***D. Approach***

**D-1 Approach Aim 1 - Convert & extend the FunSeq somatic variant pipeline for germline prioritization**

**D-1-a Preliminary results for Aim 1**

**D-1-a-i We have experience in annotating non-coding regions of the genome, including both TF-binding sites and non-coding RNAs**

Our proposed work is based on our past experience in non-coding annotation, as part of our 10-year history with the ENCODE and modENCODE projects. Our TF work includes the development of methods to define the binding peaks of TFs[1], prediction of a TF’s target genes[2], and new machine learning techniques[3]. Furthermore, we developed methods that integrate ChIP-seq, chromatin, conservation, sequence and gene annotation data to identify gene-distal enhancers[4], which we have partially validated[5]. We also constructed linear and non-linear models that utilize TF binding and histone modification signals to accurately predict the transcriptional output of a gene in different cell types of several organisms including yeast, worm, fly, and human. [6-10] We have also constructed regulatory networks for human and model organisms[11, 12], and completed many analyses on them (Fig 1)

[5, 7, 11, 13-26]. Furthermore, we conducted large-scale multi-organism regulatory and coexpression network comparisons, along with transcriptome and pseudogene lineage analyses[26-30]. We also have extensive experience conducting integrated analyses of RNA-Seq datasets generated by the ENCODE, modENCODE, BrainSpan and exRNA consortia[7, 31-34]. In particular, we developed RSEQtools and IQseq for gene model creation and transcript quantification[35, 36]. We also developed tools that specifically analyze features of ncRNAs, including incRNA and ncVAR for finding and characterizing these elements[37, 38].

[[ANS: Need better transition]]

**D-1-a-ii We have experience in allelic analyses**

A specific class of regulatory variants is one that is related to allele-specific events. These are variants that are associated with allele-specific binding (ASB), particularly of transcription factors or DNA-binding proteins, and allele-specific expression (ASE)[39, 40]. We have previously developed a tool, AlleleSeq,[24] for the detection of candidate variants associated with ASB and ASE. Using this we have generated comprehensive lists of allelic variants for ENCODE and 1000 Genomes and found that allelic variants are under differential selection from non-allelic ones[11, 22, 32]. By constructing regulatory networks based on ASB of TFs and ASE of their target genes, we further revealed substantial coordination between allele-specific binding and expression[11]. Furthermore, we have constructed a personal diploid genome and transcriptome of NA12878 on[41].

**D-1-a-iii Experience in relating annotation to variation: the FunSeq pipeline**

We have extensively analyzed patterns of variation in non-coding regions, along with their coding targets[5, 11, 38]. We used metrics, such as diversity and fraction of rare variants, to characterize selection on various classes and subclasses of functional annotations[38]. In addition, we have also defined variants that are disruptive to a TF-binding motif in a regulatory region[31]. Further studies showed relationships between selection and protein network topology (for instance, quantifying selection in hubs relative to proteins on the network periphery[21, 23]).

In recent studies[22, 42], we have integrated and extended these methods to develop a prioritization pipeline called FunSeq (Fig 2). It identifies sensitive and ultra-sensitive regions (i.e., those annotations under strong selective pressure, as determined using genomes from many individuals from diverse populations). FunSeq links each non-coding mutation to target genes, and prioritizes such variants based on scaled network connectivity. It identifies deleterious variants in many non-coding functional elements, including TF binding sites, enhancer elements, and regions of open chromatin corresponding to DNase I hypersensitive sites. It also detects their disruptiveness in TF binding sites (both loss-of and gain-of function events). Integrating large-scale data from various resources (including ENCODE and The 1000 Genomes Project) with cancer genomics data, our method is able to prioritize the known TERT promoter driver mutations, and it scores somatic recurrent mutations higher than those that are non-recurrent. Using FunSeq, we identified ~100 non-coding candidate drivers in ~90 WGS medulloblastoma, breast and prostate cancer samples [22]. Drawing on this experience, we are currently co-leading the ICGC PCAWG-2 (analysis of mutations in regulatory regions) group.

