

# Characterization of the Y Chromosome in Newfoundland and Labrador: Evidence of a Founder Effect

## Authors:

Heather Zurel<sup>1</sup>, Claude Bhérer<sup>2</sup>, Ryan Batten<sup>1</sup>, Margaret E. MacMillan<sup>1</sup>, Sedat Demiriz<sup>1</sup>, Sadra Mirhendi<sup>1</sup>, Edmund Gilbert<sup>3,4</sup>, Gianpiero L. Cavalleri<sup>3,4</sup>, Richard A. Leach<sup>1</sup>, Roderick E. M. Scott<sup>1</sup>, Gerald Mugford<sup>1</sup>, Ranjit Randhawa<sup>1</sup>, Alison L. Symington<sup>1</sup>, J. Claiborne Stephens<sup>1</sup> and Michael S. Phillips<sup>1\*</sup>

## Affiliations:

- 1) Sequence Bioinformatics, Inc., St. John's, Newfoundland and Labrador, Canada
- 2) Department of Human Genetics, Faculty of Medicine and Health Sciences, McGill University, Montréal, Québec, Canada
- 3) School of Pharmacy and Biomolecular Sciences (PBS), Royal College of Surgeons in Ireland, St. Stephen's Green, Dublin, Ireland
- 4) FutureNeuro SFI Research Centre, Royal College of Surgeons in Ireland, St. Stephen's Green, Dublin, Ireland

## Corresponding author emails:

\* Michael S. Phillips: pgxdoc@gmail.com

## Abstract

The population of Newfoundland and Labrador (NL) is largely derived from settlers who migrated primarily from England and Ireland in the 1700-1800s. Previously described as an isolated founder population, based on historical and demographic studies, data on the genetic ancestry of this population remains fragmentary. Here we describe the largest investigation of patrilineal ancestry in NL. To determine the paternal genetic structure of the population, 1,110 Y chromosomes from an NL based cohort were analyzed using 5,761 Y-specific markers. We identified 160 distinct paternal haplotypes, the majority of which (71.4%) belong to the R1b haplogroup. When NL is compared with global reference populations, the haplotype composition and frequencies of the NL paternal lineages primarily resemble the English and Irish ancestral source populations. There is also evidence for genetic contributions from Basque, French, Portuguese, and Spanish fishermen and early settlers that frequented NL. The population structure shows geographical and religious clustering that can be associated with the settlement of ancestral source populations from England and Ireland. For example, the R1b-M222 haplotype, seen in people of Irish descent, is found clustered in the Irish-settled Southeast region of NL. The clustering and expansion of Y haplotypes in conjunction with the geographical and religious clusters illustrate that limited subsequent in-migration, geographic isolation and societal factors have contributed to the genetic substructure of the NL population and its designation as a founder population.

Keywords: Population genetics, Y chromosome, Newfoundland and Labrador, ISOGG, founder population

## Introduction

The Canadian province of Newfoundland and Labrador (NL) is home to a unique population that traces its origins to the migration of European communities roughly 300 years ago. The current population is thought to be derived from approximately 25,000 immigrants in the 1700's and 1800's who settled in remote coastal communities.<sup>1,2</sup> These outports were largely isolated from each other, with little settlement in the interior of the island. Communities grew through large families but remained isolated until the 1950s with the advent of paved roads.<sup>1</sup> The population has continued to expand to its current size of 520,000 and, with the decline of the seafaring economy, is shifting from rural to urban centers.<sup>3</sup>

The main European ancestral source populations that settled NL were from communities around County Waterford and adjacent counties in Ireland and from the counties of Cornwall and Devon as well as fishing ports in Southern England.<sup>1,4</sup> Following immigration to Newfoundland, English Protestants and Irish Catholics are thought to have remained separated by attending different schools, and rarely inter-married, further isolating these communities.<sup>5,6,7</sup> Additional European influences that are also thought to have contributed to the genetic landscape of NL are the Portuguese,<sup>8,9</sup> French<sup>10</sup> and Highland Scottish.<sup>1</sup> Norse settlers were present in NL for > 100 years in around 1000 A.D.,<sup>11,12,13</sup> although it appears that they never settled permanently. Also present in NL before, during and after the time of European settlement were Indigenous peoples.<sup>1,7,14</sup> Since the 1900's, immigration to Newfoundland has been limited, and the genetic diversity in the province largely traces back to the original European settlers.<sup>15</sup>

Detailed studies of Y chromosome haplotypes have revealed male migration patterns throughout history and led to an understanding of the origins of current human populations.<sup>16,17,18,19,20</sup> These studies contributed to the development of a standardized Y-DNA phylogenetic tree maintained by the International Society of Genetic Genealogy (ISOGG).<sup>21</sup> European Y chromosomes are primarily comprised of the haplogroups E, G, I, J, N and R, with the R haplogroup comprising the majority of the Y chromosomes.<sup>22,23,24,25</sup> While many previous studies are limited by short tandem repeats (STR's) and/or low resolution single nucleotide polymorphisms (SNPs) panels,<sup>22,24,25,26,27,28</sup> they provide information on the composition and frequency of major haplogroups in Europeans.

Supported by studies on the genetic structure of the population<sup>7</sup> and the presence of numerous rare monogenic disorders,<sup>29</sup> the population of NL has been described as a founder population. However, information about the haplotypic composition, frequency of Y chromosome variation and ancestral origins across NL is limited. To address these questions, the Y chromosomes of 1,110 individuals from the Newfoundland and Labrador Genome Project (NLGP) cohort<sup>30</sup> were analyzed in order to: 1) determine the composition and frequency of haplogroups in the paternal lineages; 2) elucidate the population structure of the Y

chromosome; 3) understand how the NL population compares with the European ancestral source populations, and 4) identify evidence of founder effects based on haplotypic expansion and regional clustering.

## Materials & Methods

### Newfoundland and Labrador Cohort

Data of the initial 2,500 participants from the NLGP study, a general population cohort from NL, was used for this analysis.<sup>30</sup> As part of the participants' self-reported data, we collected information on their religion and birthplace of their ancestors. Each participant provided a saliva sample using the DNA Genotek Oragene OG-600 collection kit (DNA Genotek, Ottawa, Canada). DNA extracted from these samples was genotyped using the Illumina Global Diversity Array (GDA; Illumina, San Diego, CA). Variant calling and quality control (QC) analysis of the genotyping data set was performed using Illumina's Array Analysis Platform (IAAP) Command Line Interface (CLI) and GTCToVCF pipeline ([github.com/Illumina/GTCToVCF](https://github.com/Illumina/GTCToVCF)) (Illumina, San Diego, CA). Out of the 2.1 M variants on the Illumina GDA SNP array, 5,761 SNPs on the male specific portion of the Y chromosome were selected for analysis. QC analysis of the Y chromosome samples determined that 1,110 participants (designated NLGP<sub>1,110</sub> cohort) had fewer than 200 missing Y chromosome calls (call rate > 96.5%).

### Phylogenetic Reconstruction

The phylogenetic tree was constructed using two different methods: 1) the yHaplo software package,<sup>31</sup> and, 2) a manual method using maximum parsimony (Supplemental Materials). Although there was concordance between the methods, the manual maximum parsimony approach gave greater haplotype resolution as it enabled the incorporation of SNPs with missing data, singleton SNPs (i.e. variation only observed in a single participant), SNPs without ISOGG designations, and the resolution of phylogenetically inconsistent SNPs (Figure S1, Table S1). Of the 5,761 SNPs, 2,114 were phylogenetically informative (Supplemental Material). The 160 paternal haplotypes, their frequencies, and relationships to each other were used for subsequent analyses. Tag SNPs, associated with specific ISOGG long-form haplotypes, are reported whenever possible to facilitate comparison with the literature.

