remaining was found to have BVA grade 2). BVA grade 1 is considered to be "mildly affected", and is to be assigned to an animal showing between 1 and 25% of changes. This range is rather large, and means that any change is marked as being physically affected. In fact, of the 7 animals with a BVA grade 1 (subject to them also having a DNA test result), 5 are DNA tested normal and the remaining 2 are DNA tested carrier (all of which would be expected to be unaffected using a physical exam) indicating that this particular grade may benefit from further revision. It is expected that some of the affected results also represent only minor changes and thus suffer a similar problem to the BVA grade 1 assignment. In fact, based on personal communication, this difficulty also appears in the other direction: animals assigned a "clear" (physical exam) result may display some changes, with an ophthalmologist deciding that the animal in front of them is not "affected".

References

[Al-Shuhaib (2019)] Al-Shuhaib, Mohammed. 2019. A comprehensive in silico prediction of the most deleterious missense variants in the bovine LEP gene. *BioTechnologia*, 100(4), 429–439.

[Al-Shuhaib et al. (2018)] Al-Shuhaib, Mohammed Baqur S, Al-Kafajy, Fadhil R, Badi, Milad Ali, AbdulAzeez, Sayed, Marimuthu, Kais, Al-Juhaishi, Hussein Ali Imran, & Borgio, J Francis. 2018. Highly deleterious variations in COX1, CYTB, SCG5, FK2, PRL and PGF genes are the potential adaptation of the immigrated African ostrich population. *Comput Biol Med*, 100, 17–26.

[Al-Shuhaib et al. (2019)] Al-Shuhaib, Mohammed Baqur S, Al-Kafajy, Fadhil R, & Al-Jashami, Ghadeer S. 2019. A computational approach for explaining the effect of the prl gene polymorphism on prolactin structure and biological activity in Japanese quails. *Anim Biotechnol*, 29, 1–9.

[Burdick et al. (2006)] Burdick, Joshua T, Chen, Wei-Min, Abecasis, Gonçalo R, & Cheung, Vivian G. 2006. In silico method for inferring genotypes in pedigrees. *Nat Genet*, 38, 1002–1004.

[Dijkstra (1959)] Dijkstra, Edsger Wybe. 1959. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1, 269–271.

[Gao et al. (2009)] Gao, Guimin, Allison, David B, & Hoeschele, Ina. 2009. Haplotyping methods for pedigrees. *Hum Hered*, 248–266.

[Gellatly & Sewell (2019)] Gellatly, Nichola, & Sewell, Fiona. 2019. Regulatory acceptance of in silico approaches for the safety assessment of cosmetic-related substances. *Computational Toxicology*, 11, 82–89.

[Jensen et al. (1995)] Jensen, Claus S., Kjærulff, Uffe, & Kong, Augustine. 1995. Blocking Gibbs sampling in very large probabilistic expert systems. *International Journal of Human-Computer Studies*, 42(6), 647–666.

[Lang et al. (2018)] Lang, Annemarie, Volkamer, Andrea, Behm, Laura, Röblitz, Susanna, Ehrig, Rainald, Schneider, Marlon, Geris, Liesbet, Wichard, Joerg, & Buttgereit, Frank. 2018. In Silico Methods – Computational Alternatives to Animal Testing. *ALTEX - Alternatives to Animal Experimentation*, 35(1), 126–28. Available from: https://www.altex.org/index.php/altex/article/view/105.

[Lindholm et al. (2004)] Lindholm, Eva, Zhang, J., Hodge, Susan E, & Greenberg, David A. 2004. The Reliability of Haplotyping Inference in Nuclear Families: Misassignment Rates for SNPs and Microsatellites. *Human Heredity*, 57, 117–127.

[Oliver et al. (2017)] Oliver, J. A. C., Cottrell, B. C., Newton, J. R., & Mellersh, C. S. 2017. Gonioscopy in the dog: inter-examiner variability and the search for a grading scheme. *Journal of Small Animal Practice*, 58(11), 652–658.

[Oliver et al. (2020)] Oliver, James A C, Wright, Hattie, Massidda, Paola A, Burmeister, Louisa M, & Mellersh, Cathryn S. 2020. A variant in OLFML3 is associated with pectinate ligament abnormality and primary closed-angle glaucoma in Border Collies from the United Kingdom. *Vet Ophthalmol*, 23(1), 25–36.

[Preiss, 2019-06-19] Preiss, Judita. *Anadune Border Collie Database*. `https://www.anadune.com/abcd`. Retrieved 2019-06-19.

[Pugh et al. (2019)] Pugh, Carys A., Farrell, Lindsay L., Carlisle, Ailsa J., Bush, Stephen J., Ewing, Adam, Trejo-Reveles, Violeta, Matika, Oswald, de Kloet, Arne, Walsh, Caitlin, Bishop, Stephen C., Prendergast, James G. D., Rainger, Joe, Schoenebeck, Jeffrey J., & Summers, Kim M. 2019. Arginine to Glutamine Variant in Olfactomedin Like 3 (OLFML3) Is a Candidate for Severe Goniodysgenesis and Glaucoma in the Border Collie Dog Breed. *G3 (Bethesda)*, 9(3), 943–954.

[Real (1999)] Real, Raimundo. 1999. Tables of significant values of Jaccard's index of similarity. *Miscellania Zoologica*, 22(1), 29–40.

[Sherry et al. (2001)] Sherry, Stephen T, Ward, M H, Kholodov, M, Baker, J, Phan, L, Smigielski, E M, & Sirotkin, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, 29(1), 308–11.

[Velazco et al. (2019)] Velazco, Julio G, Malosetti, Marcos, Hunt, Colleen H., Mace, Emma S., Jordan, David R., & van Eeuwijk, Fred A. 2019. Combining pedigree and genomic information to improve prediction quality: an example in sorghum. *Theoretical and Applied Genetics*, 132, 2055–2067.