

Supplementary Figure 1: 1.48-Mb genomic amplification at 20q11.21 (29.31 – 30.78 Mb) in HSF1 hESCs as determined by CGHAnalytics 3.4. Shown is a 3-panel screenshot from this amplified region detected in all 5 HSF1 samples evaluated. The 20q11.21 locus is depicted with increasing resolution from left to right, with the amplified region centered near 30 Mb highlighted (brown rectangle) and genes within the amplified region listed (Table 1). This amplified region provides a new CNV Locus not listed within the TCAG Database of Genomic Variants.

Supplementary Figure 2: 1.48-Mb genomic amplification at 20 q11.21 (29.31 – 30.78 Mb) in HSF1 hESCs as determined by BuddhaCGH Pipeline. (A) Segmentation view mode, depicting all aCGH probe log-ratio values across the entire chromosome. Note the amplified region centered near 30 Mb. (B) Segmentation view detail, using the tool's magnification feature, focused on the amplified region. (C) Segmentation scoring mode, depicting time-series scores for intervals of various sizes across the entire chromosome – note the highest scoring ($1.0 > 0.2$ threshold) region centered near 30 Mb. (D) Segmentation scoring detail, using the tools magnification feature, focused on the highest scoring region.

Comments on Scoring Interleaved Segments using BuddhaCGH Pipeline Software

1 Introduction

In general, given a set of interleaved segments in time order, what is the best way to score, then rank, the intervals that characterize patterns of interest?

For genomic probe data, a segment represents a chromosomal region (possibly a gene) whose copy number is amplified or deleted with respect to a control. We are interested in segment patterns that show monotonic progression in time, the most simple case being a one-time increase or decrease in copy number through the course of the time series. Moreover, we are interested in the longest contiguous segment patterns of this type, covering the largest number of base pairs.

2 Setup

Let:

c	$:=$	<i>chromosome number</i> : $c \in \mathbb{N} \cap [1, 23]$
w	$:=$	<i>window size</i> : $w \in \mathbb{N} \cap [5, 30]$
E	$:=$	<i>number of experiments</i> : $E = 5$
s	$:=$	<i>experiment cursor</i> : $s \in \mathbb{N} \cap [0, E - 1]$
A_c	$:=$	<i>number of units on chromosome c</i>
u	$:=$	<i>unit cursor</i> : $u \in \mathbb{N} \cap [0, A_c - w + 1]$
μ_s	$:=$	<i>median of experiment s from u to $u + w - 1$</i>
ϵ	$:=$	<i>minimum penalty score</i> : $\epsilon = 1$
$\sigma(c, s, u)$	$:=$	<i>value of experiment s at unit u on chromosome c</i>
λ_0	$:=$	<i>penalty factor 0</i> : $\lambda_0 = 0.0$
λ_1	$:=$	<i>penalty factor 1</i> : $\lambda_1 = 0.0$
λ_2	$:=$	<i>penalty factor 2</i> : $\lambda_2 = 3.0$
λ_3	$:=$	<i>penalty factor 3</i> : $\lambda_3 = 0.0$

Then define the score function for chromosome c at unit u to be:

$$\begin{aligned}
 \mathcal{S}_u^c &= w \cdot \frac{\sum_{s=0}^{E-1} \sum_{p=u}^{w+u-1} \sigma(c, s, p)}{E \cdot w} && \text{reward large windows of high sample values} \\
 &- \lambda_0 \cdot \sum_{p=u}^{w+u-1} \sigma(c, 0, p) && \text{penalize high values in the first experiment} \\
 &- \lambda_1 \cdot \sum_{s=0}^{E-2} \max\{(\mu_{s+1} - \mu_s), \epsilon\} && \text{minimally penalize monotonic median deltas} \\
 &- \lambda_2 \cdot \sum_{s=0}^{E-2} (\mu_{s+1} - \mu_s)^2 && \text{nonlinearly penalize monotonic median deltas} \\
 \\
 \mathcal{S}_u^c &= \frac{\mathcal{S}_u^c}{w} && \text{normalize score by window size} \\
 &+ \lambda_3 \cdot w && \text{reward large windows}
 \end{aligned}$$

