JOURNAL OF ALGEBRAIC STATISTICS
Vol. 6, No. 2, 2015, 133-149
ISSN 1309-3452 – www.jalgstat.com

Uncovering Proximity of Chromosome Territories using Classical Algebraic Statistics

Javier Arsuaga[1], Ido Heskia[1], Serkan Hoşten[1,*], Tatsiana Maskalevich[1]

[1] *Department of Mathematics, San Francisco State University, San Francisco CA 94132*

Abstract. Exchange type chromosome aberrations (ETCAs) are rearrangements of the genome that occur when chromosomes break and the resulting fragments rejoin with fragments from other chromosomes or from other regions within the same chromosome. ETCAs are commonly observed in cancer cells and in cells exposed to radiation. The frequency of these chromosome rearrangements is correlated with their spatial proximity, therefore it can be used to infer the three dimensional organization of the genome. Extracting statistical significance of spatial proximity from cancer and radiation data has remained somewhat elusive because of the sparsity of the data. We here propose a new approach to study the three dimensional organization of the genome using algebraic statistics. We test our method on a published data set of irradiated human blood lymphocyte cells. We provide a rigorous method for testing the overall organization of the genome, and in agreement with previous results we find a random relative positioning of chromosomes with the exception of the chromosome pairs {1,22} and {13,14} that have a significantly larger number of ETCAs than the rest of the chromosome pairs suggesting their spatial proximity. We conclude that algebraic methods can successfully be used to analyze genetic data and have potential applications to larger and more complex data sets.

2000 Mathematics Subject Classifications: 62H17, 60J22, 13P10

Key Words and Phrases: Chromosome territories, Chromosome aberrations, Markov basis, Monte Carlo Markov Chain

1. Introduction

During the early stages of the cell cycle the mammalian genome is organized in chromosome territories [16, 39] (for a review see [18]). When DNA damaging agents, such as radiation, cross the cell nucleus they introduce double strand breaks that produce chromosome fragments. These chromosome fragments need to be rejoined with their original partners for the cell to survive. A small percentage of breaks however are incorrectly rejoined introducing exchange type chromosome aberrations (ETCAs). ETCAs between

*Corresponding author.

Email address: serkan@sfsu.edu (S. Hoşten)

non-homologue chromosomes can be detected in the laboratory by means of diverse chromosome painting and sequencing techniques (see Figure 1 for an example of ETCAs detected by the chromosome painting technique Spectral Karyotyping (SKY) [45]). It is expected that chromosomes that are in close spatial proximity form ETCAs more often than those that are far apart [6, 13, 37, 41, 50]. Therefore the frequency of ETCAs between non-homologous chromosomes is informative of their relative position and it can be used to reconstruct the three dimensional structure of the genome.

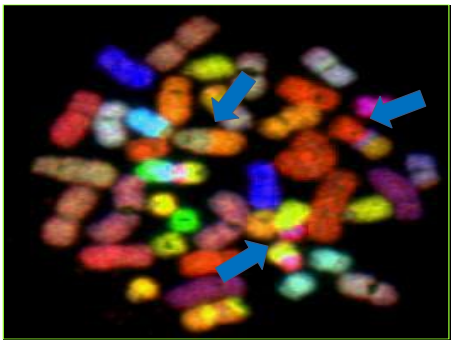

Figure 1: Examples of ETCAs found in a tumor cell. Pairs of homologue chromosomes were painted the same color using SKY. The arrows point to ETCAs. Figure kindly provided by J.L. García.

Chromosome painting techniques such Fluorescence in-situ Hybridization (FISH) or its multicolor versions such as SKY and multiplex FISH (mFISH) [46, 45] paint every pair of homologue chromosomes the same color (Figure 1). These techniques have shown that the position of chromosome territories with respect to the center of the nucleus is driven by gene density, chromosome size and/or local chromatin geometry [11, 29, 38]. In human lymphocytes gene-rich chromosomes such as $\{1, 19, 17, 22\}$ are located near the center of the nucleus [11, 18] while gene-poor chromosomes such as $\{2, 4, 13, 18\}$ are located closer to the periphery of the nucleus [11, 18]. Studies measuring the relative position of chromosomes, using radiation induced ETCAs, on the other hand have shown that chromosomes are randomly located with respect to each other with the exception of a few chromosome clusters [15, 25, 43, 44]. Interestingly this overall random relative organization has been corroborated by new sequencing techniques [33, 35, 50].

Quantitative analyses of the relative position of chromosome territories commonly use tables whose entries are the number of ETCAs detected for any two non-homologue chromosomes such as the one we use in Table 1. Several methods have been proposed to study frequency tables of radiation induced ETCAs. In [9] tables of radiation induced ETCAs detected by FISH were analyzed. In this type of studies frequency tables were densely populated and with large entries, therefore chi-square statistics were used to find significant clusters of chromosomes. In [2, 15] similar tables were generated employing mFISH data. Although more accurate in some respects, the latter data were more sparse and with small entries. In this case the proximity of chromosome territories was tested by assigning p-values to clusters of previously reported chromosome territories. Here we propose a model-based approach that builds a simple log-linear model to test the proximity

of pairs of chromosome territories, referred to as chromosome pairs from now on, and we use a Markov Chain Monte Carlo method based on the theory of algebraic statistics to assign significance. The mFISH data we analyze have already appeared in the literature [2, 15, 34].

In our study we test a no-proximity effect model and a single pair proximity model by sampling tables that have the same sufficient statistic as those observed experimentally. Sampling of the tables is performed by running a Markov Chain Monte Carlo algorithm that uses a Markov basis [21, 22] for a second hypersimplex [20]. This kind of method is one of the early contributions of algebraic statistics, hence "classical algebraic statistics" in the title of this work. Our results indicate that we could not reject the hypothesis of random relative arrangement of chromosome territories when radiation induced chromosome aberrations were analyzed. This result is in qualitative agreement with previous studies [2, 15, 34] and suggests that the specific positions of any two pairs of chromosomes do not influence the frequency of aberrations observed. However, by assuming the existence of a proximity biasing factor we found chromosome pairs $\{1, 22\}$ and $\{13, 14\}$ to be significant. We conclude that methods develop in algebraic statistics are suitable for analyzing genetic data of moderate size in which data sparsity or low numbers of measurements are present.

