How we Crunch the Numbers

The number-crunching exercise isn’t as complicated as it looks at first glance. Yes, there are some very sophisticated technical aspects to the analysis, but **the basic gist of the exercise can be simplified quite easily** as follows:

Each test result is a string of 37 numbers which correspond to the scores on the 37 markers. The genetic relationship of two participants can be assessed by comparing the test scores of one to the test scores of the other. If the test scores are exactly the same on any marker, then the mismatch for that marker is zero. However, if the scores are not exactly the same on any marker, then the mismatch for that marker is calculated as the difference between the scores (always expressed as a positive number). When all of the markers have been individually checked, **the Genetic Distance of the two participants is simply the sum of all the individual mismatches for each marker**.

For example, lets suppose that two participants “A Flan” and “B Flan” have exactly the same test scores on 35 markers but different scores on

2 markers; lets say “A Flan” scores 24 on DYS390 but “B Flan” scores 23, and “A Flan” scores 29 on DYS449 but “B Flan” scores 31. The mismatch on DYS390 is 1 (24 - 23), and the mismatch on DYS449 is 2 (31 - 29). The Genetic Distance between “A Flan” and “B Flan” is 3 (1 + 2).

Genetic Distance (GD) is probably the most common way of describing a genetic relationship, using the “total mismatch”, but another way is to express the result as a “total match” in quotient form. So, using the previous example of “A Flan” and “B Flan”, the genetic relationship would be GD = 3 or a match of 34/37. The latter can be useful to avoid ambiguity since it highlights that the genetic relationship is based on 37 markers (as opposed to 12, 25 or 67 markers, which are also sometimes used). However it is important to bear in mind that 34/37 does not mean that 34 markers matched but rather that there was a total mismatch of 3 spread across all 37 markers.

Special Markers

Most of the 37 markers are fairly basic and allow a simple comparison to be made. However, there are 6 “special” markers which are a bit tricky.

In hindsight, it would have been preferable if these special markers hadn’t been used by the lab but they provide useful information and just need to be used with care to avoid miscounting the mismatches.

2 special markers **(DYS389i and DYS389ii)** are a “nested pair”. That means that the “inner” marker (DYS389i) can mutate independently, and can be counted separately, but the “outer” marker (DYS389ii) always mutates in tandem with the “inner” marker. This can lead to a single mismatch being counted twice. In order to avoid the error, a “new” marker DYS389B is calculated as DYS389ii – DYS389i. **The genetic relationship uses DYS389i and DYS389B, and does not use DYS389ii.**

4 special markers **(DYS464a, DYS464b, DYS464c and DYS464d)** are multiple markers which, for technical reasons, are always reported in ascending order. This makes it impossible to make a direct comparison since the reported ascending sequence “abcd” might be based on test sequence “dcba” or “badc” etc. In context, this is really only a problem when using the markers to evaluate unrelated population groups where there is a reasonable expectation that the markers will be significantly

mismatched. In our own project, which examines a distinct population group (or perhaps, several distinct population groups), the consistency of matching markers elsewhere informs us that these markers are also likely to match. On that basis, **it seems reasonable to use them at face value within each distinct group** but to treat them with caution when comparing different groups.

Groups

It is immediately apparent from even the most casual review of the project test results that there are "clusters" of results which are very closely matched. Each cluster suggests a close kinship group descended from a recent common ancestor. In very general terms, **the project has yielded 4 groups so far: 2 large groups and 2 small groups, but all from the same genetic tribe known as R1b**. Each group appears to stem from a common ancestor who lived outside the scope of modern historical records. There are additional clusters or subgroups within the large groups. Each subgroup appears to stem from a common ancestor who liven within the scope of modern historical records and, in some cases, the historical documentation has been found to support this contention.

Sorting Groups

The project results are easily sorted into groups by placing all of the results into one large table. Each table column represents the test result for a particular marker (such as DYS393 etc.), and each table row represents the string of results for a particular participant. For the sake of privacy, participants remain anonymous during their lifetime and their test results may only be identified by kit reference number (such as #13653 etc.).

A final row is added to the table for the sake of computing the Modal Value of all results in the project. This is simply a theoretical result based on the most commonly occurring test score for each marker. So, for example, if there are 150 participants, and 100 participants score 13 on DYS393 and 50 participants score 12, then the Project Modal Value (PMV) will score 13 on DYS393.

The PMV is also the Group Modal Value of the largest distinct group. Each project result may be compared to the PMV by calculating the Genetic Distance (GD). Sorting the GD values in ascending order, the gradual increase in values from zero upwards eventually hits a quantum leap or “bump”. The bump usually defines the largest group, and the closely matching results form the first and largest group. The results of the largest group may then be removed from the project table, and a new PMV calculated for the remaining results. The second PMV will be the Group Modal Value of the second largest distinct group, and so on. Eventually, the process defines four groups, and a small number of ungrouped results that must be presently separately until a subsequent match with a future participant forms an additional group.

Sorting Subgroups

The group results are easily sorted into subgroups by placing all of the group results into a table, along with the Group Modal Value (GMV). Whilst it is possible to sort subgroups using numerical analysis, it is far easier to simply assign a colour code to the test scores for the GMV, and then leave all mismatched results uncoloured. The resulting pattern of “white” results is visually conspicuous and the results are very easy to group together, as subgroup results will all share a common “white” mismatch that remains coloured in all of the other group results. In most cases, documentary research confirms that the subgroup participants share a common ancestor in modern times, or traces their ancestry to the same general location (which indicates a common ancestor who lived just outside the scope of modern historical records).