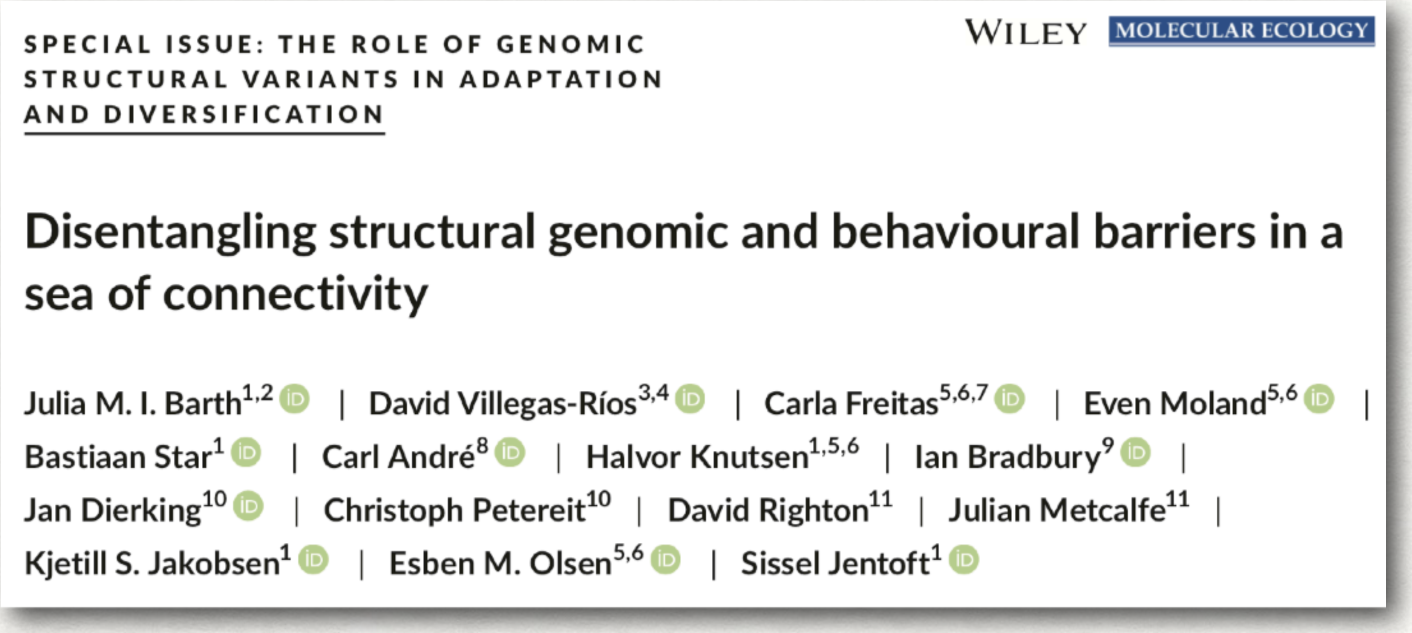
FIRST STEPS IN GENOMIC DATA ANALYSIS

Julia M.I. Barth, University of Basel

Learning goals

- Know how to apply UNIX and R commands for genomic analysis
- Understand the Variant Call Format (VCF)
- Understand and be able to apply different quality filtering steps

