



Examiners' Joint Report for a PhD Candidate			
Student's Full Name:	Sam Morris		
Student Number:	17114784	Examination for:	PhD
Thesis Title: <i>(Please enter complete thesis title)</i>	Harnessing haplotype sharing information from low coverage sequencing and sparsely genotyped data		
Date of Viva Examination:		February 25 2022	
Supervisor present? (If yes, please enter the supervisor's name in box below)			
Yes <input type="checkbox"/>		No <input checked="" type="checkbox"/>	

Please complete all sections of this report.

If you have not already submitted your preliminary report(s), please submit them with this signed joint report to [Research Degrees](#).

These reports will be forwarded by Research Degrees to the candidate, their supervisor and the Faculty Graduate Tutor.

SECTION A		
The examiners confirm that they have examined the thesis submitted by the candidate and have also examined the candidate orally on the subject of the thesis and on subjects relevant to the thesis		
<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No	
The examiners confirm that they have satisfied themselves that the candidate, as evidenced by the thesis and the viva, can communicate with the scholarly community about their areas of expertise		
<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No	
The examiners confirm that the thesis:		
is genuinely the work of the candidate:	<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
forms a distinct and significant contribution to knowledge of the subject:	<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
affords evidence of originality: a) by the discovery of new facts and/or b) by the exercise of independent critical power	<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
is an integrated whole and presents a coherent argument:	<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
gives a critical assessment of the relevant literature:	<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
gives the method of research and its findings:	<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
gives discussion of those findings and how they advance the study of the subject:	<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
demonstrates deep and synoptic understanding of the field of study, including objectivity, autonomy and the capacity for judgement in a complex situation:	<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
is satisfactory as regards literary presentation:	<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
includes a satisfactory bibliography and references:	<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
demonstrates research skills relevant to the thesis:	<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
is of a standard to merit publication in whole, in part or in revised form:	<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No

SECTION B	
The examiners confirm one of the following outcomes (please tick)	
Outcomes if the candidate will meet the required standard for a PhD award:	
1.)	The candidate has met the criteria for a PhD without the need for corrections and can be awarded the PhD <input type="checkbox"/>
2.)	The candidate is required to make specified amendments to the examiners' satisfaction within three months <input checked="" type="checkbox"/>
i)	The examiners confirm that the candidate has been provided with a list of minor amendments or an annotated thesis either immediately after the oral or within two weeks of the oral examination and has been asked to send the amended thesis for confirmation to the person nominated to check the corrections: <input checked="" type="checkbox"/>
OR	
ii)	The candidate has already made the minor amendments required to the satisfaction of designated checker: <input type="checkbox"/>
If outcome (i) has been selected, please nominate an individual to check the amendments. The person nominated should confirm these have been made satisfactorily by email to Research Degrees	
Name of corrections checker: Dr Hernán A. Burbano	
Email address: h.burbano@ucl.ac.uk	
3.)	The candidate must re-enter for the examination and resubmit the thesis in a revised form within a period not exceeding eighteen months. (The examiners may require a further viva examination) <input type="checkbox"/>
NB If examining a resubmitted thesis, this result is not applicable	
Further viva examination required: Yes <input type="checkbox"/> No <input type="checkbox"/> To be confirmed <input type="checkbox"/>	

Outcomes if candidate does not meet required standard for a PhD award:

4.) The candidate has met the criteria as outlined in the regulations and guidelines for examiners and be **awarded the degree of MPhil** ☐

5.) The candidate is required to make **specified minor amendments** to the examiners' satisfaction **within three months** for the award of the degree of **MPhil** ☐

i) The examiners confirm that the candidate has been provided with **a list of minor amendments or an annotated thesis** either immediately after the oral or within two weeks of the oral examination and has been asked to **send the amended thesis for confirmation to the person nominated to check the corrections:** ☐

OR

ii) The candidate has **already** made the minor amendments required to the satisfaction of **designated checker:** ☐

If outcome **(i)** has been selected, please nominate an individual to check the amendments. The person nominated should confirm these have been made satisfactorily by email to [Research Degrees](#)

Name of corrections checker:

Email address:

6.) The candidate is required to **enter for the degree of MPhil and to re-present the thesis** in a revised form **within twelve months**. (The examiners may require a further viva examination) ☐

NB If examining a resubmitted thesis, this result is not applicable

Further viva examination required:

Yes ☐ **No** ☐ **To be confirmed** ☐

7.) The candidate has **not fulfilled the requirements for a PhD or MPhil**. The candidate may not re-enter this thesis for examination. ☐

SECTION C

Examiners' Joint Report of the Viva

This section should be used to provide your opinion of the thesis and the candidate's performance in the viva.

Please use this section to list any minor corrections or major revisions required.

In his thesis "Harnessing haplotype sharing information from low coverage sequencing and sparsely genotyped data", Sam Morris centres on the use of haplotype-based methods to study human population history at different type scales by the combined use of present-day and ancient/historical genomes. The thesis forms a significant contribution to knowledge of the subject and presents a coherent argument throughout its chapters. The results represent a substantial advancement of general knowledge related to the topics of the thesis. The thesis starts with a technical chapter on the impact of low-coverage ancient DNA data on haplotype-based analysis pipelines and presents attempts to mitigate this impact by implementing different modifications to standard pipelines, e.g. integration of genotype likelihoods, and various SNP filtering criteria.

The thesis does a good job introducing the major theoretical concepts necessary to understand the work such as linkage disequilibrium (LD) and its use in the reconstruction of population history, as well as the algorithmics behind the software used to call SNPs, impute genotypes and reconstruct population history using linked or unlinked markers. The thesis is well structured in self-contained chapters but also integrates results from different chapters. Chapter 2 is the core of the thesis, where the thesis focuses on most of the technicalities and pipelines used in the multiple analyses presented throughout the work. It describes in a rigorous and honest way the bioinformatic approaches used to deal with aDNA data, even if the approaches did not work in particular cases.

The data analysis chapters present new results using publicly available data (chapter 3) and newly sequenced ancient genomes (chapter 4-5). We find particularly useful the analysis presented in chapter 3, since integration of multiple non-overlapping datasets with different sequence coverage is a common task for researchers. The results are clear and well presented and the outcomes can be easily implemented. The analysis presented in chapters 4 and 5 synthesizes the results obtained in chapter 2 in the analysis of newly produced dataset from European populations at two different geographical scales. The chapters illustrate well at the beginning the hypotheses that will be tested and describe in detail the new population history inferences for population in present-day Bavaria, Germany, and the migration of Slavic populations. The general summary and discussion of the thesis is useful, in particular the recommendations regarding data processing for low-coverage genomes.

The key aspects of novelty of the thesis are in providing new empirical evidence on the performance of Chromopainter based analyses with imputed genotype data from ancient genomes sequenced to different coverage range, in introduction of a new version Chromopainter that works on genotype likelihoods, in providing novel insights on population histories in Europe from unpublished ancient DNA sequences, and, in offering new insights as how best to detect ethnic-group level ancestries in non-European populations.

