Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers

Julia Höglund*, Nima Rafati, Mathias Rask-Andersen, Stefan Enroth, Torgny Karlsson, Veronica E. Ek & Åsa Johansson

Genome-wide association studies (GWAS) have identified associations between thousands of common genetic variants and human traits. However, common variants usually explain a limited fraction of the heritability of a trait. A powerful resource for identifying trait-associated variants is whole genome sequencing (WGS) data in cohorts comprised of families or individuals from a limited geographical area. To evaluate the power of WGS compared to imputations, we performed GWAS on WGS data for 72 inflammatory biomarkers, in a kinship-structured cohort. When using WGS data, we identified 18 novel associations that were not detected when analyzing the same biomarkers with genotyped or imputed SNPs. Five of the novel top variants were low frequency variants with a minor allele frequency (MAF) of <5%. Our results suggest that, even when applying a GWAS approach, we gain power and precision using WGS data, presumably due to more accurate determination of genotypes. The lack of a comparable dataset for replication of our results is a limitation in our study. However, this further highlights that there is a need for more genetic epidemiological studies based on WGS data.

Over the past decade, genome-wide association studies (GWAS) have successfully identified associations of thousands of single-nucleotide polymorphisms (SNPs) with human traits and diseases. Most of the associated alleles discovered so far are common, with a minor allele frequency (MAF) above 5%. Many SNPs are also located outside coding regions, which complicates the identification of causal mechanisms, functional variants and relevant genes. Additionally, identified SNPs collectively only explain a limited fraction of the heritability. A number of hypotheses for this “hidden heritability” have been proposed, such that part of the heritability is due to rare variants, that there is a non-negligible fraction of unmapped or untagged common variants, or that variants with very low effect sizes have not been captured in current GWAS. The extent to which rare and low-frequency coding variants (<5%) influence traits and diseases is still not completely understood. Rare variants may not be present on currently available SNP arrays nor be well tagged by the available SNPs on the array due to the low linkage disequilibrium (LD) between common SNPs and rare variants. Performing GWAS with imputed or genotyped variants is therefore not ideal for detecting associations with rare variants. Additionally, rare variants are often specific to individual populations, or even families, making them hard to detect with standard GWAS in unrelated participants.

However, the limitations in association studies can be partly reduced. Firstly, a powerful approach to identify complex trait- and disease-associated rare variants is to use populations that comprise of families or individuals from a limited geographical area. Secondly, whole genome sequencing (WGS) data can be used to better capture rare and low frequency variants and variants not in LD with SNPs on a genotyping array. WGS is superior to imputation when it comes to determining genotypes of rare variants with high accuracy. Simulation studies have shown that the mapping precision for rare variants increases considerably when using WGS data in a GWAS approach, making it an efficient way of detecting and fine-mapping rare variants simultaneously. Hence, by shifting from genotyped and imputed data to WGS data, a standard GWAS can be performed with a likely increase in both variant capture and precision. Yet, few GWAS have been performed using WGS data to date. During the last years, WGS has been performed in a variety of different populations. A recent study within...
a small kinship-structured cohort (similar to ours), tested for the burden of rare variants from WGS data on six cardiometabolic traits. The authors found novel signals that neither were captured with low-depth sequencing, nor with genome-wide genotyping with dense imputation in the same samples, supporting the notion that WGS data in kinship-structured cohorts can improve power to identify genetic associations. A number of studies have also performed GWAS in cohorts where the variants were imputed from unique reference panels based on WGS of a subset of the participants of the same cohorts, or performed GWAS on low coverage (∼4×) WGS data. For example, in a study on circulating lipid levels and five inflammatory biomarkers, WGS of 2,120 Sardinians was performed to assess the impact of the variants common in the Sardinian cohort but rare in the 1000 Genomes Project. In total, 14 signals were found, including two new loci that would have been missed if data had been imputed using the 1000 Genomes reference panel, further underlining the advantages of large-scale sequencing.

Biomarkers are often strongly genetically regulated and have been shown to be less polygenic in comparison to complex traits and diseases, which increases the power to study the effect in a smaller cohort where WGS data is available. When used for diagnosis, an ideal biomarker should be uniquely present or overexpressed in the tissue of interest and not be influenced by confounding factors, such as genetic variants. However, genetic factors commonly have a considerable effect on biomarker levels and introduce noise when biomarkers are used for diagnosis. Better characterization of the genetic contribution to variation in biomarker levels is therefore of great importance.

In this project, we used a GWAS approach to test for associations with no fewer than 72 inflammatory plasma protein biomarkers, in order to investigate the gain in precision and rare variant-capture with WGS data compared to genotyped/imputed SNPs. In total, 1,005 individuals with high coverage WGS data from the kinship-structured population-based Northern Swedish population health study (NSPHS) were included. This cohort has also been genotyped and imputed, making this a valuable opportunity to compare the relative performance of WGS and genotyping/imputation in relation to the same phenotype measurements. This study is one of few that uses WGS data with a GWAS approach in order to capture a greater number of low frequency variants associated with inflammatory protein biomarkers, and to further characterize the genetic structure underlying these associations, aiming to extend our knowledge of the genetic contribution to these biomarkers.

Results
A total of 1,005 individuals with WGS data and biomarker data were included in this study. The age of the participants ranged from 14 to 94 years with a median of 52 years, and 50.8% of the participants were females. The biomarkers were measured at two timepoints. At the first timepoint, biomarkers from Olink’s Oncology I and Cardiovascular I panels were measured of which 31 are inflammatory (from now on called ONC_CVD). At the second timepoint, 95 inflammatory biomarkers from the Inflammatory I panel were measured (from now on called INF). After quality control (QC), 72 unique biomarkers remained of which 42 were only from the INF panel, while one had been measured only on ONC_CVD. As many as 29 biomarkers were included on both the INF and ONC_CVD panels and they were considered technical replicates. The average number of individual measurements per biomarker was 915 (median 929, range: 430–957) in ONC_CVD and 829 (median 871, range: 424–892) in INF.