**D-1-b  Research plan for Aim 1**

We plan to convert and extend the current FunSeq prototype from its focus on somatic variants to allow the identification of germline variants associated with large gene expression changes (Fig 3). Our new approach called ReEnAct (Regulation of Enhancer Activity) will iteratively create a model of high impact variant.  It will have several features tailoring it to germline analysis, including 1) identifying functional sites among the conserved regions of the human genome and ncRNA regulatory elements; 2) investigating the allelic elements. We will iteratively train and test our model on the results of the Stro-seq experiments to refine its parameters.

[[LS: Remove the discussion of rare variants. Now we’re just looking at variants]]

**D-1-b-i Consistently prioritizing non-coding elements from polymorphism data**

In order to define rare variants with highly impactful events, we will use both intra-human variation data (from The 1000 Genomes Project) as well as cross-species evolutionary conservation (using classical measures such as GERP score[43]).

Due to the development of a number of massively parallel assays for identifying regulatory regions in the genomes, we have been able to identify the epigenetic signatures underpinning active enhancers.  We will use this information to make  better enhancer predictions and utilize information provided by the Epigenome Roadmap [44-46], and more recently from ENCODE projects. In particular, we will develop a new machine learning framework that combines pattern recognition within the signal of various epigenomic features and transcription of enhancer RNA (eRNA) with sequence-based features to predict active enhancers across different brain regions and other tissues in the Epigenome Roadmap project. – [[ANS: change to whatever cells HY is using. This paragraph may need a transition]]

We will first update the TF binding non-coding elements from the original FunSeq approach. Here, we will use the better enhancer definition provided by the Epigenome Roadmap [44-46], and more recently from ENCODE. In particular, we will develop a new machine learning framework that utilizes pattern recognition within the signal of various epigenomic features and transcription of enhancer RNA (eRNA) to predict active enhancers across different tissues.

[[deleted ncRNA stuff]]

**D-1-b-ii Identifying high-impact mutations: breaking & creating motifs**

For impactful events at TF binding sites, we will use motif breakers and formers to define loss-of- and gain-of-function events, respectively, as these events are more likely to have deleterious consequences[22, 38, 47-51]. Variants altering the position-weight matrix (PWM) scores for TF binding sites could potentially either decrease (loss-of-function) or increase (gain-of-function) the binding strength of TFs. A key improvement that we plan to utilize is to employ ancestral alleles to get a more accurate determination of these events.

[[deleted ncRNA stuff]]

**D-1-b-iii Variant prioritization based on allelic activity** [[SKL may add a paragraph about his new method on allelic activity]]

Allele-specific variants potentially provide a most direct readout of the functional impact of a variant. For example, if we can associate the differential binding effect of a particular transcription factor with different alleles of an SNV, then we can identify loci that have potential functional impacts in regulation. However, because allelic variants are enriched for rare variants[52], it will be difficult to match the specific variants in a personal genome of interest to prioritize against those earlier determined to be allelic in a functional genomics experiment on a cell line. Hence, instead of prioritizing by the direct overlap of allelic variants, we need to prioritize by the presence of allelic variants within 'allelic elements', or allelic regions in the genome (Fig 4).

We derive allelic elements by first identifying allelic variants from hundreds of individuals. These individuals will be amassed from The 1000 Genomes Project[53]. We will match them with their corresponding RNA-Seq and ChIP-seq experiments from multiple disparate studies, such as gEUVADIS[52] and ENCODE[31]. Because these separate studies typically have various inconsistencies in terms of tools and parameters used in processing their data, we have to reprocess and harmonize the heterogeneous data and detect allelic variants in a uniform fashion. Also, while the conventional way to detect allelic variants is using the binomial test, previous studies have found that the distributions of the allelic ratios in ChIP-seq and RNA-seq experiments have been empirically observed to give a broader, or an ‘overdispersed’, distribution than a binomial distribution[54-56]. To identify and remove problematic "outlier" datasets and to account for overdispersion of read distributions, we will extend our detection pipeline (AlleleSeq) to include the calculation of an overdispersion parameter for each ChIP-seq and RNA-seq dataset; the beta-binomial test (which parametrizes the overdispersion) will be used to detect allelic variants instead of the binomial test.

