BMC Evolutionary Biology

BioMed Central

Research article

A scale invariant clustering of genes on human chromosome 7

Wayne S Kendal*

Address: Department of Radiation Oncology, Ottawa Regional Cancer Centre, 503 Smyth, Ottawa, Ontario K1H 1C4, Canada

Email: Wayne S Kendal* - wayne.kendal@orcc.on.ca

* Corresponding author

Published: 30 January 2004

BMC Evolutionary Biology 2004, **4**:3

Received: 27 October 2003

Accepted: 30 January 2004

This article is available from: http://www.biomedcentral.com/1471-2148/4/3

Abstract

Background: Vertebrate genes often appear to cluster within the background of nontranscribed genomic DNA. Here an analysis of the physical distribution of gene structures on human chromosome 7 was performed to confirm the presence of clustering, and to elucidate possible underlying statistical and biological mechanisms.

Results: Clustering of genes was confirmed by virtue of a variance of the number of genes per unit physical length that exceeded the respective mean. Further evidence for clustering came from a power function relationship between the variance and mean that possessed an exponent of 1.51. This power function implied that the spatial distribution of genes on chromosome 7 was scale invariant, and that the underlying statistical distribution had a Poisson-gamma (PG) form. A PG distribution for the spatial scattering of genes was validated by stringent comparisons of both the predicted variance to mean power function and its cumulative distribution function to data derived from chromosome 7.

Conclusion: The PG distribution was consistent with at least two different biological models: In the *microrearrangement model*, the number of genes per unit length of chromosome represented the contribution of a random number of smaller chromosomal segments that had originated by random breakage and reconstruction of more primitive chromosomes. Each of these smaller segments would have necessarily contained (on average) a gamma distributed number of genes.

In the *gene cluster model*, genes would be scattered randomly to begin with. Over evolutionary timescales, tandem duplication, mutation, insertion, deletion and rearrangement could act at these gene sites through a stochastic birth death and immigration process to yield a PG distribution.

On the basis of the gene position data alone it was not possible to identify the biological model which best explained the observed clustering. However, the underlying PG statistical model implicated neutral evolutionary mechanisms as the basis for this clustering.

Background

Over twenty years ago Susumu Ohno postulated that gene duplication should play a major role in genomic evolution and that, consequent to eons of mutation, insertion and deletion, any surviving genes would be scattered throughout deserts of nontranscribed DNA [1]. Now, with the fruition of the Human Genome Project, his postulate could be comprehensively examined and validated [2].

The distribution of genes has more to reveal than just this scattering. Data provided from the Chromosome 7 Annotation Project [3] has revealed a highly heterogeneous

density of genes within the physical length of human chromosome 7, suggestive of non-random clustering. As well, syntenic regions within the murine genome exhibited parallel disparities in gene density, a finding that would suggest at least part of this heterogeneity could have originated through evolutionary mechanisms [3]. To confirm whether or not clustering was evident and to determine at what physical scale(s) the clustering might manifest, a quantitative analysis was required. The object here was to perform such an analysis with the intent to identify statistical and biological mechanisms that might explain the spatial distribution of genes.

The Human Genome Project has already provided an analysis of the distribution of genes within conserved segments of the human genome that appeared to be consistent with a random breakage model for chromosomal rearrangements (see Figs. 47 and 48 from [4]). This analysis was based on the assumption that differences in gene density between conserved segments could be ignored, at least to some extent. In light of the significant differences in gene density apparent to chromosome 7 [3], it seemed appropriate to question this assumption. Furthermore, an understanding of statistical rules governing the spatial distribution of genes along chromosomes should permit a more critical analysis of the biological mechanisms that might be responsible for these disparities.

An analysis of the distribution of gene structures along the physical length of a chromosome will be presented here, based upon the available data from human chromosome 7 [3]. This analysis will confirm the presence of gene clustering within chromosome 7, and the clustering will be shown to manifest over a range of measurement scales. Based on these findings, a statistical model for the spatial distribution of genes within chromosome 7 will be proposed and tested. This statistical behaviour will be interpreted in the context of two different biological models, predicated upon the evolution of either microrearrangements or gene clusters.

Results

A scale invariant clustering of genes

The Chromosome 7 Annotation Project demonstrated that the density of gene structures within the chromosome was heterogeneous [3]. Figure 1 provides the numbers of such structures contained within a sequence of equal-sized non-overlapping bins that spanned the physical length of chromosome 7. The high degree of heterogeneity in local gene density as demonstrated here seemed, at least on a qualitative level, to indicate clustering.

A quantitative test for clustering was performed upon these data. If the genes were randomly scattered throughout the chromosome, without clustering, their dispersal should reflect a Poisson distribution – the usual model for such randomness. To determine whether or not this was the case, the variance var(Z) and the mean E(Z) of the number of genes per bin, Z, were compared. At the scale of the 200 kb bins the variance and mean were, var(Z) = 6.6 and E(Z) = 2.3, with a variance/mean ratio of approximately 2.9. The variance should have equalled the mean with a Poisson distribution. This finding indicated that the genes were more dispersed than could be predicted by a random distribution, and thus there was clustering at this scale.

