Statistical Tests for Gene Conversion

Stanley Sawyer

Gene conversion is any process that causes a segment of DNA to be copied onto another DNA segment, or else appears to act in this way.

Some possible biological causes are:

- Homologous recombination
- Mosaicism in viruses

Gene conversion can involve unexpressed (or junk) DNA only, or a segment containing several genes.

Gene conversion is an important cause of the spread of

- pathogenicity
- antibiotic resistance
- vaccine resistance

in bacteria and viruses.

How can we detect gene conversion from data, specifically from an aligned set of DNA or protein sequences?

An example alignment:

```
          10        20        30        40
K  GCAG AGTGCTATAACAAGAACG GTA CCGGT GTATCTA ATGT
2  GCAG AGGGCTTTAACTTCTACG TTA GGGGT GTTTCTA ATGT
3  ACAC AACGCTGTAACTAGAAAG TTA GGGGT GTGTCTG GCGT
4  GCAC AGGGCTTTAACTTCTACG GTA CCGGT GTTTCTG GCGT
5  ACAC ACAGCTATAACATCTAAG GTA CCGGT GTATCTG GCGT
6  GCAC ATGGCGGTAACAAGTTCG GTA TTGGAATCTCTT ACAT
```

The pair of boxes is an *inner fragment*, suggesting an event in the ancestry of the 6 sequences. This amounts to a *run* of 18 matched sites in an alignment of length 41. The single box is an *outer fragment*, suggesting an exterior event.

We discard sites that are *monomorphic* in the entire alignment:

```
        10        20        30        40
K  GG GTTAAAGAAC G CCTGAA ATG
2  GG GGTTTTCTAC T GGTGTA ATG
3  AC ACTGTAGAAA T GGTGGG GCG
4  GC GGTTTTCTAC G CCTGTG GCG
5  AC CATAATCTAA G CCTGAG GCG
6  GC TGGGAAGTTC G TTAACT ACA
```

```
          10        20        30        40
K   GG GTTAAAGAAC G CCTGAA ATG
2   GG GGTTTTCTAC T GGTGTA ATG
3   AC ACTGTAGAAA T GGTGGG GCG
4   GC GGTTTTCTAC G CCTGTG GCG
5   AC CATAATCTAA G CCTGAG GCG
6   GC TGGGAAGTTC G TTAACT ACA
```

We now have a run of length 10 in an alignment of length 22. We have two ways of calculating the statistical significance of this run, a *runs test* and a *permutation test*.

For the runs test for the paired boxes, we note there are 7 differences out of 22 between sequences 2 and 4, so that the probability of a run of length 10 or longer starting at offset 3 is $p(1-p)^{10} = 0.00691$ for p=7/22. However, there are 13 possible starting points for runs of length 10 or more, leading to a "multiple-test" corrected $P = 0.1046$.

The permutation test carries out 10,000 random permutations of the 22 columns in the last alignment, and asks what proportion of these permutations has at least *one run* of length 10 or longer between sequences 2 and 4 starting at any offset. This gives a probability of $P = 0.0399$. Other runs with different starting random-number seeds gave $P = 0.0382$ and $P = 0.395$.

A second example alignment:

```
                10                  20                  30                  40
K   GCAG AGTGCT ATAACAAGAA CGGTACCGGTGT  ATCTAATGG
2   GCAG AGGGCTTTAACTTGAACGGTAGCGGTGT  CTCTAATGG
3   ACAC AACGCT GTAACTAGAA AGTTAGGGGTGT GTCTGGCGG
4   GCGC AGGGCTTTAACTTGAGCGGTAGCGGTGT  TTAAGGCAT
5   AACC ACAGCT ATAACATCTA AGTTACCAATGT ATCTGGCGG
6   GCAC ATGGCGGTAACAAGTGCGGTATTGGAAT  CTCTTACAG
```

The two boxes span offsets 5–32 inclusively, but there is a single mismatch at offset 20. Neither of the fragments on either side is significant by either the runs test ($P = 0.196$) nor the permutation test ($P = 0.120$). The fact that they are adjacent except for the mismatch suggests that an old and highly significant gene conversion event was punctuated by a later mutation.

ANOTHER CONCERN: Since there are 6 sequences, the number of sequence pairs is $6(6-1)/2 = 15$. A test procedure with a false positive rate of 5% will make at least one mistake 3/4 of the time if carries out 15 tests. A standard multiple-test correction is to multiply all P-values by the number of tests. If there were 60 aligned sequences, this would involve multiplying P-values by 1770. However, the permutation test is automatically multiple-test aware for the length of the alignment, so that there may be better ways.

