**INTRODUCTION**

Expression2Kinases (X2K) is a computational pipeline that takes as input lists of differentially expressed genes, it then performs enrichment analysis to prioritize transcription factors that most likely regulate the observed changes in expression (1). Next, known protein-protein interactions are used to connect the identified transcription factors to form a subnetwork. Finally, kinases enrichment analysis is performed to prioritize protein kinases known to phosphorylate substrates within the subnetwork of transcription factors and intermediate proteins. X2K is available as a desktop, web, and command line tools. The X2K pipeline is made of these three components (Fig. 1):

[1] *Transcription factor (TF)-Enrichment*: ChIP Enrichment Analysis (ChEA) (2) was used to rank transcription factors based on TF targets that are enriched in the input gene list.

[2] *Protein-protein Interaction (PPI) network Construction:* Genes2Networks (G2N) (3) was used to generate protein-protein interactions (PPI) subnetworks given a list of TFs.

[3] *Kinase-Enrichment*: Kinase Enrichment Analysis (KEA) (4) was used to rank protein kinases that are known to phosphorylate the proteins within the subnetwork identified by G2N.

Each step in the X2K pipeline has a number of modifiable parameters. However, there are far too many unique combinations of parameters to manually identify the optimal parameters settings to accurately predict kinase targets. We therefore developed a Genetic Algorithm (GA) to search the fitness landscape of parameter combinations in order to efficiently identify an optimal set of parameters for the X2K pipeline.

**MATERIALS & METHODS**

**GA DESCRIPTION**

The GA operates by producing a population of random binary strings that map onto modifiable parameters in each step of X2K, runs the X2K pipeline each string’s corresponding parameter combination to assess fitness (in terms of accuracy of predicting kinases), crossover the fittest individuals to create a new population, introduces random mutations into that population (to avoid the GA peaking at a sub-optimal solution), and repeating until population fitness anneals at an optimum (Fig. 2).

The following GA settings were used: 10 generations; 100 individuals/generation;3 crossover points*,* random mutation rate of 0.01; top 10 fittest individuals of each generation were bred iteratively to produce the next population and were then included within that new population.

**X2K PIPELINE PARAMETERS**

[1] TF-Enrichment:

*TF Sort*: Three mutually-exclusive sort methods could be used to sort significantly enriched TFs, including, *P-value*: sort by raw p-value from Fisher’s exact test, *Rank*: the rank by p-value minus the standard deviation, *Odds ratio*: log(p-value) x (observed overlap/the expected overlap = Odds Ratio), *Combined Score*: log(p-value) x z-score.

*TF Databases*: Three options were available for selection ChIP-seq experiment-derived TF databases, including the ChEA 2015 database (2), the combined TRANSFAC databases (5), or both ChEA and TRANSFAC.

*TF Species:* Three mutually-exclusive options to include different subsets of the TF Databases by species, including human, mouse, or both.

*Top TFs*: Three mutually exclusive options for selection the number of top-enriched TFs to be used in the PPI-construction step (5, 10 or 20).

[2] PPI-Construction:

*PPI Databases*: Eighteen non-mutually exclusive PPI databases were available, including BIND (6), BIOCARTA pathway database (www.biocarta.com), the Database of Interacting Proteins (DIP) (7), Figeys PPIs from mass-spectrometry data (8), the Human Protein Reference Database (HPRD) PPIs from mass-spectrometry data (9), InnateDB multispecies experimentally validated molecular interactions (10), IntAct molecular interaction database derived from curated literature search and experimental data (11), the Kyoto Encyclopedia of Genes and Genomes (KEGG) (12), the Molecular INTeraction Database (MINT) (13), the Munich Information Center for Protein Sequences (MIPS) highly curated database of PPIs from the literature (14), Murphy (Unsure which publication), PDZBASE which contains experimentally determined PPIs (15), PPID (unsure of source), PREDICTEDPPI (unsure of source), SNAVI molecular interaction network derived from primary papers and manually curated (16), Stelz et al. human PPI networks derived from automated yeast two-hybrid (Y2H) interaction mating experiments (17), Rolland et al. large-scale binary PPI network derived from the literature (18), and hu.MAP which is a PPI dataset derived 9,000 published mass-spectrometry experiments (19).

*PPI Path Length*: Two mutually exclusive options for the number of branches from the TF-input nodes by which to extend the PPI network (1 or 2).

[3] Kinase-Enrichment:

*Kinase Sort*: Three mutually-exclusive sort methods could be used to sort significantly enriched kinases (*same options as TF Sort*).

*Kinase Interactome*: Two different types of data are available through KEA and were used as mutually-exclusive options in the GA Kinase-Enrichment step: kinase-protein interactions (KP) and phosphorylation reactions (P). KP data is composed of data integrated from the kinase subsets of several PPI databases, including NetworkIN (20), Phospho.ELM (20), MINT (13), HPRD (9), PhosphoPoint (21) and SwissProt (22). P data was manually curated from the literature.

*Top Kinases*: The maximum number of kinases predicted by the kinase-enrichment was held constant at 20 to standardize the fitness measurements.

**X2K VALIDATION**

Two datasets were used to validate and optimize X2K results through the GA pipeline. In Run 1, whole-transcriptome data from 570 kinase perturbation experiments (e.g. knockout, knockdown, overexpression of specific kinases) in the Gene Expression Omnibus (GEO) were used to validate whether X2K was able to consistently predict the correct kinase targets that were known to have been perturbed. Fitness was calculated for each parameter combination as the percentage of experiments that contained the “correct” target kinase in the X2K predicted kinases output.

