FROST & SULLIVAN

50 Years of Growth, Innovation and Leadership

Big Science ▸ Big Data ▸ Big Collaboration... …Cancer Research in a Virtual Frontier

A Frost & Sullivan
White Paper

Virginia A. Cardin, Dr.P.H.

www.frost.com

CONTENTS

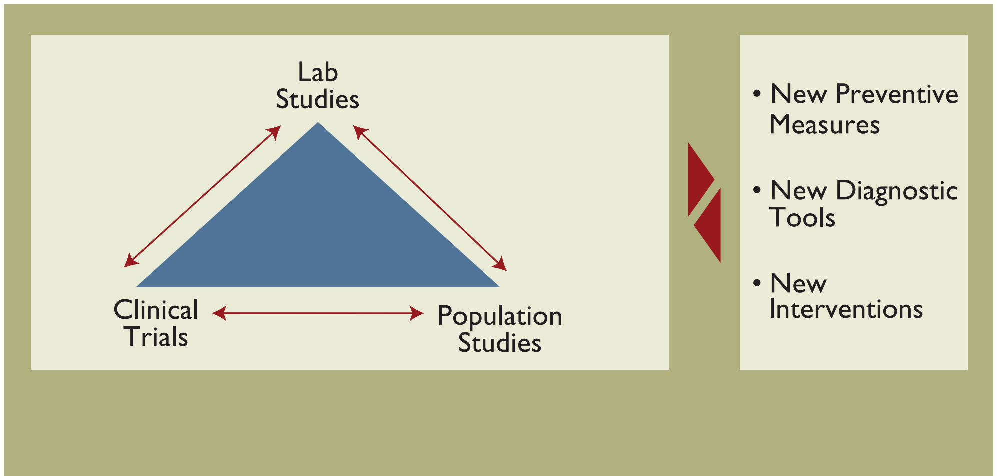
ABSTRACT

Public and private organizations engaged in cancer research are moving into the world of harvesting and managing "big data"–massive amounts of information derived from dry and traditional wet lab investigations, feasibility studies and clinical trials. Research silos are evaporating with the merging of scientific methods. In the future, traditional hypothesis-testing studies will couple with Big Science (data-driven research), and research findings will beget new refined research questions. As evidence accumulates, personalized medicine will become a reality, and patient-specific cancer interventions will become available. The catalyst will be Big Collaboration, teams of oncology specialists, researchers and bioinformaticians working in concert in a virtual frontier.

In cancer research, the phenomenon of Big Science Big Data Big Collaboration is embedded in translational cancer research, the global commitment to transform the ever-evolving continuum of scientific findings into clinical applications. The patient is at its core. From a public health perspective, the goal is to reduce cancer incidence, mortality and morbidity. On an individual level, the objective is to beat cancer.

This white paper will explore critical tools and mechanisms that accelerate knowledge discoveries along the translational cancer research continuum. What are the value propositions of these resources in a virtual frontier, and what's next?

TRANSLATIONAL CANCER RESEARCH



Source: Frost & Sullivan adapted from Translational Research Working Group (National Cancer Institute, 2011)

Cancer is a generic description of more than 100 conditions related to the uncontrolled growth of cells in the body. Whether inherited or induced by environmental factors, the growth of these atypical cells is associated with multiple

Translational science has evolved from being a linear process to being more of an iterative, rational process of approaching a problem—from bench-to-bedside-and back. (Eckhardt 2011) and specific changes at the DNA level, which are not completely understood. Simply put, we do not know the natural history of cancer. We may not need to.

Translational cancer research (TCR) is a pragmatic concept that recognizes the spectacular growth in our knowledge of basic biology and the movement of discoveries from bench research through epidemiologic and clinical studies to preventions and interventions approved for patient use. High-performance computing has enabled gene sequencing to assume a unique place in bench research—identification of relationships and pathways to understand cancer-specific molecular defects (Mathew 2007). Understanding these defects, in turn, will lead to the development of improved preventions and interventions to match each patient's genetic and clinical profile.

"The key to improving the quality of healthcare is to use large data sets to quickly recommend the best possible treatment for each patient....but, unless you have a high-performance computational platform that can efficiently combine this data into a scalable EDW, that's virtually impossible."