The data file



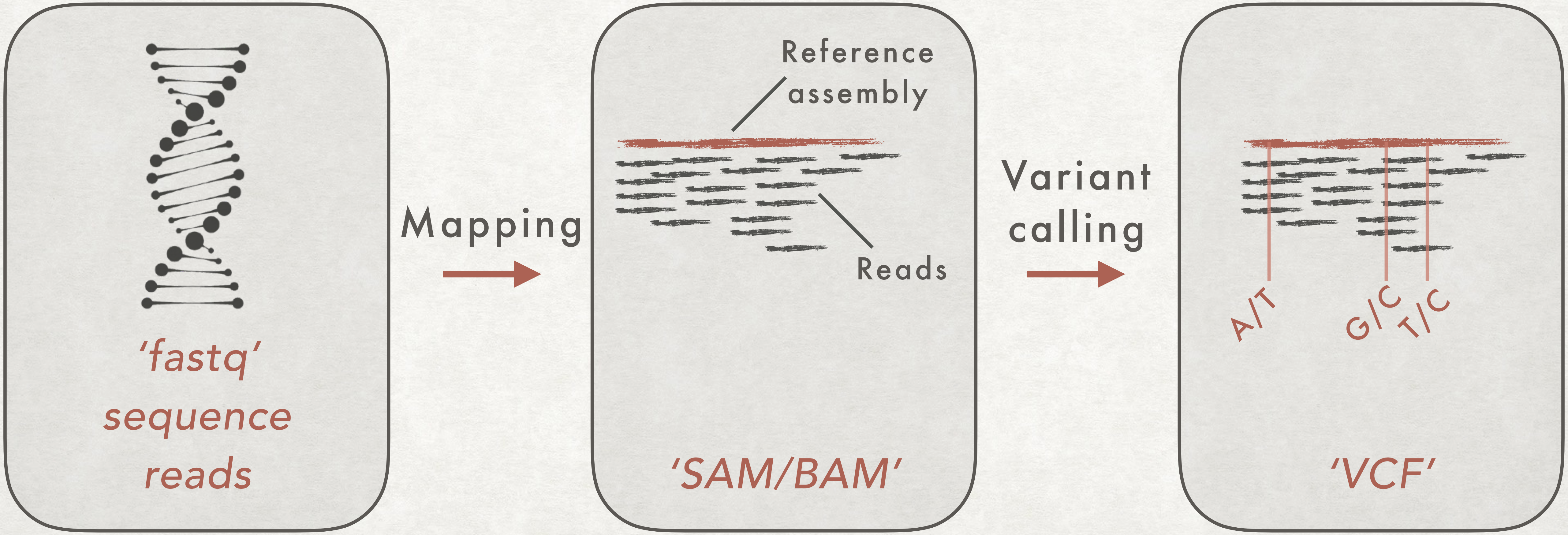
SPECIAL ISSUE: THE ROLE OF GENOMIC STRUCTURAL VARIANTS IN ADAPTATION AND DIVERSIFICATION

WILEY MOLECULAR ECOLOGY

Disentangling structural genomic and behavioural barriers in a sea of connectivity

Julia M. I. Barth[1,2] | David Villegas-Ríos[3,4] | Carla Freitas[5,6,7] | Even Moland[5,6] | Bastiaan Star[1] | Carl André[8] | Halvor Knutsen[1,5,6] | Ian Bradbury[9] | Jan Dierking[10] | Christoph Petereit[10] | David Righton[11] | Julian Metcalfe[11] | Kjetill S. Jakobsen[1] | Esben M. Olsen[5,6] | Sissel Jentoft[1]

Phto by Øystein Paulsen



'fastq' sequence reads

Mapping

Reference assembly

Reads

'SAM/BAM'