### Identification of Descendants of NL Founders

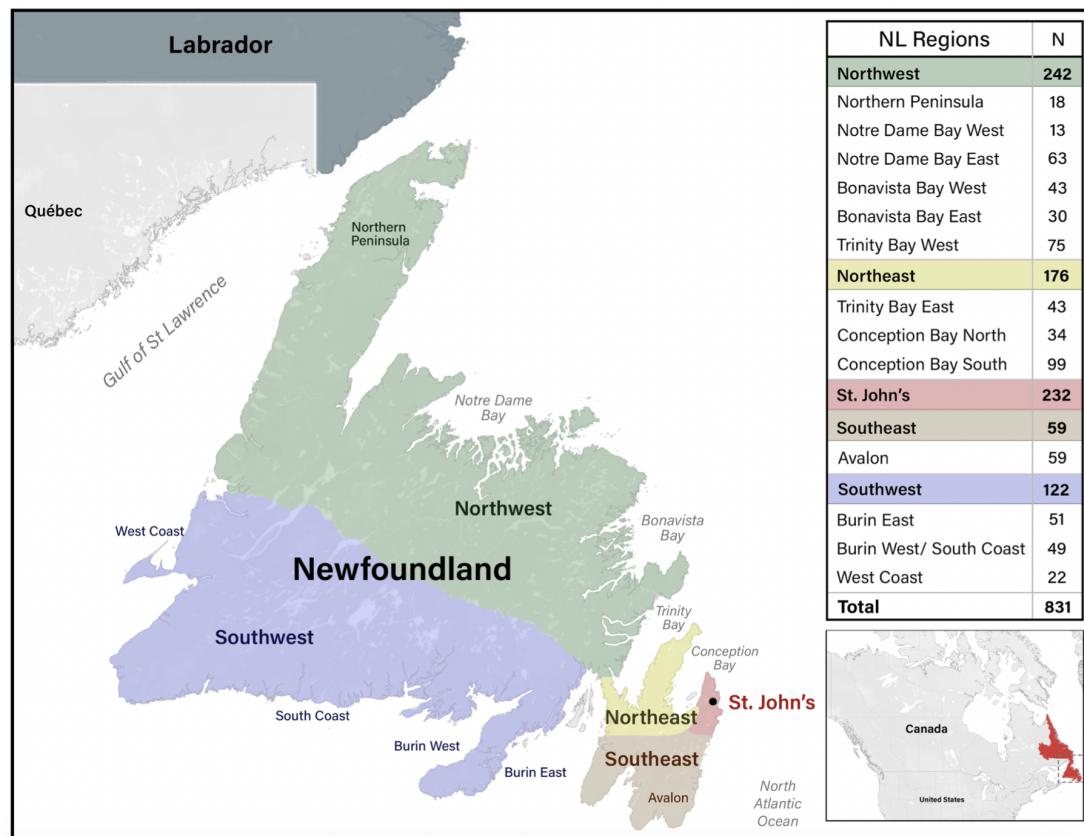
A combination of self-reported ethnicity, principal component analysis (PCA) of autosomes, and self-reported birthplaces of paternal ancestors were used to identify individuals whose ancestors descended from early European settlers. Within the NLGP<sub>1,110</sub> cohort, only 4 participants reported having Indigenous ancestry while 24 participants reported having a mixture of European and Indigenous ancestries (2.6%). Given the limited number of participants with various levels of Indigenous ancestry and the lack of an appropriate reference panel for Indigenous peoples in Eastern North America, we did not investigate the

contributions of Y-DNA from Indigenous Peoples in this study. Any participants who were recent immigrants or who reported that their paternal ancestors (up to great-grandfathers) were not from NL were excluded from this analysis. To assess continental ancestry, genotyping data from the autosomes of the NLGP participants was merged with autosome data from the 1000 Genomes project (1KGP3) before running a principal component analysis (PCA) using PLINK 2.0.<sup>32</sup> Continental ancestry was assigned using the first 5 principal components (PCs) (Figure S2).

To compare NL Y haplotypes with potential ancestral source populations, Y chromosome data from the Irish DNA Atlas<sup>33</sup> and the People of the British Isles (PoBI)<sup>34</sup> were analyzed. Of the 812 SNPs overlapped between the NLGP<sub>1,110</sub> cohort and the PoBI and Irish DNA Atlas data sets, 516 were monomorphic. The remaining 296 SNPs were used to infer major haplogroup frequencies for all 856 Y chromosomes in these data sets. For comparisons with other world populations, the gnomAD allele frequency database<sup>35</sup> was queried. Since rare variants in European populations are more likely to be population-specific, the 2,114 phylogenetically informative Y-DNA SNPs were inspected for their presence in 7 gnomAD European populations (Basque, Finnish in Finland (FIN), French, British in England and Scotland (GBR), Iberian population in Spain (IBS), Italian, and Toscani in Italia (TSI)). From these, 60 variants were observed in one or two of these populations. Analysis of these 60 variants was extended to all gnomAD populations to assess whether they were informative about potential population ancestry.

### **Characterization of the NL Y chromosome population structure**

Kinship coefficients were estimated using the KING relationship inference software<sup>36</sup> implemented in Plink2.<sup>32</sup> First degree relatives ( $0.177 < \text{kinship} < 0.354$ ) were removed from population analyses. The geographical distribution of the haplotypes was mapped using the birthplace of their most distant paternal ancestor. Regions were assigned based on historical records of settlements and societal and geographic constraints. To evaluate the regional similarities and differences across NL, the province was divided into 5 large regions along the North/South axis and East/West at the point of the Avalon Peninsula isthmus. The St. John's metropolitan area was designated as a distinct region (Figure 1). These regions were further subdivided into 15 subregions based on major geographical features. The Labrador region, with only 4 participants, was not included in clustering analyses to avoid bias from low numbers. The remaining data set consisted of 831 individuals and 133 haplotypes (designated NL<sub>831</sub> cohort). Haplotype frequencies were calculated based on geography and religion. Religious affiliation was grouped into 4 categories: Catholic, Protestant, No Religion and Other which includes all other religious/spiritual designations. Notre Dame Bay West, the Northern Peninsula and the West Coast subregions (Figure 1) had less than 25 participants which limited interpretation of this data.



**Figure 1: The division of Newfoundland and Labrador into geographical regions and subregions that were used for the analysis of regional Y haplogroup and haplotype frequencies and composition.** First degree relatives, recent immigrants and those missing information on the geographic location of paternal ancestors were not included.

