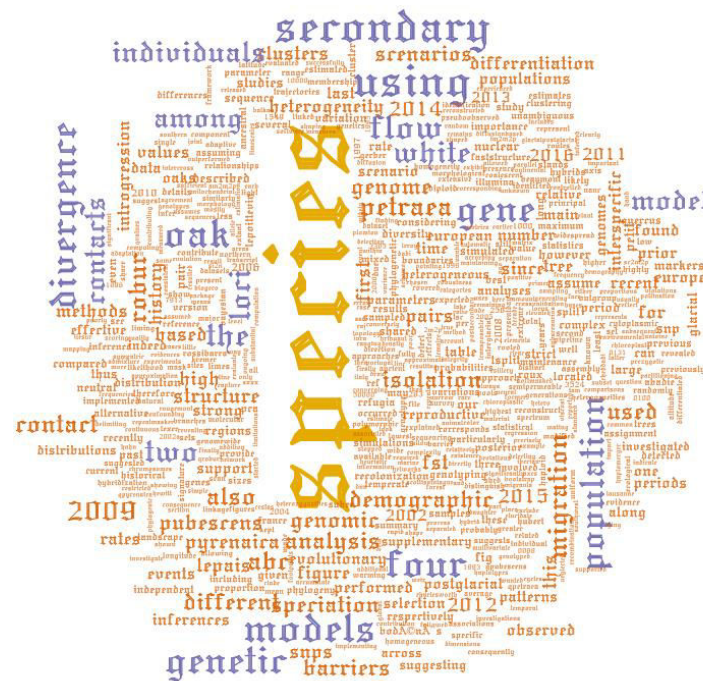


Part 1: Importance of preliminary analyses in population genomics

- Practical feedback based on my past experience in population genetics -



Thibault Leroy, University Assistant

Genomic approaches

22 April 2020



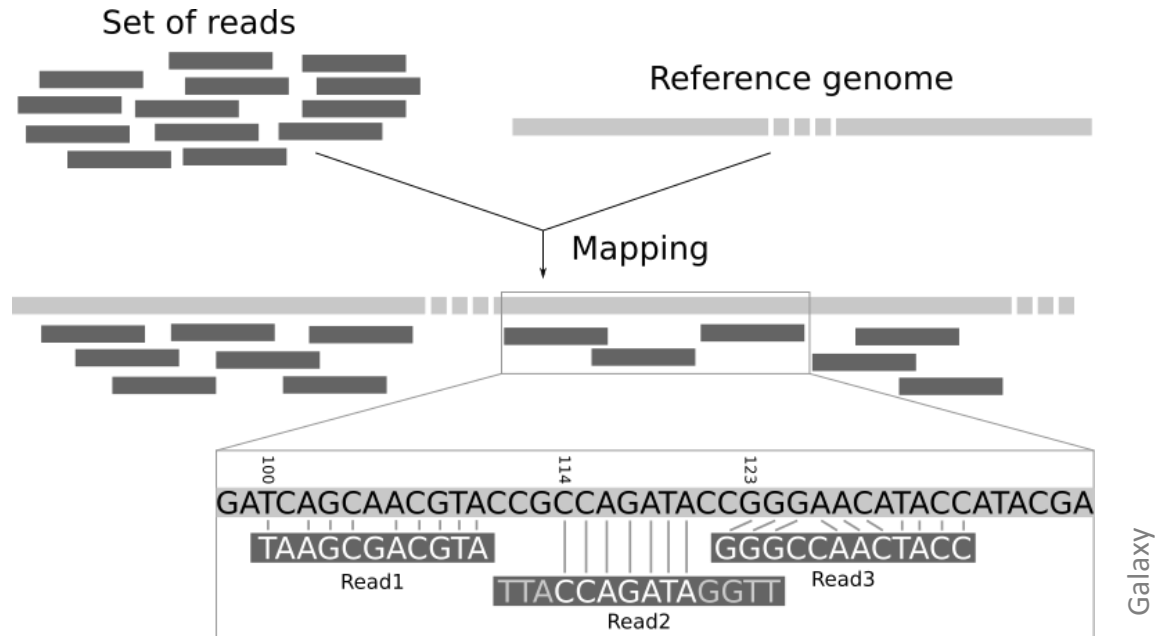
After one year of field work, library preparation and a long wait, your new project in populations genomics can concretely start.

- The sequencing of the data just finished
- You just received the first draft of the reference genome thanks to some collaborators

Understandably, you are impatient to have your results, either for your master defense, or your next PhD committee, etc... or just because it was a very long wait. (Your PI is also very impatient)

What next?

You will probably directly map the genetic variants against this reference sequence to identify variants between your different individuals/populations



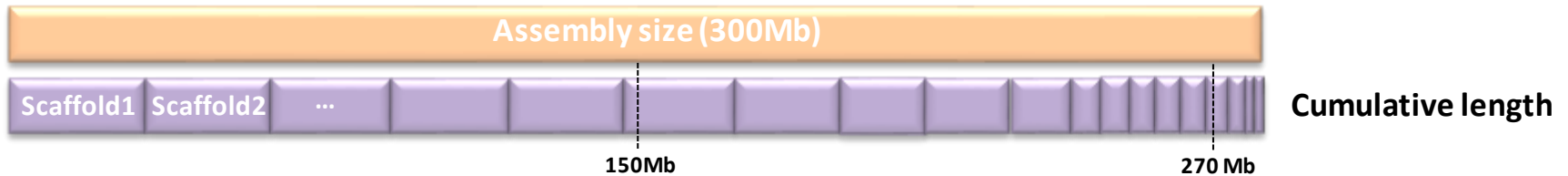
Good idea, but...