Genome-wide association for biomarker levels. In the WGS data, a total of 16,890,549 biallelic single-nucleotide variants (SNVs) were called. A MAF threshold of 0.15% was chosen in order to reach enough statistical power in the GWAS (Supplementary Fig. S1). After filtering on MAF and Hardy-Weinberg equilibrium (HWE), 12,210,410 SNVs remained for downstream analyses. For the 72 individual biomarkers analyzed, 5,812 SNV-biomarker associations were identified, and for 41 (56.9%) of the biomarkers, variants had p-values just below the genome-wide threshold (ranging from 1.17 × 10⁻⁸ to 3.43 × 10⁻¹⁰) in ONC_CVD and p-values just above the genome-wide threshold in INF (Supplementary Table S2). Here, the larger sample size of individual measurements per biomarker was 915 (median 929, range: 430–957) in ONC_CVD and 829 (median 871, range: 424–892) in INF.

Most biomarkers (67.44%) with at least one significant hit identified, had an association in cis (i.e., within 1 Mb of the gene encoding the biomarker) or even within the gene encoding the biomarker itself. The rest of the associations were in trans, all located on another chromosome than the gene encoding the biomarker (Fig. 2). Adjusting for the most significant SNV resulted in 15 biomarkers having a secondary, significant signal close to the primary signal, and adjusting for both the primary and secondary SNV resulted in seven biomarkers having a tertiary signal (Table 2, with more extensive variant data in Supplementary Tables S3 and S4). In the conditional analyses, only the SNVs that were located within each associated region (see methods) were analyzed. Due to the reduced number of variants analyzed in the conditional analyses, as compared to the primary GWAS, the power to analyze rarer variants increased and we therefore did not have a MAF threshold in the conditional analyses. Here, we then identified four variants with MAF < 0.15%.

Comparison with our previous GWAS using genotyped/imputed data suggests novel loci for many biomarkers. Twenty of the biomarkers (ADA, CASP-8, CCL11, CCL20, CCL23, CD244, CDCP1,
CST5, CX3CL1, CXCL1, CXCL11, CXCL9, FGF-5, MCP-3, ST1A1, STAMBP, TGFB1, TNFB, TNFSF14, uPA) with significant associations in the present study, did not have any significant associations in our previous GWAS when using genotyped/imputed SNP data$^{25,28}$ (Supplementary Figs S2–S20). The abundance of two of these biomarkers (CXCL9 and CXCL11) had an associated variant that in our previous studies was identified to be associated only with CXCL10 and is most likely a false positive finding for CXCL9 and CXCL11 (discussed more thoroughly in Supplementary, including Supplementary Figs S21–S25). The remaining novel biomarker associations represented 18 unique loci that were not found to be associated with the levels of the same biomarker using genotyped/imputed data in the same cohort. Of these, 15 loci (see overlap with GWAS catalog below) have not been reported in any previous study of the same biomarkers, thus making them novel loci. In the novel loci, six top variants are considered to be low-frequency variants (MAF $< 5\%$). Additional to the novel loci, four biomarkers (CD6, CXCL5, CCL4, MMP-10) had associations driven by top variants that are only in moderate LD ($R^2 < 0.8$) with the top variants from our previous studies, and might therefore be considered independent associations (Supplementary Figs S26–S30). Another 19 loci overlapped between the present study and our previous studies with SNP data$^{25,28}$, for which nine loci had the same top variant. The remaining ten overlapping loci had different top variants, although these variants were in high LD ($R^2 > 0.8$). The top variants in the overlapping loci were more strongly associated (more significant p-value) in the present study than in our previous GWAS, except for two biomarkers (MMP-10 and TRAIL), for which more significant GWAS top variants were found in the previous studies (Tables 3 and 4).

**Replication in an independent cohort.** Due to the lack of a similar dataset (WGS data and measured levels of the same inflammatory biomarkers) for replication, we could only test for replication for a subset of our results using GWAS results of circulating cytokines in a Finnish population$^{29}$ (Supplementary Table S6). Of the cytokines that were analyzed in both studies, we fully replicated our primary results for CCL11, i.e., the same top SNV was found in both cohorts. One of the two independent associations with CCL4 was also fully replicated. We also replicated one of the three independent associations with MCP-1, and one of the two independent associations with CXCL1, although with an LD between the top variants of $R^2 = 0.75$ and 0.72 for MCP-1 and CXCL1, respectively. For our second independent association with CCL4, our top variant was either monomorphic in the Finnish population or had not been analyzed. However, the most significant CCL4 associated SNV in the Finnish population is also genome-wide significant in our study ($P = 3.52 \times 10^{-12}$), even if not our most significant. Our results for MCP-3, SCF and CXCL9 did not replicate in the Finnish population ($P > 0.05$ in the Finnish cohort for our top SNVs) and our result for TNFB was only nominally significant ($P = 0.027$). On the other hand, the most significant SNVs for IL-7, IL-10, IL-18 and HGF in the Finnish population, were not genome-wide significant in our cohort, even if rs5745687 (for HGF) as well as rs385076, rs17229943, and rs71478720 (for IL-18) were nominally significant ($P < 0.05$).
Table 1. Location and annotation of significant top GWAS hits from WGS data. The raw p-values (not adjusted for multiple testing) are shown. If one biomarker had been measured twice (i.e. been measured on both INF and ONC_CVD), the SNV with the most significant p-value is presented. Novel variants are shown in bold. Additional information can be found in Supplementary Table S2. §Does not have an rs-id, †In hg19 coordinates and ONC_CVD), the SNV with the most significant p-value is presented. Novel variants are shown in bold.

Colocalization with eQTL data in blood. Colocalization (the same top variant) with cis-eQTLs in peripheral blood was found for five of the top SNVs (Table 5), associated with the levels of three different biomarkers (CD40, CXCL5 and IL-15RA). For the other top variants from our biomarker GWAS, LD was calculated between each top variant and the most significant eQTL. Two top variants, associated with CXCL5, were colocalized...
(R² > 0.8) with a variant associated with a trans-eQTL in blood. Two secondary hits associated with IL-18R1 and TNFSF14 respectively, and one tertiary hit associated with CCL23, were also found to be colocalized with a cis-eQTL. All overlapping variants had the same direction of effect except for the overlap with the trans-eQTL, where two SNVs (rs10740118 and rs7088799) in JMJD1C was associated with increased protein levels of CXCL5, but was in LD (R² = 0.86) with a variant, rs10761779 that was associated with decreased RNA levels of CXCL5.

