

Prioritizing Regions of Candidate Genes for Mutation Screening

Running title: Prioritizing Regions of Candidate Genes

Terry Braun^{1,2,4,5}, Suma Shankar², Steve Davis⁴, Brian O’Leary⁴, Todd Scheetz^{2,4,5}, Val Sheffield^{3,4,5,6}, Thomas Casavant^{1,2,4,5}, Edwin Stone^{2,4,5,6}

¹Department of Biomedical Engineering, ²Department of Ophthalmology, ³Department of Pediatrics, ⁴Center for Bioinformatics and Computational Biology, ⁵Center for Macular Degeneration, ⁶Howard Hughes Medical Institute. The University of Iowa, Iowa City, Iowa, USA. 52242.

Corresponding Author:

Edwin Stone
11190G PFP
University of Iowa
Iowa City, IA 52240
319-335-8270
FAX: 319-335-7142
Edwin-stone@uiowa.edu

Summary

The process of disease gene discovery often requires one to search regions of the genome of a large number of patients and controls for disease causing mutations. The phenotype-altering variations that have been recognized to date are not uniformly distributed in the genome, in genes, or in the population. Thus, in most experiments designed to find these variations, the resources that are consumed are not related to the likelihood of finding a phenotypically meaningful variation in a linear fashion. We hypothesized that our expanding knowledge of conserved protein motifs and functional domains could be used to predict regions of genes that would be more likely to harbor phenotype-altering variations. A bioinformatic technique was developed to prioritize regions of genes for screening. The Prioritization of Annotated Regions (PAR) technique utilizes conserved protein functional domains and protein secondary structures to predict the likelihood that a specific coding region of a gene will harbor a disease-causing mutation. The PAR technique was applied to 710 genes for which 4,498 previously identified mutations were known. Using regions 200 nucleotides in size (approximating the optimal amplicon size in an SSCP assay, excluding primers), 819 mutations were identified in 350 genes. This corresponds to a correct identification of 18% of the mutations in approximately 50% of the transcripts. More importantly, of the 1,908,911 nucleotides comprising the 710 genes, the PAR technique prioritized and selected 168,980 nucleotides for screening, which represents a 91% reduction of screening resources compared to a comprehensive screening approach. These results suggest that prioritization strategies such as PAR can accelerate the success of disease gene identification by more efficiently utilizing screening resources.

Introduction

Efforts to identify disease-causing sequence variations in humans have usually focused on the transcriptional units of the genome because it is often possible to demonstrate or infer a functional effect of such variations on the expression or regulation of a gene product. Inference of pathogenicity of these variations requires a statistically significant association between the variations and the disease phenotype. Ultimately, the challenge of finding the genetic causes of disease is to search the genomes of a sufficient number of patients and controls for evidence of variations that have a statistically significant impact on the phenotype.

The fundamental experiment used to identify sequence variations is PCR-based amplification of genomic DNA followed by an analysis of the PCR product by various methods including detection of sequence-dependent conformational changes of DNA molecules in solution (e.g., SSCP (Orita et. Al. 1998)) and automated DNA sequencing (Korkko et. al. 1998). As a result, the fundamental unit of such an experiment is the amplification and analysis of a genomic segment from a single individual, with one pair of primers, from one gene -- that is, one *amplimer*. The *search space* of this experimental approach defined in terms of this quantum has the following three dimensions: the number of *subjects*, the number of *genes*, and the number of *amplimers* analyzed within each gene.

Screening Dilemma

The screening dilemma faced by investigators is that the search space needs to be sufficiently large to ensure that the variations of interest are located, but not so large that these variations cannot be found with a realistic expenditure of time and reagents. Ideally, every exon of every candidate gene would be screened in a large set of affected subjects, as well as a matched set of control subjects. However, the multi-dimensional nature of the search space

requires that most experiments be constrained to a relatively small number in at least one dimension to make disease gene identification feasible.

