Short Visit Grant □ or Exchange Visit Grant ☒

(please tick the relevant box)

Scientific Report

Scientific report (one single document in WORD or PDF file) should be submitted online within one month of the event. It should not exceed eight A4 pages.

Proposal Title : Sequence-based characterization of farm animal biodiversity: revealing the genetic basis of body size variation in domesticated pigs.

Application Reference N°: 4128

1) Purpose of the visit

The Göttingen Minipig is one of the smallest pig breeds in the world (SIMIANER AND KÖHN, 2010). Therefore it is a promising candidate to clarify the huge variety in body size of domesticated pigs. Nowadays, the whole genome sequencing technology gives us the possibility to have a detailed look on at animals genetic information at relatively low costs. We took the decision to sequence a group of our Göttingen Minipigs and some individuals of another miniature pig breed, the Berlin Minipig (MiniLEWE) and to combine this information with publicly available sequencing information from other domesticated pigs and some wild boars as well as with several outgroup species such as the African warthog or warty pigs from the Philippines. In order to find a experienced partner to work on this project, we aimed for a collaboration with Dr. Carl-Johan Rubin from the Biomedical Center of Uppsala University. This workgroup has rich experience with the analysis of sequencing data to find proofs of domestication in animals (RUBIN ET AL., 2012) as well as access to a very powerful server cluster. A three month research stay in Uppsala was arranged to conduct the basic data preparation and to run first analyses to find differences between miniature and normal sized breeds. Simultanously, DNA samples of ten individual Göttingen Minipigs, two Berlin Minipigs, a DNA-pool from Berlin Minipigs and a pooled sample consisting of DNA from animals of the founder breeds of the Göttingen Minipig where queued at the sequencing facility in Uppsala.

2) Description of the work carried out during the visit

Unfortunately, the commisioned sequences where not ready until the end of the stay in Uppsala due to a long queue at the facility but will be processed during the summer. Alternatively we were able to download the whole genome of a lately sequenced Göttingen Minipig from another study (VAMATHEVAN ET AL., 2013) which was deposited in the European Nucleotide Archive and the genome of a Wuzhishan Minipig from China (FANG ET AL., 2012). As the representatives of normal sized pigs, we dowloaded 37 domestic pigs, either from Europe or Asia, 11 wild boars and 6 warthogs as outgroup animals, which were underlying material of the studies by GROENEN ET AL. (2012) and RUBIN ET AL. (2012)

We used the susScr3 (build 10.2, ARCHIBALD ET AL., 2010) available in the UCSC genome browser as the reference genome. This reference was indexed using BWA (LI AND DURBIN, 2009). Afterwards the downloaded sequences were aligned against this reference with the short-read aligning algorythm of BWA (LI AND DURBIN, 2009). The resulting BAM files were sorted with samtools (LI ET AL, 2009) and PCR duplicates were marked with Picard-tools (PICARD, 2009). The resulting BAM Files were indexed with Picard accordingly.

We then evaluated the depth of coverage of the single samples with GATK (MCKENNA ET AL., 2010, DEPRISTO ET AL., 2011). Since different individuals had different average depth, we normalized the results from GATK, so that every individual had the same average depth across the whole genome and summarized them in windows along the genome with a custom made script.

The final SNP calling was done with the Unified Genotyper from GATK with default options for both, single nucleotide variants (SNV) and Indels. Statistics on the quality parameters were build in order to have a basis for deciding on an appropriate filtering. Afterwards we removed all outgroup animals and all SNPs in which a variant allele was only present in the outgroups. The first subsequent filtering only on SNVs with GATK VariantFiltration used the following options: SNP Cluster were removed, if there were more than 5 SNPs in a range of 20 basepairs. A SNP was removed, if either the BaseQualityRankSum, the MappingQualityRankSum or the ReadPosition-RankSum were lower than -6. In additon FisherStrandValues higher than 26 were removed. Filtering on mapping quality was not carried out, since we wanted to keep SNPs which were only present in the minipigs or just a few more domestics or wild.