In Run 2, the NIH library of integrated network-based cellular signatures (LINCS) L1000 drug perturbation data (23) was combined with experimentally validated drug-target information curated by the Drug Repurposing Hub (DRH) (24). Only LINCS-L1000 experiments that tested known kinase-inhibitor drugs were used for Run2, and then reduced further to 570 experiments to match to match the sample size of Run 1. Fitness scores of each parameter combination are defined as the % of experiments in which the known perturbed kinase was recovered, divided by the number of kinase targets a given drug was associated with to avoid inflating fitness due to multiple targets.

**RESULTS**

In Run 1, the GA annealed after 6 generations at a peak fitness of ~20% accuracy in recovering kinase targets (~18% average population fitness). All parameters had a significant impact on fitness (p<0.0001). While there were similar patterns of increasing peak fitness over generations, the Training and Test data did differ significantly on this run (p<0.001).

In Run 1, the GA annealed after 8 generations at a peak fitness of ~28% accuracy in recovering kinase targets (~26% average population fitness). All parameters had a significant impact on fitness (p<0.0001). While there were similar patterns of increasing peak fitness over generations, the Training and Test data did differ significantly on this run (p<0.001).

**DISCUSSION**

Parameter optimization: GA Runs 1 and 2 differed in the optimal parameter combination that produced the peak fitness, including: Top TFs, TF Species, and TF Sort. In general, more sophisticated results-ordering methods (i.e. rank, combined score) for the TF Sort and Kinase Sort parameters outperformed simpler metrics such as raw p-values (though not Kinase Rank in Run 1). Both GA Runs selected TRANSFAC as the optimal TF Database, suggesting this is preferable to ChEA even when combined with TRANSFAC. While not all PPI databases overlapped between Runs, BIOCARTA, KEGG, MIPS, and PPID were selected in both Runs suggesting this combination of databases provide more robust results in X2K. Both Runs selected P as the Kinase Interactome, suggesting Phosphorylation datasets perform better for recovering kinases, though it remains to be seen whether combined the two would perform even better. Of potential concern, it is possible that X2K is biased towards constructing large PPI networks. While, both Runs did indeed create large PPI networks (~2,300-2,700 proteins) this did not appear to be related to fitness in Run 1. The addition of more databases and parameters may further improve the accuracy and consistency of the GA-optimized X2K pipeline as recovering kinase targets.

**CONCLUSIONS**

Using optimized parameters, X2K can recover the “correct” kinases with ~20-28% accuracy depending on the particular dataset. This is considerably greater than chance (4.22%). Potentially greater rates of kinase recovery and consistency across validation datasets can be achieved by adding additional databases and parameters. Further exploration of the pipeline optimization can be conducted by dividing the databases into two streams: including literature-biased resources, such as PPI and kinase-substrate interactions from extracted from low-content published studies, and high content only pipeline that includes only data from high throughput methods. The optimized parameters will inform the X2K Web application and used to infer kinases for various projects.

**References:**

1. Chen,E.Y., Xu,H., Gordonov,S., Lim,M.P., Perkins,M.H. and Ma’ayan,A. (2012) Expression2Kinases: mRNA profiling linked to multiple upstream regulatory layers. *Bioinformatics*, **28**, 105–111.  
https://doi.org/10.1093/bioinformatics/btr625  
http://www.ncbi.nlm.nih.gov/pubmed/22080467

2. Lachmann,A., Xu,H., Krishnan,J., Berger,S.I., Mazloom,A.R. and Ma’ayan,A. (2010) ChEA: Transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, **26**, 2438–2444.  
https://doi.org/10.1093/bioinformatics/btq466  
http://www.ncbi.nlm.nih.gov/pubmed/20709693

3. Berger,S.I., Posner,J.M. and Ma’ayan,A. (2007) Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. *BMC Bioinformatics*, **8**, 372.  
https://doi.org/10.1186/1471-2105-8-372  
http://www.ncbi.nlm.nih.gov/pubmed/17916244

4. Lachmann,A. and Ma’ayan,A. (2009) KEA: Kinase enrichment analysis. *Bioinformatics*, **25**, 684–686.  
https://doi.org/10.1093/bioinformatics/btp026  
http://www.ncbi.nlm.nih.gov/pubmed/19176546

5. Sandelin,A., Alkema,W., Engstrom,P., Wasserman,W.W. and Lenhard,B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, 91D–94.  
https://doi.org/10.1093/nar/gkh012  
http://www.ncbi.nlm.nih.gov/pubmed/14681366

6. Bader,G.D., Betel,D. and Hogue,C.W.V. (2003) BIND: The Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.  
https://doi.org/10.1093/nar/gkg056  
http://www.ncbi.nlm.nih.gov/pubmed/12519993

7. Xenarios,I., Salwínski,L., Duan,X.J., Patrick Higney,S.-M.K. and Eisenberg,D. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.  
https://doi.org/10.1093/nar/30.1.303  
http://www.ncbi.nlm.nih.gov/pubmed/11752321

