Using regions of homozygosity to evaluate the use of dogs as preclinical models in human drug development

Smieszek SP., Polymeropoulos MP.

¹ Vanda Pharmaceuticals Inc., Washington, DC

Corresponding Author: Sandra.Smieszek@vandapharma.com

Keywords: ROH, genomics, consanguinity, dog genome, deleterious variants, CYPD6, ADME genes

Abstract

Animals are used as preclinical models for human diseases in drug development. Dogs, especially, are used in preclinical research to support the clinical safety evaluations during drug development. Comparisons of patterns of regions of homozygosity (ROH) and phenotypes between dog and human are not well known. We conducted a genome-wide homozygosity analysis (GWA) in the human and the dog genomes.

We calculated ROH patterns across distinct human cohorts including the Amish, the 1000 genomes, WellSiderly, Vanda 1k genomes, and Alzheimer's cohort. The Amish provided a large cohort of extended kinships allowing for in depth family oriented analyses. The remaining human cohorts served as statistical references. We then calculated ROH across different dog breeds with emphasis on the beagle - the preferred breed used in drug development.

Out of five studied human cohorts we reported the highest mean ROH in the Amish population. We calculated the extent of the genome covered by ROH (F_{ROH}) (human 3.2Gb, dog 2.5Gb). Overall F_{ROH} differed significantly between the Amish and the 1000 genomes, and between the human and the beagle genomes. The mean F_{ROH} per 1Mb was ~16kb for Amish, ~0.6kb for Vanda 1k, and ~128kb for beagles. This result demonstrated the highest degree of inbreeding in beagles, far above that of the Amish, one of the most inbred human populations.

ROH can contribute to inbreeding depression if they contain deleterious variants that are fully or partially recessive. The differences in ROH characteristics between human and dog genomes question the applicability of dog models in preclinical research, especially when the goal is to gauge the subtle effects on the organism's physiology produced by candidate therapeutic agents. Importantly, there are huge differences in a subset of ADME genes, specifically cytochrome P450 family (CYPs), constituting major enzymes involved in drug metabolism. We should hesitate to generalize from dog to human, even if human and beagle are relatively close species phylogenetically

INTRODUCTION

Animals are used as preclinical model for human diseases in drug development. While central to drug development, there is lack of validity for many of such tests. Evidential weight gain accrued by the use of animal data was assessed in a recent study, evaluating the probability that a certain new tested drug may prove toxic to humans¹. The authors calculated likelihood ratios (LR) for a large set of drugs. They have shown that lack of toxicity in fact provides no weight gain for lack of adverse drug reactions, as well as large inconsistencies in LR when predicting from animals to human¹.

Dogs, especially, are used in preclinical research to support the clinical safety evaluations during drug development, however, dogs are a particularly inbred species. Recently, EMBARK project conducted large scale characterization for the dog ROH density maps for 2500 dogs comparing a range of breeds². The presented results recorded ROH down to 500 kilobases. Authors report more than '678 homozygous deleterious recessive genotypes in the panel across 29 loci, 90% of which overlapped with ROH'². The study constitutes one of the most comprehensive evaluations of dog genome across several breeds of dogs, showing great degree of variation among breeds. Patterns of ROH in dogs have seem to imply a high degree on inbreeding across multiple breeds of dogs, regions furthermore associating with deleterious variants². Long ROH are enriched in deleterious variation³. It has been further shown that inbreeding as observed via genomic measures reduces fecundity in Golden Retrievers⁴. It is well established that high degree of inbreeding results in decreased reproductive fitness as discussed initially by Charles Darwin in his initial observations in plants followed by in humans^{5, 6}. It is hence well established that inbreeding increases the incidence of recessive disease as discussed already in 1902 in the case of alkaptonuria⁷ and in cases of Mendelian disorders such as Tay-Sachs⁸. Moreover, the implications of extensive inbreeding were recently demonstrated and quantified in a sequencing study of grey wolves of Isle Royale, where inbreeding depression which has brought wolves to the brink of extinction likely an effect accounted for increased homozygosity of the deleterious variants⁹. The comparison of regions of homozygosity (ROH) patterns between dog and human has not been characterized. In the present study we focus on exactly that comparison. A ROH is defined as a continuous stretch of DNA sequence without heterozygosity in the diploid state (min ROH $\geq$ 1.5 Mb). We calculated ROH patterns across distinct human cohorts: the Amish, IGSR 1000 Genomes¹⁰, Welllderly¹¹, Vanda 1K genomes using both microarray and sequencing

