GENOMICS AND COMPUTATIONAL BIOLOGY Vol. 4 No. 2 (2018): e100044

Research Article

Of Wheat and Men: Changes in Genetic Markers for Celiac Disease Over Time

Charlotte Hewel[1,2,*], Julia Kaiser[1,2], Jan Linke[1,2]

[1] Institute for Organismic and Molecular Evolution (iOME), Johannes Gutenberg University Mainz, Institute of Anthropology, Anselm-Franz-von-Bentzel-Weg 7, 55128 Mainz, Germany
[2] These Authors contributed equally to this work

[*] Correspondence: Email: chewel@students.uni-mainz.de

Received 2017-10-27; Accepted 2018-01-28

ABSTRACT

In the recent past, sequencing of ancient human genomes has become increasingly common, leading to an immense amount of data to be explored. For this study we focused on comparing a set of ancient individuals with modern populations on behalf of markers for celiac disease. We analyzed a panel of 64 SNPs related to this disease, trying to detect changes in allele frequencies between ancient and modern individuals. We hope to make a contribution to the subject of genetic health throughout human history.

KEYWORDS

Ancient DNA; Celiac Disease; Allele Frequencies; Risk Factor

INTRODUCTION

The sprout to pre-eminence

Cereals, particularly wheat, are a mainstay of traditional western diets. Modern wheat kernels consist to 70 to 75 percent of proteins, divided into glutenin and gliadin which can form several polymers summarized under the term “gluten” [1]. They all share a unique amino acid composition with a high content of glutamine and proline and only low contents of amino acids with charged side groups [2]. In the wheat plant, gliadin and glutenin serve primarily as storage for nutrients [3]. The gluten polymer serves as “glue” in dough, providing stability and elasticity. Therefore wheat is a popular ingredient in bread, cakes and other bakery goods, beer and many more of the foods an adult might consume throughout the day. The last 50 years have seen a rate of increase in wheat consumption that is higher than for any other cereal. During this period in the United Kingdom, an average of 50% of the total carbohydrates consumed by individuals derived from cereals [4]. But the roots of wheat as a major source for carbohydrates stretch back much farther, to about 12500 years (BP), and the beginning of the Neolithic period [5].

Historical roots

One of the oldest evidences of controlled planting of crops is dated around 21000 BP [6], but in general wheat is believed to have played only a minor role in human diets during these times [7]. Until then, a nomadic lifestyle prevailed, sustained by hunting and gathering. Unprocessed, the ancient wheat was only an inferior source of carbohydrates and large scale farming was impossible, due to limiting climatic factors [7, 8]. During the span around 12500 BP the climate changed and this allowed the establishment of larger permanent settlements and an increased exploitation of grasses. The first archaeological evidence for successful cultivation of grasses in larger scales was found in the northern area of the Arabian Peninsula [7], the so called “fertile crescent” [9], and marks the onset of the Neolithic period.

The "nececereal" evil

Numerous studies have linked wheat consumption to an increase in the incidence of celiac disease, which is caused by an autoimmune reaction to gluten and other wheat proteins. Symptoms range from diarrhoea, abdominal pain, impaired growth, iron deficiency, anaemia and a decrease in bone density [10–12]. The predisposition to celiac disease, its symptoms and strength depend on lifestyle, environmental, and genetic factors. A number of genetic markers have been associated with celiac disease and its symptoms [13–15]. The SNPs with the strongest association towards celiac disease are found within a region called the HLA-loci located on chromosome 6. Two of these, HLA-DQ8 and HLA-DQ2.5, are routinely used in clinical diagnoses of celiac disease [16]. These variants usually have a weak penetrance and are found in up to 40% of the unaffected population as well [17]. Yet inheriting these specific mutations substantially increases a person’s risk of developing celiac disease. Due to its incomplete penetrance, the gold standard for celiac disease diagnosis lies in tissue biopsies and IgA anti-tissue transglutaminase antibody tests [18].

The aim of the present study is to probe changes in allele frequencies in any of the SNPs associated with celiac disease. Changes and variations in the frequency of markers associated with celiac disease, during or after the domestication of plants, can potentially be linked to the increased wheat consumption and might have had an effect on the prevalence of the disease. To gather evidence on such potential allele frequency changes over time, we compared genomes of

ancient hunter-gatherers and farmers as well as modern individuals.

MATERIAL AND METHODS

Data

The ancient dataset was formed of 57 ancient humans, consisting of 15 Hunter-Gatherers and 42 Farmers. We used the data of ancient samples from different publications (Supplement S1), focusing on samples with available whole-genome-shotgun sequencing data.

All data for modern individuals was downloaded in VCF-format from the 1000 Genomes Project (1000 Genomes Project Consortium, 2010).

The selected SNPs outside the HLA-regions were chosen by a review of literature (Supplement S2).

Data processing

The usage of data aligned to different reference genomes might lead to unreliable results. Therefore, BAM files, that were not aligned against the human reference hs37d5 (hs37d5.fa.gz file from ftp://ftp.1000genomes.ebi.ac.uk//vol1/ftp/technical/reference/phase2_reference_assembly_sequence/), were realigned. In brief, the BAM-data was reverted to fastq-format.

Then the alignment program bwa (v0.7.15) [19] in the bwa aln mode was used with standard options to map the data against the reference. The resultant BAM files underwent final processing including an indel-realignment (GATK toolkit v3.6) [20–22] and a quality-filter step (samtools v1.3.1) [23]. After every processing step, we validated the resulting BAM-files for the SAM-format specification using picard tools (v2.8.0) (http://broadinstitute.github.io/picard).

SNP-calling for all BAM files was performed using the AntCaller (v1.1) [24] with standard options, resulting in VCF files for all of the ancient data. Since the AntCaller does not create VCF header lines automatically, we had to add the lines needed for further data processing. All SNP-calling methods used for ancient data have a tendency to overestimate heterozygosity [25]; therefore, the files were filtered for a read depth over 2 and a genotype quality value over 30, using (bcftools, v1.4).

Bioinformatics Methods

The analysis of the SNPs strongest correlated towards celiac disease inside the HLA-regions was conducted using the tool HLA-VBSeq [26]. We tested the functionality of the tool for ancient DNA successful, but due to time constraints, we focused our analysis on 64 SNPs outside the HLA-regions (Supplement S2).

All SNPs connected with celiac disease outside the HLA-regions were analysed within the provided VCF files. After merging these filtered VCF-files of the ancient data with the files for the modern reference individuals, calculations for allele counts and allele frequencies were conducted using the program vcftools (v0.1.13).

