PCR, Tropical Arithmetic, and Group Testing

Hsin-Po Wang
with Ryan Gabrys and Alexander Vardy

Department of Electrical and Computer Engineering, University of California San Diego

Slides
https://h-p.wang/isit

Preprint
arXiv: 2201.05440

Motivation of This Work

Overall goal is to screen many people for covid (or for the next pandemic).

Antigen testing and antibody testing:
Cheap and fast; but not too sensitive.

PCR (polymerase chain reaction) testing:
Sensitive but expensive and slow;
keep track of variants (alpha, delta, omicron, etc).

Q: How to combine PCR testing and Group Testing (GT)?

Outline of This Talk

The working principle of PCR testing.

Variants of group testing (GT).

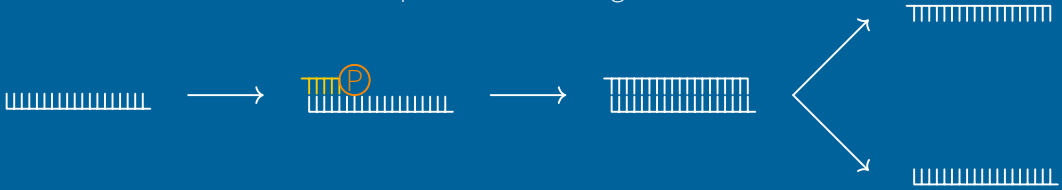
Our GT, called tropical GT.

Working Principle of PCR

A PCR machine is a sauna room for test tubes,
with three settings: cold, warm, and hot.

https://www.craiyon.com/
prompt: test tubes in sauna

Cold = a primer and a polymerase stick to a single-stranded DNA.
Warm = the polymerase synthesizes the complement strand of the DNA.
Hot = a double-stranded DNA splits into two single-stranded DNAs.

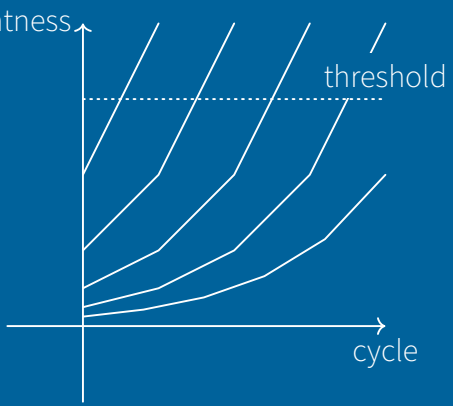


How to Detect DNA and What's Ct Value?

The amount of DNA doubles every cold-warm-hot cycle.

Insert fluorescent dyes that like to attach to DNA.
As the amount of DNA increases, the tube glows.

Ct (cycle threshold) value is
#cycles before we see the tube glowing.



So…

Can GT Make Good Use of Ct Values?

Review: Binary GT

In binary GT, a test result is either negative or positive.

Mix samples of five people.
If the mixture is negative, all five people are healthy.
If the mixture is positive, at least one is infected.

Origin = [Dorfman 1943]. Book = [Du–Hwang 1993]. Lecture note: [Ngo–Rudra 2011].
Recent survey: [Aldridge–Johnson–Scarlett 2019].

Review: Threshold GT

If less than L people are infected, the mixture is negative.
If more then U people are infected, the mixture is positive.
Inconclusive if between L and U.

Binary GT: $(L, U) = (0, 1)$.

[Damaschke 2006] [Dyachkov 2013] [Cheraghchi 2013]

Review: Quantitative GT

You have ten bags of coins, each containing many coins. Each coin weighs 5 grams.
One bag contains fake coins; each fake coin weighs 4.5 grams.
Task: Use a spring scale to find the fake bag.

Another name = coin-weighing problem.
[Hwang 1987] [Guy–Nowakowski 1995] [Bshouty 2009]

Review: Semi-Quantitative GT

The spring scale is rusty, accurate up to 1 gram.

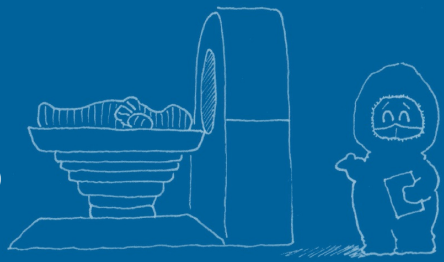
This version is basically a combination of quantitative GT and threshold GT.