We have incorporated several datasets in the analysis to have a broader view upon the potential differences in and between the two species as across breeds, across human subpopulations. The Amish with large families, well documented family history accessible via Anabaptist Genealogy Database constitute a valuable resource for genetic studies¹². Amish constitute an example of a highly inbred population, in fact ROH were first described in the Amish. The 1000 genomes serve as mixed control representation of individuals. Vanda 1000 genomes is a control set of whole genome sequencing samples with extensive phenotyping. The

Welllderly present a well-being well aged cohort of mixed ethnicity. The Alzheimer's cohort is a mixed ethnicity cohort offering yet another point of comparison. Using the same settings across datasets we calculated ROH across different dog breeds (EMBARC project) with emphasis on the beagle, as it is the preferred breed in drug development. We evaluate the degree of inbreeding in dogs compared to human populations. We furthermore focus on the CYP family, as they are the major class of enzymes involved in drug metabolism, accounting for 75% of the total metabolism¹³. The results should allow one to investigate the genetic validity of dogs as preclinical models for human drug development.

RESULTS

A ROH is defined as a continuous stretch of DNA sequence without heterozygosity in the diploid state (min ROH = established usually between 500kb and 1.5 Mb). We estimate ROH using PLINK¹⁴. The hypothesis is that deleterious genotypes will be enriched in the ROH regions as compared by background. We characterize the distribution of ROH across dog breeds, across human population cohorts as well as between human and the dog evaluated on the same parameters. **Figure 1** displays the length and the number of ROHs identified across numerous dog breeds.

Figure 1 The total number and length of ROH identified across dog breeds as compared to mixed ethnicity human cohort: Vanda 1000 genomes.

Vanda 1k and the Welllderly ROH analysis was conducted on whole genome sequencing data. When we considered both the location and the allelic form of the ROHs, we were able to separate the populations by PCA, demonstrating that ROHs contain information on the demographic history and structure of a population. We calculated the extent of the genome covered by ROH (F_{ROH}) (human 3.2Gb, dog 2.5Gb) as displayed in **Table 1**.

138 Table 1 Displays the total, mean ROH and (F_{ROH}) across human cohorts compared to dogs

	Amish	Welllderly	ADNI	1000g	Vanda1k	Beagles
total ROH (n)	13165	2404	5534	1977	1012	4896
mean ROH per individual	14.74	4.7	6.72	3.14	1.03	61.97
mean length (kb)	6343	2203	3067	2445	1863	5203
median length (kb)	3577	1858	1856	2309	1746	3137
mean F_{ROH}	0.016057	0.003233	0.006532	0.002402	0.000604	0.128986
median F_{ROH}	0.013029	0.002821	0.004122	0.001553	0.000469	0.094472

139

140

141

142 In the Amish population, one of the most inbred human populations, the F_{ROH} was 12x
143 smaller in comparison to the beagle population. F_{ROH} differed significantly between the Amish
144 and the 1000 genomes, and between the human and the beagle genomes. The mean F_{ROH} per
145 1Mb was ~16kb for Amish, ~0.6kb for Vanda 1k, and ~128kb for beagles. This result
146 demonstrated the highest degree of inbreeding in beagles, far above that of the Amish, one of the
147 most inbred human populations also displayed on **Figure 2**. Figure 2 exemplifies the
148 significantly higher observed ROH as compared to several human populations including the
149 Amish. **Figure 3** further compares the human populations and the specific regions of high ROH
150 across the human genome.

