Statistical Analysis of the Comet-Assay

Bernd-Wolfgang Igl
Principal Statistician
Bayer AG
Berlin

PSI Toxicology SIG
Webinar
17th April 2018
Contents

- Biological background
- Experimental design and data
- Classical statistical strategies
- Comparisons of different slide summaries
- Results
- Outlook
Biological background

Test batteries in genetic toxicology

Different (genetic) tests measure different endpoints.
No single test is capable to detect all types of genetic damage.

Test battery

Routinely used test battery in genetic toxicology (see Option 1, ICH S2(R1)):
I. Test for gene mutation in bacteria (e.g. Ames test);
II. Cytogenetic test for chromosome damage (in vitro metaphase chromosome aberration test or micronucleus test) or in vitro mouse lymphoma Tk gene mutation assay;
III. An in vivo test for genotoxicity (generally micronucleus test or chromosomal aberration test with haematopoietic cells)
Other in vivo tests: UDS-Test, Pig-a test in vivo, Comet assay in vivo.
First attempts:
The single cell gel electrophoresis assay (SCGE) or Comet assay traces back to the 1980s, where

- first Ostling and Johnson (1984) showed that DNA strands migrate from the nuclei to the anode, if exposed to an electric field,
- then, Singh et al. did further pioneering works to optimize this principle and yield a “highly” sensitive, specific and also reproducible test method (1988).

Today:
The Comet assay represents a widespread technique for simple, sensitive, fast and economic measurement of DNA damage on a single cell level both in vitro and in vivo.

The in vivo Comet assay is often used as a 2nd in vivo test to follow-up positive in vitro test results (substitution of the UDS-test).
Biological background

Test principle

The fundamental **principle** of the Comet assay is based on the availability of **electrically charged DNA molecules**:

- the quantity of migrated DNA molecules is assumed to be directly related to the amount of induced DNA damage
- and
- the amount of induced DNA damage is assumed to be an indicator for the genotoxic potential of the test substance.

In other words,

the goal of the Comet assay is to assess possible treatment-related direct DNA damages in isolated cells by quantifying the amount of electrically charged, migrating DNA molecules.
Biological background

Test principle (cont.)

Single cell suspensions from substance-treated animals/cell cultures are

- embedded in an agarose gel,
- lysed to liberate the DNA,
- DNA is unwound to also DNA detect single-strand breaks (alkaline version)
- placed in an electric field, neutralization and staining.

DNA molecules are polar and therefore tend to migrate in an electric field.

- DNA fragments and DNA-loops (high-molecular structures) migrate quickly off the center of the nucleoid to the anode.
- Non-damaged or repaired DNA molecules show little to no observable migration.
**Biological background**

**Test principle (cont.)**

Migration of DNA fragments is responsible for the **comet like shape** of the electrophoresis image (tail), hence the name COMET assay.

The amount of migrated DNA is directly related to treatment-induced DNA damage.

At the end of the electrophoresis, the shape of the comet is analyzed. **Excess DNA damage** (i.e., damage beyond background damage that is normal in any cell) is visualized by a **long and/or dense tail**, whereas an **untreated cell** looks like a **homogeneous disk**.

Cells suitable for measuring are identified by the scientist, measurements were done (semi)automatically using specific softwares.
Biological background

Schematic representation and typical observations
Biological background

Parameters of interest

- head length (µm),
- **tail length** (µm),
- head intensity (% of total DNA),
- **tail intensity** (% of total DNA),
- **tail moment** (tail intensity x tail length),
- tail area.

Some of the measurements are related:

- the sum of the head intensity and the tail intensity is 100%,
- tail moment is an integrated product of the tail length and the tail intensity.

Int. Workshop on Genotoxicity Test Procedures (San Francisco, 2005): **tail intensity** is the widely favored measurement for statistical analysis (Burlinson et al., 2007).
Parameter of interest (cont.)

