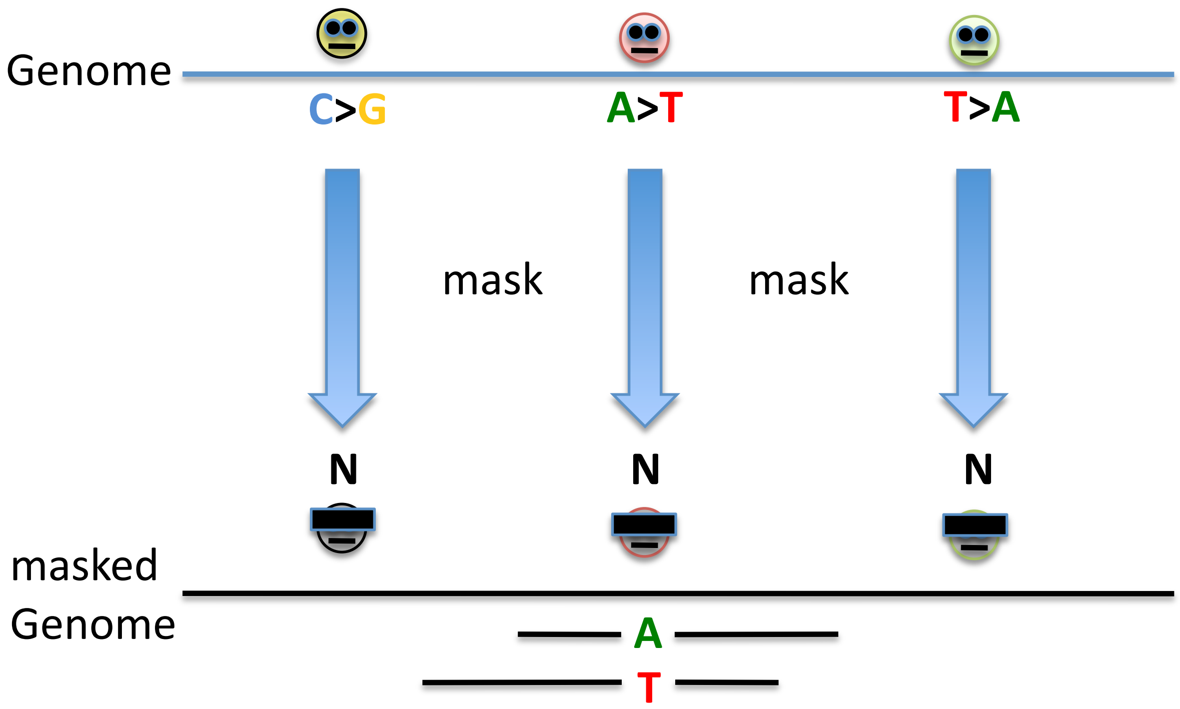
# Geuvadis TC 12/01/12:
## preliminary mapping stats
## of the Geneva data

M.Sammeth, Functional Bioinformatics
Centre d'Analisi Genòmica (CNAG)
Barcelona, Spain

# Traditional Mapping Strategy for Haplotypes

known SNP positions (e.g., allele freq > 5%)



