BMC Genomics

BioMed Central

Research article

CAG-encoded polyglutamine length polymorphism in the human genome

Stefanie L Butland[1], Rebecca S Devon[2], Yong Huang[1], Carri-Lyn Mead[1], Alison M Meynert[1], Scott J Neal[2], Soo Sen Lee[1], Anna Wilkinson[1], George S Yang[3], Macaire MS Yuen[1], Michael R Hayden[2,4], Robert A Holt[3,5], Blair R Leavitt[†2,4] and BF Francis Ouellette[*†1,4]

Address: [1]UBC Bioinformatics Centre, Michael Smith Laboratories, University of British Columbia, Vancouver, Canada, [2]Centre for Molecular Medicine and Therapeutics, Child and Family Research Institute, Department of Medical Genetics, University of British Columbia, Vancouver, Canada, [3]Canada's Michael Smith Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, Canada, [4]Department of Medical Genetics, University of British Columbia, Vancouver, Canada and [5]Department of Psychiatry, University of British Columbia, Vancouver, Canada

Email: Stefanie L Butland - butland@bioinformatics.ubc.ca; Rebecca S Devon - Rebecca.Devon@ed.ac.uk; Yong Huang - dewriver@gmail.com; Carri-Lyn Mead - cmead@bcgsc.ca; Alison M Meynert - ameynert@ebi.ac.uk; Scott J Neal - sneal@cmmt.ubc.ca; Soo Sen Lee - soo_sen@yahoo.com; Anna Wilkinson - guywil@shaw.ca; George S Yang - gyang@bcgsc.ca; Macaire MS Yuen - mackyuen@gmail.com; Michael R Hayden - mrh@cmmt.ubc.ca; Robert A Holt - rholt@bcgsc.ca; Blair R Leavitt - bleavitt@cmmt.ubc.ca; BF Francis Ouellette* - francis@bioinformatics.ubc.ca

* Corresponding author †Equal contributors

Published: 22 May 2007

BMC Genomics 2007, **8**:126 doi:10.1186/1471-2164-8-126

Received: 23 October 2006
Accepted: 22 May 2007

This article is available from: http://www.biomedcentral.com/1471-2164/8/126

Abstract

Background: Expansion of polyglutamine-encoding CAG trinucleotide repeats has been identified as the pathogenic mutation in nine different genes associated with neurodegenerative disorders. The majority of individuals clinically diagnosed with spinocerebellar ataxia do not have mutations within known disease genes, and it is likely that additional ataxias or Huntington disease-like disorders will be found to be caused by this common mutational mechanism. We set out to determine the length distributions of CAG-polyglutamine tracts for the entire human genome in a set of healthy individuals in order to characterize the nature of polyglutamine repeat length variation across the human genome, to establish the background against which pathogenic repeat expansions can be detected, and to prioritize candidate genes for repeat expansion disorders.

Results: We found that repeats, including those in known disease genes, have unique distributions of glutamine tract lengths, as measured by fragment analysis of PCR-amplified repeat regions. This emphasizes the need to characterize each distribution and avoid making generalizations between loci. The best predictors of known disease genes were occurrence of a long CAG-tract uninterrupted by CAA codons in their reference genome sequence, and high glutamine tract length variance in the normal population. We used these parameters to identify eight priority candidate genes for polyglutamine expansion disorders. Twelve CAG-polyglutamine repeats were invariant and these can likely be excluded as candidates. We outline some confusion in the literature about this type of data, difficulties in comparing such data between publications, and its application to studies of disease prevalence in different populations. Analysis of Gene Ontology-based functions of CAG-polyglutamine-containing genes provided a visual framework for interpretation of these

genes' functions. All nine known disease genes were involved in DNA-dependent regulation of transcription or in neurogenesis, as were all of the well-characterized priority candidate genes.

Conclusion: This publication makes freely available the normal distributions of CAG-polyglutamine repeats in the human genome. Using these background distributions, against which pathogenic expansions can be identified, we have begun screening for mutations in individuals clinically diagnosed with novel forms of spinocerebellar ataxia or Huntington disease-like disorders who do not have identified mutations within the known disease-associated genes.

Background

Nine different neurodegenerative disorders are known to be caused by expansions of polyglutamine-encoding CAG trinucleotide (CAGpolyQ) repeats in the following genes: the *HD* gene in Huntington disease [1], *ATN1* in dentatorubral pallidoluysian atrophy or Haw River syndrome [2,3], *AR* in spinal and bulbar muscular atrophy [4], *CACNA1A* in spinocerebellar ataxia SCA6 [5], *TBP* in SCA17 [6] and *ATXN1*, *2*, *3*, and *7* in SCA1 [7], SCA2 [8-10], SCA3 (Machado-Joseph disease) [11], and SCA7 [12]. These disorders share similar clinical features which include selective neuronal degradation associated with a progressive neurological phenotype, but their respective causative genes appear to have little functional or structural similarity, suggesting that functional genomics approaches to identifying new gene-disease associations will not be useful. The repeat expansion mechanism of pathogenesis is a shared molecular feature, and this form of mutation has only been exhaustively ruled out for a few familial forms of SCA, and has not been examined at all for the majority of patients who present with SCA or HD-like disorders.

Despite recent advances in molecular diagnosis, the majority of individuals clinically diagnosed with SCA do not have identified mutations within the known disease-associated genes [13]. There are 28 genetically distinct SCAs identified by the Human Gene Nomenclature Committee (HGNC) [14], but only 13 causative genes are known. Six genes cause SCA by CAGpolyQ expansions, but the remaining 15 clinically-defined forms of SCA have no known genetic mutation associated with them, and the search for causative genes continues. It is likely that some of these forms of SCA will be found to be caused by this common mutational mechanism. Candidate genes for SCA and HD-like disorders can be identified using a whole-genome screening approach based on the computational identification of a common sequence we have termed a Genomic Mutational Signature (GeMS). GeMS are sequence patterns occurring in the normal genome that, when mutated, cause disease – in this case CAG trinucleotide repeats that encode an extended tract of glutamine residues in the protein. A significant advantage of this approach is that novel candidate disease genes are identified and can then be screened for mutations in single cases. This approach is not constrained by any requirement for additional family members, additional affected patients, nor is a detailed family history required.

Partial lists of CAGpolyQ-containing genes identified using classical [15-20] or computational methods [21-24] have been published. Screening for CAG expansions in one such gene list, in patients with hereditary ataxias, led directly to the discovery of the causative gene for dentatorubral pallidoluysian atrophy [2,16]. To date, there has been no complete genome-wide analysis of the distributions of CAGpolyQ repeat lengths in a control population in order to set the baseline from which to detect expansions. Studies on a limited number of genes have revealed that different genes have very different polyglutamine tract (Q-tract) length distributions with some invariant (*CREBBP*) [25] some bimodal (*ATXN3*) [26], some very narrow (*ATXN2*) [26] and some broad distributions (*AR*, *ATN1*, *SMARCA2* and *THAP11*) [26-28].

The molecular nature of polyglutamine repeats

The amino acid glutamine (Q) is encoded by CAG and CAA trinucleotides. Q-tracts in proteins are typically encoded by mixtures of these two codons while expanded Q-tracts in disease-causing genes are typically composed of long uninterrupted repeats of the CAG trinucleotide only. Long uninterrupted CAG repeats are known to be a substrate for expansion mutation by a variety of mechanisms. The underlying process is currently thought to involve the generation of abnormal DNA structures induced by factors such as replication slippage, DNA repair and recombination, that can contribute to repeat instability acting either separately or in combination [29-32] and these mutations underlie pathogenic expansions [33] and genetic anticipation [34,35]. Q-tracts encoded by mixtures of CAG and CAA codons, however, are less prone to suffer expansions [30,36,37]. The precise nucleotide sequence of a repeat tract determines a particular allele's susceptibility to large expansion mutations, while the amino acid sequence – the Q-tract – in the context of the whole protein determines the effect of a length change on molecular and clinical phenotypes.

Characteristics of known disease genes

One motivation for this research was to enable us to prioritize candidate genes for polyglutamine expansion disorders. Thus, we sought to identify hallmarks among the known disease genes to which we could compare our data on CAGpolyQ genes not yet associated with disease. Disease-causing CAGpolyQ-containing genes tend to be considered a homogeneous group in terms of their repeats, with an often-cited pathogenic threshold of about 35 glutamines. In fact, a closer look at normal and pathogenic characteristics of each reveals their unique qualities. *ATXN2* has a remarkably narrow distribution of Q-tract lengths with very few alleles longer or shorter than the modal length of 22Q [26,37]. In contrast, *ATXN3* has a broad bi- (or tri-) modal distribution of Q-tract lengths [26]. Disease genes can also differ in the number of Q-residues that separate the longest normal from the shortest pathogenic allele. The longest normal *ATN1* Q-tract is 36Q and the shortest disease allele has 48Q [26,38], while a single residue separates normal (19Q) from pathogenic (20Q) Q-tracts in *CACNA1A* [26,38]. Some disease genes carry non-glutamine interruptions in their Q-tracts, though their lengths are often reported as "repeat lengths" as if they were pure Q-tracts. For example, normal *ATXN1* has one to three CAT (coding for histidine, H) interruptions near the middle of the Q-tract, but in SCA1 disease alleles the repeat tracts are pure CAGpolyQ [39]. Clearly one must be cautious in making assumptions about common features among polyglutamine expansion disease genes when seeking to identify new disease-associated genes.

At the sequence level, polyglutamine expansion disease genes share several characteristics. They have long uninterrupted CAG tracts [29] and tend to have polymorphic Q-tract lengths [26,36]. Analysis of both genomic DNA and expressed sequence tags have shown that pure CAG-tract length is correlated with Q-tract variance [36,40,41] and interruptions provide stability to repeat tracts [36,37]. Finally, comparisons of orthologous human and rodent genes show that the lengths of disease-associated Q-tracts have a low level of conservation between species compared with those that are not associated with disease [29,42].

The products of the genes causing polyglutamine expansion disorders do not all share a specific function, but the phenotypic overlap of these disorders does suggest some common functions in either their normal or mutated states, or both. As early as 1989, researchers noted the involvement of polyQ-containing genes in transcriptional regulation [43]. This connection spans organisms from yeast to humans [44-48] and known disease-causing genes like *HD*, *TBP* and *ATXN7* are directly involved in transcription and transcriptional regulation [49-55]. *ATXN1* and *ATXN2* are thought to be involved in RNA metabolism [56,57] while *CACNA1A* is the only ion channel gene known to cause a polyglutamine expansion disorder [5]. The normal function of a gene product and the role of the Q-tract in that protein determine the distribution of repeat lengths in the normal population and the threshold for pathogenic expansion for each gene. Therefore, the functions of CAGpolyQ-containing genes must be assessed in conjunction with the normal levels of repeat polymorphism in order to prioritize candidate genes for polyglutamine expansion disorders.

Summary

Using the human genome reference sequence [58,59] and Ensembl annotated genes [60] we performed a genome-wide computational identification of all candidate genes containing a specific GeMS sequence, CAGpolyQ repeats. We used fragment analysis to assess the CAG-tract lengths of these candidate genes in a large control population. We also applied two methods of analyzing the potential functions of these genes based on the Gene Ontology (GO) system of functional classification [61] in order to identify and visualize the network of functional relationships among the CAGpolyQ-containing genes in the human genome. Using related approaches, Lavoie and colleagues identified polyalanine-containing genes in the human genome and assessed their normal levels of polymorphism [62]. Functional analysis revealed that the majority of polyalanine-containing genes have roles in transcriptional regulation [62].