**Attractive features** of tail intensity:

- comparability across studies and laboratories,
- linear relation with DNA damage,
- wide range of damage that is capable of being robustly measured by image analysis software.

The tail intensity can theoretically amount to values between 0% (no DNA damage) to 100%.

**Increasing positive values** correspond to greater DNA damage.

Usually, nuclei with tail intensities larger than e.g. 85% are called “hedgehogs” and are only documented, but excluded from further analyses (Bright et. al, 2011).
Biological background

Some aspects of quality control
According to the OECD/OCDE guideline 489 (26th Sept. 2016) cells can be classified into three categories:

- scorables,
- non-scorables and
- “hedgehog”.

Only scorables with a unique defined head and tail (and no interference with neighbouring cells) should be scored for % tail DNA to avoid artefacts.

The amount of non-scorable cells must not be reported.
Biological background

Some aspects of quality control (cont.)

**Hedgehogs** (cells with a small or non-existent head, large diffuse tails of heavily damaged cells)

- represent dead or dying cells
- should be determined based on visual scoring (since automated image analysis fail in the absence of a clearly-defined head) and
- separately documented.

Moreover:

- cells close to the edge of the gel are not scored to avoid any potential “edge effects”.
- the scientist must measure every slide “blinded”, without knowledge of its treatment.

Standard practice in a regulatory environment (e.g. under GLP).
Biological background

Some advantages of the Comet assay

- quick, simple, economical
- basically every organ/cell can be analyzed (in vivo and in vitro tests)
  - local organ-specific genotoxicity
- several organs can be investigated within one experiment
- different versions are available for mechanistic analyses

Some disadvantages of the Comet assay:

- its high sensitivity can yield to high variability
- sometimes difficult to interpret
  - genetic relevance, statistics, ...
- toxic agents can lead to a downturn effect at high doses
There are a variety of experimental designs for investigations using the Comet assay (Lovell et al, 1999; Smith et al., 2008, Tice et al., 2000).

The determination of a proper design is extremely important for the statistical analysis of Comet assay data (e.g. Lovell et al, 1999; Wiklund and Agurell, 2003)).

In general, this includes the specification of:

- the experimental unit and measurement unit
- a suitable endpoint
- the distribution of measurements
- important sources of variability
- the numbers of cells, slides, animals, etc. to score
- an appropriate summary statistic (per slide, animal, ...), besides, appropriate “cut-offs” for categories of damage, a “blinded” scoring of cells, randomization, blocking and the use of historical control data.
Experimental design and data

The experimental unit and measurement unit:

For *in vitro* studies, the experimental unit is the *culture*.

for *in vivo* studies, the experimental unit is the *animal*.

In both cases, the measurement unit is the *slide*.

Endpoints:

In general, DNA strand breaks can be determined by independent endpoints such as:

- tail intensity *(primary endpoint)*,
- tail length,
- tail moment.

The % tail DNA (i.e. the tail intensity) is the primary endpoint and recommended for the evaluation and interpretation of results (OECD 489, JaCVAM).
Experimental design and data

Distribution of measurements

Distribution of individual cell data is often highly skewed incl. zeros and outliers

- mixed distribution?
- transformation?

(different for different target parameters)

Reasonable distributional assumptions are essential for specifying accurate values per measurement unit

What are relevant sources of variability?

- treatment,
- animals,
- cultures,
- slides,..
Experimental design and data

in vivo assay

A typical in vivo Comet study in the pharmaceutical industry:

five groups:
- a negative control (or “vehicle”);
- a low,
- a medium,
- a high dose of the test compound
- and a positive control (known to cause genetic damage).

Animals are randomly assigned to one of the five groups; each animal will receive a single or multiple doses of the test chemical (OECD 489, JaCVAM, Smith et al., 2008).
Experimental design and data

**in vivo assay**

According the current OECD guideline (OECD 489):

- **at least 5 animals** of one sex (per dose and per tissue) (PC: min of 3 animals)
- **at least 150 cells** (excluding hedgehogs) using
- **2 or 3 slides**.