—Denise Krause, Ph.D. Director, Research & Education IT University of Mississippi Medical Center

In practice, TCR is a global activity in early stage adoption. Stakeholders include governments, non-profit organizations, private industry, the research community and consumers. The principal driver is the level of funding required to support cancer research and demonstrate safety, efficacy and comparative effectiveness. In the United States, the National Cancer Institute alone has funded more than $100 billion in cancer research over the past 30 years, an investment of $275/person (Niederhuber 2008). In the United Kingdom, the 22-member organization of the National Cancer Research Institute funded £500m in research in 2010, twice what was spent in 2000 (National Cancer Research Institute 2011).

Relative to the costs are the inherent operational inefficiencies and redundancies in conducting research studies in silos. The Food and Drug Administration (FDA) estimates that research and development costs of a new drug are $0.8 billion to $1.7 billion (FDA 2004). There is an 8 percent chance that a new drug in a Phase I clinical trial will be approved and marketed (FDA 2004, Gilbert 2003). Failures are most often due to poor target (patient) selection or poor candidate drug selection. Specifying genome profile characteristics in addition to patient demographics and clinical history could substantially impact the success rate.

Roadblocks to full adoption include:

- Splintered multidisciplinary research silos
- Fundamental logistical challenges underlying TCR core elements described as Big Science, Big Data and Big Collaboration
- Absence of virtual environments to facilitate collaborative research, regardless of stage in the TCR continuum in sentence on solution
- Costs to develop and maintain data repositories that meet the system demands of data explosion, accession and movement
- Absence of an international, uniform research platform

- Absence of fail-safe data security mechanisms within the TCR continuum
- Patient privacy and potential misuse of patient data

While seemingly daunting tasks, stakeholders are partnering to remove these roadblocks. As an old African proverb teaches, one eats an elephant one bite at a time. In cancer research, an initial bite is resolving the logistical challenges related to Big Science Big Data Big Collaboration.

Big Science

In TCR, Big Science describes those bench studies that require massive amounts of data from which to draw inferences. These include genomic studies (DNA gene sequencing) and proteomics (the large-scale study of all the proteins).

In TCR gene sequencing there are two objectives. The first is to image and then "interrogate" human genomes to extract meaningful patterns in cancer genomes. The second is to integrate the findings with other biological and clinical data "to make sense of what it all means" (Tavare 2010). Identifying meaningful patterns has been equated to looking for the elusive needle in a haystack.

Just think of the sheer magnitude of raw data collected and then stored. Each individual's genome has 3 billion base pairs. In full gene sequencing, this would translate into an average of 1.5 gigabytes of data (2 bits per pair or 12 billion bits). On average, scientists can fully sequence the genome of 167 individuals per week, generating 250 gigabytes of images or 200 movie files (Yurkiewicz 2011). Identifying patterns expands the data by a factor of 10 to 20.

Storage of Big Science data is not an issue. Currently, clinical and research data are stored in petabytes; i.e., the number one followed by 15 zeros (Bollier 2010). Future storage may be in yottabyte (1 trillion terabytes).

Inherent challenges relate to the ability to:

- Store raw data from disparate systems seamlessly
- Store data and findings from early discovery through clinical trials to bridge the gap between life science and medical research
- Access data, in real time, by multiple users
- Retrieve data (magnetic pull) in a timely fashion
- Scale up storage capacity to accommodate data growth from next-generation sequencing

> *"No one can predict how these large repositories of information will be used on any given day, so the strategy is to make as much compute and analysis resource available to users as possible."*
>
> *—Chuck Hollis (Latamore 2011)*

- Real time security of data integrity, access and use with immediate countermeasures if a breach is detected

- The objective is to have a complete storage infrastructure for managing Big Data regardless of environment—private or public—that is fail-safe

Big Data

EMC Perspective on Big Data

–Build use cases that drive next generation business intelligence and analytics.
–Create a conceptual architecture for a more agile data platform.
–Develop a vision for applying cloud capabilities.

Big Science generates dimensions of data points and high-resolution images to be deciphered and decoded. In cancer research, Big Data often require on-demand Big Compute across settings using a private cloud, a public cloud or mix of the two. Bioinformatics has gained prominence as a principle R&D discipline, merging biology, mathematics, physics and computer science to carve the elephant into bite-size chunks.