In characterizing the Q-tract length distributions for 64 CAGpolyQ tracts in 62 genes in the human genome, we find that each Q-tract has a unique distribution of Q-tract lengths. The best predictors of known disease genes were occurrence of a long uninterrupted CAG-tract in their reference genome sequence and high Q-tract length variance in the normal population. Therefore, we used these parameters to identify eight priority candidate genes for polyglutamine expansion disorders. The majority of CAGpolyQ-containing genes are involved in transcriptional regulation and neurogenesis. We provide a visual framework for interpretation of new information on CAGpolyQ gene functions and their biomolecular interactions.

Results

Identification of CAGpolyQ-containing Genes

CAGpolyQ repeats were identified on the basis of having tandemly repeated CAG trinucleotides in the sequence within the boundaries of a known gene that had five or more tandem glutamine residues in its peptide sequence (see Methods for detailed description of approach and data sources). Build 33 of the human genome sequence [58] contained 436 CAG trinucleotide repeats in total. Sixty-six of these CAG repeats lay in glutamine-coding

sequences in genes including all nine genes in which mutation by expansion of their CAGpolyQ repeat tract is known to cause a neurodegenerative disorder (Table 1).

Distributions of Q-tract Lengths

Using PCR amplification and ABI fragment analysis we established the range of CAGpolyQ tract lengths for 64 targets (in 62 genes) in a set of healthy individuals of mixed ethnic background (Table 1, Additional file 1). We screened at least 130 normal alleles for each target (mean 162), including X-linked genes, giving us 99% confidence that 95% of the whole population lie between the minimum and maximum values in our sample (95% tolerance; see Methods), with the exception of four targets for which we screened slightly fewer alleles due to technical limitations: *ATXN2* and *CACNA1A* (94% tolerance), *FOXP2* and *RUNX2* (93% tolerance). Table 1 summarizes data for 66 CAGpolyQ repeat targets in 64 genes.

Known disease genes have long uninterrupted CAG-tracts and high Q-tract length variances

We sought in our data some hallmark of the nine known disease genes that would allow us to prioritize candidates among the 54 genes not yet associated with CAGpolyQ expansion disorders. Sorting CAGpolyQ repeats by increasing Q-tract length variance (Table 1) clustered disease genes in the top one third of 64 targets. Known disease gene Q-tract length variances ranged from 0.79 (*ATXN2*) to 29.2 (*ATXN3*). The highest Q-tract length variances of all targets were observed in four known disease genes: *ATXN3*, *ATN1*, *AR* and *HD*. The least polymorphic disease target, *ATXN2*, is distinguished from other disease genes by its previously documented tight distribution of Q-tract lengths [26].

Q-tracts are made up of lengths of CAG codons that can be pure or interspersed with one or more CAA codons. Length polymorphism tends to occur within CAG-tracts. Sorting CAGpolyQ repeats by the length of their longest uninterrupted CAG-tract in the reference genome clustered disease genes in the top half of 64 targets. This was increased to the top one third if *ATXN3* was excluded due to its reference genome sequence reflecting the low mode of a bimodal distribution of repeat tract lengths (see graph in Additional file 2). Disease gene CAG-tract lengths ranged from 10 (*ATXN7*) to 22 (*AR*) and the longest uninterrupted CAG-tracts of all targets occurred in four disease genes: *AR* (22 CAG), *HD* (19 CAG), *TBP* (19 CAG) and *ATN1* (15 CAG).

The length of the longest uninterrupted CAG-tract in the reference genome for each target (*e.g.* $CAG_{13}CAA_1CAG_9$ has CAG-tract length of 13; see Table 1) was positively correlated with its level of polymorphism (correlation = 0.62, *ATXN3* excluded; Figure 1). Given this association between long CAG-tracts and high Q-tract length variance, we divided all targets in two groups at the median CAG length of eight and tested the null hypothesis that variances were equal in the two groups. Q-tract length variances were indeed higher with longer CAG-tracts ($p = 0.002$, 1-tailed heteroscedastic t-test).

Mean or maximum Q-tract length failed to yield any significant clustering of disease genes, and mean Q-tract length was only very weakly correlated with Q-tract length variance (correlation = 0.12). Underlying this relationship is the fairly weak correlation of uninterrupted CAG-tract length with mean Q-tract length (0.31, *ATXN3* excluded). Mixtures of CAG and CAA codons making up the Q-tract account for this. One telling example is *FOXP2* which had the longest mean and maximum Q-tract lengths but relatively little variance in Q-tract length. In fact, *FOXP2* had the second-shortest uninterrupted CAG-tract of all 66 targets. Based on our analysis, this low level of polymorphism is predicted by the short pure CAG repeat length.

Sorting targets according to other parameters also failed to yield any significant clustering of disease genes. These included sorting by the proportion of alleles with Q-tract lengths longer than mean + 1 SD, and by repeat purity, which was a combined measure of both the length of the longest uninterrupted CAG-tract and the total Q-tract length.

Priority candidates for polyglutamine expansion disorders

A plot of CAG length versus Q-tract length variance for each target allowed us to identify eight genes as priority candidates for polyglutamine expansion disorders (Figure 1). We selected genes that had uninterrupted CAG-tracts equal to or longer than 10 CAG (the shortest uninterrupted CAG-tract in a known disease gene, *ATXN7*) and had Q-tract length variance equal to or higher than 0.79 (the lowest Q-tract variance in a known disease gene, *ATXN2*). All eight priority candidates: *C14orf4*, *KCNN3*, *KIAA2018*, *MEF2A*, *NCOR2*, *RAI1*, *SMARCA2*, and *THAP11* are expressed in normal brain [63-66]. This list is not meant to be exhaustive, but rather a list of the top eight genes prioritized according to two hallmarks of known disease genes.

Twelve invariant CAGpolyQ repeats have short CAG-tracts

In this set of 64 CAGpolyQ repeats, having at least four tandem CAG codons coding for five tandem glutamine residues, mean Q-tract length ranged from five to 39.8 (Table 1). Twelve repeats in eleven genes, including CREB-binding protein (*CREBBP*) for example, had no changes in Q-tract length in as many as 212 alleles tested. An additional six repeats were essentially invariant with only one out of as many as 184 alleles differing in length by one Q-residue (Table 1). The twelve invariant repeats had unin-

Table 1: Q-tract length variation in genes containing polyglutamine-encoding CAG-type trinucleotide repeats, sorted by Q-tract

Chromosome Band	Gene Name[a]	Repeat Sequence from Reference Genome (sense strand)[b]	Expected Q-tract Length from Reference Genome[c]	N[d]	Observed Q-tract Length Min-Max	Q-tract Mean	Q-tract Variance
17p13.2	MINK1*	G4N1**G5**	Q4LQ5 (SwP)	162	5 – 5	5.0	0
9q34.11	CIZ1	**G6**	Q6	154	6 – 6	6.0	0
7q36.2	PAXIP1L*	**G7**	Q7	168	7 – 7	7.0	0
11q24.3	PRDM10	**G8**	Q8	172	8 – 8	8.0	0
4q31.1	MAML3a*	**G9**	Q9	156	8 – 8	8.0	0
6p21.1	TFEB	**G6**A1G3	Q10	162	10 – 10	10.0	0
19p13.11	CHERP	**G6**A1G5	Q12	192	12 – 12	12.0	0
12q21.2	PHLDA1	G5A1**G6**A2G1	Q15	212	14 – 14	14.0	0
16p13.3	CREBBP	**G4**A1G3A2G2A1**G4**A1	Q18	158	18 – 18	18.0	0
4q31.1	MAML3b*	G3A1G3A1G1A1**G8**	Q18	166	18 – 18	18.0	0
20q11.22	NCOA6*	G4A4**G8**A2G1A1G2A2G1	Q25	166	25 – 25	25.0	0
Xq13.1	MED12*	G5A1G2A1G1A1G5A1G1A1**G7**N4G6	Q26X4Q6	205	26 – 27	26.0	0
20q13.12	PRKCBP1	**G7**A1	Q8	152	8 – 9	8.0	0.01
15q24.1	ARID3B	**G8**A2G1	Q11	212	11 – 12	11.0	0.01
22q11.21	PCQAPa	G4A1G3N1G5N3G7A1G3N8G3N5G5N1**G8**	Q8FQ5X3Q11X16Q5LQ8	152	11 – 12	11.0	0.01
3p24.3	SATB1	G1A1G3A1G1A1**G7**	Q15	174	15 – 16	15.0	0.01
6q16.2	POU3F2	G3A1G1A1G3A1G2A1**G6**A1G1	Q21	148	21 – 22	21.0	0.01
Xq22.3	FRMPD3	G4A3G4A3G3A3**G7**	Q27 (SwP)	184	26 – 27	27.0	0.01
2q35	TNS	**G9**	Q9	178	9 – 11	9.0	0.02
19p13.12	BRD4	**G5**N1G1N1A1G4A1G1A1	Q5RQEQ8	140	8 – 9	8.0	0.03
12p13.31	PHC1	**G5**A2G1A1G2A1G3	Q15	170	13 – 15	15.0	0.05
9q32	C9orf43	**G6**A1G1	Q8	168	8 – 9	8.1	0.07
1q21.3	TNRC4	A1**G8**A1G4A1	Q15	150	15 – 18	15.0	0.08
17q12	SOCS7	**G7**A1	Q8 (SwP)	134	8 – 9	8.1	0.12
1p31.1	ST6GALNAC5	**G7**A1G4	Q12	150	12 – 14	12.1	0.13
15q26.1	POLG	**G10**A1G2	Q13	164	13 – 15	13.1	0.16
22q13.1	TNRC6B	**G8**	Q8	166	7 – 8	7.8	0.17
12q13.12	MLL2*	G5N1A1G1A1G1A1N1**G7**N1A1G1A1G1A1N1G2A1G1N1A1G2A1G4N1A2G3A1G1N1A1G2A1G2N1A1G1A1G1A3G3N1A1G3A1G3	Q5LQ5LQ7LQ5LQ4LQ8LQ7LQ6LQ10FQ8	184	8 – 11	10.2	0.21
7p14.1	POU6F2	**G10**	Q10	168	6 – 11	10.0	0.22
Xq28	CXorf6	G1A1**G8**A1N92G5A1G4	Q11X92Q10	168	11 – 12	11.6	0.25
12p13.33	DCP1B	**G9**A1	Q10	136	10 – 12	10.5	0.26
17q23.2	VEZF1	**G12**A6	Q13 (through intron)	176	8 – 15	13.1	0.29
22q11.21	PCQAPb	G3A1G2N9A2G1A1**G12**	Q6X9Q16	152	12 – 18	16.1	0.34
3p14.1	MAGI1	G5A1G3A1**G10**	Q20	168	16 – 21	20.3	0.36
4q21.21	BMP2K	G8A1G1A1G4A1G1A1**G9**	Q27	148	23 – 28	26.9	0.36
16q22.1	NFAT5*	**G5**A1G3A1G3A3G1	Q17	168	11 – 19	17.0	0.37
12p13.31	ZNF384	**G14**A1G1	Q16	214	11 – 20	15.2	0.47
22q12.1	MN1*	A1**G9**A1G6A1G1A1G1A1G6	Q28	180	26 – 30	28.7	0.53
12q24.33	EP400	G6A2**G14**A1G4A1G1	Q29	158	28 – 31	28.8	0.53
12q23.2	ASCL1	**G12**	Q12	148	9 – 15	12.3	0.65
6q25.3	ARID1B	**G7**A1G7A1G1A1	Q18	152	16 – 23	17.7	0.69
11q21	MAML2	G1A1G2A1**G13**A1G5A1G1A1G1A1G1A1G2N5A2G1A1G3N5A1G5A2G5A3G1A2G6A2	Q31X5Q7X5Q27 (through intron)	168	27 – 31	28.3	0.75
12q24.12	**ATXN2**	**G13**A1G9	Q23	124	17 – 27	22.2	0.79
9p24.3	SMARCA2	G1A2G3A1**G13**A1G2	Q23	130	18 – 24	22.7	0.79
20q13.12	NCOA3	G6A1**G9**A1G1A1G1A1G1A1G2A1G2A1	Q29	150	26 – 31	28.4	0.80
17p11.2	RAI1	**G13**A1	Q14	184	11 – 17	14.6	0.84
7q31.1	FOXP2*	G4A1G4A2G2A2G3A5G2A2**G5**A1**G5**A1G1	Q40	100	34 – 40	39.8	0.85
3p14.1	**ATXN7**	**G10**	Q10	184	7 – 14	10.4	0.89
19q13.2	NUMBL	G6A1G1A1**G7**A1G2A1	Q20	156	18 – 20	18.7	0.93