3 Algorithm

Algorithm 1 SCORE-SEGMENTS : $\sigma \times \pi \rightarrow \mathcal{S}$

```
// Input:
//  $\sigma : c \times s \times u \rightarrow \mathbb{R} :=$  the segment matrix (samples,  $s$ , by units,  $u$ ) for chromosome  $c$ .
//  $\pi : c \times u \rightarrow b :=$  base pair position at center of unit  $u$  of chromosome  $c$ .
//
// Let:
//  $c_{first} \leftarrow 1, c_{last} \leftarrow 23$ 
//  $w_{first} \leftarrow 5, w_{last} \leftarrow 30, w_{step} \leftarrow 1, w_{grow} \leftarrow 1$ 
//  $\lambda_0 \leftarrow 0.0, \lambda_1 \leftarrow 0.0, \lambda_2 \leftarrow 3.0, \lambda_3 \leftarrow 0.0, \epsilon \leftarrow 1.0, \mathcal{S}_{thresh} \leftarrow 0.0$ 
//  $\mu : c \times s \times \alpha \times \beta \rightarrow \mathbb{R} :=$  median value for chromosome  $c$ , at sample  $s$ , in the regime  $[\alpha, \beta]$ .
//
// Output:
//  $\mathcal{S} :=$  the set of segment score tuples:  $\langle \pi_\alpha, \pi_\beta, \mathcal{S}_{\alpha\beta} \rangle$ .
1:  $\mathcal{S} \leftarrow \emptyset$ 
2: for  $c \leftarrow c_{first} \dots c_{last}$  step 1 do
3:   for  $w \leftarrow w_{first} \dots w_{last}$  step  $w_{grow}$  do
4:     for  $u \leftarrow 0 \dots \sigma(c).size(units) - w - 1$  step  $w_{step}$  do
5:        $\alpha \leftarrow u, \beta \leftarrow u + w - 1$ 
6:        $sum_{s_0}, sum_{all}, sum_{max}, sum_{squ}, avg_{all} \leftarrow 0$ 
7:       for  $s \leftarrow 0 \dots \sigma(c).size(samples) - 1$  step 1 do
8:          $\mu \leftarrow \mu(c, s, \alpha, \beta)$ 
9:         if  $s = 0$  then
10:           for  $p \leftarrow \alpha \dots \beta$  step 1 do
11:              $sum_{s_0} \leftarrow sum_{s_0} + \sigma(c, s, p)$ 
12:         if  $s > 0$  then
13:            $\delta_\mu \leftarrow \mu - \mu'$ 
14:            $sum_{max} \leftarrow sum_{max} + max\{\delta_\mu, \epsilon\}$ 
15:            $sum_{squ} \leftarrow sum_{squ} + \delta_\mu^2$ 
16:           for  $p \leftarrow \alpha \dots \beta$  step 1 do
17:              $sum_{all} \leftarrow sum_{all} + \sigma(c, s, p)$ 
18:            $\mu' \leftarrow \mu$ 
19:            $avg_{all} \leftarrow \frac{sum_{all}}{w \cdot \sigma(c).size(samples)}$ 
20:            $\mathcal{S}_{\alpha\beta} \leftarrow w \cdot avg_{all} - \lambda_0 \cdot sum_{s_0} - \lambda_1 \cdot sum_{max} - \lambda_2 \cdot sum_{squ}$ 
21:            $\mathcal{S}_{\alpha\beta} \leftarrow \frac{\mathcal{S}_{\alpha\beta}}{w} + \lambda_3 \cdot w$ 
22:           if  $\mathcal{S}_{\alpha\beta} > \mathcal{S}_{thresh}$  then
23:              $\mathcal{S} \leftarrow \mathcal{S} \cup \langle \pi(c, \alpha), \pi(c, \beta), \mathcal{S}_{\alpha\beta} \rangle$ 
24: return  $\mathcal{S}$ 
```
