Outlining the probability that hemizygosity of specific genomic regions are causal of some of the clinical features reported in del(1q) patients.

In order to computationally inferring the genomic segments being most likely associated with selected clinical features, we assumed that a specific trait was predominantly the outcome of the hemizygosity of specific disease locus (DL), either a protein-coding gene or a putative regulatory element, rather than the synergistic effect of the haploinsufficiency of several genomic elements.

This probability essentially depends on the penetrance of the DL and on the causative and non-causative deletions that overlap the genomic position and may be estimated by evaluating the probability for the experimental data to occur assuming that the DL intersects a given genomic location and the same probability assuming that the DL maps elsewhere in the SRO.

1) Identification of the smallest regions of overlap

At this purpose, we assigned at each *i* patient having the trait the function Y_i of the position x along the genomic segment under study:

$$y_i(x) = \begin{cases} 1 & x \in (a_i, b_i) \\ 0 & x \in (c,d) \setminus (a_i, b_i) \end{cases}$$

For i = 1,2,3..N

and for patients not associated with the trait, the function

$$\bar{y}_j(x) = \begin{cases} 1 & x \in (\bar{a}_j, \bar{b}_j) \\ 0 & x \in (c,d) \setminus (\bar{a}_j, \bar{b}_j) \end{cases}$$

For j = 1,2,3..M

where: $b_i - a_i = l_i , \bar{b}_j - \bar{a}_j = \bar{l}_j$ $\quad l_i, \bar{l}_j$ = length of the deletion of the ith, jth patient

c = most proximal boundary among the deletions

d = most distal boundary among the deletions

We ordered the set of functions y_i according to their lengths

$$l_i = b_i - a_i$$

$$Y = \{y^j(x) \text{ with } j : l_m^j \le l_n^{j+1} \; \forall m,n = 1,2,.....N\}$$

We then identified the deletion of minimal length $y^1(x)$, $y^1(x)=1$ in (a_i, b_i) , l_i= minimal length, and looked for the first element in Y, let's call it $y^{\mu}(x)$, $y^{\mu}(x)=1$ in (a_j, b_j), non-overlapping the region (a_i, b_i) :

$$(a_i, b_i) \cap (a_j, b_j) = \varnothing$$

Both regions (a_i, b_i) e (a_j, b_j) will have probability 1 to include a DL

$$P_{DL} = 1 \quad in\ (a_i, b_i) \qquad ; \qquad P_{DL} = 1 \quad in\ (a_j, b_j)$$

We then proceeded iteratively to identify any other possible region (a_k, b_k) satisfying:

$$(a_i, b_i) \cap (a_j, b_j) \cap (a_k, b_k) = \varnothing$$

Let ***v*** represents the total number of (a,b) genomic segments (primary peaks) identified.

v = total number of non-overlapping regions having probability 1 to include a DL

This set of genomic segments (primary peaks) will be then ordered according to their genomic locations :

$$(a_1, b_1)\ ,\ (a_2, b_2)\ , \ldots\ldots\ldots, (a_v, b_v) \quad of\ lengths\ l_1\ , l_2\ , \ldots\ldots\ldots, l_v\ not\ ordered$$

become

$$(a_1, b_1) \cap (a_2, b_2) \cap \ldots\ldots\ldots \cap (a_v, b_v) = \varnothing \qquad c \le a_1 < b_1 < a_2 < b_2 < \ldots\ldots\ldots < a_v < b_v \le d$$

Each peak corresponds to a genomic segment with probability 1 containing a DL, however this region may be further refined taking into account its potential overlaps with other causative deletions provided that these latter do not intercept other peaks.

Given the generic peak i (a_i, b_i) having length l_i, we identify, if any, the deletions y_k^0 :

$$y_k^0 : \begin{cases} (a_k^0, b_k^0) \cap (a_i, b_i) \neq \varnothing \\ (a_k^0, b_k^0) \cap (a_j, b_j) = \varnothing \end{cases} \quad j \neq i, j = 1,2,\ldots\ldots, v$$

The peak region becomes $(a'_i\ ,\ b'_i)$:

$$(a_i', b_i') = (a_1^0, b_1^0) \cap (a_2^0, b_2^0) \cap \ldots\ldots\ldots \cap (a_i, b_i)$$

The solution:

$$(a_i', b_i') = \varnothing$$

will determine the split of the peak into two peaks (Fig. 1).

