Probability models for detecting transgenic plants

Carlos M. Hernández-Suárez, Osval A. Montesinos-López, Graham McLaren, and José Crossa
The presence of genetically modified plants, hereafter named the adventitious presence of unwanted transgenic plants (AP), is becoming common in modern crop production systems.

This reality has created concerns regarding possible gene flow through outcrossing between AP crops and their landrace and wild relatives.

This is especially important in a country such as Mexico, a center of diversity for maize, where the effects of AP maize outcrossing with traditional maize landraces and wild relatives such as tripsacum and teocinte are unknown.
Recently, different authors have reported contrasting results in terms of detecting AP maize in Mexico.

Quist and Chapela (2001; 2002) were the first to report AP landraces collected in the Sierra Juarez region of the Mexican State of Oaxaca; they specifically identified genes from *Bacillus thuringiensis* (Bt), a soil bacterium used to create maize that is resistant to some insects.

In contrast, four years later, Ortiz-García et al. (2005a; 2005b) sampled maize landraces in the same region of Oaxaca State and failed to detect AP.
When testing for AP, two distinct activities should be emphasized

(1) The first is determining the optimal sample size \( (n) \) and sampling strategy to be used when taking seeds at random from a seed lot (farmer’s field)

(2) the second is determining the sample preparation and testing method to be used in the laboratory. The sensitivity of the analyses and specificity of the tests are important factors that may affect the rates of false negative and false positive results.
Errors and risks

All quality laboratory methods produce results with false positives and false negatives rates

→ False positive $\delta =$ probability of falsely detecting a seed with impurity and
→ False negatives $\lambda =$ probability of failing to detect seed with impurity.

These two types of errors, which commonly occur in any testing plan, can be integrated in an overall consumer and producer risk assessment. Proposed testing plans can integrate a given

→ lower quality limit (LQL) for the consumer risk and
→ acceptable quality limit (AQL) for the producer risks.
Laboratory tests

• Because laboratory tests are expensive, it is not feasible to analyze all $n$ individual seeds collected from a lot. There are several testing plans for reducing the number of samples to be analyzed.
• One plan consists of testing pooled seed samples.
• The group testing method of Dorfman (1943) is effective for reducing the number of laboratory analyses and can result in up to 80% savings in the number of laboratory analyses.
Dorfman method

• This method consists of dividing $n$ individual samples (e.g., seeds) into $g$ groups (or pools), each of size $k$. If a group tests positive, then at least one individual in the pool is positive; the author gives an approximate solution for the optimal value of $k$.

• A formula for determining the sample size ($n$) required for detecting AP can be derived from the Dorfman method.
**Dorfman method**

- Assume a population of size \( N \) in which a fraction \( p \) has AP [say type (+)]. We consider the problem of determining the optimum values of \( n \) and \( k \) such that the probability of detecting at least one individual with AP is greater than \((1-\alpha)\) (for a given small \( \alpha \)).

- For sample size \( n \) and group size \( k \), \( g=n/k \) pools can be formed. If \( X \) is the number of + individuals in a pool, then \( P(X=j) \) \((j=1, 2, \ldots, k)\) follows a binomial distribution \( X \sim Bin(k, p) \).

- The probability that a group is (+) is one minus the probability that \( k \) randomly selected individuals are negative

\[
P(X > 0) = 1 - (1 - p)^k
\]
The probability of a pool testing negative (−) is

\[ P(X = 0) = (1 - p)^k \]

Because there are \( g = \frac{n}{k} \) pools, the probability of detecting only (−) groups, given that the proportion of (+) individuals in the population is \( p \), is

\[ \left[ (1 - p)^k \right]^{n/k} = (1 - p)^n \]
• If a small probability, $\alpha$, of detecting only individuals is required, given that there is a proportion $p$ of (+) individuals in the population, then previous equation can be written as

$$(1 - p)^n < \alpha$$

• It should be pointed out that the Dorfman model was not developed with the objective of determining the sample size $n$ but rather for determining the required number of pools, $g$, and the size of the pools, $k$, that will minimize the number of laboratory tests, $T$. 
Expected number of laboratory tests ($T$) and expected relative cost

