Università degli Studi di Ferrara

DOTTORATO DI RICERCA IN
"BIOLOGIA EVOLUZIONISTICA ED ECOLOGIA"

CICLO XXXIII

Coordinatore: Prof. Barbujani Guido

UTILIZZO DELLA METODOLOGIA NGS ASSOCIATA ALLA CATTURA DEL CROMOSOMA Y PER LA CARATTERIZZAZIONE DELLA POPOLAZIONE ITALIANA NELL'ETÀ DEL FERRO

Settore Scientifico Disciplinare BIO/08

Dottoranda
Dott.ssa Vergata Chiara

Tutore
Prof. Caramelli David

Co-tutore
Prof. Francalacci Paolo

Anni 2017/2020

SOMMARIO

LEGENDA DELLE FIGURE

LEGENDA DELLE TABELLE

ABBREVIAZIONI

A	Adenina
aDNA	DNA antico
BAM	*Binary Alignment Map*
bp	*base pair*
C	Citosina
Ca^{2+}	Calcio
G	Guanina
Hg	aplogruppo
LGM	*Last Glacial Maximum*
Mb	Mega basi
MJN	*Median Joining Network*
MNPD	*Minimum Number of Pairwise Differences*
MRCA	Most Recent Common Ancestor
MSY	Regione maschio-specifica del cromosoma Y
mtDNA	DNA mitocondriale
ng	nanogrammo
NGS	*Next Generation Sequencing*
nm	nanometri
NRY	*Non-combining Region of Y chromosome*
nuDNA	DNA nucleare
NUMTs	**N***uclear-Encoded Mitochondrial Pseudogenes*
PCA	*Principal Component Analysis*
PCR	*Polymerase Chain Reaction*
rpm	*Round per minute*
SNP	*Single Nucleotide Polymorphism*
STR	*Short Tandem Repeat*
T	Timina
TC	Termociclatore
UDG	Uracil DNA Glicosilasi
VCF	*Variant Call Format*
Y-chr	Cromosoma Y
WGC	*Whole Genome Capture*
WGS	*Whole Genome Sequencing*
µg	microgrammo
µL	microlitro

CAPITOLO 1: INTRODUZIONE

Il DNA antico: resoconto di 30 anni di progressi

Da sempre l'uomo è interessato alla propria storia passata, che indaga attraverso una moltitudine di discipline, affinatesi grazie allo sviluppo, nel corso del tempo, di metodi d'indagine via via più avanzati. Tra tutti, uno dei più recenti e innovativi, è certamente l'"archeogenetica": le ricerche condotte sul DNA antico (aDNA), hanno infatti rivoluzionato l'approccio per la scoperta del passato, rappresentando la via più diretta per rispondere a molte domande in ambito evolutivo, antropologico e storico.

Con il termine "DNA antico" si fa riferimento a qualsiasi traccia di DNA degradato, proveniente da materiale organico, o estratto da campioni biologici non recenti (per esempio in una goccia di sangue coagulata, nelle tracce di cellule epiteliali lasciate su materiale inorganico, ma anche in resti fossili ossei o in tessuti mummificati). Le analisi sull'aDNA hanno raggiunto un estremo progresso nel trentennio trascorso dalla loro origine (negli anni Ottanta) ad oggi. Il primo tentativo di estrarre il DNA da un reperto antico, e più precisamente da un muscolo essiccato di quagga, avvenne nel 1984, in uno studio pionieristico portato avanti da Higuchi e collaboratori[1] che riuscirono ad isolare, dai tessuti di un campione di 140 anni, circa 230 paia basi (bp) di DNA mitocondriale (mtDNA). Prima di allora, l'interesse dell'uomo per il passato e per la propria storia evolutiva, era soddisfatto attraverso indagini non molecolari: l'antropologia, intesa come studio di tutti gli aspetti delle società umane passate e presenti, ha radici nei primi anni del 1800. Il lavoro sul quagga catturò l'attenzione internazionale dal momento che dimostrava che si potevano conservare sequenze di DNA amplificabile anche in tessuti antichi. Questa scoperta fu pertanto rapidamente seguita dall'isolamento di porzioni di DNA nucleare (nuDNA) recuperate dal tessuto muscolare di una mummia egiziana di 2400 anni[2].

L'analisi dell'aDNA è risultata tuttavia piuttosto difficoltosa prima dell'avvento di tecniche che permettessero l'amplificazione delle poche molecole presenti negli organismi antichi; infatti, in seguito alla morte di un essere vivente, si verificano molti processi diagenetici che portano il materiale biologico ad essere limitato a bassissime concentrazioni, ed altamente danneggiato. Come conseguenza, molta dell'informazione genetica che venne recuperata nei primi studi archeogenetici apparteneva a specie microbiche o fungine[3].

Solo con l'avvento di tecniche che permettessero l'amplificazione delle poche molecole presenti negli organismi antichi (prima tra tutte la PCR – *Polymerase Chain Reaction*, inventata nel 1986 da Kary Mullis) è stato possibile iniziare a ricostruire le dinamiche della storia passata. Tra le pietre miliari per lo sviluppo dell'antropologia molecolare si possono ricordare il recupero e la caratterizzazione di DNA da un osso umano[4], ma anche molti lavori condotti su larga scala che

hanno avuto lo scopo di individuare i processi migratori che hanno interessato l'Eurasia nel corso della preistoria[5,6,7,8], l'identificazione di geni con funzione specifica[9] ed anche la determinazione del sesso[10] e delle parentele[11,12].

Sebbene l'introduzione della PCR sia stata un enorme passo avanti, le dimensioni dei frammenti amplificabili di rado superavano le 100-150 bp, permettendo pertanto di ottenere solo un'esigua quantità di informazioni di sequenza. Solo con campioni molto ben preservati si poteva raggiungere l'ottenimento della sequenza completa del mtDNA, attraverso l'unione di ampliconi sovrapponibili, mentre i marcatori nucleari erano tutt'altro che di facile analisi. Inoltre, viste le concentrazioni limitate di DNA, risultava necessaria un'azione particolarmente distruttiva sul campione fossile[13].

Un punto non meno critico negli studi archeogenetici portati avanti con la metodica della PCR era rappresentato dal fatto che il DNA endogeno deposto in un organismo morto subisce processi degradativi che determinano, con alte probabilità, l'introduzione di misincorporazioni nucleotidiche nei primi cicli di arricchimento, incrementando notevolmente il rischio di letture erronee delle molecole originali[14].

La vera e propria rivoluzione negli studi sull'aDNA si è avuta con l'avvento delle metodologie di sequenziamento ultramassivo di nuova generazione (*Next Generation Sequencing* - NGS), che hanno permesso uno spostamento dell'attenzione scientifica dall'analisi di pochi e corti frammenti di DNA, a quella di interi genomi di più campioni parallelamente, durante l'"era paleogenomica"[15].

In particolare, attraverso il sequenziamento di genomi completi ad elevata profondità, o l'applicazione di metodi di arricchimento di regioni genomiche *target*, è stato possibile sormontare il problema del limitato tasso di DNA endogeno di un campione biologico degradato, rispetto al DNA ambientale e/o contaminante presente in quantità nettamente più elevate.

Al giorno d'oggi sebbene attraverso lo sviluppo tecnologico sia stato possibile recuperare l'intero DNA genomico da un enorme numero di campioni antichi, gli sforzi dei ricercatori non si sono fermati, ma continuano ad essere fondamentali per l'ottimizzazione dei processi di analisi che inevitabilmente accompagnano lo sviluppo tecnologico. Tra i perfezionamenti che hanno permesso il miglioramento della qualità del DNA sequenziabile, si possono annoverare le individuazioni dei distretti ossei che consentono la maggiore resa in termini di DNA endogeno[16], i processi di estrazione selettiva di frammenti corti[17], l'allestimento di librerie genomiche ad *hoc*[18,19,20] e le tecniche di arricchimento selettivo[21,22,23].

1.1. Le sfide dell'aDNA

1.1.1. La degradazione delle molecole

La difficoltà nella comprensione dei processi diagenetici che, come già accennato, rendono le molecole antiche estremamente danneggiate e prone alle contaminazioni da parte di numerosi substrati biologici, ha reso estremamente complicato il maneggiamento dell'aDNA e di conseguenza l'ottenimento di risultati consistenti, fino al decennio scorso.

Quando l'aDNA viene estratto e processato, tutti i fenomeni che concorrono alla degradazione molecolare si manifestano sotto diverse forme, quali una riduzione delle dimensioni dei frammenti, il blocco della replicazione e/o l'introduzione di misincorporazioni nucleotidiche da parte delle polimerasi durante la replicazione del DNA[24].

La principale causa di degradazione risiede nella perdita dei meccanismi di riparazione a cui la molecola è sottoposta *in vivo*, e che pertanto, con il passare degli anni e con l'esposizione ad ambienti avversi, determinano un progressivo danneggiamento dell'informazione contenuta nel doppio filamento. Dopo la morte di un organismo tali meccanismi vengono meno, e la combinazione delle attività di nucleasi endogene, ma anche esogene (prodotte ad esempio da microorganismi ambientali) fa sì che il doppio filamento si degradi naturalmente in molecole molto più corte. Inoltre, nelle cellule ormai morte, intervengono anche reazioni non enzimatiche che favoriscono l'ulteriore danneggiamento del DNA libero[24] (Figura 1). Con il tempo esso va incontro alla comparsa di siti abasici, a causa dell'idrolisi dei legami N-glicosidici tra la base e lo zucchero, che comportano un indebolimento dell'elica a cui consegue una ulteriore frammentazione del filamento. I principali errori nella lettura dell'informazione molecolare proveniente da reperti antichi sono le deaminazioni causate da stress idrolitici.

Alcuni studi hanno dimostrato che le deaminazioni della citosina sono piuttosto comuni nel materiale biologico estratto da materiale non fresco[14]: il risultato è la conversione della base azotata in uracile. Quando il DNA contenente queste lesioni è sottoposto ad amplificazione, si osservano transizioni C→T e G→A (le prime, tre volte più comuni delle seconde) e ciò che ne risulta è una lettura errata della sequenza nucleotidica. La perdita di residui purinici è stata dimostrata essere alla base della frammentazione del DNA, in quanto alle estremità 3' e 5' di frammenti corti si ritrova un notevole eccesso di A e T rispettivamente[25].

Esistono, in aggiunta, altre tipologie di eventi che coadiuvano l'insorgenza di modificazioni nel DNA; tra queste, molto importanti sono le lesioni ossidative causate dai radicali idrossilici, i quali producono forme mutagene della guanina che presentano maggiore complementarietà con l'adenina; come risultato, ancora una volta, i doppi filamenti possono includere misincorporazioni o addirittura, è possibile un arresto dell'estensione della molecola durante i processi replicativi[24].

Un ulteriore problema che può emergere in fase di analisi di frammenti degradati è di nuovo legato alla presenza di siti abasici, i quali possono procedere a legami crociati del DNA, tra DNA e proteine o tra lo zucchero ed un gruppo amminico situato sul filamento opposto e possono impedire l'amplificazione del DNA endogeno. Una varietà di lesioni tra cui danni ossidativi, rotture a singolo e doppio filamento, modifiche di base, legami crociati e formazione di dimeri potrebbero essere causati dalle radiazioni.

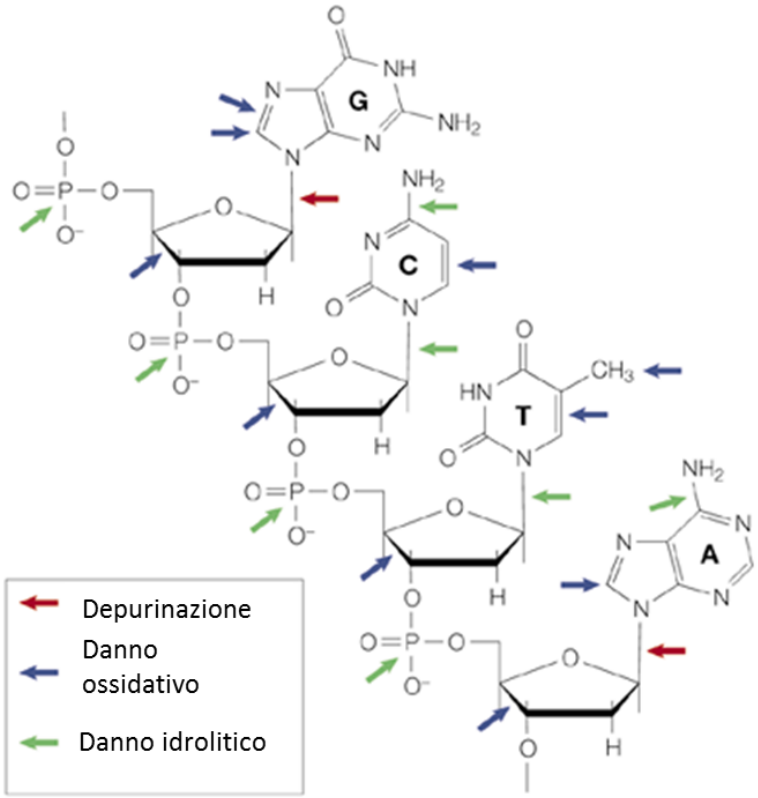

Figura 1: Principali siti di danno al DNA. (Modificata da 26).

Tutti questi processi di degradazione e frammentazione del DNA sono legati oltre che all'antichità del reperto, all'ambiente di esposizione[27]: i fattori ambientali, quali pH, temperatura, radiazioni UV, umidità, stress meccanico indotto dalla stratificazione dei sedimenti, determinano infatti un incremento della velocità di degradazione.

Oltre alle reazioni enzimatiche o non enzimatiche, un ulteriore attenzione, nell'analisi dell'aDNA, va posta alla presenza di inibitori, ovvero sostanze di natura esogena (acidi umici e tannini) o endogena (come collagene, emoglobina, e Ca^{2+}), in grado di causare un impedimento, o una diminuzione, del corretto funzionamento dei vari reattivi utilizzati nell'estrazione e/o amplificazione del DNA. Wilson, nel suo lavoro del 1997, riassume le tipologie di inibitori in tre categorie d'azione: quelli che interferiscono con la lisi cellulare, necessaria per le seguenti fasi di separazione del DNA dai siti biologici in cui si ritrova; gli inibitori che agiscono interferendo con la cattura degli acidi nucleici o agendo nella loro degradazione ed infine quelli che impediscono l'attività della polimerasi per l'amplificazione di porzioni *target* del DNA[28].

La presenza di alterazioni chimiche che derivano da quanto sopra riportato, e il conseguente inserimento di basi errate durante il processo di sequenziamento, porta a false stime dei parametri genomici come l'eterozigosità, la variabilità nucleotidica, il contenuto in GC, ed il tempo di divergenza. Per superare queste difficoltà, è necessario quindi poter disporre di numerose letture del genoma, in modo da poter inferire quella che è la reale sequenza, evitando la possibilità di associare al risultato basi diverse da quella originale.

1.1.2. Il problema delle contaminazioni

Come effetto collaterale a quanto sopra descritto, il DNA endogeno estratto da campioni antichi è spesso contaminato da una miscela di molecole endogene, in percentuali variabili, di derivazione batterica, fungina, virale, vegetale, ma anche umana[29,30]. Questo dipende dalla presenza di una quantità esigua di DNA endogeno, che favorisce il recupero del materiale genetico di organismi esogeni, portando al sequenziamento di frammenti irrilevanti per lo studio in atto. Se da un lato il problema dei contaminanti ambientali può essere aggirato utilizzando *primers* specifici in fase di amplificazione o mediante protocolli di *target enrichment*, decisamente più complesso è il problema delle contaminazioni da DNA umano moderno. Per questa ragione, sono stati definiti una serie di accorgimenti da rispettare durante tutto il processo di maneggiamento del reperto (dalla fase di scavo del campione[31,32], sino alle aree di lavoro in laboratorio[33,34,35]). Ad esempio, nelle fasi di campionamento del reperto, sono previsti processi di rimozione degli strati superficiali, e sterilizzazione mediante irradiazione con raggi UV con lunghezza d'onda a 254 nanometri (nm), che tuttavia non sempre sono applicabili (questo dipende infatti dallo stato di preservazione del campione); di recente è stata suggerita l'aggiunta, a questi *step*, di una breve fase pre-digestiva (15-20 minuti) con *buffer* costituiti prevalentemente di EDTA, che risultano efficaci per rimuovere una buona frazione di DNA esogeno[36].

Non va posta di certo minore attenzione alle contaminazioni che possono avvenire direttamente nell'ambiente di lavoro: i laboratori molecolari per lo studio dell'aDNA sono caratterizzati da importanti norme di gestione degli spazi, e del flusso di lavoro. Qui, non solo l'operatore può causare un inquinamento diretto del campione, ma anche i macchinari ed i reagenti utilizzati nelle varie fasi sperimentali, se non correttamente trattati, possono introdurre ulteriore DNA esogeno, già amplificato e presente in concentrazioni elevate su superfici ed ambienti. Per ovviare a questi problemi, oltre ad uno stringente utilizzo di dispositivi di protezione individuali da parte dei ricercatori, sono organizzate zone di lavoro, con flusso univoco, definite sulla base dello stato di amplificazione dei campioni in ambienti di pre- e post-amplificazione[33].

1.2. Tecniche e metodologie per le analisi condotte su DNA degradato

Una volta comprese le problematicità legate al maneggiamento del DNA degradato, la strada per la nascita di tecnologie estremamente sofisticate allo scopo di analizzare materiale biologico deteriorato, è stata molto breve. In questo senso, gli approcci NGS ben si sono adattati allo studio dell'aDNA, dal momento che sono in grado di sequenziare frammenti corti fino ad un minimo di circa 30 bp (dimensione comparabile con quella delle molecole estratte da materiale biologico proveniente da genomi antichi), e parallelamente determinano un aumento sostanziale dell'*output*, permettendo pertanto di superare problemi come la degradazione del campione, e la sua estrema frammentarietà, oltre alla quantità esigua. In più, le piattaforme di nuova generazione rendono possibile il sequenziamento di più librerie nella stessa corsa, grazie alla presenza di canali fisicamente separati nel *plate* di sequenziamento, o grazie all'utilizzo di *barcode* campione-specifici. Le NGS hanno permesso quindi di condurre esperimenti su interi genomi anche partendo da campioni di minor qualità attraverso approccio *shotgun*, il quale tuttavia rimane proibitivamente costoso per molti utenti[37]. Inoltre, le analisi di popolazioni antiche generalmente non si concentrano su sequenze genomiche complete, ma su loci selezionati che solitamente, in campioni antichi, si presentano in basso numero di copie. Pertanto, di pari passo all'evoluzione delle metodologie di sequenziamento ultramassivo, è stato determinante lo sviluppo di tecniche di laboratorio in grado di massimizzare la bontà del materiale *input*; tra queste, indubbiamente la tecnica di maggior successo è il *target enrichment.* La cattura del DNA tramite ibridazione consente di utilizzare in modo estremamente efficiente le tecniche NGS per analisi genetiche mirate, generando un prodotto finale che risulta selezionato in modo da contenere la maggior quantità possibile di DNA endogeno, comparato a quello esogeno[38]. In particolare, con questa tecnica è possibile generare set di dati più consistenti per più loci *target* e per più campioni in parallelo, mediante la selezione delle regioni di interesse dalle librerie genomiche prima del sequenziamento[37]. Negli approcci di cattura, in seguito alla preparazione di librerie genomiche aspecifiche, le molecole *target* vengono catturate ed immobilizzate attraverso l'uso di "esche", ovvero frammenti di DNA o RNA complementari alle regioni di interesse, al fine di arricchire il materiale selezionato all'interno dell'intera libreria precedentemente prodotta. La miscela di ibridazione (contenente la libreria di DNA e le sonde) può essere incubata in soluzione o su fase solida mediante l'uso di *array* o biglie, tuttavia la prima risulta più efficiente per le librerie di dimensioni inferiori a 500 bp[39], dal momento che le tecniche che prevedono l'uso di *array*, possono provocare un'interferenza sterica tra le molecole bersaglio e quelle non desiderate. Oltre a diverse metodologie di arricchimento, nel corso del tempo sono stati sviluppati diversi approcci: una strategia piuttosto comune per substrati che contengono quantità molto basse di DNA endogeno rispetto al DNA contaminante prevede che la stessa libreria genomica venga arricchita per più di una volta, al fine di massimizzare l'ottenimento del DNA-*target*[40].

Il raffinamento del *target enrichment* rispetto alla PCR, meccanismo d'elezione nell'epoca pre-NGS, ha prodotto diversi vantaggi. In primo luogo, con la nuova metodologia vengono notevolmente ridotti gli errori di appaiamento tra il DNA-*target* e le sonde, rispetto a quanto si verificava nel legame tra le molecole di DNA ed i *primers* in una normale reazione di PCR. Ciò è di fondamentale importanza quando si lavora con molecole degradate che, come detto, presentano *pattern* di modificazioni nucleotidiche indotte dai danni ambientali[25]. In secondo luogo, l'ibridazione è meno sensibile alla contaminazione rispetto alla PCR tradizionale: quest'ultima infatti tende a selezionare preferenzialmente molecole più lunghe (che potenzialmente rappresentano contaminanti moderni); le metodiche di arricchimento, di contro, mirano a tutte le lunghezze delle molecole di partenza in modo più equo, diminuendo la probabilità di recuperare contaminanti. Infine, attraverso la metodica della PCR, se le condizioni di legame del *primer* lo consentono, si può verificare il rischio di un'amplificazione preferenziale dei *Nuclear-Encoded Mitochondrial Pseudogenes* (NUMTs), definiti come sequenze di DNA genomico non funzionale integrate all'interno del genoma nucleare, che possiedono una stretta omologia con i reali geni mitocondriali[37].

In seguito all'ottimizzazione delle tecniche di cattura, molti studi, di varia natura (condotti su piante, animali, specie patogene e campioni recuperati in contesti archeologici) si sono concentrati sull'arricchimento del materiale di partenza. È stato così possibile recuperare interi genomi mitocondriali[41,42,43,44,45], o nucleari[46] regioni esomiche[47,48], ma anche SNPs dispersi in tutto il genoma[22,23,49,50] o interi cromosomi[23,51], che hanno portato alla produzione di una quantità di dati genetici che sarebbe risultata inconcepibile negli anni pre-NGS[52]. Inoltre, l'evoluzione di tecniche di selezione molecolare ha permesso di arricchire rigorosamente organelli o regioni genomiche di interesse per un particolare progetto di ricerca, prima del sequenziamento del DNA, favorendo un incremento della profondità del dato nelle regioni target, e parallelamente una diminuzione dei costi complessivi, aumentando esponenzialmente gli studi molecolari atti a rispondere alle domande sull'origine, l'evoluzione e la dispersione dell'uomo moderno[22,38,43,49,53,54,55,56]. Attraverso l'approccio della selezione del materiale di interesse inoltre, è stato possibile indirizzare le ricerche paleogenetiche verso il recupero di DNA anche da campioni che presentavano concentrazioni di materiale endogeno estremamente basse[21,46,57].

Grazie all'introduzione di tutti i miglioramenti tecnici descritti, l'esperienza nel campo archeogenetico si è ampliata rapidamente (Figura 2). Con gli strumenti ad oggi alla mano è quindi possibile osservare i cambiamenti nella diversità genetica degli organismi oggetto di studio attraverso il tempo e la geografia; l'aDNA può essere pertanto un'importante fonte per testare ipotesi sulle relazioni tra eventi ambientali e cambiamenti evolutivi nelle popolazioni, per risolvere controversie sui rapporti evolutivi tra specie, ma anche per rivelare relazioni altrimenti criptiche tra le popolazioni del passato e quelle attuali. Grazie all'introduzione delle tecniche NGS è ora possibile focalizzare l'attenzione delle analisi sull'aDNA su interi genomi o su specifici marcatori

molecolari di più o meno facile acquisizione, determinando l'opportunità di risolvere quesiti di nature notevolmente diverse.

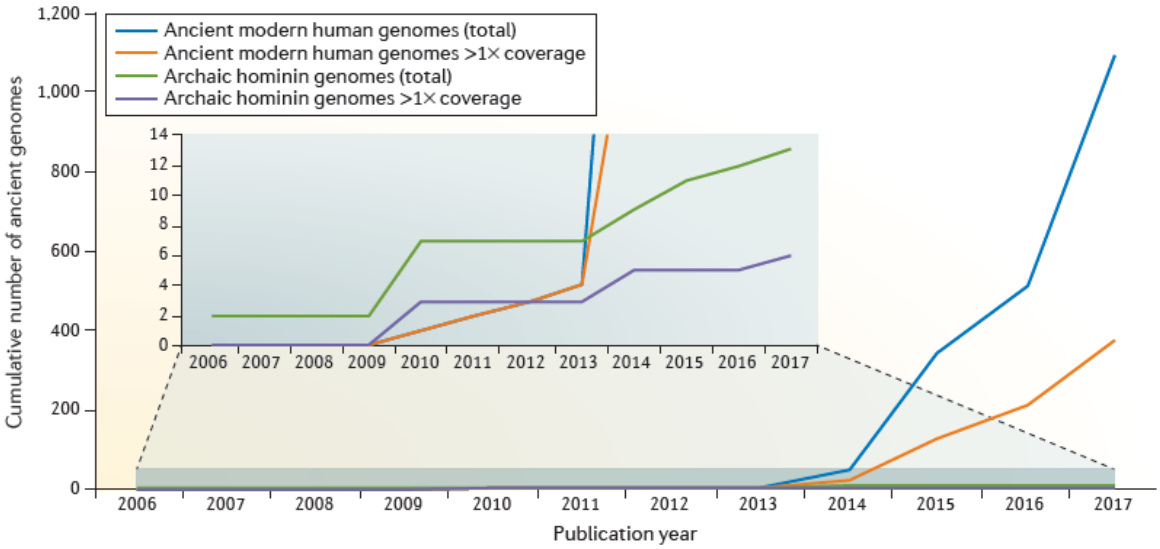

Figura 2: Andamento cumulativo del numero di dati paloegenomici disponibili (sequenziamenti di interi genomi – WGS, esomi e SNPs genomici), per anno di pubblicazione (modificata da 52).

1.3. I marcatori molecolari negli studi archeogenetici

Come già detto, difficilmente gli studi condotti sull'aDNA si basano sull'ottenimento dell'intero genoma degli individui sottoposti ad analisi, per ragioni di natura economica e pratica, ma più verosimilmente si focalizzano sull'individuazione di alcuni *markers* genetici: a seguito degli studi pionieristici condotti su marcatori classici, quali gruppi sanguigni e polimorfismi proteici[58], i progressi messi in atto nel campo della biologia molecolare, hanno permesso l'indagine diretta dei marcatori contenuti nel DNA.

Di più facile attuazione, e pertanto oggetto dei primi decenni di analisi, è lo studio dell'mtDNA, sebbene rappresenti solo lo 0,25% dell'informazione contenuta nel genoma umano attraverso un'estensione di 16569 bp. Le caratteristiche fondamentali dell'mtDNA sono l'assenza di ricombinazione (dato che si eredita esclusivamente per via materna) e la bassa presenza di meccanismi di riparazione durante i processi replicativi (che porta ad un sostanziale tasso di mutazione, utile per introdurre una buona variabilità in campioni derivanti dalla stessa linea parentale). L'accumulo di mutazioni definisce il tasso di evoluzione dell'mtDNA, ed è stato stimato attorno ad 1.57×10^{-8} sostituzioni/sito/anno per la regione codificante e 2.67×10^{-8} sostituzioni/sito/anno per l'intera molecola (tra le 10 e le 20 volte più veloce di quello attestato per il DNA nucleare)[59]. Proprio per la caratteristica delle regioni non codificanti di accumulare maggiori mutazioni, le indagini condotte sull'mtDNA si sono focalizzate inizialmente sulle differenze nelle porzioni ipervariabili[60,61,62], contenute nella regione D-loop della molecola. In seguito al miglioramento delle tecniche investigative e all'abbattimento dei costi di

sequenziamento, è stato possibile ampliare le analisi genetiche all'intero genoma circolare, al fine di ricostruire gli eventi demografici che hanno avuto luogo in Europa a partire dal Pleistocene[43,54] e che hanno modulato la variabilità mitocondriale delle popolazioni passate ed attuali[49,63,64,65,66]. Infine, l'analisi dell'mtDNA ha apportato notevoli contributi alla ricostruzione di linee di parentela: esempi di indagini ad interesse storico in questo campo, sono stati lo studio della famiglia Romanov[11], e quello eseguito su resti scheletrici attribuiti a Francesco Petrarca[12].

Sebbene un campione degradato contenga tipicamente un numero di molecole di mtDNA di tre ordini di grandezza superiore rispetto alle molecole di nuDNA, queste ultime risultano essere meno soggette alla degradazione, grazie anche all'azione preservativa delle proteine strutturali[3]. Pertanto, contestualmente alla sofisticazione degli studi sull'mtDNA, l'ottimizzazione delle tecniche di estrazione e amplificazione da substrati degradati hanno determinato il successo nell'utilizzo dei marcatori nucleari[67]. In particolare i *markers* autosomici sono stati introdotti per descrivere caratteristiche ed abitudini delle popolazioni passate attraverso l'analisi di porzioni o di interi geni, o cromosomi[68,69]. Più di recente, gli studi condotti sull'aDNA si sono concentrati sul recupero di SNPs dispersi in tutto il genoma, ovvero di porzioni del DNA in cui risulta alterato un singolo nucleotide della sequenza, con una frequenza superiore all'1%, nella popolazione esaminata, tramite l'utilizzo di centinaia di migliaia di polimorfismi contemporaneamente. Lo studio delle combinazioni di SNPs nel genoma è fondamentale per la determinazione degli aplotipi, e ha permesso di ampliare enormemente le conoscenze ad oggi disponibili sui processi di migrazione che hanno interessato l'uomo su scala globale[22]. Inoltre, tale tipologia di marcatori è di notevole interesse per le determinazione della corrispondenza tra distanze genetiche ed aree geografiche che coinvolgono diverse popolazioni. Tuttavia, l'uso degli autosomi come *markers* evolutivi complica notevolmente le ricostruzioni filogenetiche, dal momento che possono emergere misinterpretazioni generate da fattori confondenti, quali la ricombinazione genetica, la conversione genica e la selezione naturale. Inoltre, non è possibile sfruttare tutto il genoma per studi popolazionistici; risulta infatti chiaro che regioni comprendenti i geni essenziali non possono modificarsi con un rapido tasso senza generare importanti alterazioni alle funzioni metaboliche, e pertanto le regioni interessate da tali geni saranno identiche nelle popolazioni, risultando quindi utili solo per le ricostruzioni filogenetiche a più alti livelli tassonomici.

Tra i marcatori nucleari, i cromosomi sessuali sono stati in passato utilizzati principalmente al fine di ricostruire geneticamente il sesso in campioni archeologici[70]; significativo, negli ultimi anni, è stato l'affiancamento del cromosoma Y (Y-chr) all'mtDNA negli studi popolazionistici[71]. Entrambi, considerato il basso tasso di reversione, la mancanza di ricombinazione, e di conseguenza la struttura aploide, sono strumenti potenti per il recupero delle informazioni chiave relative alla storia evolutiva umana.

Le caratteristiche biologiche e genetiche del cromosoma sessuale saranno meglio eviscerate nei paragrafi seguenti, dal momento che tale marcatore è il *focus* principale di questa tesi.

1.4. Il cromosoma Y: origine, evoluzione e caratteristiche genetiche

I cromosomi sessuali umani, per quanto estremamente diversi in dimensioni e funzioni, si sono originati a partire da una coppia di autosomi ancestrali ricombinanti che tra i 160 ed i 190 milioni di anni fa sono andati incontro ad una iniziale differenziazione[72]. Quest'ultima è stata originata dalla progressiva emergenza di un gene maschio-specifico presente nel proto-cromosoma Y e nel graduale accumulo di geni a funzioni specificamente maschili, contestualmente alla soppressione, da parte della selezione naturale, dei meccanismi di ricombinazione meiotica tra i proto-cromosomi primordiali[73]. Come conseguenza di questi eventi di transizione, l'evoluzione del Y-chr è stata caratterizzata da un rapido decadimento strutturale e dalla perdita di numerosi dei suoi geni ancestrali[74]. Il risultato finale delle modifiche genetiche avvenute nei proto-cromosomi è ben saldo negli umani moderni, sebbene molti scienziati abbiano predetto la possibilità di una futura estinzione del Y-chr, come è stato evidenziato nel percorso evolutivo di alcune altre specie viventi[75,76].

Il Y-chr umano ha una lunghezza complessiva di circa 59 milioni di paia basi (Mb), che ricadono quasi interamente nella regione non ricombinante maschio-specifica (MSY); essa è composta da circa 56.4 Mb, rappresentate da una forte complessità strutturale, caratterizzata da molti elementi ripetuti e duplicazioni segmentali, e comprendente solo 78 geni codificanti proteine. Le restanti 3 Mb sono rappresentate da due regioni telomeriche pseudoautosomiche localizzate alle estremità del cromosoma, le quali possono andare incontro a fenomeni di ricombinazione con il cromosoma X (Figura 3)[77].

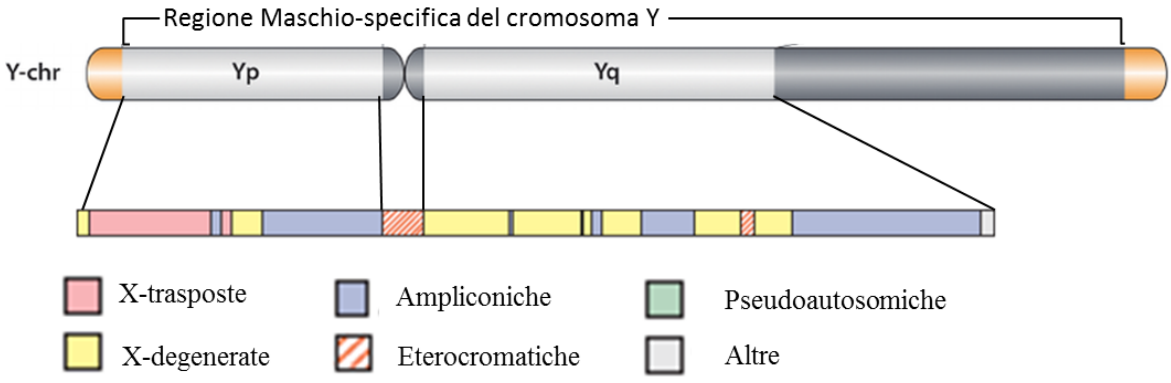

Figura 3: Rappresentazione schematica del Y-chr. La regione MSY è suddivisa in diversi colori, sulla base delle caratteristiche genetiche delle varie porzioni, come descritto in legenda. In arancione, le regioni telomeriche pseudoautosomiche del cromosoma. Immagine modificata da 78.

Nella sequenza di riferimento umana Hg 19, GRCh37 ad oggi sono state assemblate solo 23.1 Mb dell'intera regione MSY del cromosoma, dal momento che la restante porzione risulta indecifrabile essendo composta da regioni ripetute di eterocromatina costitutiva, che si possono evidenziare sia a

livello del centromero che nel braccio lungo del cromosoma, e che per ragioni pratiche risultano illeggibili[77].

La maggior parte dei tratti non ricombinanti che costituiscono l'eucromatina possono essere distinti in tre classi rappresentative (Figura 3): (i) sequenze X-trasposte (3.4 Mbp), che presentano il 99% di omologia con le sequenze localizzate nel braccio lungo del cromosoma X, e sono il risultato di fenomeni di trasposizione verificatisi dopo la divergenza tra uomo e scimpanzè; (ii) sequenze ampliconiche (9.7 Mbp); (iii) sequenze X-degenerate (8.6 Mbp), con la minore somiglianza con il cromosoma X[78]. Sebbene quest'ultime rappresentino meno della metà della porzione leggibile dell'MSY, risultano estremamente interessanti dal momento che contengono il 57.3% degli SNPs informativi[79].

Differentemente dai meccanismi di trasmissione dei cromosomi nucleari, il Y-chr non viene ereditato da entrambi i genitori; inoltre, la possibilità di evitare i fenomeni di *crossing-over* per la maggior parte della sua lunghezza lo rende il più importante fattore geneticamente dominante per la determinazione del sesso. Infatti, ad eccezione di aree ristrette, che potrebbero essere soggette ad una conversione inter-cromosomica (ICGC)[73], la mutazione è l'unico meccanismo di variazione genetica del cromosoma sessuale maschile. Le sue caratteristiche di aploidia e maschio-specificità hanno enormi influenze sulla struttura genetica, sui processi mutazionali e sulla diversità del cromosoma tra popolazioni, e all'interno della popolazione stessa[80]. Per questo il marcatore uniparentale maschile, insieme a quello femminile, sebbene sia uno strumento inappropriato per descrivere la demografia di una determinata area geografica, risulta invece ampiamente utilizzato per tracciare schemi diacronici del popolamento umano, in particolare per mettere alla luce migrazioni e contatti tra popolazioni antiche[79].

1.5. Filogenesi del cromosoma Y

Negli studi a carattere evolutivo è fondamentale datare i percorsi chiave del popolamento umano e dei cambiamenti demografici che hanno plasmato le popolazioni attuali. Tale obiettivo è particolarmente interessante nel caso della regione MSY del cromosoma, dal momento che presenta caratteristiche uniche, come già discusso nel paragrafo precedente: la sua variazione attuale si può far risiedere in un singolo individuo vissuto nel passato e portatore dell'antenato comune più recente (MRCA) di tutti i Y-chr attualmente esistenti. Poiché la maggior parte del Y-chr non ricombina durante la meiosi, è possibile definire la discendenza gerarchica di tutte le variazioni del cromosoma umano da un MRCA e, su queste basi, costruire un albero filogenetico[81]. Come conseguenza, la variabilità dell'MSY odierna riflette l'origine ed espansione di una specifica linea genetica del Y-chr, ed è pertanto determinante per permettere inferenze relative alle migrazioni e alle variazioni demografiche globali[77]. I tratti di interesse filogenetico del Y-chr contemporaneo possono essere assunti come marcatori aploidi neutri, ed è pertanto possibile applicare i principi dell'orologio molecolare per inferire le tempistiche degli eventi evolutivi. Infatti, secondo la teoria neutrale dell'evoluzione, il tasso di fissazione di una mutazione neutra è indipendente dalla dimensione della popolazione, quindi non risulta influenzata dalla deriva genetica, ma è per assunto uguale al tasso di mutazione, unico parametro che influenza il tasso di evoluzione della sequenza[82]. A tal proposito, numerosi studi hanno permesso di stimare il tasso di mutazione per gli SNPs del Y-chr attorno a 0.7×10^{-9} bp per anno[83,84,85]. Attraverso l'accumulo di mutazioni neutre un nuovo aplotipo in espansione dà vita ad una popolazione di aplotipi strettamente correlati (o sub-aplogruppi – Hg): la quantità di diversità intra-allelica tra gli aplotipi nei sub-Hg è proporzionale alla distanza temporale da un unico antenato comune.

Nei primi studi condotti allo scopo di indagare la filogenesi e i tempi di divergenza del cromosoma maschile, è stato possibile classificare le variazioni genetiche in un numero piuttosto basso di aplotipi, raggruppati in un altrettanto esiguo numero di Hg (caratterizzati da una o poche mutazioni chiave, ereditate da un antenato comune), a causa della modesta quantità di marcatori disponibili[86,87,88,89]. In particolare, sebbene il livello di risoluzione dell'albero MSY sia stato notevolmente aumentato con l'avanzare delle ricerche, per molto tempo la sua spina dorsale basale è rimasta sostanzialmente invariata; la prima ramificazione dell'albero filogenetico del MSY propose la separazione del clade specifico africano A dal clade BT, ed a seguire la suddivisione di quest'ultimo nei cladi B, principalmente africani e CT, comprendente la maggior parte dei cromosomi africani e tutti i cromosomi non africani. Questo modello di ramificazione, insieme alla distribuzione geografica dei principali cladi A, B e CT, è stato interpretato a sostegno di un'origine africana degli esseri umani anatomicamente moderni, ed ha permesso di datare inizialmente l'uscita dall'Africa in un periodo compreso tra 35000 e 89000 anni fa[86]. Successivamente, è stato possibile ampliare l'albero filogenetico del Y-chr grazie alla scoperta del gruppo A00, con caratteristiche

ancestrali rispetto al primo nodo dell'albero precedentemente noto[90,91]. La riorganizzazione dell'albero genealogico del Y-chr ha implicato quindi che le linee classificate all'interno dell'Hg A non formassero necessariamente un clade monofiletico, ma un insieme di profili, tutti caratterizzati dell'assenza dei marcatori che definiscono l'Hg BT, e all'interno dei quali sono inclusi i Y-chr degli individui più antichi. Più di recente è stato ipotizzato che la divisione iniziale del clade CT possa essere iniziata prima dell'uscita dall'Africa, e che le tre linee da essa originatisi (C, D e FT – Figura 4), fossero già portate dagli antenati dei non-africani attuali[92].

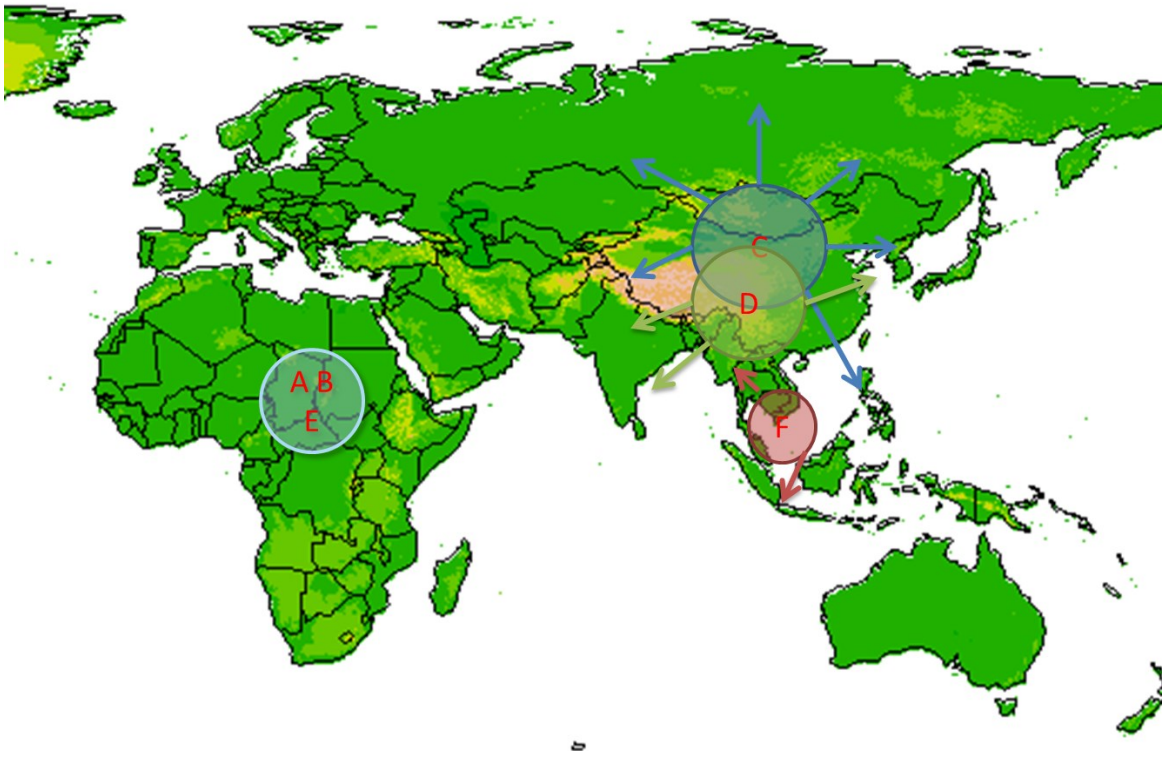

Figura 4: Rappresentazione dell'origine e diffusione dei primi Hg del Y-chr.

Molti sono stati gli alberi costruiti indipendentemente da vari gruppi di ricerca, tutti coerenti: risultano infatti descritti da una massiccia espansione di linee filogenetiche del Y-chr non africane durante l'intervallo chiave di 50000-60000[93] anni fa.

In molti casi tuttavia, gli aplotipi analizzati hanno compreso un numero molto elevato di individui con variabilità interna minima o nulla. Pertanto, la scarsa variabilità di molti Hg ha ostacolato la possibilità di utilizzare SNPs biallelici come *markers* per determinare la cronologia della filogenesi del Y-chr. Per superare questo problema, i ricercatori si sono focalizzati sull'uso di marcatori in rapida evoluzione, come gli *Short Tandem Repeats* (STRs) per fornire un'ulteriore stima cronologica. Uno dei problemi principali dell'uso degli Y-STRs per il calcolo del tasso di mutazione è che la loro variabilità si satura rapidamente, portando a valutazioni temporali troppo recenti[94].

L'effetto di casualità dovuto ai pochi marcatori MSY analizzabili si è attenuato con l'evoluzione delle analisi NGS, che hanno permesso un sostanziale incremento (nell'ordine delle migliaia) del

numero di Y-SNPs scoperti. L'uniformità nella lunghezza dei rami osservata nelle filogenesi ottenute con metodi ad alta risoluzione è coerente con un tasso filogenetico costante degli SNPs in diverse linee genealogiche e può pertanto essere efficacemente utilizzata come orologio molecolare per la datazione dei punti di diramazione[85,95]. Sono stati inoltre proposti algoritmi specifici, come la statistica rho[96] o BEAST[97], per inferire l'età dell'MRCA per una data biforcazione tra due linee filogenetiche, ma a questo scopo, il riconoscimento di un corretto tasso di mutazione è cruciale. Le stime sull'età del Y-MRCA dipendono in modo cruciale dall'Hg più arcaico noto esistente nelle popolazioni contemporanee. Le stime basate sulle informazioni ad oggi disponibili si attestano attorno ai 254000 anni (95% CI 192000-307000), compatibili con il tempo di comparsa e dispersione precoce di *Homo sapiens*[85].

La stima del Y-MRCA è in continua evoluzione, ed è strettamente dipendente da alcuni fattori chiave quali la scoperta di nuovi campioni analizzabili che potrebbero permettere di scoprire linee del Y-chr divergenti e precedentemente sconosciute, permettendo di retro-datare l'antenato comune. Inoltre, la scoperta di ulteriori mutazioni radicali in linee già conosciute potrebbe portare ad un ri-arrangiamento dell'albero genealogico, così come la revisione del tasso di mutazione del Y-chr potrebbe modificare la stima del tempo in cui è vissuto l'antenato portatore del cromosoma sessuale comune a tutti i cromosomi Y attuali.

Con le conoscenze ad oggi disponibili è stato possibile produrre un albero filogenetico estremamente dettagliato, ma non ancora completo, in cui le ramificazioni principali degli Hg sono mostrati in Figura 5[92].

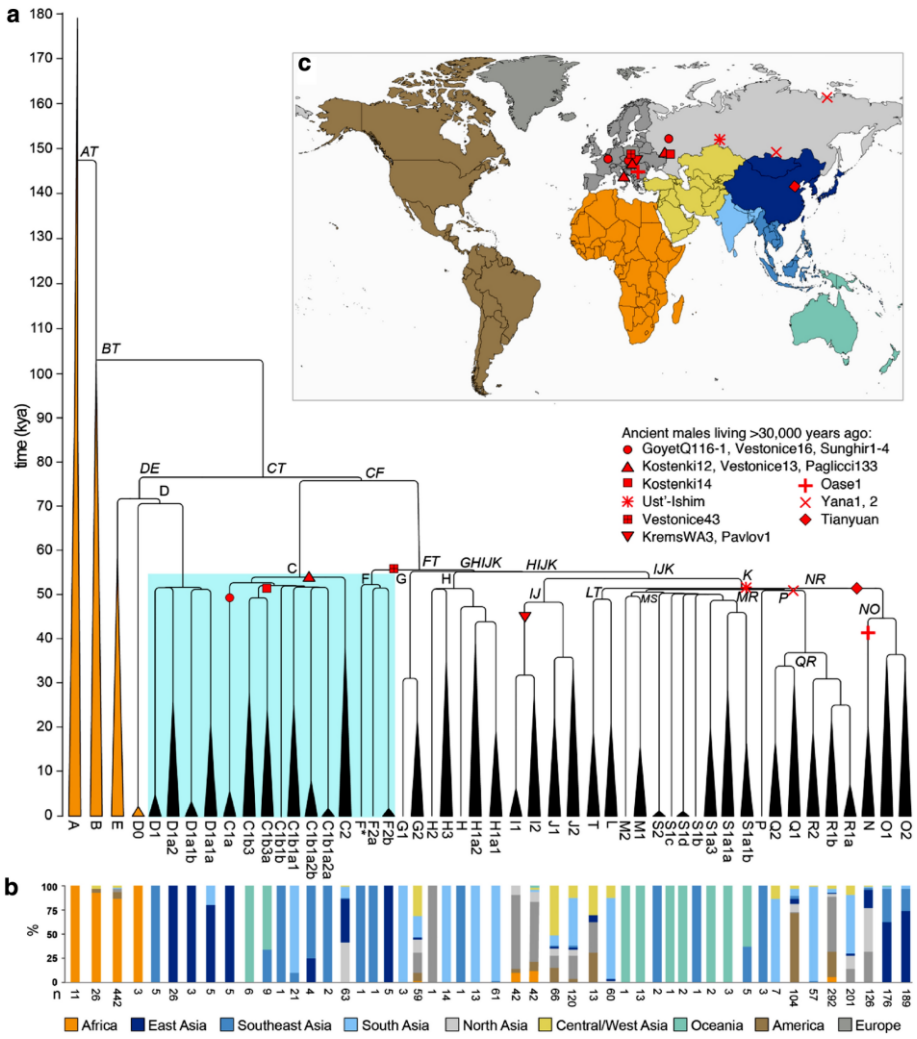

Figura 5: a) Filogenesi e distribuzione degli Hg del Y-chr ottenuta mediante l'analisi di 2319 campioni[93]. b) Proporzione di campioni analizzati che mostrano lo specifico Hg definito in a), colorati sulla base dell'origine geografica. c) Mappa mondiale integrata con i colori usati in b) e con i genomi antichi utilizzati per la produzione dell'albero. Da 93

Attraverso le conoscenze acquisite relativamente alla filogenesi del Y-chr, e all'ampliamento delle ricerche in materia, gli studiosi possono essere in grado di fare luce su argomenti peculiari come le espansioni umane e gli eventi migratori del passato, l'origine dei cognomi e molte altre caratteristiche ereditabili per via paterna[98].

1.6. Utilizzo del cromosoma Y in contesti di DNA degradato

La proprietà del Y-chr di essere specificamente maschile determina la capacità di poter evidenziare nelle popolazioni ancestrali eventi sociali, incluse la dominanza maschile in alcune culture ed eventuali *inbreeding* tra gruppi diversi, che possono aver portato ad eventi culturali di varia natura (migrazioni, unione di culture, ecc..)[99].

Definendo la distribuzione geografica delle linee filogenetiche del Y-chr (il cosiddetto approccio filo-geografico) nelle popolazioni umane del passato e del presente, è stato possibile ricostruire antiche migrazioni ed eventi concorsi nel popolamento del pianeta[100]. Non sono tuttavia molti gli studi in letteratura che si focalizzano esclusivamente sul Y-chr di campioni antichi, sebbene negli ultimi 2 anni, i dati disponibili per il cromosoma sessuale maschile abbiano iniziato ad accumularsi rapidamente, portando ad oggi, ad una disponibilità di sequenze quattro volte maggiore di quelle depositate fino al 2017[101]. Il notevole aumento di dati prodotti sul Y-chr è stato indubbiamente coadiuvato dal progresso tecnologico[80]: le NGS hanno consentito il ri-sequenziamento di grandi frammenti del Y-chr, risolvendo così il *bias* di accertamento che ha influenzato molti studi precedenti, i quali coinvolgevano un numero limitato di marcatori[102].

Gli studi disponibili in letteratura, anche se non centrati sul marcatore genetico maschile, offrono uno scenario comprensivo delle sue potenzialità in molteplici contesti di analisi. Diversi studi sull'aDNA hanno rivelato l'utilità dei dati del Y-chr in combinazione con altri marcatori (sia uniparentali che autosomici) per domande specifiche sugli eventi fondatori della storia della popolazione umana, e sui processi di dispersione e di commistione sesso-specifici[65,103]. In uno studio recente, i dati dell'intero genoma di 204 individui antichi dell'Europa Sud-Orientale hanno dimostrato la natura complessa del processo di neolitizzazione avvenuto nel continente. L'analisi e il confronto dei marcatori sul Y-chr con quelli recuperabili dall'mtDNA, dagli autosomi e dal cromosoma X, hanno suggerito che gruppi di agricoltori si sono stabiliti nell'Europa Sud-Orientale mescolandosi geneticamente con cacciatori-raccoglitori locali attraverso un bilanciamento nel numero di maschi e femmine in ingresso nel gruppo, in contrasto con quanto osservato nella regione settentrionale e occidentale del continente, in cui si assiste ad una commistione di tali gruppi entranti con cacciatori-raccoglitori prettamente di sesso maschile[66].

I cambiamenti temporali nella composizione degli Hg del Y-chr possono chiarire inoltre gli eventi di migrazione, l'introduzione di nuove linee filogenetiche ed eventuali turnover della popolazione. Ad esempio, l'analisi di genomi antichi di campioni cananei del Libano[104] ha mostrato che l'Hg J, comune nelle popolazioni odierne del Vicino Oriente, era assente negli individui neolitici del Levante ed è emerso bruscamente durante l'età del Bronzo in Libano e Giordania[65,104]. Inoltre, la recente analisi di genomi antichi da siti minoici e micenei a Creta e in Grecia ha mostrato che l'80% dei maschi disponibili appartenevano all'Hg J[65]. Questa discendenza è rara in individui

provenienti da contesti archeologici di periodi precedenti nelle aree della Grecia e dell'Anatolia occidentale, dove l'Hg G2 risulta dominante.
In alcuni casi, i dati del Y-chr possono integrare fonti archeologiche e storiche. L'analisi del Y-chr di tre tombe musulmane del primo medioevo a Nîmes, nel Sud della Francia, ha fornito ad esempio argomenti a favore di un'ascendenza nordafricana degli individui sepolti nelle tombe, corroborando prove archeologiche e storiche[105].

Infine, l'analisi del Y-chr ha recentemente contribuito a far luce sul rapporto tra gli esseri umani e i Neanderthal. Indipendentemente dagli episodi di mescolanza tra i due gruppi, che spiegano l'1-4% dell'eredità genetica neandertaliana nei genomi europei e asiatici odierni[106], l'analisi del Y-chr di un campione di sesso maschile di 49000 anni scavato ad El Sidron, in Spagna, ha dimostrato che non vi è traccia genetica del cromosoma sessuale maschile neandertaliano nella popolazione umana moderna[107]. Lo studio ha suggerito che varianti peculiari in tre geni che producono antigeni maschio-specifici avrebbero prevenuto la trasmissione dei Y-chr di Neanderthal alla progenie maschile ibrida, provocando risposte immunitarie (e aborti) nelle donne in gravidanza[107,108].

Le evidenze recuperabili dal Y-chr possono integrare e completare i modelli di variazione genetica rivelati dalla sua controparte femminile, l'mtDNA, offrendo l'opportunità di indagare eventi evolutivi influenzati dal sesso. In particolare, a causa della grande variabilità tra le popolazioni e della significativa patrilocalità in molte società umane, la variazione del Y-chr è altamente strutturata attraverso intervalli geografici e possiede segnali filogeografici più forti rispetto all'mtDNA[109].
Sebbene il Y-chr sia uno strumento essenziale e ad oggi ampiamente utilizzabile per analisi popolazionistiche e filogenetiche effettuate a partire da materiale degradato, non sono molti i lavori che si focalizzano solo su questo marcatore, il quale viene reso marginale nelle analisi che derivano da contesti di indagine di altre regioni genomiche *target*, e spesso utilizzato solamente per l'identificazione genetica del sesso degli individui in esame. Ad oggi comunque, è stato possibile tracciare la storia del popolamento e diffusione delle linee filogenetiche del Y-chr in varie regioni del globo attraverso inferenze derivanti dall'analisi del marcatore nelle popolazioni attuali.

1.7. Analisi popolazionistiche sul cromosoma Y nel bacino del Mediterraneo

Il Mar Mediterraneo è stato a lungo uno dei centri naturali più importanti per l'espansione dei geni e delle culture umane, grazie alla posizione geografica che ha rappresentato un crocevia per le migrazioni sin dalle prime dispersioni degli esseri umani anatomicamente moderni fuori dall'Africa[110].
La variazione genetica del Y-chr è ben descritta in molte popolazioni mediterranee, in particolare in Spagna[111,112] e in Italia[113,114,115]. Diversi studi hanno già affrontato questioni sulla storia genetica paterna delle popolazioni mediterranee, fornendo un quadro generale geograficamente strutturato

della variazione del Y-chr nel bacino del Mediterraneo in relazione con l'Europa, il Vicino e Medio Oriente e l'Africa[116,117,118]. La patrilocalità, una pratica residenziale in cui le femmine lasciano il loro gruppo natale per unirsi a quello dei partner maschi, è presente nel 70% delle società umane[88], ed è alla base dei comportamenti che influenzano la distribuzione della variabilità genetica analizzabile attraverso i marcatori del Y-chr.

Uno studio del 2015 mirato all'analisi della componente paterna ha coinvolto individui moderni di nove popolazioni del mediterraneo, al fine di evidenziare gli Hg del Y-chr che hanno concorso alla grande espansione in tutto il continente dopo il Neolitico[119], suggerendo un possibile fenomeno maschio-specifico veicolato in modo predominante da individui appartenenti agli Hg I1, R1a e R1b, durante l'età del Bronzo, ed associato alla selezione sociale[120]. Un'analisi simile, ma ristretta a livello regionale, di 1200 maschi sardi ha mostrato una variazione molto distintiva del marcatore genetico maschile all'interno della popolazione insulare rispetto allo spettro europeo. La topologia dell'albero filogenetico e la calibrazione molecolare della variabilità del Y-chr hanno permesso di dimostrare che la principale linea presente nella popolazione (I2a) è stata introdotta con la colonizzazione iniziale dell'isola circa 7700 anni fa, e che la popolazione si è espansa in periodi di tempo più recenti[79].

La ricerca su ampie regioni del Y-chr ha fornito anche approfondimenti su espansioni ed eventi migratori più recenti, soprattutto in regioni caratterizzate da reiterati episodi di colonizzazione, come si è verificato nell'Italia meridionale. A partire dall'VIII secolo a.C. la regione è stata oggetto di conquista da parte delle popolazioni greche; un lavoro del 2016 ha rilevato una chiara firma dell'ascendenza greca nella Sicilia orientale e ha permesso di affermare che tra i primi migranti la componente maschile risultava essere molto preponderante rispetto alla controparte femminile[121].

Proprio l'Italia, grazie alla sua posizione al centro del Mediterraneo e alla sua conformazione caratterizzata da un'estesa linea costiera, ha svolto un ruolo di fulcro nella storia del popolamento e delle migrazioni umane. Tali eventi sono ben saldi nella struttura genetica degli attuali abitanti della penisola, plasmata dai complessi movimenti umani intervenuti durante il Neolitico, l'età dei metalli e il periodo storico, che hanno apportato un mescolamento di diversi strati culturali e demici. In particolare, due indipendenti processi di neolitizzazione sembrano aver determinato una struttura genetica del Y-chr rappresentata da un gradiente Nord-Ovest/Sud-Est; a supporto di tale cline è stato messo in evidenza che più del 70% della diversità genetica italiana nel marcatore uniparentale maschile è distribuita lungo un gradiente di latitudine[122,123,124]. Altri lavori hanno suggerito una ancor maggiore distinzione della struttura genetica italiana del Y-chr secondo un'organizzazione in tre gruppi ben distinti e definiti, rappresentati dalla popolazione sarda, dagli abitanti delle regioni a Nord-Ovest, e da quelli del Sud-Est Italia[124]. Tale netta distinzione nel marcatore genetico maschile risulta estesa a tutto il bacino del Mediterraneo e suggerisce un *background* genetico condiviso tra l'Italia Sud-Orientale e il cluster del Mediterraneo Sud-Orientale da un lato, e tra l'Italia Nord-Occidentale e l'Europa Occidentale dall'altro lato[113].

I risultati che sono stati ottenuti attraverso le analisi sul Y-chr evidenziano l'importanza degli eventi demografici neolitici e post-neolitici (età dei metalli) nel plasmare l'attuale composizione della diversità paterna lungo il bacino del Mediterraneo.

1.8. Il cromosoma Y in analisi di identificazione personale

Attualmente, le nuove opportunità metodologiche, sono particolarmente sfruttate da una comunità mondiale di scienziati che studiano la variazione globale delle linee genetiche del Y-chr per definire eventuali varianti familiari e per identificare parentele su scala temporale genealogica e storica[98]. L'analisi forense del Y-DNA si è dimostrata ampiamente prolifica per l'identificazione personale: dopo il successo nell'individuazione di persone viventi in casi come violenza sessuale e test di paternità, gli sforzi degli scienziati si sono spostati anche verso il riconoscimento di persone vittime di disastri di massa, soldati dispersi o caduti di guerra[125].

Le analisi sul Y-chr sono pertanto una fonte preziosa di informazioni anche in contesti forensi, per una varietà di indagini che comprendono casi di valutazioni di paternità o di parentele di grado maggiore, ma anche identificazioni di individui di sesso maschile coinvolti in contesti criminali.

I parenti maschi condividono un profilo del Y-chr identico per diverse generazioni, il che rende possibile inferire discendenze bio-geografiche paterne, eseguire ricerche approfondite del DNA familiare e consentire la previsione del cognome in contesti forensi[98]. La tipizzazione del DNA del Y-chr può essere utilizzata anche in medicina legale per accertare le vittime per le quali i membri della famiglia di primo grado non sono più disponibili.

Negli studi forensi i marcatori d'elezione per l'analisi del Y-chr sono gli STRs, la cui importanza è determinata dall'alto grado di polimorfismo[126], e dalla semplicità di recupero mediante un'unica reazione di PCR[127]. Contestualmente inoltre, risulta sempre più diffuso l'utilizzo degli Y-SNPs, che per le loro caratteristiche permettono, attraverso la combinazione di loci dispersi in tutto il cromosoma, di identificare la discendenza geografica per linea paterna[128].

Attraverso l'analisi di migliaia di posizioni polimorfiche sul Y-chr è possibile effettuare studi di identificazione personale di individui di sesso maschile, attraverso il confronto con discendenti vivi e morti, soprattutto laddove la linea parentale femminile risulta persa, e non è quindi possibile sfruttare il più comune mtDNA; questo permette non solo di discriminare tra linee parentali, ma di individuare mutazioni private che consentono di distinguere anche tra un gruppo di individui tra loro relazionati.

CAPITOLO 2: IL POPOLAMENTO EUROPEO DAL NEOLTICO ALL'ETA' DEL FERRO

Come ampiamente descritto in precedenza, i progressi nel campo dell'aDNA, quali il miglioramento delle metodologie di sequenziamento e l'implementazione degli studi su genomi antichi, hanno permesso di addurre numerose ipotesi relativamente alla determinazione delle relazioni genetiche tra esseri umani, alle rotte migratorie che hanno portato all'espansione dell'uomo moderno, agli eventi di diversificazione e miscela genetica tra vari gruppi[129].

Dall'analisi dei genomi dei popoli dell'area europea, è emerso che i modelli odierni di variazione genomica sono stati infatti plasmati da diversi ed importanti eventi demografici del passato, inclusi il primo popolamento dell'Europa, la transizione Neolitica e le successive migrazioni durante l'età del Bronzo[130].

Indubbiamente uno dei primi eventi chiave che ha forgiato le civiltà umane è stato il passaggio da uno stile di vita dinamico come quello dei cacciatori-raccoglitori ad uno più sedentario, legato all'agricoltura, verificatosi durante la transizione Neolitica. Nell'Eurasia occidentale, tale rivoluzione culturale ha interessato dapprima la regione della Mezzaluna fertile del Vicino Oriente (circa 11000-12000 anni fa), per poi diffondersi verso l'Anatolia, e da qui in tutta Europa nei 6000 anni successivi[130]. In particolare, nel continente europeo è stato possibile ipotizzare due diverse linee di diffusione parallele, messe in evidenza da analisi genomiche condotte su campioni del Neolitico provenienti da siti dell'Egeo e dell'Anatolia nord-occidentale: una rotta migratoria verso l'Europa centro-meridionale lungo il Danubio, ed una verso la penisola iberica seguendo la costa mediterranea, rappresentate in Figura 6[131,132].