Questions:

- How can later mutations be handled? Can one assign mismatch penalties for different pairs of sequences and calculate P-values for high-scoring (rather than merely long) fragments?
- Can lengths or scores be scaled across different sequence pairs to give reasonable multiple-comparison-corrected P-values? At the same time, can mismatch penalties be coordinated in a reasonable way? (NOTE: Fragments between closely-related pairs of sequences in an alignment will have fewer differences and longer inner fragments, while a shorter fragment between a more distantly-related sequence pair may have highly-significant shorter fragments.)
- Can we avoid permutations? These can be VERY time consuming with a large number of sequences and polymorphisms. Is there a simple, easily applied approximation for pairwise P-values, with or without mismatch penalties?

For a given pair of sequences, assume we can score all sequence matches at aligned sites as 1 and mismatches as $-m$ and score each inner fragment as the sum of these match or mismatch scores. (We allow $m = \infty$, which reduces to the case of runs with scores as fragment lengths.)

How can we find the highest-scoring fragments efficiently?

More importantly, how can we find probabilities or approximate P-values for high-scoring fragments?

Fortunately, a similar problem has come up before in queueing theory:

A problem in queueing theory:

Suppose that customers arrive at a bank on the average of n_c customers per week. The single teller is very busy, but whenever he/she gets near the teller's window, then he/she serves m customers at once. The teller wanders near the teller's window around d times per week, on the average. We assume $d < n_c$ but $md > n_c$. (If $md < n_c$, the line at the teller's window will increase without limit.)

Given n_c, d, and m, what is the distribution of the maximum queue length over a year? What is the probability that the line will ever stretch out into the street?

Each new customer increases the line by one, and each act by the teller reduces it by m. When the line length becomes zero or negative, the process starts over. There are approximately $n = 52(n_c + d)$ events per year, where an event is an arrival (customer or teller). Of the events, $p = d/(n_c + d)$ correspond to mismatches and the line length drops by m. Thus the distribution of the longest queue length over a year is exactly the same as the distribution of the maximum fragment score for an alignment with $n = 52(n_c + d)$ polymorphic sites.

A Theorem from Queueing Theory:

Assume that we have a pair of sequences with

- n polymorphic sites (in the alignment)
- the two sequences differ at $d \leq n$ sites
- score 1 for a match, $-m$ for a mismatch
- $md > n - d$ (the sequence itself has a negative score).

Consider random sequences of length n such that sites are different with probability $p = d/n$ and the sites are independent. Let Score_i be the largest score for any subsegment of the first i sites. Then, there exist constants $\lambda = \lambda(p, m)$ and $K = K(p, m)$ such that

$$\lim_{n\to\infty} \Pr\left(\max_{1\leq i\leq n} \text{Score}_i \; - \; \frac{\log(nK)}{\lambda} \geq x\right) \qquad (1)$$

$$= \; 1 - \exp(-\exp(-\lambda x)) \qquad (2)$$

$$\approx \; \exp(-\lambda x) \qquad \text{if } \lambda x \gg 1$$

This holds if the mismatch penalty m is not an integer. If m is an integer, then the ratio of (1) and (2) oscillates between positive limits defined by $K_- < K_+$. If $K = K_+$, then (2) is conservative for large n.

Extreme-value-distribution approximations:

In particular, this says that

$$\max_{1 \le i \le n} \text{Score}_i \;\approx\; A \log n + B$$

where $A = 1/\lambda$ and $B = \log K/\lambda$.

This result is due to Iglehart (1972) and Karlin, Dembo, and Kawabata (1990), extending earlier work of Spitzer and others. Karlin and Altschul (1990) and Altschul et al (1990) used Iglehart's results for computationally efficient pattern-matching results for DNA and protein sequences, which became the widely-used BLAST scores for protein matches.

It has been said that someone sends a DNA or protein sequence to the National Library of Medicine server for a BLAST search of stored databases on the average of once every 3 seconds, so that Iglehart's result may be one of the most widely-used theorems in probability theory.

Karlin and Dembo (1992) have a nice mathematical treatment of these approximate P-values, and also extend the results to Markov-dependent scores.

Notes:

• This assumes that all matches should have the same positive score $= 1$, and that all mismatches should have the same negative score $= -m$. In protein search algorithms (a protein is a chain of amino acids), mismatches of amino acids with similar chemical properties are given a small positive score, and mismatches for two amino acids that have very different chemical properties are given higher penalties.