Essentially, bioinformaticians use high-performance computing (HPC) methods to navigate through dimensions of data to study how information is represented and transmitted in biological systems (Bergeron 2002) and analytics to define the questions to ask. The goal is to identify patterns, decipher relationships and offer cause and effect inferences for future analysis. For example, in gene sequencing, analytic methods are devised to look for the key genetic alterations that spur uncontrolled cellular growth; i.e., confer a growth advantage (Tavare 2010). Experts in the field suggest that bioinformatics can help build a holistic perspective of cancer at the systems level that will shape future clinical practice (Bergeron 2002).

HPC is the brain of bioinformatics, which synthesizes component lobes (networking, transmission/bandwidth, databases, visualization techniques, data mining, modeling and simulation, and artificial intelligence (AI)) and related pattern matching.

- For cancer researchers, the pivotal scientific question becomes: “Are these cellular aberrations ‘driver events’ critical to cancer progression or ‘passenger events’ that have no influence?”

- For bioinformaticians, the pivotal logistical question is: “Do you move HPC to the data or do you move the data to HPC?”

- For the patient, the pivotal question is: “How will my genetic and medical profiles be secured, and who will have access to the findings?”

Inherent challenges relate to:

- First and next generation sequencing data

- Absence of industry HPC standards

- Apriori data mining assumptions (what to keep and what to discard)

- Real-time data mining of gene sequencing or microarray results to assess data value, error rate and relevance to data from previous studies (true differences versus background noise)

- Non-standard analytics and computational power (mapping, regression, link analysis, segmentation or deviation detection)

- Continual influx of new digital streams of data from MRIs and biological sensors

- Public vs. private cloud computing

Big Collaboration

Big Science generates Big Data that demands Big Collaboration to beat cancer. Algorithms and AI enable the discovery of many of the underlying rules and relationships in biological data. Analytics enable researchers to intersect data sets that are too small to fully understand cancer subtypes and that may have measured expression set on different platforms (Russ 2010). Notwithstanding, collaborative expertise—human intelligence and intuition—are required for meaning and interpretation (Bergeron 2002).

Within the continuum of TCR, Big Collaboration extends beyond the borders of traditional research collaboration to include on-demand communication and sharing of protocols, electronic resources, data, and findings among the spectrum of stakeholders in a private or public virtual frontier. At each stage, learning and leveraging information and findings on what hasn't worked is as important as knowing what has worked. Big Collaboration embraces the "from bench to bedside" philosophy and recognizes time, manpower and R&D dollars are scarce resources.

Inherent macro challenges:

- Adoption by the scientific community
- Adoption by government, non-profit and private funding organizations
- Potential conflict between traditional silo researchers and those embracing Big Collaboration
- Compatible technologies and cloud infrastructures
- IT management of groups with different tools, requirements and expectations
- Ownership of data
- Government regulations and policies

Inherent micro challenges:

- Absence of a shared understanding of the organizational themes of the databases by creator and users; lack of consensus on data capture, analytics and communication

- Accessible data repositories and lack of transparency in findings
- Resources to support bioinformatics
- Patient privacy

