Problem Solving: DNA Data Acquisition and Analysis

By Dr. David Werrett[1], Richard Pinchin[2] and Ros Hale[2]

[1]Forensic Science Service, Birmingham B5 6QQ, UK

[2]Forensic Science Service Metropolitan Laboratory, London SE1 7LP, UK

In this paper we address some of the issues that arise when setting up and developing a large DNA database. We will also show how the knowledge gained during the process can be captured in an expert system, which we have called STRess.

MULTIPLE SUPPLIERS OF PROFILES TO A NATIONAL DNA DATABASE

In the UK, the National DNA Database data is owned by the Association of Chief Police Officers (ACPO). They provide the framework for suppliers, such as the Forensic Science Service (FSS), to submit information to the database. In addition to its role as supplier, the FSS acts as custodian for the database by the administration of a proficiency test program. Those supplying profiles to the database must take part in proficiency testing, and potential suppliers must take part in a number of validation tests. The FSS advises the ACPO (on the science used, match criteria, etc.), and the ACPO builds this information into the approval and framework for supply of profiles to the database.

The main feature of the validation and proficiency test program is that all suppliers must be accredited by an external organization to ISO9001 and ISO25 international standards, which include the standards of the National Measurement and Accreditation Service, NIS46 and NIS96. Accredited laboratories can then supply samples to the National DNA Database and take part in the ongoing quality assurance program administered by the FSS.

EMPIRICAL PROCESS FLOWS

The rapid development of the National DNA Database unit within the FSS made us acutely aware of the need for detailed examination of the process flows required to ensure that samples are analyzed correctly, results interpreted and, when necessary, samples re-analyzed. The entire process is illustrated in Figure 1.

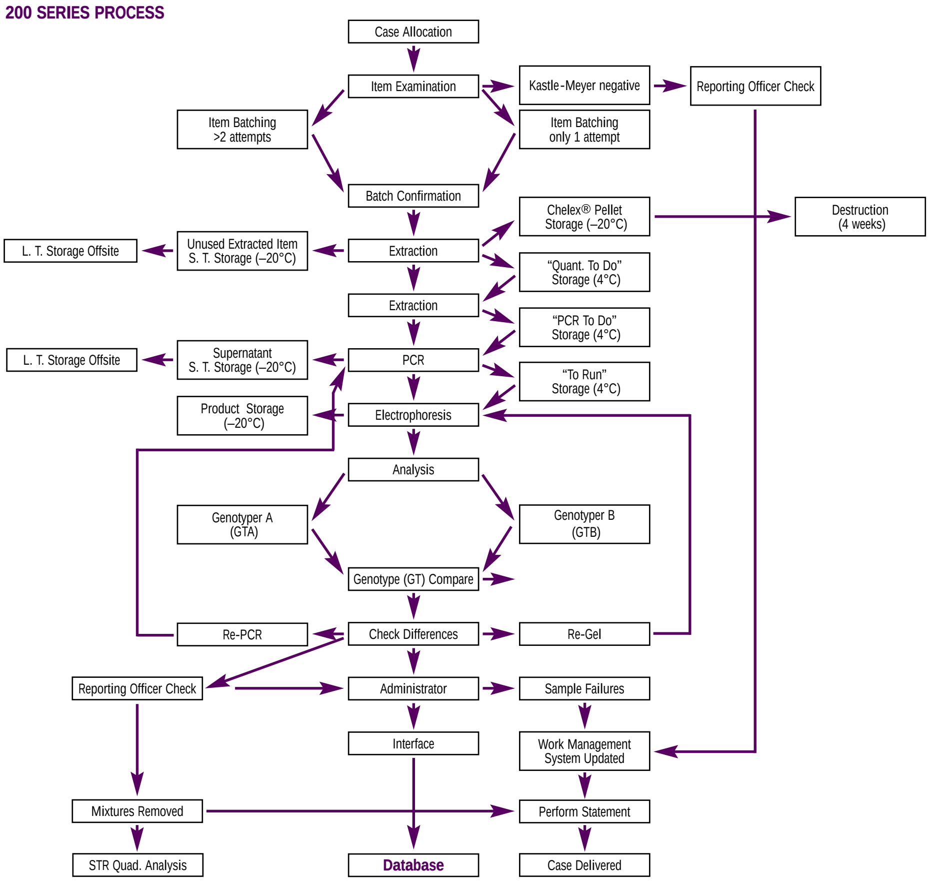


Figure 1. Overall process flow. The entire process, from case allocation to database submission of STR profiles is illustrated in this flow chart.

