Vaccine 28 (2010) 3308–3311

Contents lists available at ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

A probabilistic model for risk assessment of residual host cell DNA in biological products

Harry Yang*, Lanju Zhang, Mark Galinski

MedImmune, One MedImmune Way, Gaithersburg, MD 20878, United States

ARTICLE INFO

Article history:
Received 20 October 2009
Received in revised form 22 February 2010
Accepted 23 February 2010
Available online 10 March 2010

Keywords:
DNA oncogenicity
Enzyme inactivation
Host cell DNA
Infectivity
Risk assessment

ABSTRACT

Biological products such as viral vaccines manufactured in cells contain residual DNA derived from host cell substrates used in production. It is theoretically possible that the residual DNA could transmit activated oncogenes and/or latent infectious viral genomes to subjects receiving the product, and induce oncogenic or infective events. A probabilistic model to estimate the risks due to residual DNA is proposed. The model takes account of enzyme inactivation process. It allows for more accurate risk assessment when compared to methods currently in use. An application of the method to determine safety factor of a vaccine product is provided.

1. Introduction

In recent years, development of cell-based biological products has been in the forefront of drug research and development. Utilizing cutting edge technology, biological products can treat various conditions which defy conventional small molecule therapies. However, because biologics are produced from a cell substrate, it is inevitable that residual host cell DNA is present in the final products. There is a possibility for the residual DNA to transmit either an activated oncogene(s) or potentially an infectious viral DNA to product recipients, particularly if the biologic product is manufactured in a cell line that has tumorigenic potential [1]. Regulatory guidance suggests mitigating the risks of oncogenicity and infectivity by decreasing both the amount and the size of residual DNA [2,3]. In literature, the potential risks of residual DNA have been much researched by various researchers [4–6]. More recently, Sheng et al. [7] demonstrated that two cellular oncogenes when inoculated together could induce sarcomas in two different mouse strains. Peden et al. [8] have studied the risk associated with infectious agents in residual DNA, using HIV as a model. In their investigations, risk was quantified in terms of a safety factor, which is defined as number of doses needed to deliver an amount of oncogene (infectious agent) which induces tumor (infection). The calculation of oncogenicity risk uses the following formula in Eq. (1). A similar formula is used for calculating the safety factor of infectivity:

$$\text{safety factor} = \frac{O_m}{(OS/GS)I_0(hcDNA)}, \quad (1)$$

where O_m is the amount of oncogenes required for inducing an oncogenic event, OS and GS are average oncogene size and haploid genome size, respectively, I_0 is total number of oncogenes in the host genome, and $hcDNA$ is average amount of residual host cell DNA per dose. The expression $(OS/GS)I_0(hcDNA)$ in Eq. (1) represents the genomic mass equivalent of oncogenes in a dose.

While the calculation of the safety factor is both intuitive and easy to carry out, it does not account for disruption of the oncogene sequences through enzyme digestion; neither does it take account of the sizes of the individual oncogenes. Therefore, the risk estimates derived from their method are likely to be overstated. As a remedy, we introduce a probabilistic model to mechanistically study the relationship between the risks and characteristics of the purification process such as enzyme cutting efficiency, total amount of residual DNA in the final dose, and biological nature of the host cells including numbers and sizes of oncogenes and infectious viral DNA, amounts of oncogenes and infectious agent required to cause oncogenic and infectious events, respectively. The method is both simple and convenient to use. It is a useful tool for residual DNA risk assessment. The use of the model is illustrated through a real application.

* Corresponding author. Tel.: +1 301 398 4405.
E-mail address: yangh@medimmune.com (H. Yang).

0264-410X/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.vaccine.2010.02.099

2. Materials and methods

2.1. Data

We assess oncogenic and infective potential of residual hcDNA from a cell-based live, attenuated influenza vaccine. The product is manufactured from a production process that uses Madin Darby Canine Kidney (MDCK) as the cell. The process employs several purification steps to remove hcDNA, which include tangential flow filtration (TFF) and chromatography assay. During the TFF process step, DNA is removed from the virus based on the size difference between the virus and host cell DNA. Any residual DNA is removed or reduced in size during the affinity chromatography step. DNA does not bind to the chromatography media; however, any DNA that is associated with the virus or host cell protein that binds to the media is degraded by treating with benzonase, which is included in the chromatography buffer wash. Using a canine SINE quantitative PCR, the amount of residual hcDNA is determined to be less than 1 ng per dose. With a direct-labeling method, the size distribution of residual DNA is also examined. The median size is approximately 450 base pairs (bp); approximately 64% of residual DNA is less than 500 bp in length. The haploid genome size of the canine genome is determined to be 2.41×10^9 bp.