Colocalization with data from the GWAS catalog. The association signal for eight biomarkers (CCL19, CCL4, CD40, CD6, CXCL5, IL-12B, IL-18R1, TNFSF14) colocalized with association signals for one or several inflammatory diseases (Table 6, Supplementary Table S7). Here, we regard the signals to be colocalized when a top SNV or any SNVs in LD (R² > 0.8) with a top SNV identified in our study was also the top SNV for an inflammatory disease in the GWAS catalog (v 1.0.2). In four cases, our top SNV had been associated with an inflammatory disease in previous GWAS: rs113010081/CCL4 with inflammatory bowel disease, rs1569723/CD40 with Crohn's disease, rs4239702/CD40 with rheumatoid arthritis and rs11230563/CD6 with ulcerative colitis. The remaining top variants are in LD (R² > 0.8) with previously inflammatory disease-associated top variants. For example, the top variants for CXCL5 found in this study (rs352045 in ONC_CVD and rs425535 in INF) are both in high LD (R² = 0.944) with a variant previously associated with ulcerative colitis. Some of our top SNVs were also colocalized with associations for different blood-trait (Supplementary Table S8).

Discussion
We performed a GWAS on 72 inflammatory biomarkers in a Swedish cohort using WGS data, and identified SNVs that were associated with the plasma levels for as many as 41 biomarkers. Of the biomarkers with at least one significant hit, 67.44% had an association within 1 Mb of the gene encoding the biomarker (in cis) and the rest (32.56%) had an association on another chromosome (in trans). Many of the biomarker levels are highly

Figure 2. Circular representation of the GWAS hits. The numbers in the outer circle correspond to the chromosomes. Each biomarker is labelled at the position of the gene coding it on the cytoband. The colored lines/arrows represent the significant hits. The breadth of the line represents the size of the region associated with respective biomarker.
Table 2. Location and annotation of top GWAS hits after having conditioned on the most significant hit. The raw p-values (not adjusted for multiple testing) are shown. The adjusted p-values are based on the number of SNVs tested in each region which means that each SNV does not need to reach genome wide significance. If one biomarker had been measured twice (i.e. been measured on both INF and ONC_CVD), the SNV with the most significant p-value is presented. Additional information can be found in Supplementary Table S4. *Does not have an rs-id. †In hg19 coordinates. ‡Variant is from ONC_CVD. Either the p-value was lower, or no significant association was found in INF. ††In cis: within 1 Mb of the gene encoding the biomarker; in trans: on another chromosome of the gene encoding the biomarker.

---

| Biomarker | SNV | Conditional Signal | P-value | P adj. | Effect, beta (SE) | Effect allele (ref) | MAF (effect allele) | chr:position† | Gene | Type | Location** |
|-----------|-----|-------------------|---------|--------|------------------|------------------|-------------------|----------------|-------|------|------------|
| CCL23     | rs72831705 | secondary | 9.52 × 10^-11 | 1.07 × 10^-05 | 0.44 (0.07) | T (C) | 0.153 | 17.34321277 | CCL15-CCL14 | ncRNA...intronic | cis |
| CCL23     | rs854671 | tertiary | 1.48 × 10^-08 | 1.67 × 10^-04 | 0.23 (0.05) | C (T) | 0.475 | 17.34361300 | CCL23, CCL18 | intronic | cis |
| CCL4     | rs35199851 | secondary | 3.57 × 10^-07 | 2.10 × 10^-03 | 3.09 (0.61) | A (C) | 0.00984 | 3.51599851 | RAD54L2 | intronic | cis |
| CCL4     | rs188700215 | secondary | 1.01 × 10^-20 | 5.91 × 10^-09 | −5.19 (0.97) | A (G) | 0.00984 | 17.30092085 | MIR4365.COPR5 | intronic | cis |
| CCL4     | rs201079256 | secondary | 1.02 × 10^-13 | 5.97 × 10^-09 | −3.06 (0.05) | T (C) | 0.465 | 17.34522125 | CCL3L1, CCL3L3 | downstream | cis |
| CD40     | rs6063068 | secondary | 7.41 × 10^-08 | 6.27 × 10^-04 | 3.28 (0.61) | T (A) | 0.00984 | 20.45717496 | EYA2 | intronic | cis |
| CD40     | rs182282247 | tertiary | 9.99 × 10^-08 | 8.45 × 10^-04 | 0.82 (0.15) | A (G) | 0.021 | 20.44730041 | NCOA3; CD40 | intronic | cis |
| CST-5     | rs6138152 | secondary | 3.87 × 10^-07 | 3.13 × 10^-04 | 0.36 (0.07) | G (A) | 0.211 | 20.23850130 | CST2;CST5 | intronic | cis |
| CST-5     | rs75823487 | tertiary | 4.56 × 10^-07 | 3.69 × 10^-04 | 4.25 (0.84) | T (C) | 0.00155 | 20.29478349 | MIR663A;H. LINC01957 | intronic | cis |
| CXL1     | rs19038101 | secondary | 7.53 × 10^-07 | 6.54 × 10^-04 | −0.24 (0.05) | T (G) | 0.461 | 4.74668772 | CXL8;CXL6 | intronic | cis |
| CXL6     | rs181216093 | tertiary | 5.27 × 10^-07 | 1.86 × 10^-04 | 0.41 (0.08) | G (A) | 0.102 | 10.5700416 | ASB13 | intronic | cis |
| IL-15RA   | rs141473272 | secondary | 1.49 × 10^-11 | 1.86 × 10^-06 | −1.70 (0.25) | T (C) | 0.013 | 10.6008255 | IL15RA | missense | cis |
| IL-15RA   | rs35095871 | tertiary | 3.06 × 10^-07 | 3.81 × 10^-04 | 0.41 (0.08) | G (A) | 0.102 | 10.5700416 | ASB13 | intronic | cis |
| IL-18R1   | rs12999517 | secondary | 4.89 × 10^-10 | 1.31 × 10^-06 | −0.47 (0.05) | C (T) | 0.172 | 2.10299526 | IL18R1 | intronic | cis |
| MCP-2     | rs78432823 | secondary | 6.47 × 10^-10 | 7.79 × 10^-07 | −1.17 (0.10) | G (A) | 0.045 | 17.32335173 | LINC01989;CCL2 | intronic | cis |
| MCP-2     | rs12601658 | tertiary | 7.15 × 10^-12 | 8.61 × 10^-07 | −0.32 (0.05) | A (T) | 0.244 | 17.32533423 | LINC01989; CCL2 | intronic | cis |
| MMP-1     | rs4703588 | secondary | 9.58 × 10^-09 | 1.56 × 10^-05 | 0.29 (0.05) | T (C) | 0.397 | 11.102668702 | WATPP1 | ncRNA...intronic | cis |
| SCP       | rs8104417 | secondary | 2.66 × 10^-07 | 2.53 × 10^-04 | 0.23 (0.05) | C (T) | 0.4995 | 20.44632542 | ZNF355;MMP9 | intronic | cis |
| STI1A     | rs4149583 | secondary | 5.38 × 10^-07 | 4.56 × 10^-04 | 0.54 (0.11) | A (G) | 0.061 | 16.2862020 | SULT1A1 | UTR5 | cis |
| TNFB      | rs748688 | tertiary | 1.42 × 10^-02 | 2.18 × 10^-02 | 0.26 (0.05) | C (G) | 0.061 | 6.31540429 | TNFB | intronic | cis |
| TNFB      | rs62190519 | tertiary | 4.06 × 10^-08 | 6.22 × 10^-04 | 1.77 (0.32) | T (G) | 0.005 | 6.27190519 | MIR416; PDGFR1 | intronic | cis |
| TNFSF14   | rs2291668 | secondary | 4.71 × 10^-07 | 7.36 × 10^-04 | 0.27 (0.05) | A (G) | 0.281 | 19.6669934 | TNFSF14 | synonymous | cis |
| TRAIL     | rs812026109 | secondary | 1.01 × 10^-20 | 5.19 × 10^-09 | −5.19 (0.97) | A (G) | 0.00049 | 18.21026109 | TMEM241;RIK43 | intronic | cis |
both U.S and European populations and are in complete LD in our cohort. The minor allele for rs425535 was with the top variant (rs10190555) from our study (R² = 0.95) and the SNP most strongly associated with expression was rs11123923. However, rs11123923 is only in weak LD IL18R1 in CD4 and CD8 cells was associated with increased risk for IBD, that study, reduced transcript levels of CXCL5 protein levels. These lower levels might give a strong association signal and it is not until the top missense variant is that the missense variant affects the antibody affinity for the biomarker which in turn will give lower measured is correlated with rs1077667 that has previously been associated with multiple sclerosis. A possible explanation this variant was found to be associated with lower TNFSF14 levels. The missense variant is neither reported in the TNFSF14 (rs344560) in the primary signal is a missense variant located in the gene TNFSF14, which encodes the biomarker. This variant was found to be associated with lower TNFSF14 levels. The missense variant is neither reported in the GWAS catalog, nor is it in LD with any previously reported variants. However, the conditional SNP, rs2291668, is correlated with rs1077667 that has previously been associated with multiple sclerosis. A possible explanation is that the missense variant affects the antibody affinity for the biomarker which in turn will give lower measured protein levels. These lower levels might give a strong association signal and it is not until the top missense variant is adjusted for, that the true regulatory variant is detected.