Subsequently, allelic variants (rare and common) identified across hundreds of genomes can be aggregated into ‘allelic genomic elements’. Each element will be assigned an ‘allelicity’ score based on not only its enrichment of allelic variants within the element (in comparison to accessible variants within the elements and having sufficient coverage to make an allelic activity call), but also across the number of individuals having allelic variants in a consistent allelic direction. The scoring system by element is useful in two ways: (1) it allows continuous ranking of genomic elements based on its allelic impact across multiple individuals (as opposed to defining a threshold to make a binary decision of whether an element is ‘allelic’) and (2) it enables incorporation of ASE and ASB into the main prioritization scheme; input variants (even those which are rare, but lie in highly-ranked allelic genomic elements) will be upweighted according to their scores.

**D-1-b-iv Identifying likely target genes for distal regulatory elements & assessing the impact of variants on network connectivity**

To interpret the likely functional consequences of non-coding variants, we will comprehensively define associations between many non-coding regulatory elements and their target protein-coding genes. The correlation between enhancer and promoter activity across the ENCODE cell-lines and different tissues will be used to identify significant associations between regulatory elements and candidate target genes, as done by Yip et al[4]. A single regulatory variant may affect the expression of multiple genes, either because it directly regulates multiple genes or because the target gene is itself a regulatory factor.

We further developed an ENGINE(Enhancer Gene Interaction detection) method to detect enhancer gene linkage. ENGINE will integrate both static and dynamic genomic information. The sequence features are denoted by K-mer profile and co-occurrence matrix, while the dynamic features include DNase I,  histone modification and TF binding information. We convert histone mark/DNase signals to a pseudo image and extract informative features that can tell positive from negative datasets. Meanwhile, gene expression variance as an addition information information,is also integrated together into a statistical model to predict enhancer gene linkage.

We will use the regulatory element-target gene pairs to connect the non-coding variants into a variety of networks -- e.g. regulatory network, metabolic pathways, etc. We will examine their network centralities, such as hubs, bottlenecks and hierarchies, as we know that disruption of highly connected genes or their regulatory elements is more likely to be deleterious[21, 23].

[[Deleted RNA stuff]]

[[SKL to add something about ENGINE]]

**D-1-b-v We will use a unified weighted scoring scheme for combining all ReEnAct features to prioritize variants**

To integrate the various features mentioned above, we plan to elaborate the weighting system in FunSeq.[42]. Constrained by selective pressure, common variations tend to arise in functionally unimportant regions. Thus, features that are enriched with common polymorphisms are less likely to contribute to the deleteriousness of variants and are weighted less. In general, features can be classified into two classes: discrete (e.g., within or outside of a given functional annotation) and continuous (e.g., the PWM change in ‘motif-breaking’). We will weigh these two sets of features with different strategies.

For each discrete feature , such as sentitive region overlap, ultraconserve region overlap, and HOT region overlap, we calculate the probability that it overlaps with common polymorphisms. We then calculate the information content to denote the value of discret features + , where and can be used for score optimization.

The situation is more complex for continuous features, as different feature values have different probabilities of being observed in natural polymorphisms. Thus, one weight cannot suffice for varied feature values. For a continuous feature , such as motif gain, motif break and GERP etc, which is associated with a value , the probability is firstly estimated using common variants: . The score of continuous feature is defined as . We then fit a smoothing curve and estimate parameters ’s according to empirical distribution .

The eleVAR score (eS) is calculated by summing up the values of all its features. We will also consider the feature dependency structure when calculating the scores (e.g., removing redundant features or performing dimension reduction techniques).

**D-2-b-i Statistical framework for parameter tuning using Bayesian updates**

[[Very important, need to update. Will have many rounds of testing/training.]]

The initial feature parameter () (given number of features) assigned in D-1-b-v will be further optimized with newly available “gold standard” datasets. We plan to tune these parameters using an incremental Bayesian learning strategy. For a variant , given eleVAR score (equation 3 in D-1-b-v), the probability that is functional ( designates a positive result, whereas denotes a negative result) follows a logistic function ( are scaling parameters). To update with training data , we implement Bayes’ rule:.