To determine whether this clustering persisted at other measurement scales, the variance and mean number of gene structures per bin were estimated for a range of bin sizes. Figure 2 provides these data on a log-log plot of variance versus mean. The logarithmically transformed points seemed to describe a linear relationship. Indeed the correlation coefficient squared, estimated between the transformed variance and mean estimates, was $r^2 = 0.997$ thus substantiating a linear relationship. As well, the residuals between the logarithmically transformed variables and a trial linear relationship were essentially negligible and normally distributed about zero (Fig. 2 insert). It should be mentioned that the linear relationship tested here against the data in Fig. 2 was obtained not from the regression fit of the logarithmically transformed data, but from a statistical model that was fitted to the chromosome 7 data and that will be presented later in this article.

The strong linear relationship between the logarithmically transformed variances and means indicated that the variance and the mean were related by a power function,

$$\mathrm{var}(Z) = a \cdot \mathrm{E}(Z)^p,$$

where a and p were constants. If the gene structures had been randomly distributed along chromosome 7, as per a Poisson distribution, one would have expected an exponent $p = 1$. Since $p = 1.51$, this provided further confirmation of clustering, which now was evident over a range of measurement scales.

This variance to mean power function exhibited another property of note – *scale invariance*. This term, as used here, indicates that if one takes a small segment of a pattern and magnifies it to a larger scale, then the magnified portion should be statistically similar to the unmagnified portion. Specifically, if the measurement scale employed in the variance to mean relationship is increased by a factor c then

$$a \cdot [c \cdot \mathrm{E}(Z)]^p = (ac^p) \cdot [\mathrm{E}(Z)]^p,$$

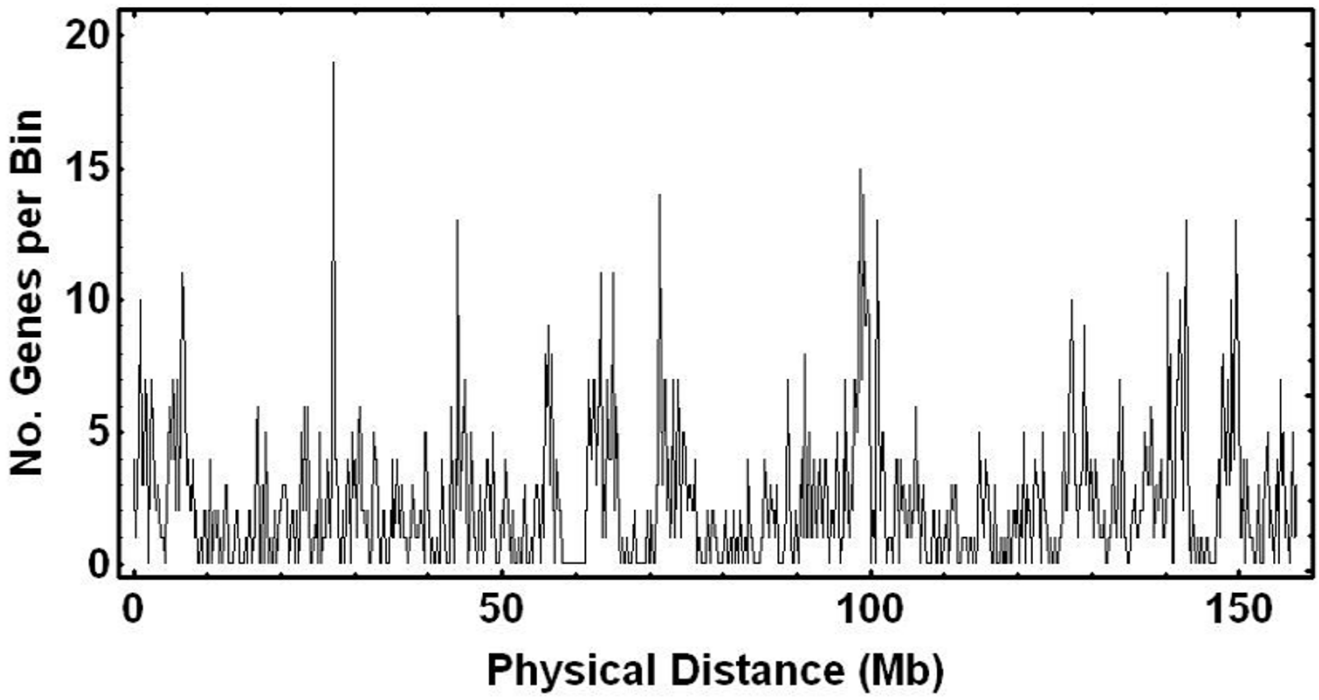

Figure 1
Gene density along human chromosome 7. The numbers of gene structures enumerated within a sequence of nonoverlapping 200 kb bins that spanned the physical length of chromosome 7 are plotted here. The density of genes appeared quite heterogeneous along the length of the chromosome.

and this relationship would retain the form of a power function with exponent *p*. In fact, this variance to mean power function is the only possible scale invariant relationship that could exist between the variance and the mean [5]. As will be seen below the variance to mean power function implicated a specific probabilistic model to represent the distribution of gene structures along chromosome 7. Before we can discuss this model, it would be useful to consider a somewhat more conventional model, for the distribution of the number of genes within chromosomal segments.

