
INSTITUTE OF MARINE RESEARCH
HAVFORSKNINGSINSTITUTTET

Identifying diagnostic SNPs in the presence of sequencing errors

Ketil Malde, Francois Besnier, Kevin Glover

INSTITUTE OF MARINE RESEARCH
HAVFORSKNINGSINSTITUTTET

The Challenge

Given two populations, P1 and P2, identify genetic variations that are *diagnostic*

(i.e. find SNPs that can be used to identify individuals as belonging to P1 or P2)

E.g.

P1 = farmed salmon

P2 = wild salmon

Standard procedure?

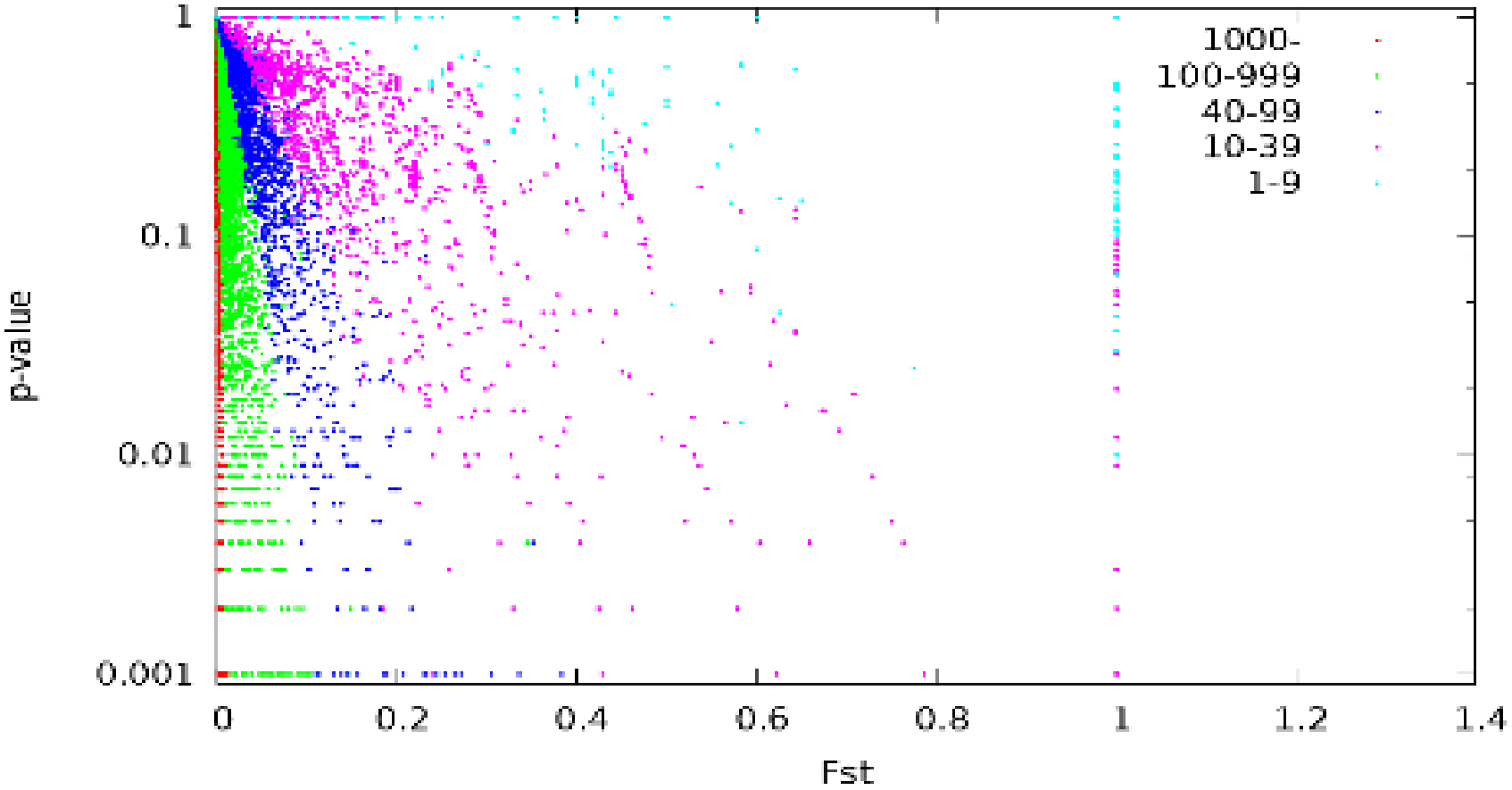
1. sequence pools from P1 and P2
2. map reads to reference genome
3. identify variant positions
4. calculate metric (Fst, significance, ...)
5. rank positions by metric
6. validate SNPs experimentally

Standard procedure

Unfortunately: it doesn't seem to work all that well.

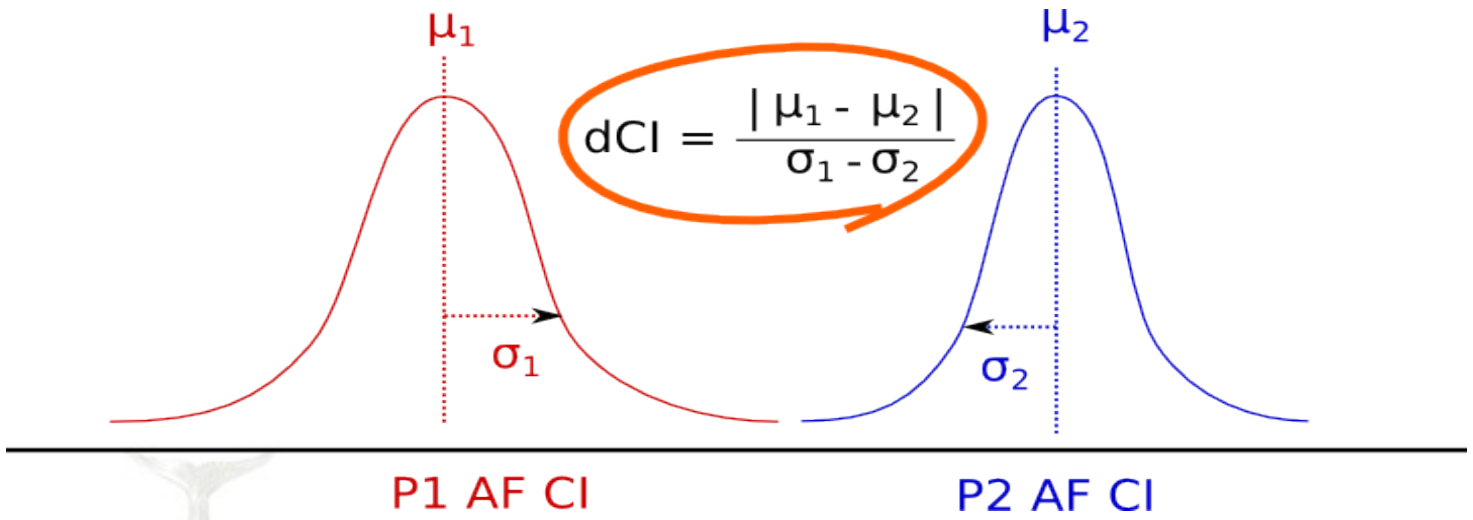
WHY?

Fst and significance by coverage



An alternative metric

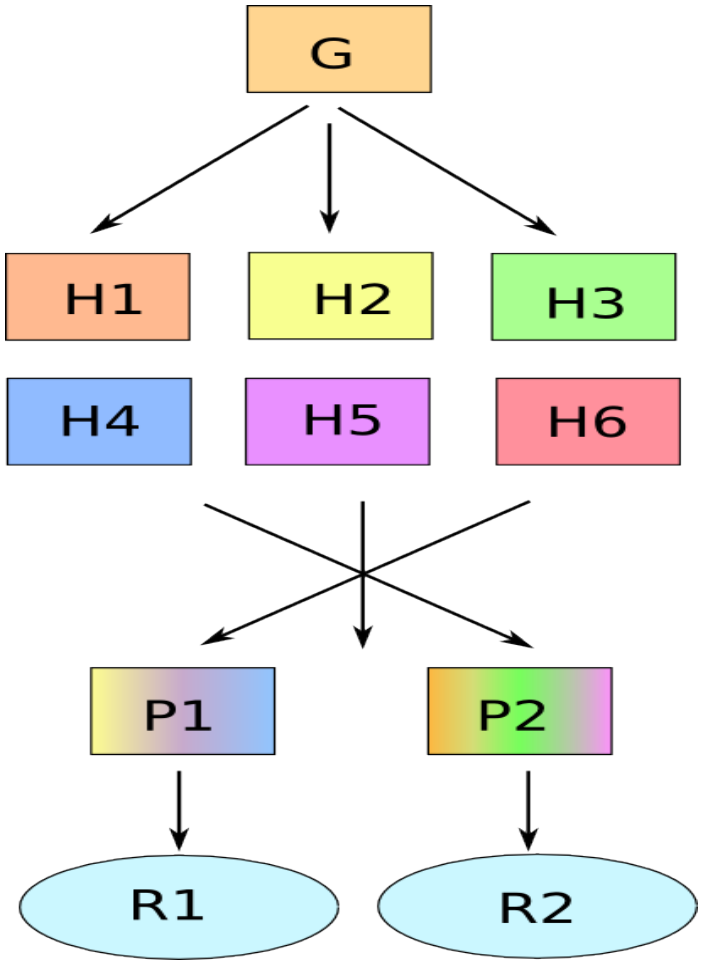
Use probability distribution for real allele frequence (based on our sample)