8. Ewing,R.M., Chu,P., Elisma,F., Li,H., Taylor,P., Climie,S., McBroom-Cerajewski,L., Robinson,M.D., O’Connor,L., Li,M., *et al.* (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.*, **3**, 1–17.  
https://doi.org/10.1038/msb4100134  
http://www.ncbi.nlm.nih.gov/pubmed/17353931

9. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A., *et al.* (2009) Human Protein Reference Database - 2009 update. *Nucleic Acids Res.*, **37**, 767–772.  
https://doi.org/10.1093/nar/gkn892  
http://www.ncbi.nlm.nih.gov/pubmed/18988627

10. Breuer,K., Foroushani,A.K., Laird,M.R., Chen,C., Sribnaia,A., Lo,R., Winsor,G.L., Hancock,R.E.W., Brinkman,F.S.L. and Lynn,D.J. (2013) InnateDB: Systems biology of innate immunity and beyond - Recent updates and continuing curation. *Nucleic Acids Res.*, **41**, 1228–1233.  
https://doi.org/10.1093/nar/gks1147  
http://www.ncbi.nlm.nih.gov/pubmed/23180781

11. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U., *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, 841–846.  
https://doi.org/10.1093/nar/gkr1088  
http://www.ncbi.nlm.nih.gov/pubmed/22121220

12. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34.  
https://doi.org/10.1093/nar/27.1.29  
http://www.ncbi.nlm.nih.gov/pubmed/10592173

13. Licata,L., Briganti,L., Peluso,D., Perfetto,L., Iannuccelli,M., Galeota,E., Sacco,F., Palma,A., Nardozza,A.P., Santonico,E., *et al.* (2012) MINT, the molecular interaction database: 2012 Update. *Nucleic Acids Res.*, **40**, 857–861.  
https://doi.org/10.1093/nar/gkr930  
http://www.ncbi.nlm.nih.gov/pubmed/17135203

14. Mewes,H.-W., Hani,J., Pfeiffer,F. and Frishman,D. (1998) MIPS: a database for protein sequences and complete genomes. *Nucleic Acids Res.*, **26**, 33–37.

15. Beuming,T., Skrabanek,L., Niv,M.Y., Mukherjee,P. and Weinstein,H. (2005) PDZBase: A protein-protein interaction database for PDZ-domains. *Bioinformatics*, **21**, 827–828.  
https://doi.org/10.1093/bioinformatics/bti098  
http://www.ncbi.nlm.nih.gov/pubmed/15513994

16. Ma’ayan,A., Jenkins,S.L., Webb,R.L., Berger,S.I., Purushothaman,S.P., Abul-Husn,N.S., Posner,J.M., Flores,T. and Iyengar,R. (2009) SNAVI: Desktop application for analysis and visualization of large-scale signaling networks. *BMC Syst. Biol.*, **3**, 10.  
https://doi.org/10.1186/1752-0509-3-10  
http://www.ncbi.nlm.nih.gov/pubmed/19154595

17. Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A., Koeppen,S., *et al.* (2005) A human protein-protein interaction network: A resource for annotating the proteome. *Cell*, **122**, 957–968.  
https://doi.org/10.1016/j.cell.2005.08.029  
http://www.ncbi.nlm.nih.gov/pubmed/16169070

18. Rolland,T., Taşan,M., Charloteaux,B., Pevzner,S.J., Zhong,Q., Sahni,N., Yi,S., Lemmens,I., Fontanillo,C., Mosca,R., *et al.* (2014) A proteome-scale map of the human interactome network. *Cell*, **159**, 1212–1226.  
https://doi.org/10.1016/j.cell.2014.10.050  
http://www.ncbi.nlm.nih.gov/pubmed/25416956

19. Drew,K., Lee,C., Huizar,R.L., Tu,F., Borgeson,B., McWhite,C.D., Ma,Y., Wallingford,J.B. and Marcotte,E.M. (2017) Integration of over 9,000 mass spectrometry experiments builds a global map of human protein complexes. *Mol. Syst. Biol.*, **13**, 932.  
https://doi.org/10.15252/msb.20167490  
http://www.ncbi.nlm.nih.gov/pubmed/28596423

20. Linding,R., Jensen,L.J., Pasculescu,A., Olhovsky,M., Colwill,K., Bork,P., Yaffe,M.B. and Pawson,T. (2008) NetworKIN: A resource for exploring cellular phosphorylation networks. *Nucleic Acids Res.*, **36**, 695–699.  
https://doi.org/10.1093/nar/gkm902  
http://www.ncbi.nlm.nih.gov/pubmed/17981841