2.2. Oncogenicity assessment

There are approximately 200 oncogenes identified in various species [9]. Using the SOURCE (located at http://smd.standford.edu) 81 expressed human oncogenes are found in 24 different tissues [8]. The average size of human oncogenes is 1925 bp with a standard deviation of 87 bp. Because the precise number of oncogenes contained in MDCK cell genome is unknown, for the oncogenic risk analysis, we restrict our evaluation to a single oncogene presumably having a size of 1925. The amount of oncogenes required to induce cancer is extrapolated from Sheng et al. [7]. They demonstrate that tumors could be formed in two different mouse strains (NIH Swiss, C57BL/6) that were co-injected with 12.5 μg each of two plasmids, each containing an activated oncogene (activated human H-ras and c-myc). This value (O_m) is calculated from the estimated size of the plasmid backbone (3186 bp) used in Sheng et al. [7], assuming that the oncogene inserted to the plasmid backbone has 1925 bp. Based on the total construct, the oncogene would account for 37.7% of the construct. If 12.5 μg of the plasmid is required for each oncogene of two oncogenes, as described by Sheng et al. [7], then the total oncogene portion amount to 9.4 μg ($25 \times 37.7\% = O_m$).

2.3. Infectivity assessment

This evaluation utilizes research results from Peden et al. [8]. Using HIV as a model, they have found that hcDNA from HIV-infected cells is infectious at 2.5 μg. In our single infective agent safety factor calculations, we make the assumptions: (1) 2.5 μg canine hcDNA is assumed to have an infectivity similar to hcDNA containing a HIV provirus; (2) the viral genome size is 7000 bp [10], which represents a smaller retrovirus genome than HIV genome of 10,000 bp; (3) a diploid canine genome size is 4.82×10^9 as there is usually a single copy of provirus per cell [8].

3. Modeling

3.1. Modeling of DNA digestion by enzyme

To facilitate introduction of our model, we will focus on the assessment of oncogenicity. The same method, once fully developed, can be directly applied to the infectivity risk evaluation. For the rest of the paper, we use Φ, Ω and c to denote the host cell genome, oncogene DNA sequence residing in the host genome and phosphate ester bond between two nucleotides, respectively. We further express Φ and Ω as

$$\Phi = B_1 c B_2 c \ldots c B_M, \quad \Omega = B_l c B_{l+1} c \ldots c B_{l+m-1}, \tag{2}$$

where M and m represent haploid size of host genome and oncogene size, respectively, and $l \geq 1$, $m > 1$ and $l + m - 1 < M$. We refer the bond c's within Ω as $c_1, c_2 \ldots c_{m-1}$. Define X_i as random variables that can take value either 0 or 1, with $P[X_i = 1] = P[c_i$ is disrupted by the enzyme$] = 1 - P[X_i = 0] = p$. The probability p represents the cutting efficiency of the enzyme. It is reasonable to assume that all X_i are independent. Therefore these $m-1$ variables X_i are independently identically distributed (*i.i.d.*) according to a Bernoulli distribution [11].

After the host cell genome Φ is enzymatically digested, for the oncogene Ω to remain intact, none of the bonds c's within the oncogene should be cut by the enzyme. That is

$$X_1 = X_2 \ldots = X_{m-1} = 0. \tag{3}$$

Thus the probability for Ω not to be disrupted is

$$Pr[X_1 = X_2 \ldots = X_{m-1} = 0] = (1-p)^{m-1}. \tag{4}$$

3.2. Residual DNA from oncogenes

Now assume that the host cell genome Φ contains I_0 oncogenes of size m_i.

$$\Omega_i = B_{l_i} c B_{l_i+1} c \ldots c B_{l_i+m_i-1}, \quad 1 \leq i \leq I_0 \tag{5}$$

By (4), the probability for Ω_i to be uncut by enzyme is given by

$$p_i = (1-p)^{m_i - 1}. \tag{6}$$

It is of interest to estimate the amount of oncogenes either fragmented or unfragmented in a final dose. To that end, we let U denote the total amount of residual host cell DNA per dose, V_i, W_i and Z_i be the total number of copies of oncogene Ω_i (either fragmented or unfragmented), the total number of copies of unfragmented oncogene Ω_i and the total number of copies of fragmented oncogene Ω_i in a dose, respectively. Clearly $V_i = W_i + Z_i$. Finally let Y be the total amount of unfragmented oncogene Ω_i in a dose. Clearly U, V_i, W_i and Y are random variables, and