At the genotyping stage there was some room for subjectivity, and therefore, two analysts (signified by GTA and GTB in Figure 1) genotype the gel independently. These two interpretations are examined by a third individual (signified by GT Compare in Figure 1) who performs a comparison of the genotyping results. Any differences are checked, and decisions are made to either accept the result or re-amplify or rerun the sample. Each part of the process flow was examined in detail, documented and timed. This initiated two projects. The goal of the first of these projects was to automate many of the manual techniques involved in the process, including extraction, quantitation and PCR. The second project sought to address the considerable time spent analyzing and genotyping gels. These efforts led to the development of a program or "expert system" known as STRess (STR expert system suite). This is a Windows®- and Macintosh®-based program that accepts raw data, generates a file of allele designations and then compares this file to one generated by a human operator. The following pages detail the development of the STRess system, and provide an overview of how the system works.

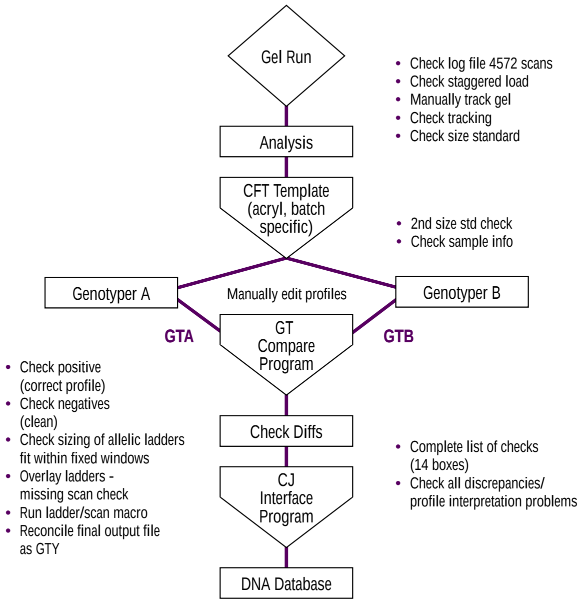

Gel Run
Analysis
CFT Template (acryl, batch specific)
Genotyper A
Manually edit profiles
Genotyper B
GTA
GTB
GT Compare Program
Check Diffs
CJ Interface Program
DNA Database
• Check log file 4572 scans
• Check staggered load
• Manually track gel
• Check tracking
• Check size standard
• 2nd size std check
• Check sample info
• Check positive (correct profile)
• Check negatives (clean)
• Check sizing of allelic ladders fit within fixed windows
• Overlay ladders - missing scan check
• Run ladder/scan macro
• Reconcile final output file as GTY
• Complete list of checks (14 boxes)
• Check all discrepancies/ profile interpretation problems

Figure 2. Detail of the genotyping process.

THE STRESS SYSTEM

Having visualized the overall process flow (Figure 1), we were in a position to focus on the particular area or domain that we believed was a potential bottleneck. We decided to concentrate on the part of the process outlined in Figure 2 – the independent analysis of the data by two individuals, genotyper A (GTA) and genotyper B (GTB).

THE OBJECTIVE

The primary objective was to provide a system that would carry out the same analytical processes as a human, to at least the same or higher standard. An additional benefit would be a decrease in processing time, but this would have to be achieved without sacrificing quality. Quality monitoring was an important part of the specification; this required a complete and documented audit trail.

The benefit of the development of this program was the achievement of an increase in throughput by making the most of available human resources and, in turn, providing value to our customers. In addition, it provided a structured process enabling us to ensure quality and allowing easy problem tracing and subsequent solving. The only remaining issue was how to implement such a system.