... given that this step will take a couple of days:

You can take this time to look at the reference genome provided by your collaborators !

Summary statistics of the assembly

- The total length of the assembly (consistent with the expectation, e.g. flow cytometry?)
- Number of scaffolds/contigs, N50, L50...

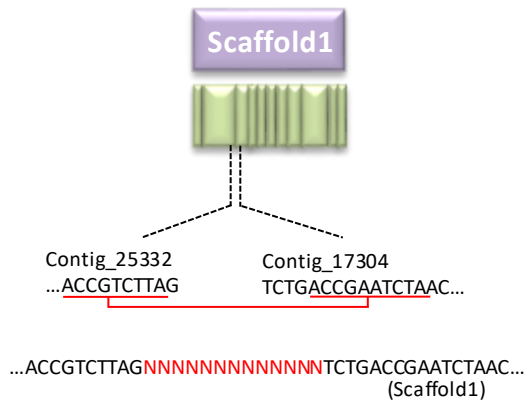


N50: sequence length of the shortest contig at 50% of the total genome length (example above: the length of the 6th longest scaffold)

L50: number of scaffolds/contigs to reach half of the genome size (example above: 6)

Over the last five years, the quality of the draft genomes were strongly improved by the availability of long-read sequencing methods. As a consequence, it is quite easy to observe excellent results for the N50. N90 and L90 are therefore more widely evaluate to further evaluate the quality of a given assembly.

- **Proportion of unknown nucleotides in the genome (N content)**



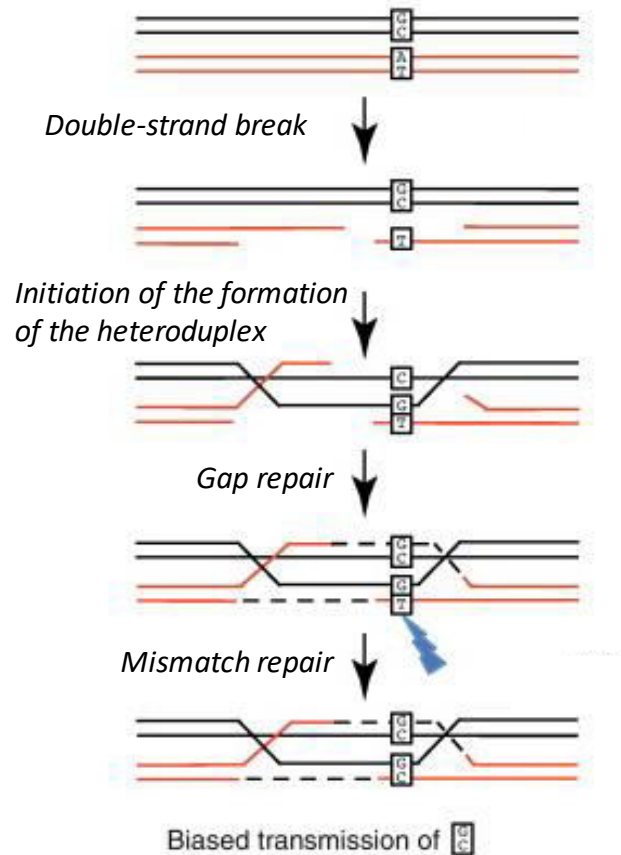
Stretches of N (contig junctions) or N in the contig sequence (depending of the parameters)

Ten years ago, it was quite frequent to observe genome sequence with up to 10% of N, now decreasing.

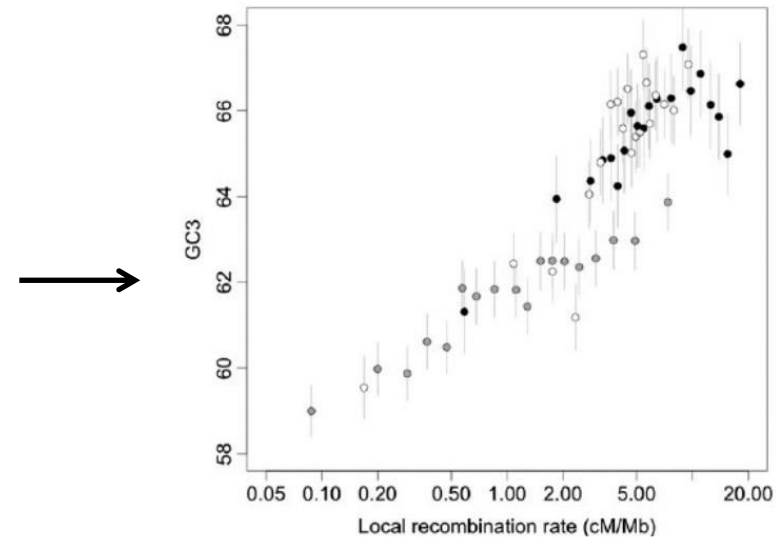
DNA composition (GC content)

Mutation is biased towards A and T
(G and C sites are more mutable than A and T sites)

Recombination is biased toward G and C (GC-biased gene conversion, gBGC)