An overdispersed Poisson distribution for conserved segments

Earlier in this article the distribution of genes within conserved segments was alluded to in the context of observations provided by the Human Genome Project [4]. Conserved segments are conventionally identified on the basis of the relative order of contiguous landmarks between the chromosomes of two different species. If one were to define the enumerative bins employed here on the basis of the limits of individual conserved segments, one might expect genes to be randomly distributed within these bins, in accordance with a Poisson distribution [6]. One might also expect some heterogeneity between different conserved segments, such that the mean number of genes per conserved segment would depend upon both the physical length of the segment and upon the local gene density. The distribution of the mean number of genes per conserved segment has been conventionally represented by a gamma distribution, giving an overdispersed Poisson distribution (i.e., a negative binomial distribution) for the actual number of genes per conserved segment [6].

How would a negative binomial distribution affect the variance to mean relationship as plotted in Fig. 2? With some simple calculation we have,

$$\mathrm{var}(Z) = \frac{1}{\xi}\mathrm{E}(Z)^2 + \mathrm{E}(Z),$$

where ξ is a constant. In Fig. 2 the optimised least squares fit of this additional variance to mean relationship was plotted as a broken line (with $\xi = 3.12$). Larger deviations

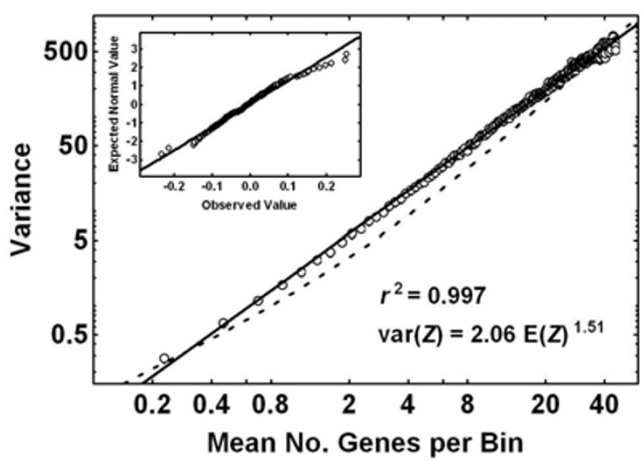

Figure 2
Variance to mean power function. Shown here is a log-log plot of the variance versus the mean number of gene structures per bin, as calculated for a range of bin sizes over chromosome 7. The transformed data points described a straight line on the log-log plot, which implied a power function relationship between the variance and the mean. The solid line represents the theoretical linear relationship determined from the fit of the PG model. A linear model fitted very well to these transformed data as evident from the high value for the correlation coefficient squared r^2, and the normal probability plot of the residuals (insert) derived from the differences between the theoretical straight line and the transformed data points. The broken line represents the best fit of a second model, intended for the distribution of genes within conserved segments that was based upon the negative binomial distribution. It did not fit the data as well as did the variance to mean power function.

were seen between this relationship and the data, relative to those seen with the variance to mean power function.

This discrepancy between the negative binomial model and the variance to mean data was probably to be expected. In the analysis presented here, the estimates for the variance and the mean had been calculated from bins of uniform size rather than from individual conserved segments, as the negative binomial model would properly have required. These uniformly sized bins would presumably have contained variable numbers of conserved segments and, since the convolution of a random number of negative binomial distributions would not result in a negative binomial distribution, a negative binomial model would have been inappropriate. In the next section, a modification of this model will be presented that will account for a variable number of conserved segments per bin.

A scale invariant Poisson-gamma model for the number of gene structures per bin

One view of chromosomal structure represents chromosomes as mosaics formed from the random fragmentation and rearrangement of more primitive chromosomes [1,7]. Possibly then, the number of genes within unit segments of vertebrate chromosomes might reflect this segmental structure. A modified model for gene distribution might thus have to account for the contribution of multiple genomic segments that would individually exhibit statistical behaviour related to the conventional gamma model for the mean number of genes.

In the initial specification of such a statistical model one might stipulate that the model be made as general as possible. The generalized linear models of Nelder and Wedderburn [8] provide a simple method to analyse a large variety of data, and these models can be further generalized into an even wider class of models called exponential dispersion models [5]. These latter models provide descriptions for a comprehensive range of normal and non-normal distributions that include the Poisson, gamma and Gaussian distributions. The reader is encouraged to refer to an excellent introduction to these models in the monograph provided by Jørgensen [5].

If one accepts the premise that some form of exponential dispersion model might describe the distribution of the number of genes per bin then, consequent to the finding of the variance to mean power function, one would be lead to a statistical model that uniquely exhibits such a power function, where its exponent p is constrained to range between the values of 1 and 2. This particular model is described by a scale invariant Poisson-gamma (PG) distribution for which its additive form has the cumulant generating function $K^*(s)$ [5],

$$K^*(s) = \lambda\kappa(\theta)\left\{\left(1+\frac{s}{\theta}\right)^{\alpha} - 1\right\}.$$

Here s is a variable used to base the generating function upon, λ is the index parameter, θ is the canonical parameter, α is a parameter related to the power function exponent such that

$$\alpha = \frac{p-2}{p-1},$$

and the cumulant function $\kappa(\theta)$ is given by,

$$\kappa(\theta) = \frac{\alpha-1}{\alpha}\left(\frac{\theta}{\alpha-1}\right)^{\alpha}.$$

The variance to mean power function,

$\mathrm{var}(Z) = \lambda^{1/(\alpha-1)} \cdot \mathrm{E}(Z)^p$,

follows as a consequence of this cumulant generating function.

