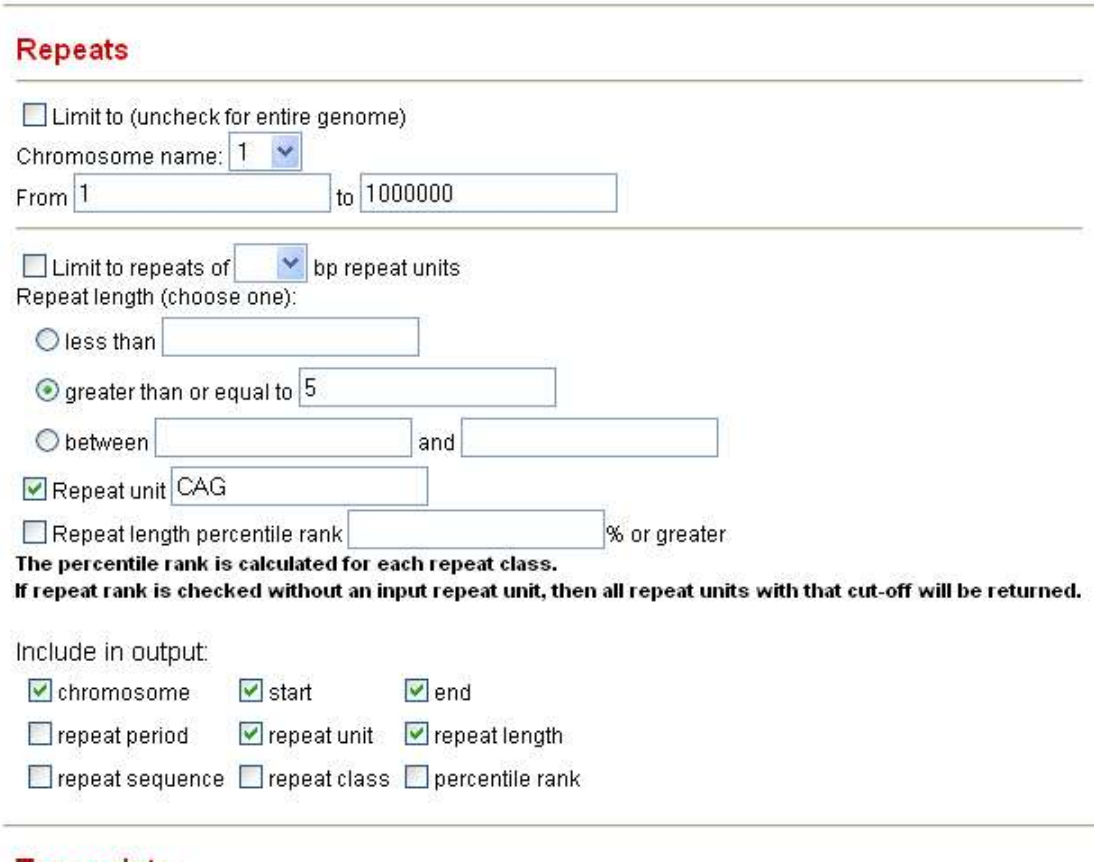
1.2 The `satellog` tutorial

The satellog tutorial provides step-by-step instructions, along with screenshots, about how to generate queries of interest to biologists. In each scenario we provide an example problem and show how to go about getting an answer from the web-based version of satellog. Power users desiring more complex queries should download satellog from Downloads and run a local instance.

1.2.1 Show me all the CAG repeats in the human genome that are repeated more than 5 times and are located within coding sequences.

First we specify the repeats in which we are interested with the repeats table.