A subset of our top SNVs overlapped with eQTLs in blood. For example, the correlated variants rs352045 and rs425535 (top SNVs for CXCL5 in the INF and ONC_CVD respectively) are in high LD with an eQTL for CXCL5 RNA levels in blood. Two variants, rs425535: a nonsynonymous exonic variant that lies within a splicing enhancer site, and rs352046: a promoter variant that is located within a transcription factor binding site for myeloid zinc finger proteins, have also previously been associated with CXCL5 mRNA expression and CXCL5 protein levels in blood. Both variants have been shown to be in almost complete LD (R² = 0.94) to complete LD (R² = 1) in both U.S and European populations and are in complete LD in our cohort. The minor allele for rs425535 was previously shown to be associated with significantly higher CXCL5 plasma concentrations and the minor allele for

Figure 3. Narrow-sense heritability estimates of the top variants. The total heritability estimate is shown in dark grey. The contribution of the top variant is shown in pink, the contribution of the first conditional top variant (secondary hit) in yellow and the second conditional (tertiary) hit in green. Light grey depicts biomarkers with no significant GWAS signal.
Table 3. Top GWAS hits with WGS data in comparison to the significant genotyped/imputed associations identified by Ahsan et al. Results from ONC_CVD are compared. The p-values for the top variants from the present study are shown (P present) as well as the p-values for the same variant in the imputed data (P Ahsan for WGS top variant). The most significant SNV from the previous study and corresponding p-value is also shown, and its LD (R²) with the most significant SNV from the present study. The comparisons have been filtered the same way as the previous study: only biallelic variants and variants not located in a spawning deletion are compared. †In hg19 coordinates. ‡Not significant in Enroth et al. **Top variant. P present variations. "Imputation quality of WGS top variant" for CXCL5 was calculated from the WGS data. ns. **Top variant calculated from WGS data. (P > 4.2e-09). 3R² with WGS top variant. **Top variant calculated from WGS data. (P > 4.2e-09). "Did not pass imputation QC or were not present in the reference panel used for the imputations.

Table 4. Top GWAS hits from WGS data in comparison to the significant genotyped/imputed associations identified by Enroth et al. Results from INF are compared. The p-values for the top variants from the present study are shown (P present) as well as the p-values for the same variant in the imputed data (P Enroth for WGS top variant). The most significant SNV from the previous study and corresponding p-value is also shown, and its LD (R²) with the most significant SNV from the present study. The comparisons have been filtered the same way as the previous study: only biallelic variants and variants not located in a spawning deletion are compared. †In hg19 coordinates. ‡Not significant in Enroth et al. (P > 4.2e-09). 3R² with WGS top variant. **Top variant calculated from WGS data. (P > 4.2e-09). "Did not pass imputation QC or were not present in the reference panel used for the imputations.