The viva took place on March on February 25 2022. It started at 12:33 and finalized at 15:30. Mr Morris had a good performance during the three hours examination. He demonstrates a very good general knowledge of the topics discussed in his thesis. Our questions covered a wide range of topics including the decay of ancient DNA, the algorithmics behind ancestry deconvolution methods, the use and integration of biobank data in genetic analysis, and the population history of different European populations. Mr Morris show a very good general knowledge of all these topics. He showed exceptional performance describing the technicalities and details of the analysis performed in his thesis, which showed his role in

the independently planning, executing and analysing the data presented in the thesis. During the discussion, Mr Morris presented the strengths of his analysis and also acknowledged the shortcomings that were pointed out by the examiners.

Overall, given the performance of Mr. Morris during the viva, and as the thesis represents a significant contribution to new knowledge and significant contribution to new knowledge, we recommend the award of a PhD subject to minor corrections.

We provide below a general and per-chapter list of minor corrections:

General

Overall, the results of the work carried out in the thesis are presented clearly and the quality of writing is good, but the thesis can gain from better proofreading, some instances show rather informal writing and not so well constructed sentences. Minor improvements can potentially be made in the areas of clarifying the details about the methods and the extent to which statements can be generalized. Some results mentioned in the text need to be linked up with relevant tables and figures.

Chapter 1

-The introduction lacks a better description of the biochemical particularities of ancient DNA. The intro would gain by describing aDNA post-mortem degradation, modelling of aDNA degradation and implementations used in the thesis to deal with it (e.g. Atlas).

p. 16 middle – 'target' haploid – is introduced as a concept that should be useful to be defined (otherwise 'haploid' has a broader meaning)

p. 16 last sentence – “*..the target is inferred to share a most recent ancestor with that reference haploid, relative to all other reference haploids,..*” – perhaps 'more recent' than 'most recent' (sensu MRCA, which all sequences share at some point), and probably only in case the other reference haploids show lower matching.

p. 21 third line a word probably missing, 'order'

p. 22. 'to confidently call heterozygous genotypes' – only heterozygous? Consider removing 'heterozygous' or replacing it with 'diploid'

p. 24 subsection title [uses of?] chromopainter [in] ancient DNA?

p. 24 first para, typos 'fruit[s]', 'difficult[y]'

p. 25, 2nd line 'the the ability'

p. 25, 2nd para, 'approximately 31' consider 'more than 30'

p. 26 4-5th lines, 'to infer identify' – one verb too many

p. 29 end of 2nd para, consider adding QUILT to methods that take BAMs and estimate likelihoods in the process

p. 30-1 section on overlaps between genotyping arrays might include reference to the ascertainment biases mentioned on p. 20, as well as commercial (Illumina vs Affymetrix) reasons for SNP selection differences. SNP ascertainment of the Human Origins Array combines, in fact, 13 different strategies (Patterson et al. 2012).

p. 31, 6th line, 'is analyse' consider replacing with 'is needed for'

Chapter 2

-In this chapter and in the thesis in general, the descriptions of pipelines are difficult to follow. We suggest the pipelines for each chapter are presented schematically through the use of figures (and tables where reporting variant numbers and precision and sensitivity by heterozygous and homozygous positions or by minor allele frequency). This will improve the readability of the chapters. Specifically, in case of the usage of GLIMPSE it needs to be specified how the ATLAS output was modified so that the input of GLIMPSE would have included all variants from the reference panel. This is a critical requirement of GLIMPSE which is particularly important for low coverage samples. The ATLAS output (unlike the

bcftools option described by the authors of GLIMPSE) by default includes only variants which have been covered at least by a read and an additional step which is not described in the thesis would have had to be implemented to introduce the reference panel positions not covered by reads in a target sample. Reporting of variant numbers that pass different steps of the pipelines

-The algorithm of ChromoPainterUncertainty is not fluently explained. Confusing writing. Confusing notation. This needs to be rewritten.

-Equation 2.1 and 2.2 are not sufficiently well explained in words. There are errors in the notation. This is fundamental for the understanding of the thesis.

p. 37, should 'Note that above (3) reduces (1)' be 'Note that above (2.2) reduces (2.1)'?

-Carefully consider whether the category Jewish can be used when comparing other populations based on geographical location and not on religious affiliation.

p. 34, 3rd line, 'each genetic region' -> 'each autosomal region'

p. 35 'Θ is the probability of a mutation occurring' – which sounds like a definition of 'mutation rate'. Would it rather be the probability of inter-individual difference in a population that is meant here?

p. 35 'a single recipient individuals copies'

p. 36 'uncertainty in imputed genotype calls'

p. 36, explain why dosage score is used rather than maxGP in defining uncertainty and whether there would be a difference between the two approaches in defining G and U? In cases where '1|1' call has low likelihood would it be clearly distinct from '1|0' cases of high likelihood when considering only dosage?

p. 37, list of ancient genomes – would be helpful to add their average depth estimates

p. 38, calling 77M sites from 1000 GP without any filter? E.g. MAC 5 typically used to exclude false positive SNPs in the reference panel.

p. 40, 'only GLIMPSE runs quickly enough to analyse SNPs with sequence-level density.' is not perhaps totally clear and correct assessment in comparison with BEAGLE

p. 40, '3202 individuals from 26 worldwide populations', note that in the appendix on page 187 there is a different number '172 worldwide populations', i.e. unclear whether what is meant is only 1000 GP data or a merger with Simons and HGDP Projects' data.

p. 41, 3rd line, 'to produced'

p. 42, explain whether when calculating precision and sensitivity only those variant positions that overlap between the target sample and the reference panel were considered

p. 45, 3rd line from bottom, 'reasonably' -> 'reasonable'

p. 47, what exactly is bias here and how measured?

p. 47, two times mitigate in one sentence

p. 50, imputation results, it would be helpful to report/table besides the accuracy other stats considered after the GLIMPSE imputation, e.g. the number of retained heterozygous sites, proportion of missingness in REF/REF, HET, ALT/ALT calls. Were het versus hom calls considered separately in sensitivity and precision estimation?

p. 51, considering the relationship with SNP frequency explain if any minor allele frequency thresholds were applied before estimating the accuracy

p. 52, "As imputation relies on matching IBD tracts between individuals, imputation accuracy increases where individuals share more IBD [104]." – explain whether this relationship is true specifically in case where target individuals share more IBD with individuals represented in the reference panel

p. 53 – what can be concluded from these analyses?

p. 62, Figure 2.8 – were all ancient samples projected?

p. 63, what imputation tool did Margaryan et al. use and how to interpret the difference of their finding that fineSTRUCTURE groupings, containing individuals as low as 0.1x coverage, were not driven by coverage?

p. 63, "Consistent with previous the results,"

p. 64, LBK excluded because anomalously poor results, in what sense?

p. 63, last section, I suppose reference to Figure 2.10 should be made?

p. 67, "However, there is a degree of noise.." needs further clarification what is meant.