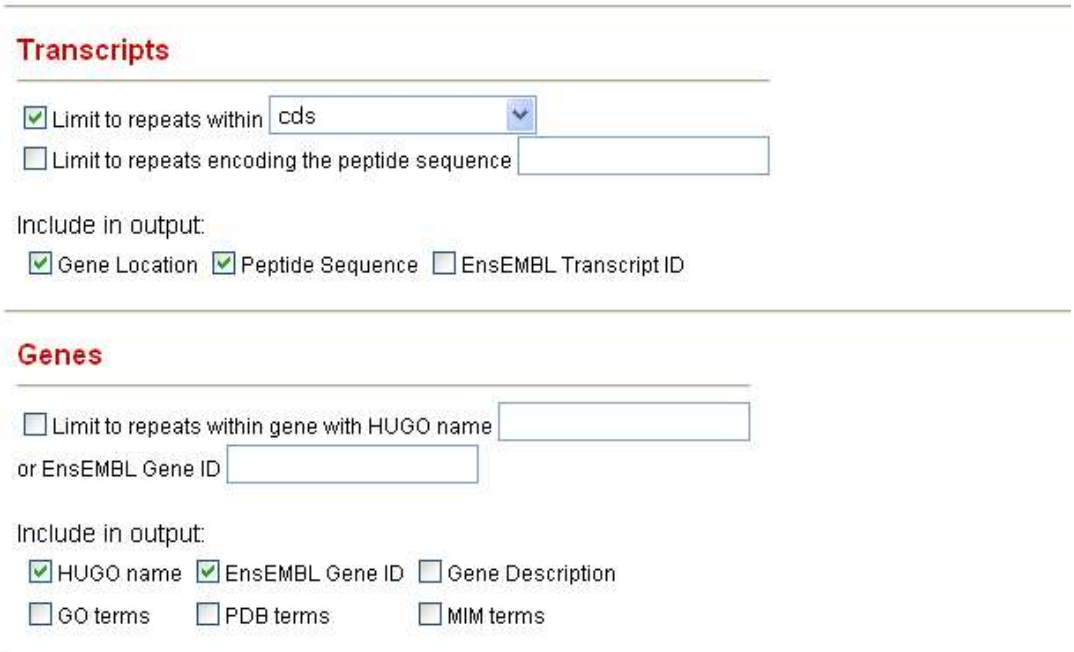
1) Uncheck the "Limit to" box because we are interested in the genome-wide distribution of CAG repeats.

2) Click the radio button "greater than or equal to" and input 5 because we are interested in all CAG repeats that are repeated 5 or more times.

3) Leave "Repeat Unit" on its default value, "CAG"

4) Select the Output variables of interest

Secondly, we specify other transcripts and gene information that is of interest to us about these CAG repeats.



5) Under transcripts, we check the "Limit to repeats within" box and select "cds" since we are only interested in coding CAG repeats.

6) Since we would like to see in the output both the location of the repeat within genes and the encoded peptide, we check "Gene Location" and "Peptide Sequence".

7) Usually a given repeat is more interesting if within certain genes.
    a. Under Genes, we can get information about which gene our repeat is in by checking selecting "HUGO name" which returns the HUGO (HUman Genome Organization) name of the gene each CAG repeat is in
    b. Next we check "EnsEMBL Gene ID" in case we wish to do more bioinformatics with this gene with the EnsEMBL Genome Browser (www.ensembl.org).

2) By default Satellog produces HTML output to screen. If you want another output format, or want the results e-mailed to you, select your preference prior to submitting your query.

Output Format

- (•) HTML
- ( ) Text, fixed width
- ( ) Text, comma separated
- ( ) Text, tab delimited

File Compression

- (•) None
- ( ) gzip (.gz)

File Info

Name for this dataset [ ]

Destination e-mail address for dataset [ ]

[Submit]

3) Click Submit.

Sample Satellog Output

This query gives all the CAG repeat co-ordinates that are within coding sequences, along with the peptide sequence they encode, the HUGO name of the gene they are within and their EnsEMBL Gene ID (output truncated for example).

Chromosome	Start	End	Repeat Unit	Repeat Length	Gene Location	Peptide Sequence	HUGO Name	EnsEMBL Gene ID
16	273195	273211	CTG	5	cds	LLLLLL	PDIP	ENSG00000185615
5	693387	693402	AGC	5	cds	EQQQQH	NULL	ENSG00000112877
4	134532390	134532405	CTG	5	cds	LCCCCC	PCDH10	ENSG00000138650
3	112111879	112111893	GCT	5	cds	PLLLLL	NULL	ENSG00000177707
11	617367	617383	AGC	5	cds	LLLLLL	SCT	ENSG00000070031
11	856188	856202	CTG	5	cds	LLLLL	TM4SF7	ENSG00000177769
6	1556462	1556477	CAG	5	cds	DSSSSS	FOXC1	ENSG00000054598
2	63823856	63823871	TGC	5	cds	KQQQQQ	NULL	ENSG00000119838
22	17512141	17512156	GCA	5	cds	ACCCCC	GSCL	ENSG00000063515
19	2202308	2202323	GCT	5	cds	PLLLLL	AMH	ENSG00000104899

