README file: "How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome" by Sampat and Williams

October 29, 2017

1 Description of Raw Data Files

Online Appendix C describes our raw data files in detail. All are publicly available with two exceptions:

1. **Thomson Innovation data on published USPTO patent applications:** We draw our patent "family size" measure from the Thomson Innovation database, which is proprietary and cannot be posted online. Details on purchasing this data are available here:

   `http://ip.thomsonreuters.com/product/derwent-world-patents-index`

2. **Pharmaprojects data:** We collect data on gene-related drug development from the Pharmaprojects database, which is proprietary and cannot be posted online. Details on purchasing this data are available here: `http://www.pharmaprojects.com`.

2 Datasets and programs used for final results

The datasets needed to replicate the figures and tables in the paper (all in STATA 14 format) are:

1. **appnum_by_round_analysis_final.dta:** This dataset is needed to generate Online Appendix Table A.3.
2. **examiner_stages_analysis_final.dta:** This dataset is needed to generate Online Appendix Tables A.1 and A.2.
3. **firststage_analysis_final.dta:** This dataset is needed to generate Figure 3; Tables 1 and 3; and Online Appendix Tables D.1 and D.5.
4. **secondstage_analysis_final.dta:** This dataset is needed to generate Figure 2; Tables 1 and 3; Online Appendix Figure 2; and Online Appendix Tables D.1, D.3, D.4, and D.6.
5. **selection_final.dta:** This dataset is needed to generate Figure 2, Online Appendix Figure D.2, Table 2, and Online Appendix Table D.2 .
6. **selection_patome_final.dta:** This dataset is needed to generate Figure 1 and Online Appendix Figure D.1.

The script final_analysis.do generates all tables and figures in the paper and associated appendices. Note that variables derived from the proprietary Thomson Innovation (family_size) and Pharmaprojects (log_nt2011, number trialYEAR and anytrialYEAR where YEAR is between 1995 and 2011) data are excluded from these replication files. In order to run the final_analysis.do script without errors, one will have to remove references to these variables from the analysis script.