p. 67, "These results suggest that imputation introduces a degree of bias into 0.1x..." – As no other imputation method was used, and the results of BEAGLE 4.0 based analyses

reported in Margaryan et al. 2020 suggest otherwise, can these results obtained with GLIMPSE within this thesis be generalized or rather should it be stated that “*These results suggest that imputation with GLIMPSE introduces a degree of bias into 0.1x...*”

p. 67, the last paragraph on ChromoPainter introduced/amplified biases – removal of poorly imputed SNPs is mentioned which emphasizes genotype imputation errors, alternatively could it be the effect of phasing errors that is behind this?

p. 68, considering the stated shifts towards the centre of the PCA it might be helpful to show the $x=0$ and $y=0$ lines on the plot.

p. 70, section 2.7.2 – besides the bias towards reference panel, can the bias to reference sequence be shown/tailed e.g. by $GP < 0.99$ call differences from high coverage calls of ALT/ALT and HET than REF/REF

p. 70, “Therefore, we can compare the amount [of] different..”

p. 71, “the largest different[ce]”

p. 72, “then calculated r -squared between copyvectors..”, correlations using r^2 ?, consider rephrasing

p. 74, RAF – unclear what it is – “variants with a low frequency” – should be specified to be either minor, derived, or non-reference allele frequency. “RAF refers to the frequency of the allele..” – of which allele?

p. 74, “cop[y]vector”; also the statement about no improvement in Table 2.3 unclear: the 1-TVD values seem to be consistently higher for r than s and u in case of 0.1x although indeed not improved in case of 0.5x.

p. 79, “As it [is] not possible..”

p. 79, unclear where the threshold for a ‘good’ SNPs being covered by a single read is coming from, considering stats in Figure 2.15, perhaps it should say ‘even’ if the threshold was one SNP? Also, the Figure 2.15 stats are about observation of two alleles which is not the same as distinguishing between het and hom genotypes.

p. 80, there is a lonely z underneath Table 2.6

p. 83, “quality score significantly can significantly”

p. 84, “Thus, [t]he results presented in”

Chapter 3

- F-STATISTIC should explain the logic of using the populations selected for a general reader without knowledge of human evolution. Please state specifically the hypothesis you are testing.

p. 86, “..at different periods in the previous three centuries..” – only just? rephrase

p. 86, “but it [is] though”

p. 88, “The latter approach is more powerful at detecting recent shared [what?] because it finds who an individual shares ancestry with overall” – consider rephrasing

p. 90, last line, UK10K is part of HRC

p. 94, “An alternative option to using Origins” – an alternative to what, Human Origins dataset?

p. 95, “the samples in the U.K. Biobank dataset do not have any associated population or ethnic group labels” – might be worth specifying that except for the Data-field 20115, Country of Birth (non-UK origin) - even though self-reported and not ethnic group level information

p. 97, last sentence ends abruptly “but had less than” and does not seem to continue on the next page

p. 99, because of specific SNP ascertainment strategies a sizeable fraction of Human Origins SNPs have a very low frequency – explain whether MAF filters were used and whether the 535,544 SNPs include also invariable positions in target samples

p. 100. “Whilst it seems counter-intuitive that there is more power using a smaller number of SNPs” – explain more explicitly whether these results suggest that there is no benefit of using more than 50K SNPs with ChromoPainter

p. 100 “Imputation methods rely on identifying reference haplotypes” – generalization that does not apply to reference-free imputation methods such as STITCH

p. 101, “one of the few west African group[s]”

p. 102, "Put together, these results suggest that using imputed data would introduce a level of bias and loss of information" – what level of bias? Some bias is indeed expected, but what impact does it have?

p. 102 "Therefore, in all later analysis, I chose to use the approximately 70,000 non-imputed SNPs" – not exactly true, rephrase, as later on, p. 116, same analyses reported also on imputed data.

p. 113-4, "There are several interesting results. For example, there are 2,263 individuals who were born in the Caribbean." – why is this interesting?

p. 118 "with recent African [ancestry] are distributed"

p. 118, mij and mj definitions check

p. 120, "Further, I found that using imputed genotypes may significantly reduce the power of a painting.." – explain what exactly does this reduced power mean, and, how generally this claim can be made – about all imputation methods, imputation from array based genotype data, specific array? Add more detail, similarly to other statements made, e.g. "I did not find evidence for structure in how African ancestry was distributed across the U.K., based on the testing centre that participants registered at."

Chapter 4

p. 124 "SNP-based studies have shown that there is only very weak substructure [152]. Questions remain as to the origin of this East-West structure; is it recent structure, or does it persist to the Middle Ages or earlier?" – unclear about the East-West structure here as the cited paper reports "that there exists a low genetic differentiation between the samples along a north-south gradient within Germany", and, unclear about the directionality of time in 'persist to the Middle Ages'

p. 127, "To determine the mtDNA and y-chromosome haplogroups for each newly sequenced ancient sample, I used Haplogrep.." note that this tool only works on mtDNA

p. 128, PLINK PCA, unclear how the pre-imputation genotypes were obtained – were these called in form of haploid/diploid genotypes or genotype likelihoods and as PLINK does not take likelihoods or haploid calls how was the uncertainty dealt with? If directly called genotypes

p. 128, Table 4.1 – date estimates given in which units, before present? Or rather BC, except for the medieval?

p. 130, Table 4.2 – mention source (HellBus) in the table footnote.

p. 134, "with other literature samples" – with other 'published samples'?

p. 135, "fineSTRUCTURE grouped Erg1 with two samples from Upper Palaeolithic/Neolithic Italy" – refer to figure/table where these results are shown, Figure 4.4.?

p. 136, Figure title "Principle component analysis of genotype matrix" – perhaps instead of genotype matrix a more informative name for the presented dataset can be used?

p. 138, the $Z > 3$ score for f4(Erg1, Din2, WHG, Mbuti) interpreted as a possible outcome of multiple testing – how does this relate to WHG admixture results in the next section 4.3.3? consider rephrasing in this context here more as an uncertainty whether admixture or multiple testing error rather than resolving for the latter

p. p. 139, qpAdm result on Erg1 – specify where shown, Figure 4.7? not referred to; Erg2 qpAdm results showing no HG admixture mentioned, but the SOURCEFIND results show some level of admixture, explain and rephrase the statements in the text

p. 140, the same, reference to relevant figures needed for MOSAIC and SOURCEFIND results within the first paragraph where the results are first mentioned.

p. 146, "from a source closest to an Alamannic-Frankish sample" – from which published source?

p. 146, "The estimated Fst" – how, where explained?

p. 148, "However, the two Germanic samples fall into a fineSTRUCTURE cluster with a set of contemporaneous samples from Northern Europe, including 10-11th century Vikings from Estonia, Sweden and Iceland," – where presented? On Figure 4.4 no Vikings are shown. If not shown anywhere include 'data not shown'. Which sources used for Early Slavic samples?

p. 148 "and has was [been] dated"

p. 150, “whilst DIN2 showed no evidence of admixture” – but it does according to figures 4.7 and 4.8, modify text to explain

p. 150, on the origin of the southern component in the Cherry Tree Cave Iron Age samples “The most plausible source of this ancestry is from Italy, with the best source in the dataset being the cluster of Renaissance samples from Antonio et al (2019) [59], date to between 282 - 354 AD. How exactly does this work considering the Cherry Tree Cave samples are from 6th century BC, i.e. more than 800 years older? Consider rephrasing/modifying the text. Similarly the time scale does not seem to fit with the Lombards in Italy and their migration to Southern Germany, at least not from Italy, unless considering the possibility of Lombard ancestors before them reaching Italy? But then this would not explain the presence of Italian ancestry in Iron Age Germans of the Cherry Tree Cave.

p. 151, “The arrival of Yamnaya-like ancestry from this early (2762BC) [period? Or what exactly?] represents..”

