

9-1-2013

# Hypothesis driven single nucleotide polymorphism search (HyDn-SNP-S)

Rebecca J. Swett

*Department of Chemistry, Wayne State University*

Angela Elias

*Department of Chemistry, Wayne State University*

Jeffrey A. Miller

*Department of Computer Science, Eastern Michigan University*

Gregory E. Dyson

*Karmanos Cancer Institute, Wayne State University School of Medicine, [dysong@karmanos.org](mailto:dysong@karmanos.org)*

G. Andres Cisneros

*Department of Chemistry, Wayne State University, [andres@chem.wayne.edu](mailto:andres@chem.wayne.edu)*

---

## Recommended Citation

Swett R J, Elias A, Miller J A, Dyson G, Cisneros G A. (2013). Hypothesis driven single nucleotide polymorphism search (HyDn-SNP-S). *DNA Repair* 12(9): 733-740. doi: [10.1016/j.dnarep.2013.06.001](https://doi.org/10.1016/j.dnarep.2013.06.001)  
Available at: <http://digitalcommons.wayne.edu/chemfrp/4>

This Article is brought to you for free and open access by the Chemistry at DigitalCommons@WayneState. It has been accepted for inclusion in Chemistry Faculty Research Publications by an authorized administrator of DigitalCommons@WayneState.

NOTICE IN COMPLIANCE WITH PUBLISHER POLICY: This is the author's final manuscript version, post-peer-review, of a work accepted for publication in *DNA Repair*. Changes resulting from the publishing process may not be reflected in this document; changes may have been made to this work since it was submitted for publication. This version has been formatted for archiving; a definitive version was subsequently published in *DNA Repair*, 12(9): 733-740 (September 2013) <http://dx.doi.org/10.1016/j.dnarep.2013.06.001>

---

# Hypothesis driven single nucleotide polymorphism search (HyDn-SNP-S)

REBECCA J. SWETT<sup>a</sup>, ANGELA ELIAS<sup>a</sup>, JEFFREY A. MILLER<sup>b</sup>, GREGORY E. DYSON<sup>c</sup>, G. ANDRÉS CISNEROS<sup>a</sup>

<sup>a</sup>Department of Chemistry, Wayne State University, 5101 Cass Avenue, Detroit, MI 48202, USA

<sup>b</sup>Department of Computer Science, Eastern Michigan University, Ypsilanti, MI 48197, USA

<sup>c</sup>Karmanos Cancer Institute, Wayne State University, Detroit, MI 48203, USA

---

**Abstract** The advent of complete-genome genotyping across phenotype cohorts has provided a rich source of information for bioinformaticians. However the search for SNPs from this data is generally performed on a study-by-study case without any specific hypothesis of the location for SNPs that are predictive for the phenotype. We have designed a method whereby very large SNP lists (several gigabytes in size), combining several genotyping studies at once, can be sorted and traced back to their ultimate consequence in protein structure. Given a working hypothesis, researchers are able to easily search whole genome genotyping data for SNPs that link genetic locations to phenotypes. This allows a targeted search for correlations between phenotypes and potentially relevant systems, rather than utilizing statistical methods only. HyDn-SNP-S returns results that are less data dense, allowing more thorough analysis, including haplotype analysis. We have applied our method to correlate DNA polymerases to cancer phenotypes using four of the available cancer databases in dbGaP. Logistic regression and derived haplotype analysis indicates that ~80 SNPs, previously overlooked, are statistically significant. Derived haplotypes from this work link POLL to breast cancer and POLG to prostate cancer with an increase in incidence of 3.01- and 9.6-fold, respectively. Molecular dynamics simulations on wild-type and one of the SNP mutants from the haplotype of POLL provide insights at the atomic level on the functional impact of this cancer related SNP. Furthermore, HyDn-SNP-S has been designed to allow application to any system. The program is available upon request from the authors.

**Keywords** SNP search, Biomarkers, Molecular dynamics, DNA polymerases, Cancer

**Corresponding author**

Tel.: +1 313 577 9229; fax: +1 313 577 8822.

E-mail address: [andres@chem.wayne.edu](mailto:andres@chem.wayne.edu) (G. Andrés Cisneros).

---

## 1. INTRODUCTION

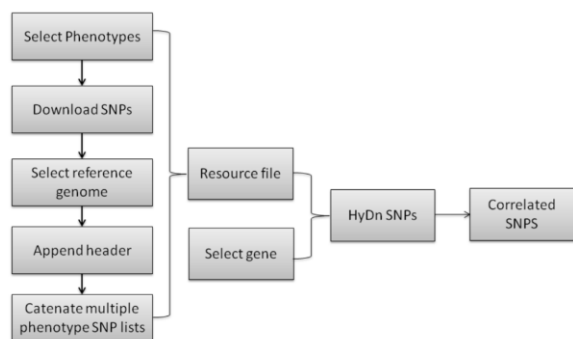
In recent years, the amount of genomic data on disease phenotypes has increased exponentially. The decreasing cost of genetic testing, along with the future promise of personalized medicine has resulted in a boom in individual genomic data [1–5]. Most bioinformatic techniques determine clusters of mutations that may be followed and used as a diagnostic tool in various diseases [6–9]. Traditional analysis of genome wide association studies (GWAS) focus on a single phenotype, and aim to find SNPs that show statistically significant association with the phenotype in any of the measured genes. In most cases these analysis do not have an a priori hypothesis of the locations of the SNPs. Therefore, very stringent statistical criteria are needed to obtain SNPs that are predictive, resulting in only a small number of SNPs being identified. Few studies have leveraged the vast information generated to identify new SNPs with clear functional impact on disease onset [10–13].

Moreover, tracking a mutation resulting from these SNPs through transcription and translation to their ultimate effects in a cell is largely left to the scientific community at large. In addition, correlating a mutation to a phenotype is

a daunting task for researchers who typically work at a cellular level. Most biochemists or molecular biologists have a biosystem of interest, and broad sweeping GWAS studies are typically intractable for their purposes. It was our intent to create a tool that would allow a user to query genomic-level data for their system of interest. There have been examples previously where this has been done on a single GWAS study, but not on combined data sets [14]. To this end, we developed an algorithm whereby a researcher can directly query one or several GWAS studies for their gene region of interest. We term this method hypothesis driven SNP search (HyDn-SNP-S). The software returns all SNP mutations within their gene region, along with information on which phenotype the mutation occurs in. Further statistical methods can then be applied to the GWAS data. Additionally, by returning a focused set of mutations, tracking the consequences of the mutations through the RNA and protein levels becomes trivial.

The workflow shown in Fig. 1 outlines the process used in HyDn-SNP-S. Researchers can select one or many GWAS for the phenotypes of interest, apply the SNP-Phenotype program to search for their gene of interest, and further analysis can be carried out as desired. Our program returns both intronic and exonic SNPs, allowing the inves-

tigation of impact on RNA, gene processing or protein sequence. The simplicity of the software implementation makes this method ideal for researchers not comfortable with large-scale bioinformatic analyses, or those who lack the resources to perform such studies.



**Fig. 1.** Flowchart of the HyDn-SNPs method. Upon development of a hypothesis, researchers select GWAS studies with relevant phenotypes, and obtain locations of the genes of interest. Following application of the algorithm, SNPs can be separated by intronic or exonic. Further analysis can be performed by *in vitro* validation or computational studies.

The database of genotypes and phenotypes (dbGaP, <http://www.ncbi.nlm.nih.gov/gap>) was developed to “archive and distribute the results of studies that have investigated the interaction of genotype and phenotype”. Data is distributed in the form of either raw genotyping data tagged with individual specific data such as gender, race and onset of phenotype; or as catenated lists of SNPs. This repository provides an ideal source of GWAS data useful to researchers with a targeted interest. Users are able to freely download the sets of SNPs, in a standard format for use with our software. Any phenotype of interest that is represented in this database would be a possible point of study for a HyDn-SNP-S study.