- Under these premises, the expected value of $T E(T) = g + kgp'$ is a function of the number of pools ($g=n/k$), plus the number of individuals in the positive pools that need to be analyzed.
- Therefore, the ratio between the expected number of laboratory tests required ($T$) and the required sample size ($n$) for each lab method is a measure of its expected relative cost ($\frac{E(T)}{n} = \frac{g + kgp'}{n} = \frac{1}{k} + p'$).
Dorfman’s assumption

• However, Dorfman’s model assumes that when \( k \) individuals in a pool are mixed, AP concentration is not diluted. Therefore, under these assumptions, the value of \( n \) that satisfies the previous equation can be obtained as

\[
n = \frac{\log(\alpha)}{\log(1 - p)}
\]
• This analysis would suggest a single group; however, in practice AP cannot always be detected when the proportion of AP seeds in the group is very small because the analytical methods used may not be sensitive enough.

• Previous equation does not give any guidelines as to how the concentration of the trait of interest (AP) (impurity) in individual seeds \(w\) could affect the pool size \(k\), or how the dilution effect could make AP undetectable by standard analytical procedures in the laboratory \(c\).
Objetives of this research

To propose probabilistic models for determining the required
→ sample size, $n$,
→ number of pools, $g$,
→ size of the pool, $k$,

that will detect individuals containing AP with a probability $\geq (1-\alpha)$
(for small $\alpha$)

The proposed models were developed within the framework of the
Dorfman model, but considering:
(1) the dilution effect when forming groups (pools) of seed to be tested,
(2) the detection limit of the laboratory test ($c$),
(3) the different rates of false positives and false negatives, and
(4) the assessment of consumer and producer risks.

The probability distributions used in this study were binomial and
negative binomial distributions for
Binomial sampling with the dilution effect

- When \( k \) individuals that form a pool are mixed or homogenized, the AP will be diluted; this dilution effect increases with the size of the pool and may decrease the AP concentration in the pool below the test’s detection limit (\( c \)), thereby increasing the number of false negatives (i.e., seed(s) with AP is not detected when in fact it is present in the group).

- We assume a reference population of size \( N \), with a proportion \( p \) of individuals with AP [type (+)]. We also assume that the concentration of AP per individual, \( w \), is known (i.e., transgenic DNA as % of the total DNA in the seed).

- When \( g \) pools are formed from a total of \( n \) individuals, the AP concentration in a single (+) individual in a pool is reduced to \( \frac{wg}{n} = \frac{w}{k} \).

- If \( c \) is the laboratory detection limit, it is required that \( \frac{w}{k} \geq c \).
The question is: what is the required sample size \( n \) and pool size \( k \) such that the probability of detecting individuals of type (+) in the population is equal to or greater than \((1-\alpha)\)?

Variable \( X = \) number of (+) individuals in the pool of size \( k \) \((X=0,1,2,\ldots,k)\) is a binomial variable with parameters \( k \) and \( p \), that is, \( X \sim Bin(k, p) \).

Hence, the probability that a group will be detected (+) is
To compute more precise probability values and avoid rounding errors when calculating $\frac{ck}{w}$ we will use the relationship between the binomial and beta distributions:

$$\sum_{j=ck/w}^{k} P(X = j) = \sum_{j=ck/w}^{k} \binom{k}{j} p^j (1-p)^{k-j}$$

Therefore

$$P(X \geq \frac{ck}{w}) = \int_{ck/w}^{k} P(X = j) = \int_{ck/w}^{k} \binom{k}{j} p^j (1-p)^{k-j}$$
For the detection of AP, two types of errors rates are important.

→ One is the proportion of false positives $\delta$, which is the probability that one individual (or group) is detected as (+) even though it is (-) ($1-\delta$ is the test specificity);

→ The other is the rate of false negatives $\lambda$, which is the probability of an individual or pool testing (-) even though it is (+) ($1-\lambda$ is the test sensitivity).
In the previous equations the probabilities of false positives and false negative and the test of specificity and the test of sensitivity are introduced.

The strategy for minimizing the number of laboratory tests is to find, for a given $n$, a value of $k$ between 1 and $\min(n, k^1 = \left\lfloor \frac{w}{c} \right\rfloor - 1)$ that satisfies the previous equations with the minimum relative cost.
Negative binomial sampling with the dilution effect considering false positives and false negatives

• Haldane (1945) proposed the inverse sampling method (or negative binomial sampling or inverse binomial sampling) for cases where $p$ is small (i.e., $p \leq 0.1$). In this method, sampling continues until $m$ individuals with AP are obtained.

