Sequence Data Processing

Workshop on
Central Resource of Data
From Genome Sequencing Projects
Why?

• Many analyses will benefit from combining information across sequencing projects

• Possibilities include ...
  – Meta-analyses that improve on analyses of any single sample
  – Case-control studies of rare variation that use many controls
  – High-resolution of analyses of natural selection

• Differences in sequence processing between projects can affect these analyses to different degrees
Case-Study #1
Rare variant in CFH and macular degeneration

• R1210C, rare variant in \textit{CFH} that abrogates C-terminal ligand binding, is associated with AMD

• What would it take to rediscover the variant in an exome wide experiment?

• We sequenced 2,348 AMD cases and 789 controls in collaboration with Washington University Genome Center
  – Variant is seen in 23 cases, 0 controls (good!)
  – P-value is about .003 (middling!)
  – Variant present 2 of 12,000+ exomes used for exome chip design (impressive!)
Case-Study #2
Comparison of Exomes Sequenced at Two Centers

- Initial calls show many differences between centers
- Calling and filtering with uniform process reduces differences
- Many differences are not intrinsic to sequence generation, but to calling
  Filtered, On-Target, Near-Target
Options for Sequencing Processing

• Laissez-Faire:
  – Each project provides its own calls
  – Focus on standard formats, queriable structures

• Central Planning:
  – Define minimum standards for calls that are deposited
  – Define analysis tools for calls that are deposited
  – Increases similarity between datasets

• Central Analyses:
  – Calls generated centrally, using data across many projects
Option #1
Using Calls Provided by Each Project

• Some valuable analyses are relatively robust to differences between sequence analysis protocols
  – Meta-analyses of association study results for quantitative traits

• Facilitating this option still requires:
  – Harmonization of phenotypes
  – Consistent use of standard formats
  – Streamlining of data access protocols
  – Data models that facilitate combining data across studies
Option #2
Minimum Standards for Calls

• A set of minimum standards for calls generated by each project could help...
  – Analyses should include variant types beyond SNPs
  – Analyses report per base coverage in addition to discovered variants

• Standards could even require that each study is processed with the same set of tools

• This would provide incremental improvement on option #1, but probably still only allow meta-analysis
  – The power of artifact filters, for example, depends on sample size
  – Old and new projects would likely be analyzed with different tools
Option #3
Joint Processing of Many Projects

• Most compute and labor intensive

• Many analyses improve with sample size
  – Power to discover variants
  – Ability to resolve complex events
  – Ability to resolve haplotypes
  – Ability to filter sequencing artifacts

• Allows benefits of new analysis tools to percolate

• Technically feasible to call 10,000s of samples ....
• ... especially if we are happy with 80% solution
Challenges for Joint Processing of Many Projects

• Uniform protocols for accessing sequence data across studies are essential
  – Much more difficult if analysis require manual intervention

• The challenges of handling corner cases can’t be underestimated
  – When are we willing to drop legacy data?
    • Shortest reads
    • Higher error rates
    • Obsolete platforms

  – A few samples with poor quality data can influence results
Sharing of “Derivates”

• Some information, like allele frequencies, could allow many benefits of joint calling without sharing raw sequence data

• Examples include:
  – Distilled summaries of haplotype structure
  – Distilled prior evidence for variant bases

• The risks of sharing these derivatives are similar to those involved in sharing allele frequencies
Final Thoughts

• All these options are likely to be pioneered by investigators with shared scientific interest
  – What happens when we combine individuals with information on a favorite trait across sequencing studies?

• Currently, not fully exploiting what can be done with calls from individuals projects (whether GWAS or sequencing)

• Many opportunities for improved sequence analysis by combining data processing across projects