LARVA-SAM computational efficiency

- LARVA-SAM is still somewhat slow when scaling above a few hundred *nrand* datasets
  - Past this, timescale stretches from hours into days or weeks
- Most compute-intensive step now is picking the position of each variant
- Current algorithm:
  - Treat the entire exome/genome as a number line from 1 to *max_coord*
  - Use a pseudorandom number generator over [1:*max_coord*]
  - Find the annotation/region that contains the coordinate produced by the rand generator

LARVA-SAM computational efficiency

Remedy

- Use the position picking code as a precomputation
  - Generate a variant pool that has the distribution implied by the exome/wg null models
- Then create random variant datasets with variants randomly drawn from this pool
  - Currently using a ~6 million variant pool. Is this enough?
- A sensible approach?