Woodward Informatics Ltd

C$WILDNA1:
DNA STR Awareness for Oracle!

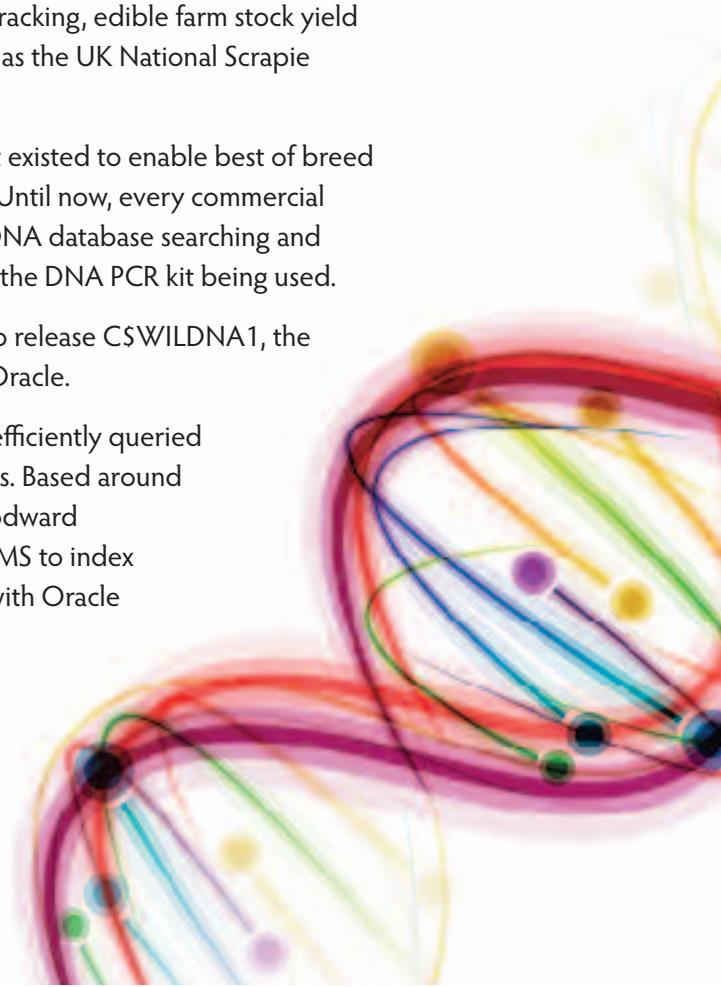
Introduction

DNA databases are often thought of as large repositories of contentious government controlled information used by the police and forensic scientists to aid in the solving of major crime such as murder, rape, and sexual abuse. DNA profile use is however far more widespread with well known citations that include a) identity verification of high profile war criminals such as Saddam and Uday Hussein, b) confirmation of the source of semen stains on a dress owned by Monica Lewinski, c) through the use of a reference DNA sample provided by Prince Philip the Duke of Edinburgh, confirmation that the remains of the last Russian tsar and family have been found, and d) body and body part pooling and victim identification after mass disasters such as the 9/11 attacks on New York or the 2004 Boxing Day Tsunami off the west coast of Sumatra. Less topical applications of DNA profile searching include insurance fraud, immigration and border control, biometric identity cards, and paternity and alimony cases. Agricultural applications include stock farm-to-shop meat tracking, edible farm stock yield improvement efforts, or ethnic cleansing programmes such as the UK National Scrapie Eradication Plan.

Until now, a single generic shrink-wrapped product has not existed to enable best of breed DBMS systems such as Oracle to become DNA STR aware. Until now, every commercial organization or academic institute that required in-house DNA database searching and manipulation functionality, had to roll their own specific to the DNA PCR kit being used.

To address this void, Woodward Informatics Ltd is proud to release C$WILDNA1, the first COTS product to add DNA STR typing awareness to Oracle.

C$WILDNA1 enables data containing DNA profiles to be efficiently queried within the DBMS for exact, partial, or near-match profile hits. Based around Oracle's extensible database indexing technology, the Woodward Informatics Ltd Data Cartridge C$WILDNA1 allows the DBMS to index and efficiently manipulate DNA STR profile data on a par with Oracle



native data types such as **VARCHAR**'s, **NUMBER**'s, and **DATE**'s. The interface to C$WILDNA1 remains SQL; any technology that can currently utilize the rich functionality of Oracle can use C$WILDNA1, eg. C# or VB.Net, Perl, PHP, C, C++, or Java.