• The inverse sampling method suggested by Haldane (1945) is more precise than binomial sampling because when $m > 1$, the coefficient of variation of $p$ decreases.
Testing seed plans with the dilution effect considering false positives, false negatives, lower quality limit (LQL), and a given acceptable quality limit (AQL)

- The AP testing plan previously outlined has the aim of computing the required sample size ($n$), pool size ($k$), and number of pools ($g$) to guarantee, with probability $\geq(1-\alpha)$, that at least one AP plant (or AP pool) will be in the sample.

- These types of testing plans have zero tolerance because they focus only on the limiting quality level (LQL=$p$) (consumer risks) and do not consider the acceptable quality level (AQL) (i.e., AQL=0), which refers to different seed production levels under normal conditions (producer risks)
## Results

The binomial sampling method with the dilution effect

**with $\lambda=0$ (false negatives), $\delta=0$ (false positives) and the dilution effect ($c=0.0001$)**

\[ (1-\alpha)=95\% \]

<table>
<thead>
<tr>
<th>$p$</th>
<th>$w=0.0002$</th>
<th>$w=0.0008$</th>
<th>$w=0.002$</th>
<th>$w=0.006$</th>
<th>$w=0.010$</th>
<th>$w=0.014$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$</td>
<td>$k$</td>
<td>$g$</td>
<td>$n$</td>
<td>$k$</td>
<td>$g$</td>
</tr>
<tr>
<td>0.005</td>
<td>19934</td>
<td>1</td>
<td>19934</td>
<td>23192</td>
<td>7</td>
<td>3314</td>
</tr>
<tr>
<td>0.01</td>
<td>7036</td>
<td>1</td>
<td>7036</td>
<td>6420</td>
<td>7</td>
<td>918</td>
</tr>
<tr>
<td>0.02</td>
<td>2480</td>
<td>1</td>
<td>2480</td>
<td>1800</td>
<td>7</td>
<td>258</td>
</tr>
<tr>
<td>0.03</td>
<td>1345</td>
<td>1</td>
<td>1345</td>
<td>862</td>
<td>7</td>
<td>124</td>
</tr>
<tr>
<td>0.04</td>
<td>871</td>
<td>1</td>
<td>871</td>
<td>519</td>
<td>7</td>
<td>75</td>
</tr>
<tr>
<td>0.05</td>
<td>621</td>
<td>1</td>
<td>621</td>
<td>344</td>
<td>7</td>
<td>50</td>
</tr>
<tr>
<td>0.06</td>
<td>471</td>
<td>1</td>
<td>471</td>
<td>253</td>
<td>7</td>
<td>37</td>
</tr>
<tr>
<td>0.07</td>
<td>372</td>
<td>1</td>
<td>372</td>
<td>190</td>
<td>7</td>
<td>28</td>
</tr>
<tr>
<td>0.08</td>
<td>303</td>
<td>1</td>
<td>303</td>
<td>155</td>
<td>7</td>
<td>23</td>
</tr>
<tr>
<td>0.09</td>
<td>253</td>
<td>1</td>
<td>253</td>
<td>127</td>
<td>7</td>
<td>19</td>
</tr>
<tr>
<td>0.1</td>
<td>215</td>
<td>1</td>
<td>215</td>
<td>106</td>
<td>7</td>
<td>16</td>
</tr>
<tr>
<td>0.11</td>
<td>186</td>
<td>1</td>
<td>186</td>
<td>85</td>
<td>7</td>
<td>13</td>
</tr>
<tr>
<td>0.12</td>
<td>163</td>
<td>1</td>
<td>163</td>
<td>78</td>
<td>7</td>
<td>12</td>
</tr>
</tbody>
</table>
Figure 1. Optimum group size ($k$) and sample size ($n$) for different values of $p$ under the binomial distribution with the dilution effect for $c=0.0001$, $\lambda=0$, $\delta=0$, $w=0.01$, and $(1-\alpha=0.95)$. 