There have been many alternative explanations for the variance to mean power function, some that represented approximations and others exact models [reviewed in [9]]. What would favour the choice of an exponential dispersion model over any of these alternatives? Much of the justification goes back to the theory of errors as developed by Gauss and others [5]. The Gaussian distribution, which has been widely applied to describe measurement error, provides a familiar description for errors of small magnitude. However, there exist processes with larger non-Gaussian errors which require a more generalized theory, such as the fluctuations apparent to the chromosome 7 data. Nelder and Wedderburn [8], with their generalized linear models, provided a simple means to analyze a large range of non-Gaussian data; exponential dispersion models represent an extension to their theory.

The utility of the exponential dispersion models becomes most apparent when one considers a class of these models characterized by the variance to mean power function. These Tweedie models, named after M.C.K. Tweedie who first studied them [10], serve as limiting distributions for a wider range of exponential dispersion models [5]. Much as the Gaussian model serves as a limiting distribution for a range of statistical processes, the Tweedie models, which include the Gaussian distribution as a special case, represent limiting distributions for a range of non-Gaussian processes. True, one may employ one of the alternative models to account for the variance to mean power function, but the burden would be then to develop a theory for the specific case which nonetheless would lack the generality offered by the exponential dispersion models. Moreover, such an alternative model, as indicated by the Tweedie convergence theorem, might be approximated by a Tweedie model. For these reasons the more general approach, allowed by the theory of exponential dispersion models, seems appropriate.

Granted these considerations in favour of exponential dispersion models, let us consider the PG distribution in more detail. It is difficult to describe the probability density function and its corresponding cumulative distribution function (CDF) for this distribution, since these expressions do not exist in closed form [5]. The probability density function $p^*(z; \theta,\lambda,\alpha)$ can be expressed in terms of the canonical statistic z such that,

$p^*(z; \theta, \lambda, \alpha) = c^*(z; \lambda) \cdot \exp[\theta \cdot z - \lambda\kappa(\theta)]$.

where

$$c^*(z;\lambda) = \begin{cases} \frac{1}{z}\sum_{n=1}^{\infty} \lambda^n \cdot \left[\left(\frac{\alpha-1}{\alpha}\right)\cdot\left(\frac{-1}{(\alpha-1)\cdot z}\right)^{\alpha}\right]^n / \left[\Gamma(-\alpha\cdot n)\cdot n!\right] & \text{for } z > 0, \text{ and} \\ 1 & \text{for } z = 0. \end{cases}$$

The CDF $P^*(z; \theta,\lambda,\alpha)$ can then be expressed:

$$P^*(z;\theta,\lambda,\alpha) = e^{-\lambda\cdot\kappa(\theta)} + \int_0^z p^*(y;\theta,\lambda,\alpha)\cdot dy. \qquad (1)$$

How well does the PG distribution fit the observed data? Figure 3 provides the empirical CDF, as obtained from a bin size of 200 kb and fitted to the theoretical PG CDF (Eq. 1). The fit was very good, with at most a 1.4% deviation between theory and observation. An analysis of the residuals (Fig. 3 insert) revealed that they were essentially negligible and normally distributed about zero. A Kolmogorov Smirnov test additionally confirmed an acceptable fit of the theoretical PG model to the empirical CDF.

Three parameters were derived from the regression of the PG CDF (Eq. 1) to the empirical CDF: α = -0.952, λ = 0.245 and θ = -0.691. These were the parameters employed to provide the theoretical variance to mean power function given in Fig. 2, with $p = (\alpha - 2)/(\alpha - 1) = 1.51$ and $a = \lambda^{1/(\alpha-1)} = 2.06$, and for which the agreement with the chromosome 7 data was also very good ($\chi^2 = 0.231$, $d.f. = 200$, $P = 1$). Thus two different tests of the PG distribution were provided here to confirm its agreement with the chromosome 7 data: the fit of the CDF and the fit of the variance to mean power function to the chromosome 7 data.

Discussion

The microrearrangement model

The PG distribution provides a statistical model that accurately describes the spatial distribution of genes within chromosome 7. A hypothesis was alluded to in the last section whereby the number of genes per bin could be represented as the summed contribution from a random (Poisson distributed) number of chromosomal segments, each with identical and independent gamma-distributed numbers of genes. However, the agreement of the statistical model with the data does not necessarily imply that the hypothesis used to interpret this model is correct. This hypothesis therefore deserves further scrutiny.

Modern chromosomes are thought to represent mosaics of genomic segments laid in sequence and drawn from more ancient chromosomes [1,7]. One might then expect a random number of such segments to be joined together within each of the equal-sized enumerative bins. These segments presumably would represent changes that mostly predated those rearrangements defined by the

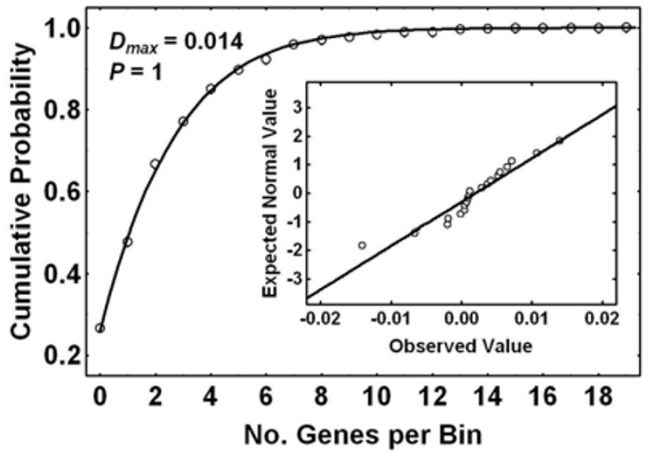

Figure 3
Cumulative distribution function. The empirical CDF, derived from the numbers of genes per 200 kb bins within chromosome 7, was plotted here as data points. The corresponding solid curve represents the least squares fit of the PG model to these data. The PG model fitted very well to these data as evident from the low value for the Kolmogorov Smirnov D_{max}, and the normal probability plot of the residuals (insert).

conserved synteny between related individual species, and so they will be distinguished here by the term, *primitive segments*. The number of genes per primitive segment would be distributed, on average, in accordance to a gamma distribution, as stipulated in the conventional model [6]. Their summed contributions would furthermore be required to obey a PG distribution.