In this contribution, we present the development of HyDn-SNP-S and its application to search for cancer related SNPs on all human DNA polymerases. These enzymes are involved in all processes related to DNA replication, repair and recombination. The efficiency and fidelity of these processes are critical since errors can lead to carcinogenesis. Numerous studies indicate that mutations in DNA polymerases affect characteristics ranging from fidelity, to nucleotide incorporation rate, to cell proliferation [15–26]. However, a direct link has not been established between these mutations and cancer onset.

Our results uncovered a large number of cancer related SNPs on DNA polymerases. Statistical analysis on selected studies reveals for the first time the possibility that DNA POL could be a major contributor to cancer risk. Molecular dynamics simulations were performed on wild-type and a SNP mutant on POL to further investigate the functional impact of the mutation.

## 2. METHODS

In this section we describe the algorithm to search for disease related SNPs based on a given hypothesis and its implementation in an easy to use software package. Subsequently we describe the statistical methods to determine the association of the SNPs with the phenotype. This is followed by a description of the graph analysis of the resulting data from HyDn-SNP-S for the present studies. Finally, the details of molecular dynamics (MD) simulations on DNA polymerase lambda structures are described.

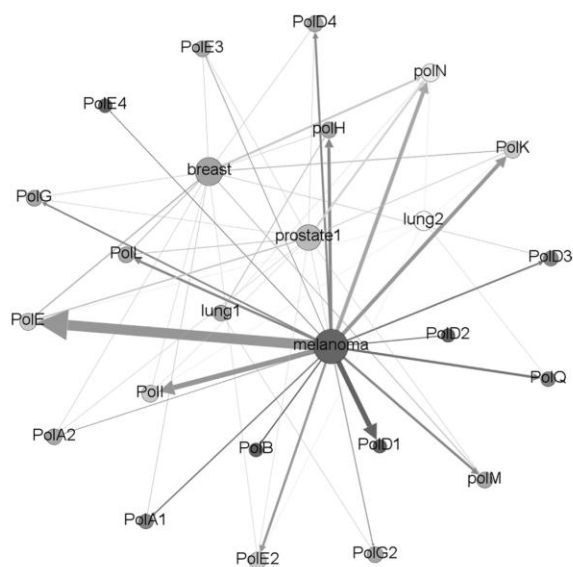
### 2.1. HyDn-SNP-S

SNP collections for each phenotype are obtained from studies deposited on the database of genotypes and phenotypes (dbGaP). A header is appended to each data set declaring the phenotype associated with each individual study. Mutations listed relative to the HuRef and Celera genomes are removed, as we are working within the frame of reference of the GrCh37 human genome reference build. All files containing the SNP collections are subsequently catenated into a single searchable resource file. This single file allows HyDn-SNP-S to search all relevant SNPs for every phenotype that was included in the resource file in one single run. Following the generation of this resource file, the program HyDn SNPs was used to search for mutations within the gene region of interest. Users enter the chromosome, and gene range for searching, and point the program to the resource file. Sample resource files are available with the HyDn SNPs download. Further information and instructions are available in the documentation for this program. Any SNPs found that match the chromosome and gene location range are deposited into a results file. This file lists all the SNP associated information, such as ss and rs number, allele, chromosome, chromosomal location, contig number, and contig location, and type of chip used in the original genotyping experiment. These can then be categorized by location; intronic, exonic, or at a splice site. For our purposes, exonic SNPs were then compared to reference SNPs to ascertain the extent of prior investigation, as well as relative allele frequency in the natural population. The consequence of any given SNP was determined either by use of the reference SNP database, or in the case of previously unreported SNPs, translated by use of a DNA codon table in conjunction with the gene sequence and protein sequence. HyDnSNPs is available upon request from the authors.

### 2.2. Statistical analysis

We utilized four publically available case/control genome wide association studies (GWAS) from dbGAP (access request #1961) across multiple cancer types (including breast, melanoma, lung and prostate cancers) [27–32] to determine if SNPs or haplotypes constructed from SNPs in our genes of interest (genes coding for all human polymerases) are associated with any of these disease phenotypes.

Additionally, we determined if any synergistic results across multiple databases exist that may imply a common cancer genesis. Multiple genetic modes of inheritance were examined: additive, dominant, recessive and genotypic in a covariate-adjusted logistic regression analysis associating each SNP with the disease phenotype. The maximum likelihood estimate of the posterior probabilities of haplotypes for each observation was produced using the EM algorithm. Score statistics for the association of the haplotypes with the cancer phenotype were constructed using these posterior probabilities. We use the R package “haplo.stats” to implement these haplotype functions [33]. Logistic regression is also used to estimate the association of a haplotype with the disease phenotype, given the genetic context. As we focus on the SNPs in only a few genes, we avoid issues with multiple testing, which are burdensome when trying to evaluate the association between genetic markers and a disease phenotype when measuring thousands or millions of genetic variants.



**Fig. 2** Edge-node network of the HyDn-SNPs results. Phenotypes and polymerases are shown as nodes, edges are weighted by total number of SNPs connecting each phenotype to each polymerase. This is also available as an interactive map at <http://www.chem.wayne.edu/cisnerosgroup/gexf-js2/index2.html>

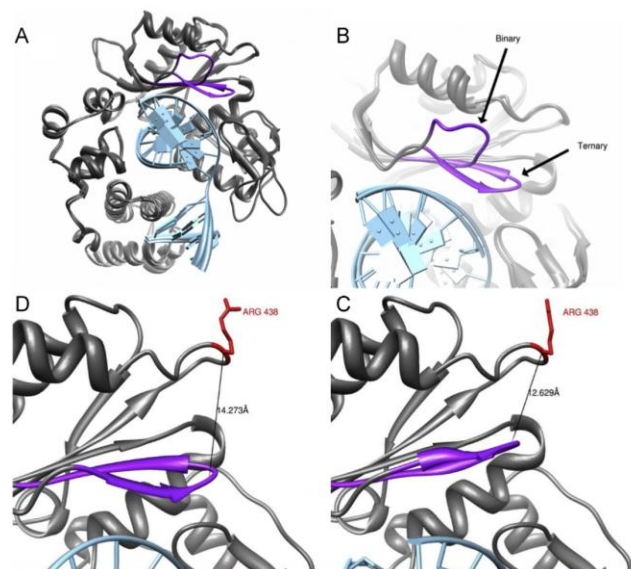
### 2.3. Graph analysis

For ease of visual analysis, the data resulting from the HyDn-SNP-S search has been transformed into edge-node format to allow visual interpretation of the networks of phenotypes and polymerases involved in tumorigenesis. Frequently more than one polymerase was found to have single point mutations within a cohort of cancer patients; network analysis allows for easy visual interpretation. Edge-node tables were csv formatted for use in Gephi [34], visualization was performed with a Fruchterman–Reingold [35] algorithm using an area of 15,000 and a gravity of 7.0.

Nodes and edges were weighted by degree; for these analyses, weight was the number of mutations occurring between each phenotype and polymerase.

### 2.4. MD simulations

MD was performed on wild-type and the R438W mutant of DNA POLI in the binary and ternary conformations (PDBID: 1RZT, 2PFQ) using NAMD. The simulations were performed using a parallel build of NAMD [36] employing the CHARMM [37] force field on the XSEDE Teragrid. The structures were solvated, and appropriate counterions were added to reach 0.5 mM NaCl. A timestep of one femtosecond was used, a Langevin thermostat was used to maintain temperature at 300 K, and a Nose–Hoover Langevin combination method was used to control pressure. The systems were solvated with TIP3P water, neutralized with counter ions and subjected to 1000 steps of conjugate gradient minimization and temperature ramped to 300 K. After equilibration, the systems were run for at least 14 ns of production time. Frames from the trajectories were written every 1 ps. The solvation boxes included a 15 Å pad on each face of the box. Long range electrostatics were calculated using particle mesh Ewald [38], and van der Waals were calculated with a nonbonded cutoff of 8 Å and a switching function between 7 and 8 Å.



