GIAB TR

Truvari changes

- Truvari refine
  - Formerly ‘rebench’
  - Major code refactor (~9x speed improvement)
  - Simplified output
- Truvari stratify
  - New tool
- Truvari bench --gtcomp
  - New genotype comparison option

Stratifications

- `stratify` counts number of variants by state for each input region
- `vcf2df` makes data science friendly results
- Note: `bench --giabreport` is removed from v4.0

```
$ truvari stratify input.bed bench/
chrom     start     end     tpbase  tp   fn   fp
chr20     280300    280900       0   0    0    0
chr20     100000    200000       1   1    0    0
chr20     642000    642350       1   1    2    1
```

```
$ truvari vcf2df -b bench/ data.jl
>>> d = joblib.load("data.jl")
>>> d.groupby(['szbin', 'state']).size().unstack()
state       tpbase   fn    tp   fp
szbin
SNP              0    0     0    0
[1,5)            0    0     0    0
[5,10)        1296  552  1296  596
[10,15)        625  256   622  278
[15,20)        285  136   284  149
[20,30)        258  177   261  182
... etc ...
```

Truvari bench --gtcomp

- Uses the genotype allele count to inform number of matches

```
CHROM POS      ID  REF ALT          base comp
chr20 17785968 ID1 A   ACGCGCGCGCG   1/1  1/0
chr20 17785968 ID2 A   ACGCGCGCGCGCG 0/0  0/1
chr20 17785969 ID3 C   CGCGCGCGCGCGC 0/0  1/1
```

Parameter	ID1 State	ID2 State	ID3 State
default	TP	FP	FP
--multimatch	TP	TP	TP
--gtcomp	TP	TP	FP

ID2 similarity to ID1: Seq=0.92 Size=0.83 Ovl=0.85 SizeDiff=-2

$$T = L(A_1) + L(A_2)$$
$$seqsim = 1 - (editDistance/T)$$

Truvari bench --gtcomp on TRGT after refinement

	default	--gtcomp
TP-base	4,553	4,623
TP-call	4,553	4,664
FP	817	698
FN	385	313
precision	0.848	0.870
recall	0.922	0.937
f1	0.883	0.902

`refine.counts.txt` analysis:

- **4.5%** regions needed refinement (951 of 21,090)

- **95%** had T/F change (933 of 951)
- **97%** had improved F1 (906 of 933)
- 2% had identical F1 (25)
- 1% had worse F1 (2)

The number of regions needing refinement will likely lower once we've curated the TR Regions down to the HG002 specific benchmark.

TR Regions Plan

1. Remake TR regions with updated Vamos Annotations
2. Refine region boundaries (expansion)
3. Remove non-TR contaminated regions
4. Simplify TR annotations
5. Refine region boundaries (contraction)
6. Add per-annotation and per-region 'purity' score
7. Add TR annotation subregion stats
8. Add CODIS and known pathogenic columns
9. Add 'ambiguity' score