How did the study come about?

The Western Australian Family Connections Genealogical Project was proposed by Dr John Bass and implemented in 2003. The project aim is to create and store a system of links representing genealogical relationships for the residents of Western Australia (WA) to be used as a research tool in conjunction with health data to help investigate familial factors in health and disease. The project exists as a supplementary system of links to the WA Data Linkage System (WADLS) that regularly links records across several population-based data sets.1 When used in conjunction with this extensive collection of health-related data, the genealogical properties can enhance genetic and familial research projects to, for example, assess the degree of relatedness between individuals within study samples, assist in locating common ancestors and allow estimates of genetic risk.

WA has developed particularly strong capabilities in the linkage of population-based data over the past three decades. Since 1995 it has supported a Data Linkage System,2 which has the primary aim of creating, storing and retrieving electronic links between records from core population-based systems, originating as early as 1966, and totalling nine data sets: birth, death and marriage registrations, electoral roll, hospital morbidity, emergency department presentations, mental health information, midwives notifications and cancers.1 Probabilistic matching techniques with extensive clerical review are used to identify records for the same individuals within and between the data sets. Dynamic updates of links for entries that relate to the same individuals are stored in a master links file. All links are created, stored and managed by the WADLS, but the detailed data remain the responsibility of the separate data custodians.

The Family Connections project was initially funded by the Medical and Health Research Council of Western Australia, and additional funding to support staff has been received from the Telethon Institute for Child Health Research. Salary for the project manager is provided as a Team Investigator on a National Health and Medical Research Council Capacity Building Grant (#254545) awarded to The University of Western Australia. Approval for Phase I of the WA Family Connections project was given by the Human Research Ethics Committee of The University of Western Australia, and by the Confidentiality of Health Information Committee which reviews applications for access to confidential data at the Health Department of Western Australia.

What does it cover?

The inheritance patterns of common diseases such as asthma, type 2 diabetes, cancer and cardiovascular disease are complex, involving multiple genes, many interactions and often strong effects of environmental determinants.3–6 It is difficult to identify causative genes, predict individual risk, estimate phenotypic outcome or propose intervention strategies based solely on genetic information. To increase the ability to detect susceptibility genes and to understand the role of environmental influences, researchers need to utilise both family-based and population-based study designs. Large, population-based information systems containing comprehensive health data; pedigree
and links to biospecimen samples are increasingly being sought as the most appropriate infrastructure to facilitate research. Genealogical information is a key component of human genome epidemiological research and enables the incorporation of biological relationships in risk assessments and epidemiological modelling. The WA Family Connections project can be used as a resource for projects where risk assessments and familial relationships are needed.

Few population-based genealogy registers exist due to the challenges of developing and maintaining the commitment to such a resource on a large scale. Two other registers exist that contain genealogical properties for a regional population; the Icelandic genealogic database managed by deCode Genetics, and the Utah Genealogical Data Base. Both include data for current residents and have the ability to combine the genealogy with data in disease registers. Other existing genealogical registers are only applicable to historical populations, are not population-based, or cannot be linked to disease registers (Table 1).