**Fig. 3.** (A) Overlay of Polλ in the binary and ternary conformations. DNA is shown in light blue, and the Loop 1 is shown in purple. (B) Differences in Loop 1 orientation between the two conformations. Distance between position 438 and Loop 1 following an interpolation between the two structures at its furthest (Panel D) and closest (Panel C) approaches.

Correlation analysis by residue was carried out for each system using the ptraj module of Amber11, across the entire simulation. An all residue correlation was performed and difference plots were calculated using Mathematica



[39]. Correlation between the mutated residue and the residues in Loop 1 were also calculated and plotted. Generalized masked Delaunay analysis was carried out using the TimeScapes software from the D.E. Shaw group [40]. Trajectories were prepared using VMD [41], and all solvent and nucleic acids were excluded from analysis. A sliding window of 5% of the total number of frames was used, and total events per frame were calculated and plotted against frame number.

### 3. RESULTS AND DISCUSSION

The HyDn-SNP-S method returns results from whole genome genotyping studies rapidly, far faster than traditional bioinformatic methods. Pre-screening with HyDn-SNP-S dramatically decreases the time required to perform statistical analysis on GWAS data, by excluding all mutations not relevant to a researcher's hypothesis. As proof of concept four genotyping studies have been statistically analyzed following application of the HyDn-SNP-S method. Additionally, one mutation, both determined to be statistically significant and of structural interest was subjected to molecular dynamics studies and consequent analysis.

Upon searching four cancer phenotype studies (melanoma, breast, lung and prostate cancer) [27–32] for mutations in all polymerase genes, a total of 708 mutations were found. Of these mutations 491 were intronic, and 217 were exonic. Additionally, four of the exonic mutations were found to be at splice sites. As per the workflow described above, all four searches were carried out simultaneously, and results were available within a few minutes. Following application of the HyDn-SNP-S analysis, the four studies were subjected to traditional biostatistical analysis. The focused nature of the search allows for relaxation of the more stringent mathematical methods, and facilitates more thorough analysis of the resulting mutations. Haplotype analysis on whole genome genotyping data is frequently not performed as the combinatorial nature of these studies across all mutations would be prohibitively computationally expensive. As the dataset used for analysis following the HyDn SNP-S method has significantly reduced complexity, these targeted studies can detect mutations of moderate significance that would be overlooked in traditional bioinformatic analyses and perform these searches more rapidly than is typically possible.

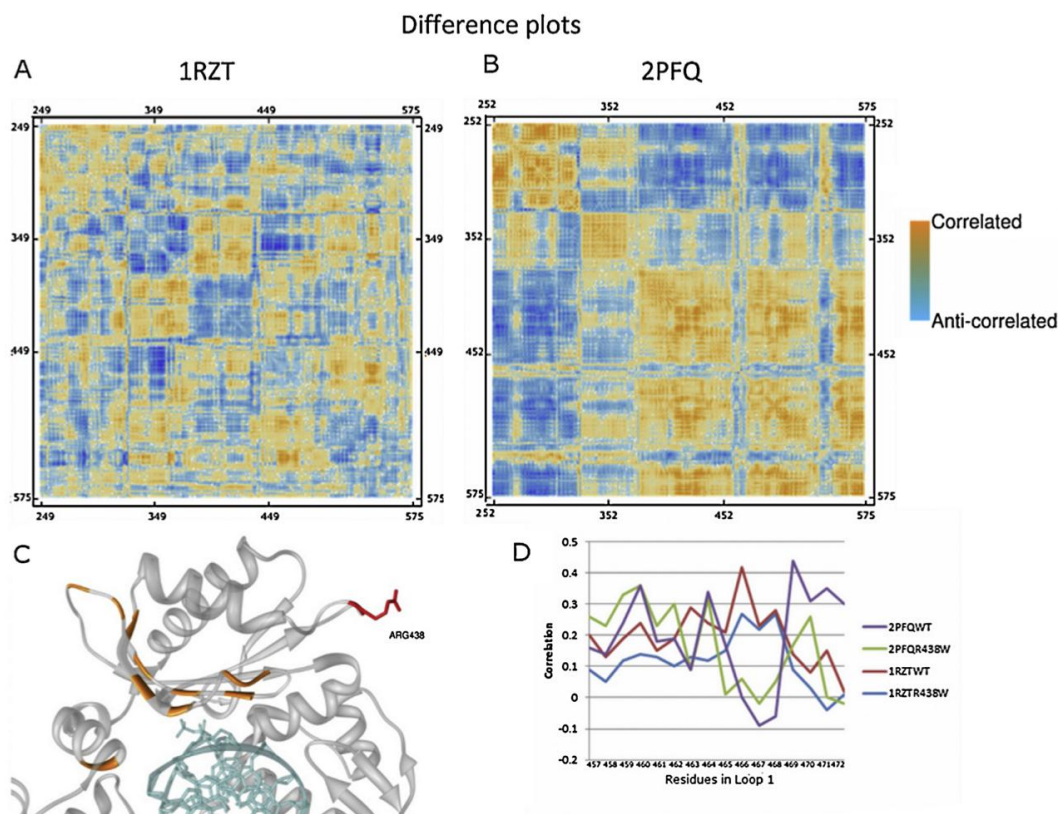
Logistical regression and haplotype analysis was performed on these studies to determine statistical significance. The prostate cancer case/control database examined yielded 69 SNPs in the genes of interest. Eleven of them were statistically significantly associated with prostate cancer status for at least one genetic model. The melanoma cancer case/control database examined yielded 215 SNPs in the genes of interest. Twenty-six of them were significantly associated with melanoma case/control status for at least one genetic model after controlling for age and gender. The breast cancer case/control database examined

yielded 100 SNPs in the genes of interest. Twenty-two of them were statistically significantly associated with prostate cancer status for at least one genetic model. The lung cancer case/control database examined yielded 51 SNPs in the genes of interest. Twenty of them were statistically significantly associated with prostate cancer status for at least one genetic model. Table S1 reports all of the significant SNPs, their p-value and corresponding POL gene.

Analysis was performed to determine the association between the derived haplotypes from each gene and disease status. No haplotypes were predictive of disease status for the lung cancer study nor the melanoma study using any of the three genetic models. However, the haplotypes constructed from SNPs on POLL were borderline significant for the breast cancer study using a recessive ( $p$ -value = 0.048) or additive ( $p$ -value = 0.091) model formulation. This haplotype is constructed from two SNPs: rs3730477 (C > T; R438W) and rs3730463 (A > C: T221P). The odds ratios from individual significant and borderline significant contrasts within each model type are reported below. In the case of the additive model, for each additional C–A haplotype observed, the odds of breast cancer are multiplied by 1.15 ( $p$ -value = 0.029). Similarly, for each additional C–C haplotype observed, the odds of breast cancer are multiplied by 0.812 ( $p$ -value = 0.062), i.e., a protective genotype. For the recessive model, having 0 or 1 copy of the C–A haplotype results in the odds of breast cancer being multiplied by 0.829 relative to having 2 copies of the C–A haplotype ( $p$  = 0.026). Having 0 or 1 copy of the C–C haplotype results in the odds of breast cancer being multiplied by 3.01 relative to having two copies of the C–C haplotype ( $p$  = 0.099). The haplotypes constructed from SNPs on the PolG genes were significant for prostate cancer. This haplotype is constructed from three SNPs: rs3087374, rs2351000 and rs2247233. The odds ratios from individual significant contrasts within each model type are reported below. For the recessive model, having 0 or 1 copy of the G–T–G haplotype results in the odds of prostate cancer being multiplied by 1.33 relative to having 2 copies of the G–T–G haplotype ( $p$  = 0.005). Having 2 copies of the G–C–A haplotype results in 9.64 of the odds of prostate cancer as compared to having 0 or 1 copy of the G–C–A haplotype ( $p$  = 0.008).