Statistics

In order to gauge the genetic risk for celiac disease per ancient individual, we implemented a simple risk score. The score for an ancient individual is calculated by first dividing all alleles present (number of existing genotypes x 2) by all alleles that can be possibly covered (number of SNPs x 2). Then this factor is taken times the sum for the risk alleles over all SNPs (n = 64), whereby a homozygote occurrence of a risk allele counts as 2 and a heterozygote occurrence counts as 1. This formula uses no log odds ratio for disease risk for SNPs and assumes independence between SNPs. We choose this method, as we did not have access to samples that had a confirmed diagnosis of celiac disease in our modern reference data set.

$$simple\ risk\ score = \frac{covered\ genotypes * 2}{128} * \sum 2 * homozygote\ for\ risk\ allele + 1 * heterozygote\ for\ risk\ allele$$

For a comparison of risk factors between ancient and modern individuals, we used a method similar to Berens et al. 2017. For every ancient genome, we subsampled all the individuals in the 1000 Genomes Project to match the coverage for that respective ancient genome. Then, the simple risk score was calculated for every subsampled modern genome. The percentile of the ancient risk score, relative to the modern individuals was generated via the ecdf function in R.

In order to create a comparable dataset of modern individuals, we bootstrapped the superpopulations of the 1000 Genomes Project. In brief, we randomly selected 20 individuals from the respective population pool, calculated allele frequency and repeated this procedure 1000 times. The mean allele frequency and the mean allele count of the complete modern reference, as well as for every subpopulation (Supplement S3), were used with the ancient population via Fisher's exact test [27].

To get an overview of the genetic distance between our ancient samples and a modern European reference we used LASER (v2.04) to create a primary component analysis (PCA). Since all ancient samples originate from Eurasia, we chose a reference dataset [28, 29], which provides a better resolution for Europe.

All of the plots in this study were created using R (v3.4.2).

To maximize time efficiency, we used GNU-parallel whenever possible [30].

RESULTS

From the 57 ancient individuals in our dataset, only 11 had at least one SNP covered. Out of the 64 selected loci outside of the HLA regions, our ancient samples contained 51 SNPs with a sufficient coverage and genotype quality (Supplement S2).

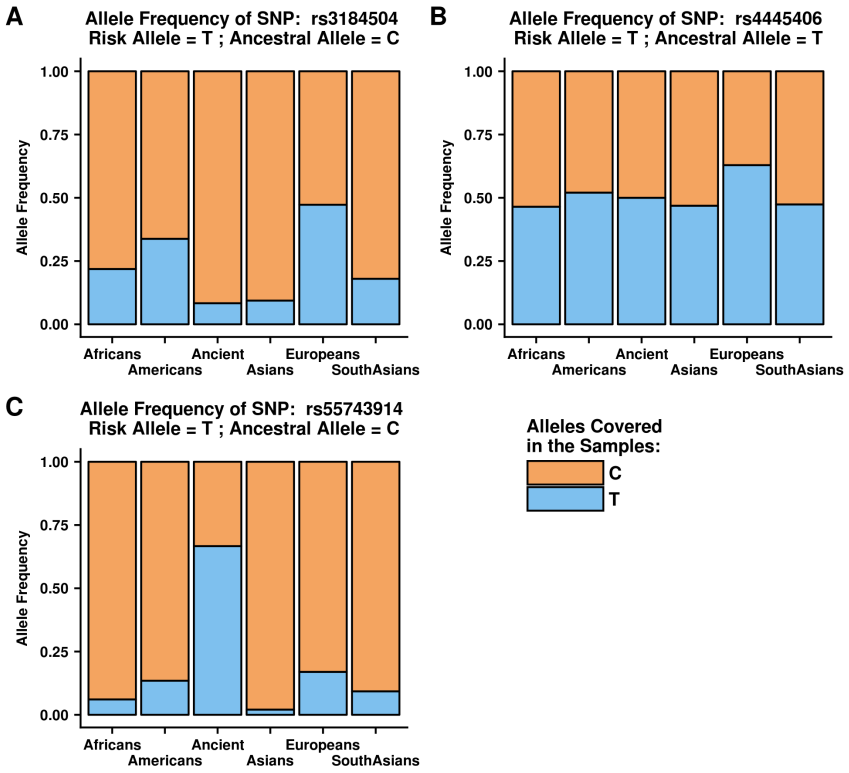

Figure 1: Allele frequencies of three selected SNPs associated with celiac disease. Risk variant in blue, non-risk variant in orange. Shown are 5 modern populations plus the ancient population.

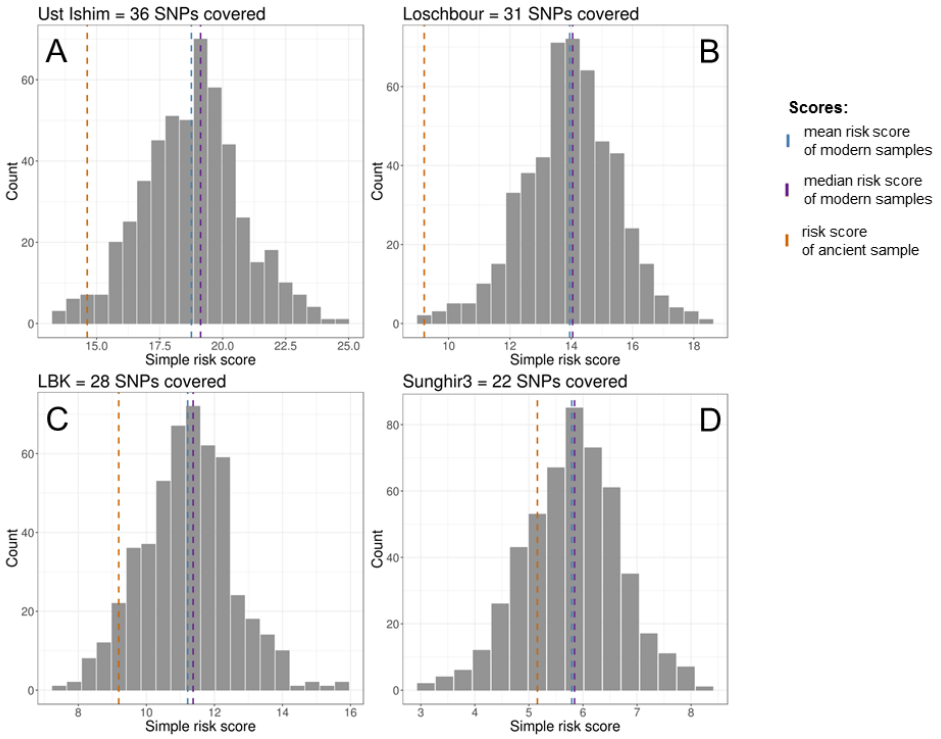

Figure 2: Subsampled risk scores for the respective ancient individual. The risk score for the ancient individual is given in orange, mean and median for the risk scores of the modern individuals are given in blue and violet. The overall distribution of the respective risk score for matched modern Europeans is represented by the grey histogram in the background. Ust Ishim: Ancient sample from Ust'-Ishim in Siberia , Loschbour: Ancient sample from Loschbour in Luxembourg, LBK (Linearbandkeramik): Ancient sample from Stuttgart in Germany, Sunghir3: Ancient sample from Sunghir in Russia.