Webster & Durst, 2012 Trends in genetics



Serres-Giardi *et al.* 2012

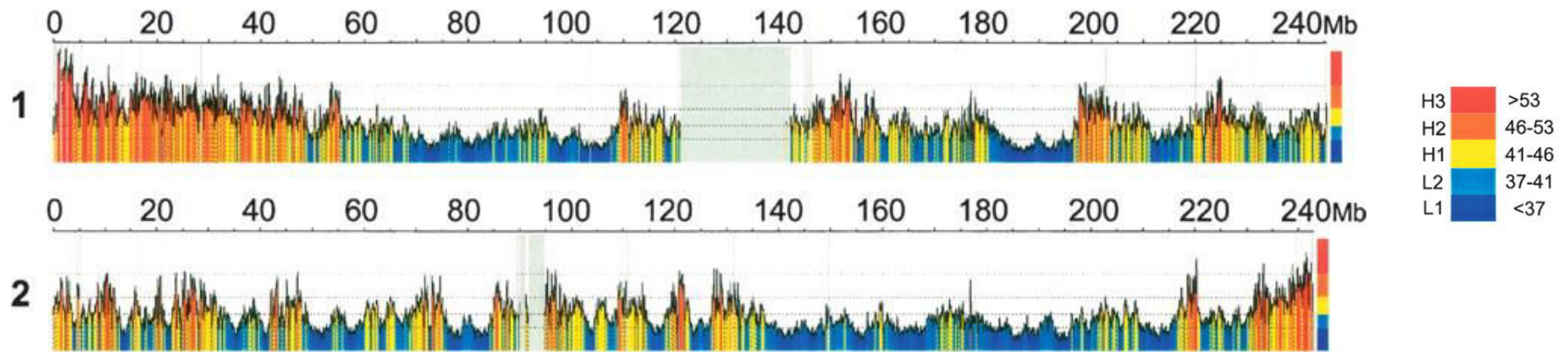
GC3 = GC content at third-codon positions of protein coding genes (redundancy in the genetic code)

DNA composition (GC content)

Mutation is biased towards A and T
(G and C sites are more mutable than A and T sites)

Recombination is biased toward G and C (GC-biased gene conversion, gBGC)

→ In species with a sexual reproduction stage, the genome-wide variation in GC content can be informative of the local recombination rates



Costantini et al. 2006 genome research

EXCITING NEWS!

After a week of computations, the mapping and SNP calling steps are finished and you have now your vcf file in hand. Tens of hundreds of thousands of variants (SNPs or INDELs), perhaps millions, and a lot of biological questions to (try to) answer!

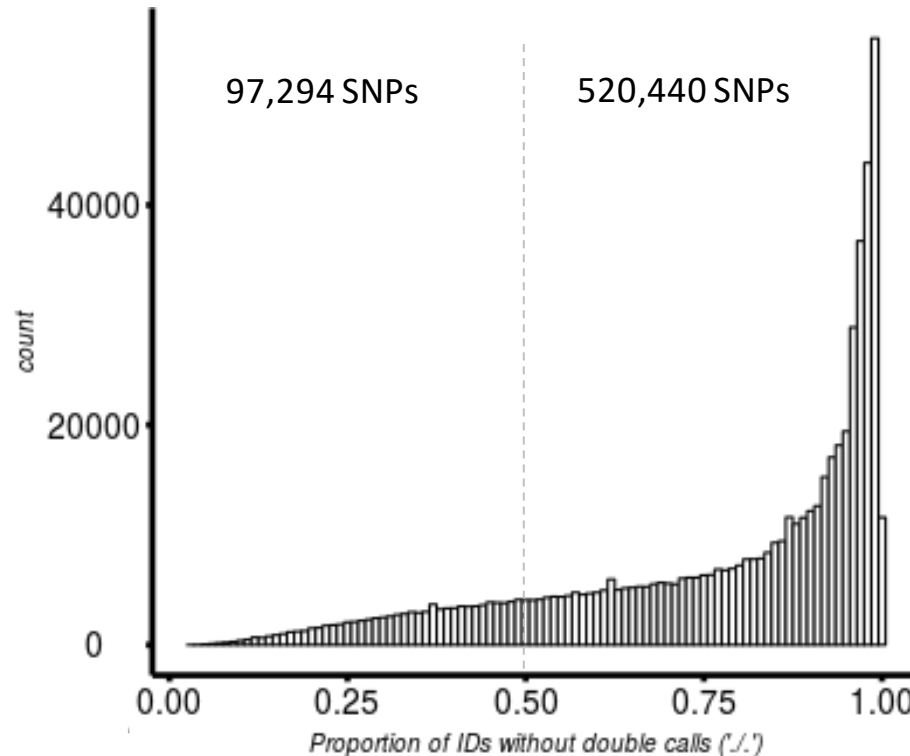
But...



(I assume here that you followed the guidelines for the mapping, SNP calling and variant filtration, e.g. see Ovidiu's presentation, GATK best practices workflows, ...)

Missingness: proportion of missing calls in the vcf file (./.)

Lack of coverage can contribute to a high proportion of variant calls!



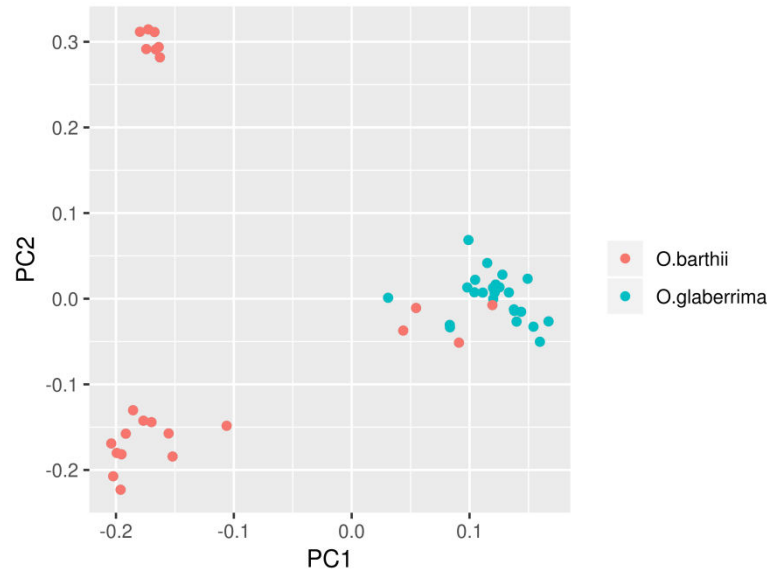
A critical example from a past collaboration