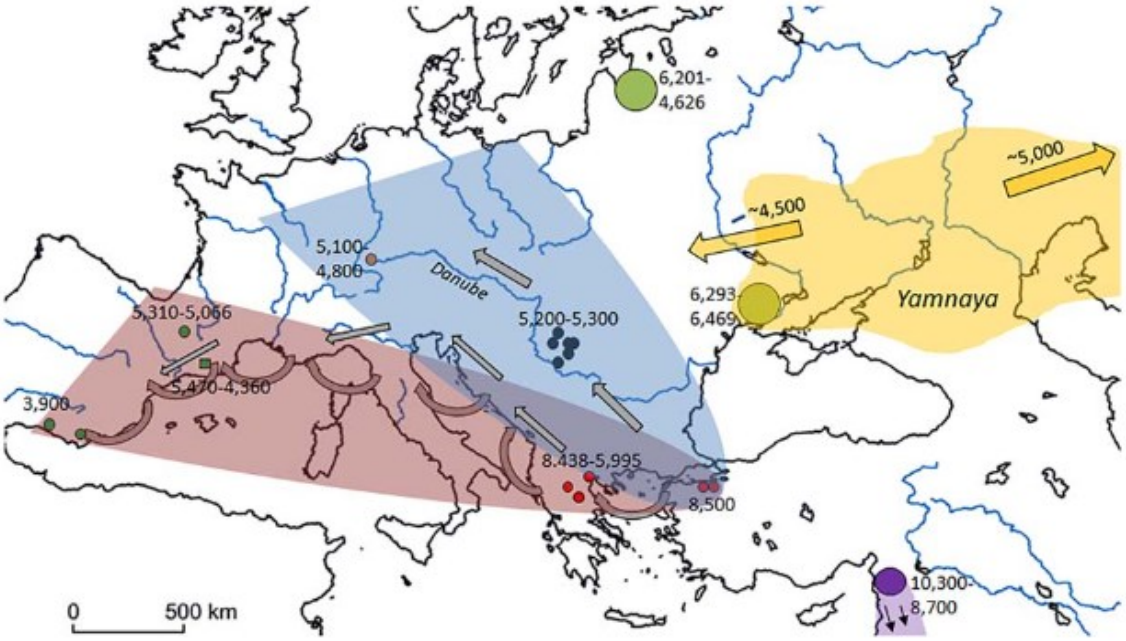

Figura 6: Corridoi migratori ipotizzati per le popolazioni neolitiche provenienti dall'Egeo e dall'Anatolia nord-occidentale (da 131).

Sono inoltre stati proposti due differenti modelli, quello della diffusione demica (prodotto da veri e propri movimenti migratori) e il modello di diffusione culturale (promosso dalla trasmissione di tecnologie agricole e stili di vita ai cacciatori-raccoglitori indigeni senza un'importante migrazione concomitante di persone), riguardanti il popolamento dell'Europa, ed in particolare l'impatto demografico che gli agricoltori del Vicino Oriente hanno avuto sui popoli della penisola europea[116]. Tra i due estremi, i modelli integrazionisti sostengono un certo grado di mescolanza tra cacciatori-raccoglitori locali e immigrati neolitici attraverso diversi meccanismi. Questi scenari si traducono in vari gradi di miscelazione genica tra popolazioni locali di cacciatori-raccoglitori e agricoltori esogeni[133].

Sebbene i primi studi condotti a tal proposito attraverso l'uso di marcatori uniparentali abbiano messo in evidenza la presenza di un cline di variabilità genetica Sud-Ovest/Nord-Est, a sostegno del modello demico, analisi genomiche più recenti, condotte attraverso l'utilizzo di campioni moderni ed antichi hanno rappresentato un quadro più complesso, nel quale insieme alla migrazione, anche i processi di assimilazione potrebbero aver giocato un ruolo importante[133].

Gli studi paleogenetici condotti su campioni attribuibili ai cacciatori-raccoglitori tardivi ed ai primi agricoltori hanno indicato una mescolanza solo limitata nelle prime popolazioni neolitiche, ma un suo aumento verso il tardo Neolitico[22,134]. Tale evidenza è risultata supportata anche dalle datazioni al radiocarbonio che hanno mostrato la presenza di comunità agricole sedentarie stabilite in Anatolia nord-occidentale e nella regione costiera dal 6600 al 6000 a.C., che non si espansero a Nord o ad Ovest dell'Egeo per altre centinaia di anni[135].

I primi agricoltori dell'Anatolia centrale, che hanno concorso alla variabilità genetica introdotta dapprima nel Vicino Oriente e poi in Europa, erano organizzati in piccoli gruppi in grado di sostenersi attraverso pratiche quali la coltivazione su piccola scala[64]. In contrapposizione ad essi, altri gruppi, quali quelli della Mezzaluna fertile orientale hanno fornito materiale genetico limitato ai primi agricoltori europei[130]. Sebbene sia probabile che lo stile di vita dei cacciatori-raccoglitori europei sia stato sostituito da quello degli agricoltori immigrati, la mescolanza tra i due gruppi si è contestualizzata in modo diverso in tutta Europa, secondo un modello che può ancora essere visto negli europei moderni[22]. Inoltre, nel Neolitico medio e nel Calcolitico precoce le popolazioni di agricoltori hanno mostrato un'ulteriore mescolanza, durata per almeno due millenni, con i cacciatori-raccoglitori mesolitici rispetto a quanto verificatosi con i primi gruppi del Neolitico[22,136].

Come nel primo Neolitico, anche durante il tardo Neolitico e la successiva età dei metalli, i dati genomici sono a supporto di eventi migratori su larga scala, che hanno determinato un impatto enorme e grandi cambiamenti sui popoli dell'Europa, tra cui la diffusione delle lingue e di alcuni tratti fenotipici. È stato infatti evidenziato che, oltre alla ricolonizzazione dell'Europa da parte dei cacciatori-raccoglitori dopo l'ultimo massimo glaciale, e gli eventi legati alla transizione neolitica, le migrazioni intervenute da Est nella prima età del Bronzo, siano stato il terzo evento che più ha influenzato la composizione e i gradienti di variazione genomica tra gli europei moderni[130]. Restano anche in questo periodo le ipotesi dei due modelli di diffusione, demica e culturale delle principali popolazioni presenti nel continente (Figura 7).

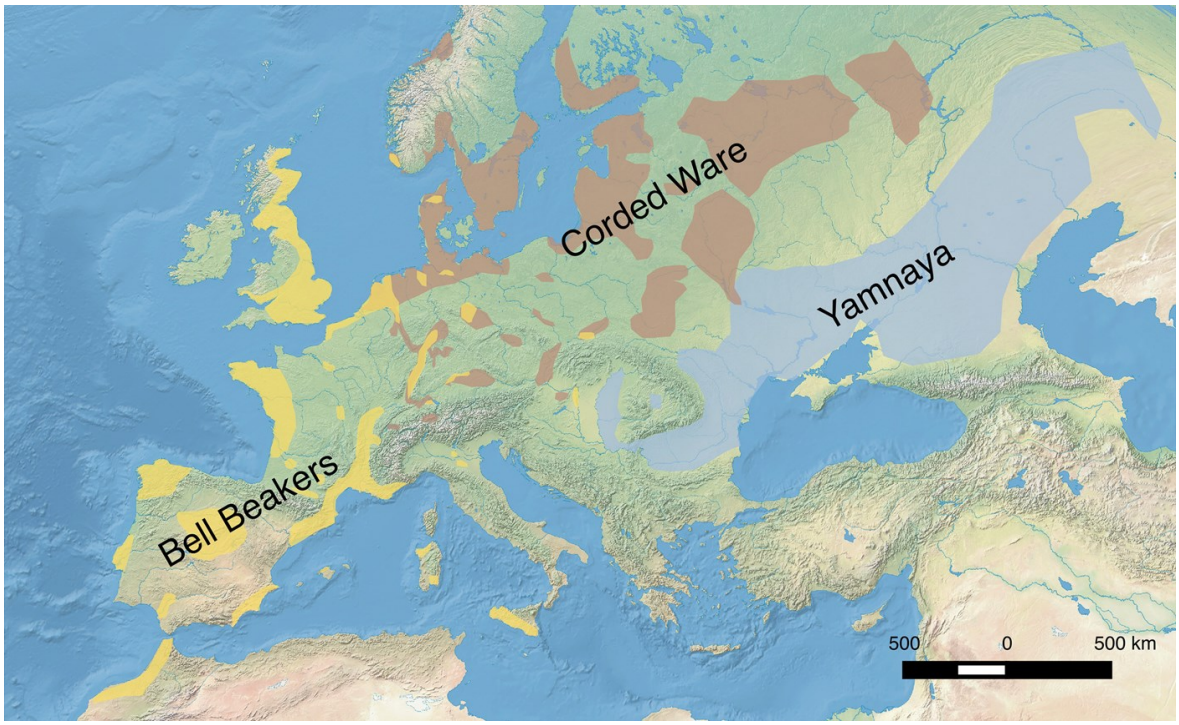

Figura 7: Distribuzione spaziale delle principali popolazioni coinvolte nei movimenti migratori dell'età del Bronzo[137].

La scoperta chiave in tal senso è stata l'individuazione di una componente genomica delle popolazioni di pastori Yamnaya delle steppe russe in popolazioni dell'Europa settentrionale[22,138]; tale comunità è responsabile di aver plasmato una nuova cultura (denominata cultura della ceramica cordata – *Corded Ware*, ma anche cultura dell'ascia da combattimento o cultura della sepoltura singola) attraverso l'introgressione, negli ultimi contadini neolitici, di una percezione completamente nuova della famiglia, della proprietà e della personalità[139]. L'informazione genetica derivante da questo contatto, sembra rivelare un'ampia mescolanza tra i popoli europei e delle steppe e non una sostituzione della popolazione su vasta scala[131].

Le analisi condotte da Allentoft e collaboratori[138] infatti, hanno evidenziato un'alta struttura genetica nell'Europa dell'età del Bronzo; in particolare, le popolazioni dell'Europa centro-settentrionale hanno mostrato una composizione derivante da una commistione dei primi gruppi di cacciatori-raccoglitori e agricoltori neolitici, a cui si è aggiunto l'*input* genetico della cultura Yamnaya all'inizio dell'età del Bronzo. Il corredo genetico euroasiatico ha raggiunto in questo modo un'importante strutturazione che, alla fine della prima età dei metalli, risultava rispecchiare già la variabilità genetica odierna[138].

Dal punto di vista culturale, l'introgressione del popolo Yamnaya in Europa corrisponde alle ipotesi sulla diffusione delle lingue indoeuropee sostenute dai linguisti, secondo cui tale trasmissione deve aver richiesto la concomitante migrazione ed il dominio (sociale o deomgrafico) dei popoli subentrati[140,141].

I risultati prodotti dalle analisi genomiche hanno indicato infine un significativo cambiamento temporale nel *pool* genetico dei popoli dell'Eurasia, e ciò ha permesso di affermare che tali gruppi fossero più strutturati delle popolazione europee contemporanee. Questo ha consentito di ipotizzare con buona affidabilità che le componenti genomiche ancestrali si siano diffuse ulteriormente nelle epoche successive all'età del Bronzo, attraverso la crescita e il continuo flusso genico tra popolazioni, che hanno portato alla bassa differenziazione osservata negli eurasiatici occidentali contemporanei[138].

Le analisi del DNA, sia su campioni moderni che su campioni antichi, hanno contribuito alla ricostruzione degli eventi demografici degli ultimi tremila anni, che hanno concorso a modellare la variabilità genetica su scala regionale[142,143]. Nelle steppe pontico-caspiche, tale periodo è risultato dominato dagli Sciti, una confederazione di diverse tribù nomadi caratterizzate dalla condivisione di elementi culturali anche su vaste distanze[144]. Nel Levante sono stati recentemente rilevati segnali di flusso genico in individui dell'età del Ferro e del periodo romano che hanno mostrato una continuità genetica con le popolazioni locali dell'età del Bronzo e odierne, ed una minore componente genetica legata a quella europea. In Europa, diversi studi si sono concentrati sulle popolazioni delle isole britanniche dell'età del Ferro e di epoche più recenti[145,146,147]. Le analisi sulle popolazioni della terraferma europea hanno messo in evidenza che esse sono state modellate da numerose migrazioni su piccola e larga scala,[143,148] che hanno prodotto il classico *pattern*

dell'isolamento a distanza, mantenuto anche negli europei moderni[149]. È stato tuttavia suggerito che, data la crescente dimensione della popolazione in Europa, questo possa aver reso le migrazioni successive meno influenti sulla demografia, dal momento che la frazione relativa dei migranti risultava progressivamente in diminuzione[130]. Nell'Europa Sud-Occidentale, gli iberici dell'età del Ferro hanno continuato a sperimentare *input* da gruppi dell'Europa centrale e settentrionale, caratterizzati da un corredo genetico ancora legato alle steppe e già diffuso attraverso il continente durante il periodo precedente[142].

Nella complicata situazione del continente euroasiatico, risulta molto interessante la posizione delle popolazioni dell'Italia dell'età del Ferro: gli individui finora analizzati hanno infatti mostrato eredità genetiche altamente eterogenee che suggeriscono molteplici eventi migratori nella regione durante questo periodo[150]; in contrasto con quanto osservato con gli individui preistorici inoltre, gli individui dell'età del Ferro sono risultati geneticamente molto simili ai moderni individui europei e mediterranei. Questo dato si è mostrato a sostegno dell'importanza centrale dell'Italia nelle connessioni con le popolazioni del Mediterraneo intervenute mediante fitte reti commerciali, di colonizzazione e di conflitti[151].

Sebbene l'attenzione negli studi archeogenetici abbia iniziato a spostarsi verso periodi storici che permettano di inquadrare la storia genetica dell'Europa e del Vicino Oriente dall'età del Ferro, ad oggi non sono molti i campioni provenienti dalla penisola italiana genotipizzati, e che possano quindi permettere di inquadrare in modo consistente il ruolo dell'Italia nella modificazione dell'assetto genetico europeo odierno.

CAPITOLO 3: SCOPO DEL LAVORO

L'analisi dell'aDNA è, dalle sue origini, limitata dalla disponibilità di informazioni recuperabili dalle molecole biologiche che inevitabilmente, a causa dei processi diagenetici, risultano estremamente danneggiate, in bassa quantità, e accompagnate da contaminanti di varia natura[24]. Sebbene negli ultimi 5 anni siano enormemente aumentati gli studi mirati alle analisi archeogenetiche attraverso l'uso di centinaia di migliaia di SNPs dispersi in tutto il genoma[22,50], questi sono principalmente legati alla disponibilità di risorse ad appannaggio di pochi gruppi di ricerca. I cromosomi a trasmissione uniparentale, mtDNA e Y-chr rimangono pertanto i marcatori elettivi per inferire le storie demografiche umane.

Nonostante le loro dimensioni, marcatamente inferiori rispetto a quelle degli autosomi e del cromosoma X, le caratteristiche biologiche che li contraddistinguono (quali l'aploidia e la mancanza di ricombinazione genetica) li rendono potenti strumenti adatti per effettuare inferenze relative ai movimenti popolazionistici, ai tempi di divergenza dal più recente antenato comune e per le ricostruzioni filogenetiche[84].

A causa dell'elevato numero di copie in ogni cellula, il mtDNA è stato per molti anni il marcatore elettivo per le analisi di substrati altamente degradati[152]; solo in seguito all'avvento di tecnologie più sofisticate è stato possibile aprire le porte al cromosoma sessuale maschile per studi di evoluzione umana, genetica di popolazione, filogenesi, medicina legale e genetica medica[80]. I due marcatori, sebbene risultino essenziali per le ricostruzioni dei movimenti migratori avvenuti nel passato, possono esprimere in alcune circostanze risultati divergenti. Questo è il caso dei dati ottenuti con le popolazioni italiane nelle quali è stata evidenziata una significativa strutturazione per le linee filogenetiche del Y-chr contrapposta ad un più omogeneo *background* della variabilità mitocondriale[114,115,116]. A questo proposito, una ricerca basata sullo studio dell'aDNA consentirebbe una migliore comprensione degli eventi passati che hanno portato all'attuale modello genetico, soprattutto dal punto di vista dell'ereditarietà maschile.

Dal punto di vista tecnico, inoltre, è importante notare che i progetti di genomica di popolazioni basati sull'arricchimento di specifiche regioni, che prendono pertanto di mira SNPs predefiniti, potrebbero generare *bias* legati alla scelta delle posizioni *target* e all'impossibilità di individuare nuove varianti, soprattutto se la selezione viene effettuata su un numero molto limitato di posizioni, o su una piccola regione, rispetto alla scelta più conservativa di catturare l'intero cromosoma di interesse[51]. Quest'ultimo è un nodo chiave in particolare per il Y-chr il quale risulta essere caratterizzato da ampie regioni ripetute di eterocromatina costitutiva, che per loro natura sono estremamente complicate da leggere, anche mediante i più sofisticati strumenti[77]. Inoltre, al giorno d'oggi, in commercio non esistono *kit* per la cattura e analisi del Y-chr ad una profondità tale da

permettere analisi filogenetiche così dettagliate da arrivare quasi all'identificazione di singolo campione ed in grado di permettere parallelamente analisi popolazionistiche.

Nella prima parte di questo dottorato, il progetto di ricerca si è pertanto focalizzato sulla messa a punto di una metodologia di selezione ed arricchimento di una larga parte del Y-chr, con l'obiettivo di dettagliare le variazioni genetiche presenti all'interno del cromosoma sessuale maschile. Lo scopo di questa prima fase di ricerca è stato pertanto duplice: dal punto di vista evolutivo, si è mirato ad ampliare le informazioni ad oggi disponibili per le ricostruzioni filogenetiche derivanti dall'analisi del marcatore paterno; inoltre, sono stati condotti differenti test sperimentali al fine di individuare l'approccio metodologico in grado di fornire maggior resa in contesti in cui il materiale biologico di partenza risulta degradato. Le sonde, prodotte per l'arricchimento delle regioni bersaglio sul Y-chr, sono state testate e successivamente utilizzate in un secondo progetto, indirizzato all'analisi genetica di una popolazione antica dal punto di vista del marcatore maschile. In particolare, questa seconda fase di ricerca si è focalizzata sulla valutazione della variabilità del Y-chr di un elevato numero di campioni recuperati in contesti archeologici associabili al periodo dell'età del Ferro in Italia, al fine di identificare eventuali clini genetici, già ipotizzati in passato[122,123,124] caratterizzanti la popolazione italiana alla fine di un lungo periodo temporale distinto da importanti movimenti migratori in tutto il bacino del Mediterraneo.

3.1 Caso studio I: La metodologia NGS associata alla cattura del cromosoma Y per analisi su aDNA

Abstract: Il Y-chr umano, fattore sessuale geneticamente dominante per la determinazione del sesso, rappresenta una fonte di informazione per studi genetici che coinvolgono diverse discipline, grazie alle caratteristiche biologiche che lo contraddistinguono dagli altri cromosomi. I principali tratti distintivi del cromosoma maschile sono la lunghezza inferiore rispetto a quella degli autosomi, la scarsa omologia con il cromosoma X, e la possibilità di andare incontro a variazioni esclusivamente attraverso eventi mutazionali. Queste proprietà, unite all'aploidia, e all'ereditarietà esclusivamente per via paterna, determinano la possibilità di individuare *pattern* di diversità genetica nelle popolazioni umane antiche, che permettono di sfruttare il cromosoma sessuale maschile come oggetto di numerose attività di ricerca. Sebbene siano molti gli studi volti all'ampliamento delle conoscenze relative alla strutturazione dell'albero filogenetico, al popolamento umano, all'identificazione personale ed a stati patologici legati al Y-chr, la bassa quantità di materiale di partenza in studi che coinvolgono l'aDNA è spesso il principale fattore limitante per la raccolta dei dati. In questo lavoro vengono confrontati due differenti disegni sperimentali e quattro protocolli per la cattura di un'ampia regione eucromatica del Y-chr, al fine di identificare la strategia con miglior resa, specificità ed uniformità nella produzione del dato. Sono stati eseguiti test su un set di campioni moderni ed antichi, per verificare l'efficienza delle sonde prodotte. I risultati ottenuti mostrano che, sebbene entrambi gli assetti abbiano una buona resa,

l'utilizzo di disegni sperimentali e protocolli specificamente sviluppati per materiale degradato aumentano drasticamente la specificità delle sonde e le informazioni deducibili, anche da campioni estremamente degradati.

***Keywords**:* aDNA, *target enrichment*, Y-chr, NGS

3.2 Caso studio II: La variabilità genetica del cromosoma Y in Italia nell'età del Ferro

Abstract: Come nelle epoche precedenti, anche durante l'età del Ferro, l'Europa e l'Italia hanno visto un periodo di estrema frammentarietà nella costituzione geopolitica, rappresentata da numerosissime civiltà e popoli coesistenti. Il ruolo di principale crocevia della penisola italiana rende il recupero e l'analisi di campioni umani di notevole interesse per poter ricostruire l'ancora vacillante storia delle migrazioni e dei rapporti tra genti che convivevano nei medesimi territori. Inoltre, l'età del Ferro fu un periodo di grandi cambiamenti culturali caratterizzati dall'adozione di nuove rotte commerciali con una piena interazione dell'Italia sia con l'Europa continentale che con le regioni del Mediterraneo orientale. Gli studi ad oggi disponibili su campioni antichi tuttavia si sono concentrati principalmente su esemplari recuperati dall'Italia continentale, in particolare Etruschi e Longobardi, o dalla Sardegna che, tuttavia, rivela una storia genetica anomala nel panorama europeo. L'obiettivo di questo lavoro è stato pertanto quello di caratterizzare geneticamente alcuni resti scheletrici provenienti da diversi siti dell'Italia peninsulare e della Sicilia, risalenti all'età del Ferro. In totale sono stati processati 44 campioni attraverso arricchimento selettivo di una vasta regione del Y-chr, mediante un set di sonde *custom* precedentemente testate per valutarne la resa. La composizione genetica paterna degli individui analizzati, presenta una variabilità divisibile in 4 principali Hg: R1b, I2, G2a e J. Sebbene le analisi prodotte fin qui siano prettamente descrittive, la variabilità geografica degli Hg osservata sembra confermare l'ipotesi di una distribuzione genetica italiana strutturata in 3 aree, rappresentative di diversi contatti con popoli stranieri: il Nord-Ovest, il Sud-Est e la Sardegna.

***Keywords**:* aDNA, Iron Age, Y-chr, Italic Population, NGS

CAPITOLO 4: CASO STUDIO I

La metodologia NGS associata alla cattura del cromosoma Y per analisi su DNA antico

4.1 Introduzione

I recenti progressi nelle tecnologie di sequenziamento ultramassivo del DNA hanno consentito di analizzare interi genomi nucleari derivanti da materiale altamente degradato. Tuttavia, il nodo chiave di tali studi restano le condizioni di preservazione dei campioni. Infatti, solo l'1% del DNA presente in librerie genomiche costruite con materiale biologico recuperato da reperti provenienti da ambienti temperati risulta essere endogeno, mentre il restante 99% è principalmente costituito da DNA contaminante di varia natura (ambientale, batterico, fungino, umano moderno)[153]. Sebbene anche con tali campioni sia possibile ottenere sufficienti informazioni da consentire analisi di popolazioni approfondite, lo sforzo economico necessario per il recupero di varianti genomiche con una copertura adeguatamente profonda risulta comunque oneroso. Per ovviare a questo problema, sono stati sviluppati numerosi metodi, applicabili a vari *step* di lavoro, per aumentare la probabilità di sequenziare la frazione endogena del campione[46]. In particolare, una prima categoria di accorgimenti sono stati messi in atto attraverso il potenziamento di protocolli *ad hoc* nelle fasi di pre-estrazione ed estrazione del DNA; tali procedimenti sono stati sviluppati per rendere più funzionali i trattamenti digestivi volti all'eliminazione delle molecole di DNA esogene[36,154] e per favorire il recupero di frammenti di DNA ultracorti[57]. Inoltre, nelle prime fasi di processamento dei campioni degradati, è risultata cruciale la messa a punto di specifici protocolli per la produzione di librerie genomiche[20]. Una seconda categoria di metodi, di più recente sviluppo, riguardano l'arricchimento del materiale *target* a partire dalle librerie genomiche aspecifiche tramite procedimenti di cattura in soluzione o su *microarrays*[22,45,47,155,156]. Attraverso queste tecniche è possibile determinare un aumento proporzionale delle sequenze di interesse, mediante una riduzione delle sequenze non *target.* I tre metodi comunemente usati (cattura su supporto solido, cattura in soluzione mediante sonde a RNA o a DNA) hanno caratteristiche diverse che possono influenzare l'efficienza, la specificità e la riproducibilità della cattura[157]. Le catture mediante *array* sono generalmente usate per arricchire *target* di dimensioni piccole o medie, come mtDNA[158], SNPs dispersi nel genoma[159], geni o esoni[160], e piccoli genomi, quali quelli di patogeni[161], mentre al momento non sono disponibili meccanismi di arricchimento basati su supporto solido per genomi di grandi dimensioni[162]. Per mezzo di questa tecnica, sonde sintetiche predisegnate, a singola elica, vengono immobilizzate in posizioni specifiche (*spots*) di particolari vetrini. Le molecole provenienti dalla libreria genomica e rappresentanti i *target*, vengono recuperati per appaiamento con i filamenti-sonda complementari, attraverso un processo che avviene a temperatura

(generalmente 65°C). Tale evento permette quindi l'eliminazione di tutti i frammenti non di interesse attraverso loro lavaggio. Normalmente, con assetti di questo tipo è possibile disporre di vetrini con più di un milione di *spots*, in ciascuno dei quali vengono immobilizzate sonde di circa 60 bp[162]. Per le caratteristiche logistiche che comporta l'ibridazione su *array* tuttavia non sono fattibili studi che coinvolgono un numero molto elevato di campioni. Inoltre, affinché la reazione abbia successo, è necessario avere concentrazioni della libreria di partenza molto alte, di circa 10-15 μg (microgrammi), indipendentemente dalle dimensioni delle molecole di interesse[163]. Pertanto, per superare questi svantaggi, sono stati sviluppati protocolli basati su metodi di arricchimento in soluzione. Il vantaggio che deriva dall'uso di questi ultimi è principalmente legato alla significativa diminuzione della quantità di materiale biologico di partenza[39]; inoltre, l'uniformità e la specificità delle sequenze ottenute da processi di cattura in soluzione tendono ad essere superiori rispetto a quelle riscontrabili per esperimenti eseguiti con supporti solidi[163]. Il principio generale su cui si basano questo secondo gruppo di processi di arricchimento è simile a quello precedentemente descritto; in particolare, sonde specifiche predefinite, costituite da corti filamenti di DNA o RNA permettono, mediante ibridazione selettiva, il recupero di frammenti complementari presenti all'interno delle librerie, mediante incubazione a temperature semi-elevate. Le sonde a DNA o RNA libere in soluzione, vengono biotinilate al fine di consentire l'isolamento dei complessi ottenuti in seguito all'ibridazione tra il DNA-*target* e la sonda, mediante il legame con sfere magnetiche ricoperte di streptavidina[39]. Finora, non esiste un protocollo standardizzato per i processi di *target enrichment*: possono essere apportate diverse modifiche alla dimensione del frammento della libreria, alle procedure di lavaggio, al numero di cicli di amplificazione e/o alla durata dell'ibridazione. Tuttavia, una serie di parametri sperimentali possono influenzare l'efficacia dei processi di arricchimento dell'aDNA. È essenziale ad esempio considerare che gli adattatori presenti alle estremità delle molecole di acido nucleico recuperate dai campioni, determinano una limitazione della *performance* nel legame tra la libreria e le sonde[20]. In secondo luogo è necessario stabilire in modo stringente le temperature di incubazione durante le fasi di cattura dal momento che rappresentano il più importante parametro che influenza l'efficienza di reazione; in particolare la specificità di legame aumenta significativamente in modo proporzionale alla temperatura, fino a circa 65°C[158]. Infine, è importante notare che per i campioni degradati, è ormai diffuso l'uso di un doppio ciclo di arricchimento della medesima libreria di partenza, al fine di incrementare la proporzione di molecole specifiche nel catturato finale[44,164]. Sono stati osservati, in tal senso, maggiori incrementi sul contenuto di materiale endogeno post-cattura a seguito di cicli di arricchimento più lunghi e a temperature inferiori, paragonati a cicli più brevi, sebbene con temperature più stringenti[165]. Questo è vero in particolare per quei campioni che mostrano un basso contenuto di DNA endogeno iniziale, mentre protocolli differenziati nelle tempistiche e temperature di ibridazione mostrano risultati paragonabili nel caso di campioni con contenuto endogeno più elevato (superiore a ~25%)[165]. Non sono state riscontrate invece sostanziali

differenze relativamente all'uso di sonde a DNA o a RNA per le catture in soluzione[166]. La tecnica della cattura per ibridazione è un approccio estremamente versatile, che consente di affrontare un'ampia gamma di studi attraverso l'uso di pannelli di sonde pronti all'uso, ma anche mediante costruzione di set di sonde specificamente disegnati per il progetto in atto, ed è risultata essenziale nei primi anni dell'era NGS per quanto riguarda le analisi sull'aDNA[158].

Sebbene negli ultimi anni i sequenziamenti di interi genomi antichi di campioni esclusivi, inclusi quelli di Neanderthal[167], dell'Homo di Denisova[53,168], di un Paleo-Eskimo[169], e di un Aborigeno australiano[170], abbiano trasformato la nostra comprensione delle migrazioni umane e rivelato mescolanze precedentemente sconosciute tra le popolazioni antiche, le catture mirate restano un potente strumento di analisi su un'ampia selezione di campioni. In questo quadro risultano estremamente interessanti i cromosomi uniparentali, in grado di fornire informazioni sugli eventi fondatori delle popolazioni attuali. Tuttavia, in confronto ai dati ad oggi prodotti attraverso arricchimento selettivo del mtDNA, i lavori pubblicati basandosi esclusivamente sul marcatore uniparentale maschile sono estremamente pochi.

Il chr-Y umano, fattore sessuale geneticamente dominante per la determinazione del sesso, rappresenta una fonte di informazione per studi genetici che coinvolgono diverse discipline, grazie alle caratteristiche biologiche che lo contraddistinguono dagli altri cromosomi. I principali tratti distintivi del cromosoma maschile sono la sua lunghezza, inferiore rispetto a quella degli autosomi e la scarsa omologia con il cromosoma X. Inoltre, ad eccezione di aree ristrette, che potrebbero essere soggette ad una conversione inter-cromosomica[73], la mutazione è l'unico meccanismo di variazione genetica del Y-chr. Queste proprietà, unite all'aploidia, e all'ereditarietà esclusivamente per via paterna, determinano la possibilità di individuare *pattern* di diversità genetica nelle popolazioni umane, che permettono di sfruttare il cromosoma sessuale maschile come oggetto di numerose attività di ricerca. Sebbene siano molti gli studi volti all'ampliamento delle conoscenze relative alla strutturazione dell'albero filogenetico[79,171,172], al popolamento umano[120], all'identificazione personale ed a stati patologici[173,174] legati al Y-chr, la bassa quantità di materiale di partenza è spesso il principale fattore limitante per la raccolta dei dati.

Pertanto questo lavoro si è proposto l'obiettivo di produrre un pannello di sonde in grado di selezionare SNPs di interesse filogenetico, localizzati nella regione X-degenerata del Y-chr, al fine di dettagliare le variazioni presenti nelle linee genetiche maschili. Sono stati prodotti due diversi disegni sperimentali delle sonde, eseguiti da due diverse aziende e volti a definire il migliore approccio per l'ottenimento di informazioni anche in campioni altamente degradati e con basse concentrazioni di DNA endogeno di partenza. Lo stesso gruppo di campioni (moderni ed antichi) sono stati arricchiti con entrambi i set di sonde prodotte, attraverso il meccanismo di ibridazione in soluzione, per il recupero di un totale di ~3Mb, al fine di testare l'efficienza e la specificità dei disegni sperimentali prodotti. Inoltre, per ciascun set di sonde sono stati testati due differenti protocolli di ibridazione, al fine di ottimizzare il recupero del materiale di interesse. I dati risultanti

da ciascun metodo sono stati analizzati utilizzando una varietà di parametri tra cui (i) la percentuale delle basi selezionate e recuperate, (ii) la percentuale di basi *on-target* sequenziate rispetto al totale (sensibilità), (iii) il numero di *reads* mappanti nelle regioni selezionate (specificità) e (iiii) la variazione generale di copertura nelle regioni di interesse (uniformità).

4.2 Materiali e metodi

4.2.1 Produzione delle sonde

Attraverso le informazioni disponibili in letteratura[79] sono stati selezionati un set di 79565 SNPs, considerati determinanti per produrre inferenze relative alla genealogia maschile. Le posizioni individuate contengono sia polimorfismi ritenuti indispensabili per l'assegnazione di una particolare linea genetica del Y-chr, sia posizioni identificabili come esclusive di singolo campione ed in grado di permettere un'identificazione a livello di aplotipo. Con questo termine si intende infatti la combinazione di varianti alleliche lungo un cromosoma o segmento cromosomico contenente loci in linkage disequilibrium; per le caratteristiche biologiche del Y-chr, gli alleli della regione non ricombinante (NRY – *Non-combining Region of Y*) sono sempre associati a formare aplotipi. Sono stati successivamente selezionati un *subset* delle posizioni totali individuate, ma in grado comunque di fornire le informazioni necessarie con un ottimo livello di risoluzione. È stato infine richiesto a due diverse aziende (Agilent Technologies e Arbor Biosciences) di produrre, sotto la nostra supervisione, un set di sonde idoneo alla cattura in soluzione dei 47551 SNPs identificati e localizzati nella regione X-degenerata della porzione eucromatica del Y-chr. Il disegno delle sonde ha previsto, per entrambe le aziende, una fase preliminare di filtraggio del dato, in cui sono stati applicati i parametri per l'eliminazione di eventuali sonde che presentavano una bassa omologia con la sequenza di riferimento e che pertanto avrebbero potuto, con maggiore probabilità, portare ad appaiamenti aspecifici lungo l'intero genoma.

Il disegno sperimentale realizzato mediante la piattaforma online SureDesign fornita dalla ditta Agilent Technologies e con il contributo tecnico degli specialisti della ditta stessa, ha previsto la produzione di 41196 sonde a RNA (SureSelect Custom DNA Target Enrichment Probes – Agilent Technologies, Santa Clara, California, USA), lunghe 120 bp, complementari alle sequenze a monte ed a valle di un subset di 29905 SNPs tra quelli precedentemente identificati (coperti da una o due sonde). Attraverso questo assetto sperimentale si è previsto di catturare ~3.5 Mb della regione NRY.

Di contro, Arbor Bioscience ha suggerito la produzione di 74809 sonde a RNA (myBaits Custom DNA-Seq –Arbor Bioscience, Michigan, USA) estese per 80 bp e con una profondità sul singolo SNPs di 3X, per la cattura di un totale di 31630 SNPs presenti nella lista fornita, ed in grado di coprire ~2.5 Mb di Y-chr.

Le principali caratteristiche dei due disegni sperimentali sono riassunte in Figura 8.

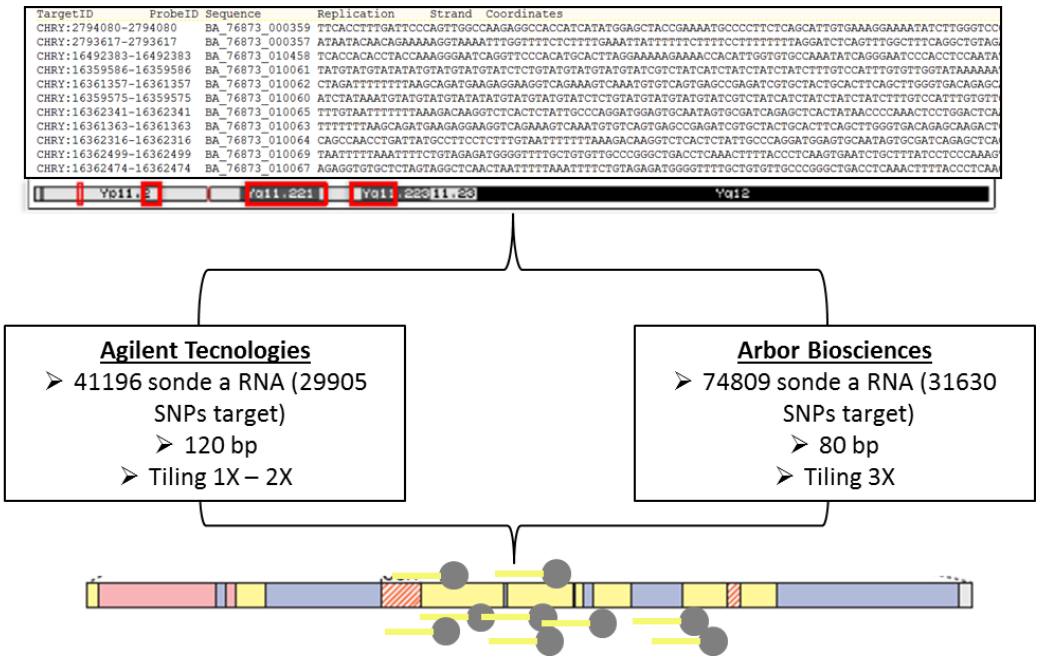

Figura 8: Principali caratteristiche distintive dei due set di sonde prodotti per la catture del Y-chr. In alto, schematizzazione del Y-chr. I box rossi rappresentano le regioni bersaglio dei disegni sperimentali.

Le sonde prodotte dalle due aziende sono state testate attraverso un set di campioni moderni ed antichi al fine di valutare la loro efficienza e specificità, sotto differenti condizioni sperimentali. Tutte le analisi sono state condotte presso il laboratorio di Antropologia Molecolare e Paleogenetica dell'Università di Firenze.

4.2.2 Selezione dei campioni di controllo

La valutazione della specificità delle sonde e dell'efficienza del protocollo adottato nelle fasi sperimentali, è stata dapprima effettuata su campioni moderni al fine di eliminare eventuali *bias* dovuti alla condizione biologica del campione, quali la perdita di siti target e l'aumento di misincorporazioni di basi azotate. Sono stati pertanto collezionati i tamponi buccali di 4 individui moderni provenienti dal centro Italia, in particolare 3 maschi tra loro imparentati con legami di primo e secondo grado (padre – figlio – nipote), ed una femmina utilizzata come controllo negativo (Tabella 1). Prima del reclutamento, è stato ottenuto un consenso informato da tutti i partecipanti (Allegato A 1), e lo studio è stato condotto in accordo con la Dichiarazione di Helsinki[175].

Sono state successivamente eseguite le stesse indagini sperimentali su campioni archeologici per valutare la specificità delle sonde anche in presenza di DNA altamente degradato e per paragonare i risultati con gli approcci sperimentali normalmente usati per analisi di genetica di popolazione.

In particolare, sono stati selezionati campioni del VI secolo, precedentemente analizzati attraverso approccio *shotgun* o di *target enrichment*, per i quali si avevano a disposizione informazioni genetiche ed analisi di parentela estremamente dettagliate[176]. Anche per i campioni antichi sono stati scelti 3 individui di sesso maschile ed una femmina utilizzata come controllo negativo. Inoltre, per avere maggiori informazioni relativamente al potere risolutivo della metodologia sviluppata,

sono stati selezionati due campioni che, dalle precedenti analisi molecolari, risultavano essere parenti di primo grado, ed un individuo appartenente ad una differente linea filogenetica (Tabella 1). I campioni antichi di sesso maschile provengono dal sito cimiteriale di Szólád, in Ungheria ed appartengono alla cultura longobarda[177]. Il campione di sesso femminile, analizzato nello stesso lavoro, è stato rinvenuto nel sito archeologico di Collegno (Torino); la necropoli di tale sito rimase in uso tra la fine del VI e l'VIII secolo[177].

ID	Tipologia	Sesso	Parentela
M1	Moderno	M	Padre
M2	Moderno	M	Figlio
M3	Moderno	M	Nipote
W	Moderno	F	-
Sz-7	aDNA	M	Figlio
Sz-11	aDNA	M	-
Sz-13	aDNA	M	Padre
Coll-87	aDNA	F	-

Tabella 1: Informazioni sui campioni selezionati per la valutazione delle sonde.

4.2.3 Preparazione dei campioni alla cattura

Gli individui moderni che hanno partecipato allo studio sono stati sottoposti al prelievo di un campione biologico attraverso il recupero di cellule epiteliali della mucosa orale mediante tampone sterile. Tutte le indagini sperimentali condotte, a partire dall'estrazione del nuDNA sono state effettuate in un laboratorio fisicamente separato da quello in cui si maneggia aDNA. Per l'estrazione del DNA da tampone salivare è stato utilizzato il QIAamp DNA Investigator Kit (Qiagen GmbH, Hilden, Germany), secondo le specifiche del protocollo fornito dall'azienda produttrice. In breve, ciascun tampone è stato inserito in un tubino da 2 mL a cui è stata aggiunta una soluzione contenente 400 µL (microlitri) di *buffer* di lisi e 20 µL di Proteinasi K, una serin-proteinasi ad ampio spettro che favorisce la digestione completa delle proteine contenute nei compartimenti cellulari, comprese quelle istoniche e le nucleasi, responsabili della degradazione del DNA endogeno una volta riversate in soluzione. L'azione lisante del *buffer* è stata coadiuvata mediante incubazione a 56°C per un'ora, durante la quale si è proceduto a miscelare ripetutamente i tubini per mezzo di un vortex, al fine di favorire il mantenimento in soluzione di tutti i componenti. Trascorso il tempo definito, si è proceduto all'aggiunta di 400 µL di *buffer* a base di guanidina, un sale caotropico in grado di veicolare il legame del DNA con la membrana in silice delle colonnine di purificazione. Sono stati poi eseguiti una serie di passaggi con *buffer* a base di etanolo, *buffer* contenenti azoturo di sodio (sostanza preservante in grado di evitare eventuali contaminazioni

microbiche nel prodotto finale) ed etanolo 100% per la purificazione massimale del DNA così estratto. Infine, il DNA legato alla membrana in silice è stato eluito in un volume finale di 50 µL con acqua di grado HPLC in nuovi tubini siliconati da 1.5 mL. Gli estratti così ottenuti per ciascun campione sono stati poi conservati in freezer, a -20°C. Durante l'estrazione è stato portato avanti anche un controllo negativo, nel quale sono stati inseriti tutti i reagenti sopra citati, ad eccezione del materiale biologico, in modo da monitorare l'eventuale presenza di contaminazioni nei reagenti utilizzati.
L'efficienza della reazione di estrazione è stata valutata mediante determinazione spettrofotometrica con Nanodrop 2000 (ThermoFisher Scientific, Waltham, Massachusetts, USA).

Per ciascun campione, il DNA estratto è stato successivamente convertito in libreria genomica attraverso SureSelectQXT Reagent Kit (Agilent Technologies, Santa Clara, California, USA)[178] sfruttando le specifiche del protocollo fornito dall'azienda. La reazione allestita si basa su un meccanismo che consente la frammentazione enzimatica delle molecole di DNA alla dimensione desiderata, e l'attacco di adattatori per analisi NGS, durante un'unica reazione.
In accordo con le istruzioni fornite dall'azienda produttrice, sono stati utilizzati un totale di 50 ng (nanogrammi) di DNA di partenza per ciascun campione. Il materiale biologico è stato inserito in tubini da 0.2 mL contenenti una soluzione composta da 2 µL di enzima e 17 µL di *buffer* necessari per la frammentazione e l'attacco degli adattatori; tale mix è stata successivamente incubata in termociclatore (TC) alla temperatura di 45°C per 10 minuti, alla conclusione dei quali sono stati addizionati 32 µL di soluzione per il blocco dell'attività delle trasposasi.
Al termine della fase di preparazione delle librerie, il materiale genomico ottenuto è stato purificato mediante biglie magnetiche Agencourt® AMPure® XP (Beckman Coulter, Brea, CA, USA), al fine di rimuovere i residui della mix di reazione. Il *buffer* in cui sono sospese le biglie magnetiche è ottimizzato per permettere il legame selettivo di frammenti di DNA di dimensioni superiori alle 100 bp alle biglie stesse. Nucleotidi, sali ed enzimi in eccesso, così come frammenti di DNA di dimensioni molto corte, possono pertanto essere rimossi utilizzando una semplice procedura di lavaggio. Il processo iniziale consiste nell'unione, in rapporto 1:1, della miscela di reazione contenente il DNA e delle biglie magnetiche, e nella loro omogeneizzazione, *step* cruciale per favorire il legame tra il DNA e le biglie. A seguito di un'incubazione di 5 minuti a temperatura ambiente, necessari a permettere il legame appena descritto, i tubini sono stati posizionati in un *rack* magnetico per circa 3 minuti, al fine di far aderire tutte le biglie alle pareti laterali e permettere la rimozione del surnatante contenente i prodotti di scarto. Successivamente, si è proceduto a due passaggi di purificazione consecutivi con 200 µL di etanolo al 70%; a seguito di ogni lavaggio, è stato eliminato il surnatante, e si è infine proceduto ad asciugare interamente il tubino in TC, con il coperchio aperto, a 37°C per 3 minuti. L'evaporazione completa dell'etanolo in questa fase è essenziale dal momento che l'alcol può inibire i successivi *step* di reazione. Il DNA purificato è stato eluito in 11 µL di acqua di grado HPLC. Il protocollo per la produzione di librerie

genomiche ha previsto infine un passaggio di arricchimento del materiale purificato mediante Herculase II Fusion Enzyme Kit (Agilent Technologies, Santa Clara, California, USA).
È stata pertanto prodotta una *mix*, costituita da 10 µL di Herculase II 5X Reaction Buffer; 0.5 µL di dNTPs mix (25mM each); 1 µL di SureSelect QXT Primer Mix; 2.5 µL di DMSO; 1 µL di Herculase II Fusion DNA Polymerase e 25 µL di acqua di grado HPLC per ciascun campione. La mix così composta è stata addizionata a 10 µL di DNA genomico e la reazione è stata incubata in TC secondo il profilo termico suggerito dall'azienda produttrice del kit: una fase iniziale di attivazione a 68°C per 2 minuti, una fase di denaturazione del DNA a 98°C per ulteriori 2 minuti, 8 cicli costituiti da denaturazione a 98°C per 30 secondi, legame dei *primer* a 57°C per 30 secondi, estensione del filamento copia a 72°C per 1 minuto, e al termine dei cicli, una fase di estensione a 72°C per 5 minuti.
Al completamento della fase di arricchimento, si è proceduto ad una ulteriore purificazione del DNA genomico attraverso biglie magnetiche Agencourt® AMPure® XP, secondo le specifiche sopra descritte, con l'unica modifica nel volume finale di eluizione (21 µL). L'eluito è stato poi quantificato mediante Agilent 2100 Bioanalyzer System (kit DNA 1000) e Nanodrop 2000, e mantenuto a -20°C fino alle successive fasi di lavoro.

Per quanto concerne i campioni antichi selezionati per la valutazione dell'efficienza e specificità delle sonde, sono state utilizzate librerie precedentemente preparate al fine di poter confrontare i dati già prodotti senza introdurre eventuali *bias* legati alle fasi sperimentali a monte della cattura del materiale di interesse. Per la produzione delle librerie era stato seguito il *workflow* normalmente applicato a campioni degradati[176]: il DNA è stato estratto mediante protocollo specifico per massimizzare l'ottenimento di DNA amplificabile ed il recupero di frammenti corti[57]. Sono state successivamente allestite librerie e incorporati indici campione-specifici come previsto per le tecniche di sequenziamento in parallelo[40].

4.2.4 Cattura del Y-chr

Per ciascun campione, gli SNPs del Y-chr precedentemente identificati sono stati selezionati dal *pool* di molecole mediante *target enrichment*, sfruttando i due set di sonde disegnati (SureSelect Custom DNA Target Enrichment Probes e myBaits Custom DNA-Seq). Per i campioni moderni sono state allestite due catture, basate sui rispettivi protocolli forniti dalle aziende produttrici delle sonde. I campioni antichi sono stati processati dapprima con i medesimi protocolli utilizzati per i campioni moderni, ed in un secondo momento con protocolli modificati al fine di favorire l'ottenimento del materiale endogeno. Ogni campione, moderno ed antico, è stato catturato in singolo, ovvero senza unire più librerie in un unico *pool* di cattura, per massimizzare il rendimento della miscela di ibridazione.

4.2.4.1 Cattura del Y-chr con sonde SureSelect

Per l'ibridazione in soluzione del DNA moderno con le sonde prodotte attraverso piattaforma online SureDesign di Agilent Technologies, è stato utilizzato il kit suggerito dall'azienda stessa e già usato nella fase di preparazione delle librerie genomiche (SureSelectQXT Reagent Kit). Per l'ibridazione e conseguente cattura delle regioni selezionate sono state seguite le specifiche indicate nel protocollo fornito dall'azienda ("protocollo A" – Figura 9). Per favorire un recupero ottimale del materiale di interesse, è necessaria una quantità di DNA di partenza, di buona qualità, compreso in un *range* tra 500 ng e 750 ng.

Il protocollo si sviluppa in una serie di passaggi sequenziali, il primo dei quali è rappresentato da un trattamento preliminare che permette di ottenere frammenti di DNA a singolo filamento attraverso il legame di oligonucleotidi complementari alle sequenze universali fiancheggianti gli inserti nelle librerie genomiche. Questa fase ha previsto l'incubazione di 12 µL di libreria con 5 µL di SureSelect QXT Fast Blocker Mix in TC a 95°C per 5 minuti (temperatura di denaturazione del DNA in filamenti singoli) seguiti da 10 minuti a 65°C. Quest'ultimo profilo termico risulta ideale affinché gli agenti bloccanti si ibridino alle regioni complementari lungo il DNA in modo stringente. La fase successiva di reazione si è svolta mediante l'aggiunta di una specifica miscela di ibridazione alle librerie genomiche bloccate a singolo filamento; con tale mix oltre alle sonde *custom* biotinilate l'ambiente di reazione viene addizionato di *buffer* che favoriscono le condizioni ottimali, ma soprattutto di proteine che inibiscono specificamente l'attività delle RNasi, evento che risulterebbe distruttivo per le sonde in uso. In particolare in ciascun tubino disposto nel TC è stata addizionata una mix di 13 µL totali, composta da 2 µL di 25% RNasi block, 2 µL di sonde *custom*, 6 µL di Fast Hyb Buffer, e 3 µL di acqua di grado HPLC. La miscela di reazione ottenuta è stata incubata in TC con un ulteriore profilo termico: 65°C per 1 minuto, 60 cicli costituiti da uno *step* di 1 minuto a 65°C ed uno di 3 secondi a 37°C.

L'isolamento e purificazione del materiale genetico catturato dalla restante miscela di DNA aspecifico e dai componenti residui della reazione, è stata effettuata mediante purificazione con biglie magnetiche. In particolare sono state utilizzate le Dynabeads™ MyOne™ Streptavidin T1 (Invitrogen, Carlsbad, California, USA), insieme ai reagenti forniti da Agilent Technologies. Brevemente, le biglie magnetiche sono state preparate al legame con il DNA *target* mediante 3 lavaggi consecutivi con 200 µL di *Binding buffer*, al termine di ciascuno dei quali il surnatante è stato scartato mediante separazione magnetica delle biglie. Il rapporto tra il volume di DNA e quello delle biglie utilizzate in questa fase di lavaggio è stato 1:1.6. Il legame tra le due componenti è stato assicurato attraverso incubazione in *plate mixer* per 30 minuti a 1800 rpm (*round per minute*). Sono seguiti 4 passaggi sequenziali di lavaggio con specifici *buffer* e la risospensione del materiale catturato in 23 µL di acqua di grado HPLC.

Infine, il protocollo SureSelect ha previsto, solo a seguito della cattura del materiale *target*, l'aggiunta a ciascun campione di indici specifici, necessari per il sequenziamento di più campioni

in parallelo, e indispensabili per associare ciascuna lettura ottenuta al campione che l'ha generata. Ogni indice è costituito da sequenze lunghe 6-8 bp ed è inserito all'interno di un *indexing-primer* che ha, da un lato, la sequenza complementare a quella degli adattatori universali Illumina (P5 e P7) e dall'altro, una sequenza utilizzabile per i successivi ed eventuali *step* di amplificazione e quantificazione. L'attacco di indici campione-specifici è possibile mediante una semplice reazione di amplificazione in cui viene prodotta una normale mix, priva però dei *primer* che vengono aggiunti separatamente per ciascun campione. Per la reazione di amplificazione è stato usato il kit Herculase II Fusion Enzyme; per ciascun campione la mix di PCR prevede: 10 µL di Herculase II 5X Reaction Buffer; 0.5 µL di dNTPs mix (100 mM); 1 µL di Herculase II Fusion DNA Polymerase; 13.5 µL di acqua di grado HPLC; 1 µL SureSelect QXT P7 dual indexing primer e 1 µL SureSelect QXT P5 dual indexing primer. Il profilo termico adottato per la reazione è stato il seguente: 98°C per 2 minuti, 12 cicli costituiti da 1 *step* a 98°C per 30 secondi, uno *step* a 58°C per 30 secondi ed 1 minuto a 72°C, ed una fase finale di estensione a 72°C per 5 minuti.
La purificazione della reazione di PCR è stata effettuata mediante biglie magnetiche Agencourt® AMPure® XP, secondo le specifiche già descritte, con due modifiche: (i) il rapporto DNA:biglie in questo caso è stato di 1:1.2 ed (ii) il volume finale di eluizione di 25 µL. L'efficienza della reazione di *indexing* è stata valutata quantitativamente e qualitativamente mediante Agilent 2100 Bioanalyzer System (kit HS).

Per i campioni antichi la cattura con sonde SureSelect è stata effettuata inserendo alcune opportune modifiche necessarie per le caratteristiche dei campioni. In primo luogo, come già detto, sono state utilizzate le librerie genomiche a doppio filamento e a doppia indicizzazione precedentemente prodotte in accordo con il protocollo di Meyer & Kircher (2010)[40]. Pertanto si è proceduto all'utilizzo del kit SureSelectQXT Reagent solo per la fase di ibridazione e cattura del materiale. In prima analisi, è stato necessario modificare la mix di ibridazione composta da SureSelect QXT Fast Blocker, addizionando la libreria genomica di oligonucleotidi per il blocco del DNA a singola elica specifici per le librerie prodotte. Inoltre, è buona regola, quando si lavora con materiale degradato aggiungere, nella miscela di reazione, agenti coadiuvanti la specificità di ibridazione, quali Human Cot-1 DNA™ (Invitrogen, Carlsbad, California, USA) e UltraPure™ Salmon Sperm DNA Solution (Invitrogen, Carlsbad, California, USA). La mix di ibridazione prodotta per i campioni antichi pertanto è stata allestita come segue: 2.5 µL Human Cot-1 DNA (1µg/µL), 0.25 µL Salmon Sperm DNA (10 µg/µL), 0.5 µL BO8.P5.part1.R (500 µM, ThermoFisher Scientific, Waltham, Massachusetts, USA), 0.5 µL BO4.P7.part1.R (500 µM, ThermoFisher Scientific, Waltham, Massachusetts, USA), 0.5 µL BO11.sslib.part2.R (500 µM, ThermoFisher Scientific, Waltham, Massachusetts, USA), 0.5 µL BO6.P7.part2.R (500 µM, ThermoFisher Scientific, Waltham, Massachusetts, USA). Infine, dal momento che il protocollo SureSelect prevede una fase di arricchimento post-cattura in cui vengono contestualmente aggiunti gli indici ai campioni, tale reazione è stata modificata, per eseguire solo un arricchimento dei catturati. Per la reazione di

amplificazione è stato usato il kit Herculase II Fusion Enzyme; per ciascun campione la mix di PCR ha previsto: 20 µL di Herculase II 5X Reaction Buffer; 1 µL di dNTPs mix (25mM each); 1 µL di Herculase II Fusion DNA Polymerase; 41 µL di acqua di grado HPLC; 4 µL IS5 (10mM); 4 µL IS6 (10mM). È stato usato il medesimo profilo termico adottato per l'arricchimento dei campioni moderni, modificando il numero di cicli complessivi a 30. L'efficienza della reazione di PCR successiva alla cattura è stata valutata quantitativamente e qualitativamente mediante Agilent 2100 Bioanalyzer System (kit HS).

Al fine di mettere a punto un protocollo in grado di garantire una resa massimale partendo da DNA altamente degradato, per i campioni antichi si è proceduto a testare un secondo metodo di arricchimento specificatamente sviluppato ("Protocollo B" – Figura 9); in particolare, la stessa procedura sperimentale è stata eseguita in doppio per ciascun campione, con l'obiettivo, nel secondo *round* di arricchimento, di favorire la specificità di legame da un sottoinsieme di molecole già selezionate. In questo caso, sono rimaste invariate le modifiche apportate al protocollo SureSelect originale, con l'esclusione dei tempi e delle temperature di ibridazione in TC: è stato eseguito un primo *round* di ibridazione a 65°C per 24 ore, al termine del quale è stata effettuata una PCR con 30 cicli di amplificazione, ed un secondo *round* a 60°C per 22 ore, al termine del quale è stata prevista un'amplificazione per 15 cicli complessivi. Nella seconda cattura, avendo già selezionato la maggior parte delle molecole di interesse da tutto il materiale biologico in soluzione, è possibile ridurre la specificità di reazione dettata dalla temperatura, per favorire una maggiore resa.

L'efficienza delle reazioni di PCR successive ad entrambi gli *step* di cattura sono state valutate quantitativamente e qualitativamente mediante Agilent 2100 Bioanalyzer System (kit HS). Infine, è stata effettuata una *reconditioning* PCR, ovvero una ri-amplificazione di un unico ciclo, del prodotto di cattura arricchito, diluito 50 volte. Ancora una volta, il risultato finale, ottenuto a seguito di purificazione con MinElute PCR purification Kit (Qiagen GmbH, Hilden, Germany), ed eluizione in 20 µL di EB Buffer (10 mM Tris-Cl, pH 8.5) è stato valutato mediante Agilent 2100 Bioanalyzer System (kit HS).

4.2.4.2 Cattura del Y-chr mediante sonde MyBaits

Le stesse librerie genomiche moderne ed antiche sono state sottoposte ad un'ulteriore cattura delle regioni di interesse del Y-chr mediante il set di sonde prodotte da Arbor Biosciences e secondo il protocollo ("Protocollo C" – Figura 9) suggerito dall'azienda (MyBaits Hybridization Capture for Targeted NGS). Per favorire un recupero ottimale del materiale di interesse, è necessario un minimo di 100 ng di DNA di partenza, di buona qualità. Il protocollo si sviluppa in una serie di passaggi sequenziali, volti a denaturare il DNA genomico e mantenerlo a singolo filamento mediante una miscela di vari acidi nucleici bloccanti e reattivi di ibridazione. La reazione finale, incubata per alcune ore in TC, ha lo scopo di favorire l'incontro tra il DNA di interesse e le sonde biotinilate a RNA. Gli ibridi DNA-sonde sono successivamente isolati dal materiale non-*target* mediante purificazione con biglie magnetiche ricoperte di streptavidina.

Nel dettaglio un totale di 7 µL di libreria genomica sono stati miscelati ad una mix di agenti bloccanti, in un volume totale di 12 µL. La reazione è stata incubata in TC a 95°C per 5 minuti e successivamente a 65°C per altri 5 minuti. Sotto tale profilo termico le molecole di DNA hanno subìto una fase di denaturazione a cui è seguito l'attacco degli oligonucleotidi alle estremità dei frammenti; l'ingombro sterico derivante dalla presenza di tali molecole ha determinato il blocco degli acidi nucleici in uno stato a singolo filamento. Al termine di tale fase alla reazione sono stati aggiunti tutti i componenti necessari per permettere l'ibridazione del DNA alle sonde (*buffer* di ibridazione e sonde), e l'intera miscela è stata incubata a 65°C per 24 ore. Al termine del periodo di incubazione si è proceduto alla purificazione del materiale ibridato dai residui di reazione. Anche tale procedimento si sviluppa in fasi sequenziali, la prima delle quali è la predisposizione delle biglie magnetiche ricoperte di streptavidina all'incontro con gli ibridi di DNA-*target*/sonde-biotinilate. Pertanto, una quantità di sonde in rapporto 1:1 con la miscela di reazione contenente il DNA sono state preparate mediante una serie di 3 lavaggi sequenziali in un *Binding buffer* a base di guanidina e successive rimozioni del surnatante per separazione della fase acquosa da quella particellare attraverso l'uso di un *rack* magnetico. Successivamente, le biglie magnetiche sono state miscelate al prodotto di ibridazione, nelle stesse condizioni di temperatura presenti nella precedente fase di incubazione, al fine di favorire la specificità di legame. Sono seguiti una serie di lavaggi con *buffer* a base di etanolo per la purificazione del materiale catturato, come previsto nel protocollo fornito dall'azienda produttrice; infine, le biglie sono state risospese in 30 µL di *buffer* TT (10 mM Tris-Cl, 0.05% TWEEN®-20 solution - pH 8.0-8.5). Per la reversione del legame tra le sonde ed il DNA ibridato, l'eluito è stato incubato in TC a 95°C per 5 minuti, al termine dei quali è stato recuperato il surnatante ed interamente utilizzato per una reazione di PCR volta ad incrementare il numero di molecole catturate. Per ciascun campione la mix di PCR ha previsto l'assemblaggio dei seguenti volumi di reagenti: 20 µL di Herculase II 5X Reaction Buffer; 1 µL di dNTPs mix (25mM each); 1 µL di Herculase II Fusion DNA Polymerase; 41 µL di acqua di grado HPLC; 4 µL IS5

(10mM); 4 µL IS6 (10mM). È stato adottato inoltre il seguente profilo termico: 95°C per 2 minuti, 15 cicli composti da una fase iniziale di denaturazione a 95°C per 30 secondi, una fase di *annealing* dei *primer* a 60°C per 30 secondi ed una fase di estensione a 72°C per 30 secondi, ed una fase finale a 72°C per 5 minuti.
I prodotti di amplificazione sono stati purificati mediante MinElute PCR purification Kit. Brevemente, i 100 µL della mix di reazione sono stati miscelati, in rapporto 1:5, ad un *buffer* di purificazione a base di guanidina (PB Buffer) per coadiuvare il legame del DNA alle membrane in silice delle colonnine MinElute. A seguito di una centrifugazione a 13200 rpm per 1 minuto, è stato scartato il percolato e sono stati effettuati 2 lavaggi sequenziali con PE Buffer, a base di etanolo. Si è proceduto infine alla eluizione del DNA in 18 µL di EB Buffer (10 mM Tris-Cl, pH 8.5), ed allo stoccaggio del purificato a -20°C.
L'efficienza della reazione di PCR successiva alla cattura è stata valutata quantitativamente e qualitativamente mediante TapeStation 4150 System (kit High Sensitivity D1000 ScreenTape – Agilent Technologies).

Per quanto riguarda i campioni antichi, l'unica modifica al protocollo che si è resa necessaria è stata l'aumento della quantità di materiale *input*; come suggerito anche da Arbor Biosciences, per campioni degradati, è stato inserito più materiale possibile, per arrivare ad una concentrazione massimale di 12 µg nei 7 µL di volume disponibile per la libreria.
Di nuovo, come già detto nel caso delle sonde precedenti, per i campioni antichi si è proceduto a testare l'efficienza anche di un protocollo di doppia-cattura ("Protocollo D" – Figura 9): la stessa procedura sperimentale è stata pertanto eseguita in doppio per ciascun campione, mantenendo invariato il protocollo appena descritto, con l'esclusione dei tempi e delle temperature di ibridazione in TC; è stato eseguito un primo *round* di ibridazione a 65°C per 24 ore, al termine delle quali si è proceduto a purificazione e arricchimento del materiale catturato per 30 cicli ed un secondo *round* a 60°C per 22 ore, seguito nuovamente da purificazione e 20 cicli di amplificazione. Anche in questo caso, i prodotti finali sono stati quantificati mediante TapeStation 4150 System (kit High Sensitivity D1000 ScreenTape).
I 4 protocolli appena descritti sono riassunti e schematizzati in Figura 9.