A general overview over the genetic distance between our samples and a modern reference is provided by a PCA (Figure 1, Supplement S5). It allows a rough estimate of which modern populations are closest to our ancient samples while showing at the same time the distance in between the ancient dataset.

For an overview over our results, we selected 3 representative SNPs with sufficient coverage that showed different trends in allele frequency.

The SNP rs3184504, located in the gene SH2B3, shows a shift towards an increase in the frequency of the risk allele in the modern European populations (p-value = 0,01811) (Figure 1A & Supplement S3).

In contrast, the SNP rs4445406, an intron variant of the gene MMEL1, shows no significant fluctuations of its frequencies between the ancient and the modern European population (p-value = 0,46268) (Figure 1B). A decrease of the risk allele frequency is detectable between ancient and modern European individuals for the SNP rs55743914 (p-value = 0,04502) associated with the gene PTPRK (Figure 1C).

Inspired by [31], who used a de-novo assembly approach to identify specific immune loci, we tried a different method, using HLA-VBSeq. This tool is a genotyper specific for HLA-regions which are highly diverse and therefore not suitable for normal alignment. We took the raw fastq data of the sample from Satsurblia Cave (SATP), of which we extracted the unmapped reads, using an aligned BAM file as reference. We successfully applied this approach to this ancient sample and found it is suitable for ancient DNA in general. We were unable to detect HLA-loci associated with celiac disease.

Eleven ancient individuals exhibited a coverage at least 1 SNP on the selected 64 positions (Supplement S2). However, most individuals had a negligible coverage of at most 3 SNPs. Only 4 individuals exhibited a coverage of more than 20 SNPs. For all of the better covered individuals, we calculated a simple genetic risk score. This risk score reflects a presence/absence of risk alleles for celiac disease in the respective ancient individual. For each ancient sample, we additionally subsampled modern Europeans from the 1000 Genomes Project to the same SNP coverage and calculated a risk score distribution for matched modern healthy genomes in order to be able to compare them to our ancient samples (Figure 2 and Supplement S4).

In all 4 cases the risk score of the ancient individuals was below the mean of the risk scores of the modern healthy Europeans from the 1000 Genomes Project.

DISCUSSION

The results of our data analysis provide a good overview of a wide spectrum of changes in allele frequencies for 64 SNPs connected to celiac disease (Figure 1) and their combined risk score for representative individuals (Figure 2). We cannot make any significant claims towards a general selection at any of the observed sites, since our ancient dataset does not provide the necessary coverage for those kinds of analyses. Nevertheless, there are certain trends observable in several SNPs.

In the case of celiac disease, an increase of the frequency in favor of the risk allele might be explained with an evolutionary advantage surpassing the negative effects of the increased risk to suffer from the disease. In cultures which are mostly sustained through agriculture, an increase of the risk for celiac disease could mean a painful disadvantage. However, certain factors might still lead to an increase of the allele frequency for the risk allele. Aside from genetic drift, a random change in allele frequencies, the risk allele might be related to another phenotype beneficial for the organism.

The risk variant of the SNP rs3184504, for example, is thought to be connected with the immune system, and to possibly enhance the prevention of bacterial infection [32] due to its presence in the gene SH2B3. This hypothesis is supported by our data, showing an increased allele frequency for the risk allele in the modern populations, especially in the European population, compared to the ancient individuals (Figure 1A). These changes in this specific allele are generally associated with a selective sweep which occurred between 1200 and 1700 years ago, possibly triggered by an infectious disease.

Aside from a shift towards an increased risk allele frequency, there is the opposite possibility: Changes against the risk allele. A decreasing risk allele frequency can be caused, apart from the ever present genetic drift, by a negative influence on the health of the individual. An example for decreasing allele frequency presents itself in form of the SNP rs55743914, located in the gene body of PTPRK. Our data shows a significant difference between the ancient population and the modern ones, with a decrease in risk allele frequency on the modern side (Figure 1B). This SNP is associated with celiac disease [33], which could be a possible factor for the shift.

After crossing out increases and decreases in the risk allele frequency there is the third possibility: no change whatsoever in the allele frequency. No significant changes in the allele frequency, besides the normal fluctuations caused by the genetic drift, might indicate an absence of any advantages or disadvantages for the affected individual. An example from our data can be seen in the SNP rs4445406, associated with the gene MMEL1 (Figure 1C). There is no known connection towards phenotypes besides celiac disease, and the small fluctuations in allele frequency might signify no strong impact of the SNP on the prevalence of the disease.

In case of the overall simple risk factor for celiac disk, the individuals with coverage greater than 20 SNPs were generally below the majority of matched healthy genomes from the European population of the 1000 Genomes Project, with regard to their risk factor. Given the fact of lacking coverage, especially for the loci that are most associated to the disease, this is a trend at

best. Even if the coverage were sufficient, it is worth to mention that all identified SNPs that increase a risk for a specific disease, were found by comparisons of modern individuals and might not hold true for comparison between ancient and modern individuals. Additionally, a heightened genetic risk does not necessarily mean that an individual will develop the disease. To date it is unknown, as to whether "genetic health" has increased or decreased over the last few thousand years and several conflicting theories exist [34, 35]. A recent study performed by Berens et al., 2017 tested over 100 ancient genomes for the presence or absence of a broad panel of disease-related SNPs. They seemed to find a trend for a decrease in genetic risk over time, albeit not significant. All in all, the genetic health of ancient individuals is certainly an interesting topic for future research.

While we did find coverage on some HLA-loci for SATP, in processing the data with HLA-VBSeq, there was no correction of the C to T shift. There are several ways to correct this shift, but completely eliminating it would have exceeded the scope of this study. Furthermore there is the possibility of misalignment and bias within the results. The fragmented state of ancient DNA may lead to spurious mapping, exacerbated by the fact that the mapping process involves a database of all known HLA-variants. None the less, we consider this to be an appropriate approach for further studies.

There are several factors in our study, which might lead to biased results. First and foremost, our sample size is very limited, which might lead to wrong signals, due to uncommon variations. Another problem presents itself in the form of demography. All our results could be influenced by founder effects and genetic bottlenecks. Additionally, it must be considered that the individuals from the ancient dataset are very widespread in terms of time, geography and ancestry. This, together with the incomplete penetrance of the disease and additional environmental factors, make it difficult to claim results as unbiased.

The next steps in research would be a repeated analysis with a bigger dataset, together with data from patients afflicted with celiac disease. Improved statistical methods and population analysis, like the calculation of FST-Values, might shed new light upon this research subject. Another important step is the inclusion of demography, which might be another influence on the distribution of the SNPs.

ACKNOWLEDGEMENTS

We would like to thank Jun. Prof. Dr. Susanne Gerber, Dr. David Fournier and Russ Hodge for giving us the chance to write this publication.

Additionally we want to express our gratitude towards our supervisors especially the Palaeogenetics Group, especially Prof. Dr. Joachim Burger, Jens Blöcher and Dr. Christian Sell for providing us with all the so much needed support.