2. Data and Methods

To test for proximity of chromosome territories we used a radiation induced exchange type chromosome aberration table published in [2, 49]. In these experiments cells from healthy donors were irradiated with sparsely ionizing γ-rays at different doses, and mFISH [46] was used to detect ETCAs. This table includes a total of 3585 records of human peripheral blood lymphocytes irradiated with sparsely ionizing radiation at different doses.

The frequency of ETCAs was summarized by recording the number of cells in which at least one exchange between two non-homologous chromosomes occurred. This quantity is robust with respect to noise introduced by apparently incomplete aberrations (i.e. those aberrations with not all fragments accounted for) and reduces false positives. Following previous publications [2, 15] we denote these values by $f(j, k)$ where j and k are the chromosomes that participate in the exchange. These values are presented in Table 1 as a 22×22 upper-triangular table.

2.1. The No-proximity and Single-pair Proximity Effects as Log-Linear Models

ETCAs between two non-homologous chromosomes j and k ($j \neq k$) were modeled by a single discrete random variable X with $\binom{22}{2} = 231$ values corresponding to all possible pairs of non-homologue chromosomes with a probability density function given by

$$p : \{(j, k) : 1 \leq j < k \leq 22\} \longrightarrow \Delta_{231}$$

Table 1: Table of ETCAs in 3585 human lymphocytes as reported in [2, 49]. Each entry $f(j,k)$ holds the number of cells in which at least one exchange between chromosomes j and k was recorded. The total number of cells in which a given chromosome was involved in at least one exchange appears in the "sum" column.

Chr	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Sum
1	44	38	42	29	26	29	18	39	29	25	18	15	18	34	31	22	12	14	22	9	27	541
2		43	37	32	30	24	25	29	16	24	30	29	9	26	8	24	8	7	12	13	15	485
3			21	31	32	24	21	26	23	25	23	21	18	18	19	21	11	17	11	12	10	465
4				23	27	28	24	26	20	13	19	23	22	20	16	18	11	6	12	10	7	425
5					17	31	26	25	24	30	25	25	15	19	8	19	13	7	16	7	4	426
6						18	22	21	31	13	30	18	15	19	14	15	13	10	9	8	7	395
7							20	20	17	28	25	13	18	8	18	23	11	9	19	6	7	396
8								13	12	24	11	25	15	16	12	16	17	4	9	7	8	345
9									21	25	7	23	23	27	20	15	22	8	9	7	10	416
10										18	21	14	14	10	19	14	9	5	11	7	3	338
11											25	5	15	16	19	15	8	10	12	3	11	364
12												9	16	9	12	16	8	13	10	5	5	337
13													29	10	10	7	16	5	6	7	9	319
14														22	13	6	10	2	6	13	11	310
15															22	13	9	7	11	7	9	332
16																12	15	12	20	8	13	321
17																	5	4	11	5	10	291
18																		2	11	9	3	223
19																			6	0	8	156
20																				7	10	240
21																					6	156
22																						193

where the set

$$\Delta_{231} \;=\; \{(p_{12}, p_{13}, \ldots, p_{2122}) \in \mathbb{R}^{231} \;:\; p_{jk} \geq 0 \;\; \sum_{j,k} p_{jk} = 1\}$$

is the probability simplex in $\mathbb{R}^{231}$ and $p_{jk} = p(j,k)$. In this no-proximity effect model the probability of observing an ETCA between the chromosomes j and k is

$$p_{jk} \;=\; \theta_j \theta_k \;\; 1 \leq j < k \leq 22.$$

where $\theta_1, \theta_2, \ldots, \theta_{22}$ are positive parameters. More precisely, the no-proximity effect model is the image of $\mathbb{R}^{22}_{+}$ in Δ_{231} under the map $\phi(\theta_1, \ldots, \theta_{22}) = (\theta_j \theta_k \;:\; 1 \leq j < k \leq 22)$. This model can be linearized by applying logarithms to both sides of the equation, which gives $\log(p_{jk}) = \log(\theta_j) + \log(\theta_k) = \beta_j + \beta_k$ and shows that our model is a log-linear model [14]. Therefore the no-proximity effect model is given by a 22×231 design matrix, denoted by $A(22)$, whose columns are $e_j + e_k$ with $1 \leq j < k \leq 22$ where e_j is the jth standard unit vector in $\mathbb{R}^{22}$. The columns of $A(22)$ are the vertices of the *second hypersymplex* in $\mathbb{R}^{22}$ [20].

For a fixed pair $\{r, s\}$ with $r \neq s$ we define an extended model that we call a *single-pair proximity effect model*. This model is given by the map $\phi' : \mathbb{R}^{23}_{+} \to \Delta_{231}$ defined as

$$p_{jk} = \phi'(\theta_1, .., \theta_{22}, \mu_{rs}) = \begin{cases} \theta_r \theta_s \mu_{rs} & \text{if } j = r \text{ and } k = s, \\ \theta_j \theta_k & \text{otherwise.} \end{cases} \tag{1}$$

The parameter μ_{rs} is a bias factor for the frequency of an observed exchange between chromosomes r and s. This bias factor corresponds to a proximity factor between the territories of two chromosomes r and s. By taking the logarithm on both sides of the

above equation we get the parametrization in logarithmic coordinates:

$$\log(p_{jk}) = \begin{cases} \beta_r + \beta_s + \alpha_{rs} & \text{if } j = r, k = s, \\ \beta_j + \beta_k & \text{otherwise.} \end{cases} \tag{2}$$

The single-pair proximity effect model for any pair of chromosomes $\{r, s\}$ is also a log-linear model which extends the no-proximity effect model since the set of probability distributions in this model are those in the image of the map ϕ' with $\alpha_{rs} = 0$. The design matrix $A[r, s]$ defining this model is a 23×231 matrix, identical to $A(22)$ in its first 22 rows, and with an extra row of all zeros except a 1 in the column corresponding to p_{rs}.

In order to assign a p-value to the goodness-of-fit test we propose to sample tables that are similar to those observed experimentally. More specifically we sample tables with the same *minimal sufficient statistic* as the data table f. For each chromosome k we have the *marginal total*

$$u_k = \sum_{j:\, j \neq k} f(j, k) \qquad \text{for } k = 1, \ldots, 22.$$

This quantity is displayed in the `Sum` column of Table 1. The marginal total of a given chromosome is a measure of the propensity of each individual chromosome to form ETCAs. In radiation studies this quantity has been associated to the sensitivity of the chromosomes to radiation [2, 15] and to repair mechanisms [51]. For the no-proximity effect model ϕ the collection of marginal totals $u = (u_k : k = 1, \ldots, 22)$ is the minimal sufficient statistic. For the single-pair proximity effect model for the chromosome pair $\{r, s\}$, the minimal sufficient statistic is the same marginal sums together with $u_{rs} = f(r, s)$. The set of tables with the same sufficient statistic is called *the fiber* of the experimentally observed table and is denoted by $\mathcal{F}(u)$. It is well-known that $\mathcal{F}(u)$ consists of lattice points in a polytope.

