The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009

Cross-species Comparison for Identifying Orthologous Simple Sequence Repeats of Disease Genes

Chih-Chia Chen[1] Tsan-Huang Shih[1] Chien-Ming Chen[1]
Chin-Hua Hu[2] Wen-Shyong Tzou[2] Tun-Wen Pai[1,*]

[1] Department of Computer Science and Engineering
[2] Institute of Bioscience and Biotechnology, National Taiwan Ocean University, Taiwan

Abstract Simple sequence repeats (SSRs) have been demonstrated to affect normal gene function to cause different genetic disorders. Several conserved and even partial functional SSR patterns were discovered in inherited orthologous disease genes. To explore a wide range of SSRs in genetic diseases, a system focuses on orthologous SSRs for disease genes through comparative genomics mechanism is constructed in this research. The system is developed by employing the OMIM (Online Mendelian Inheritance in Man) and the NCBI HomoloGene databases as the resources of human genetic diseases and homologous gene information. In addition, The Comparative Genomics for SSR discovery system (CS-SSR) is also adopted for providing annotated SSR information among various model species. By integrating these data resources and data mining technologies, biologists and doctors can retrieve novel and important conserved SSRs information among orthologous disease genes. The proposed system named as Orthologous SSR for Disease Gene (OSDG) is a comprehensive andefficient online tool for discovering conserved SSRs of disease genes and it is freely available at http://osdg.cs.ntou.edu.tw.

Keywords simple sequence repeat (SSR), Online Mendelian Inheritance in Man (OMIM), HomoloGene. comparative genomics, genetic diseases

1 Introduction

Recent studies have pointed out that repeat number variation of simple sequence repeats (SSRs) affects various genetic diseases. Several important observations of the correlation between disease phenotype and genetic heterogeneity suggest the importance of discovering conserved SSR motifs among various species. Hence, this study takes the aim of developing novel methods to identify putative functional SSRs through cross-species comparison and to discover inheritable genetic disease which are associated with the mutation of repeat tracts.

* Email: twp@mail.ntou.edu.tw

Genetic diseases can be inherited from parents and are mutations in the germ cells of the body, which involved in passing genetic information on to the next generation. The mutations of genes cause varying effects on health, depending on where they take places and whether they amend the function of crucial proteins. Four types of genetic disorders are defined in general, including single gene, multifactorial, chromosomal, and mitochondrial disorders. The single gene disorder is defined as the defects appeared in one particular gene, and often possess simple inheritance patterns. So far, more than 4,000 single gene disorders are known, such as cystic fibrosis, sickle cell anemia, and Huntington's disease. Up to date, the types of mutations include: (1) Insertion: An insertion varies the total number of DNA bases by tucking a piece of DNA in a gene. (2) Deletion: A deletion changes the number of DNA bases by eliminating a piece of DNA. (3) Duplication: A duplication consists of a segment of DNA that is abnormally copied one or more times. (4) Frame shift: A frame shift mutation causes by the addition or loss of DNA bases and changes a gene's reading frame. (5) Repeat expansion: A mutation increases the number of times of repeated short DNA sequence. All these various types of mutations bring into results of malfunctioning proteins. In this paper, we focus on single gene disorder and try to find out the relationship between a single gene diseases and genetic mutations in repeat segments.

Simple sequence repeats (SSRs, also called microsatellites or short tandem repeat) are sequences of repeated nucleotides in DNA consisting of a continuous repeated core pattern which range in length from 1 to 9 base pairs (bps) and the whole repetitive region spans less than 150 bps. Evidences show that a large number of SSRs are interspersed throughout vertebrate genomes and eukaryotic genomes [1][2][3]. Within genetic and inter-genetic regions of whole human genomes, one-tenth of the nucleotide sequences were proved to contain SSRs [4]. The mutations of SSRs appear to follow a stepwise mutation model (SMM) [5] and make SSRs become not only numerous but also polymorphous [6].