$$Y = \sum_{i=1}^{I_0} d_i W_i \tag{7}$$

where d_i is the weight of oncogene Ω_i. Given the haploid size of the host cell genome M, it is reasonable to assume that conditional on U, V_i has a Poisson distribution $P((m_i/M)(U/d_i))$ where U/d_i represents the maximum number of oncogene Ω_i which the total amount of residual DNA, U, in a dose can possibly contain. It is also reasonable to assume that conditional on V_i, W_i is distributed according to a binomial distribution $B(p_i, V_i)$ with p_i being given in Eq. (6). Using the facts [11] that

$$\begin{array}{l} E[V_i|U] = \dfrac{m_i}{M}\left(\dfrac{U}{d_i}\right) \\ E[W_i|V_i] = p_i V_i \\ E[W_i] = E_{V_i}(E_{W_i}[W_i|V_i]) = E_{V_i}[p_i V_i] = E_U(E_{V_i}[p_i V_i|U]) = \dfrac{p_i(m_i/M)E[U]}{d_i} \end{array}, \tag{8}$$

the expected value of total amount of uncut oncogenes Y can be obtained by

$$E[Y] = \sum_{i=1}^{I_0} d_i E[W_i] = \sum_{i=1}^{I_0} p_i \frac{m_i}{M} E[U]. \tag{9}$$

3.3. Safety factor estimation

Following the risk assessment in Refs. [7,8], we define safety factor (*SF*) as the number of doses required to produce an oncogenic amount O_m of oncogenes. Let Y_i be the amount of unfragmented oncogenes in dose j, $j = 1, \ldots, SF$. The safety factor is an integer such that

$$\sum_{j=1}^{SF} Y_i = O_m \tag{10}$$

When the number *SF* is large, by the Strong Law of Large Numbers [12]:

$$\frac{\sum_{j=1}^{SF} Y_j}{SF} \approx E[Y]. \tag{11}$$

Combining (6), (9)–(11), the safety factor, *SF*, can be estimated by

$$SF = \frac{O_m}{\sum_{i=1}^{I_0}(1-p)^{m_i-1}\frac{m_i}{M}E[U]}. \tag{12}$$

The safety factor is a function of amount of oncogenes, O_m, required for inducing an oncogenic event, total number of oncogenes in host genome, I_0, and their sizes m_i, average amount of residual host cell DNA $E[U]$ per dose, and finally enzyme cutting efficiency, p. The factors O_m, I_0, m_i and $E[U]$ can be experimentally determined. The average amount of host residual DNA $E[U]$ in a single dose is dependent on the efficiency of the downstream purification processes. Eq. (12) indicates that the more the processes could remove residual DNA, the larger the safety factor is. It is also evident that the higher the enzyme cutting efficiency p is, the larger the *SF*. Since p is influenced by many factors, the estimation of this quantity is not so straightforward. In the following a modeling approach is suggested to estimate the enzyme cutting efficiency. Noting that when $p = 0$, Eq. (12) is reduced to

$$SF = \frac{O_m}{\sum_{i=1}^{I_0}\frac{m_i}{M}E[U]} = \frac{O_m}{(OS/GS)I_0E[U]} \tag{13}$$

where $OS = \sum_{i=1}^{I_0} m_i/I_0$, $GS = M$ and $E[U]$ are the average oncogene size, the size of the host cell genome and the average amount of residual host cell DNA, respectively. Comparing Eq. (1) and (13), we can conclude that Eq. (1) is a special case of Eq. (12) when there are no DNA inactivation steps.

3.4. Determination of enzyme efficiency

After enzyme digestion, any DNA segment takes the form:

$$B_{r+1}cB_{r+2}c\ldots cB_{r+X} \tag{14}$$

where r is an integer and X, representing the length of the DNA segment, is a random variable. Let p denote the probability for enzyme to cleave bond c, as defined in Section 2.1. Note that the length of the above DNA segment is the same as the number of failed attempts made by the enzyme at cutting through the bonds c's before it successfully disrupts the bond c right after nucleotide B_{r+X}. The length X, in essence, can be described by a geometric distribution with parameter p [11]. In other words:

$$Pr[X = k] = (1-p)^{k-1}p, k = 1, 2, \ \ldots, M-1. \tag{15}$$

The theoretical median of X is given by

$$\text{median} = -\frac{\log 2}{\log(1-p)}. \tag{16}$$

If the residual DNA size distribution can be quantified, the median can be empirically estimated. Using Eq. (16), we could estimate the enzyme cutting efficiency p, which in turn can be used to estimate the safety factor in Eq. (12).