As the final and most important filter we used a custom made script to exclude loci with an insufficient or extremly high depth. Therefor we calculated the distribution of the depth of coverage on chromosomes 3, 13, X and Y over all domestic pigs, wild boars and the minipigs. We decided to filter away all loci with a coverage lower than approximately half of the mean coverage, i.e. 150 X and all positions with a coverage of roughly two times the mean coverage without outgroups, i.e. 630 X.

The filtered dataset was annotated with genes from the Ensembl (FLICEK ET AL., 2013) database using the software Annovar (WANG ET AL., 2010).

The next step was the calculation of F_{ST} values for the breed contrast of domestics against wild boars and minipigs to determine regions with high diversification between these groups. The F_{ST} statistic was calculated after the formula by Weir (1996):

$$F_{ST} = \frac{\begin{array}{c} n_i(p_i - p)^2 \\ \qquad\qquad (r-1)n \end{array}}{p(1-p)}$$

Where p is the allele frequency of the first allele over all subpopulations, p_i is the the allele frequency of the first allele within the subpopulation i, n_i is the number of individuals of a subpopulation i, n is the average subpopulations size and r is the number of subpopulations. The term had to be corrected, because it overestimated the F_{ST} values in a systematical manner. We exchanged $(r-1)n$ against $r * n$, so that the maximum value to occur was 1. To reduce the effect of outliers and to reduce the number of data points, the resulting F_{ST} values were summarized in 40 and 10 kilobasepair windows which were 50 % overlapping. To identify windows with extraordinary high values, a treshold was calculated. The treshold was the lowest F_{ST} value of the top 0.1 % quantile of all windows. Windows with a higher value were identified.

To determine the composition of the sequenced animals from their ancestral breeds and to figure out a reasonable number of founder breeds, we used the program "Admixture" (Alexander et al, 2009). Since "Admixture" is not able to take LD between SNPs into account properly and we had no knowledge of the actual LD structure within our samples, we performed the analysis for several marker distances (~8 kb, ~39 kb and ~117 kb) over all autosomes as well as for chromosome 1 with an average marker distance of ~1 kb.

3) Description of the main results obtained

In general, we obtained an extremly large and valuable high quality dataset for further analysis during the stay in Uppsala. It gives us the possibility to compare pigs from different continents, different levels of domestication and breeds with highly different properties with each other. On the other hand our project is lagging at the moment, because our target was to include a lot of individuals from miniature pig breeds. Since they will not be ready until August, we had to manage the first steps with only two minipigs from public sources and not of the same breed.

The analysis of the sequencing depth before filtering revealed, that the domestic, wild and outgroup animals were sequenced at an average depth of 6 up to 8 X. The two minipig genomes resulted from two assembly studies so that in total approximately 80 X were available. These samples were downsampled to an average coverage of at least 12 up to 15 X. While plotting the normalized data for every group (domestic, wild, mini, outgroup) we found regions of extremly high coverage values. The plot for chromosome 7 is shown exemplarily (**Figure 1**).