The polymorphous property of repeating length was employed to emphasis the SSRs as biomarkers in the fields of evolution and DNA fingerprinting. Nowadays, SSRs are not just applied to map genes as biomarkers, they play a critical role in gene regulation, transcription and protein function. For example, SSR variations in 5'UTRs could affect gene transcription and translation, whereas SSR expansions in 3'UTRs could cause transcription slippage and result in disrupting splicing and possibly disturbing cellular functions. If SSRs in introns, they can affect gene transcription, mRNA splicing, or export to the cytoplasm [7]. Several kinds of genetic diseases causing a number of genetic disorders are mainly due to the change of repeat numbers that may appear in the coding regions (exons) of disease genes [8]. That is to say the length polymorphisms can affect fundamental functions on human inherited genes and possibly cause genetic diseases [9][10]. There are a number of well known genetic diseases which are caused by SSRs polymorphism [11], such as Fragile X mental retardationl [12], Huntington's disease [13], *etc.*. Furthermore, microsatellite instability (MIS) cause SSR polymorphism has been proved to be the main reason that affects the DNA repair progresses and makes cancer cells to spread [14][15].

Because genetic diseases among animals provided useful information about

human counterpart disorder, it is essential to proceed with cross-species comparison to discover conserved SSRs among different homologous disease genes from particular model species. In the field of comparative genomics, homology represents the similarity originating from common ancestor, and orthology and paralogy are two major subtypes of homology. If similar genes were separated by a speciation event from a common ancestor, the genes are claimed as orthologous genest, and which genes may have the same function and even encode homologous phenotypic traits like proteins or enzymes. If they were separated by a gene duplication event in a species, the genes are called paralogous of each other [16][17]. Through the comparison of orthologous genes with probable SSRs in various model organisms, we can categorize genes from different organisms and absorb much information as possible with respect to annotations of novel conserved SSRs within orthologous genes.

The proposed system was developed by employing the OMIM (Online Mendelian Inheritance in Man) [18] and the HomoloGene databases [19] for the resources of human genetic diseases and homologous gene information. In addition, we developed efficient SSR searching algorithms for identifying prefect/imperfect SSR patterns from various genomes and constitute a significant SSR database through cross-species comparison [20].

The survey of previous work indicated that there were limited systems and databases being built for discovering the relationship between the SSRs and diseases. The TRbase system was built by Body T. *et al*., which provides all tandem repeats within disease genes of the human genome. Although TRbase system provides comprehensive tandem repeat information, it is still difficult for users to identify unknown regulators from tremendous amount of tandem repeat records [21]. On the other hand, OrthoDisease system provides the orthologous information of human genetic disease, but the SSR information of genetic diseases was not designed to present [22]. In order to provide an integrated and comprehensive system for discovering conserved SSRs from orthologous disease genes, we have combined the advantageous features of these two systems and adopted efficient SSRs retrieving technologies to establish a new online web system which is named as the Orthologous SSRs for Disease Genes (OSDG).

2 Material and Method

In this developed system, we have collected and integrated a few representative databases for identifying orthologous SSRs of disease genes. First, the genetic disease information in the developed system was retrieved from OMIM database. The OMIM is an online catalogue containing all known diseases with a genetic component, and it was shosen as the main resources of genetic diseases. The dataset of genetic diseases was rearranged and stored as an effective structure in our local database for real time searching. Currently, 11,486 genes related with 4,595 human genetic diseases were collected in the proposed system [23].

In addition to related genetic disease profiles, the annotation data for disease related genes within various species, including detailed transcripts and translate

locations, was acquired from Ensembl release 49 through the BioMart interface [24]. This information facilitates our proposed system to demarcate and comprehend the distinct part of genes (upstream, downstream, 3'UTR, 5'UTR, Intron, Coding). For SSR information, the proposed system integrated with a SSR searching system (Comparative genomics SSR discovery, CG-SSR) developed by Pai T., *et al*. [20], which contributes all SSR motifs among multiple species. Each SSR record, retrieved from CG-SSR database, includes information of core repeat pattern, core pattern length, location on the chromosome, and ratio of tolerance.

Finally, orthologous information from two databases was adopted for integration. One is the HomoloGene profile from OMIM and the other is orthologous information from Ensembl. The NCBI HomoloGene is foundation of orthologous information in the proposed system. If a gene is not included in the HomoloGene, the system checks the gene symbol and its chromosome location of gene to obtain the Gene ID. Afterward, the system proceeds to obtain the orthologous information from Ensembl according to the Gene ID automatically.