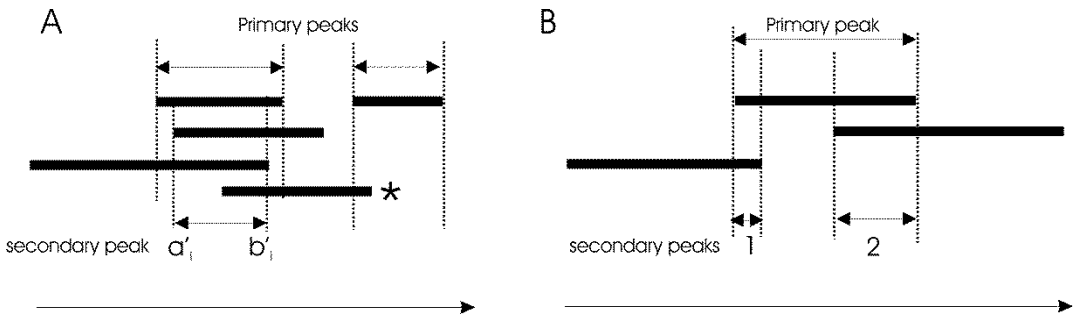

Fig. 1 Dark bars represent deletions associated with the trait. A) Secondary peaks are defined by the overlaps of causative deletions within the region of the primary peaks. Note that the deletion marked by the asterisk does not partecipate in defining the secondary peak as the DL may lie in its overlap with the second peak. B) The two deletions do not overlap, splitting the primary peak into two secondary peaks.

Each secondary peak will have probability 1 to include a DL.

2) Probability distribution inside the peaks

Up to this point, the procedure reflects the usual approach in order to identify the shortest regions of overlap in contiguous gene syndromes i.e. the graphical identification of the area of minimal overlap between deletions in patients sharing the same phenotype.

2.1) Introducing the **Δ interval**

In order to calculate this probability distribution, we scanned the genomic regions inside the peaks using a non-overlapping sliding window Δ=1 kb; each peak region **(a_i, b_i)** will be subdivided into l_i/Δ intervals having boundaries :

$$\Delta_{i,m} \equiv (a_{i,m-1}, a_{im}) \quad m = 1,2,\ldots\ldots,\frac{l_i}{\Delta} \qquad (a_{i,0} = a_i \,;\, b_i = a_{i.l_i/\Delta})$$

$$i.e. \quad a_{i,m} = a_i + m \cdot \Delta$$

Each window Δ_{im} will have an individual probability to overlap the DL which depends on the penetrance of the DL and on the number S of causative and $\overline{S}$ non-causative deletions overlapping Δ_{im}:

$$S(\Delta_{im}) = \sum_{p=1}^{N} y_p(\Delta_{im}) \qquad \overline{S}(\Delta_{im}) = \sum_{j=1}^{M} \overline{y}_j(\Delta_{im})$$

$$y_p(\Delta_{im}) = \begin{cases} 1 & \Delta_{im} \cap (a_p, b_p) \neq \varnothing \\ 0 & \Delta_{im} \cap (a_p, b_p) = \varnothing \end{cases} \qquad \bar{y}_j(\Delta_{im}) = \begin{cases} 1 & \Delta_{im} \cap (\bar{a}_j, \bar{b}_j) \neq \varnothing \\ 0 & \Delta_{im} \cap (\bar{a}_j, \bar{b}_j) = \varnothing \end{cases}$$

2.2) Baysian approach to estimate the posterior probability

The individual probability for each interval Δ_{im} inside the peaks to overlap the DL, will be estimated using the bayesian approach:

$$P(A/B) = \frac{P(A) \cdot P(B/A)}{P(B)} \qquad (1)$$

Where:

Event A*=the interval* Δ_{im} *overlaps the disease locus*

Event B*=the set of deletions overlapping* Δ_{im} *(observed experimental data)*

P(A/B)*= the posterior probability of event A occurring given that event B is true (i.e. has occurred), this probability is the probability we would like to estimate*

P(A) *= the a priori probability for the DL to overlap* Δ_{im} *.*

P(B/A): *The conditional probability for the data to be observed, given that the interval* Δ_{im} *overlaps the disease locus.*

P(B)**:** *The sum of probabilities of the events that can generate the event B (i.e. normalization factor)*

2.3 Defining the term of the probability equation

2.3.1 Introducing the term P(A)

P(A), the a priori probability for the DL to overlap Δ_{im}, depends on the size of Δ and on the length of the peak:

$$P(A) = \frac{\Delta}{l_i} \qquad l_i = \text{length of the peak}$$

2.3.2 introducing the term P(B/A) and penetrance

The probability of the event B occurring depends on the number of causative and non causative deletions and on the penetrance of the DL, through the binomial probability distribution:

$$P(B/A) = C_k^n t^{n-k}(1-t)^k \qquad n = S + \overline{S} \;\; ; \; \overline{S} = k \;\; ; \quad t = penetrance \;\; ;$$

As the real penetrance of DL is unknown, we consider the best estimator for t as the value ($0 \leq t \leq 1$) which maximizes *P(B/A)*. By differentiating *P(B/A)* with respect to t and setting the derivative function to zero we obtained :

$$t = \frac{n-k}{n}$$

2.3.3 Introducing the term *P(B)*

Event B may occur in two distinct ways:

1) DL overlaps Δ_{im} : its contribution to *P(B)* will be:

$$P(A) \cdot P(B/A)$$

2) DL does not overlap Δ_{im} (event not A= $\overline{A}$): the contribution to *P(B)* will be:

$$P(\overline{A}) \cdot P(B/\overline{A})$$

$$with \qquad P(\overline{A}) = 1 - P(A) = The\ a\ priori\ probability\ for\ the\ DL\ to\ not\ overlap\ \Delta_{im}$$

P (B) will be the sum of probabilities in 1) and in 2):

$$P(B) = P(A) \cdot P(B/A) + P(\overline{A}) \cdot P(B/\overline{A})$$

And by substituting in (1), *P(A/B)* becomes :

$$(1) \quad P(A/B) = \frac{P(A) \cdot P(B/A)}{P(A) \cdot P(B/A) + P(\overline{A}) \cdot P(B/\overline{A})}$$

We have now to calculate the term $P(B/\overline{A})$, i.e. the probability of event B occurring given that the DL does not overlap Δ_{im}.

2.3.3.1 Introducing the term $P(B/\overline{A})$

At this regard, let's consider a deletion from a patient showing the trait (causative deletion) and overlapping a peak, we may distinguish two different cases:

a) The generic deletion y_p does not overlap other DL peaks (thus, by definition, it should encompass the whole peak); given that Δ_{im} does not overlap the DL (event not A= $\overline{A}$), this

latter is included within the deletion with probability *1-P(A)* (fig. 2A) and the probability to observe such a deletion will be:

$$P(y_p / \bar{A}) = P(A)\left(\sum_{m=1}^{l_i/\Delta} t_{im} - t_{im}\right)$$

where

$$t_{im} = \frac{n-k}{n} = \frac{S_{im}}{S_{im} + \bar{S}_{im}} = \frac{\sum_{p=1}^{N} y_p}{\sum_{p=1}^{N} y_p + \sum_{j=1}^{M} \bar{y}_j}$$

b) The generic deletion y_p overlaps another (adjacent) peak . This deletion may belong to a patient showing the trait either because (1) the DL maps into the overlap between the first peak and the deletion, or (2) into the overlap with the second peak or (3) two distinct DLs are present in both overlaps (Fig 2B).
Let l_i be the length of the first peak , l_{i+1} the length of the contiguous peak, and V_{ip} , $V_{i+1,p}$ as defined by the following equations:

$$V_{ip} = \frac{(a_p, b_p) \cap l_i}{l_i} \qquad V_{i+1,p} = \frac{(a_p, b_p) \cap l_{i+1}}{l_{i+1}}$$