Only consider SNPs that are informative at the population level (excluding SNPs with e.g. 10, 25 or 50% of missing calls, depending of the number of individuals you initially sequenced)

Principal component analysis (PCA)

- Principal component analysis (PCA) is a mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the data set.
- The principal component (PC) axes are chosen to reflect major axes of variation in the data, with the first PC representing the largest variance explained, the second the second most, and so on. Each PC is perpendicular to (i.e. uncorrelated) the other PC.
- Exceptionally fast and powerful
- Widely used in genetics/genomics:
 - **Give a location to each individual data-point on each of a small number PC axes**
 - Perform genome scans for positive selection (PCAdapt, Duforet-Frebourg et al. 2016, Luu et al. 2017)

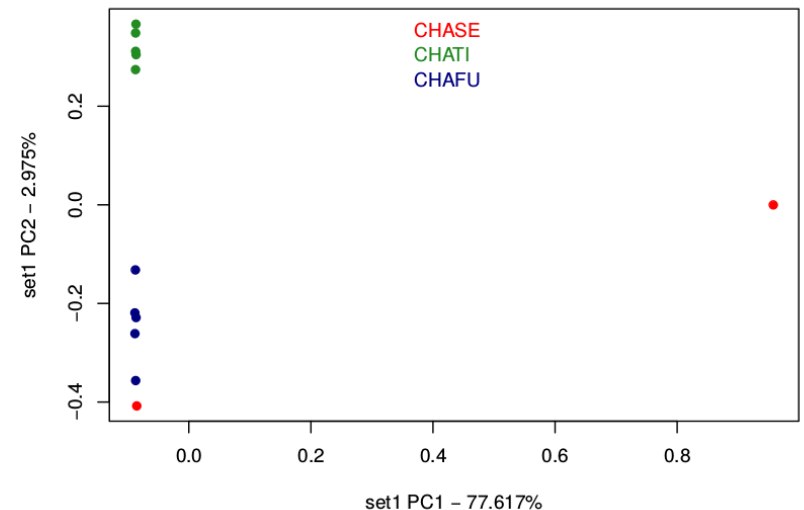
Principal component analysis (PCA)



Wet lab errors (including sequencing platforms):
Mislabelling of individuals, ...

Field work errors:

Misidentification of individuals in the field are widespread, especially in plants



Such errors are commonly observed and do not mean that it is your fault. Identify them.

(it is also possible to use Bayesian genetic clustering algorithms, such as STRUCTURE, but PCA is several orders of magnitude faster)

Principal component analysis (PCA)

The opposite situation is possible: identifying clones

Biologically:

Non-domesticated species: plants with vegetative propagation

Domesticated species: the same cultivar registered under different names

Errors:

Same individual collected twice, DNA extraction performed twice, ...



SampleIndex	sampleID	PC1	PC2	PC3
...				
9	JF3880	-0.0791949563460385	0.179937176836918	-0.183158524302421
10	RCHB1887	0.0458936978728019	0.18226398855529	0.189701803130664
11	RCHB1917	-0.503419094984371	-0.386838229450701	0.0715529676922778
12	RCKB_1917	-0.50165354153268	-0.384285318590499	0.0707883200805732