Table 1: Q-tract length variation in genes containing polyglutamine-encoding CAG-type trinucleotide repeats, sorted by Q-tract

12q24.31	NCOR2	G3A2**G12**	Q16 (through intron)	172	13 – 20	16.9	0.95
15q26.3	MEF2A	**G11**	Q11	174	8 – 16	10.2	1.13
14q24.3	C14orf4	A1G1A1G1A1G6A1**G10**A2G1	Q25 (through intron)	150	20 – 31	23.4	1.17
3q13.2	KIAA2018	**G11**A1G1A4	Q14 (through intron)	150	11 – 16	12.6	1.44
1q21.3	DENND4B	A1G5A1**G9**	Q16	156	13 – 17	15.2	2.04
6p22.3	**ATXN1**	G12T1G1T1**G14**	Q12HQHQ14	130	11 – 21	14.6	2.23
6q27	**TBP**	G3A3G8A1G1A1**G19**A1G1	Q38	158	30 – 41	36.9	2.26
19p13.3	**CACNA1A**	**G13**	Q13	112	7 – 16	12.1	2.42
16p12.1	TNRC6A	**G4**A1G3	Q8	166	4 – 8	7.2	2.50
6p21.1	RUNX2	A1G3A1G4A1**G6**A1**G6**	Q23	100	18 – 30	22.5	3.04
16q22.1	THAP11	G3A1G5A1G2A1G5A1**G10**	Q29	170	18 – 30	28.5	3.12
1q22	KCNN3	G7A1G4N25**G14**	Q12X25Q14	170	15 – 25	20.3	3.98
4p16.3	**HD**[e]	**G19**A1G1	Q21	252	9 – 33	17.2	7.18
Xq12	**AR**	**G22**A1N5G6	Q23X5Q6	180	14 – 33	23.7	9.34
12p13.31	**ATN1**	G1A1G1A1**G15**	Q19	168	11 – 27	17.6	11.6
14q32.12	**ATXN3**	G2A1N1G1A1**G8**	Q3KQ10	168	10 – 27	17.8	29.2
2q37.1	TNRC15	**G6**	Q6	n.d.	n.d.	n.d.	n.d.

[a]Boldface text marks a gene known to cause disease by expansion of a polyglutamine-encoding CAG trinucleotide repeat. 'a' and 'b' after MAML3 and PCQAP denote two targets within these genes. Genes marked with an asterisk (*) contain an additional repeat target that was not screened in this study.

[b]G denotes "CAG", A denotes "CAA" and N denotes a non-glutamine codon, each followed by the number of tandem repeats of that codon. Boldface text marks the longest uninterrupted CAG-tract.

[c]X indicates a non-glutamine amino acid; SwP indicates peptide sequence obtained from SwissProt record

[d]N denotes number of alleles screened

[e]Data for N, Observed Q-tract Length Min-Max, Q-tract Mean, Q-tract Variance, taken from Andres *et al.* (26)

terrupted CAG-tracts from four to nine repeat units long but had mean Q-tract lengths evenly distributed from five to 26 residues (Table 1). Thus, a lack of polymorphism was restricted to relatively short pure CAG-tracts but their Q-tract lengths varied widely. This again emphasizes the utility of using pure CAG-tract length rather than Q-tract length in assessments of length polymorphism.

Each CAGpolyQ repeat has a unique distribution of Q-tract lengths

The two allele frequency distributions of Q-tract lengths in Figure 2 provide examples of the 64 CAGpolyQ repeats we analyzed. *ATXN3* had a unique bi- or tri-modal distribution that is virtually identical to published data [26]. *RAI1*, a priority candidate disease gene with a long CAG-tract and relatively high Q-tract variance, had a simpler distribution that is consistent with the published Q-tract length range [62]. The 64 plots of allele frequency distributions of Q-tract lengths for each CAGpolyQ repeat illustrate clearly that there is no single pattern that is typical of Q-tract length distributions across the human genome (Additional file 2).

Functional classification of CAGpolyQ-containing genes

Browsing descriptions associated with the 64 CAGpolyQ genes suggested an over-representation of genes involved in transcriptional processes and genes involved in chromatin architecture, and thyroid hormone receptor binding. We assessed these and other observations using GO-based classification of these genes to determine whether specific functional categories are statistically overrepresented, to visualize the network of functional relationships among CAGpolyQ-containing genes, and to determine whether priority candidates for polyglutamine expansion are associated with one or more specific GO terms.

GO over-representation analysis

We used GoMiner [67] to look for statistical over-representation of CAGpolyQ genes in GO terms in the top four levels of the three GO categories: biological process, molecular function, and cellular component. GO term descriptions can be viewed at the Gene Ontology website [68]. GoMiner contained gene name-GO term annotations for 56 of our 64 genes against a background of 13,598 HGNC genes. Genes without GO term assignments at the time of this analysis were: *C14orf4*, *C9orf43*, *CXorf6*, *DENND4B*, *FRMPD3*, *KIAA2018*, *TNRC15* and *TNS*. Our null hypothesis was that the genes of interest would be distributed among the chosen GO terms in the same proportions as the background set. GO terms with p-values below the significance threshold (p = 0.05) were considered to be over-represented among CAGpolyQ genes. In negative control experiments (see Methods) we found no over-representation in GO terms under molecular function in 100 replicates. Under biological process, three out of 100 replicates each had one over-represented

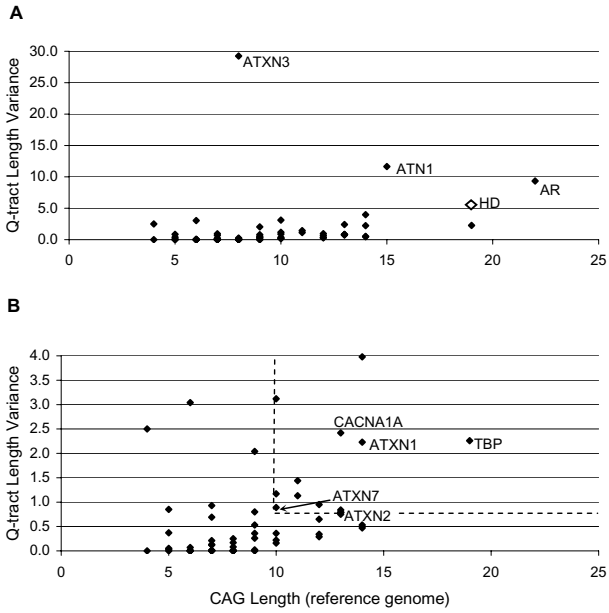

Figure 1
Relationship between length of longest uninterrupted CAG-tract and Q-tract length variance. (A) All targets. *HD* Q-tract length variance from Andres *et al.* [26]. Correlation = 0.62, not including *ATXN3*. (B) Higher resolution view of targets with Q-tract length variance < 4.0. Dashed lines at 10 CAG and 0.79 variance represent the cut-off for identifying candidate genes for polyglutamine expansion disorders. See text for list of genes falling in this area.

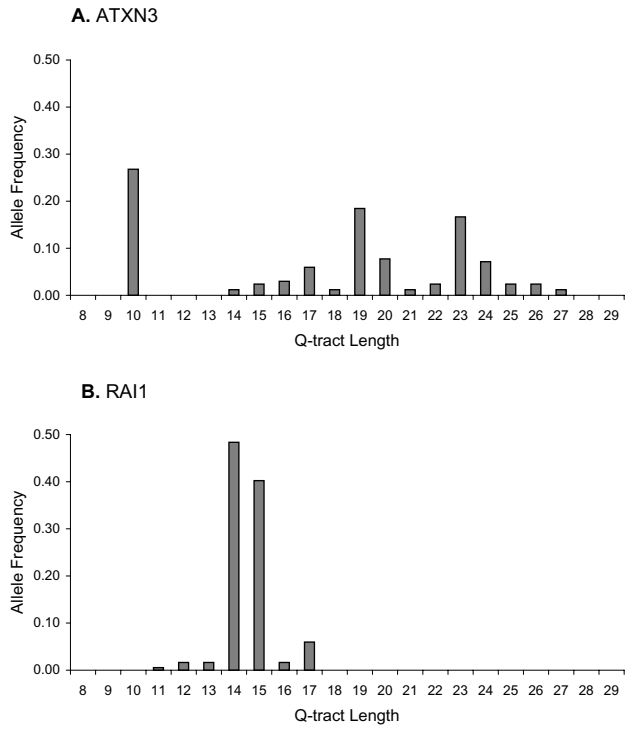

Figure 2
Example distributions of normal Q-tract lengths. (A) ATXN3, ataxin 3 (B) RAI1, retinoic acid receptor 1.

GO term. Under cellular component, one out of 100 replicates had one over-represented GO term and one out of 100 replicates had two over-represented GO terms.

Over-representation analysis confirmed these 56 CAGpolyQ genes' functional association with transcription and revealed some specific details. There were six significant GO terms under molecular function (Table 2). These included 13.4-fold over-representation of transcription coactivator activity, which is a child term of the 8.8-fold over-represented transcription cofactor activity. CAGpolyQ transcriptional coactivators on our gene list include: *ARID1B*, *CREBBP*, *MAML2*, *MAML3*, *MED12*, *MEF2A*, *NCOA3*, *NCOA6*, and *SMARCA2*. Transcription factor binding was 8.3-fold over-represented, including the transcription coactivator genes above, as well as *HD*, *NCOR2* and *TBP*. Half of the 56 genes bind DNA. There were five significant GO terms under biological process (Table 2), with the most specific, positive regulation of metabolism, 6.5-fold over-represented (*MAML2*, *CREBBP*, *RUNX2*, *ARID1B*, *NCOA6*, *NFAT5*, and *MAML3*). There were seven significant GO terms under cellular component (Table 2), with nucleoplasm 4.1-fold over-represented. Genes in over-represented GO categories are listed in Additional file 3 (Biological Process), Additional file 4 (Molecular Function) and Additional file 5 (Cellular Component).