(1) $$< t_{i,p} > \cdot (V_{ip} - P(A)) \cdot (1 - V_{i+1,p})$$

where $$< t_{i,p} >= \frac{\sum_{l=\frac{a_p - a_i}{\Delta}+1}^{l_i/\Delta} t_{il}}{\frac{b_i - a_p}{\Delta} - 1} \qquad l \neq m$$

(2) $$< t_{i+1,p} > \cdot (1 - V_{ip}) \cdot V_{i+1,p}$$

where $$< t_{i+1,p} >= \frac{\sum_{m=1}^{\frac{b_p - a_{i+1}}{\Delta}} t_{i+1,m}}{\frac{b_p - a_{i+1}}{\Delta}}$$

(3) $$(V_{ip} - P(A)) \cdot V_{i+1,p} \cdot (< t_{i,p} > + < t_{i+1,p} > + < t_{i,p} > \cdot < t_{i+1,p} >)$$

The overall probability for y_p *occurring*, given the event $\bar{A}$ is the sum of the probabilities (1),(2),(3)

$$P(y_p / \bar{A}) = V_{ip} < t_{i,p} > + V_{i+1,p} < t_{i+1,p} > + V_{ip} V_{i+1,p} < t_{i,p} >< t_{i+1,p} > - P(A)(< t_{i,p} > + V_{i+1,p} < t_{i,p} >< t_{i+1,p} > + V_{i+1,p} < t_{i+1,p} >)$$

It should be noted that for $V_{i+1,p} = 0$ (and then $V_{i,p} = 1$), the last expression equals the expression in (a)

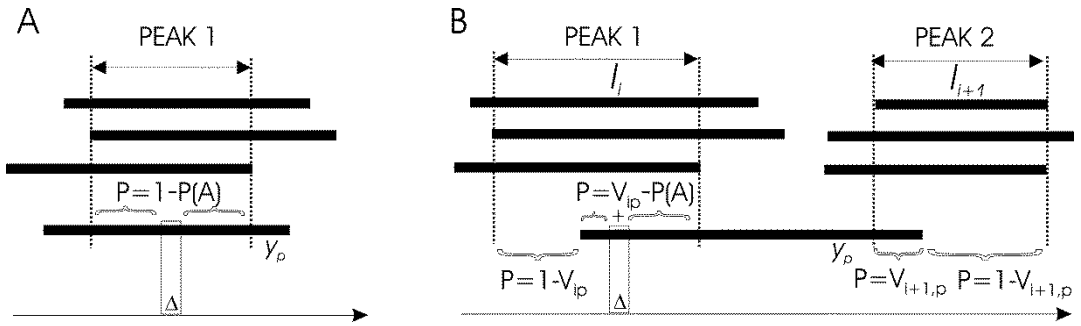

Fig. 2 Dark bars represent causative deletions. A) The deletion y_p does not overlap other peaks and B) the deletion overlaps the adjacent peak 2. The probabilities associated with overlaps between the deletion and the peak(s) are graphically represented.

Let's now consider deletions which are not associated with the trait. A generic deletion $\overline{y_p}$ may show no association with the trait either because (1) it does not overlap the DL, or (2) because, while encompassing the DL, the carrier patient does not show the trait due to the incomplete penetrance. We have also to distinguish the situations a) (fig 3A) and b) (fig 3B).

$$\bar{V}_{ip} = \frac{(\bar{a}_p, \bar{b}_p) \cap l_i}{l_i} \qquad \bar{V}_{i+1,p} = \frac{(\bar{a}_p, \bar{b}_p) \cap l_{i+1}}{l_{i+1}}$$

a) The generic deletion $\overline{y_p}$ does not overlap another peak (fig. 3A)

$$(1) \quad 1 - \bar{V}_{ip}$$

$$(2) \quad \left(\bar{V}_{ip} - P(A)\right) \cdot (1 - < t_{ip} >)$$

where

$$< t_{i,p} >= \frac{\sum_{l=\frac{\bar{a}_p - a_i}{\Delta}+1}^{l/\Delta} t_{i,l}}{\frac{\bar{b}_i - \bar{a}_p}{\Delta} - 1} \qquad l \neq m$$

