NIST AI Technology Evaluation Test Specification
Human Genome Variant Curation

Version 1.0— May 29, 2026

Test Version:	1.0
Task:	Detection
Input:	Single Image and Text
Output:	Text
Task Champion(s):	Justin Zook
Dataset:	Genome Variant Visualization v1.0
Metric:	An unweighted detection cost function will be computed.

Introduction

The predominant approach to identifying biologically and clinically relevant genomic variation in humans is to align reads and call variants against a reference human genome. Accurate, high-resolution benchmark datasets have had fundamental impact in facilitating the development of sequencing technologies and bioinformatics tools as well as the validation of clinical tests. The creation of such benchmark datasets requires careful curation of genomic variants. The task described in this test represents a basic ability for genome variant curation and, as models demonstrate their abilities and improve over time, additional, increasingly complex tasks in support of human genome variant curation will be added to the test.

Task Definition

Given an image of a small region of the genome and text describing a set of variants in this region, determine whether all of the given variants were correctly called.

Dataset Description

The test data consists of images visualizing DNA sequencing data.

1. *Modalities:* Images
2. *Data creation/collection date(s):* April 2026
3. *Derived Size:* $\sim$ 1.1 GB (10000 images and prompts)
4. *Reference labels:* It is based on highly accurate assembly from multiple sequencing technologies and confirmed by experts.
5. *Inter-annotator Agreement/Annotation and Data Quality Assessment:* Previous efforts validated the approach. Here, the data was further spot sanity checked by expert.

6. *Other pertinent features:* Due to huge amounts of the entire gene sequence where no variants take place, data was selected to focus on difficult regions of the genome.

Testing Protocol

For each image in the test dataset, provide text describing a set of variants purportedly present in the image. Ask the model whether the given variants were all correctly called.

1. *Input:* One image and a trial-specific prompt (detailing the purported variants)
2. *Output:* Decision (all_correct or some_incorrect)
3. *Scoring procedure(s):* Get output from model, compare against ground truth, compute error rates and metric value

Metric and Score Reporting

The primary metric will be an unweighted detection cost function, consisting of the average of the Type I and Type II error rates:

$$C_{Primary} = \frac{P_{Miss} + P_{FalseAlarm}}{2} \tag{1}$$

where P_{Miss} is the observed rate where the model incorrectly responds that the provided variants are all correct, and $P_{FalseAlarm}$ is the observed rate where the model incorrectly responds that some of the provided variants are incorrect.

Uncertainty will be computed using bootstraping methods[1].

Change Logs

- Version 1; May 29, 2026; Initial Release

[1]Efron, Bradley. "Bootstrap methods: another look at the jackknife." Breakthroughs in statistics: Methodology and distribution. New York, NY: Springer New York, 1992. 569-593.