Chapter 5

pp. 152-4 reference to Figure 5.1 should be made in text where talking about the three Slavic groupings, currently this figure, which is useful for presenting the relevant geographic context, is not cited anywhere in the text.

p. 156, questions – if the material is specifically focused on Czech samples then it would make more sense to formulate the questions also accordingly, e.g. “Is there evidence of genetic change between Migration period and Early Middle Ages in the area of present-day Czech Republic”. “Do the labels “Migration Period” and “Early Middle Ages” make sense” is too perhaps a bit too vague question and should be rephrased. Can the evidence from these 17 samples indeed be used to show no continuity, explain whether continuity is your H0 and change H1?

p. 156, methods “Whole genome sequence data were generated ..”. Add a disclaimer of the data generation by collaborators and explain what data were given and which initial processing was done by the author. Explain how the genotypes (likelihoods) were called (ATLAS?) and whether any filters were applied.

p. 157, Table 1, all 17 samples appear to have a relatively narrow coverage range, 5-7.3x, what was the motivation for choosing this particular coverage? And, how many samples in total were examined (assuming not all of them could be brought up to this coverage)? Would average depth be more appropriate term to use in this case than coverage?

p. 157, “imputation and phasing pipeline to generate genotype likelihoods..” – does the imputation output indeed report genotype likelihoods (GL) or genotype probabilities (GP), which (the latter) would consider the haplotypes in the reference panel

p. 158, “I retained only the 500,000 markers with the lowest amount of missingness”, by what cut-off value, was this done after the MS-POBI-HellBus merging – how many SNPs were retained after the merging?

p. 161, “the Migration Period samples are heterogeneous are [and] not likely to originate”

p. 162, a table without title or footnote. Is this table related to the last paragraph of the methods, “..and the following sets as surrogates”?

p. 164, “LIB4 and LIB5 cluster together with Early Iron Age and Renaissance samples from Italy” – cluster is perhaps a too strong word considering almost equal distance to Anatolia Neolithic and the absence of other (than Bavarian and Italian) Iron Age samples from Europe on the plots, consider rephrasing

p. 165, “Historical evidence cites alliances between Slavs and Lombards in the 5th century [195].” – consider including and discussing in this context the direct ancient DNA evidence from the Longobards (Amorim et al. 2018 Nat Comm)?

p. “LIB12 displays ancestry which is more typical of the preceding Central European Bronze Age,” – note difference between the LIB12 neighbours on Fig. 5.2 and 5.3? In the latter it is closest to bavaria_zuzana and Lombards.

p. 165, “All 12 samples cluster in the same fineSTRUCTURE group (named Slavic Early Middle Age II)” – is this with reference to Fig. 5.3? – cannot find such label there, the label shows on Fig. 5.2 but the red samples cover much wider space including Bavarian Bronze Age and GoldenHordeEuro as well as Mesolithic and other labels.

p. 165, "SOURCEFIND showed .." – is there a figure or a table? Also the claimed relationships are difficult to follow on Figure D.7.

p. 169, "affinity to present-day Greek individuals" is difficult to tell from Fig. 5.7

p. 174, Fig. 5.10, 2-way admixture involving which sources? Which samples, groups considered as the targets of admixture? What is the difference between green and blue boxes?

p. 175, "source best represented by present-day Uyghurs" – what could be the historical event that could explain this? Can Finno-Ugric (FU) substrate, that has been estimated to have Yukaghir component due to an Iron Age migration as the likeliest source, be ruled out here as the source?

p. 176, continuity exists or not statements – what exactly do these mean? No continuity between modern and Migration Period samples, does the statement mean complete replacement (no continuity at all) or significant level of change (continuity and change by drift only model can be ruled out)?

p. 177, "Alternatively, it is possible these individuals are a leftover from a historically described Lombard migration that moved from Northern Germany through Czechia, Slovakia, Hungary and ended up in Lombardia (Z. Hofmanova, personal communication)." – such migration can explain Germanic ancestry in northern Italy. Unclear how could it explain Mediterranean ancestry in Czech region at such early dates. Can migrations related to Roman armies be ruled out? Also, are there distinct Italian and Greek sources at play, as discussed in text in relation to different analyses? Modify text

Chapter 6

p. 179, "copyvectors of >0.5x downsamples show a high correspondence.." – with or without imputation? If imputed then worth also mentioning GLIMPSE

p. 179, "I also found evidence of imputation bias towards the reference (Fig. 2.13)." Worth specifying whether it is reference sequence or reference panel bias. Might be helpful to clarify this also with regards to the statement on page 133, with reference to [48], that imputation reduces reference sequence bias. In light of the negative correlations between precision and age of the ancient target samples it might be helpful to expand on whether the mentioned biases are dependent on how distant the chosen reference panel is to a sample.

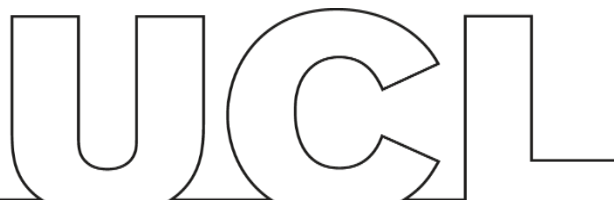
p. 181, "Iron Age Italian sourced"

p. 181, "which may be related the migrations of Lombard populations" – rephrase or clarify how Italian ancestry in pre-Roman Iron Age could be related to Lombard migration more than 800 years later. Lombards would have brought 'Germanic/Scandinavian' ancestry to North Italy rather than Italian ancestry to Germany.

p. 182, recommendations should be more specific, e.g. recommendations on GLIMPSE-based imputation more specifically rather than all imputation methods generally.

SECTION D	
Examiners' Signature Confirming Result of the Examination	
Declaration: We confirm this is the joint examiners' report for the candidate named above.	
Signed:	
Please print name:	Dr Toomas Kivisild
Date:	March 3rd 2022
Signed:	
Please print name:	Dr Hernán A. Burbano
Date:	March 3 2022
Signed*:	
Please print name:	
Date:	
(* Only for use in the exceptional cases when UCL has approved the appointment of a third examiner)	

If electronic signatures are to be used, these must be the image of a hand written signature. **We cannot accept a typed name in place of a signature.**



Research Degree Examination: Examiner's Preliminary Report

You must complete an independent preliminary report after reading the thesis but before conferring with your co-examiner. Please exchange your report with your co-examiner before the candidate's viva and sent a copy by email to [Research Degrees](mailto:researchdegrees@ucl.ac.uk).

