**Priors based on conservation would lead to an improvement of GP accuracy?**

The main concern is: The majority of functional elements are non-coding (NC) and these are not well conserved [3].

[3]: “Our finding supports those studies that have shown that conservation is a poor predictor of regulatory function [[46](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4224693/" \l "CR46)] and is consistent with findings of extensive regulatory gain and loss between lineages, indicating that there is variation in regulatory element positions across evolution”

[1]: “Regulatory elements are known to have much higher evolutionary turnover[7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5015703/" \l "R7) implying that conservation is a less important signal when interpreting variants in regulatory regions”

[2](June 2016): “enhancers have a short life and they appear to be species specific ([33](http://hmg.oxfordjournals.org/content/25/R2/R190.full" \l "ref-33)). Thus prioritizing or over-emphasizing conservation when analyzing putative enhancers is likely misleading.” What if we do comparative genomics bt breeds to capture these? Same authors point to: Key Regulatory Factors: Histones and Chromatin. Our conservation score will be compatible with the other annotations available. A Histones and Chromatin approach would be more complex than ours.

* The statement that NC are not enough-conserved is debatable:

“promoter’s half-life is comparable to that of protein coding regions’.[4]

“Functional studies demonstrated that some ultraconserved regions had a role as enhancers ([Woolfe et al. 2005](http://genome.cshlp.org/content/23/7/1063.full" \l "ref-46)).”[5]

“Evolutionary approaches infer evidence of function through both sequence conservation as well as rapidly accelerated change and have had success in identifying putative enhancers and motifs ([24](http://hmg.oxfordjournals.org/content/25/R2/R190.full" \l "ref-24),[25](http://hmg.oxfordjournals.org/content/25/R2/R190.full" \l "ref-25)).“[2]

Our estimate for the total amount of nonexonic constraint experienced by this locus is roughly twice that for exonic constraint. This small fraction of the genome includes most known protein-coding exons and the majority of known transcriptional regulatory elements ([Rat Genome Sequencing Project Consortium 2004)…](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1172034/" \l "ref53)(and keep reading). Nonexonic elements constitute a majority of the constrained 5%–6% of the human genome [8]

Exonic regions show a clear enrichment for larger, higher scoring elements, with over half of the exonic elements being larger than 18 bp, compared with only 20% of the nonexonic elements[8]

Even if NC are not enough-conserved:

The conservation approach will also allow to capture marks of accelerated evolution (“novel conserved NC”). “Elements that are highly conserved in vertebrates sometimes show accelerated evolution(i.e. human accelerated regions (HARs) (these are mainly regulatory elements)”. “These novel conserved NC elements first arose for transcription factors and developmental genes ([Lowe et al. 2011](http://genome.cshlp.org/content/23/7/1063.full" \l "ref-26))”. [5]. Developmental, important in livestock

“Constrained elements tend to cluster,”[6] so perhaps we can capture some of the not-well-conserved NC regions with clusters approach. In fact, >25%...[8]. The ST7 gene, for example, has multiple isoforms annotated as RefSeq entries and appears to harbor a large amount of nonexonic constraint immediately flanking its coding exons . Also “Constrained elements tend to cluster” “...400 of which are exclusively nonexonic”

We conclude that there is a substantial fraction of mammalian-specific noncoding constrained elements whose importance is equal to or exceeds that of pan-vertebrate elements.[8]

regulatory elements are much harder to model: This is where comparative genomics shines.[5]

Comparison of human, mouse, rat, and dog identified several hundreds of thousands of conserved noncoding elements (CNEs) that cluster near developmental genes

“GERP++ is able to better annotate longer NC [7](2010).

* Some important arguments in favor of conservation:

- “In livestock there is rapid phenotypic evolution as a response to strong directional selection” [5]. This affects the selected elements but also others that are linked. So, it is expected that there is some mobility from conserved regions to non-conserved and vice-versa. This will be captured with our approach.

-“Sequence conservation has more specificity than assays such as transcription factor binding, or even transcription,” “sequence conservation is extremely useful for prioritizing GWAS SNPs” “RNA-seq, ChiP-seq data for transcription factors, methylation or histone marks, and DNase hypersensitivity… are very useful and are often complementary to sequence conservation analysis.”[5]

-“Existing computational approaches to predicting the effect of a coding variant on protein function, such as SIFT[5](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5015703/" \l "R5) and PolyPhen[6](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5015703/" \l "R6), are largely based on quantifying constraint on the affected residue from a multiple sequence alignment.”[1]. Actually, most current tools focus on conservation-based features of variants.

-A large proportion of the human genome is under evolution-constrained[5]. Also in *sus-crofas*? “many more constrained elements remain to be pinpointed.”[1]

-“Over 94% of the coding exons in the human genome overlap at least one predicted CE”[7]

- There are a number of methods used to assess variant function that solely rely on conservation (Table two from Cooper & Shendure, [[4](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4224693/" \l "CR4)]) and others have shown that conservation can be used to discriminate functional regulatory variants from background variants [[20](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4224693/" \l "CR20)]. [3]

- The other annotations will also increase, so it is expected that accurate biological priors will play a key role in improving GP

- Genomic selection, is currently restricted to dominant and additive variants and to variants found at relatively high frequencies in the population. Conservation may point at some low-frecuency alleles

- recent completion of the genome sequences of a large number of related species may help in the establishment of sus-crofas specific conservations scores.

-Conservation study can complemented with a study of missing heterogeneity. i.e. loss of heterocigosity in high-genetic performance farms with respect to low-production farms.

1. Graham et al., 2014. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5015703/>

2. Coppola et al., 2016-June. <http://hmg.oxfordjournals.org/content/25/R2/R190.full>

3. Niamh 2014. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4224693/>

4. Villar et al., 2015. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4313353/>

5. Afoldi et al., 2013. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3698499/

6. Cooper et alt., 2005. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1172034/>

7. Davidov 2010. <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1001025>

8. Coopere et al.m 2005. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1172034/