THE CHALLENGE

The challenge was to provide a computer program that could be installed on the same Macintosh® computer as the STR analysis software and that was capable and intelligent enough to undertake the role of the second person in the process, genotyper B (GTB). In addition, the same program needed to run in an IBM® PC environment.

It became clear that the best solution was the development of an expert system. These days there is less fear of the term "expert systems", even though in the late 1980s they were seen as a universal corporate panacea that would allow the replacement of expensive, highly skilled staff with a less skilled and thus less expensive workforce. Fortunately, this has turned out to be neither practical nor desirable, and today, expert system technology is considered to be a flexible framework for holding all the relevant information about a domain (e.g., data, knowledge, other programs, reference materials, etc.). The best description we have seen to date is that expert systems are regarded as "technological glue".

Having identified a potential solution, we initiated the STRess project. After defining the domain of interest, the second stage in the construction of an expert system is the formalization of the relevant knowledge. With around 100,000 samples already processed, we felt we had acquired enough expertise to formulate rules by which our human operators worked.

It should be pointed out at this stage that this process of "knowledge engineering" is a valuable exercise in itself. Many things that are done during laboratory procedures have evolved from simple beginnings into highly complex processes. Taking a critical look at each task and asking the questions, "Why am I doing this?" and "Do I really need to do this?" at each stage often reveals inherited redundancy and, more importantly, can reveal that all-important error waiting to happen.

THE GOAL

It became the primary goal of the STRess project team to encapsulate the knowledge of the human operator (GTB) and produce a system that could perform that person's job function to an equal or greater level of competency.

The overriding concern throughout the project was the preservation of data integrity and the maintenance of a zero-error philosophy even at the expense of additional resources. We realized that to accomplish this there would be a price to pay in terms of extra staff time during the initial stages following implementation. In fact, overall processing time did increase initially, but we were confident that this cost would be more than recouped over the long term.

SYSTEM OVERVIEW

The STRess system accepts data from either of the Applied Biosystems programs, GenoTyper® or GeneScan® Analysis, in the form of a comma separated values (CSV) file. It applies rules and processes derived from the experts to this raw data and produces a number of output files. The file of designated alleles is in the same CSV format as that produced by the human operator so

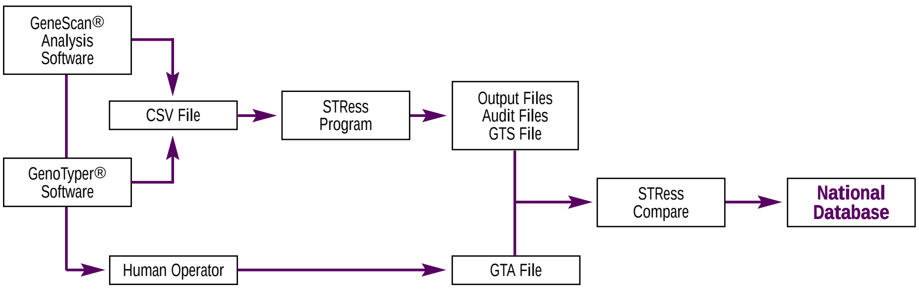


Figure 3. Outline process flow for STRess.

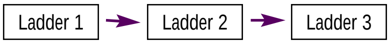
the "compare" function of the STRess program can be used to detect any differences between the two files. This is illustrated in Figure 3.

SYSTEM DETAIL

Figure 4 shows the processes invoked by the STRess program. Process 1 simply accepts data in CSV format. These data are produced from GeneScan® software by exporting a data table or from GenoTyper® software by running a raw data macro. This raw data file defines each peak in terms of position, height and area. Data from GeneScan® software are by color in peak order and data from GenoTyper® software are by locus and sample.