Effect size (μ) *and* confidence (σ)

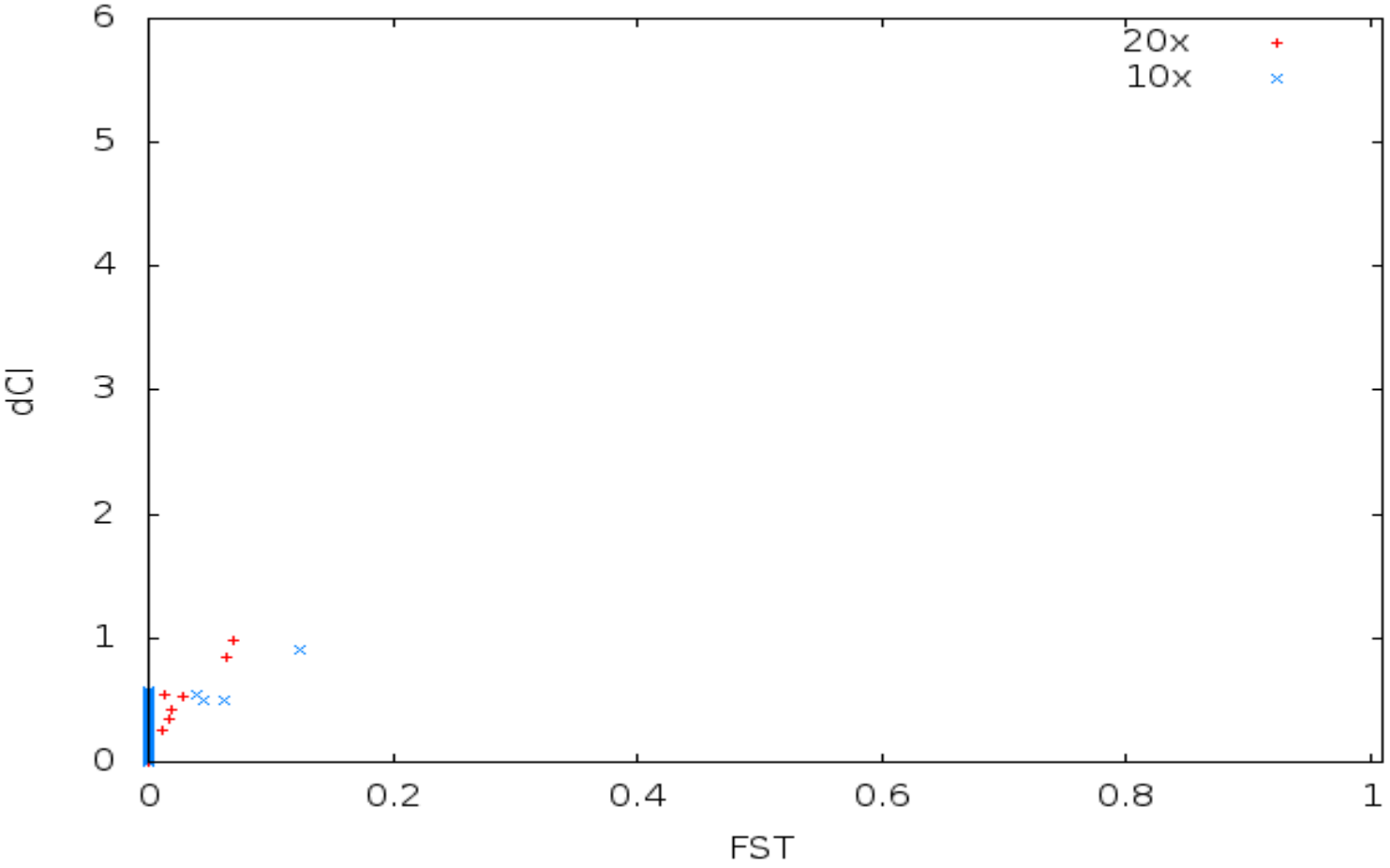
Simulation

- Genome: 10M chunk of *L.salmonis*
- 6 Haplotypes, ~37K polymorphic sites
- Mix to create two populations
- Generate simulated reads
- Mutate reads with various error rates