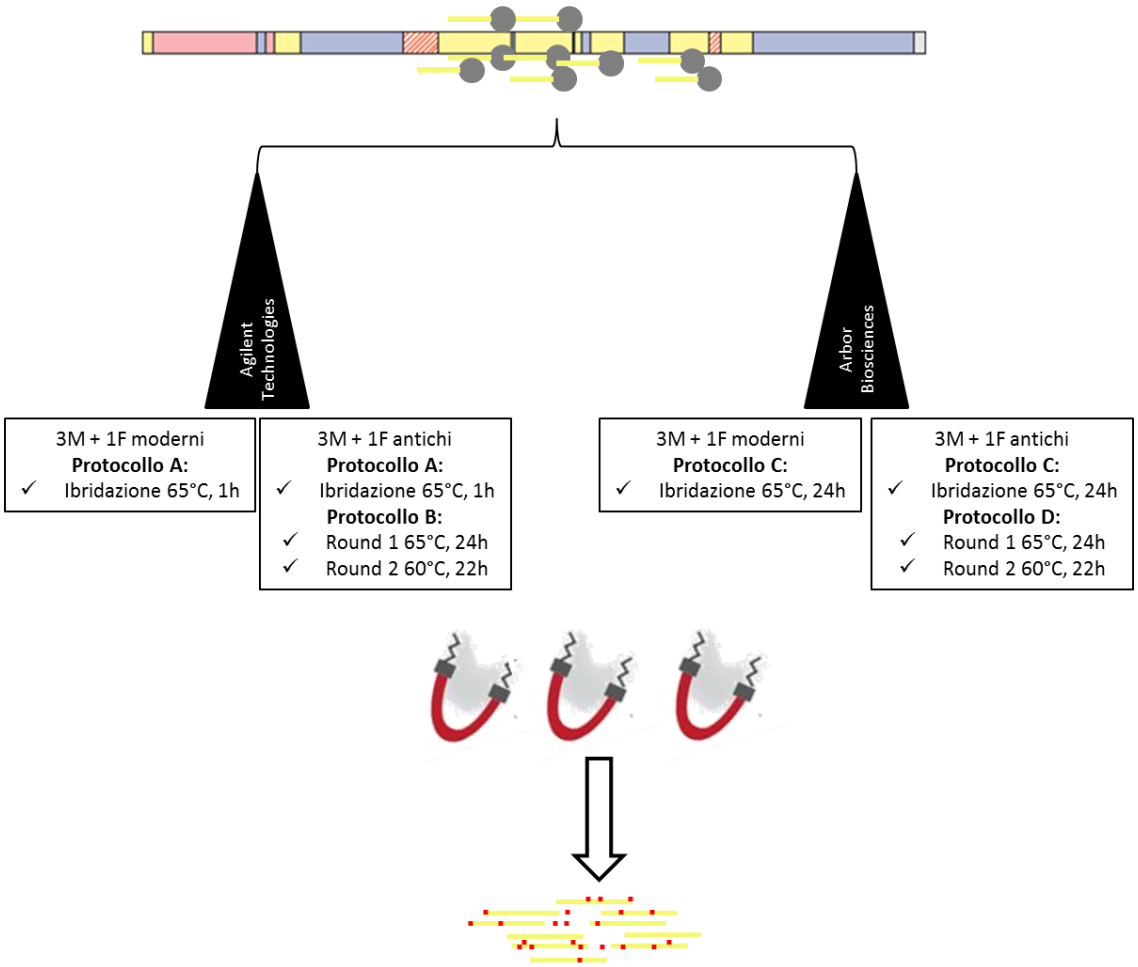

Figura 9: Schema riassuntivo dei differenti parametri di ibridazione adottati per i due diversi set di sonde disegnati.

4.2.5 Sequenziamento

Le librerie sottoposte a *target-enrichment* sono state unite in maniera equimolare in 6 distinti *pool* di sequenziamento, al fine di mantenere divisi i diversi approcci sperimentali e i set di campioni processati. Infatti, avendo utilizzato una stessa libreria di partenza, ciascun campione presenta le medesime coppie di indici in tutti i diversi protocolli di cattura in cui è stato processato.

I *pool* sono stati pertanto sequenziati mediante 6 diverse *run* (150+8+8 cicli) su piattaforma Miseq Illumina (MiSeq Reagent Kit v3, 150 cicli), utilizzando il protocollo fornito dalla casa produttrice, presso il Laboratorio di Genomica Avanzata dell'Università di Firenze. Per ogni catturato è stato previsto un *output* di 6 milioni di *reads*. I campioni moderni sono stati corsi in *single end*, mentre i campioni antichi sono stati sottoposti ad un sequenziamento in *paired end*.

Per prima cosa è stata controllata l'effettiva concentrazione di ogni catturato a 4 nM mediante Agilent 2100 Bioanalyzer System (kit HS) e Qubit™ 4 Fluorometer (dsDNA High Sensitivity Kit - Thermo Fisher Scientific – US). Una volta ottenuta la concentrazione esatta di ciascun catturato è stato possibile preparare il *pool* di sequenziamento a concentrazione finale di 2 nM. Questa fase è piuttosto critica, in quanto influenza il risultato finale del sequenziamento stesso: nel caso si vada a

sequenziare un prodotto a concentrazione troppo elevata, si crea una condizione di "*over-clustering*" che porta all'arresto della *run*; una concentrazione troppo bassa di contro, determina una condizione di "*under clustering*", che ugualmente porta ad una perdita di molte delle informazioni recuperabili.
A seguito dell'unione dei vari catturati, il *pool* è stato nuovamente controllato mediante Qubit™ 4 Fluorometer (dsDNA High Sensitivity Kit), caricato in doppio, per assicurarsi della correttezza della concentrazione finale. Per il sequenziamento, insieme al *pool* è stato caricato anche il PhiX Control v3 (Illumina, San Diego, CA, USA), in quantità pari al 5% del volume finale. Si tratta di una libreria usata come controllo del sequenziamento, costruita a partire dal piccolo e ben noto genoma del batteriofago Phi-X174 necessaria per controllare il corretto funzionamento della corsa.

4.2.6 Analisi dei dati

A seguito della generazione dei dati di sequenza, gli *output* sono stati gestiti mediante la produzione di una *pipeline* bioinformatica costruita attraverso l'uso di 3 *tools* principali (Eager[179], GATK[180], VCFtools[181]) al fine di produrre statistiche utili per la valutazione della resa dei due diversi set di sonde e degli assetti sperimentali adottati, sia per i campioni moderni che per quelli antichi. Inoltre, i campioni sono stati posizionati all'interno dell'albero filogenetico del Y-chr mediante assegnazione degli Hg caratterizzanti le sequenze di ciascun individuo.

4.2.6.1 Processamento dei dati di sequenza, mappaggio e chiamata delle varianti del Y-chr

Per ciascuna lettura, la chiamata delle basi viene effettuata dallo strumento nel corso della *run* mediante il software Real Time Analysis (RTA). I *file* prodotti in formato .bcl sono stati convertiti nel formato FastQ compresso e "demultiplexati" utilizzando il *software* Illumina CASAVA-1.8.2. Durante la fase di *demultiplexing*, ogni *read* prodotta viene assegnata al campione che l'ha generata, grazie all'associazione indice-campione. Nel caso di sequenziamenti *paired-end*, per ciascun campione si producono due *file* in formato FastQ: uno per le *reads* in *forward* (R1) ed uno per quelle in *reverse* (R2).
I *file* in FastQ sono caratterizzati da una precisa struttura a quattro righe: nella prima, il simbolo "@" precede il codice identificativo della sequenza; a seguire, nella riga sottostante è indicata la vera e propria sequenza nucleotidica, e nella terza riga il simbolo "+" è utilizzato come separatore tra la riga precedente e quella successiva[182], contenente una serie di simboli che funzionano da codice di qualità della lettura. Il codice utilizzato è quello ASCII, nel quale ogni carattere corrisponde alla probabilità (P) che l'assegnazione del nucleotide nella posizione descritta sia errata. La qualità, definita Phred Score, è descritta dalla formula:

$$Q = -10\log_{10}(P)$$

E quindi:

$$P=10^{-Q/10}$$

Pertanto, per un valore di qualità pari a 20, la probabilità d'errore è dell'1%, una qualità di 30 si traduce in una probabilità d'errore pari a 0.1%, e così via. Normalmente si sceglie una soglia di qualità minima di 30, in modo da avere una probabilità di errore della chiamata inferiore allo 0,1% e quindi delle letture con un alto grado di affidabilità.

Tutte le *reads* grezze prodotte dai sequenziamenti sono state inizialmente processate mediante una *pipeline* estremamente efficiente per l'analisi di dati di sequenza provenienti da aDNA, utilizzando il software EAGER[179], il quale comprende una serie di strumenti appositamente sviluppati per le fasi di elaborazione del dato grezzo, di analisi e di autenticazione dei risultati (Figura 10). Il medesimo *software* è stato utilizzato anche per il processamento dei dati provenienti da campioni moderni, dal momento che presenta opzioni valide anche per tali tipologie di dati di *input.*

Nella fase preliminare di processamento dei dati grezzi, la qualità di sequenziamento di ciascun campione è stata valutata mediante l'analisi delle rispettive *reads* con FastQC[183]. Questo modulo permette di ottenere una prima visione d'insieme dei dati grezzi di sequenziamento per la determinazione di importanti statistiche di base, come ad esempio il contenuto in GC e le lunghezze medie di lettura delle molecole. Le *reads* sono state successivamente processate, in prima analisi, con Clip&Merge allo scopo di rimuovere le sequenze note degli adattatori ed unire le corrispettive letture R1 e R2 in un'unica sequenza. Per i campioni moderni sono state utilizzate le impostazioni di *default* del *tool;* per quanto riguarda invece i campioni antichi, sono stati processati affinché nella fase di *merging* fossero conservate soltanto le sequenze che verosimilmente potessero essere considerate antiche; ciò è stato effettuato imponendo come condizione una sovrapposizione minima di 11 bp tra la R1 e la R2, in modo tale da scartare tutte le sequenze maggiori di 142 bp, ovvero quelle potenzialmente contaminanti dal momento che la loro lunghezza risulta superiore a quella tipica dei frammenti di aDNA. Inoltre, per tutte le tipologie di campioni, sono state scartate le *reads* di lunghezza inferiore alle 30 bp poiché in grado di mappare in modo aspecifico su più sequenze di riferimento o su più regioni dello stesso genoma. Per una valutazione iniziale della qualità dei dati, le *reads* filtrate sono state allineate e mappate sull'intero genoma umano (GRCh37/Hg19, GCF_000001405.13) incluso il mitocondrio, mediante l'algoritmo BWA[184], usando la funzione *aln* ed impostando, per l'aDNA, i parametri -l 1000, -n 0.01 e -o 2. Il parametro di filtraggio -n 0.01, indicativo del livello di *mismatch* accettato tra la sequenza nucleotidica di una *read* e la sequenza di riferimento, non risulta particolarmente stringente poiché deve tenere conto dei *patterns* di degradazione e dell'elevato contenuto in GC che caratterizzano l'aDNA. Sono state inoltre scartate tutte le *reads* con qualità di mappaggio inferiore a 30 (soglia definita dal parametro –q 30). In seguito, le letture mappanti sulla sequenza di riferimento sono state sottoposte ad una fase di rimozione dei duplicati di PCR mediante DeDup, un *software* che identifica le molecole clonali mediante le coordinate di inizio e di fine delle molecole stesse. Più in

dettaglio, tale strumento identifica tutte le *reads* che presentano le medesime coordinate iniziali, sia al 3' che al 5' della molecola, e ne conserva una sola. Dal momento che queste sequenze vengono verosimilmente prodotte da una sola molecola di partenza, che ha subìto processi di amplificazione in vari *step* di lavoro, mantenerne più di una nella fase di gestione bioinformatica del dato determinerebbe l'introduzione di importanti *bias* soprattutto in posizioni ambigue (p.es. posizioni con chiamata nucleotidica non concordante al 100%).

Per i campioni antichi inoltre, si è proceduto alla valutazione dei *patterns* di misincorporazione alle estremità 5' e 3' dei filamenti mappanti sul genoma umano. Sfruttando il *software* MapDamage2.0[185] sono stati allineati e comparati i profili di danno delle sequenze ottenute con quelli della sequenza di riferimento, per discriminare una molecola antica da una di probabile origine moderna e quindi contaminante, sulla base della percentuale di modificazioni C→T e G→A. A differenza di altri tratti diagnostici dell'aDNA, come la lunghezza delle molecole o la frammentazione in corrispondenza delle purine, l'accumulo di misincorporazioni tende ad aumentare in funzione del tempo[186,187,188]: studiando queste caratteristiche biochimiche è quindi possibile stimare il grado di incidenza delle contaminazioni moderne. La stima viene effettuata, attraverso un calcolo bayesiano, sulle *reads* che sono state precedentemente mappate sulla sequenza consenso, e quindi filtrate. Per tale analisi, sono stati usati i tre comandi richiesti come *input* dal *software*: *map*, *merge* e *plot*[189]. Il primo comando usa come *input* i *file* BAM (*Binary Alignment Map*) filtrati sulla consenso; il *merge* è necessario per permettere la compilazione di tabelle in cui vengono definite le percentuali di misincorporazioni degli ultimi 25 nucleotidi a monte ed a valle del filamento; l'ultimo comando permette infine la produzione di un *output* grafico a partire dall'informazione del *merge*.

Per ogni campione, pertanto, sono state registrate la percentuale di letture mappate sul genoma di riferimento, il numero di duplicati di PCR e le dimensioni medie dei frammenti di DNA letti.

Una volta completate le analisi preliminari sui dati di sequenziamento ottenuti, i *file* BAM filtrati sono stati usati per effettuare la chiamata delle varianti attraverso la funzione HaplotypeCaller del *toolkit* GATK 4.1[180], per la produzione di *file* VCF (*Variant Call Format* – Figura 10). Tale programma è sviluppato per la chiamata simultanea di SNPs e INDELs (inserzioni/delezioni) nelle linee germinali, mediante un assemblaggio *de-novo* degli aplotipi nella regione di analisi. In particolare, nel flusso di lavoro, viene eseguita dapprima una identificazione di ciascuna variante presente nei singoli campioni. Per l'ottimizzazione di tale processo sono stati utilizzati i parametri –ERC BP_RESOLUTION -ploidy 1 -mbq 30 -stand-call-conf 30 --output-mode EMIT_ALL_SITES -L Y --annotate-with-num-discovered-alleles. I dati di singolo campione così ottenuti sono stati gestiti mediante il programma GenotypeGVCF che ha consentito la genotipizzazione congiunta di più campioni in maniera efficiente. Il programma in particolare determina su quali regioni del genoma deve operare (regioni attive), in base alla presenza di varianti. Per ogni regione attiva, il programma costruisce un grafico simile a De Bruijn per

riassemblare la regione e identifica quali sono i possibili aplotipi presenti nei dati. Il programma quindi riallinea ogni aplotipo rispetto a quello della sequenza di riferimento utilizzando l'algoritmo Smith-Waterman al fine di identificare siti potenzialmente varianti. Per ogni regione attiva, inoltre, il programma esegue un allineamento a coppie di ciascuna lettura rispetto a ciascun aplotipo utilizzando l'algoritmo PairHMM. Ciò produce una matrice di probabilità di aplotipi sulla base dei dati letti. Infine, per ogni sito potenzialmente variante, viene applicata la regola di Bayes, per calcolare le probabilità di ciascun genotipo, per ciascun campione, al fine di assegnare quello rappresentato da un punteggio maggiore.

4.2.6.2 Statistiche di resa

Attraverso la manipolazione dei *files* genomici prodotti con GATK 4.1, è stata valutata più in dettaglio la copertura del Y-chr nelle regioni bersaglio delle sonde. In particolare attraverso i pacchetti BCFtools[190] e Vcftools[181] sono state generate statistiche sia di singolo campione che cumulative di tutti i campioni in analisi (Figura 10).
In primo luogo, sono state identificate la percentuale di copertura delle regioni di interesse. Inoltre, è stato valutato il rapporto tra le *reads* mappanti nelle regioni *target* e quelle localizzate in modo aspecifico lungo il genoma di riferimento, oltre al numero di posizioni del *subset* di SNPs selezionati coperte in ciascun campione, attraverso la comparazione tra i profili ottenuti ed il BED *file* fornito dalle aziende produttrici delle sonde, in cui sono espresse le posizioni target del disegno sperimentale.

4.2.6.3 Determinazione degli aplogruppi del Y-chr

I *file* BAM filtrati sulla sequenza consenso, prodotti mediante la pipeline EAGER, sono stati utilizzati per eseguire un'analisi preliminare sui profili mutazionali dei campioni attraverso il *software* Yleaf[191] (Figura 10), in grado di assegnare l'Hg del Y-chr attraverso la comparazione delle varianti di ciascun campione con 41560 posizioni filogeneticamente informative interne alle regioni MSY. Tali posizioni sono state recuperate a partire da tutti i *marker* presenti all'interno dell'albero del Y-chr prodotto dall'International Society of Genetic Genealogy (ISOGG), e filtrate per escludere mutazioni private, varianti in corso di studio, posizioni polialleliche e SNPs che hanno presentato chiamate alleliche anche in campioni femminili.
Le *reads* processate sono state filtrate sulla base della loro qualità; per la chiamata delle varianti sono stati imposti i parametri di copertura minima della posizione (pari a 2 *reads*) e concordanza di chiamata del polimorfismo superiore al 90%. I risultati ottenuti sono stati confrontati con quelli recuperati precedentemente attraverso analisi *shotgun* o di WGC[176] al fine di valutare la risoluzione delle sonde specificamente disegnate per analisi filogenetiche del Y-chr rispetto ad altri metodi comunemente adottati nelle analisi di aDNA.

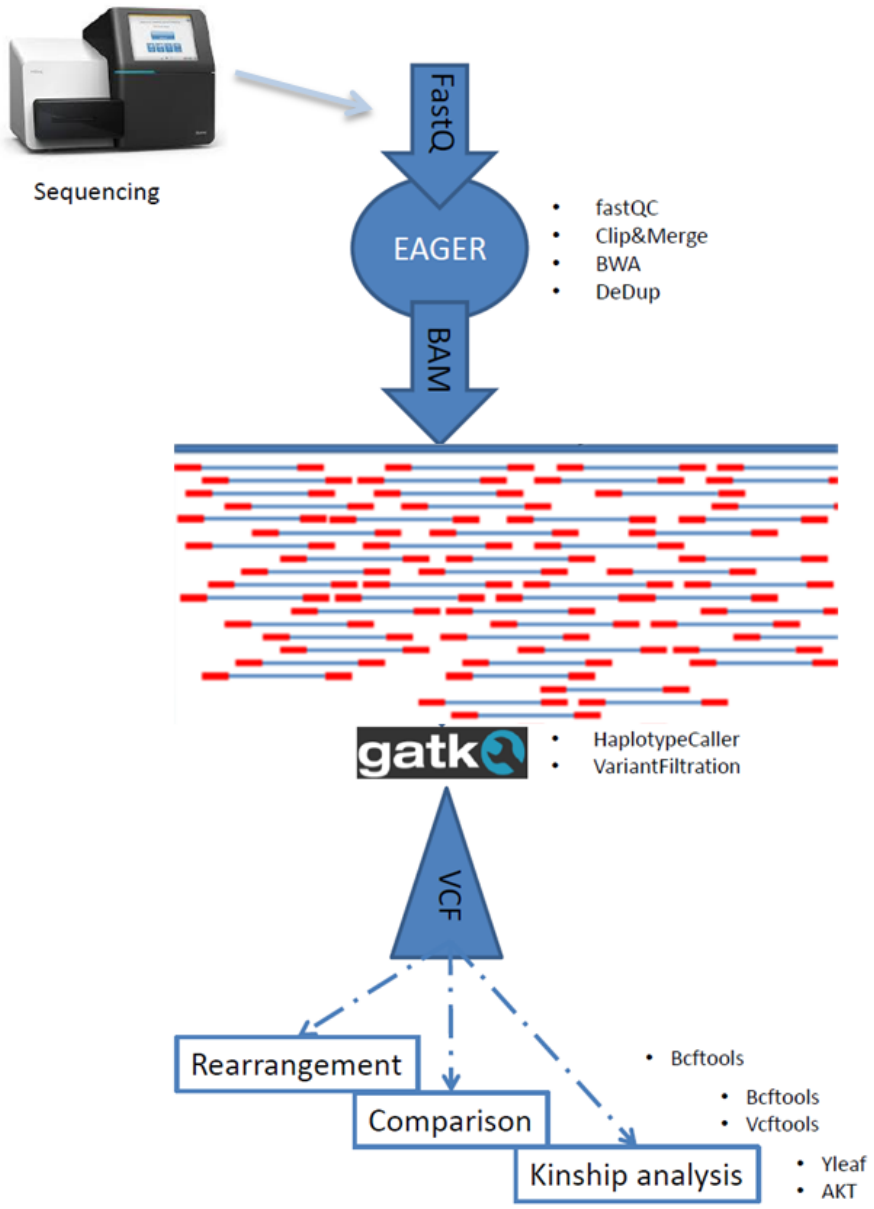

Figura 10: Workflow bioinformatico adottato nell'analisi dei dati di sequenziamento.

4.3 Risultati e discussione

Nel presente capitolo verranno esposti e discussi i risultati dei test sperimentali condotti sui set di campioni moderni ed antichi, per la valutazione delle sonde progettate e mirate alla cattura di una vasta quantità di SNPs di interesse filogenetico del Y-chr. Saranno riportati i risultati derivanti dal confronto delle diverse procedure di analisi, con particolare attenzione all'efficienza, specificità, e resa dei due set di sonde confrontati e alla funzionalità dei protocolli specificamente sviluppati per analisi su aDNA.

4.3.1 Quantificazioni pre-cattura

In seguito all'estrazione del DNA genomico dai tamponi buccali recuperati dai 4 campioni moderni (come definito nel *Paragrafo 4.2.3)*, è stata eseguita una prima valutazione della concentrazione di DNA in ciascun estratto, al fine di predisporre in modo estremamente preciso il materiale alla successiva fase di "tagmentazione" (produzione della libreria genomica e frammentazione enzimatica del DNA). In Tabella 2 sono espressi i valori di concentrazione e purezza del materiale estratto, valutati come rapporto tra le assorbanze a 260 e a 280 nm attraverso Nanodrop 2000. In estratti non contaminati da altre tipologie molecolari, quali ad esempio le proteine, il rapporto tra i due parametri si avvicina molto al valore assoluto di 1.8.

ID	ng/µL	260/280
M1	103.5	1.85
M2	290.4	1.79
M3	89.9	1.82
W	75.4	1.78

Tabella 2: Risultati di quantificazione degli estratti ottenuti mediante Nanodrop2000.

Come è possibile osservare in Tabella 2, tutti i campioni estratti presentano un valore ottimale di purezza e sono stati pertanto diluiti a 25 ng/µL per la produzione delle librerie genomiche. A seguito della frammentazione enzimatica del DNA, e all'aggiunta degli adattatori a tutte le molecole nucleiche presenti in soluzione, il materiale è stato ulteriormente quantificato con Agilent 2100 Bioanalyzer System (kit DNA 1000). In questa fase, sono stati ricontrollati anche i profili delle librerie dei campioni antichi di Szolad e Collegno, già preparate per progetti precedenti. I profili ottenuti sono mostrati in Figura 11.

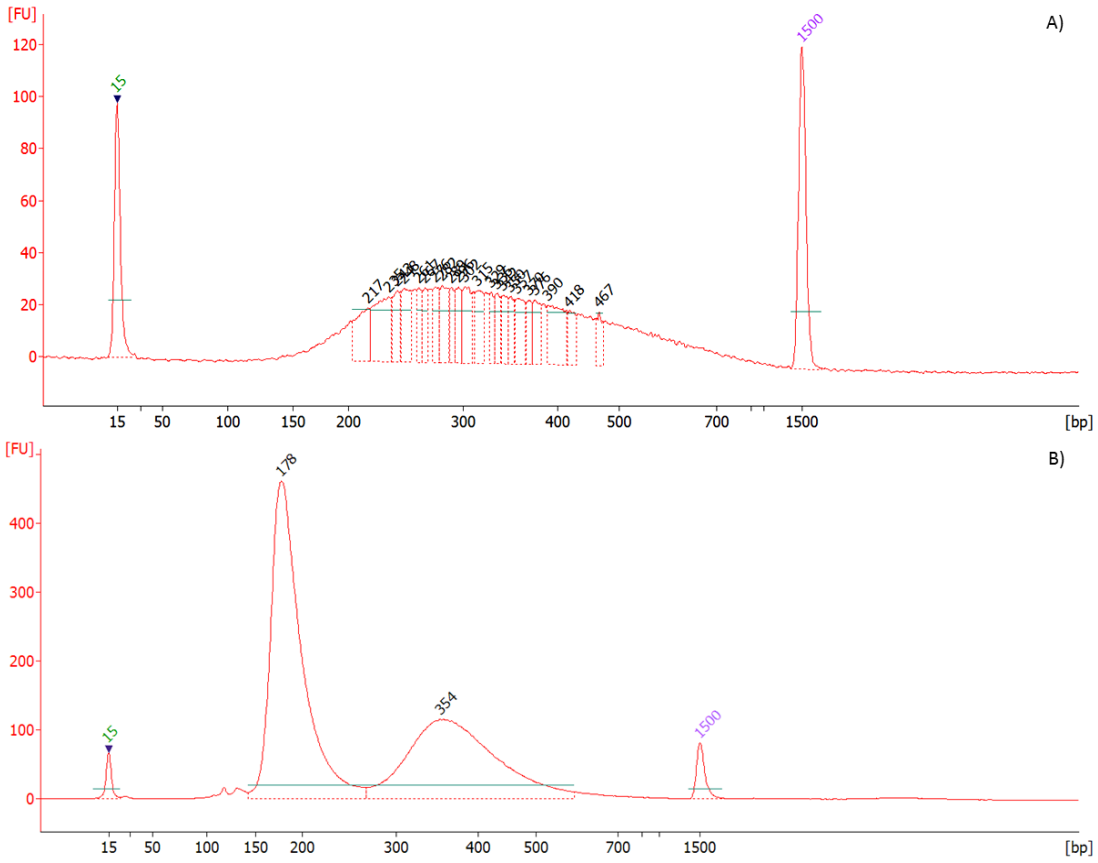

Figura 11: Risultati dell'elettroforesi quantitativa effettuata mediante Agilent Bioanalyzer 2100. A) Profilo rappresentativo dei campioni moderni "tagmentati"; B) Profilo tipico dei campioni antichi a seguito della produzione delle librerie genomiche.

Il grafico, tipico dello strumento utilizzato, viene prodotto in seguito ad una corsa elettroforetica automatizzata e mostra sulle ordinate la concentrazione del DNA espressa in unità di fluorescenza (FU), e sulle ascisse la lunghezza dei frammenti in termini di paia di basi. Al di sopra dei picchi sono precisati i valori di lunghezza maggiormente rappresentati all'interno del *pool* di molecole.

Il profilo quantitativo e qualitativo riportato in Figura 11 – A è stato ottenuto dal campione M1 ma è rappresentativo dell'andamento di tutti gli altri campioni moderni processati. Come atteso, questi presentano una varietà di frammenti molto ampia, con una concentrazione massima di molecole tra le ~250 e le ~315 bp, dimensioni compatibili con le specifiche fornite dal protocollo utilizzato per la produzione delle librerie.

Anche per i campioni antichi è stato riportato un unico profilo rappresentativo dei 4 campioni analizzati, in cui è possibile osservare una componente maggioritaria di frammenti di DNA ad una dimensione attorno alle 180 bp (lunghezza data dalla somma tra gli adattatori indicizzati, circa 130 bp, e l'inserto, mediamente lungo 50 bp). A causa della numerosità di cicli di reazione portati avanti durante il processo di aggiunta degli indici, gli enzimi nei passaggi finali, diminuiscono la loro resa e possono portare alla formazione di prodotti molecolari aspecifici. Per tale processo si possono formare *eteroduplex,* visualizzati come picchi che in questo caso si estendono dalle ~300 alle ~600 bp (Figura 11 – B).

4.3.2 Cattura del Y-chr con sonde SureSelect

La cattura eseguita con sonde prodotte da Agilent Technologies è consistita nell'ibridazione tra il DNA genomico precedentemente convertito in libreria e una serie di oligonucleotidi "esca" a RNA. Il disegno delle sonde è stato mirato alla cattura di 29905 SNPs (~63% del *subset* selezionato) del Y-chr attraverso 41196 sonde. L'azienda ha inoltre impostato una densità di sonde ottimale nella regione target compresa tra 1X e 2X. Per la valutazione della quantità di materiale di partenza da inserire in cattura, è stata integrata l'intera area sottesa dai grafici ottenuti dopo la produzione delle librerie, nell'intero *range* tra il "*lower marker"* e l' *"upper marker"*.
In Figura 12 è possibile osservare il profilo ottenuto al Bioanalyzer 2100 in seguito a cattura e indicizzazione delle librerie moderne ("protocollo A").

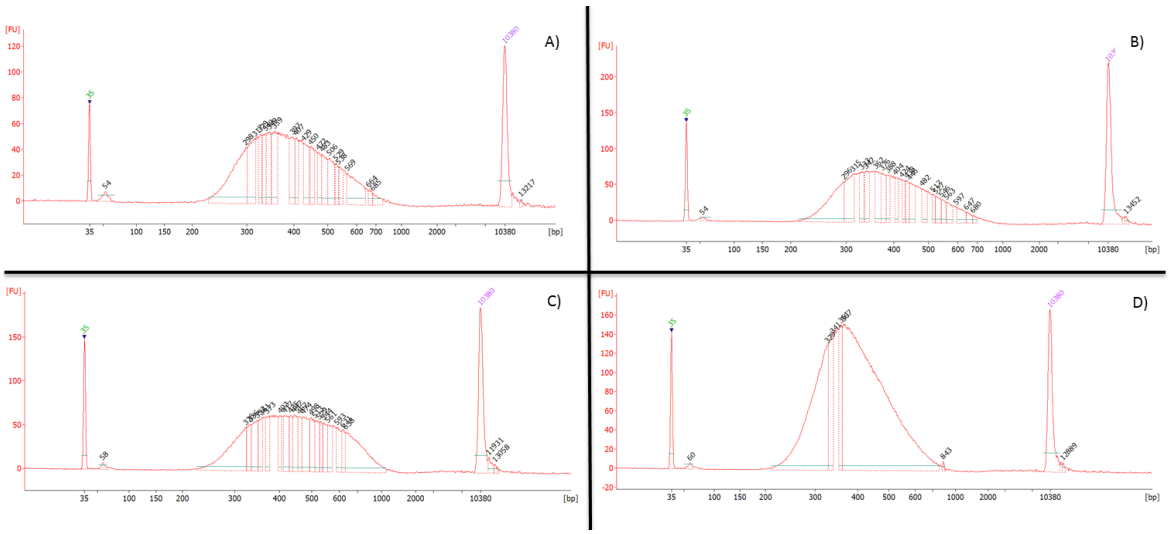

Figura 12: Risultati dell'elettroforesi quantitativa effettuata mediante Agilent Bioanalyzer 2100 sui campioni moderni a seguito della cattura del Y-chr e dell'aggiunta degli indici campioni-specifici. A) Profilo del campione M1; B) Profilo del campione M2; C) Profilo del campione M3; D) Profilo del campione W.

Tutti i campioni presentano un profilo di distribuzione dei frammenti di DNA in accordo con le linee guida fornite dal protocollo utilizzato, ed una concentrazione sufficiente per poter procedere alla produzione del *pool* di sequenziamento. In particolare, per i campioni di sesso maschile (Figura 12 – A-C), è possibile osservare una distribuzione molecolare compresa tra le ~250 e le ~700 bp, con un picco massimo attorno a 350 bp (con un inserto approssimativamente lungo 200 bp considerando le dimensioni di adattatori ed indici presenti ad entrambe le estremità di ciascun frammento completo). È interessante notare che anche il campione di sesso femminile usato come controllo (Figura 12 – D) presenta un profilo molto evidente, rappresentato in questo caso da una gaussiana più stretta, ma con una dimensione media dei frammenti simile a quella maschile, putativamente attribuibili a DNA genomico aspecifico catturato dalle sonde in mancanza del materiale *target*. La misura delle concentrazioni nell'intero *range* in cui è stata evidenziata la

presenza di molecole (150-2000 bp) ha fornito le seguenti concentrazioni: 5.776 nmol/L per M1, 10.633 nmol/L per M2, 4.109 nmol/L per M3 e 8.668 nmol/L per W. I campioni sono stati pertanto riuniti in un *pool* iniziale a 4nM, a cui è seguita, dopo ulteriore controllo quantitativo, la produzione del *pool* finale a 2nM, sottoposto a sequenziamento Illumina.

Come descritto nel *Paragrafo 4.2.4.1*, per i campioni antichi sono stati confrontati due diversi protocolli di cattura del Y-chr: il protocollo fornito dall'azienda produttrice delle sonde ("protocollo A"), ed uno specifico per la cattura in campioni caratterizzati da DNA degradato, e rappresentato da *step* di ibridazione DNA-sonde più lunghi, oltre che da processi di arricchimento più spinti ("protocollo B"). In Figura 13 è possibile osservare i profili ottenuti alla fine delle due metodologie. Per un confronto diretto, è stato presentato il profilo del medesimo campione (Sz-7).

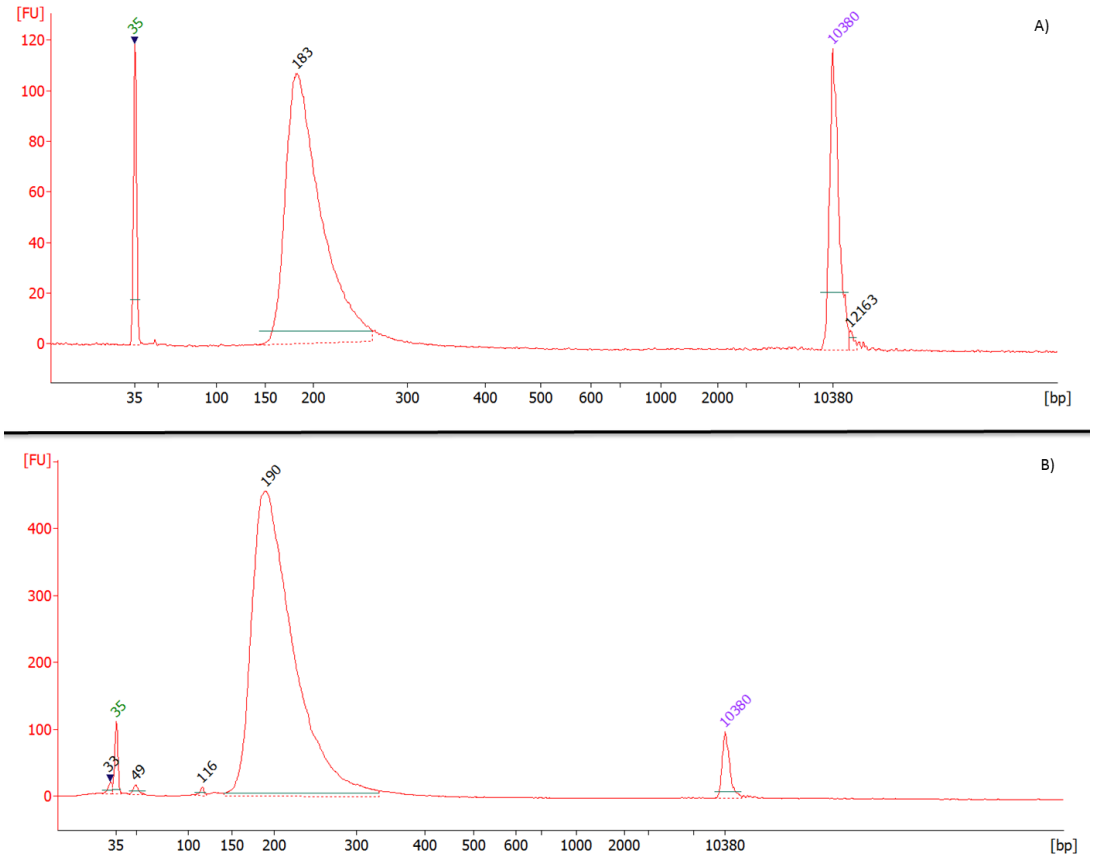

Figura 13: Risultati dell'elettroforesi quantitativa effettuata mediante Agilent Bioanalyzer 2100 su un campione antico, rappresentativo di tutti, a seguito della cattura del Y-chr con protocollo SureSelect (A) e con protocollo specificamente sviluppato per campioni antichi (B).

In generale è possibile osservare che i profili presentano lo stesso tipo di gaussiana, con un picco massimo attorno a 190 bp (dimensione standard del materiale degradato) ed un *range* di variabilità limitato rispetto a quello riscontrato nei campioni moderni a seguito della frammentazione *in vitro*. La differenza osservata nella dimensione più rappresentata del campione è da considerare effetto dell'errore strumentale. La principale difformità tra i profili è visibile in termini di FU; infatti, nel controllo quantitativo prodotto in seguito a doppia cattura, si osserva una concentrazione di

catturato quasi 4 volte maggiore di quanto osservato alla fine del "protocollo A". In Tabella 3 sono riassunte le concentrazioni ottenute per tutti i campioni antichi a seguito dei due protocolli.

ID	Protocollo A (nmol/L)	Protocollo B (nmol/L)
Sz-7	7.47	23.71
Sz-11	10.11	29.00
Sz-13	11.56	34.37
Coll-87	9.56	28.48

Tabella 3: Confronto delle concentrazioni finali post-cattura dei campioni antichi trattati con i protocolli A e B.

Come già osservato per i campioni moderni, anche qui, e con entrambi i protocolli di cattura, si osserva una perfetta sovrapposizione tra i valori ottenuti per i campioni di sesso maschile e quello femminile di controllo; anche in questo caso si può ipotizzare che le sonde abbiano partecipato a legami estremamente aspecifici in seguito alla mancanza di materiale *target*, veicolati dalla maggiore stabilità delle molecole nucleiche nella condizione a doppio filamento, sebbene con una bassa specificità nei legami ad idrogeno tra le due eliche.
I campioni sono stati riuniti in due *pool* a 4nM; è seguita, dopo ulteriore controllo quantitativo, la produzione dei *pool* finali a 2nM, entrambi sottoposti a sequenziamento Illumina.

4.3.3 Cattura del Y-chr con sonde MyBaits

La cattura eseguita con sonde prodotte da Arbor Biosciences è consistita nell'ibridazione tra il DNA genomico già convertito in libreria con il precedente protocollo, ed una serie di oligonucleotidi "esca" di RNA a singolo filamento. Il disegno sperimentale ha fornito un totale di 74809 sonde mirate alla cattura di 31630 SNPs (~66.5% del subset selezionato) del Y-chr. Il set finale è il risultato dell'applicazione dei filtri di qualità adottati dall'azienda, su un pannello iniziale di 142653 sonde volte alla cattura dei 47551 SNPs forniti alla ditta per il disegno. In particolare, sono state filtrate le sonde con sequenze tali da cadere in regioni genomiche costituite da ripetizioni intervallate e da bassa complessità; inoltre, per ridurre il numero complessivo di sonde, sono state scartate quelle rappresentanti informazioni ridondanti (p.es. per la cattura di SNPs molto vicini) e, tra le rimanenti, sono state selezionate solo quelle con almeno il 95% di identità con la sequenza fiancheggiante lo SNP di interesse, al fine di evitare una perdita di efficienza in fase di cattura. L'azienda ha infine impostato una densità di sonde ottimale nella regione *target* pari a 3X.
Per la valutazione della quantità di materiale di partenza da inserire in cattura, è stata integrata l'intera area sottesa dai grafici ottenuti dopo la produzione delle librerie, nell'intero *range* tra il "*lower marker*" e l'"*upper marker*".

In Figura 14 è possibile osservare il profilo ottenuto alla TapeStation 4150 System (kit High Sensitivity D1000 ScreenTape) in seguito a cattura e indicizzazione delle librerie moderne ("protocollo C").

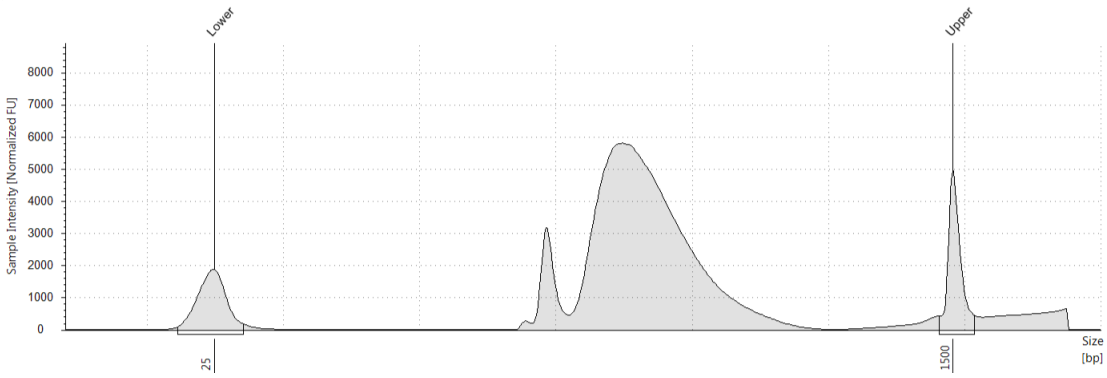

Figura 14: Risultati dell'elettroforesi quantitativa effettuata mediante TapeStation 4150 System sui campioni moderni a seguito della cattura del Y-chr e dell'aggiunta di indici campioni-specifici.

Il profilo mostrato, ottenuto dal campione M1, ma rappresentativo di tutti gli individui analizzati, evidenzia una buona sovrapposizione con quelli verificati in seguito a cattura con il "protocollo A". Risulta tuttavia interessante notare che con il "protocollo C" sia presente un picco, di dimensioni ridotte, a monte della gaussiana rappresentate la distribuzione di frammenti di DNA catturati, ed un'ulteriore *cluster* di frammenti a dimensioni superiori alle 1500 bp. È possibile ipotizzare che tali specie molecolari possano derivare da un processo di purificazione non perfettamente riuscito o da un arricchimento troppo spinto, che può aver condizionato la giusta calibrazione degli enzimi, i quali, nei passaggi finali, diminuiscono la loro resa e possono portare alla formazione di prodotti molecolari aspecifici.
Come già detto in precedenza, le concentrazioni in nmol/L ottenute per i campioni, sono state inserite in un *pool* in modo equimolare, per la corsa su piattaforma Illumina.

Anche in questo caso per i campioni test antichi, il "protocollo C", fornito dall'azienda produttrice delle sonde è stato confrontato con un protocollo modificato per massimizzare l'ottenimento di DNA *target* in campioni degradati ("Protocollo D"), riassunto nel *Paragrafo 4.2.4.2.*
Per un miglior confronto, in Figura 15 sono stati presi in considerazione i profili ottenuti dal campione Sz-7, già analizzati per il confronto tra il "Protocollo A" ed il "Protocollo B" e rappresentativi di tutti i campioni, di sesso maschile e femminile, analizzati.

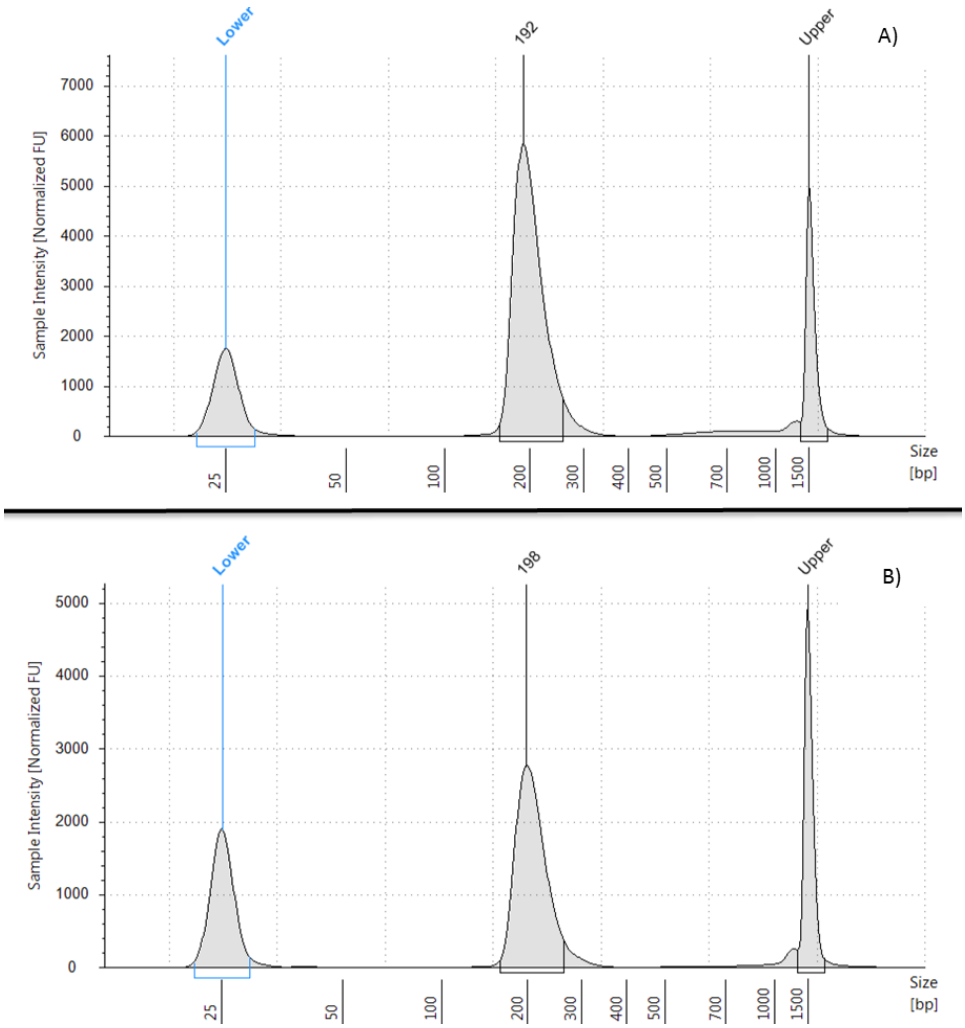

Figura 15: Risultati dell'elettroforesi quantitativa effettuata mediante TapeStation 4150 System su un campione antico, rappresentativo di tutti, a seguito della cattura del Y-chr con protocollo MyBaits (A) e con protocollo specificamente sviluppato per campioni antichi (B).

Come già discusso in precedenza, anche con questi due protocolli, i profili ottenuti dai campioni sono perfettamente sovrapponibili al termine delle due fasi di cattura, e le uniche differenze osservabili possono essere legate alla normale percentuale di errore strumentale ed a *bias* che inevitabilmente vengono introdotti durante la manipolazione dei singoli campioni. In Tabella 4 è riportato il confronto tra le concentrazioni ottenute per tutti i campioni antichi a seguito dei due protocolli.

ID	Protocollo C (nmol/L)	Protocollo D (nmol/L)
Sz-7	194.00	94.80
Sz-11	97.50	77.40
Sz-13	69.20	97.70
Coll-87	94.50	119.00

Tabella 4: Confronto delle concentrazioni finali post-cattura dei campioni antichi trattati con i protocolli A e B.

In questo caso, rispetto a quanto osservato a seguito dell'arricchimento con sonde SureSelect, si riscontrano valori di concentrazione più alti, ma con andamento variabile, soprattutto nel confronto tra cattura singola e doppia dello stesso campione.
Ovviamente non è possibile, con questi dati, effettuare alcuna inferenza relativamente alla bontà delle sonde e dei protocolli sviluppati, dal momento che si tratta di informazioni puramente descrittive. I campioni sono stati riuniti in due *pool* a 4nM; è seguita, dopo ulteriore controllo quantitativo, la produzione dei *pool* finali a 2nM, entrambi sottoposti a sequenziamento Illumina.

4.3.4 Analisi dei dati di sequenza

Le *reads* prodotte con i sequenziamenti effettuati, sono state processate in automatico dallo strumento in modo da unire tutte le letture di uno stesso campione sulla base degli indici assegnati (*demultiplexing*). Per i campioni sottoposti a sequenziamento *paired-end* sono stati prodotti due *file* di *output:* uno per le *reads* R1 (*forward)* ed uno per le R2 (*reverse).* A partire dal dato grezzo, sono state chiamate tutte le posizioni del Y-chr catturate per ciascun campione, come descritto nel capitolo "*Materiali e Metodi*".
La *pipeline* EAGER, è stata usata per il primo processamento delle *reads* grezze, al fine di mappare ciascuna *read* nella giusta regione genomica, come descritto nel *Paragrafo 4.2.6.1.* In prima analisi sono state riconosciute e tagliate le sequenze residue degli adattatori, è stato eseguito il *merging* delle *reads forward* e *reverse* – sulla base del loro codice identificativo – in modo da separarle da quelle che non rispettano i parametri definiti; infine, le *reads* filtrate sono state allineate e mappate sulla sequenza di riferimento umana hg19, incluso il mitocondrio, mediante BWA[184]. I principali risultati ottenuti per ciascun campione sono descritti in Tabella 5.

Metodo	ID	Parentela	*# Reads* Processate	*# Reads* post-Clip&Merge	*# Reads Merged*	% *Merged*	*# Reads* Mappanti	% Mappanti	# Duplicati rimossi	% DNA endogeno	*Cluster Factor*
Protocollo A	M1	Male – Father	4693173	4387148	4387148	100	3793490	86.46825	1166233	86.468	1.522
	M2	Male – Son	4322177	4043415	4043415	100	3489307	86.29604	1133991	86.296	1.571
	M3	Male - Nephew	3740238	3496532	3496532	100	3184162	91.06629	1052410	91.066	1.585
	W	Female	4428303	4122528	4122528	100	3603198	87.40263	936415	87.403	1.43
	Sz-7	Male - Son	8719556	4570554	3675867	80,42498	858824	23.36385	85738	18.79	1.172
	Sz-11	Male - Unrelated	11354662	5800247	5266004	90,78931	4861718	92.32272	294737	83.819	1.099
	Sz-13	Male - Father	7172046	3591757	3228490	89,88609	1603692	49.67313	83098	44.649	1.087
	Coll-87	Female	8159952	4075833	3655530	89,68792	1928420	52.7535	32253	47.314	1.027
Protocollo B	Sz-7	Male - Son	8943022	4848708	3740696	77,1483	3073003	82.15057	1213892	63.378	2.902
	Sz-11	Male - Unrelated	10644088	5450114	4933955	90,52939	4903292	99.37853	831338	89.967	1.377
	Sz-13	Male - Father	10538244	5361785	4821516	89,92371	4157996	86.23835	1034217	77.549	1.711
	Coll-87	Female	9312258	4687456	4314727	92,04837	3789501	87.82713	259512	80.843	1.169

Metodo	ID	Parentela	# *Reads* Processate	# *Reads* post-Clip&Merge	# *Reads Merged*	% *Merged*	# *Reads* Mappanti	% Mappanti	# Duplicati rimossi	% DNA endogeno	*Cluster Factor*
Protocollo C	M1	Male – Father	5631807	5406312	5406312	100	5009380	92.65799	1784596	93.278	1.124
	M2	Male – Son	5618830	5338788	5338788	100	4955687	92.8242	1524862	87.143	1.351
	M3	Male - Nephew	5236333	4953572	4953572	100	4748962	95.86945	1156486	94.879	1.008
	W	Female	6199624	5821861	5821861	100	5266584	90.46221	1547554	88.332	1.559
	Sz-7	Male - Son	8792436	5219739	2940040	56,32542	2018407	81.02186	912602	68.652	4.227
	Sz-11	Male - Unrelated	11114160	6131497	4457356	72,69605	4030189	71.32062	1119601	90.417	1.786
	Sz-13	Male - Father	12907864	6744736	5058116	74,99354	3080292	54.96341	765101	60.898	1.603
	Coll-87	Female	490222	250010	207211	82,88108	138096	66.64511	42323	66.645	2.337
Protocollo D	Sz-7	Male - Son	7417928	4481259	2528486	56,42356	2382075	94.20954	1946131	94.21	24.53
	Sz-11	Male - Unrelated	8356268	4645932	3312759	71,30451	3179014	95.96273	2164595	95.963	4.881
	Sz-13	Male - Father	7724958	4312422	3027836	70,21196	2780113	91.81848	1836949	91.818	5.504
	Coll-87	Female	6177832	3605052	2230087	61,86005	2012148	90.22733	806937	90.227	2.296

Tabella 5: Principali risultati ottenuti con la pipeline EAGER per tutti i campioni analizzati. Le statistiche sono suddivise sulla base del metodo utilizzato per la generazione del dato.

Tutti i sequenziamenti, se si escludono i campioni femminili utilizzati come controllo negativo, hanno generato un minimo di 3740238 *reads*. Il maggior numero di *reads* (12907864) è stato ottenuto dal sequenziamento del campione Sz-13 a seguito della cattura eseguita con il "Protocollo C". La percentuale di *reads* mappanti sulla sequenza di riferimento utilizzata (hg19, incluso il mitocondrio) è risultata compresa tra il 23.4% (Sz-7 – "Protocollo A") e 99.4% (Sz-11 – "Protocollo B). La percentuale di *reads* unite va dal 100%, ottenuto nei campioni moderni (in cui, essendo stato prodotto un sequenziamento in *single end* è presente un solo dato per ciascuna molecola di partenza), al 56.3% di Sz-7 – "Protocollo C". In generale, la quota di *reads* allineate alla sequenza di riferimento è apparsa molto alta con tutti i protocolli sviluppati; per i campioni moderni, è possibile osservare una maggiore percentuale di allineamento con il metodo portato avanti mediante l'utilizzo delle sonde prodotte da Arbor Biosciences, sebbene la differenza sia molto contenuta. Al contrario per i campioni antichi processati, il confronto tra i diversi protocolli, non mette in evidenza una netta distinzione nella quota di *reads* mappanti rispetto al totale; tuttavia, è interessante osservare che lo stesso campione, catturato con le medesime sonde, presenta una quota significativamente più alta di *reads* allineate sull'hg19 in seguito alla messa a punto del protocollo specificamente sviluppato per aDNA rispetto all'uso del protocollo definito dalle aziende produttrici delle sonde. Il campione che è risultato maggiormente favorito dalla doppia cattura, rispetto al procedimento standard, è Sz-7 quando trattato con sonde SureSelect (23.4% - 82.1%). Tale campione aveva mostrato, già in studi precedenti[176] basse percentuali di DNA endogeno; una strategia piuttosto comune per substrati che contengono quantità molto basse di DNA endogeno rispetto al DNA contaminante prevede proprio che la stessa libreria genomica venga arricchita per più di una volta, al fine di massimizzare l'ottenimento del DNA-*target*[40]. In tutti gli esperimenti condotti si osserva un alto grado di duplicati rimossi; tale valore è supportato anche da un elevato *cluster factor*, parametro che rappresenta la complessità delle librerie sulla base dell'unicità delle sequenze rilevabili al termine dei passaggi di filtraggio. I duplicati di PCR, tipicamente, si notano nei set di dati prodotti con NGS quando la profondità del sequenziamento è maggiore della complessità del campione iniziale; in questi casi pertanto, il numero di frammenti di DNA univoci è inferiore al numero di *cluster* sulla *flowcell* di sequenziamento. Ciò si verifica spesso quando la quantità di materiale biologico iniziale è molto bassa e durante l'amplificazione vengono utilizzati cicli di PCR in eccesso. Questo si è manifestato in particolare nel campione Sz-7 – "Protocollo D", ad ulteriore supporto dell'estremamente bassa preservazione del campione selezionato per l'ottenimento del dato. Infine, ad esclusione dei risultati prodotti con il "Protocollo A", sul set di individui antichi, tutte le metodologie permettono l'ottenimento di una percentuale molto alta di DNA endogeno.

4.3.5 Valutazione della resa

Sono stati valutati alcuni parametri statistici per determinare la specificità delle sonde progettate e confrontare eventuali differenze, soprattutto nel caso dei campioni antichi, tra i metodi di ibridazione selezionati. Le principali informazioni, quali la lunghezza media delle *reads* processate, la stima di copertura del Y-chr, ed il rapporto tra la copertura del cromosoma di riferimento e quella autosomica, sono espresse in Tabella 6, e la distribuzione di copertura delle *reads* mappanti, lungo il Y-chr è presentata in Figura 16. Va tuttavia tenuto di conto che il parametro di copertura media del Y-chr è puramente indicativo, dal momento che non si è proceduto alla cattura dell'intero cromosoma in analisi ma di solo una regione contenuta. Il *bias* di cui è affetto questo parametro pertanto, è stato normalizzato rendendo in termini assoluti il confronto con le altre regioni genomiche (In Tabella 6 – "rapporto Y-chr/autosomi").

Metodo	ID	Lunghezza media *read*	Copertura su Y-chr	rapporto Y-chr/autosomi	# posizioni disegno sperimentale coperte	% SNPs catturati
Protocollo A	M1	139.73	1.1303	24.92	27271	91.19
	M2	139.94	0.8438	20.14	27078	90.55
	M3	140.38	0.7969	20.68	27165	90.84
	W	140.08	0.0016	0.04	-	-
	Sz-7	58.46	0.0099	1.29	-	-
	Sz-11	61.36	0.1478	3.92	-	-
	Sz-13	64.17	0.0485	4.52	-	-
	Coll-87	61.36	0.0003	0.01	-	-
Protocollo B	Sz-7	65.93	0.0278	2.78	1162	3.89
	Sz-11	66	0.2739	10.74	18382	61.47
	Sz-13	65.85	0.1448	9.24	14230	47.58
	Coll-87	66.18	0.0005	0.027	-	-
Protocollo C	M1	139.55	1.1557	25.35	30051	95.01
	M2	140.12	0.9758	19.94	29957	94.71
	M3	140.86	0.6391	21.01	28654	90.59
	W	139.94	0.001	0.03	-	-
	Sz-7	57.95	0.011	2.3		
	Sz-11	59.79	0.14	5.2		
	Sz-13	55.16	0.072	3.27		
	Coll-87	50.44	0.0001	0.16	-	
Protocollo D	Sz-7	59.99	0.0113	7.8	1270	4.02
	Sz-11	61.94	0.1773	20.74	20295	64.16
	Sz-13	59.91	0.0765	10.93	13637	43.11
	Coll-87	61.1	0.0001	0.0077	-	-

Tabella 6: Principali dati comparativi per la valutazione della specificità delle sonde.

Per quanto riguarda i campioni moderni, la copertura media (ovvero il numero medio di volte che ciascuna base è stata chiamata attraverso le *reads* processate) del Y-chr (NC_000024.9) è compresa tra 0.64 X e 1.16 X, mentre per il campione femminile (W) è di 0.001 X. Sia il valore minimo che quello massimo di copertura sono stati ottenuti con il set di sonde MyBaits. Inoltre, il Y-chr risulta avere una copertura media almeno 20 volte maggiore rispetto alla media degli autosomi, mentre tale dato non è riscontrabile nel controllo negativo. Tale affermazione è vera per entrambi i set di sonde utilizzati, e permette di ipotizzare una buona specificità dei disegni sperimentali testati.
Dall'osservazione dei risultati prodotti per i campioni antichi, si può affermare che in ogni caso, la lunghezza media delle *reads* risulta in linea con le aspettative, trattandosi di frammenti di DNA estremamente danneggiati. La copertura media del Y-chr ottenuta per i campioni di sesso maschile è compresa tra 0.009 X (ottenuta per il campione Sz-7, "Protocollo A") e 0.27 X (Sz-11 – "protocollo B"), mentre per il campione femminile di controllo è in media 0.00025 X. Anche in questo caso, solo per i campioni di sesso maschile, il rapporto tra la copertura media del Y-chr e media autosomica è significativamente a favore del primo. I risultati inoltre indicano che sia per i campioni moderni che per quelli antichi, la distribuzione di copertura lungo il Y-chr non è uniforme ma presenta dei picchi esclusivamente in corrispondenza delle regioni in cui sono state disegnate le sonde (Figura 16). Questo dato rafforza le inferenze relative alla specificità dei disegni sperimentali.

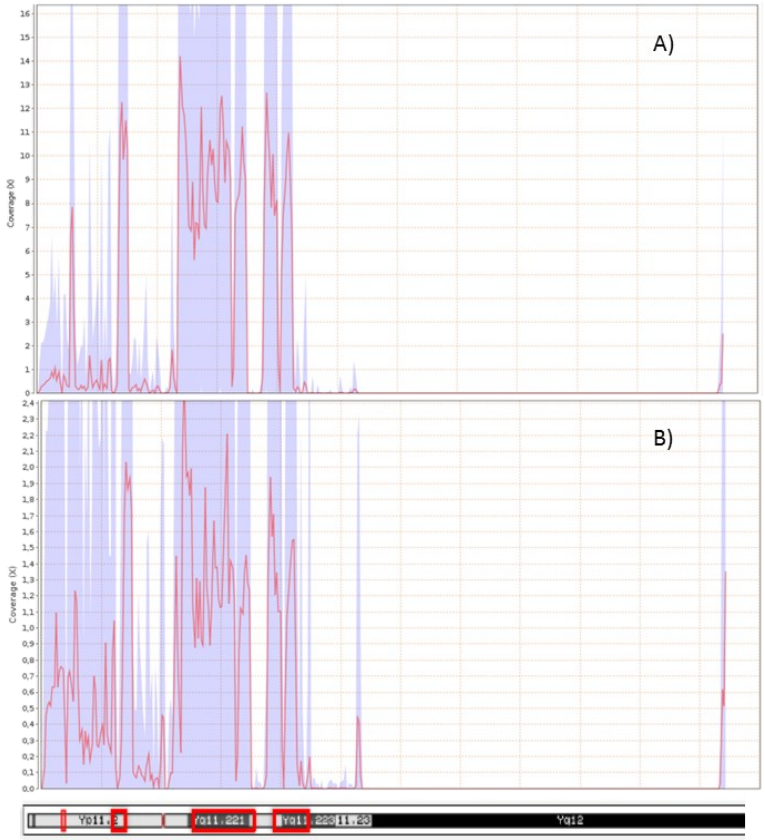

Figura 16: Copertura delle reads lungo il Y-chr. La figura in alto fornisce la distribuzione di copertura (linea rossa) e la deviazione (in blu) lungo la reference, in un campione moderno (A) e in uno antico (B). Il disegno sottostante rappresenta una schematizzazione del Y-chr. I riquadri rossi rappresentano le macroregioni in cui sono state disegnate le sonde.

Il confronto dei dati di copertura ottenuti con i diversi protocolli di ibridazione utilizzati per la cattura degli SNPs mediante le sonde prodotte da Agilent Technologies nei campioni antichi, mostra che il protocollo di arricchimento specificamente sviluppato ha una resa più di due volte maggiore rispetto al protocollo suggerito dall'azienda produttrice delle sonde, in virtù di una più lunga fase di ibridazione di queste con il DNA.
Analogamente, anche nel confronto tra i due protocolli di cattura utilizzati con le sonde disegnate da Arbor Bioscience si può valutare un'efficienza almeno 3 volte superiore quando si esegue un secondo *round* di arricchimento del materiale target. Sebbene la copertura media sul Y-chr nel protocollo di ibridazione utilizzato con le sonde MyBaits sia leggermente più bassa rispetto a quella riportata nel caso del "Protocollo B", il rapporto tra la copertura del Y-chr e quella autosomica è significativamente migliore. Tale differenza nella copertura media dipende da una diversa progettazione delle sonde, ciascuna delle quali risulta più corta di 40 bp nel caso MyBaits, come definito nel *Paragrafo 4.2.1*; il valore normalizzato visibile nella quinta colonna della Tabella 6 tuttavia, risulta indicativo di una migliore specificità delle seconde sonde disegnate.

Dal momento che le sonde sono state progettate con l'obiettivo di recuperare informazioni di interesse filogenetico relativamente alle discendenze maschili partendo da materiale biologico estremamente degradato, è stata soffermata l'attenzione sui dati prodotti per i campioni antichi.
In particolare sono stati confrontati i dati di sequenziamento ottenuti dai due protocolli di doppia cattura utilizzati con i due differenti set di sonde, i quali si sono dimostrati significativamente più performanti dei protocolli con un singolo *round* di cattura. Il confronto, effettuato sulla base delle principali statistiche di mappaggio e qualità del dato bioinformatico, è riassunto in Figura 17.
Sebbene in termini di qualità del sequenziamento i dati prodotti con il set di sonde MyBaits siano peggiori rispetto a quanto riscontrato con SureSelect (si nota infatti un aumento sensibile dei duplicati di PCR, che determina una netta riduzione del numero di *reads* mappanti sul genoma di riferimento), la percentuale di *reads* allineate specificamente sul Y-chr rispetto al totale del genoma è notevolmente aumentata con il "Protocollo D", così come la percentuale di DNA endogeno. Il peggioramento in termini di molecole clonali può essere spiegato a causa di una inferiore ottimizzazione dell'assetto sperimentale. Tale dato permette di presupporre margine per un ulteriore miglioramento nella resa delle sonde MyBaits attraverso un aggiustamento del protocollo sperimentale, in particolare durante le fasi di arricchimento post-cattura.