Out of the box, C$WILDNA1 can accommodate many existing STR typing systems including YFiler, SGM, and AmpFℓ STR® SGM Plus (Applied BioSystems) currently used by the UK National DNA database), MeowPlex (Promega), an STR marker system for the domestic cat, and the StockMarks (Applied Biosystems) equine, canine, and bovine genotyping systems. C$WILDNA1 even accommodates for trisomy and user configured, could be used for any known STR typing application, possibly even alien DNA!

C$WILDNA1 Usage

C$WILDNA1 is a component that enhances the Oracle DBMS feature set so that it becomes STR multiplex aware. The database result set returned by an Oracle SQL **SELECT** statement, data bound to one of several C$WILDNA1 operators, requires no post processing with other database logic or middleware – the SQL query result set will satisfy the query predicates provided without necessity to perform resource intensive post result retrieval filtering operations!

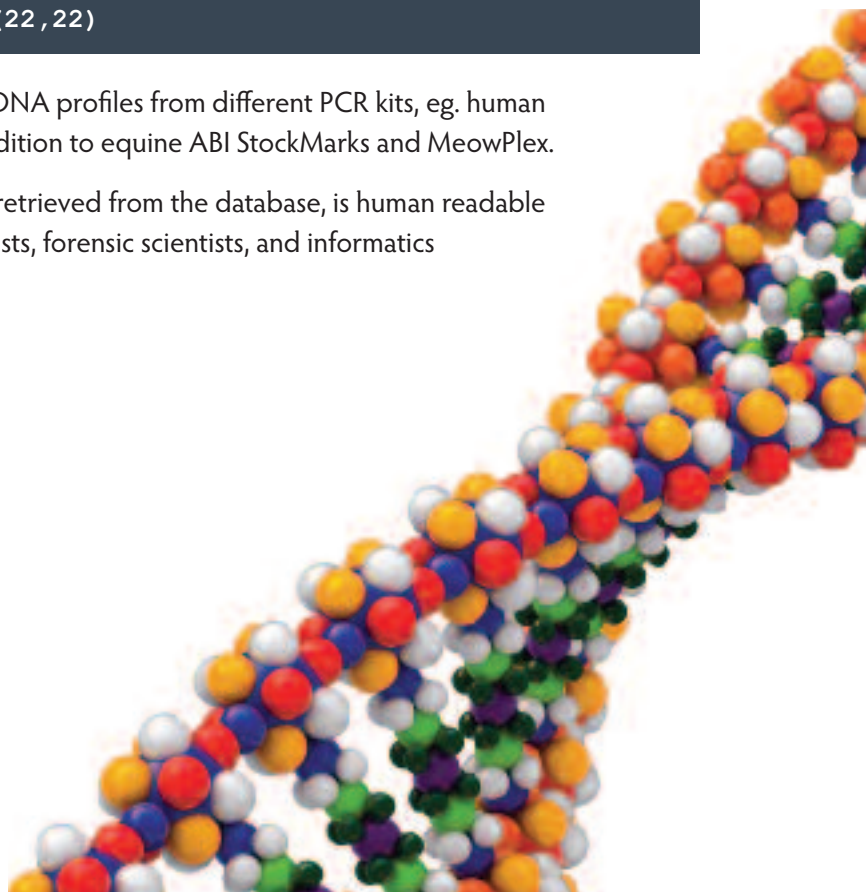
DNA Profile Representation

DNA profiles indexed and manipulated by C$WILDNA1 are represented as human readable text stored within Oracle as **VARCHAR2**'s in a single column of one or more database tables. Representative SGM Plus profiles excerpts, as they would be populated by third-party applications ultimately through SQL **INSERT** statements into Oracle tables, are included below:

```
1. D16 (15,19); Amel (X, X); TH01 (9.3,F)
2. TH01 (9, 9.2); Amel(X, Y); vWa (17,18)
3. D21 (30, 31); D8(12,13); FGA (22,22)
```

Each database table column can even contain DNA profiles from different PCR kits, eg. human SGM Plus, YFiler, and PowerPlex profiles, in addition to equine ABI StockMarks and MeowPlex.

Every DNA profile, as it is populated into and retrieved from the database, is human readable expressed in a form already familiar to geneticists, forensic scientists, and informatics practitioners.