2.2. Maximum Likelihood Estimation

The maximum likelihood estimator (MLE) tables of the data with respect to the no-proximity effect and single-pair proximity models were computed by the standard numerical algorithm *Iterative Proportional Scaling* [19]. For log-linear models, the algorithm converges to the unique MLE table $\hat{f}$ such that $\frac{1}{N}\hat{f}$ lies on the model where N is the sample size and $\hat{f}$ has the same sufficient statistic u as the data table f. The existence of this unique table is guaranteed by Birch's Theorem (see [22]).

2.3. Hypothesis testing and Monte-Carlo simulations

Our goodness-of-fit test for the no-proximity effect model uses the standard chi-square statistic

$$\chi^2(F) = \sum_{1 \leq j < k \leq 22} \frac{(F(j, k) - \hat{f}(j, k))^2}{\hat{f}(j, k)}. \tag{3}$$

where $\hat{f}(j,k)$ are the entries of the MLE table with respect to the no-proximity effect model given the data table f, and $F(j,k)$ are the entries of tables F drawn from the fiber $\mathcal{F}(u)$. Typically large $\chi^2(F)$ values would indicate that the data table f is "close" to $\hat{f}$ providing no evidence for rejecting no-proximity effect model.

The fiber $\mathcal{F}(u)$ for the no-proximity effect model is very large while at the same time the data table f has some small entries including a zero entry. These observations point to using the well-established standard Markov Chain Monte Carlo method for running the goodness-of-fit test. One might think that computing a Markov basis for the no-proximity effect model (or equivalently the second hypersimplex $A(22)$) is intractable. Luckily, a Markov basis (in fact a *Gröbner basis*) for this model is available, see [47, Theorem 9.1] and [20]. Our Markov basis is defined by the following set of moves: For each $1 \leq i < j < k < \ell \leq 22$ one defines two moves $m[i,j;k,\ell]$ and $m[i,\ell;j,k]$. The first move is a table where the (i,j) and (k,ℓ) entries are set to equal 1, the (i,k) and (j,ℓ) entries are set to equal -1, and all the other entries are set to equal 0. The second move is a table where the (i,ℓ) and (j,k) entries are set to equal 1, the (i,k) and (j,ℓ) entries are set to equal -1, and all the other entries are set to equal 0. These tables together with their negatives $-m[i,j;k,l]$ and $-m[i,\ell;j,k]$ comprise our Markov basis $\mathcal{B}$. This Markov basis contains $2 \cdot 2 \cdot \binom{22}{4} = 29,260$ moves.

Using this MCMC we generated a set of m random tables $f_1, \ldots, f_m$ using the Metropolis-Hastings algorithm and estimated the p-value of goodness-of-fit test by

$$\frac{1}{m}\left(\sum_{\chi^2(f_i)\geq\chi^2(f)} 1\right).$$

One important parameter of the Metropolis-Hastings algorithm is the number of steps it requires between each selection of tables f_i and f_{i+1}. The rule of thumb is that one needs sufficient number of steps so that the Markov chain can reach any table in $\mathcal{F}(u)$ starting from an arbitrary table. We followed the method in [24] to heuristically determine this number of steps: It is a consequence of the Gröbner basis theory that there is a unique table T_{unique} in $\mathcal{F}(u)$ where none of the Gröbner basis moves can be applied, and every table in $\mathcal{F}(u)$ is connected to T_{unique} table via the Markov basis moves. This table is the unique reduced normal form of the data table with respect to the Gröbner basis from which our Markov basis $\mathcal{B}$ is constructed. Empirically, the average number of steps one needs to go from a randomly generated table in $\mathcal{F}(u)$ to T_{unique} using the moves in $\mathcal{B}$ is about 15000. Hence, we estimate that the number of steps needed to connect two tables in $\mathcal{F}(u)$ is bounded by 30000.

2.4. Log-Ratio Test

Since the single-pair proximity effect model contains the no-proximity effect model we compared the relative fit of the two models by a *likelihood ratio test* . The likelihood ratio

test statistic is defined as:

$$G^2 = 2 \sum_{1 \le j < k \le 22} \hat{f}^1_{jk} \log \left(\frac{\hat{f}^1_{jk}}{\hat{f}^0_{jk}} \right) \tag{4}$$

where $\hat{f}^1_{jk}$ is the MLE with respect to the single-pair proximity effect model and $\hat{f}^0_{jk}$ is the MLE with respect to the no-proximity effect model. It is well known [14, Theorem 10.2.8] that, for large sample sizes N, if the null hypothesis is true (i.e. if the data fits the no-proximity effect model better than it fits the single-pair proximity effect model) then G^2 has a χ^2 distribution with degrees of freedom equal to the difference of the ranks of the nested log-linear models. In our case, the rank of the no-proximity effect model is equal to $\operatorname{rank} A(22) = 21$ and the rank of the single-pair proximity effect effect model is equal to $\operatorname{rank} A[r, s] = 22$.

3. Numerical Results

Table 2: Maximum Likelihood Estimate for experimental Table 1 obtained from irradiated human lymphocytes.