### Table 1: Characteristics of existing large genealogical databases

<table>
<thead>
<tr>
<th>Population</th>
<th>Present population size</th>
<th>Total historical population with genealogical links</th>
<th>Year range</th>
<th>Total genealogy available</th>
<th>Linkage to disease registers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utah</td>
<td>2 300 000</td>
<td>2 200 000</td>
<td>From early 19th century</td>
<td>Two centuries</td>
<td>Births, deaths, cancer</td>
</tr>
<tr>
<td>Iceland</td>
<td>295 000</td>
<td>716 000</td>
<td>From 9th century</td>
<td>11 centuries</td>
<td>Hospital and clinic records (20 years). Tissue from approximately 80 000 donors</td>
</tr>
<tr>
<td>Quebec</td>
<td>7 600 000</td>
<td>3 000 000</td>
<td>1600–1940</td>
<td>3.5 centuries</td>
<td>Data retrieval possible for temporary linkage with health information</td>
</tr>
<tr>
<td>Pennsylvania</td>
<td>12 000 000</td>
<td>295 000</td>
<td>From 18th century</td>
<td>Three centuries</td>
<td>Collection of clinical information and access to the population is available for approved projects</td>
</tr>
<tr>
<td>Hutterites</td>
<td>40 000</td>
<td>38 000</td>
<td>Earliest birth year is 1690</td>
<td>Up to 13 generations. Relatively complete back to 1981</td>
<td>Some genetic, medical and anthropometric data since 1995 are contained on the database. Able to link to existing Canadian disease registers such as mental health and cancer registries</td>
</tr>
<tr>
<td>Estonia</td>
<td>1 400 000</td>
<td>10 000 tissue donors</td>
<td>Prospectively collected</td>
<td>Unspecified</td>
<td>Medical records, questionnaire data and tissue samples</td>
</tr>
</tbody>
</table>

Information; and links to biospecimen samples are increasingly being sought as the most appropriate infrastructure to facilitate research. Genealogical information is a key component of human genome epidemiological research and enables the incorporation of biological relationships in risk assessments and epidemiological modelling. The WA Family Connections project can be used as a resource for projects where risk assessments and familial relationships are needed.

Few population-based genealogy registers exist due to the challenges of developing and maintaining the commitment to such a resource on a large scale. Two other registers exist that contain genealogical properties for a regional population; the Icelandic genealogic database managed by deCode Genetics, and the Utah Genealogical Data Base. Both include data for current residents and have the ability to combine the genealogy with data in disease registers. Other existing genealogical registers are only applicable to historical populations, are not population-based, or cannot be linked to disease registers (Table 1).

### Who is in the sample, and what is attrition like?

The resident population of WA now exceeds two million people, representing about 10% of the total Australian population. The state growth rate is approximately 1.2% per annum and over the last decade, there have been on average 25 300 births, 11 200 deaths and 10 600 marriages per annum. The WA Family Connections project aims to include as many current and historical WA residents as possible, with a current focus on all births dating back to 1950. The use of retrospective registrations will allow inclusion of an important proportion of the historical population, some of whom may have since migrated from WA, but who will be significant ancestrally to the rest of the population.

The genealogical links are made using information from original birth, death and marriage registrations. No information is known about adoptions or divorces. The birth registrations are central to defining parent–child relationships and theoretically, pedigrees may be produced for the population using birth registrations alone. However, marriage registrations are used to confirm the union between two adults and name changes, and the death and midwife records give additional information that may help to identify the relationships between offspring and parents (Figure 1). Extra information on name and address changes over time, and thus confirmation of identity, is also gained from other records in the WADLS.

Each person is represented by a chain of one or more health-related events within the WADLS and the genealogical links are a supplementary network.
used to connect individuals into family relationships. Pedigrees can be constructed for specific individuals, with selection options such as including only those individuals who are biologically but not legally related, or those who are related by a certain coefficient of relatedness, defined as the proportion of alleles held in common by two related individuals.\textsuperscript{21}

The project has no attrition rate because it is population-based and does not rely on voluntary participation. The data come from state birth, death and marriage registrations that are already received at the WADLS for other linkage purposes. There will, however, always be gaps throughout the genealogical matrix for the population, essentially for people not born in WA, those who only live in WA for a short

**Figure 1** Example of how genealogical links are created between individuals based primarily from information on birth registrations (a), supplemented with information from death, marriage and midwife registrations (b)
period of time, and for those people not in contact with any of the core data sets. While the majority of residents come in contact with at least one of the WADLS linked data sets at some point of their time living in the state, there are some individuals, mostly males born prior to 1974 or outside of WA, who have not yet come in contact with any of the data systems and therefore no data will be available for these persons. Currently, this has resulted in approximately 3% of persons for whom a genealogical link cannot be made between those people and their offspring and will remain until such time as they appear in one of the data sets that are linked regularly at the WADLS. A further 3% of individuals from the electronic birth registrations appear to have too little information (including blank paternity fields) or incorrectly recorded data on their records to enable linkage and hence remain unlinked. Recent initiatives to include nationally collected Medicare data (data from doctor visits and pharmaceutical prescriptions) within the core WADLS may in time allow some of these unlinked records to become linked, as persons living in other Australian states who use these health services can be identified. Despite the gaps, generating a genealogical matrix for the remainder of the population represents significant numbers of people on which to conduct genealogical-based research, either using the whole available population or a selected cohort.