If you have not submitted the preliminary report before the viva, please include a copy with the joint report.

Candidate and Thesis Details	
Candidate's Full Name:	Sam Morris
Student Number:	
Degree Award (PhD, etc.)	PhD
Thesis Title:	Harnessing haplotype sharing information from low coverage sequencing and sparsely genotyped data

Report
<p>The thesis is focused on genome-scale analyses of human genetic variation across a range of populations with a focus on haplotype-based methods. The key questions that are being addressed are about the best use of low coverage sequence data, as available from ancient DNA studies, or genotype array data with limited variant numbers, in the study of population structure at fine resolution. The results represent substantial advancement of general knowledge related to these questions.</p> <p>The work behind the results presented in chapters 2-5 is based on original research. This includes novel results based on down-sampling experiments, imputation analyses and</p>

simulation work (chapter 2), analyses of publicly available data (chapter3) and unpublished data (chapters 4-5). The key aspects of novelty of the thesis are in providing new empirical evidence on the performance of Chromopainter based analyses with imputed genotype data from ancient genomes sequenced to different coverage range, in introduction of a new version Chromopainter that works on genotype likelihoods, in providing novel insights on population histories in Europe from unpublished ancient DNA sequences, and, in offering new insights as how best to detect ethnic-group level ancestries in non-European populations.

The thesis is structured into six individual chapters including four that are results based, followed by a general discussion and future perspectives chapter.

Overall, the results of the work carried out in the thesis are presented clearly and I was unable to identify any major flaws in the work that would undermine the main conclusions drawn in the presented theses. However, as detailed in appendix, minor improvements can potentially be made in the areas of clarifying the details about the methods and the extent to which statements can be generalized. Overall, as the thesis represents substantial scientific novelty and significant contribution to new knowledge, I recommend the award of PhD subject to minor corrections.

Recommended corrections and questions for the viva are listed as an appendix.

Examiner Details	
Name:	Toomas Kivisild
Signature:	
Date:	24/02/2022

February 24th, 2022

Appendix to the PhD Examiner's report by Toomas Kivisild on the thesis of Sam Morris:

"Harnessing haplotype sharing information from low coverage sequencing and sparsely genotyped data"

Questions for viva and recommended corrections

Chapter 1

p. 16 middle – **'target' haploid** – is introduced as a concept that should be useful to be defined (otherwise 'haploid' has a broader meaning)

p. 16 last sentence – "*..if a target haploid matches a DNA segment to a particular reference haploid for a genomic region, the target is inferred to share a most recent ancestor with that reference haploid, relative to all other reference haploids,..*" – **perhaps 'more recent' than 'most recent'** (sensu MRCA, which all sequences share at some point), and probably only in case the other reference haploids show lower matching.

p. 21 third line a word probably missing, 'order'

p. 22. 'to confidently call heterozygous genotypes' – only heterozygous? Consider removing 'heterozygous' or replacing it with 'diploid'

p. 24 subsection title [uses of?] chromopainter [in] ancient DNA?

p. 24 first para, typos 'fruit[s]', 'difficult[y]'

p. 25, 2nd line 'the ~~the~~ ability'

p. 25, 2nd para, 'approximately 31' consider 'more than 30'

p. 26 4-5th lines, 'to infer identify' – one verb too many

p. 26, question about the use of chromopainter in archaics – how accurately can we phase the archaic genomes and does it matter?

p. 28 last section, besides low coverage are there other concern issues with aDNA, e.g. short reads, damage, contamination?

p. 29 end of 2nd para, consider adding QUILT to methods that take BAMs and estimate likelihoods in the process

p. 30-1 section on overlaps between genotyping arrays might include reference to the ascertainment biases mentioned on p. 20, as well as commercial (Illumina vs Affymetrix) reasons for SNP selection differences. SNP ascertainment of the Human Origins Array combines, in fact, 13 different strategies (Patterson et al. 2012).

p. 31, 6th line, 'is analyse' consider replacing with 'is needed for'

Chapter 2

p. 34, 3rd line, 'each genetic region' -> 'each autosomal region'

O&N IV Herestraat 49 - box 602
3000 Leuven, Belgium

Tel: +32 16 19 33 55
E-mail: toomas.kivisild@kuleuven.be

- p. 35 ' Θ is the probability of a mutation occurring' – which sounds like a definition of 'mutation rate'. Would it rather be the probability of inter-individual difference in a population that is meant here?
- p. 35 'a single recipient individual's copies'
- p. 36 'uncertainty in imputed genotype calls'
- p. 36, why use dosage score (is this the diploid dosage field in vcf?) rather than maxGP in defining uncertainty? Would there be a difference between the two approaches in defining G and U? In cases where '1|1' call has low likelihood would it be clearly distinct from '1|0' cases of high likelihood when considering only dosage?**
- p. 37, should 'Note that above (3) reduces (1)' be 'Note that above (2.2) reduces (2.1)'?
- p. 37, list of ancient genomes – would be helpful to add their average depth estimates
- p. 38, calling 77M sites from 1000 GP without any filter? E.g. MAC 5 typically used to exclude false positive SNPs in the reference panel.
- p. 39, section 2.2.3 – was the availability of shotgun rather than capture data one of the important inclusion criteria? What about treatment methods, e.g. UDG?**
- p. 40, 'only GLIMPSE runs quickly enough to analyse SNPs with sequence-level density.' is not perhaps totally clear and correct assessment in comparison with BEAGLE
- p. 40, '3202 individuals from 26 worldwide populations', note that in the appendix on page 187 there is a different number '172 worldwide populations', i.e. unclear whether what is meant is only 1000 GP data or a merger with Simons and HGDP Projects' data.
- p. 41, 3rd line, 'to produced'
- p. 41, after GLIMPSE, were any filters applied on genotypes, e.g. by the GP score, to assess accuracy/sensitivity/precision?**
- p. 42, when calculating precision and sensitivity where only those variant positions considered that overlap between the target sample and the reference panel?**
- p. 45, 3rd line from bottom, 'reasonably' -> 'reasonable'
- p. 46, was the effect of projection on PCA placements assessed?**
- p. 47, what exactly is bias here and how measured?**
- p. 47, two times mitigate in one sentence
- p. 49, in relation to mimicking the ancient DNA data at 0.15x coverage, the described simulated scenario with Human Origins modern data, was the used data in pseudohaploid format before submission to the Sanger Imputation Service?**
- p. 50, imputation results, besides the accuracy were other stats considered after the GLIMPSE imputation, e.g. the number of retained heterozygous sites, proportion of missingness in REF/REF, HET, ALT/ALT calls? Were het versus hom calls considered separately in sensitivity and precision estimation?**
- p. 51, considering the relationship with SNP frequency were any minor allele frequency thresholds applied before estimating the accuracy?**
- p. 52, "As imputation relies on matching IBD tracts between individuals, imputation accuracy increases where individuals share more IBD [104]." – perhaps this relationship is true specifically in case where target individuals share more IBD with individuals represented in the reference panel**
- p. 53 – what can be concluded from these analyses?**
- p. 62, Figure 2.8 – were all ancient samples projected?**

p. 63, what imputation tool did Margaryan et al. use and how to interpret the difference of their finding that fineSTRUCTURE groupings, containing individuals as low as 0.1x coverage, were not driven by coverage?

