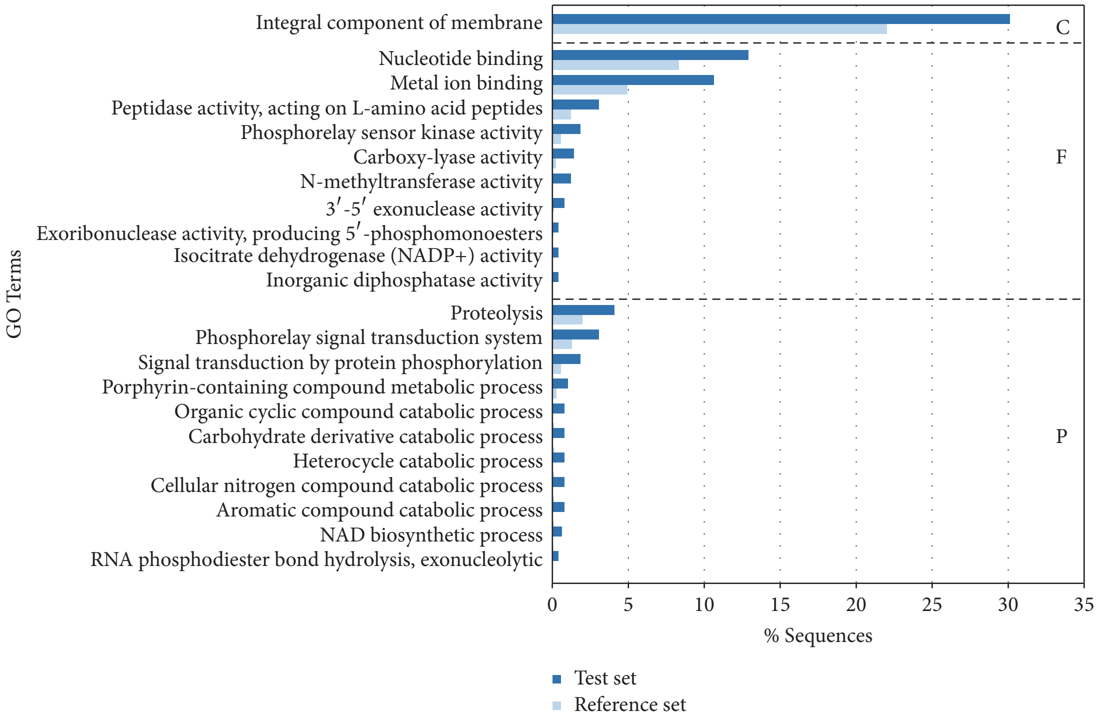

Figure 3: *GO enrichment analysis for the subtype A genes accumulating high-impact mutations*. The analysis was conducted with the genes from the type A specimens having at least one mutation of predicted high impact (test set) and their corresponding orthologs from strain JB197 (reference set). Plot shows GO terms found to be overrepresented with a p value below 0.05, ordered by percent sequences in descending order and grouped by GO class (C: cellular components, F: molecular functions, and P: biological processes).

expected, this resulted in a larger number of variants, including both SNVs and indels (Table S5). Of all the 3,553 genes annotated in the genome of strain JB197, 788 (22%) appear to have an ortholog in at least one of the subtype A specimens with a minimum of one mutation of impact predicted to be high or moderate. When considering only missense mutations altering the chemical nature of the underlying residue, the number of genes with variants was reduced to 483 (Table S6), excluding previously discussed pseudogenes. We found no evidence of positive selection in genes exhibiting variants when considering only *L. borgpetersenii* orthologs.

We also looked for gene ontology (GO) terms enriched in the selected genes when compared to all the genes in the reference genome. The refined enrichment analysis resulted in 22 enriched terms, 10 representing molecular functions, 11 representing biological processes, and one representing a cellular component (Figure 3; Table S7). The only cellular component found to be enriched was that associated with integral membrane proteins. This was also the term encompassing the largest number of proteins in the set, including 147 genes encoding a wide variety of membrane transporters, chemotaxis proteins, and peptidases. Several terms in the categories of molecular functions and biological processes, such as phosphorelay sensor kinase activity, phosphorelay signal transduction systems, and signal transduction by protein phosphorylation, can also be associated with TCSs and other regulatory proteins.

Due to the relatively high variability observed in genes encoding proteins involved in transcriptional regulation, such as members of TCSs, we used the P2RP server to assess the number of genes in the reference genomes of both types putatively encoding transcriptional regulators. Results showed that for TCSs, on average, there are at least 36 loci encoding the histidine kinase component and 26 loci encoding response regulators per genome (Table S8). Of the 62 genes predicted for the subtype A reference genome, 14 are within those found to have variants of predicted high impact and selected for the enrichment analysis described before, most of them encoding sensor histidine kinases. The server predicted nearly 80 loci encoding additional transcription factors per genome (Table S9), including sigma factors of RNA polymerase and other types of proteins with DNA-binding domains. Likewise, 10 of these genes are within those previously found to have at least one high-impact variant.

3.4. Structural Rearrangements and Copy Number Variations (CNVs). In addition to SNVs and small indels, we found relatively few structural variants between subtypes. Genomic regions that could be assembled de novo for our five specimens show high synteny with the corresponding reference genomes, thus maintaining the pattern of large-scale rearrangements previously described for the two reference genomes [17]. Two notable exceptions are the transposition of a segment of 24 kb in chromosome I of LBH-A and another one of ~40 kb from chromosome II of both LBH-A and LBH-B (Figure 1(c)). However, gene content and synteny within these segments are not globally altered by their putative rearrangement.