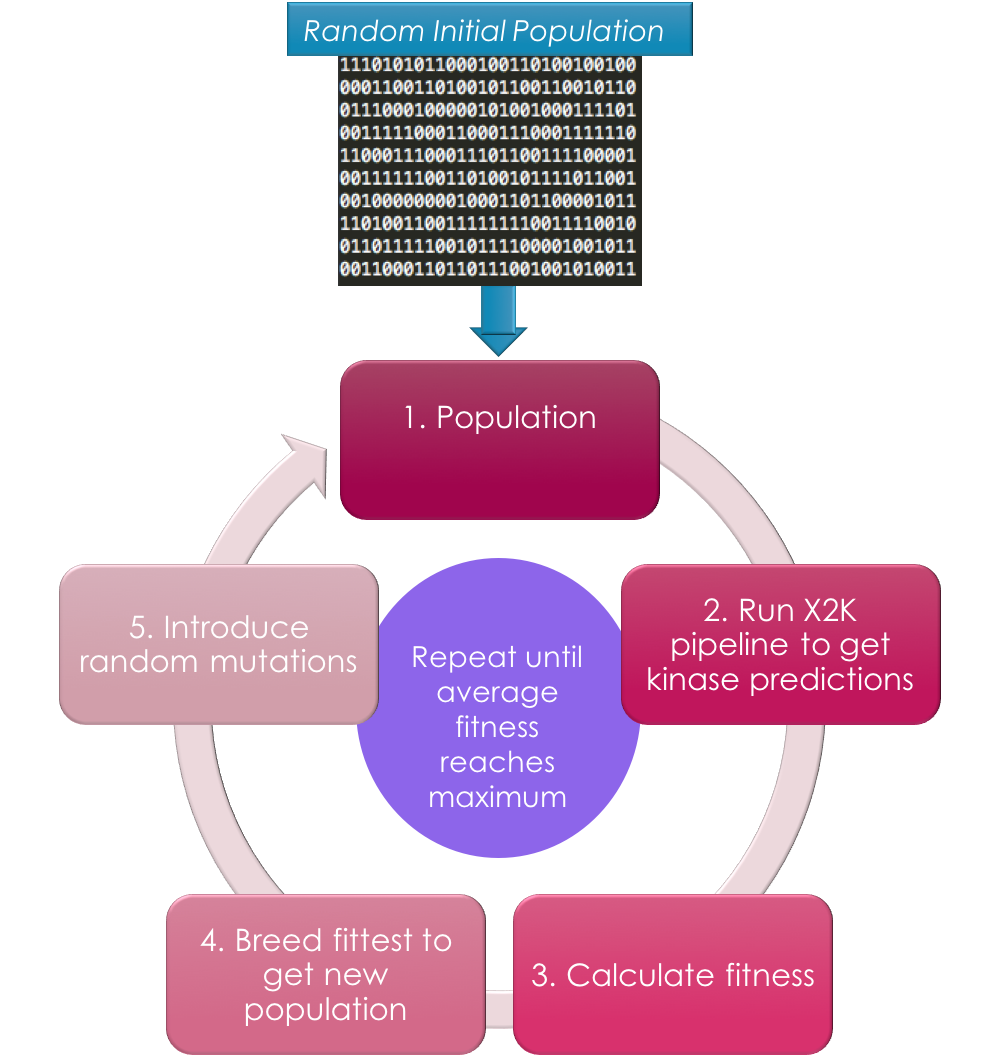
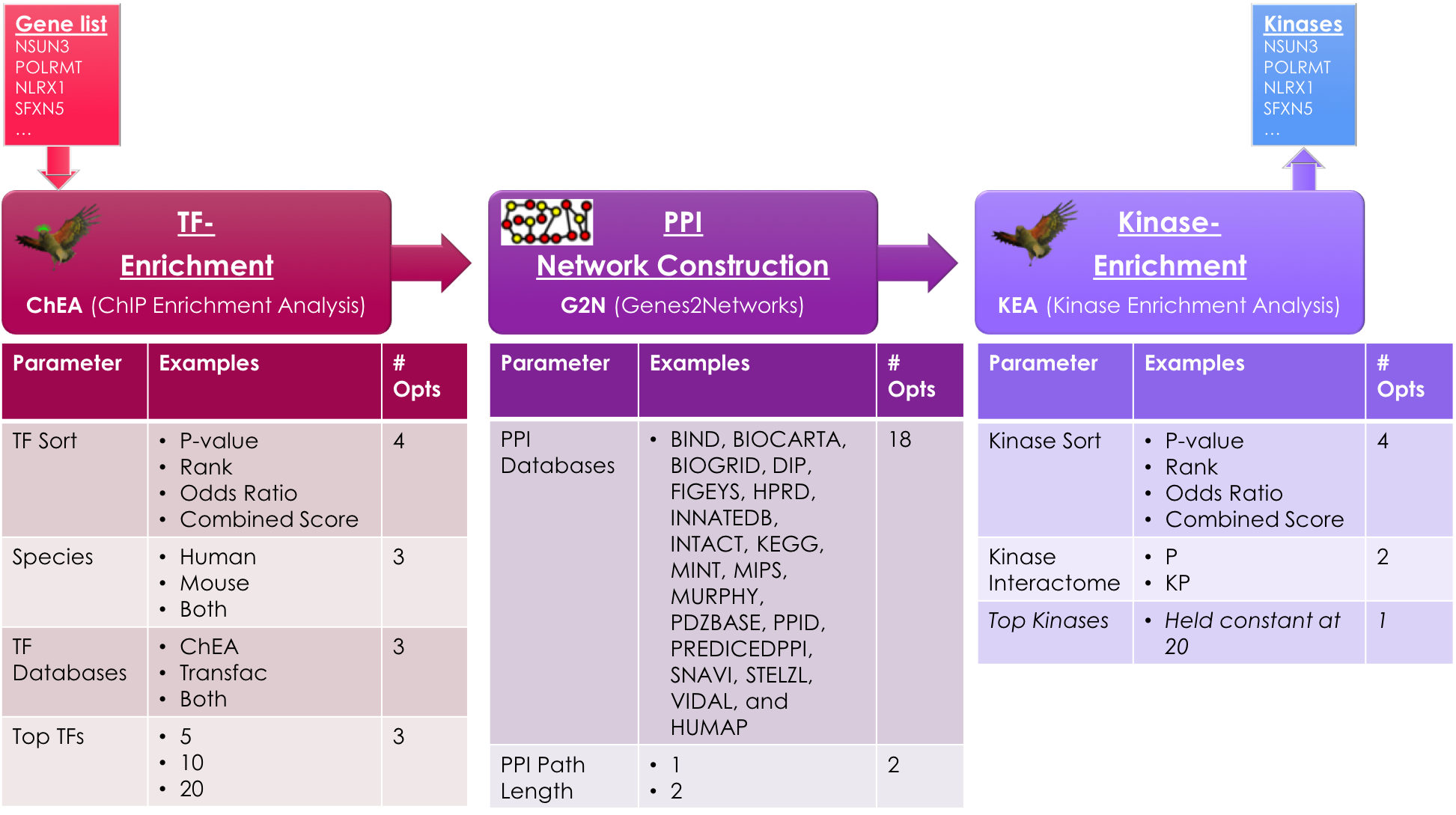
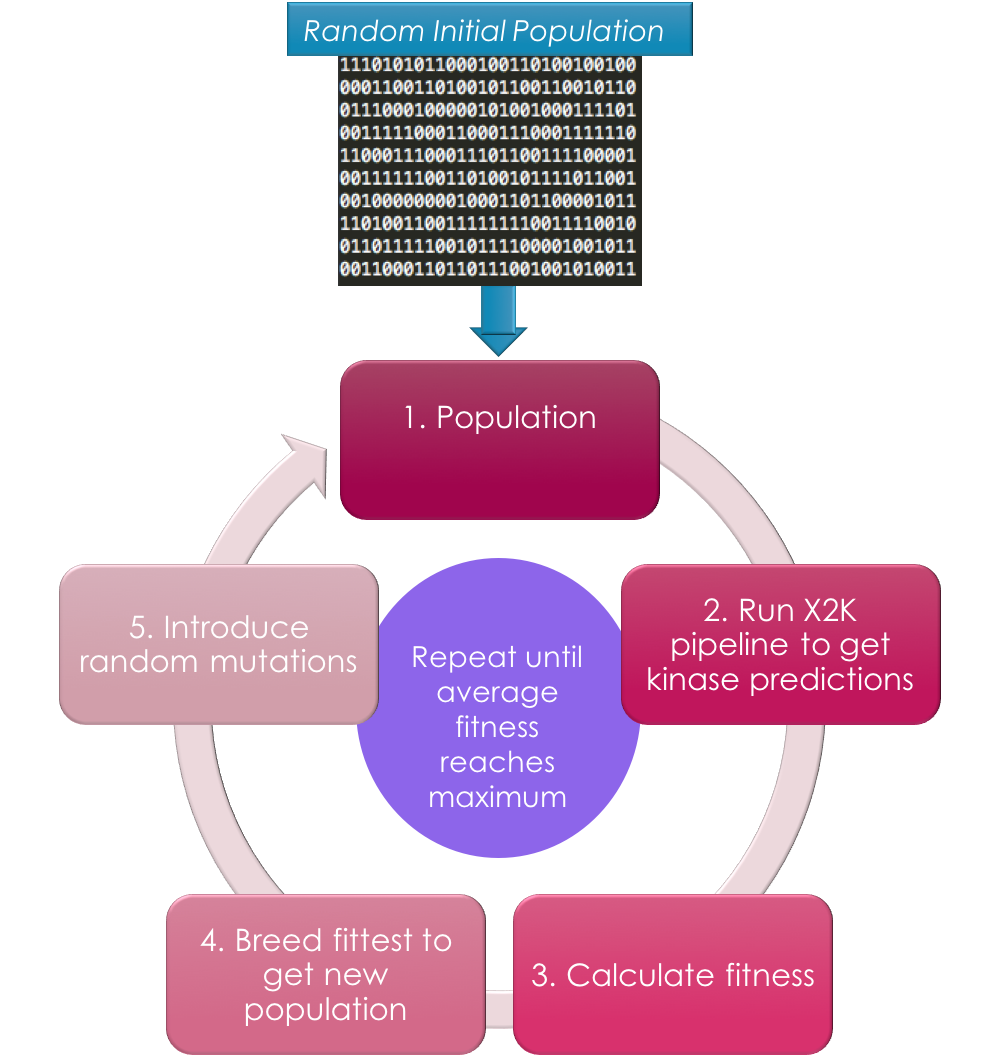
21. Yang,C.Y., Chang,C.H., Yu,Y.L., Lin,T.C.E., Lee,S.A., Yen,C.C., Yang,J.M., Lai,J.M., Hong,Y.R., Tseng,T.L., *et al.* (2008) PhosphoPOINT: A comprehensive human kinase interactome and phospho-protein database. *Bioinformatics*, **24**, 14–20.  
https://doi.org/10.1093/bioinformatics/btn297  
http://www.ncbi.nlm.nih.gov/pubmed/18689816