This work was partly funded by the center of computational sciences (CSM).

Parts of this research were conducted using the supercomputer Mogon and/or advisory services offered by Johannes Gutenberg University Mainz (hpc.uni-mainz.de), which is a member of the AHRP and the Gauss Alliance e.V. The authors acknowledge the computing time granted on the supercomputer Mogon at Johannes Gutenberg University Mainz (hpc.uni-mainz.de).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

SUPPLEMENTARY DATA

High resolution figure files and supplementary files are available at Genomics and Computational Biology online.

REFERENCES

1. Kasarda DD. **Can an increase in celiac disease be attributed to an increase in the gluten content of wheat as a consequence of wheat breeding?** Journal of agricultural and food chemistry. 2013;61(6):1155–1159. doi:10.1021/jf305122s.
2. Wieser H. **Chemistry of gluten proteins**. Food Microbiology. 2007;24(2):115–119. 3rd International Symposium on Sourdough. doi:10.1016/j.fm.2006.07.004.
3. Shewry PR, Halford NG. **Cereal seed storage proteins: structures, properties and role in grain utilization**. Journal of Experimental Botany. 2002;53(370):947–958. doi:10.1093/jexbot/53.370.947.
4. Kearney J. **Food consumption trends and drivers**. Philosophical transactions of the Royal Society of London Series B, Biological sciences. 2010;365(1554):2793–2807. doi:10.1098/rstb.2010.0149.
5. Purugganan MD, Fuller DQ. **The nature of selection during plant domestication**. Nature. 2009;457(7231):843–848. doi:10.1038/nature07895.
6. Kislev ME, Nadel D, Carmi I. **Epipalaeolithic (19,000 BP) cereal and fruit diet at Ohalo II, Sea of Galilee, Israel**. Review of Palaeobotany and Palynology. 1992;73(1-4):161–166. doi:10.1016/0034-6667(92)90054-K.
7. Cordain L. **Cereal Grains: Humanitys Double-Edged Sword**. In: Simopoulos AP, editor. Evolutionary Aspects of Nutrition and Health. vol. 84 of World Review of Nutrition and Dietetics. Basel: KARGER; 1999. p. 19–73. doi:10.1159/000059677.
8. Feynman J, Ruzmaikin A. **Climate stability and the development of agricultural societies**. Climatic Change. 2007;84(3-4):295–311. doi:10.1007/s10584-007-9248-1.
9. Clay AT. **The So-Called Fertile Crescent and Desert Bay**. Journal of the American Oriental Society. 1924;44:186. doi:10.2307/593554.
10. Green PHR. **The many faces of celiac disease: Clinical presentation of celiac disease in the adult population**. Gastroenterology. 2005;128(4):S74–S78. doi:10.1053/j.gastro.2005.02.016.
11. Fasano A. **Clinical presentation of celiac disease in the pediatric population**. Gastroenterology. 2005;128(4):S68–S73. doi:10.1053/j.gastro.2005.02.015.
12. Lionetti E, Catassi C. **New clues in celiac disease epidemiology, pathogenesis, clinical manifestations, and treatment**. International reviews of immunology. 2011;30(4):219–231. doi:10.3109/08830185.2011.602443.
13. Hunt KA, Zhernakova A, Turner G, Heap GAR, Franke L, Bruinenberg M, et al. **Newly identified genetic risk variants for celiac disease related to the immune response**. Nature genetics. 2008;40(4):395–402. doi:10.1038/ng.102.

14. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, et al. **Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease**. Nature genetics. 2011;43(12):1193–1201. doi:10.1038/ng.998.
15. Dubois PCA, Trynka G, Franke L, Hunt KA, Romanos J, Curtotti A, et al. **Multiple common variants for celiac disease influencing immune gene expression**. Nature genetics. 2010;42(4):295–302. doi:10.1038/ng.543.
16. Husby S, Koletzko S, Korponay-Szabó IR, Mearin ML, Phillips A, Shamir R, et al. **European Society for Pediatric Gastroenterology, Hepatology, and Nutrition guidelines for the diagnosis of coeliac disease**. Journal of pediatric gastroenterology and nutrition. 2012;54(1):136–160. doi:10.1097/MPG.0b013e31821a23d0.
17. Hadithi M. **Accuracy of Serologic Tests and HLA-DQ Typing for Diagnosing Celiac Disease**. Annals of Internal Medicine. 2007;147(5):294. doi:10.7326/0003-4819-147-5-200709040-00003.
18. Rubio-Tapia A, Hill ID, Kelly CP, Calderwood AH, Murray JA. **ACG clinical guidelines: Diagnosis and management of celiac disease**. The American journal of gastroenterology. 2013;108(5):656–76; quiz 677. doi:10.1038/ajg.2013.79.
19. Li H, Durbin R. **Fast and accurate short read alignment with Burrows-Wheeler transform**. Bioinformatics (Oxford, England). 2009;25(14):1754–1760. doi:10.1093/bioinformatics/btp324.
20. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. **A framework for variation discovery and genotyping using next-generation DNA sequencing data**. Nature genetics. 2011;43(5):491–498. doi:10.1038/ng.806.
21. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. Genome research. 2010;20(9):1297–1303. doi:10.1101/gr.107524.110.
22. Van der Auwera, Geraldine A, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. **From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline**. Current protocols in bioinformatics. 2013;43:11.10.1–33. doi:10.1002/0471250953.bi1110s43.
23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. **The Sequence Alignment/Map format and SAMtools**. Bioinformatics (Oxford, England). 2009;25(16):2078–2079. doi:10.1093/bioinformatics/btp352.
24. Zhou B, Wen S, Wang L, Jin L, Li H, Zhang H. **AntCaller: An accurate variant caller incorporating ancient DNA damage**. Molecular genetics and genomics : MGG. 2017;292(6):1419–1430. doi:10.1007/s00438-017-1358-5.
25. Link V, Kousathanas A, Veeramah K, Sell C, Scheu A, Wegmann D. **ATLAS: Analysis Tools for Low-depth and Ancient Samples**; 2017. doi:10.1101/105346.
26. Nariai N, Kojima K, Saito S, Mimori T, Sato Y, Kawai Y, et al. **HLA-VBSeq: Accurate HLA typing at full resolution from whole-genome sequencing data**. BMC genomics. 2015;16 Suppl 2:S7. doi:10.1186/1471-2164-16-S2-S7.
27. Fisher RA. **On the Interpretation of chi 2 from Contingency Tables, and the Calculation of P**. Journal of the Royal Statistical Society. 1922;85(1):87. doi:10.2307/2340521.
28. Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, et al. **A genetic atlas of human admixture history**. Science. 2014;343(6172):747–751. doi:10.1126/science.1243518.
29. Busby GBJ, Hellenthal G, Montinaro F, Tofanelli S, Bulayeva K, Rudan I, et al. **The Role of Recent Admixture in Forming the Contemporary West Eurasian Genomic Landscape**. Current biology. 2015;25(19):2518–2526. doi:10.1016/j.cub.2015.08.007.
30. Ole Tange. **GNU Parallel - The Command-Line Power Tool**. ;login: The USENIX Magazine. February 2011;p. 42–47.
31. Olalde I, Allentoft ME, Sánchez-Quinto F, Santpere G, Chiang CWK, DeGiorgio M, et al. **Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European**. Nature. 2014;507(7491):225–228. doi:10.1038/nature12960.
32. Zhernakova A, Elbers CC, Ferwerda B, Romanos J, Trynka G, Dubois PC, et al. **Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection**. American journal of human genetics. 2010;86(6):970–977. doi:10.1016/j.ajhg.2010.05.004.
33. Bondar C, Plaza-Izurieta L, Fernandez-Jimenez N, Irastorza I, Withoff S, Wijmenga C, et al. **THEMIS and PTPRK in celiac intestinal mucosa: Coexpression in disease and after in vitro gliadin challenge**. European journal of human genetics : EJHG. 2014;22(3):358–362. doi:10.1038/ejhg.2013.136.
34. Lynch M. **Mutation and Human Exceptionalism: Our Future Genetic Load**. Genetics. 2016;202(3):869–875. doi:10.1534/genetics.115.180471.
35. Roth FP, Wakeley J. **Taking Exception to Human Eugenics**. Genetics. 2016;204(2):821–823. doi:10.1534/genetics.116.192096.