[Emad–Milenkovic 2014] [Cheraghchi–Gabrys–Milenkovic 2021]

Review: Compressed Sensing

Very similar to semi-quantitative GT.
Want to solve $\mathbf{y} = A\mathbf{x} +$ errors.

Some meta choices:
A is zero-one matrix or with real numbers?
Usual matrix multiplication $(A \cdot B)_{ik} := \sum_j (A_{ij} \cdot B_{jk})$
or logical version $(A \wedge B)_{ik} := \bigvee_j (A_{ij} \wedge B_{jk})$?
Minimize $|A\mathbf{x} - \mathbf{y}|_2^2 + \lambda |\mathbf{x}|_1$ or other metric?



Recent works: [Ghosh et al. 2021] [Shental et al. 2020] [Mutesa et al. 2021]
Survey: [Aldridge–Ellis 2022]

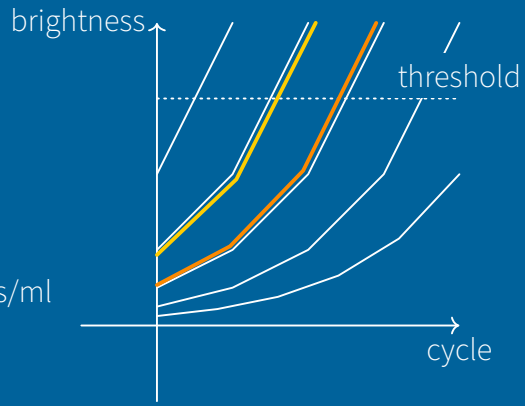
But…

Ct Values Do Not Fit.

PCR Precision Issue

DNA can double many many times.
PCR is as sensitive as 100 copies/milliliter.

On the other hand, x copies/ml and $1.9x$ copies/ml may have the same Ct value.



And the “Problem” with Logarithmic Scale

White noises of 50 dB and 30 dB combined = 50.043 dB.

Mixing pH 1 and pH 3 acids = diluting pH 0.9957 by two-fold.

Magnitude 9 and magnitude 8 earthquakes together = 9.009.

Star with apparent magnitude 1 close to star with 6 = looks like 0.9892.

Actual Question is..

How to “Add” under Logarithmic Scale?

Use Tropical Arithmetics!

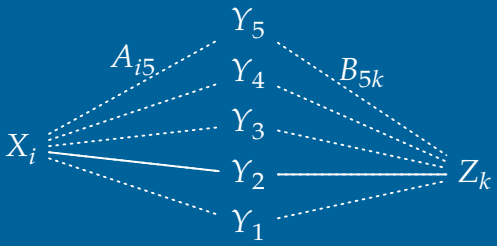
Rules are as follows:
The domain is real numbers and infinity $\mathbb{R} \cup \{\infty\}$.
Tropical addition: $x \oplus y := \min(x, y)$.
Tropical multiplication $x \odot y := x + y$.

Hint: It's all about logarithm.
$2^{-x} + 2^{-y} \approx 2^{-\min(x,y)}$, especially when $|x - y|$ is big
$2^{-x} \cdot 2^{-y} = 2^{-(x+y)}$

Extend Tropical Arithmetics to Matrix Multiplication

Let $A \odot B$ be a matrix whose (i,k)th entry is let to be $\bigoplus_j (A_{ij} \odot B_{jk}) = \min_j (A_{ij} + B_{jk})$.

Combinatorial meaning:
Suppose $X_1, \ldots, X_\ell, Y_1, \ldots, Y_m, Z_1, \ldots, Z_n$ are points on Google map.
Let the distance from X_i to Y_j be A_{ij}. Let the distance from Y_j to Z_k be B_{jk}.
$(A \odot B)_{ik}$ is the distance from X_i to Z_k via the best choice of Y_j.



Axiomize PCR and Pooling

Suppose there are n samples with Ct values $x_1, x_2, \ldots, x_N$.
The Ct value of the mixture should be $-\log_2\left(\sum_{j=1}^{N} 2^{-x_j}\right)$.

This quantity is close to, and we pretend that it is exactly,

$$\begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix} \odot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \bigoplus_{j=1}^{N} x_j = \min_{1 \le j \le N} x_j.$$

Axiomize PCR and Pooling ... and Delay!