You may be wondering why CTG, AGC, GCT, TGC, and GCA repeats came up when we asked for CAG repeats. There is a reason for this. A repeat can be represented in a number of ways in double-stranded DNA. Repeats are detected by their first tandemly repeated unit, therefore, CAGCAGCAG, AGCAGCAGC, and GCAGCAGCA are detected as repeats of CAG, AGC, and GCA respectively. Furthermore, the reference human genome sequence is only presented as the positive strand. Repeats of GTC, TCG, and CGT on the positive strand represent 5'->3 CAG, AGC and GCA repeats respectively on the negative strand. Therefore, to identify all CAG/CTG repeats in the human genome it's necessary to detect all CAG, AGC, GCA, GTC, TCG, and CGT repeats on the positive strand. To account for this, whenever we ask for a certain repeat type, all theoretical variations of the repeat unit are returned by Satellog.

1.2.2 Show me all the CAG repeats in the human genome that are repeated more than 5 times and are located within coding sequences and encode at least five glutamines.

Recently, expanding CAG repeats that encode glutamine tracts have been implicated in a number of neurodegenerative disorders. It is possible to select only those repeats that encode glutamine tracts with Satellog.

Repeats

☐ Limit to (uncheck for entire genome)
Chromosome name: 1
From 1 to 1000000

☐ Limit to repeats of bp repeat units
Repeat length (choose one):
○ less than
◉ greater than or equal to 5
○ between and
☑ Repeat unit CAG
☐ Repeat length percentile rank % or greater
The percentile rank is calculated for each repeat class.
If repeat rank is checked without an input repeat unit, then all repeat units with that cut-off will be returned.

Include in output:
☑ chromosome ☑ start ☑ end
☐ repeat period ☑ repeat unit ☑ repeat length
☐ repeat sequence ☐ repeat class ☐ percentile rank

Transcripts

☑ Limit to repeats within cds
☑ Limit to repeats encoding the peptide sequence QQQQQ

Include in output:
☑ Gene Location ☑ Peptide Sequence ☐ EnsEMBL Transcript ID

Genes

☐ Limit to repeats within gene with HUGO name
or EnsEMBL Gene ID

Include in output:
☑ HUGO name ☑ EnsEMBL Gene ID ☐ Gene Description
☐ GO terms ☐ PDB terms ☐ MIM terms

1) Repeat the steps in 1.2.1.

2) However, under Transcripts also select “Limit to repeats within” and input QQQQQ. This will return only the subset of repeats that encode five or more glutamines. Of course, this query can be run with any other plausible repeat unit and peptide combination.

3) Click Submit.

Sample Satellog Output

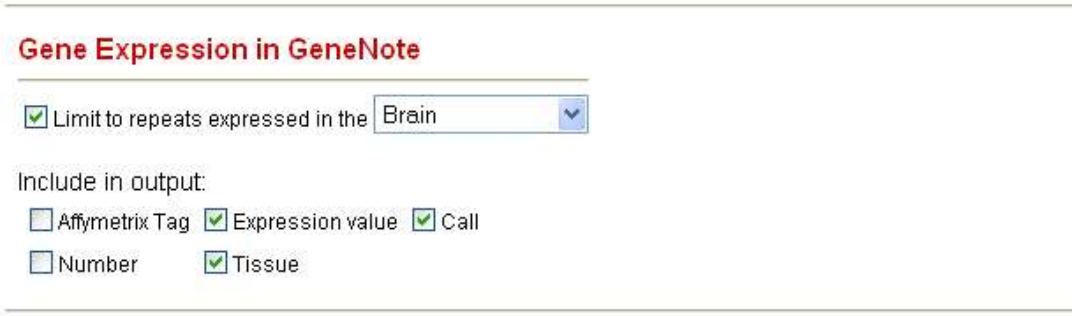
This query gives all the CAG repeat co-ordinates that are within coding sequences, along with the peptide sequence they encode, the HUGO name of the gene they are within and their EnsEMBL Gene ID (output truncated for example). However, in this example, only those repeats encoding at least five glutamines are output.