Supplementary file:

100044_Hewel_File_S1.pdf

Ancient data used in analyses.

Population	Samplename	Reference	Context/Culture	Region
Hunter-gatherers				
Ust_Ishim	Ustb	(Fu et al. 2014)	Upper Palaeolithic	Siberia
Sunghir	S1,S2,S3,S4,S5	(Sikora et al. 2017)	Upper Palaeolithic	Russia
Latvia	Latvia_HG1, Latvia_HG2, Latvia_HG3	(Jones et al. 2017)	Mesolithic	Latvia
Satsurblia	SATP	(Jones et al. 2015)	Upper Palaeolithic	Georgia
Ukraine	Ukraine_HG1	(Jones et al. 2017)	Mesolithic	Ukraine
Schela Cladovei	Schela Cladovei/SC1-Meso	(González-Fortes et al. 2017)	Mesolithic	Romania
Loschbour_HG	Loschbour	(Lazaridis et al. 2014)	Mesolithic	Luxembourg
Karelia_HG	Karelia_HG	(Fu et al. 2016)	Mesolithic	Russia
Scandinavia_HG	Ajvide58	(Skoglund et al. 2014)	Mesolithic and Neolithic hunter-gatherer/ Pitted Ware Culture	Sweden
Neolithic				
Stuttgart	Stuttgart	(Lazaridis et al. 2014)	Neolithic	Germany
Sweden_MN	GC6khem2	(Skoglund et al. 2014)	Funnelbeaker (TRB)	Sweden
Bar	Bar31	(Hofmanova et al. 2016)	early Neolithic	BarcD1n/Turkey
Rev	Rev5	(Hofmanova et al. 2016)	early Neolithic	Revenia/Greece
Bon	Bon001, Bon002, Bon004, Bon005, Tep001, Tep002, Tep003, Tep004, Tep006	(Kılınç et al. 2016)	early Neolithic(Farmer)	Boncuklu/Turkey
Remedello_CA	RISE489	(Allentoft et al. 2015)	Remedello	Italy
Afanasievo_BA	RISE509	(Allentoft et al. 2015)	Afanasievo	Russia
Yamnaya_BA	RISE547, RISE548, RISE550	(Allentoft et al. 2015)	Yamnaya	Russia
Corded_Ware_BA	RISE00	(Allentoft et al. 2015)	Corded Ware and Battle Axe	Germany/ Sweden/Estonia
Bell_Beaker_BA	RISE569	(Allentoft et al. 2015)	Bell Beaker	Germany/Czech Republic
Okunevo_BA	RISE516	(Allentoft et al. 2015)	Okunevo	Russia
Unetice_BA	RISE150, RISE577	(Allentoft et al. 2015)	Unetice	Germany/Poland/Czech Republic
Sintashta_BA	RISE392, RISE394, RISE395	(Allentoft et al. 2015)	Sintashta	Russia
Andronovo_BA	RISE500, RISE503	(Allentoft et al. 2015)	Andronovo	Russia
Karasuk_BA	RISE495, RISE496, RISE499, RISE502	(Allentoft et al. 2015)	Karasuk	Russia
Mezhovskaya_BA	RISE523	(Allentoft et al. 2015)	Mezhovskaya	Russia
Armenia_BA	RISE423	(Allentoft et al. 2015)	Middle Bronze Age	Armenia
Hungary_BA	RISE479	(Gamba et al. 2014); (Allentoft et al. 2015)	Bronze Age	Hungary
Pal	Pal7	(Hofmanova et al. 2016)	late Neolithic	Paliambela/Greece
Klei	Klei10	(Hofmanova et al. 2016)	late Neolithic	Kleitos/Greece
Iron Age				
Scandinavia_IA	RISE174	(Allentoft et al. 2015)	Iron Age	Sweden
Altai_IA	RISE600, RISE601, RISE602	(Allentoft et al. 2015)	Iron Age	Russia
Russia_IA	RISE504	(Allentoft et al. 2015)	Iron Age	Russia

Allentoft, Morten E.; Sikora, Martin; Sjögren, Karl-Göran; Rasmussen, Simon; Rasmussen, Morten; Stenderup, Jesper et al. (2015): Population genomics of Bronze Age Eurasia. In: *Nature* 522 (7555), S. 167–172. DOI: 10.1038/nature14507.

Fu, Qiaomei; Li, Heng; Moorjani, Priya; Jay, Flora; Slepchenko, Sergey M.; Bondarev, Aleksei A. et al. (2014): Genome sequence of a 45,000-year-old modern human from western Siberia. In: *Nature* 514 (7523), S. 445–449. DOI: 10.1038/nature13810.

Fu, Qiaomei; Posth, Cosimo; Hajdinjak, Mateja; Petr, Martin; Mallick, Swapan; Fernandes, Daniel et al. (2016): The genetic history of Ice Age Europe. In: *Nature* 534 (7606), S. 200–205. DOI: 10.1038/nature17993.

Gamba, Cristina; Jones, Eppie R.; Teasdale, Matthew D.; McLaughlin, Russell L.; Gonzalez-Fortes, Gloria; Mattiangeli, Valeria et al. (2014): Genome flux and stasis in a five millennium transect of European prehistory. In: *Nature communications* 5, S. 5257. DOI: 10.1038/ncomms6257.

González-Fortes, Gloria; Jones, Eppie R.; Lightfoot, Emma; Bonsall, Clive; Lazar, Catalin; Grandal-d'Anglade, Aurora et al. (2017): Paleogenomic Evidence for Multi-generational Mixing between Neolithic Farmers and Mesolithic Hunter-Gatherers in the Lower Danube Basin. In: *Current biology : CB* 27 (12), 1801-1810.e10. DOI: 10.1016/j.cub.2017.05.023.