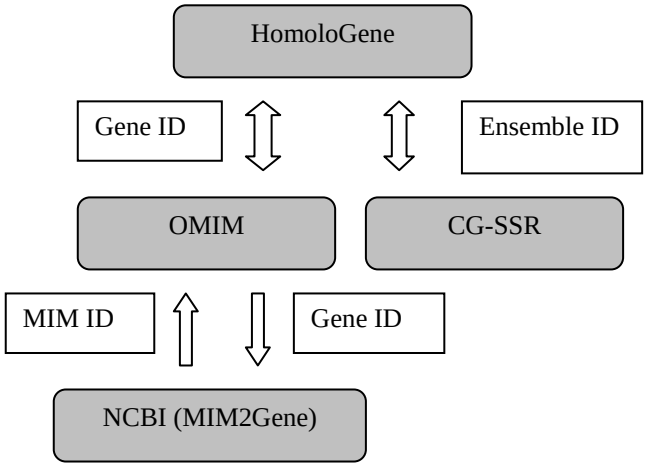

Figure 1: The framework of transforming data resources.

Each database adopts different identifiers to record different types of gene information. For combining the information from different data resources, the system has designed a transforming relationship among various distinct databases. At first, the system set up a look-up table between Gene ID and MIM ID through MIM2Gene profile belonged to NCBI. After accomplishing the mapping relationship between Gene ID and MIM ID, the link between OMIM and HomoloGene datasets was also established through Gene ID. According to the Ensemble Gene ID, the system built the relation between HomoloGene profile and CG-SSR searching results. The framework of transforming relation was shown in Figure 1. After completing the setting-up processes of different material resources, an on-line query system for identifying the orthologous SSRs of genetic diseases was accomplished by integrating the information from OMIM, HomoloGene and CG-SSR, and the data flowchart and comparative genomics mechanism were shown in Figure 2.

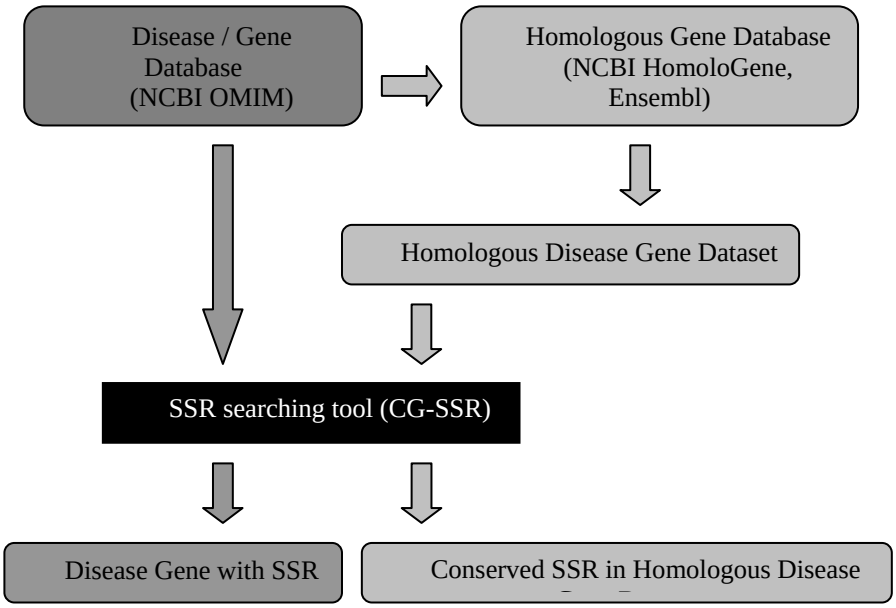

Figure 2: The flowchart of searching orthologous SSRs of genetic diseases by comparative genomics mechanism.