A literature search indicates that only one of these statistically evaluated mutations has been explored in vitro [42]. Experimental analysis of the mutations we report here is outside the scope of this work. However, the mutations arising from these SNPs present interesting targets for further experimental studies.

The data resulting from a HyDn-SNP-S search can not only be discussed at the molecular level, and in the context of predictive power, but due to the nature of these studies, on a much broader basis. Relating many phenotypes to many polymerases generates a network of data best represented by an edge-node interactive diagram. Using the number of SNPs as a weighting property, it is possible to



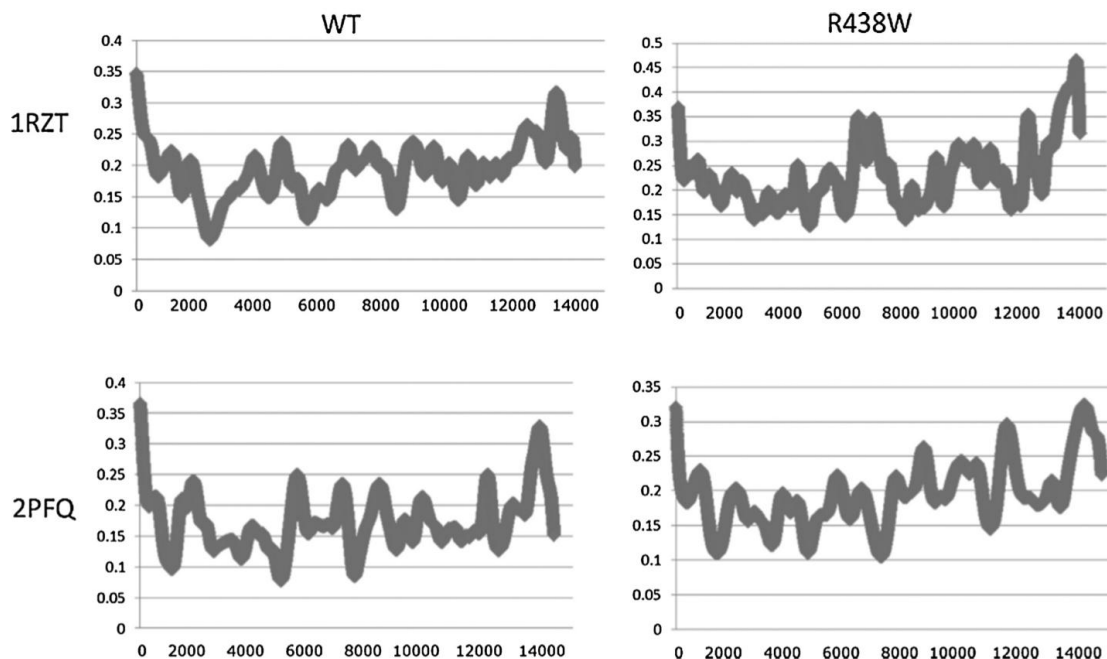
**Fig. 4.** Correlation difference plots for the binary (A) and ternary (B) conformations relative to the wild type. Increases in correlation are shown in orange, while increases in anti-correlated motions are shown in blue. In both cases, alterations in the correlation plots are visible, more notably in the ternary complex. The highest values from the ternary complex correlation plots were mapped back to the residues affected, and are colored orange in Panel C. Notably many of these residues are on Loop 1. Panel D shows the individual correlation values for each of the residues in Loop 1. While the binary complex shows moderate alteration on several, the ternary complex shows considerable differences for several residues, particularly between residues 469 and 472.

broadly examine the complete network of phenotype–polymerase interactions. Fig. 2 shows a flattened version of this data, limited to statistically explored phenotype data, which is available in interactive form online at <http://www.chem.wayne.edu/cisnerosgroup/gexf-js2/index2.html>. Both polymerases and phenotypic studies are represented as nodes, sized according to SNP density. By clicking on a node, all connections to that node will be listed. Clicking on any of those connections will return all connections to that selected node.

The interactive map allows investigation of the network associations of various phenotypes and polymerases and the complete list of connections is available from the dropdown menu at the top of the page. A complete list of all connections is available upon request from [andres@chem.wayne.edu](mailto:andres@chem.wayne.edu). This file also includes the translated mutations. Many diseases are not caused by a single point mutation, but rather by a collection of factors. As the formatting for the results of HyDn-SNP-S is well suited to network analysis, and additional data can be garnered as desired from the genotyping studies, this approach may have critical importance in searching for combinations of

factors that may be predictive for disease. Due to the targeted nature of the search, there is a significant reduction in the analytic space and thus, more thorough analysis can be performed. The haplotype described above is one example; individually the two mutations would have been overlooked by traditional analysis, but in combination they are strongly predictive.

To further validate that hypothesis driven analysis of whole genome genotyping data is valuable to researchers, we sought to study a mutation with statistical significance that would have been overlooked by traditional methods. Of the two mutations that comprise the haplotype linking POLL to breast cancer, only the mutation R438W is in the polymerase domain. This position is not close to the active site, but it is within 14 Å of Loop 1, which has been shown to be critical for fidelity [43]. The R438W SNP mutation has been previously shown to contribute to decreased fidelity *in vitro*, increased mutation frequency, and generation of chromosomal abnormalities [44]. An eightfold increase in inaccurate substitutions was observed in base substitution assays and karyotypic analysis of several cell lines carrying this mutation also reported a high level of



**Fig. 5.** GMD plots for the binary and ternary complex simulations in both wild type and mutant form. No drastic differences are apparent between the four simulations indicating that all four are showing the same general level of physical activity. This indicates that the overall motions of the polymerase are not perturbed. In light of the data presented in Fig. 4., this indicates that the significant alterations in conformational space are restricted to the Loop 1 region.

spontaneous or IR-induced chromosomal aberrations. With ample evidence to suggest a molecular basis for these results, we selected DNA polymerase lambda R438W for further study.

Four MD simulations were performed, using crystal structures 1RZT and 2PFQ. These structures were selected as they represent the binary and ternary complexes of Pol lambda, respectively. The binary complex includes Pol lambda and the template DNA, the ternary complex includes Pol lambda with both the template and incoming nucleotide. The change in Loop 1 conformation between the two structures is shown in Fig. 3. Panel A overlays the binary and ternary complexes, Loop 1 is shown in purple to illustrate the alteration in conformation. Panel B shows a closer view of the loop conformations, indicating both binary and ternary conformations. Panels B and C illustrate the relative proximity to the R438W mutation. As the structure transits between the two loop conformations, the mutation ranges from roughly 12.6 Å to 14.3 Å away. A video illustrating the position of the mutation and the visual interpolation of the binary and ternary conformations is available as Movie S1. Mutations in this loop have been shown to have no effect on catalytic rate while simultaneously increasing the number of misincorporations, thus Loop 1 is critical for polymerase fidelity [43]. Following a 14ns simulation, correlation analysis was carried out to determine whether the residues in Loop 1 were affected by the mutation.

As shown in Fig. 4, the binary complex shows little change in correlation between the wild-type and mutant structures. Conversely, the ternary complex shows high correlation and anti-correlation in two regions. The highest points of correlation are between residues at positions 438 and 569, as well as between 438 and 420. Residues showing the greatest change in correlation in the ternary complex were mapped to the structure and colored orange as shown in Fig. 4C. It is notable that a majority of these residues are on Loop 1. To further understand the impact of the SNP mutation on Loop 1, the correlation data between position 438 and all residues in Loop 1 was extracted and plotted. Fig. 4D shows that although there is higher correlation in the ternary complex between the wild-type and mutant, both complexes show altered correlation between the wild-type and mutant. The sum of these analyses suggests that the introduction of the R438W mutation alters the overall correlation pattern in the ternary complex, but more importantly, directly affects the motions of Loop 1 in the both complexes. As Loop 1 regulates fidelity, and transit between the two conformations shown in Fig. 3 is required for catalysis, the SNP leading to the R438W mutation likely has direct effects on polymerase activity *in vitro* and *in vivo*.