Phylogenetic trees

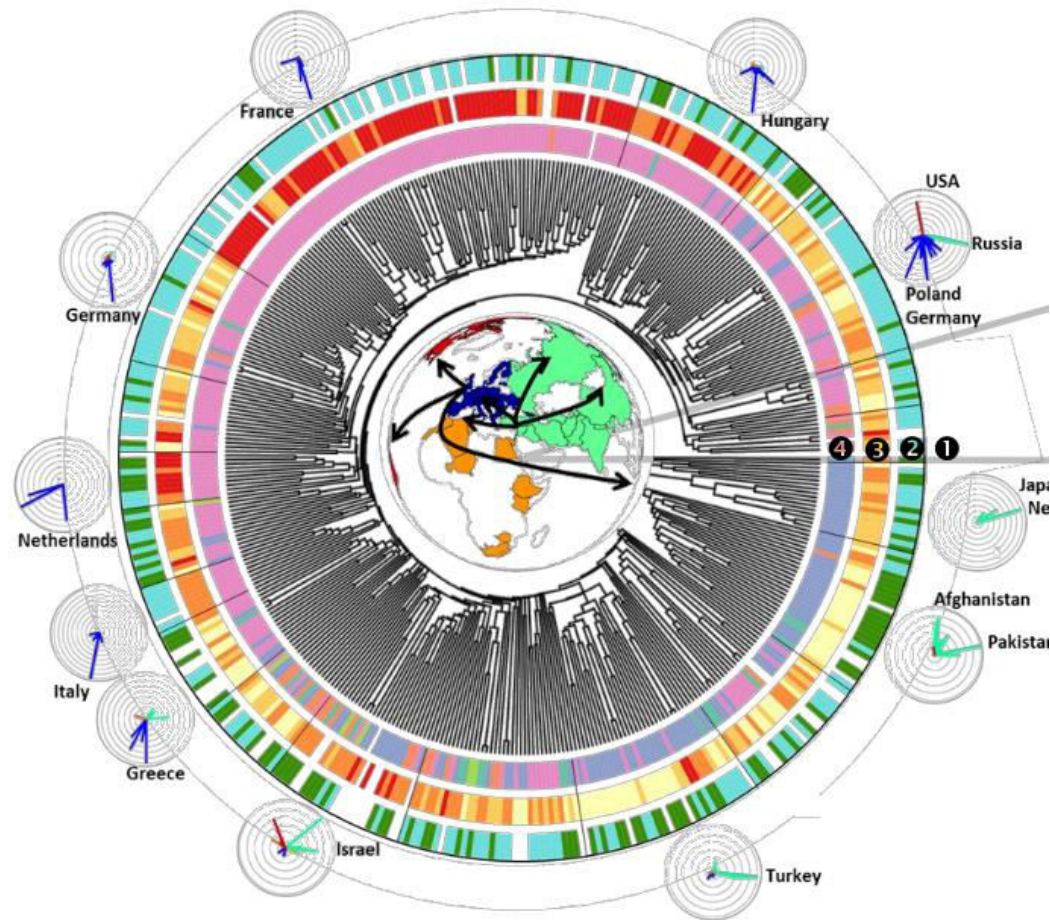
Trees can also be very informative

- Pattern of branching
- Genetic distances

Different branching than expected?
(mislabelling,
misidentification?)

Some individuals with longer branches than expected?
(individuals with poor
genotyping quality?)

Little or no distance between
two samples?
(putative clones?)



Pont et al. 2019 Nature Genetics

SNP density

The distribution of SNPs along the genome is not homogenous.

In particular, SNPs are more frequently observed non-coding regions than in coding regions or, in general, where natural selection is more acting.

But this not the only effect, genetic recombination and mutation rate also vary along the genome and contribute to this heterogeneous landscape of SNP density.

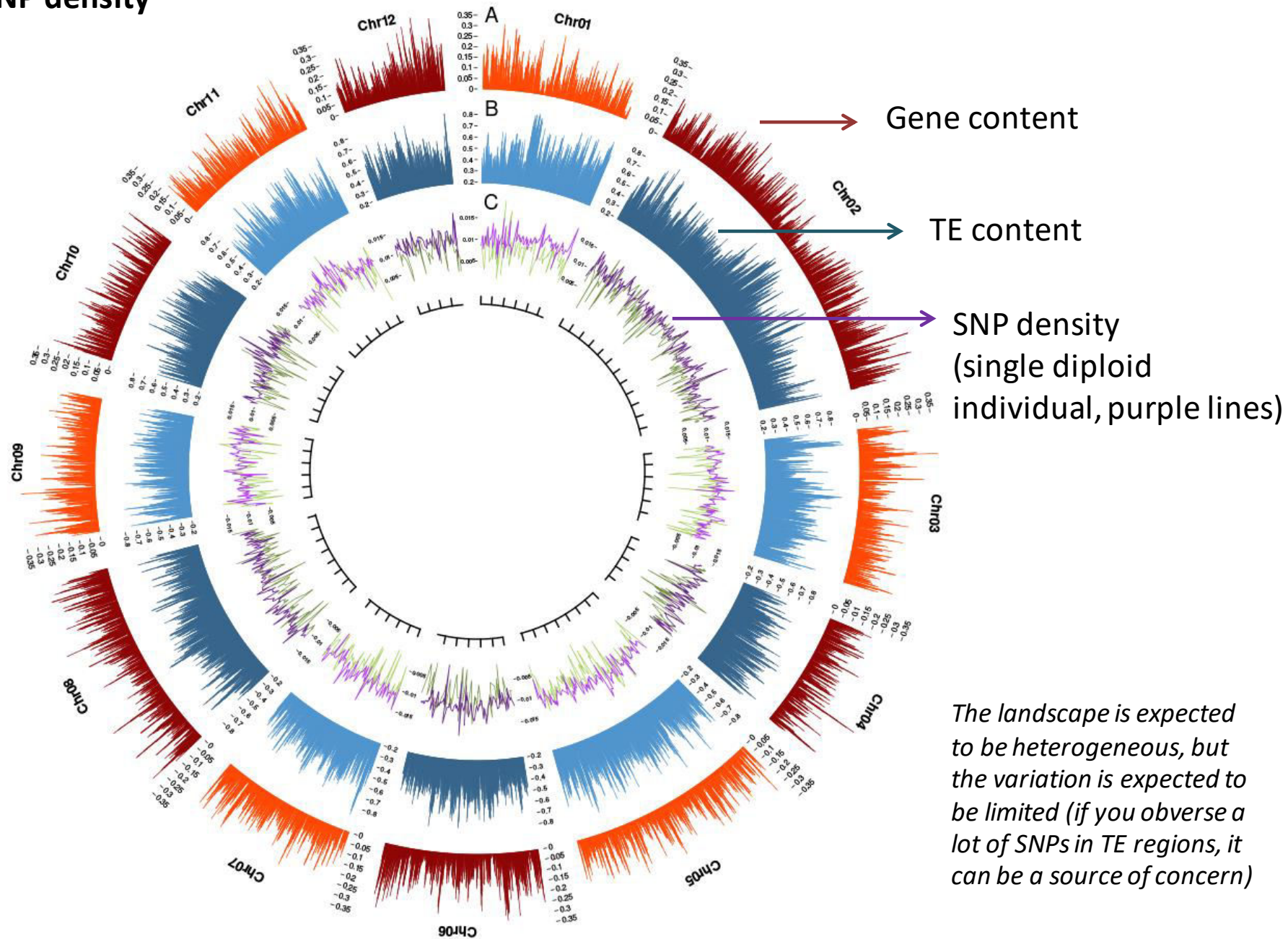
Having said that, if you observe a huge number of SNPs in some regions of the genome, you need to remain prudent, it can be due to false positive errors.