There have been several attempts to identify correlations between sequence variations, genes, and phenotypes. Disease-causing mutations are more likely to lie in structural and functional regions of genes, and a significant fraction of SNPs have been observed in these locations (Sunyaev et. al. 2000). Analyses of protein sequences have been used to predict whether a substitution affects protein function (Ng and Henikoff 2001). It is estimated that approximately 20% of non-synonymous SNPs damage the structure or function of proteins based on the impact of an amino acid substitution on the three-dimensional structure (Sunyaev et. al. 2001). Mirny and Gelfand combined information about protein-DNA complexes and sites recognized by DNA-binding proteins and found a significant correlation between site conservation and nucleotide to protein contact (Mirny and Gelfand 2002).

Other groups have used phenotype, MeshD terms, and MEDLINE articles (Perez-Iratxeta et. al. 2002) to prioritize candidate disease genes. Freudenberg and Propping prioritized candidates by clustering genes with known disease genes using Gene Ontology (GO) terms based on the assumption that similar phenotypes are likely to be caused by similar disease mechanisms (Freudenberg and Propping 2002). Goodstadt and Ponting analyzed known missense mutations and revealed more variation than expected compared to those predicted by a PAM1 evolutionary model (Goodstadt and Ponting 2002). They also reported that 91% of the known disease gene products listed in SwissProt contain a domain in Pfam or SMART. However, no determination of bias between mutations and functional domains was evaluated. Finally, a functional classification of known disease genes based on protein products has also been performed (Jimenez-Sanchez et. al. 2001).

In this study, we explore whether the third dimension of *amplimers* -- representing specific areas of interest within individual genes -- can be prioritized for mutation analysis to increase the discovery of phenotype-altering variations per unit of screening effort. If so, this would have the benefit of increasing the efficiency of disease gene identification.

We hypothesized that emerging knowledge of conserved protein motifs and functional domains can be used to predict regions of genes that would be more likely to harbor phenotype-altering variations. To test this hypothesis, we developed a bioinformatic technique for prioritizing regions of genes for screening. The Prioritization of Annotated Regions (PAR) technique utilizes conserved protein functional domains, and protein secondary structures, to predict whether specific coding regions in genes will harbor phenotype altering variants. We then quantified the benefit of the PAR technique and compared it to traditional approaches.

Materials and Methods

A retrospective study of genes with previously identified variations was performed. A list of 710 genes was obtained by considering all genes from the Online Mendelian Inheritance in Man database (OMIM) (McKusick 1998) and cross-referencing these with transcripts in Ensembl Release NCBI31 (Hubbard et. al. 2002). Only genes with single transcripts were included. This simplified the analysis, because mutations in OMIM are not clearly localized within genes containing multiple transcripts (i.e., alternatively processed or spliced). The gene structure for each gene was obtained from Ensembl. This included 10 kilobases of 5' flanking sequence, all exons, all introns, and 10 kilobases of 3' flanking sequence. Genes were uniquely identified by RefSeq numbers in Ensembl, and by a RefSeq to OMIM number mapping provided by NCBI. Annotated protein domains for each transcript were also acquired from Ensembl.

Secondary structure prediction was performed by NNpredict (Kneller 1990). The PAR value at every nucleotide for all 710 genes was calculated, where PAR is defined as the discrete convolution:

$$PAR(x) = \sum_{i=-\frac{W_s}{2}}^{\frac{W_s}{2}} W(i) \left(\sum_{j=1}^{N_x} A_f(x, j) (A_s(x, j) + A_o(x, j)) \cdot A_m(j) \right) \quad \text{Equation 1 – The PAR}$$

discrete convolution

where

x = nucleotide position

W_s = PAR window size

N_x = number of distinct annotation elements at nucleotide position x

$W(i)$ = PAR window function

$A_f(x, j)$ = annotation function for j th annotation feature at the x th position

$A_s(x, j)$ = annotation score for the j th annotation feature at the x th position

$A_o(x, j)$ = annotation scalar offset

$A_m(j)$ = annotation multiplier for j th annotation feature

In Equation 1, x represents the x th base pair in the sequence for the gene. N_x represents the number of annotation elements that exist at the x th base pair in the sequence for the gene.