Chromosome	Start	End	Repeat Unit	Repeat Length	Gene Location	Peptide Sequence	HUGO Name	EnsEMBL Gene ID
2	63823856	63823871	TGC	5	cds	KQQQQQ	NULL	ENSG00000119838
22	17747690	17747706	GCT	5	cds	RQQQQQL	HIRA	ENSG00000100084
16	3778725	3778739	TGC	5	cds	LQQQQQ	CREBBP	ENSG00000005339
3	152433507	152433521	GCA	5	cds	LQQQQQ	NULL	ENSG00000144893
3	152469023	152469037	CAG	5	cds	QQQQQ	NULL	ENSG00000144893
22	19243319	19243333	GCA	5	cds	LQQQQQ	PCQAP	ENSG00000099917
22	19243347	19243361	CAG	5	cds	QQQQQ	PCQAP	ENSG00000099917
22	19243451	19243467	GCA	5	cds	VQQQQQL	PCQAP	ENSG00000099917
17	4991604	4991620	GCA	5	cds	LQQQQQR	NULL	ENSG00000141503
17	4994558	4994573	CAG	5	cds	QQQQQL	NULL	ENSG00000141503

1.2.3 Show me all the CAG repeats in the human genome that are repeated more than 5 times and are located within coding sequences, encode at least five glutamines and are expressed in the Brain.

Often repeats are only of interest if they are within genes that are expressed in certain tissues. All genes in Satellog are cross-referenced to the GeneNote (Gene Normal Tissue Expression) database. The GeneNote database is a collection of AffyMetrix microarray experiments run with twelve normal human tissues. With GeneNote, it is possible to see if repeats are associated with a gene that is expressed in a tissue of interest.

1) Repeat the steps in 1.2.2



2) Under Gene Expression in GeneNote, check "Limit to repeats expressed in the" and select "Brain" from the drop-down box.

3) Next select the following variables:
   a. "Expression Value" – refers to the absolute intensity value on the chip, can give an indication of significant tag expression
   b. "Call" – reported by the AffyMetrix analysis software, P, M, and A refer to Present, Marginal or Absent tag expression respectively
   c. "Tissue " – refers to one of the twelve human tissues

2) "AffyMetrix Tag" and "Number" refer to the AffyMetrix Tag ID and microarray replicate number, select these if you are interested in the details

3) Click Submit.

Sample Satellog Output

This query gives all the CAG repeat co-ordinates that are within coding sequences, along with the peptide sequence they encode, the HUGO name of the gene they are within and their EnsEMBL Gene ID (output truncated for example). In this example, only those repeats encoding at least five glutamines are output and within genes expressed in the Brain are shown. Some repeats are shown more than once because their gene maps to more than one AffyMetrix tag (tag names not shown).

Chromosome	Start	End	Repeat Unit	Repeat Length	Gene Location	Peptide Sequence	HUGO Name	EnsEMBL Gene ID	Expression Value	Call	Tissue
22	17747690	17747706	GCT	5	cds	RQQQQQL	HIRA	ENSG00000100084	564.0	P	Brain
22	17747690	17747706	GCT	5	cds	RQQQQQL	HIRA	ENSG00000100084	279.5	P	Brain
16	3778725	3778739	TGC	5	cds	LQQQQQ	CREBBP	ENSG00000005339	120.4	P	Brain
16	3778725	3778739	TGC	5	cds	LQQQQQ	CREBBP	ENSG00000005339	64.5	P	Brain
22	19243451	19243467	GCA	5	cds	VQQQQQL	PCQAP	ENSG00000099917	408.6	P	Brain
22	19243451	19243467	GCA	5	cds	VQQQQQL	PCQAP	ENSG00000099917	250.2	P	Brain
17	4991604	4991620	GCA	5	cds	LQQQQQR	NULL	ENSG00000141503	1158.2	P	Brain
17	4991604	4991620	GCA	5	cds	LQQQQQR	NULL	ENSG00000141503	571.1	P	Brain
17	4994558	4994573	CAG	5	cds	QQQQQL	NULL	ENSG00000141503	1158.2	P	Brain
17	4994558	4994573	CAG	5	cds	QQQQQL	NULL	ENSG00000141503	571.1	P	Brain
1	58618041	58618056	GCT	5	cds	QQQQQP	JUN	ENSG00000177606	193.0	P	Brain
1	58618041	58618056	GCT	5	cds	QQQQQP	JUN	ENSG00000177606	11.0	A	Brain