This is particularly true in repetitive regions such as regions containing a high proportion of transposable elements (TE) due to lower quality mapped reads.

	Total		Syntenic build	
	No. SNPs	SNPs/Mb	No. SNPs	SNPs/Mb
7A	1 486 040	4077	42 041	3212
7B	1 860 295	4737	38 508	3384
7D	671 976	1939	20 563	1088

(e.g. wheat subgenomes, Lai et al. 2015)

SNP density



The landscape is expected to be heterogeneous, but the variation is expected to be limited (if you observe a lot of SNPs in TE regions, it can be a source of concern)

Transition/transversion ratios (ts/tv or ti/tv)

4 transitions possible, 8
transversions possible.

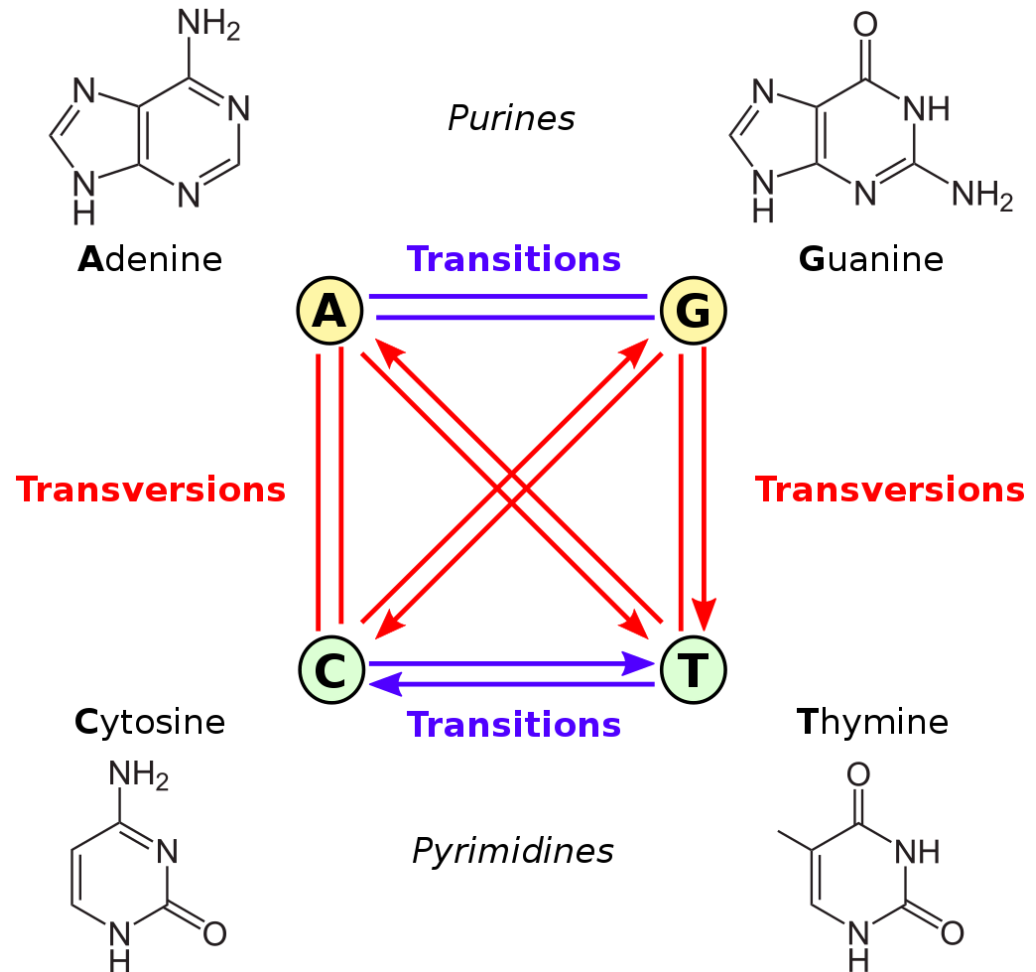
=> Assuming a totally
random process:
transitions/transversions=0.5

But again, mutation is not a
completely random process,
mutations from purines to purines
(A<->G) and from pyrimidines to
pyrimidines (C<->T) are more
frequent.

In human: ts/tv >2.1

Even if such a value depends of the
clades and the genomic regions, but a
ts/tv ratio of at least 2 is generally
observed.