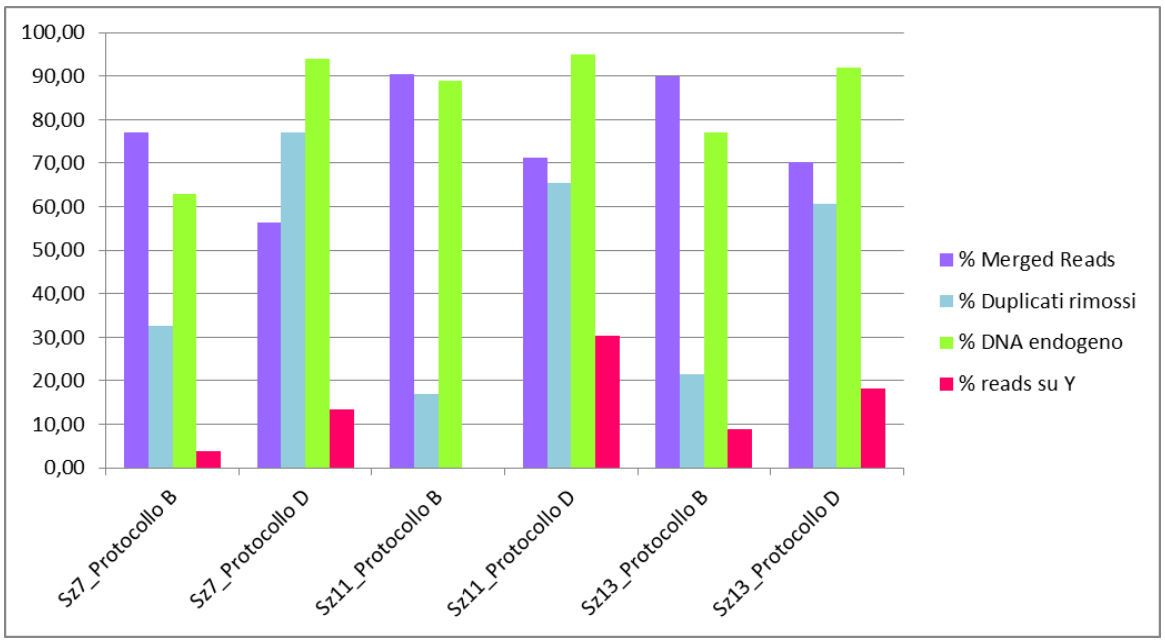

Figura 17: Principali dati di confronto sull'efficienza e specificità dei due set di sonde e protocolli di target enrichment testati.

Sono state successivamente condotte analisi più specifiche, attraverso l'utilizzo del *toolkit* GATK 4.1[180], al fine di identificare nel dettaglio le posizioni totali recuperate nei campioni moderni ed antichi. In particolare per questi ultimi, considerata la significativa miglior resa dei protocolli di doppia cattura con entrambi i set di sonde, sono state valutate le statistiche ottenute per i soli protocolli B e D. Inoltre, dal momento che il campione femminile usato come controllo negativo, come atteso, presenta dati di copertura sul Y-chr pressoché nulli, non è stato utilizzato per le successive analisi comparative. L'analisi sui campioni moderni, effettuata tramite confronto tra le posizioni mappanti sul Y-chr e quelle che si allineano in qualsiasi altra area genomica di riferimento, ha messo in evidenza che mediamente il 60% delle *reads* è localizzato al di fuori del cromosoma di interesse. Il risultato può essere spiegato attraverso una duplice motivazione: (i) le sonde non sono così specifiche come evidenziato nelle fasi preliminari di analisi, a causa di fenomeni di varia natura comprendenti problemi di complementarietà delle regioni a monte e/o a valle dello SNPs; (ii) nelle fasi sperimentali di ibridazione tra il DNA *target* e le sonde si sono verificati problemi di performance nel protocollo, o si sono create condizioni di squilibrio che hanno portato ad un esubero di sonde rispetto al materiale da catturare, e pertanto sono rapidamente subentrati fenomeni di ibridazione aspecifica con il DNA genomico. Per dare un senso a queste inferenze, attraverso la manipolazione dei *files* genomici prodotti, è stata valutata più in dettaglio la copertura del Y-chr nelle regioni bersaglio delle sonde. Mediante un'indagine diretta sui *files* VCF generati dai campioni moderni, è stato possibile asserire che la quasi totalità delle posizioni chiamate ricadono all'interno delle macroregioni dell'MSY selezionate nel disegno sperimentale delle sonde, per una quota media del 96% delle posizioni su cui è stato svolto il disegno sperimentale. In seguito a questa analisi, e con le osservazioni preliminari descritte all'inizio del

paragrafo, non è pertanto da escludere una buona specificità e performance delle sonde. Infatti, la quasi totalità delle posizioni disegnate sono state catturate e la copertura media del Y-chr, con entrambi i set di sonde risulta almeno 20 volte maggiore rispetto alla copertura media degli autosomi, i quali, essendo numericamente predominanti, potrebbero aver dato luogo ad appaiamenti aspecifici in seguito alla saturazione dei legami disponibili tra sonde e materiale genetico di interesse.

Inoltre, è stato identificato, sia per i campioni moderni che per quelli antichi, il numero di posizioni del *subset* di SNPs selezionati coperte attraverso la comparazione tra i profili ottenuti ed il BED *file* fornito dalle aziende produttrici delle sonde in cui sono espresse le posizioni *target* del disegno sperimentale (il risultato è espresso nella colonna "# posizioni disegno sperimentale coperte" – Tabella 6). Nessuno dei metodi ha raggiunto una copertura del 100%, probabilmente come risultato di una combinazione di letture mappate in modo ambiguo, tuttavia da questa analisi risulta una quota di SNPs chiamati compresa tra il 90.5% e il 91.2 % nei campioni moderni, e tra il 3.8% ed il 64.2% nei campioni antichi. Non si evidenziano nette differenze nel recupero di materiale confrontando i due set di sonde, a supporto della bontà di entrambi i disegni sperimentali prodotti.

Infine, per una valutazione complessiva dell'efficienza con cui sono state recuperate le regioni *target* sul Y-chr, si è proceduto dapprima alla determinazione della quota di posizioni del disegno sperimentale effettivamente recuperate. In questo caso si è fatto riferimento non solo alle posizioni oggetto di indagine filogenetica, bensì all'intera regione fiancheggiante lo SNPs, e recuperata attraverso il processo di cattura. Successivamente, è stata valutata e confrontata tra i vari disegni sperimentali, la quota di posizioni *on-target* chiamate rispetto alle totali (su cui pertanto risultavano presenti almeno due *reads* in fase di allineamento). I grafici a torta mostrati in Figura 18 rappresentano i risultati ottenuti attraverso questo confronto. Per quanto riguarda i dati prodotti sui campioni moderni, è possibile affermare che i due assetti sperimentali abbiano una resa pressoché confrontabile. Il grafico a sinistra nelle due immagini superiori indica infatti che per questi campioni sono state chiamate la quasi totalità delle regioni facenti parte dei disegni sperimentali, permettendo di avere a disposizione una quota compresa tra il 96.73% e il 98.18% del dato ricercato. Il "Protocollo C" in questo è risultato maggiormente performante, tuttavia, è risultato meno specifico nel recupero del materiale *target*, seppur la percentuale di chiamate al di fuori della regione di interesse risultino estremamente basse con entrambi i set di sonde.

Per i campioni antichi invece, le sonde MyBaits si sono dimostrate estremamente più specifiche, dal momento che è stato possibile recuperare oltre il doppio delle posizioni indagate rispetto a quanto ottenuto con sonde SureSelect, seppur a fronte di un disegno sperimentale condotto con sonde più corte, e di conseguenza con un numero complessivo di posizioni totali minore.

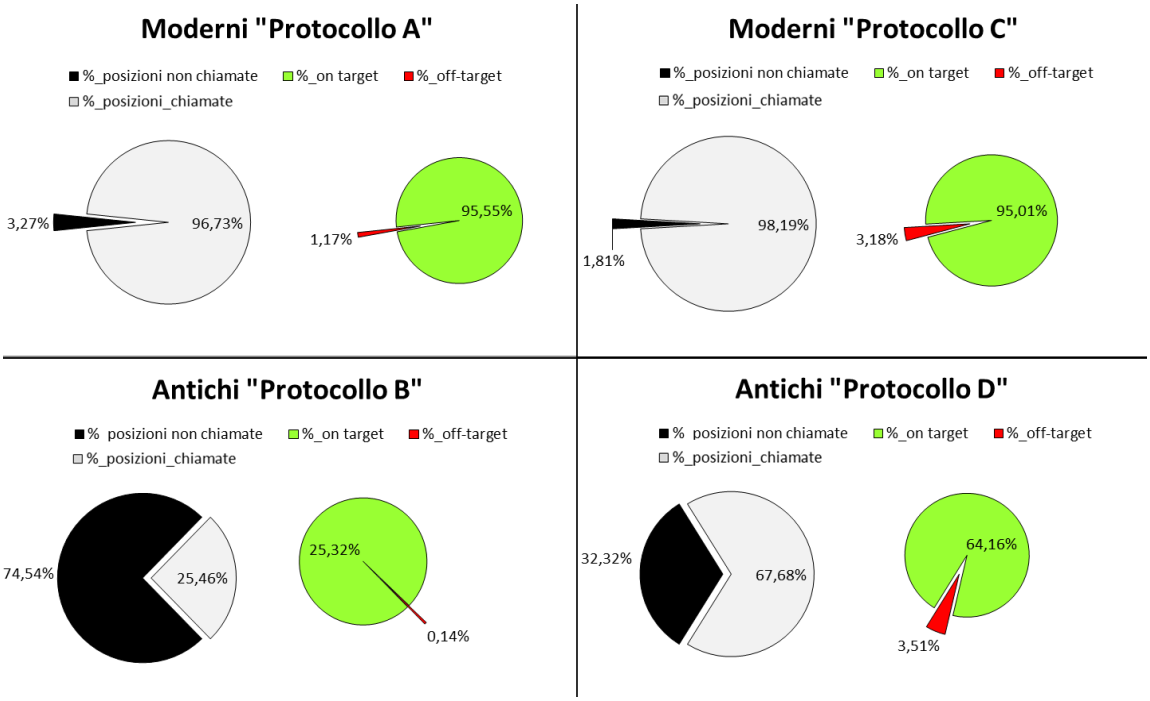

Figura 18: Statistiche medie di copertura (per base). Sono presi in confronto i dati medi ottenuti sui campioni moderni con i protocolli sviluppati dalle aziende produttrici delle sonde (Protocollo A e C) e quelli ottenuti dall'analisi dei dati di doppia cattura effettuati sui campioni antichi (Protocollo B e D).

4.3.6 Determinazione degli aplogruppi del Y-chr

È stata infine eseguita un'analisi di valutazione degli Hg del Y-chr attraverso il *software* Yleaf, al fine di confrontare la risoluzione dei dati recuperabili attraverso un set di sonde specificamente mirato all'ottenimento di informazioni filogenetiche, rispetto a quanto acquisibile con i metodi comunemente utilizzati, quali il sequenziamento profondo e l'arricchimento dell'intero genoma. Va tuttavia precisato che il *software* utilizzato è sviluppato per l'indagine di 41560 specifiche posizioni del Y-chr che, sebbene di interesse filogenetico, non risultano totalmente sovrapponibili al disegno sperimentale condotto, generando una possibile perdita di informazione nella ricostruzione filogenetica, soprattutto in rami più profondi. Per tale confronto sono stati presi in esame i soli campioni antichi, per i quali si avevano a disposizione le informazioni di appartenenza alle linee di discendenza maschili, ottenute in lavori precedenti[176]. I risultati ottenuti per i campioni antichi rispetto a quanto è risultato dall'analisi di Amorim e colleghi[176], sono mostrati in Tabella 7.

Metodo	ID	#markers (/41560)	% *reads* su chr-Y	Hg	Assegnazione precedente[176] (Metodo precedente[176])
Protocollo B	Sz-7	2130	3.93	I2a2a1b2a2a	I2a2a1b2 (cattura)
	Sz-11	12949	10.74	R1b1a1a2a1a1c2b2b1a1a	R1b1a1a2a1a1c2b2b1a (*shotgun*)
	Sz-13	10708	8.94	I2a2a	I2a2a (cattura)
Protocollo D	Sz-7	700	13.5	I	I2a2a1b2 (cattura)
	Sz-11	10755	30.29	R1b1a1a2a1a1c2b2b1a1a1	R1b1a1a2a1a1c2b2b1a (*shotgun*)
	Sz-13	7056	18.33	I2a2a	I2a2a (cattura)

Tabella 7: Principali risultati ottenuti con il software Yleaf per il set di campioni antichi selezionati. Per cattura si intende l'arricchimento del materiale biologico sui 1240K SNPs genomici. In rosso sono evidenziate le informazioni in più ottenute mediante arricchimento del Y-chr.

Il confronto è stato effettuato solo sui risultati ottenuti con i protocolli più performanti, per entrambi i set di sonde disegnati, al fine di valutare l'assetto che risulta più appropriato da utilizzare in contesti di DNA degradato. In generale, attraverso l'analisi di specifiche posizioni filogeneticamente informative, seppur disponendo di un set di sonde non specificamente disegnate per il recupero delle posizioni target indagate dal *software*, è comunque possibile assegnare sub-Hg estremamente dettagliati utili per inferenze popolazionistiche, anche per quegli individui il cui materiale genetico risulta conservato in modo peggiore. Nel caso del "Protocollo B" è possibile affermare che Sz-7 e Sz-13, parenti di primo grado, risultano appartenenti al medesimo sub-Hg (I2a2a), mentre Sz-11 presenta una variabilità genetica tipica del sub-Hg R1b1a. Nel confronto con i dati prodotti nel precedente lavoro, inoltre, è possibile osservare che con le posizioni a nostra disposizione, sia possibile fornire dati più approfonditi (in rosso in Tabella 7) in merito all'appartenenza ad uno specifico Hg del Y-chr per almeno due dei tre individui in esame.

Considerando i risultati ottenuti con il "protocollo D" è possibile osservare che sebbene la percentuale di *reads* mappanti sul Y-chr rispetto al totale del genoma sia aumentata molto, la chiamata dell'Hg viene fatta con un numero di marcatori (e quindi un dettaglio) inferiore. Questo dato, nonostante possa sembrare a prima vista contrastante, conferma la migliore resa e specificità delle sonde prodotte da Arbor Biosciences: in altre parole la quantità di *reads* mappate sul Y-chr rispetto al totale delle *reads* processate è significativamente più alta (nonostante le sonde siano più corte per via del disegno sperimentale), e la diminuzione dei marcatori di Yleaf fa presupporre che sia avvenuto un arricchimento estremamente specifico sulle regioni *on-target*.

4.4 Conclusioni e obiettivi futuri

Nelle ricerche condotte a partire da materiale biologico degradato è fondamentale tenere in considerazione che le analisi possono risultare limitate dalla disponibilità di informazioni recuperabili dalle molecole endogene e dai supporti tecnici a disposizione per lo studio; in particolare, sebbene l'evoluzione nel campo archeogenetico abbia permesso enormi conquiste tecniche, è importante considerare che i progetti di genomica di popolazioni basati sull'arricchimento di specifiche regioni, che prendono pertanto di mira SNPs predefiniti, potrebbero generare *bias* legati alla scelta delle posizioni *target* e all'impossibilità di individuare nuove varianti, soprattutto se la selezione viene effettuata su un numero molto limitato di posizioni, o su una piccola regione. Questo risulta di fondamentale importanza nella analisi che riguardano il cromosoma uniparentale maschile dal momento che la maggior parte degli studi effettuati prendono in considerazione solo pochi marcatori genetici. Il progetto di ricerca di questo dottorato si è pertanto focalizzato sulla messa a punto di una metodologia di selezione ed arricchimento di una vasta quantità di posizioni chiave del Y-chr, con l'obiettivo di dettagliare le variazioni genetiche presenti all'interno del cromosoma sessuale maschile allo scopo di ampliare le informazioni ad oggi disponibili per le ricostruzioni filogenetiche. È stata valutata l'efficienza di due set di sonde ad RNA, progettate diversamente da due aziende, in grado di bersagliare ~30000 SNPs di interesse filogenetico. Sono stati utilizzati 6 campioni, moderni ed antichi, e differenti protocolli di ibridazione. In particolare, i campioni moderni sono stati selezionati al fine di valutare l'efficienza dei due diversi set di sonde nel recupero del DNA *target.* I campioni antichi, già analizzati in precedenza in modo estremamente dettagliato, sono stati utilizzati per valutare la resa dei due assetti sperimentali nel recupero di materiale a partire da aDNA, per determinare il livello risolutivo di un disegno sperimentale volto al recupero del solo cromosoma maschile rispetto ai metodi già in uso e per ottimizzare un protocollo mirato ad aumentare il grado di efficienza nel maneggiamento di substrati degradati.

Complessivamente, entrambi i disegni sperimentali prodotti sono risultati validi per il recupero delle informazioni ricercate, soprattutto nei campioni moderni, dove è stato possibile visualizzare una copertura almeno 20 volte maggiore sul Y-chr rispetto alla media autosomica, ed isolare come minimo il 90% degli SNPs ricercati, nonostante la presenza di numerose chiamate *off-target*. Per entrambi gli approcci di arricchimento e per tutti i campioni, è stato infatti osservato che, sebbene una quota preponderante delle *reads* prodotte non risultino allineate all'interno delle regioni *target*, sono state lette la quasi totalità delle posizioni interne al disegno sperimentale. A seguito degli ottimi riscontri nelle coperture delle regioni di interesse, è possibile supporre che tale evento sia da inputare ad un fenomeno di saturazione dei legami disponibili nelle regioni di interesse sperimentale.

Inoltre, è stato messo in evidenza che con questi set di sonde è possibile ottenere almeno le stesse informazioni relativamente agli Hg del Y-chr rispetto a quanto emerso con studi basati su sequenziamenti profondi, o arricchimenti genomici, ma con il vantaggio di un abbattimento nei costi complessivi. È stato tuttavia notato che i processi di cattura tendono a determinare una notevole perdita di materiale, visualizzabile con un aumento della clonalità, soprattutto nel caso di librerie di partenza di bassa complessità. Questo punto ben conosciuto nell'ambito degli studi su aDNA, può essere modulato e controllato attraverso l'ottimizzazione di procedimenti sperimentali che non necessitino di fasi di arricchimento estremamente spinte, in modo da non aggiungere ulteriori *bias* nel caso di librerie genomiche già a bassa complessità.

Il confronto tra i protocolli di ibridazione nei campioni antichi mette in evidenza che indubbiamente l'utilizzo di fasi di ibridazione più lunghe e un secondo *round* di cattura eseguito sul materiale già selezionato, aumentano significativamente l'efficienza e la resa degli esperimenti di arricchimento selettivo.

Infine, le principali differenze che sono emerse nel confronto tra i due set di sonde sono legate ad una maggiore efficienza nei campioni antichi. In particolare, sebbene anche qui, come già riscontrato nei campioni moderni, il rapporto tra la copertura del Y-chr e quella degli autosomi sia nettamente a vantaggio del *target*, è possibile evidenziare un aumento significativo nel recupero delle informazioni di interesse in seguito all'utilizzo delle sonde MyBaits rispetto a quanto catturato con sonde SureSelect. Le sonde prodotte da Arbor Biosciences sono risultate vantaggiose quando utilizzate con aDNA anche nel recupero degli SNPs selezionati nel disegno sperimentale.

Dal momento che la differenza tra i protocolli discussa in precedenza suggerisce una specificità di legame tra regioni *target* e sonde lievemente migliore nel caso del protocollo di doppia cattura sviluppato con le sonde MyBaits, è stato scelto di selezionare tale metodologia per esperimenti futuri di cattura degli SNPs del Y-chr in campioni degradati. Restano tuttavia validi entrambi gli assetti sperimentali, su cui si renderebbero necessarie ulteriori analisi con un più ampio pannello di campioni al fine di evidenziare nette differenze di resa.

CAPITOLO 5: CASO STUDIO II

La variabilità genetica del cromosoma Y in Italia nell'età del Ferro

5.1 Introduzione

Grazie ai numerosi studi condotti sull'aDNA, oggi molto sappiamo dei processi di colonizzazione, migrazione e turnover popolazionistici che hanno interessato l'Europa dal Pleistocene ad oggi. In questi termini, la preistoria umana in Europa può essere divisa in cinque eventi fondamentali: i) la colonizzazione iniziale del continente durante il Paleolitico superiore, ii) la ri-colonizzazione di gran parte del continente a partire dai rifugi più a sud dello stesso, a seguito dell'ultimo picco glaciale (*Last Glacial Maximum,* LGM), iii) la ri-colonizzazione post-glaciazione del Mesolitico, iv) la dispersione di popolazioni provenienti dal Vicino Oriente, con lo sviluppo e diffusione del periodo Neolitico, e v) le migrazioni intra-continentali motivate da un'intensificazione delle reti commerciali iniziate nell'età dei metalli[192]. Antropologi e genetisti hanno cercato di fornire un'immagine delle possibili ondate migratorie seguite dall'uomo, a partire dalla sua comparsa in Africa, attorno ai 200000 anni fa, sino alla colonizzazione del globo, utilizzando una combinazione di dati antichi e moderni[193]. Sono molto dibattute le ipotesi relative alla dispersione umana al di fuori dell'Africa; i due fronti principali sono quelli che sostengono una singola dispersione, ovvero un'unica grande diffusione dell'uomo moderno attraverso l'Eurasia[194,195,196], e una dispersione multipla, in cui una più tardiva espansione ha contribuito alla diversità genetica delle attuali popolazioni non-africane[197,198,199,200].

Sono invece concordi le datazioni che fissano l'apparizione dell'uomo anatomicamente moderno in Europa attorno ai 45-40000 anni fa (periodo dimostrato grazie a misurazioni al radiocarbonio su fossili[201] e ad uno studio basato sull'analisi dell'mtDNA di 35 campioni paleolitici provenienti da diverse regioni dell'Europa[43] (Figura 19 - frecce verdi).

Durante l'LGM, l'Europa del Nord era prevalentemente coperta da ghiaccio, pertanto gli uomini erano relegati a vivere in ambienti estremamente poveri di risorse, che hanno portato alla necessità di spostamenti massivi, attorno ai 24000 anni fa, nei rifugi glaciali; la ri-colonizzazione delle aree abbandonate si è poi avuta da 5000 anni dopo tale evento (Figura 19 - frecce rosse) ed ha avuto profonde conseguenze nella diversificazione genetica degli Europei. È tuttavia complicato evidenziare distinzioni tra questi effetti e quelli sopraggiunti durante la transizione del Neolitico (Figura 19 - frecce blu), successiva di poche migliaia di anni, e considerata come il più importante processo demografico in Europa.

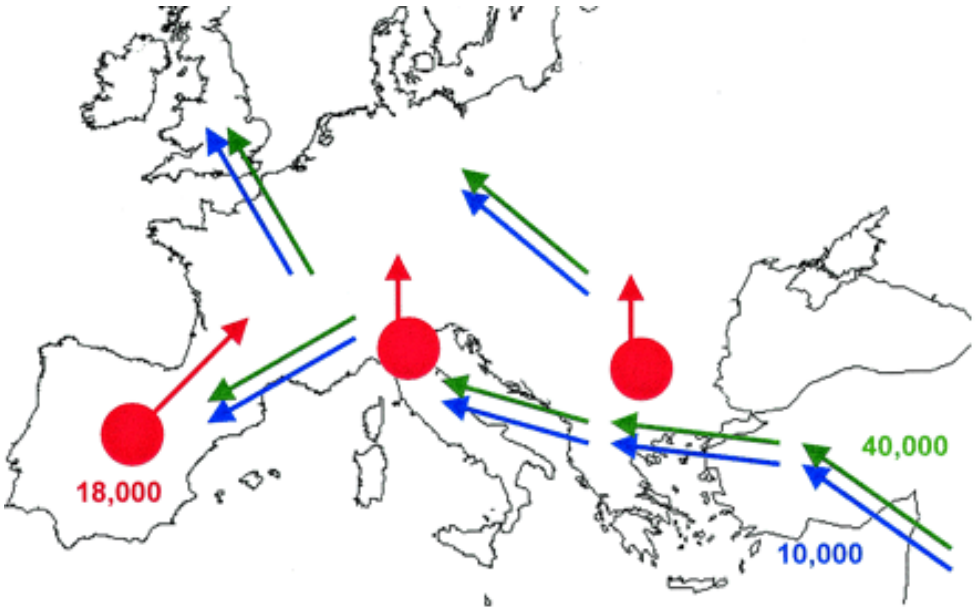

Figura 19: Principali tappe di colonizzazione del continente europeo202. Le frecce verdi rappresentano la prima colonizzazione (durante il Paleolitico), le frecce rosse la ri-espansione dopo l'LGM e le frecce blu la duplice ondata migratoria del Neolitico.

La diffusione della cultura neolitica è senza dubbio l'episodio più dibattuto e studiato delle tappe dell'evoluzione umana e i processi che l'hanno accompagnato si sono rivelati decisamente più complessi di quanto si pensava nel passato. Gli studi più recenti hanno dimostrato che il processo di neolitizzazione in Europa è avvenuto mediante due ondate migratorie dal Vicino Oriente, una per il centro ed il Nord Europa, e l'altra per il Sud Europa e l'area mediterranea[49,131,132,153,203,204]. Dopo la diffusione dell'agricoltura, la nascita di civiltà sempre più organizzate dal punto di vista geo-politico, certamente non placa gli istinti migratori dell'uomo: le espansioni, le dispersioni e le immigrazioni successive hanno lasciato traccia nel complesso mosaico genetico europeo attuale.

In epoca Storica, le più antiche tracce di civiltà, in Europa centrale, sono rappresentate dagli insediamenti dei Celti, mentre altri stanziamenti si trovavano nell'area baltica, grazie alle condizioni climatiche favorevoli. I Celti furono un insieme di popoli di origine indo-europea, tra cui si annoverano i Britanni, i Galli, i Pannoni ed i Celtiberi (che si stanziarono in un'area che si estendeva dalle isole britanniche all'Italia settentrionale, e dalla penisola iberica fino al bacino del Danubio), caratterizzati dalle medesime origini etnico-culturali, e da uno stesso fondo linguistico, ma sempre politicamente distinti. Il loro massimo apogeo si ebbe nel periodo compreso tra il IV ed il III secolo a.C.[146]. Esistevano, nello stesso periodo, dei gruppi isolati, che si spinsero ancora più a sud, toccando l'Italia centrale e l'Anatolia. In Grecia, la nascita della civiltà è attestata alla fine dell'età del Bronzo: l'organizzazione era quella di un insieme di città-stato, tra cui spiccavano Atene, Sparta, Corinto e Siracusa, molto differenti in termini socio-economici; l'ampliamento territoriale delle colonie Greche, ha portato alla nascita della cultura ellenica[205] (in Anatolia e in buona parte del Mediterraneo), la quale trovò un limite al suo sviluppo solo nell'espansione dei

Fenici. Tuttavia, il fatto che la Grecia non fosse una nazione unificata ebbe come conseguenza il susseguirsi di frequenti conflitti anche tra le varie città-stato[205].

Più in oriente, gli Sciti furono una popolazione nomade di origine iranica (indoeuropea) che dall'Ottocento a.C. si diffuse in Mesopotamia, Anatolia, Grecia, ed Italia, oltre che in tutta l'Europa centrale[144]. Gli Illiri furono un popolo di eccellenti artigiani del metallo e feroci guerrieri, che basarono i propri regni sulle lotte. Furono caratterizzati da una lingua, una cultura e tratti socio-antropologici indipendenti dagli altri popoli, e si stanziarono nella zona occidentale dei Balcani, abitando poi anche zone del Nord e Centro Europa, e fino alla Grecia[206]. Alcune tribù di Illiri migrarono e si stabilirono in Italia (in particolare, nell'attuale Puglia)[206]. La complicata situazione del continente euroasiatico nel primo millenio a.C, appena descritta, è rappresentata in Figura 20.

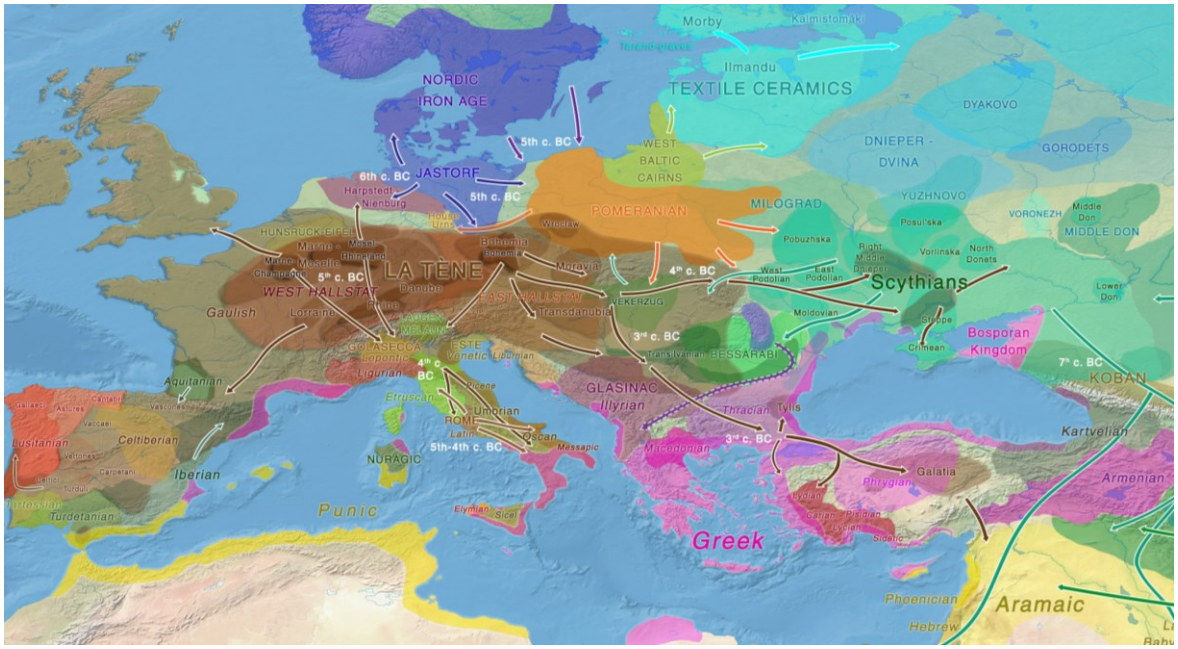

Figura 20: Mappa delle migrazioni in Eurasia nella prima età del Ferro.

5.1.1 La situazione italiana

L'Italia è sempre stata un luogo di facile approdo per le migrazioni umane a partire dall'Africa, dal Medio Oriente e da altre regioni europee. Inoltre, a causa della sua posizione geografica e delle caratteristiche geomorfologiche, fu coinvolta nel processo iniziale di popolamento dell'Europa, soprattutto nei luoghi costieri[207]. Durante il Paleolitico la navigazione nel Mediterraneo era rara, ma le isole rappresentavano ponti terrestri raggiungibili, sebbene successivamente sommersi dall'innalzamento del livello del mare a seguito dell'LGM. La situazione si stabilizzò durante il Mesolitico, quando il clima divenne più mite e permise successivi stanziamenti; molti siti risalenti a questo periodo sono stati trovati in tutta la penisola, comprese le zone interne e montuose, raggiunte attraverso due processi indipendenti: uno lungo la costa adriatica (Puglia e Italia Sud-

Orientale), risalente a circa 8.100 anni fa, e l'altro lungo il versante tirrenico, circa 7.900 anni fa, riguardante Liguria ed Italia Nord-Occidentale.

Ma è soprattutto durante l'età dei metalli che la regione italiana ha rappresentato un importante punto di contatto per le rotte migratorie di molti popoli, e per gli scambi commerciali e questo è attestato dall'enorme concomitanza di etnie eterogenee sul suolo peninsulare durante la metà del primo millennio a.C. I primi territori italiani coinvolti in questo processo sono stati la Sardegna, la Sicilia, l'area alpina e l'Italia Nord-Occidentale, per le loro posizioni strategiche e per la presenza di importanti risorse metalliche[208,209].

Le numerose dispersioni, colonizzazioni ed occupazioni intercorse in tale periodo hanno lasciato una traccia indelebile nel complesso mosaico genetico attuale, che rispecchia le peculiarità della storia del popolamento antico del territorio italiano e ne illumina alcuni passaggi fondamentali. Sebbene l'organizzazione delle popolazioni presenti nel territorio italiano sia estremamente frammentaria (Figura 21), dal punto di vista storico-archeologico è possibile distinguere tre macrocategorie: i) Popolazioni provenienti dal Nord Europa, di origine indoeuropea, arrivate in Italia in varie ondate migratorie; ii) popoli stranieri arrivati in Italia via mare (Greci, Fenici e Micenei) e via terra (come i Celti); iii) popolazioni di origine mediterranea (tra questi i Liguri, i Sicani ed i Sardi).

Tra le aree geografiche più studiate e ricche di documentazioni archeologiche vi è l'Etruria, una regione antica dell'Italia centrale, identificata con l'attuale Toscana, il Lazio settentrionale e parte dell'Umbria. Gli Etruschi, in seguito ad una prima fase espansiva (VIII-VI secolo a.C.), arrivarono a contendere a Greci e Cartaginesi il controllo delle rotte tirreniche e adriatiche e ad estendere il proprio dominio dalla Pianura Padana (VI secolo a.C.) all'alto Lazio. Sebbene non abbiano mai formato un'unità politica, le comunità etrusche condividevano una lingua e una religione non indoeuropee ed esercitavano un'influenza culturale e politica cruciale nell'area del Mediterraneo[210]. Secondo gli storici antichi, gli Etruschi non assomigliavano a nessun altro nella loro lingua, stile di vita o costumi e le loro relazioni evolutive con altre popolazioni dell'età del Bronzo non sono chiare[210]. La decadenza etrusca coincise con la crescita e l'autonomia di Roma.

Figura 21: Popoli che abitavano la penisola italiana nell'età preromana.

A contatto con questo popolo, sempre al centro del Paese, erano presenti molti popoli Italici, tra cui il numericamente più importante era quello degli Umbri. A nord-ovest, oltre i territori controllati dagli Etruschi, si erano stanziati invece i Liguri, popolo montanaro, per alcuni da considerarsi di origine iberica, per altri di provenienza celtica, ma sicuramente linguisticamente vicino all'idioma celtico[211]. Tra Veneto e Friuli Venezia Giulia erano stabiliti i Veneti, celebri nel mondo antico per i loro allevamenti di cavalli, ma famosi anche come commercianti; tessero una fitta rete di commerci incentrati sulle loro materie prime sia in direzione nord che in direzione sud: nel V secolo a.C. vennero a contatto, ad occidente, con i Galli[212]. La presenza di rilievi montuosi, quali gli Appennini, caratterizzò l'assetto delle popolazioni: un territorio estremamente frammentato poneva le basi per il possibile sviluppo di moltissime piccole comunità, tutte distinte ed indipendenti. Tra queste, i Latini erano un piccolo gruppo di pastori situato a sud del Tevere[213]; dalla fusione successiva di questi piccoli villaggi limitrofi, nacquero le fondamenta per la città e la civiltà che dominò l'antichità, e diffuse i principi della cultura occidentale: Roma. A Sud la situazione era ulteriormente frammentata: nell'attuale Calabria erano presenti insediamenti di presumibile origine Iberica, che convivevano con quelli di origine italica; la Puglia era abitata da gruppi di Illiri provenienti dai vicini Balcani. Nel resto del Sud c'erano popoli italici di cui i principali erano i Sanniti e gli Osci. A completare il quadro, la presenza greca, che si considera tutt'ora l'eredità più

forte depositata al Sud Italia, determinò una colonizzazione di massa che creò un vero e proprio terremoto culturale, tanto da portare a tutto il Sud, l'appellativo di "Magna Grecia"[214].

La popolazione italiana attuale presenta pertanto un grande grado di variabilità genomica interna, dovuta alla topografia del Paese che abita, oltre che agli eventi storici passati, che hanno portato ad importanti cambiamenti demografici[215]. Per la sua posizione cruciale al centro del bacino del Mediterraneo, l'Italia, ha sperimentato una complessa storia di colonizzazioni e migrazioni, ed il segno di questi eventi è ancora presente negli italiani moderni.

Una serie di studi, effettuati sul DNA di popolazioni moderne ed antiche hanno confermato che l'intera popolazione italiana attuale è caratterizzata dalle stesse componenti ancestrali europee, in proporzioni diverse, relative ai periodi del Mesolitico, Neolitico e dell'età del Bronzo[215].

Ovviamente, la genetica da sola non è in grado di districare il quadro estremamente complesso che si presenta, tuttavia alcuni movimenti umani hanno lasciato tracce, che possono essere seguite tramite l'analisi di marcatori genetici. Gli studi mitocondriali da questo punto di vista risultano ad oggi maggioritari, ed hanno mostrato che l'Italia antica non ha un patrimonio genetico nettamente diverso da quello che si attesta in Europa. Inoltre, in uno studio dettagliato di Brisighelli e collaboratori[115], l'analisi di 583 campioni italiani ha messo in evidenza la somiglianza dell'89% degli Hg del mtDNA, con quelli diffusi in tutta Europa, con la maggiore percentuale (40%) per la linea H. Lo studio ha anche sottolineato il fatto che la distribuzione degli Hg non è omogenea, ad esempio il tipo H passa del 59% del Nord, al 33% del Sud[115]. Tuttavia, le carte genetiche degli Italiani attuali mostrano chiari segni di remoti popolamenti, individuati soprattutto nelle zone isolate: l'influenza greca è particolarmente evidente al Sud, con un picco tra la Calabria meridionale e la Sicilia orientale[121]; nelle Marche si notano le tracce di un'antica civiltà italica del I millennio a.C. detta "osco-umbro-sabellica"; nelle vallate appenniniche meno accessibili tra Piemonte, Liguria ed Emilia, la geografia genetica mostra la presenza degli antichi Liguri; infine, in Toscana, in particolare nel Casentino, persistono ancora oggi i geni dei popoli in cui fiorì la civiltà etrusca[216,217].

Infine, la storia dei primi popolamenti della Sicilia è un tema che stimola l'interesse antropologico, tanto che nell'ultimo decennio sono aumentati i progetti di scavo nell'isola. Sebbene si consideri che le comuni rotte migratorie umane provengano dal nord, e avvengano attraverso lo stretto di Messina, alcuni autori hanno proposto l'ipotesi di un precoce popolamento dall'Africa, attraverso il canale di Sicilia (ipotesi però rifiutata). Le prime tracce umane documentate da reperti archeologici in Sicilia derivano da utensili in pietra, che non sembrano essere più antichi di 20000 anni; dal 16000 a.C. in poi, si è assistito ad un costante movimento popolazionistico tra l'isola, l'Africa e la penisola italiana, tuttavia ci sono poche tracce dell'attività umana nella regione prima della fine dell'ultima glaciazione, attorno agli 11000 anni prima di Cristo. Nonostante sia il cromosoma d'elezione per analisi popolazionistiche, il marcatore uniparentale femminile, mtDNA, fornisce una visione parziale degli eventi intercorsi ed i risultati da esso generati mostrerebbero maggiore

significatività se accompagnati da robuste analisi condotte anche sul marcatore genetico maschile. Alcuni lavori effettuati sulle popolazioni moderne, come quello portato avanti da Francalacci e colleghi[218], hanno mostrato che esiste una corrispondenza del 60% tra gli hg del Y-chr ritrovati in Sicilia e quelli del resto del Sud Italia e della Grecia[218]. Inoltre, uno studio basato sull'analisi della variazione della linea di discendenza del Y-chr, ha mostrato che esistono tracce di un flusso genetico in Sicilia, a seguito della colonizzazione da parte dei Greci, e che è presente un contributo nord africano, seppur minore[219]. Nel lavoro citato, il Y-MRCA è fatto risalire a circa 2380 anni fa, periodo che risulta comparabile con le tracce archeologiche della Grecia classica in Sicilia.
Dal punto di vista maschile, la maggior parte del *pool* genetico del Y-chr italiano può essere correlato a cinque Hg principali: R1b, J2, I, G ed E1b. Il primo di essi risulta più frequente nel Nord Italia, mentre E1b, G e J2 presentano frequenze più alte nel Sud, suggerendo una maggiore affinità con l'Europa occidentale per l'hg G e con l'Europa Sud-Orientale e centro-meridionale per J2. Infatti, nel contesto europeo, le variazioni del Y-chr osservate in Italia si attestano in linea con il cline Sud/Est-Nord/Ovest che ricalca la mescolanza genetica tra gli agricoltori del Vicino Oriente e i cacciatori-raccoglitori mesolitici precedentemente stanziati in Europa[220,221,222]. Inoltre, sebbene la maggior parte dei ricercatori interpretino la variabilità genetica del Y-chr lungo la penisola italiana, come un cline di longitudine, con una differenziazione lungo un asse Nord/Ovest-Sud/Est[122,123,124], due studi indipendenti hanno avanzato l'ipotesi di una più marcata strutturazione in tre aree principali di latitudine: Italia Nord-Occidentale, Italia Sud-Orientale e Sardegna[114,207].

Differentemente dagli studi effettuati sul DNA moderno, che possono solamente inferire eventi passati sulla base di pattern di distribuzione attuali, l'analisi dell'aDNA determina il vantaggio di fornire evidenze genetiche in modo diretto, permettendo di testare con un approccio più significativo ipotesi sull'affinità genetica di individui e popolazioni antiche. Il ruolo di principale crocevia della penisola italiana rende il recupero e l'analisi di campioni umani di notevole interesse per poter ricostruire l'ancora vacillante storia delle migrazioni e dei rapporti tra genti che convivevano nei medesimi territori. Gli studi ad oggi disponibili su campioni antichi tuttavia si sono concentrati principalmente su esemplari recuperati dall'Italia continentale, in particolare Etruschi e Longobardi, o dalla Sardegna che tuttavia, rivela una storia genetica anomala nel panorama europeo.
In questo lavoro sono stati caratterizzati geneticamente, dal punto di vista del marcatore uniparentale maschile, un buon numero di campioni italiani recuperati da diversi siti dell'Italia peninsulare e della Sicilia, e tutti attribuibili al periodo dell'età del Ferro. Gli obiettivi proposti dall'analisi di questi campioni sono due: i) lo scopo, dal punto di vista filogenetico, è quello di integrare ed ampliare i rami più a valle dell'albero filogenetico del Y-chr umano attraverso un sufficiente numero di campioni genotipizzati profondamente; ii) gli stessi individui sono stati contestualmente utilizzati per analisi di genetica di popolazione con l'intento di fornire una visuale dei movimenti migratori maschili intercorsi nella penisola italiana durante l'età del Ferro, periodo

successivo ad un ampio arco temporale ben documentato e ricco di movimenti in tutta l'area mediterranea, ma ad oggi ancora poco conosciuto dal punto di vista del marcatore uniparentale maschile.

5.2 Materiali e Metodi

È stato effettuato uno *screening* preliminare su 101 campioni provenienti da 8 siti archeologici del Centro e Sud Italia, ed i campioni geneticamente di sesso maschile sono stati successivamente arricchiti per più di 30000 SNPs filogeneticamente informativi del Y-chr, attraverso l'utilizzo di sonde *custom* precedentemente progettate e testate (Caso Studio I).
In Tabella 8 e in Figura 22 sono descritti i siti selezionati, la loro localizzazione geografica, ed il numero complessivo di campioni isolati per ciascuna località. Inoltre, nei paragrafi a seguire sono riportate brevi descrizioni dei contesti archeologici da cui sono stati recuperati i campioni oggetto di studio.
Per i campioni provenienti dai siti di Montericco e Ceretolo, e per 7 dei 18 campioni di Polizzello, precedenti analisi (dati non pubblicati) avevano già permesso di identificare il sesso genetico come maschile (Tabella A 1 in *Appendice*).

Sito	***Regione***	***Numero di campioni***
Montericco	Emilia-Romagna	8
Ceretolo	Emilia-Romagna	8
Norcia (Colle dell'Annunziata, Opaco, Campo Boario e Santa Scolastica)	Umbria	11
Gubbio	Umbria	15
Matelica	Marche	18
Castiglione	Lazio	13
Osteria dell'Osa	Lazio	10
Polizzello	Sicilia	18

Tabella 8: Elenco dei siti e relativo numero di campioni processati per le analisi genetiche.

Figura 22: Distribuzione geografica dei siti da cui sono stati recuperati i campioni oggetto di analisi. Le immagini mostrano manufatti e reperti ossei recuperati durante le campagne di scavo.

5.2.1 Contesti archeologici

5.2.1.1 Montericco, Imola (Emilia-Romagna)

Tra le più importanti scoperte archeologiche degli ultimi decenni nel territorio imolese vi sono quelle di alcuni sepolcreti d'età preromana, riferibili ai popoli "umbri" del VI e V secolo a.C. e ai gruppi "villanoviani" dell'VIII e VII secolo a.C. Tali sepolture contenevano oltre ai resti dei defunti anche numerosi oggetti personali e/o rituali, a testimonianza del ruolo sociale svolto in vita dal defunto e delle ideologie e riti legati al passaggio nell'aldilà. Lo scavo è stato condotto dalla Soprintendenza Archeologica in occasione della costruzione dell'Ospedale Nuovo di Imola in località Montericco ed ha portato in luce una delle più importanti necropoli riferibili alle popolazione umbre insediatesi in territorio romagnolo. In tale circostanza sono state rinvenute 77 tombe, del tipo a inumazione distesa entro fossa (spesso anche con residui di cassa lignea), che risultavano raggruppate (come avviene in altre necropoli italiche) a formare alcuni distinti 'circoli', della consistenza di alcune decine di tombe, ciascuno dei quali verosimilmente corrispondente a una famiglia o a un parentado.

Gli studi archeologici ed antropologici condotti sulla necropoli di Montericco hanno permesso di ipotizzare 3 fasi sequenziali (Figura 24) di permanenza della popolazione, durate circa 150 anni, e

contestualizzabili nel periodo terminale dell'età del Ferro[223]. La prima fase può essere collocata temporalmente nella metà del VI secolo a.C., e copre un periodo di due o tre generazioni; le restanti due fasi, della durata di circa 50 anni ciascuna, terminano con il V secolo a.C. Nella prima delle tre fasi, è stato possibile identificare, nelle tombe maschili, corredi tipici dei guerrieri, seppur con notevoli differenze tra le varie tombe, al contrario di quelle femminili in cui è stato possibile riscontrare un abbigliamento piuttosto uniforme. In questo contesto, si discostavano poco le ceramiche deposte con gli inumati, sia di sesso maschile che femminile. Non sono risultati invece presenti corredi comparabili con quelli degli adulti nelle tombe di infanti e uomini anziani; ciò ha permesso di ipotizzare che, in tale periodo, il prestigio di un individuo nella comunità dipendesse solamente dalla sua capacità di attuare un ruolo nella società[223].

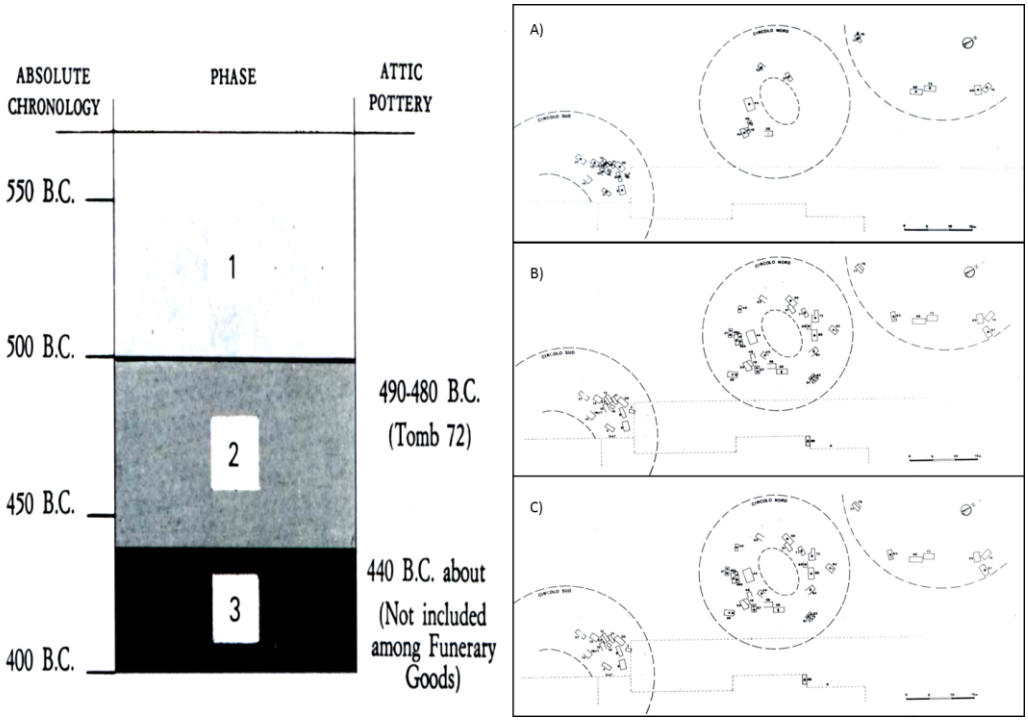

Figura 23: Cronologia delle fasi individuate nella necropoli di Montericco[223] e disposizione delle tombe nella fase 1 (a), 2 (b) e 3 (c).

Nelle tombe maschili attribuibili alla seconda fase è stato possibile individuare ancora una enfasi legata al ruolo di guerriero, mentre in quelle femminili non si sono osservati importanti cambiamenti nei corredi rispetto al periodo precedente. Tuttavia, è stata riscontrata un'importante modifica nella composizione delle ceramiche a corredo delle tombe femminili e maschili: in queste ultime è stato possibile identificare elmi, vasi di bronzo, alari e spiedi, ornamenti d'argento, punte di lancia o giavellotto, coltelli, e kylikes importati dalla Grecia. In questo secondo periodo è stato ipotizzato che le differenze sociali fossero legate in misura minore al ruolo in società; anche nelle tombe degli uomini più vecchi infatti, sono stati rinvenuti corredi comparabili a quelli dei giovani guerrieri. Inoltre, nelle tombe degli infanti con età maggiore dei 5 anni è stato possibile recuperare alcuni manufatti a corredo che permettono di ricondurre i giovani inumati allo status di guerrieri

(p.es. piccole armi, e ceramiche simili a quelle deposte nelle tombe degli adulti); in aggiunta, le loro tombe sono state individuate collocate vicine a quelle degli adulti. Non si sono osservate invece modifiche rispetto alla posizione circolare marginale e all'assenza di corredo per gli individui al di sotto dei 5 anni[223].

Nella terza fase infine, è stata osservata, per entrambi i sessi, una marcata diminuzione degli elementi personali a corredo delle tombe; sebbene gli uomini continuino ad essere identificati come guerrieri (sono state infatti recuperate una o più lance nelle tombe), il loro ruolo non è sembrato più enfatizzato come uno stato sociale. Gli inumati sono risultati normalmente accompagnati da corredi funerari di ceramica, senza grandi distinzioni tra le varie tombe. Anche i subadulti in questo periodo erano dotati di un corredo del tutto simile a quello degli adulti ed erano localizzati in vicinanza delle tombe degli individui di età superiore, dello stesso periodo. Il modello sociale di quest'ultima fase è risultato consistente con il precedente: si è osservata infatti una progressiva perdita dell'enfasi nella deposizione sulla base del ruolo sociale. Tale cambiamento può essere stato il risultato di interazioni e stimoli provenienti da gruppi stanziati nelle vicinanze e già in pieno sviluppo urbano[223].

5.2.1.2 Ceretolo, Bologna (Emilia-Romagna)

Il crescente interesse sulla storia del sito di Ceretolo nasce dalle importantissime scoperte susseguitesi dall'Ottocento in poi. Tali scoperte hanno consentito di delineare il quadro dell'evoluzione insediativa, economica e ambientale di questa zona, situata ad una decina di chilometri ad ovest di Bologna, dal Mesolitico all'età moderna.

Questa necropoli (schematizzata in Figura 23), scavata nel 1991-92, risale al primo periodo di espansione celtica nella penisola italiana, datato all'inizio del IV secolo a.C.[224,225] e consiste di 96 inumazioni e una cremazione.

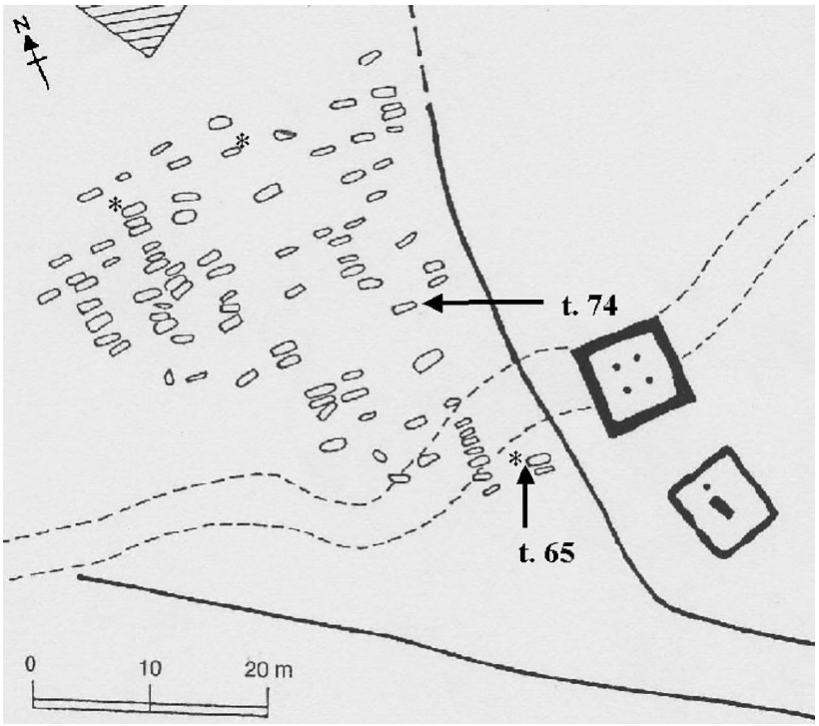

Figura 24: Mappa della necropoli celtica di Ceretolo. Da 226.

Sono stati rinvenuti corredi funerari generalmente poveri e senza alcuna prova di commistione con le popolazioni italiche, in particolare con gli Etruschi, insediatisi nella stessa regione nello stesso periodo. In quattro casi invece, è stato possibile individuare un corredo caratterizzato dalla presenza di armi (per lo più spade e punte di lancia), elemento distintivo della classe guerriera. Significativamente, l'unica tomba a cremazione è stato posizionata al centro di un recinto quadrangolare, atto chiaramente celebrativo che, accanto alla tipologia degli oggetti di corredo (tra cui un anello-sigillo in argento con castone d'oro su cui è inciso un essere fantastico), farebbe pensare alla sepoltura del più potente e prestigioso personaggio della comunità[227]. Sono stati recuperati 71 individui adulti e 23 subadulti (due nella classe di età 1-4 anni; dieci nella classe 5-9 anni; cinque nella classe 10-14 anni, e sei nella classe 15-19 anni)[228]. Di eccezionale importanza è inoltre l'area sacra con resti di strutture di culto a recinto quadrangolare, che trovano precisi confronti in contesti cimiteriali celtici di area transalpina.

5.2.1.3 Norcia (Umbria)

Ubicata nell'attuale Umbria Sud-Orientale, la città di Norcia faceva parte in antico della regione Sabina, comprendente l'alto bacino del fiume Nera e quello del Velino, ed era delimitata dai monti Sibillini, dai monti della Laga e dai monti reatini[229]. La sua fascia suburbana, in particolare, presenta due importanti zone necropolari (Figura 25) localizzate una a nord (Campo Boario) e l'altra a sud, lungo il Piano di S. Scolastica, luogo d'insediamento umano e di transito già dall'età del Ferro[230]. Varie campagne di scavo, a partire dal 1998 hanno messo alla luce 16 tombe nell'area del campo Boario e 34 inumazioni in località colle dell'Annunziata[229].

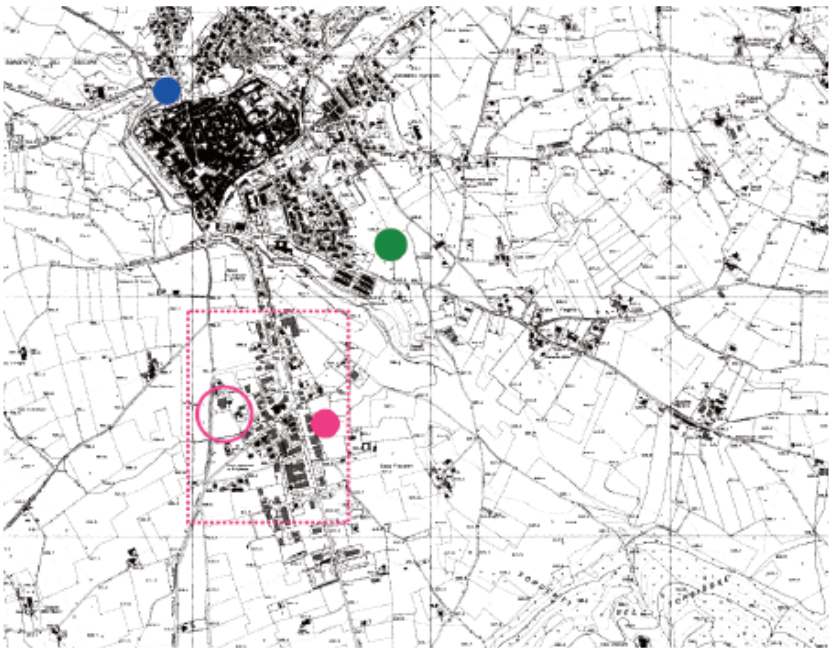

Figura 25: Necropoli di Norcia. In blu Campo Boario; il quadrato rosa rappresenta la piana di Santa Scolastica; in verde la necropoli di colle dell'Annunziata[229].

Le diverse aree destinate a necropoli si svilupparono lungo le principali direttrici viarie che si dipartivano dalla città antica: il vasto Piano di Santa Scolastica era infatti attraversato dai tracciati di raccordo con l'asse della via Salaria, sulle direttrici verso Ascoli Piceno e Cittareale, l'area del campo Boario corrisponde al punto di arrivo a Norcia della via Nursina, che la collegava con Spoleto, e il colle dell'Annunziata, posto a breve distanza dalla città, era lungo la strada che raggiungeva la conca di Castelluccio, di importanza fondamentale per i pascoli estivi delle greggi[229]. L'arco cronologico evidenziato nei diversi settori oggetto di indagine va dall'VIII secolo a.C. fino almeno al III secolo d.C. È stato possibile ricondurre alla fase arcaica (VIII-VI secolo a.C.) solo una percentuale ridotta di tombe, rispetto al numero complessivo delle deposizioni, la maggior parte delle quali sono state rinvenute in una fossa-ripostiglio scavata su uno dei lati della fossa[229].

Tra la fine del IV e gli inizi del III secolo a.C., le tombe assumono corredi ripetitivi e standardizzati, caratterizzati dalla presenza del kantharos associato all'oinochoe e alla kylix, ai quali, nella seconda metà del III secolo a.C. si aggiunge anche la situla. L'omogeneità dei corredi, ha permesso di attestare un'omogeneità anche nella struttura sociale[229].

Le tombe attribuibili alla prima generazione di coloni sono state rinvenute tutte nel Piano di Santa Scolastica, altopiano dalle importanti risorse agricole. Le inumazioni di questa prima fase di colonizzazione sono quasi sempre state ritrovate all'interno di grandi fosse rettangolari, scavate a notevole profondità all'interno del substrato ghiaioso concrezionato[229]. Era pratica comune inoltre, deporre all'interno della stessa fossa, ma a quota più alta, un secondo inumato, ed ai margini della controfossa, oppure in una nicchia laterale della fossa stessa, un'olla di ceramica comune, destinata a contenere un'offerta di liquidi[229]. È stato possibile identificare, in alcuni casi, resti organici e

chiodi di ferro ad indicazione di deposizioni all'interno di cassoni lignei. Per quanto concerne gli individui più giovani, morti in età perinatale, non sono stati individuati corredi funebri di accompagnamento. Differentemente, è stato possibile recuperare individui morti in età prepuberale nelle stesse condizioni di sepoltura degli adulti, accompagnati cioè da un corredo funebre rappresentato da forme ceramiche miniaturizzate ed asce bipenni miniaturistiche, probabilmente giocattoli infantili (se di sesso maschile), o da statuine femminili in terracotta, usate come bambole[229].

Fanno parte di questa fase, anche le tombe rinvenute in località Colle dell'Annunziata, tutte del tipo a fossa, ad eccezione della n.3 a camera, in laterizi, e della n.32 a cassone di laterizi. Risulta interessante in particolare il caso della tomba n. 32 (Figura 26), orientata grossolanamente in senso N-S, con l'inumato rinvenuto in posizione supina, deposto direttamente sul pavimento, con gli arti superiori distesi lungo i fianchi. Il cranio è risultato frantumato a causa della caduta di un filare della volta. Presso il cranio è stata rinvenuta una moneta bronzea forse collocata in bocca all'inumato, o comunque vicino alla testa; accanto alla mano sinistra sono stati evidenziati due unguentari e il manico finemente lavorato di un flabello che doveva avere almeno una lunga stecca centrale in osso. Lungo le gambe e i piedi inoltre sono state rinvenute una serie di pedine da gioco, il corredo vascolare comprendente un'olla, un'olletta e un bicchiere in ceramica comune, una coppa, una patera, un'olpe e un'olpetta in ceramica a vernice nera. La grande coppa, rinvenuta adagiata su un lato, conteneva al suo interno le due olpai e la patera. Nell'ambito del panorama locale, la sepoltura rientra tipologicamente tra le tombe destinate ad un ceto sociale ed economico elevato che ostentava il proprio *status*, mutuando modelli architettonici e forse costumi funerari dal mondo greco-orientale. Oltre all'interesse della tipologia tombaria, anche l'inumato presentava un insieme peculiare di segni patologici di particolare interesse. Lo scheletro, pressoché completo e in discreto stato di conservazione, presenta una consistenza friabile e ampie aree di erosione legate soprattutto alla giacitura in ambiente aerato. Si tratta di un soggetto anziano e di sesso maschile, con un complesso quadro patologico, tra cui entesopatie, anchilosi completa e bilaterale dell'articolazione sacroiliaca, assottigliamento dello spessore della teca cranica, dei piatti scapolari, delle ali iliache e della mandibola, oltre a neurinoma (Schwannoma) del nervo acustico[229].

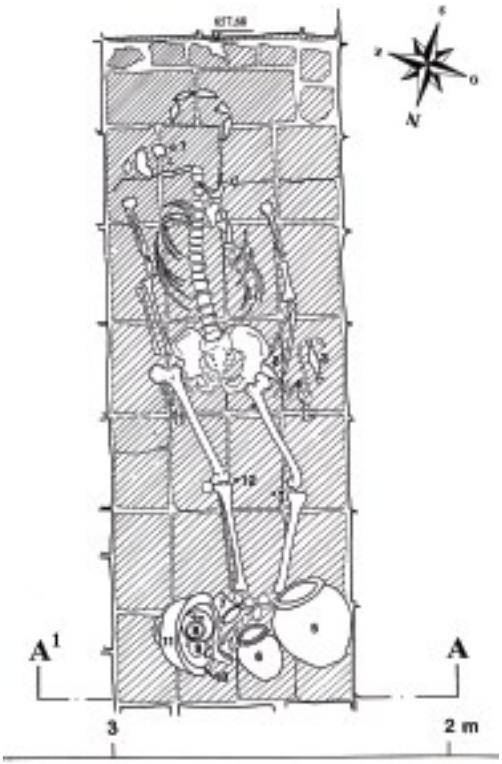

Figura 26: Disposizione dell'inumato e del corredo funerario della tomba n.32 di colle dell'Annunziata[229].