## Statistical Methods

### Haplotype Diversity and $F_{ST}$

Haplotype diversity ( $H$ ), which represents the probability that two randomly sampled haplotypes from a population are different, was calculated for each subregion, as well as the NL<sub>831</sub> cohort as a whole ( $n = 831$ , 133 haplotypes), using the following equation:

$$H = \frac{n}{n-1} \left( 1 - \sum_{i=1}^k p_i^2 \right)$$

where  $n$  is the sample size,  $k$  is the number of distinct haplotypes and  $p_i$  is the frequency of each haplotype.<sup>37</sup> All distinct haplotypes were used to estimate  $H$  for the 5 major regions and 14 remaining subregions. The haplotype diversity term,  $H$ , was used to estimate  $F_{ST}$  of each

subregion compared to the NL sample, and pairwise  $F_{ST}$  was calculated between each pair of regions as described by Slatkin<sup>38</sup> for haploid genotypes. The pairwise  $F_{ST}$  values were visualized using multidimensional scaling (MDS) analysis using the cmdscale function in R (v4.1.0).<sup>39</sup>

## Statistical Comparisons

To assess stratification of paternal lineages among the 14 subregions, a PCA was performed based on the variance-covariance matrix of haplotype frequency distribution using the PCAtools R package.<sup>40</sup> To determine the percentage of variance between populations and groupings, Analysis of MOlecular VAriance (AMOVA) was employed.<sup>41</sup> The percentage of variation and associated p-values were reported between populations, within populations between subregions and within subregions. AMOVA analyses were conducted using R version 4.1.0 and packages: ade4 package<sup>42</sup> obtaining simulated p-values (based on 1000 Monte Carlo simulations). For pairwise comparisons of haplotype composition between regions and subregions, Fisher's exact test with a simulated p-value (using 1,000 Monte Carlo simulations)<sup>43</sup> was used with a Benjamini-Hochberg correction.<sup>44</sup> R version 4.0.3 was used with the stats package<sup>39</sup> to calculate p-values, and results were visualized using the ggplot2 package.<sup>45</sup>

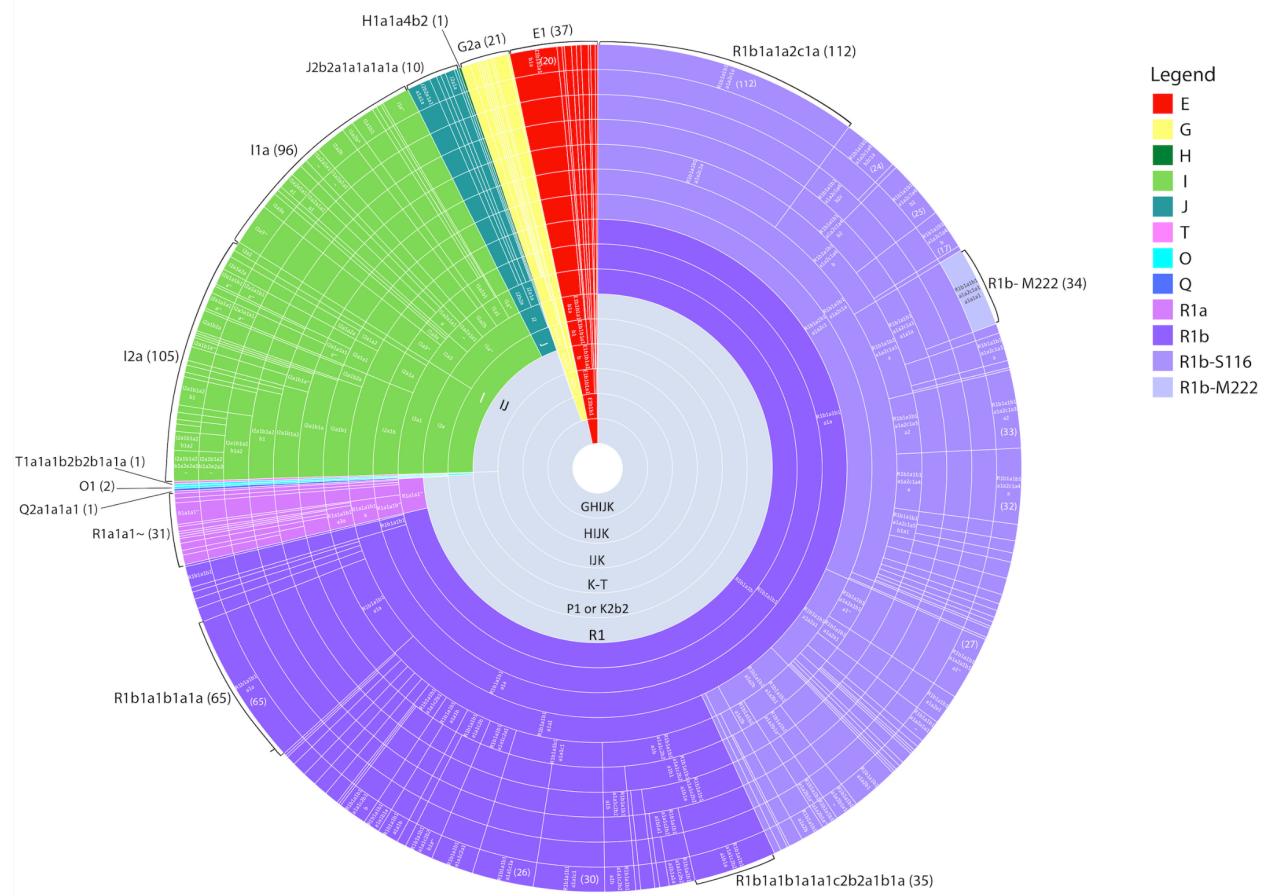
# Results

## Y Chromosome Structure of the NL population

To construct the NL phylogenetic tree, we used 2,114 phylogenetically informative SNPs in conjunction with the long-form haplotype ISOGG nomenclature to assign 1,110 NL participants to 160 specific haplotypes (Figures 2 & S3, Tables 1, S2 & S3). Seventeen major internal branch points and 7 terminal haplotypes were supported by 20 or more phylogenetically informative SNPs providing confidence in the assembly of the NL Y-DNA tree assembly (Figure S4). The majority of the Y chromosomes in the NLGP<sub>1,110</sub> cohort occur in the R haplogroup (74.2%, Table 1), predominantly within the R1b haplogroup (71.4% Figure 2, purple). The R1b-S116 haplogroup (R1b1a1b1a1a2) (light purple), comprises 46 distinct haplotypes in the NLGP<sub>1,110</sub> cohort (43.2%), including R1b-M222 (R1b1a1b1a1a2c1a1a1a1), which occurs in 3.1% of the NL Y chromosomes (Table 2). Also present are subclades of major haplogroups I2a, I1a, E1b, R1a, G2a, J2b, J2a in decreasing order of occurrence. The following 7 haplogroups, E1a, H1a, J1a, T1a, O1a, O1b, and Q2a, were detected in single participants, mostly in people who self-reported being born outside of NL.

**Table 1: List of Major Haplogroups/Haplotypes present in the NLGP<sub>1,110</sub> cohort.**

ISOGG Subclade or Haplotype	Tag SNP	Haplotype Count	Chromosome Count
E1a2a1	CTS10935	1	1
E1b1b1a1a1	V12	1	2
E1b1b1a1b1	L168	3	23
E1b1b1a1b2	L677	2	5
E1b1b1b	Z827	3	6
G2a1a1a1a1	Z6638	1	1
G2a2a1a2a1a	L166	1	6
G2a2b1	M406	1	1
G2a2b2a1	L140	6	11
G2a2b2b1a1	F872	1	2
H1a1a4b2	M2972	1	1
I1a~	CTS9857	1	12
I1a1	CTS6364	4	15
I1a2a1a1	S337	3	22
I1a2b	S296.1	2	18
I1a3~	Z63	4	29
I2a1a1	CTS595	4	18
I2a1a2	M423	2	9
I2a1b1	M223	15	64
I2a1b2a	L38	2	10
I2a2	L596	1	4
J1a	CTS5368	1	1
J2a1a	F4326	4	9
J2b2a	M241	2	13
O1a1a1	F446	1	1
O1b1a1a1a1a1	M111	1	1
Q2a1a1a1	FGC1897	1	1
R1a1a1	M417	13	31
R1b1a1b1	L23	76	791
R1b1a1b2a	GG480	1	1
T1a1a1b2b2b1a1a	CTS6507	1	1
<b>Total</b>		<b>160</b>	<b>1,110</b>



**Figure 2. Radial diagram illustrating the proportion of individuals for each haplogroup that are present in the NLGP<sub>1,110</sub> cohort.** The gray inner circles represent mega haplogroups (e.g. K-T, GHIJK) in the phylogenetic tree. Each of the major haplogroups is indicated by a unique color. Each distinct haplotype or haplogroup is separated by a straight white line. Long-form ISOGG nomenclature is provided where possible and can be seen in greater resolution in Figure S3. Each segment of each ring of the radial diagram is proportional to the number of Y chromosomes included in that haplogroup, and segments on the outer ring are proportional to the number of participants belonging to each terminal haplotype. The numbers of participants for select terminal haplotypes are indicated by the brackets on the outer ring.