In addition to the correlation analysis, generalized masked Delaunay (GMD) analysis was performed to determine the impact, if any, on the overall activity of the simulations. The results are shown in Fig. 5, events are plotted on the Y-axis, and frame number is plotted on the



X-axis. GMD defines events as persistent motions across the masked Delaunay reduced representation of the protein structure. Panels A and B show the wild type activity for both the binary and ternary complexes, with average activity levels of roughly 0.2 events per frame. These patterns are typical of stable simulations, where no major rearrangement is occurring. The alteration in the correlation plots combined with the stability of the GMD indicates that the mutation induces only local alterations in activity.

The sum of the correlation and GMD analysis indicates that the R438W mutation appears to modify only the movement of Loop 1, while the overall dynamics are not significantly altered. This provides context for the experimental work by Terrados et al. [44]. Their experiments indicated that the R438W mutation increases the error rate of Pol lambda, but does not alter the overall rate of polymerization. Our results indicate that the R438W mutation alters only the behavior of Loop 1, while leaving the overall conformational motions of Pol lambda unperturbed. This would agree with the behavior observed experimentally. The R438W mutation alters the behavior of Loop 1, thus decreasing fidelity, while the overall behavior of the polymerase is unaffected, allowing it to maintain a normal rate of polymerization.

#### 4. CONCLUSIONS

We have developed a powerful method that allows researchers to interact with whole genome genotyping data in a focused, hypothesis driven way. By allowing researchers to find data on their own systems of interest, we will expedite the study of any mutations that may logically be connected to a phenotype. Also, the focused nature of these searches will allow more thorough statistical analysis, and appropriate recognition to combinations of factors that would be difficult to fully assess in an extremely broad GWAS analysis. By applying this methodology to our system of interest we were able make the first direct statistical link between DNA polymerases and cancer, define two haplotypes with strong predictive power, and trace a cancer-associated mutation to a structural effect in the translated protein and investigate its functional impact by computational simulations.

---

#### CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest.

#### ACKNOWLEDGEMENTS

Computational time from TeraGrid and Wayne State C&IT, as well as access to genome data from dbGaP under access request #1961 is gratefully acknowledged. This work was funded by Wayne State University.

#### REFERENCES

1. The Human Genome Project: 10 years later, *Lancet* 375 (2010) 2194.
2. C.T. Caskey, Presymptomatic diagnosis: a first step toward genetic health care, *Science* 262 (1993) 48–49.
3. C.T. Caskey, Using genetic diagnosis to determine individual therapeutic utility, *Annu. Rev. Med.* 61 (2010) 1–15.
4. D. Peakall, L. Shugart, The Human Genome Project (HGP), *Ecotoxicology* 11 (2002) 7.
5. B.J. Rossiter, C.T. Caskey, Presymptomatic testing for genetic diseases of later life: pharmacoepidemiological considerations, *Drugs Aging* 7 (1995) 117–130.
6. P. Margaritte, C. Bonaiti-Pellie, M.C. King, F. Clerget-Darpoux, Linkage of familial breast cancer to chromosome 17q21 may not be restricted to early-onset disease, *Am. J. Hum. Genet.* 50 (1992) 1231–1234.
7. M.C. King, J.H. Marks, J.B. Mandell, Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2, *Science* 302 (2003) 643–646.
8. M.C. King, Localization of the early-onset breast cancer gene, *Hosp. Pract. (Off. Ed.)* 26 (1991) 121–126.
9. M.A. Austin, M.C. King, K.M. Vranizan, R.M. Krauss, Atherogenic lipoprotein phenotype. A proposed genetic marker for coronary heart disease risk, *Circulation* 82 (1990) 495–506.
10. J. Gudmundsson, P. Sulem, D.F. Gudbjartsson, T. Blondal, A. Gylfason, B.A. Agnarsson, K.R. Benediktsdottir, D.N. Magnusdottir, G. Orlygsdottir, M. Jakobsdottir, S.N. Stacey, A. Sigurdsson, T. Wahlfor, T. Tammela, J.P. Breyer, K.M. McReynolds, K.M. Bradley, B. Saez, J. Godino, S. Navarrete, F. Fuertes, L. Murillo, E. Polo, K.K. Aben, I.M. van Oort, B.K. Suarez, B.T. Helfand, D. Kan, C. Zanon, M.L. Frigge, K. Kristjansson, J.R. Gulcher, G.V. Einarsson, E. Jonsson, W.J. Catalona, J.I. Mayordomo, L.A. Kiemeny, J.R. Smith, J. Schleutker, R.B. Barkardottir, A. Kong, U. Thorsteinsdottir, T. Rafnar, K. Stefansson, Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility, *Nat. Genet.* 41 (2009) 1122–1126.
11. D.J. Hunter, P. Kraft, K.B. Jacobs, D.G. Cox, M. Yeager, S.E. Hankinson, S. Wacholder, Z. Wang, R. Welch, A. Hutchinson, J. Wang, K. Yu, N. Chatterjee, N. Orr, W.C. Willett, G.A. Colditz, R.G. Ziegler, C.D. Berg, S.S. Buys, C.A. McCarty, H.S. Feigelson, E.E. Calle, M.J. Thun, R.B. Hayes, M. Tucker, D.S. Gerhard, J.F. Fraumeni, R.N. Hoover, G. Thomas, S.J. Chanock, A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer, *Nat. Genet.* 39 (2007) 870–874.
12. C.M. Hutter, A.M. Young, H.M. Ochs-Balcom, C.L. Carty, T. Wang, C.T.L. Chen, T.E. Rohan, C. Kooperberg, U. Peters, Replication of breast cancer GWAS susceptibility loci in the Women's Health Initiative African American SHARe study, *Cancer Epidemiol. Biomark. Prev.* 20 (2011) 1950–1959.
13. T.A. Manolio, Genomewide association studies and assessment of the risk of disease, *N. Engl. J. Med.* 363 (2010) 166–176.
14. N. Manjarrez-Orduño, E. Marasco, S.A. Chung, M.S. Katz, J.F. Kiridly, K.R. Simpfendorfer, J. Freudenberg, D.H. Ballard, E. Nashi, T.J. Hopkins, CSK regulatory polymorphism is associated with systemic lupus erythematosus and influences B-cell signaling and activation, *Nat. Genet.* 44 (2012) 1227–1230.
15. W.C. Copeland, N.K. Lam, T.S. Wang, Fidelity studies of the human DNA polymerase alpha. The most conserved region among alpha-like DNA polymerases is responsible for metal-induced infidelity in DNA synthesis, *J. Biol. Chem.* 268 (1993) 11041–11049.
16. W.C. Copeland, T.S. Wang, Mutational analysis of the human DNA polymerase alpha. The most conserved region in alpha-like DNA polymerases is involved in metal-specific catalysis, *J. Biol. Chem.* 268 (1993) 11028–11040.
17. D.L. Daee, T.M. Mertz, P.V. Shcherbakova, A cancer-associated DNA polymerase  $\beta$  variant modeled in yeast causes a catastrophic increase in genomic instability, *Proc. Natl. Acad. Sci.* 107 (2010) 157–162.