The designed web system provides four different searching mechanisms named as *Orthologous*, *SSR*, *Disease*, and *Gene*, respectively. First, the *Orthologous* function displays a list of disease genes from OMIM and those genes possess orthologous relationship among selected model species. In this study, there are 8 model species considered for orthologous search including human, mouse, rat, zebrafish, chicken, cow, dog and chimpanzee. Through cross-species comparison mechanism, all potential functional SSRs in various regions of genes were annotated from all genomes of model species. Optional keyword filtering function for matching prefix of gene symbol name is available to help users to narrow down the possible searching results. Secondly, the *SSR search* function facilitates users searching all disease related genes efficiently by setting a specific SSR pattern which is considered as an important pattern found in previously published literatures. Totally 501 possible SSR basic patterns from mononucleotide to hexanucleotide can be found within whole genome sequences of various model species. It is important that users can designate a specific gene region where SSRs located at. Six regions of a gene sequence were categorized as: upstream, 5'UTR, intron, exon, 3'UTR and downstream. Thirdly, the *Disease* function provides users to get detail information about a specific genetic disease by querying with multiple keywords, prefixes of a disease name, related disease description, or simply the MIM ID parameters. Finally, the *Gene* function provides three ways to discover disease genes: (1) Disease gene symbol, such as "HTT", (2) NCBI/EMBL gene ID, such as 143100 or ENSG00000197386, and (3) MIM ID, such as 601373. A screen shot of the proposed

system was shown in Figure 3, which describes the application of retrieving associated disease genes with respect to the SSR pattern “CAG”.

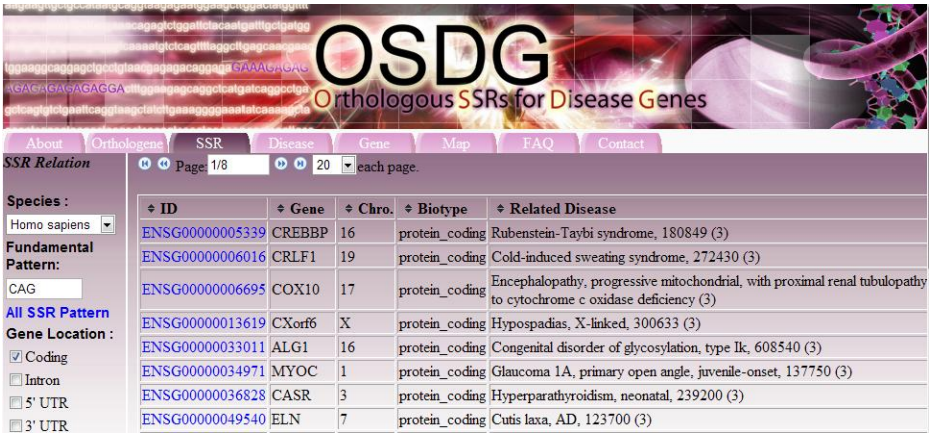

Figure 3: A screen shot from “SSR” motif function. All disease genes containing the specific SSR pattern ”CAG” were displayed in the list.

3 Results and discussion

Based on interoperability of orthologous properties and SSR searching techniques, all conserved orthologous SSR patterns can be discovered from the interested inherited disease genes. Through comparative genomics mechanism, the OSDG system possesses the ability of retrieving possible functional SSRs from orthologous disease genes of medical importance. The information of retrieved SSRs contained within 8 model species genomes includes core pattern length (length threshold >20), quality threshold (0=perfect SSR, 1=imperfect SSR), and locations (upstream, 3’UTR, 5’UTR, intron, coding and downstream).

The conserved orthologous SSRs can be defined as if the same SSR patterns are existed (the same core pattern appeared in the same region)among the target gene and its orthologous genes. System automatically highlights the SSRs within coding regions when the amount of species is over half of all organisms. Taking the HTT gene as an example, the mutation is verified and caused by the “AGC “(or “CAG”) repeat expansion and it induces the Huntington Disease symptoms [25]. The searched results from OSDG were shown in Table1.