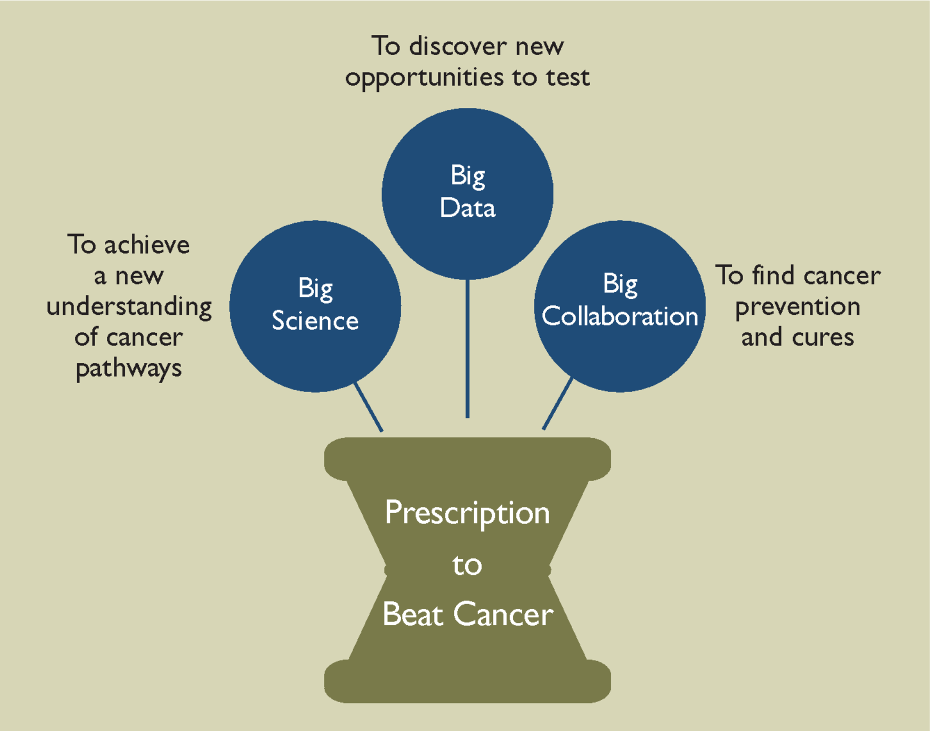
THE PRESCRIPTION TO BEAT CANCER

Big Technology

Frost & Sullivan defines "Big Technology" as an enabler that wires R&D plasticity into TCR to accelerate technology discovery to the patient. In today's cancer research environment, the critical outcome measures are time and availability of target preventions and interventions to patients.

Plasticity refers to the ability to change. R&D plasticity refers to the ability to build on early discovery research to make later stage discovery and applied research more productive and less costly. It increases the probability of successful innovations.

Big Technology will enable the virtual movement of the scientist to the data rather than the data to the scientist. As such, Big Technology provides kinetic energy to the R&D cancer community as life science and medical research merge. To those engaged in Big Science Big Data Big Collaboration activities, Big Technology will continue to fuel bioinformatics that will lead to the prescription to beat cancer.


To discover new opportunities to test
Big Data
To achieve a new understanding of cancer pathways
Big Science
Big Collaboration
To find cancer prevention and cures
Prescription to Beat Cancer

CASE STUDY SOLUTIONS

BIG SCIENCE

Lipper Center for Computational Genetics, Harvard Medical School

Challenge:
The Lipper Center needed to establish a seamless data infrastructure (central repository) for the Personal Genome Project, early discovery research of medical and non-medical data and genetic and trait data for 100,000 volunteers who agreed to publicly share their data to advance the use of genomincs to further understand health and disease (Isilon Press Release 2009).

Solution:
The Isilon OneFS Scale-out Storage Platform was selected for the operational and performance efficiencies it afforded to researchers, regardless of location, to meet the insatiable demand for data inherent in the project design. Once racked, a Network-attached Storage (NAS) cluster can be online in less than 10 minutes without additional integration services that are cost- and time-intensive.

Value:
About 13,000 individuals volunteered to participate in the project one month after it was launched in 2009.The scalability of the Isilon system has been an asset.

Oklahoma Medical Research Foundation (OMRF)

Challenge:
OMRF experienced intense data explosion with the addition of an Illumina next generation genome analyzer and server virtualization. The foundation needed a cost-effective data storage solution.

Solution:
Isilon IQ OMRF and Isilon Sync IQ unified the Foundation's DNA pipeline and virtual environment. Migrating data between sites was no longer necessary.

Value:
IT infrastructure was simplified and scaled to meet on-demand performance requirements, including 24/7/365 data protection and recovery, without costly upgrades.

BIG DATA

Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)

Jointly funded by the British Columbia Cancer Foundation and the Cancer Research UK (CRUK), Cambridge Research Institute

Challenge:
Bioinformaticians (computational biologists) are "interrogating" the genomic and transcriptional landscape of 2,000-plus clinically annotated breast tumor specimens to profile subtypes of breast cancer, based on molecular characteristics.

Solution:
Custom analytics have been designed and revised as appropriate, combining meausrements of about 2 million probes querying the genome during the sequencing process.

Value:
New subtypes of breast cancer have been defined, and the analytical methods are applicable to other cancers.

Major Research University

Challenge:
Investigators lacked a centralized view of research data. As a result, they were unable to combine decades of clinical data with molecular and genomic data due to the high costs for HPC.

Solution:
Greenplum's scale-out architecture was implemented, which eliminated the need to build separate data marts for each new research project while streamlining data and business processes.