The likelihood ratio is defined as: , and then MCMC (Monte Carlo Markov Chain) will be used to find the most probable . The updated will then be used as tuned parameters in eleVAR to prioritize variants. The procedure will be iterated in several rounds. In the first round of tuning, feature weights obtained in D-1-b-v will be used to construct priors . In subsequent rounds, the updated weights will be set as new priors.

**D-2-b-iii Generating an initial list of prioritized variants & then running them through eleVAR**

We’ll get variants from 1000G Phase 3. [[MG: remove to PCAWG]][[SKL: to cut this part?

**D-2-b-iv Round 1 of tuning based on publicly available datasets**

To perform the initial round of performance assessment and parameter tuning, we plan to use publicly available datasets from various resources. These datasets include known disease-causing mutations from molecular studies, high throughput reporter assays on enhancer activities .

The Human Gene Mutation Database (HGMD)[57] and ClinVar[58] catalogue large numbers of regulatory disease-causing mutations discovered in molecular studies. Several high-throughput technologies have also been developed to test the phenotypic impacts of non-coding genomic variants. For example, Kwasnieski et al used CRE-seq[59] to assay over 1,000 single- and double-nucleotide mutations in promoter regions. Kheradpour et al.[47] used MPRA to test variants affecting regulatory motifs in over 2,000 human enhancers. We will utilize these datasets to perform comparisons with other variant prioritization methods, such as CADD [60], to obtain a preliminary evaluation of method performance. We will then tune our parameters using the scheme described above.

**D-2-b-v Round 2 [[How many rounds of tuning?]] of tuning using high-throughput experiments done in this project**

  - gtechr01 disc.

        1. We talked about change it to 4 aims, 10K mutation first year, then 5000 mutations for Yrs 2 and 3. 10 CRISPRs per year for Yrs 2, 3, and 4. However, I also talked to John this afternoon, he wants to keep the CRISPR part in aim 2 to avoid this grant to be too similar to the one we just submitted. I think it can be a good idea because this RFA is focused on technology development and the CRISPR experiments are really just validation, so maybe it is good not to have it as a separate aim.

[[[Try to get all done by late tomorrow]]

[[/SKL: add in references w SK]]

**[[SKL try in conversation w JZ, ANS & T G to update the below as best you can taking into account the above]]**

[[MRS to read & to send another cover to HY]]

We expect an average of ~3M germline variants per genome[53]. ~~Since they rarely recur at the exact same position, we anticipate a prioritized list of ~8M variants (=40K \* 250 genomes, based on the the expected size of the prostate compendium).~~ We will select 3000 functional regions of appreciable size that contain highly ranked variants. Assuming ~8M variants are distributed evenly across the human genome, taking an average element size of 3kb, the number of variants per element will be ~4. Variants on the same element are expected to have different functional impacts. For each element, we will prioritize at least one of these variants to be of high impact, and the remaining variants to be of a lower impact. Specifically, we will have a total of 15K variants in the year1 and 5000 in the subsequence years (two fifth with a high impact and three fifth with a low impact). Subsequent tuning and refinement of the eleVAR parameters will be based on further experimental characterization of these 15K variants in first year (6K highly prioritized and 9K lowly, respectively). We will validate these variants through functional genomic screens using the [[change cloneseq]] [[SKL:Done]]STRO-seq technology coupled with luciferase reporter assays. Overall, this refinement will be accomplished in two rounds, each round per year, as detailed in Aim 3 and the timeline (Fig 6). Finally, during the last year of the proposed work, we will perform a careful assessment of our model. We will again prioritize our full list of variants and select a final set of 5000 top ranked variants for an unbiased validation. This will allow us to construct a precise ROC curve in order to measure the accuracy of our predictions.

1. Rozowsky, J., et al., *PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls.* Nat Biotechnol, 2009. **27**(1): p. 66-75.

2. Cheng, C., R. Min, and M. Gerstein, *TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles.* Bioinformatics, 2011. **27**(23): p. 3221-7.

3. Yip, K.Y. and M. Gerstein, *Training set expansion: an approach to improving the reconstruction of biological networks from limited and uneven reliable interactions.* Bioinformatics, 2009. **25**(2): p. 243-50.