Review of the NLGP<sub>1,110</sub> cohort identified 31 terminal haplotypes present in 10 or more individuals (Table S5), and more specifically, 7 terminal haplotypes that were present in 30 or more individuals (Table 2). The largest group was the R1b-DF13 terminal haplotype (R1b1a1b1a1a2c1a) which occurred in 112 individuals.

**Table 2: Terminal haplotypes and their relative frequencies in the NLGP<sub>1,110</sub> cohort with a chromosome count of 30 or more.**

ISOGG Haplotype	Tag SNP	Chromosome Count	Frequency in NL (N = 1,110)
R1b1a1b1a1a	L151	65	5.9
R1b1a1b1a1a1c1	S264	30	2.7
R1b1a1b1a1a1c2b2a1b1a	Z8	35	3.2
R1b1a1b1a1a2c1a	DF13	112	10.1
R1b1a1b1a1a2c1a1a1a1a1	M222	34	3.1
R1b1a1b1a1a2c1a3a2	CTS4466	33	3.0
R1b1a1b1a1a2c1a4a	Z255	32	2.9

### **Y Chromosome Structure of the NL population**

To understand the population structure of NL, we analyzed the haplogroup frequencies by geographical region of the 831 descendants of European founders (NL<sub>831</sub>) (Table 3). Regional differences in R1b haplogroups are observed across NL. The R1b-S116 haplogroup represents greater than 42% of the Y chromosomes except in the Northwest region (28.9%). The frequency of the R1b-M222 haplotype in comparison, is highest in the Southeast region.

**Table 3: Major haplogroup frequencies detected in the NL<sub>831</sub> cohort by geographical region.**

Each NL individual was assigned to a geographical region based on the self-reported birthplace of their most distant paternal ancestor.

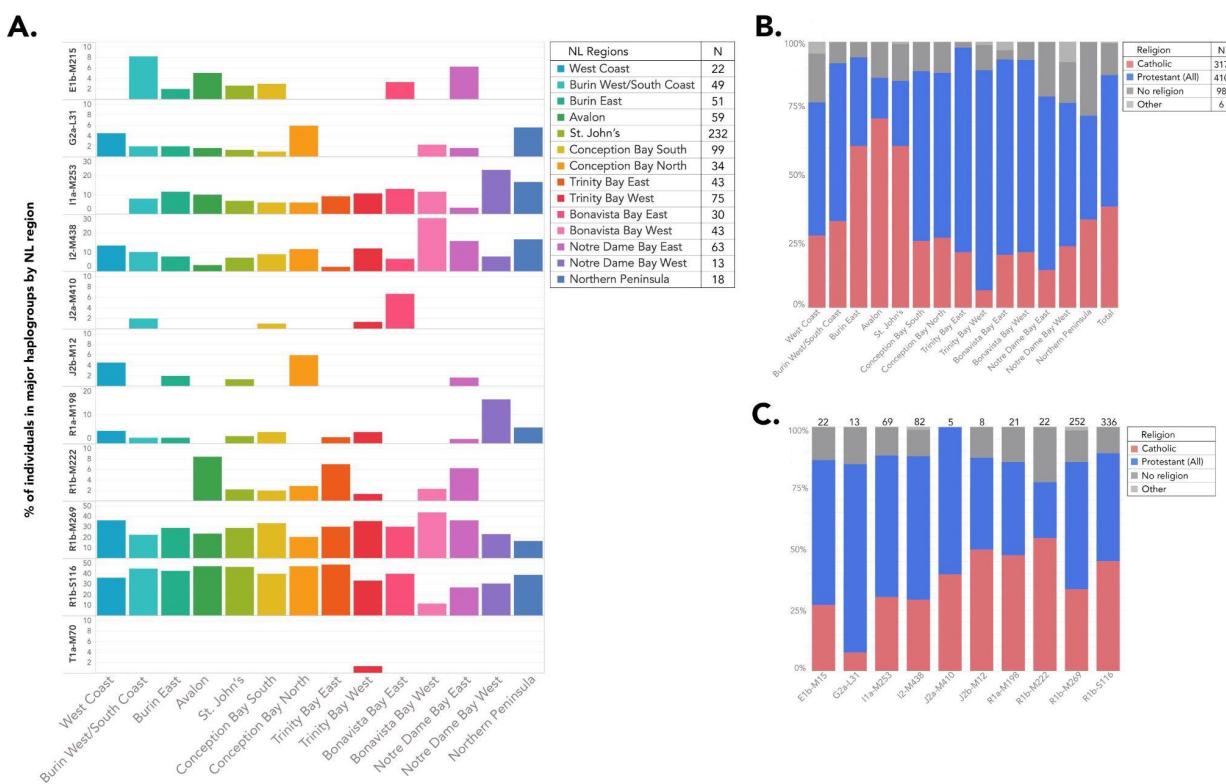
NL Region	Major Haplogroup Frequencies										Number of individuals	
	E1b-M215	G2a-L31	I1a-M253	I2-M438	J2a-M410	J2b-M12	R1a-M198	R1b-M269	R1b-S116	R1b-M222		
Southeast	5.1	1.7	10.2	3.4	0.0	0.0	0.0	23.7	47.5	8.5	0.0	59
Northeast	1.7	1.7	6.8	8.0	0.6	1.1	2.8	30.1	43.8	3.4	0.0	176
St. John's	2.6	1.3	6.9	7.3	0.0	1.3	2.6	28.9	47.0	2.2	0.0	232
Northwest	2.1	1.2	10.3	15.3	1.2	0.4	2.9	34.7	28.9	2.5	0.4	242
Southwest	4.1	2.5	8.2	9.8	0.8	1.6	2.5	27.9	42.6	0.0	0.0	122
<b>NL Cohort</b>	<b>2.6</b>	<b>1.6</b>	<b>8.3</b>	<b>9.9</b>	<b>0.6</b>	<b>1.0</b>	<b>2.5</b>	<b>30.3</b>	<b>40.4</b>	<b>2.6</b>	<b>0.1</b>	<b>831</b>

Haplotype diversity based on the frequencies of 133 haplotypes was used to calculate pairwise  $F_{ST}$  between the 5 major regions. The MDS plot (Figure S5) demonstrates that the major difference among populations (99.4% of the total variance) corresponds to an East-West axis of variation. Results from the AMOVA showed that most of the variation can be explained by the haplotype distribution within subregions (99.3%;  $p = 0.001$ ; described as “Within populations” in Table S6). A comparison of the Avalon subregion in the East with the Northwest subregions shows significant differences in haplotype composition by Fisher's exact test ( $p = 0.02$  to  $0.001$ ) (Table S7). The East-West geographical haplotype distribution is further supported by PC analysis as represented in the scree plots of the top 5 components (Figure S6).