rs352046 with higher CXCL5 expression levels, which agrees with our results where the minor alleles of rs352045 and rs425535 were associated with higher CXCL5 levels (Supplementary Table S6). In our study, rs352045 was not identified as the top variant but instead rs352045 which is in almost complete LD with rs352046 (R² = 0.97). Our top SNV is located only 137 bp from rs352046 and both are found within transcription factor binding sites. Neither the association with rs352045 nor rs352046 are reported in the GWAS catalog. However, both rs352045
and rs425535 are in high LD (R² = 0.94) with a variant, rs2457996, that has previously been associated with ulcerative colitis. Ulcerative colitis is a sub type of IBD and CXCL5 have previously been shown to play a role in IBDS, such as ulcerative colitis and Crohn’s disease. In a study by Z’Graggen et al., a preferential expression of CXCL5 mRNA in the epithelium of the intestinal tissue from patients with IBD was observed. They also found a strong expression of CXCL5 at protein level. CXCL5, which encodes an epithelial cell-derived neutrophil activating peptide (also called ENA-78), has previously been suggested as a possible candidate gene for inflammatory diseases. Since the previously mentioned variants, rs425535 and rs352046, have previously been shown to play a role in diseases and complex traits. This study found several new loci associated with inflammatory biomarkers and indeed appears promising. We compared our results to earlier studies using genotyped/imputed data as well as WGS data, we do see an increase in both power and precision despite our limited sample size, an increase which make a robust assessment, and further studies in larger cohorts, or meta-analyses needs to be performed. Another limitation is the lack of reproducibility, given the nature of the study. It is a small kinship-structured cohort, which makes the results not generalizable to more mixed population or a population of another ancestry. While this population structure increases power to detect rare variants that might be more common in an isolated population, it also makes it harder to reproduce in another cohort. As of now, there are also only a limited number of cohorts that have been measuring the levels of the same inflammatory biomarkers and that have WGS data available. Even within our own data, we fail to replicate some results, both in the technical replicates as well as in the novel associations with regards to the results from genotyped/imputed data. The limitations mentioned above are most likely also dependent on differences in biomarker quantifications between the biomarker panels. As with imputation, variant calling can be more or less precise. Only a genotype probability is given, with additional quality measures. Caution should be taken, until the associations have been validated. This applies especially to the associations containing only a few variants, or variants not yet given an official identifier. Validation has to be performed in a similar cohort, in order to obtain higher confidence and better understanding of the results. This limitation was especially apparent in our validation in the Finnish population, with a possible reason for lack of replication being the discrepancy in population size (around 950 in NSPHS compared to up to 8,293 Finns). Other possible explanations for the lack of replication are the different LD structures, the different techniques used for protein quantification, and that one cohort being based on WGS and the other on genotyped and imputed data.

In summary, we have performed GWAS in a family-based cohort with WGS data in relation to inflammatory biomarkers. By analyzing only sequencing data, we seek to further extend our knowledge on the genetic contribution to these important biomarkers. The cost-efficient solution of sequencing a few individuals and creating a reference panel to be able to do dense genotyping is becoming a well-established method in genetic studies. The use of low-depth sequencing as a way of increasing power is also more common today. By using high coverage WGS data, we do see an increase in both power and precision despite our limited sample size, an increase which indeed appears promising. We compared our results to earlier studies using genotyped/imputed data as well as to previously published WGS. This study found several new loci associated with inflammatory biomarkers and nearly 50% of the associations were only detected in the present study. The associations were also stronger, with lower p-values, compared to those identified with genotyped/imputed data, suggesting that genotypes are in general more accurately determined using WGS compared to imputed data. Our results demonstrate the need of deep coverage WGS data with deeper coverage to be able to fully understand the genetic structure of common diseases and complex traits.

| Gene name | Biomarker | Top SNV (biomarker GWAS) | Annotation | Top SNV (eQTL) | LD | P (biomarker GWAS) | P (eQTL) |
|-----------|-----------|-------------------------|------------|---------------|----|------------------|---------|
| CD40      | CD40 (INF) | rs1569723                | intergenic (NCOA5:CD40) | rs1569723 | 1 | 5.24 × 10⁻⁴⁵ | 1.06 × 10⁻⁵⁸ |
| CD40      | CD40 (ONC-CVD) | rs4239702              | intronic (CD40)      | rs4239702 | 1 | 1.01 × 10⁻⁴⁹ | 1.26 × 10⁻⁴⁹ |
| CXCL5    | CXCL5 (INF) | rs352045 (cis)          | upstream (CXCL5)     | rs352045 | 1 | 6.09 × 10⁻²⁷ | 4.25 × 10⁻²⁷ |
| CXCL5    | CXCL5 (ONC-CVD) | rs245535 (cis)         | exonic (CXCL5)       | rs245535 | 1 | 1.90 × 10⁻³⁴ | 4.50 × 10⁻³⁴ |
| IL15RA   | IL15RA     | rs3136630                | intronic (IL15RA)    | rs3136630 | 1 | 2.64 × 10⁻¹⁹ | 5.21 × 10⁻¹⁹ |
| CXCL5    | CXCL5 (INF) | rs10740118 (trans)**     | intronic (IMIDVC)    | rs10761779 | 0.856 | 2.93 × 10⁻¹⁵ | 1.82 × 10⁻¹⁷ |
| CXCL5    | CXCL5 (ONC-CVD) | rs708799 (trans)**       | intronic (IMIDVC)    | rs10761779 | 0.856 | 7.36 × 10⁻¹⁶ | 1.82 × 10⁻¹⁷ |
| IL18R1   | IL18R1     | rs12999517               | intronic (IL18R1)    | rs12999517 | 1 | 4.89 × 10⁻¹⁸ | 1.13 × 10⁻¹⁹ |
| TNFSF14  | TNFSF14    | rs2291668b               | synonymous (TNFSF14) | rs1077667 | 0.899 | 4.71 × 10⁻¹⁷ | 4.36 × 10⁻¹⁸ |
| CCL23    | CCL23      | rs854671*                | intergenic (CCL23:CCL18) | rs854671 | 1 | 1.48 × 10⁻¹⁸ | 3.21 × 10⁻¹⁰ |