Common tissues are the **liver, bone marrow, stomach** and **blood**. The liver is often of special interest because of its own metabolising capacity.

**Typical experimental design of an in vivo assay (per tissue):**

- **Vehicle** → **Low Dose** → **Medium Dose** → **High Dose** → **Positive Control**
- Each dose level is tested on at least 5 animals.
- Each tissue sample is evaluated on 2 or 3 slides.
- Each slide contains 50 cells for analysis.
Experimental design and data

**in vitro assay:**

One example of a typical experimental design of an *in vitro* assay:

![Diagram of experimental design](image)

**Note**

There are many more *in vitro* designs
Experimental design and data

Raw data on the cell level are summarized yielding one value per slide.

These are the data used for further statistical analysis.

In other words, the summary statistic per slide:

- median,
- arithmetic mean,
- geometric mean,
- 90% percentile, ...

is of fundamental importance for the final result.
Example of raw data *in vivo*

Tail intensities (lg scale):
- Vehicle, low dose, high dose, pc
- 6 animals per group
- 2 slides per animal
- 50 cells per slide

See Figure 3 in Bright et al. (2011)
Experimental design and data

Example of raw data *in vivo* (per slide) (cont.)

Many distributions, and mixtures of distributions, were analyzed to describe observed tail intensities (see e.g. Lovell D. and Omori T., 2008)

A widespread assumption is that

- tail intensities per slide can be adequately described by either a lognormal distribution or
- a zero-inflated lognormal distribution (i.e. plus a point mass at zero).

**Note:**

An increasing genotoxic treatment effect yields a decreasing number of small (or zero) observations, and the lognormal distribution will shift upwards (Bright et al., 2011).
Main statistical ingredients
The experimental design leads to a certain **hierarchical nature of the data** which must be noted when developing an appropriate statistical model (Lovell et al. 1999):

**Key** (fixed and/or random) **factors** of an inferential statistical model are:

- treatment group
- electrophoresis (?)
- animal (random?)
- slide effect (?)
- different exposure time points (?)

animal/culture (random effect?) (**mixed effect linear model**)  
animal/culture (nested within treatment group?) (**nested linear model**)
Classical Statistical Strategies

General considerations

The inferential statistical model shouldn’t involve the positive control group, neither for group comparisons nor for a trend analysis. Positive control data should be compared separately with vehicle data to assess the assay sensitivity.

Important statistical outputs for assessing treatment effects are:

- confidence intervals (one-sided),
- p-values (one-sided),
- effect sizes

and pairwise comparisons of treatment and vehicle groups and/or a trend test across all the test groups.

Confidence intervals, effect sizes, etc. support the determination of biological relevance, not the isolated presentation of single p-values!
General considerations (cont.)

Adjustments for multiplicity

This is a very hot topic in toxicology, safety pharmacology, etc.

No adjustments will control the type I error rate for only individual comparisons; adjustments for multiplicity focus on the whole experiment.

Consequently, adjusting the p-values will increase the false negatives (more genotoxic compounds will be wrongly labelled as safe);

not adjusting will increase the false positives (more safe compounds will be wrongly labelled genotoxic)

(but this also depends on the experimental design, the power and the repeatability of a test).

currently no consensus in safety assessments (e.g. Aylott et al., 2011).
General considerations (cont.)

Theoretically, one may want to fit a statistical model to the individual cell measurements.

... “statistical analyses based upon measures from the individual cells can lead to serious misinterpretation of results” (Lovell et al., 1999).

It is of fundamental importance to determine experimental and measurement units correctly (Wiklund and Agurell, 2003).

An adequate alternative would be to use a single summary statistic per slide yielding repeated measures per experimental unit (e.g. animal)

slide-to-slide variability or technical structures (e.g. slide 1 > slide 2 > ...)