p. 63, "Consistent with previous ~~the~~ results,"

p. 64, LBK excluded because anomalously poor results, in what sense?

p. 63, last section, I suppose reference to Figure 2.10 should be made?

p. 67, "However, there is a degree of noise.." needs further clarification what is meant.

p. 67, "These results suggest that imputation introduces a degree of bias into 0.1x..." – As no other imputation method was used, and the results of BEAGLE 4.0 based analyses reported in Margaryan et al. 2020 suggest otherwise, can these results obtained with GLIMPSE within this thesis be generalized or rather should it be stated that "These results suggest that imputation with GLIMPSE introduces a degree of bias into 0.1x..."

p. 67, the last paragraph on ChromoPainter introduced/amplified biases – removal of poorly imputed SNPs is mentioned which emphasizes genotype imputation errors, alternatively could it be the effect of phasing errors that is behind this?

p. 68, considering the stated shifts towards the centre of the PCA it might be helpful to show the x=0 and y=0 lines on the plot.

p. 70, section 2.7.2 – besides the potential bias towards reference panel, was the bias to reference sequence tested? E.g. were GP<0.99 calls more frequent for true ALT/ALT and HET than REF/REF calls?

p. 70, "Therefore, we can compare the amount [of] different.."

p. 71, "the largest different[ce]"

p. 72, "then calculated r-squared between copyvectors..", correlations using r²?, consider rephrasing

p. 74, RAF – unclear what it is – "variants with a low frequency" –should be specified to be either minor, derived, or non-reference allele frequency. "RAF refers to the frequency of the allele.." – of which allele?

p. 74, "cop[y]vector"; also the statement about no improvement in Table 2.3 unclear: the 1-TVD values seem to be consistently higher for r than s and u in case of 0.1x although indeed not improved in case of 0.5x.

p. 74, last paragraph – what were the proportions of variants retained for 0.1x coverage after the GP 0.99 filter, are these results tabled anywhere? And, how to explain the dramatic drop of TVD values after GP 0.99 filter in Table 2.3?

p. 79, "As it [is] not possible.."

p. 79, unclear where the threshold for a 'good' SNPs being covered by a single read is coming from, considering stats in Figure 2.15, perhaps it should say 'even' if the threshold was one SNP? Also, the Figure 2.15 stats are about observation of two alleles which is not the same as distinguishing between het and hom genotypes.

p. 80, there is a lonely z underneath Table 2.6

p. 83, second paragraph, the effects of minor allele frequency and GP filtering on ChromoPainter results, and previous studies showing significant effect – given differences of the pipelines (BEAGLE vs GLIMPSE) it might be interesting to know what the post-filtering variant numbers and accuracy estimates are as obtained here

p. 83, "quality score ~~significantly~~ can significantly"

p. 84, "Thus, [t]he results presented in"

p. 84-5, future perspectives – what about other imputation methods, ancient reference panels?, analyses of other types of biases (e.g. ref-bias, het/hom biases)?

Chapter 3

p. 86, “..at different periods in the previous three centuries..” – only just?

p. 86, “but it [is] though”

p. 88, “The latter approach is more powerful at detecting recent shared [what?] because it finds who an individual shares ancestry with overall” – consider rephrasing

p. 90, last line, UK10K is part of HRC

p. 90, was the UK Biobank’s country of origin information also used?

p. 94, “An alternative option to using Origins” – an alternative to what, Human Origins dataset?

p. 95, “the samples in the U.K. Biobank dataset do not have any associated population or ethnic group labels” – what about Data-field 20115, Country of Birth (non-UK origin)? Even though self-reported and not ethnic group level information, would it still be usable as a geographic proxy?

p. 97, last sentence ends abruptly “but had less than” and does not seem to continue on the next page

p. 99, note that because of specific SNP ascertainment strategies a large number of Human Origins SNPs have a very low frequency – was MAF filtering considered or do the 535,544 SNPs include also invariable positions?

p. 100. “Whilst it seems counter-intuitive that there is more power using a smaller number of SNPs” – do these results suggest that there is no benefit of using more than 50K SNPs with ChromoPainter?

p. 100 “imputing data results in a loss of information” – can this be generalized to this extent?

p. 100 “Imputation methods rely on identifying reference haplotypes” – generalization that does not apply to reference-free imputation methods such as STITCH

p. 101, “one of the few west African group[s]”

p. 101, “Fig 3.3 shows the amount of differential haplotype donation on a per-population basis” – what impact this -0.00006 to 0.00008 chunk length difference has on downstream analyses, e.g. ancestry inference?

p. 102, “Put together, these results suggest that using imputed data would introduce a level of bias and loss of information” – what level of bias? Some bias is indeed expected, but what impact does it have?

p. 102 “Therefore, in all later analysis, I chose to use the approximately 70,000 non-imputed SNPs” – not exactly true? Later on, p. 116, same analyses reported also on imputed data.

p. 113-4, “There are several interesting results. For example, there are 2,263 individuals who were born in the Caribbean.” – why is this interesting?

p. 116 “This is clear on the figure, as there are many population bunched around the 2% point for the imputed dataset;” – explain

p. 118 “with recent African [ancestry] are distributed”

p. 118, mij and mj definitions check

p. 120, “Further, I found that using imputed genotypes may significantly reduce the power of a painting..” – how measured and what exactly does this reduced power mean, and, how generally this claim can be made – about all imputation methods, imputation from array based genotype data, specific array? In other context, in contrast, the claims are quite detailed and precisely made, e.g. “I did not find evidence for structure in how African ancestry was distributed across the U.K., based on the testing centre that participants registered at.”

O&N IV Herestraat 49 - box 602
3000 Leuven, Belgium

Tel: +32 16 19 33 55
E-mail: toomas.kivisild@kuleuven.be

Chapter 4

p. 124 “SNP-based studies have shown that there is only very weak substructure [152]. Questions remain as to the origin of this East-West structure; is it recent structure, or does it persist to the Middle Ages or earlier?” – unclear about the East-West structure here as the cited paper reports “that there exists a low genetic differentiation between the samples along a north-south gradient within Germany”, and, unclear about the directionality of time in ‘persist to the Middle Ages’

p. 127, “To determine the mtDNA and y-chromosome haplogroups for each newly sequenced ancient sample, I used Haplogrep..” note that this tool only works on mtDNA

p. 128, PLINK PCA, unclear how the pre-imputation genotypes were obtained – were these called in form of haploid/diploid genotypes or genotype likelihoods and as PLINK does not take likelihoods or haploid calls how was the uncertainty dealt with? If directly called genotypes, with what filter and why not use projection approach typically taken with aDNA PCAs?

p. 128, Table 4.1 – date estimates given in which units, before present? Or rather BC, except for the medieval?

p. 130, Table 4.2 – mention source (HellBus) in the table footnote.