Chr	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	Sum
1	47	43	38	37	33	33	27	34	26	28	25	23	22	24	23	20	14	8.9	15	8.8	11	541
2		37	32	32	29	29	24	30	23	25	22	21	20	22	21	18	13	8.3	14	8.2	11	485
3			30	30	27	27	23	28	22	24	22	20	19	21	20	18	13	8.2	14	8.2	11	465
4				27	24	24	21	26	20	22	20	18	18	19	18	16	12	7.8	13	7.8	10	425
5					24	24	21	26	20	22	20	19	18	19	19	17	12	8	13	8	10	426
6						22	19	24	19	20	19	17	17	18	17	16	12	7.7	12	7.7	9.7	395
7							19	24	19	20	19	17	17	18	17	16	12	7.9	13	7.9	9.9	396
8								20	16	18	16	15	15	16	15	14	11	7.2	11	7.2	9	345
9									20	21	20	18	18	19	18	17	13	8.6	13	8.5	11	416
10										17	16	15	15	16	15	14	11	7.3	11	7.4	9.2	338
11											17	16	16	17	16	15	11	8	12	8	9.9	364
12												15	15	16	15	14	11	7.6	12	7.6	9.4	337
13													14	15	15	13	10	7.4	11	7.4	9.1	319
14														15	14	13	10	7.4	11	7.4	9	310
15															15	14	11	7.9	12	7.9	9.6	332
16																14	11	7.9	11	7.9	9.5	321
17																	10	7.4	11	7.4	8.9	291
18																		6	8.6	6	7.3	223
19																			6.4	4.6	5.6	156
20																				6.6	7.8	240
21																					5.7	156
22																						193

We first computed the MLE tables for the no-proximity effect and the single-pair models as discussed in the previous section. This algorithm preserves the minimal sufficient statistic which guarantees that the MLE table belongs to the fiber of the data table. The MLE table for the no-proximity effect model is displayed above. Only two significant digits are shown. The table for the no-proximity effect reveals some marked differences with the experimentally observed table. Most notably, entries for each chromosome tend to be more homogeneous than those in the experiment. This is particularly true for small chromosomes 19 to 22.

Using the MCMC approach explained earlier we generated 3×10^{10} tables which were sampled every 3×10^4 times to reduce the intrinsic correlation in the Markov Chain. We

therefore obtained a sample size of 10^6 tables. The chi-square statistic for each of the sampled table was computed. Interestingly, none of the 10^6 tables generated for Table 1 had a test statistic smaller than the experimentally observed. Hence the no-proximity effect model could not be rejected. This result is somewhat surprising but can be explained by estimating the number of tables that are contained in the ellipsoid defined by (3). We will give the details of this estimation in the Appendix. These results show that the data can be well fit by the no-proximity model and that the relative positions of chromosomes are random. We used the MCMC approach since our data table has some small entries – in fact, there is an entry with value 0. However, it is interesting to compare our result to an asymtotic chi-square test. The chi-square statistic of Table 1 is equal to 346.63 and the degree of freedom for the χ^2 distribution is $\binom{22}{2} - 21 = 210$ (the rank of the design matrix is 21). The corresponding chance probability is < 0.0001. In other words, from an asymptotic test we would have concluded that we reject the no-proximity effect hypothesis. However, this is not warranted.

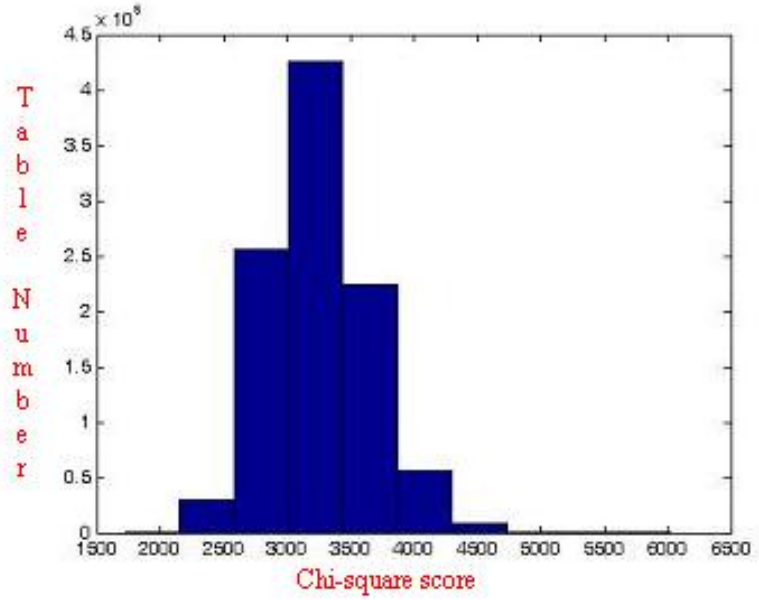

Figure 2: The resulting chi-squares of the randomly generated 1,000,000 tables in $\mathcal{F}(u)$ which give the p-value of 1 to Table 1. The chi-square statistic of Table 1 is 346.63, which is less than any of the chi-square values of the randomly generated tables. Image is taken from [24].

Microscopy observations however have shown that some groups of chromosomes tend to be close to each other and form exchange type aberrations more frequently than what one would predict using the no-proximity effect model [12]. These chromosomes include those that are found in the center of the nucleus [2, 15] and those that form the nucleolus [2]. This observation is further supported by the large positive deviations between entries in the experimentally observed table and the MLE table. The differences between the observed table and the MLE table are shown in Table 3. A positive entry indicates that

Table 3: Table of deviations between observed and MLE counts. Each entry is the difference between the observed counts and MLE for each pair of chromosomes j and k computed for Table 2.

Chr	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
1	-3.2	-5.5	4.2	-8.2	-7.2	-3.8	-9	5	3.3	-2.9	-7	-8.1	-4	10	8.4	2.2	-2	5.1	6.8	0.2	16
2		6.2	4.6	-0.1	1.2	-4.6	1.1	-0.9	-6.9	-0.9	7.6	8.2	-11	4.4	-13	5.9	-4.9	-1.3	-2.1	4.8	4.3
3			-9.4	0.8	4.7	-3.2	-1.8	-2.4	1.1	1.2	1.4	0.9	-1.3	-2.8	-0.9	3.4	-1.7	8.8	-2.8	3.8	-0.5
4				-3.9	2.5	3.6	3.3	0.3	0.0	-8.7	-0.8	4.5	4.2	0.8	-2.4	1.7	-0.9	-1.8	-0.9	2.2	-3
5					-7.5	6.6	5.2	-0.6	3.9	8.2	5.1	6.4	-3	-0.4	-11	2.5	0.9	-1	2.9	-1.0	-6.2
6						-4.4	2.8	-2.5	12	-7.2	11	0.6	-1.8	0.9	-3.4	-0.5	1.5	2.3	-3.5	0.3	-2.7
7							0.7	-3.5	-1.7	7.7	6.4	-4.5	1.1	-10	0.5	7.3	-0.7	1.1	6.3	-1.9	-2.9
8								-7.2	-4.3	6.4	-5.2	9.7	0.1	0.0	-3.5	2	6.5	-3.2	-2.4	-0.2	-1
9									1.2	3.7	-13	4.5	5.1	7.9	1.6	-1.6	9.5	-0.6	-4.5	-1.5	-0.7
10										0.8	5	-1.1	-0.7	-5.8	3.7	0.1	-1.6	-2.3	-0.4	-0.4	-6.2
11											7.8	-11	-0.9	-1.0	2.6	0.1	-3.4	2	-0.3	-5	1.1
12												-6.2	1.2	-6.8	-3.3	2	-2.8	5.4	-1.6	-2.6	-4.4
13													15	-5	-4.6	-6.4	5.6	-2.4	-5.2	-0.4	-0.1
14														7.3	-1.3	-7.1	-0.2	-5.4	-5	5.6	2
15															6.8	-0.9	-2	-0.9	-0.7	-0.9	-0.6
16																-1.6	4.2	4.1	8.5	0.1	3.5
17																	-5	-3.4	0.3	-2.4	1.1
18																		-4	2.4	3	-4.3
19																			-0.4	-4.6	2.4
20																				0.4	2.2
21																					0.4

