Analysis Pipeline Wrap-up

David Levine

August 9, 2017

Outline

- Cloud computing
  - Pros and cons
  - Benchmarks
  - Cost reduction
- Where to get DCC software

Cloud computing pros

- No/low infrastructure costs
- Pay per use model
- Minimal administration and management
- Automatic software updates
- Reliability and disaster recovery
- Scalable with increasing data set sizes
- Variety of computers (RAM, CPU, disk, GPU)

Cloud computing cons

- Ongoing monthly costs (vs. large up-front payment)
- Pay for failed & debugging runs, instance left running
- You are your own IT person (or still need one)
- Management of own security
- RAM, CPU and disk scale together
- Extra effort to minimize costs (non-uniform resources)
- Cloud vendor lock-in

Major computational need

- Run one time
  - VCF to GDS file conversion
- Run a few times
  - Relatedness analysis
- **Run many times**
  - **Association testing**

What influences cloud costs

- No. samples
- No. variants & filtering
- No. variants per aggregation unit
- No. analyses per trait
- Algorithm: Single variant, SKAT, LR, LMM
- Computer (cores, RAM, disk)
- Non-uniform resource requirements

Cloud benchmarks

- AWS On-demand pricing
- CfnCluster (old) and AWS Batch (new)
- Single variant using LMM (MAF > 1%)
- SKAT using LMM (MAF < 1%, 5kb window)

AWS cloud benchmarks

Row	Analysis	No. of Samples	No. of Variants	Time (hours)	Max Cores	Max Jobs	Cost	Parallel Software
1	Single Variant	16,503	185,970,832	4.0	512		$212	CfnCluster
2		16,503	185,970,832	3.5	500	161	$100	AWS Batch
3		16,503	185,970,832	6.0	180	59	$70	AWS Batch
4	SKAT	16,503	185,970,832	16.0	592		$787	CfnCluster
5		16,503	185,970,832	14.0	500	161	$370	AWS Batch

- N subjects, M variants
- Single variant tests
  - RAM $O(N^2)$
  - CPU $O(MN)$
- SKAT tests
  - RAM $O(N^2)$
  - CPU $O(M^2N)$

Plans to reduce costs

- Cloud computing environment
  - Spot pricing (checkpoint/restart)
  - Optimize heterogeneous computing strategy
- *fastSKAT* fast and highly-accurate approximations to SKAT
  - Reduce CPU scaling from $O(M^2N)$ to $O(MN)$
  - Preliminary: Reduce computation 2-3 orders of magnitude
- Reduce memory requirements
  - Assume subject independence in different studies/ancestry groups
  - Computation grows as largest study/ancestry group
  - More efficient sparse matrix algorithms
- Meta-analysis across studies
  - Assumes subject independence in different studies/ancestry groups
  - Computation grows as largest study/ancestry group
  - Files to share become large and burdensome (SKAT)

How to get DCC software

- Distributed as R/Python source code or Docker images
- Primary focus on R power users
- Can be integrated into other environments
- R/Bioconductor packages
  - https://bioconductor.org/packages/SeqArray
  - https://bioconductor.org/packages/SeqVarTools
  - https://bioconductor.org/packages/SNPRelate
  - https://github.com/smgogarten/GENESIS
- Docker images
  - https://hub.docker.com/r/uwgac/r-topmed
- TOPMed analysis pipeline
  - https://github.com/smgogarten/analysis_pipeline