Process 2 is the heart of the system and is responsible for cleaning the raw data ready for allelic designation. The process has been split into five components as follows:

1. Negative control lanes are checked for contamination or primer dimers.
2. Ladders are checked for artifacts and non-allelic peaks; these are removed before proceeding to Step 3.
3. Allelic ladders are compared and any differences are reported. This step will reveal any missing peaks. If there are more than two ladders present, they are compared in the order:



4. Sample lanes are cleaned using the rules contained in the knowledgebase. Examples of the rules used are shown in Figure 5.

The underlying philosophy of the system is to move data from one file to another (rather than remove the data altogether). This allows a clear audit trail that can show the fate of every peak from the input file.

Once the sample data have been cleaned, the remaining peaks have to be designated. Strictly speaking, this needs to be done by reference to the ladder lanes present on the gel. However, this presents a number of problems:

a) The gel ladder will be shifted from an ideal.

b) The gel ladder will be incomplete – not all possible alleles will be represented.

c) There may be missing ladder peaks.

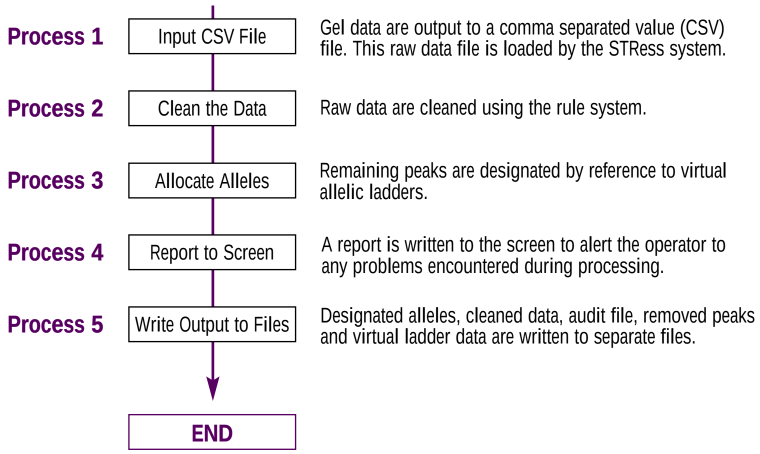

Process 1
Input CSV File
Gel data are output to a comma separated value (CSV) file. This raw data file is loaded by the STRess system.
Process 2
Clean the Data
Raw data are cleaned using the rule system.
Process 3
Allocate Alleles
Remaining peaks are designated by reference to virtual allelic ladders.
Process 4
Report to Screen
A report is written to the screen to alert the operator to any problems encountered during processing.
Process 5
Write Output to Files
Designated alleles, cleaned data, audit file, removed peaks and virtual ladder data are written to separate files.
END

Figure 4. STRess processes.