SSR expansions and/or contractions in protein-coding regions brought about a gain or loss of gene function through frame-shift mutation. Tri-nucleotide repeats located in coding region were associated with various types of cancer and implicated in various neurodegenerative disorders [26]. Hence, we experimented on all disease genes to gather statistical information about the distribution of ten kinds of triplet SSRs. The results were shown in Table2. These results not only coincided with the conclusions of previously published observation but also interpreted the importance of biological meaning. For examples, the AGC, CGG and AGG patterns are the most

core patterns appeared in coding regions, and ATT (TAA) are never appeared in the coding region since it would be translated into a stop codon [9]. Most of them in 5'UTR were CCG which served as binding sites [13]. In the intron regions, the absence of CCG in various taxonomic groups is due to highly mutable CpG pattern and results in CpG methylation [4].

Table 1: The conserved SSR motifs for HTT gene among eight model species, only tri-nucleotide patterns were displayed here.

Pattern	Cow	Dog	Zebrafish	Chicken	Human	Mouse	Chimp	Rat
AAC					Intron	Intron	Intron	Intron
AAG				Intron	Intron	Intron	Intron	Intron
ACC	Intron	Intron			Intron	Intron	Intron	Intron
AGC	**Coding**		**Coding**		**Coding**	**Coding**	**Coding**	**Coding**
AGG	Intron	Intron			Intron	Intron	Intron	Intron
CCG	**Coding**	Intron			**Coding** Intron	**Coding**	**Coding** Intron	**Coding**

Table 2: The amount of disease genes appear in six districts with specific triplet SSRs.
[Parameters: Length Threshold (20) and Quality Threshold (0.2)].

Pattern	Upstream	3'UTR	Coding	Intron	5'UTR	Downstream
AAT	118	49	0	1173	3	120
AAC	60	32	20	826	3	88
AGC	35	13	148	236	28	27
AGG	262	56	155	1168	65	152
CGG	180	27	141	170	133	18
ACG	0	0	2	4	0	1
ACT	6	1	3	120	0	4
ATC	11	11	26	386	3	23
AAG	88	31	57	846	13	65
ACC	35	15	47	440	4	33

A comprehensive annotation resource for human genes, and especially of genetic diseases with orthologous SSRs, provides one of important research directions and guidance of translational medicine. It is believed that accumulated knowledge of genes associated with genetic disorder caused by SSRs allow researchers to address more complicated issues, including the relative contributions of genetic diseases shared by all species and how sequence features (such as conservation and polymorphism) relate to disease characteristics. Accordingly, the OSDG system provides an efficient and effective tool to discover well conserved SSRs among orthologous genes and facilitate users getting potential SSR candidates as disease gene regulators, and it can be expected to discover all potential SSRs associated with

genetic diseases through cross-species comparison.

Acknowledgements

This work is supported by the Center for Marine Bioscience and Biotechnology (CMBB) in National Taiwan Ocean University, Keelung, Taiwan, and the National Science Council in Taiwan, R. O. C. (NSC97-2627-B-019-003 to T.-W. Pai).