Figure 2 Mean F_{ROH} on the y axis and mean number of ROH segments across several human and dog cohorts

Figure 3 ROH patterns across human cohorts: Human genome arranged circularly end-to-end: from inside to outside, the rings display ROH across different human cohorts starting from the innermost: 1000g, ADNI, Welllderly and Amish (blue). The outermost ring represents conservation score averaged per region of the human genome.

Currently, we classify 298 genes that encode Phase I and II drug metabolizing enzymes, transporters, and modifiers as ADME genes, with CYPs constituting a major family. To further evaluate the consequences we focused on CYP family including most common CYPs involved in metabolism of human drugs including CYP2D6, CYP2B's, CYP2C's, CYP3A's. We mapped orthologues between dogs and human using BIOMART¹⁵ (orthologues, from human to dogs). We report that beagles alone, F_ROH, would be 25% so even higher than our average in beagles, 12.9%. For humans that estimate would be 1/1000 on the same parameter. The results are presented in **Figure 4**.

Figure 4 ROH patterns in CYPs compared between humans and dogs. Especially high signal is seen for CYP2C18, CYP2D6, CYP3A4, CYP3A43 amongst others.

DISCUSSION

The study found a high degree of homozygosity across the genomes of beagle dogs, an indication of high degree of inbreeding. High degree of inbreeding creates an evolutionary bottleneck and exposing homozygous deleterious variants in the genome of the species. Furthermore due to heterozygote advantage as outlined by overdominance hypothesis, the long term loss of diversity furthermore reduced heterozygote advantage. For example, the risk for autosomal recessive disease is proportional to the degree of parental relationship in in consanguineous families¹⁶. In humans total length of ROH per individual show considerable variation across individuals and populations with higher values occurring more often in populations with known high frequencies of consanguineous unions. Even across the genome the distribution is far from uniform¹⁷. Across the genome the variation is correlated with recombination rate, as well as with signals of recent positive selection. Importantly, long ROH are more frequent in genomic regions harboring genes associated with autosomal-dominant diseases¹⁸. Interestingly the top-ranked ROH hotspot in the human genome as assessed by the authors is located on chromosome 2p in a region with low recombination rates and containing CYP26B1¹⁸.

We compared the degree of genetic inbreeding in beagle dogs using the measure of runs of homozygosity (ROH). This therefore focus on contiguous regions of the genome that are

inherited as identical from both parents. The study showed that unrelated humans show these regions of homozygosity in about 1/1000 of their genomes while beagle dogs do so in 12.8 percent of their genome. In other words beagle dogs are on average 100 times more inbred than humans. In the same study researchers examined the degree of inbreeding in the genetic isolate of the Amish in Pennsylvania which showed that they were about 12 times more inbred than the general population. It is well accepted that drug efficacy and safety studies in genetic isolates such as the Amish cannot be generalized for the population at large.

The ADME genes constitutes a highly polymorphic set of genes involved in Absorption, Distribution, Metabolism, and Excretion of drugs. This was exemplified in a recent effort to characterize the distribution of 95 polymorphisms in 31 core ADME genes in 20 populations worldwide¹⁹. The present effort to evaluate CYPs not within but in the context of between species as displayed on Figure 4, comes without a surprise. CYP2D6, another gene with allelic variants and encoding enzymes with variable degrees of activity, is associated with the development of hepatotoxicity after use of certain pharmaceutical agents. The CYP2D6 gene is polymorphic with over 150 allelic variants as defined by (<https://www.pharmvar.org/gene/CYP2D6>)²⁰. The high levels of variation in the frequencies of alleles of polymorphic pharmaco-genes between diverse ethnic groups may be responsible for severe adverse reactions to as well as altered potentially efficacy of a wide variety of drugs²¹.