Suppose there are n samples with Ct values $x_1, x_2, \ldots, x_n$.
Suppose we insert them into the PCR machine after $\delta_1, \delta_2, \ldots, \delta_n$ cycles, respectively.
The final Ct value should be $-\log_2\left(\sum_{j=1}^{N} 2^{-\delta_j - x_j}\right)$.

This is close to, and we pretend that it's exactly,

$$\begin{bmatrix}\delta_1 & \delta_2 & \cdots & \delta_N\end{bmatrix} \odot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \bigoplus_{j=1}^{N} (\delta_j \odot x_j) = \min_{1 \le j \le N} (\delta_j + x_j).$$

Problem Statement

A (T, N, D)-tropical code is a matrix $S \in (\mathbb{Z} \cup \{\infty\})^{T \times N}$ such that, for any two vectors $\mathbf{x}, \mathbf{y} \in (\mathbb{Z} \cup \{\infty\})^N$, each with at most D finite entries,

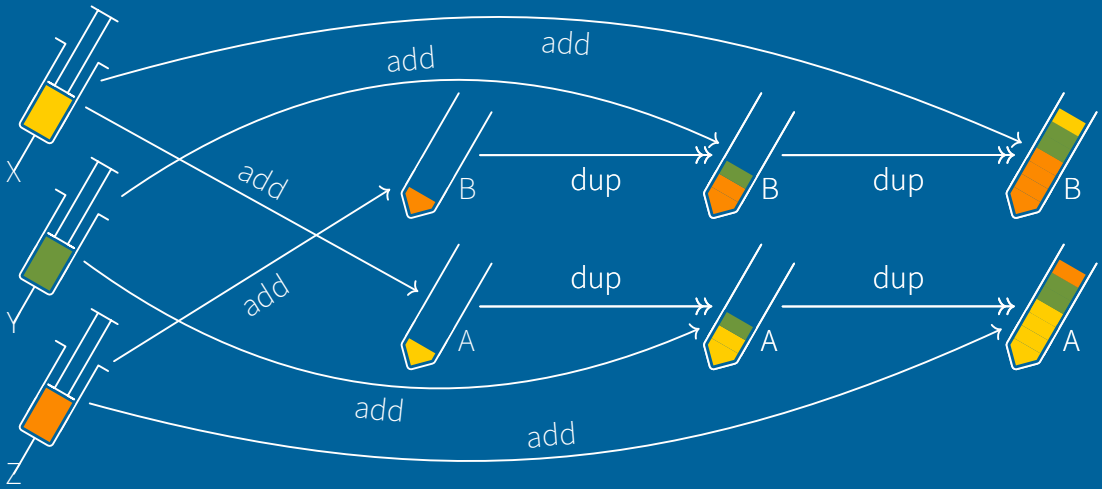
$$S \odot \mathbf{x} \neq S \odot \mathbf{y}.$$

A tropical code is said to be within maximum delay ℓ if $S \in \{0, 1, \ldots, \ell, \infty\}^{T \times N}$.

Goal: Find good tropical codes.

Why Delay?

How Does Delaying Help GT?

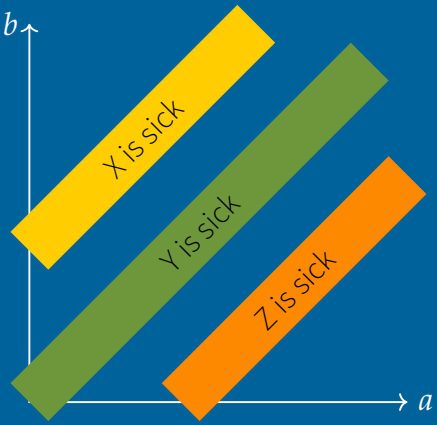


$$\begin{bmatrix} a \\ b \end{bmatrix} \coloneqq \begin{bmatrix} 0 & 1 & 2 \\ 2 & 1 & 0 \end{bmatrix} \odot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} \min(0+x,\, 1+y,\, 2+z) \\ \min(2+x,\, 1+y,\, 0+z) \end{bmatrix}$$

Decoding the Previous Slide

Suppose at most one person is infected.

X is infected iff $a - b = -2$.
Y is infected iff $a - b = 0$.
Z is infected iff $a - b = 2$.