The PAR window function $W(i)$ models typical amplicon product sizes with an arbitrary function – a square function was used. The PAR window has a size of W_s and typically ranges from 100 bases to 300 bases – the range of amplicon sizes. The notation $A(x, j)$ means the j th annotation element at the x th base pair in the gene sequence. This equation is calculated for all base pairs in the gene sequence to determine the PAR values. This equation expresses a discrete

convolution of the PAR window function across the functions representing all annotation features at every base pair in the sequence of a gene. Each sequence feature is represented by an arbitrary normal annotation function A_f . This convolution is a calculation that would not be practical to perform manually for every gene in a candidate set. Figure 1 shows a graphical representation of the PAR equation for the *EFEMP1* gene and its associated annotation.

The 25 different combinations of the PAR parameters (parameter sets) that were used in this study are summarized in Table 1. A constant weight A_m of 1.0 was used for all annotation features. For functional domains, the score parameter A_s was determined from the percentage of identity normalized to a scale of 0.0 to 1.0. The discrete convolution rewards larger domains and domains of higher sequence similarity. For the secondary structures, the score parameter A_s of 0.7 was used, based on an expected accuracy rate of 0.7 for NNpredict (Kneller, 1990).

To reduce the number of computations, only parameter sets 1-6 were used when the number of primer regions (PAR regions) for all 710 genes exceeded 10. This reduced the run time from several weeks to several days. Since any strategy will find the majority of mutations when most of the gene is screened, the performance of PAR is not relevant for larger numbers of amplimers per gene (i.e. approaching 100% coverage or resources). However, the performance of PAR is relevant for relatively few amplimers per gene (approaching 0% coverage or resources). Therefore, the PAR function was examined for additional parameter sets (sets 7-25) with 10 or fewer PAR regions.

Using the PAR values, regions in each transcript were identified that maximized the PAR function. For each gene, primer pair positions were then selected consistent with the default parameters for Primer3 (Rozen 2000) centered in the maximum PAR regions until at least one mutation was flanked.

Other strategies were also used to select regions of transcripts for comparisons to the PAR technique. The *serial* selection technique generates minimally overlapping primer pair positions for each exon with the same amplicon size requirements as the PAR technique. This technique, used for comparison, models traditional screening approaches that exhaustively examine the complete coding sequence of each candidate gene. This technique provides a conservative estimate for comparisons since all screening resources must be consumed to identify 100% of mutations in all of the genes. Another comparison technique called *random* selects a random position within the transcript to be analyzed for mutations. Additional minimally overlapping random positions are selected until at least one mutation is flanked.

Results

The PAR technique was applied retrospectively to 710 genes for which previously identified mutations were known. This set of genes contained a total of 4,097 known mutations. To measure the effectiveness of PAR, two alternative primer position selection strategies were also applied to the same set of transcripts. The first alternative strategy utilizes a naive *serial* technique that selected minimally overlapping regions to be amplified beginning at the 5' end of a transcript and proceeded linearly to the 3' end. Selection was performed such that there would be 100% coverage of the coding nucleotides by flanking primer positions. The second alternative strategy employed was a *random* approach in which regions were selected from random positions within the transcripts with the same product size requirements of the other strategies. Figure 2 compares the three strategies with regard to the number of mutations identified versus the percent coverage of the entire transcript (exon regions). As expected, the *serial* strategy identifies all 4,097 mutations when 100% of the transcripts were screened, and 90% of the

mutations (3,687) when 90% of the transcripts were screened. In comparison, the PAR technique identified 90% of the mutations (3,678) while **screening** only 66% of the transcript regions. The *random* technique falls short of identifying 100% of the mutations. This occurs because one of the parameters for primer selection is that amplified products may not overlap by more than 10 nucleotides (to avoid amplimers that are mostly redundant). Since *random* position selection cannot ensure that both the entire gene is covered and that the overlap criterion between adjacent positions is met, it would be expected that less than 100% coverage would be achieved.