The coastal communities show distinct patterns in haplogroup frequency and religious affiliation (Figure 3A & 3B). The St. John's metropolitan area has experienced immigration from many of the coastal communities. As expected, most haplotypes observed in the other regions are present in St. John's (Figure 3A and S7). In the Northeast region several haplogroup frequencies differ from those observed in the overall NL<sub>831</sub> cohort suggesting that these subregions might have been settled by immigrants originating from different European regions (Figure 3A). For example, adjacent subregions on the same Peninsula in the Northeast show differential frequencies of I2-M438 ranging from 2.3% in Trinity Bay East to 11.8% in Conception Bay North (Figure 3A). Similarly, in the adjacent Northwest region, Notre Dame Bay East, Bonavista Bay West and Bonavista Bay East subregions show significant differences in haplotype composition when compared with subregions in both the Northeast and Southeast ( $p = 0.02$  -  $0.001$  by Fisher's exact test) (Figure 3A, Table S7), further reinforcing the East-West geographical distribution of haplotypes.

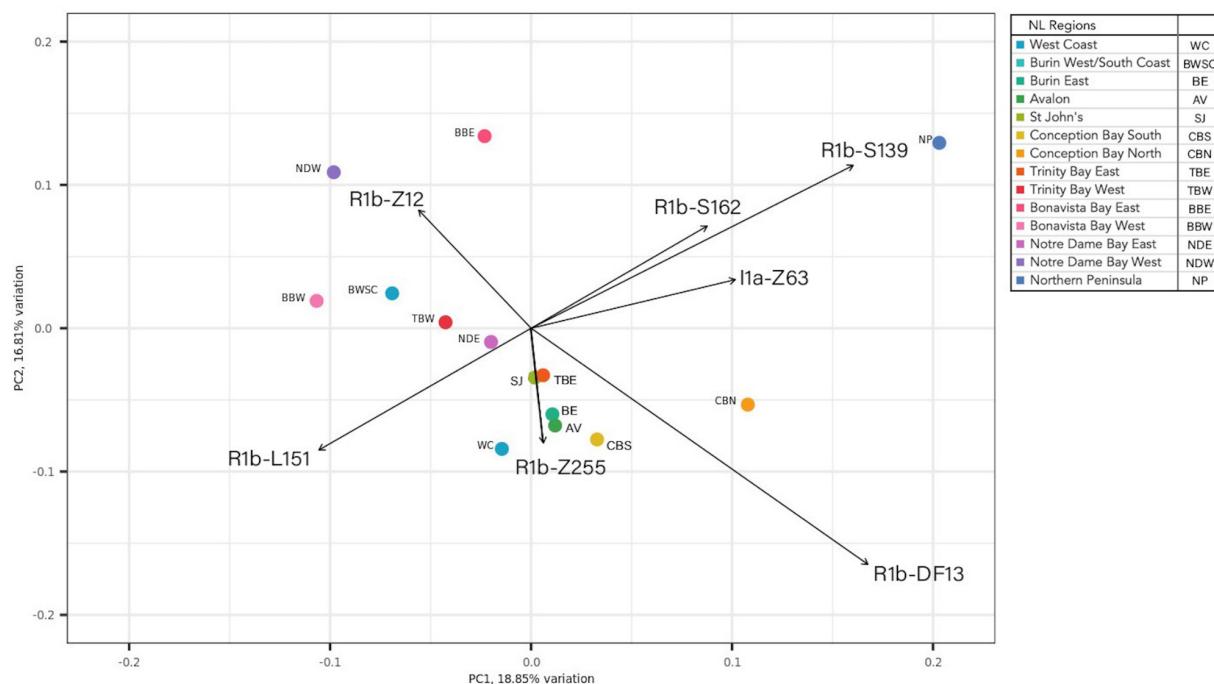
Religion displays some distinctive distribution and frequency patterns across NL as previously described.<sup>1,5,6</sup> Participants in the Southeast region are predominantly Catholic (>70%) while the Protestant religion predominates in the North (~ 70%) (Figure 3B). In the NL<sub>831</sub> cohort, in some

regions, some haplotypes appear to be associated with a specific religious affiliation (Figure 3C). For example, the elevated presence of I2-M438 and I1a-M253 haplogroups in the Northwest appears to be mainly associated with Protestant communities (Figure 3C). Similarly, the R1b-M222 haplotype, associated with Irish ancestry,<sup>28</sup> is observed mainly in Catholic communities (Figure 3C) and is primarily seen in the Avalon Peninsula. As Burin East is the closest subregion of the three to the Avalon subregion and closely resembles the Avalon subregion in terms of religious affiliation, it is noteworthy that R1b-M22, haplotype associated with catholic communities, is absent in this region (Figure 3C).



**Figure 3: Distribution of major haplogroups by religious affiliation and geographical location.** Each NL individual was assigned to a geographical region based on the self-reported birthplace of their most distant paternal ancestor and religion was self-reported. **A:** The frequencies and distribution of the major haplogroups represented as a percentage of the total number of individuals in that region. **B:** The frequencies and distribution of self-reported religious affiliation as a percentage of the total number of individuals in a given region. **C:** The frequencies of self-reported religion represented as a percentage of each major haplotype. Only one person was reported to have a haplotype of T1a-M70 and therefore is not represented on Figure 3C.

A PC analysis based on 133 terminal haplotypes was used to visualize the structure of the paternal lineages in the 14 subregions of NL. The first 5 PCs explain >70% of the variation in haplotype frequencies by subregion (Figure S6). A biplot of the first 2 PCs (Figure 4) identifies which haplotypes are the major contributors to the first and second dimensions of PC variation, and shows differentiation between the subregions in the Eastern and Western regions of NL. The R1b-Z255 haplotype, which is mainly observed in Catholics (81%) in the Southeast region is the major contributor to the clustering of the populations in Eastern NL and shows a similar distribution to the R1b-M222 haplotype. R1b-L151 and R1b-Z12 haplotypes which occur mainly in Protestant participants, located in the North Central and West coast regions of NL, appear to be the major haplotypes that are contributing to the clustering of these populations.



**Figure 4: Principal Component (PC) Analysis biplot of the haplotype frequencies in the 14 NL subregions indicating which haplotypes were the major contributors to the top 2 PC axes.** Arrows represent the loadings for the 7 haplotypes whose differences in frequencies are the largest contributors to the value of the first two PCs.

### Comparison to Ancestral Source Populations

To infer the origins of the NL paternal lineages, we compared the major Y chromosome haplotypic frequencies found in NL to those of Britain, Ireland and other European source populations using the Irish DNA Atlas,<sup>33</sup> PoBI,<sup>34</sup> and the gnomAD allele frequency database<sup>35</sup> (Table 4). The majority of Y chromosomes within the PoBI and Irish DNA Atlas data sets belong to subclades of the R1b (R1b-M343) haplogroup (Table 4). In addition, analysis of the gnomAD data in combination with data from the PoBI and the Irish DNA Atlas showed evidence of specific

haplotypes that could act as markers for the ancestral populations. For example, 3 R1b haplotypes, R1b-U198 (R1b1a1b1a1a1c2a1), R1b-L46 (R1b1a1b1a1a1c2b1a1a), and R1b-Z8 (R1b1a1b1a1a1c2b2a1b1a), seen at relatively high frequency in the NL population, are observed almost exclusively in England in the PoBI and Irish Atlas data sets and likely correspond to English paternal lines. In comparison, R1b-M222 (R1b1a1b1a1a2c1a1a1a1) and R1b-Z255 are seen primarily in the Catholic dominated areas of NL, and are almost exclusively seen in Irish populations. The majority of other Y haplotypes seen in the British and Irish populations (subclades of E1b, I1, I2, J2a, J2b, and R1a) were also seen in the NL<sub>831</sub> cohort, although at different frequencies. For example, R1b-M222 is seen at a frequency of 23.9% in the Irish DNA Atlas data set but only at a frequency of 2.6% in the NL<sub>831</sub> cohort (Table 4).

