APPENDICES A-C: The NCI and the Takeoff of Nanomedicine

Appendix A: The Globonano Scientific Literature Database and Literature Star Scientist Identification

1. Data collection

In order to obtain raw data we began downloading ISI web of science records relevant to nanotechnology using the Kostoff query beginning 19 May 2010. (Kostoff, et al. 2006) specifies a comprehensive query designed to identify nanotechnology documents with a high degree of both precision and recall; the query is widely used. The task of downloading ISI records during June 2010. Approximately 500,000 records in all were downloaded, resulting in approximately 420,000 unique nantotechnology publications, covering the complete sets of nanotech papers from 40+ nations, involving 127 countries in all.

2. Data design

The design for Globonano contains approximately 50 tables and allows for several features:

- it is *normalized*, meaning, it is updateable (specifically, third normal form)
- it has *data warehouse*-like features, meaning, it is ready for analysis;
- and it is data upload-ready, meaning, it has tables in place that assist the process of transferring raw data into the database.

The database was designed in an iterative fashion. We constructed an original scientific literature medatdata database design while collecting data records. The design was iterated upon hundreds of times during the data processing and upload process.

3. Data processing & upload

Taking the ISI metadata and transforming it into data that was atomic, query-able, and loadable into the database was a significant undertaking. Approximately 2500 lines of code (Java, bash, sed, awk, perl, and SQL) were written to automate the transformation process of raw ISI data into data ready for our database. The code not only parses the massive set of ISI data, it uploads the data into the DB, stages the data in the database, and distributes the data across the set of normalized tables. Special care was taken to automate the entirety of the process and optimize the computational process. The first complete version of the processing & upload application took approximately two weeks to process the raw data set. The application now takes less than 6 hours to process the original set of 500,000 records; starting from a raw ISI data record, the application can extract transform and load 25 ISI records per second into our database. .