22. Quintaje,S.B. and Orchard,S. (2008) The Annotation of Both Human and Mouse Kinomes in UniProtKB/Swiss-Prot. *Mol. Cell. Proteomics*, **7**, 1409–1419.  
https://doi.org/10.1074/mcp.R700001-MCP200  
http://www.ncbi.nlm.nih.gov/pubmed/18436524

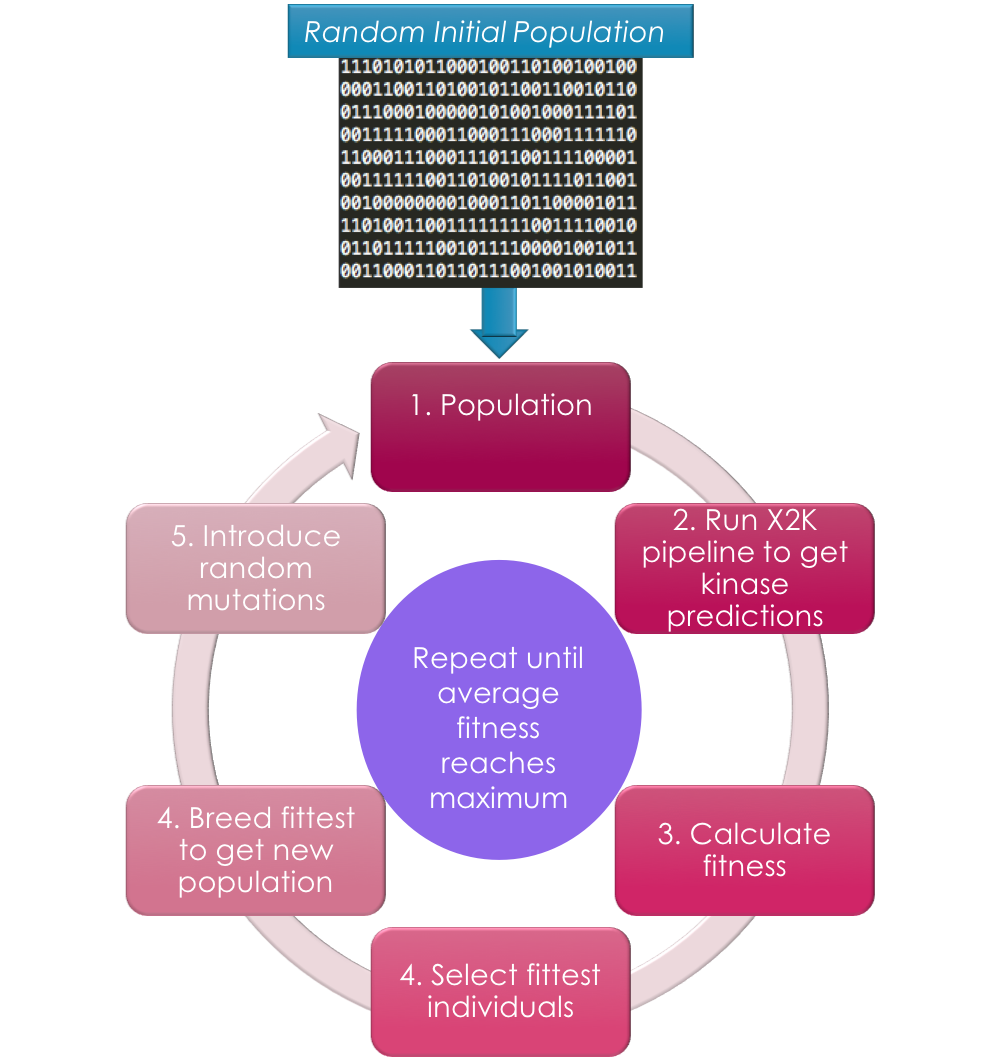
23. Duan,Q., Reid,S.P., Clark,N.R., Wang,Z., Fernandez,N.F., Rouillard,A.D., Readhead,B., Tritsch,S.R., Hodos,R., Hafner,M., *et al.* (2016) L1000CDS2: LINCS L1000 characteristic direction signatures search engine. *npj Syst. Biol. Appl.*, **2**, 16015.  
https://doi.org/10.1038/npjsba.2016.15  
http://www.ncbi.nlm.nih.gov/pubmed/28413689

24. Corsello,S.M., Bittker,J.A., Liu,Z., Gould,J., McCarren,P., Hirschman,J.E., Johnston,S.E., Vrcic,A., Wong,B., Khan,M., *et al.* (2017) The Drug Repurposing Hub: A next-generation drug library and information resource. *Nat. Med.*, **23**, 405–408.  
https://doi.org/10.1038/nm.4306  
http://www.ncbi.nlm.nih.gov/pubmed/28388612