<table>
<thead>
<tr>
<th>$p$</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>892</td>
</tr>
<tr>
<td>0.03</td>
<td>100</td>
</tr>
<tr>
<td>0.05</td>
<td>90</td>
</tr>
<tr>
<td>0.07</td>
<td>56</td>
</tr>
<tr>
<td>0.09</td>
<td>40</td>
</tr>
<tr>
<td>0.11</td>
<td>31</td>
</tr>
</tbody>
</table>
# RESULTS

The negative binomial sampling method with the dilution effect, considering false positives and false negatives

\[ (1-\alpha) = 95\% \text{ of detecting different numbers of individuals (m) with the trait of interest using the negative binomial distribution with the dilution effect and a probability of } \lambda = 0.002 \text{ of false negatives and } \delta = 0.00 \text{ of false positives} \]

\[ p = 0.01 \]

<table>
<thead>
<tr>
<th>m</th>
<th>n</th>
<th>k</th>
<th>g</th>
<th>m</th>
<th>n</th>
<th>k</th>
<th>g</th>
<th>m</th>
<th>n</th>
<th>k</th>
<th>g</th>
<th>m</th>
<th>n</th>
<th>k</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7180</td>
<td>1</td>
<td>7180</td>
<td>6553</td>
<td>7</td>
<td>937</td>
<td>3288</td>
<td>19</td>
<td>174</td>
<td>1358</td>
<td>27</td>
<td>51</td>
<td>991</td>
<td>15</td>
<td>67</td>
</tr>
<tr>
<td>3</td>
<td>15090</td>
<td>1</td>
<td>15090</td>
<td>13777</td>
<td>7</td>
<td>1969</td>
<td>6936</td>
<td>19</td>
<td>366</td>
<td>2951</td>
<td>27</td>
<td>110</td>
<td>2179</td>
<td>15</td>
<td>146</td>
</tr>
<tr>
<td>5</td>
<td>21939</td>
<td>1</td>
<td>21939</td>
<td>20035</td>
<td>7</td>
<td>2863</td>
<td>10109</td>
<td>19</td>
<td>533</td>
<td>4367</td>
<td>27</td>
<td>162</td>
<td>3169</td>
<td>15</td>
<td>212</td>
</tr>
<tr>
<td>7</td>
<td>28384</td>
<td>1</td>
<td>28384</td>
<td>25929</td>
<td>7</td>
<td>3705</td>
<td>13092</td>
<td>19</td>
<td>690</td>
<td>5665</td>
<td>27</td>
<td>210</td>
<td>4159</td>
<td>15</td>
<td>278</td>
</tr>
<tr>
<td>9</td>
<td>34598</td>
<td>1</td>
<td>34598</td>
<td>31606</td>
<td>7</td>
<td>4516</td>
<td>15961</td>
<td>19</td>
<td>841</td>
<td>6963</td>
<td>27</td>
<td>258</td>
<td>5149</td>
<td>15</td>
<td>344</td>
</tr>
<tr>
<td>11</td>
<td>40656</td>
<td>1</td>
<td>40656</td>
<td>37143</td>
<td>7</td>
<td>5307</td>
<td>18773</td>
<td>19</td>
<td>989</td>
<td>8202</td>
<td>27</td>
<td>304</td>
<td>6139</td>
<td>15</td>
<td>410</td>
</tr>
<tr>
<td>13</td>
<td>46602</td>
<td>1</td>
<td>46602</td>
<td>42575</td>
<td>7</td>
<td>6083</td>
<td>21528</td>
<td>19</td>
<td>1134</td>
<td>9441</td>
<td>27</td>
<td>350</td>
<td>7030</td>
<td>15</td>
<td>469</td>
</tr>
<tr>
<td>15</td>
<td>52460</td>
<td>1</td>
<td>52460</td>
<td>47930</td>
<td>7</td>
<td>6848</td>
<td>24226</td>
<td>19</td>
<td>1276</td>
<td>10621</td>
<td>27</td>
<td>394</td>
<td>7921</td>
<td>15</td>
<td>529</td>
</tr>
<tr>
<td>17</td>
<td>58248</td>
<td>1</td>
<td>58248</td>
<td>53215</td>
<td>7</td>
<td>7603</td>
<td>26905</td>
<td>19</td>
<td>1417</td>
<td>11801</td>
<td>27</td>
<td>438</td>
<td>8812</td>
<td>15</td>
<td>588</td>
</tr>
<tr>
<td>19</td>
<td>63978</td>
<td>1</td>
<td>63978</td>
<td>58458</td>
<td>7</td>
<td>8352</td>
<td>29565</td>
<td>19</td>
<td>1557</td>
<td>12981</td>
<td>27</td>
<td>481</td>
<td>9703</td>
<td>15</td>
<td>647</td>
</tr>
<tr>
<td>21</td>
<td>69659</td>
<td>1</td>
<td>69659</td>
<td>63645</td>
<td>7</td>
<td>9093</td>
<td>32187</td>
<td>19</td>
<td>1695</td>
<td>14161</td>
<td>27</td>
<td>525</td>
<td>10594</td>
<td>15</td>
<td>707</td>
</tr>
<tr>
<td>23</td>
<td>75299</td>
<td>1</td>
<td>75299</td>
<td>68804</td>
<td>7</td>
<td>9830</td>
<td>34790</td>
<td>19</td>
<td>1832</td>
<td>15341</td>
<td>27</td>
<td>569</td>
<td>11485</td>
<td>15</td>
<td>766</td>
</tr>
<tr>
<td>25</td>
<td>80902</td>
<td>1</td>
<td>80902</td>
<td>73921</td>
<td>7</td>
<td>10561</td>
<td>37393</td>
<td>19</td>
<td>1969</td>
<td>16462</td>
<td>27</td>
<td>610</td>
<td>12376</td>
<td>15</td>
<td>826</td>
</tr>
<tr>
<td>27</td>
<td>86473</td>
<td>1</td>
<td>86473</td>
<td>79017</td>
<td>7</td>
<td>11289</td>
<td>39977</td>
<td>19</td>
<td>2105</td>
<td>17642</td>
<td>27</td>
<td>654</td>
<td>13267</td>
<td>15</td>
<td>885</td>
</tr>
</tbody>
</table>
Figure 2. Coefficient of variation (CV%) of $p$ under the negative binomial distribution with the dilution effect for different values of $m$ and $p$ for obtaining sample sizes with $c=0.0001$, $\lambda=0$, $\delta=0$, $w=0.0008$ with 95% probability of detecting individuals with AP.
RESULTS