Genome

C>G A>T T>A

mask mask

N N N

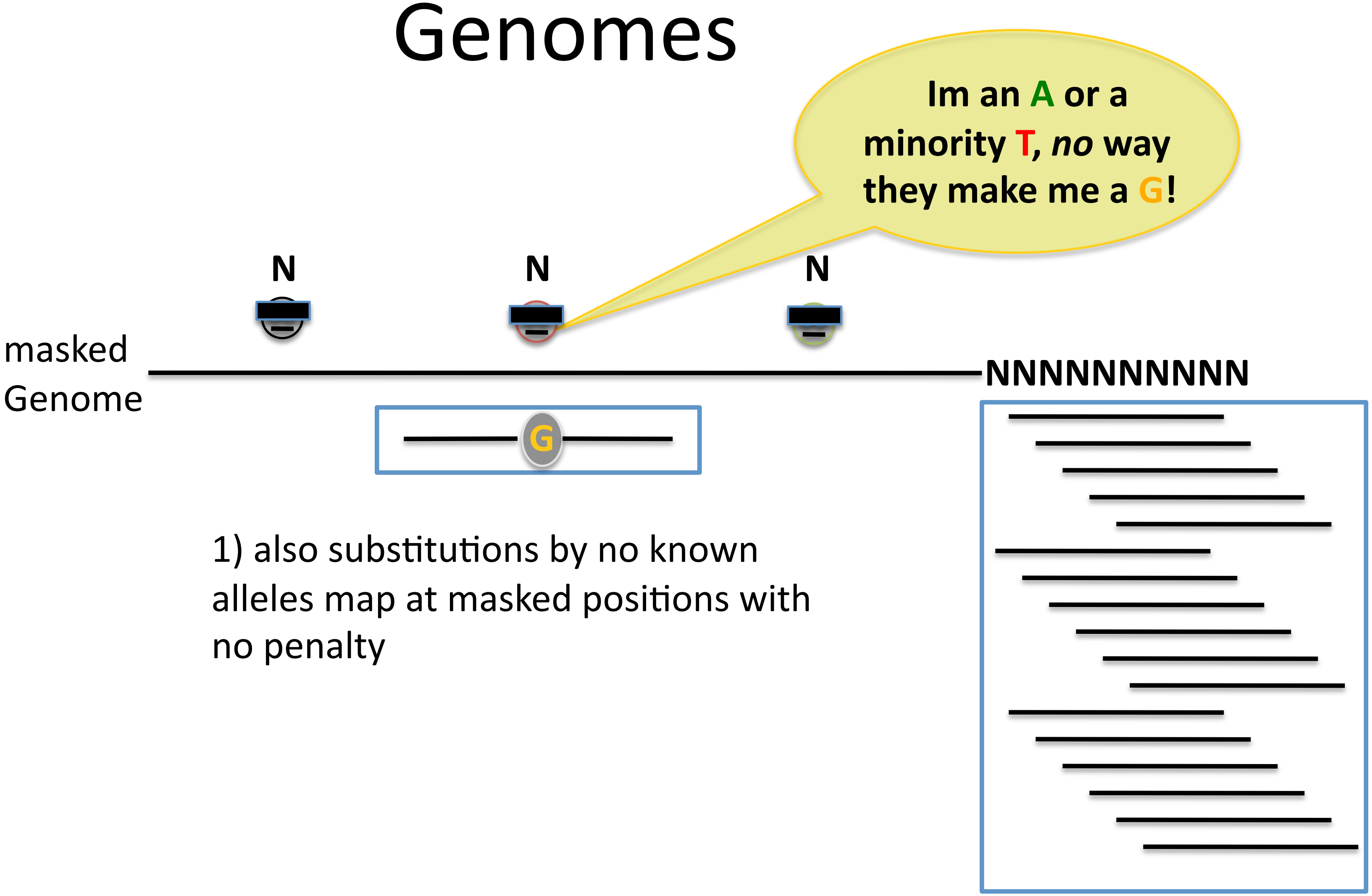
masked Genome

A

T

use mapper to ignore “N” masked position and accept mismatches there at no cost

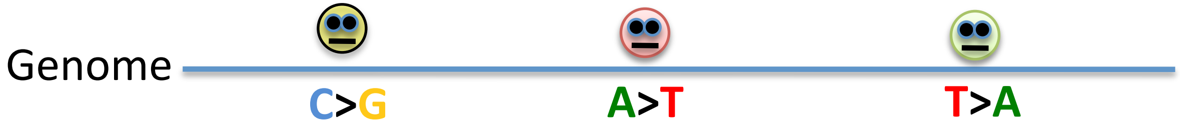
# Problems with mapping to masked Genomes