Let us look more closely at this model: Because individual genes may have sizeable lengths, they might straddle the arbitrary boundaries of the enumerative bins. In order to deal with this possibility, the genes were enumerated on the basis of the positions of their *p*-termini. One could have defined the gene positions by their *q*-termini or by their transcriptional start positions. Indeed the variance to mean power function was evident with both of these alternative definitions (data not provided).

Likewise, the packing of the primitive segments within the enumerative bins should be considered. These segments conceivably could also straddle the boundaries of the bins. The PG model, by virtue of its constituent gamma distribution, represents some degree of averaging and it should be sufficiently robust to account for minor discrepancies.

One further consideration regarding the packing of primitive segments remains. If we assume a random breakage model for chromosomal rearrangements, then the breakpoints on the ancestral chromosome would be distributed according to a Poisson process, and the lengths of the resultant segments would be exponentially distributed [7]. With rearrangement there would presumably be a random redistribution of the breakpoint positions, but these new points should be also distributed according to a Poisson distribution. Over evolutionary time periods additional rearrangements would be expected to accumulate. The eventual distribution of rearranged breakpoints would also be expected to obey a Poisson distribution, and the segment lengths would be exponentially distributed. The point here is that the assumption of a Poisson distributed number of primitive segments per enumerative bin is largely predicated upon a random breakage model for chromosomal rearrangement. If for some reason the lengths of the primitive segments were not exponentially distributed, then these segments would not readily pack together within the enumerative bins in accordance with a Poisson distribution. To some extent this requirement for random breakage could be relaxed, given averaging and the possibility that added indels might permit such a random packing of primitive segments within the bins.

We thus have a model which provides the local distribution of genes at scales larger than that of the primitive segments. This model is applicable to a range of bin sizes, and thus it is inherently scale invariant. The variance to mean power function seen in Fig. 2 would be a direct consequence of this model.

The parameters obtained from the regression fit of the PG CDF could be used to estimate the mean number of primitive segments per bin in chromosome 7, using the expression $\lambda \cdot \kappa\ (\theta) = 1.35$. Since 790 sequential bins, each 200 kb in length, spanned the 158 Mb of chromosome 7 this would imply that the total number of primitive segments within chromosome 7 would be approximately 1.1×10^3.

Pevzner and Tesler estimated that the human and mouse genomes share 281 synteny blocks of at least 1 Mb length [11]. They noted 3170 additional microrearrangements within these synteny blocks, although they conceded that many of these microrearrangements could have represented artefact. In a separate analysis Kumar *et al*. have estimated that there exist 529 conserved segments between the human and mouse genomes [6]. The relatively larger estimate for the number of primitive segments obtained here might be interpreted to indicate that either these putative primitive segments had their origin long before the divergence in evolution between humans and mice, or that many of smaller rearrangements

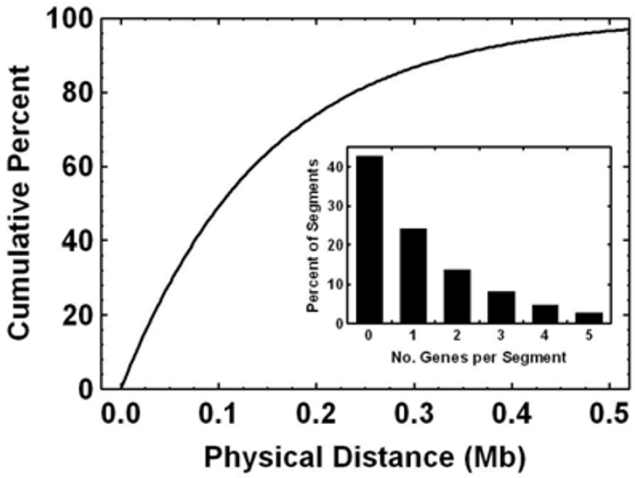

Figure 4
Predicted spacing of the segments within chromosome 7. A CDF for the physical distances between the *p*-termini of the primitive segments of the microrearrangement model is presented here, on the basis of the assumption of an underlying exponential distribution, and the parameters derived from the best fit of the PG CDF to the chromosome 7 data. Granted these assumptions, about 50% of the primitive segments should be separated by distances of 100 kb size or less. If the amounts of intervening DNA between adjacent primitive segments could be considered negligible, then this plot would correspond to the size distribution of the primitive segments. Alternatively, under the gene cluster model, this CDF would correspond to the physical distances between gene cluster sites. Insert: Frequency Histogram for the Number of Genes per Primitive Segment. The parameters provided from the PG model were used to estimate the frequency distribution of genes within the primitive segments of the microrearrangement model. More than 40% of the primitive segments would be expected to contain no recognizable gene structure, and somewhat more than 20% of segments would contain only one gene. Under the alternative gene cluster model, this histogram would represent the frequencies of the number of genes per cluster.

presumed by Pevzner and Tesler to be artefacts [11] were indeed real.