How often have they been followed up?

The success of the project relies on complete population ascertainment. Residents are not contacted for their permission to participate in the project and are not contacted in response to the genealogical data that has been created. It would not be feasible, and there would be insufficient resources, to contact the millions of individuals, many of whom have left WA or have become deceased, to invite them to participate. The project aims to include as much of the historical population as possible, and to continue adding information from new registrations for as long as the allocated resources allow.

| Table 2 Outline of the first two phases of the WA Family Connections project, as at May 2007 |

<table>
<thead>
<tr>
<th>Registration Type</th>
<th>Year range</th>
<th>Approximate number of records</th>
<th>Progress</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1 (electronic records)</td>
<td>Births (mothers, fathers and children)</td>
<td>1974–present</td>
<td>782,000</td>
</tr>
<tr>
<td></td>
<td>Deaths (parents and children of the deceased)</td>
<td>1984–present</td>
<td>237,000</td>
</tr>
<tr>
<td></td>
<td>Marriages (brides, grooms, both sets of parents)</td>
<td>1984–present</td>
<td>241,000</td>
</tr>
<tr>
<td>Phase 2 (paper records)</td>
<td>Births (mothers, fathers and children)</td>
<td>1950–1973</td>
<td>426,000</td>
</tr>
<tr>
<td></td>
<td>Deaths (parents and children of the deceased)</td>
<td>1950–1983</td>
<td>228,000</td>
</tr>
<tr>
<td></td>
<td>Marriages (brides, grooms, both sets of parents)</td>
<td>1950–1983</td>
<td>252,000</td>
</tr>
</tbody>
</table>

What has been measured?

Phase 1 of the project involves creating genealogical links from information recorded on electronic birth registrations that are available since 1974 and electronic death and marriage registrations available since 1984 (approximately 1.26 million records) (Table 2). Birth registrations include parent names; death registrations include parent and offspring names; and marriage registrations include parent names of both marriage partners.

Phase 2 will entail encoding genealogical links from earlier birth, death and marriage registrations currently held as paper records. The WA Registry of Births, Deaths and Marriages stores registrations dating back to 1841, but initially, due to the large amount of resources needed to computerise and link information from the paper records, the project will focus on birth registrations made between 1950 and 1973, and death and marriage registrations made between 1950 and 1983 (approximately 0.91 million records).

Phase 3 will incorporate a public appeal to improve the completeness of the Family Connections project data. Initial priority will be given to specific demographic groups where missing data are encountered most frequently. A public campaign will be launched to invite members of the general public to come forward with their own genealogical information, which will then be stored with data collected from official records.

The WA Family Connections project is governed and managed using the same advisory and committee structures developed for the WADLS. The entire WADLS system is designed to maximise the protection of individual privacy while providing access to information for approved health research projects. The genealogical links are represented by unique identifiers, and no actual data are stored in the WADLS or with the genealogical indices. Access to information from the WADLS may be permitted for bona fide research projects under the guidance of strict access arrangements. Requests for extracts using linked data must include clearance through an institutional ethics committee and the local Confidentiality of Health Information Committee, and have approval from all contributing data.
What has it found? Key findings and publications

The data that have been generated within the WA Family Connections project are available for research use, provided the necessary ethical and permission requirements are met. At this stage, projects have been restricted by the limited number of generations that are currently available and as a result, applications have only requested data for recent cohorts of children. Formal requests have primarily sought sibling relationships as an indication of whether siblings are full or half-siblings, when previously it had been unknown to researchers whether siblings born to the same mother had the same fathers.