1.2.4 Show me all the repeats of type GAA in the *Frataxin* gene.

GAA repeats in the *Frataxin* gene have been associated with Friedreich's Ataxia. Using Satellog, we can identify all of the GAA repeats in this gene and find the one associated with disease. Given any gene, Satellog can return either all the repeats, or certain repeats of a given size, repeat class, or within a specified genetic region.

Repeats

☐ Limit to (uncheck for entire genome)
Chromosome name: 1
From 1 to 1000000

☐ Limit to repeats of bp repeat units
Repeat length (choose one):
○ less than
○ greater than or equal to
○ between and
☑ Repeat unit GAA
☐ Repeat length percentile rank % or greater
The percentile rank is calculated for each repeat class.
If repeat rank is checked without an input repeat unit, then all repeat units with that cut-off will be returned.

Include in output:
☑ chromosome ☑ start ☑ end
☐ repeat period ☑ repeat unit ☑ repeat length
☐ repeat sequence ☐ repeat class ☐ percentile rank

Transcripts

☐ Limit to repeats within
☐ Limit to repeats encoding the peptide sequence

Include in output:
☑ Gene Location ☑ Peptide Sequence ☐ EnsEMBL Transcript ID

Genes

☑ Limit to repeats within gene with HUGO name
or EnsEMBL Gene ID ENSG00000165060

Include in output:
☑ HUGO name ☑ EnsEMBL Gene ID ☐ Gene Description
☐ GO terms ☐ PDB terms ☐ MIM terms

1) Uncheck the “Limit to” box because we are interested in the genome-wide distribution of GAA repeats.

2) Change “Repeat Unit” to “GAA”

3) Select the Repeat Output variables of interest

4) Under transcripts, check “Gene Location” and “Peptide Sequence”.

5) Also under transcripts, check “Limit to repeats within gene with...” and input the EnsEMBL Gene ID for Frataxin (ENSG00000165060). It is always safer to use the EnsEMBL Gene ID because it’s stable whereas HUGO gene names can change over time.

6) Under Genes, select “HUGO Name” and “EnsEMBL Gene ID”.

7) Click Submit.

Sample Satellog Output

This query gives all the GAA repeat co-ordinates that are within the *Frataxin* gene, along with the gene location, peptide sequence they encode, HUGO name of the gene they are within and their EnsEMBL Gene ID. In this example, chr9:67109320-67109339 is the disease-associated repeat. Note: all disease associated repeats detected by our group in Satellog are available in the Downloads section.

Chromosome	Start	End	Repeat Unit	Repeat Length	Gene Location	Peptide Sequence	HUGO Name	EnsEMBL Gene ID
9	67109320	67109339	AAG	6	intron		FRDA	ENSG00000165060
9	67127519	67127533	CTT	5	intron		FRDA	ENSG00000165060
9	67141106	67141115	GAA	3	intron		FRDA	ENSG00000165060
9	67159952	67159968	CTT	5	15000		FRDA	ENSG00000165060
9	67159982	67159992	CTT	3	15000		FRDA	ENSG00000165060

1.2.5 Show me the largest 1% of TCCCTC repeats in the genome.

Researchers are usually interested in the largest repeats of a given repeat class because these are usually the substrates for subsequent expansion. However, when interested in a class there is no way to know what the largest repeat sizes are. For instance, TCCCTC repeats range from being repeated twice to 54 times. It is possible to eliminate guesswork and select the top X% largest repeats of any class easily with Satellog.