Shared GO-term analysis

To delve deeper into the possible functional relationships among genes containing CAGpolyQ repeats, we developed a method for quantitative comparison of GO terms annotated to each gene product, based on the structure of the GO graph (AMM, SLB, BFFO, manuscript in preparation). Briefly, given a pair of genes, their GO term annotations, and a comparison scoring function for GO terms, we calculated similarity scores for every pair of GO terms for that pair of genes. GO term pairs scoring above a threshold were used to construct a graph where each node represents a gene and weighted edges between nodes represent pairs of GO term annotations and their scores. Genes were grouped by a simple visual clustering algorithm that assigns shorter lengths to edges with higher weights (i.e. more similar shared GO terms). Because a gene may have multiple shared GO terms with other genes, this method allowed us to cluster the functions of genes that share terms on different branches and at differ-

Table 2: Functional classification of CAGpolyQ genes: Gene Ontology over-representation analysis.

Gene Ontology term (levels) GO ID	Candidate genes in GO term	Fold* Enrichment
Biological Process		
regulation of biological process (**1**) GO:0050789	37	2.3
regulation of physiological process (**2**) GO:0050791	36	2.5
regulation of metabolism (**3**) GO:0019222	29	3.0
positive regulation of metabolism (**4**) GO:0009893	7	6.5
nucleobase, nucleoside, nucleotide and nucleic acid metabolism (4) GO:0006139	34	2.5
Molecular Function		
transcription regulator activity (**1**) GO:0030528	24	4.0
transcription cofactor activity (**2,**4) GO:0003712	11	8.8
transcription coactivator activity (**3**,5) GO:0003713	9	13.4
nucleic acid binding (2) GO:0003676	35	2.8
DNA binding (3) GO:0003677	28	3.1
transcription factor binding (3) GO:0008134	12	8.3
Cellular Component		
organelle (**1**) GO:0043226	43	1.7
membrane-bound organelle (2) GO:0043227	43	1.9
intracellular (2) GO:0005622	47	1.5
intracellular organelle (**2,**3) GO:0043229	43	1.7
intracellular membrane-bound organelle (**3**,4) GO:0043231	43	1.9
nucleus (3,**4**,5) GO:0005634	41	2.7
nucleoplasm (4,**5,**6) GO:0005654	11	4.1

All levels for each GO term are indicated, with boldface indicating one path through the GO
*$p < 0.00004$ for all GO terms listed except nucleoplasm, $p = 0.0001$.

ent levels of the gene ontology. Related functions go unnoticed without this clustering.

Only seven gene pairs scored above the cutoff (estimated 99th percentile; described in Methods) for the cellular component category (Additional file 6) so we did not consider this category further. There were 544 gene pairs with scores above the cutoff in the biological process category, representing 45 genes. There were 503 pairs among 42 genes in the molecular function category. The functional relationships among these CAGpolyQ genes are illustrated in Figure 3. GO terms and the genes that share them are listed in Additional file 7 (Biological Process), Additional file 8 (Molecular Function) and Additional file 6 (Cellular Component).

Based on our analysis of relationships among GO terms shared by two or more genes, CAGpolyQ genes in the human genome clustered primarily under two major biological processes: DNA dependent regulation of transcription, and neurogenesis (Figure 3A). Other processes included establishment and/or maintenance of chromatin architecture and post-translational modifications. Since there were few functional clusters, it was not surprising that all but one known disease gene and most priority candidate genes were involved in DNA dependent regulation of transcription and in neurogenesis (Figure 3A). ATXN7, the one disease gene excluded from the cluster involved in DNA dependent regulation of transcription, was recently shown to be an integral component of the TFTC (TATA-binding protein-free TAF-containing) complex and the STAGA (SPT3/TAF9/GCN5 acetyltransferase) complex involved in transcriptional regulation [52-54]. Consistent with their predominant classification in DNA dependent regulation of transcription, DNA binding was the primary shared molecular function among these 64 genes (Figure 3B). Known disease genes were involved in DNA, calcium and zinc binding and HD was classified as having transcription corepressor activity (Figure 3B). All but one priority candidate gene had DNA binding activity according to current GO annotations. CAGpolyQ genes with invariant Q-tract lengths were not limited to any one biological process or molecular function.

Discussion

Our findings build on previous work indicating that uninterrupted CAG-tract length, not the Q-tract length encoded by CAG plus CAA codons, influences the degree of polymorphism of a Q-tract. Uninterrupted CAG-tract length and Q-tract length variance are the most useful parameters in characterizing known disease genes and identifying candidate genes for expansion disorders. At one extreme, zero variance CAGpolyQ repeats – those that do not tolerate changes in Q-tract length – can likely be excluded as candidates for polyglutamine expansion disorders. The shapes of Q-tract length distributions differed widely between various loci across the genome. Thus, the data presented here for allele length distributions for 64

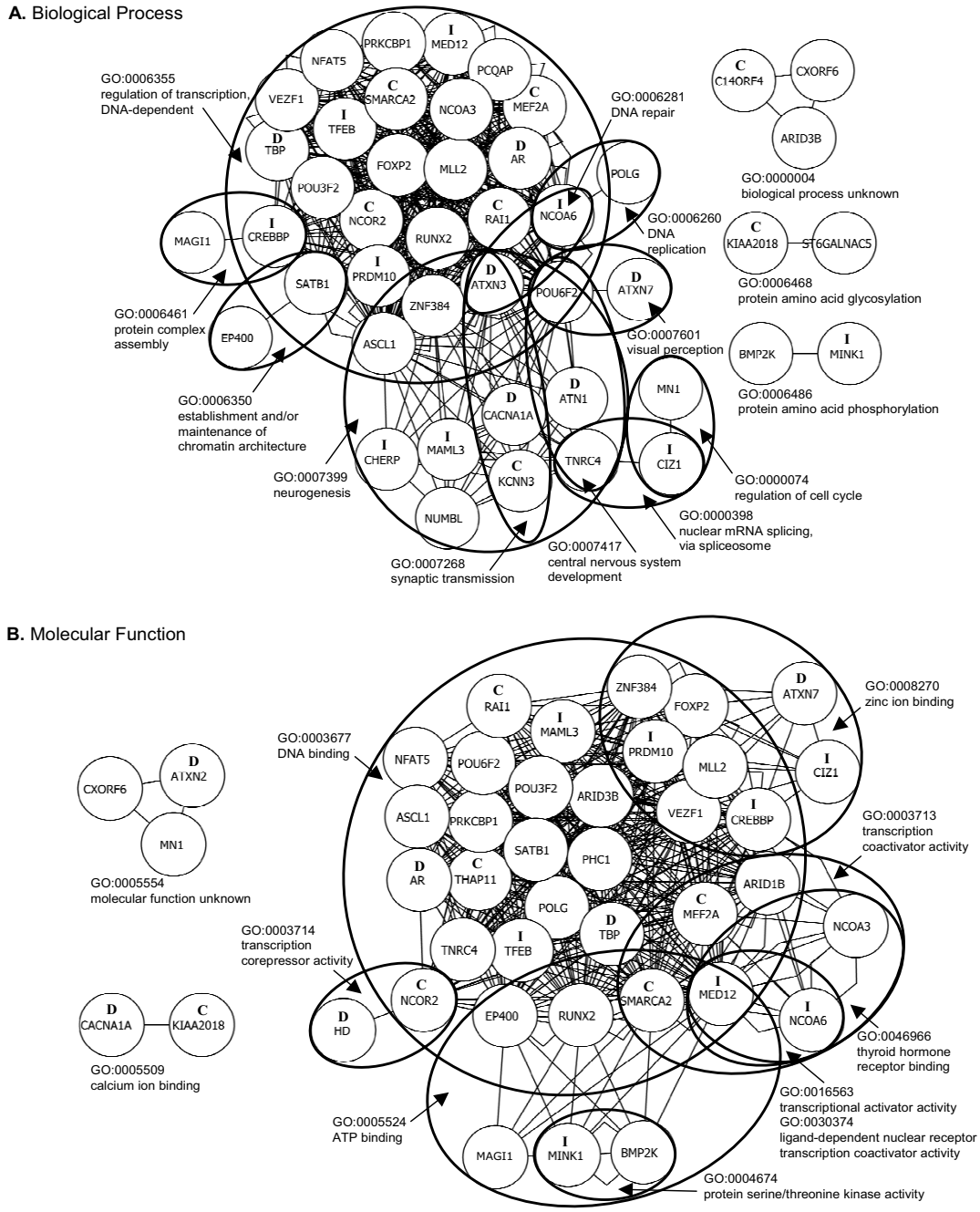

Figure 3

Functional classification of CAGpolyQ genes: shared Gene Ontology term analysis. Known disease genes are marked with a '**D**', candidate disease genes are marked with a '**C**' and genes with invariant Q-tracts (Table 1) are marked with an '**I**'. Clusters of genes are labeled with the GO terms that best described each cluster. GO terms shared by gene pairs are listed in Additional file 7 and Additional file 8. Genes not represented in a graph either had no annotation under that GO namespace or did not share a GO term with a score above the 99th percentile. (A) Biological process. Genes not represented: *ARID1B*, *ATXN1*, *ATXN2*, *BRD4*, *C9ORF43*, *DCP1B*, *HD*, *DENND4B*, *FRMPD3*, *MAML2*, *PAXIP1L*, *PHC1*, *PHLDA1*, *SOCS7*, *THAP11*, *TNRC15*, *TNRC6A*, *TNRC6B* and *TNS*. (B) Molecular function. Genes not represented: *ATN1*, *ATXN1*, *ATXN3*, *BRD4*, *C14ORF4*, *C9ORF43*, *CHERP*, *DCP1B*, *KCNN3*, *DENND4B*, *FRMPD3*, *MAML2*, *NUMBL*, *PAXIP1L*, *PCQAP*, *PHLDA1*, *SOCS7*, *ST6GALNAC5*, *TNRC15*, *TNRC6A*, *TNRC6B* and *TNS*.

Q-tracts in 62 genes with detailed conditions for their screening, will be invaluable for identifying putative expansion mutations in candidate genes not yet associated with CAGpolyQ-type neurodegenerative disorders. All nine known polyglutamine expansion disorder genes are involved in DNA-dependent regulation of transcription or in neurogenesis, as are all of the well-characterized priority candidate genes identified in this study.

Many groups have published lists of CAGpolyQ-containing genes identified using classical [15,17-20] or computational methods [21-24]. The content of each computationally-derived list differs slightly depending on the repeat detection algorithms and gene data sets used but they are largely the same. Tandem Repeat Finder, used in this study under default parameters, is not guaranteed to find all CAGpolyQ repeats, but it is likely that the vast majority of long repeats were found. Our approach is validated by its detection of all nine genes known to cause diseases by expansion of CAGpolyQ repeats. This study of the normal levels of polymorphism of human CAGpolyQ repeats is the most exhaustive conducted to date.

Our allele frequency distributions match those published for known disease genes *AR* [69], *ATN1* [2,3,26,70], *ATXN1* [26], *ATXN2* [26], *ATXN3* [26,70], *ATXN7* [70,71], *CACNA1A* [26,70], and *TBP* [70,72]. The same is true for CAGpolyQ repeats in other genes whose Q-tract lengths have been found to be invariant like *CREBBP* [25] and *MED12* [19], moderately polymorphic *FOXP2* [73], *NCOA3* [25,26,74], *POLG* [75], *RAI1* [76], *SMARCA2* [28] or highly polymorphic *THAP11* [28] and *KCNN3* [26,77]. Differences in apparent repeat lengths between this study and published data for *ATXN1* [26,70] and *ATXN3* [26,70] exist because we report repeat lengths based on the longest pure Q-tract while Andres *et al*. [26] and Juvonen *et al*. [70] report "repeat lengths" that contain non-glutamine amino acids. For *ATN1*, the shape of our distribution matches published data but our distribution is increased by two to four glutamine residues.