18. T. Lang, M. Maitra, D. Starcevic, S.-X. Li, J.B. Sweasy, A DNA polymerase  $\beta$  mutant from colon cancer cells induces mutations, *Proc. Natl. Acad. Sci. U.S.A.* 101 (2004) 6074–6079.
19. M.J. Longley, S. Clark, C. Yu Wai Man, G. Hudson, S.E. Durham, R.W. Taylor, S. Nightingale, D.M. Turnbull, W.C. Copeland, P.F. Chinnery, Mutant POLG2 disrupts DNA polymerase gamma subunits and causes progressive external ophthalmoplegia, *Am. J. Hum. Genet.* 78 (2006) 1026–1034.
20. M.J. Longley, P.A. Ropp, S.E. Lim, W.C. Copeland, Characterization of the native and recombinant catalytic subunit of human DNA polymerase gamma: identification of residues critical for exonuclease activity and dideoxynucleotide sensitivity, *Biochemistry* 37 (1998) 10529–10539.
21. T. Matsuda, K. Bebenek, C. Masutani, I.B. Rogozin, F. Hanaoka, T.A. Kunkel, Error rate and specificity of human and murine DNA polymerase  $\epsilon$ , *J. Mol. Biol.* 312 (2001) 335–346.
22. E. Ohashi, K. Bebenek, T. Matsuda, W.J. Feaver, V.L. Gerlach, E.C. Friedberg, H. Ohmori, T.A. Kunkel, Fidelity and processivity of DNA synthesis by DNA polymerase kappa, the product of the human DINB1 gene, *J. Biol. Chem.* 275 (2000) 39678–39684.
23. E. Ohashi, T. Ogi, R. Kusumoto, S. Iwai, C. Masutani, F. Hanaoka, H. Ohmori, Error-prone bypass of certain DNA lesions by the human DNA polymerase kappa, *Genes Dev.* 14 (2000) 1589–1594.
24. I.B. Rogozin, Y.I. Pavlov, K. Bebenek, T. Matsuda, T.A. Kunkel, Somatic mutation hotspots correlate with DNA polymerase  $\epsilon$  error spectrum, *Nat. Immunol.* 2 (2001) 530–536.
25. J.B. Sweasy, T. Lang, D. Starcevic, K.-W. Sun, C.-C. Lai, D. Di-Maio, S. Dalal, Expression of DNA polymerase  $\beta$  cancer-associated variants in mouse cells results in cellular transformation, *Proc. Natl. Acad. Sci. U.S.A.* 102 (2005) 14350–14355.
26. S.W. Wong, A.F. Wahl, P.M. Yuan, N. Arai, B.E. Pearson, K. Arai, D. Korn, M.W. Hunkapiller, T.S. Wang, Human DNA polymerase alpha gene expression is cell proliferation dependent and its primary structure is similar to both prokaryotic and eukaryotic replicative DNA polymerases, *EMBO J.* 7 (1988) 37–47.
27. C.I. Amos, X. Wu, P. Broderick, I.P. Gorlov, J. Gu, T. Eisen, Q. Dong, Q. Zhang, X. Gu, J. Vijaykrishnan, K. Sullivan, A. Matakidou, Y. Wang, G. Mills, K. Doheny, Y.-Y. Tsai, W.V. Chen, S. Shete, M.R. Spitz, R.S. Houlston, Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1, *Nat. Genet.* 40 (2008) 616–622.
28. D.T. Bishop, F. Demenais, M.M. Iles, M. Harland, J.C. Taylor, E. Corda, J. Randerson-Moor, J.F. Aitken, M.-F. Avril, E. Azizi, B. Bakker, G. Bianchi-Scarra, B. Bressac-de Paillerets, D. Calista, L.A. Cannon-Albright, T. Chin-A-Woeng, T. Debnjak, G. Galore-Haskel, P. Ghiorzo, I. Gut, J. Hansson, M. Hecovar, V. Hoim, J.L. Hopper, C. Ingvar, P.A. Kanetsky, R.F. Kefford, M.T. Landi, J. Lang, J. Lubinski, R. Mackie, J. Malvey, G.J. Mann, N.G. Martin, G.W. Montgomery, F.A. van Nieuwpoort, S. Novakovic, H. Olsson, S. Puig, M. Weiss, W. van Workum, D. Zelenika, K.M. Brown, A.M. Goldstein, E.M. Gillanders, A. Boland, P. Galan, D.E. Elder, N.A. Gruis, N.K. Hayward, G.M. Lathrop, J.H. Barrett, J.A. Newton Bishop, Genome-wide association study identifies three loci associated with melanoma risk, *Nat. Genet.* 41 (2009) 920–925.
29. X. Wang, V.S. Pankratz, Z. Fredericksen, R. Tarrell, M. Karaus, L. McGuffog, P.D.P. Pharaoh, B.A.J. Ponder, A.M. Dunning, S. Peock, M. Cook, C. Oliver, D. Frost EMBRACE, O.M. Similnikova, D. Stoppa-Lyonnet, S. Mazoyer, C. Houdayer, G.E.M.O. Hogervorst, F.B.L. Hoening, M.J. Ligtenberg, M.J. HEBON, A. Spurdle, G. Chenevix-Trench kConFab, R.K. Schmutzler, B. Wappenschmidt, C. Engel, A. Meindl, S.M. Domchek, K.L. Nathanson, T.R. Rebbeck, C.F. Singer, D. Gschwanter-Kaulich, C. Dressler, A. Fink, C.I. Szabo, M. Zikan, L. Foretova, K. Claes, G. Thomas, R.N. Hoover, D.J. Hunter, S.J. Chanock, D.F. Easton, A.C. Antoniou, F.J. Couch, Common variants associated with breast cancer in genome-wide association studies are modifiers of breast cancer risk in BRCA1 and BRCA2 mutation carriers, *Hum. Mol. Genet.* 19 (2010) 2886–2897.
30. K.M. Waters, L. Le Marchand, L.N. Kolonel, K.R. Monroe, D.O. Stram, B.E. Henderson, C.A. Haiman, Generalizability of associations from prostate cancer genome-wide association studies in multiple populations, *Cancer Epidemiol. Biomark. Prev.* 18 (2009) 1285–1289.
31. B.A. Weir, M.S. Woo, G. Getz, S. Perner, L. Ding, R. Beroukhi, W.M. Lin, M.A. Province, A. Kraja, L.A. Johnson, K. Shah, M. Sato, R.K. Thomas, J.A. Barletta, I.B. Borecki, S. Broderick, A.C. Chang, D.Y. Chiang, L.R. Chirieac, J. Cho, Y. Fujii, A.F. Gazdar, T. Giordano, H. Greulich, M. Hanna, B.E. Johnson, M.G. Kris, A. Lash, L. Lin, N. Lindeman, E.R. Mardis, J.D. McPherson, J.D. Minna, M.B. Morgan, M. Nadel, M.B. Orringer, J.R. Osborne, B. Ozenberger, A.H. Ramos, J. Robinson, J.A. Roth, V. Rusch, H. Sasaki, F. Shepherd, C. Sougnez, M.R. Spitz, M.S. Tsao, D. Twomey, R.G. Verhaak, G.M. Weinstock, D.A. Wheeler, W. Winckler, A. Yoshizawa, S. Yu, M.F. Zakowski, Q. Zhang, D.G. Beer, I.I. Wistuba, M.A. Watson, L.A. Garraway, M. Ladanyi, W.D. Travis, W. Pao, M.A. Rubin, S.B. Gabriel, R.A. Gibbs, H.E. Varmus, R.K. Wilson, E.S. Lander, M. Meyerson, Characterizing the cancer genome in lung adenocarcinoma, *Nature* 450 (2007) 893–898.
32. M. Yeager, N. Orr, R.B. Hayes, K.B. Jacobs, P. Kraft, S. Wacholder, M.J. Minichiello, P. Fearnhead, K. Yu, N. Chatterjee, Z. Wang, R. Welch, B.J. Staats, E.E. Calle, H.S. Feigelson, M.J. Thun, C. Rodriguez, D. Albanes, J. Virtamo, S. Weinstein, F.R. Schumacher, E. Giovannucci, W.C. Willett, G. Cancel-Tassin, O. Cussenot, A. Valeri, G.L. Andriole, E.P. Gelmann, M. Tucker, D.S. Gerhard, J.F. Fraumeni, R. Hoover, D.J. Hunter, S.J. Chanock, G. Thomas, Genome-wide association study of prostate cancer identifies a second risk locus at 8q24, *Nat. Genet.* 39 (2007) 645–649.
33. J.P. Sinnwell, D.J. Schaid, Z. Yu, haplo.stats: Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous, R Package Version 1, 2005.
34. M. Bastian, S. Heymann, M. Jacomy Gephi, An Open Source Software for Exploring and Manipulating Networks, 2009.
35. T.M.J. Fruchterman, E.M. Reingold, Graph drawing by force-directed placement, *Software: Pract. Exp.* 21 (1991) 1129–1164.
36. J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kale, K. Schulten, Scalable molecular dynamics with NAMD, *J. Comput. Chem.* 26 (2005) 1781–1802.
37. A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kucera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkerich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, M. Karplus, All-atom empirical potential for molecular modeling and dynamics studies of proteins, *J. Phys. Chem. B* 102 (1998) 3586–3616.
38. H. Wang, F. Dommert, C. Holm, Optimizing working parameters of the smooth particle mesh Ewald algorithm in terms of accuracy and efficiency, *J. Chem. Phys.* 133 (2010) 034117.
39. S. Wolfram, A. Mathematica, System for Doing Mathematics by Computer, ISBN 0, Addison Wesley, 1991, pp. 201–19334.
40. K.A.S. Willy Wriggers, Y. Shan, S. Piana, P. Maragakis, P.J.M. Kresten Lindorff-Larsen, J. Gullingsrud, A. Charles Rendleman, R.O.D.P. Michael Eastwood, E. David Shaw, Automated event detection and activity monitoring in long time-scale molecular dynamics, *J. Chem. Theory Comput.* 5 (2009) 2595–2605.
41. W. Humphrey, A. Dalke, K.V.M.D. Schulten, Visual molecular dynamics, *J. Mol. Graph.* 14 (1996) 33–38.
42. P. Woodbridge, C. Liang, R.L. Davis, H. Vandebona, C.M. Sue, POLG mutations in Australian patients with mitochondrial disease, *Int. Med. J.* 43 (2) (2013) 150–156, <http://dx.doi.org/10.1111/j.1445-5994.2012.02847.x>.
43. K. Bebenek, M. Garcia-Diaz, R.Z. Zhou, L.F. Povirk, T.A. Kunkel, Loop 1 modulates the fidelity of DNA polymerase lambda, *Nucleic Acids Res.* 38 (2010) 5419–5431.
44. G. Terrados, J.-P. Capp, Y. Canitrot, M. Garcia-Diaz, K. Bebenek, T. Kirchhoff, A. Villanueva, F.O. Boudsocq, V.R. Bergoglio, C. Caizaux, T.A. Kunkel, J.-S.B. Hoffmann, L. Blanco, Characterization of a natural mutator variant of human DNA polymerase  $\lambda$  which promotes chromosomal instability by compromising NHEJ, *PLoS ONE* 4 (2009) e7290.