1) also substitutions by no known alleles map at masked positions with no penalty

2) huge read stacks can be expected in regions with assembly problems

# Improved Strategies of Haplotype Mapping (1)

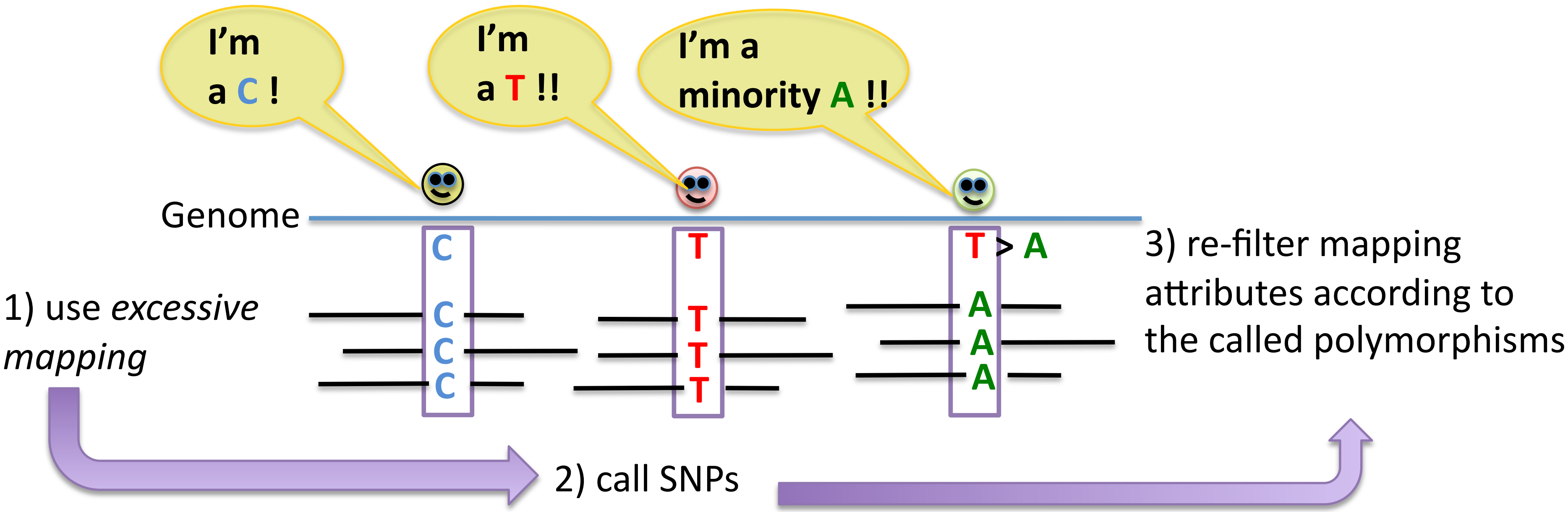


use *ambiguity codes*  **S**  **W**  **W**

- alleviates the problem but does not remove it: still 2 possibilities for mappings at every position (increase in mapping variation)

- relies on *a priori* knowledge, misses cases that are not captured by the set of masked positions