The overall probability for $\overline{y}_p$ occurring, given the event $\bar{A}$ is the sum of the probabilities (1),(2)

$$(a) \quad P(\bar{y}_p / \bar{A}) = (1 - \bar{V}_{ip}) + \left(\bar{V}_{ip} - P(A)\right) \cdot (1 - < t_{i,p} >)$$

Which is equivalent to:

$$(a') \quad P(\bar{y}_p / \bar{A}) = 1 - P(A) + < t_{i,p} > (P(A) - \bar{V}_{ip})$$

b) The generic deletion$\overline{y}_p$ overlaps a second peak: (fig. 3B)

$$(1) \qquad (1-\overline{V}_{ip})\cdot(1-\overline{V}_{i+1,p})$$

$$(2) \qquad (1-<t_{i,p}>)(\overline{V}_{ip}-P(A))(1-\overline{V}_{i+1,p})$$

$$(3) \qquad (1-<t_{i+1,p}>)\cdot(1-\overline{V}_{ip})\cdot\overline{V}_{i+1,p}$$

$$(4) \qquad (\overline{V}_{ip}-P(A))\cdot\overline{V}_{i+1,p}\cdot(1-<t_{i,p}>)\cdot(1-<t_{i+1,p}>)$$

The overall probability for $\overline{y}_p$ occurring, given the event Ā is the sum of the probabilities (1),(2),(3),(4)

$$(b) \qquad P(\overline{y}_p/\overline{A})=(1-\overline{V}_{i+1,p}\cdot<t_{i+1,p}>)\cdot[1-P(A)+<t_{i,p}>\cdot(P(A)-\overline{V}_{i,p})]$$

It should be noted that for $\overline{V}_{i+1,p}$ = 0, the expression in (b) equals the expression in (a')

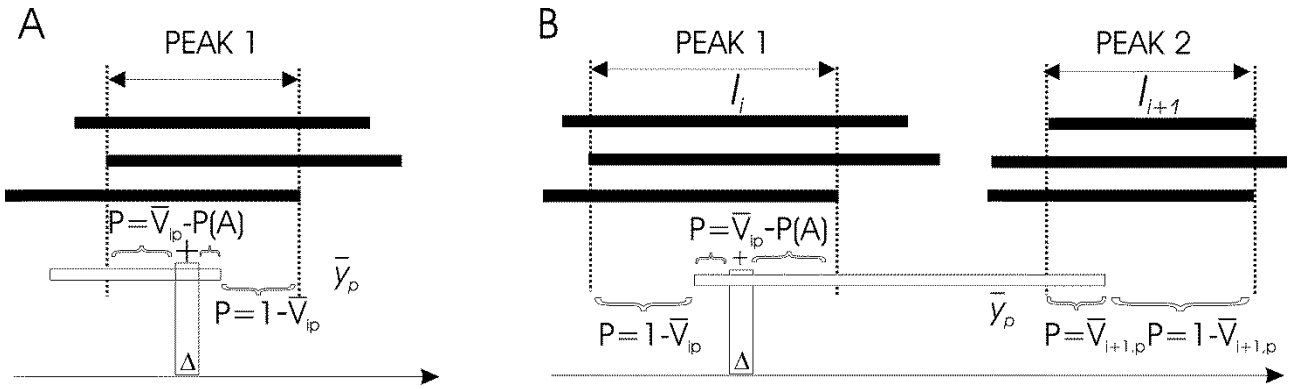

Fig. 3 Dark and white bars represent causative and non causative deletions, respectively. A) The deletion $\overline{y}_p$ does not overlap other peaks and B) the deletion overlaps the adjacent peak 2. The probabilities associated with overlaps between the non causative deletion and the peak(s) are graphically represented.