Value:
Costs of internal computing services have been significantly reduced, enabling the medical center to win more grants and attract more talent. As a by-product, researchers have also been able to identify larger pools of study participants.

> *"With Greenplum's Big Data Analytics platform, we're able to accommodate massive data volumes in a single cluster. The technology also integrates with the i2b2 platform to help us build quick and accurate cohorts for our studies."*
>
> *—Denise Krause, Ph.D. Director, Research & Education IT*

BIG COLLABORATION

Informatics for Integrating Biology and the Bedside (i2b2)

Challenge:
NCI-funded center charged with developing a scalable informatics framework for genomic data that could be used to faciltate clinical trials investigating personalized therapies for individuals with diseases having genetic origins.

Solution:
Partners Healthcare System researchers developed prototype Clinical Research Chart and associated stand-alone analytical toolkits for seven genetic conditions.

Value:
Software and toolkits are open access, have been adopted internationally, and adapted to target and expedite patient recruitment for clinical trials in cancer research.

National Cancer Institute/Cancer Therapy Research Program

Challenge:
Twenty percent of paper documents that need to be completed at the initiation of clinical trials are shipped by courier service or shipped overnight, taking three to five days per signature.

Solution:
Pilot study is under way to demonstrate time and cost savings using digital identities authentification and electronic signatures on documents that initiate clinical trials and require multiple signatures from researchers in multiple sites.

Value:
Preliminary findings have demonstrated significant cost and time savings by leveraging electronic business practices in the cloud. Phase 2 and 3 studies are planned to determine time and cost savings in larger, more complex clinical trials.

REFERENCES

American Cancer Society. (2010) "The Global Economic Cost of Cancer." Available at: http://www.cancer.org/acs/groups/content/@internationalaffairs/documents/document/acspc-026203.pdf
Accessed September 12, 2011

Bergeron B. (2002) "Bioinformatics Computing." New York: Prentice Hall.

Bollier D. (2010) "The Promise and Peril of Big Data." Washington, D.C.: The Aspen Institute.

Eckhardt SG. (2011) "Eckhardt: The new translational cancer research." Available at: http://www.coloradocancerblogs.org/target-cancer/eckhardt-newtranslationalresearch
Accessed on September 7, 2011.

FDA. (2004) "Challenge and opportunity on the critical path to new medical products." U.S. Food and Drug Administration. Available at: http://www.fda.gov/ScientificResearch/SpecialTopics/CriticalPathInitiative
Accessed on October 3, 2011.

Gilbert J., Henske P. and Singh. (2003) "Rebuilding big pharma's business model." In Vivo, the Business & Medicine Report. Windover Information. November 2003; 21(10)

Hollis C. (2011a) "Behold The New User." Chuck's Blog. Available at: http://chucksblog.emc.com/chucks_blog/2011/05/behold-the-new-user-.html
Accessed on September 14, 2011.

Isilon Press Release. (2009) "The personal genome project deploys Isilon IQ storage to transform genomic research." Available at: http://www.isilon.com/press-release/personal-genome-project-deploys-isilon-iq-storage-transform-genomic-research
Accessed on September 27, 2011.

Latamore B. (2011) "Big data is where problem becomes opportunity says EMC's Chuck Hollis." Available at: http://wikibon.org/wiki/v/Big_Data_is_Where_Problem_Becomes_Opportunity_Says_EM C's_Chuck_Hollis
Accessed on October 5, 2011.

Lewis L. (2006) "Mining cancer. Data analysis needs standards too." 2nd NCRI Conference in Birmingham. Available at: www.cancerinformatics.org.uk/Documents/.../Paul%20Lewis%202.ppt
Accessed on October 3,2011.

Lucas M. (2009) "13,000 offer up DNA to put their genomes online." Computerworld Available at:
http://www.computerworld.com/s/article/9133167/13_000_offer_up_DNA_to_put_their_genomes_online?taxonomyId=19&pageNumber=2
Accessed on October 3, 2011.

Mathews JP, Taylor BS, Bader GD, et al. (2007) "From bytes to bedside: Data integration and computational biology for translational cancer research." PloS Computational Biology 3:2, 153-162 (February 2007). Available at:
www.ploscompbio.org

National Cancer Institute. (2011a) Translational Research Working Group. TRWG Definition of Translational Research. Available at:
http://www.cancer.gov/researchandfunding/trwg/TRWG-definition-and-TR-continuum
Accessed on October 3, 2011.

