

# Using Self-Reported Data to Determine Relatedness in Biobank Subjects

Luke Rasmussen<sup>1</sup>, Stephen Turner, MS<sup>2</sup>, Carol Waudby, MS<sup>1</sup>, Marylyn Ritchie, PhD<sup>2</sup>, Catherine McCarty, PhD, MPH<sup>1</sup>

<sup>1</sup>Marshfield Clinic Research Foundation, Marshfield, WI <sup>2</sup>Vanderbilt University, Nashville, TN

Contact

Luke Rasmussen  
BIRC-ICR / ML8  
1000 N Oak Ave.  
Marshfield, WI 54449

rasmussen.luke@mcrf.mfldclin.edu  
http://www.marshfieldclinic.org/birc

## Introduction

The Personalized Medicine Research Project (PMRP)<sup>1</sup> of Marshfield Clinic is a biobank of approximately 20,000 individuals. Subjects are asked during enrollment to list first degree relatives who (at the time of enrollment) live in the study catchment area, which research coordinators attempt to link to patients in the Marshfield Clinic system.

This question asks you to provide information about your close blood relatives who live in the study area and are at least 18 years old, and who are eligible to participate in the project. This information will allow us to determine family relationships among study participants in order to integrate genetic data with the clinical and laboratory data we collect. We will use this information to determine genetic relationships between family members and to determine the distribution of genetic variation within a population, and to determine the relationship between genetic and medical information. This information will be added to the research database to protect confidentiality. Investigators will not contact your relatives based on the information you provide below.

The study area for this project is the area of Marshfield, Wisconsin, and the surrounding area. The study area includes the following zip codes: 54401 Marshfield, 54402 Marshfield, 54403 Marshfield, 54404 Marshfield, 54405 Marshfield, 54406 Marshfield, 54407 Marshfield, 54408 Marshfield, 54409 Marshfield, 54410 Marshfield, 54411 Marshfield, 54412 Marshfield, 54413 Marshfield, 54414 Marshfield, 54415 Marshfield, 54416 Marshfield, 54417 Marshfield, 54418 Marshfield, 54419 Marshfield, 54420 Marshfield, 54421 Marshfield, 54422 Marshfield, 54423 Marshfield, 54424 Marshfield, 54425 Marshfield, 54426 Marshfield, 54427 Marshfield, 54428 Marshfield, 54429 Marshfield, 54430 Marshfield, 54431 Marshfield, 54432 Marshfield, 54433 Marshfield, 54434 Marshfield, 54435 Marshfield, 54436 Marshfield, 54437 Marshfield, 54438 Marshfield, 54439 Marshfield, 54440 Marshfield, 54441 Marshfield, 54442 Marshfield, 54443 Marshfield, 54444 Marshfield, 54445 Marshfield, 54446 Marshfield, 54447 Marshfield, 54448 Marshfield, 54449 Marshfield, 54450 Marshfield, 54451 Marshfield, 54452 Marshfield, 54453 Marshfield, 54454 Marshfield, 54455 Marshfield, 54456 Marshfield, 54457 Marshfield, 54458 Marshfield, 54459 Marshfield, 54460 Marshfield, 54461 Marshfield, 54462 Marshfield, 54463 Marshfield, 54464 Marshfield, 54465 Marshfield, 54466 Marshfield, 54467 Marshfield, 54468 Marshfield, 54469 Marshfield, 54470 Marshfield, 54471 Marshfield, 54472 Marshfield, 54473 Marshfield, 54474 Marshfield, 54475 Marshfield, 54476 Marshfield, 54477 Marshfield, 54478 Marshfield, 54479 Marshfield, 54480 Marshfield, 54481 Marshfield, 54482 Marshfield, 54483 Marshfield, 54484 Marshfield, 54485 Marshfield, 54486 Marshfield, 54487 Marshfield, 54488 Marshfield, 54489 Marshfield, 54490 Marshfield, 54491 Marshfield, 54492 Marshfield, 54493 Marshfield, 54494 Marshfield, 54495 Marshfield, 54496 Marshfield, 54497 Marshfield, 54498 Marshfield, 54499 Marshfield.

Please complete the following table for any of your immediate blood relatives who live in the above zip codes and are at least 18 years old. Write down the name or maiden name of your relative. Do not include the names of adopted children or adopted siblings, or of stepchildren or step-siblings.

Last Name	First Name	Sex	DOB (month/day/year)	Zip Code	Relationship (check one)
					Parent
					Sibling
					Child
					Other

Please proceed to the next page for additional entry sheets.

Thank you for participating in Personalized Medicine Research Project

Figure 1. Page from the enrollment questionnaire that asks the participant to list blood relatives

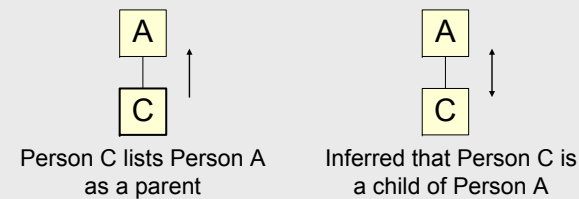
While self-reported relationships have some known limitations such as unreported relationships, reporting non-blood relatives (i.e. listing an adopted child as a child), and data entry errors by research coordinators, this data can be used in lieu of genetically verified relationships to discover subject relatedness. While other researchers have compared self-reported family relationships to genetic data for a targeted population<sup>2</sup>, we further the use of such data through the creation of a software application to address underreported relationships and data entry errors.