Table 5. Overlapping top SNVs from our biomarker GWAS with WGS data and top SNVs from the eQTL analyses by Westra et al. Linkage disequilibrium (R²) is presented for biomarkers that had different top SNVs in ONC-CVD and INF but are both in LD with a top eQTL. From the conditional analysis, adjusted for the top variant. From the second conditional analysis, adjusted for the top primary and secondary variant. "Top variant in cis, within 1 Mb of the gene encoding the biomarker. "Top variant in trans, on another chromosome than the gene encoding the biomarker."
| Biomarker | Disease trait | Mapped trait | SNV (biomarker GWAS) | Annotation (biomarker GWAS) | Associated SNP (GWAS catalog) | LD |
|----------|---------------|--------------|---------------------|----------------------------|-------------------------------|-----|
| CCL19    | Asthma, Juvenile idiopathic arthritis, Rheumatoid arthritis | Asthma; systemic, polyarticular, rheumatoid factor negative, oligoarticular juvenile idiopathic arthritis; Rheumatoid arthritis | rs149941420 | intronic (HLA-DRB1) | rs7775228 | 0.849 |
| CCL4     | Inflammatory bowel disease, Juvenile arthritis, Ulcerative colitis | Inflammatory bowel disease; systemic, polyarticular, rheumatoid factor negative, oligoarticular juvenile idiopathic arthritis; Ulcerative colitis | rs113010081 | intergenic (CCL2, CCL4) | rs11306487 | 1 |
| CD40     | Chronic hepatitis B infection, Chronic inflammatory diseases, Crohn's disease, Inflammatory bowel disease, Kawasaki disease, Multiple sclerosis, Rheumatoid arthritis, Systemic lupus erythematosus | Chronic hepatitis B infection; Ankylosing spondylitis; Psoriasis; Ulcerative colitis; Sclerosing cholangitis; Crohn's disease; Inflammatory bowel disease, mucocutaneous lymph node syndrome; Multiple sclerosis; Rheumatoid arthritis; Systemic lupus erythematosus | rs1569723, rs2439702 | intergenic (CCL2, CD40) | rs1569723, rs2439702 | 0.914, 0.914, 0.914, 0.843, 0.896, 0.914 |
| CD6      | Chronic inflammatory diseases, Crohn's disease, Inflammatory bowel disease, Ulcerative colitis | Ankylosing spondylitis; Psoriasis; Ulcerative colitis; Sclerosing cholangitis; Crohn's disease; Inflammatory bowel disease | rs11230563 | intronic (CD6) | rs11230563 | 1 |
| CXCL5    | Ulcerative colitis | Ulcerative colitis | rs352045, rs42535 | upstream (CXCL5), exonic (CXCL5) | rs2457996 | 0.944 |
| IL-12B   | Ankylosing spondylitis, Chronic inflammatory diseases, Crohn's disease | Ankylosing spondylitis; Psoriasis; Ulcerative colitis; Sclerosing cholangitis; Crohn's disease | rs10043720 | ncRNA intronic (LOC285626) | rs6556416, rs6556415 | 0.993, 0.884 |
| IL-18R1  | Celiac disease, Crohn's disease, Inflammatory bowel disease, Pediatric autoimmune diseases | Celiac disease; Crohn's disease; Inflammatory bowel disease; Autoimmune thyroid disease; Type 1 diabetes mellitus; Common variable immunodeficiency; Chronic childhood arthritis; Ankylosing spondylitis; Psoriasis; Ulcerative colitis; Autoimmune disease; Systemic lupus erythematosus | rs10190555 | intronic (IL18R1) | rs10190555 | 0.991, 0.954, 0.954, 0.954, 0.954 |
| TNFSF14  | Multiple sclerosis | Multiple sclerosis | rs22926156 | synonymous (TNFSF14) | rs1077667 | 0.879 |

**Table 6.** Disease-associations for the inflammatory biomarkers. Disease-associations with inflammatory diseases in the GWAS catalog are presented. If the variant has been reported in the catalog before, it is marked in bold. The other have not been previously reported, but are in strong LD with variants that have (R² > 0.8). A more extensive Table is found in Supplementary material, Supplementary Table S6. R² between our SNV and the previously associated variants from the GWAS catalog.

**Methods**

**Study cohort.** The NSPHS was initiated in 2006 to provide a health survey of the population in the parish of Karesuando county of Norrbotten, Sweden, to study the medical consequences of lifestyle and genetics. Additional participants were recruited in a second phase from the neighboring village Soppemo, in 2009. These parishes have about 2,000 inhabitants of which a total of 1,069 participated in the study, whereof 719 individuals participated from Karesuando (2006) while another 350 individuals participated from Soppemo (2009). For each participant in the NSPHS, blood samples were taken and serum and plasma were separated and immediately frozen and stored at ~70 °C.

**Ethical considerations.** The NSPHS was approved by the local ethics committee at the University of Uppsala (Regiona Ethisprövningsnämnden, Uppsala, 2005:325, and extension of the project was approved 2016-03-09) in compliance with the declaration of Helsinki. Informed consent to the study was given by all participants, including the examination of environmental and genetic cause of disease. If a person was not of age (<18 years), a legal guardian signed additionally. The procedure that was used to obtain informed consent and the respective informed consent form has recently been discussed in light of present ethical guidelines.

**Genetic data.** A total of 1,041 samples were successfully sequenced using Illumina short read technology (X-ten) to 30x coverage per individual. The library preparation, sequencing, and variant calling were performed as previously described. Briefly, WGS data were aligned to the GR37 using bwa-mem v0.7.12. The raw alignments were then processed according to GATK best practice using GATK v3.3. Variants were called by the GATK HaplotypeCaller 3.3 followed by variant quality score recalibration (VQSR). Sample quality control (QC) was then performed to remove genetic outliers and identify potentially contaminated samples and individuals with sex discordance errors. After QC, 1,021 unique samples with WGS data remained. Before analysis, the VCF-files were converted to PLINK-format with the PLINK software, version 1.90b4.9. Only autosomes and biallelic single nucleotide variants (SNVs) were included in the analysis. If a position had more than two alleles, PLINK keeps the two most common variants and sets the third one to a missing genotype. SNVs within a deletion were also excluded (spanning deletion/overlapping deletion, denoted #). In the same process, the variants without an rs-id were renamed to chr:position. MAF and deviation from HWE information were assessed with the --freq and --hardy commands in PLINK. The GWA analyses were performed using the GenABEL package in R. To make the files compatible with GenABEL, they were first transposed with the --recode transpose command in PLINK, and then imported into GenABEL. Variants were annotated with ANNOVAR v 2017.07.16 using the refGene database.