**A low ts/tv ratio is generally indicative of a
lot of false positive SNPs.**



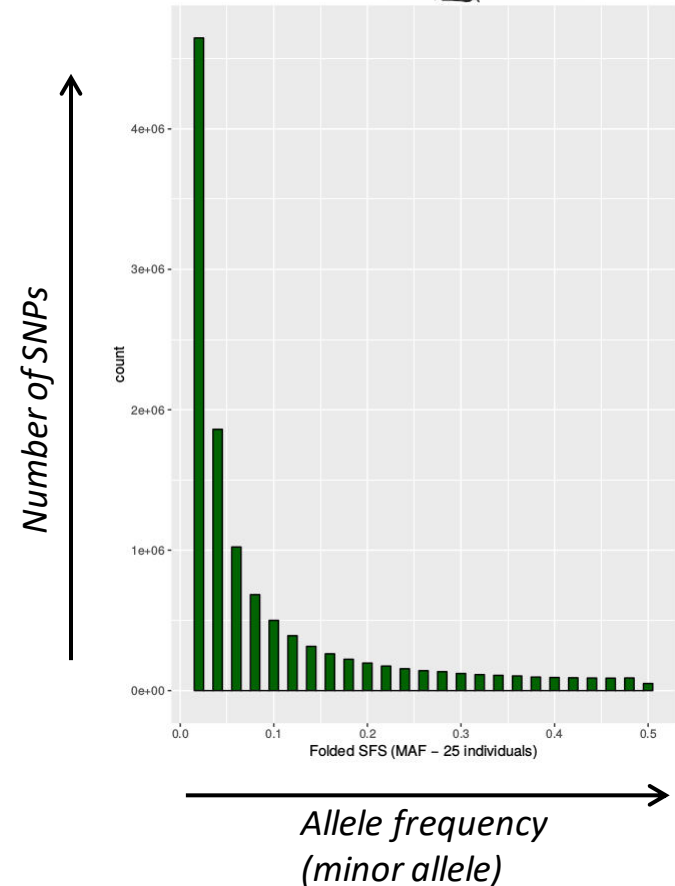
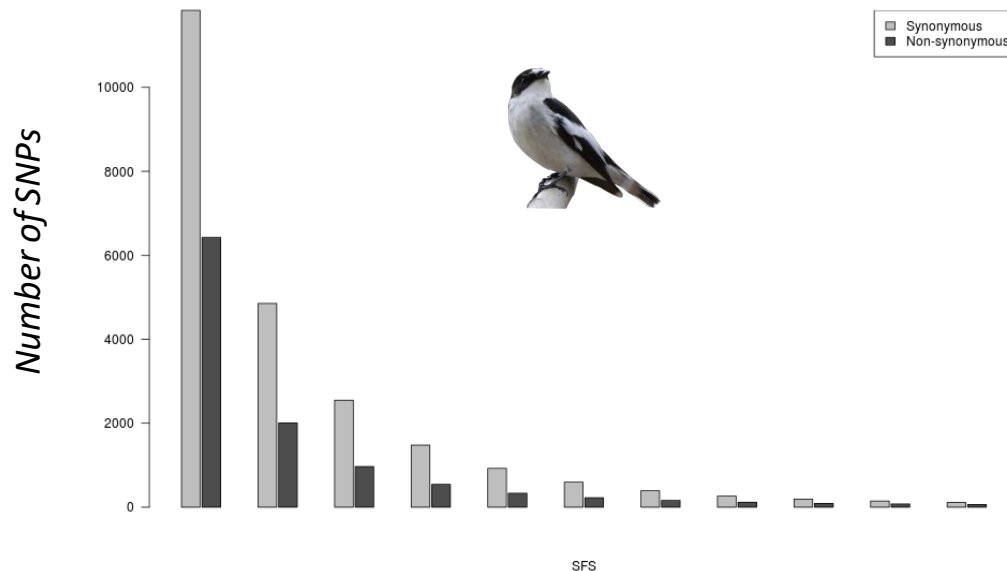
Site Frequency Spectrum (SFS)



Under selective neutrality, the site frequency spectrum is L-shaped, with a majority of variants at low frequency and few variants at high frequency.

Folded SFS = frequency of the minor allele
(minor = less common allele for a SNP)

Unfolded SFS= frequency of the derived allele
(require an outgroup species)



Do you observe deviations from these expected patterns?

If you have gene models corresponding to your assembly, you can of course more precisely look at the number of synonymous and non-synonymous SNPs.

Take home messages

Summary statistics of the assembly are quite informative.

Base content and GC content variation along the genome (for species that can reproduce sexually) is informative about recombination

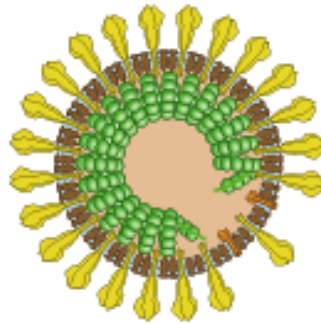
Simple summary statistics of the vcf, such as the proportion of missing data are important to consider, as well as site frequency spectra (SFS)

PCA and phylogenetic trees can help to diagnose errors (misidentification, clones, ...)

SNP density and ts/tv ratios are helpful to identify regions with a lot of false positive calls

By spending a couple of days to check this from the very start of the project, you can save a lot of time later (by avoiding some steps backwards)

Practical:
**analysis of the SARS-CoV2 reference
genome and 99 other virus sequences**