# Improved Strategies of Haplotype Mapping (2)



1) use *excessive mapping*

2) call SNPs

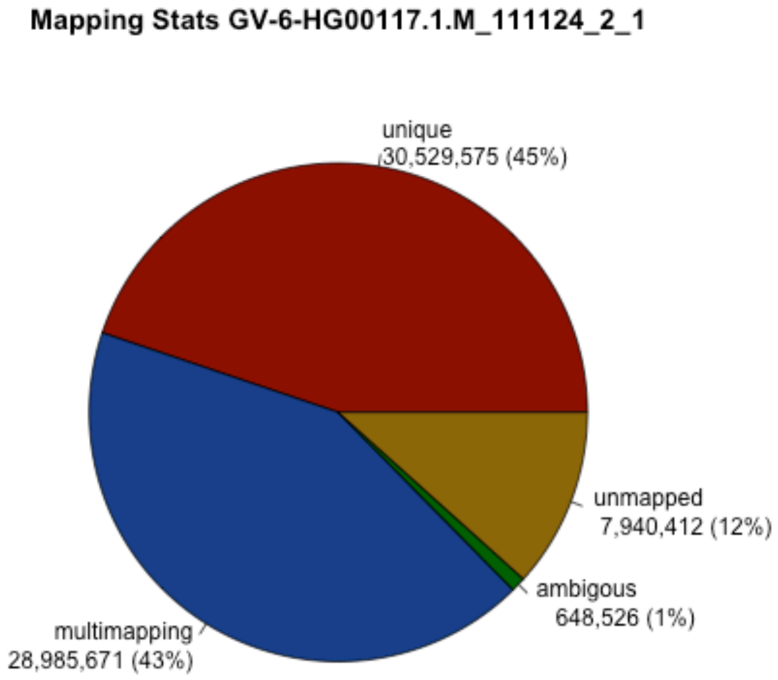
3) re-filter mapping attributes according to the called polymorphisms

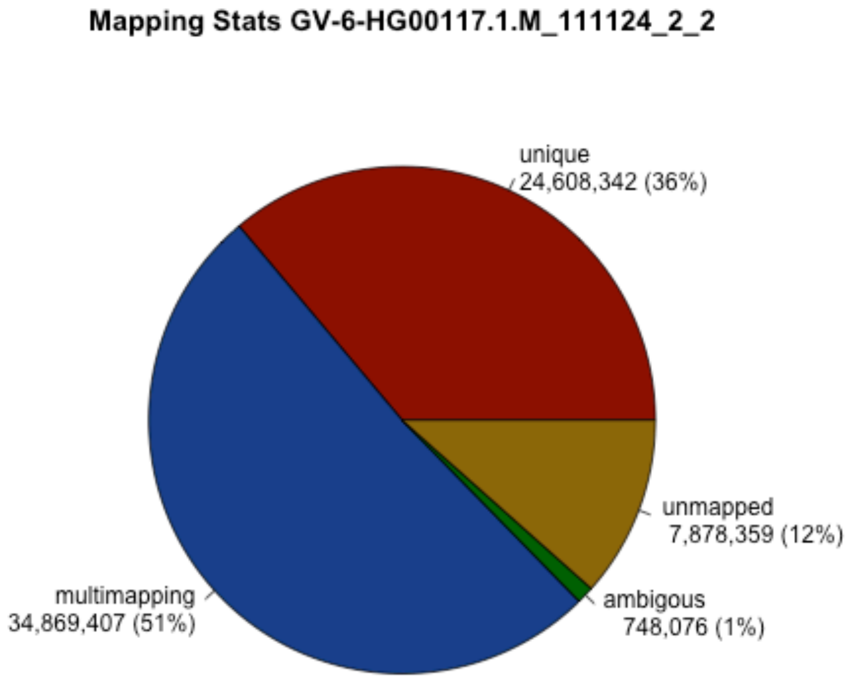
- requires a powerful mapping strategy to find all mappings beyond the usual limit (e.g., 2 mismatches)