• The literature (Karlin and Altschul, 1990; Karlin, Dembo, and Kawabata, 1990) suggests that scores with a given mismatch penalty m will be most powerful for detecting fragments with a mismatch density that depends on m, n, and d.

This means that scores with different mismatch penalties will be more powerful for detecting gene conversion events of different ages. This is unfortunate, since it means that the same score is not optimal for detecting all gene conversion events.

An Almost-Proof of the Formula:
The formula is:

$$\lim_{n\to\infty} \Pr\left(\max_{1\le i\le n} \text{Score}_i \;-\; \frac{\log(nK)}{\lambda} \ge x\right) \qquad (1)$$

$$= \; 1 - \exp(-\exp(-\lambda x)) \qquad (2)$$

Since $mp > 1-p$ by assumption, we can view MaxScore as the maximum of $n_e \approx nC$ excursions of the running score into positive values.

These excursions will be approximately exponentially distributed with some mean μ. Let $\mathcal{L} = \max_{1\le i\le n_e} X_i$ for n_e independent exponentially distributed random variables X_i with $E(X_i) = \mu$, so that $\mathcal{L} \approx \max_{1\le i\le n} \text{Score}_i$. Then

$$\Pr(\mathcal{L} \le t) \;\approx\; \Pr(X_i \le t)^{n_e} \;\approx\; \big(1 - \exp(-t/\mu)\big)^{n_e}$$

Set $t = t_n = \mu \log n_e + x = \mu \log(nC) + x$. Then

$$\lim_{n\to\infty} \Pr(\mathcal{L} \le t_n) \;\approx\; \left(1 - \left(\frac{1}{nC}\right)\exp(-x/\mu)\right)^{nC}$$

$$\approx\; \exp\big(-\exp(-x/\mu)\big)$$

This implies (2) with $\lambda = 1/\mu$ and $K = C$.

Mismatch penalties:

For a particular pair of sequences, the *score* of a fragment with possible mismatches is the number of matches at polymorphic sites minus a penalty times the number of mismatches. A significant fragment is a fragment with a significantly high score.

For different sequence pairs, mismatch penalties should be inversely proportional to the number of sequence differences. That is, the higher the proportion of sequence differences, the lower the mismatch penalty.

To be specific, we will scale

$$m \;=\; \text{int}\left(\frac{\text{gscale} * n + d - 1}{d}\right)$$

(so that $m \approx$ gscale$*n/d$ and $md > n-d$ if gscale ≥ 1) where

- n is the number of sites that are polymorphic in the alignment and the two sequences differ at $d \leq n$ sites,
- "gscale" $= 1, 2, 3, \ldots$ is a parameter.
- "gscale" $= 1$ means the lowest penalties. "gscale" $= \infty$ means that mismatches are prohibited.

This allows us to define mismatch penalties consistently across an alignment.

Remarks:

• The queueing-theory approximation applies to pairwise P-values rather than global or multiple-test aware P-values. The obvious generalizations for global P-values can be very conservative in some cases.

• In comparing or ranking significant fragments across different sequence pairs, the lengths or raw scores of fragments should not be compared directly. The longest fragments will be between the most closely related sequences. This can miss highly significant (but shorter) fragments between pairs of sequences with a denser set of sequence differences.

• Define $\text{Kscore} = \lambda * \text{MaxScore} - \log(nK)$
Then the approximation is

$$\begin{aligned} \Pr(\text{MaxScore} \geq x) &\approx 1 - \exp(-\exp(-\text{Kscore})) \\ &\approx \exp(-\text{Kscore}) \qquad \text{if Kscore} \gg 1 \end{aligned}$$

The expression Kscore can be used to define a global score for permutation tests that is not biased against sequence pairs with larger numbers of differences.

BLAST-like Test Results (Gscale = 1)