The parameters obtained from the fit of the PG distribution allowed some additional predictions regarding these primitive segments. According to the model, one would expect a Poisson distributed number of segments to be contained within each bin. If the primitive segments were distributed randomly within the bins, the distances between the *p*-termini of the primitive segments would be separated according to an exponential distribution. The CDF for this exponential distribution would be,

$$P_{\exp}(x) = 1 - \exp[\lambda \cdot \kappa(\theta) \cdot x/\Delta x],\ x \geq 0,$$

where x is the distance between the segments and Δx is the bin size. Figure 4 provides the predicted CDF for the physical spacing between the *p* termini of the primitive segments. Here we see that about half of the *p* termini were spaced at least 100 kb apart. If one assumed that there was no intervening DNA between the boundaries of the primitive segments, then these distances predicted would correspond to the lengths of the segments, and average segment length would be about $200/\lambda \cdot \kappa\ (\theta) = 150$ kb.

As well, the numbers of gene structures per primitive segment would, on discrete analysis, approximate a negative binomial distribution with the probability density,

$$p_{nb} = \frac{\Gamma(k-\alpha)}{k! \cdot \Gamma(-\alpha)} \cdot \left[\frac{1}{1-\theta}\right]^{k} \cdot \left[\frac{-\theta}{1-\theta}\right]^{-\alpha},\ k = 0,\ 1,\ 2,\ldots$$

Somewhat over 40% of segments would thus be predicted to contain no evident gene structures, and about 25% would contain only one gene (Fig 4, insert).

Two concerns regarding the microrearrangement hypothesis can be raised: If we were to assume that microrearrangements occurred at a similar frequency over the entire 3,200 Mb [4] of the human genome then we could predict a total of 2.2×10^4 such rearrangements. If the ancestral sequences had their origins with the first demonstrable microorganisms (about 3.6×10^9 yr ago [12]), and assuming that each rearrangement produced 2 primitive segments [6], this would imply about 3 rearrangements/Myr. This rate is higher than the 1.5 rearrangements/Myr estimated between the divergence of the murine and human genomes [11]. Whether this discrepancy could be explained by microrearrangements that had otherwise been dismissed as artefact [11], or by higher rates for rearrangement along other regions of the phylogenetic tree, is presently unclear.

The second concern is similarly difficult to dismiss. There is some evidence that intron and protein lengths are correlated with intergeneic distances [13]. If chromosomal rearrangements were to occur within the genome on average at about every 150 kb, as predicted from the microrearrangement hypothesis, then this relationship should not be apparent. In view of these concerns one may ask whether the agreement between the PG distribution and the chromosome 7 data could be explained by another biological model. This possibility will be considered in the next section.

The gene cluster model

The PG distribution might be attributed to mechanisms involved with the evolution of gene clusters. Under this

hypothesis each enumerative bin would contain a Poisson distributed number of gene cluster sites, and the number of genes per site would (on average) be distributed according to a gamma distribution.

The cluster sites would presumably represent the positions of ancestral genes that originally had been scattered randomly within the ancestral genome. Over evolutionary time periods the combined action of tandem duplication, mutation, large rearrangements, and indels might give rise to gene clusters. These events might be described by a stochastic birth death and immigration (BDI) process [l4]. Under this assumption, both tandem duplication and gene loss would necessarily occur at rates proportional to the number of genes within each cluster, and the introduction of new genes to the clusters through major rearrangements, would occur at a constant rate per unit time. The equilibrium distribution for such a BDI process would approximate a negative binomial form, for which the continuous equivalent is a gamma distribution.

A BDI process has a number of parameters. Let ν, β, and μ represent the birth, immigration and death rates, respectively. A finite equilibrium size for the gene clusters would require that $\nu < \mu$; otherwise with $\nu > \mu$ the cluster size would eventually become infinite. At equilibrium, the mean cluster size would be $\beta / (\mu - \nu)$, and the PG parameters α and θ would relate to the BDI rates by the equations, $\alpha = -\beta/\nu$ and $\theta = (\nu - \mu)/\nu$.

The average physical distance between cluster sites would be $200/\lambda \cdot \kappa(\theta) = 150$ kb; Fig 4 would now give the CDF for the physical spacing between these sites. The insert within Fig. 4 would describe the frequency distribution for the cluster sizes at equilibrium: over 40% of the cluster sites would contain no residual gene, about 25% would contain only one gene, and the mean cluster size would be 1.4 genes. These estimates would indicate that the primary mechanism responsible for the apparent clustering of genes would be the loss of genes. The sites where ancestral genes had been obliterated would appear as expanses of noncoding DNA between the remaining genes and gene clusters, as Ohno had originally postulated [1].

What could be the main genetic mechanisms operative in the BDI model? Gene duplication and rearrangements (that introduce new genes into cluster sites) would be necessary for equilibrium to evolve. Some insight into a major mechanism responsible for gene loss comes from a comparative analysis of the human and murine genomes where about 60% of the human genome could not be aligned to its murine analogues [16]. There is good evidence that these genomic differences can be mainly attributed to indels [17], and that indels can be used to distinguish between different branches of the phylogenetic tree [18]. Indels thus could represent a significant contributor to gene density variation, through the degradation of redundant genes and the consequent creation of noncoding DNA. Since the disparities in gene density seen within the human genome seemed to correlate with those in the murine genome [3], this would indicate that most of this putative gene loss would have occurred much earlier in the phylogenetic tree.