DNA Profile Loading

No stored procedures or packages are required in the target database schema nor is there any overhead of database triggers and associated logic on any database tables or table columns. Data is inserted into the underlying database tables using familiar SQL **INSERT** statements, eg.

```
INSERT INTO tableName(columnName)
 VALUES('TH01 (9, 9.2); Amel(X, Y); vWa (17,18)')
```

and all database indexes are updated upon the transaction being committed. The interface to the DBMS remains SQL meaning that application developers and DBA's alike need not learn a new and complex third-party API. The implication should also be clear that C$WILDNA1 could be quickly retrofitted to an existing bioscience organizations LIMS system with little new software development effort!

Exact Match DNA Profile Searching

C$WILDNA1 provides the ability to efficiently search the tables containing STR profile data through the use of the **wildNAExactMatch** operator. Again the *lingua franca* remains SQL – no complex middleware or wrapped Java stored procedures to complicate usage, impact on performance, or introduce dependencies.

```
SELECT *
 FROM tableName
  WHERE wildNAExactMatch(columnName,'D21(30, 31);D8(12,13); FGA (22,22)')>0
```

Using the functionality of the C$WILDNA1 **wildNAExactMatch** operator, the **SELECT** statement returns all the table contents retained that exactly match the DNA profile '**FGA (22,22);D21 (30, 31) ; D8(12,13); ...**'. Note both the white space and the STR designation order shown above are different yet they still satisfy the exact match predicate and all relevant hits be efficiently included in any database result set.

The exact match operator functionality provides easy to use functionality for determining DNA database genotype STR *hits* or for confirming *exclusion*.

Near and Partial Match Searching

An important requirement of any DNA database is to provide the forensic scientist or informatics practitioner with the ability to efficiently search their database for near or partial DNA profile matches. This functionality is provided by C$WILDNA1 implemented through the two operators **wildNANearMatch** and **wildNAPartialMatch**.

```
SELECT *
 FROM tableName
  WHERE wilDNANearMatch(columnName,' D21(30, 31);D8(12,13); FGA (22,22)',1)
SELECT *
 FROM tableName
  WHERE wilDNANearMatch(columnName,' D21(30, 31);D8(12,13); FGA (22,22)',2)
```

In these two examples, the C$WILDNA1 **`wilDNANearMatch`** operator filters down the result set to only those DNA profiles that match the DNA profile queried with one or two degrees of freedom respectively. In the first example, the population database search result set might include DNA profiles and other details (individual name, date of birth) for STR genotypes that match '**`D21(31, 31);D8(12,13); FGA (22,22);...`**' and '**`D21(30, 31);D8(11,13); FGA (22,22);...`**' but not '**`D21(31, 31);D8(11,13); FGA (22,22);...`**' (the latter example showing STR variation with two degrees of freedom/variation). The result set however from execution of the second SQL statement would include '**`D21(31, 31);D8(11,13); FGA (22,22);...`**'.

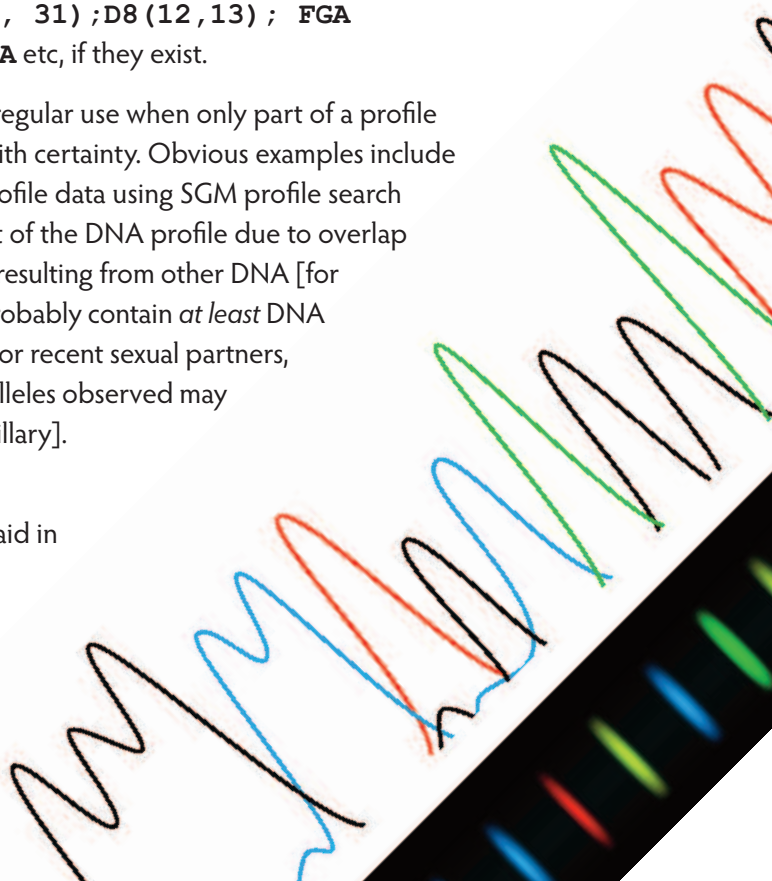
Application of these database operators in forensic and agricultural genotyping LIMS processes are numerous and include laboratory good-practice and accreditation compliance practices to ensure, for example, that samples being manipulated concurrently through the laboratory workflow have not contaminated the sample under investigation, that samples are not contaminated with the DNA of the laboratory worker, or that occasional human designation errors that occur during the interpretation of PCR Gels or capillary electrophoresis results are avoided.

```
SELECT *
 FROM tableName
  WHERE wilDNAPartialMatch(columnName,' D21(30, 31);D8(12,13); FGA (22,22)')
```