A third strategy would be to aggregate the slide summaries for each animal into a single summary statistic

only the factor “treatment” can be analyzed using a gen. linear model
Classical Statistical Strategies

Slide summaries
Widespread summary statistics are

- arithmetic mean,
- geometric mean
- median (of original or (log-) transformed data)
- further percentiles (e.g. 90th).

(Log data: a small constant (0.001) will be added to each measurement prior to the calculations to avoid the problem of taking the logarithm of 0).

Summary per animal
If there are no systematic changes between replicated slides, then summarizing across slides per animal (for a given tissue) was appropriate by taking the usual arithmetic mean (of slide summaries) by animal (Bright et al., 2011).
Classical Statistical Strategies

Widespread inferential models for group comparisons \textit{(in vivo)}

1-way linear model, with

- “treatment“ as a single fixed factor (using the animal summary statistic as the response)
- with “treatment“ as a single fixed factor and repeated measurements (“slides“) per “animal“ (experimental unit)
- (Kruskal-Wallis (and Wilcoxon tests) as a nonparametric equivalent to the 1-way linear model)

2-way linear model, with

- “treatment“ and “electrophoresis“ as fixed factors
- “treatment“ and “animal(treatment)“ as fixed factors (“animal“ nested within “treatment“)
- “treatment“ as fixed factor and “animal“ as a random factor
Classical Statistical Strategies

Widespread inferential models for group comparisons (*in vivo*)

Common parametric strategies (cont.):

**3-way linear model**
- “treatment”, “electrophoresis” and “animal” (random or nested within “treatment“)
- “treatment”, “animal” and “slide” (“animal” nested within “treatment“, “slide” nested within “animal” and “cell” nested within “slide“)

**In general:**
- usually, models use **treatment groups as categorical variables**, 
- alternatively, a **linear regression** could incorporate **actual dose levels**
  (Bright et al., 2011; Lovell et al., 1999; Smith et al., 2008; Wiklund and Agurell, 2003).
Classical Statistical Strategies

Statistical strategies for trend analysis

Common statistical trend tests include a

- **linear contrast test,**
- **Williams’ test for trend** and the
- **Jonkheere–Terpstra test** (rankbased test for a monotonic trend in the responses with increasing dose groups)

(Bright et al., 2011). A further promising methodology is the **downturn protected trend test** (Bretz and Hothorn, 2001) using

- a predefined number of interesting **dose-response shapes,**
- **linear contrasts,**
- the **correlation structure** between different endpoints (e.g. tail intensity, tail moment, tail length) and
- simulation based adjustments for multiplicity, if desired.
Classical Statistical Strategies

Statistical strategies for trend analysis (cont.)

Downturn protected trend test to determine a trend in genotoxic effects in any of the (e.g. tail intensity, tail moment, tail length) parameters.