The PAR technique was applied to the 710 genes using the 25 different parameter sets (see Methods). The average from all 25 experiments was found and these results are graphed in Figure 2 and Figure 3. With the PAR strategy, 819 mutations were identified in 350 distinct genes using a single, best, PAR-selected region per gene. This corresponds to 18% of the mutations identified in approximately half of the transcripts. More importantly, of the 1,908,911 nucleotides represented by the 710 genes, the PAR strategy prioritized and selected 168,980 nucleotides. At least one mutation was identified in 50% of the genes while screening only 9% of the total transcript size. Figure 2 shows that the number of sequence variants (mutations) that are identified is accelerated by PAR with respect to the other strategies.

Figure 3 shows the number of genes containing phenotype altering variations that were identified versus the consumption of screening resources (shown as the “Coverage Percentage”) for the three strategies. The *serial* technique shows a linear relationship between screening resource utilization and the number of genes identified to contain a mutation. The PAR technique is attractive since it is able to identify 90% of the genes as containing at least one mutation with a 60% reduction in the screening resources (amplimer coverage) required. It is also important to note that the results of the PAR technique begin relatively high on the vertical

axis (number of genes identified). This illustrates that after selecting only one primer pair in each transcript with the PAR technique, nearly 40% of the transcripts were found to contain at least one mutation by flanking primer pairs. The comparison of PAR with the *random* technique provides a more conservative evaluation of performance since the *serial* technique essentially wastes resources to provide the illusion of complete coverage. A comparison of the PAR and *random* techniques also provides some measure of the importance of the annotation used in the PAR technique. It is apparent from Figure 3 that the incorporation of annotation by the PAR technique provides valuable information for improving mutation detection for the purpose of disease gene identification by more efficiently utilizing screening resources. The PAR technique is more efficient compared to the other techniques at directing the examination of the search space. The data clearly show an advantage in using the PAR technique to select screening regions across genes.

Discussion

The PAR technique for prioritizing sub-regions of genes utilizes sequence features and annotation to focus screening resources within the search space. This analysis suggests that the PAR technique enables efficient and effective utilization of available sequence-based annotation to identify regions of genes that are most likely to harbor disease-causing mutations. On the surface, the PAR technique appears to exchange sensitivity (the ability to comprehensively detect mutations for an entire gene) for the ability to look for variations in a larger search space. There is a sense of incompleteness since only portions of genes are covered in an effort to distribute the screening coverage across more genes. However, “completeness” of screening is an illusion at best given the current method of screening primarily the coding regions of genes.

Innumerable functionally important regions go unscreened simply because their locations are not yet known. Examples may include non-coding RNA genes (Lau et. al. 2001), non-coding regions such as promoter elements, locus control regions (Nathans, et. al. 1989), transcription factor binding sites, and splice enhancer sites (Majewski 2002). An example of the functional potential for non-coding regions is the locus control region of the opsin gene cluster (Nathans, et. al. 1989) shown to cause 50% of the cases of blue cone monochromacy. The locus control region is approximately 4 kilobases upstream of the red opsin gene, and 43 kilobases upstream of the green opsin gene. The 579 base region was mapped to the X-chromosome using observed deletions upstream of the red-green opsin gene cluster in individuals with blue cone monochromacy. We believe an approach similar to PAR can be applied to analogous non-coding sequence features to prioritize candidate genes based on non-coding sequence features.

It is attractive to devise prioritization strategies, such as PAR, because known disease genes may be used to evaluate and improve variation prioritization and detection strategies. For example, the gene associated with cystic fibrosis (CF) has been known since 1989 (Riordan, et. al. 1989). It has been reported (Kerem, et al. 1989) that approximately 70% of the mutations in CF patients correspond to a specific deletion of 3 base pairs at amino acid position 508 of the putative product of the CF gene. In 1982, (Higgins, et. al. 1982) cloned and sequenced the first ATP-binding cassette (ABC) transporter, histidine permease. If we assume that sufficient evidence to identify ABC transporter domains in genes, as is the **situation** today, then in 1989 the basic strategy of PAR would have been successful in guiding the sequence variation discovery process to the most frequent mutation.