Variant calling

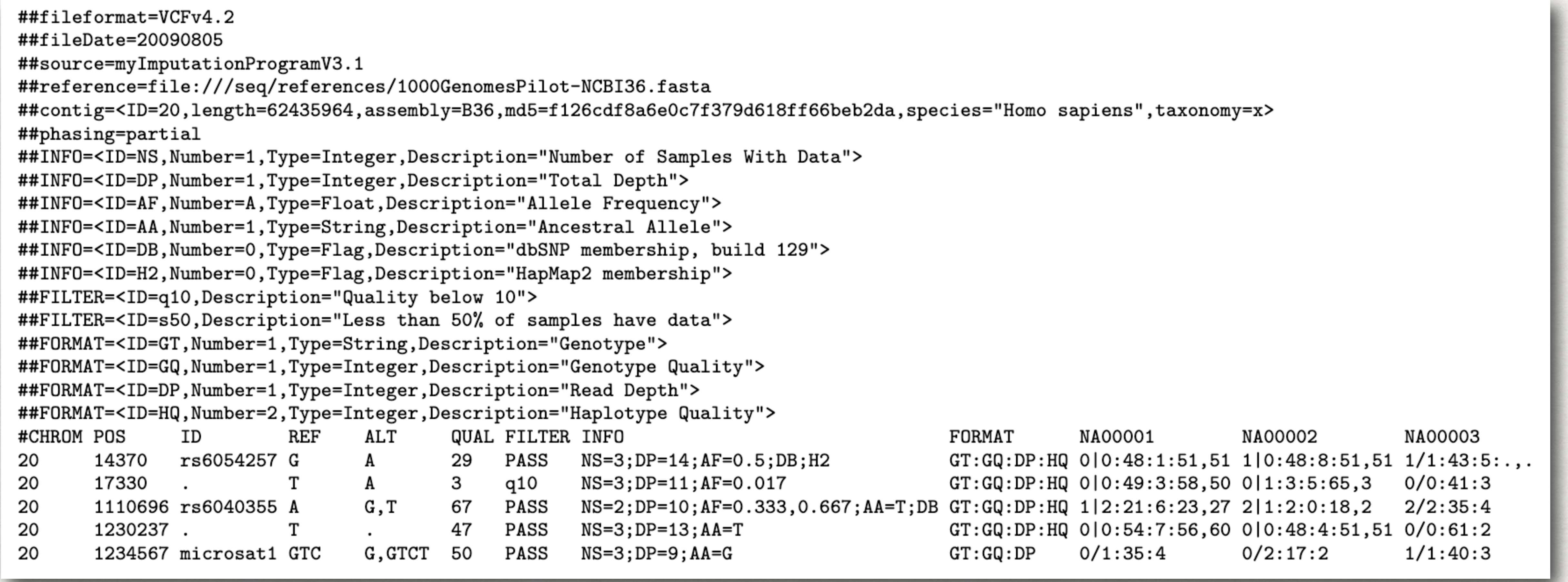
A/T

G/C

T/C

'VCF'

1. The Variant Call Format (VCF)

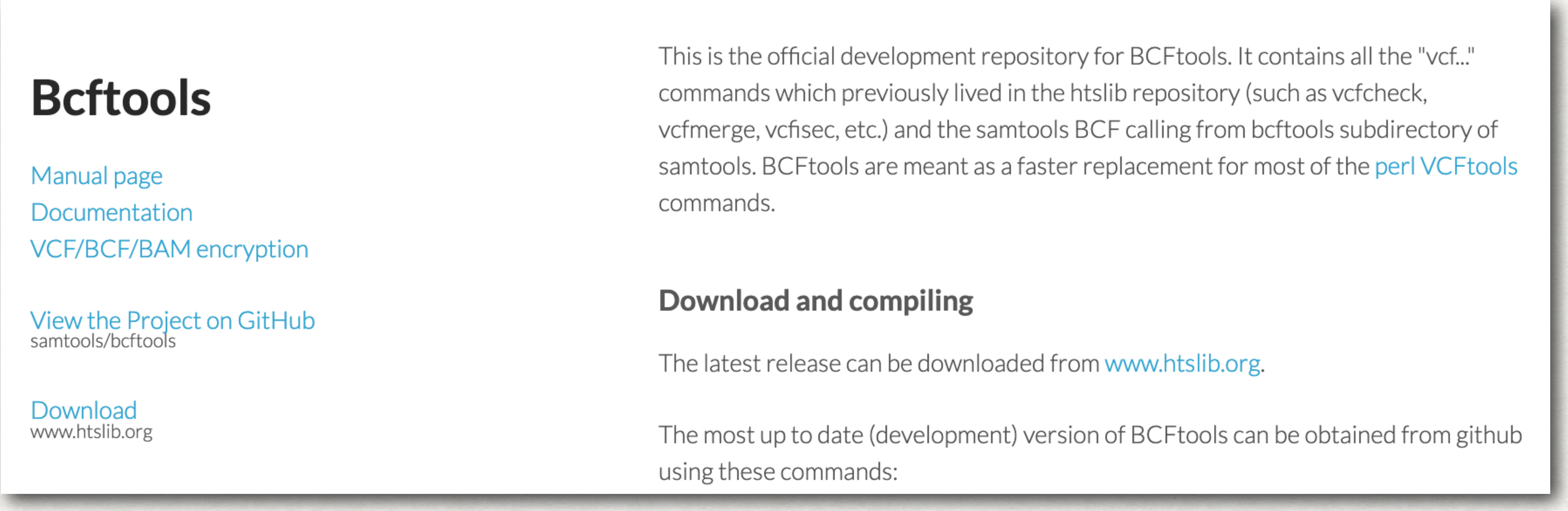


```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001        NA00002        NA00003
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:.,.
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3   0/0:41:3
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2   2/2:35:4
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:4       0/2:17:2       1/1:40:3
```

https://samtools.github.io/hts-specs/VCFv4.2.pdf

2. Hard quality filtering of variants

3. Further quality filtering



Bcftools

Manual page
Documentation
VCF/BCF/BAM encryption

View the Project on GitHub
samtools/bcftools

Download
www.htslib.org

This is the official development repository for BCFtools. It contains all the "vcf..." commands which previously lived in the htslib repository (such as vcfcheck, vcfmerge, vcfisec, etc.) and the samtools BCF calling from bcftools subdirectory of samtools. BCFtools are meant as a faster replacement for most of the perl VCFtools commands.