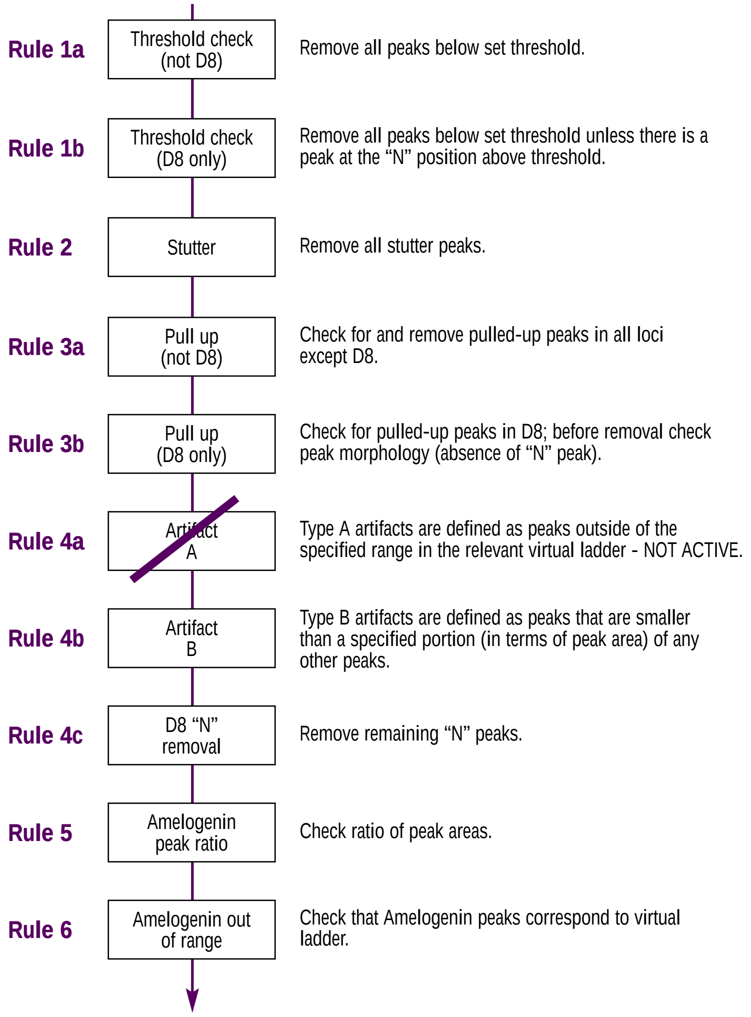

Rule 1a
Threshold check (not D8)
Remove all peaks below set threshold.
Rule 1b
Threshold check (D8 only)
Remove all peaks below set threshold unless there is a peak at the "N" position above threshold.
Rule 2
Stutter
Remove all stutter peaks.
Rule 3a
Pull up (not D8)
Check for and remove pulled-up peaks in all loci except D8.
Rule 3b
Pull up (D8 only)
Check for pulled-up peaks in D8; before removal check peak morphology (absence of "N" peak).
Rule 4a
~~Artifact A~~
Type A artifacts are defined as peaks outside of the specified range in the relevant virtual ladder - NOT ACTIVE.
Rule 4b
Artifact B
Type B artifacts are defined as peaks that are smaller than a specified portion (in terms of peak area) of any other peaks.
Rule 4c
D8 "N" removal
Remove remaining "N" peaks.
Rule 5
Amelogenin peak ratio
Check ratio of peak areas.
Rule 6
Amelogenin out of range
Check that Amelogenin peaks correspond to virtual ladder.

Figure 5. Example knowledgebase.

5. To circumvent these problems, STRess constructs a "virtual ladder." This is done by comparing the gel ladder with a known pattern of peaks determined when the acrylamide gel mix is validated. The shift between ideal ladder and the gel ladder is determined at each peak as shown by δ_1 in Figure 6, Panel A.

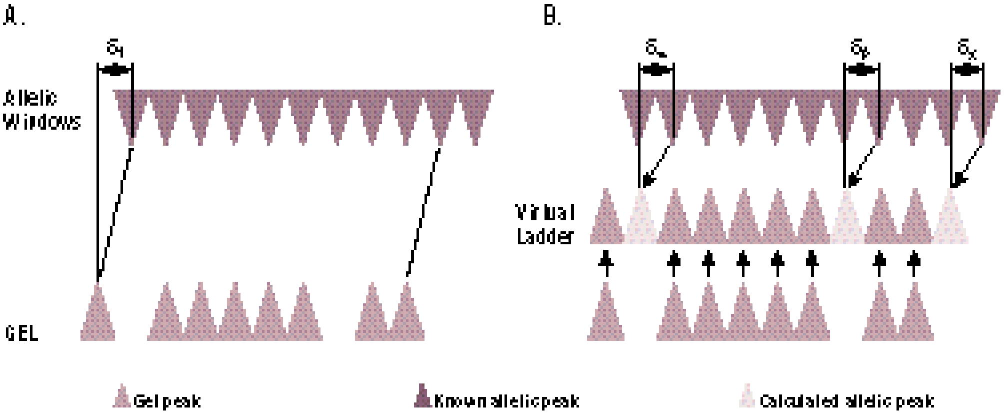


Figure 6. Panel A. Shift of gel peaks from ideal. Panel B. Creation of the virtual ladder.