General information

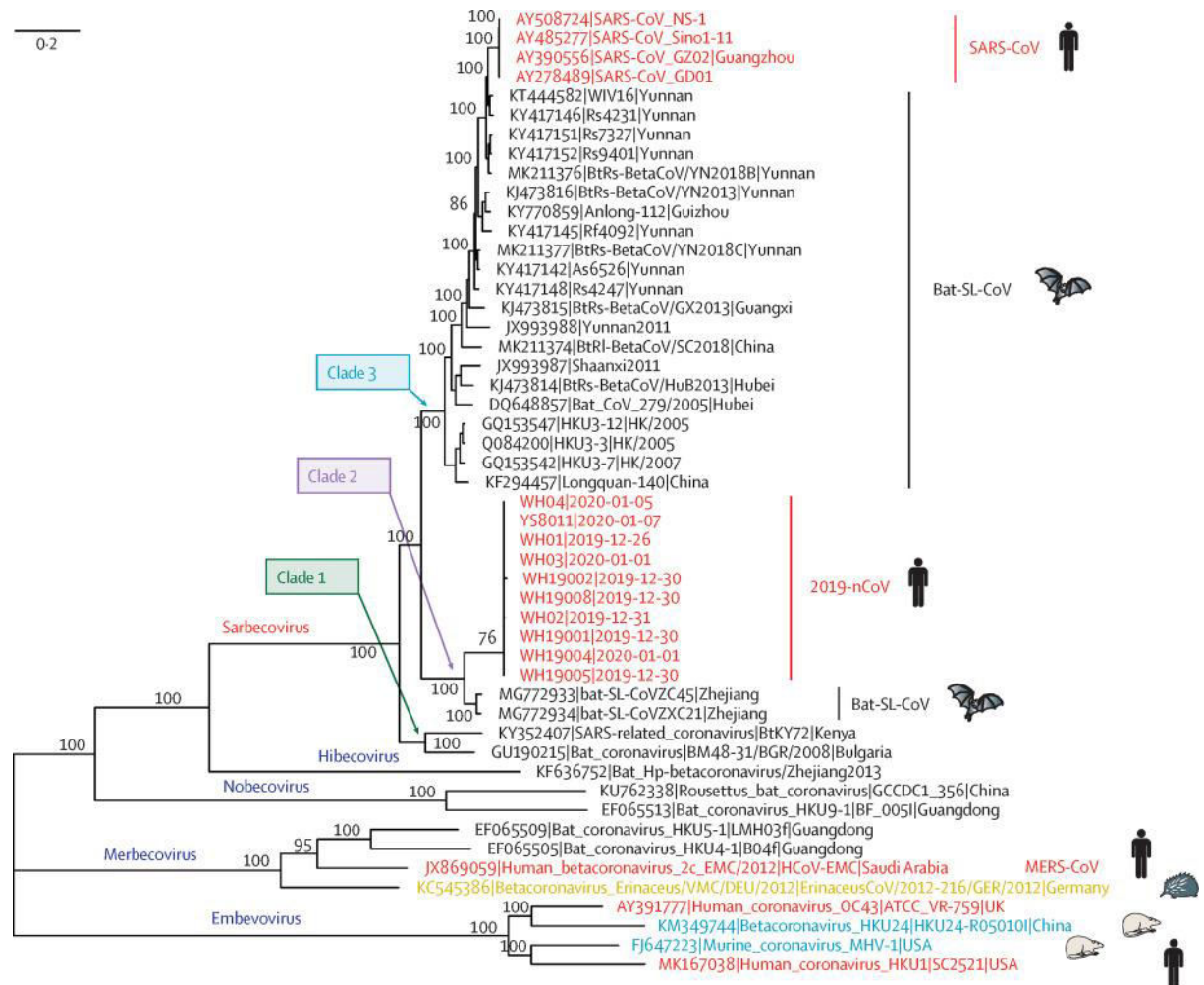
SARS-CoV-2 : Severe acute respiratory syndrome coronavirus 2 is the virus strain that causes coronavirus disease 2019 (COVID-19).

Responsible for > 160,000 deaths to date (April, 19th)

SARS-CoV-2 belongs to the broad family of coronaviruses.

89.1% nucleotide similarity) to a group of SARS-like coronaviruses (genus Betacoronavirus, subgenus Sarbecovirus)

As for other coronavirus infecting humans, zoonotic origin (bats and putatively pangolins as intermediate host between bats and humans)



Genome available

It is a positive-sense single-stranded RNA virus. It means that the RNA can directly serve as messenger RNA and can be translated into protein in the host cell.

Its RNA sequence is approximately 30,000 bases in length (29,903 for the reference genome).

Genome sequenced and released very fast :

Patient: 26 December 2019
(a 41-year-old man, worker at the seafood Wuhan market)


Published in nature: 03 February 2020

A lot of genomes were now sequenced (>4400 available on Nextstrain)

nature

Article | [Open Access](#) | Published: 03 February 2020

A new coronavirus associated with human respiratory disease in China

Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, Ming-Li Yuan, Yu-Ling Zhang, Fa-Hui Dai, Yi Liu, Qi-Min Wang, Jiao-Jiao Zheng, Lin Xu, Edward C. Holmes & Yong-Zhen Zhang 

Nature **579**, 265–269(2020) | [Cite this article](#)

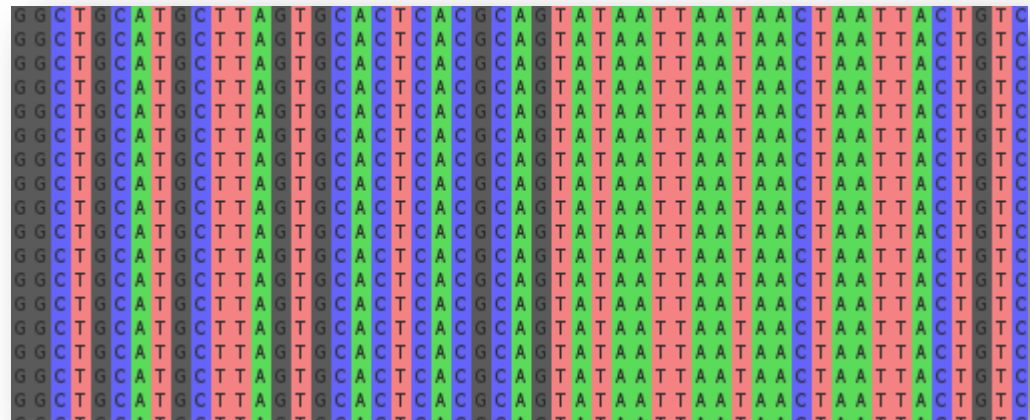
Dataset to be used today

We will only work on a subset of 100 genomes. The data for these genomes are publicly available on SRA.

The reference genome from the first patient in Wuhan is included among the 100 genomes.

Given the length of the genome, I didn't use a read mapping strategy, but a whole-genome alignment strategy. You will already work on these aligned genomes.

Now you just need to try to do the practical, and to get some information from your analyses.



Remember to save your work frequently (Rcode, plots, ...)!

Lost? Stuck? You can send me an email throughout the day! (thibault.leroy@univie.ac.at)