## APPENDIX A.

**Table S1.** Statistical results from applying HyDN-SNP-S to four cancer phenotypes.

Breast Cancer				Prostate Cancer				Lung Cancer				Melanoma Cancer			
Genotypic Genetic Model															
SNP	Gene	p.value	Location	SNP	Gene	p.value	Location	SNP	Gene	p.value	Location	SNP	Gene	p.value	Location
rs1292504	POLD3	0.006	5' UTR	rs10898995	POLD3	0.045	5' UTR	rs675450	POLD3	0.002	5' UTR	rs2908280	POLD2	0.021	intron
rs524051	POLN	0.047	intron	rs2351000	POLG	0.024	intron	rs1787091	POLD3	0.024	5' UTR	rs1945132	POLD3	0.022	3' UTR
rs3218651	POLQ	0.023	H1201R	rs3218651	POLQ	0.018	H1201R	rs645541	POLD3	0.019	5' UTR	rs1638566	POLD4	0.003	intron
								rs676537	POLD3	0.016	5' UTR	rs2514258	POLD4	0.005	intron
								rs10793091	POLD3	0.013	5' UTR	rs7951732	POLD4	0.013	intron
								rs10793092	POLD3	0.007	5' UTR	rs1790735	POLD4	0.002	intron
								rs1433972	POLD3	0.031	5' UTR	rs41549716	POLG	0.02	Y831C
								rs7113533	POLD3	0.031	5' UTR	rs3135056	POLN	0.039	intron
								rs7943085	POLD3	0.016	intron	rs1745335	POLN	0.029	intron
								rs10793095	POLD3	0.002	3' UTR	rs487848	POLQ	0.006	A581V
								rs3176175	POLG	0.013	intron	rs532411	POLQ	0.005	A2304V
								rs758130	POLG	0	intron	rs3218634	POLQ	0.004	L2538V/ L2538F
								rs6793252	POLQ	0.019	intron	rs702019	POLQ	0.045	intron
Recessive genetic model															
SNP	Gene	p.value	Location	SNP	Gene	p.value	Location	SNP	Gene	p.value	Location	SNP	Gene	p.value	Location
rs7113533	POLD3	0.05	5' UTR	rs10898995	POLD3	0.022	5' UTR	rs675450	POLD3	0.008	5' UTR	rs2908280	POLD2	0.007	intron
rs7932922	POLD3	0.038	3' UTR	rs1292504	POLD3	0.03	5' UTR	rs663016	POLD3	0.041	5' UTR	rs2887046	POLD3	0.028	no data
rs9328764	POLN	0.023	R425C	rs1638566	POLD4	0.033	intron	rs674306	POLD3	0.041	5' UTR	rs1945132	POLD3	0.022	no data
rs6830513	POLN	0.021	T378T	rs8305	POLI	0.044	A731T	rs10793091	POLD3	0.007	5' UTR	rs1638566	POLD4	0.017	intron

# HYPOTHESIS DRIVEN SNP SEARCH | R. J. SWETT ET AL.

rs10011549	POLN	0.036	G336S	rs3218651	POLQ	0.02	H1201R	rs1433972	POLD3	0.008	5' UTR	rs2514258	POLD4	0.023	intron
rs10018786	POLN	0.033	M310L					rs7943085	POLD3	0.019	intron	rs7951732	POLD4	0.013	intron
rs3117813	POLN	0.041	intron					rs10793095	POLD3	0.004	3' UTR	rs1790735	POLD4	0.025	intron
rs3117816	POLN	0.035	intron					rs6592579	POLD3	0.041	3' UTR	rs41549716	POLG	0.02	Y831C
rs3117819	POLN	0.042	intron					rs10219168	POLD3	3041	3' UTR	rs6941583	POLH	0.041	M647L
rs524051	POLN	0.015	intron					rs3176175	POLG	0.013	intron	rs6899628	POLH	0.039	3-UTR
rs1745335	POLN	0.049	intron					rs758130	POLG	0.008	intron	rs3734690	POLH	0.028	T478T
rs529966	POLN	0.042	intron					rs6793252	POLQ	0.019	intron	rs4640970	POLM	0.045	intron
rs618262	POLN	0.034	intron									rs3135056	POLN	0.025	intron
rs7659386	POLN	0.029	intron									rs6800901	POLQ	0.018	intron
rs3218651	POLQ	0.006	H1201R									rs5744990	POLE	0.036	L1903L