Repeats

☐ Limit to (uncheck for entire genome)
Chromosome name: 1
From 1 to 1000000

☐ Limit to repeats of bp repeat units
Repeat length (choose one):
○ less than
○ greater than or equal to
○ between and
☑ Repeat unit TCCCTC
☑ Repeat length percentile rank 1 % or greater
The percentile rank is calculated for each repeat class.
If repeat rank is checked without an input repeat unit, then all repeat units with that cut-off will be returned.

Include in output:
☑ chromosome ☑ start ☑ end
☐ repeat period ☑ repeat unit ☑ repeat length
☐ repeat sequence ☐ repeat class ☑ percentile rank

Transcripts

☐ Limit to repeats within
☐ Limit to repeats encoding the peptide sequence

Include in output:
☑ Gene Location ☑ Peptide Sequence ☐ EnsEMBL Transcript ID

Genes

☐ Limit to repeats within gene with HUGO name
or EnsEMBL Gene ID

Include in output:
☑ HUGO name ☑ EnsEMBL Gene ID ☐ Gene Description
☐ GO terms ☐ PDB terms ☐ MIM terms

1) Uncheck the “Limit to” box because we are interested in the genome-wide distribution of TCCCTC repeats.

2) Change “Repeat Unit” to “TCCCTC”

3) Select the Repeat Output variables of interest and ensure to check off “Repeat length percentile rank” and input “1” and “Percentile Rank” for the output.

4) Under transcripts, check “Gene Location” and “Peptide Sequence”.

5) Under Genes, select “HUGO Name” and “EnsEMBL Gene ID”.

6) Click Submit.

Sample Satellog Output

This query gives all the TCCCTC repeat co-ordinates, along with their percentile rank, gene location, peptide sequence, HUGO name of the gene they are within and their EnsEMBL Gene ID. The percentile rank refers to the fraction of TCCCTC repeats that are as large as or larger than the length of each output repeat. For example, the first repeat has a repeat length that is as large as or larger than 0.7545% of all TCCCTC repeats in the human genome.

Chromosome	Start	End	Repeat Unit	Repeat Length	Percentile Rank	Gene Location	Peptide Sequence	HUGO Name	EnsEMBL Gene ID
1	245774762	245774804	CCCTCT	7	0.007545	intron		NULL	ENSG00000177151
5	34353361	34353403	GAGAGG	7	0.007545	15000		NULL	ENSG00000184421
18	503034	503094	GAGGGA	10	0.002101	15000		COLEC12	ENSG00000158270
6	34187980	34188025	TCTCCC	7	0.007545	45000		GRM4	ENSG00000124493
14	18963549	18963615	CTCCCT	11	0.001670	15000		NULL	ENSG00000182545
1	112133833	112133927	CTCTCC	15	0.000970	15000		NULL	ENSG00000186264
17	1800553	1800597	GAGGGA	7	0.007545	15000		PRPF8	ENSG00000174231
20	2798647	2798730	TCTCCC	14	0.001077	15000		C20orf141	ENSG00000101386
20	2798647	2798730	TCTCCC	14	0.001077	15000		NULL	ENSG00000171964
20	2798647	2798730	TCTCCC	14	0.001077	15000		C20orf81	ENSG00000132635

1.2.6 Show me all the polymorphic CAC repeats (as detected in UniGene clusters).

Researchers are usually interested in repeats that already have evidence of length polymorphism and are near or within candidate genes. With Satellog, it is possible to identify candidate polymorphic sites in the human genome that are not documented elsewhere. Every transcribed repeat in Satellog has been analyzed within UniGene clusters to see if there is any evidence of repeat polymorphism. It is possible to extract just those repeats with possible polymorphism from any repeat class. Let us hypothesize that CAC repeats are an important new repeat class implicated in disease etiology and that we are interested in any potential polymorphic sites.

Repeats

☐ Limit to (uncheck for entire genome)

Chromosome name: 1

From 1 to 1000000

☐ Limit to repeats of bp repeat units

Repeat length (choose one):

○ less than

○ greater than or equal to

○ between and

☑ Repeat unit CAC

☐ Repeat length percentile rank 1 % or greater

The percentile rank is calculated for each repeat class.
If repeat rank is checked without an input repeat unit, then all repeat units with that cut-off will be returned.