References

[1] Charlesworth B, Sniegowski P, Stephan W. “The evolutionary dynamics of repetitive DNA in eukaryotes”; *Nature*, 371, 215, 1994.
[2] Sharma PC, Grover A, Kahl G. “Mining microsatellites in eukaryotic genomes”; *Trends Biotechnol*, 25, pp. 490-498, Nov, 2007.
[3] Bacolla A, Larson JE, Collins JR. ”Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties”; *Genome* , Aug. 2008.
[4] Jurka J, Pethiyagoda C. “Simple repetitive DNA sequences from primates: compilation and analysis”; *J. Mol. Evol.*. 40, pp. 120-126, Feb, 1995.
[5] Wren JD, Forgacs E, Fondon JW, Pertsemlidis A, Cheng SY, Gallardo T, Williams RS, Shohet RV, Minna JD, Garner HR. “Repeat polymorphisms within gene regions: phenotypic and evolutionary implications”; *Am. J. Hum. Genet.*. 67, pp. 345-356, Aug, 2000.
[6] Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK. ”Short tandem repeat polymorphism evolution in humans”; *European Journal of Human Genetics* 6, 38-49, 1998.
[7] Subramanian S, Madgula VM, George R, Mishra RK, Pandit MW, Kumar CS, and Singh L. ” Triplet repeats in human genome: distribution and their association with genes and other genomic regions”; *Bioinformatics* , Nov. 2002
[8] Li Y, Korol AB, Fahima T,Nevo E. “Microsatellites within genes: structure, function, and evolution”; *Mol. Biol. Evol.*. 21, pp. 991-1007, Jun, 2004
[9] http://www.ornl.gov/sci/techresources/Human_Genome/medicine/assist.shtml, Accessed date Jan. 2009.
[10] Hirschhorn, Joel N, Lohmueller, Kirk, Byrne, Edward, Hirschhorn, Kurt. “A comprehensive review of genetic association studies”; *Genetics in Medicine* - Volume 4 - Issue 2 - pp 45-61, March/April 2002.
[11] Sutherland GR, Richards RI.” Simple tandem DNA repeats and human genetic disease”; *PNAS* vol. 92 no. 9 3636-3641, Apr. 1995.
[12] Richards RI, Holman K, Yu S, Sutherland GR. “Fragile X syndrome unstable element, p(CCG)n, and other simple tandem repeat sequences are binding sites for specific nuclear proteins”; *Hum. Mol. Genet.*. 2, pp. 1429-1435, Sep, 1993.
[13] Gusella JF, Macdonald ME. “Huntington's disease: seeing the pathogenic process through a genetic lens”; *Trends Biochem. Sci.*. 31, pp. 533-540, Sep, 2006.
[14] Ionov Y, Peinado MA, Malkhosyan S, Shibata D, Perucho M. “Microsatellite instability: The mutator that mutates the other mutator ”; *Nature 363*, 558, 1993.
[15] Kashi Y, King DG. ” Simple sequence repeats as advantageous mutators in evolution”; *Trends in Genetics* Volume 22, Issue 5, 253-259, May 2006.
[16] Alexeyenko A, Lindberg J, Pe´rez-Bercoff A, Sonnhammer ELL. ” Overview and comparison of ortholog Databases”; *Drug Discovery Today*: Technol., 3:137–143, 2006.
[17] Sonnhammer EL, Koonin EV. “Orthology,paralogy and proposed classification for

paralog subtypes"; *TRENDS in Genetics* Vol.18 No.12 619-620, Dec. 2002.
[18] Online Mendelian Inheritance in Man, OMIM (TM). http://www.ncbi.nlm.nih.gov/omim/. Accessed date Jan. 2009.
[19] HomoloGene http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene. Accessed date Jan. 2009.
[20] Pai T, Chen C, Hsiao M, Cheng R, Tzou W,Hu C. "An online conserved SSR discovery through cross-species comparison"; *Computational Biology and Chemistry: Advances and Applications*. 2, pp. 23-35, 2009.
[21] Boby T, Patch AM, Aves SJ. "Trbase : a database relating tandem repeats to disease genes for the human genome"; *Bioinformatics Advance Access originally published online*, Oct. 2004.
[22] O'Brien KP, Westerlund I, and Sonnhammer ELL ," OrthoDisease: A Database of Human Disease Orthologs"; *Hum. Mutat.* 24, pp. 112–119, 2004.
[23] Hamosh A, Scott AF, Amberger JS, Bocchini CA,McKusick VA. "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders"; *Nucleic Acids Res.*. 33, p. D514-7, Jan, 2005.
[24] Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S,Flicek P. "Ensembl 2009"; Nucleic Acids Res.. 37, p. D690-7, Jan, 2009.
[25] Susan E. Andrew, Y. Paul Goldberg, Berry Kremer, Håkan Telenius, Jane Theilmann, Shelin Adam, Elizabeth Starr, Ferdinando Squitieri, Biaoyang Lin, Michael A. Kalchman, Rona K. Graham , Michael R. Hayden. "The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease"; *Nature Genetics* 4, 398 - 403 ,1993.
[26] Karl Kieburtz, Marcy MacDonald, Charles Shih, Andrew Feigin, Kim Steinberg, Kathy Bordwell, Carol Zimmerman, Jayalakshmi Srinidhi, Jenny Sotack, James Gusella, Ira Shoulson. "Trinucleotide repeat length and progression of illness in Huntington's disease"; *J Med Genet* 31:872-874.1994.