Among our eight priority candidate genes some features are already known. CAG length variation in *RAI1* is responsible for 4.1% of age of onset variability in SCA2 [76]. Huang *et al*. [42] identified RAI1 (called RAI2 in that paper) and NCOA3 as candidate disease genes by virtue of their long CAG tracts and the fact that their mouse and rat orthologues had Q-tracts less than half the size of the human repeats. In our study, NCOA3 lay just below the threshold for priority candidate disease genes, with nine CAG while priority candidates had ten CAG. *KCNN3* CAG-tract length differences have been associated with anorexia [78] and with schizophrenia and bipolar disorder but these associations are controversial [79]. *SMARCA2* and *THAP11* were previously identified as candidates by Pandey [28] based on their relatively long uninterrupted CAG-tracts. Four genes identified by Huang *et al*. [42] as candidate genes of interest fell far below our threshold of Q-tract variance so we do not consider them to be priority expansion disease candidates. These were DCP1B, MAML3 (called TNRC3 in Huang *et al*.), POLG (called NFYC in Huang *et al*.) and POU6F2 (called RPF-1 in Huang *et al*.).

Q-tract lengths for many genes do not have a normal distribution and differ widely between loci, as previously observed [27,36]. Even different disease genes have very different Q-tract length distribution shapes with different minima and maxima in normal populations and different minimum disease allele lengths so it is critical to characterize each distribution without making generalizations between loci. A gene containing more than one CAGpolyQ repeat can have two invariant repeats (*MAML3*) or a combination of invariant and variant repeats (*PCQAP*). Orthologous repeats in human and mouse genomes can have very different levels of polymorphism: human *VEZF1* has a polymorphic Q-tract (this study) while the corresponding Q-tract in its mouse orthologue is invariant [80].

Long pure repeats expand

Alba and colleagues [29,30,81] have clearly shown that, with respect to evolutionary processes, there are two classes of Q-tracts in human proteins: those whose lengths are conserved between human and mouse orthologues, and those whose lengths differ. Length-conserved polyQ repeats tend to be encoded by mixtures of CAG and CAA codons and are likely to be restricted in length by purifying selection. PolyQ repeats whose lengths vary between human and mouse tend to be encoded by longer pure CAG-tracts that evolve nearly neutrally [29,30,81]. Our data on Q-tract polymorphism within a normal human population corroborates their between species data and builds on previous work, with longer pure CAG-tracts having higher Q-tract length variance and invariant CAGpolyQ repeats having relatively short pure CAG-tracts [40,41]. Again, the extremes reinforce the rules; *FOXP2* with a short 5-CAG repeat has the longest mean Q-tract length of all candidate genes but a low level of polymorphism.

Correlation of uninterrupted CAG length with Q-tract length variance is consistent with work on dinucleotide repeats [82] and on tetra- and penta-nucleotide repeats [83]. For all of these, the level of polymorphism increases with the number of pure repeats, and non-polymorphic repeats have the shortest pure repeat tracts. Similarly, in the *HD* gene, as CAG repeat number increases, there is a significant increase in the frequency of expansion muta-

tions and the mean number of repeats added per expansion [34].

Pure CAG length is not the only factor determining repeat instability. An in-frame interruption in a CAG-tract has a stabilizing influence over and above that of reducing the pure CAG-tract length. In yeast, dinucleotide repeats with a single dinucleotide interruption in the middle of the tract are five times more stable than a pure repeat of the same length [84]. SCA1 disease alleles of the *ATXN1* gene all contain uninterrupted tracts of CAG repeats while virtually all normal alleles have one to three CAT (coding for histidine) interruptions in the middle of the Q-tract [39]. Other factors underlying repeat instability include different repair mechanisms [32], flanking sequence elements [85,86], CpG methylation, and nucleosome and replication origin positioning [86-88].

Rozanska *et al*. [89] recently published a large study that complements our results, analyzing repeat lengths and interruption patterns in a normal Polish population. They determined that the length of uninterrupted repeat tract in the most frequent allele for a locus is correlated with the degree of length polymorphism for that tract, and provide further evidence for a stabilizing effect of repeat interruptions. Trinucleotide repeat expansion disease genes were found to have a higher proportion of long repeat alleles than those not associated with disease [89].

Inferences about repeat lengths and disease prevalence

Lack of detailed reporting of repeat sequence lengths in disease genes, such as Q-tract lengths in *ATXN1* and *ATXN3* are a potential source of confusion in the literature and highlight the difficulties in comparing Q-tract length distributions for the same genes from different publications. The amino acid sequence of the most common normal *ATXN1* repeat tract is $Q_{12}H_1Q_1H_1Q_{14}$ [37] but it is frequently reported as 29 "repeats" and the *ATXN3* repeat tract, $Q_3K_1Q_{10}$, is reported as 14 "repeats" [26]. Non-glutamine interruptions in a Q-tract are critical to phenotype, so it is misleading to report these as "Q repeats" or "CAG repeats". For this reason, we reported all target Q-tract lengths based on the longest uninterrupted Q-tract (encoded by CAG/CAA) in the reference genome (Table 1, Additional file 2).

Measuring Q-tract lengths in affected individuals enables identification of putative repeat expansions outside the normal range, but more in depth characterization requires precise determination of the underlying amino acid and nucleotide sequences of individual alleles. Characterization of each allele at the nucleotide sequence level in addition to the normal (wild-type) Q-tract length distribution will be critical in better identifying candidate CAGpolyQ genes not yet associated with disease, determining which alleles at a given locus are prone to expansion, and for disease genes, characterizing allele repeat sequences with respect to disease prevalence in a given population [33]. As has been expertly laid out by Sobczak and Krzyzosiak [37] repeat interruption patterns in a given target can differ between populations, even when Q-tract length distributions are similar. Repeat interruption characteristics are not commonly studied, but reporting overall repeat lengths in the absence of repeat interruption patterns may be quite misleading in studies of allele lengths as they relate to disease prevalence in a given population [37,70,90]. Juvonen and colleagues [70] recently reported that the frequencies of large normal alleles at SCA loci were poor predictors of the prevalence of the respective diseases in Finland but Q-tract lengths were assayed without reporting CAG-tract interruption patterns in different alleles. A different picture might be revealed by characterization of repeat interruption patterns at each SCA locus in that population.

The genotype–phenotype connection

Q-tract length variance is influenced both by specific sequence characteristics and by the specific role of the Q-tract within a protein's structure and function. *AR* provides an excellent example of this balance. The *AR* Q-tract in the reference genome has a very long pure CAG-tract of 22 CAGs, consistent with its high length variance. The CAGpolyQ tract in the AR protein lies in its N-terminal transactivation domain which interacts with the C-terminal ligand binding domain (the N/C interaction). Buchanan *et al*. [69] found no changes in *in vitro* N/C interaction for Q-tract lengths of 16 to 29 but shorter or longer tracts resulted in a significant decrease in N/C interaction. Over 90% of normal alleles fall within the Q16-Q29 range both in this study and in Buchanan's re-examination of published data [69]. Q-tracts in *AR* equal to or longer than 38 glutamines cause the polyglutamine expansion disorder spinal and bulbar muscular atrophy while short Q-tracts are associated with increased risk of prostate cancer [69]. In other genes, Q-tracts with no length variation suggest the presence of strong purifying selection in which a precise Q-tract length is required to maintain a protein's structure or its biomolecular interactions, and its function. Therefore, a length change in a non-variant Q-tract is presumed to be lethal.

CAGpolyQ Gene Functions

Based on GO overrepresentation and shared-term analysis we find that CAGpolyQ genes are involved, in general, in two major biological processes, DNA dependent regulation of transcription and neurogenesis, and are enriched for transcriptional coactivator and transcription factor binding functions. Subgroups of genes such as known polyglutamine expansion disease genes, priority candidates, or genes containing invariant Q-tracts are not obvi-

ously distinguished by association with a particular process or molecular function. Polyglutamine-containing proteins in organisms from yeast to humans have been previously noted to be involved in transcriptional regulation [44-48]. In fact, most eukaryotic repeat containing proteins are involved in transcription or translation or interact directly with DNA, RNA or chromatin, irrespective of the amino acid repeat type [48]. The majority of repeat-containing proteins perform roles in processes that require the assembly of large multiprotein or protein/nucleic acid complexes [48]. Expanded Q-tracts in HD and ATN1 gene products interfere with CREBBP-activated gene transcription via interaction of their Q-rich domains [91,92] and mutant HD targets specific components of the core transcriptional machinery, in a Q-tract length-sensitive manner, to disrupt gene expression in cultured HD cells [55]. We anticipate that continual incorporation into the GO of newly published information about the normal functions of polyglutamine expansion disorder genes will reveal more specific shared functions among them.

Conclusion

We have characterized the levels of Q-tract length polymorphism in 64 CAGpolyQ repeat tracts in a normal human population, and found a strong positive correlation between uninterrupted CAG-tract length and Q-tract length variance. The best predictors of known disease genes were the occurrence of a long uninterrupted CAG-tract in the reference genome sequence and high Q-tract length variance in the normal population. Using these criteria we identified eight priority candidate genes for polyglutamine expansion disorders based on the presence of pure CAG-tracts longer and Q-tract variances higher than the smallest values in known disease genes. Twelve invariant Q-tracts (in eleven genes) are unlikely to be candidates for polyglutamine expansion disorders. Each CAGpolyQ repeat, including those in known disease genes, has a unique distribution of Q-tract lengths, emphasizing the need to characterize each distribution without making generalizations between loci. This publication makes freely available for the first time the length distributions of virtually all of the CAGpolyQ repeats in the human genome. Using these normal repeat distributions against which pathogenic expansions can be identified, we have begun screening for mutations in individuals clinically diagnosed with SCA or Huntington disease-like disorders who do not have identified mutations within known disease genes.

Methods

Selection of candidate genes

Candidate genes were identified on the basis of having a CAG-type simple repeat within the boundaries of a known gene with five or more tandem glutamine residues in the peptide sequence of that gene. To accomplish this, the Simple Repeats table (simpleRepeat.txt.gz) was downloaded from the UCSC genome annotation database [59] for build 33 (April 2003) of the human genome sequence assembly [58] and uploaded into a local MySQL database. The Simple Repeats table contained chromosomal location coordinates of all repeats detected by Tandem Repeat Finder (TRF) software [93] using default parameters. Locations of all the CAG-type repeats in this table were exported to a file using an SQL query to extract all records with the sequences 'CAG', 'AGC', 'CGA', 'CTG', 'GCT' and 'TCG' to accommodate all six potential reading frames of the repeat as they might appear in genomic sequence. This file was used as input to a Perl script that used the Ensembl Perl API [60] version 15_33 to extract all known genes (Ensembl-predicted transcripts that map to species-specific SwissProt, RefSeq or TrEMBL database entries) whose chromosomal coordinates overlapped with the repeat coordinates. For each known gene with a CAG-type repeat, if the Ensembl peptide sequence contained five or more glutamine residues in tandem, that gene was considered a candidate. A minimum glutamine repeat length of five was used since Karlin [94] determined that for a "typical" protein of 400 residues and average composition, a run of an individual amino acid is statistically significant if it is five or more residues long [94].