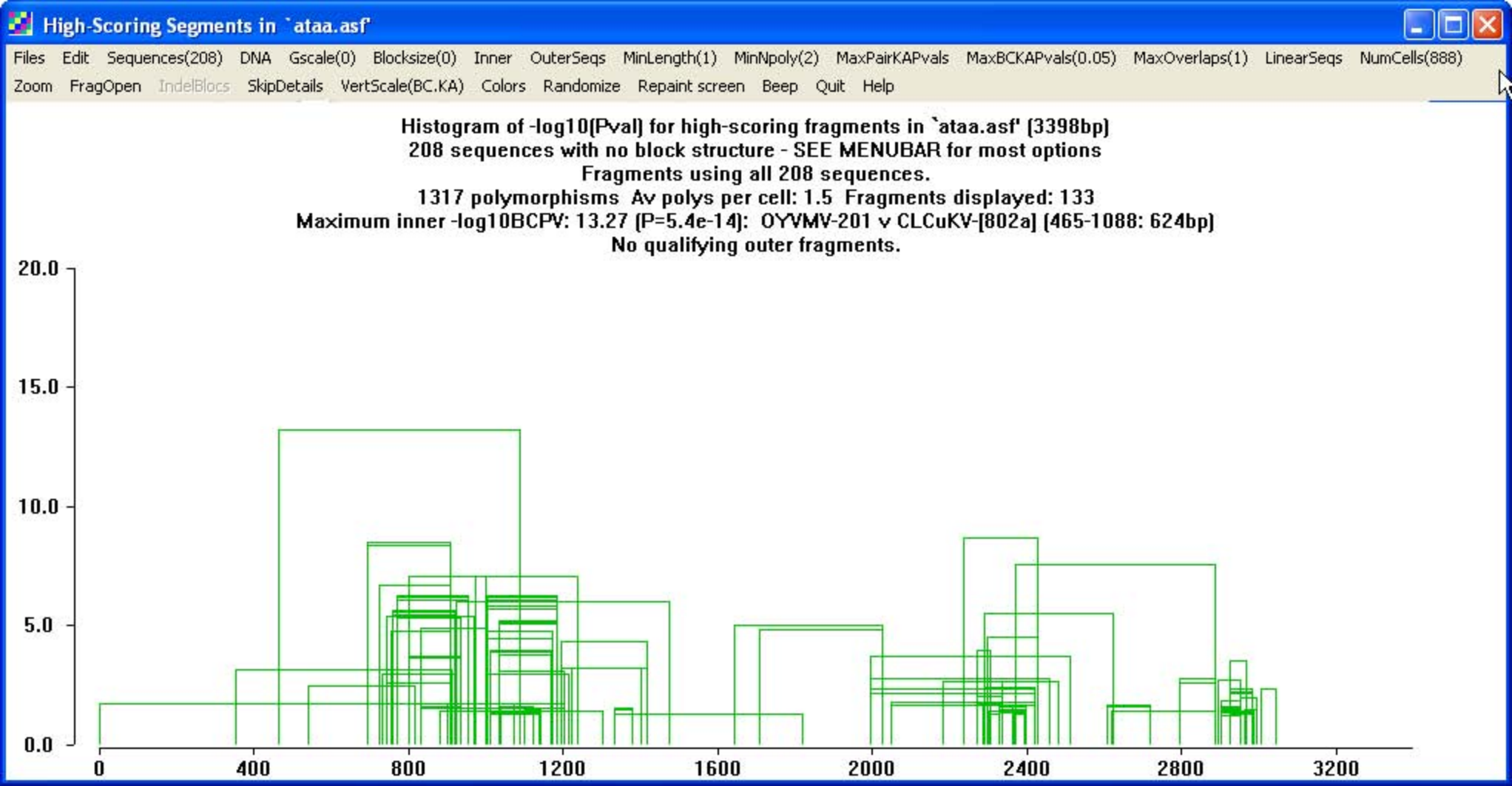
Larger alignments can be handled using the approximate P-values, and in addition P-values smaller than 1/10,000 can be estimated:

Data set (Gscale $= \infty$)	Seqs.	Bases	Number of Polys	Frags[1]	P-val.[2]
Gemini virus	64	3309	1468	61	$< 10^{-14}$
Cauliflower mosaic virus	7	8110	856	4	$< 10^{-4}$
Human γ globins	3	1763	65	2	$< 10^{-6}$
Maize actin gene	8	1008	364	2	$< 10^{-4}$

Data set (Gscale = 1)	Seqs.	Bases	Number of Polys	Frags[1]	P-val.[2]
Gemini virus	64	3309	1468	427	$< 10^{-19}$
Cauliflower mosaic virus	7	8110	856	4	$< 10^{-5}$
Human γ globins	3	1763	65	2	$< 10^{-6}$
Maize actin gene	8	1008	364	1	$< 10^{-4}$

1 – Number of inner global $P < 0.05$ (in terms of Kscore)
Overlapping fragments are ignored.