Let $\boldsymbol{P_i} = P(y_i/\overline{A})$ for i =1,2,…,n-k and $\boldsymbol{P_j} = 1 - P(\overline{y}_i/\overline{A})$ for j =n-k+1,…,n

Since $P_1, P_{2,..,} P_n$, the individual probabilities for each of the n deletions to be associated with the trait under the condition $\overline{A}$, are unlinked, one can't use the binomial coefficient as usual.

Let T_{n-h}^{n} be the sum of all the products arising from the subsets of *n-h* distinct elements from $\{P_{1,} P_{2,}.., P_n\}$

$$h = 0 \qquad T_n^n = P_1 \cdot P_2 \cdot .....P_n = \prod_{i=1}^{n} P_i$$

$$h = 1 \qquad T_{n-1}^n = \sum_{i=1}^{n} \frac{P_1 \cdot P_2 \cdot ..... P_n}{P_i} = \sum_{i=1}^{n} \frac{1}{P_i} \prod_{j=1}^{n} P_j$$

$$h = 2 \qquad T_2^n = \sum_{\substack{i=1 \\ i \neq 1}}^{n} P_1 \cdot P_i + \sum_{\substack{i=1 \\ i \neq 2}}^{n} P_2 \cdot P_i + ... + \sum_{\substack{i=2 \\ i \neq n}}^{n} P_n \cdot P_i$$

.

.

$$h = n - 1 \qquad T_1^n = \sum_{i=1}^{n} P_i$$

$$h = n \qquad T_1^n = 1$$

Finally, given *n-k* (number of associated deletions) and *k* (number of non-associated deletions) the probability $P(B/\overline{A})$, becomes:

$$P(B/\overline{A}) = \sum_{j=0}^{k} (-1)^{k-j} \cdot C_{k-j}^{n-j} \cdot T_{n-j}^{n} \quad (3)$$

Example, for n=4 and k=3, (3) becomes:

$$P(B/\overline{A}) = \sum_{j=0}^{3} (-1)^{3-j} \cdot C_{3-j}^{4-j} \cdot T_{4-j}^{4} = (-1)^3 \cdot C_3^4 \cdot T_4^4 + (-1)^2 \cdot C_2^3 \cdot T_3^4 + (-1)^1 \cdot C_1^2 \cdot T_2^4 + (-1)^0 \cdot C_0^1 \cdot T_1^4$$

$$-1 \cdot 4 \cdot P_1 \cdot P_2 \cdot P_3 \cdot P_4 + 1 \cdot 3 \cdot (P_1 \cdot P_2 \cdot P_3 + P_1 \cdot P_3 \cdot P_4 + P_1 \cdot P_2 \cdot P_4 + P_2 \cdot P_3 \cdot P_4)$$
$$-1 \cdot 2 \cdot (P_1 \cdot P_2 + P_1 \cdot P_3 + P_1 \cdot P_4 + P_2 \cdot P_3 + P_2 \cdot P_4 + P_3 \cdot P_4) + 1 \cdot 1 \cdot (P_1 + P_2 + P_3 + P_4)$$

Since the exact calculation of the term $P(B/\overline{A})$ becomes infeasible as n and k increase, we estimated $P(B/\overline{A})$ using a Monte Carlo procedure simulating $\mathbf{10^7}$ probability-weighted combinations and counting as successfully events those having n-k associated deletions. Random numbers for probability-weighted combinations were generated using the Mersenne Twister algorithm .

We can now substitute in equation (1)

$$(1) \quad P(A/B) = \frac{P(A) \cdot P(B/A)}{P(A) \cdot P(B/A) + P(\overline{A}) \cdot P(B/\overline{A})}$$

the terms $P(A)$, $P(B/A)$, $P(\overline{A})$, and $P(B/\overline{A})$ to calculate the final probability for each Δ within a specific peak.

The resulting probabilities $P_{\Delta}(A/B)$ along each peak were then normalized to 1.

The Visual Basic programming language and Excel 2007 (both from Microsoft Corporation) were respectively used to write the program and to build UCSC custom tracks (bed and bed-graph files).

In order to improve visualization at extreme probabilities values, the transformation function $y = log(P(x)) + 1 - min(log(P(x)))$, has been applied to the probability function $P(x)$ in the Log_scale track.