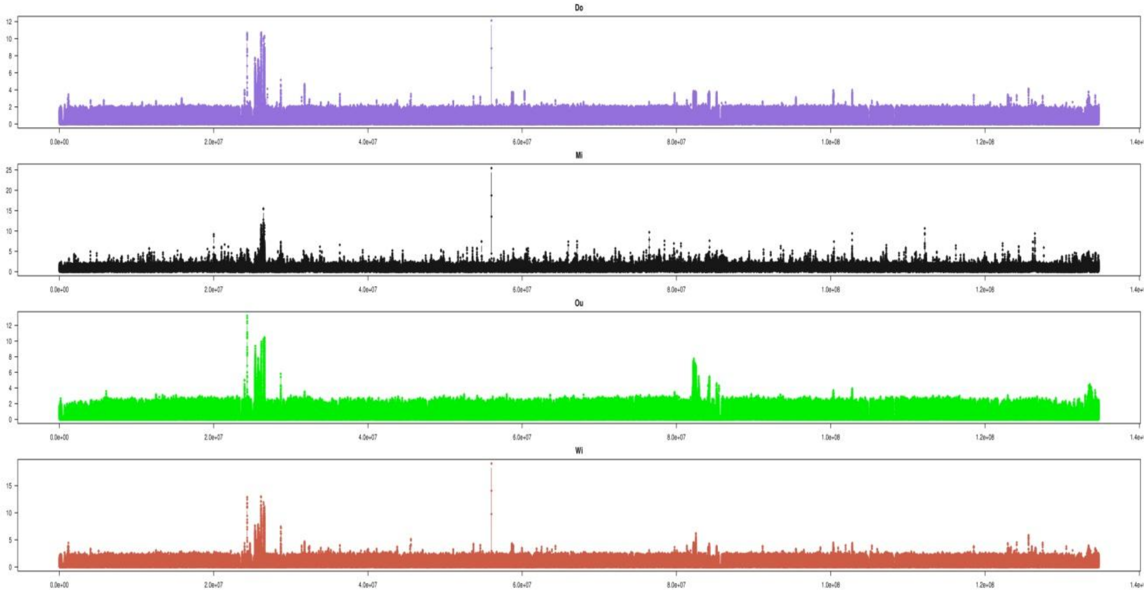

Figure 1: Sequencing depth of chromosome 7, divided by subpopulations. The plot shows the domestics, minipigs, outgroup animals and wild boars (top to bottom). It is remarkable, that there are regions, which show a different pattern between the groups, i.e. around 56000000 bp where the peak is not present in the outgroup animals or the region around 25000000 bp where the pattern in minipigs looks different than in all other groups.

After the final filtering the SNP set contained about 30.1 million SNPs on 18 autosomes and in the unknown regions. The sex chromosomes where processed seperately but not icluded in further analysis. **Table 1** shows the number of SNPs per chromosome and in total.

Table 1: Number of SNPs after filtering

Chromosome	No. of SNPs	Chromosome	No. of SNPs
chr_1	2936902	chr_11	1189587
chr_2	1920120	chr_12	865076
chr_3	1739426	chr_13	2307108
chr_4	1719589	chr_14	1846826
chr_5	1380430	chr_15	1687854
chr_6	1830387	chr_16	1184457
chr_7	1677898	chr_17	966280
chr_8	1797958	chr_18	870279
chr_9	1943120	chr_Un	990819
chr_10	1282471	**Total**	**30136587**

The calculation of F_{ST} statistics between the domestics, wild boars and minipigs revealed regions with a remarkable diversification. Taking the 0.1 %

most important windows into account, there were 116 windows. **Figure 2** gives an overview for all autosomes.