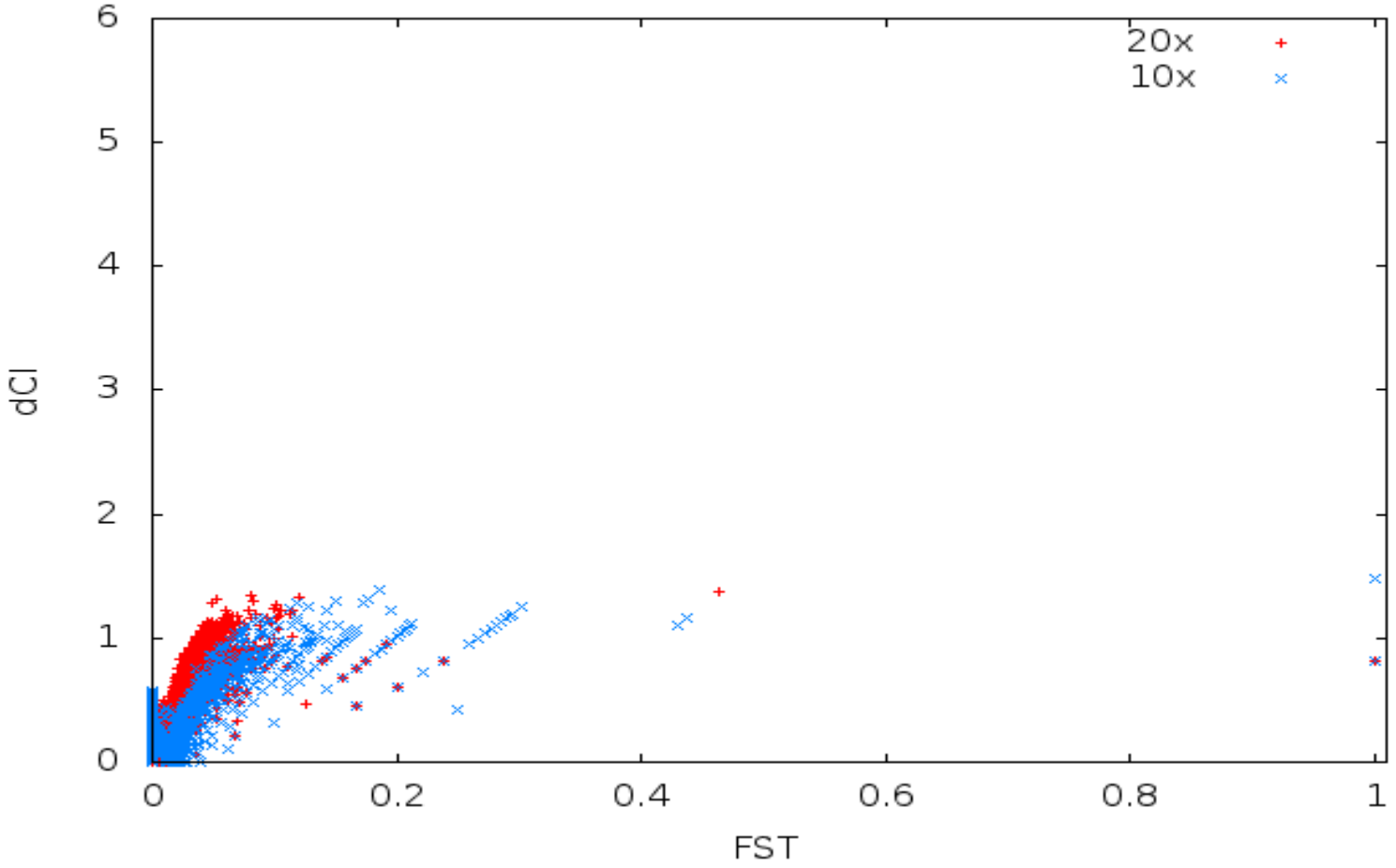
Results: Non-polymorphic sites

e=0%



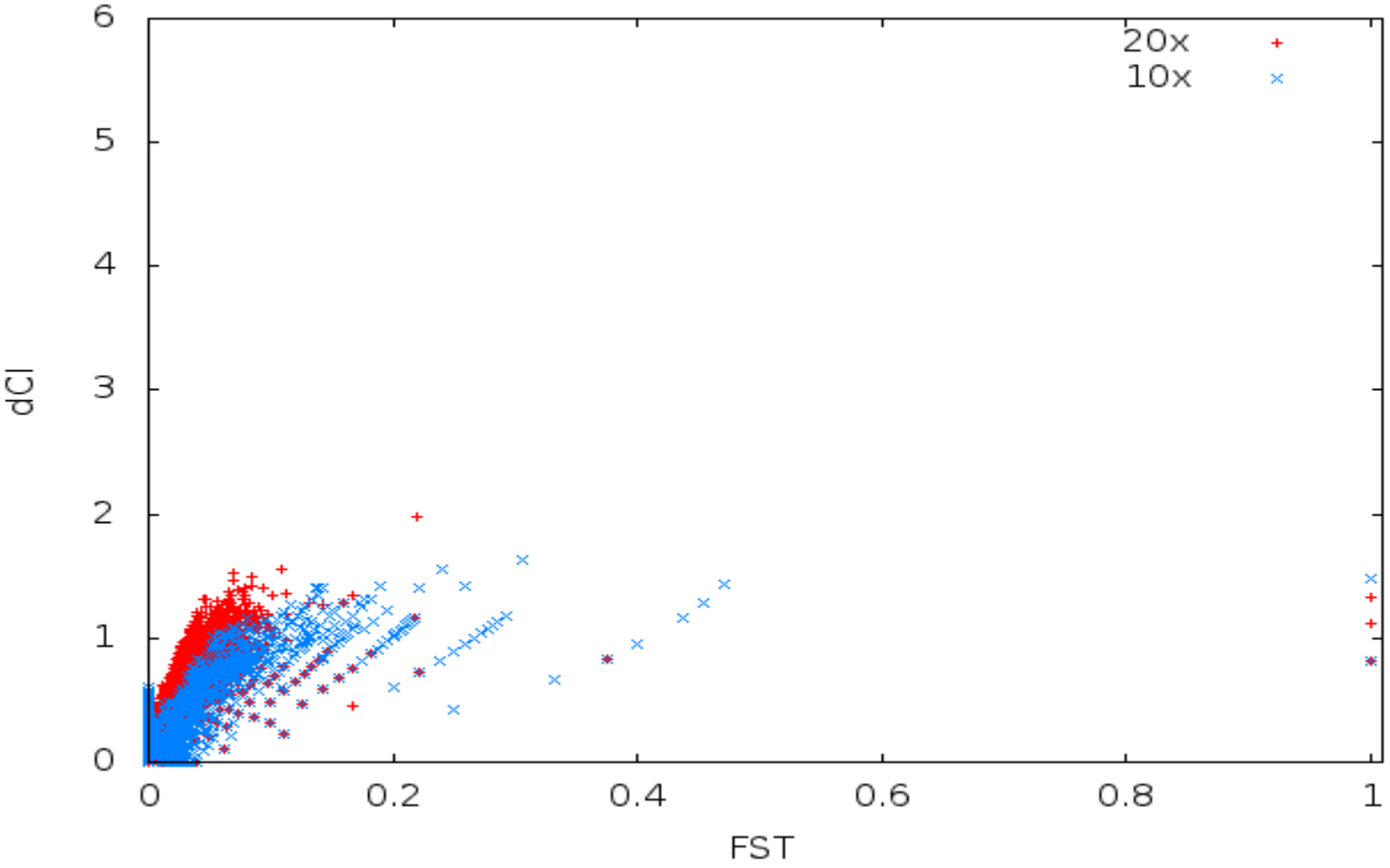
Results: Non-polymorphic sites

e=0.5%



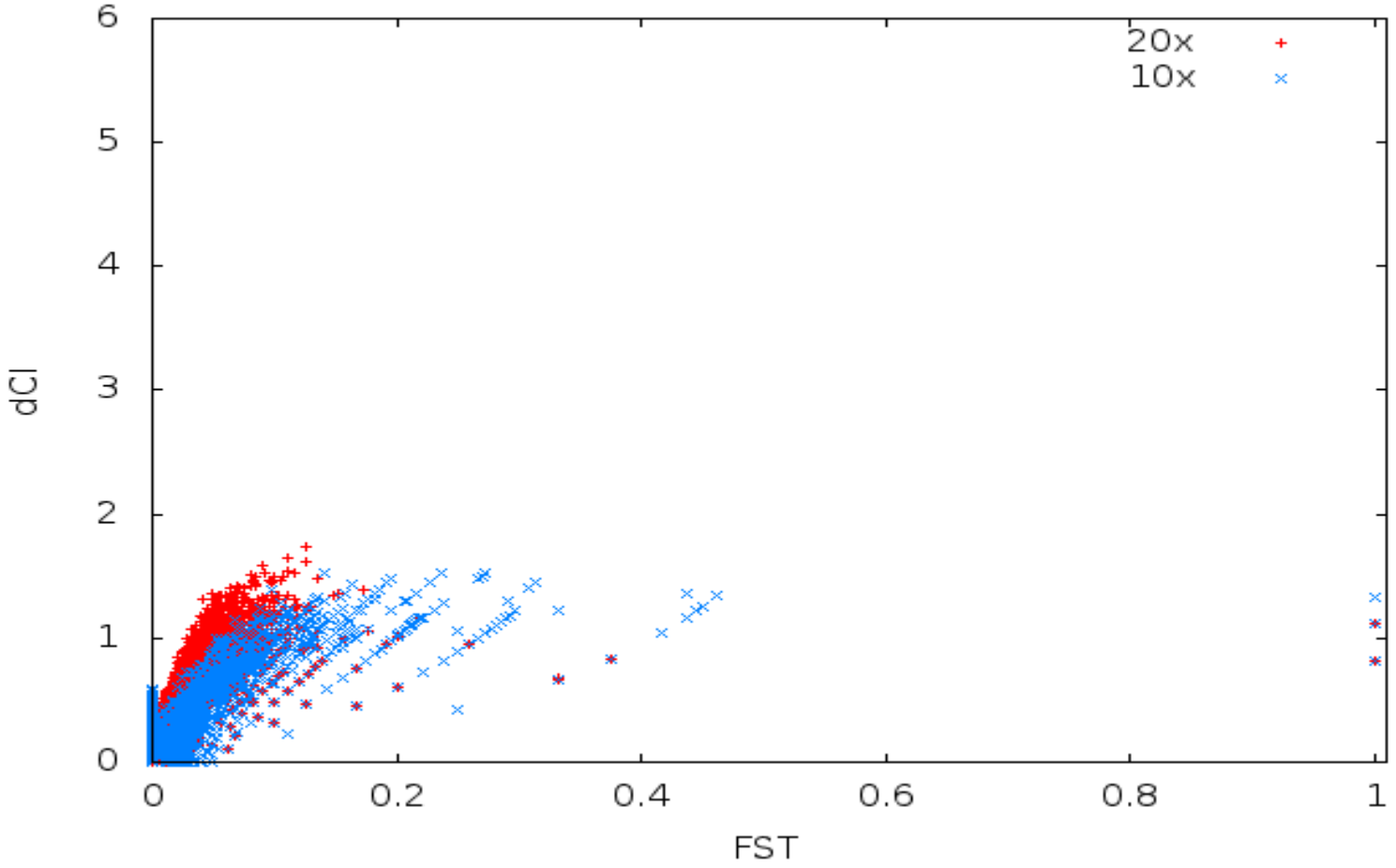
Results: Non-polymorphic sites

e=1.0%



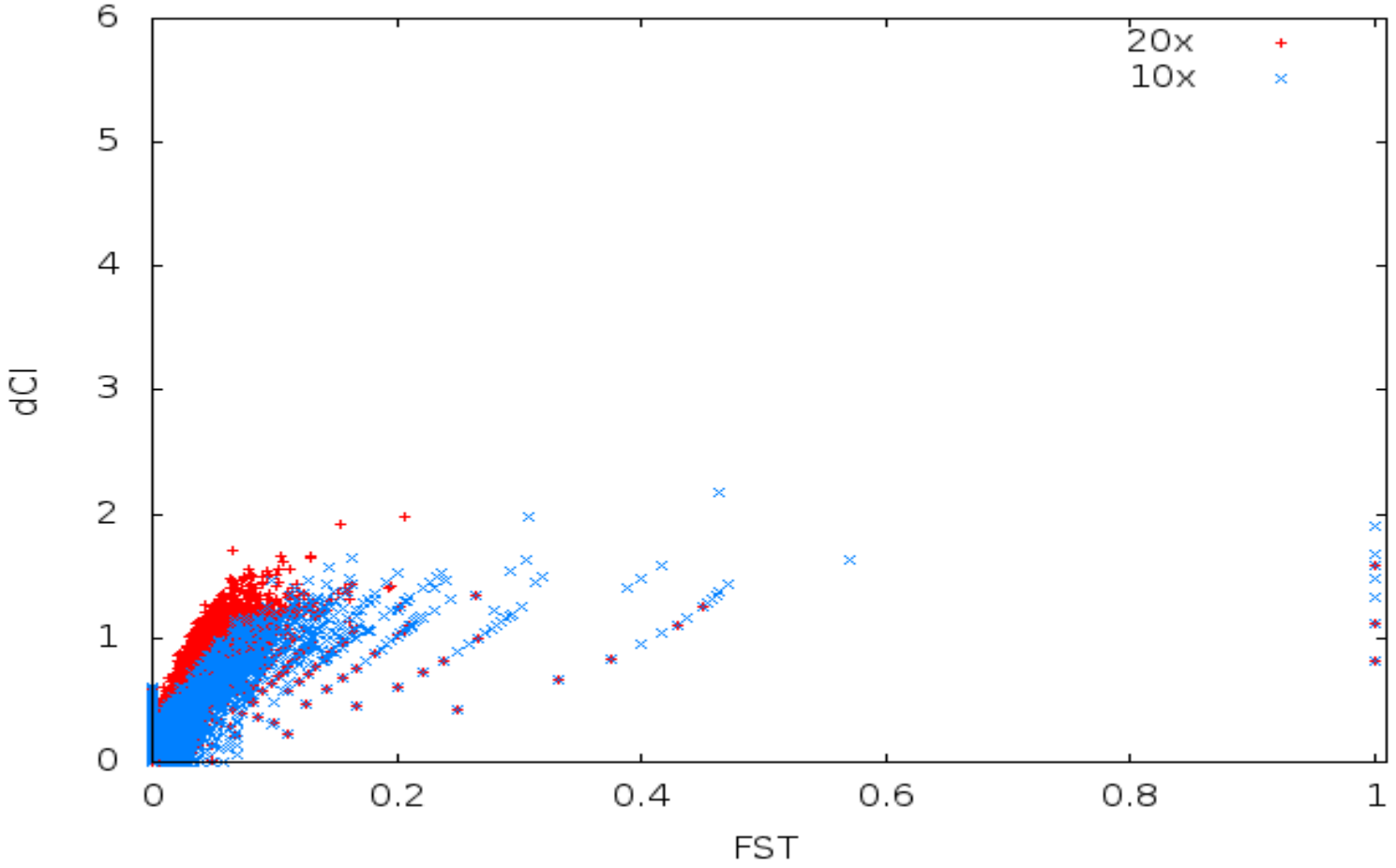
Results: Non-polymorphic sites

e=1.5%