Main Results on Nonadaptive Tropical GT

When there is $D = 1$ infected person in a population of size N,
and the delay is limited to ℓ cycles, we will use $T \approx \log_{\ell+1}(N)$ tests.

When there are $D = 2$ infected persons in a population of size N:

- The first construction uses $T \approx 2\sqrt{N}$ tests.
  In this construction, every person is present in only two tests.
- The second construction uses $T \approx 1.01 \log_2 N$ tests
  and limits the delay to $\ell \approx 3 \log_2(N)$ cycles.
  This outperforms the IT bound of binary GT.

For general D, we give one necessary condition and two sufficient conditions.

Main Results on Adaptive Tropical GT

When adaptive testing is allowed, $T = 4$ tests are sufficient to find $D = 2$ infected persons among arbitrarily many persons.

In general, $T = 3D + 1$ tests are sufficient to locate D infected persons among arbitrarily many persons. For this construction, one does not need to know D beforehand.

When delays are limited to ℓ cycles, we show that $T \approx 4D \log_\ell N$ tests suffice. For this construction, one does not need to know D beforehand.

Summary of Novelty

1. We use $x \oplus y := \min(x, y)$ to characterize the result of mixing Ct values x and y. This simplifies decoding.

2. We use $\delta \odot x := \delta + x$, i.e., delaying, to enhance GT. This inspires new combinatorics problems.

3. Tropical matrix multiplication becomes a succinct language. Nonadaptive tropical GT looks like “tropical compressed sensing.”

Future Works
Open to Questions

- Random coding approach? Asymptotics?

- Noisy/erroneous measurements?

Slides
https://h-p.wang/isit

Preprint
arXiv: 2201.05440

Appendix

PCR Error Models

Error Models of PCR

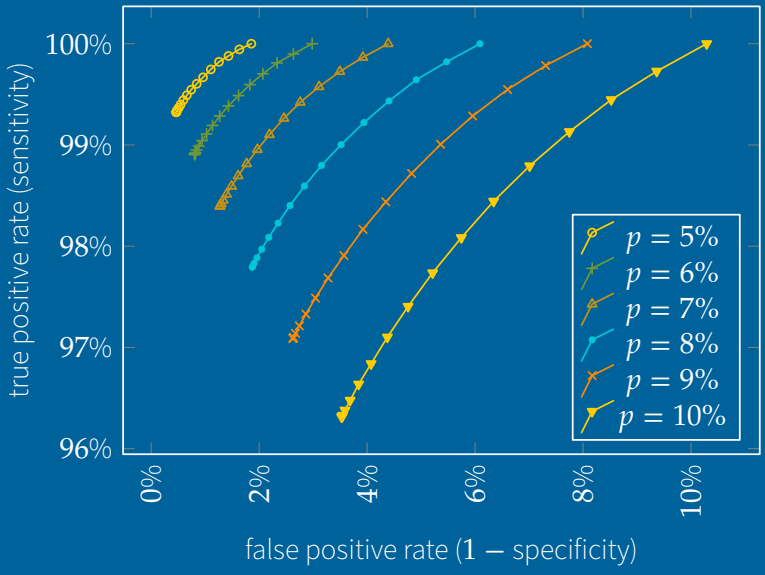
Error model 1: Ct value is rounded to the nearest integer.

Error model 2: Use fractional cycle counts (whatever that means).
(When optimizing PCR for time, earlier cycles take more time and later cycles take less time.)
Not all single-stranded DNA will be completed; DNA increases 1.9-fold or 2-fold.