**Table 4: Major haplogroup frequencies detected in the People of the British Isles and Irish DNA Atlas data sets sorted by geographical region.** The frequencies of the major haplogroups were calculated and values are represented as a percentage of the total number of individuals in that region. The “Other” column represents haplotypes (F, G, N, Q and T) that were only seen once or twice in the regions (primarily the English population). The R1b-M222 haplotype, included as part of the R1b-M343 haplogroup, is also shown separately as it represents 23.9% of the total Irish Y chromosomes.

Region	Major Haplogroup Frequencies									Number of individuals	R1b-M222
	E1b-P2	G2a-L31	I1-M253	I2-M438	J2a-M410	J2b-M12	R1a-M198	R1b-M343	Other		
England	2.1	2.3	13.6	8.2	1.9	1.3	2.8	66.7	1.1	619	1.0
Ireland	0.9	1.7	0.9	10.3	0.0	0.9	1.7	82.9	0.9	117	23.9
Northern Ireland	0.0	10.5	0.0	15.8	0.0	0.0	0.0	73.7	0.0	19	5.3
Orkney	0.0	0.0	0.0	0.0	0.0	0.0	33.3	66.7	0.0	15	0.0
Scotland	2.4	0.0	11.9	9.5	0.0	0.0	4.8	71.4	0.0	42	7.1
Wales	0.0	2.3	0.0	4.6	0.0	0.0	2.3	88.6	2.3	44	0.0
<b>NL Cohort</b>	<b>2.6</b>	<b>1.6</b>	<b>8.3</b>	<b>9.9</b>	<b>0.6</b>	<b>1.0</b>	<b>2.5</b>	<b>73.3</b>	<b>0.1</b>	<b>831</b>	<b>2.6</b>

In addition to the English and Irish ancestral populations, several other European populations are known to have fished off the coast of NL.<sup>1</sup> Under the premise that low-frequency variants are more likely to be population-specific, we looked for rare variants in gnomAD to identify potential source populations for Newfoundland's founders. We identified 26 distinct haplotypes that were present in one or two of 7 gnomAD European source populations (Basque, Finnish in Finland (FIN), French, British in England and Scotland (GBR), Iberian population in Spain (IBS), Italian, and Toscani in Italia (TSI)). Further analysis of these variants was expanded to all gnomAD populations. Multiple subclades of E1b, and I2a, observed 12 to 64 times in the NLGP<sub>1,110</sub> cohort, were mainly found in North African and Middle Eastern populations. These haplotypes were associated, at low frequency, with the Southern European populations of France, Iberia, Basque, and Italy. Given that the Y chromosome samples with population designations within gnomAD are limited, further study is required to validate these observations. Analysis of the gnomAD, PoBI and Irish DNA Atlas data revealed several examples of haplotypes that appear to have expanded over time in the NLGP<sub>1,110</sub> cohort. For example, the R1b-L46 (R1b1a1b1a1a1c2b1a1a) haplotype which is seen in only 2 samples in the English data (PoBI), and not seen in the Irish data, appears in 14 NL participants. Likewise, R1b-Z8 (R1b1a1b1a1a1c2b2a1b1a) is seen in 4 samples in the English data (2 samples in PoBI and 2 samples in gnomAD), but seen in 52 NLGP<sub>1,110</sub> cohort participants. This observation is suggestive of both possible oversampling of specific haplotypes from England in the settlers who came to NL and possible evidence of local expansion.

## Discussion

This analysis represents the most detailed study of patrilineal ancestry in NL reported to date. In order to characterize the paternal lineages within NL, a high-resolution Y-DNA tree was generated using 2,114 phylogenetically informative markers. Given this level of resolution, this study represents the most detailed study of patrilineal ancestry reported to date for NL. As discussed in the methods, we did not investigate the contributions of Y-DNA from Indigenous Peoples. To address the ancestral contributions of the Indigenous Peoples to the NL Y-DNA tree, a dedicated study of Indigenous Peoples, with and informed by these communities, would be warranted.

The majority of the Y-DNA haplogroups that were identified in the NLGP Y chromosomes appear to be of European origin and reside within the R1b haplogroup (71.4%). The frequency of R1b in the NL cohort is comparable to the English and Irish frequencies observed within the PoBI data, supporting the historical records that immigrants from both these populations settled in NL.

The remaining Y chromosomes in NL, primarily haplogroups I2a1 (9.9%), I1 (8.3%), E1b (2.6%), R1a (2.5%), and J (1.6%), are consistent with haplogroups that are seen in other Western European populations.<sup>23</sup> Many of these haplogroups have origins in specific European regions,

for example R1a and its subclades are commonly observed in Scandinavian populations.<sup>26,49</sup> It is thought that much of the R1a haplogroup in England and Ireland is associated with Viking settlement.<sup>50</sup> As the presence of the R1a haplogroup in NL appears to reflect the frequencies seen in these data sets (Table 4), it most likely originated with the English and Irish settlers. The most prevalent of the I haplogroups in the NL<sub>831</sub> cohort was I2-M438 which comprises I2a, I2b and their respective subclades. Unlike I1, the I2a haplogroup and its subclades are much less frequent in Scandinavia but are reported to comprise 10% of Irish and 6% of Basque Y haplogroups.<sup>24,25,26,28</sup> While the presence of I2a, which is clustered in the Northwest region of the province (Figure 3, Table 3), is consistent with English-settled communities in NL, it also could be indicative of the presence of Iberian/Basque Y-DNA that originated from Portuguese and Spanish fishermen.<sup>8,9</sup> Similarly, the J haplogroup, comprising ~10% of the current Portuguese population<sup>48</sup> may also have originated in NL with the presence of Portuguese ancestors.

Clustering patterns of haplotypes in specific communities and subregions in the NL<sub>831</sub> cohort appear to be associated with clustering of self-reported religious affiliation (Figure 3B). The clustering patterns of religion align with historical records of settlements in these regions, primarily Irish Catholics in the Southeast (>70% Catholic) and English Protestants in the Northwest region (70%) (Figure 3B).<sup>1,4,6</sup> Although religious affiliation can change, our data suggests that self-reported religion in the NL population can be viewed as a surrogate marker for both religion and geographic origin of the participant's paternal lineage in NL.<sup>4</sup> The R1b-M222 haplotype, a known Irish haplotype,<sup>22,25,28</sup> and R1b-Z255, speculated to be of Irish origin,<sup>51</sup> show localized clustering to known Irish Catholic communities, specifically in the Avalon subregion (Figure 4). Given that early migration of Irish Catholics to NL is well documented, it is likely that these settlers are the primary source of these haplotypes.<sup>1,6</sup> Although the R1b-M222 haplotype accounts for approximately 25% of Y chromosomes in Ireland,<sup>25</sup> it is only seen at a frequency of 3% in the NL cohort. This difference is likely because the R1b-M222 haplotype is primarily seen in Northwest Ireland<sup>25</sup> whereas the historical records suggest that NL was primarily settled by immigrants from Southeast Ireland.<sup>1,25</sup>

NL regional haplotypes exhibit differences along an East to West axis ( $p = 0.004$ ; Table S6, Figures 3, S5 & S6) and appear to be driven by the ancestral origins of the population with Irish Catholics in the South and East and English Protestants in the North and West. This observation supports the hypothesis that communities were established by settlers who originated from certain communities or specific parishes in Ireland and England and stayed isolated over time. Regions that are directly adjacent to each other, for example Bonavista Bay East and West, only separated by ~60 km of water, show significant differences in haplogroup composition, supporting the historical records of isolation of coastal communities.<sup>4,5,6</sup> As expected, the St. John's Metropolitan region, which has experienced recent immigration from many coastal communities, does not show the same patterns of geographical clustering. The data also indicate that there are Y chromosome contributions from additional European

populations such as the Basque, Portuguese, Italian and French. All these observations support the hypothesis that paternal Y haplogroups arrived from distinct European ancestral communities to specific regions within NL.