This shift is then used to compensate for missing peaks. Thus, the virtual ladder is built up from true gel peaks and peaks calculated for the observed shift. This process is carried out for each ladder on the gel and can be visualized as shown in Figure 6, Panel B.

Following the creation of the virtual ladder, the remaining peaks can be designated. This is done by reference to a list of all possible designations. As this list is under the control of the user, labels can be added to indicate such things as rare alleles. Any peak that does not have a corresponding virtual ladder peak can either be ignored or designated by a question mark.

Once the designation phase is complete, customized comments can be added depending on a range of post-designation rules – this process is known as allelic qualification. Figure 7 shows some of the qualifications used by the FSS.

OUTPUT

Once processing is complete, a number of files are generated.

GTS file - As mentioned earlier, this file is equivalent to the human operator output file (GTA) with which it is compared using the STRess compare function (see next section).

OUT file - This contains the cleaned data in the format of the input file (i.e., each peak is defined by its position, height and area).

REP file - This reports the fate of each peak discarded.

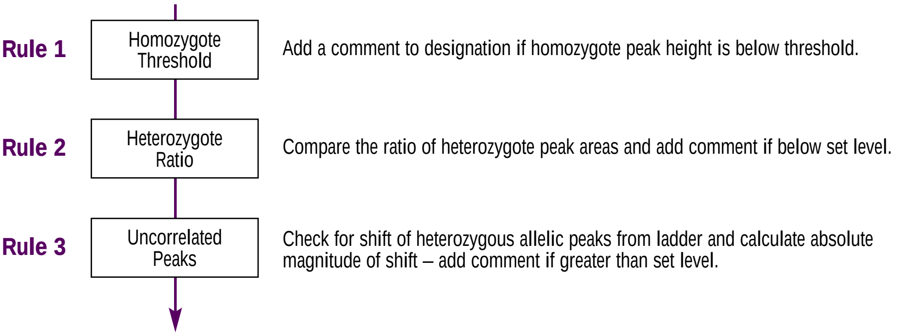
AUD file - The main audit file details all warnings issued during processing together with the operator ID. This file also contains virtual ladder summary statistics.

VLD file - This is a full listing of every peak in each of the virtual ladders together with its shift from the ideal ladder.

Each file contains a list of the version numbers of each of the numerous files that comprise the STRess environment. In this way version control is strictly monitored.

COMPARE

The final stage is to compare the human output with that from STRess. The STRess compare function produces a table of differences that allows the “check diffs” (see Figure 2) to investigate the causes of any differences and arbitrate before sending the profile to the National Database.



Add a comment to designation if homozygote peak height is below threshold.

Compare the ratio of heterozygote peak areas and add comment if below set level.

Check for shift of heterozygous allelic peaks from ladder and calculate absolute magnitude of shift – add comment if greater than set level.

Figure 7. Some qualification rules used by the FSS.

TROUBLESHOOTING

The processing of large numbers of samples (this year we will process in excess of 200,000 samples, and next year we project up to 300,000 samples) has presented us with several novel problems. We needed a troubleshooting structure that would allow us to address and document problems in each of the three FSS units performing DNA analysis and learn by them. We now have a dedicated troubleshooting procedure whereby problems are identified, documented and brought to the attention of a troubleshooting committee, which has the responsibility for implementing and following up on corrective actions as well as organizing post-implementation reviews.

It is hoped that we can all learn from an exchange of information on problem resolution as databases are implemented throughout Europe and around the world.

SUMMARY

By studying thousands of sample operations of the system and comparing them to the human operator, rules have been refined and the program tuned for maximum efficiency. To date, a saving in time of more than 30% has been achieved by use of the STRess program. This is an important saving considering that the FSS has almost 200 people in eight teams at two locations processing about 20,000 samples per month.

Chelex is a registered trademark of Bio-Rad Laboratories. GenoTyper and GeneScan are registered trademarks of the Perkin-Elmer Coroporation. IBM is a registered trademark of International Business Machines Corporation. Macintosh is a registered trademark of Apple, Inc. Windows is a registered trademark of Microsoft Corporation.