Hofmanova, Zuzana; Kreutzer, Susanne; Hellenthal, Garrett; Sell, Christian; Diekmann, Yoan; Diez-Del-Molino, David et al. (2016): Early farmers from across Europe directly descended from Neolithic Aegeans. In: *Proceedings of the National Academy of Sciences of the United States of America* 113 (25), S. 6886–6891. DOI: 10.1073/pnas.1523951113.

Jones, Eppie R.; Gonzalez-Fortes, Gloria; Connell, Sarah; Siska, Veronika; Eriksson, Anders; Martiniano, Rui et al. (2015): Upper Palaeolithic genomes reveal deep roots of modern Eurasians. In: *Nature communications* 6, S. 8912. DOI: 10.1038/ncomms9912.

Kılınç, Gülşah Merve; Omrak, Ayça; Özer, Füsun; Günther, Torsten; Büyükkarakaya, Ali Metin; Bıçakçı, Erhan et al. (2016): The Demographic Development of the First Farmers in Anatolia. In: *Current biology : CB* 26 (19), S. 2659–2666. DOI: 10.1016/j.cub.2016.07.057.

Lazaridis, Iosif; Patterson, Nick; Mittnik, Alissa; Renaud, Gabriel; Mallick, Swapan; Kirsanow, Karola et al. (2014): Ancient human genomes suggest three ancestral populations for present-day Europeans. In: *Nature* 513 (7518), S. 409–413. DOI: 10.1038/nature13673.

Raghavan, Maanasa; Skoglund, Pontus; Graf, Kelly E.; Metspalu, Mait; Albrechtsen, Anders; Moltke, Ida et al. (2014): Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. In: *Nature* 505 (7481), S. 87–91. DOI: 10.1038/nature12736.

Sikora, Martin; Seguin-Orlando, Andaine; Sousa, Vitor C.; Albrechtsen, Anders; Korneliussen, Thorfinn; Ko, Amy et al. (2017): Ancient genomes show social and reproductive behavior of early Upper Paleolithic foragers. In: *Science (New York, N.Y.)*. *DOI:* 10.1126/science.aao1807.

Skoglund, Pontus; Malmström, Helena; Omrak, Ayça; Raghavan, Maanasa; Valdiosera, Cristina; Günther, Torsten et al. (2014): Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers. In: *Science (New York, N.Y.)* 344 (6185), S. 747–750. DOI: 10.1126/science.1253448.

Supplementary file:

100044_Hewel_File_S2.pdf

rs-ID	Chromosome	Position	Ancestral Allele	Risk Allele	Associated Genes/Location
rs917997	2	103070568	C	A	IL18RAP
rs17810546	3	159665050	A	G	IL12A-AS1
rs12142280	1	172864652	T	T	Intergenic region between FASLG and TNFSF18
rs6441961	3	46352384	C	A	CCR3
rs2305764	19	17313833	A	T	MYO9B
rs13119723	4	123218313	A	A	IL21
rs6822844	4	123509421	G	C	IL21
rs9851967	3	188087628	C	T	LPP
rs3184504	12	111884608	C	T	SH2B3
rs4445406	1	2539400	t	T	C1orf39, MMEL1, TTC34
rs72657048	1	25289734	C	G	At the first exon of RUNX3
rs12068671	1	172681031	C	T	35-43 kb 5‘ of FALSG
rs859637	1	172711000	C	T	FALSG, TNFSF18, TNFSF4
rs72734930	1	192512559	G	T	32 kb 5‘ of RGS1
rs1359062	1	192541472	C	G	0-24 kb 5‘ at the first exon of RGS1
rs10800746	1	200881392	C	C	9th Intron of C1orf106
rs13003464	2	61186829	G	G	Exons 5-11 of PUS10
rs10167650	2	68645560	t	T	Intergenic region between PLEK and FBX048
rs990171	2	103086770	C	A	IL18R1, IL18RAP
rs1018326	2	182007800	T	C	Intergenic region between UBE2E3 and ITGA4
rs6715106	2	191913034	A	A	Exons 6-14 of STAT4
rs12998748	2	191948637	G	G	Intron 3 of STAT4
rs6752770	2	191973563	A	G	STAT4
rs10207814	2	204459961	C	T	111 – 121 kb 5’ *CD28*
rs1980422	2	204610396	T	C	intergenic between *CD28* and *CTLA4*
rs34037980	2	204770054	A	A	intergenic between *CTLA4* and *ICOS*

rs4678523	3	33037721	T	C	intergenic between *CCR4* and *GLB1*
rs7616215	3	46205686	C	C	LOC105377067
rs2097282	3	46378025	C	C	intergenic between *CCR3* and *CCR2*
rs61579022	3	119123278	G	A	intron 10 ARHGAP31
rs1353248	3	159623559	T	C	intergenic between *SCHIP1* and *IL12A*
rs76830965	3	159637678	C	A	intergenic
rs2561288	3	159674928	C	T	intergenic between *SCHIP1* and *IL12A*
rs2030519	3	188119901	A	A	intron 2 *LPP*
rs62323881	4	123038295	C	A	*KIAA1109, ADAD1, IL2, IL21*
rs13132308	4	123551114	A	A	*KIAA1109, ADAD1, IL2, IL21*
rs12203592	6	396321	C	C	IRF4
rs1050976	6	408079	C	C	3′ UTR *IRF4*
rs7753008	6	90809639	T	C	intron 2 *BACH2*
rs55743914	6	128293562	C	T	*PTPRK* last exon, 3′UTR
rs72975916	6	128294055	C	C	*PTPRK* exons 28-30, 3′UTR, to 24kb 3′
rs77027760	6	138002061	G	G	intergenic
rs17264332	6	138005515	A	G	intergenic between *OLIG3* and *TNFAIP3*
rs182429	6	159469574	A	A	4kb 5′ and 5′ UTR *TAGAP*
rs1107943	6	159498267	T	C	32kb 5′ *TAGAP*
rs79758729	7	37418454	G	G	ELMO1
rs10808568	8	129264060	A	A	151 - 163kb 3′ of *PVT1*
rs2387397	10	6390192	C	C	intergenic between *PFKFB3* and *PRKCQ*
rs1250552	10	81058027	g	A	ZMIZ1

rs7104791	11	111196858	C	T	*POU2AF1* , *C11orf93*
rs10892258	11	118579865	G	G	intergenic between *TREH* and *DDX6*
rs61907765	11	128391937	C	T	5kb 5′ & 1st exon *ETS1*
rs11851414	14	69259502	T	C	1kb 5′ & 1st exon *ZFP36L1*
rs1378938	15	75096443	C	A	*CLK3, CSK* and multiple genes
rs6498114	16	10964118	T	G	*CIITA*
rs243323	16	11361202	A	A	11kb 5′, all of SOCS1, 1kb 3′
rs80073729	16	11373797	G	G	Intergenic
rs9673543	16	11384956	A	G	10kb 5′ PRM1
rs11875687	18	12843137	T	C	exons 2-5 *PTPN2*
rs62097857	18	12857758	G	A	PTPN2
rs1893592	21	43855067	A	A	UBASH3A
rs58911644	21	45629121	A	A	18 - 25kb 3′ *ICOSLG*
rs4821124	22	21979289	T	C	*UBE2L3, YDJC*
rs13397	X	153248248	G	A	*HCFC1, TMEM187, IRAK1*
rs653178	12	112007756	T	G	ATXN2
rs1050152	5	131676320	C	T	LOC553103, SLC22A4