The unique characteristics of the Y haplotype population structure in NL are indicative of a founder effect. These communities increased over the last 300 years from 25K people to >520K people.<sup>1,2,3</sup> Evidence of isolation and expansion can be seen by the geographical clustering patterns, and the expansion of certain haplotypes in the NL population (Table 2, Table S4). In fact, 64% of the Y chromosomes in the NLGP<sub>1,110</sub> cohort show possible evidence of expansion over time as these haplotypes occur in 10 or more people (31 haplotypes in 709 people) (Table 2, Table S4). The expanded haplogroups of R1b-L151 and R1b-Z255 show evidence of regional clustering and expansion as the major haplogroups that differentiate subregions in the East (R1b-Z255) from subregions in the Northwest (R1b-L151) (Figure 4). These observations illustrate that specific ancestral source populations from Europe settled NL, expanded over time, and contributed to the unique clustering patterns seen today.

In summary, NL is an excellent example of a population exhibiting founder effects resulting from limited genetic input followed by generations of geographical and societal isolation which led to regional expansion of specific haplotypes. This data provides a better understanding of the NL genetic population structure which can inform both ancestral history and population structure.

## References

1. Mannion J.J. (Ed) (1977). The Peopling of Newfoundland: essays in historical geography. Institute of Social and Economic Research, Memorial University of Newfoundland.
2. Bear, J.C., Nemec, T.F., Kennedy, J.C., Marshall, W.H., Power, A.A., Kolonel, V.M., Burke, G.B., Opitz, J.M. & Reynolds, J.F. (1987). Persistent genetic isolation in Outport Newfoundland. *Am. J of Med. Genet*, 27(4), 807–830.
3. Government of Newfoundland and Labrador. (2021, September). Selected Provincial Quick Facts. Newfoundland & Labrador Statistics Agency. Retrieved November 1, 2021, from <https://www.stats.gov.nl.ca/>.
4. Handcock, W.G. (1989). So long as there comes no women: Origins of English settlement in Newfoundland. Breakwater Books.
5. Martin, L.J., Crawford M.H., Koertvelyessy, T., Keeping D., Collins, M., & Huntsman, R. (2000). The Population Structure of Ten Newfoundland Outports. *Human Biology*, 72(6), 997–1016.
6. Pope, A. M., Carr, S. M., Smith, K. N., & Marshall, H. D. (2011). Mitogenomic and microsatellite variation in descendants of the founder population of Newfoundland: high genetic diversity in an historically isolated population. *Genome*, 54(2), 110-119.
7. Zhai, G., Zhou, J., Woods, M.O., Green, J.S., Parfrey, P., Rahman, P., & Green, R.C. (2016). Genetic structure of the Newfoundland and Labrador population: founder effects modulate variability. *European Journal of Human Genetics*, 24(7), 1063-1070.
8. Teixeira, C. & Da Rosa, V.M. (2009). The Portuguese in Canada: Diasporic challenges and adjustment. (University of Toronto Press).
9. Hillier, C. (2015, October 12). A grave discovery: Portuguese fishermen return to honour White Fleet | CBC News. Retrieved from <https://www.cbc.ca/news/canada/newfoundland-labrador/a-grave-discovery-portuguese-fishermen-return-to-honour-white-fleet-1.3266931>
10. Tapper B. (2014) “An archaeological analysis of the distribution of French fishing rooms on the Petit Nord, Newfoundland,” Masters thesis (Memorial University of Newfoundland).
11. Kuitem, M., Wallace, B.L., Lindsay, C., Scifo, A., Doeve, P., Jenkins, K., Lindauer, S., Erdil, P., Ledger, P.M., Forbes, V., Vermeeren, C., Friedrich, R. & Dee, M.W. (2022). Evidence for European presence in the Americas in ad 1021. *Nature*, 601(7893), 388-391.
12. Sigurosson, G., & Kunz, K. (Ed.). (2008). The Vinland Sagas: The Icelandic Sagas about the First Documented Voyages Across the North Atlantic. (Penguin Classics).
13. Ledger, P.M., Girdland-Flink, L. & Forbes, V. (2019). New horizons at L’Anse aux Meadows. *Proceedings of the National Academy of Sciences*, 116(31), 15341-15343.
14. Bartels, D.A., and Janzen, O.U. (1990). Micmac migration to western Newfoundland. *Can. J. Native Stud.* 10, 71–94.
15. Gilbert, E., Zurel, H., MacMillan, M.E., Mirhendi. S., Merrigan, M., O'Reilly, S., Molloy, A.M., Brody, L.C., Bodmer, W., Leach, R.A. et al (2022) in preparation.

16. Kivisild. T. (2017) The study of human Y chromosome variation through ancient DNA. *Hum Genet.* 136(5), 529-546. Epub 2017 Mar 4. Erratum in: *Hum Genet.* 2018 137(10), 863.
17. Poznik, GD, Xue, Y, Mendez, FL, Willems, TF, Massaia, A, Sayres, MAW, Ayub, Q, McCarthy, SA, Narechina, A, Kashin, A. et al. (2016). Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nature genetics*, 48(6), 593-599.
18. Grugni, V., Raveane, A., Colombo, G., Nici, C., Crobu, F., Ongaro, L., Battaglia, V., Sanna, D., Al-Zahery, N., Fiorani, O. et al. (2019). Y-chromosome and surname analyses for reconstructing past population structures: The Sardinian population as a test case. *International journal of molecular sciences*, 20(22), 5763.
19. Altena, E., Smeding, R., van der Gaag, K.J., Larmuseau, M.H.D., Decorte, R., Lao, O., Kayser, M., Kraaijenbrink, T. & de Knijff, P. (2020) The Dutch Y-chromosomal landscape. *Eur J Hum Genet* 28, 287–299.
20. Batini, C, Hallast, P, Zadik, D, Delser, PM, Benazzo, A, Ghirotto, S, Arroyo-Pardo, E, Cavalleri, G, de Knijff, P, Dupuy, BM et al. (2015). Large-scale recent expansion of European patrilineages shown by population resequencing. *Nature communications*, 6(1), 1-8.
21. ISOGG 2019; International Society of Genetic Genealogy. Y-DNA Haplogroup Tree 2019, Version: 15.73, Date: July, 11, 2020, <http://www.isogg.org/tree/>.
22. Roots S, Magri C, Kivisild T, Benuzzi G, Help H, Bermisheva M, Kutuev I, Barać L, Peričić M et al. (2004) Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in europe. *Am J Hum Genet.* Jul;75(1), 128-137.
23. Myres, N., Roots, S., Lin, A. Jarve, M., King, R.J., Kutuev, I., Cabrera, V.M., Khusnutdinova, E.K., Pshenichnov, A., Yunusbayev, B. et al. (2010). A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *Eur J Hum Genet* 19, 95–101.
24. Navarro-López, B., Granizo-Rodríguez, E., Palencia-Madrid, L., Raffone, C., Baeta, M., & de Pancorbo, M.M. (2021). Phylogeographic review of Y chromosome haplogroups in Europe. *Int J Legal Med* 135, 1675–1684.
25. Capelli, C., Redhead, N., Abernethy, J. K., Gratrix, F., Wilson, J. F., Moen, T., Hervig, T., Richards, M., Stumpf, M.P.H., Underhill, P.A. et al. (2003). AY chromosome census of the British Isles. *Current Biology*, 13(11), 979-984.
26. Moore L.T, McEvoy B, Cape E, Simms K, & Bradley D.G. (2006) A Y-chromosome signature of hegemony in Gaelic Ireland. *Am J Hum Genet.* 78(2), 334-338.
27. McEvoy, B., Brady, C., Moore, L. T., & Bradley, D. G. (2006). The scale and nature of Viking settlement in Ireland from Y-chromosome admixture analysis. *European Journal of Human Genetics*, 14(12), 1288-1294.
28. Hallast P, Batini C, Zadik D, Maisano Delser P, Wetton J.H., Arroyo-Pardo E, Cavalleri GL, de Knijff P, Destro Bisol G, Dupuy BM, et al (2015). The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. *Mol Biol Evol.*, 32(3), 661-673.