4. Nanobio identification

Nanobiotechnology science articles in the Globonano database were identified using the following query, a query extending the one used in (Lenoir and Herron 2009), utilizing ISI keywordplus feature, the subject categories, and the title and abstract fields.

```
drop table if exists biocatarticles;
create table biocatarticles as
select distinct article.ut_aid as "biocatarticleid"
from article, subjectcat, article_has_subjectcat
where article.ut_aid=article_has_subjectcat.article_ut_aid
and article_has_subjectcat.subjectcat_idsubjectcat=subjectcat.idsubjectcat
and (subjectcat.subjectcat like 'Allergy' or subjectcat.subjectcat like 'Anatomy & Morphology' or subjectcat.subjectcat like 'Anesthesiology' or subjectcat.subjectcat like 'Biochemical Research Methods' or subjectcat.subjectcat like 'Biochemistry & Molecular Biology' or subjectcat.subjectcat like 'Biology' or subjectcat.subjectcat like 'Biophysics' or subjectcat.subjectcat like 'Biotechnology & Applied Microbiology' or subjectcat.subjectcat like 'Cell & Tissue Engineering' or subjectcat.subjectcat like 'Cell Biology' or subjectcat.subjectcat like 'Chemistry, Medicinal' or subjectcat.subjectcat like 'Clinical Neurology' or subjectcat.subjectcat like 'Critical Care Medicine' or subjectcat.subjectcat like 'Dentistry, Oral Surgery & Medicine' or subjectcat.subjectcat like 'Dermatology' or subjectcat.subjectcat like 'Developmental Biology' or subjectcat.subjectcat like 'Emergency Medicine' or subjectcat.subjectcat like 'Endocrinology & Metabolism' or subjectcat.subjectcat like 'Engineering, Biomedical' or subjectcat.subjectcat like 'Entomology' or subjectcat.subjectcat like 'Evolutionary Biology' or subjectcat.subjectcat like 'Fisheries' or subjectcat.subjectcat like 'Food Science & Technology' or subjectcat.subjectcat like 'Gastroenterology & Hepatology' or subjectcat.subjectcat like 'Genetics & Heredity' or subjectcat.subjectcat like 'Geriatrics & Gerontology' or subjectcat.subjectcat like 'Gerontology' or subjectcat.subjectcat like 'Health Care Sciences & Services' or subjectcat.subjectcat like 'Hematology' or subjectcat.subjectcat like 'Horticulture' or subjectcat.subjectcat like 'Imaging Science & Photographic Technology' or subjectcat.subjectcat like 'Immunology' or subjectcat.subjectcat like 'Infectious Diseases' or subjectcat.subjectcat like 'Instruments & Instrumentation' or subjectcat.subjectcat like 'Integrative & Complementary Medicine' or subjectcat.subjectcat like 'Marine & Freshwater Biology' or subjectcat.subjectcat like 'Materials Science, Biomaterials' or subjectcat.subjectcat like 'Mathematical & Computational Biology' or subjectcat.subjectcat like 'Medical Ethics' or subjectcat.subjectcat like 'Medical Informatics' or subjectcat.subjectcat like 'Medical Laboratory Technology' or subjectcat.subjectcat like 'Medicine, General & Internal' or subjectcat.subjectcat like 'Medicine, Legal' or subjectcat.subjectcat like 'Medicine, Research & Experimental' or subjectcat.subjectcat like 'Microbiology' or subjectcat.subjectcat like 'Microscopy' or subjectcat.subjectcat like 'Mycology' or subjectcat.subjectcat like 'Nanoscience & Nanotechnology' or subjectcat.subjectcat like 'Neuroimaging' or subjectcat.subjectcat like 'Neurosciences' or subjectcat.subjectcat like 'Nursing' or subjectcat.subjectcat like 'Nutrition & Dietetics' or subjectcat.subjectcat like 'Obstetrics & Gynecology' or subjectcat.subjectcat like 'Oncology' or subjectcat.subjectcat like 'Ophthalmology' or subjectcat.subjectcat like 'Ornithology' or subjectcat.subjectcat like 'Orthopedics' or subjectcat.subjectcat like 'Otorhinolaryngology' or subjectcat.subjectcat like 'Parasitology' or subjectcat.subjectcat like 'Pathology' or subjectcat.subjectcat like 'Pediatrics' or subjectcat.subjectcat like 'Peripheral Vascular Disease' or subjectcat.subjectcat like 'Pharmacology & Pharmacy' or subjectcat.subjectcat like 'Physiology' or subjectcat.subjectcat like 'Plant Sciences' or subjectcat.subjectcat like 'Psychiatry' or subjectcat.subjectcat like 'Psychology' or subjectcat.subjectcat like 'Psychology, Applied' or subjectcat.subjectcat like 'Psychology, Biological' or subjectcat.subjectcat like 'Psychology, Clinical' or subjectcat.subjectcat like 'Psychology, Developmental' or subjectcat.subjectcat like 'Psychology, Educational' or subjectcat.subjectcat like 'Psychology, Experimental' or