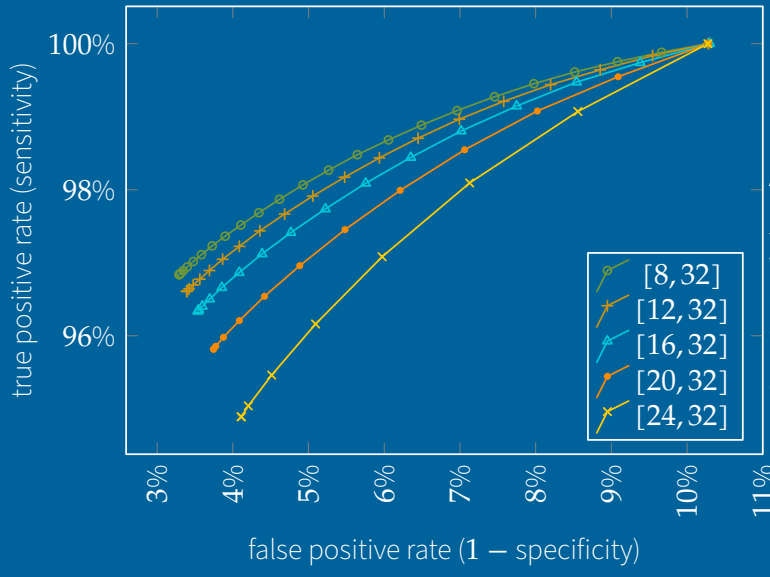
Error introduced independently: Assume tropical addition.

Appendix

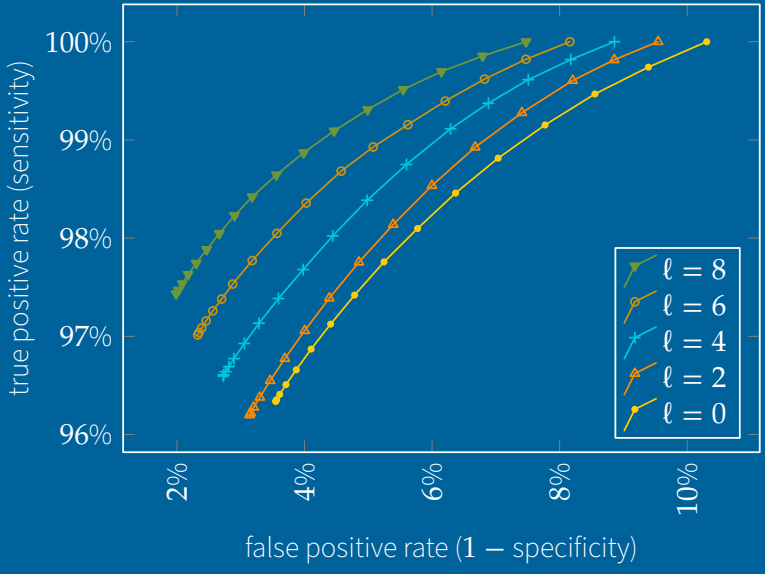
Simulation Plots



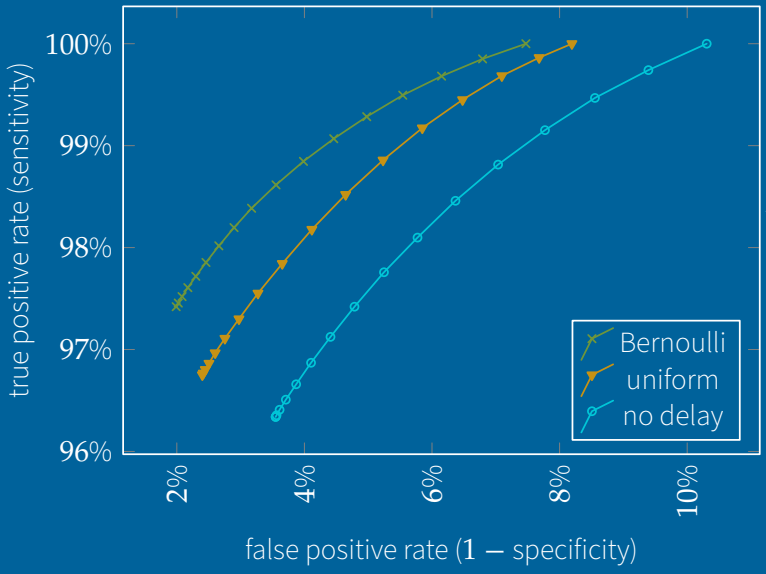
Assume uniform Ct values on the interval $[16, 32]$, 15×35 Kirkman triple system, and no delay ($\ell = 0$). We vary the prevalence rate p and plot the ROC curves.