**Biomarker data.** Out of 1,021 individuals with WGS data, up to 1,011 individuals also have measured levels of any inflammatory biomarkers using the Proximity Extension Assay (PEA) technology provided by Olink (https://www.olink.com/products/inflammation/). Inflammatory biomarkers have been measured at...
three different timepoints within the cohort. At the first two timepoints the panels Oncology I (ONC_I) and Cardiovascular I (CVD_I) were measured. These include 31 inflammatory biomarkers (ONC_CVD). A total of 1,005 samples were measured at the first two timepoints. At the third timepoint, the panel INF I was used to measure biomarker levels (INF). Here, 92 biomarkers were measured in 903 individuals. Of the inflammatory biomarkers from the ONC_CVD dataset, 30 had overlapping measurements in the INF panel, and thus served as technical replicates. The quality control of the biomarkers has been described previously. After quality control, up to 957 individuals had available biomarker data in ONC_CVD and up to 892 from INF. Biomarkers with measurements in less than 400 individuals were excluded from downstream analyses.

GWAS. The GenABEL package was used to perform GWAS adjusting for relatedness among individuals. GenABEL utilizes a genetic kinship matrix which was estimated with the ibs function. The kinship matrix was estimated based on the SNPs listed for the HumanHap300v2_A chip. This chip contains approximately 3 million SNPs.

Table 7. Location and annotation of novel top GWAS hits from the present WGS associations that were not reported in our previous studies with genotyped/imputed data. MAF is shown for both the WGS and imputed data set as well as genotype quality for the WGS data and imputation quality for the imputed data. The lowest p-value is shown from the WGS study and the p-value from combined analyses

| Biomarker chr:position | SNV                  | MAF WGS (effect allele) | MAF imputed (effect allele) | Genotype quality (sd) WGS | Imputation quality | P WGS    | P imputed |
|------------------------|----------------------|-------------------------|-----------------------------|---------------------------|-------------------|----------|----------|
| ADA                    | 20:43255220          | rs11555566               | 0.019                       | 0.016                     | 88.60 (14.78)     | 1        | 1.35 x 10^-14 |
| CASP-8                 | 2:202178477          | rs116010659              | 0.165                       | 0.141                     | 90.00 (12.94)     | 0.968    | 7.55 x 10^-15 |
| CCL1                   | 3:42306116           | rs2228467                | 0.070                       | 0.069                     | 87.10 (14.11)     | 1        | 2.19 x 10^-9  |
| CCL20                  | 11:102742761         | rs17368659               | 0.160                       | 0.154                     | 96.12 (7.33)      | 0.999    | 9.42 x 10^-9  |
| CCL23                  | 17:34326215          | rs1712084***             | 0.087                       | 0.085                     | 95.65 (7.46)      | 0.993    | 3.10 x 10^-12 |
| CD244                  | 1:160852881          | rs171517284              | 0.378                       | NA**                      | 91.14 (13.84)     | NA**     | 1.16 x 10^-19 |
| CD5CPI                 | 3:45176853           | rs1875210                 | 0.225                       | NA**                      | 94.40 (9.59)      | NA**     | 2.62 x 10^-24 |
| CST-5                  | 20:22859017          | rs4239743                | 0.499                       | 0.494                     | 94.61 (9.83)      | 0.990    | 8.61 x 10^-21 |
| CXCL1                  | 16:57374418          | rs9921681                | 0.309                       | 0.366                     | 82.93 (18.48)     | 0.984    | 3.37 x 10^-15 |
| CCL1                   | 4:74734688           | rs3171604                | 0.331                       | 0.321                     | 91.77 (12.80)     | 0.989    | 2.46 x 10^-19 |
| CCL11*                 | 4:76943947           | rs15148618               | 0.035                       | 0.034                     | 92.21 (10.60)     | 1        | 3.44 x 10^-13 |
| CCL9*                  | 4:76943947           | rs15148618               | 0.035                       | 0.035                     | 92.21 (10.60)     | 1        | 6.19 x 10^-12 |
| FGF-5                  | 4:81843413           | rs16999073               | 0.335                       | 0.334                     | 93.58 (11.26)     | 1        | 1.50 x 10^-11 |
| MCP-3                  | 1:109407135          | rs1110257                | 0.112                       | 0.103                     | 86.27 (21.81)     | 0.985    | 1.34 x 10^-17 |
| ST1A1                  | 16:2859598           | rs138534121              | 0.064                       | NA**                      | 83.89 (16.79)     | NA**     | 2.51 x 10^-12 |
| STAMBP**               | 1:53206258           | NA**                     | 29.16 (20.59)               | NA**                      | 3.22 x 10^-6     | NA**     | 2.51 x 10^-12 |
| TGFBI                  | 19:41847860          | rs1800472                | 0.040                       | 0.041                     | 88.09 (14.06)     | 1        | 1.35 x 10^-12 |
| TNFβ                   | 6:53104575           | rs2229092***             | 0.027                       | 0.027                     | 76.95 (15.90)     | 1        | 2.70 x 10^-9  |
| TNFSF14                | 19:6665020           | rs344560                 | 0.065                       | 0.066                     | 91.63 (11.59)     | 1        | 3.72 x 10^-17 |
| sPA                    | 19:44282855          | rs346058                 | 0.046                       | 0.043                     | 87.38 (14.18)     | 0.955    | 7.12 x 10^-19 |

Table 7. Location and annotation of novel top GWAS hits from the present WGS associations that were not reported in our previous studies with genotyped/imputed data. MAF is shown for both the WGS and imputed data set as well as genotype quality for the WGS data and imputation quality for the imputed data. The lowest p-value is shown from the WGS study and the p-value from combined analyses. "\*Likely to be false positive findings (STAMBP) due to low genotype quality and CXCL9/ CXCL11 is discussed in the Supplementary (including Supplementary Figs S21–S25) "**Did not pass imputation QC or were not present in the reference panel used for the imputations. ***Two associations were not reported in our previous study due to the two-stage design (discovery and replication) even though the p-value was significant in the combined analyses."
QQ-plots and Manhattan plots were produced with the qqman package in \( \text{R}^{39} \). Regional association plots were constructed using Locuszoom\(^2\). The 1000G Nov 2014 EUR population was used for the coloration based on LD. If one biomarker had different top SNVs in the technical replicates, the LD coefficient \( (R^2) \) was calculated using PLINK within the study population itself (NSPHS). To assess the size of the associated regions, an SNV clumping was performed in PLINK. A clump kb radius of 15 Mb (--clump-kb), a p-value threshold of \( 1 \times 10^{-8} \) (--clump-p1) and an \( R^2 \) cut-off set to default (--clump-r2 0.1) was used. This function clumps SNVs together based on empirical estimates of LD. The range (in bp) of the clumps was then calculated for each biomarker and used to define biomarker-based loci. We also performed conditional analysis, where the top SNV, for each marker with a significant hit, was used as covariate in the mmscore function to see whether there was more than one independent association for each biomarker. These clump-defined loci were then used to calculate a biomarker-based significance threshold for results in the conditional analyses. SNVs with a conditional p-value below 0.05/number of SNVs tested in the predefined locus, were considered as an independent association. If the biomarker had a significant secondary (conditional) signal, a third analysis was performed, adjusting for both the primary and secondary signal. No MAF cutoff was used in the conditional analyses.