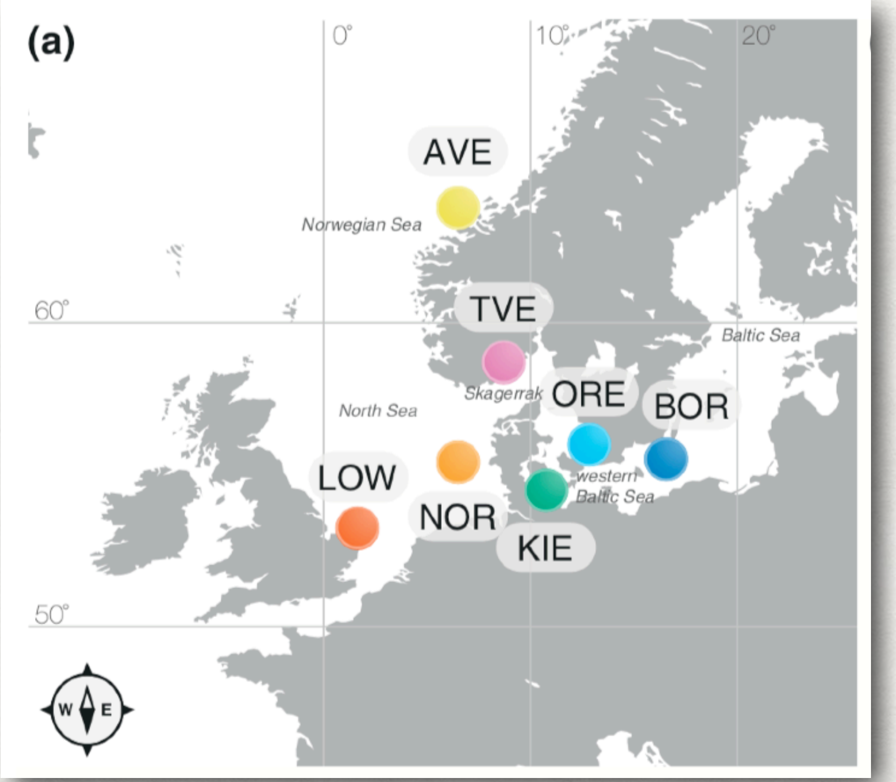
Download and compiling

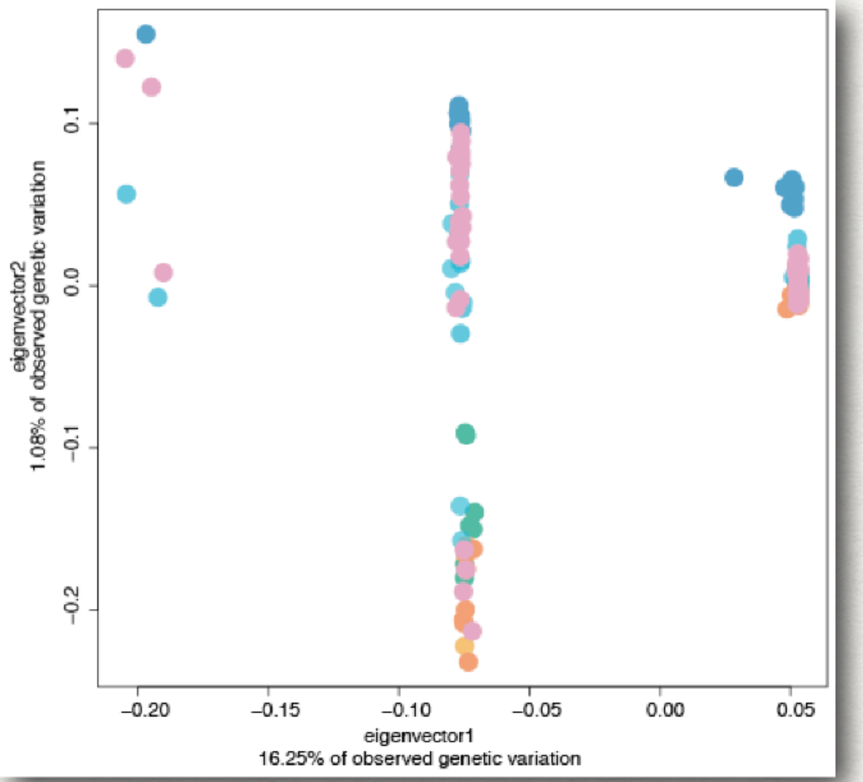
The latest release can be downloaded from www.htslib.org.