## Additive genetic model

SNP	Gene	p.value	Location	SNP	Gene	p.value	Location	SNP	Gene	p.value	Location	SNP	Gene	p.value	Location
rs7113533	POLD3	0.023	5' UTR	rs2075070	POLD2	0.037	intron	rs10793091	POLD3	0.028	5' UTR	rs1945132	POLD3	0.022	intron
rs9328764	POLN	0.018	R425C	rs10898995	POLD3	0.014	5' UTR	rs10973092	POLD3	0.007	5' UTR	rs7951732	POLD4	0.013	intron
rs6830513	POLN	0.016	T378T	rs1292504	POLD3	0.035	5' UTR	rs1433972	POLD3	0.008	5' UTR	rs41549716	POLG	0.02	Y831C
rs10011549	POLN	0.027	G336S	rs1638566	POLD4	0.044	intron	rs7943085	POLD3	0.004	intron	rs6941583	POLH	0.045	M647L
rs10018786	POLN	0.026	M310L	rs2351000	POLG	0.033	intron	rs6592581	POLD3	0.024	3' UTR	rs6899628	POLH	0.043	3-UTR
rs3117813	POLN	0.041	intron					rs3176175	POLG	0.013	intron	rs3734690	POLH	0.031	T478T
rs1923775	POLN	0.019	intron					rs6793252	POLQ	0.019	intron	rs3135056	POLN	0.046	intron
rs524051	POLN	0.025	intron					rs574316	POLN	0.041	intron	rs6800901	POLQ	0.037	intron
rs1745335	POLN	0.038	intron									rs702019	POLQ	0.018	intron
rs529966	POLN	0.032	intron									rs5744990	POLE	0.031	L1903L
rs618262	POLN	0.027	intron												
rs7659386	POLN	0.022	intron												
rs6599418	POLN	0.046	intron												

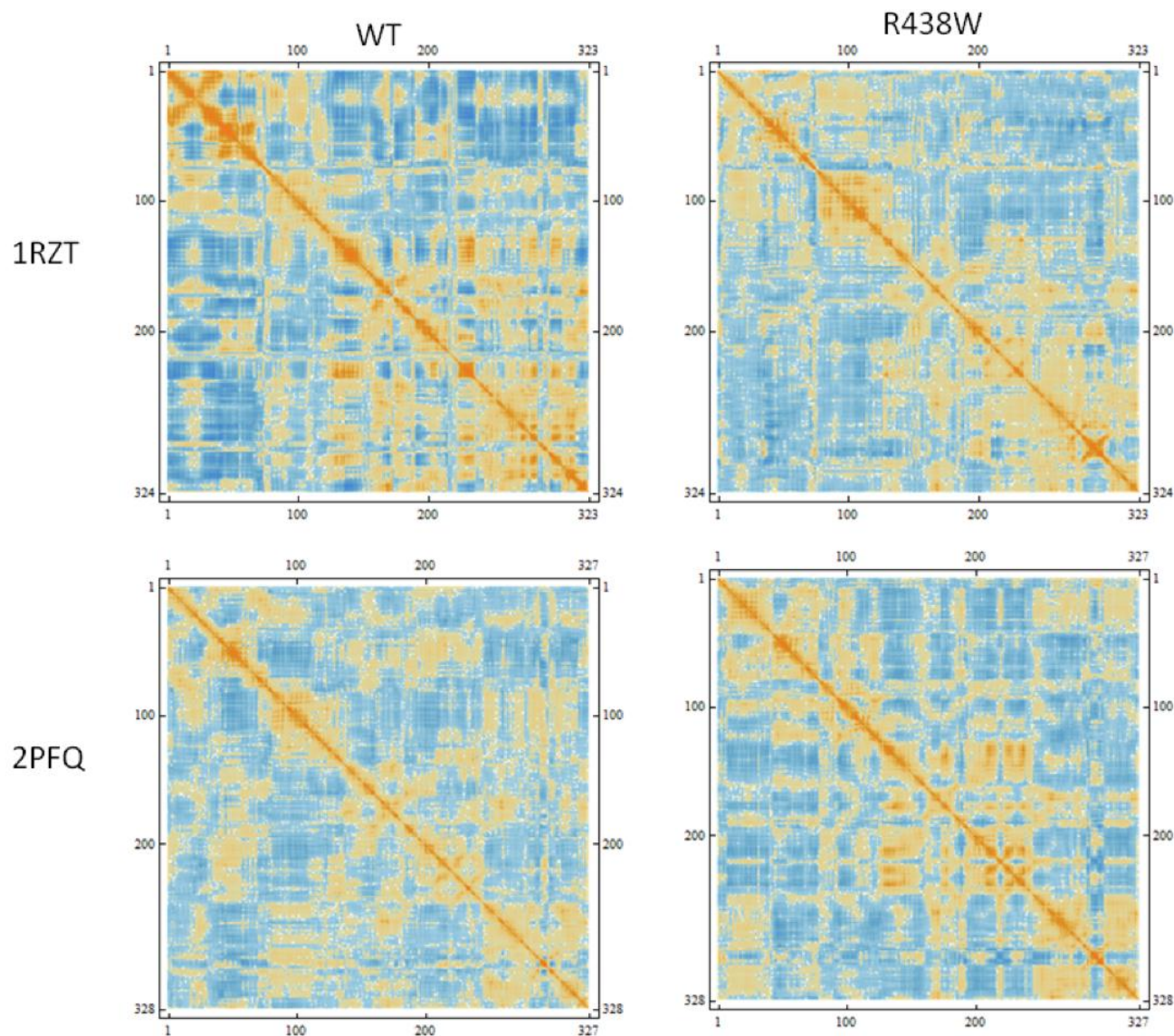


# HYPOTHESIS DRIVEN SNP SEARCH | R. J. SWETT ET AL.

rs3218651	POLQ	0.007	H1201R			
rs3730463	POLL	0.047	T221P			

## Dominant genetic model

SNP	Gene	p.value	Location	SNP	Gene	p.value	Location	SNP	Gene	p.value	Location	SNP	Gene	p.value	Location
rs1292504	POLD3	0.002	5' UTR	rs1673041	POLD1	0.029	intron	rs675450	POLD3	0.019	5' UTR	rs1638566	POLD4	0.049	intron
rs3730814	POLI	0.037	intron	rs2075070	POLD2	0.042	intron	rs1787091	POLD3	0.019	5' UTR	rs2514258	POLD4	0.049	intron
rs1923775	POLN	0.049	intron	rs3824999	POLD3	0.018	intron	rs645541	POLD3	0.008	5' UTR	rs1790735	POLD4	0.012	5' UTR
rs3821367	POLQ	0.018	intron	rs2351000	POLG	0.014	intron	rs663555	POLD3	0.041	5' UTR	rs3218784	POLI	0.041	I261M
rs3911713	POLQ	0.047	intron	rs1427463	POLG2	0.018	A169T	rs676537	POLD3	0.019	5' UTR	rs9328764	POLN	0.03	R425C
				rs607877	POLI	0.032	intron	rs10793082	POLD3	0.002	5' UTR	rs6830513	POLN	0.03	T378T
								rs7113533	POLD3	0.041	5' UTR	rs10011549	POLN	0.03	G336S
								rs7943085	POLD3	0.041	intron	rs10018786	POLN	0.03	M310L
								rs10793095	POLD3	0.041	3' UTR	rs2022302	POLN	0.03	H441H
								rs758130	POLG	0	intron	rs1745335	POLN	0.009	intron
												rs487848	POLQ	0.008	A581V
												rs532411	POLQ	0.008	A2304V
												rs3218634	POLQ	0.008	L2538V/ L2538
												rs702019	POLQ	0.035	intron



**Figure S1.** Correlation plots from the four Polymerase Lambda simulations. These original correlation plots were used to generate the difference plots shown in [Fig. 4](#).

**Movie S1.** An animation of a mathematical interpolation between the binary and ternary conformations is shown. Loop 1 is colored purple, and the site of mutation R438 is shown in red and labeled. Within the conformational space available to Loop 1 visualized by this interpolation, the mutation is between 12Å and 14Å away.