# GIAB TR Project

# Benchmarking Tools

- Truvari bench
  - Classic variant comparison
  - Limitation: 1-to-1 comparison, which misses 'split' representations
- Truvari phab
  - Variant harmonization to deal with representation issues
  - Limitations: Slow (MAFFT); assumes phased variants; regional
- Combining these two steps has shown promise
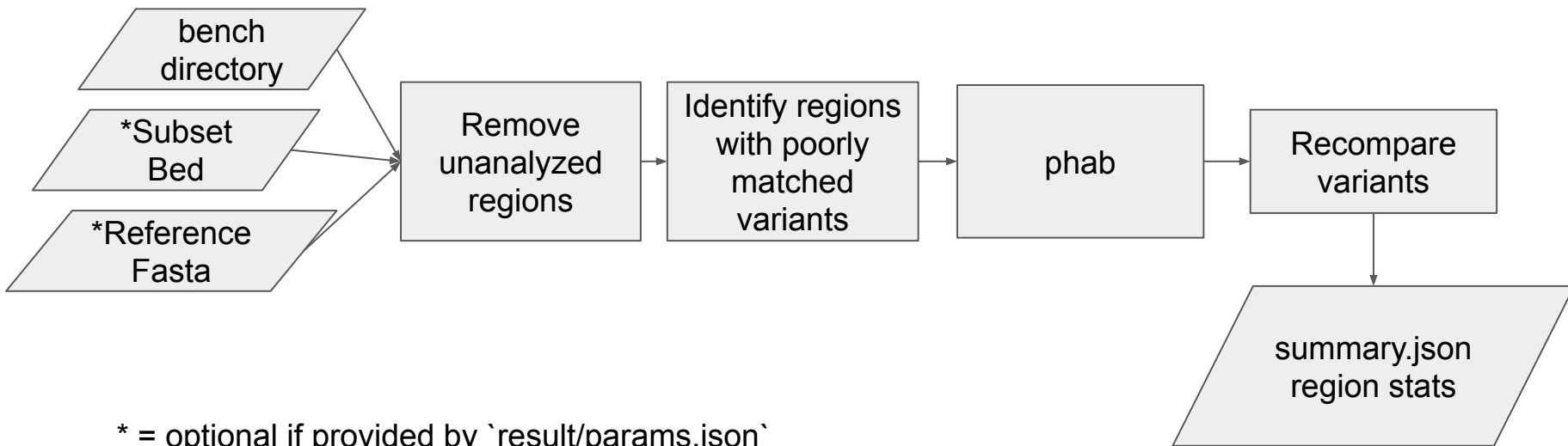  - Estimated a ~20p.p. Increase in TRGT performance