Beagle dogs are routinely used for studying the safety of new pharmaceuticals. This practice has been in place for a hundred years and followed by pharmaceutical developers and regulators. The present study is significant in that it casts a big question over the predictive validity of these commonly conducted dog studies. The very high degree of inbreeding seen would suggest that beagle dogs would be a poor model to predict toxicity even for other dog breeds and of course an inappropriate model to do so for humans. Reliance upon the beagle dog to predict human drug safety is not warranted especially if metabolism of a drug is known to be driven by CYP family in example. The results of this study should lead to the abandonment of the routine use of dogs in human drug toxicology studies and urge researchers and regulators to instead adopt appropriate and relevant scientific approaches in order to ensure human drug safety.

CONCLUSION

The fraction of the genome covered by ROH (F_{ROH}) was significantly different between human populations (Amish and 1000 genomes) and between humans and beagles. In the Amish population, one of the most inbred human populations, the F_{ROH} was 12 times smaller in comparison to the beagle population. ROH patterns in beagles increase their susceptibility to inbreeding depression and may introduce deleterious or recessive traits into their genome. Due to the high degree of inbreeding observed in beagles, preclinical research should use caution when

generalizing from dog to human (especially when the enzymes metabolizing the drug are known), despite the physiological similarities between the species.

METHODS

Datasets

ROH detection

Genomic data analysis ROH

Calculation of ROH in human cohorts including the Amish, IGSR 1000 genomes, Welllderly, Vanda 1k genomes. We detected ROH using PLINK. ROH scores reflect the probability of a stretch of SNPs being homozygous due to LOH and they are determined via the homozygosity frequency for each SNP in the genome. We detected ROH using PLINK on the EMBARK dog data. We next calculated the extent of the genome covered by ROH (F_{ROH}). ROH were defined as runs of at least 50 consecutive homozygous SNPs spanning at least 1500 kb, with less than a 1000 kb gap between adjacent ROH and a density of SNP coverage within the ROH of no more than 50 kb/SNP, with one heterozygote and 5 no calls allowed per window. Detection of ROH, depletion with confidence scores, LD in ROH, region analysis and comparison across species

DNA quantification

Incoming nucleic acid samples are quantified using fluorescent-based assays (PicoGreen) to accurately determine whether sufficient material is available for library preparation and sequencing.

DNA integrity

DNA sample size distributions are profiled by a Fragment Analyzer (Advanced Analytics) or BioAnalyzer (Agilent Technologies), to assess sample quality and integrity.

Genotyping

At the NYGC, we run the HumanCoreExome 24v1.3 array for all human DNA samples that we sequence.

WGS library preparation and sequencing, Truseq PCR-free (450bp)

Whole genome sequencing (WGS) libraries were prepared using the Truseq DNA PCR-free Library Preparation Kit

WGS Germline analysis part I

Whole Genome data were processed on NYGC automated pipeline. Paired-end 150 bp reads were aligned to the GRCh37 human reference (BWA-MEM v0.7.8) and processed with GATK best-practices workflow (GATK v3.4.0).

The mean coverage is 35.8, it reflects the samples average. The coverage for the FLG region tracks with the coverage for the whole genome. The coverage information per FLG region per sample is provided in the supplementary file.

All high quality variants obtained from GATK were annotated for functional effects (intronic, intergenic, splicing, nonsynonymous, stopgain and frameshifts) based on RefSeq transcripts using Annovar (<http://www.openbioinformatics.org/annovar/>)²². Additionally, annovar was used to match general population frequencies from public databses (Exac, GnomAD, ESP6500, 1000g) and was used to prioritize rare, loss-of-function variants.