## Methods

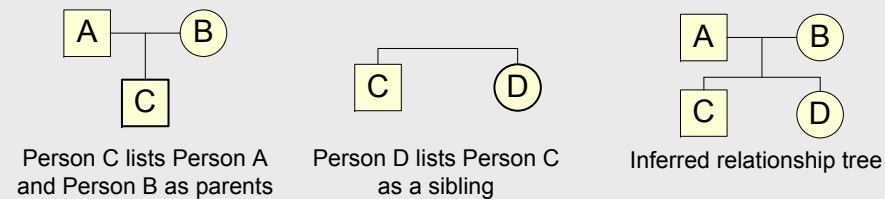
Self-reported family relationship data were processed by a custom software application to create a pedigree for a subset of PMRP subjects in a genome-wide association study. The application addressed different problems from self-reported data (see the panel "Deriving Relationships"). Identity by descent (IBD) analysis was done on the genotyped data for these subjects, and used as the gold standard to verify the pedigree.

## Deriving Relationships

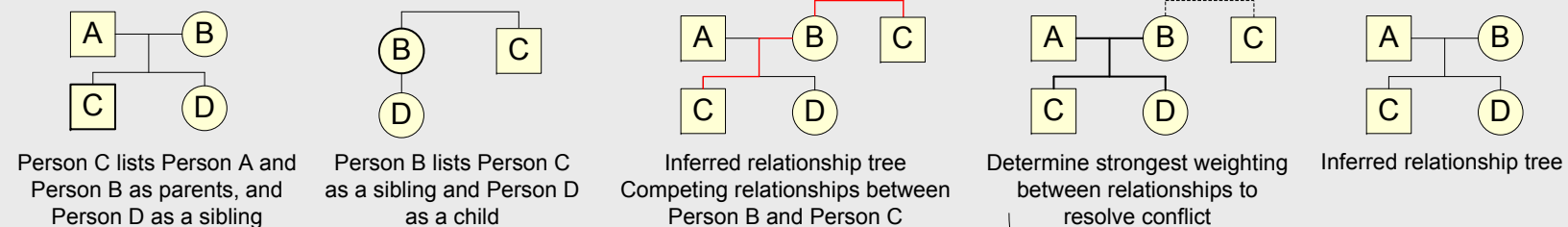
### Establishing Bi-Directional Relationships



### Building Full Relationship Trees From Partial Data



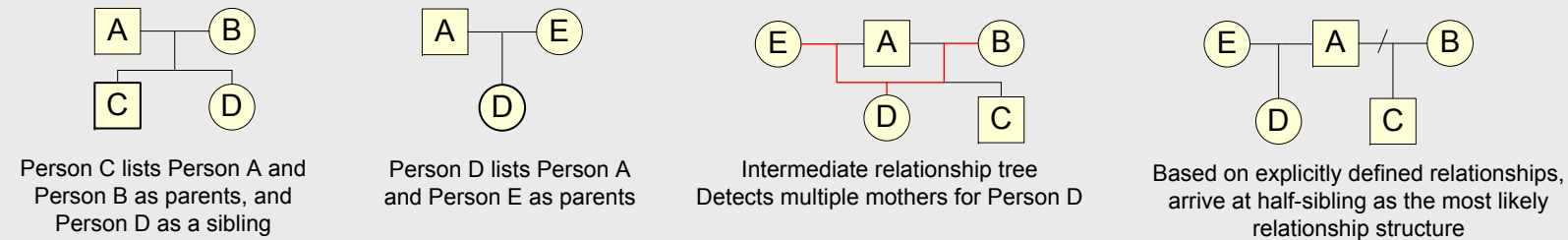
### Resolving Conflicting Relationships



If the system is not able to arrive at a conclusion, it flags suspicious relationships for manual intervention. In the course of this study, research coordinators took the list of unresolvable relationships and manually corrected them. Some of the causes of errors were:

- Data entry errors from questionnaire (i.e. listing relationship as "sibling" instead of "parent")
- Selecting the wrong person as the relative (i.e. multiple "Jane Smiths")
- Listing unrelated individuals as blood relatives (i.e. step-mother as mother)

### Determining Half-Sibling Relationships



## Results

The generated pedigree file contained 3,968 individuals from the PMRP population.

194 of the 1,300 relationships (14.9%) were created by the software using the methods described: 151 for siblings, 26 for parents and 17 for children.

	Total	Siblings	Parent/Offspring
Overall:	1,300	820	480
Missed:	56	50	6
Incorrect:	60	60	0
		26 - unrelated, listed as full	
		34 - half siblings listed as full	

## Conclusions and Future Work

Benefits of the system include:

- Ability to fill in missing relationships
- Automatically resolve some incorrect relationships/errors
- Flag potential sample errors by using genetic relationships

Limitations of the system include:

- Assumption families follow a traditional structure
- Unable to detect non-blood relatives listed as related

Future work with this system will include:

- Flagging potential relationship errors during data entry
- Exploring the use of the EHR as an additional source of information to resolve conflicting or missing relationships

## References

1. McCarty CA, Peissig P, Caldwell MD, Wilke RA. The Marshfield Clinic Personalized Medicine Research Project: 2008 scientific update and lessons learned in the first 6 years. *Personalized Medicine*. 2008;5(5):529-542
2. Lowe JK, Maller JB, Pe'er I, et al. Genome-Wide Association Studies in an Isolated Founder Population from the Pacific Island of Kosrae. *PLoS Genet*. 2009 Feb;5(2)

## Acknowledgements

This work was funded by NIH grant 5U01HG004608-02 from the National Human Genome Research Institute (NHGRI), and supported by grant 1UL1RR025011 from the Clinical and Translational Science Award (CTSA) program of the National Center for Research Resources, National Institutes of Health.