However, caution should be used when examining specific examples. A counter example to CF would include Huntington disease (HD) (Huntington's Disease Collaborative Research

Group, 1993), where the molecular cause was determined to be a trinucleotide repeat expansion. Since the PAR technique focuses on conserved functional domains and other sequences features, clearly the current implementation of PAR would not provide any benefit in locating the HD sequence variation. However, we believe the PAR approach is flexible enough to integrate other types of annotation including, but not limited to: repeats, dinucleotides and codon bias, three-dimensional structural data, GC content, and other genomic features. This could also be expanded to non-coding sequences such as repetitive elements, regulatory features, expression-based data, and pathway data.

The traditional approach to screening genes by sequentially screening coding nucleotides in contiguous order from 5' to 3' is primarily an artifact of data management convenience. Until very recently, there was tremendous overhead in obtaining the necessary data describing the genomic context of a gene to be screened for mutations. Traditional strategies for screening genes for mutations would almost necessitate complete examination of a transcript to recoup the time and effort required to obtain the gene structure and genomic context. The results presented here suggest quantitative prioritization techniques such as PAR can accelerate the success of mutation identification for the purpose of disease gene discovery by more efficiently utilizing screening resources. The challenges of finding sequence variations and inferring pathogenicity for complex traits is a motivation for developing quantitative techniques such as PAR to predict mutation potential.

Acknowledgements

We would like to acknowledge Hakeem Almabrazi, Bart Brown, John Ritchison, Rhett Sutphin, Matt Kemp, and Jason Grundstadt for their contributions in implementing the software and tools that made this work possible. Also, the experience and assistance provided by the members of the Stone lab from Jean Andorf, Heidi Sahr, and Paula Moore was invaluable.

References

Freudenberg J, Propping P. "A similarity-based method for genome-wide prediction of disease-relevant human genes." *Bioinformatics*, 2002, 18, S2,S110-S115.

Goodstadt L, Ponting C. "Sequence variation and disease in the wake of the draft human genome." *Human Mol Genet.* 2001, 10,20:2209-2214.

Higgins C, Haag P, Nikaido K, Ardeshir F, Garcia G, Ames G. "Complete nucleotide sequence and identification of membrane components of the histidine transport operon of *S. typhimurium*." *Nature* 298 (1982), pp. 723–727.

Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, et. al. "The Ensembl genome database project." *Nucleic Acids Res.* 2002 Jan 1;30(1):38-41.

Huntington's Disease Collaborative Research Group. "A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes." *Cell* 72: 971-983, 1993.

International Human Genome Sequencing Consortium. "Initial sequencing and analysis of the human genome." *Nature* 409, 860 - 921 (2001)

Jimenez-Sanchez G, Childs B, Valle D. "Human disease genes." *Nature* 2001, 409:853-855.

Kerem B, Buchanan J, Durie P, Corey M, Levison H, Rommens J, Buchwald M, Tsui L-C, “DNA marker haplotype association with pancreatic sufficiency in cystic fibrosis.” *Am J Hum Genet.* 44: 827-834, 1989.

Kneller DG, Cohen FE, Langridge R. “Improvements in protein secondary structure prediction by an enhanced neural network.” *J Mol Biol.* 1990 Jul 5;214(1):171-82.

Korkko J, Annunen S, Pihlajamaa T, Prockop DJ, Ala-Kokko L. “Conformation sensitive gel electrophoresis for simple and accurate detection of mutations: comparison with denaturing gradient gel electrophoresis and nucleotide sequencing.” *Proc Natl Acad Sci. U S A.* 1998 Feb 17;95(4):1681-5.

Lau NC, Lim LP, Weinstein EG, Bartel DP. “An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*.” *Science*, 2001 Oct 26;294(5543):858-62

Majewski J, Ott J. “Distribution and Characterization of Regulatory Elements in the Human Genome.” *Genome Res.*, 2002. 12:1827-1836.

McKusick, VA. Mendelian Inheritance in Man. [A Catalog of Human Genes and Genetic Disorders](#). Baltimore: Johns Hopkins University Press, 1998 (12th edition).

Mirney L, Gelfand M. “Structural analysis of conserved base pairs in protein-DNA complexes.” *Nucleic Acids Res.*, 2002, 30,7:1704-1711.