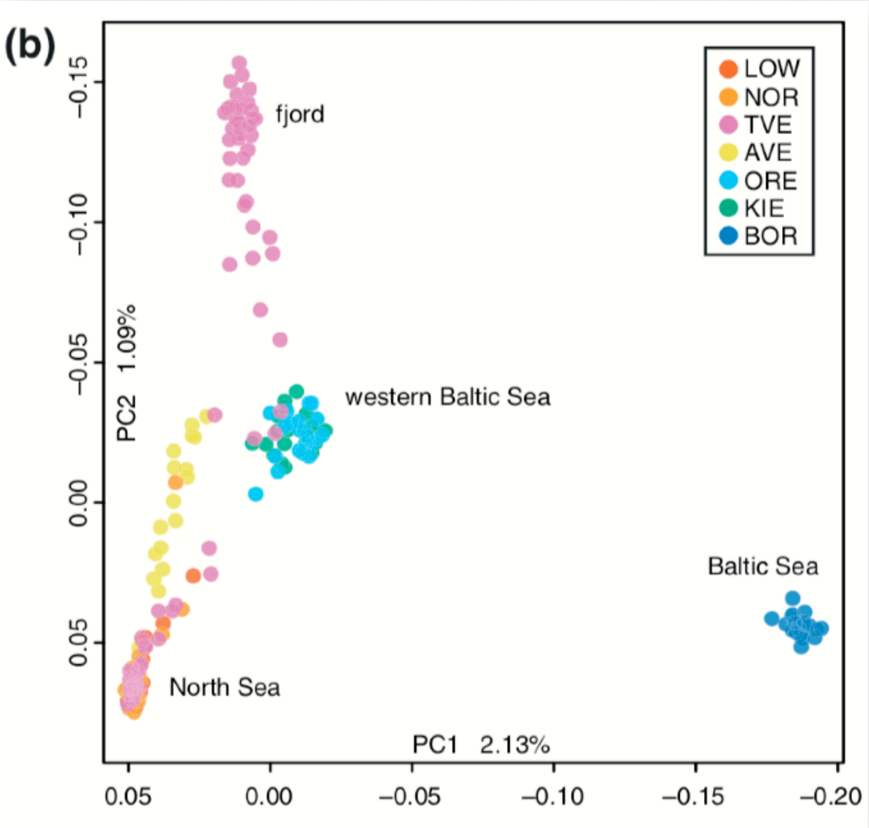
The most up to date (development) version of BCFtools can be obtained from github using these commands:

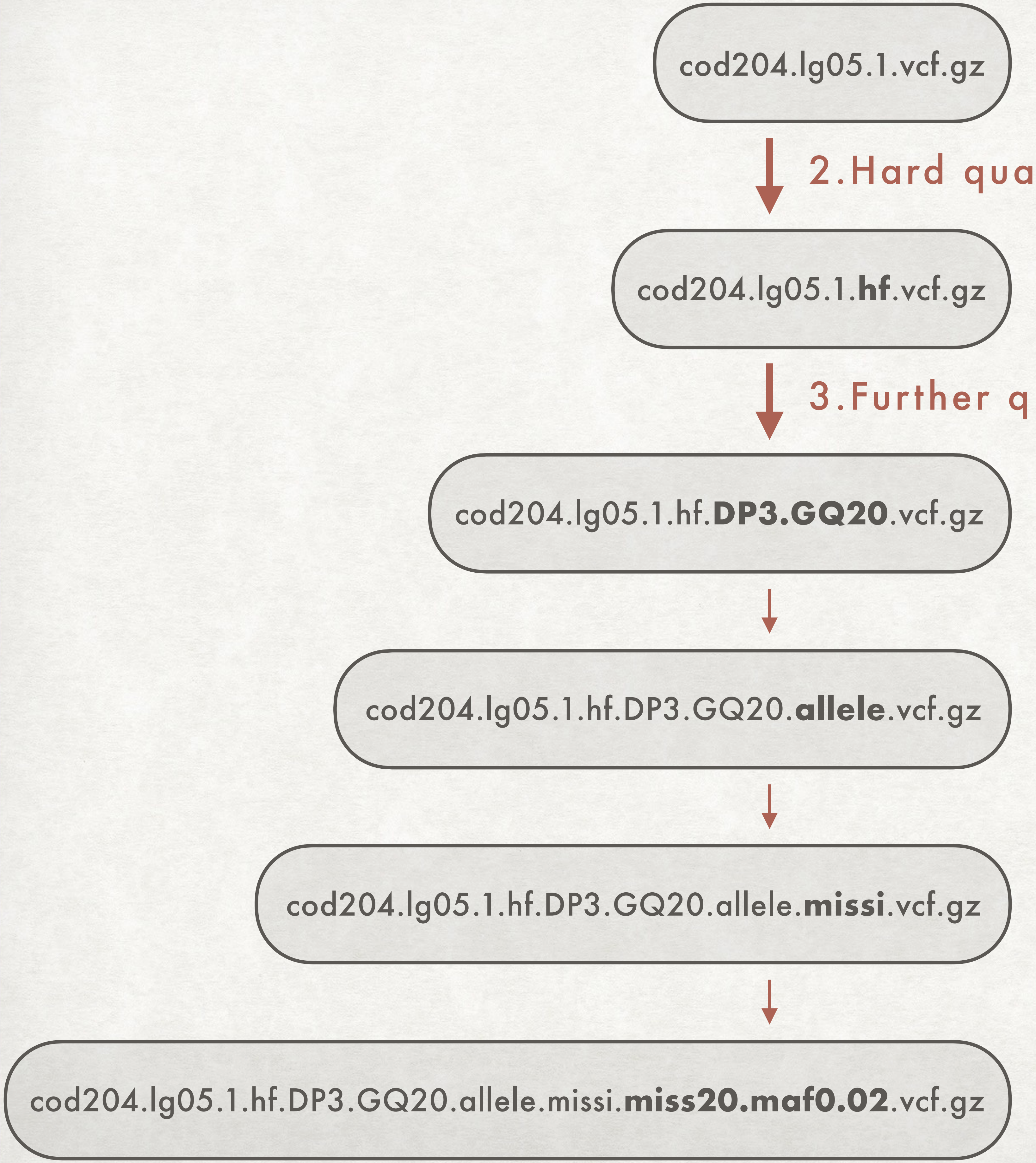
https://samtools.github.io/bcftools/

4. Performing a principal component analysis (PCA)









1. The VCF format

cod204.lg05.1.vcf.gz

- Raw VCF incl. all called variants
- Chromosome 5 of Atlantic cod
- 204 individuals, 7 sampling sites

2.Hard quality filtering of variants

cod204.lg05.1.**hf**.vcf.gz

- INFO field measurements (apply to the variant site and include information across all individuals)

3.Further quality filtering

cod204.lg05.1.hf.**DP3.GQ20**.vcf.gz

- FORMAT field measurements (apply to the single genotype of one individual at one variant site)

cod204.lg05.1.hf.DP3.GQ20.**allele**.vcf.gz

- multiallelic SNPs
- monomorphic SNPs
- indels
- SNPs in the close proximity of indels