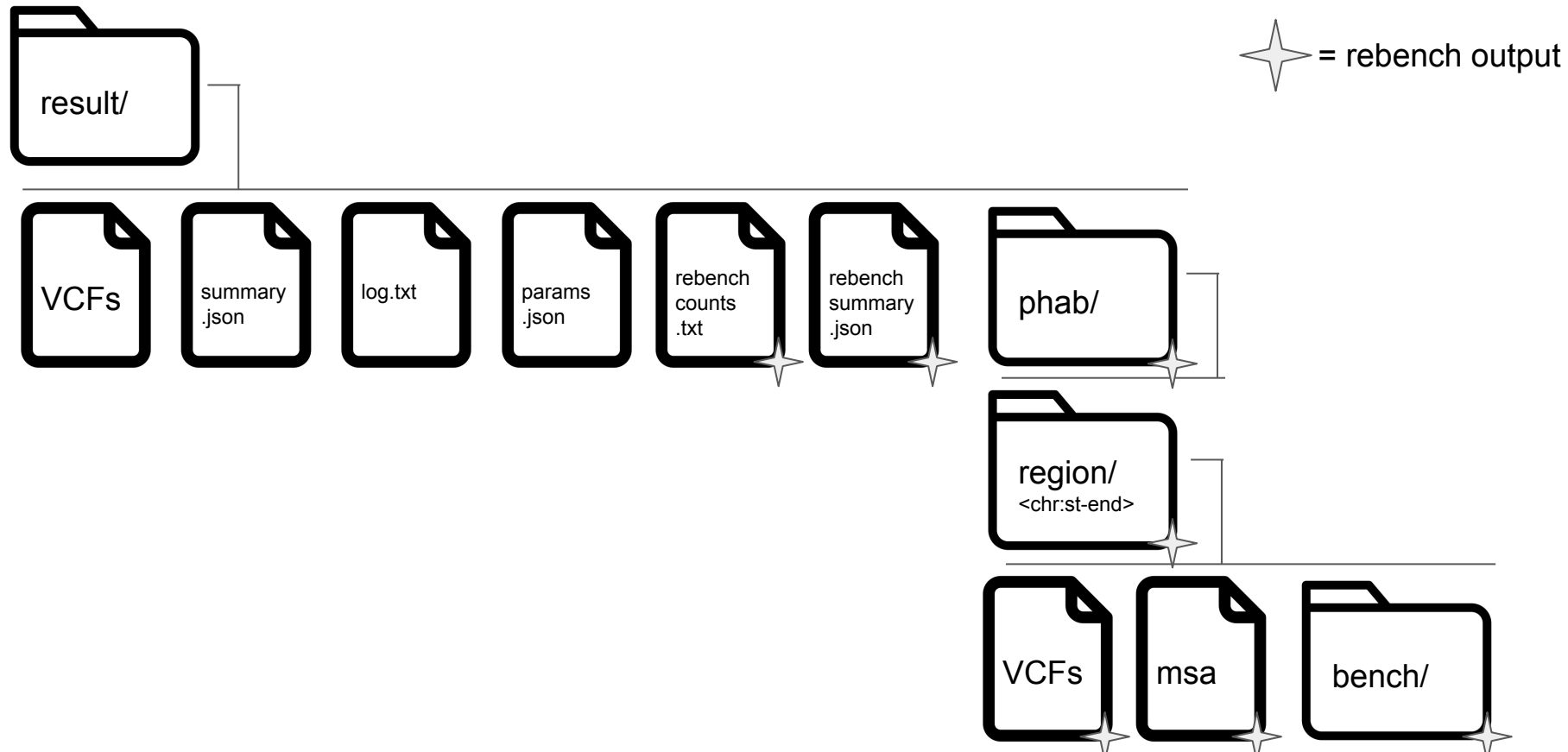|  | RTG | Truvari | Phab:GT |
|---|---|---|---|
| **tpbase** | 23,380 | 23,472 | 24,154 |
| **tp** | 25,452 | 23,472 | 24,154 |
| **fp** | 846 | 2,878 | 322 |
| **fn** | 736 | 820 | 375 |
| **recall** | 0.969 | 0.966 | 0.985 |
| **precision** | 0.968 | 0.891 | 0.987 |
| **f1** | 0.969 | 0.927 | 0.986 |



CMRG vs Adotto f1
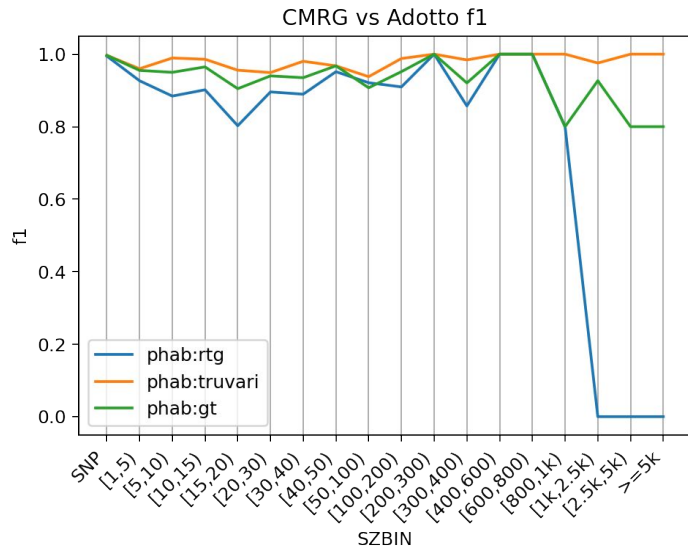
# Truvari rebench

Automate the link between `truvari bench` and `truvari phab`

```
> truvari bench -b base.vcf.gz -c comp.vcf.gz -o result/
> truvari rebench -r subset.bed -f ref.fa result/
```