subjectcat.subjectcat like 'Psychology, Multidisciplinary' or subjectcat.subjectcat like 'Psychology, Social' or subjectcat.subjectcat like 'Radiology, Nuclear Medicine & Medical Imaging' or subjectcat.subjectcat like 'Reproductive Biology' or subjectcat.subjectcat like 'Respiratory System' or subjectcat.subjectcat like 'Rheumatology' or subjectcat.subjectcat like 'Robotics' or subjectcat.subjectcat like 'Social Sciences, Biomedical' or subjectcat.subjectcat like 'Soil Science' or subjectcat.subjectcat like 'Spectroscopy' or subjectcat.subjectcat like 'Sport Sciences' or subjectcat.subjectcat like 'Substance Abuse' or subjectcat.subjectcat like 'Surgery' or subjectcat.subjectcat like 'Toxicology' or subjectcat.subjectcat like 'Transplantation' or subjectcat.subjectcat like 'Tropical Medicine' or subjectcat.subjectcat like 'Urology & Nephrology' or subjectcat.subjectcat like 'Veterinary Sciences' or subjectcat.subjectcat like 'Virology' or subjectcat.subjectcat like 'Zoology');
```

```
drop table if exists biopharmarticle;
create table biopharmarticle as
select distinct biocatarticles.biocatarticleid as "biopharmarticleid"
from biocatarticles, article, article_has_keyword, keyword, article_has_keywordplus, keywordplus
where biocatarticles.biocatarticleid=article.ut_aid
and article.ut_aid=article_has_keyword.article_ut_aid
and article_has_keyword.keyword_idkeyword=keyword.idkeyword
and article.ut_aid=article_has_keywordplus.article_ut_aid
and article_has_keywordplus.keywordplus_idkeywordplus=keywordplus.idkeywordplus
and (article.ab_abstract like '%pharm%' or article.ab_abstract like '%cyto%' or article.ab_abstract like '%immuno%' or article.ab_abstract like '%DNA%' or article.ab_abstract like '%gene%' or article.ab_abstract like '%ribo%' or article.ab_abstract like '%nucleic%' or article.ab_abstract like '%gluc%' or article.ab_abstract like '%amino%' or article.ab_abstract like '%receptor%' or article.ab_abstract like '%cell%' or article.ab_abstract like '%drug%' or article.ti_title like '%pharm%' or article.ti_title like '%cyto%' or article.ti_title like '%immuno%' or article.ti_title like '%DNA%' or article.ti_title like '%gene%' or article.ti_title like '%ribo%' or article.ti_title like '%nucleic%' or article.ti_title like '%gluc%' or article.ti_title like '%amino%' or article.ti_title like '%receptor%' or article.ti_title like '%cell%' or article.ti_title like '%drug%' or keyword.keyword like '%pharm%' or keyword.keyword like '%cyto%' or keyword.keyword like '%immuno%' or keyword.keyword like '%DNA%' or keyword.keyword like '%gene%' or keyword.keyword like '%ribo%' or keyword.keyword like '%nucleic%' or keyword.keyword like '%gluc%' or keyword.keyword like '%amino%' or keyword.keyword like '%receptor%' or keyword.keyword like '%cell%' or keyword.keyword like '%drug%' or keywordplus.keywordplus like 'cytotoxicity' or keywordplus.keywordplus like 'immunoassay' or keywordplus.keywordplus like 'glucose' or keywordplus.keywordplus like 'antibody' or keywordplus.keywordplus like 'single-molecule' or keywordplus.keywordplus like 'layered double hydroxides' or keywordplus.keywordplus like 'Ascorbic acid' or keywordplus.keywordplus like 'alpha-cyclodextrin' or keywordplus.keywordplus like 'assay' or keywordplus.keywordplus like 'expression' or keywordplus.keywordplus like 'amplification' or keywordplus.keywordplus like 'poly(acrylic acid)' or keywordplus.keywordplus like 'titanium-dioxide films' or keywordplus.keywordplus like 'cadmium-sulfide' or keywordplus.keywordplus like 'block copolymers' or keywordplus.keywordplus like 'glucose-oxidase' or keywordplus.keywordplus like 'anatase TiO2' or keywordplus.keywordplus like 'beta-cyclodextrin' or keywordplus.keywordplus like 'recombination' or keywordplus.keywordplus like 'micellization' or keywordplus.keywordplus like 'Sol-gel' or keywordplus.keywordplus like 'TiO2 films' or keywordplus.keywordplus like 'nanocrystalline tio2' or keywordplus.keywordplus like 'acrylamide' or keywordplus.keywordplus like 'fluorescence probes' or keywordplus.keywordplus like 'paste electrodes' or keywordplus.keywordplus like 'triton x-100' or keywordplus.keywordplus like 'oxidase' or keywordplus.keywordplus like 'horseradish-peroxidase' or keywordplus.keywordplus like 'binding' or keywordplus.keywordplus like 'photodegradation' or keywordplus.keywordplus like 'DNA hybridization');
```