Specify all possible shapes incl. a down-turn effect

3 dose groups + 1 vehicle 11 contrasts

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Shape</th>
<th>Hypothesis</th>
<th>Shape</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_{C-} &lt; \mu_L = \mu_M = \mu_H$</td>
<td>$\cdot \cdot \cdot$</td>
<td>$\mu_{C-} &lt; \mu_L &lt; \mu_M &lt; \mu_H$</td>
<td>$\cdot \cdot \cdot$</td>
</tr>
<tr>
<td>$\mu_{C-} = \mu_L &lt; \mu_M = \mu_H$</td>
<td>$\cdot \cdot \cdot$</td>
<td>$\mu_{C-} &lt; \mu_L = \mu_M$</td>
<td>$\cdot \cdot \cdot$</td>
</tr>
<tr>
<td>$\mu_{C-} &lt; \mu_L &lt; \mu_M = \mu_H$</td>
<td>$\cdot \cdot \cdot$</td>
<td>$\mu_{C-} &lt; \mu_L &lt; \mu_M$</td>
<td>$\cdot \cdot \cdot$</td>
</tr>
<tr>
<td>$\mu_{C-} = \mu_L = \mu_M &lt; \mu_H$</td>
<td>$\cdot \cdot \cdot$</td>
<td>$\mu_{C-} = \mu_L &lt; \mu_M$</td>
<td>$\cdot \cdot \cdot$</td>
</tr>
<tr>
<td>$\mu_{C-} &lt; \mu_L = \mu_M &lt; \mu_H$</td>
<td>$\cdot \cdot \cdot$</td>
<td>$\mu_{C-} &lt; \mu_L$</td>
<td>$\cdot \cdot \cdot$</td>
</tr>
<tr>
<td>$\mu_{C-} = \mu_L &lt; \mu_M &lt; \mu_H$</td>
<td>$\cdot \cdot \cdot$</td>
<td>$\mu_{C-} &lt; \mu_L$</td>
<td>$\cdot \cdot \cdot$</td>
</tr>
</tbody>
</table>

Table is adapted from Table 1 in Hothorn, 2004.
Classical Statistical Strategies

Statistical strategies for trend analysis (cont.)

- The most plausible contrast is the one yielding the smallest p-value ("minimum p-value approach").

Choose that contrast to describe the proposed relationship.

Note, under the null-hypothesis, multiple contrast tests follow a central multivariate t-distribution.

A bootstrap approach is used to be more robust from deviations from normality.

In order to involve the correlation structure of the data, different parameters (e.g. tail intensity, tail length, tail moment) can be analyzed simultaneously.
Classical Statistical Strategies

Statistical strategies for trend analysis (cont.)
All of the previous mentioned tests can **easily be implemented in SAS®,**
e.g.

PROC MIXED,
PROC GLIMMIX,
PROC MULTTEST,
PROC NPAR1WAY,...
Comparisons of different slide summaries

The summary-measure of data per slide is of fundamental importance for the inferential statistical outcome of the assay.

Various publications use mean or medians per slide, of originally or log scaled data (e.g. Wiklund & Agurell, 2003; Bright et al., 2011).

Sometimes, the test result mainly depends on the chosen slide summary and changes from positive to negative and vice versa (e.g. Hobbs et al., 2015)

T. Tug (2018) addresses the question, whether there are substantial differences for the final test result, if

- medians or
- geometric means

were used as slide summaries.
Comparisons of different slide summaries

Simulated tail intensity observations: log-normal distribution per group plus “outlier”

- green vertical line: chosen offset
- green arrows are simulated outliers
- procedure for each slide
- no animal and slide effect
- “outlier”: max 5 observations per slide (randomly determined),
- values stem from a log-normal distribution with a “huge” variance (plus an offset)
- For each setting: 10,000 simulations

Tug, 2018
Comparisons of different slide summaries

Basic simulation setting

Parametric 1-way linear model,

with

I. “treatment“ as a single fixed factor (using the animal summary statistic as the response)
II. with “treatment“ as a single fixed factor and repeated measurements ("slides") per “animal“ (experimental unit)

with pairwise comparisons of dose to control groups.

All computations are done in SAS® 9.4.
Comparisons of different slide summaries

It turns out that there are

- no substantially liberal or conservative effects under $H_0$
- considerable differences in power of at least 10% under $H_1$
  (power of using the geometric mean > power of using the median)
- almost no differences between strategies, if effects are large

The final statistical result might depend on whether a median or geometric mean is used

- Future works might involve further statistical strategies incl. additional:
  - data distributions
  - fixed and random factors
  - ...

April/17/2018

Statistical analysis of the Comet assay
Comparisons of different slide summaries

- Differences in power up to 15% (GeoMean > Median)
- Increasing power with increasing number of slides

Tug, 2018
Results

Usually, Comet assay data cannot be assumed to be normally distributed (Bauer et al., 1998; Wiklund and Agurell, 2003; Tug, 2018).