Table 4: Table of pairs of chromosomes with their corresponding Chi-squared value and p-value corresponding to their extended interaction model and p-value adjusted with the Bonferroni correction. Significant p-values are labeled in boldface.

	p-value		
Chromosome Pair	Chi-squared	p-value before correction	Bonferroni corrected p-value
$\{1, 22\}$	**17.27**	**0.00005**	**0.00138**
$\{13, 14\}$	**13.66**	**0.00022**	**0.01012**
$\{3, 19\}$	7.87	0.00502	0.23092
$\{6, 10\}$	7.78	0.00527	0.24242
$\{6, 12\}$	6.85	0.00888	0.40848
$\{9, 18\}$	6.51	0.0107	0.4922
$\{16, 20\}$	5.73	0.01671	0.76866
$\{8, 13\}$	5.72	0.01673	0.76958

the observed table had more exchanges than the number predicted by the MLE table while a negative entry shows cases where the MLE table had more exchanges than the observed table. The largest positive entries in the table are those for chromosome pairs {1,22}, {13,14}, {1,15}, and {9,18}. We therefore tested all chromosome pairs on the single-pair proximity effect model and performed the log-ratio tests considering the original log-linear models against the modified models for all chromosome pairs. We found eight pairs of chromosome pairs that were significant (Column 1 in Table 4). However only two of them were significant after correction for multiple testing using Bonferroni [1] (Column 4 in Table 4).

4. Discussion

In this work we have presented a model-based approach to determine the relative positioning of chromosome territories from ETCA frequency tables that are sparse and with small entries. In previous work a method for dealing with small entries was reported [15], however the assignment of p-values to specific clusters of chromosome territories was based on groups of chromosomes previously found in the literature. Our method builds on the techniques developed in classical algebraic statistics by Diaconis and Sturmfels [21].

Several theoretical models have been proposed for the large scale organization of the human genome (e.g. [6, 7, 12, 23, 30, 33, 4]). These models are mostly based on experimental observations that measure either the radial position of chromosome territories with respect to the center of the nucleus or the relative position of chromosome territories. Our results show that the overall distribution of chromosome exchanges can be simply explained by a model in which the relative position of chromosome territories is random. This finding does not quantitatively agree with the results reported in [2, 15] or with the asymptotic results reported here since a small deviation from randomness was found in those studies. However all studies agree upon a rather random organization of chromosome territories. Several sources can be contributing to this apparently random organization of the genome. The first is imposed by the limitation of the data. Chromosome painting techniques are limited by the fact that homologue chromosomes are painted the same color, and this evidently introduces unavoidable noise since territories of homologous chromosomes can be positioned in very different environments (i.e. with different neighboring chromosome territories). It is also possible that cell to cell variation is very large in these samples or that there is a severe reorganization of the chromosome territories after the radiation insult. The fact that new sequencing analysis is consistent with this overall picture suggest that radiation has a small repositioning effect [35].

Clustering of chromosome territories, on the other hand, is a consistent observation and it is believed to have functional significance, both in the form of gene expression regulation and in the form of chromosome aberration formation in genetic diseases. Our study shows that a small fraction of chromosome pairs deviate from this picture of random positioning of chromosome territories and aberration formation. We identified eight pairs of chromosome territories that were significant prior to multiple testing correction (Table 4). The first two pairs of chromosome territories (ie. $\{13, 14\}$ and $\{1, 22\}$) were also significant in [2, 15, 49]. There is an explanation for the significance of these pairs although their true functional significance remains to be determined [6] . The pair $\{13, 14\}$ is part of the cluster of chromosomes in the nucleolus $\{13, 14, 15, 21, 22\}$, an organelle that brings chromosome territories together for specific needs of the cell. The second pair $\{1, 22\}$ has been found to be part of a cluster of chromosome territories $\{1, 16, 17, 19, 22\}$ located in the center of the nucleus of lymphocyte cells. We used a Bonferroni correction method for multiple testing. This method is known to be very conservative and it is possible that we rejected some informative pairs. In fact all the pairs that were not significant after Bonferroni correction have been reported in blood malignancies suggesting that proximity of these pairs of chromosome territories may be somewhat common and furthermore may

have an important role in the development of these diseases [5, 10, 26, 27, 36, 52].

It is our intention to improve our results by including better outlier detection tools that help identify other chromosome pairs [31, 40] and by incorporating these results into the developent of biophysical models. These models are based on different properties of the genome that can obtained from basic physical priciples such as the radial organization of chromosomes using overlapping sphere or ellipsoid packings [17, 28, 48] , gene density [30] or DNA decondensation processes [42] or through the folding of chromatin fibers [4, 8, 33]

Acknowledgements This work is partially supported by NSF grant DMS-1217324 and NIH grant RO1-GM109457 to J. Arsuaga. We want to thank J. L. García from Centro de Investigació del cancer de La Universidad de Salamanca (Spain) for sharing Figure 1.

5. Appendix

In this last section we return to our remark that none of the 10^6 tables generated for Table 1 in our MCMC procedure had a test statistic smaller than $\chi^2(f)$. In order to give a heuristic explanation for this behavior we will estimate the size of $\mathcal{F}(u)$ and also estimate the size of the set of the tables F where $\chi^2(F) \leq \chi^2(f)$.