5.2.1.4 Gubbio (Umbria)

Importante centro politico e religioso umbro, Gubbio si localizza sul pendio del Monte Ingino in corrispondenza del centro medievale. Gli scavi condotti nella via centrale hanno messo in luce una necropoli di età proto-villanoviana e restituito quaranta sepolture in cinerari d'impasto. Occupata già dalla fine dell'età del Bronzo da insediamenti e necropoli, controllava il territorio con una fitta rete di castellieri e nel VII-VI secolo a.C. si è arricchita di presenze archeologiche con la necropoli di S. Biagio e le sue tombe a circolo e poi con il V secolo a.C. con le sepolture della Vittorina, tombe che continuano fino al II secolo d.C., costituendo così la necropoli più grande e longeva della città[231].

È stato possibile riferire alla fase più antica della necropoli, tre grandi circoli tombali, di ciascuno dei quali si riconosce con sicurezza il limite perimetrale delimitato da lastre di arenaria disposte verticalmente e grossolanamente squadrate. All'interno di questi limiti i defunti (presumibilmente appartenenti ad uno stesso gruppo familiare) venivano sepolti in fosse scavate nel terreno, adagiati supini sopra uno strato di breccione, con la mano destra sopra il bacino e la sinistra disposta sotto o lungo il fianco, accompagnati a volte da oggetti di corredo, tra cui vasi di differenti fogge e dimensioni, gioielli in ferro o in bronzo e più raramente armi. I corredi, per il numero dei manufatti e per la loro qualità, hanno permesso talvolta di distinguere personaggi di rango. Alcuni includono, oltre ai più frequenti vasi di impasto (olle, piattelli o ciotole), anche oggetti di importazione, quali

gocce d'ambra e oggetti d'osso finemente lavorati; talvolta è stato possibile recuperare vesti decorate con minuscoli cerchietti d'osso.

La necropoli fu utilizzata in fasi successive e alcuni dei circoli tombali andarono distrutti. Sepolture più tarde recuperarono come copertura dell'area sepolcrale quelle stesse lastre di arenaria che delimitavano i circoli più antichi. Particolarmente interessanti risultano due tombe a fossa scoperte nel 1993 con inumati sepolti entro cassa lignea, databili la prima nella seconda metà del VII secolo a.C. e la seconda nella seconda metà del VI secolo a.C.

A partire dalla fine del IV secolo a.C., l'area della necropoli di S. Biagio appare interessata da un buon numero di sepolture singole, la cui disposizione non consente di individuare raggruppamenti di tipo familiare. Si tratta di inumazioni a fossa scavate nella nuda terra o di tombe alla cappuccina. È stato possibile identificare in alcune sepolture corredi costituiti da pochi oggetti, essenzialmente da vasi, talvolta di un certo pregio.

Tra la fine del III e la metà del II secolo a.C., nell'ultima fase di utilizzo della necropoli, solo una parte ridotta dell'area è stata occupata da sepolture del tipo al1a cappuccina, quasi sempre prive di corredo, ad eccezione di sette tombe infantili rinvenute nella fascia sud-orientale dell'area indagata.

Figura 27: Mappa di distribuzione delle presenze archeologiche a Gubbio[231].

5.2.1.5 Matelica (Marche)

Matelica presenta un ricco patrimonio archeologico: le prime testimonianze di insediamento nell'area risalgono al Paleolitico medio e Paleolitico superiore come è stato evidenziato da dati stratigrafici, ma il popolamento diventa particolarmente diffuso durante la tarda età del Ferro come testimoniato delle grandi necropoli a circoli nelle località Pine dell'Incrocca e Crocifisso (Figura 28). Le comunità stanziate erano caratterizzate dalla presenza di una privilegiata classe rurale che traeva il suo benessere dallo sfruttamento agricolo delle fertili terre. Tale situazione è rimasta immutata fino all'epoca romana.

L'abitato di età picena (italico) occupava tutta l'altura che si estende dalla confluenza dell'Esino con il rio Imbrigo a nord fino alle aree situate a destra dell'ansa dell'Esino a Sud. Le tracce di un insediamento precedente la fondazione del municipio romano consistono in resti di ceramica di pregio e di uso comune databili a partire dal III secolo a.C., frequenti in tutti gli scavi archeologici condotti in profondità.

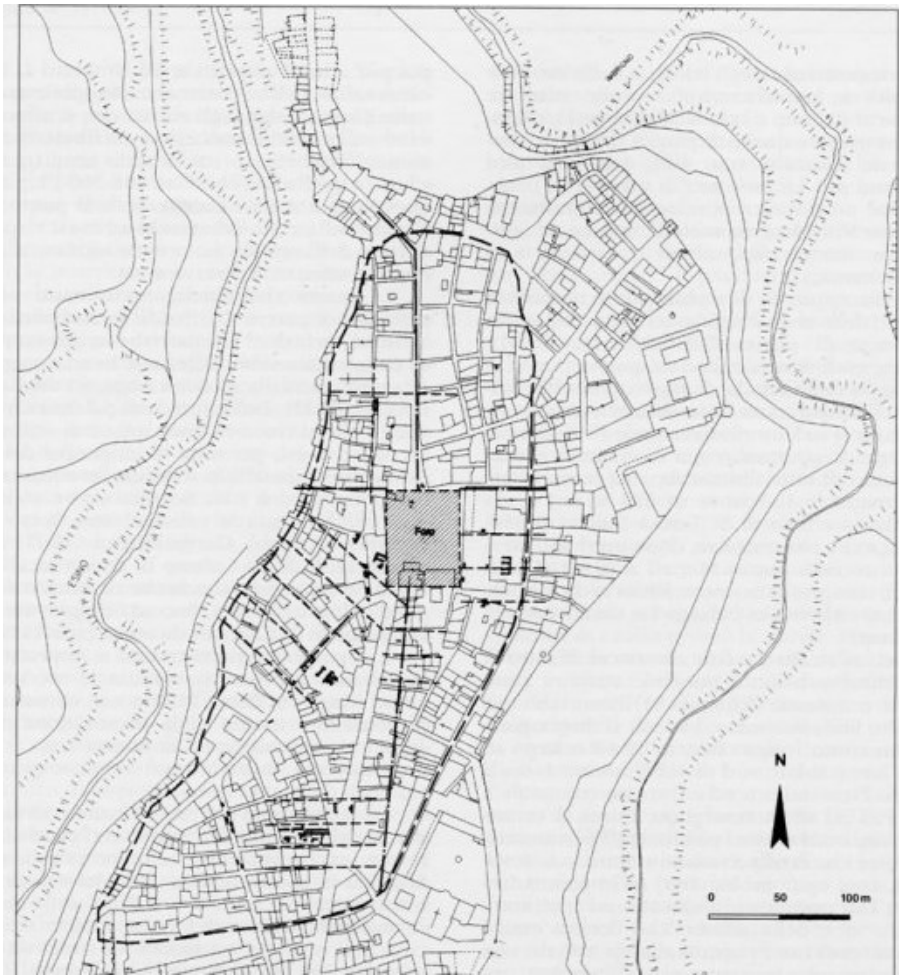

Figura 28: Necropoli di Matelica.

5.2.1.6 Castiglione (Lazio)

Situato presso la fascia interna del lato orientale del cratere di Castiglione, è possibile osservare l' abitato della fase finale della media età del Bronzo. Frammisti ai materiali della media età del Bronzo vi sono anche pochi frammenti databili al Bronzo recente. L'abitato è stato successivamente intaccato dall'omonima necropoli dell'età del Ferro[232]. La necropoli si trova sulla cresta orientale del cratere di Castiglione. L'insieme dei dati sembrerebbe indicare una comunità poco aperta ai contatti esterni. È stato infatti possibile individuare un numero minore di ornamenti, una ceramica meno curata e raramente decorata, un rituale dell'incinerazione non specializzato e una casualità dell'associazione tra sesso e corredo, rispetto ad altre necropoli dello stesso periodo[232].

5.2.1.7 Osteria dell'Osa (Lazio)

Ai margini estremi del vasto territorio dei Castelli Romani, in prossimità del chilometro diciassettesimo della Via Prenestina si erge l'importante necropoli dell'età del Ferro di Osteria dell'Osa. Datata nel periodo compreso tra il IX e i primi decenni del VI secolo a.C., è composta da un insieme di circa 600–700 tra tombe e sepolture. Le ricerche effettuate in luogo hanno permesso di riportare alla luce una quantità notevole di reperti che descrivono indizi molteplici delle pratiche funerarie esercitate nell'epoca, mediante l'incenerimento delle spoglie o la deposizione attraverso l'inumazione dell'intero corpo a contatto con il terreno. Gli scavi archeologici hanno permesso una valutazione approfondita delle caratteristiche fisiche e morfologiche dei defunti, delle condizioni di vita quotidiana, delle derrate alimentari a disposizione per il personale fabbisogno e delle professioni talora svolte dalle popolazioni dispiegate nell'insediamento. Le conoscenze estrapolate sono risultate particolarmente utili ai fini di un'ipotetica ricostruzione della struttura e dell'organizzazione comunitaria, e delle relazioni emerse con il territorio circostante e i rapporti intercorsi con le altre civiltà presenti. Le comunità erano legate tra loro da strutture di parentela, guidate da capi guerrieri e sacerdoti. Vivevano in capanne e si approvvigionavano presso il Lago di Castiglione. In alcuni periodi dell'anno lavoravano la ceramica, ed erano ottimi utilizzatori dell'argilla. Il legno era conosciuto e sfruttato. Tra la fine del IX e VIII secolo a.C. in considerazione degli intensi contatti tra l'Etruria e la Campania, in queste comunità egualitarie hanno luogo delle trasformazioni sociali; si avverte la formazione dei centri protourbani anticamera di quelli urbani tipici del territorio laziale.

5.2.1.8 Polizzello (Sicilia)

Il sito archeologico di Polizzello sorge su una montagna a 800 metri di altezza, a breve distanza da Mussomeli, nel cuore della Sikania (la parte centrale della Sicilia, antico nome della regione abitata in epoca protostorica dai Sicani). In passato, la montagna offriva le condizioni ottimali per la difesa

dagli attacchi dei nemici. Il territorio circostante, delimitato dai fiumi Platani e Salito, era molto fertile e produttivo a causa delle numerose sorgenti d'acqua. Dalla remota antichità questi fiumi furono un'importante rotta verso l'interno dal mare, spiegando la scelta di un insediamento in questo luogo e il ruolo cruciale svolto dal centro indigeno di Polizzello durante la preistoria e protostoria e la sua influenza culturale oltre i confini del territorio siciliano[233].

5.2.2 Processamento dei campioni

Le rocche petrose di 101 individui provenienti da 8 necropoli del centro e Sud Italia (Tabella A 1 in *Appendice*) sono stati selezionati sulla base delle informazioni archeologiche e dello stato di conservazione dei distretti anatomici scelti per le analisi molecolari. Tutte le tombe sono state associate, sulla base della zona stratigrafica di recupero, abbinata al corredo funerario e alla ricostruzione storica generale del sito, al periodo dell'età del Ferro. I campioni ossei sono stati recuperati (dalla sede anatomica del cranio o isolati) rispettando tutte le norme preliminari per lo studio molecolare di campioni antichi, al fine di minimizzare possibili contaminazioni con il DNA moderno degli operatori. Sono stati inoltre prelevati i tamponi buccali di tutto il personale coinvolto nello studio al fine di monitorare potenziali fonti di contaminazione.

Tutte le analisi molecolari sono state condotte presso il Laboratorio di Antropologia Molecolare e Paleogenetica dell'Università degli Studi di Firenze, attrezzato per disporre di aree di lavoro separate (fisicamente e con accessi indipendenti), per le fasi analitiche che precedono e seguono l'amplificazione del DNA genomico. Sono stati inoltre rispettati rigorosi standard di lavoro, sviluppati nel corso degli anni per evitare di apportare contaminazioni ai substrati biologici oggetto di analisi[33,234]. In particolare, durante la manipolazione dei campioni e nel corso di tutte le procedure sperimentali che precedono l'amplificazione del DNA, sono stati indossati indumenti monouso sterili (tuta da lavoro, doppio paio di guanti, sovrascarpe, maschera facciale). Il piano di lavoro e gli strumenti sono stati regolarmente puliti, prima e dopo ciascun esperimento, con NaClO al 50% ed etanolo al 96%. Inoltre, prima e dopo ogni operazione, nella zona di pre-amplificazione, sono stati utilizzati raggi UV con lunghezza d'onda a 254 nanometri (nm) come metodo di sterilizzazione fisico, in aggiunta ad un ciclo notturno di circa 6 ore predisposto per la degradazione di eventuali molecole di DNA depositate sulle superfici e apparecchiature del laboratorio. In ogni fase del lavoro sono stati utilizzati materiali sterili e pipette dedicate. Inoltre, durante ogni operazione sperimentale è stato portato avanti anche un controllo negativo, nel quale sono stati inseriti tutti i reagenti utilizzati ad eccezione del materiale biologico, in modo da monitorare l'eventuale presenza di contaminazioni nei reattivi utilizzati.

Tutti i campioni analizzati sono stati sequenziati presso il presso il Laboratorio di Genomica Avanzata dell'Università di Firenze.

5.2.2.1 Pulizia e polverizzazione dei campioni

Per prima cosa i campioni in analisi, tutti rappresentati da rocche petrose, sono stati sottoposti a pulizia mediante l'utilizzo di spazzolini da denti sterili e specilli in modo da asportare tutti i residui di terra dagli strati superficiali e dai meati acustici interni, ed eliminare eventuali tracce di materiale biologico esogeno. Lo strato esterno è infatti il più contaminato, a causa dell'esposizione all'ambiente di sepoltura, oltre che al maneggiamento durante le fasi di scavo e recupero. Tra un campione ed il successivo, il piano di lavoro e gli strumenti utilizzati sono stati puliti e sterilizzati (con le procedure descritte nel paragrafo precedente) così da evitare cross-contaminazioni: sul banco eventuali residui di polvere sono stati aspirati; successivamente il piano e lo strumento sono stati puliti con ipoclorito di sodio. Terminata la fase di pulizia dei campioni, questi sono stati irradiati con luce UV a 254 nm in un *crosslinker* (Biolink DNA crosslink), per 20 minuti su ogni lato.

Le rocche sono state successivamente tagliate longitudinalmente al margine superiore della piramide attraverso l'uso di un trapano odontoiatrico con lama circolare diamantata, al fine di esporre la zona cocleare, distretto elettivo per il prelievo di materiale osseo per le sue caratteristiche di estrema compattezza che favoriscono la preservazione del materiale biologico in modo esclusivo[16]. Nuovamente, si è proceduto a sterilizzazione della regione interna con luce UV a 254 nm per 20 minuti.

In seguito alla pulizia tutti i campioni sono stati polverizzati con un trapano odontoiatrico con fresa sterile monouso impostato a bassa velocità in modo da ridurre al minimo il surriscaldamento, il quale può provocare un ulteriore danno del materiale biologico di interesse. Per ciascun campione sono state ottenute due aliquote di circa 50-60 mg di polvere d'osso, conservate in tubini sterili da 2 mL alla temperatura di -20°C fino alla successiva fase di estrazione del DNA.

Durante queste prime due fasi sono state annotate le eventuali caratteristiche del distretto osseo che avrebbero potuto interferire nelle successive analisi: in particolare la presenza di fratture a livello dell'apice piramidale, o l'elevata porosità del distretto osseo, fattori determinanti per agevolare la penetrazione negli strati più profondi dell'osso di materiale biologico contaminante e di altri fattori ambientali che possono alterare lo stato di conservazione della componente genetica endogena.

5.2.2.2 Estrazione del DNA

Poiché l'aDNA ha già subito molti processi di danneggiamento, il metodo di estrazione utilizzato deve evitare ulteriori ed eccessivi trattamenti aggressivi, quali le alte temperature o i detergenti forti; sebbene questi processi facilitino il rilascio del DNA dalle strutture che lo proteggono, andando a solubilizzare le membrane cellulari, possono infatti determinare un'ulteriore diminuzione nella concentrazione del doppio filamento. In più i metodi che vengono utilizzati per il trattamento dei reperti antichi devono prevedere l'uso di sostanze che vadano ad interagire con gli inibitori di PCR naturalmente depositati nel campione e che possono essere recuperati con esso, impedendo le fasi successive del lavoro. Rohland e Hofreiter, nel 2007[17] hanno messo a punto un metodo di estrazione, rivisto e modificato 6 anni dopo da Dabney e collaboratori[24], per massimizzare l'ottenimento di DNA amplificabile ed il recupero di frammenti corti, a partire da materiale osseo antico, e allo stesso tempo per minimizzare la co-estrazione di sostanze che inibiscono la PCR. Si tratta di un processo in cui la polvere d'osso viene combinata con un *buffer* per la digestione dei residui cellulari e degli inquinanti, ed il lisato ottenuto viene purificato per ottenere il solo DNA tramite la reazione di legame della molecola su silice. Il protocollo che verrà descritto è stato seguito anche per questo lavoro di tesi. Il processo si articola in due giorni di lavoro, ed è caratterizzato da pochi *step* e reagenti; inoltre permette di ottenere buone quantità di DNA con poche decine di milligrammi di polvere d'osso, caratteristica estremamente vantaggiosa, date le esigue quantità di materiale di partenza di cui si può disporre in studi condotti su reperti archeologici. Il primo giorno di lavoro sono stati preparati tre *buffer*, descritti di seguito.

1. Extraction Buffer, una soluzione per la lisi delle cellule, viene fatto agire in *overnight*, e attraverso reazioni di decalcificazione e digestione enzimatica dei residui cellulari, comporta il rilascio in soluzione del materiale genetico contenuto nel campione biologico. Per ciascun campione la miscela è composta da:
   - 900 µL di EDTA (acido etilendiamminotetraacetico) 0.5 M a pH 8.0, un chelante cationico con azione sia di tipo disgregativo (legando il calcio genera un indebolimento della componente inorganica dell'osso), che inattivante (agendo su cationi che funzionano da cofattori di enzimi litici, blocca le degradazioni interne, come quella portata avanti dalle DNA-asi);
   - 25 µL di Proteinasi K (10 mg/ml), una serin-proteasi ad ampio spettro d'azione, in grado di scindere il legame peptidico adiacente al gruppo carbossilico di amminoacidi alifatici o aromatici con gruppi amminici in posizione alfa bloccati;
   - 74.5 µL di acqua di grado HPLC;
   - 0.5 µL di Tween 20, surfactante non ionico usato come detergente ed emulsionante per la solubilizzazione delle membrane proteiche.

2. Binding Buffer, utilizzato il secondo giorno, è un composto che favorisce il legame del DNA alla silice (contenuta nelle membrane delle colonnine di lavaggio). Si prepara miscelando, con le dosi definite da protocollo, i seguenti reagenti:
   - Guanidina idrocloride 5M, un sale caotropico che degrada la struttura terziaria delle proteine, e coadiuva il legame DNA-silice;
   - Acqua di grado HPLC per risospendere la guanidina;
   - Isopropanolo (40%), utilizzato per favorire il *salting out*, ovvero la precipitazione del DNA, grazie all'interferenza con la sfera di solvatazione delle molecole d'acqua presenti attorno ai gruppi fosfato del doppio filamento;
   - Tween 20.

3. TET Buffer, utilizzato anch'esso il secondo giorno, come *buffer* di eluizione finale. È prodotto *una tantum* secondo le seguenti specifiche:
   - Acqua di grado HPLC;
   - EDTA 0.5 M a pH 8.0;
   - Tween 20;
   - Tris-HCl 1 M a pH 8.0, usato come soluzione tampone per favorire il *salting-in*; in questa fase la deprotonazione dei gruppi fosfato e dei gruppi imminici fa sì che il DNA perda la propria affinità per la membrana in silice della colonnina, e possa essere eluito in un nuovo tubino.

I *buffer*, prima dell'uso, sono stati sterilizzati per 45 minuti sotto UV. Il primo giorno è stato aggiunto 1 mL di Extraction Buffer alla polvere d'osso in tubini da 2 mL sterili, e la miscela è stata incubata *overnight* in una stufa sotto rotazione continuativa a 37°C. Il giorno seguente dopo 2 minuti di centrifugazione a 15300 rpm per favorire la compattazione della polvere non disgregata in fondo al tubino (*pellet*), è stato recuperato il surnatante al quale sono stati aggiunti 10 mL di Binding Buffer e 400 µL di Sodio Acetato 3M a pH 5.2 necessario per annullare le cariche negative dei gruppi fosfato, ed in tal modo favorire l'adesione del DNA alla membrana in silice. La soluzione così ottenuta è stata trasferita all'interno di High Pure Spin Filter (Roche Molecular Systems, Inc., CA, USA), ovvero tubi conici da 50 mL contenenti un imbuto al termine del quale è collocata una colonnina con membrana in silice. Il dispositivo è stato oggetto di una prima centrifuga a 1500 rpm per 4 minuti e una seconda, in seguito a rotazione di 90°, alla stessa velocità, ma per 2 minuti. In questo modo è possibile coadiuvare il legame DNA-silice e contemporaneamente permettere la percolazione di tutta la soluzione contenente i *buffer* e la maggior parte delle macromolecole di scarto, nel fondo dei tubi da 50 mL. Successivamente, la colonnina in silice di ciascun High Pure Spin Filter è stata collocata nel *collection tube* per le fasi di lavaggio: in seguito ad una centrifuga a secco a 6000 rpm per 1 minuto, sono stati portati avanti due passaggi sequenziali di aggiunta di Buffer PE (Qiagen) per l'eliminazione delle impurità

residue, con successiva centrifuga a 6000 rpm per 30 secondi. Dopo ciascuna centrifugazione il liquido percolato è stato scartato, e a seguito del secondo lavaggio, è stata eseguita una ulteriore fase di centrifugazione a secco (13200 rpm per 1 minuto), atta ad asciugare completamente la membrana della colonnina ed eliminare del tutto ogni residuo del *buffer* di lavaggio, il quale, contenendo alcool è in grado di inibire le successive fasi di lavoro. Ogni colonnina è stata infine trasferita in nuovi tubini da 1,5 mL siliconati per l'eluizione finale del DNA in 100 µL di TET, in due ripetizioni da 50 µL, intervallate da una centrifuga a 13200 rpm. Gli estratti così ottenuti per ciascun campione sono poi stati conservati in freezer, a -20°C, fino alle successive fasi. Durante l'estrazione è stato portato avanti anche un controllo negativo, nel quale sono stati inseriti tutti i reagenti utilizzati ad eccezione del materiale biologico, in modo da monitorare l'eventuale presenza di contaminazioni.

5.2.2.3 Preparazione delle librerie per il sequenziamento Illumina

La preparazione delle librerie Illumina consiste in *step* successivi in cui si procede ad un preliminare trattamento in presenza degli enzimi uracil–DNA–glicosilasi (UDG) ed Endonucleasi VIII (progettati per scindere le citosine deaminate nella regione interna dei filamenti di aDNA e ristabilire le corrette basi azotate) e successivamente a riparare le estremità dei filamenti ai quali vengono poi attaccati adattatori universali per il sequenziamento su piattaforma Illumina; infine si procede alla rimozione di eventuali *nicks* ed alla purificazione del materiale ottenuto. Ad oggi la *library* ha rimpiazzato nettamente la PCR come metodo di arricchimento del DNA da sequenziare. Per la produzione delle librerie a partire dal DNA precedentemente estratto, è stato seguito un protocollo ottimizzato per i campioni antichi[20]. Esso è particolarmente indicato nelle analisi che coinvolgono il nuDNA, dal momento che il ridotto numero di copie per cellula, rispetto a quelle di mtDNA ed ai possibili contaminanti, determina una minore probabilità di recuperare i frammenti di interesse, e risulta pertanto essenziale per una maggior garanzia dell'autenticità del dato prodotto. A questo proposito infatti, il trattamento con l'enzima UDG viene effettuato in maniera solo parziale, al fine di mantenere un segnale di danno alle estremità delle molecole antiche.
In questa fase, oltre al controllo negativo dell'estrazione, è stato inserito un ulteriore controllo negativo (K- Library).

Trattamento con UDG e riparazione dei danni

Il trattamento preliminare del DNA genomico estratto con UDG è stato eseguito attraverso l'allestimento della reazione descritta in Tabella 9. Un volume di 30 µL di DNA (o acqua nel caso del controllo negativo) sono stati incubati in una soluzione contenente 10X CutSmart buffer, 25 mM dNTPs, 10 mM ATP e 1 U/µL USER *enzyme* (New England Biolabs, Inc., Massachusetts, USA). Quest'ultimo reagente in particolare, è una miscela degli enzimi UDG e DNA glicosilasi-

liasi endonucleasi VIII, il primo dei quali catalizza l'escissione di una base di uracile, formando un sito abasico (apirimidinico) lasciando intatta la spina dorsale del fosfodiestere; l'attività liasica dell'endonucleasi VIII rompe la spina dorsale del fosfodiestere ai lati 3 'e 5' del sito abasico in modo che venga rilasciato desossiribosio privo di basi. Tale reazione si verifica a seguito di un'incubazione a 37°C per 30 minuti, al termine dei quali l'aggiunta di 2 U/µL di UGI (Uracil Glycosylase Inhibitor - New England Biolabs, Inc.) permettono, con una nuova incubazione a 37°C per 30 minuti l'inattivazione dell'UDG.

Trattamento UDG				
Reagenti	Concentrazione Stock	Concentrazione Finale	Unità di misura	Volume per campione (µL)
H2O				6.36
CutSmart buffer	10	1	X	6
dNTPs	25000	100	µM	0.24
ATP	10	1	mM	6
USER enzyme	1	0,06	U/µL	3.6
DNA o dH2O				30

Tabella 9: Reagenti per l'allestimento della reazione in presenza di UDG.

La riparazione dei *nick* prodotti dal precedente trattamento, così come quelli alle estremità delle molecole derivanti da cause di degradazione naturali, sono stati riparati attraverso la consecutiva aggiunta, nel medesimo tubino di reazione, di 3 µL di 10 U/µL T4 PNK (New England Biolabs, Inc.) e 1.2 µL di 5 U/µL T4 DNA polimerasi. La reazione è stata ulteriormente incubata in TC a 25°C per 15 minuti e successivamente a 12°C per altri 5 minuti. Al termine della fase di reazione il prodotto ottenuto è stato purificato in colonnine MinElute (Qiagen), secondo le istruzioni descritte da manuale, con eluizione finale in 18 µL di TET. L'eluito è stato interamente usato per la fase successiva di legame degli adattatori.

Adapter Ligation

Questa fase consente il legame delle molecole di DNA con gli adattatori (P5 e P7) specifici per il sequenziamento Solexa/Illumina. In ogni eluito preparato nella reazione precedente, sono stati aggiunti 22 µL di una mix prodotta come descritto in Tabella 10.

Adapter Ligation				
Reagenti	Concentrazione Stock	Concentrazione Finale	Unità di misura	Volume per campione (µL)
Quick Ligase Buffer	2	1	X	20
Solexa adapter mix	10000	250	nM	1
Quick Ligase	5	0,125	U/µL	1
DNA o dH_2O				18

Tabella 10: Reazione di adapter ligation.

La mix di adattatori (Solexa adapter mix), è stata preparata a partire da tre oligonucleotidi, IS1_adapter_P5.R (5'-AATGATACGGCGACCACCGA), IS2_adapter P7.F (5'-CAAGCAGAAGACGGCATACGA) e IS3_adapter_P5+P7 (3'-AGATCGGAAGAGC), complementare solo a parte dei due adattatori; gli adattatori presentano un'estremità piatta e l'altra coesiva e questo favorisce la loro corretta orientazione nel legame al DNA.
L'enzima *Quick Ligase* (New England Biolabs, Inc.) catalizza la reazione di legame tra gli adattatori e le molecole di DNA a seguito di un'incubazione per 20 minuti a temperatura ambiente. Al termine della reazione è stato portato avanti un nuovo processo di purificazione con MinElute Purification Kit (Qiagen, Hilden, Germany) come descritto precedentemente. Il materiale ottenuto è stato eluito in 16 µL di TET, interamente usati nell'ultima fase di preparazione della *library*.

Adapter Fill-in

L'ultimo *step* prevede la riparazione di eventuali *nick* che possono crearsi nella reazione precedente, e la rigenerazione di estremità piatte. La mix preparata per ciascun campione (in un volume finale di 9 µL) contiene i reattivi riportati in Tabella 11.

Fill in				
Reagenti	Concentrazione Stock	Concentrazione Finale	Unità di misura	Volume per campione (µL)
Isothermal buffer	10	1	X	2,5
dNTPs	2500	125	nM	0,25
Bst ploymerase 2.0	8	0,4	U/µL	2
H_20				4,25
DNA o dH_2O				16

Tabella 11: Mix di reazione per la fase di fill-in.

I 25 µL finali, contenenti i reagenti ed il DNA, sono stati sottoposti al seguente profilo termico: 37°C per 20 minuti, e a seguire, altri 20 minuti ad 80°C. Il prodotto di reazione della Fill-in è stato interamente utilizzato per la successiva fase di incorporazione degli indici campione-specifici (senza necessitare dello *step* di purificazione).

Indexing PCR, quantificazione ed amplificazione del materiale genetico

Uno dei principali vantaggi nell'uso di NGS è quello di poter processare più campioni in parallelo. Per questa ragione è necessaria l'incorporazione di *barcodes* (indici) campione-specifici, così da associare ciascuna lettura ottenuta al campione che l'ha generata. In questo caso, la metodologia Illumina si differenzia dagli altri sequenziatori ultra-massivi, in cui l'indice viene inserito tra uno dei due adattatori e l'estremità della molecola di DNA di partenza. Nel caso di Illumina infatti, gli

indici vengono inseriti solo dopo aver completato la libreria, la quale potrà essere costruita in modo identico per tutti i campioni in studio. Ogni indice è costituito da sequenze lunghe da 6 ad 8 bp inserite all'interno di un *indexing-primer* che ha, da un lato, la sequenza complementare a quella degli adattatori (P5 e P7) e dall'altro, una sequenza utilizzabile per i successivi *step* di amplificazione e quantificazione. Durante questa fase, basata su una reazione di PCR, oltre all'attacco degli indici, si va anche ad amplificare il numero di molecole che hanno incorporato gli adattatori.

Per ciascun campione, il prodotto di Fill-in è stato suddiviso in quattro aliquote, così da aumentare l'efficienza delle singole reazioni. La mix, per un volume finale di 100 µL si compone di 10 µL di Buffer 10X; 1 µL di dNTP mix 25 mM; 1,5 µL di BSA 20 mg/mL ;1 µL di Pfu Turbo Polymerase; 72,5 µL di acqua, per un totale di 86 µL per ciascun tubino. A questa, sono stati aggiunti 4 µL di ciascun indice (secondo la Tabella A 2 in Appendice) e 6 µL di libreria. I tubini di reazione sono stati trasferiti in un TC e amplificati con il seguente profilo termico: 95 °C per 2 minuti; 15 cicli costituiti da una fase a 95°C per 30 secondi; una a 58°C per 30 secondi e un ultimo *step* a 72°C per 1 minuto; 72°C per 10 minuti.

Al termine della reazione le librerie indicizzate sono state purificate con MinElute Purification Kit (Qiagen); durante la purificazione tutte le aliquote di uno stesso campione sono state riunite in un'unica colonnina MinElute aggiungendo 500 µL di PB Buffer per ogni aliquota e centrifugando ad ogni passaggio; al termine delle quattro aliquote sono stati addizionati 750 µL di PE Buffer e 20 µL di TET sono stati utilizzati per l'eluizione. L'efficienza della reazione di *indexing* è stata valutata quantitativamente e qualitativamente con TapeStation 4150 System (kit High Sensitivity D1000 ScreenTape – Agilent Technologies).

Laddove necessario, le librerie indicizzate sono state sottoposte ad un unico ciclo di PCR, al fine di rimuovere gli eteroduplici, ovvero molecole ibride che si formano a causa dell'appaiamento di porzioni terminali complementari di filamenti diversi nei cicli finali della reazione di PCR, nel caso di una saturazione dei reattivi in soluzione. Per la reazione è stata allestita una mix, in un volume finale di 100 µL, utilizzando il kit Herculase II Fusion DNA Polymerase (Agilent Technologies), costituita da 20 µL di Fusion Buffer 5X; 1 µL di dNTPs mix (25mM each); 4 µL di IS5 10 mM; 4 µL di IS6 10 mM; 1 µL di Herculase Polymerase e 67 µL di acqua. Per una migliore resa ciascun campione è stato diviso in due aliquote, e la reazione è stata incubata in un TC usando il seguente profilo termico: 95°C per 2 minuti; 95°C per 30 secondi; 60°C per 30 secondi; 72°C per 30 secondi; 72°C per 5 minuti.

Al termine della reazione i prodotti sono stati purificati con il MinElute Purification kit ed il DNA è stato eluito in 20 µL di TET. È stata eseguita un'ulteriore quantificazione alla TapeStation 4150 System, per la valutazione finale del materiale genomico prodotto.

5.2.2.4 Sequenziamento shotgun e determinazione del sesso

Le librerie dei 78 campioni per i quali non si avevano informazioni relative al sesso ed oggetto di studio, sono state sottoposte ad un sequenziamento preliminare al fine di determinare il sesso genetico di ciascun individuo e valutare i principali parametri di qualità delle molecole recuperate, quali la percentuale di DNA endogeno e i *pattern* di degradazione delle molecole, utili per determinare l'autenticità del dato.

A tal fine, le librerie genomiche prodotte sono state diluite ad una medesima concentrazione e successivamente riunite in un *pool* equimolare (1.2 nM) di volume finale pari a 100 µL per la corsa su piattaforma Novaseq 6000 di Illumina (SP Reagent Kit, 100 cicli, 2x50 +8+8). I *file* prodotti in formato .bcl sono stati convertiti nel formato FastQ compresso e "demultiplexati" utilizzando il *software* Illumina CASAVA-1.8.2 al fine di associare ciascuna lettura al campione da cui è stata prodotta, sulla base delle coppie di indici utilizzati (Tabella A 2 in Appendice), come descritto nel *Paragrafo* 4.2.6.1.

Il software EAGER[179], è stato utilizzato per il processamento preliminare delle *reads* grezze (rimozione degli adattatori ed unione delle letture in *forward* ed in *reverse*), per l'allineamento delle *reads* sul genoma umano di riferimento hg19, incluso il mitocondrio[184] e per valutare le percentuali di DNA endogeno e di degradazione alle estremità dei frammenti processati e filtrati sulla base delle qualità di chiamata, come già ampiamente spiegato nel *Paragrafo* 4.2.6.1.

Infine, per la determinazione del sesso biologico degli individui, è stato utilizzato lo *script* prodotto da Skoglund et al.[235] al fine di calcolare il rapporto (Ry) delle letture mappate sul Y-chr rispetto al numero complessivo di sequenze allineate su entrambi i cromosomi sessuali.

I campioni di sesso maschile, per i quali i dati di qualità e di autenticità sono risultati ottimali, sono stati utilizzati per la cattura di 31630 SNPs del Y-chr di interesse filogenetico, attraverso l'uso di sonde *custom* MyBaits (Arbor Biosciences).

5.2.2.5 Cattura del Y-chr e sequenziamento

Sono stati catturati un totale di 42 campioni di sesso maschile, oltre a 2 femmine utilizzate come controllo negativo.

I campioni selezionati sono stati arricchiti secondo un numero variabile di cicli di PCR, calcolati sulla base delle concentrazioni del materiale di partenza e del fattore di arricchimento atteso ad ogni ciclo, al fine di giungere alla concentrazione genomica di partenza ottimale stimata in 12 µg. Le reazioni di PCR sono state eseguite attraverso il kit Herculase II Fusion Enzyme (Agilent Technologies), con le concentrazioni ed i volumi descritti in Tabella 12. I *primer* utilizzati per l'arricchimento (IS5-IS6) sono caratterizzati da una complementarietà di legame con le regioni

degli adattatori universali Illumina presenti a monte dell'inserto. Questo permette di amplificare correttamente le sole molecole contenenti gli adattatori e correttamente indicizzate.

Arricchimento con kit Herculase II Fusion Enzyme				
Reagenti	Concentrazione Stock	Concentrazione Finale	Unità di misura	Volume per campione (µL)
Herculase II Fusion buffer	5	1	X	20
dNTPs	25	0,25	mM	1
IS5	10	0,4	mM	4
IS6	10	0,4	mM	4
Herculase II Fusion DNA polymerase	5	0,05	U/µL	1
H_20				41
Tot				71
DNA				29

Tabella 12: Reagenti utilizzati per l'arricchimento delle librerie genomiche pre-cattura.

Per favorire una migliore efficienza di reazione, ciascun campione è stato suddivido in 3 aliquote di pari volume. La reazione è stata incubata in TC secondo il seguente profilo termico: 95°C per 2 minuti; 4-6 cicli a 95°C per 30 secondi, 60°C per 30 secondi, 72°C per 30 secondi; 72°C per 5 minuti. Al termine del processo di arricchimento, le aliquote di ciascun campione sono state purificate all'interno della medesima colonnina MinElute, secondo le specifiche precedentemente descritte. Il DNA è stato eluito in un volume finale di 18 µL e quantificato mediante TapeStation 4150 System (kit High Sensitivity D1000 ScreenTape), al fine di calcolare il volume di ciascun campione da utilizzare per la reazione di cattura, sulla base delle concentrazioni finali ottenute.
Per ciascun campione, le molecole di Y-chr su cui sono localizzati i 31630 SNPs *target* sono state selezionate dal *pool* di frammenti di DNA mediante *target enrichment,* sfruttando 74809 sonde a RNA (myBaits Custom DNA-Seq –Arbor Bioscience, Michigan, USA) estese per 80 bp e con una profondità sul singolo SNPs di 3X, secondo il protocollo fornito dall'azienda produttrice, con alcune modificazioni atte ad ottimizzare il recupero di aDNA, come già descritto nel *Paragrafo 4.2.4.2*, con l'unica ulteriore modifica di una riduzione del numero di cicli di PCR al termine delle due fasi di ibridazione (15 cicli a seguito del primo *round,* 10 cicli al termine del secondo). Per massimizzare la resa del processo di arricchimento, ciascun campione è stato catturato in singolo, al fine di utilizzare l'intero volume di reazione disponibile. Per entrambi i *round* di ibridazione il materiale genetico è stato dapprima bloccato in una configurazione a singolo filamento, grazie all'utilizzo di agenti bloccanti in grado di legarsi alle sequenze universali degli adattatori Illumina, al fine di produrre un ingombro sterico tale da non veicolare il ri-appaiamento delle doppie eliche del DNA. Successivamente al materiale così processato sono state aggiunte le sonde a RNA, oltre ad una mix di reagenti atti a ottimizzare l'ambiente di reazione; la miscela contenente il DNA *target* e le sonde è stata incubata in TC (a 65°C per 24 ore durante il primo *round* di ibridazione, e

a 60°C per 22 ore nel secondo *step* sperimentale). Al termine del periodo di incubazione il materiale genetico ibridato alle sonde è stato purificato dal resto del materiale genomico aspecifico e dai *buffer* di reazione attraverso lavaggi sequenziali portati avanti mediante l'uso di un *rack* magnetico che ha permesso la separazione fisica tra il DNA *target* (adeso a sonde biotinilate) ed il resto dei componenti sfruttando il fenomeno di complementarietà biotina/streptavidina (quest'ultima posizionata a copertura delle biglie magnetiche di lavaggio). Per la reversione del legame tra le sonde ed il DNA ibridato, il materiale purificato è stato eluito in 30 µL di *buffer* TT (10 mM Tris-Cl, 0.05% TWEEN®-20 solution - pH 8.0-8.5) ed è stato incubato in TC a 95°C per 5 minuti, al termine dei quali è stato recuperato il surnatante ed interamente utilizzato per una reazione di PCR volta ad incrementare il numero di molecole catturate.
I prodotti finali sono stati quantificati mediante TapeStation 4150 System (kit High Sensitivity D1000 ScreenTape), e riuniti in un *pool* equimolare per la corsa in *paired end* su piattaforma Novaseq 6000 di Illumina (SP Reagent Kit, 100 cicli), come riportato in precedenza.

5.2.3 Analisi dei dati

5.2.3.1 Analisi dei dati di sequenziamento

Anche in questo caso, in seguito alla produzione dei files FastQ compressi (prodotti con il *software* Illumina CASAVA-1.8.2), è stata utilizzata la pipeline descritta da Peltzer e collaboratori[179] per il processamento delle *reads* grezze ed il loro allineamento sul genoma umano di riferimento, ampiamente dettagliata nel *Paragrafo 4.2.6*. Brevemente, le caratteristiche di qualità dei dati di sequenza sono state valutate in diverse fasi utilizzando FASTQC[183]; sono stati successivamente rimossi gli adattatori, unite le *reads* R1 e R2 di ciascun campione, e scartate le sequenze con lunghezza inferiore alle 30 bp attraverso Clip&merge, impostando una qualità di base minima di 20. Le letture unite sono state mappate rispetto al genoma di riferimento umano hg19, utilizzando la funzione *samse* di BWA[184] ed impostando i parametri standard, ad eccezione del *seeding*, che è stato disabilitato per consentire una maggiore sensibilità nell'allineamento. Le letture così prodotte sono state sottoposte ad una fase di rimozione dei duplicati di PCR mediante DeDup, impostando i parametri standard. Le normali caratteristiche di degradazione del DNA, ed in particolare le misincorporazioni alle estremità dei filamenti processati, sono stati monitorati mediante il calcolo bayesiano effettuato dal pacchetto mapDamage 2.0[185].
Il *file* di allineamento BAM prodotto per ciascun campione è stato gestito mediante il *toolkit* GATK 4.1[180], al fine di eseguire le chiamate aploidi di tutte le posizioni lette sul Y-chr, con le funzioni ed i parametri già discussi in precedenza. I *file* VCF sono stati prodotti in modo da mantenere solamente le posizioni di buona qualità (≥ 30) con una copertura di almeno 3 *reads*.

5.2.3.2 Determinazione degli aplogruppi del Y-chr

Per ciascun campione di sesso maschile sono stati determinati gli hg del Y-chr attraverso 2 metodi indipendenti.

Un primo metodo è stato portato avanti attraverso l'uso del *software* Yleaf[191], in grado di assegnare l'hg del Y-chr attraverso la comparazione delle varianti di ciascun campione con 41560 posizioni filogeneticamente informative (*marker*) interne alle regioni MSY. Sono stati utilizzati i *file* allineati e ripuliti nei passaggi precedenti, e sono stati impostati i parametri di copertura minima della posizione (pari a 2 *reads*) e concordanza di chiamata del polimorfismo superiore al 90%. Il risultato prodotto dal *software,* una tabella Excel contenente le posizioni *marker* chiamate in ciascun campione processato, è stato interpretato ordinando in senso alfabetico la colonna rappresentativa degli hg e filtrando per i soli alleli derivati nella colonna che identifica lo stato di ciascuna lettura. È stato seguito l'elenco dei *marker* derivati letti in ciascun campione ed è stato assegnato l'hg sulla base dell'allele più derivato, che presentava a monte ulteriori marcatori derivati per la stessa linea filogenetica.

Il secondo metodo ha previsto l'importazione dei *file* VCF in un foglio di calcolo Excel, dove, per ogni campione e per ogni posizione, sono stati assegnati, sulla base della corrispondenza con la sequenza di riferimento, il numero 0 (per le posizioni con nucleotide identico al riferimento), il numero 1 (laddove presente l'allele alternativo) e un punto (nel caso di chiamata allelica non effettuata). Sono state successivamente elencate le posizioni polimorfiche e assegnato a ciascun campione il rispettivo genotipo in base all'associazione tra le varianti rispetto a quelle di Hg noti. La posizione filogenetica di ogni SNP è stata stabilita in base alla sua presenza in letteratura o in *database* pubblici[79,85,95,172,236,237].

5.2.3.3 Median Joining Network

Le posizioni condivise tra i campioni che hanno presentato sufficiente copertura per poter superare le analisi (posizioni non lette < 5%), sono state utilizzate per la produzione di un *network* filogenetico basato sull'algoritmo *Median Joining* (MJN) attraverso il software PopART[238]. Tale visualizzazione grafica è utilizzata per valutare le relazioni evolutive tra sequenze nucleotidiche, geni, cromosomi, genomi o specie. In particolare il MJN è un metodo in cui le sequenze intermedie (definite vettori mediani) vengono ricostruite e incluse nella rete finale. Per il calcolo delle distanze tra le sequenze originali, l'algortimo sfrutta il metodo della distanza di Hamming. Il valore ε, parametro che controlla la ricerca di nuovi vettori mediani, è stato impostato a 0 ed il peso delle trasversioni è state considerato 3 volte quello delle transizioni.

5.2.3.4 Analisi popolazionistiche

Per la valutazione della struttura genetica e della variabilità interna alle popolazioni in studio, sono stati stimati alcuni parametri basati sulle diversità nucleotidiche attraverso il software Arlequin 3.5.2.2[239]. In particolare, attraverso la costruzione di un foglio di lavoro, come descritto nel manuale di Arlequin 3.5.2.2, si è proceduto all'analisi di alcuni parametri descrittivi della variabilità intra-popolazionistica, illustrati di seguito.

→ Indici di diversità molecolare:

I. Indici di diversità standard: vengono calcolati in questo modo alcuni dei più frequenti indici di diversità, come il numero di alleli.

   i. Diversità genetica: è definita come la probabilità che due aplotipi scelti a caso nella popolazione, siano differenti. Il valore che ne deriva è pari, o inferiore al numero di campioni[240]. Tale parametro è calcolato come:

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

   Con "n" che rappresenta il numero di copie geniche nel campione, "k" pari al numero di aplotipi, p_i frequenza, all'interno del campione, dell'aplotipo i-esimo.

   ii. Numero di loci utilizzabili: sono i loci che presentano un numero di dati mancanti nelle posizioni nucleotidiche minori, rispetto ad un valore soglia definito (ad esempio 0.05), e che pertanto verranno utilizzati nelle comparazioni a coppie.

   iii. Numero di siti polimorfici (S): numero di loci utilizzabili che presentano più di un allele per locus.

II. Indici molecolari:

   i. Numero di aplotipi: è il parametro che descrive la variabilità interna della popolazione, sulla base del diverso numero di aplotipi osservati, rispetto a quelli attesi.

   ii. Numero medio di differenze a coppie[241,242]: è calcolato con la seguente formula

$$\hat{\pi} = \frac{n}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{k} p_i p_j \hat{d}_{ij}$$

Con d_{ij} che rappresenta la stima del numero di mutazioni che sono intercorse dalla divergenza di due aplotipi i e j, k che racchiude il numero totale di aplotipi, p_i frequenza, all'interno del campione, dell'aplotipo i-esimo e n che rappresenta la dimensione del campione.

iii. Valori Theta: sono parametri popolazionistici definiti come:

$$\theta=2Mu$$

con M pari ad N per popolazioni aploidi ed u che rappresenta il tasso di mutazione della popolazione a livello aplotipico;

- Theta(Hom): è una stima di theta ottenuta sulla base dell'omozigosità osservata, data da:

$$H = \frac{1}{\theta + 1}$$

- Theta(S): viene stimato sulla base dei siti segreganti. Per un DNA non ricombinante è pari a:

$$\theta = \frac{S}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

- Theta(π): viene stimato in rapporto alla media del numero di differenze a coppie. La stima di theta è pari a:

$$E_{(\pi)}=\theta$$

→ Test di neutralità selettiva di un campione: sulla base del modello a siti infiniti, compara Theta(S) e Theta(π):

$$D = \frac{\hat{\theta}_\pi - \hat{\theta}_s}{\sqrt{Var\left(\hat{\theta}_\pi - \hat{\theta}_s\right)}}$$

In una situazione di neutralità, ovvero in assenza di ricombinazione, se il valore è minore di 0, ci si aspetta una popolazione in espansione, mentre se il valore restituito è maggiore di 0 la popolazione è strutturata.

Per la comparazione della variabilità tra le popolazioni italiane dell'età del Ferro è stato utilizzato l'indice Fst, ottenuto mediante Arlequin. Questo indice calcola la distanza genetica come la varianza osservata di una frequenza allelica, divisa per il suo valore massimo possibile[243,244]. Il

risultato è restituito sotto forma di matrice; la distribuzione nulla dei valori di Fst a coppie (sotto l'ipotesi che non ci siano differenze tra le popolazioni) è ottenuta dalla permutazione degli aplotipi tra le popolazioni. Più il valore restituito è vicino allo 0, più due popolazioni risultano geneticamente vicine.

5.2.3.5 Principal Component Analysis

È stata infine prodotta una *Principal Component Analysis* (PCA), attraverso un approccio statistico multivariato che riduce la tridimensionalità dei dati mantenendo la maggior parte della variazione nel set di dati processati[245]. Attraverso questa metodologia, vengono in primo luogo identificate delle direzioni, definite componenti principali (PC), lungo le quali la variazione dei dati è massima. Utilizzando poche componenti, ogni campione può essere rappresentato da relativamente pochi numeri invece che da migliaia di variabili. I campioni possono quindi essere tracciati, rendendo possibile valutare visivamente somiglianze e differenze tra essi, in modo da determinare i migliori raggruppamenti tra gli individui analizzati. La PCA è stata prodotta sulla base delle frequenze dei principali Hg e sub-Hg ottenuti nel set di campioni in esame e confrontati con un *dataset* di campioni europei moderni ed antichi recuperati in letteratura, per un totale di 1735 individui. Questi campioni sono rappresentativi delle frequenze di alcuni tra i più diffusi Hg del Y-chr ad oggi noti, e sono stati suddivisi sia temporalmente (in un arco diacronico compreso tra il Neolitico e l'impero romano a cui sono stati aggiunti i campioni moderni di confronto), che in termini spaziali (Tabella A 3 in Appendice). Inoltre, anche i campioni sequenziati in questo studio, sono stati suddivisi in 3 macro-popolazioni, suddivise su base geografica. In particolare, i campioni provenienti dalle necropoli al di sopra del 42° latitudinale sono state considerate appartenenti al gruppo "centro-nord", mentre i campioni peninsulari recuperati a latitudini inferiori sono stati raggruppati nel "centro-sud". Infine, sono stati mantenuti separati i campioni insulari siciliani. Tale associazione è stata considerata necessaria al fine di evitare la produzione di *bias* legati ad una bassa numerosità di campioni disponibili per ciascun sito analizzato, trattandosi di analisi prodotte sulla base di frequenze e non di sequenze.

I dati, organizzati in un foglio di lavoro Excel, sono stati dapprima analizzati attraverso *FactoMineR*[246], un pacchetto di R (R-DevelopmentCoreTeam 2008) che presenta metodi esplorativi di analisi per riassumere, visualizzare e descrivere i *dataset,* e successivamente è stata eseguita la ricostruzione grafica del risultato attraverso il pacchetto R *ggplot2*[247].

5.3 Risultati e discussione

5.3.1 Quantificazione delle librerie

In seguito all'indicizzazione delle molecole, spiegata nel *Paragrafo* 5.2.2.3, è stata eseguita una prima valutazione della concentrazione e qualità del DNA convertito in libreria in ciascun campione, mediante TapeStation 4150 System (kit High Sensitivity D1000). Per semplicità in Figura 29 sono stati riportati i profili di due campioni, rappresentativi dell'andamento di tutte le librerie costruite.

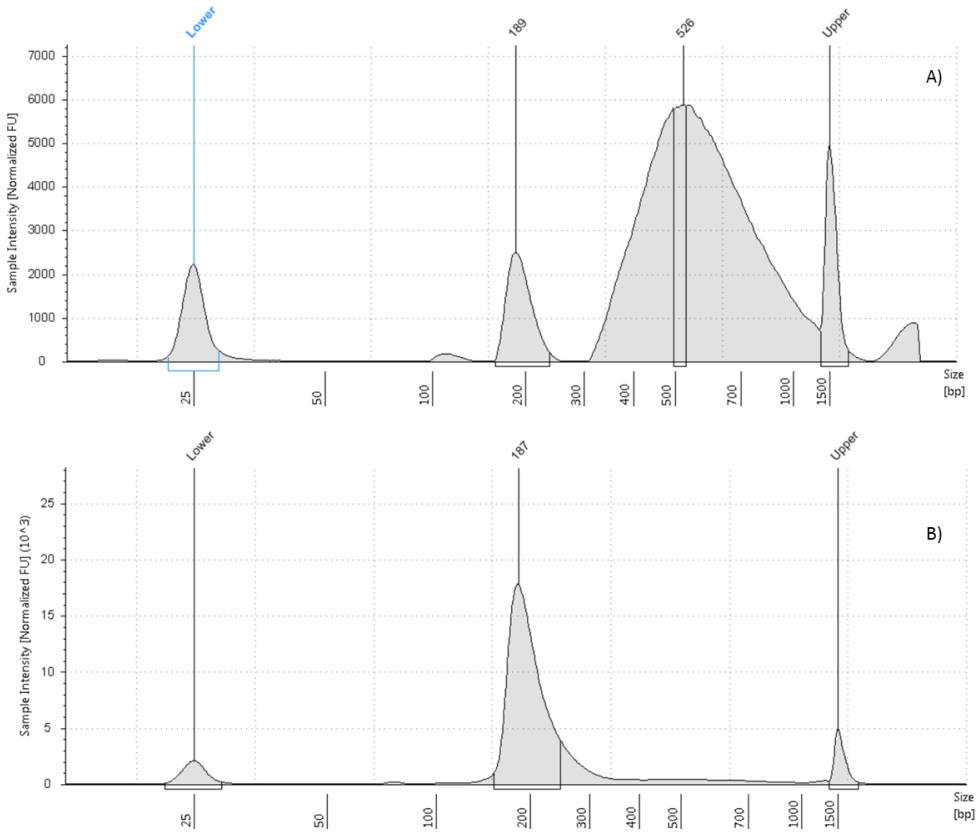

Figura 29: Risultati dell'elettroforesi quantitativa delle librerie di due campioni. A) Campione con evidenti aspecifici; B) Campione pronto per il sequenziamento Illumina.

In generale, come atteso per molecole di aDNA, in tutti i campioni è possibile visualizzare un picco di intensità del segnale di fluorescenza in corrispondenza di frammenti di dimensioni poco superiori alle 180 bp (lunghezza data dalla somma tra gli adattatori indicizzati, circa 130 bp, e l'inserto, mediamente lungo 50 bp). Nelle librerie prodotte è stato possibile osservare un andamento duplice; in alcuni campioni (rappresentati dal grafico mostrato in Figura 29 – A), a causa di uno squilibrio tra il numero di cicli di arricchimento effettuati, la concentrazione del DNA genomico iniziale, ed i reagenti, sono stati generati ponti ad idrogeno tra le porzioni complementari

degli adattatori di due o più molecole presenti in soluzione, con la conseguente formazione di molecole aspecifiche, eteroduplici, visualizzabili come picchi che in questo caso si estendono dalle 300 alle oltre 1000 bp. Inoltre, è stato possibile osservare un ulteriore picco a sinistra di quello che definisce il campione (nell'immagine sopra, attorno a 100 bp) che rispecchia la formazione di dimeri di adattatori, i quali non hanno incorporato alcun frammento di DNA (verosimilmente a causa di una concentrazione troppo esigua delle molecole biologiche) durante la produzione della libreria.

Nella maggior parte dei campioni invece, è stato possibile osservare un unico profilo, rappresentato da una gaussiana con distribuzione massima dei frammenti compresa tra 180 bp e 190 bp in cui sono presenti tutte le molecole di DNA che hanno correttamente incorporato gli adattatori e gli indici (Figura 29 – B) e che pertanto non hanno necessitato di ulteriori *step* prima del sequenziamento *shotgun*.

Per eliminare gli aspecifici dalle librerie che li presentavano, è stata allestita una reazione di PCR di un unico ciclo (*onecycle*), e la quantità e qualità dei frammenti è stata valutata nuovamente con TapeStation 4150 System (kit High Sensitivity D1000), ottenendo il risultato mostrato in Figura 30. Per un miglior confronto è stato riportato il medesimo campione già descritto precedentemente, e rappresentativo dei profili di tutti i campioni così processati.

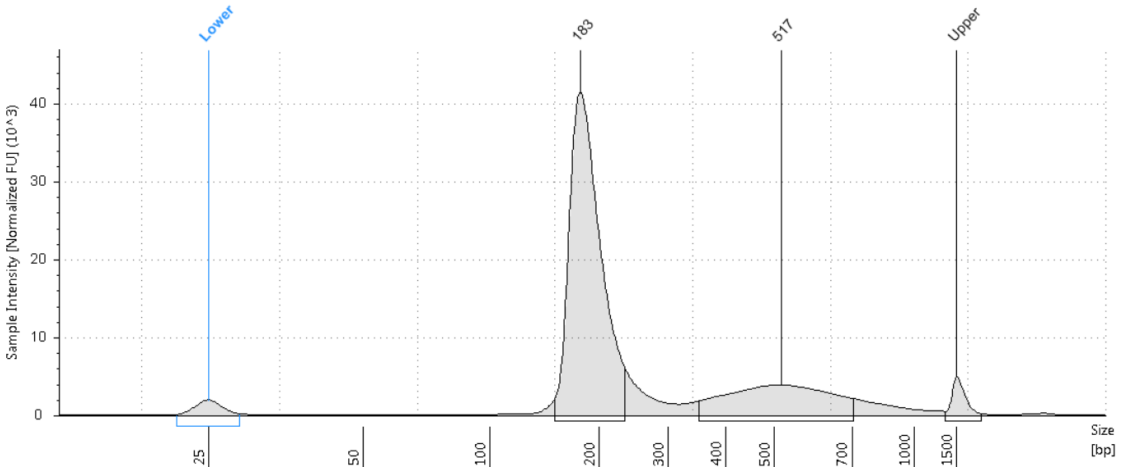

Figura 30: Profilo dei campioni sottoposti a onecycle a seguito della reazione di indexing.

Si può apprezzare da questo profilo l'effettiva eliminazione di gran parte delle molecole aspecifiche, e l'aumento della concentrazione dei frammenti *target*, indicata dall'incremento della fluorescenza misurata a ~180 bp.

Le librerie dei 78 campioni per i quali non si avevano informazioni genetiche relative al sesso (Tabella A 1), sono stati pertanto riuniti in un *pool* equimolare per il sequenziamento *shotgun* su piattaforma Novaseq 6000 di Illumina (kit SP, 100 cicli).

5.3.2 Analisi dei dati prodotti con sequenziamento *shotgun* e determinazione del sesso

Il sequenziamento *shotgun* dei 78 campioni processati ha prodotto un totale di 989'352'372 *reads*. Le principali statistiche di qualità delle librerie sono mostrate in Tabella A 4 - Appendice. Il numero maggiore di *reads* è stato ottenuto per il campione Gubbio-15 (20'890'770), mentre quello inferiore è stato osservato nel campione C_t40 (4'778'584). Per inciso, questo campione aveva già mostrato bassi livelli di DNA convertito in libreria genomica nelle fasi di quantificazione, ed ha subìto 2 differenziali processi di produzione della libreria, che egualmente hanno prodotto scarsi risultati, sebbene dalla quantificazione del materiale estratto era stata notata una elevata concentrazione di materiale biologico. Verosimilmente la presenza di una quantità elevata di inibitori ha determinato un arresto significativo delle fasi sperimentali successive all'estrazione del DNA.

In media, più del 70% delle *reads* ottenute per ciascun campione supera la fase di *merging* (percentuali comprese tra il 52.4% di Gubbio-11 e 83.9% di MAT-63) ad indicazione che la maggior parte delle *reads* è caratterizzata da una lunghezza media paragonabile alle dimensioni tipiche dei frammenti di aDNA. Tra le *reads* scartate vi sono quelle con *overlap* tra R1 e R2 minore di 10 bp, e le *reads* con lunghezza inferiore alle 30bp.

È stata in aggiunta valutata la percentuale di *reads "*mappanti" sul genoma di riferimento (hg19) rispetto al numero totali di *reads* che superano le precedenti fasi di filtraggio. In questo caso i valori ottenuti sono molto variabili nei campioni analizzati e si osservano percentuali comprese tra lo 0.031% di Osa-378 e l'87.437% di Pol-2. Per i campioni antichi si tratta di una condizione piuttosto frequente, tuttavia anomala per il sito di Osteria dell'Osa (Lazio), nel quale tutti i campioni hanno mostrato percentuali di *reads* allineate sulla sequenza di riferimento, mai superiori allo 0.1%. Tali individui, già in fase di campionamento avevano mostrato importanti caratteristiche di degradazione e frammentarietà, oltreché tracce di combustione. Sebbene si tratti di rocche petrose, distretti elettivi per l'analisi dell'aDNA per la loro capacità di preservare al meglio gli acidi nucleici, tutte le caratteristiche sopra elencate sono responsabili di un significativo danno alle molecole già degradate ed in basso numero di copie. Inoltre, dal momento che le statistiche ottenute prima del "mappaggio" delle *reads* sul genoma umano sono in linea con quelle del resto dei campioni processati, è da ipotizzare un chiaro segnale di contaminazione da parte di sorgenti esogene di varia natura, motivo per il quale la maggior parte delle *reads* processate si allineano su genomi di riferimento diversi da quello di interesse. Le stesse caratteristiche, ovvero un numero elevato di sequenze generate, con buoni filtri di qualità e un'alta percentuale di *merging*, ma una perdita complessiva molto elevata di *reads* durante l'allineamento sul genoma di riferimento umano, sono state riscontrate anche nei campioni evidenziati in rosso in Tabella A 4 - Appendice.

In ultima analisi, è stato valutato il numero di duplicati di PCR rimossi attraverso il *software DeDup.* I valori ottenuti, insieme a quelli rappresentanti il *cluster factor* evidenziano una buona complessità di tutte le librerie processate.

Il valore percentuale di DNA endogeno riscontrato nelle librerie sequenziate, ad esclusione di quelle che avevano prodotto dati di allineamento sull'hg19 molto bassi, risulta estremamente alto, ed in linea con le aspettative, trattandosi di materiale ottenuto a partire da rocche petrose. Il valore più alto di DNA endogeno (87.437%) è stato ottenuto per il campione Pol-2, mentre se si escludono i campioni sopra citati, il peggiore valore (3.092%) è stato riportato per C_t63.

Al fine di autenticare l'antichità del dato prodotto, sono state valutate le principali statistiche recuperate mediante mapDamage e mostrate in Tabella A 5 - Appendice. Queste rappresentano i principali parametri che caratterizzano le molecole antiche, ovvero la dimensione media dei frammenti ed i *patterns* di degradazione. Inoltre è stato valutato il rapporto tra il numero di *reads* allineate sul mitocondrio, ed il numero totale di *reads* che sono state "mappate" sul nuDNA. Tale parametro influenza fortemente l'accuratezza dell'estrapolazione dei livelli di contaminazione; rapporti ottimali, normalmente inferiori a 200, sono spesso associati ad alte percentuali di DNA endogeno[248]. I valori ottenuti (compresi tra 0 e 220.14) sono tipici, se si escludono quelli prodotti dai campioni che già avevano prodotto cattive statistiche di sequenziamento, di quanto si evidenzia nelle parti più dense delle ossa petrose, in cui dalla bassa attività metabolica *ante-mortem* deriva una scarsa concentrazione di molecole di mtDNA.

Nei campioni antichi i normali processi di degradazione sono evidenziabili, alle terminazioni dei frammenti, da un alto livello di sostituzioni C→T all'estremità 5' che si traducono in transizioni G→A nel filamento complementare. Nella Tabella A 5 – Appendice si possono apprezzare le percentuali di misincorporazioni ottenute al primo nucleotide dell'estremità 5' e 3'. Le prime ricadono in un *range* che va dallo 0.04 di Osa-378 allo 0.165 di Osa-417, mentre le transizioni G→A sono comprese tra lo 0.044 (OPACO_Tb.101) e lo 0.179 (Osa-417). Ad entrambe le estremità è possibile quindi osservare un'incidenza delle degradazioni *post-mortem* compatibile con l'antichità dei nostri campioni, tenendo di conto che il tipo di processamento effettuato durante la produzione della libreria genomica è responsabile della riparazione di buona parte delle misincorporazioni presenti nelle molecole antiche (come descritto nel Paragrafo 5.2.2.3). La lunghezza media dei frammenti processati varia da 38 bp a 51.61 bp e non sono stati evidenziati livelli significativi di contaminazione da DNA umano moderno.