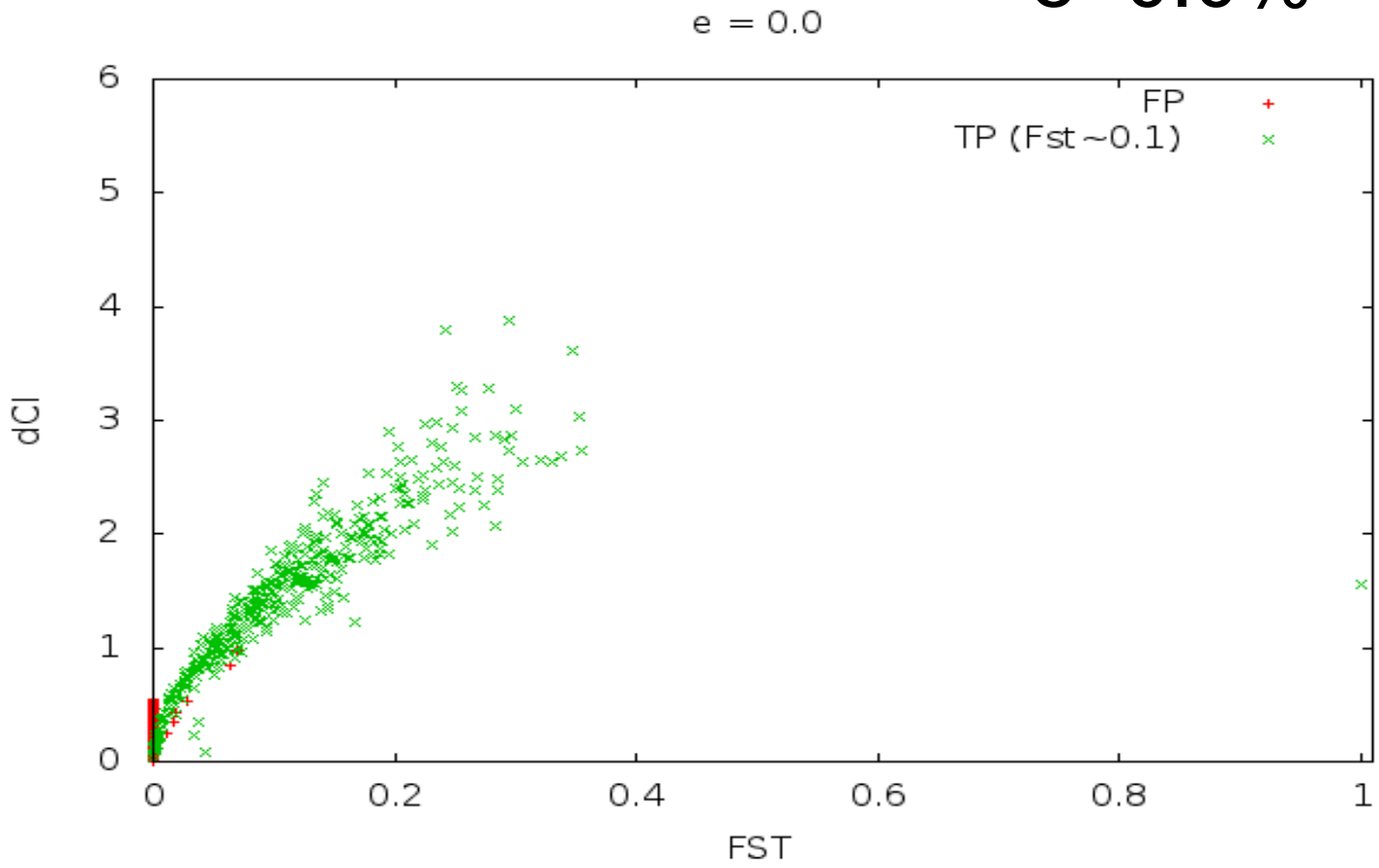
Results: Non-polymorphic sites

e=2.0%



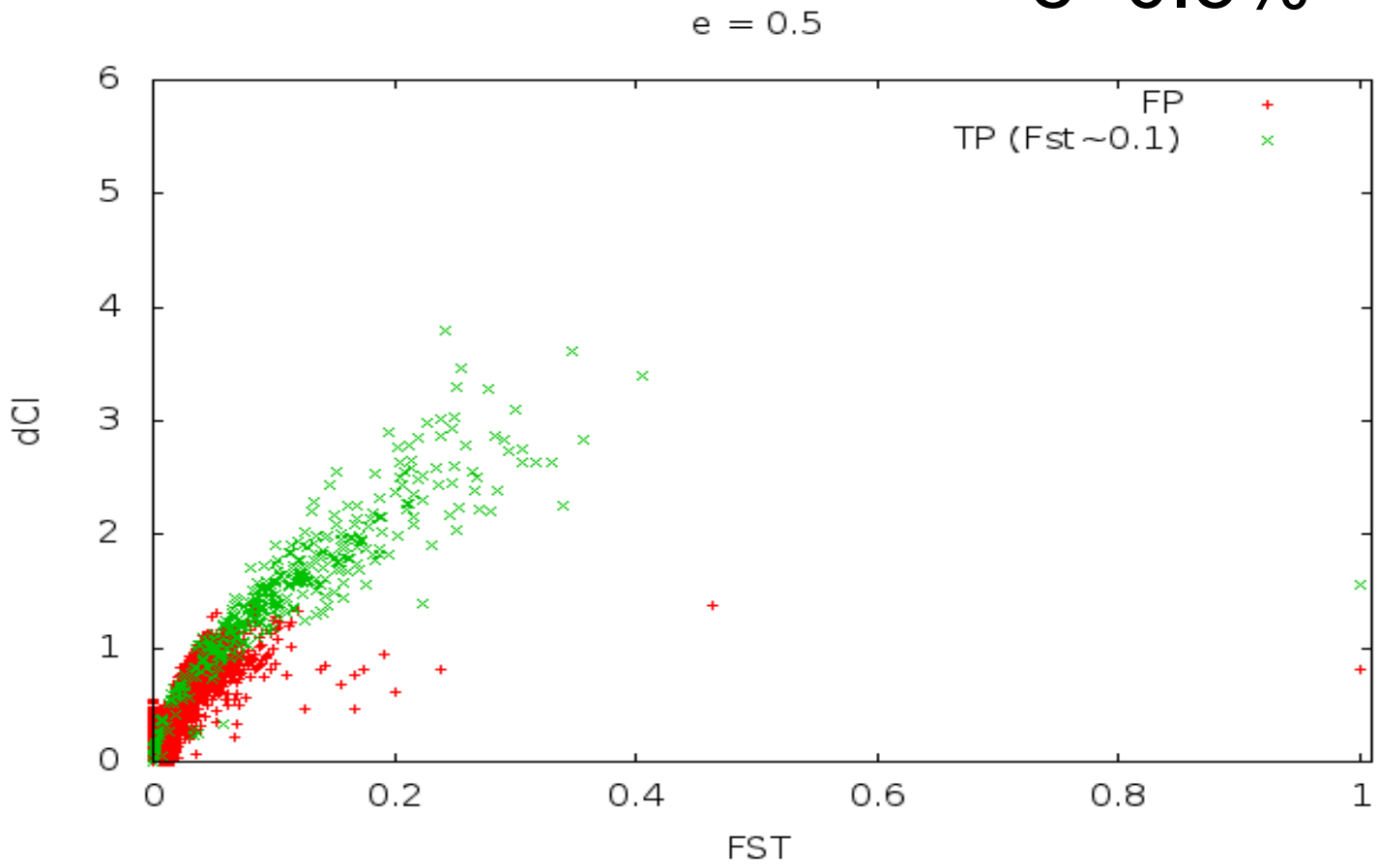
Polymorphic sites - 20x

e=0.0%



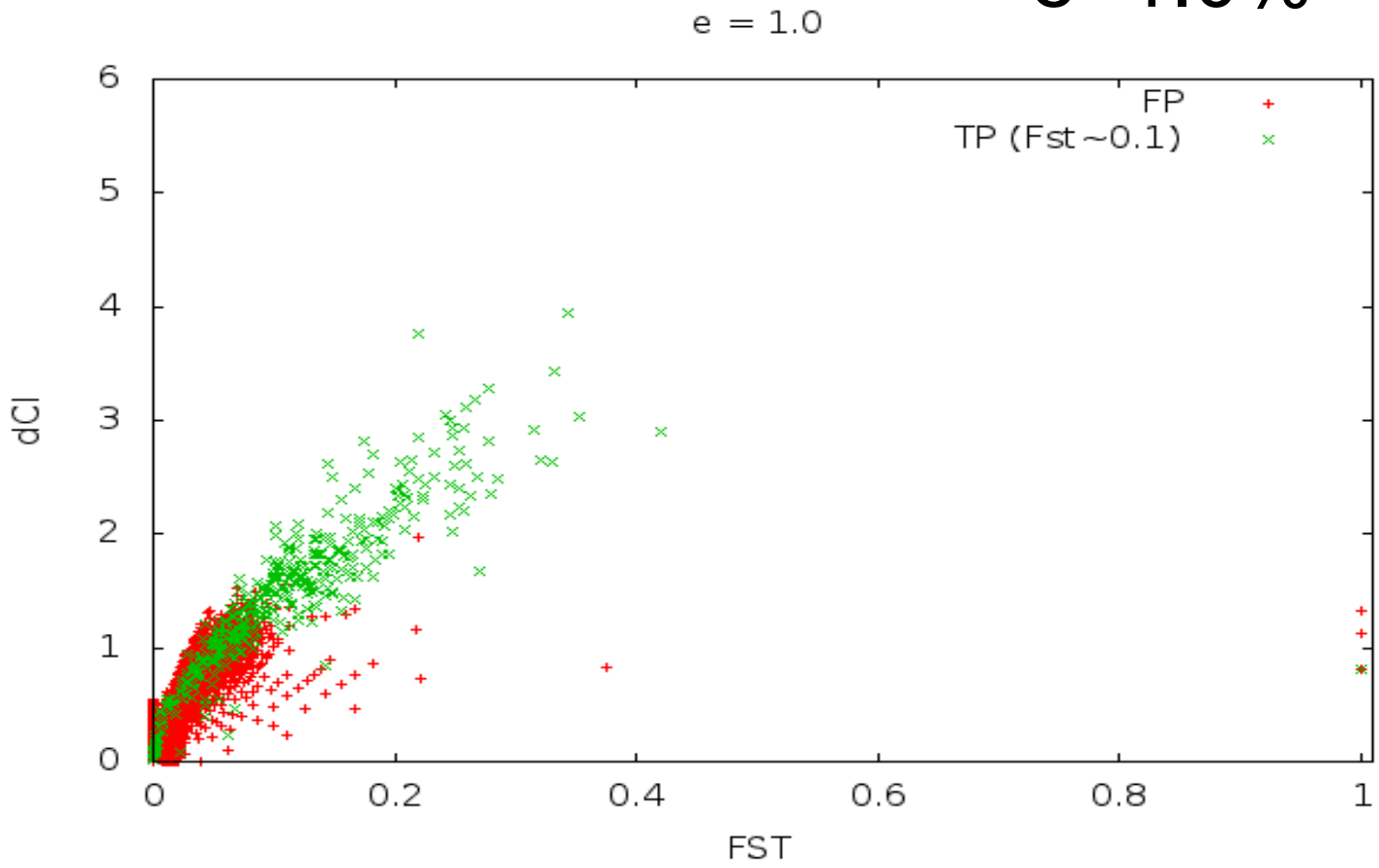
Polymorphic sites - 20x

e=0.5%



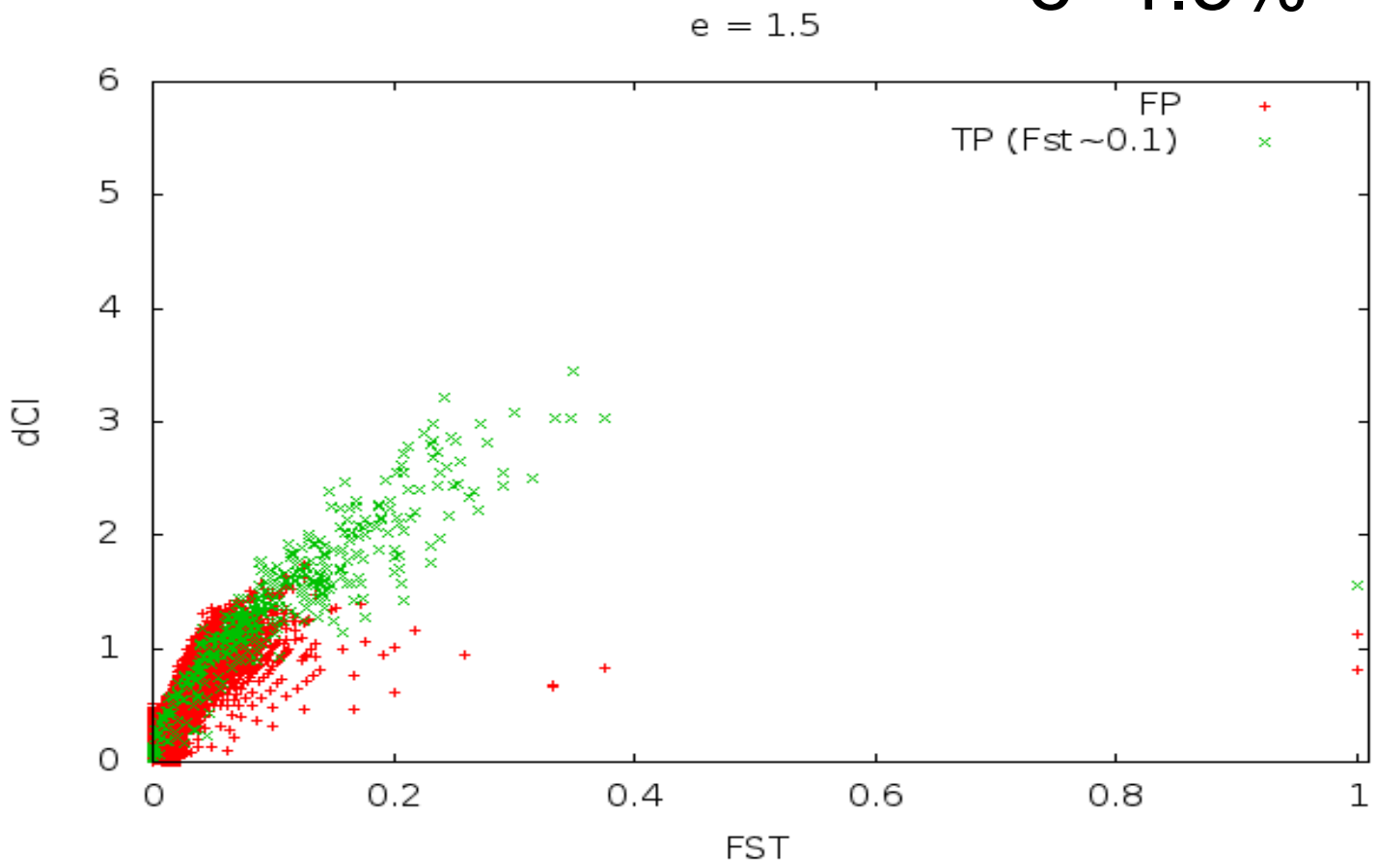
Polymorphic sites - 20x

e=1.0%



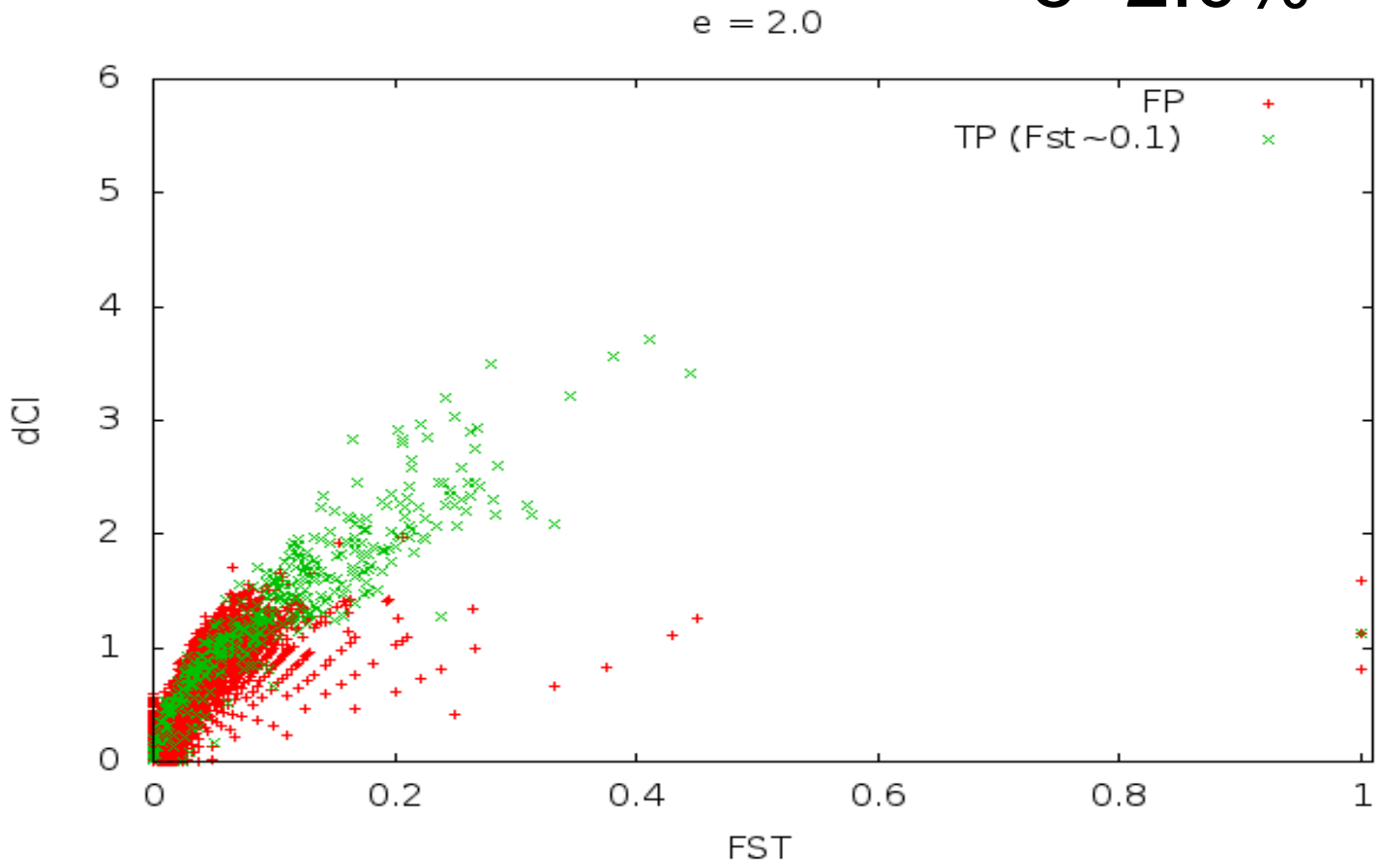