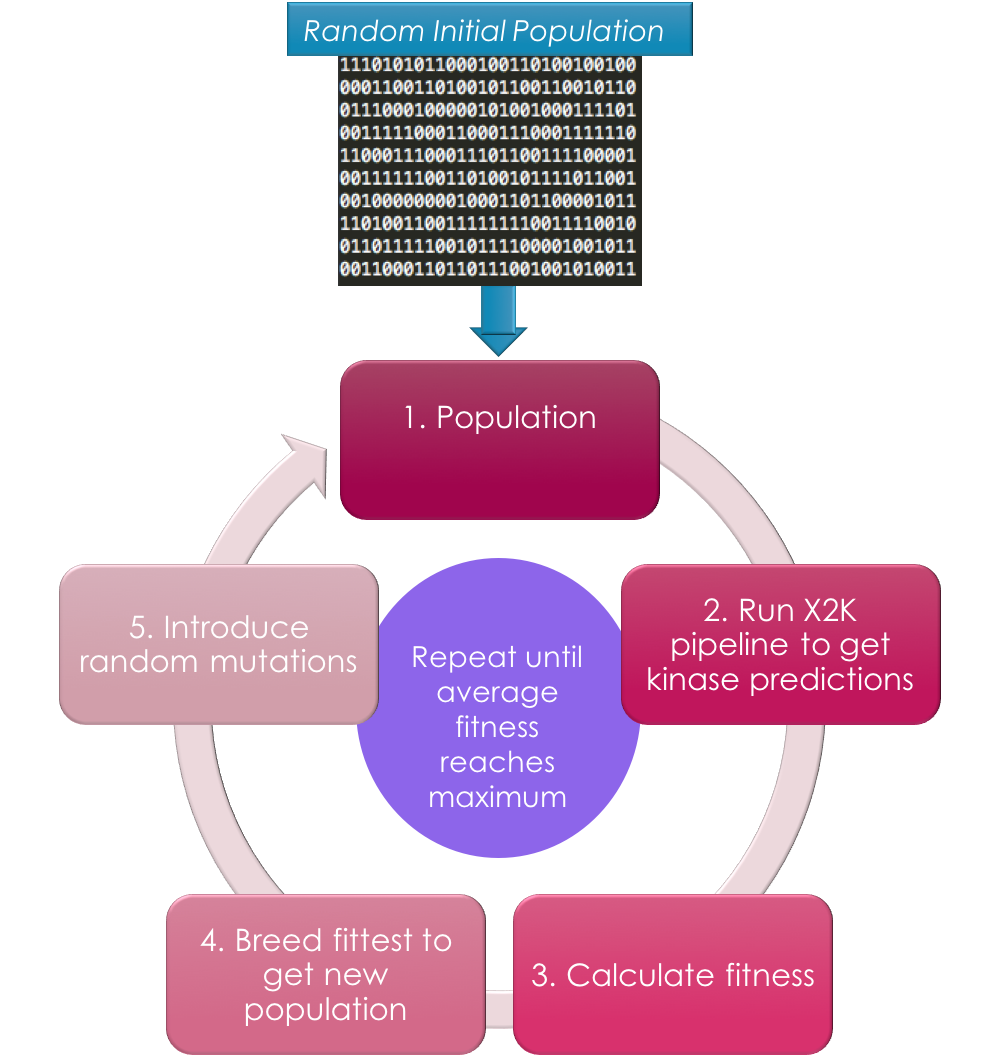
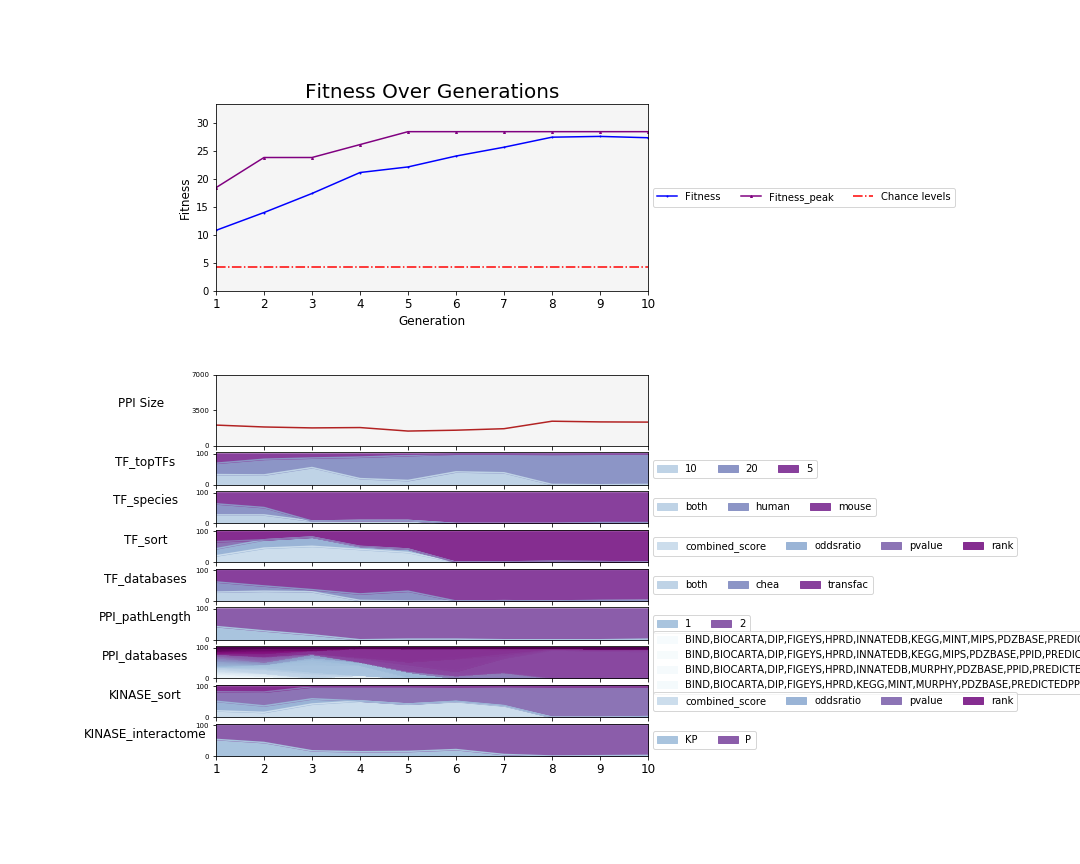
**FIGURES**

**Fig 1**. Genetic Algorithms identify improved combinations of parameters (coded as binary strings) by searching the fitness landscape.