Per una migliore valutazione dei risultati di autenticazione, in Figura 31 è mostrato il *plot* generato da *mapDamage* in seguito alla produzione del dato. Le parentesi grigie nei mini-plot rappresentano il punto iniziale della molecola frammentata, mentre le posizioni a sinistra e a destra delle parentesi rappresentano il confronto con la sequenza di riferimento. Sono rappresentate separatamente le frequenze di A, C, T e G per i primi 10 nucleotidi. Queste mostrano un aumento della frequenza di T all'estremità 5', concordante con una diminuzione di purine, ed al contrario, un aumento molto

consistente di A all'estremità 3' del filamento. Il pannello sottostante mostra la frequenza di T (in rosso) e di A (in blu) per i primi 25 nucleotidi ad entrambe le estremità del filamento. Anche qui è possibile valutare l'aumento della frequenza di T al 5' e di A al 3', concordante con le modifiche subìte dall'aDNA.

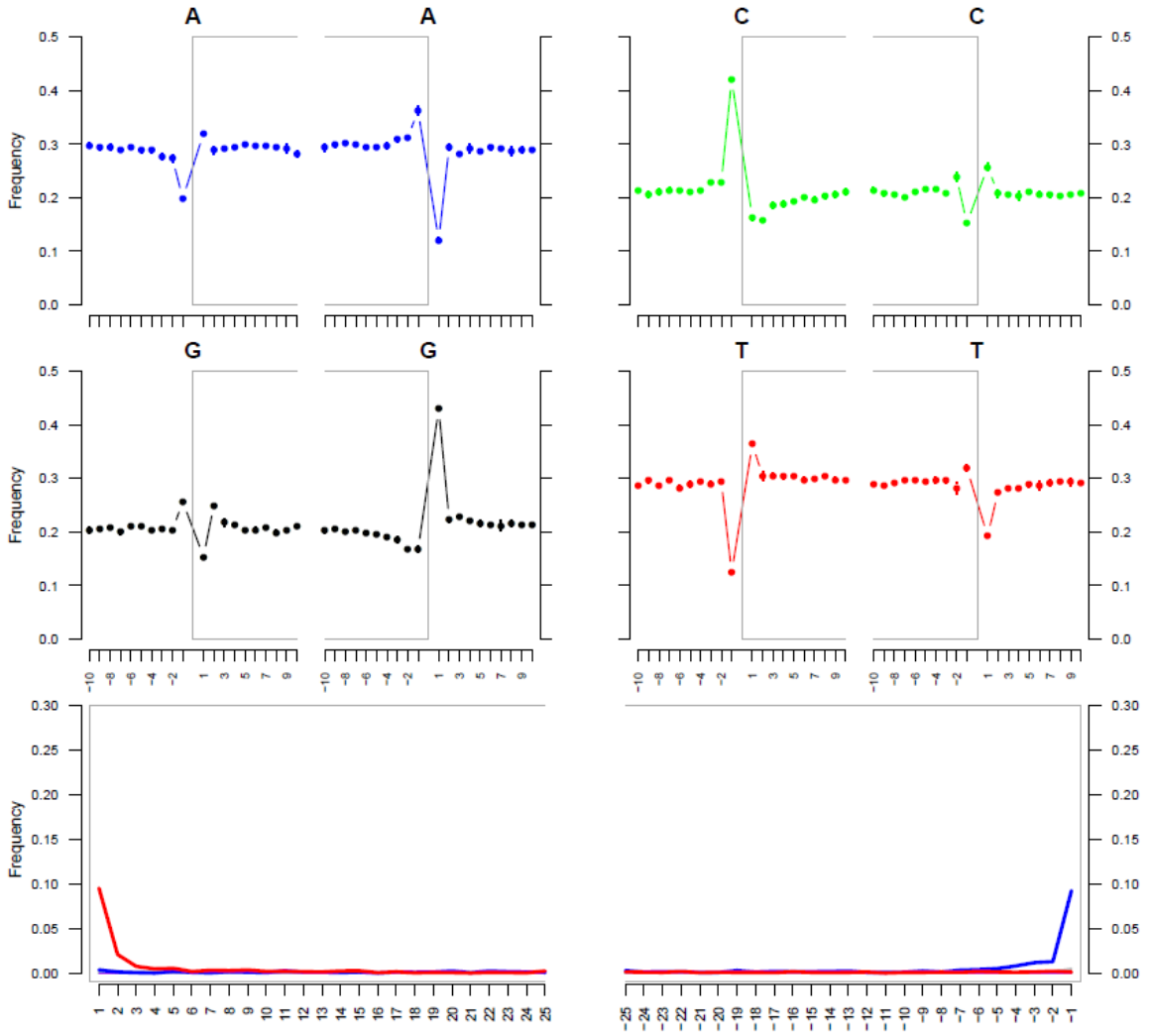

Figura 31: Plot rappresentante le modifiche alle estremità 5' e 3' della molecola. I quattro mini-grafici superiori mostrano la frequenza di ciascuna base all'esterno e all'interno della read (indicata dalla casella grigia). I grafici in basso mostrano la frequenza di sostituzioni C→T (in rosso) e G→A (in blu).

I *plot* inferiori mettono in evidenza molto chiaramente l'effetto ottenuto con l'utilizzo degli enzimi UDG ed Endonucleasi VIII, i quali, agendo, hanno determinato una riparazione pressoché totale delle misincorporazioni interne ai frammenti di DNA. L'interruzione del processo con UGI prima della fine del trattamento ha permesso di mantenere inalterati i chiari segni di danneggiamento nelle primissime posizioni del filamento.

Infine, attraverso la *pipeline* prodotta da Skoglund et al.[235] si è proceduto alla determinazione del sesso dei campioni, attraverso la valutazione del rapporto tra le *reads* allineate sul Y-chr e quelle presenti su entrambi i cromosomi sessuali. Lo *script* funziona in modo ottimale con ~100000 *reads*

di partenza, ma riesce ad assegnare un dato robusto anche con un numero inferiore di un ordine di grandezza. Tutti i 78 campioni processati sono stati valutati per avere un'informazione complessiva del dato ed i risultati ottenuti sono mostrati in Tabella A 6 – Appendice. Sono stati considerati coerenti con un genotipo maschile i valori di Ry superiori a 0.075, come suggerito dagli autori dello *script*[235].

Per il campione OPACO_Tb.50 non è stato possibile lanciare lo *script*; inoltre, anche per altri 6 campioni (in rosso in Tabella A 6) non è stato possibile ottenere assegnazioni relativamente al sesso. Per 11 campioni (2 del sito di Norcia, 1 di Gubbio, 3 di Matelica, 3 di Castiglione, 1 di Osteria dell'Osa ed 1 di Polizzello) è stato possibile solamente ipotizzare l'appartenenza al sesso maschile ("consistent with XY but not XX"): questi individui sono rappresentati da un numero molto basso di *reads* iniziali, o da valori di Ry al limite che pertanto non permettono di assegnare con assoluta certezza il sesso maschile. Infine, 25 campioni (3 del sito di Norcia, 5 di Gubbio, 9 recuperati a Matelica, 5 di Castiglione e 3 appartenenti al sito archeologico di Polizzello) sono stati assegnati con certezza al sesso maschile.

5.3.3 Cattura del Y-chr e analisi di sequenza

Sono stati catturati un totale di 42 campioni geneticamente maschi, oltre a 2 femmine utilizzate come controllo negativo. Questi sono stati selezionati sulla base delle migliori statistiche ottenute in seguito a sequenziamento *shotgun.* In particolare sono stati valutati i parametri relativi alla percentuale di DNA endogeno e all'autenticità del dato. È stata infine prestata attenzione alla selezione di un numero rappresentativo e simile di campioni per ciascun sito analizzato, al fine di avere una numerosità confrontabile. La Tabella 13 riassume i campioni utilizzati per le analisi sul Y-chr e le principali statistiche ottenute dal precedente sequenziamento, oltre a quelle dei campioni di Montericco, Ceretolo e Polizzello analizzati per progetti precedenti (dati non pubblicati).

Sono stati selezionati 7 campioni per i siti di Montericco e Matelica, 6 per Ceretolo, Castiglione e Polizzello, 5 per Gubbio e Norcia. In particolare per quest'ultimo sito sono stati selezionati tutti i campioni di sesso maschile, compreso ANN-16, sebbene la percentuale di DNA endogeno abbia mostrato un valore significativamente inferiore agli altri, ma superiore rispetto alla soglia minima normalmente imposta dello 0.1%.

Infine, i due campioni di sesso femminile selezionati per essere usati come controlli negativi, sono stati scelti in maniera aleatoria tra i campioni che avevano mostrato buone statistiche di sequenziamento *shotgun.*

Sito	Regione	IDLab	% DNA endogeno	MT/NUC	Deaminazione 3'	Deaminazione 5'
Montericco (Imola)	Emilia-Romagna	MONT-6	40.731	159.83	0.187	0.179
	Emilia-Romagna	MONT-7	53.478	130.46	0.155	0.156
	Emilia-Romagna	MONT-09	38.18	159.83	0.12	0.119
	Emilia-Romagna	MONT-15	18.976	109.96	0.085	0.086
	Emilia-Romagna	MONT-29	9.861	162.78	0.095	0.095
	Emilia-Romagna	MONT-31	44.065	95.46	0.095	0.094
	Emilia-Romagna	MONT-41	30.903	97.62	0.121	0.123
Ceretolo (Casalecchio di Reno)	Emilia-Romagna	CER-12	18.205	77.36	0.093	0.088
	Emilia-Romagna	CER-13	27.194	108.93	0.093	0.091
	Emilia-Romagna	CER-16	58.801	81.72	0.103	0.103
	Emilia-Romagna	CER-76B	20.88	104.66	0.124	0.124
	Emilia-Romagna	CER-78	87.437	70.47	0.065	0.067
	Emilia-Romagna	CER-96	74.978	73.61	0.091	0.086
Norcia (Colle dell'Annunziata)	Umbria	ANN-16	0.741	77.95	0.092	0.095
	Umbria	ANN-32	66.09	106.79	0.103	0.102
Norcia (Campo Boario)	Umbria	CB-14	15.653	66.74	0.11	0.111
	Umbria	CB-15	71.65	85.96	0.097	0.098
	Umbria	CB-20	67.145	89.32	0.097	0.097
Gubbio (San Biagio)	Umbria	Gubbio-1	38.18	109.96	0.074	0.074
	Umbria	Gubbio-3	56.27	122.33	0.105	0.101
	Umbria	Gubbio-7	44.065	208.84	0.081	0.082
	Umbria	Gubbio-8	26.353	103.88	0.105	0.105
	Umbria	Gubbio-14	10.481	102.13	0.087	0.086
Matelica (Località Crocifisso)	Marche	MAT-50	18.809	116.3	0.067	0.069
	Marche	MAT-52	31.49	100.71	0.144	0.136
	Marche	MAT-54	33.029	74.89	0.091	0.091
	Marche	MAT-58	17.754	50.08	0.073	0.075
	Marche	MAT-63	20.381	116.87	0.099	0.101
	Marche	MAT-72	10.202	157.47	0.074	0.071
	Marche	MAT-75	16.274	99.67	0.081	0.082
Castiglione	Lazio	C_IJ14	9.859	1.84	0.099	0.105
	Lazio	C_t30	11.717	0.82	0.132	0.136
	Lazio	C_t48	10.228	4.36	0.073	0.077
	Lazio	C_t71	10.909	0	0.116	0.119
	Lazio	C_t79	20.88	0.44	0.063	0.068
	Lazio	C_t83	8.742	36.8	0.084	0.087
Polizzello	Sicilia	Pol-12	64.655	83.36	0.109	0.109
	Sicilia	Pol-13	77.56	70.52	0.096	0.099
	Sicilia	Pol-15	9.765	113.72	0.121	0.122
	Sicilia	Pol-23	43.95	71.23	0.071	0.072
	Sicilia	Pol-24	11.541	73.08	0.097	0.094
	Sicilia	Pol-9	74.978	88.65	0.103	0.104
K- (F)	Umbria	SCOL-57	34.453	98.34	0.051	0.051
	Lazio	C_t63	2.185	0.15	0.016	0.014

Tabella 13: Campioni di sesso maschile selezionati per la cattura del Y-chr e statistiche di qualità ed autenticità del dato prodotto per ciascun campione attraverso sequenziamento shotgun.

Come già detto sono state eseguite catture in singolo per ciascun campione processato, come descritto nel Paragrafo 5.2.2.5, ed i prodotti finali sono stati quantificati mediante TapeStation 4150 System (kit High Sensitivity D1000), al fine di avere un controllo qualitativo oltreché quantitativo prima della riunione dei campioni in *pool* per il sequenziamento su piattaforma Novaseq6000 di Illumina. Il profilo in Figura 32 è un esempio caratteristico di ciò che è emerso per tutti i catturati quantificati. Per una migliore comparazione, è stato riportato il profilo del campione già descritto nei grafici precedenti.

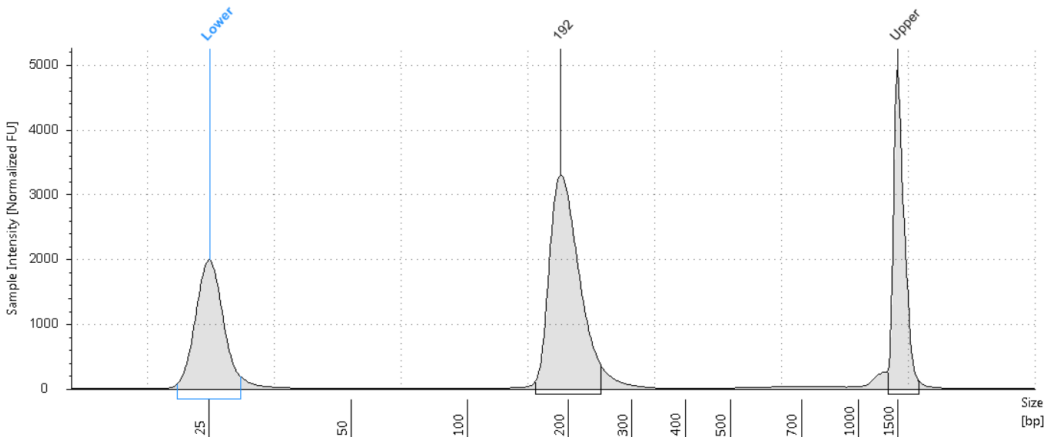

Figura 32: Profilo finale misurato alla TapeStation in seguito a cattura del Y-chr.

In generale, per tutti i campioni è stato ottenuto un profilo molto pulito, rappresentato da una distribuzione a gaussiana, con ampiezza ristretta, delle specie molecolari catturate, ed un picco massimo compreso tra le 180 bp e le 200 bp, lunghezza caratteristica dell'aDNA. Tutti i campioni sono pertanto stati riuniti in un *pool* in concentrazione equimolare, per il sequenziamento in *paired end* su piattaforma Novaseq6000 di Illumina (SP Reagent Kit, 100 cicli).

La Tabella 14 descrive le principali statistiche di sequenziamento prodotte per i 44 campioni processati. Sono state generate un totale di 434'890'988 *reads,* con un minimo di 490222 (per il controllo negativo C-t63) ed un massimo di 15'922'314 (CB-20). Per tutti i campioni, ad eccezione dei controlli negativi, il numero di *reads* processate è in linea con le aspettative (*output* previsto 6 milioni di *reads* a campione, dopo l'unione delle letture R1 ed R2). La percentuale di *reads* recuperate alla fine del *Clip&Merge* è compresa in un *range* tra l'87.52% (C_t30) e il 49.44% (Gubbio-1); per quasi tutti i campioni (ad eccezione di MONT-15 e MONT-41) è stato possibile recuperare un numero molto alto di sequenze allineate al genoma di riferimento umano hg19, a seguito di tutti i passaggi di filtraggio relativi sia alla qualità del dato, che all'eliminazione di molecole clonali. Anche le percentuali di DNA endogeno rispecchiano quanto appena detto, e risultano superiori al 90% per i ¾ dei campioni processati. Infine, i parametri di autenticazione del dato, ed in particolare il livello di deaminazioni alle estremità delle molecole, risulta in linea con i valori ottenibili per campioni antichi a seguito del trattamento delle librerie con UDG, e risulta confrontabile con quanto ottenuto a seguito del sequenziamento *shotgun.*

Sito	IDLab	# reads processate	% Merged	# Reads post-RmDup	% DNA endogeno	Deaminazione 3'	Deaminazione 5'
Montericco	MONT-6	10020106	77.86	76026	81.03	0.1777	0.1686
	MONT-7	9669438	76.98	535905	94.738	0.1436	0.1414
	MONT-09	8723728	75.61	580736	95.174	0.1162	0.1172
	MONT-15	12355532	58.87	8609	20.1	0.0726	0.0783
	MONT-29	10295926	66.4	140072	65.342	0.0877	0.0877
	MONT-31	10077694	66.48	1091203	95.942	0.1086	0.1057
	MONT-41	9479626	64.47	39551	54.208	0.1168	0.117
Ceretolo	CER-12	6034630	72.51	577180	95.063	0.0864	0.0872
	CER-13	9160120	67.96	942512	95.687	0.0737	0.0719
	CER-16	10262918	67.49	866919	89.149	0.108	0.1062
	CER-76B	9312118	71.31	1043158	93.488	0.1136	0.1129
	CER-78	11273540	61.04	1316930	94.034	0.0706	0.0697
	CER-96	10760284	58.67	729470	95.597	0.0837	0.0805
Norcia	ANN-16	7446696	78.07	100730	57.704	0.0813	0.0774
	ANN-32	12492492	68.12	1292428	94.543	0.0997	0.0987
	CB-14	11554018	60.59	584470	95.239	0.1064	0.1068
	CB-15	9787240	70.02	1100485	95.662	0.0929	0.0913
	CB-20	15922314	67.61	2674381	83.355	0.0921	0.0917
Gubbio	Gubbio-1	7803108	49.44	457670	94.696	0.067	0.0652
	Gubbio-3	11042642	58.48	1011420	94.787	0.0957	0.0908
	Gubbio-7	10638312	55.25	839101	93.193	0.0784	0.0751
	Gubbio-8	12323262	73.21	1361045	89.337	0.0993	0.0983
	Gubbio-14	11787376	62.78	741847	92.452	0.0821	0.0801
Matelica	MAT-50	7617290	62.99	457976	94.202	0.0653	0.0619
	MAT-52	11425040	66.11	824700	92.492	0.1241	0.1209
	MAT-54	8076890	70.75	804437	93.705	0.0868	0.0821
	MAT-58	12349544	54.04	547868	91.434	0.0714	0.0686
	MAT-63	9342372	69.3	279460	95.637	0.0895	0.088
	MAT-72	9533602	59.59	637131	82.657	0.0701	0.0662
	MAT-75	8267784	65.67	580736	90.202	0.0748	0.0713
Castiglione	C_IJ14	12276172	72.74	715643	93.195	0.0981	0.0993
	C_t30	9536450	87.52	450551	93.183	0.1277	0.1269
	C_t48	11627424	73.76	1151546	87.654	0.0733	0.0728
	C_t71	9714664	83.83	451850	95.505	0.1121	0.11
	C_t79	10519208	65.97	984706	95.101	0.0677	0.0665
	C_t83	8682892	78.08	609742	91.121	0.0825	0.0812
Polizzello	Pol-12	6591904	65.06	1054173	94.272	0.0875	0.0823
	Pol-13	9647294	56.33	1005900	93.66	0.0846	0.082
	Pol-15	12253050	60.32	1605879	95.072	0.0942	0.0942
	Pol-23	9011306	70.56	715296	95.924	0.0658	0.0655
	Pol-24	14895840	52.9	1494260	93.235	0.0636	0.0615
	Pol-9	8631088	71.4	1498178	95.785	0.0994	0.0984
Norcia	SCOL-57	6177832	61.86	622691	90.227	0.0498	0.0487
Castiglione	C_t63	490222	82.88	31664	66.645	0.1389	0.1273

Tabella 14: Statistiche di sequenziamento dei 44 campioni catturati per il Y-chr. # reads processate: reads grezze iniziali; %Merged: numero di reads che superano il Clip&Merge rispetto al totale; # Reads post-RmDup: numero di reads allineate alla sequenza di riferimento umana dopo l'eliminazione dei duplicati di PCR.

5.3.4 Determinazione degli aplogruppi del Y-chr

Per tutti i campioni sono stati ricostruiti gli Hg del Y-chr attraverso 2 metodi indipendenti, descritti nel Paragrafo 5.2.3.2.

I *file* BAM prodotti al termine dei processi di filtraggio, sono stati utilizzati come *input* per il software Yleaf[191], attraverso il quale sono state comparate le varianti di ciascun campione con 41560 posizioni filogeneticamente informative interne alle regioni MSY, utili per l'attribuzione degli Hg.

Inoltre, i medesimi file BAM sono stati convertiti, attraverso il *toolkit* GATK 4.1[180] in file VCF genomici, e da essi è stata recuperata l'informazione di chiamata sul solo Y-chr. Le posizioni totali ottenute per ciascun campione sono state controllate in un foglio di calcolo Excel, per l'assegnazione del genotipo a ciascun campione, sulla base dell'associazione delle varianti di ogni individuo a quelle di Hg noti.

I risultati ottenuti da Yleaf, e confermati con il secondo metodo, sono mostrati in Tabella 15.

Sito	*IDLab*	*%reads* su Y-chr	*#marker identificati*	*Hg*
Montericco	MONT-6	14.61	623	-
	MONT-7	22.98	9009	G2a2b2a(1a1b1)
	MONT-09	18.25	7963	R1b1a1a2a1a
	MONT-15	15.03	46	-
	MONT-29	6.87	455	-
	MONT-31	42.82	12860	R1b1a1a2
	MONT-41	11.46	116	-
Ceretolo	CER-12	29.34	10199	R1b1a1a2a1a2~
	CER-13	39.28	11966	R1b1a1a2a1a
	CER-16	16.04	9655	R1b1a1a2a1a
	CER-76B	30.78	11618	G2a2b2a1
	CER-78	22.98	11808	R1b1a1a2a1a2
	CER-96	29.69	11588	R1b1a1a2a1a1c1a
Norcia	ANN-16	16.65	1372	R1
	ANN-32	31.94	13005	R1b1a1a2a1a2b2b~
	CB-14	25.18	11185	I2a2a1
	CB-15	40.54	13120	R1b1a1a2a1a2
	CB-20	15.93	12302	R1b1a1a2a1a2
Gubbio	Gubbio-1	25.33	9717	R1b1a1a2
	Gubbio-3	29.76	12421	R1b1a1a2a1a
	Gubbio-7	20.37	11262	R1b1a1a2a1a2
	Gubbio-8	16.99	11051	R1b1a1a2a1a2b2b~
	Gubbio-14	21	10590	R1b1a1a2a1a2
Matelica	MAT-50	24.02	9303	J
	MAT-52	22.79	10285	J
	MAT-54	25.73	10697	J
	MAT-58	12.9	7222	J
	MAT-63	31.19	11929	J
	MAT-72	9.74	6353	R1b1a1a2a1a

Sito	*IDLab*	*%reads* su Y-chr	*#marker identificati*	*Hg*
	MAT-75	10.76	3021	J
Castiglione	C_IJ14	20.27	10105	R1b1a1a2
	C_t30	16.82	7387	R1b1a1a2b
	C_t48	15.03	10461	R1b1a1a2a1a2
	C_t71	22.54	8286	R1b1a1a2
	C_t79	37.15	12439	R1b1a1a2a1a2
	C_t83	18.27	8417	R1b1a1a2
Polizzello	Pol-12	31.37	11994	G2a2b2a1a1c1a1b
	Pol-13	50.57	13722	G2a2b2a1a1c1a1b
	Pol-15	30.49	10933	G2a2b2a1a1c1a
	Pol-23	21.34	13045	G2a2b2a1a1c1a1b2
	Pol-24	48.44	13322	G2a2b2a1a1c1a
	Pol-9	52.38	12237	G2a2b2a1a1c1a
Norcia	SCOL-57	0.02	2	-
Castiglione	C_t63	0.03	0	-

Tabella 15: Assegnazione degli Hg a ciascun campione analizzato. Nelle ultime 2 righe sono presenti i campioni femminili usati come controllo.

In totale è stato possibile attribuire l'Hg a 38 dei 42 campioni maschi analizzati. I quattro rimanenti, tutti del sito di Montericco, non hanno presentato sufficienti marcatori per l'assegnazione dell'Hg del Y-chr. Tutti i campioni di sesso maschile hanno evidenziato una percentuale molto alta di *reads* mappate sul Y-chr rispetto al totale delle sequenze allineate sul genoma umano (compresa tra 6.87% di MONT-29 e 52.38% di Pol-9).

Inoltre, i campioni di sesso femminile, come atteso, hanno mostrato una percentuale di *reads* allineate sul Y-chr rispetto al genoma di riferimento inferiore di più di due ordini di grandezza rispetto ai campioni di sesso maschile (0.02% e 0.03%) e, in modo concordante, non hanno presentato marcatori validi sul cromosoma di interesse, ad eccezione di 2 per SCOL-57.

Nei 38 campioni per cui è stata possibile l'attribuzione ad un Hg, sono state identificate 13 diverse linee filogenetiche, la cui frequenza è riportata in Figura 33.

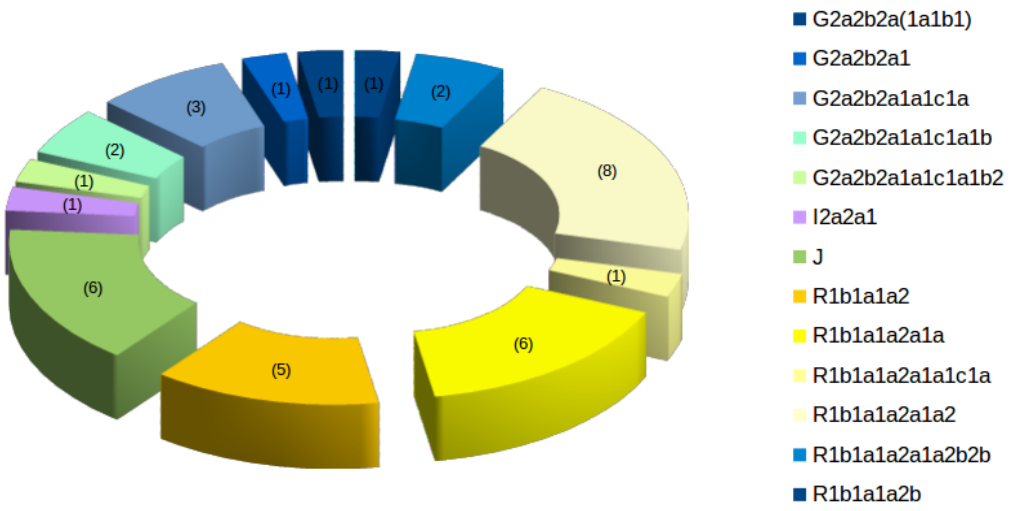

Figura 33: Frequenze degli Hg assegnati ai campioni dell'età del Ferro.

L'Hg R è il più frequente nella popolazione in esame (60.53%); in particolare, 22 dei 23 campioni (2 provenienti dal sito cimiteriale di Montericco, 5 da quello di Ceretolo, 3 da Norcia, 5 da Gubbio, 1 da Matelica ed i restanti 6 dal sito di Castiglione) racchiusi in questa linea, appartengono al sub-Hg R1b, mentre nessun campione rappresenta la variabilità degli Hg R1a e R2. Il campione ANN-16, a causa di una scarsità di marcatori, può essere ascritto solamente all'Hg R1.

L'Hg J (con una frequenza del 15.8%) è stato associato alla sola popolazione di Matelica dove è stato attribuito a 6 dei 7 individui processati. Nonostante l'elevata presenza di marcatori identificati, perfettamente in linea con quanto osservato per i campioni delle altre necropoli, per tali individui non è stato possibile assegnare con certezza l'appartenenza ad uno dei principali sub-cladi, J1 o J2. Seppur le posizioni *marker* per l'assegnazione delle linee appena citate sono presenti all'interno di quelle bersaglio delle sonde costruite, non è stato possibile, per nessuno dei campioni, collezionare informazioni sufficientemente attendibili per poter scendere maggiormente nel dettaglio dell'assegnazione, putativamente a causa di una reale mancanza di molecole nelle regioni di interesse. Nei campioni MAT-52, MAT-54, MAT-58, MAT-63 e MAT-75 è stata riscontrata la presenza delle posizioni chiave per l'assegnazione del sub-Hg J2b, tuttavia, l'assenza dei marcatori per il precedente ramo, J2, e l'esigua copertura osservata nelle posizioni derivate appartenenti alla linea filogenetica più profonda, hanno fatto propendere per la scelta più conservativa di assegnare questi campioni al solo macro-aplogruppo J. Per quanto riguarda l'individuo MAT-50, la presenza dell'allele derivato in posizione M267 permetterebbe di assegnare il campione al clade J1; analogamente al discorso precedente tuttavia, tale posizione non risulta sufficientemente coperta per poter inferire con certezza il dato, pertanto è stato scelto di assegnare anche a questo campione il solo macro-aplogruppo J.

Il restante 23.67% di variabilità del Y-chr è stato assegnato agli Hg G2a (21.05%) ed I2a (2.62%), quest'ultimo individuato in un unico campione proveniente dal sito di Campo Boario, a Norcia

(CB-14). La distribuzione geografica degli Hg lungo i siti italiani dell'età del Ferro è mostrata in Figura 34.

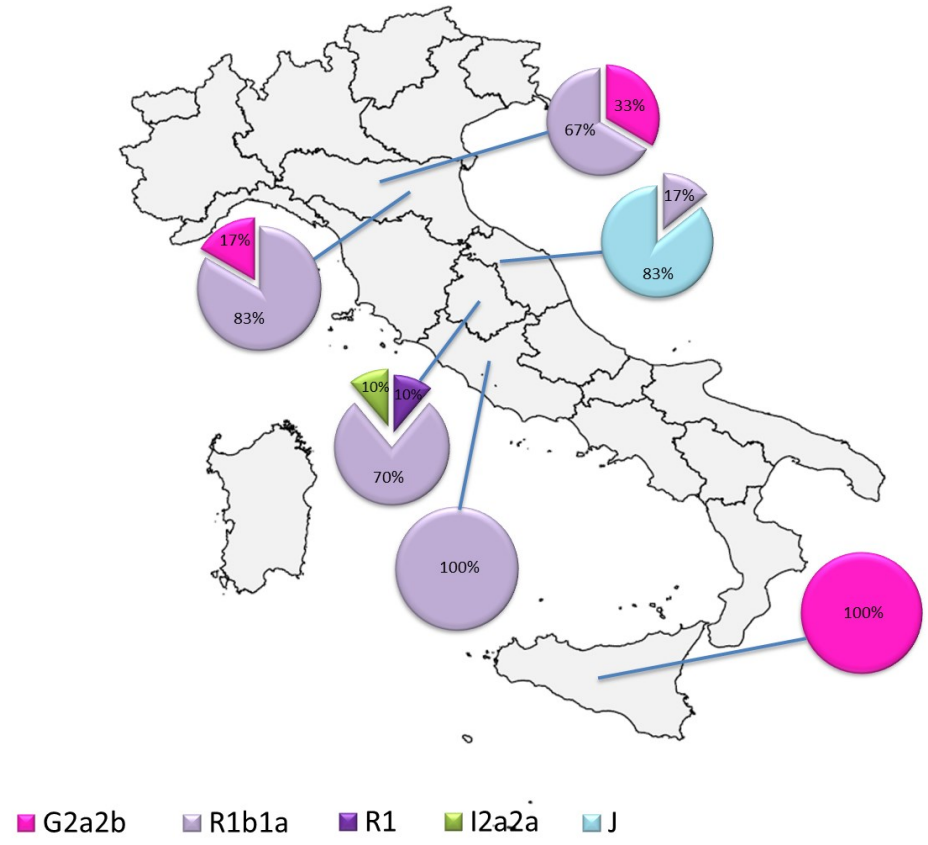

Figura 34: Distribuzione geografica dei principali Hg osservati nella popolazione italiana dell'età del Ferro.

L'Hg R è uno dei più diffusi in Europa[249], con una frequenza superiore al 50%; il sub-clade R1b, in particolare, si è ampiamente diffuso nell'Europa occidentale tra la fine del Neolitico e l'età del Bronzo[22,50,138], dopo essersi originato, presumibilmente 18500 anni fa in Eurasia[250]. Campioni recuperati dall'Europa centrale ed occidentale e datati tra il tardo Neolitico e l'età del Ferro condividono tipicamente l'Hg R1b[22,50,138,147,270]. La frequenza con cui tale Hg viene individuato nella popolazione italiana dell'età del Ferro analizzata in questo lavoro è comparabile con quella delle attuali popolazioni europee. In particolare, rappresenta la componente maggioritaria di variabilità genetica per tutti i siti del centro e Nord Italia analizzati, escluso quello di Matelica, nelle Marche. In questo sito infatti un solo campione è stato associato all'Hg R1b, mentre i restanti 6 individui processati appartengono al macro-Hg J. Sulla base dell'attuale cline di frequenza, tale Hg è stato associato alla diffusione degli agricoltori in Europa[220,222], con la linea capostipite sorta probabilmente nella penisola arabica tra 31700 e 12800 anni fa[251]. I due sub-cladi che rappresentano in modo maggioritario l'Hg, J1 e J2, sono responsabili di diverse ondate migratorie: J1 si è espanso attraverso due episodi temporalmente distinti, il più recente probabilmente associato alla diffusione del popolo arabo; la distribuzione di J2 è invece coerente con una rotta di

dispersione levantina/anatolica verso l'Europa Sud-Orientale e può riflettere la diffusione degli agricoltori anatolici[252]. Tuttavia, ad oggi, non sono stati individuati campioni europei del contesto Neolitico a supporto di questa ipotesi. Dai dati ad oggi disponibili su sequenze antiche è possibile affermare che la linea J si è iniziata ad affermare in Europa centrale ed occidentale nell'età del Bronzo, probabilmente come parte di un processo demografico di espansione a partire dal nord del Caucaso. L'Hg J è inoltre legato alle grandi migrazioni intervenute nell'età del Ferro[253]. La sua presenza è stata riportata nella parte meridionale della penisola italiana, a supporto dell'ipotesi di un flusso genico a partire da popolazioni di lingua araba del Vicino Oriente e del Nord Africa[207]. La diffusione marcata dell'Hg J nella popolazione del centro Italia, può essere legata probabilmente ad eventi migratori più recenti. Infatti è stato ipotizzato, che dal centro di radiazione iniziale, localizzato nel Caucaso e nelle regioni a nord di esso, diversi sub-cladi dell'Hg J potrebbero aver successivamente raggiunto l'Italia continentale, la Grecia e la Turchia, possibilmente seguendo percorsi e tempi differenti[254].

Otto individui, di cui il 75% recuperati nel sito di Polizzello, in Sicilia, presentano una variabilità associabile al clade G2a, Hg diffuso omogeneamente nella popolazione moderna, sebbene con frequenze basse, ma con picchi compresi tra il 15% ed il 30% nel Caucaso, in Sardegna e nel Sud Italia. I test genetici condotti su campioni antichi del periodo neolitico, in varie parti d'Europa, hanno confermato che l'Hg G2a era dominante tra gli agricoltori e pastori neolitici che migrarono dall'Anatolia all'Europa tra 9000 e 6000 anni fa[132]. La presenza dell'Hg G2a in Provenza, Italia meridionale e Ucraina potrebbe riflettere gli eventi di colonizzazione marittima greca dell'età del Ferro[255].

Infine, come detto, un solo individuo (CB_Tb14) presenta un profilo associabile all'Hg I2 (sub- Hg I2a2a1). La frequenza dei sub-cladi I1 e I2 è limitata all'Europa; è stato pertanto proposto che la ristretta area geografica di diffusione rifletta una continuità genetica con i cacciatori-raccoglitori paleolitici[256]; tale distribuzione è in netta contrapposizione con quanto determinato dal clade più affine ad I, il già citato Hg J, responsabile insieme all'Hg G della variabilità genetica veicolata dalla diffusione degli agricoltori del Vicino Oriente verso l'Europa. Si può datare l'origine dell'Hg I tra i 26000 e i 31000 anni fa, e stimare la divergenza del sub-clade I2a2a tra i 14600 ed i 3800 anni fa. L'Hg I2a2a è stato trovato in oltre il 4% della popolazione solo in alcune regioni europee, e per quanto riguarda l'Italia, in Toscana, Umbria e Lazio[256].

5.3.5 Median Joining Network

I rapporti filogenetici tra i campioni antichi italiani, prodotti attraverso comparazione dei profili mutazionali, sono mostrati nel MJN prodotto attraverso il software PopART[238] (Figura 35). Per la rappresentazione è stato scelto di utilizzare solo quei campioni che presentassero una soglia di siti non coperti minore del 5%, riducendo ad un totale di 29 i campioni analizzabili. In generale, gli aplotipi sono ampiamente raggruppati nelle loro rispettive linee filogenetiche di variazione e non sembra essere presente nessuna struttura generale associata alla geografia, sebbene si possa osservare una chiara predominanza di campioni siciliani all'interno dell'Hg G2a, ed una esclusività per quanto riguarda l'Hg J per gli individui provenienti dal sito di Matelica, nelle Marche.

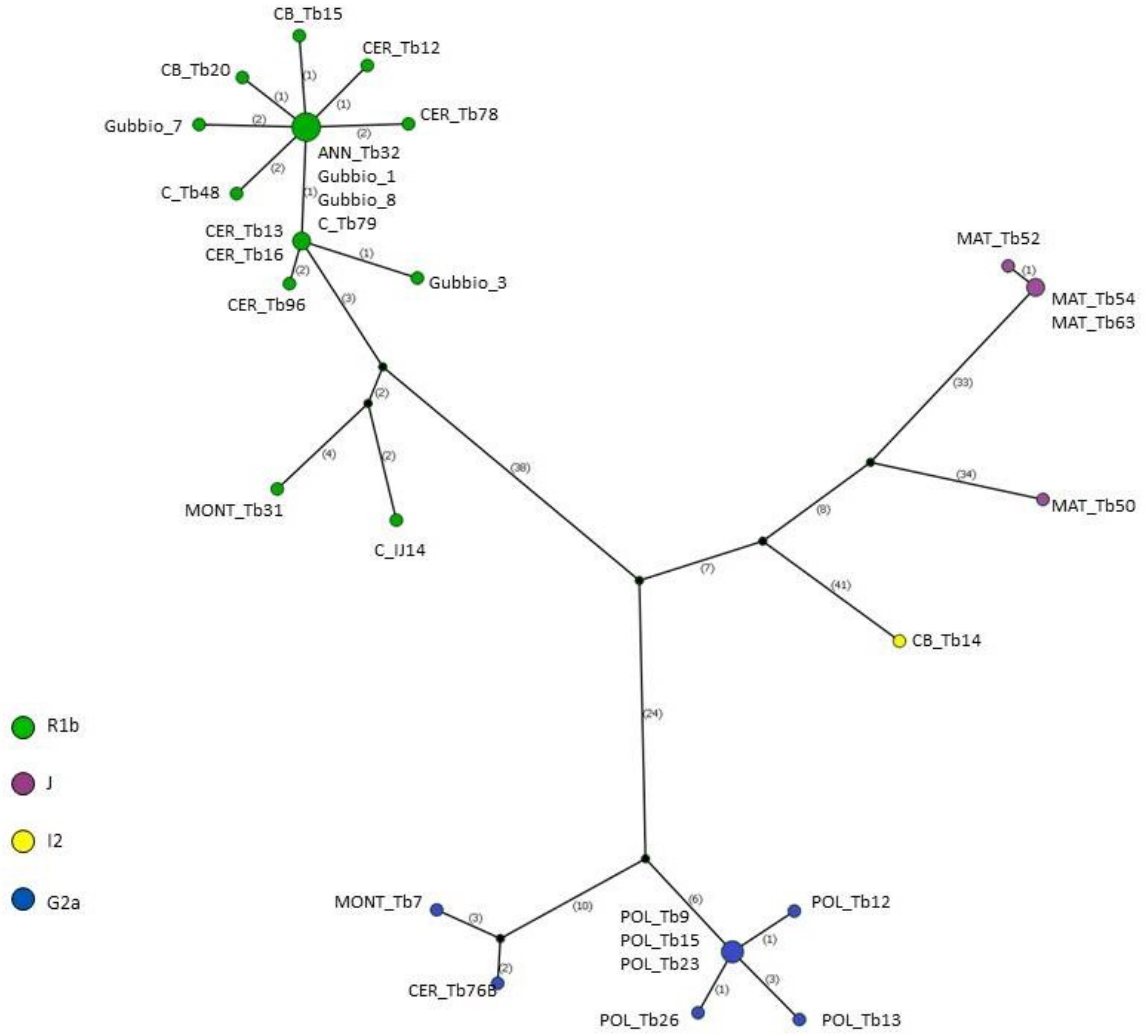

Figura 35: MJN rappresentante le relazioni filogenetiche tra i campioni italiani dell'età del Ferro sequenziati. I principali Hg sono rappresentati da diversi colori, riassunti in legenda. I numeri tra parentesi rappresentano il numero di mutazioni che distinguono i campioni contigui.

Ciò che emerge tuttavia è la netta distinzione, all'interno delle principali linee filogenetiche, di alcuni campioni rispetto al *core* centrale e più denso. Questo andamento è chiaramente rappresentato dal campione MAT-50, all'interno del gruppo di individui recuperati dal sito marchigiano; la marcata separazione genetica di tale campione dagli altri individui dello stesso sito, e la presenza di un nodo centrale di collegamento con gli altri campioni, avvalora l'ipotesi precedentemente discussa dell'appartenenza al sub-Hg J1, in contrapposizione con il resto degli individui che presenterebbero marcatori più tipici dell'Hg J2, sebbene questi non siano stati coperti con un sufficiente grado di profondità per poter eseguire un'assegnazione certa.

Le stesse considerazioni emergono nell'osservazione dei rami rappresentanti i campioni con un profilo genetico tipico dell'Hg G2a. In questo caso è possibile constatare che, sebbene i campioni presentino le stesse mutazioni *marker*, il gruppo insulare di Polizzello risulta più omogeneo, rispetto agli altri due campioni, MONT-7 e CER-76B, che si posizionano in un ramo differente e risultano maggiormente diversificati. Inoltre, anche questi due, recuperati da differenti siti, sono tra loro distinti a partire da un nodo comune. Questo potrebbe riflettere eventi di colonizzazione dell'Italia peninsulare, associati a flussi migratori diversi da quelli insulari, e provenienti da nord[255].

Infine, un discorso molto simile può essere esteso ai campioni C-IJ14 e MONT-31 all'interno del gruppo di individui che condividono i marcatori genetici tipici dell'Hg R1b. In questo caso, tali campioni differiscono, tra loro e con il più denso gruppo di individui provenienti da varie necropoli del centro Italia, per due nodi e circa 10 variazioni nucleotidiche. Altri campioni analizzati e provenienti dai medesimi siti di Castiglione (nel Lazio) e Montericco (Emilia-Romagna) tuttavia, possono essere visualizzati all'interno di un gruppo più ampio e più omogeneo; tale considerazione potrebbe essere motivata dalla presenza, non inusuale nel periodo dell'età del Ferro, di individui, quali C-IJ14 e MONT-31 che convivevano all'interno di ampi gruppi popolazionistici, sebbene non facessero parte della società fin dalla nascita[257].

Il MJN prodotto mette anche in evidenza un andamento contrario a quanto appena discusso, ovvero la condivisione del medesimo aplotipo da parte di alcuni campioni provenienti dallo stesso sito, o da siti diversi. In particolare, è stato possibile associare al medesimo aplotipo i campioni Pol-9, Pol-15 e Pol-23, per quanto riguarda il sito siciliano; nelle Marche, gli individui MAT-54 e MAT-63 condividono lo stesso profilo, così come CER-13 e CER-16 (Emilia-Romagna) ed i campioni umbri Gubbio-1 e Gubbio-8.

In generale, è possibile ipotizzare un cline di variabilità genetica distinto tra Nord e Sud Italia, già sostenuto in lavori precedenti[114,207], che mutuerebbe fenomeni di introgressione da diverse aree continentali già in tempi precedenti l'età del Ferro.

5.3.6 Analisi popolazionistiche

5.3.6.1 La variabilità genetica italiana nell'età del Ferro

Per la descrizione della variabilità genetica intra-popolazionistica dei campioni italiani dell'età del Ferro è stato utilizzato il software Arlequin 3.5.2.2, impostando i parametri descritti nel Paragrafo 5.2.3.4. In prima analisi, sono stati riassunti i principali indici di diversità standard e diversità genetica all'interno della popolazione (Tabella 16).

Popolazione	*n*	*#siti polimorfici*	*diversità nucleotidica*	*MNPD ± sd*
Emilia-Romagna	8	136	0.1535 ± 0.0849	54.5000 ± 26.4433
Umbria	9	156	0.0868 ± 0.0475	35.0000 ± 16.8681
Marche	4	95	0.1131 ± 0.0748	47.5000 ± 26.3145
Lazio	5	12	0.0189 ± 0.0128	5.4000 ± 3.1303
Sicilia	6	15	0.0107 ± 0.0069	5.5333 ± 3.0970

Tabella 16: Indici di diversità standard e genetica stimati per le popolazioni italiane antiche in esame. n: numero di campioni analizzati; MNPD: numero medio di differenze a coppie.

La diversità nucleotidica dei campioni nelle popolazioni in esame risulta piuttosto variabile. Inoltre, l'intero campione presenta un numero di siti polimorfici compreso tra 12 e 156 siti. Anche, le sequenze, prese a coppie, presentano un numero medio di siti differenti molto variabile. Le figure 36 e 37 riassumono la variabilità aplotipica osservata all'interno delle popolazioni e per tutto il campione preso in esame.

In particolare, nella prima figura sono mostrate le differenze genetiche, a coppie di individui, all'interno delle singole popolazioni analizzate. Ciascun sito è caratterizzato da un diverso numero massimo di differenze nucleotidiche, come evidenziato nella scala colorimetrica a destra di ciascun grafico. Essa è minima per i campioni di Lazio e Sicilia, in cui sono state riscontrate un massimo di 10 differenze nucleotidiche; nel primo caso il campione C-IJ14 rappresenta la maggior parte della variabilità interna, come già emerso dall'analisi del MJN. Per quanto riguarda la situazione siciliana è invece possibile osservare una maggiore eterogeneità interna; i due campioni che differiscono maggiormente tra loro sono POL-13 e POL-26.

Le popolazioni di Emilia-Romagna, Umbria e Marche, sono invece rappresentate da una differenza genetica massimale di oltre 100 nucleotidi, veicolata dalla presenza, in ciascuna popolazione, di un campione, o due (nel caso emiliano), estremamente diversificato. In particolare, in Emilia-Romagna e Umbria tale dato rispecchia la presenza di individui appartenenti a linee filogenetiche differenti rispetto a quelle della maggior parte dei campioni analizzati: MONT-7 e CER-76B sono infatti gli unici due individui del gruppo emiliano appartenenti all'Hg G2a; la differenza genetica a

coppie sottolinea tale dato, come è possibile osservare dal colore estremamente più chiaro che diversifica i campioni citati dal resto del gruppo. Analogamente, per i siti umbri, CB-14 è l'unico individuo ascrivibile al macro-Hg I, e pertanto la differenza genetica a coppie, con tutti gli altri individui risulta estremamente alta. In questo gruppo, se si esclude tale individuo, è possibile osservare una situazione estremamente omogenea, con molte coppie di individui che presentano poche, o nessuna variazione nucleotidica. Tale dato potrebbe permettere di ipotizzare, per lo meno per quegli individui recuperati dal medesimo contesto archeologico, una condizione di parentela per via paterna (per esempio per i campioni Gubbio-8 e Gubbio-14, che avevano mostrato un'eguaglianza aplotipica anche dal dato prodotto mediante MJN). Ancora una volta, la netta distinzione genetica di MAT-50 dal resto del gruppo di campioni recuperati dalla medesima necropoli, permetterebbe di confermare l'attribuzione di tale individuo ad una differente linea filogenetica. I restanti 3 individui di Matelica risultano estremamente omogenei, sebbene non sia presente un'identità di sequenza tra nessuno di essi.

Il più alto grado di diversità nucleotidica e il numero estremamente più alto di differenze a coppie (MNPD - Tabella 16) associato ai siti dell'Emilia-Romagna, dell'Umbria e delle Marche, rispetto a quanto si osserva nel siti laziale e siciliano, è pertanto da considerare principalmente come effetto di quanto appena discusso.

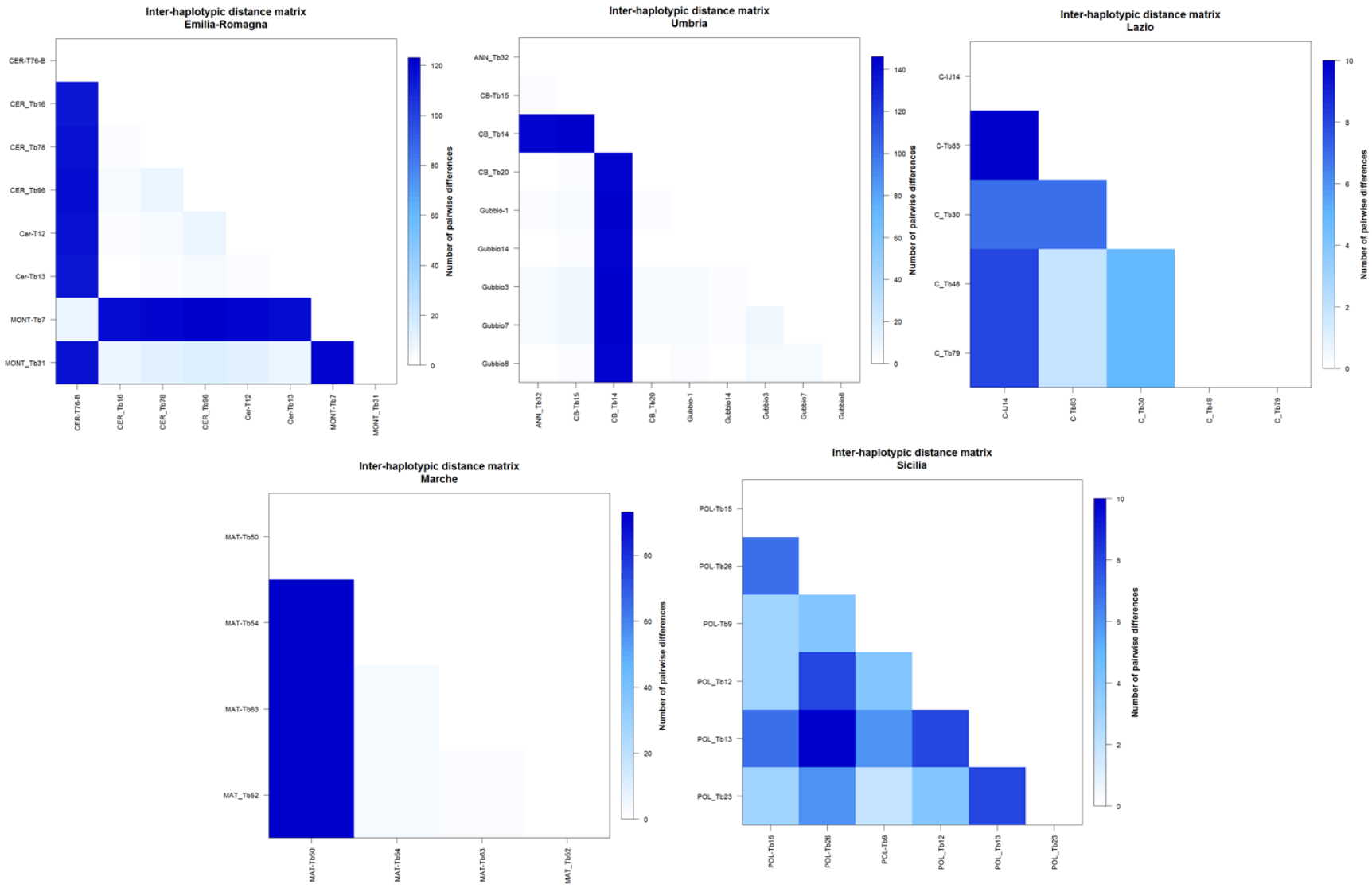

Figura 36: Distanze genetiche osservate all'interno delle popolazioni italiane analizzate.

Le stesse considerazioni possono essere estese alla Figura 37. In questo caso, tutti i campioni analizzati sono stati comparati tra loro, indipendentemente dalla necropoli da cui sono stati recuperati. Ciò che emerge è la marcata differenza del campione CB-14 da qualsiasi altro campione analizzato, per effetto dell'appartenenza ad una linea filogenetica esclusiva. Allo stesso modo è evidenziata la spiccata differenza delle popolazioni di Sicilia e Marche dal resto dei gruppi del Centro Italia, ad esclusione di MONT-7 e CER-76B che presentano una maggiore vicinanza genetica con gli individui siciliani.

Tale dato conferma a pieno le analisi effettuate attraverso il MJN e le inferenze da esso deducibili.

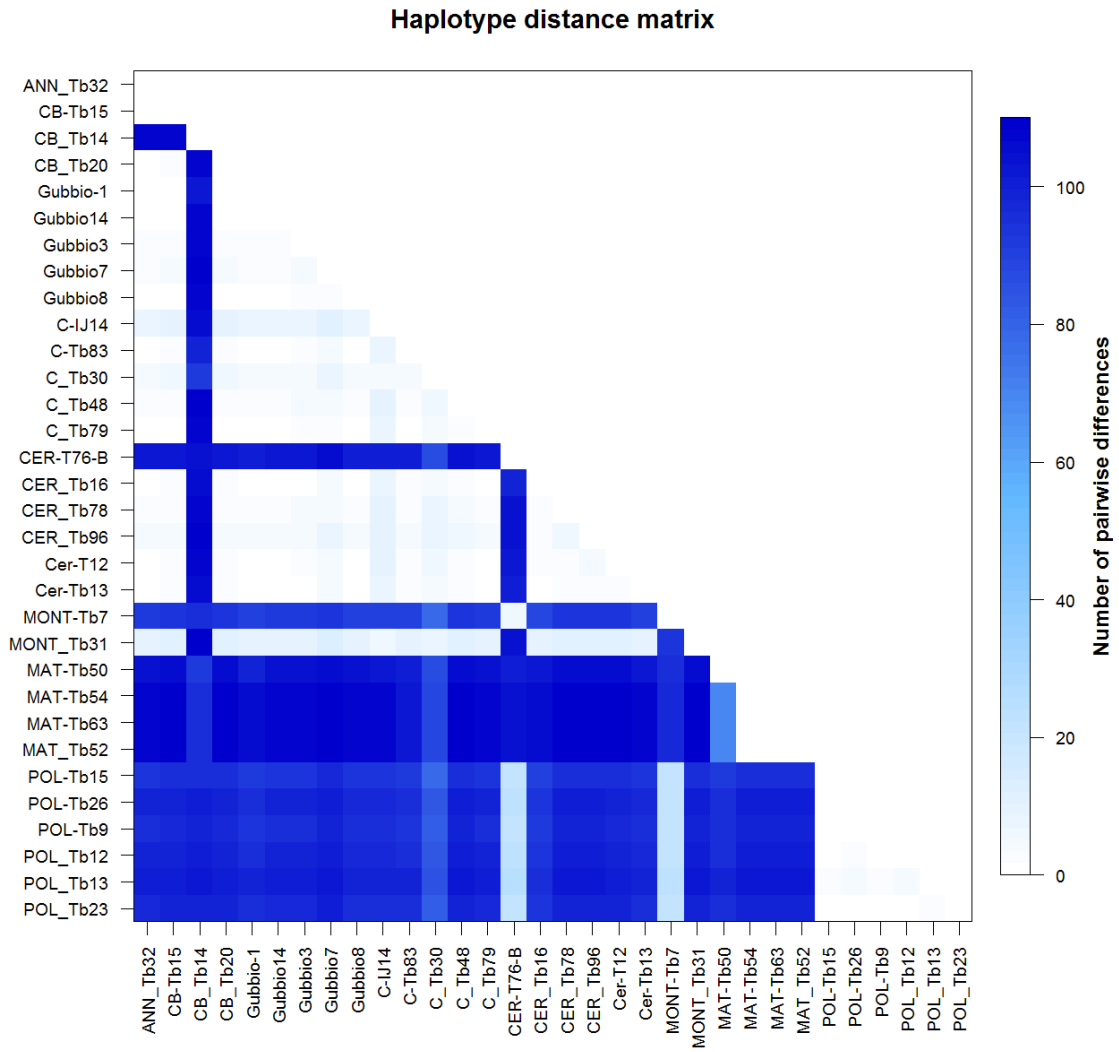

Figura 37: Distanze genetiche osservate tra i campioni italiani dell'età del Ferro.

Le principali caratteristiche di similarità e divergenza all'interno e tra le popolazioni italiane analizzate, oltre all'indice di diversità genetica di Nei, sono riassunte in Figura 38.

La diagonale della figura permette di apprezzare in modo sintetico le differenze all'interno di ciascuna popolazione analizzata. Il gruppo dell'Emilia-Romagna è quello presenta una maggiore dissomiglianza interna, mentre quello siciliano risulta il più omogeneo. A conferma di quanto precedentemente detto, nel confronto tra popolazioni, i gruppi di Umbria e Lazio condividono la maggior vicinanza genetica, mentre la situazione è più eterogenea in tutte le altre coppie di popolazioni. In particolare, le maggiori distanze genetiche si osservano tra la popolazione di Matelica e tutti gli altri gruppi analizzati. Come analisi addizionale delle distanze genetiche tra popolazioni, nella *heatmap* mostrata nel quadrante al di sotto della diagonale, sono stati rappresentati i valori dell'indice di Nei. Il risultato che ne deriva è speculare a quanto già osservato.

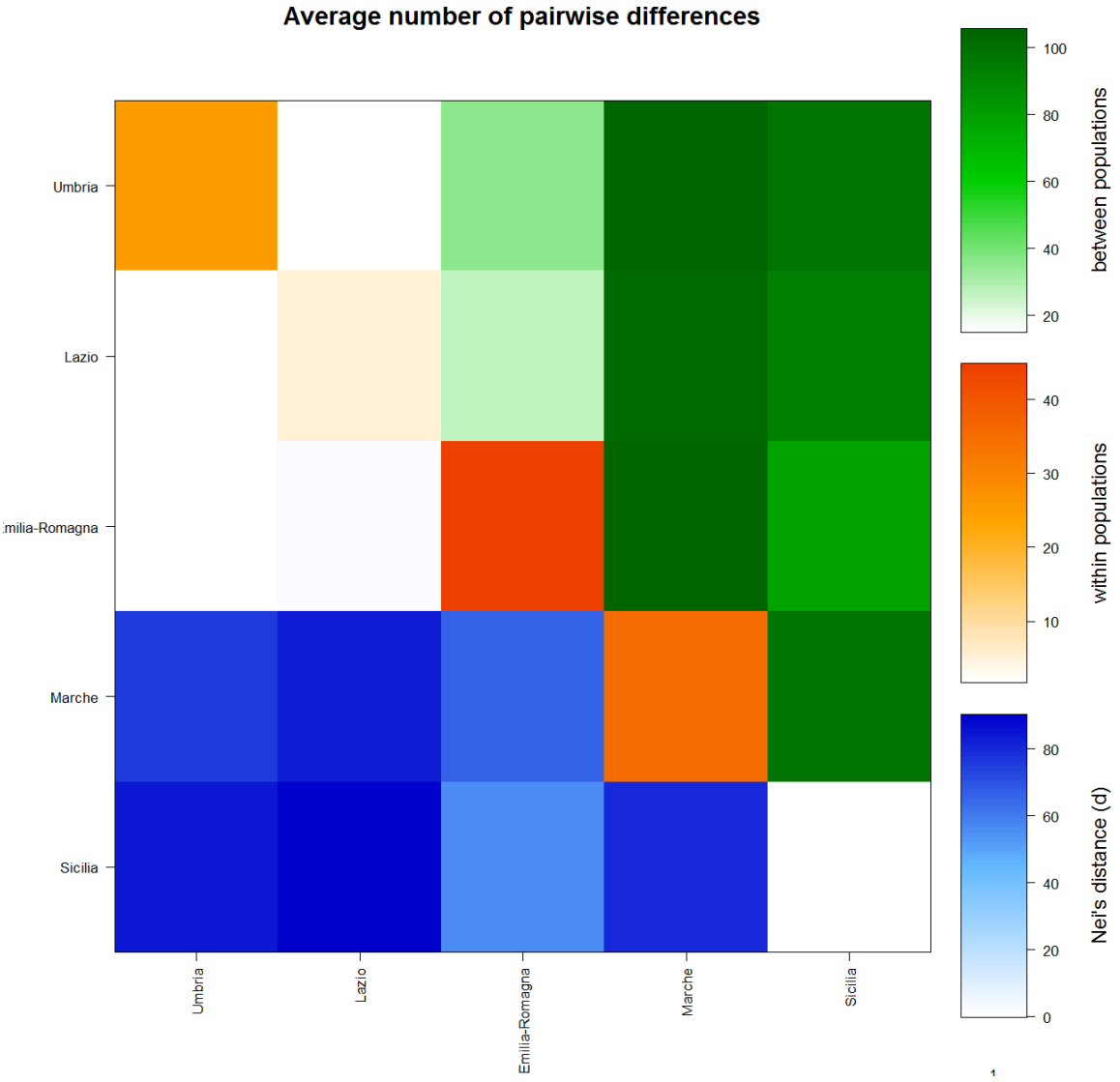

Figura 38: Differenze genetiche osservate all'interno e tra le popolazioni oggetto delle analisi molecolari effettuate.

Infine, in Figura 39 è riportata la *heatmap* relativa alla matrice dei valori Fst che relazionano le popolazioni dell'età del Ferro di Emilia-Romagna, Umbria, Marche, Lazio e Sicilia, al fine di avere una più facile visualizzazione dei rapporti di vicinanza/distanza tra le popolazioni esaminate. I valori di Fst sono compresi tra -0.07053 e 0.96549. Le popolazioni tra loro più vicine si confermano essere quelle di Umbria e Lazio (Fst = -0.07053), mentre quelle tra loro più lontane sono quelle del Lazio e della Sicilia (Fst = 0.96549).

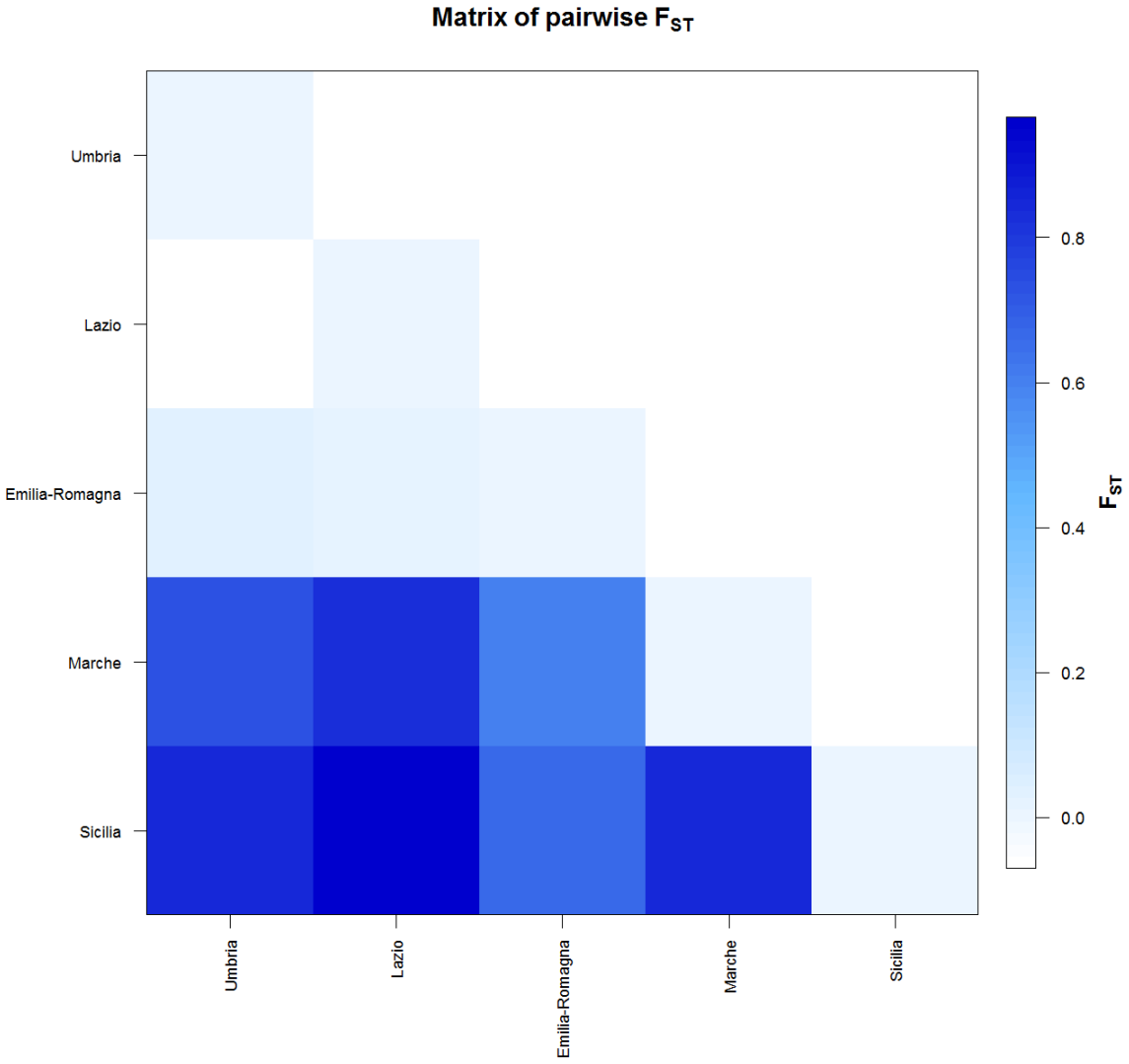

Figura 39: Valori di Fst tra le popolazioni italiane dell'età del Ferro.