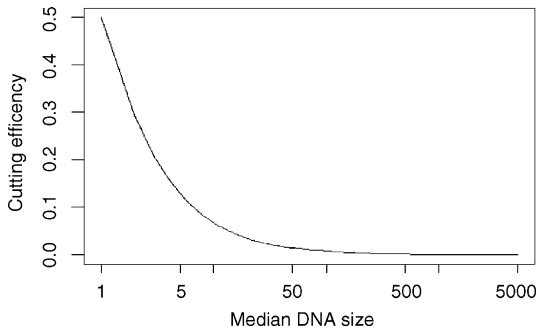

Fig. 1. Relationship between enzyme cutting efficiency and median DNA size.

3.5. Size distribution of residual DNA

In clinical research laboratories, various analytical methods such as agarose, polyacrylamide and capillary electrophoresis are used to measure the size distribution of residual DNA in biological products. These methodologies resolve purified DNA in a suitable matrix where the DNA length can be estimated relative to known DNA size markers. After the distribution of residual DNA is quantified, parameters of the distribution such as mean and median can readily be obtained. Let Med_0 denote the median size of residual DNA, determined by one of the aforesaid methods. Equating Med_0 to the theoretical median in Eq. (16) gives rise to an estimate of enzyme efficiency p:

$$\hat{p} = 1 - 2^{-1/Med_0} \tag{17}$$

The relationship between enzyme efficiency and median size of residual DNA is depicted in Fig. 1.

It is evident that the more efficient the enzyme is, the smaller the median size of residual DNA is. Combining Eq. (12) and (17), we establish the following relationship between the safety factor and other characteristics of the manufacture process:

$$SF = \frac{O_m}{\sum_{i=1}^{I_0} 2^{-(m_i-1)/Med_0}\frac{m_i}{M}E[U]}. \tag{18}$$

Since the safety factor is a decreasing factor of the median size Med_0 of residual DNA, the smaller the size of residual DNA is, the larger the safety factor is. A similar formula can be derived for safety factor concerning infectivity. It is given as follows:

$$SF_1 = \frac{Q_m}{\sum_{i=1}^{J_0} 2^{-(n_i-1)/Med_0}\frac{n_i}{N}E[U]} \tag{19}$$

where Q_m, J_0 and n_i are viral genome amount required to induce an infection, total number of proviruses contained in MDCK cell genome and their sizes n_i, respectively, and N is the diploid size of the host cell genome.

4. Results

4.1. Risk of oncogenicity

The safety factor for oncogenicity is calculated based on Eq. (18). As discussed in Section 2, the observational and experimental data suggest: (a) $O_m = 9.4\ \mu g$; (b) the amount of residual hcDNA per dose, $E[U] < 1$ ng; (c) the median size of residual DNA is approximately 450 bp; (d) the haploid genome size of the MDCK genome $M = 2.41 \times 10^9$ bp. It is also assumed that there is only one oncogene of size 1925 contained in the canine genome. The safety factor is calculated to be 2.3×10^{11}. This indicates that 230 billion doses of

vaccine would need to be administered before an oncogene dosage equivalent to 9.4 μg would be reached.

4.2. Risk of infectivity

Safety factor for infectivity due to a single provirus is similarly calculated, substituting the following values for those in Eq. (19): $Q_m = 2.5$ μg; $E[U] < 1$ ng; $Med_0 = 450$ bp; diploid size of host genome $N = 4.82 \times 10^9$ bp; $J_0 = 1$; $n_1 = 7000$ bp. The safety factor for a single provirus is calculated to be 8.3×10^{13} or the equivalent of 83 trillion doses to induce an infective event.

5. Discussion

We repeat the calculations of safety factors for the example given in Section 4, using Eq. (1), which is a method suggested in Refs. [7,8]. The safety factors of oncogenicity and infectivity are determined to be 1.2×10^{10} and 1.7×10^9, respectively. These calculations overestimate risk due to oncogenicity by more than 19-fold. The overestimation issue for risk of infectivity is even more pronounced; the risk is overstated by more than 48,000 times. The overestimation stems from the fact that enzyme inactivation is not taken into account. The method we propose in the paper clearly results in more accurate estimates of risks because of the inclusion of enzyme inactivation in its calculations. It is also worth noting that in all the calculations of safety factors, we assume that the residual hcDNA is less than 1 ng. However, the intranasal administration of the vaccine is likely to reduce the residual hcDNA found in tissues which, if shown to be true, would further lower associated risks.