The candidate gene list was generated from Build 33 of the human genome sequence assembly (April 2003), and the nucleotide/amino acid sequences of each glutamine tract reported in Table 1 were generated from Build 35 (May 2004). Two new candidate genes were identified in the later build (Ensembl known genes data set version 30_35c) that were not part of our study: *MKL1* and *C14orf43*, and additional CAGpolyQ repeats were detected in nine of our existing candidate genes: *FOXP2*, *MAML3*, *MED12*, *MINK1*, *MLL2*, *MN1*, *NCOA6*, *NFAT5*, *PAXIP1L*. These targets have been denoted by an asterisk in Table 1. Chromosome band was obtained from the UCSC Chromosome band track [95] and may differ slightly from a gene's location listed by the HGNC Database, Genew [14]. Gene names listed are official HGNC gene symbols from the HGNC website [96] (accessed March 13, 2007).

DNA samples

Control DNA samples (extracted from blood) were from a population of mixed ethnic background with individuals of Western European descent most highly represented (Additional file 1). 48 of these were from the Coriell Cell Repository [97].

PCR primers and amplification of candidate repeats

Additional file 9 lists primer sequences, annealing temperatures, specific PCR conditions and expected fragment

size (from the reference genome) for each repeat target. PCR primers for candidate repeat amplification were designed using Primer3 [98]. Forward primers were 5'-labeled with 5-HEX, 6-FAM or TAMRA fluorescent dyes (Operon) and reverse primers all had a 5'-GTTT "PIG-tail" [99]. PCR amplification was performed with standard Taq polymerase (Invitrogen) or AccuPrime Taq polymerase (Invitrogen) in 96-well plates according to the conditions specified for each target in Additional file 9. PCR products were visualized and quantitated by comparing the signal intensity of a specific volume of PCR product against 4 μl of Low DNA Mass Ladder (Invitrogen) on an agarose gel. The accuracy of this quantitation method was validated against the PicoGreen® dsDNA Quantitation assay (Molecular Probes) [100].

ABI 3700 fragment analysis and GeneMapper band calling

PCR products for fragment sizing were assembled in 96-well microtiter plates at 0.5 ng/μl in each well, with up to six PCR products multiplexed per well according to their predicted allele sizes and fluorescent labels. One microliter of the multiplexed PCR products was added to 9 μl of either 2% 400 HD [ROX] sizing standard (Applied Biosystems) or 2% 500 [ROX] sizing standard (Applied Biosystems) depending on the estimated sizes of products being analyzed. DNA fragments were separated by capillary electrophoresis using the ABI Prism 3700 DNA Analyzer (Applied Biosystems) with POP-6 polymer (Applied Biosystems). Sizing of the PCR fragments was accomplished using GeneMapper software (v.3.0, Applied Biosystems). Representative alleles from each locus were sequenced to determine the exact correspondence between fragment size and Q-tract length. In all cases (except TNRC15, for which we do not present data), fragment length polymorphism was entirely accounted for by changes in Q-tract length. At least one such sequenced allele was included on every run as a calibrator.

Data management and analysis

Repeat information, PCR conditions, sample information and analysis results were stored in a MySQL database called GeMSdb (Genomic Mutational Signature sequences database). Data was input into GeMSdb using Perl scripts and through a web interface built with PHP and Apache. Data analysis and graphics were done using PHP.

The Q-tract length of each allele was based on the difference between observed PCR fragment size from a DNA sample and expected PCR fragment size from the reference genome (plus 4 nucleotides from the primer tail). Expected fragment sizes and Q-tract lengths (reference genome Build 35) for every target are listed in Additional file 9. Q-tract $length_{Exp}$ below is that of the longest uninterrupted Q-tract in the target. For example, the *ATXN1* Q-tract ($Q_{12}H_1Q_1H_1Q_{14}$) $length_{Exp}$ is 14 because the overall repeat region of 29 residues is interrupted by two non-glutamine amino acids.

$$\text{Q-tract length}_{Obs} = (\text{Fragment size}_{Obs} - \text{Fragment size}_{Exp})/3 + \text{Q-tract length}_{Exp}$$

Repeat purity was calculated as a normalized weighted measure, nWP, combining both the length of the longest uninterrupted CAG-tract (CAG-length) and the total Q-tract length (Q-length) of each repeat. Weighted purity (WP) for each repeat was normalized by dividing by the highest WP among loci, which was 21.04 for *AR*.

$$\text{nWP} = (\text{CAG-length}/\text{Q-length}) * \text{CAG-length}/21.04$$

Statistical analysis

Because there was no *a priori* knowledge of the distribution of Q-tract lengths in each gene for the typical control population, we applied the statistics of tolerance levels to determine the number of control alleles that must be screened to distinguish a Q-tract length that occurs in the affected but not unaffected populations with a given level of confidence. Screening 130 control alleles provides us with 99% confidence that 95% of the population of interest lies between the minimum and maximum repeat lengths in our samples [101].

Gene expression

Candidate genes' expression in brain was determined according to either eVOC controlled vocabularies for gene expression data [63,64] queried through BioMart [65] or according to expression data at the GeneCards website [66] (accessed September 19, 2005).

Gene functional classification

Gene Ontology over-representation analysis

We used GoMiner [67] for GO over-representation analysis down to the fourth level in the ontology. The target and background gene sets were generated as follows. We downloaded 23,913 HGNC gene IDs on June 28, 2005 from the HGNC website [96]. All IDs ending in '~withdrawn' were removed to generate a list of 21,591 IDs used as the 'query gene file' for GoMiner. GoMiner matched 13,598 of these to GO terms. We conducted 100 negative control replicates of this experiment for the three GO categories, each replicate with 56 randomly selected genes out of the 13,598 background gene set. To correct for multiple testing we used a Bonferroni correction to adjust the threshold of significance appropriately. The raw threshold of significance was $p = 0.05$. Adjusted significance thresholds were: molecular function $p = 0.00004$; biological process $p = 0.00005$; cellular component $p = 0.00009$.

Graph-based shared Gene Ontology term analysis

For each pair of genes among our set of 64, the GO terms annotated to each gene were compared and we calculated a graph-based similarity measure (AMM, SLB, BFFO, manuscript in preparation) for all gene pairs. In order to determine significant scores and produce a meaningful subgraph, we bootstrapped an estimate of the score required to be above the 99th percentile for a set of genes of that size (64) from the background set. We randomly drew 1000 replicates from the set of 15,168 Entrez Gene human protein-coding genes and took the mean of the 99th percentile score for each GO namespace (biological process, molecular function and cellular component) as our cut-off value. Pairs of genes with shared GO terms scoring above the cut-off value were visualized using Cytoscape 2.1 [102] with the "organic" arrangement of nodes, which produced a natural set of clusters. The "organic" node arrangement treats edges as springs: the more edges among a group of nodes, the tighter they cluster. The pairwise similarity measure links GO terms via their lowest common ancestor term in the graph. These lowest common ancestor terms are output with each pair of GO terms that are scored, and can be considered as edge labels in the resulting graph. Clusters of genes joined by the same GO term edge labels were manually annotated with those GO terms.

Abbreviations

CAGpolyQ, polyglutamine-encoding CAG trinucleotide repeat; Q-tract, polyglutamine tract; HD, Huntington disease; SCA, spinocerebellar ataxia; ATN1, atrophin1; AR, androgen receptor; TBP, TATA-binding protein; ATXN, ataxin; GeMS, Genomic Mutational Signature; GO, Gene Ontology; HGNC, Human Gene Nomenclature Committee

Authors' contributions

SLB, RSD, BRL, BFFO and RAH conceived and designed the experiments. SLB, RSD, CLM, AMM, SJN, SSL, AW, GSY and MMSY performed the experiments. SLB, YH, SJN, CLM and AMM analyzed the data. CLM and YH designed and managed the database. MRH and RAH contributed reagents/materials. SLB wrote the paper and all authors read and approved the final manuscript.

Additional material

Additional file 1

Ethnic composition of control population.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-126-S1.xls]

Additional file 2

Allele length distributions in a normal population for 64 polyglutamine-encoding CAG trinucleotide repeat targets (A) – (BL). This multi-page document provides plots of allele frequency distributions.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-126-S2.pdf]

Additional file 3

Genes in over-represented GO terms under Biological Process. For each over-represented GO term and its GO ID, this document lists the CAGpolyQ repeat-containing genes that were annotated with that GO term.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-126-S3.pdf]

Additional file 4

Genes in over-represented GO terms under Molecular Function. For each over-represented GO term and its GO ID, this document lists the CAGpolyQ repeat-containing genes that were annotated with that GO term.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-126-S4.pdf]

Additional file 5

Genes in over-represented GO terms under Cellular Component. For each over-represented GO term and its GO ID, this document lists the CAGpolyQ repeat-containing genes that were annotated with that GO term.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-126-S5.pdf]

Additional file 6

Genes and their shared GO terms under Cellular Component. This document provides GO IDs, their descriptions, and the lists of CAGpolyQ repeat-containing genes that shared these annotations above the 99th percentile cutoff.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-126-S6.pdf]

Additional file 7

Genes and their shared GO terms under Biological Process. This document provides GO IDs, their descriptions, and the lists of CAGpolyQ repeat-containing genes that shared these annotations above the 99th percentile cutoff.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-126-S7.pdf]

Additional file 8

Genes and their shared GO terms under Molecular Function. This document provides GO IDs, their descriptions, and the lists of CAGpolyQ repeat-containing genes that shared these annotations above the 99th percentile cutoff.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-126-S8.pdf]

Additional file 9

Conditions for PCR amplification of CAGpolyQ repeats in 64 CAGpolyQ repeats. This table provides primer sequences, annealing temperatures and expected fragment sizes for screening these repeats.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-8-126-S9.xls]

Acknowledgements

This study has been approved by the University of British Columbia Clinical Research Ethics Board. The authors wish to thank Christopher Pearson and Simon Warby for helpful discussions, Terry Pape for suggesting a critical experiment, Ian Bosdet and Jacquie Schein for early technology development, Elizabeth Simpson for Coriell controls, and Clinical Research Support at Children's and Women's Health Centre of British Columbia for statistical consulting services. Funding for this study was provided by the Canadian Genetic Diseases Network, the National Organization for Rare Disorders, and the University of British Columbia. RAH is a Michael Smith Foundation for Health Research Scholar and AMM was funded by the Natural Sciences and Engineering Research Council of Canada.