In its current form, the WA genealogical system may be used to support both genetic and environmental epidemiological research projects utilising up to 3-generational pedigrees. Examples of familial research that could be supported by the system include the investigation of parental health history for children with obesity; the analysis of family structure following vasovasostomy procedures; the prevalence of psychiatric illness among parents of children diagnosed with autism; or the effect of close-living grandparents as a function of social structure following vasovasostomy procedures; the prevalence of psychiatric illness among parents of children diagnosed with autism; or the effect of close-living grandparents as a function of social support on the development of postnatal depression in new mothers. A limited number of 4-generation pedigrees within the WA population are currently available and when expansion to 5- or 6-generation pedigrees is accomplished, there will be a significant increase in ability to locate common ancestors and conduct research on adult populations.

Potential genetic epidemiological applications for the genealogy system include risk assessments based on population prevalence, the influence of family history on disease risk estimates, or the effect of family history on outcome/survival following drug therapy or other medical interventions. The system structure also enables it to be used as a sampling frame for genetic research studies, to locate high-risk families for genetic linkage and genotype association studies, to find families with a certain pedigree structure, or to select controls who are not closely related to case subjects or who have a family history of a particular disease. Current protocols require that consent of individuals is obtained before researchers may contact potential study participants.

In WA, over 200,000 tissue samples that were originally collected for diagnostic and research purposes are currently stored within laboratories and other facilities. A central register is being constructed to catalogue information about the type and location of tissue samples available for research. A WA Genome Health Project has also been proposed, which aims to collect detailed health data and biospecimens from a large number of WA residents. Links between the WADLS, the WA Family Connections project and tissue registers are possible with appropriate ethical approval and compliance with legal requirements. This represents significant potential to analyse genetic variants with respect to phenotype or patterns of disease history for the WA population.

To be successful in collecting and using data, genetics-based projects require careful stewardship and consultation to maintain positive public engagement from individual participants, health consumer groups and the general community. We have conducted various forms of community outreach, including seminars, consumer representation on committees, direct consultation, and information that has been made publicly available on websites dedicated to the WA population data collections, linkage capabilities, and research priorities. A critical component of our initiatives is the continuation of our ‘Genomics, Society, and Human Health’ program to address ethical, legal and social issues, maintain community outreach and to ensure community involvement in developing and managing these valuable resources.

What are the main strengths and weaknesses?

The most significant strength of the WA Family Connections Project is its relationship to the data linkage activities operating at the WADLS. The system covers all residents or visitors who are in contact with health-related services in WA and has the ability to be linked to other external datasets. Most of the core health data sets have been collecting data for at least 25 years which reduces the considerable time lags associated with achieving improved population health outcomes in prospective cohort projects and the passage of time necessary to enrol participants, collect data, accrue incident cases and analyse outcomes. It also reduces the amount of resources needed to support and sustain the project. Another advantage of the project is the use of privacy protocols that allow efficient data linkage while providing high levels of individual privacy protection in data extracts, thus increasing the amount of research projects that the system can support.

The most significant disadvantage of the project is the limited amount of electronic information that is currently available for creating the genealogical links. The unavailability of computerized birth records prior to 1974 restricts the ability to arrange WA residents in a genealogical framework much beyond three generations. The project will be unable to create genealogical links for births prior to 1974 until the relevant data are computerized and released. The project is
also somewhat disadvantaged by the inability to contact participants to confirm information about relationships or explain gaps in the matrix.

Can I get hold of the data? Where can I find out more?

Applications to use the genealogical data in health research projects are welcomed. All projects must have appropriate ethical clearance and permissions from the data custodians through the formal application process at the WADLS. Institutional ethics approval should be current and it is a requirement that a Western Australian-based researcher must be included in the research team. Interested researchers are encouraged to read the information contained on the WADLS website and questions can be directed to the manager or leader of the Family Connections project.

Conflict of interest: None declared.

References