2 – Smallest global inner P-value (BLAST-like P-value)

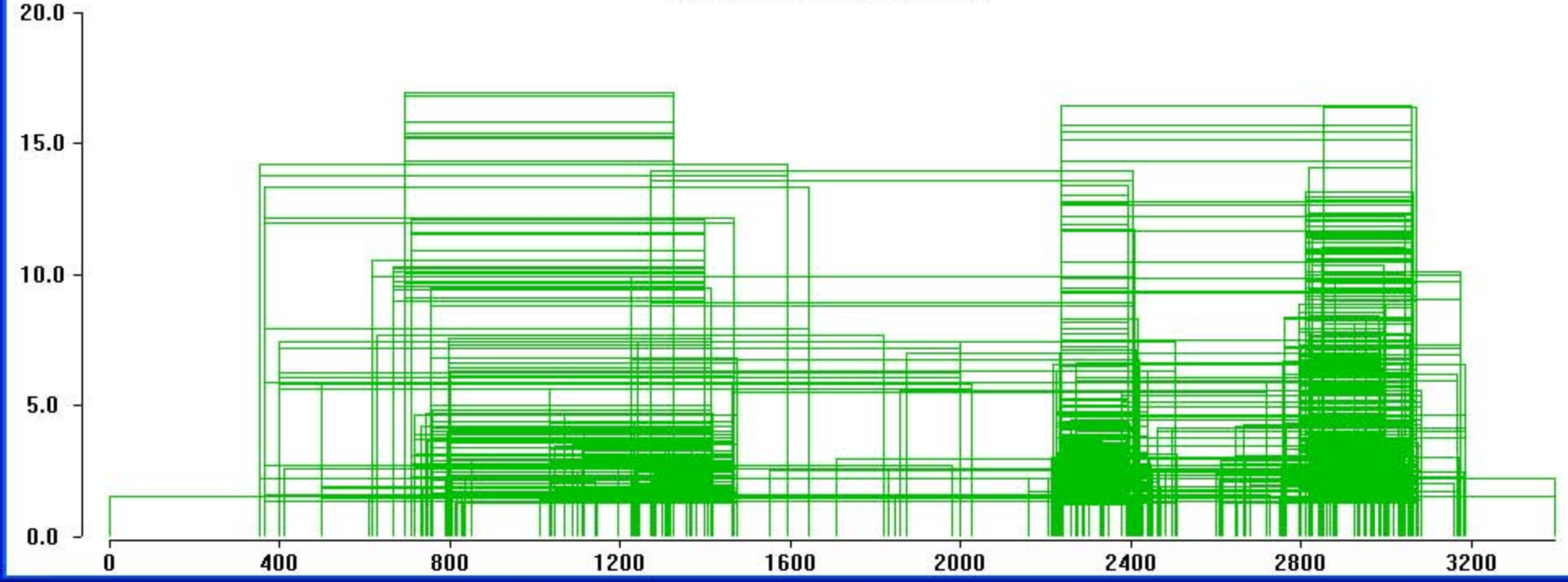

High-Scoring Segments in `ataa.asf'
Files Edit Sequences(208) DNA Gscale(0) Blocksize(0) Inner OuterSeqs MinLength(1) MinNpoly(2) MaxPairKAPvals MaxBCKAPvals(0.05) MaxOverlaps(1) LinearSeqs NumCells(888)
Zoom FragOpen IndelBlocs SkipDetails VertScale(BC.KA) Colors Randomize Repaint screen Beep Quit Help
Histogram of -log10(Pval) for high-scoring fragments in `ataa.asf' (3398bp)
208 sequences with no block structure - SEE MENUBAR for most options
Fragments using all 208 sequences.
1317 polymorphisms Av polys per cell: 1.5 Fragments displayed: 133
Maximum inner -log10BCPV: 13.27 (P=5.4e-14): OYVMV-201 v CLCuKV-[802a] (465-1088: 624bp)
No qualifying outer fragments.
20.0
15.0
10.0
5.0
0.0
0
400
800
1200
1600
2000
2400
2800
3200

High-Scoring Segments in `ataa.asf'

Files Edit Sequences(208) DNA Gscale(1) Blocksize(0) Inner OuterSeqs MinLength(1) MinNpoly(2) MaxPairKAPvals MaxBCKAPvals(0.05) MaxOverlaps(1) LinearSeqs NumCells(888)

Zoom FragOpen IndelBlocs SkipDetails VertScale(BC.KA) Colors Randomize Repaint screen Beep Quit Help

Histogram of -log10(Pval) for high-scoring fragments in `ataa.asf' (3398bp)
208 sequences with no block structure - SEE MENUBAR for most options
Fragments using all 208 sequences.
1317 polymorphisms Av polys per cell: 1.5 Fragments displayed: 1088
Maximum inner -log10BCPV: 16.97 (P=1.1e-17): EACMV-Ug2[2] v ACMV-UGMld (694-1325: 632bp)
No qualifying outer fragments.



Thank you for coming.