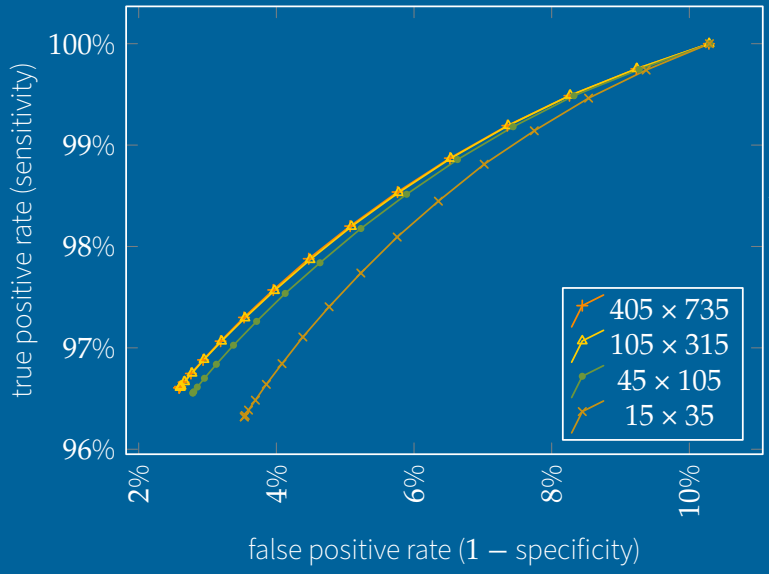
Assume prevalence rate $p = 10\%$, uniform Ct values, 15×35 Kirkman triple system, and no delay ($\ell = 0$). We vary the range of the Ct values and plot the ROC curves. Surprisingly, larger interval (consequently larger variance) is easier to decode.



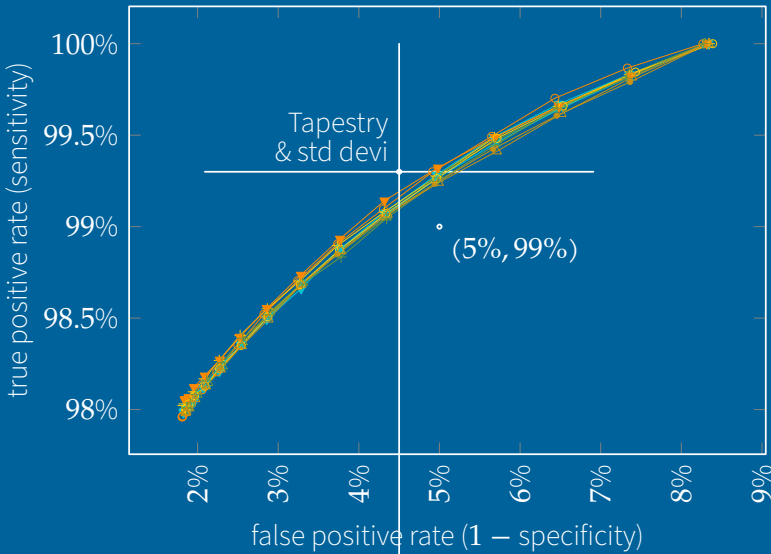
Assume prevalence rate $p = 10\%$, uniform Ct values on the interval $[16, 32]$, 15×35 Kirkman triple system, and $\ell \cdot \text{Bernoulli}(1/2)$ delay. We vary the limit of delay ℓ and plot the ROC curves.



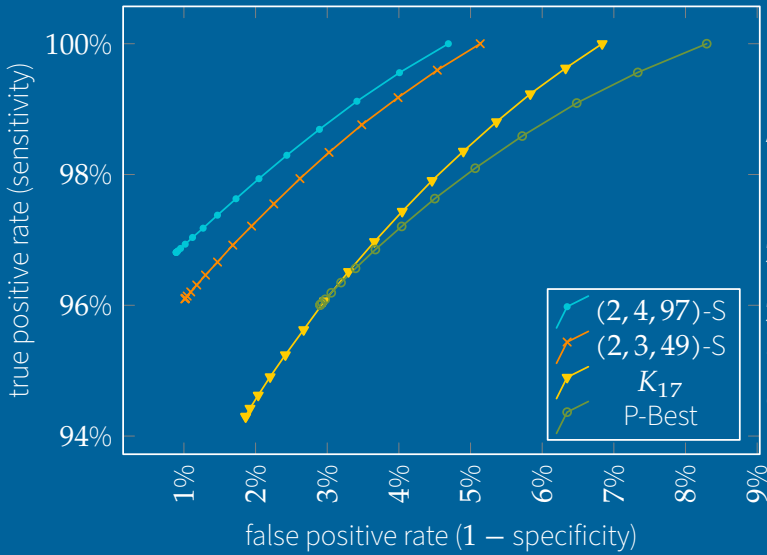
Assume prevalence rate $p = 10\%$, uniform Ct values on the interval $[16, 32]$, 15×35 Kirkman triple system, and $\ell = 8$. We vary the distribution of the random delay δ and plot the ROC curves.



