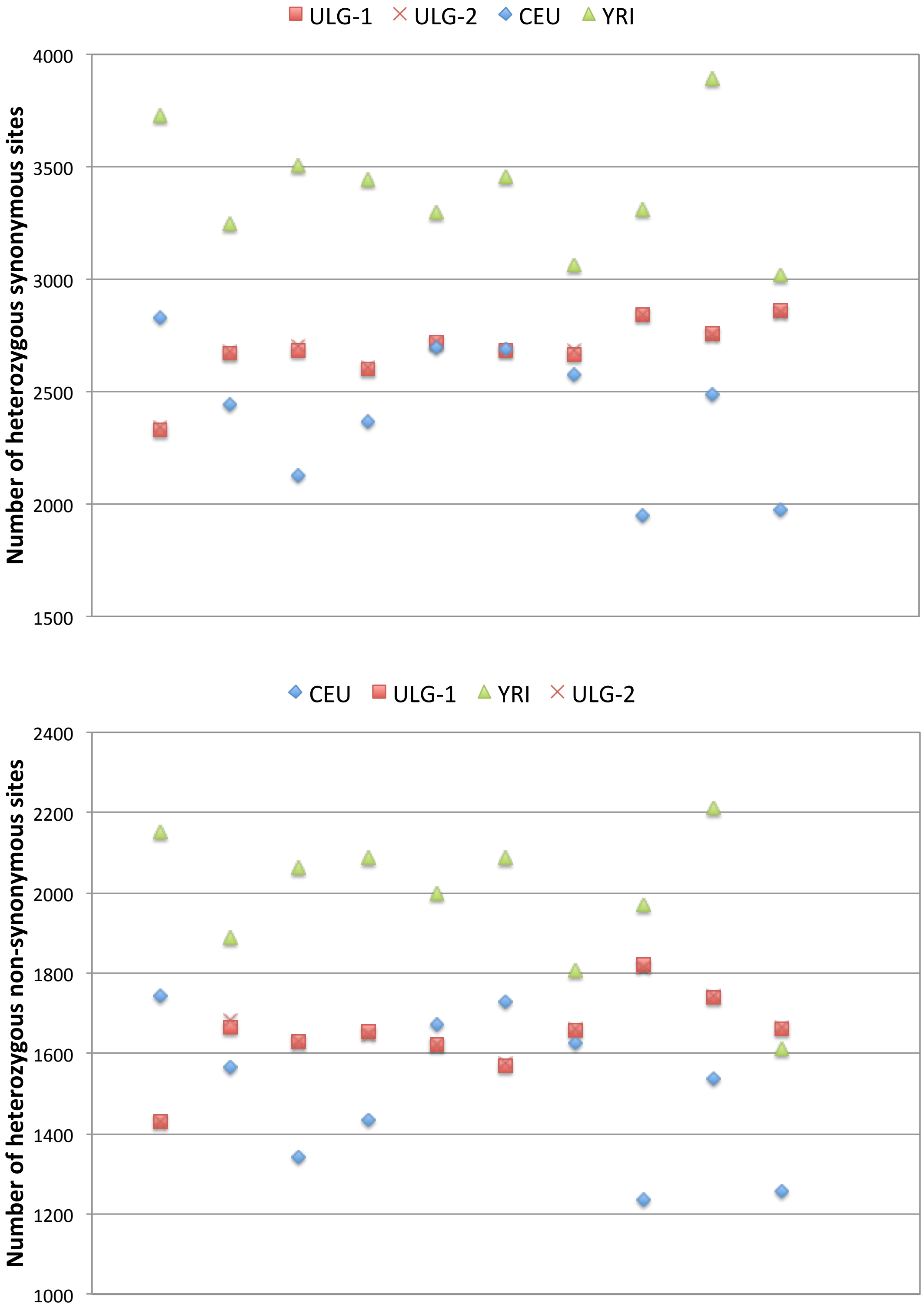
Supplemental Material S1:

Effect of dataset/sequencing center on nucleotide diversity.

To ensure that the differences in nucleotide diversity observed between the human (BAM files downloaded from the 1000 Genomes project) and bovine samples (sequenced at the University of Liège - ULg) would not be merely technical artifacts, we compared the nucleotide diversity obtained with the 1000 Genomes BAM files, with those obtained for 10 human samples sequenced at the ULg using virtually the exact same experimental conditions, i.e. chemistry, sequencers and analysis pipeline, as for the bovine samples. The ULg individuals comprised 10 unrelated individuals corresponding to members of families sampled to study a rare form of neurological cancer. They originated from Europe and South America. The sequence coverage for the ULg samples averaged 53.2 fold (range: 43.0 – 67.1). The corresponding human exomes were captured using the SureSelect Human All Exon kit (Agilent). The 1000 Genomes samples were down-sampled to 45.0 fold using GATK "downsample_to_fraction" function.

For each individual, we identified heterozygous positions using GATK and corresponding best practices. Variants were annotated using custom-made scripts and sorted into synonymous and non-synonymous variants. To compare the nucleotide diversity between populations (say A and B) while ensuring that the same exome compartment would be taken into account in the two populations, we only considered variants detected in population A if at least one individual from population B would have a genome coverage ≥ 20 at the corresponding position. Figure 1 shows the number of heterozygous synonymous (A) and non-synonymous (B) positions detected using this procedure when comparing respectively 10 CEU and 10 YRI samples with 10 ULg samples. It can be seen that very similar variant numbers were compiled for the ULg population when confronting it to either the CEU or YRI population, indicating that very similar exome compartments were explored by the CEU and YRI populations. The number of variants that were ignored in the comparisons (because not properly covered by the other population) were 4005, 5690 and 3142/3124 for the CEU, YRI and Ulg (vs CEU/YRI) population respectively. As expected, the number of synonymous and non-synonymous variants detected in individuals from the ULg population overlapped with the corresponding numbers detected in the CEU population, while being inferior to those obtained in the YRI population. Taken together, these results indicate that the observed differences between the bovine and human samples can not be explained by technical artifacts alone.



Legend: Number of synonymous (A) and non-synonymous (B) variant positions detected by exome sequencing in European CEPH samples (CEU: 10) and Yoruban samples (YRI: 10) using BAM files down-loaded from the 1000 Genomes Project (The 1000 Genomes Project; http://www.1000genomes.org/), and in European-ancestry samples sequenced at the University of Liège (ULG-1: comparison CEU vs ULg; ULG-2: comparison YRI vs ULg: 10).