- relies on sufficiently high coverage for statistical significance of SNP calls

+ discovers corresponding polymorphisms as observed in the data

# Mapping Statistics Sandbox Data





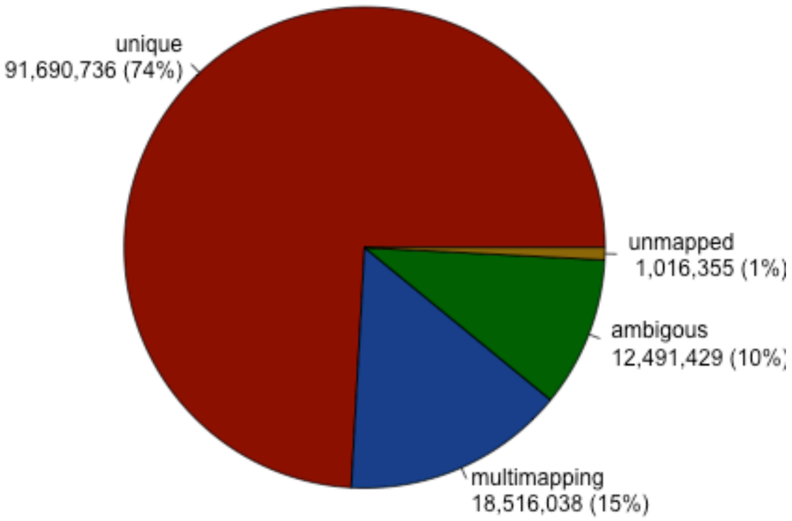
*excessive mapping*:

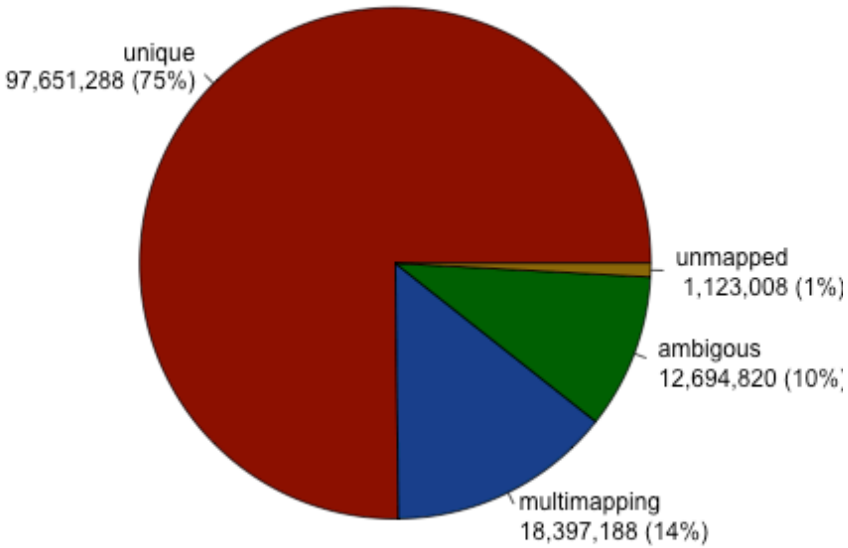
<=5 mismatches in good quality bases
split-mappings <u>not</u> contained in statistics
no read-end trimming

*first observations*:

- *high* percentage of multi-mappings (excessive)
- *low* percentage of ambigous reads (excessive)
- *too high* proportion of unmapped reads not solely explainable by split-mappings