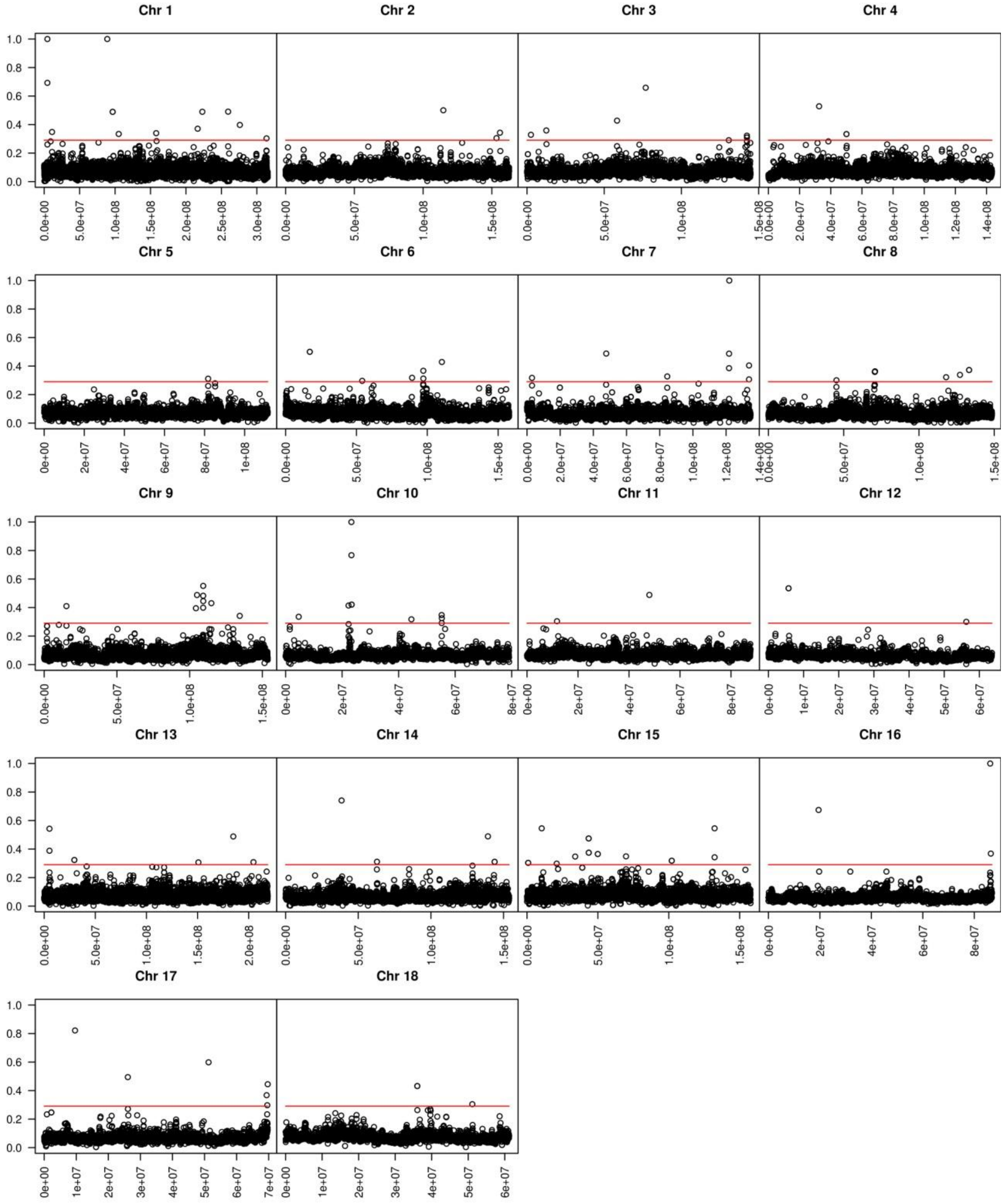

Figure 2: F_{ST} statistics for domestic pigs, wild boars and minipigs, summarized in 40 kb windows with 50 % overlap for all autosomes

To determine the breed composition of the included individuals, we used the program “Admixture”. We choose the cross-validation argument to decide for a correct K-value. **Table 2** shows the attributes of the compared runs.

Table 2: Attributes of all “Admixture” runs

Region	Chr 1	Chr 1:18	Chr 1:18	Chr 1:18
No. of SNP icluded	300000	314445	62883	20956
Sum of chromosome length [kb]	315321	2450713	2450713	2450713
Average marker distance [kb]	1.051	7.793	38.972	116.945
Results of the Cross Validation				
CV error (K=1):	0.66975	0.67278	0.67466	0.67548
CV error (K=2):	0.55316	0.56820	0.57104	0.56675
CV error (K=3):	0.54821	0.57610	0.57267	0.56856
CV error (K=4):	0.59074	0.61469	0.61545	0.61073
CV error (K=5):	0.63071	0.64771	0.65411	0.65021
CV error (K=6):	0.63207	0.64958	0.66459	0.65760