29. Rahman, P., Jones, A., Curtis, J., Bartlett, S., Peddle, L., Fernandez, B. A., & Freimer, N. B. (2003). The Newfoundland population: a unique resource for genetic investigation of complex diseases. *Human molecular genetics*, 12(suppl\_2), R167-R172.
30. Sequence Bio 2021. <https://www.nlgenomeproject.ca/>.
31. Poznik, G. (2016). Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. *bioarXiv*, doi:10.1101/088716.
32. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), s13742-015.
33. Gilbert, E., O'Reilly, S., Merrigan, M., McGettigan, D., Molloy, AM, Brody, L., Bodmer, W., Hutnik, K., Ennis, S., Lawson DJ, Wilson JF & Cavalleri, JP. (2017) The Irish DNA Atlas: Revealing Fine-Scale Population Structure and History within Ireland. *Scientific reports*, 7(1), 1-11.
34. Leslie, S., Winney, B., Hellenthal, G. Davison, D, Boumertit, A, Day, T, Hutnick, K, Rorvik, EC, Cunliffe, B, Lawson DJ et al. (2015) The fine-scale genetic structure of the British population. *Nature* 519(7543), 309–314.
35. Karczewski, KJ, Francioli, LC, Tiao, G, Cummings, BB, Alfoldi, J, Wang, Q, Collins, RL, Laricchia, KM, Ganna A, Birnbaum DP et al. (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 434–443.
36. Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M. & Chen, W-M. (2010). Robust relationship inference in genome-wide association studies, *Bioinformatics*, 26 (22), 2867–2873.
37. Nei M, & Tajima F. (1981) DNA polymorphism detectable by restriction endonucleases. *Genetics*. 97 (1), 145-163.
38. Slatkin M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*. 139 (1), 457-462
39. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
40. Blighe, K, & Lun, A (2019). PCAtools: everything Principal Components Analysis. Package 'PCAtools'.
41. Excoffier L, Smouse PE & Quattro JM. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*. 1992 131(2), 479-491.
42. Thioulouse J, Dray S, Dufour A, Siberchicot A, Jombart T & Pavoine S (2018). Multivariate Analysis of Ecological Data with ade4. Springer. New York, NY: Springer-Verlag New York.
43. Raymond, M. & Rousset, F. (1995). An Exact Test for Population Differentiation. *Evolution*. 49. 1280-1283.
44. Benjamini, Y, & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: series B (Methodological)*, 57(1), 289-300.

45. Wickham, H, Averick, M, Bryan J, Chang, W, D'Agostino McGowan L, Francois, R, Grolemund, G, Hayes, A, Henry, L, Hester, J et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686.
46. Cruciani, F, La Fratta, R, Santolamazza, P, Sellitto, D, Pascone, R, Moral, P, Watson, E, Guida, V, Colomb, EB, Zaharova, B, Lavinha, J, Vona, G, Aman, R, Cali, F, Akar, N., Richards, M, Torroni, A, Novelletto, A, & Scozzari, R. (2004). Phylogeographic analysis of haplogroup E3b (E-M215) y chromosomes reveals multiple migratory events within and out of Africa. *American journal of human genetics*, 74(5), 1014–1022.
47. Margaryan A, Lawson DJ, Sikora M, Racimo F, Rasmussen S, Moltke I, Cassidy LM, Jørsboe E, Ingason A, Pedersen MW et al. (2020) Population genomics of the Viking world. *Nature*, 585(7825), 390-396. Erratum in: (2021) *Nature*, 591, E29.
48. Manco, L, Albuquerque, J, Sousa, MF, Martiniano, R, de Oliveira, RC, Marques, S., Gomes, ., Amorim, A, Alvarez, L & Prata, MJ (2018). The Eastern side of the Westernmost Europeans: Insights from subclades within Y chromosome haplogroup JM304. *American Journal of Human Biology*, 30(2), e23082.
49. Lall, GM, Larmuseau, MHD, Wetton, JH, Batini, C, Hallast, P, Huszar, TI, Zadik, D, Aase, S, Baker, T, Balaresque, P et al. (2021) Subdividing Y-chromosome haplogroup R1a1 reveals Norse Viking dispersal lineages in Britain. *Eur J Hum Genet* 29, 512–523.
50. Bowden, G.R., Balaresque, P., King, T.E., Hansen, Z., Lee, A.C., Pergl-Wilson, G., Hurley., E., Roberts, S.J., Waite, P, Jesch, J. et al. (2008). Excavating past population structures by surname based sampling: the genetic legacy of the Vikings in northwest England. *Mol. Biol. Evol.* 25(2), 301-309.
51. FamilyTreeDNA. (n.d.). FamilyTreeDNA - Y-DNA Haplotype R-Z255. Retrieved November 17, 2021, from <https://www.familytreedna.com/public/y-dna-haplotype/R;name=R-Z255>

## Acknowledgements

The authors would like to thank all the participants who consented to participate in the Newfoundland and Labrador Genome Project for enabling this research.

This study makes use of data generated by the Irish DNA Atlas Study. A full list of the investigators who contributed to the generation of the data is available from the relevant Irish DNA Atlas papers. The work was in part funded by Science Foundation Ireland Grants 16/RC/3948 and (13/CDA/2223).

This study makes use of data generated by the PoBI project. A full list of the investigators who contributed to the generation of the data is available from the relevant PoBI papers. Part of the funding for the project was provided by the Wellcome Trust under award 088262/Z/09/Z.

## Data Availability

The genotype and sample meta-data from the Newfoundland and Labrador Genome Project (NLGP) are not publicly available due to participant recruitment conditions and consent agreements that protect the privacy of NLGP participants. Reasonable requests for access to the genotyping data should be made to Sequence Bioinformatics. Researchers interested in accessing the NLGP data are encouraged to contact Sequence Bioinformatics (rick@sequencebio.com).

## Author Contribution Statements

The authors confirm contribution to the paper as follows: study conception and design: H.Z., J.C.S., A.L.S., R.A.L., and M.S.P.; data collection: H.Z., J.C.S., S.M., R.R., R.A.L., and M.S.P.; analysis and interpretation of results: H.Z., J.C.S., C.B., R.B., M.E.M, S.D., S.M., E.G., G.L.C., R.A.L., G.M., R.R., A.L.S., and M.S.P.; draft manuscript preparation: H.Z., J.C.S., C.B., R.B., M.E.M, S.D., S.M., E.G., G.L.C., R.A.L., R.E.M.S., G.M., R.R., A.L.S., and M.S.P.. All authors reviewed the results and approved the final version of the manuscript.

## Ethical Approval

The NL cohort consists of participants recruited with informed consent under a study protocol approved by the Newfoundland and Labrador Health Research Ethics Board (Reference # 2018.243).

## Competing Interests

H.Z., M.E.M, S.D., S.M., R.A.L., G.M., R.R., and M.S.P. are full time employees and shareholders of Sequence BioInformatics, Inc.

R.B., A.L.S. and J.C.S. were paid scientific consultants employed by Sequence BioInformatics, Inc. at the time of this research.

C.B., E.G., and G.L.C. declare no competing interests.