p. 134, “with other literature samples” – with other ‘published samples’?;

p. 135, “fineSTRUCTURE grouped Erg1 with two samples from Upper Palaeolithic/Neolithic Italy” – refer to figure/table where these results are shown, Figure 4.4.?

p. 136, Figure title “Principle component analysis of genotype matrix” – perhaps instead of genotype matrix a more informative name for the presented dataset can be used?

p. 138, the $Z > 3$ score for $f_4(\text{Erg1}, \text{Din2}, \text{WHG}, \text{Mbuti})$ interpreted as a possible outcome of multiple testing – how does this relate to test results in the next section 4.3.3?

p. p. 139, qpAdm result on Erg1 – shown where, Figure 4.7? not referred to; Erg2 qpAdm results showing no HG admixture mentioned, but not the SOURCEFIND results which show some level of admixture, explain

p. 140, the same, reference to relevant figures needed for MOSAIC and SOURCEFIND results within the first paragraph where the results are first mentioned.

p. 146, “from a source closest to an Alamannic-Frankish sample” – from which published source?

p. 146, “The estimated F_{st} ” – how, where explained?

p. 148, “However, the two Germanic samples fall into a fineSTRUCTURE cluster with a set of contemporaneous samples from Northern Europe, including 10-11th century Vikings from Estonia, Sweden and Iceland,” – where presented? On Figure 4.4 no Vikings are shown. Which sources used for Early Slavic samples?

p. 148 “and has ~~was~~ [been] dated”

p. 150, “whilst DIN2 showed no evidence of admixture” – but it does according to figures 4.7 and 4.8, explain

p. 150, on the origin of the southern component in the Cherry Tree Cave Iron Age samples “The most plausible source of this ancestry is from Italy, with the best source in the dataset being the cluster of Renaissance samples from Antonio et al (2019) [59], date to between 282 - 354 AD. How exactly does this work considering the Cherry Tree Cave samples are from 6th century BC, i.e. more than 800 years older? Similarly the time scale does not seem to fit with the Lombards in Italy and their migration to Southern Germany, at least not from Italy, unless considering the possibility of Lombard ancestors before them reaching Italy? But then this would not explain the presence of Italian ancestry in Iron Age Germans of the Cherry Tree Cave.

p. 151, “The arrival of Yamnaya-like ancestry from this early (2762BC) [what?] represents..”

O&N IV Herestraat 49 - box 602
3000 Leuven, Belgium

Tel: +32 16 19 33 55
E-mail: toomas.kivisild@kuleuven.be

Chapter 5

pp. 152-4 reference to Figure 5.1 should be made in text where talking about the three Slavic groupings, currently this figure, which is useful for presenting the relevant geographic context, is not cited anywhere in the text.

p. 156, questions – if the material is specifically focused on Czech samples then it would make more sense to formulate the questions also accordingly, e.g. “Is there evidence of genetic change between Migration period and Early Middle Ages in the area of present-day Czech Republic”. “Do the labels “Migration Period” and “Early Middle Ages” make sense” is too perhaps a bit too vague question. Can the evidence from these 17 samples indeed be used to show no continuity, is continuity H0 and change H1 (which can be rejected)?

p. 156, methods “Whole genome sequence data were generated ..”. As there is disclaimer, as in 4.2.1 about the data generation by collaborators, it is to be assumed that the data were generated by the author of the thesis, in which case more details about the extraction, library preparation and quality checks should be provided. It is also not clear how the genotypes (likelihoods) were called (ATLAS?) and whether any filters were applied.

p. 157, Table 1, all 17 samples appear to have a relatively narrow coverage range, 5-7.3x, what was the motivation for choosing this particular coverage? And, how many samples in total were examined (assuming not all of them could be brought up to this coverage)? Would average depth be more appropriate term to use in this case than coverage?

p. 157, “imputation and phasing pipeline to generate genotype likelihoods..” – does the imputation output indeed report genotype likelihoods (GL) or genotype probabilities (GP), which (the latter) would consider the haplotypes in the reference panel

p. 158, “I retained only the 500,000 markers with the lowest amount of missingness”, by what cut-off value, was this done after the MS-POBI-HellBus merging – how many SNPs were retained after the merging?

p. 161, “the Migration Period samples are heterogeneous ~~are~~ [and] not likely to originate”

p. 162, a table without title or footnote. Is this table related to the last paragraph of the methods, “..and the following sets as surrogates”?

p. 164, “LIB4 and LIB5 cluster together with Early Iron Age and Renaissance samples from Italy” – cluster is perhaps a too strong word considering almost equal distance to Anatolia Neolithic and the absence of other (than Bavarian and Italian) Iron Age samples from Europe on the plots

p. 165, “Historical evidence cites alliances between Slavs and Lombards in the 5th century [195].” – why not include and discuss the direct ancient DNA evidence from the Longobards (Amorim et al. 2018 Nat Comm)?

p. “LIB12 displays ancestry which is more typical of the preceding Central European Bronze Age,” – note difference between the LIB12 neighbours on Fig. 5.2 and 5.3? In the latter it is closest to bavaria_zuzana and Lombards.

p. 165, “All 12 samples cluster in the same fineSTRUCTURE group (named Slavic Early Middle Age II)” – is this with reference to Fig. 5.3? – cannot find such label there, the label shows on Fig. 5.2 but the red samples cover much wider space including Bavarian Bronze Age and GoldenHordeEuro as well as Mesolithic and other labels.

p. 165, “SOURCEFIND showed ..” – is there a figure or a table? Also the claimed relationships are difficult to follow on Figure D.7.

p. 169, “affinity to present-day Greek individuals” is difficult to tell from Fig. 5.7

p. 174, Fig. 5.10, 2-way admixture involving which sources? Which samples, groups considered as the targets of admixture? What is the difference between green and blue boxes?

p. 175, “source best represented by present-day Uyghurs” – what could be the historical event that could explain this? Can Finno-Ugric (FU) substrate, that has been estimated to have Yukaghir component due to an Iron Age migration as the likeliest source, be ruled out here? I.e. do FU populations and the Czech have a different Asian source? Also, consider that Uyghurs are a population with mixed ancestry, roughly 50% East Asian and 50% Europe related ancestry – how would that affect the model and ancestry estimation in comparison to other populations on the list who are less admixed?

p. 176, continuity exists or not statements – what exactly do these mean? No continuity between modern and Migration Period samples, does the statement mean complete replacement (no continuity at all) or significant level of change (continuity and change by drift only model can be ruled out)?

p. 177, “Alternatively, it is possible these individuals are a leftover from a historically described Lombard migration that moved from Northern Germany through Czechia, Slovakia, Hungary and ended up in Lombardia (Z. Hofmanova, personal communication).” – such migration can explain Germanic ancestry in northern Italy. Unclear how could it explain Mediterranean ancestry in Czech region at such early dates. Can migrations related to Roman armies be ruled out? Also, are there distinct Italian and Greek sources at play, as discussed in text in relation to different analyses?