5.3.6.2 Il *pool* genetico del Y–chr italiano nel contesto mediterraneo

Al fine di visualizzare le relazioni tra i gruppi italiani analizzati ed altre 32 popolazioni europee moderne ed antiche, di vari periodi storici, estesi dal Neolitico all'impero romano, è stata effettuata una PCA sulla base delle frequenze degli Hg riportati in Tabella A 3 – Appendice (v. Paragrafo 5.2.3.5).

Come dettagliato nel Paragrafo 5.2.3.5, i campioni italiani analizzati in questo studio sono stati divisi in 3 macro-popolazioni al fine di evitare *bias* legati ad una esigua numerosità di campioni per sito, e plottati insieme ad altri 1735 individui, a loro volta suddivisi in macro-popolazioni. In Figura 40 le stelle evidenziano la localizzazione dei gruppi di campioni sequenziati in questo studio, mentre i punti rossi rappresentano le disposizioni reciproche delle altre popolazioni recuperate in letteratura. Le prime due componenti, rappresentate nel *plot* in Figura 40, complessivamente, spiegano il 18.76% della varianza.

La prima componente principale (PC1) separa chiaramente tutti i campioni neolitici (ad eccezione di quelli italiani – IT_N) e tutti i campioni dell'età del Rame, ad eccezione di quelli recuperati nell'area delle Steppe (St_CA), dai campioni analizzati in questo studio. Inoltre la PC1 sottolinea una chiara differenza anche nelle frequenze degli Hg dei campioni dell'età del Bronzo isolati in Sardegna e nell'area dei Balcani. Risulta invece una maggiore vicinanza nelle frequenze degli Hg degli individui italiani dell'età del Ferro a quelle degli individui isolati in Europa nel periodo di transizione tra Neolitico ed età del Bronzo come conseguenza di una diminuzione della variabilità genetica complessiva tra le popolazioni di tali contesti storici[138].

È interessante notare come la seconda componente principale (PC2) separi in modo netto le frequenze degli Hg dei moderni siciliani e abitanti del Sud Italia in generale (SS_M). Questo dato tuttavia può essere motivato dal fatto che gli individui sicani di Polizzello usati in questo studio fanno parte di una popolazione insulare endemica e non stupisce pertanto la bassa variabilità interna, responsabile dell'elevato differenziamento dalle popolazioni attuali delle stesse aree geografiche.

Le restanti popolazioni, assieme ai campioni analizzati in questo studio sono distribuite senza chiari segni di separazione latitudinale, e con una più ampia distribuzione dettata dalla PC2. La particolare affinità delle popolazioni analizzate con individui italiani del periodo romano (IT_R) indica che è necessaria una chiara comprensione della distribuzione temporale dei campioni analizzati sulla base di datazioni al radiocarbonio, che sono in fase di processamento, per poter assegnare chiaramente la popolazione italiana all'esatto contesto storico. In generale inoltre, la variabilità di tali popolazioni, rispecchia quella che si evidenzia lungo tutto il continente nello stesso periodo storico (IT_IA).

È accettata ad oggi l'evidenza di un'alta struttura genetica già nell'Europa dell'età del Bronzo, che risulta rispecchiare la variabilità genetica odierna, plasmata dalle componenti dei primi gruppi di cacciatori-raccoglitori, e dagli agricoltori neolitici, a cui si è aggiunto l'*input* genetico della cultura Yamnaya[138]. Nella PCA in Figura 40 è confermata la vicinanza di quest'ultima componente (Yam) ai campioni italiani dell'età del Ferro e all'attuale frequenza dei principali Hg europei (E_M).

Figura 40: PCA basata sulle frequenze degli Hg del Y-chr degli individui italiani dell'età del Ferro (suddivisi su base geografica in 3 popolazioni, rappresentate dalle stelle) e di altri 1735 campioni raggruppati in popolazioni e periodi storici come descritto in Tabella A 3.
CNIT_IA: Centro-Nord_Italy_IronAge; **CSIT_IA:** Centro-Sud_Italy_IronAge; **Sy_IA:** Sicily_IronAge; **IT_IA:** Italy_IronAge; **E_IA**: Europe_IronAge; **Nom_IA**: Nomads_IronAge; **Sy_BA:** Sicily_BronzeAge; **Sa_BA:** Sardinia_BronzeAge; **B_BA:** Balkans_BronzeAge; **CE_BA:** Centre_Europe_BronzAge; **UK_BA:** UK_BronzAge; **WE_BA:** West_Europe_BronzAge; **Yam:** Yamnaya_BronzeAge; **L_BA:** Lebanon_BronzeAge; **St_BA:** european-Steppe_BronzeAge; **BIT:** BronzeAge-IronAge_Transition (Russia); **Sy_CA:** Sicily_CopperAge; **Sa_CA:** Sardinia_CopperAge; **B_CA:** Balkans_CopperAge; **CE_CA:** Centre_Europe_CopperAge**; E_CA:** Europe_CopperAge**; St_CA:** european-Steppe_CopperAge; **St_CBT:** Steppe_CopperAge-BronzeAge_Transition; **IT_N:** Italy_Neolithic; **B_N:** Balkans_Neolithic; **CE_N:** Centre_Europe_Neolithic; **UK_N:** UK_Neolithic; **WE_N:** West_Europe_Neolithic; **EE_N:** Est_Europe_Neolithic; **E_N:** Europe_Neolithic; **E_NBT:** Europe_Neolithic-BronzeAge_Transition; **IT_R:** Italy_Roman_Period; **E_M:** Europe_modern_1000genomes; **T_M:** Tuscans_modern_1000genomes; **SS_M:** Sicily-sud_Modern.

5.4 Conclusioni e obiettivi futuri

L'Italia è sempre stata un agevole scalo per i popoli durante le proprie rotte migratorie marine e terrestri condotte a partire dall'Africa, dal Medio Oriente e da tutta l'Europa, grazie alla posizione geografica e alle caratteristiche geomorfologiche che la distinguono[207]. Questo si è accentuato soprattutto durante l'età dei metalli, quando le connessioni con il continente si sono fatte più vive anche attraverso l'aumento delle rotte commerciali. Tale evento è attestato dalla concomitante presenza sul suolo peninsulare di un'ampia varietà di popolazioni durante tutto il primo millennio a.C. Per la sua posizione cruciale pertanto, l'Italia ha sperimentato una complessa storia di colonizzazioni, migrazioni e commistioni, la cui traccia genetica è ancora presente negli italiani moderni[116].

Dal punto di vista maschile, la maggior parte del *pool* genetico del Y-chr italiano può essere correlato a cinque Hg principali: R1b, J2, I, G ed E1b. Il primo di essi risulta più frequente nel Nord Italia, mentre E1b, G e J2 presentano frequenze più alte nel Sud, suggerendo una maggiore affinità con l'Europa occidentale per l'hg G e con l'Europa Sud-Orientale e Centro-Meridionale per J2[115,207,252,255].

Al fine di approfondire le conoscenze filogenetiche relative alla variabilità del Y-chr nell'Italia antica e contribuire alla ricostruzione della complessa storia demografica della penisola, sono stati recuperati ed analizzati geneticamente un set di campioni associati, sulla base dei contesti archeologici in cui sono stati ritrovati, al periodo dell'età del Ferro, ad oggi ancora poco conosciuto dal punto di vista del marcatore uniparentale maschile.

Il Y-chr, per le proprietà biologiche che lo contraddistinguono, è un valido marcatore per l'analisi delle variazioni genetiche veicolate dagli individui di sesso maschile; infatti, caratteristiche quali la lunghezza inferiore rispetto a quella degli autosomi, la scarsa omologia con il cromosoma X, e la possibilità di andare incontro a variazioni unicamente attraverso eventi mutazionali, unitamente all'aploidia, e all'ereditarietà esclusivamente per via paterna, lo rendono, insieme alla controparte femminile, mtDNA, un importante *target* per molte applicazioni genetiche, che vanno dalle indagini forensi, alle analisi genealogiche, fino agli studi di genetica di popolazione.

Sono stati pertanto isolati, attraverso un metodo di cattura in soluzione, 31630 SNPs di interesse filogenetico da un totale di 42 campioni risultati di sesso maschile a seguito di uno *screening* genetico effettuato su un totale di 101 individui, recuperati in 8 diversi contesti archeologici del Centro Italia e della Sicilia (v. Tabella A 1 – Appendice).

Per più del 90% dei campioni processati è stato possibile, al minimo, definire le variazioni caratteristiche per l'assegnazione ai rami principali degli Hg del Y-chr, attraverso l'analisi delle posizioni *marker* interne alla regione MSY catturate.

I campioni italiani dell'età del Ferro sono stati classificati in 4 differenti Hg del Y-chr: R1 (n=23), J (n=6), G2a (n=8) e I2a (n=1). Questi ricalcano i principali Hg attestati ad oggi nella penisola;

risulta tuttavia interessante notare che nessun individuo è stato attribuito all'Hg E1b, caratteristico della variabilità italiana odierna.

La distribuzione di tali Hg sembra supportare le ipotesi precedentemente avanzate, di un cline di variabilità genetica ben distinto tra Nord e Sud Italia[114,207], a sostegno di fenomeni di introgressione da diverse aree continentali già in tempi precedenti l'età del Ferro. Escludendo gli individui assegnati all'Hg R1b (omogeneamente diffuso in tutta Europa a partire da tempi ben più remoti dell'età dei metalli), che rappresentano più della metà della popolazione esaminata, tuttavia, il numero di campioni analizzati ed assegnati ad Hg indicativi di specifiche tratte migratorie, risulta troppo esiguo per poter fornire chiare evidenze dei movimenti umani intercorsi durante la protostoria e responsabili della variabilità genetica della popolazione in esame.

Risulta comunque netta la distinzione nella variabilità genetica delle popolazioni provenienti da Sicilia e Marche, rispetto alle altre popolazioni del Centro-Nord Italia.

Le indagini a livello inter-popolazionistico, effettuate con un totale di 32 popolazioni europee antiche (di arco temporale compreso tra il Neolitico e l'impero romano) e moderne sono state condotte sulla base delle frequenze dei principali Hg del Y-chr ad oggi noti. Queste hanno messo in evidenza una netta differenza con le popolazioni neolitiche (ad eccezione dei campioni italiani) e dell'età del Rame, ed una più netta similitudine con le popolazioni europee e delle steppe del medesimo arco temporale.

Va tuttavia considerato che ad oggi non sono disponibili le datazioni al radiocarbonio per nessuno dei campioni analizzati in questo studio. Sarà pertanto fondamentale aggiungere tale informazione, unitamente ad un aumento dei campioni processati lungo tutta la penisola, per favorire una chiara contestualizzazione temporale in un paese rappresentato da numerose genti e movimenti intercorsi in brevi lassi temporali.

CAPITOLO 6: CONCLUSIONI

Nell'ultimo decennio, le nuove tecnologie e i miglioramenti nei protocolli di lavoro hanno permesso di recuperare in modo efficiente l'aDNA e di superare sfide che non sembravano sormontabili fino a poco tempo prima. Tuttavia, la frazione di DNA endogeno continua a essere un fattore limitante negli studi che coinvolgono materiale degradato. I primi sforzi per superare questa limitazione si sono concentrati sull'arricchimento dell'mtDNA, viste le sue ristrette dimensioni (~ 16 kB) e la numerosità di copie in ogni cellula, comparate con quelle degli autosomi (due copie) e del Y-chr (una copia). Inoltre, per quest'ultimo, le strategie di arricchimento presentano maggiori problematiche a causa della sua ricchezza di sequenze ripetitive e palindromiche. Nonostante queste difficoltà, il Y-chr è un marcatore elettivo, insieme alla controparte femminile, mtDNA, per tracciare schemi diacronici e sincronici del popolamento umano, ed in particolare per mettere alla luce migrazioni e contatti tra le popolazioni antiche[79].

Ad oggi è ben noto che le migrazioni a partire da diverse aree geografiche, sono state numerose nel passato e hanno plasmato la moderna variazione genetica; tuttavia, l'utilizzo di campioni moderni per tracciare gli eventi del passato non è sufficiente per mettere in luce tutte le componenti che hanno concorso a tale modellamento. Questo è un tema critico soprattutto per regioni, come l'Italia, che per la posizione geografica strategica, hanno rappresentato a lungo un importante crocevia del Mediterraneo dove popoli e culture diverse si sono mescolate nel tempo. La storia dei popoli che hanno approdato nella penisola, a partire da aree geografiche e tempi differenti, ha reso la ricostruzione della sua storia genetica e della struttura della popolazione estremamente controversa e ampiamente dibattuta. Numerosi movimenti di popolazioni si sono verificati tra il bacino del Mediterraneo e il Medio Oriente durante l'Età dei Metalli, periodo che ha determinato la trasformazione delle prime organizzazioni sociali in antiche civiltà. Ad oggi non sono disponibili informazioni sufficienti sull'origine e sui possibili eventi di commistione di queste popolazioni, e le nostre conoscenze sono ancora pressoché incomplete dal punto di vista genetico. Solo poche ricerche infatti (basate per lo più sullo studio del marcatore femminile[258,259,260], ed in pochi casi dell'intero genoma[150,261]) hanno indagato direttamente la variabilità genetica in gruppi umani antichi di questi periodi. Inoltre, come suggerito da studi precedenti sulla popolazione italiana odierna, l'attuale struttura genetica è probabilmente il risultato di storie demografiche diverse intercorse per maschi e femmine: un modello più omogeneo di variabilità del mtDNA probabilmente è da attribuire ad eventi più antichi, mentre la struttura più variegata del Y-chr è legata ad eventi migratori più recenti[114]. Ad oggi tuttavia, non esistono studi, se non relegati a pochi campioni antichi, dal punto di vista dell'eredità veicolata dal cromosoma maschile. Questo, presenta caratteristiche che hanno enormi influenze sulla struttura genetica, sui processi mutazionali e sulla diversità tra popolazioni, e all'interno della popolazione stessa[80]: la quasi totalità delle basi di cui è composto il Y-chr infatti ricadono all'interno della regione NRY, nella

quale per la maggior parte non sussistono fenomeni di *crossing-over*, ed il suo assetto aploide e maschio-specifico lo rendono più gestibile degli autosomi; pertanto, l'apporto informativo che potrebbe derivare dall'analisi di ampie regioni di tale marcatore, attraverso lo studio di campioni antichi, potrebbe risultare fondamentale per la comprensione delle tracce genetiche lasciate dagli individui di sesso maschile durante tutti i movimenti popolazionistici intervenuti nel corso del tempo.

In questo lavoro di tesi pertanto, attraverso le informazioni disponibili in letteratura[79], sono stati identificati un totale di 79565 SNPs localizzati nella regione X-degenerata del Y-chr e considerati determinanti per produrre inferenze relative alle linee di discendenza per via paterna, ma in grado anche di fornire una diversificazione tra individui che condividono la stessa linea genetica. Sono stati successivamente condotti due differenti disegni sperimentali su un *subset* delle posizioni totali individuate, ma in grado comunque di fornire le informazioni necessarie con un ottimo livello di risoluzione, ed è stata confrontata l'efficienza, specificità e resa delle sonde prodotte attraverso diversi protocolli di arricchimento delle regioni selezionate (CAPITOLO 4: CASO STUDIO I). Il migliore assetto sperimentale, individuato secondo i dati ottenuti dalle analisi statistiche condotte sui sequenziamenti prodotti, è stato applicato all'analisi di un gruppo di campioni provenienti da diversi siti italiani e cronologicamente appartenenti all'età del Ferro (CAPITOLO 5: CASO STUDIO II), con lo scopo di apportare nuovi dati per chiarire i flussi migratori intercorsi verso l'Italia e all'interno della penisola in un periodo caratterizzato da molti popoli e movimenti anche di natura commerciale.

Nel primo caso studio, sono stati eseguiti test su un set di campioni moderni (n=4) ed antichi (n=4), per verificare l'efficienza delle sonde a RNA, prodotte da due diverse aziende e con diverse caratteristiche (v. Paragrafo 4.2.1). Per entrambi i gruppi di campioni sono state confrontate le statistiche di resa prodotte con l'utilizzo dei due differenti pannelli di sonde. Per i campioni antichi inoltre, la resa ottenuta utilizzando il protocollo standard fornito dalle aziende produttrici delle sonde, è stata comparata con quella raggiunta attraverso l'allestimento di protocolli modificati, e più specifici per materiale degradato. Le statistiche a nostra disposizione ci permettono di affermare, in generale, una buona riuscita delle fasi sperimentali ed una buona efficienza delle sonde disegnate, soprattutto nell'analisi di campioni antichi e degradati.

Con un set estremamente ampio di sonde mirate alla cattura di ~1.5Mb del Y-chr, è possibile effettuare analisi di valutazione degli Hg più dettagliate rispetto a quelle ottenute da precedenti analisi effettuate mediante *shotgun* profondi o catture di SNPs genomici[176]. I risultati ottenuti mostrano che entrambi gli assetti valutati presentano una buona resa, ma l'utilizzo di disegni sperimentali e protocolli specificamente sviluppati per materiale degradato aumentano significativamente la specificità delle sonde e le informazioni deducibili, anche da campioni estremamente degradati.

Con questo approccio ottimizzato è quindi possibile giungere alla lettura di un elevato numero di posizioni del Y-chr simultaneamente, rendendo la metodologia innovativa rispetto a quanto ad oggi in uso. La selezione di un numero così elevato di posizioni informative è fondamentale in genetica di popolazioni per far fronte a possibili regioni in cui non è possibile recuperare dati, a causa dell'alta degradazione dei campioni maneggiati. Inoltre, questo approccio potrebbe essere applicato in genetica forense, in cui le analisi ad oggi sono ristrette ad un numero estremamente limitato di polimorfismi, processati attraverso le classiche metodologie di amplificazione del DNA. Infine, l'individuazione di un così elevato numero di posizioni polimorfiche può essere sfruttata in contesti storici, analogamente a quanto fatto nel passato attraverso il mtDNA[11,12], laddove vi sia una profonda conoscenza della linea genetica maschile, per chiarire questioni quali (i) l'attribuzione dei resti ad un personaggio storico importante, (ii) la validazione di opere artistiche e (iii) la conferma di ipotesi di varia natura risolvibili attraverso l'analisi del Y-chr.
Questo sistema pertanto ben si adatta a numerosi casi studio, grazie anche alla possibilità di abbattere i costi e aumentare significativamente la risoluzione rispetto a quanto conseguibile con sequenziamenti mirati all'intero genoma.

A seguito della messa a punto di un sistema in grado di fornire massima specificità e resa nelle fasi di *target enrichment*, nel secondo progetto di questa tesi sono stati processati un totale di 42 campioni di sesso maschile associati, in accordo con i contesti archeologici, al periodo dell'età del Ferro nel Centro e Sud Italia. Le analisi condotte sul Y-chr hanno messo in luce la differenziazione genetica della popolazione in 4 principali Hg: R1 (60.53%), I2 (2.62%), G2a (21.05%) e J (15.8%). Il primo di essi è il più diffuso in Europa occidentale, Eurasia ed Africa centrale, con una frequenza superiore al 50%, riscontrata anche nei campioni analizzati in questo progetto[249].
Il macro-Hg I2 è tipico della variabilità genetica di Spagna e Gran Bretagna, ed in Italia è quasi esclusivamente presente nella popolazione sarda; in modo concordante con le aspettative, un solo campione appartiene a tale Hg[256].
Particolare è la presenza del macro-Hg J, rinvenuta esclusivamente nei campioni provenienti dal sito di Matelica, nelle Marche; esso è infatti associato principalmente ai popoli del Vicino Oriente, ma è legato alle grandi espansioni occorse a partire dal Neolitico, dapprima in Nord Africa e successivamente verso Europa ed Asia[252].
Infine, l'Hg G2a risulta diffuso ovunque sebbene con bassa variabilità, ad esclusione delle regioni del Sud Italia e della Sardegna[255]. I dati in nostro possesso confermano tale distribuzione, dal momento che il 75% dei campioni attribuibili a tale Hg fanno parte della popolazione siciliana. La bassa variabilità che emerge nei campioni provenienti dal sito di Polizzello non è sorprendente, dal momento che si tratta di una popolazione insulare endemica.
Sebbene la quantità di dati attualmente disponibili a riguardo della variabilità degli Hg del Y-chr nella popolazione italiana dell'età del Ferro processata, non fornisca alcuna chiara evidenza, soprattutto per la marcata presenza dell'Hg R1b (omogeneamente diffuso in tutta Europa già da

tempi preistorici), è possibile ipotizzare un cline di variabilità genetica tra Nord e Sud Italia (sostenuto dalle popolazioni di Matelica e Polizzello), già proposto in lavori precedenti[114,207].

Infine, comparando le frequenze degli Hg del Y-chr della popolazione in analisi, con quelle di popolazioni antiche e moderne recuperate in letteratura, è stato possibile osservare una vicinanza genetica della popolazione italiana dell'età del Ferro, con quelle dello stesso periodo recuperate da diversi contesti italiani, oltreché con le popolazioni dell'età dei metalli provenienti dalle steppe pontico-caspiche. Come suggerito da Gunther e Jakobsson[130] dopo le imponenti migrazioni durante i periodi del Neolitico antico e dell'età del Bronzo, la composizione genetica delle popolazioni in alcune aree europee si è mantenuta pressoché inalterata fino ad oggi. Questa considerazione non nega le migrazioni successive, ma suggerisce che le popolazioni coinvolte non erano così diversificate come durante il Neolitico.

La particolare affinità delle popolazioni analizzate in questa tesi con individui italiani del periodo romano tuttavia indica che è necessaria una chiara comprensione della distribuzione temporale dei campioni analizzati, sulla base delle datazioni al radiocarbonio, per poter assegnare chiaramente la popolazione italica all'esatto contesto storico. Nondimeno, sarà necessario ampliare le analisi del Y-chr su un più largo set di campioni antichi italiani (ma non solo) degli stessi periodi storici, per poter chiarire con maggiore dettaglio la complessa situazione dell'età del Ferro.

APPENDICE

Allegato A 1: Informativa e consenso per il trattamento dei dati dei soggetti coinvolti nello studio.

UNIVERSITÀ
DEGLI STUDI
FIRENZE

DIPARTIMENTO
DI BIOLOGIA

Informativa e consenso per trattamento di dati sensibili

Io sottoscritto/a , residente a
Via n.
Identificato con n.

acquisite le informazioni fornite dal titolare del trattamento ai sensi dell'art. 13 del D. Lgs. n. 196/2003 e consapevole che il trattamento riguarderà i dati "sensibili" di cui all'art. 4 comma 1 lett. d,

Dichiaro il mio esplicito consenso

1. al prelievo di materiale biologico salivare sulla mia persona;
2. all'estrapolazione del profilo genetico da tale materiale biologico;
3. alla comunicazione del mio profilo genetico in forma anonima per eventuali pubblicazioni scientifiche;
4. alla conservazione di quest'ultimo secondo la vigente legge sulla privacy (Decreto Legislativo 30 giugno 2003, n. 196 "Codice in materia di protezione dei dati personali" e succ. mod. ed int.).

Dichiaro, inoltre, di essere stato reso edotto dell'utilizzabilità dei dati genetici di cui sopra esclusivamente per finalità sperimentali e non cliniche o di altra natura.

Luogo e data,

Firma

Laboratori di Antropologia
Via del Proconsolo, 12 – 50122 Firenze
+39 055 2757741 fax +39 055 2757753 | e-mail: antropologia.molecolare@unifi.it
P.IVA | Cod. Fis. 01279680480

Tabella A 1: Lista dei campioni utilizzati per le analisi molecolari, divisi per sito. *Campioni maschi identificati con analisi di determinazione genetica del sesso per progetti precedenti. °Campioni per cui sono disponibili entrambe le lateralità della rocca petrosa.

Sito	*Regione*	*Periodo*	*ID archeologico*	*IDLab*	*Distretto scheletrico*	*Informazioni*
Montericco (Imola)	Emilia-Romagna	600-300 BCE	MONT_T.13=T.47	MONT-13	Rocca petrosa sx	*
	Emilia-Romagna	600-300 BCE	MONT_T.015=T.49	MONT-15	Rocca petrosa dx	*
	Emilia-Romagna	600-300 BCE	MONT_T.09=T.43	MONT-09	Rocca petrosa sx	*
	Emilia-Romagna	600-300 BCE	MONT_T.029=T.63	MONT-29	Rocca petrosa dx	*
	Emilia-Romagna	600-300 BCE	MONT_T.6	MONT-6	Rocca petrosa sx	*
	Emilia-Romagna	600-300 BCE	MONT_T.7	MONT-7	Rocca petrosa sx	*
	Emilia-Romagna	600-300 BCE	MONT_T.41	MONT-41	Rocca petrosa dx	*
	Emilia-Romagna	600-300 BCE	MONT_T.31	MONT-31	Rocca petrosa dx	*
Ceretolo (Casalecchio di Reno)	Emilia-Romagna	200-400 CE	CER_T.16	CER-16	Rocca petrosa dx	*
	Emilia-Romagna	200-400 CE	CER_T.12	CER-12	Rocca petrosa°	*
	Emilia-Romagna	200-400 CE	CER_T.13	CER-13	Rocca petrosa sx	*
	Emilia-Romagna	200-400 CE	CER_T.97	CER-97	Rocca petrosa sx	*
	Emilia-Romagna	200-400 CE	CER_T.96	CER-96	Rocca petrosa sx	*
	Emilia-Romagna	200-400 CE	CER_T.8	CER-8	Rocca petrosa dx	*
	Emilia-Romagna	200-400 CE	CER_T.76B	CER-76B	Rocca petrosa dx	*
	Emilia-Romagna	200-400 CE	CER_T.78 US367	CER-78	Rocca petrosa°	*
Norcia (Colle dell'Annunziata)	Umbria	Iron Age	Norcia - Colle dell'Annunziata Tb.16	ANN-16	Rocca petrosa dx	Infante
	Umbria	Iron Age	Norcia - Colle dell'Annunziata Tb.25	ANN-25	Rocca petrosa sx	
	Umbria	Iron Age	Norcia - Colle dell'Annunziata Tb.32	ANN-32	Rocca petrosa dx	Adulto con patologie
Norcia (Opaco)	Umbria	Iron Age	Norcia - Opaco Tb.50	OPACO_Tb.50	Rocca petrosa sx	
	Umbria	Iron Age	Norcia - Opaco Tb.101	OPACO_Tb.101	Rocca petrosa sx	
Norcia (Campo Boario)	Umbria	Iron Age	Norcia - Campo Boario Tb.6	CB-6	Rocca petrosa dx	
	Umbria	Iron Age	Norcia - Campo Boario Tb.15	CB-15	Rocca petrosa dx	
	Umbria	Iron Age	Norcia - Campo Boario Tb.14	CB-14	Rocca petrosa sx	
	Umbria	Iron Age	Norcia - Campo Boario Tb.20	CB-20	Rocca petrosa dx	
Norcia (Santa Scolastica)	Umbria	Iron Age	Norcia - Santa Scolastica Tb. 30-rep. N°3	SCOL-30	Rocca petrosa°	Infante (18 ± 6 mesi)
	Umbria	Iron Age	Norcia - Santa Scolastica Tb. 57	SCOL-57	Rocca petrosa°	Infante (12 ± 6 mesi)
Gubbio (San Biagio)	Umbria	Iron Age	Tomba 4 Saggio G Individuo: T.4-'95	Gubbio-1	Rocca petrosa dx	
	Umbria	Iron Age	Tomba 5 Saggio H	Gubbio-2	Rocca	

Sito	*Regione*	*Periodo*	*ID archeologico*	*IDLab*	*Distretto scheletrico*	*Informazioni*
			Individuo: T.5-'95		petrosa dx	
	Umbria	Iron Age	Tomba 18 Saggio H Individuo: T.18-'95	Gubbio-3	Rocca petrosa dx	
	Umbria	Iron Age	Tomba 21 Saggio G Individuo: T.21-'95	Gubbio-4	Rocca petrosa sx	
	Umbria	Iron Age	Tomba 1 Saggio L Individuo: T.1-'97	Gubbio-5	Rocca petrosa dx	
	Umbria	Iron Age	Tomba 7 Saggio L Individuo: T.7-'97	Gubbio-6	Rocca petrosa sx	
	Umbria	Iron Age	Tomba 9 Saggio L Individuo: T.9-'97	Gubbio-7	Rocca petrosa dx	
	Umbria	Iron Age	Tomba 12 Saggio L Individuo: T.12-'97	Gubbio-8	Rocca petrosa°	
	Umbria	Iron Age	Tomba 18 Saggio L Individuo: T.18-'98	Gubbio-9	Rocca petrosa°	
	Umbria	Iron Age	Tomba 14 Saggio C Individuo: T.14 18/IX/1991	Gubbio-11	Rocca petrosa sx	
	Umbria	Iron Age	Tomba 13 Saggio C Individuo: T.13 16/IX/91	Gubbio-12	Rocca petrosa sx	
	Umbria	Iron Age	Tomba 1 Saggio D Individuo: T.1 14/IX/1992 200517	Gubbio-13	Rocca petrosa sx	
	Umbria	Iron Age	Tomba 5 Saggio D Individuo: T.5 S.D 195565	Gubbio-14	Rocca petrosa sx	
	Umbria	Iron Age	Tomba 6 Saggio D Individuo: T.6 28/IX/92 200521	Gubbio-15	Rocca petrosa dx	
	Umbria	Iron Age	Tomba 1 Saggio C Individuo: T.1 S.C 195560	Gubbio-16	Rocca petrosa sx	
Matelica (Località Crocifisso)	Marche	Iron Age	Matelica- Località Crocifisso Tb.54	MAT-54	Rocca petrosa dx	Adulto con patologie
	Marche	Iron Age	Matelica- Località Crocifisso Tb.58	MAT-58	Rocca petrosa dx	Consolidata
	Marche	Iron Age	Matelica- Località Crocifisso Tb.136	MAT-136	Rocca petrosa°	
	Marche	Iron Age	Matelica- Località Crocifisso Tb.75	MAT-75	Rocca petrosa dx	
	Marche	Iron Age	Matelica- Località Crocifisso Tb.72	MAT-72	Rocca petrosa°	Infante
	Marche	Iron Age	Matelica- Località Crocifisso Tb.68	MAT-68	Rocca petrosa sx	
	Marche	Iron Age	Matelica- Località Crocifisso Tb.21bis	MAT-21	Rocca petrosa sx	Consolidata
	Marche	Iron Age	Matelica- Località Crocifisso Tb.50	MAT-50	Rocca petrosa°	
	Marche	Iron Age	Matelica- Località Crocifisso Tb.28	MAT-28	Rocca petrosa sx	
	Marche	Iron Age	Matelica- Località Crocifisso Tb.30	MAT-30	Rocca petrosa°	
	Marche	Iron Age	Matelica- Località Crocifisso Tb.116	MAT-116	Rocca petrosa dx	
	Marche	Iron Age	Matelica- Località Crocifisso Tb.47	MAT-47	Rocca petrosa°	
	Marche	Iron Age	Matelica- Località Crocifisso Tb.52	MAT-52	Rocca petrosa°	
	Marche	Iron Age	Matelica- Località Crocifisso Tb.112	MAT-112	Rocca petrosa°	
	Marche	Iron Age	Matelica- Località	MAT-63	Rocca	

Sito	*Regione*	*Periodo*	*ID archeologico*	*IDLab*	*Distretto scheletrico*	*Informazioni*
	Marche	Iron Age	Crocifisso Tb.63 Matelica- Località Crocifisso Tb.3	MAT-3	petrosa dx Rocca petrosa dx	
	Marche	Iron Age	Matelica- Località Crocifisso Tb.80	MAT-80	Rocca petrosa dx	
	Marche	Iron Age	Matelica- Località Crocifisso Tb.121	MAT-121	Rocca petrosa°	
Castiglione	Lazio	1000-800 BCE	Castiglione IJ14	C_IJ14	Rocca petrosa	
	Lazio	1000-800 BCE	Castiglione IJ15-16	C_IJ15-16	Rocca petrosa	
	Lazio	1000-800 BCE	Castiglione t.15	C_t15	Rocca petrosa	
	Lazio	1000-800 BCE	Castiglione t.30	C_t30	Rocca petrosa	
	Lazio	1000-800 BCE	Castiglione t.40	C_t40	Rocca petrosa	
	Lazio	1000-800 BCE	Castiglione t.46	C_t46	Rocca petrosa	
	Lazio	1000-800 BCE	Castiglione t.48	C_t48	Rocca petrosa	
	Lazio	1000-800 BCE	Castiglione t.57	C_t57	Rocca petrosa	
	Lazio	1000-800 BCE	Castiglione t.58	C_t58	Rocca petrosa	
	Lazio	1000-800 BCE	Castiglione t.63	C_t63	Rocca petrosa	
	Lazio	1000-800 BCE	Castiglione t.71	C_t71	Rocca petrosa	
	Lazio	1000-800 BCE	Castiglione t.79	C_t79	Rocca petrosa	
	Lazio	1000-800 BCE	Castiglione t.83	C_t83	Rocca petrosa	
Osteria dell'Osa	Lazio	900-600 BCE	Osteria dell'Osa t.359	Osa-359	Rocca petrosa	
	Lazio	900-600 BCE	Osteria dell'Osa t.361	Osa-361	Rocca petrosa	
	Lazio	900-600 BCE	Osteria dell'Osa t.378	Osa-378	Rocca petrosa	
	Lazio	900-600 BCE	Osteria dell'Osa t.417	Osa-417	Rocca petrosa	
	Lazio	900-600 BCE	Osteria dell'Osa t.419	Osa-419	Rocca petrosa	
	Lazio	900-600 BCE	Osteria dell'Osa t.431	Osa-431	Rocca petrosa	
	Lazio	900-600 BCE	Osteria dell'Osa t.438	Osa-438	Rocca petrosa	
	Lazio	900-600 BCE	Osteria dell'Osa t.445	Osa-445	Rocca petrosa	
	Lazio	900-600 BCE	Osteria dell'Osa t.446	Osa-446	Rocca petrosa	
	Lazio	900-600 BCE	Osteria dell'Osa t.488	Osa-488	Rocca petrosa	
Polizzello	Sicilia	Iron Age	Polizzello 4	Pol-4	Rocca petrosa	*
	Sicilia	Iron Age	Polizzello 26	Pol-26	Rocca petrosa	*
	Sicilia	Iron Age	Polizzello 22	Pol-22	Rocca petrosa	*
	Sicilia	Iron Age	Polizzello 17	Pol-17	Rocca petrosa	*
	Sicilia	Iron Age	Polizzello 15	Pol-15	Rocca	*

Sito	*Regione*	*Periodo*	*ID archeologico*	*IDLab*	*Distretto scheletrico*	*Informazioni*
	Sicilia	Iron Age	Polizzello 12	Pol-12	petrosa Rocca petrosa	*
	Sicilia	Iron Age	Polizzello 10	Pol-10	Rocca petrosa	*
	Sicilia	Iron Age	Polizzello 2	Pol-2	Rocca petrosa	
	Sicilia	Iron Age	Polizzello 8	Pol-8	Rocca petrosa	
	Sicilia	Iron Age	Polizzello 9	Pol-9	Rocca petrosa	
	Sicilia	Iron Age	Polizzello 11	Pol-11	Rocca petrosa	
	Sicilia	Iron Age	Polizzello 13	Pol-13	Rocca petrosa	
	Sicilia	Iron Age	Polizzello 16	Pol-16	Rocca petrosa	
	Sicilia	Iron Age	Polizzello 19	Pol-19	Rocca petrosa	
	Sicilia	Iron Age	Polizzello 20	Pol-20	Rocca petrosa	
	Sicilia	Iron Age	Polizzello 21	Pol-21	Rocca petrosa	
	Sicilia	Iron Age	Polizzello 23	Pol-23	Rocca petrosa	
	Sicilia	Iron Age	Polizzello 24	Pol-24	Rocca petrosa	

Tabella A 2: Lista delle coppie di indici associate a ciascun campione per il sequenziamento in parallelo.

LabID	***index P5***	***index P7***
MONT-13	CTAAGCCT	GGAGCTAC
MONT-15	CGTCTAAT	GCGTAGTA
MONT-09	AAGGAGTA	GGACTCCT
MONT-29	CTAAGCCT	TAGGCATG
MONT-6	TCTCTCCG	CGAGGCTG
MONT-7	TCGACTAG	AAGAGGCA
MONT-41	TTCTAGCT	GTAGAGGA
MONT-31	TCGACTAG	TACGCTGC
CER-16	GAGCCTTA	GCGTAGTA
CER-12	CTCTCTAT	TAGGCATG
CER-13	AAGGAGTA	GTAGAGGA
CER-97	CGTCTAAT	ATCTCAGG
CER-96	TCTCTCCG	ACTCGCTA
CER-8	TCGACTAG	GGAGCTAC
CER-76B	CCTAGAGT	CGGAGCCT
CER-78	CTAAGCCT	TGCAGCTA
ANN-16	AATCGCGC	AATGCTTC
ANN-25	AATCTTCA	ACCATTAC
ANN-32	AATGCAAC	ACCGGCCG
OPACO_Tb.50	ACGTCCAT	ATCTCTAG
OPACO_Tb.101	TTATATAC	AACCAGAT
CB-6	CATAGTTC	CAATCGGT
CB-15	CCATAGCA	CCGAGCAT
CB-14	CCATGCTC	CGAGTTCC
CB-20	CCGAGCGG	CGGCGGTT
SCOL-30	TTCAAGCC	AACGGATT
SCOL-57	GCAATCTA	GATATTGA
Gubbio-1	GGACTTGA	GGATTGGA
Gubbio-2	GGATCAGG	GGTCGTCA
Gubbio-3	GGTCAAGT	GTTCAGTC
Gubbio-4	GTCATATT	TAATAACG
Gubbio-5	TGAGGCCA	TCCGCATG
Gubbio-6	TGCAGGTA	TCGGACTC
Gubbio-7	TTATATAC	TGATACGC
Gubbio-8	TTCAAGCC	TTGCGGCA
Gubbio-9	TTCGGAAG	AACCAGAT
Gubbio-11	TTCTCTCG	AACGGATT
Gubbio-12	AATCGCGC	ATCTCTAG
Gubbio-13	AATCTTCA	ATTATTCG
Gubbio-14	AATGCAAC	CAATCGGT
Gubbio-15	ACGTCCAT	CCGAGCAT
Gubbio-16	ATGGCGTT	CGAGTTCC
MAT-54	GGACTTGA	GCTCAAGG
MAT-58	GGATCAGG	GGATCCAT
MAT-136	GGTCAAGT	GGATTGGA
MAT-75	GTCATATT	GGTCGTCA
MAT-72	TGAGGCCA	GTTCAGTC

LabID	*index P5*	*index P7*
MAT-68	TGCAGGTA	TAATAACG
MAT-21	TTATATAC	TCCGCATG
MAT-50	TTCAAGCC	TCGGACTC
MAT-28	TTCGGAAG	TGATACGC
MAT-30	GGACTTGA	GGATCCAT
MAT-116	GGATCAGG	GGATTGGA
MAT-47	GGTCAAGT	GGTCGTCA
MAT-52	GTCATATT	GTTCAGTC
MAT-112	TGAGGCCA	TAATAACG
MAT-63	TGCAGGTA	TCCGCATG
MAT-3	TTATATAC	TCGGACTC
MAT-80	TTCAAGCC	TGATACGC
MAT-121	TTCGGAAG	TTGCGGCA
C_IJ14	TTCTCTCG	AACCAGAT
C_IJ15-16	CATAGTTC	CGGCGGTT
C_t15	CCATAGCA	CTGGTAAC
C_t30	CCATGCTC	CTTACCGT
C_t40	GACGTTGG	GGATTGGA
C_t46	CCTGCCGT	GATATTGA
C_t48	GCAATCTA	GGATTGGA
C_t57	GGACTTGA	GGTCGTCA
C_t58	GGATCAGG	GTTCAGTC
C_t63	GGTCAAGT	TAATAACG
C_t71	GTCATATT	TCCGCATG
C_t79	TGAGGCCA	TCGGACTC
C_t83	TGCAGGTA	TGATACGC
Osa-359	TTATATAC	TTGCGGCA
Osa-361	TTCAAGCC	AACCAGAT
Osa-378	TTCGGAAG	AACGGATT
Osa-417	GCAATCTA	GGTCGTCA
Osa-419	GGACTTGA	GTTCAGTC
Osa-431	GGATCAGG	TAATAACG
Osa-438	GGTCAAGT	TCCGCATG
Osa-445	GTCATATT	TCGGACTC
Osa-446	TGAGGCCA	TGATACGC
Osa-488	TGCAGGTA	TTGCGGCA
Pol 4	GTCATATT	GTTCAGTC
Pol 26	GGATCAGG	GGTCGTCA
Pol 22	CGGAAGCT	GCTCAAGG
Pol 17	GACGTTGG	GGATCCAT
Pol 15	CCTGCCGT	CTTACCGT
Pol 12	CGGAAGCT	GAGACCTA
Pol 10	AAGCAGCC	ATCTCTAG
Pol-2	AAGCAGCC	ACCGGCCG
Pol-8	AAGTTACG	ATCTCTAG
Pol-9	AATCGCGC	ATTATTCG
Pol-11	AATCTTCA	CAATCGGT
Pol-13	AATGCAAC	CCGAGCAT

LabID	***index P5***	***index P7***
Pol-16	ACGTCCAT	CGAGTTCC
Pol-19	ATGGCGTT	CGGCGGTT
Pol-20	CATAGTTC	CTGGTAAC
Pol-21	CCATAGCA	CTTACCGT
Pol-23	CCATGCTC	GAGACCTA
Pol-24	CCGAGCGG	GATATTGA

Tabella A 3: Frequenze in percentuale degli hg del Y-chr di 1735 campioni (inclusi quelli processati nel presente studio) utilizzati per la costruzione della PCA.

Country	Population	Pop. ID	TOT	BT	CT	C	E1a	E1b	F	G	G2a	G2a2a	G2a2b	G2a2b2a	G2a2b2b	IJK	H	I	I1	I2	I2a1	I2a2	J	J1	J2	J2a	J2b	L	L1a	P	Q1	R	R1	R1a	R1b	R1b1	R1b1a	R1b1b	T1a	Reference
Italy	Centro-Nord_Italy_IronAge	CNIT_IA	21	0	0	0	0	0	0	0	0	0	0	9.52	0	0	0	0	0	0	0	4.76	28.57	0	0	0	0	0	0	0	0	0	4.76	0	0	0	52.38	0	0	this study
Italy	Centro-Sud_Italy_IronAge	CSIT_IA	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	this study
Italy	Sicily_Iron Age	Sy_IA	6	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	this study
Italy	Italy_IronAge	IT_IA	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	150
France	Europe_IronAge	E_IA	8	0	0	0	0	12.5	0	0	12.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12.50	0	0	0	12.5	50	0	0	262; 144
Moldova-Ukraine-Russia	IronAge-Nomads	Nom_IA	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16.67	0	0	25.00	0	0	58.33	0	0	263; 144
Italy	Sicily_BronzeAge	Sy_BA	9	0	0	0	0	0	0	0	0	0	0	33.33	0	0	0	0	0	0	0	0	11.11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	55.56	0	0	264
Italy	Sardinia_BronzeAge	Sa_BA	32	0	0	0	0	0	0	0	0	3.13	0	0	12.50	0	3.13	0	0	3.13	25.00	0	0	0	0	0	0	12.50	0	0	0	0	3.13	3.13	0	0	3.13	3.13	28.13	0	264;265; 266
Bulgaria-Croatia	Balkans_BronzeAge	B_BA	9	0	0	0	0	0	0	0	0	11.11	0	0	0	0	11.11	11.11	0	0	0	44.44	0	0	0	0	11.11	0	0	0	0	0	0	11.11	0	0	0	0	0	66
Germany-France-Poland-Hungary	Centre_Europe_Bronz Age	CE_BA	69	1.45	0	0	0	0	0	0	1.45	5.80	0	0	0	0	1.45	1.45	1.45	1.45	2.90	4.35	0	0	0	1.45	0	0	0	0	0	2.90	2.90	2.90	1.45	4.35	62.32	0	0	138;262;267; 268;22;50;269
England-Ireland	UK_Bronz Age	UK_BA	46	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4.35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95.65	0	0	270;269
Spain-Portugal	West_Europe_BronzAge	WE_BA	35	2.86	2.86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2.86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	91.43	0	0	277; 269;271;272
Russia	Yamnaya_BronzeAge	Yam	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	138; 22;273

Lebanon	Lebanon_BronzeAge	L_BA	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	50	50	0	0	0	0	0	0	0	0	0	0	0	0	0	104
Russia	european-Steppe_BronzeAge	St_BA	37	0	0	0	0	0	0	0	0	2.70	0	0	0	0	0	0	0	0	0	0	2.70	0	2.70	0	5.41	2.70	8.11	0	2.7	5.41	0	5.41	27.03	0	0	35.14	0	0	138;50;273
Czech Rep-Russia	BronzeAge-IronAge_Transition	BIT	5	0	0	0	20	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	274;275
Italy	Sicily_CopperAge	Sy_CA	4	0	0	0	0	0	0	0	0	0	0	0	0	0	25.00	75.00	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	138
Italy	Sardinia_CopperAge	Sa_CA	16	0	0	0	0	6.25	0	6.25	0	37.50	0	0	0	0	6.25	18.75	0	0	6.25	12.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6.25	0	150;265;266
Bulgaria-Croatia	Balkans_CopperAge	B_CA	10	0	20	0	0	0	0	10	0	10	0	10	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	10	0	0	0	20	0	0	66
Hungary	Centre_Europe_CopperAge	CE_CA	19	0	0	10.53	0	0	0	0	0	5.26	10.53	31.58	0	0	0	5.26	0	21.05	10.53	5.26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	276
Spain-Portugal	Europe_CopperAge	E_CA	73	0	5.48	0	0	1.37	2.74	1.37	2.74	6.85	1.37	2.74	1.37	0	1.37	10.96	0	1.37	15.07	28.77	0	0	0	0	0	0	0	0	2.7	0	0	0	0	0	2.74	10.96	0	0	136;22;276;277;50;269;272
Russia	european-Steppe_CopperAge	St_CA	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14.29	0	14.29	0	0	0	0	0	14.29	0	0	14.29	0	42.86	0	0	0	50;273
Russia	Steppe_CopperAge-BronzeAge_Transition	St_CBT	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	278
Italy	Italy_Neolithic	IT_N	4	0	0	25.00	0	0	0	25.0	0	0	0	0	0	0	0	0	0	0	0	0	50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	150
Bulgaria-Croatia	Balkans_Neolithic	B_N	16	0	0	25.00	0	6.25	0	0	0	18.75	18.75	0	0	0	0	6.25	0	0	0	6.25	0	0	0	6.25	0	0	0	0	0	0	0	0	0	0	0	0	12.50	66
Germany-France-Poland-Hungary	Centre_Europe_Neolithic	CE_N	201	0.50	1.99	2.99	0	2.99	1.49	2.99	12.44	9.45	1.49	8.96	1.99	1.00	11.94	3.98	0.50	5.47	13.93	7.96	0	0	0	0.50	0	0	0	0	0	0.50	0	0	0	0	5.97	0	1.00	262;157;268;22;279;276;50;269; 280;281

Country	Population	Pop. ID	TOT	BT	CT	C	E1a	E1b	F	G	G2a	G2a2a	G2a2b	G2a2b2a	G2a2b2b	IJK	H	I	I1	I2	I2a1	I2a2	J	J1	J2	J2a	J2b	L	L1a	P	Q1	R	R1	R1a	R1b	R1b1	R1b1a	R1b1b	T1a	Reference
England-Ireland-Scotland	UK_Neolithic	UK_N	52	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21.15	0	1.92	28.85	48.08	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	282;283; 269
Spain-Portugal	West_Europe_Neolithic	WE_N	45	0	0	0	0	2.22	4.44	0	11.11	13.33	0	0	2.22	0	11.11	8.89	0	2.22	13.33	24.44	0	0	0	0	0	0	0	0	0	0	0	0	2.22	0	4.44	0	0	279;277;50;271; 272; 284
Lithuania-Ukraine	Est_Europe_Neolithic	EE_N	33	0	0	0	0	0	0	0	3.03	0	0	6.06	0	0	0	21.21	0	3.03	6.06	24.24	0	0	0	0	0	0	0	0	0	6.06	9.09	0	0	6.06	15.15	0	0	66;285
Europe	Europe_Neolithic	E_N	360	0.28	1.11	0	0	0	0	0	8.89	14.44	0.83	7.22	2.50	0	6.39	9.17	0	4.44	16.11	17.50	0.56	0	0.83	0.28	0	0	0	0	0.56	0.83	0.83	0	0.28	0.56	6.39	0	0	150;282; 262;157;268; 22;132;286;279;276;266;287;50;285;269;280;278;283;281;272;284
Europe	Europe_Neolithic-BronzeAge_Transition	E_NBT	49	2.04	4.08	0	0	0	0	0	0	2.04	0	0	0	0	0	6.12	0	6.12	4.08	20.41	0	0	0	0	0	0	0	4.08	0	0	0	40.82	0	0	10.20	0	0	138;267;22;288;277;50;285;269;289
Italy	Italy_Roman_Period	IT_R	31	0	0	0	0	0	0	16.13	0	0	0	0	0	0	3.23	6.45	0	0	0	0	54.84	0	0	0	0	0	0	0	0	19.35	0	0	0	0	0	0	0	150
Europe	Europe_modern_1000genomes	E_M	112	0	0	0	0	0	0	0	1.79	0	0	0	0	0	0	0	13.39	0	1.79	1.79	0	1.79	0	3.57	2.68	0	0	0	0	0	0	7.14	0	0	66.07	0	0	103
Italy	Tuscans_modern_1000genomes	T_M	46	0	0	0	0	0	0	0	4.35	0	0	0	0	0	0	0	4.35	0	0	0	0	2.17	0	23.91	8.70	0	0	0	0	0	0	4.35	0	0	52.17	0	0	103
Italy	Sicily-sud_Modern	SS_M	326	0	0	0	0.31	19.33	0	0.61	12.27	0	0	0	0	0	0	0	1.23	4.60	0	0	0	4.91	0.61	15.95	3.37	0.31	0	0.31	0	0	0	5.21	28.53	0	0	0	2.45	113

Tabella A 4: Risultati del sequenziamento *shotgun* dei 78 campioni italiani dell'età del Ferro. Sono mostrate le principali statistiche usate per valutare la qualità delle librerie genomiche prodotte. In rosso i campioni che non hanno prodotto statistiche di bassa qualità.

IDLab	*# reads processate*	*# reads post-Clip&Merge*	*# ReadsMerged*	*% Merged*	*# Reads mappanti*	*% Mappanti*	*# Duplicati rimossi*	*Cluster Factor*	*% DNA endogeno*
ANN-16	10464760	5123327	3991717	77.91	29559	0.74	2365	1.124	0.741
ANN-25	13013116	6712672	5006599	74.58	1322325	26.411	126806	1.142	26.412
ANN-32	12768534	6659540	5045596	75.76	3334656	66.09	289082	1.133	66.09
OPACO_Tb.50	13012584	6341360	5000835	78.86	2846	0.056	100	1.075	0.057
OPACO_Tb.101	17496238	8839702	6274581	70.98	2555707	40.731	251352	1.149	40.731
CB-6	16509976	8203658	6540507	79.73	3497733	53.478	324956	1.148	53.478
CB-15	15852204	8106210	6447232	79.53	4619424	71.649	469243	1.157	71.65
CB-14	15332790	7735541	5896183	76.22	922901	15.652	88792	1.142	15.653
CB-20	13762544	6955909	5576475	80.17	3744322	67.144	357329	1.16	67.145
SCOL-30	11854972	5997665	4109764	68.52	1494639	36.368	142923	1.146	36.368
SCOL-57	12211910	6742305	4007354	59.44	1380653	34.453	124569	1.133	34.453
Gubbio-1	11331510	5875054	3606283	61.38	1376891	38.18	125905	1.138	38.18
Gubbio-2	10129334	4921548	3462049	70.34	1337828	38.643	125517	1.138	38.643
Gubbio-3	12088714	6467212	4256141	65.81	2394916	56.27	228002	1.14	56.27
Gubbio-4	8594100	4154624	3024682	72.8	573960	18.976	53594	1.142	18.976
Gubbio-5	12935634	6996537	3728797	53.29	367682	9.861	35216	1.148	9.861
Gubbio-6	15804268	8016818	5332881	66.52	2508	0.047	86	1.059	0.047
Gubbio-7	11361034	5897506	3974538	67.39	1751380	44.065	165087	1.139	44.065
Gubbio-8	15975900	8060891	5697636	70.68	1501492	26.353	127727	1.129	26.353
Gubbio-9	9773084	5179091	3230690	62.38	31440	0.973	2991	1.153	0.973
Gubbio-11	11845234	6575489	3442960	52.36	15632	0.454	1478	1.149	0.454
Gubbio-12	14005284	6675322	5138621	76.98	1827602	35.566	151556	1.128	35.566
Gubbio-13	13030574	6502084	4690466	72.14	1449473	30.903	124425	1.126	30.903
Gubbio-14	8931310	4071975	3121391	76.66	327139	10.481	31716	1.152	10.481
Gubbio-15	20890770	9764375	7641617	78.26	2360538	30.891	256408	1.174	30.891
Gubbio-16	8486798	4051157	3206796	79.16	720564	22.47	71760	1.152	22.47
MAT-54	10729442	5788205	3767764	65.09	1244445	33.029	118190	1.142	33.029
MAT-58	11093582	6002934	3727027	62.09	661703	17.754	71215	1.163	17.754
MAT-136	10824030	5438380	3874821	71.25	213289	5.504	19400	1.138	5.504
MAT-75	10531078	5818850	3503459	60.21	570158	16.274	56672	1.147	16.274
MAT-72	10853128	5530134	3803825	68.78	388053	10.202	35844	1.134	10.202
MAT-68	8690616	4806381	2819721	58.67	514101	18.232	47102	1.135	18.232
MAT-21	8646100	4710115	3117000	66.18	1814251	58.205	184014	1.152	58.205
MAT-50	11234000	5788108	4018078	69.42	755746	18.809	69972	1.136	18.809
MAT-28	11867084	6313896	4223334	66.89	1254556	29.705	114787	1.139	29.705
MAT-30	18054708	9128520	6480155	70.99	2387645	36.845	244471	1.16	36.845
MAT-116	11570080	6203473	3784088	61	1268134	33.512	114945	1.132	33.512
MAT-47	12605346	6326062	4192650	66.28	213342	5.088	20871	1.145	5.088
MAT-52	11584976	6148355	3927975	63.89	1236935	31.49	113628	1.139	31.49
MAT-112	11614450	5626591	4193763	74.53	763478	18.205	64448	1.133	18.205
MAT-63	13248442	5548867	4655595	83.9	948871	20.381	92663	1.149	20.381
MAT-3	10858702	5422771	3963885	73.1	1500511	37.855	142628	1.143	37.855
MAT-80	10651970	5737436	3514468	61.26	955729	27.194	84863	1.13	27.194
MAT-121	14676898	7339118	4836392	65.9	392978	8.125	34639	1.136	8.125
C_IJ14	16343250	7872371	5365539	68.16	776128	14.465	70825	1.149	9.859

IDLab	*# reads processate*	*# reads post-Clip&Merge*	*# ReadsMerged*	*% Merged*	*# Reads mappanti*	*% Mappanti*	*# Duplicati rimossi*	*Cluster Factor*	*% DNA endogeno*
C_IJ15-16	15408982	7356107	5720853	77.77	4325466	75.609	417106	1.155	58.801
C_t15	13612638	6776347	4297691	63.42	721981	16.799	67699	1.151	10.654
C_t30	12753394	6077701	4155972	68.38	712116	17.135	63308	1.152	11.717
C_t40	4778584	2252085	1504527	66.81	195439	12.99	2844	1.195	8.678
C_t46	14540064	6968565	4661623	66.9	467216	10.023	39679	1.141	6.705
C_t48	14987478	7318969	5023111	68.63	748579	14.903	60906	1.132	10.228
C_t57	10440374	4695138	3521760	75.01	126472	3.591	9831	1.143	2.694
C_t58	13696528	6456147	5059060	78.36	332033	6.563	26345	1.131	5.143
C_t63	14335646	6712248	4743267	70.67	146652	3.092	10588	1.14	2.185
C_t71	12862890	6297758	4184008	66.44	687008	16.42	64135	1.156	10.909
C_t79	10123200	4770529	3580800	75.06	996084	27.817	94603	1.152	20.88
C_t83	18710824	9130695	6196384	67.86	798189	12.882	68004	1.138	8.742
Osa-359	10563876	5360292	3222090	60.11	3683	0.114	72	1.034	0.114
Osa-361	14810520	6377761	4949106	77.6	5334	0.108	185	1.055	0.108
Osa-378	19138370	9050160	6367763	70.36	1997	0.031	82	1.087	0.031
Osa-417	15162178	7379644	5151231	69.8	16383	0.318	1054	1.116	0.318
Osa-419	12677952	6404550	4038401	63.06	4248	0.105	75	1.031	0.105
Osa-431	8493420	3904643	2982023	76.37	1951	0.065	35	1.032	0.065
Osa-438	15899504	7180471	5492584	76.49	3383	0.062	93	1.046	0.062
Osa-445	16527270	7613472	5887065	77.32	3066	0.052	57	1.036	0.052
Osa-446	18070858	8035104	6272295	78.06	4035	0.064	108	1.048	0.064
Osa-488	16615798	7775695	5410845	69.59	2745	0.0507	83	1.055	0.051
Pol-2	7193130	3517146	2805319	79.76	2452875	87.437	178072	1.121	87.437
Pol-8	12894156	5733357	4043779	70.53	17553	0.434	1494	1.138	0.434
Pol-9	6431680	3118543	2477263	79.44	1857406	74.978	147856	1.125	74.978
Pol-11	15411564	7181032	5680227	79.1	3672579	64.655	336093	1.144	64.655
Pol-13	9375118	4800789	3584821	74.67	2780380	77.56	278241	1.167	77.56
Pol-16	16993794	8363106	5841264	69.85	570408	9.765	56802	1.153	9.765
Pol-19	12769204	6711632	4064824	60.56	2148	0.053	61	1.068	0.053
Pol-20	11441592	5446911	3635368	66.74	8201	0.226	564	1.127	0.226
Pol-21	8236008	3664676	2718344	74.18	385777	14.192	38875	1.151	14.192
Pol-23	12805278	5984874	4383438	73.24	1926502	43.95	192660	1.15	43.95
Pol-24	8449442	3965261	2959926	74.65	341617	11.541	33088	1.143	11.541

Tabella A 5: Risultati di autenticazione del dato.

IDLab	*MT/NUC*	*Deaminazione 3'*	*Deaminazione 5'*	*Lunghezza media*
ANN-16	77.95	0.093	0.095	43.53
ANN-25	53.55	0.126	0.127	49.67
ANN-32	106.79	0.103	0.102	50.69
OPACO_Tb.50	0	0.121	0.101	42.67
OPACO_Tb.101	130.46	0.044	0.044	51.25
CB-6	80.22	0.117	0.117	47.73
CB-15	85.96	0.097	0.098	50.68
CB-14	66.74	0.111	0.111	51.61
CB-20	89.32	0.097	0.097	50.91
SCOL-30	159.83	0.049	0.05	50.08
SCOL-57	98.34	0.051	0.051	51.43
Gubbio-1	109.96	0.074	0.074	51.5
Gubbio-2	162.78	0.111	0.113	47.31
Gubbio-3	122.33	0.105	0.101	50.52
Gubbio-4	95.46	0.087	0.091	50.42
Gubbio-5	97.62	0.131	0.13	45.92
Gubbio-6	0	0.072	0.094	44.31
Gubbio-7	208.84	0.081	0.082	50.33
Gubbio-8	103.88	0.105	0.105	49.23
Gubbio-9	90.4	0.123	0.126	39.75
Gubbio-11	0	0.125	0.116	43.11
Gubbio-12	86.85	0.119	0.119	46.67
Gubbio-13	77.36	0.106	0.108	48.81
Gubbio-14	102.13	0.087	0.086	48.97
Gubbio-15	87.78	0.112	0.114	46.19
Gubbio-16	108.93	0.093	0.091	47.74
MAT-54	74.89	0.091	0.091	50.85
MAT-58	50.08	0.073	0.075	51.44
MAT-136	73.61	0.115	0.112	48.6
MAT-75	99.67	0.081	0.082	49.54
MAT-72	157.47	0.074	0.071	49.57
MAT-68	94.92	0.081	0.08	49.62
MAT-21	99.49	0.094	0.092	49.55
MAT-50	116.3	0.067	0.07	51.59
MAT-28	71.97	0.104	0.104	49.61
MAT-30	139.97	0.088	0.088	47.47
MAT-116	104.66	0.095	0.094	49.3
MAT-47	113.72	0.081	0.081	47.72
MAT-52	100.71	0.144	0.136	46.91
MAT-112	81.72	0.123	0.124	46.66
MAT-63	116.87	0.099	0.101	47.43
MAT-3	83.36	0.135	0.135	46.51
MAT-80	76.55	0.107	0.105	48.21
MAT-121	70.47	0.122	0.124	44.27
C_IJ14	-	0.099	0.105	44
C_IJ15-16	-	0.1	0.105	45
C_t15	-	0.078	0.077	43
C_t30	-	0.132	0.136	41
C_t40	-	0.08	0.09	40
C_t46	-	0.121	0.12	38
C_t48	-	0.073	0.077	44
C_t57	-	0.108	0.106	40
C_t58	-	0.146	0.154	43
C_t63	-	0.15	0.147	39
C_t71	-	0.116	0.119	42
C_t79	-	0.063	0.068	46
C_t83	-	0.084	0.087	44
Osa-359	220.14	0.096	0.077	47.36
Osa-361	95.58	0.07	0.088	45.91
Osa-378	0	0.04	0.06	41.44
Osa-417	54.02	0.165	0.179	38.74
Osa-419	0	0.097	0.085	46.37
Osa-431	0	0.113	0.105	44.3
Osa-438	0	0.076	0.073	44.87
Osa-445	0	0.097	0.071	43.3
Osa-446	61.68	0.069	0.08	44.14
Osa-488	0	0.121	0.09	42.29
Pol-2	60.4	0.097	0.096	49.52
Pol-8	99.13	0.138	0.124	40.55
Pol-9	88.65	0.103	0.104	48.97
Pol-11	112.44	0.11	0.109	46.83

IDLab	*MT/NUC*	*Deaminazione 3'*	*Deaminazione 5'*	*Lunghezza media*
Pol-13	70.52	0.097	0.099	49.66
Pol-16	122.49	0.121	0.122	45.02
Pol-19	0	0.069	0.051	44.2
Pol-20	37.37	0.098	0.122	39.4
Pol-21	99.59	0.086	0.086	46.35
Pol-23	71.23	0.071	0.072	47.45
Pol-24	73.08	0.097	0.094	47.95

Tabella A 6: Risultati della determinazione del sesso. Le colonne mostrano il numero totale di sequenze di partenza (Nseqs), il numero di *reads* allineate sui cromosomi sessuali (NchrY+NchrX), quelle allineate solamente sul Y-chr (NchrY), il rapporto tra i valori precedenti (R_y), l'intervallo di confidenza (95%CI), ed infine l'assegnazione del sesso fornita dallo *script*.