Nathans J, Davenport C, Maumenenee I, Lewis R, Hejtmancik J, Litt M, Lovrien E, Weleber R, Bachynski B, Zwas F, Klingaman R, Fishman G. "Molecular Genetics of Human Blue Cone Monochromacy." *Science*, 1989 Aug 25;245(4920):831-383.

Ng P, Henikoff S. "Predicting Deleterious Amino Acid Substitutions." *Genome Res.*, 2001, 11:863-874.

Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T. "Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms." *Proc Natl Acad Sci. USA.* 1989 Apr;86(8):2766-70.

Riordan J, Rommens J, Kerem B, Alon N, Rozmahel R, Grzelczak Z, Zielenski J, Lok S, Plavsic N, Chou J, Drumm M, Iannuzzi M, Collins F, Tsui L-C. "Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA." *Science* 245: 1066-1073, 1989.

Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, 2000, pp 365-386.

Sunyaev S, Ramensky V, Bork P. "Towards a structural bases of human non-synonymous single nucleotide polymorphisms." *Trends Genet.* 2000,16:198-200.

Sunyaev S, Ramensky V, Koch I, Lathe III W, Kondrashov A, Bork P. "Prediction of deleterious human alleles." *Hum Mol Genet.* 2001, 15;10(6):591-7.

Table 1: The 25 experiments with unique PAR parameter sets.

Experiment	<i>PAR</i>			
	<i>window</i> $W(i)$	A_o (domains)	A_o (alpha- helix)	A_o (beta- sheets)
1	300	0.9	0.3	0.3
2	300	0.9	0.1	0.1
3	200	0.9	0.3	0.3
4	200	0.9	0.1	0.1
5	100	0.9	0.3	0.3
6	100	0.9	0.4	0.4
7	300	0.8	0.3	0.3
8	300	0.7	0.3	0.3
9	300	0.8	0.2	0.2
10	300	0.8	0.4	0.4
11	300	0.7	0.2	0.2
12	300	0.7	0.4	0.4
13	200	0.8	0.3	0.3
14	200	0.7	0.3	0.3
15	200	0.8	0.2	0.2
16	200	0.8	0.4	0.4
17	200	0.7	0.2	0.2
18	200	0.7	0.4	0.4
19	100	0.8	0.3	0.3
20	100	0.7	0.3	0.3
21	100	0.8	0.2	0.2
22	100	0.8	0.4	0.4
23	100	0.7	0.2	0.2
24	100	0.7	0.4	0.4
25	100	0.9	0.1	0.1

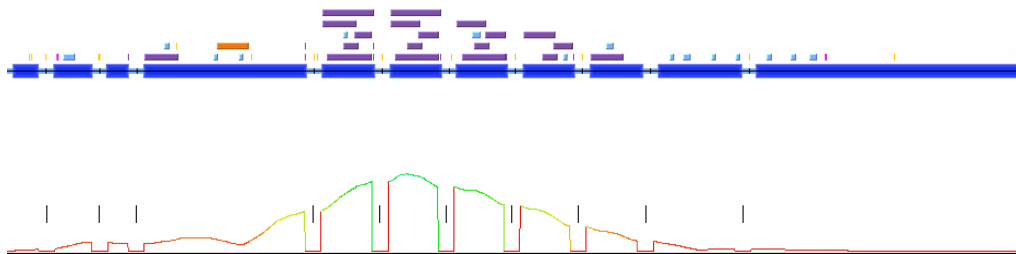


Figure 1: Gene structure and annotation for *EFEMP1* gene with corresponding graph of PAR values. Exons are illustrated as thick blue bars, functional domains are purple and secondary structures are blue. The intron sequences are truncated to a uniform length, and are not included in the PAR calculation for this graph.