bench directory

*Subset Bed

*Reference Fasta

Remove unanalyzed regions

Identify regions with poorly matched variants

phab

Recompare variants

summary.json region stats

* = optional if provided by `result/params.json`

# Outputs (v4.0-dev)

result/

✦ = rebench output

VCFs

summary .json

log.txt

params .json

rebench counts .txt ✦

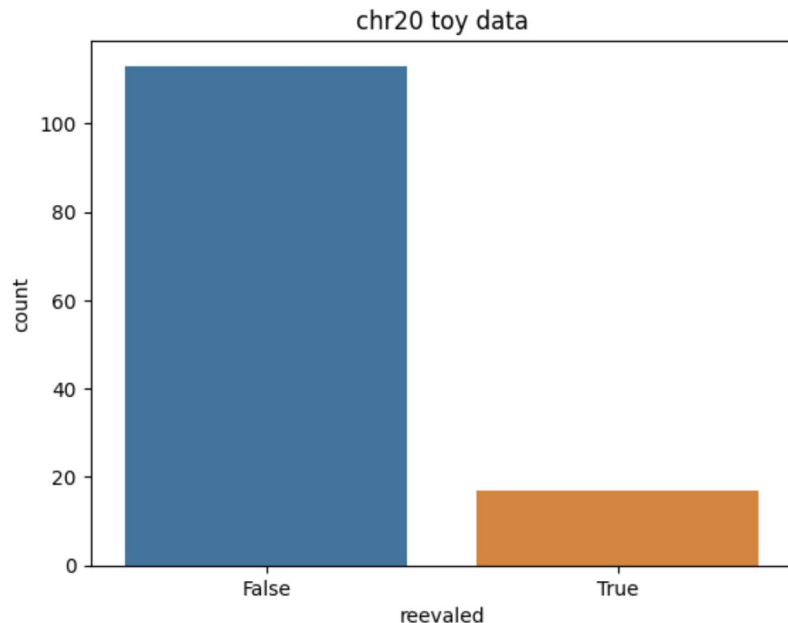rebench summary .json ✦

phab/ ✦

region/ <chr:st-end> ✦

VCFs ✦

msa ✦

bench/ ✦

# rebench.counts.txt

- Tab-separated file with variant count summaries per-region before and after re-evaluation
  - Region chrom/start/end
  - reevaled - boolean of if re-evaluated
  - base/comp counts - #variants in region
  - In_counts - #tp/fp/fn/ variants in region
  - Out_count - #tp/fp/fn variants in region after re-evaluation

```
data = pd.read_csv("rebench.counts.txt", sep='\t')
p = sb.countplot(data=data, x="reevaled")
p.set(title="chr20 toy data")
```

# rebench.summary.json

- Using HPRC HG002 chr20 VCFs from THFA and Adotto and `--sizemin 5`

```python
data = pd.read_csv("bench_output_small/rebench.counts.txt", sep='\t')
data[data["reevaled"] & (data["out_fp_count"] | data["out_fn_count"])].T
```

| | 18 | 20 | 81 | 164 |
|---|---|---|---|---|
| chrom | chr20 | chr20 | chr20 | chr20 |
| start | 5041857 | 53203987 | 64173409 | 61289625 |
| end | 5042710 | 53204474 | 64176530 | 61290556 |
| reevaled | True | True | True | True |
| base_count | 5 | 7 | 123 | 11 |
| call_count | 61 | 27 | 275 | 10 |
| in_tp_base_count | 1 | 2 | 4 | 1 |
| in_tp_call_count | 1 | 2 | 4 | 1 |
| in_fp_count | 1 | 1 | 3 | 1 |
| in_fn_count | 1 | 2 | 7 | 2 |
| out_tp_base_count | 2 | 7 | 20 | 6 |
| out_tp_call_count | 2 | 7 | 20 | 6 |
| out_fp_count | 1 | 1 | 1 | 2 |
| out_fn_count | 0 | 1 | 1 | 3 |

| | bench | rebench |
|---|---|---|
| TP-base | 390 | 588 |
| TP-call | 390 | 588 |
| FP | 69 | 28 |
| FN | 201 | 73 |
| precision | 0.849673 | 0.954545 |
| recall | 0.659898 | 0.889561 |
| f1 | 0.742857 | 0.920908 |
| base cnt | 591 | 661 |
| call cnt | 459 | 616 |

# Truvari rebench opportunities for improvement

- Identifying 'poorly matched regions' is just `FN > 0 & FP > 0`
  - More informed identification should be possible
- Incorporate hap_eval, possibly other comparison approaches?
- Choice of re-evaluation procedure is a user parameter.
  - Re-evaluation chosen by region/variants' properties