Supplementary file:

100044_Hewel_File_S3.pdf

Ancient vs. all modern Populations

rs-ID	Ancient allele count	p-value
rs10800746	8	$1{,}272*e^{-7}$
rs12142280	8	3,419*e-^{15}
rs1250552	6	0.00329
rs1359062	6	$7{,}669*e^{-12}$
rs2030519	8	0.0047
rs3184504	12	$3{,}049*e^{-7}$
rs4445406	10	0.00129
rs55743914	6	0.00306
rs5891644	6	$2{,}333*e^{-7}$
rs7104791	10	$1{,}665*e^{-5}$
rs859637	10	$7{,}224*e^{-11}$
rs9851967	6	0.0763

Ancient versus Africans

rs-ID	Ancient allele count	p-Value
rs10800746	8	0.00758
rs12142280	8	0.19093
rs1250552	6	0.10659
rs1359062	6	0.00037
rs2030519	8	0.67948
rs3184504	12	1
rs4445406	10	0.10094
rs55743914	6	0.01645
rs58911644	6	0.00228
rs7104791	10	0.40523
rs859637	10	0.1966
rs9851967	6	0.08914

Ancient versus Americans

rs-ID	Ancient allele count	p-Value
rs10800746	8	0.66555
rs12142280	8	0.03068
rs1250552	6	1
rs1359062	6	0.31132
rs2030519	8	0.2505
rs3184504	12	0.42113
rs4445406	10	1
rs55743914	6	0.03262
rs58911644	6	0.5973
rs7104791	10	0.39691
rs859637	10	0.28591
rs9851967	6	0.65515

Ancient versus Asians

rs-ID	Ancient allele count	p-Value
rs10800746	8	1
rs12142280	8	0.00036
rs1250552	6	0.37519
rs1359062	6	0.5713
rs2030519	8	0.2505
rs3184504	12	0.23077
rs4445406	10	1
rs55743914	6	0.00044
rs58911644	6	0.23734
rs7104791	10	0.15482
rs859637	10	0.00087
rs9851967	6	0.34966

Ancient versus Europeans

rs-ID	Ancient allele count	p-Value
rs10800746	8	1
rs12142280	8	0.06812
rs1250552	6	1
rs1359062	6	0.5713
rs2030519	8	0.70077
rs3184504	12	0.01811
rs4445406	10	0.46268
rs55743914	6	0.04502
rs58911644	6	0.56973
rs7104791	10	0.67059
rs859637	10	1
rs9851967	6	1

Ancient versus South Asians

Rs-ID	Ancient allele count	P-Value
rs10800746	8	0.26029
rs12142280	8	0.18677
rs1250552	6	0.67961
rs1359062	6	1
rs2030519	8	0.13174
rs3184504	12	1
rs4445406	10	1
rs55743914	6	0.00946
rs58911644	6	0.56973
rs7104791	10	0.30792
rs859637	10	1
rs9851967	6	0.34966

Supplementary file:

100044_Hewel_File_S4.pdf

sample	Risk alleles covered	risk score of ancient individual	mean	median	percentile
Ust_Ishim	36	14.625	18.75932	19.125	0.03180915
Loschbour	31	9.203125	13.95154	14.04688	0.003976143
LBK	28	9.1875	11.21409	11.375	0.08946322
Sunghir3	22	5.15625	5.793178	'5.84375	0.2902584
RISE150	3	0.09375	0.1130405	0.09375	0.5447316
Sunghir5	2	0.0625	0.04218439	0.03125	0.9025845
Sunghir4	2	0.0625	0.09350149	0.09375	0.2584493
RISE174	2	0.0625	0.06610338	0.0625	0.6481113
Sunghir2	1	0.015625	0.01428926	0.015625	0.7932406
RISE495	1	0	0	0	1
Bon002	1	0.015625	0.01208375	0.015625	0.8429423

rs-ID	Chromosome	Position	Ancestral Allele	Risk Allele	Associated Genes/Location
rs917997	2	103070568	C	A	IL18RAP
rs17810546	3	159665050	A	G	IL12A-AS1
rs12142280	1	172864652	T	T	Intergenic region between FASLG and TNFSF18
rs6441961	3	46352384	C	A	CCR3
rs2305764	19	17313833	A	T	MYO9B
rs13119723	4	123218313	A	A	IL21
rs6822844	4	123509421	G	C	IL21
rs9851967	3	188087628	C	T	LPP
rs3184504	12	111884608	C	T	SH2B3
rs4445406	1	2539400	t	T	C1orf39, MMEL1, TTC34
rs72657048	1	25289734	C	G	At the first exon of RUNX3
rs12068671	1	172681031	C	T	35-43 kb 5‘ of FALSG
rs859637	1	172711000	C	T	FALSG, TNFSF18, TNFSF4
rs72734930	1	192512559	G	T	32 kb 5‘ of RGS1
rs1359062	1	192541472	C	G	0-24 kb 5‘ at the first exon of RGS1
rs10800746	1	200881392	C	C	9th Intron of C1orf106
rs13003464	2	61186829	G	G	Exons 5-11 of PUS10
rs10167650	2	68645560	t	T	Intergenic region between PLEK and FBX048
rs990171	2	103086770	C	A	IL18R1, IL18RAP
rs1018326	2	182007800	T	C	Intergenic region between UBE2E3 and ITGA4
rs6715106	2	191913034	A	A	Exons 6-14 of STAT4
rs12998748	2	191948637	G	G	Intron 3 of STAT4
rs6752770	2	191973563	A	G	STAT4
rs10207814	2	204459961	C	T	111 – 121 kb 5′ *CD28*
rs1980422	2	204610396	T	C	intergenic between *CD28* and *CTLA4*
rs34037980	2	204770054	A	A	intergenic between *CTLA4* and *ICOS*
rs4678523	3	33037721	T	C	intergenic between *CCR4* and *GLB1*