References

1. **A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group.** *Cell* 1993, **72:**971-983.
2. Koide R, Ikeuchi T, Onodera O, Tanaka H, Igarashi S, Endo K, Takahashi H, Kondo R, Ishikawa A, Hayashi T, *et al.*: **Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA).** *Nat Genet* 1994, **6:**9-13.
3. Nagafuchi S, Yanagisawa H, Sato K, Shirayama T, Ohsaki E, Bundo M, Takeda T, Tadokoro K, Kondo I, Murayama N, *et al.*: **Dentatorubral and pallidoluysian atrophy expansion of an unstable CAG trinucleotide on chromosome 12p.** *Nat Genet* 1994, **6:**14-18.
4. La Spada AR, Wilson EM, Lubahn DB, Harding AE, Fischbeck KH: **Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy.** *Nature* 1991, **352:**77-79.
5. Zhuchenko O, Bailey J, Bonnen P, Ashizawa T, Stockton DW, Amos C, Dobyns WB, Subramony SH, Zoghbi HY, Lee CC: **Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the alpha 1A-voltage-dependent calcium channel.** *Nat Genet* 1997, **15:**62-69.
6. Nakamura K, Jeong SY, Uchihara T, Anno M, Nagashima K, Nagashima T, Ikeda S, Tsuji S, Kanazawa I: **SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein.** *Hum Mol Genet* 2001, **10:**1441-1448.
7. Orr HT, Chung MY, Banfi S, Kwiatkowski TJ Jr., Servadio A, Beaudet AL, McCall AE, Duvick LA, Ranum LP, Zoghbi HY: **Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1.** *Nat Genet* 1993, **4:**221-226.
8. Imbert G, Saudou F, Yvert G, Devys D, Trottier Y, Garnier JM, Weber C, Mandel JL, Cancel G, Abbas N, Durr A, Didierjean O, Stevanin G, Agid Y, Brice A: **Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats.** *Nat Genet* 1996, **14:**285-291.
9. Sanpei K, Takano H, Igarashi S, Sato T, Oyake M, Sasaki H, Wakisaka A, Tashiro K, Ishida Y, Ikeuchi T, Koide R, Saito M, Sato A, Tanaka T, Hanyu S, Takiyama Y, Nishizawa M, Shimizu N, Nomura Y, Segawa M, Iwabuchi K, Eguchi I, Tanaka H, Takahashi H, Tsuji S: **Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT.** *Nat Genet* 1996, **14:**277-284.
10. Pulst SM, Nechiporuk A, Nechiporuk T, Gispert S, Chen XN, Lopes-Cendes I, Pearlman S, Starkman S, Orozco-Diaz G, Lunkes A, DeJong P, Rouleau GA, Auburger G, Korenberg JR, Figueroa C, Sahba S: **Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2.** *Nat Genet* 1996, **14:**269-276.
11. Kawaguchi Y, Okamoto T, Taniwaki M, Aizawa M, Inoue M, Katayama S, Kawakami H, Nakamura S, Nishimura M, Akiguchi I, *et al.*: **CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1.** *Nat Genet* 1994, **8:**221-228.
12. David G, Abbas N, Stevanin G, Durr A, Yvert G, Cancel G, Weber C, Imbert G, Saudou F, Antoniou E, Drabkin H, Gemmill R, Giunti P, Benomar A, Wood N, Ruberg M, Agid Y, Mandel JL, Brice A: **Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion.** *Nat Genet* 1997, **17:**65-70.
13. Rudnicki DD, Margolis RL: **Repeat expansion and autosomal dominant neurodegenerative disorders: consensus and controversy.** *Expert Rev Mol Med* 2003, **2003:**1-24.
14. Wain HM, Lush MJ, Ducluzeau F, Khodiyar VK, Povey S: **Genew: the Human Gene Nomenclature Database, 2004 updates.** *Nucleic Acids Res* 2004, **32:**D255-7.
15. Gastier JM, Brody T, Pulido JC, Businga T, Sunden S, Hu X, Maitra S, Buetow KH, Murray JC, Sheffield VC, Boguski M, Duyk GM, Hudson TJ: **Development of a screening set for new (CAG/CTG)n dynamic mutations.** *Genomics* 1996, **32:**75-85.
16. Li SH, McInnis MG, Margolis RL, Antonarakis SE, Ross CA: **Novel triplet repeat containing genes in human brain: cloning, expression, and length polymorphisms.** *Genomics* 1993, **16:**572-579.
17. Riggins GJ, Lokey LK, Chastain JL, Leiner HA, Sherman SL, Wilkinson KD, Warren ST: **Human genes containing polymorphic trinucleotide repeats.** *Nat Genet* 1992, **2:**186-191.
18. Reddy PH, Stockburger E, Gillevet P, Tagle DA: **Mapping and characterization of novel (CAG)n repeat cDNAs from adult human brain derived by the oligo capture method.** *Genomics* 1997, **46:**174-182.
19. Margolis RL, Abraham MR, Gatchell SB, Li SH, Kidwai AS, Breschel TS, Stine OC, Callahan C, McInnis MG, Ross CA: **cDNAs with long CAG trinucleotide repeats from human brain.** *Hum Genet* 1997, **100:**114-122.
20. Schalling M, Hudson TJ, Buetow KH, Housman DE: **Direct detection of novel expanded trinucleotide repeats in the human genome.** *Nat Genet* 1993, **4:**135-139.
21. Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ: **Amino acid runs in eukaryotic proteomes and disease associations.** *Proc Natl Acad Sci U S A* 2002, **99:**333-338.
22. Collins JR, Stephens RM, Gold B, Long B, Dean M, Burt SK: **An exhaustive DNA micro-satellite map of the human genome using high performance computing.** *Genomics* 2003, **82:**10-19.
23. Subramanian S, Madgula VM, George R, Mishra RK, Pandit MW, Kumar CS, Singh L: **Triplet repeats in human genome: distribution and their association with genes and other genomic regions.** *Bioinformatics* 2003, **19:**549-552.
24. Jasinska A, Michlewski G, de Mezer M, Sobczak K, Kozlowski P, Napierala M, Krzyzosiak WJ: **Structures of trinucleotide repeats in human transcripts and their functional implications.** *Nucleic Acids Res* 2003, **31:**5463-5468.
25. Hayashi Y, Yamamoto M, Ohmori S, Kikumori T, Imai T, Funahashi H, Seo H: **Polymorphism of homopolymeric glutamines in coactivators for nuclear hormone receptors.** *Endocr J* 1999, **46:**279-284.
26. Andres AM, Lao O, Soldevila M, Calafell F, Bertranpetit J: **Dynamics of CAG repeat loci revealed by the analysis of their variability.** *Hum Mutat* 2003, **21:**61-70.
27. Edwards A, Hammond HA, Jin L, Caskey CT, Chakraborty R: **Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups.** *Genomics* 1992, **12:**241-253.
28. Pandey N, Mittal U, Srivastava AK, Mukerji M: **SMARCA2 and THAP11: potential candidates for polyglutamine disorders as evidenced from polymorphism and protein-folding simulation studies.** *J Hum Genet* 2004, **49:**596-602.
29. Alba MM, Santibanez-Koref MF, Hancock JM: **Conservation of polyglutamine tract size between mice and humans depends on codon interruption.** *Mol Biol Evol* 1999, **16:**1641-1644.
30. Alba MM, Santibanez-Koref MF, Hancock JM: **The comparative genomics of polyglutamine repeats: extreme differences in the codon organization of repeat-encoding regions between mammals and Drosophila.** *J Mol Evol* 2001, **52:**249-259.

31. Levinson G, Gutman GA: **Slipped-strand mispairing: a major mechanism for DNA sequence evolution.** *Mol Biol Evol* 1987, **4:**203-221.
32. Pearson CE, Edamura KN, Cleary JD: **Repeat instability: mechanisms of dynamic mutations.** *Nat Rev Genet* 2005, **6:**729-742.
33. Squitieri F, Andrew SE, Goldberg YP, Kremer B, Spence N, Zeisler J, Nichol K, Theilmann J, Greenberg J, Goto J, *et al.*: **DNA haplotype analysis of Huntington disease reveals clues to the origins and mechanisms of CAG expansion and reasons for geographic variations of prevalence.** *Hum Mol Genet* 1994, **3:**2103-2114.
34. Leeflang EP, Zhang L, Tavare S, Hubert R, Srinidhi J, MacDonald ME, Myers RH, de Young M, Wexler NS, Gusella JF, *et al.*: **Single sperm analysis of the trinucleotide repeats in the Huntington's disease gene: quantification of the mutation frequency spectrum.** *Hum Mol Genet* 1995, **4:**1519-1526.
35. Telenius H, Kremer HP, Theilmann J, Andrew SE, Almqvist E, Anvret M, Greenberg C, Greenberg J, Lucotte G, Squitieri F, *et al.*: **Molecular analysis of juvenile Huntington disease: the major influence on (CAG)n repeat length is the sex of the affected parent.** *Hum Mol Genet* 1993, **2:**1535-1540.
36. Jodice C, Giovannone B, Calabresi V, Bellocchi M, Terrenato L, Novelletto A: **Population variation analysis at nine loci containing expressed trinucleotide repeats.** *Ann Hum Genet* 1997, **61:**425-438.
37. Sobczak K, Krzyzosiak WJ: **Patterns of CAG repeat interruptions in SCA1 and SCA2 genes in relation to repeat instability.** *Hum Mutat* 2004, **24:**236-247.
38. **GeneReviews at GeneTests: Medical Genetics Information Resource** [http://www.genetests.org]
39. Chung MY, Ranum LP, Duvick LA, Servadio A, Zoghbi HY, Orr HT: **Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I.** *Nat Genet* 1993, **5:**254-258.
40. Wren JD, Forgacs E, Fondon JW 3rd, 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* 2000, **67:**345-356.
41. Mularoni L, Guigo R, Alba MM: **Mutation patterns of amino acid tandem repeats in the human proteome.** *Genome Biol* 2006, **7:**R33.
42. Huang H, Winter EE, Wang H, Weinstock KG, Xing H, Goodstadt L, Stenson PD, Cooper DN, Smith D, Alba MM, Ponting CP, Fechtel K: **Evolutionary conservation and selection of human disease gene orthologs in the rat and mouse genomes.** *Genome Biol* 2004, **5:**R47.
43. Mitchell PJ, Tjian R: **Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins.** *Science* 1989, **245:**371-378.
44. Bhandari R, Brahmachari SK: **Analysis of CAG/CTG triplet repeats in the human genome: Implication in transcription factor gene regulation.** *Journal of biosciences* 1995, **20:**613-627.
45. Karlin S, Burge C: **Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development.** *Proc Natl Acad Sci U S A* 1996, **93:**1560-1565.
46. Alba MM, Santibanez-Koref MF, Hancock JM: **Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process.** *J Mol Evol* 1999, **49:**789-797.
47. Alba MM, Guigo R: **Comparative analysis of amino acid repeats in rodents and humans.** *Genome Res* 2004, **14:**549-554.
48. Faux NG, Bottomley SP, Lesk AM, Irving JA, Morrison JR, de la Banda MG, Whisstock JC: **Functional insights from the distribution and role of homopeptide repeat-containing proteins.** *Genome Res* 2005, **15:**537-551.
49. Dunah AW, Jeong H, Griffin A, Kim YM, Standaert DG, Hersch SM, Mouradian MM, Young AB, Tanese N, Krainc D: **Sp1 and TAFII130 transcriptional activity disrupted in early Huntington's disease.** *Science* 2002, **296:**2238-2243.
50. Freiman RN, Tjian R: **Neurodegeneration. A glutamine-rich trail leads to transcription factors.** *Science* 2002, **296:**2149-2150.
51. van Roon-Mom WM, Reid SJ, Faull RL, Snell RG: **TATA-binding protein in neurodegenerative disease.** *Neuroscience* 2005, **133:**863-872.
52. Helmlinger D, Hardy S, Sasorith S, Klein F, Robert F, Weber C, Miguet L, Potier N, Van-Dorsselaer A, Wurtz JM, Mandel JL, Tora L, Devys D: **Ataxin-7 is a subunit of GCN5 histone acetyltransferase-containing complexes.** *Hum Mol Genet* 2004, **13:**1257-1265.
53. Palhan VB, Chen S, Peng GH, Tjernberg A, Gamper AM, Fan Y, Chait BT, La Spada AR, Roeder RG: **Polyglutamine-expanded ataxin-7 inhibits STAGA histone acetyltransferase activity to produce retinal degeneration.** *Proc Natl Acad Sci U S A* 2005, **102:**8472-8477.
54. McMahon SJ, Pray-Grant MG, Schieltz D, Yates JR 3rd, Grant PA: **Polyglutamine-expanded spinocerebellar ataxia-7 protein disrupts normal SAGA and SLIK histone acetyltransferase activity.** *Proc Natl Acad Sci U S A* 2005, **102:**8478-8482.
55. Zhai W, Jeong H, Cui L, Krainc D, Tjian R: **In vitro analysis of huntingtin-mediated transcriptional repression reveals multiple transcription factor targets.** *Cell* 2005, **123:**1241-1253.
56. Ralser M, Albrecht M, Nonhoff U, Lengauer T, Lehrach H, Krobitsch S: **An integrative approach to gain insights into the cellular function of human ataxin-2.** *J Mol Biol* 2005, **346:**203-214.
57. Irwin S, Vandelft M, Pinchev D, Howell JL, Graczyk J, Orr HT, Truant R: **RNA association and nucleocytoplasmic shuttling by ataxin-1.** *J Cell Sci* 2005, **118:**233-242.
58. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
59. Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu YT, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, Weber RJ, Haussler D, Kent WJ: **The UCSC Genome Browser Database.** *Nucl Acids Res* 2003, **31:**51-54.
60. Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Gilbert J, Hammond M, Herrero J, Hotz H, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Kokocinsci F, London D, Longden I, McVicker G, Melsopp

C, Meidl P, Potter S, Proctor G, Rae M, Rios D, Schuster M, Searle S, Severin J, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Birney E: **Ensembl 2005.** *Nucleic Acids Res* 2005, **33:**D447-53.

61. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
62. Lavoie H, Debeane F, Trinh QD, Turcotte JF, Corbeil-Girard LP, Dicaire MJ, Saint-Denis A, Page M, Rouleau GA, Brais B: **Polymorphism, shared functions and convergent evolution of genes with sequences coding for polyalanine domains.** *Hum Mol Genet* 2003, **12:**2967-2979.
63. Kelso J, Visagie J, Theiler G, Christoffels A, Bardien S, Smedley D, Otgaar D, Greyling G, Jongeneel CV, McCarthy MI, Hide T, Hide W: **eVOC: a controlled vocabulary for unifying gene expression data.** *Genome Res* 2003, **13:**1222-1230.
64. Hide W, Smedley D, McCarthy M, Kelso J: **Application of eVOC: controlled vocabularies for unifying gene expression data.** *C R Biol* 2003, **326:**1089-1096.
65. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, Hammond M, Rocca-Serra P, Cox T, Birney E: **EnsMart: a generic system for fast and flexible access to biological data.** *Genome Res* 2004, **14:**160-169.
66. Rebhan M, Chalifa-Caspi V, Prilusky J: **GeneCards: encyclopedia for genes, proteins and diseases.** [http://www.genecards.org].
67. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, Narasimhan S, Kane DW, Reinhold WC, Lababidi S, Bussey KJ, Riss J, Barrett JC, Weinstein JN: **GoMiner: a resource for biological interpretation of genomic and proteomic data.** *Genome Biol* 2003, **4:**R28.
68. **Gene Ontology** [http://www.geneontology.org]
69. Buchanan G, Yang M, Cheong A, Harris JM, Irvine RA, Lambert PF, Moore NL, Raynor M, Neufing PJ, Coetzee GA, Tilley WD: **Structural and functional consequences of glutamine tract variation in the androgen receptor.** *Hum Mol Genet* 2004, **13:**1677-1692.
70. Juvonen V, Hietala M, Kairisto V, Savontaus ML: **The occurrence of dominant spinocerebellar ataxias among 251 Finnish ataxia patients and the role of predisposing large normal alleles in a genetically isolated population.** *Acta Neurol Scand* 2005, **111:**154-162.
71. Gouw LG, Castaneda MA, McKenna CK, Digre KB, Pulst SM, Perlman S, Lee MS, Gomez C, Fischbeck K, Gagnon D, Storey E, Bird T, Jeri FR, Ptacek LJ: **Analysis of the dynamic mutation in the SCA7 gene shows marked parental effects on CAG repeat transmission.** *Hum Mol Genet* 1998, **7:**525-532.
72. Zuhlke C, Hellenbroich Y, Dalski A, Kononowa N, Hagenah J, Vieregge P, Riess O, Klein C, Schwinger E: **Different types of repeat expansion in the TATA-binding protein gene are associated with a new form of inherited ataxia.** *Eur J Hum Genet* 2001, **9:**160-164.
73. Bruce HA, Margolis RL: **FOXP2: novel exons, splice variants, and CAG repeat length stability.** *Hum Genet* 2002, **111:**136-144.
74. Dai P, Wong LJ: **Somatic instability of the DNA sequences encoding the polymorphic polyglutamine tract of the AIB1 gene.** *J Med Genet* 2003, **40:**885-890.
75. Rovio AT, Abel J, Ahola AL, Andres AM, Bertranpetit J, Blancher A, Bontrop RE, Chemnick LG, Cooke HJ, Cummins JM, Davis HA, Elliott DJ, Fritsche E, Hargreave TB, Hoffman SM, Jequier AM, Kao SH, Kim HS, Marchington DR, Mehmet D, Otting N, Poulton J, Ryder OA, Schuppe HC, Takenaka O, Wei YH, Wichmann L, Jacobs HT: **A prevalent POLG CAG microsatellite length allele in humans and African great apes.** *Mamm Genome* 2004, **15:**492-502.
76. Hayes S, Turecki G, Brisebois K, Lopes-Cendes I, Gaspar C, Riess O, Ranum LP, Pulst SM, Rouleau GA: **CAG repeat length in RAI1 is associated with age at onset variability in spinocerebellar ataxia type 2 (SCA2).** *Hum Mol Genet* 2000, **9:**1753-1758.
77. Figueroa KP, Chan P, Schols L, Tanner C, Riess O, Perlman SL, Geschwind DH, Pulst SM: **Association of moderate polyglutamine tract expansions in the slow calcium-activated potassium channel type 3 with ataxia.** *Arch Neurol* 2001, **58:**1649-1653.
78. Koronyo-Hamaoui M, Gak E, Stein D, Frisch A, Danziger Y, Leor S, Michaelovsky E, Laufer N, Carel C, Fennig S, Mimouni M, Apter A, Goldman B, Barkai G, Weizman A: **CAG repeat polymorphism within the KCNN3 gene is a significant contributor to susceptibility to anorexia nervosa: a case-control study of female patients and several ethnic groups in the Israeli Jewish population.** *Am J Med Genet B Neuropsychiatr Genet* 2004, **131:**76-80.
79. Tsutsumi T, Holmes SE, McInnis MG, Sawa A, Callahan C, DePaulo JR, Ross CA, DeLisi LE, Margolis RL: **Novel CAG/CTG repeat expansion mutations do not contribute to the genetic risk for most cases of bipolar disorder or schizophrenia.** *Am J Med Genet B Neuropsychiatr Genet* 2004, **124:**15-19.
80. Ogasawara M, Imanishi T, Moriwaki K, Gaudieri S, Tsuda H, Hashimoto H, Shiroishi T, Gojobori T, Koide T: **Length variation of CAG/CAA triplet repeats in 50 genes among 16 inbred mouse strains.** *Gene* 2005, **349:**107-119.
81. Hancock JM, Worthey EA, Santibanez-Koref MF: **A role for selection in regulating the evolutionary emergence of disease-causing and other coding CAG repeats in humans and mice.** *Mol Biol Evol* 2001, **18:**1014-1023.
82. Weber JL: **Informativeness of human (dC-dA)n.(dG-dT)n polymorphisms.** *Genomics* 1990, **7:**524-530.
83. Brinkmann B, Klintschar M, Neuhuber F, Huhne J, Rolf B: **Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat.** *Am J Hum Genet* 1998, **62:**1408-1415.
84. Petes TD, Greenwell PW, Dominska M: **Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevisiae*.** *Genetics* 1997, **146:**491-498.
85. Michlewski G, Krzyzosiak WJ: **Molecular architecture of CAG repeats in human disease related transcripts.** *J Mol Biol* 2004, **340:**665-679.
86. Cleary JD, Pearson CE: **The contribution of cis-elements to disease-associated repeat instability: clinical and experimental evidence.** *Cytogenet Genome Res* 2003, **100:**25-55.
87. Cleary JD, Pearson CE: **Replication fork dynamics and dynamic mutations: the fork-shift model of repeat instability.** *Trends Genet* 2005, **21:**272-280.
88. Mulvihill DJ, Edamura KN, Hagerman KA, Pearson CE, Wang YH: **Effect of CAT or AGG interruptions and CpG methylation on nucleosome assembly upon trinucleotide repeats on spinocerebellar ataxia, type 1 and fragile X syndrome.** *J Biol Chem* 2005, **280:**4498-4503.
89. Rozanska M, Sobczak K, Jasinska A, Napierala M, Kaczynska D, Czerny A, Koziel M, Kozlowski P, Olejniczak M, Krzyzosiak WJ: **CAG and CTG repeat polymorphism in exons of human genes shows distinct features at the expandable loci.** *Hum Mutat* 2007, **28:**451-458.
90. Takano H, Cancel G, Ikeuchi T, Lorenzetti D, Mawad R, Stevanin G, Didierjean O, Durr A, Oyake M, Shimohata T, Sasaki R, Koide R, Igarashi S, Hayashi S, Takiyama Y, Nishizawa M, Tanaka H, Zoghbi H, Brice A, Tsuji S: **Close associations between prevalences of dominantly inherited spinocerebellar ataxias with CAG-repeat expansions and frequencies of large normal CAG alleles in Japanese and Caucasian populations.** *Am J Hum Genet* 1998, **63:**1060-1066.
91. Shimohata T, Nakajima T, Yamada M, Uchida C, Onodera O, Naruse S, Kimura T, Koide R, Nozaki K, Sano Y, Ishiguro H, Sakoe K, Ooshima T, Sato A, Ikeuchi T, Oyake M, Sato T, Aoyagi Y, Hozumi I, Nagatsu T, Takiyama Y, Nishizawa M, Goto J, Kanazawa I, Davidson I, Tanese N, Takahashi H, Tsuji S: **Expanded polyglutamine stretches interact with TAFII130, interfering with CREB-dependent transcription.** *Nat Genet* 2000, **26:**29-36.
92. Nucifora FC Jr., Sasaki M, Peters MF, Huang H, Cooper JK, Yamada M, Takahashi H, Tsuji S, Troncoso J, Dawson VL, Dawson TM, Ross CA: **Interference by huntingtin and atrophin-1 with cbp-mediated transcription leading to cellular toxicity.** *Science* 2001, **291:**2423-2428.
93. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucl Acids Res* 1999, **27:**573-580.
94. Karlin S: **Statistical significance of sequence patterns in proteins.** *Curr Opin Struct Biol* 1995, **5:**360-371.
95. Furey TS, Haussler D: **Integration of the cytogenetic map with the draft human genome sequence.** *Hum Mol Genet* 2003, **12:**1037-1044.

96. **HUGO Gene Nomenclature Committee** [http://www.gene.ucl.ac.uk/nomenclature]
97. **Coriell Cell Repository** [http://coriell.umdnj.edu]
98. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods Mol Biol* 2000, **132:**365-386.
99. Brownstein MJ, Carpten JD, Smith JR: **Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping.** *Biotechniques* 1996, **20:**1004-6, 1008-10.
100. Ahn SJ, Costa J, Emanuel JR: **PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR.** *Nucleic Acids Res* 1996, **24:**2623-2625.
101. Mood AM: **Introduction to the theory of statistics.** New York, McGraw-Hill; 1974:516-517.
102. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13:**2498-2504.