IDLab	*Nseqs*	*NchrY+NchrX*	*NchrY*	*R_y*	*95%CI*	*Sesso*
ANN-16	19099	575	47	0.0817	0.0593-0.1041	consistent with XY but not XX
ANN-25	892447	43829	158	0.0036	0.003-0.0042	XX
ANN-32	2165942	58874	4876	0.0828	0.0806-0.085	XY
OPACO_Tb.50	-	-	-	-	-	-
OPACO_Tb.101	1692139	79658	248	0.0031	0.0027-0.0035	XX
CB-6	2201897	103477	379	0.0037	0.0033-0.004	XX
CB-15	2985125	79911	6730	0.0842	0.0823-0.0861	XY
CB-14	627246	17401	1371	0.0788	0.0748-0.0828	consistent with XY but not XX
CB-20	2231879	59402	5049	0.085	0.0828-0.0872	XY
SCOL-30	980448	45653	140	0.0031	0.0026-0.0036	XX
SCOL-57	939882	44692	138	0.0031	0.0026-0.0036	XX
Gubbio-1	915386	24508	1986	0.081	0.0776-0.0845	XY
Gubbio-2	909053	44963	103	0.0023	0.0018-0.0027	XX
Gubbio-3	1632008	44935	3738	0.0832	0.0806-0.0857	XY
Gubbio-4	378250	17782	67	0.0038	0.0029-0.0047	XX
Gubbio-5	237936	11297	38	0.0034	0.0023-0.0044	XX
Gubbio-6	1465	58	2	0.0345	-0.0125-0.0814	Not Assigned
Gubbio-7	1189614	32809	2609	0.0795	0.0766-0.0824	XY
Gubbio-8	989702	26598	2220	0.0835	0.0801-0.0868	XY
Gubbio-9	19545	535	38	0.071	0.0493-0.0928	consistent with XY but not XX
Gubbio-11	9931	461	1	0.0022	-0.0021-0.0064	XX
Gubbio-12	1187181	56620	203	0.0036	0.0031-0.0041	XX
Gubbio-13	986287	47388	145	0.0031	0.0026-0.0036	XX
Gubbio-14	209167	5553	464	0.0836	0.0763-0.0908	XY
Gubbio-15	1474120	69040	196	0.0028	0.0024-0.0032	XX
Gubbio-16	473337	22448	73	0.0033	0.0025-0.004	XX
MAT-54	834016	22854	1906	0.0834	0.0798-0.087	XY
MAT-58	435768	11862	965	0.0814	0.0764-0.0863	XY
MAT-136	140708	3770	289	0.0767	0.0682-0.0852	consistent with XY but not XX
MAT-75	386774	10464	817	0.0781	0.0729-0.0832	consistent with XY but not XX
MAT-72	267344	7374	611	0.0829	0.0766-0.0892	XY
MAT-68	348322	9387	758	0.0807	0.0752-0.0863	XY
MAT-21	1213642	58734	170	0.0029	0.0025-0.0033	XX
MAT-50	512622	14069	1173	0.0834	0.0788-0.0879	XY
MAT-28	823018	22223	1863	0.0838	0.0802-0.0875	XY
MAT-30	1529835	71889	273	0.0038	0.0033-0.0042	XX
MAT-116	869753	43103	117	0.0027	0.0022-0.0032	XX
MAT-47	144060	7083	31	0.0044	0.0028-0.0059	XX
MAT-52	818156	22643	1917	0.0847	0.081-0.0883	XY
MAT-112	485359	12826	1022	0.0797	0.075-0.0844	consistent with XY but not XX
MAT-63	620417	16430	1347	0.082	0.0778-0.0862	XY
MAT-3	998993	48295	168	0.0035	0.003-0.004	XX
MAT-80	653165	31959	98	0.0031	0.0025-0.0037	XX
MAT-121	255404	7088	578	0.0815	0.0752-0.0879	XY
C_IJ14	476735	12774	968	0.0758	0.0712-0.0804	consistent with XY but not XX
C_IJ15-16	2695220	129900	439	0.0034	0.0031-0.0037	XX
C_t15	448385	21449	71	0.0033	0.0025-0.0041	XX
C_t30	415327	10903	900	0.0825	0.0774-0.0877	XY
C_t40	14616	461	f	0.0803	0.0555-0.1051	consistent with XY but not XX
C_t46	280847	13259	40	0.003	0.0021-0.004	XX
C_t48	461100	12005	956	0.0796	0.0748-0.0845	consistent with XY but not XX

IDLab	*Nseqs*	*NchrY+NchrX*	*NchrY*	*R_y*	*95%CI*	*Sesso*
C_t57	68543	3176	11	0.0035	0.0014-0.0055	XX
C_t58	201391	5564	491	0.0882	0.0808-0.0957	XY
C_t63	75823	3643	18	0.0049	0.0027-0.0072	XX
C_t71	410777	11022	898	0.0815	0.0764-0.0866	XY
C_t79	621345	16772	1415	0.0844	0.0802-0.0886	XY
C_t83	491430	13130	1055	0.0804	0.0757-0.085	XY
Osa-359	2100	83	3	0.0361	-0.004-0.0763	Not Assigned
Osa-361	3367	145	3	0.0207	-0.0025-0.0439	consistent with XX but not XY
Osa-378	943	37	3	0.0811	-0.0069-0.169	Not Assigned
Osa-417	9122	251	10	0.0398	0.0156-0.064	consistent with XX but not XY
Osa-419	2417	84	5	0.0595	0.0089-0.1101	Not Assigned
Osa-431	1108	37	2	0.0541	-0.0188-0.1269	Not Assigned
Osa-438	2031	64	2	0.0313	-0.0114-0.0739	consistent with XX but not XY
Osa-445	1605	52	2	0.0385	-0.0138-0.0907	Not Assigned
Osa-446	2264	69	7	0.1014	0.0302-0.1727	consistent with XY but not XX
Osa-488	1504	54	1	0.0185	-0.0174-0.0545	consistent with XX but not XY
Pol-2	1477525	69934	292	0.0042	0.0037-0.0047	XX
Pol-8	10854	541	3	0.0055	-0.0007-0.0118	XX
Pol-9	1180987	32183	2752	0.0855	0.0825-0.0886	XY
Pol-11	2336577	114857	359	0.0031	0.0028-0.0034	XX
Pol-13	1661813	44292	3761	0.0849	0.0823-0.0875	XY
Pol-16	372196	18494	56	0.003	0.0022-0.0038	XX
Pol-19	896	28	0	0	0.0-0.0	consistent with XX
Pol-20	4447	214	0	0	0.0-0.0	consistent with XX
Pol-21	257824	12729	30	0.0024	0.0015-0.0032	XX
Pol-23	1286956	35818	2913	0.0813	0.0785-0.0842	XY
Pol-24	231974	6530	501	0.0767	0.0703-0.0832	consistent with XY but not XX

BIBLIOGRAFIA

[1] Higuchi R., Bowman B., Freiberger M. et al. (1984), *DNA sequences from the quagga, an extinct member of the horse family*. Nature, 312: 282–4.

[2] Pääbo S. (1985), *Molecular cloning of Ancient Egyptian mummy DNA*. Nature 314, 644–645. https://doi.org/10.1038/314644a0.

[3] Rizzi E., Lari M., Gigli E. et al. (2012), *Ancient DNA studies: new perspectives on old samples*. Genet Sel Evol 44, 21. https://doi.org/10.1186/1297-9686-44-21.

[4] Hagelberg E., et Clegg J.B. (1991), *Isolation and characterization of DNA from archaeological bone*. Proc. Biol. Sci., 244: 45-50.

[5] Edwards C.J., *et* al. (2007), *Mitochondrial DNA analysis shows a Near Eastern Neolithic origin for domestic cattle and no indication of domestication of European aurochs*. Proc Biol Sci. 274:1377–1385.

[6] Cooper A., Lalueza-Fox C., Anderson S. et al. (2001), *Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution*. *Nature* 409: 704–707. https://doi.org/10.1038/35055536.

[7] Rogaev EI, Moliaka YK, Malyarchuk BA, et al. (2006), *Complete Mitochondrial Genome and Phylogeny of Pleistocene Mammoth Mammuthus primigenius*. PLoS Biol 4(3): e73. https://doi.org/10.1371/journal.pbio.0040073.

[8] Krings M., Stone A., Schmitz R.W., et al. (1997), *Neandertal DNA sequences and the origin of modern humans.* Cell 90: 19-30. Doi:10.1016/S0092-8674(00)80310-4.

[9] Lin X., Tang W., Ahmad S., et al. (2012), *Applications of targeted gene capture and next-generation sequencing technologies in studies of human deafness and other genetic disabilities*. Hear Res. 288(1-2): 67-76.

[10] Stone A.C., Milner G.R., Paabo S. and Stoneking M. (1996), *Sex determination of ancient human skeletons using DNA*. Am. J. Phys. Anthropol. 99: 231-8.

[11] Gill P., Ivanov P.L., Kimpton C., *et* al. (1994), *Identification of the remains of the Romanov family by DNA analysis.* Nat. Genet. 6: 130-135.

[12] Caramelli D., Lalueza-Fox C., *et* al. (2007), *Genetic analysis of the skeletal remains attribuited to Francesco Petrarca.* Forensic Sci. Int. Feb, 21.

[13] Der Sarkissian C. et al. (2015), *Ancient genomics.* Phil. Trans. R. Soc. B. 370:20130387. http://dx.doi.org/10.1098/rstb.2013.0387.

[14] Hofreiter M., Jaenicke V., Serre D., et al. (2001*), DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA*. Nucleic Acid Research. 29 (23): 4793-4799.

[15] Knapp M. & Hofreiter M. (2010), *Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives.* Genes 1, no. 2: 227-243.

[16] Pinhasi R, Fernandes D, Sirak K, et al. (2015), *Optimal Ancient DNA Yields from the Inner Ear Part of the Human Petrous Bone*. PLoS ONE 10(6): e0129102. https://doi.org/10.1371/journal.pone.0129102.

[17] Rohland, N. & Hofreiter, M. (2007), *Ancient DNA extraction from bones and teeth*. Nat Protoc 2, 1756–1762. https://doi.org/10.1038/nprot.2007.247.

[18] Cold Spring Harb Protoc; 2010; doi:10.1101/pdb.prot5448.

[19] Gansauge M., Meyer M. (2013), *Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA*. Nat Protoc 8, 737–748. https://doi.org/10.1038/nprot.2013.038.

[20] Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D. (2015), *Partial uracil-DNA-glycosylase treatment for screening of ancient DNA*. Philos Trans R Soc Lond B Biol Sci. 370(1660):20130624. doi:10.1098/rstb.2013.0624.

[21] Maricic T, Whitten M, Pääbo S (2010), *Multiplexed DNA Sequence Capture of Mitochondrial Genomes Using PCR Products.* PLoS ONE 5(11): e14004. https://doi.org/10.1371/journal.pone.0014004.

[22] Haak W., Lazaridis I., Patterson N., *et* al. (2015), *Massive migration from the steppe was a source for Indo-European languages in Europe.* Nature. 522: 207-211.

[23] Fu Q., Meyer M., Gao X., et al. (2013), *DNA analysis of an early modern human from Tianyuan Cave, China.* PNAS 110 (6) 2223-2227; https://doi.org/10.1073/pnas.1221359110.

[24] Dabney J, Meyer M, Pääbo S. (2013), *Ancient DNA damage*. Cold Spring Harb Perspect Biol. 5(7):a012567. doi: 10.1101/cshperspect.a012567. PMID: 23729639; PMCID: PMC3685887.

[25] Briggs AW, Stenzel U, Johnson PLF, et al. (2007), *Patterns of damage in genomic DNA sequences from a Neandertal*. Proc Natl Acad Sci 104: 14616–14621.

[26] Hofreiter, M., Serre, D., Poinar, H. et al. (2001), *Ancient DNA*. Nat Rev Genet 2, 353–359. https://doi.org/10.1038/35072071

[27] Poinar HN, Höss M, Bada JL, Pääbo S. (1996), *Amino acid racemization and the preservation of ancient DNA.* Science 272:864–6.

[28] Wilson I.G. (1997), *Inhibition and Facilitation of Nucleic Acid Amplification*. Applied and environmental microbiology, 63(10): 3741-3751.

[29] Willerslev E., Cooper A. (2005), *Ancient DNA*. Proc. R. Soc. B. 272: 3-16.

[30] Jackson C.R., Harper J.P., et al. (1997), *A simple, efficient method for the separation of humic substances and DNA from environmental samples*. Appl. Environ. Microbiol. 58: 2458-2462.

[31] Brown TA, Brown KA. (1992), *Ancient DNA and the archaeologist*. Antiquity 66:10–23.

[32] Yang DY, Watt K. (2005), *Contamination controls when preparing archaeological remains for ancient DNA analysis*. J. Archaeol. Sci. 32:331–336.

[33] Cooper A, Poinar HN. (2000), *Ancient DNA: do it right or not at all*. Science 289:1139.

[34] Pilli E, Modi A, Serpico C, et al. (2013), *Monitoring DNA Contamination in Handled vs. Directly Excavated Ancient Human Skeletal Remains*. PLoS One 8:e52524.

[35] Llamas B, Valverde G, Fehren-Schmitz L, et al. (2017), *From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era*. STAR Sci. Technol. Archaeol. Res. 3:1–14.

[36] Damgaard PB, Margaryan A, Schroeder H, et al. (2015), *Improving access to endogenous DNA in ancient bones and teeth*. Sci. Rep. 5:11184.

[37] Shapiro B, Hofreiter M (2012), *Ancient DNA: methods and protocols*. New York, NY: Humana Press.

[38] Ávila-Arcos M., Cappellini E., Romero-Navarro J. et al. (2011), *Application and comparison of large-scale solution-based DNA capture-enrichment methods on ancient DNA*. Sci Rep 1, 74. https://doi.org/10.1038/srep00074.

[39] Gnirke A., et al (2009), *Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing*. Nat Biotechnol 27(2):182–189.

[40] Meyer M, Kircher M. (2010), *Illumina sequencing library preparation for highly multiplexed target capture and sequencing*. Cold Spring Harb Protoc. (6):pdb.prot5448.

[41] Paijmans JLA., Barnett R., Gilbert MTP., et al. (2017), *Evolutionary History of Saber-Toothed Cats Based on Ancient Mitogenomics*. Curr Biol. 27(21):3330-3336.e5. doi:10.1016/j.cub.2017.09.033.

[42] Fortes GG., Grandal-d'Anglade A., Kolbe B., et al. (2016), *Ancient DNA reveals differences in behaviour and sociality between brown bears and extinct cave bears*. Mol Ecol.25(19):4907-4918. doi:10.1111/mec.13800.

[43] Posth C., Renaud G., Mittnik A., et al. (2016), *Pleistocene Mitochondrial Genomes Suggest a Single Major Dispersal of Non-Africans and a Late Glacial Population Turnover in Europe* [published correction appears in Curr Biol. Feb 22;26(4):557-61]. Curr Biol. 26(6):827-833. doi:10.1016/j.cub.2016.01.037.

[44] Brotherton P., Haak W., Templeton J., et al. (2013), *Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans.* Nat. Commun. 4:1764.

[45] Gonzalez-Fortes G., et al. (2019), *A western route of prehistoric human migration from Africa into the Iberian Peninsula*. Proc. R. Soc. B 286:20182288. http://dx.doi.org/10.1098/rspb.2018.2288.

[46] Carpenter ML., Buenrostro JD., Valdiosera C., et al. (2013), *Pulling out the 1%: whole-genome capture for the targeted enrichment of ancient DNA sequencing libraries*. Am J Hum Genet. 93(5):852-864. doi:10.1016/j.ajhg.2013.10.002.

[47] Burbano HA., Hodges E., Green RE., et al. (2010), *Targeted investigation of the Neandertal genome by array-based sequence capture*. Science. 328(5979):723-725. doi:10.1126/science.1188046.

[48] Castellano S., Parra G., Sánchez-Quinto FA., et al. (2014), *Patterns of coding variation in the complete exomes of three Neandertals*. Proc Natl Acad Sci U S A. 111(18):6666-6671. doi:10.1073/pnas.1405138111.

[49] Lazaridis I., Patterson N., Mittinik A., *et* al. (2014), *Ancient human genomes suggest three ancestral populations for present-day Europeans.* Nature. 513: 409-413.

[50] Mathieson, I., Lazaridis, I., Rohland, N., et al. (2015), *Genome-wide patterns of selection in 230 ancient Eurasians*. Nature, 528(7583), 499–503. https://doi.org/10.1038/nature16152.

[51] Cruz-Dávalos et al. (2018), *In-solution Y-chromosome capture enrichment on ancient DNA libraries.* BMC Genomics 19:608 https://doi.org/10.1186/s12864-018-4945-x.

[52] Marciniak S., Perry G. (2017), *Harnessing ancient genomes to study the history of human adaptation*. Nat Rev Genet 18, 659–674. https://doi.org/10.1038/nrg.2017.65.

[53] Reich D., Green R.E., Kircher M., *et* al. (2010), *Genetic history an archaic hominin group from Denisova Cave in Siberia. Nature*. 468:1053–1060.

[54] Fu Q., *et* al. (2016), *The genetic history of Ice Age Europe. Nature. 534: 200–205.*

[55] Lazaridis I., Nadel D., Rollefson G., *et* al. (2016), *The genetic structure of the world's first farmers.* Nature. 10.1038/nature19310.

[56] Green R.E. et al. (2006), *Analysis of one million base pairs of Neanderthal DNA*. Nature 444: 330–336.

[57] Dabney J., Knapp M., Glocke I., et al. (2013), *Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments.* Proceedings of the National Academy of Sciences, 110 (39) 15758-15763; DOI: 10.1073/pnas.1314445110.

[58] Cavalli-Sforza LL., Menozzi P., Piazza A. (1994), *The History and Geography of Human Genes*. Princeton, USA: Princeton University Press.

[59] Fu Q., Mittnick A., et al. (2013), *A revised timescale for human evolution based on ancient mitochondrial genomes.* Curr. Biol., 23 (7): 553–559, doi:10.1016/j.cub.2013.02.044.

[60] Ramakrishnan U. & Hadly E.A. (2009), *Using phylochronology to reveal cryptic population histories: review and synthesis of 29 ancient DNA studies.* Molecular Ecology (2009) 18, 1310–1330 doi: 10.1111/j.1365-294X.2009.04092.x.

[61] Horai S., Hayasaka K., Kondo R., et al. (1995), *Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs*. Proc Natl Acad Sci USA. 92(2):532–536.

[62] Pakendorf B. & Stoneking M. (2005), *Mitochondrial DNA and human evolution*. Annu Rev Genomics Hum Genet.6:165–183.

[63] Brandt G., Haak W., Adler C.J., et al. (2013), *Ancient DNA reveals key stages in the formation of central European mitochondrial genetic diversity*. Science, 342, 257–261.

[64] Kılınc G.M., Omrak A., Ozer F., et al. (2016), *The demographic development of the first farmers in Anatolia.* Curr. Biol., 26, 2659–2666.

[65] Lazaridis I., Nadel D., Rollefson G., et al. (2016), *Genomic insights into the origin of farming in the ancient Near East*. Nature, 536, 419–424.

[66] Mathieson I., Alpaslan-Roodenberg S., Posth C., et al. (2018), *The genomic history of southeastern Europe.* Nature, 555, 197–203.

[67] Kaestle F.A. & Horsburgh K.A. (2002), *Ancient DNA in Anthropology: Methods, Applications, and Ethics.* Yearbook of Physical Anthropology 45:92–130. DOI 10.1002/ajpa.10179.

[68] Lalueza-Fox C., et al. (2007), *A Melanocortin 1 Receptor Allele Suggests Varying Pigmentation Among Neanderthals.* Science 318, 1453. DOI: 10.1126/science.1147417.

[69] Krause J., et al. (2007), *The Derived FOXP2 Variant of Modern Humans Was Shared with Neandertals*. Current Biology, Volume 17, Issue 21, Pages 1908-1912, ISSN 0960-9822. https://doi.org/10.1016/j.cub.2007.10.008.

[70] Stone AC. & Stoneking M. (1999), *Analysis of ancient DNA from a prehistoric Amerindian cemetery*. Philos Trans R Soc Lond B Biol Sci. 354(1379):153-9. doi: 10.1098/rstb.1999.0368. PMID: 10091255; PMCID: PMC1692451.

[71] Kivisild T. (2017), *The study of human Y chromosome variation through ancient DNA*. Hum Genet 136:529–546 DOI 10.1007/s00439-017-1773-z.

[72] Katsura Y., Iwase M., and Satta Y. (2012), *Evolution of genomic structures on Mammalian sex chromosomes*. Curr. Genomics 13, 115–123. doi: 10.2174/138920212799860625.

[73] Trombetta B, D'Atanasio E and Cruciani F (2017), *Patterns of Inter-Chromosomal Gene Conversion on the Male-Specific Region of the Human Y Chromosome*. Front. Genet. 8:54. doi: 10.3389/fgene.2017.00054.

[74] Hughes J.F & Page D.C (2015), *The Biology and Evolution of Mammalian Y Chromosomes*. Annual Review of Genetics 49:507-527. https://doi.org/10.1146/annurev-genet-112414-055311.

[75] Graves J. A. (2004) *The degenerate Y chromosome--can conversion save it?* Reproduction, fertility, and development 16, 527-534, doi:10.10371/rd03096.

[76] Castillo ER., Marti DA., Bidau CJ. (2010), *Sex- and neo-sex chromosomes in Orthoptera: a review*. Journal of Orthoptera Research., 213–231.

[77] Francalacci P., Sanna D., Useli A. (2016), *Human Y Chromosome Mutation Rate: Problems and Perspectives*. Anthropology: Current and Future Developments, Vol. 2, 64-90.

[78] Skaletsky H., Kuroda-Kawaguchi T., Minx PJ., et al. (2003), *The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes*. Nature 423(6942): 825-37. [http://dx.doi.org/10.1038/nature01722] [PMID: 12815422].

[79] Francalacci P., Sanna D., Useli A., et al. (2015), *Detection of phylogenetically informative polymorphisms in the entire euchromatic portion of human Y chromosome from a Sardinian sample*. BMC Research Notes. 8:174.

[80] Jobling M., Tyler-Smith C. (2017), *Human Y-chromosome variation in the genome-sequencing era*. Nat Rev Genet 18:485–497. https://doi.org/10.1038/nrg.2017.36.

[81] Hammer M.F., Karafet T.M., Redd A.J., et al. (2001), *Hierarchical Patterns of Global Human Y-Chromosome Diversity, Molecular Biology and Evolution*. 18:7 (1189–1203). https://doi.org/10.1093/oxfordjournals.molbev.a003906.

[82] Kimura & Motoo (1983), *The neutral theory of molecular evolution*. Cambridge University Press. ISBN 978-0-521-31793-1.

[83] Helgason A., Einarsson A., Guðmundsdóttir V. et al. (2015), *The Y-chromosome point mutation rate in humans.* Nat Genet 47, 453–457. https://doi.org/10.1038/ng.3171.

[84] Poznik GD, Henn BM, Yee MC, et al. (2013), *Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females*. Science. 341(6145): 562-5. [http://dx.doi.org/10.1126/science.1237619] [PMID: 23908239].

[85] Karmin M, Saag L, Vicente M, et al. (2015), *A recent bottleneck of Y chromosome diversity coincides with a global change in culture*. Genome Res; 25(4): 459-66. [http://dx.doi.org/10.1101/gr.186684.114] [PMID: 25770088].

[86] Underhill PA., Shen P., Lin AA., et al. (2000), *Y chromosome sequence variation and the history of human populations*. Nat Genet. 26(3):358-61. doi: 10.1038/81685. PMID: 11062480.

[87] Y Chromosome Consortium. (2002), *A nomenclature system for the tree of human Y-chromosomal binary haplogroups*. Genome Res. 12: 339–348.

[88] Underhill P.A. & Kivisild T. (2007), *Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations.* Annu. Rev. Genet. 41:539-564.

[89] Karafet TM., Mendez FL., Meilerman MB., et al. (2008), *New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree*. Genome Research. 18 (5): 830–38. doi:10.1101/gr.7172008. PMC 2336805. PMID 18385274.

[90] Cruciani F., Trombetta B., Massaia A., et al. (2011), *A Revised Root for the Human Y Chromosomal Phylogenetic Tree: The Origin of Patrilineal Diversity in Africa.* Report. 88(6):814-818. DOI: https://doi.org/10.1016/j.ajhg.2011.05.002.

[91] Mendez FL., Krahn T., Schrack B., et al. (2013), *An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree*. Am J Hum Genet. 7;92(3):454-9. doi: 10.1016/j.ajhg.2013.02.002.

[92] Haber M., Jones AL., Connell B.A., et al. (2019), *A rare deep-rooting D0 African Y-chromosomal haplogroup and its implications for the expansion of modern humans out of Africa*. Genetics 212:1421–1428. https://doi.org/10.1534/genetics.119.302368.

[93] Hallast P., Agdzhoyan A., Balanovsky O. et al. (2020), *A Southeast Asian origin for present-day non-African human Y chromosomes*. Hum Genet. https://doi.org/10.1007/s00439-020-02204-9.

[94] Wei W., Ayub Q., Xue Y., Tyler-Smith C. (2013), *A comparison of Y-chromosomal lineage dating using either resequencing or Y-SNP plus Y-STR genotyping.* Forensic Sci Int Genet; 7(6): 568-72. [http://dx.doi.org/10.1016/j.fsigen.2013.03.014] [PMID: 23768990].

[95] Francalacci P., Morelli L., Angius A., et al. (2013), *Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny*. Science; 341(6145): 565-9. [http://dx.doi.org/10.1126/science.1237947] [PMID: 23908240].

[96] Zerjal T., Xue Y., Bertorelle G., et al. (2003), *The genetic legacy of the Mongols*. Am J Hum Genet; 72(3):717-21. [http://dx.doi.org/10.1086/367774] [PMID: 12592608].

[97] Drummond AJ., Rambaut A. (2007), *BEAST: Bayesian evolutionary analysis by sampling trees*. BMC Evol Biol; 7: 214. [http://dx.doi.org/10.1186/1471-2148-7-214] [PMID: 17996036].

[98] Maarten HD., Larmuseau & Claudio Ottoni (2018), *Mediterranean Ychromosome 2.0—why the Y in the Mediterranean is still relevant in the postgenomic era*, Annals of Human Biology, 45:1, 20-33, DOI: 10.1080/03014460.2017.1402956.

[99] Heyer E., Chaix R., Pavard S. & Austerlitz F. (2012). *Sex-specific demographic behaviours that shape human genomic variation*. Mol. Ecol. 21, 597–612.

[100] Jobling M.A., & Tyler-Smith C. (2003), *The human Y chromosome: an evolutionary marker comes of age*. Nature reviews. Genetics, 4(8), 598–612. https://doi.org/10.1038/nrg1124.

[101] Freeman L., Brimacombe CS., Elhaik E. (2020), *aYChr-DB: a database of ancient human Y haplogroups*. NAR Genomics and Bioinformatics. 2(4). lqaa081, https://doi.org/10.1093/nargab/lqaa081.

[102] Larmuseau MHD., Van Geystelen A., Kayser M., et al. (2015). *Towards a consensus Y-chromosomal phylogeny and Y-SNP set in forensics in the next-generation sequencing era.* Forensic Sci Int Genet. 15:39–42.

[103] Poznik GD., Xue Y., Mendez FL., et al. (2016), *Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences*. Nat Genet. 12(9):809.

[104] Haber M., Doumet-Serhal C., Scheib C.L, et al. (2017), *Continuity and admixture in the last five millennia of Levantine history from ancient Canaanite and present-day Lebanese genome sequences*. Am J Hum Genet 101:274–282.

[105] Gleize Y., Mendisco F., Pemonge M-H., et al. (2016), *Early Medieval Muslim Graves in France: First Archaeological, Anthropological and Palaeogenomic Evidence*. PLoS ONE 11(2): e0148583. https://doi.org/10.1371/journal.pone.0148583.

[106] Sankararaman, S., Mallick, S., Dannemann, M. et al. (2014), *The genomic landscape of Neanderthal ancestry in present-day humans.* Nature 507, 354–357. https://doi.org/10.1038/nature12961.

[107] Fernando L., Mendez G., Poznik D., et al. (2016), *The Divergence of Neandertal and Modern Human Y Chromosomes*. The American Journal of Human Genetics. 98(4):728-734.

[108] Petr M., Hajdinjak M, Fu Q., et al. (2020), *The evolutionary history of Neanderthal and Denisovan Y chromosomes.* Science. 1653-1656.

[109] Jobling MA., Hollox E., Hurles ME., et al. (2013), *Human Evolutionary Genetics*. London/New York: Garland Science Publishing. 650.

[110] Calderon R. (2000), *Population and peopling in the Mediterranean world*. Int J Anthropol 15:271–278. doi: 10.1007/BF02445138.

[111] Villaescusa P., Illescas MJ., Valverde L., et al. (2017), *Characterization of the Iberian Y chromosome haplogroup R-DF27 in Northern Spain*. Forensic Sci Int Genet 27:142–148.

[112] Rey-Gonzalez D., Gelabert-Besada M., Cruz R, Brisighelli F., et al. (2017), *Micro and macro geographical analysis of Y-chromosome lineages in South Iberia*. Forensic Sci Int Genet 29:e9–e15.

[113] Sarno S., Boattini A., Carta M., et al. (2014), *An ancient Mediterranean melting pot: investigating the uniparental genetic structure and population history of sicily and southern Italy*. PLoS One 9:e96074.

[114] Boattini A., Martinez-Cruz B., Sarno S., et al. (2013), *Uniparental markers in Italy reveal a sex-biased genetic structure and different historical strata*. PLoS One 8:e65441.

[115] Brisighelli F., Alvarez-Iglesias V., Fondevila M., et al. (2012), *Uniparental markers of contemporary Italian population reveals details on its pre-Roman heritage*. PLoS One 7:e50794.

[116] Capelli C., Brisighelli F., Scarnicci F., et al. (2007), *Y chromosome genetic variation in the Italian peninsula is clinal and supports an admixture model for the Mesolithic-Neolithic encounter*. Mol Phylogenet Evol 44:228–239.

[117] Novelletto A. (2007), *Y chromosome variation in Europe: Continental and local processes in the formation of the extant gene pool*. Ann Hum Biol 34:139–172.

[118] Sazzini M., Sarno S., Luiselli D. (2014), *The Mediterranean human population: an anthropological genetics perspective*. In: Goffredo S, Dubinsky Z, editors. The Mediterranean Sea - Its history and present challenges. Dordrecht: Springer.

[119] Batini C., Hallast P., Zadik D., et al. (2015), *Large-scale recent expansion of European patrilineages shown by population resequencing*. Nature Commun 6:7152.

[120] Batini C., Jobling MA. (2017), *Detecting past male-mediated expansions using the Y chromosome*. Hum Genet 136:547–557.

[121] Tofanelli S., Brisighelli F., Anagnostou P., et al. (2016), *The Greeks in the West: genetic signatures of the Hellenic colonisation in southern Italy and Sicily*. Eur J Hum Genet 24:429–436.

[122] Di Giacomo F., Luca F., Anagnou N., et al. (2003), *Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects*. Mol Phylogenet Evol.28(3):387-95. doi: 10.1016/s1055-7903(03)00016-2. PMID: 12927125.

[123] Capelli C., et al. (2006), *Population Structure in the Mediterranean Basin: A Y Chromosome Perspective*. Annals of Human Genetics. 70,207–225. https://doi.org/10.1111/j.1529-8817.2005.00224.x.

[124] Sazzini M., Gnecchi Ruscone G., Giuliani C. et al. (2016), *Complex interplay between neutral and adaptive evolution shaped differential genomic background and disease susceptibility along the Italian peninsula*. Sci Rep 6, 32513. https://doi.org/10.1038/srep32513.

[125] Khan K., Siddiqi M.H., Abbas M., et al. (2017), *Forensic applications of Y chromosomal properties*. Legal Medicine 26 (2017) 86–91.

[126] Roewer L., Kayser M., Dieltjes P., et al. (1996), *Analysis of molecular variance (AMOVA) of Y-chromosome-specific microsatellites in two closely related human populations*, Hum. Mol. Genet. 5 (7) 1029–1033.

[127] Butler J.M., Schoske R., Vallone P.M., et al. (2002), *A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers*, Forensic Sci. Int. 129 (1):10–24.

[128] Kayser M. (2017), Forensic use of Y-chromosome DNA: a general overview Hum. Genet., 136 (5): 621-635.

[129] Nielsen R., Akey J.M., Jakobsson M., et al. (2017), *Tracing the peopling of the world through genomics*. Nature 541:302–310.

[130] Günther T. & Jakobsson M. (2016), *Genes mirror migrations and cultures in prehistoric Europe — a population genomic perspective.* Curr. Opin. Genet. Dev. 41:115–123.

[131] Harris EE. (2017), *Demic and cultural diffusion in prehistoric Europe in the age of ancient genomes*. Evol. Anthropol. Issues, News, Rev. 26:228–241.

[132] Hofmanová Z., et al. (2016), *Aegean origin of European Neolithic farmers.* Proceedings of the National Academy of Sciences Jun 2016, 113 (25) 6886-6891; DOI: 10.1073/pnas.1523951113.

[133] Fernández-Domínguez E. & Reynolds L. (2017), *The Mesolithic-Neolithic Transition in Europe: A Perspective from Ancient Human DNA*. In: García-Puchol O., Salazar-García D. (eds) Times of Neolithic Transition along the Western Mediterranean. Fundamental Issues in Archaeology. Springer, Cham. https://doi.org/10.1007/978-3-319-52939-4_12.

[134] Bramanti B., et al. (2009), *Genetic discontinuity between local hunter-gatherers and central Europe's first farmers*. Science 326(5949):137–140.

[135] Weninger B., et al. (2014), *Neolithisation of the Aegean and Southeast Europe during the 6600–6000 calBC period of Rapid Climate Change*. Documenta Praehistorica 41:1–31.

[136] Günther T., Valdiosera C., Malmström H., et al. (2015), *Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques.* Proc Natl Acad Sci U S A. 112(38):11917-11922. doi:10.1073/pnas.1509851112.

[137] Furholt M. (2018), *Massive Migrations? The Impact of Recent aDNA Studies on our View of Third Millennium Europe*. European Journal of Archaeology. 21(2), 159-191. doi:10.1017/eaa.2017.43.

[138] Allentoft M., Sikora M., Sjögren KG. et al. (2015), *Population genomics of Bronze Age Eurasia*. Nature 522, 167–172. https://doi.org/10.1038/nature14507.

[139] Shishlina N. (2008), *Reconstruction of the Bronze Age of the Caspian Steppes. Life Styles and Life Ways of Pastoral Nomads*. Archaeopress .Vol. 1876.

[140] Anthony D. (2007), *The Horse, The Wheel and Language. How Bronze-Age Riders from the Eurasian Steppes Shaped the Modern World.* Princeton Univ. Press.

[141] Vandkilde H. (2007), *Culture and Change in the Central European Prehistory, 6th to 1st millennium BC.* Aarhus Univ. Press.

[142] Olalde I & Posth C. (2020), *Latest trends in archaeogenetic research of west Eurasians*. Curr Opin Genet Dev.62:36-43. doi:10.1016/j.gde.2020.05.021.

[143] Busby GBJ., Hellenthal G., Montinaro F., et al. (2015), *The Role of Recent Admixture in Forming the Contemporary West Eurasian Genomic Landscape*. Curr Biol. 25(21):2878. doi:10.1016/j.cub.2015.10.037.

[144] Krzewinska M., Kilinc GM., Juras A., et al. (2018), *Ancient genomes suggest the eastern Pontic-Caspian steppe as the source of western Iron Age nomads*. Sci Adv. 4eaat4457.

[145] Martiniano R., Caffell A., Holst M., et al. (2016), *Genomic signals of migration and continuity in Britain before the Anglo-Saxons*. Nat. Commun. 7:10326.

[146] Leslie S., Winney B., Hellenthal G., et al. (2015), *The fine-scale genetic structure of the British population.* Nature 519:309–314.

[147] Schiffels S., Haak W., Paajanen P., et al. (2016), Iron Age and Anglo-Saxon genomes from East England reveal British migration history. Nat. Commun. 7:10408.

[148] Hellenthal G., Busby GBJ., Band G., et al. (2014), *A genetic atlas of human admixture history*. Science 343:747–751.

[149] Novembre J., Johnson T., Bryc K., et al. (2008), *Genes mirror geography within Europe*. Nature 456:98–101.

[150] Antonio ML., Gao Z., Moots HM., et al. (2019), *Ancient Rome: A genetic crossroads of Europe and the Mediterranean*. Science.366(6466):708-714. doi:10.1126/science.aay6826.

[151] Broodbank C. (2013), The Making of the Middle Sea. *A History of the Mediterranean from the Beginning to the Emergence of the Classical World.* Oxford Univ. Press.

[152] Ho SYW., Gilbert MTP (2010), *Ancient mitogenomics*. Mitochondrion. 10(1):1–11.

[153] Skoglund P., Malmstrom H., Raghavan M., et al. (2012), *Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe.* Science 336, 466–469.

[154] Ginolhac, A., Vilstrup, J., Stenderup, J. et al. (2012), *Improving the performance of true single molecule sequencing for ancient DNA*. BMC Genomics 13, 177. https://doi.org/10.1186/1471-2164-13-177.

[155] Enk JM., Devault AM., Kuch M, et al. (2014), *Ancient Whole Genome Enrichment Using Baits Built from Modern DNA.* Molecular Biology and Evolution. 31:5, 1292–1294, https://doi.org/10.1093/molbev/msu074.

[156] Fortes GG. & Paijmans JL (2015), *Analysis of Whole Mitogenomes from Ancient Samples*. Methods Mol Biol.1347:179-195. doi:10.1007/978-1-4939-2990-0_13.

[157] Furtwängler A, Neukamm J, Böhme L, et al. (2020), *Comparison of target enrichment strategies for ancient pathogen DNA*. Biotechniques.69(6):455-459. doi:10.2144/btn-2020-0100.

[158] Paijmans JL., Fickel J., Courtiol A., et al. (2016), *Impact of enrichment conditions on cross-species capture of fresh and degraded DNA*. Mol Ecol Resour. 16(1):42-55. doi:10.1111/1755-0998.12420.

[159] King T., Fortes G., Balaresque P. et al. (2014), *Identification of the remains of King Richard III*. Nat Commun 5, 5631. https://doi.org/10.1038/ncomms6631.

[160] Springer MS., Signore AV., Paijmans JL., et al. (2015), *Interordinal gene capture, the phylogenetic position of Steller's sea cow based on molecular and morphological data, and the macroevolutionary history of Sirenia*. Mol Phylogenet Evol. 91:178-193. doi:10.1016/j.ympev.2015.05.022.

[161] Bos KI., Jäger G., Schuenemann VJ., et al. (2015), *Parallel detection of ancient pathogens via array-based DNA capture*. Philos Trans R Soc Lond B Biol Sci. 370(1660):20130375. doi:10.1098/rstb.2013.0375.

[162] Paijmans JLA., González Fortes G., Förster DW. (2019), *Application of Solid-State Capture for the Retrieval of Small-to-Medium Sized Target Loci from Ancient DNA*. Methods Mol Biol. 1963:129-139. doi:10.1007/978-1-4939-9176-1_14.

[163] Mamanova L., Coffey AJ., Scott CE., et al. (2010), *Target-enrichment strategies for nextgeneration sequencing*. Nature methods. 7(2). DOI:10.1038/NMETH.1419.

[164] Templeton J. E., Brotherton P. M., Llamas B., et al. (2013), *DNA capture and next-generation sequencing can recover whole mitochondrial genomes from highly degraded samples for human identification.* Investig. Genet. 4:26. doi: 10.1186/2041-2223-4-26.

[165] Cruz-Dávalos DI., Llamas B., Gaunitz C., et al. (2017), *Experimental conditions improving in-solution target enrichment for ancient DNA*. Mol Ecol Resour. 17(3):508-522. doi:10.1111/1755-0998.12595.

[166] Bodi K., Perera AG., Adams PS., et al. (2013), *Comparison of commercially available target enrichment methods for next-generation sequencing*. J Biomol Tech. 24(2):73-86. doi:10.7171/jbt.13-2402-002.

[167] Green R.E., Krause J., Briggs A.W., et al. (2010), *A draft sequence of the Neandertal genome*. Science. 328:710–722.

[168] Meyer M., Kircher M., Gansauge M.T., et al. (2012), *A high-coverage genome sequence from an archaic Denisovan individual*. Science. 338:222–226.

[169] Rasmussen M., Li Y., Lindgreen S., et al. (2010), *Ancient human genome sequence of an extinct Palaeo-Eskimo*. Nature. 463:757–762.

[170] Rasmussen M., Guo X., Wang Y., et al. (2011), *An Aboriginal Australian genome reveals separate human dispersals into Asia.* Science. 334:94–98.

[171] Wei W., Ayub Q., Chen Y., et al. (2013), *A calibrated human Y-chromosomal phylogeny based on resequencing*. Genome Res. 23(2):388–395.

[172] Hallast P., Batini C., Zadik D., 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.

[173] Printzlau F., Wolstencroft J. & Skuse D. H. (2017), *Cognitive, behavioral, and neural consequences of sex chromosome aneuploidy*. J. Neurosci. Res. 95,311–319.

[174] Forsberg L. A. et al. (2014), *Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer*. Nat. Genet. 46, 624–628.

[175] WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects (2013) http://www.wma.net/en/30publications/10policies/b3/

[176] Amorim C.E.G et al. (2018), *Understanding 6th-century barbarian social organization and migration through paleogenomics*. Nature Communications. 9:3547 | DOI: 10.1038/s41467-018-06024-4.

[177] Vai S, Brunelli A, Modi A, et al. (2019), *A genetic perspective on Longobard-Era migrations*. Eur J Hum Genet. 27(4):647-656. doi:10.1038/s41431-018-0319-8.

[178] Agilent Technologies (2016), *SureSelectQXT Target Enrichment for Illumina Multiplexed Sequencing. Featuring Transposase-Based Library Prep Technology*. Protocol, G9681-90000.

[179] Peltzer A., Jäger G., Herbig A. et al. (2016), *EAGER: efficient ancient genome reconstruction*. Genome Biol 17, 60. https://doi.org/10.1186/s13059-016-0918-z.

[180] Van der Auwera G.A., et al. (2013), From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Current Protocols In Bioinformatics 43:11.10.1-11.10.33.

[181] Danecek P, Auton A., Abecasis G, et al. (2011), *The Variant Call Format and VCFtools*. Bioinformatics.

[182] Cock PJ., Fields CJ., Goto N., Heuer ML., Rice PM. (2010), *The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants*. Nucleic Acids Res. 38(6):1767-1771. doi:10.1093/nar/gkp1137.

[183] Andrews S. (2010), *FastQC: A quality control tool for high throughput sequence data*. Reference Source.

[184] Li H., & Durbin R. (2009), *Fast and accurate short read alignment with Burrows-Wheeler transform*. Bioinformatics (Oxford, England), 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

[185] Jónsson H., Ginolhac A., Schubert M., Johnson P. L., & Orlando L. (2013), *mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters*. Bioinformatics (Oxford, England), 29(13), 1682–1684.

[186] Green R.E., Malaspinas A.S., Krause J., et al. (2008), *A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing*. Cell, 134: 416-426.

[187] García-Garcerà, M., Gigli, E., Sanchez-Quinto, et al. (2011), *Fragmentation of contaminant and endogenous DNA in ancient samples determined by shotgun sequencing; prospects for human palaeogenomics*. PloS one, 6(8), e24161. https://doi.org/10.1371/journal.pone.0024161.

[188] Sawyer S., Krause J., Guschanski K., Savolainen V., Pääbo S. (2012*), Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA*. PLoS ONE 7(3): e34131. https://doi.org/10.1371/journal.pone.0034131.

[189] Ginolhac A., Rasmussen M., Gilbert MTP., et al. (2011), *mapDamage: testing for damage patterns in ancient DNA sequences.* Bioinformatics, 27(15): 2153–2155, https://doi.org/10.1093/bioinformatics/btr347.

[190] Danecek P., McCarthy SA. (2017), *BCFtools/csq: haplotype-aware variant consequences.* Bioinformatics. 33(13): 2037–2039. https://doi.org/10.1093/bioinformatics/btx100.

[191] Ralf A., González DM., Zhong K. & Kayser M., (2018), *Yleaf: Software for Human Y-Chromosomal Haplogroup Inference from Next-Generation Sequencing Data.* Molecular Biology and Evolution. 35(5): 1291–1294. https://doi.org/10.1093/molbev/msy032.

[192] Soares P., Achilli A., Semino O., *et* al. (2010), *The archaeogenetics of Europe.* Current Biology. 20: 174-183.

[193] Groucutt H.S., Petraglia M.D., Bailey G., et al. (2015), *Rethinking the dispersal of Homo sapiens out of Africa.* Evol. Anthropol. 24: 149-164.

[194] Macaulay V., Hill C., Achilli A., et al. (2005), *Single, Rapid Coastal Settlement of Asia Revealed by Analysis of Complete Mitochondrial Genomes.* Science. 308: 1034-1036.

[195] Oppenheimer S. (2012). *Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map*. Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 367(1590), 770–784. https://doi.org/10.1098/rstb.2011.0306.

[196] Mellars P., Gori K.C., Carr M., Soares P.A. et Richards M.B. (2013), *Genetic and archaeological perspectives on the initial modern human colonization of southern Asia*. Proc. Natl. Acad. Sci. 110: 10699-10704.

[197] Lahr M.M & Foley R.A, 1998, *Towards a Theory of Modern Human Origins: Geography, Demography, and Diversity in Recent Human Evolution*. Am. J. Phys. Anthropol. 27:137-176.

[198] Maca-Meyer N., González A.M., Larruga J.M., Flores C. et Cabrera V.M. (2001), *Genomic mitochondrial lineages delineate early human expansions*. BMC Genet. 2:13.

[199] Armitage S.J., Jasim S.A., Marks A.E., et al. (2011), *The Southern Route "Out of Africa": Evidence for an Early Expansion of Modern Humans into Arabia*. Science.311: 453-456.

[200] Reyes-Centeno H., Ghirotto S., Detroit F., et al. (2014), *Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia*. Proc. Natl. Acad. Sci. 111: 7248-7253.

[201] Higham T. (2011*), European Middle and Upper Palaeolithic radiocarbon dates are often older than they look: problems with previous dates and some remedies*. Antiquity. 85: 235-249.

[202] Barbujani G. & Bertorelle G. (2001), *Genetics and the population history of Europe*. Proceedings of the National Academy of Sciences. 98 (1) 22-25; DOI: 10.1073/pnas.98.1.22.

[203] Pinhasi R. *et* al. (2012), *The genetic history of Europeans.* Trends in Genetics. 28 (10): 496 – 505.

[204] Skoglund P., Malmström H., Omrak A., et al. (2014), *Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers.* Science. 344(6185): 747–750.

[205] Hooker, J.T. (1976). *Mycenaean Greece*. London: Routledge & Kegan Paul. ISBN 9780710083791.

[206] Dzino D. (2014). *'Illyrians' in ancient ethnographic discourse*. Dialogues d'histoire ancienne. 40 (2): 45–65. doi:10.3917/dha.402.0045.

[207] Grugni V., Raveane A., Mattioli F., et al. (2018), Reconstructing the genetic history of Italians: new insights from a male (Y-chromosome) perspective. Annals of Human Biology, 45:1, 44-56, DOI: 10.1080/03014460.2017.1409801

[208] Dyson S. & Rowland Jr. R. (2007), *Archaeology and History in Sardinia from the Stone Age to the Middle Ages: Shepherds, Sailors, and Conquerors*. Philadelphia, PA: University of Pennsylvania Museum of Archaeology and Anthropology. Pp. viii + 240, illus. ISBN 978-1-93453-602-5.

[209] Maggi R. & Pearce M. (2005), *Mid fourth-millennium copper mining in Liguria, north-west Italy: The earliest known copper mines in Western Europe.* Antiquity 79(303) DOI: 10.1017/S0003598X00113705.

[210] Guimaraes S., Ghirotto S., Benazzo A., et al. (2009), *Genealogical Discontinuities among Etruscan, Medieval, and Contemporary Tuscans*. Molecular Biology and Evolution. 26:9 (2157–2166). https://doi.org/10.1093/molbev/msp126.

[211] Giannattasio B.M., (2007). *I Liguri e la Liguria, Storia e archeologia di un territorio prima della conquista romana*. Longanesi, Milano.

[212] Aspes A. (1984), *Il Veneto nell'antichità: preistoria e protostoria.* Verona, Banca Popolare di Verona.

[213] Di Martino U. (1984), *Le Civiltà dell'Italia antica*. Milano.

[214] AA.VV., *I Greci in Italia. Arte e civiltà della Magna Grecia.* A cura di Fabio Bourbon e Furio Durando. Foto di Livio Bourbon). Magnus, 2004, p. 320.

[215] Fiorito G., Di Gaetano G., et al. (2015), *The Italian genome reflects the history of Europe and the Mediterranean basin.* European Journal of Human Genetics. 1–7.

[216] Ghirotto S., Tassi F., Fumagalli E., et al., (2013). *Origins and Evolution of the Etruscans' mtDNA*. PLOS ONE 8(2): e55519. doi: 10.1371/journal.pone.0055519.

[217] Tassi F., Ghirotto S., Caramelli D., and Barbujani G. (2013), *Genetic evidence does not support an Etruscan origin in Anatolia*. Am.J.Phys.Ant. 152(1): 11-18.

[218] Francalacci P., Morelli L., Underhill P.A. et al. (2003), *Peopling of three Mediterranean islands (Corsica, Sardinia, and Sicily) inferred by Y-chromosome biallelic variability*. Am J Phys Anthropol. 121: 270– 279.

[219] Di Gaetano C., et al. (2009), *Differential Greek and northern African migrations to Sicily are supported by genetic evidence from the Y chromosome*. European Journal of Human Genetics 17: 91–99.

[220] Rosser Z.H., Zerjal T., Hurles M.E., et al. (2000), Y-*chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language*. Am J Hum Genet 67:1526–1543.

[221] Semino O., Passarino G., Brega A., et al. (1996), *A view of the Neolithic demic diffusion in Europe through two Y chromosome-specific markers*. Am J Hum Genet 59:964–968.

[222] Semino O., Passarino G., Oefner P.J., et al. (2000), *The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y-chromosome perspective*. Science 290:1155–1159.

[223] Bergonzi G & Eles Masi P. (1988), *Archaeological and anthropological evidence from the Iron Age necropolis at Montericco, Imola (Emilia-Romagna, Italy): A comparison.* In Atti Simposio Internazionale. Rivista di Antropologia. Roma. LXVI: 337-348

[224] Grassi MT. (1991), *I Celti in Italia*. Longanesi & C.:Milan.

[225] Kruta V. & Manfredi VM. (1999), *I Celti in Italia*. Mondatori: Milan.

[226] Mariotti V., Dutour O., Belcastro M.G., et al. (2005), *Probable Early Presence of Leprosy in Europe in a Celtic Skeleton of the 4th-3rd Century BC (Casalecchio di Reno, Bologna, Italy).* Int. J. Osteoarchaeol. 15: 311–325.

[227] Ortalli J. (1995), *La necropoli celtica della zona ''A'' di Casalecchio di Reno (Bologna). Note preliminari sullo scavo del complesso sepolcrale e dell'area di culto*. In L'Europe celtique du Ve au IIIe sie`cle avant J.-C. Actes du deuxie`me symposium international d'Hautvillers, 8–10 October 1992. 189–283.

[228] Mariotti V. (2001), *Skeletal markers of activity in the warriors from the Celtic necropolis of Casalecchio di Reno (Bologna, Italy) (IV–III c. BC).* In Attualita` dell'Antropologia. Ricerca e insegnamento nel XXI secolo. Atti del XIII Congresso degli Antropologi Italiani, Rome, 4–8 October 1999; 103–108.

[229] Costamagna L., Turchetti M.A., Chilleri F., et al. (2011), *Le necropoli di Norcia: il caso della TOMBA N.32 della necropoli di colle dell'Annunziata a Norcia (PG)*. Bollettino di archeologia on line. II, 2-3.

[230] Cordella R. & Criniti N. (2007), *La Sabina settentrionale: Norcia, Cascia e Valnerina romane*. Ager Veleias.

[231] Cenciaioli L., Capannelli S. & Cipiciani M.L. (2011), *Gubbio, la domus di Scilla e il parco urbano: uno studio di valorizzazione.* Bollettino di archeologia on line. II, 2-3.

[232] Alessandri L. (2009), *Il Lazio centromeridionale nelle età del Bronzo e del Ferro.* Groningen: s.n.. 620.

[233] Messina A., Sineo L., Schimmenti V., Di Salvo R. (2008), *Criba orbitalia and enamel hypoplasia of the Iron-Age (IX-VII centuries b.C.) human group of Polizzello (Sicily).* Journal of Paleopathology 20 (1-3).

[234] Knapp M., Lalueza-Fox C., Hofreiter M. (2015), *Re-inventing ancient human DNA*. Inves-tig. Genet. 6:4.

[235] Skoglund P., Storå J., Götherström A. & Jakobsson M. (2013), *Accurate sex identification of ancient human remains using DNA shotgun sequencing*. J Archaeol Sci. 2013;40(12):4477–82.

[236] Alexander D. H., Novembre J. & Lange K. (2009), *Fast model-based estimation of ancestry in unrelated individuals.* Genome Res. 19, 1655–1664.

[237] International Society of Genetic Genealogy (ISOGG). Available at: https://isogg.org/tree/index.html.

[238] Leigh J.W., Bryant D. (2015), *PopART: Full-feature software for haplotype network construction*. Methods Ecol Evol 6(9):1110–1116.

[239] Excoffier L. and Lischer H.E.L. (2010), *Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows*. Molecular Ecology Resources. 10: 564-567.

[240] Nei M., (1987), *Molecular Evolutionary Genetics*. Columbia University Press, New York, NY, USA.

[241] Tajima F. (1983), *Evolutionary relationship of DNA sequences in finite populations*. Genetics 105: 437-460.

[242] Tajima F. (1993), *Measurement of DNA polymorphism*. In: Mechanisms of Molecular Evolution. Introduction to Molecular Paleopopulation Biology, edited by Takahata, N. and Clark, A.G., Tokyo, Sunderland, MA:Japan Scientific Societies Press, Sinauer Associates, Inc., p. 37-59.

[243] Reynolds J., Weir B.S., & Cockerham C.C. (1983), *Estimation for the coancestry coefficient: basis for a short-term genetic distance*. Genetics. 105:767-779.

[244] Slatkin M. (1995), *A measure of population subdivision based on microsatellite allele frequencies*. Genetics 139: 457-462.

[245] Jolliffe IT. 2002. *Principal component analysis*. Springer.

[246] Lê S., Josse J. & Husson F. (2008), *FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software*. 25(1). pp. 1-18.

[247] Wickham H. (2016), *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

[248] Furtwängler A., Reiter E., Neumann G.U., et al. (2018), *Ratio of mitochondrial to nuclear DNA affects contamination estimates in ancient DNA analysis.* Sci Rep 8, 14075. https://doi.org/10.1038/s41598-018-32083-0.

[249] Myres, N., Rootsi, S., Lin, A. et al. (2011*), A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe*. Eur J Hum Genet 19, 95–101. https://doi.org/10.1038/ejhg.2010.146.

[250] Jeong C, Balanovsky O, Lukianova E, et al. (2019), *The genetic history of admixture across inner Eurasia*. Nat Ecol Evol.3(6):966-976. doi:10.1038/s41559-019-0878-2.

[251] Heraclides A., Bashiardes E., Fernandez-Domınguez E., et al. (2017), *Y-chromosomal analysis of Greek Cypriots reveals a primarily common pre-Ottoman paternal ancestry with Turkish Cypriots*. PLoS One 12:e0179474.

[252] Semino O., Magri C., Benuzzi G., et al. (2004), *Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area.* Am J Hum Genet.74(5):1023-1034. doi:10.1086/386295.

[253] Di Giacomo F., Luca F., Popa L.O., et al. (2004), *Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe*. Hum Genet 115, 357–371. https://doi.org/10.1007/s00439-004-1168-9.

[254] Finocchio A., Trombetta B., Messina F. et al. (2018), *A finely resolved phylogeny of Y chromosome Hg J illuminates the processes of Phoenician and Greek colonizations in the Mediterranean*. Sci Rep 8, 7465. https://doi.org/10.1038/s41598-018-25912-9.

[255] Rootsi S., Myres N.M., Lin A.A., et al. (2012), *Distinguishing the co-ancestries of haplogroup G Y-chromosomes in the populations of Europe and the Caucasus*. Eur J Hum Genet. 20(12):1275-1282. doi:10.1038/ejhg.2012.86.

[256] Rootsi S., et al. (2004), *Phylogeography of Y-Chromosome Haplogroup I Reveals Distinct Domains of Prehistoric Gene Flow in Europe*. American Journal of Human Genetics. 75 (1): 128–137. doi:10.1086/422196. PMC 1181996. PMID 15162323.

[257] Catalano P., Cavazzuti C., Celant A., et al. (2015), *Analisi contestuale di alimentazione e salute nel Lazio nella I età del Ferro (II periodo laziale ca. X-IX sec. a.C.).* 50ma Riunione Scientifica dell'Istituto Italiano di Preistoria e Protostoria.

[258] Modi A., Lancioni H., Cardinali I., et al. (2020), *The mitogenome portrait of Umbria in Central Italy as depicted by contemporary inhabitants and pre-Roman remains*. Sci Rep. 10(1):10700. doi: 10.1038/s41598-020-67445-0. PMID: 32612271; PMCID: PMC7329865.

[259] Serventi P., Panicucci C., Bodega R., et al. (2018), *Iron Age Italic population genetics: the Piceni from Novilara (8th-7th century BC)*. Ann Hum Biol. 2018 Feb;45(1):34-43. doi: 10.1080/03014460.2017.1414876. PMID: 29216758.

[260] Olivieri A., Sidore C., Achilli A., et al. (2017), *Mitogenome Diversity in Sardinians: A Genetic Window onto an Island's Past*. Mol Biol Evol. 34(5):1230-1239. doi: 10.1093/molbev/msx082. PMID: 28177087; PMCID: PMC5400395.

[261] Sazzini M., Abondio P., Sarno S., et al. (2020), *Genomic history of the Italian population recapitulates key evolutionary dynamics of both Continental and Southern Europeans*. BMC Biol. 18(1):51. doi: 10.1186/s12915-020-00778-4. PMID: 32438927; PMCID: PMC7243322.

[262] Brunel S., Bennett E., Cardin L., et al. (2020), *Ancient genomes from present-day France unveil 7,000 years of its demographic history*. Proceedings of the National Academy of Sciences, 117(23), pp.12791-12798.

[263] Järve M., Saag L., Scheib C., et al. (2019), *Shifts in the Genetic Landscape of the Western Eurasian Steppe Associated with the Beginning and End of the Scythian Dominance*. Current Biology, 29(14), pp.2430-2441.e10.

[264] Fernandes D., Mittnik A., Olalde I., et al. (2020), *The spread of steppe and Iranian-related ancestry in the islands of the western Mediterranean*. Nature Ecology & Evolution, 4, pp.334-345.

[265] Marcus et al. (2019), *Population history from the Neolithic to present on the Mediterranean island of Sardinia: An ancient DNA perspective*. bioRxiv, http://dx.doi.org/10.1101/583111.

[266] Marcus J., Posth C., Ringbauer H., et al. (2020), *Genetic history from the Middle Neolithic to present on the Mediterranean island of Sardinia*. Nature Communications, 11(939).

[267] Fernandes D., Strapagiel D., Borówka P., et al. (2018), *A genomic Neolithic time transect of hunter-farmer admixture in central Poland*. Scientific Reports, 8(1).

[268] Gamba C., Jones E., Teasdale M., et al. (2014), *Genome flux and stasis in a five millennium transect of European prehistory*. Nature Communications, 5(1).

[269] Olalde I., Brace S., Allentoft M. et al. (2018), *The Beaker phenomenon and the genomic transformation of northwest Europe*. Nature. 555, pp. 190-196.

[270] Cassidy L., Martiniano R., Murphy E., et al. (2016), *Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome*. Proceedings of the National Academy of Sciences, 113(2), pp.368-373.

[271] Olalde et al. (2019), *The genomic history of the Iberian Peninsula over the past 8000 years*. Science, 363(6432), pp. 1230-1234.

[272] Valdiosera C., Günther T., Vera-Rodríguez J., et al. (2018), *Four millennia of Iberian biomolecular prehistory illustrate the impact of prehistoric migrations at the far end of Eurasia.* Proceedings of the National Academy of Sciences, 115(13), pp.3428-3433.

[273] Wang C., Reinhold S., Kalmykov A., et al. (2019), *Ancient human genome-wide data from a 3000-year interval in the Caucasus corresponds with eco-geographic regions*. Nature Communications, 10(1).

[274] Boulygina E., Tsygankova S., Sharko F., et al. (2020), *Mitochondrial and Y-chromosome diversity of the prehistoric Koban culture of the North Caucasus*. Journal of Archaeological Science: Reports, 31, p.102357.

[275] Damgaard, Peter de Barros et al. (2018), *137 ancient human genomes from across the Eurasian steppes*. Nature, 557(7705), pp.369–374.

[276] Lipson M., Szécsényi-Nagy A., Mallick S., et al. (2017*), Parallel palaeogenomic transects reveal complex genetic history of early European farmers*. Nature, 551(7680), pp.368-372.

[277] Martiniano R., Cassidy L., Ó'Maoldúin R., et al. (2017), *The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods*. PLOS Genetics, 13(7), p.e1006852.

[278] Saag et al. (2020*), Genetic ancestry changes in Stone to Bronze Age transition in the East European plain.* BioRxiv, DOI: https://doi.org/10.1101/2020.07.02.184508

[279] Lacan M., Keyser C., Ricaut F., et al. (2011). *Ancient DNA suggests the leading role played by men in the Neolithic dissemination*. Proceedings of the National Academy of Sciences, 108(45), pp.18255-18259.

[280] Rivollat M., Jeong C., Schiffels S., et al. (2020), A*ncient genome-wide DNA from France highlights the complexity of interactions between Mesolithic hunter-gatherers and Neolithic farmers.* Science Advances, 6(22), p.eaaz5344.

[281] Schroeder H., Margaryan A., Szmyt M., et al. (2019), *Unraveling ancestry, kinship, and violence in a Late Neolithic mass grave*. Proceedings of the National Academy of Sciences, 116(22), pp.10705-10710.

[282] Brace S., Diekmann Y., Booth T., et al. (2019), *Ancient genomes indicate population replacement in Early Neolithic Britain*. Nature Ecology & Evolution, 3(5), pp.765-771.

[283] Sánchez-Quinto F., Malmström H., Fraser M., et al. (2019), *Megalithic tombs in western and northern Neolithic Europe were linked to a kindred society*. Proceedings of the National Academy of Sciences, 116(19), pp.9469-9474.

[284] Villalba-Mouco V. et al. (2019), *Survival of Late Pleistocene Hunter-Gatherer Ancestry in the Iberian Peninsula.* Current Biology, 29(7), pp. 1169-1177.e12.

[285] Mittnik A. et al. (2018), *The genetic prehistory of the Baltic Sea region*. Nature Communications, 9(1), p.442.

[286] Keller A., Graefen A., Ball M., et al. (2012). *New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing*. Nature Communications, 3(1).

[287] Mary L., Zvénigorosky V., Kovalev A., et al. (2019), *Genetic kinship and admixture in Iron Age Scytho-Siberians.* Human Genetics, 138(4), pp.411-423.

[288] Malmström H., Günther T., Svensson E., et al. (2019), *The genomic ancestry of the Scandinavian Battle Axe Culture people and their relation to the broader Corded Ware horizon*. Proceedings of the Royal Society B: Biological Sciences, 286(1912).

[289] Saag L., Varul L., Scheib C., et al. (2017), *Extensive Farming in Estonia Started through a Sex-Biased Migration from the Steppe*. Current Biology, 27(14), pp.2185-2193.e6.