Model validation is an integral part of a probabilistic method development. It ensures that a method is fit for its intended use. The accuracy and reliability of the risk assessment approach we develop ideally should be validated by comparing its estimated values with observed events. However, before a biological product is approved for marketing and distributing, there are only a limited number of doses administered in human subjects during clinical development. Because the risks of oncogenicity and infectivity due to hcDNA are in general low, it would take many doses to observe some events. As a result, validation of the model based on empirical data can only be accomplished if one were to follow millions of doses for extended periods of time. This is one of the limitations the proposed method has. It is also worth pointing out that the quantity in Eq. (18) or (19) represents a point estimate of the safety factor. Because the parameters involved in the calculations are determined through analytical methods which have inherent variability, the accuracy and precision of the safety factor estimate are influenced by that of the analytical methods. It is advisable to conduct a sensitivity analysis of the safety factors. In Section 4, risks of oncogenicity and infectivity are calculated based on the estimated amount of residual DNA per dose being less than 1 ng, and the median size of DNA being 450 bp. Results of analysis of clinical materials suggest that the quantity of residual hcDNA is approximately 0.1 ng/dose. In addition, the DNA size analysis we conduct indicate that the median size of residual DNA is 450 bp with 64% of the hcDNA less than 500 bp in length and no detectable DNA above 1000 bp. Substituting $E[U] = 1$, and $Med_0 = 1000$ in Eq. (18) and Eq. (19), the safety factors of oncogenicity and infectivity are estimated to be 4.9×10^{10} and 2.2×10^{11}, which represent worst case scenario of safety factor estimates.

In general, using the analytical methods discussed in Section 3.5, variability associated with the estimate of the median size Med_0 of residual DNA can be obtained. For example, we could perform the analysis on a large number of samples, to give rise to a set of estimates of median size. The error related to the mean median size of residual DNA can be calculated. Applying Taylor expansion, the error associated with safety factor estimate can be determined. Alternatively, we could use bootstrapping method to estimate the error, based on resampling of samples from the size distribution determined by the method in [13]. This will allow us to construct one-sided confidence lower bound for the safety factor, which represents the worst case scenario.

Lastly, the theoretical model is developed in a very general context. It can easily be applied to the evaluation of oncogenic and infective risks of other biological products. The assessment of the intranasal vaccine serves as an illustration to the use of the method. As we have demonstrated, the use of the method is simple and straightforward. For interested parties a written computer code of the method can be obtained by contacting the first author.

Acknowledgements

We thank the referees for their valuable comments that have helped to improve the manuscript greatly.

References

[1] Lewis Jr AM, Krause P, Peden K. A defined-risks approach to the regulatory assessment of the use of neoplastic cells as substrates for viral vaccine manufacture. Dev Biol 2001;106:513–35.
[2] FDA. CBER Draft Guidance: Characterization and Qualification of Cell Substrates and Other Biological Starting Materials Used in the Production of Viral Vaccines for the Prevention and Treatment of Infectious Diseases; September, 2006.
[3] Griffiths E. WHO requirements for the use of animal cells as in vitro substrates for the production of biologicals: application to influenza vaccine production. In: Brown F, Robertson JS, Shcild GC, Wood JM, editors. Inactivated Influenza Vaccines Prepared in Cell Culture. Dev. Biol. Stand., 98. Karger: Basel; 1999. p. 153–7.
[4] Petricciani JC, Regan PJ. Risk of neoplastic transformation from cellular DNA: calculations using the oncogene model. Dev Biol Stand 1987;68:43–9.
[5] Petricianni JC, Horaud FN. DNA, dragons and saity. Biologicals 1995;23:233–8.
[6] Petricciani J, Loewer J. An overview of cell DNA issues. Dev Biol 2001;106:275–82.
[7] Sheng L, Cai F, Zhu Y, Pal A, Athanasiou M, Orrison B, et al. Oncogenicity of DNA in vivo: tumor induction with expression plasmids for activated H-ras and c-myc. Biologicals 2008;36(3):184–97.
[8] Peden K, Sheng L, Pal A, Lewis A. Biological activity of residual cell substrate DNA. Dev Biol (Basel) 2006;123:45–56 [discussion 55–73].
[9] Zhou Y, Ma B-G, Zhang H-Y. Human oncogene tissue-specific expression level significantly correlates with sequence compositional features. FEBS Lett 2007;581(22):4361–5.
[10] Goff SP. Retroviridae: the retroviruses and their replication. In: Knipe DM, Howley PM, editors. Fundamental virology. 4th ed. Philadelphia, PA, USA: Lippincott Williams & Wilkins; 1999. p. 843–911.
[11] Mood AM, Graybill FA, Boes DC. Introduction to the theory of statistics. McGraw-Hill Book Company; 1988.
[12] Billingsley P. Probability and measure. John Wiley & Sons; 1986.
[13] Efron B, Tibshirani RJ. An introduction to bootstrapping. Chapman & Hall; 1994.