4. Yip, K.Y., et al., *Improved reconstruction of in silico gene regulatory networks by integrating knockout and perturbation data.* PLoS One, 2010. **5**(1): p. e8121.

5. Yip, K.Y., et al., *Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors.* Genome Biol, 2012. **13**(9): p. R48.

6. Cheng, C., et al., *Genome-wide analysis of chromatin features identifies histone modification sensitive and insensitive yeast transcription factors.* Genome Biol, 2011. **12**(11): p. R111.

7. Gerstein, M.B., et al., *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.* Science, 2010. **330**(6012): p. 1775-87.

8. Cheng, C., et al., *A statistical framework for modeling gene expression using chromatin features and application to modENCODE datasets.* Genome Biol, 2011. **12**(2): p. R15.

9. Cheng, C. and M. Gerstein, *Modeling the relative relationship of transcription factor binding and histone modifications to gene expression levels in mouse embryonic stem cells.* Nucleic Acids Res, 2012. **40**(2): p. 553-68.

10. Cheng, C., et al., *Understanding transcriptional regulation by integrative analysis of transcription factor binding data.* Genome Res, 2012. **22**(9): p. 1658-67.

11. Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**(7414): p. 91-100.

12. Nègre, N., et al., *A cis-regulatory map of the Drosophila genome.* Nature, 2011. **471**(7339): p. 527-31.

13. Cheng, C., et al., *Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data.* PLoS Comput Biol, 2011. **7**(11): p. e1002190.

14. Yan, K.-K., et al., *Comparing genomes to computer operating systems in terms of the topology and evolution of their regulatory control networks.* Proc Natl Acad Sci U S A, 2010. **107**(20): p. 9186-91.

15. Yu, H., et al., *Genomic analysis of essentiality within protein networks.* Trends Genet, 2004. **20**(6): p. 227-31.

16. Yu, H., et al., *TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics.* Nucleic Acids Res, 2004. **32**(1): p. 328-37.

17. Yu, H., et al., *The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics.* PLoS Comput Biol, 2007. **3**(4): p. e59.

18. Luscombe, N.M., et al., *Genomic analysis of regulatory network dynamics reveals large topological changes.* Nature, 2004. **431**(7006): p. 308-12.

19. Gianoulis, T.A., et al., *Quantifying environmental adaptation of metabolic pathways in metagenomics.* Proc Natl Acad Sci U S A, 2009. **106**(5): p. 1374-9.

20. Yu, H., et al., *Predicting interactions in protein networks by completing defective cliques.* Bioinformatics, 2006. **22**(7): p. 823-9.

21. Kim, P.M., J.O. Korbel, and M.B. Gerstein, *Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context.* Proc Natl Acad Sci U S A, 2007. **104**(51): p. 20274-9.

22. Khurana, E., et al., *Integrative annotation of variants from 1092 humans: application to cancer genomics.* Science, 2013. **342**(6154): p. 1235587.

23. Khurana, E., et al., *Interpretation of genomic variants using a unified biological network approach.* PLoS Comput Biol, 2013. **9**(3): p. e1002886.

24. Rozowsky, J., et al., *AlleleSeq: analysis of allele-specific expression and binding in a network framework.* Mol Syst Biol, 2011. **7**: p. 522.

25. Lörcher, U., J. Peters, and J. Kollath, *[Changes in the lungs and pleura following chemoembolization of liver tumors with mitomycin-lipiodol].* Rofo, 1990. **152**(5): p. 569-73.

26. Shou, C., et al., *Measuring the evolutionary rewiring of biological networks.* PLoS Comput Biol, 2011. **7**(1): p. e1001050.

27. Boyle, A.P., et al., *Comparative analysis of regulatory information and circuits across distant species.* Nature, 2014. **512**(7515): p. 453-6.

28. Gerstein, M.B., et al., *Comparative analysis of the transcriptome across distant species.* Nature, 2014. **512**(7515): p. 445-8.

29. Sisu, C., et al., *Comparative analysis of pseudogenes across three phyla.* Proc Natl Acad Sci U S A, 2014. **111**(37): p. 13361-6.