Testing seed plans with the dilution effect, false positives, false negatives, lower quality limit (LQL), and an acceptable quality limit (AQL)

Figure 3. Operating characteristic curves for four different seed lot testing plans (with \( n=3200, \ m=8, \ \lambda=0, \ \delta=0, \ c=0.0001 \) and \( w=0.014 \)) with consumer risk LQL= 1.0% and producer risk AQL=0.5%.

Best testing plan is when \( g=40 \) and \( k=80 \) because it represents the lowest consumer (7.59 %) and producer (7.64%) risks.
RESULTS

Testing seed plans with the dilution effect, false positives, false negatives, lower quality limit (LQL), and an acceptable quality limit (AQL).

Figure 4. Operating characteristic curves for four different seed lot testing plans with detection limits of $c=0.00010$, $0.00012$, $0.00014$, $m=6$, $\lambda=0$, $\delta=0$ and $w=0.0088$ (with $n=1820$, $k=15$) and for the standard testing plan with consumer risk LQL=1.1% and producer risk AQL=0.4%.

The best testing plan is when $c=0.00012$ because it gives the lowest consumer (5.14%) and producer (6.92%) risks.
Conclusions

The method offers a strategy for forming groups (pools) from the sample that will be subjected to laboratory tests, and for determining an optimal number of pools that guarantees that if there is at least one group with AP, there is a high probability that it will be detected. The groups formed have an optimum size such that one element with AP will be detected at a low cost.

The **binomial** distribution should be used (modified Dorfman method) when it is known that the proportion of AP in the population is large, \( p > 0.1 \); otherwise, the **inverse sampling** method is recommended because it guarantees that \( m \) individuals or pools of individuals with AP will be detected with high probability.

A MatLab program is available for calculating the ideal sample size, number of pools (\( g \)) and size of the pools (\( k \)) to detect AP with probability \( (1-\alpha) \)
Conclusion

• Possible disadvantages of the method

→ it does not provide a closed solution for sample size, pool size, and total cost

→ the value of $w$ may be difficult to obtain (but an average value given by the % DNA per grain may be used); in the case of genetically modified plants, the weight of the enzyme or other protein formed by that specific DNA sequence may be estimated as a proportion of the total weight of the grain or as a % of DNA.