National Cancer Institute. (2011b) "Research collaboration in the cloud: How NCI and research partners are using interoperable digital identities, digital signatures and cloud computing to accelerate drug development." Available at:
http://www.safe-biopharma.org/infocenter/whitepaper_ResearchCollaborationinTheCloud.pdf
Accessed on September 24, 2011.

National Cancer Research Institute. (2011) "Research spend on cancer doubles within a decade and the most fatal cancers see investment." Press release, Wednesday 29 June2011. Available at:
http://www.ncri.org.uk/includes/Publications/pressreleases/2011NCRIPressRelease_RESEARCH_SPEND_DOUBLES.pdf
Accessed on September 21, 2011.

Niederhuber JE. (2008) Director's Update. The future of cancer research. What's at stake? National Cancer Institute Bulletin. May 13, 2008. 5:10. Available at:
http://www.cancer.gov/aboutnci/ncibulletin/archive/2008/051308/page4
Accessed on October 3, 2011.

Organization of European Cancer Institutes. European Economic Interest Grouping. (2009) "A platform for translational cancer research."Available at:
http://www.oeci-eeig.org/Documents/PlatformTranslationalResearch.pdf
Accessed on September 17,2011.

Pizani E. (2010) "Has the internet changed science?" Prospect. 17 November 2010, Issue 177. Available at:
http://www.prospectmagazine.co.uk/2010/11/has-the-internet-changed-science-big-date-hypothesis-driven-science/
Accessed on September 24, 2011.

Price, DJD. (1963) Little Science, Big Science. New York: Columbia University Press.

Russ AP, Aparixi SAJR and Carlton MBL. (2010) Perspective. "Large scale dataset examples." MolEcular Taxonomy of Breast Cancer International Consortium (METABRIC). Available at:
www.cancerinformatics.org.uk/Documents/.../Paul%20Lewis%202.ppt
Accessed on September 17, 2010

Tavare S. (2010) "Data mining the complex cancer landscape." Research Horizons. January2010, Issue 11. Available at:
http://www.research-horizons.cam.ac.uk/articles/print.aspx?id=368
Accessed on September 24, 2011.

Yurkiewicz I. (2011) "Fishing for Funding. Big Data is Changing Science. Are Funding Agencies Keeping Up?"
http://scienceprogress.org/2011/04/fishing-for-funding/
Accessed on September 24, 2011.

Silicon Valley
331 E. Evelyn Ave. Suite 100
Mountain View, CA 94041
Tel 650.475.4500
Fax 650.475.1570

San Antonio
7550 West Interstate 10,
Suite 400,
San Antonio, Texas 78229-5616
Tel 210.348.1000
Fax 210.348.1003

London
4, Grosvenor Gardens,
London SWIW ODH,UK
Tel 44(0)20 7730 3438
Fax 44(0)20 7730 3343

877.GoFrost • myfrost@frost.com
http://www.frost.com

ABOUT FROST & SULLIVAN

Frost & Sullivan, the Growth Partnership Company, partners with clients to accelerate their growth. The company's TEAM Research, Growth Consulting, and Growth Team Membership™ empower clients to create a growth-focused culture that generates, evaluates, and implements effective growth strategies. Frost & Sullivan employs over 50 years of experience in partnering with Global 1000 companies, emerging businesses, and the investment community from more than 40 offices on six continents. For more information about Frost & Sullivan's Growth Partnership Services, visit http://www.frost.com.

For information regarding permission, write:
Frost & Sullivan
331 E. Evelyn Ave. Suite 100
Mountain View, CA 94041

Auckland
Bangkok
Beijing
Bengaluru
Bogotá
Buenos Aires
Cape Town
Chennai
Colombo
Delhi / NCR
Dhaka
Dubai
Frankfurt
Hong Kong
Istanbul
Jakarta
Kolkata
Kuala Lumpur
London
Mexico City
Milan
Moscow
Mumbai
Manhattan
Oxford
Paris
Rockville Centre
San Antonio
São Paulo
Seoul
Shanghai
Silicon Valley
Singapore
Sophia Antipolis
Sydney
Taipei
Tel Aviv
Tokyo
Toronto
Warsaw
Washington, DC