30. Yan, K.-K., et al., *OrthoClust: an orthology-based network framework for clustering data across multiple species.* Genome Biol, 2014. **15**(8): p. R100.

31. ENCODE Project Consortium, *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

32. Djebali, S., et al., *Landscape of transcription in human cells.* Nature, 2012. **489**(7414): p. 101-8.

33. <http://brainspan.org>, *Last accessed on 21st May 2015.*

34. <http://exRNA.org>, *Last accessed on 21st May 2015th January 2014.*

35. Habegger, L., et al., *RSEQtools: a modular framework to analyze RNA-Seq data using compact, anonymized data summaries.* Bioinformatics, 2011. **27**(2): p. 281-3.

36. Du, J., et al., *IQSeq: integrated isoform quantification analysis based on next-generation sequencing.* PLoS One, 2012. **7**(1): p. e29175.

37. Lu, Z.J., et al., *Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data.* Genome Res, 2011. **21**(2): p. 276-85.

38. Mu, X.J., et al., *Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project.* Nucleic Acids Res, 2011. **39**(16): p. 7058-76.

39. Pastinen, T., *Genome-wide allele-specific analysis: insights into regulatory variation.* Nat Rev Genet, 2010. **11**(8): p. 533-8.

40. Birney, E., et al., *Allele-specific and heritable chromatin signatures in humans.* Hum Mol Genet, 2010. **19**(R2): p. R204-9.

41. <http://alleleseq.gersteinlab.org>, *Last accessed on 21st May 2015.*

42. Fu, Y., et al., *FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer.* Genome Biol, 2014. **15**(10): p. 480.

43. Cooper, G.M., et al., *Distribution and intensity of constraint in mammalian genomic sequence.* Genome Res, 2005. **15**(7): p. 901-13.

44. Roadmap Epigenomics Consortium, et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.

45. Ziller, M.J., et al., *Dissecting neural differentiation regulatory networks through epigenetic footprinting.* Nature, 2015. **518**(7539): p. 355-9.

46. Leung, D., et al., *Integrative analysis of haplotype-resolved epigenomes across human tissues.* Nature, 2015. **518**(7539): p. 350-4.

47. Kheradpour, P., et al., *Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay.* Genome Res, 2013. **23**(5): p. 800-11.

48. Horn, S., et al., *TERT promoter mutations in familial and sporadic melanoma.* Science, 2013. **339**(6122): p. 959-61.

49. Huang, F.W., et al., *Highly recurrent TERT promoter mutations in human melanoma.* Science, 2013. **339**(6122): p. 957-9.

50. Killela, P.J., et al., *TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal.* Proc Natl Acad Sci U S A, 2013. **110**(15): p. 6021-6.

51. Vinagre, J., et al., *Frequency of TERT promoter mutations in human cancers.* Nat Commun, 2013. **4**: p. 2185.

52. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-11.

53. 1000 Genomes Project Consortium, et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.

54. Meyer, C.A. and X.S. Liu, *Identifying and mitigating bias in next-generation sequencing methods for chromatin biology.* Nat Rev Genet, 2014. **15**(11): p. 709-21.

55. Srivastava, S. and L. Chen, *A two-parameter generalized Poisson model to improve the analysis of RNA-seq data.* Nucleic Acids Res, 2010. **38**(17): p. e170.

56. Diaz, A., et al., *Normalization, bias correction, and peak calling for ChIP-seq.* Stat Appl Genet Mol Biol, 2012. **11**(3): p. Article 9.

57. Stenson, P.D., et al., *Human Gene Mutation Database (HGMD): 2003 update.* Hum Mutat, 2003. **21**(6): p. 577-81.

58. Landrum, M.J., et al., *ClinVar: public archive of relationships among sequence variation and human phenotype.* Nucleic Acids Res, 2014. **42**(Database issue): p. D980-5.

59. Kwasnieski, J.C., et al., *Complex effects of nucleotide variants in a mammalian cis-regulatory element.* Proc Natl Acad Sci U S A, 2012. **109**(47): p. 19498-503.

60. Kircher, M., et al., *A general framework for estimating the relative pathogenicity of human genetic variants.* Nat Genet, 2014. **46**(3): p. 310-5.