Reproducible research and genome scale biology: approaches in Bioconductor

VJ Carey (Harvard University), R Gentleman (Genentech)

1. Software and data are dynamic. They must necessarily evolve as our knowledge improves, or as bugs and processing errors are detected and remedied. Both aspects require version tracking. The use of scripting languages, especially in environments that mix tool chains, is problematic. An analysis document should be stamped with the version numbers for all data and software components that were used to create it.
2. Biological metadata (sequences, functional annotation) are constantly evolving. Stable analyses that take time to complete benefit from frozen images of metadata. These images must be versioned, and their provenance made explicit.
3. Authoring reproducible analysis documents depends on facilities for narrating analysis steps, executing the steps and exhibiting the results. Literate programming concepts are relevant, and Sweave programs help unify narrative, software scripting, and visualization. Mixed language documents are possible, but seem to presume a coherent evaluation model and some meta-language/environment to manage inter-language communications.
4. Management of voluminous heterogeneous data must be meticulous if reproducibility is to be achieved. Object designs that unify patient-level data, genome scale assay results, experiment metadata, and assay-level metadata are feasible (Bioconductor ExpressionSet, for example.) Methods defined for such objects allow simple syntax for filtering of reporters and samples. Heavily used and validated methods for filtering, combining and interpreting large collections of microarrays are the core infrastructure of Bioconductor. The designs and their wide use contribute to reliability of the system and its basic approach.
5. Failure to adopt reproducible discipline is costly both to investigators, who have higher management and execution costs when generation of past results is not fully documented or relies upon evolving tools whose versions in use have not been recorded, and to the scientific community, who may be misled by incorrect findings whose vulnerabilities are hidden. When the medical community is misled, patients may be harmed.
6. Discovery of and proof of non-reproducibility is challenging and costly. Reliance on outside volunteer parties is not sufficient. Institutional and editorial support of reproducibility verifications is critical. Tools and criteria that help investigators choose transparently reproducible analysis workflows will help reduce costs to institutions who wish to secure reproducibility of research. Platform-independent open source statistical computing can diminish important barriers to 'checking' published results.
7. If commitments to reproducible research methods in genome scale biology are not made in a uniform way, those who adopt reproducible discipline may be at a disadvantage in certain respects, because the speed of production of results will be diminished relative to a group that is more liberal in its approach. Costs of maintenance of reproducible discipline must be brought

down; penalties for working in ways that are not verifiably reproducible must be increased. Such penalties are currently only applied after discovery of errors and potential harms. Recent work of Baggerly and Coombs in Annals of Applied Statistics defines a retrospective ``forensic bioinformatics'' whose application has led to suspension of clinical trials in cancer.

8. Bioconductor's focus on platform-independent open source statistical computing with R, integrated object designs, versioned packaging of data, metadata, and analytic software, and illustration of integrated workflows using integrated computable documents (useful for analysis audits as well as publishable vignettes and monographs) can foster prospectively reproducible analysis in genome scale biology.