Polymorphic sites - 20x

e=1.5%



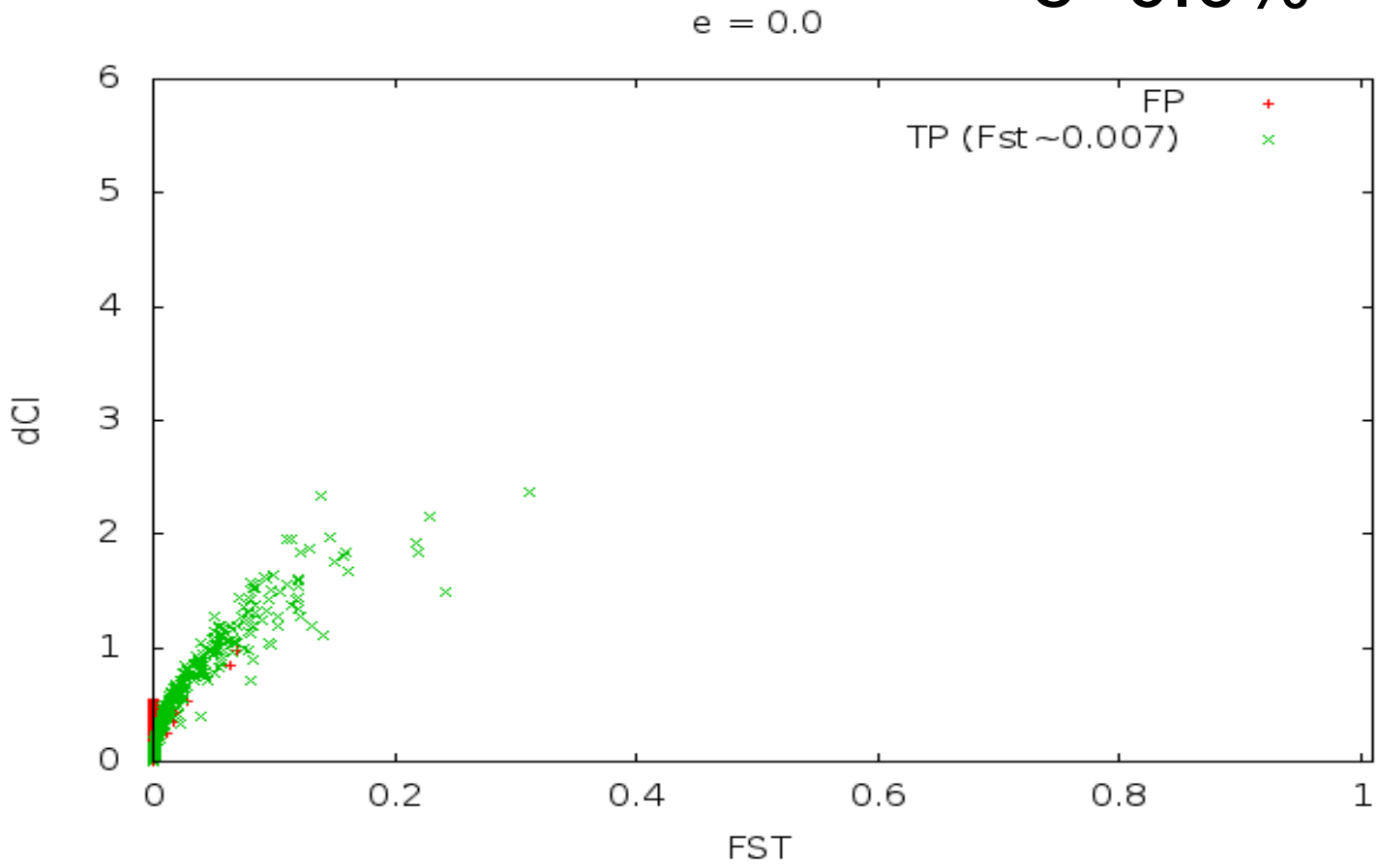
Polymorphic sites - 20x

e=2.0%



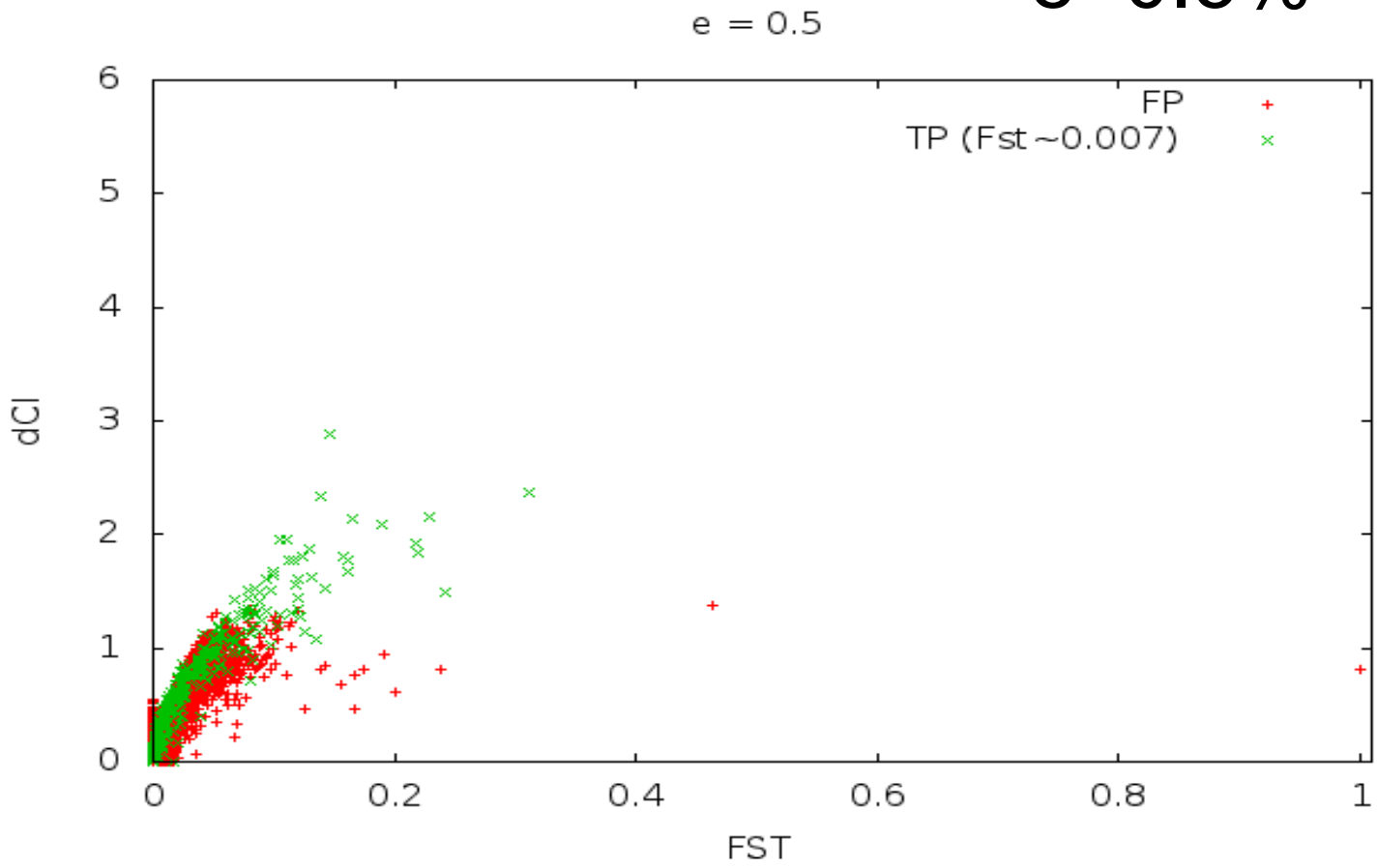
Polymorphic sites - 10x

e=0.0%



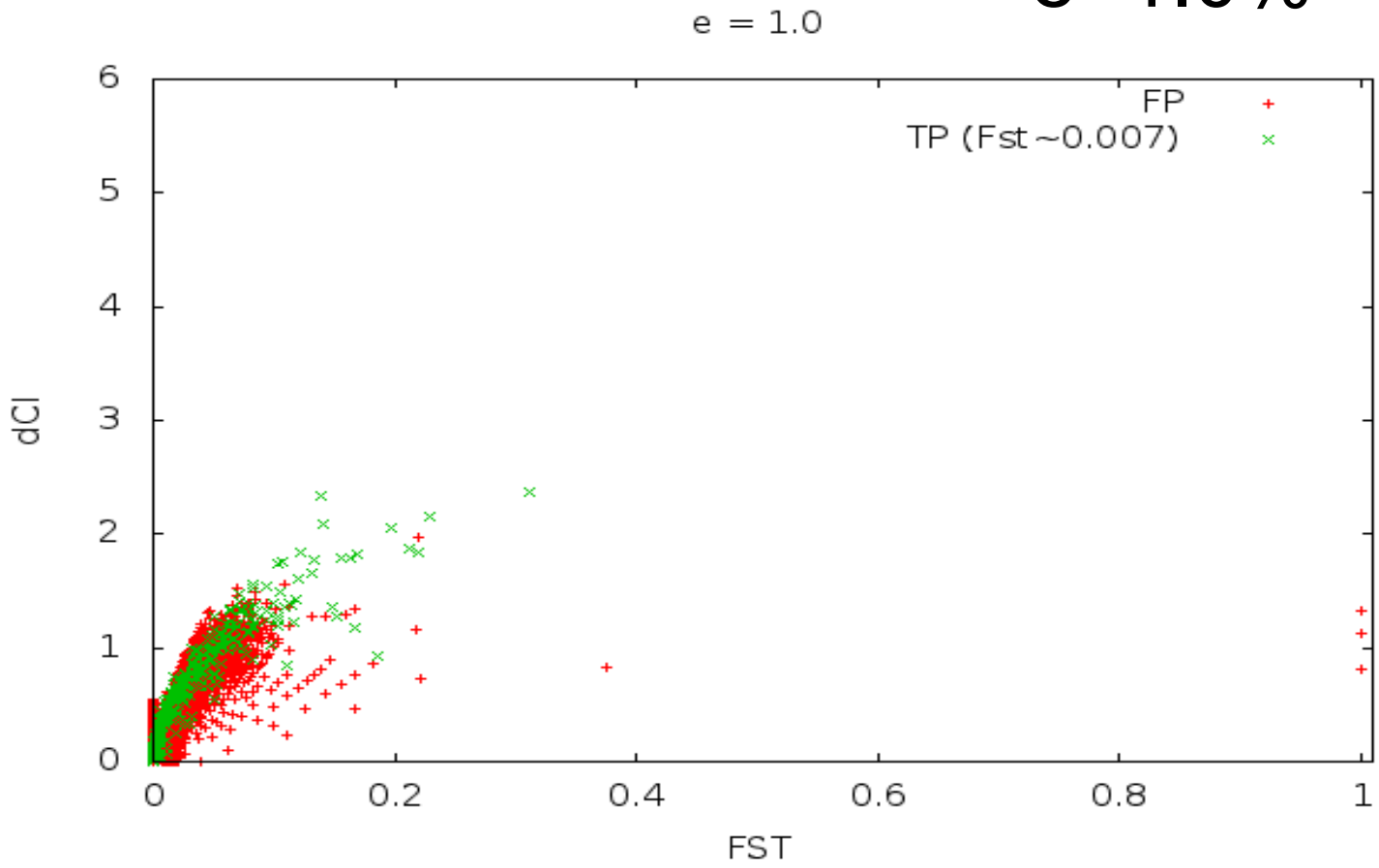
Polymorphic sites - 10x

e=0.5%



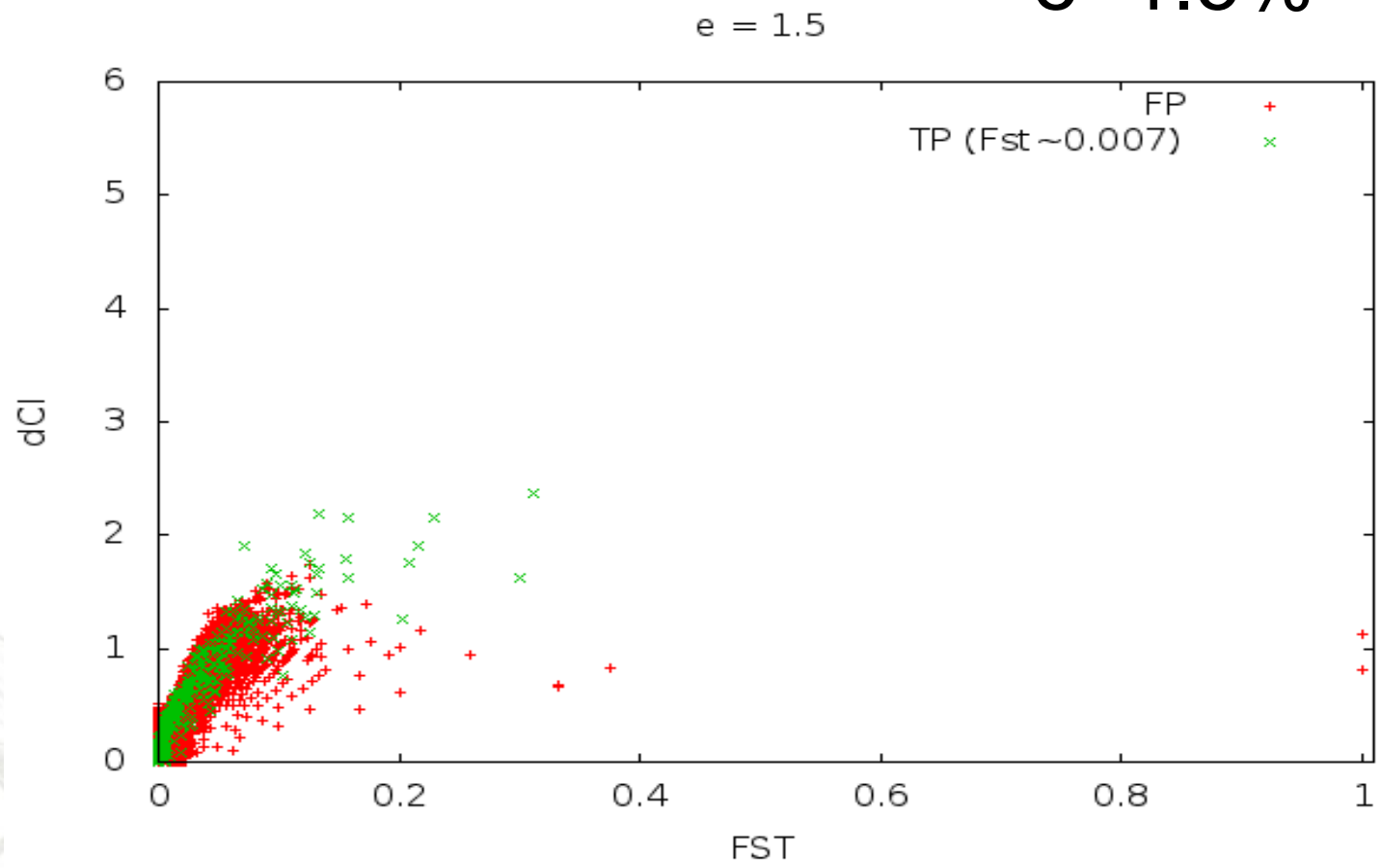
Polymorphic sites - 10x