1. Bailey, J., Thew, M. & Balls, M. An Analysis of the Use of Animal Models in Predicting Human Toxicology and Drug Safety. *Altern. to Lab. Anim.* **42**, 181–199 (2014).
2. Sams, A. J. & Boyko, A. R. Fine-Scale Resolution of Runs of Homozygosity Reveal Patterns of Inbreeding and Substantial Overlap with Recessive Disease Genotypes in Domestic Dogs. *G3 (Bethesda)*. **9**, 117–123 (2019).
3. Szpiech, Z. A. *et al.* Long Runs of Homozygosity Are Enriched for Deleterious Variation. *Am. J. Hum. Genet.* **93**, 90–102 (2013).
4. Chu, E. T. *et al.* Inbreeding depression causes reduced fecundity in Golden Retrievers. *Mamm. Genome* **30**, 166–172 (2019).
5. DYER, W. T. T. The Effects of Cross and Self-Fertilisation in the Vegetable Kingdom. *Nature* **15**, 329–332 (1877).
6. Berra, T. M., Alvarez, G. & Ceballos, F. C. Was the Darwin/Wedgwood Dynasty Adversely Affected by Consanguinity? *Bioscience* **60**, 376–383 (2010).
7. Garrod, A. E. The incidence of alkaptonuria: a study in chemical individuality. 1902. *Mol. Med.* **2**, 274–82 (1996).
8. Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. Runs of

- homozygosity: windows into population history and trait architecture. *Nat. Rev. Genet.* **19**, 220–234 (2018).
9. Robinson, J. A. *et al.* Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. *Sci. Adv.* **5**, eaau0757 (2019).
10. Clarke, L. *et al.* The 1000 Genomes Project: data management and community access. *Nat. Methods* **9**, 459–462 (2012).
11. Erikson, G. A. *et al.* Whole-Genome Sequencing of a Healthy Aging Cohort. *Cell* **165**, 1002–11 (2016).
12. Agarwala, R., Biesecker, L. G. & Schäffer, A. A. Anabaptist genealogy database. *Am. J. Med. Genet. Part C Semin. Med. Genet.* **121C**, 32–37 (2003).
13. Wang, J.-F. & Chou, K.-C. Molecular Modeling of Cytochrome P450 and Drug Metabolism. *Curr. Drug Metab.* **11**, 342–346 (2010).
14. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
15. Smedley, D. *et al.* The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res.* **43**, W589–98 (2015).
16. Sund, K. L. *et al.* Regions of homozygosity identified by SNP microarray analysis aid in the diagnosis of autosomal recessive disease and incidentally detect parental blood relationships. *Genet. Med.* **15**, 70–78 (2013).
17. Pembroke, W. G., Babbs, A., Davies, K. E., Ponting, C. P. & Oliver, P. L. Temporal transcriptomics suggest that twin-peaking genes reset the clock. *Elife* **4**, 1–15 (2015).
18. Pemberton, T. J. *et al.* Genomic patterns of homozygosity in worldwide human populations. *Am. J. Hum. Genet.* **91**, 275–92 (2012).
19. Škarić-Jurić, T. *et al.* Characterization of ADME genes variation in Roma and 20 populations worldwide. *PLoS One* **13**, e0207671 (2018).
20. Gaedigk, A., Whirl-Carrillo, M., Pratt, V. M., Miller, N. A. & Klein, T. E. PharmVar and the Landscape of Pharmacogenetic Resources. *Clin. Pharmacol. Ther.* **107**, 43–46 (2020).
21. Rodrigues, J. C. G. *et al.* Polymorphisms of ADME-related genes and their implications for drug safety and efficacy in Amazonian Amerindians. *Sci. Rep.* **9**, 7201 (2019).
22. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).

332

333 <https://www.nature.com/articles/s41598-019-43610-y>

334

335 <https://www.frontiersin.org/articles/10.3389/fgene.2019.00007/full>

336

337 https://www.fasebj.org/doi/abs/10.1096/fasebj.2019.33.1_supplement.814.10

338

339

340

341 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4013674/pdf/ijms-15-06990.pdf>

342

343 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4548351/pdf/JCTH-3-099.pdf>

344

345 <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005280>

346

347

348

349

350