We first give a lower bound for the cardinality of $\mathcal{F}(u)$, which is the set of lattice points in a polytope. A standard computational tool such as `Latte` [3] cannot compute the cardinality of this immensely large set. Instead we employ a divide-and-conquer approach where we consider the subtables consisting of the chromosomes 1 through 8 (Subtable A), chromosomes 8 through 15 (Subtable B), chromosomes 15 through 22 (Subtable C) in Table 5.

Table 5: Subtables A-C

Chr	2	3	4	5	6	7	8	Sum
1	44	38	42	29	26	29	18	226
2		43	37	32	30	24	25	235
3			21	31	32	24	21	210
4				23	27	28	24	202
5					17	31	26	189
6						18	22	172
7							20	174
8								156

Chr	9	10	11	12	13	14	15	Sum
8	13	12	24	11	25	15	16	116
9		21	25	7	23	23	27	139
10			18	21	14	14	10	110
11				25	5	15	16	128
12					9	16	9	98
13						29	10	115
14							22	134
15								110

Chr	16	17	18	19	20	21	22	Sum
15	22	13	9	7	11	7	9	78
16		12	15	12	20	8	13	102
17			5	4	11	5	10	60
18				2	11	9	3	54
19					6	0	8	39
20						7	10	76
21							6	42
22								59

The jth entry in the *Sum* column in each subtable refers to the sum of $f(j,k)$ over all chromosomes $k \neq j$ included in the subtable. See for instance $1-8$ in Subtable A in Table 5. The remaining entries of Table 1 were subdivided into six rectangular subtables. Each of these subtables are indexed by two subsets of chromosomes J and K: Subtable 1 ($J = 1-4$, $K = 9-15$), Subtable 2 ($J = 5-7$, $K = 9-15$), Subtable 3 ($J = 1-4$, $K = 16-22$), Subtable 4 ($J = 5-7$, $K = 16-22$), Subtable 5 ($J = 8-11$, $K = 16-22$), Subtable 6 ($J = 12-14$, $K = 16-22$) in Table 6.

Table 6: Subtables 1-6

Chr	9	10	11	12	13	14	15	RSum
1	39	29	25	18	15	18	34	178
2	29	16	24	30	29	9	26	163
3	26	23	25	23	21	18	18	154
4	26	20	13	19	23	22	20	143
CSum	120	80	87	90	88	67	98	

Chr	9	10	11	12	13	14	15	RSum
5	25	24	30	25	25	15	19	163
6	21	31	13	30	18	15	19	147
7	20	17	28	25	13	18	8	129
CSum	66	72	71	80	56	48	46	

Chr	16	17	18	19	20	21	22	RSum
1	31	22	12	14	22	9	27	137
2	8	24	8	7	12	13	15	87
3	19	21	11	17	11	12	10	101
4	16	18	11	6	12	10	7	80
CSum	74	85	42	44	57	44	59	

Chr	16	17	18	19	20	21	22	RSum
5	8	19	13	7	16	7	4	74
6	14	15	13	10	9	8	7	76
7	18	23	11	9	19	6	7	93
CSum	40	57	37	26	44	21	18	

Chr	16	17	18	19	20	21	22	RSum
8	12	16	17	4	9	7	8	73
9	20	15	22	8	9	7	10	91
10	19	14	9	5	11	7	3	68
11	19	15	8	10	12	3	11	78
CSum	70	60	56	27	41	24	32	

Chr	16	17	18	19	20	21	22	RSum
12	12	16	8	13	10	5	5	69
13	10	7	16	5	6	7	9	60
14	13	6	10	2	6	13	11	61
CSum	35	29	34	20	22	25	25	

In the above tables, the jth entry in the `RSum` column refers to the sum of the numbers in the jth row in the corresponding table, and the kth entry in the `CSum` row refers to the sum of the numbers in the kth column in the corresponding table. Any table that has been subdivided in a total of 9 subtables where subtables $A-C$ have same `Sum` column as in the subtables $A-C$ of the data table and where subtables $1-6$ have the same `RSum` and `CSum` columns/rows as in subtables $1-6$ of the data table is in $\mathcal{F}(u)$. So a lower bound for the cardinality of $\mathcal{F}(u)$ can be obtained by the product of the number of subtables of type A, B, C and $1, \ldots, 6$ with the given `Sum` and `RSum`/`CSum` columns/rows. The total number of such subtables range from 10^{34} for Table A to 10^{14} for Table 6. The exact number of tables calculated by Latte [3] are shown in Table 7.

Table 7: Total number subtables associated to subtables A-C and 1-6

Subtable	Size
A	29524709537992399627527976593861 90
B	252762217255461089482462934497
C	242451808378958740321921
1	384937707376563538670706387547
2	11636397863410272633
3	51895845228141509162048464
4	5538280355961059
5	336625602844011493310899
6	777971438252448

The product of these numbers is in the order of 10^{214}. This estimate however can be improved by the following arguments. Most of the $29,260$ Markov moves can be applied to Table 1 without changing the `Sum` column. The few moves that cannot be applied are those that make the $(19, 21)$ entry negative. In addition, only 1554 moves will not change the `Sum` or `RSum`/`CSum` entries of the above 9 Subtables since the moves occur completely inside each of the subtables. The remaining $27,706$ Markov moves will alter at least two of

these subtables so that their `Sum` and/or `RSum`/`CSum` entries will change. If we repeat the above `Latte` calculations for the nine subtables obtained after the application of each one of these 27,706 Markov moves (in other words those that change `Sum` and/or `RSum`/`CSum`) we will find that the number of these new tables will be again of the order 6×10^{214}. We can repeat the same argument when two or more Markov moves are considered. For instance if we applied exactly 30 Markov moves of the 27706 possible moves in sequence then we would obtain a total number of new tables in $\mathcal{F}(u)$ given by 6×10^{214} times

$$\binom{27706}{30} \quad \approx \quad 7 \times 10^{100}.$$

This is approximately 4×10^{315}. This calculation however may have some tables that are counted more than once because two distinct sequences of moves starting from Table 1 can lead to the same table. This overcounting should be more than compensated by tables that could be reached with more or fewer than 30 moves. In conclusion we estimate that a lower bound of 4×10^{315} possible tables is justified, but to be on the safe side we adopt 10^{300} as a very conservative estimation of the possible size of $\mathcal{F}(u)$.