Include in output:

☑ chromosome ☑ start ☑ end

☐ repeat period ☑ repeat unit ☑ repeat length

☐ repeat sequence ☐ repeat class ☐ percentile rank

Transcripts

☐ Limit to repeats within

☐ Limit to repeats encoding the peptide sequence

Include in output:

☑ Gene Location ☑ Peptide Sequence ☐ EnsEMBL Transcript ID

Genes

☐ Limit to repeats within gene with HUGO name

or EnsEMBL Gene ID

Include in output:

☑ HUGO name ☑ EnsEMBL Gene ID ☐ Gene Description

☐ GO terms ☐ PDB terms ☐ MIM terms

1) Uncheck the “Limit to” box because we are interested in the genome-wide distribution of CAC repeats.

2) Change “Repeat Unit” to “CAC”

3) Under transcripts, check “Gene Location” and “Peptide Sequence”.

4) Under Genes, select “HUGO Name” and “EnsEMBL Gene ID”.

Polymorphism within UniGene clusters

☑ Limit to genes with evidence of polymorphism

Include in output:

☑ Number of Hits ☑ Minimum Length ☑ Maximum Length

☑ Mean Length ☑ Standard Deviation

Include in output:

☐ UniGene cluster ☐ UniGene sequence ☐ Length within UniGene Sequence

5) Under "Polymorphism within UniGene clusters", check "Limit to ...". This will limit the output to repeats with one or more repeat length polymorphism.

6) Include the following summary statistics in the output:
   a. "Number of Hits" – refers to the total number of times the repeat has been detected in UniGene sequences
   b. "Minimum Length" – refers to the minimum repeat length detected in any of the hits.
   c. "Maximum Length" – refers to the maximum repeat length detected in any of the hits.
   d. "Mean Length" - refers to the mean length of all detected hits
   e. "Standard Deviation" – describes the standard deviation of all hits detected. Repeats with a larger standard deviation have more extreme polymorphism
   f. **Note:** "UniGene cluster", "UniGene sequence" and "Length within UniGene Sequence" provide the information about each UniGene hit and are not included in this output. For practical purposes, these are only useful if one is interested in double-checking the hits reported by Satellog.

2) Click Submit.

Sample Satellog Output

This query gives all the CAC repeat co-ordinates, along with their gene location, peptide sequence, HUGO name of the gene they are within and their EnsEMBL Gene ID. Summary statistics about their UniGene hits are also provided to give a complete picture of their polymorphism. For example, chr17:7950744-7950782 has a standard deviation of 2.12 but only has 2 hits.

Chromosome	Start	End	Repeat Unit	Repeat Length	Gene Location	Peptide Sequence	HUGO Name	EnsEMBL Gene ID	Number of Hits	Minimum Length	Maximum Length	Mean Length	Standard Deviation
3	181973881	181973892	ACC	4	cds	LPPPP	FXR1	ENSG00000114416	82	3	4	3.99	0.11
4	833274	833283	CCA	3	3utr		GAK	ENSG00000178950	45	2	3	2.89	0.32
1	53300375	53300385	CCA	3	cds	PPPP	DMRTB1	ENSG00000143006	5	2	3	2.80	0.45
20	1913627	1913657	CCA	10	3utr		PTPNS1	ENSG00000088835	15	9	12	10.07	0.59
9	4783006	4783017	CCA	4	5utr		NULL	ENSG00000120158	51	3	4	3.98	0.14
14	22900124	22900133	CAC	3	cds	QVVV	NULL	ENSG00000100445	27	2	3	2.96	0.19
1	199214717	199214726	TGG	3	cds	IGGG	TIMM17A	ENSG00000134375	332	2	3	2.90	0.30
17	7328580	7328589	TGG	3	cds	MVVV	ACADVL	ENSG00000072778	120	2	3	2.99	0.09
19	6086729	6086740	CCA	4	5utr		NULL	ENSG00000130377	16	3	4	3.94	0.25
6	43635710	43635721	CCA	4	3utr		GTPBP2	ENSG00000172432	15	3	4	3.93	0.26
6	43635710	43635721	CCA	4	intron		GTPBP2	ENSG00000172432	15	3	4	3.93	0.26
17	7683885	7683896	CCA	4	cds	PTTTS	CD68	ENSG00000129226	122	3	4	3.99	0.09