As the cross validation results show there are only slight differences between the errors in different scenarios. The final results of the admixture analysis show that there is no real dependency on the marker density in our case. In all scenarios with K=2, “Admixture” detects a clear differentiation of Asian and European breeds. The Wuzhishan minipig clusters clearly with the Asian cohort, whereas the Göttingen minipig shows genetic admixture of Asian and European origins. When we choose K=3, a clear fractioning into european wild boars, Large White related strains and asian origins could be seen. It is remarkable, that breeds like the Landrace, Pietrain and Hampshire shared important genome sections with Large White and the remaining sections with wild boars, whereas the Durocs cluster perfectly with the European wild boars. In that case, the Wuzhishan minipig still showed perfect affiliation with the Asian breed, whilst the Göttingen minipig carries minor parts from the Large White and the european wild boar clusters. **Figure 3** shows the results for the analysis of all autosomes with 20000 SNP markers exemplarily.

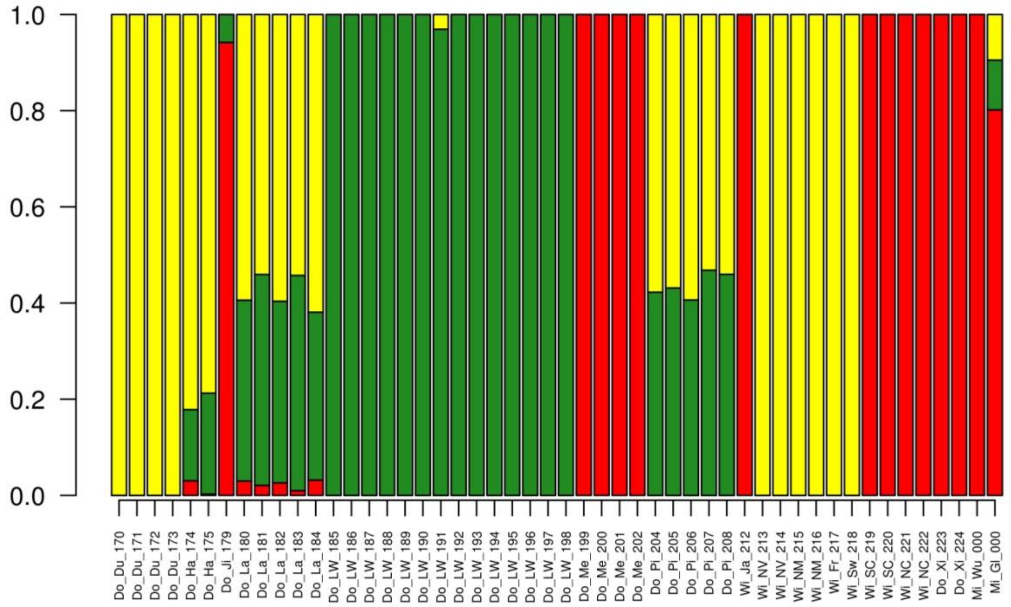

Figure 3: Genetic admixture of the project animal with K=3 estimated from 20000 markers equally distributed over all autosomes. (Group encoding: Do=domestic, Wi= Wild boar, Mi=Minipig
Breed encoding: Du=Duroc, Ha=Hampshire, Ji=Jiangquhai, La=Landrace, LW=Large White, Me=Meishan, Pi= Pietrain, Ja= Japanese wild boar, NV/NM= Wild boar Netherlands, Fr=Wild boar France, Sw= Wild boar Switzerland, SC/NC=Wild boar South/North China, Xi=Xiang, Wu= Wuzhishan, Gl= Göttingen Minipig)

4) Future collaboration with host institution (if applicable)

As mentioned before, we are awaiting another 10 whole genome sequences from our Göttingen Minipigs, two from Berlin Minipigs and two from DNA pools from Berlin Minipigs and some founder animals respectively, which are being processed in Uppsala at the moment. When the facility in Uppsala finished sequencing, the data will be transferred to the Uppmax server cluster where it will be handled with the same pipeline described in section 2.. With the additional data we will be able to do more powerful analysis, since we were limited by the low number of only two minipigs up to now.