These queries generate a comprehensive set of nanobio publications and reach into the realm of nanomaterials that while not described explicitly as bio find frequent application in nanobio.

From these tables we narrow the dataset into US-associated papers published between 2001 and 2010. The result is a list of all nanobio scientists publishing papers between 2001-2010 and the times their work has been cited by only by other nanobio papers during the same time period along with their addresses. This produced a list of 29,647 scientists who wrote 36,737 articles cited a total of 464,751 times, at an average rate of 12.7 cites per paper.

The names for the following report were selected in the following way:

Querying our nanotechnology literature database for US nanobio publications, we selected author IDs for articles listed as US papers that fell under a nanobio selection query and matched any of the following criteria:

- Ranked list of author IDs, top 300 US Nanobio by times cited
- Ranked list of author IDs, top 300 US Nanobio by number of publications
- Ranked list of author IDs, top Nanobio by quality

Further, we filtered out all author IDs with less than 5 papers. This procedure yielded 553 author IDs.

We then performed a disambiguation procedure where we:

- expanded the list of 553 author IDs by finding all other author IDs attached to names highly similar to the names associated with the 553 author names;
- retrieved addresses and papers associated with this expanded list and grouped all results by the original 553 author IDs; and
- inspected each address and each paper in every group to ensure that the author ID matched the targeted US nanobio researcher, discarding all author IDs of authors who were not the same person but had similar names.

Finally we re-ranked our results (174 names) in the following way:

- we recalculated times cited (TC), number of articles (NP), and quality (QA) for the disambiguated individuals, again filtering out all results having fewer than 5 publications.
- we assigned a rank score for each of the three quantities, 1 for the highest in that particular measure, to 174 for the lowest;
- we used the average of these three rank scores to re-rank all names. We named this measure Overall Rank, or OR.

This yielded a quantitative measure of top people in nanobio research publications.

In order to control for people whose impact is derived through their associations with other more impactful scientists not captured through our initial screen process, we took our ranked list of names and associated author IDs in our database and generated a co-authorship network, generating a new ranked list of co-authors ranked in this way by their total authorships. We generated a ranked list based total number of all nano publications associated with nanobio authors (themselves or their non-bio-nano coauthors). We then used this top 50 list to reduce the list of 174 top nanobio authors. This resulted in 24 nanobio authors, only four of which were not in the top 50 of the original 174:

- Paras Prasad (52 overall)
- Zygmunt Gryczynski (79)
- Ken-Tye Yong (145)

- Ya-Ping Sun (77?)

Appendix B: USPTO nanobio patent query method

As a general rule, our preferred technique is to rely first on existing human classification systems when identifying nanobio-relevant records in a specific data set before resorting to developing lists of keywords for identifying nanobio-relevant records. For the research literature, where there is no nanobiotechnology class assigned to the literature metadata records by library information classification specialists, we use a list of keywords that yield high precision and recall in retrieving nanobio-relevant papers. In the case of United States Patent & Trademark Office (USPTO) patent records, the USPTO classifies all patent records. More importantly, the USPTO classifies patent records in a way that allows us to construct a query that yields patents both bio-related and nano-related. We can perform this query by using the classification system, avoiding keywords altogether.

The query strategy for the construction of nanobio patent stars has two requirements, one based on when the patent was granted, and the other based on a patent's classification by the USPTO.

The query first requires that the patent was granted by the USPTO after 12/31/1999. The second requirement is that the patent is classified by the USPTO in either of the following two ways:

1. the patent was classified as 977, and the patent additionally received at least one class/subclass designation indicating bio- or medically relevant content; or
2. the patent received one of the 977 subclasses that also qualifies as bio- or medically relevant.

This means, essentially, that every nanobio patent must be Class 977, and that it have a secondary class or subclass assignment that indicates biological-related or medical-related relevance.

For example, any patent receiving the class/subclass designation 977/924 qualifies as nanobio without any additional class or subclass assignment. But, any patent that is classified as 71 would also need to be classified as 977 to qualify as nanobio. Similarly, any patent classified as 428/828 would additionally need to be classified as 977 to qualify as nanobio.

The USPTO classification scheme used was the version edited by the USPTO on 08/11/2011 at 21:44:05. The two lists of classifications we provided were revised 16 Aug 2012, 01 Sept 2012, 19 Nov 2012, and 23 Jan 2013. A complete list of USPTO classes that help us determine nanobio or nanomedicine relevance are available upon request.

Using the USPTO with the Classification Scheme

The USPTO Advanced Query Interface (http://patft.uspto.gov/netahtml/PTO/search-adv.htm) was used to perform the patent search using the above classification lists. The USPTO Advanced Query Interface limits the length of each query, so in order to query using the entire classification scheme, 10 separate queries had to be used, and then duplicate records were removed from the set produced by the 10 queries.

Appendix C: Co-Authorship Linkages between Literature Stars and Scientists at Firms in the US

Below are the co-authorship linkages between Literature Stars and Scientists at Firms in the US depicted in Figure 1. Firm names italicized and in red are founded by one of the Lit Stars. Names of Stars indicated by (*) are both Literature Stars and NCI Stars.