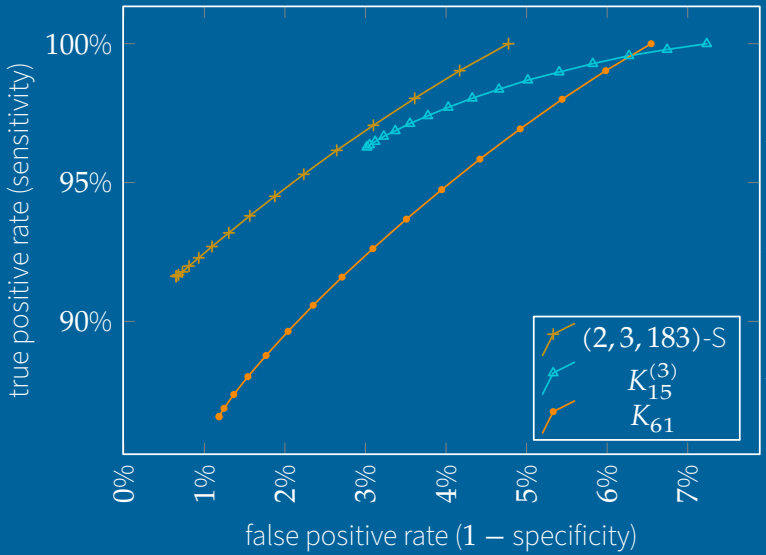
Assume prevalence rate $p = 10\%$, uniform Ct values on $[16, 32]$, and no delay ($\ell = 0$). We consider Kirkman triple systems of different size (after truncation so that the code rate $N/T = 7/3$ is fixed) and plot the ROC curves.



Assume $D = 10$ patients within $N = 105$ persons (infection rate 9.52%), uniform Ct values on $[16, 32]$, 45×105 Kirkman triple system (truncation of a 45×330 Kirkman triple system), and no delay ($\ell = 0$). We plot 10 ROC curves. Each curve is $10{,}000$ encoding–decodings, i.e., $450{,}000$ tubes, $100{,}000$ patients, and $1{,}050{,}000$ test takers. Compare this to Tapestry's data point and its standard deviations ($4.50\% \pm 2.41\%, 99.30\% \pm 2.55\%$) (Table S.XII of the preprint version [Ghosh et al. 2020]).



Assume prevalence rate $p = 2\%$, uniform Ct values on $[16, 32]$, and no delay ($\ell = 0$). We consider $(2, 4, 97)$-Steiner system (aka 2-$(97, 4, 1)$ design), $(2, 3, 49)$-Steiner system (aka 2-$(49, 3, 1)$ design), complete graph on 17 vertices, and P-BEST [Shental et al. 2020]. They all have code rate $N/T = 8$. We plot their ROC curves.



Assume prevalence rate $p = 0.5\%$, uniform Ct values on $[16, 32]$, and no delay ($\ell = 0$). We consider Kirkman triple system on 183 vertices, complete 3-uniform hypergraph on 15 vertices, and complete graph on 61 vertices. The first two have code rate $N/T = 30 + 1/3$; the last one has code rate $N/T = 30$. We plot their ROC curves.

Appendix

Comparison of GT Models

Four ways to quantify and combine test outputs. Binary tests output “negative” or “positive”; combining samples means logical OR. Quantitative tests output numbers; combining samples means addition. The other two regimes lie in between.

Regime	Reading	Remixing
Binary	Negative, Positive	Neg $\vee$ Pos = Pos
Tropical	$2^{-\infty}, \ldots, 2^{-40}, \ldots, 2^{-0}$	$\min(30, 15) = 15$
Semiquantitative	$[0,3), [3,6), [6,9), \ldots$	$[0,3) + [3,6) = [3,9)$
Quantitative	$0, 1, 2, 3, 4, 5, \ldots$	$8 + 9 = 17$