cod204.lg05.1.hf.DP3.GQ20.allele.**missi**.vcf.gz

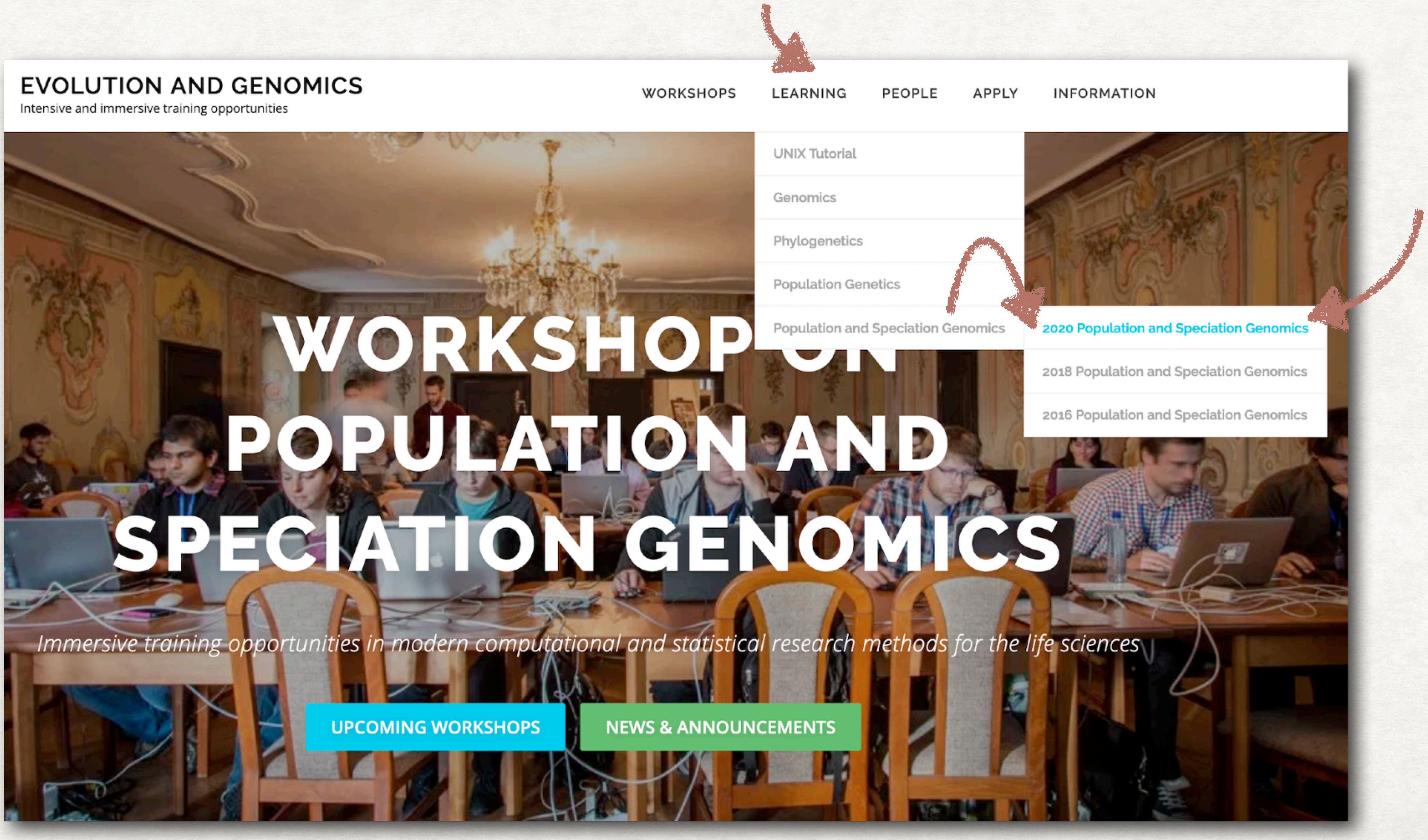
- Remove individuals with a high amount of missing data

cod204.lg05.1.hf.DP3.GQ20.allele.missi.**miss20.maf0.02**.vcf.gz

- Remove variants with a high amount of missing genotypes
- Filter on minor allele frequency

FIRST STEPS IN GENOMIC DATA ANALYSIS

http://evomics.org/learning/population-and-speciation-genomics/2020-population-and-speciation-genomics/first-steps-in-genomic-data-analysis/


EVOLUTION AND GENOMICS
Intensive and immersive training opportunities
WORKSHOPS
LEARNING
PEOPLE
APPLY
INFORMATION
UNIX Tutorial
Genomics
Phylogenetics
Population Genetics
Population and Speciation Genomics
2020 Population and Speciation Genomics
2018 Population and Speciation Genomics
2016 Population and Speciation Genomics
WORKSHOP ON POPULATION AND SPECIATION GENOMICS
Immersive training opportunities in modern computational and statistical research methods for the life sciences
UPCOMING WORKSHOPS
NEWS & ANNOUNCEMENTS