Now we provide a very liberal upper bound on the number of tables whose χ^2 value is smaller than the χ^2 value of the data table (346.63). The volume V of the ellipsoid defined by equation (3) is bounded above by 1.1×10^{266} using the volume formula for multidimensional ellipsoids. It is known that the number of lattice points in an n-dimensional ellipsoid defined by

$$\sum_{1 \le j < k \le n} a_{jk} x_{jk} \quad \le \quad r^2$$

is approximately equal to $V + O(r^{n/2})$ [32] and in our case $r^2 = 346.63$ and $n = 231$, and we arrive at 1.8×10^{293} as an upper bound on the number of tables which are inside the ellipsoid defined by (3). We note that this must be a gross estimate, since we should be counting the tables in the ellipsoid that are also in the fiber $\mathcal{F}(u)$. For this we should be counting the lattice points in another ellipsoid of lower dimension 209. We conclude that a very conservative estimate of proportion of tables F where $\chi^2(F) \le \chi^2(f) = 346.63$ in $\mathcal{F}(u)$ is extremely small:

$$\le \frac{1.8 \times 10^{293}}{10^{300}} \approx 1.8 \times 10^{-7}.$$

We believe that the true proportion is much smaller.

References

[1] H. Abdi. Bonferroni and šidák corrections for multiple comparisons. In N. J. Salkind, editor, *Encyclopedia of Measurement and Statistics*, pages 598 – 605, 2007.

[2] J. Arsuaga, K. M. Greulich-Bode, M. Vazquez, M. Bruckner, P. Hahnfeldt, D. J. Brenner, R. K. Sachs and L. Hlatky. Chromosome spatial clustering inferred from radio-

genic aberrations. *International Journal of Radiation Biology* , 80(7): 507 – 515, 2004.

[3] V. Baldoni, N. Berline, J. A. De Loera, B. Dutra, M. Köppe, G. Pinto, M. Vergne, and J. Wu. *A User's Guide for* `Latte integrale` *v1.7.1*, 2013, software package *Latte* is available at `http://www.math.ucdavis.edu/~latte/`.

[4] M. Barbieri, M. Chotalia, J. Fraser, L. M. Lavitas,J. Dostie, A. Pombo, and M. Nicodemi. Complexity of chromatin folding is captured by the strings and binders switch model. *Proc Natl Acad Sci U S A*, 109(40): 16173-16178, 2012.

[5] H. Ben-Bassat, Z. Shlomai, G. Kohn, and M. Prokocimer. Establishment of a human T-acute lymphoblastic leukemia cell line with a (16;20) chromosome translocation. *Cancer Genet Cytogenet*, 49(2): 241–248, 1990.

[6] W. A. Bickmore and P. Teague. Influences of chromosome size, gene density and nuclear position on the frequency of constitutional translocations in the human population. *Chromosome Res*, 10: 707–715, 2002.

[7] W. A. Bickmore and B. van Steensel. Genome architecture: domain organization of interphase chromosomes. *Cell.* 152(6): 1270–1284, 2013

[8] T. Blackstone ,R. Scharein, B. Borgo, R. Varela, Y. Diao , and J. Arsuaga. Modeling of chromosome intermingling by partially overlapping uniform random polygons. *J Math Biol*, 62(3): 371–389, 2011.

[9] J. J. Boei, J. Fomina, F. Darroudi, N. J. Nagelkerke, L. H. Mullenders. Interphase chromosome positioning affects the spectrum of radiation-induced chromosomal aberrations. *Radiat. Res.*, 166(2): 319–326, 2006.

[10] D. Bossi, F. Carlomagno, I. Pallavicini, G. Pruneri, M. Trubia, P. R. Raviele, A. Marinelli, S. Anaganti, M. C. Cox, G. Viale, M. Santoro, P. P. Di Fiore, and S. Minucci. Functional characterization of a novel FGFR1OP-RET rearrangement in hematopoietic malignancies. *Mol Oncol.*, 8(2): 221–231, 2014.

[11] S. Boyle, S. Gilchrist, J. M. Bridger, N. L. Mahy, J. A. Ellis, and W. A. Bickmore. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Human Molecular Genetics*, 10(3): 211–219, 2001.

[12] M. R. Branco and A. Pombo. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS Biol.*, 4(5): 780–788, 2006.

[13] A. M. Chen, J. N. Lucas, F. S. Hill, D. J. Brenner, and R. K. Sachs. Proximity effects for chromosome aberrations measured by FISH. *Int. J. Radiat. Biol.*, 69: 411–420, 1996.

[14] R. Christensen. *Log-linear models and logistic regression* 2nd ed. Springer-Verlag, New York, 1997.

[15] M. N. Cornforth, K. M. Greulich-Bode, B. D. Loucas, J. Arsuaga, M. Vázquez, R. K. Sachs, M. Brückner, M. Molls, P. Hahnfeldt, L. Hlatky, and D. J. Brenner. Chromosomes are predominantly located randomly with respect to each other in interphase human cells. *Journal of Cell Biology*, 159(2): 237–244, 2002.

[16] T. Cremer, C. Cremer, H. Baumann, E. K. Luedtke, K. Sperling, V. Teuber, and C. Zorn. Rabl's model of the interphase chromosome arrangement tested in Chinise hamster cells by premature chromosome condensation and laser-UV-microbeam experiment. *Human Genetics*, 60(1): 46–56, 1982.

[17] C. Cremer, C. Münkel, M. Granzowd et al. Nuclear architecture and the induction of chromosomal aberrations. *Mutation Research/Reviews in Genetic Toxicology*, 366(2): 97-116, 1996.

[18] T. Cremer and M. Cremer. Chromosome terri'tories. *Cold Spring Harb Perspect Biol.*, 2(3): 1–22, 2006.

[19] J. N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *Annals of Mathematical Statistics*, 43(5): 1470–1480, 1972.

[20] J. A. De Loera, B. Sturmfels and R. Thomas. Gröbner bases and triangulations of the second hypersimplex. *Combinatorica*, 15: 409–424, 1995.

[21] P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distributions. *Annals of Statistics*, 26(1): 363–397, 1998.

[22] M. Drton, B. Sturmfels and S. Sullivant. *Lectures on Algebraic Statistics*, Oberwolfach Seminars, vol. 39, Birkhäuser Basel, 2009.

[23] J. D. Halverson, J. Smrek, K. Kremer, A. Y. Grosberg. From a melt of rings to chromosome territories: the role of topological constraints in genome folding. *Rep Prog Phys.*, 77 (2): 022601, 2014.

[24] I. Heskia. *Testing proximity effect hypothesis for chromosomes using algebraic statistics*. Master's thesis, San Francisco State University, 2010.