The response distribution is usually highly skewed and often contaminated with a certain amount of outliers, sometimes plus a point mass at zero (Bright et al. 2011; Tug, 2018).

A lognormal assumption (e.g. plus a point mass at zero or a certain outlier contamination) seems to describe the response distribution adequately (Bright et al. 2011; Tug, 2018).

Experimental (EU) and measurement units (MU) must be defined properly for developing an adequate statistical model and all “relevant“ sources of variability must be addressed (e.g. EU: animal, MU: slide).
Simulation studies show that the choice of an adequate **slide summary** is extremely important for an efficient analysis (Wiklund and Agurell, 2003; Tug, 2018).

The differences in power between different statistical models (**under** $H_1$) are mainly based on the choice of different summary measures (tail intensities) per slide:

- **mean** or **median (of log data)** are clearly superior to any of the other measures;
- the **arithmetic mean of (original measurements)** leads to a very inefficient analysis with almost no power;
- the **median of (original measurements)** seems to be little sensitive to detect moderate dose-related increases (Bright et al. 2011; Tug, 2018; Wiklund and Agurell, 2003).
Results

The statistical analysis should be performed with slide as the unit of measurement and not on the cell level, which can lead to serious misinterpretations of results (Lovell et al., 1999).

The experimental design leads to a certain hierarchical nature of the data, which must be noted when developing an appropriate statistical model, e.g. a nested effects linear model (Lovell et al., 1999; Wiklund and Agurell, 2003).

A higher (e.g. 2 or 3-way) hierarchical model is to be preferred as it describes the nature of the data optimally (Wiklund and Agurell, 2003).

No single statistical method is ultimately correct, but some methods are wrong. In particular those which ignore the hierarchical nature of the data (Lovell et al., 1999).
Results

In the case of more than one tissue of interest, each should be analyzed separately.

The inferential focus of the results should be on the confidence intervals for the treatment effects to support biological relevance and not on single p-values.

Confidence intervals (and also p-values) should be one-sided and usually, comparisons are not adjusted for multiplicity (no consensus, remains a point of discussion) (Bright et al. 2011; Tug, 2018; Wiklund and Agurell, 2003).

In the present settings non-parametric statistical strategies turned out to be generally less efficient than linear models (Wiklund and Agurell, 2003).
Results

Simulations showed that statistical power increase substantially with increasing no. of cells up to 50 per slide, a further increase until 100 cells is not worthwhile (in vivo and in vitro).

*In vivo*: the effect of an additional slide on the statistical power is larger than a further doubling of the no. of cells (to 100) or adding a further animal to the test groups (tail moment and tail length).

*In vitro*: for tail length data the effect of adding a slide from 2 to 3 is similar to the effect of increasing cells from 25 to 50; for tail moment measurement the corresponding slide effect is larger than a cell effect.

Of course the practical feasibility must be taken into account, since it is „easier“ to increase the number of slides than cultures or animals.

(Tug, 2018; Wiklund and Agurell, 2003).
One **dominant factor** for the detection of a treatment effect might be the definition of an **adequate slight summary**.

"Further studies are encouraged on this apparently trivial, but obviously vital, issue of how to summarize the measurements from cells on a slide" (Wiklund and Agurell (2003)).

Do we need **different/additional summary measures** per slide?

- trimmed means,
- censored means,
- weighted means,
- semi-parametric measures,...
References


References


References


Thank you for your attention!

Special thanks to:
Andreas Sutter (Head of Genetic Toxicology, Bayer AG, Berlin)
Timur Tug (University of Dortmund)
Christina Ziemann (Fraunhofer ITEM, Hanover)

Comments, criticism, etc. are always welcome!
bernd-wolfgang.igl@bayer.com
June 19th, 2018

Prof. Dr. Ludwig A. Hothorn
Institute of Biostatistics
Leibniz University Hannover, Germany

Recent developments of statistics in toxicology-
using R