rs7616215	3	46205686	C	C	LOC105377067
rs2097282	3	46378025	C	C	intergenic between *CCR3* and *CCR2*
rs61579022	3	119123278	G	A	intron 10 ARHGAP31
rs1353248	3	159623559	T	C	intergenic between *SCHIP1* and *IL12A*
rs76830965	3	159637678	C	A	intergenic
rs2561288	3	159674928	C	T	intergenic between *SCHIP1* and *IL12A*
rs2030519	3	188119901	A	A	intron 2 *LPP*
rs62323881	4	123038295	C	A	*KIAA1109, ADAD1, IL2, IL21*
rs13132308	4	123551114	A	A	*KIAA1109, ADAD1, IL2, IL21*
rs12203592	6	396321	C	C	IRF4
rs1050976	6	408079	C	C	3′ UTR *IRF4*
rs7753008	6	90809639	T	C	intron 2 *BACH2*
rs55743914	6	128293562	C	T	*PTPRK* last exon, 3′UTR
rs72975916	6	128294055	C	C	*PTPRK* exons 28-30, 3′UTR, to 24kb 3′
rs77027760	6	138002061	G	G	intergenic
rs17264332	6	138005515	A	G	intergenic between *OLIG3* and *TNFAIP3*
rs182429	6	159469574	A	A	4kb 5′ and 5′ UTR *TAGAP*
rs1107943	6	159498267	T	C	32kb 5′ *TAGAP*
rs79758729	7	37418454	G	G	ELMO1
rs10808568	8	129264060	A	A	151 - 163kb 3′ of *PVT1*
rs2387397	10	6390192	C	C	intergenic between *PFKFB3* and *PRKCQ*
rs1250552	10	81058027	g	A	ZMIZ1
rs7104791	11	111196858	C	T	*POU2AF1* , *C11orf93*
rs10892258	11	118579865	G	G	intergenic between *TREH* and *DDX6*

rs61907765	11	128391937	C	T	5kb 5′ & 1st exon *ETS1*
rs11851414	14	69259502	T	C	1kb 5′ & 1st exon *ZFP36L1*
rs1378938	15	75096443	C	A	*CLK3, CSK* and multiple genes
rs6498114	16	10964118	T	G	*CIITA*
rs243323	16	11361202	A	A	11kb 5′, all of SOCS1, 1kb 3′
rs80073729	16	11373797	G	G	Intergenic
rs9673543	16	11384956	A	G	10kb 5′ PRM1
rs11875687	18	12843137	T	C	exons 2-5 *PTPN2*
rs62097857	18	12857758	G	A	PTPN2
rs1893592	21	43855067	A	A	UBASH3A
rs58911644	21	45629121	A	A	18 - 25kb 3′ *ICOSLG*
rs4821124	22	21979289	T	C	*UBE2L3, YDJC*
rs13397	X	153248248	G	A	*HCFC1, TMEM187, IRAK1*
rs653178	12	112007756	T	G	ATXN2
rs1050152	5	131676320	C	T	LOC553103, SLC22A4

Supplementary file:

100044_Hewel_File_S5.pdf

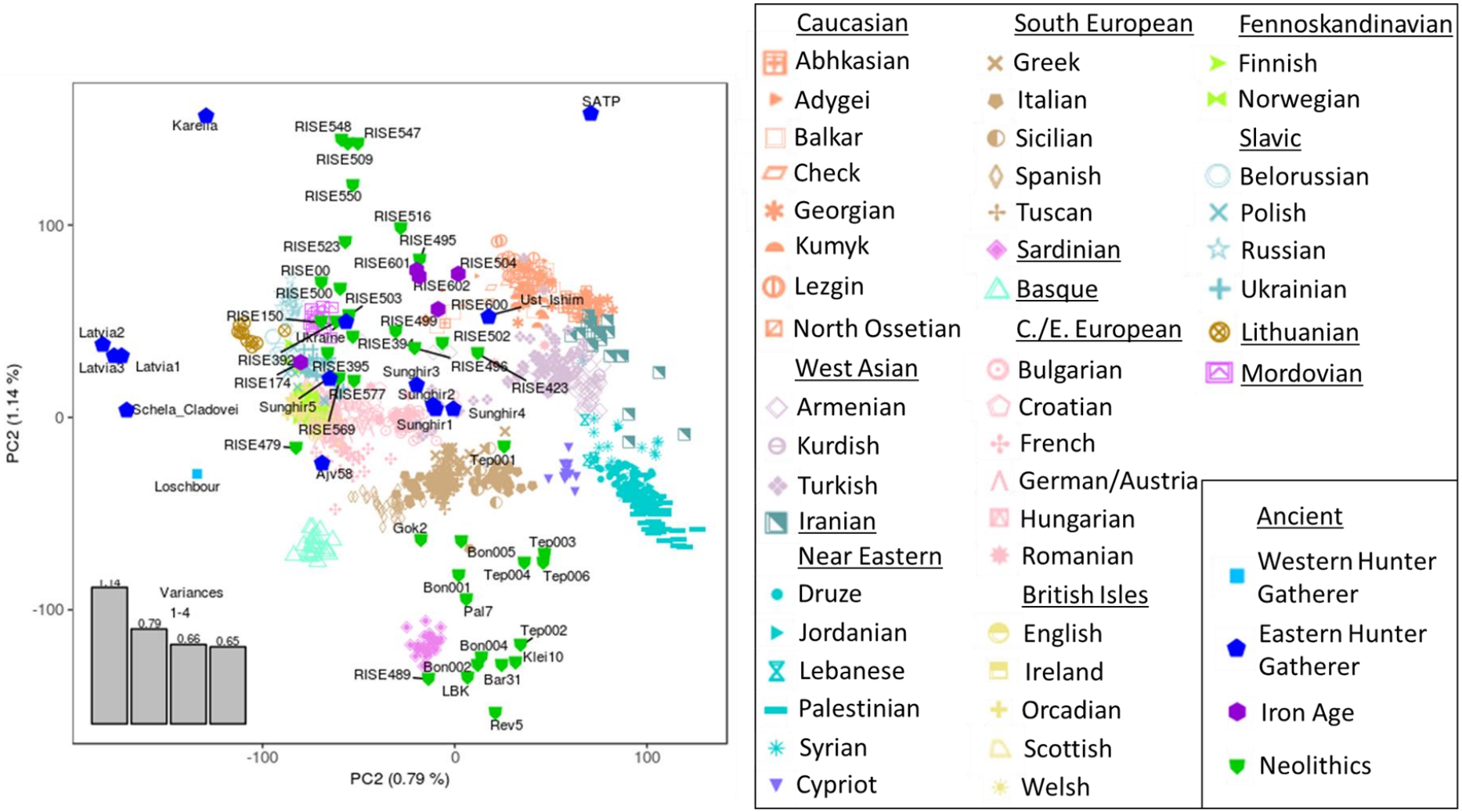

Fig. 1 Supplement S5: PCA of the ancient individuals used in this study vs. a modern European ancestry reference
The unlabeled symbols refer to the modern European ancestry reference individuals. The labeled symbols refer to all ancient individuals used in this study. Details about the origin of the ancient samples are given in Supplement S1. The bar plot in grey depicts the % of variance that is captured by the first 4 principal components.