In the gene cluster model, the expanses of noncoding DNA between gene clusters would be explained mainly on the basis of the degradation of redundant genes, rather than the generation of noncoding DNA through *de nouveau* insertions. As indicated above, indels could have a significant contribution to gene degradation. With the microrearrangement model, however, indels would represent epiphenomena without a major role in the modulation of gene density. The finding of extensive misalignments between the human and murine genomes, localized predominantly to noncoding regions [16], would indicate that indels occur with such a frequency that their potential influence should not be ignored. For this reason the gene cluster model seems more consistent with current knowledge.

Conclusions

The PG distribution provided an accurate description for the clustering of genes on human chromosome 7. It thus seemed appropriate to search for a biological mechanism based upon a scale invariant sum of a random (Poisson distributed) number of gamma distributions. Since the negative binomial distribution can be regarded as the discrete equivalent to the gamma distribution, a number of alternative statistical models could be constructed to explain the PG distribution [19].

It is conceivable that some other biological mechanism may eventually provide a more appropriate explanation for the observed clustering of genes. Indeed, one might postulate that insertions and deletions should affect the spacing of genes though their cumulative action within the intergene spacer region, rather than the degradation of redundant genes. This hypothesis is appealing, since presumably the intergene spacers should present a larger target for insertions and deletions and these events should be neutral, whereas events within gene structures would likely be selected against. How such a mechanism might account for to a Poisson-gamma distribution remains unclear, and for this reason this mechanism was not modelled here.

Regardless of the underlying biological mechanisms, the variance to mean power function that was evident to the physical distribution of genes on chromosome 7 implied an inherent scale invariance. The possible cause of this

scale invariance is worthy of consideration. Exponential dispersion models include a subclass of models characterized by this variance to mean power function, and which represents limiting forms for a broad range of statistical models [5]. The Tweedie convergence theorem, mentioned in the context of this limiting behaviour above, could be viewed as a kind of generalized Central Limit Theorem. Instead of error distributions being required to converging to a Gaussian form, they could be required to converge towards a Tweedie model, of which the Gaussian distribution is a special case. The scale invariance inherent to these models could be related to this behaviour and, in turn, could reflect the collective action of multiple complex processes.

Some of these processes might represent the gene duplication, mutation, deletion and chromosomal rearrangements that presumably have accompanied evolution. Ohno recognized that these processes, given the opportunity to manifest themselves over evolutionary timescales, could result in inhomogeneity in the physical distribution of genes, and in the presence of noncoding DNA [1]. Yet even with the understanding brought about by the Human Genome Project the question still arises: "Why are there clustered regions of high and low gene density, and are these accidents of history or driven by selection and evolution [2]?" When we observe the extent of clustering of genes on human chromosome 7 we might be tempted to interpret this clustering as evidence for some non-random force, such as selection. However, if one accepts the PG model, the clustering of genes on chromosome 7 could be attributable largely to the compound structure of the PG distribution. This compounding, of the Poisson and gamma distributions, would indicate the combined action of two random processes, and thus the resultant clustering of genes could be explained by predominantly neutral mechanisms.

There exists good evidence that the local density of genes correlates with the GC content of chromosomes [4], and that housekeeping expressed genes tend to group in clusters associated with high GC content [20,21]. These findings would indicate a non-random structure within the genome which seems difficult to reconcile with an apparent clustering of all classes of genes attributable to neutral processes. If genomic changes such as local duplications, insertions and deletions were dependent upon GC content for mechanistic reasons, then possibly these findings could be reconciled with the PG model presented here.

In summary, the physical distribution of gene structures within chromosome 7 was characterized by clustering. A variance to mean power function inherent to this clustering implicated a scale invariant PG distribution to describe the spatial distribution of genes within the chromosome. Data from the physical positions of genes on chromosome 7 provided stringent confirmation of the PG distribution in this regard. This statistical model represented the sum of a random (Poisson distributed) number of independent and identically distributed gamma distributions. The biological mechanisms to explain this statistical model remain subject to conjecture. Two hypothetical mechanisms were presented here: the microrearrangement model and the gene cluster model. Of these two hypotheses, the gene cluster model seemed to most faithfully represent Ohno's postulate that the major mechanisms behind genomic evolution are gene duplication, modulated by mutation, insertion and deletion [1].

Methods

Data collection

As part of the Chromosome 7 Annotation Project, full documentation has been provided for an assembly of 157,953,789 DNA nucleotides that comprehensively spans human chromosome 7 [3]. Of these data about 85% was derived from Celera whole-genome scaffolds, the remaining 15% came from clone-based sequences provided by the International Human Genome Sequencing Consortium. This assembly was made available at the public website, http://www.chr7.org/. The data used from this site came from columns 2, 3 and 4 of the table entitled "TCAG Annotated Genes on Chromosome 7". Column 2 identified the gene structures. In the present study all transcriptional variants and all gene/pseudogene segments from the T cell receptor loci (TRBV and TRGV) were excluded, leaving 1811 gene structures for analysis. Column 3 provided two numbers delineating the physical positions of the transcripts, and Column 4 identified the DNA strand (+/-) on which the transcripts resided. If a gene was located on the reverse (-) strand, then the first number in Column 3 represented the end of the gene, and *vice versa*. The analyses performed here used the start positions of these gene structures as the localization points for gene position.

