Synthetic Germline VCF Generation for Rare Diseases: Case Study in NF1

Sikha Pentyala
sikha@uw.edu
University of Washington Tacoma
Washington, USA

Ziwei Pan
ziwei.pan@sagebase.org
Sage Bionetworks
Washington, USA

Patrick McKeever
lysander@uw.edu
University of Washington Tacoma
Washington, USA

Jineta Banerjee
jineta.banerjee@sagebase.org
Sage Bionetworks
Washington, USA

Luca Foschini
luca.foschini@sagebase.org
Sage Bionetworks
Washington, USA

Martine De Cock
mdecock@uw.edu
University of Washington Tacoma
Washington, USA

CCS Concepts

• **Applied computing** → **Computational genomics**; • **Computing methodologies** → **Learning paradigms**; • **Security and privacy** → **Data anonymization and sanitization**.

Keywords

Synthetic Data Generation, Rare Diseases, NF1

ACM Reference Format:
Sikha Pentyala, Ziwei Pan, Patrick McKeever, Jineta Banerjee, Luca Foschini, and Martine De Cock. 2026. Synthetic Germline VCF Generation for Rare Diseases: Case Study in NF1. In *Proceedings of the 17th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB 2026), June 30–July 3, 2026, Calabria, Italy*. ACM, New York, NY, USA, 1 page.

Abstract

Background. Rare diseases collectively affect over 300 million people worldwide, yet their underlying causes are not well understood. Neurofibromatosis type 1 (NF1), affecting approximately 1 in 3,000 individuals, is one such rare disease. Despite being a genetic condition, the underlying mechanisms of malignant transformation in NF remain poorly understood, leaving patients with limited therapeutic options. While high dimensional genomic data is collected from the patients, the application of advanced computational and artificial intelligence (AI) methods to NF1 and other rare diseases is particularly challenging, as available cohorts are typically small and heterogeneous, limiting statistical power and model generalizability. Interest in rare disease research has spurred the development of techniques for generation of synthetic patient data to benchmark new statistical methods, to augment training data, and to facilitate privacy-preserving data sharing.

While synthetic data generation (SDG) has been explored in rare disease contexts, the systematic generation and evaluation of *synthetic germline genomic data* for rare diseases remains largely unexplored. Existing methods for synthetic germline data generation focus on generating single nucleotide variants (SNVs) or operate on predefined genotype matrices, while others are restricted to somatic variants. To our knowledge, no prior work generates synthetic germline variant records, *including SNVs, as well as insertions and deletions (indels)*, learned directly from real variant calls in a rare disease cohort.

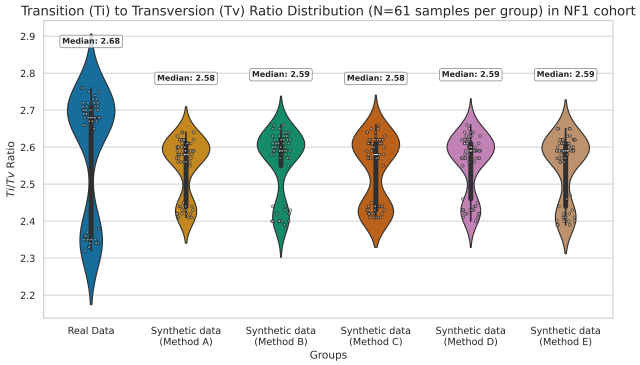

Figure 1: SNV transition/transversion (Ti/Tv) ratio comparing real and synthetic datasets.

Methods. We generated synthetic germline variant datasets in variant call format (VCF) by integrating domain knowledge (e.g., patient profiles and variant classifications) with a decision tree–based synthetic data generator (CART based SYNTHPOP) trained on 61 patients from NF Data Portal (Synapse). We developed 5 model variants (A–E) differing in incorporated domain knowledge and sampling strategy. For example, Method D relies only on common and recurrent variants, while Method E also considers patient-unique variants. Our code is available on Synapse.

Results. Fig. 1 demonstrates that the Ti/Tv ratio distribution in the synthetic datasets closely resemble the real NF1 cohort. We will further refine the models to better capture NF1 cohort characteristics, and evaluate the downstream utility and privacy risks of the synthetic data. All synthetic datasets are published on Synapse.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants Nos. 2451163 and 2523406, and by NSF NAIRR awards 240091 (TACC) and 240485 (TACC, AWS). This research was, in part, funded by the National Institutes of Health (NIH) Agreement No. 1OT2OD032581. The views expressed are those of the authors and do not necessarily reflect NIH policies. Additional support was provided by the eScience Institute.

Received 6 March 2026