Chapter 6

p. 179, “copyvectors of >0.5x downsamples show a high correspondence..” – with or without imputation? If imputed then worth also mentioning GLIMPSE

p. 179, “I also found evidence of imputation bias towards the reference (Fig. 2.13).” **Worth specifying whether it is reference sequence or reference panel bias.** Might be helpful to clarify this also with regards to the statement on page 133, with reference to [48], that imputation reduces reference sequence bias. **It should be also made clear that imputation attempts were only made on samples for which the used reference panel was not adequate (even for Europe, considering what we know of recent population histories), i.e. very old ancient DNA samples and no attempt was made to test the reference panel effects on samples down-sampled either from modern populations represented in the reference panel or ancient samples from recent past.**

p. 180, as only 40,000 SNPs are needed, **is there any value for running ChromPainter with more SNPs if available?**

p. 181, “Iron Age Italian source”

p. 181, “which may be related the migrations of Lombard populations” – **it should be clarified how Italian ancestry in pre-Roman Iron Age could be related to Lombard migration more than 800 years later that would have brought ‘Germanic/Scandinavian’ ancestry to North Italy rather than Italian ancestry to Germany.**

p. 182, recommendations should be more specific, e.g. recommendations on GLIMPSE-based imputation more specifically rather than all imputation methods generally; **it should perhaps be specified also that undesirable biases to reference panel may appear when imputing target samples for which the reference panels are not suited (whether or not there would be biases for samples for which reference panel is locally adequate has not been explored in the thesis)?**

p. 184, last paragraph, how realistic is it that ancient DNA coverage issues can be resolved?

O&N IV Herestraat 49 - box 602
3000 Leuven, Belgium

Tel: +32 16 19 33 55
E-mail: toomas.kivisild@kuleuven.be



Toomas Kivisild, February 24th, 2022

O&N IV Herestraat 49 - box 602
3000 Leuven, Belgium

Tel: +32 16 19 33 55
E-mail: toomas.kivisild@kuleuven.be



Research Degree Examination: Examiner's Preliminary Report

You must complete an independent preliminary report after reading the thesis but before conferring with your co-examiner. Please exchange your report with your co-examiner before the candidate's viva and sent a copy by email to [Research Degrees](mailto:researchdegrees@ucl.ac.uk).

If you have not submitted the preliminary report before the viva, please include a copy with the joint report.

Candidate and Thesis Details	
Candidate's Full Name:	Sam Morris
Student Number:	17114784
Degree Award (PhD, etc.)	PhD
Thesis Title:	Harnessing haplotype sharing information from low coverage sequencing and sparsely genotyped data

Report
<p>In his thesis "Harnessing haplotype sharing information from low coverage sequencing and sparsely genotyped data", Sam Morris centres on the use of haplotype-based methods to study human population history at different time scales by the combined use of present-day and ancient/historical genomes. The thesis forms a significant contribution to knowledge of the subject and presents a coherent argument throughout its chapters. The thesis starts with a technical chapter on the impact of low-coverage aDNA data on haplotype-based analysis pipelines and presents attempts to mitigate this impact by implementing different modifications to standard pipelines, e.g. integration of genotype likelihoods, and various SNP filtering criteria. The quality of writing is OK but the thesis would have gained from better proofreading, some instances show rather informal writing and not so well constructed sentences, e.g. intermixing of active and passive voice.</p> <p>The thesis does a good job introducing the major theoretical concepts necessary to understand the work such as linkage disequilibrium (LD) and its use in the</p>

reconstruction of population history, as well as the algorithmics behind the software used to call SNPs, impute genotypes and carry out population history reconstruction using linked or unlinked markers. However, it lacks a better description of the biochemical particularities of ancient DNA. The intro would gain by describing aDNA post-mortem degradation, modelling of aDNA degradation and implementations used in the thesis to deal with it (e.g. Atlas).

The thesis is well structured in self-contained chapters but also integrates results from different chapters. Chapter 2 is the core of the thesis, where the thesis focuses on most of the technicalities and pipelines used in the multiple analyses presented throughout the work. It describes in a rigorous and honest way the bioinformatic approaches used to deal with aDNA data, even if the approaches did not work in particular cases. However, in this chapter and in the thesis in general, the descriptions of pipelines are difficult to follow. I suggest the pipelines for each chapter are presented schematically through the use of figures. This will improve the readability of the chapters. Another aspect that can improve readability is a better placement of figures along the text. This would be a doable task given that the thesis was written in LaTeX.

The data analysis chapters present new results using publicly available data (chapter 3) and newly sequenced ancient genomes (chapter 4-5). I find particularly useful the analysis presented in chapter 3, since integration of multiple non-overlapping datasets with different sequence coverage is a common task for researchers. The results are clear and well presented and the outcomes can be easily implemented. The analysis presented in chapters 4 and 5 synthesize the results obtained in chapter 2 in the analysis of newly produced dataset from European populations at two different geographical scales. The chapters illustrate well at the beginning the hypotheses that will be tested and describe in detail the new population history inferences for population in present-day Bavaria, Germany, and the migration of Slavic populations.

The general summary and discussion of the thesis is useful, in particular the recommendations regarding data processing for low-coverage genomes.


Some of the questions for the viva. They will be followed with additional questions and discussion of major results and figures.

General Introduction:

- Please describe the evolution of LSM and Chromopainter - Algorithmic implementations e.g. BWT etc. Describe also advantages and limitations?
- Explain limitations of BWT (is all vs all a limitation)? Wouldn't this approach generate more unbiased results than approaches using pre-defined palettes?
- Please describe in more detail GIA (gain of informativeness for assignment)?
- Advantages and disadvantages of shared drift-based methods such as F-statistics?
- Please describe the advantages and disadvantages of pseudo-haploid calls?
- Broader genetic differentiation such as introgression between humans and archaic hominins do not require sophisticated methods. Which level of genetic differentiation will require haplotype-based methods?
- You used Atlas (incorporation of PMD) in all your dataset? Does your datasets include any single-stranded libraries? How the library preparation method will affect the substitution pattern in aDNA reads?

Analysis chapters:

- Can you explain the differences between ChromoPainter and ChromoPainterU: Page 36
Explain GL, D and U?
- Are there any of the high-coverage genomes produced by single-stranded libraries?
- Did you use other GL software such as ANGSD? Would differences in GL models affect the results?
- Is there any real possibility to call aDNA private genetic variation using aDNA?
- Evaluation of PCA rely on projections? Problems with this methodology?
- You describe a correlation between imputation accuracy and allele frequency? Why did you not follow up?
- Explain the regression to the prior concept?
- PCA location of 0.1X coverage. Why are they pulled to the origin of the PCA?
- What about using Chromopainter for low coverage genomes from haploid organisms? E.g. haploid fungi.
- Explain the problems with transferring polygenic risk scores to other populations, you said that LD structure is partially responsible, which other factors influence the lack of transferability?
- Walk us through the permutation test of page 102?
- Explain the relationship between percentage of ancestry from a particular African country and migration history in the UK?
- Describe in general terms major ancestry contributions to European populations?

Examiner Details	
Name:	Hernán A. Burbano
Signature:	
Date:	February 24 2022