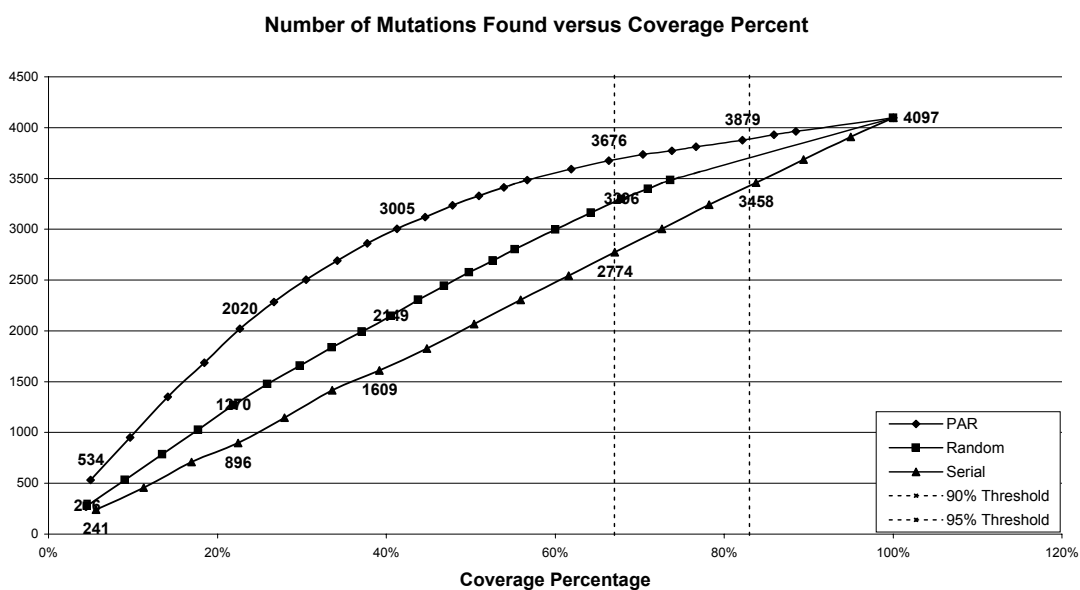


Figure 2: Number of non-redundant mutations found using PAR relative to random and serial strategies.

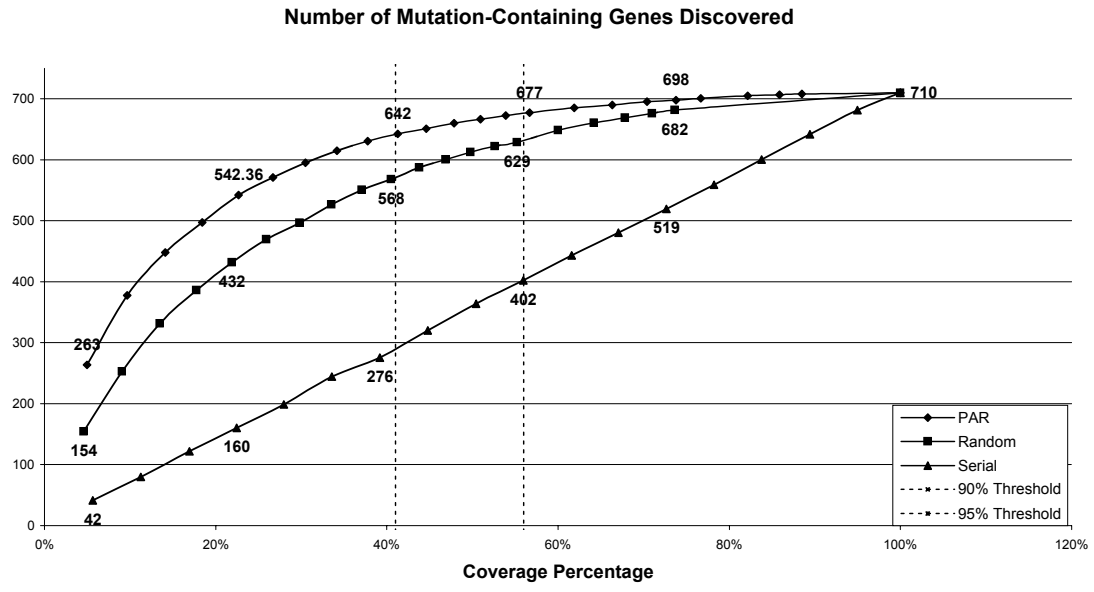


Figure 3: Number of Genes identified contain mutations by PAR, serial, and random techniques.