Analytical methods

The length of chromosome 7 was subdivided into a sequence of non-overlapping and equal-sized bins, and the gene structures within each bin were enumerated. Since the enumerative bins were rigidly defined and spaced, whereas the gene structures potentially might span more than one bin, the position of each gene was defined according to the position of its *p*-terminus. The relationship between the variance and the mean number of gene structures per bin was thus examined over a range of bin sizes, from 20 to 4000 kb. The correlation coefficient squared r^2 was used to assess the linear correlation between the logarithmically transformed variances and means.

The fit of the variance to mean power function to these data was found to be reasonably robust, whether the positions of gene structures was defined by the *p*-termini, *q*-termini or the transcriptional start positions of the genes. For reasons of simplicity the analysis presented here was confined to gene position defined by the *p*-termini.

A probabilistic model was proposed for the spatial distribution of genes structures along the physical length of chromosome 7, based upon the PG distribution. The theoretical CDF from this model was fitted to the empirical CDF at the scale of 200 kb bins, by the minimization of the sum of the squared residuals. Because there were three adjustable parameters associated with the model, the Kolmogorov Smirnov test was applied to the composite hypothesis. The critical values for the Kolmogorov Smirnov distribution were therefore estimated by Monte Carlo simulation.

Goodness of fit for the transformed variance to mean relationship and the model CDF were further assessed by analyses of residuals. Normal probability plots of the residuals were constructed such that the data points would describe a linear relationship if they were normally distributed.

Acknowledgements

The author gratefully acknowledges the support provided by the Beattie Library and the Department of Radiation Oncology (both of the Ottawa Regional Cancer Centre), and the helpful advice of Drs. Martin Lercher, Stephen W. Scherer and Jeffrey R. MacDonald.

References

1. Ohno S: **Evolution is condemned to rely upon variations of the same theme: the one ancestral sequence for genes and spacers.** *Perspec Bio Med* 1982, **25:**559-572.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al.*: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
3. Scherer SW, Cheung J, MacDonald JR, Osborne LR, Nakabayashi K, Nakabayashi K, Herbrick J-A, Carson AR, Parker-Katiraee L, Skaug J, Khaja R *et al.*: **Human chromosome 7: DNA sequence and biology.** *Science* 2003, **300:**767-772.
4. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
5. Jørgensen B: *The Theory of Dispersion Models London: Chapman & Hall*; 1997.
6. Kumar S, Gadagkar SR, Filipski A, Gu X: **Determination of the number of conserved chromosomal segments between species.** *Genetics* 2001, **157:**1387-1395.
7. Nadeau JH, Taylor BA: **Lengths of chromosomal segments conserved since divergence of man and mouse.** *Proc Natl Acad Sci* 1984, **81:**814-818.
8. Nelder JA, Wedderburn RWM: **Generalized linear models.** *J Roy Statist Soc Ser A* 1972, **135:**370-384.
9. Kendal WS: **Spatial aggregation of the Colorado potato beetle described by an exponential dispersion model.** *Ecol Model* 2002, **151:**261-269.
10. Tweedie MCK: **An index which distinguishes between some important exponential families.** In *Statistics: applications and new directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference* Edited by: *Ghosh JK, J Roy J. Calcutta, India: Indian Statistical Institute*; 1984:579-604.
11. Pevzner P, Tesler G: **Genome rearrangements in mammalian evolution: lessons from human and mouse genomes.** *Genome Res* 2003, **13:**37-45.
12. Schopf JW: **Microfossils of the early archean apex chert: new evidence of the antiquity of life.** *Science* 1993, **260:**640-646.
13. Urrutia AO, Hurst LD: **The signature of selection mediated by expression on human genes.** *Genome Res* 2003, **13:**2260-2264.
14. Kendall DG: **Stochastic processes and population growth.** *J Roy Stat Soc Ser B* 1949, **11:**230-264.
15. Medhi J: *Stochastic processes New York: Wiley*; 1982:103.
16. Mouse Genome Sequencing Consortium: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420:**520-562.
17. Britten RJ, Rowen L, Williams J, Cameron RA: **Majority of divergence between closely related DNA samples is due to indels.** *Proc Natl Acad Sci USA* 2003, **100:**4661-4665.
18. Gupta RS: **The branching order and phylogenetic placement of species from completed bacterial genomes, based on conserved indels found in various proteins.** *Int Microbiol* 2001, **4:**187-202.
19. Boswell MT, Patil GP: **Chance mechanisms generating negative binomial distributions.** In *Random Counts in Scientific Work, Vol 1 of Random counts in scientific work: Expanded from the proceedings of the biometric society Symposium, Dallas, Texas, December 1968* Edited by: *Patil GP. University Park, PA: Pennsylvania State University Press*; 1970:3-22.
20. Lercher MJ, Urrutia AO, Hurst LD: **Clustering of housekeeping genes provides a unified model of gene order in the human genome.** *Nature Genet* 2002, **31:**180-183.
21. Lercher MJ, Urrutia A, Pavlíèek A, Hurst LD: **A unification of mosaic structures in the human genome.** *Hum Mol Genet* 2003, **12:**2411-2415.