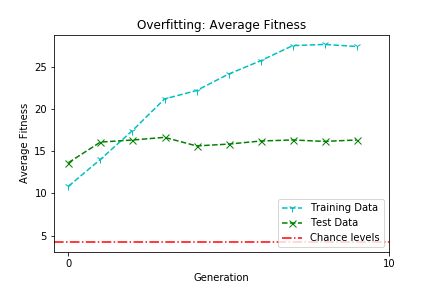


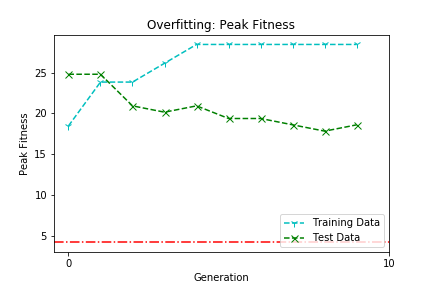
**Fig 2**. Illustration of the X2K pipeline with the modifiable parameters for each respective step, along with several example options, and the total number of options. The gene sets are first entered by the user, which then goes through three stages of the X2K pipeline (TF-enrichment, PPI network construction, and Kinase enrichment) to produce a list of predicted kinases.

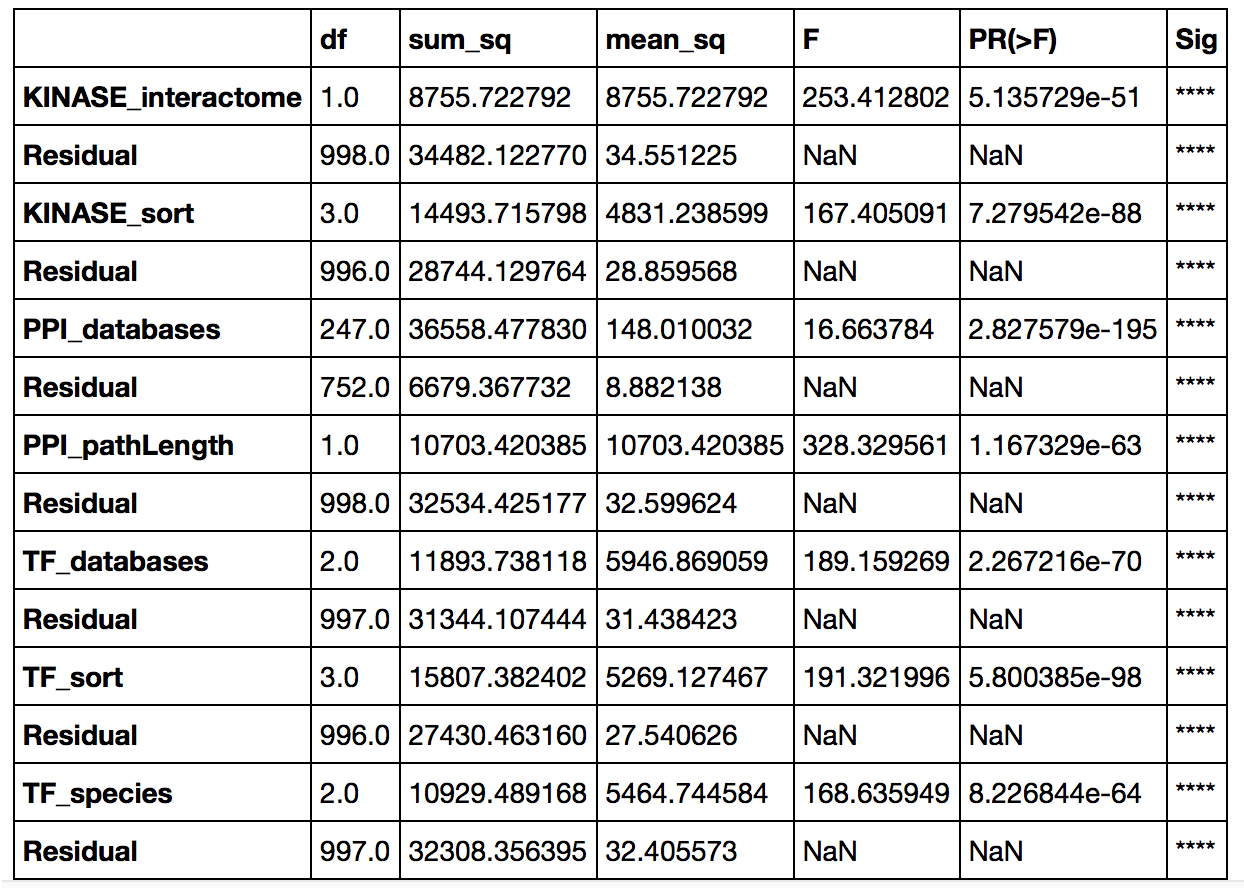
**Fig 3**. Run 1: Evolution of parameters and average population PPI size over 10 generations in the X2K GA.

**A**

**B**



**Fig 4**. Overfitting test results between training data and test data using average fitness across the entire population (B) and peak fitness.



**Table 1**. Results of a series of one-way ANOVAs statistically testing the relationship between fitness and each parameter in Run 1.