In this example, the **`wilDNAPartialMatch`** operator filters down the result set to only those DNA profiles that contain designations **`D21(30, 31);D8(12,13); FGA (22,22)`** with any other values for **`D02, D16, FGA`** etc, if they exist.

Application of these database operators will also find regular use when only part of a profile from a mixed profile is known or can be determined with certainty. Obvious examples include a) prospecting a STR database containing SGM Plus profile data using SGM profile search criteria, or b) prospecting a STR database for only part of the DNA profile due to overlap and contamination of the reference sample with peaks resulting from other DNA [for example, a vaginal swab taken from a rape victim will probably contain *at least* DNA from two people, the victim (the reference sample), prior recent sexual partners, and the rapist(s); the reference sample and other STR alleles observed may contaminate/overlap with each others in the Gel or capillary].

Finally, as biological/genetic relatives will also share considerable DNA, these database operators will also aid in the confirmation or exclusion of familial relationships between the individuals with application that includes sibling and paternity testing.



System Requirements and Performance

Hardware/Operating Systems

C$WILDNA1 will run on any Oracle DBMS edition ≥10.1, an instance of any Oracle edition (EE, SE, SE1, or XE) deployed on a number of operating systems and hardware architectures, for example x86 Windows, Linux x64, or SPARC 64 bit Solaris.

Performance

Performance will vary widely and depend on Oracle edition, host operating system, available memory, cached results, tuning and configuration parameters, and other DBMS process overhead. The following performance metrics were obtained from a test database instance loaded with 1 million SGM Plus DNA profiles. The unoptimised host server was an Amazon virtual EC2 instance running Oracle 11g SE1 on RHEL (Linux) x64, with 4GB RAM. Typical corporate enterprise servers would be expected to be of a much higher specification.

Operation	Time*
Index a database table containing 1,000,000 SGM Plus DNA profiles with C$WILDNA1	47min
wilDNAExactMatch	1.7sec
wilDNANearMatch; match for any 18 of 22 alleles specified	1.5sec
wilDNANearMatch; match for any 17 of 22 alleles specified	1.5sec
wilDNANearMatch; match for any 10 of 22 alleles specified	1.9sec
wilDNAPartial; match for specified 10 alleles	1.8sec
wilDNAPartial; match for specified 20 alleles	1.9sec
**All reported figures are averages for 250 similar operations*	

About Woodward Informatics Ltd

Woodward Informatics Ltd develops and supports bespoke Informatics software primarily for organisations in the Scientific or Life Sciences sector. The company was founded in 2011 by Dr. Michael D. O'Shea, a skilled Informatics Practitioner, Business Analyst, and Software Engineer. Dr O'Shea has published in many of the worlds most prestigious peer reviewed scientific journals and brings his first hand involvement in this bleeding edge R&D, in addition to his considerable experience as a Scientific Software Developer working on projects for bigpharma and bioscience organisations, to the portfolio of scientific IT services and products on offer at Woodward Informatics.

Woodward Informatics Limited, 17 South Row, Chilton, Oxfordshire OX11 0RT, United Kingdom
Tel +44 (0)7402 851 508 Web www.strychnine.co.uk Email woodwardinformatics@strychnine.co.uk
Company Registered in England No 07613939