e=1.0%



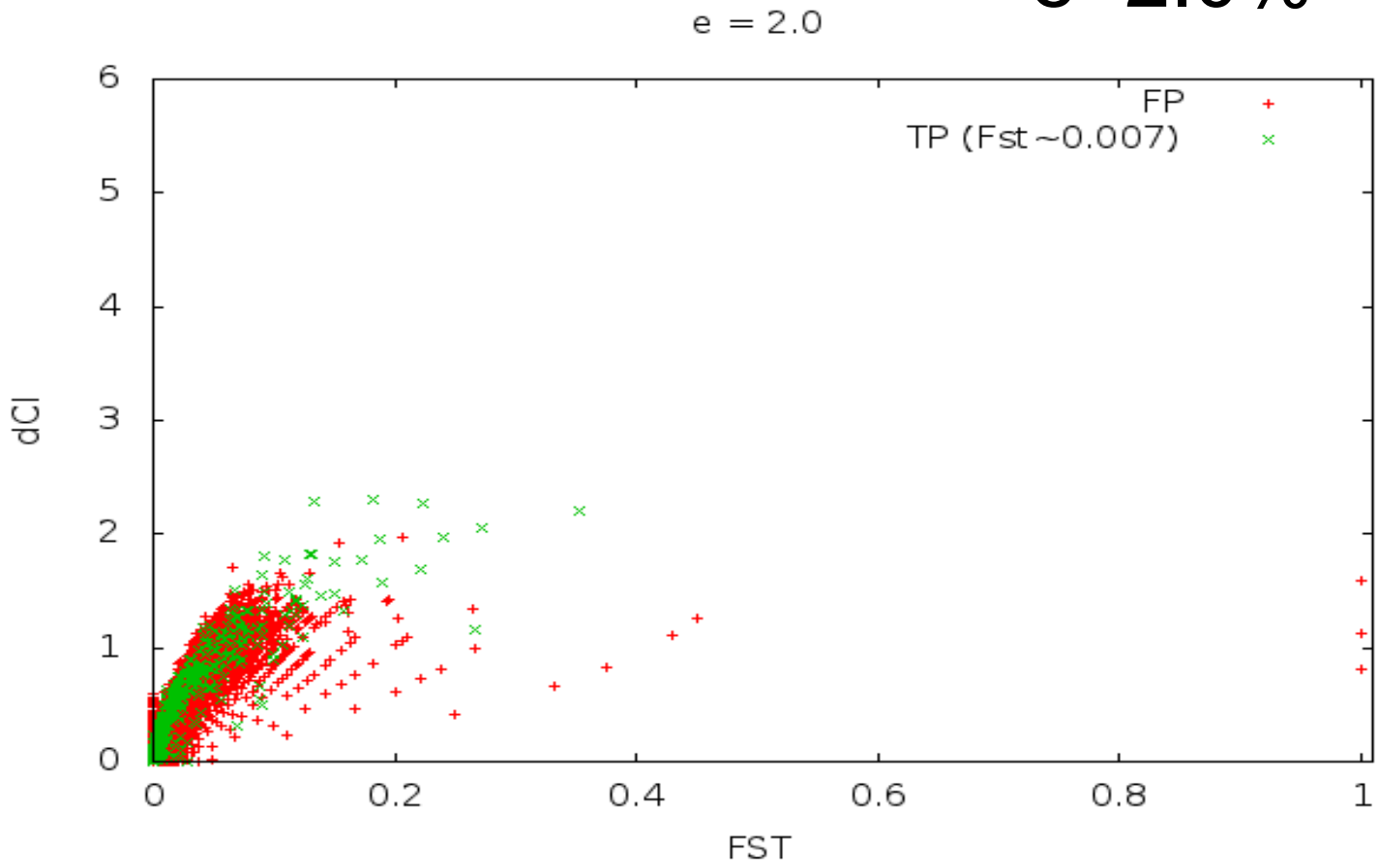
Polymorphic sites - 10x

e=1.5%



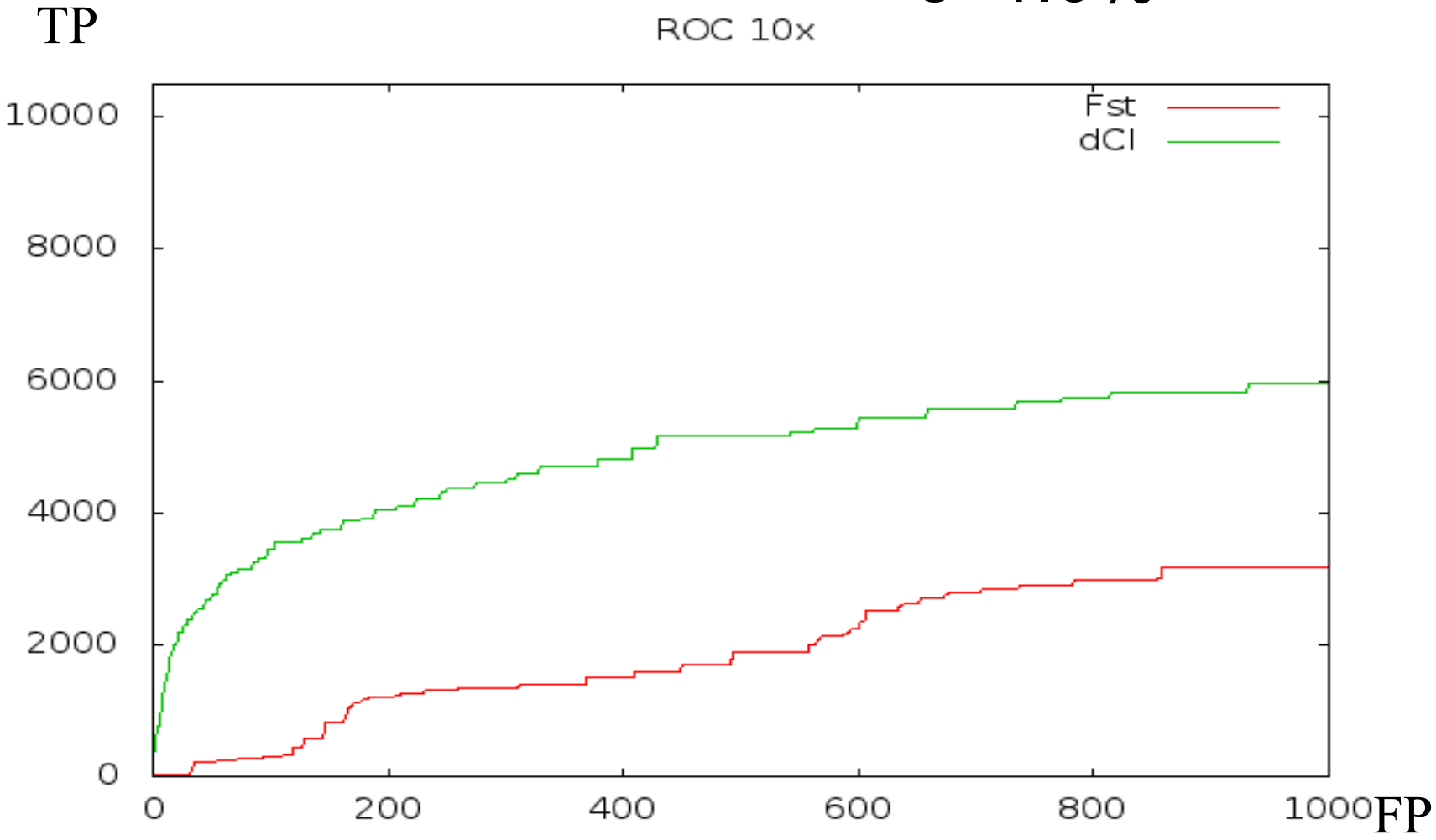
Polymorphic sites - 10x

e=2.0%



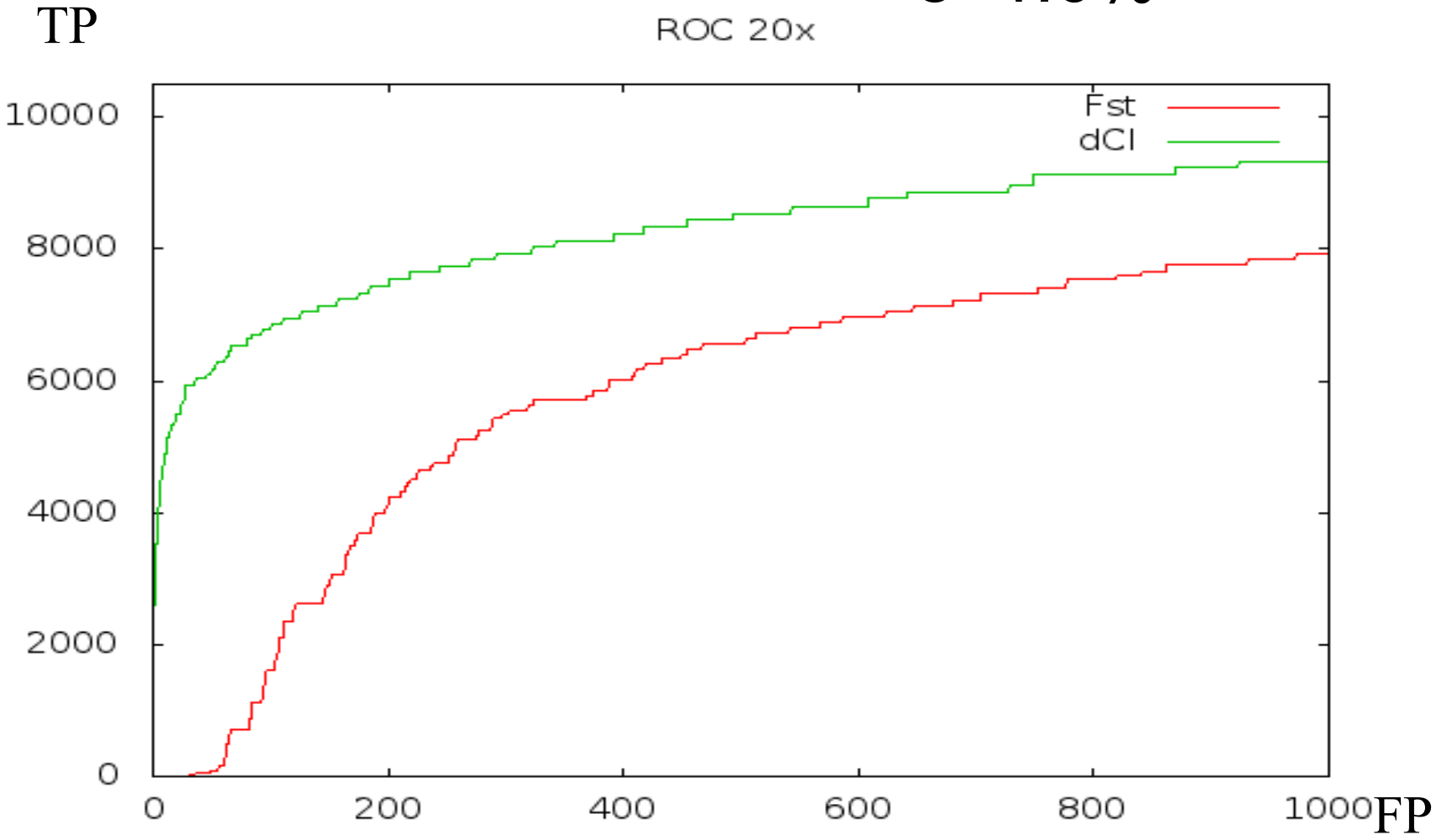
ROC 10x

e=1.0%


TP
ROC 10x
10000
8000
6000
4000
2000
0
Fst
dCl
0
200
400
600
800
1000
FP

ROC 20x

e=1.0%


TP
ROC 20x
10000
8000
6000
4000
2000
0
0
200
400
600
800
1000
FP
Fst
dCI

Caveats

Simulation is always overly optimistic

- uniform distribution of reads
- uniform distribution of SNPs
- absence of contamination and artifacts
- few repeats and low-complexity regions
- no indels or rearrangements

= Reality will be a lot *WORSE*

And yet: conclusions

- FST is *very vulnerable* to sequencing errors
- dCI is *more reliable* for selecting diagnostic SNPs
- ..but FST *might* work, too - at least for artificial data
- Estimating FST based on sequence pools seems questionable
- Sliding window averages will just make things **worse**

The End

http://malde.org/~ketil/biohaskell/varan