[25] L. Hlatky, R. K. Sachs, M. Vazquez. and M. N. Cornforth. Radiation-induced chromosome aberrations: Insights gained from biophysical modeling. *BioEssays*, 24(8): 714–723, 2002.

[26] L. Impera, A. Lonoce, D. A. Fanfulla, C. Moreilhon, L. Legros, S. Raynaud and C. T. Storlazzi. Two alternatively spliced 5'BCR/3'JAK2 fusion transcripts in a myeloproliferative neoplasm with a three-way t(9;18;22)(p23;p11.3;q11.2) translocation. *Cancer Genet.*, 204(9): 512–515, 2011.

[27] M. Jarošová, P. Rohoň, J. Živn, S. Peková, R. Nedomová, M. Holzerová, P. Mičková, S. Reptová, T. Papajk, and K. Indrák. Pathogenetic role of ETV6 fusion gene in leukemic transformation of myelodysplastic syndrome refractory anemia with excess blasts-1 with a new, rare translocation t(11;19)(q24.3;q13.12) and insertion ins(6;12)(p22.3p13). *Leuk Lymphoma.*, 55(4): 950–953, 2014.

[28] A. Khalil, J. L. Grant, L. B. Caddle, E. Atzema, K. D. Mills, and A. Arneodo. Chromosome territories have a highly nonspherical morphology and nonrandom positioning. *Chromosome Res.*, 15(7): 899–916, 2007.

[29] A. C. Kolbl, D. Weigl, M. Mulaw, T. Thormeyer, S. K. Bohlander, T. Cremer, and S. Dietzel. The radial nuclear positioning of genes correlates with features of megabase-sized chromatin domains. *Chromosome Res.*, 20(6): 735–752, 2012.

[30] G. Kreth, J. Finsterle, J. von Hase, M. Cremer, and C. Cremer. Radial arrangement of chromosome territories in human cell nuclei: a computer model approach based on gene density indicates a probabilistic global positioning code. *Biophysical J.*, 86(5): 2803–2812, 2004.

[31] S. Kuhnt, F. Rapallo, and A. Rehage. Outlier detection in contingency tables based on minimal patterns. *Statistics and Computing*, 24(3): 481–491, 2014.

[32] E. Landau. Über Gitterpunkte in mehrdimensionalen Ellipsoiden. *Mathematische Zeitschrift*, 21(4): 126–132, 1924.

[33] E. Lieberman-Aiden ,N. L. van Berkum, L. Williams, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950): 289–293, 2009.

[34] B. D. Loucas and M.N. Cornforth. Complex chromosome exchanges induced by gamma rays in human lymphocytes: An mFISH study. *Radiation Research*, 155(5): 660–671, 2001.

[35] T. Nagano, Y. Lubling, T. J. Stevens, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59-64, 2013.

[36] R. Naeem R, S. Singer, and J. A. Fletcher.Translocation t(8;13)(p11;q11-12) in stem cell leukemia/lymphoma of T-cell and myeloid lineages. *Genes Chromosomes Cancer*, 12(2): 148–151, 1995.

[37] M. N. Nikiforova, J. R. Stringer, R. Blough, M. Medvedovic, J. A. Fagin, and Y. E. Nikiforov. Proximity of Chromosomal Loci That Participate in Radiation-Induced Rearrangements in Human Cells. *Science*, 290: 138–141, 2000.

[38] L. A. Parada and T. Misteli. Chromosome positioning in the interphase nucleus. *Trends in Cell Biology*, 12(9): 425 – 432, 2002.

[39] C. Rabl. Über Zellteilung. *Morphologisches Jahrbuch*, 10, 1885.‘sxx

[40] F. Rapallo. Outliers and patterns of outliers in contingency tables with algebraic statistics. *Scandinavian Journal of Statistics*, 39(4): 784–797, 2012.

[41] J. J. Roix, P. G. McQueen, P. J. Munson, L. A. Parada, and T. Misteli. Spatial proximity of translocation-prone gene loci in human lymphomas. *Nat Genet.*, 34(3): 287–291, 2003.

[42] A. Rosa, and R. Everaers. Structure and dynamics of interphase chromosomes. *PLoS Comput Biol.*, 4(8): 2008.

[43] R K Sachs, D J Brenner, and A M Chen. Review: proximity effects in the production of chromosome aberrations by ionizing radiation. *International Journal of Radiation Biology*, 71(1): 1–19, 1997.

[44] R. K. Sachs, D. Levy, P. Hahnfeldt, and L. Hlatky. Quantitative analysis of radiation-induced chromosome aberrations. *Cytogenetic and Genome Research*, 104(1-4): 142 – 148, 2004.

[45] E. Schröck, S. du Manoir, T. Veldman, B. Schoell, J. Wienberg, M. A. Ferguson-Smith, Y. Ning, D. H. Ledbetter, I. Bar-Am, D. Soenksen, Y. Garini, and T. Ried. Multicolor spectral karyotyping of human chromosomes. *Science* 273(5274): 494–497, 2006.

[46] M. R. Speicher, S. G . Ballard, and D. C. Ward. Karyotyping human chromosomes by combinatorial multi-fluor FISH. *Nature Genetics*, 12: 368–375, 1996.

[47] B. Sturmfels. *Gröbner Bases and Convex Polytopes*, University Lectures Series, American Mathematics Society, 1996.

[48] C. Uhler and Stephen J. Wright. Packing ellipsoids with overlap. *SIAM Review*, *to appear.*

[49] S. Vives, B. Loucas, M. Vazquez, D. J. Brenner, R. K. Sachs, L. Hlatky, M. Cornforth, and J. Arsuaga. SCHIP: Statistics for chromosome interphase positioning based on interchange data. *Bioinformatics*, 21(14): 3181–3182, 2005.

[50] Y. Zhang Y., R. P. McCord, Y. J. Ho et al. Spatial oganization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* 148: 908–921, 2012.

[51] H. Wu. Probabilities of radiation-induced inter- and intrachromosomal exchanges and their dependence on the DNA content of the chromosome. *Radiat Res.*, 156: 603-606, 2001.

[52] W. Zhu, Z. Sun, Z. Zhai, K. Ding, and G. Wu. A novel t(3;19)(p21;p13) in a patient with acute myelocytic leukemia. *Cancer Genet Cytogenet.*, 179(2): 165–166, 2007.