# Mapping Statistics Standard Mapping (2 mismatches, quality, split-maps)

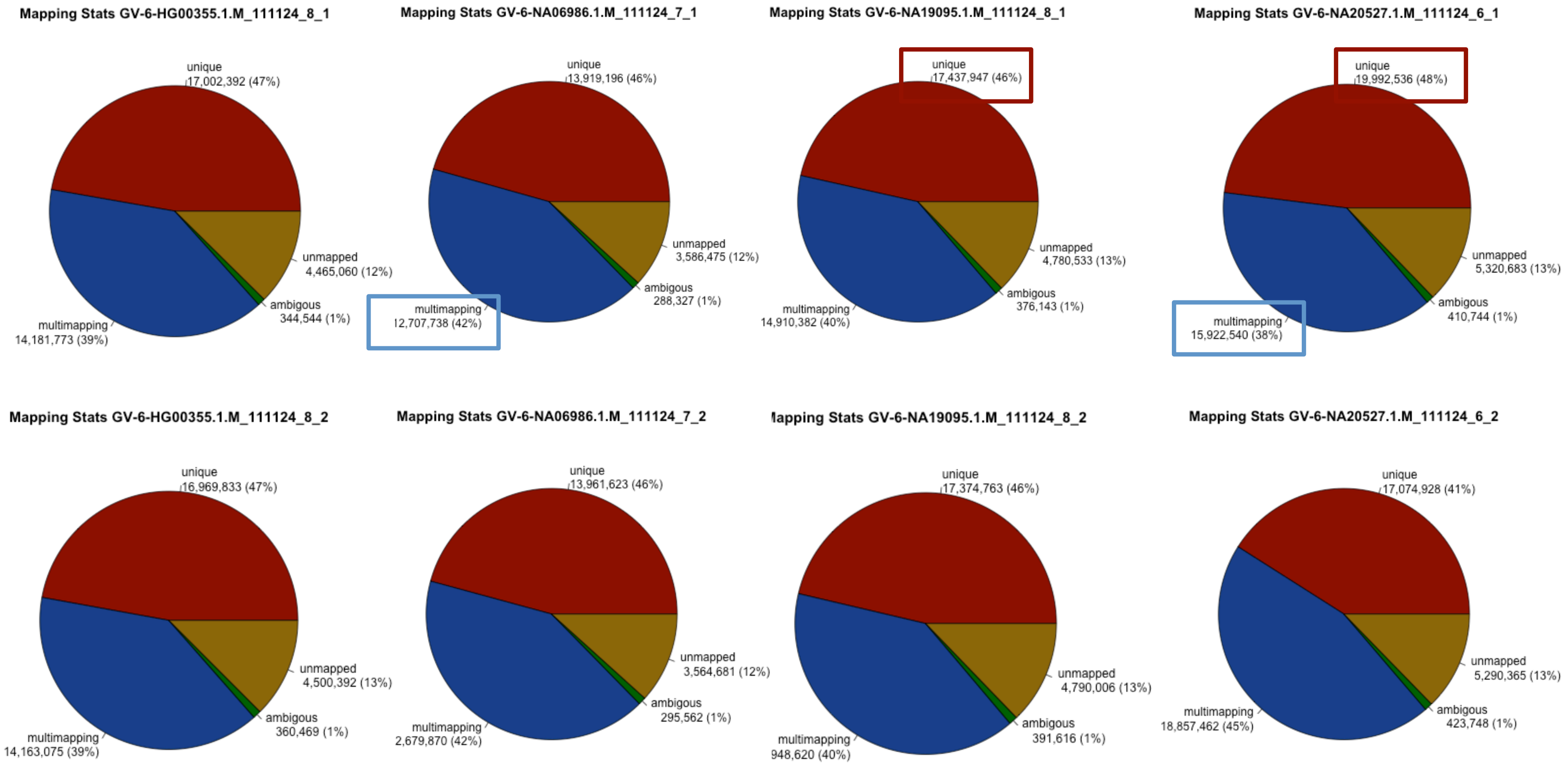




*reference datasets (75nt reads, human, not from Geuvadis)*:

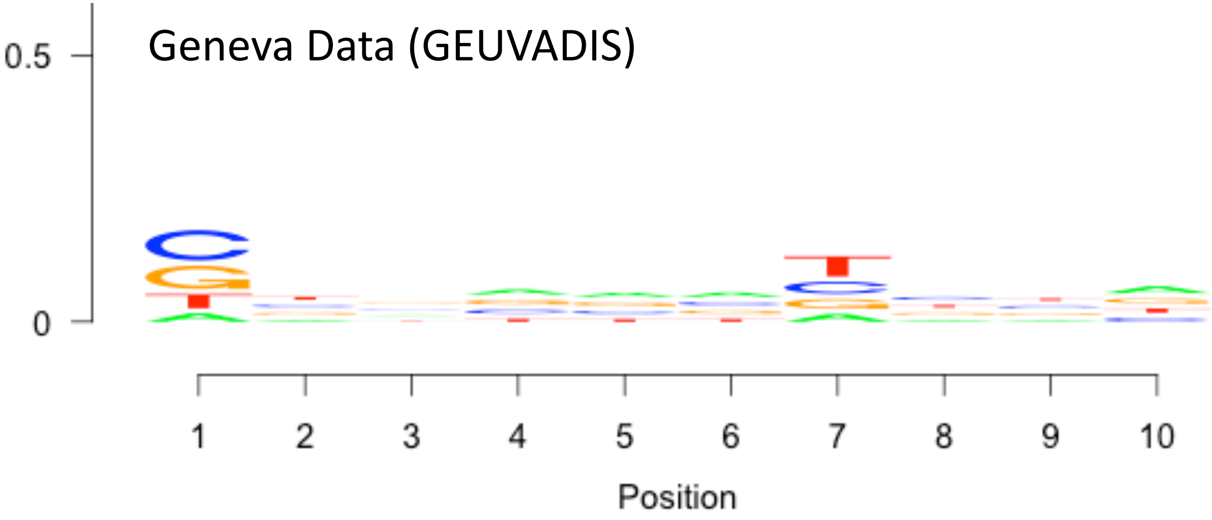
<=2 mismatches in good quality bases
split-mappings *contained* in statistics (5-6%, not 12%)
both reads merged

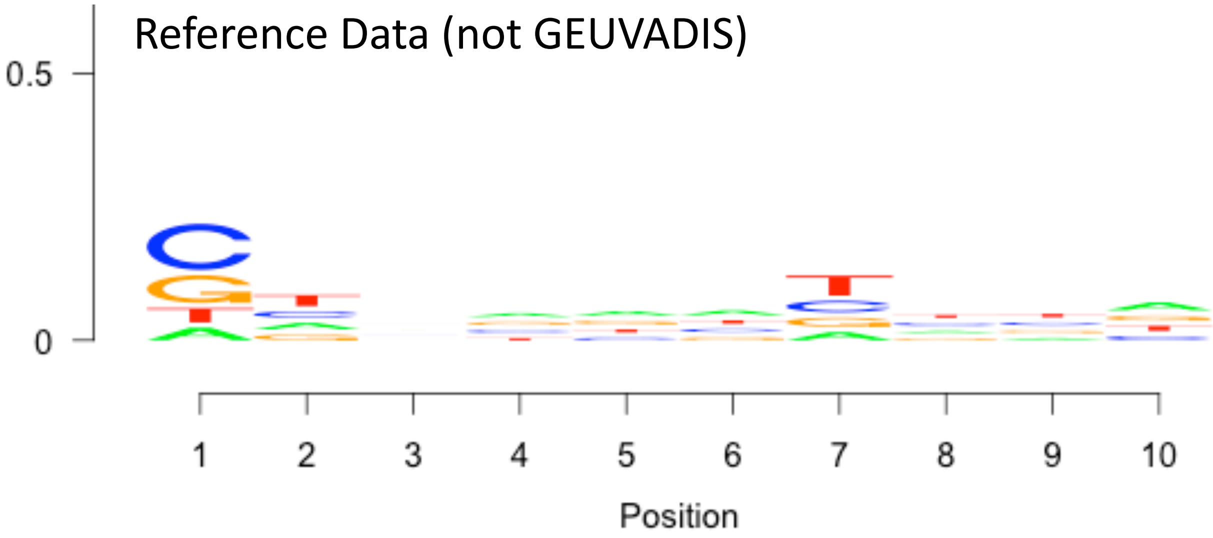
# Mapping Statistics Sandbox Data



overall similar trends, but there are differences between the datasets..

# Sequence Biases..


Geneva Data (GEUVADIS)


Reference Data (not GEUVADIS)

.. as expected.

End.