Within this collaboration we scheduled a meeting of Prof. Simianer, Dr. Rubin and Christian Reimer in Göttingen in September to discuss the approach for the new data and for Dr. Rubin to give a presentation on sequencing techniques to students and scientific stuff. In return, Prof. Simianer and Christian Reimer will visit Uppsala in October/ November for a short stay to present first results and to discuss further proceeding.

Christian Reimer will return to Uppsala for a one month stay in November/ December to conduct final analysis and to prepare the publication of our results.

5) Projected publications / articles resulting or to result from the grant *(ESF must be acknowledged in publications resulting from the grantee's work in relation with the grant)*

The results presented in this report will also be shown on the Annual meeting of the German Society of Animal Breeding in Göttingen on September 5th and 6th.

We are waiting for the new data to come to decide for a publication in a peer reviewed journal.

6) Other comments (if any)

The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project p2010044.

7) Literature

Alexander, D.H., Novembre, J., Lange, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19:1655-1664.

Archibald, A.L., Bolund, L., Churcher, C., Fredholm, M., Groenen, M.A.M., Harlizius, B., Lee, K.-T., Milan, D., Rogers, J., Rothshild, M.F., Uenishi, H., Wang, J., Schook, L.B., the Swine Genome Sequencing Consortium (2010) Pig genome sequence - analysis and publication Strategy .*BMC Genomics*, 11:438.

DePristo M, Banks E, Poplin R, Garimella K, Maguire J, Hartl C, Philippakis A, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell T, Kernytsky A, Sivachenko A, Cibulskis K, Gabriel S, Altshuler D and Daly, M (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics.* 43:491-498.

Fang, X., Mu, Y., Huang, Z., Li, Y., Han, L., Zhang, Y., Feng, Y., Chen, Y., Jiang, X., Zhao, W., Sun, X., Xiong, Z., Yang, L., Liu, H., Fan, D., Mao, L., Ren, L., Liu, C., Wang, J., Li, K., Wang, G., Yang, S., Lai, L., Zhang, G., Li, Y., Wang, J., Bolund, L., Yang, H., Wang, J., Feng, S., Li, S. and Du, Y. (2012) The sequence and analysis of a Chinese pig genome. *GigaScience*, doi:10.1186/2047-217X-1-16.

Flicek, P. et al. (2013) Ensembl 2013. *Nucleic Acids Research* 2013 41 Database issue:D48-D55, doi: 10.1093/nar/gks1236

Groenen, M.G.M et al. (2012) Analysis of pig genome provide insight into porcine demography and evolution. *Nature*; doi:10.1038/nature11622.

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25, 1754-60.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25, 2078-9.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297-303.

Picard (2009-05-18) Available: http://picard.sourceforge.net/. Accessed 2013-07-26.

Rubin, C.-J., Megens, H.-J., Barrio, A.M., Maqbool, K., Sayyab, S., Schwochow, D., Wang, C., Carlborg, Ö., Jern, P., Jørgensen, C.B., Archibald, A.L., Fredholm, M., Groenen, M.A.M., Andersson, L. (2012) Strong signatures of selection in the domestic pig genome. *PNAS*, doi:10.1073/pnas.1217149109.

Simianer, H and Köhn, F (2010) Genetic management of the Göttingen Minipig population. *Journal of Pharmacological and Toxicological Methods*, 62, 3, 221-226.

Vamathevan, J., Hall, M.D., Hasan, S., Woollard, P.M., Xu, M., Yang, Y., Li, X., Wang, X., Kenny, S., Brown, J.R., Huxley-Jones, J., Lyon, J., Haselden, J., Min, J., Sanseau, P. (2013) Minipig and beagle animal model genomes aid species selection in pharmaceutical discovery and development. *Toxicology and Applied Pharmacology*, Volume 270, Issue 2, 15 July 2013, Pages 149–157.

Wang, K., Li, M., Hakonarson, H. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, 2010, Vol. 38, No. 16, doi:10.1093/nar/gkq603.

Weir, B.S. (1996) Genetic Data Analysis II. *Sinauer Associates, Inc. Publishers*, Sunderland, Massachusetts: ISBN 0-87893-90-4.