**Narrow-sense heritability estimates.** Narrow-sense heritability \( (h^2) \) was estimated using the polygenic model in GenABEL. First, the heritability for the biomarker measurements was estimated only adjusting for age, sex, batch effects and kinship. Then, as an estimation of SNV heritability, the top variant was used as a covariate. The difference in heritability estimated between the models gives the variance explained by the top variant. If secondary and tertiary signals were present, they were added as additional covariates and SNV heritability was calculated for each separately. To test significance of heritability, the reported function minimum (twice the negative maximum log-likelihood) is compared to the reported function minimum in a polygenic model with a fixed heritability estimate set to zero. The difference gives a test approximately distributed as chi-squared with 1 degree of freedom.

**Colocalization with published GWAS data and comparison with previous biomarker studies.** If the top SNVs of two different GWASes are in LD \( (R^2 > 0.8) \), the phenotypes are considered to be colocalized. In the study cohort, the LD pattern was calculated between each top SNV and all SNVs within 2 Mb, using PLINK. All variants that were in LD \( (R^2 > 0.8) \) with a top SNV were extracted. These were used as query to test for colocalization with data from the GWAS catalog (The NHGRI-EBI Catalog of published genome-wide association studies, [https://www.ebi.ac.uk/gwas/home](https://www.ebi.ac.uk/gwas/home)) to find out whether the variants have already been published in earlier association studies of inflammatory biomarkers. Entries with p-values up to \( 1 \times 10^{-6} \) in the GWAS catalog (version 1.0.2 – downloaded 2018-10-29) were included in the comparison. All entries for the top variants and variants in LD found in the catalog was extracted together with their metadata.

Our GWAS results were also tested for colocalization with the expression quantitative trait locus (eQTL) dataset from Westra et al\(^{51} \). This dataset consists of both cis-eQTL and trans-eQTLs and is based on an eQTL meta-analysis in non-transformed peripheral blood samples. The names of the genes encoding the biomarkers were matched both in the cis and trans dataset to see if the top SNVs in this study had been reported as an eQTL. If not, LD was calculated between the top SNVs of our study and the most significant eQTL variant, using PLINK, also here within the study cohort (NSPHS).

Further replication was made using another population from Northern Europe. Circulation cytokines have been measured in Finnish populations by Ahola-Olli et al\(^{29} \). We sought to replicate the results from the 19 inflammatory biomarkers (i.e. cytokines) that were present in both studies: bNGF, CCL11, CCL3, CCL4 (MIP1b), CXCL1 (GROa), CXCL10 (IP10), CXCL9 (MIG), HGF, IL-4, IL-5, IL-7, IL-8, IL-10, IL-13, IL-18, MCP-1, MCP-3, SDF, TNFB and TRAIL. However, IL-4, IL-5 and IL-13 did not pass QC in our study, and thus, only 16 biomarkers could be compared. Only top SNVs were used in the replication.

Top variants were compared to the results from previous studies with genotyped/imputed data in the same cohort\(^{15,28} \). In the first study, the ONC_CVD biomarkers were analyzed\(^{28} \) in the whole cohort whereas in the second study\(^{29} \), the INF biomarkers were analyzed using a two-stage design by splitting the cohort into a discovery and replication cohort (genotyped using different arrays and imputed independently) followed by combined analyses for the SNVs that replicated. Even if the two previous studies were based on the same genotyping and imputation data, the quality control of the imputed genotypes where slightly different with one additional requirements of a genotype probability score of Info >0.9 in 95% of the individuals for the analyses of the INF biomarkers whereas the criteria of imputation quality >0.3 and HWE p-value >0.05/the number of SNPs within the sub cohorts based on different genotyping arrays as well as combining the two sub cohorts. More information about genotyping, imputation and quality control of imputed variants is included in the previous articles\(^{25,28} \). In this comparison, the additional filters were also applied to the imputed data similar to the WGS data, i.e. only biallelic autosomal hits were compared. If an association was previously found with an indel or a variant in a spanning deletion, the biallelic variant with the lowest p-value after the indel or spanning deletion was used for comparison. For a subset of the non-overlapping hits, we also compared the WGS genotypes with the imputed dosage values from the previous studies.

**Data availability**

Summary statistics from the GWAS will be uploaded to the GWAS catalog.

Received: 11 June 2019; Accepted: 26 October 2019;
Published online: 14 November 2019

**References**

1. Welte, D. et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* 42, 1001–1006 (2014).
2. Marouli, E. et al. Rare and low-frequency coding variants alter human adult height. *Nature* 541, 186–190 (2017).
3. Allen, H. L. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832–838 (2010).
48. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, 1–7 (2010).
49. Turner, D. S. qacman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *J. Open Source Softw.* **3**, 731 (2018).
50. Pruim, R. J. *et al.* LocusZoom: Regional visualization of genome-wide association scan results. *Bioinformatics* **27**, 2336–2337 (2011).
51. Westra, H. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Publ. Gr.* **45**, 1238–1243 (2013).

**Acknowledgements**
The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under project sens2016007. The research was funded by The Swedish Society for Medical Research (SSMF), the Swedish Medical Research Council, and the Kjell and Märta Beijers, the Marcus Borgström, the K and O F Hedströms, the Åke Wiberg, the Swedish Heart-Lung Foundation, and Science for Life Laboratory. Open access funding provided by Uppsala University.

**Author contributions**
J.H. and Å.J. designed the study; J.H. performed the data analysis, and generated the figures; J.H. and T.K. performed the statistical analysis; S.E. provided data; J.H. wrote the manuscript; J.H., T.K., M.R.-A., N.R., W.E.E., S.E. and Å.J. interpreted the data, contributed to and reviewed the manuscript.

**Competing interests**
The authors declare no competing interests.

**Additional information**
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-53111-7.

**Correspondence** and requests for materials should be addressed to J.H.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019